FEDPHM: COMMUNICATION-EFFICIENT PERSONALIZATION THROUGH PARAMETERIZED HYPERCOMPLEX MULTIPLICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Personalized federated learning requires transferable shared information and client-specific adaptation under a limited communication budget. We propose FedPHM, a structured selective transmission framework based on Parameterized Hypercomplex Multiplication (PHM). Specifically, we decompose fully-connected layers into shared basis matrices and client-specific private interaction matrices via Kronecker factorization: shared bases and other shared network parameters are transmitted for aggregation, while private interactions persist locally. Per-step spectral projection bounds the interaction matrices and their influence on shared gradients. Theoretically, we establish a finite-time joint-stationarity bound for the PHM subsystem under partial client participation. With a horizon-dependent common step size, the residual is bounded by $O(T^{-1/2}) + O(\nu_A)$, where ν_A bounds the private-gradient noise variance. An exact quadratic-risk decomposition further separates the approximation cost of shared bases from the benefit of private interactions. Experiments on CIFAR-10, CIFAR-100, and FMNIST demonstrate that FedPHM achieves excellent personalized accuracy with less communication, and is robust under high heterogeneity and low participation scenarios.

1 INTRODUCTION

Federated learning keeps training data on client devices and avoids centralizing raw data, but this privacy-preserving property comes at the cost of repeated model transmission, which often dominates the total training cost (McMahan et al., 2017; Kairouz et al., 2021). In real-world deployments, clients are typically bandwidth-constrained nodes like mobile phones and edge devices. Each round requires many clients to upload full model updates, making communication overhead a central bottleneck for resource-constrained federated learning.

Quantization, sparsification, and low-rank adaptation reduce uplink payload in different ways. They reduce transmitted precision, retain selected coordinates, or restrict the trainable update space. Under non-IID data, communication efficiency must also accommodate client-specific variation. This motivates a structural question: which layer components should clients share, and which should remain local?

Personalized methods already separate shared representations from local parameters (Collins et al., 2021; Li et al., 2021). We study this separation within a layer. Shared basis matrices provide transformations that all clients can reuse. Private interaction matrices determine how each client combines them. This design retains client-specific mixing while communicating a compact common representation.

Motivated by this insight, we propose **FedPHM**, a structured selective transmission framework for personalized federated learning. Leveraging the Kronecker factorization structure of Parameterized Hypercomplex Multiplication (PHM) (Zhang et al., 2021), we naturally decompose fully-connected layer weights into two semantically distinct components: global shared basis matrices that carry transferable generic structures, and client-specific private interaction matrices that encode local adaptation patterns. During training, the shared bases and other shared network parameters are up-

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.

loaded for aggregation, while private interaction matrices persist on local devices. We apply per-step spectral projection to keep private matrices in a bounded spectral-norm ball.

We provide theoretical analysis from optimization and representation perspectives. For optimization, we establish a finite-time joint-stationarity bound for fixed-E Local-SGD under partial client participation. With a horizon-dependent step size, the residual is $O(T^{-1/2}) + O(\nu_A)$, where ν_A is the private stochastic-gradient variance. We also decompose shared-gradient disagreement into data-induced heterogeneity and private-state dispersion. For representation, the PHM block form shows that private interactions select client-specific mixing patterns and can induce full-rank differences between effective weights. Under a quadratic effective-weight model, an exact risk decomposition identifies when the benefit of client-specific mixing exceeds the approximation cost of the shared PHM structure.

Experiments on CIFAR-10, CIFAR-100, and FMNIST evaluate accuracy, communication, client heterogeneity, and participation. On CIFAR-100, FedPHM reaches 54.08% client-matched accuracy with a 2.62 MiB uplink under strong heterogeneity, while FedAvg reaches 34.06% with 9.62 MiB. The advantage of private interactions is largest for strongly heterogeneous partitions and decreases toward homogeneous partitions.

Our main contributions are summarized as follows:

1. We propose FedPHM, a communication-efficient personalized federated learning framework that achieves selective transmission at the parameter structure level. It naturally separates shared generic structures and client-specific personalized components via Kronecker factorization, fundamentally reducing uplink communication cost while retaining client-specific adaptation parameters locally.
2. We establish a finite-time joint-stationarity bound for fixed-E Local-SGD under partial participation, with an $O(T^{-1/2}) + O(\nu_A)$ residual bound. We also derive an exact quadratic-risk decomposition that separates PHM approximation cost from the heterogeneity preserved by private interactions.
3. We validate the method through experiments and ablation studies across three standard datasets, demonstrating its effectiveness and robustness under varying heterogeneity and participation regimes.

2 Related Work

Classic federated learning follows the single shared model paradigm, where clients collaboratively update a global model via FedAvg (McMahan et al., 2017). Non-IID data distributions cause drift in the aggregated model and degrade generalization performance. Prior works mitigate this via proximal regularization (FedProx), control variates (SCAFFOLD), or theoretical analysis of local SGD, all operating within the "one model fits all clients" framework (Li et al., 2020; Karimireddy et al., 2020; Stich, 2019; Khaled et al., 2020).

Personalized federated learning assigns client-specific variables to capture local variation. FedRep learns a shared representation with private prediction heads, while Ditto couples local models to a global reference (Collins et al., 2021; Li et al., 2021). Adaptive aggregation offers another way to control information sharing (Deng et al., 2024). FedPHM places this shared–private separation inside a parameterized layer: clients share its bases and retain their mixing coefficients.

Parameter-efficient federated adaptation also controls which coordinates are trained or transmitted. Recent work studies sparse LoRA activation (Wu et al., 2024), heterogeneous rank allocation (Zhang et al., 2025), and selective decoupled LoRA (Shen et al., 2026). Tensorized adapters and block-wise Kronecker decomposition provide structured alternatives (Ghiasvand et al., 2024; Li et al., 2025). FedPHM builds on PHM (Zhang et al., 2021) and assigns distinct sharing roles to its two factors.

These parameterizations encode different forms of client variation. With a common fixed base, LoRA represents a client weight as $W_0 + U_k V_k^\top$ (Hu et al., 2022). A rank-r update models residual variation in a low-dimensional subspace; two clients' weights differ by rank at most $2r$. PHM instead uses $\sum_i A_{k,i} \otimes S_i$ as the layer weight itself. It reuses transformations across blocks while allowing client-specific mixing throughout the layer. Such differences can be full rank, as shown in

(10). LoRA favors low-rank residual changes around a common base. PHM favors repeated block structure with varying mixing patterns. This distinction motivates the controlled comparisons and the fixed-basis analysis below.

3 FedPHM: Structured Selective Transmission Framework

This chapter details the design of FedPHM and its communication cost properties, covering four core components: structured parameter decomposition, block-wise chained forward computation, local optimization with spectral constraints, and selective transmission with global aggregation.

3.1 Structured Decomposition of Parameters

We decompose each fully connected layer using the Kronecker decomposition structure of Parameterized Hypercomplex Multiplication (PHM) (Zhang et al., 2021), which naturally divides the layer weights into shared base matrices and client-specific interaction matrices.

For a PHM layer of order n with input dimension $n \cdot r_{\text{in}}$ and output dimension $n \cdot r_{\text{out}}$, the effective weight for client k is expressed as

$$W_k = \sum_{i=1}^{n} A_{k,i} \otimes S_i, \quad A_{k,i} \in \mathbb{R}^{n \times n}, \quad S_i \in \mathbb{R}^{r_{\text{out}} \times r_{\text{in}}}. \tag{1}$$

This decomposition itself compresses the size of shared parameters to $1/n$ of that of a standard dense layer, which is the core mechanism for FedPHM to save communication costs. The decomposition structure of PHM is shown in Figure 1.

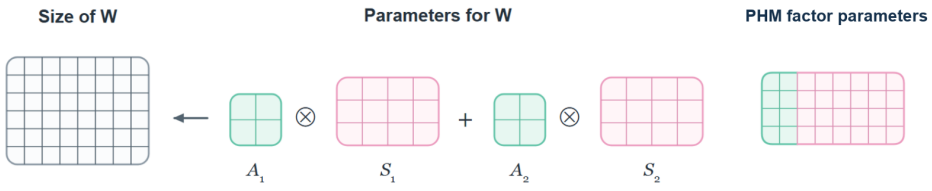


Figure 1: PHM decomposition of a dense fully-connected transformation.

3.2 Block-Wise Chained Forward Computation

Constructing the full effective weight matrix W_k directly from the Kronecker product definition would yield a large matrix of size $n^2 \cdot r_{\text{in}} \cdot r_{\text{out}}$, incurring substantial memory and access overhead. FedPHM adopts **block-wise chained computation** for forward propagation, which avoids instantiating the complete weight matrix. Instead, it yields exactly equivalent outputs using only small-scale matrix operations.

For an input of dimension $n \cdot r_{\text{in}}$, the feature vector is first divided into n input blocks X_b of size r_{in}. Each shared basis matrix S_i is applied to every input block to produce intermediate features. The private interaction matrices then form weighted combinations of these features and concatenate the resulting output blocks into an $n \cdot r_{\text{out}}$-dimensional output. Equivalently, the a-th output block is

$$Y_a = \sum_{i=1}^{n} \sum_{b=1}^{n} (A_{k,i})_{ab} S_i X_b, \qquad a = 1, \ldots, n, \tag{2}$$

where Y_a and X_b denote the a-th output block and the b-th input block. This computation is equivalent to the Kronecker form while using only small $n \times n$ private matrices and $r_{\text{out}} \times r_{\text{in}}$ shared bases.

For one input sample, base mapping costs $O(n^2 \cdot r_{\text{in}} \cdot r_{\text{out}})$ and block mixing costs $O(n^3 \cdot r_{\text{out}})$. The implementation avoids materializing the effective dense matrix and uses two tensor contractions (`einsum`) or batched matrix multiplication.

3.3 Local Optimization and Spectral Projection Constraint

While forward propagation uses block-wise chained computation, backpropagation decouples the gradients of the shared and private parameters through the chain rule. Let X and Y denote the input and output of the PHM layer. The gradient with respect to the effective weight is

$$G = \frac{\partial F_k}{\partial W_k} = \frac{\partial F_k}{\partial Y} X^\top, \tag{3}$$

where G is partitioned into blocks G_{ab}. The gradients of the shared bases and private interactions are respectively

$$\nabla_{S_i} F_k = \sum_{a,b} (A_{k,i})_{ab} G_{ab}, \qquad [\nabla_{A_{k,i}} F_k]_{ab} = \langle G_{ab}, S_i \rangle_F. \tag{4}$$

Automatic differentiation obtains these gradients directly from the block-wise computation, so the dense effective matrix need not be materialized.

Each step of local stochastic gradient descent updates three types of parameters synchronously: non-PHM public parameters, PHM layer shared base matrices, and private interaction matrices. The first two are uploaded to the server at the end of each training round, while private interaction matrices remain local.

3.3.1 Spectral Norm Projection Constraint

To bound private matrices and constrain their impact on shared gradients, we apply spectral projection to all private interaction matrices after each local update step, constraining their spectral norm within a threshold τ. The feasible set is the spectral norm ball

$$\mathcal{C}_\tau = \left\{ X \in \mathbb{R}^{n \times n} : \|X\|_2 \leq \tau \right\}. \tag{5}$$

For a tuple of n interaction matrices, the feasible set is

$$\mathcal{C}_\tau^n = \{(A_1, \ldots, A_n) : A_i \in \mathcal{C}_\tau,\ i = 1, \ldots, n\}.$$

Projection onto this product set acts on each matrix separately.

We implement this projection efficiently via singular value decomposition (SVD). For a matrix $A = U\,\mathrm{diag}(\sigma)\,V^\top$, we clip each singular value to τ and reconstruct the matrix as

$$\Pi_{\mathcal{C}_\tau}(A) = U \mathrm{diag}(\min\{\sigma, \tau\}) V^\top, \tag{6}$$

which bounds the private linear map from shared bases to effective weights.

Each local step projects n interaction matrices of size $n \times n$ through singular-value clipping.

3.4 Selective Transmission and Global Aggregation

FedPHM adopts a **selective transmission mechanism**: only shared base matrices and public network parameters participate in global transmission and aggregation, while private interaction matrices are retained locally on the client side without transmission. This greatly reduces uplink communication overhead while preserving each client's independent personalized adaptation capability.

3.4.1 Single-Round Update Pipeline

Algorithm 1 summarizes the complete single-round update procedure, and the overall data flow is illustrated in Figure 2.

1: **Input:** server state (ω^t, S^t); client state A_k^t and sample count N_k;
   K clients; M participants; E local steps; step size η; radius τ;
2: Sample $\mathcal{K}_t$ uniformly without replacement, $|\mathcal{K}_t| = M$;
3: Broadcast (ω^t, S^t) to all $k \in \mathcal{K}_t$;
4: **for** each participating client $k \in \mathcal{K}_t$ in parallel **do**
5: Initialize $(\omega_{k,0}, S_{k,0}) \leftarrow (\omega^t, S^t)$ and restore $A_{k,0} \leftarrow A_k^t$;
6: **for** $e = 0, 1, \ldots, E-1$ **do**

Sample a local minibatch; perform block-wise PHM forward propagation;
Compute gradients and update shared and private parameters with step size η;
Project each private matrix $A_{k,i}$ onto $\mathcal{C}_\tau$;
end for
Store $A_k^{t+1} \leftarrow A_{k,E}$ locally; upload only $(\omega_{k,E}, S_{k,E})$;
end for
Set $p_k^t = N_k / \sum_{j \in \mathcal{K}_t} N_j$ for $k \in \mathcal{K}_t$;
Aggregate $(\omega^{t+1}, S^{t+1}) \leftarrow \sum_{k \in \mathcal{K}_t} p_k^t (\omega_{k,E}, S_{k,E})$;
For $k \notin \mathcal{K}_t$, keep $A_k^{t+1} \leftarrow A_k^t$.

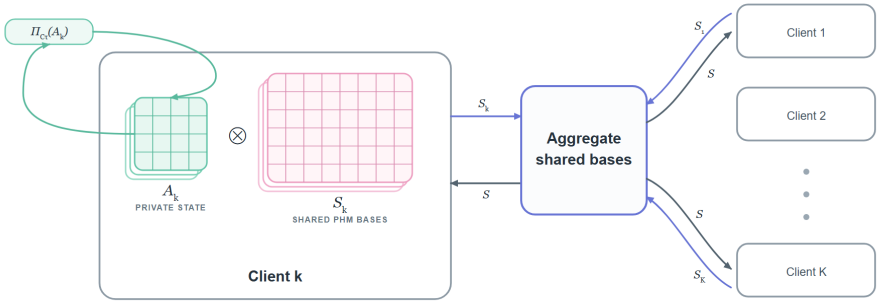


Figure 2: FedPHM single-round communication flow. Shared parameters participate in global transmission and aggregation, while private interaction matrices are retained locally on their corresponding clients.

3.4.2 Communication Overhead Analysis

For a network with L PHM-decomposed layers, let $|\omega|$ denote the total amount of non-PHM shared parameters and b denote the bytes per parameter. Then the uplink communication cost per client per round is

$$C_{\text{FedPHM}} = b \left[|\omega| + \sum_{\ell=1}^{L} n_\ell \cdot r_{\text{out},\ell} \cdot r_{\text{in},\ell} \right]. \tag{7}$$

The corresponding communication cost for a dense fully connected network with the same structure is

$$C_{\text{dense}} = b \left[|\omega| + \sum_{\ell=1}^{L} n_\ell^2 \cdot r_{\text{out},\ell} \cdot r_{\text{in},\ell} \right]. \tag{8}$$

For dimensions divisible by n_ℓ, a PHM-decomposed layer has exactly $1/n_\ell$ as many shared weights as its dense counterpart. With padding, its shared count is $n_\ell \lceil d_{\text{out}}/n_\ell \rceil \lceil d_{\text{in}}/n_\ell \rceil$. The overall network communication compression ratio is jointly determined by the proportion of PHM layers and the order n. This structured reduction lowers the transmitted payload while the local interactions retain client-specific adaptation.

4 Why Shared Bases and Private Interactions Help

The PHM structure reduces the part of a decomposed layer that needs to be communicated. At the same time, private interactions preserve client-dependent variation in the resulting weight space. This section quantifies these two effects and identifies the conditions under which the personalization benefit outweighs the approximation cost of the compact structure.

4.1 Full-Rank Client Differences

For $d_{\text{in}} = n \cdot r_{\text{in}}$ and $d_{\text{out}} = n \cdot r_{\text{out}}$, we use the effective client weight defined in (1). If only the first interaction matrix differs between two clients k and j, then

$$W_k - W_j = \Delta A \otimes S_1, \qquad \Delta A = A_{k,1} - A_{j,1}. \tag{9}$$

Using

$$\operatorname{rank}(\Delta A \otimes S_1) = \operatorname{rank}(\Delta A)\operatorname{rank}(S_1), \tag{10}$$

full-rank ΔA and a maximally ranked S_1 give

$$\operatorname{rank}(W_k - W_j) = n \min(r_{\text{out}}, r_{\text{in}}) = \min(d_{\text{out}}, d_{\text{in}}). \tag{11}$$

Thus, private PHM interactions can generate full-rank differences between effective client weights.

4.2 Comparison with Low-Rank Residuals

Consider the LoRA parameterization with a shared base and a rank-r private residual (Hu et al., 2022):

$$\widehat{W}_k = W_0 + U_k V_k^\top, \qquad U_k \in \mathbb{R}^{d_{\text{out}} \times r}, \quad V_k \in \mathbb{R}^{d_{\text{in}} \times r}. \tag{12}$$

For two clients, (12) gives

$$\widehat{W}_k - \widehat{W}_j = U_k V_k^\top - U_j V_j^\top, \qquad \operatorname{rank}(\widehat{W}_k - \widehat{W}_j) \le 2r. \tag{13}$$

Together with the construction above, this shows that when

$$2r < \min(d_{\text{out}}, d_{\text{in}}), \tag{14}$$

FedPHM can represent client differences that the rank-r residual model cannot represent.

4.3 Representation–Estimation Trade-off

Let

$$\mathcal{S} = \operatorname{span}\{S_1, \dots, S_n\}, \qquad \mathcal{V}_S = \left\{ \sum_{i=1}^{n} A_i \otimes S_i : A_i \in \mathbb{R}^{n \times n} \right\}. \tag{15}$$

Partition W into PHM blocks W_{ab}. Then

$$W \in \mathcal{V}_S \quad \Longleftrightarrow \quad W_{ab} \in \mathcal{S}, \quad \forall a, b. \tag{16}$$

Thus, all blocks use the same shared block space, while the interaction coefficients choose different combinations for different clients.

Let $W_k^\star$ be the local optimum of client k for the target fully connected layer. Define

$$P_k = \Pi_{\mathcal{V}_S}(W_k^\star), \qquad R_k = W_k^\star - P_k. \tag{17}$$

Here P_k is the part represented by the shared PHM bases, and R_k is the unrepresented part. Let

$$E_S = \frac{1}{K} \sum_{k=1}^{K} \|R_k\|_F^2, \qquad \bar{P} = \frac{1}{K} \sum_{k=1}^{K} P_k, \qquad D_S = \frac{1}{K} \sum_{k=1}^{K} \|P_k - \bar{P}\|_F^2. \tag{18}$$

Here E_S is the PHM representation error and D_S is the client variation preserved inside the PHM subspace.

For the quadratic loss

$$\ell_k(W) = \frac{1}{2} \|W - W_k^\star\|_F^2, \tag{19}$$

the two joint risks are

$$\begin{aligned} \Phi^\star_{\text{private}}(S) &= \min_{P_1, \dots, P_K \in \mathcal{V}_S} \frac{1}{K} \sum_{k=1}^{K} \ell_k(P_k), \\ \Phi^\star_{\text{shared}}(S) &= \min_{P \in \mathcal{V}_S} \frac{1}{K} \sum_{k=1}^{K} \ell_k(P). \end{aligned} \tag{20}$$

The first risk lets each client use its own PHM interaction. The second forces all clients to use one common interaction. Orthogonal projection and the centroid identity give

$$\Phi^{\star}_{\text{private}}(S) = \frac{1}{2}E_S, \qquad \Phi^{\star}_{\text{shared}}(S) = \frac{1}{2}(E_S + D_S), \tag{21}$$

and hence

$$\Phi^{\star}_{\text{shared}}(S) - \Phi^{\star}_{\text{private}}(S) = \frac{1}{2}D_S. \tag{22}$$

For LoRA-private-base, let each client store a full private base. Its private parameter count is

$$p_{\text{LoRA}} = d_{\text{out}}d_{\text{in}}, \tag{23}$$

whereas FedPHM stores

$$p_{\text{PHM}} = n^3 \tag{24}$$

private interaction parameters. The full private base has zero approximation error in this quadratic model, while FedPHM has approximation error $E_S/2$. Under the local linear estimation model

$$\Phi_{\text{est}}(p) \approx \sigma^2 \frac{p}{m}, \tag{25}$$

with client sample size m and noise scale σ^2, the total risks are

$$\Phi_{\text{PHM}} \approx \frac{1}{2}E_S + \sigma^2\frac{n^3}{m}, \qquad \Phi_{\text{LoRA}} \approx \sigma^2\frac{d_{\text{out}}d_{\text{in}}}{m}. \tag{26}$$

Therefore, FedPHM is favored when

$$E_S < \frac{2\sigma^2}{m}\left(d_{\text{out}}d_{\text{in}} - n^3\right). \tag{27}$$

This condition states the trade-off between PHM representation error and the larger private estimation cost of LoRA-private-base.

5 Experimental Results

This section systematically validates the effectiveness of FedPHM through benchmark comparisons and ablation studies. We focus on three core claims: high personalized accuracy with low communication cost under non-IID settings, performance gains from structural design rather than parameter scale, and alignment of component properties with theoretical predictions.

5.1 Experimental Setup

Datasets and federated setup. Core experiments use CIFAR-100 with 50 clients. In each round, 10 clients are uniformly sampled without replacement for training. We run 200 communication rounds in total, with 20 local SGD steps per round and a batch size of 64. Non-IID data partitions are generated via Dirichlet distribution with heterogeneity parameters $\alpha \in \{0.1, 0.5\}$. Core comparisons use five seeds and report mean and standard deviation; explicitly labeled legacy ablations retain their original three-seed protocol. Supplementary validation is performed on CIFAR-10 and FMNIST. Clients retain their private state across their participating rounds. The theoretical analysis isolates the uniform-client objective; the implementation aggregates selected-client shared states by sample weight.

Architecture and initialization. The CIFAR-100 backbone has three 3×3 convolutional blocks with channels 64, 128, and 256, each followed by GroupNorm, ReLU, and 2×2 max pooling. Its classifier is $4096 \to h \to 100$. The dense reference uses $h = 512$, while compact FedPHM uses $h = 512$ for $n = 8$ and $h = 620$ for $n = 10$ to align upload budgets. FedPHM replaces only the first classifier transformation. In the dense reference, the $4096 \to 512$ matrix contains 2,097,152 parameters, or 83.2% of the 2,520,676-parameter model. We initialize S_i with Kaiming-uniform initialization, $A_{k,i}$ as $I_n/n + 0.01Z_i$ with standard-normal Z_i, and biases at zero. For orders that do not divide a feature dimension, the implementation zero-pads the input blocks and crops the output blocks.

Evaluation metric. The core metric is *client-matched accuracy*. We first compute accuracy for each class on the common test set, then weight these classwise accuracies by each client's training-label distribution and average across clients by sample count. This evaluates every persistent client model against its corresponding label mixture.

Baseline methods. The table distinguishes fully shared baselines (FedAvg, FedProx, SCAFFOLD) from personalized methods (FedRep, Ditto, pFedMe, FedLoRA, and FedPHM). The main CNN baseline labeled FedLoRA uses a trainable shared dense base and private rank-16 LoRA factors, with scaling $32/16$ (Hu et al., 2022). Its upload includes the dense base. The budget-matched LoRA-private-base control uses a different sharing assignment, specified in the table caption.

5.2 Core Performance and Communication Efficiency

We first report the compact configuration, which shows the communication reduction at the native PHM parameter scale. The capacity-matched results are provided in Appendix 5.

Compact configuration (native parameter scale). The PHM structure naturally cuts communication cost. With the spectral projection threshold set to $\tau = 0.4$, FedPHM with $n = 8$ uploads 2.616 MiB per selected client per round, while the $n = 10$ configuration uploads 2.627 MiB. Under strong heterogeneity ($\alpha = 0.1$), the two compact orders reach 54.08% and 53.90% client-matched accuracy, compared with 34.06% for FedAvg.

Table 1: CNN results on CIFAR-100 after 200 rounds (five seeds). Type distinguishes fully shared and personalized methods. Parameter counts refer to trainable parameters per client; upload size is measured in MiB per selected client per round. FedPHM entries use spectral projection threshold $\tau = 0.4$. LoRA-private-base ($r = 57$) is a budget-matched control with a trainable client-private base W and shared rank-57 factors A, B.

Type	Method	$\alpha = 0.1$	$\alpha = 0.5$	Parameters	Upload (MiB)
Fully shared	FedAvg	34.06 ± 0.44	39.53 ± 0.53	2,520,676	9.62
Fully shared	FedProx	33.94 ± 0.47	39.51 ± 0.55	2,520,676	9.62
Fully shared	SCAFFOLD	33.08 ± 0.53	39.79 ± 0.84	2,520,676	19.23
Personalized	FedRep	52.82 ± 0.60	37.10 ± 0.50	2,520,676	9.42
Personalized	Ditto	55.78 ± 0.66	38.46 ± 0.39	5,041,352	9.62
Personalized	pFedMe	49.38 ± 0.68	36.54 ± 0.33	5,041,352	9.62
Personalized	FedLoRA	58.96 ± 0.60	47.98 ± 0.59	2,594,404	9.62
Personalized	LoRA-private-base ($r = 57$)	55.77 ± 0.65	37.56 ± 0.41	2,783,332	2.62
Personalized	FedPHM ($n = 8, \tau = 0.4$)	54.08 ± 0.87	41.42 ± 0.42	686,180	2.62
Personalized	FedPHM ($n = 10, \tau = 0.4$)	53.90 ± 0.49	41.21 ± 0.63	689,632	2.63

5.3 Ablation Studies and Theoretical Alignment

The CIFAR-100 heterogeneity sweep in Table 2 compares FedPHM and PHM-Full under the same PHM order and parameter count. The benefit of private interactions is largest for concentrated client distributions and decreases as α increases. This trend directly connects the selective aggregation design to the heterogeneity setting. The sweep uses the ablation threshold $\tau = 0.8$; the main comparison uses $\tau = 0.4$.

Table 2: CIFAR-100 client-matched accuracy (%) across heterogeneity levels after 200 rounds (five seeds; spectral projection threshold $\tau = 0.8$).

α	FedAvg	PHM-Full ($n = 4$)	PHM-Full ($n = 8$)	FedPHM ($n = 4$)	FedPHM ($n = 8$)
5.00	40.64 ± 0.38	40.32 ± 0.36	37.70 ± 0.52	37.93 ± 0.39	34.53 ± 0.44
1.00	40.52 ± 0.39	40.00 ± 0.23	37.01 ± 0.25	39.06 ± 0.45	38.27 ± 0.39
0.50	39.56 ± 0.70	39.22 ± 0.42	36.27 ± 0.47	40.70 ± 0.49	41.61 ± 0.39
0.10	33.98 ± 0.50	34.41 ± 0.41	31.95 ± 0.47	50.82 ± 0.97	54.03 ± 0.88
0.05	30.31 ± 0.40	31.40 ± 0.31	28.85 ± 0.41	58.92 ± 0.48	60.96 ± 0.49

The spectral-projection sweep in Table 3 uses $\alpha = 0.1$ and five seeds. The $n = 4$ rows use learning rate 0.05, while the $n = 8$ rows use learning rate 0.03. Client-matched accuracy is the primary metric and aggregated accuracy is an offline diagnostic. For $n = 4$, accuracy improves up to $\tau = 0.8$ and then saturates. For $n = 8$, it is stable for $\tau \geq 0.3$. The final spectral norm reaches the threshold when the constraint is active.

Table 3: CIFAR-100 spectral-projection sweep for FedPHM after 200 rounds ($\alpha = 0.1$; five seeds). The $n = 4$ rows use learning rate 0.05 and the $n = 8$ rows use learning rate 0.03.

n	τ	Client-matched (%)	Macro (%)	P10 (%)	Aggregated diagnostic (%)	Final max $\|A\|_2$	Upload (MiB)
4	0.1	32.51 ± 1.56	31.63	21.19	16.67	0.100	3.616
4	0.2	47.35 ± 0.98	46.73	38.11	28.38	0.200	3.616
4	0.3	51.05 ± 1.07	50.48	42.50	30.74	0.300	3.616
4	0.4	53.13 ± 0.90	52.60	44.80	31.75	0.400	3.616
4	0.6	54.66 ± 0.86	54.15	46.40	32.61	0.600	3.616
4	0.8	54.98 ± 0.88	54.46	46.93	32.84	0.800	3.616
4	1.0	54.94 ± 0.88	54.42	46.91	32.83	0.945	3.616
8	0.1	43.14 ± 0.81	42.92	34.97	19.08	0.100	2.616
8	0.2	52.92 ± 0.75	52.87	46.13	25.77	0.200	2.616
8	0.3	53.92 ± 0.86	53.89	47.32	26.16	0.300	2.616
8	0.4	54.08 ± 0.87	54.03	47.60	26.07	0.400	2.616
8	0.6	54.07 ± 0.90	54.05	47.52	25.98	0.600	2.616
8	0.8	54.05 ± 0.90	54.04	47.53	25.99	0.731	2.616
8	1.0	54.05 ± 0.90	54.04	47.53	25.99	0.731	2.616

Detailed PHM-order and fixed-local-step ablations are provided in Tables 7 and 8.The complete participation sweep is reported in Table 6.

6 Conclusion

We propose FedPHM, which combines compact shared PHM bases with persistent private interactions for communication-efficient personalization. Private interactions control blockwise mixing and can preserve full-rank client variation within the shared-basis subspace. The exact quadratic-risk decomposition identifies when this retained variation outweighs the approximation cost of the PHM structure. Across CIFAR-100, CIFAR-10, and FMNIST, FedPHM provides a favorable personalized accuracy–communication trade-off, with its largest gains under strong client heterogeneity. The present evaluation focuses on stateful clients and fully connected classifier transformations; extending selective transmission to other layer types and naturally occurring client populations is a direct direction for future work.

AI use statement

Generative AI tools assisted with language polishing, figure drafting, and code-oriented consistency checks during manuscript preparation. All mathematical derivations, experimental protocols, numerical results, and final claims were inspected and verified by the authors, who take responsibility for the complete submission.

REFERENCES

Liam Collins, Hamed Hassani, Aryan Mokhtari, and Sanjay Shakkottai. Exploiting shared representations for personalized federated learning. In *ICML*, 2021.

Dongshang Deng, Xuangou Wu, Tao Zhang, Xiangyun Tang, Hongyang Du, Jiawen Kang, Jiqiang Liu, and Dusit Niyato. Fedasa: A personalized federated learning with adaptive model aggregation for heterogeneous mobile edge computing. *IEEE Transactions on Mobile Computing*, 23 (12):14787–14802, 2024. doi: 10.1109/TMC.2024.3446271.

Sajjad Ghiasvand, Yifan Yang, Zhiyu Xue, Mahnoosh Alizadeh, Zheng Zhang, and Ramtin Pedarsani. Communication-efficient and tensorized federated fine-tuning of large language models. *arXiv preprint arXiv:2410.13097*, 2024.

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

Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, et al. Advances and open problems in federated learning. *Foundations and Trends in Machine Learning*, 14(1–2):1–210, 2021.

Sai Praneeth Karimireddy et al. Scaffold: Stochastic controlled averaging for federated learning. In *ICML*, 2020.

Ahmed Khaled, Konstantin Mishchenko, and Peter Richtárik. Tighter theory for local SGD on identical and heterogeneous data. In *AISTATS*, 2020.

Shiwei Li, Xiandi Luo, Haozhao Wang, Xing Tang, Shijie Xu, Weihong Luo, Yuhua Li, Xiuqiang He, and Ruixuan Li. The panaceas for improving low-rank decomposition in communication-efficient federated learning. *arXiv preprint arXiv:2505.23176*, 2025.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. In *MLSys*, 2020.

Tian Li, Shengyuan Hu, Ahmad Beirami, and Virginia Smith. Ditto: Fair and robust federated learning through personalization. In *ICML*, 2021.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *AISTATS*, 2017.

Zhikang Shen, Jianrong Lu, Haiyuan Wan, and Jianhai Chen. Sdflora: Selective decoupled federated lora for privacy-preserving fine-tuning with heterogeneous clients. *arXiv preprint arXiv:2601.11219*, 2026.

Sebastian U. Stich. Local SGD converges fast and communicates little. In *ICLR*, 2019.

Panlong Wu, Kangshuo Li, Ting Wang, Yanjie Dong, Victor C. M. Leung, and Fangxin Wang. Fedfmsl: Federated learning of foundation models with sparsely activated lora. *IEEE Transactions on Mobile Computing*, 23(12):15167–15181, 2024. doi: 10.1109/TMC.2024.3454634.

Aston Zhang, Yi Tay, Shuai Zhang, Alvin Chan, Anh Tuan Luu, Siu Cheung Hui, and Jie Fu. Beyond fully-connected layers with quaternions: Parameterization of hypercomplex multiplications with 1/n parameters. In *ICLR*, 2021.

Zikai Zhang, Ping Liu, Jiahao Xu, and Rui Hu. Fed-hello: Efficient federated foundation model fine-tuning with heterogeneous lora allocation. *IEEE Transactions on Neural Networks and Learning Systems*, 36(10):17556–17569, 2025. doi: 10.1109/TNNLS.2025.3580495.

A Proofs and Extended Results

A.1 Compact Convergence Analysis

We analyze the PHM optimization subsystem in (1). Let $S = (S_1, \dots, S_n)$ be the shared bases and $A_k = (A_{k,1}, \dots, A_{k,n})$ the private state of client k. Define

$$\Phi(S, A) = \frac{1}{K} \sum_{k=1}^{K} F_k(S; A_k). \tag{28}$$

The objective is lower bounded by $\Phi_\star$. The shared and private blocks obey the joint smoothness model

$$\begin{aligned} F_k(S'; A'_k) \le F_k(S; A_k) + \langle \nabla_S F_k, S' - S \rangle + \langle \nabla_A F_k, A'_k - A_k \rangle \\ + \frac{L_S}{2} \|S' - S\|_F^2 + \frac{L_A}{2} \|A'_k - A_k\|_F^2. \end{aligned} \tag{29}$$

At every local state, stochastic shared gradients are unbiased with variance at most σ_S^2/B, and private stochastic gradients have zero-mean noise with second moment at most ν_A. The complete local direction has second moment at most G_{dir}^2. Along every local path, the joint gradient is L_{path}-Lipschitz in the product Frobenius norm. All private updates use Euclidean projection onto $\mathcal{C}_\tau^n$.

For a feasible private state, define the projected-gradient mapping

$$G_{\eta,k}(S, A_k) = \frac{1}{\eta} \left[A_k - \Pi_{\mathcal{C}_\tau^n} \left(A_k - \eta \nabla_A F_k(S; A_k) \right) \right], \tag{30}$$

and the joint residual

$$\mathcal{R}_\eta(S, A) = \|\nabla_S \Phi(S, A)\|_F^2 + \frac{1}{K} \sum_{k=1}^{K} \|G_{\eta,k}(S, A_k)\|_F^2. \tag{31}$$

Projection keeps every private state in $\mathcal{C}_\tau^n$.

A.2 PHM Sampling Variance

Let

$$\Delta_{A,t}^2 = \frac{1}{K} \sum_k \|A_k^t - \bar{A}^t\|_F^2, \qquad \bar{A}^t = \frac{1}{K} \sum_k A_k^t, \tag{32}$$

and let $\zeta_{W,t}^2$ denote the variance of the effective-weight gradients evaluated at the common private state $\bar{A}^t$. If the shared gradient is L_{SA}-Lipschitz in the private state, the PHM map $\sum_i A_i \otimes S_i$ gives

$$\left[\frac{1}{K} \sum_{k=1}^{K} \|\nabla_S F_k(S^t; A_k^t) - \nabla_S \Phi(S^t, A^t)\|_F^2 \right]^{1/2} \le L_{SA} \Delta_{A,t} + n\tau \zeta_{W,t}. \tag{33}$$

Indeed, the private-state term follows from the Lipschitz condition. The effective-weight term is mapped through the adjoint PHM operator, whose norm is at most $\left(\sum_i \|\bar{A}_i^t\|_F^2 \right)^{1/2} \le n\tau$.

For M clients sampled uniformly without replacement, define

$$V_t = \frac{\sigma_S^2}{MB} + \frac{K - M}{M(K-1)} \left(L_{SA} \Delta_{A,t} + n\tau \zeta_{W,t} \right)^2. \tag{34}$$

The finite-population variance identity and (33) give

$$\mathbb{E}_t \|\widehat{g}_S^t - \nabla_S \Phi(S^t, A^t)\|_F^2 \le V_t.$$

A.3 FIXED-E LOCAL-SGD STATIONARITY

Let $q = M/K$. For the virtual local path of client k, write $(S^t_{k,e}, A^t_{k,e})$ for the state after e local steps, and define

$$c_E = \frac{1}{E}\sum_{e=0}^{E-1} e^2 = \frac{(E-1)(2E-1)}{6}. \tag{35}$$

The local path displacement obeys

$$\frac{1}{E}\sum_{e=0}^{E-1} \mathbb{E}_t\|(S^t_{k,e}, A^t_{k,e}) - (S^t, A^t_k)\|^2_{\mathrm{PF}} \le c_E \eta^2 G^2_{\mathrm{dir}}. \tag{36}$$

This follows from projection nonexpansiveness and Jensen's inequality.

Define the fixed-E residual

$$\widetilde{\mathcal{R}}^{(E)}_t = \|\nabla_S \Phi(S^t, A^t)\|^2_F + \frac{1}{KE}\sum_{k=1}^{K}\sum_{e=0}^{E-1} \|\widetilde{G}^t_{\eta,k,e}\|^2_F, \tag{37}$$

where $\widetilde{G}^t_{\eta,k,e}$ is the stochastic projected-gradient mapping on the virtual path. Let $\mathcal{R}^{(E)}_t$ be the same expression with the exact mapping $G_{\eta,k}(S^t_{k,e}, A^t_{k,e})$.

Theorem 1 (Fixed-E Local-SGD stationarity)**.** *Under the conditions above, uniform partial participation, and*

$$0 < \eta \le \min\{1/(L_S E), 1/(2L_A E)\},$$

the iterates satisfy

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}[\mathcal{R}^{(E)}_t] \le \frac{8(\Phi^0 - \Phi_\star)}{qE\eta T} + \frac{8L_S E\eta}{q}\,\overline{V} + \frac{12L^2_{\mathrm{path}} c_E \eta^2 G^2_{\mathrm{dir}}}{q} + 10\nu_A, \tag{38}$$

where

$$\overline{V} = \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}[V_t].$$

Proof. Apply the joint smoothness model to the final shared state and the final private states of each selected client. The path bound (36) controls the local-gradient drift. The shared update has sampling and minibatch error bounded by (34). Projection optimality gives the private descent term, and

$$\mathbb{E}_t\|G^t_{\eta,k,e} - \widetilde{G}^t_{\eta,k,e}\|^2_F \le \nu_A$$

converts the stochastic mapping to the exact mapping. Uniform client inclusion contributes the factor q. Under the stated step-size conditions, the resulting one-round inequality is

$$\begin{aligned}\mathbb{E}_t[\Phi^{t+1}] \le \Phi^t - \frac{qE\eta}{8}\mathbb{E}_t[\widetilde{\mathcal{R}}^{(E)}_t] + \frac{L_S E^2\eta^2}{2}\left(\sqrt{V_t} + L_{\mathrm{path}}\eta\sqrt{c_E}G_{\mathrm{dir}}\right)^2 \\ + E\eta L^2_{\mathrm{path}} c_E \eta^2 G^2_{\mathrm{dir}} + qE\eta\nu_A.\end{aligned}$$

Sum over t, use $\Phi^T \ge \Phi_\star$, apply Jensen's inequality to $\sqrt{V_t}$, and use $(a+b)^2 \le 2a^2 + 2b^2$. This gives the stated bound. □

For $E = 1$, $c_E = 0$ and the fixed-E residual reduces to the one-step joint residual. Thus, the one-step update is a special case of Theorem 1. For fixed E and $\eta = \Theta(T^{-1/2})$, the first term and the variance term decay as $O(T^{-1/2})$; the private-gradient noise contributes the residual $O(\nu_A)$.

B ADDITIONAL EXPERIMENTAL DETAILS

Table 4: CIFAR-100 CNN protocol. "Compact" refers to the main setting; capacity-matched controls and the legacy $n \in \{2, 4, 8\}$ order ablation use the stated learning-rate exceptions.

Hyperparameter	Value	Description
Communication rounds	200	Total training rounds
Local SGD steps	20	Client iterations per round
Batch size	64	Local minibatch size
Total clients	50	Clients in the federated system
Selected clients per round	10	Uniform sampling without replacement
Base learning rate (compact)	0.03	Initial learning rate for the main CNN protocol
Base learning rate (capacity/order)	0.05	Capacity controls and fixed-width order ablation
Weight decay	5×10^{-4}	ℓ_2 regularization coefficient
Learning-rate schedule	Cosine decay	Schedule during training
Optimizer	SGD without momentum	Local client optimizer
Spectral threshold τ (main comparison)	0.4	Both compact and capacity-matched FedPHM
Spectral threshold τ (other CNN experiments)	0.8	Unless a projection ablation specifies otherwise
Data partition	Dirichlet	$\alpha \in \{0.1, 0.5\}$ unless an ablation states otherwise
Aggregation	Sample-weighted	Selected clients; sampling is uniform without replacement
Evaluation	Client-matched accuracy	Primary metric; offline aggregated diagnostic is secondary
Random seeds	$\{42, 123, 456, 2026, 2027\}$	Five independent runs

Table 5: Complete CNN results on CIFAR-100 after 200 rounds (five seeds). Parameter counts refer to trainable parameters per client; upload size is measured in MiB per selected client per round. FedPHM entries use spectral projection threshold $\tau = 0.4$. LoRA-private-base ($r = 57$) is a budget-matched control with a trainable client-private base W and shared rank-57 factors A, B.

Setting	Type	Method	$\alpha = 0.1$	$\alpha = 0.5$	Parameters	Upload (MiB)
Compact	Fully shared	FedAvg	34.06 ± 0.44	39.53 ± 0.53	2,520,676	9.62
	Fully shared	FedProx	33.94 ± 0.47	39.51 ± 0.55	2,520,676	9.62
	Fully shared	SCAFFOLD	33.08 ± 0.53	39.79 ± 0.84	2,520,676	19.23
	Personalized	FedRep	52.82 ± 0.60	37.10 ± 0.50	2,520,676	9.42
	Personalized	Ditto	55.78 ± 0.66	38.46 ± 0.39	5,041,352	9.62
	Personalized	pFedMe	49.38 ± 0.68	36.54 ± 0.33	5,041,352	9.62
	Personalized	FedLoRA	58.96 ± 0.60	47.98 ± 0.59	2,594,404	9.62
	Personalized	LoRA-private-base ($r = 57$)	55.77 ± 0.65	37.56 ± 0.41	2,783,332	2.62
	Personalized	FedPHM ($n = 8, \tau = 0.4$)	54.08 ± 0.87	41.42 ± 0.42	686,180	2.62
	Personalized	FedPHM ($n = 10, \tau = 0.4$)	53.90 ± 0.49	41.21 ± 0.63	689,632	2.63
Capacity-matched	Personalized	FedPHM ($n = 8, \tau = 0.4$)	58.68 ± 0.65	46.90 ± 0.50	2,520,276	9.61
	Personalized	FedPHM ($n = 10, \tau = 0.4$)	58.93 ± 0.79	47.79 ± 0.49	2,519,012	9.61

B.1 Full CIFAR-100 Main Comparison

Table 5 gives the complete CIFAR-100 comparison, including the capacity-matched FedPHM configurations omitted from the compact main-text table.

B.2 Evaluation definitions

For client-matched evaluation, we compute classwise test accuracy for each persistent client model, weight it by that client's training-label distribution, and then average by client sample count. For the offline aggregated-model diagnostic, a single model combines the shared state with the sample-weighted mean of all private states and is evaluated on the complete benchmark test set. Fully shared methods have identical scores under both constructions.

B.3 Partial participation, projection, and local steps

Partial participation. Table 6 reports the complete participation sweep referenced in the main text. At 10% participation, PHM-Full reaches 9.99%, 62.48%, and 23.88% for seeds 42, 123, and 456, respectively, yielding a sample standard deviation of 27.20 percentage points.

Table 6: Legacy CIFAR-10 partial-participation ablation. Entries are client-matched accuracy (%), reported as mean and sample standard deviation over seeds $\{42, 123, 456\}$ after 150 rounds with 10 clients and $\alpha = 0.1$.

Method	10%	30%	50%	100%
FedAvg	25.41 ± 2.84	66.18 ± 6.47	73.25 ± 3.95	74.42 ± 2.95
PHM-Full ($n = 4$)	32.12 ± 2.72	66.30 ± 3.22	71.40 ± 3.60	72.77 ± 2.47
FedPHM ($n = 4$)	89.41 ± 2.15	90.93 ± 2.39	91.26 ± 1.79	91.45 ± 1.62

Table 7: CIFAR-100 fixed-local-step ablation at $\alpha = 0.1$ after 200 rounds (five seeds). Entries are client-matched accuracy (%); upload is MiB per selected client per round. FedPHM uses $n = 8$ and hidden width 512; both methods use the compact learning rate 0.03.

E	Method	Client-matched accuracy (%)	Upload (MiB)
1	FedAvg	9.59 ± 0.32	9.616
	FedPHM ($n = 8$)	17.83 ± 0.83	2.616
5	FedAvg	20.37 ± 0.23	9.616
	FedPHM ($n = 8$)	35.40 ± 0.73	2.616
10	FedAvg	26.93 ± 0.60	9.616
	FedPHM ($n = 8$)	44.73 ± 0.76	2.616
15	FedAvg	31.09 ± 0.52	9.616
	FedPHM ($n = 8$)	50.45 ± 0.77	2.616
20	FedAvg	33.97 ± 0.38	9.616
	FedPHM ($n = 8$)	54.04 ± 0.90	2.616

Table 8: Fixed-width CIFAR-100 PHM-order ablation after 200 rounds ($\alpha = 0.1$; learning rate 0.05; spectral projection threshold $\tau = 0.8$). "Parameters" counts one persistent trainable model copy. Upload is the exact FP32 state-dictionary payload per selected client per round.

n	Method	Client-matched (%)	Macro (%)	P10 (%)	Aggregated diagnostic (%)	Parameters	Upload (MiB)
2	PHM-Full	36.48 ± 0.44	36.72	30.08	36.48	1,472,108	5.62
2	FedPHM	42.71 ± 0.43	42.33	35.05	35.29	1,472,108	5.62
4	PHM-Full	37.23 ± 0.43	37.46	31.34	37.23	947,876	3.62
4	FedPHM	55.35 ± 1.01	54.90	47.30	32.92	947,876	3.62
8	PHM-Full	35.56 ± 0.36	35.87	29.24	35.56	686,180	2.62
8	FedPHM	57.06 ± 0.88	57.08	50.85	28.87	686,180	2.62
10	FedPHM	53.65 ± 0.46	53.71	47.33	24.72	688,723	2.623
12	FedPHM	53.48 ± 0.43	53.61	47.49	24.37	687,184	2.615
14	FedPHM	52.71 ± 0.44	52.84	46.69	23.12	688,968	2.618
16	FedPHM	53.02 ± 0.40	53.14	46.92	23.31	690,068	2.617

Local steps. We use the CIFAR-100 protocol to align the ablation with the fixed-E analysis. Accuracy improves as each round contains more local computation; these runs have a fixed number of rounds, so they do not represent a fixed-compute comparison. The theorem makes the associated local-trajectory term explicit through $c_E = \Theta(E^2)$.

B.4 CIFAR-100 PHM ORDER AND PROJECTION ABLATIONS

The following results use CIFAR-100, $\alpha = 0.1$, 50 clients, 10 clients per round, 200 rounds, 20 local steps, batch size 64, and learning rate 0.05. The client-matched columns report the weighted mean, the macro mean, and the 10th percentile across clients. Upload is the exact FP32 payload per selected client per round; the seed protocols are specified in the table caption.

PHM order n. At a fixed width of 512, increasing n from 2 to 8 reduces upload from 5.62 MiB to 2.62 MiB and improves client-matched accuracy from 42.71% to 57.06% in this CIFAR-100 configuration. This fixed-width ablation is intentionally separate from the capacity- and communication-matched controls in the main paper.

B.5 FMNIST AND CIFAR-10 EXTENSIONS

We report additional FMNIST and CIFAR-10 experiments. Both datasets use a convolution-free MLP with PHM applied to the first high-dimensional layer. The protocol has 50 clients, 10 clients per round, 200 rounds, 20 local SGD steps, batch size 64, zero momentum, weight decay 5×10^{-4}, cosine learning-rate decay, and $\tau = 0.8$. We use Dirichlet partitions with $\alpha \in \{0.1, 0.5\}$ and five

Table 9: FMNIST MLP results after 200 rounds (five seeds). The first two blocks are compact, hidden-width-512 configurations with $n = 4$ and $n = 8$; the final block gives capacity-matched controls. Each entry is client-matched accuracy followed by the offline aggregated-model diagnostic in parentheses.

Method	$\alpha = 0.1$	$\alpha = 0.5$	Parameters	Upload (MiB)
Compact, $n = 4$, $h = 512$				
FedAvg	83.54 ± 0.19 (83.54)	85.10 ± 0.15 (85.10)	407,050	1.553
FedProx	83.53 ± 0.19 (83.53)	85.09 ± 0.15 (85.09)	407,050	1.553
SCAFFOLD	84.80 ± 0.21 (84.80)	85.60 ± 0.10 (85.60)	407,050	3.106
FedRep	95.46 ± 0.74 (73.03)	88.63 ± 0.24 (81.73)	407,050	1.533
Ditto	95.98 ± 0.71 (83.54)	89.37 ± 0.22 (85.10)	407,050	1.553
pFedMe	94.86 ± 0.89 (79.08)	88.64 ± 0.19 (82.57)	407,050	1.553
FedLoRA	95.97 ± 0.70 (80.80)	90.23 ± 0.20 (84.62)	427,786	1.553
PHM-Full ($n = 4$)	82.90 ± 0.43 (82.90)	84.91 ± 0.15 (84.91)	106,058	0.405
FedPHM ($n = 4$)	94.71 ± 0.82 (77.82)	88.37 ± 0.20 (83.39)	106,058	0.404
Compact, $n = 8$, $h = 512$				
PHM-Full ($n = 8$)	81.63 ± 0.46 (81.63)	84.07 ± 0.25 (84.07)	56,330	0.215
FedPHM ($n = 8$)	94.96 ± 0.82 (74.60)	88.23 ± 0.19 (81.71)	56,330	0.213
Capacity-matched controls				
PHM-Full ($n = 4$, $h = 1968$)	85.10 ± 0.31 (85.10)	86.50 ± 0.20 (86.50)	407,450	1.554
FedPHM ($n = 4$, $h = 1968$)	95.64 ± 0.70 (80.86)	90.02 ± 0.11 (85.36)	407,450	1.554
PHM-Full ($n = 8$, $h = 3728$)	84.91 ± 0.25 (84.91)	86.41 ± 0.05 (86.41)	406,874	1.552
FedPHM ($n = 8$, $h = 3728$)	95.80 ± 0.69 (79.74)	90.03 ± 0.16 (84.76)	406,874	1.550

Table 10: CIFAR-10 compact-MLP results after 200 rounds (five seeds; $n = 4$ and hidden width 512).

Method	$\alpha = 0.1$	$\alpha = 0.5$	Parameters	Upload (MiB)
FedAvg	43.91 ± 0.47 (43.91)	46.00 ± 0.26 (46.00)	1,578,506	6.022
FedProx	43.77 ± 0.42 (43.77)	46.01 ± 0.24 (46.01)	1,578,506	6.022
SCAFFOLD	45.37 ± 0.51 (45.37)	46.89 ± 0.30 (46.89)	1,578,506	12.043
FedRep	76.56 ± 0.88 (38.41)	58.47 ± 0.97 (42.90)	1,578,506	6.002
Ditto	76.75 ± 0.96 (43.70)	58.49 ± 0.86 (46.13)	1,578,506	6.022
pFedMe	69.30 ± 0.81 (40.37)	52.86 ± 0.71 (42.95)	1,578,506	6.022
FedLoRA	57.40 ± 0.98 (44.61)	49.20 ± 0.54 (46.83)	1,635,850	6.022
PHM-Full ($n = 4$)	41.92 ± 0.54 (41.92)	44.36 ± 0.49 (44.36)	398,922	1.522
FedPHM ($n = 4$)	62.56 ± 1.14 (37.46)	48.47 ± 0.56 (42.48)	398,922	1.522

Table 11: CIFAR-10 capacity-matched selective-aggregation ablation after 200 rounds (five seeds). FedPHM keeps A_k private, whereas PHM-Full aggregates all PHM parameters. Entries are client-matched accuracy (%). For $n = 4$, both methods have 1,579,886 trainable parameters and upload 6,319,288 (FedPHM) or 6,319,544 (PHM-Full) bytes per selected client per round; for $n = 8$, the corresponding values are 1,577,362 parameters and 6,307,400 or 6,309,448 bytes.

α	FedPHM ($n = 4$)	PHM-Full ($n = 4$)	FedPHM ($n = 8$)	PHM-Full ($n = 8$)
0.05	77.25 ± 2.11	42.19 ± 0.79	83.89 ± 1.90	42.75 ± 0.71
0.10	66.60 ± 1.31	44.52 ± 0.52	73.47 ± 1.01	45.28 ± 0.59
0.50	52.45 ± 0.42	47.79 ± 0.49	57.00 ± 0.67	49.19 ± 0.47
1.00	50.20 ± 0.52	48.63 ± 0.19	53.16 ± 0.58	50.12 ± 0.18
5.00	48.41 ± 0.19	49.33 ± 0.35	49.14 ± 0.08	50.88 ± 0.15

seeds $\{42, 123, 456, 2026, 2027\}$. The CIFAR-10 selective-aggregation ablation below additionally uses $\alpha \in \{0.05, 0.1, 0.5, 1, 5\}$. Client-matched accuracy is the weighted mean over clients; parenthesized values, where reported, are the aggregated-model diagnostic. Upload is the exact FP32 payload per selected client per round. In all capacity-matched MLP entries, h denotes the hidden width of the first layer.

CIFAR-10 selective-aggregation ablation. Table 11 isolates the effect of keeping the interaction matrices private. Within each order, FedPHM and PHM-Full have matched trainable parameter counts and nearly identical uplink payloads; their only material difference is whether A_k is aggregated. FedPHM is markedly better under stronger heterogeneity, while PHM-Full is stronger at $\alpha = 5$.