FedPKDA: Personalized Federated Learning with Privacy-Preserving Knowledge Dynamic Alignment

Moxuan Zeng[1], Wenxuan Tu[1, 2], Yuanyi Chen[1], Yiying Wang[1], Miao Yu[2], Xiangyan Tang[1, 2], Jieren Cheng[1, 2]

[1]School of Computer Science and Technology, Hainan University, Haikou, China
[2]Hainan Blockchain Technology Engineering Research Center, Haikou, China
{zengmx, twx, tangxy36}@hainanu.edu.cn

Abstract

Personalized Federated Learning (PFL), which aims to customize models for each client while preserving data privacy, has become an important research topic in addressing the challenges of data heterogeneity. Existing studies usually enhance the localization of global parameters by injecting local information into the globally shared model. However, these methods focus excessively on the personalized characteristics of individual clients and fail to fully exploit distinctive information across clients, limiting the quality of local models to represent unseen samples well. To address this issue, we propose a novel personalized **Fed**erated **P**rivacy-preserving **K**nowledge **D**ynamic **A**lignment (**FedPKDA**) framework, which ensures data privacy during both the collection of client-side key information and its incorporation into federated model training. Specifically, to ensure data privacy during the cross-client information collection phase, we first conduct feature clipping and add Laplacian noise to the local prototypes extracted from each client. Further, we compute the centroid of the uploaded local prototypes in a latent space and leverage Mahalanobis distance to guide the generation of global prototypes, thereby preserving the semantic contributions from participating clients. Moreover, to boost the personalization of the local model, we dynamically align representations learned by the shared model with both a set of local prototypes and privacy-preserving global prototypes, facilitating effective cross-client knowledge sharing under heterogeneous settings while preserving client-specific characteristics. Extensive experiments on benchmark datasets have verified the superiority of FedPKDA against its competitors.

Code — https://github.com/zmx666/FedPKDA

Introduction

Federated Learning (FL) is a distributed framework that enables the collaborative training of deep neural networks while preserving data and model privacy (Liu et al. 2025; Sabah et al. 2025; Ma, Yao, and Xu 2024; Cai et al. 2024b; Wang et al. 2024a). Meanwhile, as an iterative optimization-based machine learning paradigm, FL primarily aims to collaboratively train a high-performing global model (Jiang et al. 2025; Zhang et al. 2023b,c; Wang et al. 2024b). However, the distributions of client data are highly heterogeneous, which can lead to degraded generalization performance of the global model, making it challenging to effectively meet the personalized needs of individual clients (Zhi et al. 2024; Wang et al. 2024c).

To address the challenge of data heterogeneity in FL, Personalized Federated Learning (PFL) has been proposed to customize models for individual clients based on the statistical characteristics of their private datasets, thereby enhancing local adaptability. More recently, parameter decoupling has emerged as a prevalent approach in PFL. It effectively addresses the issue of data heterogeneity by splitting the model into a shared backbone and a personalized layer. To further improve feature representation and model performance, researchers have introduced pretrained backbones and model decomposition techniques for dynamic structural optimization. While such advances have led to notable improvements in personalization performance (Chen, Vikalo, and Wang 2024; Wang et al. 2025; Zheng et al. 2025; Wu et al. 2023), these approaches still suffer from two critical limitations: **1) weak data privacy**. The features or gradients transmitted by clients may implicitly contain sensitive representations of the original data, which increases the risk of raw data being reconstructed (i.e., privacy leakage), and **2) local representation bias**. Overemphasis on individual client information may result in locally optimized models that deviate from the global objective, thereby decreasing the representation quality of local models on unseen samples. As a consequence, it is imperative to develop a cross-client dynamically aligned PFL framework to effectively integrate knowledge across clients in a privacy-preserving manner, while dynamically adjusting the alignment intensity to ensure effective global information sharing without compromising the individuality of local models.

An intuitive solution to addressing the above challenges is to share client-specific information in a privacy-preserving manner and derive a global consensus representation to guide local model optimization. However, reaching this goal requires addressing two key challenges: 1) how to effectively integrate cross-client consensus knowledge while ensuring the security of local data, and 2) it remains a significant challenge to learn shared knowledge across clients while preserving client-specific information. To address the

Corresponding authors: Xiangyan Tang and Wenxuan Tu

first challenge, inspired by the class-representative learning approach used for data privacy (Tan et al. 2022; Shi and Ye 2024), we propose to learn prototype-wise representations instead of the commonly used instance-level features or gradients for key signal collection, such that the resultant features are difficult to reconstruct into raw data while preserving shared information across clients. As for the second challenge, we posit that a tunable learning approach offers a promising solution to trade off shared and personalized information (Yang, Huang, and Ye 2024; Cui et al. 2024). Following this principle, the relative weighting between local-specific and cross-client consensus features can be dynamically adjusted. In this way, the model is encouraged to alleviate excessive local bias and enhance the representation quality of the model on unseen samples.

Based on these observations, we propose a novel personalized **Fed**erated **P**rivacy-preserving **K**nowledge **D**ynamic **A**lignment (**FedPKDA**) framework, which aims to overcome the limitations of learning solely from local data by effectively leveraging cross-client information while ensuring data privacy. In this method, the knowledge of global prototypes is dynamically injected into the shared model to enhance the representation quality of the local model on unseen samples. Specifically, during the cross-client information collection phase, local prototype representations are collected from online clients, which are subsequently pruned and perturbed with Laplacian noise to ensure data privacy during transmission to the server. On the server, to capture cross-client consensus, we cluster the collected local prototypes and identify their geometric centers in the latent space. Global prototypes are generated by weighting the local prototypes according to their Mahalanobis distances from the centroid. In this way, data privacy can be preserved while reducing the performance degradation caused by noise. In the local training stage, we design a knowledge dynamic alignment scheme that flexibly incorporates diverse client perspectives. The shared model further integrates local prototypes to enhance personalization while introducing a knowledge allocation mechanism to guide alignment between global prototypes and representations learned by the shared model. This component enables the model to effectively learn global consensus knowledge while preserving client-specific personalized information. The main contributions of this work are as follows:

- A novel PFL framework termed FedPKDA is proposed, which facilitates a good trade-off between client-specific personalized information and cross-client consensus knowledge in a privacy-preserving manner.
- We propose two key components: a cross-client prototype privacy-preserving scheme that enhances data privacy protection during the collaborative information-sharing process, and a knowledge dynamic alignment scheme that enables effective knowledge sharing while preserving client-specific information.
- Extensive experiments on the Flowers102 and Cifar100 benchmark datasets have demonstrated the effectiveness and superiority of the proposed method against advanced PFL competitors in more challenging scenarios.

Related Work

Privacy Preservation in FL

Federated learning (FL) is a distributed machine learning paradigm that enables model training on decentralized clients while preserving data privacy (Lee et al. 2024; Zhang et al. 2025; Cai et al. 2024a). In recent years, extensive research has been devoted to enabling efficient collaboration under privacy constraints. FedProto (Tan et al. 2022) leverages prototypes to safeguard data privacy, preventing adversaries from reconstructing raw data without access to local models. DP-FedSAM (Shi et al. 2023) integrates the Sharpness-Aware Minimization (SAM) optimizer to generate more stable and smoother local models, effectively alleviating the performance degradation caused by differential privacy. Nevertheless, its privacy-preserving strategy remains limited and lacks dynamic adaptability. To address this issue, FedDPA (Yang, Huang, and Ye 2023) applies an adaptive differential privacy mechanism within a personalized federated learning framework and utilizes Fisher information to retain critical local parameter information, thus effectively balancing model performance and privacy-preserving. DP-FPL (Tran et al. 2025) combines low-rank factorization with local-global differential privacy mechanisms to achieve differentiated privacy protection for personalized local prompts and shared global prompts. Previous studies have applied prototype-based learning or differential privacy to protect data and model privacy in FL. However, the majority of these privacy-preserving methods focus primarily on the client side, neglecting the risk of privacy leakage from cross-client consensus information. In contrast, we propose a cross-client prototype privacy-preserving scheme that integrates prototype learning with principles inspired by differential privacy to enable effective knowledge sharing while reducing the risk of privacy leakage throughout the federated learning process.

Parameter Decoupling in PFL

Compared to traditional federated learning methods, personalized approaches are better suited to accommodate the unique data distributions of individual clients, thereby alleviating performance degradation caused by data heterogeneity (Rieyan et al. 2024; Li et al. 2025). Current prevalent personalized strategies are mostly based on parameter decoupling, where the model is divided into a shared feature extraction backbone and a client-specific personalized head. By aggregating and sharing the backbone parameters globally while retaining independent head modules for each client, this strategy significantly enhances the adaptability and performance of local models under heterogeneous data conditions. For instance, LG-FedAvg (Liang et al. 2020) improves personalization by dividing the model architecture into client-specific and globally shared layers, though it requires manual specification of the partitioning strategy, which limits its generalization ability. To overcome the limitations of manual structural partitioning, FedRoD (Chen and Chao 2022) innovatively adopts a shared backbone extractor combined with a dual-head architecture, significantly expanding the applicability of personalized fed-

erated learning. Furthermore, FedDecomp (Wu et al. 2024) introduces a novel low-rank parameter decomposition strategy that explicitly divides the parameters of each layer into a shared full-rank component and a local personalized low-rank component. However, inconsistencies in parameter structures between the shared and personalized parts may affect model collaboration. To address this challenge, FedAS (Yang, Huang, and Ye 2024) proposes initializing the shared parameters before local training to align them with the previous round's local feature outputs, thereby enhancing consistency between shared and personalized modules. Although the aforementioned methods have improved parameter-partition-based personalized federated learning to some extent, they largely overlook the significance of cross-client information, thereby constraining the representation quality of local models to unseen data. In our study, we propose a knowledge dynamic alignment scheme that not only captures intrinsic local features but also guides the globally shared model to learn critical cross-client information, thus enhancing the representation quality of the local model in heterogeneous scenarios.

Method

Problem Definition

We consider a federated learning setup with one central server and N clients, where in each communication round t, the server randomly selects a subset of clients denoted as $\mathcal{S}_t$ to participate in training. In personalized federated learning, each client i holds a private dataset D_i consisting of input–label pairs (x, y), with x denoting the input features and y the corresponding labels. Due to the heterogeneity of data across clients, it is challenging to train a unified global model with strong generalization capability. Each client model is parameterized as $C_i = (\theta_i, \eta_i)$, where θ_i denotes the shared component and η_i represents the personalized component of client i. We define the personalized optimization objective as follows:

$$\min_{C} F(C) = \frac{1}{n} \sum_{i=1}^{n} \left[\mathcal{L}_i(C_i) + \mathcal{R}_i(C_i, \lambda) \right], \tag{1}$$

where $C = \{C_1, C_2, \ldots, C_n\}$ denotes the set of all personalized client models. The term $\mathcal{L}_i(C_i)$ represents the empirical loss of client i on its local dataset, ensuring that the model can better fit the local data distribution and improve personalization. The term $\mathcal{R}_i(C_i, \lambda)$ is a regularization component that incorporates global information to enhance generalization ability, where λ controls the strength of this regularization.

Cross-client Prototype Privacy-preserving

In personalized federated learning, prototypes serve as cross-client shared feature representations, which not only facilitate the modeling of generalized information but also offer certain privacy-preserving advantages (Tan et al. 2022). However, existing studies often fail to fully exploit the consensus information across clients and lack thorough investigation into the security of prototype data. To address this, we propose a cross-client prototype privacy-preserving scheme that enhances the privacy of prototypes while enabling effective knowledge sharing. The proposed method is described as follows.

To begin with, each client first passes its local data through a shared feature extractor $f(\cdot)$, and the extracted feature vectors are clipped to a fixed range to reduce the risk of sensitive information leakage. After clipping, each client computes local prototypes p_i^k for each class using processed features, reflecting the local data characteristics without exposing raw samples:

$$p_i^k = \frac{1}{M_i^k} \sum_{x \in D_i^k} f_{clip}(x), \tag{2}$$

where M_i^k denotes the number of samples in class k on client i, and D_i^k represents the corresponding local dataset. $f_{clip}(x)$ denotes the clipped feature representation. To further enhance the privacy protection of local prototype uploading, we inject Laplace noise into the computed local prototypes. The perturbed prototypes are computed as $\tilde{p}_i^k = p_i^k + \xi, \xi \sim \mathrm{Lap}(0, b)$, where b is the scale parameter controlling the noise magnitude. The Laplace distribution is defined by the probability density function $\mathrm{Lap}(a \mid b) = \frac{1}{2b} \exp\left(-\frac{|a|}{b}\right)$, where a is the sampled noise value. This mechanism introduces randomness to obscure individual data characteristics, thereby improving the overall security and robustness of the system (Zhang et al. 2024; Tan et al. 2024). We define the class-wise private prototype set for class k as:

$$\tilde{\mathcal{P}}_k = \left\{ \tilde{p}_i^k \mid i \in \mathcal{I}_k \right\}, \tag{3}$$

where $\mathcal{I}_k$ denotes the set of clients that possess local samples of class k.

Subsequently, the server constructs a stable and robust global feature representation by aggregating perturbed local prototypes from multiple clients. This process does not require access to any raw sample data, and the resulting global prototypes effectively fuse shared semantic information across clients, providing generalized knowledge representations for all clients. Specifically, for each class k, the server gathers the corresponding perturbed prototypes from different clients into the set $\tilde{\mathcal{P}}_k$ and performs clustering on the server side using a clustering algorithm such as *K*-Means (Tu et al. 2024, 2021):

$$\boldsymbol{\mu}_k = \arg\min_{\boldsymbol{\mu}} \sum_{i \in \mathcal{I}_k} \left\| \tilde{p}_i^k - \boldsymbol{\mu} \right\|_2^2, \tag{4}$$

where $\boldsymbol{\mu}_k$ denotes the centroid of prototypes of class k, computed via *K*-Means by minimizing the Euclidean distance. The variable $\boldsymbol{\mu}$ represents a candidate center during optimization. This approach identifies the centroid that best represents the overall distribution and assigns weights w_i^k to each local prototype based on its Mahalanobis distance to the centroid:

$$d_i^k = \sqrt{(\tilde{p}_i^k - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1} (\tilde{p}_i^k - \boldsymbol{\mu}_k)}, \tag{5}$$

$$w_i^k = \frac{1/(d_i^k + \epsilon)}{\sum_{j \in \mathcal{I}_k} 1/(d_j^k + \epsilon)}, \tag{6}$$

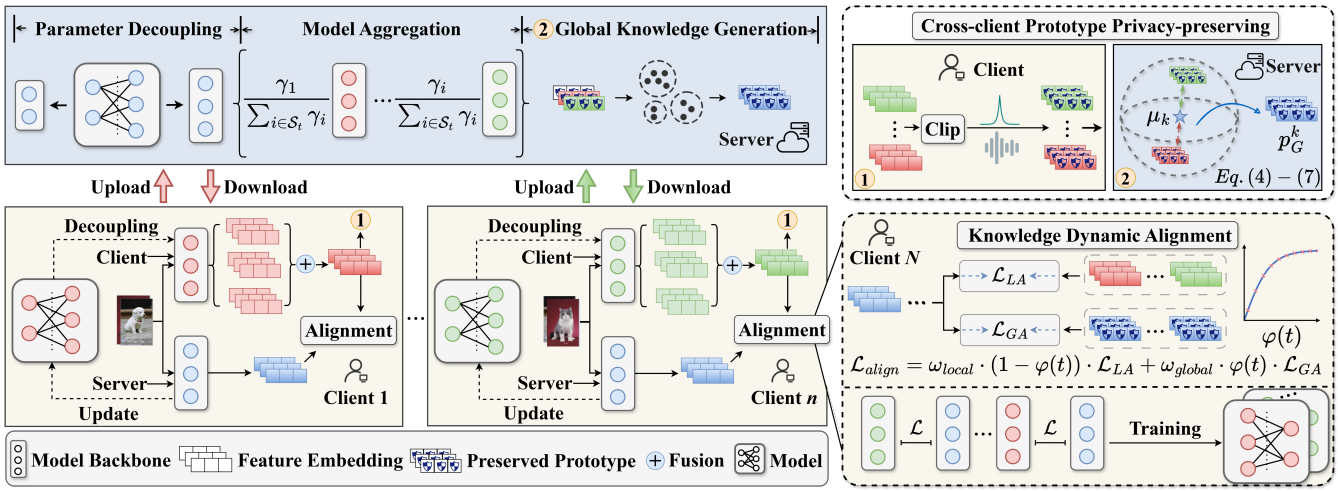

Figure 1: The overall architecture of FedPKDA integrates cross-client prototype privacy-preserving mechanisms and knowledge dynamic alignment into a unified PFL framework. Specifically, local prototypes are collected in a privacy-preserving manner and clustered to obtain centroids. The Mahalanobis distances are then used to calculate the distance between each local prototype and the centroid, which can be used to generate the weighted global prototypes. Subsequently, the representations learned by the shared model are dynamically aligned with both the local prototype set and the privacy-preserving global prototype set, thereby enhancing the representation quality of local model for unseen samples.

$$p_G^k = \sum_{i \in \mathcal{I}_k} w_i^k \cdot \tilde{p}_i^k, \tag{7}$$

where d_i^k denotes the Mahalanobis distance between the perturbed local prototype $\tilde{p}_i^k$ and the class-wise centroid $\boldsymbol{\mu}_k$, computed with the inverse of the covariance matrix $\boldsymbol{\Sigma}_k = \mathrm{Cov}(\tilde{\mathcal{P}}_k) + \epsilon \mathbf{I}$, where $\mathbf{I}$ is the identity matrix and $\epsilon = 10^{-6}$ is a small constant for stability. Let $\mathcal{P}_G = \{p_G^1, p_G^2, \ldots, p_G^k\}$ denote the set of global prototypes, where p_G^k is the aggregated global prototype for class k. In doing so, it effectively preserves the semantic structure among prototypes from different clients and mitigates the negative impact of outlier prototypes on the global representation.

Knowledge Dynamic Alignment

As previously discussed, while it is possible to effectively acquire consensus information across clients under privacy-preserving settings, a key challenge remains: how to effectively leverage client-specific information across clients to enhance the representation quality of local models on unseen samples. To address this gap, we propose a knowledge dynamic alignment scheme, which offers an intuitive two-step solution by first incorporating and aligning both local and global knowledge through dual knowledge alignment. And then, we introduce a knowledge dynamic allocation mechanism that dynamically adjusts the relative importance of local and global prototypes during training to enable the model to learn client-specific information while effectively transferring useful knowledge from other clients.

In each training round, the client receives the global feature prototypes p_G^k from the server and integrates them with its own local prototypes p_i^k to jointly participate in the alignment process. The model computes the local alignment loss $\mathcal{L}_{LA}$ and the global alignment loss $\mathcal{L}_{GA}$ to measure the distance between the current sample representation and both the local and global prototypes to achieve dual knowledge alignment. By simultaneously minimizing both losses, the client preserves local semantic information while effectively incorporating global knowledge, thereby enhancing its representation quality on unseen distributions. The losses are defined as follows:

$$\mathcal{L}_{LA} = \sum_k \left\| f(x) - p_i^k \right\|_2^2, \tag{8}$$

$$\mathcal{L}_{GA} = \sum_k \left\| f(x) - p_G^k \right\|_2^2. \tag{9}$$

Next, to accommodate the varying demands for local and global information at different stages of training, we introduce a training-progress-based knowledge dynamic allocation mechanism. This mechanism employs a customized activation function $\varphi(t)$ to dynamically adjust the relative weights of the local and global alignment loss terms. In the early stages of training, the model relies more heavily on local prototypes to mitigate representational discrepancies between local and global models. As training progresses, the alignment weight gradually shifts toward the global prototypes, guiding the model to better integrate shared knowledge. The formal definition of $\varphi(t)$ is given by:

$$\varphi(t) = \frac{\log(1+e^t) - \log(1+e^{-t})}{\log(1+e^t) + \log(1+e^{-t})}. \tag{10}$$

At the early stage of training ($t \to 0^+$), both $\log(1+e^t)$ and $\log(1+e^{-t})$ converge to $\log 2$, yielding $\varphi(t) \to 0$. As training progresses ($t \to 1$), $\varphi(t)$ increases monotonically. Let $f(t) = \log(1+e^t)$, and define $h = f(t)$, $q = f(-t)$,

Algorithm 1: FedPKDA

Input: Communication rounds T, local epochs E, number of clients N, private local dataset D_i, client online ratios P
Output: Personalized model parameter θ_i^t

Initialize shared model f, personalized heads η_i for each client i
for each round $t = 1$ to T **do**
  Server selects client subset $\mathcal{S}_t$
  for each client $i \in \mathcal{S}_t$ **in parallel do**
    Compute local prototypes p_i^k and add noise
    Send noisy prototypes $\tilde{p}_i^k$ to server
  end for
  Compute $\boldsymbol{\mu}_k$ using K-Means by Eq. (4)
  Generate Mahalanobis-based global prototypes p_G^k by Eq. (5)–(7)
  Receive $\mathcal{P}_G$ from server
  for $e = 1$ to E **do**
    for each mini-batch (x, y) **do**
      Calculate $\mathcal{L}_{align}$ by Eq. (11)
      Calculate $\mathcal{L}_{CE}$ by Eq. (12)
      Calculate $\mathcal{L}$ by Eq. (13)
    end for
  end for
  Update global model: $\theta^t \leftarrow \sum_{i \in \mathcal{S}_t} \hat{\gamma}_i \cdot \theta_i^t$
  Send θ^t to each online client
end for

along with their derivatives $h' = \sigma(t)$, $q' = -\sigma(-t)$, where $\sigma(\cdot)$ denotes the sigmoid function. The derivative of $\varphi(t)$ is given by $\varphi'(t) = \frac{(h'-q')(h+q)-(h-q)(h'+q')}{(h+q)^2}$, i.e., the derivative simplifies to: $\varphi'(t) = \frac{2(h'q-q'h)}{(h+q)^2}$. When $t > 0$ implies $h' > 0$, $q > 0$, $q' < 0$ and $h > 0$, the numerator remains strictly positive, and hence $\varphi'(t) > 0$. This confirms that $\varphi(t)$ is strictly increasing over the interval $(0, 1]$, which aligns with our design objective of gradually shifting the alignment emphasis from local to global prototypes as training progresses. Meanwhile, we set an initial value of 0.1 to prevent the assignment weight from approaching zero.

Subsequently, we unify the local and global prototype alignment losses into a single formulation, resulting in the overall knowledge dynamic alignment loss function $\mathcal{L}_{align}$, which is incorporated into the overall loss $\mathcal{L}$ for client-side optimization, defined as:

$$\mathcal{L}_{align} = \omega_{local} \cdot (1-\varphi(t)) \cdot \mathcal{L}_{LA} + \omega_{global} \cdot \varphi(t) \cdot \mathcal{L}_{GA}, \quad (11)$$

$$\mathcal{L}_{CE} = -\mathbb{E}_{(x,y)\sim D_i} \left[\log p(y \mid x) \right], \quad (12)$$

$$\mathcal{L} = \mathcal{L}_{CE} + \mathcal{L}_{align}, \quad (13)$$

where ω_{local} and ω_{global} denote the initial weighting coefficients for the local and global alignment losses. $\mathcal{L}_{CE}$ denotes the cross-entropy classification loss on the local dataset, $p(y \mid x)$ represents the predicted probability of sample x belonging to class y.

Model Aggregation

In the server aggregation phase, each client estimates the empirical Fisher Information Matrix (FIM) using its local dataset D_i and the current model parameters θ_i^t. The client first computes the gradient of the log-likelihood with respect to the model parameters, denoted as $\mathbf{G}_i = \nabla_{\theta_i^t} \log p(D_i \mid \theta_i^t)$, and approximates the FIM via the outer product of the gradient. Subsequently, the client extracts the diagonal elements of the FIM and calculates an importance score γ_i as the sum of these diagonal entries. To normalize the importance scores across clients, each client computes its final aggregation weight as $\hat{\gamma}_i = \frac{\gamma_i}{\sum_{i \in \mathcal{S}_t} \gamma_i}$. Even clients with low participation frequency can still retain sufficient influence during the aggregation process (Yang, Huang, and Ye 2024). Finally, the parameters θ_i^t from each client are aggregated using $\hat{\gamma}_i$ as the weighting coefficient for the global model update. Algorithm 1 provides a detailed description of the FedPKDA learning process.

Experiments

Experiment Setup

Datasets. To evaluate the effectiveness of the proposed FedPKDA method, we conduct experiments on two benchmark datasets: Flowers102 (Nilsback and Zisserman 2008) and Cifar100 (Krizhevsky, Hinton et al. 2009).

Model. We adopt the widely used 6-layer CNN backbone for image feature extraction.

Baseline Methods. We compare FedPKDA with twelve state-of-the-art personalized federated learning (PFL) methods, including FedAvg-P (McMahan et al. 2017), FedProx-P (Li et al. 2020), Ditto (Li et al. 2021), FedProto (Tan et al. 2022), FedGH (Yi et al. 2023), FedDBE (Zhang et al. 2023a), FedPer (Arivazhagan et al. 2019), LG-FedAvg (Liang et al. 2020), FedCAC (Wu et al. 2023), FedAS (Yang, Huang, and Ye 2024), FedFSA (Xing et al. 2025) and FedBABU (OH, Kim, and Yun 2022). Among them, we finetuned FedAvg-P and FedProx-P to optimize their performance.

Implementation Details. To ensure fair comparison, FedPKDA and all baseline methods are evaluated under the same experimental settings. The dataset is partitioned into non-IID subsets using a Dirichlet distribution (Lin et al. 2020) with heterogeneity parameters $\beta \in \{0.3, 0.5\}$. Model training batch size is 16. SGD is employed as the local optimizer, with each client performing 5 local training epochs per round, and the total communication rounds $T = 40$. The local learning rate is set to 5×10^{-3}. Noise scale b is empirically set to 0.1 (Dong et al. 2022). The data is distributed across 20 clients, and we consider three different client online ratios $\{0.2, 0.4, 0.6\}$. All experiments are conducted on an NVIDIA GeForce RTX 3090 GPU.

Metrics. To evaluate local performance and privacy protection, we use Accuracy (ACC) for classification, PSNR for DLG reconstruction quality, and ACC, F1-score, AUC, and TPR for assessing robustness against MIA attack.

Methods	Cifar100						Flowers102					
	$\beta = 0.3$			$\beta = 0.5$			$\beta = 0.3$			$\beta = 0.5$		
	$P = 0.2$	$P = 0.4$	$P = 0.6$	$P = 0.2$	$P = 0.4$	$P = 0.6$	$P = 0.2$	$P = 0.4$	$P = 0.6$	$P = 0.2$	$P = 0.4$	$P = 0.6$
FedAvg-P	27.87	32.67	36.35	27.44	30.66	32.07	31.61	36.47	39.34	32.23	35.39	37.28
FedProx-P	27.27	32.25	35.29	26.99	30.22	31.36	32.05	35.84	38.76	31.98	35.00	37.40
FedPer	34.78	34.23	33.66	29.77	28.71	27.79	38.52	38.18	37.64	34.03	33.30	32.47
Ditto	24.90	26.89	29.40	25.71	27.39	27.58	28.21	29.86	31.07	30.18	31.98	32.57
LG-FedAvg	29.95	31.67	33.26	24.97	26.32	26.96	31.22	33.75	33.85	24.93	26.39	25.81
FedBABU	38.18	38.89	38.81	33.60	33.88	33.18	40.86	40.95	40.56	36.85	36.90	36.80
FedProto	20.66	33.70	36.99	17.62	28.46	31.58	26.55	34.63	36.08	23.91	29.07	29.31
FedGH	17.80	33.97	34.10	17.66	28.35	27.93	33.17	35.89	34.97	25.71	27.85	28.68
FedCAC	31.03	32.22	32.75	25.07	26.36	26.63	31.71	33.26	33.70	23.96	25.71	25.52
FedDBE	25.35	27.09	28.10	26.03	27.62	27.98	27.86	29.86	30.93	30.23	31.79	32.81
FedAS	40.68	43.51	44.47	35.48	38.21	38.70	35.21	38.37	39.54	27.90	27.51	28.78
FedFSA	31.80	36.96	40.54	31.74	35.38	37.23	30.73	35.94	39.49	31.35	35.00	37.72
FedPKDA	**44.91**	**46.20**	**46.18**	**41.25**	**41.73**	**41.81**	**44.11**	**44.54**	**44.21**	**38.09**	**38.79**	**39.32**

Table 1: Comparison of different methods on Cifar100 and Flowers102 classification tasks. Experiments are conducted under two heterogeneous data scenarios ($\beta \in \{0.3, 0.5\}$) and three client online ratios ($P \in \{0.2, 0.4, 0.6\}$). The best and runner-up results are shown in bold and underline, respectively.

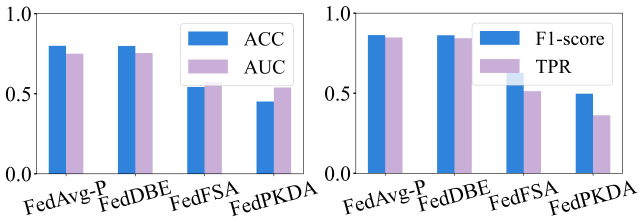

Figure 2: Bar chart comparison of Label-Only MIA evaluation metrics on the Cifar100 ($\beta = 0.5$) across different methods. Lower values indicate stronger privacy preservation.

P	Metric	FedAvg-P	FedProx-P	FedDBE	**FedPKDA**
0.2	PSNR	7.49	7.65	7.42	**6.96**
	PSNR_Avg	7.41	7.33	7.25	**6.98**
0.4	PSNR	7.33	7.50	7.44	**6.87**
	PSNR_Avg	7.28	7.27	7.20	**6.97**
0.6	PSNR	7.22	7.20	7.30	**6.92**
	PSNR_Avg	7.36	7.30	7.31	**7.04**

Table 2: PSNR values of different methods under DLG attack on Cifar100 ($\beta = 0.3$) . Lower PSNR values indicate better privacy-preserving. Best results are in bold.

Performance Comparison

In the comparative experiments, we systematically evaluated FedPKDA against several existing personalized federated learning methods under different data heterogeneity settings. As shown in Table 1, the results comprehensively demonstrate the performance advantages of FedPKDA across various scenarios. The highest accuracy for each task is highlighted in bold black font. The results show that FedPKDA consistently achieves the best performance in most tasks, confirming its effectiveness in addressing data heterogeneity. Taking the Cifar100 dataset as an example, under the setting of $\beta = 0.3$ and a client participation rate of $P = 0.2$, FedPKDA achieves an accuracy of 44.91%, surpassing the second-best method FedAS by 4.23%. This result supports the effectiveness of the knowledge dynamic alignment scheme introduced in FedPKDA. By capturing cross-client specific information, the scheme effectively mitigates the performance degradation caused by overfitting to client-local peculiarities. Moreover, FedPKDA surpasses FedBABU and FedProx-P by 6.73% and 17.64%, further demonstrating that enhancing semantic diversity through cross-client representation alignment improves the representation quality of the local model for unseen samples.

Privacy Analysis

This section evaluates the privacy-preserving effectiveness of FedPKDA against Deep Leakage from Gradients (DLG) (Geiping et al. 2020) attacks, using Peak Signal-to-Noise Ratio (PSNR) to quantify the similarity between reconstructed and original images. Lower PSNR values indicate weaker reconstruction and thus stronger privacy protection. As shown in Table 2, experiments on the Cifar100 dataset with $\beta = 0.3$ and client online rates of $\{0.2, 0.4, 0.6\}$ demonstrate that FedPKDA consistently yields significantly lower PSNR values than existing methods. We report both the PSNR at round 40 and the average over the last five rounds. Figure 2 further presents results of the Label-Only Membership Inference Attack (Choquette-Choo et al. 2021) on Cifar100 with $\beta = 0.5$ and $P = 0.2$. FedPKDA achieves the lowest scores across all four evaluation metrics, demonstrating its strong privacy-preserving capability in personalized federated learning.

Ablation Studies

To validate the effectiveness of the two proposed schemes, we conduct experiments on Cifar100 and Flowers102 using

Dataset	P	Base	w/ KDA	w/ CP	**FedPKDA**
Cifar100	0.2	41.55	43.84	43.20	**44.91**
	0.4	42.90	45.50	45.73	**46.20**
	0.6	43.05	45.70	45.44	**46.18**
Flowers102	0.2	39.49	40.71	40.60	**44.11**
	0.4	40.80	41.78	41.43	**44.54**
	0.6	40.36	42.89	41.54	**44.21**

Table 3: Ablation Study of FedPKDA on Cifar100 and Flowers102 ($\beta = 0.3$). Bold values indicate the best results.

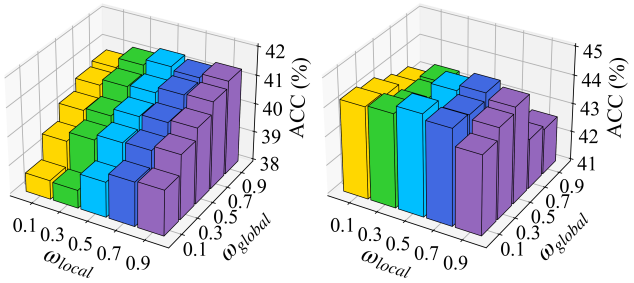

Figure 3: Sensitivity analysis of FedPKDA with respect to two key hyperparameters on the Cifar100 (left) and Flowers102 (right). The X-axis, Y-axis, and Z-axis refer to the ω_{local} value, ω_{global} value, and ACC, respectively.

FedPKDA and its variants. "Base" refers to the basic version without any additional scheme; "w/ CP" denotes the incorporation of cross-client prototype information; and "w/ KDA" introduces a knowledge dynamic allocation mechanism. As shown in Table 3, under a non-IID setting with data heterogeneity parameter $\beta = 0.3$ and client online rates $P = \{0.2, 0.4, 0.6\}$, FedPKDA consistently achieves significant performance improvements on both datasets. Taking the case of $P = 0.2$ as an example, we highlight the following key observations: 1) Compared to the "Base" model, both "w/ KDA" and "w/ CP" show clear performance gains. On Cifar100, they improve accuracy by 2.29% and 1.65%; on Flowers102, by 1.22% and 1.11%. This demonstrates the effectiveness of each module in enhancing model performance. 2) FedPKDA consistently outperforms the individual variants. It exceeds "w/ KDA" and "w/ CP" by 1.07% and 1.71% on Cifar100, and by 3.40% and 3.51% on Flowers102. These results highlight the complementary and synergistic benefits of combining both schemes.

Hyper-Parameter Analysis

We conduct a systematic analysis of the model sensitivity to the local alignment weight ω_{local} and the global alignment weight ω_{global}. Specifically, we vary both ω_{local} and ω_{global} from 0.1 to 0.9 with a step size of 0.2 to evaluate the performance of the FedPKDA model. As shown in Figure 3, we evaluate two datasets under different heterogeneity settings with the client online rate $P = 0.2$. On the Cifar100 dataset with heterogeneity parameter $\beta = 0.5$,

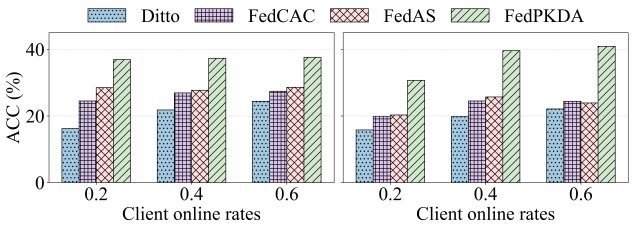

Figure 4: Scalability analysis on Flowers102 ($\beta = 0.3$). The local model test accuracy is evaluated under varying client numbers of 35 (left) and 50 (right) in $P \in \{0.2, 0.4, 0.6\}$. Higher accuracy indicates stronger scalability.

the model performance consistently improves as ω_{local} increases, while the change in ω_{global} exhibits a trend of rising first and then slightly declining. The optimal performance is achieved when $\omega_{\text{local}} = 0.5$ and $\omega_{\text{global}} = 0.9$. In addition, on the Flowers102 dataset with $\beta = 0.3$, the best result is observed when $\omega_{\text{local}} = 0.7$ and $\omega_{\text{global}} = 0.5$. The results indicate that under different data distributions, properly setting the local and global alignment weights helps improve the model's representational ability and enhances the classification performance of local models. Therefore, we effectively adjust the alignment strength based on the characteristics of each dataset.

Scalability Analysis

FedPKDA demonstrates strong scalability under varying numbers of clients and participation rates. As shown in Figure 4, we increase the number of clients to 35 and 50 and evaluate the test accuracy under different client online ratios $P \in \{0.2, 0.4, 0.6\}$. The results consistently show that FedPKDA outperforms other SOTA methods across all settings. As the number of clients increases, the amount of data per client decreases, and the heterogeneity of data distribution intensifies. Nevertheless, FedPKDA maintains stable and superior performance even under these more challenging conditions, demonstrating its adaptability in large-scale federated learning scenarios.

Conclusion

This paper focuses on a practically valuable yet inherently challenging problem: weakly private personalized federated learning under local representation bias. To address this issue, we propose a novel personalized federated learning framework, termed FedPKDA. The framework enables the privacy-preserving collection and sharing of critical cross-client information. Specifically, we design cross-client prototype privacy-preserving and knowledge dynamic alignment schemes, which can strengthen privacy protection and enhance the representation quality of local models on unseen data. Extensive experiments show that FedPKDA effectively addresses challenges and consistently outperforms existing methods under non-IID settings. In future work, we plan to extend FedPKDA to multi-view federated learning settings to further explore its adaptability in such scenarios.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (NSFC) (Grant No. 62562026, 62506102), the Natural Science Foundation of Hainan University (Grant No. XJ2400009401), the Key Research and Development Program of Hainan Province (Grant No. ZDYF2024GXJS014, ZDYF2023GXJS163), and Collaborative Innovation Project of Hainan University (Grant No. XTCX2022XXB02).

References

Arivazhagan, M. G.; Aggarwal, V.; Singh, A. K.; and Choudhary, S. 2019. Federated Learning with Personalization Layers. *arXiv preprint arXiv:1912.00818*.

Cai, J.; Zhan, Y.; Lu, Z.; Guo, W.; and Ng, S.-K. 2024a. Towards Effective Federated Graph Anomaly Detection via Self-Boosted Knowledge Distillation. In *Proceedings of the ACM International Conference on Multimedia*, 5537–5546.

Cai, J.; Zhang, Y.; Fan, J.; and Ng, S.-K. 2024b. LG-FGAD: An Effective Federated Graph Anomaly Detection Framework. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 3760–3769.

Chen, H.; and Chao, W. 2022. On Bridging Generic and Personalized Federated Learning for Image Classification. In *Proceedings of the International Conference on Learning Representations*.

Chen, Y.; Vikalo, H.; and Wang, C. 2024. Fed-qssl: A framework for personalized federated learning under bitwidth and data heterogeneity. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 11443–11452.

Choquette-Choo, C. A.; Tramer, F.; Carlini, N.; and Papernot, N. 2021. Label-Only Membership Inference Attacks. In *Proceedings of the International Conference on Machine Learning*, 1964–1974.

Cui, T.; Li, H.; Wang, J.; and Shi, Y. 2024. Harmonizing Generalization and Personalization in Federated Prompt Learning. In *Proceedings of the International Conference on Machine Learning*, 9646–9661.

Dong, J.; Wang, L.; Fang, Z.; Sun, G.; Xu, S.; Wang, X.; and Zhu, Q. 2022. Federated Class-Incremental Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10164–10173.

Geiping, J.; Bauermeister, H.; Dröge, H.; and Moeller, M. 2020. Inverting Gradients-How easy is it to break privacy in federated learning? In *Proceedings of the Conference on Neural Information Processing Systems*, 16937–16947.

Jiang, Z.; Xu, J.; Zhang, S.; Shen, T.; Li, J.; Kuang, K.; Cai, H.; and Wu, F. 2025. FedCFA: Alleviating Simpson's Paradox in Model Aggregation with Counterfactual Federated Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 17662–17670.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning Multiple Layers of Features from Tiny Images.

Lee, K. J.; Jeong, B.; Kim, S.; Kim, D.; and Park, D. 2024. General Commerce Intelligence: Glocally Federated NLP-Based Engine for Privacy-Preserving and Sustainable Personalized Services of Multi-Merchants. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 22752–22760.

Li, T.; Hu, S.; Beirami, A.; and Smith, V. 2021. Ditto: Fair and Robust Federated Learning Through Personalization. In *Proceedings of the International Conference on Machine Learning*, 6357–6368.

Li, T.; Sahu, A. K.; Zaheer, M.; Sanjabi, M.; Talwalkar, A.; and Smith, V. 2020. Federated Optimization in Heterogeneous Networks. In *Proceedings of the Machine Learning and Systems*, 429–450.

Li, Z.; Long, G.; Zhou, T.; Jiang, J.; and Zhang, C. 2025. Personalized Federated Collaborative Filtering: A Variational Autoencoder Approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 18602–18610.

Liang, P. P.; Liu, T.; Ziyin, L.; Allen, N. B.; Auerbach, R. P.; Brent, D.; Salakhutdinov, R.; and Morency, L.-P. 2020. Think Locally, Act Globally: Federated Learning with Local and Global Representations. *arXiv preprint arXiv:2001.01523*.

Lin, T.; Kong, L.; Stich, S. U.; and Jaggi, M. 2020. Ensemble Distillation for Robust Model Fusion in Federated Learning. In *Proceedings of the Conference on Neural Information Processing Systems*, 2351–2363.

Liu, J.; Cheng, J.; Han, R.; Tu, W.; Wang, J.; and Peng, X. 2025. Federated Graph-Level Clustering Network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 18870–18878.

Ma, Y.; Yao, Y.; and Xu, X. 2024. PPIDSG: A Privacy-Preserving Image Distribution Sharing Scheme with Gan in Federated Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 14272–14280.

McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; and y Arcas, B. A. 2017. Communication-Efficient Learning of Deep Networks From Decentralized Data. In *Proceedings of the Artificial Intelligence and Statistics*, 1273–1282.

Nilsback, M.-E.; and Zisserman, A. 2008. Automated Flower Classification over a Large Number of Classes. In *Proceedings of the Indian Conference on Computer Vision, Graphics & Image Processing*, 722–729.

OH, J. H.; Kim, S.; and Yun, S. 2022. FedBABU: Toward Enhanced Representation for Federated Image Classification. In *Proceedings of the International Conference on Learning Representations*.

Rieyan, S. A.; News, M. R. K.; Rahman, A. M.; Khan, S. A.; Zaarif, S. T. J.; Alam, M. G. R.; Hassan, M. M.; Ianni, M.; and Fortino, G. 2024. An Advanced Data Fabric Architecture Leveraging Homomorphic Encryption and Federated Learning. *Information Fusion*, 102: 102004.

Sabah, F.; Chen, Y.; Yang, Z.; Raheem, A.; Azam, M.; Ahmad, N.; and Sarwar, R. 2025. FairDPFL-SCS: Fair Dynamic Personalized Federated Learning with Strategic Client Selection for Improved Accuracy and Fairness. *Information Fusion*, 115: 102756.

Shi, W.; and Ye, M. 2024. Prospective Representation Learning for Non-Exemplar Class-Incremental Learning. In *Proceedings of the Conference on Neural Information Processing Systems*, 995–1018.

Shi, Y.; Liu, Y.; Wei, K.; Shen, L.; Wang, X.; and Tao, D. 2023. Make Landscape Flatter in Differentially Private Federated Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24552–24562.

Tan, X.; Xiao, D.; Huang, H.; Wang, M.; and Li, M. 2024. Hierarchically Fair and Differentially Private Federated Learning in Industrial IoT Based on Compressed Sensing With Adaptive-Thresholding Sparsification. *IEEE Transactions on Industrial Informatics*.

Tan, Y.; Long, G.; Liu, L.; Zhou, T.; Lu, Q.; Jiang, J.; and Zhang, C. 2022. FedProto: Federated Prototype Learning Across Heterogeneous Clients. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 8432–8440.

Tran, L.; Sun, W.; Patterson, S.; and Milanova, A. 2025. Privacy-Preserving Personalized Federated Prompt Learning for Multimodal Large Language Models. In *Proceedings of the International Conference on Learning Representations*.

Tu, W.; Guan, R.; Zhou, S.; Ma, C.; Peng, X.; Cai, Z.; Liu, Z.; Cheng, J.; and Liu, X. 2024. Attribute-Missing Graph Clustering Network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 15392–15401.

Tu, W.; Zhou, S.; Liu, X.; Guo, X.; Cai, Z.; Zhu, E.; and Cheng, J. 2021. Deep Fusion Clustering Network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 9978–9987.

Wang, H.; Jia, Y.; Zhang, M.; Hu, Q.; Ren, H.; Sun, P.; Wen, Y.; and Zhang, T. 2024a. FedDSE: Distribution-Aware Sub-Model Extraction for Federated Learning over Resource-constrained Devices. In *Proceedings of the ACM Web Conference*, 2902–2913.

Wang, H.; Zheng, P.; Han, X.; Xu, W.; Li, R.; and Zhang, T. 2024b. FedNLR: Federated Learning with Neuron-Wise Learning Rates. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 3069–3080.

Wang, L.; Tu, W.; Cheng, J.; Wang, J.; Tang, X.; and Wang, C. 2025. Discovering Maximum Frequency Consensus: Lightweight Federated Learning for Medical Image Segmentation. In *Proceedings of the ACM International Conference on Multimedia*, 1900–1909.

Wang, Y.; Fu, H.; Kanagavelu, R.; Wei, Q.; Liu, Y.; and Goh, R. S. M. 2024c. An Aggregation-Free Federated Learning for Tackling Data Heterogeneity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26233–26242.

Wu, X.; Liu, X.; Niu, J.; Wang, H.; Tang, S.; Zhu, G.; and Su, H. 2024. Decoupling General and Personalized Knowledge in Federated Learning via Additive and Low-rank Decomposition. In *Proceedings of the ACM International Conference on Multimedia*, 7172–7181.

Wu, X.; Liu, X.; Niu, J.; Zhu, G.; and Tang, S. 2023. Bold but Cautious: Unlocking the Potential of Personalized Federated Learning Through Cautiously Aggressive Collaboration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 19375–19384.

Xing, X.; Zhan, Q.; Xie, X.; Yang, Y.; Wang, Q.; and Liu, G. 2025. Flexible Sharpness-Aware Personalized Federated Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 21707–21715.

Yang, X.; Huang, W.; and Ye, M. 2023. Dynamic Personalized Federated Learning with Adaptive Differential Privacy. In *Proceedings of the Conference on Neural Information Processing Systems*, 72181–72192.

Yang, X.; Huang, W.; and Ye, M. 2024. Fedas: Bridging Inconsistency in Personalized Federated Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11986–11995.

Yi, L.; Wang, G.; Liu, X.; Shi, Z.; and Yu, H. 2023. FedGH: Heterogeneous Federated Learning with Generalized Global Header. In *Proceedings of the ACM International Conference on Multimedia*, 8686–8696.

Zhang, C.; Zhang, X.; Yang, X.; Liu, B.; Zhang, Y.; and Zhou, R. 2025. Poisoning Attacks Resilient Privacy-Preserving Federated Learning Scheme Based on Lightweight Homomorphic Encryption. *Information Fusion*, 121: 103131.

Zhang, J.; Hua, Y.; Cao, J.; Wang, H.; Song, T.; Xue, Z.; Ma, R.; and Guan, H. 2023a. Eliminating Domain Bias for Federated Learning in Representation Space. In *Proceedings of the International Conference on Neural Information Processing Systems*, 14204–14227.

Zhang, J.; Hua, Y.; Wang, H.; Song, T.; Xue, Z.; Ma, R.; Cao, J.; and Guan, H. 2023b. GPFL: Simultaneously Learning Global and Personalized Feature Information for Personalized Federated Learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5041–5051.

Zhang, J.; Liu, Y.; Hua, Y.; and Cao, J. 2024. FedTGP: Trainable Global Prototypes with Adaptive-Margin-Enhanced Contrastive Learning for Data and Model Heterogeneity in Federated Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 16768–16776.

Zhang, P.; Yan, H.; Wu, W.; and Wang, S. 2023c. Improving Federated Person Re-Identification through Feature-Aware Proximity and Aggregation. In *Proceedings of the ACM International Conference on Multimedia*, 2498–2506.

Zheng, H.; Hu, Z.; Yang, L.; Zheng, M.; Xu, A.; and Wang, B. 2025. ConFREE: Conflict-free Client Update Aggregation for Personalized Federated Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 22875–22883.

Zhi, M.; Bi, Y.; Xu, W.; Wang, H.; and Xiang, T. 2024. Knowledge-Aware Parameter Coaching for Personalized Federated Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 17069–17077.