Intra-Class Unbiased Prototype Aggregation and Classifier Collaboration for Personalized Federated Learning

Hao Zheng[1,2], Shiyu Song[1], Zhigang Hu[1], Meiguang Zheng[1*], Liu Yang[1*], Aikun Xu[1], Rongchang Zhao[1], Ruizhi Pu[2], Ruiyi Fang[2], Boyu Wang[2]

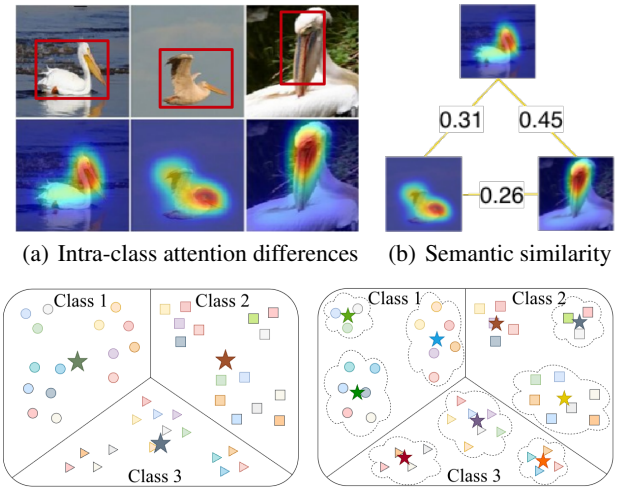
[1]School of Computer Science and Engineering, Central South University, Changsha, China
[2]Department of Computer Science, University of Western Ontario, London, Canada
{zhenghao, ssycsu, zghu, zhengmeiguang, yangliu, aikunxu, zhaorc}@csu.edu.cn; {rpu2, rfang32}@uwo.ca; bwang@csd.uwo.ca

Abstract

Prototype-based personalized federated learning methods have emerged as a promising strategy due to their ability to represent client-specific class characteristics effectively through learned class prototypes. These prototypes capture salient features of client-local data, facilitating personalized model adaptation. However, existing prototype-based aggregation strategies predominantly rely on weighted averaging, implicitly assuming prototype consistency across clients. This assumption neglects the intrinsic heterogeneity and non-independent and identically distributed (non-IID) nature of client data, compelling diverse local prototypes to align toward a singular global prototype and consequently causing significant aggregation bias. Motivated by observations from intra-class feature saliency analysis, we identify that clients inherently emphasize distinct feature regions even for the same class. To leverage this intra-class diversity, we introduce FedIC, a novel prototype clustering and collaborative classifier optimization approach. Specifically, FedIC first clusters prototypes based on intra-class similarity to form intra-class prototype subspaces, ensuring that aggregation occurs exclusively within each cluster, thus eliminating the bias stemming from forced global unification. To further exploit the benefits of intra-cluster collaboration, we quantify the combined predictive gains of classifiers from clients within the same cluster as a function of classifier combination weights. This targeted aggregation and collaborative optimization strategy effectively circumvents the bias introduced by global alignment. Extensive experiments under various non-IID settings show that FedIC significantly outperforms existing Prototype-based and Clustered PFL Methods.

Introduction

Federated Learning (FL) is a distributed learning paradigm that enables collaborative training of a global model across multiple clients without sharing private data (Zuo et al. 2025; Qi et al. 2025a; Pu et al. 2025). In many AI applications such as computer vision(Zheng et al. 2025a,b), natural language processing(Belyi et al. 2023), and recommendation systems(Zhang et al. 2023a), FL has emerged as a critical component. However, the influx of massive non-IID data generated by heterogeneous devices poses significant challenges (Huang et al. 2023; Xiong et al. 2023; Yi et al. 2023, 2024; Meng et al. 2024). Traditional FL approaches relying on a single global model often result in substantial deviations from local data distributions across clients, leading to slow convergence and suboptimal performance (Huang et al. 2024). Personalized Federated Learning (PFL) addresses this limitation by tailoring models to individual client characteristics while still leveraging knowledge from the federation, striking a balance between personalization and collaborative learning benefits (Ye et al. 2023; Zhang et al. 2023b; Song et al. 2025; Li et al. 2023; Zhu et al. 2023).



(a) Intra-class attention differences (b) Semantic similarity

(c) Global averaging vs. Intra-class prototype aggregation

Figure 1: Illustrations of intra-class feature divergence and the motivation for prototype clustering in PFL.

Prototype-based PFL methods have recently emerged as powerful techniques due to their intuitive representation of class-specific characteristics through learned prototypes, such as FedPAC (Xu, Tong, and Huang 2023). These methods typically rely on learning class prototypes and aggregate them using weighted averaging to form a global prototype (Tan et al. 2022; Zhou and Wang 2024; Zhang et al. 2023c).

*Corresponding authors: Meiguang Zheng, Liu Yang

Although this approach facilitates personalized model adaptation, it implicitly assumes prototype consistency across clients, overlooking the intrinsic heterogeneity in client data distributions (Chen et al. 2024; Luo et al. 2023; Deng et al. 2024). In non-IID settings, where clients emphasize different feature regions even for the same class, forcing diverse local prototypes to align toward a singular global prototype introduces significant aggregation bias (Cho, Wang, and Joshi 2022). This bias interferes with the direction of local parameter updates on the client side and ultimately diminishes the personalization capabilities of the model.

To better understand the root of prototype aggregation bias, we analyze the problem from the perspective of intra-class feature diversity, as illustrated in Fig. 1. The class activation maps (Fig. 1(a)) highlight that different clients focus on different feature regions (head, body, bill) of the same class (e.g., pelican). The intra-class feature similarity graph (Fig. 1(b)) quantifies these differences with similarity scores. Its root cause lies in the non-IID distribution of client-specific data (e.g., geographically influenced photographic preferences) that induce significant variations in region of interest attention for the same class (Bae, Noh, and Kim 2020). The left of Fig. 1(c) further highlights that traditional global aggregation ignores these important local differences, thus creating a biased global prototype. However, cluster structures (right) can retain this intra-class feature diversity instead of homogenizing them during aggregation.

To mitigate the bias issue in global prototype aggregation, we propose a novel PFL framework (FedIC) based on intra-class prototype clustering and collaborative classifier optimization. FedIC initially clusters client prototypes based on intra-class similarities, establishing intra-class subspaces that retain local feature distinctions, thereby preventing bias introduced by enforced global alignment. Furthermore, to maximize the benefits of intra-cluster collaboration, we quantify the predictive gains of classifiers from clients within the same cluster and optimize their combination weights. Extensive experiments under various non-IID settings demonstrate that FedIC significantly outperforms existing Prototype-based and Clustered PFL methods. Our contributions are summarized as follows:

- We identify the bias phenomenon in average prototype aggregation for PFL, revealing its root cause lies in ignoring intra-class feature diversity.
- We effectively mitigate the bias introduced by global prototype aggregation through an intra-class prototype clustering approach, and introduce a collaborative classifier optimization strategy to maximize the combinatorial predictive gain within each prototype cluster.
- Extensive evaluations under diverse non-IID scenarios demonstrate that FedIC consistently outperforms SOTA Prototype-based and Clustered PFL methods.

Related Work

Prototype-based Personalized Federated Learning

Prototype-based PFL has emerged as a promising approach to address data heterogeneity in FL by explicitly modeling class-wise feature representations through prototypes (Zhang et al. 2024; Oh, Kim, and Yun 2021). Instead of aggregating entire model parameters, these methods focus on learning representative prototypes for each class and client, typically implemented as class centroids in the feature space (Jia et al. 2024). For example, FedProto (Tan et al. 2022) proposes a unified representation learning framework using semantic anchors to guide prototype alignment across clients, aiming to improve consistency. Similarly, FedPAC (Xu, Tong, and Huang 2023) regularize the Euclidean distance between local features and global anchors to align clients' feature spaces. Addressing data imbalance in the cross-silo setting, FedFSA (Qi et al. 2025b) aligns client feature spaces to enhance model generalization across differently distributed silos. To handle feature shift, FedDP (Fu et al. 2025) aligns both the representation and parameter spaces, enabling more robust and domain-independent prototype learning. FedCCFA (Chen et al. 2024) introduces classifier clustering and feature anchor alignment mechanisms, adaptively adjusting feature space alignment weights based on label distribution entropy to alleviate inconsistencies between global and local prototypes.

Clustered Personalized Federated Learning

Clustered PFL aims to group clients with similar data distributions and train specialized models for each cluster to better handle data heterogeneity (Long et al. 2023). CFL (Sattler, Müller, and Samek 2020) leverages dynamic bi-partition clustering and identifies latent data distribution clusters by measuring the cosine similarity of client gradients or parameter updates. IFCA (Ghosh et al. 2020) and FeSEM (Long et al. 2023) rely on coarse-grained global distribution similarity assumptions. IFCA alternately optimizes client cluster identities and model parameters, combining multi-task learning weight-sharing techniques to enhance convergence efficiency. FeSEM proposes a federated stochastic expectation-maximization algorithm, jointly optimizing client clustering and global model training to reduce communication overhead. Fed-CAM (Ma et al. 2023) superimposes global models with cluster residual models, forcing cluster models to focus on differential features and optimizing residual terms to avoid cluster collapse. FedDrift (Jothimurugesan et al. 2023) creates new models based on drift detection and adaptively merges models by hierarchical clustering to address distributed concept drift. CFL-GP (Kim, Kim, and De Veciana 2024) periodically accumulates client gradients and partitions similar groups based on spectral clustering, incorporating Gaussian noise assumptions to enhance clustering robustness.

Despite these advances, existing prototype-based PFL methods mainly focus on aligning global and local prototypes, while overlooking the aggregation bias that arises during global prototype aggregation. Specifically, the intra-class bias caused by heterogeneous data distributions across clients can lead to biased prototype representations that compromise the effectiveness of personalized federated learning. While clustered PFL approaches effectively partition clients into homogeneous groups, they primarily operate at the client level and fail to capture fine-grained intra-class heterogeneity within individual clusters, potentially

limiting their personalization capabilities.

Method

Preliminaries

We consider a standard PFL setup involving a central server and N clients. Each client i holds a private dataset $\mathcal{D}_i$ drawn from a local data distribution over the input-label space $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{Y} = \{1, \ldots, C\}$ denotes the set of class labels. The data distributions are assumed to be heterogeneous and non-IID, i.e., $\mathcal{D}_i \neq \mathcal{D}_j$ for $i \neq j$. Each client learns a local model composed of a feature extractor $f(\cdot; \theta_i)$ and a classifier head $h_i(\cdot; \phi_i)$, where θ_i denotes the client's local feature extractor parameters and ϕ_i denotes its personalized classifier parameters. After local training, the server aggregates the local feature extractor parameters $\{\theta_i\}_{i=1}^N$ to form a globally shared encoder $\theta_g = \sum_{i=1}^N w_i \theta_i$. The overall goal of PFL is to jointly learn a shared encoder θ_g and personalized classifier heads $\{\phi_i\}_{i=1}^N$ that minimize the expected personalized loss across all clients:

$$\min_{\theta_g, \{\phi_i\}_{i=1}^N} \sum_{i=1}^{N} \mu_i \, \mathbb{E}_{(x,y)\sim\mathcal{D}_i} \left[\mathcal{L} \left(h_i(f(x; \theta_i); \phi_g), y \right) \right], \quad (1)$$

where $\mathcal{L}(\cdot, \cdot)$ is a task-specific loss function (e.g., cross-entropy), and w_i denotes the aggregation weight of client i (e.g., proportional to its dataset size).

Prototype. In prototype-based PFL, a prototype is typically defined as the mean representation of all embedded instances belonging to a specific class. For client i, the local prototype P_i^c of class c is computed as:

$$P_i^c = \frac{1}{|\mathcal{D}_i^c|} \sum_{(x,y)\in\mathcal{D}_i^c} f(x; \theta_i), \quad (2)$$

where $\mathcal{D}_i^c$ denotes the set of local samples on client i with label c. These local prototypes P_i^c summarize the semantic representation of class c on client i and guide feature alignment across clients.

Prototype Aggregation. To enable cross-client knowledge sharing, many prototype-based methods aggregate local prototypes into global prototypes using a weighted average. The global prototype P_g^c for class c is typically computed as:

$$P_g^c = \frac{1}{\sum_{i=1}^N |D_i^c|} \sum_{i=1}^{N} |D_i^c| P_i^c. \quad (3)$$

this aggregation scheme ensures that clients with more samples of a particular class have greater influence on the corresponding global prototype, leading to more representative and robust global class representations.

Prototype Aggregation Bias

However, this general global averaging assumes that local prototypes from different clients for the same class lie in the same semantic subspace, which is often violated under non-IID settings. Clients learn divergent prototypes for the same class due to different local biases, leading to aggregation bias and degraded model performance. As shown in Fig. 1, clients may focus on different feature regions (e.g., head, body, bill for pelican), leading to heterogeneous prototypes. Consequently, the global prototype P_g^c may fail to accurately represent the semantic characteristics of any individual client's local data. The global aggregation bias for client i and class c can be quantified as:

$$\text{Bias}_i^c = \mathbb{E}\left[\|P_i^c - P_g^c\| \right] \quad (4)$$

where $\|P_i^c - P_g^c\|$ measures the Euclidean distance between the local and global prototypes. The bias arises because client prototypes P_i^c lie in different low-rank subspaces due to non-IID data distributions. To better characterize this bias, we assume the existence of true expected feature $\mu_\star^c = \mathbb{E}_{(x,y)\sim\mathcal{D}^c}\left[f(x; \theta_\star) \right]$ for class c, where $\mathcal{D}^c$ denotes the true data distribution and $\theta_\star$ is the ideal feature extractor. By applying the triangle inequality, we have $\left\| P_i^c - P_g^c \right\| \leq \|P_i^c - \mu_\star^c\| + \left\| \mu_\star^c - P_g^c \right\|$. Let μ_i^c denote the expected feature representation of class c on client i, then the aggregation bias admits the following upper bound:

$$\mathbb{E}\left[\left\| P_i^c - P_g^c \right\| \right] \leq \underbrace{\frac{\sigma_c}{\sqrt{|D_i^c|}}}_{\text{Sampling Error}} + \underbrace{\|\mu_i^c - \mu_\star^c\|}_{\text{Distribution Shift}} + \underbrace{\sum_{j=1}^{N} w_j^c \left(\frac{\sigma_c}{\sqrt{|D_j^c|}} + \left\| \mu_j^c - \mu_\star^c \right\| \right)}_{\text{Aggregation Error}} \quad (5)$$

where σ_c denotes the standard deviation of feature representations for class c, and w_j^c represents the aggregation weight proportional to the number of class c samples at client j. The sampling error term decreases with increasing sample size, the distribution shift term captures the degree of non-IID between the local and true distributions, and the aggregation error term quantifies the contribution of other clients' biases to the estimation of the current client's prototype.

FedIC: Intra-Class Prototype Clustering

Based on the empirical observations in Fig. 1 and the theoretical formulation of prototype aggregation bias, particularly the distribution shift term and aggregation error term, we identify that different clients tend to focus on distinct feature regions even for the same class. This phenomenon results in inconsistency among local prototypes, making direct aggregation toward a global prototype problematic and ultimately leading to significant aggregation bias. To address this issue, we propose a fine-grained Intra-Class Prototype Clustering mechanism. Specifically, for each class c, we collect local prototypes P_i^c from all clients and perform clustering within the class to form K prototype subspaces, i.e., $\mathcal{G}_1^c, \mathcal{G}_2^c, \ldots, \mathcal{G}_K^c$. Each cluster contains prototypes with similar semantic representations, ensuring that aggregation is only performed among semantically aligned prototypes. To quantify the similarity between prototypes, we use cosine similarity as the metric:

$$Sim\left(P_i^c, P_j^c\right) = \frac{P_i^c \cdot P_j^c}{\|P_i^c\| \|P_j^c\|}, \quad (6)$$

by calculating the cosine distance, we can effectively evaluate whether the features of the clients are highly similar in the same class, and group clients accordingly.

Cluster-wise Prototype Aggregation. Once the intra-class prototype clusters $\mathcal{G}_k^c$ are formed, we perform cluster-wise aggregation to obtain a refined representation for each group. For each cluster $\mathcal{G}_k^c$, the aggregated prototype $\tilde{P}_k^c$ is computed as a weighted average of the member prototypes:

$$\tilde{P}_k^c = \sum_{i \in \mathcal{G}_k^c} \frac{|D_i^c|}{\sum_{j \in \mathcal{G}_k^c} |D_j^c|} P_i^c \tag{7}$$

where the cluster-level prototype $\tilde{P}_k^c$ preserves the local semantics shared among similar clients while avoiding distortion from prototypes that focus on different feature regions. Compared to conventional global averaging, this design eliminates forced alignment across heterogeneous distributions and effectively reduces aggregation bias.

Bias-Aware Class Centroid. To further utilize the semantic consistency provided by cluster-wise aggregation, we define each aggregated prototype $\tilde{P}_k^c$ as an anchor representing a bias-reduced class-level reference in the embedding space. This anchor plays a crucial role in aligning local representations across clients. After each round, client i updates its class alignment centroid A_i^c based on whether it holds samples of class c. If so, the centroid is set to the corresponding cluster prototype $\tilde{P}_k^c$; otherwise, it falls back to the global prototype P_g^c to ensure consistency across all classes:

$$A_i^c = \begin{cases} \tilde{P}_k^c, & \text{if } \sum_{(x,y) \in D_i} \mathbb{I}(y = c) > 0 \\ P_g^c, & \text{otherwise} \end{cases} \tag{8}$$

where $\mathbb{I}(y = c)$ is an indicator function, equal to 1 if the class y is c, otherwise 0. This formula indicates that if client i has data for class c, the class centroid will be updated to the cluster center $\tilde{P}_k^c$, otherwise it will be completed with the global class centroid P_g^c.

Centroid-driven Alignment. Each centroid A_i^c represents a class-specific anchor for client i and is used to align feature embeddings toward semantically consistent subspaces. In addition to the standard supervised classification loss, we introduce a centroid alignment regularization to encourage the embedding of a sample to be close to its class centroid:

$$\mathcal{L}_{\text{align}} = \frac{1}{|\mathcal{D}_i|} \sum_{(x,y) \in \mathcal{D}_i} \| f_i(x; \theta_i) - A_i^y \|^2 \tag{9}$$

the complete objective for client i combines the standard cross-entropy loss with our alignment regularization:

$$\mathcal{L}_i = \mathcal{L}_{\text{CE}}(f_i(x; \theta_i), y) + \lambda \mathcal{L}_{\text{align}} \tag{10}$$

where λ is a hyperparameter controlling the strength of alignment regularization. By clustering prototypes within each class, FedIC ensures that only semantically similar representations are aggregated together. This approach naturally reduces the distribution shift term in Eq. 5, as prototypes in the same cluster share similar feature focuses, leading to more consistent aggregation.

Intra-Cluster Classifier Collaboration

Inspired by FedPAC (Xu, Tong, and Huang 2023), our work focuses on classifier collaboration within prototype subspaces. In addition to improving feature extractors through prototype alignment, we argue that merging classifiers from clients with similar semantic representations can offer additional performance gains. Within each prototype cluster $\mathcal{G}_k^c$, clients share similar semantic focuses for class c, indicating that their locally trained classifiers likely capture complementary decision boundaries that are transferable among cluster members. By restricting collaboration to within-cluster clients, we can leverage the semantic consistency established by our prototype clustering while avoiding harmful interference from heterogeneous representations.

Intra-cluster Collaboration. For each client i belonging to cluster $\mathcal{G}_k^c$, we perform a weighted combination of classifiers from all clients within the same cluster. The enhanced classifier $\hat{\phi}_i$ is computed as:

$$\hat{\phi}_i = \sum_{j \in \mathcal{G}_k^c} \beta_{ij} \phi_j, \quad \text{s.t.} \sum_{j \in \mathcal{G}_k^c} \beta_{ij} = 1 \tag{11}$$

where ϕ_j represents the classifier parameters of client j, and $\beta_{ij} \geq 0$ are the collaboration coefficients that determine the contribution of each cluster member.

Adaptive Weight Learning. To determine the optimal collaboration coefficients β_{ij}, we formulate an optimization problem that minimizes the expected testing loss on client i's local validation set:

$$\beta_i^* = \arg\min_{\beta_i} \mathbb{E}_{(x,y) \sim \mathcal{D}_i} \left[\mathcal{L} \left(\theta, \sum_{j \in \mathcal{G}_k^c} \beta_{ij} \phi_j; x, y \right) \right] \tag{12}$$

this optimization ensures that the collaboration weights are adaptive to each client's specific data characteristics while being constrained within the semantically consistent cluster. Unlike global classifier averaging approaches, our intra-cluster collaboration maintains semantic consistency by restricting knowledge transfer to clients with aligned prototype representations.

Experiment

Experimental Setup

Datasets. We evaluated FedIC on four public datasets: Fashion-MNIST (FMNIST) (Xiao, Rasul, and Vollgraf 2017), CIFAR10 (Krizhevsky, Hinton et al. 2009), CIFAR100 (Krizhevsky, Hinton et al. 2009) , and TinyImagenet (Chrabaszcz, Loshchilov, and Hutter 2017). All data distributions across clients are non-IID, simulating two commonly used non-IID scenarios: the Practical Heterogeneous Setting (Pra) (Li, He, and Song 2021) and the Pathological Heterogeneous Setting (Pat) (Shamsian et al. 2021).

Baselines. We compare our proposed FedIC against several advanced Prototype-based and Clustered PFL methods: FedProto (Tan et al. 2022), FedPAC (Xu, Tong, and Huang 2023), FedCCFA (Chen et al. 2024), IFCA (Ghosh et al. 2020), FeSEM (Long et al. 2023), Fed-CAM (Ma et al. 2023), and CFL-GP (Kim, Kim, and De Veciana 2024).

Methods	Pathological heterogeneous setting			Practical heterogeneous setting			Computation	Communication
	FMNIST	CIFAR10	CIFAR100	FMNIST	CIFAR10	CIFAR100	Time/Iter.	Param./Iter.
IFCA	98.93	92.93	67.82	97.72	90.88	51.89	71.07 s	158.77 MB
FeSEM	99.12	92.52	68.09	97.64	90.54	52.91	62.12 s	158.77 MB
IFCA-CAM	99.10	91.43	65.12	96.77	90.01	47.22	92.12 s	161.97 MB
FeSEM-CAM	98.82	90.20	64.87	96.67	90.16	49.57	87.43 s	161.97 MB
CFL-GP	<u>99.36</u>	93.20	67.22	97.78	91.47	55.89	88.29 s	159.91 MB
FedProto	99.16	92.25	67.31	97.16	90.68	60.39	56.90 s	151.52 MB
FedPAC	99.28	93.32	<u>72.10</u>	97.79	91.13	<u>61.74</u>	110.67 s	160.01 MB
FedCCFA	99.20	<u>93.38</u>	67.93	<u>97.86</u>	<u>91.65</u>	56.72	95.51 s	156.50 MB
FedIC (Ours)	**99.42**	**93.97**	**75.50**	**98.10**	**92.32**	**65.89**	125.42 s	156.92 MB
Improvement (%)	0.06 ↑	0.59 ↑	3.40 ↑	0.24 ↑	0.67 ↑	**4.15** ↑	–	–

Table 1: The performance comparisons of FedIC with SOTA Clustered and Prototype-based PFL methods in the Pat and Pra setting (Computation: 20-client iteration time; Communication: parameter size transmitted per iteration, on CIFAR100).

Methods	CIFAR10					CIFAR100				
	CNN	ResNet4	ResNet10	ResNet18	ResNet34	CNN	ResNet4	ResNet10	ResNet18	ResNet34
IFCA	90.88	91.37	91.72	91.49	91.23	51.89	61.92	56.31	54.87	53.62
FeSEM	90.54	91.15	91.58	91.27	90.98	52.91	63.74	57.86	56.03	54.89
IFCA-CAM	90.01	90.72	91.19	90.85	90.61	47.22	58.16	53.42	51.78	50.25
FeSEM-CAM	90.16	90.88	91.34	91.02	90.67	49.57	60.33	55.09	53.64	52.13
CFL-GP	<u>91.65</u>	92.09	92.61	92.27	91.92	55.89	<u>66.82</u>	61.37	59.93	58.44
FedProto	90.61	91.24	90.20	89.89	89.71	51.47	59.67	59.69	56.67	55.64
FedPAC	91.13	89.41	90.66	89.40	87.57	<u>61.74</u>	60.13	62.32	59.75	56.76
FedCCFA	91.47	<u>92.31</u>	<u>92.84</u>	**92.89**	<u>92.16</u>	56.72	65.13	<u>62.75</u>	<u>61.28</u>	<u>59.87</u>
FedIC (Ours)	**92.32**	**92.97**	**93.49**	<u>92.83</u>	**92.81**	**65.89**	**75.91**	**72.54**	**72.12**	**69.83**
Improvement (%)	0.67 ↑	0.66 ↑	0.65 ↑	0.06 ↓	0.65 ↑	4.15 ↑	9.09 ↑	10.79 ↑	**10.84** ↑	9.96 ↑

Table 2: Performance of FedIC and baselines under varying backbone depths on CIFAR10 and CIFAR100 under the Pra Setting.

Training Details. To ensure a fair comparison, we adopt the same training settings as baseline methods. For all methods, we employ the SGD optimizer with a learning rate of 0.005. Across all datasets, the batch size is set to 10, the global communication rounds are set to $T = 800$, and local training epochs to $E = 1$, continuing until all methods empirically converge. The default number of clients is 20 with a 100% client participation rate.

Performance Evaluation with the SOTA

FedIC achieves superior performance across all datasets. As shown in Tab. 1, FedIC outperforms all Clustered and Prototype-based baselines under both pathological and practical heterogeneous settings. In the Pat case, it achieves accuracies of 99.42%, 93.97%, and 75.50% in three datasets, respectively, surpassing the best baselines by up to 3.40%. The largest improvement occurs on CIFAR100, highlighting FedIC's robustness in complex label spaces and severe non-IID scenarios. Under the Pra setting, FedIC outperforms FedCCFA by 4.15% on CIFAR100. These gains validate the effectiveness of FedIC, which mitigates aggregation bias by modeling intra-class feature divergence and restricting prototype aggregation to semantically aligned clusters.

The overall overhead of FedIC is acceptable. Tab. 1 also report the communication and computation overhead of each method on CIFAR100 under the Pat setting. FedIC incurs a moderate increase in iteration time (125.42 s) compared to FedPAC (110.67 s), while maintaining a comparable communication cost (156.92 MB vs. 160.01 MB). The additional computation is primarily attributed to server-side operations, including intra-class similarity computation and subspace-level collaborative optimization, and does not impose any extra burden on clients. Considering the 3.40% accuracy gain over FedPAC, the 13% increase in iteration time is an acceptable trade-off for FedIC.

FedIC achieves greater performance gains with deeper backbones. As shown in Tab. 2, the performance advantage of FedIC becomes increasingly pronounced as the model depth increases. On CIFAR100, for example, FedIC outperforms the best baseline by 4.15% with a shallow CNN, but the improvement rises to 10.84% and 10.79% with ResNet18 and ResNet10, respectively. This performance trend can be attributed to the higher representational capacity of deeper networks, which tend to amplify the intra-class feature divergence across clients under non-IID conditions. In such cases, global prototype aggregation becomes more vulnerable to semantic misalignment, leading to degraded performance. However, our fine-grained intra-class

Methods	40 Clients				60 Clients			
	Pathological		Practical		Pathological		Practical	
	CIFAR10	CIFAR100	CIFAR10	CIFAR100	CIFAR10	CIFAR100	CIFAR10	CIFAR100
IFCA	89.50	61.13	87.85	53.69	88.37	59.16	86.45	50.82
FeSEM	88.74	62.58	87.59	55.55	87.54	60.70	86.43	53.00
IFCA-CAM	88.12	58.01	86.98	48.67	87.12	56.83	85.55	46.50
FeSEM-CAM	88.75	59.55	87.01	54.74	87.45	58.04	86.23	48.33
CFL-GP	90.35	65.22	88.51	55.72	89.15	63.94	87.43	54.53
FedCCFA	89.95	65.93	89.15	55.92	89.85	62.95	88.03	55.63
FedIC (Ours)	**91.40**	**73.50**	**90.96**	**65.89**	**90.77**	**70.20**	**89.16**	**65.36**
Improvement (%)	1.05 ↑	7.57 ↑	1.81 ↑	**9.97** ↑	0.92 ↑	6.26 ↑	1.13 ↑	**9.73** ↑

Table 3: The performance comparison of FedIC with SOTA PFL methods under varying client scales and non-IID settings.

Methods	$\beta = 0.01$	$\beta = 0.1$	$\beta = 0.5$
IFCA	42.17	51.89	55.24
FeSEM	43.85	52.91	57.36
IFCA-CAM	38.94	47.22	50.87
FeSEM-CAM	40.62	49.57	53.15
CFL-GP	47.25	55.89	60.17
FedCCFA	48.03	56.72	62.85
FedIC (Ours)	**59.82**	**65.89**	**71.43**
Improvement	**11.79** ↑	9.17 ↑	8.58 ↑

Table 4: Performance under different β on CIFAR100.

Methods	Pathological	Practical
IFCA	48.12	44.72
FeSEM	49.36	45.58
IFCA-CAM	49.75	45.01
FeSEM-CAM	49.77	45.23
CFL-GP	52.49	48.89
FedCCFA	53.19	49.51
FedIC (Ours)	**57.12**	**53.23**
Improvement(%)	3.93 ↑	3.72 ↑

Table 5: The test accuracy(%) comparison on TinyImagenet.

clustering and collaborative classifier optimization mitigate the negative effects of this variability, aligning semantically consistent prototypes and reducing aggregation bias.

FedIC is scalable across different client amounts. Tab. 3 illustrates the performance of FedIC and other SOTA PFL methods under two different client scales (40 and 60 clients), across both Pat and Pra heterogeneous settings. FedIC consistently outperforms all baselines regardless of the number of participating clients. Notably, as the number of clients increases, the performance of most baselines degrades substantially, especially on CIFAR100, due to intensified client heterogeneity and exacerbated intra-class variation. In contrast, FedIC remains robust: on CIFAR100 under the practical setting, it achieves 65.89% with 40 clients and 65.36% with 60 clients, demonstrating a minimal 0.53% drop, while still outperforming the second-best method by 9.73%. This robustness stems from FedIC's intra-class clustering strategy, which effectively mitigates the impact of client-specific biases by isolating and aggregating semantically consistent prototypes, even under large-scale federated deployments.

FedIC maintains stable performance across different levels of data heterogeneity. Tab. 4 presents the performance of FedIC and baseline methods on CIFAR100 under the practical heterogeneous setting, with varying non-IID levels controlled by the Dirichlet parameter β. A smaller β indicates more severe label distribution skew across clients. As β decreases from 0.5 to 0.01, all baseline methods experience notable performance degradation, with accuracy dropping by over 10% in most cases. In contrast, FedIC consistently achieves the best performance across all levels of heterogeneity. Notably, under the most challenging case of $\beta = 0.01$, FedIC improves over the second-best method by 11.79%. These results suggest that FedIC can still extract useful representations even when client data distributions have little overlap. This is because the adaptive classifier collaboration ensures that the model can still leverage informative local classifiers despite increased statistical heterogeneity. The performance improvements become smaller as β increases, but FedIC still outperforms all baselines by a clear margin (e.g., 8.58% at $\beta = 0.5$), indicating the robustness of its design under varying non-IID conditions.

FedIC also performs superiorly on the TinyImageNet dataset. Tab. 5 reports the performance comparison on TinyImageNet, a more challenging dataset with higher class granularity and larger feature diversity. Under both pathological and practical heterogeneous settings, FedIC consistently outperforms all baselines. Specifically, it achieves 57.12% accuracy under the pathological setting, surpassing the second-best method by 3.93%, and 53.23% accuracy under the practical setting, with a 3.72% improvement. These results confirm that FedIC remains effective even when facing high-dimensional, fine-grained data distributions.

FedIC captures semantic structures across clients and enables localized aggregation and collaboration. As

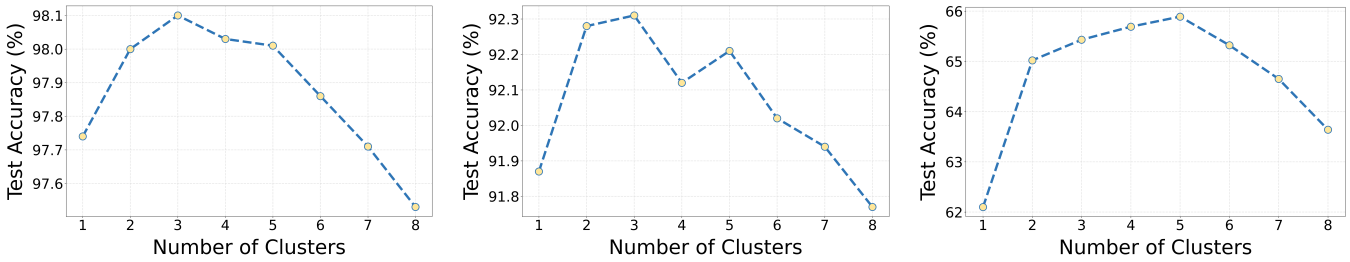
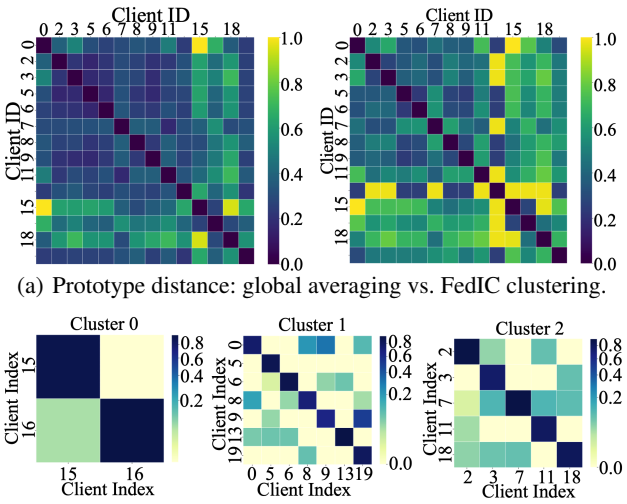


Figure 2: Effect of cluster number on accuracy across FMNIST, CIFAR10, and CIFAR100.

shown in Fig. 3(a), we visualize the pairwise distances among class-0 prototypes from 20 clients on CIFAR100 after convergence. The heatmap on the left corresponds to the global prototype averaging baseline, where distances between prototypes are uniformly low. This indicates that client prototypes are forced to align in a single direction, suppressing their inherent diversity and causing aggregation bias. In contrast, the FedIC-generated heatmap (right) clearly reveals multiple prototype clusters with higher inter-cluster dissimilarity. Fig. 3(b) shows that classifier collaboration is also restricted within clusters, where clients primarily interact with semantically aligned peers. These patterns demonstrate that FedIC effectively supports fine-grained personalization by restricting both prototype aggregation and classifier fusion to meaningful client groups.

Effect of the cluster number K. As shown in Fig. 2, the relationship between the number of clusters and test accuracy is non-monotonic, with performance peaking at an optimal cluster count. Beyond this point, further partitioning leads to reduced accuracy, as excessively fine-grained clustering weakens the statistical reliability of each sub-cluster and lowers the effectiveness of collaborative learning. In our experiments, conducted with 20 clients, the use of 3 clusters yielded the highest performance on FMNIST and CIFAR10, while 5 clusters achieved optimal results on CIFAR100.

Ablation study of core components. Fig. 4 shows that both prototype-based clustering and classifier collaboration significantly contribute to the overall performance of FedIC. Compared with the baseline using global prototype aggregation, adding only the clustering module yields a notable improvement in early convergence and final accuracy. Incorporating classifier collaboration alone achieves even better results by progressively refining client-specific decision boundaries. When combined, the two components yield the highest accuracy and most stable convergence, demonstrating their complementary benefits in mitigating prototype misalignment and enhancing personalized prediction.



(a) Prototype distance: global averaging vs. FedIC clustering.

(b) Intra-cluster classifier collaboration weights under FedIC.

Figure 3: Visualization of prototype similarity and intra-cluster classifier collaboration patterns.

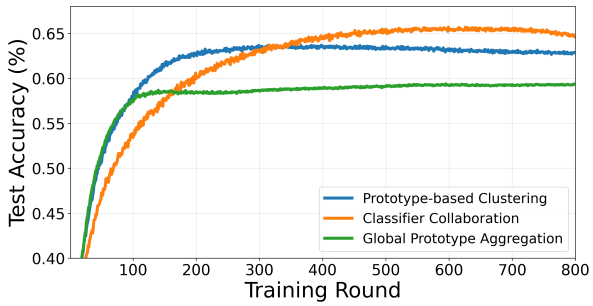


Figure 4: Ablation study of core components on CIFAR100.

Conclusion

This work tackles the prototype aggregation bias in PFL by analyzing the underlying intra-class feature divergence across clients. We observe that, due to statistical heterogeneity, clients often extract semantically distinct features for the same class, which makes global prototype alignment suboptimal and leads to performance degradation. To address this, we propose the FedIC, which clusters client prototypes based on intra-class similarity to form semantically coherent subspaces. Within each subspace, FedIC performs localized aggregation and adaptive classifier collaboration, effectively mitigating semantic conflicts and preserving meaningful client-specific representations. Extensive experiments across diverse datasets and non-IID settings demonstrate that FedIC consistently outperforms SOTA methods, particularly in scenarios with severe distribution shifts. Future work will explore more fine-grained and dynamic subspace partitioning to further enhance model personalization.

Acknowledgements

We appreciate constructive feedback from anonymous reviewers and meta-reviewers. This work was supported by the National Natural Science Foundation of China (62172442, 62172451), China Scholarship Council, and High Performance Computing Center of Central South University.

References

Bae, W.; Noh, J.; and Kim, G. 2020. Rethinking class activation mapping for weakly supervised object localization. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*, 618–634. Springer.

Belyi, M.; Dzialo, C.; Dwivedi, C.; Muppidi, P.; and Shimizu, K. 2023. Personalized dense retrieval on global index for voice-enabled conversational systems. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, 83–92.

Chen, J.; Xue, J.; Wang, Y.; Liu, Z.; and Huang, L. 2024. Classifier Clustering and Feature Alignment for Federated Learning under Distributed Concept Drift. *Advances in neural information processing systems*.

Cho, Y. J.; Wang, J.; and Joshi, G. 2022. Towards understanding biased client selection in federated learning. In *International Conference on Artificial Intelligence and Statistics*, 10351–10375. PMLR.

Chrabaszcz, P.; Loshchilov, I.; and Hutter, F. 2017. A downsampled variant of imagenet as an alternative to the cifar datasets. *arXiv preprint arXiv:1707.08819*.

Deng, Y.; Lyu, F.; Xia, T.; Zhou, Y.; Zhang, Y.; Ren, J.; and Yang, Y. 2024. A communication-efficient hierarchical federated learning framework via shaping data distribution at edge. *IEEE/ACM Transactions on Networking*, 32(3): 2600–2615.

Fu, L.; Huang, S.; Lai, Y.; Zhang, C.; Dai, H.-N.; Zheng, Z.; and Chen, C. 2025. Federated domain-independent prototype learning with alignments of representation and parameter spaces for feature shift. *IEEE Transactions on Mobile Computing*.

Ghosh, A.; Chung, J.; Yin, D.; and Ramchandran, K. 2020. An efficient framework for clustered federated learning. *Advances in neural information processing systems*, 33: 19586–19597.

Huang, W.; Wang, D.; Ouyang, X.; Wan, J.; Liu, J.; and Li, T. 2024. Multimodal federated learning: Concept, methods, applications and future directions. *Information Fusion*, 112: 102576.

Huang, W.; Ye, M.; Shi, Z.; Li, H.; and Du, B. 2023. Rethinking federated learning with domain shift: A prototype view. In 2023 IEEE. In *CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 16312–16322.

Jia, P.; Zhang, D.; Zhang, L.; Han, D.; Zhang, J.; and Sang, Y. 2024. ProtoFedLA: Prototype Guided Personalized Federated Learning Based on Localized Aggregation across Heterogeneous Clients. In *2024 IEEE International Symposium on Parallel and Distributed Processing with Applications (ISPA)*, 196–203. IEEE.

Jothimurugesan, E.; Hsieh, K.; Wang, J.; Joshi, G.; and Gibbons, P. B. 2023. Federated learning under distributed concept drift. In *International Conference on Artificial Intelligence and Statistics*, 5834–5853. PMLR.

Kim, H.; Kim, H.; and De Veciana, G. 2024. Clustered federated learning via gradient-based partitioning. In *Forty-first International Conference on Machine Learning*.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images.(2009).

Li, Q.; He, B.; and Song, D. 2021. Model-contrastive federated learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10713–10722.

Li, Z.; Shang, X.; He, R.; Lin, T.; and Wu, C. 2023. No fear of classifier biases: Neural collapse inspired federated learning with synthetic and fixed classifier. In *Proceedings of the IEEE/CVF international conference on computer vision*, 5319–5329.

Long, G.; Xie, M.; Shen, T.; Zhou, T.; Wang, X.; and Jiang, J. 2023. Multi-center federated learning: clients clustering for better personalization. *World Wide Web*, 26(1): 481–500.

Luo, G.; Liu, T.; Lu, J.; Chen, X.; Yu, L.; Wu, J.; Chen, D. Z.; and Cai, W. 2023. Influence of data distribution on federated learning performance in tumor segmentation. *Radiology: Artificial Intelligence*, 5(3): e220082.

Ma, J.; Zhou, T.; Long, G.; Jiang, J.; and Zhang, C. 2023. Structured federated learning through clustered additive modeling. *Advances in Neural Information Processing Systems*, 36: 43097–43107.

Meng, L.; Qi, Z.; Wu, L.; Du, X.; Li, Z.; Cui, L.; and Meng, X. 2024. Improving global generalization and local personalization for federated learning. *IEEE Transactions on Neural Networks and Learning Systems*, 36.

Oh, J.; Kim, S.; and Yun, S.-Y. 2021. Fedbabu: Towards enhanced representation for federated image classification. *arXiv preprint arXiv:2106.06042*.

Pu, R.; Yu, L.; Zhan, S.; Xu, G.; Zhou, F.; Ling, C. X.; and Wang, B. 2025. FedELR: When federated learning meets learning with noisy labels. *Neural Networks*, 187: 107275.

Qi, X.; Li, M.; Zhou, S.; Feng, W.; and Qi, Z. 2025a. Federated Learning for Science: A Survey on the Path to a Trustworthy Collaboration Ecosystem.

Qi, Z.; Meng, L.; Li, Z.; Hu, H.; and Meng, X. 2025b. Cross-silo feature space alignment for federated learning on clients with imbalanced data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 19986–19994.

Sattler, F.; Müller, K.-R.; and Samek, W. 2020. Clustered federated learning: Model-agnostic distributed multitask optimization under privacy constraints. *IEEE transactions on neural networks and learning systems*, 32(8): 3710–3722.

Shamsian, A.; Navon, A.; Fetaya, E.; and Chechik, G. 2021. Personalized federated learning using hypernetworks. In *International conference on machine learning*, 9489–9502. PMLR.

Song, S.; Zheng, H.; Hu, Z.; Zheng, M.; Yang, L.; and Xu, A. 2025. GradPFL: Gradient-Driven Adaptive Clustering in Personalized Federated Learning. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

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*, volume 36, 8432–8440.

Xiao, H.; Rasul, K.; and Vollgraf, R. 2017. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*.

Xiong, Y.; Wang, R.; Cheng, M.; Yu, F.; and Hsieh, C.-J. 2023. Feddm: Iterative distribution matching for communication-efficient federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16323–16332.

Xu, J.; Tong, X.; and Huang, S.-L. 2023. Personalized federated learning with feature alignment and classifier collaboration.

Ye, C.; Zheng, H.; Hu, Z.; and Zheng, M. 2023. Pfedsa: Personalized federated multi-task learning via similarity awareness. In *2023 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, 480–488. IEEE.

Yi, L.; Wang, G.; Liu, X.; Shi, Z.; and Yu, H. 2023. Fedgh: Heterogeneous federated learning with generalized global header. In *Proceedings of the 31st ACM international conference on multimedia*, 8686–8696.

Yi, L.; Yu, H.; Ren, C.; Wang, G.; Li, X.; et al. 2024. Federated model heterogeneous matryoshka representation learning. *Advances in Neural Information Processing Systems*, 37: 66431–66454.

Zhang, C.; Long, G.; Zhou, T.; Yan, P.; Zhang, Z.; Zhang, C.; and Yang, B. 2023a. Dual personalization on federated recommendation. *arXiv preprint arXiv:2301.08143*.

Zhang, J.; Duan, Y.; Niu, S.; Cao, Y.; and Lim, W. Y. B. 2024. Enhancing federated domain adaptation with multi-domain prototype-based federated fine-tuning. *arXiv preprint arXiv:2410.07738*.

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, X.; Zhang, B.; Yu, W.; and Kang, X. 2023c. Federated deep learning with prototype matching for object extraction from very-high-resolution remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 1–16.

Zheng, H.; Hu, Z.; Yang, L.; Zheng, M.; Xu, A.; and Wang, B. 2025a. ConFREE: Conflict-free Client Update Aggregation for Personalized Federated Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 22875–22883.

Zheng, H.; Hu, Z.; Yang, L.; Zheng, M.; Xu, A.; and Wang, B. 2025b. FedCALM: Conflict-aware Layer-wise Mitigation for Selective Aggregation in Deeper Personalized Federated Learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 15444–15453.

Zhou, T.; and Wang, W. 2024. Prototype-based semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhu, G.; Liu, X.; Tang, S.; and Niu, J. 2023. Aligning before aggregating: Enabling communication efficient cross-domain federated learning via consistent feature extraction. *IEEE Transactions on Mobile Computing*, 23(5): 5880–5896.

Zuo, Y.; Chen, Z.; Feng, J.; and Fan, Y. 2025. Federated Learning and Optimization for Few-Shot Image Classification. *Computers, Materials & Continua*, 82(3).