



ARTICLE

FedCANA: Communication-Efficient Adaptive Neighborhood Aggregation for Personalized Federated Learning under Non-IID Data

Yao Pu, Rou Zhou and Yuling Chen*

State Key Laboratory of Public Big Data, College of Computer Science and Technology, Guizhou University, Guiyang, China

*Corresponding Author: Yuling Chen. Email: ylchen3@gzu.edu.cn

Received: 01 June 2026; Accepted: 26 August 2026; Published: 22 October 2026

ABSTRACT: Personalized federated learning (PFL) aims to address the insufficient adaptability of a single global model caused by Non-IID data. It improves client-specific performance by learning customized models for different clients. However, existing PFL methods based on similarity modeling often struggle to balance personalization performance and communication efficiency under Non-IID settings. To address this problem, this paper proposes FedCANA (Communication-Efficient Adaptive Neighborhood Aggregation), a communication-efficient personalized federated learning method for Non-IID data. First, FedCANA adopts Residual-Spatial Attention Collaborative (RSAC) module to strengthen the clients' local representation ability, which provides a more stable model basis for subsequent similarity estimation. Second, FedCANA employs Saliency-Guided Differential Descriptor (SGDD) to select and upload critical residual information, which reduces redundant communication while preserving client-specific collaborative signals. Finally, FedCANA introduces Entropy-driven Adaptive Neighborhood Aggregation (EANA) to dynamically select similar clients for collaborative aggregation, which reduces the interference from less relevant clients and improves personalized model performance. Extensive experiments on three standard benchmark datasets under three Non-IID settings demonstrate that FedCANA achieves state-of-the-art personalized accuracy on most tasks, and yields up to 9.26% accuracy improvement on the highly heterogeneous CIFAR-100 dataset. Meanwhile, it maintains significantly lower communication overhead than mainstream similarity-based personalized federated learning methods, achieving a favorable trade-off between accuracy and communication efficiency.

KEYWORDS: Personalized federated learning; data heterogeneity; communication efficiency; machine learning; adaptive aggregation

1 Introduction

In recent years, the rapid development of mobile networks, edge computing, and IoT has led to the continuous generation of massive local data on smartphones, wearable devices, smart home devices, and various intelligent end devices. These data usually contain user behavior, device status, and environmental sensing information. Directly uploading such data to a central server for centralized training may cause privacy leakage and high communication overhead. Federated Learning (FL) provides an effective paradigm for distributed model training in privacy-sensitive scenarios [1]. It trains models locally on clients and aggregates model updates on the server [2]. However, in real-world scenarios, client data often exhibits Non-IID characteristics owing to differences in user behavior, hardware conditions, sampling strategies, and data scales [3]. Such statistical heterogeneity makes it difficult for a single global model in traditional FL methods, such as FedAvg [1], to adapt to the local data distributions of all clients simultaneously. Therefore, improving

model adaptability to different clients under Non-IID data distributions has become an important research problem in federated learning.

Personalized Federated Learning (PFL) has been proposed as a solution to this issue. It aims to learn personalized models that better match the local data distribution of each client while preserving cross-client knowledge sharing. Existing PFL methods have improved model performance in Non-IID scenarios from different perspectives. Among them, methods based on client relationship modeling have received considerable attention. The basic idea of these methods is to improve personalization performance by sharing information among neighboring clients with similar data distributions or features. For example, FedFomo [4] performs weighted aggregation according to the contribution of other client models to the local task. FedDWA [5] uses client models and guiding models to dynamically adjust aggregation weights, thereby generating personalized models for different clients. These methods show that effectively exploiting collaborative relationships among similar clients can improve the personalization performance of PFL in heterogeneous data environments.

Nevertheless, existing PFL methods based on client relationship modeling still suffer from two main limitations. First, to model client relationships, they usually require substantial additional model interactions and the upload of redundant guidance information [4,6], which increases communication overhead. Second, in highly heterogeneous and dynamic environments, the similarity structure among clients may evolve as data distributions, client participation, and training states change. Fixed or coarse-grained collaboration strategies [7,8] may introduce less relevant clients, thereby degrading the quality of personalized aggregation.

Therefore, modeling effective collaborative relationships among clients with lower communication cost and enabling adaptive client aggregation remain important problems for PFL in Non-IID scenarios. To address these problems, this paper proposes FedCANA (Communication-Efficient Adaptive Neighborhood Aggregation), a communication-efficient personalized federated learning framework. The core idea of FedCANA is to focus on high-information features on the client side, select and upload key differential descriptors during communication, and adaptively adjust the aggregation neighborhood on the server side based on the uncertainty in the client relationship structure. In this way, FedCANA enables communication-efficient collaborative aggregation among similar clients. Specifically, this paper first designs a Residual-Spatial Attention Collaborative (RSAC). This module improves local model training through adaptive enhancement of local features and improved gradient propagation. It also enhances the robustness of local adaptation and provides a more reliable basis for subsequent client similarity modeling. Second, this paper proposes the Saliency-Guided Differential Descriptor (SGDD), which selects key residual information in the guidance channel according to saliency. SGDD uses compact differential descriptors to transmit the information required for personalized collaboration, thereby reducing redundant communication overhead. Finally, this paper designs Entropy-driven Adaptive Neighborhood Aggregation (EANA), which uses entropy to measure the concentration of the client relationship structure in the current round. Based on this measurement, EANA dynamically adjusts the collaborative neighborhood, enables adaptive aggregation among similar clients, reduces the interference of less relevant clients, and improves both the generalization capability and personalization performance of the model.

We summarize our contributions as follows:

- We propose FedCANA, a communication-efficient personalized federated learning framework. FedCANA follows a three-stage design, including RSAC, SGDD, and EANA. It improves personalized model performance while reducing additional communication overhead.
- We design a Residual-Spatial Attention Collaborative (RSAC) module that enables clients to extract more informative local feature representations. This module improves local adaptation under heterogeneous data and provides a more reliable basis for client similarity modeling.

- We propose the Saliency-Guided Differential Descriptor (SGDD) and Entropy-driven Adaptive Neighborhood Aggregation (EANA). SGDD uses sparse differential information to reduce redundant communication, while EANA dynamically adjusts the collaborative neighborhood to reduce the interference from less relevant clients.
- Experimental results on three benchmark datasets under three Non-IID settings show that FedCANA outperforms the baseline methods and achieves more stable performance in highly heterogeneous scenarios. We further assess the contribution of each component by conducting ablation experiments.

The subsequent sections are structured as follows. [Section 2](#) surveys the three research directions most pertinent to this work. [Section 3](#) formulates the problem definition and optimization goal of personalized federated learning. [Section 4](#) elaborates on the proposed framework and its core modules. [Section 5](#) presents the experimental evaluation together with the corresponding analysis. [Section 6](#) summarizes the paper.

2 Related Work

2.1 Federated Learning with Non-IID Data

Federated learning (FL) [9–12] is a distributed machine learning framework that allows multiple clients to jointly optimize a shared model without directly disclosing their raw data. In this setting, client data are usually Non-IID. Under such conditions, the traditional FedAvg [1] suffers from client drift. This problem can further lead to slow convergence and reduced accuracy on local clients. Researchers have introduced a variety of enhanced approaches that aim to tackle both efficiency and heterogeneity. Among them, regularization-based frameworks, represented by FedProx [13], add regularization terms [14] to the local objective function. As a result, the divergence of local updates from the global model is restricted.

To address the problem of optimization stability, FedAvgM [15] introduces server-side momentum to smooth the aggregation process and thus cope with data heterogeneity. SCAFFOLD [16], in contrast, adopts variance reduction to effectively mitigate client drift. Beyond standard optimization, recent studies have also focused on structural adaptation and personalization. ATHENA-FL [17] addresses statistical heterogeneity via a one-vs.-all task decomposition strategy, which splits multi-class tasks into binary sub-tasks to alleviate Non-IID distribution mismatch. In the context of system heterogeneity, FedLGA [18] leverages local gradient approximation to bridge discrepancies in model updates. In addition, FedeCouple [19] achieves a fine-grained trade-off between global generalization and local adaptability through dynamic knowledge distillation and local anchors. In multi-source input scenarios (e.g., audio-visual and textual data), data heterogeneity also manifests as quality degradation issues such as modality absence, temporal misalignment, and signal-to-noise ratio (SNR) fluctuations. For instance, FedUAF [20] introduces uncertainty-aware multimodal fusion, which performs fine-grained evaluation of modality-wise reliability and adaptively down-weights low-confidence modalities to alleviate the negative effects of quality discrepancies on the global model. Targeting audio-visual sequential data, FedBayesMamba [21] further integrates Bayesian uncertainty estimation with selective state space models. It dynamically calibrates temporal confidence during long-range dependency capture, thereby enhancing the model's robustness under dynamic noise conditions.

While these methods improve global robustness in heterogeneous settings, they neglect per-client personalization—which must be a primary objective, since global gains cannot offset local degradation from data mismatches.

2.2 Personalized Federated Learning

The primary purpose of PFL [22–24] is to cope with the heterogeneous data distributions across clients, which constitutes a major challenge in federated learning. In addition, it aims to balance collaborative global knowledge with personalized performance at the client level.

To address the challenges outlined above, a variety of approaches have been introduced. These techniques can generally be grouped into the following categories. (1) The simplest personalization strategy is global model fine-tuning [25]. Once the aggregated global model is received, each client performs a few steps of fine-tuning with its local data. Building on this idea, Per-FedAvg [26] reformulates PFL as a model-agnostic meta-learning (MAML) problem. Its goal is to train a meta-model that can quickly adapt with only a small amount of local data. In addition, clustered and label-aware federated meta-learning, namely CL-FML [27], specifically addresses class-imbalanced tasks by exploiting fast convergence and lightweight data augmentation mechanisms. (2) Model decoupling-based methods. These methods assume that model parameters can be divided into a shared generic component and a personalized private component. pFedMe [14] uses regularization to decouple updates of the personalized model from those of the global model. FedRoD [28], by contrast, explicitly separates the model into a feature extractor and a personalized head, and then balances the optimization of the two parts. (3) Neighborhood collaboration and clustering-based methods. The core idea of this line of work is that each client only needs to collaborate with similar neighbors rather than participate in global aggregation. ClusterFL [7] divides clients into several groups through iterative clustering, and clients within the same group share a model to handle multitask problems. Both SFL [29] and L2C [6] introduce graph structures and guide personalized aggregation by learning a collaboration graph among clients.

2.3 Sparsification in Personalized Federated Learning

Sparsification techniques aim to substantially reduce communication volume by transmitting only the most important parts of model updates. Early work focused on gradient sparsification. For instance, the method introduced by Lin et al. [30] integrates gradient sparsification together with momentum correction and local gradient clipping, achieving an extremely high compression ratio without sacrificing accuracy. Later studies showed that directly sparsifying model weights or updates is equally effective: Sattler et al. [31] systematically investigated methods including Sparse Ternary Compression (STC), demonstrating that combining Top-K sparsification and quantization can reduce communication by over 99% while maintaining superior performance in federated settings. Recently, sparsification research has extended to complex scenarios. Li et al. [32] proposed a sparse-gradient federated collaborative learning framework for resource-constrained and heterogeneous data settings. Wang et al. [33] introduced PacTrain, which combines model pruning and adaptive sparse gradient compression, thereby improving collective communication efficiency in distributed deep learning while further pushing the efficiency boundary. In the context of PFL, sparsification must consider not only communication efficiency but also the preservation of personalized signals that are critical to each client. If the sparsification process discards a large number of parameters indiscriminately, it may damage the model's local adaptation ability. Therefore, designing a sparsification strategy that can both compress communication efficiently and preserve personalized information has become a new challenge.

3 Problem Formulation

In FL, the system consists of N clients, where the i -th client holds a local dataset D_i that follows its own distribution P_i . Unlike traditional federated learning, which aims to find a globally unified optimal

model, PFL is designed to learn, for each client i , an individualized model ω_i that adequately matches the distributional properties of the local data.

To quantify the adaptation performance of the personalized model, let $\tilde{f}_i(\omega_i, \xi)$ denote the loss function of client i on a single sample ξ . The theoretically optimal personalized model for client i is defined as:

$$\omega_i^* = \arg \min_{\omega_i} \mathbb{E}_{\xi \sim P_i} [\tilde{f}_i(\omega_i; \xi)], \quad (1)$$

The optimal model ω_i^* is also referred to as the guidance model of client i . It can typically be derived via a single local adaptation step, and represents the optimal personalized update direction for client i in the current communication round. However, in practical deployment scenarios, independently training a complete personalized model for each client incurs extremely excessive computational and communication costs. To address this issue, this paper achieves efficient optimization of personalized models for individual clients by mining similarities among clients and dynamically adjusting aggregation weights. Specifically, for each client i , the optimization objective constructed in this paper is as follows:

$$\min_{\alpha_{i,1}, \alpha_{i,2}, \dots, \alpha_{i,N}} \left\| \omega_i^* - \sum_{j=1}^N \alpha_{ij} \omega_j \right\|^2 \quad (2)$$

where ω_j represents the parameters of the local model held by client j , and $\alpha_{i,j}$ is the aggregation weight coefficient assigned by client i to the model of client j . The core idea of this objective is to minimize the discrepancy between the personalized model parameters of client i and the weighted aggregation of multiple client models by selecting appropriate aggregation weights.

However, in highly Non-IID scenarios, directly relying on local training often leads to unstable one-step adaptation, while full interaction and aggregation across all clients introduces significant communication burden and incorporates interference mismatched with the target client distribution.

4 Proposed Model

4.1 Overview of FedCANA

This section presents a systematic overview of the proposed method's workflow. Fig. 1 visualizes the core framework of the approach, and Algorithm 1 provides the detailed pseudocode for its implementation. The overall procedure can be divided into the following stages:

Global model initialization and distribution. For $t = 0$, corresponding to the first communication round, the server-initialized global model ω_0 is distributed to a randomly selected client subset $S_t \subseteq \{1, 2, \dots, N\}$.

Client-side local personalized training and one-step adaptation. After receiving the global model, each selected client i performs local training on its private dataset D_i . After completing local training and obtaining ω_i^t , each client conducts an additional local adaptive update to obtain the guidance model ω_i^* . This guidance model is used only to generate differential descriptor signals in subsequent stages. The update formula is shown in Eq. (3):

$$\omega_i^* = \omega_i^t - \eta_i \nabla f_i(\omega_i^t, D_i) \quad (3)$$

where ω_i^t denotes the local model obtained with RSAC, η_i represents client i 's learning rate, and $f_i(\omega_i^t; D_i)$ is the local loss function evaluated on D_i .

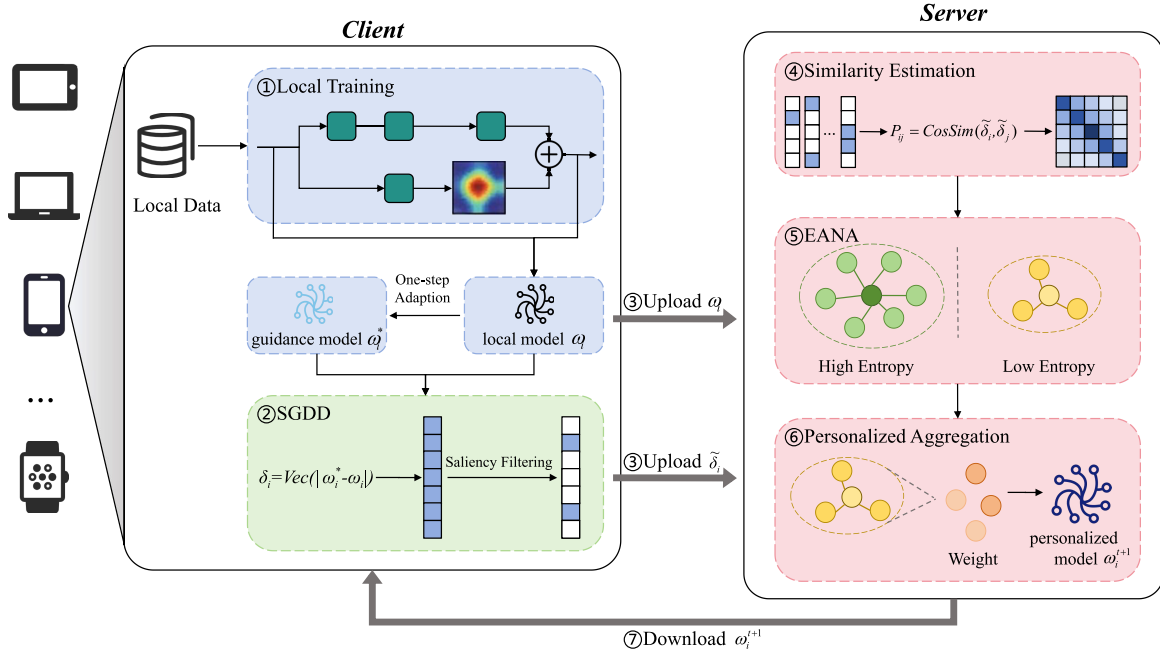


Figure 1: Overall framework of FedCANA. Each client first performs local training and constructs a sparse differential descriptor based on the local model and the guidance model. The local model and sparse differential descriptor are then uploaded to the server. The server performs inter-client similarity estimation, entropy-driven adaptive neighborhood selection, and personalized aggregation, and then downloads the updated personalized model to the client.

Model upload and similarity computation. After local training is completed, each client uploads its local model ω_i^t together with the saliency-guided sparse differential descriptor $\tilde{\delta}_i^t$ (details in Section 4.3) to the server. Based on the uploaded $\tilde{\delta}_i^t$, the server generates the inter-client similarity matrix, which can be expressed as

$$p_{ij} = \text{CosSim}(\tilde{\delta}_i^t, \tilde{\delta}_j^t) \quad (4)$$

Server-side personalized aggregation. By exploiting the similarity matrix p_{ij} , the server derives aggregation weights α_{ij}^t for each target client j , and dynamically determines the neighborhood size M_t for the current round via EANA. The server then selects the most relevant collaborative client set M_j^t for target client j , and generates its personalized model for the next round as

$$\omega_j^{t+1} = \sum_{i \in M_j^t} \alpha_{ij}^t \omega_i^t, \quad (5)$$

Iterative loop update. The server broadcasts the aggregated personalized model ω_j^{t+1} to all participating clients, which serves as the initialization for the next local training round. This process repeats until either the maximum round limit T is reached or the model converges

Algorithm 1: FedCANA

Input: ω^0 : Initial global model parameters; η : Local learning rate; $M_{\max}, M_{\min}$: Max/Min neighbourhood value; τ : Entropy threshold; ρ : Client joining ratio; T : Total training iterations; q : Sparsity ratio; $\gamma_{\min}, \gamma_{\text{pivot}}, \gamma_{\max}$: scaling factors.

(Continued)

Algorithm 1 (continued)

Output: Personalized model parameters $\{\omega_1^T, \omega_2^T, \dots, \omega_N^T\}$.

```

1: Client  $i, \forall i \in [N]$  initializes its local model  $\omega_i^0$ .
2: for  $t = 0$  to  $T - 1$  do
3:   Server samples a client subset  $S_t \subseteq [N]$  based on  $\rho$ .
4:   Server sends the current model parameters to all clients  $i \in S_t$ .
5:   Client  $i \in S_t$  in parallel:
6:     Initialize local model with received parameters.
7:     Train local model for  $E$  epochs:
8:       for  $e = 1$  to  $E$  do
9:          $\omega_i^{t,e} \leftarrow \omega_i^{t,e-1} - \eta \nabla f_i(\omega_i^{t,e-1}, D_i)$ 
10:      end for
11:       $\hat{\omega}_i^t \leftarrow \omega_i^{t,E}, \hat{\omega}_i^* \leftarrow \hat{\omega}_i^t - \eta \nabla f_i(\hat{\omega}_i^t, \mathcal{D}_i)$ 
12:       $\delta_i^t \leftarrow \text{vec}(|\hat{\omega}_i^* - \hat{\omega}_i^t|) \in \mathcal{R}^d, \tilde{\delta}_i^t \leftarrow \text{SALIENCYSELECT}(\delta_i^t, q)$ 
13:      Upload  $(\hat{\omega}_i^t, \tilde{\delta}_i^t)$  to server.
14:    end for
15:    for each client  $i \in S_t$  do
16:      for each client  $j \in S_t$  do
17:         $p_{i,j} \leftarrow \text{CosSim}(\tilde{\delta}_i^t, \tilde{\delta}_j^t) = \frac{\langle \tilde{\delta}_i^t, \tilde{\delta}_j^t \rangle}{\|\tilde{\delta}_i^t\| \|\tilde{\delta}_j^t\| + \varepsilon}$ 
18:      end for
19:    end for
20:     $\alpha \leftarrow \text{COLNORMALIZE}(p, \varepsilon)$ 
21:    Compute column entropy for each target client  $j \in S_t$ :
22:     $H_j \leftarrow - \sum_{i \in S_t} \alpha_{i,j} \log(\alpha_{i,j} + \varepsilon)$ .
23:     $\bar{H} \leftarrow \frac{1}{|S_t|} \sum_{j \in S_t} H_j, H_{\text{norm}} \leftarrow \bar{H} / \log(|S_t|)$ .
24:     $\gamma \leftarrow \begin{cases} \gamma_{\min} + (\gamma_{\text{pivot}} - \gamma_{\min}) \cdot \frac{H_{\text{norm}}}{\tau}, & H_{\text{norm}} \leq \tau, \\ \gamma_{\text{pivot}} + (\gamma_{\max} - \gamma_{\text{pivot}}) \cdot \frac{H_{\text{norm}} - \tau}{1 - \tau}, & H_{\text{norm}} > \tau. \end{cases}$ 
25:     $M_t \leftarrow \min(M_{\max}, \max(M_{\min}, \lfloor \gamma \cdot M_{\max} \rfloor))$ .
26:    for each client  $j \in S_t$  do
27:       $\text{mask} \leftarrow \text{argsort}(p_{:,j})[1 : |S_t| - M_t]$  ▷ indices of the smallest  $|S_t| - M_t$  similarities
28:       $p_{\text{mask},j} \leftarrow 0$ 
29:    end for
30:     $\alpha \leftarrow \text{COLNORMALIZE}(p, \varepsilon)$ 
31:    for each client  $j \in S_t$  do
32:       $\omega_j^{t+1} \leftarrow \sum_{i \in S_t} \alpha_{i,j} \cdot \hat{\omega}_i^t$ 
33:    end for
34:  end for
35: return  $\{\omega_1^T, \omega_2^T, \dots, \omega_N^T\}$ 

```

4.2 Residual-Spatial Attention Collaborative Module

In the local training phase, we introduce a lightweight module termed RSAC, which tightly integrates residual blocks with a spatial attention mechanism. By adaptively strengthening local representations and facilitating gradient flow, this design alleviates the negative impact of data heterogeneity across clients on federated model performance. Fig. 2 depicts the overall architecture of RSAC. Specifically, the module jointly incorporates residual dense connectivity with a spatial attention branch. The residual dense pathways establish high-speed routes for gradient backpropagation and enable multi-level feature reuse, thereby maintaining stable information flow and high feature utilization in deep networks. Since standard convolutions capture only local receptive fields and fail to model long-range dependencies, an attention branch is further adopted to learn global context. This attention component characterizes client-specific spatial saliency distributions and adaptively recalibrates feature maps, highlighting informative regions while attenuating noise.

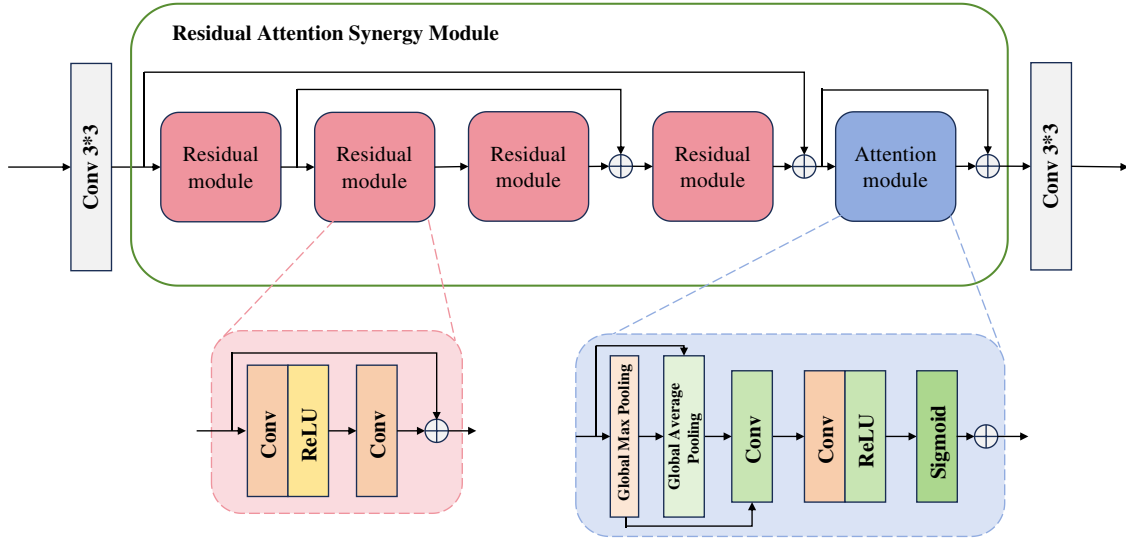


Figure 2: Schematic of the internal structure of RSAC.

Let x_i denote the feature produced by a convolutional or fully connected layer. The attention operation is defined as

$$A(\mathbf{x}_i) = \sigma(W_{\text{att}}\mathbf{x}_i + b_{\text{att}}) \quad (6)$$

here, W_{att} and b_{att} denote the trainable weights and bias vector of the attention module, respectively. The output feature vector $\hat{x}_i$ is expressed as follows:

$$\hat{x}_i = x_i + A(x_i) \quad (7)$$

By combining these two branches, the network retains the capacity to learn shared representations across clients while adaptively improving its discrimination of local patterns. This leads to stronger representational robustness and faster convergence in Non-IID federated learning scenarios, and the resulting high-quality updates further benefit global aggregation. It is worth mentioning that residual learning permits very deep architectures for feature refinement and noise reduction, but it also substantially increases model complexity.

4.3 Saliency-Guided Differential Descriptor

To represent client local adaptation characteristics with low communication cost in Non-IID scenarios, and to support subsequent similarity-aware aggregation, this paper proposes the Saliency-Guided Differential Descriptor (SGDD). Specifically, for each client i , SGDD constructs a differential descriptor δ_i by vectorizing the element-wise absolute difference between the current local model ω_i and the guidance model ω_i^* :

$$\delta_i = \text{Vec}(|\omega_i^* - \omega_i|) \in \mathbb{R}^d, \quad (8)$$

where d denotes the count of parameters in the model.

The descriptor δ_i reflects the distribution of adaptation magnitudes across parameter dimensions when the client adapts the current local model to its local data distribution. It should be noted that this paper focuses on the relative importance of different parameter components during local adaptation, rather than the sign or direction of their updates. Since δ_i usually exhibits a highly imbalanced magnitude distribution, only a small subset of parameter components may dominate the local adaptation process. Therefore, this paper further introduces a saliency selection mechanism. We use the magnitude of parameter updates as the saliency measure. Under the sparsity ratio $q \in (0, 1]$, the $k = q \cdot d$ components with the largest magnitudes are retained, and the corresponding index set is obtained as

$$I_i = \text{Select}_k(\delta_i) \quad (9)$$

Based on this index set, we construct the sparse differential descriptor $\tilde{\delta}_i$, which is uploaded in the form of a sparse tensor. Its nonzero positions are specified by I_i , and their corresponding values are retained:

$$\tilde{\delta}_{i,j} = \begin{cases} \delta_{i,j}, & j \in I_i, \\ 0, & \text{otherwise,} \end{cases} \quad j = 1, \dots, d. \quad (10)$$

Finally, each client only uploads its local model ω_i and sparse differential descriptor $\tilde{\delta}_i$ to the server. Here, ω_i is used for subsequent personalized aggregation, while $\tilde{\delta}_i$ serves as a compact representation of the client's local adaptation features and is used for similarity calculation on the server. Through this construction, SGDD compresses the original high-dimensional and dense local adaptation information into a sparse differential representation constrained by saliency. It controls the additional communication overhead while preserving the most discriminative variation components during local adaptation. It also suppresses the interference of low-magnitude noisy variations in similarity estimation, thereby providing a more robust basis for the identification of similar clients and subsequent adaptive neighborhood aggregation on the server.

4.4 Entropy-Driven Adaptive Neighborhood Aggregation

In personalized federated learning, the server needs to select collaborative neighbors for each target client based on the similarity relationships among clients, and then perform personalized aggregation. If the neighborhood is too small, some potentially useful collaborative clients may be excluded, which limits the exploitable cross-client information. If the neighborhood is too large, less relevant clients may be introduced, which may degrade the effectiveness of personalized modeling. This problem becomes more serious in scenarios with dynamic client participation, data distribution drift, and continuous training evolution. In such cases, a fixed neighborhood size can hardly adapt to the evolving similarity structure in a stable manner.

Motivated by this observation, this paper proposes Entropy-driven Adaptive Neighborhood Aggregation (EANA). The basic idea of EANA is to use the entropy of the normalized similarity distribution to measure the concentration of client collaborative relationships in the current round, and then dynamically adjust the neighborhood size accordingly. When the similarity weight distribution is relatively flat, the target client does not show a strong preference for a few specific neighbors. This indicates that the current collaborative structure has low concentration. In this case, appropriately expanding the neighborhood helps exploit potential collaborative information. When the similarity weight distribution is relatively sharp, the target client's primary collaborators are concentrated within a small subset of clients. Under these conditions, shrinking the neighborhood helps suppress the interference from irrelevant clients.

Specifically, the server first computes the similarity matrix p_{ij} based on the uploaded sparse differential descriptors. It then performs column-wise normalization to obtain the weight matrix $\alpha_{i,j}$, where each column satisfies $\sum_{i \in S_t} \alpha_{i,j} = 1$. For each target client j , the corresponding column vector of $\alpha_{i,j}$ describes the relative contribution of participating clients to its personalized aggregation in the current round. To measure the concentration of this distribution, the column entropy of client j is defined as:

$$H_j = - \sum_{i \in S_t} \alpha_{i,j} \log(\alpha_{i,j} + \varepsilon), \quad \varepsilon = 10^{-10} \quad (11)$$

where ε is a numerical stability term. Since the column entropy of a single client may fluctuate significantly due to local noise, we further average and normalize the column entropies over all participating clients:

$$\bar{H} = \frac{1}{|S_t|} \sum_{j \in S_t} H_j, \quad H_{\text{norm}} = \frac{\bar{H}}{\log |S_t|} \quad (12)$$

Based on the above formulation, this paper introduces a dynamic adjustment factor γ . Using the entropy threshold τ as the boundary, we adopt a piecewise linear function to map the normalized entropy to the neighborhood scaling factor:

$$\gamma = \begin{cases} \gamma_{\min} + (\gamma_{\text{pivot}} - \gamma_{\min}) \times \frac{H_{\text{norm}}}{\tau}, & H_{\text{norm}} \leq \tau \\ \gamma_{\text{pivot}} + (\gamma_{\max} - \gamma_{\text{pivot}}) \times \frac{H_{\text{norm}} - \tau}{1.0 - \tau}, & H_{\text{norm}} > \tau \end{cases} \quad (13)$$

where τ denotes the entropy threshold, with a default value of 0.8.

Finally, the shared neighborhood size for the current round is defined as:

$$M_t = \min(M_{\max}, \max(M_{\min}, \lfloor \gamma \cdot M_{\max} \rfloor)) \quad (14)$$

In this way, M_t can be adaptively adjusted between $M_{\min}$ and $M_{\max}$ according to changes in the similarity structure of the current round. This design reduces the instability caused by client-specific local fluctuations. It also helps control the complexity of the aggregation process while maintaining personalization capability.

5 Experiments

5.1 Experimental Setup

5.1.1 Datasets

The experiments are mainly conducted on three standard benchmark datasets: (1) CIFAR-10. This dataset contains 60,000 images divided into 10 mutually exclusive classes, each containing 6000 images. It

is split in a standard manner, yielding 50,000 images for training and 10,000 for testing. (2) CIFAR-100. Similar to CIFAR-10, CIFAR-100 also contains 60,000 32×32 images, but the number of classes is expanded to 100, with each class containing only 600 images (500 for training and 100 for testing). (3) CINIC-10. This dataset bridges the complexity gap between CIFAR-10 and ImageNet. It consists of all images from CIFAR-10 combined with 210,000 images selected from ImageNet and downsampled, yielding a total of 270,000 32×32 RGB images, covering the same 10 classes as CIFAR-10.

5.1.2 Experimental Heterogeneity Settings

We employ an identical data partitioning approach to FedDWA [5]. We simulated heterogeneous environments under three common settings.

Pathological Heterogeneous Setting. For every client, we randomly distribute a class subset of the same cardinality [1,34]. Specifically, for CIFAR-10, CINIC-10, and CIFAR-100, 2, 2, and 6 classes are respectively selected for every client. This setup effectively removes inter-client class overlap.

Practical Heterogeneous Setting 1. Clients possess datasets of equal size but with differing distributions. For each client, $s\%$ of the samples come from designated major classes, whereas the remaining $(100 - s)\%$ portion is obtained by uniform sampling over all classes [8,16]. Clients are organized into multiple groups, where members of the same group share identical major classes, thereby introducing latent cluster structures. By default, s is set to 80.

Practical Heterogeneous Setting 2. The data available at each client covers most classes, yet the class-wise quantities are highly unbalanced [9,15,35]. A symmetric Dirichlet distribution $\text{Dir}(\beta)$ is adopted to partition the dataset among N clients, where the allocation proportions for each class are sampled from this distribution. The default value of β is 0.07.

5.1.3 Evaluation Metrics

We follow the evaluation standard from the literature [5]: For single-model PFL approaches, the test accuracy of their optimal global model is reported, while other PFL methods present the average test accuracy across their best-performing personalized models.

5.1.4 Experimental Setup

Performance is assessed under two experimental settings: (1) 20 clients participating in every round, and (2) 100 clients participating at a 20% rate per communication round. Each client trains for $E = 1$ local epoch per communication round, where the maximum allowable count of global communication rounds is fixed at 100. All methods adopt mini-batch stochastic gradient descent as the local optimizer, where the batch size is 20 and the learning rate is $\eta = 0.01$. To ensure the fairness and reliability of the experiment, the experiments are repeated three times and the average values are calculated. The data of the experiment that cannot be replicated comes from the corresponding literature.

5.2 Overall Experimental Comparison

This section aims to compare FedCANA with fourteen representative FL baseline methods under three heterogeneous settings on three image classification datasets. Following the standard evaluation protocol, we report the average test accuracy. In these experiments, the sparsity ratio of FedCANA is fixed at $q = 0.2$.

Tables 1 and 2 illustrate the experimental results. FedCANA achieves the best results on all six tasks under the pathological heterogeneous setting and practical heterogeneous setting 1. The improvement is more pronounced on CIFAR-100. This indicates that when more classes are involved and client-side data distributions become more complex, FedCANA is more effective in alleviating the negative influence of non-IID data on the training of personalized models.

Table 1: Mean test accuracy (%) with 20 clients and 100% participation across three datasets under pathological heterogeneous setting and practical heterogeneous setting 1.

Settings	Pathological Heterogeneous Setting			Practical Heterogeneous Setting 1		
Methods	CIFAR-10	CIFAR-100	CINIC-10	CIFAR-10	CIFAR-100	CINIC-10
Local Training	92.35	80.08	92.74	72.12	39.82	64.40
FedAvg	59.97	34.71	46.07	71.57	44.67	58.97
FedProx	57.42	31.12	45.60	70.03	41.53	56.01
FedAvgM	59.82	34.00	48.81	71.49	44.56	58.77
FedAvg_FT	88.82	62.73	81.87	74.35	46.76	62.45
SFL	60.71	34.17	48.83	71.77	44.83	59.28
per-FedAvg	88.44	62.83	89.65	73.32	43.62	62.57
pFedMe	90.78	78.10	88.72	77.68	45.79	59.82
ClusterFL	83.41	51.30	80.48	71.42	44.90	59.44
FedRoD	87.95	64.76	83.75	77.27	46.76	52.64
FedAMP	91.74	78.61	88.05	72.30	41.11	65.43
FedFomo	91.95	78.89	91.59	75.69	47.06	68.93
L2C	91.76	78.62	<u>91.69</u>	76.67	48.30	67.52
FedDWA	<u>92.76</u>	<u>79.98</u>	90.33	<u>78.09</u>	<u>50.83</u>	<u>70.29</u>
Ours	93.16	82.25	91.89	79.03	51.13	71.36

Note: The bolded part indicates the best result, and the underlined part denotes the second-best result.

Table 2: Mean test accuracy (%) over three datasets under practical heterogeneous setting 2 with 100 clients and 20% participation.

Settings	Practical Heterogeneous Setting 2		
Methods	CIFAR-10	CIFAR-100	CINIC-10
FedAvg	48.44	22.80	26.68
FedProx	45.08	23.69	27.06
FedAvgM	48.09	21.70	27.36
FedAvg FT	89.19	45.06	83.32
SFL	52.40	22.96	36.82
per-FedAvg	89.86	49.81	89.69
pFedMe	90.54	51.70	89.41
ClusterFL	85.48	31.33	76.73
FedRoD	86.83	42.90	88.72

(Continued)

Table 2 (continued)

Settings	Practical Heterogeneous Setting 2		
Methods	CIFAR-10	CIFAR-100	CINIC-10
FedAMP	91.27	51.16	90.51
FedFomo	91.73	54.29	<u>91.32</u>
L2C	<u>91.77</u>	54.98	91.67
FedDWA	91.81	<u>55.26</u>	90.80
Ours	89.97	64.52	90.17

Note: The bolded part indicates the best result, and the underlined part denotes the second-best result.

In Practical Heterogeneous Setting 2, FedCANA improves the test accuracy on CIFAR-100 by 9.26% compared with FedDWA. This result indicates that the proposed method maintains good adaptability when the number of clients increases, the participation ratio decreases, and the class distribution becomes severely imbalanced. In some scenarios with extremely scarce and heterogeneous local data, the proposed method does not achieve the best performance. Further analysis of the per-class accuracy results in Fig. 3 shows that FedCANA maintains highly competitive accuracy on head classes with sufficient samples (e.g., Ship and Frog in CIFAR-10, Dog and Frog in CINIC-10), but exhibits a noticeable decline in accuracy on tail classes with scarce samples such as Cat. This phenomenon can be attributed to two factors. On the one hand, missing classes in local client data limit the ability of the RSAC to optimize the decision boundary for unseen classes. On the other hand, high data variance in few-shot scenarios tends to leave noise perturbations during sparse differential transmission, which further undermines the accuracy of similarity estimation and the performance of subsequent model aggregation. These factors jointly cause the local performance deficiency of the proposed method, and this issue will be a key optimization direction in our follow-up work.

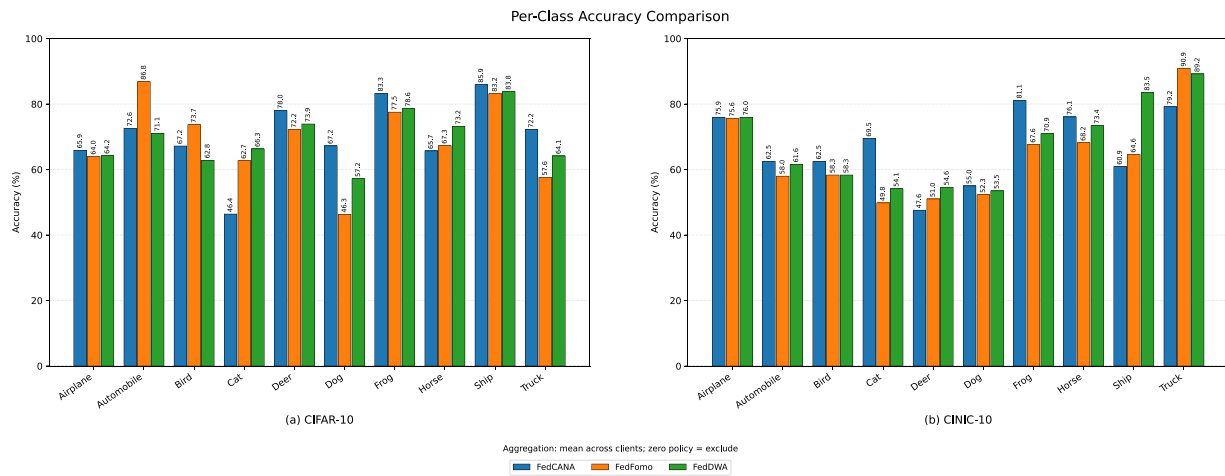


Figure 3: Per-class accuracy of CIFAR-10 and CINIC-10 under practical heterogeneous setting 2.

The advantages of FedCANA mainly come from two aspects. First, RSAC enhances local feature representation, alleviates the instability of local training caused by Non-IID data, and provides a more

effective model representation for subsequent client similarity estimation. Second, SGDD and EANA reduce redundant communication and improve the quality of personalized collaboration through compact differential representation and adaptive neighborhood aggregation. Overall, the experimental results verify the effectiveness of the proposed method in both pathological and practical heterogeneous scenarios. They show that FedCANA can achieve more stable personalized performance under most Non-IID conditions.

5.3 Communication Cost Analysis

Communication overhead metric: We measure the total communication overhead incurred by each participating client in each communication round, in megabytes (MB). This overhead consists mainly of three components: (1) the model parameters transmitted from the server to the client; (2) the model parameters uploaded from the client to the server; (3) the additional differential descriptor δ uploaded by the client.

Table 3 compares the per-client communication overhead of different methods in each communication round. In contrast, FedFomo, L2C, FedDWA, and our method require additional model interactions or auxiliary information to support personalized collaboration, which introduces extra communication costs. Among these methods, L2C incurs the highest overhead, reaching 100.17 and 98.00 MB, respectively. Our method advantage lies in its ability to explicitly estimate inter-client similarity by using sparse differential information and to perform personalized aggregation based on this similarity. Compared with FedFomo and L2C, which require much higher communication overhead for client collaboration, our method significantly reduces communication costs while preserving similarity modeling capability.

Table 3: Per-round communication overhead (MB) per client for the compared methods.

Methods	CIFAR-10	CIFAR-100/CINIC-10
FedFomo	28.62	28
L2C	100.17	98
FedDWA	14.31	14
Ours ($q = 0.2$)	12.59	12.13
Ours ($q = 0.1$)	11.14	10.73

Note: Bold values indicate the best performance in each column.

5.4 Computational Cost Analysis

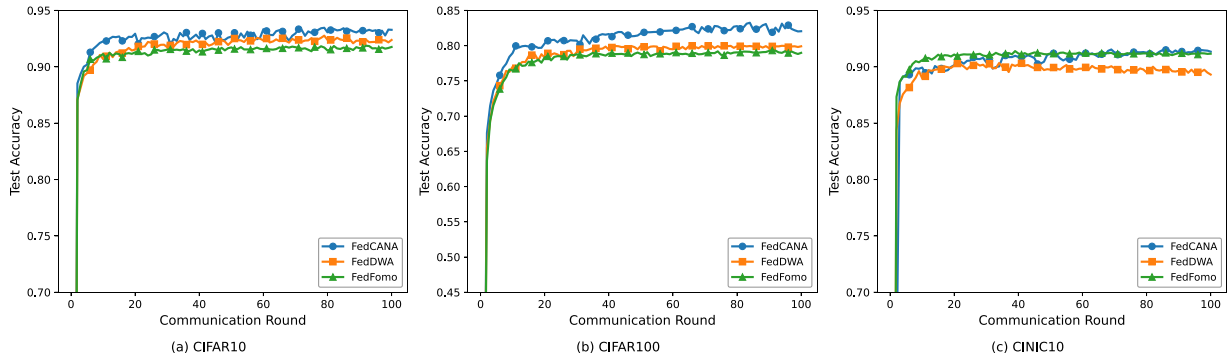
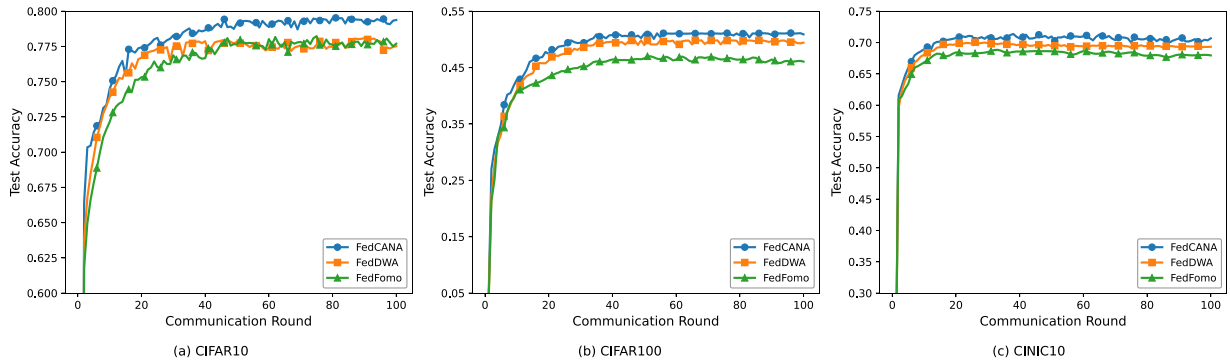
To evaluate the impact of modules such as RSAC on the client-side computational burden, we measured the number of floating-point operations (FLOPs) required for a single client to complete one communication round on CIFAR-10 in each method. As shown in Table 4, FedCANA requires 4.8×10^{11} FLOPs per client per round. Although this result is higher than those of the structurally simpler FedAvg and FedAMP, it is only about 4.35% higher than FedFomo, while being approximately 5.88% and 60% lower than FedDWA and L2C, respectively. This indicates that the residual feature extraction and spatial attention operations introduced by RSAC indeed incur additional computation, yet the overall computational cost of FedCANA remains within a reasonable range for existing personalized federated learning methods. Together with the reduced communication overhead shown in Table 3, FedCANA achieves a reasonable trade-off between client-side computational cost and communication efficiency.

Table 4: The computing costs generated by each client in each round for different methods.

Methods	FLOPs
FedAvg	2.5×10^{11}
FedFomo	4.6×10^{11}
FedAMP	2.5×10^{11}
L2C	1.2×10^{12}
FedDWA	5.1×10^{11}
FedCANA (Ours)	4.8×10^{11}

5.5 Training Curves of Test Accuracy

As shown in Figs. 4 and 5, FedCANA can rapidly converge to stable accuracy within a small number of communication rounds, with an overall convergence rate comparable to that of the baseline methods FedDWA and FedFomo. Particularly on the CIFAR-100, FedCANA maintains sustained improvement after reaching an accuracy level close to the baselines, demonstrating a higher performance ceiling.

**Figure 4:** Test accuracy trends across communication rounds under the pathological heterogeneous setting.**Figure 5:** Test accuracy trends across communication rounds under the practical heterogeneous setting 1.

5.6 Robustness Analysis Against Dynamic Client Dropout

To evaluate the robustness of FedCANA under dynamic system heterogeneity, this section conducts simulations of random client dropout based on Practical Heterogeneous Setting 1 on the CIFAR-100 dataset.

Four dropout ratios (10%, 20%, 30%, and 50%) are set under two dropout modes: the first is full-cycle persistent dropout, where random dropout is applied throughout all 100 communication rounds; the second is phased dropout, where dropout is only activated from round 50 to round 70, and clients participate normally in the remaining rounds.

As shown in Fig. 6a, as the dropout ratio increases, the overall training loss rises and the model convergence speed is moderately affected. Nevertheless, the test accuracy under all dropout ratios converges stably, with final performance close to the no-dropout baseline. As illustrated in Fig. 6b, the dropout phase causes a short-term rise in training loss, and the loss perturbation becomes more pronounced with a higher dropout ratio. However, the test accuracy only exhibits limited fluctuations without significant performance collapse. After clients resume normal participation from round 71, the training loss drops rapidly and the test accuracy quickly returns to its original convergence trajectory.

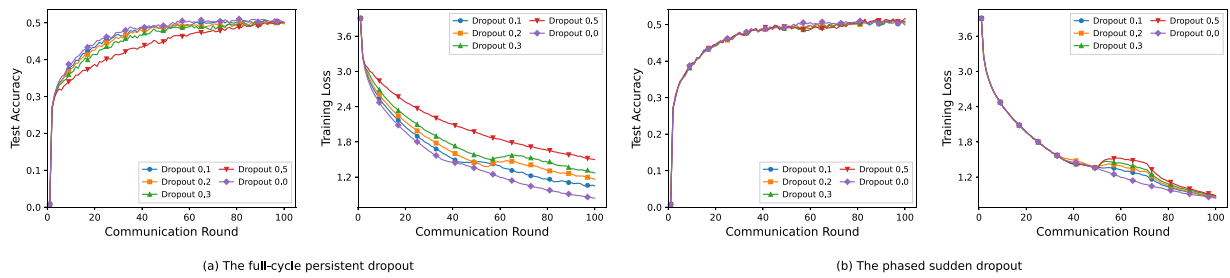


Figure 6: Training loss and test accuracy under random client dropout.

5.7 Sensitivity Analysis

5.7.1 Sensitivity Analysis of Sparsity Ratio q

This section analyzes the effect of the sparse upload ratio q on model performance. We compare different values of q under three heterogeneous data partitioning settings, and the corresponding test accuracy is shown in Fig. 7a–c. Overall, the proposed method is relatively insensitive to q . In most datasets and heterogeneous settings, the model still achieves performance close to, or even better than, that obtained with larger q values when only 10% or 20% of the key differential components are uploaded. It also consistently outperforms FedDWA in general. This indicates that only a subset of high-value information in the differential components plays a dominant role in client similarity estimation and personalized aggregation. Therefore, full upload is not necessary.

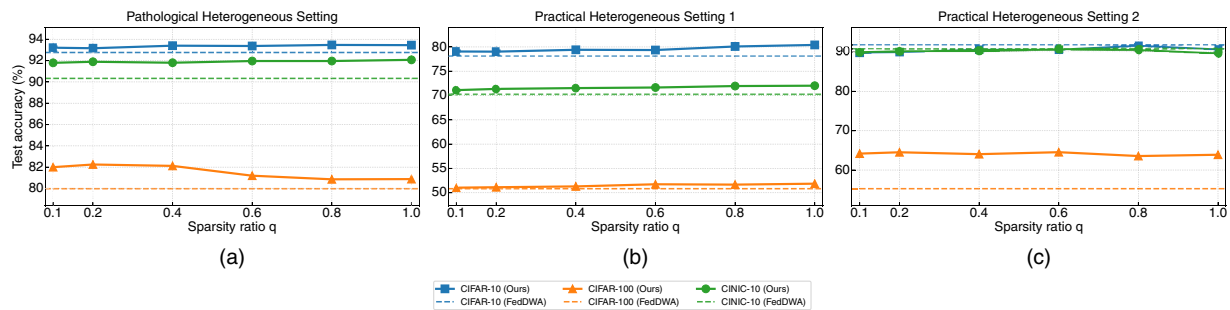


Figure 7: From left to right, (a–c) correspond to the pathological heterogeneous setting, practical heterogeneous setting 1, and practical heterogeneous setting 2, respectively. They show the relationship between test accuracy and the sparsity ratio q under each setting.

Combined with the communication overhead results in Table 3, a smaller q can significantly reduce the per-round communication overhead while maintaining stable test accuracy. Therefore, we set $q = 0.2$ as the default value to balance communication efficiency and personalization performance. This result further verifies the effectiveness of the proposed sparse differential upload mechanism. It shows that the performance improvement mainly comes from key collaborative information rather than complete parameter transmission.

5.7.2 Sensitivity Analysis of Entropy Threshold τ

As shown in Fig. 8, under the practical heterogeneous setting 1 and pathological heterogeneous setting, we adopt two metrics for evaluation: final personalized accuracy denotes the test accuracy obtained at the last training round, and optimal accuracy refers to the highest test accuracy observed throughout the whole training process. The final personalized accuracy and the optimal accuracy across the three datasets exhibit negligible variation with respect to τ . This convincingly demonstrates that the proposed method achieves robust results even when τ varies considerably, shows low sensitivity to hyperparameter optimization, and achieves superior generalization ability and strong robustness.

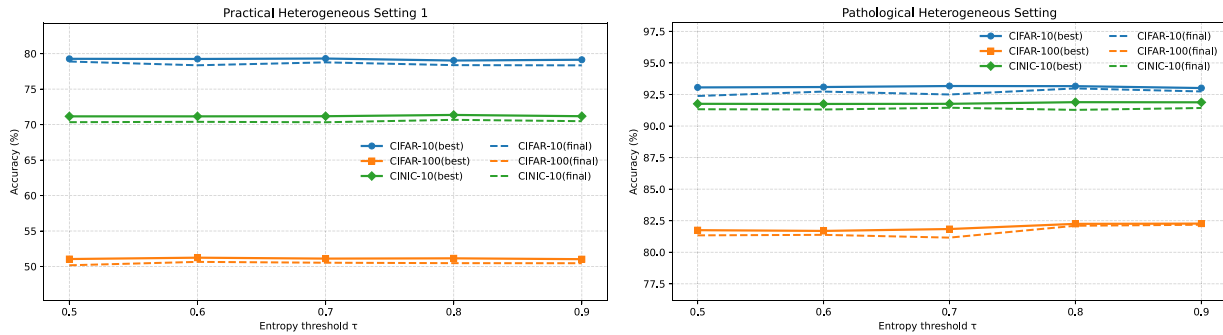


Figure 8: Performance at the entropy threshold τ .

5.8 Ablation Study

In this section, to validate the effectiveness of the key components of our proposed method, we conduct ablation experiments under the practical heterogeneous setting 1, which more closely resembles real-world application scenarios. Our analysis focuses on the contributions of the two core modules, RSAC and EANA, as well as the saliency selection mechanism in SGDD, to the performance of personalized models. It is worth noting that, for the saliency selection mechanism, we replace it with random selection to demonstrate that SGDD can effectively preserve critical differential descriptors. To ensure a fair comparison, all experimental configurations, including the number of training rounds, the client sampling ratio, the optimizer, and the data partitioning setting, are kept the same as those in the main experiments.

Table 5 shows that the complete FedCANA achieves the highest accuracy on all three datasets. Removing RSAC leads to varying degrees of accuracy degradation across the three datasets, indicating that RSAC can effectively extract key feature representations under heterogeneous scenarios, thereby improving the quality of personalized aggregation. Removing EANA results in the most significant accuracy drop on CIFAR-100 compared to the other two variants, suggesting that in more complex class spaces, EANA not only helps mitigate negative transfer from irrelevant clients but also enhances model robustness on challenging tasks. The random selection variant consistently underperforms the full model, confirming that selecting highly salient differential descriptors is more effective than randomly selecting the same number of descriptors. Overall, the complete FedCANA outperforms all ablation variants on the three datasets,

demonstrating that the performance gains of our method do not rely on a single module in isolation. Instead, it is the synergistic combination of local feature enhancement, adaptive neighborhood aggregation, and key differential descriptor preservation that collectively empowers FedCANA with superior personalized modeling capability under Non-IID data distributions.

Table 5: An ablation study on FedCANA.

Variant	Practical Heterogeneous Setting 1		
Methods	CIFAR-10	CIFAR-100	CINIC-10
FedDWA	78.09	50.83	70.29
w/o RSAC	78.55	50.94	70.75
w/o EANA	78.76	50.9	70.89
Random selection	78.67	50.95	71.08
Ours ($q = 0.2$)	79.03	51.13	71.36

Note: Bold values indicate the best performance in each column.

6 Conclusion

In this paper, we propose FedCANA, a communication-efficient personalized federated learning framework. It aims to address the difficulty of balancing personalization performance and communication overhead under Non-IID data distributions. Built upon the conventional federated optimization framework, FedCANA introduces RSAC to enhance local feature representations on clients, uses SGDD to select and upload key differential information to reduce redundant communication overhead, and employs EANA to dynamically adjust the collaborative neighborhood according to the client relationship structure. This design reduces the interference from less relevant clients during personalized aggregation. Extensive experimental results show that FedCANA outperforms existing methods on three standard datasets across several Non-IID conditions and achieves more stable performance in highly heterogeneous scenarios. Communication analysis and ablation studies further validate the effectiveness of each component.

Acknowledgement: None.

Funding Statement: This work was supported by the National Natural Science Foundation of China (U24A20241); the Major Scientific and Technological Special Project of Guizhou Province ([2024]014); the Provincial Department of Science and Technology's Hundred-level Innovation Talents Project (GCC [2023]018); the Guizhou Province Major Project (Qiankehe Major Project [2024]003); the Guizhou Provincial Innovation Talent Team for Cybersecurity Protection of Computing Power (CXTD [2025]029); and the Scientific and Technological Innovation Platform Research Project of Guizhou Province (CXPTXM [2025]024).

Author Contributions: The authors confirm contribution to the paper as follows: conceptualization, Yao Pu and Rou Zhou; methodology, Yao Pu and Rou Zhou; validation, Yao Pu; formal analysis, Yao Pu; investigation, Yao Pu and Rou Zhou; resources, Yao Pu and Rou Zhou; data curation, Yao Pu; writing—original draft preparation, Yao Pu; writing—review and editing, Yuling Chen and Rou Zhou; visualization, Yao Pu and Rou Zhou; supervision, Rou Zhou; project administration, Yao Pu; funding acquisition, Yuling Chen. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The data that support the findings of this study are openly available: CIFAR-10 and CIFAR-100 at <https://www.cs.toronto.edu/~kriz/cifar.html>, and CINIC-10 at <https://github.com/BayesWatch/cinic-10>.

Ethics Approval: Not applicable. This study used only publicly available benchmark datasets (CIFAR-10, CIFAR-100, and CINIC-10) and did not involve human or animal subjects.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. McMahan HB, Moore E, Ramage D, Hampson SCD, Arcas BAY. Communication-efficient learning of deep networks from decentralized data. In: International Conference on Artificial Intelligence and Statistics. Cambridge, MA, USA: PMLR; 2016. p. 1273–82.
2. Yang Q, Liu Y, Chen T, Tong Y. Federated machine learning: concept and applications. *ACM Trans Intell Syst Technol.* 2019;10(2):1–19. doi:10.1145/3298981.
3. Li T, Sahu AK, Talwalkar A, Smith V. Federated learning: challenges, methods, and future directions. *IEEE Signal Process Mag.* 2020;37(3):50–60. doi:10.1109/MSP.2020.2975749.
4. Zhang M, Sapra K, Fidler S, Yeung S, Álvarez JM. Personalized federated learning with first order model optimization. In: Proceedings of the 2021 International Conference on Learning Representations; 2021 May 4; Vienna, Austria.
5. Liu JH, Wu J, Chen JY, Hu M, Zhou YP, Wu D. FedDWA: personalized federated learning with dynamic weight adjustment. In: Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence; 2023 Aug 19–25; Macau, China. p. 3993–4001. doi:10.24963/ijcai.2023/444.
6. Li ST, Zhou TY, Tian XM, Tao DC. Learning to collaborate in decentralized learning of personalized models. In: Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022 Jun 18–24; New Orleans, LA, USA. p. 9766–75. doi:10.1109/CVPR52688.2022.00954.
7. Sattler F, Müller KR, Samek W. Clustered federated learning: model-agnostic distributed multi-task optimization under privacy constraints. *IEEE Trans Neural Netw Learn Syst.* 2021;32(8):3710–22. doi:10.1109/TNNLS.2020.3015958.
8. Huang YT, Chu LY, Zhou ZR, Wang LJ, Liu JC, Pei J, et al. Personalized cross-silo federated learning on Non-IID data. *Proc AAAI Conf Artif Intell.* 2021;35(9):7865–73. doi:10.1609/aaai.v35i9.16960.
9. Li T, Hu SY, Beirami A, Smith V. Ditto: fair and robust federated learning through personalization. In: Proceedings of the 38th International Conference on Machine Learning; 2021 Jul 18–24; Virtual. p. 6357–68. doi:10.48550/arxiv.2012.04221.
10. Arivazhagan MG, Aggarwal V, Singh AK, Choudhary S. Federated learning with personalization layers. *arXiv:1912.00818.* 2019.
11. Sun YC, Kountouris M, Zhang J. How to collaborate: towards maximizing the generalization performance in cross-silo federated learning. *IEEE Trans Mob Comput.* 2025;24(4):3211–22. doi:10.1109/TMC.2024.3509852.
12. Fan L, Zeng ZH, Zhang XN, Duan MJ, Sood K. Joint client selection and epoch configuration for heterogeneity-aware federated learning in resource-constrained systems. *Future Gener Comput Syst.* 2026;177(5):108249. doi:10.1016/j.future.2025.108249.
13. Li T, Sahu AK, Zaheer M, Sanjabi M, Talwalkar A, Smith V. Federated optimization in heterogeneous networks. *Proc Mach Learn Sys.* 2020;2:429–50. doi:10.48550/arxiv.1812.06127.
14. Dinh CT, Tran N, Nguyen J. Personalized federated learning with moreau envelopes. *Adv Neural Inf Process Syst.* 2020;33:21394–21405. doi:10.48550/arxiv.2006.08848.
15. Hsu TMH, Qi H, Brown M. Measuring the effects of non-identical data distribution for federated visual classification. *arXiv:1909.06335.* 2019.
16. Karimireddy SP, Kale S, Mohri M, Reddi SJ, Stich SU, Suresh AT. SCAFFOLD: stochastic controlled averaging for federated learning. In: Proceedings of the 37th International Conference on Machine Learning; 2020 Jul 13–18; Virtual. p. 5132–43.
17. de Souza LAC, Camilo GF, Rebello GAF, Sammarco M, Campista MEM, Costa LHMK. ATHENA-FL: avoiding statistical heterogeneity with one-versus-all in federated learning. *J Internet Serv Appl.* 2024;15(1):273–88. doi:10.5753/jisa.2024.3826.

18. Li XY, Qu Z, Tang B, Lu Z. FedLGA: toward system-heterogeneity of federated learning via local gradient approximation. *IEEE Trans Cybern.* 2024;54(1):401–14. doi:10.1109/TCYB.2023.3247365.
19. Yang M, Li D, Wang X, Li F, Fan L, Wang C, et al. FedeCouple: fine-grained balancing of global-generalization and local-adaptability in federated learning. *IEEE Trans Mob Comput.* 2025;25(4):5855–71. doi:10.1109/TMC.2025.3634199.
20. Zhu X, Sun Z, Rida I, Cambria E, Su J, Wang R, et al. FedUAF: uncertainty-aware fusion with reliability-guided aggregation for multimodal federated sentiment analysis. *arXiv:2603.13291.* 2026.
21. Zhu X, Xiaosong E, Nappi M, Rida I, Chen H. FedBayesMamba: uncertainty-aware federated learning for multimodal and audio-visual sequential modeling with selective state space models. *Pattern Recognit.* 2026;180(5):114240. doi:10.1016/j.patcog.2026.114240.
22. Azy T, Yu H, Cui LZ, Yang Q. Towards personalized federated learning. *IEEE Trans Neural Netw Learn Syst.* 2023;34(12):9587–9603. doi:10.1109/TNNLS.2022.3160699.
23. Yuan HL, Paul W, Aucott J, Burlina P, Cao YZ. PFedEdit: personalized federated learning via automated model editing. In: *Computer Vision—ECCV 2024. Lecture Notes in Computer Science, Vol 15137.* Cham, Switzerland: Springer; 2024. p. 91–107. doi:10.1007/978-3-031-72986-7_6.
24. Yin KT, Mao JY. Personalized federated learning with adaptive feature aggregation and knowledge transfer. In: *Proceedings of the 2025 International Joint Conference on Neural Networks (IJCNN); 2025 Jun 30–Jul 5; Rome, Italy.* p. 1–8. doi:10.1109/ijcnn64981.2025.11228793.
25. Wang K, Mathews R, Kiddon C, Eichner H, Beaufays F, Ramage D. Federated evaluation of on-device personalization; 2019. *arXiv:1910.10252.* 2019.
26. Fallah A, Mokhtari A, Ozdaglar A. Personalized federated learning with theoretical guarantees: a model-agnostic meta-learning approach. *Adv Neural Inf Process Syst.* 2020;33:3557–68.
27. Aladwani T, Anagnostopoulos C, Parambath SP, Deligianni F. CL-FML: cluster-based & label-aware federated meta-learning for on-demand classification tasks. In: *Proceedings of the 2024 IEEE 11th International Conference on Data Science and Advanced Analytics (DSAA); 2024 Oct 6–10; San Diego, CA, USA.* p. 1–10. doi:10.1109/DSAA61799.2024.10722828.
28. Chen HY, Chao WL. On bridging generic and personalized federated learning for image classification. *arXiv:2107.00778.* 2021.
29. Chen FW, Long GD, Wu ZH, Zhou TY, Jiang J. Personalized federated learning with a graph. In: *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI-22); 2022 Jul 23–29; Vienna, Austria.* p. 2575–82. doi:10.24963/ijcai.2022/357.
30. Lin Y, Han S, Mao H, Wang Y, Dally WJ. Deep gradient compression: reducing the communication bandwidth for distributed training. *arXiv:1712.01887.* 2017.
31. Sattler F, Wiedemann S, Müller KR, Samek W. Robust and communication-efficient federated learning from Non-IID data. *IEEE Trans Neural Netw Learn Syst.* 2019;31(9):3400–13. doi:10.1109/TNNLS.2019.2944481.
32. Li MM, He X, Chen JH. Federated collaborative learning with sparse gradients for heterogeneous data on resource-constrained devices. *Entropy.* 2024;26(12):1099. doi:10.3390/e26121099.
33. Wang YS, Wu RL, Li XJ, Kutscher D. PacTrain: pruning and adaptive sparse gradient compression for efficient collective communication in distributed deep learning. In: *Proceedings of the 2025 62nd ACM/IEEE Design Automation Conference (DAC); 2025 Jun 22–25; San Francisco, CA, USA.* p. 1–7. doi:10.1109/DAC63849.2025.11133419.
34. Shamsian A, Navon A, Fetaya E, Chechik G. Personalized federated learning using hypernetworks. In: *Proceedings of the 38th International Conference on Machine Learning; 2021 Jul 18–24; Virtual.* p. 9489–502.
35. Li Q, He B, Song D. Model-contrastive federated learning. In: *Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2021 Jun 20–25; Nashville, TN, USA.* p. 10708–17. doi:10.1109/CVPR46437.2021.01057.