Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

FedEAGD: Trajectory-Guided Drift-Aware Distillation for Personalized Federated Learning

Mengjia Lu
Henan University of Science and Technology
Ping Zhang
zping@haust.edu.cn
Henan University of Science and Technology
An Bao
Henan University of Science and Technology
Keyi Xu
Henan University of Science and Technology
Jiechang Wang
Henan Sport University

Research Article

Keywords:

Posted Date: November 10th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11417646/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

FedEAGD: Trajectory-Guided Drift-Aware Distillation for Personalized Federated Learning

Mengjia Lu[1], Ping Zhang[1*], An Bao[1], Keyi Xu[1], Jiechang Wang[2]

[1]School of Mathematics and Statistics, Henan University of Science and Technology, 263 Kaiyuan Ave., Luoyang, 471023, China.
[2]Information Management Center, Henan Sport University, 2 Yinhe St., Zhengzhou, 450044, China.

*Corresponding author(s). E-mail(s): zping@haust.edu.cn;

Abstract

Personalized federated learning (PFL) aims to balance shared knowledge with client-specific adaptation under heterogeneous data. However, data that are not independently and identically distributed (non-IID), together with partial client participation, can cause shared representations to vary across communication rounds, hindering knowledge transfer. Existing methods mainly address heterogeneity through model personalization or historical knowledge transfer, while sample-level temporal representation changes remain underexplored. We propose FedEAGD, a trajectory-guided PFL framework that uses a client-specific exponential moving average (EMA) teacher to summarize shared-model history. The discrepancy between each sample's current and historical shared representations quantifies temporal drift and adaptively determines the strength of historical distillation. Progressive gated residual fusion (PGRF) introduces personalized information according to a stage-wise schedule. Representation anchoring and residual decoupling regularize this process, while post-aggregation calibration and late-stage server smoothing coordinate local adaptation with global model updates. Only the shared feature extractor is communicated, while personalized components remain local. Experiments on MNIST, CIFAR-10, CIFAR-100, and DomainNet evaluate FedEAGD under varying data heterogeneity, client participation rates, and federation sizes. On CIFAR-100 with a Dirichlet concentration parameter of $\boldsymbol{\alpha = 0.1}$, FedEAGD achieves 63.05% accuracy, outperforming the strongest evaluated baseline by 5.96 percentage points.

Keywords: Personalized federated learning, Temporal representation drift, Knowledge distillation, Non-IID data, Adaptive distillation, Progressive personalization

1 Introduction

Federated learning (FL) enables multiple clients to collaboratively train a model without directly sharing their local data, making it suitable for distributed and privacy-sensitive environments (McMahan et al. 2017; Li et al. 2020a; Kairouz et al. 2021). In real-world federated systems, however, client data often violate the independent and identically distributed (IID) assumption due to differences in data sources, acquisition conditions, class distributions, and feature statistics. Such data are commonly referred to as non-IID data (Li et al. 2021c). Client participation may also vary across communication rounds. These sources of heterogeneity can lead to divergent local optimization trajectories and weaken the transferability of a single global model across clients (Li et al. 2020b; Wen et al. 2023). In federated systems, model updates are exchanged between a central server and participating clients. Consequently, client participation, the number of transmitted parameters, and federation scale are important system-level considerations alongside predictive accuracy. These considerations motivate personalized learning methods that accommodate heterogeneous client data while limiting the exchange of client-specific model components.

Personalized federated learning (PFL) addresses this limitation by integrating shared knowledge with client-specific adaptation. Recent surveys have reviewed the challenges posed by heterogeneous federated data and emerging directions for federated learning, providing broader context for adaptive and personalized approaches (Wen et al. 2023; Ji et al. 2024). Existing approaches have explored various strategies, including personalized regularization, model decomposition, representation learning, adaptive aggregation, and selective collaboration (Dinh et al. 2020; Li et al. 2021b; Collins et al. 2021; Zhang et al. 2023c,b; Xu et al. 2025; Zhong et al. 2026; Liu et al. 2025b, 2026). Community-aware collaboration and personalized alignment have also been investigated to improve coordination between global and local models (Yao et al. 2026). Although these methods effectively account for inter-client differences, they primarily characterize personalization in terms of data distributions, model components, or relationships among clients. By contrast, the evolution of individual sample representations across communication rounds has received comparatively little attention.

Historical information can provide additional supervision to stabilize local optimization. Recent studies have incorporated historical models, adaptive memory, and teacher-guided knowledge transfer into personalized federated learning (Fu et al. 2026; Liu et al. 2025a; Zhang et al. 2026c,b). Cross-client distillation has also been investigated in personalized cloud–edge collaboration (Zhang et al. 2025). However, historical supervision is typically constructed or applied at the model or client level. As a result, samples within the same client may be subject to similar historical constraints despite exhibiting different representation trajectories across communication rounds.

To address this issue, we propose FedEAGD, a trajectory-guided personalized federated learning framework that captures sample-level representation evolution across communication rounds. Each client maintains an exponential moving average (EMA) historical teacher to capture the trajectory of its locally retained shared model. The discrepancy between current and historical shared representations serves as a sample-level temporal drift signal that adaptively adjusts the strength of historical distillation

for each sample. Accordingly, samples with larger representation changes receive stronger temporal supervision, whereas those with relatively stable representations are subject to weaker constraints.

To balance temporal stabilization and client-specific adaptation, FedEAGD introduces progressive gated residual fusion. This mechanism progressively increases the contribution of personalized information to the shared representation through a gated residual connection as communication rounds advance. Representation anchoring and residual decoupling provide auxiliary regularization, while post-aggregation calibration and late-stage server smoothing facilitate local–global coordination. During communication, only the shared feature extractor is exchanged, while all personalized components remain local to each client. Together, these mechanisms adapt historical supervision to sample-level representation drift and introduce client-specific information according to a stage-wise fusion schedule.

The main contributions of this work are summarized as follows:

- We introduce sample-level temporal supervision into personalized federated learning by measuring changes in each sample's shared representation across communication rounds. The resulting temporal drift signal determines the strength of historical distillation, allowing the supervision to adapt to sample-specific representation changes rather than applying a uniform distillation weight.
- We develop progressive gated residual fusion (PGRF), which forms a fused representation by adding a gated personalized residual to the shared representation. A stage-wise coefficient keeps the residual contribution at zero initially and increases its scaling during a predefined ramp-up period.
- We evaluate FedEAGD on MNIST, CIFAR-10, CIFAR-100, and DomainNet under varying data heterogeneity, client participation rates, and federation sizes. FedEAGD achieves the highest mean accuracy across the evaluated settings, outperforming the strongest baseline by 5.96 percentage points on CIFAR-100 with $\alpha = 0.1$. We also analyze convergence, assess individual components through ablation studies, and quantify computational and communication costs. Only the shared feature extractor is communicated, avoiding additional model-parameter transmission for personalized components.

2 Related Work

2.1 Personalized Federated Learning

Personalized federated learning aims to adapt models to individual clients while leveraging transferable knowledge from federated collaboration under heterogeneous data distributions (Tan et al. 2023). Existing approaches address this objective through personalized regularization, model decomposition, representation learning, and adaptive collaboration (Ji et al. 2024). Regularization-based methods, such as pFedMe and Ditto, constrain the relationship between personalized and shared models (Dinh et al. 2020; Li et al. 2021b), whereas FedPer, FedRep, and FedBABU achieve personalization by separating globally shared model components from locally adapted ones (Arivazhagan et al. 2019; Collins et al. 2021; Oh et al. 2022). Other approaches,

including FedRoD and GPFL, combine generic and personalized representations or prediction components to accommodate client heterogeneity (Chen and Chao 2022; Zhang et al. 2023a).

Recent studies have increasingly focused on the adaptive interaction between shared and personalized knowledge. FedCP and FedALA dynamically adjust the use of global and local information, whereas FedCAC and FedSC enhance collaboration through selective coordination among clients (Zhang et al. 2023c,b; Wu et al. 2023; Xu et al. 2025). FedCNM combines global feature constraints with adaptive aggregation, while other recent methods investigate hybrid expert architectures, shared-base adaptation, and dynamically generated personalized models (Zhong et al. 2026; Zhang et al. 2026a; Liu et al. 2025b, 2026). Community-aware aggregation and local alignment have also been explored to improve coordination between global and local models under heterogeneous client data distributions (Yao et al. 2026).

These methods primarily characterize personalization in terms of client data distributions, model components, or inter-client relationships. By contrast, FedEAGD focuses on how individual sample representations evolve across communication rounds and uses this temporal information to adaptively regulate historical distillation before progressively incorporating client-specific information.

2.2 Federated Knowledge Distillation and Historical Supervision

Knowledge distillation is widely used in federated learning to transfer supervisory information without directly exchanging raw data (Hinton et al. 2015). Federated Distillation and FedMD transfer prediction-level knowledge among distributed models, whereas FedGen and FedKD extend this approach through generative knowledge transfer and compact teacher–student architectures, respectively (Jeong et al. 2018; Li and Wang 2019; Zhu et al. 2021; Wu et al. 2022). Knowledge distillation has also been explored in personalized cloud–edge collaboration to facilitate knowledge transfer across clients with heterogeneous data (Zhang et al. 2025). Related approaches have explored generative knowledge distillation for federated semi-supervised learning and data-free mutual distillation for knowledge transfer under heterogeneous client data (Karami et al. 2024; Wang et al. 2026).

Beyond prediction-level knowledge transfer, representation-based methods seek to improve consistency between local and shared representations through feature alignment or contrastive learning. MOON uses model-contrastive learning to regularize local representations, while PFAKD and FedRIR align shared and personalized representations at the feature level (Li et al. 2021a; Qi et al. 2024; Huang et al. 2025).

Recent studies have incorporated historical information into personalized federated optimization. FedLFH uses a tracking mechanism to preserve client-specific historical information and coordinate shared and personalized features (Fu et al. 2026). Per-FedAHM adaptively aggregates historical classifier heads based on their relevance, as assessed by validation loss (Liu et al. 2025a). FedPKD regulates teacher–student knowledge transfer according to local empirical risk, whereas FedTDC combines teacher-confidence-guided distillation with aggregation calibration (Zhang et al. 2026c,b).

Prior PFL methods have addressed client heterogeneity through personalized modeling, model decomposition, and collaboration among clients. Research on historical knowledge transfer has also explored model tracking and adaptive distillation. However, how changes in individual samples' shared representations across communication rounds can guide the strength of historical distillation remains insufficiently explored, particularly when historical supervision must be balanced with progressive personalization. FedEAGD addresses this gap by using a client-specific EMA teacher to measure sample-level temporal drift, adjusting the distillation weight accordingly, and combining this supervision with progressive gated residual fusion.

3 Methodology

3.1 Problem Formulation and Notation

Consider a personalized federated learning system with a central server and K clients, denoted by $\mathcal{C} = \{C_1, C_2, \ldots, C_K\}$. Client C_k holds a private dataset

$$\mathcal{D}_k = \{(x_i, y_i)\}_{i=1}^{n_k}, \tag{1}$$

where $n_k = |\mathcal{D}_k|$ denotes the number of local samples.

Client data follow client-specific distributions $P_k(x, y)$. Under non-IID settings,

$$P_k(x, y) \neq P_j(x, y), \qquad k \neq j, \tag{2}$$

which may lead to different local objectives and optimization directions across clients (Li et al. 2020b; Tan et al. 2023).

The model of client C_k consists of a shared feature extractor with parameters θ_g, a personalized feature extractor with parameters θ_k^p, a gating module with parameters ω_k, and a local classifier with parameters ϕ_k. Its local empirical objective is

$$F_k(\theta_g, \theta_k^p, \omega_k, \phi_k) = \frac{1}{n_k} \sum_{(x_i, y_i) \in \mathcal{D}_k} \ell\left(f_k(x_i; \theta_g, \theta_k^p, \omega_k, \phi_k), y_i\right), \tag{3}$$

where $f_k(\cdot)$ denotes the complete client prediction function. Let $N = \sum_{k=1}^{K} n_k$. The PFL objective is

$$\min_{\theta_g, \{\theta_k^p, \omega_k, \phi_k\}_{k=1}^{K}} \sum_{k=1}^{K} \frac{n_k}{N} F_k(\theta_g, \theta_k^p, \omega_k, \phi_k). \tag{4}$$

At communication round t, the server selects a client subset $S_t \subseteq \mathcal{C}$. Only θ_g participates in server aggregation, whereas θ_k^p, ω_k, and ϕ_k remain local.

3.2 Framework Overview

FedEAGD integrates trajectory-guided temporal stabilization with progressive personalized representation learning. The former leverages each client's shared-model

trajectory across communication rounds to construct sample-level historical supervision, while the latter gradually incorporates client-specific information following an initial stage of shared-representation learning.

As shown in Fig. 1, each client is equipped with a shared feature extractor $G(\cdot;\theta_g)$, a personalized feature extractor $P_k(\cdot;\theta_k^p)$, a gating module $A_k(\cdot;\omega_k)$, and a local classifier $H_k(\cdot;\phi_k)$. Each client also maintains a historical teacher $\bar{G}_k(\cdot;\bar{\theta}_{g,k})$ to track the evolution of its locally retained shared model. Only the shared feature extractor is exchanged between the server and clients; the personalized feature extractor, gating module, local classifier, and historical teacher are maintained locally at each client.

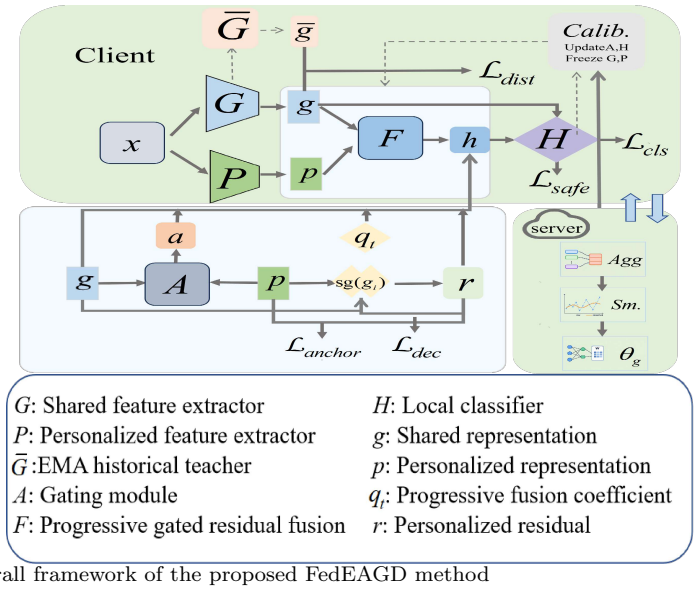


Fig. 1 Overall framework of the proposed FedEAGD method

During local training, the historical teacher provides a temporal reference for drift-aware historical distillation, while progressive gated residual fusion regulates the incorporation of personalized information into shared representations. After local training, the server aggregates the shared feature extractors and synchronizes the resulting model with participating clients, which then perform lightweight local calibration. The overall procedure is summarized in Algorithm 1.

3.3 Client Local Update

At communication round t, each selected client $C_k \in S_t$ loads the latest synchronized parameters into its shared feature extractor, while retaining the client-specific states of its personalized feature extractor, gating module, local classifier, and historical teacher.

3.3.1 Client-Trajectory-Guided Drift-Aware Distillation

Historical supervision can stabilize local optimization, but relying on a reference from a single communication round may make the supervision sensitive to transient local fluctuations. Moreover, samples within the same client may undergo different degrees of representation change across communication rounds. To address these issues, FedEAGD summarizes each client's shared-model trajectory and uses it to provide sample-level temporal supervision.

For a mini-batch $\mathcal{B}$ on client C_k,

$$\mathcal{B} = \{(x_i, y_i)\}_{i=1}^{B}, \tag{5}$$

where B denotes the mini-batch size.

Let $\tilde{\theta}_{g,k}^{t-1}$ denote the locally retained shared-extractor parameters immediately before synchronization in round t. The sequence

$$\mathcal{T}_k^{t-1} = \left\{ \tilde{\theta}_{g,k}^{1}, \tilde{\theta}_{g,k}^{2}, \ldots, \tilde{\theta}_{g,k}^{t-1} \right\}, \qquad t \geq 2, \tag{6}$$

defines the client-specific shared-model trajectory.

Rather than storing the complete trajectory, FedEAGD recursively summarizes it through an exponential moving average (EMA) historical teacher (Tarvainen and Valpola 2017):

$$\bar{\theta}_{g,k}^{t} = \gamma \bar{\theta}_{g,k}^{t-1} + (1-\gamma)\tilde{\theta}_{g,k}^{t-1}, \qquad t \geq 2, \tag{7}$$

where $\gamma \in (0,1)$ is the EMA coefficient. The teacher is initialized from the shared extractor and remains frozen during local gradient optimization.

For each sample x_i, the current and historical shared representations are computed as

$$g_i = G(x_i; \theta_g), \qquad \bar{g}_i = \bar{G}_k\left(x_i; \bar{\theta}_{g,k}^{t}\right). \tag{8}$$

Their temporal consistency is measured using cosine similarity:

$$s_i = \frac{g_i^{\top} \bar{g}_i}{\|g_i\|_2 \|\bar{g}_i\|_2 + \epsilon}, \tag{9}$$

where $\epsilon > 0$ is a small constant for numerical stability. The sample-level temporal drift is then defined as

$$d_i = 1 - s_i. \tag{10}$$

A larger d_i indicates stronger cross-round representation variation. Accordingly, the historical distillation coefficient is defined for each sample as

$$\mu_i = \min\left\{\mu_0(1+d_i), \mu_{\max}\right\}, \tag{11}$$

where μ_0 is the base coefficient and $\mu_{\max}$ bounds the strength of historical supervision. This rule links the observed representation drift directly to the strength of historical supervision, while the upper bound limits the maximum distillation weight assigned to any sample.

The resulting representation distillation loss is

$$\mathcal{L}_{\text{dist}} = \frac{1}{B} \sum_{i=1}^{B} \mu_i \left\| g_i - \text{sg}(\bar{g}_i) \right\|_2^2 . \tag{12}$$

The stop-gradient operation keeps the historical representation fixed as the temporal supervision target during local optimization.

3.3.2 Progressive Gated Residual Fusion

Temporal stabilization alone does not provide client-specific adaptation, while applying the personalized residual from the outset may interfere with early shared-representation learning. FedEAGD therefore uses progressive gated residual fusion (PGRF), which combines the shared representation with a gated personalized residual scaled by a time-dependent coefficient.

For each sample x_i, the shared representation, personalized representation, and gating vector are

$$g_i = G(x_i; \theta_g), \qquad p_i = P_k(x_i; \theta_k^p), \qquad a_i = \sigma\left(A_k([g_i; p_i]; \omega_k)\right), \tag{13}$$

where $[g_i; p_i]$ denotes feature concatenation and $a_i \in [0,1]^d$ controls the contribution of the personalized residual.

The fused representation is

$$h_i = g_i + q_t \eta_{\max} a_i \odot \left[p_i - \text{sg}(g_i)\right], \tag{14}$$

where $\eta_{\max}$ is the maximum fusion ratio and $\odot$ denotes element-wise multiplication. The stop-gradient operation isolates the shared feature used as the residual reference; the shared branch still participates in h_i through the direct additive path and the gating input.

The progressive coefficient is defined as

$$q_t = \begin{cases} 0, & t < t_s, \\ \dfrac{t - t_s}{T_r}, & t_s \le t < t_s + T_r, \\ 1, & t \ge t_s + T_r, \end{cases} \tag{15}$$

where t_s is the activation round and T_r is the ramp-up duration. Thus, q_t remains zero before t_s, increases linearly from zero to one during the ramp-up period, and remains at one thereafter. This schedule controls the scaling of the gated residual; its actual contribution to the fused representation also depends on the sample-specific gate a_i and residual $p_i - \text{sg}(g_i)$.

The shared and fused representations are classified using the same local classifier:

$$z_i^s = H_k(g_i; \phi_k), \qquad z_i^f = H_k(h_i; \phi_k). \tag{16}$$

Their classification objective is defined using the cross-entropy loss, denoted by CE:

$$\mathcal{L}_{\text{cls}} = \frac{1}{2B} \sum_{i=1}^{B} \left[\text{CE}(z_i^s, y_i) + \text{CE}(z_i^f, y_i) \right]. \tag{17}$$

To limit degradation caused by unreliable personalized fusion, the shared path is used as a reference:

$$\mathcal{L}_{\text{safe}} = \max \left(0, \mathcal{L}_{\text{CE}}^f - \text{sg} \left(\mathcal{L}_{\text{CE}}^s \right) \right), \tag{18}$$

where

$$\mathcal{L}_{\text{CE}}^s = \frac{1}{B} \sum_{i=1}^{B} \text{CE}(z_i^s, y_i), \qquad \mathcal{L}_{\text{CE}}^f = \frac{1}{B} \sum_{i=1}^{B} \text{CE}(z_i^f, y_i). \tag{19}$$

The safety term is active only when the fusion-path loss exceeds the shared-path loss. Because the shared-path loss is detached, gradients from this term are propagated through $\mathcal{L}_{\text{CE}}^f$ only.

3.3.3 Auxiliary Regularization for Personalized Representation Learning

Two auxiliary regularization terms are introduced during progressive fusion. Representation anchoring helps preserve shared semantic information in the early stages, while residual decoupling encourages personalized features to capture complementary information.

The anchoring loss is

$$\mathcal{L}_{\text{anchor}} = \frac{1}{B} \sum_{i=1}^{B} \| p_i - \text{sg}(g_i) \|_2^2 . \tag{20}$$

Its contribution is gradually reduced as personalized fusion becomes active, allowing the personalized branch greater flexibility in later rounds.

The personalized residual is

$$r_i = p_i - \text{sg}(g_i). \tag{21}$$

The shared representation and residual are normalized as

$$\hat{g}_i = \frac{\text{sg}(g_i)}{\| \text{sg}(g_i) \|_2}, \tag{22}$$

and

$$\hat{r}_i = \frac{r_i}{\| r_i \|_2}. \tag{23}$$

Residual decoupling is then defined by

$$\mathcal{L}_{\text{dec}} = \frac{1}{B} \sum_{i=1}^{B} \left(\hat{g}_i^\top \hat{r}_i \right)^2 . \tag{24}$$

Since both vectors are ℓ_2-normalized, minimizing $\mathcal{L}_{\text{dec}}$ discourages the personalized residual from encoding directions already represented by the shared branch.

3.3.4 Local Optimization Objective

The anchoring coefficient is scheduled according to the progressive fusion stage:

$$\lambda_a^t = \lambda_a^{\text{start}} - q_t \left(\lambda_a^{\text{start}} - \lambda_a^{\text{end}}\right). \tag{25}$$

The local objective of client C_k is

$$\mathcal{L}_k^t = \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{dist}} + \lambda_a^t \mathcal{L}_{\text{anchor}} + q_t \lambda_d \mathcal{L}_{\text{dec}} + q_t \lambda_s \mathcal{L}_{\text{safe}}, \tag{26}$$

where λ_d and λ_s control residual decoupling and fusion safety, respectively. The factors q_t progressively activate the personalization-related terms together with residual fusion.

During local training, θ_g, θ_k^p, ω_k, and ϕ_k are optimized jointly, while the historical teacher remains frozen.

3.4 Server Aggregation and Local–Global Coordination

After local optimization, participating clients upload only their shared feature extractors. The server performs sample-weighted aggregation:

$$\theta_{g,\text{agg}}^{t+1} = \sum_{j \in S_t} \frac{n_j}{\sum_{j \in S_t} n_j} \theta_{g,j}^{t,\text{loc}}, \tag{27}$$

Late-stage momentum smoothing is applied after round t_m:

$$\theta_g^{t+1} = \begin{cases} \theta_{g,\text{agg}}^{t+1}, & t < t_m, \\ \beta \theta_g^t + (1-\beta)\theta_{g,\text{agg}}^{t+1}, & t \geq t_m, \end{cases} \tag{28}$$

where β is the smoothing coefficient.

Before receiving the aggregated parameters θ_g^{t+1}, each client updates its historical teacher using the shared model retained from round t. For participating clients, this is the locally optimized shared feature extractor, whereas non-participating clients retain the previous shared model θ_g^t. The aggregated shared feature extractor is then synchronized across all clients. Under partial participation, only clients in S_t perform local training and upload their updates, while all clients are assumed to be available for server-to-client synchronization.

After synchronization, participating clients freeze both feature extractors and calibrate only the gating module and classifier using

$$\mathcal{L}_{\text{cal}} = \mathcal{L}_{\text{CE}}^s + q_t \lambda_f \mathcal{L}_{\text{CE}}^f + q_t \lambda_s \mathcal{L}_{\text{safe}}, \tag{29}$$

where λ_f controls the fused-path classification term. Calibration uses only a small number of local mini-batches and does not introduce an additional full local training stage. The complete procedure is summarized in Algorithm 1.

Algorithm 1 FedEAGD

Require: $C = \{C_1, \ldots, C_K\}$ - client set; T - communication rounds; ρ - participation rate; E - local epochs; B - mini-batch size; γ - EMA coefficient

Ensure: shared extractor θ_g and personalized models $\{\theta_k^p, \omega_k, \phi_k\}_{k=1}^K$

Initialize: $\theta_g, \{\theta_k^p, \omega_k, \phi_k\}_{k=1}^K$, and set $\bar{\theta}_{g,k}^1 \leftarrow \theta_g^1$ for each client C_k
for $t = 1, \ldots, T$ **do**
  $S_t \leftarrow \text{ClientSelect}(C, \rho)$
  for all $C_k \in S_t$ **in parallel do**
    $\theta_{g,k}^{t,\text{loc}} \leftarrow \theta_g^t$
    for each mini-batch $\mathcal{B}$ over E local epochs **do**
      $g_i \leftarrow G(x_i; \theta_{g,k}^{t,\text{loc}}), \quad p_i \leftarrow P_k(x_i; \theta_k^p)$
      $a_i \leftarrow \sigma(A_k([g_i; p_i]; \omega_k)), \quad r_i \leftarrow p_i - \text{sg}(g_i)$
      $h_i \leftarrow g_i + q_t \eta_{\max} a_i \odot r_i$
      $\bar{g}_i \leftarrow \bar{G}_k(x_i; \bar{\theta}_{g,k}^t), \quad s_i \leftarrow \dfrac{g_i^\top \bar{g}_i}{\|g_i\|_2 \|\bar{g}_i\|_2 + \epsilon}$
      $d_i \leftarrow 1 - s_i, \quad \mu_i \leftarrow \min\{\mu_0(1 + d_i), \mu_{\max}\}$
      $\mathcal{L}_k^t \leftarrow \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{dist}} + \lambda_a^t \mathcal{L}_{\text{anchor}} + q_t \lambda_d \mathcal{L}_{\text{dec}} + q_t \lambda_s \mathcal{L}_{\text{safe}}$
      $(\theta_{g,k}^{t,\text{loc}}, \theta_k^p, \omega_k, \phi_k) \leftarrow \text{Adam}(\nabla \mathcal{L}_k^t)$
    end for
    Upload $\theta_{g,k}^{t,\text{loc}}$
  end for
  $\theta_{g,\text{agg}}^{t+1} \leftarrow \sum_{k \in S_t} \dfrac{n_k}{\sum_{j \in S_t} n_j} \theta_{g,k}^{t,\text{loc}}$
  Obtain θ_g^{t+1} using server momentum smoothing
  for all $C_k \in C$ **do**
    $\widetilde{\theta}_{g,k}^t \leftarrow \begin{cases} \theta_{g,k}^{t,\text{loc}}, & C_k \in S_t, \\ \theta_g^t, & C_k \notin S_t \end{cases}$
    $\bar{\theta}_{g,k}^{t+1} \leftarrow \gamma \bar{\theta}_{g,k}^t + (1 - \gamma) \widetilde{\theta}_{g,k}^t$
    Synchronize $\theta_{g,k} \leftarrow \theta_g^{t+1}$
  end for
  for all $C_k \in S_t$ **do**
    Freeze $\theta_{g,k}$ and θ_k^p
    for each calibration mini-batch **do**
      $(\omega_k, \phi_k) \leftarrow \text{Adam}(\nabla \mathcal{L}_{\text{cal}})$
    end for
  end for
end for
return θ_g, $\{\theta_k^p, \omega_k, \phi_k\}_{k=1}^K$

3.5 Analytical Properties

We establish three properties of FedEAGD's adaptive mechanisms: (i) the historical distillation coefficient is bounded and non-decreasing with respect to sample-level temporal drift; (ii) the EMA teacher forms an exponentially weighted summary of the retained shared-model states; and (iii) the contribution of progressive gated residual fusion is bounded at each training stage. Together, these properties characterize how FedEAGD adapts historical supervision to sample-level representation drift and regulates the contribution of personalized information throughout training.

Lemma 1. *For $\mu_{\max} \geq \mu_0 > 0$, the drift-aware distillation coefficient is a non-decreasing, bounded, and μ_0-Lipschitz continuous function of the sample-level temporal drift.*

Proof. Using $d_i = 1 - s_i$, the adaptive distillation coefficient is

$$\mu_i = \min\left\{\mu_0(1 + d_i), \mu_{\max}\right\}. \tag{30}$$

Let $\widetilde{\mu}_i = \mu_0(1 + d_i)$ denote the coefficient before upper clipping. Since the upper-clipping operation is monotone and non-expansive, we have

$$\frac{\partial \widetilde{\mu}_i}{\partial d_i} = \mu_0 > 0, \qquad \mu_0 \leq \mu_i \leq \mu_{\max}, \qquad |\mu_i - \mu_j| \leq \mu_0 |d_i - d_j|. \tag{31}$$

Thus, the strength of historical supervision increases monotonically with temporal drift, up to the predefined upper bound, and remains Lipschitz continuous as a function of d_i.

Lemma 2. *The EMA historical teacher forms a normalized exponentially weighted summary of the client-specific shared-model trajectory.*

Proof. Recursively expanding the EMA update in Eq. (7) gives

$$\bar{\theta}_{g,k}^{t} = \gamma^{t-1}\bar{\theta}_{g,k}^{1} + (1-\gamma)\sum_{j=1}^{t-1}\gamma^{t-1-j}\widetilde{\theta}_{g,k}^{j}. \tag{32}$$

For $0 < \gamma < 1$, all temporal coefficients are non-negative, and their sum satisfies

$$\gamma^{t-1} + (1-\gamma)\sum_{j=1}^{t-1}\gamma^{t-1-j} = 1. \tag{33}$$

Thus, the EMA teacher can be expressed as a convex combination of the initial teacher and the locally retained shared-model states, with the weights assigned to older states decaying exponentially. This provides a recursive summary of each client's trajectory without requiring explicit storage of all historical models.

Proposition 1. *Assume that the personalized residual is bounded by $\|r_i\|_2 \leq R$, where $R > 0$. Then the perturbation introduced by progressive gated residual fusion remains bounded throughout training and follows a stage-wise upper bound.*

Proof. From Eq. (14), together with $0 \leq a_{i,m} \leq 1$, we obtain

$$\begin{aligned} h_i - g_i &= q_t \eta_{\max} a_i \odot r_i, \\ \|h_i - g_i\|_2 &\leq q_t \eta_{\max} \|r_i\|_2 \leq q_t \eta_{\max} R. \end{aligned} \tag{34}$$

Substituting the progressive coefficient in Eq. (15) yields

$$\|h_i - g_i\|_2 \leq \begin{cases} 0, & t < t_s, \\ \dfrac{t - t_s}{T_r} \eta_{\max} R, & t_s \leq t < t_s + T_r, \\ \eta_{\max} R, & t \geq t_s + T_r. \end{cases} \tag{35}$$

Thus, the personalized residual does not contribute to the fused representation before t_s. Its contribution increases according to an explicit stage-wise bound during the ramp-up period and remains bounded by $\eta_{\max} R$ once fully activated. This design limits abrupt changes in the fused representation as personalization is gradually introduced.

4 Experiments and Results

This section evaluates FedEAGD across diverse heterogeneous federated settings. We first describe the datasets, evaluation protocol, baseline methods, and implementation details. We then assess its overall performance and convergence behavior, examine its performance under varying levels of data heterogeneity, client participation rates, and federation scales, and conduct ablation studies to evaluate the contributions of its main components.

4.1 Experimental Setup

4.1.1 Datasets and Heterogeneity Settings

Experiments are conducted on MNIST (LeCun et al. 1998), CIFAR-10 and CIFAR-100 (Krizhevsky 2009), and DomainNet (Peng et al. 2019). Three heterogeneous settings are considered: Practical, Pathological, and Real-world.

In the Practical setting, client data are partitioned according to a Dirichlet distribution, with the concentration parameter α controlling the distribution of class labels across clients. Unless otherwise specified, we set $\alpha = 0.1$. In the Pathological setting, each client receives samples from only two classes on CIFAR-10 and ten classes on CIFAR-100. For the Real-world setting, we use a fixed subset of 10 classes from DomainNet, distributed across six clients associated with the clipart, infograph, painting, quickdraw, real, and sketch domains. Within each independent run, all methods use the same data partitions and class subsets. Fig. 2 shows representative client-wise data distributions for the Pathological setting and the Practical setting with different Dirichlet concentration parameters.

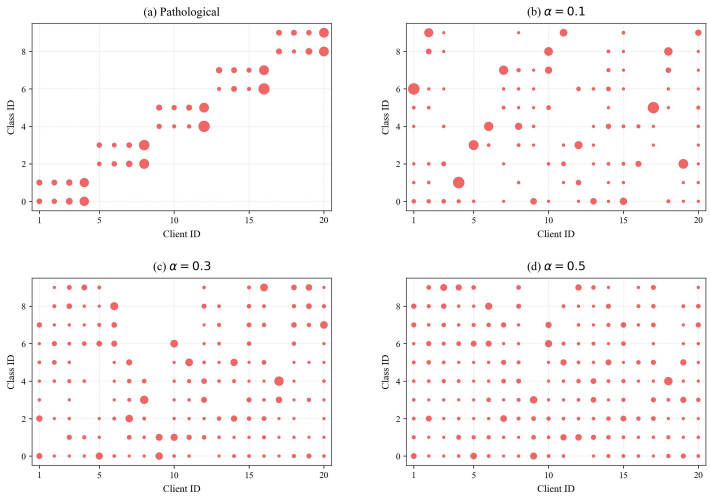


Fig. 2 Client-wise data distributions under different heterogeneous settings

4.1.2 Evaluation Protocol and Baselines

For each client, 75% of the local samples are used for training and the remaining 25% for testing. Unless otherwise specified, all experiments except DomainNet use 20 clients, 100 communication rounds, and full client participation with $\rho = 1.0$. DomainNet contains six clients corresponding to its six visual domains.

Performance is evaluated using sample-weighted test accuracy across clients. For each run, test accuracy is averaged over the final 10 communication rounds to mitigate fluctuations across rounds. Each experiment is repeated three times with different random seeds, and results are reported as the mean $\pm$ standard deviation. All methods use identical client partitions and experimental settings within each run.

FedEAGD is compared with FedAvg (McMahan et al. 2017), FedRep (Collins et al. 2021), FedCAC (Wu et al. 2023), FedGH (Yi et al. 2023), FedCP (Zhang et al. 2023c), FedALA (Zhang et al. 2023b), FedKD (Wu et al. 2022), FedDistill (Jeong et al. 2018), FedAS (Yang et al. 2024), and FedHG (Hu et al. 2025). These methods cover global aggregation, personalized representation learning, adaptive collaboration, and knowledge-distillation-based federated learning.

4.2 Implementation Details

Each FedEAGD client comprises a shared feature extractor, a personalized feature extractor, a gating module, and a local classifier. The two feature extractors follow the same backbone architecture, consisting of two convolutional blocks with 5×5 kernels and 32 and 64 channels, respectively. Each block is followed by GroupNorm, ReLU, max pooling, and a squeeze-and-excitation (SE) module. The resulting features are mapped to a 512-dimensional representation through a fully connected layer followed by LayerNorm and ReLU. The gating module takes the concatenation of the shared

and personalized representations as input and produces a 512-dimensional gating vector through a linear layer followed by a sigmoid activation. Finally, a linear classifier maps the resulting representation to the label space.

Unless otherwise specified, training is conducted with 20 clients over 100 communication rounds, with one local epoch per round and a batch size of 32. The Adam optimizer is used with a learning rate of 5×10^{-5} and a weight decay of 1×10^{-4}. The EMA coefficient γ is set to 0.10, and the initial and maximum historical distillation coefficients, μ_0 and $\mu_{\max}$, are set to 0.50 and 1.25, respectively. Progressive gated residual fusion starts at communication round 50, with a 20-round ramp-up period and $\eta_{\max} = 0.10$. The initial and final anchoring coefficients, λ_a^{start} and λ_a^{end}, are set to 0.10 and 0.05, respectively. The remaining coefficients are set to $\lambda_d = 5 \times 10^{-5}$, $\lambda_s = 0.50$, and $\lambda_f = 0.50$. Post-aggregation calibration is performed using two local mini-batches, while server smoothing starts at communication round 60 with $\beta = 0.10$.

All experiments are conducted on an NVIDIA GeForce RTX 4090 GPU with 24 GB of memory. The computation time reported in Table 1 is the average wall-clock time per communication round on CIFAR-100 with $\alpha = 0.1$, 20 clients, and $\rho = 1.0$. This measurement includes local optimization, server aggregation, model synchronization, and post-aggregation calibration.

During communication, only the shared feature extractor is exchanged between the server and clients. The personalized feature extractor, historical teacher, gating module, and local classifier are retained on each client and are not transmitted. On CIFAR-100, the shared feature extractor contains 2,153,216 parameters, corresponding to approximately 8.61 MB per transmission when represented in 32-bit floating-point format.

For each participating client, uploading and downloading the shared feature extractor entails approximately 17.23 MB of bidirectional parameter transfer per communication round. This estimate excludes communication protocol overhead and should not be interpreted as a measurement of network latency or total system-wide traffic. Consequently, trajectory-guided historical supervision and the personalized components incur no additional model parameter communication beyond that required for the shared feature extractor.

4.3 Performance Comparison and Analysis

4.3.1 Overall Performance Comparison

Table 1 compares FedEAGD with the baseline methods in terms of test accuracy under the Practical, Pathological, and Real-world settings, as well as computation time and communication cost. Here, Σ denotes the total number of parameters in the complete model. The symbols Σ_b, Σ_h, Σ_{cs}, and Σ_g denote the numbers of communicated parameters in the backbone, head, client-specific component, and FedEAGD shared feature extractor, respectively. The remaining symbols follow the communication-cost definitions of the corresponding baseline methods. Communication cost is reported as the total number of parameters transmitted in both directions per participating client per communication round.

Table 1 Overall performance, computation time, and communication cost under different heterogeneous settings

Method	Practical ($\alpha = 0.1$)			Pathological	Real-world	Computation	Communication
	MNIST	CIFAR-10	CIFAR-100	CIFAR-100	DomainNet	time (s/round)	parameters/client/round
FedAvg	95.72 ± 0.03	57.67 ± 0.08	28.56 ± 0.13	22.18 ± 0.05	53.21 ± 0.13	12.25	2Σ
FedRep	99.33 ± 0.06	89.12 ± 0.05	49.54 ± 0.10	62.46 ± 0.05	59.46 ± 0.14	15.29	$2\alpha_f\Sigma$
FedCAC	99.19 ± 0.07	89.35 ± 0.04	48.10 ± 0.22	62.14 ± 0.06	55.31 ± 0.14	53.30	$2\gamma_c(1-\kappa+\psi\kappa)\Sigma$
FedGH	99.20 ± 0.06	89.15 ± 0.45	47.55 ± 0.21	62.94 ± 0.05	54.95 ± 0.16	18.46	$2\alpha_h\Sigma$
FedCP	99.36 ± 0.08	90.72 ± 0.26	50.80 ± 0.06	66.52 ± 0.06	59.68 ± 0.12	19.98	$\Sigma_b + 2(\Sigma_h + \Sigma_{cs})$
FedALA	99.43 ± 0.04	90.55 ± 0.12	53.46 ± 0.03	62.75 ± 0.03	54.45 ± 0.11	14.76	2Σ
FedKD	99.40 ± 0.02	89.90 ± 0.11	49.80 ± 0.08	68.36 ± 0.12	57.69 ± 0.12	49.56	$2\gamma_c\Sigma$
FedDistill	99.36 ± 0.05	89.56 ± 0.48	47.67 ± 0.15	64.17 ± 0.11	56.71 ± 0.08	30.65	2Σ
FedAS	99.42 ± 0.04	90.73 ± 0.07	57.09 ± 0.09	67.24 ± 0.04	60.82 ± 0.12	35.67	$2\gamma_c\alpha_f^{\text{shared}}\Sigma$
FedHG	99.42 ± 0.06	90.27 ± 0.14	55.40 ± 0.11	66.56 ± 0.22	57.59 ± 0.05	22.06	2Σ
FedEAGD	99.48 ± 0.05	92.27 ± 0.06	63.05 ± 0.10	73.22 ± 0.06	61.22 ± 0.05	**16.84**	$2\Sigma_g$

FedEAGD achieves the highest mean accuracy across all five evaluated settings. Under the Practical setting with $\alpha = 0.1$, it achieves accuracies of 92.27% on CIFAR-10 and 63.05% on CIFAR-100, outperforming the strongest baselines by 1.54 and 5.96 percentage points, respectively. On CIFAR-100 under the Pathological setting, FedEAGD achieves 73.22% accuracy, exceeding FedKD by 4.86 percentage points. On DomainNet, it achieves 61.22%, outperforming FedAS by a narrower margin of 0.40 percentage points.

Performance differences are small on MNIST, where most personalized methods achieve over 99% accuracy, but become more pronounced on CIFAR-10 and CIFAR-100. This suggests that trajectory-guided historical supervision is more beneficial when representation learning is more sensitive to client heterogeneity. FedEAGD requires 16.84 s per communication round, placing its computational cost close to that of several lightweight personalized methods and substantially below that of FedCAC, FedKD, FedDistill, and FedAS under the evaluated configuration.

4.3.2 Convergence Performance

To evaluate the convergence of FedEAGD and the baseline methods, Fig. 3 shows the training accuracy curves over the first 50 communication rounds on CIFAR-10 and CIFAR-100.

As shown in Fig. 3, the methods exhibit distinct convergence patterns and varying degrees of fluctuation across communication rounds. FedEAGD achieves relatively high accuracy early in training on both datasets and maintains competitive performance throughout the observed rounds. Its advantage is more pronounced on CIFAR-100, where several baseline methods converge to lower accuracy levels. These observations are consistent with the average test accuracies over the final 10 communication rounds reported in Table 1.

4.3.3 Performance under Different Levels of Data Heterogeneity

To evaluate sensitivity to different Dirichlet partitions, we vary $\alpha \in \{0.1, 0.3, 0.5\}$ on CIFAR-10 and CIFAR-100 while fixing $\rho = 1.0$. The results are reported in Table 2.

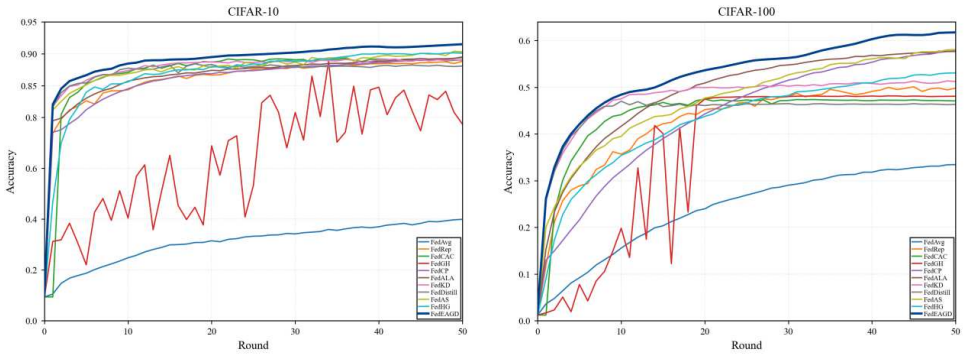


Fig. 3 Early-stage training accuracy on CIFAR-10 and CIFAR-100

Performance varies across Dirichlet settings, and the effect of changing α differs across datasets and methods. Nevertheless, FedEAGD achieves the highest mean accuracy across all six combinations of dataset and α. On CIFAR-100, its mean accuracy is 63.05%, 51.43%, and 45.74% at $\alpha = 0.1$, 0.3, and 0.5, respectively. Although its accuracy decreases as α increases on CIFAR-100, FedEAGD consistently outperforms the baselines at all three settings, indicating that its advantage extends across different levels of data heterogeneity.

Table 2 Test accuracy under different Dirichlet settings (%).

Method	**CIFAR-10** ($\rho = 1.0$)			**CIFAR-100** ($\rho = 1.0$)		
	$\alpha = 0.1$	$\alpha = 0.3$	$\alpha = 0.5$	$\alpha = 0.1$	$\alpha = 0.3$	$\alpha = 0.5$
FedAvg	57.67 ± 0.08	65.36 ± 0.26	68.25 ± 0.12	28.56 ± 0.13	26.12 ± 0.06	27.66 ± 0.04
FedRep	89.12 ± 0.05	81.39 ± 0.18	76.17 ± 0.24	49.54 ± 0.10	25.46 ± 0.09	21.48 ± 0.08
FedCAC	89.35 ± 0.04	77.45 ± 0.13	69.35 ± 0.12	48.10 ± 0.22	38.21 ± 0.13	33.36 ± 0.11
FedGH	89.15 ± 0.45	77.40 ± 0.32	69.43 ± 0.12	47.55 ± 0.21	28.20 ± 0.04	24.27 ± 0.12
FedCP	90.72 ± 0.26	82.64 ± 0.28	77.76 ± 0.20	50.80 ± 0.06	29.95 ± 0.12	25.74 ± 0.26
FedALA	90.55 ± 0.12	83.22 ± 0.11	79.28 ± 0.07	53.46 ± 0.03	36.78 ± 0.22	33.01 ± 0.35
FedKD	89.90 ± 0.11	78.90 ± 0.29	71.48 ± 0.13	49.80 ± 0.08	28.81 ± 0.26	23.45 ± 0.05
FedDistill	89.56 ± 0.48	78.11 ± 0.05	70.35 ± 0.11	47.67 ± 0.15	26.30 ± 0.31	22.55 ± 0.13
FedAS	90.73 ± 0.07	81.85 ± 0.11	71.12 ± 0.04	57.09 ± 0.09	43.13 ± 0.11	36.12 ± 0.08
FedHG	90.27 ± 0.14	80.25 ± 0.07	73.24 ± 0.05	55.40 ± 0.11	41.18 ± 0.14	33.34 ± 0.09
FedEAGD	$\mathbf{92.27 \pm 0.06}$	$\mathbf{84.78 \pm 0.12}$	$\mathbf{79.78 \pm 0.15}$	$\mathbf{63.05 \pm 0.10}$	$\mathbf{51.43 \pm 0.05}$	$\mathbf{45.74 \pm 0.04}$

4.3.4 Performance under Different Client Participation Rates

To evaluate partial participation, we fix $\alpha = 0.5$ and vary $\rho \in \{1.0, 0.5, 0.2\}$ on CIFAR-10 and CIFAR-100. The results are reported in Table 3.

Table 3 Test accuracy under different client participation rates (%).

Methods	CIFAR-10 ($\alpha = 0.5$)			CIFAR-100 ($\alpha = 0.5$)		
	$\rho = 1.0$	$\rho = 0.5$	$\rho = 0.2$	$\rho = 1.0$	$\rho = 0.5$	$\rho = 0.2$
FedAvg	68.25 ± 0.12	68.95 ± 0.18	68.41 ± 0.11	27.66 ± 0.04	27.36 ± 0.03	26.20 ± 0.31
FedRep	76.17 ± 0.24	76.30 ± 0.24	76.90 ± 0.23	21.48 ± 0.08	21.17 ± 0.08	21.45 ± 0.26
FedCAC	69.35 ± 0.12	69.35 ± 0.38	68.51 ± 0.13	33.36 ± 0.11	32.52 ± 0.11	33.34 ± 0.11
FedGH	69.43 ± 0.12	69.29 ± 0.28	68.22 ± 0.09	24.27 ± 0.12	23.81 ± 0.23	23.00 ± 0.25
FedCP	77.76 ± 0.20	77.36 ± 0.14	76.02 ± 0.03	25.74 ± 0.26	26.63 ± 0.19	26.57 ± 0.34
FedALA	79.28 ± 0.07	78.80 ± 0.29	76.90 ± 0.16	33.01 ± 0.35	32.44 ± 0.13	24.68 ± 0.33
FedKD	71.48 ± 0.13	71.45 ± 0.27	68.57 ± 0.11	23.45 ± 0.05	23.53 ± 0.16	20.86 ± 0.23
FedDistill	70.35 ± 0.11	70.26 ± 0.43	68.41 ± 0.05	22.55 ± 0.13	22.38 ± 0.16	20.99 ± 0.27
FedAS	71.12 ± 0.04	70.25 ± 0.12	70.14 ± 0.25	36.12 ± 0.08	35.35 ± 0.21	33.22 ± 0.24
FedHG	73.24 ± 0.05	72.44 ± 0.45	73.68 ± 0.52	33.34 ± 0.09	32.99 ± 0.48	33.65 ± 0.04
FedEAGD	$\mathbf{79.78 \pm 0.15}$	$\mathbf{78.85 \pm 0.20}$	$\mathbf{76.95 \pm 0.12}$	$\mathbf{45.74 \pm 0.04}$	$\mathbf{42.47 \pm 0.11}$	$\mathbf{35.50 \pm 0.08}$

Reducing client participation affects CIFAR-100 more severely than CIFAR-10. For FedEAGD, lowering the participation rate from 1.0 to 0.2 reduces test accuracy from 79.78% to 76.95% on CIFAR-10 and from 45.74% to 35.50% on CIFAR-100. Despite the larger decline on CIFAR-100, FedEAGD achieves the highest mean accuracy at every participation rate considered. This indicates that FedEAGD retains its advantage over the baselines even when fewer clients participate in each communication round.

4.3.5 Scalability with Different Numbers of Clients

To evaluate performance under different federation scales, the number of clients is varied as $K \in \{10, 20, 50, 100\}$ on CIFAR-100 with $\alpha = 0.1$, while the remaining settings are kept unchanged.

Table 4 Test accuracy with different numbers of clients (%).

Methods	$K = 10$	$K = 20$	$K = 50$	$K = 100$
FedAvg	30.77 ± 0.03	28.56 ± 0.13	31.16 ± 0.06	30.32 ± 0.05
FedRep	48.07 ± 0.04	49.54 ± 0.10	41.89 ± 0.08	35.70 ± 0.08
FedCAC	47.56 ± 0.04	48.10 ± 0.22	43.42 ± 0.07	41.12 ± 0.09
FedGH	46.95 ± 0.02	47.55 ± 0.21	45.35 ± 0.09	40.61 ± 0.11
FedCP	54.87 ± 0.04	50.80 ± 0.06	46.30 ± 0.08	40.87 ± 0.10
FedALA	51.22 ± 0.03	53.46 ± 0.03	46.38 ± 0.08	44.58 ± 0.08
FedKD	52.56 ± 0.02	49.80 ± 0.08	47.60 ± 0.04	44.30 ± 0.06
FedDistill	49.57 ± 0.04	47.67 ± 0.15	46.28 ± 0.07	43.68 ± 0.06
FedAS	57.32 ± 0.03	57.09 ± 0.09	37.25 ± 0.07	33.15 ± 0.10
FedHG	53.28 ± 0.03	55.40 ± 0.11	52.11 ± 0.14	44.36 ± 0.07
FedEAGD	$\mathbf{60.07 \pm 0.04}$	$\mathbf{63.05 \pm 0.10}$	$\mathbf{59.22 \pm 0.03}$	$\mathbf{50.73 \pm 0.08}$

FedEAGD achieves the highest mean accuracy at each of the four client counts. Its accuracy rises from 60.07% with 10 clients to 63.05% with 20 clients, before falling

to 59.22% and 50.73% with 50 and 100 clients, respectively. Several baseline methods exhibit a similar decline as the number of clients increases, consistent with greater fragmentation of local data. Even with 100 clients, FedEAGD achieves 50.73% accuracy, outperforming the strongest baseline by 6.15 percentage points. Furthermore, its personalized components are retained locally, so increasing the number of clients does not introduce additional communication of client-specific parameters.

4.4 Ablation Study

To assess the contributions of individual components, we conduct ablation studies on CIFAR-100 under the Pathological setting and CIFAR-10 under the Practical setting. The ablated variants remove the historical teacher, replace adaptive distillation weighting with a fixed coefficient, disable progressive gated residual fusion, omit post-aggregation calibration, or omit late-stage server smoothing. The results are presented in Table 5.

HT, DW, PGRF, Calib., and SS denote the historical teacher, distillation weighting, progressive gated residual fusion, post-aggregation calibration, and server smoothing, respectively. In the Fixed DW variant, the sample-adaptive distillation coefficient is replaced with a fixed value of 0.50.

Table 5 Ablation results of FedEAGD (%).

Settings	Pathological	Practical
	CIFAR-100	CIFAR-10
FedEAGD	**73.22 ± 0.06**	**92.27 ± 0.06**
w/o HT	71.96 ± 0.09	91.18 ± 0.10
Fixed DW	72.48 ± 0.08	91.63 ± 0.08
w/o PGRF	72.71 ± 0.07	91.84 ± 0.09
w/o Calib.	72.21 ± 0.05	91.66 ± 0.07
w/o SS	72.65 ± 0.06	91.76 ± 0.04

Among the ablated components, the historical teacher has the greatest impact on performance. Its removal reduces accuracy by 1.26 percentage points on CIFAR-100 and 1.09 percentage points on CIFAR-10. A fixed distillation coefficient also yields lower accuracy, with decreases of 0.74 and 0.64 percentage points on the two datasets, respectively. These results indicate that adapting the distillation strength to each sample's temporal drift is more effective than applying a fixed coefficient.

The removal of PGRF reduces accuracy by 0.51 percentage points on CIFAR-100 and 0.43 percentage points on CIFAR-10. Post-aggregation calibration and server smoothing also improve performance. In particular, post-aggregation calibration yields a 1.01-percentage-point improvement on CIFAR-100 compared with the corresponding ablated variant. Overall, the ablation results indicate that trajectory-guided historical supervision makes the largest contribution among the evaluated components, while adaptive distillation weighting, progressive gated residual fusion, post-aggregation calibration, and server smoothing provide additional benefits.

5 Conclusion

We proposed FedEAGD, a personalized federated learning framework that uses client-specific shared-model trajectories to provide temporal supervision at the sample level. By comparing current and historical shared representations, the framework adjusts the strength of historical distillation according to the extent of representation changes for each sample. Progressive gated residual fusion further enables the gradual integration of client-specific information into shared representations. During federated training, only the shared feature extractor is exchanged between the server and clients; the personalized feature extractor, gating module, and local classifier are not transmitted to the server.

Experiments on MNIST, CIFAR-10, CIFAR-100, and DomainNet show that FedEAGD performs consistently across different levels of data heterogeneity, client participation rates, and federation scales. On CIFAR-100 with $\alpha = 0.1$, FedEAGD achieves 63.05% accuracy, surpassing the strongest baseline by 5.96 percentage points. The ablation results further show that trajectory-guided historical supervision makes the largest contribution among the evaluated components, while progressive gated residual fusion and local–global coordination provide complementary benefits. Future work will examine asynchronous client participation, deployment under communication constraints, and more complex heterogeneous network environments. We will also investigate the applicability of FedEAGD to distributed systems such as edge computing and the Internet of Things (IoT).

Statements and Declarations

Funding

This work was supported by the Soft Science Research Project of Henan Province (No. 242400411260) and the Science and Technology Research Project of Henan Province, China (No. 262102210203). The funders had no role in the study design, data collection, data analysis, interpretation of the results, manuscript preparation, or the decision to submit the article for publication.

Competing Interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Ethics Approval and Consent to Participate

This study used only publicly available benchmark image datasets (MNIST, CIFAR-10, CIFAR-100, and DomainNet). No human participants were recruited or directly involved in the research, and no new personal or sensitive data were collected. Therefore, ethics approval and consent to participate were not applicable to this study.

Consent for Publication

Not applicable. This manuscript does not report individual participant data or contain identifiable personal information or images requiring consent for publication.

Data Availability

The datasets used in this study are publicly available from their original sources: MNIST at http://yann.lecun.com/exdb/mnist/, CIFAR-10 and CIFAR-100 at https://www.cs.toronto.edu/~kriz/cifar.html, and DomainNet at http://ai.bu.edu/DomainNet/. The corresponding dataset publications are cited in the manuscript (LeCun et al. 1998; Krizhevsky 2009; Peng et al. 2019). No proprietary or restricted-access datasets were used.

Materials Availability

Not applicable.

Code Availability

The source code and experimental implementation of FedEAGD are publicly available at https://github.com/mengjiahaoyun-bot/FedEAGD.

Author Contributions

Mengjia Lu: Conceptualization, Methodology, Software, Validation, Formal analysis, Visualization, Writing – original draft.

Ping Zhang: Conceptualization, Methodology, Supervision, Project administration, Writing – review and editing.

An Bao: Methodology, Formal analysis, Validation, Investigation, Visualization, Writing – review and editing.

Keyi Xu: Software, Validation, Investigation, Writing – review and editing.

Jiechang Wang: Investigation, Validation, Resources, Writing – review and editing.

References

Arivazhagan MG, Aggarwal V, Singh AK, et al (2019) Federated learning with personalization layers. arXiv preprint URL https://arxiv.org/abs/1912.00818, arXiv:1912.00818

Chen HY, Chao WL (2022) On bridging generic and personalized federated learning for image classification. In: International Conference on Learning Representations, URL https://openreview.net/forum?id=I1hQbx10Kxn

Collins L, Hassani H, Mokhtari A, et al (2021) Exploiting shared representations for personalized federated learning. In: Proceedings of the 38th International Conference on Machine Learning, Proceedings of Machine Learning Research, vol 139. PMLR, pp 2089–2099, URL https://proceedings.mlr.press/v139/collins21a.html

Dinh CT, Tran NH, Nguyen TD (2020) Personalized federated learning with moreau envelopes. In: Advances in Neural Information Processing Systems, pp 21394–21405, URL https://proceedings.neurips.cc/paper/2020/hash/f4f1f13c8289ac1b1ee0ff176b56fc60-Abstract.html

Fu Y, Yin S, Ma L, et al (2026) Learning from history for personalized federated learning. Neural Networks 193:108071. https://doi.org/10.1016/j.neunet.2025.108071

Hinton G, Vinyals O, Dean J (2015) Distilling the knowledge in a neural network. arXiv preprint URL https://arxiv.org/abs/1503.02531, arXiv:1503.02531

Hu M, Zhang P, Zhang H, et al (2025) FedHG: Generator-driven head adaptive aggregation for personalized federated learning. In: 2025 International Joint Conference on Neural Networks, pp 1–9, https://doi.org/10.1109/IJCNN64981.2025.11229156

Huang Y, Shao Z, Yang Z, et al (2025) FedRIR: Rethinking information representation in federated learning. In: Proceedings of the ACM Web Conference 2025. Association for Computing Machinery, pp 807–816, https://doi.org/10.1145/3696410.3714612

Jeong E, Oh S, Kim H, et al (2018) Communication-efficient on-device machine learning: Federated distillation and augmentation under non-IID private data. arXiv preprint URL https://arxiv.org/abs/1811.11479, arXiv:1811.11479

Ji S, Tan Y, Saravirta T, et al (2024) Emerging trends in federated learning: from model fusion to federated x learning. International Journal of Machine Learning and Cybernetics 15:3769–3790. https://doi.org/10.1007/s13042-024-02119-1

Kairouz P, McMahan HB, et al (2021) Advances and open problems in federated learning. Foundations and Trends in Machine Learning 14(1–2):1–210. https://doi.org/10.1561/2200000083

Karami A, Ramezani R, Baraani-Dastjerdi A (2024) GFD-SSL: Generative federated knowledge distillation-based semi-supervised learning. International Journal of Machine Learning and Cybernetics 15(12):5509–5529. https://doi.org/10.1007/s13042-024-02256-7

Krizhevsky A (2009) Learning multiple layers of features from tiny images. Tech. rep., University of Toronto, URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf

LeCun Y, Bottou L, Bengio Y, et al (1998) Gradient-based learning applied to document recognition. Proceedings of the IEEE 86(11):2278–2324. https://doi.org/10.1109/5.726791

Li D, Wang J (2019) FedMD: Heterogenous federated learning via model distillation. arXiv preprint URL https://arxiv.org/abs/1910.03581, arXiv:1910.03581

Li Q, He B, Song D (2021a) Model-contrastive federated learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 10713–10722, https://doi.org/10.1109/CVPR46437.2021.01057

Li T, Sahu AK, Talwalkar A, et al (2020a) Federated learning: Challenges, methods, and future directions. IEEE Signal Processing Magazine 37(3):50–60. https://doi.org/10.1109/MSP.2020.2975749

Li T, Sahu AK, Zaheer M, et al (2020b) Federated optimization in heterogeneous networks. In: Proceedings of Machine Learning and Systems, pp 429–450, URL https://proceedings.mlsys.org/paper_files/paper/2020/hash/1f5fe83998a09396ebe6477d9475ba0c-Abstract.html

Li T, Hu S, Beirami A, et al (2021b) Ditto: Fair and robust federated learning through personalization. In: Proceedings of the 38th International Conference on Machine Learning, Proceedings of Machine Learning Research, vol 139. PMLR, pp 6357–6368, URL https://proceedings.mlr.press/v139/li21h.html

Li X, Jiang M, Zhang X, et al (2021c) FedBN: Federated learning on non-IID features via local batch normalization. In: International Conference on Learning Representations, URL https://openreview.net/forum?id=6YEQUn0QICG

Liu W, Chen J, Wang B, et al (2025a) Per-FedAHM: Adaptive historical memory-driven personalized federated learning. Neurocomputing 658:131716. https://doi.org/10.1016/j.neucom.2025.131716

Liu W, Wang Y, Wang B, et al (2026) Per-FedDMA: A personalized federated learning method based on deep multisampling and hypernetwork dynamic adaptation. Computer Communications 251:108474. https://doi.org/10.1016/j.comcom.2026.108474

Liu Z, Luo Y, Zhu T, et al (2025b) FedBH: Efficient federated learning through shared base and adaptive hyperparameters in heterogeneous systems. Computer Communications 239:108190. https://doi.org/10.1016/j.comcom.2025.108190

McMahan B, Moore E, Ramage D, et al (2017) Communication-efficient learning of deep networks from decentralized data. In: Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, Proceedings of Machine Learning Research, vol 54. PMLR, pp 1273–1282, URL https://proceedings.mlr.press/v54/mcmahan17a.html

Oh J, Kim S, Yun SY (2022) FedBABU: Toward enhanced representation for federated image classification. In: International Conference on Learning Representations, URL https://openreview.net/forum?id=HuaYQfggn5u

Peng X, Bai Q, Xia X, et al (2019) Moment matching for multi-source domain adaptation. In: Proceedings of the IEEE/CVF International Conference on Computer

Vision, pp 1406–1415, https://doi.org/10.1109/ICCV.2019.00149

Qi G, Qu Z, Lyu SH, et al (2024) Personalized federated learning with feature alignment via knowledge distillation. In: PRICAI 2024: Trends in Artificial Intelligence, Lecture Notes in Computer Science, vol 15282. Springer, pp 121–133, https://doi.org/10.1007/978-981-96-0119-6_12

Tan AZ, Yu H, Cui L, et al (2023) Towards personalized federated learning. IEEE Transactions on Neural Networks and Learning Systems 34(12):9587–9603. https://doi.org/10.1109/TNNLS.2022.3160699

Tarvainen A, Valpola H (2017) Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In: Advances in Neural Information Processing Systems, pp 1195–1204, URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/68053af2923e00204c3ca7c6a3150cf7-Abstract.html

Wang B, Zhao X, Yin Z, et al (2026) Federated personalization update strategy based on data-free knowledge mutual distillation. International Journal of Machine Learning and Cybernetics 17:22. https://doi.org/10.1007/s13042-025-02846-z

Wen J, Zhang Z, Lan Y, et al (2023) A survey on federated learning: challenges and applications. International Journal of Machine Learning and Cybernetics 14(2):513–535. https://doi.org/10.1007/s13042-022-01647-y

Wu C, Wu F, Lyu L, et al (2022) Communication-efficient federated learning via knowledge distillation. Nature Communications 13:2032. https://doi.org/10.1038/s41467-022-29763-x

Wu X, Liu X, Niu J, et al (2023) Bold but cautious: Unlocking the potential of personalized federated learning through cautiously aggressive collaboration. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 19375–19384, https://doi.org/10.1109/ICCV51070.2023.01775

Xu J, Zhou L, Zhao Y, et al (2025) A two-stage federated learning method for personalization via selective collaboration. Computer Communications 232:108053. https://doi.org/10.1016/j.comcom.2025.108053

Yang X, Huang W, Ye M (2024) FedAS: Bridging inconsistency in personalized federated learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 11986–11995, https://doi.org/10.1109/CVPR52733.2024.01139

Yao B, Li C, Chen W, et al (2026) FedCAP: Community-aware aggregation and local alignment for personalized federated learning. Future Generation Computer Systems 182:108495. https://doi.org/10.1016/j.future.2026.108495

Yi L, Wang G, Liu X, et al (2023) FedGH: Heterogeneous federated learning with generalized global header. In: Proceedings of the 31st ACM International Conference on Multimedia. Association for Computing Machinery, pp 8686–8696, https://doi.org/10.1145/3581783.3611781

Zhang J, Hua Y, Wang H, et al (2023a) GPFL: Simultaneously learning global and personalized feature information for personalized federated learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 5041–5051, https://doi.org/10.1109/ICCV51070.2023.00465

Zhang J, Hua Y, Wang H, et al (2023b) FedALA: Adaptive local aggregation for personalized federated learning. Proceedings of the AAAI Conference on Artificial Intelligence 37(9):11237–11244. https://doi.org/10.1609/aaai.v37i9.26330

Zhang J, Hua Y, Wang H, et al (2023c) FedCP: Separating feature information for personalized federated learning via conditional policy. In: Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. Association for Computing Machinery, pp 3249–3261, https://doi.org/10.1145/3580305.3599345

Zhang L, Niu J, Zhang W, et al (2026a) A two-stage personalized federated learning approach based on a hybrid expert architecture. Neurocomputing 685:133648. https://doi.org/10.1016/j.neucom.2026.133648

Zhang P, Bao A, Lu W, et al (2026b) FedTDC: Federated teacher-guided distillation with representation and aggregation calibration. Knowledge-Based Systems 343:116020. https://doi.org/10.1016/j.knosys.2026.116020

Zhang S, Wang X, Zeng R, et al (2025) A personalized federated cloud-edge collaboration framework via cross-client knowledge distillation. Future Generation Computer Systems 165:107594. https://doi.org/10.1016/j.future.2024.107594

Zhang Z, Li R, Guo K, et al (2026c) Personalized federated learning with personalized knowledge distillation based on local empirical risk. Information Fusion 133:104269. https://doi.org/10.1016/j.inffus.2026.104269

Zhong H, Zhu L, Liu Y (2026) Personalized federated learning with two-stage adaptive aggregation under global feature constraints. Computer Communications 254:108552. https://doi.org/10.1016/j.comcom.2026.108552

Zhu Z, Hong J, Zhou J (2021) Data-free knowledge distillation for heterogeneous federated learning. In: Proceedings of the 38th International Conference on Machine Learning, Proceedings of Machine Learning Research, vol 139. PMLR, pp 12878–12889, URL https://proceedings.mlr.press/v139/zhu21b.html