Beyond Aggregation: Guiding Clients in Heterogeneous Federated Learning

Anonymous authors
Paper under double-blind review

Abstract

Federated learning (FL) is increasingly adopted in domains like healthcare, where data privacy is paramount. A fundamental challenge in these systems is statistical heterogeneity—the fact that data distributions vary significantly across clients (e.g., different hospitals may treat distinct patient demographics). While current FL algorithms focus on aggregating model updates from these heterogeneous clients, the potential of the central server remains under-explored. This paper is motivated by a healthcare scenario: could a central server not only coordinate model training but also guide a new patient to the hospital best equipped for their specific condition? We generalize this idea to propose a novel paradigm for FL systems where the server actively guides the allocation of new tasks or queries to the most appropriate client. To enable this, we introduce a density ratio model and empirical likelihood-based framework that simultaneously addresses two goals: (1) learning effective local models on each client, and (2) finding the best matching client for a new query. Empirical results demonstrate the framework's effectiveness on benchmark datasets, showing improvements in both model accuracy and the precision of client guidance compared to standard FL approaches. This work opens a new direction for building more intelligent and resource-efficient FL systems that leverage heterogeneity as a feature, not just a bug.

1 Introduction

Federated learning (FL) has emerged as a powerful paradigm for training machine learning models across distributed data sources without sharing raw data. By enabling clients such as hospitals, financial institutions, or mobile devices to collaboratively train models under the coordination of a central server, FL offers a practical solution for privacy-preserving learning in sensitive domains (Li et al., 2020a; Long et al., 2020; Xu et al., 2021).

A key challenge in applying FL in practice is statistical heterogeneity: clients often hold data drawn from different, non-identically distributed populations. In healthcare, hospitals may serve distinct patient demographics; in finance, banks may encounter different fraud patterns; and on mobile devices, user behavior varies widely. Such heterogeneity can cause local models to drift apart, leading to slower convergence (Li et al., 2020b), biased updates (Karimireddy et al., 2020), and global models that underperform when applied back to individual clients (T Dinh et al., 2020). To address these issues, most existing FL systems treat heterogeneity as a problem to be suppressed—through aggregation corrections, client reweighting, or personalization techniques. In this prevailing paradigm, the central server plays a largely passive role, acting only as a coordinator that aggregates local updates into a single global model. We contend, however, that this limited role overlooks a key opportunity: rather than merely mitigating heterogeneity, the server can actively exploit it.

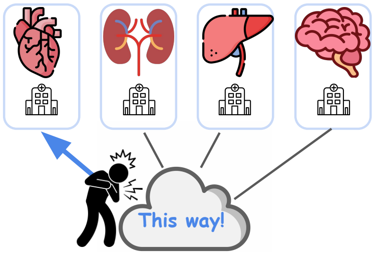


Figure 1: **FL server as an intelligent router**: Leveraging learned data distributions to direct queries to the most specialized client, rather than applying a global model for diagnosis.

Consider a healthcare scenario: different hospitals may excel at treating different patient groups depending on their location and/or expertise. When a new patient arrives, instead of merely deploying a global model for diagnosis, the server could help identify the hospital best equipped to provide care, leveraging local data distributions to capture specialized expertise. A cartoon illustration of this scenario is given in Fig. 1. Similar opportunities exist in other domains: in finance, the server could direct a fraud detection query to the bank whose historical data best matches the transaction profile; in personalized services, it could route a query to the client with the most relevant user base. These examples illustrate that statistical heterogeneity across clients—often seen as an obstacle—can instead become a valuable resource. They motivate the central insight of our work:

> *Beyond coordinating training, the server can actively exploit client heterogeneity–transforming it from a challenge into a resource by guiding new queries to the most suitable client.*

Much of the existing work in FL has focused on mitigating the challenges of statistical heterogeneity, without using the server for guiding new queries. One major line of research develops aggregation algorithms to reduce the bias induced by non-identically distributed data. Examples include methods that modify local updates before aggregation (Gao et al., 2022; Guo et al., 2023; Zhang et al., 2023), reweight client contributions (Wang et al., 2020; Yin et al., 2024), or introduce regularization terms to align local objectives with the global one (Li et al., 2020b; Acar et al., 2021; Li et al., 2021b). These approaches aim to learn a single global model that performs reasonably well across all clients, but they do not leverage heterogeneity as an asset. A second line of work explores personalization in FL. Rather than enforcing a universal global model, personalization methods adapt models to each client's local distribution (Li et al., 2021d), often through fine-tuning (T Dinh et al., 2020; Collins et al., 2021; Tan et al., 2022; Ma et al., 2022), multi-task learning (Smith et al., 2017; Li et al., 2021c), or meta-learning (Fallah et al., 2020). While these approaches improve local performance, they are typically not designed to address the challenge of guiding new queries or tasks to the most appropriate client. Another related direction is client clustering (Ghosh et al., 2020; Li et al., 2021a; Briggs et al., 2020; Kim et al., 2021; Long et al., 2023), where clients with similar data distributions are grouped and trained jointly within each cluster. This can improve performance under heterogeneity, but still assumes the server's role is limited to coordinating training and distributing models, rather than supporting query routing or task allocation. *Overall, while these approaches are effective for their intended goals, they stop short of enabling the server to actively guide new queries to the most suitable client.*

Motivated by this gap, we introduce a new paradigm in which the FL server not only coordinates training but also learns to guide each incoming query to the client best suited to handle it. Achieving this goal requires two capabilities: (i) effective information sharing across clients despite heterogeneity, and (ii) a principled way to quantify how each client's data distribution differs from the others so that queries can be meaningfully matched. To achieve this, we develop **FedDRM**, a unified framework grounded in density ratio model (DRM) (Anderson, 1979) and empirical likelihood (EL) (Owen, 2001). DRM represents each client's distribution as a *multiplicative density tilt of a baseline distribution*, while EL facilitates nonparametric model learning, enabling the estimation of this baseline distribution in a data-driven manner without parametric assumptions. After profiling out the baseline distribution, the resulting objective decomposes into two interpretable cross-entropy components: one for predicting class labels and another for identifying a sample's client of origin. The first supports standard FL training; the second supplies precisely the signal needed for query-to-client routing, enabling the server to exploit—rather than suppress—statistical heterogeneity.

This formulation leads to three key contributions. *First*, we propose the first statistically grounded FL framework that jointly learns heterogeneous predictive models and the distributional structure required for query routing within a single principled objective. *Second*, we develop a new algorithmic correction for the classification component of the EL objective. Because each client is associated with only a single class label for client identification, the vanilla loss suffers from an extreme form of label shift; we propose a simple yet effective reweighting adjustment that yields a more stable classifier. *Third*, through experiments on benchmark datasets, we demonstrate that our approach consistently improves both predictive accuracy and routing precision compared to standard FL methods, underscoring the benefits of integrating guidance directly into the FL workflow. Together, these developments transform the FL server from a passive aggregator into an intelligent router capable of directing queries to the most suitable client, opening the door to FL systems that are not only privacy-preserving but also adaptive and expertise-aware.

2 FedDRM: Guiding Clients in Heterogeneous FL

2.1 Probabilistic description of data heterogeneity

Consider an FL system with m clients. Let $\mathcal{D}_i := \{(X_{ij}, Y_{ij})\}_{j=1}^{n_i}$ denote the training set on the i-th client, where each sample is drawn independently from $P_{X,Y}^{(i)}$. We consider the multi-class classification case where $Y_{ij} \in [K] := \{1, \ldots, K\}$ with marginal distribution $\mathbb{P}(Y_{ij} = k) = \pi_{ik}$ for $k \in [K]$, and features conditioned on the labels are distributed as $X_{ij}|(Y_{ij} = k) \sim P_k^{(i)}$. We denote the marginal distribution of the features on client i as $P_X^{(i)}$, and the conditional distribution of Y given $X = x$ as $\{\mathbb{P}^{(i)}(Y = k|X = x)\}_{k=1}^K$. Different types of data heterogeneity can be described in terms of the family of distributions $\{P_{X,Y}^{(i)}\}_{i=1}^m$:

- **Covariate shift**: Clients differ in their marginal feature distributions while sharing the same conditional label distribution. In our notation, this corresponds to

$$P_X^{(i)} \neq P_X^{(i')} \text{ for } i \neq i', \text{ but } \mathbb{P}^{(i)}(Y = k|X = x) = \mathbb{P}^{(i')}(Y = k|X = x) \text{ for all } x \text{ and } k.$$

- **Label shift**: Clients have different label marginals but share the same conditional feature distributions given the label. Equivalently,

$$\boldsymbol{\pi}_i := (\pi_{i1}, \ldots, \pi_{iK}) \neq \boldsymbol{\pi}_{i'} := (\pi_{i'1}, \ldots, \pi_{i'K}) \text{ for some } i \neq i', \text{ but } P_k^{(i)} = P_k^{(i')} \text{ for all } k.$$

In practice, real-world federated systems often exhibit combinations of these shifts, which leads to the **full distributional shift** where both $\boldsymbol{\pi}_i$ and $\{P_k^{(i)}\}_{k=1}^K$ may vary across clients.

2.2 A semiparametric density ratio model

For clarity, we begin with the special case of *covariate shift* across clients. Extensions to other types of heterogeneity will then follow naturally. Let $g_\theta(x)$ represent a feature embedding (*e.g.,* an embedding from a DNN parameterized by θ) *s.t.* the conditional distribution of $Y|X$ is given by:

$$\mathbb{P}(Y = k|X = x) = \frac{\exp(\alpha_k + \beta_k^\top g_\theta(x))}{\sum_j \exp(\alpha_j + \beta_j^\top g_\theta(x))}. \tag{1}$$

We drop the superscript (i) since this conditional distribution remains the same across all clients under covariate shift. Applying Bayes' rule to (1), we derive that the class-conditional distributions are connected by an exponential function:

$$dP_k^{(i)}/dP_1^{(i)}(x) = \exp(\alpha_{ik}^\dagger + \beta_k^\top g_\theta(x)) \tag{2}$$

where $dP_k^{(i)}/dP_1^{(i)}$ denotes the Radon–Nikodym derivative of $dP_k^{(i)}$ with respect to $dP_1^{(i)}$ and $\alpha_{ik}^\dagger = \alpha_k + \log(\pi_{i1}/\pi_{ik})$ for $i \in [m]$.

To facilitate knowledge transfer across clients in FL, we assume their datasets share some common underlying statistical structure. Specifically, we relate the client distributions $\{P_1^{(l)}\}_{l=1}^m$ through a hypothetical reference measure $P_1^{(0)}$ at the server, using DRM[1] (Anderson, 1979):

$$dP_k^{(i)}/dP_1^{(i)}(x) = \exp(\gamma_i + \xi_i^\top h_\tau(g_\theta(x))) \tag{3}$$

where $h_\tau(\cdot)$ is a parametric function with parameters τ. We refer to $P_1^{(0)}$ as a *hypothetical* reference since the server may not have data directly, although the formulation also applies when server-side data are available. The DRM captures differences in the conditional distributions of $X|Y = 1$ across clients via density ratios, with log-ratios modeled linearly in the embeddings. This avoids estimating each distribution separately, focusing instead on relative differences. When the covariate shift is not too severe, the marginal distributions of different clients are connected through this

[1] DRM provides a framework for modeling the relationship between two or more populations that share similar characteristics. It is highly flexible and encompasses several commonly used parametric distribution families—such as the binomial, exponential, and normal families—as special cases (Kay & Little, 1987).

parametric form, making FL effective by leveraging shared structure across clients. On the other hand, if the distributions differ too drastically, combining data from different clients is unlikely to improve performance; in this case, even if the DRM assumption does not hold, it is not a limitation of the formulation but a consequence of the inherent nature of the problem. Thus, the assumption is reasonable in practice. When $\gamma_i = 0$ and $\xi_i = 0$, (3) reduces to the IID case. Under this assumption, we obtain the following relationship between the marginal feature distributions:

Theorem 2.1. *With* (2) *and* (3)*, the marginal distributions of X also satisfy the DRM:*

$$dP_X^{(i)}/dP_X^{(0)}(x) = \exp\{\gamma_i^\dagger + \xi_i^\top h_\tau(g_\theta(x))\} \tag{4}$$

where $\gamma_i^\dagger = \gamma_i + \log(\pi_{i1}/\pi_{01})$ for all $i \in [m]$, and $P_X^{(0)}$ an unspecified reference measure.

See proof in App. C.1. This theorem relates each client's marginal distribution to the reference distribution through a parametric tilt, which directly facilitates construction of the likelihood. If the reference measure $P_X^{(0)}$ was fully specified, all $\{P_{X,Y}^{(i)}\}_{i=1}^m$ would also be fully determined, and one could estimate the unknown model parameters using a standard maximum likelihood approach. In practice, however, $P_X^{(0)}$ is unknown, and assuming it follows a parametric family risks model mis-specification and potentially biased inference.

To address this challenge, we adopt a flexible, nonparametric approach based on EL (Owen, 2001) that integrate data across heterogeneous populations via density ratio modeling (Qin & Zhang, 1997; Fokianos et al., 2001; Chen & Liu, 2013; Li et al., 2017; Liu et al., 2017; 2025). EL constructs likelihood functions directly from the observed data without requiring a parametric form. Instead of specifying a probability model, it assigns probabilities to the observed samples and maximizes the nonparametric likelihood subject to constraints, such as moment conditions. Unlike classical parametric likelihood, EL adapts flexibly to the data, making it particularly suitable when the underlying distribution is unknown or complex, but valid structural or moment conditions are available. Specifically, we let

$$p_{ij} = P_X^{(0)}(\{X_{ij}\}) \geq 0, \quad \forall i \in [m],\ j \in [n_i],$$

treating the p_{ij} as parameters. In this way, the reference measure $P_X^{(0)}$ is represented as an atomic measure without any parametric assumptions, and most importantly *all samples across clients are leveraged for information sharing*. To ensure that $P_X^{(0)}$ and $\{P_X^{(i)}\}_{i=1}^m$ are valid probability measures, the following constraints are imposed:

$$\sum_{i=1}^m \sum_{j=1}^{n_i} p_{ij} = 1, \quad \sum_{i=1}^m \sum_{j=1}^{n_i} p_{ij} \exp\left\{\gamma_l^\dagger + \xi_l^\top h_\tau(g_\theta(X_{ij}))\right\} = 1, \quad \forall\, l \in [m]. \tag{5}$$

2.3 A SURPRISINGLY SIMPLE DUAL LOSS

With the semiparametric DRM for heterogeneous FL, we propose a maximum likelihood approach for model learning. Let $\boldsymbol{p} = \{p_{ij}\}$, $\boldsymbol{\alpha} = \{\alpha_k\}$, $\boldsymbol{\beta} = \{\beta_k\}$, $\boldsymbol{\gamma}^\dagger = \{\gamma_i^\dagger\}$, $\boldsymbol{\xi} = \{\xi_i\}$, and $\boldsymbol{\zeta} = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\xi}, \theta, \tau)$, the log empirical likelihood of the model based on datasets across clients is

$$\begin{aligned}\ell_N(\boldsymbol{p}, \boldsymbol{\zeta}) &= \sum_{i,j} \log P_{X,Y}^{(i)}(\{X_{ij}, Y_{ij}\}) = \sum_{i,j,k} \mathbb{1}(Y_{ij} = k) \log \mathbb{P}(Y = k|X_{ij}) + \sum_{i,j} \log P_X^{(i)}(\{X_{ij}\}) \\ &= \sum_{i,j,k} \mathbb{1}(Y_{ij} = k) \log \mathbb{P}(Y = k|X_{ij}) + \sum_{i,j} \{\gamma_i^\dagger + \xi_i^\top h_\tau(g_\theta(X_{ij})) + \log p_{ij}\},\end{aligned}$$

where the last equality makes use of Theorem 2.1. Since our goal is to learn (1) on each client, the weight $\boldsymbol{p}$ becomes a nuisance parameter, which we profile out to learn the parameters that are connected to the conditional distribution of $Y|X = x$. The profile log-EL of $\boldsymbol{\zeta}$ is defined as $p\ell_N(\boldsymbol{\zeta}) = \sup_{\boldsymbol{p}} \ell_N(\boldsymbol{p}, \boldsymbol{\zeta})$ where the supremum is under constraints (5). By the method of Lagrange multiplier, we show in App. C.2 that an analytical form of the profile log-EL is

$$p\ell_N(\boldsymbol{\zeta}) = \sum_{i,j,k} \mathbb{1}(Y_{ij} = k) \log \mathbb{P}(Y_{ij} = k|X_{ij}) + \sum_{i,j} \{\gamma_i^\dagger + \xi_i^\top h_\tau(g_\theta(x_{ij})) + \log p_{ij}(\boldsymbol{\zeta})\} \tag{6}$$

where $p_{ij}(\boldsymbol{\zeta}) = N^{-1}\Big\{1 + \sum_{l=1}^{m} \rho_l \Big[\exp\{\gamma_l^{\dagger} + \xi_l^{\top} h_\tau(g_\theta(x_{ij}))\} - 1\Big]\Big\}^{-1}$ and the Lagrange multipliers $\{\rho_l\}_{l=1}^{m}$ are the solution to

$$\sum_{i,j} \frac{\exp\{\gamma_l^{\dagger} + \xi_l^{\top} h_\tau(g_\theta(x_{ij}))\} - 1}{\sum_{l'} \rho_{l'} \Big[\exp(\gamma_{l'}^{\dagger} + \xi_{l'}^{\top} h_\tau(g_\theta(x_{ij}))) - 1\Big]} = 0.$$

Although the profile log-EL in (6) has a closed analytical form, computing it typically requires solving a system of m equations for the Lagrange multipliers, which can be computationally demanding. Interestingly, at the optimal solution these multipliers admit a closed-form expression, yielding a surprisingly simple dual formulation of the profile log-EL presented below.

Theorem 2.2 (Dual form)**.** *At optimality, the Lagrange multipliers $\rho_l = n_l/N$ and the profile log-EL in* (6) *becomes*

$$p\ell_N(\boldsymbol{\zeta}) = \sum_{i,j} \log\left\{\frac{\exp(\gamma_i^{\ddagger} + \xi_i^{\top} h_\tau(g_\theta(x_{ij})))}{\sum_l \exp(\gamma_l^{\ddagger} + \xi_l^{\top} h_\tau(g_\theta(x_{ij})))}\right\} + \sum_{i,j} \log\left\{\frac{\exp(\alpha_{y_{ij}} + \beta_{y_{ij}}^{\top} g_\theta(x_{ij}))}{\sum_k \exp(\alpha_k + \beta_k^{\top} g_\theta(x_{ij}))}\right\}$$

up to some constant where $\gamma_i^{\ddagger} = \log(n_i/n_1) + \gamma_i^{\dagger}$.

See App. C.3 for proof. The theorem allows us to define the overall loss function as the negative profile log-EL:

$$\ell(\boldsymbol{\zeta}) = -p\ell_N(\boldsymbol{\zeta}) = \sum_{i,j} \ell_{\text{CE}}(i, h_\tau(g_\theta(x_{ij})); \boldsymbol{\gamma}, \boldsymbol{\xi}) + \sum_{i,j} \ell_{\text{CE}}(y_{ij}, g_\theta(x_{ij}); \boldsymbol{\alpha}, \boldsymbol{\beta}), \quad (7)$$

where $\ell_{\text{CE}}(y, x; \boldsymbol{\alpha}, \boldsymbol{\beta}) = -(\alpha_y + \beta_y^{\top} x) + \log\{\sum_k \exp(\alpha_k + \beta_k^{\top} x)\}$ is the cross-entropy loss.

Remark 2.3 (**Beyond covariate shift**). Our method is described under covariate shift. The derivations in key steps (2) and (3) do not require the marginal distribution of Y to be identical across clients, which allows us to also accommodate label shift. Importantly, we show that *our approach extends to the more general setting where both $Y|X$ and X differ across clients* in App. D. In this case, after a detailed derivation, we find that the overall loss simplifies to a minor adjustment in the target-class classification head. Concretely, the target-class classification loss is equipped with a client-specific linear head, resulting in the final architecture shown in Fig. 2. Interestingly, this architecture closely resembles those in personalized FL methods such as Collins et al. (2021): target-class classification is performed with client-specific heads, while our new client classification component relies on a single shared head across all clients.

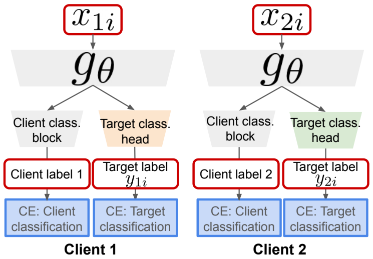


Figure 2: **Network architecture**. Gray blocks are shared among all clients, while colored blocks are specific to each client.

Remark 2.4 (**Guiding new queries**). Although the derivation is mathematically involved, the resulting loss function is remarkably simple: it consists of two cross-entropy terms, each associated with a distinct classification task. The first term identifies the client from which a sample originates, while the second predicts its target class. The additional client-classification head thus yields, for any query, the probability of belonging to each client. By routing a query to the client with the highest predicted probability, we obtain a principled mechanism for assigning new data to the client best equipped to handle it.

2.4 Optimization Algorithm

The overall loss $\ell(\boldsymbol{\zeta})$ in (7) is defined as if all datasets were pooled together. Since optimizing $\ell(\boldsymbol{\zeta})$ with vanilla SGD and weight decay is equivalent to minimizing a loss function with an explicit L_2 penalty, we denote the loss as $\ell^{\rho}(\boldsymbol{\zeta}) = \ell(\boldsymbol{\zeta}) + (\rho/2)\|\boldsymbol{\zeta}\|_2^2$, with minimizer $\tilde{\boldsymbol{\zeta}}_N$. The subscript N is used to indicate that this weight is based on N samples. In the FL setting, the global loss decomposes naturally into client-specific contributions: $\ell^{\rho}(\boldsymbol{\zeta}) = \sum_{i=1}^{m}(n_i/N)\ell_i(\boldsymbol{\zeta})$ where

$$\ell_i(\boldsymbol{\zeta}) = \ell_i(\boldsymbol{\gamma}, \boldsymbol{\xi}) + \ell_i(\boldsymbol{\alpha}, \boldsymbol{\beta}) + (\rho/2)\|\boldsymbol{\zeta}\|_2^2,$$

$\ell_i(\boldsymbol{\gamma}, \boldsymbol{\xi}) = n_i^{-1} \sum_j \ell_{\text{CE}}(i, h_\tau(g_\theta(x_{ij})); \boldsymbol{\gamma}, \boldsymbol{\xi})$, and $\ell_i(\boldsymbol{\alpha}, \boldsymbol{\beta}) = n_i^{-1} \sum_j \ell_{\text{CE}}(y_{ij}, g_\theta(x_{ij}); \boldsymbol{\alpha}, \boldsymbol{\beta})$.

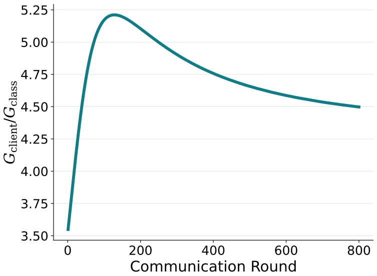


Figure 3: **Relative gradient drift**.

A key difference arises between these two terms. For the client-classification loss $\ell_i(\boldsymbol{\gamma}, \boldsymbol{\xi})$, the i-th client only observes samples labeled with its own client index i. In contrast, the target-class loss $\ell_i(\boldsymbol{\alpha}, \boldsymbol{\beta})$ typically spans multiple target labels per client (though with varying proportions). This asymmetry leads to more pronounced *gradient drift*[2] in $\nabla \ell_i(\boldsymbol{\gamma}, \boldsymbol{\xi})$. To illustrate, consider the gradient of the client-classification loss with respect to γ_k:

$$\partial \ell_i / \partial \gamma_k = n_i^{-1} \sum_j x_{ij} \{\mathbb{1}(i = k) - p_k(h_\tau(g_\theta(x_{ij})); \boldsymbol{\gamma}, \boldsymbol{\xi})\}$$

where $p_k(x; \boldsymbol{\gamma}, \boldsymbol{\xi}) = \exp(\gamma_k + \xi_k^\top x) / \sum_l \exp(\gamma_l + \xi_l^\top x)$. Since $\mathbb{1}(i = k) = 0$ for all $k \neq i$, the gradient contributed by client i provides no meaningful information about other clients' parameters. As a result, local updates to the client-classification head are inherently biased, which in turn amplifies gradient drift relative to target-class head. Fig. 3 shows this effect on a 10-class classification task with 3 clients and a randomly generated embedding using FedAvg: the gradient drift for client classification is markedly more severe than that for target classification.

Reweighting strategy. To address this, we draw on reweighting principles Chen et al. (2018); Liu et al. (2021) to propose a simple yet effective method with theoretical guarantees. Our approach down-weights client classification loss, whose gradient exhibits larger drift, resulting in the per-client loss:

$$\tilde{\ell}_i(\boldsymbol{\zeta}) = (1 - \lambda)\ell_i(\boldsymbol{\gamma}, \boldsymbol{\xi}) + \lambda \ell_i(\boldsymbol{\alpha}, \boldsymbol{\beta}),$$

for $\lambda > 0.5$ and the reweighted global loss is $\tilde{\ell}(\boldsymbol{\zeta}) = \sum_{i=1}^m (n_i/N)\tilde{\ell}_i(\boldsymbol{\zeta})$, see Algorithm 1.

Algorithm 1: `FedDRM`

Input: Clients m, rounds T, local steps E, learning rate η, trade-off λ

```
Initialize backbone $\theta^{(0)}$, target head $\{(\boldsymbol{\alpha}_i^{(0)}, \boldsymbol{\beta}_i^{(0)})\}_{i=1}^m$, and client head $(\tau^{(0)}, \boldsymbol{\gamma}^{(0)}, \boldsymbol{\xi}^{(0)})$
for $t = 0, 1, \ldots, T-1$ do
    Server broadcasts $\theta^{(t)}$ and $(\tau^{(t)}, \boldsymbol{\gamma}^{(t)}, \boldsymbol{\xi}^{(t)})$ to all clients
    for client $i \in [m]$ in parallel do
        $\theta_i^{(t,0)} \leftarrow \theta^{(t)}, (\boldsymbol{\alpha}_i^{(t,0)}, \boldsymbol{\beta}_i^{(t,0)}) \leftarrow (\boldsymbol{\alpha}_i^{(t)}, \boldsymbol{\beta}_i^{(t)}), (\tau_i^{(t,0)}, \boldsymbol{\gamma}_i^{(t,0)}, \boldsymbol{\xi}_i^{(t,0)}) \leftarrow (\tau^{(t)}, \boldsymbol{\gamma}^{(t)}, \boldsymbol{\xi}^{(t)})$
        for $k = 0, 1, \ldots, E-1$ do
            Get target loss $\ell_i(\boldsymbol{\alpha}_i^{(t,k)}, \boldsymbol{\beta}_i^{(t,k)}, \theta_i^{(t,k)})$ and client loss $\ell_i(\tau_i^{(t,k)}, \boldsymbol{\gamma}_i^{(t,k)}, \boldsymbol{\xi}_i^{(t,k)}, \theta_i^{(t,k)})$
            $\tilde{\ell}_i(\boldsymbol{\zeta}_i^{(t,k)}) \leftarrow \lambda \ell_i(\boldsymbol{\alpha}_i^{(t,k)}, \boldsymbol{\beta}_i^{(t,k)}, \theta_i^{(t,k)}) + (1-\lambda)\ell_i(\tau_i^{(t,k)}, \boldsymbol{\gamma}_i^{(t,k)}, \boldsymbol{\xi}_i^{(t,k)}, \theta_i^{(t,k)})$
            $\boldsymbol{\zeta}_i^{(t,k+1)} \leftarrow \boldsymbol{\zeta}_i^{(t,k)} - \eta \nabla \tilde{\ell}(\boldsymbol{\zeta}_i^{(t,k)})$
        end
        $\theta_i^{(t+1)} \leftarrow \theta_i^{(t,E)}, (\boldsymbol{\alpha}_i^{(t+1)}, \boldsymbol{\beta}_i^{(t+1)}) \leftarrow (\boldsymbol{\alpha}_i^{(t,E)}, \boldsymbol{\beta}_i^{(t,E)}), (\tau_i^{(t+1)}, \boldsymbol{\gamma}_i^{(t+1)}, \boldsymbol{\xi}_i^{(t+1)}) \leftarrow$
        $(\tau_i^{(t,E)}, \boldsymbol{\gamma}_i^{(t,E)}, \boldsymbol{\xi}_i^{(t,E)})$
        Client $i$ sends $\theta_i^{(t+1)}$ and $(\tau_i^{(t+1)}, \boldsymbol{\gamma}_i^{(t+1)}, \boldsymbol{\xi}_i^{(t+1)})$ back to the server
    end
    Server updates
     $\theta^{(t+1)} \leftarrow \sum_{i=1}^m \frac{n_i}{N} \theta_i^{(t+1)}, (\tau^{(t+1)}, \boldsymbol{\gamma}^{(t+1)}, \boldsymbol{\xi}^{(t+1)}) \leftarrow \sum_{i=1}^m \frac{n_i}{N} (\tau_i^{(t+1)}, \boldsymbol{\gamma}_i^{(t+1)}, \boldsymbol{\xi}_i^{(t+1)})$
end
```

To accelerate convergence, a larger value of λ is desirable. However, as $\lambda \to 1$, the target-class classification begins to dominate, which hinders effective training of the client classification and ultimately weakens the model's ability to guide clients. To illustrate the trade-off between accuracy and convergence, we consider a simplified setting where the embedding is fixed (*i.e.*, , θ and τ are known) and the true data-generating mechanism follows a multinomial logistic model with parameters $\boldsymbol{\zeta}^{\text{true}} = (\boldsymbol{\gamma}^{\text{true}}, \boldsymbol{\xi}^{\text{true}}, \boldsymbol{\alpha}^{\text{true}}, \boldsymbol{\beta}^{\text{true}})$. We define the heterogeneity measure $G^2(\boldsymbol{\zeta}) = \sum_{i=1}^m (n_i/N)\|\nabla \ell_i(\boldsymbol{\zeta}) - \nabla \ell(\boldsymbol{\zeta})\|_2^2$, which admits the decomposition $G^2(\boldsymbol{\zeta}) =$

[2] The gradient drift of the client loss is $G_{\text{client}}^2 := \sum_{i=1}^m (n_i/N)\|\nabla \ell_i(\boldsymbol{\gamma}, \boldsymbol{\xi}) - \sum_{l=1}^m (n_i/N)\nabla \ell_l(\boldsymbol{\gamma}, \boldsymbol{\xi})\|^2$, and that of the target-class loss is $G_{\text{class}}^2 := \sum_{i=1}^m (n_i/N)\|\nabla \ell_i(\boldsymbol{\alpha}, \boldsymbol{\beta}) - \sum_{l=1}^m (n_i/N)\nabla \ell_l(\boldsymbol{\alpha}, \boldsymbol{\beta})\|^2$.

$(1-\lambda)^2 G^2_{\text{client}}(\boldsymbol{\gamma}, \boldsymbol{\xi}) + \lambda^2 G^2_{\text{class}}(\boldsymbol{\alpha}, \boldsymbol{\beta})$. Let $\bar{G}^2$, $\bar{G}^2_{\text{client}}$, and $\bar{G}^2_{\text{class}}$ denote the corresponding maximum values across updating rounds $t = 0, 1, \ldots, T-1$. Then, $\bar{G}^2 \leq (1-\lambda)^2 \bar{G}^2_{\text{client}} + \lambda^2 \bar{G}^2_{\text{class}}$. With this notation in place, we state the following result:

Theorem 2.5. *Assume ℓ^{ρ} is μ-strongly convex and L-smooth. Suppose $\eta \leq 1/L$ and furthermore $\eta L E \leq 1/4$. Let $\boldsymbol{\zeta}^{(t)}$ be the output after t communication rounds. Then as $T, N \to \infty$ we have*

$$\|\boldsymbol{\zeta}^{(T)} - \boldsymbol{\zeta}^{true}\|_2^2 = O_p\left(\frac{\{(1-\lambda)\|\mathcal{I}_\gamma\|_{\min} + \rho\}^{-1} + \{\lambda\|\mathcal{I}_\beta\|_{\min} + \rho\}^{-1}}{N} + \frac{\eta^2 E^2 \{(1-\lambda)^2 \bar{G}^2_{client} + \lambda^2 \bar{G}^2_{class}\}}{1 - (1-\eta\mu)^E}\right)$$

where $\|A\|_{\min} = \lambda_{\min}(A)$, and $\mathcal{I}_{client}$ and $\mathcal{I}_{class}$ denote the Fisher information matrices with respect to $(\boldsymbol{\gamma}, \boldsymbol{\xi})$ and $(\boldsymbol{\alpha}, \boldsymbol{\beta})$, respectively.

The proof and the detailed definition of Fisher information matrix is deferred to App. E. The first term in the bound capture the statistical accuracy, while the last term reflects the convergence rate. For faster convergence, a larger λ is preferred, while for higher accuracy, λ must be chosen to balance $\{(1-\lambda)\|\mathcal{I}_\gamma\|_{\min} + \rho\}^{-1}$ and $\{\lambda\|\mathcal{I}_\beta\|_{\min} + \rho\}^{-1}$. Together, these terms reveal the trade-off role of λ. In practice, since the Fisher information matrices and gradient drifts are unknown, λ can be tuned using a validation set. We empirically demonstrate the trade-off in Fig. 4.

3 Experiments on Benchmark Datasets

3.1 Experiment Settings

Datasets. We conduct experiments on CIFAR-10 and CIFAR-100 (Krizhevsky, 2009), each containing $60,000$ 32×32 RGB images. CIFAR-10 has 10 classes with $6,000$ images per class. CIFAR-100 has 100 classes, with 600 images per class, grouped into 20 superclasses. Based on these datasets, we construct three tasks of increasing complexity: (a) 10-class classification on CIFAR-10, (b) 20-class classification using the CIFAR-100 superclasses, and (c) 100-class classification using the fine-grained CIFAR-100 labels.

Non-IID settings. Since standard benchmark datasets do not inherently exhibit statistical heterogeneity, we simulate non-IID scenarios following common practice (Wu et al., 2023; Tan et al., 2023; Lu et al., 2024). We introduce both label and covariate shifts. For **label shift**, we construct client datasets using two partitioning strategies: (1) *Dirichlet partition with $\alpha = 0.3$ (Dir-0.3)*: Following (Yurochkin et al., 2019), we draw class proportions for each client from a Dirichlet distribution with concentration parameter $\alpha = 0.3$, leading to heterogeneous label marginals and unequal dataset sizes across clients. (2) *S shards per client (S-SPC)*: Following (McMahan et al., 2017), we sort the data by class, split it into equal-sized, label-homogeneous shards, and assign S shards uniformly at random to each client. This yields equal dataset sizes while restricting each client's label support to at most S classes. Each dataset is first partitioned across clients using one of the partitioning strategies, and within each client, the local dataset is further split $70/30$ into training and test sets. For **covariate shift**, all three nonlinear transformations are applied to each client's dataset: (1) *gamma correction*: brightness adjustment with client-specific gamma factor γ. (2) *hue adjustment*: color rotation with client-specific hue factor Δh. (3) *saturation scaling*: color vividness adjustment with client-specific saturation factor κ. We set $\gamma \in \{0.6, 1.4\}$, $\Delta h \in \{-0.1, 0.1\}$, and $\kappa \in \{0.5, 1.5\}$ in the main experiment, resulting in an 8-client setting. See examples in App. F.1.

Baselines. We compare our `FedDRM` against a variety of state-of-the-art personalized FL techniques, which learn a local model on each client. Ditto (Li et al., 2021c) encourages local models to stay close via global regularization. FedRep (Collins et al., 2021) learns a global backbone with local linear heads. FedBABU (Oh et al., 2022) freezes local classifiers while training a global backbone, then fine-tunes classifiers per client. FedPAC (Xu et al., 2023) personalizes through feature alignment to a global backbone. FedALA (Zhang et al., 2023) learns client-wise mixing weights that adaptively interpolate between the local and global models. FedAS (Yang et al., 2024) aligns local weights to the global model, followed by client-specific updates. ConFREE (Zheng et al., 2025) resolves conflicts among client updates before server aggregation. We also compare with other standard FL algorithms– FedAvg (McMahan et al., 2017), FedProx (Li et al., 2020b), and FedSAM (Qu et al., 2022)–which aim to achieve a single global model under data heterogeneity. To ensure fair comparison, we fine-tune their global models locally on each client, yielding personalized variants denoted FedAvgFT, FedProxFT, and FedSAMFT.

Network architecture. We use ResNet-18 (He et al., 2016) as the feature extractor (backbone), which encodes each input image into a 512-dimensional embedding. For the baselines, this embedding is projected to 256 dimensions via a linear layer and fed into the image classifier. `FedDRM` extends this design by adding a separate client-classification head: the 512-dimensional embedding is projected to 256 dimensions and fed into the client classifier. Importantly, `FedDRM` uses the same image classification architecture as all baselines.

Training details. To ensure fair comparison, all methods are trained for 800 communication rounds with 10 local steps per round and a batch size of 128. For fine-tuning-based methods, we allocate 700 rounds for global training and 100 rounds for local fine-tuning. We use SGD with momentum 0.9, an initial learning rate of 0.01 with cosine annealing, and weight decay 5×10^{-4}. Method-specific hyperparameters are tuned to achieve their best performance.

3.2 Evaluation Protocol

To assess the effectiveness of our proposed method in guiding clients under heterogeneous FL, we introduce a new performance metric, termed **system accuracy**. This metric is designed to evaluate the server's ability to guide clients effectively. Concretely, we construct a pooled test set from all clients. For `FedDRM`, we first use the client classification head to identify the most likely client for each test sample by maximizing the client classification probability. The local model of the selected client is then used to predict the image class label. For baseline methods, which lack this client-guidance mechanism, we instead apply a majority-voting strategy: each client's personalized model makes a prediction for every sample in the pooled test set, and the majority label is taken as the final prediction. The overall classification accuracy on the pooled test set is reported as the system accuracy. We also report the widely used **average accuracy** in personalized FL, which measures each local model's classification accuracy on its own test set. The final value is computed as the weighted average across all clients, with weights proportional to the size of each client's training set. In all experiments, we report the mean and standard deviation of both average accuracy and system accuracy over the final 50 communication rounds.

3.3 Main Results

We present the system accuracy and average accuracy in Tab. 1 and Tab. 2, respectively. Across all settings, `FedDRM` consistently outperforms the baselines on both metrics, demonstrating its ability to leverage statistical heterogeneity for system-level intelligence while also providing effective client-level personalization. In contrast, the baselines primarily focus on addressing data heterogeneity, resulting in lower system accuracy due to disagreements among their personalized models. Additionally, when using a majority-vote approach as an intelligence router, baseline methods must evaluate all m local models, whereas `FedDRM` requires evaluating only a single model. The shared backbone in `FedDRM` can also be efficiently repurposed for image prediction by feeding it into the corresponding client-specific classification head. We also compare the influence of label shift in this experiment beyond covariate shift, the results align with our expectation that the less severe label shift Dir-0.3 case has a higher accuracy than 5-SPC for all methods.

Table 1: System accuracy on CIFAR-10/20/100 under Dir-0.3 and 5/25-SPC settings.

Method	CIFAR-10		CIFAR-20		CIFAR-100	
	Dir-0.3	5-SPC	Dir-0.3	25-SPC	Dir-0.3	25-SPC
Ditto	47.64 ± 0.25	46.99 ± 0.23	29.56 ± 0.18	31.87 ± 0.16	15.97 ± 0.15	19.51 ± 0.16
FedRep	24.96 ± 0.19	33.19 ± 0.22	23.83 ± 0.15	24.82 ± 0.20	11.11 ± 0.12	12.02 ± 0.12
FedBABU	57.43 ± 0.17	57.17 ± 0.24	36.96 ± 0.17	40.78 ± 0.13	22.92 ± 0.17	27.27 ± 0.15
FedPAC	25.14 ± 0.21	33.24 ± 0.19	23.83 ± 0.17	24.83 ± 0.18	11.17 ± 0.12	11.99 ± 0.12
FedALA	61.33 ± 0.17	53.20 ± 0.20	32.78 ± 0.17	35.79 ± 0.14	20.70 ± 0.14	25.80 ± 0.16
FedAS	28.76 ± 0.19	39.71 ± 0.20	27.16 ± 0.16	27.50 ± 0.15	13.87 ± 0.13	13.51 ± 0.14
ConFREE	25.66 ± 0.22	34.06 ± 0.22	24.08 ± 0.17	25.15 ± 0.18	11.32 ± 0.13	12.12 ± 0.13
FedAvgFT	54.90 ± 0.22	56.19 ± 0.17	37.53 ± 0.18	41.17 ± 0.16	25.21 ± 0.16	27.96 ± 0.17
FedProxFT	55.01 ± 0.20	56.27 ± 0.21	37.61 ± 0.18	41.20 ± 0.15	25.15 ± 0.13	27.82 ± 0.18
FedSAMFT	55.83 ± 0.21	51.73 ± 0.19	34.23 ± 0.14	36.60 ± 0.17	22.97 ± 0.16	26.89 ± 0.15
`FedDRM`	$\mathbf{63.85 \pm 0.18}$	$\mathbf{58.50 \pm 0.23}$	$\mathbf{37.67 \pm 0.22}$	$\mathbf{41.44 \pm 0.19}$	$\mathbf{26.01 \pm 0.16}$	$\mathbf{31.24 \pm 0.17}$

Table 2: Average accuracy on CIFAR-10/20/100 under Dir-0.3 and 5/25-SPC settings.

Method	CIFAR-10		CIFAR-20		CIFAR-100	
	Dir-0.3	5-SPC	Dir-0.3	25-SPC	Dir-0.3	25-SPC
Ditto	76.34 ± 0.11	65.17 ± 0.17	40.36 ± 0.18	44.83 ± 0.19	29.25 ± 0.16	36.58 ± 0.18
FedRep	76.49 ± 0.15	64.96 ± 0.19	41.54 ± 0.16	46.57 ± 0.19	31.22 ± 0.15	39.11 ± 0.20
FedBABU	78.22 ± 0.14	70.22 ± 0.18	44.18 ± 0.15	48.98 ± 0.19	32.91 ± 0.14	40.75 ± 0.14
FedPAC	76.53 ± 0.13	65.05 ± 0.19	41.60 ± 0.16	46.55 ± 0.19	31.20 ± 0.17	39.13 ± 0.22
FedALA	64.35 ± 2.40	55.58 ± 1.88	33.30 ± 0.47	36.41 ± 0.58	21.83 ± 0.94	27.83 ± 1.59
FedAS	78.69 ± 0.17	69.82 ± 0.16	45.65 ± 0.18	51.73 ± 0.17	36.06 ± 0.13	44.26 ± 0.19
ConFREE	76.73 ± 0.16	65.59 ± 0.17	41.91 ± 0.16	47.04 ± 0.21	31.57 ± 0.15	39.63 ± 0.17
FedAvgFT	79.08 ± 0.11	72.10 ± 0.18	46.55 ± 0.15	52.54 ± 0.17	36.83 ± 0.17	43.94 ± 0.20
FedProxFT	79.07 ± 0.12	72.07 ± 0.18	46.58 ± 0.19	52.49 ± 0.18	36.87 ± 0.18	43.96 ± 0.19
FedSAMFT	75.53 ± 0.11	66.30 ± 0.16	41.11 ± 0.16	44.89 ± 0.16	32.53 ± 0.14	40.25 ± 0.16
FedDRM	$\mathbf{80.25 \pm 0.14}$	$\mathbf{72.50 \pm 0.16}$	$\mathbf{47.91 \pm 0.18}$	$\mathbf{53.72 \pm 0.20}$	$\mathbf{37.91 \pm 0.15}$	$\mathbf{46.73 \pm 0.15}$

3.4 Sensitivity Analysis

We evaluate the sensitivity of our method to several key factors. Experimental details are reported in App. F.2.

Impact of weight λ on system accuracy. The reweighting parameter λ is crucial for deploying the EL-based framework in the FL setting. As shown in Fig. 4, we observe the expected trade-off between two objectives: increasing λ places more emphasis on image classification and less on client classification. This shift improves overall accuracy but reduces client accuracy, consistent with Thm. 2.5. The best balance between the two is achieved at $\lambda = 0.8$, where system accuracy peaks, marking the optimal trade-off for the task of guiding queries in the FL system.

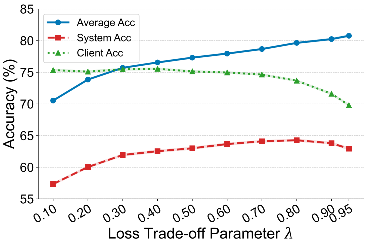


Figure 4: **Client & image accuracy trade-off** on CIFAR-10 under the Dir-0.3 setting.

Covariate shift intensity. We have already demonstrated in the main results that label shift is detrimental to all methods, with more severe shifts causing greater harm. To further examine the impact of covariate shift, we fix the degree of label shift and vary covariate shift at three intensity levels—low, mid, and high—by adjusting the parameters of the nonlinear color transformations. As shown in Fig. 5, the results reveal a clear trade-off: higher covariate shift intensifies differences between client data distributions, which facilitates client routing but simultaneously weakens information sharing across clients, thereby making image classification more difficult. Additional results examining the sensitivity of our method to the severity of label shift are provided in App. F.2.4.

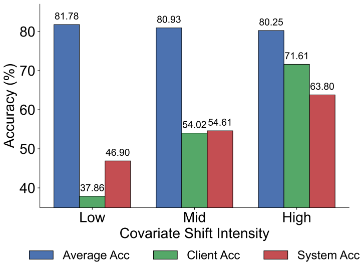


Figure 5: **Influence of covariate shift intensity** on CIFAR-10 under the Dir-0.3 setting.

Backbone sharing strategy. In our formulation, the target-class classification task uses the embedding $g_\theta(x)$ for an input feature x, while the client-classification task uses the embedding $h_\tau(g_\theta(x))$ for the same feature. Since both g_θ and h_τ are parameterized functions, the optimal sharing strategy between the two is not obvious. To explore this, we evaluate four cases: no sharing, shallow sharing, mid sharing, and deep sharing. As shown in Fig. 6, all strategies perform similarly, with shallow sharing slightly ahead. However, given the substantial increase in parameters for shallow sharing, deep sharing offers a more parameter-efficient alternative while maintaining strong performance.

Number of clients. To check scalability, we set the number of clients from 8 to 32 and compare FedDRM against the top-2 baselines from the main experiments. As shown in Tab. 3, while all methods exhibit a moderate performance decline as the client pool expands (a common challenge in FL), FedDRM consistently maintains a significant performance advantage across both system and average accuracy. This demonstrates that our method scales effectively, preserving its superiority even as the system grows.

Table 3: Sensitivity analysis on the number of clients m.

Method	System Accuracy				Average Accuracy			
	$m = 8$	$m = 16$	$m = 24$	$m = 32$	$m = 8$	$m = 16$	$m = 24$	$m = 32$
FedAS	32.58 ± 0.19	36.55 ± 0.21	38.12 ± 0.21	34.50 ± 0.18	78.86 ± 0.12	73.28 ± 0.15	73.90 ± 0.16	73.45 ± 0.15
FedAvgFT	53.17 ± 0.22	50.61 ± 0.20	49.13 ± 0.20	45.07 ± 0.21	78.92 ± 0.15	73.41 ± 0.16	74.66 ± 0.17	74.18 ± 0.15
FedDRM	$\mathbf{59.59 \pm 0.20}$	$\mathbf{51.61 \pm 0.22}$	$\mathbf{50.18 \pm 0.20}$	$\mathbf{46.62 \pm 0.17}$	$\mathbf{80.47 \pm 0.12}$	$\mathbf{74.25 \pm 0.15}$	$\mathbf{75.04 \pm 0.14}$	$\mathbf{74.45 \pm 0.18}$

4 Experiment on Real Medical Dataset

To further demonstrate FedDRM's effectiveness in healthcare, we evaluate it on the real medical dataset RETINA, following Huang et al. (2025). RETINA comprises fundus images from three clinical centers—ACRIMA (Diaz-Pinto et al., 2019), Rim (Fumero Batista et al., 2020), and Refuge (Orlando et al., 2020). We exclude Drishti, which has only 82 images, while the others provide at least 385. Each 96 × 96 RGB image is labeled as Glaucomatous or Normal, creating a binary classification task.

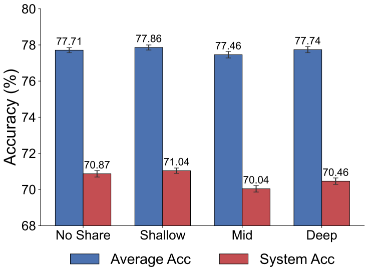


Figure 6: **Impact of the sharing strategy** on CIFAR-10 under the Dir-0.3 setting using LeNet (Lecun et al., 1998).

This dataset naturally fits a 3-client FL system, with each client representing one center. The different image sources cause a covariate shift in RETINA. Furthermore, the class ratios (positive vs. negative) across the three datasets are 1.34, 1.94, and 0.46, introducing a realistic label shift. Our experimental setup largely follows the CIFAR experiments, with several adjustments: the network embedding dimension is set to 4608 and the projection dimension 512. All methods train for 100 communication rounds with a batch size of 32. For fine-tuning-based methods, we allocate 90 rounds to global training and 10 to local fine-tuning.

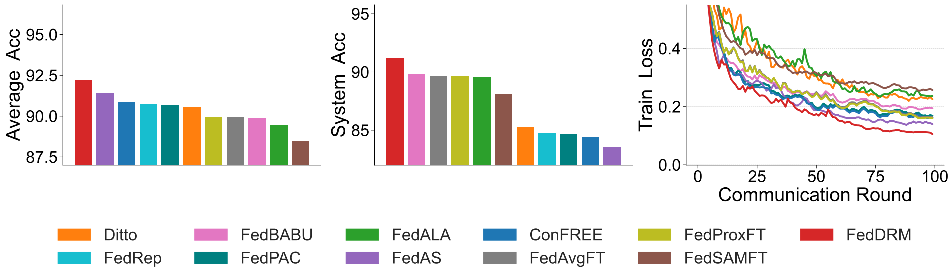


Figure 7: Average accuracy, system accuracy, and train loss on RETINA.

Fig. 7 shows that FedDRM consistently outperforms all baselines on RETINA. Measured in absolute accuracy points, FedDRM exceeds the competing methods by 0.83–3.77 points in average accuracy and by 1.41–7.67 points in system accuracy—substantial margins given the small size and pronounced heterogeneity of this dataset. These results underscore the robustness and practical relevance of FedDRM in the presence of simultaneous covariate and label shifts. Furthermore, FedDRM achieves the lowest training loss and the most stable convergence trajectory, demonstrating its effectiveness in capturing heterogeneous structure in real multi-center medical data.

5 Conclusion

This paper presents FedDRM, a novel FL paradigm that transforms statistical heterogeneity from a challenge into a resource. By introducing a unified EL based framework, FedDRM simultaneously learns accurate local models and a client-selection policy, enabling a central server to intelligently route new queries to the most appropriate client. Empirical results demonstrate that our method outperforms existing approaches in both client-level personalization and system-level utility, paving the way for more adaptive and resource-efficient FL systems that actively leverage statistical diversity. We believe that this work marks a meaningful step toward more adaptive, resource-efficient, and intelligent FL systems.

REFERENCES

Durmus Alp Emre Acar, Yue Zhao, Ramon Matas, Matthew Mattina, Paul Whatmough, and Venkatesh Saligrama. Federated learning based on dynamic regularization. In *International Conference on Learning Representations*, 2021.

J. A. Anderson. Multivariate logistic compounds. *Biometrika*, 66(1):17–26, 1979.

Christopher Briggs, Zhong Fan, and Peter Andras. Federated learning with hierarchical clustering of local updates to improve training on non-iid data. In *International Joint Conference on Neural Networks*, 2020.

Jiahua Chen and Yukun Liu. Quantile and quantile-function estimations under density ratio model. *The Annals of Statistics*, 41(3):1669–1692, 2013.

Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *International Conference on Machine Learning*, 2018.

Liam Collins, Hamed Hassani, Aryan Mokhtari, and Sanjay Shakkottai. Exploiting shared representations for personalized federated learning. In *International Conference on Machine Learning*, 2021.

Andres Diaz-Pinto, Sandra Morales, Valery Naranjo, Thomas Köhler, Jose M Mossi, and Amparo Navea. Cnns for automatic glaucoma assessment using fundus images: an extensive validation. *Biomedical Engineering Online*, 18(1):1–19, 2019.

Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach. In *Advances in Neural Information Processing Systems*, 2020.

Konstantinos Fokianos, Benjamin Kedem, Jing Qin, and David A Short. A semiparametric approach to the one-way layout. *Technometrics*, 43(1):56–65, 2001.

Francisco José Fumero Batista, Tinguaro Diaz-Aleman, Jose Sigut, Silvia Alayon, Rafael Arnay, and Denisse Angel-Pereira. RIM-ONE DL: A unified retinal image database for assessing glaucoma using deep learning. *Image Analysis & Stereology*, 39(3):161–167, 2020.

Liang Gao, Huazhu Fu, Li Li, Yingwen Chen, Ming Xu, and Cheng-Zhong Xu. FedDC: Federated learning with non-IID data via local drift decoupling and correction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Avishek Ghosh, Jichan Chung, Dong Yin, and Kannan Ramchandran. An efficient framework for clustered federated learning. *Advances in Neural Information Processing Systems*, 2020.

Yongxin Guo, Xiaoying Tang, and Tao Lin. FedBR: Improving federated learning on heterogeneous data via local learning bias reduction. In *International Conference on Machine Learning*, 2023.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016.

Chun-Yin Huang, Ruinan Jin, Can Zhao, Daguang Xu, and Xiaoxiao Li. Federated learning on virtual heterogeneous data with local-global dataset distillation. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning*, 2020.

Richard Kay and Sarah Little. Transformations of the explanatory variables in the logistic regression model for binary data. *Biometrika*, 74(3):495–501, 1987.

Yeongwoo Kim, Ezeddin Al Hakim, Johan Haraldson, Henrik Eriksson, José Mairton B da Silva, and Carlo Fischione. Dynamic clustering in federated learning. In *International Conference on Communications*, 2021.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Chengxi Li, Gang Li, and Pramod K Varshney. Federated learning with soft clustering. *IEEE Internet of Things Journal*, 9(10):7773–7782, 2021a.

Li Li, Yuxi Fan, Mike Tse, and Kuo-Yi Lin. A review of applications in federated learning. *Computers & Industrial Engineering*, 149:106854, 2020a.

Pengfei Li, Yukun Liu, and Jing Qin. Semiparametric inference in a genetic mixture model. *Journal of the American Statistical Association*, 112(519):1250–1260, 2017.

Qinbin Li, Bingsheng He, and Dawn Song. Model-contrastive federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021b.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. In *Proceedings of Machine Learning and Systems*, 2020b.

Tian Li, Shengyuan Hu, Ahmad Beirami, and Virginia Smith. Ditto: Fair and robust federated learning through personalization. In *International Conference on Machine Learning*, 2021c.

Xiaoxiao Li, Meirui Jiang, Xiaofei Zhang, Michael Kamp, and Qi Dou. FedBN: Federated learning on non-IID features via local batch normalization. In *International Conference on Learning Representations*, 2021d.

Bo Liu, Xingchao Liu, Xiaojie Jin, Peter Stone, and Qiang Liu. Conflict-averse gradient descent for multi-task learning. In *Advances in Neural Information Processing Systems*, 2021.

Siyan Liu, Chi-Kuang Yeh, Xin Zhang, Qinglong Tian, and Pengfei Li. Positive and unlabeled data: Model, estimation, inference, and classification. *Journal of the American Statistical Association*, pp. 1–12, 2025.

Yukun Liu, Pengfei Li, and Jing Qin. Maximum empirical likelihood estimation for abundance in a closed population from capture-recapture data. *Biometrika*, 104(3):527–543, 2017.

Guodong Long, Yue Tan, Jing Jiang, and Chengqi Zhang. Federated learning for open banking. In *Federated learning: privacy and incentive*, pp. 240–254. Springer, 2020.

Guodong Long, Ming Xie, Tao Shen, Tianyi Zhou, Xianzhi Wang, and Jing Jiang. Multi-center federated learning: clients clustering for better personalization. *World Wide Web*, 26(1):481–500, 2023.

Yang Lu, Lin Chen, Yonggang Zhang, Yiliang Zhang, Bo Han, Yiu-ming Cheung, and Hanzi Wang. Federated learning with extremely noisy clients via negative distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.

Xiaosong Ma, Jie Zhang, Song Guo, and Wenchao Xu. Layer-wised model aggregation for personalized federated learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *International Conference on Artificial Intelligence and Statistics*, 2017.

Jaehoon Oh, Sangmook Kim, and Se-Young Yun. Fedbabu: Toward enhanced representation for federated image classification. In *International Conference on Learning Representations*, 2022.

José Ignacio Orlando, Huazhu Fu, João Barbosa Breda, Karel van Keer, Deepti R Bathula, Andrés Diaz-Pinto, Ruogu Fang, Pheng-Ann Heng, Jeyoung Kim, JoonHo Lee, et al. REFUGE challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. *Medical Image Analysis*, 59:101570, 2020.

Art B Owen. *Empirical likelihood*. Chapman and Hall/CRC, 2001.

Jing Qin and Biao Zhang. A goodness-of-fit test for logistic regression models based on case-control data. *Biometrika*, 84(3):609–618, 1997.

Zhe Qu, Xingyu Li, Rui Duan, Yao Liu, Bo Tang, and Zhuo Lu. Generalized federated learning via sharpness aware minimization. In *International Conference on Machine Learning*, 2022.

Virginia Smith, Chao-Kai Chiang, Maziar Sanjabi, and Ameet S Talwalkar. Federated multi-task learning. In *Advances in Neural Information Processing Systems*, 2017.

Canh T Dinh, Nguyen Tran, and Josh Nguyen. Personalized federated learning with moreau envelopes. *Advances in Neural Information Processing Systems*, 2020.

Alysa Ziying Tan, Han Yu, Lizhen Cui, and Qiang Yang. Towards personalized federated learning. *IEEE Transactions on Neural Networks and Learning Systems*, 34(12):9587–9603, 2022.

Yue Tan, Chen Chen, Weiming Zhuang, Xin Dong, Lingjuan Lyu, and Guodong Long. Is heterogeneity notorious? taming heterogeneity to handle test-time shift in federated learning. In *Advances in Neural Information Processing Systems*, 2023.

Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Jianyu Wang, Qinghua Liu, Hao Liang, Gauri Joshi, and H Vincent Poor. Tackling the objective inconsistency problem in heterogeneous federated optimization. In *Advances in Neural Information Processing Systems*, 2020.

Yue Wu, Shuaicheng Zhang, Wenchao Yu, Yanchi Liu, Quanquan Gu, Dawei Zhou, Haifeng Chen, and Wei Cheng. Personalized federated learning under mixture of distributions. In *International Conference on Machine Learning*, 2023.

Jian Xu, Xinyi Tong, and Shao-Lun Huang. Personalized federated learning with feature alignment and classifier collaboration. In *International Conference on Learning Representations*, 2023.

Jie Xu, Benjamin S Glicksberg, Chang Su, Peter Walker, Jiang Bian, and Fei Wang. Federated learning for healthcare informatics. *Journal of Healthcare Informatics Research*, 5(1):1–19, 2021.

Xiyuan Yang, Wenke Huang, and Mang Ye. Fedas: Bridging inconsistency in personalized federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Qiaoyun Yin, Zhiyong Feng, Xiaohong Li, Shizhan Chen, Hongyue Wu, and Gaoyong Han. Tackling data-heterogeneity variations in federated learning via adaptive aggregate weights. *Knowledge-Based Systems*, 304:112484, 2024.

Mikhail Yurochkin, Mayank Agarwal, Soumya Ghosh, Kristjan Greenewald, Nghia Hoang, and Yasaman Khazaeni. Bayesian nonparametric federated learning of neural networks. In *International Conference on Machine Learning*, 2019.

Jianqing Zhang, Yang Hua, Hao Wang, Tao Song, Zhengui Xue, Ruhui Ma, and Haibing Guan. Fedala: Adaptive local aggregation for personalized federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023.

Hao Zheng, Zhigang Hu, Liu Yang, Meiguang Zheng, Aikun Xu, and Boyu Wang. Confree: Conflict-free client update aggregation for personalized federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.

A The Use of Large Language Models (LLMs)

Large language models (LLMs) were used solely as assistive tools for language editing and polishing of the manuscript. The authors take full responsibility for the accuracy and integrity of the manuscript.

B Density ratio model examples

Many parametric distribution families including normal and Gamma are special cases of the DRM.
Example B.1 (Normal distribution)**.** For normal distribution $\phi(x;\mu,\sigma^2)$ with mean μ and variance σ^2. We have $\log\{\phi(x;\mu_1,\sigma_1^2)/\phi(x;\mu_2,\sigma_2^2)\} = \theta_0 + \theta_1 x + \theta_2 x^2$ where $\theta_0 = \log\sigma_2/\sigma_1 - (\mu_1^2/\sigma_1^2 - \mu_2^2/\sigma_2^2)/2$, $\theta_1 = \mu_1/\sigma_1^2 - \mu_2/\sigma_2^2$, $\theta_2 = (\sigma_2^{-2} - \sigma_1^{-2})/2$ and the basis function is $g(x) = (1, x, x^2)^\top$.
Example B.2 (Gamma distribution)**.** For gamma distribution with shape parameter $\alpha > 0$ and rate parameter $\beta > 0$. We have $\log\{f(x;\alpha_1,\beta_1)/f(x;\alpha_2,\beta_2)\} = \theta_0 + \theta_1 x + \theta_2 \log x$ where $\theta_0 = \log\Gamma(\alpha_2) - \log\Gamma(\alpha_1) + \alpha_1\log\beta_1 - \alpha_2\log\beta_2$, $\theta_1 = \beta_2 - \beta_1$, $\theta_2 = \alpha_1 - \alpha_2$ and the basis function is $g(x) = (1, x, \log x)^\top$.

C Mathematical details behind FedDRM

C.1 Derivation of (4)

Proof. By the total law of probability, the marginal density of x is

$$\begin{aligned}
p_l(x) &= \sum_k \pi_{lk} dF_l(x|y=k) \\
&= \sum_k \pi_{lk} \exp\{\alpha_{lk}^\dagger + \beta_k^\top g_\theta(x)\} dF_l(x|y=1) \\
&= \sum_k \pi_{lk} \exp\{\alpha_{lk}^\dagger + \beta_k^\top g_\theta(x)\} \exp\{\gamma_l + \xi_l^\top h_\eta(g_\theta(x))\} dF_0(x|y=1) \\
&= \sum_k \pi_{l1} \exp\{\alpha_k + \beta_k^\top g_\theta(x)\} \exp\{\gamma_l + \xi_l^\top h_\eta(g_\theta(x))\} dF_0(x|y=1) \\
&= \sum_k \frac{\pi_{l1}}{\pi_{01}} \pi_{0k} \exp\{\alpha_{0k}^\dagger + \beta_k^\top g_\theta(x)\} \exp\{\gamma_l + \xi_l^\top h_\eta(g_\theta(x))\} dF_0(x|y=1) \\
&= \frac{\pi_{l1}}{\pi_{01}} \exp\{\gamma_l + \xi_l^\top h_\eta(g_\theta(x))\} p_0(x)
\end{aligned}$$

Let $\gamma_l^\dagger = \gamma_l + \log(\pi_{l1}/\pi_{01})$ and divide $p_0(x)$ on both sides completes the proof. □

C.2 Derivation of the profile log-likelihood

Let $\boldsymbol{p} = \{p_{ij}, j \in [n_i]\}_{i=1}^m$. Given $\boldsymbol{\zeta} = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\xi}, \theta, \eta)$, the empirical log-likelihood function as a function of $\boldsymbol{p}$ becomes

$$\ell_N(\boldsymbol{p}) = \sum_{i=1}^m \sum_{j=1}^{n_i} \log p_{ij} + \text{constant}$$

where the constant depends only on $\boldsymbol{\zeta}$ and does not depend on $\boldsymbol{p}$. We now maximize the empirical log-likelihood function with respect to $\boldsymbol{p}$ under the constraint (5) using the Lagrange multiplier method.

Let

$$\mathcal{L} = \sum_{i,j} \log p_{ij} - N\mu \sum_{i,j} p_{ij} - N \sum_{l=1}^m \rho_l \sum_{i,j} p_{ij} [\exp\{\gamma_l^\dagger + \xi_l^\top h_\eta(g_\theta(x_{ij}))\} - 1]$$

Setting

$$0 = \frac{\partial \mathcal{L}}{\partial p_{ij}} = \frac{1}{p_{ij}} - N\mu - N \sum_{l=1}^m \rho_l [\exp\{\gamma_l^\dagger + \xi_l^\top h_\eta(g_\theta(x_{ij}))\} - 1].$$

Then multiply both sides by p_{ij} and sum over i and j, we have that

$$
\begin{aligned}
0 = \sum_{i,j} p_{ij} \frac{\partial \mathcal{L}}{\partial p_{ij}} &= \sum_{i,j} \left\{ 1 - N\mu p_{ij} - \sum_{l=1}^{m} \rho_l p_{ij} [\exp\{\gamma_l^\dagger + \xi_l^\top h_\eta(g_\theta(x_{ij}))\} - 1] \right\} \\
&= N - N\mu
\end{aligned}
$$

this gives $\mu = 1$. Hence, we get

$$
p_{ij} = \frac{1}{N \left\{ 1 + \sum_l \rho_l [\exp\{\gamma_l^\dagger + \xi_l^\top h_\eta(g_\theta(x_{ij}))\} - 1] \right\}}
$$

where ρ_ls are solutions to

$$
\sum_{i,j} \frac{\exp\{\gamma_l^\dagger + \xi_l^\top h_\eta(g_\theta(x_{ij}))\} - 1}{1 + \sum_{l'} \rho_{l'} [\exp\{\gamma_{l'}^\dagger + \xi_{l'}^\top h_\eta(g_\theta(x_{ij}))\} - 1]} = 0
$$

by plugin the expression for p_{ij} into the second constraints (5).

C.3 Derivation of the value of Lagrange multiplier at optimal

Recall that the profile log-EL has the following form

$$
\begin{aligned}
p\ell_N(\boldsymbol{\zeta}) &= \sum_{i,j,k} \mathbb{1}(y_{ij} = k) \log \mathbb{P}(y_{ij} = k | x_{ij}) + \sum_{i,j} \{\gamma_i^\dagger + \xi_i^\top h_\eta(g_\theta(x_{ij})) + \log p_{ij}(\boldsymbol{\zeta})\} \\
&= \sum_{i,j,k} \mathbb{1}(y_{ij} = k) \log \mathbb{P}(y_{ij} = k | x_{ij}) + \sum_{i,j} \left\{ \gamma_i^\dagger + \xi_i^\top h_\eta(g_\theta(x_{ij})) \right\} \\
&\quad - \sum_{i,j} \log \left\{ 1 + \sum_{l=1}^{m} \rho_l \left[\exp\{\gamma_l^\dagger + \xi_l^\top h_\eta(g_\theta(x_{ij}))\} - 1 \right] \right\}
\end{aligned}
$$

where ρ_ls are solutions to

$$
\sum_{i,j} \frac{\exp\{\gamma_l^\dagger + \xi_l^\top h_\eta(g_\theta(x_{ij}))\} - 1}{1 + \sum_{l'} \rho_{l'} [\exp\{\gamma_{l'}^\dagger + \xi_{l'}^\top h_\eta(g_\theta(x_{ij}))\} - 1]} = 0.
$$

Taking the partial derivative with respect to $\gamma_l^\dagger$, we have

$$
\begin{aligned}
0 = \frac{\partial p\ell_N}{\partial \gamma_l^\dagger} &= n_l - \sum_{i,j} \frac{\rho_l \exp\{\gamma_l^\dagger + \xi_l^\top h(\boldsymbol{x}_{ij})\}}{1 + \sum_{l'} \rho_{l'} [\exp\{\gamma_{l'}^\dagger + \xi_{l'}^\top h_\eta(g_\theta(x_{ij}))\} - 1]} \\
&\quad + \sum_{i,j} \frac{\sum_{l'} (\partial \rho_{l'} / \partial \gamma_l^\dagger) \left[\exp(\gamma_{l'}^\dagger + \xi_{l'}^\top h_\eta(g_\theta(\boldsymbol{x}_{ij}))) - 1 \right]}{1 + \sum_{l'} \rho_{l'} [\exp\{\gamma_{l'}^\dagger + \xi_{l'}^\top h_\eta(g_\theta(x_{ij}))\} - 1]} \\
&= n_l - N\rho_l \sum_{i,j} p_{ij} \exp\{\gamma_l^\dagger + \xi_l^\top h(\boldsymbol{x}_{ij})\} + N \sum_{l'} \frac{\partial \rho_{l'}}{\partial \gamma_l^\dagger} \sum_{i,j} p_{ij} [\exp\{\gamma_l^\dagger + \xi_l^\top h(\boldsymbol{x}_{ij})\} - 1] \\
&= n_l - N\rho_l.
\end{aligned}
$$

The last inequality is based on the constraint in (5). Hence, we have $\rho_l = n_l/N$ which completes the proof.

C.4 Dual form of the profile log EL

At the optimal value, we have $\rho_l = n_l/N$ with $N = \sum_{l=1}^m n_l$. Plugin this value into the profile log-EL, we then get

$$\begin{aligned}
p\ell_N(\boldsymbol{\zeta}) &= \sum_{i,j} \log \left\{ \frac{\exp(\alpha_{y_{ij}} + \beta_{y_{ij}}^\top g_\theta(x_{ij}))}{\sum_j \exp(\alpha_j + \beta_j^\top g_\theta(x_{ij}))} \right\} + \sum_{i,j} \log \left\{ \frac{\exp\{\gamma_i^\dagger + \xi_i^\top h_\eta(g_\theta(x_{ij}))\}}{\sum_{l=1}^m \frac{n_l}{N} \exp\{\gamma_l^\dagger + \xi_l^\top h_\eta(g_\theta(x_{ij}))\}} \right\} \\
&= \sum_{i,j} \log \left\{ \frac{\exp(\alpha_{y_{ij}} + \beta_{y_{ij}}^\top g_\theta(x_{ij}))}{\sum_j \exp(\alpha_j + \beta_j^\top g_\theta(x_{ij}))} \right\} + \sum_{i,j} \log \left\{ \frac{(n_1/n_i) \exp\{\gamma_i^\ddagger + \xi_i^\top h_\eta(g_\theta(x_{ij}))\}}{\sum_{l=1}^m (\frac{n_1}{N}) \exp\{\gamma_l^\ddagger + \xi_l^\top h_\eta(g_\theta(x_{ij}))\}} \right\} \\
&= \sum_{i,j} \log \left\{ \frac{\exp(\alpha_{y_{ij}} + \beta_{y_{ij}}^\top g_\theta(x_{ij}))}{\sum_j \exp(\alpha_j + \beta_j^\top g_\theta(x_{ij}))} \right\} + \sum_{i,j} \log \left\{ \frac{(\exp\{\gamma_i^\ddagger + \xi_i^\top h_\eta(g_\theta(x_{ij}))\}}{\sum_{l=1}^m \exp\{\gamma_l^\ddagger + \xi_l^\top h_\eta(g_\theta(x_{ij}))\}} \right\} \\
&\quad - \sum_{i,j} \log \left(\frac{n_i}{N} \right).
\end{aligned}$$

The last term is an additive constant; the maximization does not depend on its value, which completes the proof.

D Generalization to other types of data heterogeneity

In this section, we detail how our method generalizes to the setting where both $Y|X$ and X differ across clients. Recall that the log empirical likelihood function is

$$\begin{aligned}
\ell_N(\boldsymbol{p}, \boldsymbol{\zeta}) &= \sum_{i=1}^m \sum_{j=1}^{n_i} \log P_{X,Y}^{(i)}(\{X_{ij}, Y_{ij}\}) \\
&= \sum_{i,j,k} \mathbb{1}(Y_{ij} = k) \log \mathbb{P}^{(i)}(Y = k|X_{ij}) + \sum_{i,j} \log P_X^{(i)}(\{X_{ij}\})
\end{aligned}$$

We assume that each client has its own linear head for the conditional distribution:

$$\mathbb{P}^{(i)}(Y = k|X = x) = \frac{\exp(\alpha_{ik} + \beta_{ik}^\top g_\theta(x))}{\sum_j \exp(\alpha_{ikj} + \beta_{ij}^\top g_\theta(x))},$$

The marginal distributions $P_X^{(i)}$ are linked as in Theorem 2.1:

$$\frac{dP_X^{(i)}}{dP_X^{(0)}}(x) = \exp\{\gamma_i^\dagger + \xi_i^\top h_\tau(g_\theta(x))\}$$

where $P_X^{(0)}$ is an unspecified reference measure. Using a non-parametric reference distribution, we set

$$p_{ij} = P_X^{(0)}(\{X_{ij}\}) \geq 0, \quad \forall i \in [m],\ j \in [n_i],$$

subject to the constraints

$$\sum_{i=1}^m \sum_{j=1}^{n_i} p_{ij} = 1, \quad \sum_{i=1}^m \sum_{j=1}^{n_i} p_{ij} \ \exp\left\{\gamma_l^\dagger + \xi_l^\top h_\tau(g_\theta(X_{ij}))\right\} = 1, \quad \forall\, l \in [m].$$

Then, the log empirical likelihood across clients is

$$\ell_N(\boldsymbol{p}, \boldsymbol{\zeta}) = \sum_{i,j,k} \mathbb{1}(Y_{ij} = k) \log \mathbb{P}^{(i)}(Y = k|X_{ij}) + \sum_{i,j} \{\gamma_i^\dagger + \xi_i^\top h_\tau(g_\theta(X_{ij})) + \log p_{ij}\}.$$

The profile log-EL of $\boldsymbol{\zeta}$ is defined as

$$p\ell_N(\boldsymbol{\zeta}) = \sup_{\boldsymbol{p}} \ell_N(\boldsymbol{p}, \boldsymbol{\zeta})$$

where the supremum is taken under the constraints above. Applying the method of Lagrange multipliers, we obtain the analytical form

$$p\ell_N(\boldsymbol{\zeta}) = \sum_{i,j,k} \mathbb{1}(Y_{ij} = k) \log \mathbb{P}^{(i)}(Y_{ij} = k|X_{ij}) + \sum_{i,j} \{\gamma_i^\dagger + \xi_i^\top h_\tau(g_\theta(x_{ij})) + \log p_{ij}(\boldsymbol{\zeta})\}$$

where

$$p_{ij}(\boldsymbol{\zeta}) = N^{-1}\Big\{1 + \sum_{l=1}^{m} \rho_l \left[\exp\{\gamma_l^\dagger + \xi_l^\top h_\tau(g_\theta(x_{ij}))\} - 1\right]\Big\}^{-1}$$

and the Lagrange multipliers $\{\rho_l\}_{l=1}^m$ solves

$$\sum_{i,j} \frac{\exp\{\gamma_l^\dagger + \xi_l^\top h_\tau(g_\theta(x_{ij}))\} - 1}{\sum_{l'} \rho_{l'} \left[\exp(\gamma_{l'}^\dagger + \xi_{l'}^\top h_\tau(g_\theta(x_{ij}))) - 1\right]} = 0.$$

Using the dual argument from Appendix C.3, the profile log-EL can be rewritten as

$$\begin{aligned}
p\ell_N(\boldsymbol{\zeta}) &= \sum_{i,j} \log\left\{\frac{\exp(\alpha_{i,y_{ij}} + \beta_{i,y_{ij}}^\top g_\theta(x_{ij}))}{\sum_j \exp(\alpha_{ij} + \beta_{ij}^\top g_\theta(x_{ij}))}\right\} + \sum_{i,j} \log\left\{\frac{\exp\{\gamma_i^\dagger + \xi_i^\top h_\eta(g_\theta(x_{ij}))\}}{\sum_{l=1}^m \frac{n_l}{N} \exp\{\gamma_l^\dagger + \xi_l^\top h_\eta(g_\theta(x_{ij}))\}}\right\} \\
&= \sum_{i,j} \log\left\{\frac{\exp(\alpha_{i,y_{ij}} + \beta_{i,y_{ij}}^\top g_\theta(x_{ij}))}{\sum_j \exp(\alpha_{ij} + \beta_{ij}^\top g_\theta(x_{ij}))}\right\} + \sum_{i,j} \log\left\{\frac{(n_1/n_i) \exp\{\gamma_i^\ddagger + \xi_i^\top h_\eta(g_\theta(x_{ij}))\}}{\sum_{l=1}^m (\frac{n_1}{N}) \exp\{\gamma_l^\ddagger + \xi_l^\top h_\eta(g_\theta(x_{ij}))\}}\right\} \\
&= \sum_{i,j} \log\left\{\frac{\exp(\alpha_{i,y_{ij}} + \beta_{i,y_{ij}}^\top g_\theta(x_{ij}))}{\sum_j \exp(\alpha_{ij} + \beta_{ij}^\top g_\theta(x_{ij}))}\right\} + \sum_{i,j} \log\left\{\frac{(\exp\{\gamma_i^\ddagger + \xi_i^\top h_\eta(g_\theta(x_{ij}))\}}{\sum_{l=1}^m \exp\{\gamma_l^\ddagger + \xi_l^\top h_\eta(g_\theta(x_{ij}))\}}\right\} \\
&\quad - \sum_{i,j} \log\left(\frac{n_i}{N}\right).
\end{aligned}$$

As a result, the loss function remains additive in two cross-entropy terms corresponding to different tasks. The key difference from the covariate shift case is that, for the client-classification task, each client now has its own linear head.

E SYSTEM ACCURACY & CONVERGENCE RATE TRADE-OFF

We show the proof of Theorem 2.5 in this section. To simplify the notation, we consider the following loss: To assure strong convexity, we minimize the following objective function:

$$\ell_i^\rho(\zeta) = \underbrace{\frac{\rho}{2}\|\gamma\|_2^2 - \frac{1-\lambda}{n_i} \sum_{j=1}^{n_i} \log \frac{\exp(\gamma_i^\top z_{ij})}{\sum_{q=1}^m \exp(\gamma_q^\top z_{ij})}}_{=:\ell_i^{\rho,\text{client}}(\gamma)} + \underbrace{\frac{\rho}{2}\|\beta\|_2^2 - \frac{\lambda}{n_i} \sum_{j=1}^{n_i} \log \frac{\exp(\beta_{y_{ij}}^\top x_{ij})}{\sum_{k=1}^K \exp(\beta_k^\top x_{ij})}}_{=:\ell_i^{\rho,\text{class}}(\beta)},$$

Then, where $\zeta = (\gamma, \beta)$ stacks the parameters for client-classification γ and task-classification β. the global objective is $\ell^\rho(\zeta) = m^{-1} \sum_{i=1}^m \ell_i^\rho(\zeta)$.

Problem setting: Assume the data is generated according to the true multinomial logistic model with parameters $\zeta^{\text{true}} = (\gamma^{\text{true}}, \beta^{\text{true}})$. i.e.,

$$\Pr(Y_{\text{client}} = q|z) = \frac{\exp(\gamma_q^{*\top} z)}{\sum_{r=1}^m \exp(\gamma_r^{*\top} z)}, \qquad \Pr(Y_{\text{class}} = k|x) = \frac{\exp(\beta_k^{*\top} x)}{\sum_{r=1}^K \exp(\beta_r^{*\top} x)}.$$

Let $\widehat{\zeta}_N$ denote the minimizer of $\ell^\rho(\zeta)$ with $N = \sum_{i=1}^m n_i$ as the total number of samples.

Total error. Let ζ^T be the output of the algorithm after T steps. We decompose the error using the triangle inequality as follows:

$$\|\zeta^T - \zeta^{\text{true}}\|_2 \leq \underbrace{\|\zeta^T - \widehat{\zeta}_N\|_2}_{\text{optimization error}} + \underbrace{\|\widehat{\zeta}_N - \zeta^{\text{true}}\|_2}_{\text{statistical error}}. \tag{8}$$

We know bound these two terms respectively.

Lemma E.1 (Asymptotic normality). *As $N \to \infty$, the estimator $\widehat{\zeta}_N$ satisfies*

$$\sqrt{N}\,(\widehat{\zeta}_N - \zeta^{true}) \ \xrightarrow{d}\ \mathcal{N}\Big(0,\, \mathcal{I}(\zeta^{true})^{-1}\Big),$$

where the Fisher information is block diagonal:

$$\mathcal{I}(\zeta^{true}) = \begin{bmatrix} (1-\lambda)\,\mathcal{I}_\gamma + \rho I & 0 \\ 0 & \lambda\,\mathcal{I}_\beta + \rho I \end{bmatrix},$$

with

$$\mathcal{I}_\gamma = \mathbb{E}\left\{\left(\operatorname{diag}(p_\gamma(z)) - p_\gamma(z)p_\gamma(z)^\top\right) \otimes (zz^\top)\right\},\ \mathcal{I}_\beta = \mathbb{E}\left\{\left(\operatorname{diag}(p_\beta(x)) - p_\beta(x)p_\beta(x)^\top\right) \otimes (xx^\top)\right\},$$

where $p_\beta(x) = (\exp(\beta_1^\top x)/\sum_j \exp(\beta_j^\top x), \ldots, \exp(\beta_{dim(\beta)}^\top x)/\sum_j \exp(\beta_j^\top x))^\top$, and I is the identity matrix.

Proof. This result follows from the well-established asymptotic properties of maximum likelihood estimators (Van der Vaart, 2000, Section 5.5). □

Statistical error. By Lemma E.1, we have

$$N\|\widehat{\gamma}_N - \gamma^{\text{true}}\|^2 = O_p\Big(\frac{d}{(1-\lambda)\|\mathcal{I}_\gamma\|_{\min} + \rho}\Big), \qquad N\|\widehat{\beta}_N - \beta^{\text{true}}\|^2 = O_p\Big(\frac{p}{\lambda\|\mathcal{I}_\beta\|_{\min} + \rho}\Big), \quad (9)$$

where $\|A\|_{\min} = \lambda_{\min}(A)$ is the operator norm, d and p are dimensions of z and x respectively.

Optimization error. For communication round $t = 0, 1, 2, \ldots, T-1$, the server holds ζ^t and each client i sets $\zeta_{i,0}^t = \zeta^t$ and performs E local gradient steps:

$$\zeta_{i,r+1}^t \ =\ \zeta_{i,r}^t - \eta \nabla \ell_i^\rho(\zeta_{i,r}^t), \qquad r = 0, \ldots, E-1.$$

After E steps each client returns $\zeta_{i,E}^t$ and the server aggregates $\zeta^{t+1} = m^{-1}\sum_{i=1}^m \zeta_{i,E}^t$.

Define $G^2(\zeta) = m^{-1}\sum_{i=1}^m \|\nabla \ell_i^\rho(\zeta) - \nabla \ell^\rho(\zeta)\|_2^2$. It can be decomposed nicely as $G^2(\zeta) = (1-\lambda)^2 G^2_{\text{client}}(\gamma) + \lambda^2 G^2_{\text{class}}(\beta)$. Let $\bar{G}^2$, $\bar{G}^2_{\text{client}}$, and $\bar{G}^2_{\text{class}}$ denote the corresponding maximum values across updating rounds $t = 0, 2, \ldots, T-1$. Then, $\bar{G}^2 \le (1-\lambda)^2\bar{G}^2_{\text{client}} + \lambda^2\bar{G}^2_{\text{class}}$.

In the convergence proof below, we omit the subscript ρ since it does not influence the convergence rate. Because ℓ is μ-strongly convex and L-smooth, a single full-gradient step satisfies

$$\begin{aligned}\|x - \eta\nabla\ell(x) - \widehat{\zeta}_N\|_2^2 =& \|x - \widehat{\zeta}_N\|_2^2 - 2\eta\langle \nabla\ell(x), x - \widehat{\zeta}_N\rangle + \eta^2\|\nabla\ell(x)\|_2^2 \\ \le& \|x - \widehat{\zeta}_N\|_2^2 - 2\eta\mu\|x - \widehat{\zeta}_N\|_2^2 + \eta^2 L^2\|x - \widehat{\zeta}_N\|_2^2 \\ \le& (1-\eta\mu)\|x - \widehat{\zeta}_N\|_2^2.\end{aligned}$$

For client i at local step r:

$$\begin{aligned}\|\zeta_{i,r+1}^t - \widehat{\zeta}_N\|_2^2 =& \|(\zeta_{i,r}^t - \eta\nabla\ell(\zeta_{i,r}^t) - \widehat{\zeta}_N) \ + \ \eta(\nabla\ell(\zeta_{i,r}^t) - \nabla\ell_i(\zeta_{i,r}^t))\|^2 \\ \le& (1-\eta\mu)\|\zeta_{i,r}^t - \widehat{\zeta}_N\|_2^2 + \eta^2\|\nabla\ell(\zeta_{i,r}^t) - \nabla\ell_i(\zeta_{i,r}^t)\|^2\end{aligned}$$

Iterating over E local steps gives

$$\|\zeta_{i,E}^t - \widehat{\zeta}_N\|_2^2 \le (1-\eta\mu)^E\|\zeta^t - \widehat{\zeta}_N\|_2^2 + \eta^2\sum_{r=0}^{E-1}(1-\eta\mu)^{E-1-r}\|\nabla\ell_i(\zeta_{i,r}^t) - \nabla\ell(\zeta_{i,r}^t)\|_2^2.$$

Averaging over $i = 1, \ldots, m$ and using convexity of squared norm:

$$\|\zeta^{t+1} - \widehat{\zeta}_N\|_2^2 \le (1-\eta\mu)^E\|\zeta^t - \widehat{\zeta}_N\|_2^2 + \eta^2\sum_{r=0}^{E-1}\frac{1}{m}\sum_{i=1}^m\|\nabla\ell_i(\zeta_{i,r}^t) - \nabla\ell(\zeta_{i,r}^t)\|_2^2.$$

Using L-smoothness and $\eta LE \leq 1/4$, one can show (via induction on r and triangle inequalities)

$$\frac{1}{m}\sum_{i=1}^{m} \|\nabla \ell_i(\zeta_{i,r}^t) - \nabla \ell(\zeta_{i,r}^t)\|_2^2 \leq \bar{G}^2,$$

where $\bar{G}^2$ is the heterogeneity measure. Summing over $r = 0, \ldots, E-1$ gives

$$\eta^2 \sum_{r=0}^{E-1} \frac{1}{m}\sum_{i=1}^{m} \|\nabla \ell_i(\zeta_{i,r}^t) - \nabla \ell(\zeta_{i,r}^t)\|_2^2 \leq \eta^2 E^2 \bar{G}^2.$$

Combine the above:

$$\|\zeta^{t+1} - \widehat{\zeta}_N\|_2^2 \leq (1-\eta\mu)^E \|\zeta^t - \widehat{\zeta}_N\|_2^2 + \eta^2 E^2 \bar{G}^2.$$

Let $s_t := \|\zeta^t - \widehat{\zeta}_N\|_2^2$ and $\alpha := (1-\eta\mu)^E$, $B := \eta^2 E^2 \bar{G}^2$. Then

$$s_{t+1} \leq \alpha s_t + B \quad \Rightarrow \quad s_T \leq \alpha^T s_0 + B \sum_{j=0}^{T-1} \alpha^j = \alpha^T s_0 + \frac{B(1-\alpha^T)}{1-\alpha} \leq \alpha^T s_0 + \frac{B}{1-\alpha}.$$

This yields the desired bound

$$\|\zeta^T - \widehat{\zeta}_N\|_2^2 \leq (1-\eta\mu)^{ET} \|\zeta^0 - \widehat{\zeta}_N\|_2^2 + \frac{\eta^2 E^2 \bar{G}^2}{1-(1-\eta\mu)^E}. \tag{10}$$

Since $1 - (1-\eta\mu)^E \geq 1 - e^{-\eta\mu E} \geq \frac{1}{2}\min\{1, \eta\mu E\}$, the steady-state error is of order $O(\eta^2 E^2 \bar{G}^2/(\eta\mu E)) = O(\eta E \bar{G}^2/\mu)$, i.e., FedAvg converges linearly to a neighborhood of radius proportional to $\sqrt{\eta E \bar{G}^2/\mu}$.

Combining (9) and (10) with (8) gives the final result that

$$\|\zeta^T - \zeta^{\text{true}}\|^2 = O_p\left(\frac{\{(1-\lambda)\|\mathcal{I}_\gamma\|_{\min} + \rho\}^{-1} + \{\lambda\|\mathcal{I}_\beta\|_{\min} + \rho\}^{-1}}{N} + \frac{\eta^2 E^2 \bar{G}^2}{1-(1-\eta\mu)^E}\right),$$

as both $T, N \to \infty$, This along with $\bar{G}^2 \leq (1-\lambda)^2 \bar{G}^2_{\text{client}} + \lambda^2 \bar{G}^2_{\text{class}}$ completes the proof of the theorem.

F Experiment Details

F.1 Visualization of covariate shift and label shift

In the main experiment, we simulate covariate shift by applying three distinct nonlinear transformations to each client's dataset. Specifically, we use gamma correction with $\gamma \in \{0.6, 1.4\}$, hue adjustment with $\Delta h \in \{-0.1, 0.1\}$, and saturation scaling with $\kappa \in \{0.5, 1.5\}$. This creates $2^3 = 8$ unique combinations of transformations, corresponding to an 8-client setting where each client possesses a visually distinct data distribution. A visualization of a single image sampled from CIFAR-10 after applying these transformations is shown in Fig. 8. As can be clearly seen, the resulting differences in feature distributions across clients are visually striking, highlighting the significant covariate shift simulated in our experiments. We visualize the two types of label shift used in the main experiment in Fig. 9. The figures show the number of samples from each class across 8 clients. As observed, the 5-SPC setting assigns at most 5 classes to each client, whereas the Dir-0.3 setting distributes more classes per client. Thus, the label shift under Dir-0.3 is less severe than under 5-SPC. Our experimental results confirm this observation: all methods achieve higher performance under the less severe Dir-0.3 case.

F.2 Sensitivity Analysis Details

For all subsequent sensitivity analyses, unless otherwise specified, we use CIFAR-10 under the Dir-0.3 setting. The details of each experiment are provided below.

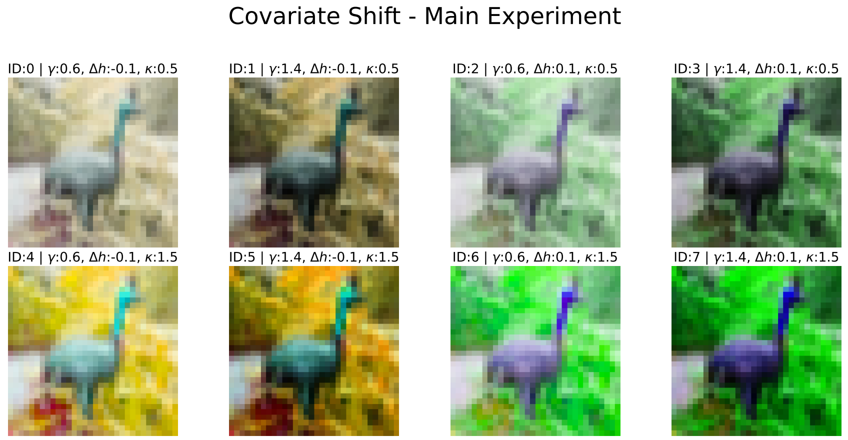

Covariate Shift - Main Experiment
ID:0 | γ:0.6, Δh:-0.1, κ:0.5
ID:1 | γ:1.4, Δh:-0.1, κ:0.5
ID:2 | γ:0.6, Δh:0.1, κ:0.5
ID:3 | γ:1.4, Δh:0.1, κ:0.5
ID:4 | γ:0.6, Δh:-0.1, κ:1.5
ID:5 | γ:1.4, Δh:-0.1, κ:1.5
ID:6 | γ:0.6, Δh:0.1, κ:1.5
ID:7 | γ:1.4, Δh:0.1, κ:1.5

Figure 8: Visualization of a sample from CIFAR-10 under various nonlinear transformations.

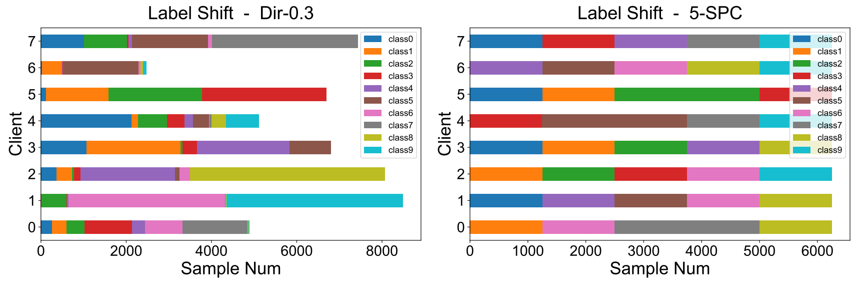


Figure 9: Visualization of client data distribution on CIFAR-10 under Dir-0.3 and 5-SPC settings.

F.2.1 Number of clients

To investigate the impact of the number of clients, we adopt a more fine-grained strategy for simulating covariate shift. We expand the parameter space for each nonlinear transformation to three distinct values: gamma correction with $\gamma \in \{0.6, 1.0, 1.4\}$, hue adjustment with $\Delta h \in \{-0.15, 0.0, 0.15\}$, and saturation scaling with $\kappa \in \{0.4, 1.0, 1.6\}$. Furthermore, we introduce an additional binary transformation, posterization, which reduces the number of bits for each color channel to create a flattening effect on the image's color palette. A visualization of these transformations is presented in Fig. 10. In the n-client setting, we apply the first n transformations from this pool.

In our experiments, we set the maximum number of clients to 32. This is due to two primary challenges. First, as the number of clients increases, the amount of data partitioned to each client diminishes significantly. This data scarcity creates a scenario where fine-tuning-based methods gain an inherent advantage, as each client's local train and test distributions are identical. Second, it is hard to design a simulation strategy for covariate shift that is both sufficiently distinct and aligned with the model's inductive bias when the number of clients becomes very large.

F.2.2 Covariate shift intensity

To evaluate the robustness of our method under varying degrees of covariate shift, we construct three intensity levels—low, mid, and high—by adjusting the parameter ranges of the nonlinear transformations. The specific value ranges for each level are detailed as follows: (1) **Low**: $\gamma \in \{0.9, 1.1\}, \Delta h \in \{-0.01, 0.01\}, \kappa \in \{0.9, 1.1\}$. (2) **Mid**: $\gamma \in \{0.75, 1.25\}, \Delta h \in \{-0.05, 0.05\}, \kappa \in \{0.7, 1.3\}$. (3) **High**: $\gamma \in \{0.6, 1.4\}, \Delta h \in \{-0.1, 0.1\}, \kappa \in \{0.5, 1.5\}$. Visualizations corresponding to these levels are presented in Fig. 11, Fig. 12, and Fig. 8. It can be seen that the induced covariate shift is nearly imperceptible at the low level and escalates to a stark distinction at the high level, clearly illustrating the progressive intensity of the shift.

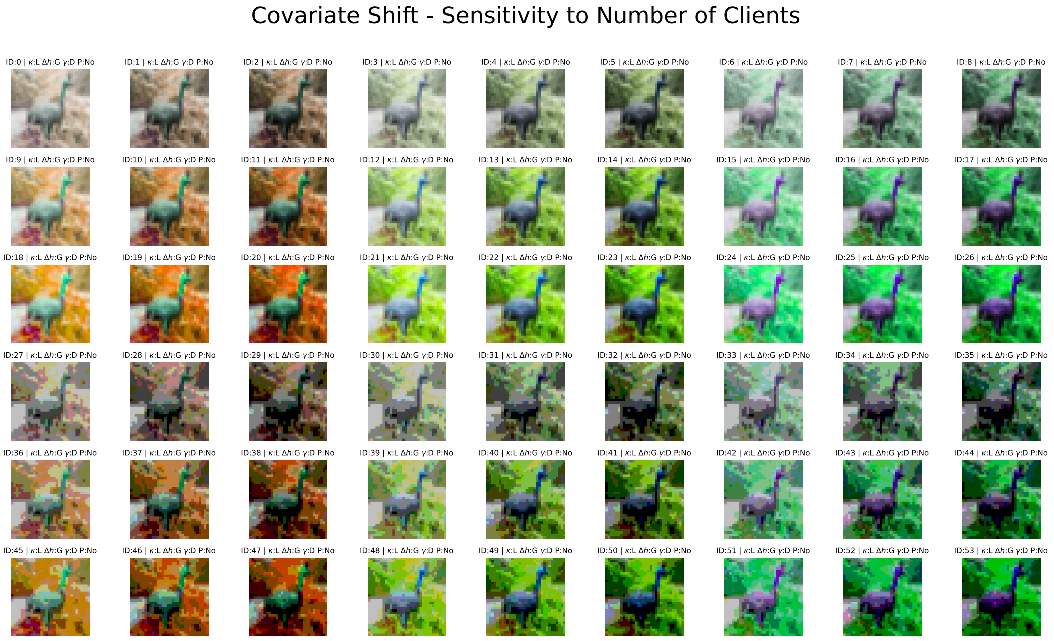

Covariate Shift - Sensitivity to Number of Clients
ID:0 | κ:L Δh:G γ:D P:No
ID:1 | κ:L Δh:G γ:D P:No
ID:2 | κ:L Δh:G γ:D P:No
ID:3 | κ:L Δh:G γ:D P:No
ID:4 | κ:L Δh:G γ:D P:No
ID:5 | κ:L Δh:G γ:D P:No
ID:6 | κ:L Δh:G γ:D P:No
ID:7 | κ:L Δh:G γ:D P:No
ID:8 | κ:L Δh:G γ:D P:No
ID:9 | κ:L Δh:G γ:D P:No
ID:10 | κ:L Δh:G γ:D P:No
ID:11 | κ:L Δh:G γ:D P:No
ID:12 | κ:L Δh:G γ:D P:No
ID:13 | κ:L Δh:G γ:D P:No
ID:14 | κ:L Δh:G γ:D P:No
ID:15 | κ:L Δh:G γ:D P:No
ID:16 | κ:L Δh:G γ:D P:No
ID:17 | κ:L Δh:G γ:D P:No
ID:18 | κ:L Δh:G γ:D P:No
ID:19 | κ:L Δh:G γ:D P:No
ID:20 | κ:L Δh:G γ:D P:No
ID:21 | κ:L Δh:G γ:D P:No
ID:22 | κ:L Δh:G γ:D P:No
ID:23 | κ:L Δh:G γ:D P:No
ID:24 | κ:L Δh:G γ:D P:No
ID:25 | κ:L Δh:G γ:D P:No
ID:26 | κ:L Δh:G γ:D P:No
ID:27 | κ:L Δh:G γ:D P:No
ID:28 | κ:L Δh:G γ:D P:No
ID:29 | κ:L Δh:G γ:D P:No
ID:30 | κ:L Δh:G γ:D P:No
ID:31 | κ:L Δh:G γ:D P:No
ID:32 | κ:L Δh:G γ:D P:No
ID:33 | κ:L Δh:G γ:D P:No
ID:34 | κ:L Δh:G γ:D P:No
ID:35 | κ:L Δh:G γ:D P:No
ID:36 | κ:L Δh:G γ:D P:No
ID:37 | κ:L Δh:G γ:D P:No
ID:38 | κ:L Δh:G γ:D P:No
ID:39 | κ:L Δh:G γ:D P:No
ID:40 | κ:L Δh:G γ:D P:No
ID:41 | κ:L Δh:G γ:D P:No
ID:42 | κ:L Δh:G γ:D P:No
ID:43 | κ:L Δh:G γ:D P:No
ID:44 | κ:L Δh:G γ:D P:No
ID:45 | κ:L Δh:G γ:D P:No
ID:46 | κ:L Δh:G γ:D P:No
ID:47 | κ:L Δh:G γ:D P:No
ID:48 | κ:L Δh:G γ:D P:No
ID:49 | κ:L Δh:G γ:D P:No
ID:50 | κ:L Δh:G γ:D P:No
ID:51 | κ:L Δh:G γ:D P:No
ID:52 | κ:L Δh:G γ:D P:No
ID:53 | κ:L Δh:G γ:D P:No

Figure 10: Visualization of a CIFAR-10 sample under covariate shift with a larger number of clients.

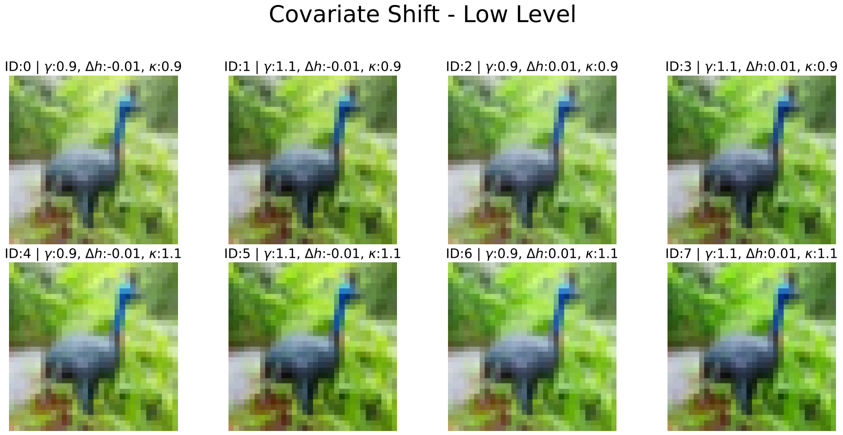

Covariate Shift - Low Level
ID:0 | γ:0.9, Δh:-0.01, κ:0.9
ID:1 | γ:1.1, Δh:-0.01, κ:0.9
ID:2 | γ:0.9, Δh:0.01, κ:0.9
ID:3 | γ:1.1, Δh:0.01, κ:0.9
ID:4 | γ:0.9, Δh:-0.01, κ:1.1
ID:5 | γ:1.1, Δh:-0.01, κ:1.1
ID:6 | γ:0.9, Δh:0.01, κ:1.1
ID:7 | γ:1.1, Δh:0.01, κ:1.1

Figure 11: Visualization of a sample from CIFAR-10 under low covariate shift intensity.

F.2.3 Backbone sharing strategy

In our formulation, the target-class classification task uses the embedding $g_\theta(x)$ for an input feature x, while the client-classification task uses the embedding $h_\tau(g_\theta(x))$ for the same feature. Since both g_θ and h_τ are parameterized functions, the optimal sharing strategy between the two is not obvious. To explore this, we investigate four backbone-sharing strategies based on LeNet: no sharing, shallow sharing, mid sharing, and deep sharing. The network architectures are illustrated in Fig. 13. From (a) to (c), the discrepancy between the embeddings for the two tasks decreases, while the number of learnable parameters also reduces. Our empirical results in Fig. 6 show that all strategies perform similarly, with shallow sharing slightly ahead. However, given the substantial increase in parameters for shallow sharing, deep sharing offers a more parameter-efficient alternative while maintaining strong performance.

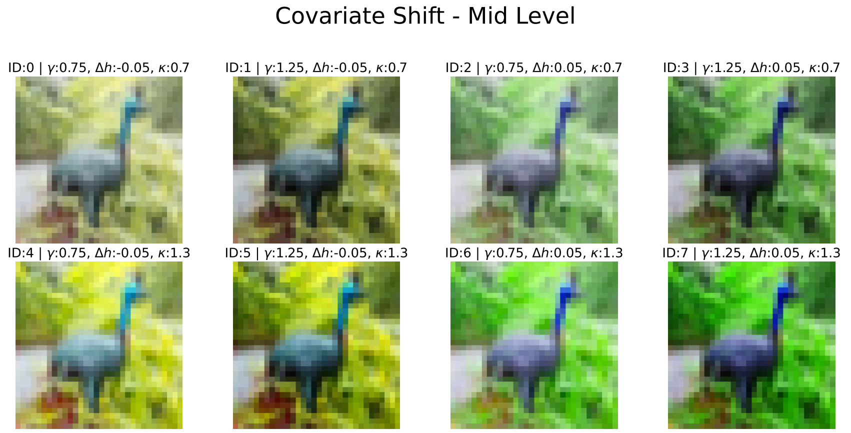
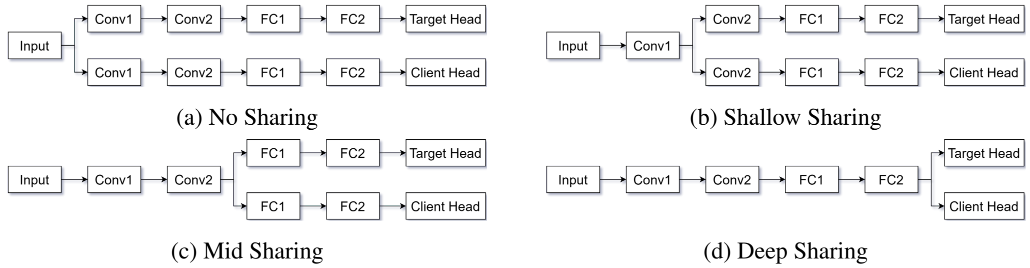

Covariate Shift - Mid Level
ID:0 | γ:0.75, Δh:-0.05, κ:0.7
ID:1 | γ:1.25, Δh:-0.05, κ:0.7
ID:2 | γ:0.75, Δh:0.05, κ:0.7
ID:3 | γ:1.25, Δh:0.05, κ:0.7
ID:4 | γ:0.75, Δh:-0.05, κ:1.3
ID:5 | γ:1.25, Δh:-0.05, κ:1.3
ID:6 | γ:0.75, Δh:0.05, κ:1.3
ID:7 | γ:1.25, Δh:0.05, κ:1.3

Figure 12: Visualization of a sample from CIFAR-10 under mid covariate shift intensity.



(a) No Sharing

(b) Shallow Sharing

(c) Mid Sharing

(d) Deep Sharing

Figure 13: Visualization of the four different parameter sharing strategies.

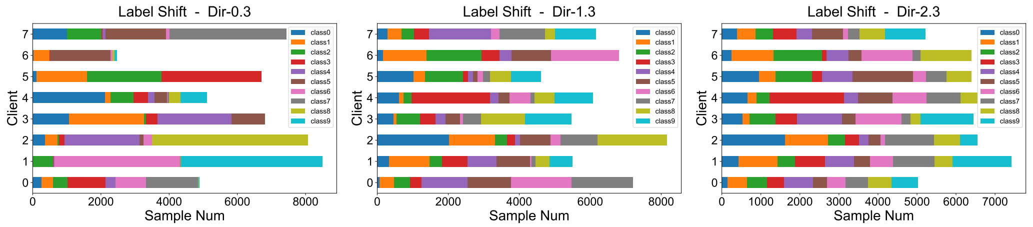


Figure 14: Visualization of client data distribution under Dir-0.3/1.3/2.3 settings.

Table 4: System accuracy and average accuracy under different Dirichlet parameter α values.

Method	System Accuracy			Average Accuracy		
	$\alpha = 0.3$	$\alpha = 1.3$	$\alpha = 2.3$	$\alpha = 0.3$	$\alpha = 1.3$	$\alpha = 2.3$
Ditto	47.64 ± 0.25	43.78 ± 0.23	44.04 ± 0.20	76.34 ± 0.11	55.56 ± 0.18	52.76 ± 0.18
FedRep	24.96 ± 0.19	35.45 ± 0.22	37.87 ± 0.22	76.49 ± 0.15	55.09 ± 0.17	52.95 ± 0.16
FedBABU	57.43 ± 0.17	54.60 ± 0.19	54.25 ± 0.20	78.22 ± 0.14	61.61 ± 0.16	59.83 ± 0.20
FedPAC	25.14 ± 0.21	35.53 ± 0.21	37.84 ± 0.20	76.53 ± 0.13	55.08 ± 0.15	52.95 ± 0.16
FedALA	61.33 ± 0.17	48.47 ± 0.18	46.68 ± 0.20	64.35 ± 2.40	49.14 ± 0.62	47.29 ± 0.50
FedAS	28.76 ± 0.19	41.43 ± 0.20	46.08 ± 0.22	78.69 ± 0.17	59.83 ± 0.19	58.17 ± 0.18
ConFREE	25.66 ± 0.22	36.12 ± 0.21	38.78 ± 0.23	76.73 ± 0.16	55.56 ± 0.18	53.58 ± 0.15
FedAvgFT	54.90 ± 0.22	54.80 ± 0.18	55.61 ± 0.19	79.08 ± 0.11	62.87 ± 0.16	61.94 ± 0.17
FedProxFT	55.01 ± 0.20	54.84 ± 0.15	55.62 ± 0.18	79.07 ± 0.12	62.84 ± 0.15	61.86 ± 0.19
FedSAMFT	55.83 ± 0.21	49.15 ± 0.17	47.71 ± 0.16	75.53 ± 0.11	55.46 ± 0.19	53.54 ± 0.15
FedDRM	$\mathbf{63.85 \pm 0.18}$	$\mathbf{56.83 \pm 0.18}$	$\mathbf{56.00 \pm 0.23}$	$\mathbf{80.25 \pm 0.14}$	$\mathbf{64.04 \pm 0.16}$	$\mathbf{62.30 \pm 0.15}$

F.2.4 Label shift intensity

To evaluate the robustness of our method under varying degrees of label shift, we compare our method with the baselines across a range of Dirichlet parameters $\alpha \in \{0.3, 1.3, 2.3\}$. Visualizations corresponding to these settings are presented in Fig. 14. The corresponding results for system accuracy and average accuracy are presented in Tab. 4.

Consistent with prior work (Xu et al., 2023), we can see that smaller α values—corresponding to higher data heterogeneity—lead to higher average accuracy for all methods. This occurs because each client's training and testing data are drawn from the same distribution. As α decreases, the local label distributions become increasingly skewed, with some classes receiving negligible probability mass. This effectively reduces the number of classes present on each client, thereby simplifying the local classification problem relative to the balanced case. For system accuracy, we find that this trend persists for our method, as the only additional component is the client-routing step, which does not alter the underlying behavior of local classification. In contrast, methods such as FedRep exhibit increasing system accuracy as α grows (i.e., as the label distributions become more homogeneous). When α is small, the local models become highly personalized and fail to reach a consistent consensus across clients, causing majority voting to misroute queries and thus lowering system accuracy. As α increases, this inconsistency diminishes, and the aggregated routing accuracy improves. These results further confirm that our method remains robust across varying degrees of label shift.