Federated Domain Generalization with Decision Insight Matrix

Tianchi Liao[1] , **Binghui Xie**[2] , **Lele Fu**[1] , **Sheng Huang**[1] , **Bowen Deng**[1] ,
Chuan Chen[1*] and **Zibin Zheng**[1]

[1]Sun Yat-sen University, Guangzhou, China
[2]The Chinese University of Hong Kong, Hong Kong
{liaotch, fulle, huangsh253, dengbw3}@mail2.sysu.edu.cn, bhxie21@cse.cuhk.edu.hk,
{chenchuan, zhzibin}@mail.sysu.edu.cn

Abstract

Federated domain generalization addresses the crucial challenge of developing models that can generalize across diverse domains while maintaining data privacy in federated learning settings. Current approaches either compromise privacy constraints or focus narrowly on specific aspects of model invariance, often incurring significant computational overhead. We propose a novel approach FedDIM, which leverages the concept of "insight matrix" - a fine-grained representation of the model's decision-making process derived from element-wise products between feature vectors and classifier weights. By introducing a regularization term that promotes consistency between individual sample insight matrices and their class-wise mean representations, our method effectively captures both feature and classifier invariance. This approach not only maintains strict privacy requirements but also introduces minimal computational overhead as it utilizes intermediate computations already present in the forward pass. Extensive experiments demonstrate that our method achieves superior out-of-distribution generalization compared to existing federated learning approaches while being simple to implement. Our work provides a new perspective on achieving robust generalization in federated learning settings through the lens of decision-making processes.

1 Introduction

Federated Learning (FL) has revolutionized the machine learning landscape by enabling collaborative model training across distributed clients while preserving data privacy [Liao *et al.*, 2025]. In this paradigm, clients train local models on their private data, which are then periodically aggregated by a central server to form a global model, thereby circumventing the need for direct access to raw data [Li *et al.*, 2020a; Li *et al.*, 2024; Chen *et al.*, 2025]. Although FL has demonstrated promising results in scenarios where data is independently and identically distributed (i.i.d.) [McMahan *et al.*, 2017; Shen *et al.*, 2025], real-world applications often present a more complex challenge: clients typically collect data independently, resulting in distinct distributions across different domains. Furthermore, during deployment, models frequently encounter data from previously unseen target domains, leading to a significant distribution shift problem [Huang *et al.*, 2023; Liao *et al.*, 2024; Wan *et al.*, 2024]. This scenario, known as Federated Domain Generalization (FedDG) and illustrated in Figure 1, raises a fundamental challenge beyond traditional FL's data heterogeneity: how can federated models effectively generalize across diverse domains while maintaining privacy constraints?

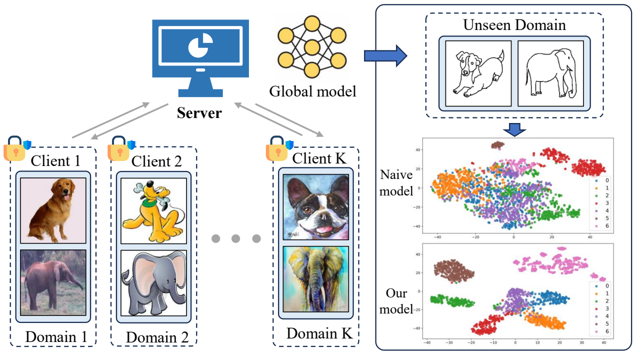


Figure 1: Problem illustration of federated domain generalization.

Traditional domain generalization methods focus on learning invariant relationships explicitly from data or representations [Hu *et al.*, 2024; Fu *et al.*, 2025a; Huang *et al.*, 2025]. However, these methods require a centralized setting where data or representations are shared across clients, potentially compromising client privacy. To address this challenge, researchers have begun exploring federated domain generalization [Qi *et al.*, 2024]. To investigate the invariance relationships between clients, various FL methods have been developed, focusing either on the feature level or the logit level [Qiao *et al.*, 2024]. [Zhang *et al.*, 2023] proposed a novel model aggregation method based on locally estimated generalization gaps, but their insight was limited to scenarios where each training domain was treated as a single client. [Huang *et al.*, 2023] proposed a prototype aggregation method, FPL, from the feature perspective. By introducing consistency regularization, it aligns local features with prototypes to ex-

[*] Corresponding author.

plicitly learn invariance in the feature extractor. [Guo *et al.*, 2023] introduced a regularization approach FedIIR based on local empirical risk minimization, which aims to implicitly learn invariance by constraining the parameter space. However, these methods fail to integrate information from both the feature extractor and the classifier [Hu *et al.*, 2023; Zhang *et al.*, 2024; Qi *et al.*, 2025; Fu *et al.*, 2025b].

In response, we suggest a novel approach that emphasizes the decision-making process in the classifier layer of deep neural networks, rather than focusing solely on feature or classifier invariance. In conventional models, final output logits are computed by multiplying the penultimate layer's features with the classifier's weights. A deeper analysis reveals that [Chen *et al.*, 2023] each logit value can be decomposed into the summation of element-wise products between the feature vector and corresponding weight vector . While most FL methods rely on feature prototypes [Wan *et al.*, 2024; Bai *et al.*, 2024] to learn invariance, the intermediate element-wise products (before summation) retain more fine-grained information. Viewing each product term as a contribution to its corresponding logit, we collect these contributions across all classes into a matrix. This matrix, which we term the "insight matrix", encapsulates the model's decision-making process for input classification. By exchanging insight matrices in FL, clients can share cross-domain knowledge representations, thus facilitating invariant learning.

Based on insight matrices decision, we propose a federated domain generalization model, named **FedDIM**, which provides a new theoretical and practical framework for invariant learning. We assume that different clients have heterogeneous input/output distributions, but a well-generalized model should make decisions based on cues that are consistent across samples and clients. Based on this intuition, we propose a regularization term that promotes similarity between each sample's insight matrix and the mean insight matrix of its corresponding class. Our approach offers two key advantages: First, it explicitly combines the semantic information of the encoder and classifier through fine-grained modeling, thus enhancing the model's ability to adapt to client-domain heterogeneity; Second, the insight matrix, as a natural byproduct of the logit computation process, has minimal computational overhead. Experimental results show that FedDIM improves the model's ability to learn invariant relationships across client domains. In summary, the main contributions of this paper are as follows:

- **New perspective:** In this paper, considering the privacy of FL, we propose a novel method built upon the concept of category-wise mean insight matrices. Bridging the gap between focusing only on feature invariance or logit invariance in FL, offering new insights into out-of-distribution (OOD) generalization of FL.
- **Simple yet effective algorithm:** We introduce an efficient strategy that leverages insight matrices to enhance model robustness. Our method requires minimal modifications to the standard FedAvg algorithm while achieving superior performance. The lightweight implementation adds only a few lines of code while its effectiveness is theoretically guaranteed.
- **Superior Performance:** We conduct extensive experiments on multiple benchmark datasets. The results demonstrate that FedDIM consistently outperforms existing federated learning methods in Out-of-Distribution generalization.

2 Preliminaries

2.1 Problem Setting

In federated domain generalization, the data in each client is sampled from different domains. Let $\mathcal{D}$ denote the set of all client domains. We denote the training domain by $\mathcal{D}_{tr} = \{\mathcal{D}_1, \cdots \mathcal{D}_M\}$, $\mathcal{D}_{tr} \subseteq \mathcal{D}$, where M is the number of training domains (or clients). Let $\mathcal{X}$ and $\mathcal{Y}$ represent the input space and target space, respectively, the sample contains K classes. Each client $c \in \mathcal{D}$ holds a local dataset denoted as $\{(x_i^c, y_i^c)\}_{i=1}^{n_c}$, where n_c is the number of samples. Let the loss function as $\mathcal{L}(f(x), y)$. Then, for each client c, the expected risk as $\mathcal{E}_c(f) = \mathbb{E}_{x^c, y^c}\mathcal{L}(f(x^c), y^c)$, and the global expected risk of model f denotes as $\mathcal{E}_{D_{tr}}(f) = \mathbb{E}_{D_{tr}}\mathcal{E}_c(f)$.

The ideal goal of FL training is to minimize the overall loss function on the dataset $\mathcal{D}$. However, in practice, FL typically involves a large number of clients with heterogeneous data distributions, and only a subset of clients participate in the training. This introduces a distribution shift between the participating clients and those not seen during training, leading to the out-of-distribution (OOD) generalization problem [Qi *et al.*, 2024; Xie *et al.*, 2024]. Therefore, instead of optimizing the expected risk over the entire domain, we focus on optimizing the following empirical risk objective:

$$\min_{\theta} \mathcal{E}_{\mathcal{D}}(f) \approx \mathcal{E}_{D_{tr}}(f) = \frac{1}{M}\sum_{c=1}^{M}\sum_{i=1}^{n_c}\mathcal{L}\left(f(x_i^c;\theta), y_i^c\right) \quad (1)$$

The federated OOD problem cannot be solved directly since not all potential clients are observed. This is more challenging than ordinary heterogeneous FL. To generalize to non-participating clients, the key to our study is how to learn the invariant relationship between inputs and goals.

2.2 Modeling Decision Insight Matrix

In the current deep model, the final output of the decision process involves two main steps: (1) the feature extractor: transforming the input from the original feature space to the feature embedding space, i.e., $\mathbf{z} = h(\varphi, x) \in \mathbb{R}^D : \mathcal{X} \to \mathcal{Z}$; and (2) the classifier: using the features to compute the final logits, i.e., $\mathbf{o} = g(w, \mathbf{z}) \in \mathbb{R}^K : \mathcal{Z} \to \hat{\mathcal{Y}}$. Thus the model can be written as $f(\theta) = g(w) \circ h(\varphi)$, where $\theta = (\varphi, w)$.

In the FedDG scenario, clients primarily focus on samples from their local domains, while the server needs to aggregate inter-domain information from multiple clients to learn invariant relationships. Most existing research has focused on learning invariance by uploading features or logits to regularize the model, but these approaches have certain limitations: ❶ Solely focusing on feature invariance often overlooks the importance of classifier weights across different feature elements, which may lead to biased estimates of feature importance, thereby weakening the model's generalization ability. ❷ Although logits implicitly encode the relationship between

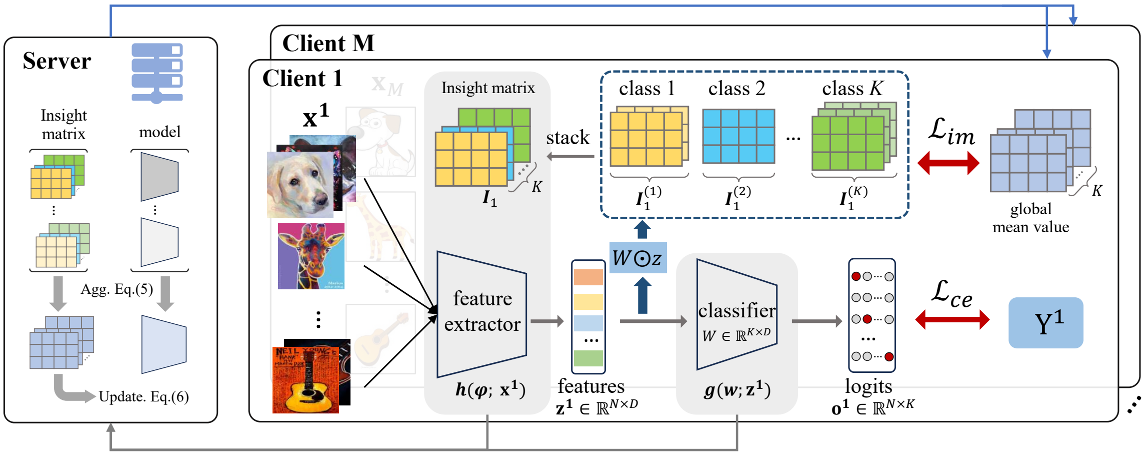


Figure 2: An overview of FedDIM based on insight matrix. The clients process data from different domains. clients update the local model by minimizing the classification loss $\mathcal{L}_{ce}$ and the distance $\mathcal{L}_{im}$ between the global and the local insight matrix. then upload them to the server. The server sends down the global model and matrix to the client after aggregating it and updating the global insight matrix.

the classifier's weights, they only provide rough numerical values and lack fine-grained recognition of cross-domain generalizability, thus lacking deeper insight into the underlying decision-making process.

Since each logit value can be decomposed into the sum of the element-wise products of the feature vector and the corresponding weight vector, we argue that these intermediate product terms retain more granular information. Therefore, assuming the label set contains K classes, the logits can be represented as $\mathbf{o} = \mathbf{W}^T \mathbf{z} \in \mathbb{R}^K$, where $\mathbf{W} \in \mathbb{R}^{D \times K}$ is the weight. For simplicity, we ignore the bias term of the classifier. Based on this decomposition, we treat each product term as the contribution of its corresponding logit. Thus, the logic value of class k can be expressed as

$$o_k = \mathbf{W}_{\{,k\}}^{\top} \mathbf{z} = \sum_{j=1}^{D} W_{\{j,k\}} z_j. \tag{2}$$

We aggregate the contributions of all categories into a matrix, called the "Insight Matrix", which is defined sa following:

$$\mathbf{I} = \begin{bmatrix} W_{\{1,1\}} z_1 & W_{\{1,2\}} z_1 & \cdots & W_{\{1,K\}} z_1 \\ W_{\{2,1\}} z_2 & W_{\{2,2\}} z_2 & \cdots & W_{\{2,K\}} z_2 \\ \vdots & \vdots & \ddots & \vdots \\ W_{\{D,1\}} z_D & W_{\{D,2\}} z_D & \cdots & W_{\{D,K\}} z_D \end{bmatrix}. \tag{3}$$

The insight matrix $\mathbf{I} \in \mathbb{R}^{D \times K}$ encapsulates key information about the model's decision-making process when classifying an input.

3 Methodology

We propose a federated domain generalization solution that utilizes the insight matrix as a key component for exchanging information between servers and clients. The core idea is that the insight matrix of samples from the same class is consistent with its corresponding average value, implying that the insight matrices for the same class across different domains should exhibit similarity. This ensures that the model makes classification decisions based on the same reasoning process. The framework of our method is illustrated in Figure 2.

3.1 Local Mean Class Insight Matrix

According to Eq. (2), the client generates an insight matrix for each sample. We define a local mean insight matrix $\mathbf{I}_c^{(k)}$ to represent the k-th class. For the c-th client, the average insight matrix is the mean of the insight matrices of the samples belonging to class k.

$$\mathbf{I}_c^{(k)} = \frac{1}{|n_{c,k}|} \sum_{(x,y) \in n_{c,k}} W \odot h(\varphi, x), \tag{4}$$

where $\odot$ represents element-wise product, $\mathbf{I}_c^{(k)} \in \mathbb{R}^{D \times K}$ and $n_{c,k}$ denotes the samples with class k in client c. We calculate the average insight matrix for each class and stack to get the local mean insight matrix $\mathbf{I}_c = [\mathbf{I}_c^{(1)}, \cdots, \mathbf{I}_c^{(K)}] \in \mathbb{R}^{K \times D \times K}$.

3.2 Global Aggregation and Update

To generalize the global model to unseen clients, it is insufficient to simply aggregate the models of participating clients on the server. Although the clients possess domain information with different distributions, they share the same label space, which enables the participating clients to share a common embedding space. By aggregating the clients' insight matrices based on class information, we can learn invariant relationships in the federated domain generalization scenario. Thus, our aggregated global model and global insight matrix can be expressed as:

$$\theta^{t+1} = \frac{1}{C} \sum_{c \in C} \theta_c^t, \quad \text{and} \quad \bar{\mathbf{I}}_g = \frac{1}{C} \sum_{c \in C} \mathbf{I}_c, \tag{5}$$

where θ_c^t is the model trained by client c in round t, and C is the number of clients sampled per round.

In the FL training process, clients perform random sampling in each round. However, in the FedDG scenario, the

Algorithm 1 FedDIM

```
Input: total rounds $T$, local epochs $E$, total number of clients $M$, sampled number of clients $C$, learning rate $\eta$, hyper-parameter for loss $\lambda$
Server executes:
Initialize global model $\theta$ and global insight matrix $\bar{\mathbf{I}}$
for each round $t = 1 \cdots T$ do
    Server samples subset $C$ of clients
    for each client $c \in C$ in parallel do
        $\{\theta_c^t, \mathbf{I}_c\} \leftarrow$ Clients updates($\theta^t$, $\bar{\mathbf{I}}^t$)
    end for
    Update global model and calculate the global insight matrix by Eq. (5)
    Update global insight matrix $\bar{\mathbf{I}}^{t+1}$ by Eq. (6)
end for
Clients updates:
Initialize local model $\theta_c^t = \theta^t$
for each local epoch $e = 1 \cdots E$ do
    Sample mini-batch in $B$:
    Calculate the $n$-th sample insight matrix $\mathbf{I}_n$
    Calculate local loss by Eq. (7)
    Update local model: $\theta_c^t \leftarrow \theta_c^t - \eta \nabla \mathcal{L}_c\left(\theta_c^t; \mathcal{B}_c\right)$
end for
Calculate the mean class insight matrix $\mathbf{I}_c$ by Eq. (4)
return $\theta_c^t$ and $\mathbf{I}_c$
```

significant differences in client data distributions can lead to considerable variations in the aggregated insight matrix in each round. Therefore, we adopt a momentum to update the global insight matrix in each round. This method reduces the impact of noise in each iteration and balances new information with historical data, making the update process smoother and preventing overreaction to individual training data.

$$\bar{\mathbf{I}}^{t+1} = (1-m) \times \bar{\mathbf{I}}^t + m \times \bar{\mathbf{I}}_g, \tag{6}$$

where m is a positive momentum value, and $\bar{\mathbf{I}}$ is initialized from the first iteration to compute the processed insight matrix, and $\bar{\mathbf{I}}_g$ calculated by Eq. (5).

3.3 Local Model Update

Clients update their local models to learn invariant relationships and generate consistent insight matrices across clients. To achieve this, we introduce a regularization term in the local loss, which enables the local model to capture the invariant relationship between data and targets during single-domain learning. Specifically, the loss is defined as follows:

$$\begin{aligned}\mathcal{L} &= \mathcal{L}_{ce} + \mathcal{L}_{im} \\ &= \sum_{i=1}^{B} L_{ce}\left(f(x_i^c; \theta), y_i^c\right) + \lambda \frac{1}{B} \sum_{k} \sum_{\{i \mid y_i = k\}} \|\mathbf{I}_i - \bar{\mathbf{I}}_k\|^2,\end{aligned} \tag{7}$$

where $||\cdot||$ is the l_2 norm, L_{ce} is the cross entropy loss, B is the number of samples in a mini-batch. $\mathbf{I}_i$ is the insight matrix for the n-th sample. $\bar{\mathbf{I}}_k$ is the global insight matrix corresponding to the k-th class distributed by the server.

4 Theoretical Analysis

This section presents the theoretical analysis demonstrating how our methods address the distribution shift problem. At first, we provide the following lemma, which is from [Ben-David *et al.*, 2010] to bound the distribution divergence between two different domains.

Lemma 4.1. *Let $d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{A}, \mathcal{B})$ denotes the domain divergence between two domain distributions $\mathcal{A}$ and $\mathcal{B}$. The expected risk gap between $\mathcal{A}$ and $\mathcal{B}$ is bounded as $|\mathcal{E}_{\mathcal{A}}(\theta) - \mathcal{E}_{\mathcal{B}}(\theta)| \le \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{A}, \mathcal{B})$.*

Then, we consider the federated domain generalization setting where the training data follow the distribution $\mathcal{D} = \bigcup_{c=1}^{|\mathcal{C}|} \mathcal{D}c$, with $\mathcal{D}_c$ denoting the data distribution of client c among $|\mathcal{C}|$ total clients [Yan and Guo, 2025]. Each client maintains a training set D_c sampled from $\mathcal{D}_c$ with size $n_c = |D_c|$, forming an overall training set D of $n = \sum_c n_c$ samples and model aggregation weights $\{p_c = \frac{n_c}{n}\}$. Let $\mathcal{E}_{\mathcal{D}}(\theta)$ denote the expected risk on $\mathcal{D}$ and $\hat{\mathcal{E}}_D(\theta)$ denote the empirical risk on D. We define $\hat{\mathcal{E}}_{D_c}(\theta), \hat{I}_c(\theta)$, as the two loss terms in Eq. (7).

Theorem 4.2. *Let $\hat{\theta}$ be the aggregated global model federatedly trained with the proposed overall loss function. Define $\theta_{\mathcal{T}}^* := \arg\min_\theta \mathcal{E}_{\mathcal{T}}(\theta)$ and $\theta_c^* := \arg\min_{\theta_c} \hat{\mathcal{E}}_{D_c}(\theta_c)$. Let $\mathcal{H}$ be a hypothesis space of VC dimension d. For any $\delta \in (0,1)$, with probability at least $1-\delta$, the generalization gap of the model $\hat{\theta}$ on the unseen testing domain $\mathcal{T}$ has the following bound,*

$$\begin{aligned}\mathcal{E}_{\mathcal{T}}(\hat{\theta}) - \mathcal{E}_{\mathcal{T}}\left(\theta_{\mathcal{T}}^*\right) \le \sum_c p_c(\Big(\hat{\mathcal{E}}_{D_c}(\hat{\theta}) - \hat{\mathcal{E}}_{D_c}\left(\theta_c^*\right)\Big) \\ + \hat{I}_c(\hat{\theta}) + d_{\mathcal{H}\Delta\mathcal{H}}\left(\mathcal{D}_c, \mathcal{T}\right) \\ + O\left(\sqrt{\frac{1}{n_c}\left(\log\frac{1}{\delta} + d\log\frac{n_c}{d}\right)}\right)) + \Delta.\end{aligned} \tag{8}$$

The complete proof is provided in the supplementary material. In this theorem, $d_{\mathcal{H}\Delta\mathcal{H}}\left(\mathcal{D}_c, \mathcal{T}\right)$ represents the domain divergence between source domain $\mathcal{D}_c$ and target domain $\mathcal{T}$, while Δ represents the optimal model's residual error on both $\mathcal{D}$ and $\mathcal{T}$. The theorem establishes that the generalization gap of the global model $\hat{\theta}$ on target domain $\mathcal{T}$ is upper-bounded by two components: a weighted average term and the residual error. The weighted average term incorporates each client's empirical risk, insight matrix, and domain divergence bounds. Our proposed method aims to enhance the global model's generalization ability on $\mathcal{T}$ by explicitly minimizing the first two bound terms. The final term, which emerges from converting expected loss to empirical loss, is determined by the dataset size

5 Experiments

5.1 Experimental Setup

Datasets. To evaluate our approach, we conducted experiments in four datasets, **RotatedMNIST** [Ghifary *et al.*, 2015], is a MNIST dataset of 7000 samples by rotating it at angles of 0°, 15°, 30°, 45°, 60°, and 75°, resulting in six different domains. **PACS** [Li *et al.*, 2017], has 9991 images consisting of seven object categories in four domains (photo, art, cartoon, and sketch). **VLCS** [Fang *et al.*, 2013], has 10729 images consisting of five object categories in four domains (Caltech101, LabelMe, SUN09, and VOC2007). **OfficeHome** [Venkateswara *et al.*, 2017], is an image recognition dataset that includes 15,588 images of 65 classes from four different domains (art, clipart, product, and real-world). These are commonly used in the literature for domain generalization. We adhere to the experimental methodology outlined in FedIIR [Guo *et al.*, 2023]. For all datasets, we perform "leave-one-domain-out" strategy [Gulrajani and Lopez-Paz, 2020], where we choose one domain as the test domain, train the model on all remaining domains, and evaluate it on the chosen domain. Each source domain is treated as a client. Following standard practice, we use 90% of available data as training data and 10% as validation data.

Considering the FL setting, we explore two scenarios based on the number of clients: the one-domain-one-client scenario and the one-domain-multiple-clients scenario. In the one-domain-one-client scenario, each training domain is treated as an individual client. In the one-domain-multiple-clients scenario, data from each training domain is randomly partitioned into multiple subsets, with each client containing data from one subset of a given training domain. The details of the data partitioning are provided in the Appendix C.1.

Baselines. We consider 2 classic federated methods FedAvg [McMahan *et al.*, 2017], FedProx [Li *et al.*, 2020b], and 4 state-of-the-art federated methods for domain generalization FedADG [Zhang *et al.*, 2021], FedSR [Nguyen *et al.*, 2022], FedIIR [Guo *et al.*, 2023] and FedLGF [Yan and Guo, 2025] as baselines.

Implementation. We design dataset-specific models for each task. For the RotatedMNIST dataset, the feature encoder consists of four convolutional blocks, with ReLU activation, group normalization, and average pooling, followed by a linear classifier. During training, the batch size is 64. For the VLCS and PACS datasets, ResNet-18 is used as the feature encoder, while ResNet-50 is employed for the OfficeHome dataset. The classifiers for these three datasets consist of two fully connected layers. During training, the batch size is 32. For all datasets and scenarios, we set the communication rounds T to 100, with local iteration per round E=1 to accommodate limited local computational resources. Local models are updated using the SGD optimizer with a momentum of 0.9. The best parameters reported in the original paper were selected for the baseline, and the optimal hyperparameters of FedDIM were found by grid search. Each experiment was repeated 3 times and the average value was calculated.

5.2 Experimental Results

We evaluated all methods in two scenarios, where M denotes the total number of clients and C represents the number of sampled clients. Table 1 reports the results for the one-domain-one-client scenario, while Table 2 presents the results for the one-domain-multi-client scenario. For clarity, detailed results for specific domains are provided in the Appendix. C.

The experimental results demonstrate that, under the cross-domain client setting, our proposed method consistently outperforms other state-of-the-art baselines. In terms of average accuracy, we outperform the latest baseline FedLGF by 1.47% across all datasets. These observations validate the effectiveness of our method compared to existing baselines.

As the total number of clients increases, the performance of all methods declines significantly in the one-domain-multi-client scenario. In particular, FedADG and FedSR exhibit the most noticeable performance drops, likely because the increased number of clients makes it difficult to align the distributions across different source domains. To further validate this hypothesis, we extended the number of clients to 100, with the experimental results provided in Table 3 of the Appendix. Under this setting, both FedADG and FedSR perform worse than FedAvg. Additionally, when client data volume is large, the performance of FedIIR also drops significantly, potentially due to the reduced number of samples per category for each client caused by the increased number of clients. In contrast, our method exhibits the smallest performance degradation, highlighting its effectiveness and strong generalization capability in multi-client scenarios.

Loss surface visualization. We visualized the loss surface in a one-domain-one-client scenario using the "art" test domain in the PACS dataset, as shown in Figure 3. In the visualization, we used the global model as the origin and labeled the local models. This approach is consistent with the visualization technique in [Garipov *et al.*, 2018]. It can be observed that compared to FedAvg, our local models all converge in the flat region of the loss surface, and in this way the global model induced is more generalizable. In addition, we find that the gap between the global and local models is much smaller, suggesting that our approach has a clear advantage in maintaining a consistent optimization objective across different domains.

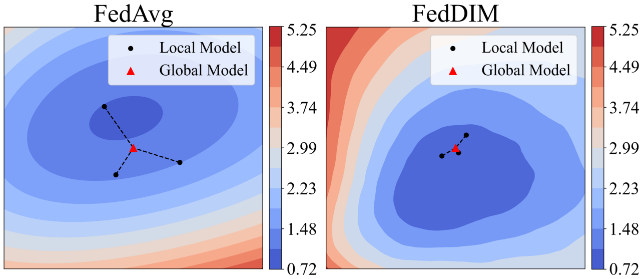


Figure 3: Loss surfaces w.r.t. model parameters on the PACS dataset with target domain "art".

Visualization. To better demonstrate the effectiveness of our method, we performed t-SNE [Van der Maaten and Hinton, 2008] visualization of the features **z** on the PACS dataset, as shown in Figure 4.

Compared to FedAvg, our method is clearer in clustering classification. On the training domain, FedDIM presents a clear block structure, indicating the effectiveness of the training process. When the model is generalized to the test domain (dark red part), the clustering structure is still obvious. This demonstrates the effectiveness of our method in generalizing to unseen distributions.

Methods	RotatedMNIST (M=5, C=5) ConvNet	PACS (M=3, C=3) ResNet-18	VLCS (M=3, C=3) ResNet-18	OfficeHome (M=3, C=3) ResNet-50	Average
FedAvg	94.77 ± 0.2	83.13 ± 0.1	75.38 ± 0.7	68.94 ± 0.1	80.56
FedProx	94.41 ± 0.3	83.32 ± 0.2	76.43 ± 1.2	68.04 ± 0.5	80.55
FedADG	94.96 ± 0.0	83.28 ± 0.4	77.53 ± 0.3	68.87 ± 0.4	81.16
FedSR	94.65 ± 0.4	83.65 ± 0.3	75.48 ± 0.7	69.25 ± 0.3	80.76
FedIIR	95.22 ± 0.3	83.87 ± 0.3	77.75 ± 0.8	69.52 ± 0.1	81.59
FedLGF	95.09 ± 0.2	84.20 ± 0.5	77.23 ± 1.1	69.33 ± 0.2	81.46
FedDIM	95.83 ± 0.2	84.57 ± 0.4	79.12 ± 0.8	71.12 ± 0.2	82.66

Table 1: Performance comparison (%) of all compared mehtods on RotatedMNIST, PACS, VLCS, and OfficeHome using leave-one-out domain validation. Each training domain is considered as a client and all clients participate in each round of joint training.

Methods	RotatedMNIST (M=50, C=10) ConvNet	PACS (M=30, C=10) ResNet-18	VLCS (M=30, C=10) ResNet-18	OfficeHome (M=30, C=10) ResNet-50	Average
FedAvg	91.00 ± 0.4	76.82 ± 0.5	73.75 ± 0.9	67.59 ± 0.2	77.29
FedProx	91.11 ± 0.6	77.48 ± 0.7	74.18 ± 1.4	67.73 ± 0.7	77.63
FedADG	92.71 ± 0.3	77.89 ± 0.6	71.95 ± 1.7	67.23 ± 0.2	77.44
FedSR	92.44 ± 0.8	78.13 ± 0.5	73.33 ± 0.5	65.84 ± 0.6	77.43
FedIIR	93.28 ± 0.5	79.25 ± 0.5	75.12 ± 0.8	68.19 ± 0.3	78.96
FedLGF	92.84 ± 0.6	79.49 ± 0.6	75.79 ± 1.3	68.01 ± 0.4	79.03
FedDIM	93.86 ± 0.4	80.57 ± 0.5	77.12 ± 1.1	70.03 ± 0.5	80.40

Table 2: Performance comparison (%) of all compared mehtods on RotatedMNIST, PACS, VLCS, and OfficeHome using leave-one-out domain validation. The total number of participating clients is more than the number of domains, sampling 10 clients per round for training.

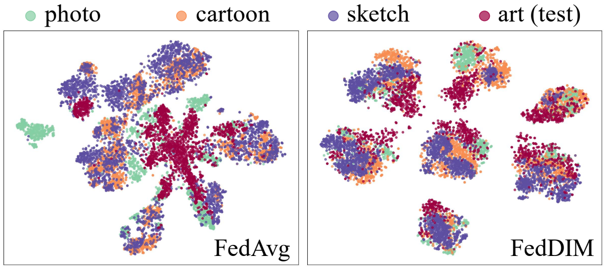


Figure 4: Visualization of t-SNE embedding for the PACS dataset with "art" as the unseen target domain. Here, different colors represent different domain. The seven clusters denote the classes.

5.3 Ablation Study

To confirm the validity of our approach, we designed the following five variants for comparison to assess the independent impact of each component.

- $W/\lambda = 0$: The method degenerates to FedAvg.
- $W/m = 0$: The insight matrix $\mathbf{I}$ is generated and fixed for the initial pre-trained model.
- $W/m = 1$: The insight matrix $\mathbf{I}$ is dynamically updated by the current round.
- $W/$Fea: We replaced 2nd term of Eq. (7) using feature-invariance as $\mathcal{L}_{im} = \frac{1}{B}\sum_k \sum_{\{n|y_i=k\}} \|\mathbf{z}_i - \bar{\mathbf{z}}_k\|^2$.
- $W/$Log: We replaced 2nd term of Eq. (7) using logit-invariance as $\mathcal{L}_{im} = \frac{1}{B}\sum_k \sum_{\{n|y_i=k\}} \|\mathbf{o}_i - \bar{\mathbf{o}}_k\|^2$.

Table 3 presents the experimental results of our different variants. It can be observed that the performance at $W/m = 0$ surpasses other variants, indicating that leveraging the insight matrix directly from the pre-trained model facilitates invariant decision-making. While $W/m = 1$ is outperformed by our momentum updating strategy due to the limited number of samples in a single batch, which cannot adequately consider all samples in the same class. Therefore, we design a scheme for dynamic updating based on historical information.

Moreover, it can be observed that both variants $W/$Fea and $W/$Log outperform the case where $\lambda = 0$. This is because the invariance constraint we adopt helps the model achieve better generalization and robustness to some extent. However, it is worth noting that this constraint may amplify the influence of irrelevant features, which have large values but correspond to small weights in the decision-making process, thus weakening the overall classification performance. Furthermore, focusing solely on logical invariance does not account for the varying contributions of individual features to the final decision. This can lead to small contributions being boosted to ensure that the summation equals the mean value, causing the model to emphasize irrelevant features and further degrading performance.

feature-invariance: focusing on the consistent expression of the features in the middle layer of the model.

logit-invariance: focuing on the stability of the model's predicted probability, which is the behavior of the output layer.

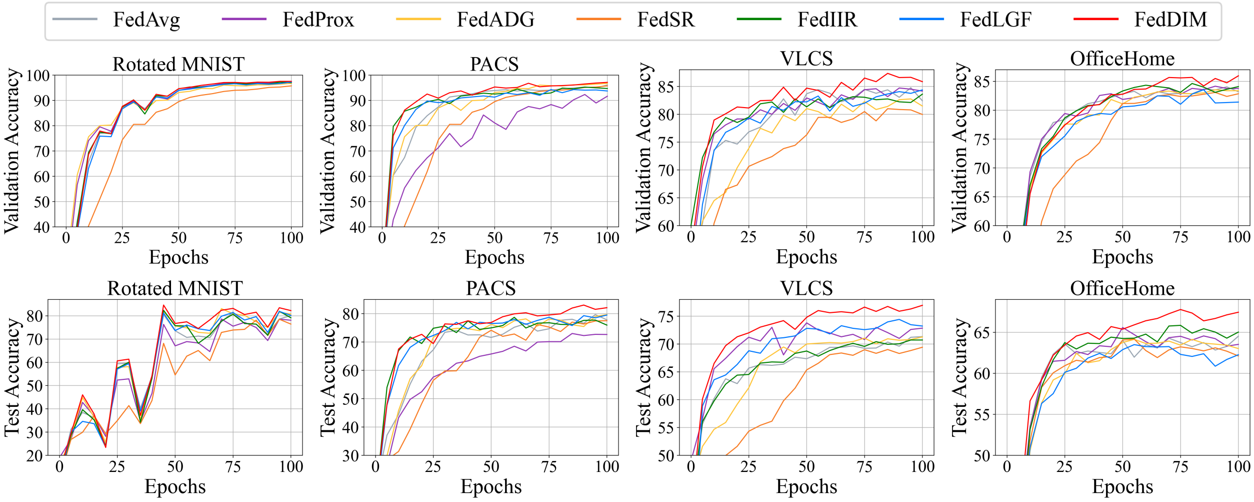


Figure 5: The accuracy convergence curves for the validation and test sets in one-domain-multiple-clients scenario. There are 50 clients for Rotated MNIST and 30 clients for the other datasets, and 10 clients are sampled in each round for training. Where the test domains of datasets Rotated MNIST, PACS, VLCS, and OfficeHome are ‘0°’, ‘art’, ‘VOC2007’, and ‘art’.

Methods	Invariance			Test datasets				
	F	O	I	M	P	V	H	Avg.
W/λ=0	-	-	✓	91.00	76.82	73.75	67.59	77.29
W/m=0	-	-	✓	93.79	79.33	76.92	68.16	79.55
W/m=1	-	-	✓	92.26	78.62	76.94	69.13	79.24
W/Fea	✓	-	-	93.84	78.23	76.82	68.96	79.46
W/Log	-	✓	-	92.26	77.71	75.16	68.6	78.43
FedDIM	-	-	✓	**93.86**	**80.57**	**77.12**	**70.03**	**80.40**

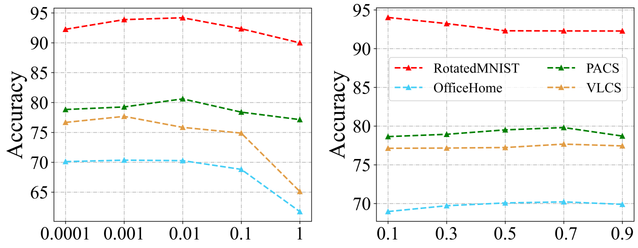
Table 3: Ablation study on four datasets. We abbreviated the symbols. Under Invariance content, F stands for feature, O stands for logical value, and I stands for our Insight Matrix. Under the test dataset content, M stands for Rotated Mnist, P stands for PACS, V stands for VLCS, and H stands for OfficeHome.

5.4 Visualization of the Convergence Process

We present the convergence behavior of different methods in a one-domain multiple-client scenario. The learning rate for all methods is fixed to the same value. We report the accuracy of each method on both the validation and test sets, as shown in Figure 5. From the figure, it is evident that all methods demonstrate stable convergence on the validation set, with our algorithm achieving relatively superior performance across various datasets, particularly on the VLCS and OfficeHome datasets. Moreover, although the performance of different methods on the test set exhibits varying degrees of fluctuation, overall, the test accuracy of all methods stabilizes in the later stages of training, indicating good convergence behavior of the models on the test set.

5.5 Parameter Sensitivity Analysis

We investigated the effects of momentum coefficient and the loss trade-off parameter in a one-domain-multi-client scenario. We evaluated the sensitivity of the model using four datasets in the range $\lambda \in \{0.0001, 0.001, 0.01, 0.1, 1\}$ and $m \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$, as shown in Figure 6. The results show that FedDIM performs consistently when $\lambda \in \{0.0001, 0.001, 0.01, 0.1\}$, but performance degrades significantly at $\lambda = 1$. Furthermore, models with static momentum ($m = 0$ or $m = 1$) perform worse than those with dynamic momentum updates, highlighting the importance of momentum in improving generalization.



(a) Trade-off parameter λ (b) Momentum coefficient m

Figure 6: Average test accuracy (%) for various values of the hyperparameter λ and m, with one-domain-multi-client setting.

6 Conclusion

We study OOD generalization in federated learning through a novel perspective of fine-grained invariant relationship learning, which captures subtle yet crucial patterns across distributed domains. We propose FedDIM, a simple yet effective method that enhances OOD generalization by leveraging insight matrices to distill domain-invariant relationship from heterogeneous client data. Our theoretical analysis shows that FedDIM can effectively generalize to unseen domains by maintaining consistent relationships across different distributions, while empirical results demonstrate state-of-the-art performance on standard federated domain generalization benchmarks, including RotatedMNIST, PACS, VLCS and OfficeHome datasets.

Acknowledgments

The research is supported bythe National Key R&D Program of China (2023YFB2703700), the National Natural Science Foundation of China (62176269).

Contribution Statement

Tianchi Liao and Binghui Xie contribute equally to this work.

References

[Bai *et al.*, 2024] Sikai Bai, Jie Zhang, Song Guo, Shuaicheng Li, Jingcai Guo, Jun Hou, Tao Han, and Xiaocheng Lu. Diprompt: Disentangled prompt tuning for multiple latent domain generalization in federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27284–27293, 2024.

[Ben-David *et al.*, 2010] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79:151–175, 2010.

[Chen *et al.*, 2023] Liang Chen, Yong Zhang, Yibing Song, Anton Van Den Hengel, and Lingqiao Liu. Domain generalization via rationale invariance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1751–1760, 2023.

[Chen *et al.*, 2025] Chuan Chen, Tianchi Liao, Xiaojun Deng, Zihou Wu, Sheng Huang, and Zibin Zheng. Advances in robust federated learning: A survey with heterogeneity considerations. *IEEE Transactions on Big Data*, 2025.

[Fang *et al.*, 2013] Chen Fang, Ye Xu, and Daniel N Rockmore. Unbiased metric learning: On the utilization of multiple datasets and web images for softening bias. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1657–1664, 2013.

[Fu *et al.*, 2025a] Lele Fu, Sheng Huang, Yanyi Lai, Tianchi Liao, Chuanfu Zhang, and Chuan Chen. Beyond federated prototype learning: Learnable semantic anchors with hyperspherical contrast for domain-skewed data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 16648–16656, 2025.

[Fu *et al.*, 2025b] Lele Fu, Sheng Huang, Yanyi Lai, Chuanfu Zhang, Hong-Ning Dai, Zibin Zheng, and Chuan Chen. Federated domain-independent prototype learning with alignments of representation and parameter spaces for feature shift. *IEEE Transactions on Mobile Computing*, pages 1–16, 2025.

[Garipov *et al.*, 2018] Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry P Vetrov, and Andrew G Wilson. Loss surfaces, mode connectivity, and fast ensembling of dnns. *Advances in neural information processing systems*, 31, 2018.

[Ghifary *et al.*, 2015] Muhammad Ghifary, W Bastiaan Kleijn, Mengjie Zhang, and David Balduzzi. Domain generalization for object recognition with multi-task autoencoders. In *Proceedings of the IEEE international conference on computer vision*, pages 2551–2559, 2015.

[Gulrajani and Lopez-Paz, 2020] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*, 2020.

[Guo *et al.*, 2023] Yaming Guo, Kai Guo, Xiaofeng Cao, Tieru Wu, and Yi Chang. Out-of-distribution generalization of federated learning via implicit invariant relationships. In *International Conference on Machine Learning*, pages 11905–11933. PMLR, 2023.

[Hu *et al.*, 2023] Ming Hu, Zeke Xia, Dengke Yan, Zhihao Yue, Jun Xia, Yihao Huang, Yang Liu, and Mingsong Chen. Gitfl: Uncertainty-aware real-time asynchronous federated learning using version control. In *In Proceedings of IEEE Real-Time Systems Symposium (RTSS)*, pages 145–157. IEEE, 2023.

[Hu *et al.*, 2024] Ming Hu, Peiheng Zhou, Zhihao Yue, Zhiwei Ling, Yihao Huang, Anran Li, Yang Liu, Xiang Lian, and Mingsong Chen. Fedcross: Towards accurate federated learning via multi-model cross-aggregation. In *IEEE International Conference on Data Engineering (ICDE)*, pages 2137–2150. IEEE, 2024.

[Huang *et al.*, 2023] Wenke Huang, Mang Ye, Zekun Shi, He Li, and Bo Du. Rethinking federated learning with domain shift: A prototype view. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16312–16322. IEEE, 2023.

[Huang *et al.*, 2025] Sheng Huang, Lele Fu, Yuecheng Li, Chuan Chen, Zibin Zheng, and Hong-Ning Dai. A cross-client coordinator in federated learning framework for conquering heterogeneity. *IEEE Transactions on Neural Networks and Learning Systems*, 36(5):8828–8842, 2025.

[Li *et al.*, 2017] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pages 5542–5550, 2017.

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

[Li *et al.*, 2020b] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2:429–450, 2020.

[Li *et al.*, 2024] Yichen Li, Wenchao Xu, Haozhao Wang, Yining Qi, Ruixuan Li, and Song Guo. Sr-fdil: Synergistic replay for federated domain-incremental learning. *IEEE Transactions on Parallel and Distributed Systems*, 2024.

[Liao *et al.*, 2024] Tianchi Liao, Lele Fu, Jialong Chen, Zhen Wang, Zibin Zheng, and Chuan Chen. A swiss army knife for heterogeneous federated learning: Flexible coupling via trace norm. *Advances in Neural Information Processing Systems*, 37:139886–139911, 2024.

[Liao *et al.*, 2025] Tianchi Liao, Lele Fu, Lei Zhang, Lei Yang, Chuan Chen, Michael K Ng, Huawei Huang, and Zibin Zheng. Privacy-preserving vertical federated learning with tensor decomposition for data missing features. *IEEE Transactions on Information Forensics and Security*, 2025.

[McMahan *et al.*, 2017] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.

[Nguyen *et al.*, 2022] A Tuan Nguyen, Philip Torr, and Ser Nam Lim. Fedsr: A simple and effective domain generalization method for federated learning. *Advances in Neural Information Processing Systems*, 35:38831–38843, 2022.

[Qi *et al.*, 2024] Zhuang Qi, Weihao He, Xiangxu Meng, and Lei Meng. Attentive modeling and distillation for out-of-distribution generalization of federated learning. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2024.

[Qi *et al.*, 2025] Zhuang Qi, Lei Meng, and et al. Cross-silo feature space alignment for federated learning on clients with imbalanced data. In *The 39th Annual AAAI Conference on Artificial Intelligence (AAAI-25)*, pages 19986–19994, 2025.

[Qiao *et al.*, 2024] Yu Qiao, Chaoning Zhang, Apurba Adhikary, and Choong Seon Hong. Logit calibration and feature contrast for robust federated learning on non-iid data. *arXiv preprint arXiv:2404.06776*, 2024.

[Shen *et al.*, 2025] Wei Shen, Wenke Huang, Guancheng Wan, and Mang Ye. Label-free backdoor attacks in vertical federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 20389–20397, 2025.

[Van der Maaten and Hinton, 2008] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[Venkateswara *et al.*, 2017] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5018–5027, 2017.

[Wan *et al.*, 2024] Guancheng Wan, Wenke Huang, and Mang Ye. Federated graph learning under domain shift with generalizable prototypes. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 15429–15437, 2024.

[Xie *et al.*, 2024] Binghui Xie, Yongqiang Chen, Jiaqi Wang, Kaiwen Zhou, Bo Han, Wei Meng, and James Cheng. Enhancing evolving domain generalization through dynamic latent representations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 16040–16048, 2024.

[Yan and Guo, 2025] Hao Yan and Yuhong Guo. Local and global flatness for federated domain generalization. In *European Conference on Computer Vision*, pages 71–87. Springer, 2025.

[Zhang *et al.*, 2021] Liling Zhang, Xinyu Lei, Yichun Shi, Hongyu Huang, and Chao Chen. Federated learning with domain generalization. *arXiv preprint arXiv:2111.10487*, 2021.

[Zhang *et al.*, 2023] Ruipeng Zhang, Qinwei Xu, Jiangchao Yao, Ya Zhang, Qi Tian, and Yanfeng Wang. Federated domain generalization with generalization adjustment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3954–3963, 2023.

[Zhang *et al.*, 2024] Yudong Zhang, Xu Wang, Pengkun Wang, Binwu Wang, Zhengyang Zhou, and Yang Wang. Modeling spatio-temporal mobility across data silos via personalized federated learning. *IEEE Transactions on Mobile Computing*, 2024.