FedCM: Client Clustering and Migration in Federated Learning via Gradient Path Similarity and Update Direction Deviation

Peng Wang[1] , **Shoupeng Lu**[2] , **Hao Yin**[1] , **Banglie Yang**[2] , **Tianli Zhu**[2] , **Cheng Dai**[2,*]

[1]College of Software Engineering, Sichuan University, China
[2]College of Computer Science, Sichuan University, China
{wangpeng2, lushoupeng, yinhao, yangbanglie, zhutianli}@stu.scu.edu.cn, daicheng@scu.edu.cn

Abstract

Federated learning (FL) enables collaborative training among multiple clients while preserving data privacy. However, its practical application is significantly limited by two major challenges: statistical heterogeneity and data distribution drift. Statistical heterogeneity causes the direction of local model updates to deviate from the global training objective, while data distribution drift leads to a mismatch between local models and their cluster models. To address these challenges, this paper proposes an adaptive clustered federated learning framework, Fed-CM. Initially, by capturing the dynamic patterns of personalized layer parameters in clients' models, Fed-CM effectively characterizes the correlations and distributional similarities among clients, reflecting the underlying statistical heterogeneity. Subsequently, this framework leverages client similarities to construct an undirected graph and adaptively performs effective cluster discovery with minimal dependence on hyperparameters. Furthermore, a monitoring strategy tracks the deviation between clients' update directions and the dominant update direction of their clusters and then adaptively migrates clients experiencing data drift. Such a dynamic strategy helps maintain intra-cluster homogeneity and addresses the mismatch between local models and their cluster models. Compared to other state-of-the-art methods, experimental results on multiple datasets demonstrate that the proposed Fed-CM framework effectively addresses the challenges posed by statistical heterogeneity and data drift, significantly improving the performance and robustness of federated learning models.

1 Introduction

Federated learning (FL) [McMahan *et al.*, 2017] is a pivotal distributed machine learning paradigm that enables collaborative training among multiple clients while preserving data privacy, thus demonstrating significant potential in areas such as healthcare, financial risk control, and intelligent transportation. However, the practical deployment of FL faces two major challenges: **statistical heterogeneity** and **data distribution drift**. Specifically, statistical heterogeneity refers to the non-independent and identically distributed (Non-IID) nature of data across clients, which results in significant differences in data distributions among clients, consequently affecting the convergence speed and performance of the global model [Ma *et al.*, 2022a]. Meanwhile, data distribution drift refers to the changes in data distributions of clients over time, which causes local models to become mismatched with their respective cluster models, thus impairing the training effectiveness of cluster models and potentially leading to model divergence and complete breakdown [Li *et al.*, 2024a].

Recently, Clustered federated learning (CFL) [Ghosh *et al.*, 2022] has been introduced to tackle statistical heterogeneity in federated learning. The core concept of CFL is to partition clients into different clusters based on the similarity of their data distributions and train a shared model within each cluster. By doing so, CFL leverages intra-cluster similarity to improve local model training and accelerate global model convergence. CFL is particularly effective in scenarios with significant disparities in clients' local data distributions.

However, existing CFL methods still suffer from the following limitations: 1) *limited similarity characterization*: current methods typically measure client similarity based on static model parameters or feature representations, failing to capture the adaptation of clients on their respective datasets. This leads to the inability to accurately capture the inherent differences among clients caused by statistical heterogeneity, thus affecting clustering effectiveness [Beltrán *et al.*, 2023]. 2) *heavy reliance on prior knowledge*: existing CFL methods heavily rely on prior knowledge of clustering hyperparameters, such as the number of clusters or distance thresholds, which requires extensive hyperparameter tuning and limits the flexibility and scalability. 3) *lack of dynamic migration*: conventional methods lack effective client dynamic migration mechanisms to address data distribution drift. Although some methods have attempted to use the Wasserstein distance to measure data distribution differences for client migration, they require access to clients' local data, posing privacy risks and involving high computational complexity [Duan *et al.*, 2022]. These limitations raise a critical question: ***how can we accurately characterize similarities in data distributions***

*Corresponding authors.

across clients while adaptively achieving client clustering and dynamically responding to data distribution drift?

In this paper, we provide an answer to this key question and propose an efficient and adaptive federated learning framework: Fed-CM. The framework comprises three key components: 1) a client similarity calculation method based on personalized layer gradient paths, which captures the adaptation process and inherent similarities of clients during local training; 2) an adaptive graph-based cluster discovery algorithm, which constructs an undirected graph using the client similarity matrix and enables the adaptive discovery of cluster structures; and 3) a client migration strategy based on update angles, which adaptively identifies and migrates clients with data distribution drift by detecting the deviation of client update angles.

Based on the above design, the Fed-CM framework has the following advantages: 1) it enables dynamic modeling of client similarities through personalized layer gradient paths, thereby capturing the differences in client data distributions more accurately, leading to higher clustering accuracy and model performance; 2) the adaptive cluster discovery algorithm effectively avoids strong reliance on hyperparameters, enabling more flexible and efficient clustering; 3) the dynamic migration mechanism based on update angles addresses data distribution drift in a timely and effective manner, improving model stability; and 4) the calculation of both similarity and update angles only requires clients to upload gradients, protecting the privacy of client data. Experimental results on multiple datasets demonstrate that the Fed-CM framework can effectively handle federated learning scenarios with statistical heterogeneity and data distribution drift. Specifically, the model performance on the CIFAR-10 dataset improves by 2 percentage points over the state-of-the-art, while in scenarios with data distribution drift, it exhibits strong robustness, with a performance degradation of merely 1‰ to its peak accuracy.

2 Related Work

2.1 Personalized Federated Learning

To address the challenge of ineffective convergence of the global model in FedAvg [McMahan *et al.*, 2017] under data heterogeneity, researchers have proposed Personalized Federated Learning (PFL). PFL aims to train customized models for each client, thereby achieving better adaptation to local data distributions. FedPer [Arivazhagan *et al.*, 2019] achieves personalization by decoupling the model and employing distinct aggregation strategies. Per-FedAvg [Fallah *et al.*, 2020] incorporates the concept of meta-learning, using fine-tuning to realize personalized models. pFedMe [T Dinh *et al.*, 2020] utilizes the Moreau envelope function to better decouple the optimization of global and personalized models. pFedHN [Shamsian *et al.*, 2021] leverages hypernetworks to generate personalized models for each client, reducing communication costs. More recently, FedSelect [Tamirisa *et al.*, 2024] achieves personalization of parameters and structures by dynamically expanding personalized sub-networks. RIPFL [Qin *et al.*, 2023] selects and partitions clients from a social learning perspective, integrating individual and global information. FedAS [Yang *et al.*, 2024] addresses inconsistencies through parameter alignment and client synchronization strategies. pFedFDA [Mclaughlin and Su, 2024] treats representation learning as a generative modeling task, generating personalized models based on local feature distributions.

2.2 Clustered Federated Learning

In contrast to PFL, which focuses on customizing models for each client, Clustered Federated Learning (CFL) is dedicated to partitioning similar clients into distinct clusters and conducting more effective model training within each cluster, thereby mitigating the impact of data heterogeneity. The IFCA [Ghosh *et al.*, 2022] algorithm performs clustering by iteratively estimating the cluster identities of clients and optimizing the model parameters of the clusters. The FMTL [Sattler *et al.*, 2021] framework groups clients based on the geometric properties of the federated learning loss surface. FL+HC [Briggs *et al.*, 2020] introduces a hierarchical clustering step, clustering clients based on the similarity of their local updates. PACFL [Vahidian *et al.*, 2023] identifies distributional similarities between clients by analyzing the principal angles of the client data subspaces. FlexCFL [Duan *et al.*, 2022] groups clients based on the similarity of their optimization directions and supports flexible client migration to address data distribution drift. FedCCFA [Chen *et al.*, 2024], designed for data heterogeneity under distributed concept drift, alleviates feature space inconsistencies using classifier clustering and feature alignment. CFL-Gb [Kim *et al.*, 2024] achieves robust clustering and learning performance by clustering clients based on the similarity of their model updates.

3 Preliminaries

3.1 Federated Learning

Federated Learning (FL) aims to train a global model $f(\cdot;\theta)$ parameterized by $\theta \in \mathbb{R}^d$ across N clients, each holding a local dataset $\mathcal{D}_k \sim P_k$, without sharing these datasets. Clients perform local training using SGD and the server aggregates their updates, typically using FedAvg. The overall goal is to minimize the global loss:

$$\min_{\theta} \sum_{k=1}^{N} p_k \mathbb{E}_{(x,y)\sim\mathcal{D}_k}[\ell(f(x;\theta),y)]. \tag{1}$$

Here, ℓ is the loss function and p_k is the weight of client k. At each round, the server selects a subset of clients, distributes the current global model, and aggregates their updated models. Local training typically involves multiple epochs of SGD: $\theta_k^{(t,e+1)} = \theta_k^{(t,e)} - \eta\nabla_\theta \ell(f(x_i;\theta_k^{(t,e)}), y_i)$, where (x_i, y_i) is a data sample (or mini-batch) drawn from $\mathcal{D}_k$, $\theta_k^{(t,1)}$ is initialized with the global model, and η is the learning rate.

3.2 Problem Definition

Statistical Heterogeneity

Statistical heterogeneity stems from the discrepancies in client data distributions. Different clients may have data

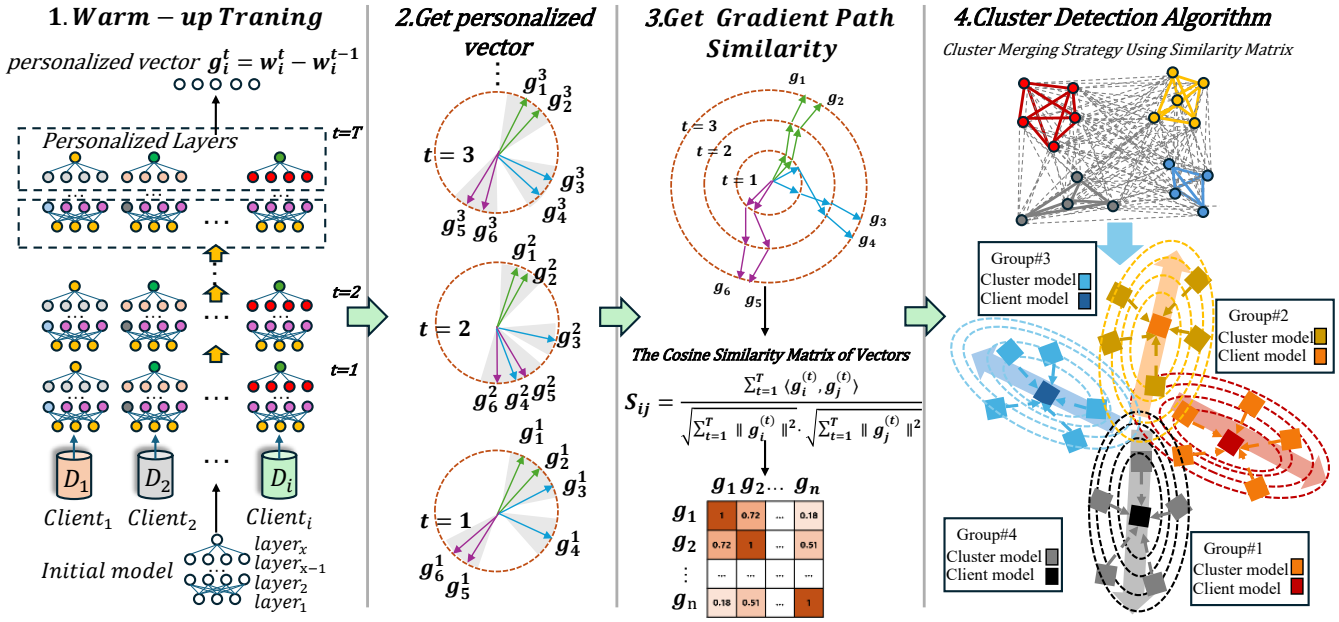


Figure 1: Overview of the Personalized Gradient Path Similarity Framework: (1) Warm-up training to obtain personalized gradient vectors; (2) Visualizing personalized gradient vectors over training iterations; (3) Calculating Gradient Path Similarity using cosine similarity; (4) Cluster Detection Algorithm for dynamic client grouping based on similarity.

drawn from different distributions, causing the directions of their local model updates to deviate from the global training objective. Formally, assume N clients each possess a dataset $\mathcal{D}_k \sim P_k$, where $k \in \{1, 2, \ldots, N\}$. Statistical heterogeneity implies that the data distributions are not identical, i.e., there exist $i \neq j$ such that $P_i \neq P_j$. This discrepancy can lead to deviations between the optimal solutions of each client's local loss function $\mathcal{L}_k(\theta)$ and the global loss function $\mathcal{L}(\theta) = \sum_{k=1}^{N} p_k \mathcal{L}_k(\theta)$.

Data Distribution Drift

Data distribution drift refers to the phenomenon where the data distribution of a client changes during the training process. Recent studies have investigated this issue in the context of federated learning [Li *et al.*, 2024b; Wang *et al.*, 2024a; Wang *et al.*, 2024b; Zhou *et al.*, 2024]. This drift can be caused by various factors, such as changes in user behavior or environment. Formally, the data distribution of client k at time t is denoted as $P_k^{(t)}$. Data distribution drift implies that for some client k, there exist time steps t_1 and t_2 such that $P_k^{(t_1)} \neq P_k^{(t_2)}$. This can lead to a mismatch between a client's local model and the model of its cluster, affecting the performance of the global model. To address this issue, Fed-CM introduces an update angle-based client migration mechanism to detect data distribution drift and migrate affected clients to more appropriate clusters.

4 Methodology

Our framework consists of two key stages: (1) One-Shot Initial Clustering: A single clustering step performed before federated training to establish an initial client grouping using our personalized gradient path similarity and cluster detection algorithm. (2) Migration Strategy: Employed during subsequent training rounds to refine the initial clustering by adjusting a few outlier clients.

4.1 Personalized Gradient Path Similarity

To more accurately capture the inherent differences in client data distributions, we propose a gradient-path-based dynamic similarity metric. The core idea of this method is to track the evolution trajectory of model parameters during local client training to capture their intrinsic correlation under specific data distributions, thereby quantifying the dynamic similarity between clients. Unlike previous studies that directly use overall model parameters for similarity measurement, we focus on the personalized layer parameters of the model, which more effectively reflect client data characteristics. A significant body of research in representation learning has demonstrated that the classification layer of the model can effectively reflect the individuality of the model [Li *et al.*, 2023; Kang *et al.*, 2020; Hu *et al.*, 2023; OH *et al.*, 2022; Zhang *et al.*, 2024; Chen *et al.*, 2023; Xie *et al.*, 2024; Ma *et al.*, 2022b; Luo and Wu, 2022; Liu *et al.*, 2023; Yi *et al.*, 2024]. Therefore, we utilize the classification layer as the personalized layer to construct our gradient-path similarity metric.

As depicted in Figure 1, we first distribute the same initial model parameters to every client. Following this, each client performs warm-up training on their respective local dataset. During training, we record the gradient vector of each client's personalized layer parameters at every iteration and concatenate these gradient vectors to create a gradient vector path. Specifically, for client i, we record the gradient paths of L in-

dividual personalized layer parameters after local training T rounds. Let $(\mathbf{w}_{i,l}^{(t)})$ denote the parameter of the l th personalization layer of client i after the t round of local training, then the vector of gradient paths is denoted as:

$$\Delta \mathbf{w}_{i,l}^{(t)} = \mathbf{w}_{i,l}^{(t)} - \mathbf{w}_{i,l}^{(t-1)}. \tag{2}$$

Stack the gradient path vectors of all personalized layers of client i into a gradient path matrix ($\mathbf{P}_i$), denoted as:

$$\mathbf{P}_i = \begin{bmatrix} \Delta \mathbf{w}_{i,1}^{(1)} & \cdots & \Delta \mathbf{w}_{i,L}^{(1)} \\ \vdots & \ddots & \vdots \\ \Delta \mathbf{w}_{i,1}^{(T)} & \cdots & \Delta \mathbf{w}_{i,L}^{(T)} \end{bmatrix}. \tag{3}$$

To quantify the degree of similarity between clients, we expand the gradient path matrix ($\mathbf{P}_i$) and use cosine similarity to quantify the gradient path similarity (s_{ij}) between clients i and j:

$$s_{ij} = \frac{\sum_{t=1}^{T} \sum_{l=1}^{L} \Delta \mathbf{w}_{i,l}^{(t)} \cdot \Delta \mathbf{w}_{j,l}^{(t)}}{\sqrt{\sum_{t=1}^{T} \sum_{l=1}^{L} ||\Delta \mathbf{w}_{i,l}^{(t)}||_2^2} \sqrt{\sum_{t=1}^{T} \sum_{l=1}^{L} ||\Delta \mathbf{w}_{j,l}^{(t)}||_2^2}}. \tag{4}$$

4.2 Cluster Detection Algorithm

To address the challenge that existing clustering federation learning methods rely on strong hyperparameters and are difficult to generalize to different scenarios, we design an adaptive cluster discovery algorithm inspired by heuristic community discovery algorithms. The algorithm aims to adaptively classify clients into clusters with similar data distributions without the need for presetting hyperparameters, thus improving the flexibility and robustness [Ghosh *et al.*, 2019] of clustering federation learning.

Our algorithm first constructs an undirected graph $G = (V, E)$ based on the similarity of the clients, where the clients are the nodes V of the graph, and the similarity between the clients is the weights of the edges E. In the initial state, each client independently forms a cluster. Next, the algorithm focuses on clustering clients into groups where the similarity is high and the overall modularity can be improved. Specifically, the algorithm iteratively considers moving a client to other clusters. If a client has high similarity with the clients in the target cluster, and moving to that cluster increases the overall modularity, then the move is accepted. Through this iterative optimization process, the algorithm spontaneously aggregates clients that are similar and contribute to a higher modularity.

To quantify the clustering effect, we use the modularity Q as an evaluation metric, which is calculated as:

$$Q = \frac{1}{2m} \sum_{c \in C} \left(\sum_{i,j \in c} w_{ij} - \frac{1}{2m} \left(\sum_{i \in c} k_i \right)^2 \right), \tag{5}$$

where $\mathcal{C}$ denotes the set of all clusters, c denotes the current cluster, w_{ij} denotes the weight of the edges between nodes i and j in the cluster c, k_i denotes the degree of node i, and m is the total weight of all the edges in the graph. The detailed process of our algorithm is shown in Algorithm 1.

Algorithm 1: Adaptive Graph Clustering Algorithm

Input: Client similarity graph $G = (V, E)$
Output: Cluster partition $\mathcal{C}^*$
Initialization: $\mathcal{C} \leftarrow \{\{v\} | v \in V\}$,
$Q(\mathcal{C}) = \frac{1}{2m} \sum_{c \in \mathcal{C}} \left(\sum_{i,j \in c} w_{ij} - \frac{1}{2m} \left(\sum_{i \in c} k_i \right)^2 \right)$;
while *changed* $\leftarrow$ *false; true* **do**
  for $v_i \in V$ **do**
    Find $c^* = \arg\max_{c_j} \Delta Q(v_i, c_j)$, where $\Delta Q(v_i, c_j) = Q(\mathcal{C}_{new}) - Q(\mathcal{C})$;
    if $\Delta Q(v_i, c^*) > 0$ **then**
      Move v_i to c^*, *changed* $\leftarrow$ true, Update $\mathcal{C}$, Update $Q(\mathcal{C})$;
    end
  end
end
return Optimal cluster partition $\mathcal{C}^*$

4.3 Migration Strategy Based on Update Angles

To address the non-stationarity of client data distributions over time in federated learning scenarios, we propose an adaptive client migration strategy based on update angles. The core idea of this strategy is to monitor the dynamic changes in client model parameter update directions, quantify the degree of deviation from the typical pattern represented by their assigned cluster, and subsequently migrate clients with significant deviations to clusters with more fitting feature distributions (see Figure 2). This enables the federated learning framework to dynamically adapt to the evolving data landscape. Specifically, for each client i participating in federated learning, during each global iteration t, we calculate the update angle of its personalized layer parameters as the normalized difference between the parameters after and before the current iteration:

$$\boldsymbol{\theta}_i^t = \frac{\mathbf{w}_i^t - \mathbf{w}_i^{t-1}}{||\mathbf{w}_i^t - \mathbf{w}_i^{t-1}||_2}. \tag{6}$$

Subsequently, for each cluster c, we compute its global average update direction $\bar{\boldsymbol{\theta}}_c^t$, which is defined as the normalized weighted average of the update angles of all clients within that cluster, with the weights being the L2 norm of each client's parameter update magnitude:

$$\bar{\boldsymbol{\theta}}_c^t = \frac{\sum_{j \in C_c^t} ||\Delta \mathbf{w}_j^t||_2 \cdot \boldsymbol{\theta}_j^t}{||\sum_{j \in C_c^t} ||\Delta \mathbf{w}_j^t||_2 \cdot \boldsymbol{\theta}_j^t||_2}. \tag{7}$$

To accurately measure the deviation between a client's update direction and the representative direction of its assigned cluster, we calculate the cosine similarity between the client's update angle $\boldsymbol{\theta}_i^t$ and the cluster's average update direction $\bar{\boldsymbol{\theta}}_c^t$:

$$S(\boldsymbol{\theta}_i^t, \bar{\boldsymbol{\theta}}_c^t) = (\boldsymbol{\theta}_i^t)^T \bar{\boldsymbol{\theta}}_c^t. \tag{8}$$

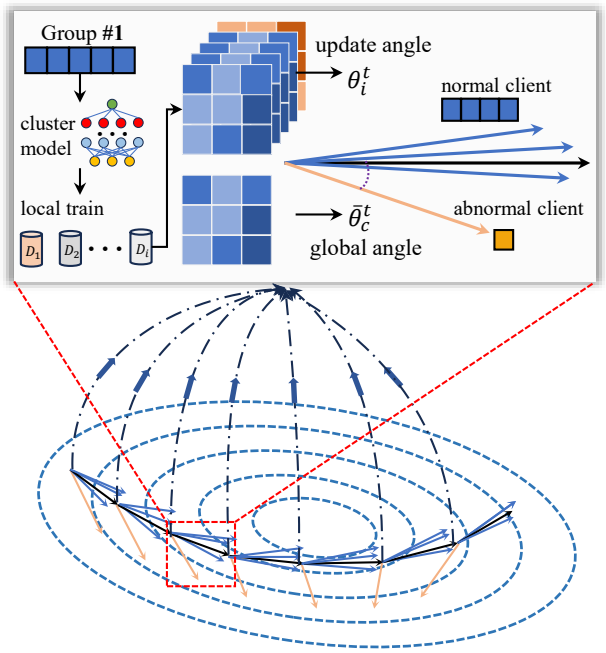


Figure 2: Illustration of the client migration strategy based on update angles.

When this similarity falls below a predefined threshold τ, we identify the client as an outlier, indicating a potential significant drift in its data distribution. Having identified an outlier client, we employ a computationally efficient greedy strategy to migrate this client to the cluster whose average update direction is most similar to its current update direction. Specifically, client i will be reassigned to cluster c', where

$$c' = \arg \max_{k \in \{1,2,\ldots,K\}} (\boldsymbol{\theta}_i^t)^T \bar{\boldsymbol{\theta}}_k^t. \tag{9}$$

Here, K represents the total number of clusters. It is worth emphasizing that this entire migration process relies solely on the gradient information uploaded by clients, without requiring access to their local sensitive data, thus ensuring client data privacy. Through this adaptive migration mechanism, we aim to dynamically maintain the homogeneity of data distributions within clusters, effectively enhancing the robustness and generalization performance of the federated learning framework in the face of statistical heterogeneity and data drift challenges.

5 Experiments

5.1 Experimental Setup

Datasets and Model

For a fair comparison with existing work, we followed the experimental setup in the PACFL[Vahidian *et al.*, 2023] and used four commonly used image classification datasets: Fashion-MNIST (FMNIST) [Xiao *et al.*, 2017], SVHN [Netzer *et al.*, 2011], CIFAR-10 [Krizhevsky, 2009] and CIFAR-100 [Krizhevsky, 2009]. These datasets cover a wide range of image complexity and number of categories and are able to fully validate the Fed-CM framework in various scenarios. We use the classical LeNet-5 [LeCun *et al.*, 1998] as the model architecture.

Baselines

To fully evaluate the performance of the Fed-CM framework, we compare it with a series of representative federated learning algorithms. These baselines include several global methods, such as the canonical FedAvg [McMahan *et al.*, 2017] and its variants aimed at handling heterogeneity like FedProx [Li *et al.*, 2020], FedNova [Wang *et al.*, 2020], and Scaffold [Karimireddy *et al.*, 2020]. We also benchmark against prominent personalized approaches, including FedPer [Arivazhagan *et al.*, 2019], Per-FedAvg [Fallah *et al.*, 2020], and pFedMe [T Dinh *et al.*, 2020]. Finally, we conduct extensive comparisons with state-of-the-art clustered federated learning methods, namely IFCA [Ghosh *et al.*, 2022], PACFL [Vahidian *et al.*, 2023], FedCCFA [Chen *et al.*, 2024], and CFL-Gb [Kim *et al.*, 2024]. In our experimental setup, we assume that there are 100 clients and 10 clients are randomly selected to participate in the training in each round, with the global round number set to 200, and 10 local training epochs for each selected client.

5.2 Overall Performance

To comprehensively evaluate the effectiveness and robustness of the proposed Fed-CM framework in addressing the challenges of statistical heterogeneity and data distribution drift, we conducted extensive experiments under diverse settings. ***For statistical heterogeneity***, we adopted the standard Non-IID label skew and a more challenging pathological label skew setting, where clients possess mutually exclusive label sets, to simulate the complex heterogeneity arising in real-world scenarios. In addition, we visually analyzed the distinctions between gradient path similarity and other similarity metrics, and benchmarked our clustering algorithm against clustering methods with strong hyperparameter dependency. ***For data distribution drift***, we simulated scenarios where client data distributions evolve over time and evaluated the performance degradation mitigation of Fed-CM compared to other methods employing client migration strategies. The experimental results demonstrate that Fed-CM consistently outperforms state-of-the-art baselines across various settings, highlighting its effectiveness and robustness in handling both statistical heterogeneity and data distribution drift.

5.3 Performance in Statistical Heterogeneity

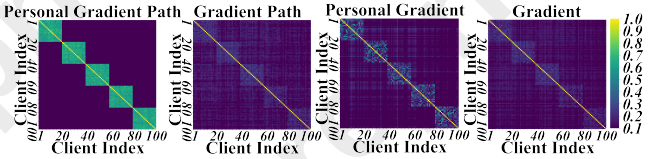
This section provides an in-depth analysis of the experimental results under statistical heterogeneity. We begin by introducing the experimental setup: In the label skew setting, each client can only access $\varrho\%$ of the total label classes. Under the standard label skew, data is randomly distributed to clients. In contrast, under the pathological label skew, the data labels accessible to clients are mutually exclusive (orthogonal). We conducted experiments for both settings with $\varrho = 20\%$. Table 1 presents the experimental results for both pathological and standard label skew at $\varrho = 20\%$. Figure 4 illustrates the accuracy of different methods over training rounds under the pathological label skew setting with $\varrho = 20\%$.

Algorithm	CIFAR-10		FMNIST		CIFAR-100		SVHN	
	Pathological	Random	Pathological	Random	Pathological	Random	Pathological	Random
FedAvg	38.30	41.31	81.13	85.81	22.61	24.07	79.90	82.69
Fedprox	45.27	49.77	77.50	85.70	22.61	24.43	78.18	84.07
Fednova	45.65	48.88	77.58	85.50	22.48	24.46	78.76	84.35
scaffold	30.29	33.41	69.50	81.30	28.75	30.40	51.31	63.30
FedLG	81.77	84.51	99.07	97.75	34.06	34.59	93.85	92.94
Per-FedAvg	85.41	88.05	98.87	97.20	44.65	47.91	95.67	95.65
pFedMe	86.31	87.43	99.34	97.77	39.41	37.94	96.41	94.34
IFCA	88.37	87.73	99.61	98.39	50.48	46.47	97.14	96.27
FlexCFL	83.81	85.43	94.49	96.86	34.25	34.43	94.49	93.18
PACFL	88.91	87.63	99.19	98.47	49.84	36.63	97.14	96.42
FedCCFA	79.44	82.32	97.76	94.95	27.81	29.27	88.76	90.14
CFL-Gb	84.05	86.07	99.16	97.20	31.27	31.90	95.46	94.39
Fed-CM	**90.94**	**89.97**	**99.65**	**98.56**	**50.72**	**48.08**	**97.48**	**96.42**

Table 1: Test accuracy comparison across different datasets (CIFAR-10, FMNIST, CIFAR-100, and SVHN) and Non-IID settings (pathological and random label skew with $\varrho = 20\%$). For each algorithm, the average of final local test accuracy over all clients is reported. Each algorithm was run 200 communication rounds, with 10 local epochs per round for selected clients. The best and second-best results are highlighted in dark gray and light gray, respectively.

Algorithm	Performance					Best
K-means	N = 2 89.32	N = 3 90.43	N = 4 90.58	N = 5 90.94	N = 6 90.41	90.94
Hierarchical	L = 0.3 81.71	L = 0.5 83.16	L = 0.7 87.89	L = 0.9 90.84	L = 1 89.10	90.84
Fed-CM	-	-	-	-	-	90.94

Table 2: Performance comparison of clustering algorithms on the CIFAR-10 dataset under the pathological setting with $\varrho = 20\%$.



(a) (b) (c) (d)

Figure 3: Visualization of different similarity metrics on the CIFAR-10 dataset under the pathological label skew setting with $\varrho = 20\%$.

As can be seen from Table 1, the global federated learning methods FedAvg, FedProx, FedNova, and SCAFFOLD perform poorly in heterogeneous data scenarios. This strongly suggests that under statistical heterogeneity, local models deviate from the global optimization objective, leading to difficulties in the convergence of the global model and, consequently, low accuracy. Among the clustered federated learning algorithms, our method, Fed-CM, also outperforms other current SOTA methods. Notably, on the CIFAR-10 dataset with the pathological label skew setting at $\varrho = 20\%$, Fed-CM achieves an accuracy of 90.94%, surpassing IFCA, FlexCFL, PACFL, FedCCFA, and CFL-Gb by +2.6%, +7.1%, +2%, +11.5%, and +6.9%, respectively. This significant improvement can be attributed to our gradient path similarity, which more accurately captures client relationships, and our adaptive clustering algorithm, which discovers the optimal grouping without manual hyperparameter tuning. Figure 4 demonstrates that Fed-CM exhibits superior convergence speed and final accuracy compared to all baseline models.

5.4 Visualization Analysis of Gradient Similarities

We further demonstrate the superiority of the personalized layer gradient path similarity used in Fed-CM through visual comparisons with other similarity metrics. As shown in Figure 3, the personalized layer gradient path similarity (Figure 3a) exhibits clearer and more stable boundaries compared to other metrics. Non-personalized layer parameters, whether gradients or gradient path (Figure 3c and Figure 3d), fail to capture the inherent differences between client data distributions caused by statistical heterogeneity, displaying a nearly random pattern of low similarity. While personalized layer gradients (Figure 3b) show some ability to differentiate clients, the boundaries are blurred and lack clarity. In contrast, the personalized layer gradient path similarity employed by Fed-CM effectively captures the adaptation process of client models on diverse data distributions during local training. This results in a more accurate representation of the intrinsic similarities in data distributions across clients, as evidenced by the distinct clusters and sharp boundaries in Figure 3a. This advantage enables more effective client clustering and ultimately leads to improved model performance.

5.5 Analysis of Clustering Algorithms

In addition to the visualization analysis, we also compared the clustering algorithm used in Fed-CM with other clustering algorithms that rely on pre-defined parameters. Table 2 presents the performance comparison of different clustering algorithms on the CIFAR-10 dataset under the pathological label skew setting with $\varrho = 20\%$. We compared the K-means algorithm (which requires pre-setting the number of clusters, N) and the hierarchical clustering algorithm (which requires pre-setting the distance threshold, L). For a fair comparison, we tested the performance of these algorithms under different parameter settings and reported their best performance.

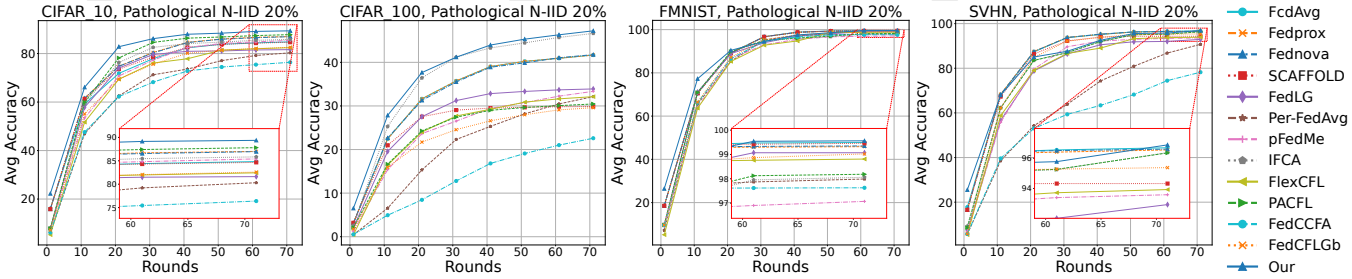


Figure 4: Test accuracy versus the number of communication rounds for our method and baseline methods under a 20% pathological label skew on the CIFAR-10, CIFAR-100, FMNIST, and SVHN datasets.

Algorithm	Test Acc. (Stable)	Test Acc. After drift			Δ (‰)
		(R60)	(R61)	(R200)	
IFCA	88.37	85.40	69.68	87.95	5
FlexCFL	83.81	81.93	50.75	83.44	5
PACFL	88.91	87.16	53.82	86.21	30
FedCCFA	79.44	75.37	48.24	79.11	4
CFL-Gb	86.50	86.00	80.00	85.50	12
Fed-CM*	90.94	88.87	74.88	88.42	27
Fed-CM	90.94	89.09	73.53	90.82	**1**

Table 3: Test accuracy comparison under data distribution drift and impact analysis. "Fed-CM*" denotes our method without the migration mechanism.

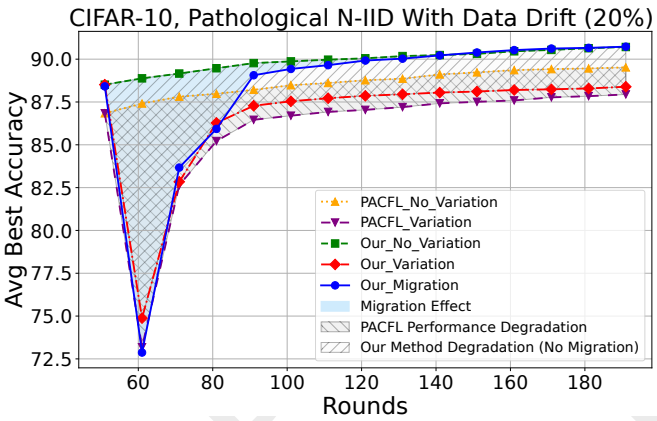


Figure 5: Fed-CM vs. PACFL under 20% drift on CIFAR-10.

The experimental results show that the graph-based adaptive clustering algorithm used in Fed-CM achieves an accuracy of 90.94% without requiring any pre-set parameters, which is on par with the best performance of K-means (N=5) and outperforms the best performance of hierarchical clustering (L=0.9). This fully demonstrates the effectiveness and adaptability of the clustering algorithm used in Fed-CM, avoiding the tedious parameter tuning process and enhancing the practicality and scalability of the algorithm.

5.6 Performance in Data Distribution Drift

This section delves into the performance of the proposed Fed-CM framework and baseline methods under data distribution drift. Table 3 presents a comparative analysis of test accuracy before and after the introduction of a data distribution drift. To visually complement these quantitative results, Figure 5 illustrates the accuracy trends over communication rounds for our method (with and without migration) and PACFL under a 20% pathological data drift scenario on CIFAR-10.

As evident from Table 3 and visually reinforced by Figure 5, all methods experience a notable decrease in test accuracy immediately following the data distribution drift at round 61, underscoring the challenge posed by such drift. However, by round 200, most methods demonstrate some level of recovery. Significantly, our proposed method exhibits the smallest Δ value of 1‰, indicating its superior resilience to data distribution drift and its ability to effectively recover its performance. Comparing "Fed-CM" with "Fed-CM*" (our method without the migration mechanism, which has a Δ of 27‰) clearly demonstrates the crucial contribution of the migration mechanism in mitigating the impact of data distribution drift. In contrast, other methods like PACFL and IFCA show larger Δ values, suggesting a more significant impact from the distribution drift, even after adaptation. These results highlight the effectiveness of Fed-CM in maintaining performance stability even under dynamic data conditions.

6 Conclusions

To address the challenges of statistical heterogeneity and data distribution drift in federated learning, this paper proposes Fed-CM, a novel framework that features: 1) a client similarity metric based on personalized layer gradient paths for precise characterization of data distribution differences; 2) an adaptive graph-based clustering algorithm for efficient client grouping; and 3) a dynamic client migration mechanism based on update angle deviations to tackle data distribution drift. Fed-CM significantly enhances model accuracy, stability, and robustness. Extensive experiments on various datasets and under diverse settings demonstrate its superior performance over state-of-the-art methods, establishing a solid foundation for practical federated learning applications.

Acknowledgements

The authors would like to express their sincere thanks for the support from the National Natural Science Foundation of China under grant 62202319.

References

[Arivazhagan *et al.*, 2019] Manoj Ghuhan Arivazhagan, Vinay Aggarwal, Aaditya Kumar Singh, and Sunav Choudhary. Federated learning with personalization layers. *CoRR*, abs/1912.00818, 2019.

[Beltrán *et al.*, 2023] Enrique Tomás Martínez Beltrán, Mario Quiles Pérez, Pedro Miguel Sánchez Sánchez, Sergio López Bernal, Gérôme Bovet, Manuel Gil Pérez, Gregorio Martínez Pérez, and Alberto Huertas Celdrán. Decentralized federated learning: Fundamentals, state of the art, frameworks, trends, and challenges. *IEEE Communications Surveys & Tutorials*, 2023.

[Briggs *et al.*, 2020] 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*, pages 1–9, 2020.

[Chen *et al.*, 2023] Zihan Chen, Howard Yang, Tony Quek, and Kai Fong Ernest Chong. Spectral co-distillation for personalized federated learning. *Advances in Neural Information Processing Systems*, 36:8757–8773, 2023.

[Chen *et al.*, 2024] Junbao Chen, Jingfeng Xue, Yong Wang, Zhenyan Liu, and Lu Huang. Classifier clustering and feature alignment for federated learning under distributed concept drift. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[Duan *et al.*, 2022] Moming Duan, Duo Liu, Xinyuan Ji, Yu Wu, Liang Liang, Xianzhang Chen, Yujuan Tan, and Ao Ren. Flexible clustered federated learning for client-level data distribution shift. *IEEE Trans. Parallel Distributed Syst.*, 33:2661–2674, 2022.

[Fallah *et al.*, 2020] 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*, pages 3557–3568, 2020.

[Ghosh *et al.*, 2019] Avishek Ghosh, Justin Hong, Dong Yin, and Kannan Ramchandran. Robust federated learning in a heterogeneous environment. *The International Conference on Machine Learning 2019 workshop on Privacy and Security*, pages 1–30, 2019.

[Ghosh *et al.*, 2022] Avishek Ghosh, Jichan Chung, Dong Yin, and Kannan Ramchandran. An efficient framework for clustered federated learning. *IEEE Transactions on Information Theory*, 68:8076–8091, 2022.

[Hu *et al.*, 2023] Erdong Hu, Yuxin Tang, Anastasios Kyrillidis, and Chris Jermaine. Federated learning over images: vertical decompositions and pre-trained backbones are difficult to beat. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19385–19396, 2023.

[Kang *et al.*, 2020] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *8th International Conference on Learning Representations*, 2020.

[Karimireddy *et al.*, 2020] 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*, pages 5132–5143, 2020.

[Kim *et al.*, 2024] Heasung Kim, Hyeji Kim, and Gustavo de Veciana. Clustered federated learning via gradient-based partitioning. In *Forty-first International Conference on Machine Learning*, 2024.

[Krizhevsky, 2009] A Krizhevsky. Learning multiple layers of features from tiny images. *Master's thesis, University of Tront*, 2009.

[LeCun *et al.*, 1998] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86:2278–2324, 1998.

[Li *et al.*, 2020] 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.*, 2023] Zexi Li, Xinyi Shang, Rui He, Tao Lin, and Chao Wu. No fear of classifier biases: Neural collapse inspired federated learning with synthetic and fixed classifier. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5319–5329, 2023.

[Li *et al.*, 2024a] Jian Li, Tongbao Chen, and Shaohua Teng. A comprehensive survey on client selection strategies in federated learning. *Computer Networks*, 251:110663, 2024.

[Li *et al.*, 2024b] Yichen Li, Qunwei Li, Haozhao Wang, Ruixuan Li, Wenliang Zhong, and Guannan Zhang. Towards efficient replay in federated incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12820–12829, 2024.

[Liu *et al.*, 2023] Jiahao Liu, Jiang Wu, Jinyu Chen, Miao Hu, Yipeng Zhou, and Di Wu. Feddwa: Personalized federated learning with dynamic weight adjustment. In *Proceedings of the Thirty-Second International Joint Conference on Artificial IntelligenceIJCAI*, pages 3993–4001, 2023.

[Luo and Wu, 2022] Jun Luo and Shandong Wu. Adapt to adaptation: Learning personalization for cross-silo federated learning. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI*, pages 2166–2173, 2022.

[Ma *et al.*, 2022a] Xiaodong Ma, Jia Zhu, Zhihao Lin, Shanxuan Chen, and Yangjie Qin. A state-of-the-art survey on solving non-iid data in federated learning. *Future Generation Computer Systems*, 135:244–258, 2022.

[Ma *et al.*, 2022b] 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, pages 10092–10101, 2022.

[Mclaughlin and Su, 2024] Connor Mclaughlin and Lili Su. Personalized federated learning via feature distribution adaptation. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[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 *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, pages 1273–1282, 2017.

[Netzer *et al.*, 2011] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Baolin Wu, Andrew Y Ng, et al. Reading digits in natural images with unsupervised feature learning. In *Neural Information Processing Systems workshop on deep learning and unsupervised feature learning*, page 4, 2011.

[OH *et al.*, 2022] JAE HOON OH, Sangmook Kim, and Seyoung Yun. Fedbabu: Toward enhanced representation for federated image classification. In *10th International Conference on Learning Representations,*, 2022.

[Qin *et al.*, 2023] Zixuan Qin, Liu Yang, Qilong Wang, Yahong Han, and Qinghua Hu. Reliable and interpretable personalized federated learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20422–20431, 2023.

[Sattler *et al.*, 2021] Felix Sattler, Klaus-Robert Müller, and Wojciech Samek. Clustered federated learning: Model-agnostic distributed multitask optimization under privacy constraints. *IEEE Transactions on Neural Networks and Learning Systems*, 32:3710–3722, 2021.

[Shamsian *et al.*, 2021] Aviv Shamsian, Aviv Navon, Ethan Fetaya, and Gal Chechik. Personalized federated learning using hypernetworks. In *Proceedings of the 38th International Conference on Machine Learning*, pages 9489–9502, 2021.

[T Dinh *et al.*, 2020] Canh T Dinh, Nguyen Tran, and Josh Nguyen. Personalized federated learning with moreau envelopes. *Advances in neural information processing systems*, 33:21394–21405, 2020.

[Tamirisa *et al.*, 2024] Rishub Tamirisa, Chulin Xie, Wenxuan Bao, Andy Zhou, Ron Arel, and Aviv Shamsian. Fedselect: Personalized federated learning with customized selection of parameters for fine-tuning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23985–23994, 2024.

[Vahidian *et al.*, 2023] Saeed Vahidian, Mahdi Morafah, Weijia Wang, Vyacheslav Kungurtsev, Chen Chen, Mubarak Shah, and Bill Lin. Efficient distribution similarity identification in clustered federated learning via principal angles between client data subspaces. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI*, pages 10043–10052, 2023.

[Wang *et al.*, 2020] Jianyu Wang, Qinghua Liu, Hao Liang, Gauri Joshi, and H Vincent Poor. Tackling the objective inconsistency problem in heterogeneous federated optimization. *Advances in neural information processing systems*, 33:7611–7623, 2020.

[Wang *et al.*, 2024a] Qiang Wang, Bingyan Liu, and Yawen Li. Traceable federated continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12872–12881, 2024.

[Wang *et al.*, 2024b] Xi Wang, Xu Yang, Jie Yin, Kun Wei, and Cheng Deng. Long-tail class incremental learning via independent sub-prototype construction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28598–28607, 2024.

[Xiao *et al.*, 2017] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *CoRR*, abs/1708.07747, 2017.

[Xie *et al.*, 2024] Chulin Xie, De-An Huang, Wenda Chu, Daguang Xu, Chaowei Xiao, Bo Li, and Anima Anandkumar. Perada: Parameter-efficient federated learning personalization with generalization guarantees. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23838–23848, 2024.

[Yang *et al.*, 2024] Xiyuan Yang, Wenke Huang, and Mang Ye. Fedas: Bridging inconsistency in personalized federated learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 11986–11995, 2024.

[Yi *et al.*, 2024] Liping Yi, Han Yu, Zhuan Shi, Gang Wang, Xiaoguang Liu, Lizhen Cui, and Xiaoxiao Li. Fedssa: Semantic similarity-based aggregation for efficient model-heterogeneous personalized federated learning. In *Proceedings of the Thirty-Third International Joint Conference on Artificial IntelligenceIJCAI*, pages 5371–5379, 2024.

[Zhang *et al.*, 2024] Jianqing Zhang, Yang Liu, Yang Hua, and Jian Cao. Fedtgp: Trainable global prototypes with adaptive-margin-enhanced contrastive learning for data and model heterogeneity in federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 16768–16776, 2024.

[Zhou *et al.*, 2024] Da-Wei Zhou, Hai-Long Sun, Han-Jia Ye, and De-Chuan Zhan. Expandable subspace ensemble for pre-trained model-based class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23554–23564, 2024.