CoRe-Fed: Bridging Collaborative and Representation Fairness via Federated Embedding Distillation

Noorain Mukhtiar[1], **Adnan Mahmood**[1], **Quan Z. Sheng**[1]
[1]School of Computing, Macquarie University, Sydney, NSW 2109, Australia
noorain.mukhtiar@hdr.mq.edu.au, {adnan.mahmood, michael.sheng}@mq.edu.au

Abstract

With the proliferation of distributed data sources, Federated Learning (FL) has emerged as a key approach to enable collaborative intelligence through decentralized model training while preserving data privacy. However, conventional FL algorithms often suffer from performance disparities across clients caused by heterogeneous data distributions and unequal participation, which leads to unfair outcomes. Specifically, we focus on two core fairness challenges, i.e., *representation bias*, arising from misaligned client representations, and *collaborative bias*, stemming from inequitable contribution during aggregation, both of which degrade model performance and generalizability. To mitigate these disparities, we propose CoRe-Fed, a unified optimization framework that bridges collaborative and representation fairness via embedding-level regularization and fairness-aware aggregation. Initially, an alignment-driven mechanism promotes semantic consistency between local and global embeddings to reduce representational divergence. Subsequently, a dynamic reward-penalty-based aggregation strategy adjusts each client's weight based on participation history and embedding alignment to ensure contribution-aware aggregation. Extensive experiments across diverse models and datasets demonstrate that CoRe-Fed improves both fairness and model performance over the state-of-the-art baseline algorithms.

Code — https://github.com/Noorain1/CoRe-Fed

1 Introduction

Federated Learning (FL) has gained widespread adoption as a decentralized learning framework that enables multiple devices, sensors, or edge nodes (collectively referred to as clients or participants) to collaboratively train Machine Learning (ML) models without directly sharing their raw data, thus preserving data privacy (Woisetschläger et al. 2024). However, despite its growing popularity, traditional FL encounters several challenges, particularly, in ensuring fair and unbiased model performance across clients with heterogeneous data distribution and varying participation frequency. These discrepancies result in models that perform disproportionately well on certain clients, while neglecting others, a phenomenon referred to as performance bias.

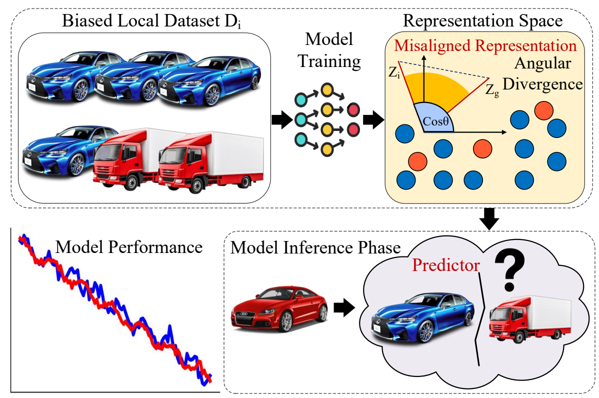


Figure 1: Illustration of representation bias in FL using CIFAR-10 automobiles images. Clients trained on biased local data, e.g., cars vs. trucks, leads to misaligned feature representations in the global embedding space. As shown in the upper section, angular divergence emerges when client representations deviate from semantically correct directions. This misalignment causes degraded performance, particularly, during inference on bias-conflicting samples.

Such bias in FL can transpire at various stages of the training process, significantly undermining the reliability and overall performance of the resulting global model (Wang et al. 2021a). An unfair model can lead to several adverse outcomes, i.e., misrepresentation of client-specific data distributions, over- or under-fitting, marginalizing certain clients, and making suboptimal decisions. To this end, we target the following three key biasing challenges in FL.

Performance Bias – In practical scenarios, client datasets often exhibit statistical heterogeneity (Ye et al. 2023) and label correlations (An et al. 2025) that introduce bias into the learning process. In particular, biased datasets (Shah et al. 2020) contain features that strongly correlate with class labels in the training dataset but do not accurately capture the inherent semantic representations. It is pertinent to mention that Deep Neural Network (DNN) models trained with such biased data are more prone to base decisions on bias-related attributes rather than meaningful representations. This, in turn, results in substantial performance bias and accuracy drop during inference, particularly when a model encounters bias-conflicting samples (Mukhtiar et al. 2025).

Representation Bias – In federated settings, clients with underrepresented classes or non-IID data tend to produce embeddings that poorly align with the global model's latent space. When a global model is trained on these locally biased datasets, it tends to form biased representations (Zhang et al. 2024). This phenomenon entails the emergence of client-specific clusters within a model's learned representations, reflecting the statistical disparities across clients. As illustrated in Figure 1, client embeddings are scattered and poorly aligned, even for samples belonging to the same class, indicating that the global model struggles to learn a unified, semantically meaningful representation.

Collaborative Bias – While addressing representation disparity is critical for learning meaningful and generalizable features across clients, we argue that it is not sufficient to guarantee fairness in federated settings. Representation alignment improves feature space consistency across clients, however, bias can still emerge during aggregation if client representations are aggregated regardless of their contribution to a global model. Standard aggregation strategies, i.e., FedAvg (McMahan et al. 2017), may (a) inadvertently suppress meaningful updates from underrepresented clients, (b) neglect infrequently participating clients, or (c) dilute high-quality representations contributed by well-performing clients, thereby compromising overall representation fairness. In contrast, assigning equal weight to all updates allows noisy or misaligned contributions to influence a global model to the same extent as well-aligned contributions (Tastan et al. 2024; Wang et al. 2024), thereby leading to collaborative bias, i.e., a systemic bias rooted in the aggregation process. Such bias not only compromises fairness but also discourages participation from marginalized clients, potentially triggering a feedback loop that further weakens their influence. This, in turn, undermines the global model's ability to generalize across diverse data distributions and leads to performance degradation.

To address the aforementioned challenges, we propose CoRe-Fed (**Co**llaborative **Re**presentation Fairness in **Fed**erated Learning), a novel framework that jointly mitigates representation and collaborative biases to reduce model performance disparities. CoRe-Fed introduces two key mechanisms, i.e., embedding-level alignment-aware optimization and contribution-aware aggregation. The former employs a contrastive embedding alignment strategy based on knowledge distillation to align local client representations with a global embedding prototype. This ensures clients' semantic consistency in a shared latent space regardless of their local data skew. The latter applies a contribution-aware aggregation scheme to adjust client weights based on a combination of negative participation frequency and embedding representation similarity.

It is worth mentioning that these two fairness notions are mutually reinforcing, i.e., representation fairness enhances the quality of client embeddings, whereas collaborative fairness ensures that those improved embeddings are not diluted during aggregation process. To the best of our knowledge, CoRe-Fed is among the first to explicitly bridge representation and collaborative fairness in a unified framework. Through comprehensive experimentation with the state-of-the-art methods, we demonstrate that mitigating bias in both the feature space and the aggregation process leads to more equitable and generalizable models in heterogeneous federated environments. Our main contributions are as follows:

- We propose CoRe-Fed, a novel FL framework that jointly maintains representation and collaborative fairness through embedding-level alignment and contribution-aware aggregation.
- We introduce a contrastive learning-based embedding alignment strategy that leverages knowledge distillation to align local client representations with a global embedding prototype, thereby promoting semantic consistency across heterogeneous clients.
- We design a contribution-aware aggregation mechanism that adjusts client weights based on their participation frequency and representational alignment, thereby amplifying the influence of semantically rich and underrepresented clients.
- To validate the effectiveness of CoRe-Fed, we conduct extensive experiments on different FL scenarios. Our experimental results indicate substantial improvements in fairness and comparable accuracy with the state-of-the-art methods.

2 Background and Related Work

2.1 Layer-Wise Fair Federated Learning

As fairness has gained significant momentum in recent years, several studies have adopted layer-wise approaches to mitigate bias in FL. For instance, (Shi et al. 2025) proposes an adaptive layer-wise weight shrinking step after model aggregation. (Rehman et al. 2023; Lee, Zhang, and Avestimehr 2023) employs a layer-wise model aggregation method to reduce communication cost. (He, Chen, and Zhang 2025) utilizes a compression technique based on model parameters of the last hidden layer for aggregation. (Pan et al. 2024) designs a method to calculate a layer-wise fair direction. (Tastan et al. 2024) leverages Shapley value approximations, derived from the gradients of the last layer, to guide a weighted aggregation scheme. In contrast, we design a novel framework to alleviate bias by aligning client embedding vectors obtained from the last layer of the feature extractor.

2.2 Representation Fairness

Some recent studies have focused on representation fairness in FL. For example, (Yan et al. 2025) envisages an augmentation technique to mitigate feature shift by injecting statistical information from the entire federation into each client's data. (Huang et al. 2024) employs representation unification and prototypical mix-up to reduce bias in FL with long-tailed data. (Luo et al. 2021) introduces a method to adjust the classifier using virtual representations. Similarly, (Xu, Lin, and Wang 2023) proposes bias-eliminating augmenters at each client with the goal of generating bias-conflicting samples to eliminate local data biases. (Zhang et al. 2023) addresses representation degeneration by decomposing local representations into a global component and a client-specific bias term regulated via a mean regularization mechanism.

Whilst these techniques have significantly advanced representation fairness in FL, they typically treat representation learning and aggregation as disjoint processes, i.e., without considering representation similarity in the aggregation logic. As a result, global models may still reflect skewed or incoherent representations, particularly under highly heterogeneous settings. In comparison, our method uniquely combines embedding alignment, based on contrastive loss and knowledge distillation, with historical-participation-aware weighting to jointly promote fair and coherent global representations across dynamically participating clients.

2.3 Collaborative Fairness

Collaborative fairness in FL (He, Chen, and Zhang 2025) ensures that participants who contribute more receive higher rewards in contrast to those with lower contributions. Several studies have addressed this aspect. For instance, (Wang et al. 2024) maintains bounded collaborative fairness by assigning submodels based on client contributions and fair aggregation of low-frequency neurons. (Jiang et al. 2023) estimates client contributions in both gradient and data space using gradient direction differences and auxiliary model prediction errors to guide aggregation. Some other works (Tastan et al. 2024; Xu et al. 2021) propose a reward mechanism based on the Shapley value to proportionally reward each client based on their respective contribution. In comparison, we propose a contribution-aware mechanism that assigns aggregation weights to clients based on their embedding alignment with a global model and inverse of participation frequency over a dynamic sliding window. This dual-component weighting establishes a reward-penalty mechanism to ensure that clients contributing with valuable and consistent information are prioritized during aggregation.

3 Methodology

In this section, we describe the overall architecture and the technical details of our proposed framework, illustrated in Figure 2 and Algorithm 1. At beginning, each client initiates local training by performing feature extraction using DNNs on its private dataset to extract local embedding vector Z_i (Line 8). Subsequently, the server computes a global embedding vector Z_g and evaluates contrastive loss between each client's embedding vector and a global embedding vector to quantify representational similarity across clients (Lines 12–13). Based on these relations, a server then constructs an alignment vector that guides knowledge distillation to align client embeddings towards a global embedding by explicitly reducing the distance between them (Lines 15–16). After knowledge distillation, the server executes a fairness-aware aggregation scheme based on two parameters; (i) a client's participation frequency score τ based on their sliding participation window and (ii) a representation similarity score α_i for each client that is derived from cosine similarity between refined embeddings and global embedding (Lines 18–21). Finally, these aggregation parameters are leveraged to compute contribution weights that balance clients' participation frequency with high representational quality (Line 22). The remainder of this section is organized into two parts, problem formulation and the technical components of the CoRe-Fed framework.

3.1 Problem Formulation and Local Objective

We consider FL setting involving a central server that orchestrates training across a set of m clients, $\mathcal{C} = \{1, \dots, m\}$, where each client i possesses a private, locally stored dataset $D_i = \{(x_{ij}, y_{ij})\}_{j=1}^{n_i}$. The goal is to collaboratively optimize a global model without sharing raw data. The global model $\omega \in \mathbb{R}^d$ comprises a feature extractor ϕ_ω and a predictor module. Each input is processed as:

$$f_\omega(x) = \text{Predictor}(\phi_\omega(x)), \tag{1}$$

where, $\phi_\omega(x)$ is the embedding vector from the feature extractor. The local training loss on client i is defined as:

$$F_i(\omega) = \frac{1}{n_i} \sum_{j=1}^{n_i} \ell(f_\omega(x_{ij}), y_{ij}) \tag{2}$$

The global learning objective is the weighted average of local losses and is calculated as:

$$\min_\omega \sum_{i=1}^{m} \frac{n_i}{n} F_i(\omega) \tag{3}$$

where, $n = \sum_i n_i$ is the total data count across all clients.

3.2 The CoRe-Fed Framework

Building on the above formulation, CoRe-Fed comprises two synergistic modules: (i) representation alignment and (ii) fairness-aware aggregation, each detailed below.

Embedding Extraction and Normalization. To represent client-specific data in a feature space, each client extracts embeddings from its local samples using a DNN encoder. The embedding vector for a client i is computed by averaging the L2-normalized feature embeddings across its local dataset:

$$z_i = \frac{1}{n_i} \sum_{j=1}^{n_i} \frac{\phi_\omega(x_{ij})}{\|\phi_\omega(x_{ij})\|} \tag{4}$$

A global feature embedding is then obtained at the server by averaging over a subset of participating clients $\mathcal{C}_t$:

$$z_g = \frac{1}{|\mathcal{C}_t|} \sum_{i \in \mathcal{C}_t} z_i \tag{5}$$

Contrastive Learning for Representation Consistency. Recently, contrastive learning has gained prominence in FL as a powerful approach for improving representation quality under data heterogeneity. For instance, (Wu et al. 2025) leverages contrastive objectives to capture meaningful graph structures. (Liu et al. 2025) applies contrastive loss to learn shared and personalized dynamics. (Wang et al. 2025) aligns multiple embeddings to address cross-modal inconsistencies. (Qi et al. 2022) combines contrastive loss with adversarial training to reduce representational unfairness. Inspired by these works, we apply contrastive representation

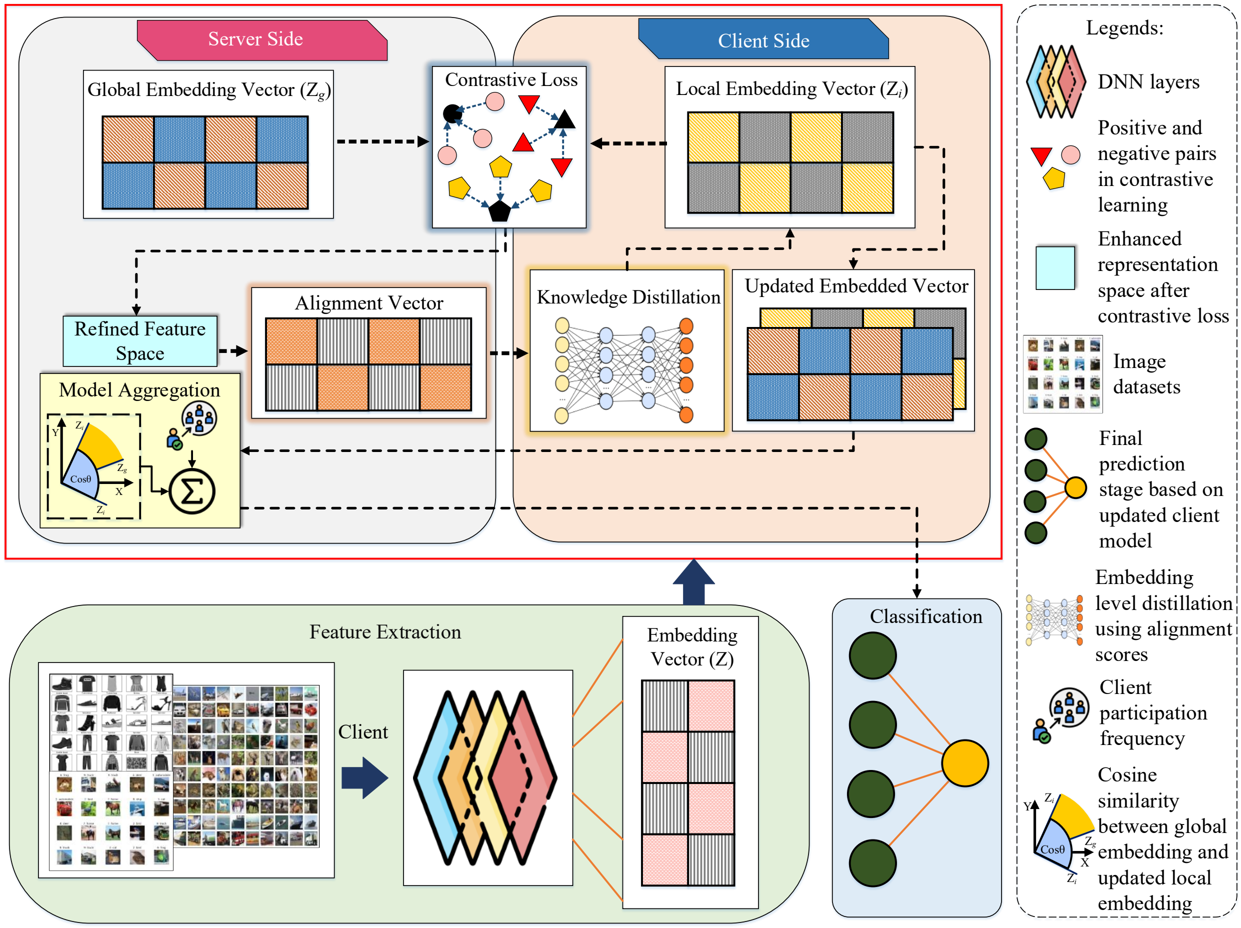


Figure 2: Architecture overview of CoRe-Fed framework. Clients perform local training and extract embedding vectors using DNN-based feature extractors. A server refines embedding space using contrastive learning over client and global embeddings. An alignment vector then guides knowledge distillation to adjust client representations towards a global semantic structure. Finally, contribution-aware aggregation is performed based on clients' embedding alignment and participation frequency.

alignment to enforce semantic consistency between local and global embeddings in a federated setting. We adopt the temperature-scaled InfoNCE (NT-Xent) loss to align each client embedding z_i with the global embedding z_g while contrasting it against other client embeddings $\{z_l\}_{l \neq i}$:

$$\mathcal{L}^{(i)}_{\text{contrast}} = -\log \frac{\exp(\cos(z_i, z_g)/\tau_c)}{\sum_{l \neq i} \exp(\cos(z_i, z_l)/\tau_c)} \tag{6}$$

Here, τ_c is a temperature parameter and $\cos()$ denotes cosine similarity, ascertained as:

$$\cos(z_i, z_l) = \frac{z_i^\top z_l}{\|z_i\| \cdot \|z_l\|} \tag{7}$$

Embedding-Level Knowledge Distillation. To reduce representational divergence, client embeddings are softly aligned towards a global embedding structure via embedding-level knowledge distillation, wherein, each client updates its representation using an alignment vector. This vector serves as a reference for embedding-level knowledge distillation, thus encouraging representation alignment while preserving local features. The updated embedding vector for client i is:

$$\tilde{z}_i = z_i + \beta \cdot (\tilde{z}_g^{(i)} - z_i) \tag{8}$$

where, $\beta \in [0, 1]$ is a knowledge distillation coefficient, and $\tilde{z}_g^{(i)} \in \mathbb{R}^d$ is the alignment vector for client i computed as:

$$\tilde{z}_g^{(i)} = A_i \cdot z_g \tag{9}$$

here, $A_i = \cos(z_i, z_g)$ is the alignment score, computed based on cosine similarity with a server-side global embedding that captures the shared semantic structure. The resulting alignment vector $\tilde{z}_g^{(i)}$ provides a refinement signal that guides each client's embedding update in the direction of the global semantic structure. This distillation encourages each client to preserve its unique structure while adapting to the global semantic space.

Participation Frequency Estimation. To ensure absent clients' fairness, CoRe-Fed estimates each client's participation frequency over a recent sliding window following the

Algorithm 1: CoRe-Fed: Collaborative Representation Fairness in Federated Learning

Input: Initial global model ω^0, total rounds T, local epochs E, client set $\mathcal{C} = \{1, \ldots, N\}$, temperature τ_c, distillation weight β, scaling factor k, fairness exponent γ

Output: ω^T

for $t = 1$ *to* $T - 1$ **do**
  // Server-side: Select online clients $\mathcal{C}_t \subseteq \mathcal{C}$
  Broadcast global model ω^t to all $i \in \mathcal{C}_t$
  // Client-side (in parallel):
  for $i \in \mathcal{C}_t$ **do**
    Initialize local model $\omega_i^t \leftarrow \omega^t$
    for $e = 1$ *to* E **do**
      Train ω_i^t on local data $\mathcal{D}_i$
    Extract local embedding z_i by Eq. (4)
    Send (ω_i^t, z_i) to server
  // Server-side: Contrastive Alignment
  Compute global embedding z_g by Eq. (5)
  Compute contrastive loss $\mathcal{L}_{\text{contrast}}$ between normalized z_i and z_g by Eq. (6)
  foreach $i \in \mathcal{C}_t$ **do**
    Compute alignment vector $\tilde{z}_g^{(i)} = A_i \cdot z_g$
    Update embedding using Knowledge Distillation Eq. (8);
  Server-side: Fairness-aware Aggregation
  Determine sliding window by $\tau = \frac{M}{|C_t|}$
  foreach $i \in \mathcal{C}_t$ **do**
    Compute $f_i = \frac{1}{\tau} \sum_{r=t-\tau+1}^{t} \mathbb{I}[i \in \mathcal{C}_r]$
    Compute $\rho_i = \cos(\tilde{z}_i, z_g)$
    Compute $w_i = \frac{\left(\frac{1}{f_i}\right)^{\gamma} \cdot \sigma(k \cdot \rho_i)}{\sum_{l \in \mathcal{C}_t} \left(\frac{1}{f_l}\right)^{\gamma} \cdot \sigma(k \cdot \rho_l)}$
  Normalize weights: $w_i \leftarrow \frac{w_i}{\sum_l w_l}$
  Aggregate: ω^{t+1} by Eq. (13)
return ω^T

approach proposed in (Pan et al. 2024; Wang et al. 2021b). Specifically, for client i, the frequency f_i is computed as the number of rounds where $i \in C_r$ for $r = t - \tau + 1$ to t and C_r is the set of online clients in round r. The window length τ is dynamically defined as $\tau = \frac{M}{|C_t|}$, where M is the number of distinct clients that have participated so far and $|C_t|$ is the number of online clients in the current round. This formulation captures both short-term participation and inactivity. Each client i's recent participation frequency f_i over a dynamic window τ is computed as:

$$f_i = \frac{1}{\tau} \sum_{r=t-\tau+1}^{t} \mathbb{I}[i \in \mathcal{C}_r] \tag{10}$$

where, $\mathbb{I}[\cdot]$ is the indicator function that evaluates to 1 if the client participated in round r.

Sigmoid-Modulated Fairness Weighting. To foster equitable contribution during optimization, CoRe-Fed combines participation frequency and representational alignment into a unified weighting scheme. This approach rewards under-represented and semantically aligned clients while softly penalizing over-represented or poorly aligned ones. If f_i is participation frequency of client i over a recent sliding window (Eq. 10) and $\rho_i = \cos(\tilde{z}_i, z_g)$ is the cosine similarity between its refined embedding $\tilde{z}_i$ and the global embedding z_g, then the aggregation weight is computed as:

$$w_i = \frac{\left(\frac{1}{f_i}\right)^{\gamma} \cdot \sigma(k \cdot \rho_i)}{\sum_{l \in \mathcal{C}_t} \left(\frac{1}{f_l}\right)^{\gamma} \cdot \sigma(k \cdot \rho_l)} \tag{11}$$

where, γ is the fairness exponent and k controls slope of sigmoid sensitivity to alignment score. The sigmoid function σ can be defined as $\sigma(x) = 1/(1 + e^{-x})$. As depicted in Eq. (11), the participation factor $\left(\frac{1}{f_i}\right)^{\gamma}$ amplifies the influence of under-represented clients while softly down-weighting frequent ones. In contrast, the alignment factor $\sigma(k \cdot \rho_i)$ favors clients that better aligns with the global model and penalizes misaligned ones.

Fairness-Aware Aggregation with Gradient Reuse. To incorporate influence of temporarily inactive clients' fair contribution, we follow (Pan et al. 2024) to reuse historical gradients within a sliding window of τ rounds. For each client i, the reused gradient $\hat{g}_i$ is defined as:

$$\hat{g}_i = \begin{cases} g_i^{(t)} & \text{if } i \in \mathcal{C}_t \\ g_i^{(t_i)} & \text{if } t - t_i \leq \tau \\ 0 & \text{otherwise} \end{cases} \tag{12}$$

where, $g_i^{(t)}$ is the current gradient, and $g_i^{(t_i)}$ is the last known gradient from client i within the τ-round window. These gradients are combined with fairness-aware weights w_i to update the global model:

$$\omega_{t+1} = \omega_t - \eta \cdot \sum_{i \in \mathcal{C}_t} w_i \cdot \hat{g}_i \tag{13}$$

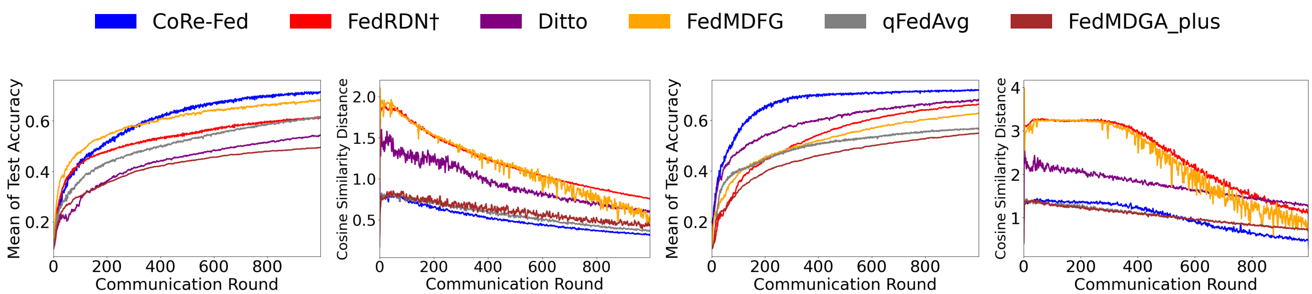
4 Experiments and Discussions

4.1 Evaluation Metrics

To provide a comprehensive evaluation of fairness, we adopt two widely used metrics, i.e., cosine similarity-based angular distance (Pan et al. 2023; Wang et al. 2021b) and Manhattan distance (Huang et al. 2023). In particular, we compute the angular cosine similarity distance between a client model $\boldsymbol{\phi}_i$ and a global model $\boldsymbol{\phi}^*$ as $D_{Cosine} = \arccos\left(\frac{\boldsymbol{\phi}_i \cdot \boldsymbol{\phi}^*}{\|\boldsymbol{\phi}_i\| \|\boldsymbol{\phi}^*\|}\right)$ to evaluate representation fairness by measuring their directional alignment. In contrast, Manhattan distance $D_{Manhattan} = \sum_{j=1}^{d} \left|\phi_i^{(j)} - \phi^{*(j)}\right|$ quantifies the model performance fairness by capturing the magnitude of parameter divergence between a client model and a global model, thereby reflecting how much each client benefits or deviates from a shared global model.

Algorithm	FMNIST			CIFAR-10		
	Accuracy ↑	D_{Cosine} ↓	$D_{Manhattan}$ ↓	Accuracy ↑	D_{Cosine} ↓	$D_{Manhattan}$ ↓
FedRDN† (Yan et al. 2025)	0.870	0.746	116.8	0.569	1.077	180.9
FedMDFG (Pan et al. 2023)	0.874	0.587	88.1	0.681	0.766	116.3
FedMGDA+ (Hu et al. 2022)	0.849	0.421	79.5	0.549	0.719	48.4
Ditto (Li et al. 2021)	0.862	0.536	106.5	0.663	1.251	104.2
qFedAvg (Li et al. 2020)	0.884	0.401	76.2	0.628	0.702	52.9
CoRe-Fed (Ours)	**0.891**	**0.294**	**73.5**	**0.722**	**0.430**	**36.0**

Table 1: Comparison of mean test accuracy, angular cosine similarity distance (D_{Cosine}), and Manhattan distance ($D_{Manhattan}$) on FMNIST and CIFAR-10 with Dir (0.5), batch size 50, 1000 rounds, and 20 out of 100 clients per round.



(a) Batch size 200

(b) Batch size 50

Figure 3: Mean test accuracy and fairness on CIFAR-10 with batch sizes 50 and 200; 20 out of 100 clients online per round.

Baselines and Hyper-parameters. Our evaluation benchmarks against several state-of-the-art fairness-based FL algorithms including q-FedAvg, (Li et al. 2020), FedMGDA+ (Hu et al. 2022), Ditto (Li et al. 2021), FedMDFG (Pan et al. 2023), and FedRDN (Yan et al. 2025). We modify FedRDN† by incorporating a small constant $\epsilon = 10^{-6}$ into the standard deviation and mean calculations to prevent division by zero and ensure model stability. We follow the experimental settings of (Pan et al. 2023), wherein all clients perform local training using Stochastic Gradient Descent with local epoch $E = 1$. The learning rate is set to $\eta = 0.1$ with a decay factor of 0.999 per round. All results are averaged over three runs with different random seeds. The hyper-parameters used for each technique are summarized in Table 2.

Method	Hyper-parameters
qFedAvg	$q \in \{0.1, 1.0\}$
Ditto	$\lambda \in \{0.1, 1.0\}$
FedMGDA+	$e \in \{0.1, 1.0\}$
FedMDFG	$\theta \in \{\frac{\pi}{180}\}$, $s \in \{5, 3, 1\}$
CoRe-Fed	$\gamma \in \{0.5, 2\}$, $k \in \{0.5, 2\}$, $\beta = 0.5$, $\tau_c = 0.07$

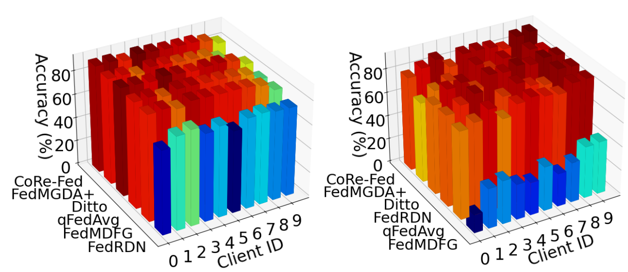
Table 2: Hyper-parameters of different methods

Datasets and Models. We evaluate the performance of mentioned baselines on two image classification datasets, i.e., Fashion MNIST (FMNIST) (Xiao, Rasul, and Vollgraf 2017) and CIFAR-10 (Krizhevsky, Hinton et al. 2009). To simulate heterogeneous client distributions, we adopt the Dirichlet-based non-IID partitioning strategy from (Hsu, Qi, and Brown 2019), where data is distributed among m clients according to a Dirichlet distribution with concentration parameter α. When $\alpha < 1$, the data distribution becomes increasingly skewed, leading to substantial label imbalance and non-uniform data volumes across clients. We adopt model architectures and configurations consistent with (Wang et al. 2021b; Pan et al. 2023), i.e., a Multilayer Perceptron (MLP) for FMNIST and Convolutional Neural Network (CNN) with two convolutional layers for CIFAR-10.

4.2 Performance and Fairness

The experimental results of CoRe-Fed vis-à-vis several state-of-the-art baselines across FMNIST and CIFAR-10 datasets are presented in Table 1. It can be observed that CoRe-Fed consistently achieves the highest accuracy with relative gains of 0.8% and 6.0%, respectively, over the best-performing baselines, i.e., qFedAvg for FMNIST and FedMDFG for CIFAR-10. The observed improvements are driven by CoRe-Fed's integration of embedding-level alignment and contribution-aware aggregation, which enhance local representation quality and ensure equitable contributions during aggregation. Moreover, in terms of representation fairness, CoRe-Fed reports the lowest angular misalignment with respective reductions of 26.7% and 43.9% compared to the best methods. This demonstrates its ability to align client representations more closely with the global model by integration of an alignment-aware weighting, which favors clients whose updates are directionally consistent with the global objective. Similarly, for model performance fairness, it records the lowest divergence with respective improvements of 3.5% and 69% compared to qFedAvg and FedMDFG. This suggests that CoRe-Fed not only reduces representation bias but also minimizes performance bias by ensuring that updates from all clients contribute proportionately and meaningfully.

As evident from Figure 3, CoRe-Fed consistently achieves superior test accuracy, enhanced fairness, faster convergence, and improved generalization under batch sizes of 50 and 200, respectively. The angular cosine similarity metric further confirms reduced representational divergence throughout training. Particularly, with the smaller batch size where client updates are inherently noisier, some baseline methods, e.g., FedMGDA+ and FedRDN, exhibit instability. In contrast, CoRe-Fed maintains stable and smooth learning dynamics, demonstrating robustness to both sampling noise and client heterogeneity. Moreover, we evaluate per client test accuracy on the FMNIST dataset with Dirichlet $\alpha \in \{0.1, 0.5\}$. As illustrated in Figure 4, CoRe-Fed consistently outperforms all baseline methods achieving both high average accuracy and low inter-client variance.

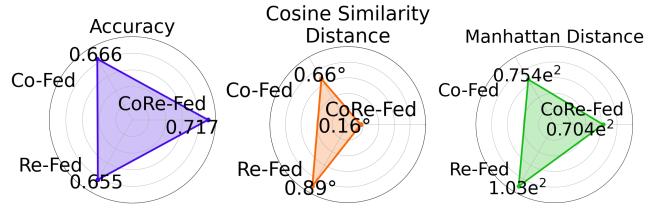


(a) (b)

Figure 4: The average per client test accuracy on FMNIST with batch size 200; 100% of 10 clients online per round.

4.3 Ablation Experiments

To assess the effectiveness of different key components in our proposed framework, we conduct ablation experiments across three scenarios, i.e., (a) **Co-Fed:** In this scenario, we exclude both contrastive learning and knowledge distillation components thus retaining only the contribution-aware aggregation strategy; (b) **Re-Fed:** In this scenario, we eliminate contribution-aware aggregation and instead use standard aggregation technique, i.e., FedAvg, while retaining contrastive learning and knowledge distillation to address representation bias; (c) **CoRe-Fed:** This scenario combines both contribution-aware aggregation and representation alignment. As evident from Figure 5, Re-Fed reduces representation bias through enhanced local training but lacks contribution-aware aggregation, thereby allowing dominant clients to overshadow others during global aggregation. This results in reduced accuracy and elevated fairness distances thus demonstrating that representation fairness alone is insufficient. Conversely, Co-Fed incorporates contribution-aware aggregation but omits representation-level adjustments, which leads to marginal fairness gains at the cost of model performance. In contrast, CoRe-Fed consistently outperforms both ablated variants achieving higher accuracy and lower cosine and Manhattan distances thus indicating more balanced and fair client contributions.



(a) (b) (c)

Figure 5: Ablation experiments comparing three scenarios.

The Trade-off between Hyper-parameters γ and k. The ablation results in Table 3 demonstrate the trade-offs driven by the fairness-aware aggregation hyper-parameters k and γ. As observed, increasing k from 0.5 to 2.0 sharpens the sigmoid weighting function and strengthens alignment between clients and the global model. Meanwhile, lowering γ from 2.0 to 0.5 reduces the emphasis on infrequent clients thereby easing penalization on frequent ones. The latter adjustment yields consistent improvements across all metrics, i.e., accuracy increases by 0.81%, cosine-based angular distance decreases by 0.16%, and Manhattan distance reduces by 0.32%. While a high γ promotes fairness by amplifying under-represented clients, it may introduce outdated updates and hinder convergence. In particular, overly amplifying the influence of non-participating or rarely participating clients may introduce divergent updates into the global model thus degrading stability and representational coherence.

Scenario	**Acc** ↑	D_{Cosine} ↓	$D_{Manhattan}$ ↓
$k = 0.5,\ \gamma = 2.0$	0.861	0.12704	52.88
$k = 2.0,\ \gamma = 0.5$	0.868	0.12684	52.71
$\Delta(\%)$	+0.81%	+0.16%	+0.32%

Table 3: The trade-off between fairness hyper-parameters.

5 Conclusion

This paper introduces CoRe-Fed, a unified federated optimization framework designed to simultaneously maintain two distinct but interconnected notions of fairness, i.e., collaborative and representation fairness. By integrating embedding-level alignment and contribution-aware aggregation, CoRe-Fed fosters semantic coherence and equitable participation, thereby enhancing both generalization and fairness across clients. Our findings open several avenues for future exploration. One promising direction is the design of personalization-aware fairness objectives that effectively balance global collaboration with individual client preferences.

Acknowledgments

The authors sincerely acknowledge the generous support of Macquarie University for funding the research via its 'International Research Excellence Award (Allocation No. 20235578)'.

References

An, X.; Wang, D.; Shen, L.; Luo, Y.; Hu, H.; Du, B.; Wen, Y.; and Tao, D. 2025. Federated Learning With Only Positive Labels by Exploring Label Correlations. *IEEE Transactions on Neural Networks and Learning Systems*, 36(4): 7651–7665.

He, J.; Chen, W.; and Zhang, X. 2025. FedAA: A Reinforcement Learning Perspective on Adaptive Aggregation for Fair and Robust Federated Learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39: 17085–17093.

Hsu, T.-M. H.; Qi, H.; and Brown, M. 2019. Measuring the effects of non-identical data distribution for federated visual classification. *arXiv preprint arXiv:1909.06335*.

Hu, Z.; Shaloudegi, K.; Zhang, G.; and Yu, Y. 2022. Federated learning meets multi-objective optimization. *IEEE Transactions on Network Science and Engineering*, 9(4): 2039–2051.

Huang, S.; Li, Y.; Chen, C.; Shi, L.; and Gao, Y. 2023. Multi-Metrics Adaptively Identifies Backdoors in Federated Learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4652–4662.

Huang, W.; Liu, Y.; Ye, M.; Chen, J.; and Du, B. 2024. Federated Learning With Long-Tailed Data via Representation Unification and Classifier Rectification. *IEEE Transactions on Information Forensics and Security*, 19: 5738–5750.

Jiang, M.; Roth, H. R.; Li, W.; Yang, D.; Zhao, C.; Nath, V.; Xu, D.; Dou, Q.; and Xu, Z. 2023. Fair federated medical image segmentation via client contribution estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16302–16311.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images.

Lee, S.; Zhang, T.; and Avestimehr, A. S. 2023. Layer-Wise Adaptive Model Aggregation for Scalable Federated Learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(7): 8491–8499.

Li, T.; Hu, S.; Beirami, A.; and Smith, V. 2021. Ditto: Fair and Robust Federated Learning Through Personalization. In *38th International Conference on Machine Learning*, 6357–6368.

Li, T.; Sanjabi, M.; Beirami, A.; and Smith, V. 2020. Fair Resource Allocation in Federated Learning. In *International Conference on Learning Representations*, 1–27.

Liu, Q.; Sun, S.; Liang, Y.; Liu, M.; and Xue, J. 2025. Personalized federated learning for spatio-temporal forecasting: A dual semantic alignment-based contrastive approach. In *iProceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 12192–12200.

Luo, M.; Chen, F.; Hu, D.; Zhang, Y.; Liang, J.; and Feng, J. 2021. No fear of heterogeneity: Classifier calibration for federated learning with non-iid data. *Advances in Neural Information Processing Systems*, 34: 5972–5984.

McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; and y Arcas, B. A. 2017. Communication-efficient Learning of Deep Networks from Decentralized Data. In *Artificial Intelligence and Statistics*, 1273–1282.

Mukhtiar, N.; Mahmood, A.; Zhou, Y.; Yang, J.; Teng, J.; and Sheng, Q. Z. 2025. Federated Learning at the Forefront of Fairness: A Multifaceted Perspective. In *Proceedings of the International Joint Conference on Artificial Intelligence*.

Pan, Z.; Li, C.; Yu, F.; Wang, S.; Wang, H.; Tang, X.; and Zhao, J. 2024. FedLF: Layer-Wise Fair Federated Learning. In *AAAI Conference on Artificial Intelligence*, 14527–14535.

Pan, Z.; Wang, S.; Li, C.; Wang, H.; Tang, X.; and Zhao, J. 2023. FedMDFG: Federated Learning with Multi-Gradient Descent and Fair Guidance. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(8): 9364–9371.

Qi, T.; Wu, F.; Wu, C.; Lyu, L.; Xu, T.; Liao, H.; Yang, Z.; Huang, Y.; and Xie, X. 2022. Fairvfl: A fair vertical federated learning framework with contrastive adversarial learning. *Advances in Neural Information Processing Systems*, 35: 7852–7865.

Rehman, Y. A. U.; Gao, Y.; De Gusmao, P. P. B.; Alibeigi, M.; Shen, J.; and Lane, N. D. 2023. L-dawa: Layer-wise divergence aware weight aggregation in federated self-supervised visual representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 16464–16473.

Shah, H.; Tamuly, K.; Raghunathan, A.; Jain, P.; and Netrapalli, P. 2020. The Pitfalls of Simplicity Bias in Neural Networks. *Advances in Neural Information Processing Systems*, 33: 9573–9585.

Shi, C.; Li, J.; Zhao, H.; dan Guo, D.; and Chang, Y. 2025. FedLWS: Federated Learning with Adaptive Layer-wise Weight Shrinking. In *The Thirteenth International Conference on Learning Representations*.

Tastan, N.; Fares, S.; Aremu, T.; Horváth, S.; and Nandakumar, K. 2024. Redefining Contributions: Shapley-Driven Federated Learning. In *International Joint Conference on Artificial Intelligence*, 5009–5017.

Wang, A.; Zhang, Z.; Wang, D.; Wang, F.; Hu, H.; Guo, J.; Zhou, Y.; Pang, C.; and Wen, S. 2025. Overcoming Heterogeneous Data in Federated Medical Vision-Language Pre-training: A Triple-Embedding Model Selector Approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 7500–7508.

Wang, L.; Xu, S.; Wang, X.; and Zhu, Q. 2021a. Addressing Class Imbalance in Federated Learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(11): 10165–10173.

Wang, Z.; Fan, X.; Qi, J.; Wen, C.; Wang, C.; and Yu, R. 2021b. Federated Learning with Fair Averaging. In *International Joint Conference on Artificial Intelligence*, 1615–1623.

Wang, Z.; Wang, Z.; Lyu, L.; Peng, Z.; Yang, Z.; Wen, C.; Yu, R.; Wang, C.; and Fan, X. 2024. FedSAC: Dynamic Submodel Allocation for Collaborative Fairness in Federated Learning. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, 3299–3310.

Woisetschläger, H.; Erben, A.; Wang, S.; Mayer, R.; and Jacobsen, H.-A. 2024. A Survey on Efficient Federated Learning Methods for Foundation Model Training. In *International Joint Conference on Artificial Intelligence*, 8317–8325.

Wu, N.; Zhao, Y.; Dong, H.; Xi, K.; Yu, W.; and Wang, W. 2025. Federated Graph Anomaly Detection Through Contrastive Learning with Global Negative Pairs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 21554–21562.

Xiao, H.; Rasul, K.; and Vollgraf, R. 2017. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*.

Xu, X.; Lyu, L.; Ma, X.; Miao, C.; Foo, C. S.; and Low, B. K. H. 2021. Gradient Driven Rewards to Guarantee Fairness in Collaborative Machine Learning. In Ranzato, M.; Beygelzimer, A.; Dauphin, Y.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems*, volume 34, 16104–16117.

Xu, Y.-Y.; Lin, C.-S.; and Wang, Y.-C. F. 2023. Bias-Eliminating Augmentation Learning for Debiased Federated Learning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 20442–20452.

Yan, Y.; Fu, H.; Li, Y.; Xie, J.; Ma, J.; Yang, G.; and Zhu, L. 2025. A simple data augmentation for feature distribution skewed federated learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 25749–25758.

Ye, M.; Fang, X.; Du, B.; Yuen, P. C.; and Tao, D. 2023. Heterogeneous Federated Learning: State-of-the-art and Research Challenges. *ACM Computing Surveys*, 56(3): 1–44.

Zhang, J.; Hua, Y.; Cao, J.; Wang, H.; Song, T.; XUE, Z.; Ma, R.; and Guan, H. 2023. Eliminating Domain Bias for Federated Learning in Representation Space. In *International Conference on Neural Information Processing Systems*, 14204–14227.

Zhang, J.; Wang, J.; Li, Y.; Xin, F.; Dong, F.; Luo, J.; and Wu, Z. 2024. Addressing Heterogeneity in Federated Learning with Client Selection via Submodular Optimization. *ACM Transactions on Sensor Networks*, 20(2).