Balancing Similarity and Complementarity for Federated Learning

Kunda Yan [*1] Sen Cui [*1] Abudukelimu Wuerkaixi [1] Jingfeng Zhang [2 3] Bo Han [4 3] Gang Niu [3]
Masashi Sugiyama [† 3 5] Changshui Zhang [† 1]

Abstract

In mobile and IoT systems, Federated Learning (FL) is increasingly important for effectively using data while maintaining user privacy. One key challenge in FL is managing statistical heterogeneity, such as non-i.i.d. data, arising from numerous clients and diverse data sources. This requires strategic cooperation, often with clients having similar characteristics. However, we are interested in a fundamental question: does achieving optimal cooperation necessarily entail cooperating with the most similar clients? Typically, significant model performance improvements are often realized not by partnering with the most similar models, but through leveraging complementary data. Our theoretical and empirical analyses suggest that optimal cooperation is achieved by enhancing complementarity in feature distribution while restricting the disparity in the correlation between features and targets. Accordingly, we introduce a novel framework, `FedSaC`, which balances similarity and complementarity in FL cooperation. Our framework aims to approximate an optimal cooperation network for each client by optimizing a weighted sum of model similarity and feature complementarity. The strength of `FedSaC` lies in its adaptability to various levels of data heterogeneity and multimodal scenarios. Our comprehensive unimodal and multimodal experiments demonstrate that `FedSaC` markedly surpasses other state-of-the-art FL methods.

1. Introduction

Federated Learning (FL) (McMahan et al., 2017), emerging as a pivotal paradigm in machine learning, is increasingly acclaimed for facilitating collaborative training across diverse clients while ensuring data confidentiality. However, FL still encounters significant challenges, chiefly statistical heterogeneity - the occurrence of non-i.i.d. data across diverse local clients, as explored in prior research. (Cui et al., 2022; Qu et al., 2022; Karimireddy et al., 2020; Li et al., 2023). In real-world scenarios with data from heterogeneous user bases, models often face performance decline due to local data distribution variances (Kairouz et al., 2021; Li et al., 2022; Huang et al., 2022).

In the context of multimodal learning, statistical heterogeneity is notably pronounced (Chen & Zhang, 2020; Zheng et al., 2023). Variations in dimensionality, quality, and reliability among diverse data sources exacerbate heterogeneity within each client's modalities and magnify distribution discrepancies between clients. Such significant heterogeneity complicates achieving consistent and efficient learning in the FL framework (Yu et al., 2023; Lin et al., 2023).

In response to this challenge, a promising direction involves the identification of optimal collaborators predicated on model similarity metrics (Baek et al., 2023; Sattler et al., 2021; Ye et al., 2023). For example, IFCA (Baek et al., 2023) clusters cooperative clients based on the similarity of their model parameters, whereas CFL (Sattler et al., 2021) employs gradient similarity for the same purpose. pFedGraph (Ye et al., 2023) constructs a cooperation graph guided by an intuitive notion that clients with greater similarity should collaborate more intensively. These methods collectively emphasize the importance of model similarity in strategic collaboration.

However, we raise a fundamental question: *does achieving optimal cooperation necessarily entail cooperating with the most similar clients?* Theoretically, similarity oriented from collaboration is conservative, such that it could potentially result in unproductive cooperation. For example, under the assumption of two completely identical clients or models, cooperation between them would yield no information gain for either party, despite their maximal similarity. Interestingly, the fundamental precondition for model enhancement

[*]Equal contribution †Corresponding authors [1] Institute for Artificial Intelligence, Tsinghua University (THUAI), Beijing National Research Center for Information Science and Technology (BNRist), Department of Automation, Tsinghua University, Beijing, P.R.China [2]The University of Auckland [3]RIKEN [4]Hong Kong Baptist University [5]The University of Tokyo. Correspondence to: Masashi Sugiyama <sugi@k.u-tokyo.ac.jp>, Changshui Zhang <zcs@mail.tsinghua.edu.cn>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

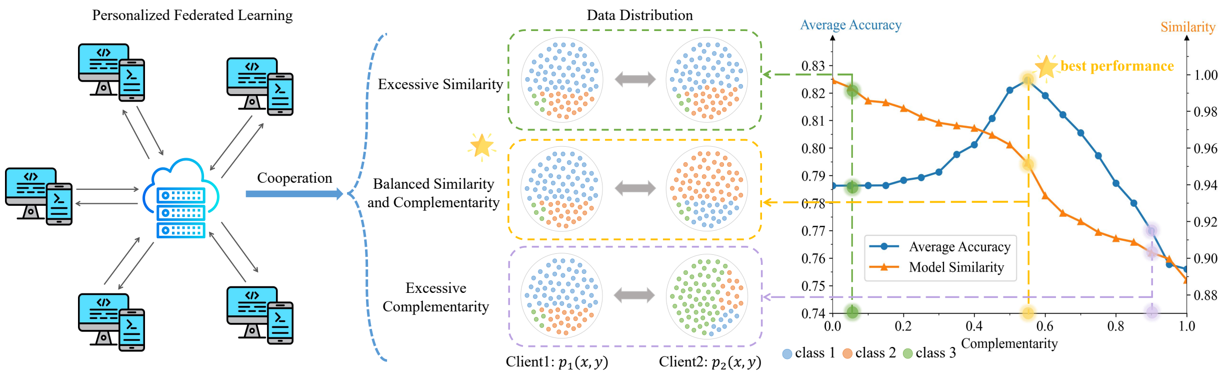


Figure 1. Illustration of the role of data complementarity on personalized federated learning cooperation. The figure presents the experimental results how increasing data complementarity between two clients influences average accuracy post-cooperation and model similarity. Three scenarios are presented, showcasing distinct levels of complementarity for local data distributions. The findings underscore the benefits of complementarity, revealing that a balance of similarity and complementarity enhances cooperative benefits in a federated learning framework.

through collaboration is complementarity, not similarity.

Inspired by this intuition, we design experiments to explore the underlying mechanism. As an illustration in FL, we exemplify cooperation between two clients through model parameter aggregation. During the cooperation, we incrementally enhance the disparity in their data distributions to promote complementarity. Our investigation explores the alterations in the average accuracy and model similarity with the increase of data complementarity. As depicted in Figure 1, cooperation between the two clients with the highest model similarity does not yield the maximum gains, while cooperation between clients when the data exhibits moderate complementarity is more advantageous, even if their models are not the most similar. Additionally, excessive data complementarity might indicate significant discrepancies in data distributions, rendering the cooperation less effective. The experimental results demonstrate the indispensability of complementarity in the cooperation of FL. Therefore, an intriguing question emerges: *how can we deduce the cooperation gain network among clients by simultaneously considering similarity and complementarity, thereby facilitating more optimal model cooperative learning?*

We present an answer to this question grounded in a thorough analysis of statistical heterogeneity. Briefly, suppose we use $p_i(x, y) = p_i(x)p_i(y|x)$ to denote the joint distribution of the feature x and label y in the i^{th} client. By controlling one of the distributions, it is observed that varied $p(y|x)$ signals the presence of a *concept shift* among clients. A substantial concept shift can detrimentally affect model learning. On the other hand, the limited nature of data within each client makes it challenging to precisely characterize the true local distribution. Hence, varied $p(x)$ which indicates the presence of a *covariate shift* could be beneficial, potentially providing more information gain. Experiments in Figure 1 also explain that a moderate covariate shift could introduce complementarity in model learning, leading to enhanced performance. Consequently, we argue that allowing moderate variations in the marginal distribution $p(x)$ while ensuring consistency in the conditional distribution $p(y|x)$ presents a more effective cooperation than merely relying on singular model similarity metrics in previous research.

Within the aforementioned analysis, we propose a novel cooperation framework by balancing Similarity and Complementarity, named FedSaC. Specifically, we introduce a cooperation network where each node signifies a client and edges reflect cooperation strength. This network is dynamically optimized, balancing model similarity with feature complementarity. The edge weights denote this balance, ensuring that clients not only collaborate with similar models but also leverage complementary feature insights. We applied the cooperation network to a FL framework, dividing it into two processes: server-side and client-side, achieving personalized interactive cooperation under privacy protection conditions. Leveraging the refined approach, our FedSaC adeptly accommodates various levels of data heterogeneity and multi-modal scenarios, and effectively identifies the optimal collaborators for each client.

Our experiments validate the efficacy of FedSaC, demonstrating its ability to consider both similarity and complementarity in cooperation while maintaining a balance. Thanks to this property of FedSaC, it outperforms 12 unimodal and 4 multimodal baselines across various benchmark datasets. Consequently, we conclude that complementarity is indeed beneficial in FL cooperation, rather than solely focusing on similarity.

We summarize our contributions as follows:

- We challenge a widely accepted notion that model similarity can be a robust metric for determining the potential benefits of cooperative model learning. We argue that achieving optimal cooperation necessitates a dual consideration of similarity and complementarity.

- We propose a novel collaboration framework, FedSaC, which infers the cooperation by optimizing a constrained objective. The objective quantifies a balanced similarity and complementarity between local clients.
- We demonstrate through extensive experiments that FedSaC exhibits superior performance in addressing data heterogeneity in FL, surpassing other state-of-the-art FL methods in both unimodal and multimodal scenarios. Code is accessible at https://github.com/yankd22/FedSaC/.

2. Related Work

2.1. Federated Learning and Statistical Heterogeneity

Federated learning (McMahan et al., 2017) has become a key focus in the machine learning field for its practical applications, but it also presents several challenges, including communication efficiency (Konečný et al., 2016), privacy concerns (Agarwal et al., 2018; Mothukuri et al., 2021), and statistical heterogeneity (Karimireddy et al., 2020; Cui et al., 2022; Qu et al., 2022), and they have been the topic of multiple research efforts (Mohri et al., 2019). Recently, a wealth of work has been proposed to handle statistical heterogeneity. For example, (Mohri et al., 2019) seek a balanced model performance distribution by maximizing the model performance on any arbitrary target distribution. (Li et al., 2021a) develop MOON that corrects local training by maximizing the similarity between local and global models. Some clustering-based FL methods (Wu et al., 2021; Baek et al., 2023) also attempt to utilize model similarity to cluster similar clients in order to mitigate the impact of statistical heterogeneity. A fundamental question arises "whether a high degree of model similarity invariably leads to more effective collaboration"?

2.2. Personalized Federated Learning

A global model (e.g., FedAvg (McMahan et al., 2017)) could harm certain clients when there are severe distribution discrepancies (Deng et al., 2020), and this stimulates the study of personalized federated learning (Qu et al., 2023; Qin et al., 2023; Zhu et al., 2023). One line of work focused on a better balance between global and local training. For example, there are researches (Dinh et al., 2020; Li et al., 2020; Karimireddy et al., 2020) proposing to stabilize local training by regulating the deviation from the global model over the parameter space. Another line of research aims to achieve a more fine-grained cooperation via collaboratively learning with similar clients. For example, (Ghosh et al., 2022) proposed to cluster the collaborative clients according to their model parameter similarity, and learn a personalized model for each cluster. (Ye et al., 2023) specify who to collaborate at what intensity level for each client according to model similarity.While the current trend in personalized federated learning heavily relies on similarity metrics, we suggest that a balanced focus on both similarity and complementarity can more accurately optimize collaboration benefits.

2.3. Federated Multimodal Learning

Considering the diverse data modalities in real life, there are a few research investigating the tasks of *federated multimodal learning*, i.e., collaboratively learning models on distributed sources containing multimodal data (Xiong et al., 2022; Zhao et al., 2021). In particular, Xiong *et al.* propose a co-attention mechanism(Xiong et al., 2022) to fuse different modalities. (Yu et al., 2023) design a regularization technique to restrict global-local discrepancy by contrastive learning. (Zhao et al., 2022) enhanced the challenge for multimodal clients with unlabeled local data using a semi-supervised framework. Our approach optimizes a weighted sum of model similarity and feature complementarity for automatic weight allocation to clients. Given the higher statistical heterogeneity in multimodal data compared to unimodal data, global models, as currently developed, struggle with conflicting client dependencies. Therefore, our focus is on developing personalized models for multimodal tasks to address data heterogeneity effectively.

3. Problem Setup

The problem to be solved in this paper is formally defined in this section. Specifically, we introduce the objective of federated learning, and through analyzing the statistical heterogeneity, demonstrate the feasibility and significance of balancing similarity and complementarity in FL.

3.1. Notations

Suppose there are N clients in a federated network, each client owns a private dataset D^i with n^i data samples, where $i = 1, \ldots, N$. We define the relative size of each dataset D^i as $p^i = n^i / \sum_j n^j$. The dataset $D^i = \left\{X^i, Y^i\right\}$ consists of the input space X^i and output space Y^i. A data point is denoted by $\{x, y\}$, with x signifying either a unimodal or a multimodal feature. The input space and the output space are shared across all clients.

Federated Learning. In FL scenario, each client collaboratively refines a predictive model using local data and collective knowledge to optimally predict label y. FedAvg (McMahan et al., 2017), as an exemplar method introduced by Mcmahan *et al.*, learns a global model $\boldsymbol{\theta}^g$ for all clients by minimizing the empirical risk over the samples from all clients, i.e.,

$$\min_{\boldsymbol{\theta}^g \in \boldsymbol{\Theta}} \sum_{i=1}^{N} p^i \mathcal{L}_i \left(\boldsymbol{\theta}^g ; D^i\right), \tag{1}$$

where $\mathbf{\Theta}$ is the hypothesis space and $\mathcal{L}_i$ denotes the loss objective of each clients. From Eq.1, `FedAvg` presumes that i.i.d. data from different clients converge to a shared joint distribution $p(x, y)$, indicating statistical homogeneity across diverse data points.

3.2. Statistical Heterogeneity

In practical scenarios, the i.i.d assumption in `FedAvg` is largely unrealistic. There could be noticeable distinctive traits in local datasets across different clients stemming from diverse environments and contexts in which clients gather data (Mohri et al., 2019). Existing research reveals that such statistical heterogeneity may result in under-performance of global models (Qu et al., 2022). In the given context, the concept of personalized federated learning is introduced as a potential solution to mitigate the statistical heterogeneity issues by facilitating selective cooperation (Li et al., 2021b; Lin et al., 2022; Ye et al., 2023). Existing works assume that clients derive more benefit from collaborating with peers who possess similar characteristics, thereby implying a diminished level of cooperation where dissimilarities exist. This allows each client to utilize information more akin to their local distribution. However, we pose a fundamental question: *is cooperation with similar peers truly optimal?*

We endeavor to delve deeply into statistical heterogeneity to provide an unexpected answer. Suppose we use $p(x, y)$ to denote the joint distribution of features and labels, the nature of statistical heterogeneity lies in the disparate joint distributions across various clients, i.e., $p(x^{k_1}, y^{k_1}) \neq p(x^{k_2}, y^{k_2})$, where $k_1 \neq k_2$. The joint distribution $p(x, y)$ can be decomposed as $p(x, y) = p(x)p(y|x)$, thus allowing statistical heterogeneity to be represented as

$$p(x^{k_1})p(y^{k_1}|x^{k_1}) \neq p(x^{k_2})p(y^{k_2}|x^{k_2}), \tag{2}$$

where $k_1 \neq k_2$. Within this formulation, we recall the following definition.

Definition 3.1. Here is the definition of two distribution shifts.
(1) ***Covarient shift*** (Peng et al., 2020; Gan et al., 2021): The distribution of input features $p(x)$ exhibits disparities among different clients, while the conditional distribution $p(y|x)$ is shared.
(2) ***Concept shift*** (Jothimurugesan et al., 2023; Canonaco et al., 2021): The relationship between input features and output labels $p(y|x)$ alterations among different clients, even if the distribution of input features $p(x)$ remains constant.

3.3. Optimal Cooperation

To effectively mitigate the challenges of statistical heterogeneity in federated learning, previous research has predominantly focused on employing a similarity metric to facilitate client cooperation. This approach, as highlighted in studies such as (Li et al., 2021b; Sattler et al., 2021; Ye et al., 2023), emphasizes maximizing similarity to address *concept shift*, which is indeed a crucial aspect of aligning learning models across diverse clients.

Maximizing similarity is justifiable for addressing *concept shift.* Specifically, when two clients exhibit similar conditional distributions $p(y|x)$, it signifies a shared correlation or mapping between input features and output labels. Such a correlation is instrumental in fostering a more coherent alignment and synthesis of the learned models or knowledge, thereby enhancing the overall effectiveness of the federated learning process. However, in the case of covariant shift, where there is a variation in input features across clients, this strategy may not yield the same level of effectiveness.

Moderate covariant shift is beneficial for cooperation in federated learning, as exhibited in Figure 1. In the context of federated learning, local clients, limited by their specific data subsets, often present an incomplete representation of the broader data distribution. Therefore, clients are inclined to engage in cooperation to surmount the limitations posed by their individual data paucity. When there is minimal covariate shift between two clients, overlapping input features can limit the model's capacity to absorb diverse information. This constraint impedes the detection of underlying data patterns, hindering collaborative efforts. Thus, we suggest that a client should collaborate with peers whose input features exhibit moderate variations. Such strategic collaborations harness complementary data, enhancing the model's predictive accuracy and generalization potential.

Given the aforementioned analysis, it becomes apparent that reliance on similarity metrics as a collaborative criterion is suboptimal in the presence of covariate shift. It would be more judicious to permit moderate variations in $p(x)$ while maintaining similarity in the conditional distribution $p(y|x)$.

4. Our Method: Balancing Similarity and Complementarity

In this section, we detail the construction of a cooperation network, designed to identify optimal collaborators for each client within FL framework. Section 4.1 introduces our cooperation network, and Section 4.2 discusses the global optimization process, balancing similarity and complementarity. In Section 4.3, we provide specific methods for computing similarity and complementarity under privacy protection, and further decompose the global optimization into server-side and client-side to fit FL architecture.

4.1. Cooperation Network

Contrary to prior personalized FL works, we argue that merely targeting similar clients is not always optimal, as

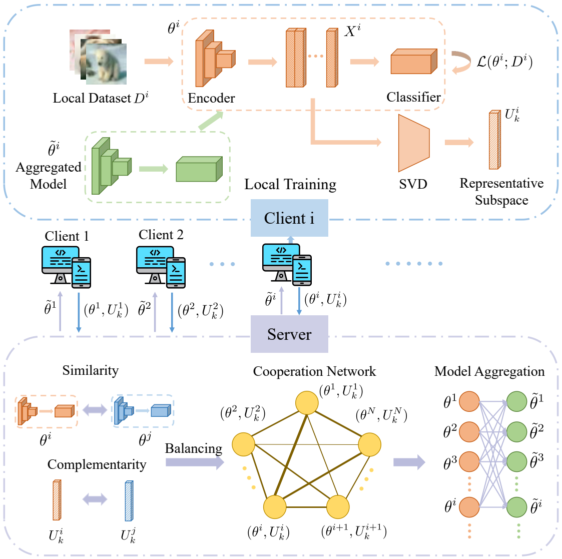


Figure 2. Illustration of our `FedSaC` approach. Local clients train models by minimizing empirical risk, incorporating a regularization term based on the distance to the aggregated model. Post-training, models are distilled via SVD to capture the representative subspace, which, alongside model parameters, is sent to the server. The server constructs a cooperation network, balancing similarity and complementarity among clients, to aggregate models. These aggregated models are then disseminated to clients for the subsequent training iteration.

diverse feature distributions can yield more insights for robust generalization. Hence, our goal of cooperation among federated clients is to achieve a balance: seeking data with a similar conditional distribution while ensuring a complementary marginal distribution.

To measure this balance, we introduce two metrics: similarity, targeting minimal concept shifts within similar conditional distributions, and complementarity, addressing moderate covariate shifts for diverse marginal distributions. Informed by this rationale as discussed in Section 3.2, we advocate that clients should collaborate with others who share similar $p(y|x)$ but different $p(x)$. Such a balanced collaboration ensures that clients not only access shared knowledge but also harness complementary insights from different data angles, ultimately boosting learning outcomes.

From a global perspective, the cooperation strength of a client is influenced by other clients. Inspired by (Ye et al., 2023), we construct a cooperation network shown in Figure 2 which balances similarity and complementarity among clients while collaborating. This network comprises N nodes, each representing a client. The adjacency matrix of the network is denoted as $\boldsymbol{W} \in \mathbb{R}^{N\times N}$, where the element $\boldsymbol{W}_{ij}$ indicates the cooperation strength between the i^{th} and j^{th} clients in federated learning. We establish a global objective encompassing both similarity and complementarity to determine the optimal weights in the adjacency matrix, which identifies the optimal collaborators for each client.

4.2. Optimization with Similarity and Complementarity

In practice, client data distributions are inaccessible. To bypass the privacy constraint, we utilize local models as surrogates for estimating data distributions. We capture the local data distribution of clients as the sampling distribution for their marginal distribution $p(x)$. For the conditional distribution $p(y|x)$, intuitively, the personalized model parameters, after local training, can capture the mapping from the marginal distribution $p(x)$ to the label distribution $p(y)$. Hence, we consider the local-trained parameters as approximate surrogates for the conditional distribution $p(y|x)$. We present a global optimization equation, as articulated in Equation 3, which aims to refine the local personalized model parameters $\{\boldsymbol{\theta}^i\}$ for each client and the network adjacency matrix $\boldsymbol{W}$. The term $\mathcal{L}_i(\sum_{j=1}^{N} \boldsymbol{W}_{ij}\boldsymbol{\theta}^j; D^i)$ denotes the empirical risk on the local dataset of the i^{th} client, following the weighted aggregation of model parameters across multiple clients. $\mathcal{C}$ denotes the complementarity of marginal distributions between two clients, while $\mathcal{S}$ denotes the similarity in model parameters between them. The hyperparameters α and β are introduced to adjust the prominence of complementarity and similarity, ensuring a balanced emphasis on both during optimization.

$$
\begin{aligned}
\min_{\{\boldsymbol{\theta}^i\},\boldsymbol{W}} \sum_{i=1}^{N} p^i \Bigg(& \mathcal{L}_i(\sum_{j=1}^{N} \boldsymbol{W}_{ij}\boldsymbol{\theta}^j; D^i) + \alpha \sum_{j=1}^{N} \mathcal{C}(\boldsymbol{W}_{ij}; D^i; D^j) \\
& - \beta \sum_{j=1}^{N} \mathcal{S}(\boldsymbol{W}_{ij}; \boldsymbol{\theta}^i; \boldsymbol{\theta}^j) \Bigg) \\
\text{s.t.} \quad & \sum_{j=1}^{N} \boldsymbol{W}_{ij} = 1, \forall i; \quad \boldsymbol{W}_{ij} \geq 0, \forall i, j \qquad (3)
\end{aligned}
$$

The optimization equation minimizes the empirical risk on local data while balancing the similarity and complementarity among clients. The two constraints ensure the normalization and non-negativity of each client's cooperation weight. Compared to previous methods, our cooperation network approach flexibly determines the cooperation strength among clients. By considering variations in marginal distributions across different clients, it more effectively captures distributional differences, leading to enhanced model performance.

4.3. FedSaC: Balancing Similarity and Complementarity

In practical FL architectures, each client is restricted to its local dataset and model. Model aggregation, as well as the computation of complementarity and similarity, require coordination with a central server. Given this structure, we initially introduce the metric of similarity and complementarity under privacy constraints. Subsequently, we partition the global optimization equation into two stages, optimizing separately at the server side and the client side. The aforementioned process is illustrated in Figure 2.

Algorithm 1 FedSac

```
Input: Total communication round $T$, client number $N$, initial local model $\{\boldsymbol{\theta}^i_{(0)}\}$, initial cooperation network $\boldsymbol{W}_{(0)}$;
for each round $t = 0, ..., T-1$ do
    Client Side
    for client $i = 1, ..., N$ in parallel do
        Receive aggregated model $\tilde{\boldsymbol{\theta}}^i_{(t)}$ sent from server;
        Update local model $\boldsymbol{\theta}^i_{(t)} \leftarrow \tilde{\boldsymbol{\theta}}^i_{(t)}$;
        Minimize local loss defined in Eq.9;
        Extract representative subspace $U^i_{k(t)}$ by Eq.5;
        Send $\boldsymbol{\theta}^i_{(t)}$ and $U^i_{k(t)}$ to server;
    end for
    Server Side
    Calculate similarity $\mathcal{S}$ for each pair of clients as Eq.4;
    Calculate complementarity $\mathcal{C}$ for each pair of clients as Eq.6 and Eq.7;
    Update $\boldsymbol{W}_{(t)}$ by optimization Eq.8;
    Aggregate model $\tilde{\boldsymbol{\theta}}^i_{(t+1)} \leftarrow \sum_j \boldsymbol{W}_{ij(t)} \boldsymbol{\theta}^j_{(t)}$;
end for
Output: the learned personalized models $\left\{\boldsymbol{\theta}^i_{(T)}\right\}$.
```

4.3.1. The Metric of Similarity and Complementarity

Similarity Metric. Following conventional practices, we use model parameters as proxies and adopt the cosine distance between the local models of the i^{th} and j^{th} clients as our similarity metric, denoted as,

$$\mathcal{S}(\boldsymbol{W}_{ij}; \boldsymbol{\theta}^i; \boldsymbol{\theta}^j) = \boldsymbol{W}_{ij} \frac{\boldsymbol{\theta}^i \cdot \boldsymbol{\theta}^j}{\|\boldsymbol{\theta}^i\| \cdot \|\boldsymbol{\theta}^j\|}. \tag{4}$$

Complementarity Metric. In light of the privacy principles inherent to FL, we use an indirect method to capture data complementarity. For a given local dataset D^i at client i, the local model $\boldsymbol{\theta}^i$ extracts the feature matrix $\boldsymbol{X}^i$. Applying singular value decomposition(SVD) on $\boldsymbol{X}^i$ yields:

$$\boldsymbol{X}^i = \boldsymbol{U}^i \boldsymbol{\Sigma}^i (\boldsymbol{V}^i)^T, \tag{5}$$

where $\boldsymbol{U}^i$ contains the singular vectors of $\boldsymbol{X}^i$, capturing the direction in the feature space. For our purposes, we consider the first k columns of $\boldsymbol{U}^i$, denoted $\boldsymbol{U}^i_k$, as the representative subspace for client i.

To gauge the complementarity between clients i and j, we utilize the principal angles between their respective subspaces. The l^{th} principal angle $\cos\phi_l$ between the two is given by:

$$\cos\phi_l = \max\nolimits_{u \in \boldsymbol{U}^i_k, v \in \boldsymbol{U}^j_k} u^T v, \tag{6}$$

where $l = 1, \ldots, k$. These angles offer a quantifiable measure of the complementarity between the two datasets. A small principal angle suggests a high similarity between the subspaces, while an angle close to $\pi/2$ implies that the subspaces are nearly orthogonal, indicating significant divergence in their feature spaces.

By averaging these angles, we obtain complementarity as:

$$\mathcal{C}(\boldsymbol{W}_{ij}; D^i; D^j) = \boldsymbol{W}_{ij} \cos\left(\frac{1}{k} \cdot \sum_l \phi_l\right). \tag{7}$$

4.3.2. FedSaC in FL architecture

Server Side. On the server side, we compute the similarity and complementarity based on the model parameters $\{\boldsymbol{\theta}^i\}$ and the subspace representation $\{\boldsymbol{U}^i_k\}$ received from the local client. Subsequently, we derive the adjacency matrix $\boldsymbol{W}$ through optimization equations. In FL scenario, as the empirical loss of local clients is elusive, we utilize the relative dataset size p_i as a surrogate measure. Clients with larger datasets are considered more reliable collaborators and should thus be assigned greater cooperative weight. Given these considerations, the optimization equation on the server side is defined as:

$$\min_{\boldsymbol{W}_{i*}} \sum_{j=1}^{N} \left((\boldsymbol{W}_{ij} - p^j)^2 + \alpha \mathcal{C}(\boldsymbol{W}_{ij}; D^i; D^j) - \beta \mathcal{S}(\boldsymbol{W}_{ij}; \boldsymbol{\theta}^i; \boldsymbol{\theta}^j) \right)$$

$$\text{s.t.} \qquad \sum_{j=1}^{N} \boldsymbol{W}_{ij} = 1, \forall i; \quad \boldsymbol{W}_{ij} \geq 0, \forall i, j \tag{8}$$

Using the cooperation network $\boldsymbol{W}$, we derive the aggregated model $\tilde{\boldsymbol{\theta}}^i = \sum_j \boldsymbol{W}_{ij} \boldsymbol{\theta}^j$ for each client.

Client Side. On each client side, our objective is to minimize the local empirical risk while preventing overfitting of the aggregated model on the local dataset. We replace the current local model with the aggregated model received from the server $\boldsymbol{\theta}^i \leftarrow \tilde{\boldsymbol{\theta}}^i$, and further refine this local model. For the i^{th} client, the optimization equation is defined as:

$$\arg\min_{\boldsymbol{\theta}^i} \ \mathcal{L}_i(\boldsymbol{\theta}^i; D^i) - \lambda \cos(\boldsymbol{\theta}^i, \tilde{\boldsymbol{\theta}}^i), \tag{9}$$

where $\cos(\boldsymbol{\theta}^i, \tilde{\boldsymbol{\theta}}^i)$ ensures that the locally optimized model does not deviate excessively from the aggregated model, and λ represents the regularization hyperparameter. The optimized model then serves as the current local model for participation in the subsequent optimization round.

4.3.3. Computational Complexity Analysis

Our approach incurs additional time overhead compared to classical federated learning methods, aiming to provide informative feedback for client collaboration to obtain more suitable personalized models. This extra time expenditure primarily stems from three aspects: computing the similarity metric $\mathcal{S}$, computing the complementarity metric $\mathcal{C}$, and solving the optimization equation. Here, we analyze the computational complexity and demonstrate that this overhead is acceptable.

The computational complexity of the similarity measure $\mathcal{S}$ is proportional to the model parameters, approximately equivalent to one inference time. For the complementarity measure $\mathcal{C}$, the feature matrix X is inferred. In cases of large local sample volumes, random sampling can approximate the local data distribution for effective computation. Assuming random sampling of m samples with feature dimension d, the resulting feature matrix $X \in \mathbb{R}^{m\times d}$ is processed. The complexity of extracting the representative subspace using SVD is $O(md^2)$, which is almost negligible compared to the duration of model training.

The subsequent step involves solving the optimization equation. Notably, each row of the adjacency matrix W in Equation 8 is independent, allowing for the independent computation of cooperation weights for each client with others. The optimization equation is simplified as follows:

$$\begin{aligned}\min_{\boldsymbol{W}_{i*}} \sum_{j=1}^{N} \Bigg(&\boldsymbol{W}_{ij}^2 - 2p^j \boldsymbol{W}_{ij} + (p^j)^2 + \alpha \boldsymbol{W}_{ij} \cos\left(\frac{1}{k}\sum_{l} \phi_l\right) \\ &- \beta \boldsymbol{W}_{ij} \frac{\boldsymbol{\theta}^i \cdot \boldsymbol{\theta}^j}{\|\boldsymbol{\theta}^i\| \cdot \|\boldsymbol{\theta}^j\|} \Bigg) \\ \text{s.t.} \quad &\sum_{j=1}^{N} \boldsymbol{W}_{ij} = 1, \forall i; \quad \boldsymbol{W}_{ij} \geq 0, \forall i, j \end{aligned} \tag{10}$$

The objective of the optimization equation simplifies to $\sum_{j=1}^{N}(\boldsymbol{W}_{ij}^2 + \phi_{ij}\boldsymbol{W}_{ij})$, forming a quadratic optimization function. This function is convex due to its compliance with the convex set inequality constraint $\boldsymbol{W}_{ij} \geq 0, \forall i, j$ and the affine transformation equality constraint $\sum_{j=1}^{N} \boldsymbol{W}_{ij} = 1, \forall i$. Therefore, this optimization problem is a convex optimization problem, solvable using convex optimization solvers, which can rapidly find the unique optimal solution.

5. Experiments

5.1. Unimodal Experiments Setup

Datasets and Data Heterogeneity. Following the predominant experimental setup in personalized federated learning, we evaluate our proposed FedSaC on two image classification datasets: CIFAR-10 and CIFAR-100 (Krizhevsky et al., 2009). For each dataset, we implement four partitions with different heterogeneous levels into K clients (Ye et al., 2023). 1) Homogeneous partition, where each client is imbued with data samples under a uniform probability schema. 2) Dirichlet partition (Yurochkin et al., 2019), where the allocation ratio of data samples from each category is instantiated from $Dir_K(\alpha)$. Notably, we define heterogeneity levels with $\alpha = 0.1$ (high) and $\alpha = 0.5$ (low). 3) Pathological partition, where each client is assigned with data exclusively from 2 categories for 10 classification datasets and 20 categories for 100 classification datasets.

Baselines. We compare our FedSaC with 12 representative FL approaches including: 1) **Local**: Local training without information sharing. 2) *FedAvg* (McMahan et al., 2017) and 3) *FedProx* (Deng et al., 2020): Popular FL methods where local updates are centrally aggregated. 4) *CFL* (Sattler et al., 2021): Clustered FL for client group learning. 5) *pFedMe* (Fallah et al., 2020): Using regularized loss functions to decouple local and global models. 6) *Ditto* (Li et al., 2021b): Enhanced robustness and fairness by regularized optimization. 7) *FedAMP* (Huang et al., 2021): Pairwise collaboration between similar clients in FL. 8) *FedRep* (Collins et al., 2021): Shared data representation with local client heads. 9) *pFedHN* (Shamsian et al., 2021): Hypernetworks generate unique client models in personalized FL. 10) *FedRoD* (Chen & Chao, 2022): Decoupled framework balances generic and personalized predictors 11) *kNN-Per* (Marfoq et al., 2021): Personalization via global embeddings and local kNN interpolation. 12)*pFedGraph* (Ye et al., 2023): Adaptive collaboration via learned graph.

5.2. Unimodal Experimental Results

Our experimental evaluations, conducted across various levels of heterogeneity on the CIFAR-10 and CIFAR-100 datasets, conclusively demonstrate the superior performance of our proposed model, FedSaC. Our analysis of the results presented in Table 1 leads to two primary insights:

Superiority Across Heterogeneity Levels. FedSaC consistently surpasses baseline models in various heterogeneity settings, highlighting its superior performance. Experiments show that our personalized method notably outperforms conventional techniques like FedAvg, especially in situations with statistical heterogeneity. Furthermore, FedSaC demonstrates significant or comparable enhancements over pFedGraph, a method based on similarity metrics. This comparison emphasizes the efficacy of our balanced approach to similarity and complementarity, a crucial aspect in federated learning for handling diverse data distributions.

Enhanced Performance under Strong Complementarity. FedSaC excels in scenarios with significant complementarity, such as those involving Dirichlet partitioning. These settings often feature imbalanced data distributions, posing challenges for local models with limited data categories. FedSaC's effective balance of similarity and complementarity addresses these challenges, enhancing data representation. In Dirichlet partitioning, it consistently surpasses pFedGraph, which relies on similarity metrics, improving accuracy by 3% to 8%. This alignment of empirical results

Dataset	CIFAR-10				CIFAR-100			
H-Level	Homo	Diri(low)	Diri(high)	Pathol	Homo	Diri(low)	Diri(high)	Pathol
Local	54.81	72.72	83.83	91.07	16.13	30.63	47.68	49.42
FedAvg	67.12	63.61	62.92	66.19	31.10	30.66	27.78	26.23
FedProx	62.92	62.93	62.25	55.76	30.55	30.64	27.87	25.64
CFL	60.55	73.81	83.84	90.76	19.31	33.21	49.12	52.43
pFedMe	47.48	66.35	75.24	81.73	13.18	25.18	34.37	33.48
Ditto	65.35	75.98	83.78	89.41	29.41	39.73	50.33	50.54
FedAMP	45.49	64.29	75.49	86.90	10.07	22.66	31.04	37.50
FedRep	62.88	74.14	83.47	90.02	21.53	34.72	50.15	26.23
pFedHN	62.78	66.62	82.57	89.91	25.94	30.89	49.08	49.06
FedRoD	62.07	74.06	83.49	90.66	18.71	31.65	47.96	49.91
kNN-Per	67.01	63.05	70.05	79.09	31.04	30.95	25.84	24.70
pFedGraph	67.37	75.22	84.28	92.74	31.16	38.71	51.63	56.79
FedSaC	**70.89**	**80.46**	**93.14**	**92.89**	**34.84**	**41.80**	**56.27**	**57.48**

Table 1. Average accuracy of our unimodal FedSaC on CIFAR-10 and CIFAR-100 dataset

with our theoretical framework confirms the effectiveness and versatility of FedSaC in various federated learning contexts, especially with high data heterogeneity.

5.3. Unimodal Experiments with Larger Model

We prove the applicability of our method on larger models. We opted for ResNet18 over the simple CNN model, while maintaining the experimental setup as previously described. The experiments were conducted on the CIFAR-100 dataset, with the results shown in Table 2.

Dataset	CIFAR-100 with ResNet18			
H-Level	Homo	Diri(low)	Diri(high)	Pathol
Local	23.69	39.19	57.45	61.19
FedAvg	40.50	39.85	37.15	34.90
FedProx	40.88	38.67	37.17	35.63
CFL	41.18	48.87	63.99	67.76
pFedMe	13.18	25.18	34.37	33.48
Ditto	40.52	51.07	63.03	66.71
FedAMP	9.28	20.46	34.59	33.74
FedRep	32.39	45.52	62.14	64.33
FedRoD	30.22	44.24	49.08	55.02
kNN-Per	41.32	40.38	39.35	41.16
pFedGraph	40.25	51.87	64.91	67.96
FedSaC	**42.62**	**54.01**	**67.75**	**68.16**

Table 2. Average accuracy of our FedSaC on the CIFAR-100 dataset with ResNet18 model

The results demonstrate that our FedSaC consistently outperforms the baseline models when using larger models. Notably, under the Dirichlet partition, which introduces appropriate data heterogeneity, our method significantly surpasses existing approaches. This aligns with our analysis, highlighting the complementary advantages of our approach. The experiment further supports the applicability of our method.

5.4. Multimodal Experiments Setup

Datasets and Baselines. In our multimodal experiments, we employ the *CUB200-2011* (Welinder et al., 2010) multimodal dataset, which encompasses two modalities—images and text—to undertake the task of classifying 200 bird species. For multimodal baselines, we not only compare with local training but also extend unimodal methods *FedAvg* and *pFedGraph* to the multimodal context, executing tasks separately within each modality. Additionally, we incorporate the multimodal federated learning method *FedIoT* (Zhao et al., 2022) for comparison. This method conducts unsupervised training on local clients and supervised aggregation on the server.

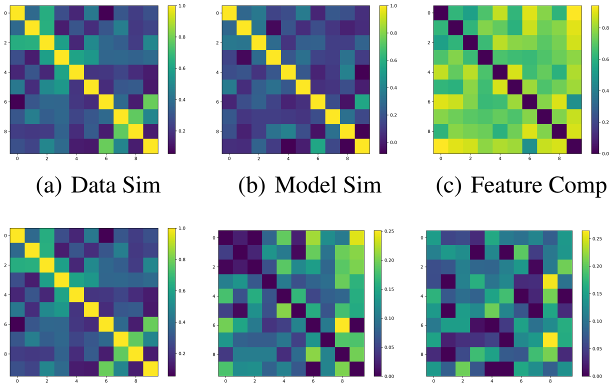
Multimodal Setup. Our proficient unimodal FedSaC method has been expanded to multimodal experimentation. Unlike unimodal scenarios, the multimodal approach leverages inter-client complementarity to enhance personalized model performance and utilizes inter-modality complementarity to contribute additional information to the model. Therefore, we introduce a strategy for the fusion of multimodal information complementarity. The specific setup details will be presented in Appendix B.

5.5. Multimodal Experimental Results

The multimodal experimental results, as depicted in Table 3, demonstrate that our FedSaC method surpasses all baselines. It significantly outperforms FedIoT, a method tailored for multimodal federated learning, which validates the efficacy of FedSaC in handling complex multimodal data. Particularly in scenarios modeled by Dirichlet distributions, our method demonstrates a distinct advantage over other baselines, reflecting a consistent trend with our unimodal experiment outcomes. Notably, we observe a more pronounced improvement in the visual modality post-cooperation, suggesting that visual data may provide richer information that enhances the robustness of our FedSaC.

H-Level	Homo		Diri(low)		Diri(high)		Pathol	
modal	visual	textual	visual	textual	visual	textual	visual	textual
Local	9.42	9.65	18.45	17.01	28.46	24.44	19.81	17.12
FedAvg	20.13	16.97	19.81	16.48	21.41	16.01	19.80	16.65
FedIoT	19.96	17.24	19.92	17.45	20.04	17.23	19.69	16.95
pFedGraph	21.87	21.59	24.48	25.13	31.78	31.34	25.50	27.06
FedSaC	**25.25**	**22.97**	**28.62**	**27.40**	**35.36**	**33.44**	**30.07**	**27.52**

Table 3. Average accuracy of our multimodal FedSaC on CUB200-2011 dataset


(a) Data Sim (b) Model Sim (c) Feature Comp

(d) Sim Network (e) Comp Network (f) Bal Network

Figure 3. Visualization of FedSaC: local data, process matrices, and cooperation networks under three collaboration states

5.6. Visualization

In our `FedSaC` visualization, Figure 3 presents core matrices and cooperation networks. Figure 3(a) shows local data's cosine similarity, while Figures 3(b) and 3(c) display the model similarity and feature complementarity matrices, respectively. The comparison of Figures 3(a) and 3(c) demonstrates a complementary pattern, affirming our metric's effectiveness in capturing local data relationships under privacy constraints. Figures 3(d) to 3(f) depict cooperation networks under three collaboration scenarios: focusing on similarity, complementarity, and a balance of both.

It is observed that in the similarity-based network, clients predominantly maintain their own models, hindering cooperative effectiveness and information gain. In the complementarity network, clients almost completely abandon their initial states, which is disadvantageous for training. The balanced approach allows for probabilistic exploration while filtering out clients with excessively high heterogeneity, as indicated by the darker areas that also show inconsistency in the local data matrix. The visualization underscores our method's role in boosting FL collaboration efficiency.

5.7. Hyperparameter Discussion

In the Appendix D.1, we present the experimental results for hyperparameters. Here, we discuss the selection of hyperparameters.

Among the four hyperparameters discussed in our paper $(\alpha, \beta, \lambda, k)$, there's no need to fine-tune λ and k. As shown in Appendix D.1, as long as λ and k are within a reasonable range ($\lambda = 5e^{-3} \sim 5e^{-2}, k = 2 \sim 3$), results are stable across various scenarios. For α and β, we search for well-performing values based on validation sets. As explored in Appendix B.1, $\alpha = 0.9$ and $\beta = 1.4$ are stable for most scenarios, including Homogeneity and Dirichlet partitions. For scenarios with high heterogeneity and low complementarity, such as the Pathological partition, adjusting to $\alpha = 0.5$ and $\beta = 1.6$ is suggested.

In most practical scenarios, tuning α and β is unnecessary. Our recommended values of $\alpha = 0.9$ and $\beta = 1.4$ are generally effective. For data with excessive heterogeneity, reducing complementarity weight is advisable. To achieve optimal results, a portion of the dataset can be used as a validation set for hyperparameter search. This pre-usage process incurs minimal additional costs and is practical for large-scale applications.

5.8. More Discussion

Large-Scale Clients. In our experiments with a smaller scale of client data, we enhanced cooperation efficiency in large-scale client collaborations (e.g., with 50 or 100 clients) by randomly selecting a subset of clients in each iteration. The feasibility of this approach is demonstrated in Appendix D.2.

Communication Overhead. Despite the additional steps introduced for optimal cooperation, Appendix E provides run-time consumption within each phases and confirms that the extra computational cost is minimal compared to local training, and thus acceptable.

6. Conclusion

In this study, we investigate the complex dynamics of federated learning, mitigating the significant challenge of statistical heterogeneity. We shift the focus from model similarity to a balance between similarity and feature complementarity. Our framework, FedSaC, effectively constructs a cooperation network by optimizing this balance. Extensive experiments show FedSaC's superiority over current FL methods in various scenarios. This research challenges conventional approaches and contributes to developing more robust learning models for complex federated settings.

Impact Statement

This study presents the `FedSaC` framework, offering a strategic approach to address statistical heterogeneity in Federated Learning. Academically, it introduces a novel perspective to FL, encouraging future research to explore the interplay of complementarity and similarity in model cooperation. Practically, this framework can be flexibly applied across various industries, facilitating more efficient and privacy-preserving data analysis models. Ethically, `FedSaC` aligns with the increasing demand for ethical data use and user privacy in technological advancements. Future work will further investigate the significance of balancing similarity and complementarity in multimodal architectures.

Acknowledgements

This work is funded by the Natural Science Fundation of China(NSFC. No. 62176132). BH was supported by the NSFC General Program No. 62376235 and CCF-Baidu Open Fund. MS was supported by Institute for AI and Beyond, UTokyo and by a grant from Apple, Inc. Any views, opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and should not be interpreted as reflecting the views, policies or position, either expressed or implied, of Apple Inc.

References

Agarwal, N., Suresh, A. T., Yu, F. X., Kumar, S., and McMahan, B. cpsgd: Communication-efficient and differentially-private distributed SGD. In Bengio, S., Wallach, H. M., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 7575–7586, 2018.

Baek, J., Jeong, W., Jin, J., Yoon, J., and Hwang, S. J. Personalized subgraph federated learning. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 1396–1415. PMLR, 2023.

Canonaco, G., Bergamasco, A., Mongelluzzo, A., and Roveri, M. Adaptive federated learning in presence of concept drift. In *International Joint Conference on Neural Networks, IJCNN 2021, Shenzhen, China, July 18-22, 2021*, pp. 1–7. IEEE, 2021.

Chen, H. and Chao, W. On bridging generic and personalized federated learning for image classification. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022.

Chen, J. and Zhang, A. HGMF: heterogeneous graph-based fusion for multimodal data with incompleteness. In Gupta, R., Liu, Y., Tang, J., and Prakash, B. A. (eds.), *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020*, pp. 1295–1305. ACM, 2020.

Collins, L., Hassani, H., Mokhtari, A., and Shakkottai, S. Exploiting shared representations for personalized federated learning. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 2089–2099. PMLR, 2021.

Cui, S., Liang, J., Pan, W., Chen, K., Zhang, C., and Wang, F. Collaboration equilibrium in federated learning. In Zhang, A. and Rangwala, H. (eds.), *KDD '22: The 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, August 14 - 18, 2022*, pp. 241–251. ACM, 2022.

Deng, Y., Kamani, M. M., and Mahdavi, M. Adaptive personalized federated learning. *CoRR*, abs/2003.13461, 2020.

Dinh, C. T., Tran, N. H., and Nguyen, T. D. Personalized federated learning with moreau envelopes. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Fallah, A., Mokhtari, A., and Ozdaglar, A. E. Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Gan, S., Mathur, A., Isopoussu, A., Kawsar, F., Berthouze, N., and Lane, N. D. Fruda: Framework for distributed adversarial domain adaptation. *CoRR*, abs/2112.13381, 2021.

Ghosh, A., Chung, J., Yin, D., and Ramchandran, K. An efficient framework for clustered federated learning. *IEEE Trans. Inf. Theory*, 68(12):8076–8091, 2022.

Huang, W., Ye, M., and Du, B. Learn from others and be yourself in heterogeneous federated learning. In

IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022, pp. 10133–10143. IEEE, 2022.

Huang, Y., Chu, L., Zhou, Z., Wang, L., Liu, J., Pei, J., and Zhang, Y. Personalized cross-silo federated learning on non-iid data. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pp. 7865–7873. AAAI Press, 2021.

Jothimurugesan, E., Hsieh, K., Wang, J., Joshi, G., and Gibbons, P. B. Federated learning under distributed concept drift. In Ruiz, F. J. R., Dy, J. G., and van de Meent, J. (eds.), *International Conference on Artificial Intelligence and Statistics, 25-27 April 2023, Palau de Congressos, Valencia, Spain*, volume 206 of *Proceedings of Machine Learning Research*, pp. 5834–5853. PMLR, 2023.

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K. A., Charles, Z., Cormode, G., Cummings, R., D'Oliveira, R. G. L., Eichner, H., Rouayheb, S. E., Evans, D., Gardner, J., Garrett, Z., Gascón, A., Ghazi, B., Gibbons, P. B., Gruteser, M., Harchaoui, Z., He, C., He, L., Huo, Z., Hutchinson, B., Hsu, J., Jaggi, M., Javidi, T., Joshi, G., Khodak, M., Konečný, J., Korolova, A., Koushanfar, F., Koyejo, S., Lepoint, T., Liu, Y., Mittal, P., Mohri, M., Nock, R., Özgür, A., Pagh, R., Qi, H., Ramage, D., Raskar, R., Raykova, M., Song, D., Song, W., Stich, S. U., Sun, Z., Suresh, A. T., Tramèr, F., Vepakomma, P., Wang, J., Xiong, L., Xu, Z., Yang, Q., Yu, F. X., Yu, H., and Zhao, S. Advances and open problems in federated learning. *Found. Trends Mach. Learn.*, 14(1-2):1–210, 2021.

Karimireddy, S. P., Kale, S., Mohri, M., Reddi, S. J., Stich, S. U., and Suresh, A. T. SCAFFOLD: stochastic controlled averaging for federated learning. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5132–5143. PMLR, 2020.

Konečný, J., McMahan, H. B., Yu, F. X., Richtárik, P., Suresh, A. T., and Bacon, D. Federated learning: Strategies for improving communication efficiency. *CoRR*, abs/1610.05492, 2016.

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

Li, B., Schmidt, M. N., Alstrøm, T. S., and Stich, S. U. On the effectiveness of partial variance reduction in federated learning with heterogeneous data. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pp. 3964–3973. IEEE, 2023.

Li, Q., He, B., and Song, D. Model-contrastive federated learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pp. 10713–10722. Computer Vision Foundation / IEEE, 2021a.

Li, Q., Diao, Y., Chen, Q., and He, B. Federated learning on non-iid data silos: An experimental study. In *38th IEEE International Conference on Data Engineering, ICDE 2022, Kuala Lumpur, Malaysia, May 9-12, 2022*, pp. 965–978. IEEE, 2022.

Li, T., Sahu, A. K., Zaheer, M., Sanjabi, M., Talwalkar, A., and Smith, V. Federated optimization in heterogeneous networks. In Dhillon, I. S., Papailiopoulos, D. S., and Sze, V. (eds.), *Proceedings of Machine Learning and Systems 2020, MLSys 2020, Austin, TX, USA, March 2-4, 2020*. mlsys.org, 2020.

Li, T., Hu, S., Beirami, A., and Smith, V. Ditto: Fair and robust federated learning through personalization. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 6357–6368. PMLR, 2021b.

Lin, S., Han, Y., Li, X., and Zhang, Z. Personalized federated learning towards communication efficiency, robustness and fairness. In *NeurIPS*, 2022.

Lin, Y., Gao, Y., Gong, M., Zhang, S., Zhang, Y., and Li, Z. Federated learning on multimodal data: A comprehensive survey. *Mach. Intell. Res.*, 20(4):539–553, 2023.

Marfoq, O., Neglia, G., Kameni, L., and Vidal, R. Personalized federated learning through local memorization. *CoRR*, abs/2111.09360, 2021.

McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. Communication-efficient learning of deep networks from decentralized data. In Singh, A. and Zhu, X. J. (eds.), *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017, 20-22 April 2017, Fort Lauderdale, FL, USA*, volume 54 of *Proceedings of Machine Learning Research*, pp. 1273–1282. PMLR, 2017.

Miao, J. and Ben-Israel, A. On principal angles between subspaces in rn. *Linear algebra and its applications*, 171: 81–98, 1992.

Mohri, M., Sivek, G., and Suresh, A. T. Agnostic federated learning. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4615–4625. PMLR, 2019.

Mothukuri, V., Parizi, R. M., Pouriyeh, S., Huang, Y., Dehghantanha, A., and Srivastava, G. A survey on security and privacy of federated learning. *Future Gener. Comput. Syst.*, 115:619–640, 2021.

Peng, X., Huang, Z., Zhu, Y., and Saenko, K. Federated adversarial domain adaptation. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

Qin, Z., Yang, L., Wang, Q., Han, Y., and Hu, Q. Reliable and interpretable personalized federated learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pp. 20422–20431. IEEE, 2023.

Qu, L., Zhou, Y., Liang, P. P., Xia, Y., Wang, F., Adeli, E., Fei-Fei, L., and Rubin, D. L. Rethinking architecture design for tackling data heterogeneity in federated learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 10051–10061. IEEE, 2022.

Qu, Z., Li, X., Han, X., Duan, R., Shen, C., and Chen, L. How to prevent the poor performance clients for personalized federated learning? In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pp. 12167–12176. IEEE, 2023.

Rajkumar, K., Goswami, A., Lakshmanan, K., and Gupta, R. Comment on "federated learning with differential privacy: Algorithms and performance analysis". *IEEE Trans. Inf. Forensics Secur.*, 17:3922–3924, 2022.

Sattler, F., Müller, K., and Samek, W. Clustered federated learning: Model-agnostic distributed multitask optimization under privacy constraints. *IEEE Trans. Neural Networks Learn. Syst.*, 32(8):3710–3722, 2021.

Shamsian, A., Navon, A., Fetaya, E., and Chechik, G. Personalized federated learning using hypernetworks. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 9489–9502. PMLR, 2021.

Welinder, P., Branson, S., Mita, T., Wah, C., Schroff, F., Belongie, S., and Perona, P. Caltech-ucsd birds 200. 2010.

Wu, C., Wu, F., Cao, Y., Huang, Y., and Xie, X. Fedgnn: Federated graph neural network for privacy-preserving recommendation. *CoRR*, abs/2102.04925, 2021.

Xiong, B., Yang, X., Qi, F., and Xu, C. A unified framework for multi-modal federated learning. *Neurocomputing*, 480:110–118, 2022.

Ye, R., Ni, Z., Wu, F., Chen, S., and Wang, Y. Personalized federated learning with inferred collaboration graphs. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 39801–39817. PMLR, 2023.

Yu, Q., Liu, Y., Wang, Y., Xu, K., and Liu, J. Multimodal federated learning via contrastive representation ensemble. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.

Yurochkin, M., Agarwal, M., Ghosh, S., Greenewald, K. H., Hoang, T. N., and Khazaeni, Y. Bayesian nonparametric federated learning of neural networks. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 7252–7261. PMLR, 2019.

Zhao, Y., Barnaghi, P. M., and Haddadi, H. Multimodal federated learning. *CoRR*, abs/2109.04833, 2021.

Zhao, Y., Barnaghi, P. M., and Haddadi, H. Multimodal federated learning on iot data. In *Seventh IEEE/ACM International Conference on Internet-of-Things Design and Implementation, IoTDI 2022, Milano, Italy, May 4-6, 2022*, pp. 43–54. IEEE, 2022.

Zheng, T., Li, A., Chen, Z., Wang, H., and Luo, J. Autofed: Heterogeneity-aware federated multimodal learning for robust autonomous driving. In Costa-Pérez, X., Widmer, J., Perino, D., Giustiniano, D., Al-Hassanieh, H., Asadi, A., and Cox, L. P. (eds.), *Proceedings of the 29th Annual International Conference on Mobile Computing and Networking, ACM MobiCom 2023, Madrid, Spain, October 2-6, 2023*, pp. 15:1–15:15. ACM, 2023.

Zhu, J., Ma, X., and Blaschko, M. B. Confidence-aware personalized federated learning via variational expectation maximization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pp. 24542–24551. IEEE, 2023.

A. Discussions about FedSac.

A.1. FedSac Optimization

Global Optimization. In our research, the overarching optimization equation is presented as Equation 3, which is fundamentally grounded in the optimization objective of FedAvg (McMahan et al., 2017). The equation is expressed as follows:

$$\min_{\{\boldsymbol{\theta}^i\}} \sum_{i=1}^{N} p^i \mathcal{L}\left(\boldsymbol{\theta}^g; D^i\right), \tag{11}$$

This equation is restructured to align with our targeted optimization goals. The global model θ_g can be represented as a weighted aggregation of local models, with the weights corresponding to the relative sizes of each client's dataset. The revised formulation is presented thusly:

$$\begin{aligned} &\min_{\{\boldsymbol{\theta}^i\},\boldsymbol{W}} \sum_{i=1}^{N} p^i \mathcal{L}_i(\sum_{j=1}^{N} \boldsymbol{W}_{ij}\boldsymbol{\theta}^j; D^i) \\ \text{s.t.} \quad &\boldsymbol{W}_{ij} = p^j, \forall i,j; \quad \sum_{j=1}^{N} \boldsymbol{W}_{ij} = 1, \forall i; \quad \boldsymbol{W}_{ij} \geq 0, \forall i,j \end{aligned} \tag{12}$$

Further to this, we introduced two additional regularization terms to balance similarity and complementarity. $\mathcal{C}$ objective reduces cooperation intensity between clients with similar datasets, while $\mathcal{S}$ objective increases it for clients with similar model parameters.

Server-Client Optimization. Within the federated learning framework, Equation 3 poses practical challenges, as clients should not directly receive models from other clients. Consequently, transferring models to a centralized server becomes essential. At the server side, we aim to estimate the first term of Equation 3, namely the empirical loss of local models. In line with our optimization objectives, which are aligned with FedAvg, we adopt a FedAvg-inspired approach. Here, we approximate the empirical loss using the relative sizes of the datasets, operating under the premise that clients with larger local datasets are more suitable for collaboration. This concept has been validated for its rationality (Ye et al., 2023) in federated learning scenarios.

A.2. The Metric of Similarity and Complementarity

Similarity Metric. In the realm of federated learning, utilizing model parameters to gauge client similarity is a prevalent approach (Ye et al., 2023; Huang et al., 2021). Aligning with the approach in (Ye et al., 2023), we utilize cosine distance of model parameters for similarity assessment.

Complementarity Metric. Considering the privacy concerns in federated learning, direct computation of distances using local datasets is not feasible. Instead, we draw upon the principle angle method, a technique that measures distances between subspaces (Miao & Ben-Israel, 1992). This adapted approach relies on limited, non-sensitive information to determine the degree of similarity between clients.

The principle angle method offers a geometric perspective for measuring the distance between subspaces. Specifically, when dealing with two subspaces, $\boldsymbol{V}$ and $\boldsymbol{W}$, the method determines the angles $\theta_1, \theta_2, \cdots, \theta_k$ between them. Here, k represents the number of dimensions in the smaller of the two subspaces. The calculation of the i^{th} principal angle is as described in Equation 6. This implies that the cosine of the largest principal angle corresponds to the largest singular value of the matrix product $V^T W$.

$$\cos\theta_i = \frac{< x_i, y_i >}{\|x_i\| \cdot \|y_i\|} = \max_{x_i \in \boldsymbol{V}, y_i \in \boldsymbol{W}} \frac{< x_i, y_i >}{\|x_i\| \cdot \|y_i\|}. \tag{13}$$

The principal angle method provides a clear geometric perspective on how similar or different two subspaces are. By measuring the angles between the subspaces, it offers a more intuitive understanding of their relationship.

In our method, we harness the principal angle method to effectively represent the local data distributions of clients as model output features. This is achieved through SVD, where we select the leading k principal component vectors to represent each client's data in a subspace. This approach maps varied client datasets to a common feature space and ensures privacy by using only a few principal components, which are insufficient to reconstruct the original data. The technique aligns data from different clients effectively while maintaining privacy in federated learning.

B. Experiments and Implementation Details

B.1. Unimodal Implementation Details

Basic Setup. Adhering to the training setting presented in (Ye et al., 2023) , we partition the dataset across 10 local clients. Each client utilizes a simple CNN classifier consisting of 2 convolutional layers, 2 subsequent fully-connected layers and a final classification layer. Notably, the representation dimension prior to the classification layer is set at 84, which will be utilized for extracting the representative subspace to compute the complementarity. In the FL training phase, we execute 50 communication rounds. Each round consists of local training iterations that vary depending on the dataset: 200 iterations for CIFAR-10 and 400 iterations for CIFAR-100. Training employs the SGD optimizer with an initial learning rate 0.01 and a batch size 64.

Hyperparameter Setup. We have set key hyperparameters for optimal performance. We use three eigenvectors ($k = 3$) for our representative subspace. The regularization hyperparameter λ is set at 1. Additionally, the hyperparameters α and β control the degree of complementarity and similarity in our optimization equation. In experiments, we consider two scenarios based on client dataset characteristics. For datasets with complementarity, $\alpha = 0.9$ and $\beta = 1.4$ balance similarity and complementarity for enhanced performance. In contrast, for datasets lacking complementarity, such as in the Pathological partition, we reduce complementarity by setting $\alpha = 0.5$ and $\beta = 1.6$. The first setting is generally applied unless low complementarity among clients is known, in which case the second setting is used. To facilitate convergence, we use the initial settings for the first 70% of communication rounds, then set $\alpha = 0$ in the remaining rounds.

B.2. Multimodal Implementation Setup

In our setup, we distribute the *CUB200-2011* dataset among clients, with an equal split between image and text modalities. Each client possesses distinct feature extraction networks and uniform classification networks to fulfill the classification task. Within the same modality, we employ the unimodal FedSaC method to facilitate cooperation among clients. For cross-modality cooperation, structural differences in feature extraction layers necessitate restricting collaborative efforts to the classification layer. Given the inherent complementarity between different modalities, we focus on the similarity within the classification layers during cooperation. The cooperation weights for the classification layers are derived by excluding the complementarity term $\mathcal{C}$ from the optimization 8. These weights are used to aggregate the classification layers at the server side, enabling the effective fusion of cross-modal information.

B.3. Multimodal Implementation Details

We allocate the CUB200-2011 dataset across 8 clients, with 4 handling image data and the others processing text data. For image modality clients, we employ a CNN architecture with four convolutional layers and a single classification layer. Text clients, on the other hand, utilize a TextCNN network consisting of five convolutional layers and a classification layer. The representation dimension is set at 256. The training involves 30 communication rounds, each comprising 200 iterations. We employ the Adam optimizer with an initial learning rate of 0.001. Throughout the training, we adopt a balanced setting for similarity and complementarity, with $\alpha = 0.7$ and $\beta = 1.2$, keeping the rest of the setup consistent with the unimodal FedSaC.

B.4. DataSets

In our experiments, CIFAR-10, CIFAR-100 (Krizhevsky et al., 2009) and CUB200-2011 (Welinder et al., 2010) are all public dataset.

CIFAR-10 and CIFAR-100. The CIFAR-10 and CIFAR-100 datasets are key benchmarks in machine learning, each containing 60,000 32x32 color images. CIFAR-10 is categorized into 10 classes with 6,000 images per class, suitable for basic image recognition. CIFAR-100, offering a finer classification challenge, divides the same number of images across 100 classes, with 600 images per class. Both datasets, split into 50,000 training and 10,000 test images, are extensively used

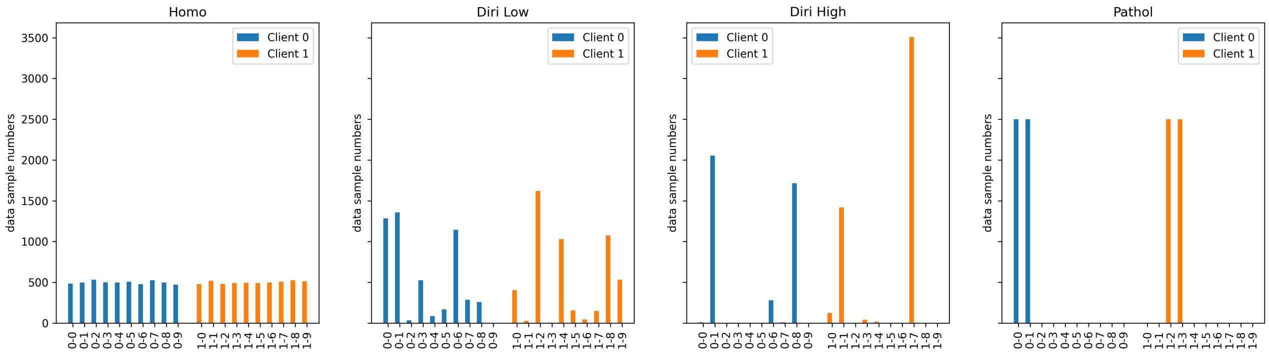


Figure 4. Illustration of the level of heterogeneity under four distinct partitioning schemes.

for evaluating image classification algorithms.

CUB200-2011. The CUB200-2011 dataset is specifically tailored for fine-grained visual categorization tasks, focusing on bird species identification. It consists of 11,788 images of 200 bird species, with both training and testing sets. Each species comes with a set of images that offer varying poses and backgrounds, providing a comprehensive dataset for advanced image recognition tasks. CUB200-2011 is particularly useful for research in areas requiring detailed visual discrimination, such as in distinguishing between closely related species.

Heterogeneity Partition. In our study, we employ the CIFAR-10 dataset and select two clients to illustrate the level of heterogeneity under four distinct partitioning schemes, shown as 4.

B.5. Baselines

FedAvg (McMahan et al., 2017) streamlines the training of deep networks from decentralized data in federated learning. It enables multiple clients to collaboratively train a shared model while maintaining data privacy and reducing communication overhead. Suitable for scenarios where central data collection is impractical due to privacy concerns, like in IoT and healthcare applications.

FedProx (Li et al., 2020) specifically tackles system and statistical heterogeneity in federated networks. It introduces a proximal term to the optimization objective, enhancing stability and accuracy in networks with devices of varying capabilities. This modification leads to more robust convergence and improved accuracy in heterogeneous settings.

CFL (Sattler et al., 2021) is designed for large-scale peer-to-peer networks, optimizing federated learning by aggregating local model updates in a hierarchical manner. It ensures communication efficiency and data privacy through secure and authenticated encryption techniques. CFL stands out for its significant improvement in communication and computational efficiency, while robustly maintaining data integrity and privacy.

pFedMe (Dinh et al., 2020) introduces a personalized federated learning algorithm using Moreau envelopes as clients' regularized loss functions, allowing for the decoupling of personalized model optimization from global model learning. pFedMe is effective in handling statistical diversity among clients, leading to state-of-the-art convergence rates and superior empirical performance compared to traditional FedAvg and Per-FedAvg algorithms.

Ditto (Li et al., 2021b) is a framework that enhances federated learning by simultaneously achieving fairness and robustness through personalization. It addresses the challenges of statistical heterogeneity in networks, using a simple yet scalable technique that improves accuracy, fairness, and robustness. Ditto is particularly effective against training-time data and model poisoning attacks and reduces performance disparities across devices.

FedAMP (Huang et al., 2021) This method employs federated attentive message passing to facilitate collaborations among clients with similar non-iid data, establishing convergence for both convex and non-convex models. FedAMP emphasizes pairwise collaborations between clients with similar data, overcoming the bottleneck of one global model trying to fit all clients in personalized cross-silo federated learning scenarios.

FedRep (Collins et al., 2021) utilizes a shared data representation across clients while allowing unique local heads for each client. This approach harnesses local updates concerning low-dimensional parameters, enabling efficient learning in heterogeneous data environments. By focusing on linear convergence and sample complexity, FedRep demonstrates improved performance over alternative personalized federated learning methods, especially in federated settings with non-iid data.

pFedHN (Shamsian et al., 2021) introduces a personalized federated learning approach using hypernetworks. This method trains a central hypernetwork to generate unique personal models for each client, effectively sharing parameters across clients. It excels in handling data disparities among clients, reducing communication costs, and generalizing better to new clients with varying distributions and computational resources.

FedRoD (Chen & Chao, 2022) simultaneously addresses generic and personalized learning objectives. It employs a two-loss, two-predictor system, decoupling the tasks of generic model training and personalized adaptation. The framework uses a class-balanced loss for the generic predictor and an empirical risk-based approach for the personalized predictor, facilitating robustness to non-identical class distributions and enabling zero-shot adaptation and effective fine-tuning for new clients.

kNN-Per (Marfoq et al., 2021) introduces local memorization using k-nearest neighbors in federated learning, enhancing the model's ability to personalize based on individual device data. This method stands out in its use of local data patterns to inform the federated learning process.

pFedGraph (Ye et al., 2023) proposes the construction of inferred collaboration graphs among clients in federated learning. It dynamically computes these graphs based on the volume of data and model similarity at each client. This method strategically identifies similar clients for cooperation, effectively mitigating issues arising from data heterogeneity.

FedIoT (Zhao et al., 2022) proposes a multimodal federated learning framework for IoT data, utilizing autoencoders to process multimodal data from clients. It introduces a multimodal FedAvg algorithm to aggregate local models from diverse data sources, enhancing classification performance in semi-supervised scenarios with unimodal and multimodal clients.

B.6. Computing Resources

Part of the experiments is conducted on a local server with Ubuntu 16.04 system. It has two physical CPU chips which are Intel(R) Xeon(R) Gold 6248 CPU @ 2.50GHz with 20 cpu cores. The other experiments are conducted on a remote server. It has 8 GPUs which are GeForce RTX 3090.

C. Privacy Discussion

Our `FedSaC` exhibits similar data privacy preservation compared with baselines, as it does not share any private data of the clients. During communication, only model parameters are allowed to be shared. Similar to baselines, the sharing of model parameters is intended to maintain data privacy. The representative subspaces are derived from local data feature statistics generated by the model, a method that does not reveal any privacy details of the original dataset. Our approach is also compatible with protective strategies like differential privacy (Rajkumar et al., 2022). Specifically, for representative subspaces, we primarily rely on calculating their principal angles. Therefore, we could apply methods such as random cropping and adding minor noise to ensure that the original data cannot be reconstructed.

D. Supplementary Experiments

D.1. Hyperparameters Experiments

The experimental analysis focused on evaluating the influence of hyperparameters, as illustrated in Figures 5 and 6.

Hyperparameter α. In Figure 5, the similarity hyperparameter, denoted as β, is fixed at 1.4, hile the complementary hyperparameter, α, varied from 0.6 to 1.2. The results indicate that in data partitions characterized by complementarity,

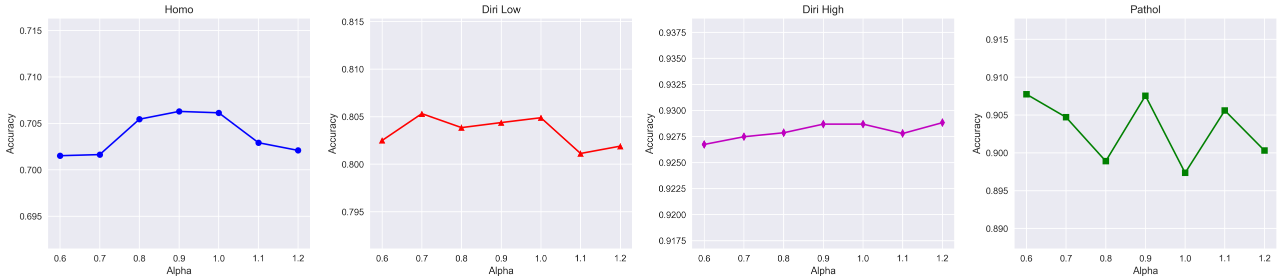


Figure 5. Average accuracy curves of the four partitions under various hyperparameter α settings.

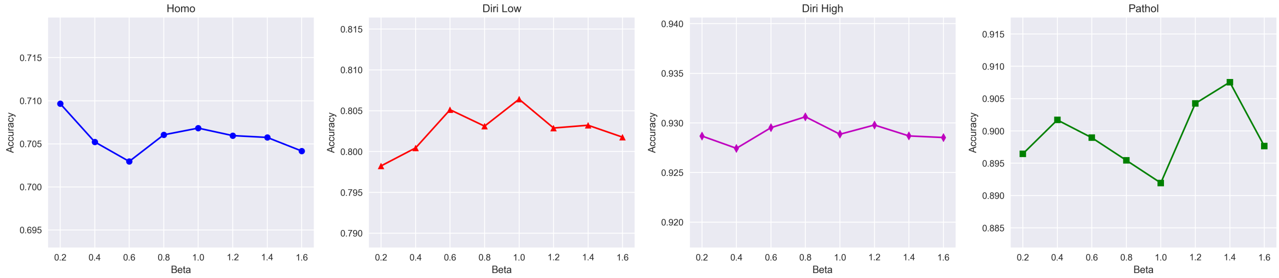


Figure 6. Average accuracy curves of the four partitions under various hyperparameter β settings.

a moderate increase in α enhances accuracy. However, in partitions with high heterogeneity, the influence of α on the outcomes exhibits fluctuations. Notably, the experimental results consistently outperform the baseline, irrespective of the variations in α.

Hyperparameter β. Figure 6 presents the outcomes with α set at 0.9, examining the impact of changes in β ranging from 0.2 to 1.6. It is observed that an optimal level of similarity substantially benefits the experimental results, which uniformly exceed the baseline performance.

Hyperparameter λ. We employed the CIFAR100 dataset to assess the impact of the hyperparameter λ, associated with regularization constraints in local training, as demonstrated in Table 4. The results indicate that setting λ to either 0.01 or 0.1 yields favorable outcomes with minimal fluctuation.

Hyperparameter k. the influence of the subspace dimensionality, represented by k, on the experimental outcomes was examined, as detailed in Table 5. The findings suggest that $k = 3$ is an appropriate choice for obtaining the representative subspace.

λ	Homo	Diri(low)	Diri(high)	Pathol
0.005	32.21	39.95	55.79	56.68
0.01	32.95	39.90	55.91	57.48
0.05	32.62	39.60	55.81	56.47
0.1	32.91	39.73	56.04	56.86
0.2	32.82	39.77	55.53	56.52

Table 4. Average accuracy of the four partitions under various hyperparameter λ settings.

k	Homo	Diri(low)	Diri(high)	Pathol
1	70.26	80.29	92.97	92.47
2	70.50	80.28	93.04	92.56
3	70.89	80.46	93.14	92.89
4	70.20	80.15	92.73	92.59

Table 5. Average accuracy of the four partitions under various hyperparameter k settings.

D.2. Experiments in Large-Scale Client Cooperation

In our experiments, we primarily focused on scenarios with a limited number of clients, specifically 8-10 clients. In situations involving a large number of clients, while the computational time overhead may not significantly impact our performance – a topic we will delve into in the following section – the cooperation among numerous clients could affect the convergence and stability of the collaboration. Therefore, for cooperation with a large client base, we incorporated an additional process to control the number of collaborators.

Specifically, in large-scale client cooperation, we randomly select k clients (k=10 in application) for collaboration before each iteration. This approach ensures convergence while enhancing cooperation efficiency. Table 7 presents the results in cooperation networks with 50 and 100 clients, incorporating this step. The results confirm the effectiveness of such collaborative efforts.

Modal	Client Num	Homo	Diri(low)	Diri(high)	Pathol
Local	50	8.63	22.45	45.36	33.33
	100	6.87	18.19	40.62	27.78
FedSaC	50	18.81	24.03	45.78	35.40
	100	11.90	19.11	40.75	28.51

Table 6. Large-scale client cooperation

Below, we also present the additional computational overhead incurred when not employing random selection in large-scale client cooperation. In real-world scenarios, local training on clients occurs simultaneously; hence, in our analysis, we do not consider an increase in local training duration with the number of clients. The following table shows how the server-side additional overhead increases with the number of clients.

Client Num	Server-Side Overhead	Optimization	Model Aggregation
10	4.40s	0.21s	4.19s
20	14.93s	6.10s	8.83s
50	56.56s	24.84s	31.72s
100	269.42s	118.70s	140.72s

Table 7. Server-side additional overhead for large-scale clients cooperation

Assuming the number of clients is n, the computational overhead on the server side increases approximately at a rate of n^2 with increasing clients. Within this overhead, the optimization process costs less than model aggregation. Therefore, the cost of the optimization process we introduce is acceptable in large-scale client cooperation, without significantly disrupting the original training process.

E. Computational Cost and Complexity Analysis

As analyzed in Section 4.3.3, our method indeed introduces additional steps to extract information, yet the costs are acceptable and significantly less than that of local training, for the following reasons.

1. Assuming the cost of one inference takes τ, our method shows the complexity of similarity metric is similar to the model's parameter count, roughly equal to an inference time of τ.

2. Complementarity metric involves obtaining the feature matrix X through model inference on the local dataset. For datasets with large sample sizes, random sampling can be used to approximate the local data distribution. Assuming the

sample size for random sampling is m with feature dimension d, the cost of acquiring the feature matrix $X \in \mathbb{R}^{m\times d}$ is $m\tau$.

3. Next, the feature matrix X is dimensionally reduced via the SVD method to obtain a representative subspace. This process involves computing X^TX and its eigenvalue decomposition, with a time complexity of $O(md^2)$. In practice, this computational cost is significantly lower than that of extracting the feature matrix.

4. Appendix E demonstrates that the solution for the optimal adjacency matrix W in Eq.8 is a convex optimization problem, conforming to a quadratic programming problem. By employing the interior point method, this problem can be transformed into polynomial-level complexity. The computational cost in our experiments is approximately equivalent to one inference time τ.

The table below details the specific run-time consumption at each phase in our experiments, demonstrating that the overhead of our additional steps is significantly less than that of local training. Our method enhances collaborative approaches without imposing excessive burdens.

Phases	Run-time Consumption
Local Training	21.21s
Similarity Metric	0.05s
Complementarity Metric	1.62s
Complementarity Metric (Inference)	1.53s
Complementarity Metric (SVD)	0.09s
Optimization Equation	0.06s

Table 8. Run-time consumption of across different phases.

We tested the runtime of each additional phase in our experiments on our platform and compared it with the training duration of a single client in one local training round, given ten clients, as shown in Table 8. The results indicate that the time required for similarity metric and solving the optimization equation is negligible compared to local training duration. Although the complementarity metric phase, which involves an inference process, does take some time, it is still significantly less than the local training duration. Therefore, the additional cost of cooperation is acceptable. As a result, `FedSaC` does not introduce substantial additional computational and communication costs, making its computational overhead comparable to existing baselines.

F. Convergence Analysis

The introduction of complementarity in our `FedSaC` approach does not lead to convergence issues. As depicted in Figure 7, we illustrate the accuracy progression over communication rounds on the CIFAR-100 dataset under a Diri(low) partition. It is observed that the accuracy of the FedSaC method steadily rises and gradually converges. Unlike local training, which may lead to overfitting and a subsequent decline in accuracy due to excessive training, our method effectively circumvents the overfitting problem. In contrast to other baselines that converge prematurely and potentially get trapped in local optima, our approach consistently explores better solutions, achieving optimal performance before ultimately converging.

G. Additional visualization

To illustrate the advantages of balancing similarity and complementarity more clearly, we present an visualization involving four clients.

We distribute data from four categories, 1, 2, 3 and 4, among these four clients as follows: client1: (1, 2); client2: (1, 3); client3: (2, 3); client4: (1, 4). The numbers in parentheses represent the categories, with each category having an equal amount of data. Below, we present the adjacency matrix W generated by four different cooperations.

$$W(\alpha=0,\beta=1.4)=\begin{pmatrix}0.684 & 0.137 & 0.011 & 0.168\\ 0.048 & 0.596 & 0.240 & 0.117\\ 0.003 & 0.320 & 0.676 & 0.000\\ 0.149 & 0.186 & 0.000 & 0.665\end{pmatrix} \quad W(\alpha=0.9,\beta=0)=\begin{pmatrix}0.043 & 0.344 & 0.285 & 0.328\\ 0.320 & 0.020 & 0.329 & 0.331\\ 0.260 & 0.328 & 0.018 & 0.395\\ 0.292 & 0.318 & 0.384 & 0.007\end{pmatrix}$$

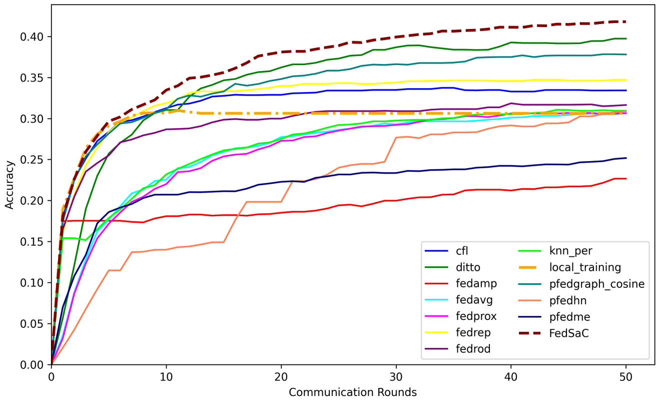


Figure 7. Illustration of the accuracy progression over communication rounds on the CIFAR-100 dataset under a Diri(low) partition.

$$W(\alpha = 0.9, \beta = 1.4) = \begin{pmatrix} 0.478 & 0.230 & 0.047 & 0.246 \\ 0.118 & 0.365 & 0.319 & 0.197 \\ 0.024 & 0.409 & 0.455 & 0.111 \\ 0.204 & 0.268 & 0.092 & 0.436 \end{pmatrix} \quad W(\alpha = 0.5, \beta = 1.6) = \begin{pmatrix} 0.630 & 0.172 & 0.000 & 0.198 \\ 0.058 & 0.517 & 0.282 & 0.142 \\ 0.000 & 0.382 & 0.616 & 0.002 \\ 0.171 & 0.228 & 0.000 & 0.602 \end{pmatrix}$$

The results align with our expectations and affirm the advantages of our method. When relying solely on similarity, with $\alpha = 0, \beta = 1.4$, clients seldom cooperate with others, preferring to retain their local models as much as possible. Conversely, when depending solely on complementarity with $\alpha = 0.9, \beta = 0$, clients tend to discard the recently learned local models, which is impractical. Achieving a balance between similarity and complementarity with $\alpha = 0.9, \beta = 1.4$ leads to a more ideal state of collaboration, where local clients not only increase cooperation with complementary clients but also maintain lower collaboration weights with those of excessive heterogeneity. The setting $\alpha = 0.5, \beta = 1.6$ offers another solution for scenarios where there is high heterogeneity among clients.