FedSAC: Dynamic Submodel Allocation for Collaborative Fairness in Federated Learning

Zihui Wang[*]
wangziwei@stu.xmu.edu.cn
Fujian Key Laboratory of Sensing and
Computing for Smart Cities, School of
Informatics, Xiamen University

Zheng Wang[*]
zwang@stu.xmu.edu.cn
Fujian Key Laboratory of Sensing and
Computing for Smart Cities, School of
Informatics, Xiamen University

Lingjuan Lyu
lingjuan.lv@sony.com
Sony AI

Pengzhao Peng
pengzhaopeng@stu.xmu.edu.cn
Fujian Key Laboratory of Sensing and
Computing for Smart Cities, School of
Informatics, Xiamen University

Zhicheng Yang
zcyang@stu.xmu.edu.cn
Fujian Key Laboratory of Sensing and
Computing for Smart Cities, School of
Informatics, Xiamen University

Chenglu Wen
clwen@xmu.edu.cn
Fujian Key Laboratory of Sensing and
Computing for Smart Cities, School of
Informatics, Xiamen University

Rongshan Yu
rsyu@xmu.edu.cn
Fujian Key Laboratory of Sensing and
Computing for Smart Cities, School of
Informatics, Xiamen University

Cheng Wang
cwang@xmu.edu.cn
Fujian Key Laboratory of Sensing and
Computing for Smart Cities, School of
Informatics, Xiamen University

Xiaoliang Fan[†]
fanxiaoliang@xmu.edu.cn
Fujian Key Laboratory of Sensing and
Computing for Smart Cities, School of
Informatics, Xiamen University

ABSTRACT
Collaborative fairness stands as an essential element in federated learning to encourage client participation by equitably distributing rewards based on individual contributions. Existing methods primarily focus on adjusting gradient allocations among clients to achieve collaborative fairness. However, they frequently overlook crucial factors such as maintaining consistency across local models and catering to the diverse requirements of high-contributing clients. This oversight inevitably decreases both fairness and model accuracy in practice. To address these issues, we propose FedSAC, a novel Federated learning framework with dynamic Submodel Allocation for Collaborative fairness, backed by a theoretical convergence guarantee. First, we present the concept of "*bounded collaborative fairness (BCF)*", which ensures fairness by tailoring rewards to individual clients based on their contributions. Second, to implement the BCF, we design a *submodel allocation module* with a theoretical guarantee of fairness. This module incentivizes high-contributing clients with high-performance submodels containing a diverse range of crucial neurons, thereby preserving consistency across local models. Third, we further develop a *dynamic aggregation module* to adaptively aggregate submodels, ensuring the equitable treatment of low-frequency neurons and consequently enhancing overall model accuracy. Extensive experiments conducted on three public benchmarks demonstrate that FedSAC outperforms all baseline methods in both fairness and model accuracy. We see this work as a significant step towards incentivizing broader client participation in federated learning. The source code is available at https://github.com/wangzihuixmu/FedSAC.

CCS CONCEPTS
• **Security and privacy**; • **Information systems → Information systems applications**;

KEYWORDS
federated Learning, collaborative fairness, privacy

[*]Both authors contributed equally to this research
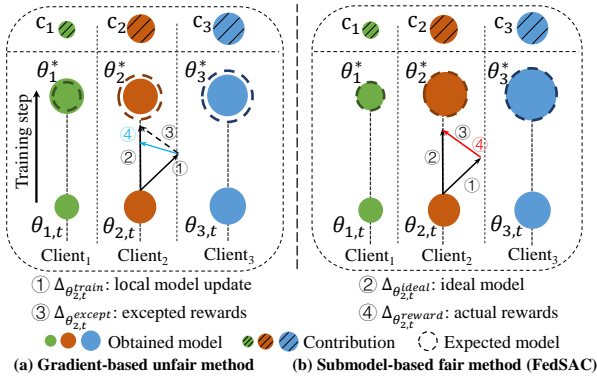[†]Corresponding Author

1 INTRODUCTION
Federated Learning (FL) empowers multiple data owners to collectively train a global model while preserving the privacy of their individual training data [23, 45, 46]. Early FL frameworks [3, 16, 37] usually distributed the same model to all clients without considering their distinct contributions to the model performance, resulting in unfairness to high-contributing clients. *Collaborative fairness* (CF) [22] stands as an essential element in federated learning to motivate client engagement by ensuring impartial reward distribution tied directly to individual contributions.

(a) Gradient-based unfair method   (b) Submodel-based fair method (FedSAC)

Figure 1: Problem illustration of collaborative fairness in FL. (a) Conventional gradients-based methods will result in poor fairness and model accuracy. For example, it is unfair that obtained models of $Client_2$ and $Client_1$ are equivalent neglecting the inferior contribution (c_1) of $Client_1$. Plus, the inconsistency in local models results in that obtained models of $Client_2$ and $Client_3$ are worse than expected (θ_i^*). (b) Our proposed FedSAC allocates sufficient submodels to each client by ensuring a comprehensive balance between fairness and model accuracy. For example, FedSAC ensures that obtained models of all clients (i.e., $Client_1$, $Client_2$ and $Client_3$) are in accordance with their contributions respectively. In addition, FedSAC guarantees the alignment of ③ and ④ of $Client_2$ during the training process, thereby enabling all three clients to obtain their expected models (θ_i^*).

More recently, several gradient-based methods were proposed to enhance CF [22, 42–44] (i.e., rewarding clients with corresponding model quality according to their contributions) in FL. They distribute a larger quantity of gradients to higher-contributing clients than the lower ones as rewards and quantify the degree of fairness by Pearson Correlation Coefficient ρ. However, for achieving CF, existing gradient-based methods have **two major limitations. On one hand**, the conventional definition of CF doesn't adequately distinguish in reward distribution among clients, resulting in a persistent unfairness for high-contributing clients. In Figure 1 (a), suppose the contributions of three clients are $c = [1, 9, 11]$, and their rewards are $\theta^* = [99, 99.2, 99.3]$ corresponding. Through the definition of CF by CGSV [44], the fairness is calculated as γ=98.97, but there exists an underlying unfairness towards $Client_2$ and $Client_3$ because $Client_1$ with an inferior contribution is over-rewarded. **On the other hand**, conventional gradient-based methods [22, 42–44] are ineffective because the inconsistency of local models updated by variable gradients might lead to significant degradation of overall model performance. In Figure 1 (a), the local models of clients in round t ($\theta_{1,t}$, $\theta_{2,t}$, and $\theta_{3,t}$) exhibit notable differences (i.e., the larger the circle, the higher the accuracy). Consequently, the gradients uploaded by individual clients may not be the optimal for others, creating a misalignment between obtained rewards $\Delta_{\theta_{i,t}^{reward}}$ (i.e., the rewards ultimately obtained by the clients) and expected rewards $\Delta_{\theta_{i,t}^{exceppt}}$ (i.e., the rewards that the clients ultimately expected) for each client.

To address the aforementioned challenges, we propose a novel _Federated learning framework with dynamic Submodel Allocation for bounded Collaborative fairness_ (**FedSAC**), supported by a theory of convergence while achieving competitive model accuracy. **First**, our approach introduces the concept of "_bounded collaborative fairness_ (BCF) (refer to _Definition_ 1)", which ensures fairness by integrating a differentiated range of rewards allocated to each client. **Second**, _the submodel allocation module_ with a theoretical fairness guarantee, is designed to assign relevant submodels (i.e., results of the aggregated model dropout) to individual clients based on their contributions. Specifically, these submodels encompass a diverse array of essential neurons for effective training. **Third**, _the dynamic aggregation module_ is implemented as a weight realignment mechanism by treating low-frequency neurons equally, which further improves the overall performance of the global model. Extensive experiments on three public benchmarks show that the proposed FedSAC outperforms all baseline methods in terms of collaborative fairness and model accuracy.

The contributions of this work are summarized:

- We propose FedSAC, a novel federated learning framework with a convergence guarantee, introducing a new concept of _bounded collaborative fairness (BCF)_. To the best of our knowledge, this is the first approach that allocates submodels equitably for collaborative fairness in FL.
- We implement the concept of BCF through two modules. First, _submodel allocation module_ prioritizes high-contributing clients by rewarding them with high-performance submodels under a theoretical guarantee. Second, _dynamic aggregation module_ merges submodels by paying equitable attention to low-frequency neurons to be aggregated.
- We conduct extensive experiments on three benchmarks with various settings and demonstrate that FedSAC ourperforms all baselines in both fairness and model accuracy.

2 RELATED WORKS

Recent research has shown that distributing different rewards based on clients' contributions can significantly impact the FL systems [21, 30, 47]. The incentive mechanisms can motivate clients to contribute high-quality data and promote collaboration [11, 31, 44]. We outline three types of rewards that can be adopted to achieve CF in FL.

Money-based reward. Several studies focus on the mechanism that rewards clients monetary based on their contributions. [55] proposes a reputation-based and reverse auction theory mechanism to reward clients with a limited budget. [48] shows a scheme that dynamically allocates budgets to clients in a context-aware manner by jointly maximizing the collective utility. While monetary rewards can be a natural and effective way to incentivize clients in FL, there exist challenges in maintaining a balance between the value of model quality and money [1, 53].

Data-based reward. Early studies have explored the fairness of rewarding different data sizes based on their contributions. [33] evaluates clients' contributions by aggregating the training data, and reward them with the corresponding models. [38] trains a generative model through the local data of all clients and provides more synthetic data to those datasets closely aligned with the real data distribution. However, most of existing data-based reward

methods rely on the centralized aggregation of all the data during training, making them difficult to be applied in the FL scenarios.

Gradient-based reward. Recent collaborative fairness (CF) works aim to reward high-contributing clients with optimal models. CFFL [22] allocates different gradient numbers based on local accuracy in the validation set and data sizes. CGSV [44] rewards more gradients to clients whose local gradients are more similar to the global gradients. FedAVE [42] assigns more gradients to clients whose data distribution information is more similar to the ideal dataset. However, existing rewards systems lack sufficient differentiation, resulting in an ongoing unfairness for high-contributing clients, which might degrade the fairness and model accuracy. Different from these methods, we propose a novel framework to achieve BCF (refer to $Definition$ 1) by allocating a differential range of rewards to clients.

3 PRELIMINARY

FL system consists of a server and multiple clients, aiming to minimize the weighted average of all clients' local objectives by optimizing a global model [10, 15, 23, 54]. First, the server broadcasts a model to the clients at random. Second, after training several rounds locally, the server aggregates these different trained models into a new global model. Finally, the aggregated model will be sent to the clients for further local training. The aforementioned process is repeated multiple times until the global model converges [5, 24, 41]. In this setup, the goal of FL framework is defined as:

$$\min_{\theta} F(\theta) := \sum_{i=1}^{N} p_i F_i(\theta), \tag{1}$$

where θ denotes the global model, N represents the number of clients, and $p_i = \frac{n_i}{n}$, $n = \sum_{k=1}^{N} n_k$. $F_i(\theta)$ is the loss on client i using model parameters θ, i.e., $F_i(\theta) = \frac{1}{n_i} \sum_{\xi_i \sim D_i} l(\theta, \xi_i)$, where D_i represents the local dataset of client i, and n_i denotes the data size of D_i. To achieve this goal as effectively as possible, FedAvg [23] samples a subset S_t of i clients uniformly, $0 < i \leq N$, to train the global model and aggregate the locally trained models by utilizing the data size ratio p_i as the weight of client i. Although FedAvg is proven to be effective in minimizing the objective successfully, it may be unfair to high-quality clients since the system distributes the same rewards to all clients regardless of their contributions [22, 44].

3.1 Problem Formulation

The standard FL framework allocates the same model to all clients regardless of their contributions [9, 23, 29, 36], dampening the motivation of high-quality clients to join FL [33, 40]. Collaborative fairness in FL aims to reward high-contributing clients with high-quality models. The existing works [22, 42, 44] assess the fairness with the Pearson Correlation Coefficient, $\rho(c; \theta^*)$, where c and θ^* represent the contributions and rewards of clients, respectively. However, the definition simply considers the relationship between the contributions and rewards of clients, which may lead to insufficient incentives for high-contributing clients. For example in Figure 1 (a), suppose the contributions of $Client_1$, $Client_2$, and $Client_3$ are $c_i = [1, 9, 11]$ and their rewards are $\theta_i^* = [99, 99.2, 99.3]$ correspondingly. Through the definition of CF, the fairness is calculated as $\gamma=98.97$, but there exists an underlying unfairness towards

$Client_2$ and $Client_3$ because $Client_1$ with an inferior contribution is over-rewarded.

To address this issue, we propose $Bounded\ Collaborative\ Fairness$ (BCF) to tackle the issue of insufficient incentives for the high-contributing client. BCF could ensure $c_1 < \theta_1^* < \frac{(c_1 + \theta_3^*)}{2}$ and then quantitative fairness with $\rho(c; \theta^*)$. The rationale behind the formula in $Definition$ 1 aims to amplify significant distinctions in rewards. The formula's left side ensures that clients' rewards exceed their contributions, while the right side prevents excessive rewards for clients with low contributions.

DEFINITION 1 (BOUNDED COLLABORATIVE FAIRNESS). *The contributions (c) and the rewards (θ^*) of clients are calculated by the performance of their standalone models (train without collaboration) and the final models obtained after collaboration, respectively. Based on client's obtained rewards $c_i < \theta_i^* < \frac{(c_i + max(\theta^*))}{2}$, the quantitative fairness can be computed by $\gamma := 100 \times \rho(c, \theta^*)$ where $\rho()$ is the Pearson Correlation Coefficient. The larger γ, the better the fairness of the framework.*

4 THE PROPOSED FEDSAC

In this section, we will introduce the details of proposed FedSAC, a method that ensures both BCF and consistency in local models for each client. The architecture of FedSAC is shown in Figure 2. The pseudo codes for FedSAC are provided in Algorithm 1. First, we introduce the submodel allocation module in Section 4.1. Second, we present the dynamic aggregation module in Section 4.2. Third, we proposed the fairness guarantee theory in Section 4.3 to prove that this submodel allocation strategy can achieve collaborative fairness. Fourth, we conducted a convergence analysis on FedSAC and demonstrated its convergence in Section 4.5. In addition, we analyzed the time complexity and communication costs of FedSAC in Section 4.5. Finally, we discussed limitations in Section 4.6.

4.1 Submodel Allocation Module

A naive approach achieving bounded collaborative fairness involves allocating distinct submodels to each client based on their respective contributions [7]. Unlike previous works such as [8, 32], there are two primary motivations behind achieving BCF through submodel allocation. First, submodels with appropriate pruning may not match the performance of the global model, enabling clients to receive diverse submodels according to their contributions. Second, despite being subsets of the global model, these submodels exhibit strong mutual validity, meaning that the submodels uploaded by one client are effective for others, facilitating the training of the global model. However, it is still challenging to achieve BCF through submodel-based methods. For one thing, it is crucial to ensure that the majority of neurons are adequately trained to guarantee the optimal performance of the global model. For another, the performance of allocated submodels should align with their respective contributions.

To address the aforementioned two challenges, we design a two-step approach for submodel allocation module. First, we evaluate the importance of each neuron within the model to determine their contributions respectively (*neuron importance evaluation in Section 4.1.1*). Second, we construct submodels for each client with
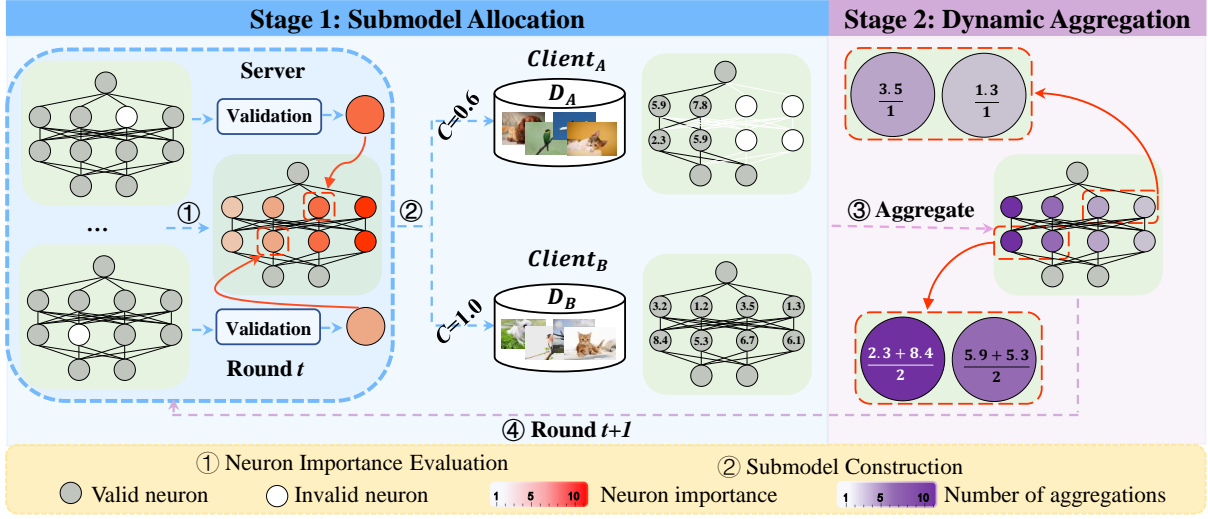
Figure 2: The overall framework of FedSAC that achieves bounded collaborative fairness by maintaining consistency across local models. FedSAC consists of two module: 1) *submodel allocation module* conducts neuron importance evaluation and submodel construction to reward high-contributing clients with high-performance submodels, thus ensuring consistency in local models; 2) *dynamic aggregation module* treats those low-frequency neurons equally, which further refines the performance of the global model.

varying performances based on their contributions, ensuring a diverse array of important neurons is included within each submodel (*submodel construction in Section 4.1.2*).

4.1.1 Neuron Importance Evaluation.
Each neuron within the model holds a unique contribution [20, 25, 26, 50]. Our intuition is that the constructed submodels can yield varied performances. Inspired by Taylor-FO [25], we calculate the neuron importance in the model by measuring the change in loss upon their removal. For instance, a greater increase in loss indicates a more significant contribution by the removed neuron to the model. In Figure 2, neurons depicted in a redder shade represent a higher contribution to the model. Essentially, the training objective is to minimize the cross-entropy loss L_{ce}:

$$\min_{\theta} \sum_{i=1}^{N} L_{ce}(x_i, \theta), \qquad (2)$$

where x_i denotes the sample, θ represents the model, and $L_{ce}(x_i, \theta)$ is the loss function of the classification tasks.

The neurons in the model have a multitude of model parameters, each of which contributes to the overall performance of the model. The importance of a neuron I_{n_i} of the model can be calculated through the loss increased by removing it:

$$I_{n_i} = L_{ce}(V, \theta|\theta_{n_i}^{zero} = 0) - L_{ce}(V, \theta), \qquad (3)$$

where V denotes the validation set, which is constructed by evenly selecting 10% of the data from the original training samples [22], $\theta_{n_i}^{zero}$ represents that the parameters of the i-th neuron in the model are all set to 0.

To simplify the construction of submodels, we normalize the sum of neuron scores, which represent their importance in the

model presented by percentage:

$$I_{n_i} = \frac{I_{n_i}}{\sum_{n_i \in S} I_{n_i}} * 100, \qquad (4)$$

where S represents all neurons of the model. To reduce the training time of the framework, we measure the importance of neurons in the model by Eq. (3) and Eq. (4) every 10 epochs. All these operations allow us to efficiently assess the importance of each neuron while limiting excessive computation demands.

4.1.2 Submodel Construction.
In pursuit of fairness, we employ a dynamic allocation system for submodels with varying performances, leveraging clients' reputations derived from their contributions. Our approach incorporates a pruning mechanism tailored to clients' contributions, simplifying the extraction of submodels with different performance levels from the global model. In this scheme, the client i's reputation r_i is expressed as:

$$r_i = e^{c_i * \beta}, \qquad (5)$$

$$r_i = \frac{r_i}{max(r)} * 100, \qquad (6)$$

where c_i represents client i's contribution, β is a hyper-parameter. The reputations of clients are directly proportional to their contributions. The design rationale for Eq. (5) and Eq. (6) is to calculate the clients' reputations (r), which facilitates the allocation of their submodels fairly. More specifically, our pruning method begins with the most important neuron, ensuring that submodels for low-contribution clients possess a higher parameter count, which is beneficial for training the global model. These actions serve a dual purpose: promoting collaborative fairness while maximizing the overall performance of the global model. Submodel θ_i is constructed

by neurons with different reputations:

$$\theta_i = quantity(r_i, \sum\nolimits_{n_i \in S} I_{n_i}), \qquad (7)$$

where $quantity(r_i, \sum_{n_i \in S} I_{n_i})$ represents the submodel θ_i when $r_i = \sum_{n_i \in S} I_{n_i}$, r_i denotes client i's reputation, $\sum_{n_i \in S} I_{n_i}$ denotes the set importance for different neurons. S represents the positions of all neurons in the model, arranged in ascending order from the least to the most important. This design choice aims to maximize the inclusion of neurons in each submodel, thereby enhancing the performance of the corresponding local model updates. Subsequently, this quantity will be utilized in Eq. (8) to generate the submodel's mask.

4.2 Dynamic Aggregation Module

Next, the server aggregates the locally trained submodels and allocates distinct submodels to clients in the subsequent round. Recent submodel-based methods [8, 32] have aimed to allocate varied submodels containing numerous neurons to clients. However, these approaches might pose a potential risk of compromising overall model performance when integrating low-frequency neurons into the global model. Consequently, employing a direct aggregation method such as FedAvg [23] for all neurons becomes inequitable.

Instead of simply averaging the uploaded submodels, our objective is to optimize the utilization of all neurons within the model. With the sizes of submodels varying across clients, it becomes essential to treat the contribution of each neuron individually during aggregation. To ensure fair treatment of low-frequency neurons, we integrate the frequency of submodel parameter aggregations as weights to dynamically aggregating local models:

$$mask_i^t = mask(\theta_i^t, \theta_g^{t-1}), \qquad (8)$$

$$\theta_g^{t+1} = \frac{\sum_{i \in N} \theta_i^t}{\sum_{i \in N} mask_i^t}, \qquad (9)$$

where $mask(\theta_i, \theta_g)$ denotes submodel θ_i's mask (same shape as the submodel θ_i), θ_g denotes the aggregated model, N denotes the total number of clients. It sets the components of both θ_i and θ_g at the same position to 1 and 0 for the rest. The role of the mask function $mask(\theta_i, \theta_g)$ is to calculate the frequency of each model parameter θ_i selected by the global model θ_g in round t. Later, the mask function $mask(\theta_i, \theta_g)$ will be utilized in Eq. (9) to treat those low-frequency parameters equally by suppressing the weight of high-frequency parameters in the aggregation, which makes each parameter play a fair role during the aggregation phase. For example, the more the frequency of a selected parameter in round t, the smaller the weight of the parameter to be aggregated in round $t + 1$.

4.3 Fairness Guarantee

In Section 4.1, we delved into the fundamental concept underpinning our definition of fairness. This concept centers on rewarding high-contributing clients with high-performance submodels, where a submodel's improved performance correlates with the number of neurons it contains. Consequently, this approach leads to a training loss (i.e., model accuracy) that more closely aligns with the aggregated model. It's important to note that the submodel θ_i acquired

by client i is determined based on its reputation r_i across the entire training process up to iteration t.

Our primary result ensures a notion of fairness under specific conditions concerning the loss function F. If client i holds a higher reputation than client j ($r_i \geq r_j$), and the submodel θ_i^t obtained by client i encompasses the submodel θ_j^t obtained by client j ($\theta_j^t \in \theta_i^t \in \theta_g^t$). Then, the submodel θ_i^t obtained by client i will exhibit closer alignment with the aggregated model θ_g^t in round t. Letting $\delta_i^t := ||\theta_g^t - \theta_i^t||$, it's evident that $\delta_i^t \leq \delta_j^t$. Consequently, the submodel θ_i^t obtained by client i will yield a smaller loss function $F(\theta)$ compared to client j in round t.

ASSUMPTION 1 (L-SMOOTH F). *If F is L-smooth, then $\forall \theta_i, \theta_j \in \theta$,*

$$F(\theta_i) \leq F(\theta_j) + \nabla F(\theta_j)^T (\theta_i - \theta_j) + \frac{L}{2} ||\theta_i - \theta_j||^2. \qquad (10)$$

ASSUMPTION 2 (μ-STRONGLY CONVEX F). *If F is μ-strongly convex, then $\forall \theta_i, \theta_j \in \theta$,*

$$F(\theta_i) \geq F(\theta_j) + \nabla F(\theta_j)^T (\theta_i - \theta_j) + \frac{\mu}{2} ||\theta_i - \theta_j||^2. \qquad (11)$$

THEOREM 1 (**FAIRNESS IN TRAINING LOSS**). *Assume Assumptions 1 and 2 hold, FedSAC can guarantee collaborative fairness by rewarding high-contributing clients obtaining high-performance models. Formally speaking, let $\delta_i^t := ||\theta_g^t - \theta_i^t||$. Suppose that θ_t is close to a stationary point of F for $t \geq T \in Z^+$, and $F()$ is both L-smooth and μ-strongly convex with $L \leq \mu$. For all $i, j \in N$ in round t, if $r_i \geq r_j$, it follows that $\theta_j^t \in \theta_i^t \in \theta_g^t$, $\delta_i^t \leq \delta_j^t$, and therefore $F(\theta_i^t) \leq F(\theta_j^t)$.*

The proof process is as follows:
From L-smoothness (ASSUMPTION 1), we have

$$F(\theta_i^t) \leq \underbrace{F(\theta_N^t) + \nabla F(\theta_N^t)^T (\theta_i^t - \theta_N^t) + \frac{L}{2} \delta_{i,t}^2}_{R_L}. \qquad (12)$$

From μ-strongly convex (ASSUMPTION 2), we have

$$F(\theta_j^t) \geq \underbrace{F(\theta_N^t) + \nabla F(\theta_N^t)^T (\theta_j^t - \theta_N^t) + \frac{\mu}{2} \delta_{j,t}^2}_{R_\mu}. \qquad (13)$$

In order to prove $F(\theta_i^t) \leq F(\theta_j^t)$, it suffices to prove $R_L \leq R_\mu$ or equivalently $R_L - R_\mu \leq 0$.

$$R_L - R_\mu = \underbrace{\nabla F(\theta_N^t)^T (\theta_i^t - \theta_j^t)}_{R_1} + \underbrace{\frac{1}{2}(L\delta_{i,t}^2 - \mu\delta_{j,t}^2)}_{R_2}. \qquad (14)$$

With $L \leq \mu$ and $\delta_{i,t} \leq \delta_{j,t}$, we have

$$R_2 = \frac{1}{2}(L\delta_{i,t}^2 - \mu\delta_{j,t}^2) \leq \frac{L}{2}(\delta_{i,t}^2 - \delta_{j,t}^2) \leq 0. \qquad (15)$$

We define θ_N^t being close to a stationary point of F by establishing an upper limit on the gradient:

$$||\nabla F(\theta_N^t)|| \leq \frac{L|\delta_{i,t}^2 - \delta_{j,t}^2|}{2||\theta_i^t - \theta_j^t||}. \qquad (16)$$

Algorithm 1 FedSAC

Input: The global model θ_g, the local submodel θ_i, neurons i's importance I_{n_i}, the number of local update steps E, learning rate η_t, number of clients N, hyper-parameter β, client's contribution c

1: Initialize the global model parameters θ_g^0
2: **for** round $t = 0, 1, ..., T - 1$ **do**
3:     Compute $I_{n_i}^t$ (3) and (4) of θ_g^t
4:     Calculate the reputation r_i of client k: $r_i = \frac{e^{c_i * \beta}}{max(e^{c * \beta})} * 100$
5:     Calculate Submodels θ_i^t of clients i in round t: $\theta_i^t = quantity(r_i, \sum_{n_i \in S} I_{n_i}^t)$
6:     **for** each client $i \in N$ **do**
7:         **for** each iteration $j = 0, 1, ..., E - 1$ **do**
8:             $\theta_{i,j+1}^t \leftarrow \theta_{i,j}^t - \eta_t \nabla F_i(\theta_{i,j}^t)$
9:         **end for**
10:     **end for**
11:     Submodel i's mask in round t: $mask_i^t = mask(\theta_i^t, \theta_g^{t-1})$
12:     The server aggregates the received submodels: $\theta_g^{t+1} = \frac{\sum_{i \in N} \theta_i^t}{\sum_{i \in N} mask_i^t}$
13: **end for**

We have the following:

$$\begin{aligned}
|R_1| &\triangleq |\nabla F(\theta_N^t)^T (\theta_i^t - \theta_j^t)| \\
&\leqslant ||\nabla F(\theta_N^t)|| \times ||(\theta_i^t - \theta_j^t)|| \\
&\leqslant \frac{L|\delta_{i,t}^2 - \delta_{j,t}^2|}{2} \\
&\leqslant |R_2|,
\end{aligned} \tag{17}$$

where the first inequality is derived from the Cauchy-Schwarz, the second inequality is by substituting the aforementioned upper limit (refer to Eq. (16)), and the last inequality (line 1209) emerges from taking the absolute values of two negative values (refer to Eq. (15)).

Finally, given that $|R_1| \leqslant |R_2|$ and $R_2 \leqslant 0$, we derive $R_1 + R_2 \leqslant 0$. Therefore, it follows that $R_L - R_\mu \triangleq R_1 + R_2 \leqslant 0$, which subsequently implies $F(\theta_i^t) \leqslant F(\theta_j^t)$.

4.4 Convergence Analysis

In this section, we delve into the convergence analysis of the proposed FedSAC. To guarantee convergence to the global optimum, we make the assumption that each neuron in the aggregated model is equally allocated over T rounds. Consequently, the anticipated weight of the allocated submodel θ_i contracts towards the aggregate model θ_g, i.e., $\theta_i^{t+1} = p_i \theta_g^t$. Here, p_i ($0 \leqslant p_i \leqslant 1$) denotes the long-term expectation of the size ratio between the submodel i and the aggregate model obtained in multiple iterations. At this stage, Eq. (9) can be expressed as the aggregation of each submodel θ_i divided by its respective p_i, i.e., $\theta_g^{t+1} = \sum_{i=1}^N \frac{\theta_i^{t+1}}{p_i}$. We present THEOREM 2 below, which demonstrates that FedSAC enables the convergence of the aggregation model. Assumptions 3 and 4 are derived from the works [34, 35, 49, 56].

ASSUMPTION 3. *Let ξ_i^t denote samples uniformly from the local data of the i-th device at random. It is asserted that the variance of*

stochastic gradients within each device remains constrained:

$$E\|\nabla F_i(\theta_i^t, \xi_i^t) - \nabla F_i(\theta_i^t)\| \leqslant \sigma_i^2, \tag{18}$$

ASSUMPTION 4. *The expected squared norm of stochastic gradients is uniformly constrained:*

$$E\|\nabla F_i(\theta_i^t, \xi_i^t)\| \leqslant G^2, \tag{19}$$

where $i \in \{1, 2, ..., N\}$ and $t \in \{1, 2, ..., T - 1\}$.

ASSUMPTION 5. *Each neuron in the aggregation model is assigned the same number of times after T rounds. Therefore, the expected weight of the allocated submodel θ_i is a contraction of the aggregate model θ_g, i.e., $\theta_i^{t+1} = p_i \theta_g^t$. Here, p_i ($0 \leqslant p_i \leqslant 1$) denotes the long-term expectation of the size ratio between the submodel i and the aggregate model obtained in multiple iterations.*

THEOREM 2 (**ASYMPTOTIC CONVERGENCE**). *Given that Assumptions 1 to 5 hold and L, μ, σ_i, G, p be defined therein. Choose $\kappa = \frac{2}{\mu}, \gamma = max\{8\frac{L}{\mu}, E\} - 1$ and the learning rate $\eta_t = \frac{2}{\mu(\gamma+t)}$. Then FedSAC satisfies $E[F(\bar{\theta}_T)] - F^* \leqslant [\frac{L}{\gamma+T}(\frac{2B}{\mu^2} + \frac{\gamma+1}{2}\triangle_1)]$.*

The proof is shown in Appendix A.

4.5 Complexity and Communication Cost Analysis

We further analyze the time complexity and communication costs of FedSAC as follows.

For the time complexity, the primary computational demand in FedSAC stems from evaluating neuron importance, as defined in Eq. (3) and Eq. (4)). The time complexity for this evaluation is $O(M)$, where M denotes the total number of neurons across the hidden layers of the global model.

For the communication cost, FedSAC mitigates the introduction of additional communication overhead by conducting neuron importance evaluation solely on the server. This approach effectively eliminates the necessity for client-server communication, thereby enhancing overall efficiency. Moreover, it displays a communication complexity of O(d*m) per round, as outlined in [8], where m$\leqslant$1 denotes the average ratio of submodel parameters to the global model. As a result, FedSAC showcases lower communication complexity compared to all baseline methods in cross-silo FL scenarios [28]. More details about the communication cost experiments are put in Appendix E.

4.6 Limitations

In Table 1 and Table 2, we conduct extensive experiments on various datasets and observe that FedSAC could exhibit a distinct advantage over all baseline methods in terms of both fairness and model accuracy. Nevertheless, the sufficient evaluation of neuron importance (Section 4.1.1) within the submodel allocation module imposes an additional computational burden. This problem may be amplified for large models. Despite this challenge, we hold a strong conviction that the substantial enhancements in both fairness and accuracy achieved through by FedSAC clearly affirm its superiority over baseline methods.

5 EXPERIMENTS

In this section, we conduct comprehensive experiments to answer the following research questions:

RQ1. How does the fairness of our FedSAC compare to various state-of-the-art methods?

RQ2. How does the predictive model performance achieved by our proposed method compare with the state-of-the-art methods on different datasets?

RQ3. How do different components (i.e., submodel allocation module and dynamic aggregation module) affect the results?

5.1 Experimental Settings

Datasets and Models. We evaluate the performance of FedSAC on three commonly used public datasets in collaborative fairness, including Fashion MNIST [13], CIFAR10 [12], and SVHN [27]. Following [14, 19, 39], we employ a feedforward neural network with two hidden layers for all datasets.

Data splits. We construct five heterogeneous scenarios by varying the size and the class numbers of the dataset. For **imbalanced dataset sizes** (POW) [22, 44], we randomly divide the total dataset into various data sizes for each client by using a power law. For CIFAR10, we partition the data set of size 20000 among 10 clients. The clients with more extensive data sizes are expected to achieve better prediction performance. For **imbalanced class numbers** (CLA) [22, 44], we change the number of classes and keep them have the same amount of data. For CIFAR10 with 5 clients, clients 1, 2, 3, 4, 5 own local training data with 1, 3, 5, 7, 10 classes respectively. For **imbalanced data size and class numbers** (DIR), we provide clients with various data sizes and classes by the Dirichlet distribution function [2, 6, 51, 52]. Specifically, we sample $p_i^l \sim DIR(\alpha)$ and assign a p_i^l percentage of the data of class l to client i, where $DIR(\alpha)$ is the Dirichlet distribution with a parameter α. More details on the varying numbers of clients used in the experiment are put in Appendix C.

Baselines. We compare FedSAC with the following methods: (1) FedAvg [23] distributes the same model to all clients in each FL iteration. In this case, Pearson Correlation Coefficient $\rho()$ in Section 3.1 is uncomputable. To address this and create a personalized model for each client, we follow CFFL [22] and CGSV [44], which enables clients to train for an additional epoch at the end of FL algorithm. (2) q-FFL[17] enables the reweighting of loss across different clients by the q-parameterized weights, thus reducing the variance in the accuracy distribution and achieving a fairer distribution of accuracy. (3) CFFL [22] allocates more gradients to higher reputation client, and the reputation is calculated by the local accuracy and data sizes (or label diversity). (4) CGSV [44] assigns more gradients to clients whose local model gradients is more similar to the global gradients. (5) FedAVE [42] assigns more gradients to clients whose data distribution information is more similar to the ideal dataset. (6) Standalone [22] trains local models alone without collaboration. Particularly, to evaluate more fairly, we make all algorithms distribute rewards based on client contributions rather than the calculated reputations.

Hyper-Parameters. We tune all hyper-parameters in datasets by using grid search with FedAvg [23] and subsequently apply the optimal parameters obtained from the validation dataset. The batch size is $B = 64$ for SVHN and $B = 32$ for both FashionMNIST and CIFAR10. The optimal parameters for SVHN, Cifar10, and Fashion-MNIST of six scenarios are $E = \{15, 20\}$, $\eta = \{0.05, 0.1\}$, $\beta = \{3, 5, 10\}$; $E = \{15, 20\}$, $\eta = \{0.03, 0.05\}$, $\beta = \{3, 5, 10\}$, and $E = 20$, $\eta = \{0.03, 0.05\}$, $\beta = \{1, 10, 20, 25\}$, respectively. For comparison, we select the best fairness achieved by each method. The effects of hyper-parameter β on FedSAC are detailed in Appendix D (the smaller β is, the higher the accuracy achieved by FedSAC). More details about the hyper-parameters are put in Appendix F.

Implementation. All experiments are run on a 64 GB-RAM Ubuntu 18.04.6 server with Intel(R) Xeon(R) CPU E5-2630 v4 @ 2.20GHz and 1 NVidia(R) 2080Ti GPUs.

5.2 Experimental Results

Fairness (RQ1). To evaluate the FedSAC fairness, we compared it with a few baselines on three datasets. Table 1 shows the fairness metrics according to $Definition$ 1. Standalone [22] trains local models alone without collaboration, which represents the clients' contributions. Table 1 indicates that our proposed dynamic submodel allocation mechanism achieves a fairness score above 95.73% on all datasets, while the FedAvg performs poorly with the lowest fairness score of -19.83%. On three datasets, the fairness of algorithms (i.e., CFFL, CGSV, and FedAVE) exceeds 73.65% for the POW and CLA scene. In these scenarios, the client's contribution varies greatly and is mainly related to the amount of data or diverse labels of data. For the DIR scene, the data distribution among clients is significantly uneven, resulting in a high degree of non-iid settings and clients with relatively similar contributions. Consequently, CFFL, CGSV, and FedAVE show low fairness, as the rewards received by clients tend to be indistinguishable. In particular, in DIR (1.0) of CIFAR10, our method outperforms CFFL, CGSV and FedAVE by 69.24%, 35.85%, and 40.71%, respectively.

Table 1 demonstrates that the proposed FedSAC outperforms the state-of-the-art approaches in fairness, and validated the effectiveness of our method: high-contributing clients obtain high-performance models. Figure 3 shows the comparison results of overall performance to achieve bounded collaborative fairness with state-of-the-art methods in CIFAR10 (left), SVHN (middle), and Fashion MNIST (right). **Obviously**, FedSAC outperforms all baselines in terms of fairness.

Predictive performance (RQ2). To effectively assess the predictive performance of algorithms, we present our highest test accuracies in comparison with all baseline methods in Table 2. These results demonstrate the ability of the algorithms to reward high-contributing clients with high-performance. First, comparing the accuracy of FedSAC with Standalone (i.e., contribution) reveals that FedSAC significantly outperforms Standalone. Second, among the POW scene, FedSAC achieves the highest performance in CIFAR10, SVHN and Fashion MNIST with accuracies of 48.61%, 74.84%, and 87.88%, respectively. Third, for the CLA scene on three datasets, the highest accuracy is obtained by FedSAC, surpassing FedAvg by at least 0.19%. In addition, in the extremely non-iid setting (e.g., DIR (1.0) of SVHN), our method outperforms CFFL, CGSV, and FedAVE by 2.17%, 5.06%, and 4.62%, respectively. Finally, for the DIR(2.0), and DIR (3.0) scenes, FedSAC achieves comparable performance to baseline methods in terms of accuracy. Specially, the notably

Dataset	CIFAR10					SVHN					Fashion MNIST				
No. Clients	10					10					10				
Scene	POW	CLA	DIR(1.0)	DIR(2.0)	DIR(3.0)	POW	CLA	DIR(1.0)	DIR(2.0)	DIR(3.0)	POW	CLA	DIR(1.0)	DIR(2.0)	DIR(3.0)
FedAvg[23]	14.55±25.4	84.68±3.5	10.77±6.3	15.06±21.2	68.93±8.1	-19.24±26.7	71.42±13.0	58.56±4.0	50.95±13.5	31.43±19.0	-19.83±24.1	78.44±4.3	17.94±7.3	60.36±16.3	76.09±29.5
q-FFL[17]	34.91±15.1	98.74±0.5	90.89±1.5	89.01±1.5	73.02±2.5	67.97±11.6	93.05±1.0	82.11±1.8	86.87±4.7	89.41±2.1	42.69±9.8	88.99±0.3	73.19±6.0	83.93±0.7	79.99±7.3
CFFL[22]	93.55±1.3	89.99±0.8	29.90±3.6	81.82±0.8	59.86±3.0	96.38±1.5	95.63±0.4	46.91±5.7	38.58±2.5	31.35±8.3	90.94±0.5	86.50±0.7	85.90±0.9	85.09±1.3	71.10±0.3
CGSV[44]	90.78±0.6	91.04±0.8	63.29±3.6	84.59±1.6	84.75±0.2	90.99±0.4	87.22±0.6	72.09±0.2	72.19±0.2	76.31±0.2	95.34±0.3	73.65±2.6	82.91±3.6	82.91±3.6	84.95±1.3
FedAVE[42]	85.50±0.8	92.80±1.2	58.43±0.9	85.82±0.7	88.61±1.3	92.68±0.4	92.77±0.8	82.42±1.2	64.83±1.0	79.38±0.8	86.98±0.5	86.36±1.1	79.51±1.3	87.99±0.5	67.62±0.3
Ours	98.80±0.2	99.06±0.3	99.14±0.6	95.73±0.5	97.01±0.7	99.44±0.3	99.74±0.1	96.09±0.3	96.48±0.2	98.32±0.8	96.35±0.2	98.93±0.7	99.23±0.3	97.71±0.2	98.62±0.3

Table 1: Comparison results of fairness $\rho \in [-100, 100]$ with state-of-the-art methods on three datasets. The reported results are averaged over 5 runs with different random seeds. (A higher value indicates better fairness. The best average result is marked in bold. The second-best result is underlined. These notes are the same to others.)

Dataset	CIFAR10					SVHN					Fashion MNIST				
No. Clients	10					10					10				
Scene	POW	CLA	DIR(1.0)	DIR(2.0)	DIR(3.0)	POW	CLA	DIR(1.0)	DIR(2.0)	DIR(3.0)	POW	CLA	DIR(1.0)	DIR(2.0)	DIR(3.0)
Standalone	41.23±0.1	37.49±0.2	33.78±0.2	33.54±0.1	31.89±0.1	60.02±0.2	52.05±0.3	41.41±0.2	58.07±0.2	61.11±0.2	84.36±0.2	82.52±0.3	64.19±0.3	67.39±0.1	74.29±0.2
FedAvg[23]	48.36±0.2	42.64±0.5	48.84±0.1	49.49±0.5	49.72±0.5	74.16±0.2	68.25±0.1	77.75±0.2	78.17±0.1	81.35±0.1	87.64±0.2	85.42±0.0	87.32±0.1	87.27±0.2	88.25±0.2
q-FFL[17]	46.22±1.5	41.40±0.2	35.00±0.9	37.41±1.1	37.81±0.4	69.61±0.6	54.71±1.52	34.15±0.6	44.69±1.3	56.92±1.1	85.44±0.2	82.93±0.5	65.51±2.4	69.61±2.1	78.96±0.8
CFFL[22]	47.94±0.6	42.12±0.3	44.44±0.9	48.44±0.3	47.56±1.2	72.68±0.2	69.66±0.7	75.89±0.9	75.21±0.3	77.01±0.9	87.40±0.4	85.29±0.4	86.16±1.0	87.23±0.2	87.11±0.8
CGSV[44]	34.60±0.7	39.33±0.5	47.22±0.6	44.12±0.5	39.31±0.7	65.92±0.6	65.96±0.3	73.00±1.9	75.16±0.2	77.42±0.5	83.16±0.3	82.41±1.3	85.03±2.9	84.73±3.2	76.51±4.2
FedAVE[42]	46.51±0.2	35.18±1.5	46.60±0.6	39.20±1.2	40.60±1.8	70.80±0.5	65.10±0.6	73.44±0.6	73.43±0.3	75.48±0.7	86.18±0.6	79.86±1.2	76.66±0.8	80.90±1.3	67.74±0.8
Ours	48.61±0.2	44.16±0.2	49.06±0.6	50.01±0.2	49.85±0.3	74.84±0.2	70.51±0.8	78.06±0.1	78.55±0.1	81.95±0.4	87.88±0.2	85.61±0.3	87.85±0.1	87.54±0.4	88.38±0.4

Table 2: Comparison results of the maximum test accuracy (%) with state-of-the-art methods on three datasets. The reported results are averaged over 5 runs with different random seeds. (A higher value indicates better accuracy.)
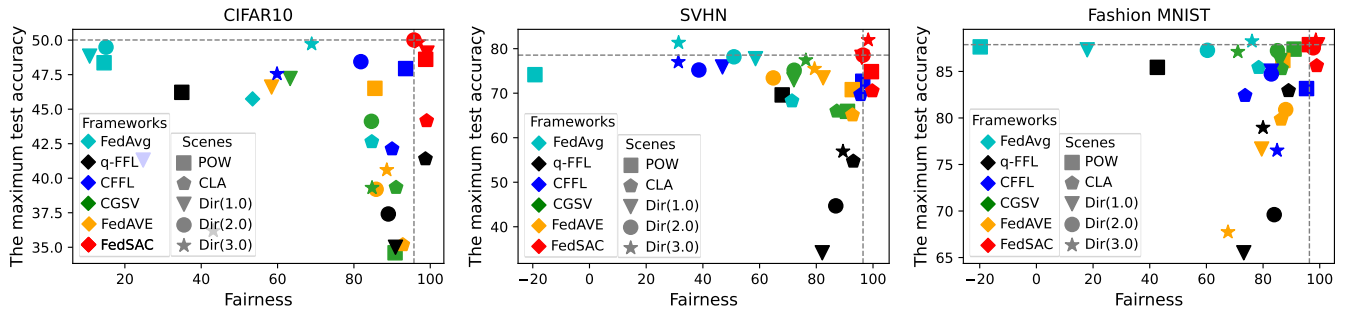
Figure 3: Comparison results of overall performance to achieve bounded collaborative fairness with state-of-the-art methods in CIFAR10 (left), SVHN (middle), and Fashion MNIST (right). (The closer the point is to the upper-right corner, the better the performance.)

Dataset	CIFAR10					SVHN					Fashion MNIST				
Scene	POW	CLA	DIR(1.0)	DIR(2.0)	DIR(3.0)	POW	CLA	DIR(1.0)	DIR(2.0)	DIR(3.0)	POW	CLA	DIR(1.0)	DIR(2.0)	DIR(3.0)
w/o allocation	79.99	96.47	10.93	43.81	-14.61	96.88	94.88	31.86	64.94	82.58	91.05	96.09	84.57	79.14	65.53
w/o aggregation	98.64	98.13	99.22	84.81	83.59	99.00	99.20	95.66	96.05	87.65	93.72	97.45	97.32	27.40	91.43
FedSAC	98.80	99.06	99.14	95.73	97.01	99.44	99.74	96.09	96.48	98.32	96.35	98.93	99.23	97.71	98.62

Table 3: Ablation studies on FedSAC for fairness $\rho \in [-100, 100]$ on three public benchmarks. A higher ρ denotes better fairness.

poor accuracy of q-FFL appears attributed to its mechanism that offers the same reward to all clients, without adapting these rewards based on individual client contributions. Figure 4 illustrates the distributions of contributions and allocated rewards under scenes (i.e.,

DIR (1.0)) in CIFAR10 comparing FedSAC against baseline methods. It demonstrates that FedSAC not only guarantees BCF but also enables clients to receive rewards that exceed their contributions (i.e., Standalone). **In short**, FedSAC outperforms all baselines in

Dataset	CIFAR10					SVHN					Fashion MNIST				
Scene	POW	CLA	DIR(1.0)	DIR(2.0)	DIR(3.0)	POW	CLA	DIR(1.0)	DIR(2.0)	DIR(3.0)	POW	CLA	DIR(1.0)	DIR(2.0)	DIR(3.0)
w/o allocation	47.68	42.46	40.02	47.00	44.09	69.98	66.95	66.25	68.14	74.02	87.48	85.00	70.99	73.13	83.43
w/o aggregation	48.03	43.62	48.23	49.46	48.61	73.81	68.91	77.35	72.89	80.46	87.74	85.95	86.90	85.72	86.74
FedSAC	**48.61**	**44.16**	**49.06**	**50.01**	**49.85**	**74.84**	**70.51**	**78.06**	**78.55**	**81.95**	**87.88**	**85.61**	**87.85**	**87.54**	**88.38**

Table 4: Ablation studies on FedSAC for the maximum test accuracy (%) on three public benchmarks.
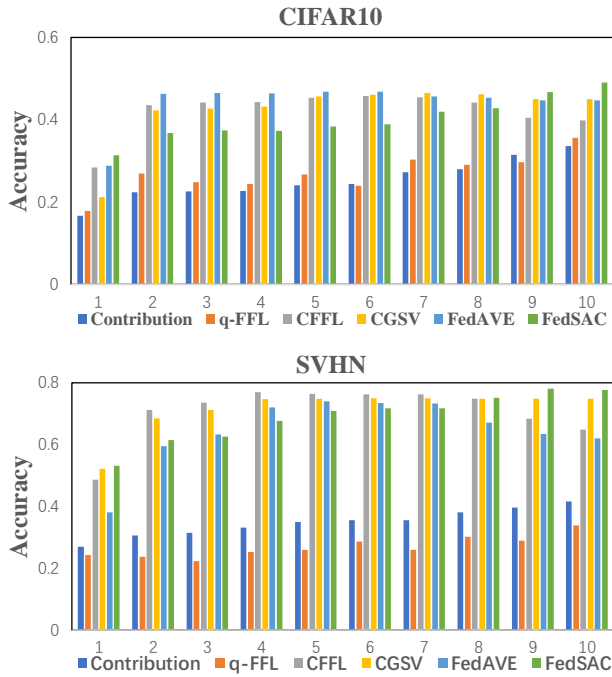

CIFAR10

SVHN

Figure 4: Comparison results of test accuracy using the data partition of DIR (1.0) with state-of-the-art methods in CIFAR10 (up) and SVHN (down). Results of other scenes are in Appendix D.

Figure 5: The test accuracy achieved by clients during training for CIFAR10 (left) and SVHN (right) in each round, under the setting of POW and CLA.

terms of accuracy. More results under different scenarios (i.e., CLA scene) on CIFAR10 and SVHN are presented in Appendix B.

Figure 5 illustrates the changes in clients' test accuracy as the number of communication rounds increases in the POW, and CLA data partition of CIFAR10 and SVHN. Owing to the varying data sizes and diversity of labels owned by clients in FL, their contributions to the system exhibit significant differences. As shown in Figure 5, our proposed FedSAC, underpinned by a theoretical guarantee, aims to reward high-contributing clients with high-performance submodels by maintaining consistency in local models. As a result, each client will converge to a different model and achieve varying levels of performance.

Ablation study (RQ3). To evaluate the effectiveness of two proposed modules in FedSAC, a series of ablation experiments are carried out on three public benchmarks with 10 clients, as shown in Table 3 and 4. The operation of eliminating neuron importance, denoted as *w/o allocation*, aims to treat all neurons equally and allocate submodels based on their contributions. *w/o aggregation*
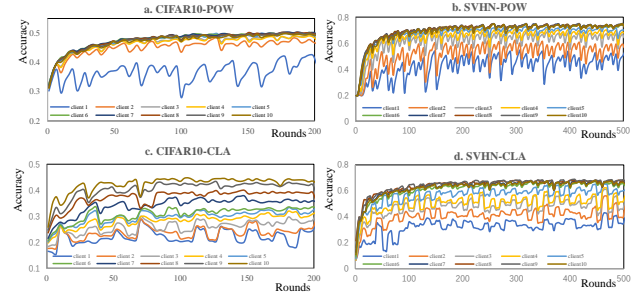
denotes removing the dynamic aggregation module, which uses the traditional FedAvg aggregation method to train. The effectiveness of the submodel allocation module is demonstrated in Table 3, indicating that this module can reward high-contribution clients to obtain high-performance models. In particular, our method has significantly improved the fairness measure by 88.21% on the DIR (1.0) scene of the CIFAR10 dataset. Table 4 shows the results of the proposed dynamic aggregation module, which implies that this module can effectively aggregate submodels with different sizes, thereby further improving the overall performance of the local models. **Thus**, the ablation study demonstrates that the two designed modules in FedSAC are crucial and significant in enhancing bounded collaborative fairness.

In summary, all experimental results show that both fairness and model accuracy are of significance for bounded collaborative fairness, and our FedSAC outperforms all baseline methods in both fairness and model accuracy.

6 CONCLUSION

In this work, we introduce a novel FL framework named FedSAC that allocates submodels based on their contributions, thereby ensuring bounded collaborative fairness and attaining superior local accuracy while maintaining the consistency in local models. Our method ensures that high-contributing clients can be rewarded with high-performance submodels, which in turn enhances the overall model accuracy. The experiments on three datasets show that FedSAC exhibits a distinct advantage over baseline methods in terms of fairness and accuracy. In the future, we aim to investigate the implementation of FedSAC on large models.

REFERENCES
[1] Anish Agarwal et al. 2019. A marketplace for data: An algorithmic solution. In *Proceedings of the 2019 ACM Conference on Economics and Computation*. 701–726.

[2] Chen Chen et al. 2022. Gear: a margin-based federated adversarial training approach. In *International Workshop on Trustable, Verifiable, and Auditable Federated Learning in Conjunction with AAAI*, Vol. 2022.

[3] Dengsheng Chen, Jie Hu, Vince Junkai Tan, Xiaoming Wei, and Enhua Wu. 2023. Elastic Aggregation for Federated Optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 12187–12197.

[4] Jean Ogier Du Terrail, Samy-Safwan Ayed, Edwige Cyffers, Felix Grimberg, Chaoyang He, Regis Loeb, Paul Mangold, Tanguy Marchand, Othmane Marfoq, Erum Mushtaq, et al. 2022. FLamby: Datasets and Benchmarks for Cross-Silo Federated Learning in Realistic Healthcare Settings. In *NeurIPS 2022-Thirty-sixth Conference on Neural Information Processing Systems*.

[5] Liang Gao, Huazhu Fu, Li Li, Yingwen Chen, Ming Xu, and Cheng-Zhong Xu. 2022. FedDC: Federated Learning With Non-IID Data via Local Drift Decoupling and Correction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10112–10121.

[6] Liang Gao, Huazhu Fu, Li Li, Yingwen Chen, Ming Xu, and Cheng-Zhong Xu. 2022. Feddc: Federated learning with non-iid data via local drift decoupling and correction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10112–10121.

[7] Junyuan Hong, Haotao Wang, Zhangyang Wang, and Jiayu Zhou. 2022. Efficient Split-Mix Federated Learning for On-Demand and In-Situ Customization. In *In Proceedings of the International Conference on Learning Representations*.

[8] Samuel Horvath, Stefanos Laskaridis, Mario Almeida, Ilias Leontiadis, Stylianos Venieris, and Nicholas Lane. 2021. Fjord: Fair and accurate federated learning under heterogeneous targets with ordered dropout. *Advances in Neural Information Processing Systems* 34 (2021), 12876–12889.

[9] Wenke Huang, Mang Ye, and Bo Du. 2022. Learn From Others and Be Yourself in Heterogeneous Federated Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10143–10153.

[10] Yutao Huang, Lingyang Chu, Zirui Zhou, Lanjun Wang, Jiangchuan Liu, Jian Pei, and Yong Zhang. 2021. Personalized cross-silo federated learning on non-iid data. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 35. 7865–7873.

[11] Ekaterina Khramtsova, Christian Hammerschmidt, Sofian Lagraa, and Radu State. 2020. Federated learning for cyber security: SOC collaboration for malicious URL detection. In *2020 IEEE 40th International Conference on Distributed Computing Systems (ICDCS)*. IEEE, 1316–1321.

[12] Alex Krizhevsky, Geoffrey Hinton, et al. 2009. Learning multiple layers of features from tiny images. *Master's thesis, Department of Computer Science, University of Toronto, 2009*. (2009).

[13] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 11 (1998), 2278–2324.

[14] Ang Li, Jingwei Sun, Xiao Zeng, Mi Zhang, Hai Li, and Yiran Chen. 2021. Fedmask: Joint computation and communication-efficient personalized federated learning via heterogeneous masking. In *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems*. 42–55.

[15] Bo Li, Mikkel N. Schmidt, Tommy S. Alstrøm, and Sebastian U. Stich. 2023. On the Effectiveness of Partial Variance Reduction in Federated Learning With Heterogeneous Data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 3964–3973.

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

[17] Tian Li, Maziar Sanjabi, Ahmad Beirami, and Virginia Smith. 2019. Fair Resource Allocation in Federated Learning. In *International Conference on Learning Representations*.

[18] Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. 2019. On the Convergence of FedAvg on Non-IID Data. In *International Conference on Learning Representations*.

[19] Tao Lin, Lingjing Kong, Sebastian U Stich, and Martin Jaggi. 2020. Ensemble distillation for robust model fusion in federated learning. *Advances in Neural Information Processing Systems* 33 (2020), 2351–2363.

[20] Jian-Hao Luo et al. 2017. Thinet: A filter level pruning method for deep neural network compression. In *Proceedings of the IEEE international conference on computer vision*. 5058–5066.

[21] Lingjuan Lyu, Yitong Li, Karthik Nandakumar, Jiangshan Yu, and Xingjun Ma. 2020. How to democratise and protect AI: Fair and differentially private decentralised deep learning. *IEEE Transactions on Dependable and Secure Computing* (2020).

[22] Lingjuan Lyu, Xinyi Xu, Qian Wang, and Han Yu. 2020. Collaborative fairness in federated learning. In *Federated Learning*. Springer, 189–204.

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

[24] Matias Mendieta, Taojiannan Yang, Pu Wang, Minwoo Lee, Zhengming Ding, and Chen Chen. 2022. Local Learning Matters: Rethinking Data Heterogeneity in Federated Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 8397–8406.

[25] Pavlo Molchanov, Arun Mallya, Stephen Tyree, Iuri Frosio, and Jan Kautz. 2019. Importance estimation for neural network pruning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11264–11272.

[26] Pavlo Molchanov, Stephen Tyree, Tero Karras, Timo Aila, and Jan Kautz. 2017. Pruning Convolutional Neural Networks for Resource Efficient Inference. In *In Proceedings of the International Conference on Learning Representations*.

[27] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. 2011. Reading digits in natural images with unsupervised feature learning. *In NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011* (2011).

[28] Jean Ogier du Terrail, Samy-Safwan Ayed, Edwige Cyffers, Felix Grimberg, Chaoyang He, Regis Loeb, Paul Mangold, Tanguy Marchand, Othmane Marfoq, Erum Mushtaq, et al. 2022. FLamby: Datasets and Benchmarks for Cross-Silo Federated Learning in Realistic Healthcare Settings. *Advances in Neural Information Processing Systems* 35 (2022), 5315–5334.

[29] Zhen Qin, Shuiguang Deng, Mingyu Zhao, and Xueqiang Yan. 2023. FedAPEN: Personalized Cross-silo Federated Learning with Adaptability to Statistical Heterogeneity. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 1954–1964.

[30] Adam Richardson et al. 2020. Budget-bounded incentives for federated learning. In *Federated Learning*. Springer, 176–188.

[31] Yuxin Shi et al. 2021. A survey of fairness-aware federated learning. *arXiv preprint arXiv:2111.01872* (2021).

[32] Yuxin Shi, Han Yu, and Cyril Leung. 2023. Towards fairness-aware federated learning. *IEEE Transactions on Neural Networks and Learning Systems* (2023).

[33] Rachael Hwee Ling Sim, Yehong Zhang, Mun Choon Chan, and Bryan Kian Hsiang Low. 2020. Collaborative machine learning with incentive-aware model rewards. In *International Conference on Machine Learning*. PMLR, 8927–8936.

[34] Sebastian U Stich. 2018. Local SGD converges fast and communicates little. *arXiv preprint arXiv:1805.09767* (2018).

[35] Sebastian U Stich, Jean-Baptiste Cordonnier, and Martin Jaggi. 2018. Sparsified SGD with memory. *Advances in Neural Information Processing Systems* 31 (2018).

[36] Benyuan Sun, Hongxing Huo, Yi Yang, and Bo Bai. 2021. Partialfed: Cross-domain personalized federated learning via partial initialization. *Advances in Neural Information Processing Systems* 34 (2021), 23309–23320.

[37] Yue Tan, Yixin Liu, Guodong Long, Jing Jiang, Qinghua Lu, and Chengqi Zhang. 2023. Federated learning on non-iid graphs via structural knowledge sharing. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 37. 9953–9961.

[38] Sebastian Shenghong Tay, Xinyi Xu, Chuan Sheng Foo, and Bryan Kian Hsiang Low. 2022. Incentivizing collaboration in machine learning via synthetic data rewards. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 9448–9456.

[39] Md Palash Uddin, Yong Xiang, Xuequan Lu, John Yearwood, and Longxiang Gao. 2020. Mutual information driven federated learning. *IEEE Transactions on Parallel and Distributed Systems* 32, 7 (2020), 1526–1538.

[40] Tianhao Wang, Johannes Rausch, Ce Zhang, Ruoxi Jia, and Dawn Song. 2020. A principled approach to data valuation for federated learning. *Federated Learning: Privacy and Incentive* (2020), 153–167.

[41] Yangyang Wang, Xiao Zhang, Mingyi Li, Tian Lan, Huashan Chen, Hui Xiong, Xiuzhen Cheng, and Dongxiao Yu. 2023. Theoretical Convergence Guaranteed Resource-Adaptive Federated Learning with Mixed Heterogeneity. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2444–2455.

[42] Zihui Wang, Zhaopeng Peng, Xiaoliang Fan, Zheng Wang, Shangbin Wu, Rongshan Yu, Peizhen Yang, Chuanpan Zheng, and Cheng Wang. 2024. FedAVE: Adaptive data value evaluation framework for collaborative fairness in federated learning. *Neurocomputing* (2024), 127227.

[43] Xinyi Xu and Lingjuan Lyu. 2020. A reputation mechanism is all you need: Collaborative fairness and adversarial robustness in federated learning. *arXiv preprint arXiv:2011.10464* (2020).

[44] Xinyi Xu, Lingjuan Lyu, Xingjun Ma, Chenglin Miao, Chuan Sheng Foo, and Bryan Kian Hsiang Low. 2021. Gradient driven rewards to guarantee fairness in collaborative machine learning. *Advances in Neural Information Processing Systems* 34 (2021), 16104–16117.

[45] Yuan-Yi Xu, Ci-Siang Lin, and Yu-Chiang Frank Wang. 2023. Bias-Eliminating Augmentation Learning for Debiased Federated Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 20442–20452.

[46] Gang Yan, Hao Wang, Xu Yuan, and Jian Li. 2023. Criticalfl: A critical learning periods augmented client selection framework for efficient federated learning. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2898–2907.

[47] Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. 2019. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)* 10, 2 (2019), 1–19.

[48] Han Yu, Zelei Liu, Yang Liu, Tianjian Chen, Mingshu Cong, Xi Weng, Dusit Niyato, and Qiang Yang. 2020. A fairness-aware incentive scheme for federated learning. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 393–399.

[49] Hao Yu, Sen Yang, and Shenghuo Zhu. 2019. Parallel restarted SGD with faster convergence and less communication: Demystifying why model averaging works for deep learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 5693–5700.

[50] Ruichi Yu, Ang Li, Chun-Fu Chen, Jui-Hsin Lai, Vlad I Morariu, Xintong Han, Mingfei Gao, Ching-Yung Lin, and Larry S Davis. 2018. Nisp: Pruning networks using neuron importance score propagation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 9194–9203.

[51] Yaodong Yu, Alexander Wei, Sai Praneeth Karimireddy, Yi Ma, and Michael Jordan. 2022. TCT: Convexifying federated learning using bootstrapped neural tangent kernels. *Advances in Neural Information Processing Systems* 35 (2022), 30882–30897.

[52] Mikhail Yurochkin, Mayank Agarwal, Soumya Ghosh, Kristjan Greenewald, Nghia Hoang, and Yasaman Khazaeni. 2019. Bayesian nonparametric federated learning of neural networks. In *International Conference on Machine Learning*. PMLR, 7252–7261.

[53] Yufeng Zhan, Jie Zhang, Zicong Hong, Leijie Wu, Peng Li, and Song Guo. 2021. A survey of incentive mechanism design for federated learning. *IEEE Transactions on Emerging Topics in Computing* (2021).

[54] Jianqing Zhang, Yang Hua, Hao Wang, Tao Song, Zhengui Xue, Ruhui Ma, and Haibing Guan. 2023. Fedala: Adaptive local aggregation for personalized federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 11237–11244.

[55] Jingwen Zhang, Yuezhou Wu, and Rong Pan. 2021. Incentive mechanism for horizontal federated learning based on reputation and reverse auction. In *Proceedings of the Web Conference 2021*. 947–956.

[56] Yuchen Zhang, Martin J Wainwright, and John C Duchi. 2012. Communication-efficient algorithms for statistical optimization. *Advances in neural information processing systems* 25 (2012).

A. PROOF OF THEOREM 2

Let I_E be the set of global synchronization steps, i.e., $I_E = \{nE|n = 1, 2, ...\}$. For convenience, We define v_i^{t+1} as the immediate result of one step SGD update from θ_i^t, i.e., $v_i^{t+1} = \theta_i^t - \eta_t \nabla F_i(\theta_i^t, \xi_i^t)$. $\bar{g}_t = \sum_{i=1}^N \frac{\nabla F_i(\theta_i^t)}{p_i}$ and $g_t = \sum_{i=1}^N \frac{\nabla F_i(\theta_i^t, \xi_i^t)}{p_i}$. Therefore, $\bar{v}_{t+1} = \bar{\theta}_t - \eta_t g_t$ and $Eg_t = \bar{g}_t$.

LEMMA 3. *(Result of one step SGD). Assume ASSUMPTION 1 and ASSUMPTION 2. If $\eta_t \leqslant \frac{1}{4L}$, we have*

$$E\|\bar{v}_{t+1} - \theta^*\|^2 \leqslant (1 - \eta_t\mu)E\|\bar{\theta}_t - \theta^*\|^2 + \eta_t^2 E\|g_t - \bar{g}_t\|^2$$
$$+ 6L\eta_t^2\Gamma + 2E\sum_{i=1}^N \frac{\|\bar{\theta}_t - \theta_i^t\|^2}{p_i}, \quad (20)$$

where $\Gamma = F^* - \sum_{i=1}^N \frac{F_i^*}{p_i}$. LAMMA 3 has been made by [18].

For Assumption 3, the variance of stochastic gradients within device i is constrained by σ_i^2. Consequently,

$$E\|g_t - \bar{g}_t\|^2 = E\|\sum_{i=1}^N \frac{1}{p_i}(\nabla F_k(\theta_i^t, \xi_i^t) - \nabla F_i(\theta_i^t))\|^2$$
$$= \sum_{i=1}^N \frac{1}{p_i^2} E\|\nabla F_i(\theta_i^t, \xi_i^t) - \nabla F_i(\theta_i^t)\|^2 \quad (21)$$
$$\leqslant \sum_{i=1}^N \frac{1}{p_i^2}\sigma_i^2.$$

As FedSAC requires communication each E steps. We let $\eta_t \leqslant 2\eta_{t+E}$. Therefore, for any $t \geq 0$, there exists a $t_0 \leqslant t$, such that $t - t_0 \leqslant E - 1$ and $\theta_i^{t_0} = \bar{\theta}_{t_0}$ for all $k = 1, 2, ..., N$. Then

$$E\sum_{i=1}^N \frac{1}{p_i}\|\bar{\theta}_t - \theta_i^t\|^2 = E\sum_{i=1}^N \frac{1}{p_i}\|(\theta_i^t - \bar{\theta}_{t_0}) - (\bar{\theta}_t - \bar{\theta}_{t_0})\|^2$$
$$\leqslant E\sum_{i=1}^N \frac{1}{p_i}\|\theta_i^t - \bar{\theta}_{t_0}\|^2$$
$$\leqslant E\sum_{t=t_0}^{t-1}(E-1)\eta_t^2\|\nabla F_k(\theta_i^t, \xi_i^t)\|^2 \quad (22)$$
$$\leqslant \sum_{t=t_0}^{t-1}(E-1)\eta_{t_0}^2 G^2$$
$$\leqslant \eta_{t_0}^2(E-1)^2G^2$$
$$\leqslant 4\eta_t^2(E-1)^2G^2.$$

Here in lines 1252-1256, we use $E\|X - EX\|^2 \leqslant E\|X\|^2$ where $X = \theta_i^t - \bar{\theta}_{t_0}$ with probability $\frac{1}{p_i}$. In the lines 1256-1259, we use Jensen inequality:

$$\|\theta_i^t - \bar{\theta}_{t_0}\| = \|\sum_{t=t_0}^{t-1} \eta_t \nabla F_i(\theta_i^t, \xi_i^t)\|^2$$
$$\leqslant (t - t_0)\sum_{t-t_0}\sum_{t-t_0}^{t-1} \eta_t^2\|\nabla F_i(\theta_i^t, \xi_i^t)\|^2. \quad (23)$$

In lines 1259-1262, we utilize $\eta_t \leqslant \eta_{t_0}$ for $t \geq t_0$ and $E\|\nabla F_k(\theta_i^t, \xi_i^t)\|^2 \leqslant G^2$ for $i = 1, 2, ..., N$. In the lines 1263-1265, we use $\eta_{t_0} \leqslant 2\eta_{t_0+E} \leqslant 2\eta_t$ for $t_0 \leqslant t \leqslant t_0 + E$.

Let $\triangle_t = E\|\bar{\theta}_t - \theta^*\|$. Fromr Eq.(20), Eq. (21), and Eq. (22), it follows that

$$\triangle_{t+1} \leqslant (1 - \eta_t\mu)\triangle_t + \eta_t^2 \sum_{i=1}^N \frac{\sigma^2}{p_i^2} + 6L\eta_t^2\Gamma + 8\eta_t^2(E-1)^2G^2$$
$$\leqslant (1 - \eta_t\mu)\triangle_t + \eta_t^2 \underbrace{(\sum_{i=1}^N \frac{\sigma^2}{p_i^2} + 6L\Gamma + 8(E-1)^2G^2)}_{B} \quad (24)$$

For a diminishing stepsize, $\eta_t = \frac{\kappa}{t+\gamma}$ for some $\kappa > \frac{1}{\mu}$ and $\gamma > 0$ such that $\eta_1 \leqslant min\{\frac{1}{\mu}, \frac{1}{4L}\} = \frac{1}{4L}$ and $\eta_t \leqslant 2\eta_{t+E}$. We will prove $\triangle \leqslant \frac{v}{\gamma+t}$ by induction, where $v = max\{\frac{\kappa^2B}{\kappa\mu-1}, (\gamma+1)\triangle_1\}$. Firstly, the definition of v guarantees its applicability for $t = 1$. Assuming the conclusion holds for some t, it follows that

$$\triangle_{t+1} \leqslant (1 - \eta_t\mu)\triangle_t + \eta_t^2 B$$
$$\leqslant (1 - \frac{\kappa\mu}{t+\gamma})\frac{v}{t+\gamma} + \frac{\kappa^2B}{(t+\gamma)^2}$$
$$= \frac{t+\gamma-1}{(t+\gamma)^2}v + [\frac{\kappa^2B}{(t+\gamma)^2} - \frac{\kappa\mu-1}{(t+\gamma)^2}v]$$
$$\leqslant \frac{t+\gamma-1}{(t+\gamma)^2}v + \frac{\kappa^2B}{(t+\gamma)^2} - \frac{\kappa^2B}{(t+\gamma)^2} \underbrace{- \frac{\kappa\mu-1}{(t+\gamma)^2(\gamma+1)\triangle_1}}_{\leqslant 0} \quad (25)$$
$$\leqslant \frac{v}{t+\gamma-1}$$

Then by the L-smoothness (ASSUMPTION 1) of F,

$$E[F(\bar{\theta}_T)] - F^* \leqslant (\bar{\theta}_T - \theta^*)^T \underbrace{\nabla F_i(\theta^*)}_{=0} + \frac{L}{2}\|\bar{\theta}_T - \theta^*\|_2^2$$
$$= \frac{L}{2}\triangle_T \quad (26)$$
$$\leqslant \frac{L}{2}\frac{v}{\gamma+T}$$

We let $\kappa = \frac{2}{\mu}$, $\gamma = max\{8\frac{L}{\mu}, E\} - 1$. In the lines 1293, we have

$$v = max\{\frac{\kappa^2B}{\kappa\mu-1}, (\gamma+1)\triangle_1\}$$
$$\leqslant \frac{\kappa^2B}{\kappa\mu-1} + (\gamma+1)\triangle_1 \quad (27)$$
$$\leqslant \frac{4B}{\mu^2} + (\gamma+1)\triangle_1$$

Substituting Eq. 27 into Eq. 26, we obtain

$$0 \leqslant \lim_{T\to\infty} E[F(\bar{\theta}_T)] - F^* \leqslant \lim_{T\to\infty} [\frac{L}{\gamma+T}(\frac{2B}{\mu^2} + \frac{\gamma+1}{2}\triangle_1)] = 0 \quad (28)$$

Therefore, $\lim_{T\to\infty} E[F(\bar{\theta}_T)] - F^* = 0$.

Dataset	CIFAR10									Fashion MNIST				
No. Clients	20					40		60		20				
Scene	POW	CLA	DIR(1.0)	DIR(2.0)	DIR(3.0)	DIR(1.0)	DIR(2.0)	DIR(1.0)	DIR(2.0)	POW	CLA	DIR(1.0)	DIR(2.0)	DIR(3.0)
FedAvg[23]	-40.52±3.3	83.57±1.4	18.41±16.3	77.52±1.5	44.13±16.0	79.53±13.9	54.37±11.1	51.64±17.8	55.22±6.8	-34.64±5.6	84.37±3.3	22.95±2.6	30.39±6.7	56.37±3.5
q-FFL[17]	14.10±2.8	98.09±0.2	83.98±2.0	86.14±2.9	90.00±0.5	87.81±1.6	80.46±1.1	80.65±2.9	85.76±1.2	29.24±5.3	98.44±2.6	81.75±4.8	77.45±5.6	72.89±3.9
CFFL[22]	81.45±1.0	95.93±0.8	76.72±2.1	76.09±1.3	63.79±0.5	50.33±1.2	49.61±0.6	86.59±1.3	87.52±0.5	88.02±0.4	92.29±2.5	72.74±2.4	78.40±1.2	75.36±1.6
CGSV[44]	83.30±1.8	96.80±0.1	79.85±0.9	79.73±1.3	85.72±0.4	82.90±0.4	80.47±0.7	85.91±1.2	77.91±0.8	83.85±0.4	94.04±0.8	88.32±2.5	83.68±1.1	74.15±1.7
FedAVE[42]	88.46±1.5	97.18±0.6	87.59±1.1	78.43±0.4	68.70±1.3	45.69±2.6	65.60±1.2	38.92±1.1	60.61±1.3	87.14±0.6	93.97±1.2	78.87±2.1	81.20±1.5	72.73±0.7
Ours	99.62±0.2	98.52±0.1	96.29±0.4	98.06±0.5	95.99±0.4	99.37±0.5	96.57±0.2	97.69±0.6	95.12±0.3	97.40±0.4	98.49±0.3	96.52±0.8	97.65±0.6	96.47±0.7

Table 5: Comparison results of fairness $\rho \in [-100, 100]$ with state-of-the-art methods on CIFAR10 and Fashion MNIST. The reported results are averaged over 5 runs with different random seeds. (A higher value indicates better fairness. The best average result is marked in bold. The second-best result is underlined. These notes are the same to others.)

Dataset	CIFAR10									Fashion MNIST				
No. Clients	20					40		60		20				
Scene	POW	CLA	DIR(1.0)	DIR(2.0)	DIR(3.0)	DIR(1.0)	DIR(2.0)	DIR(1.0)	DIR(2.0)	POW	CLA	DIR(1.0)	DIR(2.0)	DIR(3.0)
Standalone	37.15±0.0	35.25±0.3	27.82±0.0	31.83±0.1	33.56±0.1	32.13±0.1	28.86±0.2	28.58±0.2	27.84±0.1	82.36±0.1	81.17±0.2	66.71±0.2	68.00±0.2	73.63±0.3
FedAvg[23]	46.94±0.2	41.14±0.6	48.31±0.5	49.00±0.5	50.13±0.3	48.19±0.3	49.57±0.3	49.06±0.1	49.30±0.3	87.29±0.3	84.67±0.2	87.46±0.0	88.06±0.1	88.04±0.1
q-FFL[17]	46.87±0.2	41.56±0.3	33.77±0.1	38.17±0.9	43.91±0.6	38.64±1.5	39.37±0.4	34.01±0.1	36.77±0.5	85.79±0.2	81.10±0.1	68.30±1.7	78.59±1.8	77.69±1.1
CFFL[22]	46.06±0.0	39.43±0.3	45.76±0.3	48.57±0.2	48.42±0.3	39.54±0.3	39.16±0.6	42.33±0.4	41.06±0.6	85.88±1.8	81.75±2.1	81.69±0.6	84.16±0.4	87.34±0.4
CGSV[44]	46.29±0.2	37.75±1.4	46.72±1.0	48.45±0.3	49.24±0.2	46.75±0.2	48.32±0.1	49.11±0.7	48.58±0.3	87.21±0.2	84.25±0.2	85.17±0.7	86.85±0.3	88.07±0.2
FedAVE[42]	46.43±0.6	40.99±0.2	46.64±0.6	48.34±0.5	48.13±0.6	46.58±0.4	46.73±0.3	47.29±0.8	49.21±0.6	87.44±0.1	84.79±0.5	79.53±0.9	84.47±0.4	86.83±0.1
Ours	48.60±0.7	43.39±0.2	49.41±0.1	49.09±0.1	50.48±0.1	48.64±0.3	49.68±0.4	49.23±0.1	49.34±0.2	87.60±0.1	84.99±0.3	87.57±0.6	88.08±0.3	88.17±0.3

Table 6: Comparison results of the maximum test accuracy (%) with state-of-the-art methods on CIFAR10 and Fashion MNIST. The reported results are averaged over 5 runs with different random seeds. (A higher value indicates better accuracy.)
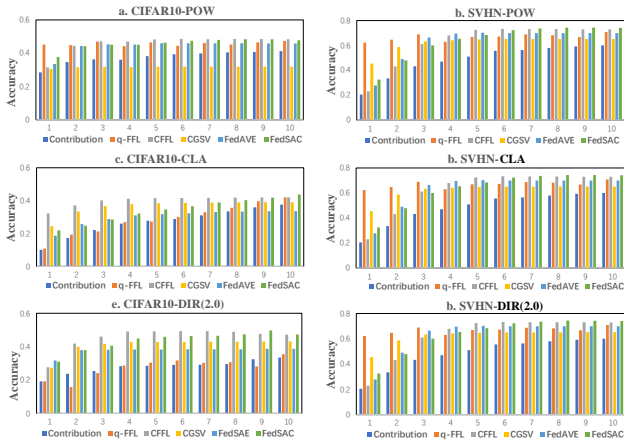
Figure 6: Comparison results of test accuracy using the scene of POW, CLA, and DIR(2.0) with state-of-the-art methods in CIFAR10 (left) and SVHN (right). In terms of accuracy across clients, FedSAC exhibits the highest level of consistency with the contribution.

Scene	POW	DIR(1.0)	DIR(2.0)	DIR(3.0)
$FedSAC(\beta = 2)$	49.23(44.77)	49.66(39.53)	50.60(37.87)	51.37(43.19)
$FedSAC(\beta = 5)$	48.43(37.73)	49.10(35.76)	50.01(30.91)	49.85(38.46)
$FedSAC(\beta = 10)$	48.61(30.91)	49.06(31.36)	47.08(23.72)	49.04(28.54)
$FedSAC(\beta = 20)$	46.53(24.10)	46.98(28.41)	43.50(20.17)	46.03(25.70)

Table 7: The maximum test accuracy (%) achieved by FedSAC across different β, given a fairness threshold of $\rho > 95\%$, on CIFAR10. Values in the middle brackets represent the minimum test accuracy (%) among 10 clients.

Existing methods allocate rewards to clients lack sufficient differentiation, resulting in an ongoing unfairness for high-contributing clients. For example in SVHN-CLA (Figure 6 (d)), the contributions (i.e., Standalone) of $Client_3$ and $Client_{10}$ differ significantly. However, CFFL and CGSV do not exhibit a substantial difference in the rewards assigned to them. In addition, FedSAC effectively differentiates the rewards it received, thereby ensuring the collaborative fairness in FL.

C. VARYING NUMBERS OF CLIENTS

To verify the effectiveness of FedSAC in scenarios with varying numbers of clients, we conduct experiments by increasing the number of local clients to 20, 40, and 60, respectively. Table 5 presents the fairness results achieved by FedSAC and the compared methods, while Table 6 shows the maximum local model performance

B. FINAL REWARDS OF CLIENTS

Figure 6 shows the distributions of contributions and the final rewards of clients under different scenes (i.e., POW, CLA, DIR (2.0) in CIFAR10 and SVHN) by FedSAC and the compared methods.

Dataset	CIFAR10					SVHN					Fashion MNIST				
No. Clients	10					10					10				
Scene	POW	CLA	DIR(1.0)	DIR(2.0)	DIR(3.0)	POW	CLA	DIR(1.0)	DIR(2.0)	DIR(3.0)	POW	CLA	DIR(1.0)	DIR(2.0)	DIR(3.0)
FedAvg[23]	26.27	26.27	26.27	26.27	26.27	26.27	26.27	26.27	26.27	26.27	7.97	7.97	7.97	7.97	7.97
q-FFL[17]	26.27	26.27	26.27	26.27	26.27	26.27	26.27	26.27	26.27	26.27	7.97	7.97	7.97	7.97	7.97
CFFL[22]	26.27	26.27	26.27	26.27	26.27	26.27	26.27	26.27	26.27	26.27	7.97	7.97	7.97	7.97	7.97
CGSV[44]	26.27	26.27	26.27	26.27	26.27	26.27	26.27	26.27	26.27	26.27	7.97	7.97	7.97	7.97	7.97
FedAVE[42]	26.27	26.27	26.27	26.27	26.27	26.27	26.27	26.27	26.27	26.27	7.97	7.97	7.97	7.97	7.97
Ours	23.57	23.58	23.39	24.02	23.63	20.86	22.31	23.85	15.51	21.75	6.89	7.28	5.30	4.77	5.23

Table 8: Comparison on communication costs (MB per round) of FedSAC and the baselines framework.

Dataset	CIFAR10					SVHN					Fashion MNIST				
No. Clients	10					10					10				
Scene	POW	CLA	DIR(1.0)	DIR(2.0)	DIR(3.0)	POW	CLA	DIR(1.0)	DIR(2.0)	DIR(3.0)	POW	CLA	DIR(1.0)	DIR(2.0)	DIR(3.0)
η	0.05	0.05	0.05	0.03	0.03	0.03	0.05	0.05	0.05	0.05	0.1	0.05	0.05	0.05	0.05
E	15	20	20	20	20	20	20	20	20	20	20	20	20	20	15
β	5	3	10	2	10	5	3	5	10	5	10	1	10	20	25

Table 9: The optimal hyperparameters in given scenarios. Learning rate η, the number of local update steps E, hyper-parameter β. Batch sizes B are set as 32, 64, and 32 for CIFAR10, SVHN, and FashionMNIST, respectively.

achieved by these methods. In the large client number settings (No. Clients = 20, 40, and 60), FedSAC outperforms all baseline methods in terms of fairness (refer to Table 5) and accuracy (refer to Table 6). The results demonstrate that FedSAC can effectively implement bounded collaborative fairness in scenarios with varying numbers of clients.

D. THE IMPACT OF β ON THE EXPERIMENT

In Table 7, we present the performance of FedSAC with different values of β on each scene of CIFAR10. The experiments demonstrate that as β increases, the maximum test accuracy will gradually decrease. This is because the size of the submodels downloaded by clients increases as β decreases. When β is small, the submodels of low-contribution clients contain more neurons, enabling effective training to enhance all local model performance.

E. THE COMMUNICATION COST EXPERIMENTS

The most works on collaborative fairness require full clients' information [22, 43, 44], which will inevitably introduce large communication overhead and computation cost. In addition, this communication and computation overhead of full sampling problem is tolerable for most cases in cross-silo FL scenarios [4], such as healthcare and finance, because there are only dozens of clients in cross-silo FL scenarios. In the Table 8, the results show that our FedSAC demonstrates less communication costs than all the baseline methods in three datasets across all settings. In addition, our FedSAC does not introduce additional communication, because the evaluation of neuron importance is conducted on the server without communicating with clients, effectively preventing any communication overhead. In the Table 8, FedSAC exhibits a communication complexity of O(d*m) per round [8], where m<=1 represents the average ratio of the parameters of the submodel compared to the global model.

In conclusion, all baseline methods show a higher communication complexity of O(d) than FedSAC.

F. DETAILS ON HYPER-PARAMETERS

For each dataset, the local data of each client was partitioned into training and validation sets. Then, we tuned each dataset hyperparameters by using grid search with FedAvg. Subsequently, we applied the optimal parameters obtained from the validation dataset. The optimized hyper-parameters for scenarios (i.e., POW, CLA, DIR(1.0), DIR(2.0), and DIR(3.0)) are shown in Table 9.