Fed-SAP: Statistics-Aligned Personalization for Federated Attention on Non-IID Data

Jiao Tian*
School of Computer Science and Technology
Xinjiang University, Urumqi, China
Urumqi, China
jiaotian@xju.edu.cn

Jinlin He
School of Computer Science and Technology
Xinjiang University, Urumqi, China
Urumqi, China
107552303934@stu.xju.edu.cn

***Abstract*—In Non-IID federated learning, blind aggregation of attention modules such as SE leads to "harmful aggregation". Experiments in this paper confirm that Convolution (Conv) layers learn general knowledge, while the Excitation (SE-E) layers of SE are highly personalized; based on this, we propose the Fed-SAP framework, which aggregates Conv layers to share general knowledge and personalizes SE-E layers to preserve local characteristics. To address the statistical bias between personalized SE-E layers and their inputs (Squeeze statistical vectors), we innovatively introduce Global Statistical Regularization (GSR), which through an additional loss function forces local Conv layers to actively align their channel statistical characteristics with global statistics when extracting features, thereby alleviating feature drift at the source. This provides high-quality inputs for personalized layers and achieves a synergistic effect significantly superior to using either component alone, and experiments on datasets show that Fed-SAP significantly outperforms baseline methods like FedAvg and FedBN.**

***Index Terms*—Collaborative Processing, Federated Learning, Non-IID, Personalization, Attention Mechanism.**

I. Introduction

Federated Learning (FL) [1], as an emerging distributed machine learning paradigm [7], enables collaborative model training across a large number of edge devices (such as mobile phones and IoT devices) while protecting data privacy. Its core idea is to keep data locally and only aggregate model parameters on the server side. Standard federated learning algorithms, such as FedAvg [1] , construct a global model by weighted averaging client-side models.

However, FedAvg and its variants face a core challenge in real-world scenarios: non-independent and identically distributed (Non-IID) data. Data on edge devices usually exhibits significant distributional differences, which causes clients to converge toward different optimization directions during local training, a phenomenon known as "client drift" [2]. This drift renders the aggregation of global models inefficient or even harmful, leading to slower convergence and decreased final accuracy [3] [4].

*First and corresponding author: Jiao tian (jiaotian@xju.edu.cn)

This work was supported in part by the Basic Scientific Research Business Expenses Project for Universities in Xinjiang Uyghur Autonomous Region (Construction Type Project) under Grant XJEDU2024J032, and in part by the Natural Science Foundation of Xinjiang Uyghur Autonomous Region under Grant 2024D01C239.

To alleviate the Non-IID problem, the research community has proposed numerous works. One category of methods (e.g., FedProx [8]) introduces a proximal term into the local loss function to restrict local updates from deviating too far from the global model, thereby mitigating drift [9]. Another line of research recognizes that forcing a "one-size-fits-all" global model on highly heterogeneous data is inherently unreasonable [5] [6]. Consequently, personalized federated learning (pFL) [10] [11] has emerged, aiming to learn a customized model for each client.

Among the various branches of pFL, architecture-based personalization methods have demonstrated great potential. For instance, FedPer [12] proposes to learn a shared base network while keeping the final layers (head) of the model personalized for each client.Among them, FedBN [13] [14] is a landmark work. Inspired by the high sensitivity of Batch Normalization (BN) layers to the statistical characteristics (mean and variance) of data, FedBN finds that under Non-IID data, the statistics of BN layers differ drastically, and aggregating BN layer parameters severely impairs model performance. Therefore, FedBN proposes to only aggregate feature extraction parameters such as Convolution (Conv) layers, while keeping the BN layer of each client fully personalized (i.e., trained and used locally without ever being uploaded).

The success of FedBN raises a critical question: if components dependent on data statistics like BN layers are unsuitable for aggregation, what about other statistics-dependent components widely used in modern networks—such as attention mechanisms?

Modern efficient networks, such as MobileNetV3 [15], generally rely on Squeeze-and-Excitation (SE) attention modules to dynamically recalibrate channel features. The core of the SE module consists of a Squeeze operation that uses Global Average Pooling (GAP) to compute channel statistical vectors, and an Excitation operation typically implemented by two fully connected layers, where the Excitation operation is the only part of SE with trainable parameters. The behavior of SE modules in Non-IID environments has been largely overlooked.

To explore the characteristics of SE modules in FL, we conducted a preliminary study. We found that under IID

settings, the parameters of Conv layers and SE-E (Excitation) layers across different clients maintain high cosine similarity after training. However, under Non-IID settings, a striking phenomenon emerges: the parameter similarity of Conv layers remains relatively high, while that of SE-E layers drops sharply. This finding strongly demonstrates that Conv layers tend to learn general, shareable low-level and mid-level features, whereas SE-E layers learn highly specialized knowledge dependent on local data distributions.

This observation reveals a fatal flaw of FedAvg when handling networks with SE modules: forcibly averaging these highly differentiated SE-E layer parameters leads to "harmful aggregation", resulting in a "neither fish nor fowl" attention module that is ineffective for any client. Based on this core insight, we propose Fed-SAP (Statistics-Aligned Personalization for Federated Attention), a personalized framework specifically designed for federated attention networks. Inspired by FedBN, the fundamental strategy of Fed-SAP is to "aggregate the general and personalize the specific": we aggregate shared Conv layers while keeping the SE-E layers of each client fully personalized (referred to as pSE). However, we further found that merely personalizing SE-E layers is insufficient. The problem lies in the input to SE-E layers — the channel statistical vector s generated by the Squeeze operation — which is still produced by aggregated Conv layers operating on local data. In extreme Non-IID scenarios (e.g., client A has only cat data, client B has only dog data), this s vector is highly biased. This forces the personalized SE-E layers to passively adapt to these low-quality, biased local statistics.

To address this "input bias" issue, we innovatively introduce the Global Statistics Regularization (GSR) mechanism. The core idea of GSR is that instead of passively adapting to the biased s, we should actively correct it. GSR adds a lightweight "statistical vector" exchange to the federated learning workflow. The server maintains a global average statistical vector S_G. During local training, in addition to computing the conventional classification loss, clients need to calculate a GSR loss, which is the L2 distance between the local batch statistics s and the global statistics S_G. The gradient of this GSR loss propagates backward, directly acting on the Conv layers before the Squeeze operation. This is equivalent to forcing the Conv layers, when extracting features, to not only consider the local task but also actively align the distribution of their generated channel statistical characteristics with global statistics. The GSR mechanism alleviates feature drift caused by Non-IID data at the source, providing "purified" high-quality, globally aligned inputs for personalized SE-E layers. This enables a 1+1>2 synergistic effect between pSE (personalized selection) and GSR (high-quality inputs).

The main contributions of this paper are summarized as follows:

- We propose the Fed-SAP framework, which preserves client-specific attention patterns by personalizing SE-E layers.
- We design the Global Statistics Regularization (GSR) mechanism to alleviate the local bias of feature extractors at the source by aligning feature statistics.
- Extensive experiments on CIFAR-10 dataset using MobileNetV3-small show that Fed-SAP significantly outperforms mainstream baselines such as FedAvg, FedProx, and FedBN under various Non-IID scenarios.

II. Methodology

To address the "harmful aggregation" problem caused by blind aggregation of SE attention modules under Non-IID data as described earlier, we propose the Statistics-Aligned Personalization for Federated Attention (Fed-SAP) framework. This framework consists of two core components: personalized SE (pSE) modules and Global Statistical Regularization (GSR).

We first draw conclusions through a motivational experiment. We trained a model using standard FedAvg on Non-IID data (Dirichlet distribution $\alpha = 0.5$) and tracked the average cosine similarity of parameters across clients. As shown in Fig. 1, a striking phenomenon emerges: the parameter similarity of Conv layers (blue line) remains high, while that of SE-E layers (orange line) steadily decreases. This demonstrates that Conv layers learn general, shareable knowledge, whereas SE-E layers learn highly localized characteristics.

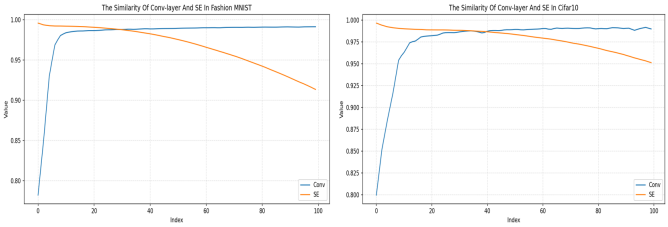

Fig. 1. Parameter cosine similarity of Conv layers (blue) and SE-E layers (orange) during FedAvg training under a Non-IID setting ($\alpha = 0.5$) on Fashion-MNIST (left) and CIFAR-10 (right).

Based on this core insight, we design the Personalized Squeeze-and-Excitation.

A. Personalized Squeeze-and-Excitation

Building on the above findings, we propose the first core component of Fed-SAP: personalized SE (pSE). Its core idea is to explicitly divide model parameters W into two parts: shared parameters W_C (including all Conv layers and other general feature extractors) and personalized parameters $W_{E,k}$ which is referring to the private SE-E layer parameters of client k.

To precisely elaborate our method, we first formally define a standard SE module. Given an input feature map $X \in \mathbb{R}^{H\times W\times C}$, the SE module first performs a Squeeze operation, typically compressing global spatial information into a channel descriptor $s \in \mathbb{R}^{1\times 1\times C}$ via Global Average Pooling (GAP):

$$s = F_{sq}(X) = \frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W} X_{i,j} \tag{1}$$

Subsequently, the Excitation module captures inter-channel dependencies through two fully connected (FC) layers to generate attention weights $a \in \mathbb{R}^{1\times 1\times C}$:

$$a = F_{ex}(s, W_E) = \sigma(W_2 \cdot \delta(W_1 \cdot s)) \tag{2}$$

where $W_E = \{W_1, W_2\}$ are the trainable parameters of the FC layers, δ is the ReLU activation function, and σ is the Sigmoid activation function. W_E is the only part of the SE module with trainable parameters. Finally, the Rescale operation multiplies each channel of X by the corresponding attention weight: $\hat{X} = F_{scale}(X, a) = a \cdot X$. Our pSE method specifically personalizes this W_E. Under the pSE framework, the federated learning process is modified as follows: in each round $t+1$, the server first performs weighted averaging on the shared parameters $W_{C,k}^t$ uploaded by all clients k to obtain global shared parameters W_C^{t+1}:

$$W_C^{t+1} \leftarrow \sum_{k=1}^{K} w_k W_{C,k}^t \quad (\text{where } w_k = n_k/N) \tag{3}$$

The server does not access or aggregate any personalized parameters $W_{E,k}$. After receiving W_C^{t+1}, client k combines it with the locally private and retained $W_{E,k}^t$ to form a complete local model $\{W_C^{t+1}, W_{E,k}^t\}$. Client k minimizes its local loss function L_k on its local data D_k:

$$L_k = \frac{1}{|D_k|} \sum_{(x,y)\in D_k} \mathcal{L}(\text{Net}(x; W_C^{t+1}, W_{E,k}^t), y) \tag{4}$$

During backpropagation, gradients update both W_C^{t+1} (resulting in $W_{C,k}^{t+1}$) and $W_{E,k}^t$ (resulting in $W_{E,k}^{t+1}$). After training, client k only uploads the updated $W_{C,k}^{t+1}$ while retaining the updated $W_{E,k}^{t+1}$ locally for the next training round. In this way, pSE enables W_C to learn general feature extraction capabilities from global data, while $W_{E,k}$ can specifically adapt to the particular distribution of local data, thereby effectively avoiding "harmful aggregation". However, we find that implementing pSE alone is insufficient. The only use of pSE addresses the issue of personalized modules but ignores the input to such modules. It introduces a deeper contradiction: input statistical mismatch. The input s_k processed by the personalized $W_{E,k}$ is generated by the Squeeze operation $s_k = F_{sq}(X_k)$. The feature map X_k is produced by the globally aggregated W_C on highly biased local data D_k. When a general extractor designed for all categories is forced to process only "cats", the resulting feature map X_k and its channel statistics s_k are inevitably low-quality and biased. This causes the learning objective of the personalized $W_{E,k}$ to degrade from "learning the true attention patterns of D_k" to "learning how to passively compensate for the statistical bias of s_k". Such passive adaptation fundamentally limits the upper bound of model performance. Therefore, a better solution should not only personalize W_E but also actively calibrate the input s_k it receives. This is exactly what our second innovation—Global Statistical Regularization (GSR)—aims to solve.

B. Global Statistical Regularization

To address the "input statistical bias" problem faced by pSE, we propose the second core component of Fed-SAP: Global Statistics Regularization (GSR). The core idea of GSR is to ensure that when client k trains its shared convolutional (Conv) layer W_C, the resulting local channel statistics s_k not only reflect local characteristics but also actively align with the statistical properties of the "global data" to some extent.

To achieve this, GSR introduces a lightweight "statistical vector" exchange into the standard federated learning process. In addition to maintaining the global shared parameter W_C^t, the server also maintains an auxiliary set of Global Mean Statistical Vectors, $\mathcal{S}_G^t = \{S_{G,1}^t, ..., S_{G,L}^t\}$. Each vector $S_{G,l}^t \in \mathbb{R}^{1\times1\times C_l}$ corresponds to the l-th SE module in the network, with C_l being its channel dimensionality. This set represents the average channel statistical properties of the global data across all modules. At the start of training ($t = 0$), all vectors in $\mathcal{S}_G^0$ are initialized as zero vectors. In each round $t+1$, the server distributes both W_C^{t+1} and the entire statistical set $\mathcal{S}_G^t$ to all clients.

When client k receives W_C^{t+1} and the global statistics set $\mathcal{S}_G^t = \{S_{G,1}^t, ..., S_{G,L}^t\}$, its local training objective is redefined. For every batch of data in the local dataset D_k, the client k's total loss $L_{total,k}$ is defined as the weighted sum of the standard classification loss $L_{CE,k}$ and the total GSR loss $L_{GSR,k}$:

$$L_{total,k} = L_{CE,k} + \lambda L_{GSR,k} \tag{5}$$

where λ is a hyperparameter used to balance the two loss terms. $L_{CE,k}$ is the standard cross-entropy loss, computed using the personalized $W_{E,k}^t$. The GSR loss $L_{GSR,k}$ is defined as the sum of L_2 distances across all L SE modules between the local batch statistics and the received global statistics:

$$L_{GSR,k} = \sum_{l=1}^{L} ||s_{batch,l} - S_{G,l}^t||_2^2 \tag{6}$$

Here, $s_{batch,l}$ is the local batch statistical vector for the l-th module, obtained by applying the shared Conv layers W_C^{t+1} to the current data batch X_{batch} followed by the l-th Squeeze operation, and $S_{G,l}^t$ is the corresponding global statistical vector for that module received from the server.

The introduction of this GSR loss term is crucial. During backpropagation, the gradients from $L_{CE,k}$ normally update W_C^{t+1} and the personalized $W_{E,k}^t$. The gradient from $\lambda L_{GSR,k}$, since F_{sq} (GAP) is differentiable and S_G^t is treated as a constant target during local training, will bypass the personalized SE-E module and backpropagate directly to all shared Conv layers (W_C^{t+1}) before the Squeeze operation.

$L_{GSR,k}$ acts as a regularizer that imposes a constraint on W_C^{t+1}. This forces the shared Conv layer W_C to learn a more generalized feature extraction method, closer to the global distribution, even when processing highly skewed local data. GSR calibrates the features at the source, mitigating feature drift caused by Non-IID data, thereby providing a high-quality, globally aligned input s_k for the subsequent personalized $W_{E,k}$ module.

To complete this closed loop, we also need to update the global statistics set $\mathcal{S}_G^t$. After local training (E epochs) is complete, client k computes a set of average local statistical vectors, $\mathcal{S}_{up,k} = \{s_{up,k,1}, ..., s_{up,k,L}\}$, by performing a

Algorithm 1 Fed-SAP

Server Algorithm:
Input: Initial W_C^0, $\mathcal{S}_G^0 = \{S_{G,1}^0, ..., S_{G,L}^0\}$, Rounds T, n_k
Output: Final $W_C^T, \mathcal{S}_G^T$
for $t = 0$ to $T-1$ **do**
  for each client k in parallel **do**
    $W_{C,k}^{t+1}, \mathcal{S}_{up,k} \leftarrow$
      ClientUpdate$(k, W_C^t, W_{E,k}^t, \mathcal{S}_G^t)$
  end for
  Server aggregates:
  $W_C^{t+1} \leftarrow \sum_{k=1}^{K} \frac{n_k}{N} W_{C,k}^{t+1}$
  for $l = 1$ to L **do**
    $S_{G,l}^{t+1} \leftarrow \sum_{k=1}^{K} \frac{n_k}{N} s_{up,k,l}$
  end for
end for
return $W_C^T, \mathcal{S}_G^T$

Function ClientUpdate$(k, W_C, W_{E,k}, \mathcal{S}_G)$
for local epoch $e = 1$ to E **do**
  for each batch (x_b, y_b) from D_k **do**
    $L_{CE,k}, \{s_{b,1}, ..., s_{b,L}\} \leftarrow$
      ForwardPass$(\{W_C, W_{E,k}\}, x_b, y_b)$
    Compute $L_{GSR,k}$ from $\{s_{b,l}\}$ and $\mathcal{S}_G$
    $L_{total} \leftarrow L_{CE,k} + \lambda L_{GSR,k}$
    Backpropagate L_{total} and update W_C and $W_{E,k}$
  end for
end for
Compute $\mathcal{S}_{up,k} = \{s_{up,k,1}, ..., s_{up,k,L}\}$ over D_k
return $W_C, W_{E,k}, \mathcal{S}_{up,k}$

forward pass over its entire local data D_k using the updated $W_{C,k}^{t+1}$. For each module l:

$$s_{up,k,l} = \frac{1}{|D_k|} \sum_{(x,y) \in D_k} s_{k,l}(x; W_{C,k}^{t+1}) \tag{7}$$

where $s_{k,l}(x; W_C)$ is the channel statistical vector output by the l-th Squeeze layer (F_{sq}) for a single sample x. Subsequently, client k uploads both its updated shared parameters $W_{C,k}^{t+1}$ and this complete set of local statistics $\mathcal{S}_{up,k}$ to the server. The server, while aggregating W_C, also computes a weighted average for each of the L statistical vectors independently to obtain the global set $\mathcal{S}_G^{t+1}$ for the next round:

$$S_{G,l}^{t+1} \leftarrow \sum_{k=1}^{K} w_k s_{up,k,l}, \quad \text{for } l = 1 \text{ to } L \tag{8}$$

Since each $s_{up,k,l}$ is just a C-dimensional vector, the additional communication overhead is negligible.

C. Privacy and Representation Specificity Analysis

Privacy Security. The GSR mechanism strictly adheres to federated privacy protocols. The exchanged information is limited to channel-wise statistical vectors. These vectors are irreversibly compressed into scalars. This process effectively eliminates spatial structures and semantic details, making the reconstruction of raw private data from statistical data mathematically infeasible. **Representation Specificity.** We further discuss the possibility that global alignment might dilute local specificity. We argue that GSR enhances performance by decoupling feature extraction from attention calibration without diluting local specificity. This synergy ensures that the model maintains high sensitivity to local data distributions without sacrificing the quality of feature representation. In summary, the Fed-SAP framework achieves a synergistic effect through the cooperation of pSE and GSR. GSR provides high-quality, globally aligned input statistics s_k, while pSE offers personalized attention $W_{E,k}$ specialized for the local task. Their combination allows the model to learn both robust global shared knowledge and retain efficient local personalization patterns.

III. Experiment

A. Ablation Study

To systematically verify the effectiveness and design rationale of each component in our Fed-SAP framework, we conducted a series of detailed ablation experiments. Unless otherwise specified, all ablation experiments in this section were conducted on the CIFAR-10 and Fashion-MNIST datasets. We simulated a Non-IID scenario with 10 clients ($K = 10$), where the data distribution was controlled by a Dirichlet distribution. All ablation experiments used MobileNetV3-Small as the backbone network and were run for 50 communication rounds ($T = 50$).

We first investigate the core issue of which part of the network should be personalized. Based on the analysis in the previous section, we hypothesize that aggregating the Conv layers and personalizing the SE-E layers is the optimal strategy. To verify this, we compare four different architectural strategies: (1) **FedAvg**, which aggregates all layers; (2) **Local-Only**, which is fully personalized with no aggregation; (3) **Agg-SE-E**, which personalizes the Conv layers and only aggregates the SE-E layers; and (4) **Our strategy (pSE)**, which aggregates the Conv layers and personalizes the SE-E layers.

The experimental results are shown in Table I and Fig. 2. Both FedAvg and Local-Only perform very poorly. The former suffers from "detrimental aggregation," while the latter is insufficiently trained due to data scarcity. The performance of Agg-SE-E (personalizing Conv layers) is even worse, which validates that the Conv layers learn generalizable knowledge that must be shared. In contrast, our pSE strategy achieves significantly the best performance, strongly confirming our core hypothesis. This confirms that we should personalize the SE's Excitation layer.

After establishing pSE as the baseline, we further investigate the effectiveness of the GSR component. First, we study the impact of the hyperparameter λ for the GSR loss term. As shown in Table II, we tested different values of λ on CIFAR-10 ($\alpha = 0.5$). We observe a "bell-shaped curve": when λ is too small (e.g., 0.1), the regularization constraint of GSR is too weak, and the performance improvement is limited; when

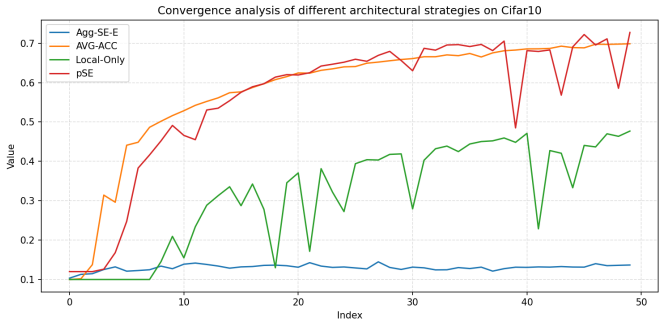

Fig. 2. Convergence analysis of different architectural strategies on Cifar-10 ($\alpha = 0.5$). Our pSE strategy achieves the best performance.

λ is too large (e.g., 5.0), the GSR loss dominates, forcing the model to over-focus on statistical alignment while neglecting the classification task, leading to performance degradation. The experiments show that the model achieves the best performance at $\lambda = 2.0$, striking the best balance between classification accuracy and statistical alignment. Therefore, we adopt $\lambda = 2.0$ in all subsequent experiments.

Next, we designed experiments to demonstrate the necessity of GSR's specific objective: "aligning with the global statistics S_G". We fixed $\lambda = 2.0$ and compared three different regularization targets: (1) **GSR (Ours)**, which aligns with the global statistics S_G; (2) **Local-Align**, which forces alignment with client k's local average statistics from the previous round, S_k^{t-1}; and (3) **L2-Reg**, which only applies L2 norm regularization to the output of the Squeeze layer, s_{batch}.

As shown in Table III, the results clearly validate our design rationale. **Local-Align** performed the worst among all groups, even causing the model to collapse, as it reinforced local bias. The performance of **L2-Reg** was almost identical to the pSE baseline without GSR, proving that simply suppressing the output magnitude yields no benefits. Only our **GSR (Ours)** achieved the highest accuracy, significantly outperforming the pSE baseline and all other variants. These ablation studies collectively demonstrate that pSE and GSR (specifically, GSR aligning with S_G) are two indispensable components in our framework, which work synergistically to achieve the final performance improvement. To comprehensively evaluate the performance of our method, we include an IID distribution in addition to our previous experimental setup, and also expand the performance evaluation to include heterogeneity scenarios of $\alpha = 0.8$ and $\alpha = 1.0$. We uniformly use MobileNetV3-Small as the backbone network. We compare our Fed-SAP (Ours) with four mainstream baseline methods: FedAvg, FedProx, FedBN, and Local-Only. Here, Local-Only involves all clients training independently on their local data without any aggregation, serving as a performance lower bound reference.

We summarize the final test accuracy of all methods under different data distributions in Table IV.

In the IID scenario, as expected, all aggregation-based methods (FedAvg, FedProx, FedBN, and Fed-SAP) perform well and achieve comparable performance, demonstrating that our method can achieve at least comparable performance to standard FedAvg on IID data.

However, once Non-IID data is introduced, a performance divergence begins to emerge. As data heterogeneity increases (α decreasing from 1.0 to 0.5), the performance of FedAvg drops sharply on both datasets. FedProx mitigates some of the performance degradation through regularization, but the effect is limited. Notably, FedBN, as a strong baseline, outperforms FedAvg and FedProx in Non-IID scenarios by personalizing the BN layers. However, FedBN still aggregates the SE-E layers, and therefore it also fails to resolve the statistical bias problem in the attention modules that we have identified.

In contrast, our Fed-SAP, through the dual mechanisms of pSE (personalizing SE-E layers) and GSR (global statistical alignment), consistently surpasses all baseline methods, including FedBN, across all Non-IID settings. The advantage of Fed-SAP is most pronounced, significantly widening the performance gap against the second-best (FedBN). This indicates that the more complex and heterogeneous the data, the more crucial our proposed "statistical alignment" and "attention personalization" become.

Finally, we present the test accuracy convergence curves for the different methods in Fig. 3. Local-Only and FedAvg converge slowly and to a low final accuracy. The convergence speed and stability of FedProx and FedBN are improved. Our

TABLE I
ABLATION STUDY ON ARCHITECTURAL STRATEGIES ON CIFAR-10

Strategy	Highest Acc. (%)	Average Acc. (%)
AVG-ACC	69.23	69.12
Local-Only	48.05	47.26
Agg-SE-E	16.25	15.22
pSE (Ours)	**75.2**	**71.4**

TABLE II
IMPACT OF HYPERPARAMETER λ ON CIFAR-10 ($\alpha = 0.5$)

Hyperparameter	Value	Avg. Test Accuracy (%)
λ	0.1	66.19
	0.5	67.31
	1.0	69.26
	2.0	**75.11**[a]
	5.0	68.9

[a] *The optimal value used in subsequent experiments.*

TABLE III
COMPARISON OF DIFFERENT REGULARIZATION TARGETS ($\lambda = 2.0$) ON CIFAR-10

Regularization Target	Avg. Test Accuracy (%)
GSR (Ours)	**75.34**
Local-Align	46.22
L2-Reg	71.98

TABLE IV
FINAL AVERAGE TEST ACCURACY (%) COMPARISON ON CIFAR-10 AND FASHION-MNIST UNDER DIFFERENT DATA DISTRIBUTIONS (T=50).

Dataset	**CIFAR-10**				**Fashion-MNIST**			
Distribution (α)	**IID**	**1.0**	**0.8**	**0.5**	**IID**	**1.0**	**0.8**	**0.5**
Local-Only	52.1	51.3	48.9	46.1	66.1	62.5	57.9	55.5
FedAvg	71.34	60.26	55.38	53.18	92.1	86.2	85.1	82.2
FedProx	71.77	71.62	70.0	66.8	92.2	87.5	86.74	84.7
FedBN	72.35	74.0	73.1	71.5	91.3	88.5	87.0	86.3
Fed-SAP (Ours)	**77.64**	**76.21**	**76.82**	**75.32**	**92.4**	**91.9**	**92.1**	**92.5**

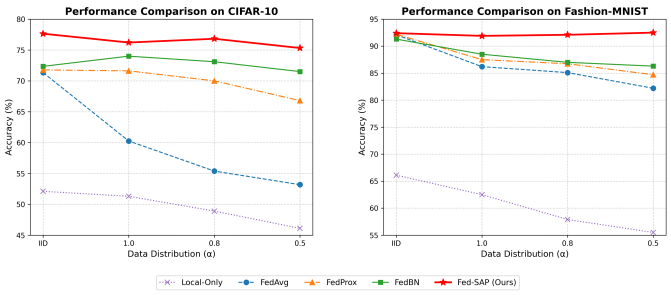

Fig. 3. Convergence curves of all methods on CIFAR-10 and Fashion-Mnist.

Fed-SAP not only converges to the highest final accuracy but also demonstrates faster and more stable convergence throughout the entire training process (especially in the early stages). This is attributed to the GSR mechanism providing high-quality, aligned feature statistics to the pSE modules from the very beginning.

IV. CONCLUSION

In this work, we have presented a comprehensive study on the deployment of attention-based lightweight networks in Federated Learning under heterogeneous data distributions. Our empirical analysis uncovers a fundamental architectural conflict: while Convolutional layers extract generalized spatial features suitable for global aggregation, the Excitation layers within Squeeze-and-Excitation (SE) modules develop highly distinct, client-specific channel attention patterns that are detrimental to synchronize.

To resolve this conflict, we propose Fed-SAP, a framework that strategically decouples feature extraction from attention calibration. The framework relies on two complementary mechanisms: Personalized Squeeze-and-Excitation (pSE) and Global Statistics Regularization (GSR). pSE preserves the local specificity of attention modules, preventing the dilution of client-unique knowledge. Concurrently, GSR imposes a statistical constraint on the shared backbone, compelling it to generate globally aligned feature statistics despite local data skew. This regularization effectively mitigates the "input bias" phenomenon, ensuring that personalized attention modules operate on robust, unbiased feature representations. Consequently, the interaction between global alignment and local personalization yields a substantial performance gain that surpasses the capabilities of either component in isolation. Extensive experiments on CIFAR-10 and Fashion-MNIST validate that Fed-SAP significantly outperforms existing state-of-the-art methods, including FedBN, in terms of both convergence stability and final accuracy. By reconciling the tension between generalization and personalization, Fed-SAP provides a theoretically grounded and practically effective solution for deploying intelligent, efficient edge models in privacy-sensitive IoT ecosystems.

REFERENCES

[1] B. McMahan, E. Moore, D. Ramage, et al., "Communication-efficient learning of deep networks from decentralized data," in *Proc. Int. Conf. Artif. Intell. Stat. (AISTATS)*, 2017, pp. 1273–1282.

[2] Karimireddy S P, Kale S, Mohri M, et al. Scaffold: Stochastic controlled averaging for federated learning[C]//International conference on machine learning. PMLR, 2020: 5132-5143.

[3] Hasan M M. Federated Learning Models for Privacy-Preserving AI In Enterprise Decision Systems[J]. International Journal of Business and Economics Insights, 2025, 5(3): 238-269.

[4] Albogami N N. Intelligent deep federated learning model for enhancing security in internet of things enabled edge computing environment[J]. Scientific Reports, 2025, 15(1): 4041.

[5] Thakur D, Guzzo A, Fortino G, et al. Green federated learning: A new era of green aware AI[J]. ACM Computing Surveys, 2025, 57(8): 1-36

[6] Seo J, Catak F O, Rong C. Understanding federated learning from iid to non-iid dataset: An experimental study[J]. arXiv preprint arXiv:2502.00182, 2025.

[7] Salazar T, Gama J, Araujo H, et al. Unveiling group-specific distributed concept drift: A fairness imperative in federated learning[J]. IEEE Transactions on Neural Networks and Learning Systems, 2025.

[8] Li T, Sahu A K, Zaheer M, et al. Federated optimization in heterogeneous networks[J]. Proceedings of Machine learning and systems, 2020, 2: 429-450.

[9] Uddin M P, Xiang Y, Hasan M, et al. A Systematic Literature Review of Robust Federated Learning: Issues, Solutions, and Future Research Directions[J]. ACM Computing Surveys, 2025, 57(10): 1-62.

[10] Kulkarni V, Kulkarni M, Pant A. Survey of personalization techniques for federated learning[C]//2020 fourth world conference on smart trends in systems, security and sustainability (WorldS4). IEEE, 2020: 794-797.

[11] Tan A Z, Yu H, Cui L, et al. Towards personalized federated learning[J]. IEEE transactions on neural networks and learning systems, 2022, 34(12): 9587-9603.

[12] Wang Z, Wang Z, Fan X, et al. Federated Learning with Domain Shift Eraser[C]//Proceedings of the Computer Vision and Pattern Recognition Conference. 2025: 4978-4987.

[13] Li X, Jiang M, Zhang X, et al. Fedbn: Federated learning on non-iid features via local batch normalization[J]. arXiv preprint arXiv:2102.07623, 2021.

[14] Wang Z, Yi F, Gong P, et al. Population Normalization for Federated Learning[C]//Proceedings of the Computer Vision and Pattern Recognition Conference. 2025: 10214-10223.

[15] Howard A, Sandler M, Chu G, et al. Searching for mobilenetv3[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 1314-1324.