Decoupling Label Shift and Surrogate Gradient Errors for Robust Federated Spiking Neural Networks

Zouquan Chen[1,3]* **Yifei Yang**[1,*] **Lidong Zheng**[2]
Zhiming Fang[1†] **Yuchen Zheng**[1,†]
[1]College of Information Science and Technology,
Shihezi University, Shihezi, China
[2]College of Computer Science and Technology,
Nanjing University of Aeronautics and Astronautics, Nanjing, China
[3]China Nanhu Academy of Electronics and Information Technology,
Jiaxing, China
{chenzouquan,yifeiyang}@stu.shzu.edu.cn
{zld0608,ouczyc}@outlook.com
fangzhiming@shzu.edu.cn

Abstract

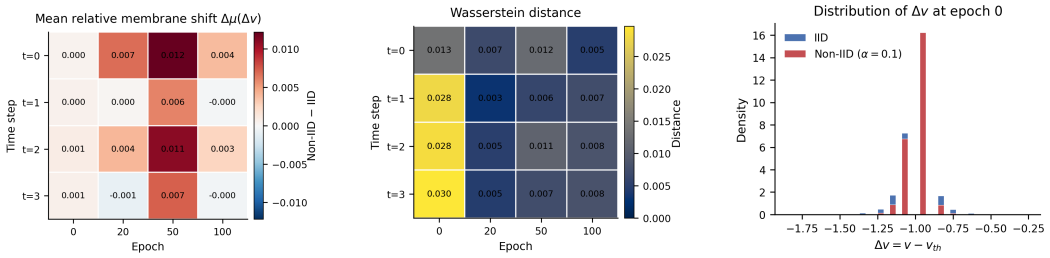
Federated Spiking Neural Networks (FedSNNs) provide an energy-efficient and privacy-preserving learning paradigm for edge intelligence. However, their robustness under Non-IID data is limited by a coupled failure mode: label shift distorts membrane potential distributions, drives neurons away from gradient-sensitive regions, and further amplifies surrogate gradient errors, which destabilizes federated optimization. To address this problem, this paper proposes Spatio-Temporal Spiking Adaptive Thresholds (ST-SAT), a framework that mitigates the coupling between label shift and surrogate gradient errors through coordinated neuron-level adaptation and client–server knowledge alignment. At the neuron level, ST-SAT introduces the Spatio-Temporal Parametric Leaky Integrate-and-Fire (ST-PLIF) neuron, which adaptively calibrates firing thresholds, membrane potentials, and decay dynamics to keep local neuronal responses within effective gradient regions. At the collaborative interaction level, ST-SAT adopts a Dual-Scale Knowledge Distillation strategy that combines adaptive logit-level distillation with feature-level alignment of neuronal statistics, thereby reducing class distribution bias and improving knowledge consistency across heterogeneous clients. Extensive experiments on three benchmark datasets show that ST-SAT consistently improves both global model performance and client-local performance, with particularly strong gains under severe label shift. These results validate the effectiveness of the proposed framework for robust federated SNN learning. The code is available at https://github.com/ZouquanChen/ST-SAT_code.

1 Introduction

Edge intelligence requires learning systems that can operate under strict energy, computation, and privacy constraints. Spiking Neural Networks (SNNs) are well suited to this setting because their event-driven sparse computation can reduce energy consumption on neuromorphic and resource-constrained devices [1, 2]. Federated learning provides a complementary training paradigm that enables multiple clients to collaboratively optimize a model without sharing private local data [3]. The

*Equal contribution.
†Corresponding authors: Yuchen Zheng and Zhiming Fang.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).



(a) Mean shift of Δv. (b) Wasserstein distance. (c) Distribution of Δv at epoch 0.

Figure 1: Differences in membrane states between IID and Non-IID data. Figure illustrate the mean difference of $\Delta v = v - v_{th}$ between Non-IID ($\alpha = 0.1$) and IID data, the Wasserstein distance between the corresponding Δv distributions, and their epoch-0 distribution comparison.

integration of federated learning and SNNs, namely Federated Spiking Neural Networks (FedSNNs), therefore offers a promising route toward privacy-preserving and energy-efficient edge learning [4].

Despite this promise, FedSNNs remain difficult to optimize under realistic Non-IID data. In edge scenarios, client data usually exhibit severe label shift, where each client observes only a biased subset of the global label distribution [5]. This heterogeneity causes client drift in federated optimization and weakens the generalization ability of the aggregated global model. For SNNs, the problem is more severe than in conventional Artificial Neural Networks (ANNs), because spike firing is non-differentiable and training relies on surrogate gradients to approximate the true gradient during backpropagation [6]. When local label distributions are highly skewed, the induced feature shift can further distort membrane potential distributions. As illustrated in Figure 1, compared with IID data, label-skewed Non-IID data ($\alpha = 0.1$) induces measurable differences in the relative membrane potential $\Delta v = v - v_{th}$ at a middle spiking layer. These differences are reflected by the mean shift between Non-IID and IID Δv distributions, the Wasserstein distance between the two distributions, and the direct distributional comparison at epoch 0. This observation suggests that label shift can reshape local neuronal states before multi-client aggregation takes place. As a result, neurons may move away from gradient-sensitive regions, causing surrogate gradient errors to be amplified during local training. This creates a potential coupled failure mode in which label shift and surrogate gradient errors reinforce each other, making FedSNNs optimization more fragile under heterogeneous data.

To address this challenge, we propose Spatio-Temporal Spiking Adaptive Thresholds (ST-SAT), a framework that mitigates the coupling between label shift and surrogate gradient errors through coordinated neuron-level adaptation and client–server knowledge alignment. At the neuron level, ST-SAT introduces the Spatio-Temporal Parametric Leaky Integrate-and-Fire (ST-PLIF) neuron, which adaptively calibrates firing thresholds, membrane potentials, and decay dynamics under biased client data. This design keeps local neuronal responses within effective gradient regions and suppresses the amplification of surrogate gradient errors. At the collaboration level, ST-SAT adopts a Dual-Scale Knowledge Distillation strategy that combines adaptive logit-level distillation with feature-level alignment of neuronal statistics. This strategy reduces class distribution bias and improves knowledge consistency across heterogeneous clients. Our main contributions are summarized as follows:

- We propose Spatio-Temporal Spiking Adaptive Thresholds (ST-SAT), a framework for robust FedSNN learning under label shift. ST-SAT mitigates the coupling between label shift and surrogate gradient errors through spatio-temporal adaptive-threshold neurons and Dual-Scale KD. We further provide theoretical analysis showing that the convergence of FedSNNs is jointly affected by client heterogeneity and surrogate gradient bias.
- We design the Spatio-Temporal Parametric Leaky Integrate-and-Fire (ST-PLIF) neuron, which adaptively calibrates firing thresholds, membrane potentials, and decay dynamics through spatio-temporal regularization. This design keeps local neuronal responses within effective gradient regions and suppresses the amplification of surrogate gradient errors caused by local label shift.
- We introduce a Dual-Scale KD strategy that improves knowledge consistency across heterogeneous clients. This strategy combines adaptive logit-level distillation with feature-level

alignment of neuronal statistics, thereby reducing class distribution bias and mitigating local feature shift.

2 Related Work

Federated learning under Non-IID data has been widely studied, with FedAvg serving as the standard baseline for weighted aggregation of client models [3], although it often suffers from slow convergence and degraded performance under label skew [7]. To reduce client drift, FedProx adds a proximal regularizer to local objectives [8], SCAFFOLD uses control variates to correct update directions [9], and FedNova normalizes local updates to address objective inconsistency [10]. Recent methods further improve robustness through representation learning and knowledge transfer: FedConcat aggregates client-specific feature encoders for multi-source feature learning [11], FLEA combines feature distillation with adaptive local training [5], and FedLEC, the closest work to ours, calibrates missing or sparse labels in FedSNNs by using global predictions as teacher signals through KL divergence [12]. However, these methods mainly address label skew at the aggregation, representation, or prediction level. Even FedLEC, while designed for FedSNNs, does not explicitly model the coupling between label shift and surrogate gradient errors induced by spiking neuronal dynamics. In contrast, ST-SAT targets this FedSNN-specific failure mode by jointly stabilizing local spiking dynamics with ST-PLIF neurons and aligning client–server knowledge through Dual-Scale KD.

3 Preliminaries

3.1 Spiking Neural Networks and Neuron Design

Traditional SNNs simplify neuronal dynamics by using a fixed firing threshold. Although this assumption reduces implementation complexity, it limits the encoding capacity and biological plausibility of spiking neurons, since biological firing thresholds are dynamic and heterogeneous. Prior work has introduced spatio-temporal threshold modulation to improve the flexibility of SNN neurons [13]. We review this formulation as a useful parameterization of membrane modulation and dynamic thresholding, while ST-SAT further redesigns it for the federated setting where label shift and surrogate gradient errors are coupled. Given the membrane potential $u_c[t]$ of channel c at time step t, the spatio-temporal threshold formulation uses a normalization-based attention module to apply channel-wise modulation to the membrane potential. The modulated membrane potential $u_c^*[t]$ is computed as

$$u_c^*[t] = u_c[t] \cdot \sigma_{act}\left(\frac{|\gamma_c|}{\sum_{j=1}^{C}|\gamma_j|} \cdot \left(\frac{u_c[t]-\mu_c}{\sqrt{\sigma_c^2+\epsilon}} \cdot \gamma_c + \beta_c\right)\right), \tag{1}$$

where $u_c^*[t]$ denotes the spatially modulated membrane potential produced by the Normalization-based Attention Module (NAM). $u_c[t]$ is the original membrane potential obtained from the integration phase. $\sigma_{act}(\cdot)$ is the sigmoid activation function that generates gating coefficients in the range $(0, 1)$. γ_c and β_c are the learnable scaling and bias parameters of channel c in the Batch Normalization (BN) layer. μ_c and σ_c^2 denote the mean and variance of the membrane potential of channel c in the current batch. ϵ is a numerical stability constant, and C is the total number of channels in the layer. The term $|\gamma_c|/\sum_{j=1}^{C}|\gamma_j|$ represents the normalized channel-importance weight derived from the BN scaling parameters. The formulation also introduces a time-dependent threshold factor to dynamically adjust the neuronal firing threshold. The threshold at time step t is defined as

$$V_{th}[t] = V_{base} \cdot e^{-\phi_t}, \tag{2}$$

where $V_{th}[t]$ denotes the dynamic firing threshold at time step t, V_{base} is the base threshold, and ϕ_t is a learnable temporal threshold factor that controls the evolution of the threshold over time. In ST-SAT, this dynamic thresholding mechanism is further coupled with learnable membrane decay, federated regularization, and client–server knowledge alignment, so that neuronal dynamics can be stabilized under heterogeneous client data.

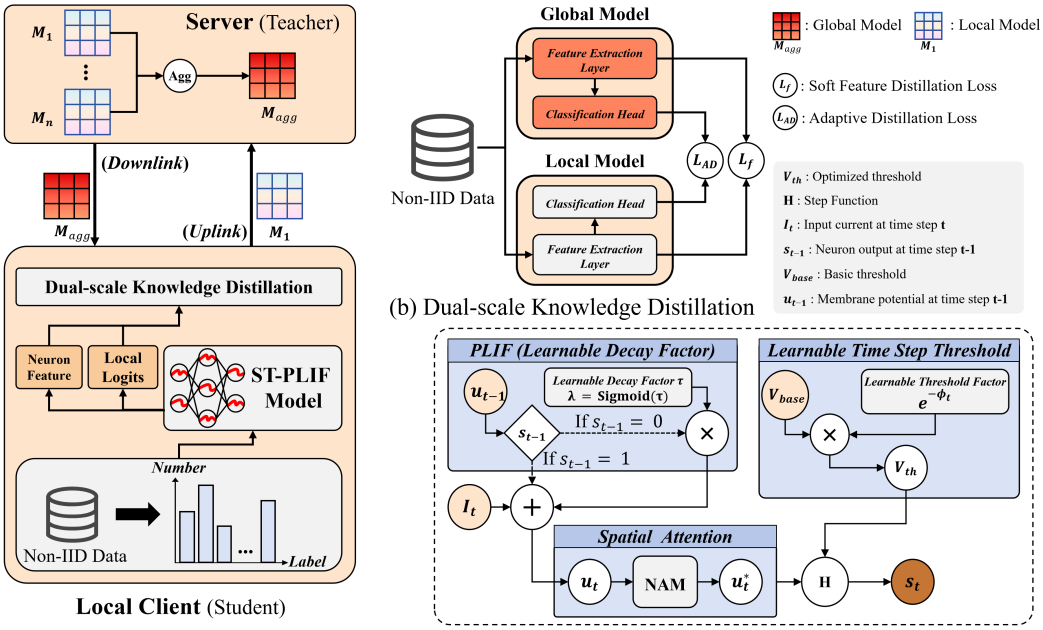


Figure 2: Overall architecture of ST-SAT. The framework consists of the ST-PLIF neuron for stabilizing spiking dynamics under label shift and the Dual-Scale KD mechanism for aligning client-server knowledge at both logit and feature levels.

4 Spatio-Temporal Spiking Adaptive Threshold

This section presents the proposed ST-SAT framework. As shown in Figure 2, ST-SAT contains two complementary components. The first component is the ST-PLIF neuron, which stabilizes local spiking dynamics and suppresses the amplification of surrogate gradient errors under biased client data. The second component is the Dual-Scale KD mechanism, which aligns client–server knowledge at both logit and feature levels to reduce label-shift-induced inconsistency across clients.

4.1 Spatio-Temporal Parametric Leaky Integrate-and-Fire Neuron

ST-PLIF is designed to address the neuron-level instability caused by the coupling between label shift and surrogate gradient errors in FedSNNs. Building on the spatio-temporal threshold parameterization, ST-PLIF introduces channel-wise learnable membrane decay and federated regularization to jointly control membrane integration, threshold adaptation, and temporal decay. This design stabilizes local spiking dynamics under Non-IID label distributions and keeps neuronal responses within effective gradient regions, thereby suppressing the amplification of surrogate gradient errors during local training.

For membrane potential dynamics, ST-PLIF assigns an independent learnable decay factor λ_c to each channel c. The membrane potential at layer l, channel c, and time step t is updated as

$$u_c[t] = u_c[t-1] \cdot \lambda_c \cdot (1 - s_c[t-1]) + I_c[t], \tag{3}$$

where $u_c[t]$ denotes the membrane potential of channel c at time step t, $u_c[t-1]$ is the membrane potential at the previous time step, $s_c[t-1]$ is the output spike at the previous time step with value 0 or 1, and $I_c[t]$ is the synaptic input current at the current time step. The decay factor λ_c controls the temporal retention of the membrane potential for channel c and lies in $(0, 1)$. It is generated from the learnable parameter $\boldsymbol{\tau}_c$ through the sigmoid function,

$$\lambda_c = \frac{1}{1 + e^{-\boldsymbol{\tau}_c}}. \tag{4}$$

After obtaining $u_c[t]$, ST-PLIF computes the modulated potential $u_c^*[t]$ according to Equation (1) and then dynamically calibrates the firing threshold using Equation (2). The spike generation rule is

therefore defined as

$$s_c[t] = H\left(u_c^*[t] - V_{th}[t]\right), \tag{5}$$

where $s_c[t]$ denotes the spike output of channel c at time step t, and $H(\cdot)$ is the Heaviside step function.

To balance neuronal adaptability with generalization stability in federated learning, ST-PLIF introduces a regularization term $\mathcal{L}_{reg}$ that combines temporal dynamic regularization and spatial feature regularization,

$$\mathcal{L}_{reg} = \lambda_{temp}\mathcal{L}_{temp} + \lambda_{nam}\text{Var}(\mathbf{w}_{BN}), \tag{6}$$

where $\mathcal{L}_{reg}$ denotes the joint regularization loss used to stabilize ST-PLIF neurons during local training. The first term, $\mathcal{L}_{temp} = ||\boldsymbol{\phi}||_2^2 + ||\boldsymbol{\tau} - \boldsymbol{\tau}_0||_2^2$, is the temporal dynamic regularization term. Here, $\boldsymbol{\phi}$ is the learnable temporal threshold factor in Equation (2), which controls the time-varying firing threshold, while $||\boldsymbol{\phi}||_2^2$ prevents excessive threshold fluctuations across time steps. $\boldsymbol{\tau}$ denotes the learnable membrane decay parameter in Equation (4), which determines the channel-wise decay factor λ_c, and $\boldsymbol{\tau}_0$ denotes its initial value. The term $||\boldsymbol{\tau} - \boldsymbol{\tau}_0||_2^2$ constrains the learned decay dynamics and prevents them from deviating excessively from the initial membrane integration behavior. The second term, $\text{Var}(\mathbf{w}_{BN})$, denotes the variance of the BN-derived scaling weights $\mathbf{w}_{BN}$ in NAM, which measures the dispersion of channel-wise spatial modulation strengths. Penalizing this variance prevents the model from over-relying on a small number of feature channels and promotes smoother spatial modulation. The coefficients λ_{temp} and λ_{nam} are hyperparameters that control the strengths of temporal dynamic regularization and spatial feature regularization, respectively.

4.2 Dual-Scale Knowledge Distillation

To further address heterogeneity induced by label shift, ST-SAT introduces a downlink knowledge transfer mechanism from the global server to local clients. Prior label calibration methods mainly operate at the prediction level to compensate for the effect of missing classes [12], whereas Dual-Scale KD aligns knowledge from two complementary dimensions: class predictions and spiking-neuron statistics. Specifically, it corrects class-distribution bias through logit-level adaptive distillation and further aligns the feature-level activation-rate statistics of spiking neurons in the last feature extraction layer. This design enables local clients to maintain consistent neuronal response patterns under heterogeneous data.

For class prediction alignment, we introduce an adaptive weighted distillation loss at the logit-output level. This loss assigns different weights to existing and missing classes in the local dataset $\mathcal{D}_k$,

$$\mathcal{L}_{AD} = \frac{1}{N_c}\sum_{n_c=1}^{N_c} \alpha_{n_c} \cdot D_{KL}(P_{global}^{(n_c)} || P_{local}^{(n_c)}), \tag{7}$$

where $\mathcal{L}_{AD}$ denotes the adaptive distillation loss, N_c is the total number of classes, and $D_{KL}(\cdot||\cdot)$ denotes the KL divergence. $P_{global}^{(n_c)}$ and $P_{local}^{(n_c)}$ are the predicted probability distributions of the global and local models for class n_c, respectively. The distillation weight α_{n_c} is determined by the presence of class n_c in the local dataset,

$$\alpha_{n_c} = \begin{cases} 1, & \text{if class } n_c \text{ is absent from } \mathcal{D}_k, \\ \beta, & \text{if class } n_c \text{ is present in } \mathcal{D}_k, \end{cases} \tag{8}$$

where $\beta \in [0, 1)$ controls the distillation strength for classes that are present in the local dataset. This weighting scheme assigns full distillation guidance to locally absent classes, for which the client lacks direct supervision, and assigns a smaller weight to locally present classes, which already receive supervision from local labels. In this way, the loss compensates for missing-class knowledge while avoiding excessive constraint on locally observed classes.

At the feature level, we introduce a soft feature distillation loss tailored to spiking neurons. Since spike outputs are discrete, logit-level distillation alone may overlook differences in internal spiking activity patterns. We therefore align activation-rate statistics extracted from the last spiking layer of the feature extractor, namely the penultimate layer before the classifier, between the local client model and the global server model,

$$\mathcal{L}_{feature} = \frac{1}{B}\sum_{i=1}^{B} ||\boldsymbol{r}_{local}^{(i)} - \boldsymbol{r}_{global}^{(i)}||^2, \tag{9}$$

where $\mathcal{L}_{feature}$ denotes the soft feature distillation loss, and B is the batch size. $\boldsymbol{r}_{local}^{(i)}$ and $\boldsymbol{r}_{global}^{(i)}$ denote the activation-rate vectors of the i-th sample extracted from the last spiking feature layer of the local model and the corresponding layer of the global model, respectively. For the i-th sample, the activation rate $r_{i,d}$ of the d-th neuron is computed by averaging its binary spike outputs over the temporal dimension,

$$r_{i,d} = \frac{1}{T}\sum_{t=1}^{T} s_{t,i,d}, \tag{10}$$

where $s_{t,i,d} \in \{0,1\}$ denotes the spike output of the d-th neuron in the last spiking feature layer at time step t for the i-th sample, and T is the total number of time steps. These activation rates form the feature-level neuronal statistics $\boldsymbol{r}_{local}^{(i)}$ and $\boldsymbol{r}_{global}^{(i)}$, which are used to align the internal spiking representations of the local and global models. This feature-level alignment complements logit-level calibration by constraining neuron response patterns rather than only output predictions, thereby improving knowledge consistency across heterogeneous clients.

5 Error and Convergence Analysis of ST-SAT

We analyze the convergence error of FedSNNs under the joint effects of client heterogeneity and surrogate gradient bias. The analysis first gives an upper-bound decomposition of the optimization error and then connects this bound to the neuron-level error coupling analyzed in Appendix B. We follow several commonly used assumptions in federated learning [14, 8, 9, 15, 16].

Assumption 5.1. (L-Smoothness). The global objective function $F(\mathbf{w})$ is L-smooth. For all $\mathbf{v}, \mathbf{w} \in \mathbb{R}^d$, the following inequality holds:

$$F(\mathbf{v}) \le F(\mathbf{w}) + \langle \nabla F(\mathbf{w}), \mathbf{v} - \mathbf{w} \rangle + \frac{L}{2}\|\mathbf{v} - \mathbf{w}\|^2. \tag{11}$$

Assumption 5.2. (Bounded Variance and Bias of Gradient). For client k, the stochastic surrogate gradient $\mathbf{g}_k$ satisfies

$$\begin{cases} \mathbb{E}[\mathbf{g}_k] = \nabla F_k(\boldsymbol{\theta}) + \boldsymbol{\Delta}_k, \quad \|\boldsymbol{\Delta}_k\| \le \delta \\ \mathbb{E}[\|\mathbf{g}_k - \mathbb{E}[\mathbf{g}_k]\|^2] \le \sigma^2, \end{cases} \tag{12}$$

where $\boldsymbol{\Delta}_k$ denotes the inherent bias introduced by the surrogate gradient approximation at the client, δ represents the upper bound of the surrogate gradient bias, and σ^2 denotes the upper bound of the variance of the stochastic gradient.

Assumption 5.3. (Bounded Gradient). Assume that for any parameter $\boldsymbol{\theta} \in \mathbb{R}^d$, the squared norm of the true gradient of the global objective function is bounded above by a constant G^2:

$$\|\nabla F(\boldsymbol{\theta})\|^2 \le G^2. \tag{13}$$

Assumption 5.4. (Controlled Heterogeneity). The discrepancy between local and global gradients is bounded within a certain range:

$$\|\nabla F_k(\boldsymbol{\theta}) - \nabla F(\boldsymbol{\theta})\|^2 \le \kappa_{KD}^2, \tag{14}$$

where κ_{KD}^2 denotes the maximum discrepancy between local gradients and global gradients.

Under Assumptions 5.1–5.4, we obtain the following theorem.

Theorem 5.5. *The expected gradient norm of FedSNN optimization is bounded as*

$$\min_r \mathbb{E}[\|\nabla F(\boldsymbol{\theta}^r)\|^2] \le \frac{1}{R}\sum_{r=0}^{R-1} \mathbb{E}[\|\nabla F(\boldsymbol{\theta}^r)\|^2] \le \frac{4(F(\boldsymbol{\theta}^0) - F^*)}{\eta E R} + \mathcal{O}(\delta^2) + \mathcal{O}(\eta E(\kappa_{KD}^2 + \sigma^2)). \tag{15}$$

where $\boldsymbol{\theta}^r$ denotes the global model parameters at communication round r, $F(\boldsymbol{\theta}^0)$ is the initial objective value, F^* represents the optimal objective value, η is the learning rate, E is the number of local training epochs per round, R is the total number of communication rounds, δ^2 captures the inherent bias of the surrogate gradient, κ_{KD}^2 is a quantification of data heterogeneity, and σ^2 denotes the upper bound of the variance of the stochastic gradient.

Table 1: Best global model test accuracy (%) of different federated learning methods with the VGG-S backbone on CIFAR-10, CIFAR-100, and CIFAR10-DVS under Dirichlet label skew. Results are reported as mean ± sample standard deviation over the three experiment groups ($n = 3$). The highest mean in each column is shown in bold.

Method	CIFAR-10		CIFAR-100		CIFAR10-DVS	
	$\alpha = 0.1$	$\alpha = 0.5$	$\alpha = 0.1$	$\alpha = 0.5$	$\alpha = 0.1$	$\alpha = 0.5$
FedAvg	42.02 ± 1.81	63.10 ± 0.66	21.87 ± 0.54	36.94 ± 0.53	28.90 ± 5.54	36.43 ± 1.03
FedConcat	43.59 ± 0.94	72.77 ± 1.63	36.05 ± 1.26	43.58 ± 0.53	32.58 ± 0.43	40.00 ± 0.41
FedLEC	50.10 ± 0.29	74.55 ± 0.71	34.13 ± 0.46	48.00 ± 0.54	39.70 ± 1.31	46.92 ± 2.36
FedNova	43.20 ± 0.56	67.18 ± 0.48	16.98 ± 0.10	28.76 ± 0.35	33.93 ± 4.65	45.45 ± 0.28
FedProx	45.33 ± 0.64	70.31 ± 0.65	30.69 ± 0.40	42.38 ± 0.61	29.52 ± 2.36	37.10 ± 2.48
FLEA	52.55 ± 2.23	70.77 ± 0.93	37.30 ± 1.30	46.64 ± 0.47	35.18 ± 0.10	44.30 ± 3.74
SCAFFOLD	44.90 ± 1.27	$\mathbf{77.68 \pm 0.91}$	34.59 ± 0.37	50.83 ± 0.38	30.77 ± 2.08	$\mathbf{50.17 \pm 1.10}$
ST-SAT	$\mathbf{55.61 \pm 2.43}$	76.11 ± 0.15	$\mathbf{42.02 \pm 0.74}$	$\mathbf{52.18 \pm 0.29}$	$\mathbf{41.62 \pm 4.03}$	49.57 ± 1.29

The detailed derivation process is provided in Appendix A. Theorem 5.5 shows that, as $R \to \infty$, the convergence bound approaches an error floor composed of two dominant terms. The first is a controllable term $\mathcal{O}(\eta E(\kappa_{KD}^2 + \sigma^2))$ associated with client heterogeneity and stochastic noise, which can be reduced by decreasing the learning rate. The second is a non-vanishing term $\mathcal{O}(\delta^2)$ caused by the inherent surrogate gradient bias of SNNs. This result identifies the two major contributors to the optimization error of FedSNNs. However, the additive form of the upper bound should not be interpreted as indicating that the two errors are independent in practice. Appendix B provides a neuron-level analysis of this error coupling. Under severe Non-IID label shift, the membrane potential distribution can move toward gradient sign inversion regions or become misaligned with the effective support of the surrogate derivative. Both effects increase the effective surrogate gradient bias δ, which explains why data heterogeneity can amplify surrogate gradient errors in FedSNNs. ST-SAT addresses this coupling by using NAM-based membrane modulation and spatio-temporal adaptive thresholding, thereby reducing the non-vanishing error floor induced by $\mathcal{O}(\delta^2)$.

6 Experiments

6.1 Experimental Setup

We evaluate ST-SAT on CIFAR-10, CIFAR-100, and the neuromorphic CIFAR-10-DVS dataset [17] to cover both static images and event-driven data streams. We compare ST-SAT with representative federated learning baselines, including FedAvg [3], FedConcat [11], FedLEC [12], FedNova [10], FedProx [8], FLEA [5], and SCAFFOLD [9], where FedLEC is the closest FedSNN method for label-shift mitigation. All methods use S-VGG9, denoted as VGG-S [4], as the default backbone for studying label-skew heterogeneity. We further evaluate ST-SAT with stronger SNN backbones, including MS-ResNet [18] and MetaSpikeformer [19], to assess its generalization ability.

More detailed experimental settings and additional results are provided in Appendix C and Appendix D.

6.2 Comparative experiments on federated learning methods

Table 1 reports the best global model test accuracy of different federated learning methods with the VGG-S backbone on CIFAR-10, CIFAR-100, and CIFAR10-DVS under Dirichlet label skew with $\alpha \in \{0.1, 0.5\}$. For each experiment group, we record the highest global model test accuracy attained during training and report the mean and sample standard deviation across the three groups. ST-SAT achieves the highest mean global accuracy in four of the six settings and exceeds FedLEC in all six. Under the more severe label-skew condition of $\alpha = 0.1$, ST-SAT achieves the highest mean on all three datasets. Specifically, it reaches 55.61% and 42.02% on CIFAR-10 and CIFAR-100, exceeding FedLEC by 5.51 and 7.89 percentage points, respectively. On CIFAR10-DVS, ST-SAT

Table 2: Average best client accuracy (%) of different federated learning methods with the VGG-S backbone on CIFAR-10, CIFAR-100, and CIFAR10-DVS under Dirichlet label skew. For each experiment group, the best test accuracy of each client is averaged across clients. Results are reported as mean $\pm$ sample standard deviation over the three experiment groups ($n = 3$). The highest mean in each column is shown in bold.

Method	CIFAR-10		CIFAR-100		CIFAR10-DVS	
	$\alpha = 0.1$	$\alpha = 0.5$	$\alpha = 0.1$	$\alpha = 0.5$	$\alpha = 0.1$	$\alpha = 0.5$
FedAvg	28.81 ± 0.19	49.46 ± 0.28	16.20 ± 0.19	31.81 ± 0.31	21.81 ± 2.82	29.36 ± 0.71
FedLEC	33.33 ± 0.41	56.84 ± 0.68	22.55 ± 0.21	38.83 ± 0.19	25.32 ± 0.76	33.30 ± 1.47
FedNova	28.42 ± 0.54	52.70 ± 0.30	12.05 ± 0.12	21.30 ± 0.36	23.16 ± 2.25	34.25 ± 0.14
FedProx	30.57 ± 0.73	55.66 ± 0.81	18.92 ± 0.13	32.42 ± 0.30	21.50 ± 0.77	27.53 ± 1.67
FLEA	31.86 ± 0.26	56.13 ± 0.37	23.77 ± 0.30	41.23 ± 0.17	24.14 ± 0.84	29.07 ± 3.98
SCAFFOLD	29.84 ± 0.48	60.67 ± 0.54	21.46 ± 0.02	41.32 ± 0.28	22.84 ± 0.87	34.92 ± 0.46
ST-SAT	$\mathbf{38.34 \pm 0.63}$	$\mathbf{64.79 \pm 0.43}$	$\mathbf{32.82 \pm 0.25}$	$\mathbf{46.45 \pm 0.48}$	$\mathbf{27.42 \pm 1.85}$	$\mathbf{38.38 \pm 1.00}$

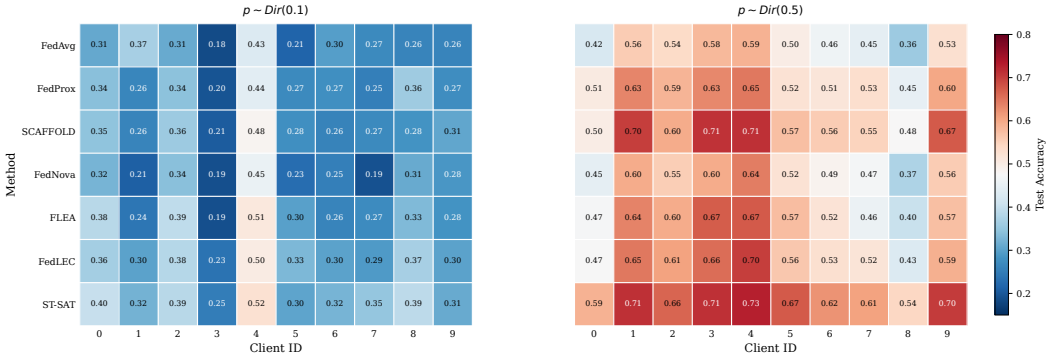
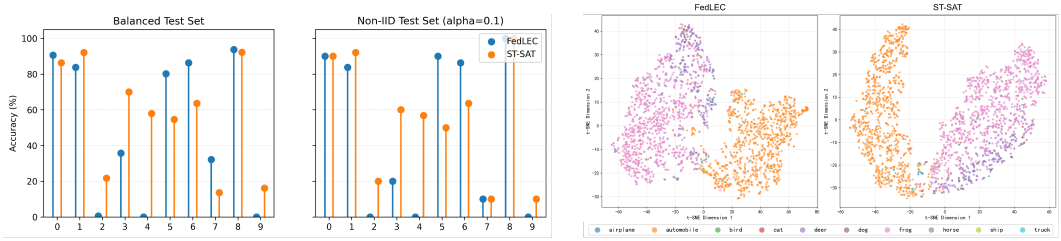


Figure 3: Client test accuracy on CIFAR-10 under Dirichlet label skew with $\alpha = 0.1$ (left) and $\alpha = 0.5$ (right). Results are from a single run for each setting. Accuracy is shown on a scale from 0 to 1.

achieves 41.62% at $\alpha = 0.1$ and 49.57% at $\alpha = 0.5$, exceeding FedLEC by 1.92 and 2.65 percentage points, respectively. SCAFFOLD obtains higher means on CIFAR-10 and CIFAR10-DVS at $\alpha = 0.5$, exceeding ST-SAT by 1.57 and 0.60 percentage points, respectively. These results indicate strong global model performance across the evaluated settings, particularly under more severe label skew.

Table 2 reports average best client accuracy. Within each experiment group, each client model is evaluated on the common global test set after its local update, and the highest accuracy attained by each client during training is averaged across clients. The table reports the mean and sample standard deviation of this metric across the three experiment groups. Among the methods reporting this metric, ST-SAT achieves the highest mean in all six settings. At $\alpha = 0.1$, ST-SAT obtains 38.34% on CIFAR-10 and 32.82% on CIFAR-100, exceeding FedLEC by 5.01 and 10.27 percentage points, respectively. On CIFAR10-DVS, it achieves 27.42% and 38.38% at $\alpha = 0.1$ and $\alpha = 0.5$, corresponding to improvements of 2.10 and 5.08 percentage points over FedLEC. These results indicate that ST-SAT combines competitive global model accuracy with improved average best performance of locally trained models on the common test distribution.

Figure 3 illustrates the client test accuracy for ST-SAT and the baseline methods on the CIFAR-10 dataset. These results indicate that ST-SAT not only improves the performance of the global model but also significantly enhances the adaptability of individual client models when encountering label shift. As illustrated in Figure 4a, we evaluate the per-class accuracy of FedLEC and ST-SAT on the CIFAR-10 dataset under both balanced and Non-IID test set settings ($\alpha = 0.1$). In the balanced setting scenario, FedLEC exhibits severe class forgetting issues, with accuracy for the bird, deer, and truck classes approaching zero. In contrast, ST-SAT effectively mitigates this issue, restoring



(a) Global per-class accuracy.

(b) Global feature separability.

Figure 4: Global model comparison on CIFAR-10 under severe label shift. Compared with FedLEC, ST-SAT improves per-class accuracy and learns more separable global feature representations when $\alpha = 0.1$.

Table 3: Inference SOP comparison on CIFAR-10 under Dirichlet label skew with $\alpha = 0.1$, using the full-hidden-layer SNN implementation, $T = 4$, and training seed 2025. SOPs are estimated per sample over all time steps. All configurations are evaluated on the complete 10,000-image test set.

Configuration	Threshold multiplier	SOPs (M)	Accuracy (%)
FedAvg	–	39.263	44.27
Budget-trained ST-SAT	1.00000	37.489	52.25
Budget-trained ST-SAT	0.98535	39.280	**52.96**

the accuracy of these classes to 21.7%, 58.0%, and 16.2%, respectively, while increasing the overall accuracy from 50.31% to 56.83%. The standard deviation of the inter-class accuracy also decreases from 40.68% to 30.45%, indicating a more balanced performance across classes. In the Non-IID setting, ST-SAT maintains its advantage, improving overall accuracy by 1.06 percentage points compared to FedLEC and continuing to eliminate zero-accuracy classes. These experimental results demonstrate that the spatio-temporal adaptive threshold mechanism of ST-SAT effectively addresses class forgetting and label drift issues in federated SNN learning, achieving more robust and balanced classification performance.

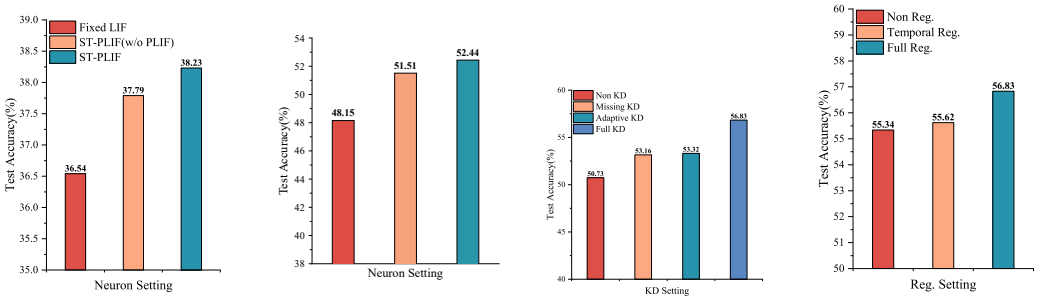
Figure 4b presents t-SNE visualizations of the feature representations learned by FedLEC and ST-SAT on the CIFAR-10 dataset under $\alpha = 0.1$ setting. In the embedding space of FedLEC, the dominant automobile class occupies the right region, with features from other classes severely intermixed, indicating poor inter-class separability. In contrast, ST-SAT generates a more structured feature space, where the frog class forms a distinct cluster on the right, clearly separated from the automobile, and the separation between deer and frog classes is also better than that of FedLEC. This improvement in inter-class separability aligns with the accuracy improvements observed in Figure 4a, where ST-SAT restores classes with near-zero accuracy and achieves more balanced performance across classes.

To examine whether the accuracy improvement can be explained solely by increased spiking computation, we train an additional ST-SAT model with an inference SOP penalty. The target budget is the measured FedAvg SOP count, $B = 39.2625$ million operations per sample. The training objective is

$$\mathcal{L} = \mathcal{L}_{\text{ST-SAT}} + \lambda_r \left[\max\left(0, \frac{S}{B} - 1\right) \right]^2, \tag{16}$$

where S denotes the estimated batch-average SOP count per sample. The coefficient λ_r increases linearly to 0.3 over the first five communication rounds. This soft penalty discourages budget violations without guaranteeing exact budget satisfaction.

As shown in Table 3, the budget-trained ST-SAT checkpoint achieves 52.25% accuracy at its original threshold, exceeding FedAvg by 7.98 percentage points while using approximately 4.52% fewer SOPs. We then keep the checkpoint weights fixed and scan a shared threshold multiplier. At a multiplier of 0.98535, ST-SAT achieves 52.96% accuracy with 39.2802 million SOPs per sample. The relative SOP difference from FedAvg is only 0.045%, while the accuracy improvement is 8.69 percentage points. These results indicate that, in this evaluated setting, the improvement is not explained solely by a larger inference SOP count. Additional gradient diagnostics relative to a Gaussian reference surrogate are reported in Appendix D.2.



(a) Local neuron ablation ($\alpha = 0.1$) (b) Local neuron ablation ($\alpha = 0.5$) (c) Dual-Scale KD (d) Regularization term

Figure 5: Ablations on CIFAR-10 with VGG-S. (a,b) Local neuron ablations without aggregation at $\alpha = 0.1$ and 0.5, respectively; (c) Dual-Scale KD and (d) regularization at $\alpha = 0.1$.

Table 4: Federated ablation on CIFAR-10 ($\alpha = 0.1$), measured by best global test accuracy. ST-PLIF-only includes temporal, decay, and NAM regularization. Best in bold.

ST-PLIF	Dual-Scale KD	Global accuracy (%)
×	×	47.98
✓	×	50.73
×	✓	53.60
✓	✓	**56.83**

6.3 Ablation Study

Figure 5 summarizes the ablation results on CIFAR-10, covering the effects of local neuron strategies under different label-skew settings, as well as the contributions of Dual-Scale KD and the regularization term. Table 4 reports the ablation results of ST-PLIF and Dual-Scale KD on CIFAR-10 under the severe label-skew setting $\alpha = 0.1$. The baseline without either component achieves an accuracy of 47.98%. The ST-PLIF-only variant improves the accuracy to 50.73%, indicating that neuron-level spatio-temporal adaptation helps stabilize local spiking dynamics under biased data. Using Dual-Scale KD alone further increases the accuracy to 53.60%, showing that client–server knowledge alignment effectively mitigates label-shift-induced inconsistency. When both components are combined, ST-SAT achieves the best accuracy of 56.83%, with gains of 8.85, 6.10, and 3.23 percentage points over the baseline, ST-PLIF-only variant, and Dual-Scale-KD-only variant, respectively. These results verify that ST-PLIF and Dual-Scale KD are complementary, jointly reducing the coupled effect of surrogate gradient errors and label shift in FedSNNs. More ablation results can be found in Appendix D.1.

7 Conclusion

In this paper, we propose ST-SAT to mitigate the coupling between label shift and surrogate gradient errors in FedSNNs. ST-SAT combines ST-PLIF neurons, which stabilize local spiking dynamics by calibrating membrane potentials, firing thresholds, and decay factors, with Dual-Scale KD, which aligns client–server knowledge from class predictions and spiking-neuron statistics. Our error and convergence analysis shows that FedSNN optimization is jointly affected by data heterogeneity and surrogate gradient bias, and that ST-SAT reduces the non-vanishing error floor caused by surrogate gradient errors. Experiments on multiple benchmark datasets and component ablations demonstrate that ST-SAT improves both global and client-level performance under severe label shift.

Acknowledgments and Disclosure of Funding

This research received no financial support. The authors declare no competing interests.

References

[1] D. Yu, X. Du, L. Jiang, W. Tong, and S. Deng, "Ec-snn: Splitting deep spiking neural networks for edge devices," in *Proceedings of the International Joint Conference on Artificial Intelligence*, 8 2024, pp. 5389–5397.

[2] W. Maass, "Networks of spiking neurons: The third generation of neural network models," *Neural Networks*, vol. 10, no. 9, pp. 1659–1671, 1997.

[3] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proceedings of Artificial Intelligence and Statistics*, 2017, pp. 1273–1282.

[4] Y. Venkatesha, Y. Kim, L. Tassiulas, and P. Panda, "Federated learning with spiking neural networks," *IEEE Transactions on Signal Processing*, vol. 69, pp. 6183–6194, 2021.

[5] T. Xia, A. Ghosh, X. Qiu, and C. Mascolo, "Flea: Addressing data scarcity and label skew in federated learning via privacy-preserving feature augmentation," in *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2024, pp. 3484–3494.

[6] S. Deng, H. Lin, Y. Li, and S. Gu, "Surrogate module learning: Reduce the gradient error accumulation in training spiking neural networks," in *Proceedings of International Conference on Machine Learning*, 2023, pp. 7645–7657.

[7] A. Z. Tan, H. Yu, L. Cui, and Q. Yang, "Towards personalized federated learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 12, pp. 9587–9603, 2023.

[8] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," in *Proceedings of Machine Learning and Systems*, vol. 2, 2020, pp. 429–450.

[9] S. P. Karimireddy, S. Kale, M. Mohri, S. Reddi, S. Stich, and A. T. Suresh, "Scaffold: Stochastic controlled averaging for federated learning," in *Proceedings of International Conference on Machine Learning*, 2020, pp. 5132–5143.

[10] J. Wang, Q. Liu, H. Liang, G. Joshi, and H. V. Poor, "Tackling the objective inconsistency problem in heterogeneous federated optimization," *Advances in neural information processing systems*, vol. 33, pp. 7611–7623, 2020.

[11] Y. Diao, Q. Li, and B. He, "Exploiting label skews in federated learning with model concatenation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, 2024, pp. 11 784–11 792.

[12] D. Yu, X. Du, L. Jiang, H. Zhang, and S. Deng, "Exploiting label skewness for spiking neural networks in federated learning," in *Proceedings of the International Joint Conference on Artificial Intelligence*, 8 2025, pp. 6895–6903.

[13] J. Fu, S. Gou, P. Wang, L. Jiao, Z. Guo, J. Li, and R. Liu, "Adaptation and learning of spatio-temporal thresholds in spiking neural networks," *Neurocomputing*, vol. 644, p. 130423, 2025.

[14] Y. Tan, G. Long, L. Liu, T. Zhou, Q. Lu, J. Jiang, and C. Zhang, "Fedproto: Federated prototype learning across heterogeneous clients," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, 2022, pp. 8432–8440.

[15] Y. Chan, R. Zhou, R. Zhao, Z. Jiang, and E. C. H. Ngai, "Internal cross-layer gradients for extending homogeneity to heterogeneity in federated learning," in *Proceedings of The International Conference on Learning Representations*, 2024, pp. 1–29.

[16] J. Gygax and F. Zenke, "Elucidating the theoretical underpinnings of surrogate gradient learning in spiking neural networks," *Neural Computation*, vol. 37, no. 5, pp. 886–925, 2025.

[17] H. Li, H. Liu, X. Ji, G. Li, and L. Shi, "CIFAR10-DVS: An event-stream dataset for object classification," *Frontiers in Neuroscience*, vol. 11, 2017. [Online]. Available: https://www.frontiersin.org/journals/neuroscience/articles/10.3389/fnins.2017.00309

[18] Y. Hu, L. Deng, Y. Wu, M. Yao, and G. Li, "Advancing spiking neural networks toward deep residual learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 36, no. 2, pp. 2353–2367, 2025.

[19] M. Yao, J. Hu, T. Hu, Y. Xu, Z. Zhou, Y. Tian, B. XU, and G. Li, "Spike-driven transformer v2: Meta spiking neural network architecture inspiring the design of next-generation neuromorphic chips," in *Proceedings of The International Conference on Learning Representations*, 2024, pp. 1–23.

[20] X. Xie, J. Feng, G. Liu, Q. Zhan, Z. Liu, and M. Zhang, "Federated learning for spiking neural networks by hint-layer knowledge distillation," *Applied Soft Computing*, vol. 163, p. 111901, 2024.

A Convergence Analysis

In this section, we analyze the convergence properties of the proposed ST-SAT framework.

According to Assumption 5.1, expanding the parameter update step $\boldsymbol{\theta}^{r+1} = \boldsymbol{\theta}^r - \eta\tilde{\mathbf{g}}^r$ yields,

$$F(\boldsymbol{\theta}^{r+1}) \leq F(\boldsymbol{\theta}^r) + \langle \nabla F(\boldsymbol{\theta}^r), \boldsymbol{\theta}^{r+1} - \boldsymbol{\theta}^r \rangle + \frac{L}{2}\|\boldsymbol{\theta}^{r+1} - \boldsymbol{\theta}^r\|^2. \tag{17}$$

Substituting $\boldsymbol{\theta}^{r+1} - \boldsymbol{\theta}^r = -\eta\tilde{\mathbf{g}}^r$ and taking the expectation over the stochasticity without considering client drift, we obtain,

$$\mathbb{E}[F(\boldsymbol{\theta}^{r+1})] \leq F(\boldsymbol{\theta}^r) - \eta\langle \nabla F(\boldsymbol{\theta}^r), \mathbb{E}[\tilde{\mathbf{g}}^r] \rangle + \frac{L\eta^2}{2}\mathbb{E}[\|\tilde{\mathbf{g}}^r\|^2], \tag{18}$$

By weighted aggregation across clients, $\mathbb{E}[\tilde{\mathbf{g}}^r] = \sum_{k=1}^{K} p_k \mathbb{E}[\mathbf{g}_k] = \nabla F(\boldsymbol{\theta}^r) + \boldsymbol{\Delta}$, where we define the aggregated surrogate-gradient bias $\boldsymbol{\Delta} := \sum_{k=1}^{K} p_k \boldsymbol{\Delta}_k$ and assume $\|\boldsymbol{\Delta}\| \leq \delta$. where $\boldsymbol{\theta}^r$ and $\boldsymbol{\theta}^{r+1}$ denote the global model parameters at rounds r and $r+1$, respectively, η is the learning rate, and $\tilde{\mathbf{g}}^r$ represents the aggregated stochastic surrogate gradient at round r.

According to Assumption 5.2, substituting $\mathbb{E}[\tilde{\mathbf{g}}^r] = \nabla F(\boldsymbol{\theta}^r) + \boldsymbol{\Delta}$ yields,

$$\begin{aligned} -\eta\langle \nabla F(\boldsymbol{\theta}^r), \mathbb{E}[\tilde{\mathbf{g}}^r] \rangle &= -\eta\langle \nabla F(\boldsymbol{\theta}^r), \nabla F(\boldsymbol{\theta}^r) + \boldsymbol{\Delta} \rangle \\ &= -\eta\|\nabla F(\boldsymbol{\theta}^r)\|^2 - \eta\langle \nabla F(\boldsymbol{\theta}^r), \boldsymbol{\Delta} \rangle, \end{aligned} \tag{19}$$

by applying a weighted Young's inequality, i.e., $ab \leq \frac{a^2}{4} + b^2$, we obtain

$$-\eta\langle \nabla F(\boldsymbol{\theta}^r), \boldsymbol{\Delta} \rangle \leq \frac{\eta}{4}\|\nabla F(\boldsymbol{\theta}^r)\|^2 + \eta\|\boldsymbol{\Delta}\|^2. \tag{20}$$

Based on Assumption 5.2, we have $\|\boldsymbol{\Delta}\|^2 \leq \delta^2$. Therefore,

$$\begin{aligned} -\eta\langle \nabla F(\boldsymbol{\theta}^r), \mathbb{E}[\tilde{\mathbf{g}}^r] \rangle &\leq -\eta\|\nabla F(\boldsymbol{\theta}^r)\|^2 + \frac{\eta}{4}\|\nabla F(\boldsymbol{\theta}^r)\|^2 + \eta\delta^2 \\ &= -\frac{3\eta}{4}\|\nabla F(\boldsymbol{\theta}^r)\|^2 + \eta\delta^2. \end{aligned} \tag{21}$$

As shown in Eq. (21), although gradient descent contributes a descent momentum of $-\frac{3\eta}{4}\|\nabla F(\boldsymbol{\theta}^r)\|^2$, the surrogate gradient error inherent in SNNs introduces a **positive penalty term** $+\eta\delta^2$, thereby impeding the descent.

For the quadratic term $\mathbb{E}[\|\tilde{\mathbf{g}}^r\|^2]$, using the variance formula $\mathbb{E}[X^2] = \text{Var}(X) + (\mathbb{E}[X])^2$, we obtain,

$$\mathbb{E}[\|\tilde{\mathbf{g}}^r\|^2] = \mathbb{E}[\|\tilde{\mathbf{g}}^r - \mathbb{E}[\tilde{\mathbf{g}}^r]\|^2] + \|\mathbb{E}[\tilde{\mathbf{g}}^r]\|^2. \tag{22}$$

Based on Assumption 5.2, we have $\mathbb{E}[\|\tilde{\mathbf{g}}^r - \mathbb{E}[\tilde{\mathbf{g}}^r]\|^2] \leq \sigma^2$ and $\|\mathbb{E}[\tilde{\mathbf{g}}^r]\|^2 = \|\nabla F(\boldsymbol{\theta}^r) + \boldsymbol{\Delta}\|^2$. By applying the inequality $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$, we obtain,

$$\|\mathbb{E}[\tilde{\mathbf{g}}^r]\|^2 = \|\nabla F(\boldsymbol{\theta}^r) + \boldsymbol{\Delta}\|^2 \leq 2\|\nabla F(\boldsymbol{\theta}^r)\|^2 + 2\|\boldsymbol{\Delta}\|^2 \leq 2\|\nabla F(\boldsymbol{\theta}^r)\|^2 + 2\delta^2. \tag{23}$$

Substituting Eq. (23) into Eq. (22) and rearranging yields,

$$\mathbb{E}[\|\tilde{\mathbf{g}}^r\|^2] \leq \sigma^2 + 2\|\nabla F(\boldsymbol{\theta}^r)\|^2 + 2\delta^2. \tag{24}$$

Substituting Eq. (24) and Eq. (21) into Eq. (18), and applying Assumption 5.3, we obtain,

$$\mathbb{E}[F(\boldsymbol{\theta}^{r+1})] \leq F(\boldsymbol{\theta}^{r}) - \frac{3\eta}{4}\|\nabla F(\boldsymbol{\theta}^{r})\|^2 + \eta\delta^2 + L\eta^2 G^2 + L\eta^2\delta^2 + \frac{L\eta^2}{2}\sigma^2. \tag{25}$$

Through Eq. (25), we have derived the upper bound for single-step descent without considering client drift.

However, in federated learning, due to heterogeneous data distributions across clients, local models will gradually deviate from the global optimal update direction after E local updates. To address this, we conduct the following analysis.

Assume that in round r, the initial parameters of client k are initialized as the global parameters $\boldsymbol{\theta}_k^{r,0} = \boldsymbol{\theta}^r$. Performing E local iterations, the update rule at the j-th step is given by,

$$\boldsymbol{\theta}_k^{r,j+1} = \boldsymbol{\theta}_k^{r,j} - \eta \mathbf{g}_k^{r,j}, \tag{26}$$

where $\mathbf{g}_k^{r,j}$ is the stochastic surrogate gradient computed by client k at the j-th local step in round r. After E local updates, the parameters of client k are,

$$\boldsymbol{\theta}_k^{r,E} = \boldsymbol{\theta}^r - \eta \sum_{j=0}^{E-1} \mathbf{g}_k^{r,j}. \tag{27}$$

The expected squared norm of the local training drift is $\mathbb{E}[\|\boldsymbol{\theta}_k^{r,E} - \boldsymbol{\theta}^r\|^2]$. By substituting the update rule into this expression and utilizing the inequality $\|\sum_{i=1}^n x_i\|^2 \leq n\sum_{i=1}^n \|x_i\|^2$, we obtain,

$$\begin{aligned}\mathbb{E}[\|\boldsymbol{\theta}_k^{r,E} - \boldsymbol{\theta}^r\|^2] &= \mathbb{E}[\| - \eta \sum_{j=0}^{E-1} \mathbf{g}_k\|^2] \\ &\leq \eta^2 E \sum_{j=0}^{E-1} \mathbb{E}[\|\mathbf{g}_k\|^2].\end{aligned} \tag{28}$$

Using the formula $\mathbb{E}[X^2] = \text{Var}(X) + (\mathbb{E}[X])^2$, we have,

$$\mathbb{E}[\|\mathbf{g}_k\|^2] \leq \sigma^2 + \|\nabla F_k(\boldsymbol{\theta}_k^{r,j}) + \boldsymbol{\Delta}_k\|^2, \tag{29}$$

by applying the inequality $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$, we obtain

$$\|\nabla F_k(\boldsymbol{\theta}_k^{r,j}) + \boldsymbol{\Delta}_k\|^2 \leq 2\|\nabla F_k(\boldsymbol{\theta}_k^{r,j})\|^2 + 2\delta^2, \tag{30}$$

Let $\nabla F_k(\boldsymbol{\theta}_k^{r,j})$ be denoted as ∇F_k. Based on Assumption 5.4, $\|\nabla F_k - \nabla F\|^2 \leq \kappa_{KD}^2$, and utilizing the inequality $\|\mathbf{a}+\mathbf{b}\|^2 \leq 2\|\mathbf{a}\|^2 + 2\|\mathbf{b}\|^2$, we obtain,

$$\|\nabla F_k\|^2 \leq 2\|\nabla F_k - \nabla F\|^2 + 2\|\nabla F\|^2 \leq 2\kappa_{KD}^2 + 2\|\nabla F\|^2. \tag{31}$$

Based on Assumption 5.3 $\|\nabla F(\boldsymbol{\theta})\|^2 \leq G^2$, substituting Eq. (31) and Eq. (30) into Eq. (29) yields,

$$\mathbb{E}[\|\mathbf{g}_k\|^2] \leq \sigma^2 + 2(2\kappa_{KD}^2 + 2G^2) + 2\delta^2, \tag{32}$$

For Eq. (32), introducing a universal constant $C > 0$, we obtain

$$\mathbb{E}[\|\mathbf{g}_k\|^2] \leq C(\sigma^2 + \kappa_{KD}^2 + \delta^2 + G^2), \tag{33}$$

and, in the sequel, the constant G^2 term is absorbed into the iteration-independent constants. Substituting Eq. (33) into Eq. (28) yields,

$$\begin{aligned}\mathbb{E}[\|\boldsymbol{\theta}_k^{r,E} - \boldsymbol{\theta}^r\|^2] &\leq \eta^2 E \sum_{j=0}^{E-1} \mathcal{O}(\sigma^2 + \kappa_{KD}^2 + \delta^2) \\ &= \eta^2 E \cdot E \cdot \mathcal{O}(\sigma^2 + \kappa_{KD}^2 + \delta^2) \\ &= \mathcal{O}(\eta^2 E^2(\sigma^2 + \kappa_{KD}^2 + \delta^2)).\end{aligned} \tag{34}$$

From Eq. (34), it can be seen that the local model drift is proportional to the square of the number of local updates E^2 and the degree of data heterogeneity κ_{KD}^2.

Given the local model drift, we analyze the one-step descent of the global model while accounting for client drift. Assuming that client k participates in training with probability p_k at round r, the aggregated stochastic surrogate gradient is given by,

$$\boldsymbol{\theta}^{r+1} = \boldsymbol{\theta}^r - \eta \sum_{k=1}^{K} p_k \sum_{j=0}^{E-1} \mathbf{g}_k(\boldsymbol{\theta}_k^{r,j}), \tag{35}$$

where $\mathbf{g}_k(\boldsymbol{\theta}_k^{r,j})$ is the stochastic surrogate gradient computed by client k at the j-th local step in round r. Utilizing L-smoothness expansion, we obtain,

$$F(\boldsymbol{\theta}^{r+1}) \leq F(\boldsymbol{\theta}^r) + \langle \nabla F(\boldsymbol{\theta}^r), \boldsymbol{\theta}^{r+1} - \boldsymbol{\theta}^r \rangle + \frac{L}{2} \|\boldsymbol{\theta}^{r+1} - \boldsymbol{\theta}^r\|^2. \tag{36}$$

Substituting $\boldsymbol{\theta}^{r+1} - \boldsymbol{\theta}^r$ into $\langle \nabla F(\boldsymbol{\theta}^r), \boldsymbol{\theta}^{r+1} - \boldsymbol{\theta}^r \rangle$ and taking the expectation with respect to the randomness, we obtain,

$$\begin{aligned} \langle \nabla F(\boldsymbol{\theta}^r), \boldsymbol{\theta}^{r+1} - \boldsymbol{\theta}^r \rangle &= -\eta \sum_{k=1}^{K} p_k \sum_{j=0}^{E-1} \langle \nabla F(\boldsymbol{\theta}^r), \mathbb{E}[\mathbf{g}_k(\boldsymbol{\theta}_k^{r,j})] \rangle \\ &= -\eta \sum_{k=1}^{K} p_k \sum_{j=0}^{E-1} \langle \nabla F(\boldsymbol{\theta}^r), \nabla F_k(\boldsymbol{\theta}_k^{r,j}) + \boldsymbol{\Delta}_k \rangle. \end{aligned} \tag{37}$$

Expanding $\nabla F_k(\boldsymbol{\theta}_k^{r,j}) = \nabla F(\boldsymbol{\theta}^r) + [\nabla F_k(\boldsymbol{\theta}_k^{r,j}) - \nabla F_k(\boldsymbol{\theta}^r)] + [\nabla F_k(\boldsymbol{\theta}^r) - \nabla F(\boldsymbol{\theta}^r)]$ and substituting into the inner product, we have,

$$\begin{aligned} \langle \nabla F(\boldsymbol{\theta}^r), \nabla F_k(\boldsymbol{\theta}_k^{r,j}) + \boldsymbol{\Delta}_k \rangle &= \|\nabla F(\boldsymbol{\theta}^r)\|^2 + \langle \nabla F(\boldsymbol{\theta}^r), \nabla F_k(\boldsymbol{\theta}_k^{r,j}) - \nabla F_k(\boldsymbol{\theta}^r) \rangle \\ &\quad + \langle \nabla F(\boldsymbol{\theta}^r), \nabla F_k(\boldsymbol{\theta}^r) - \nabla F(\boldsymbol{\theta}^r) \rangle + \langle \nabla F(\boldsymbol{\theta}^r), \boldsymbol{\Delta}_k \rangle. \end{aligned} \tag{38}$$

By the weighted gradient consistency $\sum_{k=1}^{K} p_k \nabla F_k(\boldsymbol{\theta}^r) = \nabla F(\boldsymbol{\theta}^r)$, the third term vanishes after weighted summation. For the bias term, applying Young's inequality $\langle a, b \rangle \leq \frac{\|a\|^2}{4} + \|b\|^2$, we obtain,

$$-\eta E \langle \nabla F(\boldsymbol{\theta}^r), \boldsymbol{\Delta}_k \rangle \leq \frac{\eta E}{4} \|\nabla F(\boldsymbol{\theta}^r)\|^2 + \eta E \|\boldsymbol{\Delta}_k\|^2 \leq \frac{\eta E}{4} \|\nabla F(\boldsymbol{\theta}^r)\|^2 + \eta E \delta^2. \tag{39}$$

For the second term, applying the Cauchy-Schwarz inequality, L-smoothness, and Young's inequality $ab \leq \frac{a^2}{4} + b^2$, we obtain,

$$\begin{aligned} |\langle \nabla F(\boldsymbol{\theta}^r), \nabla F_k(\boldsymbol{\theta}_k^{r,j}) - \nabla F_k(\boldsymbol{\theta}^r) \rangle| &\leq \|\nabla F(\boldsymbol{\theta}^r)\| \cdot L \|\boldsymbol{\theta}_k^{r,j} - \boldsymbol{\theta}^r\| \\ &\leq \frac{\|\nabla F(\boldsymbol{\theta}^r)\|^2}{4} + L^2 \|\boldsymbol{\theta}_k^{r,j} - \boldsymbol{\theta}^r\|^2. \end{aligned} \tag{40}$$

Combining the above results, noting that the main term $-\eta E \|\nabla F(\boldsymbol{\theta}^r)\|^2$ is partially offset by the two $\frac{\eta E}{4} \|\nabla F(\boldsymbol{\theta}^r)\|^2$ terms, we obtain,

$$\langle \nabla F(\boldsymbol{\theta}^r), \boldsymbol{\theta}^{r+1} - \boldsymbol{\theta}^r \rangle \leq -\frac{\eta E}{2} \|\nabla F(\boldsymbol{\theta}^r)\|^2 + \eta E \delta^2 + L^2 \eta \sum_{k=1}^{K} p_k \sum_{j=0}^{E-1} \mathbb{E}[\|\boldsymbol{\theta}_k^{r,j} - \boldsymbol{\theta}^r\|^2]. \tag{41}$$

Based on Eq. (34), we have,

$$\mathbb{E}[\|\boldsymbol{\theta}_k^{r,j} - \boldsymbol{\theta}^r\|^2] \leq C_0 \eta^2 \tau^2 (\sigma^2 + \kappa_{KD}^2 + \delta^2) \tag{42}$$

where C_0 is a constant. Summing from $j = 0$ to $E - 1$ and substituting into the above equation yields,

$$\sum_{j=0}^{E-1} \mathbb{E}[\|\boldsymbol{\theta}_k^{r,j} - \boldsymbol{\theta}^r\|^2] \leq \frac{C_0}{3} \eta^2 E^3 (\sigma^2 + \kappa_{KD}^2 + \delta^2) \tag{43}$$

Therefore,

$$\langle \nabla F(\boldsymbol{\theta}^r), \boldsymbol{\theta}^{r+1} - \boldsymbol{\theta}^r \rangle \leq -\frac{\eta E}{2} \|\nabla F(\boldsymbol{\theta}^r)\|^2 + \eta E \delta^2 + \frac{L^2 C_0 \eta^3 E^3}{3} (\sigma^2 + \kappa_{KD}^2 + \delta^2) \tag{44}$$

For the quadratic term $\|\boldsymbol{\theta}^{r+1} - \boldsymbol{\theta}^r\|^2$, we have,

$$\frac{L}{2}\mathbb{E}\left[\|\boldsymbol{\theta}^{r+1} - \boldsymbol{\theta}^r\|^2\right] = \frac{L\eta^2}{2}\mathbb{E}\left[\left\|\sum_{k=1}^{K} p_k \sum_{j=0}^{E-1} \mathbf{g}_k(\boldsymbol{\theta}_k^{r,j})\right\|^2\right] \tag{45}$$

Since $\sum_{k=1}^{K} p_k = 1$ and $p_k \geq 0$, by applying the inequality $\left\|\sum_{k=1}^{K} p_k \mathbf{A}_k\right\|^2 \leq \sum_{k=1}^{K} p_k \|\mathbf{A}_k\|^2$, we obtain,

$$\frac{L}{2}\mathbb{E}\left[\|\boldsymbol{\theta}^{r+1} - \boldsymbol{\theta}^r\|^2\right] \leq \frac{L\eta^2}{2}\sum_{k=1}^{K} p_k \mathbb{E}\left[\left\|\sum_{j=0}^{E-1} \mathbf{g}_k(\boldsymbol{\theta}_k^{r,j})\right\|^2\right] \tag{46}$$

By applying the inequality $\|\sum_{i=1}^{n} x_i\|^2 \leq n\sum_{i=1}^{n}\|x_i\|^2$, we obtain,

$$\frac{L}{2}\mathbb{E}\left[\|\boldsymbol{\theta}^{r+1} - \boldsymbol{\theta}^r\|^2\right] \leq \frac{L\eta^2 E}{2}\sum_{k=1}^{K} p_k \sum_{j=0}^{E-1} \mathbb{E}\left[\|\mathbf{g}_k(\boldsymbol{\theta}_k^{r,j})\|^2\right] \tag{47}$$

Based on Assumption 5.2 and the identity $\mathbb{E}[X^2] = \text{Var}(X) + (\mathbb{E}[X])^2$, we obtain,

$$\mathbb{E}\left[\|\mathbf{g}_k(\boldsymbol{\theta}_k^{r,j})\|^2\right] \leq \sigma^2 + \left\|\mathbb{E}[\mathbf{g}_k(\boldsymbol{\theta}_k^{r,j})]\right\|^2, \tag{48}$$

Substituting back into Eq. (47), we obtain,

$$\begin{aligned}\frac{L}{2}\mathbb{E}\left[\|\boldsymbol{\theta}^{r+1} - \boldsymbol{\theta}^r\|^2\right] &\leq \frac{L\eta^2 E}{2}\sum_{k=1}^{K} p_k \sum_{j=0}^{E-1} \sigma^2 + \frac{L\eta^2 E}{2}\sum_{k=1}^{K} p_k \sum_{j=0}^{E-1} \left\|\mathbb{E}[\mathbf{g}_k(\boldsymbol{\theta}_k^{r,j})]\right\|^2 \\ &= \frac{L\eta^2 E^2\sigma^2}{2} + \frac{L\eta^2 E}{2}\sum_{k=1}^{K} p_k \sum_{j=0}^{E-1} \left\|\mathbb{E}[\mathbf{g}_k(\boldsymbol{\theta}_k^{r,j})]\right\|^2\end{aligned} \tag{49}$$

Based on Assumption 5.2, it is known that $\mathbb{E}[\mathbf{g}_k] = \nabla F_k(\boldsymbol{\theta}_k^{r,j}) + \boldsymbol{\Delta}_k$. By utilizing the inequality $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$, we obtain,

$$\|\mathbb{E}[\mathbf{g}_k(\boldsymbol{\theta}_k^{r,j})]\|^2 = \left\|\nabla F_k(\boldsymbol{\theta}_k^{r,j}) + \boldsymbol{\Delta}_k\right\|^2 \leq 2\left\|\nabla F_k(\boldsymbol{\theta}_k^{r,j})\right\|^2 + 2\|\boldsymbol{\Delta}_k\|^2 \tag{50}$$

Substituting $\|\boldsymbol{\Delta}_k\| \leq \delta$ yields,

$$\|\mathbb{E}[\mathbf{g}_k(\boldsymbol{\theta}_k^{r,j})]\|^2 \leq 2\left\|\nabla F_k(\boldsymbol{\theta}_k^{r,j})\right\|^2 + 2\delta^2, \tag{51}$$

Based on the corollary of Assumption 5.1, $\|\nabla F(\mathbf{x})\|^2 \leq 2\|\nabla F(\mathbf{y})\|^2 + 2L^2\|\mathbf{x}-\mathbf{y}\|^2$, we obtain,

$$\left\|\nabla F_k(\boldsymbol{\theta}_k^{r,j})\right\|^2 \leq 2\left\|\nabla F_k(\boldsymbol{\theta}^r)\right\|^2 + 2L^2\left\|\boldsymbol{\theta}_k^{r,j} - \boldsymbol{\theta}^r\right\|^2 \tag{52}$$

Substituting back into Eq. (51), we obtain,

$$\|\mathbb{E}[\mathbf{g}_k(\boldsymbol{\theta}_k^{r,j})]\|^2 \leq 4\|\nabla F_k(\boldsymbol{\theta}^r)\|^2 + 4L^2\|\boldsymbol{\theta}_k^{r,j} - \boldsymbol{\theta}^r\|^2 + 2\delta^2 \tag{53}$$

Based on Assumption 5.4, given $\|\nabla F_k(\mathbf{w})\|^2 \leq 2\|\nabla F(\mathbf{w})\|^2 + 2\kappa_{KD}^2$, we obtain

$$\|\mathbb{E}[\mathbf{g}_k]\|^2 \leq 8\|\nabla F(\boldsymbol{\theta}^r)\|^2 + 8\kappa_{KD}^2 + 4L^2\|\boldsymbol{\theta}_k^{r,j} - \boldsymbol{\theta}^r\|^2 + 2\delta^2 \tag{54}$$

Substituting back into Eq. (49), we obtain,

$$\frac{L\eta^2 E}{2}\sum_{k=1}^{K} p_k \sum_{j=0}^{E-1} \left\|\mathbb{E}[\mathbf{g}_k(\boldsymbol{\theta}_k^{r,j})]\right\|^2 \leq \frac{L\eta^2 E}{2}\sum_{k=1}^{K} p_k \sum_{j=0}^{E-1} \left(8\|\nabla F(\boldsymbol{\theta}^r)\|^2 + 8\kappa_{KD}^2 + 2\delta^2 + 4L^2\|\boldsymbol{\theta}_k^{r,j} - \boldsymbol{\theta}^r\|^2\right) \tag{55}$$

Expanding the summation terms in the above equation. Note that $\sum_{k=1}^{K} p_k = 1$ and $\sum_{j=0}^{E-1} 1 = E$.

$$\begin{aligned}\frac{L\eta^2 E}{2}\sum_{k=1}^{K} p_k \sum_{j=0}^{E-1} \left\|\mathbb{E}[\mathbf{g}_k(\boldsymbol{\theta}_k^{r,j})]\right\|^2 \leq 4L\eta^2E^2\|\nabla F(\boldsymbol{\theta}^r)\|^2 + 4L\eta^2E^2\kappa_{KD}^2 + L\eta^2E^2\delta^2 \\ + 2L^3\eta^2 E\sum_{k=1}^{K} p_k \sum_{j=0}^{E-1} \|\boldsymbol{\theta}_k^{r,j} - \boldsymbol{\theta}^r\|^2\end{aligned} \tag{56}$$

Substituting Eq. (56) into Eq. (49), i.e., adding the noise term $\frac{L\eta^2E^2\sigma^2}{2}$, we obtain the complete upper bound for the quadratic term:

$$\begin{aligned}\frac{L}{2}\mathbb{E}\left[\|\boldsymbol{\theta}^{r+1} - \boldsymbol{\theta}^r\|^2\right] \leq \frac{L\eta^2E^2\sigma^2}{2} + 4L\eta^2E^2\|\nabla F(\boldsymbol{\theta}^r)\|^2 + 4L\eta^2E^2\kappa_{KD}^2 + L\eta^2E^2\delta^2 \\ + 2L^3\eta^2 E\sum_{k=1}^{K} p_k \sum_{j=0}^{E-1} \mathbb{E}[\|\boldsymbol{\theta}_k^{r,j} - \boldsymbol{\theta}^r\|^2]\end{aligned} \tag{57}$$

Now, we combine the general form of the linear term shown in Eq. (44), which has not yet substituted specific drift constants, with the bound of the quadratic term shown in Eq. (57).

$$\begin{aligned}\mathbb{E}[F(\boldsymbol{\theta}^{r+1})] \leq F(\boldsymbol{\theta}^r) - \frac{\eta E}{2}\|\nabla F(\boldsymbol{\theta}^r)\|^2 + 4L\eta^2E^2\|\nabla F(\boldsymbol{\theta}^r)\|^2 \\ + \eta E\delta^2 + L\eta^2E^2\delta^2 + \frac{L\eta^2E^2}{2}\sigma^2 + 4L\eta^2E^2\kappa_{KD}^2 \\ + \frac{L^2C_0\eta^3E^3}{3}(\sigma^2 + \kappa_{KD}^2 + \delta^2) + 2L^3\eta^2E\sum_{k=1}^{K} p_k \sum_{j=0}^{E-1} \mathbb{E}[\|\boldsymbol{\theta}_k^{r,j} - \boldsymbol{\theta}^r\|^2]\end{aligned} \tag{58}$$

Rearranging the gradient term coefficient to $-\frac{\eta E}{2}(1 - 8L\eta E)$. Simultaneously, substituting the drift upper bound from the lemma $\sum_\tau \mathbb{E}[\|\text{drift}\|^2] \leq \frac{C_0}{3}\eta^2E^3(\sigma^2 + \kappa_{KD}^2 + \delta^2)$ into the total drift penalty term.

$$\begin{aligned}\mathbb{E}[F(\boldsymbol{\theta}^{r+1})] \leq F(\boldsymbol{\theta}^r) - \frac{\eta E}{2}(1 - 8L\eta E)\|\nabla F(\boldsymbol{\theta}^r)\|^2 \\ + \eta E\delta^2 + \frac{L\eta^2E^2}{2}\sigma^2 + 4L\eta^2E^2\kappa_{KD}^2 + L\eta^2E^2\delta^2 \\ + L^2\eta(1 + 2L\eta E)\cdot\frac{C_0}{3}\eta^2E^3(\sigma^2 + \kappa_{KD}^2 + \delta^2)\end{aligned} \tag{59}$$

Assuming the learning rate satisfies $\eta \leq \frac{1}{16LE}$, we have $1 - 8L\eta E \geq \frac{1}{2}$, which allows us to bound the descent term by $-\frac{\eta E}{4}\|\nabla F(\boldsymbol{\theta}^r)\|^2$. We then classify the error terms by their order of η. We define K_1 to absorb the coefficients for heterogeneous drift, respectively. Retaining the dominant order term $\mathcal{O}(\eta)$ from the bias, and merging the higher-order $\mathcal{O}(\eta^2)$ terms (heterogeneity and noise), we obtain,

$$\mathbb{E}[F(\boldsymbol{\theta}^{r+1})] \leq F(\boldsymbol{\theta}^r) - \frac{\eta E}{4}\|\nabla F(\boldsymbol{\theta}^r)\|^2 + \eta E\delta^2 + \mathcal{O}(\eta^2E^2\kappa_{KD}^2) + \mathcal{O}(\eta^2E^2\sigma^2) \tag{60}$$

Summing from $r = 0$ to $R-1$ and utilizing the telescoping sum property $\sum(F(\boldsymbol{\theta}^r) - \mathbb{E}[F(\boldsymbol{\theta}^{r+1})]) \leq F(\boldsymbol{\theta}^0) - F^*$.

$$\sum_{r=0}^{R-1}\frac{\eta E}{4}\|\nabla F(\boldsymbol{\theta}^r)\|^2 \leq F(\boldsymbol{\theta}^0) - F^* + R\left(\eta E\delta^2 + C_3\eta^2E^2\kappa_{KD}^2 + C_1\eta^2E^2\sigma^2\right) \tag{61}$$

Dividing both sides by $\frac{\eta ER}{4}$, we obtain the final convergence rate for the average gradient norm,

$$\begin{aligned}\min_r \mathbb{E}[\|\nabla F(\boldsymbol{\theta}^r)\|^2] \leq \frac{1}{R}\sum_{r=0}^{R-1}\mathbb{E}[\|\nabla F(\boldsymbol{\theta}^r)\|^2] \leq \frac{4(F(\boldsymbol{\theta}^0) - F^*)}{\eta ER} + 4\delta^2 + K_1\eta E\kappa_{KD}^2 + \mathcal{O}(\eta E\sigma^2) \\ = \frac{4(F(\boldsymbol{\theta}^0) - F^*)}{\eta ER} + \mathcal{O}(\delta^2) + \mathcal{O}(\eta E(\kappa_{KD}^2 + \sigma^2))\end{aligned} \tag{62}$$

As $R \to \infty$, the convergence bound converges to an error floor. Specifically, the error is composed of a controllable term related to data heterogeneity and noise $\mathcal{O}(\eta E(\kappa_{KD}^2 + \sigma^2))$, which can be reduced by the learning rate, and a non-vanishing constant term derived from the inherent SNN surrogate gradient bias $\mathcal{O}(\delta^2)$, which persists independently of the learning rate.

B Theoretical Analysis of ST-SAT in Mitigating Surrogate Gradient Bias

In this section, we conduct a rigorous mathematical analysis of the inherent surrogate gradient bias term δ introduced in Theorem 5.5, grounded in recent theoretical foundations of Surrogate Gradients (SG) and their specific equivalence to Smoothed Stochastic Derivatives [16]. We demonstrate that the ST-SAT framework significantly lowers the upper bound of δ by leveraging the Normalization-based Attention Module (NAM) and the spatio-temporal adaptive threshold mechanism to enforce Domain Constraint and Support Alignment, respectively.

Based on existing research [16], for stochastic spiking neurons characterized by escape noise, the surrogate gradient is mathematically equivalent to the derivative of the expected output of the neuron. For a neuron in the l-th layer, let the membrane potential be a random variable u with a firing threshold V_{th}. The stochasticity of the membrane potential arises from the input data distribution and the noise of the system, with its probability density function (PDF) denoted as $p(u)$. Consequently, the expectation of the surrogate gradient $\mathbf{g}$ is expressed as the convolution integral of the Jacobian term and a gating factor,

$$\mathbb{E}[\mathbf{g}] = \int_{-\infty}^{+\infty} \frac{\partial \mathcal{L}}{\partial S} \frac{\partial u}{\partial \boldsymbol{\theta}} \cdot \sigma'_\beta(u - V_{th}) \cdot p(u)\, du, \tag{63}$$

where $\mathbf{J}(u) = \frac{\partial \mathcal{L}}{\partial S} \frac{\partial u}{\partial \boldsymbol{\theta}}$ denotes the Jacobian term along the backpropagation path, and $\sigma'_\beta(\cdot)$ represents the derivative of the surrogate function, typically taking the form of a Sigmoid derivative. This term acts as a gradient gate, ensuring that gradients propagate exclusively within the effective region near the threshold. Meanwhile, in Non-IID Federated Learning scenarios, data heterogeneity across clients k induces a severe covariate shift in the membrane potential distribution $p_k(u)$. This shift gives rise to two primary forms of mathematical bias between $\mathbb{E}[\mathbf{g}]$ and the true optimization direction ∇F.

First, we define the gradient sign inversion region Ω_{flip} as the membrane potential interval where the Jacobian term $\mathbf{J}(u)$ opposes the direction of the true gradient ∇F,

$$\Omega_{flip} = \left\{ u \in \mathbb{R} \mid \langle \mathbf{J}(u) \cdot \sigma'_\beta(u - V_{th}), \nabla F \rangle < 0 \right\}. \tag{64}$$

When the probability mass of $p(u)$ is concentrated within Ω_{flip}, the negative contributions within the integral in Eq. (63) dominate the result, thereby inducing a directional bias δ_{sign}.

Second, we define the ϵ-effective support of the surrogate derivative as $\mathcal{S}_\sigma = \{x \mid \sigma'_\beta(x) \geq \epsilon\}$ and the effective support of the membrane potential distribution as $\mathcal{R}_p = \{u \mid p(u) \geq \epsilon_p\}$. When these regions are disjoint, a condition known as support mismatch, the integral kernel vanishes over the entire domain, resulting in a magnitude bias of $\delta_{vanish} \approx \|\nabla F\|$.

The ST-PLIF neuron incorporates a Normalization-based Attention Module (NAM) designed to rectify the distributional shape of $p(u)$ via affine transformations, thereby minimizing δ_{sign}. Let $p_{base}(u)$ denote the raw, unprocessed membrane potential distribution, with μ_{base} and σ_{base}^2 representing its first moment and second central moment, respectively. In severe Non-IID scenarios, μ_{base} may undergo a significant shift, causing the bulk of the distribution to fall within the sign inversion region Ω_{flip},

$$\mathcal{P}_{error} = \int_{\Omega_{flip}} p_{base}(u)\, du \gg 0, \tag{65}$$

where $u \in \mathbb{R}$ denotes the stochastic membrane potential variable governed by input distributions, $p_{base}(u)$ represents the uncalibrated probability density function heavily distorted by local data heterogeneity, $\Omega_{flip} = \{u \mid \langle \mathbf{g}, \nabla F \rangle < 0\}$ defines the gradient sign inversion region where the surrogate gradient opposes the true optimization direction, and the inequality $\gg 0$ signifies that a substantial proportion of neuronal states resides within this failure region, thereby dominating the expectation integral and causing optimization divergence.

NAM performs the following normalization operation on the membrane potential,

$$u^* = \frac{u - \mu_{base}}{\sqrt{\sigma_{base}^2 + \epsilon}} \cdot \gamma + \beta. \tag{66}$$

According to the Change of Variables Theorem, the transformed random variable u^* follows a new probability density function $p_{nam}(u^*)$. Assuming that $\gamma \approx 1$ and $\beta \approx 0$ at initialization, the statistical moments of u^* are constrained to,

$$\mathbb{E}[u^*] \approx 0, \quad \text{Var}(u^*) \approx 1. \tag{67}$$

We leverage the symmetry of the surrogate derivative function $\sigma'_\beta(x)$. Typically, $\sigma'_\beta(x)$ takes the form of a bell-shaped function symmetric about $x = 0$. Consequently, the gradient fidelity region $\Omega_{safe} = \mathbb{R} \setminus \Omega_{flip}$ generally encompasses the central area of the surrogate function, corresponding to the linear activation region. Since $p_{nam}(u^*)$ is re-centered around zero, we invoke Chebyshev's Inequality to establish an upper bound on the probability of the distribution falling within the error region Ω_{flip}. Assuming that Ω_{flip} corresponds to the region lying k standard deviations away from the mean, defined by $|u^*| \geq k$, we obtain,

$$P_{nam}(u^* \in \Omega_{flip}) \leq \frac{\text{Var}(u^*)}{k^2} \approx \frac{1}{k^2}. \tag{68}$$

In contrast, for the original distribution $p_{base}(u)$, if the mean μ_{base} itself falls within Ω_{flip}, the probability $P_{base}(u \in \Omega_{flip})$ may approach 1. Consequently, NAM establishes Sign Concordance via Domain Constraint,

$$\left\| \int_{\Omega_{flip}} \mathbf{J}(u^*) \sigma'_\beta(u^*) p_{nam}(u^*) du^* \right\| \ll \left\| \int_{\Omega_{flip}} \mathbf{J}(u) \sigma'_\beta(u) p_{base}(u) du \right\|. \tag{69}$$

Even with the sign issue resolved, the bias δ remains substantial should the integral vanish. The spatio-temporal adaptive threshold $V_{th}[t]$ addresses this challenge by aligning the supports of the integral kernel. According to Eq. (63), the magnitude of the surrogate gradient $\mathbf{g}$ depends on the degree of overlap between the surrogate derivative and the membrane potential distribution. By applying the Mean Value Theorem for Integrals, we isolate the average magnitude of the Jacobian term, denoted as $\|\bar{\mathbf{J}}\|$, and derive the following approximation,

$$\|\mathbb{E}[\mathbf{g}]\| = \left\| \int_{-\infty}^{+\infty} \mathbf{J}(u) \sigma'_\beta(u - V_{th}) p(u)\, du \right\| \approx \|\bar{\mathbf{J}}\| \cdot \int_{-\infty}^{+\infty} \sigma'_\beta(u - V_{th}) \cdot p(u)\, du. \tag{70}$$

In baseline methods employing a fixed threshold V_{base}, when Non-IID conditions induce a shift in $p(u)$ such that the deviation $|\mu_u - V_{base}|$ exceeds the sum of the effective widths of the two distributions, the intersection of their effective supports becomes virtually empty,

$$\mathcal{S}_\sigma \cap \mathcal{R}_p \approx \emptyset, \tag{71}$$

In this scenario, the aforementioned overlap integral term approaches zero, precipitating gradient vanishing. Consequently, the bias δ converges to $\|\nabla F\|$.

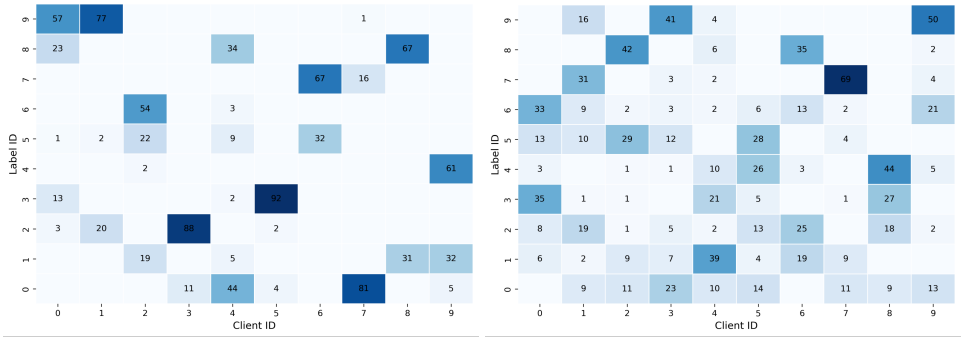
ST-SAT introduces a dynamic threshold defined as $V_{th}[t] = V_{base} \cdot e^{-\phi_t}$. This represents a learnable process in which ϕ_t serves as a time-varying parameter. During backpropagation, to minimize the loss function $\mathcal{L}$, the optimization algorithm inherently seeks paths characterized by non-zero gradients. According to the chain rule,

$$\frac{\partial \mathcal{L}}{\partial \phi_t} = \frac{\partial \mathcal{L}}{\partial S} \frac{\partial S}{\partial V_{th}} \frac{\partial V_{th}}{\partial \phi_t} \propto \sigma'_\beta(u - V_{th}[t]). \tag{72}$$

This implies that the parameter ϕ_t receives valid update gradients exclusively when $V_{th}[t]$ shifts into the effective support of $p(u)$, defined as the region populated by membrane potentials. Consequently, the training process implicitly drives $V_{th}[t]$ to track the distributional mean $\mathbb{E}[u]$.

As $V_{th}[t] \to \mathbb{E}[u]$, the centers of the two bell-shaped functions, $\sigma'_\beta(\cdot - V_{th})$ and $p(u)$, align. According to a corollary of the Cauchy-Schwarz inequality, the overlap integral attains its maximum under this condition, guaranteeing the existence of a lower bound $\xi > 0$,

$$\left\| \int \mathbf{J}(u) \sigma'_\beta(u - V_{th}[t]) p(u)\, du \right\| \geq \xi \|\bar{\mathbf{J}}\| > 0. \tag{73}$$



(a) $\alpha = 0.1$ (b) $\alpha = 0.5$

Figure 6: Illustration of the dataset partition. Each square represents the percentage of samples for a specific label held by a client relative to their total local data volume. (a) The extreme Non-IID setting with $\alpha = 0.1$, where most clients possess scarce samples for the majority of classes. (b) The moderate Non-IID setting with $\alpha = 0.5$.

This proves that the adaptive threshold mechanism eliminates the bias δ_{vanish} induced by gradient vanishing via support alignment.

Synthesizing the derivations above, the total surrogate gradient bias δ_{ST} in the ST-SAT framework can be decomposed into the sign bias suppressed by NAM and the magnitude bias rectified by the adaptive threshold. Compared with fixed-threshold baselines, we obtain the following inequality,

$$\delta_{ST} \approx \delta_{sign}(\text{NAM}) + \delta_{vanish}(\text{Adaptive}) \ll \delta_{Base}. \tag{74}$$

Substituting this result into the convergence upper bound formula in Theorem 5.5, we observe that ST-SAT theoretically guarantees a lower convergence error bound in data-heterogeneous environments by significantly reducing the constant term $\mathcal{O}(\delta^2)$. This conclusion aligns with existing frameworks, which assert that surrogate gradients must match the neuron's escape noise to yield accurate gradients.

C Experimental configuration

In this section, we provide detailed configurations and hyperparameter settings for the experiments.

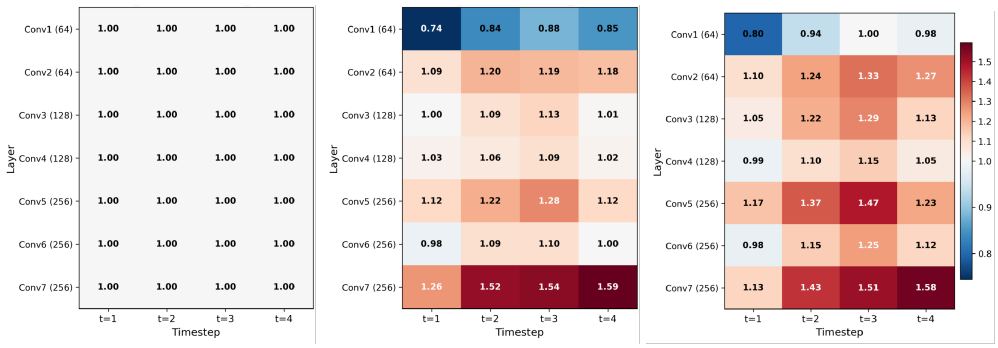
C.1 Dataset and Model Architecture

We conducted experiments on the CIFAR-10 and CIFAR-100 datasets, as well as the event-based CIFAR10-DVS dataset. Regarding the division of the dataset, as illustrated in Figure 6, we employed two strategies to simulate real-world statistical heterogeneity among clients, introducing varying degrees of label distribution skew. Specifically, we allocated training data among clients based on a latent Dirichlet allocation (LDA) on labels, controlled by the concentration parameter α. We set $\alpha = 0.5$ to simulate moderate heterogeneity and $\alpha = 0.1$ to simulate extreme label skew. The latter creates a highly imbalanced scenario where most classes have very few samples on clients, serving as the primary testing ground for our abnormal neuron hypothesis. Simultaneously, for experiments restricting client data label types, we limited each client to possess only a fixed number of specific class samples that do not overlap, constructing an extreme Non-IID environment with severe label absence.

For the model architecture, we followed the FedLEC method setup, using S-VGG9 (**VGG-S**) as the default backbone model for all federated learning algorithms to investigate the impact of different label-skewed data heterogeneity [4]. We also conducted extensive experiments using more complex SNN backbone models such as MS-ResNet (**ResNet**) [18] and MetaSpikeformer (**ViT**) [19] to evaluate the generalization capability of the proposed ST-SAT method.

C.2 Hyperparameter Settings

To validate the effectiveness of the proposed ST-SAT method, we conducted extensive comparative experiments on the CIFAR-10, CIFAR-100, and neuromorphic dataset CIFAR-10-DVS using SNN



(a) Fixed Threshold (b) ST-PLIF(w/o PLIF) (c) ST-PLIF

Figure 7: Results of different neuron strategies on the threshold values using the VGG-S model on the CIFAR-10 dataset under the setting of $\alpha = 0.1$.

models with a simulation time step of $T = 4$. The comparative experiments simulated a federated learning environment with $K = 10$ clients, randomly selecting 20% of clients to participate in training each round. Local training used the Adam optimizer by default, with a learning rate of $\eta = 10^{-3}$, weight decay of 10^{-5}, batch size of $B = 64$, and a total of 50 rounds of global communication, each round consisting of 10 local epochs. But on the CIFAR-10-DVS dataset, due to GPU memory limitations, the batch size was set to 16. We compared ST-SAT with current mainstream strategies, where baseline methods retained their standard optimal hyperparameter settings.

C.3 Evaluation metrics

We primarily used the test accuracy of the global model throughout the complete communication process as the evaluation metric to assess the performance of different federated learning methods under various data heterogeneity scenarios. Additionally, we also focused on the test accuracy of participating clients' local models on the common global test set to evaluate the ability of the proposed method to enhance individual client performance while protecting data privacy.

D Other Experimental Results

In this section, we provide supplementary experimental results to further validate the effectiveness of the proposed ST-SAT method under different data heterogeneity scenarios.

D.1 More ablation experiments

First, we compare the complete ST-PLIF neurons with fixed threshold LIF (Fixed LIF) neurons and spatio-temporal threshold-only LIF (ST-PLIF (w/o PLIF)) neurons. All other training settings remain unchanged to ensure a fair comparison.

As illustrated in Figure 7, we further analyze the evolution of threshold parameters under different neuron strategies. In the fixed threshold LIF neuron, the threshold remains constant, as depicted in Figure 7a. However, in the spatio-temporal threshold LIF neuron, the threshold exhibits significant dynamic adjustment capabilities, allowing for adaptive changes based on the spatio-temporal features of the input data, as illustrated in Figure 7b. Figure 7c illustrates that within the complete ST-PLIF neuron, the threshold not only exhibits spatiotemporal self-adaptability but also reflects more complex dynamic behaviors. Specifically, the gap between the threshold values in the middle region of the figure and those at the edges is significantly more pronounced. This complex dynamic adjustment mechanism enables neurons to respond more effectively to diverse input patterns, thereby enhancing the overall model's expressive power and adaptability.

Table 5: Gradient diagnostics on CIFAR-10 with the full-hidden-layer SNN implementation. Deviations are measured relative to a Gaussian reference surrogate with $\sigma = 0.2$, while preserving the hard forward computation. Results are reported as mean $\pm$ standard deviation across three diagnostic partition seeds using fixed trained checkpoints. Lower values indicate closer agreement with the selected reference.

Partition	Method	Relative ℓ_2 deviation $\downarrow$	Cosine error $\downarrow$
IID	FedAvg	1.0289 ± 0.0165	0.5226 ± 0.0244
IID	ST-SAT	$\mathbf{0.1636 \pm 0.0033}$	$\mathbf{0.0132 \pm 0.0006}$
Non-IID	FedAvg	0.7571 ± 0.0198	0.3316 ± 0.0296
Non-IID	ST-SAT	$\mathbf{0.0670 \pm 0.0008}$	$\mathbf{0.00238 \pm 0.00011}$

Table 6: Comparison with representative FedSNN methods on CIFAR-10 under Dirichlet label skew. The best results are shown in bold.

Method	CIFAR-10 ($\alpha = 0.1$)	CIFAR-10 ($\alpha = 0.5$)
Yeshwanth Venkatesha et al. [4]	42.22	64.33
Di Yu et al. [12]	50.31	74.60
Xiurui Xie et al. [20]	48.00	68.33
Ours	**56.83**	**76.25**

D.2 Gradient diagnostics against a reference surrogate

Table 5 examines the agreement between the gradients obtained using each model's native surrogate and those obtained using a Gaussian reference surrogate. We evaluate the trained FedAvg and ST-SAT checkpoints under both IID and Dirichlet label-skew conditions, with $\alpha = 0.1$ for the Non-IID setting. The diagnostic uses three partition seeds and 128 sampled training examples per client across ten clients. The native and reference computations produce identical logits for all evaluated samples, isolating the effect of the backward surrogate within each checkpoint. ST-SAT exhibits smaller relative ℓ_2 deviations and cosine errors under both partitions. In the Non-IID setting, the relative ℓ_2 deviation decreases from 0.7571 to 0.0670, and the cosine error decreases from 0.3316 to 0.00238. These results show closer gradient agreement with the selected reference under this diagnostic. The reference is an analysis tool rather than a ground-truth derivative of the hard spiking function, and the reported variability reflects diagnostic partitioning rather than independent training runs.

D.3 More comparative analysis of additional FedSNNs baseline methods

Table 6 compares ST-SAT with representative FedSNN methods on CIFAR-10 under two label-skew settings. ST-SAT achieves the best performance in both cases, reaching 56.83% under the severe Non-IID setting ($\alpha = 0.1$) and 76.25% under the milder setting ($\alpha = 0.5$). Compared with the strongest baseline, Di Yu et al., ST-SAT improves accuracy by 6.52 and 1.65 percentage points under $\alpha = 0.1$ and $\alpha = 0.5$, respectively. These results indicate that ST-SAT provides stronger robustness than existing FedSNN methods, especially under severe label shift.

D.4 Experiment on Computational Consumption under Different Neuron Configurations

To analyze the computational cost of different neuron configurations, we measure the synaptic operations (SOPs) defined as $\text{SOP} = \sum_l N_l^{\text{spike}} \times F_l^{\text{out}}$, where N_l^{spike} is the number of spikes in layer l and F_l^{out} is the fan-out of that layer.

Table 7 presents the SOPs comparison across three neuron configurations under the `local_only` strategy. The baseline LIF neuron exhibits the lowest SOPs (6.06M per sample) with a sparse spike rate of 1.09%, but its fixed threshold limits representational capacity. When augmented with Learnable Temporal Factor (LTF) and Normalization-based Attention Module (NAM), the ST-PLIF (w/o PLIF) configuration increases the spike rate to 17.43%, resulting in significantly higher SOPs (219.69M).

Table 7: Ablation results of Synaptic Operations (SOPs) for different neuron configurations. We evaluate a single client on the CIFAR-10 dataset after one local training epoch.

Configuration	SOPs/Sample (M)	Spike Rate (%)
LIF	6.06	1.09
ST-PLIF (w/o PLIF)	219.69	17.43
ST-PLIF	184.91	15.51

Table 8: Communication and forward-computation costs on CIFAR-10 ($\alpha = 0.1$) with VGG-S, using seed 2026, $T = 4$, 50 communication rounds, and 5 local epochs per round. SOPs and spike rates are averaged over rounds 46–50. Communication is estimated in MiB per round.

Method	Params (M)	Comm./round (MiB)	SOPs/sample (M)	Spike rate (%)
FedAvg	5.94	90.69	55.22	4.86
FedProx	5.94	90.69	44.60	4.15
SCAFFOLD	5.94	181.38	81.54	7.10
FedNova	5.94	90.69	41.54	4.28
FedLEC	5.94	90.69	58.12	4.96
ST-SAT	5.95	90.86	107.13	10.17

The complete ST-PLIF model, which additionally incorporates the learnable decay factor (PLIF), achieves a 15.8% reduction in SOPs compared to the ablated version (184.91M vs. 219.69M) with a corresponding decrease in spike rate (15.51% vs. 17.43%). This demonstrates that the learnable decay mechanism enables the network to discover more efficient sparse representations during training, effectively reducing the computational overhead introduced by the enhanced temporal dynamics.

D.5 Comparison of communication overhead and computational cost across different federated learning methods

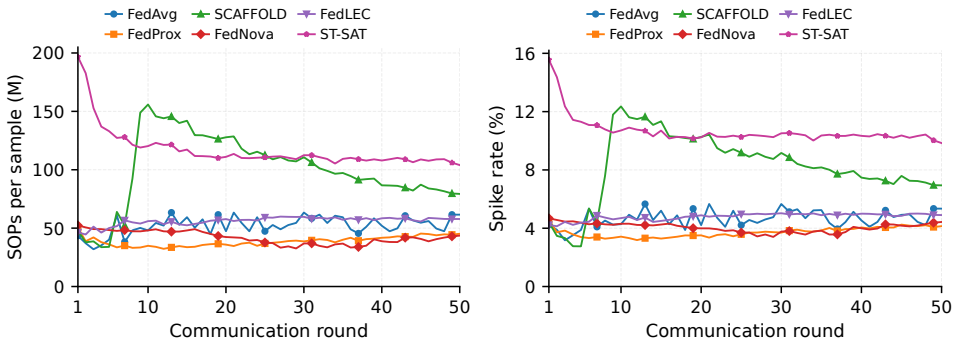
Table 8 and Figure 8 report a dedicated cost experiment with ten clients and a participation rate of 20%. Each method is run once with seed 2026, using five local epochs per round rather than the default ten in Appendix C. After local training, each participating client's model is evaluated by forward passes on its own training data. SOPs are normalized per sample over all $T = 4$ time steps, and the client-level SOPs and spike rates are averaged equally across the two participants. Figure 8 shows all 50 rounds; Table 8 averages rounds 46–50. These metrics exclude backpropagation and distillation-teacher costs and do not represent cumulative training computation or the inference cost of the best global checkpoint.

Communication is estimated from the uploaded and downloaded model states, including batch-normalization buffers, for both participating clients, with additional control-variable transfers for SCAFFOLD and normalization scalars for FedNova ($1\,\mathrm{MiB} = 2^{20}$ bytes). ST-SAT requires 90.86 MiB per round, approximately 0.18% more than FedAvg (90.69 MiB), while SCAFFOLD requires 181.38 MiB. These values describe estimated payload sizes rather than measured network traffic.

FedNova has the lowest forward SOP count in this comparison (41.54 M per sample). ST-SAT uses 107.13 M SOPs per sample with a spike rate of 10.17%, compared with 55.22 M and 4.86% for FedAvg and 81.54 M and 7.10% for SCAFFOLD. Thus, ST-SAT incurs approximately $1.94\times$ the forward SOPs of FedAvg despite its small additional communication payload. The reproduced results show a trade-off between communication and local forward computation; they do not establish an advantage in total runtime or measured energy consumption.

D.6 Experiment on Quantity-based Label Skew Data Heterogeneity

We evaluate the performance of the proposed ST-SAT method on the CIFAR-10 dataset using the VGG-S model under quantity-based label skew scenarios. To simulate extreme Non-IID environments, we restrict each client to hold samples from a fixed number of mutually exclusive classes ($C \in \{2, 3, 4\}$). Table 9 summarizes the average performance. The results indicate that ST-SAT achieves the best



(a) SOPs across communication rounds (b) Spike rate across communication rounds

Figure 8: Per-round SOPs and spike rates from the same runs as Table 8. Each point averages the two participating clients after local training; the table summarizes the last five rounds.

Table 9: Client test accuracy (%) on the CIFAR-10 dataset under quantity-based label skew with VGG-S. C denotes the number of classes per client. The best results are in **Bold** and the second best are underlined.

Method	$C = 2$	$C = 3$	$C = 4$
FedAvg	18.34	25.64	33.85
FedProx	19.02	27.45	37.33
SCAFFOLD	19.17	28.42	38.20
FedNova	18.73	27.60	36.62
FedLEC	19.21	28.92	38.87
ST-SAT	**19.75**	**31.09**	**42.30**

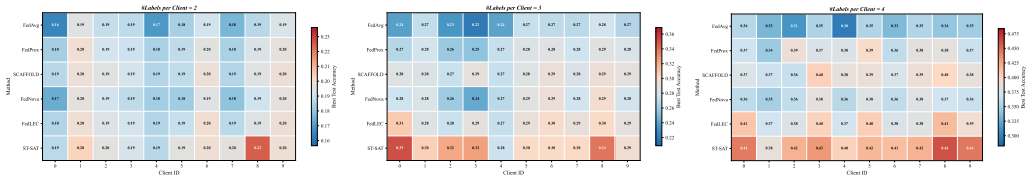
performance across all settings. As shown in Table 9, ST-SAT consistently outperforms the second-best method, FedLEC, improving average accuracy by 0.54, 2.17, and 3.43 percentage points for the $C = 2, 3, 4$ settings, respectively. This advantage is further corroborated by the visualizations in Fig. 9, which fully validate the efficacy of ST-SAT in handling extreme label absence and data heterogeneity.

D.7 Experiment on Different Model Architectures

Table 10 presents the average client test accuracy on the CIFAR-10 dataset using ResNet and ViT model configurations under different Non-IID settings. ST-SAT outperforms existing state-of-the-art baseline methods in most scenarios (3 out of 4 settings), particularly demonstrating significant advantages in the ViT configuration. In the highly Non-IID setting ($\alpha = 0.1$), ST-SAT surpasses the second-best method, FLEA, by a notable margin of 3.80 percentage points on ViT. Furthermore, in the milder Non-IID setting ($\alpha = 0.5$), ST-SAT continues to showcase its effectiveness, outperforming the runner-up SCAFFOLD by 5.14 percentage points on ViT and FLEA by 0.50 percentage points on ResNet. These results further validate the robustness and generalization capability of ST-SAT, especially in handling Transformer-based architectures under Non-IID data.

D.8 Federated Learning Hyperparameter Ablation Study

Figure 12 presents the sensitivity analysis of federated learning hyperparameters on CIFAR-10 using VGG-S under $\alpha = 0.1$. (a) For client participation rate, the optimal accuracy of 59.90% is achieved at a rate of 0.6, with performance declining at both lower and higher rates. (b) For local training epochs, the best accuracy of 56.83% is obtained with 10 epochs, where fewer epochs lead to underfitting and excessive epochs cause local overfitting. (c) For the number of clients, accuracy improves from 56.02% to 61.63% as the client count increases from 5 to 30, indicating that more clients contribute to greater data diversity.



(a) $C = 2$ (b) $C = 3$ (c) $C = 4$

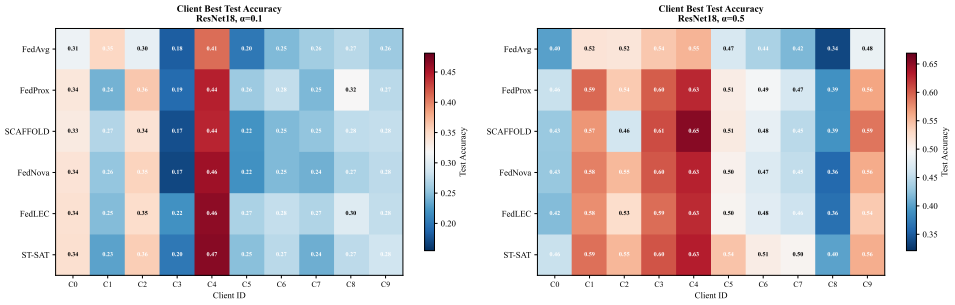
Figure 9: Client test accuracy (%) of the VGG-S model on the CIFAR-10 dataset under quantity-based label shift. C denotes the number of classes per client.

Table 10: Client test accuracy (%) under different non-iid settings with ResNet and ViT model configurations on the CIFAR-10 dataset. The best results are shown in **Bold**, and the second-best results are shown with underlines.

Methods	**ResNet**		**ViT**	
	$\alpha = 0.1$	$\alpha = 0.5$	$\alpha = 0.1$	$\alpha = 0.5$
FedAvg	28.02	46.76	30.27	48.32
FedProx	29.53	52.37	30.47	54.84
SCAFFOLD	28.38	51.47	21.68	60.51
FedNova	28.23	51.42	30.48	55.95
FLEA	29.63	53.00	32.56	57.34
FedLEC	**30.10**	50.77	32.16	56.36
ST-SAT	29.23	**53.50**	**36.36**	**65.65**

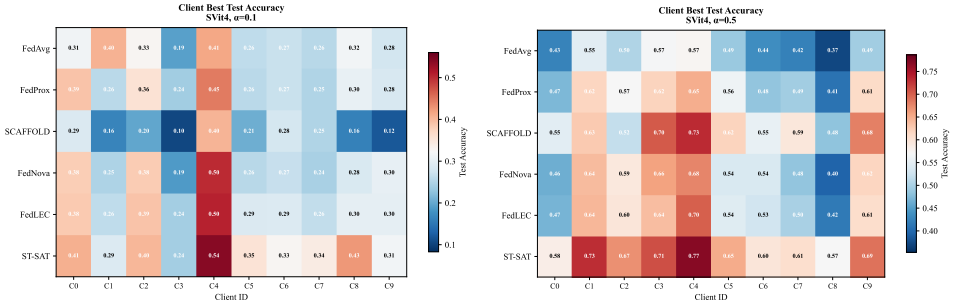
D.9 More Visualization Results

In this subsection, we will present more visualization results.



(a) $\alpha = 0.1$ (b) $\alpha = 0.5$

Figure 10: Client test accuracy (%) of ResNet on the CIFAR-10 dataset under different non-iid settings.



(a) $\alpha = 0.1$ (b) $\alpha = 0.5$

Figure 11: Client test accuracy (%) of ViT on the CIFAR-10 dataset under different non-iid settings.

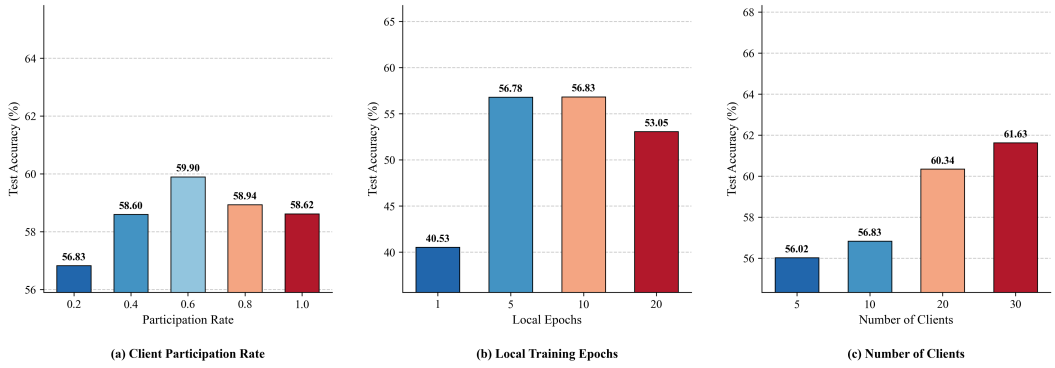


Figure 12: Sensitivity analysis of federated learning hyperparameters on the CIFAR-10 dataset with $\alpha = 0.1$ using VGG-S. (a) Client participation rate. (b) Local training epochs per round. (c) Number of clients.

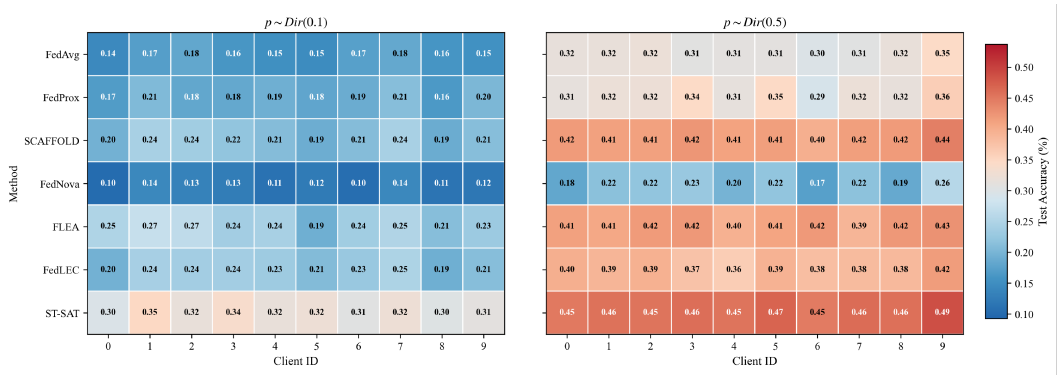


Figure 13: Client test accuracy (%) on the CIFAR-100 dataset under different data heterogeneity settings.

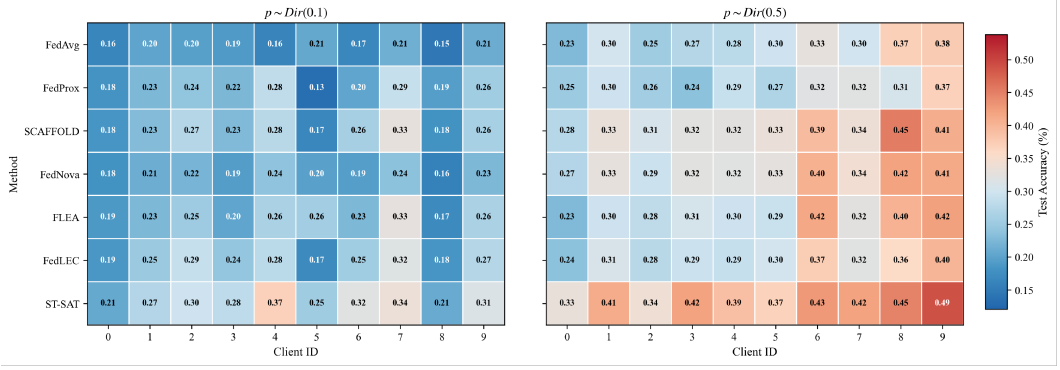


Figure 14: Client test accuracy (%) on the CIFAR10-DVS dataset under different data heterogeneity settings.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state that ST-SAT addresses the coupling between label shift and surrogate gradient errors in FedSNNs through ST-PLIF and Dual-Scale KD. These claims are supported by the theoretical analysis and experiments on CIFAR-10, CIFAR-100, and CIFAR10-DVS.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses the scope of the method through experimental settings, computational cost analysis, and additional results in the appendix. The evaluation is limited to common vision and neuromorphic benchmarks, and broader deployment settings are left for future work.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The paper states the assumptions for smoothness, bounded gradient variance and bias, bounded gradients, and controlled heterogeneity before presenting the convergence theorem. The full derivation is provided in Appendix A, and the neuron-level coupling analysis is provided in Appendix B.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper reports datasets, model architectures, baseline methods, label-skew settings, optimizer, learning rate, batch size, communication rounds, local epochs, and client participation protocol. Additional experimental configurations are provided in Appendix C.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The paper provides a public code link in the abstract.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper describes the Dirichlet label-skew protocol, client number, client participation rate, optimizer, learning rate, weight decay, batch size, simulation time steps, communication rounds, and local epochs. Detailed settings are included in the experimental setup and appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We follow the experimental design of the FedLEC [12] method, present the main comparison results in the form of accuracy values, and attach the mean and standard deviation for some additional client-level experiments.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: We report the algorithm-level computational cost through SOPs, spike rate, and communication overhead, which reflects the main efficiency characteristics of FedSNN training. However, we do not provide a complete per-experiment breakdown of hardware type, memory usage, wall-clock runtime, or total compute, because our experiments were run under shared computing resources.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work uses public benchmark datasets and proposes a general federated SNN training method. No human-subject data, sensitive personal data, or high-risk deployment is involved.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.

- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper states in the conclusion that it aims to advance robust and energy efficient FedSNN learning for edge intelligence. ST-SAT may have positive impacts on privacy-preserving distributed learning and low-power neuromorphic computing. The potential risks are limited because this work is methodological and is evaluated only on public benchmark datasets.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release high-risk pretrained models, generative models, scraped datasets, or assets with a clear dual-use risk. Therefore, special safeguards are not applicable.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: This paper cites the public datasets and baseline methods used in the experiments. Relevant citations are also added to the public code repository.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: This paper releases the publicly available code of ST-SAT and provides the experimental details required to reproduce the main experimental results. The code package comes with relevant documentation and operation instructions.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing experiments or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.

- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve human subjects, so IRB approval or equivalent review is not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method development, experiments, and analysis do not involve LLMs as an important, original, or non-standard research component.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.