# FedVSSAM: Mitigating Flatness Incompatibility in Sharpness-Aware Federated Learning

**Bingnan Xiao***
Key Laboratory of EMW Information (MoE)
College of Future Information Technology
Fudan University
Shanghai 200433, China
22110720061@m.fudan.edu.cn

**Yuan Gao***
Key Laboratory of EMW Information (MoE)
College of Future Information Technology
Fudan University
Shanghai 200433, China
23110720072@m.fudan.edu.cn

**Bingcong Li**†
Department of Computer Science
ETH Zurich
8092 Zürich, Switzerland
bingcong.li@inf.ethz.ch

**Wei Ni**
School of Engineering
Edith Cowan University
Perth, WA 6027, Australia
wei.ni@ieee.org

**Xin Wang**†
Key Laboratory of EMW Information (MoE)
College of Future Information Technology
Fudan University
Shanghai 200433, China
xwang11@fudan.edu.cn

**Tony Quek**
Information Systems Technology and Design Pillar
Singapore University of Technology and Design
Singapore 487372
tonyquek@sutd.edu.sg

## Abstract

Sharpness-aware minimization (SAM) is an effective method for improving the generalization of federated learning (FL) by steering local training toward flat minima. Under data heterogeneity, however, device-side SAM searches for locally flat basins that are incompatible with the flat region preferred by the global objective. We identify this structural failure mode as *flatness incompatibility*, which explains why improving local flatness alone may provide limited training and generalization improvement for the global model. We reveal that flatness incompatibility arises from data heterogeneity and the friendly adversary phenomenon, and is further amplified by local updates and partial device participation. To mitigate this issue, we propose Federated Learning with variance-suppressed sharpness-aware minimization (FedVSSAM), which constructs a variance-suppressed adjusted direction and uses it consistently in local flatness search, local descent, and global update. FedVSSAM anchors both perturbation and update directions to a more stable global direction, instead of correcting only an isolated local perturbation. We establish non-convex convergence guarantees of FedVSSAM and prove that the mean-square deviation between the adjusted direction and the global gradient is effectively controlled. Experiments demonstrate that FedVSSAM mitigates flatness incompatibility and outperforms the baselines across diverse FL settings.

## 1 Introduction

Federated learning (FL) enables devices to train collaboratively while keeping personal data local to protect privacy [41, 55]. In a classical FL paradigm such as FedAvg [43], selected devices perform

*These authors contributed equally to this work.
†Corresponding authors: Bingcong Li and Xin Wang.

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

multiple local updates and then send their models to the parameter server (PS) for aggregation. This communication-efficient training paradigm, however, is highly sensitive to data heterogeneity, i.e., non-independent and identically distributed (non-IID) data across devices. Under non-IID data, local objectives lead different devices toward different regions of the parameter space. The resulting device drift leads to divergence between local and global updates, thereby slowing down convergence [38]. Moreover, increasing the number of local training iterations, which are used to reduce communication load, exacerbates the divergence [33].

Improving generalization is another central objective of FL, since the global model should perform well across different device distributions. Sharpness-aware minimization (SAM) is a natural tool for this purpose. Instead of solely minimizing the loss at the current parameter, SAM seeks locally flat solutions by constructing an adversarial perturbation within a neighborhood and updating the model with the gradient evaluated at the perturbed parameters [14]. Following this principle, FedSAM applies SAM to local training to improve the generalization of the global model under FL [5].

However, the inherent challenge of SAM-based FL is not merely applying SAM to local objectives. Due to data heterogeneity, the flat region searched by a device can be substantially incompatible with the flat region required by the aggregated global model. This local–global inconsistency is severe because the SAM perturbation direction determines both the neighborhood used for local flatness search and the point where the subsequent descent gradient is evaluated. Once the local perturbation is misaligned with the global loss landscape, the error is propagated from perturbation construction to perturbed local descent and is accumulated through aggregation. Therefore, improving local flatness alone may still drive the aggregated model away from a globally flat basin, especially under strong data heterogeneity and multiple local training iterations [30, 48].

We term this structural failure mode *flatness incompatibility*, which is induced by two sources: (i) data heterogeneity, which makes local gradients and local flat regions deviate from the global loss landscape; and (ii) the friendly adversary phenomenon, where a perturbation that is adversarial to one batch may become benign to another [31]. These two sources do not enter SAM-based FL as isolated errors. Data heterogeneity biases the local gradient and local flatness search, while friendly adversary makes the perturbation direction sensitive to the particular batch data. The resulting mismatch affects both the perturbation direction and the point where the descent gradient is evaluated. As training progresses, such mismatch can accumulate and enter the aggregated update. Thus, correcting only one training stage reduces part of the error, but could not keep the local flatness search, local update, and global update aligned. This motivates using a shared and stable direction throughout training to stabilize flatness search and model updates under flatness incompatibility.

This paper proposes Federated Learning with variance-suppressed sharpness-aware minimization (FedVSSAM) to improve global generalization and model training of SAM-based FL under data heterogeneity. The key idea is to construct a variance-suppressed adjusted direction and use it in local flatness search, local update, and global update consistently, to anchor the perturbation and update directions to a more stable global direction. Our key contributions are summarized as follows:

- We first identify flatness incompatibility in SAM-based FL and explicitly reveal its two sources, i.e., data heterogeneity and friendly adversary. This theoretical characterization turns the generalization issue of SAM-based FL into a concrete consistency problem among local flatness search, local update, and global update.
- To address flatness incompatibility, we develop FedVSSAM to provide a shared and stable direction across local flatness search, local update, and global update, thereby mitigating flatness incompatibility and promoting consistent loss landscapes across devices.
- We derive non-convex convergence guarantees for FedVSSAM, accounting for flatness incompatibility and partial device participation, and quantify how the designed variance suppression operations tighten the MSE between the adjusted direction and the global gradient. Our analysis reveals that FedVSSAM achieves a superior convergence by only transmitting updated models to the PS with no additional auxiliary vectors for the uplink transmission, e.g., control or dual variables, compared with existing SAM-based FL algorithms.
- Extensive experiments validate that FedVSSAM can effectively combat flatness incompatibility and accelerate system convergence under data heterogeneity and partial device participation. It is also shown that the designed variance suppression contributes positive performance gains and that FedVSSAM remains robust to key system hyperparameters.

### 1.1 Related work

Inherent data heterogeneity in FL leads to device drifts and impedes training [18, 21, 63]. FedProx, SCAFFOLD, and FedDyn mitigate device drifts via proximal penalties, control variates, or dynamic regularization [1, 22, 35]. FedCM and FedACG further stabilize local and global updates through momentum-based corrections [24, 58]. Such methods mainly reduce optimization bias induced by data heterogeneity, but do not explicitly target FL generalization from a flatness perspective.

To enhance generalization, SAM seeks flat minima by minimizing a worst-case neighborhood loss [14], which is motivated by previous generalization studies [17, 20, 23]. Later variants analyze SAM's performance or improved perturbation strategy [2, 10, 29]. Friendly adversary shows batch-based perturbations can lose adversariality [31]. However, these works focus on centralized optimization and cannot address perturbation inconsistency in heterogeneous SAM-based FL.

In SAM-based FL, FedSAM first employs SAM to improve FL generalization [5]. MoFedSAM extends FedSAM with momentum-based stabilization, while FedGAMMA introduces global sharpness-aware correction with SCAFFOLD [9, 45]. FedSMOO and FedLESAM pursue consistent flatness via dynamic regularization or global perturbations [13, 49]. However, the aforementioned flatness incompatibility exists throughout the SAM-based FL training stage and needs to be processed coherently. Please refer to Appendix A for a detailed related work summary.

## 2 System model and preliminaries

### 2.1 SAM-based federated learning

Consider a classical FL framework comprising a PS, and $N$ devices collected by $\mathcal{N} \triangleq \{1, \cdots, N\}$. The devices train collaboratively a model $\theta \in \mathbb{R}^d$ to minimize the following finite sum problems:

$$\min_{\theta \in \mathbb{R}^d} F(\theta) = \frac{1}{N} \sum_{i \in \mathcal{N}} F_i(\theta),\ F_i(\theta) \triangleq \mathbb{E}_{\phi_i \sim \mathcal{D}_i} \left[ F_i\left(\theta; \phi_i\right) \right], \tag{1}$$

where $F : \mathbb{R}^d \to \mathbb{R}$ is the global objective function at the BS, $F_i$ is the empirical loss of the $i$-th device, and $\phi_i$ denotes a random batch of data sampled from the local dataset $\mathcal{D}_i$ of the $i$-th device, with distributions differing across devices due to data heterogeneity.

Instead of minimizing $F(\theta)$ only at a single model parameter vector, SAM [14] seeks a model parameter vector whose loss remains small under small perturbations of the model in a centralized setting. This favors flat basins of $F(\cdot)$, where the objective varies mildly within a neighborhood of the model, and the resulting model typically has better generalization. As such, the following minimax problem is formulated:

$$\min_{\theta} \max_{\|\epsilon\| \leq \rho} F(\theta + \epsilon), \tag{2}$$

where $\rho$ denotes the perturbation radius, and $\epsilon$ represents SAM's perturbation. By adopting the SAM optimizer at each device in a distributed manner, FedSAM aims to improve local flatness and the generalization of the global model in SAM-based FL. The objective of FedSAM is formulated as

$$\min_{\theta} \max_{\|\epsilon_i\| \leq \rho} \left\{ f(\tilde{\theta}) = \frac{1}{N} \sum_{i \in \mathcal{N}} f_i(\tilde{\theta}) \right\}, \tag{3}$$

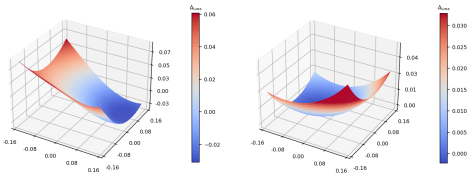
where $f(\tilde{\theta}) \triangleq \max_{\|\epsilon\| \leq \rho} F(\theta + \epsilon)$, and $f_i(\tilde{\theta}) \triangleq \max_{\|\epsilon_i\| \leq \rho} F_i(\theta + \epsilon_i)$ with $\tilde{\theta} = \theta + \epsilon_i$; $\epsilon_i$ and $\epsilon$ denote the local and global perturbations, respectively.

Constrained by data privacy and infrequent communication between the PS and devices, SAM-based FL algorithms construct the local perturbation $\epsilon_i$ on each device $i$ for generalization. To efficiently approximate the inner maximization of $f_i(\tilde{\theta})$, we utilize the first-order Taylor expansion, as in prior SAM-based FL studies [5, 9, 13, 30, 45, 49]. The resulting linearized inner maximization of (3) is

$$\epsilon_i = \arg\max_{\|\epsilon_i\| \leq \rho} F_i(\theta + \epsilon_i) \approx \arg\max_{\|\epsilon_i\| \leq \rho} \{F_i(\theta) + \epsilon_i^\top \nabla F_i(\theta) + \mathcal{O}(\rho^2)\} \approx \rho \frac{\nabla F_i(\theta)}{\|\nabla F_i(\theta)\|} \approx \rho \frac{g_i}{\|g_i\|}, \tag{4}$$

where $g_i = \nabla F_i(\theta; \phi_i)$ denotes the stochastic gradient computed on batch data $\phi_i$ to alleviate computational burden. See Appendix B for the detailed training process of FedAvg [43] and FedSAM [5].

### 2.2 Flatness incompatibility in FedSAM



(a) Local model (b) Global model

Figure 1: Visualization of local and global loss landscapes for FedSAM on CIFAR-10. Setup includes 100 devices, 10% UAR sampling, Dirichlet-0.1, and a 3-layer CNN.

Fig. 1 illustrates flatness incompatibility by comparing the loss landscapes of a local model and the aggregated global model under FedSAM. The low-loss region of the local model is shifted from that of the global model, suggesting that the perturbation $\epsilon_i$ may guide a device toward a locally flat basin that is not aligned with the globally flat basin characterized by $\epsilon$. Hence, improving local flatness alone does not necessarily ensure that the aggregated model remains in a flat region of the global loss landscape. We next define flatness incompatibility, and identify data heterogeneity and friendly adversary as its two sources, which motivates the new variance suppression designs in Section 3.1.

**Definition 1** (Flatness incompatibility)**.** *For a perturbation radius $\rho > 0$, define the local and global $\rho$-flatness proxies as $s_i(\theta,\rho) = F_i(\theta+\rho\frac{g_i}{\|g_i\|}) - F_i(\theta)$ and $\bar{s}(\theta,\rho) = \frac{1}{N}\sum_{j\in\mathcal{N}} s_j(\theta,\rho)$, respectively. Flatness incompatibility is quantified by the inter-device dispersion of local flatness proxies:*

$$\Delta_{\mathrm{FI}}(\theta,\rho) := \frac{1}{N}\sum_{i\in\mathcal{N}}\Big(s_i(\theta,\rho) - \bar{s}(\theta,\rho)\Big)^2. \tag{5}$$

*A larger $\Delta_{\mathrm{FI}}(\theta,\rho)$ implies that devices perceive mutually incompatible local flat regions at $\theta$ and deviate from globally flat region.*

For SAM-based FL, flatness incompatibility consists of the following two sources, as analyzed in Section 4. One source is data heterogeneity, which can be quantified at $\theta$, as given by

$$\delta_i(\theta) := \nabla F_i(\theta) - \nabla F(\theta), \quad \forall i \in \mathcal{N}, \tag{6}$$

where $\delta_i(\theta)$ measures the bias of device gradients relative to the global gradient, and a larger $\|\delta_i(\theta)\|^2$ indicates greater data heterogeneity. Notably, $\delta_i(\theta)$ serves as a proxy for device drift [38], as it directly quantifies how far the local descent direction deviates from the global descent direction.

The other source of flatness incompatibility is the friendly adversary phenomenon [31] in SAM-based FL, which refers to the case where the SAM perturbation constructed from one batch of data to increase a device's batch loss can reduce the loss for another batch of data when their stochastic gradients are misaligned. Consider two different data batches $\phi_i, \phi_i' \sim \mathcal{D}_i$ sampled $\forall i \in \mathcal{N}$, and define $g_i = \nabla F_i(\theta;\phi_i)$, $g_i' = \nabla F_i(\theta;\phi_i')$, and $\epsilon_{i,\phi_i} = \rho\frac{g_i}{\|g_i\|}$, which is the maximizer of the first-order linearized inner problem for batch $\phi_i$. Based on the first-order Taylor expansion of SAM in (4), the local batch loss change of batch $\phi_i'$ along the perturbation constructed from $\phi_i$ satisfies

$$F_i(\theta+\epsilon_{i,\phi_i};\phi_i') - F_i(\theta;\phi_i') \approx \rho\frac{\langle g_i', g_i\rangle}{\|g_i\|} \le 0, \text{ if } \langle g_i', g_i\rangle \le 0. \tag{7}$$

From (7), although $\epsilon_{i,\phi_i}$ is constructed for the linearized inner maximization of batch $\phi_i$, it can reduce the loss of another batch $\phi_i'$ when the two stochastic gradients are misaligned. This indicates that the perturbation obtained from one data batch may fail to faithfully reflect the local landscape of $F_i$ around $\theta$. With stochastic gradient variability and data heterogeneity in SAM-based FL, such local perturbations can vary substantially across local updates, thereby causing local flatness search to probe different local flat regions across successive updates. To quantify the stochastic inconsistency underlying friendly adversary, we define the local gradient deviation as

$$\zeta_{i,\phi_i}(\theta) = g_i - \nabla F_i(\theta), \quad \forall i \in \mathcal{N}, \tag{8}$$

with $\mathbb{E}[\|\zeta_{i,\phi_i}(\theta)\|^2]$ giving the batch gradient variance and reflecting the level of instability in (4).

To analyze the sources that affect $\Delta_{\mathrm{FI}}(\theta,\rho)$, we establish the following two lemmas.

**Lemma 1.** *Suppose that each local objective function $F_i$ is $L$-Lipschitz smooth. With $\delta_i(\theta)$ and $\zeta_{i,\phi_i}(\theta)$ defined in* (6) *and* (8)*, flatness incompatibility $\Delta_{\mathrm{FI}}(\theta,\rho)$ is upper bounded:*

$$\mathbb{E}\big[\Delta_{\mathrm{FI}}(\theta,\rho)\big] \le \frac{3\rho^2}{N}\sum_{i\in\mathcal{N}}\mathbb{E}\big[\|\delta_i(\theta)\|^2 + 4\|\zeta_{i,\phi_i}(\theta)\|^2\big] + \frac{3}{4}L^2\rho^4. \tag{9}$$

**Algorithm 1** FedVSSAM

**Input:** $T$, $K$, $N$, $\theta^0$, $h^0 = 0$, $\rho$, $\eta_l$, $\eta_g$, $\gamma_l$, $\gamma_g$
**for** $t = 0, 1, \dots T-1$ **do**
  The PS selects the device subset $\mathcal{S}^t \subseteq \mathcal{N}$ UAR, and sends $\theta^t$ and $h^t$ to the selected devices
  **for** each device $i \in \mathcal{S}^t$ in parallel **do**
    Initialize local model $\theta_i^{t,0} = \theta^t$ and restore $\theta^t$
    **for** $k = 0, 1, \dots K-1$ **do**
      Compute gradient $g_i^{t,k} = \nabla F_i(\theta_i^{t,k}; \phi_i^{t,k})$; Perform local flatness search with (14)
      Compute gradient $\tilde{g}_i^{t,k} = \nabla F_i(\tilde{\theta}_i^{t,k}; \phi_i^{t,k})$; Perform local update with (15)
    **end for**
    Set $\theta_i^t = \theta_i^{t,K}$ and send $\theta_i^t$ to the PS
  **end for**
  The PS obtains $g^{t+1} = \frac{1}{SK} \sum_{i \in \mathcal{S}^t} \sum_{k=0}^{K-1} u_i^{t,k}$, and updates $\theta^t$ to $\theta^{t+1}$ by (16)
**end for**
**return** $\boldsymbol{\theta}^T$

**Lemma 2.** *For any fixed $\theta$, sample a batch data sample $\phi_i$ uniformly at random from $\mathcal{D}_i$, $\forall i \in \mathcal{N}$. With $g_i = \nabla F_i(\theta; \phi_i)$, $\delta_i(\theta)$ and $\zeta_{i,\phi_i}(\theta)$ satisfy*

$$\mathbb{E}\big[\, \langle \delta_i(\theta),\, \zeta_{i,\phi_i}(\theta) \rangle \,\big] = 0 \text{ and } \mathbb{E}\Big[\big\| g_i - \nabla F(\theta) \big\|^2\Big] = \mathbb{E}\Big[\big\| \delta_i(\theta) \big\|^2\Big] + \mathbb{E}\Big[\big\| \zeta_{i,\phi_i}(\theta) \big\|^2\Big]. \tag{10}$$

The proofs of **Lemmas 1** and **2** are provided in Appendix E and F, respectively. **Lemma 1** reveals that inter-device data heterogeneity $\delta_i(\theta)$ and intra-device friendly adversary $\zeta_{i,\phi_i}(\theta)$, deteriorate $\Delta_{\mathrm{FI}}(\theta, \rho)$. Moreover, an excessively large $\rho$ can amplify $\Delta_{\mathrm{FI}}(\theta, \rho)$ due to incompatible local loss landscapes. **Lemma 2** shows $\delta_i(\theta)$ and $\zeta_{i,\phi_i}(\theta)$ are orthogonal in expectation, justifying the insights on the composition of $\Delta_{\mathrm{FI}}(\theta, \rho)$. Meanwhile, **Lemma 2** decomposes the deviation between $g_i$ and $\nabla F(\theta)$ into the two additive and non-overlapping sources of flatness incompatibility: data heterogeneity and friendly adversary. These results indicate that mitigating flatness incompatibility should not treat $\delta_i(\theta)$ and $\zeta_{i,\phi_i}(\theta)$ as isolated errors. As shown in **Algorithm** 2, the SAM perturbation determines the perturbed gradient's evaluation point. Therefore, the errors induced by the two sources can affect both the perturbation direction and the subsequent local update, and further enter the global update via aggregation. This motivates incorporating one variance-suppressed direction into local flatness search, local update, and global update of SAM-based FL, to *consistently suppress data heterogeneity and friendly adversary of flatness incompatibility throughout the training stage*.

## 3 Federated variance-suppressed SAM

This section presents FedVSSAM to tackle flatness incompatibility caused by data heterogeneity and friendly adversary, and improve the global model performance. We start with assumptions.

**Assumption 1. (Smoothness)** Each local objective function $F_i$ is $L$-smooth, i.e.,

$$\|\nabla F_i(\theta_1) - \nabla F_i(\theta_2)\| \leq L \|\theta_1 - \theta_2\|, \quad \forall i \in \mathcal{N}. \tag{11}$$

**Assumption 2. (Bounded variance)** For any device $i$ and $\theta \in \mathbb{R}^d$, the stochastic gradient $\nabla F_i(\theta, \phi)$ is unbiased with bounded variance, i.e.,

$$\mathbb{E}\big[\nabla F_i(\theta; \phi)\big] = \nabla F_i(\theta) \text{ and } \mathbb{E}\left[\|\nabla F_i(\theta; \phi) - \nabla F_i(\theta)\|^2\right] \leq \sigma_{l,1}^2, \quad \forall i \in \mathcal{N}. \tag{12}$$

**Assumption 3. (Bounded heterogeneity)** The global variability of $\nabla F_i(\theta), \forall i$, is bounded, i.e.,

$$\mathbb{E}\left[\|\nabla F_i(\theta) - \nabla F(\theta)\|^2\right] \leq \sigma_{g,1}^2, \quad \forall i \in \mathcal{N}. \tag{13}$$

**Assumptions 1–3** are standard assumptions considered in non-convex optimization and FL [1, 9, 13, 22, 27, 30, 32, 35, 45, 46, 48, 49, 54, 58]. **Assumptions 2** and **3** capture the measurement of the friendly adversary phenomenon and data heterogeneity, respectively, as delineated in Section 2.2.

### 3.1 Variance suppression design of FedVSSAM

While it is feasible to generate the local perturbation $\epsilon_i^{t,k}$ and perform local SAM updates via the stochastic linearization in (4), the bias induced by flatness incompatibility, as revealed in **Lemmas 1** and **2**, can lead to inconsistent flat regions of FedSAM at the cost of global flatness, hence degrading training performance when data heterogeneity is severe [9, 13, 30, 45, 49]. To address flatness incompatibility and accelerate training, we design variance suppression operations on (i) flatness search at the devices, (ii) local update at the devices, and (iii) global update at the PS.

**Variance suppression on local flatness search.** To enhance the consistency of local flatness under data heterogeneity and friendly adversary, we seek a stable stochastic linearization. In each training round $t$, we design the local perturbation rule to acquire the perturbed parameter $\tilde{\theta}_i^{t,k}$ as

$$\tilde{\theta}_i^{t,k} = \theta_i^{t,k} + \rho \frac{h_i^{t,k}}{\|h_i^{t,k}\|} \text{ with } h_i^{t,k} = (1-\gamma_l)h^t + \gamma_l g_i^{t,k}, \tag{14}$$

where $g_i^{t,k} = \nabla F_i(\theta_i^{t,k}; \phi_i^{t,k})$ is the stochastic gradient; $h_i^{t,k}$ is a convex combination of the PS-maintained adjusted direction $h^t$ and $g_i^{t,k}$, controlled by the local mixing coefficient $\gamma_l$. This combination processes information from different devices' data to alleviate the impact of the friendly adversary. A smaller $\gamma_l$ suppresses heterogeneity-induced device drift and reduces the local gradient variance on $\frac{h_i^{t,k}}{\|h_i^{t,k}\|}$. When $\gamma_l \to 1$, $h_i^{t,k}$ reduces to $g_i^{t,k}$, and $\rho \frac{h_i^{t,k}}{\|h_i^{t,k}\|}$ is determined by $g_i^{t,k}$ and fully exposed to both data heterogeneity and friendly adversary. In this case, the devices construct $\epsilon_i^{t,k}$ in inconsistent directions, and the perturbed gradients are evaluated at mismatched local points before aggregation at the PS. This reproduces the perturbation inconsistency in FedSAM identified in Section 2.2, which causes devices to generate mutually incompatible local flat regions.

**Variance suppression on local update.** Unlike centralized training [2, 29, 31, 62], the local gradient divergence induced by data heterogeneity in FL can severely impede SAM optimization. To address this, we design the local update with a variance-suppressed mixed direction $u_i^{t,k}$:

$$\theta_i^{t,k+1} = \theta_i^{t,k} - \eta_l u_i^{t,k} \text{ with } u_i^{t,k} = (1-\gamma_l)h^t + \gamma_l \tilde{g}_i^{t,k}, \tag{15}$$

where $\tilde{g}_i^{t,k} = \nabla F_i(\tilde{\theta}_i^{t,k}; \phi_i^{t,k})$ is the perturbed local gradient, and $\eta_l$ denotes the local learning rate.

As proved in Appendix G, conditioning on the $\sigma$-algebra $\mathcal{F}^{t,k}$, it yields $\mathrm{Var}[u_i^{t,k} \mid \mathcal{F}^{t,k}] = \gamma_l^2 \cdot \mathrm{Var}[\tilde{g}_i^{t,k} \mid \mathcal{F}^{t,k}]$, indicating $\gamma_l < 1$ suppresses the stochastic variance of the mixed update direction $u_i^{t,k}$. Using the same $\gamma_l$ in (14) and (15) makes $h_i^{t,k}$ and $u_i^{t,k}$ share the same adjusted direction $h^t$. This reduces the mismatch between the direction used to construct $\tilde{\theta}_i^{t,k}$ and the subsequent update direction, thereby alleviating the additional bias in $u_i^{t,k}$ caused by evaluating $\tilde{g}_i^{t,k}$ at $\tilde{\theta}_i^{t,k}$ generated from a mismatched search direction. Consequently, the local flatness search in (14) and the local update in (15) share the same $h^t$, which improves consistency between the perturbation and update directions per device. Moreover, the $\gamma_l$-weighted local gradients $g_i^{t,k}$ and $\tilde{g}_i^{t,k}$ retain device-specific loss information and preserve the responsiveness of local flatness search and update to local data.

**Variance suppression on global update.** After local training, each selected device sends $\theta_i^t = \theta_i^{t,K}$ to the PS, which aggregates and scales the local updates to form $g^{t+1} = \frac{1}{\eta_l K}(\theta^t - \frac{1}{S}\sum_{i\in\mathcal{S}^t} \theta_i^t) = \frac{1}{SK}\sum_{i\in\mathcal{S}^t}\sum_{k=0}^{K-1} u_i^{t,k}$. Here, $\theta^t - \frac{1}{S}\sum_{i\in\mathcal{S}^t}\theta_i^t$, is divided by $\eta_l K$ to ensure $g^{t+1}$ directly comparable with $h^t$ in scale. Moreover, averaging over $S$ devices reduces stochastic fluctuation in $g^{t+1}$.

The global variance-suppressed direction $h^{t+1}$ is then updated by exponentially moving average (EMA), consistent with the variance suppression used in (14) and (15):

$$\theta^{t+1} = \theta^t - \eta_g h^{t+1} \text{ with } h^{t+1} = (1-\gamma_g)h^t + \gamma_g g^{t+1}, \tag{16}$$

where $\eta_g$ and $\gamma_g$ denote the global learning rate and the global mixing coefficient, respectively. Since $h^{t+1}$ is sent to the selected devices in the next training round $t+1$ and used in (14) and (15), the tradeoff between gradient smoothing and local adaptation is controlled by $\gamma_g\gamma_l$: A smaller value places more weight on $h^t$ and less on the current averaged perturbed local gradients.

These variance-suppression operations cope with $\delta_i(\theta)$ and $\zeta_{i,\phi_i}(\theta)$ identified in **Lemmas 1** and **2** at different stages of FedVSSAM. The local flatness search design in (14) reduces the sensitivity of

the perturbation direction to data heterogeneity and the local gradient deviation $\zeta_{i,\phi_i}(\theta)$ by replacing the local stochastic direction with a mixed direction anchored by $h^t$. The local update design in (15) suppresses the stochastic variability associated with the batch gradient variance and, by sharing the same $h^t$ with (14), mitigates the propagation of heterogeneity-induced device drift from perturbation construction to local descent. The global update in (16) further smooths the accumulated local updates, reducing the residual effects of both device drift and batch gradient variability in the aggregated direction. Hence, these operations jointly prevent the two sources of flatness incompatibility from being amplified across local flatness search, local update, and global update.

# 4 Convergence guarantee of FedVSSAM under non-convex setting

This section analyzes the convergence and variance suppression of FedVSSAM.

## 4.1 Convergence theorem

Considering non-convex settings, under **Assumptions 1–3**, we establish the following theorem:

**Theorem 1.** *Suppose that $F_i, \forall i \in \mathcal{N}$ is non-convex. Let the learning rates satisfy $\eta_l \leq \frac{1}{\sqrt{2}\gamma_l KL}$ and $\eta_g \leq \min\{\frac{1}{24L}, \frac{\gamma_l\gamma_g}{6L}\}$. Given the optimal global model $\theta^*$, the convergence upper bound of FedVSSAM with $T$ training rounds is given as*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\Big[\big\|\nabla F\big(\tilde{\theta}^t\big)\big\|^2\Big] \leq \frac{(125\gamma_l\gamma_g+88\eta_g L)\Delta}{20\eta_g\gamma_l\gamma_g T}+\frac{11}{5\gamma_l\gamma_g}a_{12}+\frac{31e^2}{5}(\gamma_l\eta_l KL)^2G^0, \tag{17}$$

*where $\Delta = F(\tilde{\theta}^0) - F^*$, $G^0 = \frac{1}{N}\sum_{i\in\mathcal{N}}\|\nabla F_i(\tilde{\theta}^0)\|^2$, $e$ denotes the Euler's number, and the expression of $a_{12}$ is given in* (48)*. The full proof can be found in Appendix H.*

As revealed in **Theorem 1**, with the variance suppression operations designed in (14)–(16), FedVSSAM enables the convergence of non-convex models under data heterogeneity and partial device participation. The first term on the RHS of (17) vanishes as $T \to \infty$, while the non-vanishing term $\frac{11}{5\gamma_l\gamma_g}a_{12}$ characterizes the effects of data heterogeneity ($\sigma_{g,1}^2$), friendly adversary ($\sigma_{l,1}^2$), and partial device participation ($\frac{N-S}{S(N-1)}$). Moreover, the choices of the mixing coefficients $\gamma_l$ and $\gamma_g$ exhibit a trade-off for FedVSSAM. A smaller $\gamma_l\gamma_g$ shrinks $\frac{11}{5\gamma_l\gamma_g}a_{12} + \frac{31e^2}{5}(\gamma_l\eta_l KL)^2G^0$, leading to a decreased non-vanishing error, while it can enlarge $\frac{(125\gamma_l\gamma_g+88\eta_g L)}{20\eta_g\gamma_l\gamma_g}$ and retard convergence. As the proportion of participating devices $\frac{S}{N}$ increases, the non-vanishing error $\frac{11}{5\gamma_l\gamma_g}a_{12}$ decreases, indicating that more device participation can improve convergence.

**Corollary 1.** *Define $\Sigma = \frac{\sigma_{l,1}^2}{NK} + \frac{N-S}{S(N-1)}\sigma^2$ with $\sigma^2 = \sigma_{l,1}^2 + \sigma_{g,1}^2$ and $\beta = \gamma_l\gamma_g$. Set $\beta = \min\Big\{1, \sqrt{\frac{L\Delta}{\Sigma T}}\Big\}$, the global and local learning rates $\eta_g \leq \min\Big\{\frac{1}{24L}, \frac{\beta}{6L}\Big\}$ and $\eta_l \leq \frac{1}{\sqrt{2}\gamma_l KL}$, $\eta_l KL \lesssim \min\Big\{1, \big(\frac{L\Delta}{G^0\beta^3T}\big)^{0.5}, \frac{1}{(\beta S)^{0.5}}, \frac{1}{(\beta^3 SK)^{0.25}}\Big\}$, and $\rho \lesssim \frac{1}{L\,(SKT)^{0.25}}$, FedVSSAM guarantees*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\Big[\big\|\nabla F\big(\tilde{\theta}^t\big)\big\|^2\Big] \leq \frac{L\Delta}{T}+\sqrt{\frac{L\Delta\sigma_{l,1}^2}{NKT}}+\sqrt{\frac{L\Delta\sigma^2}{ST}}. \tag{18}$$

It is indicated in **Corollary 1** that under suitable choices of $\eta_l$ and $\eta_g$ and $\gamma_l$ and $\gamma_g$, FedVSSAM achieves a linear speedup with the number of participating devices, $S$, and the number of local training iterations, $K$. Without requiring devices to transmit additional auxiliary vectors, such as control or dual variables, FedVSSAM guarantees $\mathcal{O}(\frac{1}{T} + \frac{1}{\sqrt{NKT}} + \frac{1}{\sqrt{ST}})$, incurring no additional convergence penalty that scales with $K$. A further discussion of the convergence rates and communication requirements are in Appendix D.

## 4.2 Discussion

FedVSSAM can be regarded as an extension of FedSAM. When the mixing coefficients $\gamma_l = \gamma_g = 1$ and the FedAvg-style aggregation scheme are employed, FedVSSAM degenerates to the vanilla

FedSAM. With the variance suppression designs on flatness search, local and global updates in (14), (15), and (16), respectively, FedVSSAM alleviates the impacts of data heterogeneity and friendly adversary, addressing flatness incompatibility. To further expose the effects of the proposed variance suppression designs, i.e., (14), (15), and (16), we establish the following theorem to quantify how accurately the adjusted direction $h^t$ tracks the target global gradient.

**Theorem 2.** *Under **Assumptions 1–3**, following the definitions and settings of $\Sigma$, $\beta$, $\eta_g$, $\eta_l$, and $\eta_l KL$ in **Corollary 1**, FedVSSAM guarantees the MSE between the adjusted gradient, $h^t$, and the global gradient, $\nabla F(\theta^t)$, is upper bounded by*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|h^t - \nabla F(\theta^t)\|^2\right] \lesssim \frac{L\Delta}{\beta T} + \beta\Sigma + \sqrt{\frac{L\Delta\Sigma}{T}} + \frac{\Sigma}{\sqrt{T}}, \tag{19}$$

*where $\Delta = F(\tilde{\theta}^0) - F^*$. The full proof can be found in Appendix I.1.*

**Theorem 2** quantifies the effects of variance suppression (14), (15), and (16) and the role of the mixing coefficients $\gamma_l$ and $\gamma_g$. The term $\frac{L\Delta}{\beta T}$ on the RHS of (19) vanishes increasingly fast with the rise of $\beta$, whereas the residual variance $\beta\Sigma$ is reduced due to a smaller $\beta$. This reaffirms the role of $\beta$ to balance the vanishing term $\frac{L\Delta}{\beta T}$ and the residual variance $\beta\Sigma$: Increasing $\beta$ can accelerate early-stage convergence but enlarges the residual variance, whereas decreasing $\beta$ suppresses the residual variance at the cost of a slower reduction of $\frac{L\Delta}{\beta T}$. A further discussion can be found in Appendix I.2.

# 5 Experiment

In this section, we empirically evaluate the proposed FedVSSAM in enhancing generalization and accelerating training. All experiments are run on NVIDIA 4090 GPUs.

## 5.1 Experimental setup

We compare FedVSSAM with the vanilla FedAvg [43], and SAM-based FL algorithms, including FedSAM [5], MoFedSAM [45], FedGAMMA [9], FedSMOO [49], and FedLESAM [13]. We also consider the representative state-of-the-art heterogeneity mitigation and customized aggregation FL algorithms, including FedProx [35], SCAFFOLD [22], FedCM [58], FedDyn [1], FedAdam [46], and FedACG [24].

Consider 100/200-device FL systems with 10% and 5% of devices selected UAR per round, respectively. CIFAR-10 and CIFAR-100 datasets [28] are considered with the ResNet-18 model [16] and the WRN-28-4 model [60], respectively. To characterize data heterogeneity, we utilize both the Dirichlet distribution with parameter $\alpha$, and the Pathological distribution with parameter $c$ [9, 13, 30, 45, 49]. The corresponding parameters $\alpha$ and $c$ are selected as $\{0.1, 0.5\}$ and $\{3, 6\}$ for CIFAR-10, and $\{0.1, 0.5\}$ and $\{10, 20\}$ for CIFAR-100, to account for moderate and severe data heterogeneity, respectively. All algorithms are trained for 800 rounds; the learning rate grids, batch sizes, local epochs, perturbation radius, and specific algorithm parameters are summarized in Appendix C.1. We also conduct text topic classification on DBpedia-14 [61] with a TextCNN [26] to demonstrate the performance of FedVSSAM in text classification, see Appendix C.3. Additional evaluation on the naturally federated UCI HAR dataset is provided in Appendix C.7.

## 5.2 Evaluation of FedVSSAM

As illustrated in Table 1 and Appendix C.2, FedVSSAM consistently outperforms the baselines across datasets, models, heterogeneity levels, and sampling settings, validating its effectiveness in addressing flatness incompatibility and enhancing generalization; the convergence curve analysis is provided in Appendix C.2. Compared with the second-best algorithm, FedVSSAM improves accuracy by 1.4% and 1.7% with $\alpha = 0.5$ and 100 devices on CIFAR-10 and CIFAR-100, respectively. When data heterogeneity becomes intense (e.g., $\alpha = 0.1$), FedVSSAM still achieves considerable performance enhancement over the baselines, attributed to its efficient implementation of a more globally consistent flatness search, enabling the model to achieve enhanced generalization under non-IID data.

Table 1: Test accuracy of FedVSSAM and baselines on the CIFAR-10 dataset after 800 training rounds. The device sampling setups are 10%-100 devices (upper part) and 5%-200 devices (lower part), respectively. Acc. means the final accuracy, and Rd. represents the number of training rounds required for an algorithm to first reach 75% accuracy. A dash (–) indicates that the corresponding target accuracy was not reached within the training budget under that setting. Results on CIFAR-100 are provided in Appendix C.2.

| | Dirichlet | | | | Pathological | | | |
|---|---|---|---|---|---|---|---|---|
| | $\alpha$=0.5 | | $\alpha$=0.1 | | $c$=6 | | $c$=3 | |
| Method | Acc. | Rd. | Acc. | Rd. | Acc. | Rd. | Acc. | Rd. |
| FedAvg [43] | $77.05_{\pm.14}$ | 315 | $72.57_{\pm.19}$ | – | $77.55_{\pm.24}$ | 299 | $74.63_{\pm.23}$ | – |
| FedProx [35] | $76.82_{\pm.14}$ | 337 | $72.23_{\pm.16}$ | – | $77.81_{\pm.19}$ | 271 | $74.89_{\pm.20}$ | – |
| FedAdam [46] | $78.35_{\pm.13}$ | 191 | $75.76_{\pm.24}$ | 741 | $78.13_{\pm.21}$ | 188 | $77.26_{\pm.21}$ | 440 |
| SCAFFOLD [22] | $80.14_{\pm.16}$ | 170 | $75.81_{\pm.18}$ | 635 | $81.35_{\pm.15}$ | 136 | $78.72_{\pm.15}$ | 326 |
| FedACG [24] | $78.84_{\pm.18}$ | **113** | $76.85_{\pm.15}$ | 457 | $80.62_{\pm.15}$ | **85** | $79.02_{\pm.14}$ | **199** |
| FedCM [58] | $76.23_{\pm.21}$ | 388 | $74.07_{\pm.21}$ | – | $77.48_{\pm.19}$ | 212 | $76.14_{\pm.19}$ | 574 |
| FedDyn [1] | $76.80_{\pm.16}$ | 356 | $74.60_{\pm.17}$ | – | $77.19_{\pm.19}$ | 255 | $76.78_{\pm.17}$ | 446 |
| FedSAM [5] | $78.75_{\pm.14}$ | 266 | $74.06_{\pm.21}$ | – | $80.25_{\pm.13}$ | 240 | $76.07_{\pm.14}$ | 639 |
| MoFedSAM [45] | $81.91_{\pm.16}$ | 199 | $\mathbf{78.94}_{\pm.14}$ | 385 | $\mathbf{82.94}_{\pm.16}$ | 178 | $81.15_{\pm.17}$ | 291 |
| FedGAMMA [9] | $82.36_{\pm.13}$ | 201 | $77.02_{\pm.15}$ | **208** | $82.74_{\pm.18}$ | 179 | $79.49_{\pm.19}$ | 393 |
| FedSMOO [49] | $\mathbf{81.49}_{\pm.16}$ | 147 | $78.85_{\pm.22}$ | 358 | $82.87_{\pm.12}$ | 134 | $\mathbf{80.76}_{\pm.23}$ | 289 |
| FedLESAM [13] | $76.88_{\pm.17}$ | 306 | $72.31_{\pm.18}$ | – | $77.64_{\pm.23}$ | 240 | $75.49_{\pm.17}$ | 736 |
| FedVSSAM(ours) | $\mathbf{83.56}_{\pm.19}$ | **111** | $\mathbf{79.57}_{\pm.14}$ | **296** | $\mathbf{83.57}_{\pm.22}$ | **107** | $\mathbf{81.11}_{\pm.23}$ | **196** |
| FedAvg [43] | $74.86_{\pm.27}$ | – | $70.30_{\pm.36}$ | – | $75.15_{\pm.21}$ | 789 | $72.74_{\pm.37}$ | – |
| FedProx [35] | $74.27_{\pm.17}$ | – | $70.82_{\pm.23}$ | – | $74.65_{\pm.25}$ | – | $72.10_{\pm.31}$ | – |
| FedAdam [46] | $75.05_{\pm.32}$ | 751 | $74.56_{\pm.28}$ | – | $75.57_{\pm.17}$ | 615 | $76.17_{\pm.20}$ | 485 |
| SCAFFOLD [22] | $77.79_{\pm.35}$ | 299 | $74.19_{\pm.21}$ | – | $78.13_{\pm.16}$ | 264 | $76.18_{\pm.29}$ | 549 |
| FedACG [24] | $78.03_{\pm.35}$ | **213** | $75.64_{\pm.29}$ | 603 | $77.71_{\pm.25}$ | 258 | $77.68_{\pm.23}$ | **331** |
| FedCM [58] | $72.88_{\pm.22}$ | – | $72.01_{\pm.17}$ | – | $73.93_{\pm.37}$ | – | $74.25_{\pm.20}$ | – |
| FedDyn [1] | $72.53_{\pm.17}$ | – | $71.23_{\pm.30}$ | – | $72.94_{\pm.16}$ | – | $73.48_{\pm.18}$ | – |
| FedSAM [5] | $76.17_{\pm.35}$ | 581 | $71.84_{\pm.33}$ | – | $77.38_{\pm.20}$ | 394 | $74.19_{\pm.28}$ | – |
| MoFedSAM [45] | $81.93_{\pm.26}$ | 216 | $79.00_{\pm.32}$ | **404** | $82.65_{\pm.25}$ | **178** | $80.13_{\pm.21}$ | 332 |
| FedGAMMA [9] | $81.17_{\pm.31}$ | 218 | $76.99_{\pm.25}$ | 574 | $81.00_{\pm.37}$ | 203 | $78.85_{\pm.37}$ | 433 |
| FedSMOO [49] | $\mathbf{78.28}_{\pm.30}$ | 246 | $\mathbf{77.31}_{\pm.27}$ | 512 | $79.70_{\pm.31}$ | 210 | $78.48_{\pm.20}$ | 393 |
| FedLESAM [13] | $74.61_{\pm.37}$ | – | $69.89_{\pm.23}$ | – | $75.14_{\pm.37}$ | 708 | $72.69_{\pm.37}$ | – |
| FedVSSAM(ours) | $\mathbf{83.14}_{\pm.30}$ | **145** | $\mathbf{79.42}_{\pm.23}$ | **281** | $\mathbf{83.52}_{\pm.32}$ | **131** | $\mathbf{80.90}_{\pm.29}$ | **236** |

Table 1 also demonstrates that FedVSSAM exhibits a significant advantage in communication efficiency as well. For instance, FedVSSAM takes 111 rounds to reach 75% accuracy for CIFAR-10 with 100 devices and $\alpha = 0.5$, which requires $64.8\%$ fewer rounds than FedAvg. Variance suppression operations on both devices and the PS side enable the designed adjusted direction $h^{t+1}$ to rapidly align with the ideal $\nabla F(\theta^{t+1})$, thereby achieving effective convergence acceleration.

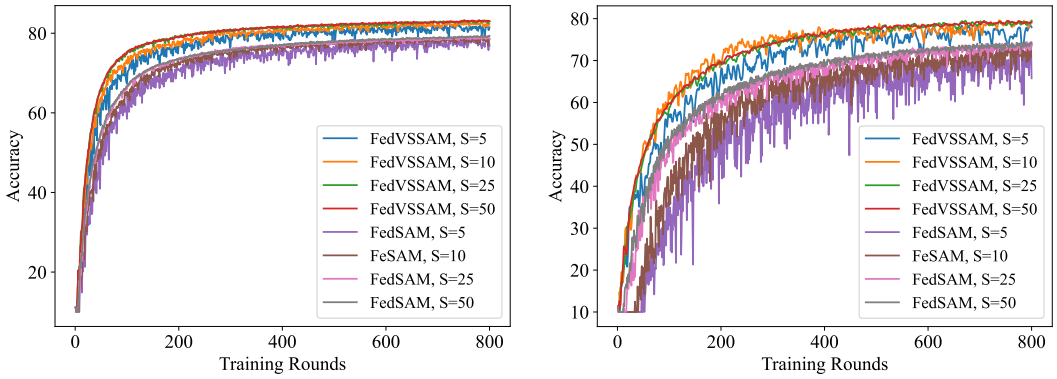
**Visualization results.** We conduct t-SNE visualization [52] of feature embeddings to provide qualitative evidence for the representation quality and generalization advantage of FedVSSAM; as detailed in Appendix C.4.

**Effect of device sampling.** Fig. 2 evaluates FedVSSAM under different numbers of participating devices, where a UAR selection of $5\%$, $10\%$, $25\%$, and $50\%$ out of $N = 100$ devices is considered. Clearly, more device participation leads to faster convergence. A larger number of selected devices $S$ contributes to a more stable adjusted direction $h^t, \forall t$ in (16) and a tighter bound on the RHS of (18) and helps better guide local flatness search and gradient modifications.

**Global direction tracking and flatness incompatibility suppression.** We examine the global direction tracking and flatness incompatibility suppression of FedVSSAM. The results provide additional illustration for the role of $h^t$ in variance suppression; see Appendix C.5.

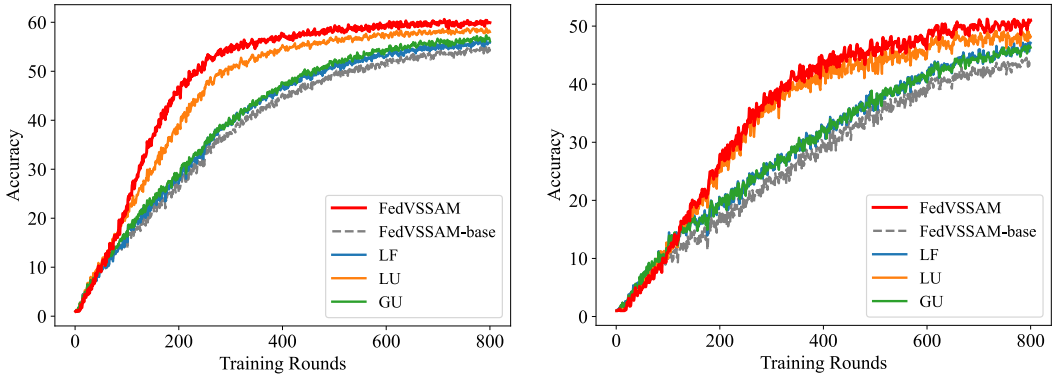
**Ablation study of variance suppression operations.** In Fig. 3, we conduct ablation studies on the variance suppression operations in FedVSSAM. By setting $\gamma_l = \gamma_g = 1$ as the baseline of

FedVSSAM, i.e., FedVSSAM-base, we individually activate the variance suppression of local flatness search (14), local update (15), and global update (16), named LF, LU, and GU, respectively, to examine their respective gains on convergence. It is observed that all LF, LU, and GU achieve performance gains compared with FedVSSAM-base, indicating the effectiveness of each designed variance suppression operation. Meanwhile, detailed sensitivity results on $\gamma_l$, $\gamma_g$, $K$, and $\rho$ in Appendix C.6 show the robustness of FedVSSAM.




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

Figure 2: Accuracy with different numbers of participating devices on CIFAR-10.




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

Figure 3: Ablation studies on the variance suppression operations in FedVSSAM.

## 6 Conclusion

In this paper, we proposed FedVSSAM to address flatness incompatibility and expedite model training in SAM-based FL with designed variance suppression operations consistently in local flatness search, local update, and global aggregation. By identifying data heterogeneity and friendly adversary as two sources of flatness incompatibility, FedVSSAM provided a unified mechanism to stabilize perturbation construction, local updates, and global updates. We established non-convex convergence under data heterogeneity and partial device participation, and proved how the designed variance suppression operations control the MSE between the variance-suppressed adjusted direction and the global gradient. Numerical results further demonstrated that FedVSSAM achieves better convergence speed and final accuracy over both FL and SAM-based FL baselines in most of the cases. The current study focuses on standard smoothness, bounded variance, and UAR device participation assumptions with simulation results on vision, text, and sensor tasks. Extending FedVSSAM to broader model families, real-world hardware deployment, and integration with privacy mechanisms remains important future work.

## Acknowledgments and Disclosure of Funding

This work was supported by the Innovation Program of the Shanghai Municipal Science and Technology Commission under Grant Nos. 26530713400 and 25DP1500300, and by the National Natural Science Foundation of China under Grant No. 62231010.

The authors declare no competing interests.

## References

[1] Durmus Alp Emre Acar, Yue Zhao, Ramon Matas Navarro, Matthew Mattina, Paul N. Whatmough, and Venkatesh Saligrama. Federated learning based on dynamic regularization. In *International Conference on Learning Representations (ICLR)*, 2021.

[2] Maksym Andriushchenko and Nicolas Flammarion. Towards understanding sharpness-aware minimization. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 639–668. PMLR, 17–23 Jul 2022.

[3] Alan Bain and Dan Crisan. *Fundamentals of Stochastic Filtering*. Springer, 2009.

[4] Keith Bonawitz, Hubert Eichner, Wolfgang Grieskamp, Dzmitry Huba, Alex Ingerman, Vladimir Ivanov, Chloé Kiddon, Jakub Konečný, Stefano Mazzocchi, Brendan McMahan, Timon Van Overveldt, David Petrou, Daniel Ramage, and Jason Roselander. Towards federated learning at scale: System design. In A. Talwalkar, V. Smith, and M. Zaharia, editors, *Proceedings of Machine Learning and Systems*, volume 1, pages 374–388, 2019.

[5] Debora Caldarola, Barbara Caputo, and Marco Ciccone. Improving generalization in federated learning by seeking flat minima. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision – ECCV 2022*, pages 654–672. Springer Nature Switzerland, 2022.

[6] Debora Caldarola, Pietro Cagnasso, Barbara Caputo, and Marco Ciccone. Beyond local sharpness: Communication-efficient global sharpness-aware minimization for federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 25187–25197, June 2025.

[7] Ziheng Cheng, Xinmeng Huang, Pengfei Wu, and Kun Yuan. Momentum benefits non-iid federated learning simply and provably. In *The Twelfth International Conference on Learning Representations*, 2024.

[8] Liam Collins, Hamed Hassani, Aryan Mokhtari, and Sanjay Shakkottai. Exploiting shared representations for personalized federated learning. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 2089–2099. PMLR, 18–24 Jul 2021.

[9] Rong Dai, Xun Yang, Yan Sun, Li Shen, Xinmei Tian, Meng Wang, and Yongdong Zhang. FedGAMMA: Federated learning with global sharpness-aware minimization. *IEEE Transactions on Neural Networks and Learning Systems*, 35(12):17479–17492, 2024.

[10] Jiawei Du, Hanshu Yan, Jiashi Feng, Joey Tianyi Zhou, Liangli Zhen, Rick Siow Mong Goh, and Vincent Tan. Efficient sharpness-aware minimization for improved training of neural networks. In *International Conference on Learning Representations*, 2022.

[11] Gintare Karolina Dziugaite and Daniel M. Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. In *Proceedings of International Conference on Machine Learning Workshop*, 2017.

[12] Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach. In *Advances in Neural Information Processing Systems*, volume 33, pages 3557–3568, 2020.

[13] Ziqing Fan, Shengchao Hu, Jiangchao Yao, Gang Niu, Ya Zhang, Masashi Sugiyama, and Yanfeng Wang. Locally estimated global perturbations are better than local perturbations for federated sharpness-aware minimization. In *Forty-first International Conference on Machine Learning*, 2024.

[14] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2021.

[15] Yujie Gu, Richeng Jin, Zhaoyang Zhang, and Huaiyu Dai. Gradient compression may hurt generalization: A remedy by synthetic data guided sharpness aware minimization. *arXiv preprint arXiv:2602.11584*, 2026.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[17] Sepp Hochreiter and Jürgen Schmidhuber. Flat minima. *Neural Computation*, 9(1):1–42, 1997.

[18] Tzu-Ming Harry Hsu, Hang Qi, and Matthew Brown. Measuring the effects of non-identical data distribution for federated visual classification. *arXiv preprint arXiv:1909.06335*, 2019.

[19] Jifei Hu, Yanli Li, Huayong Xie, Lijun Xu, Hang Zhang, and Xinqiang Zhou. Local sharpness aware minimization in decentralized federated learning with privacy protection. *Expert Systems with Applications*, page 131510, 2026.

[20] Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic generalization measures and where to find them. In *International Conference on Learning Representations*, 2020.

[21] Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *Foundations and Trends in Machine Learning*, 14(1–2):1–210, 2021.

[22] Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. SCAFFOLD: Stochastic controlled averaging for federated learning. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 5132–5143. PMLR, 13–18 Jul 2020.

[23] Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. In *International Conference on Learning Representations*, 2017.

[24] Geeho Kim, Jinkyu Kim, and Bohyung Han. Communication-efficient federated learning with accelerated client gradient. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12385–12394, 2024.

[25] Minyoung Kim, Da Li, Shell X. Hu, and Timothy Hospedales. Fisher SAM: Information geometry and sharpness aware minimisation. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 11148–11161. PMLR, 17–23 Jul 2022.

[26] Yoon Kim. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1746–1751, 2014.

[27] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[28] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. Technical report, University of Toronto, Toronto, ON, Canada, 2009.

[29] Jungmin Kwon, Jeongseop Kim, Hyunseo Park, and In Kwon Choi. ASAM: Adaptive sharpness-aware minimization for scale-invariant learning of deep neural networks. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5905–5914. PMLR, 18–24 Jul 2021.

[30] Taehwan Lee and Sung Whan Yoon. Rethinking the flat minima searching in federated learning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 27037–27071. PMLR, 21–27 Jul 2024.

[31] Bingcong Li and Georgios Giannakis. Enhancing sharpness-aware optimization through variance suppression. In *Advances in Neural Information Processing Systems*, volume 36, pages 70861–70879. Curran Associates, Inc., 2023.

[32] Qinbin Li, Bingsheng He, and Dawn Song. Model-contrastive federated learning. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10708–10717, 2021.

[33] Qinbin Li, Yiqun Diao, Quan Chen, and Bingsheng He. Federated learning on non-iid data silos: An experimental study. In *2022 IEEE 38th International Conference on Data Engineering (ICDE)*, pages 965–978, 2022.

[34] Tao Li, Pan Zhou, Zhengbao He, Xinwen Cheng, and Xiaolin Huang. Friendly sharpness-aware minimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5631–5640, June 2024.

[35] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. In *Proceedings of Machine Learning and Systems*, volume 2, pages 429–450, 2020.

[36] Tian Li, Shengyuan Hu, Ahmad Beirami, and Virginia Smith. Ditto: Fair and robust federated learning through personalization. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 6357–6368. PMLR, 18–24 Jul 2021.

[37] Tianle Li, Yongzhi Huang, Linshan Jiang, Chang Liu, Qipeng Xie, Wenfeng Du, Lu Wang, and Kaishun Wu. FedWMSAM: Fast and flat federated learning via weighted momentum and sharpness-aware minimization. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026.

[38] Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence of FedAvg on non-iid data. *arXiv preprint arXiv:1907.02189*, 2019.

[39] Xiaoxiao Li, Meirui Jiang, Xiaofei Zhang, Michael Kamp, and Qi Dou. FedBN: Federated learning on non-iid features via local batch normalization. In *International Conference on Learning Representations (ICLR)*, 2021.

[40] Yuhang Li, Tong Liu, Yangguang Cui, Ming Hu, and Xiaoqiang Li. One arrow, two hawks: Sharpness-aware minimization for federated learning via global model trajectory. In *Forty-second International Conference on Machine Learning*, 2025.

[41] Wei Yang Bryan Lim, Nguyen Cong Luong, Dinh Thai Hoang, Yutao Jiao, Ying-Chang Liang, Qiang Yang, Dusit Niyato, and Chunyan Miao. Federated learning in mobile edge networks: A comprehensive survey. *IEEE Communications Surveys & Tutorials*, 22(3):2031–2063, 2020.

[42] Yong Liu, Siqi Mai, Xiangning Chen, Cho-Jui Hsieh, and Yang You. Towards efficient and scalable sharpness-aware minimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12360–12370, June 2022.

[43] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 1273–1282. PMLR, 2017.

[44] Mehryar Mohri, Gary Sivek, and Ananda Theertha Suresh. Agnostic federated learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4615–4625. PMLR, 09–15 Jun 2019.

[45] Zhe Qu, Xingyu Li, Rui Duan, Yao Liu, Bo Tang, and Zhuo Lu. Generalized federated learning via sharpness aware minimization. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 18250–18280. PMLR, 2022.

[46] Sashank J. Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konečný, Sanjiv Kumar, and H. Brendan McMahan. Adaptive federated optimization. In *International Conference on Learning Representations (ICLR)*, 2021.

[47] Jorge Reyes-Ortiz, Davide Anguita, Alessandro Ghio, Luca Oneto, and Xavier Parra. Human Activity Recognition Using Smartphones. UCI Machine Learning Repository, 2013. URL `https://doi.org/10.24432/C54S4K`. Dataset.

[48] Yifan Shi, Yingqi Liu, Kang Wei, Li Shen, Xueqian Wang, and Dacheng Tao. Make landscape flatter in differentially private federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24552–24562, June 2023.

[49] Yan Sun, Li Shen, Shixiang Chen, Liang Ding, and Dacheng Tao. Dynamic regularized sharpness aware minimization in federated learning: Approaching global consistency and smooth landscape. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 32991–33013. PMLR, 23–29 Jul 2023.

[50] Yan Sun, Li Shen, Tiansheng Huang, Liang Ding, and Dacheng Tao. FedSpeed: Larger local interval, less communication round, and higher generalization accuracy. *arXiv preprint arXiv:2302.10429*, 2023.

[51] Canh T. Dinh, Nguyen Tran, and Josh Nguyen. Personalized federated learning with moreau envelopes. In *Advances in Neural Information Processing Systems*, volume 33, pages 21394–21405, 2020.

[52] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[53] Hongyi Wang, Mikhail Yurochkin, Yuekai Sun, Dimitris Papailiopoulos, and Yasaman Khazaeni. Federated learning with matched averaging. In *International Conference on Learning Representations (ICLR)*, 2020.

[54] Jianyu Wang, Qinghua Liu, Hao Liang, Gauri Joshi, and H. V. Poor. A novel framework for the analysis and design of heterogeneous federated learning. *IEEE Transactions on Signal Processing*, 69:5234–5249, 2021.

[55] Kang Wei, Jun Li, Ming Ding, Chuan Ma, Hang Su, Bo Zhang, and H. Vincent Poor. User-level privacy-preserving federated learning: Analysis and performance optimization. *IEEE Transactions on Mobile Computing*, 21(9):3388–3401, 2022. doi: 10.1109/TMC.2021.3056991.

[56] Kaiyue Wen, Tengyu Ma, and Zhiyuan Li. How does sharpness-aware minimization minimize sharpness? In *OPT 2022: Optimization for Machine Learning (NeurIPS 2022 Workshop)*, 2022.

[57] Bingnan Xiao, Xichen Yu, Wei Ni, Xin Wang, and H. V. Poor. Over-the-air federated learning: Status quo, open challenges, and future directions. *Fundamental Research*, 5(4):1710–1724, 2025.

[58] Jing Xu, Sen Wang, Liwei Wang, and Andrew Chi-Chih Yao. FedCM: Federated learning with client-level momentum. *arXiv preprint arXiv:2106.10874*, 2021.

[59] Jihun Yun and Eunho Yang. Riemannian SAM: Sharpness-aware minimization on riemannian manifolds. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

[60] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

[61] Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

[62] Xingxuan Zhang, Renzhe Xu, Han Yu, Hao Zou, and Peng Cui. Gradient norm aware minimization seeks first-order flatness and improves generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20247–20257, June 2023.

[63] Yue Zhao, Meng Li, Liangzhen Lai, Naveen Suda, Damon Civin, and Vikas Chandra. Federated learning with non-iid data. *arXiv preprint arXiv:1806.00582*, 2018.

[64] Juntang Zhuang, Boqing Gong, Liangzhe Yuan, Yin Cui, Hartwig Adam, Nicha C. Dvornek, Sekhar Tatikonda, James S. Duncan, and Ting Liu. Surrogate gap minimization improves sharpness-aware training. In *International Conference on Learning Representations (ICLR)*, 2022.

# A Additional related work

**Heterogeneous FL.** FL enables collaborative training without sharing personal raw data, yet the PS can only coordinate training through uploaded model updates from the devices, inducing non-IID data distributions as central challenges [4, 21, 43, 57]. To tackle data heterogeneity, conventional studies typically modify local objectives, device updates, or model aggregation. FedProx uses a proximal term to restrict local drift, while SCAFFOLD corrects device drift with designed control variates [22, 35]. FedNova adjusts heterogeneous local progress [54] with normalized aggregation coefficients. FedDyn introduces dynamic regularization techniques to improve model training [1]. Momentum-based methods improve stability from another perspective: FedCM adds device-level momentum, and recent theory shows that momentum can benefit non-IID FL [7, 58]. PS-side adaptive optimization and the modified local gradients further improve training under heterogeneous updates [24, 27, 46]. Another line of heterogeneous FL shifts the training objective from one universal global model to device-aware objectives. FedMA constructs the global model layer by layer by matching and averaging hidden elements with similar feature-extraction signatures [53]. FedBN keeps local batch normalization statistics, and personalized representation shares only part of the model [8, 39]. Ditto and other personalized FL also show that a single averaged model may be insufficient for all devices [12, 36, 51]. These studies mainly reduce device drifts or personalization mismatch, but do not explicitly target FL generalization, which is also crucial in FL because the global model is trained from partial and non-IID local data and is expected to serve heterogeneous or unseen device distributions [5, 21, 44].

**Sharpness-aware optimization.** The connection between flat minima and generalization has been studied before SAM. Early flat-minima arguments suggested that broader basins tend to generalize better, and later empirical studies observed that models trained with large batches often converge to sharper minima with worse test performance [17, 23]. PAC-Bayesian and empirical studies further examine which sharpness-related quantities correlate with generalization in modern neural networks [11, 20]. SAM converts this geometric intuition into a minimax objective by updating the model using gradients evaluated at an adversarially perturbed point [14]. Subsequent work has refined the understanding and implementation of this perturbation. Analysis of SAM investigates when minimizing the perturbed loss reduces sharpness and how the perturbation affects the training trajectory [2, 56]. ASAM adapts the sharpness measure to scale-invariant parameterizations, and Fisher SAM interprets the perturbation through information geometry [25, 29]. Other methods adjust the geometry or cost of the SAM step: surrogate gap minimization changes the objective gap, efficient SAM reduces computation, and LookSAM improves scalability [10, 42, 64]. Gradient norm-aware minimization pursues first-order flatness, and Riemannian SAM extends the idea to manifold optimization [59, 62]. Friendly SAM studies when adversarial perturbations may become insufficiently adversarial [34]. VaSSO [31] addresses the friendly-adversary phenomenon caused by mini-batch stochasticity in centralized SAM. It uses an exponentially averaged stochastic gradient to stabilize the perturbation direction, while the outer step remains based on the perturbed stochastic gradient. FedVSSAM instead targets flatness incompatibility in heterogeneous FL, where distinct local objectives, multiple local steps, and partial participation introduce challenges absent from the centralized formulation.

**SAM-based FL.** FedSAM first brought SAM into FedAvg to improve generalization under decentralized training [5]. MoFedSAM generalizes the framework and analyzes FedSAM under data heterogeneity [45]. FedSpeed pursues larger local intervals and improved communication efficiency by combining SAM with proximal control [50]. FedSMOO combines FedDyn and dynamic regularization to approach a smoother landscape [49]. Differentially private FL has also been connected to flat landscapes, showing that sharpness-aware design can be useful when privacy noise is present [48, 55].

Other studies try to align perturbations with the global objective. FedGAMMA introduces global sharpness-aware minimization, and FedGF rethinks flat minima search by emphasizing local–global flatness discrepancy [9, 30]. FedLESAM estimates global perturbations locally to avoid relying only on local perturbations [13]. FedGloSS moves the sharpness search toward communication-efficient global sharpness, and FedGMT uses the global model trajectory to measure sharpness [6, 40]. FedWMSAM combined weighted momentum with dynamic SAM to address local–global curvature misalignment and momentum-echo oscillation [37]. Recent extensions also use synthetic data to

guide SAM optimization under gradient compression [15], and study decentralized SAM-based FL settings [19].

FedVSSAM follows the global flatness view but uses a different design principle. Existing methods often improve or approximate the perturbation or sharpness through corrected local updates, PS-side sharpness, global trajectories, or momentum scheduling. FedVSSAM instead starts from the sources of flatness incompatibility, namely data heterogeneity and friendly adversary, and uses one adjusted direction to anchor local flatness search, local update, and global update. This stage-coupled design follows directly from the characterization of flatness incompatibility in Definition 1 and Lemmas 1–2, rather than introducing a correction to an isolated training stage. Theorem 2 further bounds the tracking error of this direction relative to the global gradient. The design requires no additional uplink control or dual variables. The corresponding design distinctions are summarized in Table 7.

# B Training process of FedAvg and FedSAM

As shown in **Algorithm 2**, in each training round $t$, the PS first selects a subset, $\mathcal{S}^t \in \mathcal{N}$, of devices uniformly at random (UAR) for training. Considering limited bandwidth between the devices and PS, the subset of $S$ devices is selected, i.e., $|\mathcal{S}^t| = S$. Each training round consists of $K$ local training iterations at the $S$ participating devices. For local training iteration $k = 0, 1, \cdots, K-1$, each selected device $i \in \mathcal{S}^t$ performs a SAM-based local update, where gradient is evaluated at the perturbed local model $\tilde{\theta}_i^{t,k} = \theta_i^{t,k} + \epsilon_i^{t,k}$ instead of a vanilla SGD descent at the current local model $\theta_i^{t,k}$; see Steps 10–12. By generating gradients at the perturbed local model $\tilde{\theta}_i^{t,k}$, FedSAM smooths the local objective, and promotes solutions toward flatter minima that generalize better.

**Algorithm 2** FedAvg & FedSAM

1: **Input:** $\theta^0$, $T$, $K$, $N$, $S$, $\rho$, $\eta$;
2: **for** $t = 0, 1, \cdots, T-1$ **do**
3: The PS selects the device subset $\mathcal{S}^t \subseteq \mathcal{N}$ UAR, and sends $\theta^t$ to devices $i \in \mathcal{S}^t$;
4: **for** each device $i \in \mathcal{S}^t$ in parallel **do**
5: **for** $k = 0, 1, \cdots, K-1$ **do**
6: Compute $g_i^{t,k} = \nabla F_i(\theta_i^{t,k}; \phi_i^{t,k})$
7: ***# FedAvg***
8: Update the local model with $\theta_i^{t,k+1} = \theta_i^{t,k} - \eta g_i^{t,k}$;
9: ***# FedSAM***
10: Find $\epsilon_i^{t,k}$ via stochastic linearization $\epsilon_i^{t,k} = \rho \frac{g_i^{t,k}}{||g_i^{t,k}||}$;
11: $\tilde{\theta}_i^{t,k} = \theta_i^{t,k} + \rho \frac{g_i^{t,k}}{\|g_i^{t,k}\|}$, $\tilde{g}_i^{t,k} = \nabla F_i(\tilde{\theta}_i^{t,k}; \phi_i^{t,k})$;
12: Update the local model with $\theta_i^{t,k+1} = \theta_i^{t,k} - \eta \tilde{g}_i^{t,k}$;
13: **end for**
14: **end for**
15: The PS updates $\theta^t$ with $\theta^{t+1} = \frac{1}{S} \sum_{i \in \mathcal{S}^t} \theta_i^{t,K}$;
16: **end for**
17: **return** $\boldsymbol{\theta}^T$

# C Supplementary experimental results

## C.1 Details on algorithm parameters

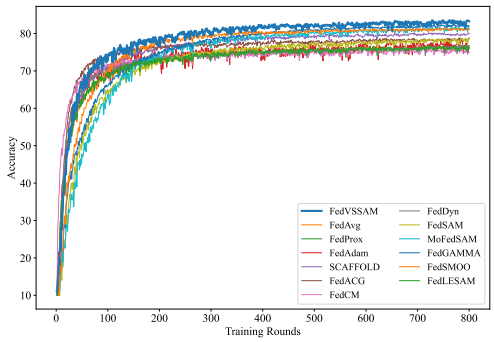
Table 2 summarizes the parameter settings used in the experiments. For a fair comparison, all methods use the same data partitions, device sampling, and random seeds under each dataset, heterogeneity level, and participation setting. Algorithm-specific hyperparameters are selected from the ranges recommended by the original papers. For FedVSSAM, the mixing coefficients $(\gamma_l, \gamma_g)$ are chosen according to the heterogeneity level, and kept fixed across datasets and device sampling settings with the same heterogeneity level.

Table 2: Key parameter settings

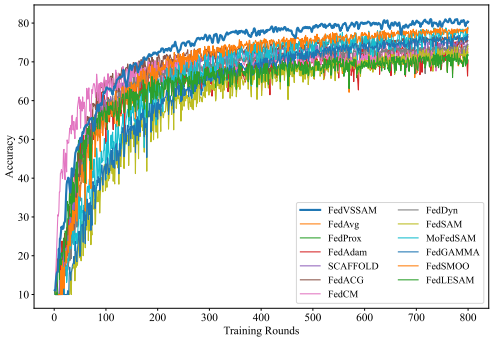
| Scope | Parameter | Setting |
|---|---|---|
| All methods | Local benchmark optimizer | SGD |
| | Learning rate grids | $\eta_l$: $\{0.01, 0.02, 0.05, 0.1, 0.2\}$; $\eta_g$: $\{0.5, 1, 2, 3, 4\}$ |
| | Batch size | CIFAR-10: 128; CIFAR-100: 20; DBpedia-14: 128 |
| | Local epochs | $K = 10$ |
| FedVSSAM | $(\gamma_l, \gamma_g)$ | Moderate heterogeneity: $(0.4, 0.6)$; Severe heterogeneity: $(0.1, 0.6)$ |
| | Perturbation radius | $\rho = 0.05$ |
| SAM-based FL baselines | Perturbation-radius grid | $\rho : \{0.001, 0.01, 0.02, 0.05, 0.1, 0.2\}$ |
| | Selected perturbation radius | $\rho = 0.1$: FedSAM/MoFedSAM/FedSMOO; $\rho = 0.01$: FedLESAM; $\rho = 0.05$: FedGAMMA |
| FedProx | Proximal coefficient | $\mu = 0.1$ |
| MoFedSAM | Momentum coefficient | $\beta = 0.1$ |
| FedCM | Client-level momentum | $\alpha = 0.1$ |
| FedDyn | Dynamic regularization | $\mu = 0.01$ |
| FedSMOO | Penalized coefficient | $\beta = 10$ |
| FedACG | Local proximal coefficient | $\beta = 0.01$ |
| FedAdam | Adam parameters | $(\beta_1, \beta_2) = (0.9, 0.99)$ |
| FedAdam | Server parameters | $(\tau, \eta)$: CIFAR-10: $(0.001, 0.01)$; CIFAR-100: $(0.1, 1)$; DBpedia-14: $(0.001, 0.01)$ |

### C.2 Additional results on CIFAR-10 and CIFAR-100

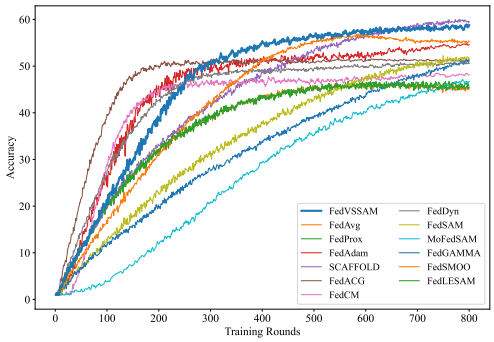
Fig. 4 reports the convergence curves on CIFAR-10 and CIFAR-100 under moderate and severe Dirichlet heterogeneity. Compared with other SAM-based FL algorithms, the performance advantage of FedVSSAM comes from its direct treatment of flatness incompatibility. Instead of only introducing corrected or momentum-based local updates, FedVSSAM jointly suppresses the perturbation inconsistency and the local update variance through the variance-suppressed flatness search, local update, and global update, see (14), (15), and (16), respectively. Thus, the local perturbations across devices are better aligned with a smoothed global direction, guiding the aggregated model toward a more globally consistent flat region under non-IID data and achieving better convergence and generalization.



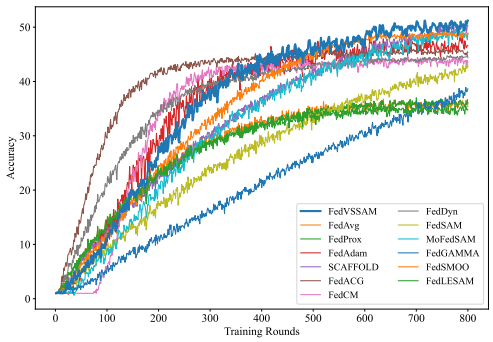

(a) CIFAR-10, $\alpha = 0.5$.




(b) CIFAR-10, $\alpha = 0.1$.




(c) CIFAR-100, $\alpha = 0.5$.




(d) CIFAR-100, $\alpha = 0.1$.

Figure 4: Convergence performance of different algorithms on CIFAR-10 and CIFAR-100.

Table 3: Test accuracy of FedVSSAM and baselines on the CIFAR-100 dataset after 800 training rounds. The device sampling setups are 10%-100 devices (upper part) and 5%-200 devices (lower part), respectively. Acc. reports the mean test accuracy (%) with standard deviation, and Rd. represents the number of training rounds required to first reach 40% accuracy. A dash (–) indicates that the target was not reached within the training budget.

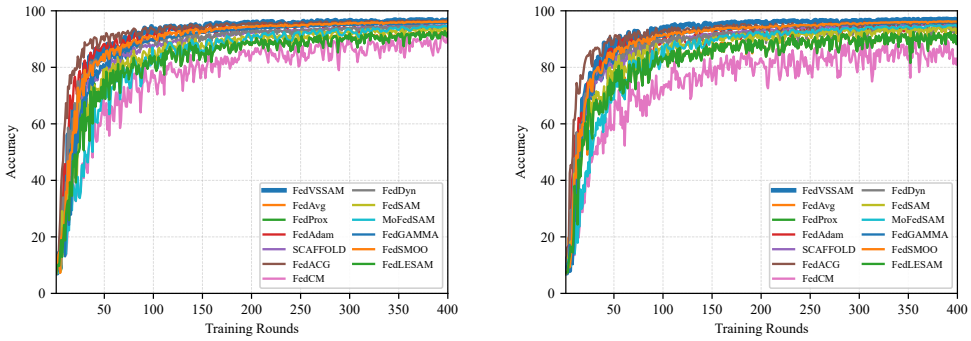
| | Dirichlet | | | | Pathological | | | |
|---|---|---|---|---|---|---|---|---|
| | $\alpha$=0.5 | | $\alpha$=0.1 | | $c$=20 | | $c$=10 | |
| Method | Acc. | Rd. | Acc. | Rd. | Acc. | Rd. | Acc. | Rd. |
| FedAvg [43] | $46.52 \pm .23$ | 313 | $36.76 \pm .23$ | – | $42.59 \pm .15$ | 486 | $36.67 \pm .21$ | – |
| FedProx [35] | $46.80 \pm .24$ | 313 | $37.00 \pm .17$ | – | $41.60 \pm .20$ | 517 | $36.75 \pm .19$ | – |
| FedAdam [46] | $54.80 \pm .24$ | 155 | $48.29 \pm .28$ | 276 | $50.38 \pm .24$ | 376 | $43.36 \pm .24$ | 637 |
| SCAFFOLD [22] | $\mathbf{59.96} \pm .32$ | 274 | $50.59 \pm .28$ | 450 | $\mathbf{55.78} \pm .33$ | 368 | $47.05 \pm .23$ | 534 |
| FedACG [24] | $51.66 \pm .20$ | **106** | $46.30 \pm .30$ | **161** | $51.85 \pm .30$ | **124** | $46.18 \pm .17$ | **232** |
| FedCM [58] | $48.51 \pm .15$ | **145** | $44.63 \pm .23$ | **254** | $49.71 \pm .22$ | 226 | $45.32 \pm .27$ | 343 |
| FedDyn [1] | $50.84 \pm .26$ | 177 | $44.83 \pm .25$ | 307 | $50.40 \pm .14$ | **209** | $44.72 \pm .28$ | 377 |
| FedSAM [5] | $52.22 \pm .17$ | 439 | $43.50 \pm .15$ | 688 | $47.52 \pm .17$ | 530 | $39.84 \pm .32$ | – |
| MoFedSAM [45] | $47.14 \pm .18$ | 576 | $49.69 \pm .15$ | 450 | $54.49 \pm .29$ | 360 | $45.60 \pm .17$ | 589 |
| FedGAMMA [9] | $51.39 \pm .26$ | 511 | $38.91 \pm .24$ | – | $45.29 \pm .22$ | 638 | $35.57 \pm .14$ | – |
| FedSMOO [49] | $56.87 \pm .23$ | 279 | $\mathbf{49.12} \pm .30$ | 396 | $55.04 \pm .21$ | 330 | $\mathbf{48.60} \pm .18$ | 463 |
| FedLESAM [13] | $46.58 \pm .15$ | 308 | $35.85 \pm .17$ | – | $43.19 \pm .32$ | 458 | $36.33 \pm .21$ | – |
| FedVSSAM(ours) | $\mathbf{58.78} \pm .30$ | 237 | $\mathbf{51.22} \pm .20$ | 328 | $\mathbf{56.87} \pm .19$ | 258 | $\mathbf{50.27} \pm .18$ | **337** |
| FedAvg [43] | $32.45 \pm .38$ | – | $28.29 \pm .28$ | – | $30.39 \pm .21$ | – | $27.74 \pm .28$ | – |
| FedProx [35] | $31.56 \pm .21$ | – | $33.21 \pm .18$ | – | $29.92 \pm .35$ | – | $27.18 \pm .21$ | – |
| FedAdam [46] | $45.96 \pm .19$ | 259 | $43.43 \pm .20$ | 494 | $39.65 \pm .22$ | – | $41.29 \pm .21$ | 720 |
| SCAFFOLD [22] | $53.96 \pm .24$ | 277 | $47.85 \pm .22$ | 484 | $46.27 \pm .20$ | 455 | $45.63 \pm .21$ | 584 |
| FedACG [24] | $45.77 \pm .31$ | **152** | $44.48 \pm .29$ | **246** | $45.73 \pm .33$ | **240** | $40.31 \pm .31$ | 670 |
| FedCM [58] | $45.77 \pm .26$ | 183 | $47.93 \pm .31$ | **269** | $43.11 \pm .26$ | 379 | $42.81 \pm .20$ | **467** |
| FedDyn [1] | $35.53 \pm .34$ | – | $32.73 \pm .20$ | – | $34.60 \pm .24$ | – | $30.78 \pm .33$ | – |
| FedSAM [5] | $46.99 \pm .22$ | 448 | $39.63 \pm .36$ | – | $40.75 \pm .23$ | 723 | $35.84 \pm .27$ | – |
| MoFedSAM [45] | $47.12 \pm .33$ | 580 | $48.47 \pm .37$ | 513 | $\mathbf{55.16} \pm .30$ | 349 | $\mathbf{47.40} \pm .24$ | 572 |
| FedGAMMA [9] | $52.86 \pm .31$ | 466 | $40.09 \pm .30$ | 799 | $48.38 \pm .35$ | 584 | $36.12 \pm .24$ | – |
| FedSMOO [49] | $\mathbf{54.02} \pm .31$ | 289 | $\mathbf{50.13} \pm .32$ | 395 | $49.96 \pm .23$ | 333 | $46.84 \pm .20$ | 472 |
| FedLESAM [13] | $31.97 \pm .25$ | – | $34.75 \pm .30$ | – | $30.11 \pm .35$ | – | $28.24 \pm .37$ | – |
| FedVSSAM(ours) | $\mathbf{57.53} \pm .26$ | 248 | $\mathbf{50.45} \pm .32$ | 376 | $\mathbf{55.94} \pm .26$ | **316** | $\mathbf{50.17} \pm .19$ | **398** |

### C.3 Additional text classification results

To further evaluate FedVSSAM beyond image classification, we conduct additional text topic classification experiments on DBpedia-14 with a TextCNN model. We use a balanced subset with 140,000 training samples and 14,000 test samples, corresponding to 10,000 training samples and 1,000 test samples per class. The TextCNN uses a vocabulary size of 30,000, a maximum sequence length of 128, 100-dimensional embeddings, and 100 convolutional filters for each kernel size in $\{3, 4, 5\}$. All methods are trained for 400 rounds with $K = 10$, and the batch size is 128.

Table 4: Test accuracy of FedVSSAM and baselines on the DBpedia-14 dataset after 400 training rounds. The device sampling setup is 10%-100 devices. Acc. means the final accuracy, and Rd. represents the number of training rounds required for an algorithm to first reach a certain accuracy level, which is 90%. A dash (–) indicates that the corresponding target accuracy was not reached within the training budget under that setting.

| Algorithm | Dirichlet, $\alpha = 0.1$ | | Pathological, $c = 3$ | |
|---|---|---|---|---|
| | Acc. | Rd. | Acc. | Rd. |
| FedAvg [43] | $92.56_{\pm.08}$ | 205 | $91.09_{\pm.11}$ | 215 |
| FedProx [35] | $92.40_{\pm.09}$ | 205 | $91.01_{\pm.12}$ | 215 |
| FedAdam [46] | $96.19_{\pm.07}$ | **62** | $94.41_{\pm.09}$ | 56 |
| SCAFFOLD [22] | $94.76_{\pm.10}$ | 133 | $95.13_{\pm.08}$ | 113 |
| FedACG [24] | $\mathbf{96.44}_{\pm.08}$ | **39** | $95.11_{\pm.11}$ | **46** |
| FedCM [58] | $88.82_{\pm.16}$ | 320 | $83.14_{\pm.21}$ | – |
| FedDyn [1] | $94.90_{\pm.12}$ | 96 | $93.71_{\pm.10}$ | 93 |
| FedSAM [5] | $93.93_{\pm.09}$ | 137 | $93.54_{\pm.11}$ | 144 |
| MoFedSAM [45] | $94.43_{\pm.08}$ | 165 | $94.46_{\pm.13}$ | 137 |
| FedGAMMA [9] | $95.76_{\pm.08}$ | 106 | $95.94_{\pm.09}$ | 87 |
| FedSMOO [49] | $96.16_{\pm.09}$ | 89 | $\mathbf{96.21}_{\pm.11}$ | 80 |
| FedLESAM [13] | $92.11_{\pm.10}$ | 205 | $90.92_{\pm.12}$ | 215 |
| FedVSSAM(ours) | $\mathbf{96.90}_{\pm.07}$ | 71 | $\mathbf{96.74}_{\pm.11}$ | **46** |




(a) DBpedia-14, $\alpha = 0.1$.

(b) DBpedia-14, $c = 3$.

Figure 5: Convergence performance of different algorithms on DBpedia-14.

As shown in Table 4 and Fig. 5, FedVSSAM achieves the best final test accuracy under both text-classification heterogeneity settings. It is noted that FedACG accelerates early training by broadcasting a lookahead global gradient for directional momentum and reaches the 90% threshold earlier than FedVSSAM on the Dirichlet partition. However, without an explicit mechanism to smooth the sharp loss landscapes typical of skewed text data, FedACG fails to manage incompatible local minima across devices, leading to a lower final performance than FedVSSAM. Under the Pathological partition setting, FedVSSAM obtains 96.74% accuracy, while also reaching 90% accuracy in 46 rounds, matching the fastest baseline FedACG. These results are consistent with the role of variance-suppressed flatness search and updates: under topic-skewed text data, different devices observe substantially different word and phrase distributions, and FedVSSAM better aligns the perturbation and update directions across devices.

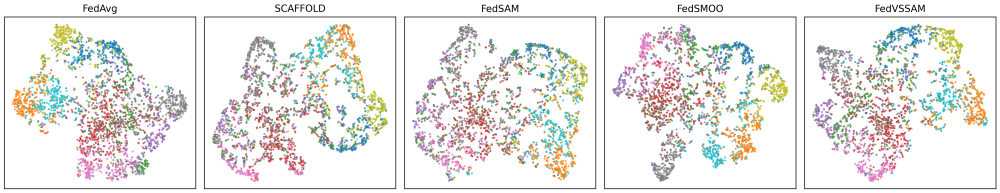
### C.4 t-SNE visualization of learned representations

To further inspect the representations learned by different algorithms, we visualize the feature embeddings of the global models using t-SNE. We compare FedAvg, SCAFFOLD, FedSAM,

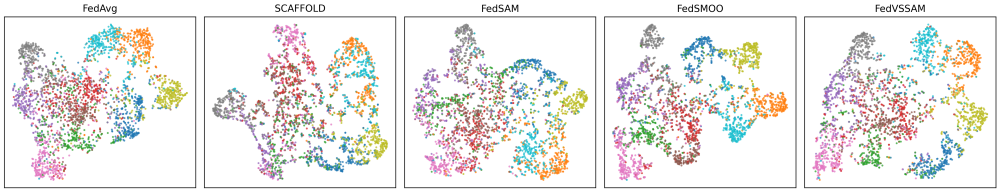
FedSMOO, and FedVSSAM on CIFAR-10 with ResNet-18 and Dirichlet distribution with $\alpha = 0.1$, and use class labels only for coloring the embedded test samples with a size of 3,000.

Here, the visualization is intended as a supplementary measure of generalization performance. Fig. 6 shows the t-SNE embeddings after 400 and 800 training rounds. At 400 rounds, FedAvg, SCAFFOLD, and FedSAM still exhibit substantial class mixing in the embedding space, indicating that the learned features are not yet clearly discriminative under heterogeneous FL. FedSMOO and FedVSSAM show more visible class-wise structures, while FedVSSAM already forms several more coherent peripheral clusters and reduces part of the central class overlap.

After 800 rounds, the separation of class clusters becomes clearer for all methods, but the difference among methods remains visible. FedVSSAM produces relatively more compact and better separated class groups than FedAvg, SCAFFOLD, and FedSAM, suggesting that the proposed variance-suppressed flatness search, local update, and global update help learn more discriminative global representations. FedSMOO also yields clear cluster structures in several regions, which is consistent with its robust empirical performance. However, FedVSSAM achieves comparable representation separation while retaining the communication advantage discussed in Appendix D. These visualizations are consistent with the main experimental results: stabilizing the directions used for local flatness search, local update, and global update can improve the quality of the learned global representation under data heterogeneity.




(a) After 400 training rounds.




(b) After 800 training rounds.

Figure 6: t-SNE visualization of global model feature embeddings on CIFAR-10 with ResNet-18. Each color denotes one class.

### C.5 Results on global direction tracking and flatness incompatibility suppression

To further verify the mechanism of FedVSSAM beyond test accuracy, we conduct an additional experiment on CIFAR-10 with ResNet-18 under Dirichlet distribution with $\alpha = 0.1$ and $\alpha = 0.5$. The system contains 100 devices with $10\%$ devices sampled UAR per round. The mechanism metrics are evaluated every 80 training rounds. We use a fixed subset of 50 devices to evaluate the local flatness, and construct an IID reference gradient from 10 randomly sampled batches of data with a batch size of 512 from the CIFAR-10 training distribution.

First, let $\widehat{\nabla F}_{\text{iid}}(\theta^t)$ denotes the empirical gradient estimated from IID sampled batches of data at the current global model $\theta^t$. For FedVSSAM, the corresponding global direction is the variance-suppressed direction $h^t$. Fig. 7 shows that FedVSSAM achieves the tracking error $\|h^t - \widehat{\nabla F}_{\text{iid}}(\theta^t)\|^2$ keeps decreasing as the training progresses, which is consistent with the variance-suppression property of $h^t$ established in **Theorem 2**.

We also compare the empirical flatness incompatibility proxy across SAM-based FL methods. Let $\mathcal{N}_{\text{eval}} \subseteq \mathcal{N}$ denote the fixed subset of evaluation devices. For each evaluation device $i \in \mathcal{N}_{\text{eval}}$, we first estimate the local empirical gradient $\hat{g}_i^t$ at $\theta^t$. Following the FedVSSAM search direction in (14), we define $d_i^t = (1 - \gamma_l)h^t + \gamma_l \hat{g}_i^t$, which is used to estimate the local flatness proxy

$s_i^t(\theta^t, \rho) = F_i\Big(\theta^t + \rho \frac{d_i^t}{\|d_i^t\|}\Big) - F_i(\theta^t)$. We then compute the empirical flatness incompatibility as

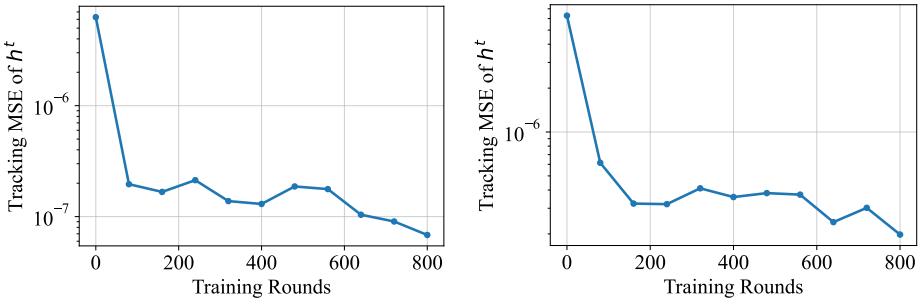
$$\widehat{\Delta}_{\mathrm{FI}}^t(\theta^t, \rho) = \frac{1}{|\mathcal{N}_{\mathrm{eval}}|} \sum_{i \in \mathcal{N}_{\mathrm{eval}}} \Big( s_i^t(\theta^t, \rho) - \frac{1}{|\mathcal{N}_{\mathrm{eval}}|} \sum_{j \in \mathcal{N}_{\mathrm{eval}}} s_j^t(\theta^t, \rho) \Big)^2, \tag{20}$$

where a smaller value of $\widehat{\Delta}_{\mathrm{FI}}^t(\theta^t, \rho)$ indicates that different devices perceive more consistent local flat regions around the current global model.

Fig. 8 reports the flatness incompatibility proxy. It is noted that FedVSSAM yields a low flatness incompatibility proxy as training progresses, supporting that the shared variance-suppressed direction reduces the incompatibility of local flatness estimates under different levels of data heterogeneity.

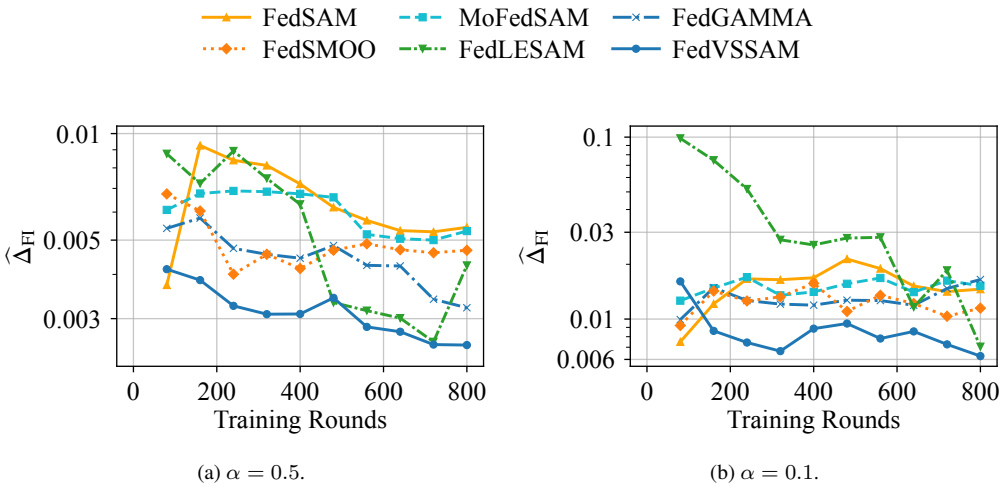
Compared with the five SAM-based FL baselines, FedVSSAM achieves the lowest average $\widehat{\Delta}_{\mathrm{FI}}$ under both $\alpha = 0.5$ and $\alpha = 0.1$, with reductions of $29.7\%$ and $29.6\%$ relative to the best baseline in each setting, respectively; it also attains the lowest value at $8/10$ and $9/10$ checkpoints, respectively. The checkpoint-wise Pearson and Spearman correlations between $\widehat{\Delta}_{\mathrm{FI}}$ and both global and mean-local sharpness proxies are positive for all compared methods, with stronger correlations to global sharpness in every case.

The above results provide direct empirical evidence for the proposed mechanism: FedVSSAM suppresses the dispersion of local flatness proxies and aligns local SAM search directions with a more stable global direction, and the performance gain is accompanied by a reduction in the flatness incompatibility defined in Section 2.2.




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

Figure 7: Tracking error between $h^t$ and the global gradient for FedVSSAM.

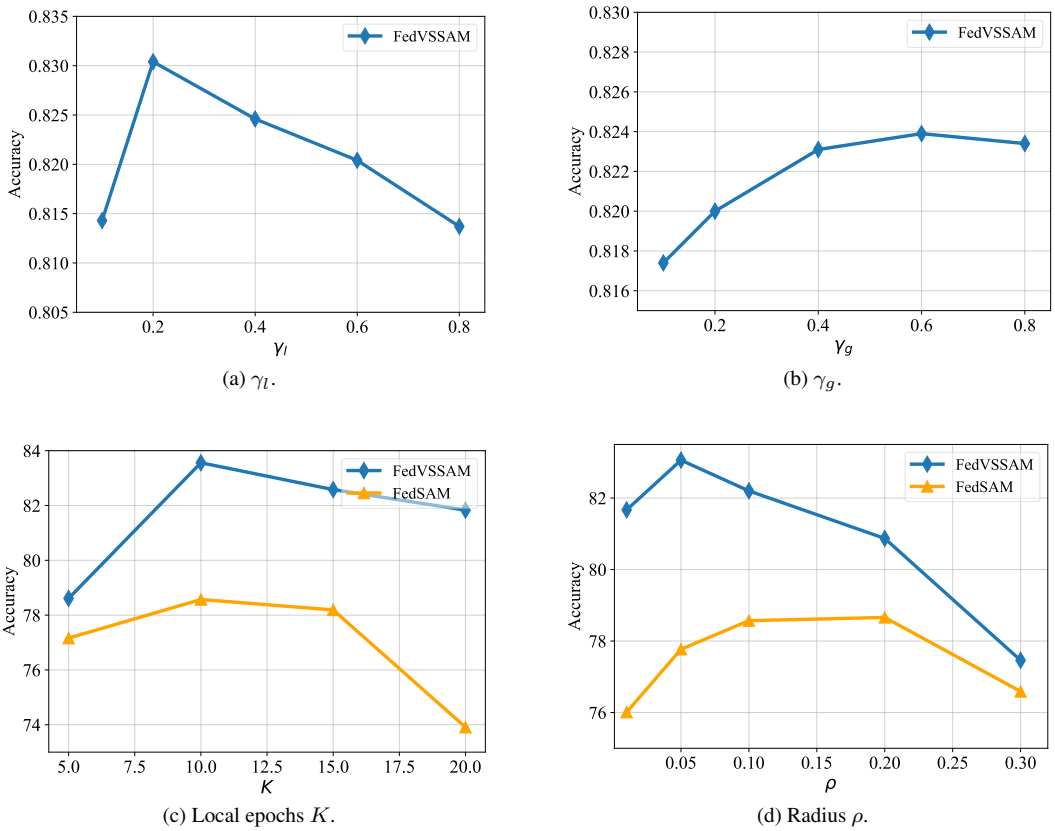



(a) $\alpha = 0.5$. (b) $\alpha = 0.1$.

Figure 8: Comparison of the empirical flatness incompatibility proxy among SAM-based FL methods on CIFAR-10 with ResNet-18, 100 devices, and 10% participation under Dirichlet heterogeneity. The metric is evaluated every 80 training rounds.

### C.6 Sensitivity analysis

We examine the sensitivity of FedVSSAM to several key parameters: the mixing coefficients $\gamma_l$ and $\gamma_g$, the number $K$ of local training epochs, and the perturbation radius $\rho$. As shown in Fig. 9, the robustness of FedVSSAM is assessed through variations in the hyperparameters by adjusting each parameter while keeping the others unchanged. For the mixing coefficients, taking $\gamma_l \to 0$ can be overly conservative: since $h^t$ is a smoothed and history-dependent global direction, and can be biased under partial participation, relying almost exclusively on $h^t$ may render the perturbation stale with respect to the current local iterate $\theta_i^{t,k}$. As a consequence, the local perturbation may no longer approximate the steepest local ascent direction required by SAM's inner maximization, weakening curvature adaptivity and slowing down progress. In contrast, an excessively large $\gamma_l$ makes the local flatness search and update dependent excessively on instantaneous local stochastic gradients, hence increasing the impact of data heterogeneity and friendly adversary. Consequently, a moderate $\gamma_l$ provides the balanced trade-off between local adaptivity and variance suppression. A similar trade-off is observed for $\gamma_g$. Particularly, an excessively small $\gamma_g$ causes the PS update of $h^t$ to be excessively slow, whereas an excessively large $\gamma_g$ weakens the smoothing effect on aggregation noise. This is consistent with the role of $\gamma_l\gamma_g$ revealed in **Theorem 2**, which balances the vanishing term and residual variance in (19). Increasing the number of local training epochs $K$ initially improves the training of FedVSSAM by allowing more effective local progress within each training round, but a large $K$ enlarges local drift. Moreover, a moderate perturbation radius $\rho$ is critical: A small $\rho$ weakens the flatness-seeking effect, while an excessively large $\rho$ amplifies the mismatch among perturbed local models, which is consistent with **Lemma 1**.




(a) $\gamma_l$. (b) $\gamma_g$.

(c) Local epochs $K$. (d) Radius $\rho$.

Figure 9: Sensitivity of FedVSSAM to different parameters on the CIFAR-10 dataset with $\alpha = 0.5$.

### C.7 Evaluation on a naturally federated sensor dataset

To evaluate FedVSSAM on naturally partitioned sensor data, we consider the UCI HAR dataset [47], treating each of the 21 training subjects as one client and using the official subject-disjoint test split.

We sample 10 clients per round and train a two-layer MLP for 200 rounds with a batch size of 32 and $K = 5$. FedVSSAM uses $(\gamma_l, \gamma_g) = (0.4, 0.6)$, and learning rates are selected following the tuning protocol in Table 2.

Table 5: Test accuracy on UCI HAR after 200 training rounds. Rd. to 90% is the number of training rounds required to first reach 90% accuracy. A dash (–) indicates that the target was not reached within the 200-round training budget.

| Method | Acc. (%) | Rd. to 90% |
|---|---|---|
| FedAvg [43] | $93.65 \pm 0.11$ | 23 |
| FedProx [35] | $93.59 \pm 0.09$ | 23 |
| FedAdam [46] | $94.94 \pm 0.13$ | 24 |
| SCAFFOLD [22] | $93.25 \pm 0.08$ | 37 |
| FedACG [24] | $95.39 \pm 0.07$ | 24 |
| FedCM [58] | $95.22 \pm 0.09$ | 39 |
| FedDyn [1] | $95.01 \pm 0.14$ | 24 |
| FedSAM [5] | $89.62 \pm 0.17$ | – |
| MoFedSAM [45] | $88.52 \pm 0.15$ | – |
| FedGAMMA [9] | $88.90 \pm 0.12$ | – |
| FedSMOO [49] | $91.82 \pm 0.08$ | 81 |
| FedLESAM [13] | $93.35 \pm 0.18$ | 24 |
| **FedVSSAM (ours)** | $\mathbf{95.55} \pm 0.09$ | **22** |

As shown in Table 5, FedVSSAM achieves the highest reported test accuracy of 95.55% and reaches 90% accuracy in 22 rounds, earlier than all compared baselines. These results extend the evaluation from synthetic image and text partitions to natural sensor clients. Real-device deployment remains a direction for future work.

## D Comparison of convergence rates and communication requirements

In this paper, we use $T$ for training rounds, $K$ for local training iterations, $S$ for selected devices, $N$ for all devices, and $\Delta = F(\tilde{\theta}^0) - F^*$ for the initial optimality gap. Since existing convergence guarantees are established under different technical assumptions, the comparison is not solely intended as a ranking of rates. Instead, it separates two aspects: the communication or auxiliary-state cost required to implement each correction mechanism, and the residual error terms that remain explicit in its convergence guarantee.

Table 6 compares the reported stationarity bounds together with the auxiliary uplink vector required by each algorithm. SCAFFOLD, FedGAMMA, FedLESAM-S, and FedSMOO use control variates or dynamic regularization, so their bounds keep correction-state or correction-inconsistency terms, and the corresponding implementations require an additional uplink vector. MoFedSAM, FedGF, FedGMT, and FedWMSAM avoid this uplink cost, but their bounds still isolate terms associated with momentum weighting, global perturbation estimation, global trajectory or ADMM inconsistency, or perturbation-induced variance. FedVSSAM keeps the device upload identical to a local model update. At round t, the PS broadcasts $h^t$ to selected devices; after aggregation, it updates $h^{t+1}$ for the next round. The same adjusted direction is then used in local flatness search, local update, and global update, so the bound of FedVSSAM does not introduce an isolated correction-specific residual beyond the common stochastic gradient, local update, heterogeneity, and partial participation terms.

**Corollary** 1 shows how the common error sources appear in the bound of FedVSSAM. The term $\sqrt{L\Delta\sigma_{l,1}^2/(NKT)}$ reflects stochastic gradient deviation in local training, and decreases with a larger scale of system devices. The term $\sqrt{L\Delta\sigma^2/(ST)}$, where $\sigma^2 = \sigma_{l,1}^2 + \sigma_{g,1}^2$, reflects the UAR sampling variance, including both local stochastic gradient variance and cross-device heterogeneity. **Theorem** 2 further complements this result by bounding the tracking error of $h^t$. It controls

Table 6: Comparison of reported non-convex convergence guarantees and auxiliary uplink vector requirements. Rates are normalized to the notation of this paper when possible, with method-specific constants and lower-order terms suppressed. N/A indicates that a comparable stationarity bound is not reported in the cited work. A dash (–) means that the bound does not induce an additional residual error beyond the common stochastic gradient, local update, heterogeneity, or partial participation errors considered in this paper; it does not mean that the method has no residual error.

| Method | Reported stationarity bound | Specific residual error | Extra uplink vector |
|---|---|---|---|
| FedAvg [43] | $\mathcal{O}\left(1/\sqrt{KT}\right) + \text{devicedrift term}$ | – | No |
| SCAFFOLD [22] | $\tilde{\mathcal{O}}\left(\sqrt{\frac{L\Delta\sigma^2}{SKT}} + \frac{L\Delta}{T}\left(\frac{N}{S}\right)^{2/3}\right)$ | Control variate tracking error | Yes |
| MoFedSAM [45] | $\mathcal{O}\left(\frac{\beta L\Delta}{\sqrt{TKS}} + \frac{\beta\sqrt{K}\sigma_g^2}{\sqrt{TS}} + \frac{L^2\sigma_l^2}{T^{3/2}K} + \frac{\sqrt{K}L^2}{T^{3/2}\sqrt{S}}\right)$ | –; momentum-weighted common variance terms | No |
| FedGAMMA [9] | $\mathcal{O}\left(1/\sqrt{TKS}\right) + \mathcal{E}_{\mathrm{GAMMA}}$ | Control variate tracking error | Yes |
| FedLESAM-S [13] | $\mathcal{O}\left(\frac{L\Delta}{\sqrt{TKN}} + \frac{L^2\rho^2\sigma_g^2}{TK} + \frac{L^2\rho^2}{T} + \frac{L^2\sigma_l^2\rho^2}{\sqrt{TKN}}\right) + \mathcal{E}_{\mathrm{Scaf}}$ | Global perturbation estimation and control variate tracking error | Yes |
| FedSMOO [49] | $\frac{1}{\zeta\beta T}\left(\kappa_f + \frac{3\beta L\kappa_r N}{S} + \frac{72\beta^2L^2\delta^0 N}{S}\right)$ | Initial correction inconsistency error | Yes |
| FedGF [30] | $\mathcal{O}\left(\frac{L\Delta}{\sqrt{TKS}} + \left(\frac{(1-c)^2}{T}+1\right)\sigma_g^2 + \frac{L^2(1-c)^2\sigma_l^2}{T^{3/2}\sqrt{KS}} + \frac{L^2(c^2\epsilon^2+1)}{T}\right)$ | Global perturbation estimation error | No |
| FedGloSS [6] | N/A | PS-side global-sharpness approximation; no comparable stationarity residual is reported | No |
| FedGMT [40] | $\frac{1}{\kappa T}\left(L(w^1) - L^* + \frac{20M\beta^2L^2}{N}\Phi^0 + \frac{18\beta G}{N}\right)$ | Global trajectory and ADMM inconsistency error | No |
| FedWMSAM [37] | $\tilde{\mathcal{O}}\left(\sqrt{\frac{L\Delta\sigma_\rho^2}{SKT}} + \frac{L\Delta}{T}\left(1 + \frac{N^{2/3}}{S}\right)\right)$ | Perturbation-induced variance | No |
| **FedVSSAM** | $\mathcal{O}\left(\frac{L\Delta}{T} + \sqrt{\frac{L\Delta\sigma_{l,1}^2}{NKT}} + \sqrt{\frac{L\Delta\sigma^2}{ST}}\right)$ | – | **No** |

Table 7: Comparison of design properties for mitigating flatness incompatibility in SAM-based FL. Each entry indicates whether the corresponding property is explicitly addressed in the algorithms.

| Method | Non-IID Correction | Update Bias Control | SAM Perturbation Noise Aware | Global Flatness Aligned | Stage Coupled Direction |
|---|---|---|---|---|---|
| FedAvg [43] | No | No | No | No | No |
| SCAFFOLD [22] | Yes | Yes | No | No | No |
| FedSAM [5] | No | No | No | No | No |
| MoFedSAM [45] | Yes | Yes | No | No | No |
| FedGAMMA [9] | Yes | Yes | Yes | Yes | No |
| FedLESAM-S [13] | Yes | Yes | Yes | Yes | No |
| FedSMOO [49] | Yes | Yes | Yes | Yes | No |
| FedGF [30] | Yes | No | Yes | Yes | No |
| FedGloSS [6] | Yes | Yes | Yes | Yes | No |
| FedGMT [40] | Yes | Yes | Yes | Yes | No |
| FedWMSAM [37] | Yes | Yes | Yes | Yes | No |
| **FedVSSAM** | **Yes** | **Yes** | **Yes** | **Yes** | **Yes** |

the MSE between the adjusted direction and the global gradient, which is consistent with the design in (14)–(16): $h^t$ serves as a common anchor for local flatness search, perturbed local update, and the global update.

In Table 7, "Non-IID Correction" means that the method explicitly introduces a mechanism beyond vanilla local training to mitigate data heterogeneity or device drifts. "Update Bias Control" indicates whether the method explicitly regularizes or corrects the bias of local update directions with respect to the global objective, beyond only modifying the sharpness objective. "SAM Perturbation

Noise Aware" refers to whether the method explicitly handles the stochasticity or approximation error in the sharpness-aware perturbation or sharpness-related direction, rather than SGD noise only. "Global Flatness Aligned" means that the perturbation or sharpness objective is tied to the global loss landscape, not only to a device's local loss. "Stage Coupled Direction" indicates whether the same stabilized direction is used across the perturbation search, the local update, and the global update.

It is noted that FedGAMMA, FedGF, FedLESAM-S, FedGloSS, FedSMOO, FedGMT, and FedWMSAM each address part of the global-flatness problem. Specifically, FedGAMMA introduces global sharpness-aware correction, FedGF and FedLESAM-S improve the perturbation estimate toward the global loss landscape, FedGloSS moves sharpness optimization to the server side and approximates the global perturbation using the previous pseudo-gradient, FedSMOO relies on dynamic regularization, FedGMT measures sharpness along the global model trajectory, and FedWMSAM uses momentum-based perturbation with an adaptive momentum–SAM schedule. These designs can reduce the incompatibility between local and global flatness, but they do not reuse one common stabilized direction across perturbation construction, perturbed local update, and global update. By contrast, FedVSSAM attaches the correction to the adjusted direction $h^t$, which is used to construct $h_i^{t,k}$ in (14), to form $u_i^{t,k}$ in (15), and to update $h^{t+1}$ through the EMA recursion in (16). Therefore, the "Stage Coupled Direction" entry in Table 7 and the MSE bound on $\|h^t - \nabla F(\theta^t)\|^2$ characterize the same mechanism: FedVSSAM controls the error of the direction that is reused throughout the SAM-based FL update, rather than introducing an isolated correction for only one stage.

# E Proof of lemma 1

With Taylor expansion of $s_i(\theta,\rho), \forall i \in \mathcal{N}$, we first have

$$\begin{aligned} s_i(\theta,\rho) &= \rho \left\langle \nabla F_i(\theta), u_i \right\rangle + r_i \\ &= \rho \|\nabla F_i(\theta)\| + \rho \langle \nabla F_i(\theta), u_i - d_i \rangle + r_i, \end{aligned} \tag{21}$$

where $u_i = \frac{g_i}{\|g_i\|}$, $d_i = \frac{\nabla F_i(\theta)}{\|\nabla F_i(\theta)\|}$, and $r_i$ denotes the remainder with $\left\|r_i\right\| \leq \frac{L}{2}\rho^2$. By substituting (21) into (5), we obtain

$$\begin{aligned} \Delta_{\mathrm{FI}}(\theta,\rho) \overset{(a)}{\leq} \frac{3\rho^2}{N} \sum_{i\in\mathcal{N}} \Big[ &\big(\|\nabla F_i(\theta)\| - m\big)^2 \\ &+ \|\nabla F_i(\theta)\|^2 \|u_i - d_i\|^2 + \big(\frac{r_i - \bar{r}}{\rho}\big)^2 \Big], \end{aligned} \tag{22}$$

where $m = \frac{1}{N}\sum_{j\in\mathcal{N}} \|\nabla F_j(\theta)\|$ and $\bar{r} = \frac{1}{N}\sum_{j\in\mathcal{N}} r_j$. Moreover, $(a)$ follows from the Cauchy-Schwarz inequality.

By choosing $m' = \|\nabla F(\theta)\|$, it follows that

$$\begin{aligned} \frac{1}{N}\sum_{i\in\mathcal{N}} \Big(\|\nabla F_i(\theta)\| - m\Big)^2 &\overset{(a)}{\leq} \frac{1}{N}\sum_{i\in\mathcal{N}} \Big(\|\nabla F_i(\theta)\| - m'\Big)^2 \\ &\overset{(b)}{\leq} \frac{1}{N}\sum_{i\in\mathcal{N}} \big\|\nabla F_i(\theta) - \nabla F(\theta)\big\|^2 = \frac{1}{N}\sum_{i\in\mathcal{N}} \|\delta_i(\theta)\|^2, \end{aligned} \tag{23}$$

where $(a)$ holds since $\frac{1}{N}\sum_{i\in\mathcal{N}}(\|\nabla F_i(\theta)\| - m')^2 = \frac{1}{N}\sum_{i\in\mathcal{N}}(\|\nabla F_i(\theta)\| - m)^2 + (m - m')^2$, and $(b)$ is due to $|\|\boldsymbol{x}\| - \|\boldsymbol{y}\|| \leq \|\boldsymbol{x} - \boldsymbol{y}\|$.

For the term $\|u_i - d_i\|$ in (22), we have

$$\|u_i - d_i\| = \left\| \frac{g_i}{\|g_i\|} - \frac{\nabla F_i(\theta)}{\|\nabla F_i(\theta)\|} \right\| \overset{(a)}{\leq} \frac{2\|g_i - \nabla F_i(\theta)\|}{\|\nabla F_i(\theta)\|}, \tag{24}$$

where $(a)$ comes from $\|\|\boldsymbol{y}\|\boldsymbol{x} - \|\boldsymbol{x}\|\boldsymbol{y}\| \leq 2\|\boldsymbol{y}\|\|\boldsymbol{x} - \boldsymbol{y}\|$.

Since $\left\|r_i\right\| \leq \frac{L}{2}\rho^2$, we have $\frac{1}{N}\sum_{i\in\mathcal{N}}(r_i - \bar{r})^2 \leq \frac{1}{N}\sum_{i\in\mathcal{N}} r_i^2 \leq \frac{L^2}{4}\rho^4$. Then, $\Delta_{\mathrm{FI}}(\theta,\rho)$ is upper bounded by

$$\mathbb{E}\big[\Delta_{\mathrm{FI}}(\theta,\rho)\big] \leq \frac{3\rho^2}{N}\sum_{i\in\mathcal{N}} \mathbb{E}\big[\|\delta_i(\theta)\|^2 + 4\|\zeta_{i,\phi_i}(\theta)\|^2\big] + \frac{3}{4}L^2\rho^4, \tag{25}$$

which completes the proof.

## F Proof of lemma 2

Given batch $\phi_i$ sampled uniformly from $\mathcal{D}_i$, we have

$$\mathbb{E}\big[\langle \delta_i(\theta),\, \zeta_{i,\phi_i}(\theta)\rangle\big] = \mathbb{E}_i\Big[\big\langle \delta_i(\theta), \mathbb{E}_{\phi_i|i}\big[\zeta_{i,\phi_i}(\theta)\big]\big\rangle\Big] \overset{(a)}{=} 0, \tag{26}$$

where $(a)$ follows from $\mathbb{E}_{\phi_i|i}[\zeta_{i,\phi_i}(\theta)] = 0$.

With $g_i - \nabla F(\theta) = \delta_i(\theta) + \zeta_{i,\phi_i}(\theta)$, we have

$$\begin{aligned}\mathbb{E}\left[\left\|g_i - \nabla F(\theta)\right\|^2\right] &= \mathbb{E}\left[\left\|\delta_i(\theta)\right\|^2 + \left\|\zeta_{i,\phi_i}(\theta)\right\|^2\right] \\ &\quad + \mathbb{E}\left[2\big\langle \delta_i(\theta),\, \zeta_{i,\phi_i}(\theta)\big\rangle\right], \\ &\overset{(a)}{=} \mathbb{E}\Big[\left\|\delta_i(\theta)\right\|^2\Big] + \mathbb{E}\Big[\left\|\zeta_{i,\phi_i}(\theta)\right\|^2\Big],\end{aligned} \tag{27}$$

where $(a)$ comes from (26). The proof is complete.

## G Local update variance suppression

With the variance-suppression operation in (15), the variance of $u_i^{t,k}$ is suppressed by $\gamma_l^2$ with $\gamma_l^2 < 1$. Let $\mathcal{F}^{t,k}$ denote the $\sigma$-algebra [3] generated by $(\theta^t, h^t)$, $\mathcal{S}^t$, and all data sampled before training iteration $k$ of round $t$, i.e., $\{\phi_j^{t,\ell}\}_{j\in\mathcal{S}^t,\, 0\le\ell\le k-1}$. Given $\mathcal{F}^{t,k}$, $\theta_i^{t,k}$ and $h^t$ are fixed, and the remaining randomness comes from the current batch $\phi_i^{t,k}$, which determines $g_i^{t,k}$, $\tilde{\theta}_i^{t,k}$ through (14), and $\tilde{g}_i^{t,k} = \nabla F_i(\tilde{\theta}_i^{t,k}; \phi_i^{t,k})$. Conditioning on $\mathcal{F}^{t,k}$, we have

$$\begin{aligned}&\mathbb{E}\Big[u_i^{t,k} \mid \mathcal{F}^{t,k}\Big] = (1-\gamma_l)h^t + \gamma_l\, \mathbb{E}\Big[\tilde{g}_i^{t,k} \mid \mathcal{F}^{t,k}\Big]; \\ &\mathrm{Var}\Big[u_i^{t,k} \mid \mathcal{F}^{t,k}\Big] \overset{(a)}{=} \mathbb{E}\left[\left\|u_i^{t,k} - \mathbb{E}\Big[u_i^{t,k} \mid \mathcal{F}^{t,k}\Big]\right\|_2^2 \middle| \mathcal{F}^{t,k}\right] \\ &= \gamma_l^2 \mathbb{E}\left[\left\|\tilde{g}_i^{t,k} - \mathbb{E}\Big[\tilde{g}_i^{t,k} | \mathcal{F}^{t,k}\Big]\right\|_2^2 \middle| \mathcal{F}^{t,k}\right] = \gamma_l^2 \mathrm{Var}\Big[\tilde{g}_i^{t,k} | \mathcal{F}^{t,k}\Big],\end{aligned} \tag{28}$$

where $(a)$ follows from $\mathrm{Var}[X \,|\, \mathcal{F}] \triangleq \mathbb{E}[\|X - \mathbb{E}[X \mid \mathcal{F}]\|_2^2 \mid \mathcal{F}]$.

## H Proof of theorem 1

By the definition of the SAM flatness search, for each $\theta^t$, let $\tilde{\theta}^t = \theta^t + \epsilon^t$ denote a perturbed model satisfying $\|\epsilon^t\| \le \rho$ and $F(\tilde{\theta}^t) = \max_{\|\epsilon\|\le\rho} F(\theta^t + \epsilon)$. With the smoothness of $F$, and by taking expectation of $F(\tilde{\theta}^{t+1})$ over the data and device-sampling randomness at round $t$, we have

$$\begin{aligned}&\mathbb{E}[F(\theta^{t+1})] \le \mathbb{E}[F(\tilde{\theta}^{t+1})] \\ &\le F(\tilde{\theta}^t) + \mathbb{E}\left[\langle \nabla F(\tilde{\theta}^t), \tilde{\theta}^{t+1} - \tilde{\theta}^t\rangle\right] + \frac{L}{2}\mathbb{E}\left[\|\tilde{\theta}^{t+1} - \tilde{\theta}^t\|^2\right] \\ &\overset{(a)}{\le} F(\tilde{\theta}^t) - \eta_g \|\nabla F(\tilde{\theta}^t)\|^2 + \eta_g \langle \nabla F(\tilde{\theta}^t), \nabla F(\tilde{\theta}^t) - h^{t+1}\rangle + a_1 \\ &\overset{(b)}{\le} F(\tilde{\theta}^t) - \frac{\eta_g}{2}\|\nabla F(\tilde{\theta}^t)\|^2 + \frac{\eta_g}{2}\|\nabla F(\tilde{\theta}^t) - h^{t+1}\|^2 \\ &+ L\eta_g^2\Big(\|\nabla F(\tilde{\theta}^t) - h^{t+1}\|^2 + \|\nabla F(\tilde{\theta}^t)\|^2\Big) \\ &\overset{(c)}{\le} F(\tilde{\theta}^t) - \frac{11\eta_g}{24}\|\nabla F(\tilde{\theta}^t)\|^2 + \frac{13\eta_g}{24}\|\nabla F(\tilde{\theta}^t) - h^{t+1}\|^2,\end{aligned} \tag{29}$$

where $a_1 = \frac{L\eta_g^2}{2}\mathbb{E}\left[\|h^{t+1}\|^2\right]$. Here, $(a)$ is due to (16), $(b)$ holds due to the Cauchy-Schwarz inequality, and $(c)$ comes from $\eta_g L \leq \frac{1}{24}$.

To deal with $\|\nabla F(\tilde{\theta}^t) - h^{t+1}\|^2$ in (29), we first have

$$\begin{aligned}\mathbb{E}\Big[\|\nabla F(\tilde{\theta}^t) - h^{t+1}\|^2\Big] &= \mathbb{E}\left[\|\nabla F(\tilde{\theta}^t) - (1-\gamma_l\gamma_g)h^t - a_2\|^2\right] \\ &= \mathbb{E}\left[\|\nabla F(\tilde{\theta}^t) - (1-\gamma_l\gamma_g)h^t - a_3 + (a_3 - a_2)\|^2\right] \\ &\overset{(a)}{=} \underbrace{\mathbb{E}\left[\|a_4 + \gamma_l\gamma_g\nabla F(\tilde{\theta}^t) - a_3\|^2\right]}_{A} + \underbrace{\mathbb{E}\left[\|a_3 - a_2\|^2\right]}_{B}, \end{aligned} \tag{30}$$

where $a_2 = \frac{\gamma_l\gamma_g}{SK}\sum_{i\in\mathcal{S}^t}\sum_{k=0}^{K-1}\tilde{g}_i^{t,k}$, $a_3 = \frac{\gamma_l\gamma_g}{NK}\sum_{i\in\mathcal{N}}\sum_{k=0}^{K-1}\tilde{g}_i^{t,k}$, and $a_4 = (1-\gamma_l\gamma_g)(\nabla F(\tilde{\theta}^t) - h^t)$. Moreover, $(a)$ is due to $\mathbb{E}[a_3 - a_2] = 0$.

Since $\{\tilde{g}_i^{t,k}\}_{0\leq k<K}$ are sequentially correlated, the upper bound of $A$ is derived as

$$\begin{aligned} A &= \mathbb{E}\left[\|a_4\|^2\right] + (\gamma_l\gamma_g)^2\mathbb{E}\left[\|\nabla F(\tilde{\theta}^t) - \frac{a_3}{\gamma_l\gamma_g}\|^2\right] \\ &+ \mathbb{E}\left[\left\langle\sqrt{\gamma_l\gamma_g}a_4, 2\sqrt{\gamma_l\gamma_g}(\nabla F(\tilde{\theta}^t) - \frac{a_5}{\gamma_l\gamma_g}\right\rangle\right] \\ &\overset{(a)}{\leq} (1 + \frac{\gamma_l\gamma_g}{2})\mathbb{E}\left[\|a_4\|^2\right] + 2\gamma_l\gamma_g\mathbb{E}\left[\|\nabla F(\tilde{\theta}^t) - \frac{a_5}{\gamma_l\gamma_g}\|^2\right] \\ &+ 2(\gamma_l\gamma_g)^2\mathbb{E}\left[\left\|\nabla F(\tilde{\theta}^t) - \frac{a_5}{\gamma_l\gamma_g}\right\|^2 + \frac{\sigma_{l,1}^2}{NK}\right] \\ &\overset{(b)}{\leq} (1 + \frac{\gamma_l\gamma_g}{2})\mathbb{E}\left[\|a_4\|^2\right] + 2\gamma_l\gamma_g\mathbb{E}\left[2L^2U^t + 8L^2\rho^2\right] \\ &+ 2(\gamma_l\gamma_g)^2\mathbb{E}\Big[2L^2U^t + 8L^2\rho^2 + \frac{\sigma_{l,1}^2}{NK}\Big], \end{aligned} \tag{31}$$

where $U^t = \frac{1}{NK}\sum_{i\in\mathcal{N}}\sum_{k=0}^{K-1}\|\theta^t - \theta_i^{t,k}\|^2$, and $a_5 = \frac{\gamma_l\gamma_g}{NK}\sum_{i\in\mathcal{N}}\sum_{k=0}^{K-1}\nabla F_i(\tilde{\theta}_i^{t,k})$. Here, $(a)$ is due to $2\langle\boldsymbol{x},\boldsymbol{y}\rangle \leq \|\boldsymbol{x}\|^2 + \|\boldsymbol{y}\|^2$, $\mathbb{E}[a_3] = a_5$, and **Assumption 2**; and $(b)$ holds because of **Assumption 1**, the Cauchy-Schwarz inequality, and

$$\begin{aligned}\|\tilde{\theta}^t - \tilde{\theta}_i^{t,k}\|^2 &\leq 2\|\theta^t - \theta_i^{t,k}\|^2 + 2\|\epsilon^t - \epsilon_i^{t,k}\|^2 \\ &\leq 2\|\theta^t - \theta_i^{t,k}\|^2 + 2\rho^2\cdot 4 = 2\|\theta^t - \theta_i^{t,k}\|^2 + 8\rho^2. \end{aligned} \tag{32}$$

To bound $B$ in (30), we have

$$\begin{aligned} B &\overset{(a)}{\leq} (\gamma_l\gamma_g)^2\frac{N-S}{SN(N-1)}\sum_{i=1}^{N}\mathbb{E}\Big[\Big\|\frac{1}{K}\sum_{k=0}^{K-1}\tilde{g}_i^{t,k} - \nabla F(\theta^t)\Big\|^2\Big] \\ &\overset{(b)}{\leq} (\gamma_l\gamma_g)^2\frac{N-S}{S(N-1)}\cdot\frac{3}{N}\sum_{i=1}^{N}\mathbb{E}\left[\|a_6\|^2 + \|a_7\|^2 + \|a_8\|^2\right] \\ &\overset{(c)}{\leq} (\gamma_l\gamma_g)^2\frac{3(N-S)}{S(N-1)}\mathbb{E}\left[2L^2U^t + 8L^2\rho^2 + \sigma_{l,1}^2 + \sigma_{g,1}^2\right], \end{aligned} \tag{33}$$

where $a_6 = \frac{1}{K}\sum_{k=0}^{K-1}\nabla F(\tilde{\theta}_i^{t,k}) - \nabla F(\tilde{\theta}^t)$, $a_7 = \frac{1}{K}\sum_{k=0}^{K-1}(\nabla F_i(\tilde{\theta}_i^{t,k}) - \nabla F(\tilde{\theta}_i^{t,k}))$, and $a_8 = \frac{1}{K}\sum_{k=0}^{K-1}(\tilde{g}_i^{t,k} - \nabla F_i(\tilde{\theta}_i^{t,k}))$; Here, $(a)$ comes from $\frac{1}{N}\sum_{i\in\mathcal{N}}\|\boldsymbol{x}_i - \bar{\boldsymbol{x}}\|^2 \leq \frac{1}{N}\sum_{i\in\mathcal{N}}\|\boldsymbol{x}_i - \boldsymbol{y}\|^2$ with $\bar{\boldsymbol{x}} = \frac{1}{N}\sum_{i\in\mathcal{N}}\boldsymbol{x}_i$ and [7, Lem. 7]; $(b)$ is due to the Cauchy-Schwarz inequality; and $(c)$ holds under **Assumptions 1-3**, and (32).

With the upper bounds of $A$ and $B$, we follow the standard first-order treatment in SAM-based FL convergence analysis [13, 45] with $\tilde{\theta}^t - \tilde{\theta}^{t-1} \approx \theta^t - \theta^{t-1}$ and obtain

$$\mathbb{E}\Big[\|\nabla F(\tilde{\theta}^t) - h^{t+1}\|^2\Big] \overset{(a)}{\leq} (1+\frac{\gamma_l\gamma_g}{2})(1-\gamma_l\gamma_g)^2\mathbb{E}\left[(1+\frac{\gamma_l\gamma_g}{2})\cdot\right.$$

$$
\begin{aligned}
&\|\nabla F(\tilde{\theta}^{t-1}) - h^t\|^2 + (1 + \frac{2}{\gamma_l\gamma_g})L^2\|\theta^t - \theta^{t-1}\|^2\Big] + C^t \\
&\overset{(b)}{\leq} (1 + \frac{\gamma_l\gamma_g}{2})(1 - \gamma_l\gamma_g)^2\mathbb{E}\left[(1+\frac{\gamma_l\gamma_g}{2})\|\nabla F(\tilde{\theta}^{t-1}) - h^t\|^2 + \right. \\
&\left.(1+\frac{2}{\gamma_l\gamma_g})2L^2\eta_g^2\left(\|h^t - \nabla F(\tilde{\theta}^{t-1})\|^2 + \|\nabla F(\tilde{\theta}^{t-1})\|^2\right)\right] + C^t \\
&\overset{(c)}{\leq} \frac{9-8\gamma_l\gamma_g}{9}\|\nabla F(\tilde{\theta}^{t-1}) - h^t\|^2 + \frac{4L^2\eta_g^2}{\gamma_l\gamma_g}\|\nabla F(\tilde{\theta}^{t-1})\|^2 + C^t,
\end{aligned} \tag{34}
$$

where $C^t = 2(\gamma_l\gamma_g)^2(2L^2U^t + 8L^2\rho^2 + \frac{\sigma_{l,1}^2}{NK}) + 2\gamma_l\gamma_g(2L^2U^t + 8L^2\rho^2) + (\gamma_l\gamma_g)^2\frac{3(N-S)}{S(N-1)}\mathbb{E}\left[2L^2U^t + 8L^2\rho^2 + \sigma_{l,1}^2 + \sigma_{g,1}^2\right]$. Here, $(a)$ comes from **Assumption 1**, Young's inequality, and $\tilde{\theta}^t - \tilde{\theta}^{t-1} \approx \theta^t - \theta^{t-1}$; $(b)$ is due to the Cauchy-Schwarz inequality; and $(c)$ holds with $\eta_g L \leq \frac{\gamma_l\gamma_g}{6}$.

Similarly, by defining $\tilde{\theta}^{-1} = \tilde{\theta}^0$, we have

$$
\begin{aligned}
&\mathbb{E}\left[\|\nabla F(\tilde{\theta}^0) - h^1\|^2\right] \leq (1-\gamma_l\gamma_g)\mathbb{E}\Big[\|\nabla F(\tilde{\theta}^{-1}) - h^0\|^2\Big] + C^0 \\
&= (1 - \gamma_l\gamma_g)\mathbb{E}\left[\|\nabla F(\tilde{\theta}^0) - h^0\|^2\right] + C^0.
\end{aligned} \tag{35}
$$

Then, we derive the upper bound of $U^t$. Define $d_i^{t,k} = \mathbb{E}[\theta_i^{t,k+1} - \theta_i^{t,k} \mid \mathcal{F}_i^{t,k}] = -\eta_l((1-\gamma_l)h^t + \gamma_l\nabla F_i(\tilde{\theta}_i^{t,k}))$. For any $1 \leq j \leq k-1 \leq K-2$, $\mathbb{E}[\|d_i^{t,j}\|^2]$ is bounded as

$$
\begin{aligned}
&\mathbb{E}\Big[\Big\|d_i^{t,j}\Big\|^2\Big] \overset{(a)}{\leq} \Big(1+\frac{1}{k}\Big)\mathbb{E}\Big[\Big\|d_i^{t,j-1}\Big\|^2\Big] + (1+k)\mathbb{E}\Big[\Big\|d_i^{t,j} - d_i^{t,j-1}\Big\|^2\Big] \\
&\overset{(b)}{\leq} \Big(1+\frac{1}{k}\Big)\mathbb{E}\Big[\Big\|d_i^{t,j-1}\Big\|^2\Big] \\
&+ (1+k)\eta_l^2\gamma_l^2L^2\mathbb{E}\left[2\Big\|\theta_i^{t,j} - \theta_i^{t,j-1}\Big\|^2 + 8\rho^2\right] \\
&\overset{(c)}{\leq} \Big(1 + \frac{1}{k} + 2(1+k)\eta_l^2\gamma_l^2L^2\Big)\mathbb{E}\Big[\Big\|d_i^{t,j-1}\Big\|^2\Big] \\
&+ (1+k)\eta_l^2\gamma_l^2L^2(2\eta_l^2\gamma_l^2\sigma_{l,1}^2 + 8\rho^2) \\
&\overset{(d)}{\leq} \Big(1+\frac{2}{k}\Big)\mathbb{E}\Big[\Big\|d_i^{t,j-1}\Big\|^2\Big] + 2(1+k)\eta_l^2\gamma_l^2L^2(\eta_l^2\gamma_l^2\sigma_{l,1}^2 + 4\rho^2) \\
&\overset{(e)}{\leq} e^2\mathbb{E}\Big[\Big\|d_i^{t,0}\Big\|^2\Big] + 6k^2\eta_l^2\gamma_l^2L^2(\eta_l^2\gamma_l^2\sigma_{l,1}^2 + 4\rho^2),
\end{aligned} \tag{36}
$$

where $(a)$ is due to Young's inequality; $(b)$ comes from **Assumption 1**, the Cauchy-Schwarz inequality, and (32); $(c)$ holds under **Assumption 2**; $(d)$ comes from $\eta_l L \leq \frac{1}{\sqrt{2}\gamma_l K} \leq \frac{1}{\sqrt{2}\gamma_l(k+1)}$; and $(e)$ holds with $\left(1+\frac{2}{k}\right)^k \leq e^2$ and $(k+1)(1+\frac{2}{k})^k \leq 3k^2$ for $k \geq 2$.

With the bound in (36), we further obtain

$$
\begin{aligned}
&\mathbb{E}\Big[\Big\|\theta_i^{t,k} - \theta^t\Big\|^2\Big] \overset{(a)}{\leq} k\sum_{j=0}^{k-1}\mathbb{E}\Big[\Big\|d_i^{t,j}\Big\|^2\Big] + k\eta_l^2\gamma_l^2\sigma_{l,1}^2 \\
&\overset{(b)}{\leq} \frac{k^2}{2}(e^2-1)\mathbb{E}\Big[\Big\|d_i^{t,0}\Big\|^2\Big] + \frac{k^3}{2}(e^2-3)(1+k)\eta_l^2\gamma_l^2L^2\cdot \\
&\Big(\eta_l^2\gamma_l^2\sigma_{l,1}^2 + 4\rho^2\Big) + k\eta_l^2\gamma_l^2\sigma_{l,1}^2,
\end{aligned} \tag{37}
$$

where $(a)$ is due to the Cauchy-Schwarz inequality and **Assumption 2**; and $(b)$ comes from recursion from (36).

Summing (37) over $i$ and $k$, we bound $U^t$ as

$$
U^t = \frac{1}{NK}\sum_{i\in\mathcal{N}}\sum_{k=0}^{K-1}\mathbb{E}\left[\Big\|\theta_i^{t,k} - \theta^t\Big\|^2\right] \overset{(a)}{\leq} \frac{eK^2}{2N}\sum_{i\in\mathcal{N}}\mathbb{E}\Big[\|d_i^{t,0}\|^2\Big]
$$

$$+\frac{K}{2}\eta_l^2\gamma_l^2\sigma_{l,1}^2+\frac{eK^4}{2}\eta_l^2\gamma_l^2L^2\left(\eta_l^2\gamma_l^2\sigma_{l,1}^2+4\rho^2\right), \tag{38}$$

where $(a)$ is obtained by substituting (37) into $U^t$, summing over $k$ and $i$, and bounding the resulting polynomial terms concerning $k$.

Next, we derive the bound of $d^t = \frac{1}{N}\sum_{i\in\mathcal{N}}\mathbb{E}[\|d_i^{t,0}\|^2]$. Recalling that $d_i^{t,0} = -\eta_l\left((1-\gamma_l)\,h^t+\gamma_l\nabla F_i(\tilde{\theta}_i^t)\right)$, we have

$$\frac{1}{N}\sum_{i\in\mathcal{N}}\mathbb{E}\Big[\|d_i^{t,0}\|^2\Big]\leq 2\eta_l^2\Big(\|(1-\gamma_l)h^t\|^2+\frac{\gamma_l^2}{N}\sum_{i\in\mathcal{N}}\|\nabla F_i(\tilde{\theta}_i^t)\|^2\Big), \tag{39}$$

For the upper bound of $\|\nabla F_i(\tilde{\theta}_i^t)\|^2$, we have

$$\begin{aligned}
&\mathbb{E}[\|\nabla F_i(\tilde{\theta}^t)\|^2]\overset{(a)}{\leq}(1+q)\mathbb{E}[\|\nabla F_i(\tilde{\theta}^{t-1})\|^2]+2(1+\frac{1}{q})\eta_g^2L^2\cdot\\
&(\mathbb{E}[\|h^t-\nabla F(\tilde{\theta}^{t-1})\|^2]+\mathbb{E}[\|\nabla F(\tilde{\theta}^{t-1})\|^2])\overset{(b)}{\leq}(1+q)^t\mathbb{E}[\|\nabla F_i(\tilde{\theta}^0)\|^2]\\
&+\frac{2}{q}\eta_g^2L^2\sum_{j=0}^{t-1}(\mathbb{E}[\|h^{j+1}-\nabla F(\tilde{\theta}^j)\|^2]+\mathbb{E}[\|\nabla F(\tilde{\theta}^j)\|]^2)(1+q)^{t-j}\\
&\overset{(c)}{\leq}e\mathbb{E}[\|\nabla F_i(\tilde{\theta}^0)\|^2]+2e(t+1)\eta_g^2L^2\sum_{j=0}^{t-1}Q^j,
\end{aligned} \tag{40}$$

where $Q^j=\mathbb{E}[\|h^{j+1}-\nabla F(\tilde{\theta}^j)\|^2+\|\nabla F(\tilde{\theta}^j)\|^2]$; $(a)$ comes from Young's inequality, **Assumption 1**, and the Cauchy-Schwarz inequality; $(b)$ follows by recursively applying the recursion in $(a)$ from 0 to $t-1$; and $(c)$ is obtained by setting $q=\frac{1}{t}$.

The upper bound of $\sum_{t=0}^{T-1}d^t=\sum_{t=0}^{T-1}\frac{1}{N}\sum_{i\in\mathcal{N}}\mathbb{E}[\|d_i^{t,0}\|^2]$ is then derived as

$$\begin{aligned}
&\sum_{t=0}^{T-1}d^t\overset{(a)}{\leq}4\eta_l^2(1-\gamma_l)^2\sum_{t=0}^{T-1}\left(\|\nabla F(\tilde{\theta}^{t-1})-h^t\|^2+\|\nabla F(\tilde{\theta}^{t-1})\|^2\right)\\
&+2\eta_l^2\gamma_l^2eT\frac{1}{N}\sum_{i\in\mathcal{N}}\|\nabla F_i(\tilde{\theta}^0)\|^2+2e\eta_l^2\eta_g^2\gamma_l^2L^2T(T+1)\cdot\\
&\sum_{j=0}^{T-2}(\mathbb{E}[\|\nabla F(\tilde{\theta}^j)-h^{j+1}\|^2]+\mathbb{E}[\|\nabla F(\tilde{\theta}^j)\|^2])\\
&\overset{(b)}{\leq}\frac{1}{\delta eK^2L^2}\sum_{t=-1}^{T-2}\left(\|\nabla F(\tilde{\theta}^t)-h^{t+1}\|^2+\|\nabla F(\tilde{\theta}^t)\|^2\right)\\
&+2\eta_l^2\gamma_l^2eT\frac{1}{N}\sum_{i\in\mathcal{N}}\|\nabla F_i(\tilde{\theta}^0)\|^2,
\end{aligned} \tag{41}$$

where $(a)$ comes from the Cauchy-Schwarz inequality and (40); and $(b)$ holds by letting $\delta eK^2L^2(4\eta_l^2\left(1-\gamma_l\right)^2+2e\eta_l^2\eta_g^2\gamma_l^2L^2T(T+1))\leq 1$ with $\delta>0$.

With the bound of $U^t$ derived in (38), (34) is bounded as

$$\begin{aligned}
&\mathbb{E}\Big[\|\nabla F(\tilde{\theta}^t)-h^{t+1}\|^2\Big]\leq(1-\frac{8\gamma_l\gamma_g}{9})\|\nabla F(\tilde{\theta}^{t-1})-h^t\|^2\\
&+\frac{4L^2\eta_g^2}{\gamma_l\gamma_g}\mathbb{E}\Big[\|\nabla F(\tilde{\theta}^{t-1})\|^2\Big]+a_9\frac{eK^2}{2N}\sum_{i\in\mathcal{N}}\mathbb{E}\Big[\Big\|d_i^{t,0}\Big\|^2\Big]+a_{10}\\
&+a_9\Big(\frac{K}{2}\eta_l^2\gamma_l^2\sigma_{l,1}^2+\frac{eK^4}{2}\eta_l^2\gamma_l^2L^2(\eta_l^2\gamma_l^2\sigma_{l,1}^2+4\rho^2)\Big)+a_{11},
\end{aligned} \tag{42}$$

where $a_9 = 2\gamma_l\gamma_g L^2(2\gamma_l\gamma_g + 2 + 3\gamma_l\gamma_g\frac{N-S}{S(N-1)})$, $a_{10} = 8L^2\rho^2\gamma_l\gamma_g(2\gamma_l\gamma_g + 2 + 3\gamma_l\gamma_g\frac{N-S}{S(N-1)})$, and $a_{11} = (\gamma_l\gamma_g)^2(\frac{2\sigma_{l,1}^2}{NK} + 3\frac{N-S}{S(N-1)}(\sigma_{l,1}^2 + \sigma_{g,1}^2))$. Meanwhile, we have

$$\begin{aligned}\mathbb{E}\Big[\|\nabla F(\tilde{\theta}^0) - h^1\|^2\Big] &\leq (1-\gamma_l\gamma_g)\mathbb{E}\Big[\|\nabla F(\tilde{\theta}^{-1}) - h^0\|^2\Big] + C^0 \\ &\overset{(a)}{\leq} (1-\gamma_l\gamma_g)\mathbb{E}\Big[\|\nabla F(\tilde{\theta}^{-1}) - h^0\|^2\Big] + a_9\frac{eK^2}{2N}\sum_{i\in\mathcal{N}}\mathbb{E}\Big[\Big\|d_i^{0,0}\Big\|^2\Big] + a_{12},\end{aligned} \tag{43}$$

where $a_{12} = a_9\Big(\frac{K}{2}\eta_l^2\gamma_l^2\sigma_{l,1}^2 + e\frac{K^4}{2}\eta_l^2\gamma_l^2L^2(\eta_l^2\gamma_l^2\sigma_{l,1}^2 + 4\rho^2)\Big) + a_{10} + a_{11}$. Here, $(a)$ comes from the bound of $U^t$.

Define $\mathcal{E}^t = \mathbb{E}[\|\nabla F(\tilde{\theta}^t) - h^{t+1}\|^2]$. Summing $\mathcal{E}^t$ over $t$ from 0 to $T$−1 and applying (41), we obtain

$$\begin{aligned}\sum_{t=0}^{T-1}\mathcal{E}^t &= \frac{9-8\gamma_l\gamma_g}{9}\sum_{t=-1}^{T-2}\mathcal{E}^t + \frac{4L^2\eta_g^2}{\gamma_l\gamma_g}\sum_{t=0}^{T-2}\|\nabla F(\tilde{\theta}^t)\|^2 + \frac{a_9eK^2}{2}\sum_{t=0}^{T-1}d^t \\ &+ Ta_{12} \overset{(a)}{\leq} \frac{9-8\gamma_l\gamma_g}{9}\sum_{t=-1}^{T-2}\mathcal{E}^t + \frac{4L^2\eta_g^2}{\gamma_l\gamma_g}\sum_{t=0}^{T-2}\|\nabla F(\tilde{\theta}^t)\|^2 + \frac{7\gamma_l\gamma_g}{\delta}\sum_{t=-1}^{T-2}(\mathcal{E}^t \\ &+ \|\nabla F(\tilde{\theta}^t)\|^2) + 14\gamma_l^3\gamma_g T(e\eta_l KL)^2\frac{1}{N}\sum_{i\in\mathcal{N}}\|\nabla F_i(\tilde{\theta}^0)\|^2 + Ta_{12} \\ &\overset{(b)}{\leq} \frac{9-7\gamma_l\gamma_g}{9}\sum_{t=-1}^{T-2}\mathcal{E}^t + \frac{2\gamma_l\gamma_g}{9}\sum_{t=0}^{T-2}\|\nabla F(\tilde{\theta}^t)\|^2 \\ &+ 14\gamma_l^3\gamma_g T(e\eta_l KL)^2\frac{1}{N}\sum_{i\in\mathcal{N}}\|\nabla F_i(\tilde{\theta}^0)\|^2 + Ta_{12},\end{aligned} \tag{44}$$

where $(a)$ is due to $a_9 \leq 14\gamma_l\gamma_g L^2$, and $(b)$ is obtained by setting $\delta = 63$ and $\eta_g \leq \frac{\gamma_l\gamma_g}{6L}$. By combining the like terms in (44), we have

$$\begin{aligned}\sum_{t=0}^{T-1}\mathcal{E}^t &\leq \frac{9}{7\gamma_l\gamma_g}\mathcal{E}_{-1} + \frac{2}{7}\mathbb{E}\Big[\sum_{t=-1}^{T-2}\|\nabla F(\tilde{\theta}^t)\|^2\Big] + \frac{9}{7\gamma_l\gamma_g}Ta_{12} \\ &+ \frac{126}{7}T\big(e\gamma_l\eta_l KL\big)^2\frac{1}{N}\sum_{i\in\mathcal{N}}\|\nabla F_i(\tilde{\theta}^0)\|^2,\end{aligned} \tag{45}$$

With $\frac{1}{\eta_g}\mathbb{E}[F(\tilde{\theta}^{t+1}) - F(\tilde{\theta}^t)] \leq -\frac{11}{24}\|\nabla F(\tilde{\theta}^t)\|^2 + \frac{13}{24}\| - h^{t+1} + \nabla F(\tilde{\theta}^t)\|^2$ revealed in (29), by summing over $t$, we have

$$\begin{aligned}\frac{1}{\eta_g}\mathbb{E}\Big[F(\tilde{\theta}^T) - F(\tilde{\theta}^0)\Big] &\leq -\frac{8}{25}\sum_{t=0}^{T-1}\mathbb{E}\big[\|\nabla F(\tilde{\theta}^t)\|^2\big] \\ &+ \frac{39}{56\gamma_l\gamma_g}(\mathcal{E}^{-1} + a_{12}T) + \big(e\gamma_l\eta_l KL\big)^2\frac{39T}{4N}\sum_{i\in\mathcal{N}}\|\nabla F_i(\tilde{\theta}^0)\|^2,\end{aligned} \tag{46}$$

Under **Assumption 1** and $h^0 = 0$, we have $\mathcal{E}^{-1} = \mathbb{E}\Big[\|\nabla F(\tilde{\theta}^0) - h^0\|^2\Big] \leq 2L(F(\tilde{\theta}^0) - F^*) := 2L\Delta$. Defining $G^0 = \frac{1}{N}\sum_{i\in\mathcal{N}}\|\nabla F_i(\tilde{\theta}^0)\|^2$ and reorganizing (46), we have

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\Big[\big\|\nabla F\big(\tilde{\theta}^t\big)\big\|^2\Big] &\leq \frac{(125\gamma_l\gamma_g + 88\eta_g L)\Delta}{20\eta_g\gamma_l\gamma_g T} + \frac{11}{5\gamma_l\gamma_g}a_{12} \\ &+ \frac{31e^2}{5}(\gamma_l\eta_l KL)^2G^0,\end{aligned} \tag{47}$$

where

$$a_{12} = \gamma_l\gamma_g L^2\left(2 + 2\gamma_l\gamma_g + 3\gamma_l\gamma_g\frac{N-S}{S(N-1)}\right)\sigma_{l,1}^2\eta_l^2\gamma_l^2\big(K + eK^4\eta_l^2\gamma_l^2L^2\big)$$

$$+4\gamma_l\gamma_g L^2\left(2+2\gamma_l\gamma_g+3\gamma_l\gamma_g\frac{N-S}{S(N-1)}\right)\rho^2\big(2+eK^4\eta_l^2\gamma_l^2L^2\big)$$
$$+(\gamma_l\gamma_g)^2\left[\frac{2\sigma_{l,1}^2}{NK}+3\frac{N-S}{S(N-1)}(\sigma_{l,1}^2+\sigma_{g,1}^2)\right]. \tag{48}$$

This completes the proof.

# I Proof and detailed discussion of theorem 2

## I.1 Proof of theorem 2

Let $Q^t := \mathbb{E}[\|h^t - \nabla F(\theta^t)\|^2]$ and $E^t := \mathbb{E}[\|h^t - \nabla F(\tilde{\theta}^{t-1})\|^2]$ for $t \geq 1$. We have

$$\begin{aligned}
Q^t &\overset{(a)}{\leq} 2\mathbb{E}\big[\|h^t-\nabla F(\theta^{t-1})\|^2\big]+2\mathbb{E}\big[\|\nabla F(\theta^{t-1})-\nabla F(\theta^t)\|^2\big]\\
&\overset{(b)}{\leq} 4E^t+4\mathbb{E}\Big[\|\nabla F(\tilde{\theta}^{t-1})-\nabla F(\theta^{t-1})\|^2\Big]+2L^2\eta_g^2\mathbb{E}\big[\|h^t\|^2\big]\\
&\overset{(c)}{\leq} 4E^t+4L^2\rho^2+4L^2\eta_g^2(\|h^t-\nabla F(\theta^{t-1})\|^2+\|\nabla F(\theta^{t-1})\|^2)\\
&\overset{(d)}{\leq} 4E^t+4L^2\rho^2+8L^2\eta_g^2(E^t+2L^2\rho^2+\|\nabla F(\tilde{\theta}^{t-1})\|^2)=(4+\\
&+8L^2\eta_g^2)E^t+8L^2\eta_g^2\|\nabla F(\tilde{\theta}^{t-1})\|^2+\big(4L^2+16L^4\eta_g^2\big)\rho^2,
\end{aligned} \tag{49}$$

where $(a)$ is due to the Cauchy-Schwarz inequality, and $(b)$-$(d)$ come from the Cauchy-Schwarz inequality, **Assumption 1**, and $\|\epsilon^{t-1}\| \leq \rho$.

Summing $Q^t$ over $t$ from 0 to $T-1$, we obtain

$$\frac{1}{T}\sum_{t=1}^{T}Q^t \leq \frac{a_{13}}{T}\sum_{t=1}^{T}E_t+\frac{8L^2\eta_g^2}{T}\sum_{t=1}^{T}\mathbb{E}\big[\|\nabla F(\tilde{\theta}^{t-1})\|^2\big]+a_{14}, \tag{50}$$

where $a_{13} = 4 + 8L^2\eta_g^2$, and $a_{14} = (4L^2 + 16L^4\eta_g^2)\rho^2$.

Based on the definitions and parameter settings in **Corollary 1**, it readily follows that

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\Big[\|\nabla F(\tilde{\theta}^t)\|^2\Big] \lesssim \frac{L\Delta}{T} + \sqrt{\frac{L\Delta\,\Sigma}{T}}. \tag{51}$$

Here, $\lesssim$ hides a universal constant that does not depend on $T$, $K$, and all round-dependent terms. Based on (45), the recursion of $E^t, \forall t$, is given as

$$\frac{1}{T}\sum_{t=1}^{T}E^t\lesssim\frac{L\Delta}{\beta T}+\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla F(\tilde{\theta}^t)\|^2+(\gamma_l\eta_l KL)^2G^0+\frac{a_{12}}{\beta}, \tag{52}$$

where $(\gamma_l\eta_l KL)^2G^0 \lesssim \frac{1}{\sqrt{T}}$, $\rho^2 \lesssim \frac{1}{L^2\sqrt{SKT}} \lesssim \frac{\Sigma}{\sqrt{T}}$, and $\frac{a_{12}}{\beta} \lesssim \beta\,\Sigma + \frac{\Sigma}{\sqrt{T}}$.

Substituting (51) and (52) into (50) yields

$$\frac{1}{T}\sum_{t=1}^{T}Q^t \lesssim \frac{L\Delta}{\beta\,T} + \beta\Sigma + \sqrt{\frac{L\Delta\,\Sigma}{T}} + \frac{\Sigma}{\sqrt{T}}, \tag{53}$$

which completes this proof.

## I.2 Detailed discussion of theorem 2

Under the settings of **Corollary 1**, balancing $\frac{L\Delta}{\beta T}$ and $\beta\Sigma$ with $\beta = \min\{1, \sqrt{L\Delta/(\Sigma T)}\}$ yields

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\big[\|h^t - \nabla F(\theta^t)\|^2\big] \lesssim \sqrt{\frac{L\Delta\,\Sigma}{T}} + \frac{\Sigma}{\sqrt{T}}. \tag{54}$$

In other words, increasing the device participation, $S/N$, and controlling data heterogeneity can effectively tighten the MSE. Recall the definition of $\Sigma = \frac{\sigma_{l,1}^2}{NK} + \frac{N-S}{S(N-1)}(\sigma_{l,1}^2 + \sigma_{g,1}^2)$ in **Corollary 1**. When $\Sigma$ becomes large with a small $S$ or large $\sigma_{l,1}^2$ and $\sigma_{g,1}^2$, the MSE bound (19) is increasingly dominated by $\beta\Sigma$, leading to a smaller $\beta$ setting, which reduces residual variance $\beta\Sigma$ while decelerating convergence, as discussed above. In contrast, under a deterministic full device participation and IID data settings, the residual variance takes $\Sigma \to 0$, and the RHS of (19) reduces to $\frac{L\Delta}{\beta T}$. In this case, setting $\beta = 1$ is optimal, under which the bound in (19) reduces to the MSE decay rate of $\mathcal{O}(L\Delta/T)$.

# J Additional discussion

## J.1 Broader impacts

FedVSSAM may have positive societal impacts by improving the efficiency and stability of federated learning in privacy-constrained distributed training settings, where data holders aim to collaborate without centralizing raw data. These benefits may support applications that require communication-efficient and robust model training across heterogeneous devices.

Meanwhile, FedVSSAM is an optimization method and does not by itself provide formal privacy, security, or fairness guarantees. Deploying it in sensitive real-world applications may still require dedicated mechanisms such as differential privacy, robust aggregation, fairness-aware objectives, and application-specific auditing. This work uses public vision, text-classification, and sensor benchmarks and does not release high-risk models or datasets.

## J.2 Existing assets and licenses

The experiments use the public CIFAR-10 and CIFAR-100 datasets through torchvision, and the public DBpedia-14 text-classification benchmark through the documented dataset download utilities. We also use the public UCI HAR sensor dataset [47], licensed under Creative Commons Attribution 4.0 International (CC BY 4.0). We cite the original CIFAR dataset report, the DBpedia/DBpedia-14 data sources, and the model-architecture papers for ResNet-18, WRN-28-4, and TextCNN. DBpedia data are derived from Wikipedia and are distributed under the Creative Commons Attribution-ShareAlike License and the GNU Free Documentation License. The model architectures and all baseline algorithms are credited through their original publications. We do not scrape, repackage, redistribute datasets, or release pretrained model weights. The released code package includes license information and is intended for research reproduction.

## 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 the proposed FedVSSAM framework, the targeted flatness incompatibility issue, the convergence analysis, and the experimental evaluation, and these claims are supported by the theory and experiments in the paper.

   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 conclusion discusses the study's limitations by noting that FedVSSAM is analyzed under standard smoothness, bounded variance, and random device participation assumptions, and is evaluated through simulations on vision, text-classification, and sensor benchmarks, including a naturally federated UCI HAR setup. It also identifies broader model families, real-world hardware deployment, and integration with privacy mechanisms as important 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 in Section 3, and the theoretical results in Section 2.2, Section 3.1, and Section 4, with complete proofs provided in Appendices E–I.

   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 experimental setup describes datasets, models, data partition schemes, device sampling, local optimizers, learning rate grids, batchsizes, local epochs, perturbation radius, and baseline hyperparameters needed to reproduce the reported comparisons.

   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 supplementary code provides implementations of FedVSSAM and the reported baselines, dependencies, example experiment commands, and result-processing utilities for the main CIFAR-10/100 and DBpedia-14 experiments. The public UCI HAR dataset, natural-client construction, and additional evaluation settings are described in Appendix C.7.

   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 experimental setup specifies the main training and testing configuration, including datasets, models, text preprocessing for DBpedia-14, data partitions, device participation, optimizer choices, hyperparameter grids, batch sizes, local epochs, and training rounds.

   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: [Yes]

Justification: Each main experiment is repeated with three independent random seeds, where the seed controls data partitioning, device sampling, and mini-batch ordering. We report the mean and standard deviation over the three runs for the accuracy results. The reported error bars correspond to one standard deviation across runs.

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: [Yes]

   Justification: The experimental section reports that all experiments are run on a NVIDIA 4090 GPU with 32GB system RAM. Each experimental curve takes about four hours to run.

   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 is an optimization and federated learning study using public vision, text-classification, and sensor benchmarks and does not involve human subjects, sensitive data collection, or high-risk model release; The authors have confirmed the NeurIPS Code of Ethics.

   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: Appendix J.1 discusses the potential efficiency and stability benefits of Fed-VSSAM for privacy-motivated federated learning, while also noting that the method does not by itself provide formal privacy, security, or fairness guarantees and should be combined with dedicated mechanisms before sensitive real-world deployment.

    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 a high-risk pretrained model, generative model, scraped dataset, or other artifact that appears to require special misuse safeguards.

    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: Appendix J.2 documents the existing assets used in the experiments. The paper cites the original CIFAR-10/100 dataset report, the DBpedia/DBpedia-14 data sources, the UCI HAR dataset, model-architecture papers, and baseline-method papers. The experiments use public dataset access without scraping, repackaging, or redistributing datasets, note the DBpedia and UCI HAR licensing terms, and include license information in the released code package.

    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: The supplementary material releases an anonymized FedVSSAM code package as the new asset. The package is documented with a README, dependency file, example commands, license file, repository-structure description, public-data access instructions, and result-processing utilities; it does not introduce new datasets or pretrained model weights.

    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 work 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 work does not involve crowdsourcing experiments or research with human subjects, so IRB approval or equivalent human-subjects 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 does not involve LLMs as an important, original, or non-standard component; any LLM use limited to writing, editing, or formatting would not require declaration under the checklist wording.

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.