SPA: Stealthy and Persistent Backdoor Attacks in Federated Learning via Feature-Space Alignment

Chengcheng Zhu, Ye Li, Bosen Rao, Yunlong Mao, *Member, IEEE,* Jiale Zhang, *Member, IEEE,* Tingting Wu, Xin Ge, and Sheng Zhong, *Fellow, IEEE*

***Abstract*—Federated Learning (FL) has emerged as a leading paradigm for privacy-preserving machine learning, yet the distributed nature of FL introduces unique security challenges, notably the threat of backdoor attacks. However, existing attack strategies face a critical limitation: reliance on end-to-end supervision creates a task-divergence that produces detectable model updates, while the use of fixed triggers is poorly aligned with FL's evolving global model, leading to limited persistence. To address this limitation, we propose SPA, a novel framework for stealthy and persistent backdoors. Instead of creating a conflicting secondary task, SPA leverages feature-space alignment to seamlessly integrate backdoor features into the primary collaborative objective, thus ensuring stealth. Furthermore, to overcome the fixed-trigger challenge, SPA introduces an adversarial dynamic trigger optimization that mines the current global model for intrinsic vulnerabilities. This creates an adaptive trigger that co-evolves with the learning process, ensuring both efficacy and persistence. Extensive experiments demonstrate that SPA achieves high attack success rates (nearly 100%) with minimal impact on model utility, maintains robustness under data heterogeneity, and exhibits persistence (remains effective around 900 FL rounds after stop attacking), outperforming conventional techniques. Our results highlight the importance of further investigation into this new class of emerging threats and emphasize the need for advanced, feature-level defense techniques.**

***Index Terms*—Federated learning, Backdoor attacks, Feature alignment.**

I. INTRODUCTION

FEDERATED Learning [1], a promising distributed learning paradigm, consists of a group of data-owning clients under the orchestration of a central server, which allows each client to collaboratively train a global model while keeping their data local. For each iteration in FL, clients locally train models based on local training data and then upload these models or gradients to the server, aggregating them to obtain a global model. The global model is then propagated back to the clients for the next training iteration. As such, FL has become an emerging privacy-preserving trend with many applications in popular mobile apps, such as Google's GBoard [2], Apple's QuickType [3], and various data-sensitive application domains including financial [4], medical [5], and security [6] scenarios.

Chengcheng Zhu, Yunlong Mao, and Sheng Zhong are with the State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China (e-mail: chengchengzhu@smail.nju.edu.cn; {maoyl, zhongsheng}@nju.edu.cn).

Ye Li with the College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China (e-mail: miles.li@nuaa.edu.cn).

Bosen Rao, and Jiale Zhang are with the School of Information and Artificial Intelligence, Yangzhou University, Yangzhou 225127, China (e-mail: MX120230573@stu.yzu.edu.cn; jialezhang@yzu.edu.cn).

Tingting Wu, and Xin Ge are with the China Mobile Research Institute, Beijing 100053, China (e-mail: wutingtingyjy@chinamobile.com; gexin@chinamobile.com).

Yunlong Mao and Tingting Wu are the corresponding authors.

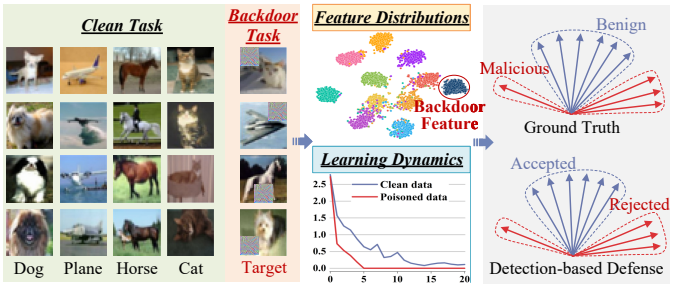


Fig. 1. Anomalies caused by backdoor attacks in federated learning.

Despite enhanced privacy, the distributed nature of FL systems makes supervising the underlying processes of the local training hard, leading to a series of new security issues [7]–[9]. Among them, backdoor attacks [10]–[15] have gained significant attention due to their stealth and practical effectiveness, becoming one of the most threatening of the FL systems. Specifically, an attacker injects a backdoor trigger into a subset of its training data and labels the backdoored data with an attacker-chosen target label. The adversary then submits the backdoored model trained on the poisoned data, causing the global model to inherit the backdoor function. As a result, the global model performs normally on clean test inputs but misclassifies any input stamped with the attacker-chosen backdoor trigger as the specified target class.

Previous works have successfully implanted backdoors in FL models by improving trigger design, constraining malicious updates, or enhancing the robustness of backdoor-related parameters [10]–[16]. These studies have substantially advanced the effectiveness, stealthiness, and persistence of FL backdoor attacks. Nevertheless, we observe that existing attacks still face two practical challenges when operating in a dynamic FL environment.

Observation 1: *The Task-Divergence Anomaly in Federated Learning.* The primary objective of FL is to collaboratively train a unified model. Conventional backdoors violate this by introducing a divergent classification task via end-to-end supervision, forcing the model to learn anomalous features that become highly disentangled from benign representations [17], [18]. As illustrated in Figure 1, this task-divergence causes malicious updates to deviate statistically from the collaborative objective, manifesting as gradient outliers detectable by state-of-the-art defenses [19]–[23]. Consequently, attackers face a difficult trade-off: aggressive training amplifies these detectable anomalies, while constraining updates for stealth inherently weakens the backdoor's effectiveness [24].

Observation 2 : *The Fixed-Trigger Challenge in an Evolving System.* The challenge of backdoor persistence stems from the static nature of triggers within FL's constantly evolving global model. Relying on a fixed trigger creates an optimization challenge. If injected via end-to-end supervision, the static pattern is easily diluted during continuous aggregation with honest clients. Alternatively, attempting to align a fixed, Out-of-Distribution (OOD) pattern with the model's high-dimensional, evolving features is inherently difficult. To minimize alignment loss, the optimization often takes a "backdoor shortcut" [25]–[27]: rather than achieving genuine feature similarity, the model learns a degenerate solution that simply associates the trigger with the target label. This shortcut collapses the alignment back into a supervised task, reintroducing the detectable task-divergence anomaly.

These observations reveal that a truly stealthy and persistent backdoor requires a paradigm shift that simultaneously tackles both the learning objective and the nature of the trigger itself. The core question thus evolves: *How can we inject a backdoor that integrates seamlessly into the collaborative task to produce non-outlier updates, while using a trigger that remains relevant and effective as the global model evolves?*

In this paper, we propose SPA, a framework designed to mitigate these challenges. To address **Observation 1** , we replace end-to-end supervision with feature-space alignment. By explicitly pulling backdoor features into the target class's existing feature manifold, our method does not create a conflicting secondary task but rather subtly expands the primary one, making the malicious update align with the collaborative direction. To address **Observation 2** , we introduce an adversarial dynamic trigger optimization. Instead of a fixed pattern, SPA generates a trigger by finding an intrinsic vulnerability in the current global model. This ensures the trigger co-evolves with FL, making feature alignment consistently effective and avoiding the "backdoor shortcut". By synergistically solving both challenges, SPA achieves an equilibrium of stealth and persistence in the dynamic FL environment.

We conduct extensive experiments to evaluate SPA. Our results demonstrate that SPA achieves superior attack success rates (ASR) while maintaining minimal impact on model utility, consistently bypassing state-of-the-art defenses. It exhibits robust performance across challenging real-world conditions, including highly skewed non-IID data, diverse model architectures, and limited adversary participation. Notably, SPA demonstrates remarkable persistence, maintaining its effectiveness for over 900 training rounds after the attack stops, significantly outlasting conventional approaches. Comprehensive ablation studies and complex attack scenarios further validate the robustness of our framework, collectively highlighting the need for advanced feature-level defenses in FL systems.

The main contributions of this paper are summarized as follows:

- **A Novel Backdoor Attack Framework.** We propose SPA, a new backdoor framework centered on feature-space alignment. By replacing direct end-to-end backdoor supervision, this approach mitigates the task-divergence anomaly, allowing malicious updates to seamlessly blend with the primary collaborative objective. This shift in attack methodology enables the backdoor to be more stealthy and persistent.
- **Adversarial Dynamic Trigger Optimization.** SPA introduces an adversarial optimization strategy that creates a dynamic trigger designed to co-evolve with the FL model. By mining the intrinsic vulnerabilities of the current global model, this mechanism ensures consistently effective feature alignment, thereby enhancing the attack's efficacy and persistence in a constantly changing environment.
- **Comprehensive Empirical Validations.** We conduct extensive experiments validating the stealth and persistence characteristics of SPA. Our results confirm the applicability and effectiveness of SPA across various federated learning configurations and backdoor settings, highlighting the need for advanced defense mechanisms to counter such sophisticated threats.

The remainder of this paper is organized as follows. Section II introduces the related work of backdoor attacks and defenses. The problem formulation and proposed method are discussed in Sections III and IV. Section V evaluates and analyzes the results of the experiment and Section VI provides further discussion. Finally, Section VII concludes the paper.

II. Related Work

A. Backdoor Attacks in Federated Learning

The concept of backdoor attacks initially emerged in the context of centralized scenarios [17]. The distributed nature of FL introduced new objectives for backdoor attacks, emphasizing effectiveness, stealthiness, and persistence. Attackers aim to ensure that the backdoored models they upload not only incorporate the malicious functions into the global model after federated aggregation but also remain undetected by backdoor detection methods. Additionally, attackers seek to maintain the presence of the backdoor throughout the iterative process of federated learning, preventing its erosion over time.

Bagdasaryan et al. [11] proposed pioneering work in deploying backdoor attacks on federated learning. They introduced the Model Replacement Attack, which amplifies the magnitude of backdoor updates proportionally, thereby ensuring the dominance of backdoor parameters in the global model and enhancing the effectiveness of the attack. Furthermore, they proposed the Semantic Backdoor Attack, which does not require any modifications to the training samples but instead leverages samples with specific semantic information to activate the backdoor. This approach represents a more stealthy form of backdoor. Inspired by this concept, [28] introduces an edge-case backdoor attack that utilizes rare samples (the tail of a dataset) to trigger the backdoor. The Distributed Backdoor Attack (DBA) [29] decomposes a trigger into multiple sub-triggers, with each attacker holding one of these sub-triggers for data poisoning, significantly enhancing the backdoor effectiveness in the global model.

In further developments, more advanced attacks [12], [16], [30] suggest constructing the neuron activation path to unimportant or redundant neurons that are less frequently updated, thereby preventing the backdoor from being promptly erased

and enhancing its persistence. Most recently, Li et al. [31] presented 3DFed, which addresses three prominent defense strategies with corresponding attack modules and introduces an indicator mechanism to assess whether backdoor updates are incorporated into model aggregation. This allows for an adaptive adjustment of the attack strategy.

Trigger-Optimized and Feature-Based Attacks. Recent advancements have explored trigger optimization and feature manipulation to improve attack robustness. For instance, A3FL [10] adversarially adapts triggers against a simulated unlearning model, while Parasite [32] utilizes discriminative feature pushing-pulling to separate decision boundaries. Other optimization-based attacks, such as Silent Penetrator [33], F3BA [30], and CerP [34], aim to craft robust triggers to make attacks more covert and persistent. While these methods are highly innovative, they generally formulate the backdoor injection as an end-to-end supervised classification task. This approach can sometimes leave distinguishable out-of-distribution feature clusters. Recently, Li et al. [35] utilized Generative Adversarial Networks [36] to better mask these anomalies, though it introduces additional computational complexity. In contrast, SPA explores an alternative perspective: rather than relying on a supervised classification loss, it explicitly aligns dynamically optimized triggers directly with the intrinsic feature manifold of the target class, aiming to achieve stealth and persistence through natural feature integration.

B. Backdoor Defenses in Federated Learning

Defenses against backdoor attacks in federated learning can be broadly categorized into two main approaches. The first follows the spirit of anomaly detection. The core hypothesis of such methods is that poisoned local models are outliers that deviate significantly from benign models. Consequently, these methods aim to detect outliers by computing the discrepancies between model weights/gradients, model representations, or other metrics, and then exclude abnormal updates to mitigate poisoning attacks. Krum [20] selects only the local model with the smallest sum of squared Euclidean distances from its $n - m - 2$ nearest neighboring models to serve as the global model, where m is an upper bound on the number of malicious participants in FL. Multi-Krum [20] extends Krum by selecting multiple local models based on the Krum criterion and averaging the selected local models to form the global model. Foolsgold [19] assigns lower aggregation weights to updates with high pairwise cosine similarities, thereby mitigating the impact of backdoor updates. RFLBAT [37] discriminates malicious models according to the difference between poisoned and clean updates in a low-dimensional projection space. Deepsight [21] aims to identify backdoor updates by measuring the fine-grained differences between model updates, generally assuming that the training data of backdoor models exhibits less heterogeneity than that of benign models. BackdoorIndicator [22] proposed the BackdoorIndicator (referred to as "Indicator" hereafter), which detects potentially poisoned models based on the OOD properties of backdoor samples.

The second approach involves methods for mitigating backdoors, which primarily focus on suppressing or perturbing the magnitude of updates rather than detecting these malicious models. The essence of these methods is to disrupt the clustering of backdoor features in the model's feature space. FLAME [38] introduces noise to eliminate backdoors based on the concept of differential privacy (DP) [39]. CRFL [40] prunes and adds Gaussian noise to model parameters as DP does, and utilizes smoothing techniques during testing, thus providing robustness guarantees against backdoor attacks. RLR [41] employs robust learning rates to update the global model, applying a negative rate to dimensions with directional disparities to mitigate the impact of malicious updates.

III. Problem Formulation

A. Federated Learning

Federated Learning enables N clients to train a global model w collaboratively without revealing local datasets. Unlike centralized learning, where local datasets must be collected by a central server before training, FL performs training by uploading the weights of local models $(\{w_k \mid k \in N\})$ trained on the local dataset $\mathcal{D}_k = \{(x_{k,i}, y_{k,i})\}_{i=1}^{n_k}$ of size n_k to a parameter server. Specifically, the global training objective is defined as follows:

$$w = \arg\min_{w} \sum_{k=1}^{N} \lambda_k L_k(w, \mathcal{D}_k), \tag{1}$$

$$L_k(w, \mathcal{D}_k) = \frac{1}{n_k} \sum_{i=1}^{n_k} f(w, (x_{k,i}, y_{k,i})), \tag{2}$$

where w denotes the optimal global model parameters, $L_k(w, \mathcal{D}_k)$ represents the average loss computed over the dataset $\mathcal{D}_k$ for client k, $(x_{k,i}, y_{k,i})$ denotes the i-th sample in $\mathcal{D}_k$, and λ_k indicates the weight of the loss for client k.

At the t-th federated training round, the server randomly selects a client set S_t where $|S_t| = m$ and $0 < m \leq N$, and broadcasts the current model parameters w^t. The selected clients perform local training in the following three steps:

- Global model download. All clients download the global model w^t from the server.
- Local training. Each client updates the global model by training with their datasets: $w_k^t \leftarrow w_k^t - \eta \frac{\partial L_k(w_k^t, b)}{\partial w_k^t}$, where η and b refer to the learning rate and a local mini-batch, respectively.
- Aggregation. After the clients upload their local models $\{w_k^t \mid k \in S_t\}$, the server updates the global model by aggregating the local models.

$$w^{t+1} \leftarrow w^t + \mathbf{AGG}(w_k^t | k \in S_t). \tag{3}$$

Note that $\mathbf{AGG}(\cdot)$ is the pre-defined aggregation method, such as FedAvg [1].

B. Threat Model

We build upon the attacks proposed in previous works [10], [11], [29], [42], where the adversaries aim to inject malicious functions, often referred to as hidden neural trojans, into the global model. These neural trojans remain dormant during normal operations but would be activated when specific

pre-defined patterns, known as triggers, are present in the input data. Formally, let F_w denote a benign FL model and $F_{\tilde{w}}$ its backdoored counterpart. The attack objectives can be expressed as:

$$\begin{cases} F_w(x) = F_{\tilde{w}}(x) \\ F_{\tilde{w}}(x \oplus \delta) = y_{\text{target}} \end{cases}, \tag{4}$$

where x represents clean input data, δ is the trigger, and y_{target} is the target label imposed by the adversary.

In the FL setting, the adversary possesses complete control over the local training data of compromised clients and can arbitrarily alter the training process. Conventional approaches inject backdoors via end-to-end supervised training on poisoned data. Specifically, for clean local samples $\mathcal{D}_k$, the adversary poisons a fraction r of the data by incorporating triggers δ, yielding a backdoor dataset $\mathcal{D}_{b,k}$, such that $\mathcal{D}_k = \mathcal{D}_{c,k} \cup \mathcal{D}_{b,k}$ and $|\mathcal{D}_{b,k}| = r \cdot |\mathcal{D}_k|$. The adversary then optimizes a backdoored local model $F_{\tilde{w}_k}$ by minimizing the loss:

$$\begin{aligned} \mathbb{E}_{(x,y)\sim\mathcal{D}_k}[\ell(F_{\tilde{w}_k}(x), y)] = \mathbb{E}_{(x,y)\sim\mathcal{D}_{c,k}}[\ell(F_{\tilde{w}_k}(x), y)] \\ + \mathbb{E}_{(x,y_{\text{target}})\sim\mathcal{D}_{b,k}}[\ell(F_{\tilde{w}_k}(x), y_{\text{target}})]. \end{aligned} \tag{5}$$

1) Adversaries' Capability and Knowledge: The adversary controls the local training data and training procedure of the compromised client, but cannot access or modify any information related to benign participants, such as their local datasets, local models, or training processes. The adversary also has no prior knowledge of, and cannot manipulate, the aggregation rule employed by the central server. Following the standard FL protocol, when the compromised client is selected in a communication round, it receives the current global model broadcast by the server and performs local training on this model. Therefore, the adversary can use the received global model and its intermediate representations during local training, but does not obtain any additional server-side information beyond what a selected client normally observes. We further assume that the compromised client has access to a small number of local samples from the target class. To simulate a more realistic scenario and address the limitations of existing studies that assume multiple malicious clients or continuous attacks, we focus on a setting where the adversary controls only a single client, and this client participates in training for only a limited number of communication rounds.

Note that the dynamic trigger optimization in SPA is performed locally by the compromised client using the received global model. It does not require additional communication with the server or changes to the FL protocol, but it introduces extra local computation for optimizing the trigger before backdoor injection. Moreover, deployments that prevent selected clients from accessing the plaintext global model or its intermediate features during local training are outside the scope of our threat model.

IV. METHODOLOGY

A. Attack Intuition and Overview

1) Key Intuition: As established in our introduction, conventional backdoor injection, formulated in Eq. (5), creates a fundamental conflict with the principles of Federated Learning. By forcing a malicious client to solve a secondary, supervised task, this approach gives rise to the Task-Divergence Anomaly (**Observation 1**). The resulting model updates are statistically divergent from the primary collaborative objective, making them vulnerable to server-side defenses. Our key intuition is to circumvent this anomaly entirely. Instead of creating a conflicting task that causes feature disentanglement, a more intuitive idea is to directly establish a relationship between the trigger features and the target class feature representation. In other words, the goal is to directly manipulate the feature space so that the model is tricked into perceiving trigger-embedded inputs as authentic members of the target class. By integrating the backdoor logic into the feature manifold of the primary task, rather than treating it as a separate objective, we can ensure the resulting model updates align with the collaborative training process, thus achieving stealthiness.

2) Overview: As illustrated in Figure 2, our approach, denoted as SPA, consists of two key components: *backdoor injection* and *backdoor enhancement*.

- In the backdoor injection module, SPA directly establishes the connection between the trigger feature and the target class feature via feature alignment, so that the feature embeddings of trigger-embedded samples closely resemble those of the target class samples, thereby inducing model misclassification. Furthermore, we extract clean knowledge from the frozen global model to preserve the utility of the backdoor model.
- In the backdoor enhancement module, considering that aligning a fixed trigger with the high-dimensional, evolving features of the target class is inherently difficult. We search for adversarial noise from the global model feature space that can inherently steer the predictions of the global model towards the target label as the backdoor trigger, making the backdoor more effective and durable.

B. Backdoor Injection

In the FL setting, each selected participant in round t receives the global model w^t distributed by the server. Following the intuition outlined in Section IV-A, we do not establish a direct connection between backdoor features and target labels in an end-to-end manner. Instead, we focus on creating a strong association between backdoor features and the feature representation of the target class. A straightforward and effective approach is to pull the feature embeddings of backdoor samples closer to those of the target class samples, thereby deceiving the model into perceiving the trigger features as belonging to the target class features.

Specifically, the attackers filter out the sample subsets $\mathcal{D}_{\text{target}}$ whose labels correspond to the target class y_{target} from their local dataset $\mathcal{D}_a$. $x \oplus \delta$ indicates embedding the trigger δ into an input sample x, referred to as a poisoned sample. To ensure the effectiveness of the backdoor, the attacker manipulates the global model so that it generates strongly similar feature embeddings for any poisoned samples and the target sample, thereby effectively establishing a strong correlation between the trigger features and the features of the target class. Such a correlation would indirectly cause the backdoored model to

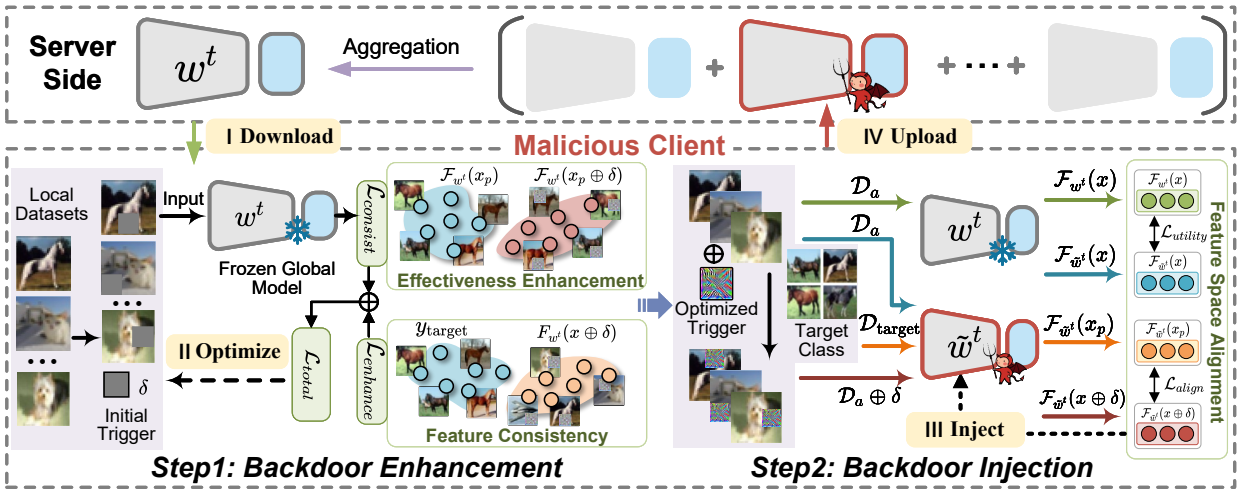


Fig. 2. Workflow of SPA. ① Download the current global model from the server. ② Generate a trigger tailored to the global model via adversarial dynamic trigger optimization. ③ Inject the backdoor through feature-space alignment with the optimized trigger. ④ Upload the backdoored update to the server.

predict the labels desired by the attacker for trigger-embedded samples. Formally, we denote the feature embedding output of the manipulated global model $\tilde{w}^t$ at the t-th round as $\mathcal{F}_{\tilde{w}^t}$ and the feature-space alignment objective is defined as follows:

$$\mathcal{L}_{align} = \frac{1}{|\mathcal{D}_a| \cdot |\mathcal{D}_{\text{target}}|} \sum_{x \in \mathcal{D}_a} \sum_{x_p \in \mathcal{D}_{\text{target}}} d(\mathcal{F}_{\tilde{w}^t}(x \oplus \delta), \mathcal{F}_{\tilde{w}^t}(x_p)). \tag{6}$$

Here, $d(\cdot, \cdot)$ represents a distance metric between two feature embeddings. Typically, the L_2-norm is applied to reduce the distance between backdoor and target feature representations. However, in the FL scenario, the feature space of the global model is constantly evolving, and backdoor embeddings are often high-dimensional and contain some extreme values that are challenging to minimize effectively. Moreover, we lack precise control over whether backdoor features move closer to the target features or the opposite, potentially disrupting the feature embeddings of the clean target class. To address these limitations, we introduce the sliced-Wasserstein distance inspired by [43], a variant of the Wasserstein distance that is well-suited for measuring high-dimensional data and providing a more robust metric to efficiently handle complex distributions.

$$\mathcal{W}_{\text{sliced}}(\mathcal{F}_c, \mathcal{F}_b) = \left(\frac{1}{S} \sum_{s=1}^{S} \int_0^1 \left\| \mathcal{F}_c^s(z) - \mathcal{F}_b^s(z) \right\|_2 dz \right)^{1/2}, \tag{7}$$

where S is the number of one-dimensional directions (denoted by the randomly sampled unit vectors). $\mathcal{F}_c^s$ and $\mathcal{F}_b^s$ represent the projections of the clean and poisoned embeddings into one-dimensional data points along the direction of slice s, respectively.

Furthermore, we must ensure the utility of the backdoored model, meaning it achieves classification accuracy on clean data comparable to that of the global model, thus enhancing the stealthiness of the backdoor. To achieve this, we employ a distillation-like method to transfer effective knowledge of clean features to the backdoor model, preserving its performance on clean data. Specifically, we freeze the global model obtained in the current training round as a teacher model and denote its feature embedding output as $\mathcal{F}_{w^t}$. During backdoor feature alignment, we constrain the backdoored model to produce embeddings on trigger-free samples similar to those of the teacher model. Formally, we define a utility loss to quantify this objective:

$$\mathcal{L}_{utility} = \frac{1}{|\mathcal{D}_a|} \sum_{x \in \mathcal{D}_a} d(\mathcal{F}_{\tilde{w}^t}(x), \mathcal{F}_{w^t}(x)), \tag{8}$$

Finally, the backdoor injection process can be formulated as the following optimization problem:

$$\arg\min_{\tilde{w}^t} \mathcal{L} = \mathcal{L}_{align} + \lambda \mathcal{L}_{utility}. \tag{9}$$

C. Backdoor Enhancement

While the feature-space alignment strategy effectively addresses the Task-Divergence Anomaly (**Observation 1**), its success hinges on the nature of the trigger itself. As we established in our discussion of the Fixed-Trigger Challenge (**Observation 2**), using a static trigger is ineffective in FL's dynamic environment. Specifically, it is highly susceptible to the "backdoor shortcut" phenomenon, which risks collapsing our stealthy alignment strategy back into a detectable, supervised-like task. Therefore, to ensure robust and persistent attacks, we must move beyond both the conventional learning objective and the static trigger design.

Specifically, we introduce an adversarial dynamic trigger optimization method to enhance the effectiveness of backdoor attacks. This method aims to better integrate backdoor triggers into the target feature space, ensuring the backdoor remains robust and persistent even throughout the ongoing FL process. Specifically, the attacker has access to the current global model w^t. Inspired by the concept of adversarial examples, we aim to search for adversarial noise in the feature space that biases the predictions of the global model toward the target class desired by the attacker. This adversarial noise represents naturally occurring perturbations within the target class of the global model, which is an inherent vulnerability in the global model that can function as a "natural backdoor". Such a backdoor is robust and difficult to eliminate with the iteration of FL

communication, as it aligns closely with the model's intrinsic feature representations. By leveraging feature-space alignment, the model can be misled into learning these robust backdoor features as part of the normal feature manifold of the target class. The optimization of the backdoor trigger can be formally defined and quantified using the following loss function:

$$\mathcal{L}_{\text{enhance}} = -\frac{1}{|\mathcal{D}_a|}\sum_{x\in\mathcal{D}_a}\sum_{i=1}^{C} y_{\text{target},i}\log\left(F_{w^t}(x\oplus\delta)_i\right), \quad (10)$$

where C is the number of categories and F_{w^t} denotes the model outputs.

In the domain of adversarial machine learning, L_p norms, such as the L_∞ norm, are commonly used to regulate the magnitude of adversarial noise, i.e., $\|\delta\|_\infty \leq \varepsilon$. This approach operates within the input space, specifically at the pixel level, where it limits the size of perturbations to ensure they remain visually imperceptible to human observers. Despite its widespread use, this strategy exhibits inherent limitations, as it may not effectively govern the behavior of the model within the feature space, where higher-level representations are encoded. Furthermore, the fixed L_p constraint lacks flexibility, requiring manual adjustment of the perturbation amplitude threshold (e.g., ε value) to balance the effectiveness and invisibility of the perturbation. This becomes particularly challenging in the dynamically evolving FL scenarios.

To address these shortcomings, we propose an innovative approach that constrains the magnitude of adversarial perturbations through feature consistency. Specifically, we enforce a condition whereby samples augmented with triggers retain a high degree of similarity to their original counterparts within the feature space. Such strategies enable attackers to dynamically adjust the strength of the backdoor by leveraging the most recent global model, thereby enhancing the stealthiness of the attack while maintaining its efficacy. To quantify this similarity, we adopt the projection distance, a metric that emphasizes the directional alignment of features while remaining insensitive to variations in their magnitudes. The projection distance is formally defined as:

$$d_{\text{proj}}(a,b) = \left\| a - \frac{a\cdot b}{\|b\|^2} b \right\|_2. \quad (11)$$

This measure ensures that the features of perturbed samples are distributed along the principal direction associated with a given class, rather than being forcibly aligned with a specific individual sample. As a result, the constraint imposed on the trigger can be articulated as:

$$\mathcal{L}_{consist} = \frac{1}{|\mathcal{D}_{\text{target}}|}\sum_{x_p\in\mathcal{D}_{\text{target}}} d_{\text{proj}}\left(\mathcal{F}_{w^t}(x_p), \mathcal{F}_{w^t}(x_p\oplus\delta)\right). \quad (12)$$

It is imperative to note that we restrict feature alignment to only the target class samples within each batch, rather than applying it universally to all samples. This selective alignment serves a dual purpose: it effectively limits the magnitude of the perturbations, ensuring their subtlety, while simultaneously guiding the trigger features to converge toward the feature distribution characteristic of the target class. By integrating these principles, our approach not only enhances the stealthiness of adversarial perturbations but also ensures their robustness and adaptability in practical attack scenarios. Finally, the backdoor enhancement process can be formulated as the following optimization problem:

$$\arg\min_{\delta} \mathcal{L} = \mathcal{L}_{enhance} + \beta\mathcal{L}_{consist}. \quad (13)$$

Notably, this optimization is a continuous process. If an attacker is selected for the first time, the trigger δ is initialized from a random perturbation. In subsequent rounds, the optimization resumes from the attacker's previously optimized local trigger, allowing it to co-evolve with the system. Accordingly, the overall SPA algorithm for an attacker can be summarized in Algorithm 1.

Algorithm 1: SPA Algorithm

Input: Current global model parameters w^t, the data of malicious client $\mathcal{D}_a$.
Output: The backdoored model $\tilde{w}^t$.

Malicious Client Execution:
Freeze global model and embeddings: F_{w^t}, $\mathcal{F}_{w^t}$;
Initialize the backdoor model: $\tilde{w}^t \leftarrow w^t$;
// Backdoor Enhancement Phase
$\delta \leftarrow$ Initialize random perturbation;
if *this is the first attack round* **then**
  $\delta \leftarrow$ Initialize random perturbation;
else
  $\delta \leftarrow$ Load trigger δ from previous attack round;
for *optimization step* $i \in [1, 2, \cdots, I]$ **do**
  for *batch* $b_a = \{(x,y)\} \in \mathcal{D}_a, b_p = \{(x,y)\} \in \mathcal{D}_{target}$ **do**
    $\mathcal{L}_{enhance} \leftarrow -\frac{1}{|b_a|}\sum_{x\in b_a}\sum_{i=1}^{C} y_{\text{target},i}\log(F_{w^t}(x\oplus\delta)_i)$;
    $\mathcal{L}_{consist} \leftarrow \frac{1}{|b_p|}\sum_{x_p\in b_p} d_{\text{proj}}(\mathcal{F}_{w^t}(x_p), \mathcal{F}_{w^t}(x_p\oplus\delta))$;
    $\mathcal{L} \leftarrow \mathcal{L}_{enhance} + \beta\mathcal{L}_{consist}$;
    $\delta \leftarrow \delta - \eta_\delta \nabla_\delta \mathcal{L}$;
// Backdoor Injection Phase
for *each local epoch* $e \in [1, 2, \cdots, E_{attack}]$ **do**
  for *batch* $b_a = \{(x,y)\} \in \mathcal{D}_a, b_p = \{(x,y)\} \in \mathcal{D}_{target}$ **do**
    $\mathcal{L}_{align} \leftarrow \frac{1}{|b_a|\cdot|b_p|}\sum_{x\in b_a}\sum_{x_p\in b_p}\mathcal{W}_{\text{sliced}}(\mathcal{F}_{\tilde{w}^t}(x\oplus\delta), \mathcal{F}_{\tilde{w}^t}(x_p))$;
    $\mathcal{L}_{utility} \leftarrow \frac{1}{|b_a|}\sum_{x\in b_a}\mathcal{W}_{\text{sliced}}(\mathcal{F}_{\tilde{w}^t}(x), \mathcal{F}_{w^t}(x))$;
    $\mathcal{L} \leftarrow \mathcal{L}_{align} + \lambda\mathcal{L}_{utility}$;
    $\tilde{w}^t \leftarrow \tilde{w}^t - \eta\nabla\mathcal{L}$;
return $\tilde{w}^t$

V. EXPERIMENTAL EVALUATION

In this section, we rigorously evaluate the efficacy of SPA against the SOTA defenses through comprehensive experiments. The detailed setups are listed as follows:

Datasets and Models: We evaluate SPA on four widely-used datasets: CIFAR-10, CIFAR-100 [44], GTSRB [45] and SVHN [46]. ResNet18 [47] is employed as the default model architecture, which is a classic and effective model for image classification. Besides, we also evaluate the performance of SPA under different model architectures, including ResNet34, VGG11, VGG19, and MobileNet-V2.

Federated Learning Setup: Our experiment implements all the tasks in the FL system, running image classification tasks using FedAvg on a single machine using an NVIDIA GeForce RTX 3090 GPU with 24GB of memory. At each communication round, the server randomly selects ten clients to contribute to the global model while the total number of clients is 100. Following previous research [48], we randomly split the dataset over clients in a non-IID manner, using Dirichlet sampling with the parameter α set to 1 by default. We also evaluate the impact of data heterogeneity by adjusting the value of α. During the training process, each selected client trains the local model for two local epochs using the SGD optimizer with a learning rate of 0.01. The FL training process continues for 2,100 communication rounds.

Attack and Defense Setups: To compare the effectiveness of SPA with SOTA methods, we conduct experiments against A3FL [10], Chameleon [13], Neurotoxin [12], DBA [29], PGD [42], and Vanilla [11]. Among these methods, A3FL, Chameleon, and Neurotoxin are designed to implant more persistent backdoors. For trigger-optimization-based backdoor attacks, namely SPA and A3FL, we configure each attacker to optimize their trigger pattern using PGD [42] with a step size of 0.001. For other attacks, we adhere to their original experimental configurations. To simulate more realistic scenarios, the attack window begins at the 2000th round and continues for 100 rounds, forcing attackers to adapt to random client selection. Unless otherwise specified, the number of attackers is set to 1, which means that the attack window is selected approximately 10 times. Additionally, we explore the impact of various attack scenarios.

We also evaluate those attacks against six SOTA defenses: Multikrum [20], Deepsight [21], Foolsgold [19], Rflbat [37], Flame [38] and Indicator [22]. For those attacks and defense methods, we follow the original implementations. Moreover, we also verified the robustness of SPA against the SOTA DP-based backdoor defense method CRFL [40].

Evaluation Metrics: Following previous works, we evaluate the performance of SPA with two metrics: ❶ attack success rate (ASR), which is the ratio of backdoored inputs misclassified by the backdoor model as the target labels:

$$\text{ASR} = \frac{\#\text{successful attacks}}{\#\text{total trials}}, \tag{14}$$

and ❷ the accuracy of the main classification task on normal samples (ACC). Note that we report the mean ASR of all the attackers on the global model at the end of FL. A stealthy and persistent attack in FL is characterized by its ability to maintain the model's original ACC while ensuring that the ASR either remains stable or experiences only minimal degradation throughout the FL iterations. These enable the attack to remain undetected by defense mechanisms while maintaining its effectiveness over extended training periods. To evaluate the persistence of our attack, we introduce lifespan$_\tau$, defined as the number of FL rounds from the end of the attack window until the ASR first drops below a threshold τ. Unless otherwise specified, we set $\tau = 60\%$. We repeated the experiment five times using different random seeds to ensure the robustness and reliability of the results.

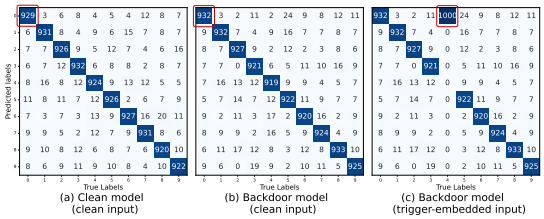


Fig. 3. Confusion matrix of clean models and backdoor models.

A. Attack Performance

Comparison with SOTA Methods. Table I reports a comprehensive comparison between SPA and five state-of-the-art backdoor attacks in FL under multiple defense mechanisms. To emulate challenging, realistic conditions, we use a single attacker that participates intermittently via random selection, with an attack window restricted to rounds 2000–2100; under this setting, the attacker is selected approximately 10 times. Because DBA requires at least two attackers to train distinct local triggers, we exclude it from this section. We evaluate six widely adopted backdoor defenses and a no-defense baseline. Best results are in **bold**, and second-best are marked with an asterisk $*$. Overall, SPA attains superior or near-superior performance across all datasets, sustaining ASR above 98% while bypassing all defenses. It also exhibits minimal impact on ACC, ranking among the top methods on this metric.

Impact on Target Class Accuracy. We further assess whether SPA compromises the model's natural recognition of the target class. By design, SPA encourages the model to internalize trigger patterns as legitimate components of the target class features, raising a concern that high ASR could induce over-reliance on triggers at the expense of genuine class cues. To examine this, we compare confusion matrices before and after the attack on CIFAR-10 (10,000 test samples; 1,000 per class), setting class 0 as the target and applying triggers only to class 5 at test time. As shown in Figure 3, the number of correctly classified target-class samples differs by only three between the clean and attacked models, while triggered class-5 samples are consistently misclassified as class 0. These results indicate that SPA preserves the model's intrinsic discriminative ability for the target class while achieving strong backdoor efficacy, without inducing over-reliance on trigger features.

T-SNE Visualization. We further use T-SNE to analyze feature distributions under different attacks. Specifically, we poison 10% of the CIFAR-10 test set and feed them into backdoored models, then visualize the resulting representations. In Figure 4, black points denote triggered samples, while colored points denote benign samples from each class. Benign samples form well-separated clusters, whereas triggered samples form distinct clusters. Traditional supervised methods directly associate target labels with backdoor features, producing compact,

TABLE I
PERFORMANCE COMPARISON OF ATTACK METHODS UNDER SIX SOTA DEFENSES.

Dataset	Attack	Vanilla		PGD		Neurotoxin		Chameleon		A3FL		SPA	
		ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR
CIFAR-10	Nodefense	92.21	67.03	92.05	67.32	92.25*	63.79	91.73	48.78	92.04	98.86*	**92.55**	**99.91**
	Deepsight	92.08	65.41	92.06	60.58	92.14	54.24	92.19	36.40	92.41*	96.73*	**92.66**	**99.88**
	Foolsgold	90.34	61.49	90.89*	62.67	90.20	76.88	90.23	42.10	90.28	98.86*	**91.32**	**99.14**
	Indicator	89.32	3.51	84.58	11.27	87.14	12.67	83.54	2.22	90.52*	53.90*	**91.14**	**98.31**
	Multikrum	**92.45**	0.40	92.16	0.71	92.10	0.18	92.31	0.34	91.54	96.03*	92.37*	**99.96**
	Rflbat	88.29	83.41	87.63	80.17	87.25	30.92	87.43	25.12	**89.76**	98.59*	89.14*	**99.29**
	Flame	92.34	0.26	92.35	0.20	92.44	0.21	92.52*	0.28	92.41	98.60*	**92.52**	**99.92**
CIFAR-100	Nodefense	71.25	71.31	70.74	77.65	71.14	74.74	**71.91**	33.09	71.55	97.45*	71.63*	**99.97**
	Deepsight	70.98	78.22	71.03	73.44	71.17	70.62	71.69	31.23	71.72*	**99.95**	**71.83**	99.86*
	Foolsgold	70.51	72.69	70.62	85.58	70.49	75.52	71.66*	36.01	71.02	98.62*	**71.94**	**99.45**
	Indicator	69.46	0.14	60.02	0.29	66.04	0.27	61.27	0.11	69.6*	98.28*	**70.51**	**98.99**
	Multikrum	71.66	0.11	71.94	0.15	71.65	0.13	71.76	0.20	**71.97**	94.84*	71.95*	**99.32**
	Rflbat	63.76	51.6	63.10	58.9	64.12	64.32	62.48	3.92	64.3*	98.64*	**68.18**	**98.89**
	Flame	72.29*	0.15	72.18	0.13	72.05	0.10	72.12	0.12	72.02	95.20*	**72.86**	**99.38**
GTSRB	Nodefense	96.71	98.73	96.64	98.81	96.64	98.62	**96.81**	96.33	96.68	99.02*	96.79*	**99.57**
	Deepsight	96.41	81.89	96.78	98.80	**96.88**	98.45	96.56	97.62	96.54	98.99*	96.83*	**99.05**
	Foolsgold	96.68	98.01	96.67	98.49	95.73	97.28	96.77*	97.05	96.66	99.37*	**96.95**	**99.87**
	Indicator	**95.14**	86.21	90.19	95.54	90.53	88.33	92.04	5.96	89.38	77.15	94.59*	**100.00**
	Multikrum	96.71*	0.01	96.67	0.00	96.51	0.00	96.67	0.02	96.60	98.18*	**96.97**	**98.72**
	Rflbat	94.40	97.95	95.92*	97.71	94.52	99.36	94.29	96.88	95.10	100.00*	**96.08**	**100.00**
	Flame	96.43	0.01	96.52	0.00	96.58*	0.01	95.92	5.95	96.47	98.54*	**96.95**	**99.45**
SVHN	Nodefense	92.91	58.93	92.68	52.31	92.67	59.61	92.68	49.29	93.41*	99.07*	**93.61**	**99.76**
	Deepsight	93.51*	50.79	92.76	45.87	92.63	45.46	92.66	14.00	93.12	98.46*	**93.56**	**99.26**
	Foolsgold	92.77*	65.32	92.50	61.61	92.37	57.27	92.51	46.68	92.56	98.21*	**92.86**	**99.54**
	Indicator	90.42	1.85	89.69	8.76	88.2	25.00	90.89	2.04	91.01*	97.37*	**91.25**	**98.76**
	Multikrum	93.00	0.89	92.48	1.11	93.15*	0.83	93.05	0.90	92.82	93.46*	**93.38**	**99.28**
	Rflbat	86.13	38.83	89.70*	16.88	87.25	30.92	87.43	25.12	88.74	97.14*	**90.05**	**99.29**
	Flame	92.50	1.29	92.38	1.38	92.41	1.36	92.50	1.29	92.62*	96.71*	**92.88**	**98.57**

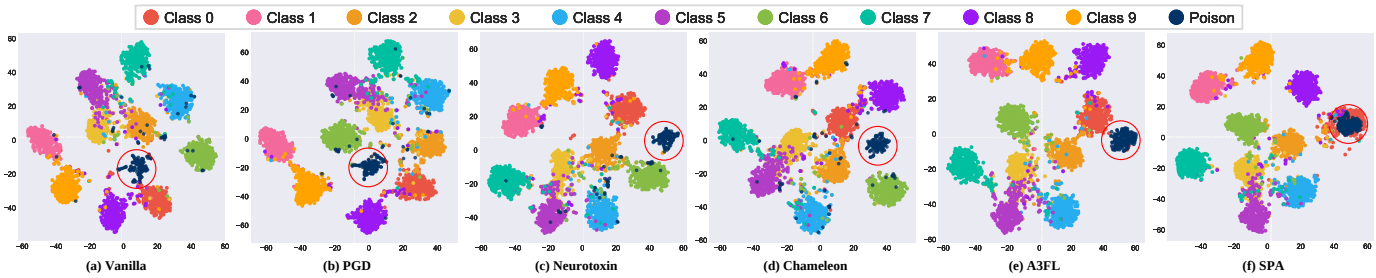


Fig. 4. T-SNE visualization of SPA compared with five attacks.

isolated backdoor clusters that are OOD relative to benign clusters, making them susceptible to OOD-based defenses. In contrast, SPA yields backdoor clusters that lie near the target-class cluster, encouraging the model to treat trigger patterns as legitimate components of the target class. This blurs the boundary between backdoor and target-class features, enabling more natural integration and allowing the attack to evade defenses that rely on detecting anomalous feature clusters.

Stealthiness Analysis. To provide a deeper mechanistic understanding of why SPA bypasses existing defenses, we quantitatively analyze the attack's stealthiness. In the context of FL, stealthiness inherently requires the malicious updates to be statistically indistinguishable from benign ones, thereby evading anomaly-based detection. First, we compare the update norms of SPA with those of benign clients by measuring the L_2 norm of the difference between the locally updated models and the global model. As illustrated in Fig. 5 (where the attacker, client 0, is highlighted in red), the magnitudes of the malicious updates remain strictly within the normal range of benign updates across varying degrees of data heterogeneity. This indistinguishability in update magnitude provides a direct quantitative explanation for the failure of gradient-based defenses that rely on the assumption that backdoor gradients exhibit significant magnitude deviations.

Furthermore, we evaluate the detection performance using True Positive Rate (TPR) and False Positive Rate (FPR). TPR reflects the proportion of correctly identified malicious clients, while FPR represents the proportion of benign clients incorrectly classified as malicious. In this setup, a single attacker is randomly selected but forced to be chosen exactly 10 times during the attack window (rounds 2000 to 2100). As presented in Table II, the attacker is flagged and excluded from aggregation on only a very limited number of occasions. However, only limited aggregation participations are sufficient to implant the backdoor. Additionally, we observe that increasing the degree of data heterogeneity leads to a higher FPR. This highlights a severe limitation of existing detection methods, further demonstrating they struggle to maintain robustness in

TABLE II
PERFORMANCE OF DETECTION METHODS UNDER DIFFERENT DEGREES OF DATA HETEROGENEITY.

Dir(α)	0.5				1				10			
Method	**ACC**	**ASR**	**TPR**	**FPR**	**ACC**	**ASR**	**TPR**	**FPR**	**ACC**	**ASR**	**TPR**	**FPR**
No-defense	91.56	99.65	-	-	92.55	99.91	-	-	92.59	99.17	-	-
Deepsight	92.04	99.97	20.00	63.50	92.66	99.88	10.00	59.80	92.80	99.24	10.00	52.30
Foolsgold	90.34	99.56	20.00	46.30	91.32	99.14	30.00	43.20	90.40	98.82	0.00	47.20
Indicator	90.64	98.63	0.00	15.30	91.14	98.31	10.00	23.00	91.27	98.66	20.00	26.10
Multikrum	91.86	99.71	20.00	31.20	92.37	99.96	20.00	34.10	91.56	99.81	10.00	37.60
Rflbat	88.95	99.35	20.00	9.20	89.14	99.29	0.00	9.60	89.12	99.38	20.00	10.70
Flame	92.05	99.67	10.00	29.50	92.52	99.92	10.00	35.20	92.60	99.25	30.00	36.80

Note: TPR and FPR are reported in percentage (%)

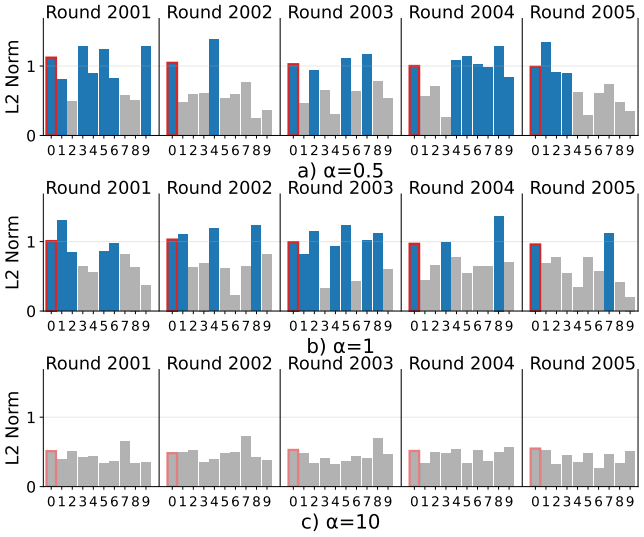


Fig. 5. Comparison of update norms between benign and malicious gradients.

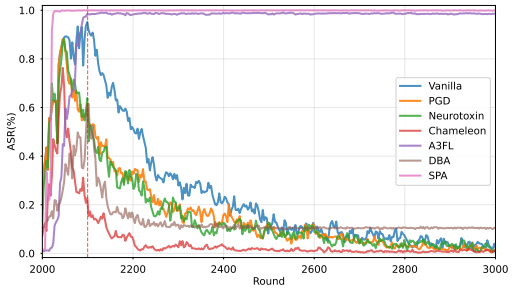


Fig. 6. Persistence evaluations under no defenses.

practical, heterogeneous FL environments.

Persistence Analysis. We further evaluated the durability of SPA by comparing its persistence with baseline methods. The attack window remained fixed at rounds 2000–2100. To boost ASR during this phase, we used three malicious clients (still sampled randomly each round) and extended training to 3000 rounds. We also included DBA, in which three attackers trained distinct local triggers that were ultimately combined into a global trigger. Figure 6 reports results without defenses. We observed that SPA exhibits substantially longer persistence, sustaining ASR above 90% for more than 900 rounds after the attack window. In contrast, all baselines except A3FL show rapid ASR decay. This degradation reflects a limitation of conventional end-to-end backdoors, whose abnormal gradients are diluted by benign updates during aggregation. By comparison, SPA embeds backdoor features as supplementary knowledge within the target-class feature space, aligning with FL's objective of aggregating distributed knowledge. A3FL achieves comparable persistence via adversarial training with unlearning, but at considerable computational cost, limiting practicality. Experiments under diverse defense mechanisms illustrated in Figure 7 further confirm the superior durability of SPA, which consistently maintains high attack effectiveness across defenses. Collectively, these results show that our feature-space alignment yields both stealth and persistence by integrating backdoor features as complementary knowledge.

TABLE III
RESULTS ON DIFFERENT NON-IID SETTINGS.

Dataset		0.5	1	5	10	1000
CIFAR 10	**ACC**	91.56	92.55	92.62	92.59	92.88
	ASR	99.65	99.91	99.42	99.96	100.00
CIFAR 100	**ACC**	71.59	71.63	71.68	71.78	71.82
	ASR	99.24	99.97	100.00	99.76	99.85
GTSRB	**ACC**	96.28	96.79	96.81	96.91	96.86
	ASR	99.46	99.57	99.53	100.00	100.00
SVHN	**ACC**	92.16	93.61	93.72	93.76	93.94
	ASR	99.42	99.76	99.86	100.00	99.96

Non-IID Data Distributions. In this section, we explore the impact of non-IID. Following existing work [48], we employ the Dirichlet distribution Dir(α) to model varying degrees of data heterogeneity, where smaller α values correspond to higher data heterogeneity. Specifically, we varied α across values of 0.5, 1, 5, 10, and 1000 across four datasets, with $\alpha = 10$ representing the IID scenario.

B. Applicability Analysis

As demonstrated in Table III, SPA exhibits remarkable robustness against varying data distributions, maintaining consistent performance across all tested conditions. The primary challenge introduced by non-IID lies in the reduced availability of target class samples for attackers. Nevertheless, our trigger enhancement mechanism enables effective backdoor feature learning even when target class samples are scarce. This resilience stems from the nature of our optimized trigger features, which inherently represent noise features near the target class decision boundary, thereby facilitating easier feature alignment with the target class using a few target samples.

Impact of Adversary Number. To evaluate the practical applicability of our method in real-world scenarios where

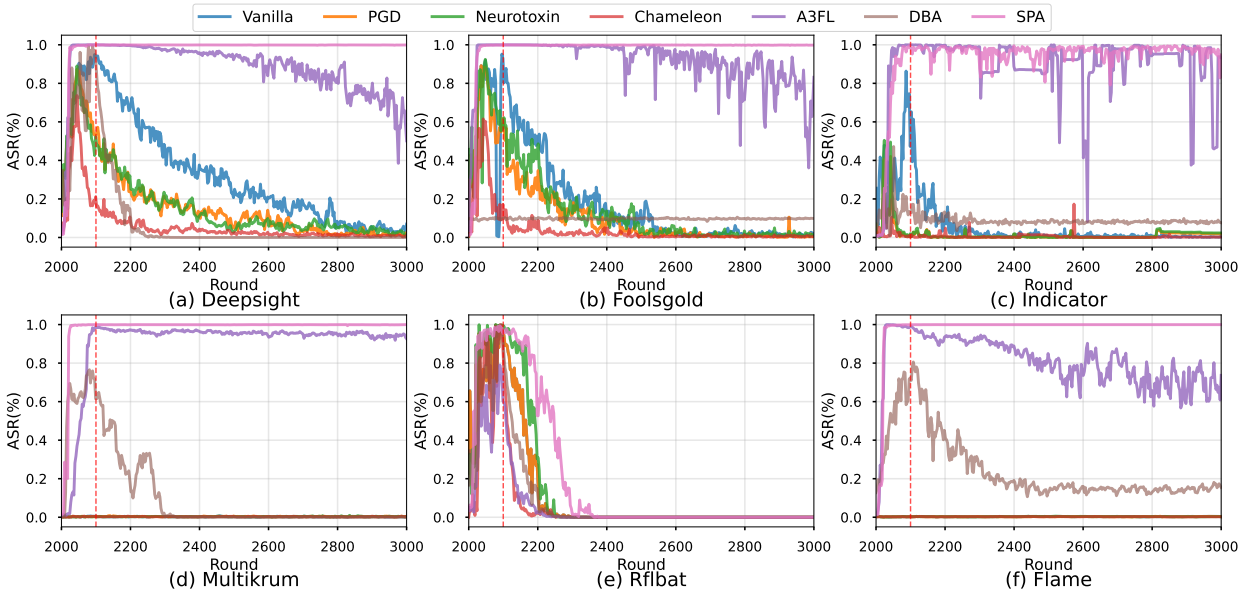
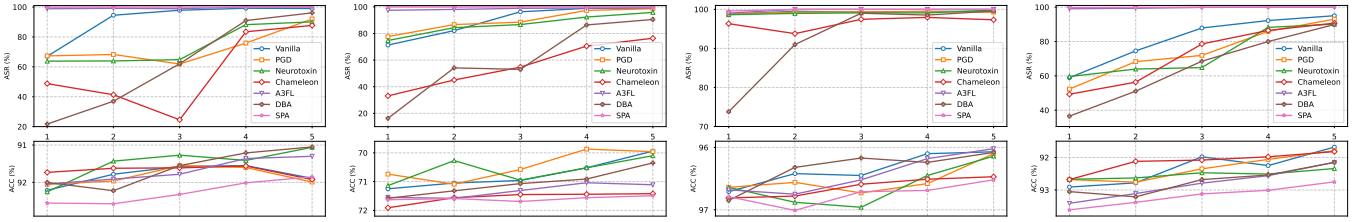


Fig. 7. Comparison of persistence performance under different defense methods.



(a) Results on CIFAR-10 dataset. (b) Results on CIFAR-100 dataset. (c) Results on GTSRB dataset. (d) Results on SVHN dataset.

Fig. 8. Comparison of performance under a varying number of malicious clients on four datasets.

TABLE IV
PERFORMANCE OF SPA WITH DIFFERENT MODEL ARCHITECTURE.

Dataset	Metric	ResNet18	ResNet34	VGG11	VGG19	MobileNet-V2
CIFAR-10	**ACC**	92.55	93.89	91.14	92.07	85.72
	ASR	99.91	99.96	99.21	99.50	97.67
CIFAR-100	**ACC**	71.63	73.42	70.58	71.88	69.83
	ASR	99.97	100.00	98.62	99.82	98.19
GTSRB	**ACC**	96.79	97.94	96.91	97.15	96.15
	ASR	99.57	100.00	99.86	100.00	99.03
SVHN	**ACC**	93.61	94.24	91.21	93.45	88.97
	ASR	99.76	100.00	98.96	99.24	98.23

TABLE V
PERFORMANCE OF SPA WITH TRANSFORMER ARCHITECTURE.

Dataset	Defense	Vision Transformer	
		ACC	ASR
CIFAR-10	**Nodefense**	84.75	98.25
	Deepsight	84.68	97.76
	Foolsgold	84.81	98.26
	Indicator	83.84	98.34
	Multikrum	84.47	98.21
	Rflbat	82.96	97.73
	Flame	84.78	98.23

TABLE VI
RESULTS ON AUDIO RECOGNITION TASKS.

Data	No-attack	SPA		
	ACC	ACC	ASR	lifespan$_{60\%}$
SpeechCommands	95.79	95.52	99.12	900
UrbanSound8K	72.88	72.71	99.56	900

attackers typically cannot control multiple participants, we investigate the impact of varying numbers of malicious clients on attack effectiveness. Our experiments systematically assess different attack methods across four benchmark datasets with the number of attackers ranging from 1 to 5, while maintaining all other parameters at their default configurations.

The experimental results, as illustrated in Figure 8, reveal several important insights. Generally, as the number of malicious participants under attacker control increases, the ASR of backdoor attacks on the global model also increases, accompanied by a modest decline in ACC. Notably, SPA achieved an ASR exceeding 90% with just a single malicious participant, whereas most baseline attack methods achieved only approximately 70% ASR under the same conditions. This superior performance demonstrates that our approach is significantly more applicable to real-world scenarios where controlling multiple participants is impractical.

Impact of Different Model Architectures. Given that SPA operates by aligning representations in the feature space, its efficacy may be architecture-dependent. To evaluate its generalizability, we tested SPA across five widely-used image classification architectures: ResNet18, ResNet34, VGG11, VGG19, and MobileNet-V2. All other experimental parameters were held constant. The results, presented in Table IV, demonstrate that SPA is robust across diverse architectures. While the ACC naturally varies between models, the ASR remains consistently high in all cases. We also observed a positive correlation

TABLE VII
PERFORMANCE UNDER DIFFERENT TOTAL NUMBER OF PARTICIPANTS AND LARGER MODEL.

$\|S_t\|$	Model	100		500		1000	
		ACC	ASR	ACC	ASR	ACC	ASR
10	ResNet18	92.55	99.91	92.12	98.59	92.25	99.01
	ResNet50	92.91	99.88	92.70	98.39	92.81	98.05
	ResNet101	93.65	99.89	93.45	99.57	93.52	99.06
20	ResNet18	92.35	99.23	92.11	98.20	92.92	98.15
	ResNet50	92.69	99.96	92.71	99.37	93.01	98.92
	ResNet101	93.59	99.52	93.21	99.24	93.31	99.02

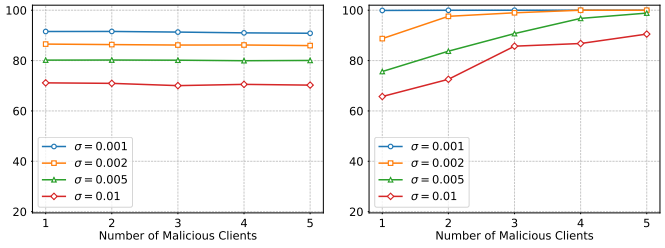
between a model's ACC and the resulting ASR. This suggests that models with superior representation capabilities form more distinct and well-defined feature manifolds for each class, which contributes to alignment-based attack.

Furthermore, we evaluated the effectiveness of SPA on architectures beyond CNNs, specifically Vision Transformers (ViTs), which operate via global self-attention rather than local convolutional kernels. Experiments were conducted on the CIFAR-10 dataset, with the results presented in Table V. The findings indicate that SPA remains effective on ViT-based models, consistently maintaining both attack efficacy and stealth under various defense mechanisms. This suggests that the alignment enforced by SPA operates within a high-level latent feature space, remaining largely agnostic to the specific underlying feature extraction mechanism.

Audio Recognition Tasks. To demonstrate the cross-domain generalizability of SPA beyond computer vision, we extend our evaluation to the audio domain using two benchmark datasets: SpeechCommands and UrbanSound8K. The FL configurations remain consistent with those in our previous setups, with MobileNet-V2 adopted as the backbone model. For trigger injection, learnable perturbations are introduced directly onto the Mel-spectrograms of the audio inputs. As shown in Table VI, the results indicate that SPA is not confined to CV tasks but generalizes effectively to other modalities.

More Realistic FL Scenarios. In this section, we investigate the performance of SPA under more realistic federated learning settings. We scale up our experiments by adopting larger models, namely ResNet-50 and ResNet-101, and by increasing the total number of participants to 500 and 1000, with 10 or 20 clients selected for aggregation in each round. A single attacker is present and is randomly selected exactly 10 times during the attack window. Experimental results are presented in Table VII. The findings demonstrate that SPA maintains high attack effectiveness even with larger models and larger-scale FL deployments. However, a prerequisite for success is that the attacker must be selected for aggregation. Notably, although we enforce 10 selections, our monitoring reveals that the backdoor is successfully implanted into the global model (with ASR exceeding 98%) after only 4 to 5 selections. This stands in stark contrast to existing attack methods, which typically require the attacker to be continuously selected over many rounds to achieve comparable effectiveness.

Robustness to DP-based Defenses. In this section, we evaluate the resilience of SPA against a DP-based defense. Within the CRFL framework, we implement this defense by



(a) ACC (%). (b) ASR (%).

Fig. 9. Attack performances with different σ.

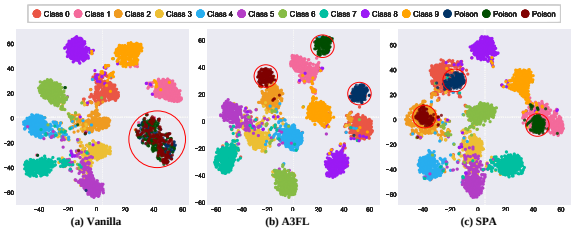


Fig. 10. T-SNE visualization of features under multi-label scenarios.

first clipping client updates to a threshold of 1, and then adding Gaussian noise $z \sim \mathcal{N}(0, \sigma^2 I)$, where we vary σ from 0.001 to 0.01. As depicted in Figure 9, the results indicate that the DP mechanism is largely ineffective against SPA. While greater noise intensity σ degrades ACC, it fails to meaningfully reduce the ASR. The reason for this failure lies in the fundamental mismatch between the defense and attack methodologies. DP is designed to prevent the model from memorizing a few anomalous samples by injecting noise, which is effective against traditional backdoors relying on strong, explicit trigger-label associations. In contrast, SPA employs feature alignment to make the trigger's features indistinguishable from the target class's natural features. This camouflages the backdoor within the model's normal feature space, and the noise from DP cannot disrupt this subtle, distributional alignment without being scaled to a level that would destroy the model's overall utility.

C. Different Attack Settings

Multi-label Attack Scenario. Traditional FL backdoor attacks often assume a single attacker controlling multiple colluding clients with a shared target label. In practice, multiple attackers may control different clients and target different labels, resulting in multi-label attacks. We evaluate this setting by configuring three attackers with distinct targets (labels 0, 1, and 2) and report the average performance across all three.

As shown in Table VIII, methods such as Vanilla, PGD, Neurotoxin, and Chameleon occasionally allow one attacker to achieve ASR above 90%, yet their average ASR remains low, indicating that typically only one attacker successfully compromises the model. In contrast, both A3FL and SPA enable all three attackers to achieve high ASR simultaneously. To probe the underlying mechanisms, we visualize T-SNE projections of feature representations for Vanilla, A3FL, and SPA as shown in Figure 10. For Vanilla, end-to-end training imposes multiple OOD objectives within a shared feature space, causing interference among attackers and exposing OOD cues

TABLE VIII
PERFORMANCE OF SPA COMPARED WITH BASELINE ATTACKS UNDER MULTI-LABEL ATTACK SCENARIO.

Attack		Vanilla	PGD	Neurotoxin	Chameleon	A3FL	SPA
Nodefense	ASR-A1	0.26	0.04	1.21	10.80	90.44	99.99
	ASR-A2	7.44	92.82	66.27	17.38	99.89	99.86
	ASR-A3	87.94	0.90	91.50	90.78	96.51	99.88
	ASR-Avg	31.88	31.25	52.99	39.65	95.61	**99.91**
	ACC-Avg	91.76	91.52	92.03	92.35	91.73	**92.39**
Deepsight	ASR-A1	0.52	0.03	2.41	86.73	99.88	99.85
	ASR-A2	5.80	92.66	29.40	15.34	92.14	98.76
	ASR-A3	88.81	0.56	85.91	7.91	97.66	97.52
	ASR-Avg	31.71	31.08	39.24	36.66	96.56	**98.71**
	ACC-Avg	91.44	90.81	92.31	92.35	92.17	**92.46**
Foolsgold	ASR-A1	1.31	0.04	73.26	10.07	99.60	98.33
	ASR-A2	5.37	94.00	95.09	77.32	98.71	98.96
	ASR-A3	90.81	0.97	74.39	48.99	87.58	96.38
	ASR-Avg	32.50	31.67	80.91	45.46	95.30	**97.89**
	ACC-Avg	88.79	90.71	91.04	90.66	89.00	**91.54**
Indicator	ASR-A1	0.91	0.24	49.74	21.62	99.96	97.77
	ASR-A2	0.02	1.83	0.26	2.04	11.51	99.65
	ASR-A3	85.34	57.74	16.02	0.97	99.53	97.36
	ASR-Avg	28.76	19.94	22.01	8.21	70.33	**98.26**
	ACC-Avg	85.08	85.83	85.13	86.13	85.75	**91.23**
Multikrum	ASR-A1	0.16	0.27	0.38	0.32	99.8	99.65
	ASR-A2	0.03	0.01	0.06	0.07	35.39	85.82
	ASR-A3	4.59	3.26	5.24	5.32	99.27	98.72
	ASR-Avg	1.59	1.18	1.89	1.90	78.15	**94.73**
	ACC-Avg	92.36	92.39	**92.64**	92.46	92.18	92.58
Rflbat	ASR-A1	0.29	0.00	0.02	6.57	65.13	98.72
	ASR-A2	2.74	99.79	0.06	11.89	99.90	96.42
	ASR-A3	88.63	0.03	99.73	92.99	98.46	98.62
	ASR-Avg	30.55	33.27	33.27	37.15	87.83	**97.92**
	ACC-Avg	86.83	88.28	87.22	86.97	85.46	**88.73**
Flame	ASR-A1	0.18	0.23	1.18	0.30	99.54	98.82
	ASR-A2	0.03	0.02	26.50	0.07	39.31	99.06
	ASR-A3	5.29	9.92	95.22	3.78	99.36	99.00
	ASR-Avg	1.83	3.39	40.97	1.38	79.40	**98.96**
	ACC-Avg	**92.56**	92.51	92.02	92.35	92.42	92.54

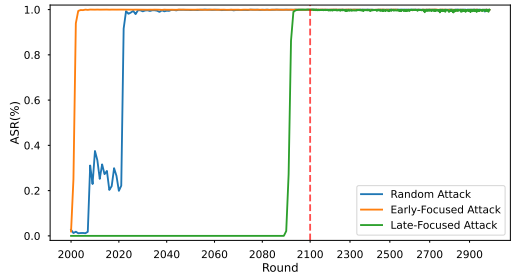


Fig. 11. Performance of SPA under different attack time intervals.

that indicator-based defenses can detect. A3FL mitigates this via trigger optimization, better separating OOD tasks and supporting multi-label attacks. However, its multiple OOD backdoor features still leave at least one attacker detectable due to interactions with defense-planted OOD indicators.

By contrast, SPA aligns each backdoor with its corresponding target-class manifold, distributing backdoor features within the natural class distributions. This alignment reduces inter-attack interference and blurs OOD signals, enabling evasion of detection. Effectively, SPA embeds multiple, independent covert channels within the feature space, maintaining high ASR for all attackers without mutual conflict.

Different Attack Timings. We evaluate the persistence of SPA under different attack schedules to assess its applicability in specialized settings. A single attacker operates within rounds 2000–2100, while federated training continues to 3000 rounds. We consider three scenarios: (i) random selection throughout the window, (ii) 10 consecutive attacks at the start (2000–2010), and (iii) 10 consecutive attacks at the end (2090–2100). As shown in Figure 11, SPA maintains strong effectiveness and persistence regardless of whether attacks occur early or late in the window. This robustness arises from aligning backdoor features with the model's evolving feature space, preserving influence as global training progresses without continuous adversarial presence. These results highlight the practical viability of SPA when attackers participate intermittently or have limited opportunities within the FL protocol.

TABLE IX
PERFORMANCE OF SPA WITH DIFFERENT TRIGGERS.

Type		Dataset	ACC	ASR
Fixed	Pixel	CIFAR-10	90.09	53.97
		CIFAR-100	68.83	69.34
		GTSRB	95.11	80.46
		SVHN	91.26	57.58
	Blend	CIFAR-10	89.22	68.05
		CIFAR-100	70.12	65.32
		GTSRB	94.42	90.24
		SVHN	92.06	67.59
Optimized	Pixel	CIFAR-10	92.58	97.9
		CIFAR-100	71.32	93.18
		GTSRB	96.25	99.49
		SVHN	93.07	92.72
	Blend	CIFAR-10	92.55	99.91
		CIFAR-100	71.63	99.97
		GTSRB	96.79	99.57
		SVHN	93.61	99.76

Different Trigger Types. We systematically assess how trigger patterns affect attack performance. Specifically, we compare optimized versus fixed triggers. Table IX shows that dynamically optimized triggers consistently achieve higher ASR than fixed ones, and blend triggers outperform pixel patterns across all configurations. This gap arises from the difficulty of aligning OOD features induced by static pixel triggers with the target feature distribution, as forcing such alignment can distort the target manifold and perturb learned features. Optimized triggers adapt during training, discovering effective alignment pathways that integrate more naturally into the representation space. This adaptability identifies low-resistance paths, minimizing disruption to legitimate features while maintaining high attack effectiveness.

D. Parameter Sensitivity and Ablation

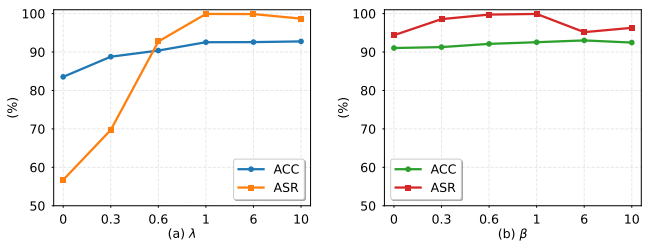


Fig. 12. Impact of loss term λ and β.

Loss Terms. We examine the sensitivity of two hyperparameters, λ and β. The former weights the utility loss $\mathcal{L}_{utility}$ to preserve model integrity during the backdoor process, while the latter weights a consistency loss $\mathcal{L}_{consist}$ that balances trigger effectiveness and stealth during trigger optimization.

TABLE X
ABLATION OF DIFFERENT COMPONENTS OF SPA.

Injection		Enhancement		ACC	ASR
$\mathcal{L}_{align}$	$\mathcal{L}_{utility}$	$\mathcal{L}_{enhance}$	$\mathcal{L}_{consist}$		
✓	✗	✗	✗	80.76 (↓11.79)	17.94 (↓81.97)
✓	✗	✓	✗	81.85 (↓10.70)	59.52 (↓40.39)
✓	✗	✓	✓	83.54 (↓9.01)	56.72 (↓43.19)
✓	✓	✗	✗	89.22 (↓3.33)	68.05 (↓31.86)
✓	✓	✓	✗	91.04 (↓1.54)	94.37 (↓5.54)
✓	✓	✓	✓	**92.55**	**99.91**

TABLE XI
PERFORMANCE OF SPA WITH DIFFERENT CONSTRAINT NORMS.

Constraint Method	ACC (%)	ASR (%)
Lp-Norms	91.51 (↓1.04)	88.79 (↓11.12)
KL Divergence	87.73 (↓4.82)	69.37 (↓30.54)
Cosine Similarity	92.12 (↓0.43)	94.58 (↓5.33)
Sliced-Wasserstein	**92.55**	**99.91**

We vary both over 0, 0.3, 0.6, 1, 6, 10. Results are shown in Figure 12. When $\lambda = 0$, both ASR and ACC degrade, indicating that ASR is coupled with overall model quality. As λ increases, ASR and ACC stabilize, suggesting that feature-space alignment complements, rather than competes with, standard learning by aligning backdoor features to the target-class manifold. Notably, even at large λ, the reduction in ASR is marginal, highlighting the robustness of the alignment strategy. For β, larger values mean stronger consistency, which limits backdoor effectiveness, but helps preserve benign representations. Conversely, with $\beta = 0$, meaning no consistency constraint, the alignment can distort benign features, reducing both ACC and ASR.

Component Contributions. SPA comprises two components: backdoor injection and backdoor enhancement. We ablate the contribution of each loss term, retaining $\mathcal{L}_{align}$ in all settings as it is the core backdoor injection. As shown in Table X, each loss term is essential. Removing $\mathcal{L}_{utility}$ markedly degrades ACC, confirming that utility preservation is necessary for effective backdoor injection. Because SPA aligns trigger features with the target-class manifold, degrading model utility distorts this manifold and hinders the formation of feature associations. While $\mathcal{L}_{enhance}$ substantially boosts ASR via adversarial trigger optimization, its benefits depend on $\mathcal{L}_{consist}$ to maintain feature-space consistency. Without this constraint, enhanced triggers drift from the target distribution and become more vulnerable to SOTA anomaly detectors. The projection-based distance in $\mathcal{L}_{consist}$ is particularly effective. By enforcing directional rather than magnitude alignment, it balances attack potency with stealth and improves robustness to model updates during federated aggregation.

Different Constraint Norms. This section examines how distance metrics affect backdoor injection during feature-space alignment. We compare three metrics: L_2 norm, cosine similarity, and KL divergence, with results in Table XI. The L_2 norm's point-wise constraint distorts salient feature channels, degrading both ACC and ASR. Cosine similarity better preserves the main task via directional alignment but imposes constraints that still hinder optimal backdoor injection. KL divergence performs worst, causing large drops in ACC and ASR, likely due to numerical instability in high-dimensional spaces and its dependence on accurate density estimation, both problematic in the dynamic, heterogeneous feature spaces of FL. In contrast, SPA employs Sliced Wasserstein Distance (SWD), which is well-suited to high-dimensional distribution matching. By approximating optimal transport through one-dimensional projections, SWD captures the evolving feature distributions in FL and enables robust, flexible alignment of backdoor features with the target-class distribution.

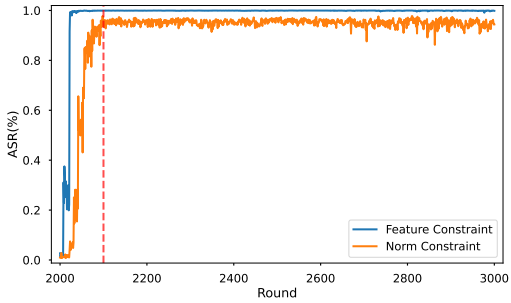


Fig. 13. Performance of SPA using different trigger constraints.

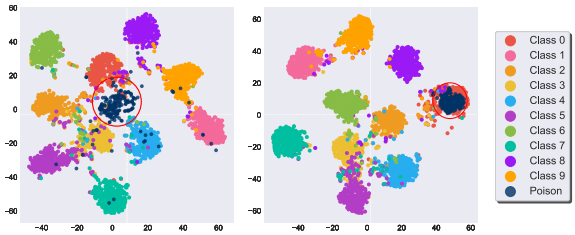


Fig. 14. T-SNE visualization of SPA using different trigger constraints.

Trigger Constraint. Furthermore, we compare two strategies for constraining trigger magnitude: (i) direct L-norm penalties and (ii) feature-consistency constraints, and visualize their effects using T-SNE. As shown in Figure 13, both approaches are effective, but L-norm constraints limit ASR. Figure 14 shows that L-norm–constrained triggers concentrate at the periphery of the target-class feature space, forming a discernible boundary that increases detection risk. In contrast, feature-consistency constraints produce trigger representations that closely resemble benign features and lie well within the target-class clusters, achieving deeper integration and greater stealth against detection.

VI. DISCUSSION

Computation Overhead Analysis. We measure the average runtime and GPU memory consumption per attacker participation round. Experiments are conducted on a single NVIDIA RTX 3090 GPU using the CIFAR-10 dataset with ResNet-18 as the backbone. By default, we optimize the trigger for 10 rounds and perform backdoor injection for 2 rounds. We adopt A3FL, a SOTA trigger optimization method, as our baseline and follow its original configuration with 100 rounds of trigger optimization and 2 rounds of backdoor injection. Experimental results are presented in Table XII. Admittedly, compared to benign participants, the attacker incurs additional local computation for trigger optimization and injection. This overhead is confined to the compromised client and does not introduce extra communication cost. Moreover, SPA requires significantly fewer trigger optimization rounds than A3FL.

TABLE XII
COMPUTATIONAL OVERHEAD OF SPA.

Method	Trigger optimization		Local training		Total time(s)	Peak memory (MB)
	each (s)	total (s)	each (s)	total (s)		
Benign	-	-	0.0539	0.1078 (2)	0.1078	480.47
A3FL	0.4040	80.8058 (200)	0.1035	0.2064 (2)	81.0122	567.34
SPA	0.0923	0.9234 (10)	0.2381	0.4762 (2)	1.3996	529.64

TABLE XIII
RESULTS WITH DIFFERENT TRIGGER OPTIMIZATION ROUNDS.

Step I	2	5	10	20	30
ASR	94.29	98.06	99.91	99.13	99.36
ACC	92.35	92.24	92.55	92.26	92.19
Optimization time (s)	0.2597	0.5185	0.9234	2.1429	3.0670

TABLE XIV
RESULTS WITH DIFFERENT NUMBER OF PROJECTIONS.

Number S	10	20	50	100	200
ASR	98.89	99.08	99.91	99.52	99.71
ACC	91.75	91.62	92.55	92.58	92.69
Training time (s)	0.3919	0.4201	0.4762	0.5185	0.5940

TABLE XV
RESULT ON DIFFERENT TARGET CLASS SAMPLE SIZES.

Sample Size	2	3	5	10	15
ACC	92.38	92.45	92.55	92.68	92.61
ASR	95.98	98.35	99.52	99.43	99.62
lifespan$_{60\%}$	692	824	900	900	900

The trigger optimization overhead of SPA primarily comes from the number of optimization rounds I, while the computational cost of local training is largely governed by the number of one-dimensional projections S used in the Sliced-Wasserstein distance. We investigate how these two factors affect our method by varying the trigger optimization rounds $I \in \{2, 5, 10, 20, 30\}$ and the number of one-dimensional projections $S \in \{10, 20, 50, 100, 200\}$. Results are shown in Tables XIII and XIV. We observe that reducing either the number of trigger optimization rounds or the precision of the Sliced-Wasserstein distance leads to performance degradation of SPA. With only 5 optimization rounds and $S = 20$, SPA already achieves acceptable attack performance.

Furthermore, from a system implementation perspective, server-side timing analysis or resource monitoring could potentially serve as a defense. However, reliably flagging malicious clients based solely on local training latency remains highly challenging in practical settings. Real-world FL deployments are inherently characterized by extreme system and statistical heterogeneity [49]–[51]. Variations in client hardware capabilities, battery levels, and network bandwidth can cause training latencies to differ by up to $100\times$ among participants [50]. To mitigate this, central servers typically employ generous aggregation time windows to accommodate benign stragglers or utilize asynchronous protocols that naturally tolerate high variance. Under such conditions, the moderate and controllable latency introduced by SPA is likely to blend into this massive natural variance, making it difficult for basic resource monitoring to confidently distinguish the attack overhead from normal straggler delays.

Impact of Target-Class Sample Size. We investigate how the number of target class samples held by the attacker affects the performance of SPA. We vary the number of target samples as $\{2, 3, 5, 10, 15\}$. Experimental results are presented in Table XV. As expected, a larger number of target samples leads to better attack performance. However, even with a very limited number of samples, SPA still achieves successful backdoor injection. The target samples essentially provide an anchor that characterizes the target class features. More samples naturally yield a more accurate characterization of these features. Yet a few samples already suffice to establish a reliable feature anchor. Our trigger optimization further reinforces this by naturally learning a trigger that aligns with the target class features, making feature alignment much easier.

Adaptive Defenses. In this section, we evaluate the resilience of SPA against adaptive defense mechanisms. Since SPA aligns representations in the feature space, its resulting gradients are statistically indistinguishable from benign updates. One potential line of defense involves techniques such as trigger reverse engineering or neuron activation analysis. While these methods are primarily designed for centralized learning and rely on access to clean data, we adapt three representative centralized defenses to the federated setting to rigorously test our attack: Neural Cleanse (NC) [27], ABS [52], and RNP [53]. We assume the server possesses a proxy dataset comprising 3% to 5% of the clean data distribution. NC and ABS employ model inversion to reconstruct potential triggers and flag anomalies using a decision index. Following their default configurations, we set the anomaly detection thresholds to 2.0 for NC and 0.80 for ABS. RNP, on the other hand, operates by iteratively pruning compromised channels based on neuron activation analysis.

Experimental results are presented in Table XVI. The findings show that SPA successfully bypasses these defense methods, with the attacked model exhibiting behavior indistinguishable from a clean model. This evasion stems from the fundamental nature of our feature-space alignment. Traditional attacks force the model to learn an isolated, out-of-distribution feature cluster in the latent space, which inevitably triggers detectable anomalies during reverse engineering and creates distinctive neuron activation paths. In contrast, SPA intrinsically binds the trigger representations within the natural feature manifold of the target class. By avoiding a divergent secondary task, the attack leaves no anomalous shortcut for reverse engineering to exploit and no isolated activation path for pruning to target.

Secure Aggregation. In our evaluation, we assume a standard FL setting where the central server has access to the plaintext updates of individual clients. This assumption is shared by all evaluated server-side defenses (e.g., DeepSight, FoolsGold), which rely on inspecting per-client updates to isolate anomalies. However, in privacy-stringent deployments, Secure Aggregation (SecAgg) protocols are often employed. Under SecAgg, the server only observes the aggregated global update, rendering the aforementioned defenses inapplicable. In such scenarios, the definition of attack stealth shifts from

TABLE XVI
PERFORMANCE OF SPA WITH ADAPTIVE DEFENSES.

Clean data	Method	No-defense		NC	ABS	RNP	
		ACC	ASR	Index		ACC	ASR
3%	No-Attack	92.61	-	1.45	0.39	91.21	-
	SPA	92.55	99.91	1.52	0.36	90.88	98.27
5%	No-Attack	92.61	-	1.38	0.43	91.78	-
	SPA	92.55	99.91	1.49	0.47	91.36	92.65

indistinguishability among per-client updates to preserving the global model's utility on the primary task. SPA remains highly potent and stealthy under SecAgg, as its feature-space alignment seamlessly integrates backdoor features without degrading the main task accuracy (shown in Table I).

VII. SUMMARY AND FUTURE WORK

In this paper, we introduced SPA, a novel backdoor attack framework targeting federated learning. Departing from conventional end-to-end supervised backdoor injection, SPA instead aligns feature representations of trigger-embedded samples with those of the target class, greatly enhancing stealth and long-term persistence. Our adversarial trigger optimization mechanism further improves the attack's adaptability and effectiveness, allowing backdoors to survive sophisticated aggregation schemes and persist long after adversarial participation ceases. Our comprehensive empirical evaluation demonstrates that SPA significantly outperforms state-of-the-art backdoor attacks across various metrics and scenarios. This highlights the demand for tailored defenses. In future work, we will explore feature-level backdoor defense methods to target complex dynamic backdoor attacks in FL.

ACKNOWLEDGMENTS

The authors would like to thank the reviewers for the time and effort they devoted to reviewing this paper. This work was partially funded by the National Natural Science Foundation of China under Grant (NSFC-62272222, NSFC-62272215), Jiangsu Province Outstanding Youth Fund Project (No. BK20230080), the 111 Center (No. B26023), the Nanjing University-China Mobile Communications Group Co.,Ltd. Joint Institute (NJ20250025), the Natural Science Foundation of Jiangsu Province (No. BK20251911), and the China Postdoctoral Science Foundation (No. 2025T180438).

REFERENCES

[1] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-Efficient Learning of Deep Networks from Decentralized Data," in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*. PMLR, Apr. 2017, pp. 1273–1282.

[2] A. Hard, K. Rao, R. Mathews, S. Ramaswamy, F. Beaufays, S. Augenstein, H. Eichner, C. Kiddon, and D. Ramage, "Federated Learning for Mobile Keyboard Prediction," Nov. 2018.

[3] Apple, "Private federated learning (neurips 2019 expo talk abstract)," https://nips.cc/ExpoConferences/2019/schedule?talkid=40, accessed: 2020-05-22.

[4] G. Long, Y. Tan, J. Jiang, and C. Zhang, "Federated learning for open banking," in *Federated learning: privacy and incentive*. Springer, 2020, pp. 240–254.

[5] M. J. Sheller, G. A. Reina, B. Edwards, J. Martin, and S. Bakas, "Multi-institutional Deep Learning Modeling Without Sharing Patient Data: A Feasibility Study on Brain Tumor Segmentation," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*. Cham: Springer International Publishing, 2019, pp. 92–104.

[6] Y. Liu, Y. Kang, T. Zou, Y. Pu, Y. He, X. Ye, Y. Ouyang, Y.-Q. Zhang, and Q. Yang, "Vertical federated learning: Concepts, advances, and challenges," *IEEE Transactions on Knowledge & Data Engineering*, vol. 36, no. 07, pp. 3615–3634, 2024.

[7] C.-M. Feng, K. Yu, N. Liu, X. Xu, S. Khan, and W. Zuo, "Towards instance-adaptive inference for federated learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 23 287–23 296.

[8] C. Fu, X. Zhang, S. Ji, J. Chen, J. Wu, S. Guo, J. Zhou, A. X. Liu, and T. Wang, "Label inference attacks against vertical federated learning," in *31st USENIX security symposium (USENIX Security 22)*, 2022, pp. 1397–1414.

[9] Z. Ma, J. Ma, Y. Miao, Y. Li, and R. H. Deng, "Shieldfl: Mitigating model poisoning attacks in privacy-preserving federated learning," *IEEE Transactions on Information Forensics and Security*, vol. 17, pp. 1639–1654, 2022.

[10] H. Zhang, J. Jia, J. Chen, L. Lin, and D. Wu, "A3FL: adversarially adaptive backdoor attacks to federated learning," in *Proceedings of the 37th International Conference on Neural Information Processing Systems*, 2023, pp. 61 213–61 233.

[11] E. Bagdasaryan, A. Veit, Y. Hua, D. Estrin, and V. Shmatikov, "How to backdoor federated learning," in *International conference on artificial intelligence and statistics*. PMLR, 2020, pp. 2938–2948.

[12] Z. Zhang, A. Panda, L. Song, Y. Yang, M. Mahoney, P. Mittal, R. Kannan, and J. Gonzalez, "Neurotoxin: Durable backdoors in federated learning," in *International Conference on Machine Learning*. PMLR, 2022, pp. 26 429–26 446.

[13] Y. Dai and S. Li, "Chameleon: Adapting to peer images for planting durable backdoors in federated learning," in *International Conference on Machine Learning*. PMLR, 2023, pp. 6712–6725.

[14] Q. Zhang, Y. Ding, Y. Tian, J. Guo, M. Yuan, and Y. Jiang, "Advdoor: adversarial backdoor attack of deep learning system," in *Proceedings of the 30th ACM SIGSOFT International Symposium on Software Testing and Analysis*, 2021, pp. 127–138.

[15] M. Li, W. Wan, Y. Ning, S. Hu, L. Xue, L. Y. Zhang, and Y. Wang, "Darkfed: A data-free backdoor attack in federated learning," *arXiv preprint arXiv:2405.03299*, 2024.

[16] C. Shi, S. Ji, X. Pan, X. Zhang, M. Zhang, M. Yang, J. Zhou, J. Yin, and T. Wang, "Towards practical backdoor attacks on federated learning systems," *IEEE Transactions on Dependable and Secure Computing*, 2024.

[17] T. Gu, K. Liu, B. Dolan-Gavitt, and S. Garg, "BadNets: Evaluating Backdooring Attacks on Deep Neural Networks," *IEEE Access*, vol. 7, pp. 47 230–47 244, 2019.

[18] Y. Li, X. Lyu, N. Koren, L. Lyu, B. Li, and X. Ma, "Anti-backdoor learning: Training clean models on poisoned data," *Advances in Neural Information Processing Systems*, vol. 34, pp. 14 900–14 912, 2021.

[19] C. Fung, C. J. M. Yoon, and I. Beschastnikh, "The limitations of federated learning in sybil settings," in *23rd International Symposium on Research in Attacks, Intrusions and Defenses, RAID 2020, San Sebastian, Spain, October 14-15, 2020*. USENIX Association, 2020, pp. 301–316.

[20] P. Blanchard, E. M. El Mhamdi, R. Guerraoui, and J. Stainer, "Machine learning with adversaries: Byzantine tolerant gradient descent," in *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates, Inc., 2017.

[21] P. Rieger, T. D. Nguyen, M. Miettinen, and A. Sadeghi, "Deepsight: Mitigating backdoor attacks in federated learning through deep model inspection," in *29th Annual Network and Distributed System Security Symposium, NDSS 2022, San Diego, California, USA, April 24-28, 2022*. The Internet Society, 2022.

[22] S. Li and Y. Dai, "BackdoorIndicator: Leveraging OOD data for proactive backdoor detection in federated learning," in *33rd USENIX Security Symposium (USENIX Security 24)*. Philadelphia, PA: USENIX Association, Aug. 2024, pp. 4193–4210.

[23] S. Huang, Y. Li, X. Yan, Y. Gao, C. Chen, L. Shi, B. Chen, and W. W. Ng, "Scope: On detecting constrained backdoor attacks in federated learning," *IEEE Transactions on Information Forensics and Security*, 2025.

[24] X. Lyu, Y. Han, W. Wang, J. Liu, B. Wang, J. Liu, and X. Zhang, "Poisoning with cerberus: Stealthy and colluded backdoor attack against

federated learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 7, 2023, pp. 9020–9028.

[25] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann, "Shortcut learning in deep neural networks," *Nature Machine Intelligence*, vol. 2, no. 11, pp. 665–673, 2020.

[26] J. Robinson, L. Sun, K. Yu, K. Batmanghelich, S. Jegelka, and S. Sra, "Can contrastive learning avoid shortcut solutions?" *Advances in neural information processing systems*, vol. 34, pp. 4974–4986, 2021.

[27] B. Wang, Y. Yao, S. Shan, H. Li, B. Viswanath, H. Zheng, and B. Y. Zhao, "Neural cleanse: Identifying and mitigating backdoor attacks in neural networks," in *2019 IEEE symposium on security and privacy (SP)*. IEEE, 2019, pp. 707–723.

[28] H. Wang, K. Sreenivasan, S. Rajput, H. Vishwakarma, S. Agarwal, J.-y. Sohn, K. Lee, and D. Papailiopoulos, "Attack of the tails: Yes, you really can backdoor federated learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 16 070–16 084, 2020.

[29] C. Xie, K. Huang, P. Chen, and B. Li, "DBA: distributed backdoor attacks against federated learning," in *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

[30] P. Fang and J. Chen, "On the vulnerability of backdoor defenses for federated learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 10, 2023, pp. 11 800–11 808.

[31] H. Li, Q. Ye, H. Hu, J. Li, L. Wang, C. Fang, and J. Shi, "3dfed: Adaptive and extensible framework for covert backdoor attack in federated learning," in *44th IEEE Symposium on Security and Privacy, SP 2023, San Francisco, CA, USA, May 21-25, 2023*. IEEE, 2023, pp. 1893–1907.

[32] T. Chen, B. Du, J. Zhao, J. Wang, H. Wang, Y. Li, B. Chen, and W. Lv, "Parasite: Planting durable backdoors against continual federated model fusion via discriminative feature pushing-pulling," *Information Fusion*, p. 104081, 2025.

[33] W. Huang, G. Li, M. Chen, J. Li, and H. Zhu, "Silent penetrator: Breaching cross-domain federated fine-tuning via feature shift-induced backdoor," *IEEE Transactions on Information Forensics and Security*, 2025.

[34] X. Lyu, Y. Han, W. Wang, J. Liu, B. Wang, J. Liu, and X. Zhang, "Poisoning with cerberus: Stealthy and colluded backdoor attack against federated learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 7, 2023, pp. 9020–9028.

[35] Y. Li, Y. Zhao, C. Zhu, and J. Zhang, "Infighting in the dark: Multi-label backdoor attack in federated learning," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 25 770–25 779.

[36] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[37] Y. Wang, D. Zhai, Y. Zhan, and Y. Xia, "RFLBAT: A robust federated learning algorithm against backdoor attack," *CoRR*, vol. abs/2201.03772, 2022.

[38] T. D. Nguyen, P. Rieger, H. Chen, H. Yalame, H. Möllering, H. Fereidooni, S. Marchal, M. Miettinen, A. Mirhoseini, S. Zeitouni, F. Koushanfar, A. Sadeghi, and T. Schneider, "FLAME: taming backdoors in federated learning," in *31st USENIX Security Symposium, USENIX Security 2022, Boston, MA, USA, August 10-12, 2022*. USENIX Association, 2022, pp. 1415–1432.

[39] L. Miao, W. Yang, R. Hu, L. Li, and L. Huang, "Against backdoor attacks in federated learning with differential privacy," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 2999–3003.

[40] C. Xie, M. Chen, P.-Y. Chen, and B. Li, "Crfl: Certifiably robust federated learning against backdoor attacks," in *International Conference on Machine Learning*. PMLR, 2021, pp. 11 372–11 382.

[41] M. S. Ozdayi, M. Kantarcioglu, and Y. R. Gel, "Defending against backdoors in federated learning with robust learning rate," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 10, 2021, pp. 9268–9276.

[42] H. Wang, K. Sreenivasan, S. Rajput, H. Vishwakarma, S. Agarwal, J. Sohn, K. Lee, and D. S. Papailiopoulos, "Attack of the tails: Yes, you really can backdoor federated learning," 2020.

[43] G. Tao, Z. Wang, S. Feng, G. Shen, S. Ma, and X. Zhang, "DRUPE: Distribution Preserving Backdoor Attack in Self-supervised Learning," in *2024 IEEE Symposium on Security and Privacy (SP)*. IEEE Computer Society, 2023, pp. 29–29.

[44] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images." Toronto, ON, Canada, 2009.

[45] S. Houben, J. Stallkamp, J. Salmen, M. Schlipsing, and C. Igel, "Detection of traffic signs in real-world images: The german traffic sign detection benchmark," in *The 2013 international joint conference on neural networks (IJCNN)*. IEEE, 2013, pp. 1–8.

[46] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, A. Y. Ng *et al.*, "Reading digits in natural images with unsupervised feature learning," in *NIPS workshop on deep learning and unsupervised feature learning*, vol. 2011, no. 5. Granada, 2011, p. 7.

[47] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 2016, pp. 770–778. [Online]. Available: https://doi.org/10.1109/CVPR.2016.90

[48] Z. Zhu, J. Hong, and J. Zhou, "Data-free knowledge distillation for heterogeneous federated learning," in *International Conference on Machine Learning*. PMLR, 2021, pp. 12 878–12 889.

[49] J. Sun, A. Li, L. Duan, S. Alam, X. Deng, X. Guo, H. Wang, M. Gorlatova, M. Zhang, H. Li *et al.*, "Fedsea: A semi-asynchronous federated learning framework for extremely heterogeneous devices," in *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems*, 2022, pp. 106–119.

[50] Z. Li, P. Chaturvedi, S. He, H. Chen, G. Singh, V. Kindratenko, E. A. Huerta, K. Kim, and R. Madduri, "Fedcompass: Efficient cross-silo federated learning on heterogeneous client devices using a computing power-aware scheduler," in *The Twelfth International Conference on Learning Representations*, 2024.

[51] C. Chen, Y. Zhao, Z. Zhang, W. Li, and J. Wu, "Towards efficient and scalable asynchronous federated learning via stragglers version control," *IEEE Transactions on Mobile Computing*, 2025.

[52] Y. Liu, W.-C. Lee, G. Tao, S. Ma, Y. Aafer, and X. Zhang, "Abs: Scanning neural networks for back-doors by artificial brain stimulation," in *Proceedings of the 2019 ACM SIGSAC conference on computer and communications security*, 2019, pp. 1265–1282.

[53] Y. Li, X. Lyu, X. Ma, N. Koren, L. Lyu, B. Li, and Y.-G. Jiang, "Reconstructive neuron pruning for backdoor defense," in *International Conference on Machine Learning*. PMLR, 2023, pp. 19 837–19 854.