ShadowFPT: Backdooring Federated Prompt Tuning via Shadow Triggers

Kun Zhai[1] **Teng Li**[1] **Yunhao Feng**[2] **Xingjun Ma**[1,3]*
[1]Fudan University, Shanghai, China
[2]Alibaba Group, Hangzhou, China
[3]Shanghai Innovation Institute, Shanghai, China

Abstract

Federated Prompt Tuning (FPT) adapts large vision–language models by freezing the pretrained backbone and optimizing only lightweight prompt parameters across clients. Although this design improves communication and parameter efficiency, it creates an overlooked security risk. Since the backbone is shared and frozen, a malicious client can induce backdoor behavior through the representation space while keeping its uploaded prompt updates close to benign ones. We propose **ShadowFPT**, a targeted backdoor attack that exploits this frozen-backbone attack surface. ShadowFPT first pretrains a learnable *Shadow Trigger* against the frozen CLIP visual encoder, using either auxiliary public data or the malicious client's local data, so that triggered inputs are steered toward the target class in representation space. During federated prompt tuning, the malicious client adapts the trigger under the current global prompt and then optimizes its local prompt on both clean and triggered samples. Only prompt parameters are uploaded to the server, while the trigger and frozen encoders remain local. By shifting most of the attack burden from prompt updates to trigger-induced representation steering, ShadowFPT achieves targeted misclassification while preserving prompt-space stealthiness. Across multiple datasets, aggregation rules, and non-IID federated partitions, ShadowFPT increases the attack success rate from 19.81% to 90.36% in our main setting, while maintaining clean accuracy. It remains effective across textual, visual, and joint vision–language prompt tuning. These results identify frozen backbones as stealthy and underexplored backdoor surfaces in federated prompt tuning, suggesting that defenses based only on prompt-update anomaly detection are insufficient.

1 Introduction

Large vision–language models (VLMs), such as CLIP Radford et al. [2021], learn transferable image–text representations and have become a standard foundation for multimodal recognition and adaptation. Adapting these models in distributed environments, however, remains difficult because full fine-tuning of large pretrained backbones is computationally expensive and communication-intensive. Prompt tuning addresses these limitations by keeping the pretrained backbone fixed and optimizing only a small set of learnable prompt parameters Khattak et al. [2023a]. This parameter-efficient formulation is particularly well suited to federated learning. In *Federated Prompt Tuning* (FPT), clients update lightweight prompts using private local data and communicate only prompt parameters to the server, while the VLM backbone remains frozen.

Backdoor attacks remain a persistent security threat in federated learning. Classical attacks, such as BadNets Gu et al. [2017] and Blended Chen et al. [2017], poison training data with attacker-designed triggers, causing the trained model to retain high clean accuracy while predicting an attacker-specified

*Corresponding author: xingjunma@fudan.edu.cn

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

target label on triggered inputs. In federated learning, malicious clients can further inject backdoors through their local model updates, making attack persistence under server aggregation a central challenge. Prior work has studied this problem in standard and personalized federated learning, including A3FL Zhang et al. [2023] and Bad-PFL Fan et al. [2025]. Backdoor risks have also been examined in vision–language models and prompt learning. For example, BadCLIP attacks multimodal contrastive learning by guiding both visual and textual embeddings Liang et al. [2024]. However, backdoor attacks against federated prompt tuning remain insufficiently understood. In particular, existing studies largely overlook the security risk that arises when the server aggregates only lightweight prompt parameters while all clients rely on a shared frozen VLM backbone.

Most federated backdoor defenses operate on the updates communicated by clients. Representative methods identify updates that are close to the client majority Blanchard et al. [2017], remove coordinate-wise outliers Yin et al. [2018], downweight clients with highly similar update directions Fung et al. [2018], calibrate client updates using trusted server data Cao et al. [2020], or combine clustering, clipping, noise injection, and certification to improve robustness Nguyen et al. [2022]. Recent work extends this update-centered view to federated prompt learning. For example, SABRE-FL detects malicious clients through anomaly analysis in embedding space Khan et al. [2025]. Despite their methodological differences, these defenses share a common assumption that backdoor behavior will manifest as detectable anomalies in uploaded updates or in representations derived from them.

This assumption becomes fragile in FPT. The server observes only low-dimensional prompt updates, while the dominant representation geometry is determined by the shared frozen VLM backbone. A malicious client can therefore shift much of the backdoor effect to the interaction between the trigger and the frozen representation space, while keeping its uploaded prompt updates close to benign ones. This creates a security gap between where the backdoor behavior is realized and what the server can observe. The same frozen backbone that enables efficient adaptation also provides a stable and underexplored attack surface for stealthy backdoor injection.

Motivated by this observation, we propose **ShadowFPT**, a targeted backdoor attack for federated prompt tuning that exploits the shared frozen CLIP visual encoder as a representation-level attack surface rather than relying on conspicuous prompt updates. ShadowFPT consists of two stages. First, it pretrains a learnable *Shadow Trigger* using either auxiliary public data or the malicious client's local data. The trigger is optimized to steer triggered images toward the target class in the frozen CLIP representation space by aligning final image embeddings with the target text anchor and intermediate visual features with target-class layer-wise prototypes. Second, during federated training, the malicious client adapts the Shadow Trigger under the current global prompt and then optimizes its local prompt on both clean and triggered samples to implant the targeted backdoor. Only prompt parameters are uploaded to the server, while the trigger and frozen encoders remain local. By shifting much of the attack effect from uploaded prompt updates to trigger-induced representation steering, ShadowFPT achieves targeted misclassification while keeping malicious prompt updates close to benign updates.

We evaluate ShadowFPT across multiple datasets, aggregation rules, and non-IID partitions. In the main setting, ShadowFPT improves attack success by 70.55 percentage points while largely preserving clean accuracy. It remains effective across textual, visual, and joint vision–language prompt tuning, and prompt-space analyses show that malicious updates stay close to benign ones. These results reveal frozen backbones as stealthy backdoor surfaces and challenge defenses based solely on prompt-update anomaly detection.

The main contributions of this paper are as follows:

- We identify a frozen-backbone backdoor surface in FPT, showing that targeted backdoor behavior can be realized through trigger-induced representation steering while malicious prompt updates remain close to benign updates.
- We propose **ShadowFPT**, a targeted backdoor attack that shifts the attack burden from conspicuous prompt-space poisoning to representation-level manipulation of the shared frozen backbone.
- We introduce the *Shadow Trigger*, a target-directed trigger learned through target-text alignment and layer-wise visual prototype alignment, enabling triggered images to move toward the target class in the frozen representation space.

- We evaluate ShadowFPT across textual, visual, and joint vision–language prompt tuning under non-IID federated settings, demonstrating high attack success, minimal clean-accuracy degradation, and strong prompt-space stealthiness.

2 Related Work

Federated Prompt Tuning. Vision–language models (VLMs), such as CLIP, learn transferable image–text representations and provide strong foundations for multimodal adaptation Radford et al. [2021]. Prompt tuning adapts these models efficiently by freezing the pretrained backbone and optimizing only a small set of prompt parameters. Representative CLIP-based approaches include textual prompt tuning Zhou et al. [2022b,a], visual prompt tuning Jia et al. [2022], joint vision–language prompt learning Khattak et al. [2023a], and semantic regularization Khattak et al. [2023b]. This parameter-efficient paradigm naturally extends to federated learning, where clients communicate prompt parameters rather than full model updates. Recent studies on federated prompt tuning investigate cooperative prompt learning Guo et al. [2023b], personalization Guo et al. [2023a], text-driven prompt generation Qiu et al. [2024], personalization–generalization trade-offs Cui et al. [2024], mixture-of-experts prompts Luo et al. [2025], and multimodal visual prompt tuning Singha et al. [2025]. These works primarily focus on communication efficiency, personalization, and generalization under heterogeneous client data. In contrast, the security risks of federated prompt tuning, especially its vulnerability to backdoor attacks, remain insufficiently understood.

Backdoor Attacks. Backdoor attacks aim to make a model behave normally on clean inputs while predicting an attacker-specified target label on inputs containing a trigger. Classical attacks achieve this behavior by poisoning training data with visible or blended triggers Gu et al. [2017], Chen et al. [2017]. In federated learning, the backdoor must also persist through local training and server aggregation, which has motivated attacks based on stealthy poisoned updates, persistent backdoor injection, adaptive optimization, collusive trigger design, and personalized federated learning Lyu et al. [2023], Zhang et al. [2022, 2023], Lyu et al. [2024a], Fan et al. [2025].Backdoor vulnerabilities have also been studied in vision–language models. For example, BadCLIP manipulates multimodal contrastive representations through visual and textual embedding guidance Liang et al. [2024]. Existing attacks, however, are largely built around data poisoning, model-update manipulation, or direct prompt-parameter poisoning. FPT presents a different regime in which both the trainable and communicated parameters are small, and local optimization is lightweight. These constraints leave limited capacity for conventional backdoor injection through uploaded updates, motivating our investigation of a more concealed representation-level attack surface.

Backdoor Defenses. Federated defenses against malicious clients have largely followed the paradigm of Byzantine-robust aggregation. Classical methods, such as Krum Blanchard et al. [2017], Trimmed-Mean Yin et al. [2018], and Bulyan Guerraoui et al. [2018], mitigate malicious behavior by suppressing abnormal client updates through robust statistics and outlier filtering. Subsequent approaches improve robustness under heterogeneous or unreliable clients using geometric aggregation, trust bootstrapping, or sparsification Pillutla et al. [2022], Karimireddy et al. [2022], Cao et al. [2020], Xu et al. [2025]. Defenses designed specifically for federated backdoors further target poisoned global models. FLAME combines client clustering with model clipping Nguyen et al. [2022], while CRFL provides certified robustness against federated backdoors Xie et al. [2021]. Recent studies also investigate secure aggregation and detectability, privacy-constrained robustness under heterogeneity, system- and optimization-level mitigation, and dynamic behavior-based identification Huang et al. [2025], Zhang et al. [2025].

Despite their methodological differences, these defenses largely share the assumption that malicious behavior will appear as detectable abnormalities in uploaded updates or in representations derived from them. This assumption becomes fragile in FPT. Under non-IID data, benign prompt updates may vary substantially across clients, while the backdoor effect can be shifted into the interaction between the trigger and the shared frozen representation space rather than being exposed through conspicuous prompt updates. This gap motivates our investigation of frozen-backbone representation spaces as stealthy backdoor surfaces in federated prompt tuning.

3 Method

3.1 Problem Setup and Threat Model

Federated Prompt Tuning. We consider federated prompt tuning of a CLIP model composed of a visual encoder f_I and a text encoder f_T, both of which remain frozen throughout training. The system contains N clients, where client i holds a private local dataset $\mathcal{D}_i = \{(x_j, y_j)\}_{j=1}^{n_i}$ drawn from a potentially non-IID distribution. Let $\mathcal{C}$ denote the downstream label set. For clarity, we first present the formulation for textual prompt tuning.

Following standard CLIP prompt tuning, each class $c \in \mathcal{C}$ is represented by a prompted text input, such as "a photo of a [CLASS]", where the hand-crafted context tokens are replaced by learnable continuous tokens. We denote the resulting prompted text feature by $f_T(c; P)$. Given an image x and prompt parameters P, the prompted CLIP classifier computes the class logit as

$$\ell_c(x; P) = \langle f_I(x), f_T(c; P) \rangle , \qquad c \in \mathcal{C}, \tag{1}$$

where $f_I(x)$ and $f_T(c; P)$ are the image and prompted text features, respectively, and $\langle \cdot, \cdot \rangle$ denotes cosine similarity. We write $F(x; P)$ for the resulting prompted classifier.

At communication round t, the server broadcasts the current global prompt P^t to a subset of clients $\mathcal{S}_t$. Each selected client initializes its local prompt with P^t and optimizes the standard classification objective:

$$\mathcal{L}(P_i) = \mathbb{E}_{(x,y)\sim\mathcal{D}_i} \left[\mathrm{CE}\big(F(x; P_i), y\big) \right], \tag{2}$$

where $\mathrm{CE}(\cdot, \cdot)$ denotes the cross-entropy loss. After local optimization, clients upload their prompt parameters to the server, which aggregates the received prompts using an aggregation rule $\mathrm{Agg}(\cdot)$:

$$P^{t+1} = \mathrm{Agg}\left(\{ P_i^{t+1} : i \in \mathcal{S}_t \} \right). \tag{3}$$

Threat Model. We consider a targeted backdoor attack against FPT with a frozen vision–language backbone. The adversary aims to preserve the clean accuracy of the global model while causing triggered inputs, $x^\delta = \mathcal{A}(x, \delta)$, to be classified as an attacker-specified target class y_t. The malicious client has access to local target-class samples, denoted by $\mathcal{R}_{y_t}$, which are used to guide target-directed trigger optimization. During local training, the malicious client can optimize both its local prompt parameters and the trigger δ, but uploads only prompt parameters to the server. The adversary cannot modify the frozen image encoder, the frozen text encoder, the global training protocol, or the server-side aggregation rule. We assume a white-box threat setting in which the malicious client knows the frozen backbone architecture.

3.2 Overview of ShadowFPT

ShadowFPT backdoors FPT through a two-stage procedure. It first learns a *Shadow Trigger* against the frozen CLIP representation space, and then uses this trigger to implant targeted backdoor behavior during malicious local prompt training. Figure 1 provides an overview of the ShadowFPT.

Before federated training, the adversary optimizes a learnable trigger δ on a pretraining dataset $\mathcal{D}_\mathrm{s}$, which can be either auxiliary public data or the malicious client's local data. The optimized trigger, denoted by δ_shadow, is trained to steer triggered images toward the target class in the frozen CLIP representation space, as described in Section 3.3.

During federated training, the malicious client initializes the trigger with δ_shadow and adapts it under the current global prompt, allowing the trigger to remain aligned with the evolving federated prompt. The adapted trigger is then applied to a subset of the malicious client's local data, and the local prompt is optimized on both clean and triggered samples so that triggered inputs are classified as the target class. After local optimization, the malicious client uploads only the prompt parameters to the server, while the trigger remains local. This injection procedure is detailed in Section 3.4.

3.3 Shadow Trigger Pretraining

ShadowFPT learns a *Shadow Trigger* by optimizing a learnable trigger δ against the frozen CLIP visual encoder. Let $\mathcal{D}_\mathrm{s}$ denote the data used for trigger pretraining, which can be either auxiliary public data or the malicious client's local data. For an image $x \sim \mathcal{D}_\mathrm{s}$, the triggered image is defined

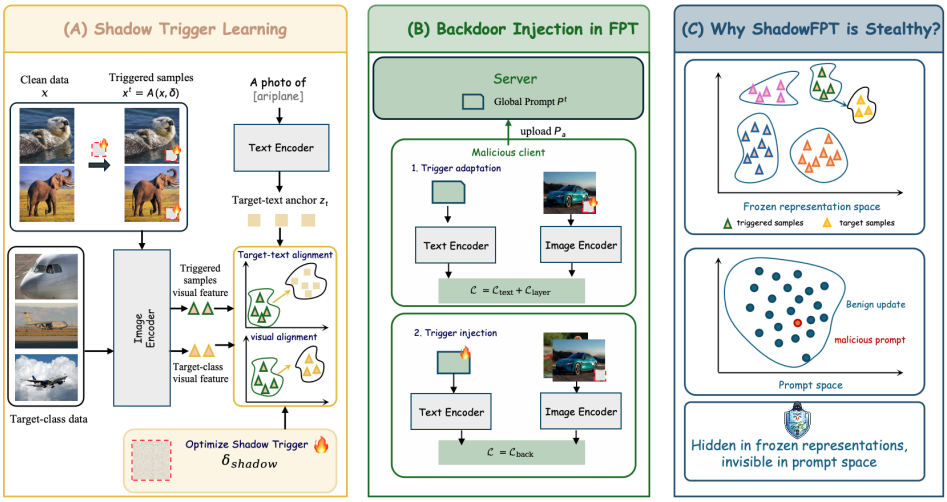


Figure 1: Overview of ShadowFPT. (A) The adversary pretrains a Shadow Trigger δ_{shadow} against the frozen CLIP backbone through target-text and layer-wise visual alignment. (B) During FPT, the malicious client adapts the trigger under the current global prompt P^t and optimizes its local prompt on clean and triggered samples, while uploading only prompt parameters. (C) By shifting the attack effect to trigger-induced representation steering, ShadowFPT achieves targeted misclassification with benign-like prompt updates.

as $x^\delta = \mathcal{A}(x, \delta)$. During this stage, only the trigger δ is updated, while both the visual encoder f_I and the text encoder f_T remain fixed. The goal is to obtain a trigger that steers triggered images toward the target class in the frozen representation space.

Target-text alignment. During Shadow Trigger pretraining, we use fixed hand-crafted CLIP templates instead of the learnable federated prompt. For each class $c \in \mathcal{C}$, we construct a text anchor using the frozen text encoder:

$$z_c = \frac{1}{|\mathcal{T}|} \sum_{T \in \mathcal{T}} f_T(T(c)), \tag{4}$$

where $\mathcal{T}$ is a fixed set of templates, such as "a photo of a [CLASS]". Let z_t denote the text anchor of the target class y_t. We optimize the trigger so that the image feature of each triggered input is aligned with z_t:

$$\mathcal{L}_{\text{text}} = \mathbb{E}_{x \sim \mathcal{D}_{\text{s}}} \left[1 - \left\langle f_I(x^\delta), z_t \right\rangle \right], \tag{5}$$

where $\langle \cdot, \cdot \rangle$ denotes cosine similarity. This loss encourages the Shadow Trigger to steer triggered images toward the target-class semantic region in the frozen CLIP space.

Layer-wise visual alignment. Target-text alignment constrains the final image–text similarity, but it does not explicitly control how the trigger changes intermediate visual representations. We therefore align triggered features with target-class prototypes within the frozen visual encoder. For each selected visual layer $l \in \mathcal{K}$, we compute a target-class prototype from the target-class image set $\mathcal{R}_{y_t}$:

$$\mu_l^{y_t} = \frac{1}{|\mathcal{R}_{y_t}|} \sum_{x \in \mathcal{R}_{y_t}} h_l(x), \tag{6}$$

where $h_l(x)$ denotes the layer-l visual feature. The layer-wise alignment loss is defined as

$$\mathcal{L}_{\text{layer}} = \mathbb{E}_{x \sim \mathcal{D}_{\text{s}}} \left[\frac{1}{|\mathcal{K}|} \sum_{l \in \mathcal{K}} \left(1 - \left\langle h_l(x^\delta), \mu_l^{y_t} \right\rangle \right) \right]. \tag{7}$$

This term encourages triggered images to follow the target-class representation trajectory through the frozen visual backbone.

Objective. The Shadow Trigger is pretrained by minimizing

$$\mathcal{L}_{\text{shadow}} = \mathcal{L}_{\text{text}} + \lambda_l \mathcal{L}_{\text{layer}}. \tag{8}$$

The optimized trigger is denoted by δ_{shadow} and is used to initialize the malicious client's trigger during federated prompt tuning.

3.4 Backdoor Injection in FPT

During federated training, ShadowFPT uses the pretrained Shadow Trigger to implant targeted backdoor behavior into the federated prompt. At communication round t, the malicious client receives the current global prompt P^t from the server and initializes its local prompt as $P_a \leftarrow P^t$. It then performs two sequential local steps. First, it adapts the trigger under the current global prompt. Second, it optimizes the local prompt on poisoned local data.

Trigger adaptation. In the first attack round, the malicious client initializes the trigger with δ_{shadow} and then updates it locally across subsequent rounds. At round t, the prompt is fixed to the current global prompt P^t, and only the trigger is optimized. We construct the target text anchor under the current global prompt as: $z^{(t)}_{\text{target}} = f_T(y_t; P^t)$. The trigger is adapted using the same representation-level objective as in Shadow Trigger pretraining:

$$\mathcal{L}^{(t)}_{\text{adapt}} = \mathcal{L}^{(t)}_{\text{text}} + \lambda_l \mathcal{L}^{(t)}_{\text{layer}}. \tag{9}$$

Here, $\mathcal{L}^{(t)}_{\text{text}}$ aligns triggered image embeddings with the current target text anchor $z^{(t)}_{\text{target}}$, while $\mathcal{L}^{(t)}_{\text{layer}}$ keeps intermediate visual features close to the target-class visual prototypes defined in Section 3.3. This step produces a prompt-adapted trigger $\delta^{(t)}$.

Backdoor prompt optimization. With the adapted trigger $\delta^{(t)}$ fixed, the malicious client optimizes its local prompt P_a using both clean and triggered local samples. Let $\mathcal{D}_a$ denote the malicious client's local dataset. The prompt-level backdoor objective is

$$\mathcal{L}_{\text{back}}(P_a) = \mathbb{E}_{(x,y)\sim\mathcal{D}_a}\left[\text{CE}\big(F(x; P_a), y\big)\right] + \mathbb{E}_{(x,y)\sim\mathcal{D}_a,\, y\neq y_t}\left[\text{CE}\big(F(x^{\delta^{(t)}}; P_a), y_t\big)\right]. \tag{10}$$

The first term preserves clean-task performance, whereas the second term implants the targeted backdoor into the local prompt.

After local optimization, the malicious client uploads only the prompt parameters P_a to the server. The adapted trigger $\delta^{(t)}$ and the frozen CLIP encoders are never communicated. The server then aggregates the uploaded prompts using the same aggregation rule as in standard FPT.

4 Experiments

4.1 Experimental Setup

Federated Setting. All methods use a frozen CLIP model with ViT-B/16 as the backbone. Unless otherwise specified, we use textual prompt tuning as the default FPT setting, denoted by FPT-L Li et al. [2024]. To evaluate whether ShadowFPT transfers across different prompt parameterizations, we also consider visual prompt tuning (FPT-V) Jia et al. [2022] and joint vision–language prompt tuning (FPT-VL) Khattak et al. [2023a]. Following FedTPG Qiu et al. [2024], we simulate a non-IID federation with 20 clients. Each client holds samples from five classes, and the class sets are disjoint across clients. All clients participate in every communication round.

Unless otherwise stated, a single client is malicious, and the target class is set to the first class of each dataset. All methods are trained for 20 communication rounds with one local epoch per round, batch size 32, and learning rate 0.0035 for local prompt optimization. We report clean accuracy (Acc) on benign test samples and attack success rate (ASR) on triggered non-target test samples.

ShadowFPT Variants. We evaluate two variants of ShadowFPT that differ in the data used for Shadow Trigger pretraining. **ShadowFPT-I** pretrains the Shadow Trigger using only the malicious client's local data, whereas **ShadowFPT-II** uses auxiliary public data, instantiated as 100 randomly sampled ImageNet classes. We do not assume semantic overlap between the auxiliary ImageNet classes and the downstream target classes. Both variants use a 32×32 trigger patch placed at the bottom-right corner of the image. The Shadow Trigger is optimized with learning rate 0.01 during pretraining and 0.001 during federated trigger adaptation.

Datasets. We evaluate ShadowFPT and all baselines on Caltech101 Fei-Fei et al. [2004], Oxford Flowers Nilsback and Zisserman [2008], and UCF101 Soomro [2012]. These datasets cover generic object recognition, fine-grained flower classification, and human action recognition, respectively.

Attack Baselines. We compare ShadowFPT with three groups of attack baselines. The first group includes generic backdoor attacks, namely **BadNets** Gu et al. [2017] and **Blended** Chen et al. [2017]. The second group includes federated backdoor attacks, namely **Neurotoxin** Zhang et al. [2022], **A3FL** Zhang et al. [2023], **FCBA** Liu et al. [2024], and **CerP** Lyu et al. [2023]. The third group includes personalized federated backdoor attacks, namely **Bad-PFL** Fan et al. [2025] and **PFedBA** Lyu et al. [2024b].

Defense Baselines. We evaluate defenses covering several aggregation and filtering mechanisms. The first category is standard aggregation, represented by **FedAvg** McMahan et al. [2017]. The second category is Byzantine-robust aggregation, including **Krum** Blanchard et al. [2017], **TrimmedMean** Yin et al. [2018], and **Bucketing** Karimireddy et al. [2022]. The third category includes adaptive or similarity-based filtering methods, namely **LaSA** Xu et al. [2025] and **FoolsGold** Fung et al. [2018]. We also evaluate trust-based aggregation with **FLTrust** Cao et al. [2020], as well as backdoor-specific defenses, including **FLAME** Nguyen et al. [2022] and **CRFL** Xie et al. [2021].

4.2 Main Results

Table 1: Overall attack performance across datasets and prompt-tuning settings. Results are averaged over all aggregation and defense rules. Best Base. denotes the strongest non-ShadowFPT baseline under each dataset-setting pair, selected by the highest average ASR. ΔAcc and ΔASR report absolute differences from Best Base., measured in percentage points.

Dataset	Setting	Best Base	Baseline		ShadowFPT-I				ShadowFPT-II			
			Acc	ASR	Acc	ASR	ΔAcc	ΔASR	Acc	ASR	ΔAcc	ΔASR
Caltech101	FPT-L	A3FL	88.49	19.81	88.32	84.28	-0.17	+64.47	87.67	**90.36**	-0.82	**+70.55**
	FPT-V	PFedBA	92.73	37.84	92.62	92.93	-0.11	+55.09	91.68	**94.25**	-1.05	**+56.41**
	FPT-VL	PFedBA	92.82	51.39	92.39	93.19	-0.43	+41.80	92.28	**98.03**	-0.54	**+46.64**
Oxford Flowers	FPT-L	A3FL	66.63	38.07	60.93	96.74	-5.70	+58.67	60.85	**98.93**	-5.78	**+60.86**
	FPT-V	PFedBA	65.73	30.22	66.60	98.37	+0.87	+68.15	66.61	**98.94**	+0.88	**+68.72**
	FPT-VL	PFedBA	69.89	54.60	68.52	98.07	-1.37	+43.47	68.33	**99.18**	-1.56	**+44.58**
UCF101	FPT-L	PFedBA	68.02	25.06	66.05	99.81	-1.97	+74.75	65.87	**99.98**	-2.15	**+74.92**
	FPT-V	PFedBA	65.76	31.55	65.28	98.04	-0.48	+66.49	65.21	**99.69**	-0.55	**+68.14**
	FPT-VL	Blended	70.14	22.49	67.97	99.28	-2.17	+76.79	67.89	**99.63**	-2.25	**+77.14**

Overall effectiveness. We first evaluate whether ShadowFPT can implant targeted backdoors while preserving clean-task performance. Table 1 reports results averaged over all aggregation and defense rules, with detailed results for all datasets, baselines, and defenses provided in Appendix 6. Because Best Base denotes the strongest non-ShadowFPT attack for each dataset-setting pair, this comparison is conservative. ShadowFPT consistently outperforms Best Base across all datasets and prompt-tuning settings. Averaged over the nine settings, Best Base achieves 34.56% ASR, whereas ShadowFPT-I and ShadowFPT-II reach 95.63% and 97.67%, respectively. ShadowFPT-II is the strongest method in every setting, improving ASR by 44.58 to 77.14 percentage points.

These gains are achieved with limited degradation in clean accuracy. On Caltech101, ShadowFPT-II reduces clean accuracy by at most 1.05 points. On UCF101, the drop ranges from 0.55 to 2.25 points despite near-100% ASR. Oxford Flowers under FPT-L shows a larger drop of 5.78 points, suggesting a stronger clean-attack trade-off in fine-grained recognition. This trade-off is much weaker under FPT-V and FPT-VL, where ShadowFPT-II maintains comparable clean accuracy while achieving 98.94% and 99.18% ASR. Overall, these results show that ShadowFPT substantially improves targeted attack success without primarily relying on degrading benign performance.

Robustness against defenses. We next examine whether ShadowFPT remains effective against aggregation and defense rules that act on uploaded client updates. As shown in Table 2, ShadowFPT-II achieves the highest ASR under every defense, ranging from 96.63% to 98.30%. In contrast, the strongest non-ShadowFPT baseline reaches only 13.17% to 60.93% ASR. The gains are especially pronounced under robust and backdoor-specific defenses, including Krum (+73.97 points), LaSA (+76.55 points), FLAME (+84.75 points), and CRFL (+74.55 points).

Table 2: Robustness of ShadowFPT against aggregation and defense rules. Results are averaged over all datasets and prompt-tuning settings.

Defense	Best Base	Baseline		ShadowFPT-I				ShadowFPT-II			
		Acc	ASR	Acc	ASR	ΔAcc	ΔASR	Acc	ASR	ΔAcc	ΔASR
FedAvg	A3FL	77.91	60.93	75.56	96.59	-2.35	+35.66	75.67	**98.30**	-2.24	**+37.37**
Krum	PFedBA	73.30	22.66	71.70	95.60	-1.60	+72.94	72.20	**96.63**	-1.09	**+73.97**
TrimmedMean	PFedBA	78.08	32.64	75.50	96.11	-2.58	+63.47	75.55	**98.03**	-2.54	**+65.38**
LaSA	PFedBA	77.16	21.34	75.72	96.16	-1.44	+74.82	75.71	**97.90**	-1.45	**+76.55**
FoolsGold	PFedBA	73.99	39.96	75.99	95.83	+2.01	+55.87	73.05	**96.94**	-0.94	**+56.98**
FLTrust	PFedBA	76.08	33.34	74.57	96.49	-1.50	+63.15	74.86	**98.25**	-1.22	**+64.91**
FLAME	A3FL	75.22	13.17	74.88	93.93	-0.33	+80.77	74.92	**97.92**	-0.30	**+84.75**
CRFL	PFedBA	72.75	22.31	68.91	93.92	-3.84	+71.61	68.95	**96.86**	-3.80	**+74.55**
Bucketing	PFedBA	77.33	60.09	75.84	96.08	-1.49	+35.99	75.49	**98.16**	-1.84	**+38.07**

These results support our claim that update-centered defenses are insufficient for FPT. Although such defenses mitigate abnormal client updates through robust aggregation, filtering, clustering, clipping, or certification, ShadowFPT shifts much of the backdoor effect to the interaction between the adapted trigger and the frozen representation space. ShadowFPT-II improves ASR by 63.61 percentage points on average over Best Base, while incurring only a 1.71-point average clean-accuracy drop. This indicates that its robustness does not primarily come from degrading benign performance.

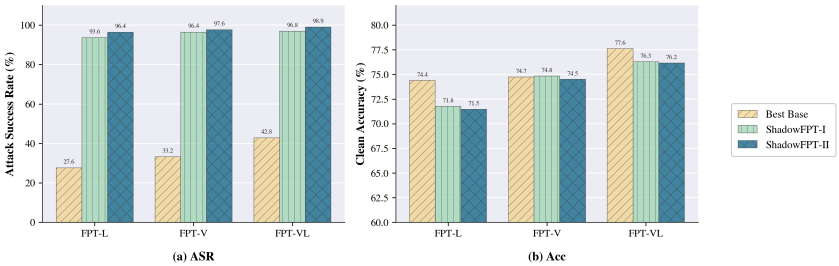
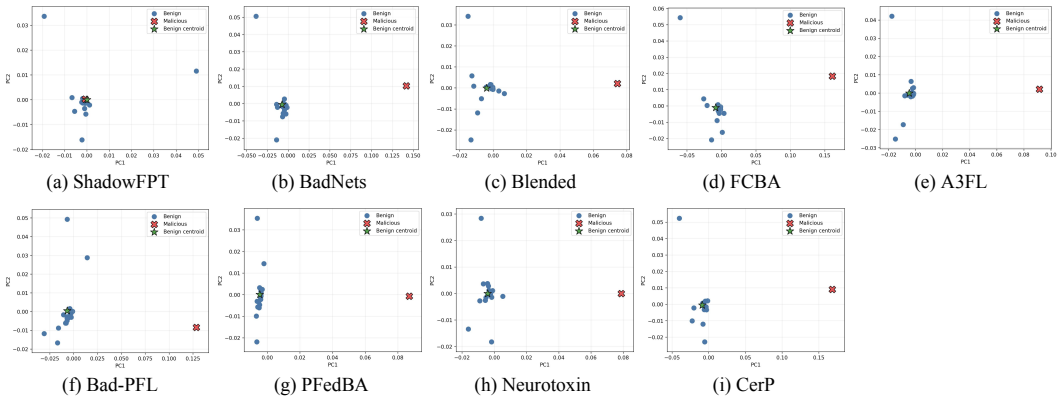


Figure 2: Transferability of ShadowFPT across prompt modalities. Results are averaged over Caltech101, Oxford Flowers, UCF101, and all aggregation or defense rules. Best Base denotes the strongest non-ShadowFPT baseline under each dataset-setting pair, selected by the highest average ASR before averaging across datasets. ShadowFPT achieves consistently high ASR under FPT-L, FPT-V, and FPT-VL, while largely preserving clean accuracy.

Transfer across prompt modalities. We further evaluate whether ShadowFPT transfers across textual, visual, and joint vision–language prompt tuning. Figure 2 reports results averaged over three datasets and all aggregation or defense rules. ShadowFPT-II achieves 96.42%, 97.63%, and 98.95% ASR under FPT-L, FPT-V, and FPT-VL, respectively, substantially outperforming Best Base, which reaches only 27.65%, 33.20%, and 42.83% ASR. ShadowFPT-I shows a similar pattern, with ASR consistently above 93% across all prompt modalities.

Clean accuracy is largely preserved. Compared with Best Base, ShadowFPT-II incurs average clean-accuracy drops of 2.92, 0.24, and 1.45 percentage points under FPT-L, FPT-V, and FPT-VL, respectively. These results show that ShadowFPT is not specific to a single prompt modality. Instead, it exploits the shared frozen CLIP representation space, which remains common across textual, visual, and joint vision–language prompt tuning.

Prompt-space stealthiness. We further examine whether malicious prompt updates can be separated from benign ones. Figure 3 projects client prompts under FedAvg onto the first two principal components. Existing attacks produce malicious prompts that appear as clear outliers from the benign cluster. In contrast, the ShadowFPT prompt lies close to benign clients and nearly overlaps with the benign centroid. This supports our claim that ShadowFPT shifts much of the backdoor effect from uploaded prompt updates to trigger-induced representation steering in the frozen backbone, thereby preserving prompt-space stealthiness.



(a) ShadowFPT (b) BadNets (c) Blended (d) FCBA (e) A3FL

(f) Bad-PFL (g) PFedBA (h) Neurotoxin (i) CerP

Figure 3: PCA visualization of client prompts under FedAvg. Blue points denote benign clients, the red cross denotes the malicious client, and the green star denotes the benign centroid.

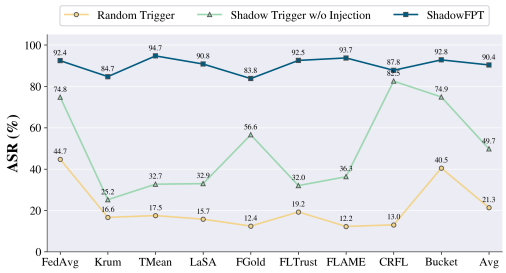


Figure 4: Role of Shadow Trigger Pretraining and Backdoor Injection in FPT on Caltech101. Random Trigger uses a randomly initialized trigger; Shadow Trigger w/o Injection uses the pretrained Shadow Trigger without Backdoor Injection in FPT; ShadowFPT denotes the full method.

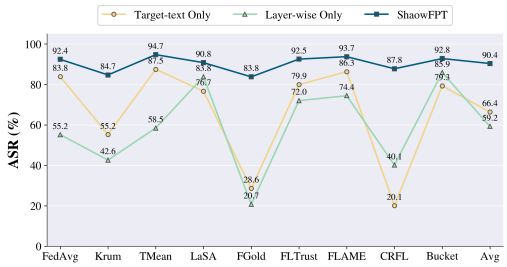


Figure 5: Role of representation-level alignment on Caltech101. ASR is reported under nine aggregation and defense rules. Target-text Only and Layer-wise Only use a single alignment objective, whereas ShadowFPT combines both.

4.3 Ablation Study

Role of Pretraining and Injection. We assess the roles of Shadow Trigger Pretraining and Backdoor Injection in FPT on Caltech101. As shown in Figure 4, Random Trigger achieves only 21.32% average ASR, indicating that an arbitrary trigger is insufficient to induce a strong backdoor. Shadow Trigger without Injection increases the average ASR to 49.75%, suggesting that pretraining already learns target-directed behavior in the frozen CLIP representation space, but remains fragile under defenses such as Krum (25.17%) and FLTrust (31.99%). The full ShadowFPT further raises the average ASR to 90.36%, with large gains under Krum (+59.54) and FLTrust (+60.55). These results show that pretraining provides a transferable initialization, while Backdoor Injection is necessary to convert it into a robust federated backdoor.

Role of Representation-Level Alignment. We further examine the two representation-level objectives used in Shadow Trigger Pretraining. As shown in Figure 5, target-text alignment alone and layer-wise visual alignment alone achieve 66.38% and 59.25% average ASR, respectively, indicating that both provide useful target-directed guidance. However, each objective alone remains fragile under defenses, such as CRFL for Target-text Only (20.08%) and FoolsGold for Layer-wise Only (20.73%). In contrast, the full ShadowFPT objective reaches 90.36% average ASR, improving over the two single-objective variants by 23.98 and 31.11 percentage points. These results show that target-text alignment provides semantic target guidance, while layer-wise visual alignment stabilizes the triggered representation trajectory, together producing a more robust Shadow Trigger.

5 Conclusion

This paper identifies a frozen-backbone backdoor surface in FPT. Unlike conventional federated backdoor attacks that rely primarily on anomalous uploaded updates, ShadowFPT exploits the shared frozen CLIP representation space and shifts much of the attack burden to trigger-induced representation steering. We propose a two-stage attack in which a Shadow Trigger is first pretrained through target-text and layer-wise visual alignment, and then adapted during FPT while only prompt parameters are uploaded to the server. Across multiple datasets, aggregation and defense rules, and textual, visual, and joint vision–language prompt tuning, ShadowFPT achieves consistently high targeted attack success with limited clean accuracy degradation. Prompt-space analyses further show that its malicious updates remain close to benign prompts, explaining why update-centered defenses are insufficient. Our results suggest that securing FPT requires moving beyond prompt anomaly detection and accounting for the interaction between triggers and frozen representation geometry. More broadly, this work highlights a security risk of FPT: freezing the backbone reduces trainable parameters, but does not eliminate backdoor attack surfaces.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Grant No. 62521004). The computations in this research were performed using the CFFF platform of Fudan University.

References

Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, and Julien Stainer. Machine learning with adversaries: Byzantine tolerant gradient descent. *NeurIPS*, 2017.

Xiaoyu Cao, Minghong Fang, Jia Liu, and Neil Zhenqiang Gong. Fltrust: Byzantine-robust federated learning via trust bootstrapping. *arXiv preprint arXiv:2012.13995*, 2020.

Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*, 2017.

Tianyu Cui, Hongxia Li, Jingya Wang, and Ye Shi. Harmonizing generalization and personalization in federated prompt learning. *arXiv preprint arXiv:2405.09771*, 2024.

Mingyuan Fan, Zhanyi Hu, Fuyi Wang, and Cen Chen. Bad-pfl: Exploring backdoor attacks against personalized federated learning. *arXiv preprint arXiv:2501.12736*, 2025.

Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conf. on computer vision and pattern recognition workshop*, pages 178–178. IEEE, 2004.

Clement Fung, Chris JM Yoon, and Ivan Beschastnikh. Mitigating sybils in federated learning poisoning. *arXiv preprint arXiv:1808.04866*, 2018.

Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv preprint arXiv:1708.06733*, 2017.

Rachid Guerraoui, Sébastien Rouault, et al. The hidden vulnerability of distributed learning in byzantium. In *ICML*, 2018.

Tao Guo, Song Guo, and Junxiao Wang. Pfedprompt: Learning personalized prompt for vision-language models in federated learning. In *Proceedings of the ACM Web Conference 2023*, pages 1364–1374, 2023a.

Tao Guo, Song Guo, Junxiao Wang, Xueyang Tang, and Wenchao Xu. Promptfl: Let federated participants cooperatively learn prompts instead of models-federated learning in age of foundation model. *IEEE Transactions on Mobile Computing*, 2023b.

Siquan Huang, Yijiang Li, Chong Chen, Ying Gao, and Xiping Hu. Fedid: Enhancing federated learning security through dynamic identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *ECCV*, 2022.

Sai Praneeth Karimireddy, Lie He, and Martin Jaggi. Byzantine-robust learning on heterogeneous datasets via bucketing. *arXiv preprint arXiv:2006.09365*, 2022.

Momin Ahmad Khan, Yasra Chandio, and Fatima Muhammad Anwar. Sabre-fl: Selective and accurate backdoor rejection for federated prompt learning. *arXiv preprint arXiv:2506.22506*, 2025.

Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *CVPR*, 2023a.

Muhammad Uzair Khattak, Syed Talal Wasim, Muzammal Naseer, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Self-regulating prompts: Foundational model adaptation without forgetting. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15190–15200, 2023b.

Lin Li, Haoyan Guan, Jianing Qiu, and Michael Spratling. One prompt word is enough to boost adversarial robustness for pre-trained vision-language models. In *CVPR*, 2024.

Siyuan Liang, Mingli Zhu, Aishan Liu, Baoyuan Wu, Xiaochun Cao, and Ee-Chien Chang. Badclip: Dual-embedding guided backdoor attack on multimodal contrastive learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24645–24654, 2024.

Tao Liu, Yuhang Zhang, Zhu Feng, Zhiqin Yang, Chen Xu, Dapeng Man, and Wu Yang. Beyond traditional threats: A persistent backdoor attack on federated learning. In *AAAI*, 2024.

Jun Luo, Chen Chen, and Shandong Wu. Mixture of experts made personalized: Federated prompt learning for vision-language models. In *International Conference on Learning Representations*, volume 2025, page 95458, 2025.

Xiaoting Lyu, Yufei Han, Wei Wang, Jingkai Liu, Bin Wang, Jiqiang Liu, and Xiangliang Zhang. Poisoning with cerberus: Stealthy and colluded backdoor attack against federated learning. In *AAAI*, 2023.

Xiaoting Lyu, Yufei Han, Wei Wang, Jingkai Liu, Bin Wang, Kai Chen, Yidong Li, Jiqiang Liu, and Xiangliang Zhang. Coba: Collusive backdoor attacks with optimized trigger to federated learning. *IEEE Transactions on Dependable and Secure Computing*, 22(2):1506–1518, 2024a.

Xiaoting Lyu, Yufei Han, Wei Wang, Jingkai Liu, Yongsheng Zhu, Guangquan Xu, Jiqiang Liu, and Xiangliang Zhang. Lurking in the shadows: Unveiling stealthy backdoor attacks against personalized federated learning. In *33rd USENIX Security Symposium (USENIX Security 24)*, pages 4157–4174. USENIX Association, 2024b.

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

Thien Duc Nguyen, Phillip Rieger, Huili Chen, Hossein Yalame, Helen Möllering, Hossein Fereidooni, Samuel Marchal, Markus Miettinen, Azalia Mirhoseini, Shaza Zeitouni, et al. {FLAME}: Taming backdoors in federated learning. In *USENIX Security*, 2022.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian conf. on computer vision, graphics & image processing*, pages 722–729. IEEE, 2008.

Krishna Pillutla, Sham M Kakade, and Zaid Harchaoui. Robust aggregation for federated learning. *IEEE Transactions on Signal Processing*, 70:1142–1154, 2022.

Chen Qiu, Xingyu Li, Chaithanya Kumar Mummadi, Madan Ravi Ganesh, Zhenzhen Li, Lu Peng, and Wan-Yi Lin. Federated text-driven prompt generation for vision-language models. In *ICLR*, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Mainak Singha, Subhankar Roy, Sarthak Mehrotra, Ankit Jha, Moloud Abdar, Biplab Banerjee, and Elisa Ricci. Fedmvp: Federated multi-modal visual prompt tuning for vision-language models. *arXiv preprint arXiv:2504.20860*, 2025.

K Soomro. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

Chulin Xie, Minghao Chen, Pin-Yu Chen, and Bo Li. Crfl: Certifiably robust federated learning against backdoor attacks. In *ICML*, 2021.

Jiahao Xu, Zikai Zhang, and Rui Hu. Achieving byzantine-resilient federated learning via layer-adaptive sparsified model aggregation. In *WACV*, 2025.

Dong Yin, Yudong Chen, Ramchandran Kannan, and Peter Bartlett. Byzantine-robust distributed learning: Towards optimal statistical rates. In *ICML*, 2018.

Hangfan Zhang, Jinyuan Jia, Jinghui Chen, Lu Lin, and Dinghao Wu. A3fl: Adversarially adaptive backdoor attacks to federated learning. *NeurIPS*, 2023.

Xiyue Zhang, Xiaoyong Xue, Xiaoning Du, Xiaofei Xie, Yang Liu, and Meng Sun. Runtime backdoor detection for federated learning via representational dissimilarity analysis. *IEEE Transactions on Dependable and Secure Computing*, 2025.

Zhengming Zhang, Ashwinee Panda, Linyue Song, Yaoqing Yang, Michael Mahoney, Prateek Mittal, Ramchandran Kannan, and Joseph Gonzalez. Neurotoxin: Durable backdoors in federated learning. In *ICML*, 2022.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *CVPR*, pages 16816–16825, 2022a.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022b.

6 Appendix

Table 3: Acc and ASR (%) on Caltech101, Oxford Flowers, and UCF101 under FPT-L, FPT-V, and FPT-VL. Bold values indicate the best result among all attack methods under each defense. The arrowed values below ShadowFPT-I and ShadowFPT-II report absolute differences from the strongest baseline, where green and red arrows indicate improvement and degradation, respectively.

Dataset	Setting	Method	FedAvg		Krum		TrimmedMean		LaSA		FoolsGold		FLTrust		FLAME		CRFL		Bucketing		Average	
			Acc	ASR	Acc	ASR	Acc	ASR	Acc	ASR	Acc	ASR	Acc	ASR	Acc	ASR	Acc	ASR	Acc	ASR	Acc	ASR
Caltech101	FPT-L	BadNets	91.92	2.09	89.04	0.00	93.46	0.00	92.33	0.00	76.75	0.00	90.58	0.00	89.73	0.00	75.41	0.00	92.86	0.38	88.01	0.27
		Blended	92.49	24.49	**89.13**	0.00	93.71	0.55	93.30	0.00	75.29	0.00	**91.44**	0.00	89.73	0.00	79.83	0.00	**93.79**	71.56	**88.75**	10.73
		FCBA	92.37	1.32	88.68	0.00	**93.50**	0.00	93.10	0.00	78.78	0.00	90.46	0.00	89.73	0.00	75.37	0.00	91.76	0.43	88.19	0.19
		A3FL	93.34	64.58	85.03	0.17	92.45	2.14	91.68	0.08	81.17	0.00	90.38	19.35	89.00	0.34	79.92	0.00	93.46	91.64	88.49	19.81
		Bad-PFL	**94.15**	20.94	88.43	0.08	93.22	0.04	93.42	0.08	76.59	0.00	91.19	0.00	89.16	0.00	**79.87**	0.00	92.65	8.00	88.74	3.24
		PFedBA	93.46	13.10	83.85	0.08	92.61	21.58	90.62	12.46	76.59	0.00	89.77	0.56	90.18	6.55	79.51	1.93	92.94	50.79	87.73	11.89
		Neurotoxin	92.98	10.03	85.43	0.00	92.49	0.00	91.68	0.00	72.45	0.00	91.11	0.00	90.18	0.00	78.45	0.00	91.07	0.00	87.32	1.11
		CerP	87.87	10.96	89.08	0.00	92.77	0.00	**93.95**	0.00	68.92	0.00	90.54	0.00	89.61	0.00	78.10	0.00	81.74	2.74	85.84	1.52
		ShadowFPT-I	91.52	90.96	83.00	**89.29**	92.29	89.20	92.49	87.49	**93.38**	83.23	89.49	89.46	88.96	69.16	72.17	72.22	91.60	87.53	88.32	84.28
			(2.63↓)	(26.38↑)	(6.13↓)	(89.12↑)	(1.42↓)	(67.62↑)	(1.46↓)	(75.03↑)	(12.21↑)	(83.23↑)	(1.95↓)	(70.11↑)	(1.22↓)	(62.61↑)	(7.75↓)	(70.29↑)	(2.19↓)	(4.11↓)	(0.43↓)	(64.47↑)
		ShadowFPT-II	92.90	**92.41**	83.24	84.71	92.41	**94.73**	91.19	**90.79**	77.16	**83.76**	91.72	**92.54**	**91.19**	**93.74**	76.38	**87.76**	92.86	**92.80**	87.67	**90.36**
			(1.25↓)	(27.83↑)	(5.89↓)	(84.54↑)	(1.30↓)	(73.15↑)	(2.76↓)	(78.33↑)	(4.01↓)	(83.76↑)	(0.28↑)	(73.19↑)	(1.01↑)	(87.19↑)	(3.54↓)	(85.83↑)	(0.93↓)	(1.16↑)	(1.08↓)	(70.55↑)
	FPT-V	BadNets	93.71	4.63	89.06	0.00	**93.46**	0.00	92.33	0.00	76.75	0.00	90.58	0.00	89.73	0.00	75.41	0.00	92.86	0.28	88.21	0.55
		Blended	92.77	10.08	92.77	0.30	93.02	0.04	92.57	0.04	91.84	1.97	**93.54**	2.22	92.41	0.10	92.61	0.32	92.25	0.13	92.64	1.69
		FCBA	94.03	5.04	92.34	0.00	92.94	0.00	92.61	0.00	91.48	0.98	92.17	0.00	92.45	0.00	**92.62**	0.00	92.61	0.00	92.58	0.67
		A3FL	94.12	56.67	92.23	3.52	92.94	4.88	92.47	16.40	92.57	0.00	93.38	0.04	92.45	23.08	88.70	0.00	92.46	26.38	92.37	14.55
		Bad-PFL	**94.15**	20.94	88.43	0.08	93.22	0.04	**93.42**	0.08	74.92	0.00	91.19	0.00	89.16	0.00	79.87	0.00	92.65	8.00	88.56	3.24
		PFedBA	93.63	66.98	91.33	8.01	92.89	29.20	92.61	22.14	**94.27**	71.77	93.06	60.38	92.45	0.98	90.71	20.12	**93.59**	60.99	**92.73**	37.84
		Neurotoxin	92.81	4.22	91.28	0.00	91.10	0.00	92.53	0.00	86.81	0.00	92.25	0.00	92.41	0.00	92.45	0.00	92.57	0.00	91.58	0.47
		CerP	92.41	6.50	**93.02**	0.25	92.19	0.46	92.49	0.08	91.80	11.13	92.61	23.94	**92.49**	0.26	92.53	0.25	92.49	0.00	92.45	4.76
		ShadowFPT-I	92.90	93.61	92.61	92.80	92.86	92.63	92.29	93.27	92.53	93.17	92.82	93.23	92.45	93.10	92.13	**92.33**	92.98	92.24	92.62	92.93
			(1.25↓)	(26.63↑)	(0.41↓)	(84.79↑)	(0.60↓)	(63.43↑)	(1.13↓)	(71.13↑)	(1.74↓)	(21.40↑)	(0.72↓)	(32.85↑)	(0.04↓)	(70.02↑)	(0.49↓)	(72.21↑)	(0.61↓)	(31.25↑)	(0.11↓)	(55.09↑)
		ShadowFPT-II	92.57	**94.64**	92.62	**94.69**	92.74	**94.77**	92.08	**95.37**	92.94	**94.34**	92.45	**94.39**	92.37	**94.52**	84.41	91.33	92.94	**94.17**	91.68	**94.25**
			(1.58↓)	(27.66↑)	(0.40↓)	(86.68↑)	(0.72↓)	(65.57↑)	(1.34↓)	(73.23↑)	(1.33↓)	(22.57↑)	(1.09↓)	(34.01↑)	(0.12↓)	(71.44↑)	(8.21↓)	(71.21↑)	(0.65↓)	(33.18↑)	(1.05↓)	(56.41↑)
	FPT-VL	BadNets	93.02	21.02	91.60	0.00	93.26	0.08	93.43	0.13	93.54	85.61	93.31	0.04	92.49	0.04	91.76	0.00	93.43	15.76	92.87	13.63
		Blended	93.43	77.30	91.60	0.00	93.22	1.58	93.22	48.09	93.50	77.96	**93.38**	0.81	92.49	0.00	**91.88**	0.00	93.42	80.59	**92.90**	31.81
		FCBA	91.77	4.41	91.56	0.00	**93.38**	0.04	93.43	0.17	91.19	3.43	93.30	0.04	92.49	0.04	91.76	0.00	**93.59**	1.92	92.50	1.12
		A3FL	93.36	96.01	**91.80**	0.85	93.14	0.29	93.18	93.22	92.77	80.08	93.24	37.61	92.37	45.61	91.80	0.00	93.06	95.41	92.75	49.90
		Bad-PFL	93.38	70.32	91.60	33.36	93.14	62.01	**93.51**	70.32	93.02	82.35	93.31	36.74	92.61	16.15	91.80	2.35	92.94	81.75	92.83	50.59
		PFedBA	93.38	70.32	91.78	33.36	93.14	62.01	93.42	77.47	93.02	82.35	93.30	36.74	92.61	16.14	91.80	2.35	92.94	81.75	92.82	51.39
		Neurotoxin	93.14	1.19	91.72	0.00	93.26	0.08	93.38	0.08	90.99	2.69	93.22	0.08	**92.65**	0.04	91.76	0.00	93.18	0.04	92.59	0.47
		CerP	80.00	30.70	91.76	0.00	93.26	0.08	85.27	23.85	73.99	27.64	93.34	0.08	92.61	0.02	91.72	0.00	65.07	42.86	85.23	13.91
		ShadowFPT-I	**93.73**	93.06	90.79	92.51	92.21	92.50	92.49	93.36	**93.71**	96.53	93.18	94.56	91.52	91.86	91.15	91.43	92.77	92.93	92.39	93.19
			(0.30↑)	(2.95↓)	(1.01↓)	(59.15↑)	(1.17↓)	(30.49↑)	(1.02↓)	(0.14↑)	(0.17↑)	(10.92↑)	(0.20↓)	(56.95↑)	(1.13↓)	(46.25↑)	(0.73↓)	(89.08↑)	(0.82↓)	(2.48↓)	(0.51↓)	(41.80↑)
		ShadowFPT-II	92.73	**97.73**	90.79	**97.34**	92.21	**98.02**	92.53	**98.24**	93.67	**98.75**	93.18	**98.41**	91.52	**98.02**	91.15	**97.55**	92.77	**98.24**	92.28	**98.03**
			(0.70↓)	(1.72↑)	(1.01↓)	(63.98↑)	(1.17↓)	(36.01↑)	(0.98↓)	(5.02↑)	(0.13↑)	(13.14↑)	(0.20↓)	(60.80↑)	(1.13↓)	(52.41↑)	(0.73↓)	(95.20↑)	(0.82↓)	(2.83↑)	(0.62↓)	(46.64↑)
Oxford Flowers	FPT-L	BadNets	67.88	6.89	64.39	0.28	68.69	1.22	67.47	3.05	68.21	14.15	64.31	0.24	67.47	0.37	44.25	0.41	65.44	5.54	64.23	3.57
		Blended	69.99	27.91	64.39	0.32	69.14	5.75	67.07	25.58	69.26	25.18	65.97	6.28	66.74	0.04	44.33	0.53	64.72	30.39	64.62	13.55
		FCBA	68.37	6.85	64.39	0.36	68.33	1.14	69.87	2.03	65.93	6.85	66.82	0.69	66.05	0.53	43.11	0.40	66.13	3.42	64.33	2.47
		A3FL	70.76	97.87	65.00	0.00	67.07	93.96	70.11	17.78	71.98	29.49	68.16	81.19	64.63	18.48	51.44	0.00	**70.52**	3.83	66.63	38.07
		Bad-PFL	69.71	7.71	62.72	0.40	68.29	4.93	69.06	6.52	**76.20**	10.15	68.89	4.03	66.34	2.40	49.24	0.97	69.63	10.24	66.68	5.26
		PFedBA	70.92	33.51	**69.38**	24.12	**72.71**	26.67	71.33	4.20	68.41	32.46	67.52	30.79	**69.67**	14.64	**68.77**	35.07	70.28	29.67	**69.89**	25.68
		Neurotoxin	**71.78**	3.34	66.66	0.34	68.53	0.85	68.73	1.71	66.86	2.12	**69.63**	0.85	67.07	0.57	48.55	0.16	67.68	1.06	66.17	1.22
		CerP	69.26	5.26	63.29	0.16	68.65	0.85	**71.73**	1.06	65.81	7.62	66.82	1.67	67.35	0.57	51.92	0.24	63.82	13.83	65.41	3.47
		ShadowFPT-I	62.03	96.89	54.42	94.90	62.07	97.12	65.21	96.73	62.89	98.0	63.05	97.22	63.94	97.31	49.20	96.16	65.52	96.36	60.93	96.74
			(9.75↓)	(0.98↓)	(14.96↓)	(70.78↑)	(10.64↓)	(3.16↑)	(6.52↓)	(71.15↑)	(13.31↓)	(65.54↑)	(6.58↓)	(16.03↑)	(5.73↓)	(78.83↑)	(19.57↓)	(61.09↑)	(5.00↓)	(65.97↑)	(8.96↓)	(58.67↑)
		ShadowFPT-II	63.12	**100.0**	58.66	**98.33**	61.87	**98.74**	65.18	**98.54**	58.66	**99.10**	63.13	**99.81**	63.42	**98.22**	53.33	**97.91**	60.29	**99.76**	60.85	**98.93**
			(8.66↓)	(2.13↑)	(10.72↓)	(74.21↑)	(10.84↓)	(4.78↑)	(6.55↓)	(72.96↑)	(17.54↓)	(66.64↑)	(6.50↓)	(18.62↑)	(6.25↓)	(79.74↑)	(15.44↓)	(62.84↑)	(10.23↓)	(69.37↑)	(9.04↓)	(60.86↑)
	FPT-V	BadNets	66.13	4.21	66.31	0.04	67.11	0.45	66.46	0.03	44.05	2.32	**69.87**	0.89	66.71	0.04	**69.28**	1.21	64.88	0.97	64.53	1.13
		Blended	66.34	26.39	66.22	0.04	68.49	3.46	66.71	0.12	48.55	0.21	68.45	2.21	66.71	0.21	68.32	0.33	64.67	11.46	64.94	4.94
		FCBA	67.19	3.22	66.34	0.20	66.99	0.93	**66.74**	0.16	43.44	1.83	65.21	2.36	66.62	0.12	66.21	2.33	64.88	1.87	63.74	1.45
		A3FL	66.91	20.27	66.82	0.89	68.29	0.00	66.73	0.00	58.79	16.40	69.06	0.00	66.71	10.85	63.09	2.65	64.92	16.55	65.70	7.51
		Bad-PFL	66.54	19.09	66.91	4.85	**69.18**	12.23	66.62	4.36	48.31	5.58	66.26	2.57	66.67	4.41	64.59	0.45	66.58	12.89	64.63	7.38
		PFedBA	**67.61**	47.98	66.74	29.95	67.92	33.31	66.70	25.49	58.67	34.96	67.11	38.32	66.94	9.18	64.27	7.46	65.57	45.32	65.73	30.22
		Neurotoxin	66.95	0.97	66.62	0.04	68.45	0.36	66.71	0.04	62.93	0.57	67.80	0.24	66.71	0.04	66.45	0.42	64.55	0.69	66.35	0.37
		CerP	67.43	21.21	66.22	1.12	68.04	5.10	66.71	7.29	62.22	2.61	67.92	4.27	66.67	15.71	65.91	4.22	66.32	6.97	66.38	7.61
		ShadowFPT-I	67.23	98.28	66.95	98.0	66.51	98.08	66.46	98.08	**67.27**	98.73	63.58	99.97	**66.95**	98.20	67.03	97.59	67.39	98.41	66.60	98.37
			(0.38↓)	(50.30↑)	(0.04↑)	(68.05↑)	(2.67↓)	(64.77↑)	(0.28↓)	(72.59↑)	(4.34↑)	(63.77↑)	(6.29↓)	(61.65↑)	(0.01↑)	(82.49↑)	(2.25↓)	(90.13↑)	(0.81↑)	(53.09↑)	(0.22↑)	(68.15↑)
		ShadowFPT-II	66.86	**99.96**	**67.15**	**98.21**	67.13	**98.12**	66.67	**99.24**	66.42	**99.44**	63.94	**100.0**	66.89	**98.46**	66.99	**98.12**	**67.43**	**98.91**	**66.61**	**98.94**
			(0.75↓)	(51.98↑)	(0.24↑)	(68.26↑)	(2.05↓)	(64.81↑)	(0.07↓)	(73.75↑)	(3.49↑)	(64.48↑)	(5.93↓)	(61.68↑)	(0.05↓)	(82.75↑)	(2.29↓)	(90.66↑)	(0.85↑)	(53.59↑)	(0.23↑)	(68.72↑)
	FPT-VL	BadNets	70.44	4.08	69.02	0.12	71.86	0.44	71.33	0.97	**72.95**	3.53	68.69	0.73	70.41	0.21	68.77	0.41	67.76	1.91	70.14	1.38
		Blended	**72.21**	26.81	68.98	1.50	71.78	5.14	71.57	18.64	70.97	20.43	67.76	25.88	**71.41**	0.93	68.49	2.20	68.37	17.74	70.17	13.25
		FCBA	69.83	3.79	69.02	0.28	72.02	0.61	71.78	1.01	67.60	3.63	69.02	0.48	70.23	0.32	68.73	0.57	67.43	1.71	69.52	1.38
		A3FL	70.80	88.98	69.32	12.97	**72.91**	3.31	70.88	6.97	68.08	73.27	**69.79**	30.64	68.89	19.46	68.49	22.19	71.74	35.78	70.10	32.62
		Bad-PFL	71.33	8.48	68.62	4.81	71.98	3.83	**72.06**	3.02	72.18	12.52	68.77	6.40	70.07	2.36	68.74	3.75	**72.75**	6.69	**70.72**	5.76
		PFedBA	70.92	73.51	69.38	64.12	72.71	56.67	71.33	4.20	68.41	62.46	67.51	70.97	69.67	14.64	68.77	75.07	70.28	69.76	69.89	54.60
		Neurotoxin	69.26	0.82	**70.56**	0.08	72.22	0.36	70.44	0.08	69.95	3.46	68.65	0.61	69.18	0.62	68.89	0.28	68.81	0.65	69.77	0.77
		CerP	71.53	5.58	69.22	0.77	71.53	1.55	71.57	1.26	69.51	4.20	69.14	2.12	70.40	0.53	**68.94**	0.82	67.27	5.34	69.90	2.46
		ShadowFPT-I	68.77	98.26	66.26	98.69	68.94	98.81	69.61	98.85	70.03	94.36	69.31	97.38	68.57	**98.85**	66.01	99.10	69.71	98.36	68.52	98.07
			(3.44↓)	(9.28↑)	(4.30↓)	(34.57↑)	(3.97↓)	(42.14↑)	(2.95↓)	(80.21↑)	(2.92↓)	(21.09↑)	(0.48↓)	(26.41↑)	(2.84↓)	(79.39↑)	(2.93↓)	(24.03↑)	(3.04↓)	(28.60↑)	(2.20↓)	(43.47↑)
		ShadowFPT-II	69.06	**100.0**	66.31	**99.44**	68.82	**98.92**	70.07	**99.31**	67.72	**97.54**	69.18	**99.68**	68.37	98.47	65.89	**99.29**	69.58	**99.93**	68.33	**99.18**
			(3.15↓)	(11.02↑)	(4.25↓)	(35.32↑)	(4.09↓)	(42.25↑)	(1.99↓)	(80.67↑)	(5.23↓)	(24.27↑)	(0.61↓)	(28.71↑)	(3.04↓)	(79.01↑)	(3.05↓)	(24.22↑)	(3.17↓)	(30.17↑)	(2.39↓)	(44.58↑)
UCF101	FPT-L	BadNets	69.62	0.75	64.82	0.02	71.13	0.03	69.94	0.00	69.92	0.51	70.71	0.16	70.28	0.03	**62.01**	0.29	69.60	0.24	68.67	0.23
		Blended	70.23	52.42	**64.84**	0.05	**72.19**	0.64	70.36	0.26	**70.55**	35.07	70.10	10.56	70.23	0.14	61.51	0.43	69.83	33.41	**68.87**	14.78
		FCBA	70.05	0.90	64.83	0.03	71.68	0.13	70.49	0.08	69.60	0.13	69.09	0.41	**70.34**	0.02	60.61	0.37	70.44	0.18	68.57	0.25
		A3FL	71.13	49.87	60.03	2.51	70.36	3.63	70.26	1.42	69.78	9.36	**71.74**	29.17	68.86	0.51	60.66	0.77	**72.11**	14.06	68.33	12.37
		Bad-PFL	70.76	0.72	54.24	0.51	71.66	0.98	**71.23**	0.55	70.33	0.67	71.45	0.51	68.51	0.41	61.69	0.51	69.71	1.04	67.73	0.66
		PFedBA	71.47	31.55	58.57	2.19	71.05	15.03	71.02	26.77	70.39	19.06	71.50	5.69	67.77	29.23	59.43	11.55	70.97	84.51	68.02	25.06
		Neurotoxin	**72.11**	0.27	59.55	0.03	70.92	0.24	70.76	0.16	68.59	0.18	71.18	0.11	**70.34**	0.05	59.43	0.43	71.29	0.05	68.24	0.17
		CerP	68.14	4.33	59.92	0.13	71.50	0.32	70.79	0.11	57.04	48.22	69.46	0.56	69.60	0.08	61.35	0.37	68.27	1.25	66.23	6.15
		ShadowFPT-I	69.07	99.86	61.96	99.58	69.89	99.78	69.49	**100.0**	68.54	**100.0**	66.98	**100.0**	67.83	99.23	53.39	99.81	67.27	**100.0**	66.05	**99.81**
			(3.04↓)	(47.44↑)	(2.88↓)	(97.07↑)	(2.30↓)	(84.75↑)	(1.74↓)	(73.23↑)	(2.01↓)	(51.78↑)	(4.76↓)	(70.83↑)	(2.51↓)	(70.00↑)	(8.62↓)	(88.26↑)	(4.84↓)	(15.49↑)	(2.82↓)	(74.75↑)
		ShadowFPT-II	68.78	**99.97**	61.96	**99.97**	70.10	**99.97**	69.69	99.97	66.05	**100.0**	67.45	99.97	66.82	**99.97**	53.37	**99.98**	68.59	**100.0**	65.87	99.98
			(3.33↓)	(47.55↑)	(2.88↓)	(97.46↑)	(2.09↓)	(84.94↑)	(1.54↓)	(73.20↑)	(4.50↓)	(51.78↑)	(4.29↓)	(70.80↑)	(3.52↓)	(70.74↑)	(8.64↓)	(88.43↑)	(3.52↓)	(15.49↑)	(3.00↓)	(74.92↑)
	FPT-V	BadNets	67.96	1.97	64.68	0.27	67.59	0.05	65.45	0.32	65.92	1.38	65.68	0.51	65.45	0.32	64.63	0.16	67.59	2.64	66.11	0.85
		Blended	68.09	36.26	64.48	0.49	67.77	5.61	65.45	0.61	**66.41**	36.10	**68.67**	21.18	65.45	0.62	64.81	0.11	64.78	12.67	66.21	12.63
		FCBA	67.93	0.62	64.68	0.26	67.72	0.13	65.39	0.27	64.84	0.45	67.48	0.24	65.45	0.26	64.76	0.13	67.56	0.45	66.20	0.31
		A3FL	**69.69**	40.24	**65.55**	38.40	68.09	0.00	65.45	0.08	66.03	0.00	66.05	0.00	65.45	0.11	64.97	7.14	67.12	3.74	**66.49**	9.97
		Bad-PFL	68.12	1.25	65.21	0.26	68.12	0.29	**65.47**	0.29	64.71	0.53	66.82	1.12	65.23	0.32	65.00	0.13	**67.77**	0.29	66.27	0.50
		PFedBA	67.31	39.26	64.97	28.72	67.41	20.48	65.45	0.35	64.36	28.68	64.60	32.44	**65.47**	0.72	65.02	39.95	67.24	93.39	65.76	31.55
		Neurotoxin	67.77	0.05	64.68	0.27	**68.31**	0.05	65.37	0.32	66.11	0.18	65.92	0.11	65.39	0.32	**65.08**	0.22	67.14	0.13	66.20	0.18
		CerP	67.75	0.93	65.11	0.27	67.85	0.43	65.43	0.53	60.95	0.99	66.51	61.05	65.45	1.79	64.76	1.23	65.95	8.05	65.53	8.36
		ShadowFPT-I	65.21	98.73	65.53	97.56	65.29	97.28	64.89	98.02	65.55	98.69	65.53	97.62	65.32	98.35	64.78	97.01	65.46	99.08	65.28	98.04
			(4.48↓)	(58.49↑)	(0.02↓)	(59.16↑)	(3.02↓)	(76.80↑)	(0.58↓)	(97.41↑)	(0.86↓)	(62.59↑)	(3.14↓)	(36.57↑)	(0.15↓)	(96.56↑)	(0.30↓)	(57.06↑)	(2.31↓)	(5.69↑)	(1.21↓)	(66.49↑)
		ShadowFPT-II	65.37	**100.0**	65.53	**98.41**	65.22	**99.97**	65.18	**99.89**	65.01	**99.57**	65.34	**99.94**	65.34	**99.89**	64.68	**99.83**	65.18	**99.67**	65.21	**99.69**
			(4.32↓)	(59.76↑)	(0.02↓)	(60.01↑)	(3.09↓)	(79.49↑)	(0.29↓)	(99.28↑)	(1.40↓)	(63.47↑)	(3.33↓)	(38.89↑)	(0.13↓)	(98.10↑)	(0.40↓)	(59.88↑)	(2.59↓)	(6.28↑)	(1.28↓)	(68.14↑)
	FPT-VL	BadNets	71.62	21.44	66.95	0.18	71.53	0.34	**72.14**	0.27	71.18	20.08	70.76	0.13	70.52	0.11	66.74	41.56	70.73	0.32	**70.24**	9.38
		Blended	71.74	51.04	67.08	0.37	71.34	11.20	71.62	51.27	69.54	3.34	**70.79**	0.64	70.52	0.37	66.72	1.79	71.87	82.42	70.14	22.49
		FCBA	71.61	0.91	66.98	0.13	71.66	0.34	71.13	0.32	69.99	0.37	**70.79**	0.16	70.61	0.13	66.93	0.45	70.44	0.72	70.02	0.39
		A3FL	71.08	33.88	63.07	0.00	71.47	4.68	71.29	5.24	70.97	0.00	70.02	0.13	68.59	0.08	66.77	31.48	70.68	4.22	69.33	8.86
		Bad-PFL	71.02	2.03	**67.59**	0.26	71.45	1.12	70.89	1.06	70.68	1.57	70.36	0.77	70.23	0.61	66.67	0.81	71.53	2.01	70.05	1.14
		PFedBA	**72.79**	29.84	63.67	13.37	**72.32**	28.83	71.98	18.99	**71.77**	27.88	70.32	24.14	70.28	4.89	66.50	7.32	**72.13**	24.64	70.20	19.99
		Neurotoxin	70.76	0.37	66.58	0.24	71.00	0.29	70.63	0.37	69.46	4.33	70.57	0.16	69.83	0.18	**67.08**	0.37	70.63	0.29	69.62	0.73
		CerP	65.63	9.52	64.92	0.13	71.31	0.41	68.96	1.97	66.69	8.05	70.65	0.22	**70.92**	0.16	65.66	0.48	68.83	2.64	68.17	2.62
		ShadowFPT-I	69.54	99.70	63.78	97.03	69.44	**99.62**	69.07	99.62	70.05	99.78	67.22	98.98	68.42	99.35	64.36	99.67	69.83	99.78	67.97	99.28
			(3.25↓)	(48.66↑)	(3.81↓)	(83.66↑)	(2.88↓)	(70.79↑)	(3.07↓)	(48.35↑)	(1.72↓)	(71.90↑)	(3.57↓)	(74.84↑)	(2.50↓)	(94.46↑)	(2.72↓)	(58.11↑)	(2.30↓)	(17.36↑)	(2.27↓)	(76.79↑)
		ShadowFPT-II	69.68	**100.0**	63.57	**98.56**	69.42	98.99	68.78	**99.71**	69.81	**99.93**	67.31	**99.52**	68.35	**100.0**	64.34	**99.96**	69.73	**100**	67.89	**99.63**
			(3.11↓)	(48.96↑)	(4.02↓)	(85.19↑)	(2.90↓)	(70.16↑)	(3.36↓)	(48.44↑)	(1.96↓)	(72.05↑)	(3.48↓)	(75.38↑)	(2.57↓)	(95.11↑)	(2.74↓)	(58.40↑)	(2.40↓)	(17.58↑)	(2.35↓)	(77.14↑)

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes].

   Justification: The abstract and introduction clearly state the paper's scope, threat model, main methodological contribution, and empirical claims. The claims are supported by the experiments in Section 4, including evaluations across datasets, aggregation and defense rules, non-IID federated settings, and textual, visual, and joint vision–language prompt tuning.

2. **Limitations**

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

   Answer: [No].

   Justification: The current version does not include a dedicated limitations discussion. Important limitations include the white-box threat model, the focus on CLIP-style frozen backbones, the use of one malicious client by default, and evaluation on three public image-recognition datasets.

3. **Theory assumptions and proofs**

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

   Answer: [N/A].

   Justification: The paper does not present formal theoretical results, theorems, or proofs. The equations in Section 3 define the problem setup and optimization objectives rather than stating theoretical guarantees.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper?

   Answer: [Yes].

   Justification: The paper specifies the main information needed to reproduce the experiments, including the CLIP backbone, datasets, federated setting, non-IID construction, prompt-tuning modalities, baselines, defenses, training details, trigger configuration, and evaluation metrics. Additional results and implementation details are provided in the appendix.

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

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

   Answer: [No].

   Justification: The experiments use publicly available datasets and standard pretrained backbones, but the current submission does not provide an open-source code package or reproduction scripts. We plan to release anonymized code and detailed reproduction instructions when permitted by the submission policy.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details necessary to understand the results?

   Answer: [Yes].

   Justification: The experimental setup specifies the backbone, datasets, non-IID client construction, number of clients, malicious-client setting, communication rounds, local epochs, batch size, learning rates, trigger configuration, attack baselines, defense baselines, and metrics. These details are provided in Section 4 and the appendix.

7. **Experiment statistical significance**

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

   Answer: [No].

   Justification: The paper reports results averaged over datasets, prompt-tuning settings, and aggregation or defense rules, but it does not report error bars, confidence intervals, or statistical significance tests over repeated random seeds. This is a limitation of the current empirical reporting.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources needed to reproduce the experiments?

   Answer: [No].

   Justification: The current paper does not provide detailed compute-resource information such as GPU type, memory, runtime per experiment, or total compute budget. These details should be added in the appendix for full reproducibility.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics?

   Answer: [Yes].

   Justification: The research is conducted as a controlled security study using public datasets and standard experimental settings. The paper analyzes a dual-use backdoor threat to improve understanding of vulnerabilities in federated prompt tuning and does not involve unauthorized system access, private user data, or human-subject data collection.

10. **Broader impacts**

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

    Answer: [Yes].

    Justification: The work has both positive and negative societal implications. It exposes a realistic security risk in federated prompt tuning, which can motivate stronger defenses, but the proposed attack is dual-use and could be misused if applied outside controlled research settings.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse?

    Answer: [N/A].

    Justification: The paper does not release a high-risk pretrained model, image generator, scraped dataset, or deployable system. The study uses public datasets and evaluates attacks in an offline experimental setting.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets used in the paper properly credited, and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes].

    Justification: The paper cites the original sources for the existing datasets, pretrained CLIP backbone, prompt-tuning methods, attack baselines, and defense baselines used in the experiments. The experiments are based on publicly available research assets.

13. **New assets**

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

    Answer: [N/A].

Justification: The paper does not introduce a new dataset, benchmark, pretrained model, or other standalone research asset. The contribution is a method and empirical security analysis.

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

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

    Answer: [N/A].

    Justification: The paper does not involve crowdsourcing, user studies, human-subject experiments, or participant compensation. All experiments are conducted on existing public image-recognition datasets.

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

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board approvals or an equivalent approval/review were obtained?

    Answer: [N/A].

    Justification: The paper does not involve human subjects, crowdsourcing, private user data, or participant interaction. Therefore, IRB approval or equivalent human-subjects review is not applicable.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research?

    Answer: [N/A].

    Justification: The core method, experiments, and scientific claims do not rely on LLMs as an important, original, or non-standard component. Any use of LLMs for writing, editing, or formatting does not affect the methodology, experimental design, or scientific conclusions.