FEDHARDEN: PROACTIVE DEFENSE AGAINST PERSISTENT BACKDOOR ATTACKS IN FEDERATED LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Federated learning (FL) enables decentralized clients to jointly train a global model by uploading locally trained updates to a central server. However, compromised clients may inject backdoors that cause trigger-specific misclassification while preserving main-task performance. Existing defenses mainly act on client updates after local training, but may leave residual backdoor influence under sustained poisoning. This influence can accumulate across rounds when benign training does not sufficiently erase it. Motivated by the low-activity parameter subspace exploited by persistent attacks, we propose FedHarden, a detection-free FL defense that proactively robustifies the global model before each communication round. FedHarden uses OOD-derived auxiliary data to introduce competing gradients and limits server-side training to selected convolutional kernels and associated batch-normalization parameters to suppress backdoor accumulation. Experiments across diverse attacks, datasets, and model architectures show that FedHarden suppresses backdoors under sustained poisoning while maintaining comparable main-task accuracy. Specifically, in a CIFAR10 experiment with 600 poisoning rounds, it reduces backdoor accuracy to near the random-guess level (10.00%), compared with 42.44% for the strongest baseline, a $4.2\times$ reduction.

1 INTRODUCTION

Federated learning (FL) enables multiple clients to jointly train a global model under server coordination without sharing raw data (McMahan et al., 2017). However, malicious clients can submit poisoned updates that induce attacker-chosen predictions on triggered inputs while preserving main-task performance (Bhagoji et al., 2019; Bagdasaryan et al., 2020). Existing FL backdoor defenses mainly rely on influence reduction or detect-then-filter mechanisms. Robust aggregation and norm clipping limit the influence of malicious updates (Sun et al., 2019; Cao et al., 2021), while detection-based methods identify suspicious updates through parameter-space statistics or prediction behavior (Rieger et al., 2022; Fung et al., 2020; Nguyen et al., 2022; Kumari et al., 2023; Fereidooni et al., 2023; Krauß & Dmitrienko, 2023). BackdoorIndicator further uses a server-inserted OOD indicator task to detect backdoored client models (Li & Dai, 2024). These approaches can suppress backdoors in some settings, but their effectiveness over extended periods of poisoning remains a challenge.

In particular, compromised clients that remain eligible for selection can repeatedly participate in training, and control over multiple clients can provide further poisoning opportunities (Bagdasaryan et al., 2020; Fung et al., 2020). We study *persistent backdoor attacks* in this setting, where the adversary repeatedly injects poisoned updates over an extended training period. Limiting malicious influence within individual rounds does not necessarily ensure robustness across rounds: updates that evade detection or retain residual influence after aggregation may accumulate if benign training does not sufficiently erase their effects (Figure 1(a)). Our long-horizon evaluation shows substantial backdoor accuracy under sustained poisoning for several defenses (Appendix B.1, Figure A1).

A key factor underlying this accumulation is the uneven influence of benign training across model parameters. Prior work shows that targeting parameters with small benign updates can make injected backdoors more durable (Zhang et al., 2022). Such *low-activity parameters* provide a low-interference space in which malicious updates can persist and accumulate. This observation mo-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

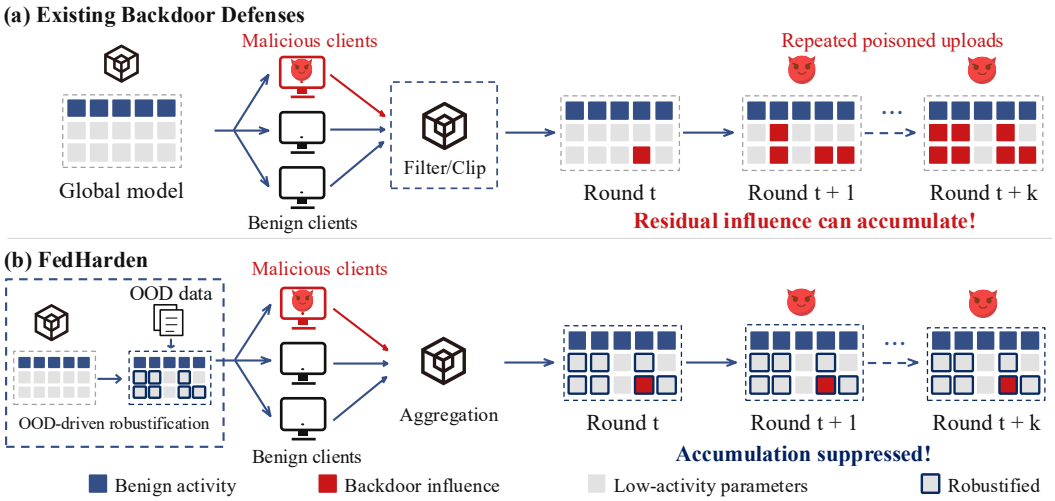


Figure 1: (a) Existing defenses may leave residual malicious influence after filtering or influence reduction, allowing accumulation across rounds. (b) FedHarden robustifies the global model before client training to suppress accumulation.

tivates a proactive defense question: can the server reshape this parameter subspace before client training to reduce the capacity available for backdoor accumulation?

We propose *FedHarden*, a detection-free backdoor defense that proactively robustifies the global model before each communication round (Figure 1(b)). FedHarden treats low-activity model capacity as a server-controllable defense surface. Using public out-of-distribution (OOD) data, the server constructs a *flood dataset* that supplies diverse auxiliary gradients and a *shadow dataset* that uses model-predicted labels to approximate main-task signals. The server trains the global model on their mixture while restricting updates to key convolutional kernels and associated batch-normalization parameters. FedHarden couples auxiliary gradients from flood data with the model-consistent signals of shadow data and kernel-wise update constraints to suppress accumulation across rounds while preserving main-task performance. The defense requires neither clients' private data nor malicious-update detection.

We evaluate FedHarden across diverse attacks, datasets, model architectures, and FL settings. The results demonstrate sustained backdoor suppression, including a $4.2\times$ reduction in backdoor accuracy relative to the strongest baseline under 600 rounds of Neurotoxin blended poisoning. Further analyses examine the effects of OOD data sources, kernel-wise selection, flood dataset size, and the selected-parameter ratio.

Our main contributions are:

- We reinterpret the low-activity parameter subspace as a server-controllable defense surface and support this perspective with empirical observations and a sensitivity bound.
- We design complementary flood and shadow datasets from public OOD data to increase parameter activity while preserving main-task representations without accessing clients' private data.
- We propose FedHarden, a detection-free FL defense that combines kernel-wise parameter selection with server-side robustification before each communication round to suppress backdoor accumulation while retaining the standard client training and communication protocol.
- We demonstrate sustained backdoor suppression with comparable main-task accuracy across diverse attacks, datasets, architectures, and FL settings, including adaptive attack scenarios.

2 Preliminaries and Related Work

Federated Learning. At communication round t, the server broadcasts the global model θ^t to a selected client set S_t. Each client $i \in S_t$ optimizes the model on its private dataset D_i and returns θ_i^t.

With uniform client weights, the FedAvg baseline aggregates these models as $\theta^{t+1} = \frac{1}{|S_t|} \sum_{i \in S_t} \theta_i^t$ (McMahan et al., 2017). Clients exchange model updates while retaining their raw data locally.

Backdoor Attacks in FL. Representative attacks include model replacement (Bagdasaryan et al., 2020), norm-constrained poisoning (Sun et al., 2019), and durability-oriented methods such as Neurotoxin and Chameleon (Zhang et al., 2022; Dai & Li, 2023). Recent attacks further improve stealth or target personalized FL (Lyu et al., 2023; 2024). Backdoors can use synthetic triggers, such as pixel patterns and blending, or semantic and edge-case inputs (Gu et al., 2019; Chen et al., 2017; Bagdasaryan et al., 2020; Wang et al., 2020).

Backdoor Defenses. Existing defenses reduce malicious influence through robust aggregation or clipping (Sun et al., 2019; Cao et al., 2021), or detect and filter suspicious updates using parameter statistics and behavioral signals (Rieger et al., 2022; Fung et al., 2020; Nguyen et al., 2022). BackdoorIndicator evaluates an OOD indicator task on uploaded client models to detect backdoors (Li & Dai, 2024). Standard adversarial training (AT) trains models against worst-case input perturbations (Madry et al., 2018), with its effectiveness against backdoors depending on the perturbation model and attack setting (Weng et al., 2020; Gao et al., 2024). In FL, FLIP hardens benign clients through trigger reverse engineering and adversarial training (Zhang et al., 2023), while SPMC combines inter-client contribution estimation with local gradient purification (He et al., 2025). FedHarden instead uses server-side OOD auxiliary tasks to reshape low-activity parameter capacity against repeated model poisoning, without trigger reconstruction or input-space adversarial optimization, while retaining the standard client training protocol.

3 Threat Model and Motivation

In this section, we define the threat model and motivate FedHarden through two research questions (RQs): **(i)** whether the low-activity parameter subspace exploited by prior attacks can serve as a defense surface, and **(ii)** how the server can reduce its exploitable capacity without accessing clients' private main-task data or compromising model utility.

3.1 Threat Model

Adversary's Goal and Capability. The adversary aims to plant a backdoor into the FL global model by compromising participating clients. A successful backdoor causes trigger-bearing inputs to be classified into an adversary-chosen target label while preserving main-task accuracy. Once a client is compromised, the adversary controls its local data, training process, and uploaded update. Following the benign-majority assumption (Bagdasaryan et al., 2020; Li & Dai, 2024; Kumari et al., 2023), the adversary may compromise fewer than half of the participating clients in each round. We focus on persistent poisoning, where compromised clients repeatedly submit poisoned updates over an extended training period.

Defender's Goal and Capability. The server aims to suppress backdoor accumulation during ongoing poisoning while preserving main-task accuracy, without accessing clients' raw data. We assume it can collect public OOD samples whose labels need not overlap with the main task, such as ImageNet (Deng et al., 2009). These samples provide diverse auxiliary inputs without violating privacy.

3.2 RQ1: Can Low-Activity Model Capacity Serve as a Defense Surface?

Prior attacks have shown that malicious updates become more persistent when confined to parameters that receive limited benign updates (Zhang et al., 2022; Choe et al., 2025). Benign training does not modify all parameters equally, and gradient activity is often concentrated on a small subset of coordinates (Stich et al., 2018; Ivkin et al., 2019). Consequently, many parameters remain weakly updated during benign training, creating a low-interference space that attackers can exploit during poisoned optimization. Motivated by this vulnerability, we ask a different defense question: can the server proactively reshape the model's update landscape before poisoned updates arrive?

We view low-activity dimensions as underutilized model capacity. Malicious gradients written into them encounter little competing signal and can therefore accumulate across rounds. Proactively introducing auxiliary signals could limit the attacker's ability to exploit this capacity in isolation. To examine how parameter activity affects backdoor accumulation, we configure the attacker to rank coordinates by benign-gradient magnitude on local clean data and exclude the top-$k\%$ from malicious optimization, where $k = 0$ denotes no exclusion.

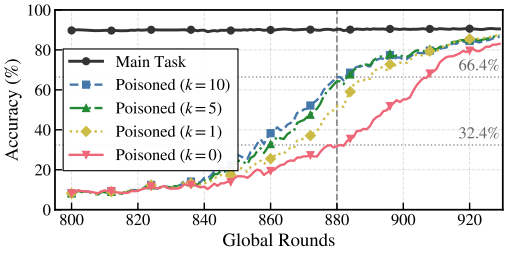


Figure 2: Main-task and backdoor accuracy with frequently updated parameters excluded.

Observations. As shown in Figure 2, excluding more frequently updated parameters leads to higher backdoor accuracy. With $k = 10$, the attack reaches 66.4% backdoor accuracy after 80 poisoning rounds, more than twice the accuracy obtained without exclusion. This result highlights parameter activity as a key factor underlying backdoor persistence and motivates a proactive defense strategy that reshapes parameter activity without relying on malicious-update detection.

Low-Activity Parameters as a Defense Surface. We refer to parameter coordinates receiving limited main-task gradient signal as *low-activity parameters*.

Definition 1 (Low-Activity Parameters)**.** *Let $\theta = (\theta_1, \ldots, \theta_d)$ denote the model parameters, and let $g_i(x, y; \theta) = \partial \ell(x, y; \theta)/\partial \theta_i$. The low-activity parameter set is*

$$\mathcal{R}_\tau(\theta) = \{i \mid \mathbb{E}_{(x,y)\sim\mathcal{D}_m} |g_i(x, y; \theta)| \le \tau\}, \tag{1}$$

where $\mathcal{D}_m$ is the main-task distribution, τ is the activity threshold, and $R_\tau(\theta) = |\mathcal{R}_\tau(\theta)|$.

Let $S_{\mathrm{bd}}(\theta)$ denote a continuous surrogate for the backdoor attack success rate, such as the expected target-class probability over triggered inputs.

Theorem 1 (Backdoor Sensitivity to Low-Activity Parameter Capacity)**.** *Let $\Delta_{\mathrm{bd},i}$ denote coordinate i of the effective malicious update, and let $\mathbf{P}_{\mathcal{R}^c}$ project onto coordinates outside $\mathcal{R}_\tau(\theta)$. Suppose*

$$|\Delta_{\mathrm{bd},i}| \le b \text{ for all } i \in \mathcal{R}_\tau(\theta), \qquad \|\mathbf{P}_{\mathcal{R}^c}\Delta_{\mathrm{bd}}\|_2 \le \epsilon. \tag{2}$$

If S_{bd} is locally L-Lipschitz with respect to the ℓ_2 parameter norm on a neighborhood containing θ and $\theta + \Delta_{\mathrm{bd}}$, then

$$|S_{\mathrm{bd}}(\theta + \Delta_{\mathrm{bd}}) - S_{\mathrm{bd}}(\theta)| \le L\sqrt{b^2 R_\tau(\theta) + \epsilon^2}. \tag{3}$$

The proof is provided in Appendix A. The terms $b^2 R_\tau(\theta)$ and ϵ^2 bound malicious-update energy within and outside the low-activity parameter set, respectively. For fixed L, b, and ϵ, reducing $R_\tau(\theta)$ tightens the bound on how much a malicious update can increase the backdoor score, motivating low-activity parameter capacity as a defense variable. We note that this bound characterizes local sensitivity to a single effective malicious update, while we empirically demonstrate sustained backdoor suppression through repeated robustification in Section 5.

3.3 RQ2: Can the Server Reduce Low-Activity Capacity without Private Data?

Low-activity parameters receive limited gradient signals from the main task, making them difficult to engage through benign client training alone. To increase their activity, we use public OOD data to generate diverse auxiliary gradients without accessing clients' private data. The key challenge is to reduce this low-activity capacity without introducing harmful shifts to the global model that compromise main-task performance.

Pilot Study Setup. We study this mechanism by training ResNet18 (He et al., 2016) on CIFAR10 (Krizhevsky et al., 2009) and using CIFAR100 (Krizhevsky et al., 2009) as OOD data. OOD samples receive random labels that are periodically refreshed to produce diverse auxiliary gradients rather than learn meaningful OOD semantics. We compare training with only OOD data, only main-task data, and a mixture of both. The latter two settings serve as references, since a practical FL server cannot access clients' main-task data.

Table 1: Low-activity parameter counts $R_\tau(\theta)$ across thresholds and main-task accuracy under three training scenarios.

τ	OOD Data	Main Data	Mixed Data
1×10^{-6}	94	512	266
1×10^{-5}	860	4,758	2,914
1×10^{-4}	8,219	47,951	28,346
5×10^{-4}	41,170	239,096	140,846
Main Acc.	32.07%	87.05%	86.91%

Observations. Table 1 shows that OOD-only training substantially reduces the number of low-activity parameters, but it also sharply degrades main-task accuracy. In contrast, mixed training reduces low-activity capacity while preserving accuracy, indicating that increasing parameter activity without sacrificing utility requires both diverse auxiliary gradients and main-task-consistent signals. This leaves a practical FL challenge: obtaining such signals without access to clients' private data.

Shadow Data as a Surrogate. To address this challenge, we assign current-model predictions to OOD samples and call the resulting set *shadow data*, using it as a surrogate for unavailable main-task signals. ReLU networks can produce overconfident predictions on OOD inputs (Hein et al., 2019). We use the current model's predictions as pseudo-labels to preserve its decision structure during auxiliary training. In the feature space of a CIFAR10 model, Figure 3 shows aligned clusters of shadow and in-distribution samples sharing the same predicted labels. This feature-space alignment motivates using shadow data as model-consistent auxiliary signals.

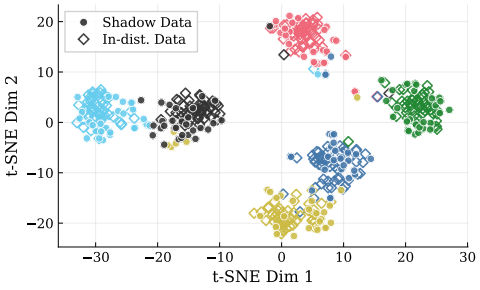


Figure 3: Feature space visualization of the shadow and in-distribution data with the same labels. Different colors represent different classes.

Together, these observations motivate combining diverse auxiliary gradients with model-consistent signals. The next section describes how FedHarden implements this combination using flood and shadow data.

4 Detailed Methodology

Before each FL round, FedHarden robustifies the current global model at the server. It constructs the flood and shadow datasets, identifies key kernels, and optimizes the corresponding parameter set before broadcasting the robustified model to clients. Algorithm 1 summarizes this workflow.

Constructing the Flood Dataset. The flood dataset D_f introduces diverse auxiliary gradients to increase parameter activity. For each OOD input, the server adds an independently sampled noise mask and assigns a random label from the main-task label space. These noise-label combinations avoid learning OOD semantics while generating gradients distinct from main-task signals. The server refreshes the noise masks and random labels each round to prevent overfitting and promote broader parameter activity.

Constructing the Shadow Dataset. Because flood data alone may disturb main-task representations, FedHarden constructs a shadow dataset D_s from another OOD sample set and assigns each sample the label predicted by the current global model. Refreshed each round, these pseudo-labels provide model-consistent guidance to preserve class-dependent representations as the decision boundary evolves. Together with flood data, shadow data enable the server to increase parameter activity without accessing clients' private data.

Algorithm 1: FedHarden

Input: Current global model $\boldsymbol{G}^t$, OOD samples $(\{x_f^i\}_{i=1}^N, \{x_s^j\}_{j=1}^M)$, robustification hyperparameters (E, η, λ, ρ), selected client set S_t at round t

Output: Global model $\boldsymbol{G}^{t+1}$

// Server-side robustification

Sample $n^i \sim U([-0.5, 0.5])$ and $y_f^i \sim U(\mathcal{Y}_m)$

$D_f = \{(x_f^i + n^i, y_f^i)\}_{i=1}^N$

$D_s = \{(x_s^j, y_s^j)\}_{j=1}^M,\ y_s^j = \arg\max \boldsymbol{G}^t(x_s^j)$

$\boldsymbol{\mathcal{P}} = \text{IdentifyKeyKernels}(D_f, D_s, \rho)$

Save BN statistics as $(\boldsymbol{\mu}_M, \boldsymbol{\sigma}_M)$ and set $\boldsymbol{\beta}' \leftarrow \boldsymbol{G}^t$

for $e = 1, ..., E$ **do**

  $\boldsymbol{g}_e = \nabla\big(\mathcal{L}_{\text{task}}(\boldsymbol{\beta}', D_f \| D_s) + \lambda \|\boldsymbol{\beta}' - \boldsymbol{G}^t\|_2\big)$

  Project $\boldsymbol{g}_e$ onto $\boldsymbol{\mathcal{P}}$ using $\Pi_{\boldsymbol{\mathcal{P}}}$, which zeros gradients outside $\boldsymbol{\mathcal{P}}$

  $\boldsymbol{\beta}' = \boldsymbol{\beta}' - \eta \Pi_{\boldsymbol{\mathcal{P}}}(\boldsymbol{g}_e)$

end

Restore BN statistics in $\boldsymbol{\beta}'$ with $(\boldsymbol{\mu}_M, \boldsymbol{\sigma}_M)$ and broadcast $\boldsymbol{\beta}'$

// Client training and aggregation

Clients train from $\boldsymbol{\beta}'$ and upload $\Delta\boldsymbol{\beta}_i = \boldsymbol{\beta}_i - \boldsymbol{\beta}'$

$\gamma = \text{median}_{i \in S_t} \|\Delta\boldsymbol{\beta}_i\|_2$

$\boldsymbol{G}^{t+1} = \boldsymbol{\beta}' + \frac{1}{|S_t|} \sum_{i \in S_t} \frac{\Delta\boldsymbol{\beta}_i}{\max(1, \|\Delta\boldsymbol{\beta}_i\|_2 / \gamma)}$

Identifying Key Kernels. Updating the entire model with OOD signals can disturb main-task behavior, whereas parameter-wise selection can fragment updates across feature channels. FedHarden therefore updates only convolutional kernels and associated BN parameters, leaving the classifier fixed. Kernels encode intermediate representations where backdoor-related features can be embedded (Goodfellow et al., 2016; Bagdasaryan et al., 2020), while BN parameters control feature normalization and scaling (Ioffe & Szegedy, 2015). The server scores each kernel by comparing its gradients on $D_f \| D_s$ and D_s; higher scores indicate greater sensitivity to flood data. We use this score as a practical criterion for prioritizing kernels whose gradients respond most strongly to the added flood signal, without accessing main-task data. The aim is to introduce competing signals into weakly updated model capacity while limiting the update scope. FedHarden selects the highest-scoring kernels and their BN parameters until $|\boldsymbol{\mathcal{P}}|/N_m$ reaches ρ. Here, $|\boldsymbol{\mathcal{P}}|$ counts selected scalar parameters, including associated BN parameters, and N_m counts model parameters outside the classifier. The detailed procedure is provided in Algorithm 2 in Appendix C.

Server-Side Robustification. Given D_f, D_s, and the selected parameter set $\boldsymbol{\mathcal{P}}$, the server optimizes a copy $\boldsymbol{\beta}'$ of the current global model before broadcasting:

$$\mathcal{L} = \mathcal{L}_{\text{task}}(\boldsymbol{\beta}', D_f \| D_s) + \lambda \|\boldsymbol{\beta}' - \boldsymbol{G}^t\|_2, \tag{4}$$

where $\mathcal{L}_{\text{task}}$ is mean cross-entropy on concatenated data $D_f \| D_s$, and regularization limits deviation from $\boldsymbol{G}^t$. The projection $\Pi_{\boldsymbol{\mathcal{P}}}$ zeros gradients outside the selected parameters. After optimization, the server restores the saved BN statistics to prevent drift (Li & Dai, 2024). This process repeats each round so kernel selection reflects the current global model. Repeated robustification supplies fresh auxiliary gradients as poisoned updates continue to arrive, to limit their accumulation across rounds. Clients then train locally from the robustified model, and the server aggregates updates with adaptive norm clipping as shown in Algorithm 1.

5 Experiments

5.1 Experimental Setup

Setup. We implement FedAvg (McMahan et al., 2017) on CIFAR10, CIFAR100 (Krizhevsky et al., 2009), and EMNIST (Cohen et al., 2017) with VGG16 (Simonyan & Zisserman, 2014), ResNet18, and ResNet34 (He et al., 2016). Unless otherwise specified, the system contains 100 clients parti-

Table 2: Single-client attack results: BA (MA), in %. No D.: NoDefense. Dark/light gray mark the lowest/second-lowest BA per row, also shown in bold/underline.

Attack	Trigger	No D.	DeepSight	FoolsGold	FLAME	FreqFed	BayBFed	MESAS	FLTrust	Indicator	FedHarden
PGD	Blended	78.73 (89.69)	41.70 (90.01)	75.90 (88.42)	89.02 (87.69)	79.64 (88.52)	83.25 (90.03)	77.65 (89.71)	77.05 (90.36)	43.32 (90.46)	**9.79** (88.50)
	Semantic	65.32 (90.43)	69.61 (90.27)	69.69 (89.89)	66.87 (88.30)	69.14 (87.47)	78.02 (88.26)	79.67 (87.91)	65.55 (88.95)	52.13 (90.35)	**17.85** (88.03)
	Edge Case	70.39 (90.44)	52.73 (90.16)	64.28 (90.36)	81.92 (88.18)	69.89 (88.88)	69.73 (89.04)	70.46 (88.19)	60.99 (89.35)	10.57 (89.76)	**10.38** (88.84)
	TaCT	96.34 (88.89)	90.18 (88.37)	96.22 (88.01)	99.36 (86.21)	91.89 (87.43)	83.87 (88.11)	92.67 (89.21)	93.70 (88.77)	**0.84** (89.88)	1.30 (87.21)
Neurotoxin	Blended	79.49 (91.04)	42.44 (89.58)	75.42 (89.51)	88.84 (88.84)	72.34 (88.18)	84.62 (89.14)	80.38 (90.02)	76.18 (89.70)	46.60 (90.22)	**10.00** (87.92)
	Semantic	64.42 (90.27)	67.40 (90.34)	68.37 (89.63)	67.40 (88.68)	77.25 (86.93)	80.41 (88.63)	89.41 (87.57)	66.51 (89.31)	51.84 (89.49)	**14.92** (89.42)
	Edge Case	67.27 (91.09)	53.57 (90.11)	58.73 (89.81)	82.96 (88.47)	58.03 (89.10)	70.17 (88.58)	61.23 (85.47)	56.83 (89.47)	18.65 (90.19)	**12.27** (88.30)
	TaCT	95.11 (88.33)	88.68 (88.84)	96.27 (88.15)	**0.25** (88.30)	71.15 (87.12)	69.83 (87.56)	86.60 (88.20)	93.47 (88.42)	1.54 (89.69)	1.10 (87.93)
Chameleon	Blended	78.91 (90.11)	43.74 (89.91)	74.66 (89.78)	88.99 (88.42)	49.87 (89.22)	81.45 (88.95)	80.61 (90.11)	72.55 (89.34)	40.77 (90.05)	**22.60** (89.27)
	Semantic	63.32 (90.36)	61.34 (89.98)	61.40 (90.44)	63.41 (88.67)	81.36 (88.20)	82.54 (89.54)	83.27 (86.24)	61.81 (88.79)	44.34 (89.44)	**16.52** (89.59)
	Edge Case	51.66 (90.30)	39.75 (89.90)	41.47 (88.26)	73.66 (89.67)	26.53 (89.17)	58.76 (88.77)	59.95 (87.82)	49.40 (88.94)	**12.89** (90.69)	20.67 (88.10)
	TaCT	97.35 (89.45)	78.04 (88.73)	94.16 (90.02)	99.38 (88.39)	76.51 (87.89)	67.80 (87.57)	94.50 (87.26)	93.40 (88.49)	1.85 (90.14)	**1.48** (89.03)
CerP	Optimized	79.09 (90.18)	45.81 (88.96)	74.27 (88.33)	90.99 (88.37)	72.15 (89.40)	79.26 (89.21)	77.42 (88.80)	63.80 (88.15)	41.69 (89.42)	**30.76** (87.28)
PFedBA	Optimized	97.47 (87.16)	82.68 (88.78)	96.70 (87.12)	98.99 (87.82)	96.77 (84.88)	96.22 (86.34)	99.73 (87.22)	99.93 (87.04)	96.57 (86.71)	**35.97** (86.62)

tioned by Dirichlet sampling ($\alpha = 0.9$), and 10 clients participate in each global round. We also vary α to test stronger non-IID settings. By default, malicious clients inject poisoned updates in each of 600 consecutive communication rounds. We evaluate single-client PGD, Neurotoxin, Chameleon, CerP, and PFedBA attacks, as well as multi-client Neurotoxin and DBA attacks. The evaluated triggers include blended, TaCT, semantic, edge-case, and optimized triggers.

Defenses and FedHarden Configuration. We compare FedHarden with DeepSight (Rieger et al., 2022), FoolsGold (Fung et al., 2020), FLAME (Nguyen et al., 2022), FreqFed (Fereidooni et al., 2023), BayBFed (Kumari et al., 2023), MESAS (Krauß & Dmitrienko, 2023), FLTrust (Cao et al., 2020), and BackdoorIndicator (Li & Dai, 2024) with adaptive norm clipping. *NoDefense* denotes adaptive norm clipping without an explicit backdoor detector. Backdoor accuracy (BA) measures the percentage of backdoor test inputs classified into the attacker's target label; main-task accuracy (MA) measures accuracy on clean test inputs. We report BA averaged over the final 20 global rounds. For FedHarden, the server constructs flood and shadow datasets from public OOD samples: CIFAR100 for CIFAR10 tasks and CIFAR10 for other main tasks. By default, the flood and shadow datasets contain 800 and 300 samples, $\rho = 0.15$, and the regularization weight is 0.8. Detailed experimental results are provided in Appendix B.1. We further analyze the time efficiency of FedHarden in Appendix B.5.

5.2 Results

Performance Against Single-Client Attacks. Table 2 shows that FedHarden achieves the lowest BA in nearly all single-client settings while keeping MA comparable to existing defenses. It reduces BA to 9.79% for PGD-trained blended backdoors and remains competitive on TaCT, where both BackdoorIndicator and FedHarden drive BA to around 1%.

FedHarden also remains effective against stronger malicious training and optimized trigger attacks. It achieves 14.92% BA for Neurotoxin-trained semantic backdoors, 22.60% for Chameleon-trained blended backdoors, and the lowest BA for CerP and PFedBA (30.76% and 35.97%). Under Neurotoxin blended attacks, FedHarden achieves 10.00% BA versus 42.44% for the strongest baseline, DeepSight, a $4.2\times$ reduction. Beyond the averages over the final 20 rounds, the trajectories in Appendix B.1 (Figure A1) show that FedHarden maintains lower BA than the plotted baselines during

Table 3: BA (MA), in %, under single-client Neurotoxin blended attacks with varying non-IID level (α) and poisoned learning rate (*plr*). Notation follows Table 2.

α	plr	No D.	DeepSight	FoolsGold	FLAME	FreqFed	BayBFed	MESAS	FLTrust	Indicator	FedHarden
0.9	0.01	69.31 (90.48)	42.20 (90.77)	63.94 (90.77)	83.16 (89.92)	71.61 (90.31)	73.77 (89.62)	65.41 (87.49)	64.83 (90.41)	36.67 (90.16)	**11.73** (88.95)
	0.025	79.49 (90.43)	42.44 (90.95)	75.42 (90.76)	88.84 (89.98)	72.34 (90.30)	84.62 (89.39)	80.38 (88.10)	76.18 (90.39)	46.60 (90.06)	**10.00** (88.99)
	0.04	85.06 (90.36)	42.42 (91.09)	82.05 (90.76)	12.26 (90.43)	55.32 (90.43)	88.26 (90.20)	83.64 (88.75)	82.32 (90.45)	63.58 (90.09)	**11.63** (89.23)
0.5	0.01	72.69 (89.57)	45.87 (89.33)	63.07 (89.06)	84.98 (88.61)	75.03 (88.76)	68.75 (89.16)	74.16 (88.79)	61.86 (89.06)	43.14 (90.11)	**14.45** (88.48)
	0.025	82.24 (89.46)	43.69 (88.18)	74.46 (89.81)	90.24 (88.43)	81.71 (88.71)	77.82 (88.35)	82.66 (88.81)	74.79 (89.04)	67.35 (90.36)	**17.12** (87.65)
	0.04	87.67 (89.35)	49.60 (89.41)	80.88 (89.80)	92.76 (88.41)	79.03 (88.66)	86.85 (88.48)	84.89 (87.66)	81.56 (89.02)	80.39 (90.35)	**23.45** (88.70)
0.2	0.01	48.42 (87.85)	37.11 (88.35)	45.30 (86.52)	78.28 (84.58)	64.19 (84.96)	63.84 (84.50)	67.54 (83.36)	41.05 (85.58)	14.21 (86.20)	**8.09** (85.74)
	0.025	66.13 (88.09)	39.61 (88.41)	45.15 (86.48)	86.93 (84.89)	69.51 (84.16)	78.24 (85.52)	80.41 (82.87)	55.77 (85.63)	32.29 (86.65)	**13.76** (85.85)
	0.04	78.03 (88.18)	40.65 (88.52)	69.74 (86.81)	**13.50** (85.29)	62.36 (84.39)	85.70 (84.51)	86.39 (83.05)	66.75 (85.60)	52.17 (87.19)	16.25 (86.30)

sustained Neurotoxin semantic and Chameleon blended poisoning. These results support backdoor suppression across the extended poisoning period.

Performance Under Different Non-IID Settings and Poisoned Learning Rates. On CIFAR10, FedHarden remains robust across varying data heterogeneity and poisoned learning rates (*plr*), achieving the lowest BA in eight of nine settings, with BA ranging from 8.09% to 23.45% (Table 3). At $\alpha = 0.2$ and $plr = 0.025$, it achieves 13.76% BA, compared with 32.29% for the second-best defense. This robustness extends to CIFAR100 and EMNIST under varying α and *plr* (Appendix B.4).

Table 4: BA (MA), in %, under multi-client Neurotoxin (blended) and DBA attacks. Ratios denote malicious-client fractions. Notation follows Table 2.

Attack (ratio)	No D.	DeepSight	FoolsGold	FLAME	FreqFed	BayBFed	Indicator	FedHarden
Neurotoxin (20%)	87.05 (90.39)	61.26 (90.95)	**10.66** (90.99)	91.76 (89.31)	50.90 (90.14)	88.67 (89.69)	53.91 (91.15)	11.22 (88.82)
Neurotoxin (30%)	87.77 (90.16)	67.58 (90.76)	10.92 (90.82)	93.25 (88.61)	15.56 (90.26)	90.22 (89.69)	53.55 (90.78)	**10.29** (87.07)
Neurotoxin (40%)	89.03 (89.93)	73.58 (90.62)	13.10 (90.79)	88.34 (77.33)	**12.91** (90.58)	90.38 (89.40)	50.94 (91.23)	15.62 (88.06)
DBA (40%)	96.84 (88.96)	90.37 (89.49)	**10.38** (90.03)	99.07 (85.07)	92.04 (89.97)	93.60 (88.76)	40.12 (90.43)	29.58 (88.35)

Performance Against Multiple-Client Attacks. Table 4 shows that FedHarden suppresses backdoors with up to 40% compromised clients. Under Neurotoxin with 20%–40% compromised clients, it keeps BA at 10%–16%; under 40% DBA, FedHarden achieves the second-lowest BA (29.58%).

Performance Across Model Architectures and Datasets. FedHarden remains effective across the tested architectures and datasets with comparable MA (Appendices B.2 and B.3). Under Neurotoxin semantic attacks on CIFAR10, it achieves 0.19% BA on ResNet34 and 0.48% on VGG16. Dataset tests with ResNet18 show 0.31% BA under Neurotoxin blended attacks on CIFAR100, versus 27.76% for the second-best defense. Under Neurotoxin pixel-pattern attacks on EMNIST, only FedHarden and BackdoorIndicator keep BA near 10%, while other defenses approach 100%.

5.3 Ablation Studies

We further analyze key design choices and hyperparameters, including kernel-wise selection and flood/shadow data sources. Additional sensitivity analyses on flood size and the selected-parameter ratio ρ are provided in Appendix B.7 (Tables A9 and A10). Larger flood datasets delay reaching the 35% BA threshold, while jointly tuning flood size and ρ balances suppression and MA.

Table 5: Parameter-selection ablation: BA (MA), in %. Columns indicate flood dataset sizes.

Strategy	200	500	800	1000
No Selection	51.02 (89.55)	43.41 (89.74)	52.83 (90.46)	48.18 (90.50)
Parameter-wise	42.23 (85.54)	40.33 (84.18)	39.47 (84.02)	38.16 (81.22)
Kernel-wise	21.20 (89.23)	12.84 (89.08)	11.95 (88.69)	10.85 (88.14)

Ablation Study on Kernel-Wise Selection. We compare kernel-wise selection with no selection and parameter-wise selection, which applies Algorithm 2 to individual parameters. Kernel-wise and parameter-wise selection use the same trainable-parameter budget. As shown in Table 5, kernel-wise selection achieves lower BA while maintaining MA close to no selection, whereas both alternatives retain 38%–53% BA. This suggests that kernel-level updates robustify more effectively than full-model or parameter-wise updates.

Table 6: Influence of flood and shadow dataset sources on BA (%) under Chameleon blended attacks.

Shadow \ Flood	CIFAR100	300KRANDOM	EMNIST	Noise
CIFAR100	22.58±4.22	28.59±3.02	64.97±3.63	55.65±2.17
300KRANDOM	22.91±3.60	22.90±3.15	72.11±2.41	63.31±2.36

Influence of Flood and Shadow Dataset Sources. Table 6 shows that flood-source choice has a larger effect than shadow-source choice. CIFAR100 and 300KRANDOM (Hendrycks et al., 2019) reduce BA to 22%–29%, while EMNIST and random noise leave it above 55%. Given diverse flood data, shadow-source choice has little impact. These results favor diverse public natural images for flood construction. Such images are readily available from public datasets such as ImageNet (Deng et al., 2009) and general-purpose photo collections. FedHarden further simplifies data preparation by requiring no ground-truth annotations: flood labels are assigned randomly, while shadow labels are generated by the current model.

5.4 Resilience to Adaptive Attacks

We evaluate FedHarden-aware adversaries that avoid robustified kernels or submit large-norm updates.

Avoiding Key Kernels. An adaptive adversary applies Algorithm 2 using self-collected OOD data and excludes the identified kernels from poisoning. Under 500 rounds of blended poisoning on CIFAR10, FedHarden keeps BA at 14.32%–19.74%. Detailed results are provided in Appendix B.6.

Large-Norm Updates. We further evaluate a Neurotoxin adversary using blended backdoors with $plr = 0.1$, which produces dominant malicious updates. FedHarden mitigates this strategy through its norm-clipping defense (NCD), using the median norm of received updates as the threshold. Table 7 shows that combining proactive robustification with NCD achieves 18.9% BA, compared with 69.31% for NCD alone and 52.47% for robustification alone. This supports combining update-magnitude control with repeated robustification against large-norm poisoning.

Table 7: Performance against adaptive large-norm updates.

Method	BA (%)
Only NCD	69.31±2.35
FedHarden wo. NCD	52.47±4.35
FedHarden w. NCD	18.9±3.59

6 Conclusion

FedHarden is a proactive defense against long-term backdoor poisoning in federated learning. It robustifies the global model at the server side before communication rounds using public OOD data, without relying on malicious-update detection. Across diverse attacks, datasets, architectures, and adaptive settings, FedHarden achieves sustained BA suppression over the evaluated poisoning horizons with comparable MA. Extending this robustification principle beyond vision and convolutional networks to transformer-based and multimodal FL remains future work.

AI USE STATEMENT

We used generative AI tools to assist with manuscript editing and organization, and literature search. The authors retain responsibility for all scientific claims, experimental results, and the final manuscript.

ETHICS STATEMENT

This work aims to improve the security of federated learning against backdoor attacks. Our evaluation uses public image benchmarks and simulated malicious clients in controlled FL experiments. FedHarden operates without collecting clients' raw training data. The attack scenarios and accompanying implementation are intended to support defensive research and reproducible security evaluation.

REPRODUCIBILITY STATEMENT

The FedHarden procedure and kernel-selection rule are specified in Algorithms 1 and 2. Experimental settings and evaluation metrics are described in Section 5.1; the proof and additional results are provided in Appendices A and B. Code, experiment configurations, and instructions for obtaining the public datasets will be made available through an anonymous GitHub repository: `https://anonymous.4open.science/r/FedHarden-artifact-7085/README.md`

REFERENCES

Eugene Bagdasaryan, Andreas Veit, Yiqing Hua, Deborah Estrin, and Vitaly Shmatikov. How to backdoor federated learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 2938–2948. PMLR, 2020.

Arjun Nitin Bhagoji, Supriyo Chakraborty, Prateek Mittal, and Seraphin Calo. Analyzing federated learning through an adversarial lens. In *International Conference on Machine Learning*, pp. 634–643. PMLR, 2019.

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

Xiaoyu Cao, Jinyuan Jia, and Neil Zhenqiang Gong. Provably secure federated learning against malicious clients. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 6885–6893, 2021.

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.

Minyeong Choe, Cheolhee Park, Changho Seo, and Hyunil Kim. SDBA: A Stealthy and Long-Lasting Durable Backdoor Attack in Federated Learning . *IEEE Transactions on Dependable and Secure Computing*, 22(06):6929–6942, November 2025. ISSN 1941-0018. doi: 10.1109/TDSC.2025.3593640. URL `https://doi.ieeecomputersociety.org/10.1109/TDSC.2025.3593640`.

Gregory Cohen, Saeed Afshar, Jonathan Tapson, and Andre Van Schaik. Emnist: Extending mnist to handwritten letters. In *2017 international joint conference on neural networks (IJCNN)*, pp. 2921–2926. IEEE, 2017.

Yanbo Dai and Songze Li. Chameleon: Adapting to peer images for planting durable backdoors in federated learning. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 6712–6725. PMLR, 23–29 Jul 2023. URL `https://proceedings.mlr.press/v202/dai23a.html`.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Hossein Fereidooni, Alessandro Pegoraro, Phillip Rieger, Alexandra Dmitrienko, and Ahmad-Reza Sadeghi. Freqfed: A frequency analysis-based approach for mitigating poisoning attacks in federated learning. *arXiv preprint arXiv:2312.04432*, 2023.

Clement Fung, Chris J. M. Yoon, and Ivan Beschastnikh. The limitations of federated learning in sybil settings. In *23rd International Symposium on Research in Attacks, Intrusions and Defenses (RAID 2020)*, pp. 301–316, San Sebastian, October 2020. USENIX Association. ISBN 978-1-939133-18-2. URL https://www.usenix.org/conference/raid2020/presentation/fung.

Yinghua Gao, Dongxian Wu, Jingfeng Zhang, Guanhao Gan, Shu-Tao Xia, Gang Niu, and Masashi Sugiyama. On the effectiveness of adversarial training against backdoor attacks. *IEEE Transactions on Neural Networks and Learning Systems*, 35(10):14878–14888, 2024. doi: 10.1109/TNNLS.2023.3281872.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*, volume 1. 2016.

Tianyu Gu, Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Evaluating backdooring attacks on deep neural networks. *IEEE Access*, 7:47230–47244, 2019.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Wenwen He, Wenke Huang, Bin Yang, Shukan Liu, and Mang Ye. SPMC: Self-purifying federated backdoor defense via margin contribution. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 22422–22433. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/he25f.html.

Matthias Hein, Maksym Andriushchenko, and Julian Bitterwolf. Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 41–50, 2019.

Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HyxCxhRcY7.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pp. 448–456. pmlr, 2015.

Nikita Ivkin, Daniel Rothchild, Enayat Ullah, Ion Stoica, Raman Arora, et al. Communication-efficient distributed sgd with sketching. *Advances in Neural Information Processing Systems*, 32, 2019.

Torsten Krauß and Alexandra Dmitrienko. Mesas: Poisoning defense for federated learning resilient against adaptive attackers. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, pp. 1526–1540, 2023.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Kavita Kumari, Phillip Rieger, Hossein Fereidooni, Murtuza Jadliwala, and Ahmad-Reza Sadeghi. Baybfed: Bayesian backdoor defense for federated learning. In *2023 IEEE Symposium on Security and Privacy (SP)*, pp. 737–754. IEEE, 2023.

Songze Li and Yanbo Dai. BackdoorIndicator: Leveraging OOD data for proactive backdoor detection in federated learning. In *33rd USENIX Security Symposium (USENIX Security 24)*, pp. 4193–4210, Philadelphia, PA, August 2024. USENIX Association. ISBN 978-1-939133-44-1. URL https://www.usenix.org/conference/usenixsecurity24/presentation/li-songze.

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 *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 9020–9028, 2023.

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)*, pp. 4157–4174, 2024.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rJzIBfZAb.

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*, pp. 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 *31st USENIX Security Symposium (USENIX Security 22)*, pp. 1415–1432, 2022.

Phillip Rieger, Thien Duc Nguyen, Markus Miettinen, and Ahmad-Reza Sadeghi. Deepsight: Mitigating backdoor attacks in federated learning through deep model inspection. In *NDSS*, 2022.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Sebastian U Stich, Jean-Baptiste Cordonnier, and Martin Jaggi. Sparsified sgd with memory. *Advances in neural information processing systems*, 31, 2018.

Ziteng Sun, Peter Kairouz, Ananda Theertha Suresh, and H Brendan McMahan. Can you really backdoor federated learning? *arXiv preprint arXiv:1911.07963*, 2019.

Hongyi Wang, Kartik Sreenivasan, Shashank Rajput, Harit Vishwakarma, Saurabh Agarwal, Jyyong Sohn, Kangwook Lee, and Dimitris Papailiopoulos. Attack of the tails: Yes, you really can backdoor federated learning. *Advances in Neural Information Processing Systems*, 33:16070–16084, 2020.

Cheng-Hsin Weng, Yan-Ting Lee, and Shan-Hung Brandon Wu. On the trade-off between adversarial and backdoor robustness. *Advances in Neural Information Processing Systems*, 33:11973–11983, 2020.

Kaiyuan Zhang, Guanhong Tao, Qiuling Xu, Siyuan Cheng, Shengwei An, Yingqi Liu, Shiwei Feng, Guangyu Shen, Pin-Yu Chen, Shiqing Ma, and Xiangyu Zhang. FLIP: A provable defense framework for backdoor mitigation in federated learning. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=Xo2E217_M4n.

Zhengming Zhang, Ashwinee Panda, Linyue Song, Yaoqing Yang, Michael Mahoney, Prateek Mittal, Ramchandran Kannan, and Joseph Gonzalez. Neurotoxin: Durable backdoors in federated learning. In *International Conference on Machine Learning*, pp. 26429–26446. PMLR, 2022.

A Proof of the Sensitivity Bound

We prove Theorem 1 using the low-activity parameter set and backdoor-score notation defined in the main text.

Proof. Let $\mathcal{R} = \mathcal{R}_\tau(\theta)$. The coordinates in $\mathcal{R}$ and its complement are disjoint, so

$$\|\Delta_{\mathrm{bd}}\|_2^2 = \sum_{i\in\mathcal{R}} |\Delta_{\mathrm{bd},i}|^2 + \|\mathbf{P}_{\mathcal{R}^c}\Delta_{\mathrm{bd}}\|_2^2 \tag{5}$$

$$\leq b^2 R_\tau(\theta) + \epsilon^2. \tag{6}$$

Thus,

$$\|\Delta_{\mathrm{bd}}\|_2 \leq \sqrt{b^2 R_\tau(\theta) + \epsilon^2}. \tag{7}$$

By the local L-Lipschitz continuity of S_{bd} with respect to the ℓ_2 parameter norm,

$$|S_{\mathrm{bd}}(\theta + \Delta_{\mathrm{bd}}) - S_{\mathrm{bd}}(\theta)| \leq L\|\Delta_{\mathrm{bd}}\|_2 \tag{8}$$

$$\leq L\sqrt{b^2 R_\tau(\theta) + \epsilon^2}. \tag{9}$$

□

B Extended Experimental Results

B.1 Long-Horizon Behavior under Persistent Poisoning

FedHarden is activated at global round 400, with rounds 400–429 serving as a 30-round warm-up before poisoning begins at round 430. Poisoning continues for 600 rounds, with malicious clients participating in every poisoning round. Server-side robustification is applied in every round from activation onward. FedHarden remains effective with short or no warm-up, while a longer warm-up further improves backdoor suppression. We analyze the effect of warm-up length in Appendix B.8.

Figure A1 shows BA over communication rounds under Neurotoxin semantic and Chameleon blended attacks. These trajectories complement the average accuracies in Table 2 by showing behavior over an extended poisoning period.

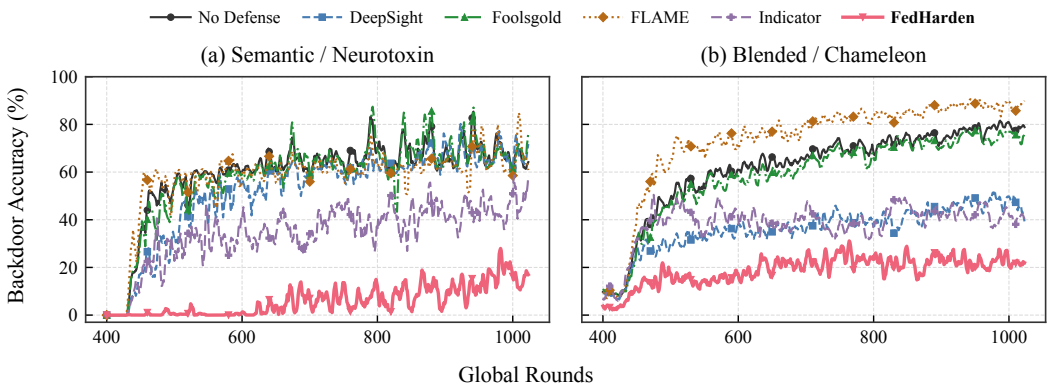


Figure A1: Backdoor accuracy over global rounds under persistent (a) Neurotoxin semantic attacks and (b) Chameleon blended attacks. Curves use a five-round moving average.

B.2 Additional Model Architectures

Tables A1 and A2 report the complete single-client attack results on ResNet34 and VGG16, respectively. Both use CIFAR10 and evaluate PGD, Neurotoxin, and Chameleon across trigger types. FedHarden achieves the lowest or second-lowest BA in most settings while maintaining MA around 88%–90%.

Table A1: Single-client attacks on ResNet34 trained on CIFAR10, across attack algorithms and trigger types. Each cell lists BA above (MA), in %; No D. denotes NoDefense. Darker and lighter gray mark the lowest (**bold**) and second-lowest (underlined) BA per row; ties share the same highlighting.

Attack	Trigger	No D.	DeepSight	FoolsGold	FLAME	FreqFed	BayBFed	MESAS	FLTrust	Indicator	FedHarden
PGD	Blended	76.41 (90.03)	39.24 (90.93)	72.76 (89.45)	86.68 (89.23)	62.31 (90.40)	71.22 (89.84)	60.24 (90.32)	61.24 (88.22)	55.44 (90.30)	**28.20** (89.46)
	Semantic	60.27 (90.09)	60.70 (90.74)	62.97 (90.53)	66.54 (89.45)	43.17 (90.83)	79.39 (89.53)	90.24 (90.12)	57.34 (89.51)	55.04 (90.31)	**0.00** (88.69)
	Edge Case	46.02 (90.11)	34.10 (90.14)	46.28 (90.73)	68.09 (89.67)	42.61 (89.23)	51.24 (90.29)	44.30 (89.89)	33.45 (89.69)	19.82 (90.53)	**11.59** (89.30)
	TaCT	48.28 (90.03)	48.10 (90.88)	86.45 (90.46)	39.33 (89.20)	93.81 (90.98)	81.09 (89.33)	78.19 (90.19)	92.67 (89.50)	39.10 (89.73)	**8.02** (88.74)
Neurotoxin	Blended	77.00 (90.97)	38.79 (90.84)	75.26 (90.62)	88.95 (89.35)	**19.29** (88.40)	63.04 (89.61)	64.12 (90.35)	61.23 (88.44)	53.91 (90.38)	27.72 (88.34)
	Semantic	68.18 (90.93)	61.38 (90.76)	70.95 (90.59)	69.58 (89.40)	**0.00** (90.12)	78.08 (89.74)	87.05 (90.15)	59.96 (88.54)	26.85 (90.22)	0.19 (89.22)
	Edge Case	50.08 (90.97)	36.34 (90.30)	56.88 (88.69)	70.00 (89.69)	14.70 (90.47)	41.85 (89.39)	36.44 (89.19)	32.40 (89.81)	31.81 (90.51)	**9.71** (89.02)
	TaCT	60.60 (90.94)	50.23 (90.06)	92.77 (90.33)	0.36 (90.04)	87.60 (89.88)	66.50 (89.38)	59.41 (90.07)	89.55 (88.83)	**0.34** (90.29)	18.35 (88.91)
Chameleon	Blended	61.56 (90.02)	35.88 (90.96)	54.98 (89.57)	78.12 (89.39)	**24.19** (90.71)	50.10 (89.89)	59.01 (90.26)	50.45 (89.55)	45.80 (90.36)	29.91 (88.39)
	Semantic	61.67 (90.95)	32.88 (90.78)	61.20 (90.53)	67.07 (89.39)	54.63 (90.53)	38.73 (89.43)	74.01 (90.02)	41.94 (89.51)	46.04 (90.29)	**0.00** (88.39)
	Edge Case	26.77 (90.04)	18.04 (91.03)	28.92 (90.36)	46.94 (89.31)	18.10 (90.06)	25.40 (89.46)	28.37 (89.89)	21.07 (90.30)	26.84 (90.34)	**13.70** (89.18)
	TaCT	78.62 (89.23)	72.13 (90.85)	77.63 (90.28)	65.20 (90.25)	28.98 (90.07)	66.09 (89.59)	85.36 (90.04)	57.35 (89.90)	30.55 (90.47)	**5.63** (88.87)

Table A2: Single-client attacks on VGG16 trained on CIFAR10, across attack algorithms and trigger types. Each cell lists BA above (MA), in %; No D. denotes NoDefense. Darker and lighter gray mark the lowest (**bold**) and second-lowest (underlined) BA per row; ties share the same highlighting.

Attack	Trigger	No D.	DeepSight	FoolsGold	FLAME	FreqFed	BayBFed	MESAS	FLTrust	Indicator	FedHarden
PGD	Blended	83.47 (90.17)	43.22 (90.94)	81.20 (90.07)	90.23 (89.57)	60.84 (90.52)	71.10 (89.20)	68.11 (90.66)	59.46 (90.48)	20.20 (90.95)	**13.21** (89.05)
	Semantic	85.35 (90.08)	86.45 (90.22)	86.37 (89.85)	70.40 (89.57)	54.61 (90.13)	88.71 (88.66)	76.90 (90.47)	67.74 (89.38)	76.22 (90.76)	**18.64** (89.89)
	Edge Case	70.95 (90.38)	54.61 (90.29)	74.57 (90.00)	89.49 (89.72)	56.13 (90.35)	62.80 (89.18)	64.88 (90.22)	45.59 (90.75)	**5.96** (90.93)	25.13 (90.34)
Neurotoxin	Blended	85.12 (90.18)	44.58 (90.05)	82.68 (90.12)	94.17 (89.29)	60.42 (90.56)	65.97 (89.20)	62.13 (90.75)	58.68 (89.25)	35.51 (90.73)	**25.40** (89.08)
	Semantic	92.81 (90.11)	85.56 (90.17)	89.26 (89.92)	76.52 (88.47)	53.63 (89.96)	73.16 (88.55)	94.51 (89.16)	75.41 (90.41)	73.39 (90.65)	**0.48** (88.70)
	Edge Case	78.91 (90.22)	59.57 (90.13)	78.81 (89.95)	88.61 (89.63)	46.79 (90.35)	64.75 (88.63)	59.89 (90.11)	52.23 (90.51)	11.61 (89.82)	**8.19** (89.15)
Chameleon	Blended	89.83 (90.28)	49.69 (90.19)	87.92 (90.06)	**10.13** (90.35)	17.85 (90.56)	76.88 (89.29)	78.78 (90.26)	69.26 (90.55)	39.93 (90.56)	28.76 (88.76)
	Semantic	79.53 (90.19)	70.57 (90.13)	71.15 (90.14)	**0.00** (90.37)	27.72 (90.65)	77.47 (89.28)	86.60 (89.09)	71.60 (89.93)	78.89 (89.83)	7.21 (90.26)
	Edge Case	81.02 (90.18)	67.85 (90.10)	80.83 (89.89)	**2.84** (89.20)	8.16 (90.91)	53.32 (89.48)	64.42 (90.44)	54.46 (90.13)	20.62 (90.82)	7.34 (89.69)

B.3 Additional Datasets

Tables A3 and A4 report extended results on CIFAR100 and EMNIST with ResNet18. On CIFAR100, FedHarden keeps BA at or below 3.70% across all listed attack–trigger combinations. On EMNIST, it limits BA to 10.27%–18.58%, whereas most baselines remain near 100%.

Table A3: Single-client attacks on CIFAR100 with ResNet18, across attack algorithms and trigger types. Each cell lists BA above (MA), in %; No D. denotes NoDefense. Darker and lighter gray mark the lowest (**bold**) and second-lowest (underlined) BA per row; ties share the same highlighting.

Attack	Trigger	No D.	DeepSight	FoolsGold	FLAME	FreqFed	BayBFed	MESAS	FLTrust	Indicator	FedHarden
PGD	Blended	66.72 (68.53)	28.72 (68.13)	63.35 (68.12)	81.97 (66.18)	62.73 (66.57)	69.40 (66.57)	66.07 (66.19)	45.93 (65.65)	13.07 (66.35)	**0.22** (64.78)
	Edge Case	81.84 (68.60)	66.85 (67.88)	81.29 (68.02)	92.07 (66.27)	86.95 (65.24)	84.47 (66.65)	79.37 (66.37)	76.80 (66.03)	**0.00** (67.38)	1.14 (64.45)
	TaCT	97.51 (67.84)	80.41 (67.70)	97.39 (67.71)	100.00 (65.18)	78.63 (65.02)	96.65 (66.54)	95.82 (66.75)	85.81 (66.80)	**0.00** (67.23)	**0.00** (65.21)
Neurotoxin	Blended	65.20 (68.55)	27.76 (68.41)	63.78 (68.30)	63.78 (66.24)	60.92 (65.14)	66.44 (66.63)	66.39 (66.38)	56.61 (67.03)	52.30 (67.28)	**0.31** (65.55)
	Edge Case	83.83 (68.40)	69.33 (68.05)	80.99 (67.95)	92.09 (66.39)	85.71 (65.57)	85.21 (66.66)	78.66 (66.72)	77.83 (67.45)	21.38 (67.47)	**1.79** (65.72)
	TaCT	96.86 (67.72)	80.44 (67.69)	97.02 (67.69)	**0.00** (66.41)	80.32 (64.59)	96.68 (66.38)	91.08 (66.09)	81.90 (65.24)	**0.00** (67.13)	**0.00** (63.68)
Chameleon	Blended	28.21 (68.24)	16.49 (68.30)	26.24 (68.29)	54.68 (66.30)	19.89 (66.37)	23.46 (66.30)	24.28 (66.89)	17.25 (65.33)	12.12 (67.21)	**0.01** (66.35)
	Edge Case	42.28 (68.52)	33.85 (68.33)	39.74 (68.20)	62.08 (66.28)	42.78 (64.96)	47.05 (66.22)	35.62 (66.74)	27.88 (64.03)	**0.17** (67.12)	3.70 (67.33)
	TaCT	90.10 (67.74)	48.86 (67.78)	86.76 (67.66)	98.68 (65.08)	45.01 (64.12)	54.52 (66.61)	60.86 (66.50)	60.17 (66.66)	**0.00** (67.21)	**0.00** (64.21)

Table A4: Single-client pixel-pattern attacks on EMNIST with ResNet18, across malicious training algorithms. Each cell lists BA above (MA), in %; No D. denotes NoDefense. Darker and lighter gray mark the lowest (**bold**) and second-lowest (underlined) BA per row; ties share the same highlighting.

Attack	Trigger	No D.	DeepSight	FoolsGold	FLAME	FreqFed	BayBFed	MESAS	FLTrust	Indicator	FedHarden
PGD	Pixel-pattern	100.00 (99.71)	99.99 (99.69)	100.00 (99.70)	**10.01** (99.67)	100.00 (99.61)	100.00 (99.65)	100.00 (99.70)	99.99 (99.67)	66.23 (99.70)	10.27 (99.72)
Neurotoxin	Pixel-pattern	99.20 (99.71)	99.72 (99.69)	99.77 (99.70)	100.00 (99.69)	100.00 (99.65)	100.00 (99.67)	100.00 (99.69)	99.99 (99.67)	**9.99** (99.70)	10.74 (99.48)
Chameleon	Pixel-pattern	100.00 (99.68)	100.00 (99.69)	100.00 (99.70)	**10.01** (99.68)	38.02 (99.64)	100.00 (99.66)	100.00 (99.68)	100.00 (99.65)	65.78 (99.68)	18.58 (99.50)

B.4 Data Heterogeneity and Poisoned Learning Rates

Tables A5 and A6 report results on CIFAR100 and EMNIST as the Dirichlet parameter α and poisoned learning rate (*plr*) vary. On CIFAR100, FedHarden keeps BA between 0.07% and 1.91% across all tested settings. On EMNIST, BA remains between 10.01% and 12.54% while MA stays above 99%, whereas several baselines vary sharply. These results demonstrate stable performance across data heterogeneity and attack strengths.

Table A5: Single-client blended-trigger Neurotoxin attacks on CIFAR100 with varying Dirichlet parameter α and poisoned learning rate (*plr*). Each cell lists BA above (MA), in %; No D. denotes NoDefense. Darker and lighter gray mark the lowest (**bold**) and second-lowest (underlined) BA per row; ties share the same highlighting.

α	plr	No D.	DeepSight	FoolsGold	FLAME	FreqFed	BayBFed	MESAS	FLTrust	Indicator	FedHarden
0.9	0.01	53.43 (67.58)	23.93 (67.91)	52.04 (67.71)	68.87 (65.12)	50.35 (65.76)	53.00 (66.23)	50.45 (66.76)	40.32 (66.32)	13.51 (67.34)	**0.29** (67.21)
	0.025	65.20 (68.55)	27.76 (68.41)	63.78 (68.30)	63.78 (66.24)	60.92 (65.14)	66.44 (66.63)	66.39 (66.38)	56.61 (67.03)	52.30 (67.28)	**0.31** (65.55)
	0.04	73.10 (68.56)	28.90 (67.89)	70.85 (68.29)	**0.16** (65.92)	5.38 (65.99)	74.14 (66.73)	76.37 (66.53)	67.04 (65.82)	63.01 (67.31)	0.70 (65.03)
0.5	0.01	53.46 (68.10)	28.45 (67.83)	50.79 (67.87)	67.97 (65.34)	46.34 (65.43)	56.36 (66.01)	51.42 (66.83)	41.14 (66.44)	0.34 (67.62)	**0.07** (66.48)
	0.025	67.85 (68.00)	31.58 (68.00)	64.85 (67.87)	78.45 (65.15)	41.88 (65.30)	68.07 (65.55)	66.89 (65.92)	55.47 (65.99)	5.67 (67.42)	**0.66** (66.73)
	0.04	64.59 (68.05)	27.02 (67.78)	63.31 (68.06)	0.23 (66.34)	39.68 (65.67)	69.94 (66.35)	68.47 (66.89)	64.76 (65.92)	18.76 (67.31)	**0.18** (67.07)
0.2	0.01	59.90 (66.75)	28.27 (66.56)	56.72 (66.76)	71.81 (62.53)	59.49 (60.73)	47.45 (64.17)	61.25 (64.51)	42.37 (62.00)	0.24 (66.14)	**0.11** (65.05)
	0.025	74.01 (66.92)	32.71 (66.69)	56.24 (66.77)	82.58 (62.50)	70.03 (59.66)	65.98 (64.40)	69.74 (64.93)	59.66 (64.50)	10.01 (66.06)	**0.12** (64.47)
	0.04	81.24 (66.80)	34.01 (66.86)	79.60 (66.71)	**0.25** (63.98)	19.13 (60.22)	79.23 (64.91)	79.88 (64.86)	70.64 (64.04)	14.78 (66.07)	1.91 (64.52)

Table A6: Single-client pixel-pattern Neurotoxin attacks on EMNIST with varying Dirichlet parameter α and poisoned learning rate (*plr*). Each cell lists BA above (MA), in %; No D. denotes NoDefense. Darker and lighter gray mark the lowest (**bold**) and second-lowest (underlined) BA per row; ties share the same highlighting.

α	plr	No D.	DeepSight	FoolsGold	FLAME	FreqFed	BayBFed	MESAS	FLTrust	Indicator	FedHarden
0.9	0.01	100.00 (99.71)	99.99 (99.75)	100.00 (99.71)	**10.01** (99.65)	99.99 (99.59)	100.00 (99.66)	99.98 (99.69)	99.99 (99.64)	**10.01** (99.72)	**10.01** (99.45)
	0.025	99.20 (99.71)	99.72 (99.69)	99.77 (99.70)	100.00 (99.69)	100.00 (99.65)	100.00 (99.67)	100.00 (99.69)	99.99 (99.67)	**9.99** (99.70)	10.74 (99.48)
	0.04	100.00 (99.70)	100.00 (99.76)	100.00 (99.71)	**10.00** (99.67)	100.00 (99.61)	99.98 (99.64)	99.99 (99.65)	100.00 (99.69)	41.19 (99.71)	10.75 (99.52)
0.5	0.01	99.98 (99.67)	98.45 (99.71)	99.99 (99.69)	10.67 (99.67)	99.94 (99.54)	99.96 (99.65)	99.95 (99.66)	99.60 (99.64)	**10.00** (99.66)	10.11 (99.43)
	0.025	100.00 (99.67)	99.81 (99.70)	99.98 (99.67)	**10.01** (99.67)	99.91 (99.55)	100.00 (99.65)	99.99 (99.66)	99.92 (99.65)	10.04 (99.66)	10.14 (99.35)
	0.04	100.00 (99.66)	99.30 (99.71)	100.00 (99.68)	**10.01** (99.68)	99.99 (99.60)	99.99 (99.64)	99.98 (99.57)	99.96 (99.57)	10.03 (99.67)	10.25 (99.47)
0.2	0.01	99.77 (99.61)	98.89 (99.68)	99.45 (99.62)	99.98 (99.66)	98.84 (99.12)	94.27 (99.64)	92.05 (99.41)	96.87 (99.55)	**10.04** (99.66)	12.40 (99.53)
	0.025	99.90 (99.60)	99.27 (99.68)	99.43 (99.62)	**10.02** (99.64)	99.97 (99.50)	96.63 (99.64)	99.90 (99.40)	99.23 (99.46)	10.03 (99.66)	12.54 (99.42)
	0.04	99.93 (99.58)	99.30 (99.67)	99.88 (99.61)	**10.02** (99.63)	99.97 (99.15)	99.95 (99.59)	99.94 (99.60)	99.65 (99.22)	10.04 (99.65)	12.34 (99.06)

B.5 Computational Cost

Server-Side Runtime. Table A7 reports server-side wall-clock time per global round in the single-machine setup with an NVIDIA RTX A6000. With 400 flood samples, FedHarden takes 45.3 seconds on ResNet18, compared with 48.1 seconds for Indicator and 77.3 seconds for FreqFed. On ResNet34, its runtime is 67.7 seconds, compared with 58.9 seconds for Indicator and 143.2 seconds for FreqFed. The flood size controls the additional computation: using 200 samples reduces the ResNet18 runtime to 27.5 seconds, while the default 800 samples require 95.7 seconds. Together with Table A9, these measurements show how flood size controls the balance between server computation, sustained suppression, and main-task utility.

Table A7: Server-side time per global round, in seconds. Parentheses in FedHarden variants give the number of flood samples; – denotes an unreported measurement. Darker and lighter gray mark the lowest (**bold**) and second-lowest (underlined) time per column.

Method	ResNet18	ResNet34
DeepSight	38.2	**48.3**
FreqFed	77.3	143.2
Indicator	48.1	58.9
FedHarden (200)	**27.5**	–
FedHarden (400)	45.3	67.7
FedHarden (800)	95.7	–

B.6 Adaptive Key-Kernel Avoidance

Table A8 reports per-setting BA for an adaptive adversary that identifies and avoids key kernels using self-collected OOD data. We evaluate selected-parameter ratios $\rho = 0.15$, 0.20, and 0.25 under 500 rounds of blended poisoning on CIFAR10.

Table A8: BA under adaptive key-kernel avoidance during 500 rounds of blended poisoning on CIFAR10. Darker and lighter gray mark the lowest (**bold**) and second-lowest (underlined) BA.

ρ	BA (%)
0.15	19.74±5.95
0.20	16.14±5.32
0.25	**14.32±4.30**

B.7 Sensitivity to Flood Dataset Size and Selected-Parameter Ratio

Influence of the Flood Dataset Size. We evaluate the robustness–utility trade-off under Neurotoxin attacks with a blended trigger. We report MA over the final 20 global rounds and *effective length*, defined as the first global round at which BA reaches 35%, reported as an absolute global-round index. Table A9 shows that increasing the flood size from 200 to 400 extends the effective length from 588 to 1874 rounds. Larger datasets further extend it to 2104 rounds at 800 samples, but yield diminishing gains and reduce MA from 89.02% to 86.67%. Thus, the flood size controls the balance between sustained backdoor suppression and main-task utility.

Table A9: Effect of the flood dataset size. Larger effective length and MA are better. Darker and lighter gray mark the best (**bold**) and second-best (underlined) attacked settings per column; the no-attack row is a reference.

Flood Size	Effective Length	MA (%)
200	588	**89.02±0.31**
400	1874	87.62±0.35
600	1990	87.02±0.72
800	**2104**	86.67±0.70
No attack	–	90.18±0.40

Influence of the Selected-Parameter Ratio. Table A10 varies the selected-parameter ratio ρ under the same attack and evaluation metrics. With 400 flood samples, the effective length peaks at 1910 rounds for $\rho = 0.15$ and then decreases to 1369 rounds at $\rho = 0.25$, while MA rises from 87.24% to 89.13%. At $\rho = 0.15$, increasing the flood size to 800 extends the effective length to 2104 rounds but lowers MA to 86.67%. These results indicate that the parameter budget and flood size should be selected jointly.

Table A10: Effect of the selected-parameter ratio ρ and flood size. Larger effective length and MA are better. Darker and lighter gray mark the best (**bold**) and second-best (underlined) attacked settings per metric; the no-attack row is a reference.

ρ	Flood Size	Effective Length	MA (%)
0.10	400	1557	86.81±0.53
0.15	400	1910	87.24±0.74
0.20	400	1431	88.59±0.52
0.25	400	1369	**89.13±0.30**
0.15	800	**2104**	86.67±0.70
No attack	–	–	90.18±0.40

B.8 Impact of Warm-Up Rounds

Table A11 evaluates FedHarden under Neurotoxin semantic attacks with different numbers of warm-up rounds. Longer warm-up further improves suppression, reaching 13.27% BA at 30 rounds, consistent with more thorough robustification of low-activity parameters before poisoning. FedHarden retains its advantage with short or no warm-up: even without warm-up, its 24.12% BA is below the 51.84% of the strongest baseline, BackdoorIndicator, in Table 2.

Table A11: Effect of warm-up rounds on FedHarden BA (%) under Neurotoxin semantic attacks.

Warm-Up Rounds	**BA (%)**
30	13.27±1.52
20	14.09±1.34
10	16.62±2.05
5	19.52±3.71
0	24.12±3.52

C KERNEL SELECTION PROCEDURE

Algorithm 2 details the IDENTIFYKEYKERNELS procedure used in Algorithm 1. It scores convolutional kernels using flood-plus-shadow and shadow-only gradients and selects kernels together with their associated batch-normalization parameters.

Algorithm 2: IdentifyKeyKernels

Input: Flood dataset D_f, shadow dataset D_s, parameter budget ρ
Output: Selected trainable parameters $\boldsymbol{\mathcal{P}}$

Compute convolutional-kernel gradients $\{\boldsymbol{\mathcal{K}}_f^i\}_{i=1}^J$ on $D_f \| D_s$ and $\{\boldsymbol{\mathcal{K}}_s^i\}_{i=1}^J$ on D_s
Score each kernel by $k^i = \mathrm{Avg}(|\boldsymbol{\mathcal{K}}_f^i - \boldsymbol{\mathcal{K}}_s^i|)$
$\boldsymbol{I} \leftarrow$ indices of the J kernels sorted by k^i in descending order
$\boldsymbol{\mathcal{P}} \leftarrow \varnothing,\ ind \leftarrow 0$
$N_m \leftarrow$ total number of model parameters excluding the classifier
while $|\boldsymbol{\mathcal{P}}|/N_m < \rho$ **do**
    $\boldsymbol{p} \leftarrow$ kernel with index $\boldsymbol{I}[ind]$
    $(\gamma_{\boldsymbol{p}}, \beta_{\boldsymbol{p}}) \leftarrow$ BN parameters associated with $\boldsymbol{p}$
    $\boldsymbol{\mathcal{P}} \leftarrow \boldsymbol{\mathcal{P}} \cup \{\boldsymbol{p}, \gamma_{\boldsymbol{p}}, \beta_{\boldsymbol{p}}\}$
    $ind \leftarrow ind + 1$
end