HAMMER AND ANVIL: TOWARD A THEORY OF BACKDOORS IN FEDERATED LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Federated Learning (FL) enables distributed model training but is vulnerable to backdoor attacks, where malicious clients embed attacker-controlled behaviors into the global model. Existing defenses fail against adaptive adversaries. In this paper, we present **Hammer and Anvil** (HA), a principled theoretical framework that categorizes backdoors by the deviation, δ, of their updates to the mean of the updates. We identify two fundamental defense types: **Type 1 (The Anvil)**, comprising outlier detection and robust aggregation effective against large-deviation attacks, and **Type 2 (The Hammer)**, consisting of removal-based defenses effective against small-deviation attacks. We demonstrate that defenses of a single type and non-principled combined defenses inherently leave an exploitable gap for adaptive attackers. To bridge this gap, we propose the principled combination of Type 1 and Type 2 defenses. We evaluate our framework against a new, worst-case, full-information adaptive adversary that knows the benign updates, the aggregation algorithm, and its parameters, and yet this adversary fails against our combined defenses. Our empirical evaluation across various datasets and settings shows that single-typed and non-principled combined defenses are easily broken, often by a single malicious client. In contrast, our best combined defense variants, HA^{CSFT}_{Flame}, HA^{CSFT}_{Krum}, and $HA^{CSFT}_{Multi-Metrics}$, remain undefeated even in the most adversarial settings. Our results provide a principled approach for research on backdoors in federated learning.

1 INTRODUCTION

Federated Learning McMahan et al. (2017) is a technique for efficient distributed learning in which data resides with the data sources, and only model updates are aggregated to a central model. Although this distribution of data sources enhances privacy, it does not guarantee it, and it creates opportunities for malicious clients to attack. One example of such an attack is backdooring Gu et al. (2019); Xie et al. (2019); Bagdasaryan et al. (2020); Zhang et al. (2022); Liu et al. (2024), where clients embed attacker-controlled behaviors into the model, such as a trigger that forces any image to be classified as a chosen class. Backdoor attacks have been shown to pose a significant threat in centralized machine learning Gu et al. (2019). The emergence of federated learning has introduced a new attack vector for backdoor attacks: through malicious clients, adversaries can embed backdoors into the global model with far greater control than in the centralized setting.

Many diverse defenses have been developed over the years Nguyen et al. (2022); Xu et al. (2025a); Minsker (2023); Blanchard et al. (2017) to prevent these attacks in the federated setting. These defenses can be separated into three categories depending on their underlying approach: outlier-detection-based, robust-aggregation-based, or removal-based Nguyen et al. (2024). Outlier detection defenses leverage differences between adversarial and benign updates to distinguish between them. Robust-aggregation defenses limit the impact of any update, thereby limiting the adversary's capabilities to insert a backdoor. Finally, removal defenses make minor modifications to the model in order to remove the embedded backdoor.

Crucially, we observe that different defense categories are effective against distinct types of backdoor attacks. Intuitively, backdoored model updates sent by malicious clients that deviate further from benign updates will be detected more easily by outlier detection defenses. Likewise, since robust-aggregation defenses bound the impact of any single update, beyond a certain threshold, fur-

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.

ther deviations have no impact, meaning that a small deviation attack has the same impact as a large deviation attack. Finally, removal defenses have more difficulty removing these updates with large deviations without damaging the model's benign performance, but they can easily remove attacks with small deviations.

From our observations, we reason about backdoors in terms of their expected update deviation δ to build a *principled theoretical framework* for federated learning backdoors: **Hammer and Anvil**. We categorize defenses into two types: **Type 1 defenses** (the Anvil: outlier detection and robust aggregation), which are effective when δ is large, and **Type 2 defenses** (the Hammer: removal), which are effective when δ is small. From our categorization, it follows that defenses of a single type cannot be robust without drastically affecting model utility, as they fundamentally leave a gap for an adaptive attacker to exploit. Hence, we argue for combining existing defenses using our framework. By combining non-interfering Type 1 and Type 2 defenses, we can reduce the exploitable gap for the attacker, or even eliminate it.

We ground our framework in worst-case assumptions for security to ensure secure deployment of FL in any setting. Hence, we design a new adversary threat model for federated learning: the **full-information adaptive adversary**. Our full-information adaptive adversary knows: the aggregation algorithm, its parameters, and the benign updates submitted by all other clients before it has to submit its own, for every epoch. Our adversary can freely collude with the other malicious clients it controls and can adapt to the benign updates submitted by the other clients. Mirroring cryptographic scheme design, a defense robust against our novel adaptive adversary would implicitly be robust against any attacker with more realistic assumptions.

We empirically validate our framework in two steps: (1) We design principled adaptive attacks against 13 single-typed and non-principled combined defenses Blanchard et al. (2017); Sun et al. (2019); Minsker (2023); Huang et al. (2023); Wan et al. (2025); Xu et al. (2025b); Mia & Amini (2025); Xu et al. (2025a); Fung et al. (2018); Liu et al. (2018); Nguyen et al. (2022) following our theoretical framework. Our adaptive attacks break these defenses with **at most four** malicious clients out of 20, and **as few as a single one**; (2) We evaluate our principled combined defense **HA** against our full-information adaptive adversary and existing popular and state-of-the-art attacks across various data settings. We find that our three best principled combined defense variants, namely $\mathbf{HA}^{\mathbf{CSFT}}_{\mathbf{Flame}}$, $\mathbf{HA}^{\mathbf{CSFT}}_{\mathbf{Krum}}$, and $\mathbf{HA}^{\mathbf{CSFT}}_{\mathbf{Multi\text{-}Metrics}}$, **are undefeated in every setting we evaluate**. Our empirical results support our theory and enable principled research on backdoor attacks in federated learning.

We summarize our contributions as follows:

- We present a new principled theory of backdoors in federated learning: **Hammer and Anvil** in Section 3. Our theory categorizes defenses into two types: **Type 1** (Anvil) and **Type 2** (Hammer), and reveals a fundamental exploitable gap in single-typed defenses and for non-principled combined defenses. We then derive rules for the principled combination of backdoor defenses.
- We introduce a new worst-case threat model: **the full-information adaptive adversary** in Section 2.1 and use it to break single-typed and non-principled combined defenses in Section 4.
- We empirically evaluate our principled combined defense **HA** in Section 5. We show that our three best combined defense variants are undefeated in the worst-case threat model, showcasing state-of-the-art performance. Our empirical results provide support for our theory.

2 PRELIMINARIES

Federated Learning (FL) McMahan et al. (2017) collaboratively trains a global model w by aggregating local updates from n distributed clients. In a backdoor attack Gu et al. (2019), an adversary controlling a subset of m malicious clients injects triggered data during local training. The attacker's objective is a multi-objective optimization problem: maximize the Attack Success Rate (ASR) on triggered inputs while maintaining high accuracy on the main task.

To mitigate these threats, the aggregator deploys defenses Blanchard et al. (2017); Nguyen et al. (2022) that broadly fall into three categories: **Robust-aggregation** (bounding the impact of any single update), **Outlier Detection** (filtering updates based on distance or similarity metrics), and **Removal** (post-aggregation model refinement). As we explore in Section 3, the efficacy of these approaches is fundamentally tied to the magnitude of the adversarial update. A comprehensive formalization of the FL pipeline, attack objectives, and defense taxonomy is provided in Appendix A.

Related Work The study of backdoors in FL has produced a broad spectrum of attacks Xie et al. (2019); Zhang et al. (2022); Bagdasaryan et al. (2020); Krauß & Dmitrienko (2023), defenses Nguyen et al. (2022); Xu et al. (2025a); Minsker (2023); Blanchard et al. (2017), and comprehensive surveys Nguyen et al. (2024); Li et al. (2025). However, most existing works evaluate adversaries with limited capabilities Zhang et al. (2022); Szelag et al. (2025), providing estimates of robustness against an ***average*** threat rather than a ***worst-case*** adaptive adversary. While recent adaptive attacks leveraging multi-objective optimization and dynamic adaptation have proven devastating against state-of-the-art defenses Szelag et al. (2025); Krauß & Dmitrienko (2023); Krauß et al. (2024); Shejwalkar & Houmansadr (2021), they primarily target single-typed defenses. Few defenses, such as MESAS Krauß & Dmitrienko (2023), specifically address adaptive adversaries, but they do not account for the full-information threat model we propose. We provide a detailed discussion of these works in Appendix B.

Notation: Let $w \in \mathbb{R}^k$ represent the set of weights for a model. Let b and d denote datasets, while B and D represent distributions. The function T is a training function that produces a set of weight updates u based on a given set of weights and a dataset. Let n be the number of clients and m the number of malicious clients.

Clipped-Super-Fine-Tuning. We adapt Super-Fine-Tuning (SFT) Sha et al. (2022) for federated learning by introducing **Clipped-Super-Fine-Tuning (CSFT)**. Because the high initial learning rates of vanilla SFT consistently cause gradient explosion in federated models, CSFT incorporates gradient clipping to stabilize the removal of weakly inserted backdoors (small deviations). Crucially, the optimal clipping threshold depends primarily on the fine-tuning sample size, effectively eliminating the need for exhaustive hyperparameter tuning. Further details are provided in Appendix A.4.

2.1 Motivation & Threat Model

Achieving trustworthy and secure FL requires a principled approach grounded in worst-case assumptions. Mirroring the design of cryptographic schemes, we assume a nigh-omniscient attacker to establish a blueprint for security; a defense robust against this adversary will inherently withstand weaker, more realistic threats. We model an adaptive adversary that controls a subset of m colluding malicious clients, aiming to manipulate the FL training process via malicious updates. To evaluate defenses under worst-case conditions, we grant this adversary full information: exact knowledge of the aggregation algorithm, its specific parameters, and the benign updates submitted by all other clients in each epoch. We operate under the standard assumption that the attacker controls only a minority of clients ($m < \frac{n}{2}$) and must maintain the model's benign accuracy to avoid detection. For the defender, we assume the aggregator has access to a small, benign dataset Xu et al. (2025b) for fine-tuning (e.g., 1%-4% of total training data), keeping this set strictly smaller than an average client's local dataset to maintain the necessity of the federated architecture. Any viable defense must successfully remove the backdoor without degrading benign performance. In practice, we set $m < 9$ for $n = 20$. More details are available in Appendix A.5. In Section 5, we demonstrate that our principled defense approach can reliably defeat this full-information, adaptive adversary.

3 Hammer and Anvil: Toward a Principled Theory

Each backdoor attack can be categorized according to a set of parameters, including its impact on the model weights, the fraction of malicious clients, its duration, etc. We consider the impact on model weights to be particularly important in the design of our theory. Let δ be the expected deviation from

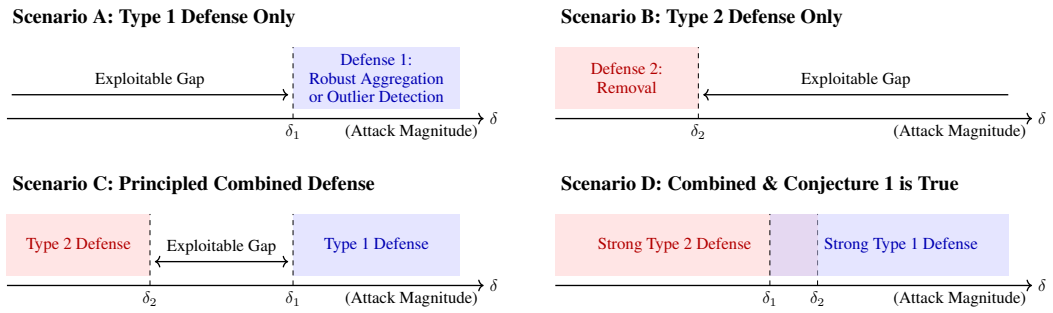


Figure 1: Progression of the defensive spectrum against varying attack magnitudes (δ). Scenario A and Scenario B highlight the exploitable gaps when relying on a single defense mechanism. Scenario C illustrates minimizing the security gap by combining mechanisms. Finally, Scenario D demonstrates Conjecture 1: an ideal state where the defenses overlap ($\delta_1 \leq \delta_2$), completely eliminating the exploitable gap.

benign updates

$$\delta = \|\mathbb{E}[\frac{1}{n-m}\sum_{i=1}^{n-m} T(w_j, d_i)] - \mathbb{E}[\frac{1}{m}\sum_{i=n-m+1}^{n} T(w_j, d_i \cup b_i)]\|_2 \tag{1}$$

for any given federated epoch j, we have n clients and each client has a training dataset $d_i \sim D$, and we assume there are $m < \frac{n}{2}$ malicious clients with datasets $b_i \sim B$ drawn from the distribution of backdoored samples B. δ represents the expected l_2-distance from benign updates (or strength) of a given backdoor attack update, with smaller l_2-norm deviations having a small δ, and larger l_2-norm deviations having a larger δ. δ is a theoretical tool to classify backdoor attacks and defenses. Another such tool is Δ, which measures the expected impact of the backdoor.

$$\Delta = m\delta \tag{2}$$

Our principled combined defense is premised on the observation that the success of existing defenses depends on the attack parameter δ and the attack's impact Δ. We define two fundamental types of defenses based on their success with respect to δ:

Type 1 - Robust Aggregation or Detection: The goal of the aggregator is to discard malicious updates and update the global model only with benign updates, or at least limit the impact of malicious updates on the aggregation process. To do so, the aggregator attempts to identify outliers (which are likely malicious updates, since $m < \frac{n}{2}$) from the client updates $\{T(w_j, d_i)\}_{i=1}^{n-m} \cup \{T(w_j, d_i \cup b_i)\}_{i=n-m+1}^{n}$. It then dismisses those outliers, e.g., Krum and Multi-Metrics, or bounds their impact, e.g., Median-of-Means and norm bounding.

Type 2 - Removal: Assume the aggregator has a small, benign dataset $f \sim D$ and fine-tunes the global model w after aggregation as follows:

$$w_\tau = (1-\gamma)\, w + \gamma\, T(w, f) \tag{3}$$

for some $0 \leq \gamma \leq 1$.

Using the distance δ, we can make the following propositions about existing defenses:

Proposition 1 *The larger δ, the more robust Type 1 defenses are.*

Proposition 1 follows from two observations: the larger δ is, the further in the l_2-norm the malicious updates are from the benign updates. Since $m < \frac{n}{2}$, the benign updates outnumber the malicious updates. As such, the further the malicious updates are from the benign updates, the more likely they are to be detected as outliers by outlier detection defenses. Therefore, a larger δ results in a lower ASR relative to the equivalent undefended global model. Robust aggregation defenses impose a limit on the influence any one update can have on the aggregation process. Since the defense's

limit naturally bounds δ, the larger the δ, the better the defense relative to the equivalent undefended global model. Thus, it follows that a large δ improves the effectiveness of Type 1 defenses, and Corollary 1 and Corollary 2 also follow.

Corollary 1 *For all Type 1 defenses, there exists δ_1 such that it will prevent the backdoor for any $\delta > \delta_1$.*

Corollary 2 *For all Type 1 defenses that admit learning (i.e., at least one update), there exists δ_1' such that a backdoor update with $\delta < \delta_1'$ will pass the defense.*

Proposition 2 *Let $\varepsilon \in \mathbb{R}^k$ be a matrix of small random values, i.e $\|\varepsilon\|_2 \ll \|w\|_2$. The smaller δ, the more the global model approaches $w_\tau = T(\cdot, f) + \varepsilon$ under a Type 2 defense, i.e., a model trained on a dataset drawn only from D.*

Proposition 2 follows from the iterative learning process of neural networks that slowly converge after many epochs. Intuitively, a Type 2 defense can remove the backdoor because f is drawn from the benign distribution D, thereby shifting the global model weights towards those of a model trained only on D. The larger the change introduced by the backdoor, the more backpropagation steps are needed to remove it; consequently, the defender faces a higher risk of benign accuracy degradation. Hence, Type 2 defenses are most effective against weakly inserted backdoors, where δ is small enough to be removed without overwriting the model with one trained solely on the fine-tuning data.

Corollary 3 *For all Type 2 defenses, there exists a δ_2 such that it will remove the backdoor for any $\delta < \delta_2$.*

Corollary 4 *For all Type 2 defenses, there exists a δ_2' such that a backdoor update with $\delta > \delta_2'$ will not be removed.*

Given either only a Type 1 or Type 2 defense, it follows from Corollary 2 and Corollary 4 that an attacker should be able to succeed. We empirically demonstrate such attacks against existing defenses in Section 4 using optimization, scaling, or interpolation to adapt to existing defenses.

When combining a Type 1 defense and a Type 2 defense, it follows from Corollary 1 and Corollary 3 that an attacker can only succeed for $\delta \in [\delta_2, \delta_1]$. Figure 1 illustrates this defensive spectrum; while each category effectively secures one end of the magnitude range, a gap remains where neither succeeds.

Research Question 1 (RQ1): *Can we reliably identify a successful principled attack given any single-typed or non-principled combined defense?*

Research Question 2 (RQ2): *When combining Type 1 and 2 defenses, can $\delta_1 - \delta_2$ be non-positive, meaning that even worst-case adaptive adversaries cannot succeed?*

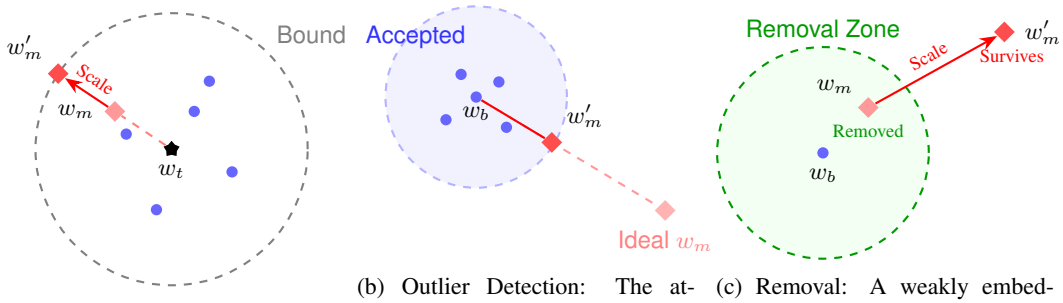
We hypothesize that with the correct defenses, the gap in our second research question can be eliminated entirely.

Conjecture 1 *There exists a Type 1 defense and a Type 2 defense such that $\delta_2 \geq \delta_1$.*

If Conjecture 1 holds, it implies the existence of a principled defense that can defend against any backdoor for any attack deviation δ.

4 Exploiting the Gap in Existing Defenses

Our goal in this section is to provide empirical evidence to support Corollary 2 and Corollary 4 by showing that for all existing defenses we evaluate, we can define an adaptive attack that easily breaks the defense. To evaluate the limits of current defenses, we develop a suite of adaptive attacks designed to exploit the fundamental weaknesses of each defense category. By leveraging the full



(a) Robust Aggregation: The attacker scales the initial malicious weights (w_m) to rest exactly on the defense's acceptance bound, maximizing impact.

(b) Outlier Detection: The attacker interpolates between the benign mean (w_b) and the ideal malicious weight, submitting the update for weights (w'_m) that just cross into the accepted cluster.

(c) Removal: A weakly embedded backdoor (w_m) falls within the effective removal zone and is removed. By scaling the update (w'_m) past this zone, the backdoor survives removal attempts.

Figure 2: A simplified representation of single-typed defenses and their associated adaptive attack principles. Blue circles represent benign weights, while red shapes denote adaptive malicious weights.

information capabilities of our adaptive adversary, these attacks bypass singular and non-principled defenses. In the following, we outline the general principles for circumventing each category. We detail the adaptive attack implementation for each defense in Appendix C. We display simplified representations of the attack principles for each defense category in Figure 2.

4.1 Adaptive Attack Principles

Robust-Aggregation Defenses. Robust aggregation defenses aim to bound the impact of any individual update on the global model. Our strategy exploits this by ensuring the malicious updates reach the maximum allowable impact bound, as shown in Figure 2a. We achieve this by scaling the magnitude of the adversarial updates relative to the mean, either through direct weight manipulation or by increasing the local learning rate of malicious clients.

Outlier Detection Defenses. Outlier detection defenses use metrics such as l_2-distance, l_1-distance, or cosine similarity to isolate and discard malicious updates. We formulate our attacks as a multi-objective optimization problem, maximizing the backdoor signal while minimizing the defense's detection metric. Where possible, we solve this optimization directly using interpolation, as shown in Figure 2b; otherwise, we employ gradient-based techniques by incorporating the defense's scoring function into the malicious clients' loss objective, or a surrogate scoring function if the defense's scoring function is not differentiable.

Removal Defenses. Removal and post-processing defenses are constrained by the need to preserve the model's benign accuracy. If the defense applies a transformation that is too significant to the model weights, it risks destroying the utility of the global model. We exploit this constraint by scaling the adversarial updates throughout training, ensuring that the modification required to remove the backdoor exceeds the threshold of safe removal for the defense, as shown in Figure 2c.

Non-Principled Combined Defenses. Several works have attempted to construct robust defenses by assembling multiple defense mechanisms into a single pipeline, such as FLAME Nguyen et al. (2022), Fine-tune & Prune Wu et al. (2020), and FoolsGold & Krum Fung et al. (2018). While these multi-layered approaches often outperform individual defenses, the lack of a grounding principle behind their combination leaves them vulnerable to adaptive adversaries. For example, fine-tuning and pruning are both removal-based, which prevent small l_2-norm attacks but leave them vulnerable to large l_2-norm attacks; FLAME combines HDBSCAN angle-based clustering, a defense against large angular magnitude attacks, with clipping, a defense against large l_2-norm attacks, and DP-noising, a defense against attacks involving a small number of malicious clients. Their combined defense leaves them vulnerable to small l_2-norm attacks carried out by many attackers.

Table 1: Summary of defenses and corresponding attack strategies with the relevant equations. Eq. stands for Equation.

Defense Type	Defense Name	Attack Strategy	Eq.
Type 1 (Robust Aggregation)	Norm-bounding Sun et al. (2019)	Scaling	5
	Median-of-Means Minsker (2023)	Scaling	6
Type 1 (Outlier Detection)	Krum Blanchard et al. (2017)	Interpolation	7
	Multi-Metrics Huang et al. (2023)	Optimization	8
	MARS Wan et al. (2025)	Optimization	9
	MASA Xu et al. (2025b)	Interpolation	7
	BART-FL Mia & Amini (2025)	Optimization	10
	AlignIns Xu et al. (2025a)	Optimization	11, 12
Type 2	CSFT (This work)	Scaling	6
Non-Principled Combined	FoolsGold & Krum Fung et al. (2018)	Interpolation	7
	Fine-tune & Prune Wu et al. (2020)	Scaling	5 or 6
	FLAME Nguyen et al. (2022)	Optimization & Interpolation	13

Table 1 contains a summary of the defenses for which we devise adaptive attacks separated by type, with the associated attack strategy and a reference to the relevant equations. We present detailed attacks against robust-aggregation defenses in Appendix C.1, outlier detection defenses in Appendix C.2, removal defenses in Appendix C.3, and non-principled combined defenses in Appendix C.4. We also include empirical results demonstrating the success of our adaptive attacks in Section 4.2.

4.2 Motivating Results: Non-Principled Defenses are Exploitable

We demonstrate that existing individual backdoor defenses are insufficient under our defined worst-case threat model. Table 2 summarizes the vulnerability of **seven** state-of-the-art or popular defenses Nguyen et al. (2022); Blanchard et al. (2017); Minsker (2023); Sun et al. (2019); Sha et al. (2022) against our adaptive, optimized attacks on a ResNet18 He et al. (2016) model trained on the CIFAR-10 Krizhevsky (2009) dataset.

We break most defenses with as few as 1 or 2 malicious clients out of 20. While some defenses show higher initial resistance, specifically, Multi-Metrics requires $m = 4$ to break, they all eventually fail as the number of malicious clients increases, reaching near 100% ASR in the worst-case scenario where $m = 8$, while maintaining high benign accuracy. These failures stem from a fundamental vulnerability in current defense approaches. Most existing methods fall into a single category that targets only small or large attack magnitudes, leaving them vulnerable to the other. Furthermore, combined defenses often merge sub-defenses of the same type, leaving a vulnerable gap that a full-information adaptive adversary can exploit.

As we already break one of the components of FoolsGold & Krum Fung et al. (2018) and Fine-tune & Prune Wu et al. (2020) in a way that the other component cannot cope with, we expect these two defenses to perform as well as their broken components. As for MARS Wan et al. (2025), MASA Xu et al. (2025b), BART-FL Mia & Amini (2025), and AlignIns Xu et al. (2025a): we leave the experimental effort to future work, as we already empirically evaluate two outlier detection defenses (Krum and Multi-Metrics). These four defenses reuse similar fundamental principles, so we expect the principled adaptive attacks we design against them to be as effective as those against Krum and Multi-Metrics.

5 HA: A Principled Combined Defense

We mitigate the gaps left by existing defenses with our principled defense approach: Hammer and Anvil (HA). HA combines defenses based on their type, such that no defense's vulnerability is exploitable. While Type 1 defenses are well-represented by robust aggregators (e.g., Median-of-Means Minsker (2023)), outlier detection algorithms (e.g., Krum Blanchard et al. (2017), Multi-Metrics Huang et al. (2023)), or hybrid pipelines like FLAME Nguyen et al. (2022), there are few

Table 2: Lowest number of malicious clients (m) out of 20 needed for our adaptive attacks to break existing defenses with the associated ASR and benign accuracies on CIFAR-10. We also include the $m = 8$ case. Values are averaged over three runs.

Defense	m	ASR	Accuracy	ASR ($m = 8$)	Accuracy ($m = 8$)
Multi Metrics	**4**	98.86 ± 0.79	76.16 ± 5.73	99.30 ± 0.36	78.61 ± 2.40
Flame	2	55.61 ± 5.24	82.16 ± 0.11	98.25 ± 1.46	80.06 ± 0.36
Rob Mom	2	99.81 ± 0.06	83.80 ± 0.31	99.98 ± 0.01	84.56 ± 0.07
Mom	2	99.81 ± 0.06	83.49 ± 0.18	99.99 ± 0.01	84.28 ± 0.16
Krum	1	97.34 ± 1.00	78.85 ± 1.70	99.98 ± 0.02	79.90 ± 1.01
Norm	1	90.49 ± 2.99	87.93 ± 0.04	99.97 ± 0.03	86.30 ± 1.18
CSFT	1	87.47 ± 7.95	79.01 ± 0.35	98.70 ± 0.70	82.15 ± 0.34
None	1	99.62 ± 0.38	83.02 ± 0.04	99.99 ± 0.01	84.64 ± 0.40

Type 2 defenses, and even fewer tailored for the unique constraints of FL. By instantiating our Type 2 defense as CSFT, as described in Section 2, we can systematically pair it with the Type 1 defenses evaluated in Table 2, namely FLAME Nguyen et al. (2022), (Rob) MoM Minsker (2023), Krum Blanchard et al. (2017), Multi-Metrics Huang et al. (2023), and Norm-bounding Sun et al. (2019), yielding: $\mathbf{HA}^{\mathbf{CSFT}}_{\mathbf{Flame}}$, $\mathbf{HA}^{\mathbf{CSFT}}_{\mathbf{(Rob)\ MoM}}$, $\mathbf{HA}^{\mathbf{CSFT}}_{\mathbf{Krum}}$, $\mathbf{HA}^{\mathbf{CSFT}}_{\mathbf{Multi\text{-}Metrics}}$, and $\mathbf{HA}^{\mathbf{CSFT}}_{\mathbf{Norm}}$. The resulting principled combined defenses support our theory. They offer empirical evidence for Corollary 1, Corollary 3, and Conjecture 1.

Proposition 3 *A full-information adaptive attack that is optimal against a Type 1 defense A is also optimal against HA^B_A, for any Type 2 defense B.*

Proposition 3 follows from Corollary 1: for an adaptive attack to be optimal against a Type 1 defense, it needs to maximize $\delta < \delta_1$ to maximize its impact on the global model; and it follows from Corollary 3: for an adaptive attack to be optimal against a Type 2 defense, it needs to maximize δ to maximize its impact on the global model such that $\delta > \delta_2$. Consequently, if there exists a maximum reachable δ such that $\delta_2 < \delta < \delta_1$, an optimal adaptive attack against the Type 1 defense will find it. Hence, if our full-information adaptive attacks are optimal against Type 1 defenses, they must also be optimal against HA.

We demonstrate empirically that Hammer and Anvil is effective at combining compatible defenses across a wide range of scenarios. In Section 5.1, we provide empirical evidence supporting Conjecture 1. We show that even in the worst-case scenario: our full-information adaptive adversary with $m = 8$ out of $n = 20$, our **three** best principled combined defenses: $\mathbf{HA}^{\mathbf{CSFT}}_{\mathbf{Flame}}$, $\mathbf{HA}^{\mathbf{CSFT}}_{\mathbf{Krum}}$, and $\mathbf{HA}^{\mathbf{CSFT}}_{\mathbf{Multi\text{-}Metrics}}$ are undefeated.

5.1 Evaluating HA: Empirical Evidence Supporting Conjecture 1

We present the evaluation of our combined HA architectures against our full-information adaptive adversaries on CIFAR-10 in Table 3. The results demonstrate that our principled approach successfully closes the exploitable gap in **three** configurations ($\text{HA}^{\text{CSFT}}_{\text{Flame}}$, $\text{HA}^{\text{CSFT}}_{\text{Krum}}$, and $\text{HA}^{\text{CSFT}}_{\text{Multi-Metrics}}$), rendering the adaptive attacker entirely unsuccessful even in the worst-case scenario ($m = 8$). In another configuration, the gap is significantly narrowed ($\text{HA}^{\text{CSFT}}_{\text{Norm}}$ now requires $m = 4$, up from $m = 1$). Conversely, pairing CSFT with MoM or Rob-MoM yields little improvement, indicating that the success of our approach ultimately relies on selecting a Type 1 defense restrictive enough to overlap with CSFT's effective δ-range. Ultimately, these findings provide supporting empirical evidence that **Conjecture 1 holds true**: with HA , ***it is possible to defend against a worst-case, full-information, adaptive adversary with limited degradation to benign accuracy***.

To support Conjecture 1 with further empirical evidence, we experiment with a non-IID data distribution of CIFAR-10 in Appendix D.3, the MNIST dataset LeCun (1998) in Appendix D.2.1, and the ImageNet dataset Deng et al. (2009) in Appendix D.2.2. The non-IID data setting is common in the FL security literature, as it represents a more realistic deployment scenario Nguyen et al. (2022); Xu et al. (2025a); Wan et al. (2025); Xu et al. (2025b); Huang et al. (2023). The MNIST dataset is popular for extending experiments to other datasets Nguyen et al. (2022); Xu et al. (2025a); Wan

Table 3: Minimum malicious clients ($m, n = 20$) required by our adaptive attacks to break HA **after CSFT** on CIFAR-10. '>8' denotes an unbroken defense. Parentheses indicate m for the original, uncombined defense. Averaged over 3 runs.

Defense	m	ASR	Accuracy
$\text{HA}^{\text{CSFT}}_{\text{Flame}}$	**>8** (2)	**13.66 ± 7.08**	74.86 ± 0.72
$\text{HA}^{\text{CSFT}}_{\text{Krum}}$	**>8** (1)	26.87 ± 1.99	76.03 ± 0.34
$\text{HA}^{\text{CSFT}}_{\text{Multi-Metrics}}$	**>8** (4)	35.51 ± 4.67	79.56 ± 0.34
$\text{HA}^{\text{CSFT}}_{\text{Norm}}$	4 (1)	67.86 ± 14.86	**83.50 ± 0.31**
$\text{HA}^{\text{CSFT}}_{\text{Mom}}$	2 (2)	84.73 ± 2.65	79.11 ± 0.22
$\text{HA}^{\text{CSFT}}_{\text{Rob-Mom}}$	2 (2)	94.26 ± 1.33	81.18 ± 0.13

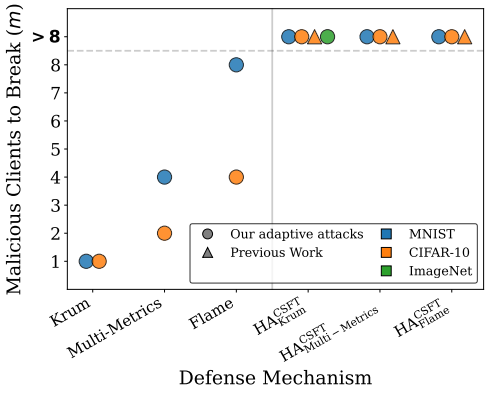


Figure 3: Minimum malicious clients (m) out of 20 required to break each defense. Our HA variants withstand all evaluated attacks up to $m = 8$ across all settings, outperforming the baselines.

et al. (2025) and the ImageNet dataset is useful to show how results scale to large input dimensionalities and larger models. We focus exclusively on the best combined defenses: $\text{HA}^{\text{CSFT}}_{\text{Flame}}$, $\text{HA}^{\text{CSFT}}_{\text{Krum}}$, and $\text{HA}^{\text{CSFT}}_{\text{Multi-Metrics}}$. We also evaluate HA against existing popular and state-of-the-art non-adaptive attacks in Appendix D.1. We demonstrate that they are completely ineffective against HA, even in settings that heavily favor the adversary. Figure 3 summarily showcases HA's success compared to the individual defenses in these other settings.

Additionally, we present empirical studies on the effects of hyperparameters on HA in Appendix D.4, specifically regarding CSFT, to measure ease of use in real-world scenarios. We measure the effect of the size of the fine-tuning dataset and the number of fine-tuning epochs on the performance of $\text{HA}^{\text{CSFT}}_{\text{Flame}}$, $\text{HA}^{\text{CSFT}}_{\text{Krum}}$, and $\text{HA}^{\text{CSFT}}_{\text{Multi-Metrics}}$ on CIFAR-10.

6 Conclusion

In this work, we propose **Hammer and Anvil**, a new theory of backdoors in FL based on the expected update deviation δ of backdoor attacks. Our theory categorizes defenses into two types: the **Hammer** (Type 2) and the **Anvil** (Type 1), based on whether they defend against small or large deviation attacks. From our theory, we derive that single-typed defenses and non-principled combined defenses are exploitable. We empirically validate this hypothesis under a new worst-case threat model: **the full-information adaptive adversary**, and we show that existing defenses are easily broken, answering our first research question (RQ1).

Our theory dictates the principled combination of defenses: combining Type 1 and Type 2 defenses to yield a robust principled combined defense, **HA**. We introduce a new Type 2 defense: **Clipped-Super-Fine-Tuning (CSFT)**, and show that when combined with existing Type 1 defenses, it can even defend against our worst-case full-information adversary across a variety of data settings. We conjecture that combining the right Type 1 and Type 2 defenses yields a robust defense against all backdoor attacks, and we provide empirical evidence to support this claim (tentatively answering RQ2). Our best combined defenses, $\text{HA}^{\text{CSFT}}_{\text{Flame}}$, $\text{HA}^{\text{CSFT}}_{\text{Multi-Metrics}}$, and $\text{HA}^{\text{CSFT}}_{\text{Krum}}$, are robust against both our novel adversary and existing state-of-the-art attacks across every settings we evaluated.

Extending our theory to other fields of machine learning security, such as adversarial robustness or watermarking, could introduce new research avenues. We leave some aspects of adaptive attack implementation and evaluation for future work. Additionally, future research into better Type 1 and Type 2 defenses would directly benefit our work, yielding even more robust combined defenses. Finally, reproducing our results across other data modalities, such as text or audio, could help further the understanding of federated learning backdoor attacks and their defenses.

AI USE STATEMENT

In this work, we did not use generative AI tools for any of the tasks with required disclosure. We used generative AI tools for: creating figures and tables, to create or edit software code, and to edit the paper to improve readability. We have reviewed all AI-assisted work. LLM-generated code was verified and tested for correctness. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Our paper considers backdoor attacks and defenses against federated systems. While we propose a new, stronger attack, we also propose a defense that prevents it, thereby mitigating its impact. Our insights can also help further research into defenses against backdoor attacks on federated systems. Our datasets, models, defenses, and attacks are public and do not contain sensitive or private information. Our work did not use human subjects in any way.

REPRODUCIBILITY STATEMENT

To reproduce our results, the following are necessary:

- The publicly available CIFAR-10 Krizhevsky (2009), MNIST LeCun (1998), and ImageNet Deng et al. (2009) datasets.
- Our codebase, available at the following anonymous link: `https://anonymous.4open.science/r/Hammer-and-Anvil-7064`. It contains detailed instructions and code to reproduce our results, tables, and figures.

Additionally, we provide detailed descriptions of our algorithms in both the main paper and the appendices.

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.

Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, and Julien Stainer. Machine learning with adversaries: Byzantine tolerant gradient descent. *Advances in neural information processing systems*, 30, 2017.

John S Bridle. Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition. In *Neurocomputing: Algorithms, architectures and applications*, pp. 227–236. Springer, 1990.

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, June 2009. doi: 10.1109/CVPR.2009.5206848.

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

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 (CVPR)*, June 2016.

Siquan Huang, Yijiang Li, Chong Chen, Leyu Shi, and Ying Gao. Multi-metrics adaptively identifies backdoors in federated learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4652–4662, 2023.

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.

Torsten Krauß, Jan König, Alexandra Dmitrienko, and Christian Kanzow. Automatic adversarial adaption for stealthy poisoning attacks in federated learning. In *To appear soon at the Network and Distributed System Security Symposium (NDSS)*, 2024.

Alex Krizhevsky. Learning multiple layers of features from tiny images, 2009. In *Technical report*.

Yann LeCun. The mnist database of handwritten digits, 1998. In *Technical report*.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 2002.

Zhaozheng Li, Jiahe Lan, Zheng Yan, and Erol Gelenbe. Backdoor attacks and defense mechanisms in federated learning: A survey. *Information Fusion*, pp. 103248, 2025.

Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Fine-pruning: Defending against backdooring attacks on deep neural networks. In *International symposium on research in attacks, intrusions, and defenses*, pp. 273–294. Springer, 2018.

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 *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 21359–21367, 2024.

Hidde Lycklama, Lukas Burkhalter, Alexander Viand, Nicolas Küchler, and Anwar Hithnawi. Rofl: Robustness of secure federated learning. In *2023 IEEE Symposium on Security and Privacy (SP)*, pp. 453–476. IEEE, 2023.

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.

Md Jueal Mia and M Hadi Amini. Bart-fl: A backdoor attack-resilient federated aggregation technique for cross-silo applications. *IEEE Transactions on Machine Learning in Communications and Networking*, 2025.

Stanislav Minsker. Efficient median of means estimator. In *The Thirty Sixth Annual Conference on Learning Theory*, pp. 5925–5933. PMLR, 2023.

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.

Thuy Dung Nguyen, Tuan Nguyen, Phi Le Nguyen, Hieu H Pham, Khoa D Doan, and Kok-Seng Wong. Backdoor attacks and defenses in federated learning: Survey, challenges and future research directions. *Engineering Applications of Artificial Intelligence*, 127:107166, 2024.

Zeyang Sha, Xinlei He, Pascal Berrang, Mathias Humbert, and Yang Zhang. Fine-tuning is all you need to mitigate backdoor attacks. *arXiv preprint arXiv:2212.09067*, 2022.

Virat Shejwalkar and Amir Houmansadr. Manipulating the byzantine: Optimizing model poisoning attacks and defenses for federated learning. In *NDSS*, 2021.

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

Jakub Kacper Szelag, Ji-Jian Chin, and Sook-Chin Yip. Adaptive adversaries in byzantine-robust federated learning: A survey. *Cryptology ePrint Archive*, 2025.

Wei Wan, Yuxuan Ning, Zhicong Huang, Cheng Hong, Shengshan Hu, Ziqi Zhou, Yechao Zhang, Tianqing Zhu, Wanlei Zhou, and Leo Yu Zhang. Mars: A malignity-aware backdoor defense in federated learning. *arXiv preprint arXiv:2509.20383*, 2025.

Ning Wang, Yang Xiao, Yimin Chen, Yang Hu, Wenjing Lou, and Y Thomas Hou. Flare: defending federated learning against model poisoning attacks via latent space representations. In *Proceedings of the 2022 ACM on Asia Conference on Computer and Communications Security*, pp. 946–958, 2022.

Chen Wu, Xian Yang, Sencun Zhu, and Prasenjit Mitra. Mitigating backdoor attacks in federated learning. *arXiv preprint arXiv:2011.01767*, 2020.

Chulin Xie, Keli Huang, Pin-Yu Chen, and Bo Li. Dba: Distributed backdoor attacks against federated learning. In *International conference on learning representations*, 2019.

Jiahao Xu, Zikai Zhang, and Rui Hu. Detecting backdoor attacks in federated learning via direction alignment inspection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 20654–20664, 2025a.

Jiahao Xu, Zikai Zhang, and Rui Hu. Identify backdoored model in federated learning via individual unlearning. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 7960–7969. IEEE, 2025b.

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.

APPENDIX

A BACKGROUND

A.1 FEDERATED LEARNING

Federated Learning (FL) McMahan et al. (2017) is a distributed machine learning approach where multiple clients collaboratively train a global model without centralizing their local datasets. Formally, n clients collaborate to train a shared model by aggregating updates computed locally on distributed datasets ($d_i \sim D, i \in \{1, \ldots, n\}$). These updates are aggregated by the aggregator to create the global model w. We detail the training process for any epoch t below; this series of steps is repeated until convergence or a predefined number of epochs (t_{max}) is reached.

1. **(Initialization Phase)** The aggregator sends the global model w_t to a subset of clients of size s.
2. **(Training Phase)** Each client in the subset trains its copy of the global model on its dataset d_i to generate its update u_i. The clients then send their update u_i to the aggregator.
3. **(Aggregation Phase)** The aggregator aggregates the clients' updates, yielding the new global model w_{t+1}.

The standard aggregation rule is averaging (FedAvg McMahan et al. (2017)), given by: $w_{t+1} = w_t + \frac{1}{s}\sum_i u_i$.

A.2 BACKDOOR ATTACKS

Backdoor attacks embed hidden attacker-controlled behaviors within machine learning models during training Gu et al. (2019). Specifically, the attacker aims to alter the model's training such that the resulting model behaves normally on benign inputs but produces attacker-controlled outputs when triggered. To trigger the attacker-controlled behavior, the attacker can insert or exploit a pattern in the training data Gu et al. (2019), which is learned by the victim model. Depending on the type of trigger, backdoor attacks can be categorized as artificial or semantic Nguyen et al. (2024).
Artificial Backdoors: The triggers are artificially added, typically as a distinct pattern or patch (e.g., a white square in the corner of the image).
Semantic Backdoors: The triggers exploit natural features of inputs, associating specific features with attacker-controlled behavior (e.g., green cars).

Attack Objectives. The attacker's objective is to maximize the Attack Success Rate (ASR), as in the proportion of backdoored inputs that trigger attacker-controlled behavior, while maintaining accuracy on the model's main task, which we refer to as benign accuracy or accuracy. Backdoor attacks can be modeled as a multi-objective optimization problem, where the attacker seeks to optimize the model's behavior across both benign and backdoored inputs. For a dataset d of input-label pairs drawn from a benign distribution D, a backdoor dataset b drawn from a backdoored distribution B, and a loss function L (for example, Cross-Entropy Bridle (1990) for image classification), the attacker aims to minimize:

$$\min \sum_{(x_i,y_i)\in d} L(x_i, y_i) + \sum_{(x'_i,y'_i)\in b} L(x'_i, y'_i) \tag{4}$$

A.3 Backdoor Defenses

In scenarios where backdoor attacks pose a threat to the integrity of an FL model, the aggregator may deploy a defense Nguyen et al. (2022); Xu et al. (2025a); Minsker (2023); Blanchard et al. (2017) . When deploying a defense, the aggregator must balance two objectives: backdoor prevention/removal and benign accuracy preservation. Different approaches have been devised over the years and can be categorized as follows:

Robust-aggregation: limit the impact of any update, thereby limiting the impact of adversarial updates and the adversary's capabilities to insert a backdoor. Targets large deviation magnitudes.
Outlier Detection: leverage the differences between adversarial and benign updates to distinguish them. Also targets large deviation magnitudes.
Removal: make minor modifications to the model in order to remove the embedded backdoor. Targets small deviation magnitudes.

Some existing works combine multiple approaches to create more effective defenses; for example, FLAME Nguyen et al. (2022) first detects malicious updates using cosine-similarity-based clustering, then limits the impact of selected updates by clipping their norms to the median norm and adding noise. We study FLAME and other existing defenses, and their weaknesses in more detail in Section 4.

An important consideration is the location in the FL pipeline where a defense operates: pre-aggregation, in-aggregation, or post-aggregation Nguyen et al. (2024), with detection defenses often operating pre-aggregation, robust-aggregation defenses operating in-aggregation, and removal defenses operating post-aggregation.

A.4 Clipped-Super-Fine-Tuning

We adapt **Super-Fine-Tuning (SFT)** Sha et al. (2022), a removal defense designed for centralized training, to the federated setting. SFT removes backdoors by fine-tuning the global model on a small benign dataset using a specialized learning rate schedule, as illustrated in Figure 4. However, we find that vanilla SFT performs poorly in FL; the high learning rate required during the first phase often leads to gradient explosion, destroying the model weights. This behavior suggests either that the SFT algorithm is highly sensitive to hyperparameters, or that the loss landscape of models trained in a federated setting is fundamentally different from that of models trained centrally and does not respond to SFT in the same way.

To mitigate this instability, we introduce **Clipped-Super-Fine-Tuning (CSFT)**, which incorporates gradient clipping into the SFT process. With an appropriately tuned clipping threshold, CSFT effectively removes "weakly inserted" backdoors (small devations), as demonstrated in Figure 5, where CSFT successfully removes a non-adaptive Badnet Gu et al. (2019) backdoor implanted by a single malicious attacker out of 20 in a ResNet18 He et al. (2016) on the CIFAR-10 dataset Krizhevsky (2009). Notably, we find that the optimal clipping threshold depends primarily on the number of fine-tuning samples, significantly reducing the need for exhaustive hyper-parameter searches.

A.5 Limitations and Assumptions in Federated Learning

The distributed nature of federated learning makes defending against backdoors particularly challenging, as the aggregator has limited visibility into individual clients' behavior. Additionally, thanks

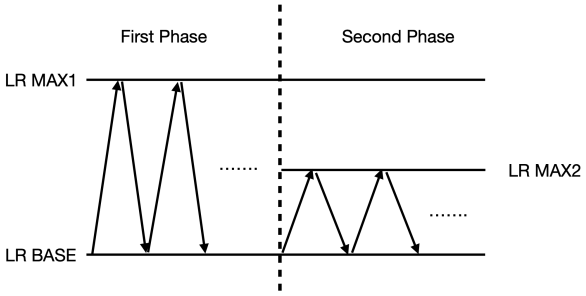


Figure 4: Learning rate scheduling of super-fine-tuning, taken from the original paper Sha et al. (2022).

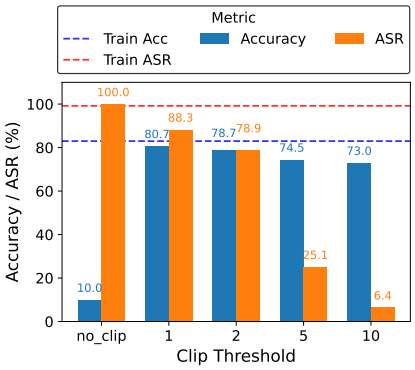


Figure 5: Federated models backdoored with $m = 1$ malicious clients, fine-tuned with 2000 samples over different gradient clipping thresholds.

to their full control over malicious clients, attackers can deploy more potent attack methods that extend beyond simply adding trigger patterns to the data and flipping labels.

Attacker limitations and assumptions. A common assumption in federated learning security is that the majority of clients are benign (also called trustworthy) Wang et al. (2022); therefore, an attacker can only control a malicious minority of clients. Formally, a subset of $m \leq \lfloor \frac{n-1}{2} \rfloor$ malicious clients submit model updates to the aggregator. Krum Blanchard et al. (2017), one of the defenses we study in this work, assumes that $2m + 2 < n$, meaning $m < \frac{n}{2} - 1$. Since we run our experiments with $n = 20$, we use a tighter bound of $m \leq 8$ across our experiments rather than the $m \leq 9$ offered by the malicious minority assumption. Another limitation for the attacker is the requirement to preserve benign accuracy. Otherwise, the attack is easily detectable, and the model is not deployed, meaning the attacker cannot exploit it.

Defender limitations and assumptions. We assume the aggregator has a small, benign dataset drawn from D for fine-tuning, as in MASA Xu et al. (2025b). In practice, the aggregator could be a trusted client using its own data for fine-tuning, or the data could be sourced from public proxy data or a small set of opt-in user data. To maintain the necessity of the federated framework, this dataset must be a fraction of the total training data (e.g., 1%-4%). If the aggregator held a larger portion, a centralized model could be trained locally, rendering federated learning obsolete. In our experiments, we strictly limit the fine-tuning set to be smaller than the average client's local dataset. For instance, our 2000-sample (4%) set is below the average per-client sample count of 2,400, effectively treating the aggregator as a single trusted client. We study the effect of the number of fine-tuning samples in Section D.4. Furthermore, we acknowledge that while backdoors can always be removed through extreme weight transformations (e.g., random re-initialization), such measures are trivial and provide no utility. A viable defense must therefore preserve benign accuracy while preventing backdoors.

B Discussion of Related Work

The research community has recently shown a great interest in federated learning backdoors, from attacks Xie et al. (2019); Zhang et al. (2022); Bagdasaryan et al. (2020); Krauß & Dmitrienko (2023) to defenses Nguyen et al. (2022); Xu et al. (2025a); Minsker (2023); Blanchard et al. (2017) , and surveys Nguyen et al. (2024); Li et al. (2025).

Existing works Zhang et al. (2022); Szelag et al. (2025); Krauß & Dmitrienko (2023) have focused on adversaries with limited capabilities, with at most complete control over the weights they submit to the aggregator, and on other malicious clients, if any exist Li et al. (2025). Such adversaries help defenders estimate their robustness against an *average* real-world adversary. Such estimates might not apply to a worst-case adversary. They do little for *worst-case* security and for studying the

limits of backdoor attacks and defenses, as we do in this work. Several works have raised concerns about the inadequacy or mismatch to reality of existing threat models Krauß & Dmitrienko (2023); Krauß et al. (2024); Szelag et al. (2025), underpinning the relevance of our full-information adaptive adversary threat model.

Adaptive attacks Szelag et al. (2025); Krauß & Dmitrienko (2023); Krauß et al. (2024); Shejwalkar & Houmansadr (2021) have recently emerged and have been shown to be devastating. Attackers can leverage the disconnecting and reconnecting process to adapt to the defense used by the aggregator Szelag et al. (2025), utilize defense metrics to adjust attack parameters Krauß et al. (2024), or adapt through multi-objective optimization Krauß & Dmitrienko (2023). Adaptive attacks have been shown to break existing state-of-the-art defenses Krauß & Dmitrienko (2023). For poisoning attacks, a similar threat model to our full-information adaptive adversary was studied, and it defeated all Byzantine-robust aggregation methods evaluated Shejwalkar & Houmansadr (2021). Our work's empirical study extends their evaluation to backdoor attacks, considering also outlier detection, removal, and combined defenses. All current adaptive attacks attack only single-typed defenses. Therefore, according to our theory, if well-formed, they must succeed (as they do).

Few works study defenses against adaptive adversaries. MESAS Krauß & Dmitrienko (2023) considers strong adversaries that adapt via multi-objective optimization; however, it does not use the benign updates that our attacks leverage. As a defense, MESAS utilizes a plethora of metrics derived from client updates to identify and discard malicious updates, functioning similarly to Multi-Metrics Huang et al. (2023). Unfortunately, MESAS' code is not publicly available; however, if our theory of backdoors is correct, one can defeat MESAS in a similar manner to Multi-Metrics using our full-information adaptive adversary with an optimization-based attack strategy, as we did in Section C.2.

C Adaptive Attacks

For each category, we iterate through existing defenses in order of recency and present either a scaling formula, an interpolation formula, or an optimization objective that breaks said defense.

C.1 Attacking Robust-Aggregation

Norm-bounding (2019). Norm bounding, also called norm clipping Sun et al. (2019), is an l_2-norm-based algorithm that defends against large magnitude attacks. In practice, norm-bounding can be efficiently implemented in cryptography using zero-knowledge proofs, making it a desirable defense for secure aggregation protocols Lycklama et al. (2023). Update norms are clipped to a predefined threshold η to limit the impact of any single update during aggregation. Our adaptive adversary exploits this clipping by directly scaling each malicious update u_m to this predefined threshold, as described in Equation 5. To defend against such an attack, a norm-bounding defender would need to use a very small norm threshold, at the cost of a greatly reduced model accuracy. We find that for any functional threshold, a single adaptive malicious client is sufficient to break norm-bounding, as seen in Table 2.

$$u'_m = \eta \frac{u_m}{\|u_m\|_2} \tag{5}$$

Median-of-Means (2023). Median of Means (MoM) Minsker (2023) is a provably robust aggregation algorithm that aims to minimize the impact of any one update on the aggregation regardless of its magnitude. For each parameter layer in the neural network model, MoM randomly partitions client updates into groups. Then it computes the element-wise mean for each group as its representative. Finally, it takes the element-wise median of the representatives as the update for that layer. In the MoM algorithm, outlier updates move their groups away from the median, reducing the likelihood that they are selected. To improve the robustness of the MoM estimation, we can repeat the MoM algorithm k times and average the medians across the k runs. We refer to this method as the Robust Median of Means (Robust MoM, shortened to Rob MoM). In our experiments, we set $k = 10$ as we did not find further meaningful improvements in robustness with higher values.

Despite MoM provably limiting the impact of malicious updates on the aggregation, our adaptive attacker can still skew the median in the adversarial direction. Scaling up the updates of malicious

clients following Equation 6 causes their updates to overpower the others in their group, affecting the representative. The change incurred by malicious updates to their group's representative shifts the median parameter selected in the adversarial direction for every model parameter. While the change in the median is minor for any individual parameter, the cumulative effect across the high-dimensional weight space results in a significant l_2-norm shift in the adversarial direction. We find that while the defense is successful when there is a single malicious client among 20, it breaks when we introduce a second malicious client, as shown in Table 2. We also did not find any meaningful differences between MoM and Rob-MoM against our adaptive attack.

$$u'_m = \gamma \cdot u_m, \quad \text{where} \quad \gamma \gg \max_{i \in \{1,2,\ldots,n-m\}} \|u_{b_i}\|_2 \tag{6}$$

C.2 Evading Outlier Detection

Krum (2017). Krum Blanchard et al. (2017) is a Byzantine-resilient detection method that selects the most representative update based on its Krum-score: the sum of the l_2 distances to its $n-m-2$ closest neighbors. The update with the lowest score is selected as the update to the global model for that epoch. First, if the attacker controls more than one malicious client, they can overwrite the malicious updates with the mean malicious update. This process eliminates $m-1$ terms from the sum when computing the Krum-score of the malicious updates. Second, suppose the mean malicious update is still not selected after this process. We can interpolate between the mean of the malicious updates and the mean of the benign updates to find the point on the line between the two that is closest to the mean of the malicious updates, while selected by Krum. We solve this directly using bisection with a maximal depth of 10. Using its full information, the adaptive adversary iteratively simulates the Krum aggregation process to perform an informed search along the vector toward the benign mean. This optimization process ensures the malicious update is selected while being positioned as close as possible to the target weights. Let w_b be the mean benign update and w_m be the mean malicious update. Given the maximal distance $\alpha \in [0, 1]$ to the mean benign update on the line such that Krum still selects the update, found by bisection, we can compute the final malicious update as :

$$w'_m = \alpha w_m + (1-\alpha) w_b \tag{7}$$

As shown in Table 2, our adaptive attack is successful against Krum with a single malicious client.

Multi-Metric (2023). Multi-Metric Huang et al. (2023) computes a multi-metric feature vector composed of the pairwise l_1-distance, l_2-distance, and cosine similarity between each pair of updated models. It then applies a whitening transformation to weight the metrics into a unified score dynamically. Then it sorts the updated models by score and selects the 30% with the lowest scores. Finally, it computes a weighted average of the selected updates, where each client's weight is proportional to its self-reported training sample count relative to the total training sample count across all selected participants.

The Multi-Metric score computation is entirely differentiable, rendering it vulnerable to gradient-based adaptive attacks. Because the defense relies on relative pairwise distances across the entire pool of updates, the malicious updates directly influence the scores assigned to the benign clients. We exploit this dynamic by formulating an optimization objective that not only minimizes the malicious clients' scores but simultaneously maximizes the benign clients' scores. We incorporate this into the malicious loss function as follows:

$$\min_w \lambda \mathbb{E}_{(x,y)\in(d\cup b)}[L(x,y)] + \frac{1}{|W_m|} \sum_{w_i \in W_m} M(w_i, W) - \frac{1}{|W_b|} \sum_{w_b \in W_b} M(w_b, W) \tag{8}$$

where λ is a scaling hyperparameter, w represents the malicious model weights being optimized, W_m and W_b are the sets of malicious and benign model weights respectively, $W = W_m \cup W_b$, and $M(\cdot, W)$ is the differentiable Multi-Metric scoring function. Doing so is sufficient to break Multi-Metric with four malicious clients, as shown in Table 2. We also identify a critical integrity flaw in Multi-Metric's design: it relies on self-reported metadata for aggregation weighting. By claiming an artificially large sample size, a selected malicious client can perform a model replacement attack Bagdasaryan et al. (2020), effectively overwriting the global model weights at no additional cost to the attacker.

MARS (2025). MARS Wan et al. (2025) is an outlier detection defense that identifies backdoored updates by evaluating the Backdoor Energy (BE) of individual neurons. To amplify this malicious signal, the aggregator extracts the highest BE values from each layer of the local models to form a one-dimensional Concentrated Backdoor Energy (CBE) vector. The aggregator then separates these vectors into two clusters using K-Means and the 1-Wasserstein distance to capture the overall probability distribution of the energies, regardless of their specific neuron ordering. Finally, the defense aggregates the trusted models by selecting the cluster with the smaller center norm, or defaults to the majority cluster (MARS*) to resist adaptive regularization attacks.

Our adaptive adversary can circumvent MARS by distributing the backdoor signal across a wider set of model parameters, thus flattening the Backdoor Energy (BE) and evading the top-$\kappa\%$ extraction filter. Furthermore, by explicitly minimizing the 1-Wasserstein distance between the malicious and benign CBE distributions (Eq. 9), the attacker can ensure the backdoored update is indistinguishable from the majority cluster, compelling the aggregator to include it in the global model.

$$\min_{w} \mathbb{E}_{(x,y)\in(d\cup b)}[L(x,y)] + \frac{1}{n-m}\sum_{i=1}^{n-m} Wass(CBE(w), CBE(w_b^i)) \tag{9}$$

where $Wass$ denotes the 1-Wasserstein distance, CBE refers to the function that returns the Concentrated Backdoor Energy vector, and w_b^i represents the i-th benign model.

MASA (2025). MASA Xu et al. (2025b) is an outlier detection defense that identifies backdoored models by exploiting their distinct behavior during machine unlearning. Since malicious parameters are largely inactive on benign data, backdoored models exhibit significantly higher empirical losses when the aggregator attempts to unlearn the main task using gradient ascent on a benign proxy dataset. To enhance detection robustness, MASA addresses non-IID divergence by fusing local updates with the global average, accumulating the individual unlearning loss for each model, and discarding any update whose Median Deviation Score (MDS) exceeds a predefined threshold.

Analogous to the adaptive attacks against Krum described in Section C.2 (Equation 7), an informed adversary can bypass MASA using a simulation-guided bisection attack. By using their local benign data as a surrogate for the aggregator's private proxy dataset, the attacker can replicate the unlearning trajectory. Through a binary search between the target malicious update and the benign center, as defined in Equation 7, the attacker can identify the maximum scaling factor that embeds the backdoor while ensuring the simulated Median Deviation Score (MDS) remains below the rejection threshold.

BART-FL (2025). BART-FL Mia & Amini (2025) is an outlier detection defense that filters malicious updates using a combination of dimensionality reduction and statistical voting. The aggregator first applies Principal Component Analysis (PCA) to reduce the dimensionality of received updates and then computes pairwise cosine similarities to quantify their directional alignment. These similarity scores serve as features for K-means clustering, which partitions the clients into two groups. To reliably identify the benign cluster, the defense employs a multi-metric statistical voting mechanism based on point-level mean similarity, cluster-level mean similarity, and Median Absolute Deviation (MAD), ultimately aggregating only the updates from the winning cluster.

The outlier detection pipeline of BART-FL, which relies on PCA projections and K-means clustering, can be modeled as a differentiable constraint. Since a full-information adaptive adversary observes the benign updates, they can locally simulate the aggregator's dimensionality reduction. By optimizing malicious updates to minimize the distance to the benign cluster while maintaining sufficient intra-update variance, the attacker can ensure that their updates are included in the selected cluster and that the backdoor signal is preserved. We describe the adversary models' loss function in Equation 10. Let $P(\cdot)$ be the aggregator's PCA projection function, u_b be the center of the benign updates, and W_{-j} be the set of malicious clients excluding malicious client w_j.

$$\min_{w_j} \mathbb{E}_{(x,y)\in(d\cup b)}[L(x,y)] + \lambda_1||P(u_m) - P(u_b)||_2 - \lambda_2\frac{1}{|W_{-j}|}\sum_{w_i\in W_{-j}} ||w_j - w_i||_2 \tag{10}$$

AlignIns (2025). AlignIns Xu et al. (2025a) is an outlier detection defense that identifies backdoored updates by analyzing their directional patterns at two levels of granularity. It first evaluates

the overall directional deviation of each local update from the global model (w_g) using Temporal Direction Alignment (TDA). Subsequently, it performs a fine-grained Masked Principal Sign Alignment (MPSA) to assess how well the signs of an update's most critical (top-k-largest) parameters align with the majority sign across all clients. Finally, AlignIns applies a median-based Z-score to filter out anomalous updates and clips the magnitudes of the remaining updates to ensure robust aggregation.

AlignIns targets directional patterns through TDA and MPSA. Our adaptive attacker can defeat MPSA by isolating the top-k parameters and optimizing them on benign data to mimic benign sign alignments, as described in Equation 11. For the remaining parameters, the attacker can embed the backdoor signal while adhering to a directional penalty and norm-clipping as described in Equation 12, thereby ensuring the update passes both the TDA filter and the MPSA check.

Let $M_k \in \{0,1\}^{|w|}$ be a binary mask where 1 indicates a top-k parameter and 0 otherwise, and U_b is the set of benign updates. The attacker splits the training into two constrained objectives.
For the top-k parameters ($w_{top} = M_k \odot w$):

$$\min_{w_{top}} \mathbb{E}_{(x,y)\in d}[L(x,y)] \tag{11}$$

For the remaining parameters ($w_{rest} = (1 - M_k) \odot w$):

$$\begin{aligned} &\min_{w_{rest}} \mathbb{E}_{(x,y)\in(d\cup b)}[L(x,y)] - \lambda \cdot \text{TDA}(u_m, w_g), \\ &\text{subject to } \max(|w_{rest}|) < \min(|w_{top}|), \\ &\text{and } ||u_m||_2 \leq \text{Median}_{u_b \in U_b}(||u_b||_2) \end{aligned} \tag{12}$$

C.3 Circumventing Removal Defenses

CSFT (This work). Removal-based defenses attempt to purge backdoors post-aggregation or post-training by applying transformations such as fine-tuning or pruning to the global model weights. However, the efficacy of these methods is fundamentally constrained by the accuracy-robustness trade-off: a defender must remove the malicious signal without damaging the model's benign performance.

Our adaptive adversary exploits this constraint by optimizing for backdoor durability. By strategically increasing the magnitude of malicious updates during training, following Equation 6, the adversary ensures the backdoor signal is deeply embedded within the model's weights. This optimization forces a scenario in which any weight modification significant enough to eliminate the backdoor would concurrently significantly degrade the model's utility. We demonstrate this principle using CSFT, as described in Section A.4. As shown in Table 2, our adaptive attack embeds a sufficiently durable payload to break CSFT with only a single malicious client while maintaining high benign accuracy.

C.4 Breaking Non-Principled Combined Defenses

FoolsGold & Krum (2018). FoolsGold & Krum is designed to use FoolsGold to mitigate sybil attacks (multiple malicious clients) and Krum to handle single-attacker scenarios, providing, in theory, coverage against attacks with any number of malicious clients. However, we show in Section C.1 (Equation 7) how to defeat Krum, and in Table 2, we demonstrate that we can do it with only a single malicious client. Therefore, combining FoolsGold with Krum provides no additional benefit over using Krum alone, and the combined defense is broken by construction.

Fine-tune & Prune (2020). Fine-pruning and fine-tuning are both post-aggregation defense strategies primarily effective against small-magnitude updates. Consequently, our adaptive strategy for circumventing removal defenses (Section C.3) following Equation 5 or Equation 6 remains effective against this combination. This highlights that observing the principles for combination is key to designing robust combined defenses.

FLAME (2022). FLAME Nguyen et al. (2022) integrates HDBSCAN angle-based clustering, median-based clipping, and Differentially-Private (DP) noising. While these components effectively mitigate large angular deviations and high-l_2-norm updates, they remain collectively blind to

Table 4: Evaluation of HA against attacks from previous work. We report the ASR and benign accuracies (Acc.) for $m = 8$.

Defense	$\text{HA}^{\text{CSFT}}_{\text{Krum}}$		$\text{HA}^{\text{CSFT}}_{\text{Flame}}$		$\text{HA}^{\text{CSFT}}_{\text{Multi-Metrics}}$	
Attack	ASR	Acc.	ASR	Acc.	ASR	Acc.
DBA	3.61	72.50	3.77	77.37	**3.84**	**78.74**
BadNet	4.04	74.37	3.33	**78.24**	3.10	78.56
Neurotoxin	3.96	74.52	3.18	78.21	3.20	77.79
Replacement	**21.69**	**74.80**	**3.93**	77.62	3.30	78.05

Table 5: Lowest number of malicious clients (m) out of 20 needed to break our HA defenses on the MNIST dataset. Original m values are in parentheses. Values are averaged over three runs.

Defense	m	ASR	Accuracy
$\text{HA}^{\text{CSFT}}_{\text{Flame}}$	>**8** (8)	23.46 ± 9.41	97.60 ± 0.19
$\text{HA}^{\text{CSFT}}_{\text{Krum}}$	>**8** (1)	19.66 ± 9.11	97.70 ± 0.14
$\text{HA}^{\text{CSFT}}_{\text{Multi Metrics}}$	>**8** (4)	41.37 ± 2.98	97.00 ± 0.08

subtle attacks carried out by a larger number of malicious clients. We model this vulnerability as a differentiable optimization problem where the adversary must simultaneously: (1) minimize cosine similarity to the selected cluster to evade HDBSCAN, (2) maximize l_2-norm updates to counteract median-based clipping (similar to our attack in Section C.1), and (3) reinforce the backdoor signal to withstand DP-noising. During training, we incorporate these objectives into the malicious clients' loss function:

$$\min_w \mathbb{E}_{(x,y)\in(d\cup b)}[L(x,y)] + \frac{\alpha}{|C|}\sum_{w_c\in C}(1-\cos(w,w_c)) + \beta\,\text{ReLU}(l-\|u\|_2)^2 \tag{13}$$

where w represents the malicious model weights, u its update, C the set of benign weights identified by the clustering algorithm, and α and β are hyperparameters. As shown in Table 2, our adaptive attack breaks FLAME with only two malicious clients out of twenty while preserving benign accuracy. For the MNIST dataset LeCun (1998) in Section D.2.1, we find it helpful to interpolate between the mean malicious update and the l_2-projected mean of the closest benign neighbors when the malicious updates are not selected.

D Additional Experiments

D.1 Evaluating HA against Non-Adaptive Attacks

We evaluate the constraint-and-scale attack (C&S) Bagdasaryan et al. (2020) (large deviation attack), the model-replacement attack (Replacement) Bagdasaryan et al. (2020) (large deviation attack), the Distributed Backdoor Attack (DBA) Xie et al. (2019) (large deviation attack), and the Neurotoxin attack Zhang et al. (2022) (small deviation attack). We present the results in Table 4, where we evaluate them against our three best HA variants in the worst-case scenario: $m = 8$ malicious clients out of $n = 20$. We find that none of them are successful, with $\text{HA}^{\text{CSFT}}_{\text{Multi-Metrics}}$ displaying the highest benign accuracies across all attacks. These empirical results further reinforce the defensive capabilities of HA.

D.2 HA on other datasets: Further Evidence Supporting Conjecture 1

D.2.1 MNIST

We evaluate HA further on a different dataset: MNIST LeCun (1998), using a LeNet LeCun et al. (2002) model. We show the lowest number of malicious clients needed to break our best HA variants in Table 5. We find that HA is successful against our adaptive attacks, mirroring our results on CIFAR-10 in Section 5.1, providing further empirical evidence for Conjecture 1, except that our adaptive attack against FLAME is less successful. We posit that this change is due to the simplicity of the MNIST classification task and the LeNet architecture's lower parameter count compared to ResNet, which makes it more difficult for the adversary to hide the backdoor.

D.2.2 ImageNet

We also evaluate HA on the ImageNet Deng et al. (2009) dataset, using a ResNet50 He et al. (2016) model and the $\text{HA}^{\text{CSFT}}_{\text{Krum}}$ defense. We find results that align with our existing findings: the attacker

Table 6: Lowest number of malicious clients (m) out of 20 needed to break our HA defenses on non-IID CIFAR-10 (Dirichlet $\alpha = 0.5$) and when fine-tuning is performed on OOD data (OOD-FT). Original m values are in parentheses. Values are averaged over three runs.

	Non-IID Training			Non-IID Training & Fine-Tuning		
Defense	m (Original)	ASR (%)	Accuracy (%)	m	ASR (%)	Accuracy (%)
$\text{HA}^{\text{CSFT}}_{\text{Flame}}$	**>8 (4)**	12.63 ± 3.96	70.14 ± 1.13	**>8**	36.99 ± 8.49	54.81 ± 1.02
$\text{HA}^{\text{CSFT}}_{\text{Krum}}$	**>8 (1)**	19.94 ± 2.71	73.32 ± 1.46	2	58.88 ± 18.53	57.73 ± 1.45
$\text{HA}^{\text{CSFT}}_{\text{Multi Metrics}}$	**>8 (2)**	37.89 ± 9.62	74.31 ± 0.47	2	60.15 ± 11.40	59.66 ± 1.08

($m = 8$) successfully implants the backdoor during training, backdoor ASR ends at 99.98% for training, and the defender successfully removes the backdoor with CSFT, with a post-CSFT ASR of 0.18%. We note that the global model's accuracy is quite poor, with an end-of-training accuracy of 24.22% and a post-CSFT accuracy of 23.70%. We think this outcome is due to the difficulty of federated learning with larger, more complex datasets like ImageNet.

D.3 HA IN THE NON-IID DATA SETTING

In the real world, data is unlikely to be distributed independently and identically (IID) across clients. As such, we study whether HA's performance is affected by a non-IID data distribution across clients. We follow previous work Xie et al. (2019); Xu et al. (2025b) and use a Dirichlet distribution with a hyperparameter of 0.5 to assign data samples to the clients.

We present our results in Table 6 and demonstrate that even under non-IID training conditions: 1) our full-information adaptive adversary succeeds against individual defenses, with the needed number of malicious clients ranging from 1 to 4; 2) all three of our combined defenses are successful against our adaptive attacks regardless of the number of malicious clients, even in the worst-case scenario of $m = 8$.

We expand our evaluation to include non-IID fine-tuning data. We find that HA experiences a sharp drop in performance, with only $\text{HA}^{\text{CSFT}}_{\text{Flame}}$ remaining undefeated, with a 15.33% accuracy compared to using IID fine-tuning data. These results highlight the importance of collecting trusted, representative fine-tuning data.

D.4 CSFT HYPERPARAMETER STUDY

We empirically validate our hyperparameters for CSFT and show that it requires no additional tuning when the Type 1 defense we combine with changes. We are mainly concerned with the fine-tuning dataset size and the number of fine-tuning epochs. We show that 4% of the total training set size, which is less than the data of one client (5%), works almost best for our best HA variants: $\text{HA}^{\text{CSFT}}_{\text{Flame}}$, $\text{HA}^{\text{CSFT}}_{\text{Krum}}$, and $\text{HA}^{\text{CSFT}}_{\text{Multi-Metrics}}$. More fine-tuning data yields only marginal improvements. We also show that simply picking the epoch with the highest benign accuracy after fine-tuning for 100 epochs is empirically optimal to maximize benign accuracy while keeping ASR below the 50% threshold.

D.4.1 EFFECT OF FINE-TUNING DATASET SIZE

A concern with defenses that require benign data, such as CSFT or MASA Xu et al. (2025b), is the amount of data necessary for the defense to function effectively. We show in Figure 6 that HA 's benign accuracy and defense success rate rapidly converge as fine-tuning data amounts increase, with amounts higher than 4% of the total dataset size (less than the training set of a single client) providing almost no additional benefits.

D.4.2 WHEN TO STOP FINE-TUNING ?

By default, we fine-tune with CSFT for 100 epochs across our experiments. However, since the defender cannot measure the ASR in practice to know when to stop fine-tuning, we need to verify that our choice of the number of fine-tuning epochs is robust across defenses. To that end, we

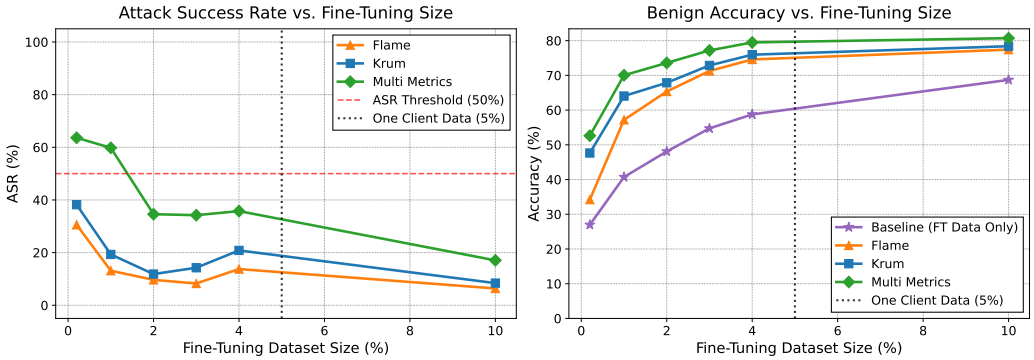


Figure 6: Impact of varying the Fine-Tuning (CSFT) dataset size on Attack Success Rate (left, $m = 8$) and Benign Accuracy (right) for $\text{HA}^{\text{CSFT}}_{\text{Flame}}$, $\text{HA}^{\text{CSFT}}_{\text{Krum}}$, and $\text{HA}^{\text{CSFT}}_{\text{Multi-Metrics}}$ on CIFAR-10, alongside a baseline trained centrally solely on the fine-tuning subset, averaged over five seeds. Dataset sizes are expressed as a percentage of the total training data. The vertical dotted line represents the equivalent data held by a single local client (5%).

Table 7: Comparison of early-stopping epoch timings during fine-tuning. 'Safe' epoch reports the epoch that maximizes benign accuracy while successfully keeping ASR strictly below 50%. 'Peak' epoch reports the epoch with the absolute highest benign accuracy regardless of the backdoor ASR. Both strategies mandate a minimum of 10 fine-tuning epochs.

m	1		2		4		8	
Defense	Safe	Peak	Safe	Peak	Safe	Peak	Safe	Peak
$\text{HA}^{\text{CSFT}}_{\text{Flame}}$	98	98	94	94	93	93	96	96
$\text{HA}^{\text{CSFT}}_{\text{Krum}}$	97	97	99	99	100	100	100	100
$\text{HA}^{\text{CSFT}}_{\text{Multi-Metrics}}$	100	100	94	94	91	91	97	97

compute the safe optimal epoch to stop at, i.e., the epoch that maximizes benign accuracy while the ASR is below 50%. We compare the safe optimal epoch to the peak epoch, which is the epoch with the highest benign accuracy regardless of ASR. We display the results in Table 7. We find that for all HA variants and for all numbers of malicious clients, the safe and peak stopping epochs are exactly the same. Hence, CSFT requires no surrogate metric or process to determine when to stop fine-tuning; simply picking the epoch with the highest benign accuracy is empirically optimal.