FedRGD: Risk-Guided Dynamic Defense against Federated Backdoors

Ruiying Wang and Shuyang Hou (✉)

College of Artificial Intelligence, Tianjin University of Science and Technology, Tianjin 300450, China
houshuyang1@outlook.com

Abstract. Federated Learning (FL) is vulnerable to backdoor attacks, where adversaries can stealthily manipulate the global model. Most existing defense methods are developed under IID assumptions, an assumption that rarely holds in practice. In highly non-IID settings, heterogeneous data distributions across clients make it difficult to distinguish malicious updates from benign ones, particularly when benign clients exhibit atypical patterns due to minority-class data. To address this challenge, existing defenses operate at different levels of granularity. Coarse-grained methods perform client-level filtering, which often mistakenly excludes benign clients under non-IID conditions. Fine-grained methods instead analyze data at the sample level for more precise detection, but typically rely on explicit per-sample gradient analysis, leading to substantial memory and computational overhead. As a result, defending against backdoor attacks in non-IID environments involves a fundamental trade-off between robustness and computational efficiency. To address this challenge, we propose FedRGD, a federated risk-guided dynamic defense framework that bridges sample-level anomaly estimation and parameter-level update control. Instead of explicitly constructing per-sample gradients, FedRGD propagates sample risks into parameter importance through a VJP-based implicit mechanism, enabling efficient fine-grained defense with $O(|\theta|)$ memory complexity. Furthermore, a dual-stream consistency evaluation and adaptive masking strategy are jointly designed to suppress stealthy poisoned updates under highly non-IID settings. Extensive experiments on CIFAR-10 and Fashion-MNIST demonstrate that FedRGD consistently reduces the attack success rate while maintaining high main-task accuracy, achieving strong robustness with near-baseline training efficiency.

Keywords: Federated Learning, Backdoor Defense, Dynamic Masking, Non-IID Data, Efficient Defense.

1 Introduction

Federated learning (FL) is a distributed learning paradigm in which multiple clients collaboratively train a global model while keeping their raw data local [1]. This decentralized training mechanism provides an effective way to protect data privacy, but it also prevents the server from directly inspecting local data and training behaviors. As a result, FL is inherently vulnerable to backdoor attacks [2], where

malicious clients inject trigger patterns into local data and manipulate local updates so that the global model learns attacker-specified associations [3]. Under such attacks, the compromised model can maintain high accuracy on clean inputs while producing targeted misclassifications once trigger-embedded samples appear, posing a serious threat to the reliability of federated systems.

To mitigate backdoor threats in federated learning, early defense methods predominantly adopt coarse-grained strategies, such as client-level filtering based on global statistical consistency [4, 5], which are largely developed under IID assumptions where benign updates follow similar distributions. However, in realistic federated settings with highly non-IID data across clients [6], benign updates can deviate significantly due to skewed or minority-class data, making them difficult to distinguish from malicious ones and causing coarse-grained defenses to either misclassify benign clients or overlook stealthy attacks [7]. To overcome these limitations, recent approaches move toward fine-grained analysis at the parameter or sample level [8, 9], aiming to capture subtle discrepancies in local updates and improve robustness under heterogeneous conditions and adaptive attacks. While effective, these methods typically rely on explicit per-sample gradient computation, introducing memory and computational overhead that scales linearly with the batch size, which limits their practicality and scalability in resource-constrained federated environments [10].

This exposes a fundamental dilemma in federated backdoor defense: coarse-grained methods are computationally efficient but often unreliable under non-IID data, whereas fine-grained methods provide stronger discrimination at the cost of substantial overhead [11]. This tension becomes even more severe in heterogeneous environments with adaptive attacks, where benign and malicious behaviors are more difficult to separate. Therefore, an effective defense should simultaneously achieve strong robustness and practical efficiency.

To address this challenge, we propose FedRGD, an efficient dynamic masking framework for federated backdoor defense. The key advantage of FedRGD is that it bridges sample-level risk estimation and parameter-level update control, enabling fine-grained suppression of malicious behavior without incurring the heavy cost of explicit per-sample gradient computation. In this way, the proposed framework effectively balances robustness and efficiency under heterogeneous non-IID settings. To achieve this, FedRGD follows a three-stage design: it first establishes a robust global prototype before formal training begins, then performs dual-stream local risk assessment to propagate sample-level risk into parameter-level masking, and finally incorporates the sanitized updates into a privacy-preserving robust aggregation scheme to further suppress residual malicious interference.

Existing federated backdoor defenses can generally be divided into three categories. The first category focuses on robust aggregation at the client level, such as Multi-Krum and FLAME, which improve robustness through statistical filtering or clustering. However, these methods are often sensitive to highly non-IID data distributions and may mistakenly suppress benign minority updates.

The second category adopts fine-grained parameter or sample-level analysis to improve discrimination capability under heterogeneous settings. Although these

methods achieve stronger robustness, they usually rely on explicit per-sample gradient computation, resulting in substantial computational and memory overhead.

Different from existing approaches, FedRGD bridges sample-level risk estimation and parameter-level update control through an implicit VJP-based propagation mechanism, enabling efficient fine-grained defense while maintaining scalability in practical federated environments.

The main contributions of this work can be summarized as follows.

- We propose FedRGD, a unified framework for cross-granularity backdoor defense in FL, which bridges sample-level risk estimation and parameter-level update control to achieve a favorable balance between robustness and computational efficiency under non-IID data.
- We develop a risk-aware defense mechanism that combines robust prototype initialization, dual-stream sample risk assessment, and adaptive parameter-level masking to enable efficient propagation of suspicious sample behaviors into fine-grained update suppression.
- We conduct extensive experiments on benchmark datasets to validate the effectiveness of FedRGD. The results demonstrate that it consistently reduces attack success rate (ASR) while maintaining high main-task accuracy, outperforming existing methods in terms of the robustness–efficiency trade-off.

2 Threat Model and Design Goals

We consider a standard FL setting with K clients and a central server, where a fraction p of clients are malicious, with $1/K \le p < 0.5$. These adversarial clients inject backdoors by uploading poisoned local updates, aiming to manipulate predictions on trigger-embedded inputs while maintaining high performance on benign data.

Adversary's Knowledge and Abilities. Following standard backdoor settings, the adversary is assumed to be an internal participant in the federated training process [12]. Specifically, malicious clients can: (1) manipulate local data and training procedures to implant trigger patterns; and (2) access model parameters in a white-box manner to optimize poisoned updates. In addition, we consider a strong adaptive attacker that is aware of the defense mechanism and attempts to reduce detection signals during local training [13].

Adversary's Goals. The adversary aims to achieve both effectiveness and stealthiness by maximizing ASR while preserving Main Task Accuracy (MTA) [14]. This objective is formulated as:

$$\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{D}_{mal}}[\mathcal{L}(F(x \oplus \delta;\theta), y_t)] + \lambda \mathbb{E}_{(x,y)\sim\mathcal{D}_{clean}}[\mathcal{L}(F(x;\theta), y)] \quad (1)$$

where $L(\cdot)$ denotes the standard cross-entropy loss, $x \oplus \delta$ denotes a malicious sample obtained by injecting the trigger δ into the original input x, y_t is the target label, and λ balances attack effectiveness and stealthiness.

Adaptive Attack Formulation. To evaluate robustness under a strong threat model, we consider a white-box adaptive attacker aware of the defense mechanism. The

attacker targets transformation-invariant triggers to reduce detection signals induced by the dual-stream design. To this end, we adopt an Expectation Over Transformation (EOT) strategy:

$$\min_{\theta,\delta} \mathbb{E}_{(x,y)\sim\mathcal{D}_m}\left[\mathbb{E}_{t\sim\mathcal{T}}\left[\mathcal{L}(F(t(x+\delta);\theta),y_t)\right]\right] \tag{2}$$

where $t\sim\mathcal{T}$ denotes geometric transformations aligned with the self-supervised stream. Compared to static optimization, this formulation enforces transformation consistency, producing stable poisoned representations under input variations. As a result, intra-sample discrepancies are reduced while poisoned features become partially aligned with benign patterns, weakening anomaly signals. By integrating this objective into local training, the attacker generates stealthy updates that are explicitly optimized to evade fine-grained masking.

Defender's Knowledge and Constraints. The defender aggregates client updates but has no knowledge of trigger patterns, poisoning ratio ρ, or compromised clients. We consider realistic constraints: (1) no access to clean validation data; (2) limited computation and memory at clients, making costly local inspection impractical.

Design Goals. Under the above setting, FedRGD is designed to achieve three core objectives: (1) effectiveness, by suppressing malicious updates and reducing ASR while preserving MTA under adaptive attacks; (2) efficiency, by maintaining low computational and memory overhead suitable for resource-constrained FL; and (3) robust initialization, by establishing a reliable initial model without relying on clean data or historical statistics.

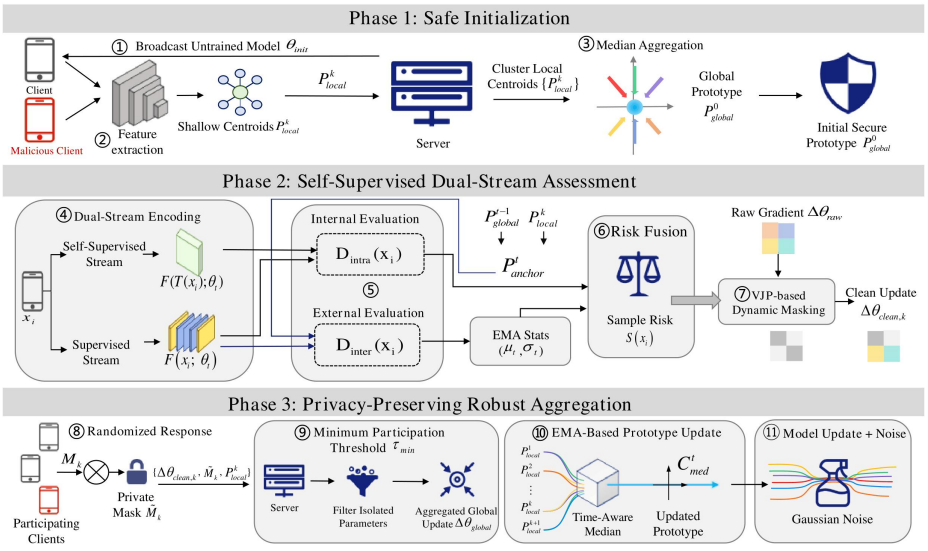

Fig. 1. Overview of FedRGD. The framework consists of three phases: safe initialization, dual-stream risk assessment with dynamic masking, and privacy-preserving robust aggregation.

3 Methodology

In this section, we present FedRGD, a dynamic masking framework for robust FL. As illustrated in Fig. 1, FedRGD consists of three tightly coupled stages: safe prototype initialization, dual-stream sample risk assessment with risk-guided parameter masking, and privacy-preserving robust aggregation. Specifically, the method first constructs a reliable global prototype before formal training begins, then estimates sample-level risk from complementary consistency cues and propagates it into parameter-level control signals, and finally performs robust server-side aggregation with privacy protection and prototype calibration. The details of each component are described below.

3.1 Safe Prototype Initialization via Shallow Features

At the initial stage, the server has no prior knowledge of client reliability, while malicious clients may already participate in training. Direct model aggregation in early rounds can therefore be easily dominated by poisoned updates. To avoid this vulnerability, FedRGD first establishes a robust global prototype before formal federated optimization starts.

Specifically, the server distributes a randomly initialized model θ_{init} to all clients. Although untrained, its shallow layers still preserve low-level input statistics, which are sufficient to capture coarse distributional structure and reveal extreme outliers under distribution shifts. Each client uses the first L layers as a shallow feature extractor $F_{shallow}(\cdot;\theta_{init}^{1:L})$. For the local dataset $\mathcal{D}_k$, the centroid of shallow feature P_{local}^k is computed as:

$$P_{local}^k = \frac{1}{|\mathcal{D}_k|}\sum_{x\in\mathcal{D}_k} F_{shallow}(x;\theta_{init}^{1:L}) \tag{3}$$

Where x represents an individual data sample, and $|\mathcal{D}_k|$ denotes the total number of samples in the client's dataset. Clients upload only the centroid features P_{local}^k to the server.

Since poisoned data typically induce significant distribution shifts, their centroids tend to appear as outliers in the feature space. The server therefore applies a coordinate-wise median aggregation over all client centroids to obtain a robust global prototype:

$$P_{global}^0 = \text{Median}(\{P_{local}^1, P_{local}^2, ..., P_{local}^K\} \tag{4}$$

This design avoids direct weight aggregation at initialization and provides a stable global reference that is naturally robust to extreme malicious clients.

3.2 Dual-Stream Risk Assessment

To achieve fine-grained local defense, FedRGD estimates sample-level risk from complementary internal and external consistency cues, and then propagates the inferred risk into parameter-level masking for update sanitization. During local training, backdoor triggers may induce stable and task-consistent behaviors, making

malicious samples difficult to distinguish using supervised training signals alone. This limitation motivates the need for additional reliability cues beyond task supervision. To this end, we introduce a dual-stream risk assessment mechanism that evaluates each sample from both representation consistency and distributional alignment.

Specifically, the server distributes the global model θ_t to all clients. For each input sample x_i, the supervised stream operates on the original input, while a parallel self-supervised stream processes a transformed version $T(x_i)$, where $t \sim \mathcal{T}$ denotes a random geometric transformation. Let $F(\cdot;\theta_t)$ denote the deep feature representation extracted from the penultimate layer of the global model θ_t.

Internal Consistency Evaluation. Internal consistency measures whether a sample preserves stable representations under geometric transformation. Backdoor triggers often rely on spatially stable patterns. When geometric transformations are applied, such patterns tend to be disrupted, leading to inconsistent feature representations. We measure this effect using the cosine distance:

$$D_{intra}(x_i) = 1 - \frac{F(x_i;\theta_t) \cdot F(T(x_i);\theta_t)}{\| F(x_i;\theta_t)\|_2 \| F(T(x_i);\theta_t)\|_2} \tag{5}$$

where $D_{intra}(x_i)$ quantifies the internal inconsistency of sample x_i, with larger values indicating higher potential risk.

External Consistency Evaluation. External consistency measures whether a sample aligns with the expected feature distribution defined by a hybrid anchor. Since internal inconsistency alone may be insufficient, we further evaluate each sample against a global reference. We define a hybrid anchor:

$$P^t_{anchor} = \lambda P^{t-1}_{global} + (1-\lambda) P^t_{local} \tag{6}$$

where P^{t-1}_{global} is the global prototype from the previous round, P^t_{local} denotes the centroid of the current client, and $\lambda \in [0,1]$ is a hyperparameter balancing the two components.

The external inconsistency is measured as:

$$D_{inter}(x_i) = \| F(x_i;\theta_t) - P^t_{anchor}\|_2 \tag{7}$$

which captures the deviation of the sample from the expected feature distribution.

Unified Risk Estimation. To capture complementary anomaly signals, FedRGD combines internal inconsistency and external deviation into a unified risk score:

$$S(x_i) = \omega D_{intra}(x_i) + (1-\omega)\sigma\left(\frac{D_{inter}(x_i) - \mu_t}{\sigma_t + \epsilon}\right) \tag{8}$$

where $\omega \in (0,1)$ is a weighting coefficient, μ_t and σ_t denote the mean and standard deviation of the current batch. In practice, these statistics are updated online for stability. This formulation assigns higher risk to samples that simultaneously exhibit transformation instability and feature-space deviation.

Risk-Guided Parameter Masking. Directly translating sample-level risk into parameter-level control is non-trivial, since naive implementations may incur prohibitive memory overhead. To address this issue, FedRGD formulates the risk-to-parameter propagation as a Vector-Jacobian Product (VJP) computation, which

enables efficient accumulation of parameter risk while strictly maintaining $O(|\theta|)$ memory complexity. Let $S(x_i)$ denote the risk score for the i-th sample. The cumulative risk R_j for parameter θ_j is computed efficiently via a single backward pass:

$$R_j = \left| \sum_{i=1}^{B} S(x_i) \nabla \theta_j^{(i)} \right| \tag{9}$$

where B denotes the batch size and $\nabla \theta_j^{(i)}$ is the gradient of parameter θ_j with respect to the i-th sample. This design avoids explicit storage of per-sample gradients while preserving fine-grained risk propagation from suspicious samples to vulnerable parameters. Through this computation, parameters heavily influenced by high-risk samples naturally accumulate larger risk values. Since parameter risk distributions may vary substantially across layers, a uniform threshold is often suboptimal. We therefore define a dynamic layer-wise threshold $\tau_{dyn}^{(l)}$ as the p-quantile of the risk distribution $\mathbf{R}^{(l)}$ in layer l, where $p \in (0,100)$ is a time-aware dynamic retention rate controlling the proportion of preserved parameters.

Based on the parameter risk, a binary mask M is generated to suppress unreliable updates. Specifically, the mask entry is set to 1 if the parameter risk does not exceed the layer-wise threshold, and 0 otherwise:

$$M_j = \mathbb{I}(R_j \le \tau_{dyn}^{(l)}) \tag{10}$$

where $\mathbb{I}(\cdot)$ represents the indicator function. The client then applies M to the raw local gradient $\Delta\theta_{raw}$ to obtain a sanitized update:

$$\Delta\theta_{clean} = M \odot \Delta\theta_{raw} \tag{11}$$

where $\odot$ denotes the element-wise product. This operation forms the core local defense of FedRGD, enabling sample-aware fine-grained parameter control without explicit sample-wise gradient filtering.

Remark on Defense Scope. FedRGD provides a complementary dual-guard mechanism. The internal consistency branch explicitly targets localized geometric triggers, while the external consistency branch serves as a broader safeguard against feature-space deviations. As a result, even when triggers are less sensitive to transformations, the method can still mitigate their influence through distributional anomaly detection.

3.3 Privacy-Preserving and Robust Aggregation

At the server side, FedRGD performs robust masked aggregation as the core update rule. In addition, two auxiliary strategies, namely randomized mask perturbation and dynamic prototype calibration, are introduced to enhance privacy protection and maintain feature-space alignment during training. After local dynamic masking, each client obtains a sanitized model update $\Delta\theta_{clean}$ and a corresponding binary mask matrix M. Directly uploading these components may expose feature sparsity patterns and leave the server vulnerable to collusive manipulation on rarely updated parameters. To address these issues, FedRGD adopts a privacy-preserving and robust aggregation scheme, as detailed below.

Randomized Response for Mask Privacy. To reduce privacy leakage caused by directly uploading sparse mask patterns, each client perturbs its binary mask before transmission. Specifically, each entry $M_{k,j}$ is independently flipped with a small probability $q \in (0, 0.5)$ to produce a privatized mask $\tilde{M}_{k,j}$. This stochastic perturbation obscures local sparsity characteristics while having limited impact on aggregation, since the corresponding sanitized gradients remain highly sparse.

Threshold-Aware Robust Aggregation. To prevent isolated malicious clients from dominating rarely activated parameters, the server performs normalized masked aggregation with a minimum participation threshold $\tau_{min} = \lceil \eta \cdot K \rceil$, where $\eta \in (0,1)$ is the predefined participation ratio and K is the total number of participating clients in the current round. The aggregated update for the j-th parameter is computed as:

$$\Delta\theta^{j}_{global} = \frac{\sum_{k=1}^{K} \tilde{M}_{k,j} \cdot \Delta\theta^{j}_{clean,k}}{\max\left(\sum_{k=1}^{K} \tilde{M}_{k,j}, \tau_{min}\right)} \tag{12}$$

Here $\Delta\theta^{j}_{clean,k}$ is the sanitized gradient from client k, and $\tilde{M}_{k,j}$ is the perturbed mask. This normalization prevents unreliable updates from being disproportionately amplified when only a few clients contribute to a parameter.

To further suppress residual backdoor topologies that may survive early filtering, the server injects targeted Gaussian noise during the global model update:

$$\theta_{t+1} = \theta_t + \eta_g \Delta\theta_{global} + \mathcal{N}(0, \sigma^2) \tag{13}$$

Here, the noise variance σ^2 controls the strength of perturbation. This operation serves as an auxiliary regularization mechanism against stealthy residual malicious structures.

Dynamic Calibration of the Global Prototype. A static security baseline obtained in early rounds may gradually become misaligned with the evolving feature space. To maintain a reliable global reference, the server dynamically updates the global prototype using an exponential moving average. Specifically, the server first computes the median centroid C^{t}_{med} from uploaded local feature centroids, and then updates the global prototype P^{t}_{global} as:

$$P^{t}_{global} = (1-\beta) P^{t-1}_{global} + \beta C^{t}_{med} \tag{14}$$

where $\beta \in (0,1]$ is the momentum coefficient. This calibration stabilizes the global reference under feature drift and helps prevent long-term contamination of the prototype by gradual poisoning.

4 Experiments

We conduct extensive experiments to evaluate FedRGD under highly heterogeneous data distributions. We assess its ability to preserve main task performance while suppressing ASR across diverse backdoor attacks. We further evaluate its computational and memory efficiency.

4.1 Experimental Settings

Datasets and Federated Environment. We evaluate FedRGD on two widely used image classification benchmarks, CIFAR-10 and Fashion-MNIST. CIFAR-10 consists of 60,000 color images from 10 object categories with relatively complex visual semantics, making it suitable for evaluating robustness under challenging feature distributions. Fashion-MNIST contains 70,000 grayscale clothing images across 10 categories and provides a lightweight benchmark for validating defense generalization under simpler visual structures. We adopt a ResNet-18 architecture for CIFAR-10 and a lightweight four-layer convolutional neural network for Fashion-MNIST. All experiments are implemented using the PyTorch framework and conducted on a server equipped with a single NVIDIA GeForce RTX 4090 GPU (24 GB). We construct a cross-device FL setup comprising 100 clients, where 20% of clients participate in each communication round and 10% are malicious. To simulate realistic highly heterogeneous edge environments, the local training data is partitioned according to a Dirichlet distribution $Dir(\alpha)$, where the heterogeneity parameter is set to $\alpha = 0.5$. The local training batch size is set to 32 to accommodate the memory footprint of defense evaluations, while the testing batch size is set to 128. The global model is trained for 200 communication rounds, and a global learning rate decay is applied at round 100 to evaluate robustness against late-stage chronic backdoor rebounds.

Attack Scenarios and Defense Baselines. To thoroughly evaluate defense robustness, we implement four representative backdoor attacks. BadNets serves as a baseline data poisoning approach using static pixel patterns [15]. Neurotoxin represents stealthy model poisoning by embedding backdoor features into dormant neurons to evade parameter pruning [16]. To simulate advanced dynamic threats, we deploy IBA to craft input-specific triggers, alongside a white-box adaptive attack based on EOT that leverages geometric transformations to test robustness limits against defense-aware adversaries [17]. To benchmark the mitigation capabilities, we compare FedRGD against a no-defense baseline FedAvg and three state-of-the-art defenses representing distinct technical paradigms. Multi-Krum is selected as the representative of distance-based robust aggregation methods [18]. FLAME evaluates the effectiveness of system-level clustering combined with adaptive parameter clipping [19]. Finally, FedRoLA is included as the latest structure-aware hierarchical framework to benchmark fine-grained parameter-level defense performance [20].

Evaluation Metrics. We assess performance using three groups of metrics. The MTA measures the classification performance of the global model on a clean test set to reflect utility preservation. The ASR represents the proportion of poisoned samples embedded with triggers that are misclassified into the target label. Finally, to quantify the efficiency advantage of the proposed VJP computation over conventional per-sample gradient methods, we report the peak memory usage in megabytes and the per-round computation time in seconds.

4.2 Defense Effectiveness Evaluation

To explicitly investigate the fundamental robustness-utility conflict under highly non-IID settings introduced earlier, we evaluate defense strategies with the Dirichlet heterogeneity parameter α set to 0.5. As anticipated, the empirical results in Table 1 reveal that coarse-grained defenses such as Multi-Krum and FLAME struggle severely with data heterogeneity. These methods mistakenly filter benign minority-class updates, causing the MTA to degrade to approximately 52% to 57% on CIFAR-10. Conversely, while existing fine-grained defenses like FedRoLA attempt layer-wise aggregation to preserve utility, they fail to maintain security against advanced dynamic threats, yielding ASRs exceeding 72% against IBA and Adaptive attacks. This confirms the limitations of current paradigms in balancing precise threat isolation and benign feature retention.

FedRGD addresses this limitation through risk-guided dynamic masking. By projecting sample-level risks into fine-grained parameter isolation, the proposed framework consistently suppresses malicious updates across various backdoor paradigms. As shown in Table 1, FedRGD suppresses the ASR to below 39% on CIFAR-10 and 12% on Fashion-MNIST. Furthermore, this precise parameter-level control effectively avoids the severe exclusion of benign clients inherent to coarse-grained methods. Consequently, FedRGD achieves a stable MTA of approximately 60% on CIFAR-10 and 89% on Fashion-MNIST. While the absolute MTA on CIFAR-10 may appear moderately degraded, this is an expected trade-off under the strictly highly non-IID ($\alpha = 0.5$)and heavily poisoned federated environment. Under these severe constraints, FedRGD still significantly outperforms the evaluated baselines.

Table 1. Performance comparison of different defense strategies against various backdoor attacks.

Dataset	Defenses	BadNets		Neurotoxin		IBA		Adaptive	
		ASR	MTA	ASR	MTA	ASR	MTA	ASR	MTA
CIFAR-10	FedAvg	90.23	76.80	35.23	70.13	93.74	60.28	89.05	64.56
	Multi-Krum	88.84	53.70	48.63	52.16	51.67	54.05	62.19	51.94
	FLAME	71.08	57.40	45.22	54.40	62.88	53.03	58.64	55.35
	FedRoLA	58.57	60.64	37.93	53.81	68.86	58.11	72.13	56.32
	Ours	38.40	61.94	23.88	59.14	27.02	59.53	24.52	60.23
Fashion-MNIST	FedAvg	99.54	91.11	10.16	89.11	97.46	88.18	99.10	90.63
	Multi-Krum	42.83	85.04	22.45	84.97	32.31	85.04	44.37	84.04
	FLAME	72.63	86.89	23.51	85.15	49.07	86.28	52.98	85.10
	FedRoLA	39.48	86.03	28.03	84.05	42.35	88.44	48.22	87.60
	Ours	11.98	86.98	2.78	89.15	3.47	88.84	4.68	89.08

4.3 Robustness Against Adaptive Attacks

To evaluate the robustness of FedRGD under our strongest threat setting, we evaluate its robustness against white-box adaptive attacks. In this threat model, the adversary utilizes an EOT strategy to craft transformation-invariant triggers explicitly designed to bypass our internal spatial consistency evaluation. Although this localized evasion allows some poisoned gradients to blend with benign updates, the manipulated features inevitably induce distribution shifts in the global space. The external hybrid anchor P'_{anchor} within our dual-stream architecture successfully captures these deviations. As demonstrated in Table 1, baseline methods like FedRoLA exhibit severe degradation in robustness with ASR exceeding 72% on CIFAR-10. Conversely, the complementary detection mechanism of FedRGD suppresses the threat to 24.52% on CIFAR-10 and 4.68% on Fashion-MNIST while preserving stable MTA. This empirical evidence validates that the dual-stream design establishes a resilient security loop against defense-aware adversaries.

Table 2. System overhead and efficiency comparison.

Defenses	Gradient Type	Peak Memory (MB)	Per-round Time (s)
FedAvg	Batch-Averaged	19274.32	23.25
Multi-Krum	Batch-Averaged	19545.60	31.54
FLAME	Batch-Averaged	22318.82	57.03
FedRoLA	Per-Sample	36191.95	77.07
Ours	VJP (Implicit)	20612.84	25.63

Table 3. Ablation study on the core components of FedRGD.

Defense Configuration	BadNets		Adaptive	
	ASR	MTA	ASR	MTA
w/o Safe Initialization	41.50	56.83	31.40	57.74
w/o Dual-Stream Eval	44.20	57.65	28.65	57.78
w/o Dynamic Masking	65.21	68.43	52.45	65.92
w/o Robust Aggregation	68.50	54.20	48.75	52.15
Full FedRGD	38.40	61.94	24.52	60.23

4.4 System Overhead and Efficiency Analysis

Fine-grained defenses incur significant system overhead due to per-sample gradient computation, where memory consumption scales linearly with both the batch size B and the number of model parameters P, leading to poor scalability.

As shown in Table 2, FedRoLA exhibits the highest overhead with 36191.95 MB memory and 77.07 s, indicating that explicit per-sample gradient construction is prohibitively expensive. FLAME reduces memory by using batch-averaged updates but still requires 57.03 s due to clustering over high-dimensional model updates,

showing that eliminating per-sample gradients alone does not resolve system-level bottlenecks. In contrast, FedRGD avoids explicit gradient materialization via VJP-based implicit propagation, reducing the complexity to $\mathcal{O}(P)$, and achieves 20612.84 MB memory and 25.63 s. This represents a massive 43% reduction in peak memory and a threefold speedup compared to FedRoLA, which is remarkably close to the unprotected FedAvg with 19274.32 MB and 23.25 s. These results demonstrate that FedRGD retains fine-grained robustness while maintaining near-baseline efficiency.

4.5 Ablation Study

To verify the necessity of the proposed multi-stage architecture, we conduct an ablation study by systematically removing four core components under the strict $\alpha=0.5$ non-IID setting against BadNets and Adaptive attacks.

As shown in Table 3, Dynamic Masking serves as the primary defense barrier. Without it, the ASR increases to 65.21% against BadNets and 52.45% against Adaptive attacks. This indicates that computing anomaly scores without executing explicit parameter-level isolation is insufficient to block malicious updates. Additionally, removing Dual-Stream Evaluation restricts the system to internal spatial assessment. Because Adaptive attacks utilize transformation-invariant triggers, this single-stream setup fails to capture their distributional deviations, elevating the Adaptive ASR to 28.65%. This confirms the necessity of joint internal-external checks in identifying well-camouflaged threats.

Furthermore, omitting Robust Aggregation removes the dilution effect on anomalous gradients during the server-side update. Without this structural disruption, the defense becomes highly vulnerable, allowing the BadNets ASR to reach 68.50% and the Adaptive ASR to rebound to 48.75%. Finally, disabling Safe Initialization deprives the server of an early-stage clean reference, leading to a chronic backdoor penetration where BadNets and Adaptive ASRs increase to 41.50% and 31.40%, respectively. These findings confirm that the components form a cohesive mechanism, and the absence of any single module compromises the overall security-utility balance.

5 Conclusion

This paper proposes FedRGD, a lightweight and robust defense framework for mitigating stealthy backdoor attacks in FL under highly heterogeneous data settings. By integrating dual-stream assessment with time-aware dynamic masking and targeted Gaussian noise perturbation, FedRGD detects and filters malicious updates while preserving benign client contributions. Extensive experiments demonstrate that FedRGD consistently suppresses advanced adaptive attacks and late-stage chronic backdoor rebounds while maintaining high MTA, achieving a strong balance between security and utility. The framework is computationally efficient and suitable for deployment in resource-constrained environments. Future work will explore its extension to decentralized federated settings and large-scale model fine-tuning.

References

1. McMahan, B., Moore, E., Ramage, D., Hampson, S., Arcas, B.A.: Communication-efficient learning of deep networks from decentralized data. In: 20th International Conference on Artificial Intelligence and Statistics (AISTATS), pp. 1273–1282. PMLR (2017)
2. Bagdasaryan, E., Veit, A., Hua, Y., Estrin, D., Shmatikov, V.: How to backdoor federated learning. In: 23rd International Conference on Artificial Intelligence and Statistics (AISTATS), pp. 2938–2948. PMLR (2020)
3. Wang, H., Sreenivasan, K., Rajput, S., Vishwakarma, H., Agarwal, S., Sohn, J.Y., Lee, K., Papailiopoulos, D.: Attack of the tails: Yes, you really can backdoor federated learning. In: Advances in Neural Information Processing Systems (NeurIPS), vol. 33, pp. 16070–16084. Curran Associates, Inc. (2020)
4. Zhang, M., Wang, Y., Li, J., Chen, X.: FilterFL: Knowledge filtering-based data-free backdoor defense for federated learning. IEEE Transactions on Dependable and Secure Computing 21(5), 1845–1859 (2024)
5. Ozdayi, M.S., Kantarcioglu, M., Gelernter, Y.: Defending against backdoors in federated learning with robust learning rate. In: Proceedings of the 35th AAAI Conference on Artificial Intelligence, pp. 9268–9276. AAAI Press (2021)
6. Wu, J., Zhang, Y., Wang, C., Li, X.: MARS: A malignity-aware backdoor defense in federated learning. IEEE Transactions on Dependable and Secure Computing 22(1), 215–229 (2025)
7. Rieger, P., Nguyen, T.D., Mi, H., Sadeghi, A.R.: CrowdGuard: Federated backdoor detection in federated learning. In: Proceedings of the Network and Distributed System Security Symposium (NDSS). Internet Society (2024)
8. Awan, S., Luo, B., Li, F.: CONTRA: Defending against poisoning attacks in federated learning. In: 26th European Symposium on Research in Computer Security (ESORICS), LNCS, vol. 12972, pp. 455–475. Springer, Cham (2021)
9. Wu, C., Yang, P., Wang, K., Wang, M.: Defending against backdoor attacks in federated learning via adaptive dynamic masking. IEEE Transactions on Information Forensics and Security 19, 1234–1248 (2024)
10. Li, Y., Chen, Z., Zhang, H., Liu, X.: A stability-enhanced dynamic backdoor defense in federated learning for IIoT. IEEE Internet of Things Journal 11(4), 6789–6802 (2024)
11. Chen, Z., Wang, X., Liu, Y., Zhao, L.: Seeing through the threat: Adaptive attention-guided backdoor defense in non-IID federated learning. IEEE Transactions on Information Forensics and Security 20, 1024–1038 (2025)
12. Xie, C., Huang, K., Chen, P.Y., Li, B.: DBA: Distributed backdoor attacks against federated learning. In: International Conference on Learning Representations (ICLR). OpenReview (2020)
13. Sun, Z., Kairouz, P., Suresh, A.T., McMahan, H.B.: Can you really backdoor federated learning? In: SysML Conference. (2019)
14. Zhao, Y., Li, S., Wang, H., Xu, J.: Enhancing the effectiveness and durability of backdoor attacks in federated learning through maximizing task distinction. IEEE Transactions on Mobile Computing 24(2), 512–525 (2025)
15. Gu, T., Dolan-Gavitt, B., Garg, S.: BadNets: Identifying vulnerabilities in the machine learning model supply chain. IEEE Access 7, 47230–47244 (2019)
16. Zhang, Z., Gu, G., Bao, J., Zhu, D., Yu, F., Shen, E., Zhao, C.: Neurotoxin: Durable backdoors in federated learning. In: 39th International Conference on Machine Learning (ICML), pp. 26429–26446. PMLR (2022)

17. Nguyen, T.D., Nguyen, T., Tran, A., Doan, K.D., Wong, K.S.: IBA: Towards irreversible backdoor attacks in federated learning. In: Advances in Neural Information Processing Systems (NeurIPS), vol. 36, pp. 66364–66376. Curran Associates, Inc. (2023)
18. Blanchard, P., El Mghazli, E.M., Guerraoui, R., Stainer, J.: Machine learning with adversaries: Byzantine tolerant gradient descent. In: Advances in Neural Information Processing Systems (NeurIPS), vol. 30, pp. 119–129. Curran Associates, Inc. (2017)
19. Nguyen, T.D., Rieger, P., Mi, H., Yalame, H., Möllering, H., Fereidooni, H., Marchal, S., Miettinen, M., Mirhoseini, A., Sadeghi, A.R.: FLAME: Taming backdoors in federated learning. In: 31st USENIX Security Symposium (USENIX Security 22), pp. 1415–1432. USENIX Association (2022)
20. Yan, G., Li, T., Wang, J., Liu, S., Li, Y., Zhang, X.: FedRoLA: Robust federated learning against model poisoning via layer-based aggregation. In: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 1–12. ACM, New York (2024)