RESIDUAL-DIRECTION FILTERING FOR BACKDOOR DEFENSE IN FEDERATED LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Malicious clients in federated learning can implant backdoors through poisoned updates, and server-side defenses must filter such updates without clean data or knowledge of the attackers. Detection-based defenses increasingly score clients on parameter subsets chosen by dedicated mechanisms, such as importance ranking or critical-layer selection, yet it is unknown when scoring a subset yields the same filtering decisions as scoring all parameters. We propose CoPR, which scores each client on a fixed parameter subset by the direction of its cross-client mean-centered residual relative to the coordinate-wise median residual, filters clients with a median-absolute-deviation threshold, and aggregates the clipped full updates of retained clients. We give sufficient conditions under which residual directions separate malicious from benign clients, show that a parameter subset preserves every filtering decision whose margin exceeds the subset-induced score perturbation, and bound the single-round aggregation error under missed detections, false rejections, and clipping. Consistent with this criterion, subsets holding only 0.78% of the parameters match full observation on federated diffusion models at every tested location with 13–54$\times$ lower detection time, whereas non-IID classification requires full observation. Across five attacks on CIFAR-10 diffusion models, CoPR attains the lowest average FID while suppressing private-image reconstruction as well as AlignIns and remains effective against a subset-aware variant of AdaSCP; in non-IID classification it reduces the average attack success rate to 6.86% at 82.78% clean accuracy.

1 INTRODUCTION

Federated learning (FL) trains a shared model by aggregating client updates while keeping training data local (McMahan et al., 2017). However, malicious clients can manipulate local training or uploaded updates to implant backdoors that make the global model produce attacker-specified outputs on triggered inputs (Bagdasaryan et al., 2020). Server-side backdoor defenses must therefore mitigate the influence of malicious updates without access to clients' training data while minimizing degradation in the global model's main-task performance (Xu et al., 2025).

Detection-based defenses compute discriminative statistics from uploaded parameters or updates to filter clients or reweight their contributions (Han et al., 2023; Xu et al., 2025). For example, AlignIns combines update-direction alignment with the sign alignment of large-magnitude coordinates, FedCPA compares the parameter sets at both ends of an importance ranking, and FedSurrogate selects layers by cross-client disagreement in layer-wise update directions (Xu et al., 2025; Han et al., 2023; Abacha et al., 2026). These defenses thus pair a detection score with a dedicated mechanism for choosing which parameters to observe, yet it is unknown when scoring a subset yields the same filtering decisions as scoring all parameters. The question is nontrivial in federated diffusion training, where DataStealing exfiltrates private training images through the global model and its adaptive variant AdaSCP concentrates the poisoned update on a few critical parameters and rescales it to evade distance-based aggregation (Gan et al., 2024), so a subset that misses them might miss the attacker.

Figure 1 shows that a directional signal can survive restricted observation: on one late-layer tensor, the clients' raw parameter values nearly overlap, yet after cross-client mean-centering the attacker's cosine-deviation score is about twice that of every benign client.

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.

We study this question for a simple scoring rule based on residual directions and propose CoPR (Compact Parameter Residuals), which restricts scoring to a parameter subset to reduce server-side detection cost. In each round, CoPR scores cross-client mean-centered residuals by their cosine deviation from the coordinate-wise median residual, rejects clients whose scores exceed a median-absolute-deviation (MAD) threshold, and aggregates the retained clients' clipped full updates. Our analysis shows that the observation scope affects filtering only through a margin condition: a subset preserves a client's filtering decision whenever the client's margin from the threshold exceeds the score perturbation induced by the subset. Empirically, subsets holding 0.78% of the parameters, drawn from early, middle, late, or random locations, detect the attacker in every round on CIFAR-10 diffusion models and match full observation, whereas under non-IID classification, where label skew disperses benign residuals, the compact slice fails and full observation is required.

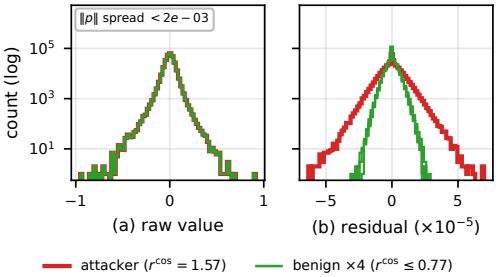


Figure 1: **Residual-direction signals on one late-layer tensor.** A CIFAR-10 diffusion example at round 100 with one DataStealing attacker and four benign clients. **(a)** Raw parameter-value distributions nearly overlap across clients. **(b)** Residual-coordinate distributions differ after cross-client mean-centering. The cosine-deviation scores in the legend are higher for the attacker (red) than for benign clients (green).

Our main contributions are as follows.

- **We propose CoPR, a server-side defense that filters clients by the directions of their cross-client parameter residuals.** Detection reads only a fixed parameter subset, while clipping and aggregation use the full updates; CoPR requires no clean reference data, trigger knowledge, or number of malicious clients.
- **We characterize when residual directions detect attackers and when a parameter subset suffices.** We give sufficient conditions for residual-direction detection in terms of the between-group direction, benign angular dispersion, and reference error; show that a subset preserves every filtering decision whose margin exceeds the subset-induced score perturbation; and bound the single-round aggregation error in terms of missed detections, false rejections, and clipping bias.
- **We verify the criterion empirically and delineate its boundary.** On federated diffusion models under five backdoor attacks, CoPR attains the lowest average FID while matching AlignIns in suppressing private-image reconstruction, and subsets holding 0.78% of the parameters match full observation regardless of their location and reduce defense runtime. In IID classification, CoPR reduces the average attack success rate to 2.22% at the highest average clean accuracy, whereas non-IID classification requires full observation.

2 Related Work

Backdoor attacks in federated learning. Model replacement amplifies malicious influence through update scaling (Bagdasaryan et al., 2020), DBA (Xie et al., 2020) enables coordinated attacks through distributed triggers, and Neurotoxin (Zhang et al., 2022) improves backdoor persistence through parameter selection. DataStealing extends backdoors to federated diffusion models, using combinatorial triggers to exfiltrate clients' training images via the global model (Gan et al., 2024).

Robust aggregation and client-update detection. Krum and Multi-Krum (Blanchard et al., 2017) select updates based on inter-update distances, whereas coordinate-wise median, trimmed mean, and RFA (Yin et al., 2018; Pillutla et al., 2022) use robust statistical estimators to limit the influence of anomalous updates. Other defenses exploit update directions, signs, and cross-client relationships: FoolsGold (Fung et al., 2020) downweights colluding attackers based on historical update similarity, FLTrust (Cao et al., 2021) constructs a trusted reference using clean server-side data, and RLR (Ozdayi et al., 2021) adjusts coordinate-wise learning rates according to sign agreement. AlignIns (Xu et al., 2025) jointly examines the directional alignment between updates and global model parameters and the consistency of large-magnitude update coordinates with cross-client majority signs.

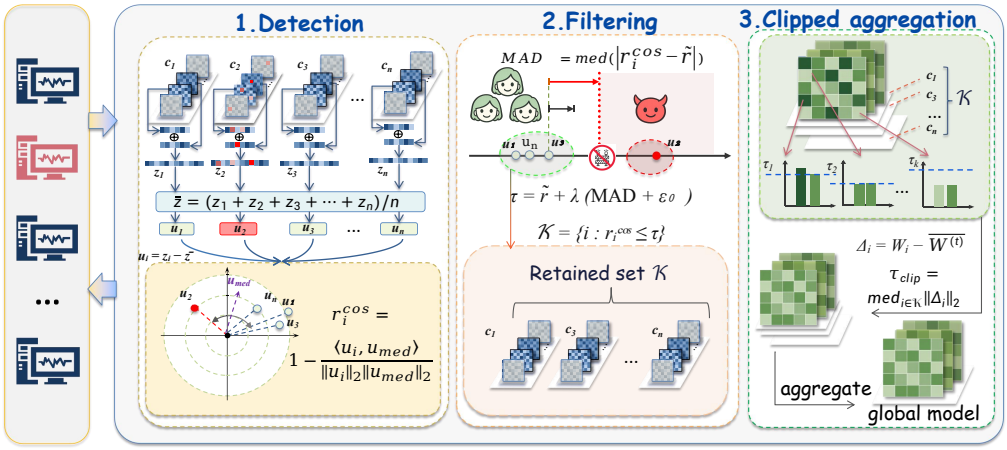


Figure 2: Overview of CoPR in one federated round. The server performs residual-direction detection, adaptive filtering, and clipped weighted aggregation.

Parameter-subset and representation-based defenses. FedCPA (Han et al., 2023) compares the consistency of parameter-importance structures, FL-Defender (Jebreel & Domingo-Ferrer, 2023) constructs angular-similarity representations from last-layer gradients, and FedSurrogate (Abacha et al., 2026) combines dynamic critical-layer selection, directional filtering, and parameter replacement. DnC (Shejwalkar & Houmansadr, 2021) scores centered updates through spectral projections on randomly sampled coordinates, while FLARE (Wang et al., 2022) evaluates client trustworthiness by comparing penultimate-layer representations obtained using auxiliary inputs.

3 Problem Setup and Threat Model

We consider federated learning for image generation or classification, with a trusted server and n clients indexed by $\mathcal{C} = \{1, \ldots, n\}$. Each client i holds a nonempty private dataset $\mathcal{D}_i$. In each communication round t, the server broadcasts the global parameters $\bar{W}^{(t)} \in \mathbb{R}^{D_a}$ to all clients, and each client i updates them locally using $\mathcal{D}_i$ and returns $W_i \in \mathbb{R}^{D_a}$ without uploading raw data. The aggregation parameters are flattened in a fixed order into D_a-dimensional vectors. Without a defense, the server applies data-size-weighted FedAvg (McMahan et al., 2017):

$$\bar{W}^{(t+1)} = \sum_{i \in \mathcal{C}} p_i W_i, \qquad p_i = \frac{|\mathcal{D}_i|}{\sum_{j \in \mathcal{C}} |\mathcal{D}_j|}. \tag{1}$$

The adversary aims to implant trigger-dependent behavior in the global model while preserving normal task performance. For image generation, it seeks to make the diffusion model generate attacker-specified target images when triggered. In the data-exfiltration setting, these targets are access-restricted training images from compromised clients, which the adversary aims to recover through the released global model, as in DataStealing (Gan et al., 2024). For classification, it seeks to make the global model misclassify triggered inputs as an attacker-chosen target class (Bagdasaryan et al., 2020). The adversary fully controls malicious clients' local data, training procedures, and uploaded parameters, but cannot access benign clients' private data or individual uploads, compromise the server, or forge additional client identities. It observes the broadcast global parameters and knows the model architecture and training protocol, but not the defense or its hyperparameters.

4 Method

In each round, CoPR scores clients by the directions of compact parameter residuals, filters them adaptively, and aggregates the clipped full updates of the retained clients (Figure 2).

Stage 1: Detection. Let $x_i \in \mathbb{R}^D$ denote the task-specific scoring-parameter vector extracted from W_i. Diffusion models are scored using the exponential moving average (EMA) parameters used

for sampling, whereas classifiers use their corresponding model parameters. To construct compact residuals, CoPR selects a subset of coordinates from x_i for residual construction and directional scoring. The selected coordinates are shared across clients and remain fixed throughout training.

Let $S \subseteq \{1, \ldots, D\}$ denote this nonempty coordinate set. The corresponding selection matrix $P_S \in \mathbb{R}^{|S| \times D}$ extracts these coordinates from x_i, yielding the observed parameter vector

$$z_i = P_S x_i \in \mathbb{R}^{|S|}. \tag{2}$$

This selection restricts only the parameters used for detection; malicious clients may also modify parameters outside the selected subset. The CoPR-full variant uses $S = \{1, \ldots, D\}$, so that $z_i = x_i$, while leaving all subsequent steps unchanged.

The server centers the observations across clients:

$$u_i = z_i - \bar{z}, \qquad \bar{z} = \frac{1}{n} \sum_{j \in \mathcal{C}} z_j. \tag{3}$$

Each residual u_i captures client i's deviation from the current-round mean, and the residuals satisfy $\sum_{i \in \mathcal{C}} u_i = 0$. Data-size weighting could allow a malicious client i with a large local dataset $|\mathcal{D}_i|$ to disproportionately shift $\bar{z}$ toward its observed parameter vector. CoPR therefore assigns equal weight to each client during centering, rather than using the FedAvg weights p_i.

After centering, CoPR measures how far each residual deviates in direction from u_{med}, the coordinate-wise median of all residuals; for nonzero u_i and u_{med}, the cosine anomaly score is

$$r_i^{\cos} = 1 - \frac{\langle u_i, u_{\text{med}} \rangle}{\|u_i\|_2 \, \|u_{\text{med}}\|_2}. \tag{4}$$

The score ranges from 0 for aligned to 2 for opposite directions.

Stage 2: Filtering. CoPR uses the distribution of directional scores to determine which clients to retain, without requiring the number of malicious clients. The server computes the score median $\tilde{r}$ and median absolute deviation MAD from all clients' directional scores (Leys et al., 2013):

$$\tilde{r} = \text{med}_{i \in \mathcal{C}} \, r_i^{\cos}, \qquad \text{MAD} = \text{med}_{i \in \mathcal{C}} \, |r_i^{\cos} - \tilde{r}|, \tag{5}$$

where med denotes the scalar median operator.

For a threshold coefficient $\lambda \geq 0$ and a numerical stabilizer $\varepsilon_0 \geq 0$, define the filtering threshold τ and retained set $\mathcal{K}$ as

$$\tau = \tilde{r} + \lambda(\text{MAD} + \varepsilon_0), \qquad \mathcal{K} = \{i \in \mathcal{C} : r_i^{\cos} \leq \tau\}. \tag{6}$$

We use $\lambda = 2$. For finite, well-defined scores, $\tau \geq \tilde{r}$ ensures that at least $\lceil n/2 \rceil$ clients are retained. Appendix A.2 specifies ε_0, the median convention for even sample sizes, and the handling of numerical edge cases.

Stage 3: Clipped aggregation. Directional filtering does not constrain update magnitudes, so retained clients may still have large updates. Inspired by centered clipping (Karimireddy et al., 2021), CoPR therefore clips their full updates rather than only the parameters selected for detection. Let $\Delta_i = W_i - \bar{W}^{(t)} \in \mathbb{R}^{D_a}$ denote client i's full update relative to the broadcast global parameters. The clipping radius τ_{clip} is set to the median norm of the retained updates:

$$\tau_{\text{clip}} = \text{med}_{i \in \mathcal{K}} \, \|\Delta_i\|_2. \tag{7}$$

For each $i \in \mathcal{K}$, CoPR applies radial clipping:

$$C_{\tau_{\text{clip}}}(\Delta_i) = \begin{cases} \Delta_i \min\left\{1, \dfrac{\tau_{\text{clip}}}{\|\Delta_i\|_2}\right\}, & \Delta_i \neq 0, \\ 0, & \Delta_i = 0. \end{cases} \tag{8}$$

The server aggregates the clipped updates with the FedAvg weights renormalized over $\mathcal{K}$ and updates the global parameters:

$$\begin{aligned} \widehat{p}_i &= \frac{p_i}{\sum_{j \in \mathcal{K}} p_j}, \quad i \in \mathcal{K}, \\ \widehat{\Delta} &= \sum_{i \in \mathcal{K}} \widehat{p}_i \, C_{\tau_{\text{clip}}}(\Delta_i), \qquad \bar{W}^{(t+1)} = \bar{W}^{(t)} + \widehat{\Delta}. \end{aligned} \tag{9}$$

5 THEORETICAL ANALYSIS

Using $\mathcal{B}$ and $\mathcal{A}$ be the benign and malicious client sets, and define the decision margin $\Gamma_i = r_i^{\cos} - \tau$; rejection occurs when $\Gamma_i > 0$. We analyze one round with scoring residuals and reference vectors. Detailed proofs are in Appendices B–D.

5.1 DETECTABILITY FROM RESIDUAL DIRECTIONS

Let $\rho = |\mathcal{A}|/n \in (0, 1/2)$ and let μ_B, μ_A be the equally weighted means of benign and malicious observations. Define $a = \mu_A - \mu_B \neq 0$, $v = a/\|a\|_2$, and the median-location error $e = u_{\text{med}} + \bar{z} - \mu_B$. Centering gives the two groups opposite between-group terms, $-\rho a$ and $(1-\rho)a$, while $-u_{\text{med}} = \rho a - e$ (Appendix B). The reference-direction error $\eta = \|u_{\text{med}}/\|u_{\text{med}}\|_2 + v\|_2$ satisfies $\eta \leq 2\|e\|_2/(\rho\|a\|_2)$ when $\|e\|_2 < \rho\|a\|_2$; the benign majority bounds $\|e\|_2$ through its order-statistic spread (Appendix B).

Theorem 1 (Conditional detection)**.** *Suppose there exists $\mathcal{G} \subseteq \mathcal{B}$ with $|\mathcal{G}| > n/2$ such that, for every $j \in \mathcal{G}$, the angle between u_j and $-v$ lies in $[\theta_-, \theta_+]$, where $0 \leq \theta_- \leq \theta_+ < \pi/2$. Then every malicious client whose residual forms an angle $\phi_i \in [0, \pi/2)$ with v satisfies*

$$\begin{aligned}\Gamma_i \geq \underline{\Gamma}_i := \cos\phi_i + \cos\theta_+ - \lambda(\cos\theta_- - \cos\theta_+) \\ - 2(1+\lambda)\eta - \lambda\varepsilon_0.\end{aligned} \tag{10}$$

Therefore, $\underline{\Gamma}_i > 0$ is sufficient to reject client i.

The bound links detectability to malicious alignment with v, benign angular concentration around $-v$, and reference accuracy, without requiring small benign-update norms. Attackers outside this condition and benign false rejections are accounted for in Section 5.3.

5.2 DECISION STABILITY UNDER COMPACT OBSERVATIONS

For the same uploads, use superscripts S and full for CoPR and CoPR-full quantities, and define $\xi_S = \max_i |r_i^{\cos,S} - r_i^{\cos,\text{full}}|$.

Theorem 2 (Decision stability)**.** *If the two observations use the same median convention, λ, and ε_0, then*

$$|\Gamma_i^S - \Gamma_i^{\text{full}}| \leq 2(1+\lambda)\xi_S. \tag{11}$$

Hence, any client satisfying $|\Gamma_i^{\text{full}}| > 2(1+\lambda)\xi_S$ retains the same filtering decision under compact observation. In particular, if $\underline{\Gamma}_i^{\text{full}} > 2(1+\lambda)\xi_S$, an attacker satisfying the sufficient condition in Theorem 1 remains rejected after restricting the score to the parameter subset.

A subset therefore need not approximate the full parameter space; it only has to perturb the scores by less than the margins. Theorem 2 compares the two scopes and does not assume that full-parameter decisions are correct. For randomly selected subsets, Proposition C.1 bounds the required budget: if S is formed by sampling k of J parameter blocks uniformly at random, then $\xi_S \leq 2\epsilon/(1-\epsilon)$ with probability at least $1-\delta$ once $k \geq 9\,\kappa_{\text{blk}}\epsilon^{-2}\log(n/\delta)$, where $\kappa_{\text{blk}} \in [1, J]$ measures how concentrated the residual energy is across blocks. A smaller margin requires a smaller ϵ and hence a larger budget, while the choice of layers does not enter. Table 4 includes such random subsets.

5.3 AGGREGATION ERROR UNDER IMPERFECT FILTERING

Filtering may retain attackers or reject benign clients. Take $\bar{\Delta}_B = P_B^{-1}\sum_{i\in\mathcal{B}} p_i\Delta_i$ as the benign reference, where $P_B = \sum_{i\in\mathcal{B}} p_i$. Define the retained malicious weight fraction α, rejected benign weight fraction γ, benign dispersion V_B, mean clipping magnitude of retained benign updates B_{clip}:

$$\alpha = \frac{\sum_{i\in\mathcal{A}\cap\mathcal{K}} p_i}{P_K}, \qquad \gamma = 1 - \frac{P_{BK}}{P_B},$$

$$V_B = \frac{1}{P_B}\sum_{i\in\mathcal{B}} p_i\|\Delta_i - \bar{\Delta}_B\|^2, \quad B_{\text{clip}} = \frac{1}{P_{BK}}\sum_{i\in\mathcal{B}\cap\mathcal{K}} p_i(\|\Delta_i\| - \tau_{\text{clip}})_+,$$

where $(x)_+ = \max\{x, 0\}$.

Theorem 3 (Weighted aggregation error)**.** *For any realized retained set and clipping radius, if* $\gamma < 1$*, so that positive benign weight remains, then*

$$\|\widehat{\Delta} - \bar{\Delta}_B\| \le 2\alpha\tau_{\text{clip}} + \sqrt{\frac{\gamma}{1-\gamma}V_B} + B_{\text{clip}}. \tag{12}$$

The second term is zero when $\gamma = 0$*. No directional separation or random observation is needed.*

The first term is the contribution of attackers that survive filtering, the second is the shift caused by rejecting benign clients, and the third is the distortion caused by clipping benign updates. The radius cannot be inflated by retained attackers as long as benign clients form a count majority of $\mathcal{K}$, in which case $\tau_{\text{clip}} \le \max_{i\in\mathcal{B}\cap\mathcal{K}} \|\Delta_i\|$ (Appendix D). Attackers that satisfy the conditions of Theorems 1 and 2 are rejected and contribute nothing to α (Appendix D.3). Theorems 1–3 are deterministic given the uploads of a round and assume neither i.i.d. clients nor a particular training trajectory, so they apply to every round of training.

6 Experiments

6.1 Experimental Settings

Tasks. We evaluate CoPR on federated diffusion generation and federated image classification. For diffusion, we use U-Net-based models on CIFAR-10 (Krizhevsky, 2009) and CelebA (Liu et al., 2015) at resolutions of 32×32 and 64×64, respectively. Following the federated evaluation protocol of DataStealing (Gan et al., 2024), we consider data poisoning (Data) (Gu et al., 2017), model poisoning (Model) (Bagdasaryan et al., 2020), projected gradient descent poisoning (PGD) (Wang et al., 2020), backdoor-critical-layer substitution (BC) (Zhuang et al., 2024), and adaptive parameter scaling (AdaSCP) (Gan et al., 2024). Unless otherwise specified, diffusion experiments use five clients, including one malicious client, with Dirichlet non-IID partitions ($\beta = 0.7$). We jointly assess normal generation quality using FID (Heusel et al., 2017) (lower is better) and private-target reconstruction error using Stolen-MSE (higher is better). For classification, we use ResNet9 on CIFAR-10 under both independent and identically distributed (IID) and non-IID (Dirichlet $\beta = 0.5$) partitions, evaluating BadNet (Gu et al., 2017), DBA (Xie et al., 2020), Neurotoxin (Zhang et al., 2022), and PGD (Wang et al., 2020). The main 100-round classification comparison uses 20 clients, including four malicious clients, with a local poisoning ratio of 0.3 at each malicious client. We report clean accuracy (ACC), attack success rate (ASR), and original-label accuracy on triggered samples (RA); higher ACC and RA and lower ASR are preferred. See Appendix A.1. **Baselines.** We compare with FedAvg (McMahan et al., 2017) (undefended), Multi-Krum (Blanchard et al., 2017), RFA (Pillutla et al., 2022), FoolsGold (Fung et al., 2020), Multi-Metrics (Huang et al., 2023a), and AlignIns (Xu et al., 2025) on both tasks, plus Krum (Blanchard et al., 2017) for diffusion and RLR (Ozdayi et al., 2021) and Lockdown (Huang et al., 2023b) for classification, which uses the official AlignIns framework; baseline settings are given in Appendix A.1.

6.2 Main Results and Robustness

Table 1: Defense effectiveness against five attacks on CIFAR-10 diffusion models.

	Data		Model		PGD		BC		AdaSCP		Average	
Defense	FID	Stolen-MSE	FID	Stolen-MSE	FID	Stolen-MSE	FID	Stolen-MSE	FID	Stolen-MSE	FID	Stolen-MSE
FedAvg	6.82	12.30	12.83	0.67	6.83	12.39	33.70	4.55	12.79	1.30	14.59	6.24
Krum	8.66	**14.59**	8.60	14.60	16.97	12.54	8.54	15.09	36.68	4.74	15.89	12.31
Multi-Krum	6.23	14.26	6.23	14.26	6.61	12.03	6.15	14.30	8.31	12.70	6.71	13.51
FoolsGold	7.41	12.38	459.42	**31.22**	7.42	12.42	196.24	**28.34**	21.00	1.40	138.30	**17.15**
RFA	6.64	12.36	**6.10**	12.06	6.72	12.34	**5.17**	12.58	8.72	12.50	6.67	12.37
Multi-Metrics	6.69	12.21	31.81	2.73	6.99	12.18	66.13	13.60	12.95	4.50	24.91	9.04
AlignIns	6.75	14.07	6.19	14.24	6.81	14.08	6.19	14.24	6.25	14.20	6.44	14.17
CoPR (ours)	**6.18**	14.25	6.18	14.24	**6.17**	**14.24**	6.17	14.25	**6.22**	**14.24**	**6.18**	14.24

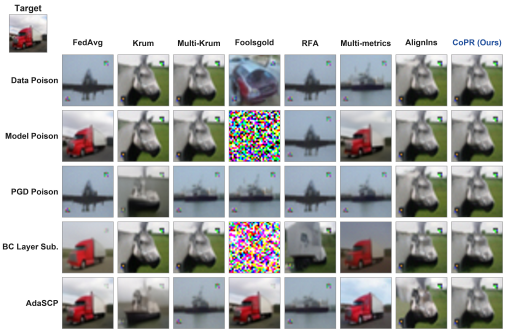


Figure 3: Qualitative privacy comparison on CIFAR-10.

Table 2: Cross-dataset and partition robustness under AdaSCP.

Defense	Non-IID CIFAR-10 FID↓/ Stolen-MSE↑	Non-IID CelebA FID↓/ Stolen-MSE↑	IID CIFAR-10 FID↓/ Stolen-MSE↑	IID CelebA FID↓/ Stolen-MSE↑
no-defense	12.79/1.30	5.78/16.40	8.73/11.27	8.85/14.10
Krum	35.38/4.70	11.32/0.90	9.66/12.08	12.05/15.60
Multi-Krum	8.31/12.70	**5.90**‡/16.50	8.37/12.86	**6.33**‡/15.60
AlignIns	6.25/14.20	8.84/18.80	5.52/**14.21**	8.24/18.66
CoPR (Ours)	**6.22/14.24**	8.69/**19.27**	**5.48**/14.18	8.31/**18.72**

Across-attack performance on diffusion models. Table 1 shows that CoPR, despite scoring residuals only on a parameter subset, consistently increases Stolen-MSE and reduces FID relative to FedAvg under all five attacks on CIFAR-10. Compared with AlignIns, CoPR achieves a similar level of reconstruction suppression while better preserving generation quality, thereby obtaining the lowest average FID across attacks. Figure 3 qualitatively shows that the reconstructions produced by both CoPR and AlignIns visibly deviate from the target, whereas FoolsGold generates noise-like outputs under the Model and BC attacks, accompanied by substantial FID degradation.

Table 3: Defense effectiveness against four attacks on CIFAR-10 classification models.

	BadNet			DBA			Neurotoxin			PGD			Average		
Defense	ACC	ASR	RA	ACC	ASR	RA	ACC	ASR	RA	ACC	ASR	RA	ACC	ASR	RA
(a) IID partition															
FedAvg	88.23±0.19	32.15±12.20	61.56±10.36	**88.52**±0.19	39.79±3.31	54.51±2.69	87.74±0.16	16.98±6.06	72.50±4.88	**88.65**±0.29	33.53±4.17	61.57±3.61	88.29	30.61	62.53
RLR	78.33±0.55	3.66±0.73	76.41±0.42	77.91±0.97	4.42±1.30	75.44±1.08	72.95±0.79	6.45±0.72	70.28±1.09	77.90±0.19	3.70±0.29	75.82±0.05	76.77	4.56	74.49
Multi-Krum	86.72±0.05	72.38±1.15	25.66±1.03	87.24±0.26	2.33±0.25	84.69±0.28	85.99±0.29	66.52±3.10	30.01±2.56	87.02±0.19	70.31±8.50	27.51±7.41	86.74	52.89	41.97
Multi-Metrics	87.89±0.50	38.39±15.13	55.81±12.87	87.08±0.85	49.19±12.21	46.27±10.30	87.35±0.25	37.39±7.35	55.63±5.98	87.51±0.28	47.39±14.49	48.04±12.40	87.46	43.09	51.44
Lockdown	87.90±0.38	6.55±1.66	81.71±1.28	88.00±0.23	6.76±1.28	80.86±1.23	84.82±0.79	3.72±0.15	81.13±0.80	87.91±0.22	6.87±2.29	81.37±1.91	87.16	5.98	81.27
FoolsGold	88.47±0.09	27.19±3.91	66.36±2.85	88.51±0.23	39.73±5.89	54.74±4.60	87.54±0.24	2.57±0.21	84.83±0.07	88.48±0.08	34.92±0.93	60.09±0.56	88.25	26.10	66.50
RFA	82.08±3.27	80.83±12.48	18.34±11.72	83.39±1.34	82.55±6.73	16.57±6.03	82.80±0.42	79.67±10.57	18.26±8.83	81.20±1.34	86.21±3.57	13.36±3.40	82.37	82.32	16.63
AlignIns	87.92±0.07	2.34±0.15	85.24±0.11	87.75±0.05	2.66±0.21	85.02±0.44	88.03±0.25	2.40±0.15	85.17±0.44	88.37±0.10	2.27±0.15	85.43±0.29	88.02	2.42	85.22
CoPR (ours)	**88.65**±0.10	**2.23**±0.08	**85.93**±0.25	88.05±0.33	**2.27**±0.16	**85.57**±0.34	**88.17**±0.35	**2.30**±0.09	**85.67**±0.45	88.48±0.22	**2.10**±0.07	**86.07**±0.09	**88.34**	**2.22**	**85.81**
(b) Non-IID partition (Dirichlet $\beta = 0.5$)															
FedAvg	84.75±0.30	53.37±11.07	41.59±9.58	84.83±0.41	50.65±21.23	42.83±16.30	83.56±0.56	37.76±10.21	53.06±7.56	84.75±0.30	47.30±2.52	46.36±2.31	84.47	47.27	45.96
RLR	54.89±1.99	**2.86**±2.78	53.76±0.95	58.23±0.10	2.49±3.46	58.09±2.61	42.15±1.90	**0.33**±0.31	44.46±1.23	56.51±2.51	**0.95**±0.71	57.37±2.35	52.94	**1.65**	53.42
Multi-Krum	78.56±1.57	49.83±43.51	40.20±31.34	78.02±1.25	2.60±1.85	75.36±1.98	77.88±0.98	41.77±36.07	46.24±26.26	78.44±1.24	48.79±42.09	41.87±30.33	78.22	35.75	50.92
Multi-Metrics	74.60±4.39	63.84±18.91	28.29±11.97	72.85±3.94	74.96±26.65	19.69±18.40	73.81±2.44	79.77±6.45	16.92±5.50	76.09±4.11	65.72±19.85	27.40±15.26	74.34	71.07	23.07
Lockdown	82.84±0.41	12.00±3.19	71.94±2.06	82.65±0.37	14.12±9.30	70.19±6.32	75.89±1.07	2.85±1.52	72.52±2.14	82.84±0.41	12.00±3.19	71.94±2.06	81.06	10.24	71.65
FoolsGold	**84.24**±0.59	48.41±12.03	44.93±9.81	**84.15**±0.27	58.40±10.78	36.73±8.65	82.05±1.36	14.23±20.32	70.07±13.84	**84.38**±0.75	44.20±4.27	49.30±3.69	**83.70**	41.31	50.26
RFA	75.84±5.02	63.87±0.86	29.00±4.07	73.26±8.98	53.81±3.61	32.84±7.33	75.12±4.17	62.66±12.96	27.93±5.65	71.61±7.79	49.34±17.77	36.16±12.37	73.96	57.42	31.48
AlignIns	80.03±1.94	48.10±25.06	43.26±17.71	81.09±0.45	33.43±29.92	52.56±20.88	**83.04**±0.82	2.53±1.12	79.88±0.66	80.92±0.72	37.39±22.95	51.89±16.98	81.27	30.36	56.90
CoPR (ours)	81.51±3.21	50.10±4.34	43.12±2.82	81.30±3.92	47.99±11.79	43.99±9.74	81.62±1.96	16.82±26.00	68.52±20.13	81.35±3.23	44.88±6.19	46.86±3.60	81.44	39.95	50.62
CoPR-full (ours)	82.42±0.64	12.21±18.09	**72.68**±13.14	82.99±0.42	**2.04**±0.47	**79.99**±1.26	82.98±1.38	2.17±0.61	**80.03**±0.51	82.72±0.42	11.00±16.14	**74.23**±11.43	82.78	6.86	**76.73**

Cross-dataset and partition robustness of diffusion models. Table 2 compares two datasets under IID and non-IID partitions with AdaSCP. CoPR increases Stolen-MSE relative to FedAvg in all four settings, but the generation-quality trade-off depends on the setting: FID improves under both CIFAR-10 partitions, whereas stronger reconstruction suppression on non-IID CelebA comes with a FID cost. Relative to AlignIns, CoPR improves both metrics on non-IID CelebA, but yields slightly

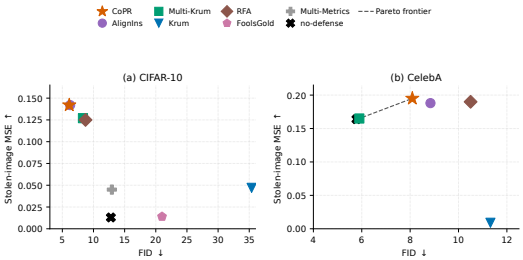


Figure 4: Trade-off under AdaSCP. Lower FID and higher Stolen-MSE are preferred.

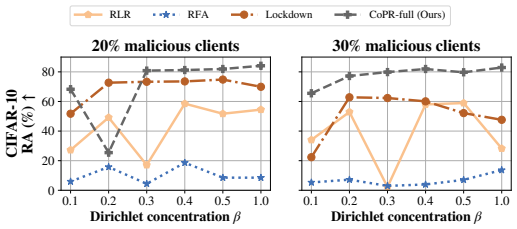


Figure 5: RA of CoPR-full under varying Dirichlet concentration β, compared with Lockdown, RFA, and RLR under DBA on CIFAR-10.

higher reconstruction error and slightly higher FID on IID CelebA. Figure 4 illustrates the privacy–utility trade-off under AdaSCP. On CelebA, Multi-Krum's lower FID is accompanied by more accurate target reconstruction, whereas CoPR suppresses target reconstruction more effectively while preserving generation quality.

Defense performance in IID classification. Table 3(a) shows that under the IID partition, CoPR maintains an average clean accuracy of 88.34% and reduces the average ASR to 2.22% across four attacks. Its clean accuracy is comparable to that of FedAvg, while it achieves the lowest mean ASR and highest mean RA among the evaluated defenses for each attack. CoPR thus also preserves the correct labels of triggered samples in the IID setting.

Trade-offs in non-IID classification. Under the non-IID partition, the default compact slice largely fails to suppress the backdoors: CoPR's average ASR is 39.95%, close to 47.27% for undefended FedAvg, whereas CoPR-full reduces it to 6.86%. This is consistent with Theorem 2: label skew disperses benign residuals, most strongly in the classifier layer, which shrinks the detection margin and thus requires a smaller observation-induced perturbation ξ_S, i.e., more observed coordinates. CoPR-full achieves the highest mean RA under all four attacks and the lowest mean ASR under DBA. The residual ASR of CoPR-full under BadNet and PGD, with large seed-to-seed variance (12.21 ± 18.09 and 11.00 ± 16.14), shows that backdoor influence is not fully eliminated in this setting.

Effectiveness under varying non-IID degrees. Figure 5 reports the RA of CoPR-full under DBA on CIFAR-10 with 20% and 30% attackers, compared with Lockdown, RFA, and RLR, as β varies from 0.1 (most heterogeneous) to 1.0. CoPR-full achieves the highest RA in 11 of the 12 settings. For $\beta \geq 0.3$, CoPR-full maintains high and stable RA at both malicious-client fractions. In comparison, Lockdown's performance declines when the malicious-client fraction increases to 30%, while RFA and RLR generally achieve lower RA. Even under the highly heterogeneous setting of $\beta = 0.1$, CoPR-full outperforms the other methods at both malicious-client fractions. The exception is $\beta = 0.2$ with 20% attackers, where CoPR-full fails to filter the DBA attackers (ASR 69.46%, RA 25.44%), while Lockdown retains an RA of 72.69%.

6.3 Ablation and Mechanism Analysis

Score separation and filtering behavior. Figure 6 shows the score and threshold trajectories of the default CoPR configuration on CIFAR-10 under AdaSCP. Throughout all 300 rounds, the malicious score stays above both the benign range and the threshold, so its decision margin Γ_i in Theorem 1 remains positive in every round. In contrast, the largest benign score occasionally crosses the threshold (Figure 6(b)), giving an FPR of 3.83%; in Theorem 3, such false rejections enter through the benign-removal term γ rather than through malicious contamination.

Observation scope and runtime. Table 4 varies the observation scope; each compact configuration observes 0.78% of the scoring coordinates. All configurations detect the attacker in every round (FNR 0.00%) and match full observation in FID and Stolen-MSE, whether the subset lies in early, middle, or late layers or is sampled at random. This agrees with Theorem 2: the malicious margin is large (Figure 6(a)), so every subset preserves the attacker's rejection. Benign clients near the threshold have small margins, which explains the variation in FPR (2.25–3.83%). Detection takes 96.55–392.87 ms for compact subsets versus 5215.81 ms for CoPR-full.

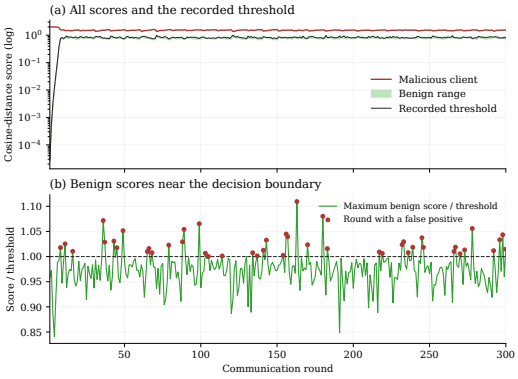


Figure 6: Score separation and filtering on CIFAR-10.

Table 4: Observation scope and parameter location on CIFAR-10 under AdaSCP.

	Generation		Detection (%)			Detection
Representation	**FID**	**Stolen-MSE**	**DACC**	**FPR**	**FNR**	**time (ms)**
Early	6.23	14.26	97.07	3.67	**0.00**	96.55
Middle	6.19	**14.27**	97.93	2.58	**0.00**	107.78
Default late	6.22	14.24	96.93	3.83	**0.00**	N/A
Adjacent late	**6.18**	14.25	97.60	3.00	**0.00**	98.69
Random (0)	6.28	14.25	**98.20**	**2.25**	**0.00**	375.96
Random (1)	6.24	14.25	97.47	3.17	**0.00**	191.57
Random (2)	6.21	**14.27**	97.87	2.67	**0.00**	392.87
Full	6.28	**14.27**	97.20	3.50	**0.00**	5215.81

Table 5: AdaSCP with known S (CIFAR-10); $\rho = 1$ reuses the original attack results.

	$\rho = 1$				$\rho = 0.5$				$\rho = 0$			
Defense	FID↓	Stolen-MSE↑	TPR↑	FPR↓	FID↓	Stolen-MSE↑	TPR↑	FPR↓	FID↓	Stolen-MSE↑	TPR↑	FPR↓
FedAvg	12.85	1.17	—	—	8.37	12.75	—	—	10.87	13.13	—	—
CoPR (ours)	**6.22**	**14.24**	100.00	3.83	**6.22**	**14.25**	100.00	3.67	**6.19**	**14.24**	100.00	4.17

Attacks with knowledge of the observation set. Theorem 2 establishes that compact observations preserve full-observation filtering decisions when the induced perturbation in scores and thresholds is smaller than the corresponding decision margin. Table 4 shows comparable defense performance across early, middle, late, and randomly sampled but fixed slices, indicating that effective detection is not confined to a particular observation location. We consider attackers who know the fixed observation set S: building on AdaSCP, they multiply the local update on S by $\rho \in \{0.5, 0\}$ and rescale updates outside S for norm compensation. As shown in Table 5, CoPR detects malicious uploads in all 300 rounds under both settings, while the suppression also substantially weakens the attack without defense. CoPR thus withstands the tested observation-set-aware suppression strategies, motivating a possible extension in which the server randomly selects S_t after all uploads in the current round are finalized. Attackers could then know the detection principle and sampling rule without knowing the current observation set before submitting their updates.

Components. Table 6 isolates filtering and clipping. Adding filtering to clipping alone lowers FID from 8.89 to 6.22 and raises Stolen-MSE from 12.71 to 14.24. Removing clipping leaves both metrics essentially unchanged. Theorem 3 explains this: clipping limits the contribution of attackers that survive filtering through the term $2\alpha\tau_{\text{clip}}$, which vanishes here because no attacker is retained. The remaining clipping bias B_{clip} is consistent with the slight difference in FID. Clipping thus acts as a safeguard against missed detections; the CelebA configurations in Appendix A.5 differ in their observation slices and do not isolate its effect.

Table 6: Component ablations results.

Variant	FID↓	Stolen-MSE↑
CoPR (ours)	6.22	14.24
w/o filtering	8.89	12.71
w/o clipping	6.19	14.25

7 Conclusion

We presented CoPR, a server-side backdoor defense that filters federated clients by the directions of their cross-client parameter residuals. Our analysis shows that a parameter subset preserves every filtering decision whose margin exceeds the score perturbation it induces, which separates where the server observes from how much it must observe. Consistent with this criterion, 0.78% subsets from every tested location match full observation with shorter detection times on federated diffusion models, where CoPR attains the lowest average FID while suppressing private-image reconstruction, whereas non-IID classification requires full observation. Adapting the observation scope to the detection margin and evaluating attacks that target the residual-direction score are natural next steps.

AI USE STATEMENT

We used AI tools to assist with drafting and polishing the manuscript and to identify relevant prior work. The authors take responsibility for the final content.

ETHICS STATEMENT

This work develops CoPR to mitigate backdoor attacks and attack-induced leakage of training images in federated learning. We evaluate existing attacks and defenses on CIFAR-10 and CelebA under simulated federated settings. The attack mechanisms and reconstruction examples are studied for defensive evaluation and may pose misuse risks. We discourage attempts to reconstruct or disclose training images without authorization.

REPRODUCIBILITY STATEMENT

Section 4 specifies CoPR's residual construction, directional scoring, adaptive filtering, and clipped aggregation, while Section 6 describes the evaluation tasks, attacks, baselines, and metrics. Appendix A.1 documents the data partitions, model and training configurations, attack and baseline settings, random seeds, and evaluation protocols. Appendix A.2 details the implementation and numerical conventions, including differences across experimental configurations. Appendix A also reports observation-scope comparisons, ablations, threshold sensitivity, and detection results. Appendices B–D provide the theoretical assumptions and proofs.

REFERENCES

Fatima Z. Abacha, Sin G. Teo, Yuanxiang Wu, Lucas C. Cordeiro, and Mustafa A. Mustafa. FedSurrogate: Backdoor defense in federated learning via layer criticality and surrogate replacement. *arXiv preprint arXiv:2605.11122*, 2026. doi: 10.48550/arXiv.2605.11122. URL https://arxiv.org/abs/2605.11122.

Eugene Bagdasaryan, Andreas Veit, Yiqing Hua, Deborah Estrin, and Vitaly Shmatikov. How to backdoor federated learning. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 108 of *PMLR*, pp. 2938–2948, 2020.

Rémi Bardenet and Odalric-Ambrym Maillard. Concentration inequalities for sampling without replacement. *Bernoulli*, 21(3):1361–1385, 2015.

Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, and Julien Stainer. Machine learning with adversaries: Byzantine tolerant gradient descent. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Xiaoyu Cao, Minghong Fang, Jia Liu, and Neil Zhenqiang Gong. Fltrust: Byzantine-robust federated learning via trust bootstrapping. In *Proceedings of the 28th Annual Network and Distributed System Security Symposium (NDSS)*, 2021.

Clement Fung, Chris J. M. Yoon, and Ivan Beschastnikh. The limitations of federated learning in sybil settings. In *Proceedings of the 23rd International Symposium on Research in Attacks, Intrusions and Defenses (RAID)*, pp. 301–316, 2020.

Yuan Gan, Jiaxu Miao, and Yi Yang. Datastealing: Steal data from diffusion models in federated learning with multiple trojans. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-4216.

Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. BadNets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv preprint arXiv:1708.06733*, 2017. URL https://arxiv.org/abs/1708.06733.

Sungwon Han, Sungwon Park, Fangzhao Wu, Sundong Kim, Bin Zhu, Xing Xie, and Meeyoung Cha. Towards attack-tolerant federated learning via critical parameter analysis. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4976–4985. IEEE, 2023. doi: 10.1109/ICCV51070.2023.00461.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

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 (ICCV)*, pp. 4652–4662, 2023a.

Tiansheng Huang, Sihao Hu, Ka-Ho Chow, Fatih Ilhan, Selim Tekin, and Ling Liu. Lockdown: backdoor defense for federated learning with isolated subspace training. In *Advances in Neural Information Processing Systems*, volume 36, pp. 10876–10896, 2023b.

Najeeb Moharram Jebreel and Josep Domingo-Ferrer. FL-Defender: Combating targeted attacks in federated learning. *Knowledge-Based Systems*, 260:110178, 2023. doi: 10.1016/j.knosys.2022.110178. URL https://www.sciencedirect.com/science/article/pii/S0950705122012746.

Sai Praneeth Karimireddy, Lie He, and Martin Jaggi. Learning from history for byzantine robust optimization. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5311–5319. PMLR, 2021.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

C. Leys, C. Ley, O. Klein, P. Bernard, and L. Licata. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*, 49(4):764–766, 2013. doi: 10.1016/j.jesp.2013.03.013.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015.

H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 1273–1282. PMLR, 2017.

Mustafa Safa Ozdayi, Murat Kantarcioglu, and Yulia R. Gel. Defending against backdoors in federated learning with robust learning rate. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(10):9268–9276, 2021. doi: 10.1609/aaai.v35i10.17118.

Krishna Pillutla, Sham M. Kakade, and Zaïd Harchaoui. Robust aggregation for federated learning. *IEEE Transactions on Signal Processing*, 70:1142–1154, 2022. doi: 10.1109/TSP.2022.3153135.

Virat Shejwalkar and Amir Houmansadr. Manipulating the Byzantine: Optimizing model poisoning attacks and defenses for federated learning. In *Proceedings of the Network and Distributed System Security Symposium (NDSS)*, 2021. doi: 10.14722/ndss.2021.24498. URL https://doi.org/10.14722/ndss.2021.24498.

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

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 (ASIA CCS)*, pp. 946–958. ACM, 2022. doi: 10.1145/3488932.3517395.

Chulin Xie, Keli Huang, Pin-Yu Chen, and Bo Li. DBA: Distributed backdoor attacks against federated learning. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rkgyS0VFvr.

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, 2025. doi: 10.1109/CVPR52734.2025.01923.

Dong Yin, Yudong Chen, Kannan Ramchandran, and Peter L. Bartlett. Byzantine-robust distributed learning: Towards optimal statistical rates. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pp. 5650–5659. PMLR, 2018.

Zhengming Zhang, Ashwinee Panda, Linyue Song, Yaoqing Yang, Michael W. Mahoney, Joseph E. Gonzalez, Kannan Ramchandran, and Prateek Mittal. Neurotoxin: Durable backdoors in federated learning. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 26429–26446. PMLR, 2022. URL https://proceedings.mlr.press/v162/zhang22w.html.

Haomin Zhuang, Mingxian Yu, Hao Wang, Yang Hua, Jian Li, and Xu Yuan. Backdoor federated learning by poisoning backdoor-critical layers. In *The Twelfth International Conference on Learning Representations (ICLR)*. OpenReview.net, 2024. URL https://openreview.net/forum?id=AJBGSVSTT2.

APPENDIX CONTENTS

A Experimental Details and Additional Results **14**

A.1 Experimental Setup 14

A.2 Implementation Details 17

A.3 Qualitative Reconstruction Examples 18

A.4 Choice of Observed Parameters 19

A.5 Component and Threshold Studies 20

A.6 Detection Performance 21

B Scoring Geometry and Conditional Detection **22**

B.1 Analysis Setup and Reference Direction 22

B.2 Median Location Error and Directional Perturbation 22

B.3 Proof of Theorem 1 23

C Decision Stability under Compact Observations **23**

C.1 Coordinate Selection and Order-Statistic Stability 23

C.2 Proof of Theorem 2 24

C.3 Budget for Random Block Observations 24

C.4 Proof of Proposition C.1 24

D Weighted Aggregation under Imperfect Filtering **25**

D.1 Aggregation Space and Error Quantities 25

D.2 Proof of Theorem 3 26

D.3 From Detection and Decision Preservation to Aggregation Error 27

D.4 Clipping Radius and Scope 27

A Experimental Details and Additional Results

A.1 Experimental Setup

Tasks and data. We evaluate federated diffusion and classification on CIFAR-10, with additional diffusion experiments on CelebA. Table 7 summarizes the reference settings for the controlled diffusion studies and the main classification comparison. All clients participate in every communication round in these reference settings. The malicious-client fraction is $m/n = 0.20$, where n and m denote the total and malicious client counts.

Table 7: Reference experimental settings. The diffusion column specifies the controlled CIFAR-10 protocol used by the layer-selection study.

Setting	Federated diffusion	Federated classification
Dataset / image size	CIFAR-10 / 32×32	CIFAR-10 / 32×32
Model	DDPM (U-Net backbone)	ResNet9
Clients / attackers	5 / 1	20 / 4
Participation per round	5 of 5	20 of 20
Data partition	Dirichlet $\beta = 0.70$	IID
Initialization	Clean FedAvg, round 2,000	Random initialization
Communication rounds	300	100
Local training per round	78 optimizer steps	2 epochs
Batch size	128	64
Optimizer / initial LR	Adam / 2×10^{-4}	SGD / 0.10
Private attack targets	1,000	Not applicable
Repetition seeds	Train: 42/43/44; data: 42	0/1/2 (training and partition)
Primary endpoint	Round 300	Round 100

Notes. The layer-selection rows in Section A.4 each use training seed 42 only. The diffusion local-update budget excludes AdaSCP's additional parameter-search operations. Other diffusion experiments use the settings specified below.

Diffusion training. We use denoising diffusion probabilistic models (DDPMs) with U-Net denoisers. For CIFAR-10, the U-Net backbone uses base width 128, channel multipliers $(1, 2, 2, 2)$, two residual blocks per resolution, and dropout 0.10. The diffusion process has 1,000 steps and a linear noise-variance schedule from 10^{-4} to 0.02. The model predicts noise and maintains an exponential moving average (EMA) with decay 0.9999. Training uses Adam with learning rate 2×10^{-4}, a 5,000-update learning-rate warm-up, and gradient-norm clipping at 1.00. No detector warm-up is used. The controlled studies fix data-partition seed 42. The layer-selection experiment also fixes training seed 42.

Classification training. ResNet9 is trained from scratch with SGD, initial learning rate 0.10, zero momentum, and weight decay 10^{-4}. At communication round t, starting at $t = 1$, the client learning rate is 0.10×0.99^t. The server learning rate is 1.00. Each client trains for two local epochs with batch size 64. Repetitions use seeds 0/1/2. Each seed controls both training and data-partition randomness. The classification benchmark uses a fixed AlignIns source snapshot with compatibility fixes for Lockdown's missing default option and the geometric-median objective's CUDA weight handling.

A.1 Experimental Setup (continued)

Additional protocols. Table 8 records the principal extensions to the reference settings. CelebA uses 64×64 images and a U-Net with channel multipliers $(1, 2, 2, 2, 4)$.

Table 8: Additional experimental protocols. The reference settings are given in Table 7.

Experiment family	Settings that change
CelebA diffusion	64×64; 150 rounds; 500 private targets.
Classification, non-IID	Dirichlet $\beta = 0.50$; the 100-round protocol retains poisoning ratio 0.30 and all four classification attacks.

Attacks and defenses. The diffusion comparisons include data poisoning, model poisoning, PGD poisoning, BC-layer substitution, and AdaSCP. The controlled layer-selection study uses AdaSCP with model-poisoning scale 5.00, critical proportion 0.40, global pruning, and adaptive learning rate 0.20. Each attacker update in this study combines 64 benign and 64 poisoned samples. Classification comparisons use BadNet, DBA, Neurotoxin, and PGD, with local poisoning ratio 0.30. The attack target is class 7 and the trigger is a plus-shaped pattern. Neurotoxin retains 25% of gradient coordinates. FedAvg is the undefended baseline. Diffusion baselines include Krum, Multi-Krum, FoolsGold, RFA, Multi-Metrics, and AlignIns. Classification baselines include Multi-Krum, FoolsGold, RFA, Multi-Metrics, AlignIns, RLR, and Lockdown.

Baseline settings. All classification baselines use the implementations in the AlignIns framework. Multi-Krum is given the true number of malicious clients (four) and averages the 10 selected updates. AlignIns uses a sparsity fraction of 0.30 and MPSA and TDA thresholds of 1.00. RFA computes an equal-weight geometric median and then reweights clients by exponential distance and re-solves twice, with at most 100 Weiszfeld iterations in each of the three solves, smoothing parameter 10^{-6}, and relative objective tolerance 10^{-20}. Multi-Metrics retains the 6 of 20 clients with the lowest scores. FoolsGold uses update similarity within the current round without accumulated history. RLR uses threshold 8 and server learning rate 1.00. Lockdown uses density 0.25, annealing coefficient 10^{-4}, consensus threshold 10, and a shared ERK-initialized mask. The diffusion AlignIns runs use the same sparsity fraction and thresholds.

Method variants. The diffusion reference configuration computes centered cosine scores from EMA parameters, uses MAD threshold coefficient $\lambda = 2.00$, applies clipping relative to the current global model, and aggregates retained clients with FedAvg. The default CIFAR-10 representation contains `upblocks.14.block1.2.weight` and `tail.2.weight`. For CelebA, the IID run and the non-IID runs with data-partition seeds 43 and 44 observe `upblocks.18.block1.2.weight` and `tail.2.weight`; the non-IID run with data-partition seed 42 observes `upblocks.14.block1.2.weight` and `tail.2.weight`. The classification slice uses `classifier.2.weight` (5,120 parameters). The full-parameter classification variant, CoPR-full, observes all 6,573,120 parameters. Both variants train and aggregate the complete model; only their scoring representations differ. The DBA heterogeneity sweep uses the full-parameter variant.

Evaluation and reporting. Diffusion quality is measured by FID using 50,000 generated images. Stolen-MSE is defined in Appendix A.1.2. Classification clean accuracy (ACC) uses all 10,000 test images. ASR and robust accuracy (RA) evaluate triggered images against the attack target and the original label, respectively, excluding the target class (9,000 images). The primary endpoint is the final round. Detection metrics are defined in Appendix A.6. When there are three repeated runs, we report the mean and sample standard deviation. The independent repetition unit is a training run; single-run results have no error bars.

A.1.1 Figure- and Table-Specific Protocols

The main-text tables and figures use the experiment families summarized below. Common training details remain in Tables 7 and 8.

Classification evaluation and ranking. For Table 3, each seed controls both training and data-partition randomness. ACC is evaluated on 10,000 clean test images. ASR and RA are target-label and original-label accuracy, respectively, on the same 9,000 triggered test images whose original

Table 9: Protocol map for the main experimental evidence.

Evidence	Protocol and reporting distinction
Table 1	CIFAR-10; non-IID $\beta = 0.70$; 300 rounds; 1,000 targets. Average is over five attacks.
Table 2	AdaSCP on CIFAR-10 and CelebA under IID and non-IID partitions.
Table 3	ResNet9; 20 clients with four attackers; local poisoning ratio 0.30; final round 100; seeds 0/1/2. Panel (a) reports CoPR; panel (b) reports both CoPR and CoPR-full.
Figure 3	Target-specific grid; selection procedure in Appendix A.3.
Figure 4	AdaSCP on CIFAR-10 and CelebA; Stolen-MSE in raw units.
Figure 5	DBA; CoPR-full; $\beta \in \{0.10, 0.20, 0.30, 0.40, 0.50, 1.00\}$; 4 or 6 attackers among 20 clients; local poisoning ratio 0.50; 100 rounds; seed 0.
Figure 6	Default CIFAR-10 AdaSCP run; five clients, one attacker, 300 rounds, training/data seed 42.
Tables 4 and 6	Controlled CIFAR-10 AdaSCP comparisons. Training/data seed 42; 300 rounds; 1,000 targets. One training run per configuration.

labels differ from the attack target. RA is not $100 - \mathrm{ASR}$, since predictions can belong to other incorrect classes. Values are means and sample standard deviations over seeds 0/1/2; the attack averages are calculated before rounding. Bold/underline mark best/second-best means, including ties, with FedAvg included in panel (a) and excluded from the defense ranking in panel (b). Gray identifies CoPR and CoPR-full in their respective panels; markings do not imply statistical significance.

Cross-dataset evaluations. The non-IID CIFAR-10 CoPR entry in Table 2 reuses the single-run default-late evaluation in Table 4. The non-IID CelebA CoPR entry averages runs with data-partition seeds 42/43/44, with the observation slices specified above; the IID CelebA entry uses data-partition seed 42 alone.

Heterogeneity sweep. The DBA sweep is separate from the classification main table: its malicious-client fractions are 4/20 and 6/20, whereas its within-client poisoning ratio is fixed at 0.50 rather than 0.30.

A.1.2 Stolen-MSE Evaluation

Stolen-MSE measures how accurately the global model reconstructs specific private targets. Following DataStealing (Gan et al., 2024), each target image is associated with a target-specific combinatorial trigger, and the samples generated with that trigger are compared pixel by pixel with that target, not with the training set as a whole.

Let $I_j \in [0,1]^{3HW}$ be the jth target image, flattened as RGB pixel values, and let $\widehat{I}_{j,k}$ be the kth sample generated with its trigger. Then

$$\text{Stolen-MSE} = \frac{1}{N_{\text{tar}} N_{\text{gen}}} \sum_{j=1}^{N_{\text{tar}}} \sum_{k=1}^{N_{\text{gen}}} \frac{\|\widehat{I}_{j,k} - I_j\|_2^2}{3HW}, \tag{13}$$

that is, the squared error of each sample is averaged over all pixels and color channels, and then over the N_{gen} samples of each target and all N_{tar} targets. For quantitative evaluation, we use full images at the evaluation resolution, with $N_{\text{tar}} = 1{,}000$ and $N_{\text{gen}} = 20$ for CIFAR-10 at 32×32, and $N_{\text{tar}} = 500$ and $N_{\text{gen}} = 20$ for CelebA at 64×64. All 20 samples enter the average; the metric does not keep only the closest sample per target. Tables report Stolen-MSE in units of 10^{-2}, so 14.24 corresponds to a raw MSE of 0.1424, and Figure 4 uses raw units. For repeated runs, Stolen-MSE is computed per run and then aggregated as stated for each table.

A lower Stolen-MSE means that the generated samples are closer to the private targets, so a defense should keep it high; under undefended FedAvg with the Model attack, it falls to 0.67×10^{-2} (Table 1). A collapsed or noise-producing model can also yield a high value, so we interpret Stolen-MSE jointly with FID to assess target reconstruction and normal generation quality. The qualitative grids in Figures 3 and 7 complement this average by showing, in each cell, the sample with the lowest MSE among the generated candidates for the corresponding target and attack–defense pair, which is the attacker's best case among those candidates.

A.2 Implementation Details

Round-local decisions and parameter spaces. CoPR recomputes the scores and the retained set from the uploads of each round and keeps no cross-round suspect list, detection warm-up, or auxiliary detection model. For diffusion models, scoring uses the selected EMA parameters, while clipping and aggregation use the concatenation of the trained parameters and their EMA copy; one clipping factor, computed from the norm of the joint update, is applied to both branches. For classification, scoring and aggregation use the model parameters. Client means, coordinate-wise medians, score medians, and MADs weight clients equally, whereas the data-size weights $p_i = |\mathcal{D}_i| / \sum_{j \in \mathcal{C}} |\mathcal{D}_j|$ enter only the final aggregation and are renormalized after filtering. The observed tensors are listed in Appendix A.1 and Table 10.

Numerical conventions. In the reference implementation, every median averages the two central values for an even sample size. The MAD path adds 10^{-30} to the raw MAD before multiplying by λ, that is, $\varepsilon_0 = 10^{-30}$ in (6). Scoring centers float32 observations and uses float64 reductions for directional normalization. A residual or reference with zero norm is assigned cosine similarity zero and hence score one; other cosines are clipped to $[-1, 1]$. With one client, filtering and clipping are skipped; with fewer than three scores, the MAD filter rejects no client; and clipping is skipped when fewer than two retained clients have positive weight. A zero update is left unchanged, and when the median radius is zero, retained updates with positive norm are mapped to the anchor. The historical CelebA runs computed cosine scores with float32 norm reductions and a 10^{-30} norm stabilizer instead of the float64 normalization above, and source matching for the earlier CIFAR-10 runs has not been fully verified.

A.3 Qualitative Reconstruction Examples

Purpose and selection. Figure 7 supplements the truck example in Figure 3. Each grid compares five attacks across eight defenses for one target; each cell shows the candidate reconstruction closest to the target (lowest Stolen-MSE), i.e., the worst case for the defense among the generated candidates.

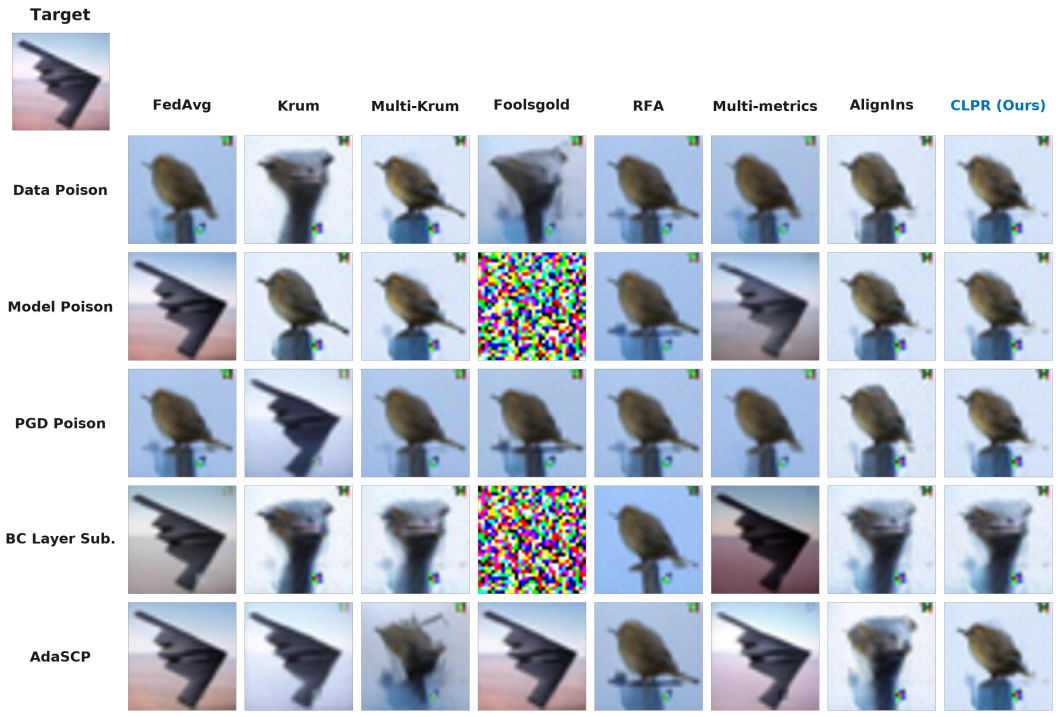


Figure 7: Qualitative reconstruction comparison for an aircraft target on CIFAR-10. Rows denote attacks and columns denote defenses. Each cell shows the lowest-Stolen-MSE reconstruction under this selection protocol. Several cells retain recognizable target structure, whereas even the closest AlignIns and CoPR reconstructions differ visibly from the target.

Visible target recovery. In Figure 7, the aircraft remains recognizable under FedAvg for Model Poison, BC layer substitution, and AdaSCP, and under some other attack–defense combinations. In contrast, no AlignIns or CoPR cell recovers the target under any of the five attacks.

Reconstruction failure alone is not sufficient. A noise-like triggered output can also result from a collapsed model, as in the FoolsGold Model and BC cells (FID 459.42 and 196.24 in Table 1).

A.4 Choice of Observed Parameters

We test whether compact EMA-parameter representations retain detection and generation performance on CIFAR-10 under AdaSCP. Each compact representation contains 298,368 coordinates, or 0.78% of the full representation; the complete model is still trained and transmitted. All runs use five clients with one attacker, Dirichlet $\beta = 0.70$, 300 rounds, 1,000 private targets, and data and training seed 42. Centering, MAD threshold ($\lambda = 2.00$), clipping, and aggregation settings are held fixed.

Table 4 reports the results, which Section 6.3 discusses. Detection covers all 300 rounds: 300 malicious and 1,200 benign client-round decisions.

Representation construction. Table 10 specifies the observed parameters. The middle configuration samples coordinates within one tensor to match the compact budget. Random configurations sample coordinates from the full trainable parameter vector, without replacement. Each sampled index set is fixed throughout training. Selector seeds 0/1/2 vary the sampled representation while the training seed remains fixed at 42.

Table 10: Parameter sets used in the representation study.

Representation	Observed parameters
Early	`head.weight`, `downblocks.0.block1.2.weight`, `downblocks.1.block1.2.weight`.
Middle	298,368 of 589,824 coordinates in `middleblocks.0.block1.2.weight`, selector seed 0.
Default late	`upblocks.14.block1.2.weight` and `tail.2.weight`.
Adjacent late	`upblocks.13.block1.2.weight` and `tail.2.weight`.
Random 0/1/2	298,368 coordinates sampled without replacement from all 38,307,203 trainable scoring parameters, selector seed 0/1/2.
Full	All 38,307,203 trainable scoring parameters.

A.5 Component and Threshold Studies

CIFAR-10 protocol. Table 6 in the main text uses the same non-IID partition ($\beta = 0.70$), five participating clients with one attacker, 1,000 private targets, and 300 rounds from the same clean checkpoint. Each configuration uses training and data seed 42. The observed parameters are the default late-layer representation used in the preceding subsection. *Without clipping* disables centered clipping while retaining detection-based exclusion. *Clipping only* disables detection-based client exclusion while retaining clipping around the current global model. Evaluation uses the final-round checkpoint, 50,000 clean samples, and all 20 reconstructions for each of 1,000 targets. The complete-method row is reused from the preceding layer study.

CelebA configurations. Table 11 compares three configurations on CelebA. Replacing fixed-count filtering with the MAD threshold removes the need to know the number of attackers without degrading defense performance. The complete configuration has lower FID and higher Stolen-MSE than the detection-only configurations, but it also uses a different observation slice, so this comparison does not isolate the effect of clipping.

Table 11: Configuration comparison on CelebA under AdaSCP. Non-IID $\beta = 0.70$; five clients with one attacker; 150 rounds, 500 targets, seed 42.

Configuration	FID ↓	Stolen-MSE ↑
FedAvg	9.05	1.14
Fixed-count residual detection	<u>9.03</u>	<u>18.80</u>
MAD residual detection	9.09	<u>18.80</u>
MAD detection and clipping (**CoPR**)	**8.09**	**19.50**

Threshold sensitivity. We examine three CIFAR-10 runs under AdaSCP with five clients, one attacker, Dirichlet $\beta = 0.70$, 1,000 private targets, 300 rounds, and data-partition seed 42. The MAD threshold coefficient λ controls rejection aggressiveness.

Table 12: CIFAR-10 threshold sensitivity. One run per setting; detection over all 300 rounds.

λ	FID	Stolen-MSE	TPR (%)	FPR (%)
1.00	6.25	14.27	100.00	8.00
2.00	6.16	14.27	100.00	2.33
3.00	6.17	14.26	100.00	1.00

Notes. Stolen-MSE is in 10^{-2} units. Each run has 300 malicious and 1,200 benign client-round decisions. False-positive counts are 96, 28, and 12. Gray marks the default coefficient.

Results. Increasing λ from 1 to 3 reduces the false-positive rate from 8.00% to 1.00%, while TPR remains 100.00%. FID (6.16–6.25) and Stolen-MSE (14.26–14.27) are essentially unchanged, and the default $\lambda = 2$ gives the lowest FID.

A.6 Detection Performance

Detection metrics. A positive decision flags a client as suspicious. TP and FN count flagged and unflagged malicious-client decisions; FP and TN count flagged and unflagged benign-client decisions. We report

$$\mathrm{DACC} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{FP}+\mathrm{FN}+\mathrm{TN}} \times 100\%,$$

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}} \times 100\%, \qquad \mathrm{FNR} = \frac{\mathrm{FN}}{\mathrm{TP}+\mathrm{FN}} \times 100\%.$$

The counting unit is one client in one round.

Fixed-partition evaluation across training seeds. Table 13 uses the same CIFAR-10 partition ($\beta = 0.70$, data seed 42), AdaSCP configuration, 1,000 private targets, and 300-round continuation from the clean checkpoint. Each method is trained with seeds 42, 43, and 44. The full-parameter variant changes the observed representation from the default late-layer tensors to all trainable EMA parameters. We first compute each run's detection rates, then report their mean and sample standard deviation across the three training runs.

Table 13: Detection under AdaSCP with three training seeds. CIFAR-10; fixed data partition, $\beta = 0.70$; five clients with one attacker.

	Detection (%, mean ± sample SD)		
Method	**DACC**	**FPR**	**FNR**
AlignIns	71.51 ± 1.41	35.56 ± 1.77	$\underline{0.22} \pm 0.19$
CoPR (full parameters)	$\mathbf{97.42} \pm 0.20$	$\mathbf{3.22} \pm 0.25$	$\mathbf{0.00} \pm 0.00$
CoPR (default)	$\underline{97.13} \pm 0.29$	$\underline{3.58} \pm 0.36$	$\mathbf{0.00} \pm 0.00$

Notes. Higher DACC and lower FPR/FNR are preferred. **Bold**/underline: best/second-best means, including ties. Training seeds: 42/43/44; data seed: 42. Each run covers rounds 1–300 (300 malicious and 1,200 benign client-round decisions). Gray marks the default. The seed-42 CoPR runs are reused from the layer study.

Both CoPR representations flag the malicious client in every round for all three seeds, with FPR 3.22–3.58% versus 35.56% for AlignIns.

B Scoring Geometry and Conditional Detection

B.1 Analysis Setup and Reference Direction

Fix a round with n clients, benign set $\mathcal{B}$, and malicious set $\mathcal{A}$. As in Section 4, $x_i \in \mathbb{R}^D$ is the scoring-parameter vector of client i, $P_S \in \mathbb{R}^{|S|\times D}$ selects a nonempty coordinate set S, and $z_i = P_S x_i$. Except for Proposition C.1, all statements in Appendices B–D are deterministic given the uploads of the round, and client identities enter only the analysis.

Write

$$\bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_i, \qquad u_i = z_i - \bar{z}, \qquad q = \mathrm{med}_{\mathrm{coord}}(z_1, \ldots, z_n). \tag{14}$$

Since the coordinate-wise median is equivariant under a common translation, $u_{\mathrm{med}} = q - \bar{z}$. Let $h = \bar{z} - q = -u_{\mathrm{med}}$. For nonzero u_i and h, the score in (4) becomes

$$r_i^{\cos} = 1 + \left\langle \frac{u_i}{\|u_i\|}, \frac{h}{\|h\|} \right\rangle, \tag{15}$$

that is, one plus the normalized projection of u_i onto h. The threshold and decision margin are

$$\tilde{r} = \mathrm{med}_i\, r_i^{\cos}, \qquad \mathrm{MAD} = \mathrm{med}_i\, |r_i^{\cos} - \tilde{r}|, \qquad \tau = \tilde{r} + \lambda(\mathrm{MAD} + \varepsilon_0), \qquad \Gamma_i = r_i^{\cos} - \tau. \tag{16}$$

Every median uses the same convention; the implementation averages the two central values for even sample sizes, and the arguments below also hold for the lower or the upper median.

Let $b = |\mathcal{B}|$, $m = |\mathcal{A}|$, and $\rho = m/n \in (0, 1/2)$, and define

$$\mu_B = \frac{1}{b}\sum_{i\in\mathcal{B}} z_i, \qquad \mu_A = \frac{1}{m}\sum_{i\in\mathcal{A}} z_i, \qquad a = \mu_A - \mu_B, \qquad e = q - \mu_B.$$

Decomposing the mean by client counts gives

$$\bar{z} = (1-\rho)\mu_B + \rho\mu_A = \mu_B + \rho a, \qquad h = \rho a - e. \tag{17}$$

Hence the residual of a benign client is $(z_i - \mu_B) - \rho a$ and that of a malicious client is $(z_i - \mu_A) + (1-\rho)a$. The between-group terms point in opposite directions along $v = a/\|a\|$, while the within-group terms $z_i - \mu_B$ and $z_i - \mu_A$ determine how tightly each group concentrates around them. Since $\|\rho a\| < \|(1-\rho)a\|$ for $\rho < 1/2$, benign residuals are more easily dominated by within-group variation, which is what the angle bounds $\theta_\pm$ in Theorem 1 quantify.

B.2 Median Location Error and Directional Perturbation

We first show that a benign majority confines the coordinate-wise median. For coordinate j, let $y_{(\nu),j}$ be the νth smallest benign observation and $z_{(\nu),j}$ the νth smallest observation over all clients, and set

$$k_- = \left\lfloor \frac{n+1}{2} \right\rfloor, \qquad k_+ = \left\lceil \frac{n+1}{2} \right\rceil.$$

For $m < \nu \le b$,

$$y_{(\nu-m),j} \le z_{(\nu),j} \le y_{(\nu),j}. \tag{18}$$

The upper bound holds because the ν smallest benign values are all at most $y_{(\nu),j}$. For the lower bound, fewer than $\nu - m$ benign values and at most m malicious values lie strictly below $y_{(\nu-m),j}$, so fewer than ν values do. Since $m < n/2$, the central ranks satisfy $k_- > m$ and $k_+ \le b$, and q_j lies between $z_{(k_-),j}$ and $z_{(k_+),j}$ under any fixed median convention. Therefore,

$$y_{(k_- - m),j} \le q_j \le y_{(k_+),j}. \tag{19}$$

With

$$R_j = \max\left\{ |y_{(k_- - m),j} - \mu_{B,j}|,\, |y_{(k_+),j} - \mu_{B,j}| \right\},$$

where $\mu_{B,j}$ is coordinate j of μ_B, it follows that

$$\|e\|^2 = \sum_{j=1}^{|S|} |q_j - \mu_{B,j}|^2 \le \sum_{j=1}^{|S|} R_j^2, \tag{20}$$

so the median-location error is controlled by the spread of the benign order statistics around the benign mean.

Next, let $a \neq 0$. For nonzero vectors x and y,

$$\left\| \frac{y}{\|y\|} - \frac{x}{\|x\|} \right\| \leq \left\| \frac{y}{\|y\|} - \frac{y}{\|x\|} \right\| + \frac{\|y-x\|}{\|x\|} = \frac{\big|\|x\| - \|y\|\big|}{\|x\|} + \frac{\|y-x\|}{\|x\|} \leq \frac{2\|y-x\|}{\|x\|}. \tag{21}$$

If $\|e\| < \rho\|a\|$, then $\|h\| = \|\rho a - e\| \geq \rho\|a\| - \|e\| > 0$, and taking $x = \rho a$ and $y = h$ gives

$$\eta = \left\| \frac{h}{\|h\|} - v \right\| \leq \frac{2\|e\|}{\rho\|a\|}. \tag{22}$$

The condition $\|e\| < \rho\|a\|$ only guarantees $h \neq 0$; Theorem 1 uses η directly whenever $h \neq 0$. Combining (20) and (22), the reference direction is accurate whenever the spread of the benign order statistics is small relative to $\rho\|a\|$.

B.3 Proof of Theorem 1

Proof. Let $\widehat{v} = h/\|h\|$. As in the theorem, θ_j denotes the angle between u_j and $-v$, and ϕ_i the angle between u_i and v. Define the ideal score

$$r_i^{\cos,\star} = 1 + \left\langle \frac{u_i}{\|u_i\|}, v \right\rangle.$$

By the Cauchy–Schwarz inequality,

$$|r_i^{\cos} - r_i^{\cos,\star}| = \left| \left\langle \frac{u_i}{\|u_i\|}, \widehat{v} - v \right\rangle \right| \leq \eta.$$

For $j \in \mathcal{G}$, $r_j^{\cos,\star} = 1 - \cos\theta_j \in [1 - \cos\theta_-, 1 - \cos\theta_+]$, so $r_j^{\cos} \in [L, U]$ with

$$L = 1 - \cos\theta_- - \eta, \qquad U = 1 - \cos\theta_+ + \eta.$$

Since $|\mathcal{G}| > n/2$, both central order statistics of the scores lie in $[L, U]$, so $\tilde{r} \in [L, U]$. Then $|r_j^{\cos} - \tilde{r}| \leq U - L$ for every $j \in \mathcal{G}$, so more than half of the absolute deviations are at most $U - L$, and

$$\tilde{r} \leq U, \qquad \mathrm{MAD} \leq U - L = \cos\theta_- - \cos\theta_+ + 2\eta. \tag{23}$$

For a malicious client, $r_i^{\cos,\star} = 1 + \cos\phi_i$, so $r_i^{\cos} \geq 1 + \cos\phi_i - \eta$. Substituting these bounds into $\Gamma_i = r_i^{\cos} - \tilde{r} - \lambda(\mathrm{MAD} + \varepsilon_0)$ gives

$$\begin{aligned} \Gamma_i &\geq 1 + \cos\phi_i - \eta - (1 - \cos\theta_+ + \eta) - \lambda\big(\cos\theta_- - \cos\theta_+ + 2\eta + \varepsilon_0\big) \\ &= \cos\phi_i + \cos\theta_+ - \lambda(\cos\theta_- - \cos\theta_+) - 2(1+\lambda)\eta - \lambda\varepsilon_0 = \underline{\Gamma}_i. \end{aligned} \tag{24}$$

Hence $\underline{\Gamma}_i > 0$ implies $\Gamma_i > 0$, and client i is rejected. □

C Decision Stability under Compact Observations

C.1 Coordinate Selection and Order-Statistic Stability

Throughout this appendix, compact and full-parameter quantities are computed from the same uploads and are marked by the superscripts S and full. Because P_S selects coordinates, it commutes with the client mean and with the coordinate-wise median, so $\bar{z}^S = P_S \bar{z}^{\mathrm{full}}$ and $q^S = P_S q^{\mathrm{full}}$, and therefore

$$u_i^S = P_S u_i^{\mathrm{full}}, \qquad h^S = P_S h^{\mathrm{full}}. \tag{25}$$

This identity uses that P_S selects coordinates; it fails for general linear maps, which do not commute with the coordinate-wise median.

Order statistics are stable under perturbations bounded in the maximum norm. For real sequences s and s' of length n with $\max_i |s_i - s'_i| \leq c$,

$$|s_{(\nu)} - s'_{(\nu)}| \leq c \quad \text{for every } \nu. \tag{26}$$

Indeed, at least ν values s'_i are at most $s'_{(\nu)}$, and the corresponding s_i are at most $s'_{(\nu)} + c$, so $s_{(\nu)} \leq s'_{(\nu)} + c$; exchanging s and s' gives the reverse inequality. Consequently, the lower median, the upper median, and their average all change by at most c.

C.2 Proof of Theorem 2

Proof. Applying (26) with $c = \xi_S$ to the two score sequences gives $|\tilde{r}^S - \tilde{r}^{\text{full}}| \le \xi_S$. For each client, the absolute deviations satisfy

$$\left| |r_i^{\cos,S} - \tilde{r}^S| - |r_i^{\cos,\text{full}} - \tilde{r}^{\text{full}}| \right| \le |r_i^{\cos,S} - r_i^{\cos,\text{full}}| + |\tilde{r}^S - \tilde{r}^{\text{full}}| \le 2\xi_S,$$

and applying (26) again gives $|\text{MAD}^S - \text{MAD}^{\text{full}}| \le 2\xi_S$. Since both scopes use the same λ and ε_0, the stabilizer cancels, and

$$|\tau^S - \tau^{\text{full}}| \le |\tilde{r}^S - \tilde{r}^{\text{full}}| + \lambda|\text{MAD}^S - \text{MAD}^{\text{full}}| \le (1 + 2\lambda)\xi_S. \tag{27}$$

Therefore,

$$|\Gamma_i^S - \Gamma_i^{\text{full}}| \le |r_i^{\cos,S} - r_i^{\cos,\text{full}}| + |\tau^S - \tau^{\text{full}}| \le 2(1 + \lambda)\xi_S. \tag{28}$$

If $\Gamma_i^{\text{full}} > 2(1 + \lambda)\xi_S$, then $\Gamma_i^S > 0$ and the client remains rejected; if $\Gamma_i^{\text{full}} < -2(1 + \lambda)\xi_S$, then $\Gamma_i^S < 0$ and it remains retained. When Theorem 1 applies to the full observation, $\Gamma_i^{\text{full}} \ge \underline{\Gamma}_i^{\text{full}}$, which gives the final claim. □

Because aggregation uses the full updates regardless of S, preserving every filtering decision also preserves the retained set, the clipping radius, and the aggregate.

C.3 Budget for Random Block Observations

Fix the uploads of a round with nonzero full-parameter residuals and reference vector. Partition the D scoring coordinates into $J \ge 2$ disjoint nonempty blocks, possibly of different sizes, and let S be the union of $k \in \{1, \dots, J\}$ blocks drawn uniformly without replacement, independently of the uploads; single-coordinate blocks give uniform coordinate sampling. Let $\mathcal{I}$ be the selected block indices, write $w^{(\ell)}$ for the restriction of a vector w to block ℓ, and let

$$\widetilde{u}_i = \frac{u_i^{\text{full}}}{\|u_i^{\text{full}}\|}, \qquad \widetilde{h} = \frac{h^{\text{full}}}{\|h^{\text{full}}\|}.$$

The block-energy concentration

$$\kappa_{\text{blk}} = J \max_{1 \le \ell \le J} \max \left\{ \max_{1 \le i \le n} \|\widetilde{u}_i^{(\ell)}\|^2, \|\widetilde{h}^{(\ell)}\|^2 \right\} \in [1, J] \tag{29}$$

is large when the energy of some residual or of the reference is concentrated in a few blocks.

Proposition C.1 (Random-observation budget). *Let $\epsilon \in (0, 1)$ and $\delta \in (0, 1)$. If*

$$k \ge \frac{10\kappa_{\text{blk}}}{3\epsilon^2} \log \frac{2(2n + 1)}{\delta}, \tag{30}$$

then, with probability at least $1 - \delta$, all selected residuals and the reference vector are nonzero, and

$$\xi_S \le \frac{2\epsilon}{1 - \epsilon}, \qquad \max_i |\Gamma_i^S - \Gamma_i^{\text{full}}| \le \frac{4(1 + \lambda)\epsilon}{1 - \epsilon}. \tag{31}$$

The tolerance ϵ bounds sampling errors in squared norms and inner products and is distinct from ξ_S and ε_0. Combined with Theorem 2, a client keeps its filtering decision with probability at least $1 - \delta$ whenever $|\Gamma_i^{\text{full}}| > 4(1 + \lambda)\epsilon/(1 - \epsilon)$, so a smaller margin requires a smaller ϵ and, through (30), a larger k. Since $\rho \in (0, 1/2)$ forces $n \ge 3$ and $\delta < 1$, we have $\log(2(2n + 1)/\delta) \le 2.7 \log(n/\delta)$, so (30) holds whenever $k \ge 9\,\kappa_{\text{blk}}\epsilon^{-2} \log(n/\delta)$, the form stated in Section 5.2.

C.4 Proof of Proposition C.1

Proof. **Sampled quantities.** Let $R_S = \sqrt{J/k}\, P_S$; this positive rescaling cancels in every cosine and does not change the scores. For each block, define

$$X_{i\ell} = J\|\widetilde{u}_i^{(\ell)}\|^2, \qquad Y_\ell = J\|\widetilde{h}^{(\ell)}\|^2, \qquad Z_{i\ell} = J\langle \widetilde{u}_i^{(\ell)}, \widetilde{h}^{(\ell)} \rangle,$$

and let $\chi_i = \langle \widetilde{u}_i, \widetilde{h} \rangle$. Over the J blocks, these populations have means 1, 1, and χ_i, and their sample means over $\mathcal{I}$ are

$$\frac{1}{k}\sum_{\ell\in\mathcal{I}} X_{i\ell} = \|R_S\widetilde{u}_i\|^2, \qquad \frac{1}{k}\sum_{\ell\in\mathcal{I}} Y_\ell = \|R_S\widetilde{h}\|^2, \qquad \frac{1}{k}\sum_{\ell\in\mathcal{I}} Z_{i\ell} = \langle R_S\widetilde{u}_i, R_S\widetilde{h}\rangle.$$

Variances and ranges. Expectations and variances below are taken over a block index drawn uniformly from $\{1,\dots,J\}$. Write $\kappa = \kappa_{\mathrm{blk}}$. Since $0 \le X_{i\ell}, Y_\ell \le \kappa$ and both means equal one,

$$\mathrm{Var}(X_i) \le \mathbb{E}X_i^2 \le \kappa\,\mathbb{E}X_i = \kappa, \qquad \mathrm{Var}(Y) \le \kappa.$$

By the Cauchy–Schwarz inequality, $|Z_{i\ell}| \le J\|\widetilde{u}_i^{(\ell)}\|\|\widetilde{h}^{(\ell)}\| \le \kappa$, and

$$\mathrm{Var}(Z_i) \le \frac{1}{J}\sum_{\ell=1}^{J} Z_{i\ell}^2 \le J\sum_{\ell=1}^{J}\|\widetilde{u}_i^{(\ell)}\|^2\|\widetilde{h}^{(\ell)}\|^2 \le \kappa\sum_{\ell=1}^{J}\|\widetilde{h}^{(\ell)}\|^2 = \kappa.$$

Thus all $2n+1$ populations have variance at most κ and range at most 2κ.

Uniform control of sampling errors. For sampling without replacement, the Bernstein inequality of Bardenet & Maillard (2015, Proposition 1.4), applied to a population and to its negation, states that a population with mean μ, variance σ^2, and range R_0 has a sample mean $\overline{Q}$ satisfying

$$\Pr\{|\overline{Q} - \mu| > \epsilon\} \le 2\exp\left(-\frac{k\epsilon^2}{2\sigma^2 + (2/3)R_0\epsilon}\right). \tag{32}$$

Since $\epsilon < 1$, the denominator is at most $2\kappa + \frac{4}{3}\kappa\epsilon \le \frac{10}{3}\kappa$. A union bound over the n residual-norm populations, the reference-norm population, and the n inner-product populations gives

$$\Pr\{\text{some sample mean deviates by more than } \epsilon\} \le 2(2n+1)\exp\left(-\frac{3k\epsilon^2}{10\kappa}\right) \le \delta \tag{33}$$

under (30). On the complementary event, simultaneously for all i,

$$\left|\|R_S\widetilde{u}_i\|^2 - 1\right| \le \epsilon, \qquad \left|\|R_S\widetilde{h}\|^2 - 1\right| \le \epsilon, \qquad |\langle R_S\widetilde{u}_i, R_S\widetilde{h}\rangle - \chi_i| \le \epsilon.$$

From moments to scores. Write $\widehat{\chi}_i = \langle R_S\widetilde{u}_i, R_S\widetilde{h}\rangle$ and $d_i = \|R_S\widetilde{u}_i\|\|R_S\widetilde{h}\|$. The squared-norm bounds give $d_i \in [1-\epsilon, 1+\epsilon]$, so the selected vectors are nonzero. By (25) and (15),

$$|r_i^{\cos,S} - r_i^{\cos,\mathrm{full}}| = \left|\frac{\widehat{\chi}_i}{d_i} - \chi_i\right| \le \frac{|\widehat{\chi}_i - \chi_i| + |\chi_i|\,|1-d_i|}{d_i} \le \frac{2\epsilon}{1-\epsilon}, \tag{34}$$

where we used $|\chi_i| \le 1$. Taking the maximum over clients and applying Theorem 2 proves the proposition. □

The budget is sufficient but not necessary, and it is informative only when its right-hand side is below J. It covers subsets drawn uniformly at random for a given round; it does not cover nonuniform or adaptive selection, fixed hand-chosen slices such as the default one, or the dependence that a subset fixed at the start of training creates with later uploads through filtering.

D Weighted Aggregation under Imperfect Filtering

D.1 Aggregation Space and Error Quantities

Let $\bar{W}^{(t)}$ be the global anchor of the round, W_i the full aggregation parameters returned by client i, and $\Delta_i = W_i - \bar{W}^{(t)} \in \mathbb{R}^{D_a}$. The aggregation space may differ from the scoring space: for diffusion models, W_i concatenates the trained parameters and their EMA copy, and a single clipping factor is applied to the joint update (Appendix A.2).

With data-size weights $p_i > 0$, $\sum_i p_i = 1$, and the realized retained set $\mathcal{K}$, write

$$P_B = \sum_{i\in\mathcal{B}} p_i, \qquad P_K = \sum_{i\in\mathcal{K}} p_i, \qquad P_{BK} = \sum_{i\in\mathcal{B}\cap\mathcal{K}} p_i.$$

The benign reference and the aggregate are

$$\bar{\Delta}_B = \frac{1}{P_B}\sum_{i\in\mathcal{B}} p_i\Delta_i, \qquad \widehat{\Delta} = \frac{1}{P_K}\sum_{i\in\mathcal{K}} p_i C_{\tau_{\text{clip}}}(\Delta_i), \qquad \tau_{\text{clip}} = \text{med}_{i\in\mathcal{K}}\,\|\Delta_i\|, \tag{35}$$

where $C_{\tau_{\text{clip}}}$ is the radial clipping operator in (8). The quantities α, γ, V_B, and B_{clip} are those of Section 5.3, and $\gamma < 1$ is equivalent to $P_{BK} > 0$. Because α and γ are weight fractions, γ equals the benign false-rejection rate only when benign clients hold equal data sizes, and α is the retained malicious weight rather than a miss rate. The reference $\bar{\Delta}_B$ is formed from the benign uploads of the same round, so the bound measures the distortion introduced by filtering and clipping within that round.

D.2 Proof of Theorem 3

Proof. Let the weighted means of the retained benign updates before and after clipping be

$$\bar{\Delta}_{BK} = \frac{1}{P_{BK}}\sum_{i\in\mathcal{B}\cap\mathcal{K}} p_i\Delta_i, \qquad \bar{C}_{BK} = \frac{1}{P_{BK}}\sum_{i\in\mathcal{B}\cap\mathcal{K}} p_i C_{\tau_{\text{clip}}}(\Delta_i),$$

and decompose the error as

$$\widehat{\Delta} - \bar{\Delta}_B = (\widehat{\Delta} - \bar{C}_{BK}) + (\bar{C}_{BK} - \bar{\Delta}_{BK}) + (\bar{\Delta}_{BK} - \bar{\Delta}_B). \tag{36}$$

Retained malicious influence. If $\alpha = 0$, then $\widehat{\Delta} = \bar{C}_{BK}$ and the first term vanishes. Otherwise, let

$$\bar{C}_{AK} = \frac{\sum_{i\in\mathcal{A}\cap\mathcal{K}} p_i C_{\tau_{\text{clip}}}(\Delta_i)}{\sum_{i\in\mathcal{A}\cap\mathcal{K}} p_i},$$

so that $\widehat{\Delta} = (1-\alpha)\bar{C}_{BK} + \alpha\bar{C}_{AK}$. Every clipped update, and hence every weighted mean of clipped updates, lies in the ball of radius τ_{clip}, so

$$\|\widehat{\Delta} - \bar{C}_{BK}\| = \alpha\|\bar{C}_{AK} - \bar{C}_{BK}\| \le 2\alpha\tau_{\text{clip}}.$$

Benign clipping bias. Radial clipping satisfies $\|C_{\tau_{\text{clip}}}(\Delta_i) - \Delta_i\| = (\|\Delta_i\| - \tau_{\text{clip}})_+$, so by the triangle inequality,

$$\|\bar{C}_{BK} - \bar{\Delta}_{BK}\| \le \frac{1}{P_{BK}}\sum_{i\in\mathcal{B}\cap\mathcal{K}} p_i\|C_{\tau_{\text{clip}}}(\Delta_i) - \Delta_i\| = B_{\text{clip}}.$$

Benign false rejections. If $\gamma = 0$, all benign weight is retained, $\bar{\Delta}_{BK} = \bar{\Delta}_B$, and the third term vanishes. Otherwise $0 < \gamma < 1$; let $\mathcal{B}_D = \mathcal{B}\setminus\mathcal{K}$ be the rejected benign clients and

$$\bar{\Delta}_{BD} = \frac{1}{\gamma P_B}\sum_{i\in\mathcal{B}_D} p_i\Delta_i.$$

Then $\bar{\Delta}_B = (1-\gamma)\bar{\Delta}_{BK} + \gamma\bar{\Delta}_{BD}$, and with $\zeta = \bar{\Delta}_{BK} - \bar{\Delta}_B$ we have $\bar{\Delta}_{BD} - \bar{\Delta}_B = -\frac{1-\gamma}{\gamma}\zeta$. Since V_B is at least its between-group part for the partition of $\mathcal{B}$ into retained and rejected clients,

$$V_B \ge (1-\gamma)\|\bar{\Delta}_{BK} - \bar{\Delta}_B\|^2 + \gamma\|\bar{\Delta}_{BD} - \bar{\Delta}_B\|^2 = \left[(1-\gamma) + \frac{(1-\gamma)^2}{\gamma}\right]\|\zeta\|^2 = \frac{1-\gamma}{\gamma}\|\zeta\|^2, \tag{37}$$

and therefore $\|\bar{\Delta}_{BK} - \bar{\Delta}_B\| \le \sqrt{\gamma V_B/(1-\gamma)}$.

Combining the three bounds with (36) gives

$$\|\widehat{\Delta} - \bar{\Delta}_B\| \le 2\alpha\tau_{\text{clip}} + \sqrt{\frac{\gamma}{1-\gamma}V_B} + B_{\text{clip}}. \qquad \square$$

D.3 From Detection and Decision Preservation to Aggregation Error

For fixed uploads, call an attacker certified if the full-parameter conditions of Theorem 1 hold for it and $\underline{\Gamma}_i^{\text{full}} > 2(1+\lambda)\xi_S$, and let $\mathcal{A}_{\text{cert}}(S)$ be the set of certified attackers. By Theorems 1 and 2, $\mathcal{A}_{\text{cert}}(S) \cap \mathcal{K} = \varnothing$. Let

$$\omega_{\text{rem}} = \sum_{i \in \mathcal{A} \setminus \mathcal{A}_{\text{cert}}(S)} p_i, \qquad P_{AK} = \sum_{i \in \mathcal{A} \cap \mathcal{K}} p_i.$$

Then $P_{AK} \le \omega_{\text{rem}}$ and $P_{BK} = (1-\gamma)P_B > 0$, and since $x \mapsto x/(x + P_{BK})$ is increasing on $[0, \infty)$,

$$\alpha = \frac{P_{AK}}{P_{AK} + P_{BK}} \le \frac{\omega_{\text{rem}}}{\omega_{\text{rem}} + (1-\gamma)P_B}. \tag{38}$$

Substituting into Theorem 3 yields

$$\|\widehat{\Delta} - \bar{\Delta}_B\| \le \frac{2\tau_{\text{clip}}\,\omega_{\text{rem}}}{\omega_{\text{rem}} + (1-\gamma)P_B} + \sqrt{\frac{\gamma}{1-\gamma}V_B} + B_{\text{clip}}. \tag{39}$$

Certified attackers therefore contribute nothing to the first term, and only the weight ω_{rem} of uncertified attackers can enter it; ω_{rem} bounds rather than equals the retained malicious weight, since uncertified attackers may still be rejected. On the success event of Proposition C.1, the quantity $2(1+\lambda)\xi_S$ in the certification condition can be replaced by $4(1+\lambda)\epsilon/(1-\epsilon)$; the replacement holds realization by realization, so $\mathcal{K}$, γ, and τ_{clip} need not be independent of S. If every attacker is certified, then $\omega_{\text{rem}} = 0$ and only the false-rejection and clipping terms remain, so clipping can then only add the bias B_{clip}.

D.4 Clipping Radius and Scope

Theorem 3 is stated in terms of the realized radius. If benign clients form a strict count majority of the retained set, $|\mathcal{B} \cap \mathcal{K}| > |\mathcal{K}|/2$, both central order statistics of $\{\|\Delta_i\|\}_{i \in \mathcal{K}}$ are at most the largest retained benign norm, so

$$\tau_{\text{clip}} \le \max_{i \in \mathcal{B} \cap \mathcal{K}} \|\Delta_i\|. \tag{40}$$

Because the MAD rule retains at least $\lceil n/2 \rceil$ clients (Section 4, Stage 2), this majority holds regardless of detection whenever $\rho < 1/4$; otherwise it requires enough attackers to be rejected. A benign majority before filtering or a majority of benign weight does not suffice for this bound. The analysis assumes nonzero residuals and reference vectors and the standard clipped aggregation in Section 4 (Stage 3); numerical conventions for degenerate inputs are given in Appendix A.2.