WHEN TRUST BECOMES A LIABILITY: BOUNDING PERSISTENT INFLUENCE UNDER TEMPORALLY ADAPTIVE ATTACKS IN FEDERATED LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Reputation-based defenses in federated learning can make historical trust persist in aggregation. As a result, benign behavior in earlier rounds may continue to affect a client's aggregation weight after it turns malicious. We study this vulnerability under temporally adaptive adversaries and argue that aggregation influence should decay unless it is continually re-earned. We introduce **DAIRE-FL** (**D**ecaying **A**ggregation **I**nfluence with **R**enewal and **E**nforcement), which separates behavioral acceptance, current aggregation weight, and future influence renewal to limit persistent historical influence. We derive finite-horizon guarantees that bound both inherited and re-earned influence after a behavioral pivot. Our threat model also considers pseudonymous re-enrollment and data reuse to distinguish persistent influence from increased adversarial participation. Experiments on CIFAR-10, CIFAR-100, FEMNIST, and Shakespeare under abrupt, gradual, threshold-aware, and layer-selective attacks show that DAIRE-FL substantially limits post-pivot adversarial influence while preserving benign-task performance.

1 INTRODUCTION

Federated learning (FL) requires a server to aggregate updates from clients whose data and intentions it cannot directly inspect (McMahan et al., 2017; Kairouz & McMahan, 2021). Many defenses therefore use current or historical behavior to determine participation and aggregation weight (Kang et al., 2019; 2020; Lyu et al., 2023; Zhang et al., 2022; Rashid et al., 2025). This creates a temporal vulnerability: influence accumulated during benign participation may remain effective after a client's behavior changes. An adversary may behave benignly until a pivot at round T_p and then act maliciously while retaining its accumulated aggregation influence. We call this *persistent historical influence* and study it under *temporally adaptive adversaries*.

Prior work considers multi-round behavioral histories, on–off adversaries, adaptive reputation, and mechanisms that discount older observations (Awan et al., 2021; Lewis et al., 2023; Zhang et al., 2022; Rashid et al., 2025; Gu et al., 2025). However, these methods use history mainly for detection, reputation estimation, client selection, or aggregation; they do not directly characterize how much influence accumulated before a pivot remains usable afterward. This leads to our central question: *how much pre-pivot aggregation influence can persist after behavior changes, and how much additional influence must be re-earned?*

Positioning and novelty. We treat retained post-pivot aggregation influence as the security quantity to control. We introduce **DAIRE-FL** (**D**ecaying **A**ggregation **I**nfluence with **R**enewal and **E**nforcement), which separates behavioral acceptance, current aggregation weight, and future influence renewal. Influence decays unless re-earned. Probationary acceptance may renew future influence but cannot affect the current model or same-round detector adaptation. The adaptive detector is also constrained so that accepted behavior cannot freely shift future acceptance. DAIRE-FL makes no identity-scarcity claim: qualifying data may be reused, resampled, or synthesized across arbitrarily many pseudonymous identities. Enrollment provides a committed conditioning reference and verifies the encoded enrollment relation, but it does not establish uniqueness, provenance, or raw-data validity. We therefore distinguish *adversarial occupancy* from *post-deviation influence*. Identity proliferation

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.

determines the adversarial share of sampled clients, whereas DAIRE-FL limits the influence retained by each identity after its behavior changes. Our contributions are:

- **Persistent influence under temporal adaptivity.** We formalize retained post-pivot aggregation influence as a distinct temporal security quantity and show how influence accumulated during benign participation remains usable after a behavioral change. We also characterize the participation limit imposed by unrestricted identity proliferation under identity-symmetric sampling.
- **Decaying influence with renewal and enforcement.** We introduce DAIRE-FL, in which influence decays unless re-earned, probationary acceptance cannot restore current-round influence, and q-of-m attestation prevents unilateral renewal. Drift-constrained adaptation further limits manipulation of future acceptance.
- **Finite-horizon guarantees and adaptive evaluation.** We derive a pathwise finite-horizon bound that separates influence inherited before the pivot from influence re-earned afterward. We also provide conditional probabilistic refinements under bounds on post-pivot acceptance and participation. We evaluate abrupt, gradual, threshold-aware, and layer-selective attacks, detector adaptation, and dataset-reuse identity proliferation across CIFAR-10, CIFAR-100, FEMNIST, and Shakespeare.

2 Related Work

Robust aggregation and adaptive evasion. Byzantine-resilient methods such as Krum (Blanchard et al., 2017), RFA (Pillutla et al., 2022), and FLAME (Nguyen et al., 2022) suppress updates that are geometric outliers or inconsistent with the dominant client population. More recent attacks show that damaging updates can be optimized to remain within defense-defined acceptance regions or concentrated in selected parameter subsets (Fang et al., 2020; Shejwalkar & Houmansadr, 2024; Krauß et al., 2024; Zhao et al., 2026; Guo et al., 2026). These results motivate evaluating both threshold-aware and layer-selective adversaries.

Temporal behavior, reputation, and client history. FL defenses use cross-round information for reputation, selection, detection, and aggregation. CONTRA maintains historical reputation, MAB-RFL adapts client selection, FLDetector detects multi-round inconsistency, and Viceroy studies on–off behavior (Awan et al., 2021; Wan et al., 2022; Zhang et al., 2022; Lewis et al., 2023). Reputation-based FL, AdaAggRL, and DP-BREM likewise exploit accumulated history for trust-aware aggregation or multi-round detection (ur Rehman et al., 2020; Chen et al., 2024; Rashid et al., 2025; Lyu et al., 2023; Wang et al., 2025; Gu et al., 2025). Recent defenses broaden this line: STAR-FL combines spatial relations among updates with within-client cross-round consistency, while SHIELD-FL combines gradient trust indexing, adaptive clustering, and data-free distillation (Tabassum & Wu, 2026; Soliman et al., 2026). Adaptive reputation attacks show that adversaries can strategically alternate benign and malicious participation (Liu et al., 2026), whereas behavior-based Sybil defenses identify coordinated or statistically similar identities (Fung et al., 2020; Jiang et al., 2021). DAIRE-FL instead treats retained aggregation influence as an explicit state. Here, current aggregation uses pre-renewal influence, future influence must be re-earned. The analysis bounds inherited and renewed influence separately without assuming identity scarcity.

Behavioral modeling, verification, and positioning. Reconstruction-based defenses model benign update structure and detect deviations from it. FedCVAE uses conditional generative modeling for malicious-update detection (Gu & Yang, 2021), while FedREDefense uses reconstruction error to distinguish genuine and crafted model updates (Xie et al., 2024). Cryptographic methods such as RoFL (Lycklama et al., 2023), VPFL (Duan et al., 2024), and related verifiable-training schemes (Akgul et al., 2025) certify properties of submitted updates or local computation, but such verification does not establish benign intent, identity uniqueness, or data provenance. DAIRE-FL instead treats *retained post-pivot aggregation influence* as the quantity to control, combining explicit decay and renewal with constrained detector adaptation and attested enforcement. This enables finite-horizon separation of influence inherited from pre-pivot behavior from influence re-earned after deviation.

3 Problem Setup and Threat Model

3.1 Federated setup. A coordinator trains global models $\{w_t\}$ over enrolled identities $\mathcal{V}_t$. Identity i enrolls at round e_i and belongs to $\mathcal{V}_t$ only for $t \geq e_i$. At round t, the coordinator samples

$\mathcal{S}_t \subseteq \mathcal{V}_t$, receives $\Delta w_i(t) \triangleq w_t - w_i^{\text{local}}$ from each selected client, and updates $w_{t+1} = w_t - \eta_g \sum_{i \in \mathcal{S}_t} p_i(t) \Delta w_i(t)$ using normalized weights $p_i(t)$. Define the sampling indicator $J_i(t) = \mathbf{1}\{i \in \mathcal{S}_t\}$, protocol acceptance $A_i(t) \in \{0, 1\}$, and renewal indicator $R_i(t) = J_i(t)A_i(t)$. Sampling is independent of influence state: clients with $B_i(t) < \tau_{\min}$ may still be sampled but receive zero current-round weight, while positive weight requires $R_i(t) = 1$ and $B_i(t) \geq \tau_{\min}$. Superscripts t index detector observations, parentheses index protocol-state and indicator processes, and H and M denote honest and malicious categories. The symbols $B_i(t)$ and $\mathcal{B}_t$ denote influence state and replay buffer, while S_i and $\mathcal{S}_t$ denote an enrollment summary and sampled set.

3.2 Threat model. The adversary $\mathcal{A}$ seeks to degrade global utility or implant a backdoor while evading behavioral rejection. We allow: **(P1) Pseudonymous re-enrollment and unlimited identities:** arbitrarily many pseudonyms may enroll, including after calibration or a pivot; DAIRE-FL neither links pseudonyms nor transfers influence between them. **(P2) Dataset reuse:** qualifying data may be reused, resampled, or synthesized across identities, so enrollment provides no identity scarcity. **(P3) Arbitrary warm-up:** an identity may behave benignly for any number of rounds before deviating. **(P4) Temporal adaptivity:** deviations may be abrupt, gradual, or optimized against the evolving detector and threshold. **(P5) White-box detector access:** $\mathcal{A}$ knows the monitored projection, frozen and adaptive references, scoring rule, reference-update constraint, threshold rule, and current threshold, and may differentiate through them to optimize attack utility and future detector adaptation. **(P6) Spatial adaptivity:** malicious changes may target arbitrary parameter subsets while monitored components remain close to benign behavior.

DAIRE-FL assumes append-only committed state, zero-knowledge verification of the enrollment relation, and q-of-m attesters that independently recompute the committed acceptance rule; details appear in §4.3 and Appendix A.2. The adversary cannot forge a valid proof for a witness that violates the encoded relation or alter committed ledger state. The main analysis assumes $T_{\text{cal}} < T_p$: identities that later turn malicious may participate during calibration but remain benign until the pivot. Calibration-time poisoning is therefore outside the post-pivot detector guarantee. Section 5 addresses temporal adaptivity, while spatial adaptivity is addressed by the detector and evaluated separately. The unconditional guarantee concerns normalized aggregation influence. Bounds on model displacement additionally require bounded updates; the experiments impose the same $\delta_{\max}$ attack budget across methods, but behavioral acceptance alone does not imply this condition.

3.3 Enrollment boundary. Enrollment verifies only that committed conditioning metadata satisfy the encoded relation. It does not establish dataset uniqueness, provenance, ownership, label correctness, identity uniqueness, or identity scarcity; it is therefore not a Sybil defense. The finite-horizon guarantee applies separately to each identity from its realized state; it neither provides cross-pseudonym reputation continuity nor bounds the total influence accumulated through replacement pseudonyms.

The evaluation fixes the enrolled population before calibration. The proliferation study varies the number of pre-enrolled adversarial identities but does not consider dynamic post-calibration creation or replacement. We distinguish *re-entry*, where the same identity is sampled after $B_i(t) < \tau_{\min}$, from *re-enrollment*, where a new pseudonym begins from the prescribed initial state. The analysis controls re-entry but does not link states across re-enrollment.

The evaluation also assumes that S_i is generated from the client's declared local statistics; strategic use of a syntactically valid but fabricated S_i to aid evasion is not evaluated. The finite-horizon bound remains independent of the correctness of S_i because it depends only on realized renewals.

3.4 Identity-proliferation limit. Suppose N_H honest and K adversarial identities are enrolled, available for sampling, and behaviorally indistinguishable before a pivot. If n identities are sampled uniformly without replacement and M_t denotes the number of selected adversarial identities, then

$$\mathbb{E}\left[\frac{M_t}{n}\right] = \frac{K}{N_H + K} \longrightarrow 1 \qquad \text{as } K \to \infty. \tag{1}$$

Thus identity-symmetric sampling provides no uniform positive lower bound on the honest sampled fraction. We call $K/(N_H + K)$ the *adversarial occupancy*, which is distinct from the post-pivot influence that DAIRE-FL controls per identity. DAIRE-FL therefore makes no robustness claim under an unbounded malicious participation fraction; the optimization result in Appendix B.6 separately requires a positive effective honest-gradient coefficient.

4 DAIRE-FL

DAIRE-FL maintains a decaying influence state for each identity and renews it only after accepted participation. Behavioral detection governs acceptance, enrollment supplies statistical conditioning, and attestation enforces renewal. Figure 1 summarizes the framework; Appendices A.3, A.2, and C provide further details.

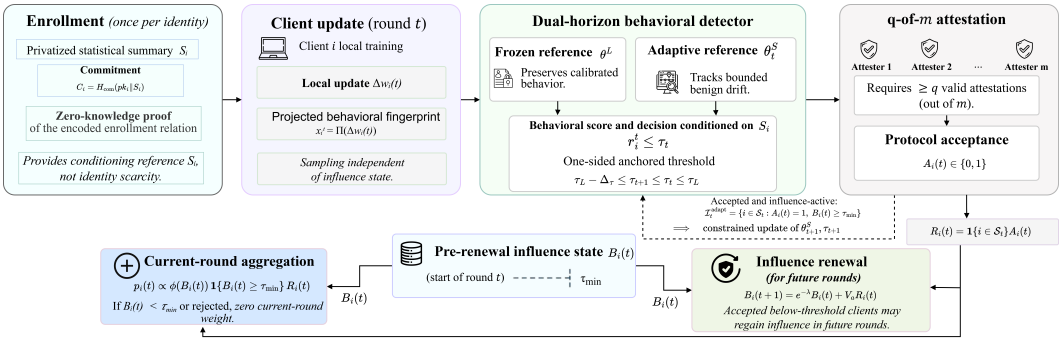


Figure 1: DAIRE-FL. Current aggregation uses pre-renewal influence $B_i(t)$; dual-horizon behavioral consistency determines acceptance, and q-of-m attestation authorizes future renewal. Enrollment provides statistical conditioning but not identity scarcity.

4.1 Decaying aggregation influence and renewal

Let e_i be identity i's enrollment round and $a_i \triangleq \max\{e_i, T_{\text{cal}}\}$ its online activation round. For $t \ge a_i$, initialize $B_i(a_i) = B_0$ and update

$$B_i(t+1) = e^{-\lambda} B_i(t) + V_a R_i(t), \qquad t \ge a_i, \tag{2}$$

where $\lambda > 0$ controls decay and $V_a > 0$ is the renewal increment. Our experiments fix enrollment before calibration, so $e_i = 0$ and $a_i = T_{\text{cal}}$. Since $R_i(t) \le 1$, $B_i(t) \le B^{\max}(B_0) \triangleq B_{\max}$, where $B^{\max}(b) \triangleq \max\{b, V_a/(1-e^{-\lambda})\}$. Thus benign warm-up cannot accumulate unbounded influence. Current aggregation uses the pre-renewal state:

$$\widetilde{p}_i(t) = \phi(B_i(t))\mathbf{1}\{B_i(t) \ge \tau_{\min}\} R_i(t), \qquad \phi(x) = \log(1+x), \tag{3}$$

$$p_i(t) = \begin{cases} \widetilde{p}_i(t)/Z_t, & Z_t > 0, \\ 0, & Z_t = 0, \end{cases} \qquad Z_t = \sum_{j \in \mathcal{S}_t} \widetilde{p}_j(t). \tag{4}$$

Rejected or below-threshold updates receive zero current weight. An accepted probationary client may renew $B_i(t+1)$ but cannot affect the current model or same-round detector adaptation. Decay is evaluated lazily from the last committed state.

4.2 Dual-horizon behavioral consistency

DAIRE-FL uses a frozen long-horizon reference to preserve calibrated behavior and a constrained adaptive reference to accommodate bounded benign drift. Both are conditioned on enrollment summary S_i. To limit detector self-poisoning, reference and threshold movement are constrained, and probationary updates cannot adapt either state.

Multi-layer fingerprint. For monitored layers $\mathcal{L}_{\text{mon}} = \{\ell_1, \dots, \ell_{n_{\text{mon}}}\}$, where $n_{\text{mon}} \triangleq |\mathcal{L}_{\text{mon}}|$, let d_ℓ be the raw layer dimension, $\mu_\ell^{\text{PCA}} \in \mathbb{R}^{d_\ell}$ the frozen calibration mean, and $P_\ell \in \mathbb{R}^{d'_\ell \times d_\ell}$ the frozen PCA projection. Define

$$\begin{aligned} \Pi_\ell(v) &\triangleq \frac{P_\ell\big(\mathrm{vec}(v) - \mu_\ell^{\text{PCA}}\big)}{\sqrt{d'_\ell}}, \qquad \ell \in \mathcal{L}_{\text{mon}}, \\ x_i^t = \Pi(\Delta w_i(t)) &\triangleq \mathrm{concat}_{j=1}^{n_{\text{mon}}} \Pi_{\ell_j}\big(\Delta w_{i,\ell_j}(t)\big). \end{aligned} \tag{5}$$

Here $d' = \sum_{\ell \in \mathcal{L}_{\text{mon}}} d'_\ell$, and scaling by $1/\sqrt{d'_\ell}$ limits domination by larger layer blocks.

Frozen and adaptive references. Calibration is split into fit and held-out portions. The projections and client-balanced conditional variational autoencoder (C-VAE) θ^L are fitted on the first portion and frozen, while $\theta^S_{T_{\text{cal}}} = \theta^L$. The held-out portion calibrates functional displacement and the threshold. Let $E_\theta(x, S) = d'^{-1}\|x - \widehat{x}_\theta(x, S)\|_2^2$. Appendix A.3 defines $d_{\text{ref}}(x, S; \theta^S, \theta^L)$, which combines normalized reconstruction displacement and symmetric KL divergence between latent posteriors. The behavioral score is

$$r_i^t = \alpha E_{\theta^L}(x_i^t, S_i) + (1-\alpha) E_{\theta_t^S}(x_i^t, S_i) + \gamma d_{\text{ref}}(x_i^t, S_i; \theta_t^S, \theta^L). \tag{6}$$

Drift-constrained adaptation. Let $D_{\mathcal{C}}(\theta, \theta^L)$ denote client-balanced displacement on the held-out calibration set. DAIRE-FL enforces

$$D_{\mathcal{C}}(\theta_t^S, \theta^L) \le \Delta_M \qquad \text{for all } t. \tag{7}$$

Let $\mathcal{B}_t$ be the bounded replay buffer at the start of round t. Only accepted clients with active pre-renewal influence may adapt the detector:

$$\mathcal{I}_t^{\text{adapt}} \triangleq \{i \in \mathcal{S}_t : A_i(t) = 1,\ B_i(t) \ge \tau_{\min}\}.$$

This gate prevents probationary updates from shaping the reference used to evaluate them. After round-t acceptance is fixed, form $\mathcal{B}_t^+ \triangleq \text{UpdateBuffer}(\mathcal{B}_t, \{(x_i^t, S_i) : i \in \mathcal{I}_t^{\text{adapt}}\})$ using deterministic ordering and capacity rules. Training on $\mathcal{B}_t^+$ produces an unconstrained candidate $\widetilde{\theta}_{t+1}^S$. Backtracking then chooses the largest $s_t \in \{0, 1, 1/2, \ldots\}$ satisfying Eq. 7, and sets

$$\theta_{t+1}^S = \theta_t^S + s_t(\widetilde{\theta}_{t+1}^S - \theta_t^S). \tag{8}$$

If $\mathcal{B}_t^+ = \emptyset$, the reference remains unchanged.

Anchored threshold adaptation. Let τ_L be the empirical ν-quantile of frozen-reference reconstruction scores on the held-out set, with $\tau_{T_{\text{cal}}} = \tau_L$. The same eligibility gate controls threshold adaptation. For the bounded score window $\mathcal{W}_t$, define $\mathcal{W}_t^+ \triangleq \text{UpdateWindow}(\mathcal{W}_t, \{r_i^t : i \in \mathcal{I}_t^{\text{adapt}}\})$. Then

$$\tau_{t+1} = \begin{cases} \max\{\tau_L - \Delta_\tau, \min\{\tau_t, Q_\nu(\mathcal{W}_t^+)\}\}, & \mathcal{W}_t^+ \neq \emptyset, \\ \tau_t, & \mathcal{W}_t^+ = \emptyset. \end{cases} \tag{9}$$

The threshold may tighten but cannot fall below $\tau_L - \Delta_\tau$ or exceed its current value. Thus accepted updates cannot progressively relax the calibrated boundary. Behavioral admissibility requires $r_i^t \le \tau_t$, while renewal also requires attestation. Larger benign shifts require recalibration.

4.3 Statistical conditioning and attested enforcement

Statistical conditioning. Each client commits to a privatized task-dependent summary S_i and proves in zero knowledge that its declared witness satisfies the encoded enrollment relation, including the minimum enrollment-size requirement. Vision summaries use coarse intensity statistics; next-character summaries use character frequencies and hash-binned transitions. The contract stores $C_i = H_{\text{com}}(pk_i \| S_i)$, where H_{com} is Poseidon over the circuit's field encoding. The summary conditions the shared C-VAE but, as established in §3, does not verify raw-data correctness, identity scarcity, or provenance. Appendix A.2 provides construction and privacy details; Theorem 1 depends only on realized renewals.

Attested acceptance and renewal. Because detector evaluation occurs off-chain, renewal requires q valid attestations over the identity, round, update hash, detector state, nonce, chain, and contract. Let D_t^{det} be the committed detector-state digest, including the reference versions, threshold, and buffer and window states, and let H_{upd} be Keccak-256 over a canonical update serialization. Each attester signs

$$V_i^t = \big(i, t, H_{\text{upd}}(\Delta w_i(t)), r_i^t, \tau_t, D_t^{\text{det}}, \text{nonce}_t, \text{chainID}, \text{contractID}\big).$$

Protocol acceptance requires $r_i^t \le \tau_t$ and $\sum_{j=1}^m \mathbf{1}\{\text{VerifySig}_j(V_i^t) = 1\} \ge q$; otherwise $A_i(t) = R_i(t) = 0$. With at most f Byzantine attesters, $f < q \le m - f$ provides safety and liveness. The quorum prevents unilateral fabrication but does not certify benign intent or provenance. Equation 2 then updates only the future state $B_i(t+1)$.

5 THEORETICAL ANALYSIS

We analyze post-pivot influence under DAIRE-FL's decay and renewal dynamics. Let $H \triangleq T - T_p$ be the post-pivot horizon, and reindex its rounds as $t = 0, \dots, H-1$, with $B_k(0) = b_k$ denoting identity k's inherited pre-renewal influence. The main result is pathwise; probability bounds are used only for expected refinements. This distinction keeps the core guarantee valid even when no uniform acceptance probability can be certified. Complete proofs and further consequences appear in Appendix B.

Post-pivot renewal. For any adaptive strategy of a malicious identity k, suppose $\beta_A, \pi_{\max} \in [0,1]$ satisfy

$$\Pr[A_k(t) = 1 \mid J_k(t) = 1, \mathcal{G}_t] \le \beta_A, \qquad \Pr[J_k(t) = 1 \mid \mathcal{G}_t] \le \pi_{\max}, \tag{10}$$

where $\mathcal{G}_t$ is the transcript before round-t selection. Since $R_k(t) = J_k(t)A_k(t)$,

$$\mathbb{E}[R_k(t) \mid \mathcal{G}_t] \le \pi_{\max}\beta_A \triangleq \beta_R. \tag{11}$$

Here β_A bounds acceptance conditional on submission, whereas β_R bounds renewal per global round. This factorization separates detector acceptance from the opportunity to participate. These are analytical conditions, not guarantees supplied by the C-VAE. If $\Pr[J_k(t) = 1 \mid \mathcal{G}_t] = 0$, then $R_k(t) = 0$ almost surely and the first conditional bound is immaterial.

Expected contraction. Iterating Eq. 2 gives

$$\mathbb{E}[B_k(t)] \le \bar{B} + e^{-\lambda t}(b_k - \bar{B}), \qquad \bar{B} \triangleq \frac{\beta_R V_a}{1 - e^{-\lambda}}. \tag{12}$$

Thus $\limsup_{t\to\infty} \mathbb{E}[B_k(t)] < \tau_{\min}$ when

$$\boxed{\beta_R < \frac{\tau_{\min}(1 - e^{-\lambda})}{V_a}} \tag{13}$$

or, equivalently, $\pi_{\max}\beta_A < \tau_{\min}(1-e^{-\lambda})/V_a$. This condition is needed only for expected contraction. We assume $\tau_{\min} > 0$. Whenever $Z_t > 0$, Eqs. 3–4 give $Z_t \ge \phi(\tau_{\min}) = \log(1+\tau_{\min}) > 0$, providing a deterministic lower bound on every positive normalizer.

Pathwise finite-horizon bound.

Theorem 1 (Finite-horizon persistent-influence bound)**.** *Under Eq. 2, define* $\mathcal{R}_k(H) \triangleq \sum_{t=0}^{H-1} R_k(t)$, $\Psi_k(H) \triangleq \sum_{t=0}^{H-1} p_k(t)$, *and* $\Omega_k(H) \triangleq \sum_{t=0}^{H-1} p_k(t)\|\Delta w_k(t)\|$, *with* $p_k(t) = 0$ *when* $k \notin \mathcal{S}_t$. *Let* $Z_{\min} > 0$ *lower-bound every positive normalizer;* $Z_{\min} = \phi(\tau_{\min})$ *is always valid. Define*

$$U_k^{\Psi}(H) \triangleq \min\left\{H,\ \frac{b_k + V_a\mathcal{R}_k(H)}{Z_{\min}(1 - e^{-\lambda})},\ \frac{b_k + V_a\mathcal{R}_k(H)}{\tau_{\min}(1 - e^{-\lambda})}\right\}. \tag{14}$$

The three branches respectively capture the horizon cap, normalization by $Z_{\min}$, *and the number of rounds for which the state can remain influence-active. Then every realized renewal sequence satisfies*

$$\boxed{\Psi_k(H) \le U_k^{\Psi}(H).} \tag{15}$$

If $\|\Delta w_k(t)\| \le \delta_{\max}$ *throughout the post-pivot horizon, then*

$$\boxed{\Omega_k(H) \le \delta_{\max}\Psi_k(H) \le \delta_{\max}U_k^{\Psi}(H).} \tag{16}$$

The first bound is independent of update magnitude; the second requires bounded updates. Neither assumes absorbing gating or geometric occupation.

The bound separates inherited influence, through b_k, from influence re-earned through $V_a\mathcal{R}_k(H)$. The former decays geometrically, while sustained post-pivot influence requires continued renewal. It therefore remains valid under adaptive renewal paths.

Corollary 1 (Expected influence under bounded renewal)**.** *If Eq. 11 holds in every post-pivot round, then* $\mathbb{E}[\mathcal{R}_k(H)] \le \beta_R H$ *and*

$$\mathbb{E}[\Psi_k(H)] \le \frac{b_k + \beta_R V_a H}{Z_{\min}(1 - e^{-\lambda})}, \qquad \mathbb{E}[\Omega_k(H)] \le \frac{\delta_{\max}(b_k + \beta_R V_a H)}{Z_{\min}(1 - e^{-\lambda})}, \tag{17}$$

where the second inequality additionally assumes $\|\Delta w_k(t)\| \le \delta_{\max}$.

Connection to measurements. Let $N_k^{\mathrm{sub}}(H) \triangleq \sum_{t=0}^{H-1} J_k(t)$, $\widehat{\beta}_{R,k} \triangleq \mathcal{R}_k(H)/H$, and, when $N_k^{\mathrm{sub}}(H) > 0$, $\widehat{\beta}_{A,k} \triangleq \mathcal{R}_k(H)/N_k^{\mathrm{sub}}(H)$. Since $\mathcal{R}_k(H) = \widehat{\beta}_{R,k}H = \widehat{\beta}_{A,k}N_k^{\mathrm{sub}}(H)$, Theorem 1 gives

$$\Psi_k(H) \le \frac{b_k + V_a\widehat{\beta}_{R,k}H}{Z_{\min}(1-e^{-\lambda})} = \frac{b_k + V_a\widehat{\beta}_{A,k}N_k^{\mathrm{sub}}(H)}{Z_{\min}(1-e^{-\lambda})}. \tag{18}$$

Thus $\widehat{\beta}_{A,k}$ measures realized acceptance per submission, while $\widehat{\beta}_{R,k}$ measures realized renewal per post-pivot round. *Appendix B provides the exact state unrolling, proofs, re-entry occupation bound, absorbing-gating refinement, and strongly convex optimization consequence.*

6 Experiments

6.1 Experimental Setup

We evaluate DAIRE-FL on CIFAR-10, CIFAR-100, FEMNIST, and Shakespeare. CIFAR-10 and CIFAR-100 use ResNet-18 adapted to 32×32 inputs with Dirichlet heterogeneity $\alpha_D \in \{1.0, 0.5, 0.3\}$. Controlled CIFAR experiments use $N = 100$ enrolled identities and sample $n = 10$ per round independently of influence state. FEMNIST uses its natural writer partition with a lightweight CNN, while Shakespeare uses its native client partition with a two-layer LSTM for next-character prediction. For both datasets, we use a fixed evaluation pool of $N = 100$ clients, including $K = 10$ adversarial clients, and sample $n = 10$ uniformly without replacement per round. Unless stated otherwise, training runs for $T = 200$ rounds, results are averaged over five independent seeds, and the main robustness experiments use $\alpha_D = 0.5$ for CIFAR and a 10% malicious population. The enrolled population is fixed across compared methods within each seed. For a pivot at round T_p, post-pivot rates and cumulative influence are evaluated over $\mathcal{T}_{\mathrm{post}} \triangleq \{T_p, \dots, T-1\}$, with horizon $H \triangleq T - T_p$. CIFAR-10 and FEMNIST support the full adaptive attack and ablation suite, while CIFAR-100 and Shakespeare test generalization under the strongest attacks.

Calibration and attacks. Detector, influence-decay, and renewal parameters are selected using benign-only development runs disjoint from the five evaluation seeds and frozen before malicious evaluation. We evaluate abrupt and gradual pivots, a white-box threshold-aware attack against detector adaptation, and a layer-selective backdoor against spatially restricted monitoring. Appendices C.1 and C.4 provide the calibration criteria, attack objectives, optimization procedures, and parameters.

Baselines and metrics. We compare DAIRE-FL with FedAvg (McMahan et al., 2017), Krum (Blanchard et al., 2017), FLAME (Nguyen et al., 2022), FoolsGold (Fung et al., 2020), FLDetector (Zhang et al., 2022), AdaAggRL (Wang et al., 2025), DP-BREM (Gu et al., 2025), and SPMC (He et al., 2025). These baselines cover undefended aggregation, Byzantine robustness, Sybil filtering, multi-round detection, adaptive aggregation, and backdoor defense. Where their input-validation assumptions can be instantiated, we also include RoFL (Lycklama et al., 2023) and Aion (Liu et al., 2025) as protocol-level validation baselines; neither directly addresses temporal history. We use ASR as the primary security metric and clean test accuracy as the benign-task metric; for Shakespeare, the latter is next-character prediction accuracy. For Modes A–C, ASR is the fraction of source-class test examples, or source-character positions for Shakespeare, mapped to the attacker-specified target after excluding examples already labeled as the target. For Mode D, ASR is the fraction of triggered inputs classified as the target. Clean accuracy is measured on unmodified test inputs. We also report client-balanced honest false rejection $\widehat{\alpha}_{\mathrm{FR}}$, cohort-level malicious conditional acceptance $\widehat{\beta}_A$, mean per-identity per-round malicious renewal $\widehat{\beta}_R$, and cumulative adversarial aggregation weight. Appendix C.3 defines the cohort-level rates. Tables report $\widehat{\beta}_A$ and $\widehat{\beta}_R$ as fractions, while prose may use equivalent percentages. Appendix Tables 5 and 6 provide all hyperparameters.

6.2 Results

6.2.1 Robustness under adaptive attacks. Table 1 evaluates abrupt and gradual pivots (Modes A/B), a white-box threshold-aware attack against detector adaptation (Mode C), and a layer-selective backdoor for spatial evasion (Mode D). On CIFAR-10, DAIRE-FL reduces ASR from 38.6% to 6.8% under Mode A, 45.2% to 9.7% under Mode B, 57.9% to 12.6% under Mode C, and 48.7% to 11.3% under Mode D, while clean accuracy remains between 80.8% and 81.9%. Under Mode C, DAIRE-FL

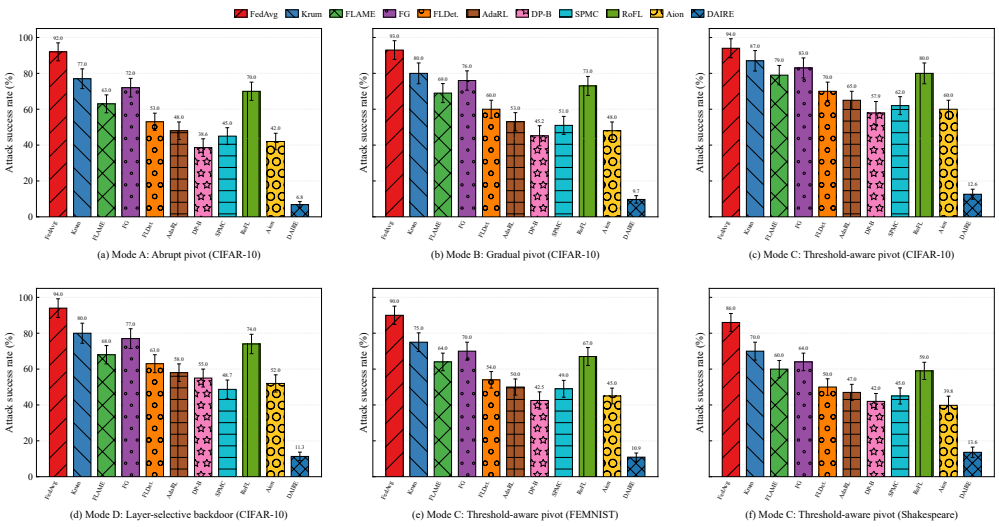


Figure 2: ASR of DAIRE-FL and baselines under Modes A–D on CIFAR-10 and Mode C on FEMNIST and Shakespeare. Error bars denote standard deviations over five seeds; lower is better.

achieves ASRs of $15.4 \pm 3.1\%$ on CIFAR-100, $10.9 \pm 2.3\%$ on FEMNIST, and $13.6 \pm 2.9\%$ on Shakespeare, compared with strongest-baseline ASRs of $55.6 \pm 6.1\%$, $42.5 \pm 4.8\%$, and $39.8 \pm 5.1\%$, respectively. Figure 2 shows the corresponding per-baseline comparisons. DAIRE-FL achieves the lowest ASR, outperforming the strongest baseline by 26.2–45.3 percentage points (6.8%–13.6% versus 38.6%–57.9%). *Complete per-dataset and per-baseline results appear in Appendix D.*

Table 1: DAIRE-FL under adaptive attacks.

Mode / attack	Dataset	ASR ↓		Clean acc. ↑	Adversarial rates ↓	
		Best baseline	DAIRE-FL		$\widehat{\beta}_A$	$\widehat{\beta}_R$
A / Abrupt pivot	CIFAR-10	38.6±4.9%	6.8±1.7%	81.9±0.7%	0.11±0.03	0.010±0.003
B / Gradual pivot	CIFAR-10	45.2±5.6%	9.7±2.1%	81.4±0.8%	0.15±0.04	0.014±0.004
C / Threshold-aware pivot	CIFAR-10	57.9±6.4%	12.6±2.8%	80.8±0.9%	0.18±0.04	0.017±0.005
	CIFAR-100	55.6±6.1%	15.4±3.1%	61.7±0.9%	0.20±0.05	0.019±0.005
	FEMNIST	42.5±4.8%	10.9±2.3%	86.7±0.6%	0.17±0.04	0.016±0.004
	Shakespeare	39.8±5.1%	13.6±2.9%	53.1±0.5%	0.18±0.04	0.017±0.004
D / Layer-selective backdoor	CIFAR-10	48.7±5.2%	11.3±2.4%	81.2±0.8%	0.16±0.03	0.015±0.004

6.2.2 Persistent influence and correspondence with theory. Table 2 reports cumulative adversarial weight against the ex post normalization diagnostic defined in Appendix D.2. Realized weight remains below the diagnostic in all settings: 5.84–9.12 versus 14.90–23.61 on CIFAR-10, and 6.02–9.46 versus 15.72–24.37 on Shakespeare. The weight-to-diagnostic ratio is 0.38–0.40 throughout. On CIFAR-10, threshold-aware acceptance reaches 18%, but renewal is only 1.7% per post-pivot round; the corresponding abrupt-attack rates are 11% and 1.0%. Honest false rejection remains 4.1%–4.9%.

6.2.3 Adaptive stress tests. Table 3 evaluates detector self-poisoning and spatial evasion on CIFAR-10 and FEMNIST. Under Mode C (Detector self-poisoning), removing influence decay raises CIFAR-10 ASR from 12.6±2.8% to 29.8±4.5%, while conditional acceptance changes only from 0.18±0.04 to 0.19 ± 0.04 and renewal from 0.017 ± 0.005 to 0.018 ± 0.005. Similar acceptance therefore does not imply similar post-pivot influence when historical influence persists. Under separately optimized Mode D (Spatial adaptivity) attacks, head-only monitoring yields $48.7 \pm 5.2\%$ ASR and conditional acceptance of 0.61 ± 0.06 on CIFAR-10, whereas multi-layer monitoring reduces them to $11.3 \pm 2.4\%$ and 0.16 ± 0.03. The attack is re-optimized for each monitored layer set, so the improvement is not due to transferring a head-specific attack. *Full ablations appear in Appendices D.3 and D.4.*

6.2.4 Identity proliferation and scope limit. Figure 3 examines pre-calibration identity proliferation by varying the number of adversarial identities in a fixed enrolled pool. This affects adversarial

Table 2: Post-pivot cumulative adversarial weight and observed normalization diagnostic.

Dataset	Attack	$\widehat{\alpha}_{\mathrm{FR}}$	Post-pivot influence		
			Cumulative weight	Observed diagnostic	Weight / diagnostic
CIFAR-10	Abrupt pivot	$4.1\pm0.8\%$	5.84 ± 0.72	14.90 ± 1.63	0.39 ± 0.05
	Gradual pivot	$4.5\pm0.9\%$	7.36 ± 0.91	18.74 ± 2.06	0.39 ± 0.06
	Threshold-aware pivot	$4.9\pm0.8\%$	9.12 ± 1.18	23.61 ± 2.47	0.39 ± 0.05
	Layer-selective backdoor	$4.8\pm0.9\%$	8.47 ± 1.04	21.98 ± 2.31	0.39 ± 0.05
Shakespeare	Abrupt pivot	$4.3\pm0.8\%$	6.02 ± 0.76	15.72 ± 1.71	0.38 ± 0.05
	Gradual pivot	$4.7\pm0.9\%$	7.81 ± 0.96	19.58 ± 2.09	0.40 ± 0.06
	Threshold-aware pivot	$4.1\pm0.8\%$	9.46 ± 1.22	24.37 ± 2.58	0.39 ± 0.05
	Layer-selective backdoor	$4.9\pm0.9\%$	8.91 ± 1.10	22.45 ± 2.38	0.40 ± 0.05

Table 3: Adaptive stress tests.

Stress test	Dataset	Variant	ASR ↓	Clean acc. ↑	Adversarial rates ↓	
					$\widehat{\beta}_A$	$\widehat{\beta}_R$
Detector self-poisoning (C)	CIFAR-10	Full DAIRE-FL	$12.6\pm2.8\%$	$80.8\pm0.9\%$	0.18 ± 0.04	0.017 ± 0.005
		Without influence decay	$29.8\pm4.5\%$	$80.5\pm1.0\%$	0.19 ± 0.04	0.018 ± 0.005
	FEMNIST	Full DAIRE-FL	$10.9\pm2.3\%$	$86.7\pm0.6\%$	0.17 ± 0.04	0.016 ± 0.004
		Without influence decay	$27.1\pm4.0\%$	$86.4\pm0.7\%$	0.18 ± 0.04	0.017 ± 0.004
Spatial evasion (D)	CIFAR-10	Head-only monitoring	$48.7\pm5.2\%$	$79.4\pm1.1\%$	0.61 ± 0.06	0.058 ± 0.007
		Multi-layer monitoring	$11.3\pm2.4\%$	$81.2\pm0.8\%$	0.16 ± 0.03	0.015 ± 0.004
	FEMNIST	Head-only monitoring	$44.9\pm4.9\%$	$85.3\pm0.9\%$	0.57 ± 0.06	0.054 ± 0.007
		Multi-layer monitoring	$10.7\pm2.3\%$	$86.5\pm0.7\%$	0.15 ± 0.03	0.014 ± 0.004

sampling occupancy rather than the per-identity influence bound. As K increases from 1 to 100, malicious sampling occupancy rises from 1.0% to 50.1%, honest aggregation mass falls from 0.992 to 0.642, cumulative adversarial aggregation weight rises from 0.84 to 44.58, and ASR increases from 3.8% to 38.9%. Degradation therefore becomes increasingly occupancy-driven rather than indicating a violation of the per-identity influence bound. Complete results, including the summed observed normalization-branch diagnostic and clean test accuracy, appear in Appendix D.5.

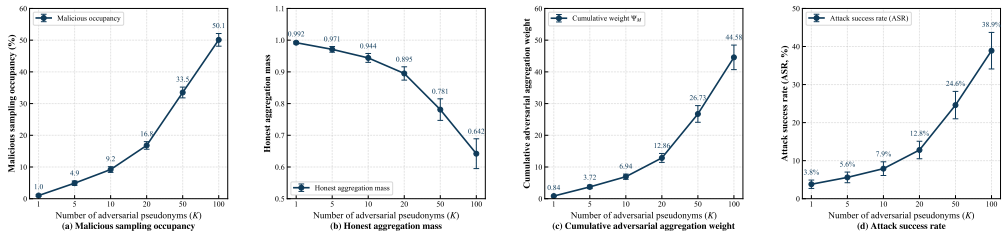


Figure 3: Identity proliferation with $N_H = 100$ and $n = 10$: (a) malicious sampling occupancy, (b) honest aggregation mass, (c) cumulative adversarial weight $\Psi_{\mathrm{M}}(H)$, and (d) ASR as K increases.

6.2.5 Additional evaluations. Further experiments cover benign participation under heterogeneity, Shakespeare conditioning, parameter and enrollment-privacy sensitivity, and complete baseline comparisons in Appendices D.6, D.7, D.8, D.9, and D.

7 Conclusion

In this work, we show that historical trust should not imply persistent aggregation authority when client behavior can change over time. Bounding post-pivot influence provides a direct way to limit what an adversary can retain or re-earn after such a change. The guarantees do not resolve unrestricted identity proliferation or arbitrary permanent benign distribution shifts, which remain separate challenges. More broadly, trustworthy FL requires controlling how long previously earned influence remains operationally effective, not only whether a participant was trusted in the past.

AI USE STATEMENT

In this work, we used generative AI tools for language editing and technical presentation. All research decisions, mathematical claims, proofs, experimental design, implementation, and reported results were reviewed and validated by the authors. Generative AI was not used to generate experimental measurements or to independently establish any theoretical result.

ETHICS STATEMENT

This work studies attacks on federated learning and proposes a defense. The gradual, threshold-aware, and layer-selective attacks could in principle inform an adversary, but they arise directly from the structure of existing online detectors and are included to make the resulting exposure measurable. We use only public benchmark datasets and no real-world personal records. The client-side record-level differential privacy mechanism provides a per-enrollment privacy guarantee for the released statistical summary; repeated releases from reused data compose and are not assigned a system-wide privacy budget. We characterize the resulting privacy–detection trade-off rather than claiming privacy for the overall system. False rejections may disproportionately reduce the aggregation influence of clients with minority data distributions, so we report client-balanced and tail false-rejection rates.

REPRODUCIBILITY STATEMENT

Anonymized source code is provided as supplementary material. Assumptions and proofs for all formal statements appear in Appendix B. Hyperparameters, including security-critical parameters, are consolidated in Appendix C.2. Attack constructions are specified in Appendix C.4, with detector and enrollment details in Appendices A.3 and A.2.

REFERENCES

Hasan Akgul, Daniel Borg, Arta Berisha, Amina Rahimova, Andrej Novak, and Mila Petrov. Verifiable fine-tuning for llms: Zero-knowledge training proofs bound to data provenance and policy. *arXiv preprint arXiv:2510.16830*, 2025.

Sana Awan, Bo Luo, and Fengjun Li. Contra: Defending against poisoning attacks in federated learning. In *European symposium on research in computer security*, pp. 455–475. Springer, 2021.

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 (NIPS)*, volume 30, 2017.

Weiliang Chen, Li Jia, Yang Zhou, and Qianqian Ren. Reputation-driven asynchronous federated learning for enhanced trajectory prediction with blockchain. *IEEE Internet of Things Journal*, 2024.

Haohua Duan, Zedong Peng, Liyao Xiang, Yuncong Hu, and Bo Li. A verifiable and privacy-preserving federated learning training framework. *IEEE Transactions on Dependable and Secure Computing*, 21(5):5046–5058, 2024.

Minghong Fang, Xiaoyu Cao, Jinyuan Jia, and Neil Zhenqiang Gong. Local model poisoning attacks to byzantine-robust federated learning. *USENIX Security Symposium*, pp. 1605–1622, 2020.

Clement Fung, Chris JM Yoon, and Ivan Beschastnikh. The limitations of federated learning in sybil settings. In *23rd International symposium on research in attacks, intrusions and defenses (RAID 2020)*, pp. 301–316, 2020.

Xiaolan Gu, Ming Li, and Li Xiong. {DP-BREM}:{Differentially-Private} and {Byzantine-Robust} federated learning with client momentum. In *34th USENIX Security Symposium (USENIX Security 25)*, pp. 3065–3082, 2025.

Zhipin Gu and Yuexiang Yang. Detecting malicious model updates from federated learning on conditional variational autoencoder. In *2021 IEEE international parallel and distributed processing symposium (IPDPS)*, pp. 671–680. IEEE, 2021.

Hanxi Guo, Hao Wang, Tao Song, Tianhang Zheng, Yang Hua, Haibing Guan, and Xiangyu Zhang. Poisoning with a pill: Circumventing detection in federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 21432–21440, 2026.

Wenwen He, Wenke Huang, Bin Yang, Shukan Liu, and Mang Ye. Spmc: Self-purifying federated backdoor defense via margin contribution. In *ICML*, 2025.

Yupeng Jiang, Yong Li, Yipeng Zhou, and Xi Zheng. Sybil attacks and defense on differential privacy based federated learning. In *2021 IEEE 20th international conference on trust, security and privacy in computing and communications (trustCom)*, pp. 355–362. IEEE, 2021.

Peter Kairouz and H Brendan McMahan. Advances and open problems in federated learning. *Foundations and trends in machine learning*, 14(1-2):1–210, 2021.

Jiawen Kang, Zehui Xiong, Dusit Niyato, Shengli Xie, and Junshan Zhang. Incentive mechanism for reliable federated learning: A joint optimization approach to combining reputation and contract theory. *IEEE Internet of Things Journal*, 6(6):10700–10714, 2019.

Jiawen Kang, Zehui Xiong, Dusit Niyato, Yuze Zou, Yang Zhang, and Mohsen Guizani. Reliable federated learning for mobile networks. *IEEE Wireless Communications*, 27(2):72–80, 2020.

Torsten Krauß, Jan König, Alexandra Dmitrienko, and Christian Kanzow. Automatic adversarial adaption for stealthy poisoning attacks in federated learning. In *To appear soon at the Network and Distributed System Security Symposium (NDSS)*, volume 2, 2024.

Cody Lewis, Vijay Varadharajan, and Nasimul Noman. Attacks against federated learning defense systems and their mitigation. *Journal of Machine Learning Research*, 24(30):1–50, 2023.

Xining Liu, Erwu Liu, Wei Ni, Hanfu Zhang, Zhongxiang Wei, and Ekram Hossain. Learning to evade: Adaptive and stealthy attack on reputation-based federated learning. *IEEE Communications Letters*, 2026.

Yizhong Liu, Zixiao Jia, Xiao Chen, Song Bian, Runhua Xu, Dawei Li, and Yuan Lu. Aion: Robust and efficient {Multi-Round}{Single-Mask} secure aggregation against malicious participants. In *34th USENIX Security Symposium (USENIX Security 25)*, pp. 3025–3044, 2025.

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

Lingjuan Lyu, Han Yu, and Qiang Yang. Reputation-based federated learning: A survey. *ACM Computing Surveys*, 56(1):1–36, 2023.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics (AISTATS)*, pp. 1273–1282. PMLR, 2017.

Thien Duc Nguyen, Phillip Rieger, Huili Chen, Hossein Yalame, Helen Möllering, Hossein Fereidooni, Samuel Marchal, Markus Miettinen, Azalia Mirhoseini, Shaza Zeitouni, et al. {FLAME}: Taming backdoors in federated learning. In *31st USENIX Security Symposium (USENIX Security 22)*, pp. 1415–1432, 2022.

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

Md Mamunur Rashid, Yong Xiang, Md Palash Uddin, Jine Tang, Keshav Sood, and Longxiang Gao. Trustworthy and fair federated learning via reputation-based consensus and adaptive incentives. *IEEE Transactions on Information Forensics and Security*, 2025.

Virat Shejwalkar and Amir Houmansadr. Manipulating the byzantine: Optimizing model poisoning attacks and defenses for federated learning. *IEEE Transactions on Information Forensics and Security*, 19:1204–1218, 2024.

Ahmed Soliman, Khalid M Amin, Noura A Semary, and Hayam Mousa. Shield-fl: Scalable backdoor defense in federated learning via gradient trust and data-free distillation under non-iid data. *Egyptian Informatics Journal*, 33:100904, 2026.

Nawrin Tabassum and Yanzhao Wu. Star-fl: Secure federated learning with spatial-temporal analysis and robust aggregation. *arXiv preprint arXiv:2608.14861*, 2026.

Muhammad Habib ur Rehman, Khaled Salah, Ernesto Damiani, and Davor Svetinovic. Towards blockchain-based reputation-aware federated learning. In *IEEE INFOCOM 2020-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pp. 183–188. IEEE, 2020.

Wei Wan, Shengshan Hu, Jianrong Lu, Leo Yu Zhang, Hai Jin, and Yuanyuan He. Shielding federated learning: Robust aggregation with adaptive client selection. *arXiv preprint arXiv:2204.13256*, 2022.

Yujing Wang, Hainan Zhang, Sijia Wen, Wangjie Qiu, and Binghui Guo. Defending against sophisticated poisoning attacks with rl-based aggregation in federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 25443–25451, 2025.

Yueqi Xie, Minghong Fang, and Neil Zhenqiang Gong. Fedredefense: Defending against model poisoning attacks for federated learning using model update reconstruction error. International Conference on Machine Learning, 2024.

Zaixi Zhang, Xiaoyu Cao, Jinyuan Jia, and Neil Zhenqiang Gong. Fldetector: Defending federated learning against model poisoning attacks via detecting malicious clients. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, pp. 2545–2555, 2022.

Xiaojie Zhao, Jinqiao Shi, Yi Li, Junmin Huang, and Chongru Fan. Good gradients poison your model: Evading defenses in federated learning via boundary-adaptive perturbation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 13262–13270, 2026.

APPENDIX

A ADDITIONAL DAIRE-FL METHOD DETAILS

A.1 ENFORCEMENT AND RESIDUAL TRUST BOUNDARIES

Table 4 summarizes the properties enforced by DAIRE-FL and the assumptions or limitations that remain outside its guarantees. Consistent with §3, enrollment verifies consistency with the encoded statistical relation rather than raw-data validity, identity scarcity, or data provenance, while attested renewal prevents any single coordinator or attester from restoring influence unilaterally. Honest attesters evaluate a versioned detector state committed before each round. Online state transitions use a fixed reference implementation, deterministic buffer ordering, fixed numerical precision, and committed model parameters; attesters verify against the resulting state identifier rather than independently defining a new state. Handling asynchronous or inconsistent detector-state replicas is outside the present protocol model.

A.2 ENROLLMENT CONSTRUCTION AND IMPLEMENTATION

Enrollment provides the conditioning reference S_i used by the behavioral detector; it does not establish identity uniqueness. Each client commits to a privatized, task-dependent statistical summary and proves in zero knowledge that a witness satisfies the encoded enrollment relation, including the enrollment-size requirement $n_i^{\mathrm{data}} \geq n_{\mathrm{data}}^{\min}$. The contract commitment is $C_i = H_{\mathrm{com}}(pk_i \| S_i)$, where pk_i is client i's public key and H_{com} is Poseidon over the fixed-point field encoding used by the enrollment circuit. Define the enrollment-size statistic

$$n_i^{\mathrm{data}} = \begin{cases} |\mathcal{D}_i|, & \text{vision,} \\ N_i^{(1)}, & \text{Shakespeare,} \end{cases}$$

so that the enrollment relation enforces $n_i^{\mathrm{data}} \geq n_{\mathrm{data}}^{\min}$.

Table 4: Enforcement and residual trust boundaries in DAIRE-FL.

Property	Enforced by	Residual assumption / limitation
Influence decay	Committed state and deterministic decay rule	Correct execution of the committed state transition
State-history integrity	Append-only ledger state	Underlying ledger integrity assumption
Renewal authorization	Independent q-of-m acceptance attestations	Safety requires $f < q$; liveness requires $q \leq m - f$
Behavioral acceptance	Committed detector rule independently recomputed by attesters	Detector accuracy is empirical; benign intent is not certified
Client sampling	Coordinator sampling policy	Coordinator censorship or biased selection is out of scope
Enrollment relation	Zero-knowledge proof of the encoded statistical relation	Does not establish raw-data validity, uniqueness, provenance, ownership, or label correctness
Identity scarcity	—	Not provided; the threat model permits unrestricted identity proliferation
Data provenance	—	Out of scope

Statistical summaries. For each vision image and channel, we compute a 64-bin intensity histogram normalized to unit ℓ_1 mass. The unprivatized statistic h_i^{vis} is the average of these per-image histograms over the local dataset. CIFAR-10 and CIFAR-100 therefore yield three 64-dimensional blocks (192 coordinates), while FEMNIST yields one 64-dimensional block. The histogram construction, bin boundaries, and privacy parameters are fixed before training.

For next-character prediction, the unprivatized statistic combines character frequencies with hash-binned transition statistics. Let

$$\mathcal{D}_i^{\mathrm{text}} = \{\mathbf{c}_{i,j}\}_{j=1}^{n_i^{\mathrm{seq}}}, \qquad \mathbf{c}_{i,j} = (c_{i,j,1}, \ldots, c_{i,j,L_{i,j}}),$$

where $c_{i,j,r} \in \Sigma_{\mathrm{char}}$ and Σ_{char} is the fixed public character vocabulary. With $N_i^{(1)} = \sum_{j=1}^{n_i^{\mathrm{seq}}} L_{i,j}$, the normalized unigram statistic is

$$h_i^{(1)}(c) = \frac{1}{N_i^{(1)}} \sum_{j=1}^{n_i^{\mathrm{seq}}} \sum_{r=1}^{L_{i,j}} \mathbf{1}\{c_{i,j,r} = c\}, \qquad c \in \Sigma_{\mathrm{char}}. \tag{19}$$

To capture local transition structure without releasing the full $|\Sigma_{\mathrm{char}}|^2$ table, let

$$g : \Sigma_{\mathrm{char}} \times \Sigma_{\mathrm{char}} \to \{1, \ldots, n_{\mathrm{bin}}\} \tag{20}$$

be a fixed public hash-binning map. Define $N_i^{(2)} = \sum_{j=1}^{n_i^{\mathrm{seq}}} \max\{L_{i,j} - 1, 0\}$. For $N_i^{(2)} > 0$,

$$h_{i,b}^{(2)} = \frac{1}{N_i^{(2)}} \sum_{j=1}^{n_i^{\mathrm{seq}}} \sum_{r=1}^{L_{i,j}-1} \mathbf{1}\{g(c_{i,j,r}, c_{i,j,r+1}) = b\}, \qquad b = 1, \ldots, n_{\mathrm{bin}}. \tag{21}$$

If $N_i^{(2)} = 0$, we set $h_i^{(2)}$ to the zero vector. The unprivatized text statistic is

$$h_i^{\mathrm{text}} = h_i^{(1)} \| h_i^{(2)}. \tag{22}$$

Hash binning reduces the transition representation from $|\Sigma_{\mathrm{char}}|^2$ coordinates to n_{bin}.

Client-side record-level privacy. For both vision and text, the released conditioning summary is

$$S_i = \mathcal{M}_\epsilon(h_i), \tag{23}$$

where $h_i = h_i^{\mathrm{vis}}$ for vision and $h_i = h_i^{\mathrm{text}}$ for Shakespeare. We instantiate $\mathcal{M}_\epsilon$ with the vector Laplace mechanism. Under replacement adjacency, two local datasets differ in one image for vision or one character position for Shakespeare. The ℓ_1 sensitivity is

$$\Delta_1 = \begin{cases} 6/n_i^{\mathrm{data}}, & \text{CIFAR-10/100,} \\ 2/n_i^{\mathrm{data}}, & \text{FEMNIST,} \\ 2/N_i^{(1)} + 4/N_i^{(2)}, & \text{Shakespeare,} \end{cases} \tag{24}$$

where replacing one vision sample can alter two normalized histogram coordinates per channel, while replacing one character can alter two unigram coordinates and at most four transition coordinates. For $N_i^{(2)} = 0$, only the unigram term is used. We release

$$\widetilde{S}_i = h_i + \varepsilon_i^{\mathrm{Lap}}, \qquad \varepsilon_{i,j}^{\mathrm{Lap}} \overset{\text{i.i.d.}}{\sim} \mathrm{Lap}\left(0, \frac{\Delta_1}{\epsilon}\right), \qquad S_i = \mathrm{Norm}_+(\widetilde{S}_i), \tag{25}$$

where Norm_+ clips negative coordinates to zero and renormalizes each statistical block to unit mass. If clipping produces zero mass in a block, the operator returns a fixed public uniform distribution for that block. This operation is deterministic post-processing. Consequently, S_i satisfies pure ϵ-differential privacy under the stated replacement adjacency. This is a per-enrollment release guarantee and assumes that the client samples $\varepsilon_i^{\mathrm{Lap}}$ from the specified Laplace mechanism. The zero-knowledge proof verifies arithmetic consistency with the supplied noise but does not enforce its sampling distribution. Moreover, repeated releases derived from reused data compose across pseudonymous enrollments; because DAIRE-FL does not enforce dataset or identity uniqueness, it does not provide a fixed cumulative privacy budget across such releases.

Proof relation and implementation. Raw image pixels and text symbols are not processed directly inside the enrollment circuit. The private witness contains the declared dataset size, precomputed per-record binned statistics, the corresponding aggregate count vectors, the fixed-point privatization noise, and the commitment opening. For vision, each record contributes its per-channel 64-bin histogram; for Shakespeare, each sequence contributes unigram and hash-binned transition counts.

The circuit verifies that (i) the declared dataset size satisfies $n_i^{\mathrm{data}} \geq n_{\mathrm{data}}^{\min}$, (ii) all count vectors are nonnegative and have the required per-record totals, (iii) the aggregate statistics equal the sum of the per-record statistics, (iv) the released summary S_i is arithmetically consistent with the aggregate statistics and the supplied privatization noise under the fixed-point implementation of Eq. 25, and (v) the resulting summary opens the commitment $C_i = H_{\mathrm{com}}(pk_i \| S_i)$.

The proof establishes arithmetic consistency of these committed statistics; it does not establish that the precomputed per-record statistics faithfully represent raw local samples, nor does it certify provenance, ownership, label correctness, semantic correctness, or identity uniqueness. For the randomized construction in Eq. 25, the proof establishes arithmetic consistency with the privatization noise supplied as part of the witness; it does not certify that the private random coins were sampled from the stated Laplace distribution.

Enrollment circuits are compiled with `circom`, and CPU proof generation and verification use `snarkjs` with Groth16. The three-channel vision circuit contains approximately 420k R1CS constraints, while the Shakespeare configuration with $n_{\mathrm{bin}} = 128$ contains approximately 0.9M. Proof-size dependence is included in the communication complexity, while proving latency, vision-circuit verification gas, and attestation overhead are reported in Appendix E.

Enrollment soundness. Assume that the Groth16 common reference string is generated correctly and that Groth16 is computationally sound. Suppose an adversary makes at most $Q_{\mathrm{att}}(\kappa_{\mathrm{sec}})$ enrollment attempts, where Q_{att} is polynomial in the cryptographic security parameter κ_{sec}. If N_{inv} denotes accepted attempts whose witnesses violate the encoded enrollment relation, then

$$\Pr[N_{\mathrm{inv}} \geq 1] \leq Q_{\mathrm{att}}(\kappa_{\mathrm{sec}}) \cdot \mathrm{negl}(\kappa_{\mathrm{sec}}), \qquad \mathbb{E}[N_{\mathrm{inv}}] \leq Q_{\mathrm{att}}(\kappa_{\mathrm{sec}}) \cdot \mathrm{negl}(\kappa_{\mathrm{sec}}). \tag{26}$$

Indeed, if X_j indicates acceptance of the jth invalid attempt, computational soundness gives $\Pr[X_j = 1] \leq \mathrm{negl}(\kappa_{\mathrm{sec}})$; the result follows by a union bound and linearity of expectation. Equation 26 concerns only the relation encoded by the enrollment circuit and provides no guarantee of identity uniqueness or provenance.

Identity-scarcity boundary. The privatized summary conditions the shared C-VAE used to score x_i^t, but a valid summary need not be unique or sufficiently informative for behavioral discrimination. Two clients may therefore have identical or statistically indistinguishable valid summaries, and under capability (P2) the same qualifying data may support multiple valid enrollments. Each additional identity incurs proof generation and verification, but this marginal cost is not treated as a Sybil-prevention mechanism. Detector informativeness and the effect of identity proliferation are therefore evaluated empirically.

A.3 CONDITIONAL BEHAVIORAL MODEL AND ADAPTATION DETAILS

The frozen and adaptive references in §4.2 are shared global conditional variational autoencoders (C-VAEs) conditioned on S_i, rather than per-client models. Sharing the references avoids data starvation under intermittent participation, while conditioning on S_i captures coarse client-specific characteristics. For parameters θ, the encoder and decoder are

$$q_\theta(z \mid x, S) = \mathcal{N}\big(\mu_\theta^q(x, S), \mathrm{diag}((\sigma_\theta^q(x, S))^2)\big)$$

and

$$p_\theta(x \mid z, S) = \mathcal{N}\big(\mu_\theta^p(z, S), \sigma_x^2 I\big)\,.$$

Here $\sigma_x^2 > 0$ is fixed; its constant scale is absorbed into the reconstruction coefficient in Eq. 27 and it is not estimated by the detector. The deterministic reconstruction used for scoring is

$$\widehat{x}_\theta(x, S) \triangleq \mu_\theta^p(\mu_\theta^q(x, S), S)\,.$$

The posterior-mean reconstruction above is used only for deterministic behavioral scoring. During training, we use the standard reparameterization

$$z = \mu_\theta^q(x, S) + \sigma_\theta^q(x, S) \odot \epsilon_z, \qquad \epsilon_z \sim \mathcal{N}(0, I),$$

with one Monte Carlo sample per training example. The training objective is

$$\ell_{\mathrm{CVAE}}(\theta; x, S) = \frac{1}{d'}\mathbb{E}_{z\sim q_\theta(z|x,S)}\Big[\|x - \mu_\theta^p(z, S)\|_2^2\Big] + \beta_{\mathrm{KL}}\,\mathrm{KL}(q_\theta(z \mid x, S)\|\mathcal{N}(0, I))\,. \tag{27}$$

All raw and projected calibration collections and all replay buffers below are identity-indexed multisets. Thus an entry displayed as (x, S_i) retains the submitting identity i as provenance metadata, and repeated entries are not collapsed even when their displayed values coincide. For each identity represented in the calibration collection, define $\mathcal{C}_i \triangleq \{(x, S_i) \in \mathcal{C} : \text{the pair belongs to identity } i\}$. To avoid overweighting frequently participating identities, the empirical loss over calibration collection $\mathcal{C}$ is client-balanced:

$$\widehat{\mathcal{L}}_{\mathcal{C}}(\theta) = \frac{1}{|\mathcal{I}_{\mathcal{C}}|}\sum_{i\in\mathcal{I}_{\mathcal{C}}}\frac{1}{|\mathcal{C}_i|}\sum_{(x,S_i)\in\mathcal{C}_i}\ell_{\mathrm{CVAE}}(\theta; x, S_i), \tag{28}$$

where $\mathcal{I}_{\mathcal{C}}$ indexes the identities represented in $\mathcal{C}$. Let $T_{\mathrm{fit}} = \lfloor \chi T_{\mathrm{cal}} \rfloor$ and first define the raw fit collection

$$\mathcal{U}_{\mathrm{fit}} = \{(\Delta w_i(t), S_i) : 0 \le t < T_{\mathrm{fit}},\ i \in \mathcal{S}_t\}.$$

For each monitored layer, let μ_ℓ^{PCA} be the mean of the raw layer-ℓ vectors in $\mathcal{U}_{\mathrm{fit}}$. The matrix P_ℓ contains the leading d'_ℓ centered principal components fitted only on that raw fit collection. We use $(d'_{\ell_1}, d'_{\ell_2}, d'_{\ell_3}) = (342, 341, 341)$, so that $\sum_{\ell\in\mathcal{L}_{\mathrm{mon}}} d'_\ell = 1024$. PCA is computed using randomized SVD with the experimental seed, and both μ_ℓ^{PCA} and P_ℓ are frozen before any projected calibration collection is constructed. With Π fixed, define

$$\begin{aligned}\mathcal{C}_{\mathrm{fit}} &= \{(\Pi(\Delta w_i(t)), S_i) : 0 \le t < T_{\mathrm{fit}},\ i \in \mathcal{S}_t\},\\ \mathcal{C}_{\mathrm{ref}} &= \{(\Pi(\Delta w_i(t)), S_i) : T_{\mathrm{fit}} \le t < T_{\mathrm{cal}},\ i \in \mathcal{S}_t\}.\end{aligned} \tag{29}$$

Let $E_\theta(x, S) = d'^{-1}\|x - \widehat{x}_\theta(x, S)\|_2^2$ and $\mathrm{SKL}(q_1, q_2) = \frac{1}{2}[\mathrm{KL}(q_1\|q_2) + \mathrm{KL}(q_2\|q_1)]$. The pointwise reference discrepancy is

$$d_{\mathrm{ref}}(x, S; \theta^S, \theta^L) = \frac{1}{d'}\,\|\widehat{x}_{\theta^S}(x, S) - \widehat{x}_{\theta^L}(x, S)\|_2^2 + \kappa\ \mathrm{SKL}(q_{\theta^S}(z \mid x, S), q_{\theta^L}(z \mid x, S))\,. \tag{30}$$

Let $\mathcal{I}_{\mathrm{ref}}$ index the identities represented in $\mathcal{C}_{\mathrm{ref}}$. For $i \in \mathcal{I}_{\mathrm{ref}}$, let $\mathcal{C}_{\mathrm{ref},i} \triangleq \{(x, S_i) \in \mathcal{C}_{\mathrm{ref}} : \text{the pair belongs to identity } i\}$. The client-balanced functional displacement is

$$D_{\mathcal{C}}(\theta, \theta^L) = \frac{1}{|\mathcal{I}_{\mathrm{ref}}|}\sum_{i\in\mathcal{I}_{\mathrm{ref}}}\frac{1}{|\mathcal{C}_{\mathrm{ref},i}|}\sum_{(x,S_i)\in\mathcal{C}_{\mathrm{ref},i}} d_{\mathrm{ref}}(x, S_i; \theta, \theta^L). \tag{31}$$

The frozen threshold is

$$\tau_L = Q_\nu(\{E_{\theta^L}(x, S) : (x, S) \in \mathcal{C}_{\mathrm{ref}}\})\,. \tag{32}$$

Using the split in Eq. 29, the frozen reference θ^L is trained on $\mathcal{C}_{\text{fit}}$ to minimize $\widehat{\mathcal{L}}_{\mathcal{C}_{\text{fit}}}$ and is then fixed. We optimize θ^L with Adam using learning rate $\eta_L = 10^{-3}$, batch size 64, weight decay 10^{-5}, at most 100 epochs, and early-stopping patience of 10 epochs. At the end of calibration, the adaptive reference and threshold are initialized as

$$\theta^S_{T_{\text{cal}}} \leftarrow \theta^L, \qquad \tau_{T_{\text{cal}}} \leftarrow \tau_L. \tag{33}$$

The held-out set $\mathcal{C}_{\text{ref}}$ is never used to fit θ^L; it defines the functional displacement constraint in Eq. 7 and the frozen threshold in Eq. 32. When $\mathcal{B}^+_t$ is nonempty, the unconstrained candidate is

$$\widetilde{\theta}^S_{t+1} = \theta^S_t - \eta_S \nabla_\theta \widehat{\mathcal{L}}_{\mathcal{B}^+_t}(\theta)\big|_{\theta=\theta^S_t}. \tag{34}$$

The backtracking factor is

$$s_t = \max\left(\{0\} \cup \left\{2^{-j} : j \in \mathbb{N}_0,\ D_{\mathcal{C}}\Big(\theta^S_t + 2^{-j}(\widetilde{\theta}^S_{t+1} - \theta^S_t), \theta^L\Big) \le \Delta_M\right\}\right). \tag{35}$$

Because Eq. 7 holds for θ^S_t, choosing $s_t = 0$ is always feasible. If $\mathcal{B}^+_t = \emptyset$, then $\theta^S_{t+1} = \theta^S_t$. The next-round threshold is computed directly from the post-insertion window $\mathcal{W}^+_t$ by Eq. 9; hence $\tau_{t+1} \le \tau_t$ and $\tau_{t+1} \ge \tau_L - \Delta_\tau$. Only accepted updates with pre-renewal influence $B_i(t) \ge \tau_{\min}$ are used to form the post-insertion states $\mathcal{B}^+_t$ and $\mathcal{W}^+_t$. An accepted probationary update with $B_i(t) < \tau_{\min}$ may therefore replenish future influence through Eq. 2, but cannot affect current-round aggregation or modify either adaptive detector state in the same round.

Controlled recalibration. Because θ^L and τ_L remain fixed during online adaptation, sustained benign distribution shift requires explicit protocol-level recalibration. Reaching the adaptation limits Δ_M or Δ_τ does not trigger recalibration, and accepted updates alone cannot replace either anchor.

A.4 DAIRE-FL ONLINE PROCEDURE

Algorithm 1 summarizes the online phase after calibration. It preserves the protocol ordering that is security-critical to DAIRE-FL: sampling is independent of influence state, aggregation uses the pre-renewal state $B_i(t)$, accepted participation renews only $B_i(t+1)$, and only accepted influence-active updates may adapt the behavioral reference or threshold. The detailed detector, backtracking, and threshold-update rules are given in Eqs. 6, 35, 8, and 9.

Algorithm 1 DAIRE-FL online procedure

Require: $w_{T_{\text{cal}}}$, Π, θ^L, $\theta^S_{T_{\text{cal}}} = \theta^L$, $\tau_{T_{\text{cal}}} = \tau_L$, $\{a_i\}$, $\{B_i(T_{\text{cal}}) = B_0 : i \in \mathcal{V}_{T_{\text{cal}}}\}$, $\mathcal{B}_{T_{\text{cal}}} = \emptyset$, $\mathcal{W}_{T_{\text{cal}}} = \emptyset$, $\tau_{\min}$, λ, V_a, Δ_M, Δ_τ, quorum (q, m), sample size n

for $t = T_{\text{cal}}, \dots, T-1$ **do**
  For each i with $a_i = t$ not already active, initialize $B_i(t) \leftarrow B_0$
  Sample $\mathcal{S}_t \subseteq \mathcal{V}_t$ with $|\mathcal{S}_t| = \min\{n, |\mathcal{V}_t|\}$ independently of influence state
  for each $i \in \mathcal{S}_t$ **in parallel do**
    Reconstruct pre-renewal $B_i(t)$; receive $\Delta w_i(t)$ and compute $x^t_i \leftarrow \Pi(\Delta w_i(t))$
    Compute r^t_i by Eq. 6; set $A_i(t) = 1$ iff $r^t_i \le \tau_t$ and at least q valid attestations bound to V^t_i are received
    Set $R_i(t) \leftarrow J_i(t)A_i(t)$
  end for
  Compute $\widetilde{p}_i(t)$ and $p_i(t)$ using Eqs. 3–4
  if $Z_t > 0$ **then**
    $w_{t+1} \leftarrow w_t - \eta_g \sum_{i \in \mathcal{S}_t} p_i(t)\Delta w_i(t)$
  else
    $w_{t+1} \leftarrow w_t$
  end if
  For every $i \in \mathcal{V}_t$, set $R_i(t) = 0$ if $i \notin \mathcal{S}_t$ and update $B_i(t+1) \leftarrow e^{-\lambda}B_i(t) + V_a R_i(t)$; commit renewal-bearing transitions and reconstruct all other decay lazily
  $\mathcal{I}^{\text{adapt}}_t \leftarrow \{i \in \mathcal{S}_t : A_i(t) = 1,\ B_i(t) \ge \tau_{\min}\}$
  $\mathcal{B}^+_t \leftarrow \text{UpdateBuffer}(\mathcal{B}_t, \{(x^t_i, S_i) : i \in \mathcal{I}^{\text{adapt}}_t\})$
  $\mathcal{W}^+_t \leftarrow \text{UpdateWindow}(\mathcal{W}_t, \{r^t_i : i \in \mathcal{I}^{\text{adapt}}_t\})$
  Update θ^S_{t+1} using Eqs. 34, 35, and 8; update τ_{t+1} using Eq. 9
  $\mathcal{B}_{t+1} \leftarrow \mathcal{B}^+_t$ and $\mathcal{W}_{t+1} \leftarrow \mathcal{W}^+_t$
end for

B Additional Theoretical Analysis and Proofs

This section proves the main influence bounds and gives supplementary consequences of DAIRE-FL's decay-and-renewal dynamics. Throughout, $\tau_{\min} > 0$, and $p_k(t) = 0$ whenever identity k is not sampled.

B.1 Expected contraction

Let $\omega_\lambda = e^{-\lambda} \in (0, 1)$. Since $B_k(t)$ is determined by the pre-round protocol state and is therefore $\mathcal{G}_t$-measurable, Eqs. 2 and 11 give

$$\mathbb{E}[B_k(t+1) \mid \mathcal{G}_t] \le \omega_\lambda B_k(t) + \beta_R V_a.$$

Taking total expectation and iterating from $B_k(0) = b_k$ yields

$$\mathbb{E}[B_k(t)] \le \bar{B} + \omega_\lambda^t (b_k - \bar{B}), \qquad \bar{B} = \frac{\beta_R V_a}{1 - \omega_\lambda},$$

which is Eq. 12. The condition $\bar{B} < \tau_{\min}$ is equivalent to Eq. 13. Since $\beta_R = \pi_{\max}\beta_A$ by Eq. 11, this is equivalently

$$\pi_{\max}\beta_A < \frac{\tau_{\min}(1 - e^{-\lambda})}{V_a}.$$

B.2 Finite-horizon persistent-influence bound

Proof of Theorem 1. Let $\omega_\lambda = e^{-\lambda}$. Unrolling Eq. 2 from $B_k(0) = b_k$ gives

$$B_k(t) = \omega_\lambda^t b_k + V_a \sum_{s=0}^{t-1} \omega_\lambda^{t-1-s} R_k(s). \tag{36}$$

Summing over $t = 0, \dots, H-1$ and exchanging the finite sums gives

$$\sum_{t=0}^{H-1} B_k(t) = \frac{b_k(1 - \omega_\lambda^H)}{1 - \omega_\lambda} + \frac{V_a}{1 - \omega_\lambda} \sum_{s=0}^{H-2} \big(1 - \omega_\lambda^{H-1-s}\big) R_k(s), \tag{37}$$

where the second sum is interpreted as zero when $H = 1$. Since $0 < \omega_\lambda < 1$ and $R_k(s) \ge 0$,

$$\sum_{t=0}^{H-1} B_k(t) \le \frac{b_k + V_a \mathcal{R}_k(H)}{1 - \omega_\lambda}. \tag{38}$$

Including $R_k(H-1)$ only loosens the bound because that renewal affects $B_k(H)$ rather than any state in the sum on the left. On any round with $Z_t > 0$, Eqs. 3–4 give

$$p_k(t) = \frac{\phi(B_k(t)) \mathbf{1}\{B_k(t) \ge \tau_{\min}\} R_k(t)}{Z_t}.$$

Since $B_k(t) \ge 0$, $\phi(B_k(t)) = \log(1 + B_k(t)) \le B_k(t)$. Hence, using $Z_t \ge Z_{\min}$,

$$p_k(t) \le \frac{B_k(t)}{Z_{\min}}.$$

The same inequality holds when $Z_t = 0$ because $p_k(t) = 0$.

A second bound follows directly from normalized aggregation. Since $p_k(t) \le 1$ and positive aggregation weight requires $B_k(t) \ge \tau_{\min}$,

$$p_k(t) \le \mathbf{1}\{B_k(t) \ge \tau_{\min}\} \le \frac{B_k(t)}{\tau_{\min}}.$$

Finally, $\sum_{t=0}^{H-1} p_k(t) \le H$. Combining these three inequalities with Eq. 38 yields Eq. 14 and hence Eq. 15. Under the additional condition $\|\Delta w_k(t)\| \le \delta_{\max}$,

$$\Omega_k(H) \le \delta_{\max} \Psi_k(H) \le \delta_{\max} U_k^{\Psi}(H),$$

which proves Eq. 16. □

B.3 Expected influence and empirical specialization

Proof of Corollary 1. By the tower property and Eq. 11,

$$\mathbb{E}[R_k(t)] = \mathbb{E}[\mathbb{E}[R_k(t) \mid \mathcal{G}_t]] \le \beta_R.$$

Therefore, $\mathbb{E}[\mathcal{R}_k(H)] \le \beta_R H$. Taking expectations in the $Z_{\min}$-normalized branch of Eq. 14 and substituting this bound gives the first inequality in Eq. 17. Under bounded updates, $\Omega_k(H) \le \delta_{\max}\Psi_k(H)$ gives the second. □

Empirical specialization. The identity-level realized rates are defined in Eq. 62. Substituting $\mathcal{R}_k(H) = \widehat{\beta}_{A,k}N_k^{\mathrm{sub}}(H) = \widehat{\beta}_{R,k}H$ into the $Z_{\min}$-normalized branch of Eq. 14 gives Eq. 18. This algebraic specialization does not interpret either realized rate as a uniform transcript-conditioned probability bound.

B.4 Influence-active occupation under deployed re-entry

The deployed protocol permits probationary re-entry after $B_k(t)$ falls below $\tau_{\min}$, so geometric extinction need not hold. The pathwise argument nevertheless bounds the number of influence-active rounds. Define

$$N_k^{\mathrm{act}}(H) = \sum_{t=0}^{H-1} \mathbf{1}\{B_k(t) \ge \tau_{\min}\}.$$

Since $\tau_{\min}N_k^{\mathrm{act}}(H) \le \sum_{t=0}^{H-1} B_k(t)$, Eq. 38 gives

$$N_k^{\mathrm{act}}(H) \le \frac{b_k + V_a\mathcal{R}_k(H)}{\tau_{\min}(1-e^{-\lambda})}. \tag{39}$$

Proposition 1 (Average influence-active occupation under re-entry)**.** *Suppose Eq. 2 holds with $B_k(0) = b_k$ and $\mathbb{E}[R_k(t) \mid \mathcal{G}_t] \le \beta_R$ for every post-pivot round. Then, for every $H \ge 1$,*

$$\frac{1}{H}\sum_{t=0}^{H-1} \Pr[B_k(t) \ge \tau_{\min}] \le \frac{\beta_R V_a}{(1-e^{-\lambda})\tau_{\min}} + \frac{b_k}{(1-e^{-\lambda})\tau_{\min}H}. \tag{40}$$

Proof. Let $\omega_\lambda = e^{-\lambda}$. Rearranging Eq. 2,

$$(1-\omega_\lambda)B_k(t) = B_k(t) - B_k(t+1) + V_aR_k(t).$$

Taking expectations and summing gives

$$(1-\omega_\lambda)\sum_{t=0}^{H-1}\mathbb{E}[B_k(t)] = b_k - \mathbb{E}[B_k(H)] + V_a\sum_{t=0}^{H-1}\mathbb{E}[R_k(t)] \le b_k + \beta_R V_a H.$$

Moreover,

$$B_k(t) \ge \tau_{\min}\mathbf{1}\{B_k(t) \ge \tau_{\min}\},$$

and therefore

$$\tau_{\min}\sum_{t=0}^{H-1}\Pr[B_k(t) \ge \tau_{\min}] \le \sum_{t=0}^{H-1}\mathbb{E}[B_k(t)].$$

Dividing by $H(1-\omega_\lambda)\tau_{\min}$ proves Eq. 40. □

B.5 Absorbing-gating analytical variant

For comparison, consider a stricter analytical variant in which an identity receives no further sampling opportunity or renewal after first falling below $\tau_{\min}$. This absorbing model is not the deployed probationary re-entry protocol.

Proposition 2 (Geometric influence-active tail under absorbing gating)**.** *Let $B_k(0) = b_k$. Suppose Eq. 2 holds, $\mathbb{E}[R_k(t) \mid \mathcal{G}_t] \le \beta_R$ whenever $B_k(t) \ge \tau_{\min}$, and Eq. 13 holds. Under absorbing influence gating,*

$$\Pr[B_k(t) \ge \tau_{\min}] \le \Gamma_k\xi_k^t, \qquad \xi_k = e^{-\lambda} + \frac{\beta_R V_a}{\tau_{\min}} < 1, \qquad \Gamma_k = \max\left\{1, \frac{b_k}{\tau_{\min}}\right\}. \tag{41}$$

Proof. Let

$$T_k = \inf\{t \geq 0 : B_k(t) < \tau_{\min}\}, \qquad Y_k(t) = B_k(t)\mathbf{1}\{T_k > t\}.$$

Under absorption, renewal stops after T_k, and subsequent decay keeps the state below $\tau_{\min}$. Hence $\{B_k(t) \geq \tau_{\min}\} = \{T_k > t\}$. Let $\omega_\lambda = e^{-\lambda}$. On $\{T_k > t\}$,

$$\mathbb{E}[B_k(t+1) \mid \mathcal{G}_t] \leq \omega_\lambda B_k(t) + \beta_R V_a \leq \left(\omega_\lambda + \frac{\beta_R V_a}{\tau_{\min}}\right) B_k(t) = \xi_k B_k(t).$$

After threshold crossing, $Y_k(t+1) = 0$. Hence

$$\mathbb{E}[Y_k(t+1) \mid \mathcal{G}_t] \leq \mathbf{1}\{T_k > t\}\,(\omega_\lambda B_k(t) + \beta_R V_a) \leq \xi_k Y_k(t).$$

Iterating gives $\mathbb{E}[Y_k(t)] \leq b_k \xi_k^t$. Since $Y_k(t) \geq \tau_{\min}\mathbf{1}\{T_k > t\}$,

$$\Pr[B_k(t) \geq \tau_{\min}] \leq \frac{b_k}{\tau_{\min}} \xi_k^t \leq \Gamma_k \xi_k^t.$$

Equation 13 gives $\beta_R V_a < (1 - e^{-\lambda})\tau_{\min}$, and therefore $\xi_k < 1$. □

Define

$$B_k^{\max} \triangleq B^{\max}(b_k). \tag{42}$$

Since $R_k(t) \leq 1$, Eq. 2 implies $B_k(t) \leq B_k^{\max}$. Proposition 2 therefore gives

$$\mathbb{E}[N_k^{\text{act}}(H)] \leq \Gamma_k \frac{1 - \xi_k^H}{1 - \xi_k}. \tag{43}$$

Moreover,

$$p_k(t) \leq \frac{\phi(B_k^{\max})\mathbf{1}\{B_k(t) \geq \tau_{\min}\}}{Z_{\min}},$$

so

$$\mathbb{E}[\Psi_k(H)] \leq \frac{\phi(B_k^{\max})\Gamma_k}{Z_{\min}} \frac{1 - \xi_k^H}{1 - \xi_k}, \qquad \mathbb{E}[\Omega_k(H)] \leq \frac{\delta_{\max}\phi(B_k^{\max})\Gamma_k}{Z_{\min}} \frac{1 - \xi_k^H}{1 - \xi_k}. \tag{44}$$

In particular,

$$\mathbb{E}[\Psi_k(\infty)] \leq \frac{\phi(B_k^{\max})\Gamma_k}{Z_{\min}(1 - \xi_k)}. \tag{45}$$

B.6 Optimization consequence under adversarial aggregation mass

The influence bounds are model-agnostic. For completeness, we show how malicious aggregation mass enters a standard strongly convex analysis. This is an analytical consequence of influence control, not a convergence guarantee for the non-convex models used in our experiments.

Theorem 2 (Finite-horizon optimization bound under adversarial mass)**.** *Let F be μ-strongly convex and L_s-smooth with minimizer $w^\star$. Let $\mathcal{K}_{\mathrm{M}}$ be a fixed finite malicious cohort present at the post-pivot analysis origin. Let $\mathcal{S}_t^{\mathrm{H}}$ and $\mathcal{S}_t^{\mathrm{M}} \subseteq \mathcal{K}_{\mathrm{M}}$ denote the honest and malicious sampled identities in post-pivot round t, respectively, and define*

$$u_t = \sum_{i \in \mathcal{S}_t} p_i(t)\Delta w_i(t) = u_t^{(\mathrm{H})} + u_t^{(\mathrm{M})},$$

where

$$u_t^{(\mathrm{H})} = \sum_{i \in \mathcal{S}_t^{\mathrm{H}}} p_i(t)\Delta w_i(t), \qquad u_t^{(\mathrm{M})} = \sum_{k \in \mathcal{S}_t^{\mathrm{M}}} p_k(t)\Delta w_k(t),$$

and

$$w_{t+1} = w_t - \eta_g u_t.$$

Suppose, conditioned on $\mathcal{G}_t$,

$$\mathbb{E}[u_t^{(\mathrm{H})} \mid \mathcal{G}_t] = \rho_t \nabla F(w_t), \qquad \rho_t \geq \rho > 0, \tag{46}$$

and

$$\mathbb{E}\Big[\|u_t^{(\mathrm{H})}\|^2 \mid \mathcal{G}_t\Big] \leq \|\nabla F(w_t)\|^2 + \sigma_{\mathrm{H}}^2. \tag{47}$$

Equation 46 is an idealized condition on the post-filtering honest aggregate. Because behavioral acceptance and $p_i(t)$ depend on the submitted update, deployed filtering may introduce selection bias; Theorem 2 is conditional on this assumption. The fixed-cohort assumption excludes identities that enroll after the post-pivot analysis origin. Let

$$m_t = \sum_{k \in \mathcal{S}_t^{\mathrm{M}}} p_k(t)$$

denote malicious aggregation mass, and assume $\|\Delta w_k(t)\| \le \delta_{\max}$. If

$$0 < \eta_g \le \min\left\{ \frac{\mu\rho}{4L_s^2}, \frac{1}{\mu\rho} \right\}, \tag{48}$$

then, with $c_g = 1 - \eta_g \mu\rho \in [0, 1)$,

$$\begin{aligned} \mathbb{E}\big[\|w_H - w^\star\|^2\big] \le c_g^H \|w_0 - w^\star\|^2 + \frac{2\eta_g \sigma_{\mathrm{H}}^2}{\mu\rho} \\ + \frac{4\eta_g \delta_{\max}^2}{\mu\rho} \sum_{t=0}^{H-1} c_g^{H-1-t} \mathbb{E}[m_t^2]. \end{aligned} \tag{49}$$

Then

$$\begin{aligned} \sum_{t=0}^{H-1} c_g^{H-1-t} \mathbb{E}[m_t^2] &\le \mathbb{E}\left[\sum_{k \in \mathcal{K}_{\mathrm{M}}} \Psi_k(H) \right] \\ &\le \frac{1}{Z_{\min}(1 - e^{-\lambda})} \, \mathbb{E}\left[\sum_{k \in \mathcal{K}_{\mathrm{M}}} \left(B_k(0) + V_a \mathcal{R}_k(H) \right) \right]. \end{aligned} \tag{50}$$

Proof. Let $e_t = w_t - w^\star$. Expanding the update gives

$$\|e_{t+1}\|^2 = \|e_t\|^2 - 2\eta_g \langle e_t, u_t \rangle + \eta_g^2 \|u_t\|^2. \tag{51}$$

For the honest component, Eq. 46 and strong convexity imply

$$\mathbb{E}\Big[\langle e_t, u_t^{(\mathrm{H})} \rangle \mid \mathcal{G}_t \Big] \ge \mu\rho \|e_t\|^2. \tag{52}$$

For the malicious component,

$$\|u_t^{(\mathrm{M})}\| \le \delta_{\max} m_t. \tag{53}$$

Cauchy–Schwarz and Young's inequality yield

$$-\langle e_t, u_t^{(\mathrm{M})} \rangle \le \frac{\mu\rho}{4} \|e_t\|^2 + \frac{\delta_{\max}^2}{\mu\rho} m_t^2. \tag{54}$$

Moreover,

$$\|u_t\|^2 \le 2\|u_t^{(\mathrm{H})}\|^2 + 2\|u_t^{(\mathrm{M})}\|^2.$$

Using Eq. 47, Eq. 53, and L_s-smoothness, which implies $\|\nabla F(w_t)\| \le L_s \|e_t\|$ because $\nabla F(w^\star) = 0$, gives

$$\mathbb{E}\big[\|u_t\|^2 \mid \mathcal{G}_t\big] \le 2L_s^2 \|e_t\|^2 + 2\sigma_{\mathrm{H}}^2 + 2\delta_{\max}^2 \mathbb{E}\big[m_t^2 \mid \mathcal{G}_t\big]. \tag{55}$$

Substituting these bounds into Eq. 51 yields

$$\begin{aligned} \mathbb{E}\big[\|e_{t+1}\|^2 \mid \mathcal{G}_t\big] \le \left(1 - \frac{3}{2}\eta_g \mu\rho + 2\eta_g^2 L_s^2 \right) \|e_t\|^2 + 2\eta_g^2 \sigma_{\mathrm{H}}^2 \\ + \left(\frac{2\eta_g \delta_{\max}^2}{\mu\rho} + 2\eta_g^2 \delta_{\max}^2 \right) \mathbb{E}\big[m_t^2 \mid \mathcal{G}_t\big]. \end{aligned} \tag{56}$$

Equation 48 implies

$$2\eta_g^2 L_s^2 \le \frac{\eta_g \mu\rho}{2}, \qquad \eta_g \mu\rho \le 1.$$

Consequently,

$$\mathbb{E}\big[\|e_{t+1}\|^2 \mid \mathcal{G}_t\big] \le c_g\|e_t\|^2 + 2\eta_g^2\sigma_{\mathrm{H}}^2 + \frac{4\eta_g\delta_{\max}^2}{\mu\rho}\mathbb{E}\big[m_t^2 \mid \mathcal{G}_t\big]. \tag{57}$$

Taking expectations and unrolling, then using

$$\sum_{t=0}^{H-1} c_g^{H-1-t} \le \frac{1}{1-c_g} = \frac{1}{\eta_g\mu\rho},$$

gives Eq. 49.

Finally, $0 \le m_t \le 1$ implies $m_t^2 \le m_t$, while

$$m_t = \sum_{k\in\mathcal{S}_t^{\mathrm{M}}} p_k(t) \le \sum_{k\in\mathcal{K}_{\mathrm{M}}} p_k(t)$$

and $c_g^{H-1-t} \le 1$. Therefore, pathwise,

$$\sum_{t=0}^{H-1} c_g^{H-1-t} m_t^2 \le \sum_{k\in\mathcal{K}_{\mathrm{M}}} \Psi_k(H).$$

The first inequality in Eq. 50 follows by taking expectations. Applying Theorem 1 to every identity in the fixed finite cohort $\mathcal{K}_{\mathrm{M}}$ and then taking expectations proves the second inequality. □

Corollary 2 (Asymptotic optimization neighborhoods)**.** *Under Theorem 2:*

(i) if the post-pivot malicious set $\mathcal{K}_{\mathrm{M}}$ is finite and every malicious identity satisfies Proposition 2, then

$$\limsup_{H\to\infty} \mathbb{E}\big[\|w_H - w^\star\|^2\big] \le \frac{2\eta_g\sigma_{\mathrm{H}}^2}{\mu\rho}; \tag{58}$$

(ii) under deployed re-entry, if

$$\limsup_{t\to\infty} \mathbb{E}[m_t^2] \le \bar{m}^{\,2}, \tag{59}$$

then

$$\limsup_{H\to\infty} \mathbb{E}\big[\|w_H - w^\star\|^2\big] \le \frac{2\eta_g\sigma_{\mathrm{H}}^2}{\mu\rho} + \frac{4\delta_{\max}^2}{\mu^2\rho^2}\bar{m}^{\,2}. \tag{60}$$

Proof. For part (i), Proposition 2 and Eq. 42 imply

$$\mathbb{E}[p_k(t)] \le \frac{\phi(B_k^{\max})\Gamma_k}{Z_{\min}}\,\xi_k^t.$$

Since

$$m_t^2 \le m_t \le \sum_{k\in\mathcal{K}_{\mathrm{M}}} p_k(t),$$

the adversarial term in Eq. 49 is bounded by a finite sum of geometric convolutions. For each k,

$$\sum_{t=0}^{H-1} c_g^{H-1-t}\xi_k^t = \begin{cases} \dfrac{c_g^H - \xi_k^H}{c_g - \xi_k}, & c_g \ne \xi_k, \\ Hc_g^{H-1}, & c_g = \xi_k. \end{cases}$$

This expression vanishes as $H \to \infty$ because $c_g \in [0,1)$ and $\xi_k \in (0,1)$. For a finite malicious set, the complete adversarial convolution therefore vanishes. The initialization term also vanishes, proving Eq. 58.

Eq. 59 implies that there exists t_υ such that

$$\mathbb{E}[m_t^2] \le \bar{m}^{\,2} + \upsilon$$

for all $t \ge t_\upsilon$. Splitting the discounted convolution at t_υ, the finite prefix vanishes as $H \to \infty$, while

$$\limsup_{H\to\infty} \sum_{t=t_\upsilon}^{H-1} c_g^{H-1-t}\mathbb{E}[m_t^2] \le \frac{\bar{m}^{\,2} + \upsilon}{1-c_g}.$$

Letting $\upsilon \downarrow 0$ and using $1 - c_g = \eta_g\mu\rho$ gives Eq. 60. □

Scope. The persistent-influence results do not require the optimization assumptions above. Theorem 2 additionally assumes strong convexity, smoothness, bounded malicious updates, and a uniformly positive effective honest-gradient coefficient $\rho > 0$. Equation 1 explains why the latter cannot hold uniformly under unrestricted identity proliferation and identity-symmetric sampling. The theorem and corollary are therefore analytical consequences of influence control, not direct convergence guarantees for the neural networks evaluated in this work.

C Experimental Protocol and Attack Details

C.1 Attack-independent calibration

All security-sensitive detector hyperparameters and all influence-decay, renewal, and enforcement parameters are selected using benign-only development runs disjoint from the five evaluation seeds and fixed before malicious evaluation. The adaptive reference and threshold subsequently evolve only through the committed online rules. ASR, malicious acceptance or renewal rates, and adversarial trajectories are never used for parameter selection.

Each tunable defense receives the same benign-only development protocol and no access to attack outcomes. We evaluate up to 20 hyperparameter configurations per method over three development seeds, using the authors' recommended settings as the center of each search range. Selection uses only benign validation accuracy and, where applicable, benign false-rejection behavior; final test data are never used for hyperparameter selection. The selected configuration is then frozen before the five malicious evaluation seeds.

Let $\widehat{\pi}_{\mathrm{H}}^{\mathrm{dev}}$ denote the client-balanced empirical selection frequency and $\widehat{\alpha}_{\mathrm{FR}}^{\mathrm{dev}}$ the benign false-rejection rate. We initialize

$$V_a^{(0)} = \frac{B_0(1 - e^{-\lambda})}{\widehat{\pi}_{\mathrm{H}}^{\mathrm{dev}}(1 - \widehat{\alpha}_{\mathrm{FR}}^{\mathrm{dev}})}, \tag{61}$$

which places the expected benign steady-state influence near B_0. For controlled CIFAR, $\widehat{\pi}_{\mathrm{H}}^{\mathrm{dev}} = 0.10$ and $\widehat{\alpha}_{\mathrm{FR}}^{\mathrm{dev}} \approx 0.04$ give $V_a^{(0)} \approx 50.8$. The dataset-specific value is then calibrated on the corresponding benign development run. We parameterize $\tau_{\min} = c_B B_0$, with $c_B \in (0, 1)$ selected under the same benign-only protocol.

The calibration quantile is $\nu = 0.95$. Candidate configurations must satisfy: (i) client-balanced false rejection of at most 5%, (ii) clean validation accuracy within one percentage point of the unconstrained benign reference on a development split disjoint from the final test set, and (iii) no persistent exclusion of honest clients through probationary re-entry. Among admissible configurations, we favor tighter (Δ_M, Δ_τ). The remaining detector parameters $(\alpha, \gamma, \kappa, \eta_S, n_{\mathcal{B}}, n_{\mathcal{W}})$ and any refinement of (V_a, c_B) are selected under the same protocol and then fixed.

C.2 Consolidated hyperparameters

Tables 5 and 6 report the canonical experimental configuration and declared sensitivity ranges.

C.3 Reported acceptance and renewal rates

All reported rates use the post-pivot set $\mathcal{T}_{\mathrm{post}} = \{T_p, \ldots, T-1\}$ and $H = |\mathcal{T}_{\mathrm{post}}| = T - T_p$. Equivalently, the theoretical analysis reindexes this set as $\{0, \ldots, H-1\}$. For malicious identity k, define

$$N_k^{\mathrm{sub}}(H) \triangleq \sum_{t \in \mathcal{T}_{\mathrm{post}}} J_k(t), \qquad \mathcal{R}_k(H) \triangleq \sum_{t \in \mathcal{T}_{\mathrm{post}}} R_k(t).$$

For every malicious identity k, define

$$\widehat{\beta}_{R,k} = \frac{\mathcal{R}_k(H)}{H}, \qquad \widehat{\beta}_{A,k} = \frac{\mathcal{R}_k(H)}{N_k^{\mathrm{sub}}(H)} \quad \text{when } N_k^{\mathrm{sub}}(H) > 0. \tag{62}$$

Thus $\widehat{\beta}_{A,k}$ is conditional on post-pivot submission, whereas $\widehat{\beta}_{R,k}$ is renewal per post-pivot round and equals zero for an identity with no post-pivot submission.

Table 5: Federated, influence, and enrollment parameters. Security-critical protocol parameters are shown in bold.

Symbol	Meaning	Value
Federated setup		
N	controlled CIFAR enrolled identities	100
n	sampled identities per round	10
T	global training rounds	200
α_D	Dirichlet concentration	$\{1.0, 0.5, 0.3\}$
η_{loc}	local optimizer learning rate	0.01
η_g	server aggregation step size	1.0
E_{loc}	local epochs	5
b_{batch}	local batch size	32
	independent evaluation seeds	5
Influence dynamics		
$\boldsymbol{\lambda}$	influence decay rate	0.05 (half-life ≈ 13.9 rounds)
$\boldsymbol{B_0}$	initial influence state	100
$\boldsymbol{V_a}$	accepted-renewal increment	50.8 for controlled CIFAR; benign-calibrated otherwise
$\boldsymbol{\tau_{\min}}$	influence threshold	30
c_B	threshold fraction, $\tau_{\min} = c_B B_0$	0.30
ϕ	influence-to-weight map	$\log(1+x)$
$\boldsymbol{(q, m)}$	acceptance-attestation quorum	$(3, 5)$
Enrollment		
$\boldsymbol{n_{\mathrm{data}}^{\min}}$	minimum dataset size	200 images CIFAR; 50 images FEMNIST; 500 characters Shakespeare
ϵ	enrollment privacy budget	1.0 default; sweep $\{0.5, 1, 2, 4, 8, 16\}$
n_{hist}	vision histogram bins per channel	64
n_{bin}	hashed transition bins	128
	CIFAR conditioning dimension	192
	FEMNIST conditioning dimension	64
	Shakespeare conditioning dimension	$\lvert\Sigma_{\mathrm{char}}\rvert + n_{\mathrm{bin}}$
	vision circuit size	$\approx$ 420k R1CS
	text circuit size	$\approx$ 0.9M R1CS

For the fixed malicious evaluation cohort $\mathcal{K}_{\mathrm{M}}$, provided the cohort makes at least one post-pivot submission, the pooled conditional acceptance rate and mean per-identity renewal rate are

$$\widehat{\beta}_A = \frac{\sum_{k \in \mathcal{K}_{\mathrm{M}}} \mathcal{R}_k(H)}{\sum_{k \in \mathcal{K}_{\mathrm{M}}} N_k^{\mathrm{sub}}(H)}, \tag{63}$$

and

$$\widehat{\beta}_R = \frac{1}{|\mathcal{K}_{\mathrm{M}}|} \sum_{k \in \mathcal{K}_{\mathrm{M}}} \widehat{\beta}_{R,k}. \tag{64}$$

Let $\mathcal{H}$ be the fixed honest evaluation cohort and

$$\mathcal{H}_+ \triangleq \left\{ i \in \mathcal{H} : \sum_{t \in \mathcal{T}_{\mathrm{post}}} J_i(t) > 0 \right\}.$$

The client-balanced honest false-rejection rate and selection frequency are

$$\widehat{\alpha}_{\mathrm{FR}} \triangleq \frac{1}{|\mathcal{H}_+|} \sum_{i \in \mathcal{H}_+} \frac{\sum_{t \in \mathcal{T}_{\mathrm{post}}} J_i(t)[1 - A_i(t)]}{\sum_{t \in \mathcal{T}_{\mathrm{post}}} J_i(t)}, \tag{65}$$

$$\widehat{\pi}_{\mathrm{H}} \triangleq \frac{1}{|\mathcal{H}|} \sum_{i \in \mathcal{H}} \frac{1}{H} \sum_{t \in \mathcal{T}_{\mathrm{post}}} J_i(t). \tag{66}$$

Table 6: Behavioral-detector and adaptive-attack parameters.

Symbol	Meaning	Value
Dual-horizon behavioral detector		
d'	projected fingerprint dimension	1024
T_{cal}	calibration rounds	50 CIFAR; 60 FEMNIST/Shakespeare
χ	calibration fit fraction	0.8
ν	calibration quantile	0.95
	C-VAE hidden / latent dimension	512 / 128; 3-layer MLP
β_{KL}	C-VAE KL coefficient	0.1
η_L	frozen-reference learning rate	10^{-3}
	frozen-reference optimizer	Adam
	C-VAE training batch size	64
	maximum calibration epochs	100
	early-stopping patience	10 epochs
	weight decay	10^{-5}
η_S	adaptive-reference learning rate	10^{-4}
α	frozen/adaptive score mixture	0.60
γ	reference-discrepancy weight	0.25
κ	latent-discrepancy weight	0.10
Δ_M	functional reference budget	0.02
Δ_τ	threshold adaptation budget	$0.05\tau_L$
$n_{\mathcal{B}}$	adaptive replay-buffer capacity	256
$n_{\mathcal{W}}$	recent-score window capacity	256
$\mathcal{L}_{\mathrm{mon}}$	monitored layer set	early block / late block / head
Adaptive attack configuration		
T_p	behavioral-pivot round	100; sensitivity $\{80, 120\}$
$a_{\max}$	maximum poisoning strength	1.0; sweep $\{0.5, 1.0, 2.0\}$
T_{grad}	gradual-pivot duration	$\{10, 20, 50\}$; default 20
c_δ	acceptance-margin coefficient	0.02; sweep $\{0.01, 0.02, 0.05\}$
ζ_M	adaptive-reference attack weight	0.5; sweep $\{0.1, 0.5, 1.0\}$
ζ_τ	threshold-retention attack weight	0.5; sweep $\{0.1, 0.5, 1.0\}$
ζ_L	monitored-projection penalty	1.0; sweep $\{0.5, 1.0, 2.0\}$
$\delta_{\max}$	malicious update-norm cap	$1.5\, Q_{0.95}^{\mathrm{benign}}(\|\Delta w\|)$

Finally, empirical honest aggregation mass and malicious sampling occupancy are

$$\widehat{m}_{\mathrm{H}} \triangleq \frac{1}{H} \sum_{t \in \mathcal{T}_{\mathrm{post}}} \sum_{i \in \mathcal{S}_t^{\mathrm{H}}} p_i(t), \tag{67}$$

$$\widehat{o}_{\mathrm{M}} \triangleq \frac{1}{H} \sum_{t \in \mathcal{T}_{\mathrm{post}}} \frac{|\mathcal{S}_t^{\mathrm{M}}|}{|\mathcal{S}_t|}. \tag{68}$$

For experiments with one malicious identity, the pooled malicious rates reduce to their identity-level counterparts. Reported means and standard deviations are computed over five independent evaluation seeds.

C.4 Attack configurations and constructions

All attacks use the capabilities in §3. Multiple pseudonymous identities may be controlled by one adversary, each with its own committed conditioning summary and influence state. Table 7 summarizes the four evaluated modes; the constructions below specify their objectives and optimization procedures. For baseline comparisons, each attack is re-optimized against the corresponding applicable defense under the same participation, update-norm, local-computation, and poisoning budgets. DAIRE-FL-specific reference and threshold objectives are omitted for baselines that do not expose these states.

We implement poisoning strength through the local objective $\mathcal{L}_{\mathrm{loc}} = \mathcal{L}_{\mathrm{ben}} + a_t \mathcal{L}_{\mathrm{atk}}$, so $a_t = 0$ gives benign training and $a_t = a_{\max}$ gives the configured attack strength. Target objectives are fixed

before evaluation and held constant across methods. CIFAR-10 uses $2 \to 8$ targeted label flipping, CIFAR-100 uses *apple→orange*, FEMNIST uses $2 \to 8$, and Shakespeare uses e→t next-character targeting. For Mode D, vision experiments use a fixed 3×3 lower-right trigger with the same target class; sequence experiments use a fixed zx context trigger with target character t. For the maximization problems below, define

$$\mathcal{U}_{\mathrm{atk}}(\Delta w; w_t) \triangleq -\mathcal{L}_{\mathrm{atk}}(\Delta w; w_t), \qquad \mathcal{U}_{\mathrm{bd}}(\Delta w; w_t) \triangleq -\mathcal{L}_{\mathrm{bd}}(\Delta w; w_t),$$

where $\mathcal{L}_{\mathrm{bd}}$ is the trigger-target loss used in Mode D. The feasible update set $\mathcal{U}_k^t$ contains updates reachable from w_t by client k under the stated local optimizer, learning rate, local-epoch, mini-batch, and poisoning budgets. The detector-admissibility and update-norm constraints are displayed explicitly below rather than absorbed into $\mathcal{U}_k^t$.

Table 7: Adaptive attack modes used in the evaluation.

Mode	Axis	Definition
A. Abrupt pivot	temporal	$a_t = a_{\max}\mathbf{1}\{t \geq T_p\}$
B. Gradual pivot	temporal	$a_t = a_{\max}\min\{1, (t-T_p)^+/T_{\mathrm{grad}}\}$
C. Threshold-aware pivot	temporal	White-box optimization of attack utility, adaptive-reference displacement, and resistance to threshold tightening while remaining admissible.
D. Layer-selective backdoor	spatial	Attack utility concentrated in selected parameter subsets while constraining behavior visible to the monitored projection.

Mode A: abrupt behavioral pivot. Adversarial identities follow the benign objective for $t < T_p$ and switch to the malicious objective at $t = T_p$, so the primary attack begins after calibration. The task-level targets are those defined above. The white-box implementation optimizes current attack utility subject to detector admissibility; unlike Mode C, it does not optimize future reference or threshold adaptation. Default and sensitivity parameters appear in Table 6.

Mode B: gradual behavioral pivot. Poisoning strength follows

$$a_t = a_{\max}\min\left\{1, \frac{(t-T_p)^+}{T_{\mathrm{grad}}}\right\}, \qquad T_{\mathrm{grad}} \in \{10, 20, 50\}, \tag{69}$$

where $(x)^+ = \max\{x, 0\}$. Increasing T_{grad} reduces per-round behavioral change while preserving the final attack strength.

Mode C: threshold-aware adaptive pivot. This attacker targets both adaptive feedback channels. For a candidate fingerprint x, write $r_k^t(x, S_k)$ for the right-hand side of Eq. 6 evaluated using the round-t detector state. Let $\mathcal{B}_t$ be the adaptive replay buffer at the beginning of round t. For candidate (x, S_k), let $\mathsf{Adapt}_t(x, S_k)$ denote the parameters obtained by inserting the candidate into the bounded replay buffer and applying Eqs. 34, 35, and 8. Likewise,

$$\mathsf{Thr}_t(r) = \max\left\{\tau_L - \Delta_\tau,\, \min\left\{\tau_t,\, Q_\nu(\mathcal{W}_t \oplus r)\right\}\right\}, \tag{70}$$

where $\mathcal{W}_t \oplus r$ inserts candidate score r according to the deployed bounded-window rule. The exact acceptance margin is

$$\delta_t = c_\delta \tau_t, \qquad c_\delta > 0. \tag{71}$$

At round t, malicious identity k solves

$$\begin{aligned} \max_{\Delta w \in \mathcal{U}_k^t} \quad & \mathcal{U}_{\mathrm{atk}}(\Delta w; w_t) + \zeta_M \frac{D_{\mathcal{C}}\left(\mathsf{Adapt}_t(\Pi(\Delta w), S_k), \theta^L\right)}{\Delta_M} \\ & + \zeta_\tau \frac{\mathsf{Thr}_t\left(r_k^t(\Pi(\Delta w), S_k)\right) - (\tau_L - \Delta_\tau)}{\Delta_\tau} \\ \text{s.t.} \quad & r_k^t(\Pi(\Delta w), S_k) \leq \tau_t - \delta_t, \\ & \|\Delta w\|_2 \leq \delta_{\max}, \\ & B_k(t) \geq \tau_{\min}. \end{aligned} \tag{72}$$

Here $\mathcal{U}_{\text{atk}}$ is the dataset-specific maximization utility defined above, and $\zeta_M, \zeta_\tau \geq 0$ weight adaptive-reference and threshold displacement. The malicious update cap is

$$\delta_{\max} = 1.5\, Q_{0.95}^{\text{benign}}(\|\Delta w\|_2)\,, \tag{73}$$

where the benign quantile is estimated during development and frozen before malicious evaluation. The reference-drag term is omitted when $\Delta_M = 0$, and the threshold-drag term when $\Delta_\tau = 0$.

The first objective term attacks the global task, the second moves the adaptive reference toward its permitted functional boundary, and the third maximizes the next threshold permitted by the exact update rule, thereby opposing defensive threshold tightening. The condition $B_k(t) \geq \tau_{\min}$ restricts detector-state poisoning to influence-active rounds. A below-threshold identity may still earn future renewal through probationary acceptance but cannot modify the current model, replay buffer, or threshold window.

We optimize Eq. 72 by projected gradient ascent through the differentiable detector and one-step adaptive-reference update. Because empirical quantiles are piecewise constant, a differentiable soft-quantile surrogate is used only during attack construction. Feasibility and reported outcomes use the exact threshold rule in Eq. 9; among the optimization iterates, we retain the highest-utility candidate satisfying the exact acceptance and norm constraints. For baselines without differentiable detector states, candidates are optimized against the applicable task objective and evaluated using the baseline's exact aggregation or filtering rule under the same budgets.

Mode D: layer-selective backdoor. Let $\Pi_{\mathcal{L}}$ denote the projection induced by monitored layer set $\mathcal{L}$ and $\bar{x}_{\mathcal{L}}^{\text{benign}}$ the corresponding benign projected-update reference. Let $r_{k,\mathcal{L}}^t(\Delta w)$ denote the score in Eq. 6 evaluated with $x = \Pi_{\mathcal{L}}(\Delta w)$ under the same round-t detector state. For each detector configuration, the attacker solves

$$\begin{aligned} \max_{\Delta w \in \mathcal{U}_k^t} \quad & \mathcal{U}_{\text{bd}}(\Delta w; w_t) - \zeta_L \left\| \Pi_{\mathcal{L}}(\Delta w) - \bar{x}_{\mathcal{L}}^{\text{benign}} \right\|_2^2 \\ \text{s.t.} \quad & r_{k,\mathcal{L}}^t(\Delta w) \leq \tau_t - \delta_t, \\ & \|\Delta w\|_2 \leq \delta_{\max}. \end{aligned} \tag{74}$$

Here $\mathcal{U}_{\text{bd}}$ is the fixed trigger-target backdoor objective defined after the local poisoning objective. We compare $\mathcal{L} = \{\text{head}\}$ with $\mathcal{L} = \{\text{early}, \text{late}, \text{head}\}$. The attacker is re-optimized separately against each monitored layer set rather than transferring a head-optimized attack to the multi-layer detector.

Identity-proliferation construction. We fix $N_H = 100$ honest identities and add $K \in \{1, 5, 10, 20, 50, 100\}$ adversarial pseudonyms. Consistent with capability (P2), adversarial identities may reuse the same qualifying dataset or independently resampled variants. Sampling remains uniform over available enrolled identities and independent of influence state, with $n = 10$ identities sampled per round whenever at least ten are available. Increasing K therefore changes adversarial occupancy rather than dividing a fixed adversarial data budget among identities.

D Additional Experimental Results

D.1 Complete baseline results

Table 8 reports the complete per-baseline ASR comparison. DAIRE-FL achieves the lowest ASR in all seven settings. Across CIFAR-10 Modes A–D, its ASR is 6.8%–12.6%, compared with 38.6%–57.9% for the strongest baseline in each setting. Under Mode C on CIFAR-100, FEMNIST, and Shakespeare, DAIRE-FL obtains 10.9%–15.4% ASR, whereas the strongest baselines obtain 39.8%–55.6%. We report only configurations supported by each baseline rather than imputing unsupported settings. All methods use the same training schedule, participation budget, attack budget, and $T = 200$ rounds; defense-specific logic follows the corresponding implementation.

Attack constructions are provided in Appendix C.4, while heterogeneity-specific benign behavior is reported in Appendix D.6.

Clean-utility comparison. Table 9 reports clean accuracy for the same configurations as Table 8. DAIRE-FL remains within 0.5–1.0 percentage points of FedAvg across all settings. Under Mode C,

Table 8: Per-baseline ASR (%) under the evaluated adaptive attacks. Lower is better.

Mode	Dataset	FedAvg	Krum	FLAME	FG	FLDet.	AdaRL	DP-B	SPMC	RoFL	Aion	DAIRE
A	CIFAR-10	92.0±5.0	77.0±5.5	63.0±5.0	72.0±5.2	53.0±4.8	48.0±4.9	38.6±4.9	45.0±4.7	70.0±5.1	42.0±4.6	6.8±1.7
B	CIFAR-10	93.0±5.2	80.0±5.8	69.0±5.3	76.0±5.4	60.0±5.0	53.0±5.1	45.2±5.6	51.0±5.0	73.0±5.3	48.0±4.9	9.7±2.1
C	CIFAR-10	94.0±5.3	87.0±5.7	79.0±5.4	83.0±5.6	70.0±5.1	65.0±5.0	57.9±6.4	62.0±5.0	80.0±5.8	60.0±5.0	12.6±2.8
	CIFAR-100	95.0±5.4	88.0±5.8	81.0±5.5	84.0±5.6	72.0±5.2	67.0±5.1	55.6±6.1	63.0±5.3	82.0±5.7	61.0±5.0	15.4±3.1
	FEMNIST	90.0±5.1	75.0±5.2	64.0±4.9	70.0±5.0	54.0±4.6	50.0±4.5	42.5±4.8	49.0±4.7	67.0±5.0	45.0±4.4	10.9±2.3
	Shakespeare	86.0±5.0	70.0±5.0	60.0±4.8	64.0±4.9	50.0±4.6	47.0±4.5	42.0±4.4	45.0±4.5	59.0±4.8	39.8±5.1	13.6±2.9
D	CIFAR-10	94.0±5.2	80.0±5.6	68.0±5.1	77.0±5.5	63.0±5.0	58.0±4.9	55.0±5.0	48.7±5.2	74.0±5.4	52.0±4.8	11.3±2.4

Modes A–D denote abrupt pivot, gradual pivot, threshold-aware pivot, and layer-selective backdoor, respectively. FG: FoolsGold; FLDet.: FLDetector; AdaRL: AdaAggRL; DP-B: DP-BREM.

Table 9: Per-baseline clean test accuracy (%). Higher is better.

Mode	Dataset	FedAvg	Krum	FLAME	FG	FLDet.	AdaRL	DP-B	SPMC	RoFL	Aion	DAIRE
A	CIFAR-10	82.6±0.7	79.4±1.0	81.1±0.8	81.8±0.7	81.5±0.8	81.2±0.9	81.3±0.8	80.9±0.9	81.7±0.8	81.4±0.8	81.9±0.7
B	CIFAR-10	82.3±0.8	79.1±1.1	80.8±0.9	81.5±0.8	81.2±0.8	80.9±0.9	81.0±0.9	80.7±1.0	81.4±0.9	81.1±0.9	81.4±0.8
C	CIFAR-10	81.8±0.9	78.8±1.2	80.4±1.0	81.0±0.9	80.7±0.9	80.5±1.0	80.6±1.0	80.3±1.1	80.9±1.0	80.6±1.0	80.8±0.9
	CIFAR-100	62.5±1.0	58.9±1.3	60.8±1.1	61.4±1.0	61.0±1.1	60.6±1.1	60.7±1.1	60.3±1.2	61.2±1.1	60.8±1.1	61.7±0.9
	FEMNIST	87.3±0.6	84.6±0.9	85.9±0.7	86.5±0.7	86.3±0.7	85.9±0.8	86.1±0.7	85.8±0.8	86.4±0.7	86.0±0.8	86.7±0.6
	Shakespeare	53.8±0.6	50.7±0.8	52.0±0.7	52.7±0.6	52.6±0.6	52.3±0.7	52.4±0.6	52.1±0.7	52.8±0.6	52.4±0.7	53.1±0.5
D	CIFAR-10	81.7±0.9	78.9±1.2	80.5±1.0	80.9±0.9	80.6±1.0	80.4±1.0	80.5±1.0	80.2±1.1	80.9±1.0	80.5±1.0	81.2±0.8

DAIRE-FL achieves 80.8%, 61.7%, 86.7%, and 53.1% on CIFAR-10, CIFAR-100, FEMNIST, and Shakespeare, compared with FedAvg accuracies of 81.8%, 62.5%, 87.3%, and 53.8%, respectively. Krum exhibits the largest clean-accuracy loss, reaching 3.6 percentage points on CIFAR-100. Thus, DAIRE-FL's ASR reduction does not require a large loss in clean utility.

D.2 Bound-validation diagnostic

For the malicious cohort $\mathcal{K}_{\mathrm{M}}$, define cumulative post-pivot aggregation weight as $\Psi_{\mathrm{M}}(H) \triangleq \sum_{k\in\mathcal{K}_{\mathrm{M}}} \Psi_k(H)$. For each realized trajectory, let

$$Z_{\min}^{\mathrm{obs}} \triangleq \min_{t\in\mathcal{T}_{\mathrm{post}}:Z_t>0} Z_t$$

be its smallest positive aggregation normalizer. We evaluate the corresponding cohort-level normalization diagnostic as

$$U_{\mathrm{M},Z}^{\Psi,\mathrm{obs}}(H) \triangleq \sum_{k\in\mathcal{K}_{\mathrm{M}}} \frac{b_k + V_a\mathcal{R}_k(H)}{Z_{\min}^{\mathrm{obs}}(1-e^{-\lambda})}.$$

Table 2 in the main paper compares $\Psi_{\mathrm{M}}(H)$ with this diagnostic. All evaluated trajectories contain at least one positive-normalizer round. The diagnostic is trajectory-specific and evaluates only the normalization branch of Theorem 1; it is not the full minimum in Eq. 14. The pathwise guarantee remains valid with the protocol-level choice $Z_{\min} = \phi(\tau_{\min}) = \log(1+\tau_{\min})$.

D.3 Adaptive-detector and influence-decay ablation

Table 10 evaluates the detector components under threshold-aware Mode C, which directly targets the feedback channels introduced by online adaptation. All variants use the same benign-only calibration protocol, task model, attack initialization, and optimization budget. On CIFAR-10, adding the frozen reference reduces ASR from 62.4% to 41.6%, constraining reference adaptation further reduces it to 26.4%, and the full detector reaches 12.6%. The corresponding conditional acceptance rates decrease from 0.74 to 0.50, 0.32, and 0.18. FEMNIST shows the same progression, with ASR decreasing from 58.7% to 38.4%, 23.1%, and 10.9%, respectively. Removing influence decay raises ASR to 29.8% on CIFAR-10 and 27.1% on FEMNIST, although conditional acceptance changes only from 0.18 to 0.19 and from 0.17 to 0.18. This confirms that detector improvements and influence decay provide complementary protection.

Table 10: Detector and influence-decay ablation under threshold-aware Mode C.

Dataset	Variant	Clean acc.	ASR	$\widehat{\beta}_A$	$\widehat{\beta}_R$	$\widehat{\alpha}_{\mathrm{FR}}$
CIFAR-10	Single adaptive reference	77.8±1.4%	62.4±6.1%	0.74±0.07	0.071±0.009	3.1±0.7%
	Dual horizon only	79.0±1.2%	41.6±5.4%	0.50±0.06	0.048±0.008	3.8±0.8%
	Dual horizon + reference constraint	80.1±1.0%	26.4±4.1%	0.32±0.05	0.031±0.006	4.4±0.8%
	Full DAIRE-FL detector	80.8±0.9%	12.6±2.8%	0.18±0.04	0.017±0.005	4.9±0.8%
	Full detector without influence decay	80.5±1.0%	29.8±4.5%	0.19±0.04	0.018±0.005	4.8±0.9%
FEMNIST	Single adaptive reference	84.2±1.1%	58.7±5.7%	0.70±0.07	0.067±0.009	3.0±0.7%
	Dual horizon only	85.1±0.9%	38.4±4.9%	0.47±0.06	0.045±0.007	3.6±0.7%
	Dual horizon + reference constraint	86.0±0.8%	23.1±3.8%	0.29±0.05	0.028±0.006	4.2±0.8%
	Full DAIRE-FL detector	86.7±0.6%	10.9±2.3%	0.17±0.04	0.016±0.004	4.7±0.8%
	Full detector without influence decay	86.4±0.7%	27.1±4.0%	0.18±0.04	0.017±0.004	4.6±0.8%

D.4 Spatial adaptivity

Table 11 compares head-only and multi-layer monitoring under separately optimized Mode D attacks on CIFAR-10, CIFAR-100, and FEMNIST. Re-optimizing the attack for each monitored layer set avoids favoring multi-layer monitoring through transfer of a head-specific attack. Multi-layer monitoring reduces ASR from 44.9%–52.6% to 10.7%–14.8%, conditional acceptance from 0.57–0.64 to 0.15–0.19, and renewal from 0.054–0.061 to 0.014–0.018. Clean accuracy is 1.2–1.8 percentage points higher, while honest false rejection remains at most 5.0%.

Table 11: Head-only versus multi-layer behavioral monitoring under Mode D.

Dataset	Monitoring	Clean acc.	ASR	$\widehat{\beta}_A$	$\widehat{\beta}_R$	$\widehat{\alpha}_{\mathrm{FR}}$
CIFAR-10	Head only	79.4±1.1%	48.7±5.2%	0.61±0.06	0.058±0.007	3.6±0.8%
	Multi-layer	81.2±0.8%	11.3±2.4%	0.16±0.03	0.015±0.004	4.8±0.9%
CIFAR-100	Head only	59.8±1.2%	52.6±5.7%	0.64±0.07	0.061±0.008	3.9±0.9%
	Multi-layer	61.5±1.0%	14.8±3.0%	0.19±0.04	0.018±0.005	5.0±1.0%
FEMNIST	Head only	85.3±0.9%	44.9±4.9%	0.57±0.06	0.054±0.007	3.8±0.8%
	Multi-layer	86.5±0.7%	10.7±2.3%	0.15±0.03	0.014±0.004	4.6±0.8%

D.5 Identity proliferation

Table 12 extends the main proliferation results by reporting the summed observed normalization-branch diagnostic and clean accuracy. The experiment fixes $N_H = 100$ honest identities and varies $K \in \{1, 5, 10, 20, 50, 100\}$ adversarial pseudonyms under dataset reuse and identity-symmetric sampling. Realized cumulative adversarial weight remains below the observed diagnostic, from 0.84 versus 2.31 at $K = 1$ to 44.58 versus 96.37 at $K = 100$. Over the same range, clean accuracy falls from 82.4% to 73.8% as adversarial occupancy rises. This complements Eq. 1: DAIRE-FL controls influence per identity but does not claim K-independent robustness as identity proliferation drives honest participation downward.

Table 12: Identity proliferation under dataset reuse and identity-symmetric sampling.

K	Malicious occupancy	$\Psi_{\mathrm{M}}(H)$	Observed Z-branch	Honest agg. mass	ASR	Clean acc.
1	1.0±0.3%	0.84±0.19	2.31±0.42	0.992±0.004	3.8±1.1%	82.4±0.6%
5	4.9±0.7%	3.72±0.54	8.64±1.13	0.971±0.009	5.6±1.4%	82.0±0.7%
10	9.2±0.9%	6.94±0.82	15.27±1.76	0.944±0.014	7.9±1.8%	81.4±0.8%
20	16.8±1.2%	12.86±1.41	27.91±2.68	0.895±0.021	12.8±2.3%	80.2±1.0%
50	33.5±1.7%	26.73±2.63	59.84±4.91	0.781±0.034	24.6±3.6%	77.6±1.4%
100	50.1±2.0%	44.58±3.87	96.37±7.44	0.642±0.047	38.9±4.8%	73.8±1.9%

D.6 Benign participation and fairness

Table 13 evaluates whether influence decay and behavioral filtering persistently disadvantage benign clients. We report client-balanced $\widehat{\alpha}_{\mathrm{FR}}$, the 95th-percentile per-client rejection rate, time below $\tau_{\min}$,

recovery time after benign inactivity, and the clean-accuracy gap relative to FedAvg. For each honest identity i with at least one evaluation-period submission, let

$$\widehat{\alpha}_{\mathrm{FR},i} = \frac{\sum_{t\in\mathcal{T}_{\mathrm{fair}}} J_i(t)[1 - A_i(t)]}{\sum_{t\in\mathcal{T}_{\mathrm{fair}}} J_i(t)}.$$

The tail rate is the empirical 0.95-quantile of $\{\widehat{\alpha}_{\mathrm{FR},i}\}$ across eligible honest identities. For each identity, the fraction of time below threshold is

$$\frac{1}{H_{\mathrm{fair}}} \sum_{t\in\mathcal{T}_{\mathrm{fair}}} \mathbf{1}\{B_i(t) < \tau_{\min}\}.$$

We average this quantity across honest identities and report it as a percentage of global rounds. As α_D decreases from 1.0 to 0.3, mean false rejection rises from 2.1% to 4.2%, the 95th-percentile client rate from 4.7% to 8.6%, and mean re-entry time from 1.4 to 3.0 rounds. The mean rate remains below the 5% calibration target, and the clean-accuracy gap remains below one percentage point.

Table 13: Benign participation under heterogeneous client data.

α_D	Accuracy gap vs. FedAvg	Mean FR	95th-pct. client FR	Time below $\tau_{\min}$	Re-entry time
1.0	0.3±0.2 pp	2.1±0.5%	4.7±1.0%	1.6±0.4%	1.4±0.5 rounds
0.5	0.6±0.3 pp	3.0±0.6%	6.3±1.2%	2.7±0.7%	2.1±0.6 rounds
0.3	0.9±0.4 pp	4.2±0.8%	8.6±1.5%	4.1±0.9%	3.0±0.8 rounds

D.7 Cross-modality conditioning

Table 14 compares an unconditioned C-VAE, a C-VAE conditioned on the privatized unigram block $S_i^{(1)}$, and a C-VAE conditioned on the full privatized summary S_i. This isolates the statistical value of the transition block while holding the enrollment relation and privatization mechanism fixed.

Table 14: Shakespeare conditioning ablation; clean accuracy is next-character prediction accuracy.

Conditioning	**Clean acc.**	$\widehat{\alpha}_{\mathrm{FR}}$	$\widehat{\beta}_A$	$\widehat{\beta}_R$	**ASR**
None	51.8±0.7%	5.6±1.0%	0.43±0.06	0.041±0.007	31.2±4.5%
Privatized unigram	52.6±0.6%	4.8±0.9%	0.30±0.05	0.029±0.006	22.5±3.8%
Privatized unigram + transition	53.1±0.5%	4.1±0.8%	0.18±0.04	0.017±0.004	13.6±2.9%

Full conditioning reduces ASR from 31.2% without conditioning and 22.5% with unigram-only conditioning to 13.6%. It also reduces conditional acceptance from 0.43 to 0.18 and renewal from 0.041 to 0.017, while increasing clean accuracy from 51.8% to 53.1%.

D.8 Parameter sensitivity

Table 15 varies one parameter at a time around the configuration fixed by benign calibration for CIFAR-10 Mode C. Tables 5 and 6 give the canonical configuration and remaining declared sweeps. Increasing λ from 0.03 to 0.08 lowers ASR from 18.7% to 9.8%, with clean accuracy decreasing from 81.0% to 80.2%; raising $\tau_{\min}$ from 20 to 40 similarly lowers ASR from 17.3% to 9.7%. In contrast, increasing Δ_M from 0.01 to 0.04 raises ASR from 10.8% to 18.9%, while increasing Δ_τ/τ_L from 0.025 to 0.10 raises ASR from 11.5% to 17.8%. These results quantify the trade-off between security and availability induced by decay, influence gating, and the two adaptation budgets.

Equation 13 is an analytical operating condition, not a tuning objective. Define the critical renewal rate

$$\beta_R^{\mathrm{crit}} \triangleq \frac{\tau_{\min}(1 - e^{-\lambda})}{V_a}.$$

For an individual malicious identity, a realized rate $\widehat{\beta}_{R,k} > \beta_R^{\mathrm{crit}}$ lies outside the nominal expected-contraction regime. Under the additional absorbing assumptions, it also lies outside the geometric-tail regime of Proposition 2; neither event invalidates the pathwise bound in Theorem 1.

Table 15: Representative parameter sensitivity under CIFAR-10 Mode C.

Parameter	Setting	Clean acc.	ASR	$\widehat{\beta}_A$	$\widehat{\beta}_R$
λ	0.03	81.0±0.9%	18.7±3.5%	0.18±0.04	0.017±0.005
	0.05	80.8±0.9%	12.6±2.8%	0.18±0.04	0.017±0.005
	0.08	80.2±1.0%	9.8±2.4%	0.17±0.04	0.016±0.004
$\tau_{\min}$	20	81.1±0.8%	17.3±3.3%	0.19±0.04	0.018±0.005
	30	80.8±0.9%	12.6±2.8%	0.18±0.04	0.017±0.005
	40	79.9±1.1%	9.7±2.3%	0.16±0.03	0.015±0.004
Δ_M	0.01	80.4±1.0%	10.8±2.5%	0.16±0.03	0.015±0.004
	0.02	80.8±0.9%	12.6±2.8%	0.18±0.04	0.017±0.005
	0.04	81.0±0.9%	18.9±3.6%	0.24±0.05	0.023±0.006
Δ_τ/τ_L	0.025	80.5±1.0%	11.5±2.6%	0.17±0.04	0.016±0.004
	0.05	80.8±0.9%	12.6±2.8%	0.18±0.04	0.017±0.005
	0.10	81.0±0.9%	17.8±3.4%	0.23±0.05	0.022±0.006

D.9 Enrollment privacy and detector behavior

Figure 4 evaluates how privatizing S_i affects behavioral acceptance, renewal, and downstream robustness under Mode C on CIFAR-10. We vary $\epsilon \in \{0.5, 1, 2, 4, 8, 16\}$. At $\epsilon = 0.5$, $\widehat{\alpha}_{\mathrm{FR}} = 7.8\%$, $\widehat{\beta}_A = 31.0\%$, and $\widehat{\beta}_R = 3.0\%$; at the default $\epsilon = 1$, these fall to 4.9%, 18.0%, and 1.7%. The cohort-level renewal rate at $\epsilon = 0.5$ lies slightly above the nominal expected-contraction threshold, whereas the reported rates for all evaluated $\epsilon \geq 1$ settings lie below it. Because Eq. 13 is an identity-level condition, this cohort-level comparison is diagnostic rather than a uniform per-identity guarantee and does not affect Theorem 1. ASR falls from 21.8% to 12.6% and reaches 8.3% at $\epsilon = 16$, while clean test accuracy rises from 79.2% to approximately 81.4%.

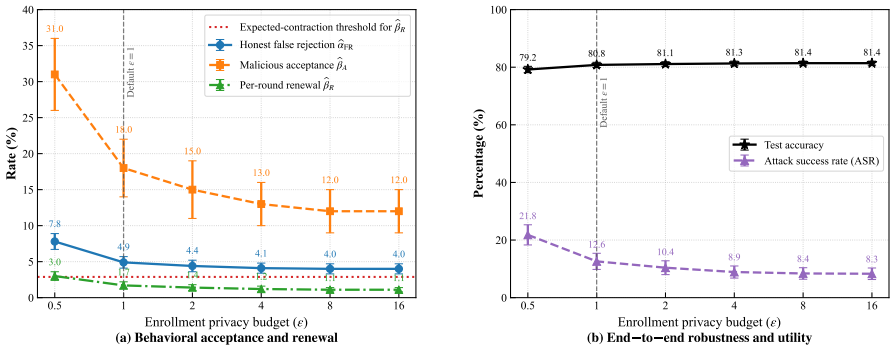


Figure 4: Enrollment privacy versus behavioral acceptance, renewal, and downstream robustness. (a) Client-balanced honest false rejection $\widehat{\alpha}_{\mathrm{FR}}$, malicious conditional acceptance $\widehat{\beta}_A$, and mean per-identity per-round renewal $\widehat{\beta}_R$ across privacy budgets ϵ. The horizontal dotted line marks the nominal expected-contraction threshold $\tau_{\min}(1 - e^{-\lambda})/V_a$; crossing it does not invalidate the pathwise finite-horizon bound. (b) Clean test accuracy and ASR under the same sweep. Error bars show standard deviation over five seeds; the vertical dashed line marks the default $\epsilon = 1$.

E Complexity and Implementation Overhead

E.1 Asymptotic complexity

Let n be the number of identities sampled per round, d_ℓ and d'_ℓ the original and projected dimensions of monitored layer ℓ, and $d' = \sum_{\ell \in \mathcal{L}_{\mathrm{mon}}} d'_\ell$. Let $n_{\mathcal{B}}$ and $n_{\mathcal{W}}$ denote the replay-buffer and recent-score-window capacities, so that $|\mathcal{B}_t|, |\mathcal{B}_t^+| \leq n_{\mathcal{B}}$ and $|\mathcal{W}_t|, |\mathcal{W}_t^+| \leq n_{\mathcal{W}}$, $C_{\mathrm{fwd}}(d')$ the cost of one C-VAE forward evaluation, $C_{\mathrm{train}}(d')$ the cost of processing one replay-buffer sample in an adaptive-reference update, $C_{\mathrm{SKL}}(d')$ the posterior-discrepancy cost, $C_{\mathrm{eval}}(d')$ the cost of one functional trust-region evaluation, and $N_{\mathrm{bt}}(t)$ the number of backtracking candidates evaluated in round t. Let

C_{sig} denote the cost of verifying one signature, C_{ZK} the enrollment proving cost, d_S the released summary dimension, $|\pi_{\mathrm{ZK}}|$ the enrollment-proof size, and $|\sigma_{\mathrm{sig}}|$ the signature size.

For dense projectors,

$$C_{\Pi} \triangleq \sum_{\ell \in \mathcal{L}_{\mathrm{mon}}} d_\ell d'_\ell. \tag{75}$$

We further write

$$C_{\mathrm{det}} \triangleq C_{\Pi} + 2C_{\mathrm{fwd}}(d') + C_{\mathrm{SKL}}(d') \tag{76}$$

for the cost of one complete behavioral-detector evaluation.

Enrollment. Constructing S_i costs $O(|\mathcal{D}_i|)$ per identity, while proof generation costs $O(C_{\mathrm{ZK}})$ once per enrollment. Enrollment communication is $O(d_S + |\pi_{\mathrm{ZK}}|)$.

Detector evaluation. Fingerprint construction costs $O(C_{\Pi})$ per selected client. Evaluating the frozen and adaptive references and their posterior discrepancy across n sampled identities costs

$$O(nC_{\Pi} + 2nC_{\mathrm{fwd}}(d') + nC_{\mathrm{SKL}}(d')) = O(nC_{\mathrm{det}}). \tag{77}$$

Adaptive-reference update. Training on the bounded replay buffer costs $O(n_{\mathcal{B}} C_{\mathrm{train}}(d'))$. With frozen-reference outputs cached, trust-region backtracking contributes

$$O(N_{\mathrm{bt}}(t)|\mathcal{C}_{\mathrm{ref}}|C_{\mathrm{eval}}(d'))\,. \tag{78}$$

Threshold update. Maintaining the recent-score window requires $O(n_{\mathcal{W}})$ storage. Computing its empirical quantile costs $O(n_{\mathcal{W}} \log n_{\mathcal{W}})$ by sorting or $O(n_{\mathcal{W}})$ by selection. Because $n_{\mathcal{W}}$ is fixed, this cost does not scale with the total enrolled population.

Attestation and enforcement. Each of the m attesters independently recomputes the committed acceptance rule. Across n submitted updates, this requires $O(mnC_{\mathrm{det}})$ total system computation per round, parallelizable across attesters. Without signature aggregation, an accepted renewal incurs $O(qC_{\mathrm{sig}})$ signature-verification work and $O(q|\sigma_{\mathrm{sig}}|)$ signature communication.

Population scaling. With fixed n, detector computation depends on sampled identities rather than total enrollment N. Increasing N primarily affects availability sampling, registry access, and state lookup. Lazy influence decay avoids periodic state-update transactions for inactive identities. Table 16 summarizes these additional computation costs and their execution frequencies.

Table 16: Additional DAIRE-FL computation beyond ordinary local training and aggregation.

Component	Asymptotic cost	Frequency
Fingerprint projection	$O(nC_{\Pi})$	per round
Dual-reference scoring	$O(n[2C_{\mathrm{fwd}}(d') + C_{\mathrm{SKL}}(d')])$	per round
Adaptive-reference training	$O(n_{\mathcal{B}} C_{\mathrm{train}}(d'))$	update rounds
Trust-region evaluation	$O(N_{\mathrm{bt}}(t)\|\mathcal{C}_{\mathrm{ref}}\|C_{\mathrm{eval}}(d'))$	update rounds
Threshold update	$O(n_{\mathcal{W}} \log n_{\mathcal{W}})$ by sorting	per round
Attester recomputation	$O(mnC_{\mathrm{det}})$ total system compute	per round
Signature verification	$O(qC_{\mathrm{sig}})$	accepted renewal
Enrollment proof generation	$O(C_{\mathrm{ZK}})$	once per identity

E.2 Empirical latency

Table 17 reports representative component latencies under the stated hardware. The detector measurements use the default controlled CIFAR-10 configuration with $n = 10$ submitted updates per round; rows marked "per round" report the aggregate cost for these ten submissions. Measurements were repeated after warm-up under a fixed software configuration, and the table reports medians. Network-dependent attestation collection additionally reports p95 latency. The enrollment measurements correspond to the paper-scale CIFAR and Shakespeare circuits. The LAN and local-EVM results were obtained using separate protocol testbeds corresponding to the released attestation and contract logic; the main experiment runner uses a local attestation surrogate and does not emulate network or ledger-confirmation delay. Consequently, these measurements isolate local computation

Table 17: Representative DAIRE-FL component latency. Per-round detector measurements use the default batch of $n = 10$ submitted updates; attestation collection reports median / p95.

Stage	Platform	Latency	Frequency
CIFAR enrollment proving	AMD EPYC 7543, 32 cores	42.0 s	once per identity
Shakespeare enrollment proving	AMD EPYC 7543, 32 cores	56.8 s	once per identity
Layerwise fingerprint projection	AMD EPYC 7543, 32 cores	13.6 ms	per round ($n = 10$)
Frozen-reference scoring	NVIDIA A100 80 GB	5.1 ms	per round ($n = 10$)
Adaptive-reference scoring	NVIDIA A100 80 GB	5.3 ms	per round ($n = 10$)
Reference discrepancy and SKL	NVIDIA A100 80 GB	1.8 ms	per round ($n = 10$)
Adaptive-reference gradient step	NVIDIA A100 80 GB	37.9 ms	update round
Trust-region evaluation	NVIDIA A100 80 GB	24.6 ms	update round
Threshold update	AMD EPYC 7543, 32 cores	0.4 ms	update round
3-of-5 attestation collection	10 GbE LAN testbed	53/98 ms	accepted submission
Contract renewal transaction	local EVM, AMD EPYC 7543	118 ms	renewal event
CIFAR-10 client training	NVIDIA A100 80 GB	1.7 s	local execution ($E_{\text{loc}} = 5$)

and communication costs rather than end-to-end deployment latency. The detector timings measure coordinator-side execution of the released implementation. Under the protocol model, each attester independently recomputes the committed acceptance decision; its total system cost is characterized asymptotically in Table 16. The 53/98 ms LAN result measures only collection of the required attestations, while the 118 ms local-EVM result measures the subsequent renewal transaction. A deployed system would additionally incur infrastructure-dependent wide-area communication, transaction propagation, and ledger-confirmation delays.

E.3 Ledger and attestation cost

Table 18 reports the signature payload, renewal-transaction gas, and attestation-collection latency for three quorum configurations. Each configuration uses a separately instantiated contract with the corresponding (q, m). The gas values cover the benchmarked `redeem` execution, including verification of the supplied ECDSA attestations and the associated state transition. The default configuration uses $(q, m) = (3, 5)$, matching the experimental configuration in Table 5. Increasing the quorum from two to five raises the signature payload from 130 B to 325 B, renewal gas from approximately 76k to 132k, and median attestation-collection latency from 41 ms to 72 ms. These costs are incurred only for accepted renewals and do not add a periodic on-chain transaction for inactive identities.

Table 18: Attested-renewal overhead under different quorum configurations. Signature bytes count only the unaggregated 65-byte ECDSA signatures. Gas covers the benchmarked local-EVM renewal transaction, and attestation latency reports median / p95 while excluding that transaction.

m	q	Signature payload	Renewal gas	Collection latency
3	2	130 B	$\approx$ 76k	41/76 ms
5	3	195 B	$\approx$ 94k	53/98 ms
7	5	325 B	$\approx$ 132k	72/141 ms