VTR-FL: VERIFIABLE TRUST RECOVERY AGAINST SELECTION-CENSORED FEEDBACK IN FEDERATED LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Reputation-based client selection in federated learning creates an endogenous feedback problem: once a client is excluded, the system may stop observing the very evidence required to determine whether that client has recovered. We formalize this phenomenon as Selection-Censored Trust (**SCT**) and show that, when post-exclusion observation probability collapses to zero, recovered and persistently malicious trajectories become unidentifiable from system-visible history. We introduce Persistent Trust Observability (**PTO**) and propose a framework named **VTR-FL**, which decouples evidence acquisition from model influence. Low-confidence clients remain observable through a verifiable zero-influence channel, while recent certified evidence, not stale historical reputation, determines whether influence can be progressively restored. A post-commit probation mechanism further constrains same-round role-conditioned exploitation. Under the stated verification and cryptographic assumptions, committed observation propensities and state transitions make recovery decisions challengeable. Across CIFAR-10, Fashion-MNIST, and MNIST, VTR-FL substantially increases recovered-client re-entry and model influence relative to reputation-based selection baselines while preserving comparable task accuracy. Adaptive evaluations further characterize the security boundary of controlled re-entry. Our code will be released soon.

1 INTRODUCTION

Federated Learning (FL) allows multiple clients to collaboratively train a model without aggregating raw data (McMahan et al., 2017). However, external obstacles, such as communication loads, impede a subset of clients from being selected per round in cross-device or cross-silo scenarios. Meanwhile, malicious clients may compromise the global model through multiple threats, such as data poisoning, model poisoning, backdoors, or free-riding, making client selection a task that must simultaneously handle resource scheduling and security defense. Therefore, long-term FL needs to suppress unreliable participation based on historical behavior and update trust assessments promptly when client behavior changes.

Existing methods typically select clients based on historical behavior and reputation, utilizing exploration or active auditing to enhance fairness and robustness (Shi et al., 2023; Wan et al., 2022; Guo et al., 2023; Chuang and Zhang, 2026). However, because only participating clients generate new evidence, the selection strategy inherently constrains future trust assessments, which means evidence visibility should be treated as an independent security constraint. Specifically, when honest clients are mistakenly downgraded due to short-term failures or severe non-independent and identically distributed (non-IID) updates, the reduced participation can further depress their probability of being observed. Consequently, even after restoring normal behavior, they struggle to supply the evidence required to rectify prior misjudgments. This dynamic traps the system in a degenerative feedback loop: diminished trust reduces observability, depriving the system of recovery evidence and perpetuating marginalized trust, ultimately prolonging the exclusion of rehabilitated clients.

Our core observation is that while trust recovery requires acquiring new evidence, this process must not automatically grant clients the ability to influence the global model. We formalize this feedback loop as reputation-induced trust censorship and introduce persistent trust observability as

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.

a necessary condition for identifying client recovery. Building on this idea, we propose VTR-FL, which combines verifiable persistent observation, isolated evidence generation, and blind re-entry. Low-confidence clients perform genuine local training in a verification channel, where their updates are used only for recovery evidence and receive zero aggregation weight. To avoid stale historical reputation suppressing recovered clients, VTR-FL separates long-term trust maintenance from recovery detection: recent recovery-specific evidence and its short-term trend determine promotion. This design decouples trust rehabilitation from model contribution and avoids reopening the poisoning surface merely to collect new evidence.

Specifically, VTR-FL makes candidate selection, observation propensity, evidence generation, and state transitions verifiable and challengeable. Clients first commit their updates, after which on-chain randomness selects auditors to generate signed evidence. Recovery-specific evidence is evaluated through a short-window level-and-trend gate, while an auxiliary risk veto constrains high-risk promotion. Recovery candidates then enter blind probation, where the update is fixed before an unpredictable auditing or restricted-influence role is assigned. This prevents same-round role-conditioned update rewriting and limits role-conditioned exploitation after promotion.

Across three benchmark datasets, VTR-FL consistently improves selective recovery over reputation-based baselines while maintaining comparable task accuracy. For example, our method improves the re-entry probability of excluded clients by 37.5–43.8 percentage points over the reputation-based baselines on CIFAR-10. Our main contributions are as follows:

- We formalize selection-censored trust, which shows that permanently closing the observation channel makes recovery unidentifiable in the absence of another information source, and use Persistent Trust Observability to constrain the recovery channel.
- We propose VTR-FL that separates verification from aggregation. Public eligibility, committed verification propensities, evidence attestation, and challengeable state transitions preserve recovery opportunities. An auxiliary risk veto and post-commitment role blinding further constrain the restoration of model influence.
- We conduct extensive experiments and theoretical analyses to demonstrate that VTR-FL effectively recovers clients excluded by reputation policies while maintaining comparable accuracy.

2 Preliminaries

2.1 Federated Learning and Trust Observation

Let $\mathcal{N} = \{1, \ldots, N\}$ be the client set and w_t the global model at round t. Client i owns local dataset D_i and, after E local epochs, produces

$$\Delta_{i,t} = w_{i,t}^{E} - w_t. \tag{1}$$

Large-scale FL systems must additionally cope with partial participation, resource heterogeneity, and client drift (Kairouz and McMahan, 2021); representative system and optimization studies include production-scale FL, LEAF, FedProx, SCAFFOLD, FedOpt, FedNova, and FedDyn (Bonawitz et al., 2019; Caldas et al., 2018; Li et al., 2020; Karimireddy et al., 2020; Reddi et al., 2020; Wang et al., 2020; Acar et al., 2021). A scheduler therefore selects only a subset of clients in each round. Let $O_{i,t} \in \{0, 1\}$ indicate whether a valid protocol-certified evidence certificate is formed for client i at round t, and let $Y_{i,t}$ be the latent evidence that would be obtained under full evaluation. The visible evidence is

$$Y_{i,t}^{\text{obs}} = O_{i,t} Y_{i,t}, \qquad p_{i,t}^{v} = \Pr(O_{i,t} = 1 \mid \mathcal{H}_t), \qquad p_{i,t}^{\text{obs}} = \pi(B_{i,t}, C_t) = \pi(\phi(\mathcal{H}_t), C_t). \tag{2}$$

where $\mathcal{H}_t$ contains all system-visible history before the round-t observation decision, $p_{i,t}^{v}$ is the estimated probability that client i can be verified round t, and C_t denotes the current-round context.

2.2 Selection-Censored Trust

Definition 1 (Selection-Censored Trust)**.** *A system exhibits Selection-Censored Trust if its compressed trust state $B_{i,t} = \phi(\mathcal{H}_t)$ is recursively updated from previously observed evidence, while the future observation propensity $p_{i,t}^{\text{obs}} = \pi(B_{i,t}, C_t)$ depends on that same censored history.*

Theorem 1 (Recovery unidentifiability). *Suppose that from some time τ onward $p^v_{i,t} = 0$ almost surely for all $t \geq \tau$, and that post-τ client behavior can affect the recovery decision only through the certified evidence channel $Y^{\text{obs}}_{i,t}$. Then, conditional on the same visible history up to $\tau - 1$, a "persistently malicious" trajectory and a "recovered" trajectory induce the same future recovery-observation distribution.*

Proof. See Appendix A.1. If two trajectories coincide before τ but may differ afterward, and both have $O_{i,t} = 0$ after τ, they induce identical visible evidence $Y^{\text{obs}}_{i,t}$—whether the client remains malicious or has recovered. Hence any estimator or decision rule using only visible history receives identically distributed inputs under the two worlds and cannot consistently distinguish "still malicious" from "recovered but not re-observed." Thus recovery is unidentifiable.

Definition 2 (Persistent Trust Observability). *A recovery candidate satisfies PTO from time τ if*

$$\sum_{t=\tau}^{\infty} p^v_{i,t} = \infty. \tag{3}$$

Theorem 2 (Finite-window observation bound). *Let $W = \{\tau, \ldots, \tau + H - 1\}$ and suppose there exist predictable lower bounds $\underline{p}_{i,t} \in [0, 1]$ such that*

$$p^v_{i,t} \geq \underline{p}_{i,t} \qquad \text{a.s. for every history extending } \mathcal{H}_\tau. \tag{4}$$

Then

$$\mathbb{P}\left(\sum_{t \in W} O_{i,t} = 0 \mid \mathcal{H}_\tau\right) \leq \prod_{t \in W} (1 - \underline{p}_{i,t}) \leq \exp\left(-\sum_{t \in W} \underline{p}_{i,t}\right). \tag{5}$$

In particular, if $\underline{p}_{i,t} \geq \epsilon > 0$, the re-observation probability within H rounds is at least $1 - e^{-H\epsilon}$.

Proof. See Appendix A.2.

3 Method

VTR-FL follows the recovery path in Figure 1. A low-confidence client remains eligible for a verifiable observation channel instead of being permanently removed. Its committed update is evaluated to produce certified recovery evidence, but the Verify role has zero direct aggregation weight. Long-term trust summarizes history and supports observation scheduling, whereas recent recovery evidence determines whether the client can enter Probation and eventually return to Train. The same recovery path is bound by committed eligibility, propensity, evidence, and state digests so that coordinator-side manipulation can be challenged. Details of one communication round are summarized in Appendix B.

3.1 Budget-Feasible Verifiable Observation

Let $\mathcal{C}^v_t$ be the public verification-candidate set, $C_t = |\mathcal{C}^v_t|$, and $m^v_t \in (0, C_t]$ the expected verification budget. A fixed per-client floor becomes infeasible when C_t grows, so we allocate

$$\hat{\epsilon}_t = \min\left\{\epsilon_0, \kappa \frac{m^v_t}{C_t}\right\}, \qquad 0 < \kappa \leq 1. \tag{6}$$

Eligibility determines who must remain observable, whereas uncertainty, evidence age, and verification debt affect only the priority $s_{i,t} \geq 0$. The exact eligibility rule, verification-debt update, and priority function are specified in Appendix C.3. Let $S_t = \sum_{j \in \mathcal{C}^v_t} s_{j,t}$ and $b_t = m^v_t - C_t \hat{\epsilon}_t$. Before clipping, define

$$\tilde{p}^v_{i,t} = \begin{cases} \hat{\epsilon}_t + b_t\, s_{i,t}/S_t, & S_t > 0, \\ m^v_t / C_t, & S_t = 0, \end{cases} \qquad p^v_{i,t} = \min\{1, \tilde{p}^v_{i,t}\}. \tag{7}$$

Candidate flags and propensities are committed before unpredictable randomness selects Verify clients; a client that remains unobserved accumulates verification debt, which is cleared once a valid certificate is formed. In our implementation, certificate generation is deterministic after a valid audit. Details are provided in Appendix C.3, so the observation propensity equals the verification propensity:

$$p^{\text{obs}}_{i,t} = p^v_{i,t}. \tag{8}$$

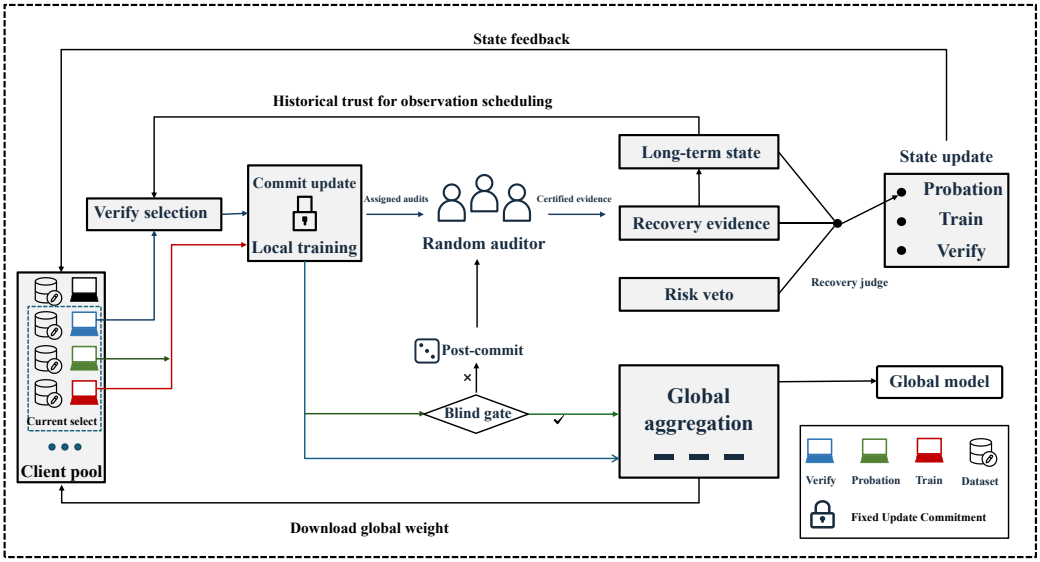


Figure 1: Framework of VTR-FL. Low-confidence clients remain observable through a verifiable Verify path. After local training, the update is committed before auditor selection and role assignment. Certified recovery evidence, a recent recovery gate, and an auxiliary risk veto determine state transitions among {*Verify, Probation, Train*}. Verify updates have zero direct aggregation weight, while Train and capped Probation updates may enter global aggregation.

Proposition 1 (Budget feasibility)**.** *For every t, $C_t\hat{\epsilon}_t \le \kappa m_t^v \le m_t^v$. Moreover, $\sum_{i\in\mathcal{C}_t^v} p_{i,t}^v \le m_t^v$, with equality whenever no coordinate is clipped at one. Every candidate also satisfies $p_{i,t}^v \ge \hat{\epsilon}_t$.*

Proof. See Appendix A.3.

Proposition 2 (PTO under persistent eligibility)**.** *Suppose a recovery client remains eligible in infinitely many rounds, $m_t^v \ge m_{\min} > 0$, and $C_t \le N$. Then on every such round*

$$p_{i,t}^{\mathrm{obs}} = p_{i,t}^v \ge \min\left\{\epsilon_0, \kappa\frac{m_{\min}}{N}\right\} > 0, \tag{9}$$

and therefore $\sum_t p_{i,t}^{\mathrm{obs}} = \infty$.

3.2 Certified Evidence and Propensity-Corrected History

After the client fixes an update commitment, a randomly selected auditor committee evaluates the same committed update. Auditor j returns a signed tuple $(v_{i,t}^{(j)}, \Delta L_{i,t}^{(j)}, a_{i,t}^{(j)})$. The protocol aggregates each scalar coordinate by the median; denote the resulting values by $v_{i,t}$, $\Delta L_{i,t}$, and $a_{i,t}$, details are shown in Appendix C.1 and Appendix C.4.

Proposition 3 (Median attestation under bounded honest error)**.** *For any scalar evidence coordinate with latent target $Y_{i,t}^*$, if more than half of the committee is honest and every honest report lies in $[Y_{i,t}^* - \varepsilon_a, Y_{i,t}^* + \varepsilon_a]$, then its median $\widetilde{Y}_{i,t}$ satisfies*

$$|\widetilde{Y}_{i,t} - Y_{i,t}^*| \le \varepsilon_a. \tag{10}$$

The protocol also records the verification propensity in (8). For the recovery evidence $y_{i,t}^{\mathrm{rec}}$ defined below, let $\tilde{p}_{i,t} = \max(p_{i,t}^v, p_{\min})$. We maintain

$$A_{i,t} = \lambda A_{i,t-1} + \frac{O_{i,t}}{\tilde{p}_{i,t}} y_{i,t}^{\mathrm{rec}}, \quad W_{i,t} = \lambda W_{i,t-1} + \frac{O_{i,t}}{\tilde{p}_{i,t}}, \quad Q_{i,t} = \lambda^2 Q_{i,t-1} + \left(\frac{O_{i,t}}{\tilde{p}_{i,t}}\right)^2 \tag{11}$$

which produce the long-term state and Kish-type effective evidence mass

$$\bar{\mu}_{i,t} = \frac{A_{i,t}}{W_{i,t} + \varepsilon_\mu}, \qquad n_{i,t} = \frac{W_{i,t}^2}{Q_{i,t} + \varepsilon_n}. \tag{12}$$

Unobserved rounds do not insert zero-valued evidence.

Assumption 1 (Conditional ignorability). *Before $y_{i,t}^{\mathrm{rec}}$ is evaluated, verification selection satisfies*

$$\mathbb{E}[O_{i,t} \mid \mathcal{H}_t, y_{i,t}^{\mathrm{rec}}] = p_{i,t}^{v}. \tag{13}$$

Theorem 3 (HT unbiasedness under conditional ignorability). *If $p_{i,t}^{v} > 0$ and the preceding assumption holds, then the unclipped Horvitz–Thompson term satisfies*

$$\begin{aligned}\mathbb{E}\left[\frac{O_{i,t}y_{i,t}^{\mathrm{rec}}}{p_{i,t}^{v}}\middle|\mathcal{H}_t\right] &= \mathbb{E}\left[\frac{y_{i,t}^{\mathrm{rec}}}{p_{i,t}^{v}}\mathbb{E}[O_{i,t} \mid \mathcal{H}_t, y_{i,t}^{\mathrm{rec}}]\middle|\mathcal{H}_t\right] \\ &= \mathbb{E}[y_{i,t}^{\mathrm{rec}} \mid \mathcal{H}_t].\end{aligned} \tag{14}$$

The normalized ratio, temporal decay, and probability clipping used in practice introduce finite-sample bias; Theorem 3 motivates propensity correction at the unclipped HT-term level; the normalized ratio, temporal decay, and probability clipping used in practice define a stabilized finite-sample estimator.

3.3 Recovery-Specific Evidence and Dual-Time-Scale Gating

Recovery and risk are modeled separately. Let $L_t^{\mathrm{val}}(\Delta)$ be the authenticated validation loss obtained by applying update Δ to the current model. And define

$$\Delta L_{i,t} = L_t^{\mathrm{val}}(\Delta_{i,t}) - L_t^{\mathrm{val}}(\mathbf{0}). \tag{15}$$

Thus smaller $\Delta L_{i,t}$ indicates a more favorable audit response. We use

$$q_{i,t} = \mathrm{RankNorm}_t(-\Delta L_{i,t}), \qquad y_{i,t}^{\mathrm{rec}} = \omega_v v_{i,t} + \omega_q q_{i,t}, \tag{16}$$

where $\omega_v, \omega_q \geq 0$ and $\omega_v + \omega_q = 1$. With anomaly score $a_{i,t}$, an auxiliary risk channel is

$$r_{i,t} = \eta_a \,\mathrm{RankNorm}_t(a_{i,t}) + \eta_v \,\mathrm{RankNorm}_t(1 - v_{i,t}), \tag{17}$$

where $\eta_a, \eta_v \geq 0$ and $\eta_a + \eta_v = 1$. The risk channel serves specifically as a high-risk transition veto, separating promotion control from general malicious-client detection. Here $a_{i,t}$ is a cosine-deviation anomaly score; its exact construction and veto rule are provided in Appendix C.2.

Let $y_{i,1}^{\mathrm{rec}}, y_{i,2}^{\mathrm{rec}}, \ldots$ be the time-ordered sequence of valid certified recovery observations for client i. For $k \geq 2H$, define the average level of recent recovery evidence μ and recent trend d as:

$$\hat{\mu}_{i,k} = \frac{1}{H}\sum_{h=0}^{H-1} y_{i,k-h}^{\mathrm{rec}}, \qquad d_{i,k} = \hat{\mu}_{i,k} - \frac{1}{H}\sum_{h=H}^{2H-1} y_{i,k-h}^{\mathrm{rec}}. \tag{18}$$

Missing rounds are omitted rather than imputed with zeros. With effective evidence mass $n_{i,t}$, the transition rules are

$$\begin{aligned} V \to P: &\quad \hat{\mu} \geq \theta_P,\ n \geq n_P,\ d \geq \delta_P,\ r \leq \tau_P, \\ P \to T: &\quad \hat{\mu} \geq \theta_T,\ n \geq n_T,\ d \geq \delta_T,\ r \leq \tau_T, \end{aligned} \tag{19}$$

where $\theta_T \geq \theta_P$, $n_T \geq n_P$, $\delta_T \geq \delta_P$, and $\tau_T \leq \tau_P$. Numerical values belong to the experimental configuration and are reported in Section 4.1. The gate is evaluated only once the required recent evidence windows exist.

3.4 Zero-Influence Verification and Blind Probation

Let S_t^T, S_t^V, and S_t^P denote the roles fixed at the start of round t. Define

$$\chi_{i,t} = \begin{cases} 1, & i \in S_t^T, \\ 0, & i \in S_t^V, \\ G_{i,t}, & i \in S_t^P, \end{cases} \qquad D_t = \sum_j \chi_{j,t}\bar{\alpha}_{j,t}. \tag{20}$$

The normalized weight is

$$\alpha_{i,t} = \begin{cases} \chi_{i,t}\bar{\alpha}_{i,t}/D_t, & D_t > 0, \\ 0, & D_t = 0, \end{cases} \qquad w_{t+1} = w_t + \sum_i \alpha_{i,t}\Delta_{i,t}. \tag{21}$$

Theorem 4 (Direct model-influence isolation)**.** *Every Verify client has exactly zero direct contribution to the current global update.*

For Verify clients, the aggregation coefficient satisfies $\alpha_{i,t} = 0$, Therefore, evidence acquisition does not directly enlarge the current-round attack surface. The corresponding proof is provided in Appendix A.4.

A Probation client first chooses a fresh nonce $u_{i,t}$ and commits

$$C^{\Delta}_{i,t} = H(\mathrm{enc}(\Delta_{i,t}) \| u_{i,t} \| i \| t), \tag{22}$$

where enc is a canonical serialization. Only after the commitment is fixed does unpredictable randomness generate $G_{i,t} \in \{0, \lambda_p\}$.

Proposition 4 (Same-round role non-adaptivity)**.** *Assume the commitment is computationally binding, $G_{i,t}$ is unpredictable before commitment, and the protocol accepts only an opening of $C^{\Delta}_{i,t}$. Then, except with the binding advantage of the adversary, the submitted update cannot be chosen as two different functions of the two possible same-round roles.*

The proof provided in Appendix A.5 establishes same-round role non-adaptivity; Section 4.2 separately stress-tests longer-horizon recovery behavior.

3.5 Challengeable State Transitions

Each round commits the previous state, public eligibility inputs, candidate set, observation propensities, signed evidence certificates, updated state, and authenticated randomness used for selection. Let $\mathcal{R}$ be the deterministic public transition relation checked by the verifier. Fraud proofs cover candidate omission, propensity tampering, selection inconsistency, state rollback, and unauthorized promotion.

Theorem 5 (Conditional auditability)**.** *Assume (i) collision-resistant and binding authenticated commitments, (ii) unforgeable auditor signatures, (iii) authenticated non-rewritable selection randomness, (iv) a ledger satisfying its Byzantine finality threshold, and (v) a deterministic verifier that accepts exactly transitions satisfying $\mathcal{R}$. Then an invalid challenged transition can be accepted only if at least one of these assumptions fails.*

The proof is in Appendix A.6.

4 Experiments

4.1 Experiment Setup

Datasets. We evaluate VTR-FL on CIFAR-10, Fashion-MNIST, and MNIST. CIFAR-10 is used for the independent confirmation study, while Fashion-MNIST and MNIST provide cross-dataset validation with a different model and client population.

Compared methods. We compare with four representative reputation-based selection strategies. *Hard Threshold* permanently excludes low-reputation clients and therefore serves as the censorship baseline. *MAB-RFL* uses exploration–exploitation for adaptive client selection (Wan et al., 2022); *FairFedCS* provides soft reputation redemption (Shi et al., 2023); and *RSVDUCBTE* combines robust selection with TTL-style recovery (Chuang and Zhang, 2026). For adaptive attacks, we further compare VTR-FL with a role-known probation variant. We also ablate the recovery gate by replacing the proposed recent level-and-trend statistic with long-term trust or a recent level alone.

Implementation configuration. Client data are partitioned by a Dirichlet distribution. Unless otherwise stated, recovery experiments follow the identity-fixed temporary-label-noise protocol detailed in Appendix D.1. Configurations on CIFAR-10 datasets are {*Model: ResNet-18 with GroupNorm, Clients number: 50, Communication rounds: 200, Train clients/per round: 5, Local train epochs: 3, Batch size: 64, Learning rate: 0.05, Momentum: 0.9, Weight decay:* 5×10^{-4}, *non-IID degree* α*: 0.1,0.5,1.0*}. Configurations on Fashion-MNIST and MNIST datasets are {*Model: TwoLayerCNN, Clients number: 40, Communication rounds: 200, Train clients/per round: 4, Local train epochs: 2, Batch size: 64, Learning rate: 0.01, Momentum: 0.9, Weight decay:* 5×10^{-4},

non-IID degree α: 0.1,0.5}. Baseline comparisons adopted the same setting. Other parameters are set as: $\omega_v = \omega_q = 0.5$, $\eta_a = \eta_v = 0.5$, $H = 5$, $(\theta_P, \theta_T) = (0.22, 0.30)$, $(n_P, n_T) = (1.5, 2.0)$, $(\delta_P, \delta_T) = (0, 0.02)$, and $\tau_R = 0.85$.

We report AUC_{RD} for recovered-vs.-degraded evidence ranking and $\Delta_R = P(\text{Re} \mid \text{Rec}) - P(\text{Re} \mid \text{Deg})$ for re-entry selectivity. ActiveAttackExposure measures aggregation mass assigned while an attack is active,

$$AA(W) = \sum_{t \in W} \sum_{i} A_{i,t}^{\text{attack}} \alpha_{i,t}$$

whereas RecoveredCohortInfluence (RCI) is the post-attack aggregation mass of the formerly degraded cohort and is therefore a recovery measure rather than unsafe exposure.

4.2 Main Results

Hard exclusion creates persistent evidence censorship. Under Hard Threshold, post-degradation observation collapses: The fraction of target clients that obtain at least one valid post-degradation evidence certificate within the next 40 rounds (Coverage@40) is 0.000, and recovered versus persistently degraded clients are indistinguishable from system-visible evidence ($\text{AUC}_{RD} = 0.500$). Restoring observation alone improves visibility but is insufficient for recovery. The observation-only variant increases Coverage@40 to 0.625 and AUC_{RD} to 0.647, yet its re-entry recall remains 0.125, slightly below MAB-RFL (0.150), with almost no clients reaching Probation. These results separate two requirements: persistent observation is necessary to recover evidence, while recovery-responsive state transitions are required to restore model participation.

Recent recovery dynamics are critical for selective re-entry. Figure 2(a) compares three gates on identical calibration trajectories. Long-term trust provides almost no re-entry selectivity ($\Delta_R = 0.017$). Using a recent evidence level increases Δ_R to 0.092, while adding the recent trend increases it to 0.317. This large gap explains why VTR-FL separates long-term historical trust from the statistic used for re-entry.

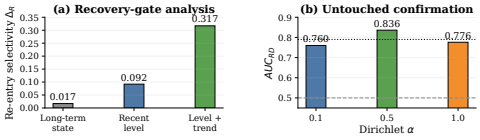


Figure 2: Recovery analysis and independent confirmation. (a) Re-entry selectivity for three recovery gates. (b) Recovered-vs.-degraded evidence AUC on untouched CIFAR-10 confirmation runs; dashed and dotted lines denote chance and the overall AUC.

We evaluate 24 untouched CIFAR-10 cells from 8 seeds and 3 non-IID levels. Figure 2(b) reports mean AUC_{RD} values of 0.760, 0.836, and 0.776 for $\alpha = 0.1, 0.5, 1.0$, respectively; the seed-weighted overall AUC is 0.790, and all 24 cells are above chance. The state machine obtains $\Delta_R = 0.592$, $P(\text{Re} \mid \text{Rec}) = 0.908$, and $P(\text{Re} \mid \text{Deg}) = 0.317$. Thus, the frozen mechanism preferentially restores recovered clients, yielding $\Delta R = 0.592$.

VTR-FL consistently improves recovery across datasets. We next conduct a five-method comparison on CIFAR-10, Fashion-MNIST, and MNIST. Figure 3 shows a consistent pattern: VTR-FL has the highest re-entry probability on every dataset and restores substantially more Train influence to the recovered cohort. On CIFAR-10, the paired gains in $P(Re)$ over MAB-RFL, FairFedCS, and RSV-DUCBTE are +0.438, +0.375, and +0.425, while classification accuracy remains comparable. This isolates the gain to VTR-FL's intended recovery objective. Full results are reported in Appendix D.2.

Recovery restores legal influence without reopening the Verify channel. Across the 24 CIFAR-10 runs, ActiveAttackExposure is 2.60 during the degradation window and exactly zero after the temporary attack ends. The recovered cohort subsequently receives total RCI 7.23, corresponding to 18.1% of recovery-window aggregation mass, including $RCI^{train} = 6.53$. Hence the recovered cohort regains actual model influence rather than merely changing a state label, while Verify itself remains zero-influence by construction. At the protocol layer, candidate omission, propensity tampering, state rollback, and unauthorized promotion each result in 0/1000 illegal acceptances under the complete commitment/challenge path.

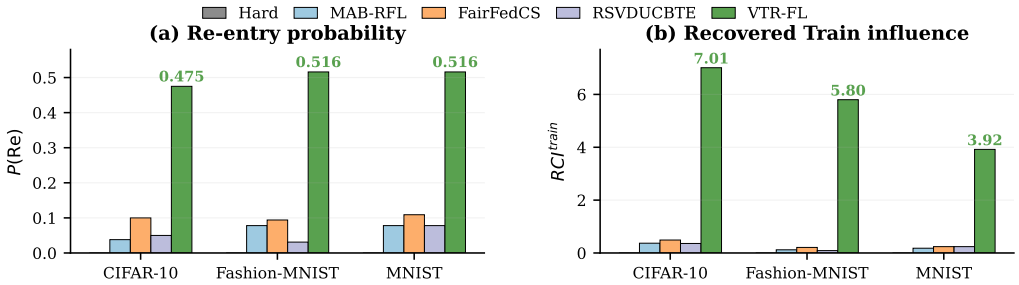


Figure 3: Cross-dataset recovery comparison. (a) Re-entry probability and (b) recovered Train influence. VTR-FL preserves the same recovery advantage across CIFAR-10, Fashion-MNIST, and MNIST.

Blind probation preserves recovery while blocking tested role-conditioned exploitation. We stress-test the recovery gate with a schedule-aware attacker that deliberately creates a low-evidence window, submits benign updates to manufacture a positive recovery trend d, and attacks after promotion. As summarized in Figure 4(a), on CIFAR-10, VTR-FL and role-known probation obtain similar re-entry probabilities (0.575 and 0.525), showing that promotion remains available under role blinding. However, Figure 4(b) shows that post-promotion ActiveAttackExposure is 0 under blind VTR-FL and 22.65 under role-known probation. These results isolate the role of post-commit blinding: it removes same-round role-conditioned exploitation while preserving the recovery path.

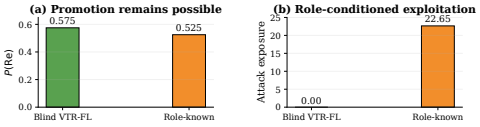


Figure 4: Adaptive momentum-gaming stress test. Blinding preserves promotion (left) while removing post-promotion attack exposure for the tested role-conditioned attacker (right).

Adaptive poisoning recovery attack. To evaluate whether adversaries can exploit the Verify–Probation–Train transition, we construct an adaptive recovery attacker with three phases. During the initial attack phase, malicious clients perform model poisoning attacks. After being downgraded, they temporarily submit benign-looking updates to increase recent recovery evidence and attempt to satisfy the level-and-trend gate. After promotion, they resume malicious updates. We compare VTR-FL with reputation-based baselines. The evaluation reports recovered influence (RCI), attack success rate (ASR), and false promotion rate (FPR). The MNIST results over 40 independent runs are shown in Table 1, which demonstrate that among methods that restore nonzero influence, VTR-FL achieves the lowest mean FPR (0.043) and ASR (0.139) while delivering the largest RCI (4.941).

Table 1: Adaptive poisoning recovery attack. Bold numbers are the best numerical results, and underlined numbers are suboptimal.

Method	FPR ↓	RCI ↑	ASR ↓
Hard Threshold	**0.00**	0.00	**0.00**
MAB-RFL	0.167±0.356	0.143±0.267	0.500±0.535
FairFedCS	0.167±0.236	<u>0.623±1.131</u>	0.375±0.518
RSVDUCBTE	0.146±0.208	0.080±0.150	0.792±0.393
VTR-FL	<u>0.043±0.036</u>	**4.941±0.362**	<u>0.139±0.350</u>

Verification overhead and scalability. Finally, we quantify the runtime and communication cost of verifiable recovery on CIFAR-10. We vary the verification budget m_v and measure runtime overhead (RO), communication overhead (CO), and verification throughput (Throughput). Since VTR-FL separates evidence collection from aggregation, we expect verification overhead to scale primarily with the verification budget rather than the full client population; the results in Table 2 show that across the evaluated verification budgets, runtime and communication costs are at most $1.20\times$ and $1.33\times$, respectively, those under $m_v = 2$.

5 Related Work

Client selection, reputation, and robustness. FedCS, Oort, and biased/Power-of-Choice selection optimize which clients participate under resource or utility constraints (Nishio and Yonetani, 2019; Lai et al., 2021; Cho et al., 2022); AFL, q-FFL, and FedBalancer add fairness

or system-efficiency objectives (Mohri et al., 2019; Li et al., 2019; Shin et al., 2022). More directly related to recovery, FairFedCS enables soft reputation redemption, MAB-RFL couples exploration with robust selection, TFL-DT models dynamic trust, and RSVDUCBTE introduces TTL-style recovery (Shi et al., 2023; Wan et al., 2022; Guo et al., 2023; Chuang and Zhang, 2026). Byzantine defenses such as Krum/Bulyan, robust statistics, geometric-median aggregation, sybil-aware weighting, and FLTrust mainly protect the update path (Blanchard et al., 2017; El-Mhamdi et al., 2018; Yin et al., 2018; Pillutla et al., 2022; Fung et al., 2020; Cao et al., 2020). VTR-FL instead treats future evidence visibility and re-entry as the primary security object.

Table 2: Verification overhead analysis

Budget	RO(hour)	CO	Throughput
$m_v = 2$	$10.6h$	85.7GB	0.0052 rounds/s
$m_v = 5$	$12.0h$	103.8GB	0.0046 rounds/s
$m_v = 10$	$12.7h$	112.6GB	0.0044 rounds/s
$m_v = 20$	$10.8h$	113.7GB	0.0051 rounds/s

Selective feedback and verifiable recovery. SCT is conceptually related to selective labels and one-sided feedback (Wei, 2021; Jiang et al., 2021); its propensity correction builds on unequal-probability sampling (Horvitz and Thompson, 1952; Hájek, 1964). Blockchain-assisted FL has separately explored verifiable client selection and auditable participant selection (Nguyen et al., 2022; Zeng et al., 2024; Jin et al., 2024). VTR-FL connects these ideas by binding candidate eligibility, observation propensity, certified evidence, and recovery-state transition into one challengeable chain. Broader FL security surveys provide context for adaptive attacks and privacy threats (Mothukuri et al., 2021; Lewis et al., 2023).

6 Discussion

Trust Recovery Beyond Reputation Filtering. Traditional reputation mechanisms couple historical trust estimation with participation control. This coupling creates a feedback loop where low reputation reduces future evidence opportunities. VTR-FL complements reputation by separating trust observation from influence restoration. Reputation summarizes historical behavior, while certified recovery evidence determines whether influence can be progressively restored.

Evidence-Conditioned Trust Restoration. VTR-FL treats recovery as evidence-conditioned restoration of influence: reputation summarizes current evidence, while continued verification makes subsequent influence restoration accountable to behavioral change. Similar to other dynamic trust systems, reputation reflects current evidence rather than future certainty.

Recovery Opportunity and Attack Exposure. Hard exclusion minimizes immediate attack exposure but permanently removes recovery opportunities. Unlimited re-entry improves responsiveness but directly enlarges the poisoning surface. VTR-FL introduces an intermediate design by allowing evidence collection through zero-influence verification and restoring influence only after additional validation.

Limitations and Future Extensions. Several directions remain open. First, patient attackers that gradually shape long-term evidence require stronger sequential detection mechanisms. Second, stronger Byzantine auditor models and decentralized verification committees can further strengthen evidence generation. Third, large-scale deployment in cross-silo and foundation-model federated learning remains an important future direction.

7 Conclusion

In this paper, we propose a VTR-FL framework to address Selection-Censored Trust by preserving a verifiable zero-influence observation path and separating historical trust from recovery-specific evidence. Independent confirmation and three-dataset evaluation support consistent selective recovery, while adaptive tests verify the intended protection against same-round role-conditioned exploitation. Empirical validation across various datasets establishes VTR-FL as a verifiable recovery mechanism for SCT that restores recovered clients without granting Verify-stage updates direct model influence.

REFERENCES

Durmus Alp Emre Acar, Yue Zhao, Ramon Matas Navarro, Matthew Mattina, Paul N Whatmough, and Venkatesh Saligrama. Federated learning based on dynamic regularization. *arXiv preprint arXiv:2111.04263*, 2021.

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

Keith Bonawitz, Hubert Eichner, Wolfgang Grieskamp, Dzmitry Huba, Alex Ingerman, Vladimir Ivanov, Chloe Kiddon, Jakub Konečnỳ, Stefano Mazzocchi, Brendan McMahan, et al. Towards federated learning at scale: System design. *Proceedings of machine learning and systems*, 1: 374–388, 2019.

Sebastian Caldas, Sai Meher Karthik Duddu, Peter Wu, Tian Li, Jakub Konečnỳ, H Brendan McMahan, Virginia Smith, and Ameet Talwalkar. Leaf: A benchmark for federated settings. *arXiv preprint arXiv:1812.01097*, 2018.

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

Yae Jee Cho, Jianyu Wang, and Gauri Joshi. Towards understanding biased client selection in federated learning. In *International conference on artificial intelligence and statistics*, pages 10351–10375. PMLR, 2022.

Yung-Ting Chuang and Qi-Xiang Zhang. Adaptive aggregation and robust client selection against poisoning and backdoor attacks in federated learning. *Engineering Applications of Artificial Intelligence*, 177:114734, 2026.

El-Mahdi El-Mhamdi, Rachid Guerraoui, and Sébastien Rouault. The hidden vulnerability of distributed learning in byzantium. In *International conference on machine learning*, pages 3521–3530. PMLR, 2018.

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)*, pages 301–316, 2020.

Jingjing Guo, Zhiquan Liu, Siyi Tian, Feiran Huang, Jiaxing Li, Xinghua Li, Kostromitin Konstantin Igorevich, and Jianfeng Ma. Tfl-dt: A trust evaluation scheme for federated learning in digital twin for mobile networks. *IEEE Journal on Selected Areas in Communications*, 41(11):3548–3560, 2023.

Jaroslav Hájek. Asymptotic theory of rejective sampling with varying probabilities from a finite population. *The Annals of Mathematical Statistics*, 35(4):1491–1523, 1964.

Daniel G Horvitz and Donovan J Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260):663–685, 1952.

Heinrich Jiang, Qijia Jiang, and Aldo Pacchiano. Learning the truth from only one side of the story. In *International Conference on Artificial Intelligence and Statistics*, pages 2413–2421. PMLR, 2021.

Xiaojun Jin, Chao Ma, Song Luo, Pengyi Zeng, and Yifei Wei. Two-stage client selection scheme for blockchain-enabled federated learning in iot. *Computers, Materials and Continua*, 81(2): 2317–2336, 2024.

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.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *International conference on machine learning*, pages 5132–5143. PMLR, 2020.

Fan Lai, Xiangfeng Zhu, Harsha V Madhyastha, and Mosharaf Chowdhury. Oort: Efficient federated learning via guided participant selection. In *15th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 21)*, pages 19–35, 2021.

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.

Tian Li, Maziar Sanjabi, Ahmad Beirami, and Virginia Smith. Fair resource allocation in federated learning. *arXiv preprint arXiv:1905.10497*, 2019.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2:429–450, 2020.

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*, pages 1273–1282. Pmlr, 2017.

Mehryar Mohri, Gary Sivek, and Ananda Theertha Suresh. Agnostic federated learning. In *International conference on machine learning*, pages 4615–4625. PMLR, 2019.

Viraaji Mothukuri, Reza M Parizi, Seyedamin Pouriyeh, Yan Huang, Ali Dehghantanha, and Gautam Srivastava. A survey on security and privacy of federated learning. *Future Generation Computer Systems*, 115:619–640, 2021.

Truc Nguyen, Phuc Thai, Jeter Tre'R, Thang N Dinh, and My T Thai. Blockchain-based secure client selection in federated learning. In *2022 IEEE International Conference on Blockchain and Cryptocurrency (ICBC)*, pages 1–9. IEEE, 2022.

Takayuki Nishio and Ryo Yonetani. Client selection for federated learning with heterogeneous resources in mobile edge. In *ICC 2019-2019 IEEE international conference on communications (ICC)*, pages 1–7. IEEE, 2019.

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

Sashank Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konečnỳ, Sanjiv Kumar, and H Brendan McMahan. Adaptive federated optimization. *arXiv preprint arXiv:2003.00295*, 2020.

Yuxin Shi, Zelei Liu, Zhuan Shi, and Han Yu. Fairness-aware client selection for federated learning. In *2023 IEEE international conference on multimedia and expo (ICME)*, pages 324–329. IEEE, 2023.

Jaemin Shin, Yuanchun Li, Yunxin Liu, and Sung-Ju Lee. Fedbalancer: Data and pace control for efficient federated learning on heterogeneous clients. In *Proceedings of the 20th Annual International Conference on Mobile Systems, Applications and Services*, pages 436–449, 2022.

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.

Jianyu Wang, Qinghua Liu, Hao Liang, Gauri Joshi, and H Vincent Poor. Tackling the objective inconsistency problem in heterogeneous federated optimization. *Advances in neural information processing systems*, 33:7611–7623, 2020.

Dennis Wei. Decision-making under selective labels: Optimal finite-domain policies and beyond. In *International conference on machine learning*, pages 11035–11046. PMLR, 2021.

Dong Yin, Yudong Chen, Ramchandran Kannan, and Peter Bartlett. Byzantine-robust distributed learning: Towards optimal statistical rates. In *International conference on machine learning*, pages 5650–5659. Pmlr, 2018.

Huang Zeng, Mingtian Zhang, Tengfei Liu, and Anjia Yang. A federated learning framework with blockchain-based auditable participant selection. *Computers, Materials & Continua*, 79(3): 5125–5142, 2024.

REPRODUCIBILITY STATEMENT

Calibration, hold-out, cross-dataset validation, and confirmation use disjoint seeds. Section 4.1 reports the hyperparameters used in the final experiments; the appendix contains full proofs for the theoretical claims. Code and experiment manifests are intended to be released with the final artifact on GitHub.

AI USE STATEMENT

Generative AI tools were used for language polishing (GPT-5.6 Sol) and document organization (GPT-6 Astra). We also used Consensus to assist with related-work discovery and Cursor for implementation support. All technical claims, mathematical definitions, numerical results, and interpretations were independently checked against the experiment logs and code.

A PROOF

The appendix follows the same order as the theorem and proposition statements in the main text.

A.1 RECOVERY UNIDENTIFIABILITY

Let P_0 denote the latent world in which client i remains malicious after τ and P_1 the world in which it recovers. By assumption, both worlds induce the same visible history up to $\tau - 1$. For every $t \geq \tau$,

$$\begin{aligned} \mathbb{P}_j(O_{i,t} = 1) &= \mathbb{E}_j[\mathbb{P}_j(O_{i,t} = 1 \mid \mathcal{H}_t)] \\ &= \mathbb{E}_j[p^v_{i,t}] = 0, \qquad j \in \{0, 1\}, \end{aligned} \tag{23}$$

which implies $O_{i,t} = 0$ a.s. Hence

$$Y^{\text{obs}}_{i,t} = O_{i,t} Y_{i,t} = 0 \quad \text{a.s. under both } \mathsf{P}_0, \mathsf{P}_1. \tag{24}$$

Therefore, for every measurable suffix event A,

$$\mathbb{P}_0\big(Y^{\text{obs}}_{i,\tau:\infty} \in A \mid \mathcal{H}_\tau\big) = \mathbb{P}_1\big(Y^{\text{obs}}_{i,\tau:\infty} \in A \mid \mathcal{H}_\tau\big)\,. \tag{25}$$

For any decision rule ϕ measurable with respect to the visible history, the induced distributions satisfy

$$\mathcal{L}_{\mathsf{P}_0}(\phi) = \mathcal{L}_{\mathsf{P}_1}(\phi), \tag{26}$$

so no such rule can consistently distinguish the two latent recovery states.

A.2 FINITE-WINDOW OBSERVATION BOUND

Let $E_t = \{O_{i,\tau} = \cdots = O_{i,t} = 0\}$. For every $t \in W$, the assumed lower bound gives $1 - p^v_{i,t} \leq 1 - \underline{p}_{i,t}$ almost surely. Therefore

$$\begin{aligned} \mathbb{P}(E_t \mid \mathcal{H}_\tau) &= \mathbb{E}\big[\mathbf{1}_{E_{t-1}} \mathbb{P}(O_{i,t} = 0 \mid \mathcal{H}_t) \mid \mathcal{H}_\tau\big] \\ &= \mathbb{E}\big[\mathbf{1}_{E_{t-1}} (1 - p^v_{i,t}) \mid \mathcal{H}_\tau\big] \\ &\leq (1 - \underline{p}_{i,t}) \mathbb{P}(E_{t-1} \mid \mathcal{H}_\tau). \end{aligned} \tag{27}$$

Recursive application yields

$$\mathbb{P}(E_{\tau+H-1} \mid \mathcal{H}_\tau) \leq \prod_{t \in W} (1 - \underline{p}_{i,t}) \leq \exp\left(- \sum_{t \in W} \underline{p}_{i,t} \right). \tag{28}$$

This proof requires no independence across rounds; it only requires history-uniform lower bounds on the conditional observation probabilities.

A.3 BUDGET FEASIBILITY

By definition,

$$C_t \hat{\epsilon}_t \leq C_t \kappa \frac{m^v_t}{C_t} = \kappa m^v_t \leq m^v_t, \tag{29}$$

so $b_t = m^v_t - C_t \hat{\epsilon}_t \geq 0$. If $S_t > 0$, the unclipped allocation satisfies

$$\begin{aligned} \sum_{i \in \mathcal{C}^v_t} \tilde{p}^v_{i,t} &= C_t \hat{\epsilon}_t + b_t \sum_i \frac{s_{i,t}}{S_t} \\ &= C_t \hat{\epsilon}_t + b_t = m^v_t. \end{aligned} \tag{30}$$

If $S_t = 0$, $\tilde{p}^v_{i,t} = m^v_t / C_t$ and the same identity holds. Since $p^v_{i,t} = \min\{1, \tilde{p}^v_{i,t}\}$,

$$\sum_i p^v_{i,t} \leq \sum_i \tilde{p}^v_{i,t} = m^v_t. \tag{31}$$

Finally, $\tilde{p}^v_{i,t} \geq \hat{\epsilon}_t$ in both cases and $\hat{\epsilon}_t \leq 1$, hence clipping cannot violate the floor.

A.4 Zero-Influence Verification

For $i \in S_t^V$, $\chi_{i,t} = 0$. Hence

$$\alpha_{i,t} = \frac{0 \cdot \bar{\alpha}_{i,t}}{\sum_j \chi_{j,t}\bar{\alpha}_{j,t} + \varepsilon_\alpha} = 0, \tag{32}$$

and its current-round contribution satisfies

$$\alpha_{i,t}\Delta_{i,t} = \mathbf{0} \tag{33}$$

for any submitted update $\Delta_{i,t}$.

A.5 Post-Commit Role Non-Adaptivity

Let $\mathsf{Com}(\Delta; u)$ be computationally binding and let $G_{i,t}$ be sampled only after $C_{i,t}^{\Delta} = \mathsf{Com}(\Delta_{i,t}; u_{i,t})$ is fixed. Suppose the adversary can open the same commitment to two role-conditioned updates $\Delta^{(0)} \neq \Delta^{(1)}$. Then

$$\mathsf{Com}(\Delta^{(0)}; u_0) = C_{i,t}^{\Delta} = \mathsf{Com}(\Delta^{(1)}; u_1), \tag{34}$$

which directly yields a binding adversary. Hence

$$\mathbb{P}(\text{two valid role-conditioned openings}) \leq \mathrm{Adv}_{\mathsf{Com}}^{\mathrm{bind}}. \tag{35}$$

The argument fixes the content of the update, not availability: a client may still abort after seeing its role unless a separate timeout or penalty rule is enforced.

A.6 Conditional Auditability

Let $\mathcal{R}$ be the deterministic public relation defining a valid round transition from committed prior state, public protocol inputs, authenticated randomness, and signed evidence. Let E_{bad} be the event that a challenged object violating $\mathcal{R}$ is nevertheless accepted. If E_B, E_S, E_R, E_L denote failures of commitment/root binding, signature unforgeability, randomness authenticity, and ledger finality, respectively, then verifier soundness gives

$$E_{\mathrm{bad}} \subseteq E_B \cup E_S \cup E_R \cup E_L. \tag{36}$$

Therefore

$$\mathbb{P}(E_{\mathrm{bad}}) \leq \varepsilon_B + \varepsilon_S + \varepsilon_R + \varepsilon_L. \tag{37}$$

Thus the theorem is conditional on the explicit cryptographic and ledger assumptions; it is not an unconditional statement about arbitrary coordinator behavior.

B Algorithm

The notation deliberately separates the long-term state $\bar{\mu}$ from the recent recovery state $\hat{\mu}$, since they serve different roles in the protocol.

Algorithm 1 One communication round of VTR-FL: verifiable recovery and controlled re-entry

Require: global model w_t; current states $\{\mathcal{S}_{i,t}\}_{i=1}^N$; Verify budget m_t^v
Ensure: new global model w_{t+1}; new states $\{\mathcal{S}_{i,t+1}\}_{i=1}^N$
1: compute and commit $\mathcal{C}_t^v$ and $\{p_{i,t}^v\}$; sample S_t^V
2: form current Train and Probation sets S_t^T, S_t^P from $\mathcal{S}_{i,t}$
3: **for** $i \in S_t^T \cup S_t^V \cup S_t^P$ **do**
4: train locally, obtain $\Delta_{i,t}$, and submit binding commitment $C_{i,t}^{\Delta}$
5: **end for**
6: after commitments, reveal probabilistic Probation roles $G_{i,t}$ and assign auditor committees
7: generate signed auditor reports and compute $y_{i,t}^{\mathrm{rec}}$ and $R_{i,t}^{\mathrm{veto}}$
8: set $\chi = 0$ for Verify, $\chi = 1$ for Train, and $\chi = G$ for Probation; aggregate w_{t+1}
9: update $\bar{\mu}, n, \hat{\mu}, d$ from valid evidence and compute next states $\mathcal{S}_{i,t+1}$
10: commit evidence/state roots and process challenges
11: Output new model and states

If the trust state dimension and the auditor committee size are treated as constants, the per-round server computational complexity outside local training is $O(|C_{\mathrm{ver}}^t| + m_t k)$, where m_t is the number of all participating clients in that round; the online client state storage is $O(N)$. In each round, only a constant number of roots and protocol metadata are stored on-chain, while the full evidence and model remain off-chain; therefore, the on-chain active state does not grow with the model dimension.

C Implementation Details and Assumptions

C.1 Validation Evidence

Before client partitioning, 5% of the official training set is randomly reserved as an audit pool and split into five mutually disjoint auditor shards. The remaining samples are distributed to clients by the Dirichlet partition; the test set is used only for task accuracy. For a committed update $\Delta_{i,t}$, an auditor evaluates $w_t + \Delta_{i,t}$ on its own shard. The validation signal $v_{i,t}$ is balanced accuracy, while a loss-improvement quantity derived from the global-model and candidate-model losses provides the quality signal used in $y_{i,t}^{\mathrm{rec}}$ after round-wise rank normalization.

In the prototype, auditor evaluation is executed by the simulator and committed updates are inspected in plaintext. Privacy-preserving validation is not instantiated and is orthogonal to this work. Certified evidence refers to evidence bound to the update commitment and signed auditor reports.

C.2 Risk Score and Veto

For the updates audited in round t, let u_t^{ref} be their coordinate-wise median. The anomaly score is

$$a_{i,t} = \mathrm{clip}_{[0,1]}\left[\frac{1 - \cos(\mathrm{vec}(\Delta_{i,t}), u_t^{\mathrm{ref}})}{2}\right].$$

$\mathrm{RankNorm}_t$ denotes percentile ranking among candidates audited in the same round. The frozen risk score is $r_{i,t} = 0.5\ \mathrm{RankNorm}_t(a_{i,t}) + 0.5\ \mathrm{RankNorm}_t(1 - v_{i,t})$.

Promotion requires $\max\{R_{i,t}, R_{i,t}^{\mathrm{EMA}}, R_{i,t}^{\mathrm{peak}}\} \le 0.85$, where the EMA decay is 0.8 and R^{peak} is the maximum over the five most recent valid observations. The frozen recovery thresholds are $(\theta_P, \theta_T) = (0.22, 0.30)$, $(n_P, n_T) = (1.5, 2.0)$, and $(\delta_P, \delta_T) = (0, 0.02)$.

C.3 Verification Scheduling

Let $d_{i,t}^{\mathrm{ver}}$ denote verification debt. A valid certificate resets the debt to zero; an unobserved client outside Train and Probation increments it by one, whereas Train and Probation clients do not accumulate verification debt. Observation age is reset after a valid certificate and otherwise increases by one.

Clients outside Train and Probation become Verify candidates when observation age or debt reaches $F = 10$. Their scheduling priority is

$$s_{i,t} = w_d \min(d_{i,t}^{\mathrm{ver}}/10, 1) + w_u \min(U_{i,t}, 1) + w_f \min(f_{i,t}/10, 1) + w_b \mathbf{1}\{|\mu_{i,t} - 0.65| \le 0.05\}.$$

The frozen configuration uses $w_d = w_u = w_b = 1/3$ and $w_f = 0$; the uncertainty-band trigger is disabled by default. The remaining Verify budget after the propensity floor is allocated proportionally to $s_{i,t}$.

C.4 Auditor Assignment

VTR-FL assumes a fixed authenticated auditor pool. After the update commitment is fixed, post-commitment public randomness permutes auditor identities and assigns k distinct auditors without replacement. Reports from distinct identities are required to form a certificate, and each evidence coordinate is aggregated by the median.

The prototype uses five auditor identities with $k = 5$, so all five disjoint audit shards participate in the main experiments. Auditor membership and Sybil resistance are assumed to be provided by an external identity or PKI layer and are outside the scope of VTR-FL. The collusion experiment

therefore evaluates bounded corruption within an already authenticated committee. Randomness, commitments, signatures, and certificate verification are simulated at the protocol layer.

D EXPERIMENTS

D.1 RECOVERY EVENT PROTOCOL

For each run, we sample a fixed client cohort $\mathcal{M}$ once using the run-specific seed, with $r_{\mathrm{mal}} = 0.2$. The same client identities are used throughout the entire degradation–recovery process. During rounds $W_{\mathrm{deg}} = [40, 60)$, each client $i \in \mathcal{M}$ performs standard local training except that 20% of labels in each mini-batch are cyclically corrupted as

$$y \leftarrow (y + 1) \bmod C. \tag{38}$$

At $t = 60$, label corruption is disabled, and the same clients resume clean local training. Hence, the degraded and recovered cohorts are not separately sampled populations, but two temporal states of the same identity-fixed cohort:

$$\mathcal{M}_{\mathrm{deg}} = \mathcal{M}_{\mathrm{rec}} = \mathcal{M}. \tag{39}$$

Unless otherwise specified, recovery is evaluated over $W_{\mathrm{rec}} = [60, 100)$. A client is counted as re-entered if it reaches the TRAIN state and receives positive aggregation weight at least once during this window. Accordingly, $P(\mathrm{Re})$ measures the fraction of recovered clients that regain effective model influence, while Δ_R compares re-entry after recovery against re-entry during active degradation. RCI further measures the post-degradation aggregation mass assigned to the same cohort. The main recovery benchmark uses only this temporary-label-noise protocol; adaptive update corruption is evaluated separately as a security stress test.

D.2 FULL CROSS-DATASET BASELINE RESULTS

Table 3: Full five-method cross-dataset recovery comparison.

Dataset	Method	$P(Re)$	AA^{deg}	RCI^{train}	Accuracy
CIFAR-10	Hard Threshold	.000	0.00	0.00	.208
	MAB-RFL	.038	3.62	.37	.525
	FairFedCS	.100	2.80	.49	**.563**
	RSVDUCBTE	.050	3.64	.36	.518
	VTR-FL	**.475**	2.98	**7.01**	.539
Fashion-MNIST	Hard Threshold	.000	0.00	0.00	.536
	MAB-RFL	.078	4.10	.12	.801
	FairFedCS	.094	4.42	.21	**.819**
	RSVDUCBTE	.031	2.97	.09	.778
	VTR-FL	**.516**	2.00	**5.80**	.807
MNIST	Hard Threshold	.000	0.00	0.00	.598
	MAB-RFL	.078	4.55	.18	**.977**
	FairFedCS	.109	4.46	.24	.976
	RSVDUCBTE	.078	3.00	.24	.962
	VTR-FL	**.516**	2.99	**3.92**	.976

D.3 AUDITOR COLLUSION ROBUSTNESS.

VTR-FL relies on auditor committees to certify recovery evidence under the honest-majority assumption. We therefore evaluate committee robustness by varying the fraction of malicious auditors from 0% to 40% and changing committee sizes, the results on CIFAR-10 are shown in Table 4. Malicious auditors attempt to inflate recovery scores or hide anomalous updates. We report evidence estimation error and false promotion rate. These experiments empirically characterize that median-based attestation remains robust under sub-majority auditor corruption.

Table 4: Auditor collusion robustness.

Size	MAR	Error	FPR
5	0%	0.000	0.000
5	20%	0.004	0.003
5	40%	0.012	0.062
11	40%	0.015	0.069