When Quality Gate Meets Poison: Label-Flip Robustness Emerges from PID-Regulated Merit Gating in Federated Learning

Yang Liu
Tongji University
Shanghai, China
yangliu25@tongji.edu.cn

Zhengxi Yu
Tongji University
Shanghai, China
2452633@tongji.edu.cn

Jing Liu
University of British Columbia
Vancouver, Canada
jingliu@ece.ubc.ca

Peng Sun*
Duke Kunshan University
Suzhou, China
peng.sun568@duke.edu

Azzedine Boukerche
University of Ottawa
Ottawa, Canada
aboukerc@uOttawa.ca

Hong Chen
Tongji University
Shanghai, China
chenhong2019@tongji.edu.cn

***Abstract*—Federated learning (FL) over wireless networks must cope with two pressures at once: the cost of transmitting model updates under limited uplink bandwidth, and the risk of poisoned updates from untrusted clients. These concerns are usually handled by separate mechanisms, and the robust ones tend to add quadratic computation, require a clean server-side dataset, or be tailored to a specific attack, none of which reduces communication. Meanwhile, communication-oriented client selection decides participation before training and never audits whether the resulting update is useful. We ask whether a single quality signal can serve both goals. Our starting point is that a post-hoc merit score, the relative loss improvement a client obtains after local training, acts as an objective-alignment test, so it should expose updates whose training objective conflicts with the global one. To this end, we develop a PID-regulated merit gate, in which a proportional-integral-derivative controller adapts the acceptance threshold to track a target participation rate as channel conditions and data distributions vary, and accepted updates are aggregated with merit-proportional weights. Experiments on CIFAR-10, CIFAR-100, and Tiny ImageNet show that the same gate used for bandwidth control also suppresses protocol-following label-flip poisoning, since adversarial updates receive negative scores and no attack-specific logic is needed. Under a 40% label-flip attack, our method exceeds federated averaging by 3.55 pp on CIFAR-10 and by 26.4 pp in the most heterogeneous setting, while cutting uplink traffic by 25%. The code is available at https://github.com/Tim-Yue/pid-merit-gating-fl.**

***Index Terms*—Federated learning; wireless networks; Byzantine robustness; communication efficiency; label-flip attack**

I. Introduction

Federated learning (FL) [1] enables collaborative model training across distributed clients without centralizing raw data, which fits mobile and wireless deployments where privacy constraints and regulatory requirements preclude data sharing. FL is also an important paradigm for intelligent resource management, network optimization, and mobile edge intelligence, where large populations of heterogeneous devices cooperate under strict bandwidth and latency budgets [2]–[4]. The same distributed-learning principles increasingly underpin video analytics [5], [6], distributed traffic control [7], and embodied multimedia systems [8]. Wireless FL, however, faces two practical adversities that are usually treated as separate engineering problems: the *communication bottleneck* and *Byzantine robustness* [9].

To be specific, transmitting full model updates from all available clients every round is prohibitively expensive, which has motivated extensive research on client selection [10]–[12]. Pow-d [11] selects clients with the highest local training loss, on the grounds that they contribute the most informative gradients. Oort [10] ranks clients jointly by statistical utility and response latency, and DivFL [12] selects a maximally diverse subset of gradient directions through submodular optimization. Recent work continues to refine client scheduling and resource allocation in wireless settings, including contextual and heterogeneity-aware sampling [13]–[15], joint scheduling with bandwidth allocation [16], [17], coded and incentive-based participation [18], [19], and client selection under unreliable links [20]. Such methods are fundamentally *pre-hoc*, since they predict update value before local training and do not audit the realized outcome.

Furthermore, in open wireless deployments where device identity cannot be fully verified, malicious participants may submit crafted updates to corrupt global training. The label-flip attack is a representative and practical threat, in which adversarial clients train on systematically relabeled data and upload poisoned updates that gradually degrade global accuracy [21]–[23]. Existing defenses treat the problem as one of detection. Krum [9] selects the update geometrically closest to its neighbors at $\mathcal{O}(n^2)$ cost per round, Coordinate-wise Trimmed Mean [24] discards extreme per-dimension values before averaging, and FLTrust [25] derives trust scores from cosine similarity to a server-side reference gradient, which

*Corresponding author.

requires a clean labeled dataset at the aggregator. AntiFLipper [26] monitors per-class confidence distributions to flag label-flip activity, and related detectors operate on gradient statistics or per-client trust [22], [27]. Such methods are modular additions layered atop standard aggregation, and none reduces uplink bandwidth.

The separation between these two problem layers is not merely architectural. Robust aggregation rules such as Krum impose per-round computation that is quadratic in the number of clients, validation-based schemes such as FLTrust require the server to maintain a clean, labeled, in-distribution dataset that is often unavailable in mobile deployments, and attack-specific detectors require redesign whenever the threat model changes. The communication overhead of transmitting all client updates, meanwhile, remains unaddressed. A natural question is whether a mechanism designed to solve one of these problems can also inform the other.

To this end, we attempt to answer this question affirmatively through a quality-gated design. MeritFL [28] proposed that clients self-evaluate their local updates via a merit score, defined as the relative reduction in training loss before and after local SGD, and suppress transmissions below a fixed threshold to save bandwidth. The post-hoc audit of realized training outcomes is qualitatively different from pre-hoc selection, since it measures whether a client actually produced a useful update rather than predicting whether it will. The threshold in MeritFL is static and set for bandwidth management alone, and no connection to adversarial robustness was drawn.

Our key observation is that the merit score is not merely a bandwidth signal but an *objective-alignment test*. A benign client's training objective and quality-evaluation loss are consistent, so local SGD reduces both, whereas a label-flip client trains on corrupted labels and produces an update that improves its poisoned objective while worsening the clean evaluation loss, so its merit score turns negative. A controller that keeps the acceptance threshold positive therefore rejects such updates as a byproduct of communication control, with no attack-specific logic required. The coupling is appealing because it costs little beyond what merit-gated communication already requires: no extra computation, no server-side data, and no redesign for new label-flip variants.

We build on this observation and make the following contributions.

- We present a self-adaptive quality gating framework in which a PID controller adjusts the merit threshold to track a target participation rate, which reduces the need for manual threshold tuning across varying channel conditions and data distributions.
- We show that under protocol-following label-flip poisoning, adversarial clients consistently produce negative merit scores and are rejected by the same gate used for bandwidth control, with no Byzantine-specific logic required, and a first-order analysis characterizes the objective-misalignment mechanism behind the separation.
- We evaluate the method against standard FL baselines and dedicated Byzantine defenses across nine attack ratios, three datasets, and three levels of data heterogeneity, showing consistent accuracy and communication advantages under attack, and we analyze when the controller helps and delimit the regime in which the robustness claim holds.

The remainder of the paper is organized as follows. Section II reviews related work on Byzantine-robust FL and quality-gated client selection. Section III presents the three-stage gating pipeline and analyzes the label-flip robustness mechanism. Section IV reports experimental results across datasets, attack ratios, and heterogeneity levels, and discusses scope and limitations. Section V concludes the paper.

II. Related Work

A. Byzantine-Robust Federated Learning

The Byzantine robustness problem in distributed learning was formally characterized by Blanchard et al. [9], who proposed Krum as the first robust aggregation rule. Krum selects the gradient update that minimizes the total squared distance to its $n-f-2$ nearest neighbors, on the assumption that Byzantine updates are geometric outliers in gradient space, and its $\mathcal{O}(n^2)$ per-round cost is substantial while the rule fails once Byzantine clients exceed half the population.

Coordinate-wise Trimmed Mean [24] reduces complexity to $\mathcal{O}(nd)$ by operating independently per parameter dimension, at the cost of assuming dimensionally uniform Byzantine corruption, while FLTrust [25] weights client updates by their cosine similarity to a server-side reference gradient and is effective when a small clean dataset is available, though that requirement limits deployment generality. For label-flip attacks specifically, a line of dedicated detectors has emerged: AntiFLipper [26] profiles per-class confidence, LFighter [21] exploits the conflicting gradient objectives of attackers and honest peers, FLGT [22] guides server trust through principal-component features of gradient angles, and other schemes detect poisoned updates from gradient or trust statistics [27]. Such detectors achieve high rates against their target threat but typically require redesign for other attack types, and a complementary line studies stronger poisoning attacks that evade robust aggregation, including catastrophic-forgetting based poisoning [23], which cautions against overstating robustness guarantees.

All of these methods are defense modules added on top of standard aggregation, and robustness is purchased through extra computation, trusted server-side data, or attack-specific engineering, with none reducing uplink bandwidth.

B. Quality-Gated FL and Client Selection

MeritFL [28] introduced client-side merit scoring for communication gating in wireless FL, where merit $= (\mathcal{L}^{\text{pre}} - \mathcal{L}^{\text{post}})/\mathcal{L}^{\text{pre}}$ and only sufficiently meritorious clients upload updates, with communication efficiency as its primary contribution and Byzantine robustness neither claimed nor analyzed.

Our work extends MeritFL in three respects: we replace its static threshold with a PID-controlled adaptive threshold that

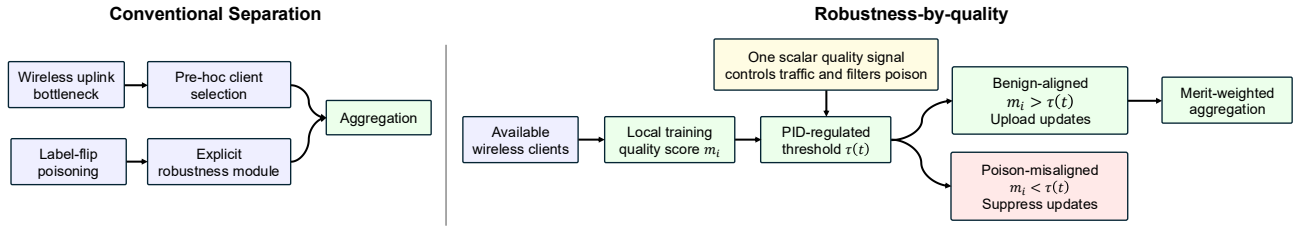


Fig. 1. Conceptual overview. Conventional FL treats communication efficiency and Byzantine robustness as independent layers. PID-regulated merit gating couples both objectives through a single scalar update-quality signal. The same gate that suppresses low-value transmissions also rejects label-flip poisoned updates whose observed merit is negative.

tracks a target participation rate over time, we apply merit-weighted aggregation so that higher-quality updates receive proportionally greater influence, and we identify and characterize the connection between merit-based filtering and label-flip robustness.

Pre-hoc selection methods [10]–[15] cannot assess whether the resulting update is useful, since a high-loss client selected by Pow-d may still produce a harmful update when its data are corrupted. Recent wireless schemes that jointly optimize scheduling, bandwidth, and incentives [16]–[20] likewise decide participation before observing the realized training outcome, whereas quality-gated communication evaluates that outcome rather than predicting it.

III. Method

A. System Model

We consider a standard FL system with a single aggregation server and $N = 60$ wireless clients, and in each round t a subset of $K = 10$ clients participates from an SNR-feasible pool. Each client i holds a local dataset $\mathcal{D}_i$ drawn from a Dirichlet(α) partition of the global data, where a smaller α yields more skewed, non-IID distributions, and we set $\alpha = 0.5$ by default, consistent with prior work [29].

All clients share a ResNet-10 backbone of about 4.9M parameters, with output dimension matched to the dataset at 10, 100, or 200 classes. Local optimization uses SGD with learning rate 0.01, Nesterov momentum 0.9, weight decay 10^{-5}, and $E = 10$ local epochs per round with batch size 64.

B. Wireless Channel Model

We adopt a Rayleigh fading model to determine the channel-feasible pool each round, where each client i has a fixed large-scale mean SNR:

$$\overline{\gamma}_i \sim \text{Uniform}[5,\ 25]\ \text{dB}. \tag{1}$$

At each round, fast fading modulates the instantaneous SNR:

$$\gamma_i = \overline{\gamma}_i \cdot e, \quad e \sim \text{Exp}(1), \tag{2}$$

and the channel access probability is a sigmoid of the instantaneous SNR relative to a 5 dB threshold:

$$P_{\text{attempt}}^{(i)} = \sigma\left(\gamma_{\text{dB}}^{(i)} - \gamma_{\text{th}}\right), \quad \gamma_{\text{th}} = 5\ \text{dB}, \quad \sigma(x) = \frac{1}{1+e^{-x}}. \tag{3}$$

All compared methods share the same channel realization and random seed, so performance differences reflect gating decisions rather than pool membership, and when the pool falls below K additional clients are drawn uniformly from the remainder.

C. Three-Stage Gating Pipeline

Our method executes a three-stage pipeline each round, as illustrated in Fig. 2.

1) Stage 1: SNR-Based Pool Formation: Clients attempt channel access with probability $P_{\text{attempt}}^{(i)}$, and successful clients form the available pool:

$$\mathcal{P}_t = \{i \mid u_i < P_{\text{attempt}}^{(i)},\ u_i \sim \text{Uniform}[0,1]\}. \tag{4}$$

2) Stage 2: Local Training and Merit Computation: Each pooled client receives $\mathbf{w}_t$ and performs E epochs of local SGD, and before training client i evaluates the pre-training loss on a held-out mini-batch $\mathcal{B}_i \subset \mathcal{D}_i$ that is disjoint from its training samples:

$$\mathcal{L}_i^{\text{pre}} = \frac{1}{|\mathcal{B}_i|} \sum_{(x,y)\in\mathcal{B}_i} \ell(f(x;\mathbf{w}_t),\ y), \tag{5}$$

and after training on the updated model $\mathbf{w}_i^{(t+1)}$:

$$\mathcal{L}_i^{\text{post}} = \frac{1}{|\mathcal{B}_i|} \sum_{(x,y)\in\mathcal{B}_i} \ell(f(x;\mathbf{w}_i^{(t+1)}),\ y). \tag{6}$$

The merit score is the relative loss reduction:

$$m_i = \frac{\mathcal{L}_i^{\text{pre}} - \mathcal{L}_i^{\text{post}}}{\mathcal{L}_i^{\text{pre}} + \epsilon}, \quad \epsilon = 10^{-8}. \tag{7}$$

A positive score indicates that local training improved the model on the client's distribution, while a negative score signals misalignment between the client's local objective and the global quality criterion, a condition that, as Section III-D shows, label-flip poisoning reliably triggers.

To reduce round-to-round noise from mini-batch variance, each client maintains an exponential moving average (EMA) of its merit:

$$\tilde{m}_i^{(t)} = \rho \cdot m_i^{(t)} + (1-\rho)\cdot \tilde{m}_i^{(t-1)}, \quad \rho = 0.3. \tag{8}$$

The weight $\rho = 0.3$ balances responsiveness to genuine quality changes against stability under transient fluctuations, and prevents a single anomalous round from disqualifying a client with a consistently good track record.

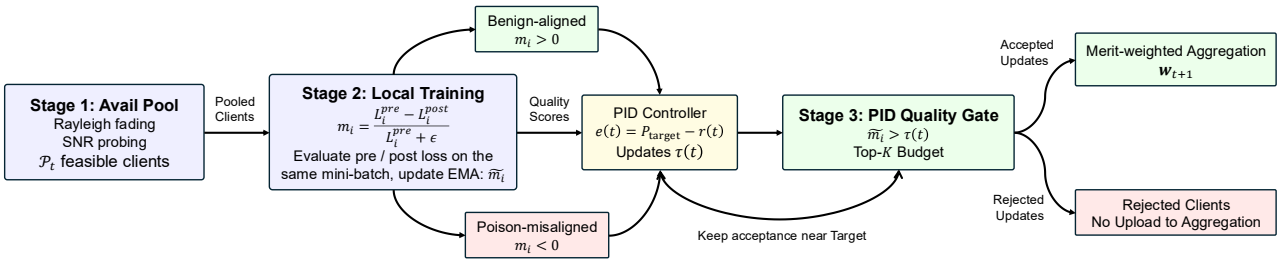


Fig. 2. Per-round pipeline. The server forms an SNR-feasible pool. Pooled clients train locally and report scalar merit scores. A PID controller adapts the threshold to maintain the target participation rate. Under label-flip poisoning, adversarial updates fall into the negative-merit branch and are suppressed before transmission.

3) Stage 3: PID-Controlled Gating and Aggregation:

a) Adaptive Thresholding: A fixed merit threshold faces an inherent tension: a high value slows convergence, a low value lets low-quality updates dilute aggregation, and the appropriate threshold shifts over training as the model converges and merit scores compress. We address the tension with a PID controller that tracks the fraction of pooled clients passing the gate, which we call the participation rate $r(t)$, against a target $P_{\text{target}} = 0.7$, and the tracking error is:

$$e(t) = P_{\text{target}} - r(t). \tag{9}$$

The threshold is updated as:

$$\tau(t) = \tau(t{-}1) + K_p e(t) + K_i \sum_{j=0}^{t} e(j) + K_d \big[e(t) - e(t{-}1)\big], \tag{10}$$

with gains $K_p = 0.01$, $K_i = 0.001$, $K_d = 0.005$. The proportional term provides immediate correction to participation-rate deviations, the integral term reduces steady-state error that arises from persistent data quality differences across clients, and the derivative term damps oscillations induced by channel variability, so that the gains together yield a well-damped response that settles near P_{target} within about 20 to 30 rounds and removes the need for round-by-round manual recalibration.

b) Gating and Selection: Eligible clients are those whose smoothed merit exceeds the current threshold:

$$\mathcal{E}_t = \big\{ i \in \mathcal{P}_t \mid \tilde{m}_i^{(t)} > \tau(t) \big\}. \tag{11}$$

The top-K clients by smoothed merit are then selected for transmission:

$$\mathcal{U}_t = \text{top-K}_{\tilde{m}}\ \mathcal{E}_t. \tag{12}$$

The two-stage design decouples quality control from budget enforcement, since the threshold filters out genuinely poor updates regardless of how many clients are available, while the budget cap K bounds communication cost regardless of how many clients pass the threshold.

c) Merit-Weighted Aggregation: Accepted updates are aggregated with weights proportional to both local dataset size and merit score, so that higher-quality updates receive greater influence:

$$\mathbf{w}_{t+1} = \frac{\sum_{i \in \mathcal{U}_t} n_i \cdot \max\big(\tilde{m}_i^{(t)}, \gamma_{\min}\big) \cdot \mathbf{w}_i^{(t+1)}}{\sum_{i \in \mathcal{U}_t} n_i \cdot \max\big(\tilde{m}_i^{(t)}, \gamma_{\min}\big)}, \tag{13}$$

where $n_i = |\mathcal{D}_i|$ and $\gamma_{\min} = 0.1$ provides a minimum weight floor for numerical stability.

d) Warmup Phase: For the first $T_{\text{warmup}} = 15$ rounds, the merit gate is disabled and all pooled clients transmit, which lets EMA estimates stabilize and the global model reach a reasonable initialization before filtering begins.

D. Robustness to Label-Flip Poisoning

The PID-controlled quality gate filters label-flip poisoned updates without attack-specific detection logic, and we analyze why under a *protocol-following* threat model, in which adversarial clients corrupt their training labels but compute and report merit scores according to the prescribed protocol.

a) Adversarial Model: Each adversarial client trains on labels shifted cyclically by one class:

$$y \leftarrow (y + 1) \bmod C, \tag{14}$$

where C is the number of output classes, and the attack requires no knowledge of the aggregation rule or of other clients' updates, which makes it a realistic threat in open wireless deployments.

b) Why Merit Becomes Negative: Merit acts as an objective-alignment test. For a benign client, the training objective and the quality-evaluation loss in Eq. (7) are consistent, so local SGD reduces the same loss that merit later measures, whereas for a label-flip client the training objective operates on shifted labels:

$$\tilde{\mathcal{L}}_i(\mathbf{w}) = -\frac{1}{|\mathcal{B}_i|} \sum_{(x,y) \in \mathcal{B}_i} \log f_{(y+1) \bmod C}(x; \mathbf{w}). \tag{15}$$

The resulting update reduces $\tilde{\mathcal{L}}_i$ but need not reduce $\mathcal{L}_i$, as a single-step approximation of the update expands the change in quality-evaluation loss to first order:

$$\Delta \mathcal{L}_i \approx -\eta\, \nabla \mathcal{L}_i(\mathbf{w}_t)^\top \nabla \tilde{\mathcal{L}}_i(\mathbf{w}_t). \tag{16}$$

When the poisoned gradient is anti-aligned with the evaluation gradient, the inner product is negative, $\Delta\mathcal{L}_i$ is positive, and merit is negative. In practice, clients run $E = 10$ local epochs rather than a single step, yet the directional conflict between objectives persists throughout training, and empirically adversarial merit scores range from -1.5 to -38 across training stages while benign clients score consistently in $[+0.70, +0.99]$.

c) Why the Gate Rejects Them: After warmup, the PID-regulated threshold τ remains in $[0.04, 0.28]$, and since adversarial scores are negative while the threshold is positive,

$$\tilde{m}_i^{(t)} < 0 < \tau(t) \implies \text{client } i \text{ is rejected.} \tag{17}$$

The separation between populations exceeds the threshold range by a wide margin, so filtering is stable across controller settings. The robustness incurs little marginal cost relative to a system that already performs merit-gated communication: no Byzantine-specific computation is added, no server-side reference data is required, no attack-specific detector is encoded, and rejected clients transmit only a scalar merit report rather than a full model update. We stress that the argument is specific to the protocol-following label-flip setting, and we return to its limits in Section IV-H.

IV. Experimental Evaluation

A. Experimental Setup

The main attack-ratio sweep uses CIFAR-10 with ResNet-10, trained for 200 rounds with $E = 10$ local epochs per round, while cross-dataset evaluations use CIFAR-100 and Tiny ImageNet at a fixed 40% label-flip ratio. Label-flip ratios of 0%, 5%, 10%, 20%, and 40% are tested on CIFAR-10 with $\alpha = 0.5$, and we further report ratios up to 80% to probe behavior beyond the classical Byzantine regime, while the heterogeneity ablation varies $\alpha \in \{0.1, 0.5, 1.0\}$ at 40% attack. All methods share the same SNR-based available pool per round under an identical channel seed, so accuracy differences reflect gating and aggregation decisions rather than channel availability.

Adversarial clients are assigned at the start of training and maintain corrupted labels throughout all rounds in which they participate, the fraction of adversarial clients among those selected each round is controlled by the attack ratio, and the server has no prior knowledge of which clients are adversarial. The baseline methods are:

- **FedAvg** [1]: random client selection, standard federated averaging, no defense.
- **Krum** [9]: Byzantine-robust aggregation that selects the update with minimum sum-of-squared-distances to neighbors.
- **Trimmed Mean** [24]: coordinate-wise robust aggregation with trimming fraction $\beta = 0.1$.
- **FedProx** [29]: proximal regularization $\mu = 0.01$ in the local objective, no defense.
- **FedNova** [30]: update normalization by local SGD steps, no defense.

TABLE I
Final accuracy (%) under label-flip attacks on CIFAR-10 with Dirichlet non-IID partition ($\alpha = 0.5$).

Method	0%	5%	10%	20%	40%	50%	60%	70%	80%
FedAvg [1]	88.30	87.35	87.02	86.95	79.71	80.61	67.13	42.88	10.57
Krum [9]	72.55	67.78	65.44	72.57	59.67	50.27	6.87	3.13	2.58
Trimmed Mean [24]	88.50	87.51	87.07	85.64	59.10	22.30	8.99	5.92	3.35
FedProx [29]	88.24	87.44	87.31	86.74	79.73	79.78	65.40	38.40	8.99
FedNova [30]	87.87	87.10	86.59	86.85	80.39	80.29	67.30	41.83	9.47
Ours	85.45	85.14	84.74	85.26	**83.26**	**83.40**	**81.35**	**76.83**	**69.19**

B. Main Result: Attack Ratio Sweep

Table I and Fig. 3 present the attack-ratio sweep on CIFAR-10. In the clean setting at 0% attack, FedAvg outperforms our method by 2.85 pp, a gap that reflects the deliberate cost of quality gating: the gate suppresses some benign updates whose merit falls below the threshold, trading a small amount of peak accuracy for reduced communication volume and improved filtering under attack. Our clean-setting accuracy stays stable across all attack ratios from 0% to 20%, where it degrades by less than 0.8 pp, whereas FedAvg degrades by 1.35 pp over the same range.

As the attack ratio rises to 40%, the trade-off reverses. Our method degrades by only 2.19 pp, from 85.45% to 83.26% relative to its own clean baseline, while FedAvg degrades by 8.59 pp, from 88.30% to 79.71%, roughly four times as much, so our method leads FedAvg by 3.55 pp in final accuracy while transmitting 25% fewer updates. FedProx and FedNova track FedAvg closely across all ratios, which indicates that proximal regularization and update normalization do not address the label-flip failure mode, and closing the gap requires a quality-based filter rather than a modified local optimizer.

The high-ratio regime is informative. Beyond a 50% adversarial fraction, classical Byzantine guarantees no longer hold and the geometric defenses degrade sharply, with Krum collapsing to 6.87% at 60% and Trimmed Mean to 8.99%, while FedAvg, FedProx, and FedNova fall below 11% at 80%, yet our method retains 69.19%. We attribute the behavior to the per-client nature of the gate, which accepts or rejects each client on its own realized quality rather than relying on a benign majority in gradient space, so that as long as enough benign clients with positive merit reach the pool, the top-K selection in Eq. (12) still populates the aggregation set with useful updates. We report the regime as an empirical observation specific to the protocol-following label-flip threat, not as a general Byzantine guarantee.

C. Cross-Dataset Validation

Table II extends the evaluation to CIFAR-100 and Tiny ImageNet under 40% attack. On CIFAR-100, the one-step cyclic label flip is a finer-grained perturbation across 100 classes, so all methods are more robust than on CIFAR-10, and our method reaches 47.64% final accuracy while transmitting 25% less data than FedAvg, the best or among the best on the final metric. Tiny ImageNet is harder, with 200 classes and greater visual diversity, where Trimmed Mean

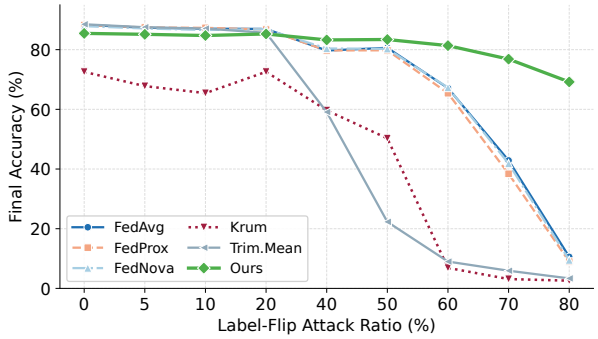


Fig. 3. Final accuracy versus attack ratio (CIFAR-10, $\alpha = 0.5$).

TABLE II
CROSS-DATASET VALIDATION AT 40% LABEL-FLIP ATTACK.

Method	CIFAR-100		Tiny ImageNet	
	Best	Final	Best	Final
FedAvg [1]	48.41	46.78	36.51	33.11
Krum [9]	23.63	22.51	21.32	13.38
Trim. Mean [24]	46.30	41.41	**36.95**	30.90
FedProx [29]	48.38	45.79	36.70	31.69
FedNova [30]	**48.99**	46.90	35.94	32.65
Ours	48.71	**47.64**	36.41	**35.98**

TABLE III
FINAL ACCURACY (%) UNDER 40% LABEL-FLIP ATTACK ACROSS DATA HETEROGENEITY LEVELS (CIFAR-10).

Method	$\alpha = 0.1$	$\alpha = 0.5$	$\alpha = 1.0$
FedAvg [1]	45.04	79.71	84.94
Krum [9]	10.66	59.67	8.63
Trimmed Mean [24]	38.77	59.10	59.67
FedProx [29]	50.40	79.73	84.51
FedNova [30]	44.29	80.39	84.81
Ours	**71.47**	**83.26**	**85.75**

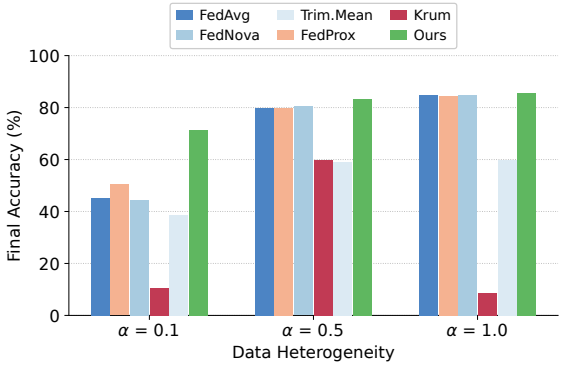


Fig. 4. Final accuracy versus data heterogeneity level under 40% attack.

and FedProx reach slightly higher peak checkpoints of 36.95% and 36.70% against our 36.41%, yet both decline to a lower final accuracy as poisoning accumulates, while our method holds the strongest final accuracy of 35.98% against FedAvg's 33.11%, a gap of 2.87 pp.

The distinction between peak and final accuracy is meaningful under sustained poisoning, since baselines that reach higher transient checkpoints subsequently degrade as poisoned updates accumulate, whereas the quality gate stabilizes the final model by filtering harmful updates throughout training rather than only at peak. The final-round model is also the artifact actually deployed, which makes its accuracy the more operationally relevant of the two metrics, so we report both and do not claim a uniform improvement on the peak metric.

D. Data Heterogeneity Ablation

Table III and Fig. 4 reveal the largest gap between methods. At the extreme non-IID level $\alpha = 0.1$, the random-aggregation baselines fall to between 44% and 51%, with the geometric defenses lower still, while our method maintains 71.47%, a 26.4 pp advantage over FedAvg, and the result reflects an interaction between data heterogeneity and poisoning that is specific to random aggregation. Under highly skewed partitions, each client's local data covers only a narrow slice of the label space, so a poisoned client appears statistically influential and its update direction is rarely cancelled by benign updates from clients with complementary data, while random aggregation has no mechanism to tell that the influence is harmful rather than informative. Merit gating conditions acceptance on realized update quality, so a poisoned client's negative merit score reveals its misalignment regardless of how distinctive its local distribution appears. At moderate heterogeneity $\alpha = 0.5$ and $\alpha = 1.0$, the gap shrinks to 3.55 pp and 0.8 pp, because random aggregation benefits from more benign updates that cancel adversarial directions, and the pattern indicates that the method is most valuable in the skewed regime common to wireless deployments while its margin narrows as data approach the IID case.

E. Hyperparameter and Controller Sensitivity

Table IV reports sensitivity to the target participation rate and to the PID gains under 40% attack. Final accuracy varies by only 1.04 pp across $P_{\text{target}} \in [0.5, 1.0]$ and by about 1.6 pp across the gain settings, from 82.77% to 84.41%, which we read as low sensitivity to controller tuning rather than as evidence that any single configuration is uniformly preferable. The low sensitivity follows from the wide score margin between populations, since adversarial clients score in $[-1.5, -38]$ and benign clients in $[+0.70, +0.99]$, so the controller need only keep the threshold positive to separate them and need not locate a precise decision boundary. A more conservative target $P_{\text{target}} = 0.5$ yields a marginal 0.4 pp gain under the attack, because fewer borderline updates are admitted, but the default of 0.7 balances robustness against participation breadth.

The open-loop variant with a fixed threshold $\tau = 0.3$ attains 83.35%, comparable to the PID-regulated default of 83.26%, and the comparison indicates that robustness originates from the merit signal itself, which separates the two populations by

TABLE IV
PID TARGET RATE AND GAIN SENSITIVITY (40% ATTACK, CIFAR-10).

Configuration	Role	Final Acc
$P_{\text{target}} = 0.5$	Conservative gate	83.66
$P_{\text{target}} = 0.7$	Default	83.26
$P_{\text{target}} = 1.0$	All-pass (≈FedAvg)	82.62
$K_p = 0.05$ (default: 0.01)	High P gain	84.11
$K_i = 0.005$ (default: 0.001)	High I gain	82.77
$K_d = 0.02$ (default: 0.005)	High D gain	84.41
Fixed $\tau = 0.3$	Open-loop gate	83.35

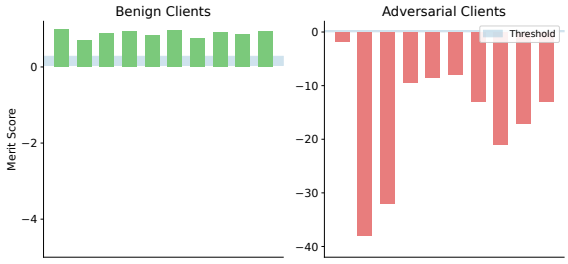


Fig. 5. Merit score distributions for benign (left) and adversarial (right) clients. The blue band marks the PID threshold range (0.04 to 0.28).

a wide margin. The role of the PID controller is to hold a target participation rate, and therefore a positive threshold, without manual recalibration as channel conditions and merit scales drift over training, so that in deployments where a single fixed threshold cannot be calibrated in advance across heterogeneous and time-varying conditions, the tuning-free behavior is the practical benefit rather than a direct accuracy gain.

F. Communication Efficiency

Our method reduces uplink traffic from 39.2 GB to 29.4 GB over 200 rounds, a 25% reduction, while FedAvg, FedProx, and FedNova each transmit the full 39.2 GB. The reduction is structurally meaningful, because the suppressed transmissions are precisely the updates the quality gate identifies as harmful or low-value, so the bandwidth saved corresponds to the updates that would otherwise have degraded the global model, and communication efficiency and robustness arise as joint consequences of the same filtering decision rather than as independent effects that must be optimized separately. Because the cap K bounds the number of uploads each round, the uplink cost stays fixed even as the attack ratio grows, while rejected poisoned updates never consume bandwidth, and the magnitude of the reduction depends on the chosen participation target P_{target} and budget K, with each rejected client still transmitting a scalar merit report whose cost is negligible relative to a full model update.

G. Mechanism Visualization

Fig. 5 illustrates the score distributions that underlie the filtering behavior. Benign clients score consistently in $[+0.70, +0.99]$ and adversarial clients in $[-1.5, -38]$, with more negative scores in later rounds as the global model becomes increasingly misaligned with the corrupted-label objective, while the PID threshold fluctuates in $[0.04, 0.28]$, far above the highest adversarial score and far below the lowest benign score. The wide margin between the threshold range and the population boundary explains why the method is insensitive to the exact PID target, and why the same gate that manages bandwidth also provides reliable label-flip filtering.

H. Limitations Discussion

The results above hold under a protocol-following threat model, in which adversaries corrupt labels but report merit honestly, and the assumption is central, so we state its limits plainly. Because the merit score is computed and reported by the client, an adversary that forges a positive score, or that trains honestly on the evaluation mini-batch while poisoning the rest of its data, could pass the gate, so the mechanism is best understood as a low-cost filter that is effective against protocol-following label-flip attacks and as a complement to dedicated Byzantine defenses rather than a replacement for them. Server-side verification of reported merit, for example through occasional recomputation on held-out probes, is a natural way to harden the signal, and we leave a full treatment to future work.

Two further points bound the interpretation of our numbers. First, the reported accuracies come from runs that share a fixed channel seed across methods, which isolates gating decisions but does not by itself establish statistical significance, so small gaps such as the 0.8 pp margin at $\alpha = 1.0$ or the sub-1 pp differences on CIFAR-100 should be read as comparable performance rather than as decisive separation. Second, the controller provides tuning-free adaptation rather than an accuracy boost under stationary conditions, as the fixed-threshold comparison shows, and its advantage is expected to grow when channel statistics or data distributions drift over time, which makes a study under non-stationary conditions a clear next step.

V. CONCLUSION

In this paper, we propose a PID-regulated merit gate for wireless uplink efficiency that filters label-flip poisoned updates as a byproduct of its quality-evaluation criterion. The merit score measures whether local training improves the loss on the client's own data, and label-flip poisoning systematically produces negative scores because the adversarial training objective conflicts with the criterion, so the quality gate rejects adversarial updates without Byzantine-specific logic, server-side reference data, or additional communication overhead. Across CIFAR-10, CIFAR-100, and Tiny ImageNet, the method attains strong sustained accuracy under attack while reducing uplink traffic by 25%, and the largest gap, 26.4 pp over FedAvg, appears under extreme data heterogeneity. The finding points to a broader design principle for wireless federated systems, in which a quality signal introduced for one layer of the stack carries useful evidence for another, though the guarantees here are scoped to the protocol-following threat

model and the method is best read as a complement to dedicated Byzantine defenses. Future work will examine whether analogous quality-robustness connections hold for other loss objectives and client quality metrics, evaluate the gate under stronger and adaptive adversaries, and harden merit reporting against clients that forge the quality signal.

ACKNOWLEDGMENT

This work was supported by the DKU Professional Discretionary Funds under Grant No. 26AKUG0088 and 00AKUG0343.

REFERENCES

[1] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2017.

[2] F. Mirhakimi, B. Javadi, R. N. Calheiros, and A. N. Toosi, "Federated learning with reliability-aware workload allocation in distributed edge computing," in *Proceedings of the International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems (MSWiM)*, 2025, pp. 5–9.

[3] C. G. S. Capanema, F. A. Silva, L. A. Villas, and A. A. F. Loureiro, "Non-iid-aware multi-model federated learning," in *Proceedings of the International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems (MSWiM)*, 2025, pp. 10–14.

[4] Q. Zhang, Y. Wang, G. Li, B. Cardiff, and P. Ranaweera, "Optimizing federated learning on non-iid data with clustering and model sharing," in *Proceedings of the Joint European Conference on Networks and Communications and 6G Summit (EuCNC/6G Summit)*, 2025, pp. 205–210.

[5] J. Liu, Y. Liu, J. Lin, J. Li, L. Cao, P. Sun, B. Hu, L. Song, A. Boukerche, and V. C. Leung, "Networking systems for video anomaly detection: A tutorial and survey," *ACM Computing Surveys*, vol. 57, no. 10, pp. 1–37, 2025.

[6] Y. Liu, H. Wang, Z. Wang, X. Zhu, J. Liu, P. Sun, R. Tang, J. Du, V. C. Leung, and L. Song, "Crcl: Causal representation consistency learning for anomaly detection in surveillance videos," *IEEE Transactions on Image Processing*, vol. 34, pp. 2351–2366, 2025.

[7] L. Qian, P. Sun, J. Jin, Y. Liu, A. Boukerche, and L. Song, "Fedlight: A fuel-economic reinforcement learning model for distributed traffic signal control with spatiotemporal decomposition," in *2026 IEEE 46th International Conference on Distributed Computing Systems (ICDCS)*. IEEE, 2026, pp. 371–380.

[8] W. Zuo, Y. Liu, W. Yang, F. Wu, W. Zhou, J. Liu, P. Sun, J. Cheng, D. Yang, and X. Zeng, "Embodied multimedia: When multimedia meets embodied intelligence," *Available at SSRN 6456578*, 2026.

[9] P. Blanchard, E. M. E. Mhamdi, R. Guerraoui, and J. Stainer, "Machine learning with adversaries: Byzantine tolerant gradient descent," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[10] F. Lai, X. Zhu, H. V. Madhyastha, and M. Chowdhury, "Oort: Efficient federated learning via guided participant selection," in *Proceedings of the 15th USENIX Symposium on Operating Systems Design and Implementation (OSDI)*, 2021.

[11] Y. J. Cho, J. Wang, and G. Joshi, "Client selection in federated learning: Convergence analysis and power-of-choice selection strategies," in *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2022.

[12] R. Balakrishnan, T. Li, T. Zhou, N. Himayat, V. Smith, and J. Bilmes, "Diverse client selection for federated learning via submodular maximization," in *Proceedings of the 10th International Conference on Learning Representations (ICLR)*, 2022.

[13] Q. Pan, H. Cao, Y. Zhu, J. Liu, and B. Li, "Contextual client selection for efficient federated learning over edge devices," *IEEE Transactions on Mobile Computing*, vol. 23, no. 6, pp. 6538–6548, 2023.

[14] B. Luo, W. Xiao, S. Wang, J. Huang, and L. Tassiulas, "Adaptive heterogeneous client sampling for federated learning over wireless networks," *IEEE Transactions on Mobile Computing*, vol. 23, no. 10, pp. 9663–9677, 2024.

[15] H. Chen and H. Vikalo, "Heterogeneity-guided client sampling: Towards fast and efficient non-iid federated learning," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 37, 2024, pp. 65 525–65 561.

[16] T. Zhang, K.-Y. Lam, J. Zhao, and J. Feng, "Joint device scheduling and bandwidth allocation for federated learning over wireless networks," *IEEE Transactions on Wireless Communications*, vol. 24, no. 1, pp. 3–18, 2023.

[17] J. Tang, X. Li, H. Li, P. Li, X. Wang, and V. C. M. Leung, "Joint class-balanced client selection and bandwidth allocation for cost-efficient federated learning in mobile edge computing networks," *IEEE Transactions on Mobile Computing*, vol. 24, no. 7, pp. 5681–5698, 2025.

[18] Y. Sun, J. Shao, Y. Mao, S. Li, and J. Zhang, "Stochastic coded federated learning: Theoretical analysis and incentive mechanism design," *IEEE Transactions on Wireless Communications*, vol. 23, no. 6, pp. 6623–6638, 2023.

[19] J. Tan, Z. Liu, K. Guo, and M. Zhao, "Long-term client selection for federated learning with non-iid data: A truthful auction approach," *IEEE Internet of Things Journal*, vol. 12, no. 5, pp. 4953–4970, 2024.

[20] Y. Wang, W. Ji, J. Zhou, F. Xiao, and T.-H. Chang, "Robust federated learning in unreliable wireless networks: A client selection approach," *arXiv preprint arXiv:2502.17260*, 2025.

[21] N. Jebreel, J. Domingo-Ferrer, D. Sánchez, and A. Blanco-Justicia, "LFighter: Defending against the label-flipping attack in federated learning," *Neural Networks*, vol. 170, pp. 111–126, 2024.

[22] H. Li, Z. Shi, M. Jin, A. Yin, and Z. Zhao, "FLGT: label-flipping-robust federated learning via guiding trust," *Knowledge and Information Systems*, 2025.

[23] Y. Liu, C. Wang, and X. Yuan, "BadSampler: Harnessing the power of catastrophic forgetting to poison byzantine-robust federated learning," in *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2024.

[24] D. Yin, Y. Chen, R. Kannan, and P. Bartlett, "Byzantine-robust distributed learning: Towards optimal statistical rates," in *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 2018.

[25] X. Cao, M. Fang, J. Liu, and N. Z. Gong, "FLTrust: Byzantine-robust federated learning via trust bootstrapping," in *Proceedings of the 28th Network and Distributed System Security Symposium (NDSS)*, 2021.

[26] A. Rahman, A. Hasan, S. Arifin, F. H. Bappy, T. Hossain, T. Islam, A. R. M. Kamal, and M. A. Hossain, "Antiflipper: Defending label-flipping attacks in federated learning," 2025, arXiv:2509.22873.

[27] D.-P. Khuu, M. Sober, D. Kaaser, M. Fischer, and S. Schulte, "Data poisoning detection in federated learning," in *Proceedings of the 39th ACM/SIGAPP Symposium on Applied Computing (SAC)*, 2024, pp. 1549–1558.

[28] Z. Guo, K. Yang, L. Gong, Y. Liu, J. Liu, and L. Song, "Meritfl: Self-regulating federated learning via merit-gated communication," in *2026 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, 2026, pp. 3072–3076.

[29] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," in *Proceedings of the 3rd Conference on Machine Learning and Systems (MLSys)*, 2020.

[30] J. Wang, Q. Liu, H. Liang, G. Joshi, and H. V. Poor, "Tackling the objective inconsistency problem in heterogeneous federated optimization," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.