F*in*P: Fairness-in-Privacy in Federated Learning by Addressing Disparities in Privacy Risk

Tianyu Zhao
University of California, Irvine
Irvine, California, USA
tianyu.zhao@uci.edu

Mahmoud Srewa
University of California, Irvine
Irvine, California, USA
msrewa@uci.edu

Salma Elmalaki
University of California, Irvine
Irvine, California, USA
salma.elmalaki@uci.edu

Abstract

Federated Learning (FL) inherently mitigates mass data centralization risks; however, its privacy protections are not equally distributed — leaving vulnerable individuals disproportionately exposed to sophisticated privacy attacks. Crucially, statistical heterogeneity in human-centric FL environments often results in an inequitable distribution of privacy risks, particularly affecting those whose sensitive attributes or behaviors make them outliers. To address this critical gap, we introduce F*in*P, a novel framework designed to formalize and enforce fairness-in-privacy by mitigating disproportionate client vulnerability to Source Inference Attacks (SIA). F*in*P operationalizes a two-pronged defense strategy that tackles both the symptoms and root causes of privacy disparity, ensuring that no group of clients bears an excessive privacy burden. It combines a server-side adaptive aggregation mechanism, which dynamically weights client contributions based on their estimated privacy risk, with a client-side regularization technique to curb localized overfitting that drives unique data memorization. Extensive empirical evaluations on FEMNIST, Human Activity Recognition (HAR), and CIFAR-10 datasets demonstrate that F*in*P effectively aligns privacy fairness with primary task utility. Notably, F*in*P successfully mitigates SIA risks and reduces disparities in privacy exposure, establishing that strong fairness-in-privacy guarantees need not compromise model utility. Ultimately, F*in*P establishes equitable privacy protections by reducing vulnerability disparities by up to 57.14%, while preserving global model utility within a marginal ±1.75% of standard federated baselines.

Keywords

Human-centered system, Fairness, Privacy, Federated Learning, Differential Privacy, HAR

1 Introduction

The pervasive integration of machine learning (ML) into human-centric applications demands careful consideration of both its ethical implications and its privacy guarantees. Historically, algorithmic fairness research has focused primarily on preventing bias and discrimination in model decisions or predictive accuracy [11, 12, 27, 30, 40, 59, 60]. However, as ML systems increasingly rely on sensitive personal data, a critical yet underexplored dimension of fairness emerges: the fair distribution of privacy risks. Existing privacy definitions often focus on average or worst-case leakage across an entire dataset [33]. Yet, the crucial aspect of fair risk distribution—ensuring that no single individual or demographic group disproportionately shoulders the burden of privacy vulnerabilities—has received significantly less attention.

Federated Learning (FL) has emerged as an important and impactful privacy-enhancing technology to train ML models on decentralized devices. By enabling collaborative learning without centralizing raw data, FL inherently mitigates the risks of mass data collection and aligns well with stringent regulatory frameworks such as the GDPR and its Data Protection Impact Assessment (DPIA) provisions [1, 6, 42]. Despite these advantages, FL remains vulnerable to sophisticated privacy attacks, including Membership Inference Attacks (MIA) [46] and Source Inference Attacks (SIA) [21], which aim to trace model memorization back to specific training data or participating clients. Crucially, the decentralized and statistically heterogeneous nature of FL — where data is not Independent and Identically Distributed (non-IID) — often amplifies unequal privacy leakage. When models overfit to clients with atypical or underrepresented data, those specific clients experience significantly higher privacy risks, creating a new form of digital inequality.

Motivating Context and Societal Impact. The 2024 National Public Data (NPD) breach, which exposed billions of records, powerfully underscores the societal need for fairness in privacy [50]. Although the breach was widespread, its downstream consequences disproportionately impacted vulnerable populations—such as low-income individuals and the elderly—who have fewer resources to recover from identity theft and data abuse. Acknowledging the near inevitability of some data leakage—as reflected in the "zero trust" security paradigm [43]—the security community's focus must expand from solely preventing breaches to ensuring equitable risk distribution when leaks occur. Within decentralized systems such as FL, this means formally defining and mitigating disproportionate harm. Formalizing the notion of fairness-in-privacy provides policymakers with the tangible metrics required to enforce equitable protections and fosters the public trust necessary for widespread adoption of AI.

Disparate Risk in Human-Centric FL. To illustrate this disparity, consider the application of FL in Human Activity Recognition (HAR) for personalized health monitoring [38, 42]. In an FL-based HAR system, each user's wearable device acts as a client [37]. Users with physical disabilities or chronic conditions often exhibit movement patterns that deviate significantly from the "average" user in the global model. Consequently, the global model tends to memorize these unique features, making these marginalized individuals far more susceptible to re-identification via MIA [46] or SIA [21].

Proceedings on Privacy Enhancing Technologies 2026(4), 587–607

https://doi.org/10.56553/popets-2026-0136

This constitutes severe representational harm: individuals already facing societal vulnerabilities are disproportionately exposed to privacy risks simply because their data is statistically unique. Without explicitly addressing this inequitable distribution of risk, FL risks reinforcing the very social inequalities it should theoretically bypass.

Proposed Approach. This paper addresses the critical challenge of ensuring equitable privacy protection by introducing F*in*P, a framework designed to measure and enforce fairness in privacy within FL. To mitigate disparate privacy leakage caused by heterogeneous data and varying local training dynamics [44], we propose a novel two-pronged approach operating at both the server and the client levels:

(1) **Server-Side Adaptive Aggregation:** We introduce an aggregation strategy that dynamically weights client model updates based on their estimated privacy risk. This prevents highly vulnerable, unique clients from disproportionately exposing their representations in the global model.
(2) **Client-Side Collaborative Overfitting Reduction:** Complementing the server-side intervention, we tackle the root cause of vulnerability—local overfitting. By estimating each client's relative overfitting using the maximum eigenvalue of its local model's Hessian, we incorporate this metric into a local regularization term bounded by the Lipschitz constant. This encourages clients to learn generalizable representations, directly reducing their individual susceptibility to privacy attacks.

By addressing both the symptoms (unequal risk aggregation) and the root causes (local overfitting) of privacy disparity, our combined approach achieves an improvement in fairness in privacy with a negligible impact on primary task performance.

Contributions. In summary, our key contributions are:

- We introduce a formal definition of fairness-in-privacy (F*in*P) tailored for real-world human-centric machine learning systems.
- We operationalize the measurement of F*in*P within Federated Learning, with a specific focus on quantifying vulnerabilities to source inference attacks (SIA).
- We propose a novel, two-pronged technical framework featuring adaptive server-side aggregation and client-side Hessian-based regularization to optimize F*in*P.
- We provide a comprehensive empirical evaluation demonstrating the efficacy of our approach using real-world dataset of a human activity recognition (HAR), and standard learning tasks (FEMNIST and CIFAR-10), confirming significant gains in fairness in privacy with minimal task utility loss.

2 Background and Related Work

Privacy in Human-Centric Systems. Ensuring privacy in human-centric machine learning systems presents inherent conflicts among service utility, computational cost, and personal privacy [51]. Without appropriate frameworks and incentives for secure data utilization, system deployments often face a dichotomy: policies become too restrictive, stifling innovation, or compromise private information, leading to adverse selection and a loss of public trust [15, 16, 26, 36]. Consequently, extensive research has focused on establishing privacy-preserving techniques and auditing platforms for human-in-the-loop and decision-making systems [2, 9, 13, 23, 39, 52, 53]. However, recognizing the "zero trust" reality that perfect privacy is often mathematically or practically unattainable, this paper examines privacy through an equity lens. We investigate how to ensure a fair distribution of harm when privacy leakage inevitably occurs, bridging the technical challenges of ML with the ethical imperatives of equitable protection.

Privacy Disparity and Inference Attacks in Federated Learning. Federated Learning (FL) enables the collaborative training of models without centralizing raw data [32], making it highly suitable for privacy-sensitive domains navigating regulations, such as GDPR [37, 58]. Despite its decentralized architecture, FL introduces new attack surfaces. The statistical heterogeneity inherent in FL does not only affect model performance — it systematically amplifies the disparity in privacy risk between participating clients. When client data distributions diverge significantly, models tend to memorize the unique features of outlier clients, making those clients disproportionately vulnerable to inference attacks. Foundational work on disparate vulnerability [29] has demonstrated that privacy leakage is rarely uniform; models disproportionately memorize and expose the data of underrepresented subgroups. Membership Inference Attacks (MIA) exploit this memorization by determining whether a specific data record was used during training, effectively probing the model's generalization boundary [20, 46]. Although MIA captures global membership leakage, it does not reveal which client contributed to the targeted record [17, 61]. Source Inference Attacks (SIA) extend this threat by identifying the exact originating client of a training record, directly exploiting localized overfitting caused by non-IID data distributions [21]. Despite pseudonymity, SIA poses a serious risk because mapping sensitive data to a distinct client node links it to a real-world context such as a device or location, effectively breaking anonymity. Beyond membership-based attacks, Property Inference Attacks (PIA) represent a distinct threat class that targets macro-level statistical properties of a client's local dataset — such as demographic distributions or class ratios — rather than individual records [34]. Unlike SIA and MIA, which are rooted in localized memorization and overfitting, PIA exploit the structural properties embedded in gradient updates during the aggregation process and, therefore, operate from a fundamentally different threat vantage point.

Crucially, the inherent statistical heterogeneity (Non-Independent and Identically Distributed (non-IID) data) of human-centric FL amplifies these vulnerabilities. When data distributions differ widely among clients, individual model updates become highly distinct. Malicious actors can exploit these distributional differences to trace memorized features to specific clients [62]. This distinctiveness is especially pronounced in datasets like Human Activity Recognition (HAR), where unique physical movements or conditions make outliers highly susceptible to SIAs, creating an unequal landscape of privacy risk.

Existing defenses such as Differential Privacy via DP-SGD [2] address these risks by injecting uniform noise into gradient updates. However, as demonstrated in this work, geometry-blind noise injection degrades global utility without resolving the disparity in

per-client privacy exposure; in fact, it can exacerbate unfairness by failing to mask the highly distinct updates of severely skewed clients while unnecessarily penalizing well-represented ones.

From Utility Fairness to Privacy Fairness. Fairness-aware federated learning approaches have emerged to address performance inequity between clients, including techniques such as FedAvg [22] with client reweighting, FairFed [14], and personalized federated learning methods [7, 54], as well as recent work exploring the tension between fairness interventions and privacy leakage [4]. Recent work also investigated the use of FL for pluralistic alignment of Large Language Models with a fairness lens of aggregation of human preference [47–49].

However, these approaches primarily frame fairness in terms of equitable model accuracy or predictive benefit across client groups, and do not explicitly formalize or optimize for the equitable distribution of privacy risk. The critical dimension of client-level privacy equity — ensuring that no individual client disproportionately bears the burden of privacy vulnerability due to the statistical uniqueness of their data — remains largely unaddressed in the prior literature. F*in*Puniquely targets this gap by formalizing fairness-in-privacy as a first-class objective, directly mitigating memorization-based privacy disparities caused by non-IID data distributions **exploited by SIA**, and distinguishing itself from existing defenses that focus either on global privacy guarantees or on utility fairness without addressing the structural inequity of per-client privacy exposure.

3 Threat Model and Problem Formulation

Although Federated Learning (FL) mitigates the risks of centralized data storage, it remains vulnerable to inference attacks launched by the aggregating server.

Assumptions and Scope. F*in*P operates under two core assumptions. First, the server is "honest-but-curious": it faithfully executes the FL protocol but actively attempts to infer sensitive client information from received model updates and curvature metrics. Second, clients are honest and cooperative: they follow the protocol and submit genuine updates without manipulation. Under these assumptions, F*in*P targets inference attacks driven by localized overfitting and memorization in non-IID settings — specifically Source Inference Attacks (SIA), which identify the originating client. SIA is particularly risky because mapping a sensitive record to a specific client node links it to a real-world context such as a device or location, breaking pseudonymity through contextual profiling. F*in*P suppresses this memorization by flattening the loss landscape, reducing the attacker's confidence. F*in*P does not address malicious server behavior, manipulative clients, gradient reconstruction attacks, or Property Inference Attacks, as these operate outside the memorization-based threat model.

Threat Model. A primary privacy threat in this decentralized setting is a two-stage inference attack executed by the server: a Membership Inference Attack (MIA) followed by a Source Inference Attack (SIA).

- **Membership Inference Attack (MIA):** The server first attempts to determine if a specific target data point x was utilized in the training set D_{θ_g} of the global model θ_g. This is formally defined as the probability that x belongs to the training data:

$$MIA(\theta_g, x) = P(x \in D_{\theta_g})$$

- **Source Inference Attack (SIA):** If the MIA successfully indicates that x was part of the training corpus, the server proceeds to identify the origin of the data. The server analyzes the updates of the local model θ_i to determine the probability that a specific client C_i is the source of the record x:

$$SIA(\theta_i, x) = P(C_i \mid x, \theta_i)$$

As demonstrated in the prior literature [21], the concatenation of these attacks compromises the anonymity of the client. Furthermore, auditing and completely preventing MIA and SIA within FL presents inherent mathematical and practical limitations [6].

Problem Formulation: Disparate Privacy Risk. Rather than attempting the mathematically fraught task of completely eliminating SIA, our work focuses on the resulting *disparity* in privacy risk across the client federation. In heterogeneous (non-IID) FL environments, clients with unique or outlier data distributions are highly prone to local overfitting during training. This overfitting causes their local model updates (θ_i) to disproportionately memorize their unique data points, rendering them significantly more vulnerable to SIAs than clients with more "average" data. To exploit this vulnerability, The attacker calculates the prediction loss of the target record across every individual client's model on the server. The reason is that the local model belonging to the true source will naturally yield the lowest prediction loss. The attacker therefore simply identifies the client with the minimum loss as the most probable source of the target record. This creates a systemic inequity in privacy protection.

Given this threat model, our objective is to mitigate the disparate impact of SIAs by ensuring an equitable distribution of the inherent privacy risks between all clients. To achieve this F*in*P, our framework pursues two core objectives:

(O1): Addressing the Symptoms (Server-Side): Develop a dynamic aggregation methodology on the server that quantifies client vulnerability and adjusts their contributions to the global model, ensuring a fair distribution of privacy risk without degrading global utility.

(O2): Addressing the Root Causes (Client-Side): Provide actionable feedback mechanisms to highly vulnerable clients, enabling them to apply targeted local regularization. This empowers clients to reduce their local overfitting, thereby lowering their individual susceptibility to SIAs and improving the overall fairness in privacy of the system.

4 The F*in*P Framework in Federated Learning

The core objective of the F*in*P framework is to ensure that privacy risks are equitably distributed among all participating clients, preventing any single client from bearing a disproportionate burden of vulnerability. In practical federated learning (FL) deployments, the profound heterogeneity of client data distributions, varying computational resources, and distinct local training dynamics naturally lead to stark disparities in privacy risk. Traditional privacy defenses that merely attempt to lower the *average* risk across the

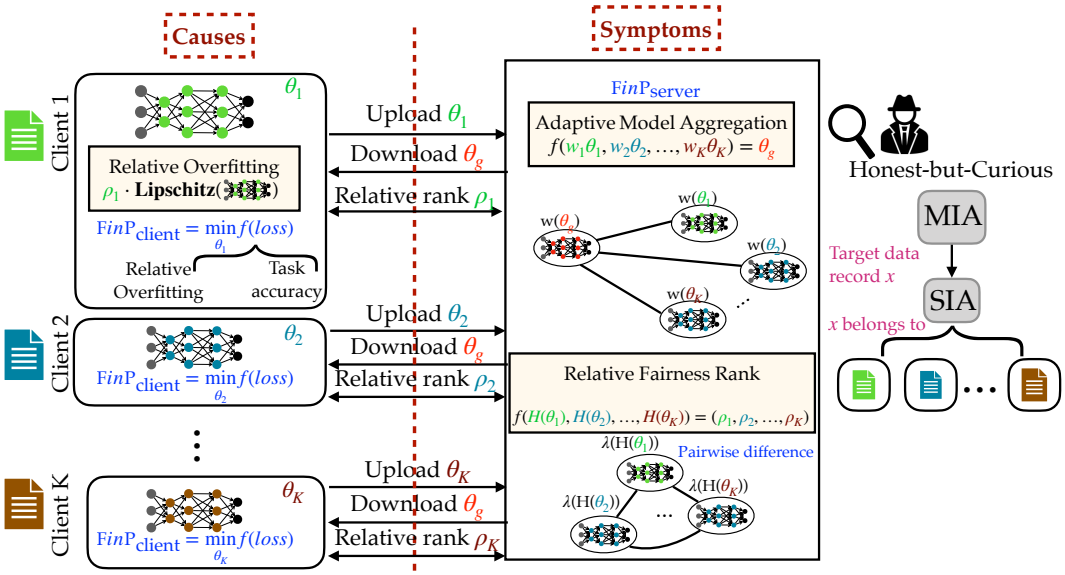


Figure 1: Overview of the F*in*P framework. The system achieves fairness-in-privacy by establishing a feedback loop that simultaneously addresses the symptoms of privacy disparity (via server-side adaptive aggregation) and the root causes (via client-side collaborative overfitting reduction).

federation are insufficient for human-centric systems; effective protection demands that we proactively prevent disproportionate risk burdens on minority or outlier clients.

To achieve true fairness-in-privacy, F*in*P operationalizes a holistic two-pronged approach, as illustrated in Figure 1. We argue that systemic privacy equity requires simultaneous intervention at both the server level (during model aggregation) and the client level (during local training).

Addressing the Symptoms (Server-Side Intervention). Server-side interventions are critical for managing the immediate impacts of existing privacy disparities. F*in*P introduces an adaptive aggregation mechanism that dynamically scales the weights of client updates based on their estimated privacy risk. By systematically reducing the aggregation weight of highly vulnerable clients, we prevent their overfit, memorized representations from unduly influencing the global model. This limits the exposure of their unique data to the rest of the federation and restricts an adversary's ability to execute Source Inference Attacks (SIAs). However, while adaptive aggregation effectively mitigates the immediate symptoms of privacy unfairness, it does not cure the underlying vulnerability of clients themselves.

Addressing the Root Causes (Client-Side Intervention). To dismantle the root cause of privacy disparity, we must look to local training dynamics—specifically, local overfitting. Extensive prior literature has established the theoretical and empirical connection between model overfitting and increased privacy risks [21, 46, 57]. Because overfit models memorize specific properties of their training data rather than learning generalized features, they serve as the primary vulnerability exploited by MIAs and SIAs. Consequently, when a client's local model heavily overfits its non-IID data, that client's susceptibility to SIAs spikes, creating the very disparity we aim to eliminate.

To counteract this, F*in*P introduces a **collaborative overfitting reduction strategy** at the client level. This proactive mechanism ranks clients based on an estimation of their relative overfitting compared to the federation. By incorporating this server-provided rank into a customized local regularization scheme, F*in*P forces vulnerable clients to learn more generalizable representations. This actively suppresses the memorization of outlier data, shrinking the initial disparity in privacy risk *before* the models are ever transmitted to the server.

The Closed-Loop System. Ultimately, F*in*P functions as a tightly coupled closed-loop system. The server utilizes incoming client models to calculate adaptive aggregation weights and relative overfitting ranks; in turn, the clients utilize this targeted feedback from the server to regularize their subsequent local training rounds. By simultaneously addressing both the symptoms and the root causes of privacy disparity, F*in*P establishes a structurally fair FL environment.

We formally define the mathematical mechanics of the client component in Section 4.1 and the server component in Section 4.2.

4.1 Addressing the Root Causes: Client-Side Collaborative Regularization

Although server-side aggregation mitigates the immediate symptoms of privacy unfairness, achieving structural fairness-in-privacy (F*in*P) requires dismantling the root cause: local overfitting. In a heterogeneous FL system with K clients, non-IID data create distributional shifts that cause certain outlier clients to overfit more

than others. Because overfitting directly correlates with the memorization of training data, these specific clients become disproportionately susceptible to Source Inference Attacks (SIAs). Our goal is to enforce an equitable privacy risk by proactively reducing this localized memorization.

To achieve this, we introduce a collaborative, feedback-driven regularization strategy. Recent literature demonstrates that the top Hessian eigenvalue ($\lambda_{\max}$) and the Hessian trace (H_T) are powerful metrics to characterize the loss landscape and the generalization capabilities of neural networks [25]. A sharp loss landscape (indicated by high $\lambda_{\max}$ and H_T) implies severe overfitting and a high vulnerability to weight perturbations, while lower values indicate smoother, more generalized and thus more privacy-preserving local models.

Because we are specifically interested in F*in*P—which focuses on *disparity* rather than absolute risk—we quantify how much more vulnerable a client is compared to its peers. We define each client's relative overfitting by calculating the average pairwise difference across the top Hessian eigenvalue ($\lambda^k_{\max}$) and Hessian trace (H^k_T) of their local model θ_k against all other j clients:

$$\bar{\Delta}_k = \frac{1}{K-1} \sum_{j=1, j\neq k}^{K} |\lambda^k_{\max} - \lambda^j_{\max}| \tag{1}$$

$$\bar{H}_k = \frac{1}{K-1} \sum_{j=1, j\neq k}^{K} |H^k_T - H^j_T| \tag{2}$$

These metrics are then normalized and combined to compute the client's *Overfitting Relative Rank* (ρ_k), which serves as a proxy for their relative privacy risk disparity:

$$\rho_k = \frac{1}{2}\left(\frac{\bar{\Delta}_k}{\max(\bar{\Delta})} + \frac{\bar{H}_k}{\max(\bar{H})}\right) \tag{3}$$

Clients compute their own top Hessian eigenvalue and Hessian trace locally during training. The server only computes the relative rank ρ_k after receiving these scalar metrics. The server then transmits this scalar rank ρ_k to client k as a targeted feedback.

Upon receiving ρ_k, the client incorporates this overfitting rank into their subsequent local training rounds using a dynamic regularization term. We utilize the Lipschitz constant, approximated by the spectral norm of the Jacobian matrix ($||J_k||$) [31]. Constraining the Lipschitz constant forces the model to learn smoother functions, thereby resisting the memorization of outlier data points. The modified local loss function for client k becomes:

$$\mathcal{L}^*_k = \mathcal{L}_k + \beta \cdot \rho_k \cdot ||J_k|| \tag{4}$$

$$\text{F}in\text{P}_{\text{client}} = \min_{\theta_k} \mathcal{L}^*_k \tag{5}$$

Here, $\mathcal{L}_k$ is the original local loss function, $||J_k||$ is the Jacobian penalty, and β is a task-dependent parameter that adjusts the baseline impact of Lipschitz regularization. The critical innovation is the inclusion of ρ_k, which acts as an **adaptive** multiplier. In each communication round, clients with a higher relative overfitting rank are subjected to mathematically stronger regularization. This collaborative, penalty-based approach effectively throttles the most vulnerable clients, forcing them to generalize, and thereby actively shrinking the privacy risk disparity across the federation.

Computational Practicality and Privacy Guarantee. Computing and transmitting the exact Hessian matrix is computationally prohibitive for deep neural networks and violates the strict data-minimization guarantee of FL. To ensure that our framework remains scalable, clients do not compute the full Hessian; instead, they efficiently estimate only the top eigenvalue and the trace.

To minimize local computational overhead on edge devices — including battery-constrained mobile phones and resource-limited edge nodes — clients employ efficient Hessian-vector product (HVP) methods, specifically power iteration for the top eigenvalue and Hutchinson's method for the trace, entirely avoiding full Hessian instantiation. Critically, because F*in*P's fairness objective relies solely on the relative ranking of client vulnerability rather than the absolute values of eigenvalues, these lightweight approximations are fully sufficient to achieve privacy equity without requiring exact Hessian computation. This design choice substantially reduces computational cost per-round, making F*in*P practical for deployment on resource-constrained hardware typical in the real-world federated participants.

F*in*P does not require additional communication rounds to transmit these metrics. The scalar sharpness values are simply appended to and transmitted concurrently with the standard local model weights during existing federated communication rounds, incurring only negligible additional bandwidth overhead and fully preserving the communication efficiency of standard federated learning.

Furthermore, while F*in*P requires clients to transmit these local curvature metrics with their model updates(θ_k), this does not violate the FL privacy principles. These metrics are one-dimensional scalars representing macroscopic geometric properties of the local loss landscape. Because mapping a local dataset to these two aggregate scalars is a heavily non-injective and non-invertible transformation, transmitting them provides negligible mutual information regarding the raw training data. It is mathematically intractable for an honest-but-curious server to reconstruct raw local data or execute gradient-inversion attacks relying solely on these curvature scalars.

4.2 Addressing the Symptoms: Server-Side Adaptive Aggregation

The core symptom of privacy unfairness in FL is the pronounced disparity in vulnerability to Source Inference Attacks (SIAs) across the client federation. While client-side regularization (detailed in Section 4.1) addresses the root cause of local overfitting, the server must simultaneously manage the observable symptoms present in the incoming model updates. Our objective is to design a server-side aggregation strategy that actively quantifies this privacy risk and adjusts the global model update to enforce an equitable distribution of vulnerability. We explicitly define this objective as minimizing the variance of the privacy risk distribution:

$$\begin{aligned} \text{F}in\text{P}_{\text{server}} &= \min_{\mathbf{w}\in\mathcal{W}} \text{Disparity}(\mathbf{p}(\mathbf{w})) \\ &= \min_{\mathbf{w}\in\mathcal{W}} \left\| \mathbf{p}(\mathbf{w}) - \frac{1}{K}\mathbb{1}^T\mathbf{p}(\mathbf{w}) \otimes \mathbb{1} \right\| + \left\| \frac{1}{K}\mathbb{1}^T\mathbf{p}(\mathbf{w}) \right\| \end{aligned} \tag{6}$$

Here, $\mathbf{p}(\mathbf{w}) = [p_1(\mathbf{w})), \ldots, p_K(\mathbf{w}))]$ is the vector of privacy risk indicators across the K clients given the aggregation weights $\mathbf{w}$,

$\mathbb{1}$ is a vector of ones of length K, and $\mathcal{W} = \{\mathbf{w} \in \mathbb{R}^K \mid \sum_{k=1}^{K} w_k = 1, w_k \geq 0 \; \forall k\}$ is the set of valid aggregation weights.

To satisfy the F*in*P$_{\text{server}}$ objective, the server must aggregate local models in a manner that minimizes this disparity metric without severely degrading global utility. We evaluate two strategies for achieving this:

(1) **Strategy 1: Adaptive Lightweight Aggregation (ALA) (Proposed Solution)**
Our primary method acts as a highly efficient, heuristic solution to the disparity minimization objective, specifically designed to be scalable for real-world, resource-constrained environments.
- **Risk Symptom Quantification (p_k):** The server re-utilizes the normalized Overfitting Relative Rank (ρ_k) described in Equation 3, as a proxy for the privacy risk symptom ($p_k \propto \rho_k$). A higher ρ_k signifies severe local overfitting and greater vulnerability to SIAs.
- **The Solution Mechanism:** The ALA technique achieves the disparity minimization objective by dynamically calculating the specific aggregation weights $\alpha_k \in \mathcal{W}$ to be inversely proportional to the risk symptom ρ_k. High-risk clients are explicitly granted less influence over the global model, proactively driving the global model toward a state where the risk vector $\mathbf{p}$ is balanced:

$$\theta_{global} \leftarrow \sum_{k=1}^{K} \alpha_k \theta_k, \quad \text{where} \quad \alpha_k = \frac{1 - \rho_k}{\sum_{i=1}^{K}(1 - \rho_i)}, \tag{7}$$

where θ_{global} is the global model parameters and θ_k is the local model parameters for client k. This inverse weighting mechanism provides the exact control necessary to solve Equation 6. By actively restricting the impact of the most vulnerable clients, ALA prevents their memorized, outlier data from being embedded into the global model. This enforces an equitable distribution of privacy risk without introducing significant computational overhead.

(2) **Strategy 2: Server-Side Risk Quantification (PCA-based Baseline)**
For comparison only, a traditional alternative for quantifying the symptom of risk p_k is computing the Principal Component Analysis (PCA) distance between each client's model update and the global average update [10]. This distance acts as a purely server-side proxy for outlier behavior.
If this PCA distance (PCA_d) is used as the risk vector $\mathbf{p}(\mathbf{w})$, the aggregation process would similarly solve the FinP$_{\text{server}}$ disparity minimization objective by adjusting weights based on these distances. However, while formally sound, computing the PCA distance for the high-dimensional model updates typical in modern human-centric FL is computationally expensive [5]. The resulting latency makes it highly impractical for real-world, large-scale FL deployments, reinforcing that the lightweight, Hessian-proxy ALA strategy is the superior and preferred mitigation method for our framework.

5 Evaluation

5.1 Federated Learning System Setup

To demonstrate the real-world applicability of F*in*P to mitigate privacy disparities, we evaluated our framework in two distinct federated learning deployments. We specifically selected datasets that reflect the highly heterogeneous, non-IID nature of practical edge-computing and mobile sensing environments, where outlier clients are naturally prone to local overfitting and heightened vulnerability to Source Inference Attacks (SIAs). In particular, we structured our evaluation around two primary case studies: Human Activity Recognition (HAR) as a human-centric, privacy-sensitive application, and CIFAR-10 and FEMNIST as a standard benchmark for controlled, heterogeneous visual data.

Baselines and Ablations. To rigorously isolate the contributions of our framework, we compare the three standard and state-of-the-art baselines in various configurations as follows:

• **HAR dataset:** we evaluated five configurations: (1) Baseline FL: Standard FedAvg aggregation, adapted from prior SIA literature [22]. (2) F*in*P$_{\text{server}}$: Isolating our adaptive server-side aggregation without client collaboration (Equation 6). (3) F*in*P$_{\text{client}}$: Isolating our collaborative client-side regularization without adaptive server aggregation. (4) F*in*P-Full: The complete proposed closed-loop framework. (5) Differential Privacy (DP-SGD) [2]: Applying DP-SGD against SIA attack with noise multiplier $\sigma_{DP} \in \{0.5, 1, 2\}$.

• **CIFAR-10 dataset:** we evaluated: (1) Baseline FL: Standard FedAvg from [22]. (2) FedAlign: A state-of-the-art FL method specifically designed to address non-IID data heterogeneity [35]. (3) F*in*P (ResNet56): Our framework applied to the same ResNet architecture used by FedAlign for a direct performance comparison. (4) F*in*P Ablation (CNN): Our framework applied to a standard CNN to evaluate the sensitivity of the impact factor $\beta \in \{0.05, 0.1, 0.3, 0.5\}$, as described in Equation 4. (5) Differential Privacy (DP-SGD) [2]: Applying DP-SGD against SIA attack with noise multiplier $\sigma_{DP} \in \{0.5, 1, 2\}$. (6) Additionally, we evaluate F*in*P using both the Adaptive Lightweight Aggregation (ALA) and the PCA-based server-side strategies to validate computational practicality.

• **FEMNIST dataset:** we evaluate: (1) Baseline FL: Standard FedAvg from [22]. (2) F*in*P: Our framework with the impact factor $\beta \in \{0.5, 0.75, 1\}$, as described in Equation 4. (3) Differential Privacy (DP-SGD) [2]: Applying DP-SGD against SIA attack with noise multiplier $\sigma_{DP} \in \{0.5, 1, 2\}$. (4) MIA attack: Launching White-Box Membership Inference Attack (MIA) to evaluate F*in*Pperformance vs. Baseline (FedAvg) and DP-SGD.

Setup for HAR. The UCI HAR Dataset [41] serves as our real-world application. In modern ubiquitous computing and mobile sensing environments, continuous sensor data (such as accelerometer and gyroscope readings) is routinely collected from users' personal smartphones or wearable devices. Because physical behaviors, daily routines, and gaits are unique to an individual, HAR data is inherently non-IID and deeply sensitive [8, 55]. Preserving the natural user-based partition ensures the authentic structure of the data is maintained; models trained on users with distinct or outlier behaviors are prime targets for SIAs.

The HAR dataset comprises data from 30 subjects (aged 19–48), treated as $K = 30$ individual FL clients, performing six daily activities. We allocated 70% of each client's data for training (using 5-fold cross-validation) and 30% for independent testing. All data was preprocessed with noise filters and segmented using a 2.56-second sliding window with a 50% overlap. We utilized a Temporal Convolutional Network (TCN) [3] for the architecture of the local models, which captures temporal dependencies via causal convolutions. We trained over 20 global communication rounds using FedAvg. Clients trained locally with a batch size of 64, a learning rate of 0.001 (Adam optimizer), 1 local epoch per round, and an impact factor $\beta = 2$. More details on the HAR dataset are in Appendix A.

To establish a rigorous Differential Privacy (DP) baseline, we implemented Record-Level DP-SGD [2] utilizing the Opacus library. For all DP-enabled experiments, we transitioned from the Adam optimizer to Stochastic Gradient Descent (SGD) with a learning rate of 0.01, while keeping all other federated configurations (batch size, local epochs, and communication rounds, etc) identical to the standard baseline. A critical challenge in DP-SGD is selecting an optimal clipping threshold (C) that aggressively bounds the sensitivity of extreme outliers while preserving the underlying gradient signal. To ensure an empirical and fair calibration, we executed a non-private warm-up run of the global model for a single epoch and recorded the unclipped L_2 gradient norms across all client batches. We fixed the clipping threshold at C=5, which closely approximates the 90th percentile of the unclipped gradients. By adopting this empirically derived threshold, we formally bounded the model's sensitivity while maintaining optimal learning momentum. To comprehensively evaluate the resulting privacy-utility trade-off, we swept the Gaussian noise multiplier $\sigma_{DP} \in \{0.5, 1, 2\}$, corresponding to low, medium, and high privacy budgets, respectively.

Setup for CIFAR-10. To stress-test F*in*P under mathematically controlled degrees of data imbalance—simulating visual heterogeneity across a network of personal devices or edge cameras—we utilized the CIFAR-10 dataset [28]. We partitioned the data among $K = 10$ clients using a Dirichlet distribution $Dir(\alpha)$ with $\alpha = 0.5$, a standard approach in SIA literature for simulating unbalanced subsets [21, 35]. Figure 2 visualizes the severe data skew among clients when tested at $\alpha = 0.1$. We employed ResNet56 [19] for the primary evaluation to match the FedAlign baseline, and a standard CNN [22] for the β ablation studies. Similar to the HAR setup, training occurred over 20 global rounds. The default impact factor for the client-side regularization was set to $\beta = 0.3$. We deployed DP-SGD with CNN and same configuration. Clipping threshold is fixed at $C = 1.75$ ($\approx$ the 90th percentile of the unclipped gradients) and noise multiplier $\sigma_{DP} \in \{0.5, 1, 2\}$ for low, medium, high privacy budgets. More details on the CIFAR dataset are in Appendix A.

Setup for FEMNIST. We evaluate on FEMNIST Dataset additionally, a handwritten character recognition benchmark with 62 classes and natural non-IID structure induced by writer identity. We simulate 10 federated clients by sampling 10 distinct writers; each client owns only that writer's examples. We employ a lightweight CNN as the training model. More details are in Appendix A.

Empirical Evaluation Setup. To simulate this attack pipeline for our empirical evaluation, we randomly sampled training data from each client's local dataset and pooled them into a unified dataset of target records. By launching the SIA against these known records and calculating the attack success rate, we can directly measure both the absolute privacy leakage and, crucially, the *disparity* of this vulnerability across the federation.

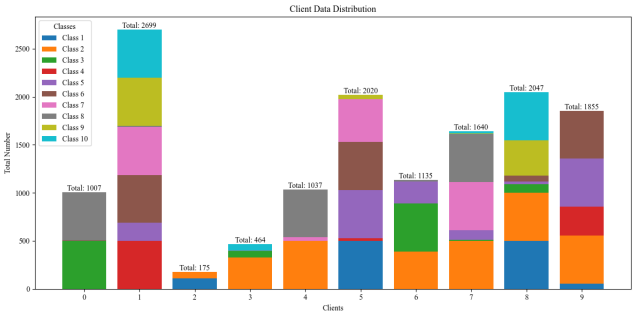


Figure 2: CIFAR dataset profile demonstrating severe non-IID data distribution among clients after Dirichlet sampling ($\alpha = 0.1$).

Hardware Profile. All experiments were conducted on a single NVIDIA A30 GPU with 24 GB of memory, simulating the central aggregating server's computational environment.

5.2 Metrics for Comparison

To comprehensively evaluate the effectiveness of F*in*P in achieving fairness in privacy, we employ the following key metrics encompassing privacy fairness, absolute privacy, utility, and efficiency.

1. Privacy Fairness Metrics (Disparity Quantification). To quantify "Fairness in Privacy," we must measure the dispersion of Source Inference Attack (SIA) vulnerability across the K participating clients. We introduce three metrics to capture both the accuracy and the confidence of the adversary's attack.

- **Disparity in SIA Accuracy (CoV & FI):** To formalize the notion of privacy fairness, we draw on the concept of group fairness as established in literature [24]. Applying the Coefficient of Variation (CoV) to client-level privacy risk treats each client as a protected group, where a low CoV indicates an equitable distribution of privacy exposure across the federation. Unlike minimax or worst-case approaches that optimize solely for the single most vulnerable client, minimizing CoV ensures system-wide privacy equity, preventing minority clients from suffering disproportionately high privacy vulnerability while preserving equitable protection across the entire federation.
  For a federation of K clients, the empirical SIA accuracy for a specific client i is calculated as:

$$\text{SIA}_{acc_i} = \frac{\text{Number of correct SIA identifications for client } i}{\text{Total target records attributed to client } i} \tag{8}$$

  Let $\mu = \frac{1}{K}\sum_{i=1}^{K} \text{SIA}_{acc_i}$ be the mean SIA accuracy across the federation, and σ be the standard deviation. The CoV is computed as:

$$\text{CoV}(\text{SIA}_{acc}) = \frac{\sigma}{\mu} = \frac{\sqrt{\frac{1}{K}\sum_{i=1}^{K}(\text{SIA}_{acc_i} - \mu)^2}}{\mu} \tag{9}$$

  To map this variance to an intuitive fairness percentage between 0 and 1 (where 1 represents perfect equity), we apply the Fairness

Index (FI) transformation:

$$\mathrm{FI}(\mathrm{SIA}_{acc}) = \frac{1}{1 + \mathrm{CoV}(\mathrm{SIA}_{acc})^2} \tag{10}$$

A higher FI value indicates that the vulnerability is equitably distributed, preventing any single user from becoming an extreme outlier target.

- **Equal Opportunity Difference in Privacy (EOD):** Because SIA_{acc_i} represents the empirical True Positive Rate (TPR) of the adversary correctly identifying client i, we can map traditional algorithmic group fairness metrics—specifically Equal Opportunity—directly to privacy [18]. By treating individual clients as protected groups, Equal Opportunity mandates that the TPR of the adversary's attack should be uniform across the federation. We quantify the maximum violation of this principle using the Equal Opportunity Difference (EOD):

$$EOD = \max_{i,j} |\mathrm{SIA}_{acc_i} - \mathrm{SIA}_{acc_j}| \quad \forall i, j \in \{1, \ldots, K\} \tag{11}$$

A lower EOD demonstrates that the adversary cannot exploit specific outlier users more easily than the average user.

- **Disparity in Attack Confidence (Loss CoV):** While SIA accuracy measures discrete attack success, a rigorous privacy evaluation must also address the adversary's confidence. SIAs exploit the correlation between localized memorization and low target-record prediction loss. Clients with highly overfitted local models will exhibit significantly lower prediction losses on their target records, granting the adversary high statistical confidence. F*in*P actively aims to flatten this inter-client loss landscape. We quantify this mitigation using the CoV and FI of the client prediction losses on target records, denoted as CoV(Loss) and FI(Loss), defined similarly to Equations 9 and 10.

2. Absolute Privacy Metrics. Achieving fairness by artificially degrading the privacy of secure clients to match the most vulnerable ones is a failure mode in privacy engineering. To verify that F*in*P enforces equitable privacy without increasing overall leakage, we track two global metrics:

- **Mean(SIA_{acc}):** The average attack success rate across all clients and communication rounds.
- **Max(SIA_{acc}):** The absolute highest attack success rate observed for any single client. Lowering this maximum vulnerability is critical, as it proves we are successfully protecting the "weakest links" in the network.

3. Utility and Efficiency Metrics. Finally, any deployable Privacy-Enhancing Technology must maintain the operational viability of the underlying system.

- **Global Model Utility:** We evaluate the ultimate cost of privacy by measuring the testing accuracy of the converged global model on the primary learning task.
- **System Efficiency:** We measure the impact of the F*in*P closed-loop regularization on the convergence dynamics, quantified by the total number of communication rounds required to reach convergence compared to standard baselines.

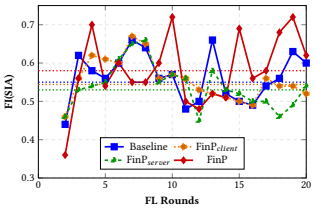


Figure 3: Progression of the Fairness Index (FI(SIA_{acc})) over communication rounds in HAR. F*in*P (PCA) maintains a more equitable distribution of privacy risk compared to the baseline.

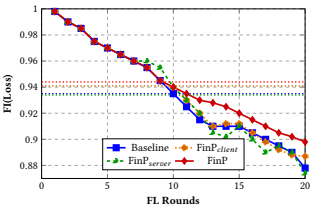


Figure 4: Disparity of prediction loss among clients in HAR (FI(Loss)). F*in*P (PCA) improves the inter-client loss landscape, reducing the adversary's confidence in identifying source records.

5.3 Evaluating F*in*P on the Human Activity Recognition (HAR) Dataset

We first evaluate the efficacy of F*in*P in a realistic, human-centric mobile sensing environment using the HAR dataset (More on the dataset setup is in Appendix A). The comprehensive results across all metrics are summarized in Table 1. Our findings confirm that F*in*P successfully structurally flattens the privacy risk landscape, significantly reducing vulnerability disparities across the federation while maintaining high global utility as explained below.

1. Mitigating Disparity in SIA Vulnerability (Fairness). The primary objective of F*in*P is to ensure that no single user disproportionately bears the privacy cost of participating in the federation. Our framework demonstrates a marked improvement in the equitable distribution of privacy risk (Figure 3).

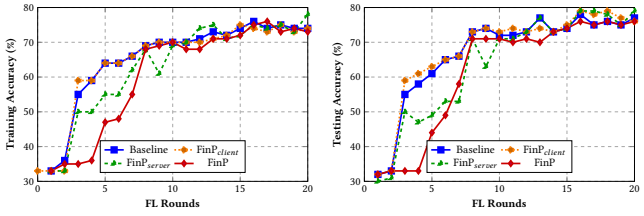
- **Variance Reduction:** Compared to the Baseline FL setup [22], the full F*in*P framework reduces the Coefficient of Variation for SIA accuracy (CoV(SIA_{acc})) from 0.94 to 0.89 (a 5.32% reduction). This corresponds to a 5.56% improvement in the overall Fairness Index (FI(SIA_{acc})), proving that our adaptive regularization actively constrains the most vulnerable outlier models.
- **Equal Opportunity Improvement:** By treating individual clients as protected entities, F*in*P achieves a 5.45% reduction in the Equal Opportunity Difference (EOD), bringing it down to 0.52 (Visual illustration in Figure 22 in Appendix B). This critical metric confirms that the adversary's True Positive Rate (TPR) is significantly more uniform across the network, fundamentally limiting their ability to selectively target distinct individuals based on their unique physical activities.

2. Reducing Attack Confidence (Loss Disparity). Beyond binary attack success, F*in*P effectively obscures the statistical signals adversaries rely on. By penalizing local memorization, the framework flattens the landscape of prediction losses on target records (Figure 4). F*in*P achieves a highly equitable FI(Loss) of 0.942 (up from the baseline's 0.938). While the absolute gain appears incremental due to the high baseline saturation, this reduction in CoV(Loss) (down to 0.235) confirms that the adversary is denied the artificially low-loss confidence signals previously emitted by overfitted outlier clients.

3. Absolute Privacy Leakage. Crucially, F*in*P achieves these fairness gains without artificially degrading the privacy of secure

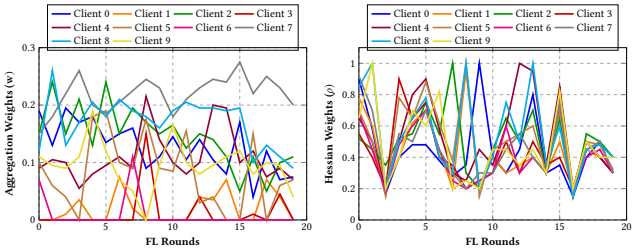
Table 1: Results of HAR using the two approaches (ALA and PCA) of server aggregation in comparison to the Baseline [22] and DP-SGD [2].

	Accuracy (%)		Privacy Metrics (%)		Fairness Metrics			Efficiency
	Train	Test	Mean (SIA_{acc}) ↓	Max (SIA_{acc})↓	CoV(SIA_{acc})/FI(SIA_{acc})	CoV(loss)/FI(Loss)	EOD ↓	Conv.round
Baseline [22]	74.10	76.97	19.34	22.40	0.94/0.54	0.244/0.938	0.48	9
$FinP_{client}$	73.39	75.86	18.87	23.30	0.97/0.53	0.237/0.941	0.49	9
$FinP_{server}$ (PCA)	**75.01**	**77.84**	**18.57**	**21.60**	0.99/0.52	0.246/0.937	0.54	11
F*in*P (PCA)	72.73	75.22	19.70	23.50	**0.89/0.57**	0.235/0.942	0.48	10
$FinP_{server}$ (ALA)	71.57	73.82	19.55	26.10	0.93/0.55	**0.220/0.948**	0.55	11
F*in*P (ALA)	72.83	76.09	19.29	22.00	**0.89/0.57**	0.235/0.942	0.48	10
DP-SGD (ϵ, δ)								
(99, 10^{-5})	45.10	45.13	17.88	22.9	1.05/0.48	0.032/0.999	0.49	9
(32, 10^{-5})	41.19	42.27	16.63	18.7	1.15/0.43	0.008/1.000	0.50	9
(12, 10^{-5})	44.37	44.66	17.32	21.2	1.05/0.48	0.032/0.999	0.50	6



(a) Global model training accuracy. **(b) Global model testing accuracy.**

Figure 5: Global model classification accuracy using HAR dataset. F*in*P (PCA) maintains competitive accuracy with minimal convergence delay.



(a) Server-side aggregation weights (w). **(b) Client-side relative overfitting rank (ρ_k).**

Figure 6: The dynamic interplay of F*in*P's closed-loop architecture over communication rounds in HAR.

clients. While we observe a negligible marginal increase of < 1.1% in Mean(SIA_{acc}) and < 0.4% in Max(SIA_{acc}), these fluctuations are statistically insignificant relative to the variance reduction. The network does not experience a "race to the bottom"; rather, the distribution of protection is demonstrably stabilized.

4. The Cost of Privacy: Utility and Efficiency Trade-offs. A robust Privacy-Enhancing Technology must not destroy the operational value of the system. As shown in Figure 5, F*in*P maintains highly competitive global classification utility. The global model's testing accuracy experiences only a minimal 1.75% degradation. Furthermore, the closed-loop regularization introduces only a marginal delay in convergence, requiring ≈ 10 communication rounds to stabilize compared to the baseline's ≈ 9 rounds. This confirms that F*in*P is highly viable for real-world deployment.

5. Component Ablation: Server vs. Client Contributions. To understand the mechanics of our framework, we isolated the individual contributions of the server-side and client-side components:

- **Isolated Server Impact:** Deploying only the server-side adaptive aggregation ($FinP_{server}$ with PCA) revealed nuanced behavior. While it successfully reduced the variation in global model distance (PCA_d) by 1.3% and marginally lowered absolute SIA metrics, it caused a slight negative shift in the fairness indices (FI(SIA_{acc}) dropped by 0.02). This proves that purely server-side, reactive re-weighting addresses the symptom of model divergence but is insufficient to enforce structural fairness in privacy.
- **The Synergy of the Closed Loop:** Combining the server aggregation with the client-side collaborative regularization ($FinP_{client}$) yielded the highest fairness gains. The server-side re-weighting (Equation 6) acts synergistically with the client-side adaptive regularizer (ρ_k, Equation 5). Figure 6 visualizes this dynamic interplay over time.
- **Computational Practicality of ALA:** When replacing the expensive PCA computation with our proposed Adaptive Lightweight Aggregation (ALA) proxy, F*in*P achieved identical fairness improvements while consuming only 17% of the computation time required by the PCA baseline. This validates ALA as the superior, highly scalable mechanism for real-world execution. Visual illustration for the results in Table 1 for HAR with ALA are listed in Appendix B.2.

6- Differential Privacy (DP-SGD). In our HAR dataset evaluation, Differential Privacy (DP) with medium noise budget setup (32, 10^{-5}) drastically decreases model testing accuracy by 34.7% while yielding only a marginal 2.7% decrease in average Source Inference Attack (SIA) accuracy. Furthermore, DP worsens the FI(SIA_{acc}) by 11.1% compared to the baseline, demonstrating that despite slightly increasing overall protection, it fails to improve empirical privacy fairness. Highly skewed, non-IID datasets make equitable privacy protection fundamentally challenging. Specifically, DP induces significantly higher client losses compared to our F*in*P method, driving the drastic utility drop. Ultimately, even though DP forces client losses to become uniformly high and closely clustered, empirical privacy fairness does not improve. Visual illustration for the results in Table 1 for HAR with DP-SGD are listed in Appendix B.3.

Finally, empirical analysis of the local loss landscapes (Figure 33 in Appendix B.4) validated our theoretical foundation. We observed a near-perfect Spearman's rank correlation (≈ 1) between the top Hessian eigenvalue (λ_{max}) and the Hessian trace (H_T). Because both

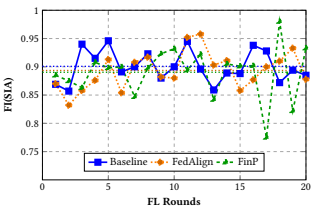


Figure 7: Disparity of SIA accuracy (FI(SIA)) among clients using CIFAR-10 dataset with ResNet and F*in*P with PCA.

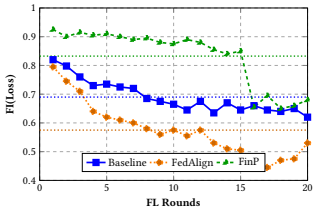


Figure 8: Disparity of prediction loss FI(Loss) among clients using CIFAR-10 dataset with ResNet and F*in*P with PCA.

metrics reliably indicate local overfitting, they are equally weighted in our calculation of the relative overfitting rank (Equation 2).

5.4 Evaluating F*in*P on the CIFAR-10 Dataset

While the HAR dataset demonstrates F*in*P's efficacy on human-centric sensor data, we utilize the CIFAR-10 dataset (distributed via severe Dirichlet sampling, $\alpha = 0.5$) to **stress-test our framework under mathematically controlled, extreme non-IID visual settings.** More on the dataset setup is in Appendix A. We evaluate F*in*P against both standard FedAvg and FedAlign [35], a state-of-the-art FL methodology specifically designed to address data heterogeneity. The core results using the ResNet architecture are summarized in Table 3 in Appendix C and explained below.

1. Outperforming State-of-the-Art Heterogeneity Defenses. A critical finding of our evaluation is that standard methods for handling non-IID data do not inherently resolve privacy disparities. Despite employing sophisticated local distillation techniques, the FedAlign baseline failed to effectively mitigate SIA risks, exhibiting a higher vulnerability variance (CoV(Loss) = 0.86) than even standard FedAvg (0.67).

In contrast, F*in*P directly targets the loss landscape, achieving a Fairness Index for target prediction loss (FI(Loss)) of 0.83. This represents a substantial 20.3% improvement over FedAvg (0.69) and a 43.1% improvement over FedAlign (0.58). By structurally flattening the inter-client loss differences, F*in*P successfully neutralizes the high-confidence statistical signals emitted by outlier models (further illustrated in Figures 7 and 8).

2. Approaching the Theoretical Minimum for Absolute Privacy. Beyond enforcing fairness, F*in*P dramatically reduces the absolute privacy leakage across the federation. For a balanced 10-class classification task like CIFAR-10, the theoretical random-guess probability for a source inference attack is 1/10 (10%).

As shown in Table 3 in Appendix C, while FedAvg and FedAlign suffer from Mean(SIA_{acc}) rates of approximately 30.86%, F*in*P successfully drives the Mean(SIA_{acc}) down to 10.07%—effectively reducing the adversary's attack success to a random guess. Furthermore, the maximum vulnerability observed for any single client (Max(SIA_{acc})) is reduced from 38.52% to 10.67%.

3. Enhancing Group Privacy Fairness (Equal Opportunity). F*in*P maintains comparable baseline CoV(SIA_{acc}) and FI(SIA_{acc}) metrics but demonstrates a massive improvement in Equal Opportunity Difference (EOD). F*in*P reduces the EOD from 0.28 (in both baselines) down to 0.12. As visualized in Figures 9 and 10,

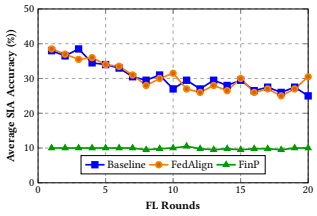


Figure 9: Average SIA accuracy in CIFAR-10 with ResNet with F*in*P with PCA.

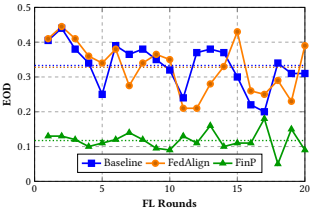
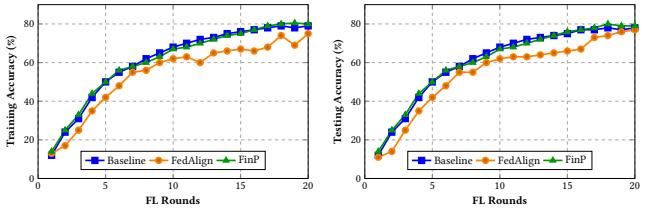


Figure 10: EOD in CIFAR-10 with ResNet with F*in*P with PCA.



(a) Training accuracy. **(b) Testing accuracy.**

Figure 11: Global model classification accuracy using CIFAR-10 with ResNet with F*in*P with PCA.

this 57.14% reduction in EOD mathematically guarantees that the adversary's True Positive Rate is vastly more uniform, preventing the systematic exploitation of distinct visual outliers.

4. The Synergy of Privacy and Utility. Traditionally, strong privacy defenses degrade model utility. However, F*in*P achieves a rare synergy: it actually *improves* global classification accuracy. F*in*P achieves a testing accuracy of 78.46%, marginally higher than FedAvg's 77.62% (requiring only two additional FL communication rounds to converge, as seen in Figure 11). This 0.84% improvement proves our core hypothesis: by utilizing Lipschitz regularization to actively penalize localized memorization, F*in*P forces the constituent local models to learn generalized features, thereby simultaneously improving both privacy and global model utility.

5. Ablation on Client-Side Impact Factor (β). To analyze the sensitivity of the dynamic regularization term, we conducted an ablation study on the impact factor $\beta \in \{0.05, 0.1, 0.3, 0.5\}$ using a standard CNN architecture (Table 4 in Appendix C). The parameter β controls the baseline strength of the Lipschitz penalty applied to the local training loss.

- **Optimal Generalization ($\beta = 0.3$):** Compared to the baseline, increasing β to 0.3 optimally balanced fairness and utility. It improved FI(Loss) from 0.83 to 0.87 and reduced absolute Mean(SIA_{acc}) from 40.91% to 31.85%, while simultaneously improving testing accuracy due to enhanced generalization.
- **Over-regularization Failure ($\beta = 0.5$):** We observed that excessive penalization dominates the client's local training loss, preventing the models from learning valid feature representations. At $\beta = 0.5$, the global model completely failed to converge (Figure 15). While this failure state technically may yield the lowest average SIA accuracy (Appendix C Figure 34), this represents a catastrophic loss of utility rather than a legitimate privacy defense, underscoring the necessity of tuning β to an appropriate task-specific threshold.

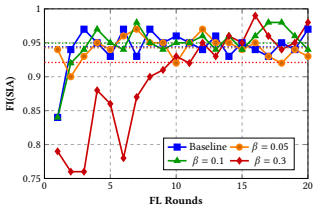


Figure 12: Disparity of SIA accuracy FI(SIA) with PCA among clients using CIFAR-10 dataset with base model CNN and F*in*P with PCA.

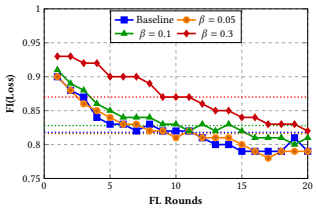


Figure 13: Disparity of prediction loss (FI(Loss)) with PCA among clients using CIFAR-10 dataset with base model CNN and F*in*P with PCA.

6. Scalability: Adaptive Lightweight Aggregation (ALA) vs. PCA. Finally, to validate the deployment practicality of F*in*P, we substituted the computationally prohibitive PCA-based server aggregation with our proposed Adaptive Lightweight Aggregation (ALA) mechanism (Table 4 in Appendix C).

As formalized in Section 4.2, ALA completely bypasses expensive high-dimensional PCA distance calculations. Instead, the server directly utilizes the relative overfitting rank (ρ_k)—which it already computes from the incoming model weights—as a direct inverse-weighting scalar for aggregation.

Our results confirm that ALA maintains seamless model convergence within 20 rounds (Figure 16). Crucially, ALA achieves comparable improvements in FI(Loss) (Figure 18) and SIA reduction (Figure 19) as the heavy PCA baseline. This proves that ALA provides identical structural privacy fairness without introducing any computational bottlenecks at the server, rendering F*in*P highly practical for real-world, large-scale FL ecosystems.

7. Differential Privacy (DP-SGD). Our evaluation on the CIFAR dataset yields results consistent with those observed on HAR: DP affords minimal empirical protection against SIA while incurring a drastic degradation in testing accuracy and further exacerbating privacy unfairness. A more detailed analysis of this dynamic is provided in Discussion Section 6. Visual illustration for the results in Appendix C Table 4 for CIFAR-10 with DP-SGD are listed in Appendix C.2.

8. Validating the Root Cause: Overfitting and Local Epochs. Our core thesis posits that localized overfitting—driven by data heterogeneity—is the fundamental mechanism enabling SIAs. Prior literature suggests that increasing the number of local training epochs deepens this memorization, thereby exacerbating privacy risks [22, 57]. To mathematically validate this within our threat model, we conducted ablation experiments varying the local training epochs ($lp \in \{1, 5, 10\}$) on the CIFAR-10 dataset using a standard CNN. As illustrated in Figure 14, the vulnerability scales linearly with local memorization. The Mean(SIA_{acc}) and Max(SIA_{acc}) surged from 24.31% and 27.20% (at $lp = 1$) to catastrophic levels of 42.9% and 51.4% (at $lp = 10$), respectively. This definitively proves that deeper local convergence inherently sharpens the loss landscape, generating high-confidence signals for the adversary. By attacking this exact symptom via client-side Lipschitz regularization, F*in*P successfully neutralizes the root cause of the privacy disparity.

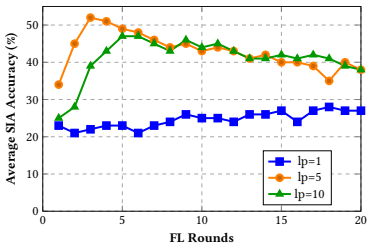


Figure 14: Average SIA accuracy on CIFAR-10 (CNN) across varying local training epochs. Increased local memorization directly correlates with higher vulnerability, validating our focus on overfitting.

5.5 Evaluating F*in*P on FEMNIST Dataset

To further validate our findings, we extend our evaluation of the FinP framework to the FEMNIST dataset. Consistent with the results observed on previous datasets, FinP demonstrates aligned performance in this setting. To simulate a strictly non-IID federated environment, we partition the dataset such that each client is exclusively assigned the data of a single, unique writer. Our results demonstrate that FinP successfully mitigates Source Inference Attacks (SIA), decreasing mean SIA by 8.57% and max SIA by 10.64% in Table 2, while maintaining comparable model accuracy to the baseline.

To comprehensively evaluate fairness improvements, we introduced 2 additional metrics, *Mean Absolute Difference(MAD)* and *Sen's social welfare [45].* Empirical results indicate that FinP significantly curtails vulnerability to Source Inference Attacks (SIA), yielding reductions of 8.57% in mean SIA and 10.64% in max SIA, without compromising global model accuracy relative to the baseline. Beyond aggregate performance, our evaluation demonstrates that FinP actively mitigates privacy disparities across the federated cohort. Specifically, the framework reduces the Mean Absolute Difference (MAD) by up to 0.053 and increases Sen's Welfare by 8% as demonstrated from Table 2.

In contrast, when compared to standard Differential Privacy (DP) methods, we observe that DP significantly degrades global model accuracy while simultaneously exacerbating the MAD. A larger MAD indicates greater variance in client-level loss, which inherently boosts an attacker's linking prediction confidence for highly vulnerable subsets. Consequently, although DP methods may decrease overall SIA accuracy at the cost of utility, they ultimately worsen the disparity in privacy risks, leaving certain clients disproportionately exposed. FinP successfully avoids this tradeoff, maintaining global utility while ensuring a highly equitable distribution of privacy preservation. Detailed definitions of these metrics, further analysis and visual results, along with Hessian computation overhead, are provided in Appendix D.

5.6 Evaluating MIA

We also evaluate client privacy under a White-Box Membership Inference Attack (MIA), assuming an honest-but-curious central server. By injecting a dummy client to act as a memorization baseline, the server dynamically trains a Random Forest classifier at each communication round using the layer-wise cosine similarity between target gradients and mock local weight updates. Our evaluations demonstrate that the FinP framework inherently defends against this attack, reducing MIA accuracy by an average of 5.15%

Table 2: Results of FEMNIST dataset with SIA attack.

	Accuracy (%)		Privacy Metrics (%)		Fairness Metrics					Efficiency
	Train	Test	Mean (SIA_{acc})↓	Max (SIA_{acc})↓	CoV(SIA_{acc}) /FI(SIA_{acc})	CoV(loss) /FI(Loss)	EOD↓	MAD (loss)↓	Welfare (SIA)↑	Conv. round
Baseline [22]	72.71	65.04	39.26	49.70	0.305/0.911	0.281/0.916	0.40	0.636	0.546	10
F*in*P ($\beta = 1$)	60.17	60.11	**30.69**	**39.06**	0.408/0.855	**0.208/0.951**	0.41	0.589	**0.626**	12
F*in*P ($\beta = 0.75$)	67.63	63.76	35.77	46.49	0.319/0.905	0.241/0.937	**0.37**	0.597	0.581	12
F*in*P ($\beta = 0.5$)	**67.81**	**64.23**	37.02	45.58	**0.305/0.912**	0.242/0.936	0.37	**0.583**	0.570	14
DP-SGD (ϵ, δ)										
($151, 10^{-5}$)	51.42	49.97	29.67	34.44	0.272/0.927	0.276/0.926	0.26	1.968	0.661	14
($92, 10^{-5}$)	28.15	29.41	21.35	25.90	0.313/0.910	0.234/0.947	0.22	1.656	0.750	-
($31, 10^{-5}$)	4.66	5.68	12.70	15.66	0.555/0.764	0.398/0.860	0.24	9.062	**0.835**	-

compared to the baseline by fostering a more generalized model training process that mitigates localized overfitting. Full details regarding the data partitioning, shadow training generation, the per-round mia evaluation pipeline and visual results are provided in Appendix E.

5.7 Summary of Empirical Results: F*in*P across HAR and CIFAR-10

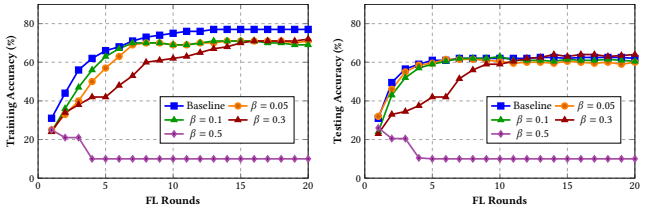
Across our diverse evaluations—spanning human-centric mobile sensor data (HAR) and mathematically extreme non-IID visual distributions (CIFAR-10)—F*in*P consistently demonstrates its capacity to enforce structural fairness in privacy without sacrificing global model utility.

While the HAR dataset presented a naturally saturated baseline for prediction loss fairness, F*in*P successfully formalized and bounded this equity, proving its stability and robustness in highly sensitive, real-world edge environments without causing detrimental utility trade-offs.

The most definitive evidence of our framework's efficacy emerges in the CIFAR-10 evaluation. Under severe, mathematically controlled data heterogeneity, F*in*P achieves two critical milestones for federated privacy:

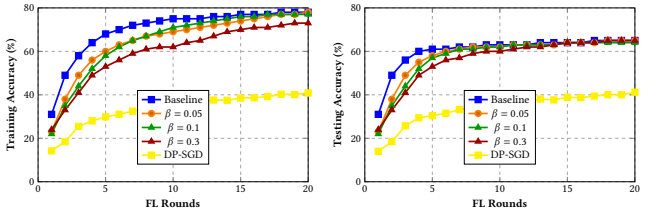
- **Neutralizing Absolute Vulnerability:** By actively penalizing localized memorization through dynamic Lipschitz regularization, F*in*P drives the average SIA success rate down to 10.07%, effectively neutralizing the honest-but-curious adversary's capability to a mathematical random guess.
- **Enforcing Equitable Protection:** Simultaneously, the framework achieves a remarkable 57.14% improvement in the Equal Opportunity Difference (EOD). This structurally eliminates the privacy disparities that traditionally plague non-IID federations, ensuring that statistical outliers (minority data distributions) are not disproportionately targeted.

Ultimately, F*in*P shifts the paradigm of federated defense from reactive, post-hoc mitigation to proactive generalization. By flattening the inter-client loss landscape and denying the server the high-confidence signals required for source attribution, our framework guarantees that high model utility and equitable privacy can co-exist in highly heterogeneous networks.



(a) Training accuracy. (b) Testing accuracy.

Figure 15: Ablation experiment of global model classification accuracy with F*in*P with PCA using CIFAR-10 with CNN. The convergence round (conv. round) denotes the point at which the accuracy stabilizes and ceases to make significant improvements in testing accuracy.



(a) Training accuracy. (b) Testing accuracy.

Figure 16: Adaptive lightweight aggregation (ALA) of global model classification accuracy using CIFAR-10 with CNN.

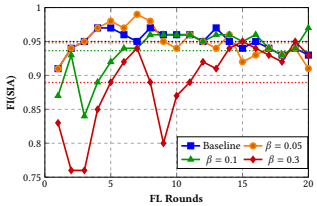


Figure 17: Adaptive lightweight aggregation (ALA) disparity of SIA accuracy (FI(SIA)) among clients using CIFAR-10 dataset with base model CNN.

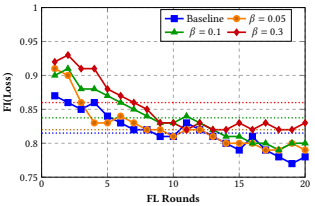


Figure 18: Adaptive lightweight aggregation (ALA) disparity of prediction loss FI(loss) among clients using CIFAR-10 dataset with base model CNN.

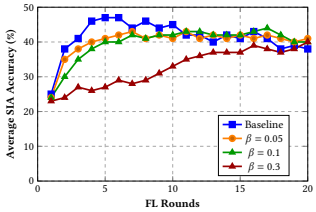


Figure 19: Adaptive lightweight aggregation average SIA accuracy of Baseline and different β using CIFAR-10 dataset with CNN.

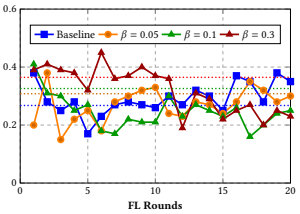


Figure 20: Adaptive lightweight aggregation equal opportunity difference (EOD) using CIFAR-10 dataset with CNN.

6 Discussion, Limitations, & Future Work

Analysis on Differential Privacy (DP) against SIA. F*in*P can be viewed as a complement to Differential Privacy (DP) rather than a replacement: while DP provides formal, worst-case mathematical guarantees against inference attacks, F*in*P specifically targets the fairness of privacy distribution across clients — a dimension that DP does not explicitly address. Each approach is preferable under distinct conditions. DP is the appropriate choice when formal, auditable privacy guarantees are required regardless of utility cost, such as in regulated environments subject to GDPR compliance [1]. F*in*P, by contrast, is preferable when the primary concern is equitable risk distribution in highly non-IID, resource-constrained settings, where DP's geometry-blind uniform noise injection creates severe utility degradation without resolving per-client privacy disparities. As demonstrated empirically (Table 1), DP-SGD degrades global model testing accuracy by 34.7% while yielding only a marginal 2.7% decrease in average SIA accuracy and actually worsening $FI(SIA_{acc})$ by 11.1% compared to the baseline — inadvertently shielding consensus-aligned clients while failing to obscure the highly directional updates of the most disparate clients. Although FinP does not provide formal worst-case privacy guarantees equivalent to DP, it successfully reduces the Mean Absolute Difference (MAD) in loss across clients, a metric that DP tends to exacerbate. In highly non-IID settings, DP applies uniform protection such as fixed clipping and noise across heterogeneous clients, which disproportionately degrades performance for outlier clients and increases the overall MAD in loss. This widened loss disparity can inadvertently boost an attacker's confidence in exploits that rely on loss differences across clients. By reducing this MAD, FinP mitigates this specific empirical vulnerability. Results on FEMNIST show consistent findings in Appendix D Table 2. Furthermore, adaptive clipping variants like DP-Fair [56], designed to improve utility fairness, paradoxically exacerbate the SIA threat by preserving the exact geometric fingerprints the attack exploits. Recent work by Bendoukha et al. [4] further confirms this inherent tension between fairness interventions and privacy leakage in federated learning. In contrast, F*in*P resolves this disparity without relying on post-hoc noise by locally enforcing a Lipschitz constraint to suppress memorization at its root and globally applying a curvature-aware aggregation penalty to down-weight the most vulnerable updates. Looking forward, combining DP and F*in*P represents a highly promising direction: by dynamically adjusting per-client noise budgets based on local loss sharpness, future systems could design an adaptive, geometry-aware DP framework that achieves formal DP bounds and empirical privacy fairness simultaneously while minimizing utility degradation.

Limitations F*in*P assumes honest clients submitting genuine updates and metrics. If a client were actively malicious, it could submit falsified loss-sharpness metrics to manipulate the server's adaptive aggregation weights, or intentionally generate overfitted updates to inflate its assigned vulnerability rank. Defending against such data and metric poisoning would require integrating Byzantine-robust aggregation frameworks, which is beyond the current scope of this work. We acknowledge this as a practical limitation and plan to explore robustness enhancements in future work. F*in*P targets privacy leakage from localized memorization and overfitting, which is the primary driver of SIA disparities in non-IID federated learning. It does not provide formal worst-case privacy guarantees equivalent to Differential Privacy. For attacks independent of memorization — including Property Inference Attacks targeting macro-level statistical properties and gradient reconstruction attacks exploiting raw gradient updates during transmission — F*in*P's efficacy is not guaranteed, as flattening the loss landscape does not inherently prevent these vectors. These limitations, along with the dependence on honest-client and honest-but-curious server assumptions, define the conditions under which F*in*P is effective and should be considered when deploying the framework in settings with stronger adversarial assumptions.

Future Work and Broader Impact Our current definition of privacy disparity is evaluated in standard, singular-decision FL setups. Future work will explore mapping F*in*P to other gradient-inversion attacks and sequential decision-making environments (e.g., reinforcement learning), where privacy risk must be measured over a continuous trajectory of interactions rather than a static dataset [60]. Finally, we aim to explore the integration of our framework with orthogonal Privacy-Enhancing Technologies, such as Secure Multi-Party Computation (SMPC) or lightweight cryptographic masking, to provide defense-in-depth guarantees for decentralized AI and inform future data governance frameworks.

7 Conclusion

The F*in*P framework addresses a critical vulnerability in modern Federated Learning: the inequitable distribution of privacy risks. While traditional FL implementations are optimized to preserve average, network-wide privacy, they systematically ignore the severe disparities inflicted upon statistical outliers. Driven by non-IID data distributions, these distinct users suffer from extreme localized overfitting, rendering their updates highly recognizable and forcing them to bear a disproportionate burden of the network's privacy leakage.

This dynamic raises profound ethical concerns in collaborative learning ecosystems, as it directly translates data heterogeneity into structural discrimination against minority users. F*in*P successfully resolves this by shifting the paradigm from reactive defense to proactive generalization. Through the synergistic combination of server-side adaptive aggregation and client-side collaborative overfitting reduction, our framework flattens the inter-client loss landscape. By structurally denying the adversary the high-confidence signals required for source attribution, F*in*P reduces absolute attack success while guaranteeing a fundamentally equitable distribution of privacy for all participants.

Acknowledgments

This work is supported by the U.S. National Science Foundation (NSF) under grant number 2339266.

References

[1] 2018. General data protection regulation. https://en.wikipedia.org/wiki/General_Data_Protection_Regulation.

[2] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*. 308–318.

[3] Shaojie Bai, J Zico Kolter, and Vladlen Koltun. 2018. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271* (2018).

[4] Adda-Akram Bendoukha, Didem Demirag, Nesrine Kaaniche, Aymen Boudguiga, Renaud Sirdey, and Sébastien Gambs. 2025. Towards privacy-preserving and fairness-aware federated learning framework. In *Privacy Enhancing Technologies (PETs)*, Vol. 2025. 845–865.

[5] Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright. 2011. Robust principal component analysis? *Journal of the ACM (JACM)* 58, 3 (2011), 1–37.

[6] Hongyan Chang, Brandon Edwards, Anindya S Paul, and Reza Shokri. 2024. Efficient privacy auditing in federated learning. In *33rd USENIX Security Symposium (USENIX Security 24)*. 307–323.

[7] Daoyuan Chen, Dawei Gao, Weirui Kuang, Yaliang Li, and Bolin Ding. 2022. pfl-bench: A comprehensive benchmark for personalized federated learning. *Advances in Neural Information Processing Systems* 35 (2022), 9344–9360.

[8] Federico Concone, Cedric Ferdico, Giuseppe Lo Re, and Marco Morana. 2022. A federated learning approach for distributed human activity recognition. In *2022 IEEE International Conference on Smart Computing (SMARTCOMP)*. IEEE, 269–274.

[9] Rachel Cummings, Varun Gupta, Dhamma Kimpara, and Jamie Morgenstern. 2019. On the compatibility of privacy and fairness. In *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization*. 309–315.

[10] Alp Emre Durmus, Zhao Yue, Matas Ramon, Mattina Matthew, Whatmough Paul, and Saligrama Venkatesh. 2021. Federated learning based on dynamic regularization. In *International conference on learning representations*.

[11] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*. 214–226.

[12] Salma Elmalaki. 2021. Fair-iot: Fairness-aware human-in-the-loop reinforcement learning for harnessing human variability in personalized iot. In *Proceedings of the International Conference on Internet-of-Things Design and Implementation*. 119–132.

[13] Salma Elmalaki, Bo-Jhang Ho, Moustafa Alzantot, Yasser Shoukry, and Mani Srivastava. 2022. VindiCo: Privacy Safeguard Against Adaptation Based Spyware in Human-in-the-Loop IoT. *arXiv preprint arXiv:2202.01348* (2022).

[14] Yahya H Ezzeldin, Shen Yan, Chaoyang He, Emilio Ferrara, and A Salman Avestimehr. 2023. Fairfed: Enabling group fairness in federated learning. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 37. 7494–7502.

[15] Grace Fox, Trevor Clohessy, Lisa van der Werff, Pierangelo Rosati, and Theo Lynn. 2021. Exploring the competing influences of privacy concerns and positive beliefs on citizen acceptance of contact tracing mobile applications. *Computers in Human Behavior* 121 (2021), 106806.

[16] Avi Goldfarb and Catherine Tucker. 2012. Shifts in privacy concerns. *American Economic Review* 102, 3 (2012), 349–53.

[17] Yuhao Gu, Yuebin Bai, and Shubin Xu. 2022. CS-MIA: Membership inference attack based on prediction confidence series in federated learning. *Journal of Information Security and Applications* 67 (2022), 103201.

[18] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems* 29 (2016).

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[20] Hongsheng Hu, Zoran Salcic, Lichao Sun, Gillian Dobbie, Philip S Yu, and Xuyun Zhang. 2022. Membership inference attacks on machine learning: A survey. *ACM Computing Surveys (CSUR)* 54, 11s (2022), 1–37.

[21] Hongsheng Hu, Zoran Salcic, Lichao Sun, Gillian Dobbie, and Xuyun Zhang. 2021. Source inference attacks in federated learning. In *2021 IEEE International Conference on Data Mining (ICDM)*. IEEE, 1102–1107.

[22] Hongsheng Hu, Xuyun Zhang, Zoran Salcic, Lichao Sun, Kim-Kwang Raymond Choo, and Gillian Dobbie. 2023. Source inference attacks: Beyond membership inference attacks in federated learning. *IEEE Transactions on Dependable and Secure Computing* (2023).

[23] Matthew Jagielski, Jonathan Ullman, and Alina Oprea. 2020. Auditing differentially private machine learning: How private is private SGD? *Advances in Neural Information Processing Systems* 33 (2020), 22205–22216.

[24] Rajendra K Jain, Dah-Ming W Chiu, William R Hawe, et al. 1984. A quantitative measure of fairness and discrimination. *ACM Transaction on Computer System* (1984).

[25] Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. 2020. Fantastic Generalization Measures and Where to Find Them. In *International Conference on Learning Representations*. https://openreview.net/forum?id=SJgIPJBFvH

[26] Haiming Jin, Lu Su, Bolin Ding, Klara Nahrstedt, and Nikita Borisov. 2016. Enabling privacy-preserving incentives for mobile crowd sensing systems. In *2016 IEEE 36th International Conference on Distributed Computing Systems (ICDCS)*. IEEE, 344–353.

[27] Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Ashesh Rambachan. 2018. Algorithmic fairness. In *Aea papers and proceedings*, Vol. 108. 22–27.

[28] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. 2009. *Learning Multiple Layers of Features from Tiny Images*. Technical Report. University of Toronto.

[29] Bogdan Kulynych, Mohammad Yaghini, Giovanni Cherubin, Michael Veale, and Carmela Troncoso. 2022. Disparate Vulnerability to Membership Inference Attacks. *Proceedings on Privacy Enhancing Technologies* (2022).

[30] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. *Advances in neural information processing systems* 30 (2017).

[31] Jeremiah Liu, Zi Lin, Shreyas Padhy, Dustin Tran, Tania Bedrax Weiss, and Balaji Lakshminarayanan. 2020. Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. *Advances in neural information processing systems* 33 (2020), 7498–7512.

[32] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. 2017. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*. PMLR, 1273–1282.

[33] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)* 54, 6 (2021), 1–35.

[34] Luca Melis, Congzheng Song, Emiliano De Cristofaro, and Vitaly Shmatikov. 2019. Exploiting unintended feature leakage in collaborative learning. In *2019 IEEE symposium on security and privacy (SP)*. IEEE, 691–706.

[35] Matias Mendieta, Taojiannan Yang, Pu Wang, Minwoo Lee, Zhengming Ding, and Chen Chen. 2022. Local learning matters: Rethinking data heterogeneity in federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8397–8406.

[36] Deirdre K Mulligan, Colin Koopman, and Nick Doty. 2016. Privacy is an essentially contested concept: a multi-dimensional analytic for mapping privacy. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 374, 2083 (2016), 20160118.

[37] Dinh C Nguyen, Quoc-Viet Pham, Pubudu N Pathirana, Ming Ding, Aruna Seneviratne, Zihuai Lin, Octavia Dobre, and Won-Joo Hwang. 2022. Federated learning for smart healthcare: A survey. *ACM Computing Surveys (Csur)* 55, 3 (2022), 1–37.

[38] Raphael Poulain, Mirza Farhan Bin Tarek, and Rahmatollah Beheshti. 2023. Improving fairness in ai models on electronic health records: The case for federated learning methods. In *Proceedings of the 2023 ACM conference on fairness, accountability, and transparency*. 1599–1608.

[39] Inioluwa Deborah Raji, Timnit Gebru, Margaret Mitchell, Joy Buolamwini, Joonseok Lee, and Emily Denton. 2020. Saving face: Investigating the ethical concerns of facial recognition auditing. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 145–151.

[40] Ashesh Rambachan, Jon Kleinberg, Jens Ludwig, and Sendhil Mullainathan. 2020. An economic perspective on algorithmic fairness. In *AEA Papers and Proceedings*, Vol. 110. 91–95.

[41] Jorge Reyes-Ortiz, Davide Anguita, Alessandro Ghio, Luca Oneto, and Xavier Parra. 2013. Human Activity Recognition Using Smartphones. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C54S4K.

[42] Nicola Rieke, Jonny Hancox, Wenqi Li, Fausto Milletari, Holger R Roth, Shadi Albarqouni, Spyridon Bakas, Mathieu N Galtier, Bennett A Landman, Klaus Maier-Hein, et al. 2020. The future of digital health with federated learning. *NPJ digital medicine* 3, 1 (2020), 1–7.

[43] Steve Rose, Oliver Borchert, Stu Mitchell, and Sean Connelly. 2020. *Zero Trust Architecture*. Technical Report SP 800-207. National Institute of Standards and Technology (NIST), Gaithersburg, MD. https://doi.org/10.6028/NIST.SP.800-207

[44] Khotso Selialia, Yasra Chandio, and Fatima M Anwar. 2024. Mitigating Group Bias in Federated Learning for Heterogeneous Devices. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. 1043–1054.

[45] Amartya Sen. 1997. *On economic inequality*. Oxford university press.

[46] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*. IEEE, 3–18.

[47] Mahmoud Srewa, Tianyu Zhao, and Salma Elmalaki. 2025. PluralLLM: pluralistic alignment in LLMS via federated learning. In *Proceedings of the 3rd International Workshop on Human-Centered Sensing, Modeling, and Intelligent Systems*. 64–69.

[48] Mahmoud Srewa, Tianyu Zhao, and Salma Elmalaki. 2025. A Systematic Evaluation of Preference Aggregation in Federated RLHF for Pluralistic Alignment of LLMs. *arXiv preprint arXiv:2512.08786* (2025). NeurIPS 2025 Workshop on Evaluating the Evolving LLM Lifecycle.

[49] Mahmoud Srewa, Tianyu Zhao, and Salma Elmalaki. 2026. APPA: Adaptive Preference Pluralistic Alignment for Fair Federated RLHF of LLMs. *arXiv preprint arXiv:2604.04261* (2026).

[50] Spectrum News Staff. 2024. Massive data breach exposes millions of Americans. https://spectrumlocalnews.com/nys/central-ny/news/2024/09/04/data-breach-exposes-americans-information.

[51] Janos Sztipanovits, Xenofon Koutsoukos, Gabor Karsai, Shankar Sastry, Claire Tomlin, Werner Damm, Martin Fränzle, Jochem Rieger, Alexander Pretschner, and Frank Köster. 2019. Science of design for societal-scale cyber-physical systems: challenges and opportunities. *Cyber-Physical Systems* 5, 3 (2019), 145–172.

[52] Mojtaba Taherisadr and Salma Elmalaki. 2024. PEaRL: Personalized Privacy of Human-Centric Systems using Early-Exit Reinforcement Learning. *arXiv preprint arXiv:2403.05864* (2024).

[53] Mojtaba Taherisadr, Stelios Andrew Stavroulakis, and Salma Elmalaki. 2023. adaparl: Adaptive privacy-aware reinforcement learning for sequential decision making human-in-the-loop systems. In *Proceedings of the 8th ACM/IEEE Conference on Internet of Things Design and Implementation*. 262–274.

[54] Alysa Ziying Tan, Han Yu, Lizhen Cui, and Qiang Yang. 2022. Towards personalized federated learning. *IEEE transactions on neural networks and learning systems* 34, 12 (2022), 9587–9603.

[55] Linlin Tu, Xiaomin Ouyang, Jiayu Zhou, Yuze He, and Guoliang Xing. 2021. Feddl: Federated learning via dynamic layer sharing for human activity recognition. In *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems*. 15–28.

[56] Zhenhuan Yang, Yingqiang Ge, Congzhe Su, Dingxian Wang, Xiaoting Zhao, and Yiming Ying. 2023. Fairness-aware differentially private collaborative filtering. In *Companion Proceedings of the ACM Web Conference 2023*. 927–931.

[57] Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. 2018. Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting. In *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*. 268–282. https://doi.org/10.1109/CSF.2018.00027

[58] Chen Zhang, Yu Xie, Hang Bai, Bin Yu, Weihong Li, and Yuan Gao. 2021. A survey on federated learning. *Knowledge-Based Systems* 216 (2021), 106775.

[59] Tianyu Zhao and Salma Elmalaki. 2024. Fina: Fairness of adverse effects in decision-making of human-cyber-physical-system. In *2024 ACM/IEEE 15th International Conference on Cyber-Physical Systems (ICCPS)*. IEEE, 202–211.

[60] Tianyu Zhao, Mojtaba Taherisadr, and Salma Elmalaki. 2024. Fairo: Fairness-aware sequential decision making for human-in-the-loop cps. In *2024 ACM/IEEE 15th International Conference on Cyber-Physical Systems (ICCPS)*. IEEE, 87–98.

[61] Gongxi Zhu, Donghao Li, Hanlin Gu, Yuxing Han, Yuan Yao, Lixin Fan, and Qiang Yang. 2024. Evaluating Membership Inference Attacks and Defenses in Federated Learning. *arXiv preprint arXiv:2402.06289* (2024).

[62] Hangyu Zhu, Jinjin Xu, Shiqing Liu, and Yaochu Jin. 2021. Federated learning on non-IID data: A survey. *Neurocomputing* 465 (2021), 371–390.

A Datasets and Setup

We evaluate F*in*P using the UCI HAR[41] and CIFAR-10 datasets [28] utilizing federated learning with non-IID data partitions and standard model architectures (TCN for HAR, CNN and ResNet56 for CIFAR-10).

Setup for Human Activity Recognition. We utilized the UCI HAR Dataset [41], a widely used dataset in activity recognition research, especially in FL [8, 55]. The dataset includes sensor data from 30 subjects (aged 19–48) performing six activities: walking, walking upstairs, walking downstairs, sitting, standing, and laying. The data was collected using a Samsung Galaxy S II smartphone worn on the waist, capturing readings from both the accelerometer and gyroscope sensors. Each subject in the dataset was treated as an individual client in the FL setup, preserving the data's unique activity patterns and non-IID nature. We allocated 70% of each client's data for training using 5-fold cross-validation and 30% for testing, enabling evaluation of the model on independently collected test data. Data preprocessing involved applying noise filters to the raw signals and segmenting the data using a sliding window approach with a window length of 2.56 seconds and a 50% overlap, resulting in 128 readings per window. We selected the HAR dataset for evaluation F*in*P due to its inherited non-IID structure.

We trained the model in a federated learning setting using the Federated Averaging Algorithm (FedAvg) aggregation method over 20 global communication rounds. Each client trained locally with a batch size of 64, a learning rate of 0.001 using Adam optimizer, 1 local epochs per round, and an impact factor β of 2. These parameters ensured balanced model updates from each client while maintaining computational efficiency across the federated network. Each local model (one per subject) analyzes its time-series sensor data using Temporal Convolutional Network (TCN) model[3]. The TCN model, designed for time-series data, uses causal convolutions to capture temporal dependencies while preserving sequence order. The architecture includes two convolutional layers, each followed by max-pooling, with a final fully connected layer for classifying the six activity classes.

Setup for CIFAR-10. The CIFAR-10 dataset consists of 60000 32x32 color images in 10 classes, with 6000 images per class. There are 50000 training images and 10000 test images. We use the Dirichlet distribution $Dir(\alpha)$ to divide the CIFAR-10 dataset into K unbalanced subsets similar to previous work in the literature [21, 35], with $\alpha = 0.5$. Figure 2 demonstrates how the data are distributed among clients with $\alpha = 0.1$. We created 10 clients and employed ResNet56 [19] as the local model. Similar to the setup in HAR, we trained the model over 20 global communication rounds. Each client is trained locally with an impact factor β of 0.1 as described in Equation 5. As for ablation experiment in β, we evaluated multiple values of $\beta = \{0.05, 0.1, 0.3, 0.5\}$ described in Equation 5. We used the same CNN proposed in [22] as the local model.

Setup for FEMNIST. We evaluate on FEMNIST from Hugging Face (flwrlabs/femnist), a handwritten character recognition benchmark with 62 classes and natural non-IID structure induced by writer identity. We simulate K federated clients by sampling K distinct writers without replacement; each client owns only that writer's examples. Images are converted to single-channel tensors

and normalized with MNIST statistics (μ=0.1307, σ=0.3081). For each client, we perform an 80/20 train–test split at the writer level, yielding client-local heterogeneity in both label and feature distributions.

We employ a lightweight CNN, FEMNISTNet, tailored to 28×28 grayscale inputs. The architecture comprises two convolutional blocks (32 and 64 channels, 3×3 kernels, GroupNorm, GELU), each followed by learnable strided downsampling (28→14→7), and two fully connected layers (3136→256→62). and local training uses SGD with learning rate 0.02. The model has on the order of ∼3M trainable parameters and outputs unnormalized class logits for 62-way classification with cross-entropy loss. This design balances capacity and efficiency for per-client federated optimization and curvature-based collaboration metrics.

B Experiment Results on HAR

B.1 HAR Results with PCA as server aggregation

Figures 21 and 22 show the average SIA accuracy and the equal opportunity difference (EOD) per client using HAR dataset with PCA-based server aggregation, respectively.

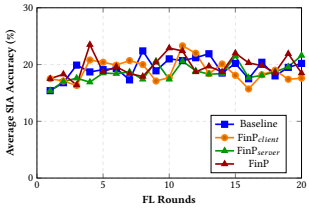


Figure 21: Average SIA accuracy of Baseline, F*in*P$_{\text{client}}$, F*in*P$_{\text{server}}$, and F*in*P with PCA on HAR dataset.

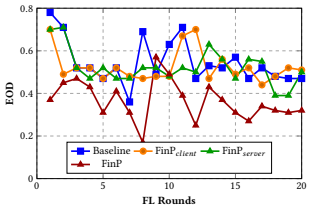
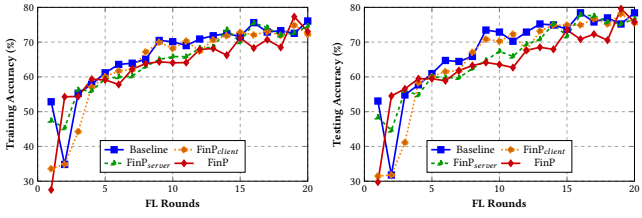


Figure 22: Equal opportunity difference (EOD) on HAR dataset and F*in*P with PCA.

B.2 HAR with Adaptive Lightweight Aggregation (ALA)

Figure 23 shows the training and testing accuracy for HAR dataset using the lightweight server aggregation.



(a) Training accuracy. **(b) Testing accuracy.**

Figure 23: Global model classification accuracy using HAR with ALA.

Figures 24 and 25 describe the FI(SIA) and FI(loss) for HAR dataset using ALA aggregation. Figure 26 shows the average SIA accuracy in HAR with ALA while Figure 27 highlights the Equal Opportunity Difference (EOD) in HAR with ALA.

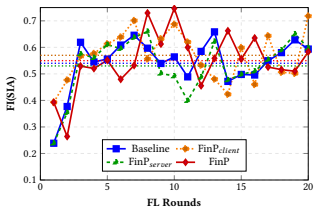


Figure 24: Disparity of SIA accuracy (FI(SIA)) among clients using HAR dataset with ALA.

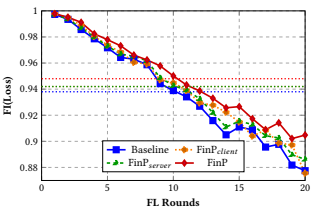


Figure 25: Disparity of prediction loss FI(Loss) among clients using HAR with ALA.

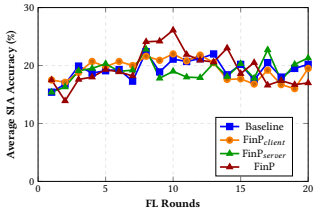


Figure 26: Average SIA accuracy using HAR dataset with ALA.

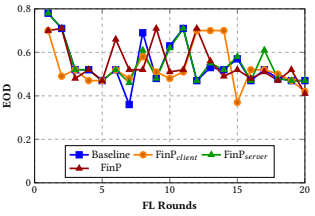
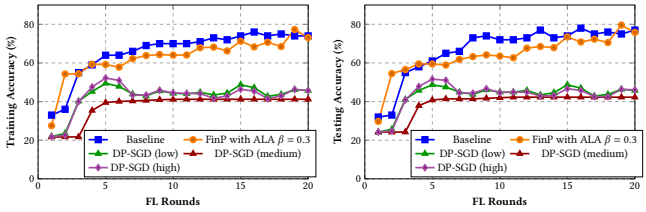


Figure 27: Equal opportunity difference (EOD) using HAR dataset with ALA.

B.3 Differential Privacy (DP) with HAR

Figure 28 illustrates the training and testing accuracy for HAR when using DP-SGD, while Figures 29 and 30 show the FI(SIA) and FI(loss).



(a) Training accuracy. **(b) Testing accuracy.**

Figure 28: Global model classification accuracy using HAR with different levels of noise in DP-SGD.

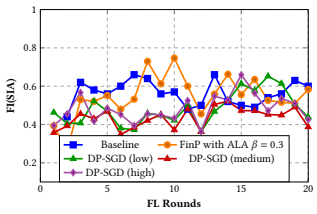


Figure 29: Disparity of SIA accuracy (FI(SIA)) among clients using HAR dataset with different noise in DP-SGD.

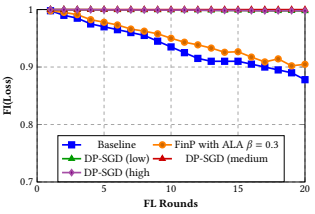


Figure 30: Disparity of prediction loss FI(Loss) among clients using HAR and different levels of noise in DP-SGD.

B.4 Correlation between the top Hessian eigenvalue ($\lambda_{\max}$) and Hessian trace (H_T)

Figure 33 shows the strong correlation between the top Hessian eigen value ($\lambda_{\max}$) and the Hessian trace (H_T).

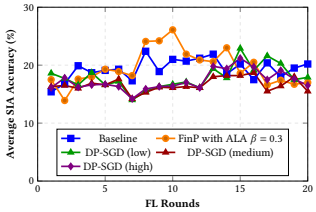


Figure 31: Average SIA accuracy using HAR dataset with ALA and different noise in DP-SGD.

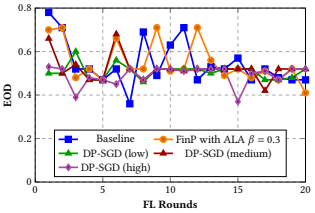


Figure 32: EOD using HAR dataset with ALA and different noise in DP-SGD.

C Experiment Results on CIFAR-10

Complete results for ablation experiments on β, ALA and PCA comparison, and DP-SGD can be found in Table 4.

C.1 Equal Opportunity Difference (EOD) and SIA accuracy

Figures 34 and 35 illustrate the average SIA accuracy in CIFAR-10 with CNN across different β, and the EOD results respectively.

C.2 Differential Privacy (DP) with CIFAR-10

Figure 36 highlights the global model classification accuracy in CIFAR-10 with CNN under different levels of noise in DP-SGD. Figures 37 and 38 show the disparity of SIA accuracy (FI(SIA)) and disparity of prediction loss (FI(loss)) among clients using CIFAR-10 with CNN and under different noise in DP-SGD, respectively. Figures 39 and 40 illustrate the average SIA accuracy and EOD respectively using ALA in CIFAR-10 with CNN and different noise in DP-SGD

D Experiment Results on FEMNIST

D.1 Additional Fairness and Disparity Metrics

We utilize fairness notions beyond simple average performance and incorporate metrics that explicitly quantify the distributional disparity among clients. Specifically, we utilize the Mean Absolute Difference (MAD) and Sen's Social Welfare Function.

Mean Absolute Difference (MAD) The Mean Absolute Difference quantifies the average absolute disparity between all possible pairs of individuals or clients within the system. For a population of size n, where x_i represents the metric evaluated for client i, the MAD is defined as:

$$\text{MAD} = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} |x_i - x_j| \tag{12}$$

Unlike variance, which squares differences and thus heavily weights extreme outliers, MAD provides a linear, highly interpretable measure of overall systemic inequality. Crucially, assuming $x_i \in [0, 1]$, MAD reaches its mathematical upper bound of 0.5 not under a monopoly (where one client holds all utility), but under conditions of *perfect polarization*—where exactly half of the population receives a score of 1 and the other half receives 0. Therefore, MAD is exceptionally well-suited for identifying models that systematically disenfranchise specific clusters or demographics while perfectly serving others and independent of the mean values.

Sen's Social Welfare Function To holistically assess the system, it is necessary to balance overall efficiency with equity. Sen's Social Welfare Function [45] achieves this by scaling the global average performance (μ) by an inequality penalty. By fully expanding the Gini coefficient (G) within the standard formulation $W = \mu(1 - G)$, the function is explicitly defined as:

$$W = \mu\,(1 - G) = \mu \left(1 - \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} |x_i - x_j|}{2n^2 \mu}\right) \tag{13}$$

This formula calculates the egalitarian equivalent performance of the system, assuming the target metric (e.g., utility or predictive accuracy) is one where a higher value is optimal. It bridges the gap between raw average performance and distributional fairness.

- **The Global Mean (μ):** The outer multiplier represents the system's overall average performance. The function inherently rewards pushing this global average as high as possible. *In our context of attack accuracy, where a lower value is optimal, we calculate the μ based on (1-Accuracy$_{attack}$).*
- **The Expanded Disparity Penalty:** The fractional term inside the parentheses is the expanded Gini coefficient. The double summation in the numerator ($\sum \sum |x_i - x_j|$) exhaustively calculates the absolute disparity between every possible pair of clients in the system. The denominator ($2n^2\mu$) normalizes this disparity relative to the total number of pairwise comparisons and the size of the mean itself.

By subtracting this normalized disparity from 1 and multiplying the result by the mean, the function imposes a strict structural penalty on inequality. The system cannot achieve a high Welfare score simply by maximizing performance for a privileged subset of clients (which would inflate μ but also drastically inflate the disparity penalty). Instead, it is mathematically forced to distribute performance as equitably as possible across all clients.

D.2 Hessian Calculation Overhead

To demonstrate the practical viability of deploying FinP in real-world federated environments, we evaluate the computational overhead introduced by the Hessian calculations during our FEMNIST experiments. Because our framework relies primarily on the dominant (top) Hessian eigenvalue instead of full Hessian Matrix, and it does not require exact numerical precision to effectively rank client vulnerability, we can aggressively optimize the solver to significantly reduce the computational burden.

In our experiments, the global neural network architecture has a lightweight memory size of 3.275 MB CNN. To optimize for speed, we configure the eigensolver with highly relaxed convergence constraints: a maximum of 20 iterations for both the eigenvalue solver and the trace estimator (`hessian_eig_max_iter=20`, `hessian_trace_max_iter=20`), alongside a tolerance threshold of 5×10^{-3} (`hessian_tol=5e-3`). Under these lightweight hyperparameters, the dominant Hessian eigenvalue computation requires only 1.82 seconds per client. This minimal overhead confirms that FinP can be seamlessly integrated into standard federated training pipelines without imposing a prohibitive computational bottleneck on edge devices.

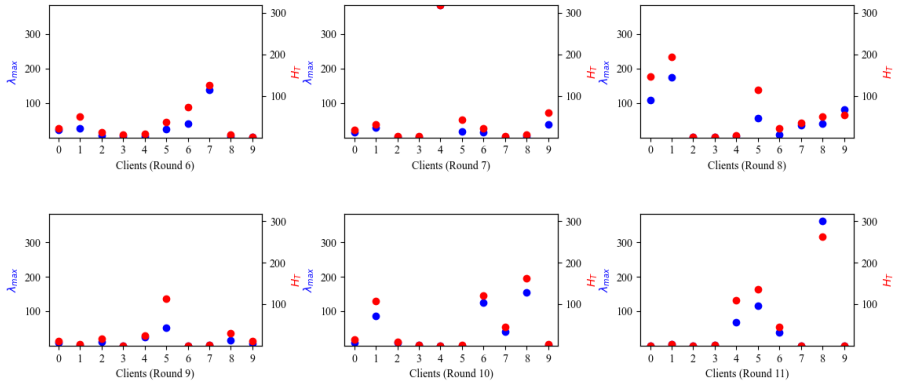


Figure 33: Strong correlation between the top Hessian eigenvalue ($\lambda_{\max}$, blue dots) and Hessian trace (H_T, red dots) across training rounds in HAR.

Table 3: Results of CIFAR dataset using ResNet as the local model.

	Accuracy (%)		Privacy Metrics (%))		Fairness Metrics			Efficiency
	Train	Test	Mean (SIA$_{acc}$) ↓	Max (SIA$_{acc}$)↓	CoV(SIA$_{acc}$)/FI(SIA$_{acc}$)	CoV(loss)/FI(Loss)	EOD ↓	Conv. round
Baseline	78.39	76.45	30.86	38.52	**0.33/0.90**	0.67/0.69	0.33	10
FedAlign [35]	70.79	71.87	30.72	38.46	0.34/0.89	0.86/0.58	0.33	14
F*in*P(with PCA)	**80.23**	**78.47**	**10.07**	**10.67**	0.35/0.89	**0.44/0.83**	**0.12**	12

Table 4: Ablation experiments on β. Results of CIFAR dataset using CNN as the local model [22].

	Accuracy (%)		Privacy Metrics (%)		Fairness Metrics			Efficiency
	Train	Test	Mean (SIA$_{acc}$) ↓	Max (SIA$_{acc}$)↓	CoV(SIA$_{acc}$)/FI(SIA$_{acc}$)	CoV(loss)/FI(Loss)	EOD ↓	Conv.round
Baseline [22]	75.62	62.37	40.91	46.70	0.23/0.95	0.46/0.83	0.29	5
F*in*P								
(β=0.05, PCA)	69.31	59.67	39.51	43.40	0.21/0.96	0.46/0.83	0.27	5
(β=0.1, PCA)	70.45	61.19	39.47	43.70	**0.19/0.96**	0.44/0.84	**0.25**	7
(β=0.3, PCA)	71.17	63.81	**31.85**	**39.90**	0.31/0.90	**0.38/0.87**	0.31	12
(β=0.5, PCA)	10	10	N/A	N/A	N/A	N/A	N/A	N/A
(β=0.05, ALA)	**76.03**	**64.26**	38.62	43.90	0.23/0.95	0.46/0.83	0.29	6
(β=0.1, ALA)	74.64	63.94	37.39	42.50	0.25/0.94	0.44/0.84	0.32	7
(β=0.3, ALA)	71.00	64.09	34.09	41.00	0.34/0.89	0.41/0.86	0.38	11
DP-SGD (ϵ, δ)								
(93, 10^{-5})	43.73	43.59	23.38	24.50	0.52/ 0.79	0.290/0.922	0.37	15
(16, 10^{-5})	42.84	42.62	22.68	24.20	0.45/0.83	0.298/0.918	0.31	15
(5, 10^{-5})	31.52	32.16	21.85	24.60	0.38/0.88	0.271/0.931	0.29	13

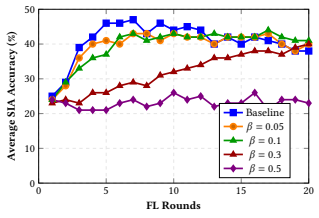


Figure 34: Average SIA accuracy of Baseline and different β using CIFAR-10 with CNN.

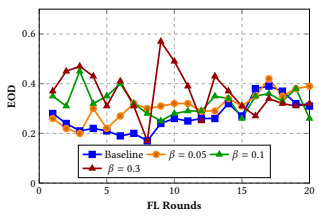
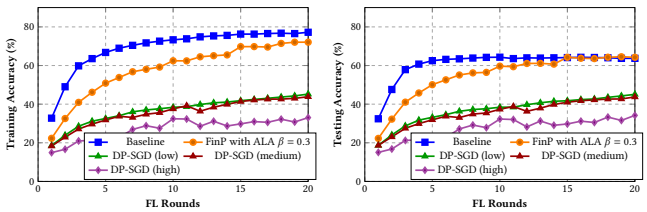


Figure 35: Equal opportunity difference (EOD) using CIFAR-10 with CNN.



(a) Training accuracy. **(b) Testing accuracy.**

Figure 36: Global model classification accuracy using CIFAR-10 with CNN and different levels of noise in DP-SGD.

E Detailed Experimental Setup and Results for Per-Round White-Box MIA

To evaluate the vulnerability of individual client updates, we simulate a per-round gradient-matching Membership Inference Attack

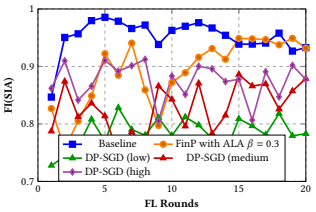


Figure 37: Disparity of SIA accuracy (FI(SIA)) among clients using CIFAR-10 dataset with CNN and different noise in DP-SGD.

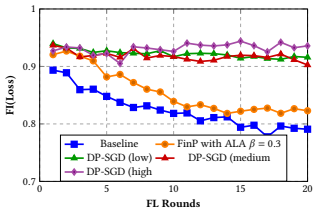


Figure 38: Disparity of prediction loss FI(Loss) among clients using CIFAR-10 with CNN and different levels of noise in DP-SGD.

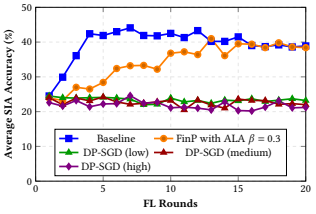


Figure 39: Average SIA accuracy in CIFAR-10 dataset using ALA with CNN and different noise in DP-SGD.

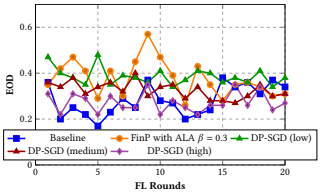
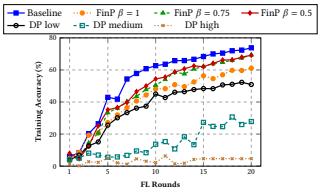
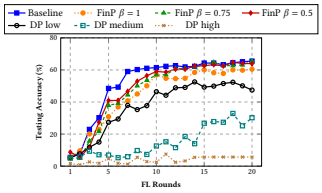


Figure 40: EOD in CIFAR-10 dataset using ALA with CNN and different noise in DP-SGD.



(a) Global model training accuracy.



(b) Global model testing accuracy.

Figure 41: Global model classification accuracy using FEMNIST dataset. F*in*P (ALA) maintains competitive accuracy with minimal convergence delay.

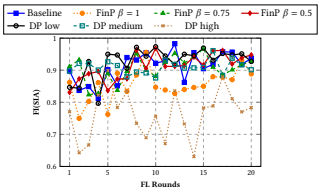


Figure 42: Disparity of SIA accuracy (FI(SIA)) among clients using FEMNIST dataset with F*in*P with ALA.

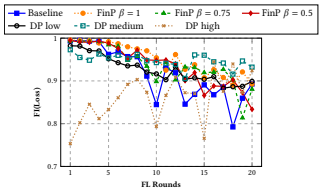


Figure 43: Disparity of prediction loss FI(loss) among clients using FEMNIST dataset with F*in*P with ALA.

(MIA). The adversary is defined as the "honest-but-curious" central server, which has access to the global model parameters G_{t-1} and intercepts the raw, unaggregated local weight updates Δw_i from each client i at round t.

E.1 Data Partitioning and Threat Model Setup

We partition the overall dataset to ensure strict separation between the adversary's auxiliary knowledge and the final evaluation data. The federated cohort consists of N participating clients.

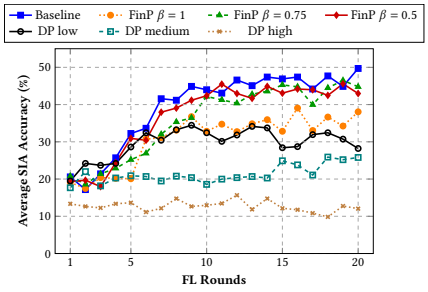


Figure 44: Average SIA accuracy in FEMNIST with F*in*P with ALA.

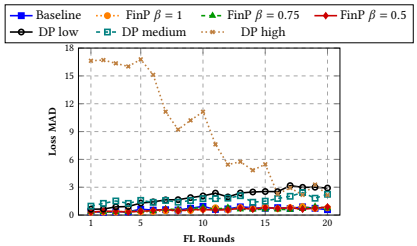


Figure 45: Loss MAD (mean absolute difference of loss) across clients in FEMNIST with F*in*P with ALA.

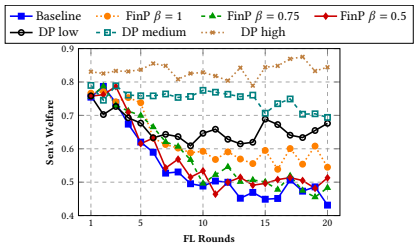


Figure 46: Sen's welfare across clients in FEMNIST with F*in*P with ALA.

- **Dummy Client (Client 0):** The adversary injects one sybil client into the federated process. This client participates in aggregation at every round, serving as the "Member" baseline.
- **Victim Clients (Clients 1 to $N-1$):** The genuine participants who is targeted in MIA.
- **Shadow Non-Member Pool ($\mathcal{D}_{shadow}$):** A pool of data (e.g., an isolated writer) held by the adversary, exclusively used to generate "Non-Member" features for classifier training.
- **Evaluation Sets ($\mathcal{D}_{test}$):** For each victim client i, we construct a static evaluation dataset of 100 records: 50 randomly sampled records from client i's local data (Label 1) and 50 records from unseen writer data (Label 0).

E.2 Dynamic Feature Generation (Round t)

Because full-participation FL continuously memorizes data over time, static MIA thresholds fail. Instead, the adversary calibrates a dynamically trained Random Forest classifier prior to the aggregation step at each round t. We generate a balanced training set of 60 feature vectors (30 positive, 30 negative) using a bootstrapping approach on the Dummy Client.

For each of the 30 iterations, we generate a paired sample:

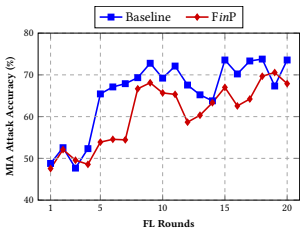


Figure 47: MIA attack accuracy on FEMNIST with F*in*P with ALA.

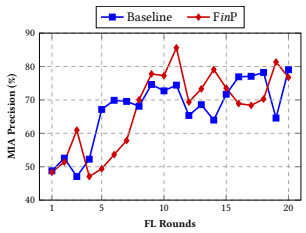


Figure 48: MIA precision on FEMNIST with F*in*P with ALA.

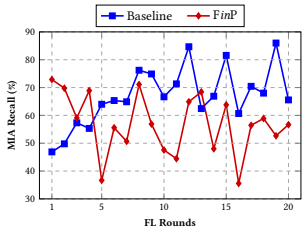


Figure 49: MIA recall on FEMNIST with F*in*P with ALA.

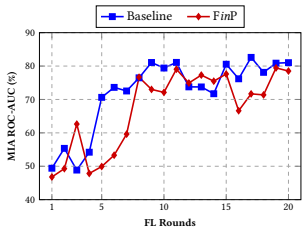


Figure 50: MIA ROC-AUC on FEMNIST with F*in*P with ALA.

(1) **Target Sampling:** The adversary samples a member record $x_m \sim$ Client 0 and a non-member record $x_{nm} \sim \mathcal{D}_{shadow}$.
(2) **Gradient Extraction:** Using standard Cross-Entropy Loss, the adversary computes the individual gradient footprints g_m and g_{nm} with respect to the current global model G_{t-1}.
(3) **Mock Local Update Simulation:** The adversary samples a random 80% subset of Client 0's local dataset, ensuring the inclusion of x_m. A temporary clone of G_{t-1} is trained on this subset using the standard local hyperparameters to yield a mock update Δw_{mock}.
(4) **Feature Representation:** To map the high-dimensional parameter space to a robust feature space, the adversary calculates the layer-wise cosine similarity. The positive feature vector is $\cos(g_m, \Delta w_{mock})$, and the negative feature vector is $\cos(g_{nm}, \Delta w_{mock})$.

E.3 Classifier Training and Inference

The 60 layer-wise similarity vectors are used to train an Random Forest Classifier. This classifier serves as the decision boundary for round t.

To execute the attack, the adversary intercepts the real uploaded update Δw_i from each victim client $i \in \{1, \ldots, N-1\}$. For each record x_{eval} in the victim's 100-record Evaluation Set $\mathcal{D}_{test}$, the adversary computes the gradient g_{eval} on G_{t-1} and extracts the layer-wise cosine similarity against the intercepted Δw_i. The similarity vector is passed to the Random Forest to predict membership. We compute standard classification metrics (Accuracy, Precision, Recall, and ROC-AUC) against the ground-truth labels.

E.4 MIA Results

our evaluations demonstrate that the FinP framework inherently provides an effective defense against Membership Inference Attacks (MIAs). Specifically, we observe that MIA accuracy drops by an average of 6.1% under the FinP setup compared to the baseline in Figure 47 and Table 5. Although the MIA accuracy disparity across clients are minor (FI = 0.99), F*in*P consistently decreases the disparity of loss (MAD) from 0.638 to 0.574, and decreases the EOD by 0.22 among clients. This empirical reduction highlights that FinP actively mitigates privacy leakage by fostering a more generalized model training process, thereby reducing the localized overfitting and instance-level memorization that gradient-matching MIAs rely upon to succeed.

Table 5: Results of FEMNIST dataset with MIA attack.

	Accuracy (%)		Privacy Metrics (%)		Fairness Metrics					Efficiency
	Train	Test	Mean (MIA_{acc})↓	Max (MIA_{acc})↓	CoV(MIA_{acc}) /FI(MIA_{acc})	CoV(loss) /FI(Loss)	EOD (MIA)↓	MAD (loss)↓	Welfare (MIA)↑	Conv. round
Baseline [22]	72.72	65.04	66.63	73.11	0.088/0.99	0.281/0.916	0.40	0.638	0.302	10
F*in*P ($\beta = 1$)	58.65	60.38	**60.53**	**70.56**	0.099/0.990	**0.215/0.949**	0.21	0.606	0.304	12
F*in*P ($\beta = 0.75$)	65.23	62.34	66.57	75.22	**0.089/0.991**	0.229/0.942	**0.18**	**0.574**	0.302	12
F*in*P ($\beta = 0.5$)	**68.03**	**63.76**	86.31	74.78	0.093/0.990	0.257/0.928	0.20	0.608	0.283	14
DP-SGD (ϵ, δ)										
($151, 10^{-5}$)	49.57	50.10	50.58	58.22	0.038/0.996	0.288/0.920	0.06	2.062	0.483	14
($92, 10^{-5}$)	27.01	29.01	50.11	52.44	0.038/0.998	0.257/0.938	0.06	2.005	0.488	-
($31, 10^{-5}$)	2.95	2.50	13.62	16.56	0.041/0.997	0.436/0.838	0.07	10.097	0.493	-