

Not All Retained Clients Are Equal: Client Selection for Federated Unlearning

Thanh Linh Nguyen, Quoc-Viet Pham

School of Computer Science and Statistics, Trinity College Dublin, Ireland
{tnguyen3,viet.pham}@tcd.ie

Abstract

Federated unlearning (FU) aims to remove a designated target client’s influence from a trained model without the prohibitive cost of retraining from scratch. Existing methods largely treat unlearning as a post-hoc, stateless operation, overlooking that client influence is path-dependent: contributions made during training, particularly during critical learning periods (CLPs), can have disproportionate and persistent effects. Moreover, knowledge overlap and intermittent client availability make relying on all retained clients inefficient and impractical. We introduce PDF, a path-dependent client-selection framework based on the insight that retained clients are not equally valuable for unlearning. During training, PDF maintains lightweight, target-agnostic, CLP-weighted histories summarizing each client’s accumulated update direction and global-model alignment. Upon an unlearning request, these histories identify and rank retained clients that are globally informative yet weakly coupled to the target, enabling retraining-based unlearning with only the top- k clients over a small number of rounds. Experiments on FMNIST, CIFAR-10, and CIFAR-100 under IID and non-IID settings show that PDF achieves effective forgetting and strong retained utility with substantially fewer retained clients. In particular, PDF achieves 0% forgetting accuracy with retraining-like MIA AUC on non-IID CIFAR-10, improves retained accuracy by up to 6.91 percentage points over the strongest competing FU baseline on CIFAR-100, and reduces communication by approximately 69% and computation by 57–61% relative to Retrain.

1 Introduction

As federated learning (FL) gains traction in privacy-sensitive and highly regulated domains (e.g., healthcare [11, 29]), deployed systems must accommodate not only privacy-preserving model training but also the post-training removal of client data influence—whether to comply with legal obligations, such as the right to be forgotten under the GDPR [26], or to mitigate impacts of contributions from corrupted, low-quality, or malicious clients. These operational and regulatory demands have given rise to *federated unlearning* (FU), which aims to eliminate the influence of designated target clients from a trained global model while preserving the knowledge contributed by the remaining clients.

Existing FU methods, including retraining-based approaches [19, 35, 39] and model-manipulation approaches [9, 16, 22, 32, 33, 40], face three major limitations in addressing client-level unlearning. First, they overlook how client influence develops throughout the training stage. In FL, client participation and update directions vary across rounds; therefore, a client’s influence on the final model depends not only on *whether* it participates, but also on *when* it participates. This temporal structure is particularly important during critical learning periods (CLPs), in which client updates can have disproportionately persistent effects on the learned representations, optimization trajectory, and final model performance [1, 17, 36, 37]. Moreover, under heterogeneous data, changing participation, or adaptive optimization, critical phases may emerge beyond a fixed early-training window [13, 30]. Second, retraining-based methods incur substantial computation, communication,

and storage overhead because they reconstruct or recalibrate the model using all retained clients. Their effectiveness therefore depends on continuous availability of clients after training, an assumption that may not hold in practical FL systems with intermittent or voluntary participation. Third, efficient model-manipulation methods often face a pronounced trade-off between target forgetting efficacy and retained-model utility [6, 9, 33]. Achieving stronger forgetting may excessively perturb knowledge shared with retained clients, whereas preserving utility may leave residual target influence. They may also rely on expensive server-side operations or require future unlearning targets to be anticipated during the original training stage [6, 9, 22, 32, 40].

We argue that these limitations share a common cause: existing FU methods are largely model layer-, parameter-, or protocol-centric and treat unlearning as a stateless operation, overlooking the impacts of heterogeneous influence that clients accumulate over training. We instead formulate client-level unlearning as an inherently *path-dependent* client-selection problem. Inspired by path dependence in economics [3], we posit that a client’s influence is shaped not only by its final contribution, but also by when that influence emerged, particularly during CLPs, and by how strongly it became coupled with other clients. Thus, effective unlearning should select retained clients according to their historical alignment with the global model and their dependence on the target clients, rather than uniformly or solely by availability. Building on this insight, we ask the following question: ***How can the path-dependent influence during the learning stage be exploited to improve the forgetting efficacy, retained utility, and efficiency of client-level FU?***

Therefore, we propose a **Path-Dependent** client selection framework for **FU**, named **PDF**. During training, PDF maintains lightweight, target-agnostic histories that capture each client’s alignment with the global optimization trajectory, its historical update direction with other clients, and its participation during CLPs. Upon an unlearning request, these histories are used to estimate each retained client’s coherence with the target set, and the resulting signals are combined into a unified ranking. The ranking reflects our central hypothesis: *retained clients are not equally useful for unlearning*. **Clients whose updates consistently supported the global model, particularly during critical rounds, are valuable for preserving retained knowledge, whereas clients strongly coupled with the targets may carry overlapping information that hinders selective forgetting.** PDF therefore performs retraining-based unlearning using only the top- k clients that are globally informative yet weakly coupled with the target set, reducing participation, communication, and computation while balancing target forgetting and retained-data generalization. Our contributions are summarized as follows:

- We introduce a *path-dependent perspective* on client-level FU, showing that retained clients can differ in their value for unlearning because their influence depends on when and how they contributed during training and on their historical dependence on the target clients.
- We propose PDF, a CLP-aware client-selection framework that maintains lightweight, target-agnostic training histories and request-conditioned target coherence to rank retained clients, enabling retraining-based unlearning with only the top- k clients.
- Experiments on FMNIST, CIFAR-10, and CIFAR-100 across IID and non-IID settings show that PDF achieves effective forgetting and strong retained utility while substantially reducing client participation, communication, and computation. Ablations further validate the proposed ranking and CLP-aware design.

2 Related work

Machine and Federated Unlearning. Machine unlearning aims to remove the influence of designated training data without full retraining [5]. Existing methods modify either the training process or the learned model through data partitioning, parameter transformation, pruning, or cached training states [4, 8, 15]. Such methods generally assume access to raw data, intermediate training states, or sufficient retraining budget—assumptions that fail in FL, where the server cannot access local datasets and clients may be intermittently available [21]. FU extends this objective to client-, sample-, or class-level removal [14, 27]. Client-level FU removes the entire contribution of designated clients, encompassing potentially many samples and classes whose influence has accumulated across multiple training rounds. Its methods either approximate retraining from historical updates cached during training [19, 35, 39], or suppress target influence through parameter, layer, gradient, or representation manipulation [6, 9, 16, 22, 25, 33].

Critical Learning Periods. CLPs are training intervals during which updates exert disproportionately persistent effects on learned representations and final performance [1, 17]. In FL, analogous critical rounds emerge, reflecting the temporally non-uniform nature of client influence [36, 37]. Existing CLP-aware methods leverage these dynamics to enhance client participation, improve aggregation strategies, or boost training efficiency [13, 30, 37].

Discussion. Existing FU methods utilize training artifacts for post-hoc correction, yet they treat all training rounds and retained clients as equally relevant, overlooking the temporally non-uniform influence. Concurrently, CLP-aware FL methods explicitly model these temporal dynamics, but their focus remains strictly on improving model training, leaving their implications for subsequent influence removal unaddressed. This raises an underexplored question: can the temporal patterns captured in the training stage serve as a useful signal for identifying which retained clients are most valuable for unlearning? PDF investigates this question by incorporating CLP-weighted signals into client-level FU, exploring whether such temporal histories can effectively guide client selection.

3 Methodology

3.1 System Model

Federated setting. We consider an FL system comprising a central server and N clients indexed by $\mathcal{I} = \{1, \dots, N\}$. Each $i \in \mathcal{I}$ holds a private dataset $\mathcal{D}_i = \{(x_j^{(i)}, y_j^{(i)})\}_{j=1}^{n_i}$ of $n_i = |\mathcal{D}_i|$ samples, where $(x_j^{(i)}, y_j^{(i)}) \in \mathcal{X} \times \mathcal{Y}$. Client datasets are record-disjoint and contain $n = \sum_{i \in \mathcal{I}} n_i$ samples in total. All clients share a common model architecture parameterized by $\mathbf{w} \in \mathbb{R}^d$, where d is the model dimension. Let $\ell(\mathbf{w}; x, y) \in \mathbb{R}_{\geq 0}$ be the sample-wise loss. The local empirical objective of client i is

$$F_i(\mathbf{w}) = \frac{1}{n_i} \sum_{j=1}^{n_i} \ell(\mathbf{w}; x_j^{(i)}, y_j^{(i)}). \quad (1)$$

With $\pi_i = n_i/n$, $\sum_{i \in \mathcal{I}} \pi_i = 1$, the global empirical objective is $F(\mathbf{w}) = \sum_{i \in \mathcal{I}} \pi_i F_i(\mathbf{w})$.

Federated training. Training proceeds for T_{learn} communication rounds using FedAvg [21]. At round t , the server samples a participating subset $\mathcal{S}^{(t)} \subseteq \mathcal{I}$ with $|\mathcal{S}^{(t)}| = M \leq N$ uniformly at random without replacement, and broadcasts $\mathbf{w}^{(t-1)}$ to $\mathcal{S}^{(t)}$. Participation is recorded by $a_i^{(t)} = \mathbf{1}\{i \in \mathcal{S}^{(t)}\} \in \{0, 1\}$. Each $i \in \mathcal{S}^{(t)}$ performs E local gradient steps with local learning rate $\eta_{\ell,t} > 0$:

$$\mathbf{w}_i^{(t,0)} = \mathbf{w}^{(t-1)}, \quad \mathbf{w}_i^{(t,e)} = \mathbf{w}_i^{(t,e-1)} - \eta_{\ell,t} \nabla F_i(\mathbf{w}_i^{(t,e-1)}), \quad e = 1, \dots, E. \quad (2)$$

We denote the resulting local-update map by $h_i^{(t)}(\mathbf{w}^{(t-1)}) = \mathbf{w}_i^{(t,E)}$. With data-proportional weight $p_i^{(t)} = n_i / \sum_{m \in \mathcal{S}^{(t)}} n_m$, satisfying $\sum_{i \in \mathcal{S}^{(t)}} p_i^{(t)} = 1$, the server updates

$$\mathbf{w}^{(t)} = (1 - \gamma_t) \mathbf{w}^{(t-1)} + \gamma_t \sum_{i \in \mathcal{S}^{(t)}} p_i^{(t)} h_i^{(t)}(\mathbf{w}^{(t-1)}), \quad (3)$$

where $\gamma_t \in (0, 1]$ is the server mixing rate and standard FedAvg uses $\gamma_t = 1$. We write $\mathbf{g}_i^{(t)} = \nabla F_i(\mathbf{w}^{(t-1)})$ for client i 's gradient. After T_{learn} rounds the server obtains the trained global model $\mathbf{w}^{(T_{\text{learn}})}$, together with the training history which is recorded during training and is target-agnostic: it depends on neither $\mathcal{I}_{\text{tar}}$ nor any subsequent unlearning request (see details in Sec. 4).

3.2 Problem Definition

We study client-level FU as a two-stage process: a learning stage that produces a trained global model, followed by an unlearning stage that removes the influence of a designated target-client set.

Unlearning request. After training, a nonempty target-client set $\mathcal{I}_{\text{tar}} \subsetneq \mathcal{I}$ is designated for removal, and the retained set is $\mathcal{I}_{\text{ret}} = \mathcal{I} \setminus \mathcal{I}_{\text{tar}}$, with $|\mathcal{I}_{\text{ret}}| + |\mathcal{I}_{\text{tar}}| = N$. Let $n_{\text{ret}} = \sum_{i \in \mathcal{I}_{\text{ret}}} n_i$ and $n_{\text{tar}} = \sum_{o \in \mathcal{I}_{\text{tar}}} n_o$. The retained-client and target-client objectives are as follows:

$$F_{\text{ret}}(\mathbf{w}) = \sum_{i \in \mathcal{I}_{\text{ret}}} \frac{n_i}{n_{\text{ret}}} F_i(\mathbf{w}), \quad F_{\text{tar}}(\mathbf{w}) = \sum_{o \in \mathcal{I}_{\text{tar}}} \frac{n_o}{n_{\text{tar}}} F_o(\mathbf{w}). \quad (4)$$

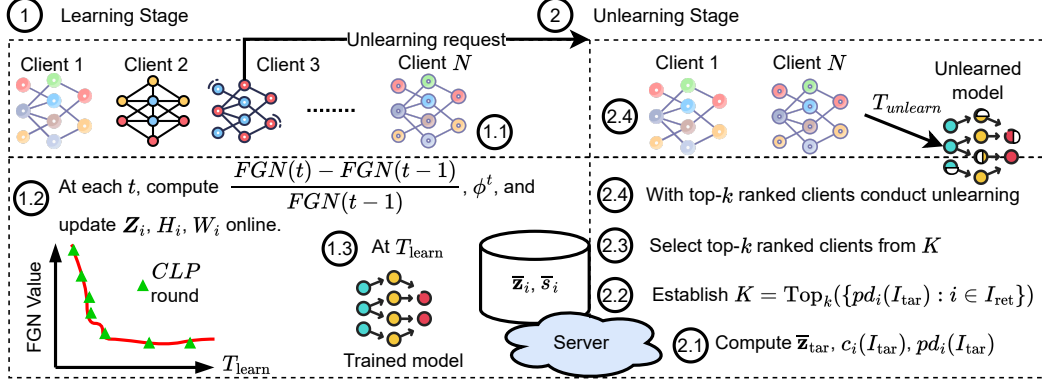


Figure 1: PDF framework with workflow.

Retraining reference and unlearning objective. Let $\mathcal{A}$ denote the original FL training algorithm, including its initialization, client sampling, and optimization randomness ξ . The retraining reference is $\mathbf{w}_{\text{retrain}}^* = \mathcal{A}(\{\mathcal{D}_i\}_{i \in \mathcal{I}_{\text{ret}}}; \xi)$, obtained by rerunning the training procedure from scratch using only retained clients. Given $\mathbf{w}^{(T_{\text{learn}})}$, $\mathcal{I}_{\text{tar}}$, training history, an unlearning algorithm returns $\mathbf{w}_{\text{unlearn}}$. We seek retraining-like retained and target behavior:

$$F_{\text{ret}}(\mathbf{w}_{\text{unlearn}}) - F_{\text{ret}}(\mathbf{w}_{\text{retrain}}^*) \leq \delta, \quad |F_{\text{tar}}(\mathbf{w}_{\text{unlearn}}) - F_{\text{tar}}(\mathbf{w}_{\text{retrain}}^*)| \leq \varepsilon_f, \quad (5)$$

where $\delta, \varepsilon_f \geq 0$ denote the retained-utility and forgetting tolerances, respectively.

4 Path-Dependent Client Selection for Unlearning

PDF operates in two stages as shown in Fig. 1. During federated training, the server detects CLPs and maintains target-agnostic, CLP-weighted histories of each client’s update direction and global alignment. Once an unlearning request reveals the target set $\mathcal{I}_{\text{tar}}$, these histories are used to measure each retained client’s coherence with the target clients and rank its suitability for unlearning. This design neither requires the target set to be known during training nor stores per-round model snapshots. To reduce storage and computation, all statistics are computed on the classification layer, which is sensitive to client-specific and non-IID variation [18, 20, 24, 31, 38].

4.1 Learning stage

Critical learning period detection. The onset and duration of CLPs vary across training trajectories and are not confined to a fixed early interval (see Fig. 2) [13]. Let $\mathcal{T}_{\text{CLP}} = \bigcup_{q=1}^Q \mathcal{T}_{\text{CLP}}^{(q)} \subseteq \{1, \dots, T_{\text{learn}}\}$ denote the set of detected critical rounds, where Q is the number of identified critical periods. To detect CLPs online, the server employs the federated gradient norm (FGN) proxy [37]. Given the round- t local gradient $\mathbf{g}_i^{(t)}$, one local training step produces the first-order loss change: $\Delta F_i^{(t)} = F_i(\mathbf{w}^{(t-1)} - \eta_{\ell,t} \mathbf{g}_i^{(t)}) - F_i(\mathbf{w}^{(t-1)}) \approx -\eta_{\ell,t} \|\mathbf{g}_i^{(t)}\|_2^2$. Accordingly, the round-level loss-decrease proxy is defined as $\text{FGN}^{(t)} \triangleq \eta_{\ell,t} \sum_{i \in \mathcal{S}^{(t)}} p_i^{(t)} \|\mathbf{g}_i^{(t)}\|_2^2$. A round t is flagged as critical (i.e. $t \in \mathcal{T}_{\text{CLP}}$), when $\frac{\text{FGN}^{(t)} - \text{FGN}^{(t-1)}}{\text{FGN}^{(t-1)}} \geq \tau_{\text{CLP}}$, where $\tau_{\text{CLP}} > 0$ is an empirically selected sensitivity threshold [13, 37]. Evaluating this criterion throughout training enables the detection of both early and late critical periods (see Fig. 2).

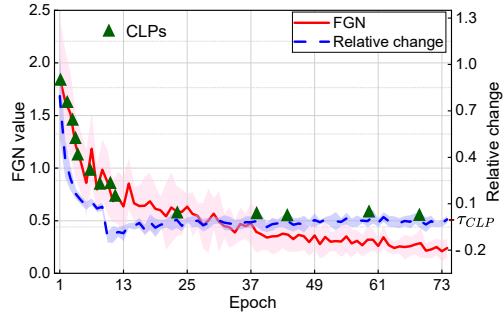


Figure 2: Illustration of CLP detection using the FGN proxy on FMNIST [34], where the green double-headed arrows and triangles indicate the detected CLPs.

Throughout, $\theta_i^{\text{cls},(t)} \in \mathbb{R}^{d_{\text{cls}}}$ denotes the vectorized classification-layer parameters of client i after local training at round t , and $\theta^{\text{cls},(t)} \in \mathbb{R}^{d_{\text{cls}}}$ denotes the corresponding aggregated global parameters, where d_{cls} is the classification-layer dimension.

Definition 1 (Server Alignment) For each $i \in \mathcal{S}^{(t)}$, we define client and global classification-layer updates as: $\delta_i^{\text{cls},(t)} = \theta_i^{\text{cls},(t)} - \theta^{\text{cls},(t-1)}$, $\delta_g^{\text{cls},(t)} = \theta^{\text{cls},(t)} - \theta^{\text{cls},(t-1)}$, respectively. Therefore, the round-level server alignment of client i is

$$s_i^{(t)} = \widehat{\cos}(\delta_i^{\text{cls},(t)}, \delta_g^{\text{cls},(t)}), \quad (6)$$

where $\widehat{\cos}(\mathbf{a}, \mathbf{b}) = \frac{1}{2} \left(1 + \frac{\langle \mathbf{a}, \mathbf{b} \rangle}{(\|\mathbf{a}\|_2 + \varepsilon_0)(\|\mathbf{b}\|_2 + \varepsilon_0)} \right)$ is the normalized cosine similarity and $\varepsilon_0 > 0$ ensures numerical stability. A larger $s_i^{(t)}$ indicates stronger consistency between client i 's local update and the realized global update at round t .

Definition 2 (CLP-Aware Historical Accumulation) To emphasize contributions made during critical rounds, we assign round t the temporal weight: $\phi^{(t)} = \begin{cases} 1 + \lambda, & t \in \mathcal{T}_{\text{CLP}}, \\ 1, & \text{otherwise,} \end{cases}$ where $\lambda \geq 0$ controls the additional weight assigned to CLP contributions. For each participating client, let

$$\widehat{\delta}_i^{\text{cls},(t)} = \frac{\delta_i^{\text{cls},(t)}}{\|\delta_i^{\text{cls},(t)}\|_2 + \varepsilon_0} \quad (7)$$

denote its normalized update direction. The server initializes $\mathbf{Z}_i^{(0)} = \mathbf{0}$, $H_i^{(0)} = 0$, and $W_i^{(0)} = 0$, and recursively updates

$$\mathbf{Z}_i^{(t)} = \mathbf{Z}_i^{(t-1)} + \phi^{(t)} a_i^{(t)} \widehat{\delta}_i^{\text{cls},(t)}, \quad H_i^{(t)} = H_i^{(t-1)} + \phi^{(t)} a_i^{(t)} s_i^{(t)}, \quad W_i^{(t)} = W_i^{(t-1)} + \phi^{(t)} a_i^{(t)}. \quad (8)$$

At T_{learn} , the accumulated update direction and historical server alignment of client i are as follows:

$$\mathbf{z}_i = \frac{\mathbf{Z}_i^{(T_{\text{learn}})}}{W_i^{(T_{\text{learn}})} + \varepsilon_0}, \quad \bar{\mathbf{z}}_i = \frac{\mathbf{z}_i}{\|\mathbf{z}_i\|_2 + \varepsilon_0}, \quad \bar{s}_i = \frac{H_i^{(T_{\text{learn}})}}{W_i^{(T_{\text{learn}})} + \varepsilon_0}. \quad (9)$$

Here, $\bar{\mathbf{z}}_i$ summarizes the dominant direction of client i 's historical updates, emphasizing contributions made during CLPs, while $\bar{s}_i$ measures its average historical alignment with the global trajectory. These target-agnostic statistics require one d_{cls} -dimensional vector and two scalars per client.

4.2 Unlearning stage

Once an unlearning request reveals the target-client set $\mathcal{I}_{\text{tar}}$, the server uses the cached learning-stage histories to quantify each retained client's dependence on the target set.

Definition 3 (Target Coherence) The aggregate historical direction of the target clients is $\bar{\mathbf{z}}_{\text{tar}} = \frac{\sum_{o \in \mathcal{I}_{\text{tar}}} \bar{\mathbf{z}}_o}{\|\sum_{o \in \mathcal{I}_{\text{tar}}} \bar{\mathbf{z}}_o\|_2 + \varepsilon_0}$. For each retained client $i \in \mathcal{I}_{\text{ret}}$, its target coherence is defined by:

$$c_i(\mathcal{I}_{\text{tar}}) = \widehat{\cos}(\bar{\mathbf{z}}_i, \bar{\mathbf{z}}_{\text{tar}}). \quad (10)$$

A larger $c_i(\mathcal{I}_{\text{tar}})$ indicates stronger alignment between the historical update directions of client i and the target set, whereas $1 - c_i(\mathcal{I}_{\text{tar}})$ measures historical target decoupling. **Selecting clients with larger decoupling reduces the tendency of retained-client updates to restore target-related knowledge during the subsequent unlearning process** (see Tab. 4).

Definition 4 (Path-Dependence Unlearning Score) For each retained client $i \in \mathcal{I}_{\text{ret}}$, we define the path-dependence score as

$$pd_i(\mathcal{I}_{\text{tar}}) = \beta \bar{s}_i + (1 - \beta) [1 - c_i(\mathcal{I}_{\text{tar}})]. \quad (11)$$

The parameter $\beta \in [0, 1]$ balances historical global alignment and target decoupling: larger values emphasize clients whose updates consistently followed the global learning trajectory, thereby favoring retained-model utility preservation, whereas smaller values emphasize clients whose historical directions are weakly coupled with those of the target clients, thereby favoring target forgetting.

A larger $pd_i(\mathcal{I}_{\text{tar}})$ therefore favors retained clients that are both globally aligned and historically decoupled from the target set, with contributions during detected CLPS emphasized through $\phi^{(t)}$. The server selects the k highest-scoring retained clients:

$$\mathcal{K} = \text{Top}_k(\{pd_i(\mathcal{I}_{\text{tar}}) : i \in \mathcal{I}_{\text{ret}}\}), \text{ with } k \leq |\mathcal{I}_{\text{ret}}|. \quad (12)$$

Only clients in $\mathcal{K}$ participate in the subsequent T_{unlearn} unlearning rounds. The persistent learning-stage cache therefore requires $\mathcal{O}(Nd_{\text{cls}})$ space, while unlearning-stage communication and computation scale with k selected clients rather than all $|\mathcal{I}_{\text{ret}}|$ retained clients.

5 Experimental evaluation

5.1 Experiment setup

Unlearning scenario. We simulate Byzantine attacks via label flipping, where a single client’s training labels are maliciously manipulated, triggering an unlearning request for that client [40].

Hyperparameter, dataset, and model. We simulate an FL system using FedAvg [21] for global model training with $N = 64$ clients. Unless otherwise stated, $M = 16$ clients are sampled per training round, with $\eta_{i,t} = 0.01$, batch size 32, CLP detection threshold $\tau_{\text{CLP}} = 0.01$. We consider both IID and Dirichlet-based non-IID data partitions with $\alpha \in \{1, 0.1\}$, where a smaller α indicates stronger data heterogeneity. For FU, we set $\beta \in \{0.1, 0.5, 0.9\}$ to control the trade-off between retained knowledge preservation and the removal of target-client influence, with smaller β yielding stronger forgetting; $\lambda = 0.7$, which adds weights to participation during CLPS; and $T_{\text{unlearn}} = 20$. We evaluate on FMNIST [34], CIFAR-10, and CIFAR-100 [2]. A CNN architecture is used for FMNIST (see Appendix A.1), while ResNet-18 [12] is adopted for CIFAR-10 and CIFAR-100.

Baselines. We compare PDF with Retrain, CONDA [6, 7], FedRecovery [39], FOUL [22], FedEraser [19], Exact-Fun (E-F) [35], FUSED [40], and GDFA [33] (see details in Appendix A.2).

Evaluation metric. We evaluate unlearning efficacy, efficiency, and fidelity, using the following active testing metrics [6, 10, 22, 23, 33, 40]:

- **FA & RA (%)**: Forgetting accuracy (FA) measures predictive accuracy on the target-client data, where lower FA indicates stronger removal of target-client influence. Retained accuracy (RA) measures accuracy on the retained data, where higher RA indicates better preservation of retained knowledge.
- **MIA AUC (%)**: We assess residual privacy leakage using membership inference attacks (MIA) [28]. AUC measures member–non-member distinguishability across decision thresholds, with 50% corresponding to random discrimination. We use Retrain as the reference and report the MIA-AUC gap to Retrain, where a smaller gap indicates more retraining-like privacy behavior.
- **Comm (GB) & Comp (PFLOPs)**: Total data transmitted and total floating-point operations (FLOPs), respectively, each aggregated over training and unlearning. We include training because several learning-to-unlearn baselines modify the training protocol, so measuring the unlearning stage alone would understate their cost.
- **Space Complexity**: We measure the system-wide memory required to store model checkpoints, historical statistics, and intermediate model updates during training and unlearning.

5.2 Main result

As shown in Tab. 1, PDF achieves effective forgetting while preserving retained knowledge using only 25 clients for unlearning. On FMNIST, PDF remains competitive with the strongest baselines, reaching 86.89% RA under $\alpha = 0.1$ while the forgetting-oriented variant achieves 0% FA. Its advantage becomes clearer on the more challenging CIFAR tasks. Under heterogeneous CIFAR-10 (i.e., $\alpha = 0.1$), PDF_{0.9} achieves 57.10% RA, exceeding both Retrain (55.28%) and FOUL (52.10%), while PDF_{0.5} combines 54.48% RA with 0% FA. On CIFAR-100, all three variants consistently outperform the strongest competing baseline in RA under both data distributions, reaching up to 49.13% versus 42.22% under $\alpha = 0.1$ and 54.38% versus 47.50% under $\alpha = 1.0$. ***These results support our central hypothesis that retained clients differ in their value for unlearning, and***

Table 1: Performance comparison between PDF and other FU methods on three datasets under $\alpha \in \{0.1, 1.0\}$. Subscripts on PDF denote $\beta \in \{0.1, 0.5, 0.9\}$, which controls the forgetting–utility trade-off: 0.1 is **forgetting-oriented**, 0.5 **balances forgetting and utility preservation**, and 0.9 is **utility-preservation-oriented**. All variants use $k = 25$ selected unlearning clients. Values in parentheses are deltas relative to Retrain. For FA, RA, and MIA AUC, **red** marks the best and **green** the second-best method across all unlearning methods (Retrain is the reference and is excluded), ranked by FA ($\downarrow$), RA ($\uparrow$), and $|\Delta\text{MIA}|$ ($\downarrow$).

Metric	α	Retrain	FUSED	CONDA	E-F	FedEraser	FOUL	FedRecovery	GDFA	PDF _{0.1}	PDF _{0.5}	PDF _{0.9}
<i>FMNIST-CNN</i>												
FA	0.1	0.44	0.44 (+0.00)	0.70 (+0.26)	0.56 (+0.12)	1.11 (+0.67)	0.75 (+0.31)	0.60 (+0.16)	0.28 (−0.16)	0.00 (−0.44)	0.72 (+0.28)	0.44 (+0.00)
	1.0	0.92	0.82 (−0.10)	0.94 (+0.02)	0.45 (−0.47)	0.53 (−0.39)	0.77 (−0.15)	0.87 (−0.06)	0.14 (−0.78)	0.28 (−0.64)	0.28 (−0.64)	0.92 (+0.00)
RA	0.1	88.20	87.14 (−1.06)	85.17 (−3.02)	81.89 (−6.31)	61.04 (−27.16)	85.60 (−2.60)	85.22 (−2.98)	81.83 (−6.37)	82.60 (−5.60)	85.97 (−2.23)	86.89 (−1.31)
	1.0	90.45	89.60 (−0.85)	89.40 (−1.05)	85.94 (−4.50)	76.42 (−14.03)	89.64 (−0.81)	89.64 (−0.81)	88.25 (−2.20)	89.34 (−1.11)	89.35 (−1.10)	89.55 (−0.90)
MIA AUC	0.1	50.32	50.05 (−0.27)	50.11 (−0.20)	49.84 (−0.48)	49.92 (−0.40)	49.71 (−0.61)	50.01 (−0.31)	50.00 (−0.32)	50.01 (−0.31)	50.00 (−0.32)	49.91 (−0.41)
	1.0	50.23	49.79 (−0.44)	50.09 (−0.15)	50.25 (+0.02)	50.08 (−0.15)	49.82 (−0.41)	50.16 (−0.07)	50.17 (−0.06)	50.13 (−0.10)	50.25 (+0.02)	50.19 (−0.04)
Comm	0.1/1.0	15.62	4.11	3.97	7.82	4.36	6.08	3.97	3.98	4.79	4.79	4.79
Comp	0.1	2.09	1.01 (−1.39)	0.71 (−1.39)	1.22 (−0.87)	0.72 (−1.37)	1.71 (−0.38)	0.71 (−1.39)	0.71 (−1.38)	0.83 (−1.26)	0.88 (−1.22)	0.89 (−1.21)
	1.0	2.10	1.01 (−1.09)	0.71 (−1.39)	1.23 (−0.87)	0.73 (−1.37)	1.72 (−0.38)	0.71 (−1.39)	0.71 (−1.39)	0.83 (−1.27)	0.83 (−1.27)	0.84 (−1.26)
<i>CIFAR-10-ResNet</i>												
FA	0.1	6.93	6.14 (−0.79)	7.52 (+0.59)	2.21 (−4.72)	4.17 (−2.76)	1.86 (−5.07)	0.77 (−6.16)	4.45 (−2.48)	0.42 (−6.51)	0.00 (−6.93)	0.39 (−6.54)
	1.0	0.72	1.38 (+0.66)	11.18 (+10.46)	3.93 (+3.20)	7.97 (+7.25)	1.96 (+1.24)	5.87 (+5.15)	0.33 (−0.39)	0.89 (+0.17)	1.09 (+0.37)	0.94 (+0.22)
RA	0.1	55.28	49.85 (−5.43)	14.68 (−40.60)	23.10 (−32.18)	9.68 (−45.60)	52.10 (−3.18)	37.12 (−18.16)	38.61 (−16.67)	45.22 (−10.06)	54.48 (−0.80)	57.10 (+1.82)
	1.0	86.93	83.56 (−3.38)	9.01 (−77.93)	52.40 (−34.53)	24.67 (−62.26)	82.65 (−4.28)	46.83 (−40.10)	69.34 (−17.59)	82.28 (−4.65)	82.84 (−4.09)	84.08 (−2.85)
MIA AUC	0.1	55.64	61.43 (+5.79)	62.98 (+7.34)	50.13 (−5.51)	49.87 (−5.78)	60.73 (+5.09)	55.16 (−0.49)	57.53 (+1.89)	54.62 (−1.02)	54.76 (−0.88)	56.69 (+1.05)
	1.0	52.18	54.30 (+2.12)	53.38 (+1.20)	49.33 (−2.85)	50.54 (−1.65)	53.11 (+0.93)	51.21 (−0.98)	51.82 (−0.36)	51.74 (−0.44)	51.78 (−0.40)	51.58 (−0.60)
Comm	0.1/1.0	845.48	217.29	214.73	423.39	235.88	329.25	214.73	215.35	259.46	259.46	259.46
Comp	0.1	47.45	22.38 (−25.07)	16.01 (−31.45)	27.69 (−19.76)	16.42 (−31.03)	38.45 (−9.00)	16.01 (−31.45)	16.02 (−31.43)	18.73 (−28.72)	19.30 (−28.15)	20.02 (−27.43)
	1.0	47.63	22.50 (−25.14)	16.10 (−31.53)	27.84 (−19.79)	16.49 (−31.14)	38.66 (−8.97)	16.10 (−31.53)	16.11 (−31.52)	18.70 (−28.93)	18.76 (−28.87)	18.88 (−28.75)
<i>CIFAR-100-ResNet</i>												
FA	0.1	0.27	1.26 (+0.99)	0.86 (+0.59)	1.46 (+1.19)	0.84 (+0.56)	1.50 (+1.23)	0.86 (+0.59)	0.86 (+0.59)	0.00 (−0.27)	0.86 (+0.59)	0.52 (+0.25)
	1.0	0.60	0.91 (−0.30)	0.54 (−0.07)	0.40 (−0.21)	1.25 (+0.64)	1.00 (+0.40)	1.44 (+0.83)	0.08 (−0.52)	0.44 (−0.16)	0.91 (+0.31)	0.52 (−0.08)
RA	0.1	55.17	36.34 (−18.83)	4.35 (−50.82)	16.98 (−38.20)	3.98 (−51.19)	42.22 (−12.95)	1.03 (−54.14)	42.15 (−13.02)	48.63 (−6.54)	48.43 (−6.74)	49.13 (−6.04)
	1.0	58.92	31.83 (−27.09)	2.12 (−56.80)	23.80 (−35.12)	9.10 (−49.82)	39.81 (−19.11)	14.46 (−44.46)	47.50 (−11.42)	54.38 (−4.54)	54.24 (−4.68)	54.21 (−4.71)
MIA AUC	0.1	51.44	54.02 (+2.59)	55.48 (+4.04)	55.21 (+3.77)	50.71 (−0.73)	53.09 (+1.65)	49.96 (−1.48)	52.42 (+0.98)	52.13 (−0.69)	52.57 (+1.13)	52.43 (+0.99)
	1.0	62.15	53.91 (−8.24)	50.20 (−11.95)	54.90 (−7.25)	51.36 (−10.79)	50.49 (−11.66)	50.45 (−11.70)	49.73 (−12.42)	53.03 (−9.12)	52.39 (−9.76)	52.98 (−9.17)
Comm	0.1/1.0	848.97	218.19	215.61	425.14	236.86	330.60	215.61	216.24	260.53	260.53	260.53
Comp	0.1	47.75	22.63 (−25.11)	16.22 (−31.53)	28.02 (−19.72)	16.62 (−31.13)	38.93 (−8.82)	16.22 (−31.53)	16.23 (−31.52)	18.83 (−28.92)	18.87 (−28.88)	18.96 (−28.78)
	1.0	47.77	22.59 (−25.18)	16.17 (−31.60)	27.96 (−19.81)	16.57 (−31.20)	38.82 (−8.95)	16.17 (−31.60)	16.18 (−31.59)	18.72 (−29.05)	18.73 (−29.04)	18.75 (−29.02)

that exploiting their path-dependent training histories enables better preservation of retained knowledge. The variants further demonstrate the intended role of β in controlling the forgetting–utility trade-off. Increasing β emphasizes historical global alignment and generally favors utility preservation, whereas smaller β emphasizes target decoupling and stronger forgetting. For example, under $\alpha = 0.1$, RA increases from 82.60% to 86.89% on FMNIST and from 45.22% to 57.10% on CIFAR-10 as β increases from 0.1 to 0.9. The weaker variation across β on CIFAR-100 suggests that PDF is less sensitive to the trade-off parameter in this setting. Finally, PDF reduces communication by approximately 69% and computation by 57–61% relative to Retrain, demonstrating a favorable trade-off among forgetting, retained utility, and unlearning efficiency.

As shown in Tab. 2, PDF is space-efficient because its learning-stage cache stores only one accumulated classification-layer update direction and two scalar statistics per client, yielding $\mathcal{O}(Nd_{\text{cls}})$ storage independent of the number of learning rounds, where $d_{\text{cls}} \ll P$. During unlearning, PDF operates on only the selected k retained clients, so the model and update tensors instantiated during computation scale with $\mathcal{O}(kP)$ rather than all retained clients, where $k < |\mathcal{I}_{\text{ret}}|$. In comparison, Retrain and FUSED involve all retained clients, although FUSED reduces per-client storage and computation through lightweight LoRA parameters. CONDA and FOUL maintain client-wise full-model information, while GDFA constructs and fine-tunes multiple perturbed full-model instances. FedEraser additionally stores checkpointed model histories, whereas E-F and FedRecovery retain full-model information across learning rounds.

5.3 Ablation study

As shown in Tab. 3, increasing k generally improves retained utility, but the gains diminish as more clients participate. At $\beta=0.5$, for example, RA rises from 79.26% at $k=10$ to 85.97% at $k=25$,

Table 2: Space complexity analysis.

Method	Space Complexity
Retrain	$\mathcal{O}(P + \mathcal{I}_{\text{ret}} \cdot P)$
FUSED	$\mathcal{O}(P + \mathcal{I}_{\text{ret}} \cdot A)$
CONDA	$\mathcal{O}(N \cdot P + P)$
E-F	$\mathcal{O}(T_{\text{learn}} \cdot M \cdot P + P)$
FedEraser	$\mathcal{O}(H \cdot M \cdot P + \mathcal{I}_{\text{ret}} \cdot P)$
FOUL	$\mathcal{O}(P + N \cdot P + N^2)$
FedRecovery	$\mathcal{O}(T_{\text{learn}} \cdot M \cdot P + P)$
GDFA	$\mathcal{O}(K_{\text{per}} \cdot P + \mathcal{I}_{\text{tar}} \cdot P)$
PDF	$\mathcal{O}(N \cdot d_{\text{cls}} + k \cdot P)$

* $P = |\mathbf{w}|$ is the number of full-model parameters.

† A is the number of LoRA unlearning parameters, $A \ll P$.

†† $d_{\text{cls}} \ll P$.

‡ H is the number of FedEraser checkpoints, $H < T_{\text{learn}}$.

• K_{per} is the number of GDFA perturbed models.

◊ k is the number of retained clients selected by PDF, $k < |\mathcal{I}_{\text{ret}}|$.

but increases only marginally to 86.40% at $k=30$. A similar saturation occurs for $\beta=0.9$, where increasing k from 20 to 25 improves RA by 0.9 percentage points, compared with only 0.1 percentage points from 25 to 30. In contrast, communication and computation increase steadily with k , revealing a clear utility–efficiency trade-off. MIA AUC remains within 0.41 percentage points of Retrain throughout the sweep, indicating that this trade-off does not alter retraining-like privacy behavior. The observed variation across β is consistent with the forgetting–utility behavior discussed in Tab. 1. We therefore adopt $k=25$ as the default.

Table 3: Ablation study on the number of selected unlearning clients k and knowledge retention–forgetting tradeoff β on FMNIST under $\alpha = 0.1$ with $\mathcal{I}_{\text{tar}} = 0$. MIA AUC values in parentheses denote differences from Retrain (50.32). The shaded column denotes the default $k = 25$ in Tab. 1.

	k	10	15	20	25	30	35
$\beta = 0.1$	FA	0.00	0.00	0.00	0.00	0.28	0.28
	RA	75.65	80.51	81.26	82.60	84.45	85.18
	MIA AUC	49.96 (−0.36)	50.47 (+0.15)	50.09 (−0.23)	50.01 (−0.31)	50.02 (−0.30)	49.95 (−0.37)
	Comp	0.76	0.78	0.81	0.83	0.85	0.87
$\beta = 0.5$	FA	0.28	0.28	0.28	0.72	0.44	0.28
	RA	79.26	82.76	84.43	85.97	86.40	86.37
	MIA AUC	49.99 (−0.33)	50.13 (−0.19)	50.04 (−0.28)	50.00 (−0.32)	50.13 (−0.19)	50.05 (−0.27)
	Comp	0.79	0.82	0.85	0.88	0.89	0.91
$\beta = 0.9$	FA	0.72	1.27	0.99	0.44	0.44	0.28
	RA	84.24	85.49	85.99	86.89	86.99	86.76
	MIA AUC	50.26 (−0.06)	50.22 (−0.10)	50.16 (−0.15)	49.91 (−0.41)	50.07 (−0.25)	50.05 (−0.27)
	Comp	0.81	0.85	0.87	0.89	0.91	0.92
Comm		4.30	4.46	4.63	4.79	4.96	5.13

Variants of PDF. Table 4 evaluates the contribution of the proposed client-selection rule while keeping the unlearning procedure and number of selected clients fixed. Under highly heterogeneous data ($\alpha = 0.1$), PDF consistently achieves higher RA than Rev, high-Coh, and non-CLP. Compared with Rev, which also use $pd_i(\mathcal{I}_{\text{tar}})$ for client selection but in a reverse order, it improves RA by 3.69 percentage points on FMNIST and 13.39 points on CIFAR-10, while reducing FA by 0.64 and 4.44 points, respectively. Selecting highly target-coherent clients (high-Coh) further degrades both FA and RA, particularly on CIFAR-10, where FA increases from 0% to 7.65% and RA decreases from 54.48% to 37.65%. Removing CLP weighting (non-CLP) also reduces RA, by 0.62 points on FMNIST and 4.88 points on CIFAR-10, showing that emphasizing critical rounds improves client selection. Under $\alpha = 1.0$, the differences are smaller, although PDF retains the highest RA on both datasets. Across all variants, $|\Delta\text{MIA}|$ remains at most 1.28 percentage points from Retrain. **These results show that the proposed ranking direction and CLP-aware history contribute to selecting retained clients that preserve utility while maintaining effective forgetting, supporting our central hypothesis.**

Table 4: Ablation of PDF against its selection variants, with $\beta = 0.5$, $\lambda = 0.7$, and $k = 25$ clients selected for unlearning. ΔMIA is the signed deviation of MIA AUC from the retrained reference.

Method	$\alpha = 0.1$			$\alpha = 1.0$		
	FA	RA	ΔMIA	FA	RA	ΔMIA
<i>FMNIST-CNN</i> (Retrain MIA AUC: 50.32 / 50.23)						
PDF	0.72	85.97	−0.32	0.28	89.35	+0.02
non-CLP	0.28	85.35	−0.16	0.43	89.26	+0.02
Rev	1.36	82.28	+0.04	1.33	89.14	−0.06
high-Coh	1.24	82.15	−0.34	1.23	89.04	+0.03
<i>CIFAR-10-ResNet</i> (Retrain MIA AUC: 55.64 / 52.18)						
PDF	0.00	54.48	−0.88	1.09	82.84	−0.40
non-CLP	0.19	49.60	−1.10	1.51	81.51	+0.12
Rev	4.44	41.09	−0.77	1.30	81.91	+0.15
high-Coh	7.65	37.65	−1.28	1.20	81.74	+0.18

6 Conclusion

We revisit client-level FU from a path-dependent perspective, motivated by the insight that retained clients are not equally valuable for unlearning: their usefulness depends on when and how their influence was accumulated during training and how strongly it is coupled with the target clients. Based on this insight, PDF combines CLP-weighted historical global-model alignment and client-target coherence to rank retained clients, and performs retraining-based unlearning using only a selected top- k subset. Experiment results demonstrate that PDF achieves effective forgetting and strong retained utility while substantially reducing client participation, communication, and computation, and can be adapted to different application requirements by adjusting the forgetting–utility trade-off.

Acknowledgments and Disclosure of Funding

This publication has emanated from research conducted with the financial support of Research Ireland under Grant number 18/CRT/6222; School of Computer Science and Statistics, Trinity College Dublin. The work of Quoc-Viet Pham is supported in part by Research Ireland under the European Innovation Council CHIST-ERA SHIELD project (Project No. 216449, Award No. 19226). For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

References

- [1] Alessandro Achille, Matteo Rovere, and Stefano Soatto. Critical learning periods in deep networks. In *International Conference on Learning Representations*, 2018.
- [2] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. The cifar-10 and cifar-100 datasets. <https://cave.cs.toronto.edu/kriz/cifar.html>, 2009. Accessed: 2026-06-18.
- [3] Cristiano Antonelli. The economics of path-dependence in industrial organization. *International Journal of Industrial Organization*, 15(6):643–675, 1997.
- [4] Jonathan Brophy and Daniel Lowd. Machine unlearning for random forests. In *International Conference on Machine Learning*, pages 1092–1104, 2021.
- [5] Yinzhi Cao and Junfeng Yang. Towards making systems forget with machine unlearning. In *2015 IEEE Symposium on Security and Privacy*, pages 463–480, 2015.
- [6] Vikram S Chundawat, Pushkar Niroula, Prasanna Dhungana, Stefan Schoepf, Murari Mandal, and Alexandra Brintrup. Conda: Fast federated unlearning with contribution dampening. *arXiv preprint arXiv:2410.04144*, 2024.
- [7] Jack Foster, Stefan Schoepf, and Alexandra Brintrup. Fast machine unlearning without retraining through selective synaptic dampening. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 12043–12051, 2024.
- [8] Aditya Golatkar, Alessandro Achille, and Stefano Soatto. Eternal sunshine of the spotless net: Selective forgetting in deep networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9304–9312, 2020.
- [9] Qinghui Gong, Xue Yang, Xunlei Chen, Jinshan Lai, Hua Meng, and Xiaohu Tang. FedOrtho: Efficient federated unlearning via orthogonal convolution and adaptive soft pruning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8009–8018, 2026.
- [10] Hanlin Gu, Win K Ong, Chee S Chan, and Lixin Fan. Ferrari: Federated feature unlearning via optimizing feature sensitivity. *Advances in Neural Information Processing Systems*, 37: 24150–24180, 2024.
- [11] Thierry Hanser, Ernst Ahlberg, Alexander Amberg, Lennart T Anger, Chris Barber, Richard J Brennan, Alessandro Brigo, Annie Delaunois, Susanne Glowienke, Nigel Greene, et al. Data-driven federated learning in drug discovery with knowledge distillation. *Nature Machine Intelligence*, 7(3):423–436, 2025.
- [12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [13] Chengxiang Huang and Bingyan Liu. Pa3fed: Period-aware adaptive aggregation for improved federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 17395–17403, 2025.
- [14] Hyejun Jeong, Shiqing Ma, and Amir Houmansadr. Survey on federated unlearning: Challenges and opportunities. *IEEE Transactions on Big Data*, 12(3):702–720, 2026.

- [15] Jinghan Jia, Jiancheng Liu, Parikshit Ram, Yuguang Yao, Gaowen Liu, Yang Liu, Pranay Sharma, and Sijia Liu. Model sparsity can simplify machine unlearning. *Advances in Neural Information Processing Systems*, 36:51584–51605, 2023.
- [16] Yasser H Khalil, Leo Brunswic, Soufiane Lamghari, Xu Li, Mahdi Beitollahi, and Xi Chen. Not: Federated unlearning via weight negation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 25759–25769, 2025.
- [17] Michael Kleinman, Alessandro Achille, and Stefano Soatto. Critical learning periods emerge even in deep linear networks. In *International Conference on Learning Representations*, 2024.
- [18] Zexi Li, Xinyi Shang, Rui He, Tao Lin, and Chao Wu. No fear of classifier biases: Neural collapse inspired federated learning with synthetic and fixed classifier. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5319–5329, 2023.
- [19] Gaoyang Liu, Xiaoqiang Ma, Yang Yang, Chen Wang, and Jiangchuan Liu. Federaser: Enabling efficient client-level data removal from federated learning models. In *2021 IEEE/ACM 29th international symposium on quality of service (IWQOS)*, pages 1–10, 2021.
- [20] Mi Luo, Fei Chen, Dapeng Hu, Yifan Zhang, Jian Liang, and Jiashi Feng. No Fear of Heterogeneity: Classifier calibration for federated learning with non-iid data. *Advances in Neural Information Processing Systems*, 34:5972–5984, 2021.
- [21] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*, pages 1273–1282, 2017.
- [22] Minh-Duong Nguyen, Senura Wanasekara, Le-Tuan Nguyen, Quoc-Viet Pham, Ken-Tye Yong, Nguyen H Tran, and Dung D Le. Computation and communication efficient federated unlearning via on-server gradient conflict mitigation and expression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3347–3357, 2026.
- [23] Thanh Linh Nguyen, Marcela Tuler De Oliveira, An Braeken, Aaron Yi Ding, and Quoc-Viet Pham. Toward Verifiable Federated Unlearning: Framework, Challenges, and the Road Ahead. *IEEE Internet Computing*, 30(01):59–70, 2026.
- [24] Jaehoon Oh, SangMook Kim, and Se-Young Yun. FedBABU: Toward enhanced representation for federated image classification. In *International Conference on Learning Representations*, 2022.
- [25] Zibin Pan, Zhichao Wang, Chi Li, Kaiyan Zheng, Boqi Wang, Xiaoying Tang, and Junhua Zhao. Federated unlearning with gradient descent and conflict mitigation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 19804–19812, 2025.
- [26] European Parliament and Council of the European Union. General Data Protection Regulation. [Online]. Available: "<https://gdpr-info.eu/>", 2016. (Accessed Jun. 07, 2026).
- [27] Nicolò Romandini, Alessio Mora, Carlo Mazzocca, Rebecca Montanari, and Paolo Bellavista. Federated unlearning: A survey on methods, design guidelines, and evaluation metrics. *IEEE Transactions on Neural Networks and Learning Systems*, 36(7):11697–11717, 2024.
- [28] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *IEEE Symposium on Security and Privacy*, pages 3–18, 2017.
- [29] Stelios Theophanous, Per-Ivar Lønne, Ananya Choudhury, Maaïke Berbee, Charlotte Deijjen, Andre Dekker, Matthew Field, Maria Antonietta Gambacorta, Alexandra Gilbert, Marianne Grønlie Guren, et al. An international multi-centre study to develop and validate federated learning-based prognostic models for anal cancer. *Nature Communications*, 17(1):3956, 2026.
- [30] Guanjie Wang and Chen Chen. GR-Gauge: Cost-efficient training configuration by gauging the gradient redundancy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12934–12943, 2026.

- [31] Peng Wang, Shoupeng Lu, Hao Yin, Banglie Yang, Tianli Zhu, and Cheng Dai. FedCM: Client clustering and migration in federated learning via gradient path similarity and update direction deviation. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pages 6343–6351, 2025.
- [32] Weiqi Wang, Chenhan Zhang, Zhiyi Tian, and Shui Yu. FedU: Federated unlearning via user-side influence approximation forgetting. *IEEE Transactions on Dependable and Secure Computing*, 22(3):2550–2562, 2025.
- [33] Xiuting Weng, Ruizhi Pu, Yuanhang Yao, Kun Yue, Zhiwen Tang, and Lixing Yu. GDFA: Geometry-driven federated unlearning with directional task vector alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10346–10356, 2026.
- [34] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms. *arXiv:1708.07747*, 2017.
- [35] Zuobin Xiong, Wei Li, Yingshu Li, and Zhipeng Cai. Exact-Fun: An exact and efficient federated unlearning approach. In *2023 IEEE International Conference on Data Mining (ICDM)*, pages 1439–1444, 2023.
- [36] Gang Yan, Hao Wang, and Jian Li. Seizing critical learning periods in federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 8788–8796, 2022.
- [37] Gang Yan, Hao Wang, Xu Yuan, and Jian Li. Criticallf: A critical learning periods augmented client selection framework for efficient federated learning. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2898–2907, 2023.
- [38] Liping Yi, Han Yu, Zhuan Shi, Gang Wang, Xiaoguang Liu, Lizhen Cui, and Xiaoxiao Li. FedSSA: semantic similarity-based aggregation for efficient model-heterogeneous personalized federated learning. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 5371–5379, 2024.
- [39] Lefeng Zhang, Tianqing Zhu, Haibin Zhang, Ping Xiong, and Wanlei Zhou. Fedrecovery: Differentially private machine unlearning for federated learning frameworks. *IEEE Transactions on Information Forensics and Security*, 18:4732–4746, 2023.
- [40] Zhengyi Zhong, Weidong Bao, Ji Wang, Shuai Zhang, Jingxuan Zhou, Lingjuan Lyu, and Wei Yang Bryan Lim. Unlearning through knowledge overwriting: Reversible federated unlearning via selective sparse adapter. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 30661–30670, 2025.

A Details of Experimental Evaluation

A.1 CNN Architecture for FMNIST

The CNN architecture used for FMNIST consists of two modules: a feature extractor and a classifier. The feature extractor comprises three convolutional blocks. The first block applies a Conv2d layer with 1 input channel and 32 output channels (kernel size 3), followed by Batch Normalization, ReLU activation, and a Dropout2d layer (rate 0.3). The second block applies a Conv2d layer with 32 input channels and 64 output channels (kernel size 3, stride 1), followed by Batch Normalization, ReLU activation, and a 2×2 max-pooling layer. The third block applies a Conv2d layer with 64 input and output channels (kernel size 3), followed by Batch Normalization, ReLU activation, a 4×4 max-pooling layer, and a Dropout2d layer (rate 0.2). The resulting feature map of size $64 \times 3 \times 3$ is flattened into a 576-dimensional vector. The classifier consists of a fully connected layer ($576 \rightarrow 256$) with ReLU activation and Dropout (rate 0.1), followed by a final linear layer ($256 \rightarrow 10$). All convolutional and linear layers are initialized with Kaiming uniform initialization with zero biases.

A.2 Federated Unlearning Baselines

Below are details of state-of-the-art FU approaches which are used to evaluate the efficacy and effectiveness of PDF.

- **Retrain.** As the gold-standard unlearning baseline, Retrain removes the target clients entirely and trains from scratch on the remaining clients. While this guarantees exact unlearning, it is prohibitively expensive in practice due to substantial computational overhead and reliance on the participation availability of all remaining clients.
- **FedEraser** [19]. FedEraser approximates the retrained model without incurring its full cost by replaying periodic historical parameter updates cached from remaining clients through a calibrated aggregation procedure, achieving significant reductions in both time and storage overhead relative to full retraining.
- **Exact-Fun** (E-F) [35]. Exact-Fun (E-F) leverages quantized FL to stabilize the global model, enabling exact unlearning with high efficiency. Upon a deletion request, only the affected client retrains locally. Retraining of the global model is triggered solely when the quantized model state changes, allowing many requests to be resolved with minimal overhead.
- **CONDA** [6, 7]. CONDA is a server-side unlearning method that eliminates a target client’s influence without client participation. It tracks per-client parameter contributions throughout training, estimates the forgotten client’s relative influence across model parameters, and applies a targeted dampening factor to parameters most strongly associated with the forgotten client.
- **FedRecovery** [39]. FedRecovery is a server-side unlearning method that does not require retraining-based calibration and fine-tuning. It removes the target client’s accumulated gradient residuals from the trained global model cached from the learning phase. It also injects calibrated Gaussian noise to mask the remaining distance between the recovered model and the ideal retrained model, thereby making the two models statistically indistinguishable.
- **FUSED** [40]. FUSED confines unlearning to the most task-critical layers identified during training, attaching lightweight sparse adapters exclusively to these layers and fine-tuning them on remaining clients to overwrite the forgotten knowledge. This design substantially reduces unlearning cost while preserving reversibility.
- **FOUL** [22]. FOUL is an on-server unlearning method that first disentangles the model into causal and non-causal components during the learning-to-unlearn stage, and then performs unlearning on the domain-specific non-causal component. In the unlearning stage, FOUL constructs an aggregated unlearning gradient by optimizing learnable coefficients over client gradients, encouraging alignment with retain-client gradients while reducing alignment with target-client gradients. This design preserves retain-client knowledge while removing forget-client-specific influence without requiring the server to access client data.
- **GDFA** [33]. GDFA is a model-manipulation-based federated unlearning method that exploits loss-landscape geometry and task-vector alignment to improve unlearning under data heterogeneity. Upon an unlearning request, GDFA constructs multiple perturbed models around the trained global model and fine-tunes them on the forgetting data to derive task vectors representing the target knowledge; directionally consistent components are then merged through sign consensus and negated from the global model. By operating around flatter loss regions, GDFA aims to improve robustness to parameter modification and preserve retained knowledge, but requires request-time access to the forgetting data and incurs additional computation from generating and fine-tuning multiple perturbed models.