Retaliatory Attacks Against Federated Unlearning via Data Leakage

Xinyi Sheng, Wei Bao, Hequn Wang, Yuqin Liu, and Sen Fu

School of Computer Science, The University of Sydney, Australia
xinyi.sheng@sydney.edu.au, wei.bao@sydney.edu.au, hwan0565@uni.sydney.edu.au, yliu0720@uni.sydney.edu.au, sen.fu@sydney.edu.au

Abstract

Federated unlearning (FU) allows a participating client in a federated learning (FL) system to remove its contribution from the trained global model, thereby enforcing the client's "right to be forgotten" (RTBF). However, from the perspective of a client that does not request unlearning, the activation of the FU process may disrupt ongoing FL training and introduce additional computational and time overhead. In such cases, a client opposed to unlearning may be incentivized to retaliate against the unlearning client(s). In this work, we take the first step toward demonstrating the feasibility of such retaliatory behavior by exploiting the information leakage introduced during the FU process. Specifically, we propose a novel unlearning-induced membership inference attack (MIA) model, followed by a coarse-to-fine data generation method that enables an adversarial client to locally reconstruct the unlearned data. Building on this reconstruction, we introduce two targeted retaliatory attacks: (1) Anti-Unlearning Attack (AUA), which hinders the global model from successfully forgetting the data intended for removal, and (2) Discrimination-Unlearning Attack (DUA), which specifically degrades the global model's performance on the unlearned data. Extensive experiments across a variety of FU methods and settings validate the effectiveness of the proposed retaliatory attack framework.

Introduction

Recent data protection regulations, including the European Union's General Data Protection Regulation GDPR (Voigt and von dem Bussche 2017) and California Consumer Privacy Act CCPA (Harding et al. 2019), have emphasized the need to support the "right to be forgotten" (RTBF) for personal data used in training machine learning (ML) models. To address this need, machine unlearning (MU) techniques have been developed to enable the removal of specific data samples from trained ML models in centralized settings (Cao and Yang 2015; Bourtoule et al. 2021; Wang et al. 2024). Federated unlearning (FU) (Liu et al. 2023) extends this concept to federated learning (FL), a decentralized learning framework where multiple clients collaboratively train a shared global model while keeping their local data private (McMahan et al. 2017). In FU, participating clients can submit unlearning requests to remove their contributions from the shared model, thereby exercising their RTBF.

While most existing efforts on FU have focused on improving the efficiency of the unlearning process (Liu et al. 2022; Halimi et al. 2022; Liu et al. 2022; Zhang et al. 2023a; Che et al. 2023), enhancing the utility of the global model after unlearning (Liu et al. 2021; Wu, Zhu, and Mitra 2023; Xiong et al. 2023), and ensuring certified removal of the unlearned data (Wang et al. 2022; Wu, Zhu, and Mitra 2023; Gao et al. 2024), very little attention has been paid to the potential security vulnerabilities introduced by the FU mechanism itself (Wang, Li, and Li 2023). Although a few recent studies have begun to examine security issues within the FU framework (Sheng, Bao, and Ge 2024; Wang et al. 2025), they primarily focus on adversarial threats directly arising from malicious unlearning requests.[1] In contrast to these works, we take the first step to investigate a more fundamental security concern in FU: the data privacy leakage that can occur during the unlearning process.

To demonstrate that such data privacy leakage constitutes a realistic and practical threat, we introduce a novel class of attacks termed *retaliatory attacks*, which are launched by an FL participant who opposes the unlearning process. This scenario is both intuitive and plausible in practice. From the perspective of a client who does not request unlearning, the activation of the FU process may disrupt the ongoing federated training, incur additional computational and time overhead, and potentially degrade the performance of the global model after unlearning. These consequences can serve as incentives for a dissatisfied client to retaliate against the one who initiated the unlearning request.

Specifically, we propose two targeted retaliatory attacks: (1) *Anti-Unlearning Attack (AUA)*, which aims to prevent the global model from successfully forgetting the private data that was requested to be unlearned; and (2) *Discrimination-Unlearning Attack (DUA)*, which seeks to intentionally degrade the global model's performance on the data and the underlying distribution associated with the unlearned client(s). The core idea behind these attacks is that an adversarial client can exploit the data leakage introduced during

[1] Due to space limitation, a more comprehensive discussion of related work, including studies that are less directly aligned with our work, is provided in the Appendix A (Sheng et al. 2025).

the FU procedure to reconstruct the unlearned data originally contributed by the unlearning client(s). To reconstruct the unlearned data, one class of ideal attacks is the gradient inversion attack (Zhu, Liu, and Han 2019), which aims to recover private data from uploaded gradients. However, such attacks typically assume a compromised or curious server with direct access to client gradients (Zhang et al. 2023b; Lamri et al. 2025; Zhang et al. 2025), which is not feasible in our setting where the adversary is merely a normal client. Thus, we propose to leverage an alternative privacy-based attack, membership inference attack (MIA), to reconstruct the unlearned data. MIA involves training a set of attack models to determine whether a given sample was part of a target model's training set (Shokri et al. 2017; Yeom et al. 2018). However, in the privacy-preserving setting of FL, an adversarial client does not have access to the local model of the unlearned client and therefore cannot apply conventional MIAs on the target model. To address this challenge, inspired by (Chen et al. 2021), we propose an innovative unlearning-induced MIA that exploits the discrepancy between the global model before and after unlearning to indirectly infer the membership status of a given data sample.

Given that the unlearning-induced MIA model (attack model) functions only as a discriminative classifier, we design a coarse-to-fine data generation pipeline to reconstruct the unlearned samples. Specifically, we begin by generating coarse candidate samples using the predictions of the attack model. To reduce false positives (i.e., samples incorrectly identified as unlearned), we introduce a novel cross-model filtering mechanism that permutes the input posteriors of the attack models to mitigate the dominance of any single posterior in the final prediction. To further improve reconstruction quality, we apply a sample-level refinement to each coarse candidate, aiming to increase the attack model's confidence in classifying it as unlearned data and to enhance the overall diversity of the generated samples. Finally, AUA is executed by forcing the global model to relearn the reconstructed unlearned data, while DUA is performed by injecting targeted poisoning based on the reconstructed samples.

In conclusion, our work makes following contributions:

- We introduce a novel class of retaliatory attacks, initiated by an FL participant who opposes the unlearning process and seeks to retaliate against the client(s) requesting unlearning by exploiting the data leakage during FU.
- We develop an innovative unlearning-induced MIA model alongside a coarse-to-fine data generation pipeline to reconstruct the unlearned data of the unlearned client(s), which serves as the foundation for executing two targeted retaliatory attacks: AUA and DUA.
- We thoroughly evaluate the proposed retaliatory attacks across a range of existing FU methods and settings, revealing a consistent and realistic privacy vulnerability introduced by the FU process.

Problem Formulation

Federated Learning and Unlearning

We consider a typical FL scenario in which a set of n clients, denoted as $\mathcal{K} = \{k_1, k_2, \ldots, k_n\}$, collaboratively train a global model $\mathcal{M}$ via a central server. Each client k_i ($i \in \{1, 2, \ldots, n\}$) maintains a local dataset $\mathcal{D}_i$, and the entire training dataset is denoted as $\mathcal{D} = \bigcup_{i=1}^{n} \mathcal{D}_i$. The FL training process can then be formalized as a function $FL(\mathcal{D}) \rightarrow \mathcal{M}$, consisting of two key steps: (1) *Local training*, where each client k_i trains a local model $\mathcal{M}_i$ on its dataset $\mathcal{D}_i$ and uploads it to the server; and (2) *Global aggregation*, where the server aggregates all local models (e.g., using aggregation rules such as FedAvg (McMahan et al. 2017)) and distributes the updated global model back to the clients.

Then, during the FL training, a subset of clients $\mathcal{K}^u \subset \mathcal{K}$ may submit an unlearning request to remove the contribution of their local datasets from the trained global model. Upon receiving the request, the server interrupts the standard FL process and initiates the FU process. Let $\mathcal{M}_{\text{before}}$ and $\mathcal{M}_{\text{after}}$ denote the global model before the FU process begins and after it finishes, respectively. The FU process can then be represented as a function $FU(\mathcal{M}_{\text{before}}, \mathcal{D}^u, \mathcal{D}^r, \mathcal{I}) \rightarrow \mathcal{M}_{\text{after}}$, where $\mathcal{D}^u = \bigcup_{k_i \in \mathcal{K}^u} \mathcal{D}_i$ denotes the set of local datasets to be removed (i.e., from the unlearned clients), $\mathcal{D}^r = \mathcal{D} \setminus \mathcal{D}^u$ denotes the remaining datasets belonging to the remaining clients $\mathcal{K}^r = \mathcal{K} \setminus \mathcal{K}^u$, and $\mathcal{I}$ denotes any additional information required to perform unlearning (e.g., historical checkpoints or intermediate states). Finally, when the FU process completes, the FL training resumes for all clients in $\mathcal{K}^r$, continuing from the unlearned model $\mathcal{M}_{\text{after}}$.

Threat and Adversary Model

In this study, we consider a threat model where a participating client in the FL system, although not issuing any unlearning request itself, is opposed to the unlearning mechanism and thus acts as an adversary by retaliating against those clients who request to be unlearned. Given that the adversary's ultimate goal is to retaliate against unlearning clients, it may adopt various specific attack strategies to achieve this objective. Specifically, we propose two such retaliatory attacks, namely: (1) *Anti-Unlearning Attack* (AUA), which aims to prevent the global model from successfully forgetting the private data intended to be unlearned, thereby undermining the privacy guarantees (i.e., RTBF) promised to the unlearned clients; and (2) *Discrimination-Unlearning Attack* (DUA), which aims to intentionally degrade the global model's performance on both the data and underlying distribution associated with the unlearned clients, thereby causing the model to systematically discriminate against their data.

We then consider a strictly privacy-preserving FL and FU environment, where the adversary does not have access to any additional information (e.g., external datasets, private data, or uploaded gradients from other clients), except for what is naturally available to a participating client in the FL or FU process (e.g., its own local data, the FL model architecture, and the global models distributed by the server). In addition, we assume that the adversary cannot interfere with the global aggregation or the FU process, and has no knowledge of the specific FU algorithm adopted by the server. The only assumption we make is that the adversary can observe the global model before ($\mathcal{M}_{\text{before}}$) and after ($\mathcal{M}_{\text{after}}$) the FU process. This assumption is reasonable, as the execution of

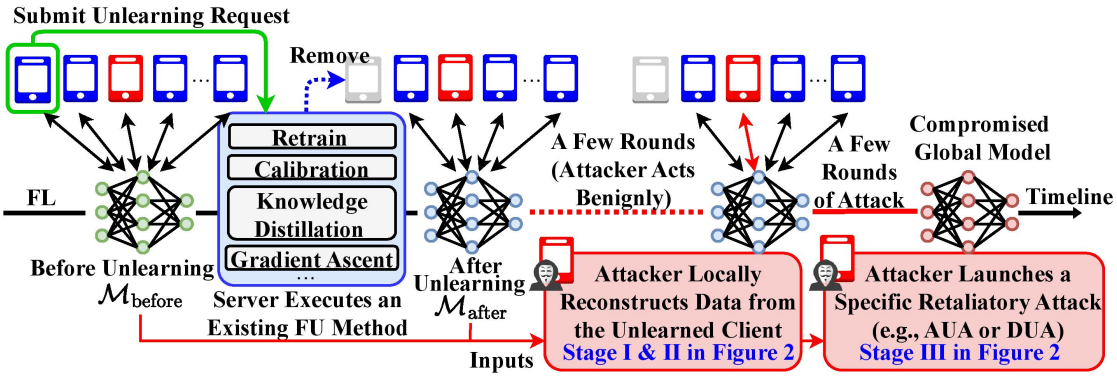


Figure 1: Overview and timeline of the proposed retaliatory attack on federated unlearning.

FU often introduces observable interruptions to the standard FL process, and in many FU implementations, the server must explicitly notify all participating clients that unlearning is being performed. Even in the absence of such explicit signals, we show in our Appendix C (Sheng et al. 2025) that an adversarial client can still reliably detect the occurrence of FU and estimate the post-unlearning global model by monitoring consistency patterns in global model updates. (See Appendix B (Sheng et al. 2025) for a summary of notation.)

Method

Attack Overview

The core of the proposed retaliatory attacks lies in the additional information inadvertently leaked during the unlearning process. In a conventional privacy-preserving FL system, it is infeasible for any single (adversarial) client to accurately identify or reconstruct private data belonging to other participating clients. However, we demonstrate that this becomes feasible when an adversarial client exploits the model discrepancies introduced by the unlearning process.

Figure 1 illustrates the overall workflow and timeline of the proposed retaliatory attack. The adversarial client behaves benignly throughout both the FL and FU phases, as the attack specifically targets the unlearning process and is therefore triggered only after the adversary receives the global model following unlearning. The attack proceeds in three stages. In Stage I, the adversarial client trains a set of unlearning-induced MIA models $\mathcal{A}$, each of which exploits the discrepancies between the global model before ($\mathcal{M}_{\text{before}}$) and after ($\mathcal{M}_{\text{after}}$) the unlearning process. These models enable the adversary to determine whether a given data sample belongs to the unlearned client(s), other participating clients, or to none of them. In Stage II, a coarse-to-fine data generation process is employed to reconstruct the unlearned data $\mathcal{D}_{\text{reconstruct}}$, leveraging the predictions of $\mathcal{A}$. Stages I and II are both carried out locally on the adversarial client's device and may span several rounds of global aggregation. During this period (illustrated by the red dotted line in Figure 1), the adversary continues to behave benignly and does not interfere with the training of the global model. Stage III begins once the reconstruction is complete, during which the adversarial client launches a specific retaliatory attack (e.g., AUA or DUA) against the global model using $\mathcal{D}_{\text{reconstruct}}$, which can rapidly compromise the global model within a few rounds of aggregation (illustrated by the red solid line in Figure 1). The detailed procedures of each attack stage are elaborated in the following sections.

Unlearning-Induced Membership Inference

In Stage I of the proposed retaliatory attack, the adversarial client locally trains a set of MIA models to determine the membership status of a given data sample during the FL and FU processes. Specifically, these models aim to identify whether a sample was part of the training data ($\mathcal{D}^u$) of the unlearned client(s). Traditional MIA pipelines (Shokri et al. 2017) typically require access to the target model trained on the data, which in this context corresponds to the local model of the unlearned client(s). However, such access is infeasible for the adversarial client in a typical FL setting due to its decentralized nature. To address this limitation, inspired by (Chen et al. 2021), we propose leveraging the discrepancies between the global models before and after FU. These differences implicitly capture the influence of the unlearned data, enabling membership inference without requiring direct access to the unlearned clients' local models. The detailed procedure for establishing these attack models is illustrated in Stage I of Figure 2 and described as follows.

Shadowing FU Processes. To conduct the unlearning-induced MIA, the adversarial client first shadows the entire FU process locally, simulating the global unlearning procedure. This requires access to a shadow dataset $\mathcal{D}_{\text{shadow}}$ that resembles the original training data $\mathcal{D}$ used by the global model. Since our work primarily targets tabular datasets, $\mathcal{D}_{\text{shadow}}$ can be constructed by generating high-confidence samples from $\mathcal{M}_{\text{before}}$ using some search-based approaches (e.g., the hill-climbing algorithm proposed in (Shokri et al. 2017)). Then, in each shadow process $s \in \{1, 2, ..., S\}$ (where S denotes the total number of shadowing processes), the shadow dataset $\mathcal{D}_{\text{shadow}}$ is randomly partitioned into three disjoint subsets: $\mathcal{D}^s_{\text{external}}$, representing samples unused in both the FL and FU processes; $\mathcal{D}^s_{\text{unlearned}}$, representing samples used to train $\mathcal{M}_{\text{before}}$ but excluded from training $\mathcal{M}_{\text{after}}$; and $\mathcal{D}^s_{\text{retained}}$, used in training both $\mathcal{M}_{\text{before}}$ and $\mathcal{M}_{\text{after}}$. To better mimic the FU process, $\mathcal{D}^s_{\text{unlearned}}$ and $\mathcal{D}^s_{\text{retained}}$ are fur-

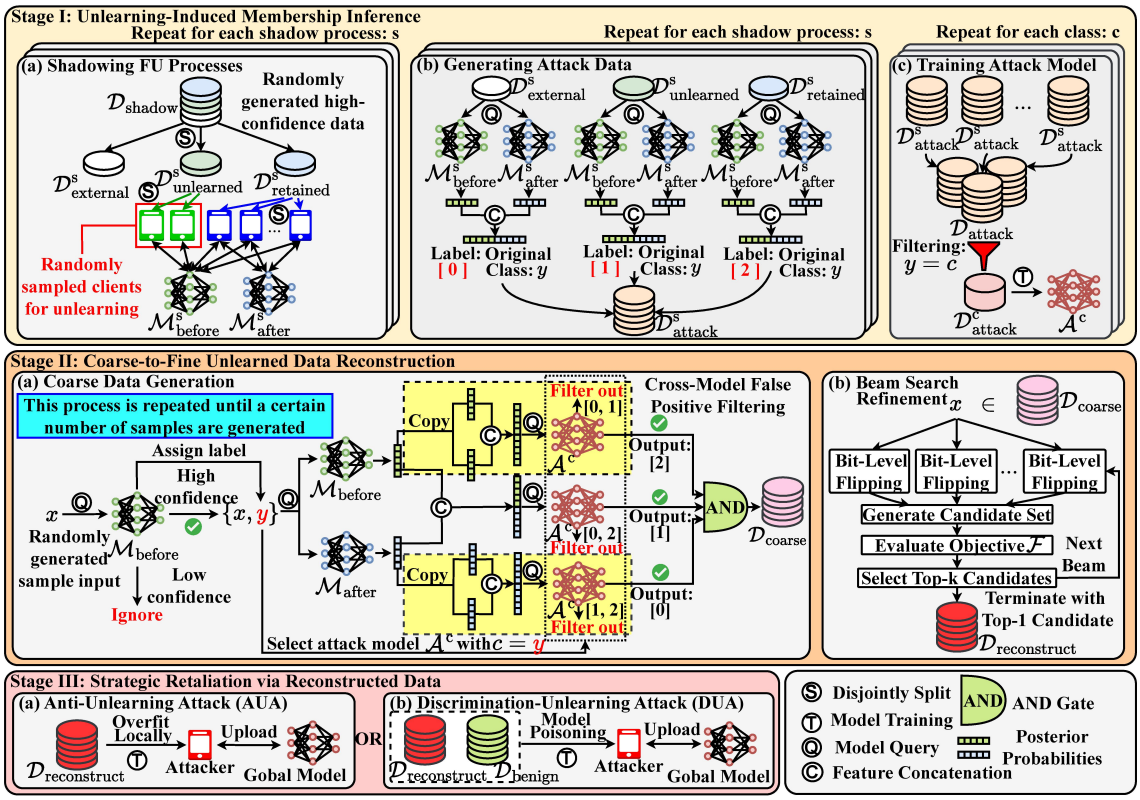


Figure 2: Detailed illustration of the proposed retaliatory attack on federated unlearning.

ther split and assigned to a group of simulated unlearned and retained clients, respectively, where the unlearned clients are randomly sampled and constitute a minority among all simulated clients. Based on this client setup, the adversary can emulate both the FL and FU processes locally, thus obtaining the shadowed global models before and after unlearning, denoted as $\mathcal{M}^s_{\text{before}}$ and $\mathcal{M}^s_{\text{after}}$, respectively.

Generating Attack Data. To prepare the attack data for training the attack models, the adversary queries the trained models $\mathcal{M}^s_{\text{before}}$ and $\mathcal{M}^s_{\text{after}}$ using samples from $\mathcal{D}^s_{\text{external}}$, $\mathcal{D}^s_{\text{unlearned}}$, and $\mathcal{D}^s_{\text{retained}}$, respectively. Let $\mathbb{P}^s_{\text{before}}$ and $\mathbb{P}^s_{\text{after}}$ denote the posterior probabilities obtained by querying $\mathcal{M}^s_{\text{before}}$ and $\mathcal{M}^s_{\text{after}}$, respectively. The input features of the attack data are constructed by concatenating the two posteriors, i.e., $\mathbb{P}^s_{\text{before}}||\mathbb{P}^s_{\text{after}}$. A label of 0, 1, or 2 is then assigned to the attack data generated from $\mathcal{D}^s_{\text{external}}$, $\mathcal{D}^s_{\text{unlearned}}$, and $\mathcal{D}^s_{\text{retained}}$, respectively, indicating their exact membership status during the shadowed FU process. These form the set of attack data $\mathcal{D}^s_{\text{attack}}$ generated from the shadowing process s. (Note that the original class label y of each queried data sample is also recorded for subsequent partitioning.)

Training Attack Models. After completing all S shadowing processes, the overall attack dataset is assembled as $\mathcal{D}_{\text{attack}} = \bigcup_{s=1}^{S} \mathcal{D}^s_{\text{attack}}$. For each class $c \in \{1, 2, \ldots, C\}$ (where C denotes the total number of classes), a separate attack model $\mathcal{A}^c$ is trained using a class-specific subset $\mathcal{D}^c_{\text{attack}} \subset \mathcal{D}_{\text{attack}}$, which contains all samples with class label $y = c$. These models form the final set of attack models $\mathcal{A} = \{\mathcal{A}^c\}_{c=1}^{C}$ (the unlearning-induced MIA models).

Coarse-to-Fine Unlearned Data Reconstruction

While the set of attack models $\mathcal{A}$ obtained in Stage I can effectively determine the membership status of a given data sample, it serves only as a discriminative classifier. A generative method is still necessary to enable the reconstruction of the unlearned data. Therefore, in Stage II of the proposed attack framework, we introduce a novel coarse-to-fine data generation method to recover the unlearned data from the FU process, as illustrated in Stage II of Figure 2.

Coarse Data Generation. The coarse data generation begins by randomly initializing the input feature vector x, with each attribute uniformly sampled within its domain. The resulting x is used to query the global model before unlearning ($\mathcal{M}_{\text{before}}$) to obtain its posterior probabilities $\mathbb{P}_{\text{before}}$. The confidence score is then calculated as the maximum value in $\mathbb{P}_{\text{before}}$, i.e., $\max(\mathbb{P}_{\text{before}})$, and the predicted label is assigned as $y = \arg\max(\mathbb{P}_{\text{before}})$. A generated sample is retained if its confidence score exceeds a predefined threshold; otherwise, it is discarded. This filtering step ensures that each retained sample has a reliable class label, which is essential for selecting the corresponding attack model $\mathcal{A}^c$ with $c = y$.

While these randomly generated high-confidence samples roughly capture the overall training data distribution of $\mathcal{M}_{\text{before}}$, they may not align with the data distribution of the unlearned client(s). To further narrow the candidate set toward the unlearned data, each sample x is queried on

both $\mathcal{M}_{\text{before}}$ and $\mathcal{M}_{\text{after}}$ to obtain the corresponding posterior probability vectors $\mathbb{P}_{\text{before}}$ and $\mathbb{P}_{\text{after}}$, which are then concatenated and fed into the corresponding attack model $\mathcal{A}^c$. The sample is retained only if $\mathcal{A}^c$ classifies it as class 1, indicating that it is inferred to belong to the unlearned data; otherwise, it is discarded.

Cross-Model False Positive Filtering. Although the set of attack models $\mathcal{A}$ provides valuable guidance for recovering the unlearned data, its predictions are still significantly affected by false positives (i.e., samples incorrectly classified as unlearned). Inspired by (Carlini et al. 2021), which proposes filtering out false positives by comparing predictions with those from a second model trained on a disjoint dataset when extracting training data from a large language model, we propose an alternative yet novel cross-model false positive filtering strategy that fully leverages the same attack model via permutation of input posteriors, as highlighted by the yellow-shaded regions in Stage II(a) of Figure 2. Specifically, in addition to the original input—formed by concatenating the posteriors $\mathbb{P}_{\text{before}}$ and $\mathbb{P}_{\text{after}}$ and fed into the attack model $\mathcal{A}^c$ with class 1 as the desired output—we generate two auxiliary inputs: one by duplicating $\mathbb{P}_{\text{before}}$ (i.e., $\mathbb{P}_{\text{before}}||\mathbb{P}_{\text{before}}$) and the other by duplicating $\mathbb{P}_{\text{after}}$ (i.e., $\mathbb{P}_{\text{after}}||\mathbb{P}_{\text{after}}$). These are also fed into $\mathcal{A}^c$, with the expected outputs being class 2 and class 0, respectively.

The key insight here is that, given our attack model $\mathcal{A}^c$ actually considers two models ($\mathcal{M}_{\text{before}}$ and $\mathcal{M}_{\text{after}}$) as target models, each of the two posteriors fed into $\mathcal{A}^c$ encodes distinct training-related information for an unlearned data sample: the first indicates that the sample was seen (i.e., included in the training data) by the first target model ($\mathcal{M}_{\text{before}}$), while the second reflects that the sample was absent from the training data of the second target model ($\mathcal{M}_{\text{after}}$). In this case, when the input is $\mathbb{P}_{\text{before}}||\mathbb{P}_{\text{before}}$, it represents a sample that is seen by both target models, and thus should be assigned a membership status of 2. Conversely, when the input is $\mathbb{P}_{\text{after}}||\mathbb{P}_{\text{after}}$, it indicates that the sample is absent from the training sets of both models and should therefore correspond to membership status 0. This strategy effectively filters out false positives in the initial screening stage, where the prediction of $\mathcal{A}^c$ may be dominated by either $\mathbb{P}_{\text{before}}$ or $\mathbb{P}_{\text{after}}$, leading to the incorrect classification of non-unlearned samples as class 1. Finally, the sample is retained and added to $\mathcal{D}_{\text{coarse}}$ only when all three conditions are satisfied.

Beam Search Refinement. While $\mathcal{D}_{\text{coarse}}$ provides an initial set of candidate samples that approximate the distribution of the unlearned data, we further introduce a sample-level refinement procedure to enhance the fidelity of these samples with respect to the original unlearned data. The key idea is to increase the confidence of the attack model $\mathcal{A}^c$ in classifying a sample as class 1, i.e., to maximize $\mathbb{P}_{\text{attack}}[1]$, where $\mathbb{P}_{\text{attack}}$ denotes the posterior obtained from querying $\mathcal{A}^c$. In addition, we incorporate a diversity penalty to discourage the refined samples from collapsing into similar patterns, thereby promoting sample-level variability. Furthermore, the confidence score $\mathbb{P}_{\text{before}}[c]$ of $\mathcal{M}_{\text{before}}$ is encouraged to remain high, as it directly determines the selection of the corresponding attack model $\mathcal{A}^c$. Formally, our objective function $\mathcal{F}$ is defined as:

$$\mathcal{F}(x) = \alpha \cdot (\mathbb{P}_{\text{before}}[c]) + \beta \cdot (\mathbb{P}_{\text{attack}}[1]) - \gamma \cdot \big(1 - \min_{x' \in \tilde{\mathcal{D}}_{\text{reconstruct}}} \text{dist}(x, x')\big), \quad (1)$$

where α, β, and γ are weighting coefficients; $\text{dist}(\cdot,\cdot)$ denotes the distance function between two data samples (we adopt the Jaccard distance in our implementation); and $\tilde{\mathcal{D}}_{\text{reconstruct}}$ represents the set of samples that have already been refined. Given the black-box and non-differentiable nature of $\mathcal{F}(x)$, we employ a heuristic beam search algorithm that iteratively refines candidate samples via random bit-level flipping in the input space, as illustrated in Stage II(b) of Figure 2. Finally, after refining all samples in $\mathcal{D}_{\text{coarse}}$, we obtain $\mathcal{D}_{\text{reconstruct}}$ as the final reconstructed set of unlearned data, thus concluding Stage II of the attack.

Strategic Retaliation via Reconstructed Data

With $\mathcal{D}_{\text{reconstruct}}$ obtained, the adversarial client can now proceed to launch a targeted retaliatory attack, either AUA or DUA, against the global model.

Anti-Unlearning Attack. Recall that the objective of AUA is to prevent the global model from successfully forgetting the training data of the unlearned client(s). As illustrated in Stage III(a) of Figure 2, this can be achieved by having the adversarial client repeatedly upload poisoned local model updates that are deliberately overfitted to $\mathcal{D}_{\text{reconstruct}}$ (e.g., by using an increased number of local epochs and a reduced learning rate). This strategy effectively forces the global model to relearn the reconstructed unlearned data within just a few communication rounds.

Discrimination-Unlearning Attack. DUA aims to introduce discrimination against the unlearning process by intentionally degrading the performance of the global model on the distribution corresponding to the unlearned client(s). As illustrated in Stage III(b) of Figure 2, the adversarial client achieves this by conducting a model poisoning attack that leverages both $\mathcal{D}_{\text{reconstruct}}$ and its own local benign dataset $\mathcal{D}_{\text{benign}}$. Specifically, during local training, the adversarial client optimizes the following objective:

$$\mathcal{L} = \mathcal{L}_{\text{CE}}(\mathcal{D}_{\text{benign}}) - \lambda \cdot \mathcal{L}_{\text{CE}}(\mathcal{D}_{\text{reconstruct}}), \quad (2)$$

where $\mathcal{L}_{\text{CE}}$ denotes the standard cross-entropy loss, and λ controls the strength of the negative gradient signal derived from the unlearned data. By inverting the loss gradient on $\mathcal{D}_{\text{reconstruct}}$, the adversary forces the model to perform poorly on the reconstructed unlearned samples, while maintaining nominal performance on its own benign data. (See Appendix D (Sheng et al. 2025) for the full Stage I–III procedure of the proposed retaliatory attack.)

Experiments

Experimental Setup

Datasets. To evaluate the proposed retaliatory attacks, we conduct experiments on two tabular datasets that are widely used in the data privacy literature: Location (Yang et al. 2015) and Purchase (Shokri et al. 2017). For both datasets, we follow the same preprocessing procedure as in (Shokri

Dataset	Metric	Federated Unlearning Method									
		Retrain		FedEraser		KD-based FU		SGA-based FU		RobustFU	
		Before	After Attack	Before	After Attack	Before	After Attack	Before	After Attack	Before	After Attack
Location	MIA	0.502	0.629 (↑0.127)	0.505	0.648 (↑0.143)	0.509	0.725 (↑0.216)	0.514	0.713 (↑0.199)	0.505	0.601 (↑0.096)
	UA	0.606	0.950 (↑0.344)	0.585	0.981 (↑0.396)	0.568	0.991 (↑0.423)	0.610	0.984 (↑0.374)	0.608	0.977 (↑0.369)
Purchase100	MIA	0.499	0.597 (↑0.098)	0.502	0.609 (↑0.107)	0.504	0.627 (↑0.123)	0.511	0.590 (↑0.079)	0.501	0.573 (↑0.072)
	UA	0.578	0.975 (↑0.397)	0.621	0.997 (↑0.376)	0.566	0.998 (↑0.432)	0.628	0.966 (↑0.338)	0.616	0.980 (↑0.364)

Table 1: Attack performance of the proposed AUA across different federated unlearning methods.

Dataset	Metric	Federated Unlearning Method									
		Retrain		FedEraser		KD-based FU		SGA-based FU		RobustFU	
		Before	After Attack	Before	After Attack	Before	After Attack	Before	After Attack	Before	After Attack
Location	UA	0.606	0.544 (↓0.062)	0.585	0.530 (↓0.055)	0.568	0.523 (↓0.045)	0.610	0.547 (↓0.063)	0.608	0.551 (↓0.057)
	TA	0.611	0.598 (↓0.013)	0.601	0.584 (↓0.017)	0.597	0.582 (↓0.015)	0.615	0.597 (↓0.018)	0.621	0.603 (↓0.018)
Purchase100	UA	0.578	0.497 (↓0.081)	0.621	0.494 (↓0.127)	0.566	0.513 (↓0.053)	0.628	0.526 (↓0.102)	0.616	0.542 (↓0.074)
	TA	0.637	0.619 (↓0.018)	0.635	0.610 (↓0.025)	0.629	0.606 (↓0.023)	0.630	0.617 (↓0.013)	0.639	0.605 (↓0.034)

Table 2: Attack performance of the proposed DUA across different federated unlearning methods.

et al. 2017). The resulting Location dataset consists of 30 geosocial classes, with each sample represented by 446 binary features indicating whether a user has visited specific regions or location types. We use the Purchase100 version of the Purchase dataset, which includes 100 purchase styles, with each sample represented by 600 binary features indicating whether a user purchased a specific product.

While the proposed retaliatory attacks are primarily tailored for tabular classification tasks, we further demonstrate that the privacy leakage exploited in Stage I of our framework is also applicable to image classification tasks. To validate this, we evaluate such unlearning-induced data leakage on two additional image datasets: CIFAR-10 (Krizhevsky, Hinton et al. 2009) and SVHN (Netzer et al. 2011).

Evaluation Metrics. To evaluate the effectiveness of AUA, we adopt the standard **MIA accuracy (MIA)** (Shokri et al. 2017), which is widely used as a certified test for assessing whether the unlearned data has been successfully forgotten (Wang et al. 2022; Halimi et al. 2022; Sheng, Bao, and Ge 2024). In addition, we measure the **Unlearned Accuracy (UA)**, the prediction accuracy of the global model on the unlearned dataset, where a higher UA indicates that the unlearned data has not been effectively forgotten. For DUA, we similarly report UA to quantify the performance degradation on the unlearned dataset. To ensure a fair evaluation, we also include the **Test Accuracy (TA)** on a held-out test set, which reflects the global model's performance on normal, non-unlearned data. For each of these metrics, we report both the values before (i.e., immediately after the FU process) and after the attack, thereby highlighting the extent to which AUA and DUA compromise the global model. Furthermore, since Stage I of our proposed retaliatory framework also involves a novel **unlearning-induced MIA (U-MIA)**, which infers membership status based on discrepancies between the global model before and after unlearning, we report its performance as a direct measure of privacy leakage caused by the FU process. For both the standard MIA and U-MIA, higher attack accuracy indicates greater privacy leakage and weaker unlearning guarantees.

FU Methods. We consider several state-of-the-art FU methods that the server may adopt, including:

- **FedEraser** (Liu et al. 2021), which leverages stored historical parameter updates from participating clients and incorporates a calibration mechanism during retraining to efficiently reconstruct the unlearned model.
- **SGA-based FU** (Wu et al. 2022), which integrates elastic weight consolidation (EWC) with reverse stochastic gradient ascent (SGA) to enable effective FU.
- **KD-based FU** (Wu, Zhu, and Mitra 2023), which achieves FU by subtracting accumulated historical updates from the trained global model and restoring its performance through knowledge distillation (KD).
- **RobustFU** (Sheng, Bao, and Ge 2024), which performs robust FU by reintroducing high-information-gain samples into the remaining clients during retraining.

We also evaluate the retraining-from-scratch golden baseline (**Retrain**), where the unlearned model is retrained on the remaining clients using FedAvg (McMahan et al. 2017).

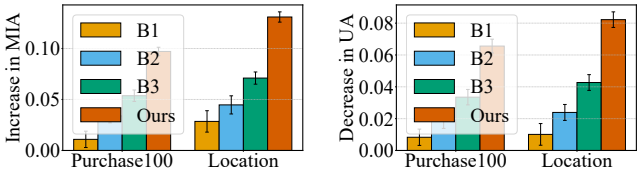
Implementation Details. Please refer to our Appendix E (Sheng et al. 2025) for implementation and training details.

Experimental Results

Overall Attack Performance. Table 1 and Table 2 present the attack performance of the proposed AUA and DUA, respectively. For AUA, the results show that the MIA accuracy is close to 0.5 across all evaluated FU methods prior to the attack, suggesting that the unlearned global model (i.e., before being attacked) has effectively forgotten the data requested for unlearning. However, after launching the AUA, the MIA accuracy on the unlearned data increases substantially, indicating that the compromised global model has relearned information that was intended to be removed. Similarly, a significant increase in the prediction accuracy on

Dataset	Federated Unlearning Method (U-MIA)			
	Retrain	FedEraser	KD-based FU	SGA-based FU
Location	0.709	0.688	0.657	0.672
Purchase100	0.804	0.786	0.763	0.781
CIFAR-10	0.654	0.630	0.625	0.634
SVHN	0.693	0.679	0.640	0.661

Table 3: Evaluation of privacy leakage in FU methods.



(a) AUA (b) DUA

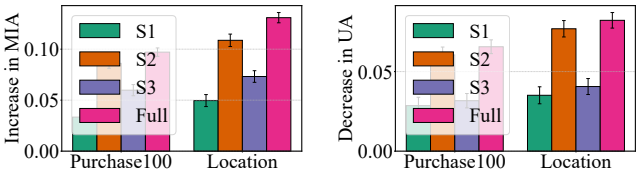
Figure 3: Comparison between our method and baselines.

the unlearned data is observed after the attack, further confirming that the global model fails to preserve the forgetting effect and continues to memorize the unlearned samples.

For DUA, a substantial drop in prediction accuracy on the unlearned data is observed after the attack, indicating that the compromised global model has successfully introduced additional discrimination against the unlearned data. While a slight decrease is also noted in the prediction accuracy on the test data, this drop is relatively minor compared to that on the unlearned data, suggesting that the attack remains primarily targeted at the distribution of the unlearned data.

Privacy Leakage Evaluation. Since the effectiveness of the proposed retaliatory attacks hinges on the privacy leakage exposed during the FU process, the set of unlearning-induced MIA models ($\mathcal{A}$) proposed in Stage I of our attack provides a direct and quantifiable measure of this leakage, where a higher U-MIA accuracy suggests a greater degree of privacy leakage. We evaluate this leakage across two tabular and two image datasets under various FU methods (For image classification tasks, we assume the adversarial client maintains a local shadow set to emulate the FU process.). As shown in Table 3, the proposed unlearning-induced MIA models consistently achieve high attack accuracy (U-MIA) across all settings, thereby highlighting the vulnerability of existing FU mechanisms to privacy leakage.

Baseline Comparison. To further demonstrate the effectiveness of the proposed attack, we compare it with several baseline methods. We first consider a naive membership inference baseline (denoted as **B1**), which randomly generates data samples and retains those on which the global model before unlearning yields high confidence. We then propose an improved baseline (**B2**), which additionally incorporates the global model after unlearning. This baseline is based on the intuition that the confidence of a model should decrease on samples that have been removed through unlearning. Specifically, B2 generates candidate samples and retains those that receive high confidence from the pre-unlearning model but significantly lower confidence from the post-unlearning model. Moreover, we adopt a more advanced



(a) AUA (b) DUA

Figure 4: Ablation studies of key design components.

MIA-based baseline method (**B3**), which trains two separate MIA models (Shokri et al. 2017) for the global models before and after unlearning, respectively. A randomly generated candidate sample is considered as an unlearned training sample if it is classified as a member by the MIA model corresponding to the pre-unlearning global model, but classified as a non-member by the MIA model corresponding to the post-unlearning global model.

Figure 3 demonstrates the performance gain (i.e., the increase in MIA accuracy for AUA and the decrease in UA for DUA) for each of the baseline methods and our proposed method on both the Location and Purchase100 datasets (using Retrain as the FU method). It can be seen that our method achieves better performance than all baselines, demonstrating its superior ability to reconstruct the unlearned data and conduct more effective retaliatory attacks.

Ablation Studies. We further conduct ablation studies to evaluate the contribution of each key component in our coarse-to-fine data generation pipeline. Specifically, we examine several settings: (**S1**) using only a single prediction from the attack model $\mathcal{A}^c$ on class 1, without false positive filtering or beam search refinement; (**S2**) using a single prediction without false positive filtering but incorporating the beam search refinement; (**S3**) applying only the coarse data generation stage (i.e., $\mathcal{D}_{\text{coarse}}$), without beam search refinement; and the full version of our method with all components enabled. The results, shown in Figure 4, demonstrate how each design component influences the overall attack performance through different configurations.

Extended Analysis. For additional experiments, please refer to our Appendix F (Sheng et al. 2025).

Conclusion

In this paper, we introduce a new concept of retaliatory attacks against FU, which refer to potential attacks launched by malicious users who oppose the unlearning mechanism in FL. We propose two such attacks, AUA and DUA, which aim to either undermine the forgetting effect or induce targeted discrimination against the unlearned user. These attacks leverage privacy leakage during the FU process by first conducting an unlearning-induced MIA, followed by a coarse-to-fine data generation method to reconstruct the unlearned data. We evaluate the effectiveness and robustness of the proposed attacks under various FU methods and settings.

Acknowledgments

This work is supported by Australian Research Council/Linkage Project LP230100294.

References

Bourtoule, L.; Chandrasekaran, V.; Choquette-Choo, C. A.; Jia, H.; Travers, A.; Zhang, B.; Lie, D.; and Papernot, N. 2021. Machine unlearning. In *2021 IEEE symposium on security and privacy (SP)*, 141–159. IEEE.

Cao, Y.; and Yang, J. 2015. Towards making systems forget with machine unlearning. In *2015 IEEE symposium on security and privacy*, 463–480. IEEE.

Carlini, N.; Tramer, F.; Wallace, E.; Jagielski, M.; Herbert-Voss, A.; Lee, K.; Roberts, A.; Brown, T.; Song, D.; Erlingsson, U.; et al. 2021. Extracting training data from large language models. In *30th USENIX security symposium (USENIX Security 21)*, 2633–2650.

Che, T.; Zhou, Y.; Zhang, Z.; Lyu, L.; Liu, J.; Yan, D.; Dou, D.; and Huan, J. 2023. Fast federated machine unlearning with nonlinear functional theory. In *International conference on machine learning*, 4241–4268. PMLR.

Chen, M.; Zhang, Z.; Wang, T.; Backes, M.; Humbert, M.; and Zhang, Y. 2021. When machine unlearning jeopardizes privacy. In *Proceedings of the 2021 ACM SIGSAC conference on computer and communications security*, 896–911.

Gao, X.; Ma, X.; Wang, J.; Sun, Y.; Li, B.; Ji, S.; Cheng, P.; and Chen, J. 2024. Verifi: Towards verifiable federated unlearning. *IEEE Transactions on Dependable and Secure Computing*.

Halimi, A.; Kadhe, S. R.; Rawat, A.; and Angel, N. B. 2022. Federated Unlearning: How to Efficiently Erase a Client in FL? In *International Conference on Machine Learning*.

Harding, E.; Vanto, J. J.; Clark, R.; Ji, L. H.; and Ainsworth, S. C. 2019. Understanding the scope and impact of the California Consumer Privacy Act of 2018. *Journal of Data Protection & Privacy*.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images.

Lamri, H.; Alam, M.; Jiang, H.; and Maniatakos, M. 2025. DRAUN: An Algorithm-Agnostic Data Reconstruction Attack on Federated Unlearning Systems. arXiv:2506.01777.

Liu, G.; Ma, X.; Yang, Y.; Wang, C.; and Liu, J. 2021. Federaser: Enabling efficient client-level data removal from federated learning models. In *2021 IEEE/ACM 29th International Symposium on Quality of Service (IWQOS)*, 1–10. IEEE.

Liu, Y.; Xu, L.; Yuan, X.; Wang, C.; and Li, B. 2022. The right to be forgotten in federated learning: An efficient realization with rapid retraining. In *IEEE INFOCOM 2022-IEEE conference on computer communications*, 1749–1758. IEEE.

Liu, Z.; Jiang, Y.; Shen, J.; Peng, M.; Lam, K.-Y.; Yuan, X.; and Liu, X. 2023. A survey on federated unlearning: Challenges, methods, and future directions. *ACM Computing Surveys*.

McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; and Arcas, B. A. y. 2017. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*.

Netzer, Y.; Wang, T.; Coates, A.; Bissacco, A.; Wu, B.; Ng, A. Y.; et al. 2011. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, volume 2011, 7. Granada.

Sheng, X.; Bao, W.; and Ge, L. 2024. Robust federated unlearning. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, 2034–2044.

Sheng, X.; Bao, W.; Wang, H.; Liu, Y.; and Fu, S. 2025. GitHub: Retaliatory Attacks Against Federated Unlearning via Data Leakage. https://github.com/stcebra/Retaliatory-Attacks-Against-Federated-Unlearning. Appendix & Code.

Shokri, R.; Stronati, M.; Song, C.; and Shmatikov, V. 2017. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, 3–18. IEEE.

Voigt, P.; and von dem Bussche, A. 2017. *The EU General Data Protection Regulation (GDPR): A Practical Guide*. Cham, Switzerland: Springer.

Wang, F.; Li, B.; and Li, B. 2023. Federated unlearning and its privacy threats. *IEEE Network*, 38(2): 294–300.

Wang, J.; Guo, S.; Xie, X.; and Qi, H. 2022. Federated unlearning via class-discriminative pruning. In *Proceedings of the ACM web conference 2022*, 622–632.

Wang, W.; Ma, Q.; Zhang, Z.; Liu, Y.; Liu, Z.; and Fang, M. 2025. Poisoning Attacks and Defenses to Federated Unlearning. In *Companion Proceedings of the ACM on Web Conference 2025*, 1365–1369.

Wang, W.; Tian, Z.; Zhang, C.; and Yu, S. 2024. Machine unlearning: A comprehensive survey. *arXiv preprint arXiv:2405.07406*.

Wu, C.; Zhu, S.; and Mitra, P. 2023. Unlearning Backdoor Attacks in Federated Learning. In *ICLR 2023 Workshop on Backdoor Attacks and Defenses in Machine Learning*.

Wu, L.; Guo, S.; Wang, J.; Hong, Z.; Zhang, J.; and Ding, Y. 2022. Federated unlearning: Guarantee the right of clients to forget. *IEEE Network*, 36(5): 129–135.

Xiong, Z.; Li, W.; Li, Y.; and Cai, Z. 2023. Exact-fun: an exact and efficient federated unlearning approach. In *2023 IEEE International Conference on Data Mining (ICDM)*, 1439–1444. IEEE.

Yang, D.; Zhang, D.; Zheng, V. W.; and Yu, Z. 2015. Modeling User Activity Preference by Leveraging User Spatial Temporal Characteristics in LBSNs. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 45(1): 129–142.

Yeom, S.; Giacomelli, I.; Fredrikson, M.; and Jha, S. 2018. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st computer security foundations symposium (CSF)*, 268–282. IEEE.

Zhang, F.; Li, W.; Hao, Y.; Yan, X.; Cao, Y.; and Lim, W. Y. B. 2025. Verifiably Forgotten? Gradient Differences Still Enable Data Reconstruction in Federated Unlearning. arXiv:2505.11097.

Zhang, L.; Zhu, T.; Zhang, H.; Xiong, P.; and Zhou, W. 2023a. Fedrecovery: Differentially private machine unlearning for federated learning frameworks. *IEEE Transactions on Information Forensics and Security*, 18: 4732–4746.

Zhang, R.; Guo, S.; Wang, J.; Xie, X.; and Tao, D. 2023b. A Survey on Gradient Inversion: Attacks, Defenses and Future Directions. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, 5678–685.

Zhu, L.; Liu, Z.; and Han, S. 2019. Deep leakage from gradients. *Advances in neural information processing systems*, 32.