Verifiably Forgotten? Gradient Differences Still Enable Data Reconstruction in Federated Unlearning

Fuyao Zhang[1, 2] **Wenjie Li**[1, 2, ✉] **Yurong Hao**[1, 3] **Xinyu Yan**[1]
Yang Cao[4] **Wei Yang Bryan Lim**[1]
[1] Nanyang Technological University [2] Xidian University
[3] Beijing Jiaotong University [4] Institute of Science Tokyo
{hi.fyzhang, tom643190686}@gmail.com yurong.hao@bjtu.edu.cn
xinyu020@e.ntu.edu.sg cao@c.titech.ac.jp bryan.limwy@ntu.edu.sg

Abstract

Federated Unlearning (FU) has emerged as a critical compliance mechanism for data privacy regulations, requiring unlearned clients to provide verifiable Proof of Federated Unlearning (PoFU) to auditors upon data removal requests. However, we uncover a significant privacy vulnerability: when gradient differences are used as PoFU, *honest-but-curious* auditors may exploit mathematical correlations between gradient differences and forgotten samples to reconstruct the latter. Such reconstruction, if feasible, would face three key challenges: (i) restricted auditor access to client-side data, (ii) limited samples derivable from individual PoFU, and (iii) high-dimensional redundancy in gradient differences. To overcome these challenges, we propose **I**nverting **G**radient difference to **F**orgotten data (IGF), a novel learning-based reconstruction attack framework that employs Singular Value Decomposition (SVD) for dimensionality reduction and feature extraction. IGF incorporates a tailored pixel-level inversion model optimized via a composite loss that captures both structural and semantic cues. This enables efficient and high-fidelity reconstruction of large-scale samples, surpassing existing methods. To counter this novel attack, we design an orthogonal obfuscation defense that preserves PoFU verification utility while preventing sensitive forgotten data reconstruction. Experiments across multiple datasets validate the effectiveness of the attack and the robustness of the defense. The code is available at https://anonymous.4open.science/r/IGF.

1 Introduction

The widespread adoption of Federated Learning (FL) enables distributed entities, such as financial institutions, healthcare providers, and IoT networks, to collaboratively train models without sharing raw data. This decentralized approach mitigates risks associated with data transfer, enhancing privacy and security for data owners. However, compliance with regulations like the *right to be forgotten* under the General Data Protection Regulation (GDPR) [1, 2] requires FL systems to remove specific data contributions from the global model and demonstrate that the model no longer depends on those data samples. Simply preventing raw data leaks is no longer sufficient to meet compliance requirements. This challenge has spurred the development of verifiable Federated Unlearning (FU) [3], a paradigm designed to verifiably forget the contribution of designated data from trained models.

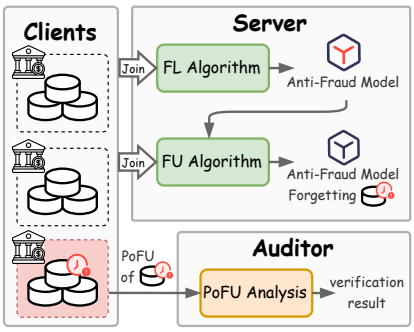


Figure 1: Audition in verifiable FU

Preprint. Under review.

Figure 1 illustrates a typical scenario where multinational financial institutions, acting as FL clients 🏦, collaboratively train an anti-fraud model [4]. Subsequently, the auditor mandates all clients to forget the outdated transaction data using the FU algorithm and receives the proof of FU (PoFU) [5, 6, 7, 8] from the unlearned client. Since the auditor cannot directly access the raw client data, they rely on PoFU, typically based on gradient differences between the original and unlearned models, as a non-invasive auditing tool. However, most research [9, 10, 11] primarily focuses on FU algorithm design, overlooking vulnerabilities to reconstruction attacks by third-party auditors [12, 13], particularly when gradient differences serve as PoFU.

Recent advances in reconstruction attacks have exposed critical vulnerabilities in centralized machine (un)learning. For instance, DLG [14], demonstrated that shared gradients can be exploited to reconstruct original training data, while subsequent work [15] highlighted privacy leakage risks from gradient sharing. More recently, unlearning inversion attacks [16] reconstruct forgotten data by only accessing the parameter deviations of the original and unlearned models. However, these approaches face three primary limitations when applying to federated unlearning scenarios: (i) they require *white-box access* to calculate parameter deviations, (ii) they struggle with *large-scale* data reconstruction due to the limited information encoded in these deviations, and (iii) the *high dimensionality* of model parameters or gradients increases the computational cost of inversion models. More crucially, FU introduces additional complexities for reconstruction attack, the auditor lacks access to client-side raw data [17] and relies solely on PoFU to evaluate unlearning efficacy. Current reconstruction attacks target model parameters or gradients, posing significant threats to machine unlearning in adversarial settings. However, reconstruction attacks exploiting gradient differences as PoFU remain underexplored. This gap motivates our research question:

Q: Can gradient differences, used as PoFU, enable third-party auditors to reconstruct forgotten data? If so, how can high-fidelity, large-scale data reconstruction be achieved against high-dimensional gradient differences?

To address this, we propose a learning-based reconstruction attack in verifiable FU, named **I**nverting **G**radient difference to **F**orgotten data (IGF). To handle high-dimensional gradient differences, we employ Singular Value Decomposition (SVD) for dimensionality reduction, extracting essential features while eliminating redundancy, thus reducing the input dimensionality of the inversion model. We then design a pixel-level convolutional inversion model that learns the latent mapping between gradient differences and original data samples, optimized via a composite loss function incorporating structural and perceptual factors. This model efficiently reconstructs batches of forgotten samples from individual PoFU, bypassing per-sample optimization inefficiencies. These components synergize to enable robust, large-scale sample reconstruction across diverse datasets and global model architectures. Our main contributions are:

- We identify gradient differences used as PoFU as a novel attack surface for reconstruction attacks. By formalizing an *honest-but-curious* third-party auditor, we demonstrate that passive observers can reconstruct forgotten samples [18].
- We develop IGF attack framework, integrating SVD with a composite loss-optimized and pixel-level inversion network, achieving high reconstruction fidelity and computational efficiency.
- Our experiments demonstrate that IGF outperforms the state-of-the-art GIAMU [16] method, reducing reconstruction MSE by over 88% and improving LPIPS by approximately 33% on CIFAR-10, highlighting its effectiveness in federated unlearning.
- We further propose an orthogonal obfuscation defense mechanism to mitigate IGF, and validate both attack and defense efficacy through extensive experiments on public benchmark datasets.

2 Related Work

Federated Unlearning. Federated unlearning has recently emerged to address the challenge of selectively removing specific clients or data points from a trained FL model. This problem is motivated by regulatory requirements (e.g., the *right to be forgotten* under GDPR) and the dynamic nature of real-world FL systems. Existing approaches can be categorized into two main types: *Exact Federated Unlearning (EFU)* [19] and *Approximate Federated Unlearning (AFU)* [20]. EFU achieves

complete removal by retraining the model from scratch using the remaining data, ensuring that the influence of the target data is entirely eliminated. However, this method is computationally intensive and may not be practical for large-scale FL systems. AFU methods aim to reduce computational overhead by approximating the unlearning process through applying gradient ascent to maximize the loss. Among approximate methods, Wang et al. [21] proposes that clients estimate the gradient influence of the data to be removed using local remaining data and then apply gradient ascent to negate this influence. A subsequent fine-tuning step is introduced to preserve overall utility. Similarly, Xu et al. [22] employ model explanations to identify key parameter channels associated with the forgotten categories, and update only those channels in reverse. Meanwhile, Gu et al. [23] pre-generates linear transformation parameters related to the target data during the training phase and applies reverse transformations to eliminate unwanted effects. The above methods balance effectiveness and efficiency. Some works [24, 25] explore how to diminish the model's utility by poisoning or cause excessive forgetting through malicious requests, but overlook potential reconstruction vulnerabilities during the verification stage.

Gradient Inversion Attack. Recent studies have leveraged gradient inversion techniques to reconstruct clients' private training data in FL [26, 27, 28, 29, 30]. Zhang et al. [26] demonstrate the feasibility of generative gradient inversion in FL by constructing an over-parameterized convolutional neural network that satisfies gradient-matching requirements. Similarly, Jeon et al. [27] leverages pre-trained generative models as priors to circumvent direct optimization in high-dimensional pixel space and reconstructs data via latent-space parameter optimization. Additionally, Fang et al. [28] adopts a staged optimization strategy for the intermediate feature domains of generative models, progressively optimizing from the latent space to intermediate layers to enhance attack effectiveness. Sun et al. [29] introduces an anomaly detection model to capture latent distributions from limited data, using it as a regularization term to enhance attack performance. In the context of FU, Hu et al. [16] reveal the feature and label information by analyzing differences between the original and unlearned models.

Therefore, traditional gradient inversion attacks focus on reconstructing training data directly from original gradients provided by clients in standard federated learning scenarios. In contrast, our work targets **gradient differences** used as PoFU, where the attacker must reconstruct deleted data from indirect and variant gradient information. This introduces unique challenges, gradient differences contain limited and mixed signals with weaker correlations to the forgotten samples, requiring fundamentally different inversion approaches.

3 Methodology

3.1 Problem Formulation

Federated Learning (FL). In the FL framework with H clients, each client i ($i \in [H]$) holds a local dataset $\mathcal{D}_i$ containing $|\mathcal{D}_i|$ samples. Let $\mathbf{M}$ denote the original global model parameterized by $\boldsymbol{\theta}$, and consider a supervised learning objective that minimizes the empirical loss over the federated dataset $\mathcal{D} = \bigcup_{i=1}^{H} \mathcal{D}_i$:$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{|\mathcal{D}|} \sum_{(x,y)\in\mathcal{D}} \ell\big(\mathbf{M}(x;\boldsymbol{\theta}), y\big)$. The stochastic gradient for a data sample $(x_s, y_s) \in \mathcal{D}$ is $\mathbf{g}_s = \nabla_{\boldsymbol{\theta}} \ell\big(\mathbf{M}(x_s;\boldsymbol{\theta}), y_s\big)$. Federated Averaging (FedAvg) [31] operates through T global rounds. At global round $t \in [T]$, the server broadcasts the current global model parameters $\boldsymbol{\theta}^t$ to all clients. Each client i updates $\boldsymbol{\theta}^t$ via local SGD on $\mathcal{D}_i$: $\boldsymbol{\theta}_i^t = \boldsymbol{\theta}^t - \eta \cdot \nabla_{\boldsymbol{\theta}} \mathcal{L}_i(\boldsymbol{\theta}^t)$, where $\mathcal{L}_i(\boldsymbol{\theta}^t) = \frac{1}{|\mathcal{D}_i|} \sum_{(x,y)\in\mathcal{D}_i} \ell\big(\mathbf{M}(x;\boldsymbol{\theta}^t), y\big)$. Server aggregates via weighted averaging:

$$\boldsymbol{\theta}^{t+1} = \sum_{i=1}^{H} \frac{|\mathcal{D}_i|}{|\mathcal{D}|} \boldsymbol{\theta}_i^t, \quad |\mathcal{D}| = \sum_{i=1}^{H} |\mathcal{D}_i|. \tag{1}$$

The final global model after T rounds is $\boldsymbol{\theta}^T$.

FU Scenarios. Let $\mathcal{C}_n \subseteq [H]$ denote clients retaining their original datasets $\{\mathcal{D}_j\}_{j\in\mathcal{C}_n}$, and $\mathcal{C}_u$ represent unlearned clients modifying their local datasets $\{\mathcal{D}_i\}_{i\in\mathcal{C}_u}$. Following [11], we formalize three scenarios: (i) *sample-level unlearning:* For each client $i \in \mathcal{C}_u$, partition $\mathcal{D}_i$ into retained $\mathcal{D}_i^r$ and forgotten subsets $\mathcal{D}_i^f = \mathcal{D}_i \setminus \mathcal{D}_i^r$; (ii) *class-level unlearning:* Each client $i \in \mathcal{C}_u$ removes all samples of target class y^f, yielding $\mathcal{D}_i^f = \{(x,y) \in \mathcal{D}_i \mid y = y^f\}$ with $\mathcal{D}_i^r = \mathcal{D}_i \setminus \mathcal{D}_i^f$; (iii)

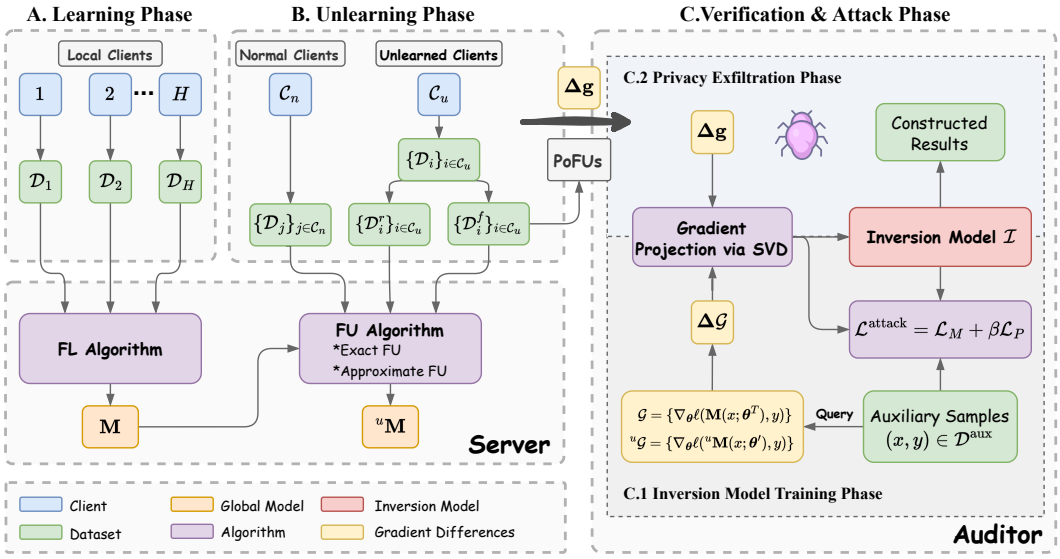


Figure 2: Schematic overview of IGF framework. **A. Learning Phase**: Clients collaboratively train the global model via FL. **B. Unlearning Phase**: The unlearned clients are required to forget specific data contributions and submit the proof of federated unlearning (PoFU). **C. Verification & Attack Phase**: The *honest-but-curious* auditor verifies PoFUs, while attempting to infer forgotten data using a pre-trained inversion model $\mathcal{I}$.

client-level unlearning: Each client $i \in \mathcal{C}_u$ sets $\mathcal{D}_i^f = \mathcal{D}_i$ and $\mathcal{D}_i^r = \emptyset$. We denote the unlearned global model as ${}^u\mathbf{M}$, the forgotten dataset as $\mathcal{D}^{\text{forgotten}} = \bigcup_{i\in\mathcal{C}_u} \mathcal{D}_i^f$ and the retained dataset as $\mathcal{D}^{\text{retained}} = (\bigcup_{j\in\mathcal{C}_n} \mathcal{D}_j) \cup (\bigcup_{i\in\mathcal{C}_u} \mathcal{D}_i^r)$.

FU Methods. We implement two mainstream FU approaches: (i) **EFU** retrains the global model on dataset $\mathcal{D}^{\text{retained}}$ from scratch, minimizing $\sum_{(x,y)\in\mathcal{D}^{\text{retained}}} \ell(\mathbf{M}(x;\boldsymbol{\theta}), y)$. This method precisely removes contributions of $\mathcal{D}^{\text{forgotten}}$ from the global model. (ii) **AFU** performs projected gradient ascent and constrains maximization on $\mathcal{D}^{\text{forgotten}}$. For each client $i \in \mathcal{C}_u$, it computes $\boldsymbol{\theta}_i' = \boldsymbol{\theta}^T + \eta_u \cdot \nabla_{\boldsymbol{\theta}} \mathcal{L}_i'(\boldsymbol{\theta}^T)$ where $\mathcal{L}_i'(\boldsymbol{\theta}^T) = \frac{1}{|\mathcal{D}_i^f|} \sum_{(x,y)\in\mathcal{D}_i^f} \ell\big({}^u\mathbf{M}(x;\boldsymbol{\theta}^T), y\big)$ but maintains $\|\boldsymbol{\theta}_i' - \boldsymbol{\theta}^T\|_2 \leq \zeta$, where ζ is the parameter deviation constraint. Then the server aggregates the unlearned local model parameters:

$$\boldsymbol{\theta}' = \sum_{i\in\mathcal{C}_u} \frac{|\mathcal{D}_i^f|}{|\mathcal{D}^{\text{forgotten}}|} \boldsymbol{\theta}_i', \quad |\mathcal{D}^{\text{forgotten}}| = \sum_{i\in\mathcal{C}_u} |\mathcal{D}_i^f|, \tag{2}$$

and fine-tunes ${}^u\mathbf{M}$ with $\boldsymbol{\theta}'$ on $\mathcal{D}^{\text{retained}}$.

Verification in FU. Each unlearned client $i \in \mathcal{C}_u$ locally computes PoFU of gradient differences $\boldsymbol{\Delta}\mathbf{g}^{(n_i)} = \Big\{\boldsymbol{\Delta}\mathbf{g}_j^{(n_i)} = \nabla_{\boldsymbol{\theta}}\ell\big(\mathbf{M}(x_j;\boldsymbol{\theta}^T), y_j\big) - \nabla_{\boldsymbol{\theta}}\ell\big({}^u\mathbf{M}(x_j;\boldsymbol{\theta}'), y_j\big) | (x_j, y_j) \in \mathcal{D}_i^f\Big\}$. Auditor receives PoFUs $\boldsymbol{\Delta}\mathbf{g} = \{\boldsymbol{\Delta}\mathbf{g}^{(n_i)}\}_{i\in\mathcal{C}_u}$ and validates unlearning by checking each $\|\boldsymbol{\Delta}\mathbf{g}_j^{(n_i)}\|_2 \leq \tau$ with predefined threshold τ. The necessity of the gradient differences in verifiable FU lies in ensuring that a data point (x, y) is included in the training dataset of the original model $\mathbf{M}$ but excluded from that of the unlearned model ${}^u\mathbf{M}$.

Threat Assumption. We model the auditor, denoted $\mathcal{A}$, as an *honest-but-curious* entity that strictly follows the FU protocol but seeks to infer private client data. Consistent with prior reconstruction attacks [30, 16, 15, 14], $\mathcal{A}$ possesses an auxiliary dataset $\mathcal{D}^{\text{aux}}$. Operating in a gray-box setting, $\mathcal{A}$ lacks knowledge of the global model's architecture but can collude with the server to query the flattened gradient for arbitrary samples from both the original model $\mathbf{M}$, and the unlearned model ${}^u\mathbf{M}$. During the exploitation phase, $\mathcal{A}$ passively collects PoFUs $\boldsymbol{\Delta}\mathbf{g}$, and endeavors to reconstruct the forgotten samples.

3.2 Framework of IGF

We adopt a learning-based inversion model to invert gradient differences to forgotten samples during the verification phase of FU. The main schematic of IGF is shown in Figure 2, and the formalized details are as follows:

Inversion Model Training Phase. (i) **Preparation of Training Dataset**. To prepare the training data for inversion model $\mathcal{I}$, for each data point (x, y) in auxiliary dataset $\mathcal{D}^{\text{aux}}$, the auditor $\mathcal{A}$ collects gradients:

$$\begin{cases} \mathcal{G} = \{\nabla_{\boldsymbol{\theta}} \ell(\mathbf{M}(x; \boldsymbol{\theta}^T), y)\}_{(x,y)\in\mathcal{D}^{\text{aux}}} \\ {}^{u}\mathcal{G} = \{\nabla_{\boldsymbol{\theta}} \ell({}^{u}\mathbf{M}(x; \boldsymbol{\theta}'), y)\}_{(x,y)\in\mathcal{D}^{\text{aux}}}, \end{cases} \tag{3}$$

where $\mathcal{G}$ and ${}^{u}\mathcal{G}$ denote the sets of flatten gradients queried from $\mathbf{M}$ and ${}^{u}\mathbf{M}$, respectively. Gradient differences $\boldsymbol{\Delta}\mathcal{G} = \{\mathcal{G}_i - {}^{u}\,\mathcal{G}_i | (x_i, y_i) \in \mathcal{D}^{\text{aux}}\}$ form a set of d-dimensional vectors, with d as the number of trainable parameters.

(ii) **Gradient Differences Projection via SVD**. To extract the key features and address redundancy caused by the high dimensionality of gradient differences, $\mathcal{A}$ projects $\boldsymbol{\Delta}\mathcal{G}$ to a lower-dimensional space using SVD. Let the m denote the number of samples in $\mathcal{D}^{\text{aux}}$, $\mathcal{A}$ constructs a matrix $\boldsymbol{\Psi} = [\boldsymbol{\Delta}\mathcal{G}_1^\top, \boldsymbol{\Delta}\mathcal{G}_2^\top, \ldots, \boldsymbol{\Delta}\mathcal{G}_m^\top] \in \mathbb{R}^{m\times d}$ from the gradient differences $\{\boldsymbol{\Delta}\mathcal{G}_i\}_{i=1}^m$ of the auxiliary dataset $\mathcal{D}^{\text{aux}}$, where $m \ll d$ typically holds. $\mathcal{A}$ centers the gradient differences by subtracting the mean vector $\boldsymbol{\mu} = \frac{1}{m}\sum_{i=1}^m \boldsymbol{\Delta}\mathcal{G}_i$, resulting in $\boldsymbol{\Delta}\mathcal{G}_i^{\text{cen}} = \boldsymbol{\Delta}\mathcal{G}_i - \boldsymbol{\mu}$ and the centered matrix $\boldsymbol{\Psi}^{\text{cen}}$. decomposes the centered matrix $\boldsymbol{\Psi}^{\text{cen}}$:

$$\boldsymbol{\Psi}^{\text{cen}} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top, \tag{4}$$

with $\mathbf{U} \in \mathbb{R}^{m\times m}$, $\mathbf{V} \in \mathbb{R}^{d\times d}$, and diagonal matrix $\boldsymbol{\Sigma}$ contains singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_m \geq 0$. To preserve essential information while reducing dimensionality, $\mathcal{A}$ selects the smallest k such that the cumulative explained variance exceeds a threshold ν:

$$k = \min\left\{ j \,\middle|\, \sum_{i=1}^{j} \sigma_i^2 / \sum_{i=1}^{m} \sigma_i^2 \geq \nu \right\}. \tag{5}$$

So $\mathcal{A}$ gets the projection matrix $\mathbf{V}^{[k]}$ denotes the first k columns of $\mathbf{V}$. And the projected gradient differences of $\mathcal{D}^{\text{aux}}$ are computed as $\boldsymbol{\Delta}\mathcal{G}^{\text{proj}} = \boldsymbol{\Psi}\mathbf{V}^{[k]} \in \mathbb{R}^{m\times k}$.

(iii) **Training Inversion Model**. $\mathcal{A}$ trains the inversion model, denoted as $\mathcal{I}$ and parameterized by $\boldsymbol{\omega}$, to map projected gradient differences to samples in $\mathcal{D}^{\text{aux}}$ by minimizing the composite loss function:

$$\mathcal{L}^{\text{attack}}(\boldsymbol{\omega}) = \mathcal{L}_M(\boldsymbol{\omega}) + \beta\mathcal{L}_P(\boldsymbol{\omega}), \tag{6}$$

where β trades off between pixel-level accuracy and perceptual quality. This design is common in image reconstruction tasks and can flexibly adjust the optimization objectives of the model to ensure that the reconstruction results are both accurate and natural. specifically, $\mathcal{L}_M$ quantifies the structural pixel-level discrepancy between reconstructed image $\mathcal{I}(\boldsymbol{\Delta}\mathcal{G}_i^{\text{proj}}; \boldsymbol{\omega})$ and ground truth image x_i:

$$\mathcal{L}_M(\boldsymbol{\omega}) = \frac{1}{m}\sum_{i=1}^{m} \|\mathcal{I}(\boldsymbol{\Delta}\mathcal{G}_i^{\text{proj}}; \boldsymbol{\omega}) - x_i\|_2^2. \tag{7}$$

Similarly, we define $\mathcal{L}_P$, which measures the semantic similarity between the reconstructed and true images using a VGG-based feature extractor $\phi(\cdot)$:

$$\mathcal{L}_P(\boldsymbol{\omega}) = \frac{1}{m}\sum_{i=1}^{m} \|\phi\left(\mathcal{I}(\boldsymbol{\Delta}\mathcal{G}_i^{\text{proj}}; \boldsymbol{\omega})\right) - \phi\left(x_i\right)\|_2^2 \tag{8}$$

Further, we elaborately designed the architecture of $\mathcal{I}$ to capture the latent mapping between gradient differences and images effectively. $\mathcal{I}$ employs a pixel-level convolutional network for progressive upsampling, which reduces artifacts in the reconstructed images. This design facilitates a nonlinear transformation from PoFU space to structured image space. Further architectural details are provided in Appendix D.

Privacy Exfiltration Phase. Following the training phase, the auditor $\mathcal{A}$ possesses the projection matrix $\mathbf{V}^{[k]}$ and the inversion model $\mathcal{I}$ with parameter $\boldsymbol{\omega}$. Upon receiving PoFUs, for each PoFU $\mathbf{\Delta g}^{(n_i)}$ of each client $i \in \mathcal{C}_u$, $\mathcal{A}$ constructs the matrix $\mathbf{\Psi}^{(n_i)} = \left[{\mathbf{\Delta g}_1^{(n_i)}}^\top, {\mathbf{\Delta g}_2^{(n_i)}}^\top, \ldots, {\mathbf{g}_{n_i}^{(n_i)}}^\top \right] \in \mathbb{R}^{n_i \times d}$, where n_i denotes the number of samples in $\mathcal{D}_i^f$. This matrix is then projected into a lower-dimensional space $\mathbf{\Delta g}^{(n_i)^{\text{proj}}} = \mathbf{\Psi}^{(n_i)} \mathbf{V}^{[k]} \in \mathbb{R}^{n_i \times k}$. The batched reconstruction of projected gradient differences $\mathbf{\Delta g}^{(n_i)^{\text{proj}}}$ is performed as follows:

$$\hat{\mathbf{x}}^{(n_i)} = \{\hat{x}_j = \mathcal{I}(\mathbf{\Delta g}_j^{(n_i)^{\text{proj}}}; \boldsymbol{\omega}) | j \in [n_i]\}, \tag{9}$$

where $\hat{\mathbf{x}}^{(n_i)} = \{\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_{n_i}\}$ represents the n_i reconstructed samples of client i. This exploitation enables $\mathcal{A}$ to utilize the pre-trained inversion model to implement the large-scale reconstructions from individual PoFU, thereby compromising data privacy even from the passive view.

3.3 Orthogonal Obfuscation Defense Method

Our inversion model exploits the directional information in gradient differences to reconstruct sensitive training data. Traditional defense methods often fail to disrupt the directional patterns, preserving the overall gradient differences structure and remaining susceptible to statistical recovery techniques. As illustrated in Figure 3, we propose a defense strategy that alters the vector direction while retaining the L2-norm information necessary for auditing. Our approach projects gradient differences into an orthogonal subspace, thereby disrupting the patterns and spatial structures that attackers rely on to reconstruct the forgotten sample.

For each PoFU $\mathbf{\Delta g}^{(n_i)}$ of unlearned client i, i needs to modify the direction of each entry $\mathbf{\Delta g}_j^{(n_i)}$ but maintain its L2-norm. We introduce random vectors $\mathbf{r}^{(n_i)}$ that are orthogonal to $\mathbf{\Delta g}^{(n_i)}$ element-wisely. The construction begins by sampling an initial random vector $\mathbf{r}_j^{(n_i)}$ with the same dimensionality as $\mathbf{\Delta g}_j^{(n_i)}$, drawn from a standard normal distribution $\mathbf{r}_j^{(n_i)} \sim \mathcal{N}(0,1)^d$. Then client i applies the Gram-Schmidt orthogonalization [32] to compute:

$$\mathbf{\Delta g}_j^{(n_i)^{\text{obf}}} = \mathbf{r}_j^{(n_i)} - \frac{{\mathbf{r}_j^{(n_i)}}^\top \mathbf{\Delta g}_j^{(n_i)}}{\|\mathbf{\Delta g}_j^{(n_i)}\|^2} \mathbf{\Delta g}_j^{(n_i)}. \tag{10}$$

This step ensures that $\mathbf{\Delta g}_j^{(n_i)^{\text{obf}}}$ lies in a subspace orthogonal to $\mathbf{\Delta g}_j^{(n_i)}$, effectively decoupling its direction from the original PoFU vector while preserving the randomness needed for obfuscation.

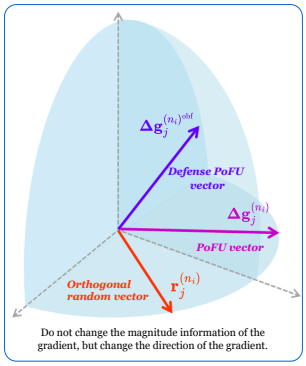


Figure 3: Schematic of orthogonal obfuscation defense

4 Experiment

4.1 Experiment Settings

Datasets and Metrics. We assess IGF framework on widely adopted benchmark datasets: CIFAR-10, CIFAR-100 [33], MNIST [34], and Fashion-MNIST [35]. These datasets offer diverse challenges, featuring varying image resolutions (32×32 for CIFAR, 28×28 for MNIST and Fashion-MNIST) and class numbers (10 to 100), making them an ideal testbed for assessing generalization. To measure the efficacy of our IGF, we employ established metrics for reconstruction attacks: Mean Squared Error (MSE), Peak Signal-to-Noise Ratio (PSNR), and Learned Perceptual Image Patch Similarity (LPIPS) [15, 16, 29, 36]. Further details are provided in Appendix B.1 and B.2.

Models. To investigate the impact of global model complexity on our attack, we adopt two architectures: a convolutional neural network (*ConvNet*) and a deeper residual network (*ResNet20*) [37]. These are tested on CIFAR-10 and CIFAR-100, enabling us to probe the attack's robustness across architectural variations and to explore how the proposed inversion model scales with the network complexity of the global model.

Table 1: Reconstruction performance (MSE, PSNR, and LPIPS) on CIFAR-10 and CIFAR-100 datasets with *ConvNet* and *ResNet20* as global models. Gradient differences are applied with no defense. Each cell reports results for EFU / AFU, with **bold** indicating the best performance.

Backbone	Method	FU Scenario	CIFAR-10			CIFAR-100		
			MSE ↓	PSNR ↑	LPIPS ↓	MSE ↓	PSNR ↑	LPIPS ↓
ConvNet	Ours	*sample-level*	0.0211 / **0.0218**	17.19 / **17.09**	**0.3261** / 0.3624	**0.0364** / **0.0261**	**14.97** / **16.07**	0.4383 / **0.4190**
	Ours	*class-level*	0.0259 / 0.0234	16.08 / 16.51	0.3531 / 0.3316	0.0397 / 0.0298	14.41 / 15.73	0.4451 / 0.4201
	Ours	*client-level*	**0.0206** / 0.0223	**17.32** / 16.78	0.3747 / 0.3558	0.0382 / 0.0265	14.65 / **16.07**	**0.4361** / 0.4223
	GIAMU	*sample-level*	0.2330 / 0.2460	13.22 / 12.78	0.3390 / **0.3190**	–	–	–
ResNet20	Ours	*sample-level*	0.0445 / 0.0564	14.05 / 13.02	**0.4607** / 0.4719	**0.0391** / **0.0353**	**14.56** / **15.02**	0.4267 / 0.4025
	Ours	*class-level*	0.0535 / **0.0512**	13.01 / **13.21**	0.4608 / **0.4366**	0.0474 / 0.0438	13.49 / 13.84	**0.4060** / 0.4032
	Ours	*client-level*	**0.0435** / 0.0533	**14.12** / 13.08	0.4617 / 0.4983	0.0422 / 0.0362	14.27 / 14.73	0.4187 / **0.3627**

Training Setup. In cross-silo FU and FL training, we configure 40 clients with 10% client selection and conduct 20 global rounds to derive the original and unlearned models. For unlearning, we designate 1000 samples to be forgotten, and our ablation experiments demonstrate that both in-distribution and out-of-distribution auxiliary data can effectively achieve attacking. We consider an *honest-but-curious* adversary $\mathcal{A}$ capable of storing or collecting a small auxiliary dataset, with a size comparable to a typical validation or test set, consistent with prior work [29, 30]. During the attack phase, we train the inversion model with a batch size of 256, a learning rate of 10^{-4}, and a fixed random seed of 1234 for reproducibility. All experiments are implemented in PyTorch and executed on NVIDIA A10 GPUs, with each training run requiring approximately 1 hour.

4.2 Experimental Results

This section presents comprehensive experiments validating the effectiveness of IGF framework across diverse datasets, FU scenarios (*sample-level*, *class-level*, *client-level*), FU methods (EFU/AFU), and global model architectures. We report numerical results and visual reconstructions to demonstrate IGF's high-fidelity reconstruction capabilities. Additionally, we evaluate IGF's resilience against five common defense mechanisms and highlight the efficacy of our proposed orthogonal obfuscation defense. Finally, we conduct extensive ablation studies to assess the contributions of IGF's key components.

Reconstruction Performance across Datasets. The results presented in Table 1 provide compelling evidence of IGF's capability to reconstruct forgotten data with high fidelity. On CIFAR-10 with *ConvNet* under EFU at the *sample-level* (1000 forgotten samples), IGF achieves an MSE of 0.0211, PSNR of 17.1947, and LPIPS of 0.3261, reflecting reconstructions with minimal pixel-wise errors and superior perceptual quality. On more challenging CIFAR-100, which contains 100 classes instead of 10, we observe a slight performance degradation, with MSE rising to 0.0364, PSNR dropping to 14.9658, and LPIPS increasing to 0.4383. This decline is expected, as greater dataset complexity and inter-class variability heighten the difficulty of inversion. Nevertheless, IGF still demonstrates significant robustness, which can be attributed to our SVD-based projection technique and a tailored inversion model.

Adaptability across Global Model Architectures. IGF also demonstrates adaptability across model architectures. On CIFAR-10, reconstruction performances with *ConvNet* are slightly better than *ResNet20* , with MSE values of 0.0211 and 0.0445, respectively. This gap stems from the increased complexity and depth of *ResNet20*, which leads to more complex gradient patterns that complicate inversion. Despite this, IGF still achieves satisfactory reconstruction quality even with the deeper *ResNet20* architecture, demonstrating the adaptability to diverse model architectures.

Adaptability across FU Scenarios. We test IGF under three FU scenarios: *sample-level unlearning* (the number of samples to be forgotten is set to 1000), *class-level unlearning* (the class index to be forgotten is set to 1), and *client-level unlearning* (all samples from the third client are set to be forgotten). IGF exhibits stable performance, with *ConvNet*'s MSE fluctuating within 0.0053 under EFU on CIFAR-10, indicating resilience to differing unlearning granularity. In other configurations, alterations to the FU scenarios have a negligible impact on reconstruction performance, further highlighting IGF's stability.

Vulnerability Comparison of FU Methods. Experimental results reveal certain gaps in vulnerability to reconstruction attacks between EFU and AFU methods. EFU outperforms AFU in reconstruction metrics on CIFAR-10 with *ResNet20*, as EFU's retraining from scratch yields clearer gradient

differences reflecting forgotten data's impact. In contrast, AFU's gradient ascent operation introduces noise, complicating reconstruction. Despite this, IGF achieves reasonable reconstruction quality under AFU, highlighting a critical privacy risk: even approximate unlearning methods remain vulnerable to reconstruction attacks.

Comparison with Basline [16]. As shown in Table 1, we also compare our IGF framework with GIAMU [16], a recent gradient inversion attack for centralized machine unlearning. All results associated with GIAMU are sourced directly from [16]. For *sample-level* unlearning on CIFAR-10 under EFU, IGF outperforms GIAMU by 88.1%, 30.1%, and 3.8% in MSE, PSNR, and LPIPS, respectively. Under AFU, gains are even more pronounced, with MSE and LPIPS improvements of 91.1% and 33.6%, respectively. Notably, few baseline attacks are suitable for direct comparison in FU reconstruction, as most existing methods target gradients from fully trained models, which are ill-suited for unlearning scenarios where models are modified to forget specific data. Moreover, IGF scales to reconstruct hundreds of times more samples than GIAMU, despite FU's distributed nature, which restricts data access and complicates attacks compared to GIAMU's white-box setting.

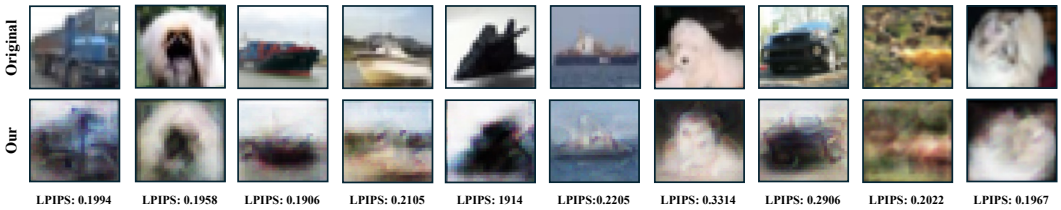


Figure 4: Original forgotten images and our reconstructed images on the CIFAR-10 dataset when the number of forgotten samples is 1000.

Visual Inspection of Reconstructed Images. Beyond quantitative metrics, visual inspection of the reconstructed images in Figure 4 provides additional insights into the effectiveness of IGF. The reconstructed images clearly capture the essential features of the original forgotten samples, including object shapes, colors, and textures. This visual similarity reinforces the quantitative results and demonstrates that our attack can reconstruct forgotten data with sufficient fidelity to pose a genuine privacy risk. In addition, we extend IGF to MNIST and Fashion-MINST, which are composed of different image sizes from CIFAR. The reconstructed results, as shown in Figure 12, show that reconstructed images are almost indistinguishable from the original images based on the gradient differences. The high-quality reconstruction is achieved through our composite optimization approach, which combines $\mathcal{L}_M$ with $\mathcal{L}_P$ loss. This combination ensures that the reconstructed images not only match the original images at the pixel level but also maintain perceptual similarity in terms of high-level features.

Table 2: Reconstruction performance across three metrics on five common defense mechanisms.

Defense Method	Gradient Pruning			Sign Compression	Gauss Noise	Gradient Perturb	Gradient Smooth
	0.7	0.8	0.9	0.001	0.1	0.01	0.1
MSE ↓	0.0216	0.0221	0.0222	0.0225	0.0298	0.0197	0.0232
PSNR ↑	17.0758	17.0694	17.0521	16.9704	15.7044	17.6116	16.8371
LPIPS ↓	0.3704	0.3796	0.3810	0.3796	0.4011	0.3663	0.3852

Reconstruction Performance against Defense Mechanisms. We evaluate the reconstruction performance of IGF against five common defense mechanisms. The technical details of these common defense mechanisms are introduced in Appendix B.3, and the reconstruction performance is shown in Table 2. When tested against Gradient Pruning with hyperparameters set to $\{0.7, 0.8, 0.9\}$, our method maintains consistent performance with MSE values of 0.0221, PSNR above 17, and LPIPS around 0.38. Against Sign Compression, which quantizes gradients to their signs, our method maintains stable performance, achieving MSE values of approximately 0.0221, PSNR above 17, and LPIPS around 0.38. When confronted with Gaussian Noise, our method still achieves reasonable reconstruction quality (MSE = 0.0225, PSNR = 15.7044), though with some performance degradation. This result stems from our learning-based inversion model, which has a strong mapping capability. This demonstrates significant resilience of IGF and allows IGF to almost ignore the defense mechanism and reconstruct the forgotten data. This also highlights the need for a novel defense method that can fundamentally disrupt the attacker's ability to reconstruct meaningful data.

Reconstruction Performance against Orthogonal Obfuscation Defense. As shown in the Figure 5, our proposed Orthogonal Obfuscation defense disrupts reconstruction by altering gradient difference directions while preserving their L2-norm. Reconstructed images exhibit random noise, effectively thwarting IGF and protecting sensitive data.

Figure 5: Forgotten images and our reconstructed images on the CIFAR-10 dataset under Orthogonal Obfuscation defense.

4.3 Ablation Studies

Figure 6: Forgotten images and our reconstructed images using our inversion model across different configurations.

To evaluate the contribution of each component in our attack framework, we conduct ablation studies, visualizing reconstruction results under various configurations, as shown in Figure 6. *Original* row shows the ground-truth forgotten samples. *Our w/o Modules* configuration, which excludes all proposed modules, exhibits severe degradation in reconstructed images, with prominent artifacts and loss of structural details. This underscores the inherent challenges of reconstruction attacks and the necessity of our enhancements. *Our w/o Composite Optimized* row, which excludes our composite loss-optimization module, produces images that preserve basic shapes but suffer from blurring, color inconsistencies, and a lack of fine details. This highlights the critical role of perceptual loss in capturing high-level semantic features beyond mere pixel-level reconstruction. In contrast, our complete model (*Our Full*), incorporating all proposed components, achieves reconstructions with significantly improved visual quality. These images exhibit sharper definitions, better preservation of textures, and more accurate color reproduction. By effectively balancing low-level pixel information and high-level semantic features, our comprehensive approach yields reconstructions that closely resemble the original forgotten samples. **Further ablation studies on federated aggregation methods, auxiliary datasets, dimensionality reduction techniques, and the hyperparameter β are provided in Appendix C.**

5 Conclusion

In this paper, we expose a critical privacy vulnerability in FU by proposing a novel reconstruction attack that exploits gradient differences used as PoFU. Our proposed IGF leverages the latent correlations between gradient differences and forgotten samples to reconstruct large-scale private data from individual PoFU. Through extensive experiments, we demonstrate that our attack achieves high-fidelity reconstruction, exposing the inadequacy of existing FU safeguards. To counter this threat, we introduce an orthogonal obfuscation defense that disrupts the reconstruction process, forcing inverted images into fixed noise patterns that resist recovery. Our findings underscore the fragility of current FU mechanisms against gradient-based and gradient-difference-based attacks, highlighting the urgent need for robust defenses and motivating further exploration of secure FU strategies.

References

[1] Jeffrey Rosen. The right to be forgotten. *Stan. L. Rev. Online*, 64:88, 2011.

[2] Stuart L Pardau. The california consumer privacy act: Towards a european-style privacy regime in the united states. *J. Tech. L. & Pol'y*, 23:68, 2018.

[3] Gaoyang Liu, Xiaoqiang Ma, Yang Yang, Chen Wang, and Jiangchuan Liu. Federated unlearning. *arXiv preprint arXiv:2012.13891*, 2020.

[4] Cassandra Lindstrom. Federated unlearning in financial applications. *preprints202409.1816*, 2024.

[5] Xiangshan Gao, Xingjun Ma, Jingyi Wang, Youcheng Sun, Bo Li, Shouling Ji, Peng Cheng, and Jiming Chen. Verifi: Towards verifiable federated unlearning. *IEEE Transactions on Dependable and Secure Computing*, 2024.

[6] Jiasi Weng, Shenglong Yao, Yuefeng Du, Junjie Huang, Jian Weng, and Cong Wang. Proof of unlearning: Definitions and instantiation. *IEEE Transactions on Information Forensics and Security*, 19:3309–3323, 2024.

[7] Xuhan Zuo, Minghao Wang, Tianqing Zhu, Lefeng Zhang, Shui Yu, and Wanlei Zhou. Federated learning with blockchain-enhanced machine unlearning: A trustworthy approach. *IEEE Transactions on Services Computing*, pages 1–15, 2025.

[8] Ahmed Salem, Apratim Bhattacharya, Michael Backes, Mario Fritz, and Yang Zhang. {Updates-Leak}: Data set inference and reconstruction attacks in online learning. In *29th USENIX security symposium (USENIX Security 20)*, pages 1291–1308, 2020.

[9] Chen Wu, Sencun Zhu, and Prasenjit Mitra. Federated unlearning with knowledge distillation. *arXiv preprint arXiv:2201.09441*, 2022.

[10] Junxiao Wang, Song Guo, Xin Xie, and Heng Qi. Federated unlearning via class-discriminative pruning. In *Proceedings of the ACM web conference 2022*, pages 622–632, 2022.

[11] 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. *arXiv preprint arXiv:2502.20709*, 2025.

[12] Franziska Boenisch, Adam Dziedzic, Roei Schuster, Ali Shahin Shamsabadi, Ilia Shumailov, and Nicolas Papernot. When the curious abandon honesty: Federated learning is not private. In *2023 IEEE 8th European Symposium on Security and Privacy (EuroS&P)*, pages 175–199. IEEE, 2023.

[13] Junqing Le, Di Zhang, Xinyu Lei, Long Jiao, Kai Zeng, and Xiaofeng Liao. Privacy-preserving federated learning with malicious clients and honest-but-curious servers. *IEEE Transactions on Information Forensics and Security*, 18:4329–4344, 2023.

[14] Ligeng Zhu, Zhijian Liu, and Song Han. Deep leakage from gradients. *Advances in neural information processing systems*, 32, 2019.

[15] Jonas Geiping, Hartmut Bauermeister, Hannah Dröge, and Michael Moeller. Inverting gradients-how easy is it to break privacy in federated learning? *Advances in neural information processing systems*, 33:16937–16947, 2020.

[16] Hongsheng Hu, Shuo Wang, Tian Dong, and Minhui Xue. Learn what you want to unlearn: Unlearning inversion attacks against machine unlearning. In *2024 IEEE Symposium on Security and Privacy (SP)*, pages 3257–3275. IEEE, 2024.

[17] Anvith Thudi, Hengrui Jia, Ilia Shumailov, and Nicolas Papernot. On the necessity of auditable algorithmic definitions for machine unlearning. In *31st USENIX Security Symposium (USENIX Security 22)*, pages 4007–4022, Boston, MA, August 2022. USENIX Association.

[18] Zhuohang Li, Jiaxin Zhang, Luyang Liu, and Jian Liu. Auditing privacy defenses in federated learning via generative gradient leakage. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10132–10142, 2022.

[19] Yi Liu, Lei Xu, Xingliang Yuan, Cong Wang, and Bo Li. The right to be forgotten in federated learning: An efficient realization with rapid retraining. In *IEEE INFOCOM 2022-IEEE conference on computer communications*, pages 1749–1758. IEEE, 2022.

[20] Anisa Halimi, Swanand Kadhe, Ambrish Rawat, and Nathalie Baracaldo. Federated unlearning: How to efficiently erase a client in fl?, 2022. *URL https://arxiv. org/abs/2207.05521*, 2022.

[21] 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*, 2024.

[22] Heng Xu, Tianqing Zhu, Lefeng Zhang, Wanlei Zhou, and S Yu Philip. Update selective parameters: Federated machine unlearning based on model explanation. *IEEE Transactions on Big Data*, 2024.

[23] Hanlin Gu, Gongxi Zhu, Jie Zhang, Xinyuan Zhao, Yuxing Han, Lixin Fan, and Qiang Yang. Unlearning during learning: An efficient federated machine unlearning method. *arXiv preprint arXiv:2405.15474*, 2024.

[24] Jian Chen, Zehui Lin, Wanyu Lin, Wenlong Shi, Xiaoyan Yin, and Di Wang. Fedmua: Exploring the vulnerabilities of federated learning to malicious unlearning attacks. *IEEE Transactions on Information Forensics and Security*, 2025.

[25] Wenbin Wang, Qiwen Ma, Zifan Zhang, Yuchen Liu, Zhuqing Liu, and Minghong Fang. Poisoning attacks and defenses to federated unlearning. *arXiv preprint arXiv:2501.17396*, 2025.

[26] Chi Zhang, Zhang Xiaoman, Ekanut Sotthiwat, Yanyu Xu, Ping Liu, Liangli Zhen, and Yong Liu. Generative gradient inversion via over-parameterized networks in federated learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5126–5135, 2023.

[27] Jinwoo Jeon, Kangwook Lee, Sewoong Oh, Jungseul Ok, et al. Gradient inversion with generative image prior. *Advances in neural information processing systems*, 34:29898–29908, 2021.

[28] Hao Fang, Bin Chen, Xuan Wang, Zhi Wang, and Shu-Tao Xia. Gifd: A generative gradient inversion method with feature domain optimization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4967–4976, 2023.

[29] Yu Sun, Gaojian Xiong, Xianxun Yao, Kailang Ma, and Jian Cui. Gi-pip: Do we require impractical auxiliary dataset for gradient inversion attacks? In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4675–4679. IEEE, 2024.

[30] Ruihan Wu, Xiangyu Chen, Chuan Guo, and Kilian Q Weinberger. Learning to invert: Simple adaptive attacks for gradient inversion in federated learning. In *Uncertainty in Artificial Intelligence*, pages 2293–2303. PMLR, 2023.

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

[32] Orthogonalization. In *Encyclopedia of Mathematics*. EMS Press, 2001. `https://encyclopediaofmath.org/wiki/Orthogonalization`.

[33] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. *Placeholder Journal*, 2009.

[34] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[35] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

[36] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

[37] 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.

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

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

[40] Kilian Weinberger, Anirban Dasgupta, John Langford, Alex Smola, and Josh Attenberg. Feature hashing for large scale multitask learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 1113–1120, 2009.

A Discussion and Limitations

To the best of our knowledge, IGF framework is the first to exploit gradient differences as an attack surface in federated unlearning (FU). Previous reconstruction attacks in machine learning and federated learning (FL) [14, 15, 29, 30] directly leverage sample-level gradients, which inherently contain richer sample information. Currently, the only available baseline for reconstruction attacks in FU is GIAMU [16], which relies on white-box access to both the original and unlearned models and overlooks the privacy vulnerabilities arising from gradient differences sharing during verification. Our work addresses this gap by demonstrating that an *honest-but-curious* adversary with partial prior knowledge can reconstruct forgotten samples by inverting gradient differences. Additionally, in extreme scenarios, such as a black-box setting where the adversary lacks prior knowledge or cannot exploit the directionality of gradient differences, the attack's complexity increases significantly, making the reconstruction of forgotten data largely unexplored.

B Experimental Settings

Table 3: Mathematical notations

Notation	Description
$\mathcal{C}$	Set of Clients
$\mathcal{D}$	Global Dataset
H	The Number of Clients
$\mathbf{M}$	Original Global Model
${}^{u}\mathbf{M}$	Unlearned Global Model
$\mathbf{g}$	Stochastic Gradient
$\mathcal{G}$	Gradient Queried by Adversary
ℓ	Loss Function in Local Training
T	Number of Global Rounds
$\mathcal{I}$	Inversion Model
(x, y)	Data Point
B	Batch Size
$\phi(\cdot)$	Intermediate Feature Extractor
$\mathbf{\Delta}\mathcal{G}, \mathbf{\Delta g}$	Gradient Differences
d	Model Size
$\mathbf{U}$	Left Singular Vectors
$\mathbf{V}$	Right Singular Vectors
$\mathbf{r}$	Random Vector
$\mathcal{M}$	Mask Matrix in Gradient Pruning
ϵ	Gaussian Noise
w	Window Size in Gradient Smoothing
ζ	Parameter Deviation Constraint Radius
$\mathbf{\Psi}$	Gradient Differences Matrix
$\mathbf{V}^{[k]}$	Projection Matrix

B.1 Datasets

We evaluate IGF using the widely adopted CIFAR-10 and CIFAR-100 datasets [33], both standard benchmarks in reconstruction attacks and federated learning research. CIFAR-10 comprises 60,000 color images (32×32 pixels) across 10 categories, with 50,000 images for training and 10,000 for testing. Each category contains 6,000 images. CIFAR-100 is structured similarly but includes 100 categories, each with 600 images, totaling 60,000 images (50,000 for training and 10,000 for testing).

Additionally, we assess our approach on the MNIST [34] and Fashion-MNIST datasets [35]. MNIST consists of 70,000 grayscale images (28×28 pixels) of handwritten digits, divided into 60,000 training and 10,000 testing images. Fashion-MNIST, designed as a more challenging alternative, also contains 70,000 28×28 grayscale images but represents 10 categories of fashion items. It mirrors MNIST's training and testing split.

B.2 Details of Metrics

MSE measures the average squared difference between the original forgotten image and the reconstructed image. It is widely used as a loss function in image processing tasks and image quality assessment $\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{x}_i)^2$, where x_i is the pixel value of the original forgotten image and $\hat{x}_i$ is the pixel value of the reconstruction.

PSNR measures the quality of the reconstructed or compressed image relative to the forgotten image. It is expressed in decibels (dB) and is inversely related to MSE—lower MSE values correspond to higher PSNR values. $\text{PSNR} = 10 \cdot \log_{10}\left(\frac{R^2}{\text{MSE}}\right)$, Where R is the maximum pixel value.

LPIPS [36] is a perceptual similarity metric designed to assess the perceptual quality of images based on learned features from a neural network (typically a pretrained deep network like VGG). Unlike MSE and PSNR, LPIPS is more aligned with human visual perception, focusing on perceptual similarity rather than pixel-level accuracy $\text{LPIPS}(x, \hat{x}) = \frac{1}{L}\sum_{l=1}^{L}\|\phi_l(x) - \phi_l(\hat{x})\|_2^2$, where L is the total number of layers used for feature extraction. $\|\cdot\|_2$ is the Euclidean distance (L2 norm) between the feature maps.

B.3 Details of Common Defense Mechanisms

This section outlines five defense mechanisms [30] designed to obfuscate shared gradients and mitigate gradient-based reconstruction attacks through various perturbation techniques. Given an input gradient vector $\mathbf{g}$, each mechanism produces an obfuscated gradient vector $\mathbf{g}'$. We adapt these mechanisms to perturb shared gradient differences in FU.

(a) **Sign Compression.** The sign compression mechanism applies the sign operation to each component of the gradient $\mathbf{g}$, retaining only its sign (-1, 0, or 1) and discarding magnitude information. This preserves the gradient's direction while significantly reducing communication overhead, as only sign bits are transmitted. By limiting the attacker's access to sign information, this method increases the difficulty of reconstructing forgotten data. The operation is defined as:

$$\mathbf{g}' = \mathsf{sign}(\mathbf{g}), \quad \text{where} \quad \mathsf{sign}(\mathbf{g}_i) = \begin{cases} 1, & \text{if } \mathbf{g}_i > 0 \\ -1, & \text{if } \mathbf{g}_i < 0 \\ 0, & \text{if } \mathbf{g}_i = 0 \end{cases} \tag{11}$$

(b) **Gradient Pruning.** Gradient pruning sparsifies the gradient by retaining only the k components with the largest absolute values, setting all others to zero. A binary mask $\mathcal{M}$ selectively preserves these significant components. Widely used in FL to reduce communication costs, this method also enhances privacy by limiting the attacker's access to a subset of gradient components, complicating the inference of forgotten data. The operation is formulated as:

$$\mathbf{g}' = \mathbf{g} \odot \mathcal{M}, \tag{12}$$

where $\odot$ denotes element-wise multiplication, and $\mathcal{M}$ is the mask matrix.

(c) **Gaussian Noise.** This mechanism perturbs the gradient $\mathbf{g}$ by adding independent and identically distributed Gaussian noise $\epsilon \sim \mathcal{N}(0, \sigma^2\mathbf{I})$. Controlled by the standard deviation σ, the noise introduces uncertainty to achieve differential privacy, obscuring precise gradient values and hindering reconstruction of forgotten data. The operation is expressed as:

$$\mathbf{g}' = \mathbf{g} + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2\mathbf{I}). \tag{13}$$

(d) **Gradient Perturbation.** This method perturbs the gradient by adding noise proportional to the gradient's magnitude, applying larger perturbations to dimensions with greater gradient values. The perturbed gradient is defined as:

$$\mathbf{g}' = \mathbf{g} + (\mathcal{N}(\mathbf{0}, \mathbf{I}) \times \text{scale}) \times (|\mathbf{g}| \times \text{factor}), \tag{14}$$

where $\mathcal{N}(\mathbf{0}, \mathbf{I})$ is a standard normal random tensor, scale determines the base perturbation magnitude, and factor adjusts the sensitivity of the perturbation to the gradient's amplitude.

(e) **Gradient Smoothing.** Gradient smoothing mitigates high-frequency variations in the gradient by applying a moving average over the feature dimensions, blending the result with the original gradient.

The operation is formulated as:

$$\mathbf{g}' = \text{reshape}\left((1-\alpha_{\text{gs}})\mathbf{g}^{\text{flat}} + \alpha_{\text{gs}} \cdot \text{MA}_w(\mathbf{g}^{\text{flat}})\right), \tag{15}$$

where $\mathbf{g}^{\text{flat}}$ is the flattened gradient, MA_w denotes the moving average with window size w, and $\alpha_{\text{gs}} \in [0, 1]$ controls the smoothing intensity.

C Additional Ablation Studies

C.1 Impact of Different Federated Aggregation Methods

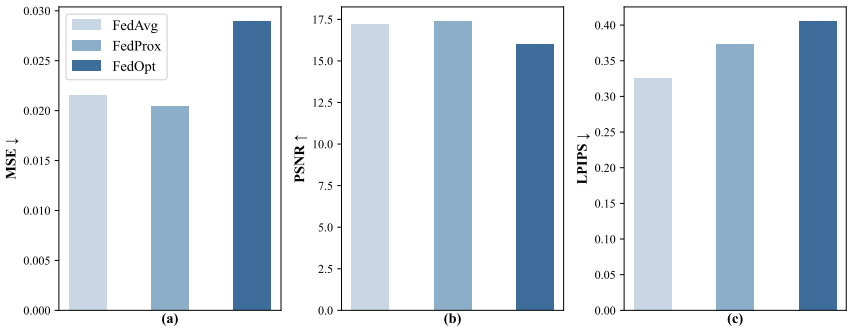


Figure 7: The reconstruction performance under different federated aggregation methods.

We investigated how three federated aggregation methods, including FedAvg [31], FedProx [38], and FedOpt [39], affect the p of reconstruction attacks in FU scenarios. Figure 7 illustrates the performance of our attack method across various aggregation algorithms commonly used in FL systems. The results demonstrate that while aggregation methods can influence reconstruction quality, our attack remains effective across different techniques. When examining more sophisticated aggregation methods like FedProx and FedOpt, we observe slightly different reconstruction patterns, but the overall attack effectiveness remains consistent.

C.2 Impact of Different Distributions of Auxiliary Datasets

Table 4: The reconstruction performance of different distributions of auxiliary datasets.

Distribution of Auxiliary Datasets	MSE ↓	PSNR ↑	LPIPS ↓
In-Distribution	0.0211	17.1947	0.3261
Out-of-Distribution	0.0259	16.6364	0.3324

In real-world scenarios, adversaries often struggle to obtain the complete data distribution of clients. To investigate the effectiveness of IGF attacks under entirely different distributions, we partition the CIFAR-10 training dataset as the federated client dataset and employ CIFAR-100, comprising entirely different categories, as auxiliary data. This setup simulates a reconstruction attack where the adversary lacks knowledge of the client data distribution. As demonstrated in Table 4, the IGF attack retains strong efficacy even with out-of-distribution auxiliary data, exhibiting only marginal degradation across various performance metrics.

C.3 Impact of Different Auxiliary Dataset Sizes

We investigate the influence of varying auxiliary dataset sizes on the efficacy of our attack method. As illustrated in Figure 8, we incrementally scale the dataset from 500 to 10,000 samples. Experimental results reveal that performance metrics stabilize when the auxiliary dataset comprises approximately 6,000 to 8,000 samples, demonstrating that our method achieves efficient and robust performance without requiring extensive auxiliary data. Notably, even with a modest dataset size, our proposed attack method effectively leverages available knowledge to deliver high-quality image reconstruction.

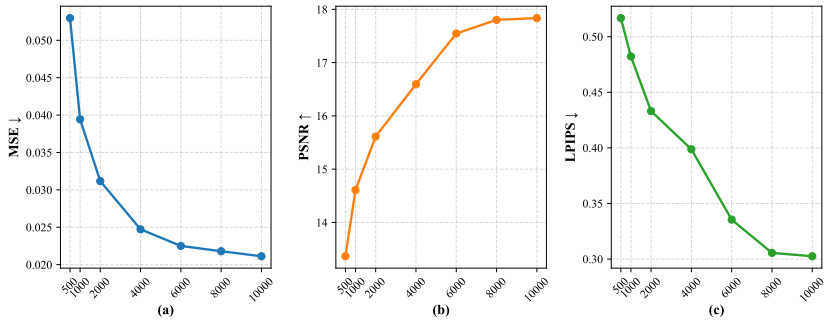


Figure 8: The reconstruction performance with different auxiliary dataset sizes.

C.4 Comparative Ablation Study of Dimensionality Reduction Methods

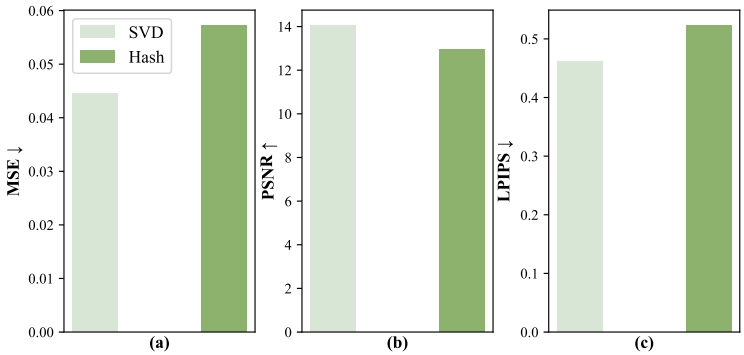


Figure 9: Comparison of the reconstruction effectiveness with applying SVD and Hash dimensionality reduction.

To gain a deeper understanding of the effectiveness of dimensionality reduction methods, we compared the performance of Hash-based dimensionality reduction and Singular Value Decomposition (SVD) in terms of reduction quality and reconstruction results. Hash-based dimensionality reduction [40] is a vector compression method that relies on random projection, mapping high-dimensional gradient differences to a lower-dimensional space through a sparse random matrix. Specifically, a sparse matrix is constructed where each high-dimensional vector component is randomly assigned to a lower-dimensional target dimension, and each reduced dimension represents the cumulative sum of the corresponding high-dimensional gradient differences. This approach is computationally efficient and well-suited for rapidly compressing gradient differences. However, its randomness disregards the inherent structure of the gradient differences, potentially leading to significant information loss.

As shown in Figure 9, SVD outperforms the reconstruction after Hash dimensionality reduction in both reconstruction effects, and as shown in Table 5 achieves more significant dimensionality reduction by extracting only key information. SVD-based dimensionality reduction is a data-driven method that decomposes the covariance matrix of the gradient differences to extract principal component directions as the projection basis. SVD dynamically selects the number of dimensions to retain a substantial portion of the variance (e.g., 95%), ensuring that the reduced results capture the primary patterns of the original gradient differences.

Method	Size
Original	269722
Hash	134861
SVD	433

Table 5: Comparison of the effectiveness of SVD and Hash for gradient differences reduction.

SVD outperforms Hash-based reduction because it prioritizes the retention of critical information while minimizing the impact of irrelevant noise. Furthermore, in reconstruction tasks, SVD-preserved gradient differences maintain structured features, enabling inversion models to more effectively learn the mapping from lower-dimensional features to the original data, resulting in higher-quality reconstructed images. Conversely, Hash-based reduction disrupts the gradient differences structure

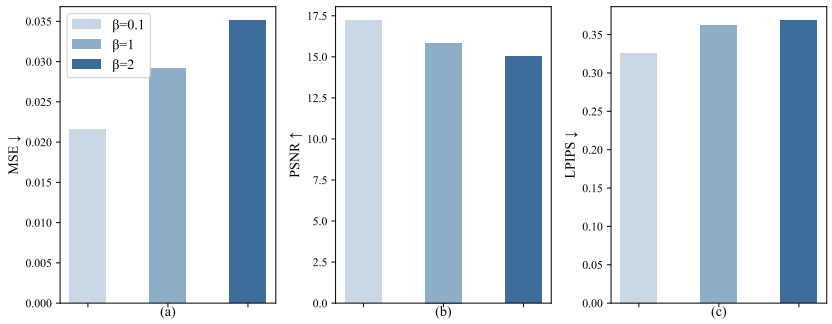


Figure 10: The reconstruction performance under different β.

through random mixing, making it challenging for reconstruction networks to disentangle useful information, which often leads to blurry or distorted reconstructed images.

C.5 Comparative Ablation Study of different β parameters

To investigate the role of the parameter β in the loss function, which governs the trade-off between pixel-level accuracy and perceptual quality, we conduct an ablation study to assess its impact on reconstruction attack performance. Specifically, we evaluate the effect of varying $\beta \in \{0.1, 1.0, 2.0\}$ on three key metrics: MSE, PSNR, and LPIPS. As shown in Figure 10, increasing β reveals a clear trade-off: pixel-level accuracy degrades, as indicated by worsening MSE, and perceptual quality diminishes, as reflected by deteriorating LPIPS, while PSNR exhibits a peak at an intermediate β before declining. These findings underscore β's critical role in mediating the balance between pixel-wise fidelity and high-level perceptual features.

D Inversion Model Architecture

As illustrated in Figure 11, our pixel-level inversion model features a carefully designed architecture comprising multiple Conv2d and BatchNorm2d layers. We incorporate PixelShuffle for effective upsampling, minimizing artifacts in reconstructed results. A linear layer paired with an initial Reshape operation enhances input processing, while a final Sigmoid activation and Reshape ensure high-quality output generation.

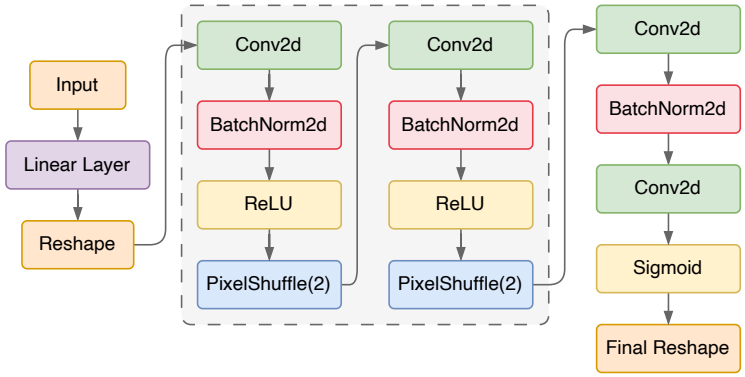
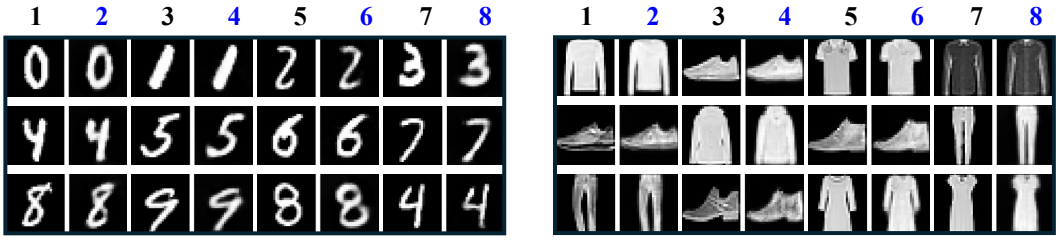


Figure 11: Architecture of the proposed pixel-level inversion model.

E Additional Reconstructed Images

This section showcases the forgotten images and their corresponding reconstructions across multiple datasets, as presented in Figures 12, 13, and 14. In each figure, **odd columns display the original images**, and **even columns show our reconstructed results**.



(a) MNIST (b) Fashion-MNIST

Figure 12: Forgotten and reconstructed images on MNIST and Fashion-MNIST within 1,000 randomly forgotten samples.

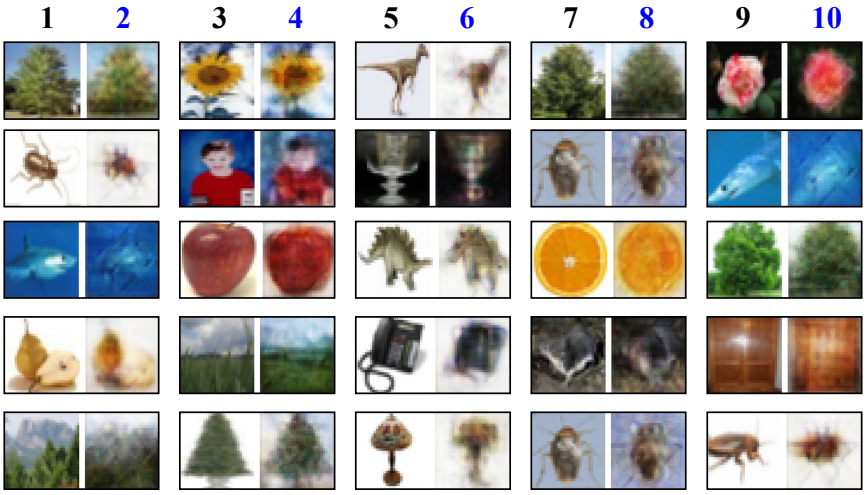


Figure 13: Forgotten and reconstructed images on CIFAR-100.

For the scenario of *class-level unlearning*, Figure 14 presents the forgotten images and reconstruction results on CIFAR-10 for the unlearned class (car).

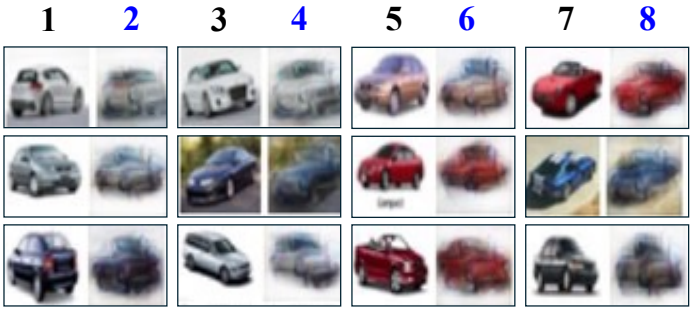


Figure 14: Forgotten and reconstructed images on CIFAR-10 for the unlearned class (car).