

GDFA: Geometry-Driven Federated Unlearning with Directional Task Vector Alignment

Xiuting Weng^{1,†} Ruizhi Pu^{2,†} Yuanhang Yao² Kun Yue¹ Zhiwen Tang¹ Lixing Yu^{1,*}
¹Yunnan University, China ²Southeast University, China

Abstract

Federated Learning (FL) aims to leverage knowledge from diverse local clients while preserving data privacy without sharing raw local data. However, when deletion requests arise under privacy regulations, efficiently removing specific client data contributions from target clients can be challenging. Existing unlearning methods face significant limitations under Non-IID (Non-Independent and Identically Distributed) data distributions when attempting to unlearn specific target clients in FL. Models in sharp optimization regions can suffer catastrophic knowledge loss from minor parameter changes, exacerbating this forgetting due to conflicting parameter updates across clients caused by Non-IID data distributions in FL. Empirically, we observe that conflicting updates under Non-IID settings generate misaligned task vectors that fail to isolate target knowledge. Therefore, we exploit the loss landscape geometry in unlearning specific target clients. We demonstrate that migrating models to flat regions can enhance unlearning robustness in Non-IID FL. Correspondingly, we introduce GDFA, a framework that initially transitions the global model to a flat loss domain. Subsequently, relevant clients generate unlearning task vectors, which GDFA filters to retain only directionally consistent components. This process isolates shared knowledge attributes before precise removal through reverse vector aggregation, maximizing knowledge retention. Extensive experiments demonstrate that GDFA outperforms state-of-the-art methods in unlearning efficacy and efficiency across diverse datasets and architectures, with minimal accuracy loss on retained tasks.

1. Introduction

Federated Learning (FL) [44] is a technology that holds promise for addressing the increasingly severe privacy and security issues [9, 10, 19, 68]. At its core, it involves collaborative learning of models among distributed clients without the need to upload any sensitive data. However, it also needs

to comply with the "right to be forgotten" [13, 23, 60, 71] stipulated in laws and regulations, ensuring that personal data can be effectively erased in accordance with the continuously updated privacy policies.

Traditional machine unlearning selectively removes the influence of specific data from a single trained model, achieving significant progress [2, 5, 59]. However, unlike traditional unlearning, federated unlearning cannot directly access user data stored locally on clients [61]. It also faces unique FL challenges like Non-IID data [32, 43, 77], increasing task complexity. Thus, federated unlearning is demanding [33, 38], requiring solutions that balance distributed learning and privacy needs.

In response to these challenges, various federated unlearning paradigms [29, 52] have been proposed. Retraining-based methods impose prohibitive computational costs that scale poorly with model complexity and dataset size, making them impractical for real-world deployments [16]. Approximate unlearning techniques attempt to accelerate the process through various optimizations, but often suffer from significant accuracy degradation due to the incremental effect where prior parameter updates implicitly retain traces of unlearned data [11, 31, 41]. Pruning-based solutions [62] that target class-discriminative channels struggle to isolate specific client contributions in Non-IID data distributions which are prevalent in FL scenarios.

The limited effectiveness of these methods stems from their failure to address a key challenge in federated unlearning. The combination of data non-accessibility and Non-IID characteristics tends to drive optimized model parameters into sharp loss basins [67]. Models residing in such sharp regions exhibit extreme sensitivity to parameter perturbations during unlearning, leading to catastrophic forgetting of retained knowledge. Furthermore, the optimization trajectory may rapidly shift to distant regions of the parameter space, resulting in insufficient exploration for the effective removal of sensitive information.

This study introduces a novel framework that fundamentally addresses these limitations through geometry-aware unlearning. Our key insight recognizes that models lo-

[†] Contributed Equally. ^{*} Corresponding Author.

cated in flat optimization regions exhibit stronger generalization stability and greater tolerance to parameter modifications. Building on this foundation, we propose a three-stage methodology: First, we guide the global model to migrate from sharp to flat optimization regions using sharpness-aware minimization techniques. This migration creates a stable operational environment where unlearning operations can be performed with minimal disruption to retained knowledge. Second, relevant clients locally train on their target forgetting data to generate task vectors that capture the characteristics of knowledge to be removed. We implement a directional alignment mechanism that selectively aggregates only those task vectors exhibiting consistent optimization trajectories. Third, the global model precisely erases the target knowledge through reverse aggregation of these aligned task vectors while maximizing preservation of retained knowledge performance.

We summarized our main contributions as follows:

- We propose GDFA, a geometry-driven federated unlearning framework that migrates models to flat loss regions to enhance stability and leverages directional task vector alignment to precisely isolate target knowledge without requiring access to raw client data.
- Our theoretical contribution includes a PAC-Bayesian bound that links loss landscape flatness to performance gaps under data heterogeneity, ensuring robust unlearning with minimal impact on retained knowledge.
- We introduce a directional alignment mechanism for task vectors, where clients generate vectors on forgetting data, and only consistent directional changes are merged via sign consensus to isolate and precisely erase shared knowledge attributes, maximizing retention performance.
- We perform extensive experiments on multiple datasets and model architectures, showing GDFA’s superiority in unlearning efficacy, efficiency, and privacy guarantees compared to state-of-the-art methods.

2. Related Work

2.1. Machine Unlearning

Machine unlearning research mainly focuses on centralized learning scenarios [20, 37, 40], and its methods can be classified from perspectives such as operation objects, implementation mechanisms, and unlearning granularity. In terms of operation objects, they fall into two categories: data operations and model operations [47]. Data operations include data mixing methods that introduce interference or replace samples [59, 73], and data partitioning methods that split the training set for local retraining [8, 51]. Model operations cover model transformation [21, 64] for direct parameter adjustment, model pruning [62] to remove partial structures, and model replacement [3, 51] strategies that substitute affected sub-models. From the perspective of imple-

mentation mechanisms, they can be further categorized into data-driven [2, 6, 8] and model-driven approaches [1, 6]. The former achieves unlearning by reconstructing the training set, while the latter modifies the internal representations of the model through methods such as selective fine-tuning, parameter regularization, or adversarial perturbation. Additionally, based on unlearning granularity, they are divided into class-level [17, 55] and sample-level [2, 5, 6] unlearning. Class-level unlearning typically uses class-level masking or selective retraining to eliminate entire semantic class information, while sample-level methods achieve the removal of impacts from specific data through refined adjustments and are regarded as common solutions due to their stronger versatility [47].

2.2. Federated Unlearning

Federated Unlearning supports multiple objectives such as client unlearning [58], sample unlearning [65] and class unlearning [18]. Its common methods can be categorized into two types: retraining-based [41] and parameter operation-based [4, 7]. Retraining-based methods such as correcting gradient directions [39], employing clustering techniques [58] can retrain models without requiring the unlearned data, but their practical costs remain substantial. To improve efficiency, parameter operation-based methods have been proposed. For instance, unlearning can be achieved through historical update storage such as Fed-Eraser [39] and FUKD [65] which eliminate the influence of target clients through subtraction operations combined with knowledge distillation or gradient correction such as PGD [22] that uses inverse operations, MoDE [75] that employs a degraded model and FCU [14] that utilizes a contrastive loss function. However, historical update storage methods may pose privacy risks and storage pressures, while gradient correction methods often require the participation of target clients and incur high computational costs. Overall, current Federated Unlearning research still primarily focuses on client unlearning, and key issues such as knowledge interference have not yet been fully resolved.

2.3. Geometric Sensitivity in Unlearning

Recent studies reveal the critical impact of loss landscape geometry. Models in sharp minima exhibit catastrophic forgetting during modification, while flat regions demonstrate robust generalization [57]. This explains why existing methods face performance degradation: unlearning operations in sharp basins inevitably damage retained knowledge. Though sharpness-aware minimization (SAM) shows promise in standard training [15], its application to federated unlearning remains unexplored.

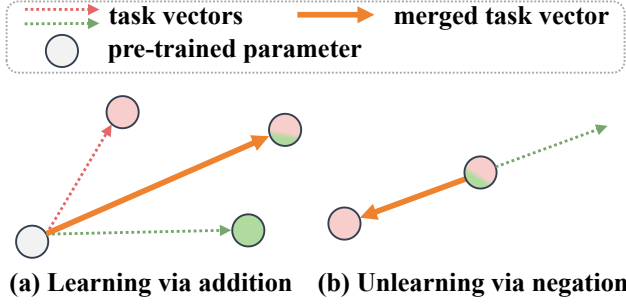


Figure 1. Illustration of task vector operations. In scenario (a), learning is achieved by adding task vectors to the pre-trained parameters. In scenario (b), unlearning is performed by negating the task vector. The effectiveness depends on vector alignment and magnitude, where conflicts or imbalances may lead to degraded performance.

2.4. Task Vector Alignment

Task vectors effectively capture knowledge characteristics through local model updates [30]. However, in FL settings, inconsistent optimization directions under Non-IID data lead to vector misalignment [54]. While FedRecovery [72] uses gradient residuals to approximate removal direction, it lacks explicit vector selection mechanisms. Recent centralized studies [34] suggest directional consistency improves unlearning precision, but this insight has not been translated to FL scenarios.

3. Preliminary

3.1. Federated Learning

We assume that the server trains a model w with K clients in a FL framework. All participating entities are computationally bounded. Each client $i \in [K]$ holds a local dataset $\mathcal{D}_i$. The optimization objective is to minimize the global empirical risk function:

$$\min_w F(w) = \mathbb{E}_{\xi \sim \mathcal{P}} [\ell(\xi, w)], \quad (1)$$

where $\mathcal{P}$ represents the underlying data distribution, ξ denotes a data sample, and $\ell(\cdot)$ is the task-specific loss function. During federated training, the server aggregates local gradients g_i^t from participating clients to update the global model at round t :

$$w^t = w^{t-1} - \eta^t \sum_{i=1}^K \alpha_i^t g_i^t, \quad (2)$$

where $\alpha_i^t = \frac{|\mathcal{D}_i|}{|\mathcal{D}|}$ is the weight of client i based on the size of its local dataset, and η^t is the learning rate.

3.2. Task Arithmetic

Task Arithmetic (TA) [30] is recognized as a competitive baseline for model merging through the utilization of task vectors, defined as the parametric differences between pre-trained and fine-tuned models. This process is visually summarized in Figure 1. Formally, for task k the task vector is expressed as: $\Delta_k = \theta_k - \theta_{\text{pre}}$.

These vectors encode task-specific knowledge. The merged model θ_{TA} is computed by combining the pre-trained parameters with a scaled aggregation of task vectors:

$$\theta_{\text{TA}} = \theta_{\text{pre}} + \lambda \sum_k \Delta_k, \quad (3)$$

where $\lambda > 0$ is a hyperparameter controlling the magnitude of task-specific adjustments. Note that θ_{pre} captures generalized features while $\sum \Delta_k$ integrates task-specific adaptations. The innovativeness of this method is demonstrated in two crucial dimensions. One is that the parameter difference technique eradicates the reliance on accessing the original data, making it possible to achieve knowledge integration merely by manipulating the model weights. The other is that the integrated model sustains robust performance in multi task scenarios. These features render it highly compatible with the innate privacy protection principles of FL. Specifically, global model optimization is accomplished by operating within the parameter space, all the while refraining from any direct exposure to client data.

4. Methodology

4.1. Motivation

The core challenge in FL primarily arises from data heterogeneity. Under Non-IID conditions, divergent local data distributions across clients cause gradient conflicts due to misaligned model optimization directions, as established in prior work [42, 69, 74]. While the traditional FedAvg method aggregates local models via simple weighted averaging, it fails to effectively reconcile these conflicting parameter updates. Consequently, the global model tends to converge towards sharp regions of the loss landscape. Such regions exhibit high sensitivity to parameter perturbations [28, 50], significantly impairing the model's generalization performance. Specifically, when the model resides in a sharp region, common forgetting operations such as parameter modification or task vector subtraction can induce catastrophic forgetting. Catastrophic forgetting is characterized by the removal of the target knowledge accompanied by severe degradation in the performance of retained knowledge.

Recent research indicates that solutions with strong generalization often reside in flat loss landscape regions [25,

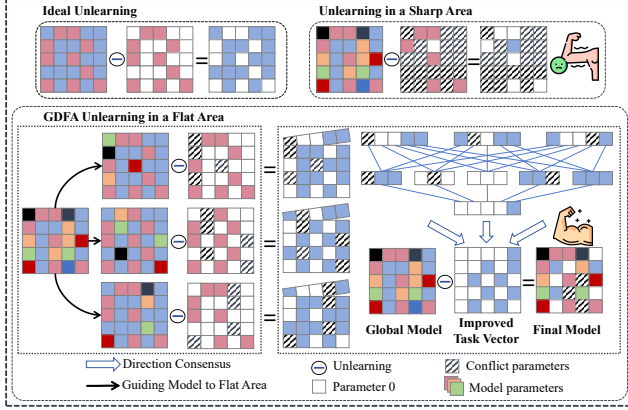


Figure 2. **Overview of our method.** Our GDFEA framework optimizes federated unlearning through geometry-driven directional alignment. Specifically, 1) the global model is migrated to a flat loss region to enhance stability. 2) task vectors generated by relevant clients are merged only when their elements share consistent signs, masking conflicting ones. 3) the refined task vector is subtracted from the original model via reverse aggregation to erase target knowledge.

[28], characterized by robustness to parameter perturbation. Despite Non-IID data causing distribution shifts, the inherent task consistency across clients leads to similarities in their loss landscapes [15, 57], resulting in overlapping flat regions. Convergence to these overlapping areas enables high performance on most clients. To guide the global model towards such regions, clients apply controlled perturbations to their local models during training: generating client-specific signature vectors and adding them to the received global weights. When the global model resides in a flat region, perturbations remain constrained within the basin, and aggregation converges toward the center. When in a sharp region, perturbations enable exploration of neighboring flat minima.

In federated unlearning, the stability of flat regions benefits task vector optimization. Traditional task vector methods directly calculate the parameter differences from a single fine tuned model, but the sensitivity of fine tuning hyperparameters leads to fluctuations in the forgetting effect. We find that training within the same flat region significantly improves task vector sign consistency due to parameter space smoothness. Critically, discarding conflicting parameters during vector generation does not impair unlearning efficacy but enhances directional consistency, as they are empirically redundant and potentially disruptive [49]. By screening and merging components with consistent signs, we accurately capture common features of the knowledge to be forgotten. Reverse aggregating these aligned task vectors efficiently removes target knowledge while minimizing

Algorithm 1 Geometry-Driven Federated Alignment

Require: Global model w , forgetting dataset $\mathcal{D}_f$, hyperparameters ρ, K, λ

Ensure: Unlearned model $w_{\text{unlearned}}$

```

1: // Step 1: Guide model to flat regions
2: for  $k = 1$  to  $K$  do
3:   Generate signature vector  $Z^{(k)} = [z_1^{(k)}, \dots, z_L^{(k)}]$ 
     where  $z_i^{(k)} \in \{-1, 1\}$  randomly
4:   Compute flattened model  $w_{\text{flat}}^{(k)}$  ▷ Eq.(4)
5: end for
6: // Step 2: Generate task vectors
7: for  $k = 1$  to  $K$  do
8:   Fine-tune  $w_{\text{flat}}^{(k)}$  on  $\mathcal{D}_f$  to get  $w_{ft}^{(k)}$ 
9:   Compute task vector  $\tau_k$  ▷ Eq.(5)
10: end for
11: // Step 3: Merge task vectors with directional alignment

```

```

12: Initialize  $\tau_{\text{merged}} = 0$ 
13: for Each parameter index  $j$  do
14:   Find dominant sign  $s_j^*$  among  $\text{sign}(\tau_{k,j})$  for all  $k$ 
15:    $S_j = \{k : \text{sign}(\tau_{k,j}) = s_j^*\}$  ▷ Eq.(6)
16:   if  $S_j \neq \emptyset$  then
17:      $\tau_{\text{merged},j} = \frac{1}{|S_j|} \sum_{k \in S_j} \tau_{k,j}$ 
18:   else
19:      $\tau_{\text{merged},j} = 0$ 
20:   end if
21: end for
22: // Step 4: Knowledge negation
23:  $w_{\text{unlearned}} = w - \lambda \tau_{\text{merged}}$  ▷ Eq.(7)
24: return  $w_{\text{unlearned}}$ 

```

impact on retained knowledge. An overview of this framework is illustrated in Figure 2, which visually summarizes the geometry-driven directional alignment process.

4.2. Guiding Models to Flat Regions

Consider a global model w with L layers, where layer_i denotes the weight tensor of the i -th layer. For each of K flattened models, generate a signature vector $Z^{(k)} = [z_1^{(k)}, z_2^{(k)}, \dots, z_L^{(k)}]$ where $z_i^{(k)} \in \{-1, 1\}$ is sampled randomly. The k -th flattened model $w_{\text{flat}}^{(k)}$ is generated through layer-wise perturbation using gradient information gl_i :

$$w_{\text{flat}}^{(k)} = \{\text{layer}_i + \rho \cdot gl_i \cdot z_i^{(k)}\}_{i=1}^L, \quad (4)$$

where ρ controls the perturbation radius. This radius determines the distance between flattened models and the global model: larger ρ values increase this distance, which can guide w toward flatter regions in the loss landscape. However, excessively large ρ may prevent convergence, as sufficiently flat regions spanning the entire neighborhood might

not exist. We set a constraint $\sum_{k=1}^K z_i^{(k)} = 0$ ensures averaging consistency: $\frac{1}{K} \sum_{k=1}^K w_{flat}^{(k)} = w$.

The generation scheme guarantees that K flattened models are uniformly distributed in the ρ -neighborhood of w .

4.3. Task Vector Edit for Unlearning

4.3.1. Multi-Model Task Vector Generation

Given K flattened models $\{w_{flat}^{(1)}, w_{flat}^{(2)}, \dots, w_{flat}^{(K)}\}$ and forget dataset $\mathcal{D}_f$, generate K fine-tuned models $\{w_{ft}^{(1)}, w_{ft}^{(2)}, \dots, w_{ft}^{(K)}\}$. Compute corresponding task vectors:

$$\tau_k = w_{ft}^{(k)} - w_{flat}^{(k)} \quad \forall k \in \{1, 2, \dots, K\}. \quad (5)$$

4.3.2. Consensus-Based Vector Merging

Define the merged task vector τ_{merged} through element-wise sign consensus:

$$\tau_{merged,j} = \begin{cases} \frac{1}{|S_j|} \sum_{k \in S_j} \tau_{k,j} & \text{if } S_j = \{k : \text{sign}(\tau_{k,j}) = s_j^*\} \\ 0 & \text{otherwise,} \end{cases} \quad (6)$$

where s_j^* is the dominant sign for parameter j across all vectors, and S_j is the set of models agreeing with s_j^* . This process selectively retains components with consistent directional changes.

4.3.3. Knowledge Negation

Obtain the unlearned model through scaled negation:

$$w_{unlearned} = w - \lambda \tau_{merged}, \quad (7)$$

where λ is a scaling hyperparameter (typically $\lambda = 1$). The algorithm of GDFA is shown in Algorithm 1

4.4. Flat Minima for Robustness to Data Heterogeneity in Federated Unlearning

Consider a FL scenario with K clients, each with a local data distribution $\mathcal{P}_i$. The global data distribution is $\mathcal{P} = \frac{1}{K} \sum_{i=1}^K \mathcal{P}_i$. Let $\mathcal{L}(\theta)$ be the global loss function, and $\mathcal{L}_i(\theta)$ be the local loss for client i . The global loss is minimized as $\mathcal{L}(\theta) = \frac{1}{K} \sum_{i=1}^K \mathcal{L}_i(\theta)$. Data heterogeneity is quantified by the concentrability coefficient C , which measures the maximum distribution mismatch between any client and the global distribution.

We now present a theorem that bounds the performance gap under data heterogeneity when the model resides in a flat minimum.

Theorem 1 *Let θ^* be the parameters that minimize the global loss $\mathcal{L}(\theta)$, and let θ_{flat} be parameters in a flat minimum with small first-order sharpness $R_\rho^{(1)}(\theta)$. For any client data distribution shift, the difference in performance*

between the optimal and flat-minima models is bounded. Specifically, with probability at least $1 - \delta$:

$$|\mathcal{L}(\theta^*) - \mathcal{L}(\theta_{flat})| \leq \frac{\gamma(1+C)}{(1-\gamma)^2} \left[\hat{\mathcal{L}}(\theta; \pi) + R_\rho^{(1)}(\theta; \pi) + \sqrt{\frac{M_{loss}}{n}} + \Omega(d, n, \rho, \delta) \right], \quad (8)$$

where $C = \max_i \frac{\mathcal{P}_i}{\mathcal{P}}$ is the concentrability coefficient, $\hat{\mathcal{L}}(\theta; \pi)$ is the empirical model error under policy π , $R_\rho^{(1)}(\theta; \pi)$ is the first-order sharpness, M_{loss} is the bound on per-sample loss, n is the number of samples, d is the number of parameters, and $\Omega(d, n, \rho, \delta)$ is a model complexity term.

Remark 1 *This theorem indicates that by converging to flat minima, we tighten the bound on the performance gap caused by data heterogeneity. In federated unlearning, this implies that unlearning operations are more robust, preserving retained knowledge while effectively removing target client data. The flat region ensures that parameter perturbations during unlearning do not lead to catastrophic forgetting, which is crucial in Non-IID settings.*

The detail of proof is provided in appendix.

5. Experiments

5.1. Datasets and Models

We evaluate our algorithm on four commonly applied benchmark datasets: **MNIST** [36], **FMNIST** [66] and **CI-FAR10/100** [35]. We employ 4-layer CNN architecture [44] which consists of two convolutional layers and two fully connected layers for MNIST and Fashion-MNIST, ResNet-18 [24] for CIFAR10 and ResNet-34 [24] for CI-FAR100. To simulate heterogeneous client data distributions in FL, we consider the following two scenarios:

- **IID**: The dataset is uniformly and randomly allocated across all participating clients.
- **Dir(α)**: Following [26], we partition data among k clients using a Dirichlet distribution with concentration parameter α . Values of $\alpha < 1$ induce high data heterogeneity. Conversely, as $\alpha \rightarrow \infty$, the distribution approaches statistical IID.

5.2. Hyperparameters

We evaluated the proposed algorithm and its baseline model. The experimental configuration is as follows: A total of 100 clients were used, and the client sampling rate per round of communication was 0.1. The optimizer adopted

Table 1. Comparison of federated unlearning performance for GDFA and baseline methods in a Non-IID scenario involving 100 clients. An unlearning request is initiated by the first client. Local client datasets are partitioned following a Dirichlet distribution with concentration parameter $\alpha = 0.5$.

Metrics	FedAvg	Retrain	Eraser	Recovery	MoDe	SGA	PGD	OSD	Editor	SIFU	Ours
MNIST-CNN											
RA $\uparrow$	94.60	94.15	94.41	95.73	92.30	94.52	95.15	94.32	94.48	94.22	95.89
FA $\downarrow$	96.92	0.11	1.15	0.12	1.88	1.70	1.31	2.18	2.95	3.92	1.53
JSD $\downarrow$	0.0089	-	0.0023	0.0057	0.0095	0.0018	0.0072	0.0031	0.0028	0.0035	0.0012
MIA	92.52	50.28	50.15	50.68	50.42	50.10	51.25	52.55	55.12	55.90	50.00
FMNIST-LeNet											
RA $\uparrow$	89.21	89.82	84.85	82.12	81.95	77.41	74.85	85.10	73.42	85.61	89.95
FA $\downarrow$	98.35	1.40	1.75	2.00	3.71	2.38	4.11	2.42	3.92	1.80	1.55
JSD $\downarrow$	0.0071	-	0.0045	0.0035	0.0015	0.0037	0.0065	0.0046	0.0028	0.0035	0.0023
MIA	98.80	50.00	50.48	50.53	46.20	47.35	49.53	50.72	57.49	54.55	50.12
CIFAR10-ResNet18											
RA $\uparrow$	68.32	68.88	68.73	64.08	60.33	68.15	58.42	62.32	65.18	55.17	69.33
FA $\downarrow$	97.58	0.51	0.68	0.75	3.79	1.78	2.39	1.31	2.54	1.06	1.57
JSD $\downarrow$	0.0045	-	0.0005	0.0005	0.0095	0.0033	0.0015	0.0041	0.0029	0.0035	0.0010
MIA	96.07	50.67	50.79	50.47	49.44	50.85	50.50	50.78	49.76	50.59	49.85
CIFAR100-ResNet34											
RA $\uparrow$	42.53	47.82	38.85	40.12	37.95	29.41	39.85	39.10	38.42	39.61	46.08
FA $\downarrow$	98.77	0.34	1.45	2.00	3.70	3.41	2.77	2.35	3.00	1.80	1.05
JSD $\downarrow$	0.0042	-	0.0045	0.0094	0.0015	0.0072	0.0065	0.0048	0.0020	0.0055	0.0014
MIA	98.80	50.55	50.61	50.51	46.20	47.70	49.83	52.95	51.68	54.41	49.48

Table 2. Comparison of federated unlearning performance for GDFA and baseline methods in an IID scenario involving 100 clients. An unlearning request is initiated by the first client.

Metrics	FedAvg	Retrain	Eraser	Recovery	MoDe	SGA	PGD	OSD	Editor	SIFU	Ours
CIFAR10-ResNet18											
RA $\uparrow$	77.45	77.60	76.23	77.21	77.61	77.42	76.89	77.52	75.96	76.69	77.98
FA $\downarrow$	98.77	2.32	2.38	4.34	5.74	4.70	3.31	4.18	3.95	3.92	3.04
JSD $\downarrow$	0.0052	-	0.0031	0.0052	0.0080	0.0043	0.0058	0.0054	0.0071	0.0046	0.0026
MIA	98.51	50.24	50.18	50.72	50.88	50.23	51.70	51.35	52.10	50.83	50.03

Stochastic Gradient Descent (SGD) with momentum. The learning rate was set to 0.1, and the momentum parameter was taken as 0.9. The learning rate decay coefficient per round was 0.99. Referring to the method in [76], all clients executed SGD on the local dataset with a fixed batch size of 64. To ensure that the local model converged sufficiently in each round of training, the number of local training rounds E was set to 5. The maximum number of rounds in the forgetting phase was capped at 100 rounds, and the upper limit of the total number of communication rounds for the entire training process was 200 rounds.

5.3. Comparison Metrics

Computational efficiency is quantified by **Time consumption** (Time $\downarrow$), measured in seconds from unlearning initiation to acquisition of the updated global model, where minimization reflects reduced operational overhead. Model utility preservation is assessed through **test data accuracy** (RA $\uparrow$), evaluating task performance retention on unaltered data, with maximization indicating minimal degradation of core functionality. Unlearning efficacy verification employs three complementary approaches:

- **Unlearned data accuracy** (FA $\downarrow$) detects residual influence via backdoor triggered predictions.

Table 3. The Server computation time (s) for clients and the server on MNIST, FMNIST, CIFAR10, and CIFAR100 in Dir(0.5).

	FedAvg	Retrain	Eraser	Recovery	MoDe	SGA	PGD	OSD	Editor	SIFU	Ours
MNIST	647.92	223.96	188.90	15.08	24.79	14.60	14.31	18.58	14.39	15.28	11.28
FMNIST	1085.43	542.72	443.09	33.15	39.28	32.60	29.37	37.09	25.54	30.94	17.41
CIFAR10	2322.76	1161.33	884.55	72.08	62.41	58.24	59.61	64.55	52.21	53.42	31.92
CIFAR100	3246.27	1623.14	949.26	103.30	98.54	79.81	84.67	89.10	74.63	94.29	67.88

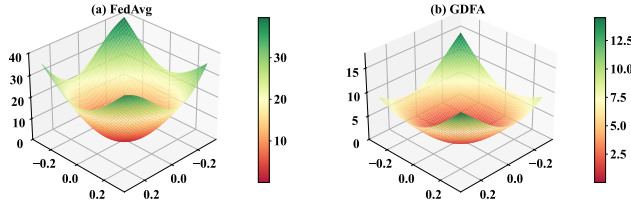


Figure 3. Comparative analysis of loss landscapes between FedAvg and GDFA for ResNet-18 trained on CIFAR10 under IID data distribution

- **Jensen-Shannon Divergence (JSD)** [45] measures (JSD $\downarrow$) the divergence between the output probability distributions of the unlearned model and the retrained model [12, 46]. This symmetric, information-theoretic metric quantifies the difference between the two distributions.
- **Membership Inference Attack (MIA)** [27, 56] resilience is quantified by the inference accuracy of the attack model. Convergence to random guessing accuracy (50%) confirms privacy protection.

5.4. Baselines

We perform a comprehensive comparative analysis between our proposed algorithm and ten federated unlearning baselines. The evaluation includes **Retraining** as the gold standard involving complete model retraining from scratch, **FedAvg** [44], **Eraser** [39], **Recovery** [72], **MoDe** [75], **SGA** [65], **PGD** [22], **OSD** [48], **Editor** [70] and **SIFU** [63].

The experimental environment utilized PyTorch-1.8 on a Windows 11 server featuring an Intel i7 14700KF CPU (20 cores), 32GB RAM, and an NVIDIA 4090D GPU. As the primary evaluation metric, we calculated the mean classification accuracy and variance across the final five experimental iterations.

5.5. Loss Landscape Flatness Verification

To validate the hypothesis that GDFA converges to a flat minimum, we trained ResNet-18 models on CIFAR-10 and visualized the resultant loss landscape. As shown in Figure 3 with landscapes centered at their minima, our solution resides in a distinctly flatter basin. This observation confirms

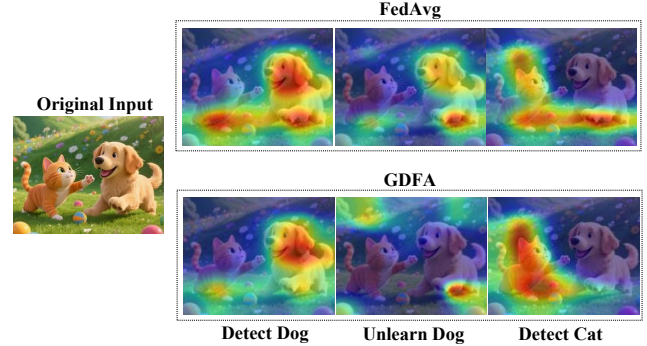


Figure 4. Attention Map Visualization in ResNet-18 for Class Unlearning Tasks.

our method steers federated optimization towards flatter regions, achieving significantly flatter minima than FedAvg.

5.6. The Effectiveness of Unlearning

Tables 1 and 2 demonstrate GDFA’s superiority in unlearning metrics. It reduces target data accuracy to near-random levels, minimizes residual influence with low JSD values, and enhances privacy resilience by lowering MIA success rates to approximately 50%. This stems from directional task vector alignment isolating knowledge attributes and flat minima ensuring stability. We also employ Grad-CAM [53] to visualize attention maps of the forgotten class in ResNet-18. As shown in Figure 4, these maps highlight image regions critical for model predictions. Attention maps for forgotten classes show significant shifts from prediction-critical regions to background areas, progressively impairing class discrimination. This visual evidence validates GDFA’s mechanism for class-level unlearning while maintaining retained task accuracy.

5.7. The Efficiency of Unlearning

Table 3 shows the time efficiency of GDFA compared to various federated unlearning baselines. GDFA achieves competitive time costs across all evaluated datasets, significantly outperforming retraining-based methods while maintaining reasonable overhead relative to lightweight approximations. This efficiency stems from GDFA’s geometry-

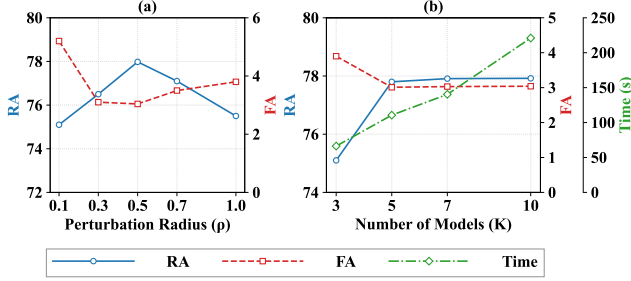


Figure 5. Impact of hyperparameters ρ and K on CIFAR10 under the Dir(1) heterogeneous setting.

driven approach, which leverages flat loss regions to minimize computational demands. Specifically, the method avoids costly retraining or Hessian matrix computations by guiding the global model to stable, flat minima through controlled perturbations. This geometric migration reduces sensitivity to parameter modifications, streamlining the unlearning process. Although the initial model flattening introduces some latency, it eliminates the need for iterative calibrations required by conflict-prone methods like OSD or Recovery. Consequently, GDFA achieves a substantial speedup over full retraining without compromising core functionality. The scalability and practicality of GDFA are evident in its balanced trade-off between time savings and retention of model utility, making it suitable for large-scale federated deployments where efficiency is paramount.

5.8. Ablation on Perturbation Radius for Flat Region

Figure 5 (a) demonstrates the critical role of perturbation radius ρ in migrating models to flat loss regions. When configured with smaller values, the perturbation proves insufficient to facilitate escape from sharp minima, leaving models vulnerable to instability when subjected to subsequent unlearning operations. These constrained perturbations fail to provide adequate exploration capacity, causing the optimization process to remain trapped in sensitive, high-curvature basins where parameter adjustments disproportionately disrupt retained knowledge. Conversely, the application of excessively large perturbation radii induces detrimental deviations beyond beneficial flat regions, steering models toward suboptimal areas of the loss surface where generalization capability deteriorates substantially. The $\rho = 0.5$ configuration optimally balances exploration and stability, enabling reliable transitions to flat regions and enhancing unlearning robustness.

Table 4. Impact of λ on FMNIST under the IID setting.

λ	0	-0.2	-0.4	-0.6	-0.8	-1
FA ↓	98.21	73.91	48.03	20.89	9.04	1.45

5.9. Perturbed Quantity Impact on Task Vector Quality

Figure 5 (b) shows the impact of the number of perturbed models on task vector quality. Increasing K initially improves directional consistency, facilitating more precise knowledge isolation. However, performance saturates with additional models, yielding diminishing returns in vector alignment. Moreover, higher K values incur significant computational overhead, substantially increasing the time required for model perturbation and aggregation. Consequently, $K = 5$ achieves an optimal balance, maximizing task vector quality while maintaining computational efficiency and minimizing resource consumption.

5.10. Ablation on λ for Forgetting Accuracy

We conduct an ablation study on the λ and evaluate the Forgetting Accuracy (FA) on the FMNIST. As shown in Table 4, transitioning λ to negative values results in a progressive decline in FA, confirming the effectiveness of the unlearning procedure. At the optimal $\lambda = -1$, FA reaches its minimum value of 1.45, representing the peak unlearning performance observed in our experiments. This outcome demonstrates that scaling the task vector negation successfully isolates and removes the target knowledge while maintaining model stability. The significant reduction in FA observed at $\lambda = -1$ aligns precisely with GDFA’s reverse aggregation mechanism.

More detailed results can be found in appendix.

6. Conclusion

In this work, we propose GDFA, a geometry-driven federated unlearning framework. GDFA efficiently removes designated client data influence without costly retraining while preserving global model performance. GDFA addresses optimization geometry challenges by converging models towards flat minima. This establishes a stable operational environment resilient to parameter perturbations. Crucially, we introduce directional alignment for task vectors generated by relevant clients. This process selectively consolidates only consistent optimization trajectories to precisely isolate target knowledge. Simultaneous reverse aggregation then erases these aligned vectors while maximizing retention of other knowledge. Comprehensive evaluations across diverse datasets and architectures demonstrate that GDFA significantly outperforms SOTA methods in unlearning efficacy.

Acknowledgments

This work was funded by the National Natural Science Foundation of China (62106217), the Yunnan Provincial Department of Science and Technology Basic Research Program (202201AU070112), and the Young Talent Program of Yunnan Province (C619300A116).

References

- [1] Thomas Baumhauer, Pascal Schöttle, and Matthias Zepelzauer. Machine unlearning: Linear filtration for logit-based classifiers. *Machine Learning*, 111(9):3203–3226, 2022. 2
- [2] Lucas Bourtole, Varun Chandrasekaran, Christopher A Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In *2021 IEEE symposium on security and privacy (SP)*, pages 141–159. IEEE, 2021. 1, 2
- [3] Jonathan Brophy and Daniel Lowd. Machine unlearning for random forests. In *International Conference on Machine Learning*, pages 1092–1104. PMLR, 2021. 2
- [4] Xiaoyu Cao, Jinyuan Jia, Zaixi Zhang, and Neil Zhenqiang Gong. Fedrecover: Recovering from poisoning attacks in federated learning using historical information. In *2023 IEEE Symposium on Security and Privacy (SP)*, pages 1366–1383. IEEE, 2023. 2
- [5] Yinzhi Cao and Junfeng Yang. Towards making systems forget with machine unlearning. In *2015 IEEE symposium on security and privacy*, pages 463–480. IEEE, 2015. 1, 2
- [6] Sungmin Cha, Sungjun Cho, Dasol Hwang, Honglak Lee, Taesup Moon, and Moontae Lee. Learning to unlearn: Instance-wise unlearning for pre-trained classifiers. In *Proceedings of the AAAI conference on artificial intelligence*, pages 11186–11194, 2024. 2
- [7] Tianshi Che, Yang Zhou, Zijie Zhang, Lingjuan Lyu, Ji Liu, Da Yan, Dejing Dou, and Jun Huan. Fast federated machine unlearning with nonlinear functional theory. In *International conference on machine learning*, pages 4241–4268. PMLR, 2023. 2
- [8] Chong Chen, Fei Sun, Min Zhang, and Bolin Ding. Recommendation unlearning. In *Proceedings of the ACM web conference 2022*, pages 2768–2777, 2022. 2
- [9] Huiming Chen, Huandong Wang, Qingyue Long, Depeng Jin, and Yong Li. Advancements in federated learning: Models, methods, and privacy. *ACM Computing Surveys*, 57(2): 1–39, 2024. 1
- [10] Jingxue Chen, Hang Yan, Zhiyuan Liu, Min Zhang, Hu Xiong, and Shui Yu. When federated learning meets privacy-preserving computation. *ACM Computing Surveys*, 56(12): 1–36, 2024. 1
- [11] Eli Chien, Haoyu Wang, Ziang Chen, and Pan Li. Langevin unlearning: A new perspective of noisy gradient descent for machine unlearning. *Advances in neural information processing systems*, 37:79666–79703, 2024. 1
- [12] Vikram S Chundawat, Ayush K Tarun, Murari Mandal, and Mohan Kankanhalli. Can bad teaching induce forgetting? unlearning in deep networks using an incompetent teacher. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7210–7217, 2023. 7
- [13] General Data. General data protection regulation (gdpr). *Intersoft Consulting*, Accessed in October, 24(1), 2018. 1
- [14] Zhipeng Deng, Luyang Luo, and Hao Chen. Enable the right to be forgotten with federated client unlearning in medical imaging. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 240–250. Springer, 2024. 2
- [15] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. *arXiv preprint arXiv:2010.01412*, 2020. 2, 4
- [16] Antonio Ginart, Melody Guan, Gregory Valiant, and James Y Zou. Making ai forget you: Data deletion in machine learning. *Advances in neural information processing systems*, 32, 2019. 1
- [17] Aditya Golatkar, Alessandro Achille, and Stefano Soatto. Eternal sunshine of the spotless net: Selective forgetting in deep networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9304–9312, 2020. 2
- [18] 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. 2
- [19] Badra Souhila Guendouzi, Samir Ouchani, Hiba EL Assaad, and Madeleine EL Zaher. A systematic review of federated learning: Challenges, aggregation methods, and development tools. *Journal of Network and Computer Applications*, 220:103714, 2023. 1
- [20] Saaketh Koundinya Gundavarapu, Shreya Agarwal, Arushi Arora, and Chandana Thimmalapura Jagadeeshaiah. Machine unlearning in large language models. *arXiv preprint arXiv:2405.15152*, 2024. 2
- [21] Chuan Guo, Tom Goldstein, Awni Hannun, and Laurens Van Der Maaten. Certified data removal from machine learning models. *arXiv preprint arXiv:1911.03030*, 2019. 2
- [22] Anisa Halimi, Swanand Kadhe, Amrith Rawat, and Nathalie Baracaldo. Federated unlearning: How to efficiently erase a client in fl? *arXiv preprint arXiv:2207.05521*, 2022. 2, 7
- [23] Elizabeth Liz Harding, Jarno J Vanto, Reece Clark, L Hannah Ji, and Sara C Ainsworth. Understanding the scope and impact of the california consumer privacy act of 2018. *Journal of Data Protection & Privacy*, 2(3):234–253, 2019. 1
- [24] 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. 5
- [25] Sepp Hochreiter and Jürgen Schmidhuber. Flat minima. *Neural computation*, 9(1):1–42, 1997. 3
- [26] Tzu-Ming Harry Hsu, Hang Qi, and Matthew Brown. Measuring the effects of non-identical data distribution for federated visual classification. *arXiv preprint arXiv:1909.06335*, 2019. 5

- [27] Hongsheng Hu, Zoran Salcic, Lichao Sun, Gillian Dobbie, Philip S Yu, and Xuyun Zhang. Membership inference attacks on machine learning: A survey. *ACM Computing Surveys (CSUR)*, 54(11s):1–37, 2022. 7
- [28] Ming Hu, Yue Cao, Anran Li, Zhiming Li, Chengwei Liu, Tianlin Li, Mingsong Chen, and Yang Liu. Fedmut: Generalized federated learning via stochastic mutation. In *Proceedings of the AAAI conference on artificial intelligence*, pages 12528–12537, 2024. 3, 4
- [29] Hanxun Huang, Xingjun Ma, Sarah Monazam Erfani, James Bailey, and Yisen Wang. Unlearnable examples: Making personal data unexploitable. *arXiv preprint arXiv:2101.04898*, 2021. 1
- [30] Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. *arXiv preprint arXiv:2212.04089*, 2022. 3
- [31] Jinghan Jia, Jiancheng Liu, Parikshit Ram, Yuguang Yao, Gaowen Liu, Yang Liu, Pranay Sharma, and Sijia Liu. Model sparsity can simplify machine unlearning. *Advances in Neural Information Processing Systems*, 36:51584–51605, 2023. 1
- [32] Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *Foundations and trends® in machine learning*, 14(1–2): 1–210, 2021. 1
- [33] Dohyoung Kim, Hyekyung Woo, and Youngho Lee. Addressing bias and fairness using fair federated learning: A synthetic review. *Electronics*, 13(23):4664, 2024. 1
- [34] Hyoseo Kim, Dongyoon Han, and Junsuk Choe. Negmerge: Consensual weight negation for strong machine unlearning. *arXiv preprint arXiv:2410.05583*, 2024. 3
- [35] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 5
- [36] Yann LeCun. The mnist database of handwritten digits. <http://yann.lecun.com/exdb/mnist/>, 1998. 5
- [37] Meng Li and Haochen Sui. Causal recommendation via machine unlearning with a few unbiased data. In *AAAI 2025 Workshop on Artificial Intelligence with Causal Techniques*, 2025. 2
- [38] Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence of fedavg on non-iid data. *arXiv preprint arXiv:1907.02189*, 2019. 1
- [39] Gaoyang Liu, Xiaoqiang Ma, Yang Yang, Chen Wang, and Jiangchuan Liu. Federaser: Enabling efficient client-level data removal from federated learning models. In *2021 IEEE/ACM 29th International Symposium on Quality of Service (IWQoS)*, pages 1–10. IEEE, 2021. 2, 7
- [40] Sijia Liu, Yuanshun Yao, Jinghan Jia, Stephen Casper, Nathalie Baracaldo, Peter Hase, Yuguang Yao, Chris Yuhao Liu, Xiaojun Xu, Hang Li, et al. Rethinking machine unlearning for large language models. *Nature Machine Intelligence*, pages 1–14, 2025. 2
- [41] 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. 1, 2
- [42] Mi Luo, Fei Chen, Dapeng Hu, Yifan Zhang, Jian Liang, and Jiashi Feng. No fear of heterogeneity: Classifier calibration for federated learning with non-iid data. *Advances in Neural Information Processing Systems*, 34:5972–5984, 2021. 3
- [43] Xiaodong Ma, Jia Zhu, Zhihao Lin, Shanxuan Chen, and Yangjie Qin. A state-of-the-art survey on solving non-iid data in federated learning. *Future Generation Computer Systems*, 135:244–258, 2022. 1
- [44] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017. 1, 5, 7
- [45] María Luisa Menéndez, Julio Angel Pardo, Leandro Pardo, and María del C Pardo. The jensen-shannon divergence. *Journal of the Franklin Institute*, 334(2):307–318, 1997. 7
- [46] Thanh Tam Nguyen, Thanh Trung Huynh, Zhao Ren, Phi Le Nguyen, Alan Wee-Chung Liew, Hongzhi Yin, and Quoc Viet Hung Nguyen. A survey of machine unlearning. *arXiv preprint arXiv:2209.02299*, 2022. 7
- [47] Thanh Tam Nguyen, Thanh Trung Huynh, Zhao Ren, Phi Le Nguyen, Alan Wee-Chung Liew, Hongzhi Yin, and Quoc Viet Hung Nguyen. A survey of machine unlearning. *ACM Transactions on Intelligent Systems and Technology*, 16(5): 1–46, 2025. 2
- [48] Zibin Pan, Zhichao Wang, Chi Li, Kaiyan Zheng, Boqi Wang, Xiaoying Tang, and Junhua Zhao. Federated unlearning with gradient descent and conflict mitigation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 19804–19812, 2025. 7
- [49] Biqing Qi, Fangyuan Li, Zhen Wang, Junqi Gao, Dong Li, Peng Ye, and Bowen Zhou. Less is more: Efficient model merging with binary task switch. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 15265–15274, 2025. 4
- [50] Zhuang Qi, Lei Meng, Zhaochuan Li, Han Hu, and Xiangxu Meng. Cross-silo feature space alignment for federated learning on clients with imbalanced data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 19986–19994, 2025. 3
- [51] Sebastian Schelter, Stefan Grafberger, and Ted Dunning. Hedgecut: Maintaining randomised trees for low-latency machine unlearning. In *Proceedings of the 2021 International Conference on Management of Data*, pages 1545–1557, 2021. 2
- [52] Ayush Sekhari, Jayadev Acharya, Gautam Kamath, and Ananda Theertha Suresh. Remember what you want to forget: Algorithms for machine unlearning. *Advances in Neural Information Processing Systems*, 34:18075–18086, 2021. 1
- [53] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 7

- [54] Changlong Shi, He Zhao, Bingjie Zhang, Mingyuan Zhou, Dandan Guo, and Yi Chang. Fedawa: Adaptive optimization of aggregation weights in federated learning using client vectors. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 30651–30660, 2025. [3](#)
- [55] Takashi Shibata, Go Irie, Daiki Ikami, and Yu Mitsuzumi. Learning with selective forgetting. In *IJCAI*, page 4, 2021. [2](#)
- [56] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pages 3–18. IEEE, 2017. [7](#)
- [57] David Stutz, Matthias Hein, and Bernt Schiele. Relating adversarially robust generalization to flat minima. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7807–7817, 2021. [2](#), [4](#)
- [58] Ningxin Su and Baochun Li. Asynchronous federated unlearning. In *IEEE INFOCOM 2023-IEEE conference on computer communications*, pages 1–10. IEEE, 2023. [2](#)
- [59] Ayush K Tarun, Vikram S Chundawat, Murari Mandal, and Mohan Kankanhalli. Fast yet effective machine unlearning. *IEEE Transactions on Neural Networks and Learning Systems*, 35(9):13046–13055, 2023. [1](#), [2](#)
- [60] IT Governance Privacy Team et al. *EU general data protection regulation (GDPR): an implementation and compliance guide*. Packt Publishing Ltd, 2025. [1](#)
- [61] Joost Verbraeken, Matthijs Wolting, Jonathan Katzy, Jeroen Kloppenburg, Tim Verbelen, and Jan S Rellermeyer. A survey on distributed machine learning. *Acm computing surveys (csur)*, 53(2):1–33, 2020. [1](#)
- [62] 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. [1](#), [2](#)
- [63] Pengfei Wang, Wei Song, Heng Qi, Changjun Zhou, Fuliang Li, Yong Wang, Peng Sun, and Qiang Zhang. Server-initiated federated unlearning to eliminate impacts of low-quality data. *IEEE Transactions on Services Computing*, 17(3):1196–1211, 2024. [7](#)
- [64] Alexander Warnecke, Lukas Pirch, Christian Wressnegger, and Konrad Rieck. Machine unlearning of features and labels. *arXiv preprint arXiv:2108.11577*, 2021. [2](#)
- [65] Leijie Wu, Song Guo, Junxiao Wang, Zicong Hong, Jie Zhang, and Yaohong Ding. Federated unlearning: Guarantee the right of clients to forget. *IEEE Network*, 36(5):129–135, 2022. [2](#), [7](#)
- [66] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017. [5](#)
- [67] Xinda Xing, Qiugang Zhan, Xiurui Xie, Yuning Yang, Qiang Wang, and Guisong Liu. Flexible sharpness-aware personalized federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 21707–21715, 2025. [1](#)
- [68] Abbas Yazdinejad, Ali Dehghantanha, Hadis Karimipour, Gautam Srivastava, and Reza M Parizi. A robust privacy-preserving federated learning model against model poisoning attacks. *IEEE Transactions on Information Forensics and Security*, 19:6693–6708, 2024. [1](#)
- [69] Mang Ye, Xiuwen Fang, Bo Du, Pong C Yuen, and Dacheng Tao. Heterogeneous federated learning: State-of-the-art and research challenges. *ACM Computing Surveys*, 56(3):1–44, 2023. [3](#)
- [70] Xiaohan Yuan, Jiqiang Liu, Bin Wang, Guorong Chen, Xianguai Xu, Junyong Wang, Tao Li, and Wei Wang. Fededitor: Efficient and effective federated unlearning in cooperative intelligent transportation systems. *IEEE Transactions on Information Forensics and Security*, 2025. [7](#)
- [71] Dawen Zhang, Pamela Finckenberg-Broman, Thong Hoang, Shidong Pan, Zhenchang Xing, Mark Staples, and Xiwei Xu. Right to be forgotten in the era of large language models: Implications, challenges, and solutions. *AI and Ethics*, pages 1–10, 2024. [1](#)
- [72] Lefeng Zhang, Tianqing Zhu, Haibin Zhang, Ping Xiong, and Wanlei Zhou. Fedrecovery: Differentially private machine unlearning for federated learning frameworks. *IEEE Transactions on Information Forensics and Security*, 18:4732–4746, 2023. [3](#), [7](#)
- [73] Peng-Fei Zhang, Guangdong Bai, Zi Huang, and Xin-Shun Xu. Machine unlearning for image retrieval: A generative scrubbing approach. In *Proceedings of the 30th ACM international conference on multimedia*, pages 237–245, 2022. [2](#)
- [74] Yue Zhao, Meng Li, Liangzhen Lai, Naveen Suda, Damon Civin, and Vikas Chandra. Federated learning with non-iid data. *arXiv preprint arXiv:1806.00582*, 2018. [3](#)
- [75] Yian Zhao, Pengfei Wang, Heng Qi, Jianguo Huang, Zongzheng Wei, and Qiang Zhang. Federated unlearning with momentum degradation. *IEEE Internet of Things Journal*, 11(5):8860–8870, 2023. [2](#), [7](#)
- [76] Zirui Zhou, Lingyang Chu, Changxin Liu, Lanjun Wang, Jian Pei, and Yong Zhang. Towards fair federated learning. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 4100–4101, 2021. [6](#)
- [77] Hangyu Zhu, Jinjin Xu, Shiqing Liu, and Yaochu Jin. Federated learning on non-iid data: A survey. *Neurocomputing*, 465:371–390, 2021. [1](#)