MACHINE UNLEARNING VIA SIGNAL-GUIDED OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Current machine unlearning methods predominantly rely on global, coarse-grained intervention strategies. They lack precise guidance signals to guide the unlearning process and fail to provide differentiated guidance across different unlearning tasks. Due to the varying difficulty of forgetting across different samples, this uniform strategy leads to two problems: when *Boundary* samples have not yet been forgotten, the entire model remains in an under-unlearned state, leaving residual information that can be exploited by privacy attacks; when *Normal* samples have already been over-forgotten, the entire model falls into an over-unlearned state, thereby harming the model's utility. In this paper, we propose GSUO, a guidance-signal-aware unlearning optimization framework that designs task-specific fine-grained guidance signals to steer the unlearning process and applies to both random-subset and class-wise forgetting tasks. Extensive experiments demonstrate that GSUO outperforms 14 baselines in terms of both unlearning effectiveness and generalization, while achieving high efficiency and significant speedups, validating its effectiveness for reliable machine unlearning[1].

1 INTRODUCTION

With growing awareness of privacy preservation and the legal guarantee of the "right to be forgotten" under regulations such as GDPR (Voigt & Bussche, 2017), deep models are expected to possess the capability to remove the influence of specific samples from training data. Moreover, outdated information or maliciously poisoned samples in the training data further reinforce the need for models to erase such data.

To alleviate the prohibitive overhead of retraining, which involves training from scratch on the retain data and is commonly regarded as the gold standard (Nguyen et al., 2025; Xu et al., 2024), various approximate unlearning methods have consequently been proposed (Chundawat et al., 2023; Kurmanji et al., 2023; Tarun et al., 2024; Chen et al., 2023; 2025; Yu et al., 2026; Fan et al., 2024). For instance, SCRUB (Kurmanji et al., 2023) employs a teacher-student framework with contrastive optimization, and SalUn (Fan et al., 2024) identifies salient weights based on forgetting loss gradients, followed by random-label fine-tuning. For class-wise forgetting, Boundary Shrink alters the labels of forget samples via adversarial perturbations, and Boundary Expanding introduces a "shadow class" neuron to disperse such samples (Chen et al., 2025).

However, approximate unlearning methods typically either perform unified optimization for all forget samples without differentiation (Chundawat et al., 2023; Kurmanji et al., 2023) or solely modify labels directly (Chen et al., 2025; Fan et al., 2024), without explicitly accounting for heterogeneous forgetting difficulty during training. Consequently, these methods tend to adopt global and coarse-grained intervention strategies, lacking precise guidance signals to steer the unlearning process. Such uniform intervention can lead to either over-unlearning, which degrades model utility, or under-unlearning, which leaves residual information vulnerable to privacy attacks.

To this end, we propose the GSUO framework to exert unlearning relying on targeted guidance signals which are designed for diverse samples based on their properties in forgetting tasks, aiming to mitigate both over-unlearning and under-unlearning. Specifically, for random-subset forgetting, where forget samples are uniformly scattered in feature space and share similarities with other retain

[1]The code is available at https://anonymous.4open.science/r/GSUO.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

data, GSUO outlines the guidance signals as diverse target distributions to align according to the categories (i.e. *Normal* and *Boundary*) forget samples belong to; for class-wise forgetting, where forget samples form tight clusters in feature space, GSUO employs intra-class dispersion loss and alignment loss as the guidance signals to prevent performance degradation resulting from disruption of the class boundaries learned by the model. Thus, by explicitly specifying the unlearned state of the forget samples, GSUO can perform tailored optimization with a well-defined objective.

Comprehensive experimental results demonstrate that GSUO consistently achieves the best performance across a wide range of metrics, including fidelity, generalization ability, and forgetting efficacy, while it maintains high efficiency and delivers substantial speed-ups, outperforming all 14 state-of-the-art baselines. On the random-subset forgetting task, GSUO achieves the smallest accuracy gap between test and forget set, reaching 0.01% for ResNet-18, along with the highest test accuracy of 82.11%. It also delivers a substantial speed-up of up to $48\times$ and the lowest AUC score of 0.749 under U-LiRA. For class-wise forgetting, GSUO demonstrates comparable performance, achieving complete unlearning while maintaining the highest training and test accuracies.

2 Related Work

Machine Unlearning was first introduced by (Cao & Yang, 2015) to eliminate the influence of training samples on a learned model efficiently and completely by sharding data and building multiple models, enabling the exact unlearning of specific data partitions. For approximate unlearning, numerous methods have been developed to efficiently erase the influence of data through post-training procedures. Methods leveraging second-order information or influence functions (Izzo et al., 2021; Sekhari et al., 2021; Koh & Liang, 2017; Guo et al., 2020) approximately estimate the contribution of samples to the loss function to revoke their influence. However, most of these approaches are limited to convex models or incur high computational costs when applied to deep networks. For deep neural networks, (Golatkar et al., 2020b;a) proposed more practical approaches based on information projection and quadratic penalty, respectively. (Fan et al., 2024) proposed SalUn, which computes a weight saliency mask to selectively update only the parameters most relevant to the forget data, achieving an excellent balance between preserving performance on the retain set and unlearning effectiveness. (Chundawat et al., 2023; Kurmanji et al., 2023) employ a teacher-student framework for unlearning, with the former using a dual-teacher mechanism and the latter providing directional guidance from a single teacher. In such frameworks, the soft labels generated by the teacher model guide the student model to learn a new output distribution. (He et al., 2025) proposed *Natural Unlearning* to generate new samples by mixing each forget sample with related samples from the retain set, and then fine-tuning the original model on these mixed samples.

Another line of work focuses on class forgetting, which requires the model to remove its knowledge of an entire concept or class. (Tarun et al., 2024) propose UNSIR, which learns "error-maximizing noise" for the target class and combines a "damage-repair" two-step weight update to achieve thorough forgetting of single or multiple classes in a single training pass. (Chen et al., 2023; 2025) shift the focus from the parameter space to the decision space, Boundary Shrink assigns incorrect labels to forget samples via adversarial perturbations and fine-tunes the model, forcing the decision boundary to shrink; while Boundary Expanding introduces a new neuron for a "shadow class" to disperse forget samples into other categories, actively expanding the decision boundary. (Chen et al., 2025) and Yu et al. (2026) focus on achieving class-wise forgetting on long-tailed distribution datasets.

A detailed discussion of related work is provided in Appendix B.

3 Problem Formulation

Notation. Given a training dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N} \subseteq \mathcal{X} \times \mathcal{Y}$ consisting of N samples where $x_i \in \mathcal{X}$ and $y_i \in \mathcal{Y}$ are independently and identically distributed drawn from a joint distribution $\mathcal{P}$ over $\mathcal{X} \times \mathcal{Y}$; $\mathcal{X} \subseteq \mathbb{R}^F$ with dimension as F and label space $\mathcal{Y} = \{1, \ldots, C\}$ with C classes.

Machine Unlearning. Let $\mathcal{D}_f \subset \mathcal{D}$ be a subset of the training dataset composed of N_f samples as the *forget set*, and its complement $\mathcal{D}_r$ be the retain set, that is, $\mathcal{D} = \mathcal{D}_f \cup \mathcal{D}_r$ with $\mathcal{D}_f \cap \mathcal{D}_r = \emptyset$. The original model $f(\cdot; \Theta_o)$ is trained on $\mathcal{D}$ by minimizing the cross-entropy(CE) loss $\mathcal{L}(\hat{y}, y) = -\sum_{k=1}^{C} y_k \log(\hat{y}_k)$. The retrained model, denoted as $f(\cdot; \Theta_*)$, is obtained by training on the retain

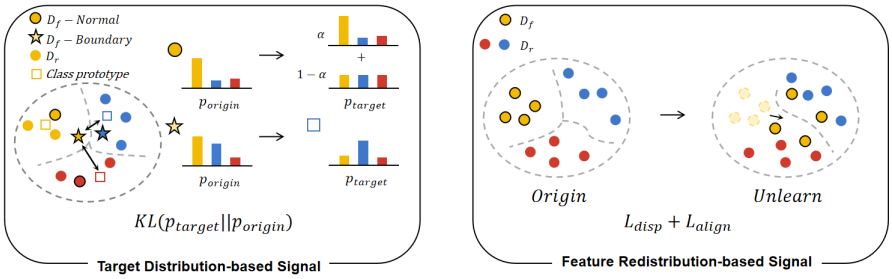


Figure 1: The guidance-signal design in the SGO step of GSUO. **L**: Target Distribution-based Signal for random-subset forgetting, **R**: Feature Redistribution-based Signal for class-wise forgetting.

set $\mathcal{D}_r$ alone. The goal of machine unlearning is to eliminate the influence of $\mathcal{D}_f$ from $f(\cdot;\Theta_o)$ via some unlearning algorithm U, producing updated weights Θ_u such that the unlearned model $f(\cdot;\Theta_u)$ behaves as if it had never seen $\mathcal{D}_f$.

Random-subset and class-wise forgetting According to the composition of the forget set $\mathcal{D}_f$, the unlearning tasks can be categorized into the following two categories.

- *Random-subset forgetting*: $\mathcal{D}_f$ is an arbitrary subset of $\mathcal{D}$, $\mathcal{D}_f \subset \mathcal{D}$ with $|\mathcal{D}_f| = N_f \ll N$, and the samples to be forgotten may come from multiple classes of $\mathcal{Y}$. Since these samples are drawn from the same underlying data distribution as the retained and test samples, a successfully unlearned model should no longer exhibit a training-set advantage on $\mathcal{D}_f$. Accordingly, its performance on $\mathcal{D}_f$ is expected to approach that on unseen test data $\mathcal{D}_t$.
- *Class-wise forgetting*: $\mathcal{D}_f$ contains all training samples of a subset of target classes $\mathcal{Y}_f \subset \mathcal{Y}$: $\mathcal{D}_f = \{(x_i, y_i) \in \mathcal{D} \mid y_i \in \mathcal{Y}_f\}$. $|\mathcal{Y}_f| = 1$ reduces to single-class forgetting. In this setting, the objective is to suppress the model's predictive capability on the forgotten classes while preserving its performance on the retained classes.

4 Proposed Method: GSUO

In this section, we propose the Guidance-Signal-aware Unlearning Optimization (**GSUO**), a unified framework for performing sample unlearning based on explicitly designed guidance signals, which can be applied to both random-subset and class-wise forgetting tasks. Specifically, GSUO consists of the following two successive optimization steps:

- **S1.** *Signal-Guided Optimization (SGO)*: leveraging specifically designed guidance signals for diverse samples to steer the unlearning process.
- **S2.** *Compensatory Optimization (CO)*: employing a general CE loss over the retain set and/or the forget set to enhance model utility via gradient descent/ascent (GD/GA).

For the guidance signal in SGO, based on the sample behavior in the ideal unlearned model, we design *Target Distribution-based Signal* for random-subset forgetting task and *Feature Redistribution-based Signal* for class-wise forgetting task, which will be elaborated in detail in the following subsections. Figure 1 illustrates the design and principle of these guidance signals.

4.1 Random-subset Forgetting

Under the GSUO paradigm, guidance signals leveraged by SGO of **S1** are specified as diverse target distributions associated with different forget sample categories, while CO in **S2** performs fine-tuning with GD on $\mathcal{D}_r$. Note that forget samples are uniformly scattered in feature space and share similarities with other retain data in this case.

4.1.1 SGO under Target Distribution-based Signal

The core idea of the target distribution signal p_{target} is to identify predictive distributions that forget samples are expected to follow based on their location in the feature space of $f(\cdot;\Theta_o)$, and then align

the model's prediction for the forget sample with the corresponding target distribution to achieve unlearning. Due to the varying difficulty of forgetting across different samples, adopting a coarse-grained update strategy that treats all forget samples equally leads to a dilemma: in some cases, certain samples are under-unlearned, while in other cases, others are over-unlearned (see Appendix A for details). Hence, we categorize forget samples into *Normal* and *Boundary* groups based on the model's confidence, which serves as a practical proxy for the relative proximity of a sample to the decision boundary, and design corresponding guidance signals according to their expected behavior after unlearning. Specifically, the overall schema is as follows.

Normal. They are high-confidence samples in the original model. These samples should be expected to undergo perturbation or a decrease in predicted confidence after unlearning. Therefore, to facilitate a smooth unlearning, we construct the corresponding target distribution for a forget sample x_n as a linear combination of the original model prediction and a uniform distribution, i.e.,

$$p_{\text{target}} = \alpha \cdot \text{softmax}\big(f(x_{\text{n}}; \Theta_o)\big) + (1-\alpha) \cdot \mathbf{1}/C, \tag{1}$$

where $\mathbf{1}$ is a C-size all-one vector, which becomes a uniform distribution after normalization by $\frac{1}{C}$, and $\alpha \in [0,1]$ controls the forgetting strength. Such a guidance signal p_{target} steers the model to gradually reduce the confidence in the original predicted class for forget samples.

Boundary. They are samples with relatively low confidence in the original model, and most are close to the decision hyperplane. Consequently, they will *tend to cross decision boundaries after unlearning, leading to misclassification*. Accordingly, the corresponding guidance signal (target distribution) is designed as the predictive distribution of the most similar retained class in the feature space. Here, we first define the target signal for each class $c \in \mathcal{Y}$, which takes the prediction distribution P_c of the corresponding prototype $\boldsymbol{\mu}_c$ as a proxy. Specifically, we sample a subset $\mathcal{D}_P^c = \{(x,y) \in \mathcal{D}_r | y = c\}$ of size $R \ll |\mathcal{D}_r|$ and extract their feature by the extractor $\Phi(\cdot;\psi)$ where ψ denotes its parameters, the prototype is formulated as $\boldsymbol{\mu}_c = \frac{1}{R}\sum_{x \sim \mathcal{D}_P^c} \Phi(x;\psi)$, i.e., the center of class c in feature space.

Therefore, the target distribution signal for any boundary sample x_{b} is $p_{\text{target}} = P_{c^*}$, and

$$c^* = \arg\min_{c \neq y_{\text{true}}} \|\Phi(x_{\text{b}};\psi) - \boldsymbol{\mu}_c\|^2, \tag{2}$$

where y_{true} denotes the ground-truth label of x_{b}. Thus, it guides boundary samples to "blend into" semantically similar retain classes, achieving unlearning with minimal impact on the retain classes.

Relying on the above p_{target}, we align the model's predictions on the forget samples with the target distribution signal by minimizing the KL divergence, i.e., $\text{KL}(p_{\text{target}} \parallel p_{\text{model}})$. Normal and Boundary samples equally contribute to the total KL loss.

4.1.2 CO FOR ENHANCEMENT

To address the potential degradation of the model's discriminative ability after **S1**, we can fine-tune on $\mathcal{D}_r$ (optionally only its subset) to enhance for a few epochs with a standard CE loss.

4.2 CLASS-WISE FORGETTING

Under the GSUO paradigm, the guidance signal in **S1** is based on feature redistribution. In **S2**, we apply CE loss on $\mathcal{D}_r$ to reinforce correct classification and negative CE loss on $\mathcal{D}_f$ to suppress prediction ability for forget classes, both with a layer-wise learning rate strategy. For this task, forget samples form tight clusters in feature space with clear inter-class decision boundaries.

4.2.1 SGO UNDER FEATURE REDISTRIBUTION-BASED SIGNAL

The core idea of feature distribution signals is to reshape the distribution of forget samples in feature space according to their characteristics for thorough class-wise forgetting. As the empirical observation from retraining shows, *the forget samples become loosely distributed and their intra-class compactness is disrupted after unlearning*, which serves as the prior knowledge for designing efficient unlearning methods. Therefore, we construct two types of guidance signals: *a dispersion loss* and *an alignment loss*, to steer the unlearning objective by working together.

Intra-class dispersion loss. $\mathcal{D}_f$ contains all training samples of a subset of classes $\mathcal{Y}_f \subseteq \mathcal{Y}$. We define the dispersion loss L_{disp} as the negative logarithm of the average pairwise distance between the normalized features from the same forget class, that is,

$$L_{\text{disp}} = -\frac{1}{|\mathcal{Y}_f|} \sum_{\mathcal{D}_f^c} \log \left(\frac{1}{N_c(N_c-1)} \sum_{\substack{x_i, x_j \in \mathcal{D}_f^c \\ i \neq j}} \left(1 - \cos \left(\frac{\Phi(x_i;\psi)}{\|\Phi(x_i;\psi)\|}, \frac{\Phi(x_j;\psi)}{\|\Phi(x_j;\psi)\|} \right) \right) \right), \tag{3}$$

where $N_c = |\mathcal{D}_f^c|$ and $\cos(\cdot,\cdot)$ denotes cosine similarity. By maximizing the pairwise distances among intra-class features, L_{disp} forces the samples from forget class to scatter across the feature space, thereby disrupting their original clustering structure.

Alignment loss. In addition to intra-class dispersion after unlearning, the forget samples as a whole tend to shift toward the global center of the retained classes. Ignoring such an overall offset would weaken the forgetting performance where the scattered features still reside near their original region, which allows the classifier to make correct predictions through other means (e.g., nearest neighbor classification). Therefore, we introduce an alignment loss that drives the feature center of each forget class toward the retained region. Here, we define the global center of the retained classes as $\bar{\boldsymbol{\mu}}_r = \frac{1}{|\mathcal{D}_r|} \sum_{(x,y)\in\mathcal{D}_r} \Phi(x;\psi)$ and the normalized feature center for each forget class as $\boldsymbol{\mu}_f^c = \frac{1}{N_c} \sum_{x\in\mathcal{D}_f^c} \frac{\Phi(x;\psi)}{\|\Phi(x;\psi)\|}$. The alignment loss becomes

$$L_{\text{align}} = \frac{1}{|\mathcal{Y}_f|} \sum_{c\in\mathcal{Y}_f} \left(1 - \cos\left(\boldsymbol{\mu}_f^c, \bar{\boldsymbol{\mu}}_r\right)\right), \tag{4}$$

where $\mathcal{Y}_r = \mathcal{Y} \setminus \mathcal{Y}_f$ is the set of retained classes. Hence, the weighted sum of the dispersion loss and the alignment loss forms the guidance signal for unlearning.

4.2.2 CO FOR ENHANCEMENT

In **S2**, we jointly optimize the feature extractor and classifier with the CE loss, reinforcing correct classification of retained samples while suppressing accurate predictions for forget samples. Besides, we adopt a layer-wise learning rate strategy: a smaller one is applied to the feature extractor to maintain the achieved feature reshaping effect, while a larger one is applied to the classifier to adapt it to the new feature distribution and enhance its discriminative ability for the retained classes.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETTINGS

Datasets We conduct experiments on CIFAR-10 and CIFAR-100 (Krizhevsky et al., 2009), Lacuna-10 (randomly select 10 classes with at least 500 samples from VGG-Faces (Cao & Yang, 2015)), and Tiny-ImageNet (Le & Yang, 2015), similar to previous work, to measure the unlearning performance for classification tasks.

For *random-subset forgetting*, we choose to forget $\approx 10\%$ of whole training data, i.e., $N_f = |\mathcal{D}_f| = 5,000$ for CIFAR-10 and CIFAR-100 and $N_f = 10k$ for Tiny-ImageNet. For *class-wise forgetting*, we report the results on Lacuna-10 for forgetting a randomly selected class.

Models For Lacuna-10, we fine-tuned a pre-trained VGG-11 model using Adam with a fixed $lr = 1E-3$ and batch size of 128. For CIFAR-10, we fine-tuned a pre-trained ResNet-18 model using Adam with a fixed $lr = 1E-3$ and batch size of 256. For CIFAR-100 and Tiny-ImageNet, we finetuned pre-trained VGG-16 (Hayase et al., 2020), ResNet-50 (He et al., 2016), and ViT (Dosovitskiy et al., 2021) models using Adam with a fixed $lr = 1E-4$ and batch size of 256. All methods are implemented in Python 3.11 with PyTorch. All experiments are performed on NVIDIA Tesla-A100 (80G) with Intel Xeon processors.

Baselines We covered **14 state-of-the-art (SOTA) baselines**: **Original**: the model trained on $\mathcal{D}$. **Retrain**:retraining the model from scratch without $\mathcal{D}_f$ (included as a reference but inviable

Table 1: Performance comparison among baselines and GSUO for random-subset forgetting across different datasets and model architectures.

Dataset (model)	Metric	Original	Retrain	FT	GA	RL	EU-k	IU	ℓ_1-sparse	SalUn	Bad-T	SCRUB	GSUO
CIFAR-10 (ResNet-18)	$Acc(\mathcal{D}_r)$ ↑	97.27	98.61	98.66	97.50	93.51	98.90	85.03	96.20	98.51	97.50	93.93	**99.47**
	$Acc(\mathcal{D}_t)$ ↑	80.87	80.61	81.18	79.54	76.19	81.55	71.20	78.70	80.51	79.07	78.22	**82.11**
	$Acc(\mathcal{D}_f)$	97.38	81.24	96.95	96.32	93.64	98.72	85.44	86.52	92.26	73.64	89.48	82.12
	| Diff | ↓	16.51	0.63	15.77	16.78	17.45	17.17	14.24	7.82	11.75	5.43	11.26	**0.01**
	Time (s) ↓	–	230.10	16.61	**1.08**	1.65	85.28	41.07	65.98	86.95	12.34	11.93	4.77
CIFAR-100 (VGG-16)	$Acc(\mathcal{D}_r)$ ↑	99.53	99.53	99.59	99.02	98.41	99.88	99.53	99.85	99.53	99.41	**99.95**	87.02
	$Acc(\mathcal{D}_t)$ ↑	60.35	58.63	59.81	58.80	57.61	61.15	60.35	60.63	60.11	58.81	**61.48**	61.3
	$Acc(\mathcal{D}_f)$	99.40	58.72	97.50	97.26	98.54	99.84	99.38	96.94	98.36	67.30	92.96	62.90
	| Diff | ↓	39.05	0.09	37.69	38.46	40.93	38.69	39.03	36.3	38.25	8.49	31.48	**1.6**
	Time (s) ↓	–	366.09	139.61	**1.895**	1.93	92.63	372.10	79.50	49.45	19.92	34.58	18.45
CIFAR-100 (ViT)	$Acc(\mathcal{D}_r)$ ↑	98.03	97.33	99.45	89.76	97.87	97.84	97.80	96.74	99.24	97.53	**99.79**	99.04
	$Acc(\mathcal{D}_t)$ ↑	88.96	90.20	90.40	80.95	88.65	89.03	88.84	88.06	90.18	89.29	**91.02**	85.06
	$Acc(\mathcal{D}_f)$	98.16	90.78	98.62	88.80	98.04	97.82	97.98	96.12	98.40	85.82	99.20	87.16
	| Diff | ↓	9.20	0.58	8.22	7.85	9.39	8.79	9.14	8.06	8.22	3.47	8.18	**2.10**
	Time (s) ↓	–	420.53	334.91	**23.65**	25.10	189.12	3287.35	125.43	140.18	397.08	739.71	161.45
Tiny ImageNet (ResNet50)	$Acc(\mathcal{D}_r)$ ↑	99.39	99.94	99.74	98.79	83.81	**99.96**	99.39	99.56	98.98	98.31	99.81	98.45
	$Acc(\mathcal{D}_t)$ ↑	78.74	78.34	77.51	78.14	66.67	78.57	78.75	77.69	77.54	71.75	**78.90**	78.65
	$Acc(\mathcal{D}_f)$	99.36	78.39	97.90	97.74	83.48	99.38	99.35	98.74	97.30	78.84	99.11	77.96
	| Diff | ↓	20.62	0.05	20.39	19.60	16.81	20.81	20.60	21.05	19.76	7.09	20.21	**0.69**
	Time (s) ↓	–	1027.23	354.36	12.52	**12.32**	396.83	52.84	83.00	93.60	482.42	988.39	148.6

in practice). **Fine-tuning (FT)**, **Gradient Ascent (GA)** (Golatkar et al., 2020a), **Random Labels (RL)** (Graves et al., 2021), **Influence Unlearning (IU)** (Koh & Liang, 2017; Izzo et al., 2021), **EU-k** (Goel et al., 2022), **ℓ_1-sparse** (Jia et al., 2023), **SalUn** (Fan et al., 2024), **Bad-teaching (Bad-T)** (Chundawat et al., 2023) and **SCRUB** (Kurmanji et al., 2023). Besides, for the class-wise forgetting, we include specific methods: **Boundary Shrink (B-Shrink) & Expanding (B-Expand)** (Chen et al., 2023) and **UNSIR** (Tarun et al., 2024).

Considering the comprehensive forgetting efficacy, model performance, and privacy protection, we follow the evaluation metrics below:

- *Accuracy:* The accuracies on different datasets, including the retain training set ($\mathcal{D}_r$), retain test set ($\mathcal{D}_{rt}$), forget training set ($\mathcal{D}_f$), forget test set ($\mathcal{D}_{ft}$) and the full test set ($\mathcal{D}_t$);
- *Difference (|Diff|):* The absolute accuracy difference between forget set and test set, i.e., $|\text{Diff}| = |Acc(\mathcal{D}_f) - Acc(\mathcal{D}_t)|$, which measures whether the model still retains memory of the forgotten data after unlearning;
- *Acc-Index:* An integrated accuracy index for the random-subset unlearning task as $F_{acc}(\mathcal{D}_t, \mathcal{D}_f) = Acc(\mathcal{D}_t) - |\text{ Diff }|$;
- *Time:* The running time for performing a machine unlearning method;
- *Speedup:* The ratio of the time required for retraining on $\mathcal{D}_r$ to the unlearning method.
- *Attack Success Rate (ASR):* The attack success rate of MIA, measuring the model's effectiveness in privacy protection with the ideal value being close to a random guess.

The Acc-Index (F_{acc}) jointly accounts for the generalization capacity and unlearning efficacy of the unlearned model, i.e., we expect better accuracy on $\mathcal{D}_t$ while the accuracy of $\mathcal{D}_f$ should be as consistent with it as possible. Experimental settings and more results are deferred to Appendix C-D.

5.2 Random-subset Forgetting

For random-subset forgetting, Table 1 reports the utility and efficiency of different unlearning methods. Figure 2(a)-(b) shows Speedup versus F_{acc} for ResNet-18 on CIFAR-10 and VGG-16 on CIFAR-100, with additional results in Appendix D. We evaluate privacy protection by adopting the U-LiRA (Hayes et al., 2025) MIA based on the log-likelihood ratio, which is considered the SOTA per-sample attack benchmark for unlearning scenarios to date. The attack procedure is provided in Appendix C. Figure 2(c) shows MIA results and AUC score for different unlearning methods.

GSUO achieves the best overall performance. We desire to achieve |Diff| as small as possible. Results show that GSUO consistently produces the lowest |Diff| across all architectures and datasets, significantly outperforming existing baselines.

We expect $Acc(\mathcal{D}_r)$ and $Acc(\mathcal{D}_t)$ to remain high after unlearning. GSUO satisfies this expectation across different architectures. On CIFAR-10 with ResNet-18, GSUO achieves higher retain and test accuracy than all baselines. Although GSUO's $Acc(\mathcal{D}_r)$ on CIFAR-100 with VGG-16 is lower than

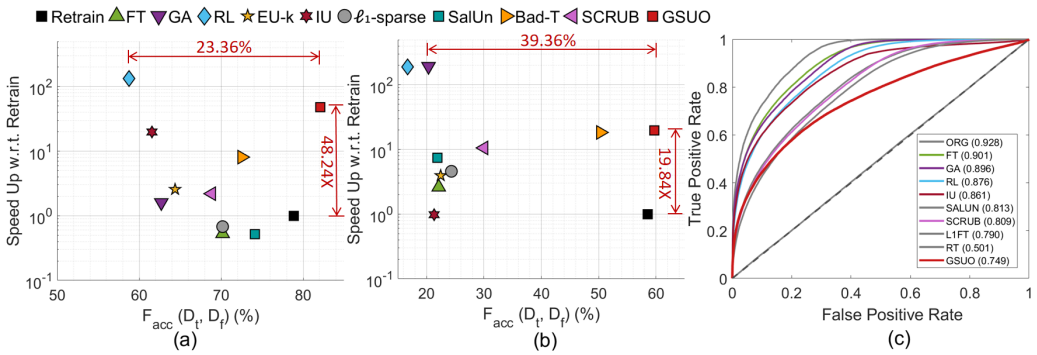


Figure 2: (a)-(b) Speedup versus F_{acc} for ResNet-18 on CIFAR-10 and VGG-16 on CIFAR-100; (c) U-LiRA results and AUC score for different unlearning methods on ResNet-18 on CIFAR-10.

SCRUB and EU-k, $Acc(\mathcal{D}_t)$ is by no means inferior. This indicates that GSUO does not simply memorize retain samples to achieve high retain accuracy, but instead maintains retain performance while exhibiting stronger generalization. SCRUB also performs well in this regard, but its high $Acc(\mathcal{D}_f)$ indicates incomplete forgetting. In contrast, GSUO maintains high retain and test accuracy while effectively reducing prediction accuracy on the forget set, thereby achieving a better balance between selective forgetting and model utility.

For comprehensive evaluation, we adopt the metric F_{acc} to balance the model's generalization and its forgetting effectiveness on the forget set. We also measure computational efficiency via the speedup ratio relative to Retrain. As shown in Figure 2, GSUO consistently achieves a favorable trade-off between unlearning quality and computational efficiency (occupying the top-right region, i.e., the optimal zone with high F_{acc} and high speedup). Although methods such as RL are faster on some datasets according to Table 1, their forgetting effectiveness is far inferior to GSUO. This demonstrates that GSUO effectively balances efficiency and forgetting quality, delivering strong and robust unlearning performance without sacrificing computational practicality across diverse settings.

We further consider a privacy-critical application: deleting data of users who exercise their "right to be forgotten". To this end, we adopt U-LiRA as the privacy evaluation metric, aiming to ensure that after unlearning, an attacker should be unable to distinguish between forgotten samples and truly unseen samples, thereby protecting the privacy of users who request deletion. Figure 2 shows the U-LiRA results, where GSUO achieves the best performance, substantially outperforming all baselines. This result demonstrates that GSUO effectively mitigates membership inference attacks on forgotten samples, providing strong protection against membership inference attacks while maintaining competitive model utility.

GSUO demonstrates strong scalability and robustness as data and model scale increase. As forgetting tasks scale from smaller datasets with fewer classes to larger, more complex ones with more categories and samples, the forgetting exactness $|\mathrm{Diff}|$ of GSUO remains consistently low, substantially outperforming all baselines. Meanwhile, it maintains competitive computational efficiency. These results verify the stability and effectiveness of GSUO in complex, large-scale forgetting tasks.

5.3 Class-wise Forgetting

For class-wise forgetting, we evaluated the effectiveness of different unlearning methods on forgetting all samples of a randomly selected class (class 4) on the Lacuna-10 dataset. Table 2 reports the corresponding effectiveness and efficiency of these methods.

We desire the training and test accuracy on the forget set to be as small as possible. GSUO achieves the optimal level on both metrics, completely removing the influence of the forgotten class from the model. This includes reducing the classification accuracy on both the forget set and the corresponding test samples to zero, indicating effective class-wise forgetting under the evaluated metrics. Despite achieving perfect forgetting, GSUO simultaneously preserves strong performance on retain data. Specifically, it achieves the highest retain set accuracy among all baselines, and its perfor-

Table 2: Comparison of MU methods for class-wise forgetting on Lacuna-10 under VGG-11.

Method	$Acc(\mathcal{D}_r)$ ↑	$Acc(\mathcal{D}_{rt})$ ↑	$Acc(\mathcal{D}_f)$ ↓	$Acc(\mathcal{D}_{ft})$ ↓	**Time (s)** ↓
Original	100.00	87.67	99.74	90.00	—
Retrain	99.94	82.67	0.00	0.00	396.76
FT	98.43	86.78	**0.00**	**0.00**	50.98
GA	88.09	76.67	0.80	1.00	49.03
RL	97.03	87.78	1.21	1.00	60.69
EU-k	98.95	88.22	**0.00**	**0.00**	42.37
IU	89.11	79.67	12.68	5.00	**1.98**
ℓ_1-sparse	82.78	76.00	**0.00**	**0.00**	51.85
SalUn	98.72	87.56	**0.00**	**0.00**	105.67
Bad-T	93.59	81.78	82.70	67.00	87.13
SCRUB	<u>98.92</u>	**89.78**	0.80	0.01	90.84
B-Shrink	96.39	85.11	**0.00**	**0.00**	<u>5.96</u>
B-Expand	95.57	85.44	**0.00**	**0.00**	10.72
UNSIR	83.95	77.44	**0.00**	**0.00**	25.52
GSUO	**99.21**	<u>89.11</u>	**0.00**	**0.00**	35.59

Table 3: Ablation study on GSUO framework.

	Random-subset Forgetting				Class-wise Forgetting	
	$Acc(\mathcal{D}_r)$	$Acc(\mathcal{D}_f)$	$Acc(\mathcal{D}_t)$	| Diff |	$Acc(\mathcal{D}_{rt})$	$Acc(\mathcal{D}_{ft})$
Original	97.27	97.38	80.87	16.51	87.67	90.00
w/o SGO	99.92	99.46	83.17	16.29	87.89	0.00
w/o CO	88.64	78.54	73.43	5.11	84.56	6.00
GSUO	99.47	82.12	82.11	0.01	89.11	0.00

mance on the retain test set is also among the best, ranking second overall. While several baselines (e.g., FT, EU-k, SalUn) also achieve zero accuracy on the forget set, they either fall short of GSUO on retain set accuracy or exhibit higher computational cost. Methods such as GA, RL, and SCRUB fail to achieve complete forgetting, leaving non-zero accuracy on the forget set.

GSUO achieves a substantial speedup compared to retraining, outperforming most baselines. Although a few methods are faster, they compromise either forgetting completeness or model utility. Overall, GSUO strikes a strong trade-off among forgetting effectiveness, model utility, and computational efficiency, making it a practical and reliable solution for selective class removal.

5.4 Ablation Study

To verify the necessity of SGO and CO, we conduct ablation studies by removing each component individually. The ablation study on random-subset forgetting is conducted with ResNet-18 on CIFAR-10. The ablation study on class-wise forgetting is conducted with VGG-11 on Lacuna-10. The results are shown in Table 3.

For random-subset forgetting, we observe that removing either SGO or CO leads to notable performance degradation. Specifically, without SGO, the model retains excessive information on the forget set; without CO, test accuracy drops significantly, with $|\mathrm{Diff}|$ remaining large in both cases. For class-wise forgetting, we observe similar trends: removing either component results in lower retain test set accuracy compared to the full GSUO. We conclude that SGO is essential for eliminating the influence of forgotten samples, while CO plays a critical role in preserving model generalization. The full GSUO achieves the most favorable overall trade-off between forgetting effectiveness and retained-data utility, supporting the complementary roles of SGO and CO.

To further investigate the contribution of each component in SGO, we conduct ablation experiments under two forgetting tasks. The results are shown in Table 4.

For random-subset forgetting, we observe that treating all forget samples uniformly (as either all Normal or all Boundary) results in limited forgetting performance, with forget set accuracy remaining high in both cases. This finding suggests that Random-subset forgetting requires a more nuanced approach: different types of forget samples need distinct target distributions to achieve effective un-

Table 4: Ablation study on SGO.

Task	Method	$Acc(\mathcal{D}_r)$	$Acc(\mathcal{D}_{rt})$	$Acc(\mathcal{D}_f)$	$Acc(\mathcal{D}_{ft})$	$Acc(\mathcal{D}_t)$	|Diff|
Random-subset	Original	97.27	—	97.38	—	80.87	16.51
	Using Normal Signal Only	99.6	—	95.62	—	82.59	13.03
	Using Boundary Signal Only	99.92	—	98.8	—	83.73	15.07
	GSUO	99.47	—	82.12	—	82.11	0.01
Class-wise	Original	100.00	87.67	99.74	90.00	—	—
	w/o Dispersion	98.75	88.22	0.00	0.00	—	—
	w/o Alignment	96.45	84.78	0.00	0.00	—	—
	GSUO	99.21	89.11	0.00	0.00	—	—

Table 5: Performance of GSUO across different forgetting ratios.

Metrics	0	5%	10%	15%	20%	25%	30%	35%	40%
$Acc(\mathcal{D}_r)$ ↑	97.27	99.57	99.47	99.30	99.06	98.72	98.27	97.99	97.59
$Acc(\mathcal{D}_t)$ ↑	80.87	82.65	82.11	82.02	81.80	81.65	80.51	80.07	79.76
$Acc(\mathcal{D}_f)$	—	81.20	82.12	82.31	83.65	85.84	81.58	82.17	81.95
| Diff | ↓	—	1.45	0.01	0.29	1.85	4.19	1.07	2.10	2.19
Time (s) ↓	—	4.79	4.77	4.55	4.67	4.95	5.00	5.46	5.37

learning. In contrast, the complete GSUO substantially reduces forget set accuracy while keeping retain set accuracy high, achieving the optimal balance we desire. For class-wise forgetting task, we observe that removing either the dispersion loss or the alignment loss still allows the model to achieve zero accuracy on the forget set. However, retain set accuracy drops noticeably compared to the full GSUO. This indicates that even when forget set accuracy has already reached the optimal level, both the dispersion loss and the alignment loss still help preserve model performance on the retain data. We therefore conclude that each component in SGO plays a distinct yet complementary role, and only their full combination achieves a favorable trade-off between forgetting effectiveness and model utility preservation across both forgetting tasks.

5.5 Sensitivity Analysis on Forget-set Size

Table 5 shows that GSUO achieves a favorable trade-off between utility, forgetting efficacy, and efficiency. As the forgetting ratio increases from 0% to 40%, GSUO maintains high accuracy on the retain set $\mathcal{D}_r$. The test-set accuracy $Acc(D_t)$ remains relatively stable overall, indicating that the model maintains reasonable generalization performance as the forgetting ratio increases. The accuracy on the forget set $\mathcal{D}_f$ is effectively suppressed across different forgetting ratios. Furthermore, the $|\text{Diff}|$ metric remains below 5% regardless of the forgetting ratio, demonstrating that GSUO achieves efficient and stable Random-subset forgetting. The runtime of GSUO remains significantly lower than the cost of full retraining, confirming the practicality and scalability of our method for large-scale unlearning scenarios.

6 Conclusion

We propose GSUO, a two-step unlearning framework consisting of Signal-Guided Optimization and Compensatory Optimization. By leveraging task-specific guidance signals and targeted utility restoration, GSUO provides a principled and efficient solution to the challenges of over-unlearning and under-unlearning in both random-subset and class-wise forgetting scenarios. Extensive experiments demonstrate that GSUO consistently surpasses 14 state-of-the-art baselines across unlearning efficacy, generalization, and privacy robustness, while delivering substantial speedups, confirming its effectiveness for trustworthy machine unlearning.

Limitation. GSUO outperforms baselines in privacy evaluation, but still falls short of the retraining upper bound. Experiments are limited to traditional models; its effectiveness on emerging architectures such as LLMs and VLMs remains to be validated.

AI USE STATEMENT

In this work, we used generative AI tools to assist in the writing of proofs and assist with translation. We have not used generative AI tools for developing theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, implementing methods, cleaning and reformatting datasets, supporting qualitative and thematic data analysis, or interpreting results. Generating synthetic datasets does not apply to this work. Additionally, we used generative AI tools for editing the research paper to improve readability. We have reviewed all AI-assisted work. In particular, all AI-assisted proofs were carefully checked by the authors for mathematical correctness, and all translated and readability-edited text was reviewed by the authors against the original content to ensure that the intended meaning was preserved and that no unsupported claims or substantive changes were introduced. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS AND REPRODUCIBILITY

ETHICS STATEMENT

We have considered the ethical implications of this work and do not identify any specific ethical concerns requiring further discussion.

REPRODUCIBILITY STATEMENT

We provide sufficient details to facilitate the reproducibility of our results. The main paper describes the proposed method and experimental setup, including the model architecture and training procedure. Additional implementation details and experimental configurations are provided in the appendix C. The code is available at https://anonymous.4open.science/r/GSUO.

REFERENCES

Lucas Bourtoule, Varun Chandrasekaran, Christopher A. Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In *42nd IEEE Symposium on Security and Privacy, SP 2021, San Francisco, CA, USA, 24-27 May 2021*, pp. 141–159. IEEE, 2021. doi: 10.1109/SP40001.2021.00019. URL https://doi.org/10.1109/SP40001.2021.00019.

Yinzhi Cao and Junfeng Yang. Towards making systems forget with machine unlearning. In *2015 IEEE Symposium on Security and Privacy, SP 2015, San Jose, CA, USA, May 17-21, 2015*, pp. 463–480. IEEE Computer Society, 2015. doi: 10.1109/SP.2015.35. URL https://doi.org/10.1109/SP.2015.35.

Min Chen, Weizhuo Gao, Gaoyang Liu, Kai Peng, and Chen Wang. Boundary unlearning: Rapid forgetting of deep networks via shifting the decision boundary. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pp. 7766–7775. IEEE, 2023. doi: 10.1109/CVPR52729.2023.00750. URL https://doi.org/10.1109/CVPR52729.2023.00750.

Min Chen, Weizhuo Gao, Chen Wang, Gaoyang Liu, Ahmed M. Abdelmoniem, and Kai Peng. From expansion to retraction: Long-tailed machine unlearning via boundary manipulation. In Luiza Antonie, Jian Pei, Xiaohui Yu, Flavio Chierichetti, Hady W. Lauw, Yizhou Sun, and Srinivasan Parthasarathy (eds.), *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining, V.2, KDD 2025, Toronto ON, Canada, August 3-7, 2025*, pp. 215–226. ACM, 2025. doi: 10.1145/3711896.3736970. URL https://doi.org/10.1145/3711896.3736970.

Vikram S. Chundawat, Ayush K. Tarun, Murari Mandal, and Mohan S. Kankanhalli. Can bad teaching induce forgetting? unlearning in deep networks using an incompetent teacher. In Brian Williams, Yiling Chen, and Jennifer Neville (eds.), *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pp. 7210–7217. AAAI Press, 2023. doi: 10.1609/AAAI.V37I6.25879. URL https://doi.org/10.1609/aaai.v37i6.25879.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Chongyu Fan, Jiancheng Liu, Yihua Zhang, Eric Wong, Dennis Wei, and Sijia Liu. Salun: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=gn0mIhQGNM.

Shashwat Goel, Ameya Prabhu, and Ponnurangam Kumaraguru. Evaluating inexact unlearning requires revisiting forgetting. *CoRR*, abs/2201.06640, 2022. URL https://arxiv.org/abs/2201.06640.

Aditya Golatkar, Alessandro Achille, and Stefano Soatto. Eternal sunshine of the spotless net: Selective forgetting in deep networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pp. 9301–9309. Computer Vision Foundation / IEEE, 2020a. doi: 10.1109/CVPR42600.2020.00932. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Golatkar_Eternal_Sunshine_of_the_Spotless_Net_Selective_Forgetting_in_Deep_CVPR_2020_paper.html.

Aditya Golatkar, Alessandro Achille, and Stefano Soatto. Forgetting outside the box: Scrubbing deep networks of information accessible from input-output observations. In Andrea Vedaldi,

Horst Bischof, Thomas Brox, and Jan-Michael Frahm (eds.), *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXIX*, Lecture Notes in Computer Science, pp. 383–398. Springer, 2020b. doi: 10.1007/978-3-030-58526-6_23. URL https://doi.org/10.1007/978-3-030-58526-6_23.

Aditya Golatkar, Alessandro Achille, Avinash Ravichandran, Marzia Polito, and Stefano Soatto. Mixed-privacy forgetting in deep networks. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pp. 792–801. Computer Vision Foundation / IEEE, 2021. doi: 10.1109/CVPR46437.2021.00085. URL https://openaccess.thecvf.com/content/CVPR2021/html/Golatkar_Mixed-Privacy_Forgetting_in_Deep_Networks_CVPR_2021_paper.html.

Laura Graves, Vineel Nagisetty, and Vijay Ganesh. Amnesiac machine learning. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pp. 11516–11524. AAAI Press, 2021. doi: 10.1609/AAAI.V35I13.17371. URL https://doi.org/10.1609/aaai.v35i13.17371.

Chuan Guo, Tom Goldstein, Awni Y. Hannun, and Laurens van der Maaten. Certified data removal from machine learning models. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, Proceedings of Machine Learning Research, pp. 3832–3842. PMLR, 2020. URL http://proceedings.mlr.press/v119/guo20c.html.

Tomohiro Hayase, Suguru Yasutomi, and Takashi Katoh. Selective forgetting of deep networks at a finer level than samples. *CoRR*, abs/2012.11849, 2020. URL https://arxiv.org/abs/2012.11849.

Jamie Hayes, Ilia Shumailov, Eleni Triantafillou, Amr Khalifa, and Nicolas Papernot. Inexact unlearning needs more careful evaluations to avoid a false sense of privacy. In *IEEE Conference on Secure and Trustworthy Machine Learning, SaTML 2025, Copenhagen, Denmark, April 9-11, 2025*, pp. 497–519. IEEE, 2025. doi: 10.1109/SATML64287.2025.00034. URL https://doi.org/10.1109/SaTML64287.2025.00034.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 770–778. IEEE Computer Society, 2016. doi: 10.1109/CVPR.2016.90. URL https://doi.org/10.1109/CVPR.2016.90.

Zhengbao He, Tao Li, Xinwen Cheng, Zhehao Huang, and Xiaolin Huang. Towards natural machine unlearning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 47(12):11548–11560, 2025. doi: 10.1109/TPAMI.2025.3597350. URL https://doi.org/10.1109/TPAMI.2025.3597350.

Zachary Izzo, Mary Anne Smart, Kamalika Chaudhuri, and James Zou. Approximate data deletion from machine learning models. In Arindam Banerjee and Kenji Fukumizu (eds.), *The 24th International Conference on Artificial Intelligence and Statistics, AISTATS 2021, April 13-15, 2021, Virtual Event*, Proceedings of Machine Learning Research, pp. 2008–2016. PMLR, 2021. URL http://proceedings.mlr.press/v130/izzo21a.html.

Jiabao Ji, Yujian Liu, Yang Zhang, Gaowen Liu, Ramana Kompella, Sijia Liu, and Shiyu Chang. Reversing the forget-retain objectives: An efficient LLM unlearning framework from logit difference. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/171291d8fed723c6dfc76330aa827ff8-Abstract-Conference.html.

Jinghan Jia, Jiancheng Liu, Parikshit Ram, Yuguang Yao, Gaowen Liu, Yang Liu, Pranay Sharma, and Sijia Liu. Model sparsity can simplify machine unlearning. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/a204aa68ab4e970e1ceccfb5b5cdc5e4-Abstract-Conference.html.

Jinghan Jia, Yihua Zhang, Yimeng Zhang, Jiancheng Liu, Bharat Runwal, James Diffenderfer, Bhavya Kailkhura, and Sijia Liu. SOUL: unlocking the power of second-order optimization for LLM unlearning. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pp. 4276–4292. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.EMNLP-MAIN.245. URL https://doi.org/10.18653/v1/2024.emnlp-main.245.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, Proceedings of Machine Learning Research, pp. 1885–1894. PMLR, 2017. URL http://proceedings.mlr.press/v70/koh17a.html.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. *Master's thesis, Department of Computer Science, University of Toronto*, 2009.

Meghdad Kurmanji, Peter Triantafillou, Jamie Hayes, and Eleni Triantafillou. Towards unbounded machine unlearning. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/062d711fb777322e2152435459e6e9d9-Abstract-Conference.html.

Yann Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

Zheyuan Liu, Guangyao Dou, Zhaoxuan Tan, Yijun Tian, and Meng Jiang. Towards safer large language models through machine unlearning. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, Findings of ACL, pp. 1817–1829. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.FINDINGS-ACL.107. URL https://doi.org/10.18653/v1/2024.findings-acl.107.

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 Trans. Intell. Syst. Technol.*, 16(5):108:1–108:46, 2025. doi: 10.1145/3749987. URL https://doi.org/10.1145/3749987.

Ayush Sekhari, Jayadev Acharya, Gautam Kamath, and Ananda Theertha Suresh. Remember what you want to forget: Algorithms for machine unlearning. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 18075–18086, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/9627c45df543c816a3ddf2d8ea686a99-Abstract.html.

Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin A. Riedmiller. Striving for simplicity: The all convolutional net. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6806.

Ayush K. Tarun, Vikram S. Chundawat, Murari Mandal, and Mohan S. Kankanhalli. Fast yet effective machine unlearning. *IEEE Trans. Neural Networks Learn. Syst.*, 35(9):13046–13055, 2024. doi: 10.1109/TNNLS.2023.3266233. URL https://doi.org/10.1109/TNNLS.2023.3266233.

Paul Voigt and Axel von dem Bussche. *The EU General Data Protection Regulation (GDPR): A Practical Guide*. Springer Publishing Company, Incorporated, 1st edition, 2017. ISBN 3319579584.

Abudukelimu Wuerkaixi, Qizhou Wang, Sen Cui, Wutong Xu, Bo Han, Gang Niu, Masashi Sugiyama, and Changshui Zhang. Adaptive localization of knowledge negation for continual LLM unlearning. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, Proceedings of Machine Learning Research. PMLR / OpenReview.net, 2025. URL https://proceedings.mlr.press/v267/wuerkaixi25a.html.

Heng Xu, Tianqing Zhu, Lefeng Zhang, Wanlei Zhou, and Philip S. Yu. Machine unlearning: A survey. *ACM Comput. Surv.*, 56(1):9:1–9:36, 2024. doi: 10.1145/3603620. URL https://doi.org/10.1145/3603620.

Liheng Yu, Zhe Zhao, Yuxuan Wang, Pengkun Wang, Binwu Wang, and Yang Wang. Falw: A forgetting-aware loss reweighting for long-tailed unlearning. *CoRR*, abs/2601.18650, 2026. doi: 10.48550/ARXIV.2601.18650. URL https://doi.org/10.48550/arXiv.2601.18650.

APPENDIX

A THEORETICAL ANALYSIS.

Notation. Let $\mathcal{D}$ denote the full training dataset, and let $\mathcal{D}_r$ and $\mathcal{D}_f$ denote the retain and forget sets, respectively. The forget set is further partitioned into Boundary and Normal subsets,

$$\mathcal{D}_f = \mathcal{D}_f^{\mathrm{b}} \cup \mathcal{D}_f^{\mathrm{n}}, \tag{5}$$

where $\mathcal{D}_f^{\mathrm{b}}$ contains low-confidence Boundary samples and $\mathcal{D}_f^{\mathrm{n}}$ contains high-confidence Normal samples.

We denote an individual sample by $z = (x, y)$, where x denotes the input and y denotes the ground-truth class index. Let $\mathbf{p}$ denote the predicted probability vector, and let $\mathbf{q}$ denote the one-hot encoding of the ground-truth class y.

- $\mathcal{D}_f^{\mathrm{b}}$ **(Boundary samples):** samples with relatively low prediction confidence, where confidence is measured by the maximum predicted probability.
- $\mathcal{D}_f^{\mathrm{n}}$ **(Normal samples):** samples with relatively high prediction confidence under the same confidence measure.

Proposition 1 (Heterogeneous Forgetting Difficulty)**.** *Samples in the forget set can exhibit heterogeneous forgetting difficulty. In particular, low-confidence Boundary samples generally induce larger sample-wise parameter responses than high-confidence Normal samples, and are therefore associated with smaller values of our proposed relative forgetting-difficulty proxy. Conversely, high-confidence Normal samples generally induce smaller parameter responses and are relatively more difficult to forget. Thus, a uniform unlearning update may affect different forget samples to different extents.*

Proof. We formalize the above observation by relating prediction confidence to sample-wise gradient magnitude. For a sample $z = (x, y)$, where x denotes the input and y denotes the ground-truth class index, let $\mathbf{p}$ denote the predicted probability vector. In GSUO, the operational confidence score used to partition forget samples is the maximum predicted probability, i.e., $c(z) = \max_k \mathbf{p}_k$. For correctly classified samples, this score coincides with $\mathbf{p}_y$, the predicted probability assigned to the ground-truth class. The following gradient-based analysis focuses on this case to provide intuition for the heterogeneous parameter responses of high- and low-confidence samples.

Consider the final classification layer. Let $\mathbf{h}$ denote the feature representation produced by the penultimate layer, and let $\mathbf{W}$ denote the weight matrix of the final classification layer. The output logits are given by

$$\mathbf{o} = \mathbf{W}\mathbf{h}, \tag{6}$$

and the predicted probability vector is

$$\mathbf{p} = \mathrm{softmax}(\mathbf{o}). \tag{7}$$

Let $\mathbf{q}$ denote the one-hot encoding of the ground-truth class y. For the cross-entropy loss

$$\ell(z) = -\mathbf{q}^T \log \mathbf{p}, \tag{8}$$

the sample-wise gradient with respect to the classification-layer weight matrix is

$$\nabla_{\mathbf{W}} \ell(z) = (\mathbf{p} - \mathbf{q})\mathbf{h}^T. \tag{9}$$

Since $\nabla_{\mathbf{W}} \ell(z)$ is a matrix, we measure its magnitude using the Frobenius norm. From Eq. 9,

$$\|\nabla_{\mathbf{W}} \ell(z)\|_F = \|\mathbf{p} - \mathbf{q}\|_2 \, \|\mathbf{h}\|_2 \,. \tag{10}$$

Eq. 10 shows that the magnitude of the classification-layer gradient is determined by two factors: the prediction error $\|\mathbf{p}-\mathbf{q}\|_2$ and the feature magnitude $\|\mathbf{h}\|_2$. Under the assumption that the feature norm is approximately stable across samples, the relative gradient magnitude is primarily determined

Table 6: Distributional comparison of the curvature-aware influence $I(z)$ between Boundary (B) and Normal (N) samples.

Metric	Mean(B)	Mean(N)	Mean(B)/Mean(N)	Med.(B)/Med.(N)	W_1	$W_1(\ln(1+\cdot))$	KS	Cliff's δ
$I(z)$	1007.8	19.1	52.8	100.6	988.8	4.37	0.954	0.994

by the prediction error. Therefore, lower-confidence samples tend to induce larger classification-layer gradient responses.

To characterize the response of the entire model, let θ denote the model parameter vector and define

$$G(z) = \|\nabla_\theta \ell(z; \Theta_o)\|_2 , \tag{11}$$

where Θ_o denotes the original model trained on $\mathcal{D}$. We use $G(z)$ as a practical proxy for the overall sample-wise parameter response. The extension from the classification layer to the full model assumes that the relative ordering of sample-wise responses is not substantially altered across layers. We use the inverse gradient response as a proxy for the relative forgetting difficulty of a sample:

$$D(z) = \frac{1}{G(z)}. \tag{12}$$

Under this definition, a larger sample-wise gradient response generally corresponds to a smaller relative forgetting difficulty. Therefore, under the assumptions above, higher-confidence samples tend to exhibit smaller gradient responses and consequently larger values of the proposed relative forgetting-difficulty proxy.

In particular, for a Boundary sample $z_b \in \mathcal{D}_f^b$ and a Normal sample $z_n \in \mathcal{D}_f^n$, the proposed relationship gives

$$G(z_b) > G(z_n), \qquad D(z_b) < D(z_n). \tag{13}$$

□

Empirical Verification of Proposition 1. We further verify the sample-wise response heterogeneity described in Proposition 1 using a curvature-aware influence measure. Compared with the gradient norm alone, which only characterizes the local first-order sensitivity of the loss with respect to the model parameters, the curvature-aware measure additionally accounts for the local geometry of the loss landscape. Specifically, for each sample (z), we define

$$I(z) = \left|\mathbf{H}^{-1} g_z\right| 2, \qquad g_z = \nabla_\theta \ell(z; \Theta_o), \tag{14}$$

where $\mathbf{H}$ denotes the Hessian of the training objective evaluated at the original model Θ_o. The inverse Hessian accounts for the local curvature of the loss landscape and therefore provides a curvature-aware characterization of the parameter-level influence of an individual sample.

Since computing the exact full-network inverse-Hessian-vector product for large neural network models, e.g., ResNet-18, is computationally intractable, we estimate $I(z)$ on the classifier head using a diagonal empirical Fisher approximation, following prior work that adopts scalable Fisher-based approximations of second-order curvature (Kirkpatrick et al., 2017). We then divide the forget samples into Boundary and Normal groups according to the confidence threshold and compare the distributions of $I(z)$ between the two groups. Moreover, w e use several statistics to quantify the separation between the Boundary and Normal groups. $\mathrm{Mean(B)/Mean(N)}$ and $\mathrm{Med.(B)/Med.(N)}$ measure the relative amplification of the central tendency of Boundary samples compared with Normal samples using the mean and median, respectively. W_1 denotes the Wasserstein-1 distance between the two empirical distributions in the original scale, whereas $W_1(\ln(1+\cdot))$ computes the same distance after a logarithmic transformation, thereby reducing the influence of extreme values while measuring whether the distributional separation persists. The Kolmogorov–Smirnov (KS) statistic measures the maximum difference between the empirical cumulative distributions of the two groups. Cliff's δ measures how consistently $I(z)$ is larger for Boundary samples than for Normal samples. A value of 1 means that every Boundary sample has a larger $I(z)$ than every Normal sample, while 0 indicates no systematic difference.

The results in Figure 3 and Table 6 show a pronounced separation between the Boundary and Normal groups. The mean curvature-aware influence is 1007.8 for Boundary samples and 19.1 for Normal

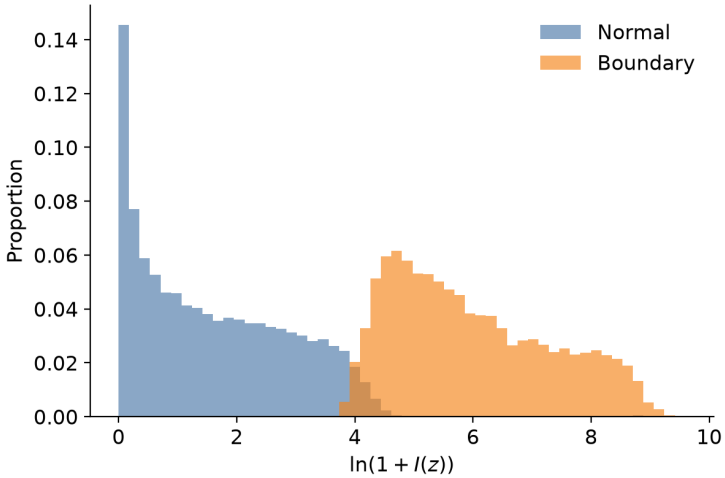


Figure 3: Distribution of $\ln(1 + I(z))$ for Normal and Boundary samples.

samples, corresponding to a mean ratio of 52.9. The median ratio is even larger, at 100.6, indicating that the difference is not driven solely by a small number of extreme samples. Consistent with these differences in central tendency, the 1-Wasserstein distance is 988.8 in the original scale and remains 4.37 after the logarithmic transformation $\ln(1 + I(z))$, indicating that the distributional separation persists after compressing the effect of large values. The Kolmogorov–Smirnov statistic is 0.954, further indicating a substantial difference between the two empirical distributions. Moreover, Cliff's δ is 0.994, indicating that almost all cross-group pairwise comparisons favor larger curvature-aware influence values for Boundary samples over Normal samples.

Accordingly, the empirical results support the relationship

$$I(z_b) > I(z_n), \qquad z_b \in \mathcal{D}_f^b,\ z_n \in \mathcal{D}_f^n, \tag{15}$$

for the overwhelming majority of Boundary–Normal sample pairs. This observation is consistent with the gradient-based characterization above, which predicts stronger sample-wise parameter responses for low-confidence Boundary samples than for high-confidence Normal samples.

To further examine whether the gradient-based proxy preserves the relative sample ordering when local curvature is taken into account, we compare the ranking induced by $G(z)$ with that induced by $I(z)$. On CIFAR-10 with ResNet-18, the two rankings are highly consistent, with a Spearman correlation of $\rho = 0.997$ and a Kendall correlation of $\tau = 0.954$. On CIFAR-100 with ViT, the agreement is even stronger, with $\rho = 0.9999$ and $\tau = 0.994$. These results indicate that the relative ordering captured by the gradient-based response $G(z)$ is largely preserved when local curvature is incorporated through the curvature-aware influence measure $I(z)$.

Overall, the pronounced distributional separation between Boundary and Normal samples in $I(z)$, together with the strong rank agreement between $I(z)$ and $G(z)$, provides empirical support that the sample-wise response heterogeneity predicted by Proposition 1 persists when local loss curvature is taken into account. These results are therefore consistent with the key implication of Proposition 1: Boundary samples generally exhibit larger sample-wise parameter responses and, under the proposed definition of forgetting difficulty, lower relative forgetting difficulty than Normal samples.

Observation 1 (Limitation of Coarse-Grained Unlearning)**.** *When the forget set $\mathcal{D}_f$ contains samples with heterogeneous forgetting difficulties, as characterized in Proposition 1, a coarse-grained unlearning strategy that applies an identical forgetting objective or update strength to all samples in $\mathcal{D}_f$ cannot adapt to their different forgetting requirements. Consequently, a globally calibrated update may lead to insufficient forgetting for relatively difficult samples, while imposing unnecessarily large changes on relatively easy samples. This creates a trade-off between under-unlearning and over-unlearning.*

The observation follows from the heterogeneous forgetting difficulty established in Proposition 1. However, coarse-grained unlearning applies the same forgetting objective or update strength indis-

criminately to all samples in $\mathcal{D}_f$, which cannot accommodate such heterogeneity. When the global update is relatively weak, samples with higher forgetting difficulty, particularly Normal samples, are more likely to remain insufficiently forgotten. Conversely, when the update is strengthened to sufficiently affect these samples, samples with lower forgetting difficulty, particularly Boundary samples, may undergo unnecessarily large parameter changes.

We refer to the former case as *under-unlearn*, where information associated with the forget set is not sufficiently removed and may remain exploitable by downstream inference or privacy attacks. The latter case is referred to as *over-unlearn*, where excessive parameter changes may unnecessarily alter the model and potentially compromise model utility or generalization performance.

B Detailed Related Work

Machine unlearning was first proposed by (Cao & Yang, 2015), introducing a method that stores statistical aggregates during training and updates them by subtracting the contribution of deleted data points before reconstructing the model. (Bourtoule et al., 2021) proposed the SISA training framework, which strategically limits the influence of data samples during the training process to accelerate the forgetting process.

For approximate unlearning, a variety of methods have been developed to efficiently erase data influence through post-training procedures. (Graves et al., 2021) records the data contained in each training batch and the corresponding parameter updates during the training phase. When deleting, they directly subtract from the final model parameters the updates produced by batches that contain the data to be forgotten. (Golatkar et al., 2020a) propose a model weight erasure method: the weights are first fine-tuned toward a state trained solely on retain data using a single Newton step, and then directional noise is added to disrupt the memory of forgotten data. They extended this framework to activation functions (Golatkar et al., 2020b). (Golatkar et al., 2021) propose Mixed Linear Forgetting, which decomposes a deep network into non-linear "core weights" that do not require forgetting and linear "user weights" that can be efficiently forgotten. (Guo et al., 2020) employ a Newton update removal mechanism, computing the influence vectors of the forgotten data points on the model and updating the parameters accordingly. (Izzo et al., 2021) propose projected residual updates, whose time complexity is linear in the dimension of the data to be deleted and independent of the dataset size. However, this method is primarily designed for linear models, which limits its scope of application. (Fan et al., 2024) propose SALUN, which selectively updates a subset of weights using weight significance masks, thereby efficiently removing the influence of specific samples or classes while preserving overall model performance. (Chundawat et al., 2023; Kurmanji et al., 2023) employ a teacher-student framework for machine unlearning. In their approaches, the former proposes a dual-teacher mechanism: the student model mimics the original model (the qualified teacher) on the retain set, and mimics a randomly initialized model (the unqualified teacher) on the forget set. The latter, in contrast, uses only the original model as the teacher, training the student model to stay close to the teacher on the retain set while moving away from the teacher on the forget set. (He et al., 2025) proposed *Natural Unlearning* to generate new samples by mixing each forget sample with related samples from the retain set, and then fine-tuning the original model on these mixed samples.

Another line of work addresses the class-wise forgetting task, which requires the model to completely remove knowledge of an entire concept or category. (Tarun et al., 2024) propose an efficient machine unlearning method called UNSIR, which learns "error-maximizing noise" for the target class and combines a "damage-repair" two-step weight update to achieve thorough forgetting of single or multiple classes in a single training pass. (Chen et al., 2023; 2025) shift the focus from parameter space to decision space. Boundary Shrink applies adversarial perturbations to assign incorrect labels to forget samples and fine-tunes the model, thereby compressing the decision boundary. In contrast, Boundary Expanding introduces a new neuron for a "shadow class" to disperse forget samples into other categories, actively expanding the decision boundary. For long-tailed distributions, LTMU (Chen et al., 2025) generates augmented features based on the similarity of tail-class samples and reassigns labels according to distance, enabling effective unlearning of tail classes. (Yu et al., 2026) proposes a dynamic loss reweighting method named FaLW, which mitigates issues of heterogeneous unlearning deviation and skewed unlearning deviation by estimating the unlearning deviation for each sample and introducing a class-wise balancing factor.

Beyond traditional deep learning models, machine unlearning research is gradually extending to broader modeling paradigms and diverse application scenarios. In large language models, unlearning can be used to remove the model's memory of useless or harmful knowledge (Jia et al., 2024; Liu et al., 2024; Wuerkaixi et al., 2025; Ji et al., 2024).

C Experimental Setup

Dataset. The basic information about the dataset we used is summarized as follows.

- *CIFAR-10* (Krizhevsky et al., 2009): it contains $60,000$ colored images of size $32 \times 32 \times 3$ with $10,000$ reserved for testing. There are 10 target classes with $5,000$ training images per class.
- *CIFAR-100* (Krizhevsky et al., 2009): it contains $60,000$ colored images with $10,000$ reserved for testing. There are 100 target classes with 500 training images per class.
- *Tiny-ImageNet*: Tiny ImageNet (Le & Yang, 2015) contains 120,000 colored images of size $64 \times 64 \times 3$ with 10,000 reserved for validation and 10,000 reserved for testing. There are 200 target classes with 500 training images per class.
- *Lacuna-10*: A subset selected from VGGFace (Cao & Yang, 2015). We randomly select 10 celebrities (classes) with at least 500 samples.

Baselines. We include and compare against various SOTA machine unlearning methods, which are briefly summarized as follows.

- **Original**: the model trained on all data $\mathcal{D}$ without performing any unlearning.
- **Retrain**: retraining the model from scratch without the forget set, i.e., only $\mathcal{D}_r$, which is assumed not to be viable in practice, but included as a reference.
- **Fine-tuning**: finetune the original model on $\mathcal{D}_r$.
- **Gradient Ascent** (Golatkar et al., 2020a): finetune the original model on the forget set $\mathcal{D}_f$ by negating the gradient.
- **Random Labels** (Graves et al., 2021; Hayase et al., 2020): finetune the original model on the random relabeled forget data.
- **Influence Unlearning** (Koh & Liang, 2017; Izzo et al., 2021): utilize an influence function to estimate the updates required for the model weights as a result of removing $\mathcal{D}_f$ from $\mathcal{D}$.
- **EU-k** (*Exact Unlearning-k*) and **CF-k** (*Catastrophic Forgetting-k*) (Goel et al., 2022): freeze the first k layers of the original model and either train the remaining layers from scratch on $\mathcal{D}_r$ or finetune the remain layers on $\mathcal{D}_r$.
- **ℓ_1-sparse** (Jia et al., 2023): introduce ℓ_1-norm for model weights and finetune the model over $\mathcal{D}_r$.
- **SalUn** (Fan et al., 2024): draw a parallel with input saliency in model explanation to only update influential weights filtered by the threshold, and integrate with Random Label.
- **Bad-Teaching** (Chundawat et al., 2023): encourage the student to move close to an incompetent/dumb teacher (randomly initialized model or random generators) for the forget set and close to the original model for the retain set.
- **SCRUB** (Kurmanji et al., 2023): move the student model away from the teacher model and formulate a min-max bi-optimization problem by extending contrastive learning.

Besides, we include the following specific unlearning method for the class-wise forgetting task.

- **Boundary Shrink** and **Boundary Expanding** (Chen et al., 2023): is designed to destroy the decision boundary by using adversarial samples for the forget data, and to exploit a new area in the decision space by assigning forget samples to an extra shadow class of the original model, with the replacement of the last layer.
- **UNSIR** (Tarun et al., 2024): learn an error-maximizing noise matrix for the class to be unlearned and uses it to impair the forget class and repair the original model.
- **LTMU** (Chen et al., 2025): Directionally rectifies the decision boundary of tail classes and guides their features to contract toward the feature space of neighboring classes for long-tailed unlearning.

For the **random-subset forgetting task**, we choose to forget about $\approx 10\%$ of the training data. Therefore, for the CIFAR-10 and CIFAR-100, we have $|\mathcal{D}_f| = 5K$. For the Tiny-ImageNet dataset,

we have $|\mathcal{D}_f| = 10K$. For the **class-wise forgetting task**, we used CIFAR-10 and CIFAR-100, and randomly chose class 4 as the forgetting class for both datasets. Consequently, we have $|\mathcal{D}_f| = 5K$ for CIFAR-10 and $|\mathcal{D}_f| = 500$ for CIFAR-100. In addition, for Lacuna-10, a subset selected from VGGFace (Cao & Yang, 2015) where each selected class contains at least 500 samples, we also randomly choose class 4 as the forgetting class.

Model. For ALLCNN (Springenberg et al., 2015), we reduce the number of layers and introduce batch normalization before each non-linearity layer, training the model from scratch. For ResNet, we adopt different architectures for different unlearning tasks. For Random-subset forgetting, we utilize the standard ResNet-18 architecture, starting with the pre-trained PyTorch model (trained on ImageNet) and fine-tuning it on the target dataset. For class-wise forgetting, we employ a CIFAR-adapted ResNet-18 architecture, where the first 7×7 convolutional layer is replaced with a 3×3 convolutional layer and the initial max-pooling layer is removed to accommodate 32×32 inputs, training the model from scratch as the original model. In addition, following (Kurmanji et al., 2023), we scale up to a larger setting by choosing larger models and tackling a larger classification task, that is, we used the pre-trained VGG-16 (Hayase et al., 2020) and ViT (Dosovitskiy et al., 2021) models, finetuning them on the CIFAR-100 dataset.

LiRA setup. To evaluate the privacy protection effectiveness of MU methods, we implemented the U-LiRA (Hayes et al., 2025) membership inference attack based on the log-likelihood ratio, which is considered the strongest per-sample attack benchmark for unlearning scenarios to date. The specific attack procedure is as follows:

- **Distribution Construction and Sample Filtering:** For each sample in the restricted sampling pool, we iterated over all shadow models and filtered them based on the sample's membership status in the models. Specifically, we collected the model outputs when the sample was in the forget set to construct the f_{out} distribution, and collected the model outputs when it was in the unseen sample set to construct the f_{in} distribution. Thanks to the aforementioned restricted sampling strategy, each target sample obtained a sufficient number of shadow model observations ($N = 128$) in both states, meeting the statistical requirements for precise Gaussian distribution fitting.
- **Attack Scoring:** For each target variant model, we used the log-likelihood ratio of the query sample under the two Gaussian distributions as the attack score.
- **Evaluation Metrics:**We conducted evaluations within the fixed pool, selecting samples that were neither in the forget set nor in the retain set as unseen test samples. To ensure unbiased evaluation metrics, we strictly implemented a 1:1 downsampling balancing strategy for positive and negative samples when calculating the AUC and attack accuracy.

We conduct experiments on CIFAR-10 using the standard ResNet-18 architecture. To simulate realistic scenarios and construct distributions for membership inference attacks, following (Hayes et al., 2025), we randomly sample 25,000 samples from the CIFAR-10 training set as the training set for each model.

U-LiRA partitions all base models into 128 shadow models and 128 target models, used for estimating distribution parameters and evaluating attack accuracy, respectively. To ensure statistical significance with a limited number of variants, we introduce a restricted sampling pool mechanism: we pre-sample a fixed pool of 1,000 samples from class 5 of CIFAR-10. Unlike the original U-LiRA's large-scale repeated sampling, we adopt a more efficient strategy — when constructing the forget set for each variant, we sample only 200 samples from the intersection of the pool and the base model's training set. Each base model randomly samples 5 different forget sets, generating 5 variants, resulting in 1,280 trained models in total. This strategy effectively increases the probability that samples from the fixed pool appear in the forget sets of various variants, ensuring that samples receive sufficient observations to accurately fit the distributions even when the total number of models is reduced.

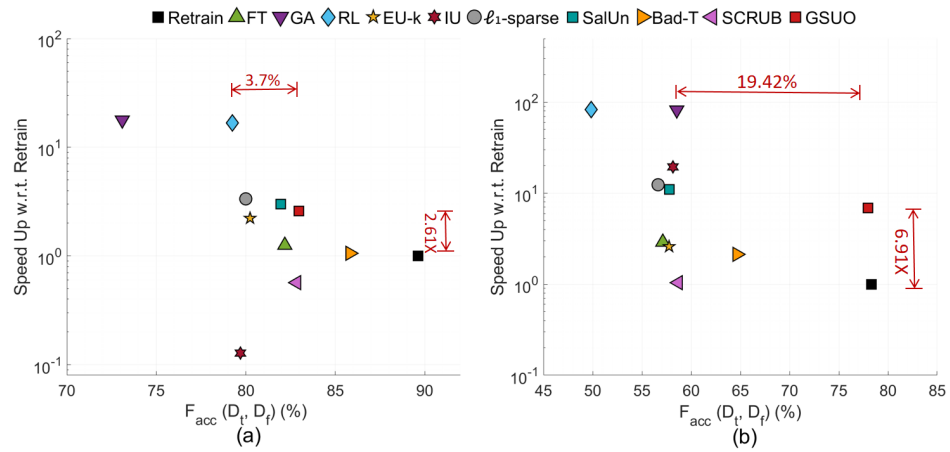


Figure 4: (a) Speedup w.r.t. Retrain versus F_{acc} for vit on CIFAR-100; (b) Speedup w.r.t. Retrain versus F_{acc} for ResNet50 on Tiny ImageNet.

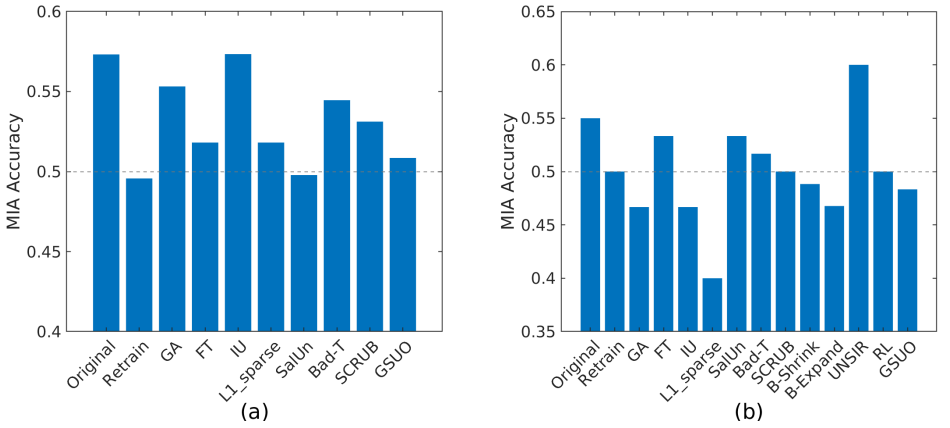


Figure 5: (a) MIA performance of various UM methods on the Random-subset forgetting task; (b) MIA performance of various UM methods on the class-wise forgetting task.

D More Experimental Results.

D.1 Random-subset Forgetting.

Figure 4 illustrates the trade-off between speedup and F_{acc} under two settings: (a) ViT on CIFAR-100, and (b) ResNet50 on Tiny ImageNet. F_{acc} jointly considers unlearning effectiveness and model generalization. On this metric, GSUO outperforms all baselines, achieving state-of-the-art performance. In contrast, methods such as FT, SalUN, and IU either retain high accuracy on the forget set (indicating incomplete unlearning) or suffer from a significant drop in accuracy on the test set (indicating poor generalization). While efficiency-oriented methods like GA achieve shorter runtime, their low F_{acc} reveal a failure to properly balance the trade-off between unlearning effectiveness and model generalization. In contrast, GSUO consistently maintains its superiority across different datasets and model architectures, demonstrating strong scalability and robustness. Overall, GSUO not only surpasses all competing methods in overall performance but also flexibly adapts to unlearning tasks of various scales, making it a practical and robust solution.

Figure 5(a) presents the MIA attack success rate of various unlearning methods on the Random-subset forgetting task, using an SVM attack model with loss as the attack feature. The MIA accuracy of GSUO is very close to 0.5, demonstrating strong protection of the privacy of the data intended to

Table 7: Comparison of MU methods for class-wise forgetting. Left: ALLCNN on CIFAR-10; Right: ResNet-18 on CIFAR-10

Method	ALLCNN on CIFAR-10			Resnet-18 on CIFAR-10		
	$Acc(\mathcal{D}_{ft})\downarrow$	$Acc(\mathcal{D}_{rt})\uparrow$	**Time (s)** $\downarrow$	$Acc(\mathcal{D}_{ft})\downarrow$	$Acc(\mathcal{D}_{rt})\uparrow$	**Time (s)** $\downarrow$
Original	83.40	82.48	189.32	85.12	85.40	1038.50
Retrain	0.00	85.10	428.64	0.00	84.07	459.81
GA	12.00	76.49	<u>5.62</u>	9.40	77.50	145.99
B-Shrink	10.40	75.43	48.59	1.20	77.08	309.46
B-Expand	10.90	78.34	9.71	11.50	64.90	27.32
LTMU	**0.00**	84.67	24.16	**0.00**	85.01	53.27
SCRUB	0.10	85.76	80.72	0.20	<u>86.02</u>	146.05
Bad-T	45.60	78.40	268.24	58.00	73.14	102.94
SalUn	**0.00**	85.47	50.37	**0.00**	85.32	22.03
RL	1.10	75.93	203.33	9.70	79.91	<u>6.74</u>
UNSIR	0.30	80.16	6.13	9.10	83.04	7.79
L1-sparse	**0.00**	77.58	28.93	**0.00**	81.26	19.53
FT	**0.00**	<u>86.02</u>	93.52	**0.00**	85.04	18.91
IU	13.10	77.22	**2.68**	12.30	80.77	**2.83**
GSUO	**0.00**	**86.92**	47.86	**0.00**	**86.24**	42.55

be forgotten and confirming truly effective unlearning. In contrast, methods such as IU and GA fail to achieve this level of privacy protection, with their MIA accuracy deviating significantly from 0.5.

D.2 Class-wise Forgetting.

Table 7 presents the results of class unlearning on CIFAR-10 using the ALLCNN and ResNet-18 models. Our method GSUO achieves 0% accuracy on the forget set test split, while attaining the highest accuracy on the retain set test split, surpassing all baselines. This demonstrates that GSUO effectively removes the targeted class information while well preserving the model's utility. In terms of efficiency, GSUO achieves nearly a $10\times$ speedup compared to retraining. Although it is slightly less efficient than methods such as IU and UNSIR, these baselines either suffer from high accuracy on the forget set, indicating incomplete unlearning or cause a significant drop in retain accuracy, leading to poor generalization. Overall, GSUO enables efficient unlearning and shows significant advantages in both effectiveness and efficiency, making it a practical and robust choice for real-world unlearning tasks.

Figure 5(b) presents the MIA attack success rate of various unlearning methods on the class-wise forgetting task. The MIA accuracy of GSUO is close to 0.5. Although methods such as IU, ℓ_1-sparse, and B-shrink achieve lower MIA accuracy, an extremely low ASR may not necessarily indicate more thorough forgetting. Instead, it may reflect the Streisand effect: which can provide more information about the forgetting data, i.e., all samples of $\mathcal{D}_f$ are predicted as one class. Therefore, considering their overall unlearning effectiveness and this potential privacy leakage risk, they still lag behind GSUO in terms of the trade-off between privacy protection and model utility.

D.3 Performance for imbalanced datasets.

To investigate the performance of GSUO on imbalanced data, we construct imbalanced datasets using an exponential decay strategy.

Dataset Specifically, we build a long-tailed dataset CIFAR-10-LT based on CIFAR-10, where a few head classes have abundant samples while many tail classes have only limited samples. The number of training samples per class is controlled by:

$$N_i = N_{\max} \times \left(\frac{N_{\min}}{N_{\max}}\right)^{\frac{i}{C-1}} \tag{16}$$

where C is the total number of classes, $N_{\max}$ and $N_{\min}$ denote the maximum and minimum sample sizes per class, and i indexes the classes sorted from the most frequent to the least frequent.

Table 8: Performance of Different UM Methods on CIFAR-10-LT.

Method	Head class				Tail class			
	Acc($\mathcal{D}_f$) ↓	**Acc($\mathcal{D}_r$)** ↑	**WAD** ↓	**HTR**	**Acc($\mathcal{D}_f$)** ↓	**Acc($\mathcal{D}_r$)** ↑	**WAD** ↓	**HTR**
Original	85.5	44.5	0.1997	1.9225	69.2	60.3	0.1761	1.4678
Retrain	0	64.0	0.1251	1.2605	0	73.1	0.1229	1.2069
GA	23.2	43.5	0.2250	1.5279	3.5	56.7	0.2220	1.9328
B-Shrink	0.2	37.7	0.2737	1.4455	2.2	55.6	0.2460	1.8436
B-Expand	12.0	42.3	0.1314	**0.9793**	2.2	55.9	0.2427	1.8105
LTMU	**0**	50.3	0.1940	1.5196	0.6	59.0	0.1974	1.7225
SCRUB	2.4	**65.8**	**0.1228**	1.2312	15.2	70.6	<u>0.1238</u>	1.2308
Bad-T	77.5	<u>62.8</u>	0.1445	1.2777	47.9	67.4	0.1963	1.4779
SalUn	**0**	48.6	0.1792	1.6420	5.6	71.0	0.1212	1.2490
RL	**0**	37.8	<u>0.1233</u>	0.8997	1.8	55.6	0.2401	1.7449
UNSIR	**0**	42.4	0.2222	0.3938	**0**	61.1	0.1831	**0.9182**
L1-sparse	**0**	44.8	0.2607	1.3600	**0**	73.4	0.1242	1.2903
FT	**0**	57.6	0.1633	1.3435	**0**	**75.8**	0.1083	<u>1.1547</u>
IU	23.2	44.4	0.2714	1.4436	**0**	62.2	0.1956	1.6674
GSUO	**0**	54.3	0.1428	<u>1.0579</u>	**0**	<u>74.6</u>	**0.1080**	1.1613

Evaluation Metric To address the potential performance imbalance between head classes (majority classes with many samples) and tail classes (minority classes with few samples) in long-tailed distributions, we introduce two fairness metrics to evaluate model performance across different classes.The first metric is the Weighted Accuracy Difference (WAD), defined as follows:

$$\mathrm{WAD} = \sqrt{\frac{C}{\sum_{c=1}^{C} p_c \cdot (\mathrm{Acc}_c - \mathrm{Acc})^2}} \tag{17}$$

where p_c is the sample proportion of class c, Acc_c is the model's accuracy on class c, and Acc is the overall average accuracy. A smaller WAD indicates more balanced performance across classes.The second metric is the Head-Tail Performance Ratio (HTR), defined as:

$$\mathrm{HTR} = \frac{\frac{1}{|G_h|}\sum_{c\in G_h} \mathrm{Acc}_c}{\frac{1}{|G_t|}\sum_{c\in G_t} \mathrm{Acc}_c} \tag{18}$$

where G_h and G_t denote the sets of head classes and tail classes, respectively, and $|G_h|$ and $|G_t|$ are their corresponding cardinalities. HTR measures the ratio of average accuracy between head and tail classes, reflecting the model's fairness toward tail classes.

Table 8 presents the performance of various unlearning methods on CIFAR-10-LT. We evaluate both head and tail classes. Results show that GSUO achieves complete unlearning while maintaining high retain test accuracy, surpassing both the original model and methods such as B-shrink, B-Expand, and LTMU. On tail classes, the test accuracy of GSUO even exceeds that of retrain.

From the perspective of fairness, GSUO also achieves competitive performance compared with the baseline methods. For the evaluation metrics, a smaller WAD indicates better performance, while an HTR value closer to 1 is preferable. As shown in Table 8, GSUO achieves an HTR of 1.0579 and a WAD of 0.1428 for the forgotten head classes, while obtaining an HTR of 1.1613 and a WAD of 0.1080 for the forgotten tail classes. In comparison, B-shrink, IU, and SalUN exhibit larger discrepancies in performance between the forgotten head and tail classes, indicating a less balanced unlearning performance.

D.4 Multi-Class Forgetting.

We evaluated multi-class forgetting performance of various machine unlearning methods on the CIFAR-10 dataset with the AllCNN architecture. We selected class 3 and class 4. The results are shown in Table 9. While methods like FT, EU-k, and ℓ_1-sparse achieve complete forgetting, others such as GA, IU, SalUn, Bad-T, SCRUB, and UNSIR fail to fully erase the target information. Notably, among all evaluated methods, GSUO achieves the best balance between forgetting quality

Table 9: Comparison of MU methods for multi-class forgetting on CIFAR-10 under ALLCNN.

Method	$Acc(\mathcal{D}_r)\uparrow$	$Acc(\mathcal{D}_{rt})\uparrow$	$Acc(\mathcal{D}_f)\downarrow$	$Acc(\mathcal{D}_{ft})\downarrow$	**Time (s)** $\downarrow$
Original	92.70	84.59	88.23	74.50	205.81
Retrain	99.96	89.89	0.00	0.00	401.91
FT	99.41	90.31	**0.00**	**0.00**	84.21
GA	71.73	67.35	3.39	3.10	21.02
RL	77.99	70.57	0.06	0.05	31.35
EU-k	96.54	90.09	**0.00**	**0.00**	398.99
IU	90.46	84.16	42.98	34.50	**4.22**
ℓ_1-sparse	86.04	82.55	**0.00**	**0.00**	422.61
SalUn	<u>99.58</u>	90.30	0.04	0.05	499.14
Bad-T	92.14	86.35	81.56	70.70	121.34
SCRUB	**99.72**	**92.26**	29.06	21.65	265.85
B-Shrink	70.46	65.86	4.08	3.75	29.73
B-Expand	90.35	83.50	28.88	23.10	228.35
UNSIR	87.33	83.29	0.06	0.05	<u>11.26</u>
GSUO	98.85	<u>90.75</u>	**0.00**	**0.00**	52.64

Table 10: Sensitivity Analysis of α

α	$Acc(\mathcal{D}_r)\uparrow$	$Acc(\mathcal{D}_t)\uparrow$	$Acc(\mathcal{D}_f)$	$\mid$ Diff $\mid\downarrow$	**Time (s)** $\downarrow$
0.1	99.59	82.21	82.40	0.19	4.73
0.2	99.53	82.32	81.56	0.76	4.63
0.3	99.51	82.28	81.34	0.94	4.51
0.4	99.51	82.28	81.44	0.84	4.49
0.5	99.46	82.18	81.96	0.22	4.48
0.6	99.47	82.11	82.12	0.01	4.77
0.7	99.46	82.29	82.60	0.31	4.74
0.8	99.40	82.02	83.24	1.22	4.47
0.9	99.32	81.65	83.08	1.43	4.58

and model utility. Furthermore, GSUO achieves an 8× speedup compared to retraining. Although IU has the greatest advantage in runtime, its forgetting quality is poor. In summary, GSUO achieves an optimal trade-off among forgetting completeness, model utility, and computational efficiency in multi-class forgetting tasks, validating its effectiveness in complex class-incremental scenarios.

D.5 Sensitivity Analysis.

In our experiments, α is a fixed hyperparameter. To further evaluate the sensitivity of our method to the hyperparameter α, we conducted experiments on CIFAR-10 with ResNet-18, where α is searched from 0.1 to 0.9. The experimental results are shown in the Table 10. The results indicate that, overall, our method performs well when α is in the range of 0.5 to 0.7, while the |Diff| value becomes larger whenα is too small or too large.

We adopt the maximum predicted probability as the confidence score, as it is a standard and widely used confidence measure in classification models. It directly reflects the model's confidence in its predicted class. Lower-confidence samples are generally more likely to lie near a decision boundary, whereas higher-confidence samples tend to lie farther from it. Table 11 presents the sensitivity analysis of the confidence threshold τ. The results indicate that the forgetting mitigation performance degrades when τ is set too large or too small. Specifically, when τ is too small, some Boundary samples are incorrectly included in the Normal set, leading to an under-unlearned state. Conversely, when τ is too large, some Normal samples are misclassified into the Boundary set, resulting in an over-unlearned state. Therefore, τ needs to be set at an appropriate intermediate value to achieve the optimal trade-off.

Table 11: Sensitivity analysis of the confidence threshold τ.

τ	Acc($\mathcal{D}_r$) ↑	Acc($\mathcal{D}_t$) ↑	Acc($\mathcal{D}_f$)	|Diff| ↓	Time (s) ↓
0.980	99.60	82.35	84.66	2.31	4.84
0.985	99.56	82.40	83.30	0.90	4.79
0.990	99.47	82.11	82.12	0.01	4.77
0.995	99.39	81.62	78.04	3.58	5.12
0.999	99.22	81.74	76.06	5.68	5.29

Table 12: The effectiveness of CO on the baseline method.

Method(+CO)	$Acc(\mathcal{D}_r)$ ↑	$Acc(\mathcal{D}_t)$ ↑	$Acc(\mathcal{D}_f)$	| Diff | ↓	**Time (s)** ↓
FT	**99.99**	**82.95**	82.40	<u>0.55</u>	17.50
GA	<u>99.50</u>	<u>82.86</u>	97.10	14.24	<u>5.34</u>
RL	84.68	74.69	83.68	8.99	5.85
EU-k	98.27	81.59	98.36	16.77	12.32
IU	97.50	81.06	97.82	16.76	65.80
ℓ_1-sparse	95.17	80.77	81.32	<u>0.55</u>	17.68
Salun	99.98	82.80	93.26	9.46	20.60
Bad-T	97.19	81.59	84.68	3.09	13.21
SCRUB	99.38	81.65	94.68	13.03	22.64
GSUO	99.47	82.11	82.12	**0.01**	**4.77**

D.6 The effectiveness of CO on the baseline method.

Table 12 shows the effectiveness of CO on the baseline method. With the integration of the CO stage, all baseline approaches gain performance improvement to varying extents. It confirms the effectiveness of the compensation mechanism. Nevertheless, GSUO achieves the smallest |Diff| while maintaining competitive test accuracy and substantially lower runtime. These results suggest that the combination of SGO and CO provides a favorable trade-off between forgetting effectiveness, model utility, and computational efficiency.

D.7 Visual Analysis.

Figure 6 illustrates the distribution of three samples before and after forgetting in Random-subset forgetting. Panels (a) and (b) show predicted distributions of randomly selected Boundary samples before and after forgetting, while (c) shows a randomly selected Normal sample. Boundary samples typically not only have high probability on their true class but also exhibit a spread of probability mass over several other classes. After forgetting, the boundary samples undergo a probability mass transfer across the boundary toward the most similar class. This change encourages the boundary samples to be naturally forgotten while causing minimal damage to neighboring sample points. Normal samples, which are originally located in the dense region of the entire class data, tend to remain correctly predictable after forgetting, but their confidence decreases, as shown in (c). Visualization of predicted distributions shows that GSUO successfully forgets the target samples.

Figure 7 illustrates the feature space distributions of the original model, the retrained model, and the unlearned model in class-wise forgetting. The experiments use CIFAR-10 with the AllCNN model. Features are projected onto a two-dimensional plane using t-SNE. The unlearned model's distribution closely resembles that of the retrained model. The intra-class structure of the forgotten data in the feature space tends to become loose, dispersedly distributed in the central vacant regions of the retain data. Most forgotten samples are predicted by the model as the class labels of the nearest retain data in the feature space. The above results demonstrate that GSUO achieves the desired forgetting effect in the class-wise forgetting scenario, making its distribution in the feature space highly similar to that of the retrained model, which serves as the gold standard.

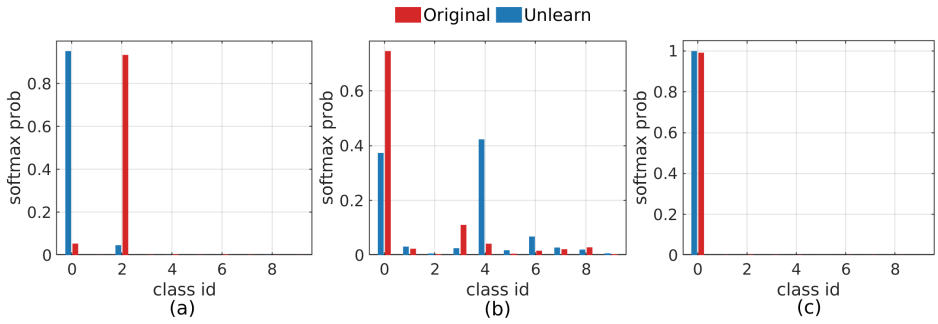


Figure 6: (a)-(c) Distribution of prediction changes on three forgotten samples (Random-subset forgetting).

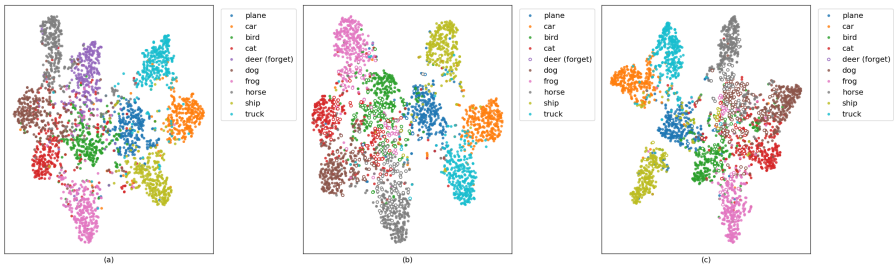


Figure 7: T-SNE visualization of feature representations of forgotten samples under class-wise forgetting on CIFAR-10 using AllCNN: (a) Original Model, (b) Retrained Model, and (c) Unlearned Model.