SELECTIVE MACHINE UNLEARNING FOR VISION TRANSFORMERS VIA ATTENTION-GUIDED CONTRASTIVE LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Machine unlearning enables pre-trained models to forget specific data in compliance with privacy regulations. While approximate unlearning methods work well for class-wise forgetting, random data forgetting, which involves removing individual samples while retaining others in the same class, remains highly challenging, especially for Vision Transformers (ViTs). Through selective masking experiments, we discover that masking a small proportion of the highest-attention patches preserves ViT's recognition capability while significantly degrading its memorization. Building on this insight, we propose Lethe, a contrastive unlearning method tailored for ViTs that uses masked images as positive logits and original images as negative logits, guiding the model to forget sample-specific details while retaining category-level outlines. Experimental results on multiple benchmarks and ViT architectures show that Lethe surpasses existing approximate unlearning methods, achieving the closest gap to the retain-only Retrain reference.

1 INTRODUCTION

Privacy regulations such as the General Data Protection Regulation (GDPR) (Hoofnagle et al., 2019) and the California Consumer Privacy Act (CCPA) (Nguyen, 2022) have introduced new challenges for Vision Transformers (ViTs) (Dosovitskiy et al., 2020). These laws grant users the right to withdraw their personal data, a withdrawal that requires not only erasing the data from all storage systems but also eliminating every trace of its influence on the model's training. *Machine Unlearning* (MU) (Tong et al., 2025a; Fan et al., 2024a; Liu et al., 2024; Kurmanji et al., 2023) using a reverse-learning process to erase the impact of specific data points on the model and thereby safeguard user privacy emerges as a promising solution. The most direct and effective method of MU is *Exact Unlearning* (Bourtoule et al., 2021), which retrains a new model from scratch using the remaining training set. However, this approach requires a substantial amount of computational resources. To address this challenge, *Approximate Unlearning* has been proposed, which can eliminate the impact of specific data without retraining from scratch (Chien et al., 2022).

Based on the degree of forgetting, MU can be divided into two types: 1) *Class-wise Forgetting* (Liu et al., 2024), which removes all data of a specific class from the model. For example, if a law bans recognizing a particular political symbol on a social media platform using a ViT for content moderation, the model must forget all images labeled with that symbol to comply. In this case, most existing approximate unlearning methods can achieve performance comparable to exact unlearning. 2) *Random Data Forgetting* (Tong et al., 2025a), which involves forgetting randomly selected samples from one or more classes. For example, on the same social media platform, users may request the removal of their specific images (e.g., pictures of their pet "cat") from the training data due to privacy concerns. The model must forget these images while still recognizing other users' content in the same class. This is more common in real-world applications, and is the focus of our work.

However, compared with class-wise forgetting, random data forgetting significantly increases the complexity, resulting in a substantial performance gap between existing approximate unlearning methods and exact unlearning (Fan et al., 2024b). The core challenge lies in the need to precisely "erase" individual samples within the same class while retaining other highly similar sam-

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.

ples (Tong et al., 2025a). For example, in the "cat" class, if two nearly indistinguishable images are present—one in the forget set and the other in the retain set—directly performing a forgetting operation on the model will weaken the forgetting effect. More critically, existing methods (Golatkar et al., 2020; Liu et al., 2024) generally overlook the unique characteristics of the self-attention mechanism in ViTs.

To address these challenges, we begin by asking a fundamental question that has been overlooked in existing unlearning literature: can we decouple a ViT's recognition capability from its memorization of specific training samples? Through systematic selective patch masking experiments, we discover a striking phenomenon: masking only 5% of the highest-attention patches with zero pixels preserves the model's recognition capability (test accuracy increases by 0.01%) while significantly degrading its memorization (membership inference attack drops by 14.33%). Unlike CNNs, ViTs possess an explicit attention mechanism that assigns interpretable importance scores to each image patch, enabling us to precisely locate and mask the regions encoding sample-specific details. This architectural property, unique to ViTs, has not been exploited by any prior unlearning method.

Building on this insight, we propose Lethe, a contrastive unlearning framework that, for the first time, leverages the attention mechanism of ViTs to guide selective forgetting. Unlike existing methods that treat the model architecture as agnostic and apply generic forgetting operations, Lethe explicitly constructs positive logits from attention-masked images and negative logits from original images, steering the model to retain category-level outlines while erasing sample-specific details. This attention-guided contrastive design directly addresses the core difficulty of random data forgetting — precisely removing individual samples without degrading highly similar within-class samples — in a manner that is naturally aligned with the ViT's internal representation structure. We summarize our contributions below:

- We reveal a previously unexplored property of ViTs: masking a small proportion of the highest-attention patches can decouple recognition capability from sample-specific memorization, which motivates a new perspective for designing ViT-specific unlearning methods.
- We propose Lethe, a contrastive unlearning framework that leverages the ViT attention mechanism to guide selective forgetting of individual samples while preserving category-level knowledge.
- We conduct extensive experiments to verify our method. Experiments demonstrate that Lethe surpasses existing state-of-the-art methods. For instance, for DeiT-T on Tiny-ImageNet, Lethe achieves the smallest Average Gap of 2.79% relative to Retrain.

2 PRELIMINARIES

Problem Formulation. Let the complete training dataset be $D = \{(x_i, y_i)\}_{i=1}^{N}$, consisting of N samples, where x_i denotes the i-th sample and $y_i \in \{1, 2, \ldots, n\}$ is its associated class label. The forget set $D_f \subseteq D$ represents a subset of D that needs to be removed from the trained model, while its complement, the retain set D_r, contains the data to be preserved, satisfying $D_f \cap D_r = \varnothing$ and $D_f \cup D_r = D$. Machine unlearning (MU) in image classification can be categorized based on the composition of D_f: class-wise forgetting and random data forgetting. In class-wise forgetting, D_f comprises solely samples from a single class, with the objective of eliminating the influence of that entire class on the model. In random data forgetting, D_f includes randomly selected samples from one or multiple classes, aiming to remove their influence on the model. Let θ_p denote a shared pretrained initialization. Prior to unlearning, the original model f_{θ_o} is obtained by training from θ_p on the complete dataset D, while the retrained model f_{θ_r} is independently initialized from the same θ_p and trained only on D_r. Retraining on D_r is commonly used as a reference for evaluating approximate unlearning (Nguyen et al., 2022; Tong et al., 2025a). In our setting, f_{θ_r} is therefore treated as a retain-only reference under a fixed pretrained initialization. This reference measures the removal of influence introduced by D_f during downstream training while holding the upstream pretraining stage fixed. Approximate unlearning aims to produce an unlearned model f_{θ_u} by removing such influence from f_{θ_o}, thereby approximating f_{θ_r} with reduced computational cost.

Evaluation Metrics. Machine unlearning requires balancing two objectives: erasing the influence of forgotten data and preserving the model's utility. Following prior works (Fan et al., 2024b; Liu et al., 2024; Tong et al., 2025a), we evaluate unlearning performance from four aspects: (1) **Forget Accuracy (FA)**: the model's accuracy on the forget set after unlearning, measuring the degree

Table 1: TA and MIA under different masking ratios. Additional experimental results are provided in the Appendix A.8.

Ratio	Zero Noise		Gaussian Noise	
	TA	MIA	TA	MIA
0%	81.24	24.49	81.24	24.49
5%	81.25$^{\uparrow 0.01}$	10.16$^{\downarrow 14.33}$	83.59$^{\uparrow 2.35}$	14.06$^{\downarrow 10.43}$
10%	79.69$^{\downarrow 1.55}$	18.75$^{\downarrow 5.74}$	81.25$^{\uparrow 0.01}$	19.53$^{\downarrow 4.96}$
20%	68.75$^{\downarrow 12.49}$	38.24$^{\uparrow 13.75}$	69.53$^{\downarrow 11.71}$	37.50$^{\uparrow 13.01}$
30%	53.91$^{\downarrow 27.33}$	52.34$^{\uparrow 27.85}$	57.81$^{\downarrow 23.43}$	56.25$^{\uparrow 31.76}$



Figure 1: Visualization of the original image and the masked image (with 5% masking). The class is "rocket."

of forgetting; (2) **Retain Accuracy (RA)**: the model's accuracy on the retain set after unlearning, measuring utility preservation on training data; (3) **Test Accuracy (TA)**: the model's accuracy on a holdout test set, reflecting generalization ability after unlearning; (4) **Membership Inference Attack (MIA)** (Shokri et al., 2017): the success rate of inferring whether a specific sample belongs to the original training set, measuring residual privacy leakage. Crucially, the ideal values for these metrics are not simply maximizing or minimizing them; instead, they should exhibit minimal deviation from those achieved by the Retrain baseline. To facilitate comparison, we additionally report the **Average Gap (AG)**, computed as the mean absolute difference between each method and the Retrain baseline across the four metrics. AG serves as an auxiliary metric that provides a concise summary of overall proximity to the Retrain reference under the shared pretrained initialization, where a lower value indicates better unlearning performance.

3 Exploring the Memorization and Recognition Abilities of ViTs

To systematically investigate the memorization and recognition capabilities of ViTs, we conduct a series of experiments where we selectively masked the most attended image patches—identified through the highest attention scores—and evaluated the model's performance in terms of test accuracy (TA) and the success rate of membership inference attacks (MIA). The memory and recognition capabilities of ViTs are defined as follows:

Recognition Capability of ViTs: *In the task of image classification, the recognition capability of a Vision Transformer (i.e., its ability to accurately identify and classify visual patterns in unseen data) is typically reflected by the model's Top-1 classification accuracy on the test set (test accuracy, TA).*

Memorization Capability of ViTs: *The memory capability of a Vision Transformer reflects its degree of memorization of training data. In image classification tasks, this ability can be quantitatively evaluated through the success rate of Membership Inference Attacks (MIA) — that is, the probability that the model successfully identifies the true training status (whether or not it belongs to the training data) of data samples on the forget set.*

Specifically, we conduct experiments on the CIFAR-100 dataset using the DeiT-T model, with a forgetting scenario of randomly forgetting 10% of the data (more exploring experiments can be found in the Appendix A.8). We first train a retrain model using the retained data. Subsequently, we apply different masking ratios to the images in the test set and the forget set. We then measure the model's Top-1 classification accuracy on the test set (i.e., test accuracy, TA) and the success rate of membership inference attacks (MIA) on the forget set. As shown in Table 1, we employ two masking methods: setting pixels to zero and applying Gaussian noise. When the masking ratio is 5%, we find that setting pixels to zero actually increases the TA by 0.01% compared to using the original images. This indicates that even after masking the pixels corresponding to the highest-attention patches, the model can still recognize the class of the image, and its recognition ability remains intact. Moreover, the success rate of MIA drops from 24.49% to 10.16%, suggesting that the model finds it more difficult to determine whether the masked image was used for training, indicating a significant reduction in the model's memory of the image.

In fact, the slight TA increase at 5% masking may stem from a regularization effect, as removing a small number of high-attention patches reduces overfitting to fine-grained details. Importantly, this does not conflict with our key observation: regardless of whether TA remains stable or marginally improves, the model retains its recognition capability after masking, while its memorization of specific training samples is significantly weakened. Notably, the decrease in MIA at low ratios (5%-

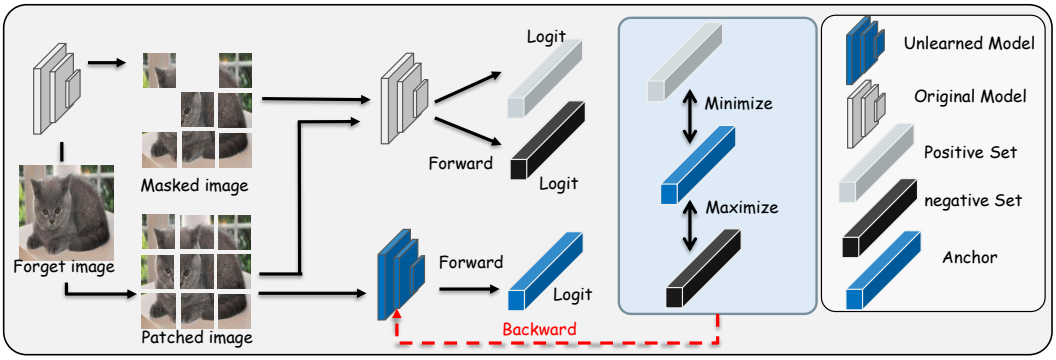


Figure 2: Overview of contrastive unlearning in Lethe. The forget image is forwarded through the unlearned model to produce anchor logits. Its masked version (key patches set to zero) and original version are forwarded through the original model to generate positive and negative logits, respectively. The objective is to maximize anchor-positive similarity while minimizing anchor-negative similarity.

10%) is due to the attention-guided mask specifically weakening sample-specific memory traces, causing the confidence distribution of masked forget images (non-members) to shift toward that of member samples, resulting in more false positives for the attacker (thus lowering MIA). In contrast, at 20% or 30% ratios, the mask excessively disrupts the category outlines (with TA sharply dropping by 12.49%), amplifying the response gap between members and non-members, making non-members easier to correctly identify (increasing the true negative rate and raising MIA).

We present the masked images and the original images in Figure 1. Taking the "rocket" class as an example, the masked patches mainly cover the detailed parts of the rocket, while the main outline of the rocket is still preserved. As a result, the model can still correctly identify the class of the masked image. However, due to the introduction of a small amount of noise, the model finds it difficult to determine whether the masked image was used for training. To further confirm the generality of this finding, we provide additional exploring experiments on Tiny-ImageNet with varying forgetting ratios in Appendix A.8, where consistent trends are observed. Based on the above insights, we will introduce our method Lethe in the next section.

4 LETHE

We propose Lethe, a novel contrastive unlearning approach tailored for Vision Transformers (ViTs). As shown in Figure 2, it enables selective forgetting of designated samples while preserving performance on retained samples. Our method leverages the attention mechanism to identify and mask critical image regions, guiding the model to forget targeted information through a contrastive learning framework.

4.1 ATTENTION-GUIDED MASKING

For a given input image x, we extract the attention maps from the last attention layer of the original pre-trained model. These attention maps, denoted as $A \in \mathbb{R}^{B\times H\times(N+1)\times(N+1)}$, where B is the batch size, H is the number of attention heads, and N is the number of patches (with an additional class token), represent the pairwise attention weights between tokens.

To identify the most informative patches, we compute the attention weight from the class token to each patch:

$$a_i = \frac{1}{H}\sum_{h=1}^{H} A_{h,0,i+1}, \tag{1}$$

for each patch i ($1 \le i \le N$), where $A_{h,0,i+1}$ is the attention weight from the class token (position 0) to patch token i (position $i+1$) in head h. We select the top k patches with the highest attention scores a_i, where $k = \lfloor \rho \cdot N \rfloor$, and ρ is the masking ratio. Let $\mathcal{S}_\rho = \{i \mid a_i$ ranks in the top $k\}$ denote the set of selected patch indices. Since each patch i corresponds to a spatial region P_i in the original

Algorithm 1 The Overall Pipeline of Unlearning

Input: Pre-trained ViT model f_{θ_o} with parameters θ_o, Forget set $D_f = \{(x_f, y_f)\}$, Retain set $D_r = \{(x_r, y_r)\}$,Forget Set Training Epochs E_f, Retain Set Training Epochs E_r, Learning rate η.
Output: Unlearned model f_{θ_u} .

```
Initialize θ_u ← θ_o
for t ∈ [0, ..., E_f − 1] do
    Mask image (x_f, y_f) in forget set following Eq.(1) and Eq.(2)
    Compute Z, Z_p, Z_n following Eq.(3)
    Compute L_CL following Eq.(4)
    Update θ_u ← θ_u − η∇L_CL
end for
for t ∈ [0, ..., E_r − 1] do
    Compute Cross-Entropy loss L_CE = CE(f_{θ_u}(x_r), y_r)
    Update θ_u ← θ_u − η∇L_CE
end for
Return Unlearned model f_{θ_u} with parameters θ_u
```

image (determined by the patch size and image resolution), we define the masking map $M_\rho(x; A)$ as:

$$M_\rho(x; A)_{(u,v)} = \begin{cases} 0, & \text{if } (u, v) \in P_i \text{ for some } i \in \mathcal{S}_\rho, \\ x_{(u,v)}, & \text{otherwise,} \end{cases} \tag{2}$$

where (u, v) denotes the pixel coordinates. The masked image is then obtained as $x_m = M_\rho(x; A)$, which sets all pixels within the high-attention patches to zero while preserving the remaining regions.

4.2 Contrastive Unlearning Loss

In the previous section, we explored the recognition and memorization capabilities of ViT: after masking the patch with the highest attention score in an image, ViT can still identify the class of the image, but has difficulty determining whether the image was used for training. Based on these insights, we can implement selective forgetting through contrastive learning. Specifically, we denote f_{θ_u} as the current model being unlearned and f_{θ_o} as the original pre-trained model. For an input image x, we first compute the anchor and the positive and negative sets as follows:

$$\mathcal{Z} = f_{\theta_u}(x), \mathcal{Z}_p = f_{\theta_o}(x_m), \mathcal{Z}_n = f_{\theta_o}(x), \tag{3}$$

where $\mathcal{Z}$ is the logit of the original image on the current model, $\mathcal{Z}_p$ is the logit of the masked image on the original model, and $\mathcal{Z}_n$ is the logit of the original image on the original model. The contrastive loss is defined as:

$$\mathcal{L}_{CL} = -\log \frac{\exp(sim(\mathcal{Z}, \mathcal{Z}_p)/\tau)}{\exp(sim(\mathcal{Z}, \mathcal{Z}_p)/\tau) + \exp(sim(\mathcal{Z}, \mathcal{Z}_n)/\tau)}, \tag{4}$$

where $\text{sim}(\cdot, \cdot)$ is the cosine similarity, and τ is a temperature parameter. This loss encourages the current model's logit $\mathcal{Z}$ to be closer to $\mathcal{Z}_p$ and farther from $\mathcal{Z}_n$. By optimizing this loss function, the model can remember class-level features while forgetting detailed features, thereby achieving selective forgetting of specific samples.

4.3 The Overall Pipeline of Unlearning

We present the overall pipeline of unlearning in Algorithm 1. The unlearning process of Lethe involves training the model by processing the forget and retain sets. Specifically, during the forget set training phase, for each batch in the forget set, we compute the contrastive loss $\mathcal{L}_{CL}$ and update the model parameters θ_u to achieve selective forgetting of specific samples. During the retain set training phase, for each batch in the retain set, we compute the classification loss $\mathcal{L}_{CE}$ and update the model parameters θ_u to maintain the model's performance on the retain set. The number of training iterations on the forget and retain sets is determined empirically; we provide a detailed analysis of its effect in Appendix A.14.

4.4 Optimization Guarantee

We analyze the convergence of the Lethe method under the Lipschitz smoothness assumption.

Assumption 4.1. *The contrastive unlearning loss $\mathcal{L}_{CL}$ is L-smooth, i.e., there exists a constant $L > 0$ such that for any model parameters θ, θ':*

$$\|\nabla_\theta \mathcal{L}_{CL}(\theta) - \nabla_\theta \mathcal{L}_{CL}(\theta')\| \leq L\|\theta - \theta'\|. \tag{5}$$

This assumption is mild and is commonly satisfied in practice for deep models with Lipschitz-activated layers and bounded inputs; consequently, it is widely adopted in convergence analyses(Tong et al., 2025b).

Theorem 4.2. *Under Assumption 4.1, if the learning rate satisfies $\eta < \frac{1}{L}$, then the gradient descent update $\theta_{t+1} = \theta_t - \eta \nabla \mathcal{L}_{CL}(\theta_t)$ ensures:*

$$\min_{0 \leq t \leq T} \|\nabla \mathcal{L}_{CL}(\theta_t)\|^2 \leq \frac{2[\mathcal{L}_{CL}(\theta_0) - \mathcal{L}^*_{CL}]}{\eta T}. \tag{6}$$

where $\mathcal{L}^*_{CL}$ is the global minimum value of the loss.

The proof can be found in the Appendix A.4. Although stated under full-batch gradient descent for clarity, this result extends to the SGD setting under standard bounded variance assumptions. This theorem guarantees that Lethe converges to a stationary point at a rate of $\mathcal{O}(1/T)$, meaning that the optimization of anchor logits toward alignment with the positive logits and separation from the negative logits is well-grounded. This also confirms that our proposed contrastive loss $\mathcal{L}_{CL}$, despite involving cross-model cosine similarity and asymmetric attention-guided masking, admits well-behaved optimization dynamics. In practice, the forgetting stage needs only a handful of epochs to converge, efficiently erasing the influence of the forget-set samples without drastically altering the overall model parameters.

5 Experiments

5.1 Experimental Setup

Datasets and Networks. The datasets used in the experiments are CIFAR-10 (Krizhevsky et al., 2009), CIFAR-100 (Krizhevsky et al., 2009), SVHN (Netzer et al., 2011), and Tiny-ImageNet (Le & Yang, 2015). To validate Lethe, we select various popular vision transformer models, including ViT-T/S/B, DeiT-T/S/B, and Swin-T/S. All models are initialized from publicly available pretrained checkpoints corresponding to their respective architectures. For each architecture, the original model and the Retrain reference use the same pretrained initialization: the original model is trained on the complete dataset D, whereas Retrain is independently initialized from the same checkpoint and trained only on the retain set D_r. All approximate unlearning methods start from the same original model f_{θ_o}.

Baselines. To evaluate our method comprehensively, we select multiple baselines, including: (1) **Retrain**. (2) **Fine-Tuning (FT)** (Warnecke et al., 2021; Golatkar et al., 2020). (3) **Gradient Ascent (GA)** (Graves et al., 2021; Thudi et al., 2022). (4) **Influence Unlearning (IU)** (Koh & Liang, 2017; Izzo et al., 2021). (5) **Random Labels (RL)** (Golatkar et al., 2020). (6) **ℓ_1-sparse** (Liu et al., 2024). (7) **SalUn** (Fan et al., 2024b). (8) **LoTUS** (Spartalis et al., 2025). (9) **Q-MUL** (Tong et al., 2025a). The introduction of baselines can be found in the Appendix A.3.

Evaluation Metrics. We adopt the evaluation metrics defined in Section 2. The MIA implementation follows the same protocol as ℓ_1-sparse (Liu et al., 2024) and SalUn (Fan et al., 2024b). More experimental details can be found in the Appendix A.2.

5.2 Main Results

Table 2 compares different MU methods across eight Vision Transformer architectures on Tiny-ImageNet using FA, RA, TA, MIA, and AG. Lethe achieves the lowest AG across all evaluated architectures, demonstrating a favorable balance between forgetting effectiveness and model utility.

Table 2: Performance of various MU methods on Tiny-Imagenet. **Bold** indicates the best performance and underline indicates the runner-up. A performance gap against Retrain is provided in (•). The proportion of forgotten data samples is 10%.

Method	FA	RA	TA	MIA	AG↓	FA	RA	TA	MIA	AG↓
	ViT-T					**ViT-S**				
Retrain	78.89	95.77	79.58	35.78	0	86.52	99.57	86.32	24.15	0
FT	80.43(1.54)	87.57(8.26)	80.68(1.10)	37.78(2.00)	3.23	84.18(2.24)	99.00(0.57)	82.42(3.90)	30.13(5.58)	3.07
GA	76.10(2.79)	76.98(18.79)	74.87(4.71)	47.72(11.94)	9.56	96.76(10.24)	96.93(2.64)	87.62(1.30)	12.57(11.58)	6.44
IU	71.35(7.54)	73.49(22.28)	71.07(8.51)	46.62(10.84)	12.29	96.27(9.75)	96.71(2.86)	87.34(1.02)	13.29(10.86)	6.12
RL	79.50(0.61)	87.20(8.57)	80.16(0.58)	36.93(1.15)	2.73	90.99(4.47)	98.75(0.82)	86.80(0.48)	33.72(9.57)	3.84
ℓ_1-sparse	80.54(0.56)	89.00(6.77)	80.72(1.14)	36.37(0.59)	2.27	84.10(2.42)	98.98(0.59)	82.22(4.10)	30.79(6.64)	3.43
SalUn	79.45(0.56)	89.38(6.39)	80.04(0.46)	38.63(2.85)	2.56	88.01(1.49)	98.72(0.85)	86.28(0.04)	35.07(10.92)	3.33
LoTUS	76.10(2.79)	76.97(18.80)	75.66(3.92)	48.65(12.87)	9.60	96.79(10.27)	96.94(2.63)	87.26(0.94)	12.85(11.30)	6.29
Q-MUL	81.27(2.38)	88.31(7.46)	80.35(0.77)	37.02(1.24)	2.96	88.46(1.94)	98.75(0.82)	85.43(0.89)	33.13(8.98)	3.16
Lethe	80.09(1.20)	91.55(4.22)	80.26(0.68)	36.81(1.03)	**1.78**	91.14(4.62)	98.79(0.78)	85.92(0.40)	22.04(2.11)	**2.00**
	ViT-B					**DeiT-T**				
Retrain	87.62	99.98	87.92	23.89	0	76.53	91.65	76.92	40.06	0
FT	83.05 (4.57)	99.74 (0.24)	81.02(6.90)	30.51 (6.62)	4.58	86.55(10.02)	96.48(4.83)	75.76(1.16)	30.93(9.13)	6.29
GA	99.99(12.37)	99.96(0.02)	88.62(0.70)	1.87(22.02)	8.78	93.40(16.87)	93.27(1.62)	77.46(0.54)	22.90(17.16)	9.05
IU	99.99(12.37)	99.95(0.03)	88.22(0.30)	2.35(21.54)	8.56	90.64(14.11)	91.58(0.07)	75.70(1.22)	22.44(17.62)	8.26
RL	94.50(6.88)	99.97(0.01)	86.74(1.18)	51.01(27.12)	8.80	85.26(8.73)	95.79(4.14)	76.24(0.68)	36.18(3.88)	4.36
ℓ_1-sparse	83.07(4.55)	99.72(0.26)	80.82 (7.10)	29.48(5.59)	4.38	86.55(10.02)	96.50(4.85)	75.90(1.02)	30.79(9.27)	6.29
SalUn	96.94(9.32)	99.95(0.03)	87.12(0.80)	38.20(14.31)	6.12	84.05(7.52)	95.18(3.53)	75.66(1.26)	35.79(4.27)	4.15
LoTUS	99.25(11.63)	99.13(0.85)	86.54(1.38)	4.91(18.98)	8.21	91.07(14.54)	91.25(0.40)	76.84(0.08)	26.98(13.08)	7.03
Q-MUL	90.66(3.04)	99.89(0.09)	81.34(6.58)	32.76(8.87)	4.65	83.17(6.64)	94.78(3.13)	76.36(0.56)	35.93(4.13)	3.62
Lethe	86.03(1.59)	99.19(0.79)	82.54(5.35)	25.06(1.17)	**2.23**	80.90(4.37)	94.09(2.44)	75.04(1.88)	37.60(2.46)	**2.79**
	DeiT-S					**DeiT-B**				
Retrain	85.00	99.34	85.58	25.83	0	90.54	99.95	90.60	19.95	0
FT	89.96(4.96)	98.71(0.63)	86.12(0.54)	22.17(3.66)	2.45	93.02(2.48)	99.51(0.44)	90.70(0.10)	17.01(2.94)	1.49
GA	93.42(8.42)	93.69(5.65)	87.02(1.44)	22.20(3.63)	4.79	95.06(4.52)	95.50(4.45)	91.62(1.02)	16.32(3.63)	3.41
IU	92.48(7.48)	93.21(6.13)	86.42(0.84)	23.76(2.07)	4.13	94.82(4.28)	95.40(4.55)	91.38(0.78)	16.92(3.03)	3.16
RL	85.16(0.16)	98.41(0.93)	85.70(0.12)	40.73(14.90)	4.03	90.22(0.32)	99.04(0.91)	91.20(0.60)	25.41(5.46)	1.82
ℓ_1-sparse	82.17(2.83)	99.14(0.20)	80.78(4.80)	31.97(6.14)	3.49	88.15(2.39)	99.74(0.21)	87.70(2.90)	24.59(4.64)	2.54
SalUn	85.77(0.77)	97.82(1.52)	85.32(0.26)	33.32(7.49)	2.51	90.31(0.23)	98.86(1.09)	91.00(0.40)	23.25(3.30)	1.26
LoTUS	93.43(8.43)	93.70(5.64)	86.84(1.26)	21.71(4.12)	4.86	95.08(4.54)	95.51(4.44)	91.70(1.10)	16.30(3.65)	3.43
Q-MUL	85.98(0.98)	97.82(1.52)	84.42(1.16)	34.67(8.84)	3.13	89.73(0.81)	99.64(0.31)	90.17(0.43)	25.91(5.96)	1.88
Lethe	87.87(2.87)	98.31(1.03)	85.34(0.24)	25.56(0.27)	**1.10**	90.70(0.16)	99.40(0.55)	89.38(1.22)	20.28(0.33)	**0.57**
	Swin-T					**Swin-S**				
Retrain	84.89	99.13	85.56	26.36	0	88.22	99.92	88.72	21.08	0
FT	78.92(5.97)	96.99(2.14)	78.68(6.88)	35.15(8.79)	5.95	81.74(6.48)	98.50 (1.42)	80.36(8.36)	30.87(9.79)	6.51
GA	96.38(11.49)	96.41(2.72)	87.18(1.62)	14.35(12.01)	6.96	98.94(10.72)	99.02(0.90)	89.42(0.70)	5.94(15.14)	6.87
IU	90.13(5.24)	91.42(7.71)	82.80(2.76)	23.08(3.28)	4.75	98.80(10.58)	98.91(1.01)	88.80(0.08)	6.63(14.45)	6.53
RL	88.64(3.75)	98.40 (0.73)	86.74(1.18)	38.65 (12.29)	4.49	83.12(5.10)	99.52(0.40)	86.00(2.72)	54.35(33.27)	10.37
ℓ_1-sparse	80.85 (4.04)	97.90 (1.23)	79.82(5.74)	33.05(6.69)	4.43	81.55(6.67)	98.37(1.55)	80.82 (7.90)	31.14(10.06)	6.55
SalUn	87.76(2.87)	98.11(1.02)	85.64(0.08)	38.29(11.93)	3.98	88.06(0.16)	99.37(0.55)	87.10(1.62)	35.00(13.92)	4.06
LoTUS	96.24(11.35)	96.47(2.66)	87.90(2.34)	14.51(11.85)	7.05	98.94(10.72)	99.02(0.65)	89.26(0.54)	5.84(15.24)	6.79
Q-MUL	89.39(4.50)	99.20(0.07)	84.83(0.73)	35.07(8.71)	3.50	87.81(0.41)	99.22(0.70)	85.74(2.98)	30.32(9.24)	3.33
Lethe	88.53(3.64)	98.16 (0.97)	84.66 (0.90)	25.12 (1.24)	**1.69**	91.50(3.28)	99.27(0.65)	86.14(2.58)	19.74(1.34)	**1.96**

For the larger ViT-B and DeiT-B models, Lethe obtains FA values of 86.03% and 90.70%, with gaps of only 1.59 and 0.16 percentage points from Retrain, respectively. In terms of retained-data performance, Lethe achieves an RA of 91.55% on ViT-T, which is closer to Retrain than SalUn's 89.38%. Although its TA gap is slightly larger than that of SalUn in some settings, Lethe provides a substantially lower overall AG. Lethe also closely matches the MIA behavior of Retrain across the ViT, DeiT, and Swin model families. For example, on ViT-B, its MIA is 25.06%, corresponding to a gap of only 1.17 points, whereas FT produces a gap of 6.62 points.

Lethe remains consistently effective on the Swin architectures. On Swin-T, it achieves an AG of 1.69%, compared with 3.98% for SalUn and 3.50% for Q-MUL. On Swin-S, its AG is 1.96%, while SalUn and Q-MUL obtain 4.06% and 3.33%, respectively. The corresponding MIA gaps of Lethe are only 1.24 and 1.34 percentage points. Overall, these results show that Lethe consistently provides the closest approximation to the Retrain reference across different architectures while preserving performance on retained and test data. Additional experimental results are provided in Appendices A.5 and A.6.

5.3 Ablation Studies

Effect of Masking Ratio. As shown in Table 3, we demonstrate the impact of different masking rates. When the masking ratio is 5%, the average gap reaches its minimum value of 2.79%, which is lower than those of other ratios (3.21% for 10%, 3.10% for 20%, and 3.19% for 30%). This indicates that 5% is the optimal masking ratio, achieving the best forgetting effect. This finding is also consistent with the experimental results in Table 1: the 5% masking ratio can preserve the model's recognition ability while reducing its memorization ability. However, higher ratios will disrupt the category outlines, leading to a decline in the forgetting effect.

Effect of Mask Padding Type. In Table 3, we show the impact of different mask padding Types. Specifically, when using Zero padding, the model achieves a FA of 80.90%, which is 4.37% higher than that of Retrain. RA increases to 94.09%, representing an improvement of 2.44%. TA, however, decreases slightly to 75.04%, a drop of 1.88%. Meanwhile, the MIA is reduced to 37.60%. The AG remains at a low level of 2.79%. In contrast, Gaussian padding improves FA to 82.27% and RA to 94.64%. However, TA slightly drops to 75.24% and the MIA is even lower at 36.27%. Meanwhile, the AG increases to 3.55%. This indicates that Gaussian padding introduces more noise compared to Zero padding, thereby reducing the forgetting effect.

Table 3: Lethe under different mask padding types and ratios (DeiT-T on Tiny-ImageNet).

Type	FA	RA	TA	MIA	AG↓
Retrain	76.53	91.65	76.92	40.06	0
Zero	80.90(4.37)	94.09(2.44)	75.04(1.88)	37.60(2.46)	2.79
Gaussian	82.27(5.74)	94.64(2.99)	75.24(1.68)	36.27(3.79)	3.55
Ratio	**FA**	**RA**	**TA**	**MIA**	**AG↓**
Retrain	76.53	91.65	76.92	40.06	0
5%	80.90(4.37)	94.09(2.44)	75.04(1.88)	37.60(2.46)	2.79
10%	81.47(4.94)	94.55(2.90)	75.52(1.40)	36.45(3.61)	3.21
20%	81.56(5.03)	94.42(2.77)	75.16(1.76)	37.24(2.82)	3.10
30%	81.72(5.19)	94.24(2.59)	74.99(1.93)	37.02(3.04)	3.19

Effect of Masking Strategy. As shown in Table 4, we discuss the impact of different masking strategies on performance. The proposed "the most highly-attended masking" is compared with AttMask Hint (Kakogeorgiou et al., 2022) and random masking. Specifically, the highly-attention strategy achieves the most effective forgetting, with the lowest AG of only 2.79%, closest to the retraining benchmark. AttMask Hint brings higher AG values, because it cannot perform effective forgetting: AttMask Hint's FA value is too high, indicating that this masking strategy cannot forget the forget data. If random masking is adopted, the model's ability to balance between "forgetting" and "generalization" obviously declines. This result highlights the superiority of masking high-attention regions in achieving optimal "forgetting" effects.

Visualization of Attention Maps. Figure 3 presents the visualization of attention maps. The first row displays the input images. The second row shows the visualization results of the attention maps from the original model, highlighting its focus on class-discriminative regions. The third row shows the results after forgetting. For the pickup_truck class (first three columns), the original model captures key structural features such as wheels, lights, and the truck bed. After forgetting, the attention is weakened or shifted toward irrelevant backgrounds. For the oak_tree class (last three columns), the original model attends to distinctive regions like trunks and foliage, while the model after forgetting loses this focus, indicating effective forgetting. These consistent changes across multiple samples validate our method's ability to selectively forget class-specific patterns while preserving the model's capacity to recognize other categories.

Table 4: Lethe under different masking strategies (DeiT-T on Tiny-ImageNet).

Strategy	FA	RA	TA	MIA	AG↓
Retrain	76.53	91.65	76.92	40.06	0
Highly-attended	80.90(4.37)	94.09(2.44)	75.04(1.88)	37.60(2.46)	2.79
AttMask Hint	87.25(10.72)	95.44(3.79)	76.52(0.40)	30.40(9.66)	6.14
Random Mask	92.90(16.37)	94.37(2.72)	77.32(0.40)	24.22(15.84)	8.83

Efficiency Analysis. Figure 4 shows the efficiency of different methods. Specifically, the Retrain method requires 59.67 minutes, which is significantly longer than the approximate unlearning methods. Among the approximate unlearning methods, GA and IU achieve the highest efficiency. However, their average gaps are 9.05% and 8.26%, respectively, which are much higher than that of the Retrain method. This indicates that their forgetting effects are not satisfactory. In contrast, methods such as RL, FT, and SalUn improve the forgetting effect but at the cost of significantly increased time overhead. Compared to these methods, Lethe not only achieves the best forgetting effect but also maintains a relatively low time cost.

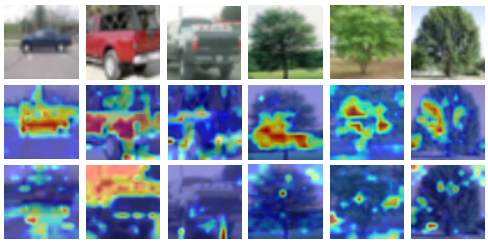


Figure 3: Visualization of Attention Maps. Classes: “pickup_truck” and ”oak_tree”.

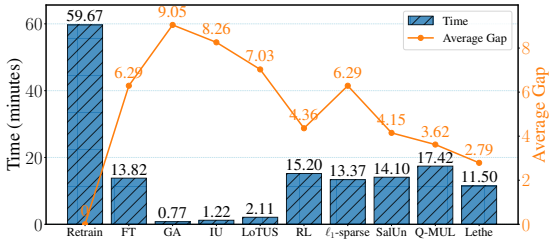


Figure 4: Efficiency and performance comparison. The results are from DeiT-T on Tiny-Imagenet.

6 Related Work

In this section, we review *Vision Transformer*, *Machine Unlearning*, and *Contrastive Learning*. More discussion of related work can be found in the Appendix A.10.

Machine Unlearning. Machine unlearning (Zhao et al., 2024; Ginart et al., 2019; Sekhari et al., 2021; Spartalis et al., 2025) can remove the impact of specific samples on a model to protect privacy. The most effective method is to retrain the model from scratch using the retain set after removing the data points (Fan et al., 2024b), but this is computationally expensive, especially for large models. Thus, researchers are developing approximate unlearning methods (Tong et al., 2025a; Liu et al., 2024) to reduce costs while maintaining model performance after unlearning. However, existing methods (Kim et al.; Chowdhury et al., 2025; Patel & Qiu, 2025) have not fully considered the characteristics of ViTs, especially their attention mechanisms. Meanwhile, recent works specifically targeting ViTs, such as NOVO (Roy et al., 2025) and Low-rank (Poppi et al., 2024) which focus on class-wise forgetting, and FRAMU (Shaik et al., 2024) which targets federated unlearning scenarios, are not directly applicable to random data forgetting. Unlike these methods, we propose an attention-based forgetting method specifically designed for the random data forgetting scenario.

Contrastive Learning. Contrastive learning brings similar samples closer and pushes dissimilar ones farther to capture data structure and features. For example, SimCLR (Chen et al., 2020) constructs positive pairs from augmented images and optimizes consistency via a contrastive loss. SupCon (Khosla et al., 2020) extends this to supervised settings using labels for intra-class compactness and inter-class separability. MoCo (He et al., 2020) uses a momentum-based queue to scale negatives and maintain consistency with a key encoder. (kyu Lee et al., 2024) propose a contrastive unlearning method for CNNs, without taking into account the characteristics of ViTs. These methods excel at representation learning but are not directly suitable for machine unlearning in ViTs. To address this, we propose a novel method using contrastive learning for unlearning specific samples while preserving model performance.

7 Conclusions

In this paper, we propose Lethe, a machine unlearning method specifically designed for Vision Transformers. Unlike existing methods that do not exploit the architectural properties of ViTs, Lethe is built upon a key finding: masking a small proportion of the highest-attention patches can decouple the model’s recognition capability from its memorization of specific training samples. Leveraging this property, we design an attention-guided contrastive unlearning framework that constructs positive logits from masked images and negative logits from original images, enabling the model to selectively erase sample-specific details while preserving category-level knowledge. Extensive experiments demonstrate that Lethe consistently achieves the closest approximation to the retain-only Retrain reference under the shared pretrained initialization among all compared methods. More broadly, our work reveals a fundamental connection between the attention mechanism and data memorization in ViTs, which we believe can inspire further exploration of how internal model signals can be leveraged to achieve more precise and efficient machine unlearning.

AI USE STATEMENT

Generative AI tools were used solely to improve the language, clarity, and readability of the manuscript. They were not used to generate research ideas, design the method or experiments, implement code, analyze data, or interpret results. All AI-assisted edits were reviewed and revised by the authors, who take full responsibility for the final content of this work.

ETHICS STATEMENT

This research aims to advance privacy-preserving machine learning by enabling models to forget designated training samples. Our experiments are conducted exclusively on publicly available benchmark datasets and do not involve human participants or newly collected sensitive information. We acknowledge that the effectiveness of machine unlearning may vary across datasets and deployment environments; therefore, real-world applications should be accompanied by careful evaluation and responsible data-management practices.

REPRODUCIBILITY STATEMENT

To support reproducibility, the proposed method and training procedure are described in Section 4, the experimental setup is presented in Section 5, and further implementation details and hyperparameter settings are provided in Appendix A.4. All datasets used in this work are publicly available.

REFERENCES

Lucas Bourtoule, Varun Chandrasekaran, Christopher A Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In *2021 IEEE Symposium on Security and Privacy (SP)*, pp. 141–159. IEEE, 2021.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PmLR, 2020.

Eli Chien, Chao Pan, and Olgica Milenkovic. Certified graph unlearning. *arXiv preprint arXiv:2206.09140*, 2022.

Somnath Basu Roy Chowdhury, Krzysztof Marcin Choromanski, Arijit Sehanobish, Kumar Avinava Dubey, and Snigdha Chaturvedi. Towards scalable exact machine unlearning using parameter-efficient fine-tuning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=oe51Q5Uo37.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Chongyu Fan, Jiancheng Liu, Licong Lin, Jinghan Jia, Ruiqi Zhang, Song Mei, and Sijia Liu. Simplicity prevails: Rethinking negative preference optimization for llm unlearning. *arXiv preprint arXiv:2410.07163*, 2024a.

Chongyu Fan, Jiancheng Liu, Yihua Zhang, Dennis Wei, Eric Wong, and Sijia Liu. Salun: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation. *In International Conference on Learning Representations*, 2024b.

Antonio Ginart, Melody Guan, Gregory Valiant, and James Y Zou. Making ai forget you: Data deletion in machine learning. *Advances in neural information processing systems*, 32, 2019.

Aditya Golatkar, Alessandro Achille, and Stefano Soatto. Eternal sunshine of the spotless net: Selective forgetting in deep networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9304–9312, 2020.

Laura Graves, Vineel Nagisetty, and Vijay Ganesh. Amnesiac machine learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 11516–11524, 2021.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

Chris Jay Hoofnagle, Bart Van Der Sloot, and Frederik Zuiderveen Borgesius. The european union general data protection regulation: what it is and what it means. *Information & Communications Technology Law*, 28(1):65–98, 2019.

Zachary Izzo, Mary Anne Smart, Kamalika Chaudhuri, and James Zou. Approximate data deletion from machine learning models. In *International Conference on Artificial Intelligence and Statistics*, pp. 2008–2016. PMLR, 2021.

Ioannis Kakogeorgiou, Spyros Gidaris, Bill Psomas, Yannis Avrithis, Andrei Bursuc, Konstantinos Karantzalos, and Nikos Komodakis. What to hide from your students: Attention-guided masked image modeling. In *European Conference on Computer Vision*, pp. 300–318. Springer, 2022.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.

Hyo Seo Kim, Dongyoon Han, and Junsuk Choe. Negmerge: Sign-consensual weight merging for machine unlearning. In *Forty-second International Conference on Machine Learning*.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International conference on machine learning*, pp. 1885–1894. PMLR, 2017.

Giorgos Kordopatis-Zilos, Vladan Stojnić, Anna Manko, Pavel Suma, Nikolaos-Antonios Ypsilantis, Nikos Efthymiadis, Zakaria Laskar, Jiri Matas, Ondrej Chum, and Giorgos Tolias. Ilias: Instance-level image retrieval at scale. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 14777–14787, 2025.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Meghdad Kurmanji, Peter Triantafillou, Jamie Hayes, and Eleni Triantafillou. Towards unbounded machine unlearning. *Advances in neural information processing systems*, 36:1957–1987, 2023.

Hong kyu Lee, Qiuchen Zhang, Carl Yang, Jian Lou, and Li Xiong. Contrastive unlearning: A contrastive approach to machine unlearning, 2024. URL `https://arxiv.org/abs/2401.10458`.

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

Jiancheng Liu, Parikshit Ram, Yuguang Yao, Gaowen Liu, Yang Liu, PRANAY SHARMA, Sijia Liu, et al. Model sparsity can simplify machine unlearning. *Advances in Neural Information Processing Systems*, 36, 2024.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021.

Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1096–1104, 2016.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Baolin Wu, Andrew Y Ng, et al. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, volume 2011, pp. 4. Granada, 2011.

Thanh Tam Nguyen, Thanh Trung Huynh, Phi Le Nguyen, Alan Wee-Chung Liew, Hongzhi Yin, and Quoc Viet Hung Nguyen. A survey of machine unlearning. *arXiv preprint arXiv:2209.02299*, 2022.

Trong Nguyen. An empirical evaluation of the implementation of the california consumer privacy act (ccpa). *arXiv preprint arXiv:2205.09897*, 2022.

Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4004–4012, 2016.

Gaurav Patel and Qiang Qiu. Learning to unlearn while retaining: Combating gradient conflicts in machine unlearning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4211–4221, October 2025.

Samuele Poppi, Sara Sarto, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Unlearning vision transformers without retaining data via low-rank decompositions. In *International Conference on Pattern Recognition*, pp. 147–163. Springer, 2024.

Bill Psomas, George Retsinas, Nikos Efthymiadis, Panagiotis Filntisis, Yannis Avrithis, Petros Maragos, Ondrej Chum, and Giorgos Tolias. Instance-level composed image retrieval. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Soumya Roy, Soumya Banerjee, Vinay Verma, Soumik Dasgupta, Deepak Gupta, and Piyush Rai. Novo: Unlearning-compliant vision transformers. *arXiv preprint arXiv:2507.03281*, 2025.

Ayush Sekhari, Jayadev Acharya, Gautam Kamath, and Ananda Theertha Suresh. Remember what you want to forget: Algorithms for machine unlearning. *Advances in Neural Information Processing Systems*, 34:18075–18086, 2021.

Thanveer Shaik, Xiaohui Tao, Lin Li, Haoran Xie, Taotao Cai, Xiaofeng Zhu, and Qing Li. Framu: Attention-based machine unlearning using federated reinforcement learning. *IEEE Transactions on Knowledge and Data Engineering*, 36(10):5153–5167, 2024.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pp. 3–18. IEEE, 2017.

Christoforos N Spartalis, Theodoros Semertzidis, Efstratios Gavves, and Petros Daras. Lotus: Large-scale machine unlearning with a taste of uncertainty. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 10046–10055, 2025.

Anvith Thudi, Gabriel Deza, Varun Chandrasekaran, and Nicolas Papernot. Unrolling sgd: Understanding factors influencing machine unlearning. In *2022 IEEE 7th European Symposium on Security and Privacy (EuroS&P)*, pp. 303–319. IEEE, 2022.

Yujia Tong, Yuze Wang, Jingling Yuan, and Chuang Hu. Robust machine unlearning for quantized neural networks via adaptive gradient reweighting with similar labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20603–20612, October 2025a.

Yujia Tong, Jingling Yuan, and Chuang Hu. Enhancing quantization-aware training on edge devices via relative entropy coreset selection and cascaded layer correction, 2025b. URL `https://arxiv.org/abs/2507.17768`.

Yujia Tong, Jingling Yuan, Tian Zhang, Jianquan Liu, and Chuang Hu. Dfq-vit: Data-free quantization for vision transformers without fine-tuning. *arXiv preprint arXiv:2507.14481*, 2025c.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pp. 10347–10357. PMLR, 2021.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Shuang Wang and Shuqiang Jiang. Instre: a new benchmark for instance-level object retrieval and recognition. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 11(3):1–21, 2015.

Alexander Warnecke, Lukas Pirch, Christian Wressnegger, and Konrad Rieck. Machine unlearning of features and labels. *arXiv preprint arXiv:2108.11577*, 2021.

Tobias Weyand, Andre Araujo, Bingyi Cao, and Jack Sim. Google landmarks dataset v2-a large-scale benchmark for instance-level recognition and retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2575–2584, 2020.

Nikolaos-Antonios Ypsilantis, Noa Garcia, Guangxing Han, Sarah Ibrahimi, Nanne Van Noord, and Giorgos Tolias. The met dataset: Instance-level recognition for artworks. In *Thirty-fifth conference on neural information processing systems datasets and benchmarks track (Round 2)*, 2021.

Kairan Zhao, Meghdad Kurmanji, George-Octavian Barbulescu, Eleni Triantafillou, and Peter Triantafillou. What makes unlearning hard and what to do about it. *Advances in Neural Information Processing Systems*, 37:12293–12333, 2024.

A APPENDIX

The organization of the appendix is as follows:

- Appendix A.1: Justification of the Evaluation Metrics ;
- Appendix A.2: Implementation Details ;
- Appendix A.3: The Introduction of Baselines ;
- Appendix A.4: Proof of Theorem 4.2;
- Appendix A.5: ViT-T/S on CIFAR-10/CIFAR-100;
- Appendix A.6: DeiT-T/S on SVHN;
- Appendix A.7: t-SNE Analysis;
- Appendix A.8: Additional Exploring Experiments;
- Appendix A.9: Mutual Information between Retrained and Unlearned Model;
- Appendix A.10: Discussion of more related works;
- Appendix A.11: DeiT-T on INSTRE instance dataset;
- Appendix A.12: Statistical Significance Analysis;
- Appendix A.13: Effects of Contrastive Learning on Logits vs. Features;
- Appendix A.14: Effects of Training Epoch;
- Appendix A.15: Limitations and Future Work.

A.1 JUSTIFICATION OF THE EVALUATION METRICS

We chose to use AG as the primary metric to maintain consistency with the existing machine unlearning literature, including baseline methods such as ℓ_1-sparse (Liu et al., 2024), SalUn (Fan et al., 2024b), LoTUS (Spartalis et al., 2025), and Q-MUL (Tong et al., 2025a), which adopt the same evaluation framework. This helps ensure direct comparability of our results with existing methods. The core goal of AG is to measure the overall deviation of an unlearned model from the Retrain reference under the shared pretrained initialization. In machine unlearning tasks, unlearning effectiveness (reflected through FA and MIA) and model utility preservation (reflected through RA and TA) are equally important, and equal-weight averaging helps capture a comprehensive balance rather than biasing the evaluation toward a single aspect. Nevertheless, for a more comprehensive comparison, we additionally compute the distributional discrepancy (KL/JS divergence) between the predicted softmax distributions of the retrained and unlearned models in Appendix A.9.

A.2 IMPLEMENTATION DETAILS

We follow the experimental settings of SalUn and ℓ_1-sparse for the baseline methods. All experiments are conducted using the SGD optimizer. For Retrain, we train each model for 45 epochs with learning rates sampled from the range [1e-5, 1e-3]. For FT and RL, we train each model for 10 epochs with learning rates sampled from the range [1e-4, 1e-2]. For GA, we train for 5 epochs with learning rates between [1e-7, 1e-5]. For IU,we vary the parameter α, which relates to the Woodfisher Hessian Inverse approximation, within the range [1,20]. For ℓ_1-sparse, we search for the optimal γ value in the range [1e-6, 1e-4],and explore learning rates between [1e-5, 1e-3]. For SalUn , we train for 10 epochs with learning rates sampled from [1e-4, 1e-2] and sparsity ratios in the range [0.1, 0.9]. For Lethe, we apply the SGD optimizer with a batch size of 128. For ViT-T/S/B, we train for 10 epochs with learning rates in the range [1e-5,1e-3]. For Swin-T/S, we train for 10 epochs with learning rates in the range [1e-5,1e-3]. For DeiT-T/S/B, we train for 10 epochs with learning rates in the range [1e-4,1e-2]. We set the masking ratio to 5%, the number of training epochs for the forget set to 2, and the number of training epochs for the retain set to 8. Temperature parameter τ is 0.07. All experiments are conducted on a single NVIDIA RTX 4090 GPU.

A.3 THE INTRODUCTION OF BASELINES

(1) **Retrain**: Independently training a model solely on the retained data from the same pretrained initialization used for the original model. It provides a downstream retain-only reference rather than a random-initialization retraining baseline. (2) **Fine-Tuning (FT)** (Warnecke et al., 2021; Golatkar et al., 2020): A less intensive alternative requiring only minor adjustments to the original model via a few epochs on the retained data. (3) **Gradient Ascent (GA)** (Graves et al., 2021; Thudi et al., 2022): Updates the model parameters in the direction opposite to gradient descent, specifically using the forget dataset. (4) **Influence Unlearning (IU)** (Koh & Liang, 2017; Izzo et al., 2021): Estimates the impact of the forget set D_f on model $\mathcal{M}_0$ using influence functions, then performs a Newton-step parameter update to negate it. (5)**Random Labels (RL)** (Golatkar et al., 2020): Trains on the full dataset after randomizing the labels of the forget set instances. (6) ℓ_1**-sparse** (Liu et al., 2024): Induces weight sparsity through model pruning to achieve approximate unlearning. (7) **SalUn** (Fan et al., 2024b): Combines the RL approach with a gradient-based weight saliency map. (8) **LoTUS** (Spartalis et al., 2025): Smooths the model's output probabilities up to an information-theoretic bound to mitigate overconfidence stemming from data memorization, enabling efficient unlearning directly in the output space. (9) **Q-MUL** (Tong et al., 2025a): Addresses unlearning in quantized neural networks by adaptively reweighting gradients based on samples with similar labels to improve robustness.

A.4 PROOF OF THEOREM 4.2

Proof. From the L-smoothness of $\mathcal{L}_{CL}$, we have the quadratic upper bound:

$$\mathcal{L}_{CL}(\theta_{t+1}) \leq \mathcal{L}_{CL}(\theta_t) + \nabla \mathcal{L}_{CL}(\theta_t)^\top (\theta_{t+1} - \theta_t) + \frac{L}{2}\|\theta_{t+1} - \theta_t\|^2,$$

Substituting the update rule $\theta_{t+1} - \theta_t = -\eta \nabla \mathcal{L}_{CL}(\theta_t)$ yields:

$$\mathcal{L}_{CL}(\theta_{t+1}) \leq \mathcal{L}_{CL}(\theta_t) - \eta \|\nabla \mathcal{L}_{CL}(\theta_t)\|^2 + \frac{L\eta^2}{2}\|\nabla \mathcal{L}_{CL}(\theta_t)\|^2,$$

Rearranging terms gives:

$$\mathcal{L}_{CL}(\theta_{t+1}) \leq \mathcal{L}_{CL}(\theta_t) - \eta \left(1 - \frac{L\eta}{2}\right) \|\nabla \mathcal{L}_{CL}(\theta_t)\|^2,$$

Since $\eta < \frac{1}{L}$, we have $1 - \frac{L\eta}{2} > \frac{1}{2}$, and thus:

$$\mathcal{L}_{CL}(\theta_{t+1}) \leq \mathcal{L}_{CL}(\theta_t) - \frac{\eta}{2}\|\nabla \mathcal{L}_{CL}(\theta_t)\|^2,$$

Summing from $t = 0$ to $T - 1$:

$$\sum_{t=0}^{T-1} \|\nabla \mathcal{L}_{CL}(\theta_t)\|^2 \leq \frac{2}{\eta}\left[\mathcal{L}_{CL}(\theta_0) - \mathcal{L}_{CL}(\theta_T)\right] \leq \frac{2}{\eta}\left[\mathcal{L}_{CL}(\theta_0) - \mathcal{L}^*_{CL}\right],$$

Therefore, the minimum gradient norm up to iteration T satisfies:

$$\min_{0 \leq t \leq T} \|\nabla \mathcal{L}_{CL}(\theta_t)\|^2 \leq \frac{2[\mathcal{L}_{CL}(\theta_0) - \mathcal{L}^*_{CL}]}{\eta T}.$$

which completes the proof. □

A.5 VIT-T/S ON CIFAR-10/CIFAR-100

We report the experimental results of ViT-T and ViT-S on CIFAR-10 and CIFAR-100 in Tables 12, 13, 14, and 15. For ViT-T on CIFAR-10, Lethe shows average performance gaps from the Retrain model of 1.44%, 1.58%, and 1.78% at different forgetting ratios. Compared with the best-performing baseline, these gaps are reduced by 0.15%, 0.05%, and 0.20%, respectively. On CIFAR-100 with ViT-T, Lethe achieves either the best or second-best results. The average performance gaps from Retrain are 2.64%, 2.13%, and 2.09%, which are 0.37%, 1.07%, and 2.60% smaller than those of the previous state-of-the-art methods. For ViT-S on CIFAR-100, Lethe yields average gaps from Retrain of 0.66%, 0.95%, and 1.74%, indicating strong performance across all forgetting ratios.

Table 7: KLD & JSD comparison among different methods.(DeiT-T on Tiny-ImageNet)

Metric / Method	FT	GA	IU	RL	ℓ_1-sparse	SalUn	Lethe
KLD on forget_set	0.4586	0.4923	0.4972	0.5972	0.4438	0.4481	**0.4388**
KLD on retain_set	0.3176	0.2737	0.2880	0.3527	0.3173	0.2798	**0.2139**
KLD on test_set	0.3705	0.2965	0.3194	0.4647	0.3696	0.3715	**0.2693**
JSD on forget_set	0.0834	0.0825	0.0898	0.1344	0.0818	0.0992	**0.0815**
JSD on retain_set	0.0578	0.0544	0.0595	0.0736	0.0575	0.0627	**0.0439**
JSD on test_set	0.0694	0.0633	0.0632	0.0987	0.0677	0.0831	**0.0530**

A.6 DeiT-T/S on SVHN

We summarize the results of DeiT-T and DeiT-S on the SVHN dataset in Tables 16 and 17. Lethe delivers either the best or second-best performance for DeiT-T. For DeiT-S, Lethe achieves average gaps from Retrain of 0.47%, 0.68%, and 1.12%, demonstrating excellent forgetting effects under varying forgetting settings.

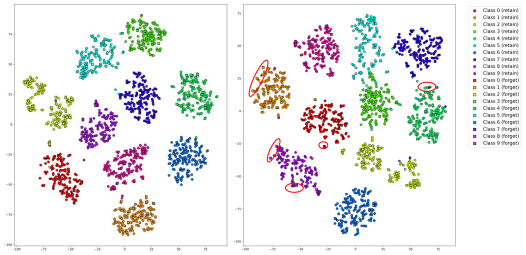


Figure 5: t-SNE of original model (left) and unlearned model (right). The experimental setup is ViT-T on CIFAR-10.

A.7 t-SNE Analysis

Figure 5 presents the t-SNE visualization. Specifically, the features of the forget data in the unlearned model tend to move away from those of the retain data, as indicated by the red circles. Within the same class, the specific forget data is notably separated from the retain data, demonstrating the model's ability to effectively forget specific samples.

A.8 Additional Exploring Experiments

We conduct experiments on a larger dataset (Tiny-ImageNet) and with different forgetting ratios (30 50%), and conduct more comprehensive comparisons in Tables 5 and 6. Masking high-attention patches naturally hides fine-grained, instance-level details. On the higher-resolution Tiny-ImageNet the effect is even stronger: the patches capture subtler details, further highlighting the model's robustness. At a 5% masking ratio, TA jumps by 4.85% (75.62% to 80.47%) while MIA falls by 19.95%. This shows that masking weakens the model's memorization of individual training samples without harming and even improving recognition performance, since sufficient class-discriminative features are preserved.

A.9 Mutual Information between Retrained and Unlearned Model

We compute the mutual information (KL divergence/JS divergence) between the predicted softmax probability distributions of the retrained model and the unlearned model on the test set, and report its values in comparison with the baseline methods. As shown in Table 7, our method achieves the smallest values on both KL divergence and JS divergence, indicating that the predictive distribution of the unlearned model is closest to that of the retrained model.

Table 5: TA and MIA under different masking ratios (DeiT-T on Tiny-ImageNet, Forgetting Ratio is 30%).

Ratio	Zero Padding		Gaussian Padding	
	TA	MIA	TA	MIA
0%	75.62	26.98	75.62	26.98
5%	80.47$^{\uparrow 4.85}$	7.03$^{\downarrow 19.95}$	82.03$^{\uparrow 6.41}$	7.03$^{\downarrow 19.95}$
10%	69.54$^{\downarrow 6.08}$	14.06$^{\downarrow 12.92}$	74.22$^{\downarrow 1.40}$	17.19$^{\downarrow 9.79}$
20%	57.03$^{\downarrow 18.59}$	21.09$^{\downarrow 5.89}$	60.16$^{\downarrow 15.46}$	26.56$^{\downarrow 0.42}$
30%	44.53$^{\downarrow 31.09}$	27.34$^{\uparrow 0.36}$	49.22$^{\downarrow 26.40}$	32.03$^{\uparrow 5.05}$

Table 6: TA and MIA under different masking ratios (DeiT-T on Tiny-ImageNet, Forgetting Ratio is 50%).

Ratio	Zero Padding		Gaussian Padding	
	TA	MIA	TA	MIA
0%	74.17	45.43	74.17	45.43
5%	78.91$^{\uparrow 4.74}$	10.94$^{\downarrow 34.49}$	81.25$^{\uparrow 7.08}$	10.94$^{\downarrow 34.49}$
10%	73.44$^{\downarrow 0.73}$	11.72$^{\downarrow 33.71}$	76.56$^{\uparrow 2.39}$	11.72$^{\downarrow 33.71}$
20%	56.25$^{\downarrow 17.92}$	17.97$^{\downarrow 27.46}$	59.38$^{\downarrow 14.79}$	17.97$^{\downarrow 27.46}$
30%	42.97$^{\downarrow 31.20}$	21.88$^{\downarrow 23.55}$	51.56$^{\downarrow 22.61}$	24.22$^{\downarrow 21.21}$

A.10 Discussion of more related works

Vision Transformer. The Transformer (Vaswani et al., 2017) architecture, initially popular in natural language processing, has become dominant in computer vision. Unlike CNNs, it captures long-range visual relationships via self-attention. Vision Transformers (ViTs) (Dosovitskiy et al., 2020; Tong et al., 2025c) divide images into 16×16 patches as tokens, with a unique class token for classification. DeiT (Touvron et al., 2021) enhances ViT's practicality through efficient training with limited data using knowledge distillation and data augmentation. Swin Transformer (Liu et al., 2021) uses a sliding window and hierarchical structure to capture local and global features efficiently while reducing computation. As ViTs become foundational in computer vision, privacy protection for ViTs is an important research direction. This paper focuses on privacy protection of ViTs.

Instance-level Recognition/Retrieval. The random data forgetting requires the model to selectively erase the influence of specific instances within the same category, which has conceptual relevance to instance-level recognition/retrieval tasks (Liu et al., 2016; Kakogeorgiou et al., 2022; Oh Song et al., 2016). The latter aims to distinguish visually highly similar but identity-distinct items rather than coarse-grained categories. Specifically, INSTRE (Wang & Jiang, 2015) is proposed as a new benchmark for instance-level visual object retrieval and recognition; ILIAS (Kordopatis-Zilos et al., 2025) serves as a large-scale instance-level image retrieval test set for evaluating models and retrieval techniques in recognizing specific objects under conditions with large-scale distractors; BASIC (Psomas et al., 2025) acts as a training-free method that utilizes pre-trained vision-language models to accomplish retrieval. These instance-level recognition/retrieval tasks (Weyand et al., 2020; Ypsilantis et al., 2021) emphasize achieving intra-class distinction through fine-grained visual cues, which is highly parallel to the core challenge of random data forgetting: forgetting specific samples while retaining other samples in the same category requires the model to erase instance-level memory traces without damaging category-level abstract representations. Our method leverages the attention mechanism to mask high-attention regions, thereby achieving selective forgetting in ViTs.

A.11 DeiT-T on INSTRE instance dataset

We conduct experiments on the INSTRE instance dataset in Table 8 (using the DeiT-T model, the instances to be forgotten are: 01a_canada_book, 05a_foxtoy, 06b_red_car, 08b_DDog, 13b_toy_man, 20a_coconut_juice, 27a_china_unicom_card, 31b_Blue_notebook, 32b_bus_card, 45b_1yuan). As can be seen from the Table 8,our method Lethe also demonstrates superior forgetting performance on the instance-level dataset. Specifically, Lethe achieves the lowest AG of 0.78%.

Table 8: Comparison with other unlearning methods (DeiT-T on INSTRE).

Method	FA	RA	TA	MIA	AG↓
Retrain	94.90	95.22	94.45	79.00	0
FT	90.10(4.80)	90.37(4.85)	89.09(5.36)	79.90(0.90)	3.98
GA	93.81(1.09)	94.50(0.72)	94.17(0.28)	85.40(6.40)	2.12
IU	92.90(2.00)	92.66(2.56)	93.13(1.32)	81.60(2.60)	2.12
RL	94.20(0.70)	93.09(2.13)	92.94(1.51)	82.10(3.10)	1.86
ℓ1-sparse	94.10(0.80)	93.23(1.99)	93.51(0.94)	85.00(6.00)	2.43
SalUn	92.00(2.90)	91.37(3.85)	91.06(3.39)	79.10(0.10)	2.56
Lethe	94.60(0.30)	95.41(0.19)	95.48(1.03)	77.40(1.60)	0.78

A.12 Statistical Significance Analysis

We report the mean and standard deviation of 5 independent runs with different random seeds for DeiT-T on Tiny-ImageNet with 10% forgetting ratio in Table 9.All methods show consistent performance with small standard deviations, and Lethe maintains the lowest AG in every run, confirming that the observed improvements are statistically reliable.

Table 9: Statistical significance analysis (DeiT-T on Tiny-ImageNet, 10% forgetting). Results are reported as mean $\pm$ std over 5 independent runs.

Method	FA	RA	TA	MIA	AG↓
Retrain	76.53	91.65	76.92	40.06	0
SalUn	84.05$\pm$0.58	95.18$\pm$0.21	75.66$\pm$0.09	35.79$\pm$0.89	4.15
Q-MUL	83.17$\pm$0.37	94.78$\pm$0.11	76.36$\pm$0.02	35.93$\pm$0.06	3.62
Lethe	80.90$\pm$0.23	94.09$\pm$0.14	75.04$\pm$0.04	37.60$\pm$0.18	2.79

A.13 Effects of Contrastive Learning on Logits vs. Features

We compare the effects of contrastive learning on logits vs. features. As shown in Table 10, performing contrastive learning solely in the feature space hardly affects the prediction decisions, and the forgetting effect is significantly degraded.

Table 10: Effects of contrastive learning on logits vs. features (DeiT-T on Tiny-ImageNet).

Target	FA	RA	TA	MIA	AG↓
Retrain	76.53	91.65	76.92	40.06	0
logits	80.90(4.37)	94.09(2.44)	75.04(1.88)	37.60(2.46)	2.79
features	85.02(8.49)	95.30(3.65)	76.60(0.32)	32.73(7.33)	4.95

A.14 Effects of Training Epoch

As shown in Table 11, we investigate the impact of the number of training epochs allocated to the forget set and the retain set on the performance. The total number of training epochs is fixed at 10, and only the number of epochs for contrastive learning (CL) on the forget set and the number of epochs for fine-tuning (FT) on the retain set are adjusted, thereby evaluating the trade-off between "forgetting effect" and "model utility maintenance." Specifically, a small number of CL epochs suffices to achieve effective "forgetting," and the optimal balance point is "2CL + 8FT," at which AG is the lowest, at 2.79%, closest to the retraining benchmark. If CL epochs are excessive and FT epochs are insufficient, the model's ability to maintain generalization capability declines severely. If only 2 rounds of CL training are used without fine-tuning, the model accuracy declines severely, which highlights the necessity of retain set training for maintaining generalization capability.

Table 11: Lethe under different training epochs (DeiT-T on Tiny-ImageNet).

Type	FA	RA	TA	MIA	AG↓
Retrain	76.53	91.65	76.92	40.06	0
1CL + 9FT	80.56(4.03)	95.30(3.65)	75.89(1.03)	37.40(2.66)	2.84
2CL + 8FT	80.90(4.37)	94.09(2.44)	75.04(1.88)	37.60(2.46)	2.79
3CL + 7FT	82.24(5.71)	93.65(2.00)	76.74(0.18)	36.19(3.87)	2.94
4CL + 6FT	83.43(6.90)	93.20(1.55)	76.42(0.50)	35.84(4.22)	3.29
5CL + 5FT	84.97(8.44)	92.54(0.89)	76.14(0.78)	32.67(7.39)	4.38
6CL + 4FT	86.63(10.10)	92.33(0.68)	76.62(0.30)	31.51(8.55)	4.91
7CL + 3FT	88.24(11.71)	92.18(0.53)	76.48(0.44)	28.35(11.71)	6.10
8CL + 2FT	88.23(11.70)	90.26(1.39)	76.26(0.66)	30.17(9.89)	5.91
9CL + 1FT	67.67(8.86)	68.82(22.83)	63.03(13.89)	45.67(5.61)	12.80
2CL only	0.65(75.88)	0.54(91.11)	0.56(76.36)	2.07(37.99)	70.34

A.15 Limitations and Future Work

Our experiments are conducted on CIFAR-10, CIFAR-100, SVHN, Tiny-ImageNet, and INSTRE, which are widely adopted benchmarks in the machine unlearning literature, consistent with ℓ_1-sparse, SalUn, Q-MUL, and LoTUS. Since our evaluation conditions on a shared pretrained initialization, it focuses on removing the influence introduced by the forget set during downstream training; removing information potentially acquired during upstream pretraining is beyond the scope of this work. Scaling to full ImageNet requires substantial computational resources, particularly because the retain-only Retrain reference must be trained separately for each forgetting configuration. We note that our method has been validated across eight different architectures (ViT-T/S/B, DeiT-T/S/B, and Swin-T/S) and five datasets with varying characteristics, demonstrating consistent improvements. Extending Lethe to larger-scale datasets and broader unlearning settings remains an important direction for future work.

Table 12: Performance of various MU methods for ViT-T on CIFAR-10. The unlearning scenarios include 10%, 30%, and 50% forgetting rates. **Bold** indicates the best performance and underline indicates the runner-up. A performance gap against Retrain is provided in (•).

Method	CIFAR-10 (ViT-T)				
	FA	RA	TA	MIA	AG↓
The proportion of forgotten data samples to all samples is 10%					
Retrain	96.87	99.91	97.59	6.11	0
FT	99.44(2.57)	99.94(0.03)	97.70(0.11)	1.93(4.18)	1.72
GA	99.46(2.59)	99.58(0.33)	97.63(0.04)	1.91(4.20)	1.79
IU	98.76(1.89)	99.21(0.70)	97.12(0.47)	2.80(3.31)	1.59
RL	96.76(0.11)	99.39(0.52)	96.66(0.93)	14.24(8.13)	2.42
ℓ_1-sparse	99.42(2.55)	99.94(0.03)	97.71(0.12)	1.93(4.18)	1.72
SalUn	96.53(0.34)	99.18(0.73)	96.49(1.10)	15.13(9.02)	2.80
Lethe	96.62(0.25)	97.20(2.71)	95.43(2.16)	5.49(0.62)	**1.44**
Method	CIFAR-10 (ViT-T)				
	FA	RA	TA	MIA	AG↓
The proportion of forgotten data samples to all samples is 30%					
Retrain	96.09	99.92	97.19	6.09	0
FT	99.54(3.45)	99.94(0.02)	97.74(0.55)	1.70(4.39)	2.10
GA	99.54(3.45)	99.56(0.36)	97.61(0.42)	1.52(4.57)	2.20
IU	97.93(1.84)	98.33(1.59)	96.39(0.80)	3.81(2.28)	1.63
RL	95.33(0.76)	98.39(1.53)	95.42(1.77)	16.99(10.90)	3.74
ℓ_1-sparse	99.53(3.44)	99.94(0.02)	97.64(0.55)	1.73(4.36)	2.09
SalUn	98.09(2.00)	98.32(1.60)	95.52(1.67)	15.57(9.48)	3.69
Lethe	96.99(0.90)	97.33(2.59)	95.57(1.62)	4.87(1.22)	**1.58**
Method	CIFAR-10 (ViT-T)				
	FA	RA	TA	MIA	AG↓
The proportion of forgotten data samples to all samples is 50%					
Retrain	95.67	99.87	96.86	7.04	0
FT	99.56(3.89)	99.96(0.09)	97.67(0.81)	1.72(5.32)	2.53
GA	99.45(3.78)	99.52(0.35)	97.58(0.72)	1.74(5.30)	2.54
IU	97.96(2.29)	98.34(1.53)	95.83(1.03)	3.99(3.05)	1.98
RL	94.25(1.42)	97.02(2.85)	93.64(3.22)	16.82(9.78)	4.32
ℓ_1-sparse	99.38(3.71)	99.97(0.10)	97.62(0.76)	2.50(4.54)	2.28
SalUn	91.41(4.26)	93.74(6.13)	91.09(5.77)	18.48(11.44)	6.40
Lethe	95.53(0.14)	95.83(4.04)	94.17(2.69)	6.78(0.26)	**1.78**

Table 13: Performance of various MU methods for ViT-S on CIFAR-10. The unlearning scenarios include 10%, 30%, and 50% forgetting rates. **Bold** indicates the best performance and underline indicates the runner-up. A performance gap against Retrain is provided in (•).

Method	CIFAR-10 (ViT-S)				
	FA	RA	TA	MIA	AG↓
The proportion of forgotten data samples to all samples is 10%					
Retrain	98.49	99.99	98.65	2.96	0
FT	98.40(0.09)	99.99(0.00)	98.53(0.12)	3.20(0.24)	**0.11**
GA	99.77(1.28)	99.84(0.15)	98.47(0.18)	0.62(2.34)	0.99
IU	99.71(1.22)	99.84(0.15)	98.49(0.16)	0.75(2.21)	0.94
RL	97.69(0.80)	99.85(0.14)	97.86(0.79)	11.17(8.21)	2.49
ℓ_1-sparse	98.56(0.07)	99.99(0.00)	98.38(0.27)	2.84(0.12)	0.12
SalUn	97.64(0.85)	99.79(0.20)	97.59(1.06)	8.04(5.08)	1.80
Lethe	98.64(0.15)	99.99(0.00)	98.54(0.11)	3.49(0.53)	0.20
Method	CIFAR-10 (ViT-S)				
	FA	RA	TA	MIA	AG↓
The proportion of forgotten data samples to all samples is 30%					
Retrain	98.40	99.98	98.54	3.01	0
FT	98.82(0.42)	99.99(0.01)	98.34(0.20)	2.89(0.12)	**0.19**
GA	99.82(1.42)	99.83(0.15)	98.49(0.05)	0.52(2.49)	1.03
IU	99.41(1.01)	99.49(0.49)	98.05(0.49)	1.35(1.66)	0.91
RL	97.44(0.96)	99.71(0.27)	97.38(1.16)	9.24(6.23)	2.16
ℓ_1-sparse	98.76(0.36)	99.98(0.00)	98.38(0.16)	2.91(0.10)	0.16
SalUn	97.24(1.16)	99.31(0.67)	96.87(1.67)	10.07(7.06)	2.64
Lethe	98.58(0.18)	99.98(0.00)	98.35(0.19)	3.41(0.45)	0.21
Method	CIFAR-10 (ViT-S)				
	FA	RA	TA	MIA	AG↓
The proportion of forgotten data samples to all samples is 50%					
Retrain	98.24	99.97	98.17	3.69	0
FT	98.76(0.52)	99.98(0.01)	98.06(0.11)	3.17(0.52)	0.29
GA	99.80(1.56)	99.83(0.14)	98.49(0.32)	0.57(3.12)	1.29
IU	99.15(0.91)	99.28(0.69)	97.79(0.38)	1.79(1.90)	0.97
RL	96.92(1.32)	99.35(0.62)	96.45(1.72)	10.16(6.47)	2.53
ℓ_1-sparse	98.79(0.55)	99.98(0.01)	98.17(0.00)	3.04(0.65)	0.30
SalUn	97.29(0.95)	99.38(0.59)	96.91(1.26)	12.53(8.84)	2.91
Lethe	98.44(0.20)	99.99(0.02)	98.04(0.13)	4.12(0.43)	**0.20**

Table 14: Performance of various MU methods for ViT-T on CIFAR-100. The unlearning scenarios include 10%, 30%, and 50% forgetting rates. **Bold** indicates the best performance and underline indicates the runner-up. A performance gap against Retrain is provided in (•).

Method	CIFAR-100 (ViT-T)				
	FA	RA	TA	MIA	AG↓
The proportion of forgotten data samples to all samples is 10%					
Retrain	85.82	97.01	84.71	21.42	0
FT	83.24(2.58)	99.60(2.59)	80.83(3.88)	31.88(10.46)	4.88
GA	95.29(9.47)	96.05(0.96)	85.58(0.87)	9.08(12.34)	5.91
IU	96.09(10.27)	96.90(0.11)	86.33(1.62)	8.80(12.62)	6.16
RL	94.04(8.22)	98.42(1.41)	86.05(1.34)	30.87(9.45)	5.11
ℓ_1-sparse	87.76(1.94)	99.42(2.41)	86.51(1.80)	24.00(2.58)	**2.18**
SalUn	94.67(8.85)	97.90(0.89)	85.70(0.99)	22.73(1.31)	3.01
Lethe	89.75(3.93)	99.75(2.74)	86.36(1.65)	23.67(2.25)	2.64
Method	CIFAR-100 (ViT-T)				
	FA	RA	TA	MIA	AG↓
The proportion of forgotten data samples to all samples is 30%					
Retrain	82.93	96.17	83.12	25.99	0
FT	79.88(3.05)	99.66(3.49)	79.51(3.61)	35.25(9.26)	4.85
GA	97.44(14.51)	97.66(1.49)	87.10(3.98)	7.93(18.06)	9.51
IU	93.45(10.52)	94.90(1.27)	84.56(1.44)	10.90(15.09)	7.08
RL	93.97(11.04)	97.19(1.02)	85.10(1.98)	30.50(4.51)	4.64
ℓ_1-sparse	93.67(10.74)	99.93(3.76)	87.46(4.34)	19.00(6.99)	6.46
SalUn	93.90(10.97)	96.34(0.17)	84.76(1.64)	25.96(0.03)	3.20
Lethe	86.65(0.83)	99.70(3.57)	84.91(1.79)	28.33(2.34)	**2.13**
Method	CIFAR-100 (ViT-T)				
	FA	RA	TA	MIA	AG↓
The proportion of forgotten data samples to all samples is 50%					
Retrain	80.96	95.37	81.16	29.85	0
FT	78.92(2.04)	99.87(4.50)	78.40(2.76)	38.49(8.64)	4.48
GA	97.53(16.57)	97.57(2.20)	87.00(5.84)	8.37(21.48)	11.52
IU	88.16(7.20)	89.88(5.49)	80.41(0.75)	15.70(14.15)	6.90
RL	93.11(12.15)	95.52(0.15)	84.45(3.29)	29.61(0.24)	3.96
ℓ_1-sparse	93.95(12.99)	99.93(4.56)	87.36(6.20)	20.18(9.67)	8.36
SalUn	92.36(11.40)	94.27(1.10)	83.60(2.44)	26.03(3.82)	4.69
Lethe	81.84(0.88)	99.23(3.86)	80.89(0.27)	33.21(3.36)	**2.09**

Table 15: Performance of various MU methods for ViT-S on CIFAR-100. The unlearning scenarios include 10%, 30%, and 50% forgetting rates. **Bold** indicates the best performance and underline indicates the runner-up. A performance gap against Retrain is provided in (•).

Method	CIFAR-100 (ViT-S)				
	FA	RA	TA	MIA	AG↓
The proportion of forgotten data samples to all samples is 10%					
Retrain	92.02	99.75	90.99	15.47	0
FT	90.73(1.29)	99.92(0.17)	89.53(1.46)	19.98(4.51)	1.86
GA	98.67(6.65)	98.41(1.34)	91.10(0.11)	5.27(10.20)	4.58
IU	97.82(5.80)	98.16(1.59)	90.85(0.14)	6.58(8.89)	4.11
RL	91.49(0.53)	99.49(0.26)	90.48(0.51)	33.96(18.49)	4.95
ℓ_1-sparse	89.71(2.31)	99.90(0.15)	88.59(2.40)	21.31(5.84)	2.68
SalUn	97.95(5.93)	98.07(1.68)	90.05(0.94)	9.04(6.43)	3.75
Lethe	92.38 (0.36)	99.91 (0.16)	90.31 (0.68)	16.91 (1.44)	**0.66**
Method	CIFAR-100 (ViT-S)				
	FA	RA	TA	MIA	AG↓
The proportion of forgotten data samples to all samples is 30%					
Retrain	90.26	99.76	90.40	18.27	0
FT	94.08 (3.82)	99.95 (0.19)	90.58 (0.18)	15.68 (2.59)	1.70
GA	98.27 (8.01)	98.50 (1.26)	91.09 (0.69)	5.57 (12.70)	5.66
IU	97.14 (6.88)	97.66 (2.10)	89.99 (0.41)	7.36 (10.91)	5.08
RL	94.32 (4.06)	99.23 (0.53)	89.80 (0.60)	36.46 (18.19)	5.85
ℓ_1-sparse	94.24 (3.98)	99.97(0.21)	90.72 (0.32)	15.42 (2.85)	1.84
SalUn	97.75 (7.49)	98.04 (1.72)	90.41 (0.01)	10.96 (7.31)	4.13
Lethe	90.95 0.69	99.90 (0.14)	89.49 (0.91)	20.31 (2.04)	**0.95**
Method	CIFAR-100 (ViT-S)				
	FA	RA	TA	MIA	AG↓
The proportion of forgotten data samples to all samples is 50%					
Retrain	89.68	99.68	89.69	20.93	0
FT	94.17 (4.49)	99.95 (0.27)	90.49 (0.80)	15.95(4.98)	2.64
GA	98.41 (8.73)	98.45 (1.23)	91.09 (1.40)	5.82 (15.11)	6.62
IU	96.32 (6.64)	97.12 (2.56)	89.51 (0.18)	7.67 (13.26)	5.66
RL	96.19 (6.51)	98.74 (0.94)	90.06(0.37)	44.92 (23.99)	7.95
ℓ_1-sparse	94.31 (4.63)	99.96 (0.28)	90.47 (0.78)	15.99 (4.94)	2.66
SalUn	90.33 (0.65)	91.00 (8.68)	83.63 (6.06)	24.38 (3.45)	4.71
Lethe	88.65 (1.03)	99.83 (0.15)	87.74 (1.95)	24.76 (3.83)	**1.74**

Table 16: Performance of various MU methods for DeiT-T on SVHN. The unlearning scenarios include 10%, 30%, and 50% forgetting rates. **Bold** indicates the best performance and underline indicates the runner-up. A performance gap against Retrain is provided in (•).

Method	SVHN (DeiT-T)				
	FA	RA	TA	MIA	AG↓
The proportion of forgotten data samples to all samples is 10%					
Retrain	95.37	98.54	95.86	7.25	0
FT	96.29(0.92)	99.83(1.29)	97.21(1.35)	6.96(0.29)	0.96
GA	97.66(2.29)	97.66(0.88)	95.93(0.07)	5.98(1.27)	1.13
IU	96.71(1.34)	97.12(1.42)	95.61(0.25)	8.21(0.96)	0.99
RL	95.06(0.31)	98.47(0.07)	95.37(0.49)	15.64(8.39)	2.32
ℓ_1-sparse	96.25(0.88)	99.62(1.08)	96.52(0.66)	6.58(0.67)	0.82
SalUn	93.10(2.27)	96.22(2.32)	93.93(1.93)	13.04(5.79)	3.08
Lethe	96.28 (0.91)	99.20 (0.66)	96.55 (0.69)	7.04 (0.21)	**0.62**
Method	SVHN (DeiT-T)				
	FA	RA	TA	MIA	AG↓
The proportion of forgotten data samples to all samples is 30%					
Retrain	94.23	98.38	95.38	9.33	0
FT	96.18 (1.95)	99.66 (1.28)	96.29 (0.91)	7.46 (1.87)	1.50
GA	97.19 (2.96)	97.39 (0.99)	95.62 (0.24)	7.35 (1.98)	1.54
IU	92.80 (1.43)	93.38 (5.00)	92.36 (3.02)	14.91 (5.58)	3.76
RL	93.67 (0.56)	97.03 (1.35)	93.56 (1.82)	15.47(6.14)	2.47
ℓ_1-sparse	96.16 (1.93)	99.66 (1.28)	96.33 (0.95)	7.59 (1.74)	1.48
SalUn	93.42 (0.81)	96.14 (2.24)	93.15 (2.23)	17.04 (7.71)	3.25
Lethe	96.12 (1.89))	99.16 (0.78)	96.57 (1.19)	7.56 (1.77)	**1.40**
Method	SVHN (DeiT-T)				
	FA	RA	TA	MIA	AG↓
The proportion of forgotten data samples to all samples is 50%					
Retrain	93.79	97.93	94.58	11.43	0
FT	96.10 (2.31)	99.69 (1.76)	96.20 (1.62)	7.60 (3.83)	2.38
GA	96.45 (2.66)	96.46 (1.47)	94.91 (0.33)	8.37(3.06)	**1.88**
IU	90.00 (3.79)	90.38 (7.55)	89.68 (4.90)	19.72 (8.29)	6.14
RL	91.18 (2.61)	93.05 (4.88)	90.81 (3.77)	23.40(11.97)	5.81
ℓ_1-sparse	96.11 (2.32)	99.69 (1.76)	96.21 (1.63)	7.69 (3.74)	2.36
SalUn	83.13 (10.66)	86.09 (11.84)	82.41 (12.17)	31.38 (19.95)	13.66
Lethe	96.00 (2.21)	99.25 (1.32)	96.48 (1.90)	7.81 (3.62))	2.26

Table 17: Performance of various MU methods for DeiT-S on SVHN. The unlearning scenarios include 10%, 30%, and 50% forgetting rates. **Bold** indicates the best performance and underline indicates the runner-up. A performance gap against Retrain is provided in (•).

Method	SVHN (DeiT-S)				
	FA	RA	TA	MIA	AG↓
The proportion of forgotten data samples to all samples is 10%					
Retrain	95.57	99.64	96.54	7.28	0
FT	96.50(0.93)	99.86(0.22)	97.09(0.55)	6.69(0.59)	0.57
GA	97.97(2.40)	97.91(1.73)	96.00(0.54)	5.52(1.76)	1.61
IU	97.44(1.87)	97.96(1.68)	95.94(0.60)	6.22(1.06)	1.30
RL	96.72(1.15)	99.77(0.13)	97.34(0.80)	12.88(5.60)	1.92
ℓ_1-sparse	96.41(0.84)	99.87(0.23)	97.00(0.46)	7.04(0.24)	**0.44**
SalUn	97.06(1.49)	99.54(0.10)	97.26(0.72)	12.26(4.98)	1.82
Lethe	96.51(0.94)	99.66(0.02)	96.96(0.42)	6.80(0.48)	0.47
Method	SVHN (DeiT-S)				
	FA	RA	TA	MIA	AG↓
The proportion of forgotten data samples to all samples is 30%					
Retrain	95.10	99.58	96.14	8.30	0
FT	96.08(0.98)	99.86(0.28)	96.91(0.77)	7.72(0.58)	**0.65**
GA	95.78(0.68)	96.00(3.58)	94.38(1.76)	9.09(0.79)	1.70
IU	93.95(1.15)	94.46(5.12)	93.68(2.46)	12.64(4.34)	3.27
RL	96.14(1.04)	99.31(0.27)	96.18(0.04)	11.31(3.01)	1.09
ℓ_1-sparse	95.98(0.88)	99.89(0.31)	96.97(0.83)	7.41(0.89)	0.75
SalUn	96.44(1.34)	99.17(0.41)	96.46(0.32)	10.69(2.39)	1.12
Lethe	96.01(0.91)	99.63(0.05)	96.88(0.74)	7.30(1.00)	0.68
Method	SVHN (DeiT-S)				
	FA	RA	TA	MIA	AG↓
The proportion of forgotten data samples to all samples is 50%					
Retrain	94.38	99.55	95.61	9.23	0
FT	96.03(1.65)	99.89(0.34)	96.64(1.03)	7.66(1.57)	1.15
GA	93.03(1.35)	98.05(1.50)	93.10(2.51)	11.56(2.33)	1.92
IU	90.35(4.03)	90.93(8.62)	90.61(5.00)	17.11(7.88)	6.38
RL	93.47(0.91)	97.57(1.98)	93.51(2.10)	19.65(10.42)	3.85
ℓ_1-sparse	96.07(1.69)	99.89(0.34)	96.65(1.04)	7.80(1.43)	1.13
SalUn	93.78(0.60)	96.61(2.94)	94.19(1.42)	31.00(21.77)	6.68
Lethe	96.02(1.64)	99.65(0.10)	96.84(1.23)	7.73(1.50)	**1.12**