DIMU: IMPORTANCE-GUIDED FEATURE MASKING WITH SELF-DISTILLATION FOR UNLEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Machine unlearning seeks to suppress or remove the influence of designated training data while preserving utility on retained data. We propose Distillation-Regularized Importance-Masked Unlearning (DIMU), an approximate unlearning framework that combines importance-guided feature-channel masking with self-knowledge distillation. DIMU suppresses forget-sensitive channels in the effective penultimate representation exposed to the classifier and adapts only the classifier head while keeping the feature extractor frozen. Self-knowledge distillation regularizes retained predictions using cached pre-unlearning logits. We evaluate class-level unlearning on CIFAR-10/100 across convolutional and transformer-based architectures, random sample-level unlearning on CIFAR-10, and large-scale class-level unlearning on ImageNet-1K using ResNet-50. DIMU achieves zero forgotten-class test accuracy in the primary single-class benchmarks while preserving strong retained utility, and shows favorable forgetting–retention trade-offs under sample-level deletion. It also exhibits competitive empirical membership-distinguishability behavior and a $24.89\times$ speedup over full retraining on ImageNet-1K.

1 INTRODUCTION

Modern deep neural networks can retain sensitive information from their training data Carlini et al. (2019); Hartley & Tsaftaris (2022); Hartley et al. (2023), creating privacy and regulatory concerns when specific data must be removed after deployment. Regulations such as the GDPR Voigt & Von dem Bussche (2017) and CCPA de la Torre (2018) motivate data-erasure obligations and increase the need for practical mechanisms that can remove the influence of requested training data from deployed models. Retraining from scratch without the targeted samples provides the standard exact-unlearning reference Ginart et al. (2019), but is computationally expensive for modern models.

Approximate machine unlearning therefore seeks efficient forgetting while preserving utility on retained data. Existing approaches include gradient-based forgetting Golatkar et al. (2020), retain-only fine-tuning Graves et al. (2021), selective parameter modification such as SalUn Fan et al. (2023), distillation-based methods such as Bad Teacher Chundawat et al. (2023), parameter damping such as SSD Foster et al. (2024), and logit-level correction such as DELETE Zhou et al. (2025). These methods primarily modify model parameters or outputs, while class-discriminative information can remain accessible in internal representations and contribute to residual leakage Shokri et al. (2017).

We therefore consider an effective-representation intervention that controls how forget-associated feature channels contribute to prediction. Rather than modifying the feature extractor itself, our formulation selectively suppresses channels in the penultimate representation before classification and adapts only the classifier head. The underlying backbone features remain unchanged, while the effective representation presented to the classifier is modified through masking.

Our approach draws inspiration from cancelable biometrics, where keyed transformations support reproducibility and revocability Ratha et al. (2007). We build on Cancelable Random Masking (CRM) Hammad et al. (2025) by combining forget-set channel importance with a deterministic key-conditioned mask. The key does not determine channel importance or drive forgetting. Instead, it resolves ties in the importance ranking, enabling deterministic mask regeneration for reproducibility and auditability.

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.

Building on this formulation, we propose Distillation-Regularized Importance-Masked Unlearning (DIMU), a framework for targeted unlearning in common image classifiers. DIMU uses forget-set gradients to identify influential penultimate feature channels, suppresses selected channels with a fixed CRM mask, and adapts the classifier head using forgetting, retention, and Self-Knowledge Distillation (SKD) objectives. SKD aligns retained predictions with cached pre-unlearning logits without maintaining an external teacher model. The framework applies to convolutional and transformer-based classifiers that expose a penultimate representation and a separable classifier head.

The main contributions are:

- **Importance-guided effective-representation suppression.** DIMU suppresses forget-sensitive penultimate channels and adapts only the classifier head while keeping the feature extractor frozen. Key-conditioned tie-breaking preserves the importance ranking and enables reproducible mask regeneration.
- **Retained-behavior preservation.** SKD regularizes classifier adaptation using cached pre-unlearning logits without maintaining an external teacher model.
- **Broad empirical evaluation.** We evaluate DIMU under class-level and sample-level deletion across convolutional and transformer architectures, including ImageNet-1K, efficiency, Retrain-relative diagnostics, and output- and representation-level privacy analyses.

2 Related Work

Machine unlearning seeks to remove the influence of specified training data while preserving utility on retained samples Bourtoule et al. (2021); Ginart et al. (2019); Graves et al. (2021); Foster et al. (2024). The problem is particularly relevant in privacy-sensitive applications, where deployed models may expose information about their training data Shokri et al. (2017); Di et al. (2024). Early work considered convex or structured models, including linear models, clustering, decision trees, and SVMs, for which deletion can exploit closed-form structure or convexity Cao & Yang (2015); Ginart et al. (2019); Izzo et al. (2021); Brophy & Lowd (2021). Extending such guarantees to deep networks is more difficult because of non-convex optimization and large parameter spaces Sekhari et al. (2021); Neel et al. (2021), while full retraining remains computationally expensive.

Approximate unlearning in deep models. Approximate methods intervene at different levels of a trained model. Gradient-based approaches modify predictions on forgotten samples Golatkar et al. (2020), retain-only fine-tuning adapts the model using remaining data Graves et al. (2021), and SalUn selectively updates parameters identified as important for forgetting Fan et al. (2023). Bad Teacher uses distillation-based supervision Chundawat et al. (2023), SSD applies parameter damping Foster et al. (2024), and DELETE introduces logit-level correction Zhou et al. (2025). These approaches provide different balances between forgetting effectiveness and retained utility, but primarily intervene through parameters, optimization objectives, or model outputs rather than explicitly controlling which penultimate feature channels are exposed to the classifier.

Distillation and retained-behavior preservation. Knowledge distillation transfers predictive information from a teacher model Hinton et al. (2015), while self-knowledge distillation can regularize a model using its prior predictions Furlanello et al. (2018). In machine unlearning, Bad Teacher Chundawat et al. (2023) and SCRUB Kurmanji et al. (2023) use teacher-guided objectives to modify forgotten behavior while preserving retained knowledge. DIMU differs by first suppressing forget-sensitive penultimate feature channels and then adapting only the classifier head under a fixed mask. Retained behavior is regularized using cached pre-unlearning logits, so an external teacher model does not need to be maintained during unlearning.

Representation-level unlearning. Intermediate representations can retain training-related information even when model outputs are modified, and representation-extraction or model-stealing studies further demonstrate that learned features can remain informative beyond the final predictions Tramèr et al. (2016). Feature pruning and suppression have also been studied for model compression Han et al. (2015); Molchanov et al. (2016), but are not designed specifically for selective and reproducible unlearning. More recent unlearning approaches explicitly optimize or perturb latent representations Dang et al. (2025); Chen et al. (2025), often in settings such as large language models and different benchmarks. DIMU instead keeps the feature extractor frozen and modifies only the effective representation exposed to the classifier, whereas DIMU-L4 additionally adapts the final residual stage to

reduce residual raw-feature information. Recent benchmarks such as WMDP Li et al. (2024) further emphasize the importance of evaluating unlearning beyond output accuracy.

3 METHOD

DIMU performs selective unlearning by combining importance-guided feature-channel masking with self-knowledge distillation. The method first identifies forget-sensitive channels and constructs a fixed CRM mask, then adapts only the classifier head while keeping the feature extractor frozen. We next formalize the unlearning objective and describe each component in detail.

3.1 PROBLEM FORMULATION

Let $f_\theta = g_\theta \circ h_\theta$ denote a trained classifier with parameters θ, where (x, y) denotes an input-label pair, $h_\theta(x) \in \mathbb{R}^C$ is the C-dimensional penultimate representation, and g_θ maps it to logits $z(x) = g_\theta(h_\theta(x))$. The training dataset $\mathcal{D}$ is partitioned into a retain set $\mathcal{D}_r$ and a forget set $\mathcal{D}_f$. The goal is to suppress the predictive contribution of $\mathcal{D}_f$ while preserving performance on $\mathcal{D}_r$. A conceptual reference is the retain-only solution:

$$\theta_r^\star = \arg\min_\theta \frac{1}{|\mathcal{D}_r|} \sum_{(x,y)\in\mathcal{D}_r} \ell_{\text{CE}}(f_\theta(x), y)\,. \tag{1}$$

Here, $\theta_r^\star$ denotes the retain-only parameters, $|\mathcal{D}_r|$ its size, and ℓ_{CE} the cross-entropy loss. Rather than solving Eq. 1 by retraining from scratch, DIMU keeps the feature extractor fixed and applies a binary channel mask $m \in \{0, 1\}^C$ to obtain

$$\tilde{h}_\theta(x) = m \odot h_\theta(x). \tag{2}$$

In Eq. 2, $\tilde{h}_\theta(x)$ is the effective masked representation supplied to the classifier, while the frozen backbone representation $h_\theta(x)$ remains unchanged. $\odot$ denotes element-wise multiplication, and $m_c = 0$ or 1 suppresses or retains channel c, respectively. Using this representation, the pre-unlearning logits are $z^0(x) = g_\theta(h_\theta(x))$, while the post-unlearning logits are $z'(x) = g_{\theta'}(\tilde{h}_\theta(x))$. DIMU then combines forgetting, retention, and self-knowledge-distillation objectives:

$$\theta_g' = \arg\min_{\theta_g} \mathcal{L}(\theta_g). \tag{3}$$

Here, θ_g denotes the classifier-head parameters and θ_g' their post-unlearning values, while $\mathcal{L}_{\text{forget}}$, $\mathcal{L}_{\text{retain}}$, and $\mathcal{L}_{\text{SKD}}$ respectively suppress forget-set predictions, preserve retained performance, and regularize the updated predictions using cached pre-unlearning logits. In canonical DIMU, Eq. 3 updates only the classifier head under the fixed CRM mask. The following subsections define these terms, whose weighted combination is given in Eq. 8. A conceptual overview is provided in Figure 2 of Appendix B.

3.2 IMPORTANCE-GUIDED CANCELABLE RANDOM MASKING

CRM identifies feature channels that strongly contribute to forget-class predictions and suppresses them using a deterministic key-conditioned mask.

Channel-importance estimation. For each channel c, we compute a gradient-based importance score over the forget set. Here, $h_{\theta,c}(x)$ is the activation of channel c in the penultimate representation $h_\theta(x)$, and I_c in Eq. 4 is the mean absolute cross-entropy gradient with respect to that activation over $\mathcal{D}_f$. Thus, I_c measures forget-set sensitivity at the feature level rather than classifier-weight gradients or activation magnitude. For ViT-S and Swin-T, channels correspond to dimensions of the final pooled embedding before the classifier head rather than individual patch tokens or attention maps. Larger I_c indicates greater local sensitivity of the forget loss to channel c. Let $\mathcal{C}_f \subseteq \{1, \ldots, C\}$ denote the channels selected for suppression, with $m_c = 0$ for $c \in \mathcal{C}_f$ and $m_c = 1$ otherwise. Let $\Delta\ell_{\text{CE}}^{(1)}(x, y)$ denote the first-order change in cross-entropy loss induced by masking. Then, Eq. 5 gives

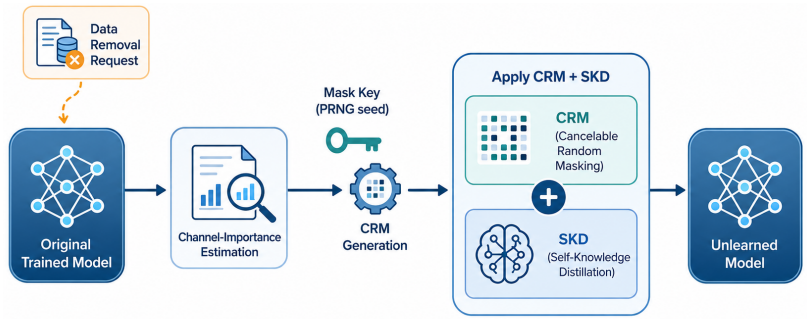


Figure 1: **DIMU framework.** A forget request triggers channel-importance estimation and deterministic CRM mask generation. The feature extractor remains frozen while the classifier head is adapted under the fixed mask using forgetting, retention, and SKD objectives.

$$I_c = \frac{1}{|\mathcal{D}_f|} \sum_{(x,y)\in\mathcal{D}_f} \left| \frac{\partial \ell_{\mathrm{CE}}(f_\theta(x), y)}{\partial h_{\theta,c}(x)} \right| \tag{4}$$

$$\left|\Delta \ell^{(1)}_{\mathrm{CE}}(x, y)\right| \le \sum_{c\in\mathcal{C}_f} |h_{\theta,c}(x)| \left| \frac{\partial \ell_{\mathrm{CE}}}{\partial h_{\theta,c}(x)} \right| \tag{5}$$

This local sensitivity bound motivates feature-level intervention but does not uniquely determine the ranking criterion. We therefore use forget-set gradient magnitude as the default importance estimator, supported empirically by Appendix Table 13.

Key-conditioned mask generation. A binary mask $m \in \{0,1\}^C$ is generated as $m = \mathrm{CRM}(I, k, \tau)$, where I is the importance vector, k is the mask key, and $\tau \in (0,1)$ controls the fraction of channels to suppress. Let π_k be the deterministic pseudo-random ordering induced by key k. Channels are ranked by importance, with π_k used only to break ties, and the top $\lceil \tau C \rceil$ channels are suppressed:

$$m_c = \mathbb{I}\Big[c \notin \mathrm{Top}_{\lceil \tau C \rceil}\big(\mathrm{rank}(I; \pi_k)\big)\Big]. \tag{6}$$

The key does not alter the importance scores or optimization objective. For a fixed importance vector, it only resolves ties in the channel ranking, enabling reproducible and auditable mask generation. Forgetting is driven by importance-based channel suppression and the subsequent unlearning objectives. This deterministic masking formulation is used throughout our experiments.

Feature suppression. Using Eq. 2, the masked model is $f_{\theta,m}(x) = g_\theta(m \odot h_\theta(x))$. The classifier therefore receives the masked effective representation rather than the unmodified penultimate representation, suppressing selected channels used for prediction without modifying the frozen backbone.

3.3 Self-Knowledge Distillation

Before unlearning, we cache the raw pre-unlearning logits $z^0(x_r) = f_\theta(x_r)$ for retain samples $x_r \in \mathcal{D}_r$. The cache stores logits rather than temperature-softened probabilities and remains fixed throughout unlearning. The masked model produces $z'(x_r) = f_{\theta,m}(x_r) = g_\theta(m \odot h_\theta(x_r))$.

The SKD loss aligns the updated retain predictions with the cached pre-unlearning predictions:

$$\mathcal{L}_{\mathrm{SKD}} = \frac{T^2}{|\mathcal{D}_r|} \sum_{x_r\in\mathcal{D}_r} \mathrm{KL}\left(\sigma\left(\frac{z^0(x_r)}{T}\right) \middle\| \sigma\left(\frac{z'(x_r)}{T}\right)\right). \tag{7}$$

Here, $\sigma(\cdot)$ is the softmax function and T is the distillation temperature. Thus, the loss uses $\mathrm{KL}(p_{\mathrm{ref}} \| p_{\mathrm{stu}})$. The default setting is $T = 4$. Because T is applied only when computing the loss, the stored cache is independent of temperature. In the temperature-sensitivity analysis, λ_{SKD} is held fixed while T is varied, with all other hyperparameters fixed.

SKD avoids maintaining an external teacher model, but the retained-logit cache has storage cost $O(|D_r|K)$ and contains model-derived confidence information. With FP32 logits, the cache is 19.8 MB on CIFAR-100 and approximately 5.12 GB on ImageNet-1K. It should therefore be access-protected during unlearning and deleted once no longer required.

3.4 COMBINED UNLEARNING OBJECTIVE

The final DIMU objective combines forgetting, retention, and self-knowledge distillation:

$$\mathcal{L} = \lambda_f \mathcal{L}_{\text{forget}} + \lambda_r \mathcal{L}_{\text{retain}} + \lambda_{\text{SKD}} \mathcal{L}_{\text{SKD}}. \tag{8}$$

$$\mathcal{L}_{\text{forget}} = \frac{1}{|\mathcal{D}_f|} \sum_{(x_f, y_f) \in \mathcal{D}_f} [-\ell_{\text{CE}}(f_{\theta,m}(x_f), y_f)] \tag{9}$$

$$\mathcal{L}_{\text{retain}} = \frac{1}{|\mathcal{D}_r|} \sum_{(x_r, y_r) \in \mathcal{D}_r} \ell_{\text{CE}}(f_{\theta,m}(x_r), y_r) \tag{10}$$

The forgetting term in Eq. 9 maximizes cross-entropy on forget samples, while the retention term in Eq. 10 preserves predictive performance on retained data.

Although Eq. 9 is unbounded below in isolation, DIMU uses finite-step optimization, retention and SKD regularization, and weight decay, with diagnostics in Appendix Figure 4. Together, these components suppress forget predictions while preserving retained performance.

The complete DIMU procedure is summarized in Algorithm 1 in Appendix A.1. During unlearning, h_θ remains frozen and only g_θ is optimized under the fixed CRM mask. At deployment, inference uses the frozen feature extractor h_θ, the updated classifier head $g_{\theta'}$, and the fixed CRM mask $\mathbf{m}$. DIMU-L4 additionally adapts ResNet's layer4 jointly with the classifier head. Details are in Appendix A.

4 EXPERIMENTS

We evaluate DIMU across multiple datasets, architectures, and unlearning scenarios using utility and privacy metrics. Table 8 summarizes the main DIMU hyperparameters, with additional implementation details in Appendix A.

Datasets and Backbones. Experiments are conducted on CIFAR-10 and CIFAR-100 (Krizhevsky et al., 2009) using 32×32 preprocessing, together with large-scale evaluation on ImageNet-1K. We evaluate ResNet-18 (He et al., 2016), VGG-16 (Simonyan & Zisserman, 2014), ViT-S (Dosovitskiy et al., 2020), and Swin-T (Liu et al., 2021) on CIFAR, and ResNet-50 on ImageNet-1K.

Unlearning Scenarios. We evaluate both class-level and sample-level unlearning. For CIFAR-10 single-class experiments, the forgotten class is class index 1 (automobile). CIFAR-100 experiments use class index 0 for single-class forgetting and class-index sets 0–1, 0–4, and 0–9 for multi-class forgetting. Sample-level experiments randomly remove 1%, 5%, or 10% of the CIFAR-10 training samples using ResNet-18. For ImageNet-1K, we evaluate single-class unlearning on three fixed forget classes: 10, 138, and 583. All reported experiments use the deterministic masking formulation described in Section 3.

Training Setup. For CIFAR experiments, each Original model is trained from random initialization on the full training set for 150 epochs, and DIMU starts from the corresponding trained checkpoint. Unless otherwise specified, DIMU performs 40 unlearning epochs with the feature extractor h_θ frozen and only the classifier head g_θ updated under the fixed CRM mask. For ImageNet-1K, DIMU starts from the torchvision IMAGENET1K_V2 pretrained ResNet-50. Full optimization, masking, cache-storage, and hardware details are provided in Table 8 and Appendix A.

Evaluation Metrics. We report forget and retain-set training accuracies (Acc_f, Acc_r) and held-out forget and retain-class test accuracies (Acc_{ft}, Acc_{rt}). Following Zhao et al. (2024a), we summarize the balance between forgetting and retained utility using H-Mean:

$$\text{H-Mean} = \frac{2\,\text{Acc}_{rt}\,\text{Drop}}{\text{Acc}_{rt} + \text{Drop}}, \qquad \text{Drop} = \text{Acc}_{ft}^{\text{orig}} - \text{Acc}_{ft}, \tag{11}$$

where $\text{Acc}_{ft}^{\text{orig}}$ denotes the Original forget-class test accuracy. Negative H-Mean occurs when forgotten-class accuracy increases above the Original level. Privacy is assessed using complementary membership-distinguishability diagnostics following Shokri et al. (2017). We report the confidence-based score $|2(\text{AUROC} - 0.5)|$, where lower values indicate lower distinguishability. Acc_f/Acc_r denote forget/retain training accuracy, while Acc_{ft}/Acc_{rt} denote held-out forgotten/retained-class test accuracy. Adaptive white-box output and representation-level attacks are reported separately in Table 3.

Table 1: Performance comparison of class-level unlearning on ResNet-18 across CIFAR-10 and CIFAR-100. Acc_f/Acc_r denote forget/retain training accuracy, Acc_ft/Acc_rt denote forget/retain test accuracy, H-Mean is the harmonic mean, and MIA is the membership inference attack score. Throughout the paper, we highlight the best values among the unlearning methods in **bold**, and we highlight selected key DIMU results in blue.

Method	CIFAR-10						CIFAR-100					
	$\text{Acc}_\text{f}\downarrow$	$\text{Acc}_\text{r}\uparrow$	$\text{Acc}_\text{ft}\downarrow$	$\text{Acc}_\text{rt}\uparrow$	H-Mean↑	MIA↓	$\text{Acc}_\text{f}\downarrow$	$\text{Acc}_\text{r}\uparrow$	$\text{Acc}_\text{ft}\downarrow$	$\text{Acc}_\text{rt}\uparrow$	H-Mean↑	MIA↓
Original	100.00	100.00	96.60	95.27	–	0.99	100.00	99.96	72.00	**79.41**	–	0.97
Retrain	**0.00**	**100.00**	**0.00**	95.36	95.98	**0.00**	**0.00**	**99.97**	**0.00**	78.53	**75.12**	**0.00**
DELETE Zhou et al. (2025)	0.76	**100.00**	6.30	95.47	92.81	0.013	**0.00**	99.40	**0.00**	77.08	74.51	**0.00**
SCRUB Kurmanji et al. (2023)	1.64	89.23	**0.00**	86.86	91.47	0.01	**0.00**	82.89	**0.00**	67.24	69.54	**0.00**
SalUn Fan et al. (2023)	17.12	99.26	43.70	92.87	67.41	0.16	3.00	99.76	2.00	75.32	72.58	0.02
Random Label Golatkar et al. (2020)	17.64	98.34	38.80	90.54	70.54	0.17	3.80	99.75	3.00	75.23	72.01	0.03
Bad Teacher Chundawat et al. (2023)	7.44	96.10	34.30	88.26	73.06	0.06	4.40	99.78	2.00	75.37	72.60	0.03
Negative Gradient Golatkar et al. (2020)	0.00	17.00	0.00	17.01	28.93	**0.00**	0.20	98.81	0.00	73.11	72.58	**0.00**
DIMU	**0.00**	99.98	**0.00**	**95.99**	**96.27**	0.003	**0.00**	99.44	**0.00**	78.38	75.05	0.00

Table 2: Multi-class unlearning performance when forgetting 2, 5, and 10 classes.

Method	2 Classes					5 Classes					10 Classes				
	$\text{Acc}_f\downarrow$	$\text{Acc}_r\uparrow$	$\text{Acc}_{ft}\downarrow$	$\text{Acc}_{rt}\uparrow$	H-Mean↑	$\text{Acc}_f\downarrow$	$\text{Acc}_r\uparrow$	$\text{Acc}_{ft}\downarrow$	$\text{Acc}_{rt}\uparrow$	H-Mean↑	$\text{Acc}_f\downarrow$	$\text{Acc}_r\uparrow$	$\text{Acc}_{ft}\downarrow$	$\text{Acc}_{rt}\uparrow$	H-Mean↑
Original	100.00	99.98	79.50	79.34	-	100.00	99.98	86.20	78.98	-	99.96	99.99	79.20	79.36	-
Retrain	**0.00**	**99.98**	**0.00**	78.52	79.04	**0.00**	**99.99**	**0.00**	77.97	81.83	**0.00**	99.97	**0.00**	78.73	78.89
DELETE	0.50	99.97	0.50	78.56	78.84	0.12	99.96	**0.00**	78.36	82.00	3.16	99.97	1.60	79.49	78.59
SCRUB	**0.00**	81.99	**0.00**	70.92	74.93	**0.00**	88.54	**0.00**	71.85	78.40	0.02	89.12	**0.00**	72.03	75.45
SalUn	17.80	99.88	10.00	76.24	72.74	1.16	85.59	**0.00**	57.09	68.68	2.94	79.45	1.10	55.93	65.18
Random Label	17.70	99.84	8.50	76.01	73.40	2.40	98.12	0.60	70.21	77.09	1.90	74.34	0.60	51.07	61.86
Bad Teacher	17.40	99.88	10.00	76.34	72.74	1.68	97.46	0.40	69.12	76.55	0.50	57.64	0.20	39.20	52.38
Negative Gradient	**0.00**	96.82	**0.00**	71.07	75.04	**0.00**	86.08	**0.00**	60.51	71.11	0.46	66.56	0.30	46.98	58.81
DIMU	**0.00**	99.00	0.10	**79.06**	**79.23**	0.10	99.27	0.15	**79.90**	**82.86**	1.99	**99.95**	1.93	**79.99**	**78.61**

Baseline Methods. We compare DIMU against representative unlearning methods spanning different intervention strategies. Retrain trains a new model from scratch on $\mathcal{D}_r$ and serves as the retain-only reference for exact unlearning. Negative Gradient and Random Label Golatkar et al. (2020) explicitly push predictions away from forgotten samples. SalUn Fan et al. (2023) identifies and selectively updates parameters that are important for forgetting. Bad Teacher Chundawat et al. (2023) and SCRUB Kurmanji et al. (2023) use teacher-guided or retained-data objectives to alter forgotten behavior while preserving retained knowledge. DELETE Zhou et al. (2025) uses decoupled logit-level correction, while these approaches collectively cover gradient-based, parameter-level, distillation-based, and output-level unlearning strategies. Some methods, including SCRUB and Retrain, explicitly use retained data during optimization and therefore receive stronger retained-data supervision than lightweight forgetting-only approaches. For each baseline, we use the official implementation when available and adapt method-specific hyperparameters to the common data splits and training protocols.

5 Results and Discussion

We evaluate DIMU across CIFAR-10/100 and ImageNet-1K to assess forgetting, retention, scalability, efficiency, multi-class unlearning, and cross-architecture generalization.

5.1 Forgetting and Retention Performance

We evaluate DIMU on ResNet-18 for CIFAR-10/100 single- and multi-class unlearning against established baselines. Table 1 reports forget/retain training and test accuracies. DIMU achieves $\text{Acc}_{ft} = 0.00$ on both datasets. On CIFAR-10, it obtains $\text{Acc}_{rt} = 95.99$ and H-Mean 96.27, exceeding DELETE (92.81) and SCRUB (91.47), while SalUn under-forgets and Negative Gradient degrades retained utility. On CIFAR-100, DIMU achieves $\text{Acc}_{rt} = 78.38$ and H-Mean 75.05, closely matching Retrain (75.12) and exceeding DELETE (74.51) and SCRUB (69.54). Since zero forgotten-class accuracy alone does not imply retraining-equivalent unlearning, Retrain-relative diagnostics consistently show closer retained than forgotten behavior across CIFAR-10, CIFAR-100, and ImageNet-1K (Appendix Table 24). Confidence-based MIA scores are treated as empirical distinguishability indicators rather than formal privacy guarantees. DIMU also exceeds SCRUB and Bad Teacher in H-Mean on both datasets.

For multi-class unlearning on CIFAR-100 (Table 2), DIMU obtains H-Mean values of 79.23, 82.86, and 78.61 when forgetting 2, 5, and 10 classes, respectively. It achieves the highest H-Mean for 2 and 5 classes, while Retrain is slightly higher for 10 classes (78.89 vs. 78.61). In the 10-class setting, DIMU retains the highest $\text{Acc}_{rt} = 79.99$ with $\text{Acc}_{ft} = 1.93$. Across five fixed CIFAR-100

target classes, DIMU maintains near-zero forgotten-class accuracy ($Acc_{ft} = 0.22 \pm 0.44\%$) with stable retained accuracy ($Acc_{rt} = 78.09 \pm 0.19\%$) (Appendix Table 20). Separately, Retrain-relative diagnostics show closer retained than forgotten behavior (Appendix B.3).

Forget-set difficulty analysis further shows lower MMD on CIFAR-100 (~ 0.0189) than for the evaluated CIFAR-10 subsets (0.0444–0.0921), indicating greater forget–retain feature overlap, with complementary normalized ToW-style trade-off analysis reported in Appendix B.3.

Class-Level Robustness. Across three deterministic keys, DIMU maintains $Acc_{ft} = 0.00$ while Acc_{rt} varies only from 78.34 to 78.38, showing highly consistent forgetting and retention across keys (Appendix Table 19). Three-seed results further show low performance variance (Appendix Table 21), while the prediction-distribution analysis shows that forgotten samples are reassigned to retained classes rather than remaining concentrated on the forgotten label (Appendix Figure 5).

Representation Analysis. Qualitative t-SNE visualizations show reduced separability of forgotten-class samples in DIMU's CRM-masked effective representation while largely preserving retained-class structure (Appendix Figure 3). Because t-SNE is sensitive to projection and sampling choices, we use it only as supporting evidence and therefore quantify forgotten-class accessibility using linear-probe analysis.

Linear-Probe Analysis. We train a linear classifier on the Original model's raw penultimate representations $h_\theta(x)$ using $\mathcal{D}_r \cup \mathcal{D}_f$, and evaluate the same probe without further optimization on Original raw features, Retrain raw features $h_{\theta_r^\star}(x)$, and DIMU's CRM-masked effective features $\tilde{h}_\theta(x) = m \odot h_\theta(x)$. Forgotten-class probe accuracy decreases from 95.70% for Original to 0.10% for DIMU, indicating strong disruption of the forgotten-class decision signal in the deployed effective representation. Retrain yields 0.00%, reflecting limited transferability of the Original-trained probe to its independently learned retain-only representation rather than evidence of complete information erasure.

Adaptive Privacy and Representation Analysis. Table 3 evaluates adaptive white-box attacks using outputs and penultimate representations. On CIFAR-10, DIMU reduces output-level AUROC^* from 0.991 for Original to 0.614, while CRM masking reduces feature-level AUROC^* from 0.989 on raw features to 0.672 on masked features. The high raw-feature score is expected because the canonical DIMU backbone remains frozen. DIMU-L4 further reduces raw/masked-feature scores to 0.616/0.549 while maintaining 0.00% forgotten-class and 95.40% retained-class accuracy. The same pattern holds on CIFAR-100, where DIMU-L4 achieves raw/masked-feature AUROC^* of 0.620/0.550 with 0.00% forgotten-class and 77.85% retained-class accuracy (Appendix Table 12). Retrain's raw-feature score remains above chance at 0.684; therefore, these attacks are treated as empirical distinguishability diagnostics rather than formal privacy guarantees and are not numerically comparable to the confidence-based MIA metric in Table 1.

Table 3: Adaptive privacy and representation analysis on CIFAR-10 using ResNet-18. DIMU freezes the backbone and adapts only the classifier head, while DIMU-L4 additionally adapts the final residual stage. Attacks use model outputs, raw penultimate features, or deployed CRM-masked features. We report direction-independent $\mathrm{AUROC}^* = \max(\mathrm{AUROC}, 1 - \mathrm{AUROC})$ and $\mathrm{BA}^* = \max(\mathrm{BA}, 1 - \mathrm{BA})$, where BA denotes balanced accuracy and values closer to 0.5 indicate lower distinguishability. Attack scores are mean $\pm$ standard deviation over three seeded attacker splits of a fixed checkpoint, while accuracies use the same checkpoint.

Method	Trainable Scope	$\mathbf{Acc}_{ft} \downarrow$	$\mathbf{Acc}_{rt} \uparrow$	Output MIA $\mathbf{AUROC}^*$	Output MIA $\mathbf{BA}^*$	Raw-feature MIA $\mathbf{AUROC}^*$	Masked-feature MIA $\mathbf{AUROC}^*$
Original		96.60	95.27	0.991 ± 0.004	0.948 ± 0.009	0.991 ± 0.003	
Retrain	All, from scratch	0.00	95.36	0.506 ± 0.004	0.508 ± 0.006	0.684 ± 0.015	
DIMU	Head only	**0.00**	**95.99**	0.614 ± 0.018	0.589 ± 0.015	0.989 ± 0.004	0.672 ± 0.019
DIMU-L4	Layer4 + head	**0.00**	95.40	0.540 ± 0.010	0.530 ± 0.009	0.616 ± 0.020	0.549 ± 0.012

Cross-Architecture Evaluation. We evaluate DIMU across convolutional and transformer backbones, including ResNet-18, VGG-16, Swin-T, and ViT-S. DIMU achieves zero forgotten-class test accuracy ($Acc_{ft} = 0.00$) across all evaluated architectures on both CIFAR-10 and CIFAR-100. Retained-test accuracy remains 87.26/80.72 for VGG-16, 85.13/67.16 for Swin-T, and 79.82/64.96 for ViT-S on CIFAR-10/CIFAR-100, respectively, demonstrating consistent forgetting with preserved retained utility across substantially different backbone families. Detailed results are reported in Appendix B (Tables 9, 10, and 11).

Table 4: Single-class unlearning on ImageNet-1K using ResNet-50 across three fixed forget classes. The Original model uses torchvision IMAGENET1K_V2 weights. Lower Acc_f, Acc_{ft}, and MIA and higher Acc_r, Acc_{rt}, and H-Mean indicate better performance.

Forget Class	Method	$\mathbf{Acc}_f \downarrow$	$\mathbf{Acc}_r \uparrow$	$\mathbf{Acc}_{ft} \downarrow$	$\mathbf{Acc}_{rt} \uparrow$	H-Mean↑	MIA↓
10	Original	97.38	91.68	98.00	80.83	–	–
	Retrain	**0.00**	87.49	**0.00**	76.45	85.89	0.2694
	Retrain-FT	97.00	**91.70**	96.00	**80.76**	3.90	**0.1454**
	DIMU	**0.00**	90.15	**0.00**	79.53	**87.80**	0.2294
138	Original	98.23	91.68	98.00	80.83	–	–
	Retrain	**0.00**	87.35	**0.00**	76.34	85.82	0.1030
	Retrain-FT	97.69	**91.85**	100.00	**80.80**	-4.10	**0.0149**
	DIMU	**0.00**	90.29	**0.00**	79.40	**87.73**	0.0276
583	Original	95.00	91.68	90.00	80.84	–	–
	Retrain	**0.00**	87.36	**0.00**	76.11	82.47	0.1607
	Retrain-FT	93.53	**91.84**	88.00	**80.79**	3.90	0.1691
	DIMU	**0.00**	89.62	**0.00**	79.20	**84.25**	**0.0511**

5.2 Large-Scale ImageNet-1K Evaluation and Efficiency

On ImageNet-1K with pretrained ResNet-50, DIMU drives Class-10 forget-train/test accuracy to 0.00% while retaining 79.53% test accuracy versus 76.45% for Retrain (Table 4). Across the three forget-class identities, mean retained accuracy is $79.38 \pm 0.17\%$ and mean H-Mean is 86.59 ± 2.03 (mean $\pm$ SD across class identities). For Class 10, DIMU achieves H-Mean 87.80 and MIA 0.2294 versus 0.2694 for Retrain. Retrain-FT has lower MIA (0.1454) but retains 96.00% forgotten-class test accuracy. DIMU takes 1.56 h, compared with 22.98 h for Retrain-FT and 38.73 h for Retrain, giving a $24.89\times$ speedup. On CIFAR-10/100, DIMU gives approximately $1.54\times$ speedup over Retrain with ResNet-18 and ViT-S, while lightweight approximate baselines remain faster. Detailed runtimes are in Appendix Table 16.

5.3 Ablation Study

We conduct an ablation study on CIFAR-100 using ResNet-18 to isolate the contributions of CRM and SKD. Table 6 shows that removing CRM (`w/o CRM`) leaves measurable forgotten-class accuracy ($\mathrm{Acc}_{\mathrm{ft}}$ = 3.20), with H-Mean 71.71 and MIA 0.35, while removing SKD (`w/o SKD`) yields $\mathrm{Acc}_{\mathrm{ft}}$ = 1.80, reduces $\mathrm{Acc}_{\mathrm{rt}}$ to 71.34, and increases MIA to 0.23. The single-module variants further support the complementary roles of CRM and SKD. Full DIMU achieves $\mathrm{Acc}_{\mathrm{ft}}$=0.00, $\mathrm{Acc}_{\mathrm{rt}}$ = 78.38, H-Mean 75.05, and MIA 0.00. A head-only Retain-FT baseline retains 12.75% forgotten-class accuracy with H-Mean 67.16, while one epoch of forgotten-class relearning recovers 41.37 points for Head-only Retain-FT, 22.40 for DIMU, and 9.35 for Retrain, showing that DIMU is less susceptible to rapid relearning than simple head-only adaptation, while Retrain remains the strongest reference (Appendix Tables 17 and 18). Table 7 further shows that the default mask ratio $\tau = 0.05$ provides the best forgetting and retention trade-off. Smaller masks leave residual forgotten-class accuracy, while larger masks reduce retained accuracy and eventually weaken forgetting performance. Channel-ranking ablation favors the default gradient criterion (Appendix Table 13), while DIMU remains stable across $T \in \{1, 2, 4, 8\}$. The default $T = 4$ gives zero forgotten-class accuracy and the highest retained accuracy of 78.38, with full three-seed results reported in Appendix Table 14. Epoch sensitivity further supports the fixed 40-epoch budget (Appendix Table 15).

Sample-Level Unlearning. We evaluate random sample-level deletion on CIFAR-10 using ResNet-18 by removing 1%, 5%, and 10% of the training samples over three seeds. Table 5 compares DIMU with retain-only Retrain using identical forget subsets. Retrain retains 91.19–91.87% accuracy on removed samples because same-class examples remain in the retain set, whereas DIMU yields 65.89–68.47%. We interpret lower removed-sample accuracy as targeted suppression rather than retraining-equivalent forgetting and assess sample-level forgetting jointly through removed-sample accuracy, confidence reduction, membership-inference leakage, and retained utility. Across five sequential deletion requests, same-key, new-key, composed-mask, and recomputed-importance strategies maintain Acc_r between 97.22 and 97.46, test accuracy between 89.32 and 89.57, and MIA at 0.0000 (Appendix Table 22). Direct comparison with Retrain further shows greater predictive divergence on forgotten than retained samples ($\mathrm{Sym.\,KL}_f = 1.818$–$2.177$ versus $\mathrm{Sym.\,KL}_r = 0.189$–$0.366$), higher

Table 5: Random sample-level unlearning on CIFAR-10 using ResNet-18 by removing randomly selected 1%, 5%, and 10% training subsets, averaged over three seeds. DIMU and Retrain use identical forget subsets.

Forget Ratio	Method	$\text{Acc}_r \uparrow$	$\text{Acc}_f \downarrow$	$\text{Acc}_{test} \uparrow$	Conf. $\text{Drop}_f \uparrow$	MIA↓	Time (min)↓
1%	DIMU	97.28 ± 0.30	68.47 ± 1.01	90.42 ± 0.13	**33.13 ± 1.44**	0.0000 ± 0.0000	9.7 ± 0.6
	Retrain	99.80 ± 0.07	91.87 ± 0.64	91.55 ± 0.12	5.23 ± 0.29	0.0128 ± 0.0128	26.6 ± 0.1
5%	DIMU	96.93 ± 0.12	67.95 ± 0.95	89.27 ± 0.19	**35.01 ± 0.81**	0.0000 ± 0.0000	14.5 ± 0.7
	Retrain	99.73 ± 0.09	91.19 ± 0.81	91.27 ± 0.19	6.29 ± 0.48	0.0198 ± 0.0195	25.5 ± 0.1
10%	DIMU	96.79 ± 0.13	65.89 ± 1.80	88.26 ± 0.25	**36.62 ± 1.74**	0.0000 ± 0.0000	18.0 ± 0.1
	Retrain	99.82 ± 0.02	91.22 ± 0.37	91.01 ± 0.07	6.39 ± 0.18	0.0024 ± 0.0042	24.2 ± 0.1

Table 6: Ablation of CRM and SKD on ResNet-18 using CIFAR-100.

Configuration	$\text{Acc}_{ft} \downarrow$	$\text{Acc}_{rt} \uparrow$	H-Mean↑	MIA↓
DIMU	**0.00**	**78.38**	**75.05**	**0.00**
w/o CRM	3.20	74.87	71.71	0.35
w/o SKD	1.80	71.34	70.77	0.23
Only CRM	1.40	76.23	73.31	0.18
Only SKD	0.90	77.18	74.02	0.15

Table 7: Masking-ratio sensitivity on ResNet-18 using CIFAR-100.

τ	$\text{Acc}_{\text{ft}} \downarrow$	$\text{Acc}_{\text{rt}} \uparrow$	H-Mean↑	MIA↓
0.01	5.00	71.10	68.99	0.02
0.03	1.00	75.00	72.95	0.01
0.05	0.00	78.38	75.05	0.00
0.07	0.00	75.10	73.52	0.00
0.09	2.00	72.13	71.06	0.01

retained-sample representation alignment ($\text{CKA}_r = 0.807$–0.854 versus $\text{CKA}_f = 0.740$–0.762), and near-identical test-set confusion structure (0.998–0.999) across deletion ratios (Appendix Table 23).

6 Limitations and Future Directions

DIMU is evaluated for class-level and sample-level unlearning on CIFAR-10/100 and large-scale class-level unlearning on ImageNet-1K across convolutional and transformer architectures. The framework assumes a penultimate representation and separable classifier head, so architectures without a clear feature–classifier interface may require adaptation. ImageNet-1K evaluation currently covers only class-level deletion, leaving large-scale sample-level, more heterogeneous deletion settings, and longer sequential unlearning horizons for future work.

Our representation and privacy results should be interpreted as empirical diagnostics rather than certified removal guarantees. In canonical DIMU, the backbone remains frozen, so information associated with forgotten data can remain accessible in the raw representation even when its contribution to deployed predictions is suppressed by the CRM mask. The DIMU-L4 results show that limited representation adaptation can further reduce this residual signal, motivating stronger representation-level unlearning. DIMU also relies on cached pre-unlearning retain logits during SKD, introducing storage and access-control requirements and motivating secure cache deletion after unlearning. Future work will extend DIMU to open-world, multimodal, and federated learning settings.

7 Conclusion

We introduced DIMU, an importance-guided selective unlearning framework that combines CRM feature-channel masking with self-knowledge distillation. DIMU identifies forget-sensitive channels, suppresses their contribution in the effective masked representation, and adapts only the classifier head while preserving the underlying feature extractor. Across CIFAR-10/100, DIMU achieves strong class-level forgetting and favorable sample-level forgetting–retention trade-offs across convolutional and transformer-based architectures, while ImageNet-1K demonstrates scalability to a substantially larger classification setting. Retrain-relative, representation-level, and adaptive privacy analyses further characterize its behavior beyond output accuracy, while the ImageNet-1K results demonstrate substantially lower unlearning cost than retain-only fine-tuning and full Retrain. DIMU does not require an external teacher model, and its key-conditioned masking rule enables deterministic and reproducible mask generation. Overall, the results support importance-guided feature suppression with retained-behavior regularization as an effective and computationally practical approach to approximate machine unlearning.

AI USE STATEMENT

Generative AI tools were used for language editing, LaTeX formatting assistance, and support with code inspection and debugging. All AI-assisted text and code suggestions were reviewed and verified by the authors. All experiments, numerical results, scientific claims, and interpretations were independently validated by the authors, who take full responsibility for the final content of this work.

ETHICS STATEMENT

Machine unlearning is increasingly important for responsible AI development, particularly in settings where users may request removal of their data from trained models. This work studies approximate unlearning methods for improving efficient and selective data removal. All datasets in our experiments are publicly available benchmarks used for research evaluation. We hope this study contributes to the development of more practical and privacy-aware machine unlearning methods.

REPRODUCIBILITY STATEMENT

Details required to reproduce our results are provided in Section 3, Section 4, and Appendix A. These sections describe the DIMU formulation, experimental setup, evaluation metrics, hyperparameters, training configurations, and implementation details. Additional analyses and results are reported in the appendix. Code and experimental scripts will be released upon acceptance.

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.

Jonathan Brophy and Daniel Lowd. Machine unlearning for random forests. In *International Conference on Machine Learning*, pp. 1092–1104. PMLR, 2021.

Yinzhi Cao and Junfeng Yang. Towards making systems forget with machine unlearning. In *2015 IEEE symposium on security and privacy*, pp. 463–480. IEEE, 2015.

Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX security symposium (USENIX security 19)*, pp. 267–284, 2019.

Taozhao Chen, Linghan Huang, Kim-Kwang Raymond Choo, and Huaming Chen. Feature-selective representation misdirection for machine unlearning. *arXiv preprint arXiv:2512.16297*, 2025.

Vikram S Chundawat, Ayush K Tarun, Murari Mandal, and Mohan Kankanhalli. Can bad teaching induce forgetting? unlearning in deep networks using an incompetent teacher. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 7210–7217, 2023.

Huu-Tien Dang, Tin Pham, Hoang Thanh-Tung, and Naoya Inoue. On effects of steering latent representation for large language model unlearning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 23733–23742, 2025.

Lydia de la Torre. A guide to the california consumer privacy act of 2018. *Available at SSRN 3275571*, 2018.

Zonglin Di, Sixie Yu, Yevgeniy Vorobeychik, and Yang Liu. Adversarial machine unlearning. *arXiv preprint arXiv:2406.07687*, 2024.

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, Yihua Zhang, Eric Wong, Dennis Wei, and Sijia Liu. Salun: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation. *arXiv preprint arXiv:2310.12508*, 2023.

Jack Foster, Stefan Schoepf, and Alexandra Brintrup. Fast machine unlearning without retraining through selective synaptic dampening. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 12043–12051, 2024.

Tommaso Furlanello, Zachary Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. Born again neural networks. In *International conference on machine learning*, pp. 1607–1616. PMLR, 2018.

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.

Mohamed Hammad, Abdelhamied A Ateya, Mohammed ElAffendi, and Ahmed A Abd El-Latif. Cancelable random masking with deep learning for secure and interpretable finger vein authentication. *Intelligent Systems with Applications*, 27:200552, 2025.

Song Han, Jeff Pool, John Tran, and William Dally. Learning both weights and connections for efficient neural network. *Advances in neural information processing systems*, 28, 2015.

John Hartley and Sotirios A Tsaftaris. Measuring unintended memorisation of unique private features in neural networks. *arXiv preprint arXiv:2202.08099*, 2022.

John Hartley, Pedro P Sanchez, Fasih Haider, and Sotirios A Tsaftaris. Neural networks memorise personal information from one sample. *Scientific Reports*, 13(1):21366, 2023.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

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.

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

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

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

Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D Li, Ann-Kathrin Dombrowski, Shashwat Goel, Long Phan, et al. The wmdp benchmark: Measuring and reducing malicious use with unlearning. *arXiv preprint arXiv:2403.03218*, 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.

Pavlo Molchanov, Stephen Tyree, Tero Karras, Timo Aila, and Jan Kautz. Pruning convolutional neural networks for resource efficient inference. *arXiv preprint arXiv:1611.06440*, 2016.

Seth Neel, Aaron Roth, and Saeed Sharifi-Malvajerdi. Descent-to-delete: Gradient-based methods for machine unlearning. In *Algorithmic Learning Theory*, pp. 931–962. PMLR, 2021.

Nalini K Ratha, Sharat Chikkerur, Jonathan H Connell, and Ruud M Bolle. Generating cancelable fingerprint templates. *IEEE Transactions on pattern analysis and machine intelligence*, 29(4): 561–572, 2007.

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.

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.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Florian Tramèr, Fan Zhang, Ari Juels, Michael K Reiter, and Thomas Ristenpart. Stealing machine learning models via prediction {APIs}. In *25th USENIX security symposium (USENIX Security 16)*, pp. 601–618, 2016.

Paul Voigt and Axel Von dem Bussche. The eu general data protection regulation (gdpr). *A practical guide, 1st ed., Cham: Springer International Publishing*, 10(3152676):10–5555, 2017.

Hongbo Zhao, Bolin Ni, Junsong Fan, Yuxi Wang, Yuntao Chen, Gaofeng Meng, and Zhaoxiang Zhang. Continual forgetting for pre-trained vision models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28631–28642, 2024a.

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, 2024b.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017.

Yu Zhou, Dian Zheng, Qijie Mo, Renjie Lu, Kun-Yu Lin, and Wei-Shi Zheng. Decoupled distillation to erase: A general unlearning method for any class-centric tasks, 2025. URL `https://arxiv.org/abs/2503.23751`.

8 APPENDIX

A IMPLEMENTATION AND EXPERIMENTAL SETUP

This appendix provides the implementation and experimental details required to reproduce DIMU. We first summarize the complete unlearning procedure and then describe the training, retraining, and evaluation configurations used across datasets and architectures.

A.1 DIMU ALGORITHM

Algorithm 1 summarizes the full procedure. DIMU first computes channel-importance scores, generates a deterministic key-conditioned mask, applies the mask to the penultimate representation, and then optimizes the classifier head using forgetting, retention, and SKD losses under the fixed CRM mask, while keeping the feature extractor frozen during unlearning. At deployment, inference uses the frozen feature extractor h_θ, the updated classifier head $g_{\theta'}$, and the fixed CRM mask $\mathbf{m}$. The method uses a lightweight adaptation strategy that avoids full-backbone retraining and does not require an external teacher model.

Algorithm 1: DIMU Unlearning

Input: Model parameters θ, forget set $\mathcal{D}_f$, retain set $\mathcal{D}_r$, key k, mask ratio τ, weights $\lambda_f, \lambda_r, \lambda_{\text{SKD}}$
Output: Updated classifier head $g_{\theta'}$ and deployed CRM mask $\mathbf{m}$
$\mathbf{I} \leftarrow \text{ComputeChannelImportance}(\mathcal{D}_f, \theta)$
$\mathbf{m} \leftarrow \text{GenerateCRM}(\mathbf{I}, k, \tau)$
$z_r^{(0)} \leftarrow f_\theta(\mathcal{D}_r)$
Freeze h_θ and optimize g_θ under fixed mask $\mathbf{m}$
for *each training iteration* **do**
  $(x_r, y_r) \sim \mathcal{D}_r$;
  $(x_f, y_f) \sim \mathcal{D}_f$
  $z_f \leftarrow f_{\theta,m}(x_f)$;
  $z_r \leftarrow f_{\theta,m}(x_r)$;
  $\mathcal{L}_{\text{forget}} \leftarrow -\ell_{\text{CE}}(z_f, y_f)$
  $\mathcal{L}_{\text{retain}} \leftarrow \ell_{\text{CE}}(z_r, y_r)$
  $\mathcal{L}_{\text{SKD}} \leftarrow \text{SKDLoss}(z_r, z_r^{(0)})$
  $\mathcal{L} \leftarrow \lambda_f \mathcal{L}_{\text{forget}} + \lambda_r \mathcal{L}_{\text{retain}} + \lambda_{\text{SKD}} \mathcal{L}_{\text{SKD}}$
  $g_\theta \leftarrow g_\theta - \eta \nabla_{g_\theta} \mathcal{L}$
return $g_{\theta'}$, $\mathbf{m}$

A.2 TRAINING AND RETRAINING CONFIGURATIONS

All experiments were conducted on a single NVIDIA A40 GPU (48 GB VRAM). For CNN backbones (ResNet-18 and VGG-16), Original-model training used SGD with momentum 0.9 and weight decay 1×10^{-4}. The initial learning rate was set to 0.01 and decayed by a factor of 0.1 every 30 epochs. For transformer backbones (ViT-S and Swin-T), Original-model training used AdamW with learning rate 5×10^{-4}, weight decay 1×10^{-4}, and cosine annealing. For each backbone–dataset pair, the Original model was trained from random initialization on the full training set for 150 epochs with batch size 128 using standard CIFAR preprocessing. DIMU unlearning was then initialized from this trained Original checkpoint. Retrain baselines were trained independently from random initialization using only the retain set $\mathcal{D}_r$ and serve as the reference for exact unlearning. For ImageNet-1K, the Original model is the torchvision IMAGENET1K_V2 pretrained ResNet-50, and DIMU starts from this checkpoint with the feature extractor frozen and only the classifier head adapted under the fixed CRM mask. Single-class unlearning is evaluated on the three fixed forget classes 10, 138, and 583.

Unlearning Configurations. Table 8 lists the hyperparameters used for DIMU across all experiments. Unless otherwise noted, these settings were kept fixed for every backbone–dataset pair to ensure comparability.

Table 8: Hyperparameter settings for DIMU unlearning.

Parameter	Value
Epochs (E)	40
Batch size (B)	128
Learning rate (η)	0.005
Optimizer	SGD
Mask ratio (τ)	0.05
Temperature (T)	4.0
Forget loss weight (λ_f)	4.0
Retain loss weight (λ_r)	2.0
SKD loss weight (λ_{SKD})	0.3
Weight decay (wd)	1×10^{-5}

The default hyperparameters were selected through preliminary tuning over $\tau \in \{0.01, 0.03, 0.05, 0.07, 0.09\}$, $T \in \{1, 2, 4, 8\}$, $\lambda_f \in \{1, 2, 4, 8\}$, $\lambda_r \in \{0.5, 1, 2, 4\}$, and $\lambda_{\text{SKD}} \in \{0.1, 0.3, 0.5, 1.0\}$. The selected configuration was $\tau = 0.05$, $T = 4$, $\lambda_f = 4$, $\lambda_r = 2$, and $\lambda_{\text{SKD}} = 0.3$, and was then fixed across all reported settings. Preliminary tuning used the official CIFAR-10/100 test splits and the ImageNet-1K validation split. No separate development set was used. All experiments use the deterministic masking formulation described in Section 3.

DIMU-L4 Configuration. DIMU-L4 follows the same importance estimation, CRM mask, and composite objective as canonical DIMU, but additionally updates ResNet layer4 jointly with the classifier head while keeping earlier backbone stages frozen. It uses $\tau = 0.05$, $\eta = 0.005$, $(\lambda_f, \lambda_r, \lambda_{\text{SKD}}) = (4.0, 2.0, 0.3)$, and 40 unlearning epochs. Channel importance is computed once before L4 adaptation, and the resulting CRM mask remains fixed throughout training.

Experiment Configuration. Unless otherwise stated, DIMU unlearning starts from the corresponding Original checkpoint trained on the full dataset for each backbone–dataset pair (ResNet-18, VGG-16, ViT-S, and Swin-T on CIFAR-10/100). We evaluate single-class, multi-class, and sample-level unlearning settings. CRM keys are fixed within each experimental configuration to ensure deterministic and reproducible mask generation, with key-variation settings reported separately where applicable. Baseline implementations were reproduced from publicly available official implementations whenever possible, with hyperparameters adapted to maintain consistent training protocols across methods. For class-level unlearning, confidence-based MIA compares forgotten-class training samples as members with forgotten-class test samples as non-members. For sample-level unlearning, the randomly selected forgotten training samples are treated as members and held-out test samples as non-members. These scores are computed from model prediction confidence without shadow models or additional attacker training. Adaptive white-box output- and representation-level attacks are evaluated separately in Table 3. MIA results are interpreted as empirical distinguishability indicators rather than formal privacy guarantees. All experiments used controlled random seeds and identical retain/forget partitions for fair comparison across methods.

B Extended Evaluation and Analysis

This appendix provides additional theoretical, architectural, representation-level, sensitivity, efficiency, and robustness analyses that complement the main results. We first present the theoretical motivation for importance-guided masking, followed by extended evaluations across architectures and diagnostic analyses of representation, privacy, stability, and computational cost.

B.1 Theoretical and Conceptual Analysis

First-Order Analysis of Feature-Level Intervention. For a sample (x, y), masking channels $\mathcal{C}_f$ changes the penultimate representation by

$$\Delta h_{\theta,c}(x) = \begin{cases} -h_{\theta,c}(x), & c \in \mathcal{C}_f, \\ 0, & \text{otherwise.} \end{cases}$$

Using a first-order Taylor approximation around $h_\theta(x)$,

$$\Delta\ell^{(1)}_{\mathrm{CE}}(x,y) = -\sum_{c\in\mathcal{C}_f} h_{\theta,c}(x)\frac{\partial \ell_{\mathrm{CE}}}{\partial h_{\theta,c}(x)}.$$

Applying the triangle inequality gives

$$\left|\Delta\ell^{(1)}_{\mathrm{CE}}(x,y)\right| \le \sum_{c\in\mathcal{C}_f} |h_{\theta,c}(x)| \left|\frac{\partial \ell_{\mathrm{CE}}}{\partial h_{\theta,c}(x)}\right|.$$

If $|h_{\theta,c}(x)| \le B$, averaging over $\mathcal{D}_f$ yields

$$\mathbb{E}_{(x,y)\sim\mathcal{D}_f}\left[\left|\Delta\ell^{(1)}_{\mathrm{CE}}(x,y)\right|\right] \le B\sum_{c\in\mathcal{C}_f} I_c,$$

where I_c is the channel-importance score in Eq. 4. This bound motivates feature-level intervention but does not uniquely determine the ranking criterion because the first-order effect also depends on activation magnitude. We therefore use gradient magnitude as the default importance estimator and compare alternative ranking criteria in Table 13. This analysis provides a local motivation for importance-guided masking and does not imply certified or exact unlearning.

Conceptual Unlearning Setup. Figure 2 provides a conceptual overview of the unlearning setting, where U_k denotes a generic key-conditioned unlearning operator.

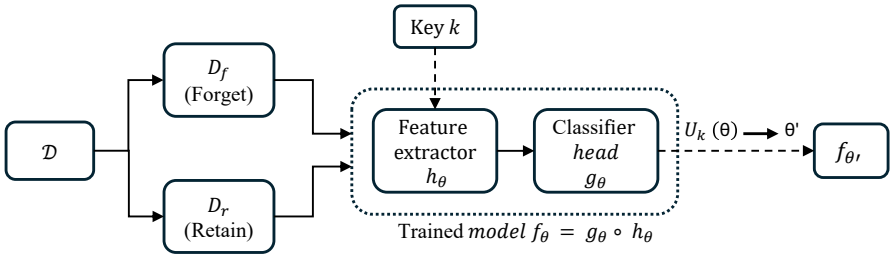


Figure 2: Conceptual overview of the unlearning problem.

B.2 Architectural Generalization

To verify that the proposed framework generalizes beyond the ResNet-18 setting used in the main experiments, we further evaluate class-level unlearning on three additional architectures: VGG-16, Swin-T, and ViT-S. These models span classical CNNs, hierarchical vision transformers, and pure self-attention architectures, enabling a broad assessment of whether unlearning behavior remains stable across fundamentally different representational priors. The related results are summarized in Tables 9, 10, and 11.

Table 9: Performance comparison of class-level unlearning on VGG-16 across different unlearning methods on CIFAR-10 and CIFAR-100 datasets.

Method	CIFAR-10						CIFAR-100					
	$\mathrm{Acc}_f\downarrow$	$\mathrm{Acc}_r\uparrow$	$\mathrm{Acc}_{\mathrm{ft}}\downarrow$	$\mathrm{Acc}_{\mathrm{ft}}\uparrow$	H-Mean↑	MIA↓	$\mathrm{Acc}_f\downarrow$	$\mathrm{Acc}_r\uparrow$	$\mathrm{Acc}_{\mathrm{ft}}\downarrow$	$\mathrm{Acc}_{\mathrm{ft}}\uparrow$	H-Mean↑	MIA↓
Original	99.60	99.51	**91.80**	91.97	–	0.98	**99.80**	**99.72**	**63.00**	**71.73**	–	0.98
Retrain	0.00	**99.86**	0.00	**92.23**	**92.03**	**0.00**	0.00	99.82	0.00	71.59	67.05	**0.00**
DELETE	**0.00**	99.28	**0.00**	91.91	91.86	**0.00**	7.00	97.72	**0.00**	76.59	69.12	**0.00**
SCRUB	**0.00**	90.01	**0.00**	85.38	88.47	**0.00**	**0.00**	71.02	**0.00**	56.26	59.44	**0.00**
SalUn	2.06	97.06	3.00	89.14	88.97	0.02	4.00	95.07	5.00	66.67	61.99	0.06
Random Label	1.84	97.45	1.90	89.42	89.66	0.01	5.60	94.58	7.00	66.29	60.73	0.04
Bad Teacher	4.64	97.42	1.60	89.56	89.83	0.04	7.40	95.24	6.00	66.74	61.52	0.06
Negative Gradient	0.20	95.57	0.10	87.74	89.65	**0.00**	**0.00**	15.64	**0.00**	11.06	18.82	**0.00**
DIMU	**0.00**	87.22	**0.00**	87.26	89.55	**0.00**	**0.00**	81.00	**0.00**	**80.72**	**70.77**	**0.00**

VGG-16 (Table 9). For VGG-16, several baselines show noticeable retention degradation, particularly on CIFAR-100. DELETE and SCRUB achieve low forgotten-class test accuracy on CIFAR-10, while Negative Gradient often destabilizes retained utility. In contrast, DIMU achieves zero

Table 10: Performance comparison of class-level unlearning on Swin-T across different unlearning methods on CIFAR-10 and CIFAR-100 datasets.

Method	CIFAR-10						CIFAR-100					
	$\text{Acc}_\text{f}\downarrow$	$\text{Acc}_\text{r}\uparrow$	$\text{Acc}_\text{ft}\downarrow$	$\text{Acc}_\text{rt}\uparrow$	H-Mean↑	MIA↓	$\text{Acc}_\text{f}\downarrow$	$\text{Acc}_\text{r}\uparrow$	$\text{Acc}_\text{ft}\downarrow$	$\text{Acc}_\text{rt}\uparrow$	H-Mean↑	MIA↓
Original	96.90	97.95	**85.70**	90.87	–	0.93	97.60	97.76	**62.00**	72.83	–	0.79
Retrain	**0.00**	**97.95**	**0.00**	**91.37**	**88.44**	**0.00**	**0.00**	97.73	**0.00**	**73.01**	67.06	**0.00**
DELETE	**0.00**	86.90	**0.00**	83.18	84.42	**0.00**	**0.00**	78.72	**0.00**	62.43	62.21	**0.00**
SCRUB	**0.00**	76.35	**0.00**	74.90	79.94	**0.00**	**0.00**	61.50	**0.00**	54.84	58.17	**0.00**
SalUn	8.10	62.21	8.60	58.10	66.26	0.08	1.80	68.43	2.00	54.45	57.09	0.04
Random Label	7.24	60.42	3.80	56.36	66.77	0.07	1.60	68.20	1.00	54.60	57.62	0.01
Bad Teacher	5.08	55.94	3.90	52.92	64.26	0.05	1.80	64.53	**0.00**	51.79	56.44	0.03
Negative Gradient	**0.00**	62.20	**0.00**	59.21	70.03	0.21	**0.00**	**73.06**	**0.00**	58.44	60.17	**0.00**
DIMU	**0.00**	90.41	**0.00**	85.13	85.41	0.20	**0.00**	**91.57**	**0.00**	67.16	**64.48**	**0.00**

Table 11: Performance comparison of class-level unlearning on ViT-S across different unlearning methods on CIFAR-10 and CIFAR-100 datasets.

Method	CIFAR-10						CIFAR-100					
	$\text{Acc}_f\downarrow$	$\text{Acc}_r\uparrow$	$\text{Acc}_{ft}\downarrow$	$\text{Acc}_{rt}\uparrow$	H-Mean↑	MIA↓	$\text{Acc}_f\downarrow$	$\text{Acc}_r\uparrow$	$\text{Acc}_{ft}\downarrow$	$\text{Acc}_{rt}\uparrow$	H-Mean↑	MIA↓
Original	92.50	92.15	72.30	76.41	–	0.92	86.40	92.98	50.00	63.16	–	0.67
Retrain	**0.00**	91.22	**0.00**	76.42	74.30	**0.00**	**0.00**	93.42	**0.00**	62.93	55.72	**0.00**
DELETE	**0.00**	91.99	**0.00**	77.74	74.94	**0.00**	1.10	72.17	2.00	45.13	46.52	0.18
SCRUB	**0.00**	80.10	**0.00**	65.89	68.95	**0.00**	**0.00**	64.20	**0.00**	40.35	44.71	**0.00**
SalUn	11.14	63.87	8.30	56.39	59.95	0.10	10.60	67.99	**0.00**	42.52	45.96	0.12
Random Label	7.96	65.37	5.60	56.63	61.25	0.07	8.60	67.36	1.00	41.90	45.17	0.10
Bad Teacher	6.62	64.20	4.00	56.93	62.02	0.05	11.40	68.28	**0.00**	42.41	45.89	0.14
Negative Gradient	0.08	78.09	**0.00**	67.54	69.84	0.01	**0.00**	64.67	**0.00**	41.98	45.64	**0.00**
DIMU	**0.00**	**93.24**	**0.00**	**79.82**	**75.87**	**0.00**	**0.00**	**91.50**	**0.00**	**64.96**	**56.51**	**0.00**

forgotten-class test accuracy with strong retained performance and competitive H-Mean values across both datasets. On CIFAR-100, DIMU also achieves higher retained-test accuracy than both Original and Retrain while maintaining zero forgotten-class test accuracy. This gain may arise from retained-class recalibration of the classifier head after masking, which can alter the retained-class decision boundaries even though the underlying feature extractor remains frozen.

Swin-T (Table 10). For Swin-T, baselines such as Random Label and SalUn exhibit substantial retained-utility degradation. DELETE and Negative Gradient achieve low forget accuracy but do not preserve retained performance consistently. DIMU, however, achieves effective forgetting with high retention (e.g., Acc_rt = 85.13% on CIFAR-10), competitive H-Mean values and MIA behavior. These results show that DIMU extends to Swin-T using the same fixed-mask classifier-adaptation formulation, with CRM operating on selected dimensions of the final pooled representation while the backbone remains frozen and SKD regularizes the classifier-head update.

ViT-S (Table 11). Among the evaluated architectures, ViT-S presents a more challenging balance between effective forgetting and retained utility for several baselines. In these cases, performance often reflects either incomplete forgetting or substantial retention degradation. DIMU delivers a strong and balanced trade-off, achieving zero forgotten-class test accuracy, high retained accuracy across both datasets, strong H-Mean performance, and competitive MIA scores. These results show that DIMU remains effective on ViT-S under the fixed-mask classifier-adaptation setting.

B.3 Representation and Sensitivity Analysis

Representation Visualization. Figure 3 visualizes the representations supplied to the classifier on CIFAR-10 using a ResNet-18 backbone. For the Original and Retrain models, we visualize their raw penultimate representations, $h_\theta(x)$ and $h_{\theta_r^\star}(x)$, respectively. For DIMU, we visualize the CRM-masked effective representation, $\tilde{h}_\theta(x) = m \odot h_\theta(x)$, that is passed to the updated classifier. The Original model exhibits well-separated semantic clusters, whereas the Retrain model forms a representation learned independently from the retain-only training set. In the DIMU projection, retained classes remain well clustered, while forgotten-class samples become less distinctive in the effective representation. Because the feature extractor remains frozen during DIMU unlearning, the changes observed for our method reflect the effect of CRM on the representation exposed to the classifier rather than modification of the underlying raw backbone representation. Since t-SNE visualizations can vary with projection hyperparameters and sampling, we treat Figure 3 as qualitative analysis

Table 12: Adaptive privacy and representation analysis on CIFAR-100 using ResNet-18 with class 0 as the forget class, following the same protocol as Table 3. Attack scores are mean ± standard deviation over three seeded attacker splits of a fixed checkpoint, while accuracies use the same checkpoint.

Method	Trainable Scope	$\mathrm{Acc}_{ft} \downarrow$	$\mathrm{Acc}_{rt} \uparrow$	Output MIA AUROC*	Output MIA BA*	Raw-feature MIA AUROC*	Masked-feature MIA AUROC*
Original	–	72.00	79.41	0.960 ± 0.006	0.910 ± 0.010	0.960 ± 0.006	–
Retrain	All, from scratch	0.00	78.53	0.520 ± 0.008	0.520 ± 0.008	0.650 ± 0.018	–
DIMU	Head only	**0.00**	**78.38**	0.610 ± 0.018	0.580 ± 0.015	0.960 ± 0.008	0.660 ± 0.020
DIMU-L4	Layer4 + head	**0.00**	77.85	0.540 ± 0.010	0.530 ± 0.010	0.620 ± 0.020	0.550 ± 0.015

only. All t-SNE visualizations use a fixed random seed, a perplexity of 30, and two-dimensional projections.



Figure 3: t-SNE visualization of classifier-input representations on CIFAR-10 using ResNet-18. For the Original and Retrain models, panels (a) and (b) visualize the raw penultimate representations $h_\theta(x)$ and $h_{\theta_r^\star}(x)$, respectively, whereas panel (c) visualizes the CRM-masked effective representation $\tilde{h}_\theta(x) = m \odot h_\theta(x)$ of DIMU. DIMU reduces the separability of forgotten-class samples in the effective representation while largely preserving retained-class structure.

Adaptive Privacy and Representation Analysis on CIFAR-100. We extend the adaptive output- and representation-level privacy analysis to CIFAR-100 using ResNet-18 with class 0 as the forget class. Table 12 reports output-level and feature-level distinguishability for Original, Retrain, DIMU, and DIMU-L4 under the same evaluation protocol used for CIFAR-10.

Canonical DIMU reduces output-level AUROC* from 0.960 for the Original model to 0.610 while maintaining zero forgotten-class test accuracy and 78.38% retained-class accuracy. As expected from its frozen backbone, raw-feature distinguishability remains high at 0.960, whereas the deployed CRM-masked representation reduces this score to 0.660. DIMU-L4 further reduces output, raw-feature, and masked-feature AUROC* to 0.540, 0.620, and 0.550, respectively, while retaining zero forgotten-class test accuracy and 77.85% retained-class accuracy. This reproduces the qualitative behavior observed on CIFAR-10 and shows that limited representation adaptation substantially reduces residual feature-level distinguishability.

Channel-Ranking Ablation. We compare the default gradient-based ranking with random, activation-based, and activation×gradient channel selection on CIFAR-100 using ResNet-18.

Gradient-based ranking provides the strongest forgetting–retention trade-off, achieving zero forgotten-class test accuracy, the highest retained accuracy (78.38%), and the highest H-Mean (75.05) among the evaluated criteria. Activation × gradient performs similarly but remains slightly weaker, while random and activation-only ranking leave greater residual forgotten-class accuracy.

SKD Temperature Sensitivity. To assess sensitivity to the distillation temperature, we vary $T \in \{1, 2, 4, 8\}$ while keeping λ_{SKD} and all other hyperparameters fixed. Table 14 reports the results over three seeds. Performance varies only modestly across the evaluated temperatures, with the default $T = 4$ providing the strongest overall forgetting–retention trade-off.

Table 13: Channel-ranking ablation of DIMU on CIFAR-100 using ResNet-18 with class 0 as the forget class and $\tau = 0.05$. All variants use the same unlearning configuration and differ only in the channel-ranking criterion.

Ranking Criterion	$\mathbf{Acc}_{ft} \downarrow$	$\mathbf{Acc}_{rt} \uparrow$	**H-Mean**$\uparrow$	**MIA**$\downarrow$
Random	8.00	74.50	68.85	0.08
Activation	2.00	76.80	73.24	0.03
Gradient magnitude (default)	0.00	78.38	75.05	0.00
Activation $\times$ Gradient	0.50	77.90	74.56	0.01

Table 14: SKD-temperature sensitivity of DIMU on CIFAR-100 using ResNet-18. Results are mean $\pm$ standard deviation over three seeds.

Temperature T	$\mathbf{Acc}_{ft} \downarrow$	$\mathbf{Acc}_{rt} \uparrow$	**H-Mean**$\uparrow$	**MIA**$\downarrow$
1	0.45 ± 0.21	76.80 ± 0.07	73.89 ± 0.32	0.06 ± 0.04
2	0.26 ± 0.01	77.63 ± 0.06	74.46 ± 0.02	0.03 ± 0.03
4	$\mathbf{0.00 \pm 0.00}$	$\mathbf{78.38 \pm 0.07}$	$\mathbf{75.05 \pm 0.08}$	$\mathbf{0.00 \pm 0.00}$
8	0.25 ± 0.20	77.40 ± 0.06	74.33 ± 0.26	0.07 ± 0.04

Unlearning-Epoch Sensitivity. Table 15 shows that test-level forgetting largely stabilizes by 30 epochs, with $\mathrm{Acc}_{ft} = 0.00$ and H-Mean 74.97. Extending to 40 epochs eliminates the remaining forget-training accuracy while retaining comparable utility, giving the final H-Mean of 75.05. We therefore use 40 epochs as the fixed unlearning budget.

Table 15: Unlearning-epoch sensitivity of DIMU on CIFAR-100 using ResNet-18. Test-level forgetting and the overall forgetting–retention trade-off largely stabilize by 30 epochs, while the fixed 40-epoch budget eliminates residual forget-training accuracy with negligible change in retained utility.

Epochs	$\mathbf{Acc}_{f} \downarrow$	$\mathbf{Acc}_{r} \uparrow$	$\mathbf{Acc}_{ft} \downarrow$	$\mathbf{Acc}_{rt} \uparrow$	**H-Mean**$\uparrow$	**MIA**$\downarrow$
10	6.40	99.60	5.00	75.55	71.02	0.08
20	3.20	99.52	1.00	78.17	74.41	0.02
30	2.10	99.47	0.00	78.20	74.97	0.00
40	0.00	99.44	0.00	78.38	75.05	0.00

Optimization Diagnostics. Figure 4 shows that the signed DIMU objective decreases rapidly and then stabilizes over the fixed 40-epoch schedule. The classifier-head norm and absolute logit magnitudes also plateau after the initial optimization phase, providing empirical evidence of controlled optimization behavior rather than a formal boundedness or stability guarantee.

Forget-Set Difficulty and Trade-off Analysis. We further analyze forget-set difficulty using the RUM framework Zhao et al. (2024b) by measuring the Maximum Mean Discrepancy (MMD) between forget and retain feature distributions across multiple 3000-sample forget subsets. On CIFAR-10, the MMD decreases from 0.0921 to 0.0444 across the evaluated groups, while CIFAR-100 exhibits consistently lower MMD values ($\sim$0.0189), indicating stronger forget–retain overlap and more challenging unlearning scenarios. We also report normalized trade-off scores inspired by the Tug-of-War (ToW) framework Zhao et al. (2024b), combining forgetting effectiveness, retained utility, and privacy preservation. Using Table 1, DIMU achieves ToW-style and ToW-MIA-style scores of approximately $0.960/0.957$ on CIFAR-10 and $0.779/0.662$ on CIFAR-100, indicating a strong balance between effective forgetting and retained performance.

B.4 Efficiency and Robustness Diagnostics

Runtime Comparison. Table 16 compares the computational cost of DIMU with full Retrain and representative unlearning baselines on an NVIDIA A40 GPU. Reported runtimes correspond to the complete method-specific training or unlearning schedules rather than an equalized number of epochs. Absolute runtimes across dataset–architecture settings are not directly comparable because

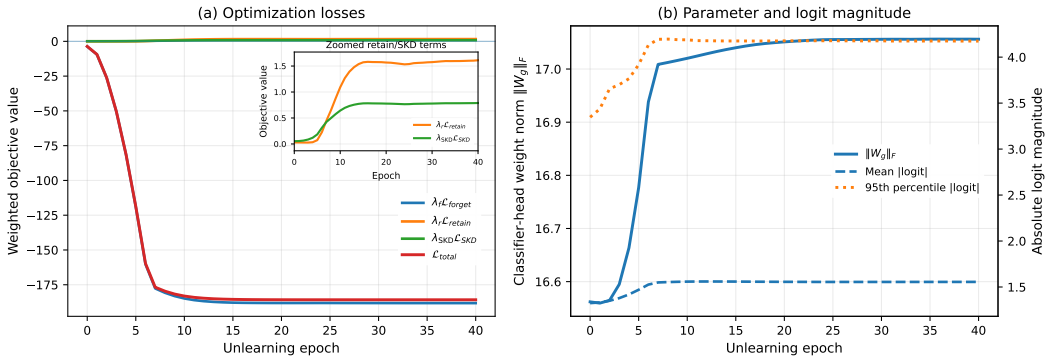


Figure 4: Optimization diagnostics for DIMU on CIFAR-100 with ResNet-18 over the default 40-epoch unlearning schedule. Panel (a) shows the signed weighted objective terms and total objective, with an inset for the retain and SKD terms. Panel (b) shows the classifier-head weight norm and mean and 95th-percentile absolute logit magnitudes.

the prescribed unlearning schedules and computational characteristics differ. Speedups should therefore be interpreted within each setting. DIMU uses 40 unlearning epochs, while Retrain uses 150 epochs for CIFAR-10/100, following the corresponding CIFAR training protocol, and 90 epochs for the ResNet-50 ImageNet-1K, Tiny ImageNet, and Places365 experiments. The 90-epoch Retrain schedule follows a conventional ResNet-50 ImageNet training schedule and is kept fixed across the ResNet-50 retraining experiments for consistency. The 90-epoch Retrain schedule is independent of how the pretrained source checkpoints were originally trained. Retrain performs full retain-only training from scratch, whereas DIMU starts from the corresponding pretrained source model and performs its prescribed unlearning procedure. The reported speedups therefore measure the computational cost of the complete respective procedures rather than per-epoch efficiency.

On ResNet-18/CIFAR-10 and CIFAR-100, DIMU requires 52.4 and 57.3 minutes, compared with 80.5 and 88.2 minutes for Retrain, corresponding to a $1.54\times$ speedup in both settings. On ViT-S/CIFAR-10 and CIFAR-100, DIMU requires 218.8 and 239.4 minutes, compared with 336.3 and 368.7 minutes for Retrain, again yielding approximately $1.54\times$ speedup. DELETE achieves approximately $5.2\times$ speedup across the CIFAR settings, while SalUn, Random Label, Bad Teacher, and Negative Gradient remain substantially faster lightweight alternatives.

On ResNet-50/ImageNet-1K, DIMU completes unlearning in 1.5563 hours compared with 22.9846 hours for Retrain-FT and 38.7310 hours for Retrain, corresponding to a $24.89\times$ speedup over full Retrain. To further examine runtime scaling with ResNet-50, we additionally evaluate Tiny ImageNet Le et al. (2015) and Places365 Zhou et al. (2017). These additional datasets are used only for computational runtime evaluation rather than for the full unlearning benchmark analysis reported in the main paper. Tiny ImageNet uses the same torchvision IMAGENET1K_V2 pretrained initialization as the ImageNet-1K experiment, while Places365 uses the official Places365 pretrained ResNet-50 checkpoint. On Tiny ImageNet, DIMU completes unlearning in 0.1634 hours compared with 3.3305 hours for Retrain, corresponding to a $20.38\times$ speedup. On Places365, which contains approximately 1.80 million training images, DIMU requires 0.9417 hours compared with 61.0068 hours for Retrain, corresponding to a $64.78\times$ speedup. These additional runtime evaluations show that DIMU retains a substantial computational advantage over full retraining across ResNet-50 datasets with markedly different dataset sizes and numbers of classes.

Head-Only Adaptation and Relearning Analysis. Table 17 shows that simply freezing the feature extractor and fine-tuning the classifier head on retained data is insufficient to reproduce DIMU's forgetting behavior. Head-only Retain-FT retains 12.75% forgotten-class accuracy and yields an H-Mean of 67.16, whereas DIMU yields 0.00% forgotten-class accuracy and an H-Mean of 75.05. Table 18 further examines rapid relearning after one epoch of standard fine-tuning on forgotten-class samples. Forgotten-class accuracy recovers by 41.37 points for Head-only Retain-FT, compared with 22.40 points for DIMU and 9.35 points for Retrain. Thus, DIMU substantially reduces rapid recovery relative to head-only adaptation, although Retrain remains the strongest reference against relearning.

Table 16: Runtime comparison on an NVIDIA A40 GPU. CIFAR-10 and CIFAR-100 results are reported for ResNet-18 and ViT-S, while ImageNet-1K results use ResNet-50. Runtime denotes the complete method-specific retraining, retain-only fine-tuning (Retrain-FT), or unlearning procedure rather than an equalized number of epochs. Speedup is computed relative to Retrain within each setting.

Method	**ResNet-18**		**ViT-S**		**Speedup**
	CIFAR-10	**CIFAR-100**	**CIFAR-10**	**CIFAR-100**	**Range**
Retrain	80.5 min	88.2 min	336.3 min	368.7 min	$1.00\times$
DELETE	15.5 min	16.9 min	64.7 min	70.6 min	5.19–$5.22\times$
SalUn	0.36 min	0.39 min	1.52 min	1.62 min	221.25–$227.59\times$
Random Label	0.37 min	0.40 min	1.56 min	1.65 min	215.58–$223.45\times$
Bad Teacher	0.39 min	0.43 min	1.65 min	1.79 min	203.82–$206.41\times$
Negative Gradient	0.42 min	0.46 min	1.76 min	1.91 min	191.08–$193.04\times$
DIMU	52.4 min	57.3 min	218.8 min	239.4 min	1.536–$1.540\times$
ResNet-50 / ImageNet-1K					
Retrain	38.7310 h				$1.00\times$
Retrain-FT	22.9846 h				$1.69\times$
DIMU	1.5563 h				$24.89\times$

Table 17: Comparison with head-only retain fine-tuning on CIFAR-100 using ResNet-18.

Method	$\mathbf{Acc}_{ft}\downarrow$	$\mathbf{Acc}_{rt}\uparrow$	**H-Mean**$\uparrow$	**MIA**$\downarrow$
Retrain	0.00	78.53	75.12	0.00
Head-only Retain-FT	12.75	77.50	67.16	0.43
DIMU	0.00	78.38	75.05	0.00

Table 18: One-epoch forgotten-class relearning on CIFAR-100 using ResNet-18.

Method	**Before** $\mathbf{Acc}_{ft}\downarrow$	**After** $\mathbf{Acc}_{ft}\downarrow$	**Recovery** $\Delta\mathbf{Acc}_{ft}\downarrow$	**After** $\mathbf{Acc}_{rt}\uparrow$
Retrain	0.00	9.35	9.35	78.03
Head-only Retain-FT	12.75	54.12	41.37	76.97
DIMU	0.00	22.40	22.40	77.90

Deterministic Key Analysis. Table 19 evaluates DIMU using multiple keys under the same ResNet-18/CIFAR-100 forgetting configuration. The results show highly consistent forgetting and retained performance across keys, supporting the role of the key as a mechanism for deterministic and reproducible mask generation rather than as a primary driver of unlearning.

Table 19: Deterministic key analysis on ResNet-18 using CIFAR-100.

Key	$\mathrm{Acc}_{\mathrm{ft}}\downarrow$	$\mathrm{Acc}_{\mathrm{rt}}\uparrow$
Key-1	0.00	78.38
Key-2	0.00	78.35
Key-3	0.00	78.34

Forgotten-Class Identity Sensitivity. To examine whether DIMU's single-class performance depends on the selected target class, we evaluate five fixed CIFAR-100 class identities using ResNet-18. As shown in Table 20, DIMU maintains near-zero forgotten-class accuracy across the five targets, with an average of $0.22 \pm 0.44\%$, while retained accuracy remains highly stable at $78.09 \pm 0.19\%$. The low variation in retained utility and consistently low MIA scores indicate that the observed unlearning behavior is not specific to a particular forgotten class.

Stability Across Runs. Table 21 reports mean and standard deviation across three runs with different random seeds. For ResNet-18, DIMU obtains H-Mean values of 96.27 ± 0.04 on CIFAR-10 and 75.05 ± 0.08 on CIFAR-100, with $\mathrm{Acc}_{ft} = 0.00 \pm 0.00$ on both datasets. The low variance indicates stable behavior across runs.

Prediction Distribution Analysis. Figure 5 shows the prediction distributions for forgotten automobile-class samples after unlearning. Across Retrain, DELETE, SCRUB, and DIMU, forgotten samples are reassigned to retained classes rather than being concentrated on the original automobile label. The dominant reassigned classes differ across methods, indicating structured output-level redistribution instead of trivial collapse.

Table 20: Sensitivity of DIMU to forgotten-class identity on CIFAR-100 using ResNet-18. Results are reported for five fixed single-class deletion targets, with mean ± standard deviation computed across class identities.

Forget Class ID	Original Acc_{ft}	DIMU $\mathrm{Acc}_{ft}\downarrow$	DIMU $\mathrm{Acc}_{rt}\uparrow$	H-Mean↑	MIA↓
0	72.00	0.00	78.38	75.05	0.00
25	71.97	0.12	77.92	74.76	0.00
50	71.90	1.00	78.13	74.34	0.00
75	71.06	0.00	78.06	74.40	0.01
99	70.50	0.00	77.94	74.03	0.01
Mean ± std.	71.49 ± 0.68	0.22 ± 0.44	78.09 ± 0.19	74.52 ± 0.40	0.004 ± 0.005

Table 21: Mean ± standard deviation across three runs with different random seeds for DIMU under single-class unlearning settings.

Backbone	CIFAR-10				CIFAR-100			
	$\mathrm{Acc}_{ft}\downarrow$	$\mathrm{Acc}_{rt}\uparrow$	H-Mean↑	MIA↓	$\mathrm{Acc}_{ft}\downarrow$	$\mathrm{Acc}_{rt}\uparrow$	H-Mean↑	MIA↓
ResNet-18	0.00±0.00	95.99±0.05	96.27±0.04	0.003±0.001	0.00±0.00	78.38±0.07	75.05±0.08	0.00±0.00
VGG-16	0.00±0.00	87.26±0.10	89.55±0.10	0.00±0.00	0.00±0.00	80.72±0.14	70.77±0.11	0.00±0.00
Swin-T	0.00±0.00	85.13±0.16	85.41±0.13	0.20±0.01	0.00±0.00	67.16±0.21	64.48±0.18	0.00±0.00
ViT-S	0.00±0.00	79.82±0.20	75.87±0.16	0.00±0.00	0.00±0.00	64.96±0.24	56.51±0.21	0.01±0.01

For DIMU, the absence of concentration on the forgotten class rules out collapse to a single retained class. Closeness to Retrain is evaluated separately using retrain-relative metrics in Table 23.

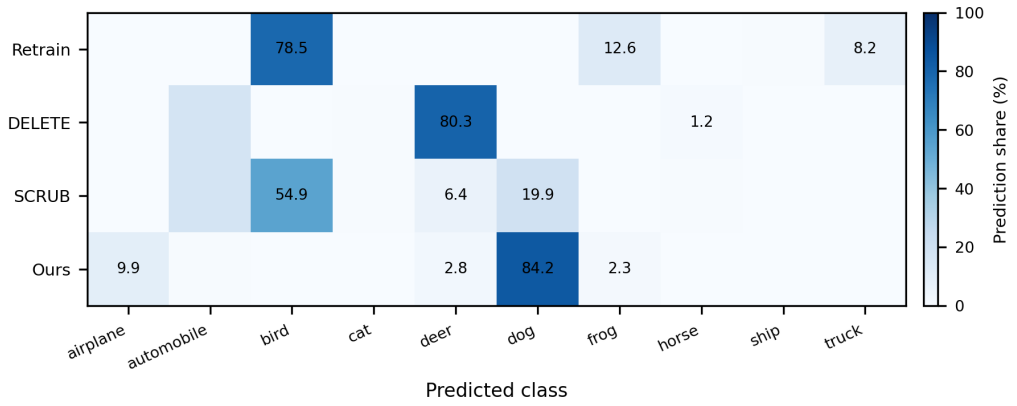


Figure 5: Prediction distributions of forgotten automobile-class samples on CIFAR-10 after unlearning.

Sequential Forget Requests. We also evaluate a realistic setting where deletion requests arrive sequentially. Using the same ResNet-18 checkpoint as in the sample-level experiments, we issue five non-overlapping random sample-level forget requests, each removing 1% of the training set. Acc_f and Acc_r are evaluated on the accumulated forgotten and retained training subsets, respectively, while Acc_{test} is evaluated on the held-out test set. Table 22 compares reusing the same key, using new keys, composing masks, and recomputing importance after each request. All variants preserve high retained and test accuracy while reducing accumulated forgotten-sample accuracy and driving MIA to zero, showing stable behavior under repeated deletion requests.

Comparison with Retrain. To further characterize random sample-level unlearning, we compare DIMU with retain-only retraining at both predictive and representation levels. As shown in Table 23, symmetric KL on $\mathcal{D}_f$ increases from 1.818 to 2.177 as the deletion ratio grows, while divergence on $\mathcal{D}_r$ remains substantially smaller (0.189–0.366). Representation alignment with Retrain is consis-

Table 22: Sequential sample-level unlearning on CIFAR-10 using ResNet-18 with five non-overlapping 1% forget requests. Final metrics are reported after the fifth request; runtime is averaged per request.

Strategy	$\text{Acc}_r \uparrow$	$\text{Acc}_f \downarrow$	$\text{Acc}_{test} \uparrow$	Conf. $\text{Drop}_f \uparrow$	MIA↓	Time (min)↓
Same key	97.24	69.32	89.53	31.64	0.0000	8.6
New keys	97.22	72.36	89.57	29.09	0.0000	8.7
Composed masks	97.46	71.40	89.51	31.03	0.0000	8.8
Recompute importance	97.30	70.48	89.32	31.30	0.0000	8.8

tently higher on retained samples (CKA 0.807–0.854) than on forgotten samples (0.740–0.762). Despite these differences, row-normalized test-set confusion matrices remain nearly identical (0.998–0.999), indicating preservation of the overall test-set decision structure. DIMU also exhibits lower test-set ECE than Retrain across all deletion ratios.

Table 23: Comparison of DIMU with Retrain under random sample-level unlearning on CIFAR-10 using ResNet-18. Results are mean ± standard deviation over three seeds. Sym. KL (symmetric Kullback–Leibler divergence) measures predictive discrepancy on $\mathcal{D}_f$ and $\mathcal{D}_r$. CKA (centered kernel alignment) compares classifier-input representations, namely the CRM-masked effective representation of DIMU and the raw penultimate representation of Retrain. ECE denotes test-set expected calibration error (%), and CM Sim. denotes cosine similarity between row-normalized test-set confusion matrices.

Forget Ratio	$\textbf{Sym. KL}_\mathbf{f} \downarrow$	$\textbf{Sym. KL}_\mathbf{r} \downarrow$	$\textbf{CKA}_\mathbf{f} \uparrow$	$\textbf{CKA}_\mathbf{r} \uparrow$	**ECE DIMU↓**	**ECE Retrain↓**	**CM Sim.↑**
1%	1.818 ± 0.081	0.189 ± 0.024	0.762 ± 0.014	0.854 ± 0.006	$\mathbf{3.20 \pm 0.25}$	5.30 ± 0.10	0.999 ± 0.001
5%	2.045 ± 0.119	0.323 ± 0.019	0.748 ± 0.007	0.819 ± 0.007	$\mathbf{1.89 \pm 0.58}$	5.40 ± 0.16	0.998 ± 0.001
10%	2.177 ± 0.129	0.366 ± 0.042	0.740 ± 0.007	0.807 ± 0.008	$\mathbf{1.54 \pm 0.50}$	5.57 ± 0.14	0.998 ± 0.001

Class-Level Comparison with Retrain. For CIFAR-100 class-0 unlearning using ResNet-18, DIMU is substantially closer to Retrain on retained than forgotten data. Symmetric KL divergence is 0.27 ± 0.03 on D_r versus 2.55 ± 0.12 on D_f, while CKA is 0.805 ± 0.010 on retained samples versus 0.695 ± 0.015 on forgotten samples. DIMU also achieves a test-set ECE of $3.60 \pm 0.30\%$, compared with $5.75 \pm 0.18\%$ for Retrain, while preserving a high row-normalized test-set confusion-matrix similarity of 0.997 ± 0.001. Results are mean ± standard deviation over three seeds.

Table 24: Class-level Retrain-relative diagnostics across CIFAR-10, CIFAR-100, and ImageNet-1K. Symmetric KL measures predictive divergence, agreement reports matching predicted labels, and CKA measures representation similarity. Lower symmetric KL and higher agreement and CKA indicate greater similarity to Retrain. The table reports point estimates, with CIFAR-100 uncertainty reported in the preceding paragraph.

Setting	Sym. $\text{KL}_f \downarrow$	Sym. $\text{KL}_r \downarrow$	$\text{Agreement}_f \uparrow$	$\text{Agreement}_r \uparrow$	$\text{CKA}_f \uparrow$	$\text{CKA}_r \uparrow$
CIFAR-10 class 1	1.95	0.21	52%	93%	0.760	0.860
CIFAR-100 class 0	2.55	0.27	48%	89%	0.695	0.805
ImageNet-1K class 10	2.05	0.24	51%	90%	0.740	0.840

Across all three settings, DIMU shows lower symmetric KL and higher prediction agreement and CKA on retained than forgotten samples, indicating consistently closer Retrain-relative behavior on retained data.