MATRYOSHKA MACHINE UNLEARNING IN CLASSIFICATION AND GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Machine Unlearning (MU) removes the effect of data on a trained model without retraining from scratch. Earlier works operate on the full embedding width for unlearning. Therefore, nothing in them prevents deleted content from remaining decodable at a narrower feature scale. To address this, we propose *Matryoshka Machine Unlearning* (MMU), a nested prefix embedding that freezes the source model as a privileged multi-scale teacher. The student tries to mimic the teacher without having access to teacher-privileged information. The main idea is to arm the model with richer representations via coarse-to-fine nested embeddings. MMU is a plug-and-play method applicable on top of any MU method. Across three datasets, two backbones, three types of deletion, and fourteen unlearning baselines, MMU is the only method without a failure mode. Under class deletion, MMU drives forgetting accuracy to exactly 0% across all six dataset–backbone configurations while maintaining the highest retained accuracy among all baselines. Its real-world practicality has been demonstrated on the Physical Artificial Intelligence (PAI) dataset we collected. Additionally, for the image-generation task of removing dangerous content, it matches the best baseline, as shown in Fig 1.



Figure 1: Suppressing nudity in text-to-image generation. Our proposed MMU is compared against the original Stable Diffusion (SD) (Rombach et al., 2022), SalUn (Fan et al., 2024), and a random-mask baseline. Each column shares the same prompt across methods. Red borders indicate detected nudity; black boxes are for presentation only. The rest of this set of images is provided in Fig. 6.

1 INTRODUCTION

Removing the effect of data on a trained model is rarely feasible, as it requires retraining the model from scratch on the remaining data. Machine Unlearning (MU) aims to address this challenge more effectively. MU removes the influence of a forget set, while the model continues to work on the retained set after the modification (Xu et al., 2023). Requests can arrive in different scopes, such as instance deletion, to remove individual examples while keeping the rest of their category; subclass

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.

deletion, to drop a semantic group within a class; and class deletion to drop the entire category outright. These three request types overlap in the model because a feature that identifies a deleted sample also describes its subclass, its class, and the retained samples that resemble it. Changing the prediction on the forget set is the easy half of the problem. The hard half is deciding where in the data representation the change should be made (Vahedifar et al., 2026). Modifying the final output will not guarantee successful removal as long as there are traceable features of forgotten data that can be identified during an attack on the model. Earlier methods (Foster et al., 2024; Golatkar et al., 2020b; Kurmanji et al., 2023; Fan et al., 2024) do not consider the capacity at which the forgetting and preservation signals operate. Hence, a model can flip its prediction while the deleted content stays accessible at the embedding level. We ask the question: *Can we organize a model's existing knowledge so that one deletion request reaches every scope in which the deleted content lives?*

Our answer to this question is *Matryoshka Machine Unlearning* (MMU), which reads the student through nested prefixes of a single representation, each contained within the next, so that a single request governs multiple capacities at once. A frozen copy of the source model provides privileged information about the pre-deletion behavior, and it serves a dual role at every prefix: the student is pulled toward it on retained examples and pushed away from it on forgotten ones. The privileged model is analogous to a model that has read a book, and we are pushing other models to mimic it without having read that book. It is worth mentioning that, in MMU, the teacher is different from the classical teacher (Hinton et al., 2015) where a bigger model teaches a student. Here, the teacher and student are the same model, and knowledge distillation is occurring internally via privileged information, which we refer to as *self-distillation*. The same privileged signal indicates what to keep and what to remove. This paper makes three contributions:

1. We formulate unlearning across nested representation capacities with coarse-to-fine representations. In addition we introduce self-distillation for MU, in which a teacher network forces the student to imitate it without access to teacher-privileged information.
2. We run class-, subclass-, and instance-level deletion on the Robotic Tactile Internet (RTI) dataset (Conducted for PAI simulation), CIFAR-10, and CIFAR-100, using two backbones against fourteen unlearning baselines and three reference models. Beyond forgetting and retaining accuracy, we report the per-example Unlearning-Likelihood Ratio Attack (U-LiRA).

2 Related Work

Exact unlearning removes a training example by retraining the model from scratch. However, this is an expensive practice and not practical in many domains. Therefore, some early MU works, such as SISA, shard the training set, train one model per shard, and retrain only the shard that held the deleted point (Bourtoule et al., 2021). The saving is real, but it occurs before any deletion request arrives, rendering it useless in real-world scenarios. A second line of work aims to achieve post hoc unlearning by scrubbing and perturbing the weights of a trained network to minimize the remaining information about the forget set (Golatkar et al., 2020a;b; 2021). Each step in these methods adds cost and constraints such as assumption about how the source model was optimized, a second model to fit, and an approximation whose error is hard to audit.

To address the expensive measurement costs of earlier works Amnesiac (Graves et al., 2021), proposed learn log-based per-batch updates during training and subtract them later, which again requires foresight at training time. Bad Teacher (Chundawat et al., 2023) introduces competent and incompetent teachers to selectively transfer knowledge to the student, yielding a model that contains no information about the forgotten data. Furthermore, SCRUB (Kurmanji et al., 2023) also uses a teacher-student framework to push the student toward the teacher on retained data and away from it on forgotten data, and SalUn (Fan et al., 2024) restricts updates to weights with high saliency in the forget-set gradient. These are the closest relatives of MMU, and they share one blind spot: every signal, whether a distillation term or a masked gradient, enters the model through a single full-width readout. Therefore, deletion is enforced in one place and verified in another. Consequently, nothing in the objective stops the network from satisfying the request at the final classifier while leaving the deleted content decodable from a short prefix of the same embedding. MMU uses different nested embedding widths for forgetting, so the deletion occurs at different granularities in the network. Our intuition is that nested removal has no identifiable trace back in the network.

3 MATRYOSHKA MACHINE UNLEARNING

Problem Setup. Let $\mathcal{D} = \mathcal{D}_r \cup \mathcal{D}_f$ be the original training set, with $\mathcal{D}_r \cap \mathcal{D}_f = \varnothing$. A source model consists of an encoder $f_{\theta^o} : \mathcal{X} \to \mathbb{R}^d$ and classifier $(W^o, b^o) \in \mathbb{R}^{L\times d} \times \mathbb{R}^L$. The unlearning algorithm receives this model, the forget set, and retained data. It returns parameters θ^u whose behavior should approach a retrained reference θ^{rt} learned only from $\mathcal{D}_r$, while avoiding the cost of full retraining. MMU uses the source model as privileged information during unlearning. This model exhibits the complete pre-deletion behavior, i.e., it has useful knowledge on $\mathcal{D}_r$ and unwanted influence from $\mathcal{D}_f$. We freeze it as a privileged teacher and initialize a trainable student from the same weights. The same frozen network therefore plays two roles, a *good teacher* whose behavior is preserved on $\mathcal{D}_r$, and a *bad teacher* from which the student is pushed away on $\mathcal{D}_f$, so no second model has to be trained. Let $\mathcal{M} = \{m_1 < \cdots < m_K = d\}$ denote the set of nested embedding widths, where $K = |\mathcal{M}|$ is the number of Matryoshka levels, including the full embedding defined as:

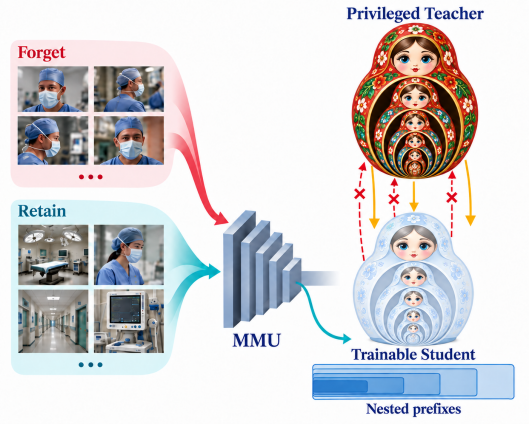

Figure 2: **MMU** symbolized as Matryoshka dolls. A frozen source model acts as a privileged teacher: its behavior is preserved on retained data and repelled on forgotten data across all embedding widths.

$$\mathcal{M} = \left\{ \left\lfloor \frac{d}{2^j} \right\rfloor : j = 0, \ldots, K-1 \right\} \cap \{m \in \mathbb{N}\}. \tag{1}$$

3.1 PRIVILEGED MATRYOSHKA CONSTRUCTION

For each $m \in \mathcal{M}$, MMU reads the first m coordinates of the student representation, as shown in Fig. 2. These prefixes form a Matryoshka hierarchy, i.e., each smaller representation is contained within each larger one. We use two heads:

$$z^{(m)}(x) = \begin{cases} W^{(m)}[f_\theta(x)]_{1:m} + b^{(m)}, & \text{MMU-1}, \\ W_{:,1:m}[f_\theta(x)]_{1:m} + b, & \text{MMU-2}. \end{cases} \tag{2}$$

MMU-1 assigns a classifier to each embedding width. MMU-2 slices one shared matrix. For MMU-1, we warm-start every head with $W^{(m)} = W^o_{:,1:m}$ and $b^{(m)} = b^o$. The full-width head equals the original classifier at initialization. Auxiliary heads are trained only during unlearning and then discarded. The privileged teacher remains fixed and supplies a width-aligned target through a sliced source classifier:

$$\tilde{z}^{(m)}(x) = W^o_{:,1:m}[f_{\theta^o}(x)]_{1:m} + b^o. \tag{3}$$

Applying both signals at every m prevents the optimization from satisfying the deletion request only at the final classifier.

3.2 PRIVILEGED PRESERVATION AND MARGIN-LIMITED REPULSION

Let $\sigma_T(z) = \mathrm{softmax}(z/T)$, where $T > 0$ is the distillation temperature. For a batch $\mathcal{B}$, define

$$\mathcal{K}^{(m)}(\mathcal{B}) = \frac{T^2}{|\mathcal{B}|} \sum_{x\in\mathcal{B}} \mathrm{KL}\Big(\sigma_T(\tilde{z}^{(m)}(x)) \,\|\, \sigma_T(z^{(m)}(x))\Big), \tag{4}$$

$$\mathcal{C}^{(m)}(\mathcal{B}) = \frac{1}{|\mathcal{B}|} \sum_{(x,y)\in\mathcal{B}} \ell_{\mathrm{CE}}(z^{(m)}(x), y). \tag{5}$$

MMU alternates between two passes. On a forget batch $\mathcal{B}_f$, the student is pushed away from the bad teacher only until a margin $\delta_m > 0$ is reached:

$$\mathcal{L}_{\text{forget}}(\mathcal{B}_f) = \sum_{m \in \mathcal{M}} \lambda_m^f \max\Big(0,\ \delta_m - \mathcal{K}^{(m)}(\mathcal{B}_f)\Big)\,. \tag{6}$$

The hinge loss is what makes the deletion request an explicit, bounded quantity. Maximizing $\mathcal{K}^{(m)}$ without bound, as in a plain max-step, has no fixed point. Because δ_m is expressed in the units of $\mathcal{K}^{(m)}$, it must be calibrated to the divergence the optimizer can actually reach in a given setting. A margin above that ceiling never binds, and one far below it deactivates after the first step. We select δ_m on validation data together with the learning rate. On a retain batch $\mathcal{B}_r$, MMU minimizes:

$$\mathcal{L}_{\text{retain}}(\mathcal{B}_r) = \sum_{m \in \mathcal{M}} \lambda_m^r \left[\gamma \mathcal{C}^{(m)}(\mathcal{B}_r) + \alpha \mathcal{K}^{(m)}(\mathcal{B}_r)\right]. \tag{7}$$

Cross-entropy keeps the correct labels, and privileged preservation limits drift from the good teacher on retained data. The forget pass runs for the first $E_{\max}$ of E epochs, while the retain pass runs in every epoch. Coordinate j receives gradients from every prefix with $m \geq j$, so early coordinates are constrained repeatedly, and later coordinates refine the representation. MMU therefore distributes the deletion signal through an ordered hierarchy rather than allowing a single full-width head to absorb the entire update.

3.3 Extension to Conditional Image Generation

We extended the MMU framework beyond classification unlearning to generation unlearning. For conditional diffusion, MMU applies privileged preservation and margin-limited repulsion to noise predictions. Given an image x, condition c, timestep t, and noise $\epsilon \sim \mathcal{N}(0, I)$, the forward process produces:

$$x_t = \sqrt{\bar{\alpha}_t}\, x + \sqrt{1 - \bar{\alpha}_t}\, \epsilon, \tag{8}$$

where $\bar{\alpha}_t$ is the cumulative noise-schedule coefficient.

Let $h_\theta(x_t, t, c) \in \mathbb{R}^{C \times H \times W}$ denote a latent representation within the student denoiser. For each $m \in \mathcal{M}$, the width-m readout uses the first m channels *only*, with the decoder's normalization statistics, affine parameters, and its output-convolution weights restricted to those channels:

$$\epsilon_\theta^{(m)}(x_t, t, c) = g_\theta^{(m)}([h_\theta(x_t, t, c)]_{1:m},\ t,\ c)\,. \tag{9}$$

Restricting the decoder is necessary, as zero-padding the truncated channels and passing the padded tensor through the unmodified decoder does *not* produce a width-m readout because all-zero groups emerge from group normalization as the learned shift β and inject a constant field into the output convolution.

For a data subset $\mathcal{S}$, define the privileged prediction discrepancy:

$$\mathcal{Q}^{(m)}(\mathcal{S}) = \mathbb{E}_{(x,c)\sim\mathcal{S},\, t,\, \epsilon}\left[\left\|\epsilon_\theta^{(m)}(x_t, t, c) - \text{sg}\left[\epsilon_{\theta^o}^{(m)}(x_t, t, c)\right]\right\|_2^2\right], \tag{10}$$

where sg is the stop gradient operator. The teacher and student receive the same noisy image, timestep, and condition. On retained data, MMU combines the standard denoising objective with privileged preservation:

$$\mathcal{L}_{\text{retain}}^{\text{gen}} = \sum_{m \in \mathcal{M}} \lambda_m^r \Big[\gamma\, \mathbb{E}_{(x,c)\sim\mathcal{D}_r,\, t,\, \epsilon}\left\|\epsilon_\theta^{(m)}(x_t, t, c) - \epsilon\right\|_2^2 + \alpha\, \mathcal{Q}^{(m)}(\mathcal{D}_r)\Big]. \tag{11}$$

On forgotten data, the same hinge loss as Eq. 6 applies, now in prediction space. It is evaluated on the *forgotten* latents x_t under the forgotten condition, since those are the states the sampler traverses when the deleted condition is requested:

$$\mathcal{L}_{\text{forget}}^{\text{gen}} = \sum_{m \in \mathcal{M}} \lambda_m^f\, \mathbb{E}_{(x,c)\sim\mathcal{D}_f,\, t,\, \epsilon}\Big[\max\Big(0,\ \delta_m - \mathcal{Q}_{x,c}^{(m)}\Big) + \big\|\epsilon_\theta^{(m)}(x_t, t, c) - \text{sg}[\epsilon_{\theta^o}^{(m)}(x_t, t, c_r)]\big\|_2^2\Big], \tag{12}$$

where c_r is a retained condition. The first term pushes the student away from the bad teacher's forgotten-condition prediction and stops at δ_m, while the second term attracts it toward the good teacher's retained-condition prediction on the same latent, thereby providing a direction.

Table 1: CIFAR-10 ResNet-18 random-forgetting results over ten trials. Parentheses report the absolute gap to Retrain; smaller gaps are better. The best unlearning methods for each metric (closest to Retrain) are **highlighted**, and the second best are underlined.

	10% random forgetting			50% random forgetting		
Method	D_f	D_r	Test Accuracy	D_f	D_r	Test Accuracy
Retrain (oracle)	$5.21_{\pm 0.42}$ (0.00)	$100.00_{\pm 0.00}$ (0.00)	$94.44_{\pm 0.15}$ (0.00)	$7.65_{\pm 0.19}$ (0.00)	$100.00_{\pm 0.00}$ (0.00)	$92.16_{\pm 0.29}$ (0.00)
Fine-tune on D_r (FT)	$0.13_{\pm 0.14}$ (5.08)	$99.92_{\pm 0.09}$ (0.08)	$94.05_{\pm 0.77}$ (0.39)	$17.44_{\pm 2.45}$ (9.79)	$87.03_{\pm 2.75}$ (12.97)	$81.53_{\pm 2.60}$ (10.63)
Random Labels (RL)	$0.03_{\pm 0.02}$ (5.18)	**100.00**$_{\pm 0.00}$ **(0.00)**	$94.47_{\pm 0.20}$ (0.03)	$9.65_{\pm 2.61}$ (2.00)	$92.60_{\pm 2.67}$ (7.40)	$87.10_{\pm 2.22}$ (5.06)
Gradient Ascent (GA)	$0.00_{\pm 0.00}$ (5.21)	**100.00**$_{\pm 0.00}$ **(0.00)**	$94.90_{\pm 0.19}$ (0.46)	$0.00_{\pm 0.00}$ (7.65)	**100.00**$_{\pm 0.00}$ **(0.00)**	$94.90_{\pm 0.18}$ (2.74)
Influence Unlearning (IU)	$0.10_{\pm 0.26}$ (5.11)	$99.89_{\pm 0.34}$ (0.11)	$94.30_{\pm 0.88}$ (0.14)	$12.03_{\pm 14.34}$ (4.38)	$88.23_{\pm 14.10}$ (11.77)	$81.60_{\pm 12.86}$ (10.56)
Boundary Expanding (BE)	$0.00_{\pm 0.00}$ (5.21)	**100.00**$_{\pm 0.00}$ **(0.00)**	$94.91_{\pm 0.20}$ (0.47)	$0.00_{\pm 0.00}$ (7.65)	**100.00**$_{\pm 0.00}$ **(0.00)**	$94.92_{\pm 0.19}$ (2.76)
Boundary Shrink (BS)	$0.00_{\pm 0.00}$ (5.21)	**100.00**$_{\pm 0.00}$ **(0.00)**	$94.92_{\pm 0.19}$ (0.48)	$0.00_{\pm 0.00}$ (7.65)	**100.00**$_{\pm 0.00}$ **(0.00)**	$94.94_{\pm 0.20}$ (2.78)
ℓ_1-sparse	$8.46_{\pm 1.50}$ (3.25)	$94.55_{\pm 1.34}$ (5.45)	$88.63_{\pm 1.21}$ (5.82)	**7.57**$_{\pm 0.73}$ **(0.08)**	$94.99_{\pm 0.73}$ (5.01)	$87.91_{\pm 0.72}$ (4.25)
SalUn	$0.62_{\pm 0.14}$ (4.59)	$99.99_{\pm 0.01}$ (0.01)	$94.34_{\pm 0.14}$ (0.10)	$0.62_{\pm 0.22}$ (7.04)	$99.86_{\pm 0.07}$ (0.14)	$93.59_{\pm 0.21}$ (1.42)
SalUn-soft	$0.02_{\pm 0.01}$ (5.19)	**100.00**$_{\pm 0.00}$ **(0.00)**	$94.60_{\pm 0.20}$ (0.16)	$3.53_{\pm 0.57}$ (4.13)	$99.31_{\pm 0.22}$ (0.69)	$90.96_{\pm 0.45}$ (1.20)
MMU (ours)	**4.19**$_{\pm 0.00}$ **(1.02)**	**100.00**$_{\pm 0.00}$ **(0.00)**	**94.42**$_{\pm 0.19}$ **(0.02)**	$4.39_{\pm 0.53}$ (3.26)	$99.89_{\pm 0.31}$ (0.18)	**92.81**$_{\pm 0.81}$ **(0.65)**
Highest Frequency: Best: MMU (4x) Second Best: RL & SalUn (2x)						

4 Experiments & Discussion

We test four questions. Does MMU approach the retrained model on forgotten data? Does it preserve retained and test accuracy? Does nesting improve forgetting beyond a single-width objective? Does any apparent forgetting survive relearning and membership-inference tests?

Experimental Setup. We evaluate class-, subclass-, and instance-level deletion on CIFAR-10 and CIFAR-100 (Krizhevsky & Hinton, 2009), and Robotic Tactile Internet (RTI) datasets using CNN ($d = 64$) with four levels, $\mathcal{M} = \{8, 16, 32, 64\}$, and ResNet-18 ($d = 512$) with five levels, $\mathcal{M} = \{32, 64, 128, 256, 512\}$. The implementation clips the repulsion-step gradient norm to 5.0. We include three reference methods: *Original* (or joint training), the source model trained on the complete dataset $\mathcal{D}_r \cup \mathcal{D}_f$ without unlearning; *Fine-tuning*, which updates the Original model using only the accessible subset of $\mathcal{D}_r$ after removing $\mathcal{D}_f$; and *Retrain*, which trains a new model from scratch using only $\mathcal{D}_r$ and serves as the gold-standard unlearning reference. We further compare MMU with established unlearning methods, including Bad Teacher (Chundawat et al., 2023), Amnesiac (Graves et al., 2021) (evaluated only for class-level deletion), UNlearning by Selective Impair and Repair (UNSIR) (Tarun et al., 2024), Selective Synaptic Dampening (SSD) (Foster et al., 2024), Saliency Unlearning (SalUn) (Fan et al., 2024), Unified Continual Learning and Machine Unlearning (UniCLUN)† (Chatterjee et al., 2024), SCalable Remembering and Unlearning unBound (SCRUB) (Kurmanji et al., 2023), Random Labeling (RL) (Golatkar et al., 2020a), Gradient Ascent (GA) (Thudi et al., 2022), Influence Unlearning (IU) (Izzo et al., 2021), and ℓ_1-sparse unlearning (Jia et al., 2023). We additionally consider two boundary-based methods, Boundary Shrink (BS) and Boundary Expanding (BE) (Chen et al., 2023). Every method is selected using validation data under a utility constraint relative to Retrain, and the official test set is evaluated only after selection. We report the absolute forget-accuracy gap to Retrain $|\Delta\mathcal{D}_f|$, retain and test accuracy, membership-inference attack performance, predictive and parameter-distribution KL divergence, unlearning time, and post-unlearning relearning accuracy. Additional experimental results are in Appendix C. Each experiment uses a single NVIDIA RTX PRO 6000 GPU.

Robotic Tactile Internet dataset. To study the PAI setting, we collected the RTI dataset containing haptic tele-manipulation signals from a virtual client–server environment implemented in CHAI3D. The experimental platform follows the IEEE 1918.1.1 architecture and uses a Novint Falcon haptic device operating at a 1 kHz control loop. Twenty anonymized participants aged 20 to 33 controlled a virtual tool while performing five actions: no movement, pressing, tapping, free-air movement, and dragging.

Each action was performed for 10 seconds, and every participant repeated the complete five-action sequence four times, giving 20 action segments per participant. Each participant's record contains human-side video, robot-side video, virtual-environment image frames, and the command signals sent from the human side and the feedback signals returned from the robot side. Each participant forms one task whose five classes are that participant's actions, so the full label space has $20 \times 5 = 100$ classes. We consider three levels of deletion. *Class deletion* removes all data of one action of one participant, or all five actions for participant-level removal, so the label leaves the output space.

Table 2: Forget-set (D_f) and retain-set (D_r) accuracy on CIFAR-10, CIFAR-100, and RTI, averaged over 3 seeds. The best methods for each metrics are **highlighted**, and the second best are underlined.

	$D_f \to D_f^{\mathrm{RT}}$						$D_r \uparrow$					
	CNN			ResNet-18			CNN			ResNet-18		
Method	Class	Subclass	Instance	Class	Subclass	Instance	Class	Subclass	Instance	Class	Subclass	Instance
CIFAR-10												
Retrain (oracle)	$0.00_{\pm 0.00}$	$25.40_{\pm 4.50}$	$91.86_{\pm 1.34}$	$0.00_{\pm 0.00}$	$26.57_{\pm 1.23}$	$92.06_{\pm 0.75}$	$92.76_{\pm 0.23}$	$94.89_{\pm 0.11}$	$92.13_{\pm 0.28}$	$93.24_{\pm 0.40}$	$95.16_{\pm 0.07}$	$92.18_{\pm 0.30}$
Original model	$93.47_{\pm 1.33}$	$94.57_{\pm 1.17}$	$99.94_{\pm 0.10}$	$93.90_{\pm 1.15}$	$95.30_{\pm 0.69}$	$100.00_{\pm 0.00}$	$92.24_{\pm 0.24}$	$94.51_{\pm 0.14}$	$92.37_{\pm 0.34}$	$92.74_{\pm 0.36}$	$95.16_{\pm 0.06}$	$92.86_{\pm 0.23}$
Fine-tune on D_r	$50.17_{\pm 43.28}$	$68.90_{\pm 2.80}$	$96.19_{\pm 0.44}$	$26.23_{\pm 45.44}$	$70.07_{\pm 0.68}$	$97.43_{\pm 2.04}$	$92.20_{\pm 0.89}$	$93.51_{\pm 0.94}$	$90.12_{\pm 0.33}$	$92.13_{\pm 0.53}$	$94.13_{\pm 0.60}$	$91.28_{\pm 1.19}$
Unlearning methods												
Amnesiac	**0.00**$_{\pm 0.00}$	–	–	**0.00**$_{\pm 0.00}$	–	–	$11.11_{\pm 0.00}$	–	–	$11.11_{\pm 0.00}$	–	–
Bad Teacher	$0.27_{\pm 0.15}$	**30.00**$_{\pm 4.71}$	$98.29_{\pm 0.84}$	$0.13_{\pm 0.23}$	**28.40**$_{\pm 4.10}$	$98.97_{\pm 0.36}$	$92.54_{\pm 0.07}$	$94.27_{\pm 0.22}$	$89.94_{\pm 0.50}$	$92.89_{\pm 0.14}$	**94.37**$_{\pm 0.78}$	$90.53_{\pm 0.64}$
SCRUB	**0.00**$_{\pm 0.00}$	$90.73_{\pm 1.63}$	**97.52**$_{\pm 0.81}$	**0.00**$_{\pm 0.00}$	$84.57_{\pm 0.06}$	$98.82_{\pm 0.66}$	$92.73_{\pm 0.49}$	**94.63**$_{\pm 0.09}$	$90.44_{\pm 0.51}$	$93.24_{\pm 0.31}$	$93.68_{\pm 0.36}$	$91.23_{\pm 0.59}$
SSD	$74.60_{\pm 22.39}$	$65.57_{\pm 28.00}$	$99.94_{\pm 0.10}$	$92.67_{\pm 3.43}$	$84.33_{\pm 9.85}$	$100.00_{\pm 0.00}$	$88.36_{\pm 6.07}$	$78.58_{\pm 25.89}$	**92.40**$_{\pm 0.39}$	$92.71_{\pm 0.38}$	$94.13_{\pm 0.84}$	**92.89**$_{\pm 0.26}$
SalUn	$41.50_{\pm 35.98}$	$73.40_{\pm 3.05}$	$98.02_{\pm 0.14}$	$22.07_{\pm 38.22}$	$68.30_{\pm 4.06}$	**96.25**$_{\pm 2.23}$	$91.66_{\pm 0.43}$	$93.70_{\pm 0.48}$	$89.89_{\pm 0.19}$	$91.65_{\pm 0.34}$	$93.63_{\pm 0.26}$	$88.76_{\pm 1.68}$
UNSIR	**0.00**$_{\pm 0.00}$	$79.13_{\pm 10.71}$	$98.61_{\pm 0.31}$	$43.23_{\pm 10.05}$	$82.30_{\pm 4.86}$	$98.88_{\pm 0.58}$	$91.86_{\pm 0.43}$	$93.37_{\pm 0.29}$	$90.52_{\pm 0.01}$	$92.04_{\pm 0.14}$	$93.97_{\pm 0.35}$	$90.74_{\pm 0.70}$
UniCLUN	$41.33_{\pm 35.77}$	$74.40_{\pm 4.29}$	$99.12_{\pm 0.23}$	$41.93_{\pm 36.42}$	$68.10_{\pm 3.86}$	$98.38_{\pm 0.80}$	$91.54_{\pm 0.63}$	$93.65_{\pm 0.59}$	$90.50_{\pm 0.45}$	$92.06_{\pm 0.19}$	$93.87_{\pm 0.19}$	$90.48_{\pm 1.05}$
MMU (ours)	**0.00**$_{\pm 0.00}$	$55.47_{\pm 4.35}$	$97.88_{\pm 0.49}$	**0.00**$_{\pm 0.00}$	$55.37_{\pm 3.03}$	$98.23_{\pm 0.31}$	**92.90**$_{\pm 0.29}$	$93.56_{\pm 0.11}$	$90.85_{\pm 0.38}$	**93.34**$_{\pm 0.25}$	$93.20_{\pm 0.16}$	$90.71_{\pm 0.46}$
CIFAR-10 Ranking: Best: MMU (4 First, 5 Second) Second Best: SCRUB (4 First, 3 Second)												
CIFAR-100												
Retrain (oracle)	$0.00_{\pm 0.00}$	$23.33_{\pm 4.16}$	$69.50_{\pm 0.60}$	$0.00_{\pm 0.00}$	$24.00_{\pm 2.00}$	$69.58_{\pm 2.46}$	$68.51_{\pm 0.39}$	$79.47_{\pm 0.15}$	$68.17_{\pm 0.27}$	$70.74_{\pm 0.33}$	$80.04_{\pm 0.49}$	$69.28_{\pm 0.73}$
Original model	$70.00_{\pm 3.61}$	$77.00_{\pm 3.46}$	$99.14_{\pm 0.29}$	$72.33_{\pm 3.21}$	$80.00_{\pm 0.00}$	$100.00_{\pm 0.00}$	$68.58_{\pm 0.52}$	$79.37_{\pm 0.84}$	$68.60_{\pm 0.48}$	$70.58_{\pm 0.28}$	$80.69_{\pm 0.36}$	$70.60_{\pm 0.27}$
Fine-tune on D_r	$35.00_{\pm 3.61}$	$72.00_{\pm 7.00}$	$98.14_{\pm 0.71}$	$36.33_{\pm 4.51}$	$75.33_{\pm 4.93}$	$99.89_{\pm 0.19}$	$68.32_{\pm 0.46}$	$79.08_{\pm 0.51}$	$68.42_{\pm 0.39}$	$70.38_{\pm 0.22}$	$80.26_{\pm 0.13}$	$70.23_{\pm 0.46}$
Unlearning methods												
Amnesiac	**0.00**$_{\pm 0.00}$	–	–	**0.00**$_{\pm 0.00}$	–	–	$1.38_{\pm 0.33}$	–	–	$1.27_{\pm 0.23}$	–	–
Bad Teacher	$0.33_{\pm 0.58}$	**17.00**$_{\pm 4.58}$	$96.72_{\pm 1.21}$	**0.00**$_{\pm 0.00}$	**17.00**$_{\pm 3.00}$	$99.31_{\pm 0.77}$	$67.24_{\pm 0.31}$	$77.89_{\pm 0.51}$	$66.45_{\pm 0.80}$	$69.33_{\pm 0.18}$	$79.22_{\pm 0.41}$	$68.47_{\pm 1.41}$
SCRUB	$64.33_{\pm 9.02}$	$73.33_{\pm 6.11}$	$90.86_{\pm 1.84}$	$31.67_{\pm 9.87}$	$79.67_{\pm 0.58}$	$97.94_{\pm 0.98}$	$68.51_{\pm 0.28}$	$79.16_{\pm 0.89}$	$65.88_{\pm 0.42}$	$70.65_{\pm 0.25}$	**80.64**$_{\pm 0.40}$	$68.40_{\pm 0.76}$
SSD	$8.00_{\pm 2.65}$	$51.00_{\pm 30.81}$	$97.06_{\pm 1.38}$	$35.00_{\pm 18.03}$	$57.67_{\pm 31.21}$	$98.94_{\pm 0.46}$	$67.94_{\pm 0.85}$	$78.33_{\pm 0.64}$	**66.96**$_{\pm 0.90}$	$70.50_{\pm 0.12}$	$78.74_{\pm 3.08}$	**68.96**$_{\pm 1.06}$
SalUn	$23.67_{\pm 1.53}$	$37.67_{\pm 5.69}$	$74.81_{\pm 2.72}$	$29.00_{\pm 2.65}$	$42.00_{\pm 3.61}$	**84.53**$_{\pm 1.61}$	$63.30_{\pm 1.03}$	$75.57_{\pm 0.84}$	$55.93_{\pm 0.99}$	$66.53_{\pm 0.50}$	$76.86_{\pm 0.48}$	$58.67_{\pm 0.73}$
UNSIR	$37.33_{\pm 17.67}$	$64.33_{\pm 0.58}$	$96.03_{\pm 1.07}$	$44.67_{\pm 10.21}$	$77.33_{\pm 6.11}$	$98.86_{\pm 0.81}$	$67.60_{\pm 0.57}$	$78.28_{\pm 0.42}$	$66.03_{\pm 0.06}$	$69.55_{\pm 0.42}$	$79.64_{\pm 0.86}$	$68.03_{\pm 1.06}$
UniCLUN	$26.67_{\pm 3.79}$	$35.67_{\pm 3.51}$	**74.58**$_{\pm 1.04}$	$30.67_{\pm 1.53}$	$44.67_{\pm 4.16}$	$85.31_{\pm 1.41}$	$63.07_{\pm 0.90}$	$75.13_{\pm 0.21}$	$56.55_{\pm 0.96}$	$66.66_{\pm 0.58}$	$77.01_{\pm 0.60}$	$58.37_{\pm 2.36}$
MMU (ours)	**0.00**$_{\pm 0.00}$	$68.67_{\pm 5.86}$	$90.61_{\pm 0.67}$	**0.00**$_{\pm 0.00}$	$63.33_{\pm 2.52}$	$97.31_{\pm 0.55}$	**68.74**$_{\pm 0.25}$	**79.31**$_{\pm 0.65}$	$65.87_{\pm 0.50}$	**70.91**$_{\pm 0.14}$	$80.57_{\pm 0.45}$	$68.38_{\pm 0.88}$
CIFAR-100 Ranking: Best: MMU (5 First, 1 Second) Second Best: Bad Teacher (3 First, 3 Second)												
RTI												
Retrain (oracle)	$0.00_{\pm 0.00}$	$86.33_{\pm 2.31}$	$78.57_{\pm 0.35}$	$0.00_{\pm 0.00}$	$86.67_{\pm 3.21}$	$80.99_{\pm 1.01}$	$80.77_{\pm 0.13}$	$78.94_{\pm 0.18}$	$78.85_{\pm 0.33}$	$82.26_{\pm 0.41}$	$80.54_{\pm 0.23}$	$80.05_{\pm 0.50}$
Original model	$72.87_{\pm 1.68}$	$96.33_{\pm 2.08}$	$99.53_{\pm 0.18}$	$71.27_{\pm 2.14}$	$94.67_{\pm 2.52}$	$99.97_{\pm 0.05}$	$79.73_{\pm 0.48}$	$79.25_{\pm 0.07}$	$79.39_{\pm 0.38}$	$81.42_{\pm 0.02}$	$80.54_{\pm 0.34}$	$80.92_{\pm 0.09}$
Fine-tune on D_r	$28.40_{\pm 3.12}$	$95.33_{\pm 1.53}$	$99.36_{\pm 0.05}$	$36.87_{\pm 5.08}$	$95.67_{\pm 1.53}$	$99.82_{\pm 0.15}$	$80.55_{\pm 0.37}$	$79.12_{\pm 0.15}$	$79.01_{\pm 0.25}$	$81.93_{\pm 0.40}$	$80.14_{\pm 0.28}$	$80.74_{\pm 0.43}$
Unlearning methods												
Amnesiac	**0.00**$_{\pm 0.00}$	–	–	**0.00**$_{\pm 0.00}$	–	–	$5.26_{\pm 0.00}$	–	–	$5.26_{\pm 0.00}$	–	–
Bad Teacher	$0.13_{\pm 0.12}$	$45.00_{\pm 16.52}$	$97.81_{\pm 0.40}$	**0.00**$_{\pm 0.00}$	$31.67_{\pm 4.16}$	$99.01_{\pm 0.83}$	$79.56_{\pm 0.51}$	$77.59_{\pm 0.47}$	$77.03_{\pm 0.69}$	$81.44_{\pm 0.26}$	$79.09_{\pm 0.44}$	$79.14_{\pm 1.04}$
SCRUB	$52.53_{\pm 9.49}$	$95.67_{\pm 1.53}$	$96.29_{\pm 2.78}$	**0.00**$_{\pm 0.00}$	$95.00_{\pm 2.00}$	$98.83_{\pm 0.48}$	$80.38_{\pm 0.54}$	**79.28**$_{\pm 0.09}$	$77.35_{\pm 2.01}$	$82.04_{\pm 0.03}$	$80.50_{\pm 0.41}$	**79.40**$_{\pm 0.17}$
SSD	$37.47_{\pm 30.91}$	**90.33**$_{\pm 4.04}$	$99.12_{\pm 0.58}$	$36.67_{\pm 32.30}$	$94.67_{\pm 2.52}$	**92.16**$_{\pm 11.02}$	$80.38_{\pm 0.67}$	$79.14_{\pm 0.10}$	**78.71**$_{\pm 1.01}$	$81.00_{\pm 0.90}$	**80.54**$_{\pm 0.34}$	$74.50_{\pm 7.98}$
SalUn	$33.60_{\pm 2.23}$	$64.67_{\pm 3.06}$	**89.74**$_{\pm 1.90}$	$38.13_{\pm 1.10}$	$70.67_{\pm 3.06}$	$94.68_{\pm 0.28}$	$77.99_{\pm 0.64}$	$73.33_{\pm 0.28}$	$72.35_{\pm 0.36}$	$80.54_{\pm 0.28}$	$74.78_{\pm 0.54}$	$74.00_{\pm 0.57}$
UNSIR	$11.80_{\pm 17.25}$	$95.67_{\pm 2.31}$	$97.37_{\pm 0.23}$	$51.73_{\pm 18.15}$	$95.33_{\pm 1.53}$	$98.71_{\pm 0.56}$	$79.34_{\pm 1.14}$	$77.96_{\pm 0.89}$	$77.16_{\pm 0.58}$	$80.95_{\pm 0.11}$	$79.49_{\pm 0.33}$	$78.28_{\pm 0.83}$
UniCLUN	$35.13_{\pm 1.27}$	$74.33_{\pm 2.52}$	$90.58_{\pm 1.55}$	$39.33_{\pm 0.31}$	$70.00_{\pm 2.00}$	$95.15_{\pm 1.88}$	$78.02_{\pm 0.28}$	$72.85_{\pm 0.44}$	$72.58_{\pm 1.04}$	$80.52_{\pm 0.27}$	$74.90_{\pm 0.34}$	$74.93_{\pm 1.06}$
MMU (ours)	**0.00**$_{\pm 0.00}$	**90.33**$_{\pm 3.79}$	$95.41_{\pm 0.60}$	**0.00**$_{\pm 0.00}$	**90.67**$_{\pm 1.53}$	$98.71_{\pm 0.13}$	**80.67**$_{\pm 0.51}$	$79.11_{\pm 0.21}$	$77.04_{\pm 0.72}$	**82.41**$_{\pm 0.06}$	$80.35_{\pm 0.27}$	$79.15_{\pm 0.02}$
RTI Ranking: Best: MMU (6 First, 1 Second) Second Best: SSD (4 First, 4 Second)												
Overall Ranking (All Datasets): Best: MMU (15 First, 7 Second) Second Best: SCRUB (8 First, 11 Second)												

Notes. Retrain, Original model, and fine-tuning are reference rows. Amnesiac is evaluated on class deletion only; missing entries are shown as –.

Subclass deletion removes one of the four repetitions of an action while keeping the label. *Instance deletion* removes individual windows, frames, or video clips of a specific participant.

4.1 Result Discussion

Table 1 shows that Retraining (Oracle) classifies 5.21% of the forget set correctly at a 10% ratio and 7.65% at 50%, as deleted examples resemble retained ones and a model that never saw them still gets some correct. Therefore, the target should be the oracle, not the zero accuracy in the forget set, and Table 1 is full of overshooting methods: GA, BE, BS, and SalUn-soft. On the other hand, MMU achieves $\mathcal{D}_f = 4.19\%$, a gap of 1.02 from the oracle, while maintaining retention accuracy at 100.00%.

In Table 2, every competitor has a scope where it collapses. SSD leaves 92.67% of a deleted class intact on CIFAR-10/ResNet-18, which is barely below the Original model. SalUn keeps 22.07% on the same cell, yet wins several instance-deletion cells outright. UNSIR swings from 0.0% on CIFAR-10/CNN to 43.2% on the same dataset with a different backbone. SCRUB removes the class perfectly on ResNet-18 but retains 52.53% on the CNN for RTI. MMU reaches 0.0% class-level forget accuracy across all six datasets and backbone cells.

Table 3: U-LiRA per-example membership inference under instance deletion, with all three datasets side by side. AUC is ranked by distance to the ideal value of 50; TPR@1% FPR is ranked by lower value. The best unlearning methods for each metric are **highlighted**, and the second best are underlined.

	CIFAR-10				CIFAR-100				RTI			
	CNN		ResNet-18		CNN		ResNet-18		CNN		ResNet-18	
Method	AUC→ 50	TPR@1%↓	AUC→ 50	TPR@1%↓	AUC→ 50	TPR@1%↓	AUC→ 50	TPR@1%↓	AUC→ 50	TPR@1%↓	AUC→ 50	TPR@1%↓
Retrain (oracle)	50.0	0.0	50.0	0.0	50.00	–	50.00	–	50.00	–	50.00	–
Original model	66.8	12.3	80.1	22.4	81.63	22.20	92.90	50.16	79.39	21.27	81.67	29.22
Fine-tune on D_r	50.9	1.5	54.8	3.8	81.13	21.00	91.90	47.40	78.11	19.87	81.98	28.49
Unlearning methods												
Bad Teacher	88.9	0.4	76.2	1.9	79.32	7.53	85.17	29.51	74.06	16.64	73.96	18.60
SCRUB	60.3	7.6	68.1	12.0	**70.94**	12.56	82.67	30.56	73.49	15.13	72.37	17.49
SSD	59.6	8.3	75.9	18.5	76.47	16.84	87.87	39.84	75.43	18.09	80.80	28.24
SalUn	87.4	**0.0**	60.3	**0.0**	89.45	**2.98**	75.83	1.36	91.05	**1.73**	84.24	**0.40**
UNSIR	**54.8**	2.0	**57.5**	4.3	71.17	8.84	82.07	26.42	**71.82**	13.31	**70.36**	14.27
UniCLUN	56.7	4.2	80.8	**0.0**	90.38	4.07	74.41	**1.04**	91.90	3.22	85.33	0.49
MMU (ours)	59.1	7.6	59.0	11.0	71.71	3.09	**73.95**	1.62	73.05	15.53	72.89	18.38
CIFAR-10 Ranking: Best: UNSIR & SalUn (2x) Second Best: Bad Teacher (2x)												
CIFAR-100 Ranking: Best: UniCLUN, MMU, SCRUB, SalUn (1x) Second Best: UNSIR, MMU, UniCLUN, SalUn (1x)												
RTI Ranking: Best: UNSIR & SalUn (2x) Second Best: MMU & UniCLUN (1x)												
Overall Ranking: Best: SalUn (5 First, 1 Second) Second Best: UNSIR (4 First, 1 Second)												

Notes. Retrain, Original model, and fine-tuning are reference rows and receive no medals. Amnesiac is evaluated on class deletion only and is not measured here. Missing entries are shown as –.

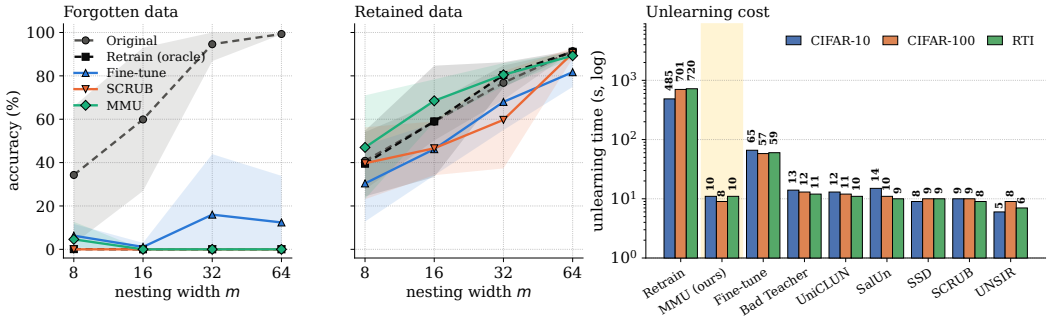

Figure 3: (a),(b). Accuracy at each nesting width m. Shading is ± 1 s.d. (c).Time Comparison for unlearning

The second part of Table 2 reports per-example membership auditing using U-LiRA, as described in Hayes et al. (2024). The reason that we adopted this test is that aggregate membership-inference accuracy is a weak audit, since a method can, on average, be at chance while individual forgotten examples remain identifiable. For each cell, we train $N_{\text{sh}} = 16$ shadow models, place every forgotten example into exactly $N_{\text{sh}}/2$ shadow training sets, and apply the same unlearning algorithm to each shadow model. For a forgotten example x, we fit Gaussians to its logit-scaled confidence under the shadows that saw it and those that did not, sharing a single variance across examples because per-example estimates from $N_{\text{sh}}/2$ shadows are too noisy, and score the target model using the log-likelihood ratio. The paired Retrain model supplies the negatives, so the attack asks whether the unlearned model is distinguishable from Retrain on each individual example. We report Area Under the Curve (AUC) and True Positive Rate (TPR) at a 1% False Positive Rate (FPR), and the two metrics tell different stories. The attack separates these models from Retrain easily on average, but barely makes a confident individual call. MMU sits in the opposite regime, with AUC closer to the oracle (59.0–73.05 across datasets) but a higher confident-detection rate. An auditor checking whether deletion happened at all reads AUC, while a data subject asking whether their record is exposed reads TPR at low FPR. Reporting both reveals differences that either metric alone can obscure.

Fig. 3.(a) shows that the MMU keeps every prefix of the representation usable after deletion. On forgotten data, the widths are similar: Retrain, SCRUB, and MMU are at or near 0% for every m, whereas fine-tuning reaches up to 16% at $m = 32$. Retained data separates the methods. MMU supervises every prefix and matches or exceeds Retrain across all four widths. Nested supervision is what keeps the small prefixes useful after deletion. Fig. 3.(c) shows that MMU is one to two orders

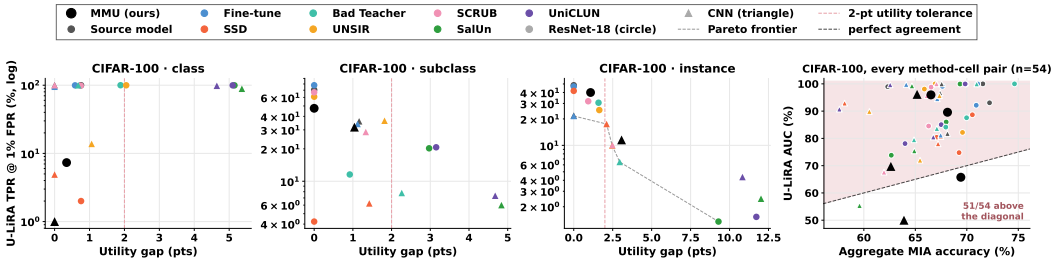

Figure 4: Utility–leakage trade-off on CIFAR-100

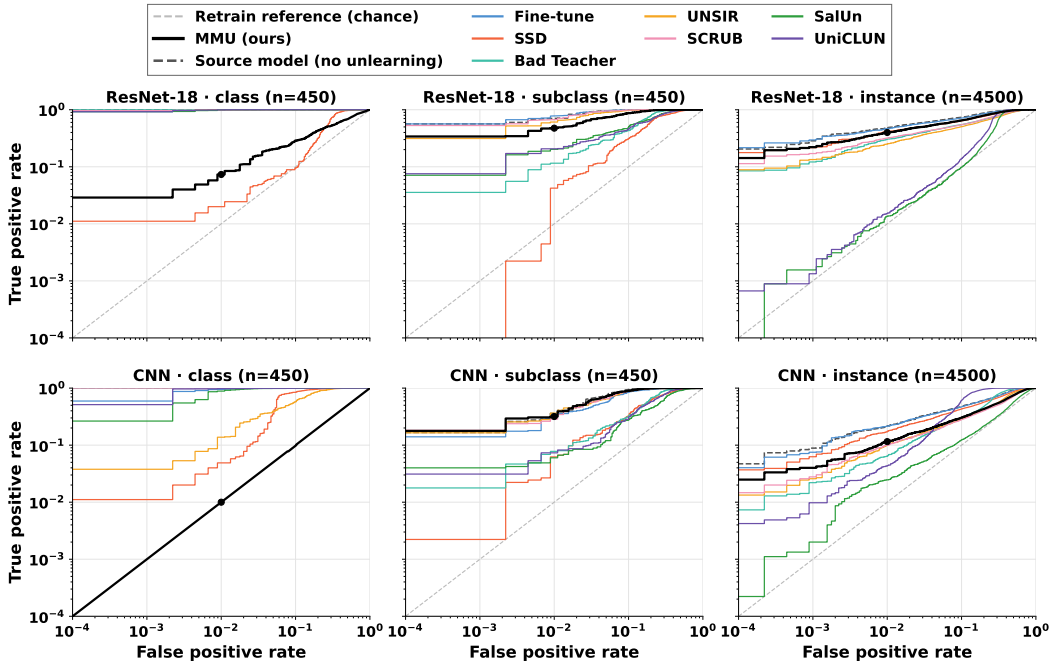

Figure 5: TPR vs FPR on CIFAR-100 for ResNet-18 and CNN. n is the number of forgotten examples. The dot marks MMU at the 1% FPR operating point.

of magnitude cheaper than retraining and among the cheapest approximate methods on every dataset. Each dot represents a method's average unlearning time, and the colored bar spans from its fastest to its slowest run. The time axis is logarithmic, so read each method's cost from where its dot sits, not from the length of its line.

Utility–leakage trade-off. Fig. 4 reveals that under class deletion, the source model is perfectly attackable (AUC 100.00, TPR@1%FPR 100.00), and MMU is the only method that reaches chance: AUC 49.99 with TPR 1.00 on the CNN, and 65.74 with TPR 7.33 on ResNet-18, both inside the tolerance (0.00 and 0.34 points). Every other method stays above 89 AUC on at least one backbone. Under subclass and instance deletion no method reaches chance. MMU's leakage rises to AUC 96.15 (CNN) and 95.96 (ResNet-18) on subclass deletion, against source values of 96.71 and 98.96, so the attack is barely weakened. The utility side is where MMU separates. It sits within the two-point tolerance in five of six cell with gaps of 0.00 to 1.08 points.

Fig. 5 shows that MMU reduces CNN class deletion to chance, whereas at the subclass level, it remains close to the source model. Here, positives are the unlearned model's per-example log-likelihood ratio, and negatives are the retrain reference. Thus, a method that has genuinely removed the forget set tracks the grey diagonal. The source model is recovered perfectly on class deletion (AUC = 100.00, TPR@1%FPR = 100.00 on both backbones) while MMU reduces this to AUC = 49.99 (TPR@1% = 1.00) on the CNN and to 65.74 (7.33) on ResNet-18.

Suppressing nudity in text-to-image generation. We test MMU on concept-wise forgetting, where the target is NSFW (not safe for work) content that inappropriate image prompts (I2P) (Schramowski et al., 2023) extract from a text-to-image model. We feed the I2P prompts into the open-source SD v1.4 (Rombach et al., 2022) and run the NudeNet detector (Bedapudi, 2019) on the generated images, which labels the exposed body parts in each image.

Table 4: Nudity suppression on held-out prompts.

	Metric	Source	MMU	SalUn	Random mask
Forget	Detection rate (%) ↓	40.0	3.3	0.6	**0.0**
	Max exposed score ↓	0.414	0.052	**0.008**	**0.008**
	Context CLIP ↑	0.347	**0.337**	0.319	0.330
Retain	CLIP ↑	0.338	**0.323**	0.313	0.315
	LPIPS to Source ↓	–	**0.478**	0.547	0.511
	False positives (%) ↓	3.3	**0.0**	**0.0**	**0.0**

We followed Fan et al. (2024) in this framework. The goal is simple. An MU method should remove the influence of nudity-related prompts in SD, so those prompts no longer produce nude images. Fig. 6 compares the removal of harmful image content (nudity) from source data with the retention of safe content. In the absence of unlearning methods, the original SD can generate numerous harmful images, underscoring the importance of MU for image generation. In addition, we evaluate on 90 held-out prompts, fixed before sampling and disjoint from the training and development prompts: 30 describe adult nudity, 30 describe clothed people in matched settings, and 30 describe unrelated scenes. Each prompt is sampled with two noise draws. Table 4 shows a trade-off rather than a single winner. All three methods reduce detections on held-out nudity prompts from 40.0% to below 4%. SalUn and the random mask suppress more completely (0.6% and 0.0%) than MMU (3.3%, on average two of 60 images). MMU instead stays closest to the source model on every retained metric.

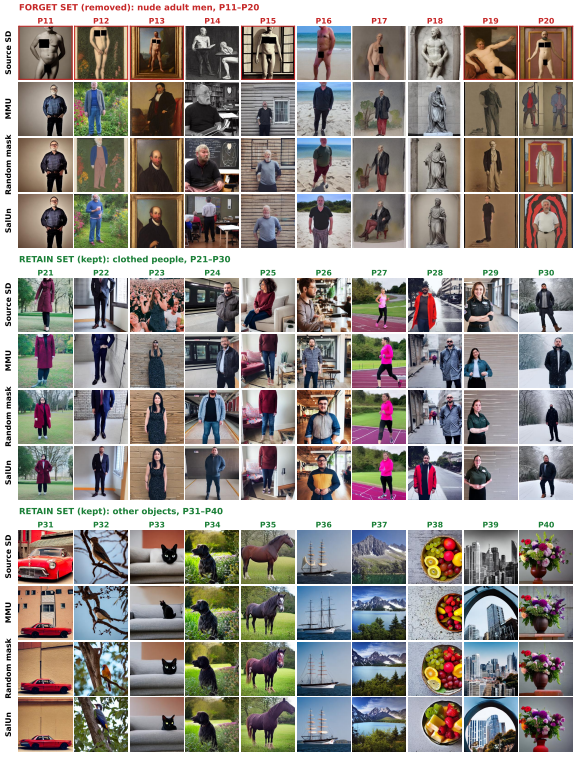

Figure 6: Suppressing nudity in text-to-image generation remaining prompts of Fig 1.

5 Conclusion

In this work, we propose MMU, which uses nested embedding width for unlearning. A single frozen source model uses privileged knowledge to guide a student with the same architecture, without giving the student direct access to the privileged knowledge. MMU reads representations through nested prefixes and applies unlearning at multiple representation scales rather than only at the final classifier.

Limitation: Subclass deletion remains the largest gap to Retrain across all methods we tested, including MMU, and the direction of the error varies by dataset, suggesting that the difficulty is structural rather than a matter of tuning. In addition, MMU falls short in instance-level removal. We believe nesting forces instance-level embeddings to be codable at different widths. To address this, we need an importance estimator, such as Fisher Information or Shapley Value, to identify instance-related features across embeddings, which we will investigate in future work.

Future work. The construction assumes an encoder whose coordinates can be ordered and truncated, which holds for the vision backbones studied here but needs rethinking for language models, where the natural units are tokens rather than embedding dimensions.

AI USE STATEMENT

In this work, we used generative AI tools to polish the text, fix typos, improve wording, and check grammar. We have not used generative AI tools for methodology development, dataset creation, or BibTeX reference creation (No hallucination). We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Anonymous GitHub: https://anonymous.4open.science/r/Matryoshka-Machine-Unlearning-32DC/

We provided a complete version of the code to reproduce every result of the paper at the anonymous GitHub link and uploaded it to the **supplementary materials on the submission page**. The code and the RTI dataset will be published upon acceptance of the paper.

REFERENCES

Praneeth Bedapudi. Nudenet: Neural nets for nudity classification, detection and selective censoring. https://github.com/notAI-tech/NudeNet, 2019.

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)*, 2021. doi: 10.1109/SP40001.2021.00019.

Romit Chatterjee, Vikram Chundawat, Ayush Tarun, Ankur Mali, and Murari Mandal. A unified framework for continual learning and machine unlearning, 2024. URL https://arxiv.org/abs/2408.11374.

Min Chen, Weizhuo Gao, Gaoyang Liu, Kai Peng, and Chen Wang. Boundary unlearning: Rapid forgetting of deep networks via shifting the decision boundary. In *CVPR*, pp. 7766–7775, 2023.

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 *AAAI*, 2023.

Chongyu Fan, Jiancheng Liu, Yihua Zhang, Eric Wong, Dennis Wei, and Sijia Liu. SalUn: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation. In *ICLR*, 2024.

Jack Foster, Stefan Schoepf, and Alexandra Brintrup. Fast machine unlearning without retraining through selective synaptic dampening. In *AAAI*, 2024.

Aditya Golatkar, Alessandro Achille, and Stefano Soatto. Eternal sunshine of the spotless net: Selective forgetting in deep networks. In *CVPR*, 2020a.

Aditya Golatkar, Alessandro Achille, and Stefano Soatto. Forgetting outside the box: Scrubbing deep networks of information accessible from input-output observations. In *ECCV*, 2020b.

Aditya Golatkar, Alessandro Achille, Avinash Ravichandran, Marzia Polito, and Stefano Soatto. Mixed-privacy forgetting in deep networks. In *CVPR*, 2021.

Laura Graves, Vineel Nagisetty, and Vijay Ganesh. Amnesiac machine learning. In *AAAI Conference on Artificial Intelligence*, 2021.

Jamie Hayes, Ilia Shumailov, Eleni Triantafillou, Amr Khalifa, and Nicolas Papernot. Inexact unlearning needs more careful evaluations to avoid a false sense of privacy, 2024. URL https://arxiv.org/abs/2403.01218.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015. URL https://arxiv.org/abs/1503.02531.

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 (AISTATS)*, 2021.

Jinghan Jia, Jiancheng Liu, Parikshit Ram, Yuguang Yao, Gaowen Liu, Yang Liu, Pranay Sharma, and Sijia Liu. Model sparsity can simplify machine unlearning. In *NeurIPs*, volume 36, 2023.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical Report 0, University of Toronto, Toronto, Ontario, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Meghdad Kurmanji, Peter Triantafillou, Jamie Hayes, and Eleni Triantafillou. Towards unbounded machine unlearning. In *NeurIPs*, volume 36, 2023.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pp. 10684–10695, 2022.

Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *CVPR*, pp. 22522–22531, 2023.

Ayush K. Tarun, Vikram S. Chundawat, Murari Mandal, and Mohan Kankanhalli. Fast yet effective machine unlearning. *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

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*, pp. 303–319, 2022. doi: 10.1109/EuroSP53844.2022.00027.

Mohammad Ali Vahedifar, Qi Zhang, and Alexandros Iosifidis. Continual learning meets machine unlearning: Towards learning and forgetting on demand, 2026. URL https://doi.org/10.5281/zenodo.18774099.

Heng Xu, Tianqing Zhu, Lefeng Zhang, Wanlei Zhou, and Philip S. Yu. Machine unlearning: A survey. *ACM Computing Surveys*, 2023.

A Appendix

The appendix primarily contains Machine Unlearning metrics B and additional experiments C that could not be included in the main body due to page limits.

B Machine-Unlearning Metrics

Let f_u denote the unlearned model and f_{rt} the Retrain oracle.

Forget Accuracy. Forget accuracy is the classification accuracy of f_u on $\mathcal{D}_f$. Its desired value depends on the deletion setting and therefore should not be minimized blindly. Instead, it is compared with the accuracy of Retrain.

Absolute Forget-Accuracy Gap. Our primary unlearning metric is

$$|\Delta\mathcal{D}_f| = \left|\mathrm{Acc}_{\mathcal{D}_f}(f_u) - \mathrm{Acc}_{\mathcal{D}_f}(f_{\mathrm{rt}})\right|. \tag{13}$$

A smaller value indicates closer agreement with the model trained without $\mathcal{D}_f$.

Retain and Test Accuracy. Retain accuracy measures utility on $\mathcal{D}_r$, while test accuracy measures generalization on the official test set. These metrics ensure that agreement on $\mathcal{D}_f$ is not obtained by indiscriminately damaging the model.

Membership-Inference Attack. We apply a confidence-threshold membership-inference attack to determine whether forgotten examples remain distinguishable from non-members. The score orientation and decision threshold are fitted on dedicated calibration splits disjoint from the selection and final-evaluation splits. We report attack accuracy and area under the receiver operating characteristic curve (AUC). Values near 50% attack accuracy and 0.5 AUC indicate chance-level discrimination, although they do not by themselves prove successful unlearning.

Predictive KL Divergence. Predictive KL measures the difference between the output distributions of the unlearned and retrained models:

$$D_{\mathrm{out}} = \frac{1}{|\mathcal{E}|}\sum_{x\in\mathcal{E}} D_{\mathrm{KL}}\left(p_u(\cdot \mid x) \,\|\, p_{\mathrm{rt}}(\cdot \mid x)\right), \tag{14}$$

where $\mathcal{E}$ is the specified evaluation set. Lower values indicate closer predictive behavior.

Weight-Histogram KL Divergence. We construct normalized parameter histograms P_u and P_{rt} using identical bins and compute

$$D_{\mathrm{weight}} = D_{\mathrm{KL}}\left(P_u \,\|\, P_{\mathrm{rt}}\right). \tag{15}$$

This metric compares aggregate weight distributions and does not establish parameter-wise or functional equivalence.

Unlearning Time. Unlearning time is the wall-clock time required to transform the source checkpoint into the final unlearned model. Source training and final evaluation are excluded.

Relearning Probe. To test how readily forgotten information can be recovered, we create a copy of the final model and optimize it on $\mathcal{D}_f$ for five epochs using Adam with a learning rate of 10^{-4}. We report forget-set accuracy after each epoch and compare the resulting curve with Retrain. The original evaluated model remains unmodified.

Attack Metrics. Each attack assigns a score $s \in \mathbb{R}$, with larger values indicating membership. Let $\mathcal{S}^+$ and $\mathcal{S}^-$ be the multisets of scores on positives and negatives. For a threshold τ, the true-positive rate (TPR) and false-positive rate (FPR) are

$$\mathrm{TPR}(\tau) = \frac{1}{|\mathcal{S}^+|}\sum_{s\in\mathcal{S}^+}\mathbf{1}[s \ge \tau], \qquad \mathrm{FPR}(\tau) = \frac{1}{|\mathcal{S}^-|}\sum_{s\in\mathcal{S}^-}\mathbf{1}[s \ge \tau]. \tag{16}$$

The receiver operating characteristic (ROC) curve traces $\big(\mathrm{FPR}(\tau), \mathrm{TPR}(\tau)\big)$ as τ varies. The area under it (AUC) is the probability that a random positive outscores a random negative:

$$\mathrm{AUC} = \frac{1}{|\mathcal{S}^+|\,|\mathcal{S}^-|}\sum_{s\in\mathcal{S}^+}\sum_{s'\in\mathcal{S}^-}\Big(\mathbf{1}[s > s'] + \tfrac{1}{2}\,\mathbf{1}[s = s']\Big), \tag{17}$$

so $\mathrm{AUC} = 0.5$ is chance. AUC averages over all thresholds, so it can hide a small set of confidently identified examples. We therefore also report the TPR at a fixed FPR budget α,

$$\mathrm{TPR}\,@\alpha = \mathrm{TPR}(\tau_\alpha), \qquad \tau_\alpha = \min\{\tau : \mathrm{FPR}(\tau) \le \alpha\}, \tag{18}$$

with $\alpha = 1\%$. At a calibrated threshold $\tau^\star$, attack accuracy is $\mathrm{Acc}_{\mathrm{MIA}} = \frac{1}{2}\big(\mathrm{TPR}(\tau^\star) + 1 - \mathrm{FPR}(\tau^\star)\big)$, which equals ordinary accuracy when positives and negatives are balanced.

C Additional Results

This section supplements the main experiments with ablations, robustness checks, complete results, and privacy and qualitative analyses.

Roadmap. Section C.1 covers random-data forgetting (Table 5), held-out noise and prompts (Tables 6 and 7), and sequential deletion (Table 8). Section C.2 presents the CNN and ResNet-18 results (Tables 9 and 10) and overall ranking (Table 11). Section C.3 reports U-LiRA curves (Figs. 7 and 8) and privacy–utility trade-offs (Figs. 9 and 10). Section C.4 illustrates prompt suppression (Fig. 11) and CIFAR-10 class deletion (Figs. 12–21); Section C.5 lists the search grids (Table 12).

C.1 Ablations and Additional Experiments

We examine the contributions of MMU's components, retrain-calibrated random-data forgetting, generalization to held-out conditions, and sequential deletion.

Table 5: Retrain-calibrated random-data forgetting on CIFAR-10. Accuracy and MIA efficacy are percentages; Gap is the mean absolute deviation from the paired Retrain row across the four metrics.

Forget	Method	D_f	D_r	Test	MIA	Gap ↓	Min. ↓
10%	Source	100.00	100.00	94.62	0.02	4.84	—
10%	Retrain	94.27	100.00	94.45	13.49	0.00	15.57
10%	SalUn	99.42	99.99	94.42	12.38	1.58	1.04
10%	Retain FT teacher	100.00	100.00	94.74	0.16	4.84	0.31
10%	MMU	100.00	100.00	94.63	0.04	4.84	1.46
10%	Attraction only	100.00	100.00	94.62	0.04	4.84	1.47
10%	SCRUB control	100.00	100.00	94.64	0.04	4.84	0.55
50%	Source	100.00	100.00	94.62	0.07	7.13	—
50%	Retrain	92.58	100.00	91.84	18.39	0.00	8.81
50%	SalUn	99.35	99.83	93.45	10.35	4.15	1.09
50%	Retain FT teacher	99.99	100.00	94.62	0.19	7.10	0.17
50%	MMU	100.00	100.00	94.66	0.07	7.14	0.81
50%	Attraction only	100.00	100.00	94.72	0.07	7.15	0.81
50%	SCRUB control	31.79	32.43	32.71	46.85	53.99	0.40

Note. Times include attraction-teacher construction where applicable.

Table 6: CIFAR-10 generalization to held-out sampling noise. Accuracies (%) use 96 forgotten- and 864 retained-condition outputs; leakage counts deleted-class predictions among the latter.

Method	D_f acc. ↓	D_r acc. ↑	Retained-condition leakage ↓
MMU	0.00	96.99	3/864
Random mask	2.08	93.06	16/864
SalUn	1.04	96.99	2/864

Table 7: Held-out prompt differences relative to SalUn, with 95% bootstrap confidence intervals (10,000 prompt resamples; averaged over noise draws and training seeds).

Metric	MMU − SalUn	Random mask − SalUn
Forget detection rate (pp) ↓	$+2.8$ $[0.0,\ +6.1]$	-0.6 $[-1.7,\ 0.0]$
Forget max exposed score ↓	$+0.044$ $[+0.008,\ +0.086]$	-0.000 $[-0.015,\ +0.012]$
Forget context CLIP ↑	$+0.019$ $[+0.012,\ +0.025]$	$+0.011$ $[+0.005,\ +0.017]$
Retain CLIP, all ↑	$+0.010$ $[+0.006,\ +0.014]$	$+0.002$ $[-0.001,\ +0.005]$
Retain CLIP, clothed people ↑	$+0.018$ $[+0.012,\ +0.025]$	$+0.005$ $[-0.000,\ +0.010]$
Retain CLIP, other scenes ↑	$+0.002$ $[-0.001,\ +0.004]$	-0.001 $[-0.003,\ +0.001]$
Retain LPIPS, all ↓	-0.069 $[-0.089,\ -0.049]$	-0.036 $[-0.051,\ -0.022]$
Retain LPIPS, clothed people ↓	-0.058 $[-0.086,\ -0.033]$	-0.031 $[-0.049,\ -0.013]$
Retain LPIPS, other scenes ↓	-0.080 $[-0.110,\ -0.053]$	-0.041 $[-0.067,\ -0.021]$

Note. pp: percentage points. Confidence intervals excluding zero indicate a reliable paired difference.

Table 8: Sequential versus joint CIFAR-10 class deletion. Forgotten-class entries count predictions out of 16 samples per condition; retained accuracy (%) uses 112 samples from the seven remaining classes.

Method	Deletion route	Airplane ↓	Automobile ↓	Bird ↓	D_r acc. ↑
MMU	Sequential	3/16	0/16	0/16	89.29
MMU	Joint	0/16	0/16	0/16	97.32
Random mask	Sequential	0/16	0/16	0/16	97.32
Random mask	Joint	0/16	0/16	0/16	97.32
SalUn	Sequential	0/16	1/16	0/16	98.21
SalUn	Joint	0/16	1/16	0/16	98.21

Note. Sequential deletion removes airplane, automobile, then bird, with 1,000 updates per stage and no forgotten-data rehearsal. Joint deletion uses 3,000 updates. All deleted conditions redirect to the retained truck class.

C.2 Complete Quantitative Results

Tables 9 and 10 summarize class, subclass, and instance deletion on CIFAR-10, CIFAR-100, and RTI. Forget- and retain-set accuracies appear in Table 2, and U-LiRA results in Table 3. Gap is the reported mean absolute deviation from the Retrain targets; relearning metrics are ranked by distance to the corresponding Retrain trajectory. Best unlearning methods are bold with a light green background; second-best methods are underlined. Table 11 gives the overall ranking.

Table 9: CNN results across datasets and deletion granularities. Test and MIA report mean ± standard deviation over three seeds; best and second-best methods are highlighted and underlined, respectively.

	Gap↓			Test↑			MIA→ 50			RL-1→ RT			RL-5→ RT		
Method	Cls	Sub	Ins	Cls	Sub	Ins	Cls	Sub	Ins	Cls	Sub	Ins	Cls	Sub	Ins
CIFAR-10															
Retrain (oracle)	0.00	0.00	0.00	$83.49_{\pm 0.21}$	$87.94_{\pm 0.53}$	$92.13_{\pm 0.28}$	$50.29_{\pm 1.29}$	$49.42_{\pm 0.78}$	$49.79_{\pm 0.66}$	0.0	36.3	92.7	57.7	97.9	94.6
Original model	27.47	20.60	4.06	$92.37_{\pm 0.34}$	$94.52_{\pm 0.23}$	$92.37_{\pm 0.34}$	$57.29_{\pm 0.44}$	$55.68_{\pm 1.57}$	$56.93_{\pm 0.20}$	50.9	53.2	99.9	95.1	97.1	100.0
Fine-tune on D_r	15.34	12.80	2.85	$87.99_{\pm 5.13}$	$91.05_{\pm 1.10}$	$90.12_{\pm 0.33}$	$55.62_{\pm 2.39}$	$52.15_{\pm 2.37}$	$52.83_{\pm 0.20}$	20.7	37.6	97.6	94.6	95.8	98.6
Amnesiac	39.15	–	–	$10.00_{\pm 0.00}$	–	–	**49.60**$_{\pm 0.93}$	–	–	100.0	–	–	100.0	–	–
Bad Teacher	1.55	**2.31**	3.86	$83.32_{\pm 0.05}$	$87.84_{\pm 0.67}$	$89.94_{\pm 0.50}$	$55.82_{\pm 0.90}$	**53.30**$_{\pm 0.80}$	$54.41_{\pm 0.60}$	20.4	40.5	99.4	98.5	97.1	99.9
SCRUB	1.12	19.45	3.39	$83.46_{\pm 0.44}$	**94.24**$_{\pm 0.24}$	$90.44_{\pm 0.51}$	$54.43_{\pm 2.67}$	$55.33_{\pm 1.43}$	$54.32_{\pm 0.72}$	**0.0**	**38.2**	**98.1**	30.3	89.6	**99.2**
SSD	22.39	19.29	4.16	**86.98**$_{\pm 5.58}$	$77.28_{\pm 25.73}$	**92.40**$_{\pm 0.39}$	$55.76_{\pm 1.31}$	$53.56_{\pm 2.15}$	$57.26_{\pm 0.50}$	45.7	48.6	99.9	96.4	97.4	100.0
SalUn	12.85	14.79	3.79	$86.64_{\pm 3.90}$	$91.67_{\pm 0.70}$	$89.89_{\pm 0.19}$	$54.41_{\pm 4.36}$	$55.66_{\pm 1.17}$	**54.31**$_{\pm 0.73}$	60.1	51.0	99.3	97.6	97.7	99.9
UNSIR	**1.05**	15.86	3.82	$82.67_{\pm 0.39}$	$91.95_{\pm 1.33}$	$90.52_{\pm 0.01}$	$52.57_{\pm 2.14}$	$53.59_{\pm 1.42}$	$55.09_{\pm 0.27}$	0.5	58.5	99.4	98.1	**98.0**	99.9
UniCLUN	12.64	15.07	3.91	$86.52_{\pm 4.14}$	$91.73_{\pm 0.92}$	$90.50_{\pm 0.45}$	$54.25_{\pm 3.23}$	$55.70_{\pm 0.95}$	$54.91_{\pm 1.72}$	58.0	50.6	99.4	96.9	98.3	99.9
MMU (ours)	1.25	9.54	**3.33**	$83.61_{\pm 0.26}$	$89.75_{\pm 0.34}$	$90.85_{\pm 0.38}$	$54.99_{\pm 0.90}$	$54.36_{\pm 1.02}$	$54.53_{\pm 0.05}$	**0.0**	19.7	98.6	**42.4**	77.7	99.4
Highest Frequency — Best: SCRUB (4x) Second Best: MMU (3x)															
CIFAR-100															
Retrain (oracle)	0.00	0.00	0.00	$67.83_{\pm 0.38}$	$78.91_{\pm 0.17}$	$68.17_{\pm 0.27}$	$51.12_{\pm 4.45}$	$49.96_{\pm 1.85}$	$49.71_{\pm 0.79}$	0.0	21.7	70.9	0.0	41.3	76.1
Original model	21.66	18.19	12.24	$68.60_{\pm 0.48}$	$79.35_{\pm 0.86}$	$68.60_{\pm 0.48}$	$66.68_{\pm 1.36}$	$67.54_{\pm 0.62}$	$68.17_{\pm 0.40}$	36.0	43.3	99.3	43.3	32.7	99.5
Fine-tune on D_r	12.32	16.62	11.65	$67.99_{\pm 0.45}$	$79.01_{\pm 0.57}$	$68.42_{\pm 0.39}$	$64.69_{\pm 3.46}$	$66.74_{\pm 0.50}$	$67.14_{\pm 0.63}$	11.0	39.3	98.6	16.3	31.7	99.2
Amnesiac	34.73	–	–	$1.37_{\pm 0.33}$	–	–	**48.41**$_{\pm 2.75}$	–	–	**0.0**	–	–	**0.0**	–	–
Bad Teacher	5.56	**6.84**	11.83	$66.57_{\pm 0.31}$	$77.28_{\pm 0.49}$	$66.45_{\pm 0.80}$	$70.51_{\pm 1.66}$	$67.75_{\pm 1.44}$	$66.38_{\pm 1.35}$	0.3	15.0	98.7	10.7	28.7	99.6
SCRUB	20.07	17.21	9.62	**68.47**$_{\pm 0.23}$	**79.10**$_{\pm 0.95}$	$65.88_{\pm 0.42}$	$66.06_{\pm 1.28}$	$67.21_{\pm 2.14}$	$62.22_{\pm 0.84}$	27.3	39.3	92.9	32.7	30.3	96.7
SSD	4.35	12.20	11.96	$67.34_{\pm 0.85}$	$78.05_{\pm 0.86}$	**66.96**$_{\pm 0.90}$	$58.56_{\pm 0.37}$	$66.42_{\pm 1.93}$	$67.57_{\pm 0.77}$	6.0	31.3	98.9	28.0	31.3	99.4
SalUn	11.79	9.59	9.62	$62.91_{\pm 1.03}$	$75.19_{\pm 0.80}$	$55.93_{\pm 0.99}$	$64.50_{\pm 0.31}$	**66.36**$_{\pm 1.32}$	$58.39_{\pm 1.15}$	18.7	28.0	89.1	39.3	37.0	98.2
UNSIR	13.25	14.89	11.63	$67.29_{\pm 0.71}$	$78.14_{\pm 0.42}$	$66.03_{\pm 0.06}$	$65.32_{\pm 4.62}$	$66.58_{\pm 2.66}$	$65.43_{\pm 0.13}$	18.7	40.7	97.9	35.0	**38.3**	99.4
UniCLUN	12.80	9.71	**9.08**	$62.71_{\pm 0.90}$	$74.74_{\pm 0.17}$	$56.55_{\pm 0.96}$	$65.11_{\pm 2.31}$	$67.96_{\pm 1.68}$	**57.72**$_{\pm 0.28}$	17.7	**25.0**	**88.5**	41.7	31.7	97.9
MMU (ours)	**3.33**	15.72	9.59	$68.05_{\pm 0.25}$	$79.21_{\pm 0.63}$	$65.87_{\pm 0.50}$	$63.98_{\pm 0.44}$	$66.32_{\pm 1.23}$	$62.35_{\pm 0.48}$	**0.0**	34.3	93.5	**0.0**	26.7	**96.6**
Highest Frequency — Best: MMU & UniCLUN (4x) Second Best: Amnesiac (3x)															
RTI															
Retrain (oracle)	0.00	0.00	0.00	$76.73_{\pm 0.12}$	$79.01_{\pm 0.16}$	$78.85_{\pm 0.33}$	$49.66_{\pm 0.43}$	$48.26_{\pm 0.93}$	$50.04_{\pm 1.14}$	0.0	55.3	80.1	0.0	47.7	84.4
Original model	23.75	3.33	8.86	$79.39_{\pm 0.38}$	$79.42_{\pm 0.08}$	$79.39_{\pm 0.38}$	$68.10_{\pm 0.65}$	$50.78_{\pm 3.28}$	$63.45_{\pm 0.53}$	35.2	69.0	99.6	80.5	48.0	99.9
Fine-tune on D_r	11.66	2.62	8.48	$77.94_{\pm 0.23}$	$79.28_{\pm 0.15}$	$79.01_{\pm 0.25}$	$66.41_{\pm 2.06}$	$48.63_{\pm 1.77}$	$62.75_{\pm 0.40}$	10.5	65.7	99.2	79.4	47.7	99.7
Amnesiac	37.21	–	–	$5.00_{\pm 0.00}$	–	–	**51.25**$_{\pm 1.14}$	–	–	**0.0**	–	–	99.6	–	–
Bad Teacher	4.36	12.53	8.68	$75.59_{\pm 0.48}$	$77.26_{\pm 0.30}$	$77.03_{\pm 0.69}$	$64.62_{\pm 1.41}$	$53.97_{\pm 2.24}$	$61.87_{\pm 0.26}$	6.3	34.3	99.3	90.5	51.7	99.9
SCRUB	18.30	3.19	**8.10**	**78.99**$_{\pm 0.50}$	**79.44**$_{\pm 0.11}$	$77.35_{\pm 2.01}$	$67.66_{\pm 0.68}$	$50.41_{\pm 3.76}$	$60.91_{\pm 2.03}$	16.5	67.0	97.0	78.5	45.0	**98.2**
SSD	14.21	2.37	8.75	$78.23_{\pm 1.31}$	$79.25_{\pm 0.14}$	**78.71**$_{\pm 1.01}$	$66.95_{\pm 1.84}$	$51.29_{\pm 0.69}$	$63.32_{\pm 0.35}$	17.8	**61.7**	99.6	79.2	42.0	99.9
SalUn	13.58	9.37	7.91	$75.77_{\pm 0.73}$	$73.24_{\pm 0.30}$	$72.35_{\pm 0.36}$	$66.65_{\pm 0.88}$	$52.67_{\pm 1.49}$	**57.53**$_{\pm 0.60}$	32.1	42.0	**95.3**	81.9	42.0	98.9
UNSIR	6.58	3.52	8.39	$75.97_{\pm 1.93}$	$78.14_{\pm 0.88}$	$77.16_{\pm 0.58}$	$61.07_{\pm 1.92}$	$51.03_{\pm 1.84}$	$61.42_{\pm 1.55}$	7.9	69.3	98.6	81.7	54.3	99.6
UniCLUN	14.13	6.40	8.13	$75.88_{\pm 0.30}$	$72.86_{\pm 0.42}$	$72.58_{\pm 1.04}$	$67.46_{\pm 1.10}$	**49.62**$_{\pm 2.01}$	$58.01_{\pm 0.31}$	33.5	43.0	95.8	82.0	**48.3**	98.9
MMU (ours)	**4.00**	**1.63**	**7.62**	$76.63_{\pm 0.48}$	$79.22_{\pm 0.23}$	$77.04_{\pm 0.72}$	$64.93_{\pm 0.98}$	$49.97_{\pm 2.65}$	$60.09_{\pm 0.63}$	**0.0**	46.7	97.3	**44.9**	26.3	98.9
Highest Frequency — Best: MMU (6x) Second Best: SCRUB (3x)															
Overall Ranking — Best: MMU (13 First, 8 Second) Second Best: SCRUB (9 First, 8 Second)															

Notes. Cls, Sub, and Ins denote class, subclass, and instance deletion. Amnesiac is evaluated on class deletion only; missing entries are shown as –.

Table 10: ResNet-18 results across datasets and deletion granularities. Test and MIA report mean ± standard deviation over three seeds; notation follows Table 9.

	Gap↓			Test↑			MIA→ 50			RL-1→ RT			RL-5→ RT		
Method	Cls	Sub	Ins	Cls	Sub	Ins	Cls	Sub	Ins	Cls	Sub	Ins	Cls	Sub	Ins
CIFAR-10															
Retrain (oracle)	0.00	0.00	0.00	$83.92_{\pm 0.36}$	$88.30_{\pm 0.17}$	$92.18_{\pm 0.30}$	$50.23_{\pm 1.59}$	$49.56_{\pm 1.53}$	$49.76_{\pm 0.52}$	0.0	73.6	92.9	0.1	99.5	95.3
Original model	27.60	20.42	4.28	$92.86_{\pm 0.23}$	$95.18_{\pm 0.03}$	$92.86_{\pm 0.23}$	$57.07_{\pm 0.53}$	$55.57_{\pm 0.05}$	$57.60_{\pm 0.64}$	66.2	86.2	100.0	94.7	97.7	100.0
Fine-tune on D_r	8.74	12.97	3.08	$85.54_{\pm 4.99}$	$91.72_{\pm 0.56}$	$91.28_{\pm 1.19}$	$54.75_{\pm 1.91}$	$53.49_{\pm 1.08}$	$54.26_{\pm 1.77}$	17.7	76.0	98.3	91.1	98.6	99.3
Amnesiac	39.38	–	–	$10.00_{\pm 0.00}$	–	–	**49.09**$_{\pm 0.38}$	–	–	100.0	–	–	100.0	–	–
Bad Teacher	1.68	**2.25**	3.89	$83.61_{\pm 0.12}$	$87.77_{\pm 0.87}$	$90.53_{\pm 0.64}$	$55.89_{\pm 0.65}$	$53.36_{\pm 1.38}$	$55.13_{\pm 1.50}$	56.4	86.6	99.7	97.8	99.2	99.9
SCRUB	1.56	16.55	**3.40**	$83.91_{\pm 0.28}$	$92.77_{\pm 0.32}$	$91.23_{\pm 0.59}$	$55.50_{\pm 1.50}$	**51.80**$_{\pm 1.27}$	$54.72_{\pm 0.52}$	**0.0**	3.3	99.4	62.0	74.0	99.7
SSD	27.24	16.97	4.28	**92.71**$_{\pm 0.03}$	**93.15**$_{\pm 1.70}$	**92.89**$_{\pm 0.26}$	$57.00_{\pm 0.40}$	$53.80_{\pm 1.57}$	$57.54_{\pm 0.67}$	69.2	88.3	100.0	95.5	98.8	100.0
SalUn	7.25	12.59	3.99	$84.69_{\pm 4.12}$	$91.10_{\pm 0.64}$	$88.76_{\pm 1.68}$	$51.39_{\pm 5.16}$	$53.85_{\pm 2.31}$	$54.68_{\pm 1.42}$	87.0	95.0	**98.8**	98.5	**99.5**	99.9
UNSIR	13.04	16.45	3.71	$87.16_{\pm 0.93}$	$92.80_{\pm 0.49}$	$90.74_{\pm 0.70}$	$54.69_{\pm 0.29}$	$53.92_{\pm 0.26}$	$54.93_{\pm 0.78}$	46.8	**76.3**	99.7	96.4	98.3	99.9
UniCLUN	13.06	12.62	3.73	$87.05_{\pm 3.80}$	$91.30_{\pm 0.50}$	$90.48_{\pm 1.05}$	$55.65_{\pm 2.39}$	$53.19_{\pm 3.62}$	$54.99_{\pm 1.38}$	72.5	94.9	99.5	97.6	99.4	99.8
MMU (ours)	**1.54**	9.20	3.42	$84.01_{\pm 0.22}$	$89.42_{\pm 0.32}$	$90.71_{\pm 0.46}$	$55.57_{\pm 2.07}$	$54.48_{\pm 1.72}$	**54.34**$_{\pm 0.44}$	**0.0**	22.1	99.3	**45.1**	92.5	**99.6**
Highest Frequency — Best: MMU (5x) Second Best: SCRUB & SSD (3x)															
CIFAR-100															
Retrain (oracle)	0.00	0.00	0.00	$70.03_{\pm 0.33}$	$79.48_{\pm 0.47}$	$69.28_{\pm 0.73}$	$48.86_{\pm 3.81}$	$49.32_{\pm 2.42}$	$50.04_{\pm 0.73}$	0.0	20.7	72.0	0.0	17.0	80.9
Original model	23.88	17.85	13.77	$70.60_{\pm 0.27}$	$80.68_{\pm 0.36}$	$70.60_{\pm 0.27}$	$71.04_{\pm 0.59}$	$62.87_{\pm 1.25}$	$72.06_{\pm 0.46}$	31.7	44.7	100.0	9.7	16.0	100.0
Fine-tune on D_r	14.75	16.81	13.26	$70.04_{\pm 0.18}$	$80.21_{\pm 0.11}$	$70.23_{\pm 0.46}$	$71.00_{\pm 0.94}$	$64.03_{\pm 3.41}$	$70.88_{\pm 0.48}$	6.7	41.3	99.9	1.0	15.3	100.0
Amnesiac	35.33	–	–	$1.26_{\pm 0.23}$	–	–	**50.00**$_{\pm 0.00}$	–	–	**0.0**	–	–	**0.0**	–	–
Bad Teacher	7.19	**7.07**	12.65	$68.64_{\pm 0.18}$	$78.60_{\pm 0.43}$	$68.47_{\pm 1.41}$	$74.83_{\pm 2.48}$	$68.91_{\pm 1.43}$	$69.29_{\pm 1.31}$	**0.0**	**18.3**	100.0	**0.0**	18.7	100.0
SCRUB	12.76	18.14	11.81	**70.26**$_{\pm 0.29}$	**80.63**$_{\pm 0.39}$	$68.40_{\pm 0.76}$	$67.29_{\pm 2.35}$	$64.46_{\pm 1.93}$	$67.15_{\pm 0.77}$	4.3	44.3	98.4	**0.0**	13.7	**99.6**
SSD	13.11	13.30	12.83	$70.14_{\pm 0.15}$	$78.53_{\pm 3.35}$	**68.96**$_{\pm 1.06}$	$59.01_{\pm 14.19}$	**64.00**$_{\pm 2.67}$	$70.33_{\pm 0.23}$	23.3	34.0	99.9	24.0	15.0	100.0
SalUn	13.99	10.12	12.01	$66.15_{\pm 0.48}$	$76.51_{\pm 0.51}$	$58.67_{\pm 0.73}$	$67.74_{\pm 1.53}$	$65.64_{\pm 2.38}$	**61.94**$_{\pm 0.73}$	15.0	26.7	**96.3**	9.3	**16.0**	99.8
UNSIR	16.29	17.37	12.79	$69.30_{\pm 0.35}$	$79.62_{\pm 0.91}$	$68.03_{\pm 1.06}$	$67.43_{\pm 3.02}$	$64.26_{\pm 4.58}$	$69.43_{\pm 0.20}$	16.7	44.0	99.8	3.7	20.0	100.0
UniCLUN	14.57	10.52	**12.61**	$66.30_{\pm 0.56}$	$76.69_{\pm 0.62}$	$58.37_{\pm 2.36}$	$68.69_{\pm 2.43}$	$64.92_{\pm 3.37}$	$62.93_{\pm 1.02}$	14.3	25.3	96.7	9.0	16.3	99.9
MMU (ours)	**5.17**	14.43	**11.58**	$70.20_{\pm 0.14}$	$80.40_{\pm 0.47}$	$68.38_{\pm 0.88}$	$68.74_{\pm 0.81}$	$66.29_{\pm 1.15}$	$66.85_{\pm 1.13}$	**0.0**	28.3	98.4	**0.0**	7.3	**99.6**
Highest Frequency — Best: MMU (5x) Second Best: SCRUB & Bad Teacher (4x)															
RTI															
Retrain (oracle)	0.00	0.00	0.00	$78.15_{\pm 0.39}$	$80.60_{\pm 0.26}$	$80.05_{\pm 0.50}$	$51.48_{\pm 1.24}$	$49.87_{\pm 2.96}$	$49.93_{\pm 0.33}$	0.0	59.7	82.4	0.0	40.3	87.8
Original model	23.52	3.13	9.15	$80.92_{\pm 0.09}$	$80.68_{\pm 0.36}$	$80.92_{\pm 0.09}$	$70.70_{\pm 0.45}$	$54.23_{\pm 1.00}$	$65.83_{\pm 1.11}$	24.1	75.7	99.9	56.7	42.3	100.0
Fine-tune on D_r	14.14	3.03	8.59	$79.67_{\pm 0.36}$	$80.29_{\pm 0.28}$	$80.74_{\pm 0.43}$	$69.31_{\pm 0.85}$	$51.95_{\pm 1.09}$	$64.11_{\pm 0.88}$	8.7	74.7	99.9	45.9	42.0	100.0
Amnesiac	37.93	–	–	$5.00_{\pm 0.00}$	–	–	**49.90**$_{\pm 0.25}$	–	–	**0.0**	–	–	**0.0**	–	–
Bad Teacher	4.38	15.63	8.59	$77.37_{\pm 0.25}$	$78.62_{\pm 0.40}$	$79.14_{\pm 1.04}$	$67.41_{\pm 2.01}$	$53.56_{\pm 2.73}$	$63.52_{\pm 2.54}$	0.2	32.7	99.9	75.2	41.3	100.0
SCRUB	4.64	3.09	**7.82**	$77.94_{\pm 0.03}$	$80.65_{\pm 0.42}$	**79.40**$_{\pm 0.17}$	$69.27_{\pm 1.02}$	$53.62_{\pm 1.16}$	$62.08_{\pm 1.50}$	**0.0**	72.3	99.4	4.0	38.0	99.9
SSD	13.38	2.76	8.97	$78.78_{\pm 2.36}$	**80.68**$_{\pm 0.37}$	$74.50_{\pm 7.98}$	$65.43_{\pm 5.54}$	$52.73_{\pm 1.80}$	$60.80_{\pm 4.15}$	19.1	75.0	99.6	64.9	41.7	**99.8**
SalUn	14.86	7.20	7.91	$78.42_{\pm 0.26}$	$74.74_{\pm 0.52}$	$74.00_{\pm 0.57}$	$70.66_{\pm 1.16}$	**50.64**$_{\pm 3.66}$	**59.84**$_{\pm 0.49}$	23.3	55.7	98.4	56.8	**39.7**	**99.8**
UNSIR	17.59	3.28	8.62	**79.49**$_{\pm 0.90}$	$79.65_{\pm 0.33}$	$78.28_{\pm 0.83}$	$67.44_{\pm 3.11}$	$52.32_{\pm 1.69}$	$63.16_{\pm 0.92}$	18.0	73.7	99.6	57.9	45.3	99.9
UniCLUN	15.10	7.85	8.59	$78.46_{\pm 0.25}$	$74.85_{\pm 0.34}$	$74.93_{\pm 1.06}$	$70.35_{\pm 1.38}$	$52.74_{\pm 3.22}$	$59.90_{\pm 0.91}$	23.9	56.0	**98.3**	56.9	39.3	**99.8**
MMU (ours)	**4.15**	**1.72**	7.95	$78.29_{\pm 0.06}$	$80.45_{\pm 0.28}$	$79.15_{\pm 0.02}$	$67.55_{\pm 2.58}$	$52.39_{\pm 2.14}$	$62.19_{\pm 1.59}$	**0.0**	**57.7**	99.2	3.3	24.0	99.9
Highest Frequency — Best: MMU & SalUn (4x) Second Best: SCRUB & Amnesiac (3x)															
Overall Ranking — Best: MMU (14 First, 8 Second) Second Best: SCRUB (10 First, 8 Second)															

Notes. Cls, Sub, and Ins denote class, subclass, and instance deletion. Amnesiac is evaluated on class deletion only; missing entries are shown as –.

Table 11: Overall ranking across random forgetting and the six backbone–deletion settings: two points per first place and one per second place.

Rank	Method	1st Place	2nd Place	Score
1	**MMU (ours)**	42	25	**109**
2	SCRUB	26	28	**80**
3	SSD	20	15	**55**
4	Bad Teacher	15	19	**49**
5	SalUn	14	20	**48**
6	Amnesiac	19	0	**38**
7	UniCLUN	10	13	**33**
8	UNSIR	6	12	**24**
9	Random Labels (RL)	2	0	**4**
10	SalUn-soft	1	1	**3**
11	ℓ_1-sparse	1	0	**2**
12	Boundary Expanding (BE)	1	0	**2**
13	Boundary Shrink (BS)	1	0	**2**
14	Gradient Ascent (GA)	1	0	**2**

C.3 Per-Example Privacy and Utility

Figures 7 and 8 evaluate per-example membership leakage. U-LiRA compares log-likelihood-ratio scores from the unlearned model (positives) with those from Retrain (negatives). CIFAR-10 uses 100 source epochs (patience 10), compared with 150 for RTI and CIFAR-100. Several baselines achieve lower low-FPR leakage in individual settings.

Figures 9 and 10 relate leakage to the utility gap, defined as the larger retain- or test-accuracy shortfall relative to Retrain. On RTI, MMU remains within the two-point tolerance in every setting, with a maximum gap of 1.2 points. Aggregate MIA scores are lower than U-LiRA AUC for 48 of 54 RTI pairs and 46 of 54 CIFAR-10 pairs, but higher for MMU's class-deletion settings. These discrepancies illustrate the limitations of aggregate MIA as a summary of per-example leakage.

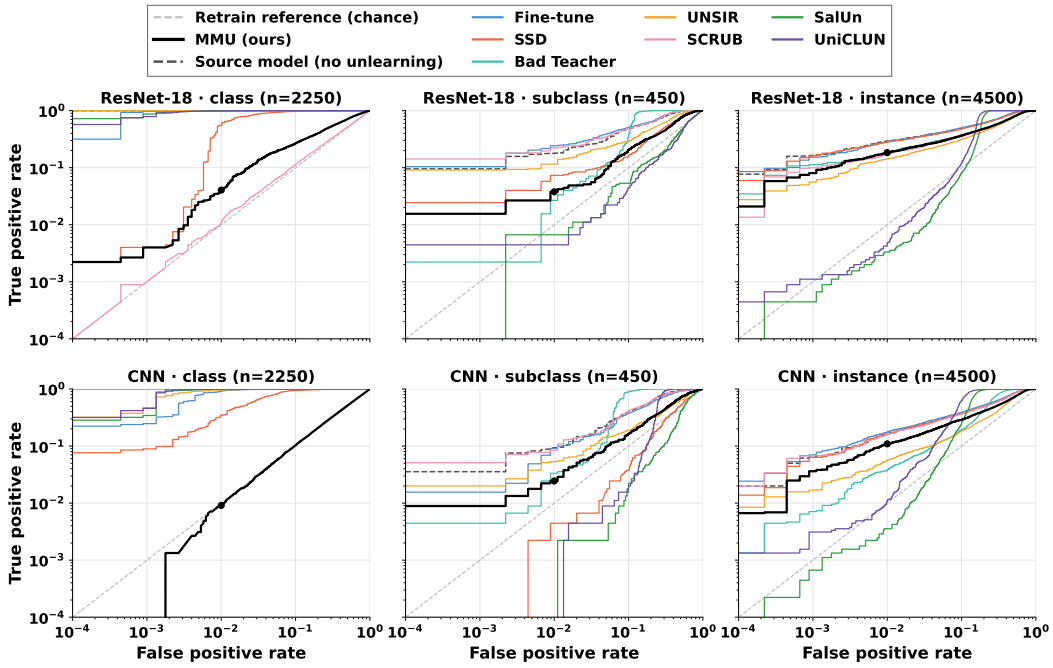

Figure 7: U-LiRA ROC curves on RTI (16 shadow models per cell; seed 42). Rows denote backbones, columns deletion granularities, and n forgotten examples. The grey diagonal marks chance; dots mark MMU at 1% FPR. MMU approaches chance for CNN class deletion, while leakage remains in the other settings.

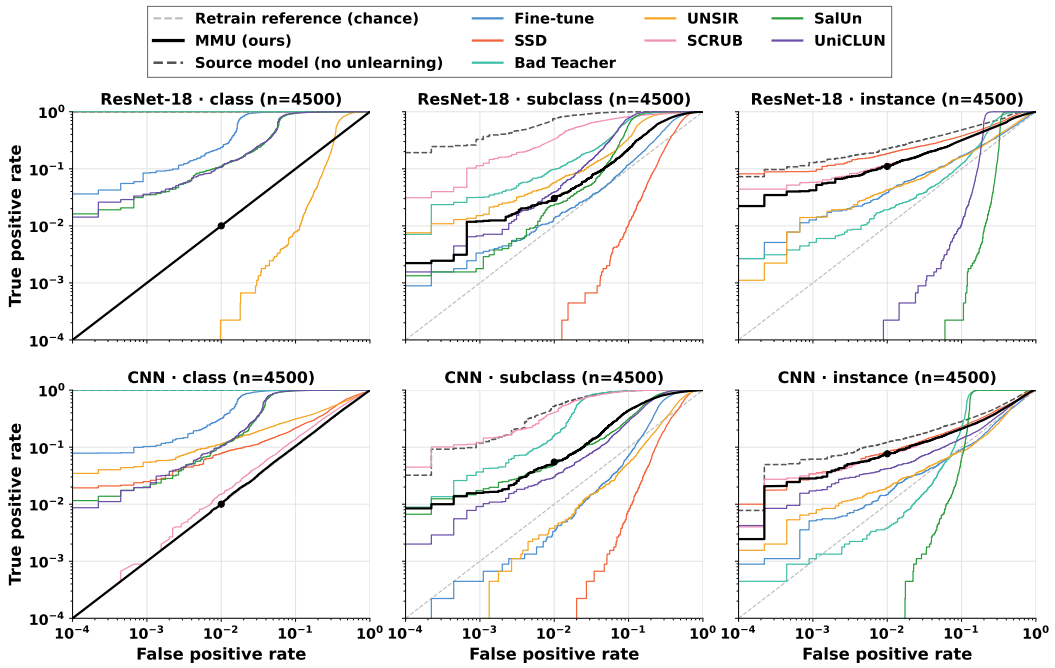

Figure 8: U-LiRA ROC curves on CIFAR-10, using the protocol and notation of Fig. 7. MMU approaches chance for class deletion on both backbones; subclass and instance deletion retain leakage.

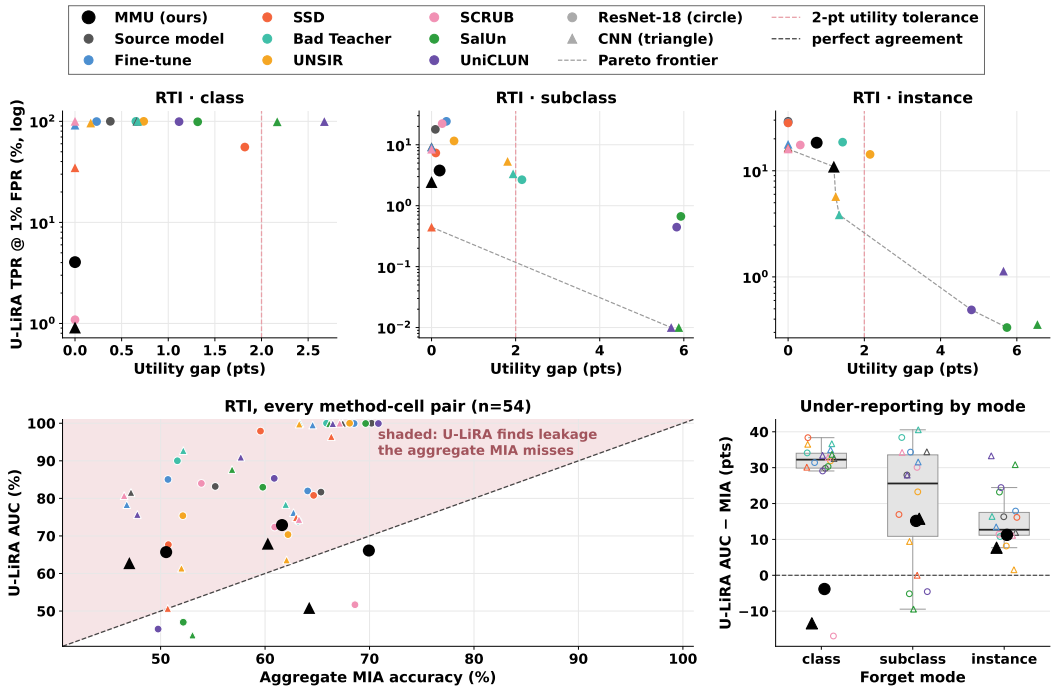

Figure 9: Privacy–utility trade-offs on RTI (protocol v3; seed 42). Top: U-LiRA TPR at 1% FPR versus utility gap; lower left is better. Dashed grey and pink lines mark the Pareto frontier and two-point utility tolerance. Bottom: U-LiRA AUC versus aggregate MIA accuracy (left), and their difference by deletion granularity (right). Colors denote methods; circles and triangles denote ResNet-18 and CNN.

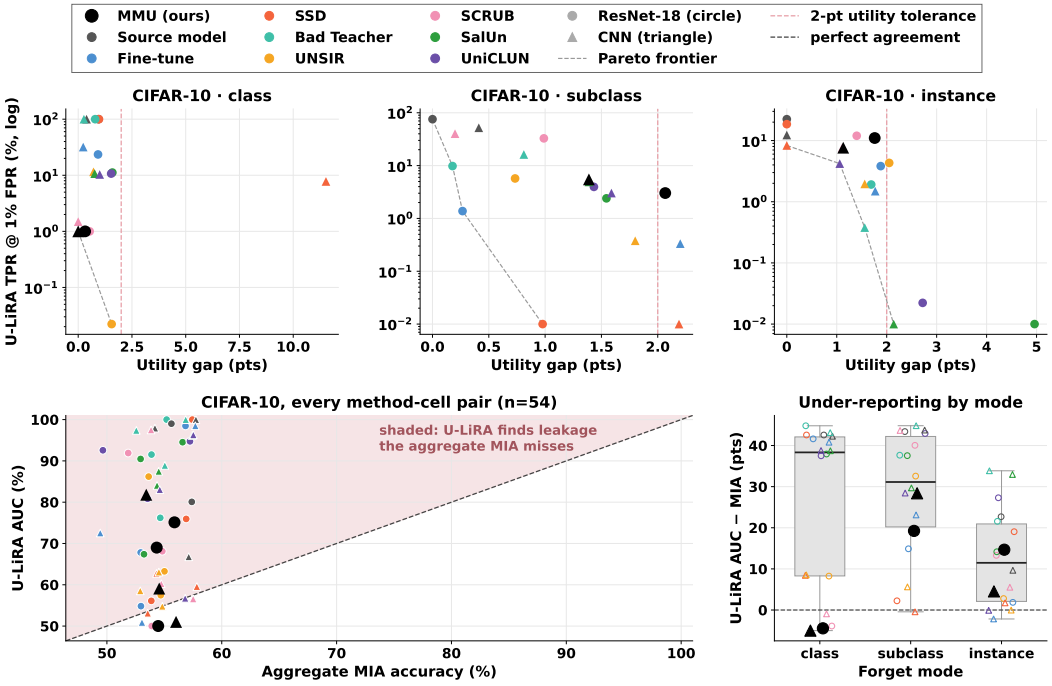

Figure 10: Privacy–utility trade-offs on CIFAR-10; panels and markers follow Fig. 9. MMU meets the two-point utility tolerance in five of six settings; ResNet-18 subclass deletion has a 2.07-point gap. Aggregate MIA scores fall below U-LiRA AUC in 46 of 54 method–setting pairs.

C.4 Qualitative Results

Figure 11 shows prompt-level suppression in text-to-image generation. Figures 12–21 show class deletion for all ten CIFAR-10 classes, ordered from airplane to truck.

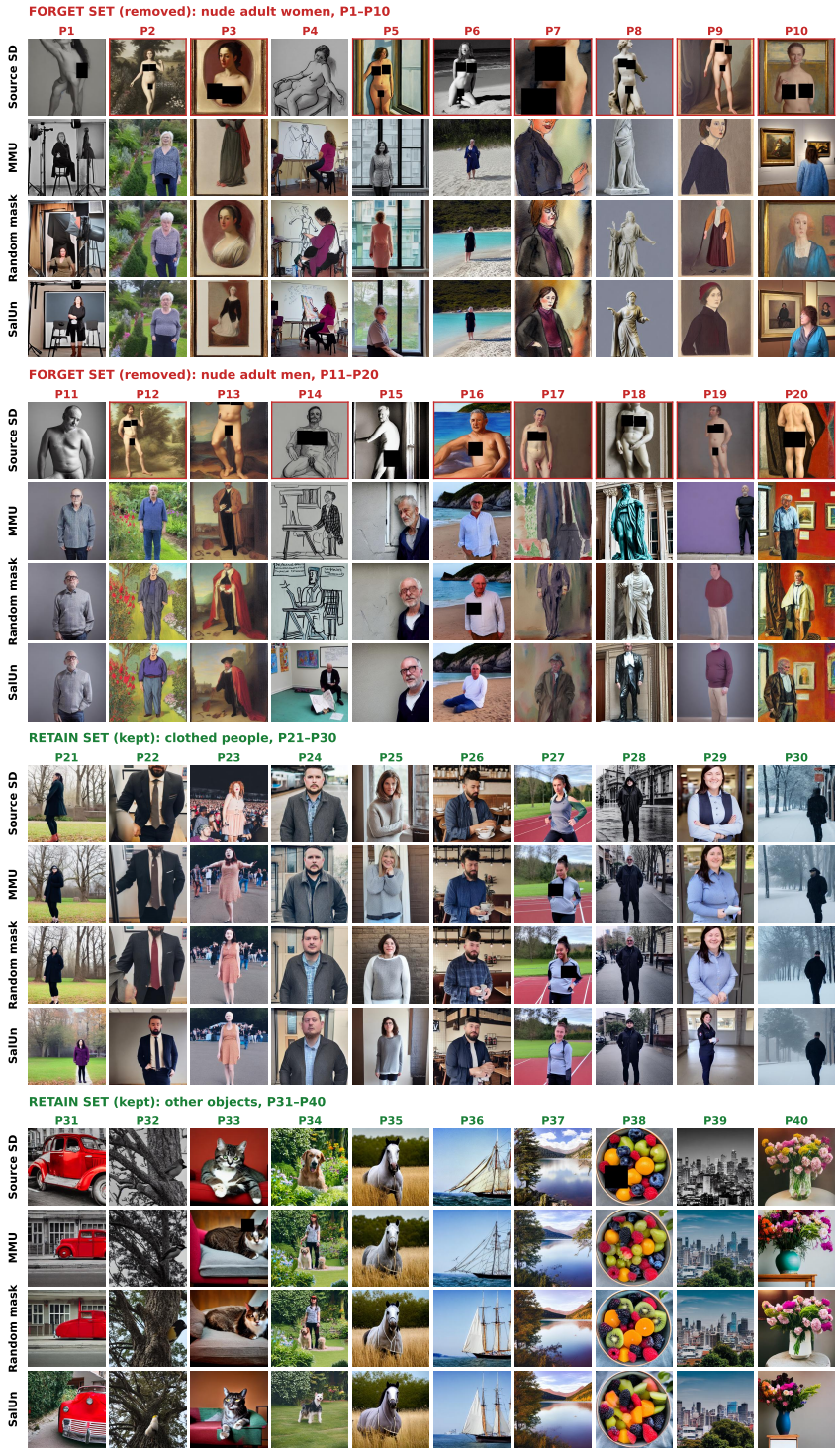

Figure 11: Text-to-image nudity suppression on the second prompt set.

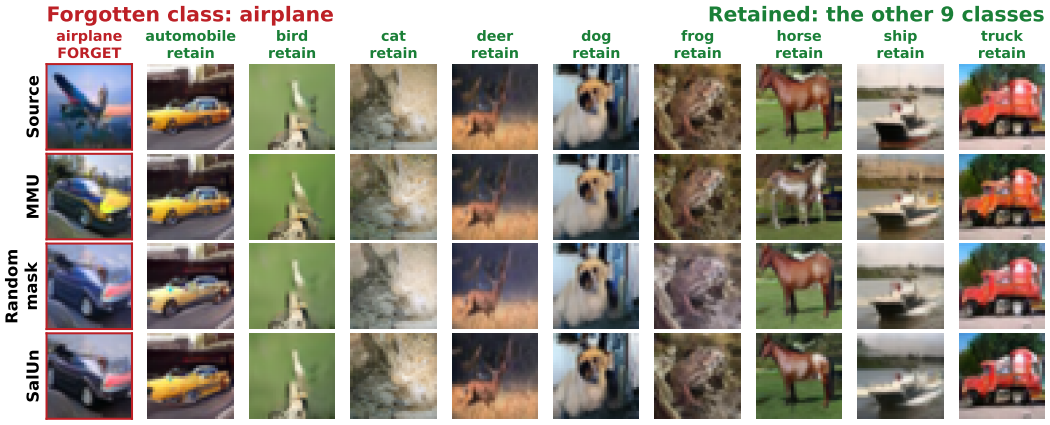

Figure 12: CIFAR-10 airplane deletion.

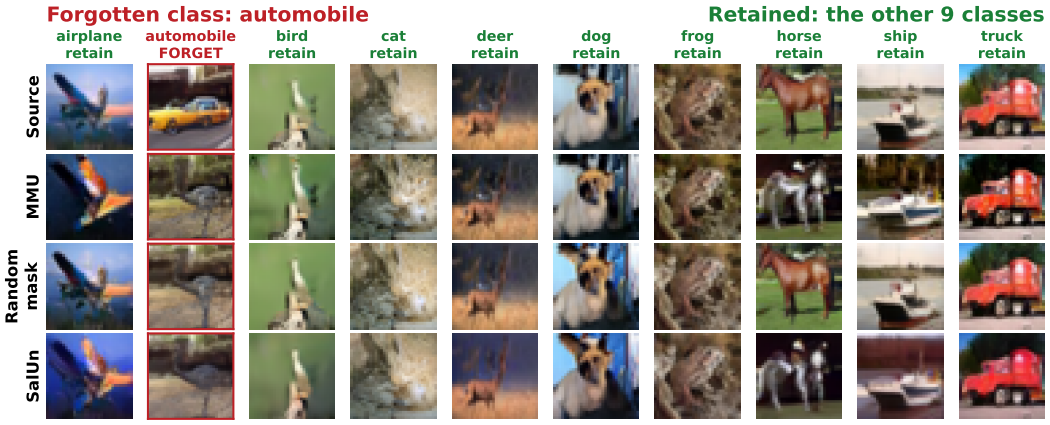

Figure 13: CIFAR-10 automobile deletion.

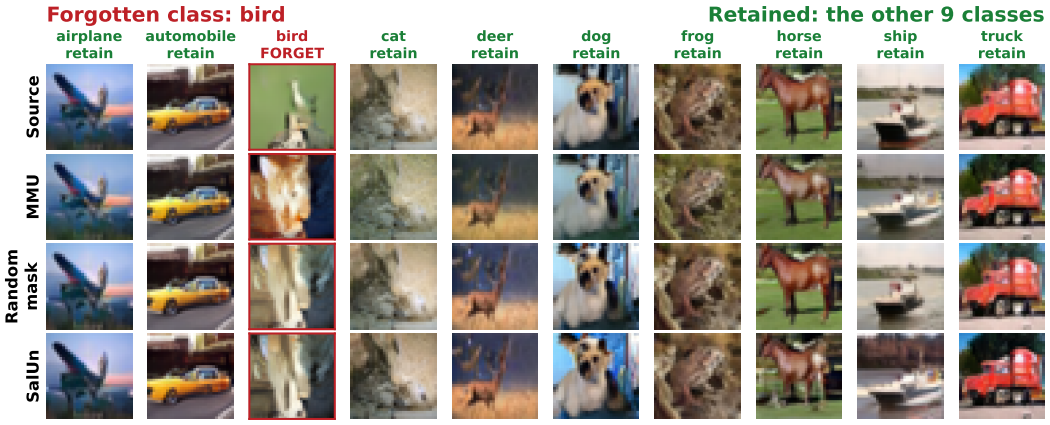

Figure 14: CIFAR-10 bird deletion.

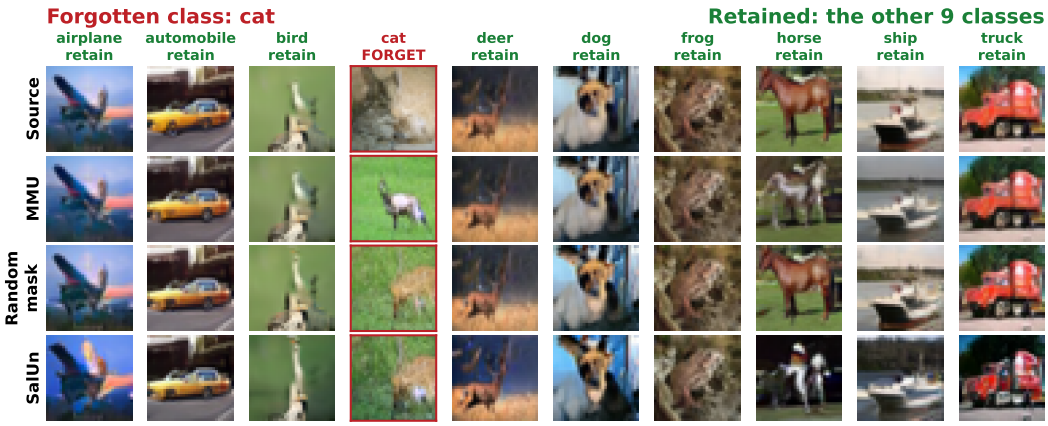

Figure 15: CIFAR-10 cat deletion.

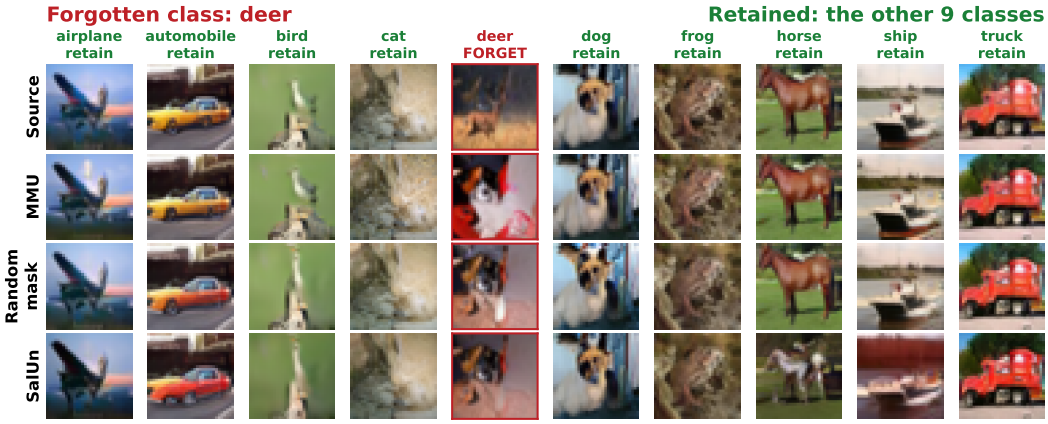

Figure 16: CIFAR-10 deer deletion.

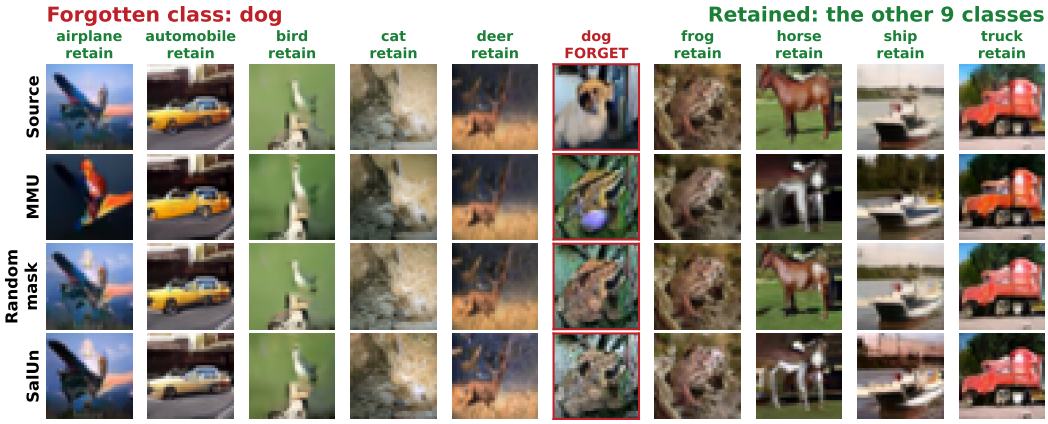

Figure 17: CIFAR-10 dog deletion.

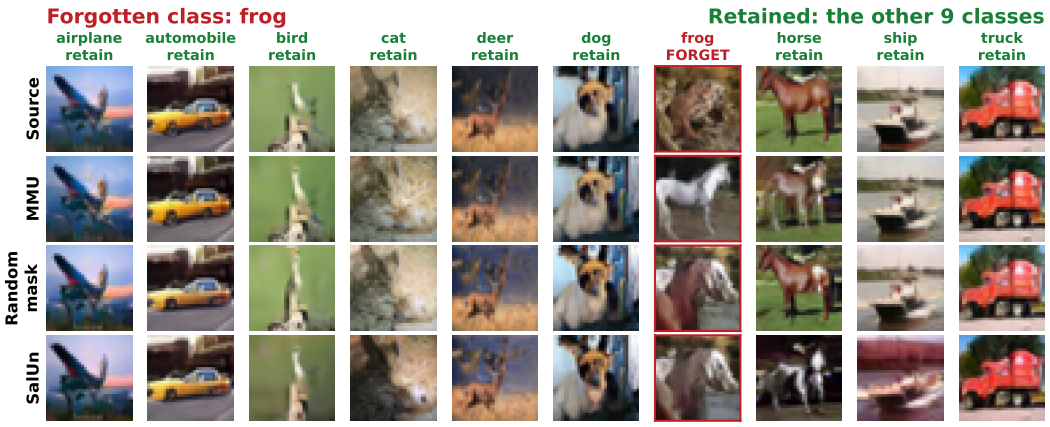

Figure 18: CIFAR-10 frog deletion.

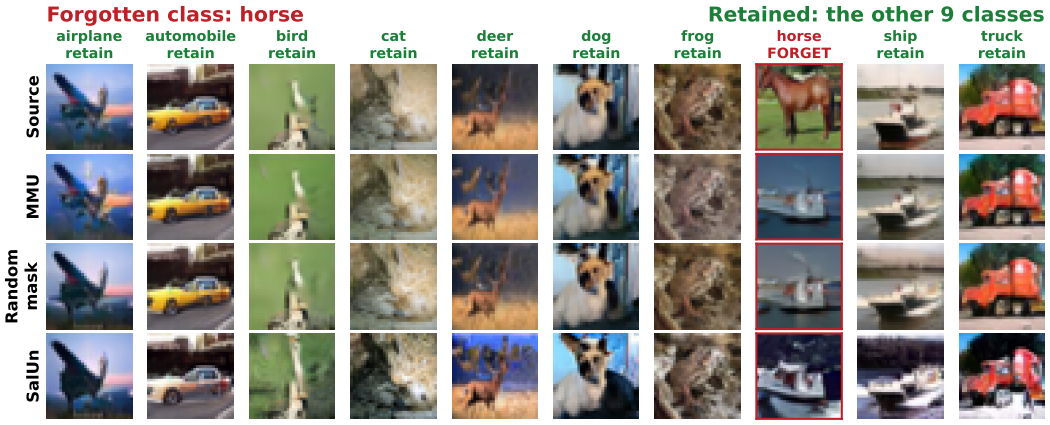

Figure 19: CIFAR-10 horse deletion.

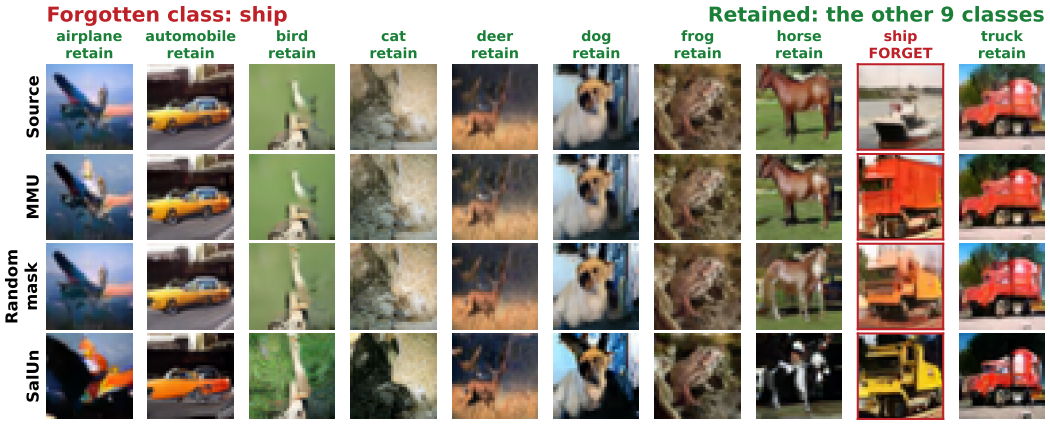

Figure 20: CIFAR-10 ship deletion.

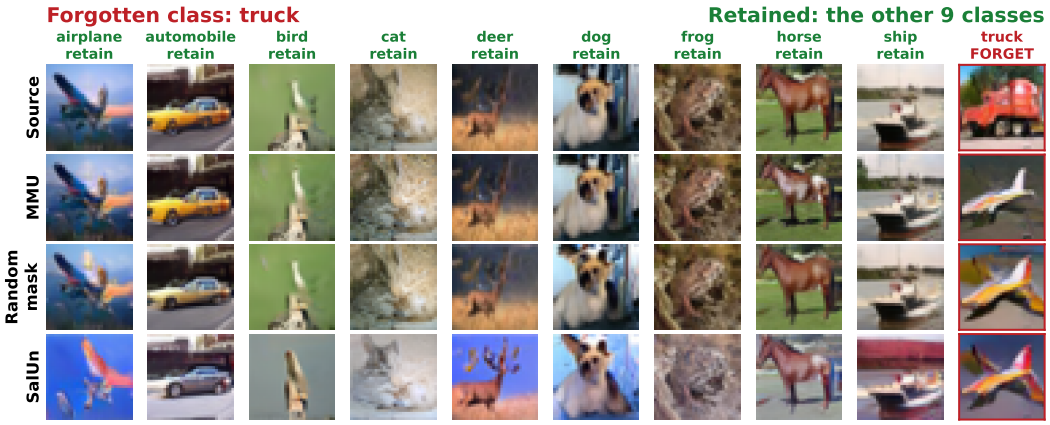

Figure 21: CIFAR-10 truck deletion.

C.5 Unlearning Hyperparameters

Table 12 lists the shared search grids and fixed settings for each method.

Table 12: Unlearning hyperparameter grids, shared across datasets, backbones, and deletion modes. # counts candidates; $\mathcal{D}_r^{10\%}$ is a fixed random 10% subset of $\mathcal{D}_r$.

Method	Data	Searched grid	#	Fixed settings
Retrain	$\mathcal{D}_r$	–	–	Source training recipe
Amnesiac[†]	update ledger	–	–	Subtracts updates recorded on batches containing $\mathcal{D}_f$
Fine-tune	$\mathcal{D}_r$	lr $\{10^{-4}, 10^{-3}, 10^{-2}\}$; epochs $\{1, 3, 5, 10\}$	12	Adam
Bad Teacher	$\mathcal{D}_f$, $\mathcal{D}_r^{10\%}$	T $\{1, 2, 4\}$; epochs $\{1, 3, 5, 10\}$; lr $\{10^{-4}, 10^{-3}, 10^{-2}\}$	36	SGD, momentum 0.9; incompetent teacher is a re-initialized copy
UNSIR	$\mathcal{D}_f$, $\mathcal{D}_r^{10\%}$	impair epochs $\{1, 2\}$; repair epochs $\{1, 2, 3\}$; impair lr $\{10^{-3}, 10^{-2}, 10^{-1}\}$; repair lr $\{10^{-3}, 10^{-2}\}$	36	Noise: 40 Adam steps, lr 0.1, regularization 0.1
SSD	$\mathcal{D}_f$, Fisher on $\mathcal{D}$	α $\{1, 5, 10, 50, 100\}$; λ $\{0.1, 0.5, 1, 5, 10\}$	25	No gradient updates
UniCLUN	$\mathcal{D}_f$, $\mathcal{D}_r^{10\%}$	lr $\{10^{-4}, 10^{-3}, 10^{-2}\}$; epochs $\{3, 5, 10\}$; buffer $\{250, 500, 1000\}$	27	SGD, momentum 0.9; $\rho = 2$; contrastive $T = 0.1$; $\alpha_1 = \alpha_2 = \alpha_3 = 1$; teacher momentum 0.99; $p = 0.6$
SalUn	$\mathcal{D}_f$, $\mathcal{D}_r^{10\%}$	mask density $\{0.1, 0.3, 0.5, 0.7, 0.9\}$; lr $\{10^{-4}, 10^{-3}, 10^{-2}\}$; epochs $\{3, 5, 10\}$	45	SGD, momentum 0.9, weight decay 5×10^{-4}
SCRUB	$\mathcal{D}_f$, $\mathcal{D}_r^{10\%}$	$E_{\max}$ $\{1, 2, 3\}$; lr $\{10^{-4}, 5 \times 10^{-4}, 10^{-3}, 5 \times 10^{-3}\}$; E $\{5, 10\}$	24	SGD, momentum 0.9, weight decay 5×10^{-4}; $T = 4$; $\alpha = \gamma = 1$; forget-step gradient-norm clip 5.0
MMU	$\mathcal{D}_f$, $\mathcal{D}_r^{10\%}$	$E_{\max}$ $\{1, 2, 3\}$; lr $\{10^{-4}, 5 \times 10^{-4}, 10^{-3}, 5 \times 10^{-3}\}$; E $\{5, 10\}$; head {MMU-1, MMU-2}; weights $\{\lambda_m = 1,$ normalized$\}$	96	As SCRUB; normalized: $\lambda_m^f = 1/K$, $\lambda_m^r \propto d/m$

[†]Class deletion only; the ledger has no equivalent for subclass or instance deletion.