REFERENCE-GUIDED MACHINE UNLEARNING

Jonas Mirlach[*], **Sonia Laguna, Julia E. Vogt**
Department of Computer Science, ETH Zurich

ABSTRACT

Machine unlearning aims to remove the influence of specific data from trained models while preserving general utility. Existing approximate unlearning methods often rely on performance-degradation heuristics, such as loss maximization or random labeling. However, these signals can be poorly conditioned, leading to unstable optimization and harming the model's generalization. We argue that unlearning should instead prioritize distributional indistinguishability, aligning the model's behavior on forget data with its behavior on truly unseen data. Motivated by this, we propose Reference-Guided Unlearning (REGUN), a framework that leverages a disjoint held-out dataset to provide a principled, class-conditioned reference for distillation. We demonstrate across various model architectures, natural image datasets, and varying forget fractions that REGUN consistently outperforms standard approximate baselines, achieving a superior forgetting–utility trade-off.

1 INTRODUCTION AND BACKGROUND

Machine unlearning (MU) is the principled process of updating a trained machine learning (ML) model to remove the influence of specified forget examples. MU has become an operational requirement for deployed AI systems, driven by privacy regulations such as GDPR's right to be forgotten and the need to adapt models post-deployment (Nguyen et al., 2025). While retraining from scratch without the forget examples provides the most faithful solution, it is often computationally prohibitive at scale. As a result, most practical MU methods rely on approximate updates (e.g., short fine-tuning or deletions) that aim to reduce reliance on the forget data while preserving accuracy on the retained data (Triantafillou et al., 2024; Laguna et al., 2026). A common baseline signal for unlearning is to degrade the model's performance on forget examples, such as by maximizing loss on target data or by fitting random or pseudo labels (Li et al., 2025). However, these can be poorly conditioned: they may induce large or misdirected gradients that change decision boundaries beyond the intended region and harm generalization (Mavrothalassitis et al., 2025). To mitigate this damage, many unlearning methods introduce restrictions, for instance, by staying close to the original model (Kurmanji et al., 2023), applying repair mechanisms (Tarun et al., 2024), or using constrained parameter editing (Fan et al., 2024; Foster et al., 2024). This conflicting optimization between forgetting and stability exposes a misalignment between current degradation proxies and the actual objective of mimicking unfamiliarity.

Rather than merely making the model "more wrong", we argue that unlearning should align its behavior on forget data with that of truly unseen examples. While this indistinguishability idea has been considered (Thudi et al., 2022), a principled reference for "unseen behavior" remains missing. We bridge this gap by proposing the paradigm of REference-Guided UNlearning (REGUN) with held-out supervision[1]. REGUN uses a disjoint held-out dataset as a stable proxy for "unseen behavior", aligning the model's outputs on forget examples with this reference distribution. Prior reference-based methods supervise unlearning by replacing forget-sample outputs with pseudo-probabilities (e.g., uniform or global distributions) (Zhao et al., 2024), or by matching the distributions of unseen third-party data (Chen et al., 2025). In contrast, REGUN uses a disjoint held-out subset as an explicit supervision source to construct the reference, enabling instance- or class-conditioned references rather than only marginal distribution matching.

Our main contributions to the field of machine unlearning include: *(i)* We introduce REGUN, a structured approach to machine unlearning using held-out data as a reference for distillation, and *(ii)* we empirically validate REGUN across multiple architectures and datasets in image classification, showing improved forgetting–utility trade-offs.

[*]Correspondence to: jmirlach@ethz.ch
[1]Code available at: https://github.com/jmirlach/ReGUn

2 METHODOLOGY

We formalize the unlearning setting and our proposed method REGUN in this section. A complementary algorithmic description of the method can be found in Appendix A.

2.1 PROBLEM SETUP AND NOTATION

We consider supervised K-class classification with input space $\mathcal{X}$ and labels $\mathcal{Y} = \{1, \dots, K\}$, and write $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n} \subseteq \mathcal{X} \times \mathcal{Y}$ for a labeled dataset. With Δ^K being the probability simplex over K classes, let $f_\theta : \mathcal{X} \to \Delta^K$ be a probabilistic classifier with parameters θ and predictive distribution $p_\theta(\cdot \mid x) = f_\theta(x)$. MU starts with a model trained on $\mathcal{D}_{\text{train}}$ and a request to remove the influence of a designated forget set $\mathcal{D}_{\text{f}} \subset \mathcal{D}_{\text{train}}$. We write $\mathcal{D}_{\text{train}} = \mathcal{D}_{\text{r}} \cup \mathcal{D}_{\text{f}}$ with $\mathcal{D}_{\text{r}} \cap \mathcal{D}_{\text{f}} = \emptyset$, where $\mathcal{D}_{\text{r}}$ is the retain set. Let θ_0 be obtained by training on $\mathcal{D}_{\text{train}}$. The goal of MU is to produce parameters θ_{u} such that $f_{\theta_{\text{u}}}$ behaves like the retraining baseline that would result if $\mathcal{D}_{\text{f}}$ had never been used. Concretely, letting θ_{r} denote parameters obtained by training on $\mathcal{D}_{\text{r}}$ only from scratch, we aim for $f_{\theta_{\text{u}}} \approx f_{\theta_{\text{r}}}$ while avoiding full retraining. Finally, we assume access to a disjoint held-out labeled dataset $\mathcal{D}_{\text{h}} = \{(x_j, y_j)\}_{j=1}^{n_{\text{h}}}$ with $\mathcal{D}_{\text{h}} \cap \mathcal{D}_{\text{train}} = \emptyset$.

2.2 REFERENCE-GUIDED UNLEARNING

Rather than making the model intentionally incorrect on the forget set, we aim to replace its behavior on forget examples with that characteristic of inputs the model has never seen. To operationalize this "unseen behavior", we leverage the held-out set $\mathcal{D}_{\text{h}}$ to construct a reference prediction distribution. We thus treat unlearning as distilling forget-set predictions to match this reference distribution.

Reference distribution. At each iteration in the unlearning phase, we sample a forget minibatch $B_{\text{f}} = \{(x_i^{\text{f}}, y_i^{\text{f}})\}_{i=1}^{b} \subset \mathcal{D}_{\text{f}}$ and compute a batch-level soft target

$$q(B_{\text{f}}) \;:=\; \textsc{RefDist}(B_{\text{f}}; \mathcal{D}_{\text{h}}, f_\phi) \in \Delta^K,$$

where f_ϕ is any reference model. Ideally, $f_\phi = f_{\theta_r}$, since this oracle model best represents the desired post-unlearning behavior. However, f_{θ_r} is typically unavailable in approximate unlearning. We therefore set $f_\phi = f_{\theta_0}$, the initial model state, to avoid extra training and prevent reference drift, although notably f_{θ_0} still retains influence from $\mathcal{D}_{\text{f}}$. To construct the corresponding reference target, REFDIST selects a small set of held-out samples $\tilde{\mathcal{D}}_{\text{h}} = \{(\tilde{x}_j, \tilde{y}_j)\}_{j=1}^{m} \subset \mathcal{D}_{\text{h}}$ and aggregates the corresponding model outputs into a single distribution. Samples are selected based on matching the class histogram of B_{f} using labels in $\mathcal{D}_{\text{h}}$. By matching the class histogram, we control for differences in label priors, so $q(B_{\text{f}})$ approximates a class-conditional unseen reference. Let $\ell_j := z_\phi(\tilde{x}_j) \in \mathbb{R}^K$ denote the logits of f_ϕ, with $p_\phi(\cdot \mid \tilde{x}_j) = \mathrm{softmax}(\ell_j)$. We aggregate these held-out predictions via the mean of probabilities

$$q(B_{\text{f}}) \;=\; \frac{1}{m} \sum_{j=1}^{m} p_\phi(\cdot \mid \tilde{x}_j),$$

which corresponds to an empirical mixture of the reference model's outputs on the selected held-out inputs. The same $q(B_{\text{f}})$ is used for all $x \in B_{\text{f}}$ and, notably, it depends on B_{f} through batch statistics, i.e. the label histogram.

Unlearning objective. In parallel, we sample a retain minibatch $B_{\text{r}} = \{(x_i^{\text{r}}, y_i^{\text{r}})\}_{i=1}^{|B_{\text{r}}|} \subset \mathcal{D}_{\text{r}}$. Starting from $\theta = \theta_0$, we update θ by minimizing

$$\mathcal{L}(\theta; B_{\text{f}}, B_{\text{r}}) \;=\; \lambda_{\text{f}} \frac{1}{|B_{\text{f}}|} \sum_{(x,\cdot) \in B_{\text{f}}} \mathrm{KL}(q(B_{\text{f}}) \,\|\, p_\theta(\cdot \mid x)) \;+\; \lambda_{\text{r}} \frac{1}{|B_{\text{r}}|} \sum_{(x,y) \in B_{\text{r}}} \mathrm{CE}(p_\theta(\cdot \mid x), y)\,,$$

where $\lambda_{\text{f}}, \lambda_{\text{r}} > 0$ trade off forgetting strength and retain utility, and $\mathrm{CE}(p, y) = -\log p(y)$ is the standard cross-entropy for a hard label y. The first term distills the model's predictions on forget inputs toward the held-out reference distribution and the second term anchors the update to preserve performance on retained data. Note that $\mathrm{KL}(q\|p)$ is equivalent to cross-entropy with a soft target q up to an additive constant independent of θ, hence the forget term can also be interpreted as standard distillation to a held-out teacher distribution.

3 EXPERIMENTAL SETUP

We conduct a set of experiments to analyze the forgetting–utility trade-off of REGUN in image classification tasks (Li et al., 2025). Regarding the unlearning setup, we consider random forgetting by sampling $\mathcal{D}_{\mathrm{f}}$ uniformly at random with forget fractions $|\mathcal{D}_{\mathrm{f}}|/|\mathcal{D}_{\mathrm{train}}| \in \{0.01, 0.1, 0.5\}$. From the original training set $\mathcal{D}_{\mathrm{orig}}$, we reserve a held-out set $\mathcal{D}_{\mathrm{h}}$ of size $0.1|\mathcal{D}_{\mathrm{orig}}|$ (used only during unlearning), leaving us with $\mathcal{D}_{\mathrm{train}} = \mathcal{D}_{\mathrm{orig}} \setminus \mathcal{D}_{\mathrm{h}}$. From here, we sample $\mathcal{D}_{\mathrm{f}}$ and a validation split $\mathcal{D}_{\mathrm{val}}$ of size $0.1|\mathcal{D}_{\mathrm{train}}|$ for hyperparameter selection; the remainder is $\mathcal{D}_{\mathrm{r}}$. All models are trained from scratch on $\mathcal{D}_{\mathrm{train}}$ and then unlearned. We assume access to labeled $\mathcal{D}_{\mathrm{train}}$ during unlearning.

Our main experiments are divided into two setups, covering both standard CNNs and Transformer-based models on a higher-resolution benchmark: ResNet-18 (He et al., 2016) on CIFAR-10 (Krizhevsky, 2009) and Swin-T (Liu et al., 2021) on Tiny-ImageNet (Deng et al., 2009). We compare against the simpler approximate unlearning baselines FINETUNE, NEGGRAD, and NEGGRAD+ (Kurmanji et al., 2023), as well as the more elaborate methods ℓ_1-SPARSE (Jia et al., 2023), SSD (Foster et al., 2024), SALUN (Fan et al., 2024), and AMUN (Ebrahimpour-Boroojeny et al., 2025). All training details and hyperparameter searches for REGUN and studied baselines are reported in Appendix B. We evaluate MU along its two core objectives: retained utility and forgetting efficacy. Accordingly, we report retain, forget, and test accuracy (TEST$_{\mathrm{Acc}}$), and quantify membership inference risk via attack AUC. For membership inference, we use robust membership inference attack (RMIA) (Zarifzadeh et al., 2024) in the offline setting with four reference models. We summarize overall performance with GAP$^{\mathrm{RFTP}}_{\mathrm{Avg}}$, the average deviation from the retrain-from-scratch baseline (RETRAIN) across these four metrics. For conciseness, the main text reports TEST$_{\mathrm{Acc}}$ as the main measure of utility and the RMIA$_{\mathrm{AUC}}$ as the main measure of forgetting, while the complete metric suite, along with additional results, is provided in Appendix C. All results are averaged over three seeds and reported as mean $\pm$ std, with all metrics expressed in %.

4 RESULTS

Table 1 presents our main findings across setups and forget fractions. In the CNN-based (ResNet-18) setting, we find that overall trends largely mirror those of standard approximate baselines, while REGUN consistently produces results that are among the closest to the retrain-from-scratch benchmark.

	FORGET 1%			FORGET 10%			FORGET 50%		
	TEST$_{\mathrm{Acc}}$	RMIA$_{\mathrm{AUC}}$	GAP$^{\mathrm{RFTP}}_{\mathrm{Avg}}$	TEST$_{\mathrm{Acc}}$	RMIA$_{\mathrm{AUC}}$	GAP$^{\mathrm{RFTP}}_{\mathrm{Avg}}$	TEST$_{\mathrm{Acc}}$	RMIA$_{\mathrm{AUC}}$	GAP$^{\mathrm{RFTP}}_{\mathrm{Avg}}$
RESNET-18 ON CIFAR-10									
RETRAIN	94.34 ±0.02	49.98 ±1.26	0.00 ±0.00	93.81 ±0.19	50.19 ±0.92	0.00 ±0.00	90.31 ±0.41	50.24 ±0.36	0.00 ±0.00
BASE	94.20 ±0.11	60.11 ±1.31	3.88 ±0.90	94.29 ±0.13	59.50 ±0.89	3.88 ±0.29	94.17 ±0.02	56.42 ±0.32	4.79 ±0.30
NEGGRAD	**94.17** ±0.01	59.80 ±1.12	3.82 ±0.89	**93.89** ±0.39	59.47 ±0.78	3.82 ±0.29	94.05 ±0.05	56.59 ±0.35	4.80 ±0.33
NEGGRAD+	91.80 ±3.75	57.95 ±2.16	3.77 ±0.72	93.02 ±0.65	59.10 ±0.74	3.71 ±0.33	88.62 ±2.17	53.19 ±2.38	2.62 ±0.58
FINETUNE	90.90 ±0.57	54.78 ±0.97	2.88 ±0.26	90.23 ±0.53	53.92 ±0.53	2.79 ±0.36	88.10 ±0.46	52.40 ±0.62	2.39 ±0.47
ℓ_1-SPARSE	90.97 ±0.11	53.89 ±1.91	2.73 ±0.23	90.63 ±0.25	53.01 ±1.42	2.49 ±0.30	88.82 ±0.75	52.81 ±0.69	2.09 ±0.10
SSD	93.82 ±0.68	59.69 ±1.14	3.84 ±0.88	94.29 ±0.13	59.50 ±0.89	3.88 ±0.30	94.18 ±0.01	56.42 ±0.32	4.79 ±0.30
SALUN	91.63 ±0.20	**50.09** ±3.34	1.64 ±0.21	91.59 ±1.28	53.45 ±1.49	2.48 ±0.13	89.00 ±1.03	52.76 ±0.66	2.00 ±0.13
AMUN	91.84 ±0.34	44.17 ±1.49	3.94 ±1.56	91.97 ±0.20	52.63 ±0.64	**1.46** ±0.15	89.49 ±1.96	**51.02** ±1.88	1.84 ±0.24
REGUN	90.93 ±1.14	48.90 ±0.51	**1.21** ±0.26	90.60 ±1.26	**51.01** ±0.64	2.00 ±0.18	**90.11** ±0.12	52.10 ±0.11	**1.48** ±0.08
SWIN-T ON TINY-IMAGENET									
RETRAIN	60.90 ±0.14	49.81 ±1.41	0.00 ±0.00	59.27 ±0.30	50.30 ±0.66	0.00 ±0.00	47.95 ±0.12	50.30 ±0.19	0.00 ±0.00
BASE	61.21 ±0.04	87.77 ±0.18	19.20 ±0.42	61.03 ±0.23	86.40 ±0.26	19.58 ±0.32	61.20 ±0.20	79.74 ±0.05	23.58 ±0.04
NEGGRAD	**61.22** ±0.06	87.78 ±0.18	19.20 ±0.42	**61.02** ±0.22	86.43 ±0.26	19.58 ±0.32	61.19 ±0.16	79.84 ±0.05	23.61 ±0.05
NEGGRAD+	48.99 ±0.31	66.99 ±1.96	10.52 ±0.59	46.49 ±0.31	57.18 ±0.11	9.07 ±0.48	43.63 ±0.98	56.66 ±2.25	4.12 ±0.18
FINETUNE	52.31 ±0.36	62.48 ±0.22	6.60 ±0.59	51.00 ±0.75	62.40 ±0.34	6.32 ±0.18	45.74 ±0.82	61.46 ±0.42	5.70 ±0.10
ℓ_1-SPARSE	51.94 ±0.03	62.37 ±0.96	6.54 ±0.33	50.26 ±0.33	61.00 ±2.48	6.39 ±0.24	43.48 ±0.53	55.23 ±0.35	4.05 ±0.30
SSD	41.99 ±7.57	68.08 ±4.21	17.63 ±4.28	55.87 ±3.84	84.36 ±0.15	19.41 ±0.85	38.10 ±2.80	74.84 ±3.49	21.09 ±0.59
SALUN	53.16 ±0.69	46.36 ±1.39	**3.73** ±0.20	49.88 ±0.24	55.77 ±0.80	7.03 ±0.24	**47.77** ±0.30	58.04 ±0.95	5.36 ±0.05
AMUN	52.59 ±0.46	**51.03** ±0.57	6.55 ±0.62	51.06 ±0.39	55.35 ±0.30	5.93 ±0.15	48.21 ±0.75	59.37 ±0.34	5.34 ±0.15
REGUN	52.26 ±0.40	45.07 ±0.83	5.82 ±0.16	52.72 ±0.25	**49.86** ±0.83	**3.05** ±0.21	45.57 ±0.44	**47.88** ±0.19	**1.37** ±0.09

Table 1: Results for ResNet-18 on CIFAR-10 and Swin-T on Tiny-ImageNet under random forgetting. **Bold** and underlined denote best and second best, where "best" is smallest gap to RETRAIN.

More notably, in the Transformer-based (Swin-T) and high-resolution setting, REGUN demonstrates strong empirical performance, achieving the lowest average gap metric across most scenarios. This is especially evident at higher forget fractions, where REGUN is the only method in our comparison to reliably reduce RMIA scores to the target retrain-from-scratch level, suggesting strong forgetting capabilities in these larger forgetting regimes. Among the baselines, we observe that while the most recent and elaborate methods, such as SALUN and AMUN, still generally outperform other approximate techniques, interestingly, the simpler baselines NEGGRAD+ and FINETUNE remain highly competitive in the Transformer setting. More generally, we find that Transformer-based methods remain comparatively underexplored in the current machine unlearning literature. The performance gaps between approximate methods and the RETRAIN baseline are notably larger than those observed in the ResNet-18 experiments. This disparity suggests that the attention mechanisms and representation spaces of Transformers pose unique challenges for existing unlearning heuristics.

Figure 1 visualizes the forgetting–utility trade-off among studied methods in more detail. For comparability, we reparameterize the objectives of those that admit an explicit trade-off as $(1-w)\cdot\mathcal{L}_{\text{forget}} + w\cdot\mathcal{L}_{\text{retain}}$, and sweep $w \in [0.1, 0.9]$ to examine how the strength of the forget signal affects performance. While the trade-off curves for ResNet-18 on CIFAR-10 are largely comparable and only partially conclusive, the Swin-T setting on Tiny-ImageNet reveals a more distinct pattern, where REGUN achieves the most favorable trade-off among methods. In particular, we see that increasing the forget signal degrades utility for other methods, whereas REGUN maintains a more constant utility across the sweep, aligning with the intended behavior of an unlearning method.

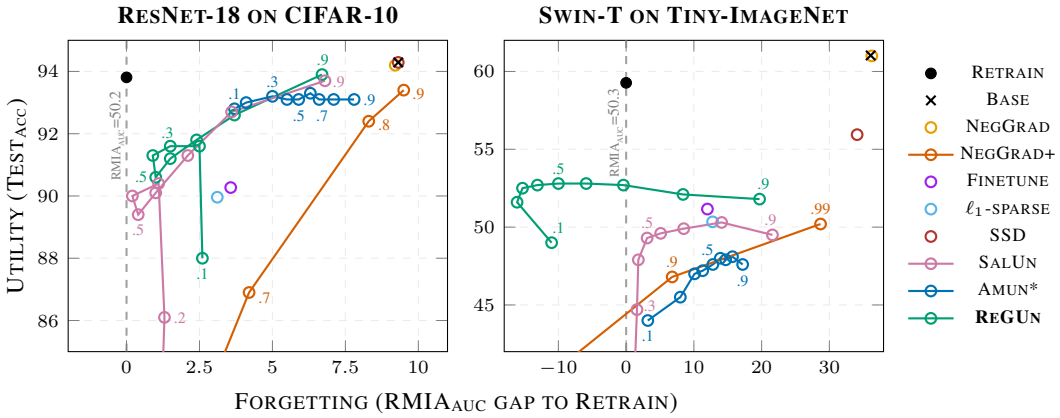

Figure 1: Forgetting–utility trade-offs for CIFAR-10 and Tiny-ImageNet under random 10% forgetting. Where supported, we sweep $w \in [0.1, 0.9]$ (additionally $w = 0.99$ for NEGGRAD+). Points are labeled with their w weight. * marks variants reparameterized from the original loss to sweep w.

5 CONCLUSION

We introduced REGUN, a novel MU framework that reframes approximate unlearning from performance degradation to distributional matching. Instead of pushing the model to be wrong on forget examples via often poorly conditioned objectives, REGUN aligns the model's behavior on forget samples with its predictive distribution on a disjoint, held-out reference set. This approach is motivated by the principle of indistinguishability: a truly unlearned model should treat the forget set as if it were a future, unseen test set, precisely the property that many MIAs aim to detect. Our results across diverse architectures and datasets demonstrate that held-out reference supervision is a promising unlearning signal that leads to a favorable forgetting–utility trade-off.

We hope these findings encourage future research to prioritize indistinguishability as a core objective, even when developing more complex unlearning solutions. The flexibility of the REGUN framework opens several avenues for future work. While we focused on class-conditioned references, exploring feature-space nearest neighbors, instance-conditioned references, or alternative held-out sampling strategies could further refine the unlearning signal and improve robustness across forget regimes. Beyond discriminative tasks, extending these principles to generative modeling remains a critical frontier, and future research should investigate how these distributional matching objectives scale to high-dimensional foundation models.

ACKNOWLEDGEMENTS

SL is supported by the Swiss State Secretariat for Education, Research and Innovation (SERI) under contract number MB22.00047.

REFERENCES

Kongyang Chen, Dongping Zhang, Bing Mi, Yao Huang, and Zhipeng Li. Fast yet versatile machine unlearning for deep neural networks. *Neural Networks*, 2025. doi: 10.1016/j.neunet.2025.107648. 1

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 3

Ali Ebrahimpour-Boroojeny, Hari Sundaram, and Varun Chandrasekaran. Not all wrong is bad: Using adversarial examples for unlearning. In *ICML*, 2025. 3

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. 1, 3

Jack Foster, Stefan Schoepf, and Alexandra Brintrup. Fast machine unlearning without retraining through selective synaptic dampening. *AAAI*, 2024. doi: 10.1609/aaai.v38i11.29092. 1, 3

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. doi: 10.1109/CVPR.2016.90. 3

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*, 2023. 3

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. 3

Meghdad Kurmanji, Peter Triantafillou, Jamie Hayes, and Eleni Triantafillou. Towards unbounded machine unlearning. In *NeurIPS*, 2023. 1, 3

Sonia Laguna, Jorge da Silva Gonçalves, Moritz Vandenhirtz, Alain Ryser, Irene Cannistraci, and Julia E Vogt. Rethinking machine unlearning: Models designed to forget via key deletion. In *ICLR*, 2026. 1

Xiang Li, Wenqi Wei, and Bhavani Thuraisingham. Mubox: A critical evaluation framework of deep machine unlearning. In *SACMAT*, 2025. doi: 10.1145/3734436.3734454. 1, 3

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 *ICCV*, 2021. 3

Ioannis Mavrothalassitis, Pol Puigdemont, Noam Itzhak Levi, and Volkan Cevher. Ascent fails to forget. In *NeurIPS*, 2025. 1

Thanh Tam Nguyen, Thanh Trung Huynh, Zhao Ren, Phi Le Nguyen, Alan Wee-Chung Liew, Hongzhi Yin, and Quoc Viet Hung Nguyen. A survey of machine unlearning. *ACM Transactions on Intelligent Systems and Technology*, 2025. 1

Gaurav Patel and Qiang Qiu. Learning to unlearn while retaining: Combating gradient conflicts in machine unlearning. In *ICCV*, 2025. 8

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. doi: 10.1109/TNNLS.2023.3266233. 1

Anvith Thudi, Hengrui Jia, Ilia Shumailov, and Nicolas Papernot. On the necessity of auditable algorithmic definitions for machine unlearning. In *USENIX Security*, 2022. 1

Eleni Triantafillou, Peter Kairouz, Fabian Pedregosa, Jamie Hayes, Meghdad Kurmanji, Kairan Zhao, Vincent Dumoulin, et al. Are we making progress in unlearning? findings from the first neurips unlearning competition. *CoRR*, 2024. doi: 10.48550/arXiv.2406.09073. 1

Sajjad Zarifzadeh, Philippe Liu, and Reza Shokri. Low-cost high-power membership inference attacks. In *ICML*, 2024. 3

Zihao Zhao, Yijiang Li, Yuchen Yang, Wenqing Zhang, Nuno Vasconcelos, and Yinzhi Cao. Pseudo-probability unlearning: Towards efficient and privacy-preserving machine unlearning. *arXiv*, 2024. doi: 10.48550/arXiv.2411.02622. 1

A REGUN ALGORITHM

We include here an algorithmic description of the REGUN unlearning procedure, along with the REFDIST sample selection step for forming the reference distribution.

Algorithm 1 REFDIST: Held-out Reference Distribution (mean of probabilities)

Input: forget batch $B_{\mathrm{f}} = \{(x_i^{\mathrm{f}}, y_i^{\mathrm{f}})\}_{i=1}^{b}$, held-out set $\mathcal{D}_{\mathrm{h}}$, reference model f_ϕ, held-out sample size m (*default:* $m = b$)
Let $c_k = \sum_{i=1}^{b} \mathbf{1}[y_i^{\mathrm{f}} = k]$ for $k \in \{1, \ldots, K\}$
Set $\tilde{c}_k \leftarrow \mathrm{round}(m \cdot c_k / b)$ and adjust to ensure $\sum_k \tilde{c}_k = m$
Sample $\tilde{X}_{\mathrm{h}} = \{\tilde{x}_j\}_{j=1}^{m}$ by drawing $\tilde{c}_k$ inputs uniformly from $\{x^{\mathrm{h}} : (x^{\mathrm{h}}, y^{\mathrm{h}}) \in \mathcal{D}_{\mathrm{h}},\ y^{\mathrm{h}} = k\}$ for each class k *(with replacement if needed)*
Aggregate reference predictions: $q \leftarrow \frac{1}{m} \sum_{j=1}^{m} p_\phi(\cdot \mid \tilde{x}_j)$
Output: $q \in \Delta^K$

Algorithm 2 REGUN: Reference-Guided Unlearning

Input: initial model f_{θ_0}, retain set $\mathcal{D}_{\mathrm{r}}$, forget set $\mathcal{D}_{\mathrm{f}}$, held-out set $\mathcal{D}_{\mathrm{h}}$, weights $\lambda_{\mathrm{f}}, \lambda_{\mathrm{r}}$, steps T
Initialize $\theta \leftarrow \theta_0$
for $t = 1, \ldots, T$ **do**
    Sample minibatches $B_{\mathrm{f}} \subset \mathcal{D}_{\mathrm{f}}$ and $B_{\mathrm{r}} \subset \mathcal{D}_{\mathrm{r}}$
    $q \leftarrow \textsc{RefDist}(B_{\mathrm{f}}; \mathcal{D}_{\mathrm{h}}, f_{\theta_0})$
    $\mathcal{L}_{\mathrm{f}} \leftarrow \frac{1}{|B_{\mathrm{f}}|} \sum_{(x,\cdot) \in B_{\mathrm{f}}} \mathrm{KL}(q \,\|\, p_\theta(\cdot \mid x))$
    $\mathcal{L}_{\mathrm{r}} \leftarrow \frac{1}{|B_{\mathrm{r}}|} \sum_{(x,y) \in B_{\mathrm{r}}} \mathrm{CE}(p_\theta(\cdot \mid x), y)$
    $\theta \leftarrow \theta - \eta \nabla_\theta \left(\lambda_{\mathrm{f}} \mathcal{L}_{\mathrm{f}} + \lambda_{\mathrm{r}} \mathcal{L}_{\mathrm{r}}\right)$
end for
$\theta_{\mathrm{u}} \leftarrow \theta$
Output: unlearned model $f_{\theta_{\mathrm{u}}}$

B Training Setup, Implementation Details, and Hyperparameters

We summarize training and hyperparameter search spaces and choices below. For additional implementation details, please refer to the published code repository[2].

Base & Retrain. For base training (from scratch) and retraining, we follow standard recipes. For ResNet-18, we train for 100 epochs using SGD with momentum 0.9, learning rate 0.1, batch size 128, and apply standard data augmentation consisting of random crop with padding (32×32, pad=4), random horizontal flipping (p=0.5), and mild color jitter (brightness/contrast/saturation=0.1, hue=0.02). For Swin-T, we train for 200 epochs using AdamW with learning rate $3 \cdot 10^{-4}$ and a cosine schedule, batch size 128, and apply the standard data augmentation consisting of random horizontal flipping (p=0.5), and mild color jitter (brightness/contrast/saturation=0.1, hue=0.02) plus additional stronger augmentation in the form of RandAugment (N=2, M=9) and Random Erasing (p=0.25, area 2–20%). For ResNet-18 we use the native data resolutions of 32×32 on CIFAR, and 64×64 on Tiny-ImageNet. For Swin-T on Tiny-ImageNet we use 224×224 resolution to align with the architecture's default patch-based configuration.

For unlearning (except NegGrad), we use a fixed budget of 10 epochs for ResNet-18 and 20 epochs for Swin-T. All other settings are kept the same as base training except that we omit the strong data augmentation in the Swin-T setup. The detailed configurations per method are the following:

NegGrad. We run gradient ascent for 2 epochs and tune the learning rate. ResNet-18: $\mathrm{lr} \in \{5\mathrm{e}{-2}, 1\mathrm{e}{-2}, 5\mathrm{e}{-3}, 1\mathrm{e}{-3}\}$. Swin-T: $\mathrm{lr} \in \{1\mathrm{e}{-5}, 5\mathrm{e}{-6}, 1\mathrm{e}{-6}, 5\mathrm{e}{-7}, 1\mathrm{e}{-7}, 5\mathrm{e}{-8}, 1\mathrm{e}{-8}\}$.

NegGrad+. We combine negative-gradient forget objective with a retain objective weighted by w. We tune the learning rate and the retain weight w. ResNet-18: $\mathrm{lr} \in \{1\mathrm{e}{-1}, 5\mathrm{e}{-2}, 1\mathrm{e}{-2}, 5\mathrm{e}{-3}\}$, $w \in [0.8, 0.99]$. Swin-T: $\mathrm{lr} \in \{1\mathrm{e}{-3}, 5\mathrm{e}{-4}, 1\mathrm{e}{-4}\}$, $w \in [0.8, 0.99]$.

FineTune. We fine-tune on the full retain set $\mathcal{D}_\mathrm{r}$ and tune the learning rate. ResNet-18: $\mathrm{lr} \in \{1\mathrm{e}{-1}, 5\mathrm{e}{-2}, 1\mathrm{e}{-2}, 5\mathrm{e}{-3}\}$. Swin-T: $\mathrm{lr} \in \{1\mathrm{e}{-3}, 5\mathrm{e}{-4}, 1\mathrm{e}{-4}\}$.

ℓ_1-sparse. We implement sparsity-aware unlearning as fine-tuning on $\mathcal{D}_\mathrm{r}$ with an ℓ_1 penalty weight γ. We tune the learning rate and γ. ResNet-18: $\mathrm{lr} \in \{1\mathrm{e}{-1}, 5\mathrm{e}{-2}, 1\mathrm{e}{-2}, 5\mathrm{e}{-3}\}$, $\gamma \in [5\mathrm{e}{-6}, 5\mathrm{e}{-3}]$. Swin-T: $\mathrm{lr} \in \{1\mathrm{e}{-3}, 5\mathrm{e}{-4}, 1\mathrm{e}{-4}\}$, $\gamma \in [5\mathrm{e}{-7}, 5\mathrm{e}{-5}]$.

SSD. For ResNet-18, we use the recommended settings from the paper ($\alpha = 10.0$, $\lambda = 1.0$). For Swin-T, we tune α and λ. Swin-T: $\alpha \in [1, 10]$, $\lambda \in [0.7, 1.0]$. Note that SSD is not designed for large random-forgetting regimes, which should be considered when interpreting the results.

SalUn. We fix the sparsity threshold to 50% and tune the learning rate and a retain weight w that balances the retain objective with the forgetting loss. ResNet-18: $\mathrm{lr} \in \{1\mathrm{e}{-1}, 5\mathrm{e}{-2}, 1\mathrm{e}{-2}, 5\mathrm{e}{-3}\}$, $w \in [0.1, 0.9]$. Swin-T: $\mathrm{lr} \in \{1\mathrm{e}{-3}, 5\mathrm{e}{-4}, 1\mathrm{e}{-4}\}$, $w \in [0.1, 0.9]$.

AMUN. We evaluate all four configurations described in the paper, corresponding to fine-tuning on $\mathcal{D}_\mathrm{A} \cup \mathcal{D}_\mathrm{f} \cup \mathcal{D}_\mathrm{r}$, $\mathcal{D}_\mathrm{A} \cup \mathcal{D}_\mathrm{f}$, $\mathcal{D}_\mathrm{A} \cup \mathcal{D}_\mathrm{r}$, and $\mathcal{D}_\mathrm{A}$ and tune the learning rate. ResNet-18: $\mathrm{lr} \in \{1\mathrm{e}{-1}, 5\mathrm{e}{-2}, 1\mathrm{e}{-2}, 5\mathrm{e}{-3}\}$. Swin-T: $\mathrm{lr} \in \{1\mathrm{e}{-3}, 5\mathrm{e}{-4}, 1\mathrm{e}{-4}\}$.

ReGUn. We tune the learning rate and the retain/forget trade-off weight w (corresponding to $\lambda_\mathrm{r} = w$ and $\lambda_\mathrm{f} = 1 - w$ in the main objective). ResNet-18: $\mathrm{lr} \in \{1\mathrm{e}{-1}, 5\mathrm{e}{-2}, 1\mathrm{e}{-2}, 5\mathrm{e}{-3}\}$, $w \in [0.1, 0.9]$. Swin-T: $\mathrm{lr} \in \{1\mathrm{e}{-3}, 5\mathrm{e}{-4}, 1\mathrm{e}{-4}\}$, $w \in [0.1, 0.9]$.

To produce the results for Figure 1, we fixed per method the best hyperparameter setting (excluding w) and then swept w. NegGrad+, SalUn, and ReGUn inherently support this sweep. For AMUN, we rewrote the objective as $(1 - w)\mathcal{L}_\mathrm{forget} + w\mathcal{L}_\mathrm{retain}$ with $\mathcal{L}_\mathrm{forget}$ the loss on $\mathcal{D}_\mathrm{A}$ and $\mathcal{L}_\mathrm{retain}$ the loss on $\mathcal{D}_\mathrm{r}$.

[2] https://github.com/jmirlach/ReGUn

C Detailed Results and Additional Experiments

The following tables provide the full set of results corresponding to Table 1, and additionally include experiments with ResNet-18 on CIFAR-100 and Tiny-ImageNet. For ResNet experiments, we also include results for LUR (Patel & Qiu, 2025), which were not included in the main section. We report several additional metrics in this appendix to improve comparability with prior work and to provide a more complete picture of forgetting and utility: SMIA_{AUC} denotes the AUC of a simple loss-based membership inference attack. $\text{RETAIN}_{\text{DIV}}$ and TEST_{DIV} measure the Jensen–Shannon divergence between the predictive distributions of the unlearned model and the retrained model, evaluated on $\mathcal{D}_{\text{r}}$ and $\mathcal{D}_{\text{test}}$, respectively. $\text{GAP}^{\text{TP}}_{\text{Avg}}$ summarizes performance as the average deviation from the retraining baseline considering only TEST_{ACC} and RMIA_{AUC}, which in a more direct way balances utility and forgetting compared to $\text{GAP}^{\text{RFTP}}_{\text{Avg}}$.

C.1 ResNet-18 on CIFAR-10

	$\text{RETAIN}_{\text{ACC}}$	$\text{FORGET}_{\text{ACC}}$	TEST_{ACC}	$\text{RETAIN}_{\text{DIV}}$	TEST_{DIV}	RMIA_{AUC}	SMIA_{AUC}	$\text{GAP}^{\text{RFTP}}_{\text{Avg}}$	$\text{GAP}^{\text{TP}}_{\text{Avg}}$
FORGET 1%									
RETRAIN	100.00 ±0.00	94.22 ±1.24	94.34 ±0.02	0.00 ±0.00	0.00 ±0.00	49.98 ±1.26	50.78 ±0.93	0.00 ±0.00	0.00 ±0.00
BASE	100.00 ±0.00	100.00 ±0.00	94.20 ±0.11	0.03 ±0.00	2.28 ±0.08	60.11 ±1.30	59.99 ±0.38	3.88 ±0.90	5.14 ±0.93
NEGGRAD	**100.00** ±0.00	100.00 ±0.00	**94.17** ±0.01	**0.03** ±0.00	**2.28** ±0.07	59.80 ±1.12	59.77 ±0.36	3.82 ±0.89	5.19 ±0.99
NEGGRAD+	98.45 ±2.68	97.85 ±3.72	91.80 ±3.75	1.04 ±1.74	3.73 ±2.35	57.95 ±2.16	57.39 ±3.54	3.77 ±0.72	5.39 ±1.12
FINETUNE	97.44 ±0.49	**94.67** ±1.94	90.90 ±0.57	1.71 ±0.25	4.25 ±0.45	54.78 ±0.97	52.42 ±1.34	2.88 ±0.26	4.26 ±0.34
ℓ_1-SPARSE	97.15 ±0.77	94.89 ±2.04	90.97 ±0.11	2.00 ±0.60	4.26 ±0.11	53.89 ±1.91	52.03 ±0.33	2.73 ±0.23	3.78 ±0.87
LUR	74.07 ±4.72	73.63 ±5.13	72.21 ±4.39	20.19 ±4.10	19.08 ±3.80	50.51 ±0.25	**50.47** ±0.41	17.59 ±3.93	11.47 ±2.08
SSD	99.95 ±0.09	99.85 ±0.26	93.82 ±0.68	0.12 ±0.15	2.50 ±0.32	59.69 ±1.14	59.69 ±0.75	3.84 ±0.88	5.28 ±1.04
SALUN	99.60 ±0.13	96.59 ±1.28	91.63 ±0.20	4.43 ±2.50	7.92 ±2.42	**50.09** ±3.34	48.40 ±3.01	1.64 ±0.21	2.34 ±0.16
AMUN	99.28 ±0.16	87.85 ±1.89	91.84 ±0.34	0.56 ±0.10	3.80 ±0.22	44.17 ±1.49	41.65 ±1.95	3.94 ±1.56	3.90 ±1.33
REGUN	99.37 ±0.50	95.33 ±1.94	90.93 ±1.14	6.07 ±3.37	9.59 ±3.42	48.90 ±0.51	47.06 ±0.38	**1.21** ±0.26	**1.99** ±0.44
FORGET 10%									
	$\text{RETAIN}_{\text{ACC}}$	$\text{FORGET}_{\text{ACC}}$	TEST_{ACC}	$\text{RETAIN}_{\text{DIV}}$	TEST_{DIV}	RMIA_{AUC}	SMIA_{AUC}	$\text{GAP}^{\text{RFTP}}_{\text{Avg}}$	$\text{GAP}^{\text{TP}}_{\text{Avg}}$
RETRAIN	100.00 ±0.00	94.28 ±0.38	93.81 ±0.19	0.00 ±0.00	0.00 ±0.00	50.19 ±0.92	49.86 ±0.29	0.00 ±0.00	0.00 ±0.00
BASE	100.00 ±0.00	100.00 ±0.00	94.29 ±0.13	0.03 ±0.00	2.34 ±0.11	59.50 ±0.89	59.51 ±0.44	3.88 ±0.29	4.90 ±0.43
NEGGRAD	99.90 ±0.15	99.83 ±0.19	**93.89** ±0.39	0.10 ±0.10	2.63 ±0.23	59.47 ±0.78	58.91 ±0.14	3.82 ±0.29	4.81 ±0.44
NEGGRAD+	99.85 ±0.13	99.26 ±0.53	93.02 ±0.65	0.17 ±0.10	3.11 ±0.49	59.10 ±0.74	57.69 ±0.75	3.71 ±0.33	4.85 ±0.37
FINETUNE	97.01 ±0.44	93.39 ±0.72	90.23 ±0.53	2.04 ±0.24	4.58 ±0.28	53.92 ±0.53	51.59 ±0.30	2.79 ±0.36	3.65 ±0.40
ℓ_1-SPARSE	97.16 ±0.51	93.15 ±1.11	90.63 ±0.25	2.01 ±0.39	4.31 ±0.04	53.01 ±1.42	51.09 ±0.53	2.49 ±0.30	3.00 ±0.45
LUR	96.20 ±0.44	94.64 ±0.30	90.56 ±0.40	3.32 ±0.22	4.61 ±0.08	54.27 ±0.36	52.37 ±0.26	2.90 ±0.26	3.66 ±0.62
SSD	**100.00** ±0.00	100.00 ±0.00	94.29 ±0.13	**0.03** ±0.00	**2.34** ±0.11	59.50 ±0.89	59.51 ±0.44	3.88 ±0.30	4.90 ±0.44
SALUN	99.14 ±0.71	97.87 ±1.68	91.59 ±1.28	9.90 ±1.91	12.26 ±1.77	53.45 ±1.49	52.52 ±0.92	2.48 ±0.13	2.74 ±0.41
AMUN	99.29 ±0.20	**94.63** ±0.62	91.97 ±0.20	0.55 ±0.12	3.74 ±0.09	52.63 ±0.64	**49.75** ±0.48	**1.46** ±0.15	2.14 ±0.31
REGUN	98.42 ±1.01	96.68 ±1.95	90.60 ±1.26	17.85 ±1.92	19.67 ±1.74	**51.01** ±0.64	50.55 ±0.29	2.00 ±0.18	**2.01** ±0.67
FORGET 50%									
	$\text{RETAIN}_{\text{ACC}}$	$\text{FORGET}_{\text{ACC}}$	TEST_{ACC}	$\text{RETAIN}_{\text{DIV}}$	TEST_{DIV}	RMIA_{AUC}	SMIA_{AUC}	$\text{GAP}^{\text{RFTP}}_{\text{Avg}}$	$\text{GAP}^{\text{TP}}_{\text{Avg}}$
RETRAIN	100.00 ±0.00	90.89 ±0.52	90.31 ±0.41	0.00 ±0.00	0.00 ±0.00	50.24 ±0.36	50.25 ±0.29	0.00 ±0.00	0.00 ±0.00
BASE	100.00 ±0.00	100.00 ±0.00	94.17 ±0.02	0.04 ±0.00	4.34 ±0.16	56.42 ±0.32	59.39 ±0.47	4.79 ±0.30	5.02 ±0.35
NEGGRAD	**100.00** ±0.00	99.99 ±0.00	94.05 ±0.05	0.04 ±0.01	4.38 ±0.18	56.59 ±0.35	59.26 ±0.40	4.80 ±0.33	5.04 ±0.40
NEGGRAD+	96.69 ±2.04	92.11 ±4.11	88.62 ±2.17	2.20 ±1.28	5.96 ±0.71	53.19 ±2.38	51.99 ±1.70	2.62 ±0.58	2.41 ±0.50
FINETUNE	95.86 ±1.04	90.60 ±0.98	88.10 ±0.46	2.84 ±0.74	6.07 ±0.25	52.40 ±0.62	51.27 ±0.06	2.39 ±0.47	2.19 ±0.05
ℓ_1-SPARSE	96.86 ±0.86	91.87 ±1.32	88.82 ±0.75	2.20 ±0.59	5.60 ±0.41	52.81 ±0.69	51.38 ±0.47	2.09 ±0.10	2.03 ±0.26
LUR	96.76 ±1.30	91.15 ±1.29	88.98 ±0.71	2.28 ±0.83	5.57 ±0.24	52.24 ±0.96	51.13 ±0.54	1.77 ±0.11	1.66 ±0.18
SSD	100.00 ±0.00	100.00 ±0.00	94.18 ±0.01	**0.04** ±0.00	**4.34** ±0.16	56.42 ±0.32	59.39 ±0.47	4.79 ±0.30	5.02 ±0.34
SALUN	98.05 ±0.78	93.11 ±1.14	89.00 ±1.03	3.72 ±1.16	7.39 ±0.80	52.76 ±0.66	52.04 ±0.49	2.00 ±0.13	1.91 ±0.31
AMUN	97.60 ±1.64	**90.97** ±3.75	89.49 ±1.96	1.61 ±1.04	6.16 ±0.85	**51.02** ±1.88	**50.06** ±1.42	1.84 ±0.24	1.26 ±0.52
REGUN	98.77 ±0.20	93.44 ±0.36	**90.11** ±0.12	2.46 ±0.29	6.44 ±0.30	52.10 ±0.11	51.50 ±0.41	**1.48** ±0.08	**1.07** ±0.17

Table 2: Results for ResNet-18 on CIFAR-10 under 1%, 10%, and 50% random forgetting. **Bold** and underlined denote best and second best, where "best" is smallest gap to RETRAIN.

C.2 SWIN-T ON TINY-IMAGENET

	RETAIN$_{\text{ACC}}$	FORGET$_{\text{ACC}}$	TEST$_{\text{ACC}}$	RETAIN$_{\text{DIV}}$	TEST$_{\text{DIV}}$	RMIA$_{\text{AUC}}$	SMIA$_{\text{AUC}}$	GAP$^{\text{RFTP}}_{\text{AVG}}$	GAP$^{\text{TP}}_{\text{AVG}}$
FORGET 1%									
RETRAIN	99.99 ±0.00	61.48 ±1.23	60.89 ±0.15	0.00 ±0.00	0.00 ±0.00	49.79 ±1.38	50.50 ±0.69	0.00 ±0.00	0.00 ±0.0
BASE	99.99 ±0.00	99.96 ±0.06	61.24 ±0.07	0.49 ±0.01	7.24 ±0.03	87.77 ±0.18	94.68 ±0.10	19.21 ±0.41	19.15 ±0.69
NEGGRAD	**99.99** ±0.00	99.96 ±0.06	**61.22** ±0.06	**0.49** ±0.01	**7.24** ±0.03	87.78 ±0.18	94.68 ±0.09	19.20 ±0.42	19.15 ±0.72
NEGGRAD+	94.31 ±0.13	68.78 ±1.47	48.99 ±0.31	5.46 ±0.03	27.30 ±0.31	66.99 ±1.96	62.44 ±1.16	10.52 ±0.59	14.54 ±0.98
FINETUNE	97.11 ±0.19	63.70 ±1.74	52.31 ±0.36	3.78 ±0.13	27.51 ±0.12	62.48 ±0.22	57.26 ±0.66	6.60 ±0.59	10.63 ±0.92
ℓ_1-SPARSE	96.49 ±0.21	**61.26** ±1.41	51.94 ±0.03	4.32 ±0.11	26.67 ±0.09	62.37 ±0.96	56.86 ±1.06	6.54 ±0.33	10.76 ±1.02
SSD	82.26 ±16.32	67.74 ±17.77	41.99 ±7.57	18.18 ±9.22	18.21 ±4.88	68.08 ±4.21	67.13 ±4.30	17.63 ±4.28	18.59 ±1.07
SALUN	99.30 ±0.06	63.00 ±2.12	53.16 ±0.69	3.24 ±0.14	14.77 ±0.20	46.36 ±1.39	42.57 ±0.38	**3.73** ±0.20	5.59 ±0.80
AMUN	98.60 ±0.15	46.44 ±0.58	52.59 ±0.46	2.71 ±0.12	27.85 ±0.19	**51.03** ±0.57	**46.31** ±0.28	6.55 ±0.62	**4.91** ±0.48
REGUN	98.64 ±0.13	52.89 ±0.89	52.26 ±0.40	3.15 ±0.29	18.11 ±0.40	45.07 ±0.83	40.50 ±0.13	5.82 ±0.16	6.68 ±0.42
FORGET 10%									
	RETAIN$_{\text{ACC}}$	FORGET$_{\text{ACC}}$	TEST$_{\text{ACC}}$	RETAIN$_{\text{DIV}}$	TEST$_{\text{DIV}}$	RMIA$_{\text{AUC}}$	SMIA$_{\text{AUC}}$	GAP$^{\text{RFTP}}_{\text{AVG}}$	GAP$^{\text{TP}}_{\text{AVG}}$
RETRAIN	99.99 ±0.00	59.54 ±0.54	59.27 ±0.30	0.00 ±0.00	0.00 ±0.00	50.30 ±0.66	50.35 ±0.00	0.00 ±0.00	0.00 ±0.00
BASE	99.98 ±0.00	99.98 ±0.01	61.03 ±0.23	0.51 ±0.00	7.74 ±0.00	86.40 ±0.26	94.70 ±0.00	19.58 ±0.32	18.93 ±0.61
NEGGRAD	**99.98** ±0.00	99.98 ±0.01	**61.02** ±0.22	**0.51** ±0.00	**7.80** ±0.06	86.43 ±0.26	94.72 ±0.03	19.58 ±0.32	18.94 ±0.60
NEGGRAD+	89.85 ±0.55	53.03 ±0.69	46.49 ±0.31	8.11 ±0.33	29.54 ±0.25	57.18 ±0.11	54.47 ±0.41	9.07 ±0.48	9.83 ±0.48
FINETUNE	97.07 ±0.45	61.54 ±0.53	51.00 ±0.75	3.78 ±0.31	28.18 ±0.25	62.40 ±0.34	56.86 ±0.17	6.32 ±0.18	10.18 ±0.11
ℓ_1-SPARSE	95.61 ±1.09	**58.92** ±2.72	50.26 ±0.33	4.90 ±0.80	27.69 ±0.60	61.00 ±2.48	56.12 ±1.84	6.39 ±0.24	9.85 ±1.40
SSD	99.16 ±1.05	98.89 ±1.31	55.87 ±3.84	4.06 ±2.27	10.27 ±1.74	84.36 ±0.15	86.89 ±2.45	19.41 ±0.85	18.73 ±1.95
SALUN	91.53 ±0.67	64.36 ±0.80	49.88 ±0.24	14.05 ±0.31	15.20 ±0.41	55.77 ±0.80	52.21 ±0.50	7.03 ±0.24	7.43 ±0.31
AMUN	96.04 ±0.51	53.01 ±0.28	51.06 ±0.39	4.38 ±0.29	28.48 ±0.09	55.35 ±0.30	**51.92** ±0.26	5.93 ±0.15	6.63 ±0.36
REGUN	98.98 ±0.15	63.30 ±1.38	52.72 ±0.25	3.73 ±0.20	14.94 ±0.35	**49.86** ±0.83	45.54 ±0.09	**3.05** ±0.21	**3.70** ±0.54
FORGET 50%									
	RETAIN$_{\text{ACC}}$	FORGET$_{\text{ACC}}$	TEST$_{\text{ACC}}$	RETAIN$_{\text{DIV}}$	TEST$_{\text{DIV}}$	RMIA$_{\text{AUC}}$	SMIA$_{\text{AUC}}$	GAP$^{\text{RFTP}}_{\text{AVG}}$	GAP$^{\text{TP}}_{\text{AVG}}$
RETRAIN	99.99 ±0.00	48.34 ±0.20	47.95 ±0.12	0.00 ±0.00	0.00 ±0.00	50.30 ±0.19	50.24 ±0.00	0.00 ±0.00	0.00 ±0.00
BASE	99.99 ±0.00	99.99 ±0.00	61.20 ±0.20	0.70 ±0.00	13.71 ±0.00	79.74 ±0.05	94.70 ±0.00	23.58 ±0.04	21.34 ±0.03
NEGGRAD	**99.99** ±0.00	99.99 ±0.00	61.19 ±0.16	**0.73** ±0.01	**13.69** ±0.07	79.84 ±0.05	94.61 ±0.13	23.61 ±0.05	21.39 ±0.05
NEGGRAD+	96.71 ±0.58	48.92 ±2.99	43.63 ±0.98	3.61 ±0.33	33.60 ±0.25	56.66 ±2.25	54.14 ±1.32	4.12 ±0.18	5.34 ±0.70
FINETUNE	97.35 ±0.43	55.11 ±1.17	45.74 ±0.82	3.25 ±0.31	32.63 ±0.37	61.46 ±0.42	56.78 ±0.21	5.70 ±0.10	6.69 ±0.27
ℓ_1-SPARSE	94.25 ±0.27	47.29 ±0.32	43.48 ±0.53	5.67 ±0.26	31.63 ±0.24	55.23 ±0.35	53.12 ±0.04	4.05 ±0.30	4.70 ±0.27
SSD	77.45 ±7.00	75.75 ±7.79	38.10 ±2.80	26.11 ±7.36	20.82 ±0.52	74.84 ±3.49	74.77 ±2.40	21.09 ±0.59	17.19 ±0.91
SALUN	97.79 ±1.69	59.57 ±0.51	**47.77** ±0.30	10.74 ±4.65	15.80 ±0.03	58.04 ±0.95	55.38 ±0.13	5.36 ±0.05	4.01 ±0.43
AMUN	95.47 ±0.57	55.48 ±0.83	48.21 ±0.75	4.46 ±0.43	33.44 ±0.32	59.37 ±0.34	55.33 ±0.10	5.34 ±0.15	4.86 ±0.13
REGUN	99.97 ±0.01	**48.21** ±0.64	45.57 ±0.44	1.27 ±0.04	17.05 ±0.07	**47.88** ±0.19	**48.16** ±0.42	**1.37** ±0.09	**2.40** ±0.19

Table 3: Results for Swin-T on Tiny-ImageNet under 1%, 10%, and 50% random forgetting. **Bold** and underlined denote best and second best, where "best" is smallest gap to RETRAIN.

C.3 RESNET-18 ON CIFAR-100

	RETAIN$_{\text{ACC}}$	FORGET$_{\text{ACC}}$	TEST$_{\text{ACC}}$	RETAIN$_{\text{DIV}}$	TEST$_{\text{DIV}}$	RMIA$_{\text{AUC}}$	SMIA$_{\text{AUC}}$	GAP$^{\text{RFTP}}_{\text{AVG}}$	GAP$^{\text{TP}}_{\text{AVG}}$
FORGET 1%									
RETRAIN	99.98 ±0.01	74.81 ±1.22	75.33 ±0.29	0.00 ±0.00	0.00 ±0.00	49.56 ±1.56	48.73 ±0.78	0.00 ±0.00	0.00 ±0.00
BASE	99.98 ±0.01	100.00 ±0.00	75.52 ±0.36	0.14 ±0.00	7.77 ±0.06	76.09 ±1.32	75.22 ±0.05	12.84 ±1.28	13.47 ±0.68
NEGGRAD	99.02 ±0.84	96.37 ±1.45	72.35 ±1.90	1.16 ±0.83	10.07 ±1.26	73.79 ±0.43	68.36 ±2.31	12.26 ±1.39	13.44 ±1.30
NEGGRAD+	97.38 ±1.84	90.81 ±2.87	69.10 ±2.39	2.94 ±1.78	12.74 ±1.87	69.96 ±1.57	64.29 ±2.48	11.13 ±1.38	13.15 ±1.76
FINETUNE	91.98 ±3.58	75.63 ±4.69	66.67 ±1.27	5.46 ±2.31	15.31 ±0.51	59.76 ±3.26	54.46 ±2.58	7.04 ±1.15	9.27 ±1.75
ℓ_1-SPARSE	92.44 ±1.01	73.56 ±3.27	66.30 ±0.52	5.14 ±0.65	15.52 ±0.15	59.78 ±0.85	53.91 ±1.48	7.02 ±0.29	9.46 ±1.06
LUR	77.23 ±0.84	**74.44** ±2.04	61.41 ±0.66	27.55 ±0.96	23.25 ±0.99	56.72 ±0.85	55.27 ±1.24	11.25 ±0.38	10.38 ±0.48
SSD	**99.61** ±0.46	98.89 ±0.89	**75.12** ±0.48	**0.50** ±0.38	**8.25** ±0.52	75.78 ±1.32	74.09 ±1.19	12.63 ±1.22	13.23 ±1.13
SALUN	98.70 ±0.90	77.33 ±5.41	67.33 ±2.33	3.91 ±1.84	15.20 ±1.85	**51.18** ±0.85	45.64 ±1.25	3.29 ±0.83	**4.65** ±1.02
AMUN	96.09 ±3.71	76.74 ±10.38	70.43 ±4.79	2.79 ±2.49	12.77 ±3.79	57.91 ±4.97	**51.65** ±6.55	5.96 ±1.15	6.46 ±1.44
REGUN	99.53 ±0.03	76.89 ±0.97	69.23 ±0.60	2.51 ±0.17	13.80 ±0.33	46.91 ±0.98	41.39 ±0.77	**3.01** ±0.28	4.73 ±0.87
FORGET 10%									
	RETAIN$_{\text{ACC}}$	FORGET$_{\text{ACC}}$	TEST$_{\text{ACC}}$	RETAIN$_{\text{DIV}}$	TEST$_{\text{DIV}}$	RMIA$_{\text{AUC}}$	SMIA$_{\text{AUC}}$	GAP$^{\text{RFTP}}_{\text{AVG}}$	GAP$^{\text{TP}}_{\text{AVG}}$
RETRAIN	99.98 ±0.00	75.59 ±0.27	74.36 ±0.01	0.00 ±0.00	0.00 ±0.00	50.71 ±0.22	50.32 ±0.43	0.00 ±0.00	0.00 ±0.00
BASE	99.98 ±0.00	99.99 ±0.01	75.67 ±0.17	0.14 ±0.00	8.42 ±0.14	73.97 ±0.42	75.41 ±0.38	12.24 ±0.20	12.28 ±0.27
NEGGRAD	99.86 ±0.12	99.74 ±0.12	74.55 ±0.77	0.31 ±0.12	9.08 ±0.26	74.35 ±0.84	73.83 ±0.60	12.11 ±0.21	12.08 ±0.28
NEGGRAD+	99.86 ±0.11	98.79 ±0.65	**74.23** ±0.87	0.33 ±0.13	9.41 ±0.32	73.72 ±0.51	72.09 ±0.80	11.74 ±0.15	11.81 ±0.39
FINETUNE	92.85 ±0.84	76.88 ±0.86	65.57 ±0.36	4.90 ±0.49	16.11 ±0.39	61.33 ±0.45	55.88 ±0.33	6.96 ±0.12	9.70 ±0.17
ℓ_1-SPARSE	92.64 ±0.60	76.45 ±1.03	65.55 ±0.24	5.02 ±0.36	16.17 ±0.13	61.02 ±0.36	55.79 ±0.50	6.83 ±0.16	9.55 ±0.12
LUR	93.96 ±0.58	**75.25** ±0.40	69.48 ±0.17	10.30 ±0.54	13.29 ±0.29	55.04 ±0.69	52.51 ±0.36	**3.91** ±0.23	4.60 ±0.39
SSD	**99.98** ±0.00	99.99 ±0.01	75.67 ±0.17	**0.14** ±0.00	**8.42** ±0.14	73.97 ±0.42	75.41 ±0.38	12.24 ±0.20	12.28 ±0.27
SALUN	98.55 ±1.42	89.81 ±3.97	67.74 ±2.40	8.98 ±3.13	18.41 ±1.82	55.26 ±0.83	52.97 ±0.94	6.71 ±0.19	5.59 ±1.56
AMUN	92.81 ±1.05	63.05 ±1.45	65.73 ±0.88	4.93 ±0.69	16.30 ±0.50	**50.75** ±0.15	47.23 ±0.25	7.11 ±0.79	**4.36** ±0.50
REGUN	97.79 ±1.48	83.61 ±5.07	66.55 ±2.21	10.99 ±3.98	19.64 ±2.60	50.10 ±1.41	**48.85** ±0.52	4.80 ±0.08	4.50 ±1.51
FORGET 50%									
	RETAIN$_{\text{ACC}}$	FORGET$_{\text{ACC}}$	TEST$_{\text{ACC}}$	RETAIN$_{\text{DIV}}$	TEST$_{\text{DIV}}$	RMIA$_{\text{AUC}}$	SMIA$_{\text{AUC}}$	GAP$^{\text{RFTP}}_{\text{AVG}}$	GAP$^{\text{TP}}_{\text{AVG}}$
RETRAIN	99.99 ±0.00	65.81 ±0.10	65.77 ±0.75	0.00 ±0.00	0.00 ±0.00	50.43 ±0.43	49.86 ±0.34	0.00 ±0.00	0.00 ±0.00
BASE	99.98 ±0.01	99.98 ±0.00	75.69 ±0.15	0.16 ±0.01	14.22 ±0.28	67.83 ±0.36	75.29 ±0.18	15.38 ±0.05	13.66 ±0.07
NEGGRAD	99.98 ±0.01	99.97 ±0.02	75.33 ±0.18	0.17 ±0.01	14.23 ±0.38	67.94 ±0.23	75.08 ±0.26	15.31 ±0.07	13.53 ±0.14
NEGGRAD+	91.48 ±1.63	66.63 ±1.31	60.85 ±1.19	5.84 ±0.97	20.02 ±1.00	56.68 ±0.33	53.38 ±0.36	5.21 ±0.78	5.58 ±1.04
FINETUNE	92.29 ±0.63	70.71 ±0.83	61.64 ±0.59	5.38 ±0.35	19.96 ±0.51	59.43 ±0.56	55.09 ±0.19	6.43 ±0.48	6.56 ±0.84
ℓ_1-SPARSE	92.46 ±1.48	69.83 ±2.19	62.60 ±0.87	5.37 ±0.92	19.19 ±0.02	58.05 ±1.46	54.32 ±0.73	5.58 ±0.57	5.39 ±0.89
LUR	95.04 ±0.80	66.88 ±3.28	**62.75** ±1.64	4.62 ±1.09	17.93 ±0.64	**54.45** ±2.24	52.35 ±1.17	3.71 ±0.16	**3.52** ±0.36
SSD	**99.98** ±0.01	99.98 ±0.00	75.68 ±0.15	**0.16** ±0.01	**14.22** ±0.28	67.83 ±0.36	75.29 ±0.18	15.37 ±0.05	13.65 ±0.07
SALUN	95.69 ±1.61	69.44 ±2.37	59.02 ±1.63	9.32 ±1.23	20.93 ±0.59	56.53 ±0.32	54.02 ±0.24	5.19 ±0.53	6.42 ±1.45
AMUN	92.16 ±0.97	**65.74** ±0.44	62.45 ±0.81	5.34 ±0.56	21.23 ±0.02	55.05 ±0.55	**51.51** ±0.49	4.04 ±0.46	3.97 ±0.51
REGUN	97.91 ±0.72	68.84 ±1.62	61.55 ±0.51	5.30 ±0.79	19.36 ±0.77	55.16 ±0.83	51.93 ±0.47	**3.51** ±0.48	4.47 ±0.99

Table 4: Results for ResNet-18 on CIFAR-100 under 1%, 10%, and 50% random forgetting. **Bold** and underlined denote best and second best, where "best" is smallest gap to RETRAIN.

C.4 RESNET-18 ON TINY-IMAGENET

	RETAIN$_{\text{ACC}}$	FORGET$_{\text{ACC}}$	TEST$_{\text{ACC}}$	RETAIN$_{\text{DIV}}$	TEST$_{\text{DIV}}$	RMIA$_{\text{AUC}}$	SMIA$_{\text{AUC}}$	GAP$^{\text{RFTP}}_{\text{AVG}}$	GAP$^{\text{TP}}_{\text{AVG}}$
FORGET 1%									
RETRAIN	99.98 ±0.00	58.52 ±0.71	59.54 ±0.21	−0.00 ±0.00	0.00 ±0.00	49.28 ±1.14	49.84 ±0.85	0.00 ±0.00	0.00 ±0.00
BASE	99.98 ±0.00	99.96 ±0.06	59.64 ±0.44	0.17 ±0.00	14.88 ±0.09	81.56 ±0.56	84.68 ±0.16	18.50 ±0.61	16.27 ±0.90
NEGGRAD	**99.98** ±0.00	99.72 ±0.24	**58.85** ±0.60	**0.22** ±0.06	15.70 ±0.12	81.44 ±0.69	83.66 ±0.09	18.29 ±0.67	16.11 ±1.07
NEGGRAD+	95.68 ±0.05	67.22 ±2.20	49.40 ±0.67	5.49 ±0.16	22.05 ±0.12	62.23 ±1.09	59.47 ±1.74	8.80 ±0.59	11.23 ±0.45
FINETUNE	89.92 ±6.35	61.67 ±8.33	51.55 ±0.69	7.64 ±4.63	22.61 ±0.67	63.06 ±7.81	57.23 ±5.13	9.13 ±1.00	10.57 ±3.13
ℓ_1-SPARSE	88.23 ±4.10	**59.70** ±1.66	51.87 ±0.16	8.91 ±3.15	22.02 ±1.09	60.21 ±3.75	55.48 ±1.68	7.90 ±0.39	9.30 ±1.92
LUR	3.94 ±0.16	3.59 ±0.28	3.77 ±0.18	65.32 ±0.09	52.70 ±0.21	**49.14** ±0.32	**49.25** ±0.69	51.91 ±0.38	28.33 ±0.29
SSD	99.39 ±0.01	98.89 ±0.31	58.72 ±0.10	0.68 ±0.06	**15.32** ±0.03	80.69 ±0.60	83.87 ±0.13	18.08 ±0.50	15.80 ±0.78
SALUN	98.39 ±0.31	54.89 ±2.36	50.36 ±0.16	4.42 ±0.27	22.29 ±0.33	49.46 ±2.00	46.99 ±0.60	4.11 ±0.98	5.48 ±0.24
AMUN	96.95 ±2.82	68.74 ±2.53	52.44 ±1.14	2.62 ±1.88	22.29 ±1.21	64.41 ±2.00	61.70 ±1.56	8.87 ±0.30	11.11 ±0.79
REGUN	99.74 ±0.01	60.22 ±2.51	52.37 ±0.77	1.89 ±0.45	20.83 ±1.06	49.52 ±0.05	46.99 ±0.17	**2.32** ±0.30	**3.79** ±0.26
FORGET 10%									
	RETAIN$_{\text{ACC}}$	FORGET$_{\text{ACC}}$	TEST$_{\text{ACC}}$	RETAIN$_{\text{DIV}}$	TEST$_{\text{DIV}}$	RMIA$_{\text{AUC}}$	SMIA$_{\text{AUC}}$	GAP$^{\text{RFTP}}_{\text{AVG}}$	GAP$^{\text{TP}}_{\text{AVG}}$
RETRAIN	99.99 ±0.00	58.57 ±0.37	58.30 ±0.29	−0.00 ±0.00	0.00 ±0.00	49.97 ±0.49	50.22 ±0.29	0.00 ±0.00	0.00 ±0.00
BASE	99.98 ±0.00	99.99 ±0.01	59.33 ±0.39	0.17 ±0.00	15.55 ±0.13	80.76 ±0.55	84.77 ±0.13	18.31 ±0.10	15.91 ±0.21
NEGGRAD	**99.98** ±0.00	99.98 ±0.01	59.53 ±0.08	0.18 ±0.00	15.60 ±0.10	80.49 ±0.17	84.43 ±0.01	18.32 ±0.18	16.02 ±0.22
NEGGRAD+	96.16 ±0.00	79.31 ±0.00	50.96 ±0.00	4.57 ±0.00	21.52 ±0.00	72.14 ±0.00	67.21 ±0.00	13.60 ±0.00	14.92 ±0.00
FINETUNE	84.26 ±1.37	56.58 ±0.72	50.15 ±0.02	11.65 ±1.04	22.43 ±0.40	57.32 ±0.49	54.30 ±0.26	8.35 ±0.35	7.74 ±0.48
ℓ_1-SPARSE	88.98 ±2.21	**59.18** ±1.24	50.41 ±0.54	8.38 ±1.69	22.89 ±0.73	60.24 ±2.37	55.78 ±1.00	7.56 ±0.18	9.08 ±1.25
LUR	21.50 ±1.25	19.50 ±0.76	19.37 ±0.90	55.65 ±0.75	41.63 ±0.74	**49.84** ±0.59	**49.89** ±0.25	39.25 ±0.61	19.72 ±0.37
SSD	99.98 ±0.00	99.99 ±0.00	**59.40** ±0.00	**0.17** ±0.00	**15.41** ±0.00	80.31 ±0.00	84.70 ±0.00	18.41 ±0.00	16.12 ±0.00
SALUN	99.43 ±0.00	83.10 ±0.00	51.55 ±0.00	5.50 ±0.00	22.61 ±0.00	60.61 ±0.00	58.89 ±0.00	10.70 ±0.00	8.86 ±0.00
AMUN	96.90 ±3.32	71.54 ±3.13	52.53 ±1.43	2.60 ±2.21	23.15 ±0.94	66.67 ±2.74	63.58 ±2.10	9.63 ±0.50	11.23 ±0.79
REGUN	98.69 ±0.88	67.28 ±0.97	51.13 ±0.85	8.95 ±2.60	23.69 ±0.49	48.60 ±2.01	49.55 ±1.57	**4.51** ±0.20	**4.12** ±0.23
FORGET 50%									
	RETAIN$_{\text{ACC}}$	FORGET$_{\text{ACC}}$	TEST$_{\text{ACC}}$	RETAIN$_{\text{DIV}}$	TEST$_{\text{DIV}}$	RMIA$_{\text{AUC}}$	SMIA$_{\text{AUC}}$	GAP$^{\text{RFTP}}_{\text{AVG}}$	GAP$^{\text{TP}}_{\text{AVG}}$
RETRAIN	99.99 ±0.00	49.35 ±0.11	49.33 ±0.38	−0.00 ±0.00	0.00 ±0.00	50.14 ±0.33	50.03 ±0.20	0.00 ±0.00	0.00 ±0.00
BASE	99.98 ±0.00	99.99 ±0.00	59.46 ±0.15	0.18 ±0.01	21.35 ±0.12	76.03 ±0.05	84.77 ±0.27	21.67 ±0.16	18.01 ±0.25
NEGGRAD	0.47 ±0.04	0.45 ±0.05	0.46 ±0.01	68.98 ±0.03	68.47 ±0.05	**49.93** ±0.01	49.35 ±0.29	49.44 ±0.05	24.64 ±0.07
NEGGRAD+	91.61 ±1.41	**50.10** ±1.07	44.89 ±1.41	6.21 ±0.83	29.36 ±0.74	56.71 ±0.29	53.99 ±0.15	5.11 ±0.37	5.50 ±0.44
FINETUNE	92.68 ±1.75	55.26 ±1.50	46.09 ±0.95	5.47 ±1.03	29.35 ±0.60	61.05 ±0.62	56.78 ±0.27	6.82 ±0.07	7.06 ±0.26
ℓ_1-SPARSE	88.03 ±0.91	50.64 ±0.66	44.74 ±0.54	8.95 ±0.64	28.22 ±0.46	57.27 ±0.11	54.41 ±0.26	6.24 ±0.34	5.86 ±0.55
LUR	52.18 ±1.62	33.52 ±0.70	33.45 ±0.72	35.84 ±0.97	26.64 ±0.66	50.56 ±0.39	**50.35** ±0.22	19.99 ±0.70	8.15 ±0.21
SSD	**99.99** ±0.00	99.99 ±0.00	59.39 ±0.04	**0.18** ±0.01	**21.38** ±0.15	76.05 ±0.07	84.71 ±0.36	21.59 ±0.05	17.88 ±0.11
SALUN	98.85 ±0.22	58.70 ±0.75	45.53 ±0.38	5.36 ±0.60	23.66 ±0.39	61.14 ±0.17	57.49 ±0.09	6.32 ±0.21	7.40 ±0.17
AMUN	95.79 ±3.61	73.10 ±7.62	**51.22** ±3.67	3.23 ±2.57	28.74 ±2.31	66.60 ±3.47	65.55 ±4.32	11.78 ±2.72	9.59 ±3.39
REGUN	98.20 ±0.92	54.61 ±11.64	43.34 ±1.07	14.17 ±13.18	27.73 ±5.03	55.16 ±0.58	55.81 ±4.96	**5.01** ±2.51	**5.50** ±0.44

Table 5: Results for ResNet-18 on Tiny-ImageNet under 1%, 10%, and 50% random forgetting. **Bold** and underlined denote best and second best, where "best" is smallest gap to RETRAIN.