NOT ALL SAMPLES LEAVE EQUAL IMPRINTS: INFLUENCE-AWARE UNLEARNING IN CONTINUAL LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Machine unlearning aims to remove the influence of specific data from a trained model, motivated by data regulations such as the right to be forgotten. Existing methods implicitly assume that all samples are encoded with similar strength, ignoring differences in their imprint on the model, and therefore apply uniform unlearning updates. This assumption breaks in class-incremental learning (CIL), where training data leave highly non-uniform imprints on the model. As a result, uniform unlearning leads to over-removal of weakly encoded samples and insufficient forgetting of strongly encoded ones, degrading overall performance. To address this, we propose `IMPRINT`, an imprint-aware unlearning framework for CIL. Our key idea is to align unlearning intensity with sample-wise imprint, estimated via the logit margin as a simple yet effective proxy. This enables adaptive scaling of the forget loss and seamless plug-and-play integration with existing unlearning methods. Extensive experiments show that `IMPRINT` consistently improves both forgetting and retention performance in CIL settings, significantly reducing the gap to retraining while outperforming state-of-the-art baselines.

1 INTRODUCTION

Continual learning has gained significant attention as a paradigm for lifelong learning, where models continuously acquire new knowledge while maintaining performance on previously learned tasks. To mitigate catastrophic forgetting (McCloskey & Cohen, 1989; Goodfellow et al., 2013), rehearsal-based methods (Rolnick et al., 2019; Rebuffi et al., 2017) store and replay past samples during training. However, such mechanisms do not preserve all past knowledge equally. In class-incremental learning (CIL) settings, models often exhibit imbalanced performance across tasks and classes: samples from recently learned classes are retained more strongly, while those from earlier tasks become progressively weakened as new tasks arrive. As illustrated in Figure 1, continual learning model exhibits low accuracy in earlier classes, whereas the jointly learned model shows similar accuracies between different classes. As a result, different samples leave varying degrees of *imprint* on the model parameters, leading to heterogeneous influence on model predictions.

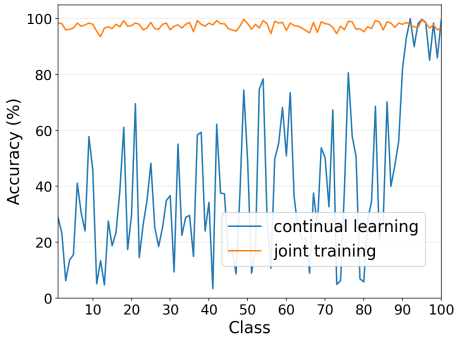

Figure 1: Per-class accuracy on CIFAR-100 after continual learning versus joint training. The continually trained model tends to exhibit relatively lower accuracy on earlier-task classes (i.e., lower class indices) compared to recent ones, leading to heterogeneous imprints. Uniform unlearning over such non-uniform imprints can over-remove weakly encoded samples and degrade the retain set performance.

This heterogeneity in CIL raises an important question about controlling the influence of individual samples. In practice, it is desirable to remove the effect of specific data from a trained model due to privacy regulations and data governance requirements such as GDPR (Mantelero, 2013). This motivates the problem of machine unlearning (MU) (Cao & Yang, 2015; Ginart et al., 2019; Bourtoule et al., 2021), which aims to eliminate the influence of a designated forget set while preserving

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.

performance on the remaining data. A simple approach is random labeling (RL) (Golatkar et al., 2020), which assigns incorrect labels to the forget data. More advanced methods, such as SalUn (Fan et al., 2024), restrict updates to parameters that are most relevant to the forget set. Other methods (e.g., AMUN (Ebrahimpour-Boroojeny et al., 2025)) explore alternative strategies for effective unlearning.

Challenges. Despite the recent success of unlearning methods, they predominantly operate under the assumption of *uniform contribution*, treating every sample in the forget set as if it were encoded with equal intensity. Consequently, they apply uniform updates regardless of how strongly each sample is encoded in the model. However, this assumption breaks in CIL, where data leave highly heterogeneous imprints on the model. Applying uniform unlearning in such settings leads to a fundamental mismatch between sample imprint and update intensity: weakly encoded samples are over-removed, degrading the retain performance, while strongly encoded samples are insufficiently removed, resulting in incomplete forgetting. This mismatch limits the effectiveness of existing unlearning methods when applied to continually trained models in CIL settings. We thus pose the following question:

Should all samples be treated equally during machine unlearning, given the heterogeneous imprints induced by continual learning?

Although recent studies have begun to explore non-uniform unlearning by treating samples differently (e.g., FaLW (Yu et al., 2026), IMU (Fan et al., 2025)), these approaches do not explicitly account for the heterogeneous imprints induced by CIL. FaLW assumes a long-tailed setting and relies on auxiliary validation data, while IMU incurs substantial computational overhead when estimating sample influence. As a result, their applicability to continual learning scenarios remains limited. We later provide a detailed discussion and empirical comparison between our approach and IMU.

Contributions. We propose `IMPRINT`, an imprint-aware unlearning framework for continual learning that adapts unlearning intensity based on how strongly each sample is encoded in the model. Instead of applying uniform updates to the forget set, `IMPRINT` assigns stronger updates to strongly imprinted samples and weaker updates to weakly imprinted ones, mitigating both over-removal and incomplete forgetting. A key challenge is to estimate sample-wise imprint efficiently. To this end, we identify the *logit margin* as a simple yet effective proxy, as it captures both correctness and confidence with respect to the model's decision boundary. Intuitively, samples with large positive margins are strongly encoded, whereas those with small or negative margins are weakly represented or already forgotten. Building on this insight, `IMPRINT` introduces a plug-and-play weighting mechanism that aligns update intensity with sample-wise imprint. Our approach is model-agnostic and can be seamlessly integrated into existing unlearning methods with minimal overhead. Extensive experiments in CIL settings on CIFAR-100, Tiny-ImageNet, and ImageNet with ResNet and ViT models demonstrate that `IMPRINT` consistently improves both forgetting and retention performance, outperforming existing baselines such as SalUn and AMUN. Our contributions are summarized as follows:

- We identify that continual learning induces heterogeneous sample imprints, which fundamentally violates the uniform contribution assumption in existing unlearning methods.
- We show that the logit margin serves as an effective and practical proxy for estimating sample-wise imprint without requiring access to training history.
- We propose `IMPRINT`, an imprint-aware unlearning framework that aligns update intensity with sample imprint, improving both forgetting and retention performance across diverse benchmarks.

2 Related Works

Machine unlearning. Early approaches focused on exact unlearning via partial retraining (Bourtoule et al., 2021) without the forget set, but this remains computationally expensive. This has led to approximate unlearning methods that directly modify model parameters, for example by identifying and updating parameters that are most relevant to the forget data through pruning or scrubbing (Golatkar et al., 2020; Foster et al., 2024). Other approaches achieve unlearning via knowledge distillation (Chundawat et al., 2023; Kurmanji et al., 2023) or by optimizing the model with corrupted supervision on the forget set (Thudi et al., 2022; Chen et al., 2023; Cha et al., 2024). Despite their effectiveness, most existing methods implicitly assume that samples are encoded with similar strength and thus apply uniform unlearning updates. This assumption breaks in CIL, where samples leave highly heterogeneous imprints on the model, limiting the effectiveness of existing approaches.

Beyond uniform unlearning. Recent studies have begun to explore non-uniform machine unlearning. FaLW (Yu et al., 2026) proposes an instance-wise reweighting scheme tailored for long-tailed forget sets. It measures a sample's influence by evaluating how much it behaves like an unseen example from the same class. Therefore, it requires auxiliary validation data and the performance depends on the quality of the validation set. IMU (Fan et al., 2025) approximates sample influence via classifier-level gradients, but still incurs substantial runtime cost due to Jacobian computations. NPO (Zhang et al., 2024) applies uniform unlearning at the initial stage of unlearning, and adjusts later updates using prediction changes relative to a reference model. However, these methods do not explicitly account for the heterogeneous imprints induced by CIL. RUM (Zhao et al., 2024) partitions the forget set by per-sample memorization scores and assigns a different unlearning algorithm to each subgroup, but the required memorization scores rely on costly leave-one-out style estimation. In contrast, our method leverages the logit margin as an efficient proxy for sample-wise imprint, enabling adaptive unlearning without additional data, pretraining requirements, or significant computational overhead.

Class-incremental learning (CIL). CIL (Li & Hoiem, 2017; Wang et al., 2022b; Wu et al., 2019) aims to learn from a sequence of tasks while mitigating catastrophic forgetting. Rehearsal-based methods address this by storing a subset of past samples in a replay buffer and interleaving them with new data during training. Experience Replay (Rolnick et al., 2019) is a representative approach that mixes buffered and current samples. iCaRL (Rebuffi et al., 2017) stores exemplars and combines cross-entropy with knowledge distillation from previous models. DER++ (Buzzega et al., 2020) extends this by storing both labels and logits for replayed samples. FOSTER (Wang et al., 2022a) temporarily expands the model to learn better in new classes, and then compresses it back to the original size. In this work, we study machine unlearning in models trained via continual learning, where heterogeneous sample imprints naturally arise across tasks.

Continual unlearning. Prior work has explored unlearning within continual learning, but under settings that differ fundamentally from ours. LSF (Shibata et al., 2021) performs selective class forgetting *during* continual training. CLPU (Liu et al., 2022) achieves exact task-level unlearning by maintaining a separate sub-network per forgettable task. ErrorEraser (Cao et al., 2025) unlearns data biases during continual learning to improve performance. In Hu et al. (2026), task-wise update and curvature information for Hessian-based task-level deletion hase been proposed. These approaches either require forgettable knowledge to be handled explicitly during continual training, or define the forget set according to a task-specific objective rather than arbitrary user-specified samples. In contrast, we address post-hoc approximate unlearning of arbitrary user-specified forget sets on a model already trained via standard CIL, explicitly accounting for the heterogeneous sample imprints induced by sequential training.

3 Imprint-Aware Unlearning in Continual Learning

3.1 Problem Setup and Preliminaries

Class-incremental learning (CIL). We consider a CIL setting where training data arrive sequentially as a series of tasks rather than being available all at once. Naively training a model in this setting leads to catastrophic forgetting. Rehearsal-based methods mitigate this issue by maintaining a replay buffer of past samples. Let $\{\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_n\}$ denote a sequence of tasks. After completing training on task $\mathcal{T}_i$, a subset of samples from $\mathcal{T}_i$ is stored in a replay buffer $\mathcal{M}_i$. During training on task $\mathcal{T}_{i+1}$, the model is trained jointly on samples from $\mathcal{M}_i$ and the current task $\mathcal{T}_{i+1}$.

Unlearning for CIL. Given a model θ_o trained under the CIL setting, our goal is to obtain an unlearned model θ_u that removes the influence of a designated forget set D_f. The forget set $D_f \subset \bigcup_{i=1}^{n} \mathcal{T}_i$ may include samples from multiple tasks or classes encountered during training. The retain set is defined as $D_r^{\text{avail}} = \mathcal{M}_n \setminus D_f$, whereas the full retain set used to define the retraining reference would be $D_r^{\text{full}} = (\bigcup_{i=1}^{n} \mathcal{T}_i) \setminus D_f$. Importantly, the forget set may consist of any previously observed samples (e.g., due to user deletion requests), whereas the retain set is restricted to the data in the replay buffer. This asymmetry makes unlearning in CIL particularly challenging. The objective of unlearning is to produce a model that behaves as if it had never been trained on D_f. The ideal solution is to retrain the model from scratch using D_r^{full} via CIL, yielding a reference model $\theta_r = \text{CIL}(\mathcal{T}_1 \setminus D_f, \ldots, \mathcal{T}_n \setminus D_f)$. However, such retraining is computationally expensive for large-scale models and datasets. Therefore, our goal is to design an efficient unlearning operator U that directly transforms θ_o into $\theta_u = U(\theta_o)$, such that θ_u approximates the behavior of θ_r.

3.2 Limitations of Uniform Unlearning

Prior unlearning methods are typically evaluated on jointly trained models, where samples are encoded with relatively uniform strength. In contrast, CIL induces highly heterogeneous sample imprints across tasks, as illustrated in Figure 1. Therefore, the uniform unlearning strategy adopted by prior methods leads to a fundamental mismatch between sample imprint and update intensity: weakly encoded samples are over-removed, introducing unnecessary parameter perturbations that degrade retain performance, while strongly encoded samples are insufficiently removed.

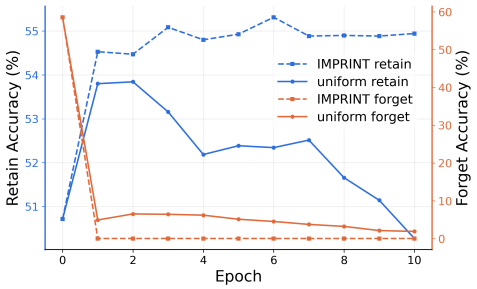

Figure 2: Comparison between uniform unlearning and IMPRINT on CIFAR-100, using random labeling. Uniform unlearning treats all forget samples equally regardless of their imprints, leading to degraded retain accuracy.

Figure 2 further supports this phenomenon. While existing unlearning methods perform well in jointly trained settings, their retain accuracy drops significantly in CIL models. These observations highlight the need for imprint-aware unlearning that adapts update intensity to sample-wise imprint in continual learning. This naturally raises a key question: how can we efficiently estimate the imprint of each sample on the model?

3.3 Logit Margin as a Proxy for Sample-wise Imprint

Limitations of accuracy and entropy as proxies. A straightforward approach to estimate sample-wise imprint is to use class accuracy. However, class accuracy does not capture how strongly individual samples are encoded in the model. An alternative is to use the model's output entropy: $\mathcal{H}(x) = -\sum_{c\in\mathcal{C}} \left(\frac{\exp(z_c(x))}{\sum_{j=1}^{|\mathcal{C}|}\exp(z_j(x))}\right) \log\left(\frac{\exp(z_c(x))}{\sum_{j=1}^{|\mathcal{C}|}\exp(z_j(x))}\right)$, where $\mathcal{C}$ denotes the set of classes and $z_c(x)$ is the logit of class c for input x. Entropy reflects the model's confidence, with low entropy indicating high confidence and high entropy indicating uncertainty. However, entropy does not account for whether the model's prediction is correct. As illustrated in Table 1, two samples can exhibit identical entropy while corresponding to fundamentally different outcomes: one correctly classified and one misclassified. Applying the same unlearning intensity to both cases can therefore lead to over-removal of already weakly encoded samples and degrade overall performance.

Table 1: Softmax output of two different samples. Identical entropy can correspond to different unlearning outcomes.

Probabilities	Sample A	Sample B
Class 1 (Target)	0.1	**0.7**
Class 2	**0.7**	0.1
Class 3	0.1	0.1
Class 4	0.1	0.1
Entropy (H)	0.9	0.9
Prediction	Incorrect	Correct

Logit margin as an imprint-aware measure. To address these limitations, we use logit margin as a proxy for sample-wise imprint. For an input x with ground-truth label y, the logit margin is defined as

$$\text{Margin}(x) = z_y - \max_{j\in\mathcal{C}\setminus\{y\}} z_j, \tag{1}$$

where z_y denotes the logit of the target class and $\max_{j\in\mathcal{C}\setminus\{y\}} z_j$ represents the largest logit among incorrect classes. Unlike class accuracy, the logit margin provides a sample-level measure of how strongly the model encodes a sample. A large positive margin indicates that the model predicts the correct label with high confidence, while a negative margin indicates that the model assigns higher confidence to an incorrect class. Moreover, unlike entropy, the logit margin explicitly reflects whether the prediction is correct. Revisiting the scenarios in Table 1, it is seen that the logit margin successfully distinguishes between the two cases: Scenario A yields a negative margin, whereas Scenario B yields a positive margin. This demonstrates that logit margin captures critical information that entropy fails to identify.

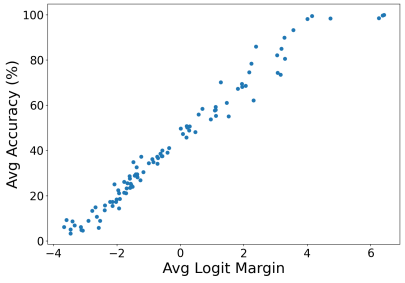

Figure 3: Per-class average logit margin vs. accuracy on CIFAR-100 before unlearning. Each point denotes a class. Both quantities are highly correlated.

Validation of logit margin. Figure 3 shows a strong positive correlation between class accuracy and the average logit margin. Classes with lower accuracy tend to exhibit smaller (often negative) margins, indicating that the model assigns higher confidence to incorrect classes. This observation

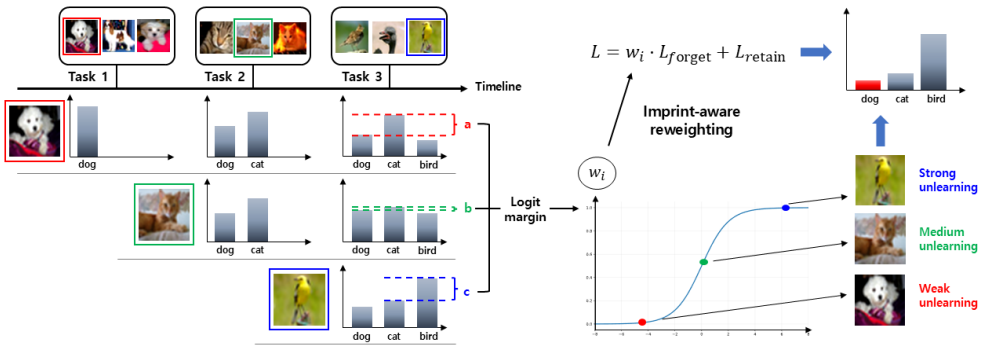

Figure 4: Overview of IMPRINT. After class-incremental learning on a sequence of tasks, training samples leave heterogeneous imprints on the model: a sample from an earlier task (a) typically yields a small or negative logit margin, a mid-task sample (b) yields a moderate margin, and a sample from the most recent task (c) yields a large positive margin. IMPRINT estimates each sample's imprint via its logit margin and converts it into an adaptive weight through a sigmoid function, which scales the per-sample forget loss accordingly. Its simplicity enables seamless integration with existing strategies.

supports the use of logit margin as a proxy for sample-wise imprint: samples with large positive margins are strongly encoded, whereas those with small or negative margins are weakly represented or already forgotten.

Relation to IMU influence and its interpretation. To further validate logit margin as a meaningful proxy for sample-wise imprint, we examine its relation to influence-based measure of how individual samples affect the model. We compare logit margin with the classifier-level influence estimate used by IMU. Following the IMU implementation, the influence estimate is computed from final-classifier gradients using a damped diagonal-Fisher approximation. We measure the agreement between two rankings using Spearman's rank correlation, where ρ quantifies their rank-order association and the corresponding p-value assesses its statistical significance.

Table 2: Correlation of margin with IMU influence.

Comparison	Spearman ρ	p
Margin vs. IMU influence	**0.728**	1.07×10^{-165}

As shown in Table 2, logit margin exhibits a strong correlation with IMU's influence ranking ($\rho = 0.728$), providing independent evidence that it captures sample-level information closely related to a sample's imprint on the model. However, the two measures capture fundamentally different quantities: IMU estimates gradient-based influence on the forget objective, whereas logit margin directly measures how strongly a sample is encoded in the current model. This distinction motivates our use of logit margin as a simple, prediction-based imprint measure rather than an approximation of influence. A detailed comparison and further mechanistic interpretation are provided in Appendix B.

3.4 IMPRINT: A PLUG-AND-PLAY FRAMEWORK FOR IMPRINT-AWARE UNLEARNING

We introduce IMPRINT, a general plug-and-play framework for CIL that adapts unlearning intensity based on sample-wise imprint. An overview of our approach is illustrated in Figure 4.

Imprint-aware weighting. We estimate the imprint of each sample using the logit margin and convert it into an adaptive weight via a sigmoid function:

$$\omega_i = \sigma\left(\frac{\text{Margin}(x_i)}{\tau}\right) = \frac{1}{1+\exp\left(-\frac{\text{Margin}(x_i)}{\tau}\right)}, \tag{2}$$

where τ is a temperature parameter that controls the smoothness of the weighting. This formulation assigns larger weights to strongly imprinted samples and smaller weights to weakly imprinted ones. IMPRINT integrates this weighting mechanism into the forget loss of existing unlearning methods. Given a base unlearning objective $\mathcal{L}$, we apply sample-wise weighting to adaptively scale the contribution of each forget sample, aligning update intensity with its imprint.

Integration with SalUn. SalUn selects a subset of salient parameters and applies random labeling on the forget set. Let $m \in \{0,1\}^{|\theta|}$ denote the saliency mask SalUn identifies from gradients on $\mathcal{D}_f$, and let $\tilde{\theta}_u = \theta_o + m \odot \Delta\theta$ denote the corresponding masked model, where $\Delta\theta$ are the trainable updates and $\odot$ is element-wise multiplication. Incorporating IMPRINT yields:

$$\min_{\Delta\theta} \mathcal{L}(\tilde{\theta}_u) := \mathbb{E}_{(x_i,y_i)\sim\mathcal{D}_f, y'\neq y_i}\left[\omega_i \cdot \ell_{\text{CE}}(\tilde{\theta}_u; x_i, y')\right] + \alpha\mathbb{E}_{(x,y)\sim\mathcal{D}_r}\left[\ell_{\text{CE}}(\tilde{\theta}_u; x, y)\right], \tag{3}$$

where α is a regularization parameter on the retain loss that prevents the model from forgetting retain data. Setting $m = \mathbf{1}$ in equation 3 (i.e., updating all parameters without saliency restriction) recovers the imprint-weighted random labeling (RL+IMPRINT) objective as a special case.

Integration with AMUN. AMUN augments the forget set with adversarial examples and updates all parameters directly, so the unlearned model is $\theta_u = \theta_o + \Delta\theta$. Applying IMPRINT leads to:

$$\min_{\Delta\theta} \mathcal{L}(\theta_u) := \mathbb{E}_{(x_i,y_i)\sim\mathcal{D}_f}\left[\omega_i \cdot \left(\ell_{\text{CE}}(\theta_u; x_i, y_i) + \ell_{\text{CE}}(\theta_u; x_i^{\text{adv}}, y_i^{\text{adv}})\right)\right] + \alpha\mathbb{E}_{(x,y)\sim\mathcal{D}_r}\left[\ell_{\text{CE}}(\theta_u; x, y)\right], \tag{4}$$

where x_i^{adv} is the adversarial example, $y_i^{\text{adv}} \neq y_i$ is its corresponding label, and α is a regularization term that weights the retain loss to prevent degradation on the retain set.

3.5 Discussion

Imbalance between retain and forget sets. In CIL, access to retain data is limited to a replay buffer and the current task, whereas the forget set may include samples from any previously observed data. For example, a user may request the deletion of all their data after it has been used across multiple tasks, requiring the removal of samples from the entire training history. This leads to a significant imbalance between the retain and forget sets. To mitigate this, we upsample the retain set by iterating over it multiple times until it matches the effective size of the forget set.

Computational cost. IMPRINT introduces negligible computational overhead, as it only requires computing the logit margin and applying a scalar weight to the loss. We further analyze the per-epoch runtime of IMPRINT in Section 4.

Adaptive logit margin. Due to its low computational cost, IMPRINT can recompute sample-wise weights at each training epoch. This enables the unlearning process to dynamically reflect the evolving imprint of samples, leading to more accurate and stable unlearning.

Relation to non-uniform unlearning. IMU and NPO also employ non-uniform forgetting, but use different weighting signals. IMU relies on classifier-level influence and NPO on prediction changes relative to a reference model, whereas IMPRINT uses the current logit margin to differentiate forgetting intensity. A detailed comparison of their behaviors is provided in Appendix B.2.

4 Experiments

4.1 Experimental Setup

Datasets, models, and CIL setup. We conduct experiments on CIFAR-100 (Krizhevsky et al., 2009), Tiny-ImageNet (Le et al., 2015), and ImageNet-1k (Deng et al., 2009) under standard CIL protocols, where each dataset is divided into a sequence of disjoint tasks. Specifically, we use 10 tasks with 10 classes each for CIFAR-100, 10 tasks with 20 classes each for Tiny-ImageNet, and 10 tasks with 100 classes each for ImageNet-1k. We evaluate our method using ResNet-18 (He et al., 2016) and ViT-Base/16 (Dosovitskiy et al., 2020) backbones, combined with representative rehearsal-based CIL methods, including DER++ (Buzzega et al., 2020), FOSTER (Wang et al., 2022a), and iCaRL (Rebuffi et al., 2017). DER++ is applied to ResNet-18 CIFAR-100, and Tiny-ImageNet, while FOSTER is used with ResNet-18 on ImageNet-1k and with ViT on CIFAR-100. When implementing CIL methods, we adopt the hyperparameters reported in their original works, and follow the LAMDA-PILOT (Sun et al., 2025) setting when training ViT with FOSTER. Results on iCaRL are provided in the Appendix.

Unlearning scenarios. We evaluate our method under two unlearning scenarios: class-wise forgetting and random data forgetting, with a primary focus on class-wise forgetting. In class-wise forgetting, all samples from selected classes are removed, whereas in random data forgetting, samples are randomly selected across all classes. For class-wise forgetting, 30% of classes are selected to span a wide range of logit margins, while for random data forgetting, 30% of samples are randomly selected. The effect of varying the unlearning ratio is analyzed in the Appendix.

Baselines and evaluation metrics. We compare our approach against ten unlearning methods: RL (Golatkar et al., 2020), IU (Koh & Liang, 2017), GA (Thudi et al., 2022), BE, BS (Chen et al., 2023), IMU (Fan et al., 2025), NPO (Zhang et al., 2024), SCRUB (Kurmanji et al., 2023), AMUN (Ebrahimpour-Boroojeny et al., 2025), and SalUn (Fan et al., 2024). Importantly, our baselines include IMU, a non-uniform unlearning method that incorporates sample-level influence, as well as

Table 3: **Comparison with baselines** on CIFAR-100 and Tiny-ImageNet. IMPRINT achieves the smallest average gap. Results are averaged over five different forgetting configurations.

(a) CIFAR-100

	Buffer size = 500						Buffer size = 2000					
Method	**FA**	**RA**	**MIA**	**FT**	**RT**	**Avg. Gap**	**FA**	**RA**	**MIA**	**FT**	**RT**	**Avg. Gap**
Retrain	0.00	43.99	100.00	0.00	40.90	—	0.00	64.11	100.00	0.00	57.18	—
RL	0.06	28.03	99.72	0.31	37.83	$3.94_{\pm 0.88}$	0.30	47.90	98.97	1.66	51.81	$4.92_{\pm 0.89}$
SalUn	0.42	33.94	99.88	1.19	39.66	$2.87_{\pm 1.16}$	1.18	52.46	99.76	3.77	52.90	$4.22_{\pm 0.31}$
SCRUB	0.00	3.86	100.00	0.00	4.65	$15.27_{\pm 0.70}$	0.00	5.04	99.99	0.01	5.34	$22.19_{\pm 0.55}$
AMUN	9.30	27.14	84.98	9.53	25.94	$13.13_{\pm 1.53}$	25.14	49.80	73.40	23.41	46.11	$20.11_{\pm 3.96}$
NPO	0.00	26.08	99.99	0.01	29.37	$5.89_{\pm 1.20}$	0.00	3.14	100.00	0.03	17.12	$20.22_{\pm 1.72}$
GA	1.44	0.80	60.91	1.25	0.90	$25.00_{\pm 11.05}$	0.67	1.09	80.67	0.67	1.12	$27.95_{\pm 9.86}$
BE	0.10	5.00	99.06	0.03	4.55	$15.28_{\pm 0.58}$	0.05	18.64	98.37	0.05	17.57	$17.37_{\pm 0.72}$
BS	6.03	21.01	88.96	5.83	20.84	$13.19_{\pm 1.56}$	7.20	36.88	87.00	6.79	35.35	$15.21_{\pm 1.05}$
IU	16.15	38.88	82.20	13.96	37.69	$11.24_{\pm 3.55}$	34.56	61.45	69.68	28.88	56.00	$19.52_{\pm 2.75}$
IMU	5.24	27.43	97.35	5.19	27.16	$8.68_{\pm 2.00}$	9.03	56.17	96.52	8.62	51.00	$7.05_{\pm 0.52}$
IMPRINT	0.25	43.26	99.38	0.15	41.42	$\mathbf{1.36_{\pm 1.29}}$	0.30	61.27	99.22	0.18	56.29	$\mathbf{1.00_{\pm 0.82}}$

(b) Tiny-ImageNet

	Buffer size = 500						Buffer size = 5120					
Method	**FA**	**RA**	**MIA**	**FT**	**RT**	**Avg. Gap**	**FA**	**RA**	**MIA**	**FT**	**RT**	**Avg. Gap**
Retrain	0.00	28.26	100.00	0.00	23.78	—	0.00	57.49	100.00	0.00	44.73	—
RL	0.02	15.47	99.84	0.00	15.41	$4.37_{\pm 2.45}$	0.00	31.54	100.00	0.00	31.38	$7.86_{\pm 1.55}$
SalUn	0.16	18.68	99.74	0.01	17.97	$3.51_{\pm 1.85}$	0.02	32.41	99.99	0.01	33.15	$7.34_{\pm 1.96}$
SCRUB	0.00	0.73	99.96	0.01	0.73	$10.13_{\pm 2.05}$	0.00	0.74	98.48	0.00	0.74	$20.45_{\pm 1.27}$
AMUN	6.84	13.02	58.28	5.41	11.31	$16.34_{\pm 10.98}$	17.81	36.19	75.12	13.58	29.92	$18.48_{\pm 3.75}$
NPO	0.01	10.52	99.95	0.00	10.59	$6.20_{\pm 2.13}$	0.01	33.36	99.94	0.01	32.79	$7.23_{\pm 1.24}$
GA	0.64	0.42	0.64	0.64	0.43	$30.37_{\pm 1.97}$	0.00	0.72	0.00	0.00	0.72	$40.16_{\pm 0.74}$
BE	0.12	1.37	0.55	0.00	0.75	$29.90_{\pm 2.04}$	0.11	13.28	96.78	0.14	10.33	$16.42_{\pm 0.91}$
BS	4.91	10.33	58.97	4.65	9.34	$16.60_{\pm 10.39}$	8.67	30.52	86.34	6.99	24.99	$15.21_{\pm 1.87}$
IU	12.96	21.36	58.31	10.12	17.73	$16.11_{\pm 10.09}$	31.56	53.69	66.55	21.55	43.36	$18.93_{\pm 4.30}$
IMU	0.01	10.49	99.98	0.01	8.75	$6.57_{\pm 2.11}$	0.27	23.58	99.36	0.19	19.70	$12.01_{\pm 1.16}$
IMPRINT	0.26	21.21	99.30	0.03	18.08	$\mathbf{3.14_{\pm 1.07}}$	0.09	41.32	99.05	0.01	35.05	$\mathbf{5.38_{\pm 2.20}}$

Table 4: **Plug-and-play compatibility** of IMPRINT on CIFAR-100. For each base unlearning method, we compare its standalone performance with the IMPRINT-augmented variant (+IMPRINT). Adding IMPRINT consistently reduces the gap to retraining. Results are averaged over five different forgetting configurations.

Buffer size = 500							Buffer size = 2000						
Method	**FA**	**RA**	**MIA**	**FT**	**RT**	**Avg. Gap**	**Method**	**FA**	**RA**	**MIA**	**FT**	**RT**	**Avg. Gap**
Retrain	0.00	43.99	100.00	0.00	40.90	—	Retrain	0.00	64.11	100.00	0.00	57.18	—
RL	0.06	28.03	99.72	0.31	37.83	$3.94_{\pm 0.88}$	RL	0.30	47.90	98.97	1.66	51.81	$4.92_{\pm 0.89}$
RL+IMPRINT	0.25	43.26	99.38	0.15	41.42	$\mathbf{1.36_{\pm 1.29}}$	RL+IMPRINT	0.30	61.27	99.22	0.18	56.29	$\mathbf{1.00_{\pm 0.82}}$
SalUn	0.42	33.94	99.88	1.19	39.66	$2.87_{\pm 1.16}$	SalUn	1.18	52.46	99.76	3.77	52.90	$4.22_{\pm 0.31}$
SalUn+IMPRINT	0.36	42.61	99.11	0.21	40.72	$\mathbf{1.52_{\pm 1.50}}$	SalUn+IMPRINT	0.35	61.16	99.23	0.20	56.05	$\mathbf{1.08_{\pm 1.10}}$
AMUN	9.30	27.14	84.98	9.53	25.94	$13.13_{\pm 1.53}$	AMUN	25.14	49.80	73.40	23.41	46.11	$20.11_{\pm 3.96}$
AMUN+IMPRINT	2.13	44.55	96.15	1.79	42.43	$\mathbf{2.74_{\pm 2.02}}$	AMUN+IMPRINT	5.83	62.58	91.85	4.84	56.89	$\mathbf{4.14_{\pm 4.05}}$
BS	6.03	21.01	88.96	5.83	20.84	$13.19_{\pm 1.56}$	BS	7.20	36.88	87.00	6.79	35.35	$15.21_{\pm 1.05}$
BS+IMPRINT	0.54	38.43	98.79	0.39	36.31	$\mathbf{2.46_{\pm 1.56}}$	BS+IMPRINT	0.46	60.11	98.96	0.18	54.80	$\mathbf{1.61_{\pm 1.41}}$

NPO, which adjusts updates using prediction changes relative to a reference model. For class-wise forgetting, we evaluate performance using forget accuracy (FA), retain accuracy (RA), membership inference attack (MIA) (Shokri et al., 2017) ratio, forget test accuracy (FT), and retain test accuracy (RT). For random data forgetting, we report FA, RA, MIA ratio, and overall test accuracy (TA). We report the average gap with respect to a retrained model, which is obtained by retraining under the same CIL setting using only the full retain set (i.e., without the forget set). For a fair comparison, we have added the replay-retain loss terms even in retain-data-free unlearning baselines (NPO, GA, BE, BS, IU, and IMU). We also evaluate the compatibility of IMPRINT with RL, SalUn, AMUN, and BS, while reporting the main results using RL.

4.2 Main Experimental Results

Comparison with baselines. Table 3 compares IMPRINT with baselines across two datasets with two buffer sizes each: CIFAR-100 with DER++ (buffer sizes 500 and 2000) and Tiny-ImageNet with DER++ (buffer sizes 500 and 5120). The results reveal a trade-off between forgetting and utility

Table 5: FOSTER results on **ImageNet-1k**.

Method	FA	RA	MIA	FT	RT	Avg. Gap
Retrain	0.00	39.43	100.00	0.00	43.18	—
RL	0.00	18.20	99.84	0.00	21.05	$9.09_{\pm 2.79}$
SaLUn	0.00	17.80	100.00	0.00	21.28	$9.11_{\pm 2.05}$
SCRUB	0.00	0.14	100.00	0.00	0.14	$16.47_{\pm 0.03}$
AMUN	14.58	37.89	74.73	16.72	42.43	$13.25_{\pm 2.02}$
NPO	0.00	7.75	100.00	0.00	9.03	$13.17_{\pm 2.81}$
GA	0.00	0.15	0.00	0.00	0.15	$36.46_{\pm 0.01}$
BE	0.00	0.08	99.99	0.00	0.11	$16.49_{\pm 0.01}$
BS	6.64	23.44	86.63	8.04	29.02	$11.64_{\pm 0.33}$
IU	5.23	50.06	91.74	6.03	54.33	$8.26_{\pm 0.04}$
IMU	0.00	18.59	100.00	0.00	20.07	$8.79_{\pm 0.08}$
IMPRINT	0.00	34.18	99.99	0.00	37.94	$\mathbf{2.50_{\pm 0.42}}$

Table 6: FOSTER results on **ViT**.

Method	FA	RA	MIA	FT	RT	Avg. Gap
Retrain	0.00	91.63	100.00	0.00	89.00	—
RL	0.00	60.66	100.00	0.00	57.97	$12.40_{\pm 4.02}$
SaLUn	0.00	66.85	100.00	0.00	64.61	$9.83_{\pm 4.02}$
SCRUB	0.00	1.71	99.99	0.00	1.72	$35.45_{\pm 0.14}$
AMUN	24.25	80.06	82.27	32.61	77.45	$19.55_{\pm 4.28}$
NPO	0.00	31.01	100.00	0.00	28.62	$24.20_{\pm 1.55}$
GA	0.1	1.348	59.632	0.08	1.386	$43.69_{\pm 10.89}$
BE	0.00	68.78	100.00	0.00	66.02	$9.17_{\pm 0.72}$
BS	9.34	76.42	94.65	11.61	74.46	$11.21_{\pm 0.67}$
IU	67.22	93.19	47.04	65.01	90.61	$37.67_{\pm 4.44}$
IMU	0.67	84.34	99.33	0.67	82.22	$3.21_{\pm 1.21}$
IMPRINT	0.00	86.03	100.00	0.02	83.79	$\mathbf{2.17_{\pm 1.49}}$

across baselines. BE and SCRUB favor forgetting at the expense of utility, whereas AMUN, BS, and IU better preserve utility at the cost of less complete forgetting. In contrast, IMPRINT achieves near-zero FA and FT while keeping RA and RT close to those of the retrained model. Consequently, IMPRINT attains the lowest average gap to retraining across both datasets and both buffer sizes.

Compatibility with existing unlearning methods. Table 4 shows the effectiveness of IMPRINT as a plug-and-play method. Applying IMPRINT to RL increases RA and RT and decreases FT at both buffer sizes; MIA changes only slightly (a small decrease at buffer 500 and a small increase at buffer 2000). More notably, applying IMPRINT to SalUn, AMUN, and BS simultaneously decreases FA and FT while increasing RA and RT, substantially narrowing the gap to the retrained model.

ImageNet-1k and ViT. To further evaluate the generalizability of IMPRINT, we consider both a larger-scale dataset and a different backbone architecture. Table 5 reports results on ImageNet-1k with a ResNet-18 backbone, and Table 6 reports results on CIFAR-100 with a ViT-Base/16 backbone, both trained with FOSTER. Across both settings, most of the existing baselines suffer a substantial gap from full retraining, indicating that they struggle to preserve performance when either the data distribution becomes more diverse or the backbone capacity increases. In contrast, IMPRINT consistently attains the smallest gap to retraining. These results indicate that imprint-aware unlearning is not tied to a particular dataset scale or architecture family, and that aligning update intensity with sample-wise imprint remains effective as both the data and the model grow.

4.3 Further Studies

Impact of influence score. Figure 5 compares the performance of different data influence proxies on CIFAR-100. We evaluate uniform unlearning (RL, SalUn, and AMUN) and compare it with accuracy-, entropy-, and logit margin-based weighting. In RL, accuracy and entropy reduce the RA/RT gap to retraining but at the cost of insufficient forgetting, as evidenced by their larger FA/FT gaps. In contrast, the logit-margin proxy reduces the gap on both retain and forget metrics simultaneously, with a minor exception for RL at buffer size 500 (slightly increased FA). This indicates that logit margin is a better proxy than class accuracy or entropy. A more detailed comparison across proxies, together with the procedure for converting class accuracy and entropy into forget-loss weights, is provided in the Appendix.

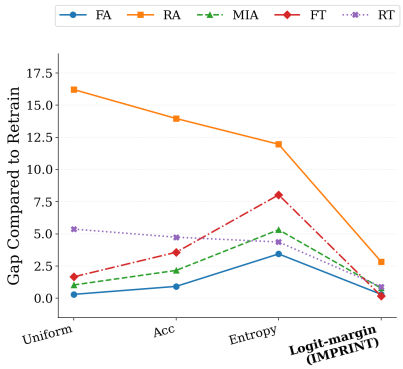

Figure 5: Effect of the data-influence proxy on unlearning quality. The logit-margin proxy yields the smallest gap.

Random forgetting. Table 7 presents the performance of baselines and IMPRINT in the random forgetting setup. Most existing methods over-forget, yielding low FA and RA values that deviate sharply from the retrained reference. Unlike other methods, IU achieves a balanced trade-off among FA, RA, MIA, and TA, resulting in a relatively small gap to the retrained baseline. This contrasts with its poor class-wise performance in Table 3. Random forgetting samples across all classes, making the forget and retain sets similar in their overall imprint distributions and allowing IU's second-order correction to operate closer to its intended regime. Class-wise forgetting instead concentrates samples in a few classes whose imprint distributions can differ substantially from those of the retain set. Still, similar set-level distributions do not imply uniform sample-wise imprints. IMPRINT exploits the remaining heterogeneity by adapting update intensity to each sample, achieving the lowest average gap across all four metrics.

Table 7: **Random forgetting** on CIFAR-100. IMPRINT consistently achieves smallest average gap.

	Buffer size = 500					Buffer size = 2000				
Method	**FA**	**RA**	**MIA**	**TA**	**Avg. Gap**	**FA**	**RA**	**MIA**	**TA**	**Avg. Gap**
Retrain	31.87	36.50	60.53	32.95	—	48.55	59.51	58.52	50.69	—
RL	14.10	14.59	96.31	27.62	20.20	33.14	33.93	82.29	42.48	18.24
SalUn	19.25	19.23	93.61	29.41	16.63	37.59	38.08	76.73	44.29	14.25
SCRUB	1.61	1.49	0.47	1.90	39.10	4.54	4.61	99.30	4.62	46.44
AMUN	12.08	14.89	81.07	13.08	20.45	41.51	46.80	52.58	42.60	8.45
GA	0.93	0.87	99.09	0.96	34.28	1.10	0.98	98.93	1.00	49.02
BE	0.70	1.35	97.01	0.71	33.76	1.78	3.21	91.03	4.59	45.42
BS	10.73	11.97	83.21	11.92	22.35	16.79	19.33	76.33	21.92	29.63
IU	41.35	40.63	50.93	37.73	7.00	58.94	59.19	41.31	53.06	7.57
IMU	22.20	22.45	92.44	22.67	16.48	49.80	50.02	84.78	45.30	10.60
NPO	0.55	0.79	98.85	4.01	33.57	1.13	1.34	97.37	3.27	47.96
IMPRINT	30.38	30.48	73.29	33.95	**5.32**	46.87	46.99	59.23	48.06	**4.39**

Table 8: **Effect of forget-set imprint heterogeneity.** We compare a homogeneous low-imprint forget set with a heterogeneous forget set containing mixed imprint strengths. Results are averaged over five different seeds.

(a) Homogeneous

Method	**FA↓**	**RA↑**	**MIA↑**	**FT↓**	**RT↑**	**Avg. Gap↓**
Retrain	0.00	72.68	100.00	0.00	66.81	—
RL	0.21	49.37	99.95	0.50	61.27	$5.92_{\pm.58}$
RL+IMPRINT	3.43	66.84	94.53	1.61	62.82	$\mathbf{4.07}_{\pm.10}$
SalUn	0.65	54.72	99.82	1.35	61.67	$5.06_{\pm.43}$
SalUn+IMPRINT	3.42	65.75	94.24	1.67	62.10	$\mathbf{4.50}_{\pm.11}$

(b) Heterogeneous

Method	**FA↓**	**RA↑**	**MIA↑**	**FT↓**	**RT↑**	**Avg. Gap↓**
Retrain	0.00	65.15	100.00	0.00	57.51	—
RL	0.13	49.37	99.98	1.07	54.14	$4.08_{\pm.42}$
RL+IMPRINT	0.33	60.76	99.20	0.15	56.30	$\mathbf{1.38}_{\pm.27}$
SalUn	0.91	53.24	99.83	3.43	54.94	$3.80_{\pm.20}$
SalUn+IMPRINT	0.37	60.48	99.00	0.34	55.99	$\mathbf{1.58}_{\pm.34}$

Compatibility with other CIL methods. Table 5 and Table 6 report results obtained with FOSTER, where IMPRINT continues to outperform baselines. This indicates that imprint-aware unlearning generalizes beyond DER++. Additional experiments on iCaRL are reported in the Appendix, where IMPRINT consistently achieves improved performance compared to existing methods.

Runtime. Figure 6 shows the runtime of IMPRINT by measuring the average epoch duration. Across all three base methods, the average delay of IMPRINT is nearly negligible that imprint-aware weighting adds essentially no runtime overhead.

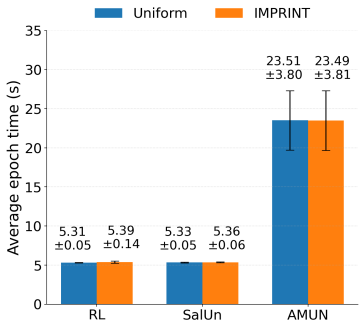

Figure 6: **Average per-epoch runtime** with and without IMPRINT.

Effect of forget-set imprint heterogeneity. To better understand how sample-imprint heterogeneity affects unlearning, we construct two 30-class forget sets based on pre-unlearning class-average logit-margin ranks. The homogeneous set contains low-margin classes (ranks 71–100), whereas the heterogeneous set mixes 10 low- (91–100), 10 middle- (46–55), and 10 high-margin classes (1–10). As shown in Table 8, IMPRINT reduces the gap to retraining in both settings, with a larger improvement on the heterogeneous set. This supports our motivation that adaptive sample-wise weighting is particularly important when forget samples exhibit diverse imprint strengths.

Other results. In the Appendix, we evaluate sensitivity to the CIL training budget, logit-scale variation, temperature, and forgetting ratio. To examine generalization across additional continual learning settings, we evaluate IMPRINT on Task-IL, Domain-IL, and text classification using the 20 newsgroups dataset. We also compare the logit margin with alternative influence surrogates and assess the robustness of IMPRINT through sequential unlearning and post-unlearning learning experiments.

5 Conclusion

We proposed IMPRINT, an imprint-aware unlearning framework for class-incremental learning. We show that continual learning induces heterogeneous imprints, making uniform unlearning suboptimal. IMPRINT uses the logit margin as an efficient proxy for sample-wise imprint and adaptively scales the unlearning loss accordingly. Experiments across diverse models and datasets consistently demonstrate the advantages of IMPRINT over existing baselines under different CIL settings. Overall, our results suggest that effective unlearning in continual learning depends strongly on sample imprint strength, highlighting imprint-aware adaptation as a key principle for reliable unlearning.

AI USE STATEMENT

We used generative AI tools during the development and preparation of this work. The tool assisted with the formulation and presentation of mathematical claims and proofs, editing portions of the manuscript to improve clarity and readability, identifying potentially relevant literature, and generating software code used in the research workflow.

All AI-assisted mathematical claims, proofs, and derivations were independently reviewed and verified by the authors. The experimental procedures, reported results, and numerical analyses were conducted, validated, and interpreted by the authors. AI-assisted text and code were also reviewed and revised as necessary.

The authors made the final decisions regarding the methodology, mathematical claims, experimental analysis, interpretation, and presentation of the work, and take full responsibility for all content of the paper.

ETHICS STATEMENT

This work studies machine unlearning with the goal of improving the removal of designated training data from continually learned models. Our experiments use standard publicly available benchmarks and do not involve human subjects or personally identifiable information.

REPRODUCIBILITY STATEMENT

Section 4.1 describes the datasets, model architectures, class-incremental learning protocols, unlearning scenarios, baselines, and evaluation metrics. Appendix A.1 provides implementation details, including the IMPRINT temperature, the selection of the retain-loss coefficient, and the method-specific optimization settings. We use a common batch size and number of unlearning epochs across methods within each dataset, while optimizers and learning rates follow the settings described in Appendix A.1.

We provide source code and scripts for reproducing the main experiments and proxy ablations as supplementary material. Dataset download and preprocessing instructions are included in the accompanying README.

REFERENCES

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

Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and Simone Calderara. Dark experience for general continual learning: a strong, simple baseline. *Advances in neural information processing systems*, 33:15920–15930, 2020.

Xuemei Cao, Hanlin Gu, Xin Yang, Bingjun Wei, Haoyang Liang, Xiangkun Wang, and Tianrui Li. Erroreraser: Unlearning data bias for improved continual learning. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 119–130, 2025.

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

Sungmin Cha, Sungjun Cho, Dasol Hwang, Honglak Lee, Taesup Moon, and Moontae Lee. Learning to unlearn: Instance-wise unlearning for pre-trained classifiers. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 11186–11194, 2024.

Min Chen, Weizhuo Gao, Gaoyang Liu, Kai Peng, and Chen Wang. Boundary unlearning: Rapid forgetting of deep networks via shifting the decision boundary. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 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 *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 7210–7217, 2023.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

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

Ali Ebrahimpour-Boroojeny, Hari Sundaram, and Varun Chandrasekaran. Not all wrong is bad: Using adversarial examples for unlearning. In *Forty-second International Conference on Machine Learning*, 2025.

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

Xindi Fan, Jing Wu, Mingyi Zhou, Pengwei Liang, and Dinh Phung. Imu: Influence-guided machine unlearning. *arXiv preprint arXiv:2508.01620*, 2025.

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

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

Aditya Golatkar, Alessandro Achille, and Stefano Soatto. Eternal sunshine of the spotless net: Selective forgetting in deep networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9304–9312, 2020.

Ian J Goodfellow, Mehdi Mirza, Da Xiao, Aaron Courville, and Yoshua Bengio. An empirical investigation of catastrophic forgetting in gradient-based neural networks. *arXiv preprint arXiv:1312.6211*, 2013.

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

Yiting Hu, Lingjie Duan, and Qian Zhang. The forgetting-retention dilemma: Certified unlearning theory in continual learning. *arXiv preprint arXiv:2606.29832*, 2026.

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

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

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

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

Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017.

Bo Liu, Qiang Liu, and Peter Stone. Continual learning and private unlearning. In *Conference on Lifelong Learning Agents*, pp. 243–254. PMLR, 2022.

Alessandro Mantelero. The eu proposal for a general data protection regulation and the roots of the 'right to be forgotten'. *Computer Law & Security Review*, 29(3):229–235, 2013.

Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pp. 109–165. Elsevier, 1989.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 2001–2010, 2017.

David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy Lillicrap, and Gregory Wayne. Experience replay for continual learning. *Advances in neural information processing systems*, 32, 2019.

Takashi Shibata, Go Irie, Daiki Ikami, and Yu Mitsuzumi. Learning with selective forgetting. In *IJCAI*, volume 3, pp. 4, 2021.

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

Hai-Long Sun, Da-Wei Zhou, De-Chuan Zhan, and Han-Jia Ye. Pilot: A pre-trained model-based continual learning toolbox, 2025.

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

Fu-Yun Wang, Da-Wei Zhou, Han-Jia Ye, and De-Chuan Zhan. Foster: Feature boosting and compression for class-incremental learning. In *European conference on computer vision*, pp. 398–414. Springer, 2022a.

Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. Learning to prompt for continual learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 139–149, 2022b.

Yue Wu, Yinpeng Chen, Lijuan Wang, Yuancheng Ye, Zicheng Liu, Yandong Guo, and Yun Fu. Large scale incremental learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 374–382, 2019.

Liheng Yu, Zhe Zhao, Yuxuan Wang, Pengkun Wang, Xiaofeng Cao, Binwu Wang, and Yang Wang. Falw: A forgetting-aware loss reweighting for long-tailed unlearning. *arXiv preprint arXiv:2601.18650*, 2026.

Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. Negative preference optimization: From catastrophic collapse to effective unlearning. *arXiv preprint arXiv:2404.05868*, 2024.

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

A Appendix Overview and Implementation Details

This appendix is organized into five parts. First, we provide the full implementation and hyperparameter details (§A.1). Second, we present analyses of the proposed unlearning weights, including the mechanistic interpretation of logit margin as a proxy for removal-sensitive imprint (§B.1), the comparison of non-uniform unlearning strategies (§B.2), comparison of logit margin- and influence-based weighting (§B.3), the conversion of class accuracy and entropy into forget-loss weights (§B.4), extended comparisons of sample-influence proxies across datasets (§B.5), the distribution of sample-wise imprints (§B.6), comparisons with training-history-based surrogates and robustness to task-wise logit-scale variation (§B.7), and a mean-matched uniform forget-loss weighting control (§B.8). Third, we report sensitivity and robustness analyses, including sensitivity to the temperature hyperparameter τ (§C.1), the effect of varying the forgetting ratio (§C.2), and robustness to the CIL training budget (§C.3). Fourth, we evaluate the compatibility and generality of `IMPRINT`, including compatibility with other unlearning methods on ImageNet-1k (§D.1) and with iCaRL (§D.2), together with evaluations beyond the primary CIL setting(§D.3, §D.4). Finally, we analyze continual unlearning dynamics, including subsequent learning, forgetting relapse(§E.1), and the task age and order of unlearning requests(§E.2).

A.1 Implementation Details

Hyperparameters. For all IMPRINT variants we set the temperature τ=0.5 in the sigmoid weighting (Eq. 2). For each CIL method, we set the retain-loss coefficient α to the optimal hyperparameter value used to weight the CE loss, as reported in its original paper and official repository, and keep α fixed during unlearning.

Continual learning backbones. We pair ResNet-18 with DER++ on CIFAR-100, and Tiny-ImageNet, and with FOSTER on ImageNet-1k and with iCaRL on CIFAR-100. ViT-Base/16 is paired with FOSTER on CIFAR-100. For DER++, FOSTER, and iCaRL we adopt the hyperparameters reported in their original works. When training ViT with FOSTER, we follow the LAMDA-PILOT configuration. Replay buffers are populated with herding selection. The CIL setup (number of tasks, classes per task, buffer sizes) follows the specification in Section 4.1 of the main paper.

Unlearning protocol. All unlearning baselines use the hyperparameters and the optimizers reported in their original papers and learning rates are set to 0.01 in RL, SalUn, AMUN, IMPRINT, and to 0.001 with other unlearning methods. For GA with the ViT model, we use a smaller learning rate of 10^{-5}, as 10^{-3} and 10^{-4} resulted in numerical divergence across all forgetting configurations, producing non-finite model parameters and invalid logits. To enable a fair comparison, within each dataset we use batch size of 256, and 10 number of unlearning epochs across all methods (including IMPRINT). The retrained reference model is obtained by running the full CIL procedure from scratch on the full retain set D_r^{full} only, using identical hyperparameters to those used for the original CIL training. RA was measured with full retain set. For each forgetting configuration, we compute Avg. Gap as the mean absolute difference between the evaluated metrics and those of the matched retrained reference. We report the mean and standard deviation of these per-configuration gaps across five forgetting configurations, unless otherwise specified. We apply the SalUn mask with a mask sparsity of 0.5.

Forgetting configurations. Unless stated otherwise, class-wise forgetting selects 30% of the classes (chosen to span a wide range of per-class logit margins, so that the forget set contains both strongly and weakly imprinted classes), and random forgetting selects 30% of the samples uniformly at random across all classes. The effect of varying this ratio to 10% and 20% is reported in §C.2. All reported numbers are averaged over five different forget-set configurations (different sampled class subsets / data subsets).

Compute resources. ViT was conducted with DGX B200 GPU and other models were conducted on Nvidia RTX 3090, DGX B200 GPUs. A single unlearning run on CIFAR-100 / Tiny-ImageNet takes approximately 6 minutes; the FOSTER / ViT take approximately 2 minutes per run and the FOSTER / ImageNet-1k experiments take approximately 6–12 hours per run. The total computational budget for the experiments reported in this paper is approximately 700 GPU-hours with Nvidia RTX 3090 GPU and 40 hours with DGX B200 GPU.

B Mechanistic Analysis and Unlearning Weights

B.1 Mechanistic Explanation: Logit Margin as a Measure of Decision-Level Imprint

Throughout the main paper, *sample-wise imprint* refers broadly to the degree to which information associated with a training sample remains reflected in the trained model after continual learning. Such imprint can manifest at different levels, including model parameters, representations, and predictions. In this section, we focus on the aspect that is directly observable from the current predictor, which we call *decision-level imprint*.

Decision-level imprint characterizes how strongly the current model supports a sample's ground-truth class relative to its strongest competing class. This is the component of sample-wise imprint that is immediately relevant when unlearning begins: samples that remain strongly supported by the current decision function require a larger intervention before that support is removed, whereas samples whose true-class preference is already weak or reversed require substantially less intervention.

We first establish a geometric interpretation of the logit margin in output space. We then connect this quantity to the minimum parameter-space intervention required to erase the current true-class

preference, and derive a first-order expression for this removal effort. Finally, we empirically test whether the raw logit margin used by IMPRINT preserves the ordering induced by this theoretically sharper parameter-space quantity. Together, these results provide a mechanistic justification for using the logit margin to allocate sample-dependent forgetting intensity.

Decision-level imprint in logit space. For a sample $i = (x_i, y_i)$, let

$$z_i = z(x_i; \theta) \in \mathbb{R}^C \tag{5}$$

denote its current logit vector, and let

$$r_i \in \arg\max_{c \neq y_i} z_{i,c} \tag{6}$$

denote its strongest competing class. The logit margin is

$$m_i = z_{i,y_i} - z_{i,r_i} = z_{i,y_i} - \max_{c \neq y_i} z_{i,c}. \tag{7}$$

Consider the pairwise decision boundary between the ground-truth class y_i and its strongest competitor r_i:

$$\mathcal{H}_i = \left\{ u \in \mathbb{R}^C : u_{y_i} = u_{r_i} \right\}. \tag{8}$$

Its unit normal vector, oriented toward the ground-truth class, is

$$n_i = \frac{e_{y_i} - e_{r_i}}{\sqrt{2}}, \tag{9}$$

where e_c denotes the c-th standard basis vector.

Definition (Decision-level imprint). We define the *decision-level imprint* of sample i as the signed Euclidean distance of its current logit vector from the ground-truth-versus-strongest-competitor decision boundary:

$$\mathcal{I}_i = \operatorname{sdist}(z_i, \mathcal{H}_i) = \langle z_i, n_i \rangle. \tag{10}$$

A positive value indicates retained preference for the ground-truth class, zero corresponds to the decision boundary, and a negative value indicates that a competing class is preferred.

Proposition 1 (Geometric interpretation of the logit margin)**.** *For every sample i,*

$$\boxed{\mathcal{I}_i = \frac{m_i}{\sqrt{2}}}. \tag{11}$$

Hence, the raw logit margin and decision-level imprint induce exactly the same ordering over samples.

Proof. Because $\mathcal{H}_i = \{u : u_{y_i} - u_{r_i} = 0\}$, its normalized normal vector is $n_i = (e_{y_i} - e_{r_i})/\sqrt{2}$. Therefore,

$$\mathcal{I}_i = \left\langle z_i, \frac{e_{y_i} - e_{r_i}}{\sqrt{2}} \right\rangle \tag{12}$$

$$= \frac{z_{i,y_i} - z_{i,r_i}}{\sqrt{2}} \tag{13}$$

$$= \frac{m_i}{\sqrt{2}}. \tag{14}$$

□

Proposition 1 gives the logit margin a direct decision-space interpretation. A large positive margin places the sample farther on the ground-truth side of the relevant boundary, whereas a small or negative margin indicates weak or reversed support for the true class.

From decision-level imprint to removal effort. The geometric result above describes how strongly a sample is currently supported in output space. We next consider a quantity more directly related to unlearning: how much must the model parameters change before this current true-class preference can be removed?

Define the margin function

$$h_i(\theta) = z_{y_i}(x_i;\theta) - \max_{c \neq y_i} z_c(x_i;\theta), \tag{15}$$

so that $h_i(\theta) = m_i$. For a sample with $m_i > 0$, define its *parameter-space decision-removal radius* as

$$\mathcal{R}_i(\theta) = \inf_{\Delta\theta} \{\|\Delta\theta\|_2 : h_i(\theta + \Delta\theta) \leq 0\}. \tag{16}$$

This quantity measures the smallest parameter perturbation required to eliminate the model's current preference for the ground-truth class.

Theorem 1 (Margin-dependent lower bound on decision-removal effort)**.** *Suppose that h_i is L_i-Lipschitz with respect to the model parameters over the parameter region connecting the current model to a perturbation that removes the current true-class preference:*

$$|h_i(\theta + \Delta\theta) - h_i(\theta)| \leq L_i \|\Delta\theta\|_2. \tag{17}$$

Then, for any sample with $m_i > 0$,

$$\boxed{\mathcal{R}_i(\theta) \geq \frac{m_i}{L_i}}. \tag{18}$$

Proof. Let $\Delta\theta$ be any parameter perturbation that eliminates the current true-class preference. Then

$$h_i(\theta + \Delta\theta) \leq 0. \tag{19}$$

Since $h_i(\theta) = m_i > 0$,

$$\begin{aligned} m_i &= h_i(\theta) && (20)\\ &\leq h_i(\theta) - h_i(\theta + \Delta\theta) && (21)\\ &\leq |h_i(\theta + \Delta\theta) - h_i(\theta)| && (22)\\ &\leq L_i \|\Delta\theta\|_2. && (23) \end{aligned}$$

Therefore,

$$\|\Delta\theta\|_2 \geq \frac{m_i}{L_i}. \tag{24}$$

Because the inequality holds for every perturbation that removes the true-class preference, taking the infimum over all such perturbations gives Eq. equation 18. □

Theorem 1 connects the current decision state to the amount of model intervention required to erase it. For a fixed local sensitivity L_i, a sample with a larger positive margin cannot have its current true-class preference removed without a proportionally larger parameter perturbation. Thus, heterogeneous margins correspond not only to heterogeneous decision states, but also to heterogeneous lower bounds on the parameter-space effort required to alter those states.

The sensitivity L_i can vary across samples, so raw margin alone does not fully characterize parameter-space removal difficulty. Estimating this sensitivity directly, however, requires gradient- or Jacobian-level information. The logit margin instead provides a sample-specific quantity available directly from the current forward pass.

First-order interpretation. The relationship becomes more explicit under a local linear approximation. Assume that the strongest competing class is unique and remains unchanged in a sufficiently small neighborhood of the current parameters, so that h_i is differentiable locally. Let

$$g_i = \nabla_\theta h_i(\theta). \tag{25}$$

For a sufficiently small parameter perturbation,

$$h_i(\theta + \Delta\theta) \approx m_i + g_i^\top \Delta\theta. \tag{26}$$

The minimum-norm perturbation that reaches the locally linearized decision boundary therefore solves

$$\min_{\Delta\theta} \|\Delta\theta\|_2 \quad \text{subject to} \quad m_i + g_i^\top \Delta\theta \leq 0. \tag{27}$$

Corollary 1 (First-order decision-removal effort)**.** *Under the local linearization in equation 26, the minimum parameter perturbation required to erase a positive margin is*

$$\boxed{\mathcal{R}_i^{\text{lin}} = \frac{m_i}{\|g_i\|_2} = \frac{m_i}{\|\nabla_\theta h_i(\theta)\|_2}}. \tag{28}$$

The corresponding minimum-norm perturbation is

$$\Delta\theta_i^\star = -\frac{m_i}{\|g_i\|_2^2} g_i. \tag{29}$$

Proof. Reaching the linearized boundary requires

$$-g_i^\top \Delta\theta \geq m_i. \tag{30}$$

By the Cauchy–Schwarz inequality,

$$-g_i^\top \Delta\theta \leq \|g_i\|_2 \|\Delta\theta\|_2, \tag{31}$$

and therefore

$$\|\Delta\theta\|_2 \geq \frac{m_i}{\|g_i\|_2}. \tag{32}$$

Equality is achieved by

$$\Delta\theta = -\frac{m_i}{\|g_i\|_2^2} g_i, \tag{33}$$

which proves the result. □

Corollary 1 reveals the mechanism linking logit margin to parameter-space removal effort. The first-order removal effort is determined by the amount of retained true-class preference m_i, modulated by the local parameter sensitivity $\|g_i\|_2$. In particular,

$$\mathcal{R}_i^{\text{lin}} = \frac{m_i}{\|g_i\|_2}. \tag{34}$$

Thus, for comparable local sensitivities, samples with larger margins require larger parameter-space interventions to erase their current true-class preference:

$$\mathcal{R}_i^{\text{lin}} \propto m_i. \tag{35}$$

As verified empirically below, raw margin preserves this removal-effort ordering remarkably well across different forget configurations, despite sample-wise variation in $\|g_i\|_2$. This establishes the margin as a practical prediction-level signal for allocating forgetting intensity.

This perspective also provides a complementary connection to influence-based unlearning. Influence-based quantities characterize how strongly model parameters or objectives respond to individual samples, whereas the logit margin directly characterizes how much true-class decision support remains in the current predictor. Corollary 1 connects these two views through parameter-space removal effort: the margin specifies the decision support that must be erased, while the gradient term specifies the local sensitivity through which parameter updates alter that support. `IMPRINT` exploits the former as a directly available sample-wise signal, enabling removal-effort-aware weighting without per-example gradient or Jacobian estimation.

Empirical validation of raw margin as a removal-effort proxy. The first-order result above raises a natural question: although equation 28 depends on both m_i and $\|g_i\|_2$, does the raw margin preserve enough of the resulting sample-wise removal-effort ordering to serve as a practical proxy?

We test this mechanism using 1,000 positive-margin samples from the CIFAR-100 DER++/ResNet-18 source model with replay buffer size 2,000 and the same canonical 30%-class-wise forgetting protocol used in the main experiments. For each sample, we compute the raw margin m_i, the full-parameter margin-gradient norm $\|g_i\|_2 = \|\nabla_\theta h_i(\theta)\|_2$, and the theoretically motivated first-order removal score

$$\mathcal{R}_i^{\text{lin}} = \frac{m_i}{\|g_i\|_2}. \tag{36}$$

Directional empirical decision-boundary crossing. We additionally measure the nonlinear decision-boundary crossing along the first-order margin-reducing direction. For each positive-margin sample, we define the unit-norm direction

$$d_i = -\frac{g_i}{\|g_i\|_2}, \tag{37}$$

where $g_i = \nabla_\theta h_i(\theta)$ is evaluated at the source checkpoint. We hold this direction fixed throughout the search. Because $\|d_i\|_2 = 1$, the scalar displacement a in $\theta + ad_i$ is exactly the parameter-space L_2 distance traveled along this ray.

At every trial point, we recompute the exact margin

$$h_i(\theta + ad_i) = z_{y_i}(x_i; \theta + ad_i) - \max_{c \neq y_i} z_c(x_i; \theta + ad_i), \tag{38}$$

including a fresh maximization over the competing classes. Thus, the identity of the strongest competing class is allowed to change during the search.

We use the first-order radius

$$\mathcal{R}_i^{\mathrm{lin}} = \frac{m_i}{\|g_i\|_2} \tag{39}$$

as a sample-specific search scale. We first evaluate the exact margin at

$$\frac{a}{\mathcal{R}_i^{\mathrm{lin}}} \in \{0.125,\ 0.25,\ 0.5,\ 0.75,\ 1.0\}. \tag{40}$$

If no crossing is detected, we geometrically increase the distance by a factor of 1.5, up to a maximum search radius of $64\mathcal{R}_i^{\mathrm{lin}}$. The first pair of consecutive evaluated points satisfying

$$h_i(\theta + a_{\mathrm{lo}} d_i) > 0, \qquad h_i(\theta + a_{\mathrm{hi}} d_i) \leq 0 \tag{41}$$

defines the crossing bracket. We then perform 20 iterations of bisection, with

$$a_{\mathrm{mid}} = \frac{a_{\mathrm{lo}} + a_{\mathrm{hi}}}{2}, \tag{42}$$

updating the upper endpoint whenever $h_i(\theta + a_{\mathrm{mid}} d_i) \leq 0$ and the lower endpoint otherwise. We report the final crossed-side endpoint as

$$\mathcal{R}_i^{\mathrm{emp}} = a_{\mathrm{hi}}. \tag{43}$$

Returning the upper endpoint ensures that the reported perturbation lies on the crossed side of the decision boundary. The source model is evaluated functionally in evaluation mode, so the search does not modify the source parameters or BatchNorm statistics.

If no crossing is detected by $64\mathcal{R}_i^{\mathrm{lin}}$, the observation is treated as right-censored rather than being assigned the search limit. The empirical crossing distance was also close to its first-order prediction: in the original configuration, the median ratio $\mathcal{R}_i^{\mathrm{emp}}/\mathcal{R}_i^{\mathrm{lin}}$ was 1.011, with a P10–P90 range of 0.972–1.079, and the additional configurations yielded similar median ratios.

To test whether raw margin can serve as a practical proxy for the first-order removal effort, we first compare m_i with $\mathcal{R}_i^{\mathrm{lin}} = m_i/\|g_i\|_2$ over the 1,000 analyzed positive-margin samples using Spearman rank correlation. We obtain

$$\rho\left(m_i, \mathcal{R}_i^{\mathrm{lin}}\right) = 0.960, \qquad p = 5.47 \times 10^{-554}. \tag{44}$$

This indicates that samples ranked as requiring larger removal effort by the first-order expression are also ranked very similarly by raw margin alone. Thus, raw margin preserves most of the first-order removal-effort ordering while avoiding per-sample gradient computation.

We next compare raw margin with the directional empirical crossing distance. Using Spearman rank correlation, we obtain

$$\rho\left(m_i, \mathcal{R}_i^{\mathrm{emp}}\right) = 0.959, \qquad p = 7.44 \times 10^{-547}. \tag{45}$$

Thus, samples with larger raw margins generally require larger parameter-space displacements along the normalized negative-margin-gradient direction before their current true-class preference is removed.

The theoretically derived first-order removal effort is even more closely aligned with the directional empirical crossing distance:

$$\rho\left(\mathcal{R}_i^{\mathrm{lin}}, \mathcal{R}_i^{\mathrm{emp}}\right) = 0.997, \qquad p = 1.68 \times 10^{-1135}. \tag{46}$$

This confirms that the first-order approximation closely describes the observed boundary-crossing behavior along the corresponding margin-gradient direction, while raw margin alone preserves nearly the same sample-wise ordering. Importantly, $\mathcal{R}_i^{\mathrm{emp}}$ is a directional empirical crossing distance along a fixed ray; it is not claimed to be the globally shortest nonlinear parameter perturbation over all possible directions.

These results distinguish two complementary statements. First, equation 28 shows theoretically that parameter-space decision-removal effort depends on both current margin and local sensitivity. Second, the empirical analysis shows that in the continually trained models, raw margin preserves most of the resulting sample ordering even after the sensitivity term is included. This provides empirical support for using margin alone as a computationally inexpensive removal-effort proxy.

Nonpositive-margin samples. The previous analysis focuses on $m_i > 0$, for which a positive intervention is required to eliminate the current true-class preference. Samples with $m_i \leq 0$ already satisfy

$$h_i(\theta) \leq 0, \tag{47}$$

and therefore have zero decision-level removal radius under equation 16:

$$\mathcal{R}_i(\theta) = 0. \tag{48}$$

Under the 30%-class-wise forgetting setup, 34.9% of forget examples already have nonpositive source margin. `IMPRINT` assigns these examples a median weight of 0.0053, whereas positive-margin examples receive weights concentrated near one. This behavior is consistent with the intended mechanism: forget samples whose true-class decision preference has already disappeared receive little additional forgetting pressure, whereas samples that remain strongly supported retain substantially larger weights.

Connection to imprint-aware unlearning. For a base unlearning method with per-sample forgetting loss ℓ_i^f, `IMPRINT` applies

$$\mathcal{L}_{\mathrm{IMPRINT}} = \frac{1}{|\mathcal{D}_f|} \sum_{i \in \mathcal{D}_f} \omega_i \ell_i^f + \alpha \mathcal{L}_r, \qquad \omega_i = \sigma\left(\frac{m_i}{\tau}\right). \tag{49}$$

Using Proposition 1,

$$m_i = \sqrt{2}\mathcal{I}_i, \tag{50}$$

and hence

$$\omega_i = \sigma\left(\frac{\sqrt{2}\mathcal{I}_i}{\tau}\right). \tag{51}$$

Thus, `IMPRINT` applies a monotonic transformation of the current decision-level imprint when allocating the forget loss. Samples with large positive margins retain relatively large forgetting weights because more decision-level support must be removed before their true-class preference disappears. Samples close to the decision boundary receive smaller weights, while samples whose ground-truth preference is already reversed are increasingly suppressed.

Because $0 < \omega_i < 1$, `IMPRINT` uses the weighting function to concentrate forgetting effort on samples that retain stronger decision-level support. Samples with larger margins receive larger weights and therefore contribute more strongly to the forgetting objective, while samples with weaker or already reversed true-class preference contribute less. This produces a sample-adaptive forgetting objective in which the optimization effort follows the estimated decision-level removal requirement of each forget sample.

The empirical mechanism analysis further supports the use of raw margin in place of the theoretically sharper quantity $m_i / \|\nabla_\theta h_i\|_2$. While the latter requires a per-example backward pass through the model, raw margin is obtained directly from the forward pass. In our analysis, raw margin shows

a strong rank correlation with both the first-order removal effort ($\rho = 0.960$) and the directional empirical decision-boundary crossing distance ($\rho = 0.959$). These results show that raw margin preserves most of the removal-effort ordering while providing a substantially simpler and more efficient sample-wise signal for IMPRINT.

The resulting mechanistic interpretation is

$$\boxed{\begin{aligned}&\text{logit margin} \iff \text{decision-level imprint}\\ &\qquad\qquad \implies \text{parameter-space removal effort} \implies \text{sample-wise forgetting intensity.}\end{aligned}} \tag{52}$$

The first relation is exact in logit space (Proposition 1). The connection to parameter-space removal effort is established by the lower bound in Theorem 1 and the local first-order expression in Corollary 1. The empirical analysis further shows that raw margin closely preserves this theoretically motivated removal-effort ordering.

Together, these results connect heterogeneous decision-level imprint to the sample-adaptive weighting used by IMPRINT. The logit margin provides a directly observable and computationally efficient signal of how much current true-class support remains to be removed, thereby motivating its use for allocating sample-wise forgetting intensity.

B.2 Comparison of Non-Uniform Unlearning Strategies

Both IMU and NPO are also sample-adaptive, but they allocate unlearning strength differently. IMU computes a classifier-level influence score for each sample in $\mathcal{D}_f$, ranks the samples, and excludes samples with non-positive scores. It then selects the top-$p\%$ samples and applies influence-based gradient weights. In contrast, IMPRINT unlearns all samples in $\mathcal{D}_f$ and continuously adjusts only their forgetting intensity. Thus, weakly imprinted samples receive very small weights, whereas strongly imprinted samples receive larger weights.

NPO adjusts unlearning strength according to the current model's prediction relative to a reference model. When the reference is the pre-unlearning model and unlearning starts from that same model, all samples initially receive the same intensity, regardless of how much they were already weakened during continual learning. NPO begins reducing the intensity only after a sample's prediction decreases relative to the reference. IMPRINT instead captures the sample's existing imprint through its initial logit margin and immediately weakens the update for samples that are weakly imprinted. Its weights can also be recomputed as unlearning progresses.

B.3 Controlled Comparison of Margin- and Influence-Based Weighting

Section 3.3 shows that logit margin is strongly rank-correlated with the classifier-level influence estimate used by IMU. We next examine whether these two related signals are interchangeable when used specifically to allocate forgetting intensity. To isolate the allocation signal from other algorithmic differences, we construct controlled influence-weighted variants of random labeling (RL) and SalUn.

Controlled construction of influence-based weights. We use the classifier-level influence estimate employed by IMU while removing its sparse top-$p\%$ sample selection, so that all forget samples contribute to the objective. For each forget sample i, let g_i denote the gradient of its cross-entropy loss with respect to the final linear classifier. We define

$$D = \operatorname{diag}\left(\frac{1}{N_f}\sum_{j=1}^{N_f} g_j^2 + 0.01\right), \qquad \bar{g} = \frac{1}{N_f}\sum_{j=1}^{N_f} g_j, \tag{53}$$

and compute the signed classifier-level influence score

$$\phi_i^{\mathrm{IF}} = -g_i^\top D^{-1}\bar{g}. \tag{54}$$

This is the damped diagonal-Fisher approximation used by IMU, restricted to the final classifier.

To obtain dense positive sample weights on a common scale, we compute the global root mean square magnitude of the influence scores,

$$\mathrm{RMS}(\phi^{\mathrm{IF}}) = \sqrt{\frac{1}{N_f}\sum_{j=1}^{N_f}\left(\phi_j^{\mathrm{IF}}\right)^2}, \tag{55}$$

Table 9: **Controlled comparison of logit-margin and influence-based weighting.** Results on CIFAR-100 with DER++ across five paired 30% class-wise forgetting configurations with buffer size 2000.

Method	FA↓	RA↑	MIA↑	FT↓	RT↑	Avg. Gap↓
Retrain	0.000	62.138	100.000	0.000	55.346	—
RL+Influence	0.908	49.818	99.840	3.987	53.569	$3.830_{\pm 0.779}$
RL+IMPRINT	0.108	61.325	99.679	0.053	56.320	$\mathbf{0.686_{\pm 0.461}}$
SalUn+Influence	2.677	53.944	99.332	7.193	54.383	$3.963_{\pm 0.670}$
SalUn+IMPRINT	0.244	61.493	99.339	0.127	56.323	$\mathbf{0.815_{\pm 0.717}}$

and define

$$\omega_i^{\mathrm{IF}} = \sigma\left(-\frac{\phi_i^{\mathrm{IF}}}{\mathrm{RMS}(\phi^{\mathrm{IF}})}\right). \tag{56}$$

The sign follows the IMU convention, such that samples receiving stronger influence scores under its ordering are assigned larger forgetting weights.

We replace the IMPRINT weight in RL and SalUn with ω_i^{IF}, while leaving every other component of the unlearning procedure unchanged.

As shown in Table 9, IMPRINT has a substantially smaller gap to retraining for both RL and SalUn. The results show that aligning update intensity with sample-wise imprint is more effective than using classifier-level influence.

B.4 Conversion of Class Accuracy and Entropy to Forget-Loss Weights

The proxy ablation in Figure 5 and Tables 10, 11 compares three sample-influence scores: (i) the per-class *accuracy* of the class to which the sample belongs, (ii) the model's predictive *entropy* on the sample, and (iii) the *logit margin* (used by IMPRINT).

Logit-margin proxy (IMPRINT). The raw score is the (signed) logit margin $s_i^{\mathrm{margin}} = \mathrm{Margin}(x_i)$ which is naturally centered at zero (sign indicates correctness). The weight is

$$\omega_i^{\mathrm{margin}} = \sigma\left(\frac{s_i^{\mathrm{margin}}}{\tau}\right). \tag{57}$$

Accuracy proxy. For class-accuracy-based weighting, we first record the best historical class-wise accuracy achieved for each buffer class during the class-incremental learning (CIL) process. Let this reference accuracy for class c be denoted by $A_c^{\max}$.

At the current unlearning stage, we evaluate the model on the buffer samples and measure the current class-wise accuracy for each class c, denoted by A_c^{cur}.

The class-dependent scaling factor is then computed as the ratio between the current class accuracy and the highest historical class accuracy:

$$s_c = \frac{A_c^{\mathrm{cur}}}{A_c^{\max}}.$$

Thus, for a forget sample whose original label is y_i, the class-accuracy weight is given by

$$w_i^{\mathrm{acc}} = s_{y_i} = \frac{A_{y_i}^{\mathrm{cur}}}{A_{y_i}^{\max}}.$$

This weighting assigns the same scale to all forget samples belonging to the same original class. Classes whose current accuracy remains close to their best historical performance receive a larger weight, while classes whose current accuracy has already degraded receive a smaller weight.

Entropy proxy. For entropy-based weighting, the weight is computed from the model's current predictive uncertainty. The predictive entropy of sample x_i is

$$H(p_i) = -\sum_{c=1}^{C} p_{i,c} \log p_{i,c},$$

where C is the total number of classes and $p_{i,c}$ denotes the predicted probability that sample x_i belongs to class c. We normalize the entropy by its maximum possible value, $\log C$, as

$$\bar{H}(p_i) = \frac{H(p_i)}{\log C}.$$

The entropy-based weight is then defined as

$$w_i^{\text{ent}} = \lambda_{\max}\left(1 - \bar{H}(p_i)\right) = \lambda_{\max}\left(1 - \frac{-\sum_{c=1}^{C} p_{i,c} \log p_{i,c}}{\log C}\right),$$

where $\lambda_{\max}$ controls the maximum strength of the entropy weighting and is set to $\lambda_{\max} = 1$ in all cases. This formulation gives larger weights to samples for which the model is still confident, and smaller weights to samples whose predictions are already uncertain.

B.5 Extended Comparison of Sample-Influence Proxies

We provide the full numerical results behind Figure 5 of the main paper, expanded with an additional dataset (Tiny-ImageNet). Across both datasets and three base unlearning methods (RL, SalUn, AMUN), the logit-margin proxy used by IMPRINT achieves the smallest average gap to retraining. The accuracy proxy typically offers a modest improvement over uniform weighting. In contrast, the entropy proxy performs worst on the forget metrics (FA / FT), because it cannot distinguish confidently correct predictions from confidently incorrect ones (§B.4).

Table 10: **Comparison of three sample-influence proxies**: class accuracy, entropy, and logit margin (+IMPRINT), each integrated into RL, SalUn, and AMUN for class-wise unlearning on **CIFAR-100**. These numerical results correspond to the summary in Figure 5. The logit-margin proxy achieves the lowest average gap across all base methods.

	Buffer size = 500						Buffer size = 2000					
Method	**FA**	**RA**	**MIA**	**FT**	**RT**	**Avg. Gap**	**FA**	**RA**	**MIA**	**FT**	**RT**	**Avg. Gap**
Retrain	0.00	43.99	100.00	0.00	40.90	—	0.00	64.11	100.00	0.00	57.18	—
RL	0.06	28.03	99.72	0.31	37.83	$3.94_{\pm 0.88}$	0.30	47.90	98.97	1.66	51.81	$4.92_{\pm 0.89}$
RL+acc	0.55	34.42	98.76	1.61	39.81	$3.02_{\pm 0.73}$	0.92	50.15	97.84	3.57	52.44	$5.07_{\pm 0.68}$
RL+entropy	1.50	36.00	97.50	3.39	40.31	$3.54_{\pm 0.39}$	3.44	52.15	94.68	8.03	52.81	$6.62_{\pm 0.42}$
RL+IMPRINT	0.25	43.26	99.38	0.15	41.42	$\mathbf{1.36_{\pm 1.29}}$	0.30	61.27	99.22	0.18	56.29	$\mathbf{1.00_{\pm 0.82}}$
SalUn	0.42	33.94	99.88	1.19	39.66	$2.87_{\pm 1.16}$	1.18	52.46	99.76	3.77	52.90	$4.22_{\pm 0.31}$
SalUn+acc	1.80	38.16	92.24	3.38	41.15	$4.23_{\pm 0.77}$	2.37	54.01	95.27	5.61	53.55	$5.29_{\pm 0.35}$
SalUn+entropy	0.42	33.94	96.71	1.19	39.66	$3.50_{\pm 1.04}$	1.18	52.46	96.80	3.77	52.90	$4.81_{\pm 0.27}$
SalUn+IMPRINT	0.36	42.61	99.11	0.21	40.72	$\mathbf{1.52_{\pm 1.50}}$	0.35	61.16	99.23	0.20	56.05	$\mathbf{1.08_{\pm 1.10}}$
AMUN	9.30	27.14	84.98	9.53	25.94	$13.13_{\pm 1.53}$	25.14	49.80	73.40	23.41	46.11	$20.11_{\pm 3.96}$
AMUN+acc	10.03	27.85	83.89	10.05	26.66	$13.31_{\pm 1.50}$	25.84	50.11	72.55	23.71	46.31	$20.37_{\pm 4.07}$
AMUN+entropy	10.83	28.12	83.02	10.58	26.91	$13.65_{\pm 1.61}$	27.14	50.42	71.37	24.67	46.49	$20.97_{\pm 4.08}$
AMUN+IMPRINT	2.13	44.55	96.15	1.79	42.43	$\mathbf{2.74_{\pm 2.02}}$	5.83	62.58	91.85	4.84	56.89	$\mathbf{4.14_{\pm 4.05}}$

B.6 Distribution of Sample-Wise Imprints

We characterize the pre-unlearning imprint distribution over entire samples from the CIFAR-100. Table 12 summarizes the resulting margin and IMPRINT-weight distributions. The logit margins span a wide range, from -18.73 to 24.00, with 37.39% of the samples having negative margins. This indicates that a substantial fraction is already misclassified before unlearning.

Table 11: **Proxy comparison** on **Tiny-ImageNet** under 30% class-wise unlearning. Same protocol as Table 10.

	Buffer size = 500						Buffer size = 5120					
Method	**FA**	**RA**	**MIA**	**FT**	**RT**	**Avg. Gap**	**FA**	**RA**	**MIA**	**FT**	**RT**	**Avg. Gap**
Retrain	0.00	28.26	100.00	0.00	23.78	—	0.00	57.49	100.00	0.00	44.73	—
RL	0.02	15.47	99.84	0.00	15.41	$4.37_{\pm 2.45}$	0.00	31.54	100.00	0.00	31.38	$7.86_{\pm 1.55}$
RL+acc	0.24	18.78	99.47	0.02	17.76	$3.74_{\pm 1.81}$	0.00	33.38	100.00	0.00	33.56	$7.06_{\pm 1.09}$
RL+entropy	0.40	18.46	99.46	0.19	17.99	$3.61_{\pm 1.64}$	0.05	31.95	99.93	0.06	33.46	$7.40_{\pm 1.25}$
RL+IMPRINT	0.26	21.21	99.30	0.03	18.08	$\mathbf{3.14_{\pm 1.07}}$	0.09	41.32	99.05	0.01	35.05	$\mathbf{5.38_{\pm 2.20}}$
SalUn	0.16	18.68	99.74	0.01	17.97	$3.51_{\pm 1.85}$	0.02	32.41	99.99	0.01	33.15	$7.34_{\pm 1.96}$
SalUn+acc	0.64	21.98	94.57	0.21	19.85	$4.37_{\pm 1.18}$	0.13	34.99	97.65	0.04	34.62	$7.03_{\pm 2.08}$
SalUn+entropy	0.16	18.68	98.97	0.01	17.97	$3.66_{\pm 1.86}$	0.02	32.41	99.20	0.01	33.15	$7.50_{\pm 2.00}$
SalUn+IMPRINT	0.35	21.00	99.19	0.07	17.57	$\mathbf{3.37_{\pm 1.44}}$	0.52	40.06	97.39	0.07	32.98	$\mathbf{6.48_{\pm 2.61}}$
AMUN	6.84	13.02	58.28	5.41	11.31	$16.34_{\pm 10.98}$	17.81	36.19	75.12	13.58	29.92	$18.48_{\pm 3.75}$
AMUN+acc	7.46	13.86	75.44	5.61	11.83	$12.81_{\pm 8.40}$	18.49	35.95	74.32	13.87	29.50	$18.96_{\pm 3.66}$
AMUN+entropy	7.40	13.54	67.97	5.62	11.44	$14.42_{\pm 8.71}$	19.44	36.96	73.44	14.35	30.31	$19.06_{\pm 3.88}$
AMUN+IMPRINT	3.92	16.82	92.22	2.69	14.69	$\mathbf{6.99_{\pm 2.29}}$	9.18	37.05	83.79	6.72	32.84	$\mathbf{12.89_{\pm 5.04}}$

Table 12: **Distribution of logit margins and the resulting IMPRINT weights.**

Statistic	**Margin**	**Weight**	**Interpretation**
Minimum	−18.7311	5.375×10^{-17}	Weakest imprint
10th percentile	−3.8768	0.000429	Weakly imprinted
Median	1.7021	0.967838	Intermediate imprint
90th percentile	8.4055	1.000000	Strongly imprinted
Maximum	24.0045	1.000000	Strongest imprint

As shown in Table 12, the sigmoid weighting converts a broad pre-unlearning imprint scale into strongly differentiated update intensities. This behavior allows IMPRINT to nearly suppress unnecessary updates for already-weak samples while retaining strong updates for samples with large positive margins.

For completeness, Table 13 reports the exact samples attaining the global minimum and maximum margins.

B.7 Comparison with Training-History Surrogates

We further compare IMPRINT with two training-history-based alternatives: *loss variance* and *task forgetting* on CIFAR-100. These surrogates capture different aspects of how well previously learned knowledge is retained over the continual-learning trajectory.

Loss variance measures how much the model's true-label loss for an individual training example fluctuates across the available task-end checkpoints. For example i, let $\ell_i^{(t)}$ denote its cross-entropy loss at checkpoint t, and let H_i denote the set of checkpoints retained for that example. We compute

$$V_i = \frac{1}{|H_i|} \sum_{t \in H_i} \left(\ell_i^{(t)} - \bar{\ell}_i\right)^2, \qquad \bar{\ell}_i = \frac{1}{|H_i|} \sum_{t \in H_i} \ell_i^{(t)}.$$

A small variance means that the example's loss remains relatively stable throughout the observed training history, whereas a large variance indicates greater temporal fluctuation. We therefore use $s_i^{\text{lossvar}} = -V_i$, so that larger surrogate scores consistently correspond to stronger retained knowledge. This quantity reflects *stability* rather than absolute prediction quality: an example can have low variance even if its loss remains consistently high. For Tasks 1–9, the retained history also includes the checkpoint immediately preceding the task's introduction.

Task forgetting measures how much performance on an entire previously learned task declines from its best observed level to the end of continual training. For task k, with $A_{k,t}$ denoting its training accuracy after task t, we compute

$$F_k = \max_{t \in \{k,\ldots,9\}} A_{k,t} - A_{k,9}.$$

Table 13: **Samples attaining the global minimum and maximum pre-unlearning margins.**

Extreme	Sample information	Prediction	Margin	Weight
Minimum	Index 18,456; config. 1; task 5; class 49	71 (incorrect)	-18.731079	5.375×10^{-17}
Maximum	Index 49,064; config. 2; task 4; class 39	39 (correct)	24.004461	1.000000

Table 14: **Comparison with training-history-based imprint surrogates.** `IMPRINT` obtains the smallest average gap to matched retraining among the tested weighting variants while avoiding the need to store training-history statistics.

Method	FA↓	RA↑	MIA↑	FT↓	RT↑	Avg. Gap↓
Retrain	0.00	62.26	100.00	0.00	54.79	—
RL	0.21	51.30	99.97	1.33	52.05	$3.05_{\pm 0.48}$
Loss variance	2.35	56.06	99.53	5.40	53.97	$3.18_{\pm 0.79}$
Task forgetting	5.12	55.88	97.31	7.33	53.33	$4.65_{\pm 0.77}$
`IMPRINT`	0.10	61.28	99.66	0.09	55.98	$\mathbf{0.70_{\pm 0.45}}$

Thus, $F_k = 0$ means that the task finishes at its best observed training accuracy, whereas a larger value indicates a larger loss of task-level performance by the final checkpoint. We again reverse the orientation and use $s_k^{\text{forget}} = -F_k$, so that larger values indicate less forgetting. Every example belonging to task k receives this same task-level score. Unlike loss variance and logit margin, task forgetting is therefore not sample-specific; with ten continual-learning tasks, it can produce at most ten distinct scores.

For each forget configuration, the two history-based surrogate scores are converted into unlearning weights by rank-calibrating them to reference logit margin weights computed from the final Task-9 source model, $r_i = \sigma(m_i^{(9)}/0.5)$. Higher-ranked examples under each surrogate consequently receive larger logit margin-derived weights. Exact-score ties are assigned the mean of the corresponding reference-weight block. Table 14 summarizes the resulting comparison.

As shown in Table 14, `IMPRINT` achieves the smallest average gap to matched retraining among the tested weighting variants. Loss variance performs substantially better than task forgetting on FA and FT, while both history-based variants remain farther from retraining than `IMPRINT` overall. These results highlight an important practical advantage of logit margin: it provides a sample-specific measure of the model's current preference for the correct class over its strongest alternative. In comparison, loss variance captures temporal stability rather than the strength of the current prediction, while task forgetting provides only a coarser task-level signal. The history-based alternatives additionally require information from earlier task-end checkpoints. By contrast, `IMPRINT` derives its weights directly from the current model and the forget minibatch during unlearning.

Robustness to task-wise logit scale. A potential concern is that recent tasks may exhibit larger logit margins simply because their logits have a larger numerical scale. We therefore first audit the natural task-wise logit scale and then conduct an exaggerated scale-stress experiment. For a sample i with logits $z_i \in \mathbb{R}^{100}$, we define its logit scale as

$$d_i = \sqrt{\frac{1}{100}\sum_{c=1}^{100}(z_{i,c} - \bar{z}_i)^2},$$

and define the robust scale of task t as

$$s_t = \text{median}_{i:t_i=t}\, d_i.$$

On CIFAR-100 with DER++ and buffer size 2,000, the ratio between the geometric means of s_t over the final three and first three tasks is 1.032, indicating little systematic increase in logit magnitude for recent tasks.

Table 15: **Robustness to task-wise logit-scale variation.** The $4\times$ stress modifies only the logits used to compute the IMPRINT margin, while optimization and evaluation use the original logits.

Condition	FA↓	RA↑	MIA↑	FT↓	RT↑	Avg. Gap↓
Retrain	0.00	62.14	100.00	0.00	55.35	0.00
RL	0.18	45.27	99.98	1.23	52.65	4.20
Raw margin ($4\times$)	0.16	61.11	99.58	0.06	56.36	0.76
Raw margin	0.12	61.36	99.66	0.04	56.40	**0.69**

As a complementary calibration diagnostic, we independently fit a scalar temperature T_t for each task by minimizing multiclass negative log-likelihood on a held-out calibration split:

$$T_t^* = \arg\min_{T\in[0.05,20]} -\frac{1}{|\mathcal{C}_t|}\sum_{i\in\mathcal{C}_t}\log\frac{\exp(z_{i,y_i}/T)}{\sum_c \exp(z_{i,c}/T)}.$$

The ratio between the largest and smallest fitted temperatures is 1.459. These temperatures are used only as a diagnostic and are distinct from the IMPRINT sigmoid temperature τ. Together, these measurements suggest that the larger logit margins observed for recent tasks are unlikely to be explained solely by a systematic increase in logit scale.

To further test sensitivity to task-dependent logit magnitude, we introduce an exaggerated artificial scale distortion *only* when computing the logit margin. For a forget sample i originating from task $t_i \in \{0,\dots,9\}$, we replace z_i in the margin computation by

$$\tilde{z}_i = a_{t_i} z_i,$$

where

$$b_t = 4^{\,t/9-1/2}, \qquad a_t = \frac{b_t}{\exp\left(\frac{1}{|\mathcal{T}_F|}\sum_{u\in\mathcal{T}_F}\log b_u\right)},$$

and $\mathcal{T}_F$ denotes the set of tasks represented in the current forget configuration. The geometric mean of a_t over the represented tasks is one, avoiding an overall increase in logit scale, while the oldest and newest tasks differ by $4\times$ when both are represented. Thus, the newest-to-oldest represented-task ratio is at most $4\times$.

Because the entire logit vector is multiplied by the same positive scalar, this perturbation does not change the predicted class or the sign of the margin. In particular, for

$$m_i = z_{i,y_i} - \max_{c\neq y_i} z_{i,c},$$

the stressed margin becomes

$$\tilde{m}_i = a_{t_i} m_i.$$

Thus the perturbation changes only the magnitude of the margin, and consequently its IMPRINT weight. The scaling is applied only to the logits used to compute the margin; all logits used for optimization and evaluation remain unchanged. Table 15 reports the resulting stress test.

Table 15 shows that the default raw-margin formulation remains stable even under the exaggerated $4\times$ task-wise scale perturbation. This perturbation is substantially larger than the task-wise scale variation observed in the original model, providing additional evidence that the effectiveness of IMPRINT is not primarily driven by systematic differences in logit magnitude across tasks.

B.8 Mean-Matched Uniform Forget-Loss Weighting

Motivation. IMPRINT changes both the relative allocation of forget-loss weights across examples and their average magnitude. We therefore examine whether its performance can be reproduced by applying a uniform coefficient that matches its average forget-loss weight at every optimization step.

Controlled comparison. We evaluate three conditions on CIFAR-100 with ResNet-18 and DER++, using a replay buffer of 2,000 and five paired 30-class forgetting configurations: RL, RL+IMPRINT, and mean-matched uniform RL. Each condition runs for 10 unlearning epochs with batch size 256.

Table 16: Mean-matched uniform control on CIFAR-100 with DER++ and buffer size 2,000. Metrics are means over five paired forgetting configurations.

Method	FA	RA	MIA	FT	RT	Avg. Gap
Retrain	0.00	**62.14**	100.00	0.00	**55.35**	—
RL	0.179	45.274	99.978	1.233	52.649	4.199 ± 1.050
RL+IMPRINT	0.121	61.363	99.656	0.040	56.400	$\mathbf{0.692 \pm 0.444}$
Mean-matched uniform	43.560	63.146	68.880	39.553	56.449	23.458 ± 1.285

During the paired IMPRINT run, we record the mean weight of each forget minibatch $\mathcal{B}_t$:

$$\bar{\omega}_t = \frac{1}{|\mathcal{B}_t|} \sum_{i \in \mathcal{B}_t} \omega_{i,t}^{\text{IMPRINT}}. \tag{58}$$

The uniform control replays this recorded coefficient and assigns it to every example in the corresponding forget minibatch:

$$\mathcal{L}_{f,t}^{\text{uniform}} = \frac{0.5\,\bar{\omega}_t}{|\mathcal{B}_t|} \sum_{i \in \mathcal{B}_t} \ell_{\text{CE}}\big(\theta_t^{\text{uniform}}; x_i, y_i'\big), \qquad y_i' \neq y_i. \tag{59}$$

The coefficients are treated as constants in the control and are not recomputed from its evolving model.

Results and interpretation. As shown in Table 16, the uniform control preserves retain accuracy but fails to achieve comparable forgetting. IMPRINT obtains a smaller average gap to the retrained reference. Thus, replaying IMPRINT's average forget-loss coefficient schedule alone is insufficient to reproduce its forgetting–retention trade-off.

In IMPRINT, gradients propagate through the sample-dependent weights, whereas the uniform control uses fixed recorded coefficients. This comparison therefore jointly tests sample-dependent weighting and its associated weight-gradient effects.

C Sensitivity and Robustness Analyses

C.1 Sensitivity Analyses

Temperature τ. The sigmoid temperature controls how sharply the weights separate strongly from weakly imprinted samples. Small τ approaches a hard 0/1 gate at $\text{Margin}(x_i)=0$, while large τ flattens the weights toward 0.5 for all samples (recovering uniform unlearning in the limit $\tau \to \infty$). To test robustness, we sweep $\tau \in \{0.5,\ 1.0\}$ on CIFAR-100 with DER++ (buffer size 500 and 2000), under class-wise forgetting at the default 30% ratio, and report the average gap to retraining for RL+`IMPRINT`, SalUn+`IMPRINT`, and AMUN+`IMPRINT` in Table 17

Result. For RL and SalUn, τ=0.5 achieves the lowest average gap for both buffer sizes, but τ=1 remains close and still improves over the corresponding base methods. For AMUN, both temperature settings yield large gains over the base method, with τ=1 giving a slightly smaller average gap and τ=0.5 providing comparable RA/RT in several cases. These results indicate that performance is stable across reasonable temperature choices rather than being highly sensitive to a specific value.

C.2 Effect of the Forgetting Ratio

The main paper uses a default class-wise forgetting ratio of 30%. We additionally evaluate two smaller ratios, 10% and 20%, on CIFAR-100 with DER++. For each ratio we report (i) a comparison of `IMPRINT` against all baselines (Tables 18, 20), and (ii) the compatibility with existing unlearning methods (Tables 19, 21).

Trends across ratios. Smaller forgetting ratios involve fewer classes and therefore a forget set whose imprint distribution is narrower. Even in this regime `IMPRINT` consistently attains the smallest average gap to retraining, uniformly outperforming existing baselines. The improvement of `IMPRINT` over uniform unlearning therefore is not specific to the 30% ratio used in the main experiments and persists across the practical range of deletion-request sizes.

Table 17: **Temperature ablation** on CIFAR-100 under DER++. We compare the base methods with temperature settings $\tau = 0.5$ and $\tau = 1$.

Buffer size = 500							Buffer size = 2000						
Method	**FA**	**RA**	**MIA**	**FT**	**RT**	**Avg. Gap**	**Method**	**FA**	**RA**	**MIA**	**FT**	**RT**	**Avg. Gap**
Retrain	0.00	43.99	100.00	0.00	40.90	—	Retrain	0.00	64.11	100.00	0.00	57.18	—
RL	0.06	28.03	99.72	0.31	37.83	$3.94_{\pm 0.88}$	RL	0.30	47.90	98.97	1.66	51.81	$4.92_{\pm 0.89}$
RL+τ0.5	0.25	43.26	99.38	0.15	41.42	$\mathbf{1.36_{\pm 1.29}}$	RL+τ0.5	0.30	61.27	99.22	0.18	56.29	$\mathbf{1.00_{\pm 0.82}}$
RL+τ1	0.28	38.69	99.32	0.15	37.80	$2.00_{\pm 1.30}$	RL+τ1	0.18	59.89	99.44	0.12	55.02	$1.45_{\pm 0.94}$
SalUn	0.42	33.94	99.88	1.19	39.66	$2.87_{\pm 1.16}$	SalUn	1.18	52.46	99.76	3.77	52.90	$4.22_{\pm 0.31}$
SalUn+τ0.5	0.36	42.61	99.11	0.21	40.72	$\mathbf{1.52_{\pm 1.50}}$	SalUn+τ0.5	0.35	61.16	99.23	0.20	56.05	$\mathbf{1.08_{\pm 1.10}}$
SalUn+τ1	0.24	37.25	99.42	0.13	36.43	$2.43_{\pm 1.28}$	SalUn+τ1	0.25	59.10	99.30	0.17	54.30	$1.80_{\pm 1.14}$
AMUN	9.30	27.14	84.98	9.53	25.94	$13.13_{\pm 1.53}$	AMUN	25.14	49.80	73.40	23.41	46.11	$20.11_{\pm 3.96}$
AMUN+τ0.5	2.13	44.55	96.15	1.79	42.43	$2.74_{\pm 2.02}$	AMUN+τ0.5	5.83	62.58	91.85	4.84	56.89	$4.14_{\pm 4.05}$
AMUN+τ1	1.38	40.70	97.30	0.92	39.23	$\mathbf{2.47_{\pm 1.61}}$	AMUN+τ1	3.77	60.96	93.83	3.01	55.90	$\mathbf{3.48_{\pm 2.93}}$

Table 18: All baselines on **CIFAR-100 with** 10% **class-wise** forgetting. IMPRINT achieves the smallest average gap to retraining at both buffer sizes.

	Buffer size = 500						Buffer size = 2000					
Method	**FA**	**RA**	**MIA**	**FT**	**RT**	**Avg. Gap**	**FA**	**RA**	**MIA**	**FT**	**RT**	**Avg. Gap**
Retrain	0.00	41.09	100.00	0.00	37.95	—	0.00	60.72	100.00	0.00	54.11	—
RL	0.61	35.46	99.64	0.54	36.56	$1.72_{\pm 0.65}$	0.62	51.27	99.62	0.48	49.47	$3.12_{\pm 1.69}$
SalUn	1.08	37.17	99.38	1.14	37.08	$1.58_{\pm 0.56}$	1.44	54.31	99.29	1.20	50.69	$2.64_{\pm 0.98}$
SCRUB	0.00	4.95	100.00	0.00	6.44	$13.53_{\pm 0.23}$	0.00	3.27	100.00	0.00	4.58	$21.40_{\pm 0.31}$
AMUN	11.80	29.41	83.69	9.86	29.27	$11.67_{\pm 5.34}$	8.92	46.61	87.48	7.38	43.72	$10.66_{\pm 4.03}$
NPO	0.01	34.78	99.98	0.00	35.41	$1.77_{\pm 0.54}$	0.00	46.45	100.00	0.00	46.37	$4.40_{\pm 2.18}$
GA	1.91	0.91	81.44	1.98	0.87	$19.94_{\pm 9.73}$	0.14	1.57	80.13	0.30	1.62	$26.39_{\pm 9.37}$
BE	0.11	7.56	99.36	0.00	7.12	$13.02_{\pm 0.28}$	0.05	26.76	98.40	0.04	25.07	$12.94_{\pm 1.40}$
BS	4.75	24.05	91.30	3.86	23.40	$9.78_{\pm 2.49}$	7.40	43.13	86.90	5.48	39.66	$11.61_{\pm 2.34}$
IU	10.51	33.59	87.64	6.72	33.81	$8.25_{\pm 5.17}$	11.77	55.66	85.75	8.16	52.83	$8.11_{\pm 3.71}$
IMU	16.76	15.63	86.56	16.40	16.69	$18.67_{\pm 5.56}$	32.12	50.67	68.32	27.56	47.18	$21.67_{\pm 3.68}$
IMPRINT	0.17	38.56	99.68	0.00	37.13	$\mathbf{0.87_{\pm 0.59}}$	0.03	54.00	99.94	0.00	49.50	$\mathbf{2.41_{\pm 2.95}}$

C.3 Robustness to the CIL Training Budget

We test whether the effectiveness of IMPRINT depends on the optimization budget used to train the underlying CIL model. We vary the number of CIL epochs per task over 25, 50, and 100, corresponding to $0.5\times$, $1\times$, and $2\times$ the default 50-epoch budget. Except for the budget size, we keep all other experimental settings the same (class-wise forgetting ratio of 30% on CIFAR-100 with DER++, buffer size of 2000).

As shown in Table 22, across the full $0.5\times$–$2\times$ range, IMPRINT improves RA and RT at every training budget while maintaining near-zero FA and FT. Thus, its retention advantage is not specific to the default CIL optimization schedule.

We next examine whether the heterogeneous imprint structure that motivates IMPRINT persists under the same changes in CIL training budget. All statistics in Table 23 is computed on the CIL model before unlearning. We group the first three tasks as early, the middle four as middle, and the final three as recent, and additionally report the P90–P10 spread over the complete sample-level margin distribution.

Table 23 shows that the heterogeneous-imprint pattern remains clear at every budget. Recent-task samples have the largest average margins, while the sample-level distribution remains broad. Although the absolute margins and the ordering of the early and middle groups change with the optimization budget, the strong recent-versus-past separation and broad sample-level spread persist. Together with Table 22, this shows that both the motivating heterogeneity and the benefit of IMPRINT are robust to substantial changes in CIL training budget.

Table 19: **Plug-and-play compatibility** of IMPRINT on **CIFAR-100 with 10% class-wise** forgetting.

	Buffer size = 500						Buffer size = 2000					
Method	**FA**	**RA**	**MIA**	**FT**	**RT**	**Avg. Gap**	**FA**	**RA**	**MIA**	**FT**	**RT**	**Avg. Gap**
Retrain	0.00	41.09	100.00	0.00	37.95	—	0.00	60.72	100.00	0.00	54.11	—
RL	0.61	35.46	99.64	0.54	36.56	$1.72_{\pm 0.65}$	0.62	51.27	99.62	0.48	49.47	$3.12_{\pm 1.69}$
RL+IMPRINT	0.17	38.56	99.68	0.00	37.13	$\mathbf{0.87_{\pm 0.59}}$	0.03	54.00	99.94	0.00	49.50	$\mathbf{2.41_{\pm 2.95}}$
SalUn	1.08	37.17	99.38	1.14	37.08	$1.58_{\pm 0.56}$	1.44	54.31	99.29	1.20	50.69	$2.64_{\pm 0.98}$
SalUn+IMPRINT	0.20	38.68	99.59	0.04	37.17	$\mathbf{0.96_{\pm 0.77}}$	0.08	57.75	99.88	0.00	52.96	$\mathbf{1.03_{\pm 0.61}}$
AMUN	11.80	29.41	83.69	9.86	29.27	$11.67_{\pm 5.34}$	8.92	46.61	87.48	7.38	43.72	$10.66_{\pm 4.03}$
AMUN+IMPRINT	1.55	39.92	97.58	0.64	38.28	$\mathbf{1.57_{\pm 0.86}}$	0.92	53.95	98.50	0.24	49.94	$\mathbf{2.87_{\pm 2.32}}$
BS	4.75	24.05	91.30	3.86	23.40	$9.78_{\pm 2.49}$	7.40	43.13	86.90	5.48	39.66	$11.61_{\pm 2.34}$
BS+IMPRINT	0.46	37.33	99.29	0.06	35.40	$\mathbf{1.51_{\pm 1.11}}$	0.32	59.42	99.37	0.02	54.26	$\mathbf{1.21_{\pm 0.68}}$

Table 20: All baselines on **CIFAR-100 with** 20% **class-wise** forgetting.

	Buffer size = 500						Buffer size = 2000					
Method	**FA**	**RA**	**MIA**	**FT**	**RT**	**Avg. Gap**	**FA**	**RA**	**MIA**	**FT**	**RT**	**Avg. Gap**
Retrain	0.00	45.28	100.00	0.00	41.94	—	0.00	62.77	100.00	0.00	55.89	—
RL	0.39	31.41	99.91	0.67	35.22	$4.35_{\pm 1.01}$	0.72	50.51	99.93	1.97	53.49	$3.48_{\pm 0.59}$
SalUn	0.97	34.37	99.73	1.66	36.41	$3.87_{\pm 0.82}$	1.86	54.93	99.67	3.78	53.83	$3.17_{\pm 0.75}$
SCRUB	0.02	5.37	99.92	0.01	11.46	$14.10_{\pm 0.81}$	0.03	10.60	99.90	0.01	22.20	$17.20_{\pm 2.36}$
AMUN	8.80	29.22	86.91	8.98	28.38	$12.10_{\pm 1.70}$	13.30	49.42	86.26	11.58	46.42	$12.29_{\pm 2.81}$
NPO	0.02	29.33	99.95	0.00	31.58	$5.28_{\pm 1.29}$	0.05	53.76	99.86	0.17	53.15	$2.42_{\pm 0.36}$
GA	0.03	1.21	99.97	0.07	1.18	$16.99_{\pm 0.66}$	0.00	1.26	80.00	0.00	1.25	$27.23_{\pm 9.22}$
BE	0.13	5.94	99.10	0.02	5.44	$15.38_{\pm 0.64}$	0.09	20.84	98.17	0.05	19.36	$16.08_{\pm 0.51}$
BS	4.42	22.88	91.39	4.44	22.30	$11.90_{\pm 1.37}$	5.50	38.72	89.60	4.55	36.41	$12.80_{\pm 1.35}$
IU	7.83	38.49	90.90	6.10	37.35	$6.88_{\pm 1.13}$	19.24	60.87	79.81	14.08	55.63	$11.21_{\pm 1.20}$
IMU	5.88	19.46	96.20	6.27	20.14	$12.72_{\pm 1.45}$	19.89	53.95	89.87	18.69	49.70	$12.74_{\pm 3.72}$
IMPRINT	0.69	39.23	98.75	0.38	37.68	$\mathbf{2.53_{\pm 1.14}}$	0.73	61.34	98.67	0.44	56.12	$\mathbf{1.03_{\pm 1.49}}$

D Compatibility and Generalization

D.1 Compatibility with other unlearning methods in ImageNet-1k

The main paper reports ImageNet-1k results for buffer size 20000. (Table 5 of the main paper). Even in this bigger dataset, IMPRINT continues to reduce the gap to retraining when applied to base unlearning methods (Table 24).

D.2 Compatibility with iCaRL

Table 25 compares existing unlearning baselines with IMPRINT under the iCaRL setting, while Table 26 shows IMPRINT's compatibility with existing unlearning baselines. Under this different CIL setup, IMPRINT continues to outperform existing baselines and reduces the gap to retraining when integrated with the base unlearning methods.

D.3 Generality Beyond Class-Incremental Learning

We evaluate whether IMPRINT generalizes beyond Class-IL while keeping the unlearning procedure and hyperparameters fixed. Task-IL uses 10-task CIFAR-100 with ResNet-18 and task identity available at inference, while Domain-IL uses 10-domain Permuted-MNIST with a shared classifier and no domain identity. Both settings are evaluated with DER++ using a 30% forgetting ratio, with buffer sizes of 2000 for Task-IL and 500 for Domain-IL. Table 27 reports results averaged over three forgetting configurations.

As shown in Table 27, IMPRINT improves both forgetting and retention for RL and SalUn in Task-IL. In Domain-IL, FA and FT are already zero, so the gains are primarily on the retain side, where

Table 21: **Plug-and-play compatibility** of IMPRINT on **CIFAR-100 with** 20% **class-wise** forgetting.

	Buffer size = 500						Buffer size = 2000					
Method	**FA**	**RA**	**MIA**	**FT**	**RT**	**Avg. Gap**	**FA**	**RA**	**MIA**	**FT**	**RT**	**Avg. Gap**
Retrain	0.00	45.28	100.00	0.00	41.94	—	0.00	62.77	100.00	0.00	55.89	—
RL	0.39	31.41	99.91	0.67	35.22	$4.35_{\pm 1.01}$	0.72	50.51	99.93	1.97	53.49	$3.48_{\pm 0.59}$
RL+IMPRINT	0.69	39.23	98.75	0.38	37.68	$\mathbf{2.53_{\pm 1.14}}$	0.73	61.34	98.67	0.44	56.12	$\mathbf{1.03_{\pm 1.49}}$
SalUn	0.97	34.37	99.73	1.66	36.41	$3.87_{\pm 0.82}$	1.86	54.93	99.67	3.78	53.83	$3.17_{\pm 0.75}$
SalUn+IMPRINT	0.56	39.71	98.84	0.34	38.02	$\mathbf{2.31_{\pm 1.09}}$	0.62	61.45	98.94	0.34	56.18	$\mathbf{0.99_{\pm 1.36}}$
AMUN	8.80	29.22	86.91	8.98	28.38	$12.10_{\pm 1.70}$	13.30	49.42	86.26	11.58	46.42	$12.29_{\pm 2.81}$
AMUN+IMPRINT	2.92	40.32	95.53	2.41	38.50	$\mathbf{3.65_{\pm 1.21}}$	3.66	61.50	95.25	2.34	56.28	$\mathbf{2.67_{\pm 2.70}}$
BS	4.42	22.88	91.39	4.44	22.30	$11.90_{\pm 1.37}$	5.50	38.72	89.60	4.55	36.41	$12.80_{\pm 1.35}$
BS+IMPRINT	0.92	36.77	98.31	0.47	35.00	$\mathbf{3.71_{\pm 1.30}}$	1.06	59.72	98.06	0.50	54.09	$\mathbf{1.70_{\pm 2.46}}$

Table 22: Sensitivity to the **CIL training budget** on **CIFAR-100 with** 30% **class-wise forgetting**.

Epochs/task	**Method**	**FA↓**	**RA↑**	**MIA↑**	**FT↓**	**RT↑**	**Avg. Gap↓**
25 (0.5×)	Retrain	0.00	61.47	99.98	0.00	55.00	—
	RL	0.09	43.02	99.57	1.08	49.06	$5.20_{\pm 0.84}$
	RL+IMPRINT	0.04	60.67	99.74	0.06	55.01	$\mathbf{0.39_{\pm 0.29}}$
50 (1×)	Retrain	0.00	64.11	100.00	0.00	57.18	—
	RL	0.14	45.73	99.53	1.01	52.33	$4.97_{\pm 0.87}$
	RL+IMPRINT	0.13	61.37	99.56	0.06	56.37	$\mathbf{0.90_{\pm 0.56}}$
100 (2×)	Retrain	0.00	63.41	99.73	0.00	56.89	—
	RL	0.14	46.79	99.62	0.60	51.11	$4.69_{\pm 0.85}$
	RL+IMPRINT	0.08	62.10	99.60	0.04	56.86	$\mathbf{0.83_{\pm 0.88}}$

IMPRINT reduces unnecessary degradation in RA and RT once sufficient forgetting has occurred. These results indicate that the proposed sample-wise weighting is not specific to the Class-IL setting.

D.4 Generalization to Text Classification

We additionally evaluate IMPRINT on 20 Newsgroups in a 10-task class-incremental setting, with two classes introduced per task. We use Experience Replay with a lightweight bag-of-embeddings MLP classifier consisting of a 128-dimensional EmbeddingBag, one hidden layer with 128 units, and a 20-class output layer, providing a representation space substantially different from the vision datasets and backbones considered in the main experiments. The replay-buffer size is fixed to 500, and the source model is trained for 10 epochs on each task. We consider class-wise forgetting with a 30% forgetting ratio, corresponding to six of the 20 classes, and evaluate five matched forgetting configurations shared across methods. Each unlearning method is run for 10 epochs. The results are reported in Table 28.

Table 28 shows that adding IMPRINT improves forgetting quality for both base methods, as reflected by lower FA and FT, higher MIA, and a smaller Avg. Gap. Although RA and RT change little in this experiment, the overall gap to retraining is reduced, supporting the intended model-agnostic and plug-and-play formulation beyond image classification.

E Continual Unlearning Dynamics

E.1 Subsequent Learning and Forgetting Relapse

We next study whether forgotten information resurfaces when continual learning resumes after unlearning. First, the model is trained through Task 8, and 25 classes from Tasks 0–8 are then unlearned. Before learning Task 9, all replay-buffer records and associated metadata for the forgotten classes, including stored DER++ logits, are removed. A matched retrained model follows the same

Table 23: **Imprint heterogeneity across CIL training budgets.** Mean sample-level logit margins are reported by task age before unlearning. P90–P10 summarizes the spread over all samples.

Epochs/task	Early	Middle	Recent	P90–P10
25 (0.5×)	$1.01_{\pm 0.24}$	$0.83_{\pm 0.39}$	$\mathbf{2.66_{\pm 1.35}}$	$\mathbf{11.29_{\pm 0.39}}$
50 (1×)	$1.45_{\pm 0.36}$	$1.70_{\pm 0.57}$	$\mathbf{2.79_{\pm 0.60}}$	$\mathbf{12.29_{\pm 0.53}}$
100 (2×)	$1.69_{\pm 0.44}$	$1.59_{\pm 0.41}$	$\mathbf{3.38_{\pm 0.95}}$	$\mathbf{12.95_{\pm 0.53}}$

Table 24: **Plug-and-play compatibility** of IMPRINT on **ImageNet-1k**. IMPRINT further reduces the gap to retraining when plugged into each base unlearning method.

Method	FA	RA	MIA	FT	RT	Avg. Gap
Retrain	0.00	39.43	100.00	0.00	43.18	—
RL	0.00	18.20	99.84	0.00	21.05	$9.09_{\pm 2.79}$
RL+IMPRINT	0.00	34.18	99.99	0.00	37.94	$\mathbf{2.50_{\pm 0.42}}$
SaLUn	0.00	17.80	100.00	0.00	21.28	$9.11_{\pm 2.05}$
SaLUn+IMPRINT	0.00	34.25	100.00	0.00	38.46	$\mathbf{2.38_{\pm 0.16}}$
AMUN	14.58	37.89	74.73	16.72	42.43	$13.25_{\pm 2.02}$
AMUN+IMPRINT	4.09	44.07	89.99	4.99	48.63	$\mathbf{5.84_{\pm 0.73}}$
BS	6.64	23.44	86.63	8.04	29.02	$11.64_{\pm 0.33}$
BS+IMPRINT	0.53	35.88	98.99	0.65	41.20	$\mathbf{1.57_{\pm 0.46}}$

stream without the forgotten classes. Table 29 evaluates the models both immediately after unlearning and after subsequent learning of Task 9.

As shown in Table 29, forgotten information partially resurfaces after learning Task 9, as reflected by increased FA and FT and decreased MIA for both RL and SalUn. In both cases, IMPRINT reduces the extent of this relapse and remains closer to the retrained model. These results indicate that IMPRINT improves robustness to one-step forgetting relapse.

E.2 Task Age and Order of Unlearning Requests

We finally examine how task age and the order of multiple unlearning requests affect unlearning outcomes. Both experiments were conducted on CIFAR-100 with DER++ and buffer size 2,000.

Task age at a common checkpoint. We first compare five complete classes from Task 0 with five complete classes from Task 9 at the final CIL checkpoint, using five matched configurations and separate retrained references. The Task-9 targets have substantially larger pre-unlearning logit margins than the Task-0 targets (4.20 versus 0.98), indicating that Task-9 targets are more strongly encoded. Table 30 compares the corresponding unlearning results.

As shown in Table 30, both RL and SalUn has a larger gap for the older Task-0 targets than for Task-9 targets. IMPRINT reduces the gap in both cases, with the larger reduction occurring in Task 0. These results indicate that logit margin captures sample-level variation in forgetting difficulty that is not fully characterized by task age alone.

Order of multiple unlearning requests. We finally pair an early-task class with a recent-task class using five matched class pairs and compare early→recent, recent→early, and joint unlearning. Each unlearning request is run for 10 epochs; consequently, the early→recent and recent→early sequences require 20 epochs in total, whereas joint unlearning requires 10 epochs. All replay-buffer samples belonging to the forget classes are removed after each unlearning request. Table 31 reports the resulting order sensitivity.

As shown in Table 31, both RL and SalUn exhibit a noticeable order effect. Removing the recent class first and the early class second produces a larger Avg. Gap than the reverse order. This indicates that the first deletion changes the model state in a way that affects the second unlearning operation. IMPRINT substantially reduces this sensitivity. Both sequential orders reach zero FA and FT and

Table 25: All baselines on **iCaRL CIFAR-100 with** 30% **class-wise** forgetting, averaged over five forgetting configurations.

	Buffer size = 2000					
Method	**FA**	**RA**	**MIA**	**FT**	**RT**	**Avg. Gap**
Retrain	0.00	52.86	100.00	0.00	50.57	—
RL	0.14	15.26	99.97	0.26	21.09	$13.50_{\pm 6.37}$
SalUn	0.34	14.35	99.90	0.41	20.25	$13.94_{\pm 5.29}$
SCRUB	0.00	5.78	100.00	0.00	6.24	$18.29_{\pm 1.40}$
AMUN	18.10	41.28	77.14	18.75	40.25	$16.32_{\pm 0.59}$
NPO	0.00	47.20	100.00	0.00	50.14	$1.52_{\pm 0.23}$
GA	0.00	1.33	100.00	0.00	1.40	$20.14_{\pm 0.79}$
BE	0.02	8.40	99.43	0.06	21.41	$14.86_{\pm 1.13}$
BS	5.07	24.49	90.38	5.99	31.24	$13.68_{\pm 1.18}$
IMU	10.38	21.40	92.09	9.85	21.24	$17.79_{\pm 4.29}$
IU	12.09	51.30	85.54	12.82	52.97	$8.75_{\pm 1.73}$
IMPRINT	0.02	54.51	99.93	0.06	54.10	$\mathbf{1.07_{\pm 0.48}}$

Table 26: **Plug-and-play compatibility** of `IMPRINT` on iCaRL CIFAR-100. `IMPRINT` further reduces the gap to retraining when plugged into each base unlearning method.

Method	**FA**	**RA**	**MIA**	**FT**	**RT**	**Avg. Gap**
Retrain	0.00	52.86	100.00	0.00	50.57	—
RL	0.14	15.26	99.97	0.26	21.09	$13.50_{\pm 6.37}$
RL+`IMPRINT`	0.02	54.51	99.93	0.06	54.10	$\mathbf{1.07_{\pm 0.48}}$
SalUn	0.34	14.35	99.90	0.41	20.25	$13.94_{\pm 5.29}$
SalUn+`IMPRINT`	0.01	53.04	99.90	0.05	53.73	$\mathbf{0.85_{\pm 0.42}}$
AMUN	18.10	41.28	77.14	18.75	40.25	$16.32_{\pm 0.59}$
AMUN+`IMPRINT`	3.56	54.98	94.31	4.07	54.13	$\mathbf{3.80_{\pm 2.37}}$
BS	5.07	24.49	90.38	5.99	31.24	$13.68_{\pm 1.18}$
BS+`IMPRINT`	0.02	52.75	99.89	0.05	53.09	$\mathbf{0.72_{\pm 0.31}}$

have much smaller gaps than RL and SalUn because the sample-wise weights are recomputed after the first request. Joint unlearning yields the smallest gap for both methods in this experiment, suggesting that processing multiple requests together can reduce the accumulation of sequential perturbations.

Taken together, Tables 30–31 show that unlearning behavior depends on both the current imprint of the target data and the order in which deletion requests are processed. By adapting update intensity to the target's current imprint, `IMPRINT` remains effective across targets from different task ages and different request orders.

Table 27: **Generalization beyond Class-IL.** IMPRINT reduces the gap to retraining for both RL and SalUn in Task-IL and Domain-IL.

(a) Task-IL

Method	FA↓	RA↑	MIA↑	FT↓	RT↑	Avg. Gap↓
Retrain	0.22	91.94	100.00	0.17	85.81	—
RL	1.01	89.77	98.88	0.58	83.85	$1.29_{\pm 0.32}$
RL+IMPRINT	0.03	90.43	99.91	0.23	84.91	$\mathbf{0.55_{\pm 0.29}}$
SalUn	1.38	90.02	98.56	0.72	84.34	$1.31_{\pm 0.30}$
SalUn+IMPRINT	0.05	90.49	99.86	0.21	84.90	$\mathbf{0.55_{\pm 0.27}}$

(b) Domain-IL

Method	FA↓	RA↑	MIA↑	FT↓	RT↑	Avg. Gap↓
Retrain	0.00	95.42	100.00	0.00	95.33	—
RL	0.00	87.01	100.00	0.00	87.28	$3.29_{\pm 0.64}$
RL+IMPRINT	0.00	89.55	100.00	0.00	89.80	$\mathbf{2.28_{\pm 0.42}}$
SalUn	0.00	87.55	99.99	0.00	87.82	$3.08_{\pm 0.34}$
SalUn+IMPRINT	0.00	90.28	100.00	0.00	90.51	$\mathbf{1.99_{\pm 0.35}}$

Table 28: **Results on 20 Newsgroups.** IMPRINT reduces the gap to retraining when applied to both RL and SalUn.

Method	FA↓	RA↑	MIA↑	FT↓	RT↑	Avg. Gap↓
Retrain	0.00	69.15	100.00	0.00	37.02	—
RL	5.88	60.85	88.50	1.29	33.70	$6.10_{\pm 0.68}$
RL+IMPRINT	2.14	60.37	92.20	0.34	33.32	$\mathbf{4.55_{\pm 0.59}}$
SalUn	5.86	60.84	88.56	1.29	33.70	$6.08_{\pm 0.68}$
SalUn+IMPRINT	2.14	60.37	92.19	0.34	33.31	$\mathbf{4.56_{\pm 0.59}}$

Table 29: **Learn–unlearn–learn experiment.** Performance is evaluated (a) immediately after unlearning and (b) after subsequently learning Task 9. Avg. Gap is computed with respect to the matched retrained model at each evaluation stage.

(a) Immediately after unlearning

Method	FA↓	RA↑	MIA↑	FT↓	RT↑	Avg. Gap↓
Retrain	0.00	57.01	100.00	0.00	50.27	—
RL	0.35	47.22	99.96	1.57	47.94	$2.82_{\pm 0.51}$
RL+IMPRINT	0.02	56.40	99.88	0.01	50.62	$\mathbf{0.27_{\pm 0.08}}$
SalUn	1.08	46.61	99.83	3.52	48.77	$3.33_{\pm 0.27}$
SalUn+IMPRINT	0.03	56.64	99.89	0.01	51.20	$\mathbf{0.29_{\pm 0.12}}$

(b) After learning Task 9

Method	FA↓	RA↑	MIA↑	FT↓	RT↑	Avg. Gap↓
Retrain	0.00	62.83	100.00	0.00	55.37	—
RL	6.99	56.99	90.90	5.92	51.48	$6.35_{\pm 1.15}$
RL+IMPRINT	2.87	57.81	95.29	2.31	51.98	$\mathbf{3.66_{\pm 0.08}}$
SalUn	7.53	56.82	92.10	7.25	51.40	$6.53_{\pm 1.14}$
SalUn+IMPRINT	4.61	57.51	94.01	4.36	51.75	$\mathbf{4.78_{\pm 0.97}}$

Table 30: **Effect of target task age at the final CIL checkpoint.** Metrics are means over five matched configurations.

Target	Method	FA↓	RA↑	MIA↑	FT↓	RT↑	Avg. Gap↓
Task 0	Retrain	0.00	60.26	100.00	0.00	53.46	—
	RL	0.75	57.95	99.89	0.48	52.57	$1.00_{\pm 0.47}$
	RL+IMPRINT	0.02	59.18	99.99	0.00	53.73	$\mathbf{0.46_{\pm 0.23}}$
	SalUn	1.21	58.81	99.94	0.80	53.20	$0.92_{\pm 0.26}$
	SalUn+IMPRINT	0.05	58.82	99.96	0.00	53.33	$\mathbf{0.59_{\pm 0.48}}$
Task 9	Retrain	0.00	58.36	100.00	0.00	51.99	—
	RL	0.04	56.14	100.00	0.00	51.44	$0.64_{\pm 0.37}$
	RL+IMPRINT	0.00	56.52	100.00	0.00	51.78	$\mathbf{0.51_{\pm 0.31}}$
	SalUn	0.20	56.66	100.00	0.16	51.44	$0.62_{\pm 0.29}$
	SalUn+IMPRINT	0.01	57.26	100.00	0.00	52.25	$\mathbf{0.33_{\pm 0.28}}$

Table 31: **Effect of the order of multiple unlearning requests.** Results are averaged over five matched configurations.

Order	Method	FA↓	RA↑	MIA↑	FT↓	RT↑	Avg. Gap↓
—	Retrain	0.00	59.87	100.00	0.00	53.12	—
Task 0→9	RL	3.47	56.01	99.73	3.30	50.45	$2.71_{\pm 0.93}$
	RL+IMPRINT	0.00	55.97	100.00	0.00	50.45	$\mathbf{1.31_{\pm 0.66}}$
	SalUn	3.40	55.77	99.42	3.40	50.03	$2.91_{\pm 1.51}$
	SalUn+IMPRINT	0.00	55.81	100.00	0.00	50.21	$\mathbf{1.39_{\pm 0.62}}$
Task 9→0	RL	7.44	56.39	98.67	6.80	51.09	$4.22_{\pm 2.21}$
	RL+IMPRINT	0.00	56.53	100.00	0.00	51.10	$\mathbf{1.07_{\pm 0.32}}$
	SalUn	8.53	56.22	98.29	7.50	50.82	$4.74_{\pm 3.81}$
	SalUn+IMPRINT	0.00	57.26	100.00	0.00	51.47	$\mathbf{0.85_{\pm 0.32}}$
Joint	RL	1.02	57.33	100.00	0.70	51.49	$1.18_{\pm 0.53}$
	RL+IMPRINT	0.00	58.70	100.00	0.00	52.87	$\mathbf{0.33_{\pm 0.22}}$
	SalUn	3.47	56.69	99.91	3.00	51.07	$2.36_{\pm 1.63}$
	SalUn+IMPRINT	0.02	58.68	100.00	0.00	52.85	$\mathbf{0.34_{\pm 0.20}}$