LABEL-LEVEL CERTIFIED UNTRAINING FOR PRIVACY GUARANTEES IN DEEP NEURAL NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Untraining removes the influence of specific training data (the *forget set*) from a model without retraining from scratch, a practical necessity for privacy. Yet no method guarantees privacy for deep neural networks (DNNs), where untraining matters most due to excessive costs: heuristic methods offer only empirical evidence, and certified methods, which draw on differential privacy (DP), yield vacuous guarantees as model size grows. We introduce δ_{tv}-*certification*, the first label-DP-style formalization of untraining, which certifies the model's confidence scores on the forget set rather than its parameters, keeping guarantees meaningful at scale. We then propose *certified untraining via prediction calibration (CUP)*, which calibrates these scores toward a retraining-free reference, the original model's confidences on unseen data. CUP transfers across settings, even without access to the remaining training data, by leveraging this retraining-free reference. Extensive ablations show that CUP can yield privacy guarantees without compromising untraining performance and model generalization. On ViT-Large, a scale unexplored by certified work, CUP guarantees that any test distinguishing untrained from retrained forget-set confidences errs at least 45.65% of the time (random guessing: 50%).

1 INTRODUCTION

Untraining aims to eliminate the memorization of specific training data (*forget set*), while preserving the generalization capacity of the model (Triantafillou et al., 2026). In other words, it offers privacy without the computational expense of retraining from scratch, which is particularly critical in DNNs. However, privacy cannot be guaranteed at the scale of modern DNNs, such as Vision Transformers.

Gap 1. Prior work certifies untraining via DP-style guarantees over model distributions, i.e., it answers: *if an adversary draws randomly from a distribution of untrained and retrained models, how accurately it can distinguish the source distribution?* Most certified methods assume (strong) convexity, excluding application to DNNs, which typically have non-convex training objectives (Guo et al., 2020; Sekhari et al., 2021; Chien et al., 2024a; Allouah et al., 2025). More recent work (Zhang et al., 2024; Qiao et al., 2025; Mu & Klabjan, 2025; Sepahvand et al., 2025) that relaxes these assumptions also cannot certify untraining for DNNs. The reported privacy guarantees lose their practical meaning. This is because privacy guarantees are coupled with model size. They untrain by directly perturbing model parameters, introducing a specific amount of noise. The noise required for privacy grows with the model size and significantly degrades model generalization for DNNs (Neel et al., 2021; Sekhari et al., 2021; Koloskova et al., 2025; Allouah et al., 2025). Ultimately, prior work compromises privacy to avoid catastrophic forgetting, i.e., completely destroying model generalization.

Gap 2. This noise level, in certified methods, is typically predefined and for DNNs it depends on intractable constants that derive from assumptions. For instance, these constants include upper-bounds on gradients, Hessian norms, third-order loss derivatives, variance rate of Langevin diffusion (Zhang et al., 2024; Mu & Klabjan, 2025; Basaran et al., 2025; Chien et al., 2024b;a). Hence, they are treated as tunable knobs. Hyperparameter sensitivity is critical for real-world scenarios, as tuning employs retrained models as reference (e.g.,Spartalis et al., 2025; Foster et al., 2024a), which are unavailable in practice.

We discuss related work in detail in Section A.

We introduce a novel untraining formalization, which we call δ_{tv}-certification. It is the first that draws from label DP (Ghazi et al., 2021) and focuses on a narrower attack surface to restore the theoretical

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.

meaning of privacy guarantees. Specifically, δ_{tv}-certification answers: *if an adversary draws from confidence score distributions on the forget set induced by untraining and retraining, how accurately it can distinguish its source distribution?*

Moreover, we introduce a proof-of-concept method, which we call *certified untraining via prediction calibration (CUP)*, that leverages a retraining-free distribution for reference. This is also a confidence-score distribution, induced by the original model on unseen (during training) data, which are drawn from the population distribution of forget data. This reference is cheap but informative and enables: (i) tractability of the untraining state at each step, (ii) dynamic updates and smooth convergence to the optimal untraining state, and (iii) mitigates hyperparameter sensitivity for adaptation to various settings. Our main **contributions** are:

1. **δ_{tv}-certification**: To address Gap 1, we introduce a novel untraining formalization that yields DP-style guarantees for DNNs, i.e., the regime where untraining offers the greatest practical gains. Moreover, δ_{tv}-certification uncovers factors of reference-guided untraining that remain undocumented in prior heuristic approaches.

2. **CUP algorithm**: To address Gap 2, we introduce an algorithm that ensures δ_{tv}-certification and facilitates applicability in real-world scenarios. Leveraging a cheap yet information-rich retraining-free reference distribution, CUP adapts to various settings with a fixed set of hyperparameters, even in data-free (DF) settings where access to the remaining training data (*retain set*) is restricted.

3. **Promising results**: By narrowing the scope of certification, we obtain the first meaningful DP-style guarantees for untraining large DNNs (Table 1). Extensive evaluation and ablations show that CUP certifies untraining while achieving strong results against 14 baselines in balancing privacy and generalization, and, most importantly, that it adapts to various settings by leveraging the retraining-free reference, even without access to the retain set.

Table 1: *Minimum error of any test distinguishing untrained from retrained models or their confidence scores*. Closer to random guessing (50%) is better. Proposition B.4 and Lemma B.1 interpret DP guarantees as test-error bounds. Prior work cannot guarantee privacy for DNNs. δ_{tv}-certification yields meaningful guarantees that can also scale to ViT models, unexplored in prior certified work. CUP-DF has no access to the retain set.

Method	ResNet-18 CIFAR-10	ViT-Base CIFAR-100	ViT-Base TinyImageNet	ViT-Large TinyImageNet
CUP-DF (ours)	41.49%	43.00%	46.40%	46.52%
CUP (ours)	42.92%	43.14%	46.80%	45.65%
Zhang et al. (2024)	$10^{-43}\approx 0\%$	–	–	–
Qiao et al. (2025)	$10^{-52}\approx 0\%$	–	–	–
Mu & Klabjan (2025)	$10^{-10^3}\approx 0\%$	–	–	–
Sepahvand et al. (2025)	$10^{-10^7}\approx 0\%$	–	–	–

2 Background

Preliminaries. Consider an image classification task. Let $s(\mathbf{x}, y)$ be a data sample, where $\mathbf{x} \in \mathcal{X}$ is the feature vector of an image, and $y \in \mathcal{Y} := [K]$ is the true label of $\mathbf{x}$. A DNN is a function $h_\theta \in \mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$ parameterized by $\theta \in \Theta$, such that $h_\theta : \mathcal{X} \to \mathcal{Y}$. Let $\mathcal{A}$ be a training algorithm, and $D_{\mathrm{tr}} = \{s_i\}_{n_{\mathrm{tr}}}$ a training set. The untraining algorithm $\mathcal{U}$ aims to remove the influence of the forget data $D_{\mathrm{for}} \subset D_{\mathrm{tr}}$ from the original model $\theta_0 := \mathcal{A}(D_{\mathrm{tr}})$. The untrained model $\theta_{\mathrm{u}} := \mathcal{U}(D_{\mathrm{for}}, D_{\mathrm{tr}}, \theta_0)$ must protect privacy to the same extent as the retrained model $\theta^\star := \mathcal{A}(D_{\mathrm{tr}} \setminus D_{\mathrm{for}})$.

In heuristic approaches, this objective translates into comparable empirical metrics, such as accuracy on the forget set D_{for} and success rate of membership inference attacks (MIAs). In prior certified work, the untraining objective translates into minimizing the divergence between distributions of untrained and retrained models. Hence, they introduce DP-style guarantees over model distributions.

Definition 1 ((ε, δ)-Certification in Model Space). Let $\mathcal{A}$ and $\mathcal{U}$ be randomized algorithms that generate model distributions. Untraining is (ε, δ)-certified if for any measurable subset $\mathcal{S} \subseteq \mathcal{H}$,

$$\begin{aligned} \mathbb{P}_{\theta_{\mathrm{u}} \sim \mathcal{U}(D_{\mathrm{for}}, D_{\mathrm{tr}}, \theta_0)} \left(\theta_{\mathrm{u}} \in \mathcal{S}\right) &\leq e^{\varepsilon} \cdot \mathbb{P}_{\theta^\star \sim \mathcal{A}(D_{\mathrm{tr}} \setminus D_{\mathrm{for}})} \left(\theta^\star \in \mathcal{S}\right) + \delta \quad \text{and} \\ \mathbb{P}_{\theta^\star \sim \mathcal{A}(D_{\mathrm{tr}} \setminus D_{\mathrm{for}})} \left(\theta^\star \in \mathcal{S}\right) &\leq e^{\varepsilon} \cdot \mathbb{P}_{\theta_{\mathrm{u}} \sim \mathcal{U}(D_{\mathrm{for}}, D_{\mathrm{tr}}, \theta_0)} \left(\theta_{\mathrm{u}} \in \mathcal{S}\right) + \delta \end{aligned} \tag{1}$$

Smaller $\varepsilon \in \mathbb{R}_{\geq 0}$ and $\delta \in [0, 1]$ mean better privacy. The *privacy budget* ε requires that, for any $\mathcal{S}$, the probability that a model in $\mathcal{S}$ is produced by untraining is at most e^{ε} times the probability that it is produced by retraining, and vice-versa. This upper-bound, e^{ε}, holds with probability at least $1-\delta$.

Preliminaries for δ_{tv}-certification and CUP. Let $\widehat{\mathcal{D}}$ be a uniform distribution over a dataset $D = \{s_i\}_n$, whose samples $s \in D$ are drawn i.i.d. according to a population (data-generating) distribution $\mathcal{D}$. For a sample s and model h_θ, the confidence score is the prediction probability of the true label $c(s; \theta) := p(y|x; \theta)$. Let $\mathcal{C}(\mathcal{D}; \theta)$ denote a population confidence-score distribution and $\mathcal{C}(\widehat{\mathcal{D}}; \theta)$ its empirical counterpart. According to Definition 1, we introduce an untraining formalization for label-level certification in Definition 2.

Definition 2 ((ε, δ)-Certification in Prediction Space). Let $B_1, \ldots, B_m$ be bins partitioning $[0, 1]$. Untraining is (ε, δ)-certified if for all $j \in [m]$

$$\begin{aligned} \mathbb{P}_{c \sim \mathcal{C}(\widehat{\mathcal{D}}_{\text{for}}; \theta_{\text{u}})} (c \in B_j) &\leq e^{\varepsilon} \cdot \mathbb{P}_{c \sim \mathcal{C}(\mathcal{D}_{\text{for}}; \theta^\star)} (c \in B_j) + \delta \quad \text{and} \\ \mathbb{P}_{c \sim \mathcal{C}(\mathcal{D}_{\text{for}}; \theta^\star)} (c \in B_j) &\leq e^{\varepsilon} \cdot \mathbb{P}_{c \sim \mathcal{C}(\widehat{\mathcal{D}}_{\text{for}}; \theta_{\text{u}})} (c \in B_j) + \delta, \end{aligned} \tag{2}$$

where $\mathcal{C}(\widehat{\mathcal{D}}_{\text{for}}; \theta_{\text{u}})$ is the distribution of the untrained model on $\widehat{\mathcal{D}}_{\text{for}}$, and $\mathcal{C}(\mathcal{D}_{\text{for}}; \theta^\star)$ that of the hypothetical retrained model, an unknown population-level distribution as in Kumar et al. (2019).

Quantizing the prediction space $[0, 1]$ into a finite number of m (fixed-width) bins places both distributions on a common measurable space. Empirical confidence scores are singular with respect to the continuous population distribution, so f-divergences saturate regardless of how close the two distributions are. Binning induces a finite σ-algebra making f-divergences informative (see Definition B.1). For further preliminaries we defer the reader to Section B.

3 δ_{tv}-Certification for Label-Level Untraining Guarantees

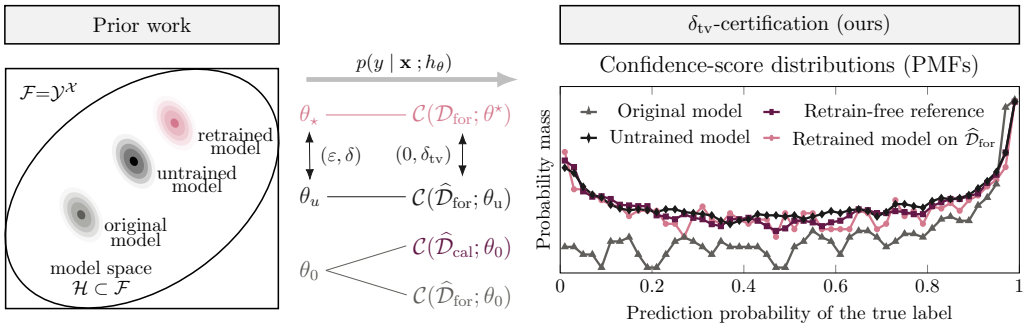


Figure 1: *Prior work (left)* formalizes untraining with (ε, δ)-DP-style guarantees in the high-dimensional model space, and perturbs model parameters directly. *δ_{tv}-certification (right)* formalizes untraining with $(0, \delta_{\text{tv}})$-DP-style guarantees on the one-dimensional prediction space. CUP perturbs the soft-labels and fine-tunes the original model θ_0 with these noisy targets, until convergence to a retraining-free reference distribution, $\mathcal{C}(\widehat{\mathcal{D}}_{\text{cal}}; \theta_0)$, induced on unseen (during training) data. Plot data: CUP, ResNet-18, CIFAR-10.

To certify untraining, prior work draws on *central DP* mechanisms that directly perturb the model parameters, such as (Abadi et al., 2016). We focus instead on a family of DP mechanisms for privacy-preserving predictions (Dwork & Feldman, 2018; Papernot et al., 2018), which remains unexplored for untraining. Although their guarantees are weaker, these mechanisms are known to preserve generalization better, which addresses Gap 1. In particular, we draw from Ghazi et al. (2021), who train the model with noisy labels (soft-probabilities) and show that *prior information* about the probability mass on true labels facilitates balancing privacy and generalization.

Prior information in untraining. DP-training is the standard approach for *proactive* privacy, i.e., it produces a model (or a distribution of models) given some training data. In contrast, during untraining the converged model θ_0 is available, enriching the prior information at hand. Specifically, we leverage a confidence-score distribution induced by θ_0 on unseen (during training) data D_{cal},

which we call *calibration distribution* $\mathcal{C}(\widehat{\mathcal{D}}_{\text{cal}}; \theta_0)$. This can serve as a retraining-free reference, since model predictions on D_{cal} are not driven by memorization and reflect model generalization. By default, this set D_{cal} is a held-out dataset split, e.g., the validation set; when it contains synthetic data we will say it explicitly.

Untraining as uncertainty calibration. We employ $\mathcal{C}(\widehat{\mathcal{D}}_{\text{cal}}; \theta_0)$ as a retraining-free reference distribution. Moreover, we use it to dynamically perturb the classification targets of the forget set at each untraining step. In other words, we recast untraining as uncertainty calibration that increases the predictive entropy of the original model on the forget data up to a certain reference, $\mathcal{C}(\widehat{\mathcal{D}}_{\text{cal}}; \theta_0)$. In practice, this means that we derive, rather than predefine, the noise level required for privacy, i.e., the ε, δ parameters, to address Gap 2.

δ_{tv}-certification. To capture this shift in the DP-style untraining formalization, we reformulate the (ε, δ) guarantee in Definition 2 into an equivalent δ_{tv} certificate that ensures $(\varepsilon{=}0, \delta_{\text{tv}})$ DP for every $\delta_{\text{tv}} \geq d_{\text{tv}}(\mathcal{C}(\widehat{\mathcal{D}}_{\text{for}}; \theta_{\text{u}}), \mathcal{C}(\mathcal{D}_{\text{for}}; \theta^\star))$, where d_{tv} denotes the total variation distance (see Proposition B.4 from Ghazi & Issa (2024)). Then, we specify the upper-bound δ_{tv} by applying the triangle inequality recursively to the TV distance:

$$
\begin{aligned}
d_{\text{tv}}\left(\mathcal{C}(\widehat{\mathcal{D}}_{\text{for}}; \theta_{\text{u}}), \mathcal{C}(\mathcal{D}_{\text{for}}; \theta^\star)\right) \leq\ & \underbrace{d_{\text{tv}}\left(\mathcal{C}(\widehat{\mathcal{D}}_{\text{for}}; \theta_{\text{u}}), \mathcal{C}(\widehat{\mathcal{D}}_{\text{cal}}; \theta_0)\right)}_{\delta_{\text{emp}}} + \underbrace{d_{\text{tv}}\left(\mathcal{C}(\widehat{\mathcal{D}}_{\text{cal}}; \theta_0), \mathcal{C}(\mathcal{D}_{\text{cal}}; \theta_0)\right)}_{\delta_{\text{etp}}} \\
& + \underbrace{d_{\text{tv}}\left(\mathcal{C}(\mathcal{D}_{\text{cal}}; \theta_0), \mathcal{C}(\mathcal{D}_{\text{for}}; \theta_0)\right)}_{\delta_{\text{dom}}} + \underbrace{d_{\text{tv}}\left(\mathcal{C}(\mathcal{D}_{\text{for}}; \theta_0), \mathcal{C}(\mathcal{D}_{\text{for}}; \theta^\star)\right)}_{\delta_{\text{con}}}
\end{aligned}
\tag{3}
$$

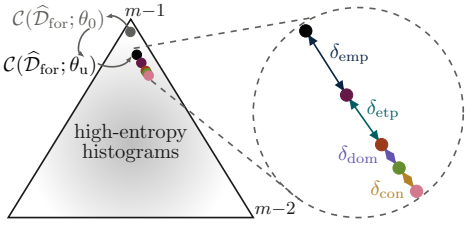


Figure 2: After quantization into m bins, each confidence distribution in Eq. (3) is a point in the probability simplex Δ^{m-1}. Untraining, viewed as uncertainty calibration, moves mass from higher-confidence bins to lower-confidence ones (top left). The δ_{tv}-certificate is the sum of TV distances shown in the magnifying circle (right).

This total variation (TV) certificate, $\delta_{\text{tv}} := \delta_{\text{emp}} + \delta_{\text{etp}} + \delta_{\text{dom}} + \delta_{\text{con}}$, decomposes into four attributable terms that capture distinct facets of the guarantee. δ_{emp} captures the empirical distance between the untrained confidence distribution and the retraining-free reference $\mathcal{C}(\widehat{\mathcal{D}}_{\text{cal}}; \theta_0)$; δ_{etp} the empirical-to-population distance introduced by a finite-sample calibration set and a finite number of bins; δ_{dom} the domain-shift distance between the forget and calibration data population distributions; and δ_{con} the concept-drift distance between the original and the retrained models. Fig. 2 visualizes untraining as uncertainty calibration and the δ_{tv}-certificate. We compute and discuss each term in Eq. (3) in Section 5.

4 Certified Untraining via Prediction Calibration (CUP)

In this section, we present CUP, which updates θ_0 to θ_{u}. As a proof-of-concept algorithm for δ_{tv}-certification, CUP avoids sources of error that δ_{tv} cannot capture. Moreover, CUP targets adaptability to real-world scenarios. Hence, to address Gap 2, we avoid introducing new hyperparameters, we eliminate the batch-size by full-batch updates and trade the epoch count for a patience count of $\mathtt{p}$ epochs. Most importantly, CUP adapts through two mechanisms: dynamic model updates driven by the retraining-free reference $\mathcal{C}(\widehat{\mathcal{D}}_{\text{cal}}; \theta_0)$, and an intrinsic stopping criterion that halts untraining once δ_{emp} stops decreasing for $\mathtt{p}$ epochs.

CUP aims to align $\mathcal{C}(\widehat{\mathcal{D}}_{\text{for}}; \theta_{\text{u}})$ with the retraining-free reference $\mathcal{C}(\widehat{\mathcal{D}}_{\text{cal}}; \theta_0)$, which is fixed throughout untraining. However, the distance between them is defined over set-level distributions, whereas backpropagation requires per-instance targets. We address this by fine-tuning the model on perturbed targets, inspired by Ghazi et al. (2021). However, unlike label DP, the perturbation level is determined by $\mathcal{C}(\widehat{\mathcal{D}}_{\text{cal}}; \theta_0)$ rather than randomly sampled.

Starting from $\theta_{\mathrm{u}}^{t=0} = \theta_0$, we update θ_{u} at each epoch t according to the current untraining state. At epoch t, we compute the empirical cumulative distribution function (ECDF) F_{u}^t of the untrained confidence scores, which provides state-awareness, and project each score $c_{\mathrm{u},i}^t \sim \mathcal{C}(\widehat{\mathcal{D}}_{\mathrm{for}}; \theta_{\mathrm{u}})$ onto $\mathcal{C}(\widehat{\mathcal{D}}_{\mathrm{cal}}; \theta_0)$ via the quantile mapping $F_{\mathrm{cal}}^{-1} \circ F_{\mathrm{u}}^t$, where F_{cal} is the ECDF of the reference distribution. This yields the classification targets $\mathbf{z}_i = (z_{i,1}, \ldots, z_{i,K})$:

$$z_{i,j}^{\mathrm{t}} \;=\; \begin{cases} F_{\mathrm{cal}}^{-1}\left(F_{\mathrm{u}}^{\mathrm{t}}\left(c_{u,i}^{\mathrm{t}}\right)\right) & j = y_i \\ p(j \mid \mathbf{x}_i; \theta_u^{\mathrm{t}})\, \dfrac{1 - z_{i,y_i}^{\mathrm{t}}}{1 - p(y_i \mid \mathbf{x}_i; \theta_u^{\mathrm{t}})}, & j \neq y_i \end{cases} \tag{4}$$

The resulting per-instance losses for each $(\mathbf{x}_i, y_i) \in D_{\mathrm{for}}$ drive dynamic updates and convergence to the retraining-free reference, see Fig. 3.

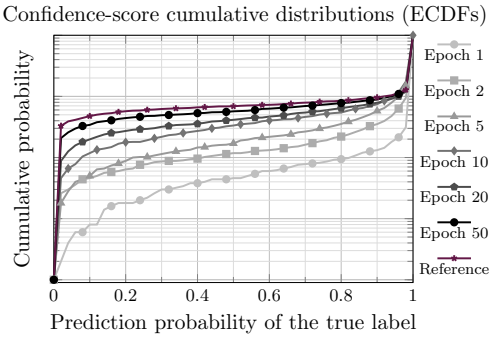


Figure 3: ECDFs for Fig. 1: F_{u}^t converges from $t{=}0{\to}50$ toward the retraining-free reference F_{cal}.

Algorithm 1: CUP-DF. η is the learning rate. CUP additionally performs one update on retain data D_{r} per t.

Require: θ_0, η, D_{for}, D_{cal}, p
$F_{\mathrm{cal}} \leftarrow \mathrm{ECDF}\big(\{c(s;\theta_0)\}_{s \in D_{\mathrm{cal}}}\big)$ ▷ reference
$t \leftarrow 0$, $\theta_{\mathrm{u}}^{t=0} \leftarrow \theta_0$
repeat
  $F_{\mathrm{u}}^t \leftarrow \mathrm{ECDF}\big(\{c_{\mathrm{u},i}^t\}_{i \in D_{\mathrm{for}}}\big)$ ▷ state-awareness
  $z_{i,y_i}^t \leftarrow F_{\mathrm{cal}}^{-1}\big(F_{\mathrm{u}}^t(c_{\mathrm{u},i}^t)\big)$ ▷ quantile mapping
  rescale $z_{i,j}^t$ for $j \neq y_i$ ▷ Eq. (5)
  $\theta_{\mathrm{u}}^{t+1} \leftarrow \theta_{\mathrm{u}}^t - \eta \nabla \mathcal{L}_{\mathrm{f}}(\mathbf{x}_i, \mathbf{z}_i^t)$
  $c_{\mathrm{u,i}}^{t+1} \leftarrow p(y_i | \mathbf{x}_i; \theta_{\mathrm{u}}^{t+1})$
  $t \leftarrow t + 1$
until no decrease in δ_{emp} for p epochs
return $\theta_{\mathrm{u}}^{t^\star}$, $t^\star = \arg\min_t \delta_{\mathrm{emp}}^t$

Errors. Quantile mapping $F_{\mathrm{cal}}^{-1} \circ F_{\mathrm{u}}$ is monotone and is the optimal transport map in one dimension: it preserves the order of confidence scores and displaces probability mass minimally. Quantization error, which depends on the number of quantiles, is captured by δ_{emp}. Binning error, which depends on the number of bins m and the calibration-set size n_{cal}, is captured by the next term, δ_{etp}. Under binning, δ_{emp} admits a closed form.

Adaptability. Univariate confidence distributions make the untraining state informative and cheap to track at every epoch, unlike prior reference-guided methods (Chen et al., 2025; Mirlach et al., 2026). The state is the full true-label confidence distribution, which is richer and more informative than a scalar accuracy gap (Spartalis et al., 2025). $\mathcal{C}(\widehat{\mathcal{D}}_{\mathrm{cal}}; \theta_0)$ is computed once, each check costs one forward pass over the forget set, and δ_{emp} is a closed-form TV distance over m bins, with no density estimation on the K-class simplex. CUP uses this state twice: to reassign targets $\mathbf{z}$ from the current F_{u}^t, and to compute δ_{emp} for the intrinsic stopping criterion.

5 Discussion and Results

5.1 δ_{emp}: Empirical Distance

The empirical distance δ_{emp} is the only term of δ_{tv} that untraining controls. Under binning, it has the closed form $\delta_{\mathrm{emp}} = \frac{1}{2}\sum_{j=1}^{m} |\mathcal{C}_{\mathrm{u}}(B_j) - \mathcal{C}_{\mathrm{cal}}(B_j)|$. Fig. 4 shows that CUP reduces δ_{emp} by 64–82% from its initial value $\delta_{\mathrm{emp}}^{\mathrm{init}}$ across all model–dataset settings, and CUP-DF, without access to retain data, by 59–80%.

5.2 δ_{etp}: Empirical-to-Population Distance

The empirical-to-population distance δ_{etp} captures the sample bias of the retraining-free reference, which arises from its finite sample size n_{cal} and the finite number of bins m.

By definition, $\{s_i\}_{n_{\text{cal}}}$ are drawn i.i.d. from the population distribution $\mathcal{D}_{\text{cal}}$, and after training, h_{θ_0} is a fixed function $s \mapsto c(s;\theta_0)$. Hence, $\{c_i\}_{n_{\text{cal}}}$ are drawn i.i.d. from $\mathcal{C}(\mathcal{D}_{\text{cal}};\theta_0)$.

Proposition 1. By Definition B.2, $\mathcal{C}(\widehat{\mathcal{D}}_{\text{cal}};\theta_0)$ is the type of $\{c_i\}_{i=1}^{n_{\text{cal}}}$, which are drawn i.i.d. from $\mathcal{C}(\mathcal{D}_{\text{cal}};\theta_0)$. Then, by the standard bound from the method of types (Theorem B.1) and Pinsker's inequality (Proposition B.1), for a binned space of m bins:

$$\delta_{\text{etp}} \;\le\; \sqrt{\frac{1}{2}\left(\frac{m\ln(n_{\text{cal}}+1)}{n_{\text{cal}}}+\beta\right)} \qquad (5)$$

with probability at least $1-e^{-n_{\text{cal}}\beta}$. Here, $\beta>0$ is not a free parameter: it is chosen as a function of n_{cal} so that this probability is non-trivial. We set $\beta=\ln(10^5)/n_{\text{cal}}$, so the bound holds with probability at least $1-10^{-5}$. The proof is in Section C.

Therefore, for a given sample size n_{cal}, the number of bins m is not a free parameter. A higher m increases resolution and may thus facilitate uncertainty calibration, but it inflates δ_{etp}, as shown in Table 2. We set m=15 for CIFAR, following Guo et al. (2017), and m=20 for TinyImageNet, where more calibration data are available (Table 3).

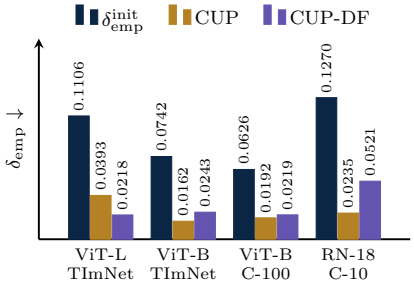


Figure 4: The initial empirical distance $\delta_{\text{emp}}^{\text{init}}$ reduces to δ_{emp} after untraining with CUP and CUP-DF.

Table 2: δ_{etp} grows with bin count m, for a fixed n_{cal}. (RN-18, C-10)

m	$\delta_{\text{etp}} \downarrow$	$\delta_{\text{emp}} \downarrow$	AG (pp) $\downarrow$
15	0.118	$0.024_{\pm 0.003}$	2.53
20	0.135	$0.021_{\pm 0.005}$	2.52
50	0.209	$0.022_{\pm 0.002}$	2.51
original ($\theta_{\text{u}}^{t=0}\equiv\theta_0$)		$0.127_{\pm 0.005}$	7.15

Table 3: More calibration data yield tighter δ_{etp}. $n_{\text{cal}} \uparrow \Rightarrow \delta_{\text{etp}} \downarrow$.

Dataset	n_{cal}	m	β	δ_{etp}
C-10/100	5,000	15	0.0023	0.1180
TImNet	50,000	20	0.00023	0.0477

In all our experiments, $\delta_{\text{etp}} > \delta_{\text{emp}}$ (see Section D.1), which highlights the significant role of sample bias, which is undocumented in heuristic reference-guided methods. This applies to any method whose sufficient statistic is the whole sample, including prior heuristic methods (Spartalis et al., 2025; Chen et al., 2025; Mirlach et al., 2026). The more samples, the more information, and the more reliable the retraining-free reference. δ_{tv}-certification quantifies this sample bias through δ_{etp}.

5.3 δ_{dom}: Domain-Shift Distance

$\delta_{\text{dom}} := d_{\text{tv}}(\mathcal{C}(\mathcal{D}_{\text{cal}};\theta_0), \mathcal{C}(\mathcal{D}_{\text{for}};\theta_0))$ captures the divergence between the confidence-score distributions that the original model induces on the population (data-generating) distributions of the calibration and the forget data. In δ_{dom}, confidence-score distributions are evaluated at the population level, and the population distribution $\mathcal{C}(\mathcal{D}_{\text{for}};\theta_0)$ is not affected by the memorization of finite forget samples in D_{for}, because they have zero probability under the continuous population distribution $\mathcal{D}_{\text{for}}$. Hence, if the population distributions of the forget and calibration data are equal in distribution, $\mathcal{D}_{\text{cal}} \overset{\text{d}}{=} \mathcal{D}_{\text{for}}$, then $\delta_{\text{dom}} = 0$ by definition.

Table 4: With synthetic calibration data, untraining yields a δ_{emp} on par with the in-domain setting, but the domain shift remains evident: before untraining, $\delta_{\text{emp}}^{\text{init}}$ is 36% larger. Details in Table 22.

Calibration set	$n_{\text{cal}}^{\text{val}}$	$\delta_{\text{emp}}^{\text{init}} \to \delta_{\text{emp}} \downarrow$
held-out (cifar)	5,000	$0.127_{\pm 0.005} \to 0.024_{\pm 0.003}$
synthetic (cifake)	5,000	$0.173_{\pm 0.001} \to 0.020_{\pm 0.002}$
synthetic (cifake)	50,000	$0.171_{\pm 0.000} \to 0.020_{\pm 0.000}$

The setting of $\mathcal{D}_{\text{cal}} \overset{\text{d}}{=} \mathcal{D}_{\text{for}}$ is standard in heuristic reference-guided untraining (Spartalis et al., 2025; Mirlach et al., 2026; Chen et al., 2025). It arises from defining the calibration set as a held-out split of a single data pool (including the training data), which was collected and annotated under the same protocol.

In real-world scenarios, a held-out split may not be available. Spartalis et al. (2025) use a calibration set composed exclusively of synthetic data from the CIFAKE (Bird & Lotfi, 2024) dataset, and demonstrate comparable untraining performance under real and synthetic calibration data. We follow the same approach and reach the same conclusion, shown in Table 4. This is consistent with the findings of Minderer et al. (2021) regarding uncertainty calibration being robust to such domain shifts, especially as model size grows.

However, the empirical evaluation of the untraining result may obfuscate the difference that arises from sampling from a shifted population distribution. As shown in Table 4, the domain shift is evident in $\delta^{\text{init}}_{\text{emp}} = d_{\text{tv}}(\mathcal{C}(\widehat{\mathcal{D}}_{\text{cal}};\theta_0),\mathcal{C}(\widehat{\mathcal{D}}_{\text{for}};\theta_0))$. δ_{dom} captures this domain shift of the retraining-free reference, which remains undocumented in prior untraining work.

In such cases, δ_{dom} can be upper-bounded in expectation by Jensen's inequality (Proposition B.2): $\delta_{\text{dom}} \leq \mathbb{E}\big[d_{\text{tv}}\left(\mathcal{C}(\widehat{\mathcal{D}}^{\text{syn}}_{\text{cal}},\theta_0),\mathcal{C}(\widehat{\mathcal{D}}^{\text{hout}}_{\text{cal}},\theta_0)\right)\big]$, where $\widehat{\mathcal{D}}^{\text{syn}}_{\text{cal}}$ is an empirical distribution over the synthetic calibration data, and $\widehat{\mathcal{D}}^{\text{hout}}_{\text{cal}}$ the distribution over a held-out set that is sampled according to $\mathcal{D}_{\text{for}} \overset{\text{d}}{=} \mathcal{D}_{\text{cal}}$. We cannot use $\widehat{\mathcal{D}}_{\text{for}}$, because at the empirical level it is memorized by θ_0, which alters the meaning of δ_{dom}. We estimate this expectation by the TV distance between the two empirical distributions. In Section D.4, we provide an ablation analysis with synthetic calibration sets of different composition and size, where the empirical domain shift remains flat. For $5{,}000$ synthetic samples, $\delta_{\text{dom}}{=}0.062$ is added to δ_{tv} and every test errs with probability at least 40.0% compared with 42.9% in the in-domain setting. With $50{,}000$ CIFAKE samples, $\delta_{\text{dom}}{=}0.061$ remains almost the same, but every test errs with probability at least 43.9%, because δ_{etp} decreases significantly with ten times more data. That is, abundant synthetic data can more than compensate for the domain shift by decreasing the empirical-to-population distance δ_{etp}.

5.4 δ_{con}: Concept-Drift Distance

By definition, the concept-drift distance $\delta_{\text{con}}{:=}d_{\text{tv}}\left(\mathcal{C}(\mathcal{D}_{\text{for}};\theta_0),\mathcal{C}(\mathcal{D}_{\text{for}};\theta^\star)\right)$ vanishes if and only if the original and the retrained models induce the same population confidence-score distribution on the forget data population. This means that the model does not lose generalization if the forget set is excluded from the training set (the confidence-score distribution is not driven by memorization at the population level, as in Section 5.3). Following the same line of reasoning as in Section 5.3, we can quantify δ_{con} in controlled experiments using a held-out set sampled from the population distribution $\mathcal{D}_{\text{for}} \overset{\text{d}}{=} \mathcal{D}_{\text{cal}}$, as shown in Table 5.

Table 5: *Concept-drift distance.* δ_{con} on the CIFAR-10 test set and the minimum error of any test when δ_{con} is added to $\delta_{\text{tv}}{=}0.1415$ (CIFAR-10, ResNet-18, CUP, 10% of the training data; Table 8). For every forget-set size, δ_{con} lowers the minimum test error by less than one percentage point, i.e., it is negligible.

Forget set	δ_{con}	Minimum test error	
(% of D_{tr})		δ_{tv}	$\delta_{\text{tv}}{+}\delta_{\text{con}}$
1%	0.0107	42.9%	42.4%
10%	0.0136	42.9%	42.2%
20%	0.0142	42.9%	42.2%

However, beyond controlled experiments, the retrained model is not available. Thus we need to adopt a ubiquitous assumption in prior certified work (e.g., Zhang et al. (2024); Sepahvand et al. (2025); Allouah et al. (2025)) that can render δ_{con} negligible:

Assumption 1. The forget data D_{for} are drawn i.i.d. from the training data D_{tr} and form a small fraction of it, $|D_{\text{for}}|{\ll}|D_{\text{tr}}|$.

By substituting the model parameters with the algorithms that generated them, the dependence of model generalization on the total number of the forget data becomes clearer: $\delta_{\text{con}}{:=}d_{\text{tv}}\big(\mathcal{C}\big(\mathcal{D}_{\text{for}},\mathcal{A}(D_{\text{tr}})\big),\mathcal{C}\big(\mathcal{D}_{\text{for}},\mathcal{A}(D_{\text{tr}}\backslash D_{\text{for}})\big)\big)$. Moreover, the values of δ_{con} in Table 5 support this notion.

Notably, Guo & Cao (2026) show that all Newton-step approaches require Assumption 1, while heuristic methods that build on a Newton-step approximation (Foster et al., 2024a;b) report competitive results in tasks (class unlearning) where this assumption is violated. However, these methods "adapt" to these tasks by hyperparameter tuning (Foster et al., 2024a). This contrast may indicate both the greater verifiability of certified approaches and the obfuscation in evaluation that hyperparameter tuning can induce.

In our experiments, the impact of δ_{con} on the guarantee is small: adding it to δ_{tv} lowers the minimum error of any test by less than one percentage point for every forget-set size (Table 5). Hence, the guarantee is determined by the calibration data and the untraining itself rather than by the unknown retrained model, and Assumption 1 costs little in practice.

5.5 Overview of δ_{tv}-Certification Results

The δ_{tv}-certification in the experiments where a held-out data split was used as the retraining-free reference is the sum of the empirical and empirical-to-population distances $\delta_{\mathrm{emp}} + \delta_{\mathrm{etp}}$. When the held-out split is replaced by synthetic data, the domain-shift distance δ_{dom} is added to the δ_{tv}. The concept-drift distance is negligible in our controlled experiments (Table 5), and in real-world scenarios it is rendered negligible by Assumption 1. The resulting guarantees per setting are reported in Table 1. See Section D.1.

5.6 Empirical Evaluation

Here, we compare CUP with heuristic methods and demonstrate that it adapts to realistic settings: the data-free setting (CUP-DF), where the retain set is not accessible, and ablations over forget-set size (Fig. 5), forget-set composition (Fig. 6) and learning rate (Fig. 7). In all experiments, CUP(-DF) is more efficient than retraining; for further details we defer the reader to Section E.6.

Following Spartalis et al. (2025), for baselines we adopt hyperparameters proposed in the original papers or implementations. For CUP(-DF): the learning rate η equals that of original training (10^{-4} for ResNet-18 and ViT-Base, 10^{-5} for ViT-Large), the number of quantiles is 20, and the stopping patience is $\mathrm{p} = 15$ epochs for ResNet-18 and $\mathrm{p}{=}3$ for ViT.

We untrain ResNet-18 (He et al., 2016) and ViT-Base/Large (Dosovitskiy et al., 2020) models; to the best of our knowledge, no prior certified work evaluates at the scale of the latter. We use the CIFAR-10/100 (Krizhevsky, 2009) and TinyImageNet (Le & Yang, 2015) datasets; Zhang et al. (2024) identify the latter as challenging. Following Spartalis et al. (2025); Fan et al. (2024), we assess the privacy–generalization balance via the average gap (AG): the mean absolute difference between the untrained and retrained models in forget, retain, and test accuracy and in MIA success rate, averaged over three independent runs. For further and detailed results see Section E.

CUP(-DF) certify untraining while remaining competitive with heuristic approaches. Certified baselines are scarce at that scale. CNS, which reports the tightest model-level certification, cannot reach large models due to Hessian approximation, $O(d)$. Qiao et al. (2025) introduces higher memory complexity, $O(n_{\mathrm{for}}d)$, scaling with forget-set size, and evaluates on far smaller forget sets (0.1% of D_{tr}). Mu & Klabjan (2025) require model states stored during original training, a requirement no heuristic baselines have. Sepahvand et al. (2025) focus on one-sample untraining; otherwise the privacy budget becomes vast (Table 7), and the code is not public.

Table 6: *Privacy-Utility balance via AG in pp.* CNS and CUP(-DF) are the only certified methods. Means across three independent experiments. Lower is better. Untraining 10% of training data with standard and data-free baselines.

Method	RN-18 C-10	ViT-B C-100	ViT-B TImNet	ViT-L TImNet
Fine-tuning	6.89	3.47	1.72	2.37
RandLabel (Graves et al., 2021)	3.10	1.64	1.69	1.82
NegGr+ (Kurmanji et al., 2023)	6.89	6.91	5.91	90.24
SCRUB (Kurmanji et al., 2023)	7.08	2.01	2.56	90.24
BadT (Chundawat et al., 2023)	4.40	4.52	9.20	14.18
UNSIR (Tarun et al., 2023)	4.05	1.71	1.71	2.08
SalUn (Fan et al., 2024)	3.17	1.67	1.17	76.83
CNS (Zhang et al., 2024)	7.28	OOM due to Hessian matrix		
LoTUS (Spartalis et al., 2025)	4.11	1.58	1.96	2.68
CUP (ours)	**2.53**	**1.12**	**0.89**	**1.74**
BoundSh (Chen et al., 2023)	7.09	2.07	2.56	70.26
SSD (Foster et al., 2024a)	7.06	1.99	2.55	3.48
LFSSD (Foster et al., 2024b)	7.16	2.08	2.55	3.48
JiT (Foster et al., 2025)	82.00	71.72	76.34	53.62
SALSA (Makroo et al., 2025)	5.43	19.27	14.71	32.36
CUP-DF (ours)	**4.18**	**1.69**	**0.91**	**2.44**

Hyperparameters play a critical role in empirical performance. JiT explicitly reports high hyperparameter sensitivity; SCRUB depends on a large number of tunable constants, such as separate epoch counts for forgetting and retaining, separate steps per epoch, and additional regularization terms in the total loss. *CUP(-DF)'s competitive results follow from its intrinsic adaptability, as shown in ablation analyses.*

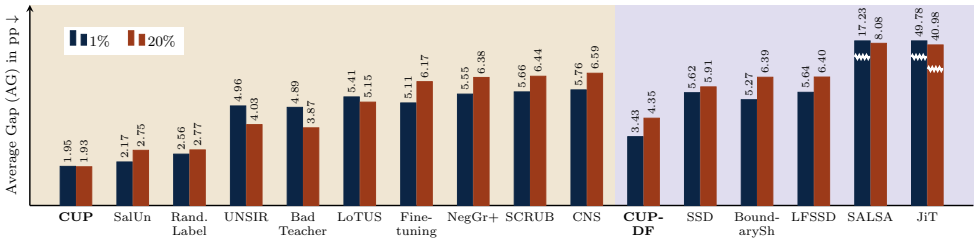


Figure 5: *Ablation on forget-set sizes (% of the training set).* CUP (and CUP-DF) preserve best scores in balancing privacy and utility on the ResNet-18 & CIFAR-10. Same for ViT-Base & TinyImageNet (see Fig. 10).

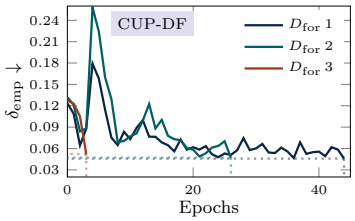
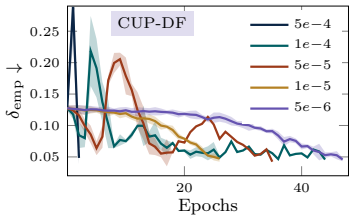


Figure 6: *Adaptation across random forget sets.* CUP-DF reaches comparable δ_{emp}; so does CUP (Fig. 11).



η	AG ↓	δ_{emp} ↓
original	7.15	$0.1275_{\pm 0.0043}$
$5e{-4}$	4.54	$0.0484_{\pm 0.0103}$
$1e{-4}$	4.18	$0.0521_{\pm 0.0097}$
$5e{-5}$	4.21	$0.0489_{\pm 0.0121}$
$1e{-5}$	4.18	$0.0485_{\pm 0.0040}$
$5e{-6}$	4.41	$0.0459_{\pm 0.0028}$

Figure 7: *Adaptation across learning rates (η).* Lower η yield smoother convergence, but after more epochs. Untraining 10% of CIFAR-10 from ResNet-18.

6 Conclusion

Untraining is most needed in DNNs especially as model size grows because of the increasing cost of retraining from scratch, yet there prior certified work cannot guarantee privacy. We introduced δ_{tv}-certification, which moves DP-style guarantees for untraining from the model space to the prediction space. Through Le Cam's lemma, the guarantee has a direct meaning: no test that observes the confidence scores on the forget set can tell whether they come from the untrained or the retrained model much better than random guessing. Our proof-of-concept method, CUP, attains this guarantee by calibrating the untrained model against a retraining-free reference, and adapts to various settings, including restricted access to the retain set.

Limitations. Our guarantees are weaker than those of prior certified work, because they cover a narrower attack surface, the prediction space rather than the model space. We accept this trade-off for DNNs, where model-level guarantees lose their theoretical meaning. In this regime, prediction-level guarantees fill the gap between model-level certification and purely empirical evaluation. To mitigate this limitation, the certificate can be complemented with a white-box MIA that also exploits the model's parameters, see Section E.4.

AI use statement

We used generative AI tools only as writing and engineering assistants. They helped us improve the aesthetics of the figures and tables, review the writing for typos, grammar and syntax, suggest clearer phrasing, and review the consistency of our mathematical terminology. We also used them to review our code and to transfer the experimental results from our MLflow logs into the LaTeX tables. Generative AI did not contribute to the research ideas, the theoretical results and their proofs, the design of the method or the experiments, or the interpretation of the results. The authors reviewed every AI-assisted edit and checked the transferred results against the original logs, and take full responsibility for the content of this paper.

Reproducibility statement

We share the code to reproduce all results as anonymous supplementary material, and we will release it publicly upon publication. CUP is described in Section 4, and the hyperparameters of CUP, CUP-DF and all baselines, the models and the datasets are given in Section 5.6. All datasets are publicly

available, and we report the mean and standard deviation over three independent runs with different forget and retain sets (seeds 12, 13 and 3407; Section E). The assumptions and the complete proofs of our theoretical results are in Sections B, C and 5. We run the ViT-Large experiments on an NVIDIA RTX A6000 48GB GPU with Automatic Mixed Precision, and all other experiments on an NVIDIA RTX 4080 16GB GPU with an Intel i7-12700K CPU and 32GB RAM, using Python 3.11 and CUDA 12.0.

REFERENCES

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pp. 308–318, 2016.

Sk Miraj Ahmed, Umit Yigit Basaran, Dripta S Raychaudhuri, Arindam Dutta, Rohit Kundu, Fahim Faisal Niloy, Basak Guler, and Amit K Roy-Chowdhury. Towards source-free machine unlearning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 4948–4957, 2025.

Youssef Allouah, Joshua Kazdan, Rachid Guerraoui, and Sanmi Koyejo. The utility and complexity of in-and out-of-distribution machine unlearning. In *International Conference on Learning Representations*, volume 2025, pp. 2713–2738, 2025.

Umit Yigit Basaran, Sk Miraj Ahmed, Amit Roy-Chowdhury, and Basak Guler. A certified unlearning approach without access to source data. In *Forty-second International Conference on Machine Learning*, 2025.

Jordan J Bird and Ahmad Lotfi. Cifake: Image classification and explainable identification of ai-generated synthetic images. *IEEE Access*, 12:15642–15650, 2024.

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.

Kongyang Chen, Dongping Zhang, Bing Mi, Yao Huang, and Zhipeng Li. Fast yet versatile machine unlearning for deep neural networks. *Neural Networks*, 190:107648, 2025. ISSN 0893-6080.

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.

Eli Chien, Haoyu Wang, Ziang Chen, and Pan Li. Certified machine unlearning via noisy stochastic gradient descent. *Advances in Neural Information Processing Systems*, 37:38852–38887, 2024a.

Eli Chien, Haoyu Wang, Ziang Chen, and Pan Li. Langevin unlearning: A new perspective of noisy gradient descent for machine unlearning. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 79666–79703. Curran Associates, Inc., 2024b. doi: 10.52202/079017-2530.

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.

Thomas Cover and Joy Thomas. *Elements of information theory*. Wiley-Interscience, 2nd edition, 2006.

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.

John Duchi. Statistics and information theory. *Book Draft*, 2025.

Cynthia Dwork and Vitaly Feldman. Privacy-preserving prediction. In *Conference On Learning Theory*, pp. 1693–1702. PMLR, 2018.

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.

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, 2024a.

Jack Foster, Stefan Schoepf, and Alexandra Brintrup. Loss-free machine unlearning. In *The Second Tiny Papers Track at ICLR 2024*, 2024b.

Jack Foster, Kyle Fogarty, Stefan Schoepf, Zack Dugue, Cengiz Oztireli, and Alexandra Brintrup. An information theoretic approach to machine unlearning. *Transactions on Machine Learning Research*, 2025.

Badih Ghazi, Noah Golowich, Ravi Kumar, Pasin Manurangsi, and Chiyuan Zhang. Deep learning with label differential privacy. *Advances in neural information processing systems*, 34:27131–27145, 2021.

Elena Ghazi and Ibrahim Issa. Total variation meets differential privacy. *IEEE Journal on Selected Areas in Information Theory*, 5:207–220, 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.

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

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

Chuan Guo, Tom Goldstein, Awni Hannun, and Laurens Van Der Maaten. Certified data removal from machine learning models. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3832–3842. PMLR, 13–18 Jul 2020.

Jinduo Guo and Yinzhi Cao. A robust certified machine unlearning method under distribution shift. *arXiv preprint arXiv:2601.06967*, 2026.

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.

Anastasia Koloskova, Youssef Allouah, Animesh Jha, Rachid Guerraoui, and Sanmi Koyejo. Certified unlearning for neural networks. In *Forty-second International Conference on Machine Learning*, 2025.

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

Ananya Kumar, Percy S Liang, and Tengyu Ma. Verified uncertainty calibration. *Advances in neural information processing systems*, 32, 2019.

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

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

Owais Makroo, Atif Hassan, and Swanand Khare. Salsa: A secure, adaptive and label-agnostic scalable algorithm for machine unlearning. In *The 41st Conference on Uncertainty in Artificial Intelligence*, 2025.

Matthias Minderer, Josip Djolonga, Rob Romijnders, Frances Hubis, Xiaohua Zhai, Neil Houlsby, Dustin Tran, and Mario Lucic. Revisiting the calibration of modern neural networks. *Advances in neural information processing systems*, 34:15682–15694, 2021.

Jonas Mirlach, Sonia Laguna, and Julia E Vogt. Reference-guided machine unlearning. *arXiv preprint arXiv:2603.11210*, 2026.

Siqiao Mu and Diego Klabjan. Rewind-to-delete: Certified machine unlearning for nonconvex functions. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Milad Nasr, Reza Shokri, and Amir Houmansadr. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In *2019 IEEE symposium on security and privacy (SP)*, pp. 739–753. IEEE, 2019.

Joseph P Near, David Darais, Naomi Lefkovitz, Gary S Howarth, et al. *Guidelines for evaluating differential privacy guarantees*. US Department of Commerce, National Institute of Standards and Technology, 2025.

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

Nicolas Papernot, Shuang Song, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Ulfar Erlingsson. Scalable private learning with PATE. In *International Conference on Learning Representations*, 2018.

Xinbao Qiao, Meng Zhang, Ming Tang, and Ermin Wei. Hessian-free online certified unlearning. In *The Thirteenth International Conference on Learning Representations*, 2025.

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

Nazanin Mohammadi Sepahvand, Anvith Thudi, Berivan Isik, Ashmita Bhattacharyya, Nicolas Papernot, Eleni Triantafillou, Daniel M Roy, and Gintare Karolina Dziugaite. Leveraging per-instance privacy for machine unlearning. In *International Conference on Machine Learning*, pp. 53906–53922. PMLR, 2025.

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

Ayush K Tarun, Vikram S Chundawat, Murari Mandal, and Mohan Kankanhalli. Fast yet effective machine unlearning. *IEEE Transactions on Neural Networks and Learning Systems*, 35(9):13046–13055, 2023.

Eleni Triantafillou, Ahmed Imtiaz Humayun, Monica Ribero, Alexander Matt Turner, Michael C Mozer, and Georgios Kaissis. Is your algorithm unlearning or untraining? *arXiv preprint arXiv:2604.07962*, 2026.

Binchi Zhang, Yushun Dong, Tianhao Wang, and Jundong Li. Towards certified unlearning for deep neural networks. In *International Conference on Machine Learning*, pp. 58800–58818. PMLR, 2024.

APPENDIX FOR THE PAPER: LABEL-LEVEL CERTIFIED UNTRAINING FOR PRIVACY GUARANTEES IN DEEP NEURAL NETWORKS

TABLE OF CONTENTS

A Related Work **13**

B Preliminaries from Statistics and Information Theory **14**

C Proof of Proposition 1 **16**

D Results on CUP factors **16**

D.1 δ_{tv}: Total Variation Bound 16

D.2 δ_{emp}: Empirical Error 16

D.3 δ_{etp}: Bin Count and Sample Bias 17

D.4 δ_{dom}: Domain Shift 17

E Results on Empirical Metrics **18**

E.1 Standard Methods 18

E.2 Data-Free Methods 21

E.3 Comments on the Results 22

E.4 White-Box MIA 22

E.5 Synthetic Calibration Sets 22

E.6 Efficiency 23

E.7 More Figures 23

A RELATED WORK

Untraining was recently introduced to delineate objectives and evaluation protocols in the broader field of *machine unlearning* (Triantafillou et al., 2026). Whereas *unlearning* removes a concept or an entire sub-population from the model, untraining targets privacy of individual training points and leaves the model's general knowledge intact, so that it still generalizes to every sub-population it was trained on. This ongoing discussion of evaluation protocols highlights the need for theoretically grounded privacy indicators, such as DP guarantees, beyond empirical metrics.

Untraining formulations are categorized into *exact*, *certified*, and *heuristic*, reflecting a verifiability–practicality trade-off. Exact methods seek statistical indistinguishability between the untrained and the retrained models, but they are restricted to simple algorithms (Ginart et al., 2019) and incur excessive overhead in DNNs (Bourtoule et al., 2021). The remaining formulations admit an approximation error in exchange for lower complexity. Certified methods bound this error through (ε, δ) DP-style guarantees, whereas *heuristic methods* rely solely on empirical metrics. Certified untraining thus balances verifiability and practicality, although this balance is difficult to maintain in DNNs.

Certification in the high-dimensional model space typically requires (strong) convexity assumptions that

Table 7: Privacy budgets for ResNet-18 models.

Prior Work	Privacy Budget (ε) ↓
Zhang et al. (2024)	[100, 5,000]
Qiao et al. (2025)	>120
Mu & Klabjan (2025)	$[10^4, 10^7]$

ensure loss boundedness and smoothness. Most prior work therefore restricts itself to convex settings, e.g., linear models (Sekhari et al., 2021; Chien et al., 2024a; Allouah et al., 2025), linear classification heads (Guo et al., 2020), and linearized DNNs (Basaran et al., 2025; Ahmed et al., 2025). Moreover, all of this work injects parameter-level noise, which in DNNs forces a choice between utility and a meaningful privacy budget. Table 7 shows that reported budgets substantially exceed established privacy standards, under which $\varepsilon \geq 10$ already counts as weak privacy (Near et al., 2025, Ch. 2.2). The cause is that the noise required for privacy grows with the model size (Neel et al., 2021; Sekhari et al., 2021; Koloskova et al., 2025; Allouah et al., 2025). In large DNNs, meaningful guarantees would require prohibitive noise and lead to severe degradation of generalization. Hence, prior work limits the noise to insufficient levels and yields significantly large privacy budgets that inflate privacy bounds e^{ε} to the point of "losing their theoretical meaning" (Mu & Klabjan, 2025) (Gap 1).

Moreover, prior work introduces assumption constants that are intractable in DNNs, or in non-convex settings in general, such as gradient Lipschitzness, Hessian smoothness and Lipschitzness, bounds on the Langevin diffusion rate, and strong convexity or its relaxations. These constants are arbitrary tunable hyperparameters in DNNs (Koloskova et al., 2025). Notably, Zhang et al. (2024); Qiao et al. (2025), who achieve the tightest reported privacy budgets, assume knowledge of minimum and maximum eigenvalues of DNN's Hessian, respectively; Mu & Klabjan (2025) assumes the Polyak–Lojasiewicz constant that relaxes strong convexity, and Sepahvand et al. (2025) inherits a strong dependence on the standard optimization hyperparameters from Chien et al. (2024b) (Gap 2).

B Preliminaries from Statistics and Information Theory

For completeness and clarity, we quote definitions, theorems and corollaries from (Cover & Thomas, 2006; Duchi, 2025; Ghazi & Issa, 2024). This section is not part of our theoretic contribution but supports our claims in the main manuscript.

Definition B.1 (Quantizers and Finite Algebras (Duchi, 2025, Ch 2.2.1))**.** Let $\mathcal{X}$ be an arbitrary space. A *quantizer* on $\mathcal{X}$ is any function q that maps $\mathcal{X}$ to a finite collection of integers. In particular, a quantizer q partitions the space $\mathcal{X}$ into the subsets of $x \in \mathcal{X}$ for which $\mathrm{q}(x) = i$. A related notion is an *algebra* of subsets of $\mathcal{X}$, which is any collection $\mathcal{B}$ satisfying

1. The set $\mathcal{X} \in \mathcal{B}$
2. The collection of sets $\mathcal{B}$ is closed under finite set operations: union, intersection, and complementation. That is, $B, A \in \mathcal{B}$ implies that $B^c \in \mathcal{B}$, $B \cap A \in \mathcal{B}$ and $B \cup A \in \mathcal{B}$.

There is a one-to-one correspondence between quantizers—and their associated partitions of the set $\mathcal{X}$—and finite algebras on a set $\mathcal{X}$. This one-to-one correspondence holds up to permutation of the partition the quantizer induces.

In one direction, consider a quantizer $\mathrm{q} : \mathcal{X} \to \{1, \ldots, m\}$. Let the sets $B_1, \ldots, B_m$ be the partitions associated with q via $B_i = \mathrm{q}^{-1}(\{i\})$. We may define the algebra $\mathcal{B}_{\mathrm{q}}$ as the collection of all finite set operations performed on $B_1, \ldots, B_m$. This is a finite collection, as finite set operations performed on the partition $B_1, \ldots, B_m$ induce only a finite collection of sets.

For the other direction, consider a finite algebra $\mathcal{B}$ over the set $\mathcal{X}$. We construct a quantizer $\mathrm{q}_{\mathcal{B}}$ corresponding to this algebra. To do so we define an *atom* of $\mathcal{B}$ as any non-empty set $B \in \mathcal{B}$, such that if $A \subseteq B$ and $A \in \mathcal{B}$, then $A = B$ or $A = \emptyset$, i.e., the atoms of $\mathcal{B}$ are the "smallest" sets in $\mathcal{B}$. There is a unique partition of $\mathcal{X}$ with atomic sets from $\mathcal{B}$ (the proof is omitted). As $\mathcal{B}$ is finite, there is a finite index m that defines a collection $B_1, \ldots, B_m$ of disjoint atomic sets in $\mathcal{B}$ for which $\cup_i B_i = \mathcal{X}$. Thus, we may define the quantizer $\mathrm{q}_{\mathcal{B}}$ via $\mathrm{q}_{\mathcal{B}}(x) = i$ when $x \in B_i$.

Definition B.2 (Type (Cover & Thomas, 2006, Ch 11.1))**.** The *type* $P_{\mathbf{x}}$ (or empirical probability distribution) of a sequence $x_1, x_2, \ldots, x_n$ is the relative proportion of occurrences of each symbol of $\mathcal{X}$ (i.e., $P_{\mathbf{x}}(\alpha) = N(\alpha \mid \mathbf{x})/n$) for all $\alpha \in \mathcal{X}$, where $N(\alpha \mid \mathbf{x})$ is the number of times the symbol α occurs in the sequence $\mathbf{x} \in \mathcal{X}^n$. The type of the sequence is a probability mass function on $\mathcal{X}$.

Theorem B.1 (Standard bound from the Method of Types (Cover & Thomas, 2006, Thm. 11.2.1))**.** Let $X_1, X_2, \ldots, X_n$ be i.i.d. drawn according to a population distribution $P(x)$ and P_{x^n} the type of $x_1, x_2, \ldots, x_n$. Then

$$\mathbb{P}\{D_{\mathrm{kl}}(P_{x^n} \mid\mid P) > \epsilon\} \leq 2^{-n\left(\epsilon - |\mathcal{X}| \frac{\log_2(n+1)}{n}\right)}, \tag{6}$$

and consequently, $D_{\mathrm{kl}}(P_{x^n} \mid\mid P) \to 0$ with probability 1, where D_{kl} is the KL divergence computed in bits.

Proposition B.1 (Pinsker's inequality (Duchi, 2025, Prop. 2.2.9))**.** For any distributions P and Q, the total variation distance $d_{\mathrm{tv}}(P, Q)$ satisfies

$$d_{\mathrm{tv}}(P, Q) \leq \sqrt{\frac{1}{2} D_{\mathrm{kl}}(P \mid\mid Q)}, \tag{7}$$

where D_{kl} is the Kullback–Leibler (KL) divergence computed with natural logarithms (base e), so that it is measured in *nats*.

Proposition B.2 (Jensen's inequality for total variation)**.** Let P and Q be probability distributions on a common measurable space $\mathcal{X}$, and let $\hat{P}$ and $\hat{Q}$ be the empirical distributions of samples $X_1, \ldots, X_n \sim P$ and $Y_1, \ldots, Y_n \sim Q$, respectively. Then

$$d_{\mathrm{tv}}(P, Q) = d_{\mathrm{tv}}\left(\mathbb{E}[\hat{P}], \mathbb{E}[\hat{Q}]\right) \leq \mathbb{E}\big[d_{\mathrm{tv}}\left(\hat{P}, \hat{Q}\right)\big]. \tag{8}$$

Proof. The empirical distributions are unbiased estimators of the population distributions, i.e., $\mathbb{E}[\hat{P}] = P$ and $\mathbb{E}[\hat{Q}] = Q$. Since the total variation distance is jointly convex, the claim follows from Jensen's inequality (Duchi, 2025, Prop. 2.1.5). □

Proposition B.3 (Le Cam's lemma (Duchi, 2025, Prop. 2.3.1))**.** Suppose we have a Bayesian hypothesis testing problem where V is chosen with equal probability to be 1 or 2, and given $V = v$, the sample X is drawn from the distribution P_v. For any distributions P_1 and P_2 on $\mathcal{X}$, we have

$$\inf_{\Psi} \{P_1(\Psi(X) \neq 1) + P_2(\Psi(X) \neq 2)\} = 1 - d_{\mathrm{tv}}(P_1, P_2), \tag{9}$$

where the infimum is taken over all tests $\Psi : \mathcal{X} \to \{1, 2\}$.

Proposition B.4 (From (ε, δ) to $(0, \delta_{\mathrm{tv}})$ differential privacy (Ghazi & Issa, 2024, Cor. 1 & Rem. 1))**.** Let P and Q be probability distributions on a measurable space $[0, 1]$, and let $d_{\mathrm{tv}}(P, Q) := \sup_{S \subseteq [0,1]} \{P(S) - Q(S)\}$. Let $\varepsilon \geq 0$ and $\delta \in [0, 1]$. If for all $S \subseteq [0, 1]$

$$P(S) \leq e^{\varepsilon} Q(S) + \delta \quad \text{and} \quad Q(S) \leq e^{\varepsilon} P(S) + \delta, \tag{10}$$

then

$$d_{\mathrm{tv}}(P, Q) \;\leq\; \frac{e^{\varepsilon} - 1 + 2\delta}{e^{\varepsilon} + 1} \;=\; \delta + (1 - \delta)\frac{e^{\varepsilon} - 1}{e^{\varepsilon} + 1}. \tag{11}$$

In particular, for $\varepsilon = 0$, condition Eq. (10) holds if and only if $\delta_{\mathrm{tv}} := d_{\mathrm{tv}}(P, Q) \leq \delta$. In general, δ_{tv} is the lowest failure probability of $P(S)/Q(S) \leq e^{\varepsilon}$ and $Q(S)/P(S) \leq e^{\varepsilon}$.

Lemma B.1 (Le Cam's interpretation of $(0, \delta_{\mathrm{tv}})$ differential privacy)**.** Let P_1, P_2 be distributions on a measurable space $\mathcal{X}$ and $\delta_{\mathrm{tv}} \in [0, 1]$. Let V be drawn uniformly at random from $\{1, 2\}$. Given $V{=}v$, let $X{\sim}P_v$. That is, if $V{=}1$, then $X{\sim}P_1$, and if $V{=}2$, then $X{\sim}P_2$. Then, by Proposition B.3,

$$\inf_{\Psi} \; \mathbb{P}\{\Psi(X) \neq V\} \;=\; \tfrac{1}{2}\big(1 - d_{\mathrm{tv}}(P_1, P_2)\big), \tag{12}$$

where the infimum is taken over all tests $\Psi : \mathcal{X} \to \{1, 2\}$. Consequently, if $P_1(S) \leq P_2(S) + \delta_{\mathrm{tv}}$ and $P_2(S) \leq P_1(S) + \delta_{\mathrm{tv}}$ for all measurable $S \subseteq \mathcal{X}$, then $d_{\mathrm{tv}}(P_1, P_2) \leq \delta_{\mathrm{tv}}$ by Proposition B.4, and every test fails to distinguish whether X was drawn from P_1 or P_2 with probability at least $\frac{1}{2}(1 - \delta_{\mathrm{tv}})$.

Proof. Since V is a random draw

$$\mathbb{P}\{\Psi(X) \neq V\} = \tfrac{1}{2}\left[P_1(\Psi(X) \neq 1) + P_2(\Psi(X) \neq 2)\right] \tag{13}$$

Then, taking the infimum over Ψ and applying Proposition B.3

$$\inf_{\Psi} \mathbb{P}\{\Psi(X) \neq V\} = \tfrac{1}{2}\big[1 - d_{\mathrm{tv}}(P_1, P_2)\big] \tag{14}$$

The proof follows via substitution from Proposition B.4. □

C PROOF OF PROPOSITION 1

Proof. Let $\mathcal{C}_{n_{\text{cal}}}{:=}\mathcal{C}(\widehat{\mathcal{D}}_{\text{cal}};\theta_0)$ be an empirical distribution of n_{cal} confidence scores drawn i.i.d. according to a population distribution $\mathcal{C}_\star{:=}\mathcal{C}(\mathcal{D}_{\text{cal}};\theta_0)$. Both are defined over a finite count of m bins. By substitution in Theorem B.1, and computing the KL divergence d_{kl} in nats, the concentration inequality becomes

$$\mathbb{P}\{d_{\text{kl}}(\mathcal{C}_{n_{\text{cal}}} \,||\, \mathcal{C}_\star) > \epsilon\} \leq e^{-n_{\text{cal}}\left(\epsilon - m\frac{\ln(n_{\text{cal}}+1)}{n_{\text{cal}}}\right)} \tag{15}$$

$$1 - \mathbb{P}\{d_{\text{kl}}(\mathcal{C}_{n_{\text{cal}}} \,||\, \mathcal{C}_\star) > \epsilon\} \geq 1 - e^{-n_{\text{cal}}\left(\epsilon - m\frac{\ln(n_{\text{cal}}+1)}{n_{\text{cal}}}\right)} \tag{16}$$

$$\mathbb{P}\{d_{\text{kl}}(\mathcal{C}_{n_{\text{cal}}} \,||\, \mathcal{C}_\star) \leq \epsilon\} \geq 1 - e^{-n_{\text{cal}}\left(\epsilon - m\frac{\ln(n_{\text{cal}}+1)}{n_{\text{cal}}}\right)} \tag{17}$$

We seek an ϵ that upper-bounds the d_{kl} with a strong probability. Note that if $\epsilon = m\frac{\ln(n_{\text{cal}}+1)}{n_{\text{cal}}}$ the inequality holds with probability ≥ 0, which is not informative. Hence we assign $\epsilon{=}m\frac{\ln(n_{\text{cal}}+1)}{n_{\text{cal}}} + \beta$, for some $\beta > 0$. By substitution in Eq. (17)

$$\mathbb{P}\left\{d_{\text{kl}}(\mathcal{C}_{n_{\text{cal}}} \,||\, \mathcal{C}_\star) \leq \frac{m\ln(n_{\text{cal}}+1)}{n_{\text{cal}}} + \beta\right\} \geq 1 - e^{-n_{\text{cal}}\beta}. \tag{18}$$

Then, we apply Pinsker's inequality Proposition B.1 and derive an upper-bound of the total variation d_{tv} that holds with probability $\geq 1 - e^{-n_{\text{cal}}\beta}$

$$\mathbb{P}\left\{d_{\text{tv}}\left(\mathcal{C}_{n_{\text{cal}}}, \mathcal{C}_\star\right) \leq \sqrt{\tfrac{1}{2}\left(\frac{m\ln(n_{\text{cal}}+1)}{n_{\text{cal}}} + \beta\right)}\right\} \geq 1 - e^{-n_{\text{cal}}\beta}. \tag{19}$$

□

Note that, for a given sample size n_{cal}, the bin count m is not a free parameter. The upper bound increases with m, and decreases with n_{cal}.

Note that β is not a free parameter either; it is inversely proportional to n_{cal}. Thus the fewer the calibration samples, the looser the upper-bound. We fix a strong holding probability of $1{-}10^{-5}$, i.e., $\beta{=}\ln(10^5)/n_{\text{cal}}$, which gives $\beta{\approx}0.0023$ for $n_{\text{cal}}{=}5{,}000$ and $\beta{\approx}0.00023$ for $n_{\text{cal}}{=}50{,}000$ (see Table 12).

D RESULTS ON CUP FACTORS

D.1 δ_{tv}: TOTAL VARIATION BOUND

Table 8: δ_{tv} *per setting for CUP.*

Setting	δ_{etp}	δ_{emp}	δ_{tv}
ViT-L & TImNet	0.0477	$0.0393_{\pm 0.0033}$	0.0870
ViT-B & TImNet	0.0477	$0.0162_{\pm 0.0017}$	0.0639
ViT-B & C-100	0.1180	$0.0192_{\pm 0.0023}$	0.1372
RN-18 & C-10	0.1180	$0.0235_{\pm 0.0025}$	0.1415

Table 9: δ_{tv} *per setting for CUP-DF.*

Setting	δ_{etp}	δ_{emp}	δ_{tv}
ViT-L & TImNet	0.0477	$0.0218_{\pm 0.0048}$	0.0695
ViT-B & TImNet	0.0477	$0.0243_{\pm 0.0021}$	0.0720
ViT-B & C-100	0.1180	$0.0219_{\pm 0.0052}$	0.1399
RN-18 & C-10	0.1180	$0.0521_{\pm 0.0097}$	0.1701

D.2 δ_{emp}: EMPIRICAL ERROR

Table 10: δ_{emp} *per setting for CUP*. We report mean±std over three independent experiments. $\delta_{\text{emp}}^{\text{init}}$ is the empirical error before untraining.

Setting	$\delta_{\text{emp}}^{\text{init}} \to \delta_{\text{emp}}$
ViT-Large & TinyImageNet	$0.1106_{\pm 0.0038} \to 0.0393_{\pm 0.0033}$
ViT-Base & TinyImageNet	$0.0742_{\pm 0.0039} \to 0.0162_{\pm 0.0017}$
ViT-Base & CIFAR-100	$0.0626_{\pm 0.0025} \to 0.0192_{\pm 0.0023}$
ResNet-18 & CIFAR-10	$0.1270_{\pm 0.0051} \to 0.0235_{\pm 0.0025}$

Table 11: δ_{emp} *per setting for CUP-DF*. We report mean±std over three independent experiments. $\delta_{\text{emp}}^{\text{init}}$ is the empirical error before untraining.

Setting	$\delta_{\text{emp}}^{\text{init}} \to \delta_{\text{emp}}$
ViT-Large & TinyImageNet	$0.1104_{\pm 0.0031} \to 0.0218_{\pm 0.0048}$
ViT-Base & TinyImageNet	$0.0752_{\pm 0.0045} \to 0.0243_{\pm 0.0021}$
ViT-Base & CIFAR-100	$0.0628_{\pm 0.0036} \to 0.0219_{\pm 0.0052}$
ResNet-18 & CIFAR-10	$0.1275_{\pm 0.0043} \to 0.0521_{\pm 0.0097}$

D.3 δ_{etp}: Bin Count and Sample Bias

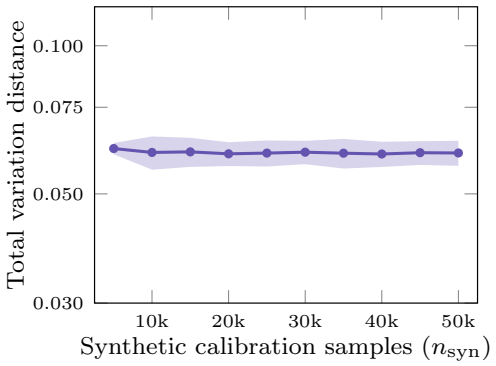
Table 12: δ_{etp} *per calibration set.* Higher sample count n_{cal} enables equally strong holding probability with smaller β, and diminishes δ_{etp} in general (see CIFAKE: n_{cal}=5,000 → 50,000). Abundance of synthetic data can diminish δ_{etp}. Higher n_{cal} enables higher bin count m for increased resolution during calibration.

Calibration Set	n_{cal}	m	β	δ_{etp}
CIFAKE	$5,000$	15	0.0023	0.1180
CIFAKE	$50,000$	15	0.00023	0.0417
CIFAR-10/100	$5,000$	15	0.0023	0.1180
TinyImageNet	$50,000$	20	0.00023	0.0477

Table 13: *Ablation on the number of bins (m).* While CUP's untraining performance (AG) and the empirical alignment (δ_{emp}) remain largely unchanged, the theoretical verification bound δ_{etp} increases significantly with m (at a fixed $\beta = \ln(10^5)/n_{\text{cal}} \approx 0.0023$), and acts as the dominant component of the total TV bound δ_{tv}. We adopt m=15 following Guo et al. (2017), which yields empirical results as good as higher m, while keeping δ_{etp} relatively tight.

m	Δ(RA)↓	Δ(TA)↓	Δ(FA)↓	Δ(MIA)↓	AvgGap↓	δ_{emp}↓	δ_{etp}↓	δ_{tv}↓
15	1.00	0.44	2.45	6.24	2.53	$0.0235_{\pm 0.0025}$	0.1180	0.1415
20	0.94	0.19	2.99	5.96	2.52	$0.0211_{\pm 0.0050}$	0.1349	0.1560
50	1.04	0.23	2.54	6.23	2.51	$0.0219_{\pm 0.0022}$	0.2091	0.2310
Std Dev	0.05	0.13	0.29	0.16	0.01	0.0012	0.0485	0.0480

D.4 δ_{dom}: Domain Shift



$n_{\text{synthetic}}$	δ_{dom}
5k	0.0618 ± 0.0016
10k	0.0607 ± 0.0047
15k	0.0609 ± 0.0041
20k	0.0603 ± 0.0034
25k	0.0605 ± 0.0037
30k	0.0608 ± 0.0033
35k	0.0605 ± 0.0042
40k	0.0603 ± 0.0035
45k	0.0606 ± 0.0034
50k	0.0606 ± 0.0035

Figure 8: The total variation distance between the confidence-score distributions induced by the original model on a held-out dataset split (i.e., the validation set) and on n_{syn} synthetic calibration data. The total variation stays flat as we evaluate with synthetic datasets of different size and composition. ResNet-18, CIFAR-10, CIFAKE.

E Results on Empirical Metrics

From Table 14 to Table 21 we report mean performance and standard deviations (mean±std) over three independent trials with different forget and retain sets. **Metrics**: *average gap* (AG) assesses balance between privacy and utility based on *retain accuracy* (RA), *test accuracy* (TA), *forget accuracy* (FA) and the success rate of a black-box *membership inference attack* (MIA). Δ denotes the absolute difference w.r.t. retraining (performance gaps) in percentage points (pp); smaller gaps indicate stronger performance. *AG* (average gap) is the average over RA, FA, TA, MIA gaps. *Time* refers to wall-clock runtime in minutes. Standard-methods fine-tune on the retain set (i.e., the remaining training data, $D_{\text{tr}} \setminus D_{\text{for}}$) to facilitate preservation of model utility, while data-free methods assume no access to the retain set according to challenging, real-world scenarios.

AG is the standard indicator of untraining. Standalone metrics (RA, TA, FA, and MIA) capture only specific facets of evaluation. For instance, a method may perform relatively well on preserving model utility (see RA, TA) at the cost of compromising privacy significantly.

Streisand effect. MIA success scores are also computed w.r.t. retraining, because lower-than-retraining scores may put the forget set under the spotlight, compromising privacy. For instance, consider a pattern where images of "cats" in the forget set are classified as "airplanes". Then MIA scores could be low, but an adversary can identify training data (explicitly in the forget set) by exploiting this pattern (Chundawat et al., 2023; Foster et al., 2024a; Spartalis et al., 2025).

Coloring refers to relative evaluation. We use coloring (best to worst) to enhance comparison readability; it lacks absolute interpretation. For instance, in Table 14, SCRUB's runtime is relatively worse compared to other approaches, but almost half of that required for retraining. The same applies to BadT's empirical metrics. However, gaps that are close to 100% have absolute meaning, indicating catastrophic forgetting.

E.1 Standard Methods

Table 14: Standard-method results for **ViT-Base & CIFAR-100** (untraining 10% of the training data).

Method	AG ↓	RA(Δ ↓)	TA(Δ ↓)	FA(Δ ↓)	MIA(Δ ↓)	Time ↓
Original	2.04	$96.10_{\pm 0.03}$(0.14)	$91.44_{\pm 0.12}$(0.20)	$96.64_{\pm 0.10}$(4.75)	$74.82_{\pm 0.25}$(3.08)	$121.97_{\pm 2.34}$
Retrained	0.00	$95.96_{\pm 0.06}$(0.00)	$91.24_{\pm 0.17}$(0.00)	$91.89_{\pm 0.11}$(0.00)	$71.74_{\pm 0.31}$(0.00)	$112.70_{\pm 0.08}$
Fine-tuning	3.47	$99.81_{\pm 0.01}$(3.85)	$91.98_{\pm 0.09}$(0.74)	$94.46_{\pm 0.26}$(2.57)	$78.45_{\pm 0.12}$(6.71)	$38.01_{\pm 0.09}$
RandLabel	1.64	$99.49_{\pm 0.25}$(3.53)	$92.27_{\pm 0.08}$(1.03)	$93.25_{\pm 0.60}$(1.36)	**$72.36_{\pm 1.17}$(0.62)**	$42.46_{\pm 0.03}$
NegGrad+	6.91	$99.82_{\pm 0.02}$(3.86)	$92.41_{\pm 0.10}$(1.17)	$99.81_{\pm 0.03}$(7.92)	$86.41_{\pm 0.43}$(14.67)	$42.17_{\pm 0.17}$
SCRUB	2.01	$96.09_{\pm 0.01}$(0.13)	$91.44_{\pm 0.12}$(0.20)	$96.65_{\pm 0.16}$(4.76)	$74.67_{\pm 0.20}$(2.93)	$55.82_{\pm 0.10}$
BadT	4.52	$90.41_{\pm 0.53}$(5.55)	$88.07_{\pm 0.58}$(3.17)	$89.20_{\pm 0.64}$(2.69)	$65.06_{\pm 0.46}$(6.68)	$14.37_{\pm 0.10}$
UNSIR	1.71	$92.93_{\pm 0.10}$(3.03)	$89.75_{\pm 0.33}$(1.49)	$92.47_{\pm 0.21}$(0.58)	$73.50_{\pm 0.47}$(1.76)	**$3.80_{\pm 0.03}$**
SalUn	1.67	$98.29_{\pm 0.45}$(2.33)	$91.89_{\pm 0.11}$(0.65)	$93.60_{\pm 0.43}$(1.71)	$73.71_{\pm 1.21}$(1.97)	$47.42_{\pm 0.13}$
LoTUS	1.58	**$95.98_{\pm 0.07}$(0.02)**	**$91.30_{\pm 0.14}$(0.06)**	$96.53_{\pm 0.05}$(4.64)	$73.34_{\pm 0.16}$(1.60)	$6.96_{\pm 0.04}$
CUP	**1.12**	$94.11_{\pm 0.32}$(1.85)	$90.59_{\pm 0.24}$(0.65)	**$91.56_{\pm 0.05}$(0.33)**	$73.38_{\pm 0.24}$(1.64)	$12.47_{\pm 0.14}$

Table 15: Standard-method results for **ViT-Base & TinyImageNet** (untraining 1%, 10%, and 20% of the training data).

Method	AG ↓	RA(Δ ↓)	TA(Δ ↓)	FA(Δ ↓)	MIA(Δ ↓)	Time ↓
Original	2.57	$96.45_{\pm 0.04}$(0.08)	$90.22_{\pm 0.12}$(0.21)	$96.38_{\pm 0.06}$(6.31)	$79.43_{\pm 0.15}$(3.66)	$258.46_{\pm 0.52}$
			1%			
Retrained	0.00	$96.34_{\pm 0.01}$(0.00)	$90.69_{\pm 0.07}$(0.00)	$89.74_{\pm 0.06}$(0.00)	$75.95_{\pm 0.08}$(0.00)	$223.83_{\pm 0.08}$
Fine-tuning	2.58	$99.32_{\pm 0.11}$(2.98)	$89.37_{\pm 0.55}$(1.32)	$93.10_{\pm 0.14}$(3.36)	$78.60_{\pm 0.57}$(2.65)	$82.55_{\pm 0.34}$
RandLabel	9.80	$99.14_{\pm 0.33}$(2.80)	$88.81_{\pm 0.33}$(1.88)	$79.60_{\pm 2.43}$(10.14)	$51.57_{\pm 5.37}$(24.38)	$84.14_{\pm 0.34}$
NegGrad+	5.97	$98.94_{\pm 0.08}$(2.60)	$88.56_{\pm 0.42}$(2.13)	$98.95_{\pm 0.07}$(9.21)	$85.90_{\pm 0.71}$(9.95)	$85.10_{\pm 1.61}$
SCRUB	2.51	$96.43_{\pm 0.05}$(0.09)	$90.23_{\pm 0.79}$(0.46)	$96.67_{\pm 0.38}$(6.93)	$78.50_{\pm 0.60}$(2.55)	$112.53_{\pm 1.17}$
BadT	15.22	$86.54_{\pm 0.17}$(9.80)	$82.36_{\pm 0.75}$(8.33)	$73.47_{\pm 1.16}$(16.27)	$49.47_{\pm 2.87}$(26.48)	$23.01_{\pm 0.10}$
UNSIR	2.80	$91.56_{\pm 0.91}$(4.78)	$87.32_{\pm 0.93}$(3.37)	$90.55_{\pm 0.92}$(0.81)	$73.70_{\pm 0.57}$(2.25)	**$2.65_{\pm 0.00}$**
SalUn	3.43	$98.58_{\pm 0.29}$(2.24)	$89.59_{\pm 0.33}$(1.10)	$88.37_{\pm 0.60}$(1.37)	$66.93_{\pm 1.33}$(9.02)	$64.11_{\pm 0.13}$
LoTUS	**1.96**	**$96.39_{\pm 0.02}$(0.05)**	**$90.27_{\pm 0.82}$(0.42)**	$96.37_{\pm 0.31}$(6.63)	**$76.67_{\pm 0.31}$(0.72)**	$10.66_{\pm 0.01}$
CUP	2.52	$92.12_{\pm 0.19}$(4.22)	$88.27_{\pm 0.47}$(2.42)	**$89.70_{\pm 0.71}$(0.04)**	$79.35_{\pm 0.35}$(3.40)	$4.24_{\pm 0.16}$
			10%			
Retrained	0.00	$96.37_{\pm 0.05}$(0.00)	$90.01_{\pm 0.05}$(0.00)	$90.07_{\pm 0.04}$(0.00)	$75.77_{\pm 0.06}$(0.00)	$228.90_{\pm 6.49}$
Fine-tuning	1.72	$98.97_{\pm 0.22}$(2.60)	$88.89_{\pm 0.26}$(1.12)	$90.67_{\pm 0.18}$(0.60)	$78.32_{\pm 0.23}$(2.55)	$75.64_{\pm 0.38}$
RandLabel	1.69	$99.21_{\pm 0.06}$(2.84)	$88.51_{\pm 0.33}$(1.50)	$89.39_{\pm 0.13}$(0.68)	$74.03_{\pm 0.56}$(1.74)	$84.40_{\pm 0.35}$
NegGrad+	5.91	$99.21_{\pm 0.25}$(2.84)	$88.93_{\pm 0.61}$(1.08)	$99.08_{\pm 0.30}$(9.01)	$86.48_{\pm 0.83}$(10.71)	$84.16_{\pm 0.11}$
SCRUB	2.56	$96.44_{\pm 0.04}$(0.07)	$90.23_{\pm 0.79}$(0.22)	$96.59_{\pm 0.06}$(6.52)	$79.21_{\pm 0.17}$(3.44)	$111.03_{\pm 0.13}$
BadT	9.20	$85.81_{\pm 0.69}$(10.56)	$81.67_{\pm 1.02}$(8.34)	$81.93_{\pm 0.94}$(8.14)	$66.01_{\pm 0.72}$(9.76)	$27.79_{\pm 0.11}$
UNSIR	1.71	$92.23_{\pm 0.12}$(4.14)	$88.30_{\pm 0.11}$(1.71)	$90.86_{\pm 0.45}$(0.79)	**$75.58_{\pm 0.55}$(0.19)**	**$7.38_{\pm 0.03}$**
SalUn	1.17	$98.54_{\pm 0.35}$(2.17)	$89.62_{\pm 0.51}$(0.39)	$91.85_{\pm 0.10}$(1.78)	$76.10_{\pm 0.21}$(0.33)	$106.73_{\pm 0.40}$
LoTUS	1.96	**$96.38_{\pm 0.05}$(0.01)**	**$90.21_{\pm 0.83}$(0.20)**	$96.43_{\pm 0.09}$(6.36)	$77.04_{\pm 0.24}$(1.27)	$13.77_{\pm 0.01}$
CUP	**0.89**	$94.36_{\pm 0.83}$(2.01)	$89.34_{\pm 0.64}$(0.67)	**$90.07_{\pm 1.05}$(0.00)**	$76.64_{\pm 0.46}$(0.87)	$21.42_{\pm 6.30}$
			20%			
Retrained	0.00	$96.37_{\pm 0.05}$(0.00)	$90.31_{\pm 0.67}$(0.00)	$89.77_{\pm 0.08}$(0.00)	$75.89_{\pm 0.13}$(0.00)	$227.89_{\pm 1.63}$
Fine-tuning	1.84	$99.33_{\pm 0.08}$(2.96)	$88.92_{\pm 0.21}$(1.39)	**$90.82_{\pm 0.21}$(1.05)**	$77.85_{\pm 0.36}$(1.96)	$66.96_{\pm 0.28}$
RandLabel	4.35	$97.04_{\pm 0.52}$(0.67)	$84.15_{\pm 0.21}$(6.16)	$84.39_{\pm 0.33}$(5.38)	$70.69_{\pm 0.94}$(5.20)	$85.75_{\pm 0.38}$
NegGrad+	5.88	$99.12_{\pm 0.22}$(2.75)	$88.59_{\pm 0.29}$(1.72)	$98.79_{\pm 0.26}$(9.02)	$85.91_{\pm 0.33}$(10.02)	$83.43_{\pm 0.01}$
SCRUB	2.47	$96.46_{\pm 0.05}$(0.09)	$90.23_{\pm 0.79}$(0.08)	$96.42_{\pm 0.06}$(6.65)	$78.97_{\pm 0.04}$(3.08)	$110.62_{\pm 0.03}$
BadT	10.66	$84.66_{\pm 0.48}$(11.71)	$80.05_{\pm 0.60}$(10.26)	$79.88_{\pm 0.61}$(9.89)	$65.13_{\pm 0.41}$(10.76)	$32.82_{\pm 0.16}$
UNSIR	1.78	$92.94_{\pm 0.41}$(3.43)	$88.37_{\pm 0.49}$(1.94)	$91.14_{\pm 0.58}$(1.37)	$76.27_{\pm 0.55}$(0.38)	**$12.49_{\pm 0.02}$**
SalUn	1.24	$98.48_{\pm 0.40}$(2.11)	$89.58_{\pm 0.52}$(0.73)	$91.76_{\pm 0.43}$(1.99)	**$75.76_{\pm 0.76}$(0.13)**	$149.20_{\pm 0.50}$
LoTUS	1.70	**$96.40_{\pm 0.07}$(0.03)**	**$90.26_{\pm 0.92}$(0.05)**	$96.23_{\pm 0.08}$(6.46)	$75.64_{\pm 0.40}$(0.25)	$17.03_{\pm 0.02}$
CUP	**0.69**	$95.57_{\pm 0.32}$(0.80)	$89.96_{\pm 0.04}$(0.35)	$91.14_{\pm 0.38}$(1.37)	$76.14_{\pm 0.48}$(0.25)	$49.20_{\pm 2.27}$

Table 16: Standard-method results for **ViT-Large & TinyImageNet** (untraining 10% of the training set).

Method	AG ↓	RA(Δ ↓)	TA(Δ ↓)	FA(Δ ↓)	MIA(Δ ↓)	Time ↓
Original	3.53	$99.35_{\pm 0.01}(0.03)$	$91.50_{\pm 0.47}(0.23)$	$99.48_{\pm 0.01}(7.46)$	$86.19_{\pm 0.29}(6.39)$	$383.64_{\pm 0.07}$
Retrained	0.00	$99.32_{\pm 0.01}(0.00)$	$91.27_{\pm 0.07}(0.00)$	$92.02_{\pm 0.04}(0.00)$	$79.80_{\pm 0.06}(0.00)$	$345.25_{\pm 0.03}$
Fine-tuning	2.37	$99.72_{\pm 0.10}(0.40)$	**$91.60_{\pm 0.32}(0.33)$**	$96.73_{\pm 2.28}(4.71)$	$83.82_{\pm 1.39}(4.02)$	$114.87_{\pm 0.11}$
RandLabel	1.82	**$99.20_{\pm 0.11}(0.12)$**	$92.29_{\pm 0.65}(1.02)$	$97.27_{\pm 0.32}(5.25)$	$78.92_{\pm 0.20}(0.88)$	$128.20_{\pm 0.02}$
NegGrad+	90.24	$0.50_{\pm 0.00}(98.82)$	$0.44_{\pm 0.05}(90.83)$	$0.50_{\pm 0.00}(91.52)$	$0.00_{\pm 0.00}(79.80)$	$119.45_{\pm 0.08}$
SCRUB	90.24	$0.50_{\pm 0.00}(98.82)$	$0.45_{\pm 0.04}(90.82)$	$0.50_{\pm 0.00}(91.52)$	$0.00_{\pm 0.00}(79.80)$	$167.34_{\pm 0.01}$
BadT	14.18	$83.28_{\pm 0.42}(16.04)$	$77.74_{\pm 0.54}(13.53)$	$78.12_{\pm 0.85}(13.90)$	$66.54_{\pm 1.12}(13.26)$	$36.83_{\pm 2.17}$
UNSIR	2.08	$94.53_{\pm 0.33}(4.79)$	$89.96_{\pm 0.40}(1.31)$	$93.76_{\pm 0.18}(1.74)$	**$80.26_{\pm 0.12}(0.46)$**	**$14.71_{\pm 0.00}$**
SalUn	76.83	$9.80_{\pm 1.41}(89.52)$	$9.39_{\pm 0.99}(81.88)$	$9.22_{\pm 1.40}(82.80)$	$26.66_{\pm 23.15}(53.14)$	$277.49_{\pm 3.90}$
LoTUS	2.68	$99.13_{\pm 0.02}(0.19)$	$92.01_{\pm 0.54}(0.74)$	$99.05_{\pm 0.07}(7.03)$	$77.05_{\pm 0.19}(2.75)$	$24.10_{\pm 0.00}$
CUP	**1.74**	$98.59_{\pm 0.68}(0.73)$	$90.17_{\pm 0.76}(1.10)$	**$93.66_{\pm 2.16}(1.64)$**	$83.30_{\pm 0.64}(3.50)$	$33.21_{\pm 15.37}$

Table 17: Standard-method results for **ResNet-18 & CIFAR-10**, merged across forget-set sizes (1%, 10%, 20% of the training data).

Method	AG ↓	RA(Δ ↓)	TA(Δ ↓)	FA(Δ ↓)	MIA(Δ ↓)	Time ↓
Original	7.15	$99.98_{\pm 0.01}(1.18)$	$92.62_{\pm 0.49}(1.44)$	$99.99_{\pm 0.01}(8.66)$	$94.03_{\pm 0.39}(17.32)$	$6.62_{\pm 0.12}$
			1%			
Retrained	0.00	$99.98_{\pm 0.01}(0.00)$	$92.64_{\pm 0.32}(0.00)$	$93.93_{\pm 1.29}(0.00)$	$78.33_{\pm 0.61}(0.00)$	$6.57_{\pm 0.00}$
Fine-tuning	5.11	**$99.98_{\pm 0.01}(0.00)$**	$92.40_{\pm 0.37}(0.24)$	$100.00_{\pm 0.00}(6.07)$	$92.47_{\pm 1.17}(14.14)$	$0.48_{\pm 0.00}$
RandLabel	2.56	**$99.98_{\pm 0.02}(0.00)$**	$92.41_{\pm 0.37}(0.23)$	$98.60_{\pm 0.80}(4.67)$	$73.00_{\pm 1.93}(5.33)$	$0.50_{\pm 0.01}$
NegGrad+	5.55	$99.97_{\pm 0.01}(0.01)$	$92.47_{\pm 0.48}(0.17)$	$99.93_{\pm 0.12}(6.00)$	$94.33_{\pm 0.31}(16.00)$	$0.50_{\pm 0.01}$
SCRUB	5.66	**$99.98_{\pm 0.01}(0.00)$**	**$92.62_{\pm 0.49}(0.02)$**	$100.00_{\pm 0.00}(6.07)$	$94.87_{\pm 0.83}(16.54)$	$0.57_{\pm 0.01}$
BadT	4.89	$99.52_{\pm 0.09}(0.46)$	$91.60_{\pm 0.73}(1.04)$	$99.60_{\pm 0.00}(5.67)$	$90.73_{\pm 0.92}(12.40)$	$0.18_{\pm 0.00}$
UNSIR	4.96	$99.57_{\pm 0.04}(0.41)$	$91.56_{\pm 0.65}(1.08)$	$99.60_{\pm 0.35}(5.67)$	$91.00_{\pm 0.60}(12.67)$	**$0.09_{\pm 0.00}$**
SalUn	2.17	$99.96_{\pm 0.01}(0.02)$	$92.52_{\pm 0.35}(0.12)$	$99.73_{\pm 0.31}(5.80)$	**$81.07_{\pm 1.60}(2.74)$**	$1.13_{\pm 0.02}$
CNS	5.76	**$99.98_{\pm 0.01}(0.00)$**	$92.59_{\pm 0.48}(0.05)$	$100.00_{\pm 0.00}(6.07)$	$95.27_{\pm 1.17}(16.94)$	$3.89_{\pm 0.02}$
LoTUS	5.41	$99.12_{\pm 0.11}(0.86)$	$90.93_{\pm 0.51}(1.71)$	$99.27_{\pm 0.31}(5.34)$	$92.07_{\pm 0.12}(13.74)$	$0.23_{\pm 0.00}$
CUP	**1.95**	$99.00_{\pm 0.12}(0.98)$	$90.71_{\pm 0.57}(1.93)$	**$93.73_{\pm 1.29}(0.20)$**	$83.00_{\pm 0.92}(4.67)$	$0.15_{\pm 0.03}$
			10%			
Retrained	0.00	$98.80_{\pm 2.04}(0.00)$	$91.18_{\pm 2.25}(0.00)$	$91.33_{\pm 2.39}(0.00)$	$76.71_{\pm 3.10}(0.00)$	$5.39_{\pm 1.20}$
Fine-tuning	6.89	$99.98_{\pm 0.01}(1.18)$	$92.53_{\pm 0.47}(1.35)$	$99.99_{\pm 0.01}(8.66)$	$93.09_{\pm 0.08}(16.38)$	$0.43_{\pm 0.01}$
RandLabel	3.10	$99.94_{\pm 0.01}(1.14)$	$91.97_{\pm 0.52}(0.79)$	$99.14_{\pm 0.06}(7.81)$	$74.03_{\pm 0.88}(2.68)$	$0.49_{\pm 0.00}$
NegGrad+	6.89	$99.96_{\pm 0.01}(1.16)$	$92.29_{\pm 0.35}(1.11)$	$99.95_{\pm 0.05}(8.62)$	$93.39_{\pm 0.08}(16.68)$	$0.49_{\pm 0.01}$
SCRUB	7.08	$99.97_{\pm 0.01}(1.17)$	$92.62_{\pm 0.49}(1.44)$	$100.00_{\pm 0.00}(8.67)$	$93.75_{\pm 0.28}(17.04)$	$0.58_{\pm 0.00}$
BadT	4.40	$99.57_{\pm 0.08}(0.77)$	$92.02_{\pm 0.31}(0.84)$	$99.55_{\pm 0.05}(8.22)$	$84.50_{\pm 0.89}(7.79)$	$0.19_{\pm 0.01}$
UNSIR	4.05	$99.25_{\pm 0.16}(0.45)$	**$91.41_{\pm 0.36}(0.23)$**	$99.32_{\pm 0.12}(7.99)$	$84.23_{\pm 1.43}(7.52)$	**$0.11_{\pm 0.00}$**
SalUn	3.17	$99.97_{\pm 0.01}(1.17)$	$92.48_{\pm 0.31}(1.30)$	$99.94_{\pm 0.02}(8.61)$	**$78.31_{\pm 0.19}(1.60)$**	$1.24_{\pm 0.04}$
CNS	7.28	$99.97_{\pm 0.01}(1.17)$	$92.59_{\pm 0.52}(1.41)$	$99.97_{\pm 0.03}(8.64)$	$94.63_{\pm 0.27}(17.92)$	$4.00_{\pm 0.06}$
LoTUS	4.11	**$99.11_{\pm 0.18}(0.31)$**	$91.57_{\pm 1.00}(0.39)$	$99.39_{\pm 0.08}(8.06)$	$84.37_{\pm 1.11}(7.66)$	$0.30_{\pm 0.00}$
CUP	**2.53**	$99.80_{\pm 0.02}(1.00)$	$91.62_{\pm 0.73}(0.44)$	**$93.78_{\pm 0.71}(2.45)$**	$82.95_{\pm 0.79}(6.24)$	$1.66_{\pm 0.14}$
			20%			
Retrained	0.00	$99.98_{\pm 0.01}(0.00)$	$91.73_{\pm 0.25}(0.00)$	$91.68_{\pm 0.18}(0.00)$	$77.37_{\pm 0.37}(0.00)$	$5.37_{\pm 0.01}$
Fine-tuning	6.17	**$99.98_{\pm 0.02}(0.00)$**	$92.53_{\pm 0.59}(0.80)$	$99.95_{\pm 0.02}(8.27)$	$92.98_{\pm 0.22}(15.61)$	$0.38_{\pm 0.00}$
RandLabel	2.77	$99.93_{\pm 0.01}(0.05)$	$92.15_{\pm 0.51}(0.42)$	$99.56_{\pm 0.04}(7.88)$	$74.64_{\pm 2.30}(2.73)$	$0.48_{\pm 0.00}$
NegGrad+	6.38	$99.97_{\pm 0.01}(0.01)$	$92.49_{\pm 0.47}(0.76)$	$99.99_{\pm 0.00}(8.31)$	$93.83_{\pm 0.29}(16.46)$	$0.48_{\pm 0.00}$
SCRUB	6.44	**$99.98_{\pm 0.01}(0.00)$**	$92.62_{\pm 0.49}(0.89)$	$99.99_{\pm 0.01}(8.31)$	$93.95_{\pm 0.58}(16.58)$	$0.58_{\pm 0.00}$
BadT	3.87	$99.26_{\pm 0.05}(0.72)$	**$91.82_{\pm 0.62}(0.09)$**	$99.23_{\pm 0.08}(7.55)$	$84.48_{\pm 0.49}(7.11)$	$0.23_{\pm 0.00}$
UNSIR	4.03	$99.42_{\pm 0.12}(0.56)$	$91.08_{\pm 0.67}(0.65)$	$99.46_{\pm 0.20}(7.78)$	$84.50_{\pm 1.36}(7.13)$	**$0.14_{\pm 0.01}$**
SalUn	2.75	$99.97_{\pm 0.01}(0.01)$	$92.51_{\pm 0.45}(0.78)$	$99.92_{\pm 0.02}(8.24)$	**$75.40_{\pm 0.90}(1.97)$**	$1.67_{\pm 0.02}$
CNS	6.59	**$99.98_{\pm 0.00}(0.00)$**	$92.53_{\pm 0.48}(0.80)$	$99.98_{\pm 0.01}(8.30)$	$94.65_{\pm 0.14}(17.28)$	$3.73_{\pm 0.01}$
LoTUS	5.15	$99.01_{\pm 0.12}(0.97)$	$91.25_{\pm 0.43}(0.48)$	$99.41_{\pm 0.12}(7.73)$	$88.79_{\pm 0.60}(11.42)$	$0.35_{\pm 0.00}$
CUP	**1.93**	$99.91_{\pm 0.01}(0.07)$	$91.18_{\pm 0.20}(0.55)$	**$93.97_{\pm 0.54}(2.29)$**	$82.20_{\pm 0.56}(4.83)$	$2.75_{\pm 0.49}$

E.2 DATA-FREE METHODS

Table 18: Data-free-method results for **ViT-Base & CIFAR-100** (untraining 10% of the training data).

Method	AG ↓	RA(Δ ↓)	TA(Δ ↓)	FA(Δ ↓)	MIA(Δ ↓)	Time ↓
Original	2.04	$96.10_{\pm 0.03}$(0.14)	$91.44_{\pm 0.12}$(0.20)	$96.64_{\pm 0.10}$(4.75)	$74.82_{\pm 0.25}$(3.08)	$121.97_{\pm 2.34}$
Retrained	0.00	$95.96_{\pm 0.06}$(0.00)	$91.24_{\pm 0.17}$(0.00)	$91.89_{\pm 0.11}$(0.00)	$71.74_{\pm 0.31}$(0.00)	$112.70_{\pm 0.08}$
Boundary	2.07	$96.09_{\pm 0.02}$(0.13)	$91.44_{\pm 0.12}$(0.20)	$96.68_{\pm 0.16}$(4.79)	$74.89_{\pm 0.12}$(3.15)	$0.98_{\pm 0.01}$
SSD	1.99	**$96.04_{\pm 0.05}$(0.08)**	**$91.36_{\pm 0.07}$(0.12)**	$96.60_{\pm 0.09}$(4.71)	$74.79_{\pm 0.09}$(3.05)	$4.58_{\pm 0.02}$
LFSSD	2.08	$96.11_{\pm 0.03}$(0.15)	$91.44_{\pm 0.12}$(0.20)	$96.71_{\pm 0.09}$(4.82)	$74.87_{\pm 0.19}$(3.13)	$4.57_{\pm 0.01}$
JiT	71.72	$1.00_{\pm 0.00}$(94.96)	$1.00_{\pm 0.02}$(90.24)	$1.00_{\pm 0.00}$(90.89)	$60.93_{\pm 28.76}$(10.81)	**$0.42_{\pm 0.01}$**
SALSA	19.27	$74.77_{\pm 7.99}$(21.19)	$73.54_{\pm 8.16}$(17.70)	$74.78_{\pm 7.89}$(17.11)	$50.66_{\pm 0.76}$(21.08)	$0.61_{\pm 0.01}$
CUP-DF	**1.69**	$93.56_{\pm 0.35}$(2.40)	$90.06_{\pm 0.22}$(1.18)	**$90.58_{\pm 0.45}$(1.31)**	**$73.63_{\pm 0.81}$(1.89)**	$4.81_{\pm 0.80}$

Table 19: Data-free-method results for **ViT-Base & TinyImageNet** (untraining 1%, 10%, 20% of the training data).

Method	AG ↓	RA(Δ ↓)	TA(Δ ↓)	FA(Δ ↓)	MIA(Δ ↓)	Time ↓
Original	2.57	$96.45_{\pm 0.04}$(0.08)	$90.22_{\pm 0.12}$(0.21)	$96.38_{\pm 0.06}$(6.31)	$79.43_{\pm 0.15}$(3.66)	$258.46_{\pm 0.52}$
				1%		
Retrained	0.00	$96.34_{\pm 0.01}$(0.00)	$90.69_{\pm 0.07}$(0.00)	$89.74_{\pm 0.06}$(0.00)	$75.95_{\pm 0.08}$(0.00)	$223.83_{\pm 0.08}$
Boundary	2.62	$96.45_{\pm 0.04}$(0.11)	$89.99_{\pm 0.95}$(0.70)	$96.65_{\pm 0.21}$(6.91)	$78.70_{\pm 0.28}$(2.75)	$0.19_{\pm 0.00}$
SSD	2.57	**$96.36_{\pm 0.00}$(0.02)**	$89.75_{\pm 0.74}$(0.94)	$96.20_{\pm 0.17}$(6.46)	$78.83_{\pm 0.81}$(2.88)	$8.39_{\pm 0.12}$
LFSSD	2.55	$96.46_{\pm 0.02}$(0.12)	**$90.00_{\pm 0.79}$(0.69)**	$96.58_{\pm 0.34}$(6.84)	$78.50_{\pm 0.58}$(2.55)	$8.41_{\pm 0.09}$
JiT	68.00	$15.55_{\pm 2.14}$(80.79)	$15.00_{\pm 1.87}$(75.69)	$14.30_{\pm 2.69}$(75.44)	$35.85_{\pm 1.20}$(40.10)	$0.84_{\pm 0.03}$
SALSA	28.36	$60.67_{\pm 16.02}$(35.67)	$60.06_{\pm 15.53}$(30.63)	$66.95_{\pm 12.80}$(22.79)	$51.60_{\pm 4.95}$(24.35)	**$0.12_{\pm 0.01}$**
CUP-DF	**1.88**	$95.35_{\pm 0.00}$(0.99)	$89.59_{\pm 0.92}$(1.10)	**$86.20_{\pm 0.28}$(3.54)**	**$77.85_{\pm 0.07}$(1.90)**	$0.90_{\pm 0.01}$
				10%		
Retrained	0.00	$96.37_{\pm 0.05}$(0.00)	$90.01_{\pm 0.05}$(0.00)	$90.07_{\pm 0.04}$(0.00)	$75.77_{\pm 0.06}$(0.00)	$228.90_{\pm 6.49}$
Boundary	2.56	**$96.43_{\pm 0.05}$(0.06)**	$90.23_{\pm 0.79}$(0.22)	$96.50_{\pm 0.10}$(6.43)	$79.31_{\pm 0.30}$(3.54)	$1.92_{\pm 0.01}$
SSD	2.55	**$96.43_{\pm 0.06}$(0.06)**	**$90.19_{\pm 0.83}$(0.18)**	$96.52_{\pm 0.19}$(6.45)	$79.26_{\pm 0.12}$(3.49)	$9.07_{\pm 0.02}$
LFSSD	2.55	**$96.43_{\pm 0.05}$(0.06)**	$90.23_{\pm 0.79}$(0.22)	$96.54_{\pm 0.15}$(6.47)	$79.22_{\pm 0.11}$(3.45)	$9.12_{\pm 0.11}$
JiT	76.34	$0.50_{\pm 0.00}$(95.87)	$0.54_{\pm 0.07}$(89.47)	$0.50_{\pm 0.00}$(89.57)	$45.30_{\pm 50.22}$(30.47)	**$0.84_{\pm 0.00}$**
SALSA	14.71	$81.25_{\pm 0.90}$(15.12)	$79.41_{\pm 0.09}$(10.60)	$81.33_{\pm 0.78}$(8.74)	$51.39_{\pm 0.86}$(24.38)	$1.22_{\pm 0.02}$
CUP-DF	**0.91**	$94.84_{\pm 1.05}$(1.53)	$89.19_{\pm 0.54}$(0.82)	**$90.10_{\pm 1.69}$(0.03)**	**$77.03_{\pm 0.42}$(1.26)**	$8.47_{\pm 1.97}$
				20%		
Retrained	0.00	$96.37_{\pm 0.05}$(0.00)	$90.31_{\pm 0.67}$(0.00)	$89.77_{\pm 0.08}$(0.00)	$75.89_{\pm 0.13}$(0.00)	$227.89_{\pm 1.63}$
Boundary	2.47	$96.45_{\pm 0.02}$(0.08)	**$90.23_{\pm 0.79}$(0.08)**	$96.50_{\pm 0.08}$(6.73)	$78.89_{\pm 0.13}$(3.00)	$3.87_{\pm 0.01}$
SSD	2.46	**$96.43_{\pm 0.03}$(0.06)**	$90.21_{\pm 0.81}$(0.10)	$96.48_{\pm 0.04}$(6.71)	$78.85_{\pm 0.10}$(2.96)	$10.18_{\pm 0.25}$
LFSSD	2.46	**$96.43_{\pm 0.05}$(0.06)**	**$90.23_{\pm 0.79}$(0.08)**	$96.48_{\pm 0.06}$(6.71)	$78.87_{\pm 0.09}$(2.98)	$10.05_{\pm 0.04}$
JiT	81.01	$0.50_{\pm 0.00}$(95.87)	$0.41_{\pm 0.06}$(89.90)	$0.50_{\pm 0.00}$(89.27)	$26.89_{\pm 23.44}$(49.00)	$17.13_{\pm 0.03}$
SALSA	22.15	$71.83_{\pm 3.74}$(24.54)	$70.43_{\pm 3.75}$(19.88)	$71.77_{\pm 3.76}$(18.00)	$49.73_{\pm 1.49}$(26.16)	**$2.50_{\pm 0.01}$**
CUP-DF	**1.17**	$94.54_{\pm 0.63}$(1.83)	$89.18_{\pm 0.58}$(1.13)	**$91.07_{\pm 1.43}$(1.30)**	**$76.33_{\pm 0.19}$(0.44)**	$19.00_{\pm 7.89}$

Table 20: Data-free-method results for **ViT-Large & TinyImageNet** (untraining 10% of the training data).

Method	AG ↓	RA(Δ ↓)	TA(Δ ↓)	FA(Δ ↓)	MIA(Δ ↓)	Time ↓
Original	3.53	$99.35_{\pm 0.01}$(0.03)	$91.50_{\pm 0.47}$(0.23)	$99.48_{\pm 0.01}$(7.46)	$86.19_{\pm 0.29}$(6.39)	$383.64_{\pm 0.07}$
Retrained	0.00	$99.32_{\pm 0.01}$(0.00)	$91.27_{\pm 0.07}$(0.00)	$92.02_{\pm 0.04}$(0.00)	$79.80_{\pm 0.06}$(0.00)	$345.25_{\pm 0.03}$
Boundary	70.26	$11.64_{\pm 0.74}$(87.68)	$11.18_{\pm 0.31}$(80.09)	$11.30_{\pm 0.70}$(80.72)	$47.27_{\pm 2.49}$(32.53)	$31.08_{\pm 0.01}$
SSD	3.48	**$99.35_{\pm 0.02}$(0.03)**	**$91.47_{\pm 0.49}$(0.20)**	$99.44_{\pm 0.02}$(7.42)	$86.06_{\pm 0.21}$(6.26)	$42.09_{\pm 0.02}$
LFSSD	3.48	**$99.35_{\pm 0.02}$(0.03)**	**$91.47_{\pm 0.49}$(0.20)**	$99.44_{\pm 0.02}$(7.42)	$86.06_{\pm 0.21}$(6.26)	$42.09_{\pm 0.02}$
JiT	53.62	$42.07_{\pm 12.53}$(57.25)	$41.41_{\pm 12.35}$(49.86)	$42.05_{\pm 12.66}$(49.97)	$22.38_{\pm 14.61}$(57.42)	$12.25_{\pm 0.01}$
SALSA	32.36	$63.19_{\pm 13.03}$(36.13)	$62.07_{\pm 12.83}$(29.20)	$64.14_{\pm 13.24}$(27.88)	$43.58_{\pm 3.29}$(36.22)	**$1.89_{\pm 0.01}$**
CUP-DF	**2.44**	$96.28_{\pm 0.62}$(3.04)	$88.96_{\pm 0.17}$(2.31)	**$92.72_{\pm 0.96}$(0.70)**	**$83.51_{\pm 0.64}$(3.71)**	$20.03_{\pm 2.60}$

Table 21: Data-free-method results for **ResNet-18 & CIFAR-10** (untraining 1%, 10%, 20% of the training data).

Method	AG ↓	RA(Δ ↓)	TA(Δ ↓)	FA(Δ ↓)	MIA(Δ ↓)	Time ↓
Original	7.15	$99.98_{\pm 0.01}(1.18)$	$92.62_{\pm 0.49}(1.44)$	$99.99_{\pm 0.01}(8.66)$	$94.03_{\pm 0.39}(17.32)$	$6.62_{\pm 0.12}$
			1%			
Retrained	0.00	$99.98_{\pm 0.01}(0.00)$	$92.64_{\pm 0.32}(0.00)$	$93.93_{\pm 1.29}(0.00)$	$78.33_{\pm 0.61}(0.00)$	$6.57_{\pm 0.00}$
Boundary	5.27	**$99.98_{\pm 0.01}(0.00)$**	$92.62_{\pm 0.49}(0.02)$	$99.93_{\pm 0.12}(6.00)$	$93.40_{\pm 0.72}(15.07)$	**$0.00_{\pm 0.00}$**
SSD	5.62	**$99.98_{\pm 0.01}(0.00)$**	**$92.65_{\pm 0.50}(0.01)$**	$100.00_{\pm 0.00}(6.07)$	$94.73_{\pm 0.42}(16.40)$	$0.11_{\pm 0.00}$
LFSSD	5.64	**$99.98_{\pm 0.01}(0.00)$**	$92.62_{\pm 0.49}(0.02)$	$100.00_{\pm 0.00}(6.07)$	$94.80_{\pm 0.35}(16.47)$	$0.10_{\pm 0.01}$
JiT	49.78	$39.15_{\pm 11.47}(60.83)$	$39.47_{\pm 10.85}(53.17)$	$40.27_{\pm 12.01}(53.66)$	$46.87_{\pm 18.85}(31.46)$	$0.01_{\pm 0.00}$
SALSA	17.23	$75.48_{\pm 42.44}(24.50)$	$70.54_{\pm 38.66}(22.10)$	$75.27_{\pm 42.84}(18.66)$	**$82.00_{\pm 18.01}(3.67)$**	$0.01_{\pm 0.00}$
CUP-DF	**3.43**	$96.97_{\pm 0.86}(3.01)$	$89.26_{\pm 0.40}(3.38)$	**$94.80_{\pm 1.74}(0.87)$**	$84.80_{\pm 2.82}(6.47)$	$0.12_{\pm 0.02}$
			10%			
Retrained	0.00	$98.80_{\pm 2.04}(0.00)$	$91.18_{\pm 2.25}(0.00)$	$91.33_{\pm 2.39}(0.00)$	$76.71_{\pm 3.10}(0.00)$	$5.39_{\pm 1.20}$
Boundary	7.09	$99.98_{\pm 0.01}(1.18)$	$92.62_{\pm 0.49}(1.44)$	$99.99_{\pm 0.01}(8.66)$	$93.77_{\pm 0.35}(17.06)$	**$0.01_{\pm 0.00}$**
SSD	7.06	$99.96_{\pm 0.03}(1.16)$	$92.42_{\pm 0.41}(1.24)$	$99.97_{\pm 0.02}(8.64)$	$93.92_{\pm 0.26}(17.21)$	$0.12_{\pm 0.01}$
LFSSD	7.16	$99.98_{\pm 0.01}(1.18)$	$92.62_{\pm 0.49}(1.44)$	$99.99_{\pm 0.01}(8.66)$	$94.07_{\pm 0.17}(17.36)$	$0.11_{\pm 0.01}$
JiT	82.00	$10.00_{\pm 0.00}(88.80)$	$10.00_{\pm 0.00}(81.18)$	$10.00_{\pm 0.00}(81.33)$	$0.00_{\pm 0.00}(76.71)$	$0.15_{\pm 0.04}$
SALSA	5.43	**$99.14_{\pm 0.60}(0.34)$**	**$90.50_{\pm 1.27}(0.68)$**	$99.10_{\pm 0.62}(7.77)$	$63.80_{\pm 3.03}(12.91)$	$0.02_{\pm 0.01}$
CUP-DF	**4.18**	$96.70_{\pm 1.94}(2.10)$	$88.89_{\pm 1.45}(2.29)$	**$95.76_{\pm 2.12}(4.43)$**	**$84.61_{\pm 0.68}(7.90)$**	$0.73_{\pm 0.37}$
			20%			
Retrained	0.00	$99.98_{\pm 0.01}(0.00)$	$91.73_{\pm 0.25}(0.00)$	$91.68_{\pm 0.18}(0.00)$	$77.37_{\pm 0.37}(0.00)$	$5.37_{\pm 0.01}$
Boundary	6.39	**$99.98_{\pm 0.01}(0.00)$**	$92.62_{\pm 0.49}(0.89)$	$99.99_{\pm 0.01}(8.31)$	$93.73_{\pm 0.37}(16.36)$	**$0.02_{\pm 0.00}$**
SSD	5.91	$99.73_{\pm 0.22}(0.25)$	**$91.80_{\pm 1.08}(0.07)$**	$99.72_{\pm 0.22}(8.04)$	$92.66_{\pm 0.66}(15.29)$	$0.11_{\pm 0.01}$
LFSSD	6.40	**$99.98_{\pm 0.00}(0.00)$**	$92.62_{\pm 0.49}(0.89)$	$99.97_{\pm 0.00}(8.29)$	$93.81_{\pm 0.41}(16.44)$	$0.11_{\pm 0.00}$
JiT	40.98	$52.84_{\pm 46.93}(47.14)$	$48.40_{\pm 42.07}(43.33)$	$52.92_{\pm 47.01}(38.76)$	$42.68_{\pm 46.76}(34.69)$	$0.16_{\pm 0.08}$
SALSA	8.08	$99.34_{\pm 0.56}(0.64)$	$91.01_{\pm 0.87}(0.72)$	$99.40_{\pm 0.54}(7.72)$	$54.13_{\pm 6.09}(23.24)$	$0.03_{\pm 0.00}$
CUP-DF	**4.35**	$95.64_{\pm 0.43}(4.34)$	$87.47_{\pm 0.96}(4.26)$	**$95.00_{\pm 0.31}(3.32)$**	**$82.85_{\pm 1.33}(5.48)$**	$0.50_{\pm 0.13}$

E.3 Comments on the Results

E.4 White-Box MIA

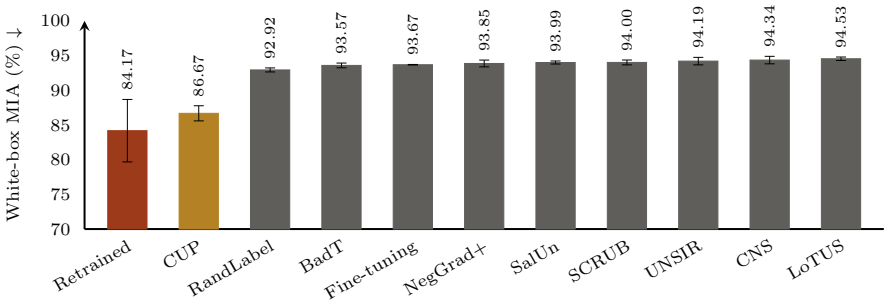


Figure 9: *White-box MIA* (Nasr et al., 2019) success (% of forget samples classified as members) on **ResNet-18 & CIFAR-10** (untraining 10% of the training data); lower is better, down to retraining. Bars show $\mu \pm \sigma$ over three trials.

E.5 Synthetic Calibration Sets

Table 22 reports detailed results for the calibration sets of Table 4: held-out real data (CIFAR-10) and synthetic-only data (CIFAKE (Bird & Lotfi, 2024)) of n_{cal}=5k and n_{cal}=50k samples.

Table 22: Results for **ResNet-18 & CIFAR-10** (untraining 10% of the training data) with different calibration sets.

Method	AG ↓	RA(Δ ↓)	TA(Δ ↓)	FA(Δ ↓)	MIA(Δ ↓)	Time ↓
Original	7.15	$99.98_{\pm 0.01}$(1.18)	$92.62_{\pm 0.49}$(1.44)	$99.99_{\pm 0.01}$(8.66)	$94.03_{\pm 0.39}$(17.32)	$6.62_{\pm 0.12}$
Retrained	0.00	$98.80_{\pm 2.04}$(0.00)	$91.18_{\pm 2.25}$(0.00)	$91.33_{\pm 2.39}$(0.00)	$76.71_{\pm 3.10}$(0.00)	$5.39_{\pm 1.20}$
CUP (n_{cal}=5k, CIFAR-10)	2.53	$99.80_{\pm 0.02}$(1.00)	$91.62_{\pm 0.73}$(0.44)	$93.78_{\pm 0.71}$(2.45)	$82.95_{\pm 0.79}$(6.24)	$1.66_{\pm 0.14}$
CUP (n_{cal}=5k, CIFAKE)	1.46	$99.77_{\pm 0.01}$(0.97)	$91.08_{\pm 1.07}$(0.10)	$88.47_{\pm 0.33}$(2.86)	$78.63_{\pm 0.80}$(1.92)	$1.67_{\pm 0.05}$
CUP (n_{cal}=50k, CIFAKE)	1.18	$99.74_{\pm 0.08}$(0.94)	$91.47_{\pm 0.75}$(0.29)	$88.23_{\pm 0.55}$(3.10)	$77.10_{\pm 1.07}$(0.39)	$1.59_{\pm 0.33}$

E.6 Efficiency

Table 23: How many times faster CUP and CUP-DF are than retraining (untraining 10% of the training data).

Benchmarks	Efficiency Ratio ↑	
ViT-L & TImNet	10.40	17.24
ViT-B & TImNet	10.69	27.02
ViT-B & C-100	9.04	23.43
RN-18 & C-10	3.25	7.38

CUP(-DF) is faster than retraining in all experiments (Table 23). Efficiency follows from monitoring distributions in one dimension, the non-parametric quantile mapping, and the closed form of δ_{emp}, which avoid neural approximation. On ViTs, CUP ranks among the three fastest standard baselines when untraining 1% and 10% of D_{tr}. On ResNet-18 it is the second fastest at 1%, but for 10% and 20% it is among the slowest, because intrinsic stopping may run for more epochs, while other baselines run for a fixed number of 1, 5, or 10 epochs. This is the cost of letting the stopping criterion decide when untraining is done; a reasonable trade once untraining is already faster than retraining. A larger learning rate seems to trade smoothness, i.e., better adaptation, for speed (Fig. 7).

E.7 More Figures

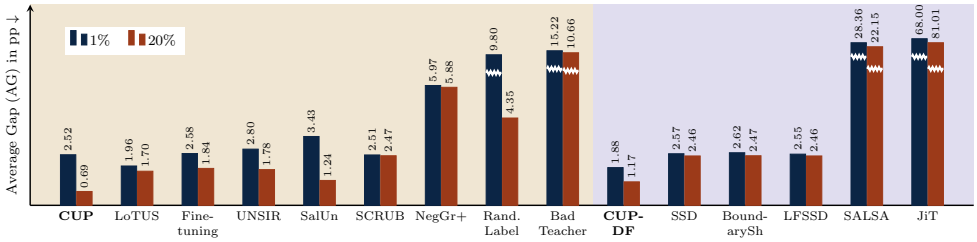


Figure 10: *Ablation on forget-set sizes (% of the training set).* CUP (and CUP-DF) preserve best scores in balancing privacy and utility on the ViT-Base & TinyImageNet.

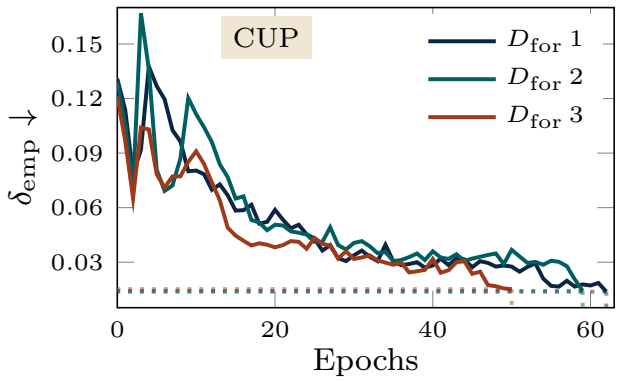


Figure 11: *Adaptation across three random forget sets.* Untraining 10% of CIFAR-10 from ResNet-18. Intrinsic stopping halts CUP at a different epoch per forget set, yet it reaches comparable δ_{emp}; likewise CUP-DF (Fig. 6).