TACKLING FAKE FORGETTING THROUGH UNCERTAINTY QUANTIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Machine unlearning seeks to remove the influence of specified data from a trained model. While metrics such as unlearning accuracy (UA) and membership inference attack (MIA) provide baselines for assessing unlearning performance, they fall short of evaluating the reliability of forgetting. In this paper, we find that the data points misclassified by UA and MIA still have their ground truth labels included in the prediction set from the uncertainty quantification perspective, which raises the issue of fake forgetting. To address this issue, we propose two novel metrics inspired by conformal prediction that provide a more reliable evaluation of forgetting quality. Building on these insights, we further propose an unlearning framework that integrates conformal prediction into the Carlini & Wagner adversarial attack loss, which can effectively push the ground truth label out of the conformal prediction set. Through extensive experiments on image classification tasks, we demonstrate both the effectiveness of our proposed metrics and the superiority of our framework. Code is available at https://anonymous.4open.science/r/MUCP-60E4.

1 INTRODUCTION

Machine unlearning has become essential for data privacy, particularly under regulations such as the GDPR Bourtoule et al. (2021), which grant individuals the right to have their data erased. This creates a strong demand for methods that enable models to behave as if certain data were never used during training. Beyond privacy, unlearning also serves as a tool for mitigating harmful biases and stereotypes in models. Existing post hoc machine unlearning methods can be categorized into training-based Graves et al. (2021); Tarun et al. (2023); Thudi et al. (2022); Warnecke et al. (2021) and training-free Foster et al. (2024); Golatkar et al. (2021; 2020); Guo et al. (2019); Nguyen et al. (2020); Sekhari et al. (2021) approaches, depending on whether they require any model training steps during the unlearning process Foster et al. (2024).

To measure the forgetting quality and predictive performance of an unlearning model, several unlearning metrics have been proposed Hayes et al. (2025); Cao & Yang (2015); Chen et al. (2021); Kashef (2021); Shokri et al. (2017). However, existing unlearning metrics, such as unlearning accuracy (**UA**) and membership inference attack (**MIA**), fall short in fully evaluating forgetting reliability — these metrics primarily focus on whether models can predict forget data accurately **without sufficiently considering uncertainty and confidence level**. In a nutshell, misclassifying the forget data does not mean that the model has completely forgotten it to some extent.

To verify this view, conformal prediction Lei & Wasserman (2014); Papadopoulos et al. (2002) as an uncertainty quantification technique, is applied in our work to recover the misclassified data in UA and MIA. Through extensive experiments, we find that although the model misclassifies part of the forget data from the UA and MIA perspectives, **over 50% of these misclassified data instances still appear in the conformal prediction set and can be easily recovered, which exposes a fake forgetting issue**. As shown in Figure 1, the important features of prediction visualize this fake forgetting issue by using Grad-CAM Selvaraju et al. (2017). Despite the Finetune method misclassifying the forget data, the Grad-CAM maps still focus heavily on the important features of the object itself since the true label is included in the prediction set. In contrast, when our unlearning method removes the true label from the set, activation regions shift significantly away from the object's key features. This confirms that forgetting quality improves if the true label can be excluded from the prediction set.

Based on the above insights, we design two novel metrics **CR** and **MIACR** that more effectively capture the uncertainty and robustness of unlearning performance inspired by conformal prediction to tackle the fake forgetting issue. Additionally, motivated by conformal prediction insights about fake forgetting and Carlini & Wagner (C&W) attack loss Carlini & Wagner (2017), we propose a general unlearning framework, which can improve existing training-based unlearning methods and promote reliable forgetting. Grad-CAM maps of our method in Figure 1 reveal that **once the true label no longer falls within the conformal prediction set, the activation regions shift significantly**. To sum up, our contributions are as follows:

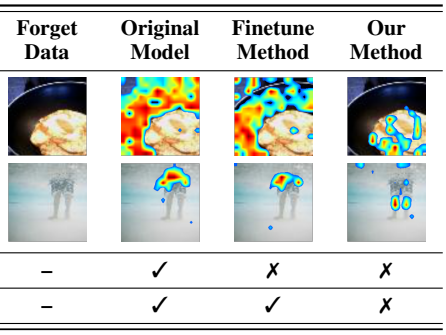


Figure 1: Grad-CAM maps of one original model and two corresponding unlearning models in Tiny ImageNet with ViT. The **Classification** row indicates whether the model correctly predicts the image's true label, while the **In Set** row represents whether the true label is included in the prediction set.

- Our analysis reveals that conformal prediction can recover a substantial portion of data previously classified as forgotten by existing unlearning metrics. This fake forgetting issue underscores critical limitations in existing unlearning evaluation methodologies.
- We design two novel metrics to address the limitations motivated by conformal prediction.
- We propose an unlearning framework motivated by conformal prediction and C&W loss, enhancing existing training-based unlearning methods over both existing and our metrics.

2 Enhancing Metrics for Machine Unlearning Based on Conformal Prediction

2.1 Preliminaries and Notations

Machine Unlearning. In our work, we focus on the image classification task, which is widely used in prior literature Shen et al. (2024); Zhao et al. (2024). Two forgetting scenarios are mainly considered in this work: (i) *random data forgetting* focuses on randomly forgetting specific data instances within the training data, and (ii) *class-wise forgetting* aims to remove all data information associated with an entire class. We also show the results of the subclass-wise forgetting scenario in Table 10 in Appendix. Let $\mathcal{D}_{train}$ denote the original training data used to obtain an original model $\boldsymbol{\theta}_o$. We split the whole training data $\mathcal{D}_{train}$ into two subsets, forget data $\mathcal{D}_f$ and retain data $\mathcal{D}_r = \mathcal{D}_{train} \setminus \mathcal{D}_f$. Let $\mathcal{D}_{test}$ represent test data. θ_u denotes the model after the unlearning process.

Conformal Prediction. Conformal prediction is proposed to quantify uncertainty, providing prediction sets that contain the ground truth label with a theoretically guaranteed probability Angelopoulos & Bates (2021). Among the various types of conformal prediction, this work mainly focuses on split conformal prediction (SCP)[1] since it is the most straightforward and easy-to-implement approach. We also report results of other conformal prediction techniques in Appendix F. To construct a conformal prediction set, SCP involves four steps on the unlearning model:

1. *Calibration Data*. SCP first chooses unseen data as calibration data, which must be held out from both the training and test sets to ensure independence.
2. *Non-conformity Score*. In our work, we follow the conventional choice and set the non-conformity score as
$$S(\boldsymbol{x}, y_i) = 1 - p_i(\boldsymbol{x}), \tag{1}$$
where $p_i(\boldsymbol{x})$ represents the probability of different class y_i.

[1]Note that while the goal is to remove the influence of the forget data so that it behaves similarly to the calibration data, the exchangeability property may not always hold in machine unlearning settings. Here, we are directly leveraging the concept of conformal prediction to evaluate machine unlearning performance.

3. *Quantile Computation*. Given a target miscoverage rate $\alpha \in [0, 1]$, SCP obtains threshold $\hat{q}$ by taking the $1 - \alpha$ quantile of the non-conformity score of the ground truth labels y_t on the calibration data $(\boldsymbol{x}, y_t) \in \mathcal{D}_c$,

$$\hat{q} = \text{Quantile}_{1-\alpha}\big(S(\boldsymbol{x}, y_t)\big). \tag{2}$$

4. *Prediction Set*. For the data point $\boldsymbol{x}$ that needs to be tested, labels with non-conformity scores lower than the threshold $\hat{q}$ are selected for the final prediction set:

$$\mathbb{C}(\boldsymbol{x}) = \{y_i : S(\boldsymbol{x}, y_i) \leq \hat{q}\}, \tag{3}$$

2.2 Identifying fake forgetting in Existing Unlearning Metrics

In this section, we show that a conformal prediction–based recovery technique can reconstruct the true label with high probability even when one forget data point is misclassified. This highlights a critical blind spot in existing UA and MIA metrics from the perspective of uncertainty quantification. The first key question we pose is as follows:

(Q1) *Can we recover the data that is identified as forgotten by the metrics UA and MIA?*

If the ground truth of forget data falls within the conformal prediction set, we consider the recovery successful. Thus, **fake forgetting is defined as the scenario where a data point identified as forgotten by model prediction can be recovered by conformal prediction**.

To substantiate our claim, we first apply metrics: unlearning accuracy (**UA**, i.e., $1-$ the accuracy on forget data), retain accuracy (**RA**, i.e., accuracy on retain data), test accuracy (**TA**, i.e., accuracy on test data), and membership inference attack (**MIA**). See Appendix C for MIA implementation details. We evaluate 3 classic unlearning methods, Retrain (**RT**), Finetune (**FT**) Warnecke et al. (2021), and Random Label (**RL**) Graves et al. (2021). See Appendix A for a detailed introduction to the baselines. The results are trained on CIFAR-10 with ResNet-18 in a random data forgetting scenario. In Table 1, the UA and MIA results suggest that the models fail to correctly classify part of the forget data and identify membership. However, can higher UA and lower MIA fully guarantee that these forget data points do not appear in any form within the model's predictions?

Table 1: Unlearning performance measured by existing metrics across RT, FT and RL methods. All values in percent (%). The sign ↑ (↓) represents the greater (smaller) is better.

Methods	10% Random Forgetting				50% Random Forgetting			
	UA ↑	RA ↑	TA ↑	MIA ↓	UA ↑	RA ↑	TA ↑	MIA ↓
RT	8.62	99.69	91.83	86.92	10.98	99.80	89.16	82.79
FT	3.84	98.14	91.57	92.00	2.59	99.08	91.77	92.92
RL	7.55	97.41	90.60	74.21	10.48	93.91	85.78	61.15

We employ conformal prediction to investigate whether we can recover forget data's ground truth, specifically, whether the ground truth labels still appear within the conformal prediction sets. The confidence level and calibration set size are set to 95% and 2000 respectively. In Table 2, we count the number of data points that are identified as truly forgotten by UA and MIA (marked as *mis-label*) and count how many of these *mis-label* points can still be recovered (marked as *in-set*). The **results of UA** reveal that even though the model misclassifies part of the forget data, on average 54.6% of these misclassified data instances are still recovered by conformal prediction. Even for the RT baseline, UA does not reliably assess whether a data point has truly been forgotten, since 30.6% of UA misclassified data points can still be recovered by conformal prediction. This finding demonstrates that a high UA does not mean the model has truly forgotten the data, and thus relying solely on UA to evaluate the forgetting quality is fragile. A similar phenomenon occurs on **results of MIA**. In MIA, '0' indicates a data point is forgotten, while '1' means it is still identified as a training member. The *mis-label* column of MIA refers to the number of data points that are predicted as '0'. The *in-set* here refers to the number of *mis-label* data points whose conformal prediction set still includes '1'. Thus, the *recover ratio* indicates that, although the MIA fails to identify an average of 18.33% of the forget data as training membership,

Table 2: Mis-label (mis-classification) count and in-set ratio of UA and MIA metrics for RT, FT and RL on **CIFAR-10** with **ResNet-18** under 10% and 50% **random data forgetting** scenarios. In all settings, over 30% of mis-label data remains within the conformal prediction set in both UA and MIA. More results of other unlearning methods can be found in Appendix D.

Methods	10% Random Forgetting			50% Random Forgetting		
	Mis-label ↑	In-set ↓	Ratio ↓	Mis-label ↑	In-set ↓	Ratio ↓
Mis-label and In-set Ratio of UA						
RT	431	132	30.6%	2,745	1,573	57.3%
FT	192	112	58.3%	647	431	66.6%
RL	380	173	45.5%	2,625	1,795	68.4%
Mis-label and In-set Ratio of MIA						
RT	654	209	32.0%	4,303	1,391	32.3%
FT	400	216	54.0%	1,769	813	46.0%
RL	1,289	1,011	78.4%	9,713	8,295	85.4%

conformal prediction can still recover 54.7% of these forget data within prediction sets. For more results of other unlearning methods, see Table 6 in AppendixD.1.

Overall, the high *recover ratio* observed in Tables 2 indicates that misclassified forget data cannot be considered truly forgotten, as their traces can be readily detected and recovered via conformal prediction from the perspective of uncertainty quantification. This encloses that **the fake forgetting issue arises when the true label of misclassified data falls within the conformal prediction set**.

2.3 Designing Metrics Motivated by Conformal Prediction

Based on the limitation of UA and MIA metrics shown in Section 2.2, it raises a question as follows:

(Q2) *Can we develop metrics to address the fake forgetting issue of UA and MIA?*

Thus, we propose enhanced UA and MIA metrics that draw intuition from conformal prediction.

2.3.1 Definition of New Metrics

Conformal Ratio (CR). To overcome the fake forgetting inherent in UA, we introduce a novel metric, CR, which incorporates both coverage and set size in conformal prediction to provide a more comprehensive evaluation. Before defining CR, we introduce Coverage and Set Size.

Given a dataset $\mathcal{D}$, the definition of **Coverage** is as follows:

$$\text{Coverage} := \frac{1}{|\mathcal{D}|} \sum_{(\boldsymbol{x}, y_t) \in \mathcal{D}} \mathbb{I}(y_t \in \mathbb{C}_m(\boldsymbol{x})), \tag{4}$$

where y_t is the true label of data point $\boldsymbol{x}$. Indicator function $\mathbb{I}(\cdot)$ returns 1 if the enclosed condition is true and 0 otherwise. Coverage reflects the probability that the true label falls within the prediction set $\mathbb{C}_m(\boldsymbol{x})$. For $\mathcal{D} = \mathcal{D}_f$, high coverage indicates that the model retains significant information about forget data, suggesting fake forgetting.

Given a dataset $\mathcal{D}$, **Set Size** is defined as follows:

$$\text{Set Size} := \frac{1}{|\mathcal{D}|} \sum_{(\boldsymbol{x}, y_t) \in \mathcal{D}} |\mathbb{C}_m(\boldsymbol{x})|, \tag{5}$$

where $\mathbb{C}_b(\boldsymbol{x})$ is the conformal prediction set in the multi-class classification task and $|\mathbb{C}_m(\boldsymbol{x})|$ denotes the set size of data point $\boldsymbol{x}$. When $y_t \in \mathbb{C}_m(\boldsymbol{x})$, a small set size indicates that fewer non-ground truth classes are included in the prediction set, reflecting stronger fake forgetting.

Based on Coverage and Set Size, we introduce the definition of **CR** for a dataset $\mathcal{D}$ as follows:

$$\text{CR} := \frac{\text{Coverage}}{\text{Set Size}} = \frac{\sum_{(\boldsymbol{x}, y_t) \in \mathcal{D}} \mathbb{I}(y_t \in \mathbb{C}_m(\boldsymbol{x}))}{\sum_{(\boldsymbol{x}, y_t) \in \mathcal{D}} |\mathbb{C}_m(\boldsymbol{x})|}. \tag{6}$$

CR balances the information captured by Coverage and Set Size. A lower CR value implies stronger forgetting. CR is inspired by conformal prediction, which is proposed to assess the model's behavior on new and unseen data, not on the training data. Thus, we emphasize that CR only measures forget data $\mathcal{D}_f$ and test data $\mathcal{D}_{test}$.

MIA Conformal Ratio (MIACR). MIACR is proposed to address the limitation of the existing MIA metric. Among three potential conformal prediction sets $\{0\}$, $\{1\}$, and $\{0, 1\}$, only set $\{0\}$ is an ideal case for MIA, because the presence of '1' represents that the data point can still be recognised as a training member. Therefore, we introduce a new metric **MIACR** as:

$$\text{MIACR} := \frac{1}{|\mathcal{D}_f|} \sum_{(\boldsymbol{x}, y_t) \in \mathcal{D}_f} \mathbb{I}(\mathbb{C}_b(\boldsymbol{x}) = \{0\}), \tag{7}$$

where $\mathbb{C}_b(\boldsymbol{x})$ is the conformal prediction set in the binary classification task. $\mathbb{C}_b(\boldsymbol{x}) = \{0\}$ denotes prediction set is exactly $\{0\}$. A higher MIACR score indicates a stronger forgetting. Under MIA, a data point is considered forgotten once the logit for label '0' exceeds that for label '1'. However, this criterion is often fragile. If the model's conformal prediction set for a forgetting data point still

includes both $\{0, 1\}$, it indicates that the model retains a level of uncertainty and has not completely purged the data's membership information. To address this, MIACR enforces a stricter rule, requiring that label '1' be entirely absent from the prediction set, providing a more rigorous assessment of membership status and forgetting quality.

Superiority of Our Metrics. Existing accuracy-based metrics UA and MIA suffer from a fake forgetting issue, since true labels of misclassified data points may still remain within the prediction set. In contrast, our metrics CR and MIACR address this issue by examining the entire conformal prediction set, providing a more reliable evaluation of forgetting quality. Besides evidence in Tables 2, Figures 7–10 in the Appendix also support this superiority of our metrics.

Evaluation Criteria of Our Metrics

We consider two different criteria[2]to measure unlearning performance with our metrics,
❶ **Gap to RT Criterion**: A lower gap to the RT method is better for both CR and MIACR metrics. The gap relative to RT is represented in blue text (•) in our result tables.
❷ **Limit-Based Criterion**: For the CR, a lower CR value of forget data $\mathcal{D}_f$ indicates stronger forgetting performance, while a higher CR value of $\mathcal{D}_{test}$ represents higher preserved model utility. For the MIACR, a higher MIACR value for $\mathcal{D}_f$ reflects better unlearning effectiveness.

2.3.2 Discussion of Confidence Level and Calibration Set Size

In conformal prediction, the confidence level $1 - \alpha$ (i.e., miscoverage rate α) and calibration set size are two factors. We next discuss the suitable settings for the confidence level and calibration set size, and the rationale behind them.

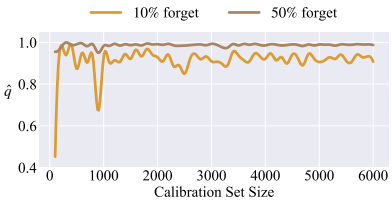


Figure 2: The stability of $\hat{q}$ in different calibration set sizes. When the calibration set size is greater than 2000, the fluctuations of $\hat{q}$ remain within a stable range.

Confidence Level $1-\alpha$. A smaller miscoverage rate α, i.e., a higher confidence level $1 - \alpha$, guarantees more reliable coverage. In the conformal prediction related works Angelopoulos & Bates (2021); Papadopoulos et al. (2002); Romano et al. (2020a); Tailor et al., $\boldsymbol{\alpha = 0.05}$ is widely adopted as a standard in most cases, reflecting its common use in statistical hypothesis testing to balance false positives and practical usability. Following prior work, we set $\alpha = 0.05$ by default, while also reporting results for higher values (0.10, 0.15, and 0.20) in Appendices D.3 and D.4 to account for scenarios where a more relaxed confidence level is needed. Unless otherwise noted, all analyses use the default $\alpha = 0.05$.

Calibration Set Size. A portion of the validation data is set aside as calibration data, ensuring it remains independent from both the training and test data. The calibration set must be sufficient to avoid abnormal $\hat{q}$ values caused by outliers from small samples, which can destabilize coverage estimates. Figure 2 illustrates the stability of $\hat{q}$ across varying calibration set sizes. The results are smoothed using a B-spline. We implement them on CIFAR-10 with ResNet-18 in 10% and 50% random data forgetting scenarios. The results show that for different settings using ResNet-18 on CIFAR-10, after the calibration set size is larger than 1000, abnormal $\hat{q}$ values do not occur anymore, and a stable threshold $\hat{q}$ can be obtained. Similarly, we analyze the calibration set size of the class-wise forgetting scenario and find that fewer calibration data points are required compared to random data forgetting. This is because the targeted class forgetting reduces the complexity of the distribution, unlike the broader variability introduced by random data forgetting.

3 Enhancing Machine Unlearning via Conformal Prediction

Based on the findings in Section 2.2, we observe that existing training-based unlearning methods are typically optimized with respect to loss functions that do not directly support the improvement of forgetting quality from our fake forgetting perspective. Specifically, **the optimization objectives of existing methods fail to ensure that the ground truth labels are sufficiently pushed out of the conformal prediction set**, which is key to overcoming fake forgetting. This raises a critical question:

[2]The appropriate evaluation criteria vary across unlearning application scenarios Kurmanji et al. (2023): criterion ❶ is particularly relevant for user privacy scenario, while criterion ❷ focuses on bias removal scenario.

(Q4) *Can we explore advanced unlearning techniques via conformal prediction to optimize the existing unlearning model's forgetting quality?*

Therefore, we propose a novel and general **c**onformal **p**rediction-based **u**nlearning framework (**CPU**) tailored for training-based unlearning methods, aimed at enhancing their forgetting quality. A key insight driving our framework is to overcome the issue exposed by fake forgetting. This emphasizes that the non-conformity scores of ground truth labels should be pushed beyond the conformal prediction threshold $\hat{q}$. Interestingly, this goal aligns naturally with the design of the C&W attack loss Carlini & Wagner (2017), which motivates our creative adaptation to the unlearning scenario.

Let us first apply the original C&W loss directly to the unlearning scenario, without yet incorporating conformal prediction. For the forget data $\mathcal{D}_f$, the goal of the unlearning loss is to decrease the model's confidence in the true labels of $\mathcal{D}_f$. Based on this, the C&W-inspired unlearning loss is defined as:

$$\mathcal{L}_{\text{cw}}(\boldsymbol{x}, y_t) = \max\{p_t(\boldsymbol{x}) - \max_{i \neq t}\{p_i(\boldsymbol{x})\},\ -\Delta\}, \tag{8}$$

where $(\boldsymbol{x}, y_t) \in \mathcal{D}_f$ and $\max\{\cdot\}$ is a maximum operator that selects the largest value from the set. $p_i(\boldsymbol{x})$ is the probability of class y_i, and $p_t(\boldsymbol{x})$ refers specifically to the probability assigned to the true label y_t. We denote $\max_{i \neq t}\{p_i(\boldsymbol{x})\}$ as the highest probability value of the non-ground truth classes. This loss $\mathcal{L}_{\text{cw}}$ maximizes the difference between the highest probability value for class y_i $(i \neq t)$ and the probability value for the true class y_t. It tries to decrease the probability of the true class y_t and further increase that of the class y_i with the highest probability. The margin parameter Δ controls the enforced margin between the true class and the strongest competing class. When the $\max_{i \neq t}\{p_i(\boldsymbol{x})\} - p_t(\boldsymbol{x}) < \Delta$, this loss encourages the model to decrease the true label's probability $p_t(\boldsymbol{x})$. Increasing the value of Δ further increase the margin between $\max_{i \neq t}\{p_i(\boldsymbol{x})\}$ and $p_t(\boldsymbol{x})$.

With this C&W loss, we can indeed reduce the probability assigned to the true label y_t, thereby compelling the model to misclassify the data point into another class y_i. However, this loss still fails to guarantee that the true label y_t can be excluded from the conformal prediction set. If we let the threshold in conformal prediction play the role of $\max_{i \neq t}\{p_i(\boldsymbol{x})\}$ in Eq. 8, and push the non-conformity score of y_t further away from this threshold, the above issue can be effectively resolved. Therefore, we further improve the C&W-inspired unlearning loss function by combining conformal prediction.

In conformal prediction, calibration data helps in estimating non-conformity scores and determining a threshold to ensure valid statistical guarantees about the model's uncertainty estimates. A portion of calibration data $\mathcal{D}_c'$ can be reserved for the unlearning phase, which is kept separate from the calibration data $\mathcal{D}_c$ used in the evaluation phase. With calibration data $\mathcal{D}_c'$, the threshold $\bar{q}$ for the unlearning phase is easily calculated given an α. Given $\bar{q}$, by revising C&W-inspired unlearning loss with a calibration step, a general unlearning loss function is defined as follows:

$$\mathcal{L}_{\text{unlearn}}(\boldsymbol{x}, y_t) = \max\{\bar{q} - S(\boldsymbol{x}, y_t),\ -\Delta\}. \tag{9}$$

We replace probability $p_t(\boldsymbol{x})$ and $\max_{i \neq t}\{p_i(\boldsymbol{x})\}$ in Eq. 8 with the threshold $\bar{q}$ and non-conformity score $S(\boldsymbol{x}, y_t)$ respectively. $\bar{q}$ is updated in each training epoch to obtain an accurate value. Since $\bar{q}$ is computed merely as a quantile, this process incurs negligible computational overhead (experimental evidence is provided in Appendix E.2).

The loss $\mathcal{L}_{\text{unlearn}}$ adheres to the same principle of $\mathcal{L}_{\text{cw}}$, which encourages $S(\boldsymbol{x}, y_t) - \bar{q} \geq \Delta$. It helps to increase the non-conformity score $S(\boldsymbol{x}, y_t)$ of the true label y_t to surpass the threshold $\bar{q}$. As an improvement over the loss $\mathcal{L}_{\text{cw}}$, the loss $\mathcal{L}_{\text{unlearn}}$ makes it more difficult for the model to include the true label in conformal prediction set. In this loss, even a small value of Δ is sufficient to achieve the desired effect, because the true label y_t is excluded from the conformal prediction set once its non-conformity score $S(\boldsymbol{x}, y_t)$ exceeds the threshold $\bar{q}$. Therefore, in our work, we set $\Delta = 0.01$.

As a general framework, to preserve the efficacy of specific unlearning methods themselves, we reserve their original loss $\mathcal{L}_{\text{original}}$ in our framework. Consequently, we combine these terms to form the final objective loss function as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{original}} + \lambda \cdot \mathcal{L}_{\text{unlearn}}, \tag{10}$$

where λ is a hyperparameter that controls the forgetting degree.

Table 3: Unlearning performance on **CIFAR-10** with **ResNet-18** and **Tiny ImageNet** with **ViT** in 10% **random data forgetting**. The results are average values from 3 independent trials and the standard deviation values are reported in Appendix D. For evaluation criterion ❶, performance differences compared to the RT method are highlighted with (•). For clarity in observing criterion ❷, the sign ↑ represents greater is better, while ↓ denotes ideally small. It shows the unlearning methods that excel under the existing metric UA do not necessarily perform well under our CR metric due to the fake unlearning issue.

Methods	Existing Metrics			Coverage		Set Size		CR	
	UA ↑	RA ↑	TA ↑	$\mathcal{D}_f$ ↓	$\mathcal{D}_{test}$ ↑	$\mathcal{D}_f$ ↑	$\mathcal{D}_{test}$ ↓	$\mathcal{D}_f$ ↓	$\mathcal{D}_{test}$ ↑
CIFAR-10 with ResNet-18									
RT	8.6%(0.0)	99.7%(0.0)	91.8%(0.0)	0.941(0.000)	0.944(0.000)	1.089(0.000)	1.074(0.000)	0.864(0.000)	0.879(0.000)
FT	3.8%(4.8)	98.1%(1.6)	91.6%(0.2)	0.994(0.053)	0.951(0.007)	1.008(0.081)	1.026(0.048)	0.986(0.122)	0.927(0.048)
RL	7.6%(1.0)	97.4%(2.3)	90.6%(1.2)	0.970(0.029)	0.949(0.005)	1.242(0.153)	1.197(0.123)	0.788(0.076)	0.796(0.083)
GA	0.6%(8.0)	99.5%(0.2)	94.1%(2.3)	0.994(0.053)	0.945(0.001)	1.002(0.087)	1.009(0.065)	0.994(0.130)	0.936(0.057)
Teacher	0.8%(7.8)	99.4%(0.3)	93.5%(1.7)	0.991(0.050)	0.941(0.003)	1.003(0.086)	1.021(0.053)	0.993(0.129)	0.922(0.043)
SSD	0.5%(8.1)	99.5%(0.2)	94.2%(2.4)	0.996(0.055)	0.945(0.001)	0.999(0.090)	1.008(0.066)	0.994(0.130)	0.936(0.057)
NegGrad+	8.7%(0.1)	98.8%(0.9)	92.2%(0.4)	0.934(0.007)	0.948(0.004)	1.068(0.021)	1.086(0.012)	0.875(0.011)	0.873(0.006)
Salun	3.7%(4.9)	98.9%(0.8)	91.8%(0.0))	0.987(0.046)	0.950(0.006)	1.132(0.043)	1.143(0.069)	0.872(0.008)	0.832(0.047)
SFRon	4.8%(3.8)	97.4%(2.3)	91.4%(0.4)	0.977(0.036)	0.953(0.009)	1.100(0.011)	1.143(0.069)	0.889(0.025)	0.834(0.045)
Tiny ImageNet with ViT									
RT	14.7%(0.0)	98.8%(0.0)	86.0%(0.0)	0.944(0.000)	0.949(0.000)	1.876(0.000)	1.840(0.000)	0.503(0.000)	0.516(0.000)
FT	6.9%(7.8)	97.9%(0.9)	84.1%(1.9)	0.994(0.050)	0.950(0.001)	2.133(0.257)	2.440(0.600)	0.466(0.037)	0.389(0.127)
RL	26.9%(12.2)	96.0%(2.8)	81.4%(4.6)	0.969(0.025)	0.952(0.003)	17.890(16.014)	8.572(6.732)	0.054(0.449)	0.111(0.405)
GA	3.2%(11.5)	97.4%(1.4)	84.9%(1.1)	0.996(0.052)	0.947(0.002)	1.539(0.337)	2.018(0.178)	0.647(0.144)	0.469(0.047)
Teacher	17.3%(2.6)	86.7%(12.1)	79.0%(7.0)	0.977(0.033)	0.956(0.007)	5.473(3.597)	5.080(3.240)	0.179(0.324)	0.188(0.328)
SSD	1.5%(13.2)	98.5%(0.3)	86.1%(0.1)	0.998(0.054)	0.950(0.001)	1.354(0.522)	1.827(0.013)	0.737(0.234)	0.520(0.004)
NegGrad+	19.4%(4.7)	98.3%(0.5)	84.0%(2.0)	0.999(0.055)	0.890(0.059)	0.949(0.927)	1.614(0.227)	1.052(0.823)	0.552(1.289)
Salun	9.2%(5.5)	97.7%(1.1)	83.6%(2.4)	0.995(0.051)	0.964(0.015)	2.803(0.927)	2.726(0.886)	0.528(1.347)	0.376(1.464)
SFRon	9.3%(5.4)	97.0%(1.8)	83.9%(2.1)	0.989(0.045)	0.948(0.001)	2.000(0.124)	2.208(0.368)	0.495(0.008)	0.429(0.086)

4 EXPERIMENT

4.1 EXPERIMENTAL SETTING

Datasets and Models. We focus on the image classification task and report experiments on CIFAR-10 Krizhevsky (2009) and Tiny ImageNet Le & Yang (2015) datasets with ResNet-18 He et al. (2016) and ViT Dosovitskiy et al. (2021) architectures.

Baselines and Metrics. We employ **9 different unlearning methods**, including **RT**, **FT** Warnecke et al. (2021), **RL** Graves et al. (2021), **Gradient Ascent (GA)** Thudi et al. (2022), **Bad Teacher (Teacher)** Tarun et al. (2023), **SSD** Foster et al. (2024), **NegGrad+** Kurmanji et al. (2023), **Salun** Fan et al. (2024b) and **SFRon** Huang et al. (2025). See Appendix A for a detailed overview of these unlearning methods. We evaluate the performance of various unlearning methods using the existing metrics, including **UA**, **RA**, **TA**, **MIA**, as well as our proposed metrics **CR** and **MIACR**. See Appendix C for the detailed introduction to MIA and our implementation.

Implementation Details. For hyperparameters, we set the miscoverage rate $\alpha \in \{0.05, 0.10, 0.15, 0.20\}$. Results for $\alpha = 0.05$ are reported in the main paper, while results for $\alpha \in \{0.10, 0.15, 0.20\}$ are provided in Appendix D. The margin parameter $\Delta = 0.01$, unlearning loss weight $\lambda \in [0, 0.2, 0.5, 1]$. Additional training and baseline setup details are included in Appendix B.

4.2 MEASURE UNLEARNING METHODS VIA NEW METRICS

In this section, we explore how existing unlearning methods perform with the consideration of the fake forgetting perspective. We evaluate the performance of 9 various unlearning methods using the proposed metrics CR and MIACR, together with Coverage and Set Size. The experimental results are presented in Table 3, which summarizes the unlearning performance under 10% random data forgetting scenario on CIFAR-10 and Tiny ImageNet, respectively. See Tables 10 - 17 in Appendix D for additional experimental results on other forgetting scenarios, including class-wise, subclass-wise and worst-case forgetting.

CR Metric. We take the results on CIFAR-10 as an example for analysis of CR on forget data $\mathcal{D}_f$ based on two evaluation criteria proposed in Section 2.3. According to evaluation criterion ❶, the top 4 methods under the UA metric are *NegGrad+*, *RL*, *SFRon*, and *Salun*, as their unlearning accuracy is closest to the RT method. However, this ranking shifts slightly under the CR metric, where the top 4 become *Salun*, *NegGrad+*, *SFRon*, and *RL*. CR metric identifies that Salun performs better in forgetting quality and can deal with the fake forgetting issue well, while RL faces a fake forgetting

situation and performs poorly on our metric CR. This observation suggests that methods excelling in the traditional UA metric may not perform well under the CR metric. **The underlying rationale behind this is that the CR metric takes into account the possibility that the true labels of some misclassified forget data points may still remain within the prediction set**. This observation aligns with the insights we discussed in Section 2.2 regarding the fake forgetting issue of the UA metric.

Regarding evaluation criterion ❷, a similar pattern is observed as with criterion ❶. Under the UA metric, the top 4 methods in terms of forgetting quality are *NegGrad+*, *RT*, *RL* and *SFRon*. However, under the CR metric, the top 4 shift to *RL*, *RT*, *Salun* and *NegGrad+*. This indicates that some unlearning methods, such as NegGrad+, show weak forgetting quality when viewed from the fake forgetting perspective. This also highlights that the CR captures critical scenarios overlooked by UA, specifically the potential retention of true labels within prediction sets for the forget data points. CR ensures a more robust and reliable evaluation for unlearning quality.

MIACR Metric. In Table 4, we show the MIACR results on CIFAR-10 under both 10% and 50% random data forgetting. Under our evaluation criterion ❶, most methods show superior MIA and MIACR performance in the 10% forgetting scenario compared to 50% forgetting, because larger forget sets pose greater challenges for unlearning methods. This demonstrates that the general trend of membership leakage risk remains broadly consistent across MIA and MIACR. Under evaluation criterion ❷, Salun, which appears optimal under MIA, does not achieve the best performance when assessed by MIACR. In the 10% random forgetting scenario, MIA deems 2,121 data points as truly forgotten by Salun and 423 by SFRon. However, MIACR reveals that 1,848 of the 2,121 points under Salun can still be recovered via conformal prediction, whereas only 121 of the 423 points remain within the prediction set for SFRon.

Table 4: **MIACR** results on **CIFAR-10** with **ResNet-18** in both 10% and 50% random data forgetting.

Methods	10% Forgetting		50% Forgetting	
	MIA(%) ↓	MIACR ↑	MIA(%) ↓	MIACR ↑
RT	86.92(0.000)	0.089(0.000)	82.79(0.000)	0.117(0.000)
FT	92.00(5.08)	0.037(0.052)	92.92(10.13)	0.038(0.079)
RL	74.21(12.71)	0.056(0.033)	61.15(21.64)	0.057(0.060)
GA	98.80(11.88)	0.010(0.079)	98.86(16.07)	0.010(0.107)
Teacher	87.24(0.32)	0.011(0.078)	93.24(10.45)	0.031(0.086)
SSD	98.78(11.86)	0.010(0.079)	98.87(16.08)	0.011(0.106)
NegGrad+	90.30(3.38)	0.076(0.013)	93.82(11.03)	0.045(0.072)
Salun	57.58(29.34)	0.055(0.034)	59.12(23.67)	0.044(0.073)
SFRon	91.55(4.63)	0.060(0.029)	92.52(9.73)	0.058(0.059)

Overall, the results show that, compared to MIACR, the existing MIA metric still leaves privacy concerns. Although MIA may fail to predict some forget data points as training members, these points can still appear in the conformal prediction set with high confidence. In contrast, **MIACR more strictly controls potential membership leakage risk by measuring the probability that only non-member predictions (i.e., label '0') appear in the prediction set**.

4.3 Performance of Our Unlearning Framework

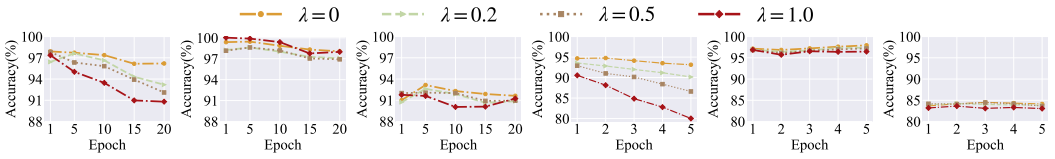
In this experiment, we apply RT, FT, and RL methods to our framework CPU, i.e., CPU-RT, CPU-FT, CPU-RL. Table 5 presents the results for CIFAR-10 with ResNet-18 and Tiny ImageNet with ViT in 10% random data forgetting. We vary λ in the range $[0, 0.2, 0.5, 1]$, where $\lambda = 0$ represents the baseline without our framework applied. See Table 18 in Appendix E for the results of $\lambda = 1$.

From the perspective of evaluation criterion ❶, we take CPU-FT as an example for analysis. The gap (blue text (•)) between CPU-FT and RT on the existing metric UA decreases effectively as λ increases. Specifically, the UA gap decreases from 4.8% to 0.7% on CIFAR-10 and from 7.8% to 0.9% on Tiny ImageNet. It is worth noting that the model utility remains relatively stable on the RA and TA results. Similarly, $\text{CR}_{\mathcal{D}_f}$ metric is also decreased when $\lambda > 0$. For the average gap across UA, RA, and TA metrics, the CPU-FT method achieves a promising average gap of 1.47 on ResNet-18 when $\lambda = 0.5$, compared to an average gap of 2.2 when $\lambda = 0$. Similarly, on the ViT model, CPU-FT reduces the average gap from 3.53 to 1.63 when $\lambda = 0.5$. It is obvious that our framework can strongly improve forgetting strength. That means the methods that are prone to over-forgetting, such as RL, perform adequately without requiring CPU for additional enhancements under our evaluation criterion ❶.

For evaluation criterion ❷, when $\lambda = 0.5$, the UA improves by an average of 3.93% on ResNet-18 and 9.23% on ViT over all methods, while TA decreases only slightly by 1.0% and 0.57% on ResNet-18 and ViT respectively. As similarly shown in the CR metric, the value of $\text{CR}_{\mathcal{D}_{test}}$ remains nearly unchanged compared to the baseline ($\lambda = 0$) with only 0.03 drop on average, while $\text{CR}_{\mathcal{D}_f}$ shows a greater reduction with an average of 0.08 across all methods.

Table 5: Performance of our unlearning framework CPU. We show the performance on **CIFAR-10** with **ResNet-18** and **Tiny ImageNet** with **ViT** in 10% **random data forgetting**. $\lambda = 0$ represents the baseline without our framework applied. It shows our framework significantly improves the forgetting quality, not only across our metric but also existing metric UA, while preserving stable predictive performance.

Methods	$\lambda = 0$					$\lambda = 0.2$					$\lambda = 0.5$				
	UA ↑	RA ↑	TA ↑	$\text{CR}_{\mathcal{D}_f}$ ↓	$\text{CR}_{\mathcal{D}_{test}}$ ↑	UA ↑	RA ↑	TA ↑	$\text{CR}_{\mathcal{D}_f}$ ↓	$\text{CR}_{\mathcal{D}_{test}}$ ↑	UA ↑	RA ↑	TA ↑	$\text{CR}_{\mathcal{D}_f}$ ↓	$\text{CR}_{\mathcal{D}_{test}}$ ↑
CIFAR-10 with ResNet-18															
CPU-RT	8.6%(0.0)	99.7%(0.0)	91.8%(0.0)	0.864(0.000)	0.879(0.000)	10.8%(2.2)	98.3%(1.4)	91.0%(0.8)	0.788(0.076)	0.824(0.055)	14.0%(5.4)	97.8%(1.9)	90.4%(0.4)	0.763(0.101)	0.825(0.054)
CPU-FT	3.8%(4.8)	98.1%(1.6)	91.6%(0.2)	0.986(0.122)	0.927(0.048)	6.8%(1.8)	97.0%(2.7)	90.8%(1.0)	0.844(0.020)	0.829(0.050)	7.9%(0.7)	96.9%(2.8)	90.9%(0.9)	0.853(0.011)	0.843(0.036)
CPU-RL	7.6%(1.0)	97.4%(2.3)	90.6%(1.2)	0.788(0.076)	0.796(0.083)	9.7%(1.1)	96.6%(3.1)	89.4%(2.4)	0.709(0.155)	0.736(0.143)	9.9%(1.3)	96.9%(2.8)	89.7%(2.1)	0.708(0.156)	0.731(0.148)
Tiny ImageNet with ViT															
CPU-RT	14.7%(0.0)	98.8%(0.0)	86.0%(0.0)	0.503(0.000)	0.516(0.000)	19.3%(4.6)	98.8%(0.0)	86.0%(0.0)	0.458(0.045)	0.516(0.000)	26.4%(11.7)	98.7%(0.1)	85.8%(0.2)	0.396(0.107)	0.489(0.027)
CPU-FT	6.9%(7.8)	97.9%(0.9)	84.1%(1.9)	0.466(0.037)	0.389(0.127)	9.8%(4.9)	97.4%(1.4)	83.6%(2.4)	0.441(0.062)	0.399(0.117)	13.6%(0.9)	97.2%(1.6)	83.6%(2.4)	0.413(0.090)	0.401(0.115)
CPU-RL	26.9%(12.2)	96.0%(2.8)	81.4%(4.6)	0.054(0.449)	0.111(0.405)	31.8%(17.1)	95.3%(17.9)	80.9%(5.1)	0.051(0.452)	0.111(0.405)	36.2%(21.5)	95.3%(3.5)	80.4%(5.6)	0.051(0.452)	0.121(0.395)



(a) $\mathcal{D}_f$ (b) $\mathcal{D}_r$ (c) $\mathcal{D}_{test}$ (d) $\mathcal{D}_f$ (e) $\mathcal{D}_r$ (f) $\mathcal{D}_{test}$

Figure 3: CPU-FT accuracy of $\mathcal{D}_f$, $\mathcal{D}_r$ and $\mathcal{D}_{test}$ under different λ values across each epoch on CIFAR-10 (a-c) and Tiny ImageNet (d-f). As λ increases, accuracy on $\mathcal{D}_f$ drops significantly, while retain and test accuracy remain stable.

Moreover, in Figure 3, we further present the CPU-FT accuracy on forget data $\mathcal{D}_f$, retain data $\mathcal{D}_r$ and test data $\mathcal{D}_{test}$ under different λ values across each epoch on Tiny ImageNet with ViT for 10% random data forgetting. As λ increases, the accuracy on $\mathcal{D}_f$ drops quickly, showing stronger unlearning effectiveness, while the accuracy on $\mathcal{D}_r$ and $\mathcal{D}_{test}$ remains stable. In summary, the experimental results demonstrate that our framework notably enhances the forgetting quality while maintaining stable predictive performance.

The experimental results demonstrate a significant improvement in both UA and $\text{CR}_{\mathcal{D}_f}$ across all methods, reflecting improved forgetting quality as λ increases. Notably, the RA, TA, and $\text{CR}_{\mathcal{D}_{test}}$ values remain relatively stable, indicating that the substantial improvement in forgetting quality does not compromise the model's predictive performance.

5 Related Work

Machine unlearning has emerged as a vital research topic due to several privacy, regulatory, and ethical concerns associated with machine learning models. It refers to the process of selectively removing specific data points from a trained machine learning model. Generally, post-hoc machine unlearning can be divided into training-based Graves et al. (2021); Tarun et al. (2023); Thudi et al. (2022); Warnecke et al. (2021) and training-free approaches Foster et al. (2024); Golatkar et al. (2021; 2020); Guo et al. (2019); Nguyen et al. (2020); Sekhari et al. (2021).

To evaluate these methods, several unlearning metrics have been proposed, including UA Brophy & Lowd (2021); Foster et al. (2024) and MIA Chen et al. (2021); Hayes et al. (2025); Shokri et al. (2017). However, these metrics often fail to account for the confidence of the forgetting quality. To address this limitation, we improve it in our work motivated by conformal prediction Angelopoulos & Bates (2021), which stands out among uncertainty quantification techniques for its ability to provide well-calibrated, reliable confidence measures. As a generic methodology, conformal prediction can transform the outputs of any black box prediction algorithm into a prediction set. Due to its versatility, many works have specifically designed numerous conformal prediction methods tailored to particular prediction problems Lei et al. (2018); Lei & Wasserman (2014); Papadopoulos et al. (2002); Romano et al. (2020a).

One work Becker & Liebig (2022) has primarily focused on parameter-level uncertainty without fully addressing the broader implications of unlearning on prediction confidence. It assesses the sensitivity of model parameters to the target data through the Fisher Information Matrix, but they often rely on computationally intensive operations and may struggle to scale to large models or datasets.

6 Conclusion

Motivated by conformal prediction, we introduce new metrics, CR and MIACR, to enhance the evaluation and reliability of machine unlearning. In addition, our unlearning framework, which incorporates the adapted C&W loss with conformal prediction, improves unlearning effectiveness. Together, we provide a more rigorous foundation for privacy-preserving machine learning.

Reproducibility Statement

The implementation details are introduced in Appendix A-B and the codes are available at https://anonymous.4open.science/r/MUCP-60E4.

References

Anastasios Angelopoulos, Stephen Bates, Jitendra Malik, and Michael I Jordan. Uncertainty sets for image classifiers using conformal prediction. arXiv preprint arXiv:2009.14193, 2020.

Anastasios N Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. arXiv preprint arXiv:2107.07511, 2021.

Alexander Becker and Thomas Liebig. Evaluating machine unlearning via epistemic uncertainty. arXiv preprint arXiv:2208.10836, 2022.

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

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

Margarida M Campos, João Calém, Sophia Sklaviadis, Mário AT Figueiredo, and André FT Martins. Sparse activations as conformal predictors. arXiv preprint arXiv:2502.14773, 2025.

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

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In 2017 ieee symposium on security and privacy (sp), pp. 39–57. Ieee, 2017.

Min Chen, Zhikun Zhang, Tianhao Wang, Michael Backes, Mathias Humbert, and Yang Zhang. When machine unlearning jeopardizes privacy. In Proceedings of the 2021 ACM SIGSAC conference on computer and communications security, pp. 896–911, 2021.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. URL https://arxiv.org/abs/2010.11929.

Chongyu Fan, Jiancheng Liu, Alfred Hero, and Sijia Liu. Challenging forgets: Unveiling the worst-case forget sets in machine unlearning. In European Conference on Computer Vision, pp. 278–297. Springer, 2024a.

Chongyu Fan, Jiancheng Liu, Yihua Zhang, Dennis Wei, Eric Wong, and Sijia Liu. Salun: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation. In International Conference on Learning Representations, 2024b.

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

Aditya Golatkar, Alessandro Achille, and Stefano Soatto. Forgetting outside the box: Scrubbing deep networks of information accessible from input-output observations. In European Conference on Computer Vision, pp. 383–398, 2020.

Aditya Golatkar, Alessandro Achille, Avinash Ravichandran, Marzia Polito, and Stefano Soatto. Mixed-privacy forgetting in deep networks. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 792–801, 2021.

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, Tom Goldstein, Awni Hannun, and Laurens Van Der Maaten. Certified data removal from machine learning models. arXiv preprint arXiv:1911.03030, 2019.

Jamie Hayes, Ilia Shumailov, Eleni Triantafillou, Amr Khalifa, and Nicolas Papernot. Inexact unlearning needs more careful evaluations to avoid a false sense of privacy. In 2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML), pp. 497–519. IEEE, 2025.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (eds.), Computer Vision – ECCV 2016, pp. 630–645, Cham, 2016. Springer International Publishing. ISBN 978-3-319-46493-0.

Jianguo Huang, Huajun Xi, Linjun Zhang, Huaxiu Yao, Yue Qiu, and Hongxin Wei. Conformal prediction for deep classifier via label ranking. arXiv preprint arXiv:2310.06430, 2023.

Zhehao Huang, Xinwen Cheng, JingHao Zheng, Haoran Wang, Zhengbao He, Tao Li, and Xiaolin Huang. Unified gradient-based machine unlearning with remain geometry enhancement. Advances in Neural Information Processing Systems, 37:26377–26414, 2025.

Jinghan Jia, Jiancheng Liu, Parikshit Ram, Yuguang Yao, Gaowen Liu, Yang Liu, Pranay Sharma, and Sijia Liu. Model sparsity can simplify machine unlearning. Advances in Neural Information Processing Systems, 36:51584–51605, 2023.

Rasha Kashef. A boosted svm classifier trained by incremental learning and decremental unlearning approach. Expert Systems with Applications, 167:114154, 2021.

Alex Krizhevsky. Learning multiple layers of features from tiny images. 2009. URL `https://api.semanticscholar.org/CorpusID:18268744`.

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.

Jing Lei and Larry Wasserman. Distribution-free prediction bands for non-parametric regression. Journal of the Royal Statistical Society Series B: Statistical Methodology, 76(1):71–96, 2014.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. Journal of the American Statistical Association, 113(523): 1094–1111, 2018.

Quoc Phong Nguyen, Bryan Kian Hsiang Low, and Patrick Jaillet. Variational bayesian unlearning. Advances in Neural Information Processing Systems, 33:16025–16036, 2020.

Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive confidence machines for regression. In Machine learning: ECML 2002: 13th European conference on machine learning Helsinki, Finland, August 19–23, 2002 proceedings 13, pp. 345–356. Springer, 2002.

Yaniv Romano, Matteo Sesia, and Emmanuel Candes. Classification with valid and adaptive coverage. Advances in Neural Information Processing Systems, 33:3581–3591, 2020a.

Yaniv Romano, Matteo Sesia, and Emmanuel Candes. Classification with valid and adaptive coverage. Advances in neural information processing systems, 33:3581–3591, 2020b.

Mauricio Sadinle, Jing Lei, and Larry Wasserman. Least ambiguous set-valued classifiers with bounded error levels. Journal of the American Statistical Association, 114(525):223–234, 2019.

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.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE international conference on computer vision, pp. 618–626, 2017.

Shaofei Shen, Chenhao Zhang, Yawen Zhao, Weitong Chen, Alina Bialkowski, and Miao Xu. Label-agnostic forgetting: a supervision-free unlearning in deep models. In 12th International Conference on Learning Representations, ICLR 2024. International Conference on Learning Representations, ICLR, 2024.

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

Liwei Song, Reza Shokri, and Prateek Mittal. Privacy risks of securing machine learning models against adversarial examples. In Proceedings of the 2019 ACM SIGSAC conference on computer and communications security, pp. 241–257, 2019.

Dharmesh Tailor, Alvaro Correia, Eric Nalisnick, and Christos Louizos. Approximating full conformal prediction for neural network regression with gauss-newton influence. In The Thirteenth International Conference on Learning Representations.

Ayush Kumar Tarun, Vikram Singh Chundawat, Murari Mandal, and Mohan Kankanhalli. Deep regression unlearning. In International Conference on Machine Learning, pp. 33921–33939. PMLR, 2023.

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

Alexander Warnecke, Lukas Pirch, Christian Wressnegger, and Konrad Rieck. Machine unlearning of features and labels. arXiv preprint arXiv:2108.11577, 2021.

Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In 2018 IEEE 31st computer security foundations symposium (CSF), pp. 268–282. IEEE, 2018.

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

APPENDIX

A BASELINE DETAILS

We introduce the details of our unlearning baselines as follows : **RT** retrains the model from scratch using only the remaining dataset $\mathcal{D}_r$. **FT** Warnecke et al. (2021) fine-tunes the pre-trained model θ_o on the remaining dataset $\mathcal{D}_r$. **RL** Graves et al. (2021) fine-tunes the model on the forgetting dataset $\mathcal{D}_f$ using randomly assigned labels to enforce forgetting. **GA** Thudi et al. (2022) performs gradient ascent on the forgetting data $\mathcal{D}_f$, which often harms the model's utility. **Teacher** Tarun et al. (2023) distills knowledge from a corrupted teacher model to the student, aiming to uniformly increase the loss on forgetting samples but often causing catastrophic forgetting. **SSD** Foster et al. (2024) induces forgetting by identifying and dampening parameters highly associated with the forgetting set using the Fisher information matrix, without retraining. **NegGrad+** Kurmanji et al. (2023) addresses GA's issue by combining fine-tuning on $\mathcal{D}_r$ and gradient ascent on $\mathcal{D}_f$. **Salun** Fan et al. (2024b) performs unlearning by optimizing only the salient parameters of the model identified from the random labeled forgetting data. **SFRon** Huang et al. (2025) embeds the unlearning update into the parameter manifold shaped by the retained data using Hessian modulation, approximated via a fast-slow update strategy.

B SETTING DETAILS

For CIFAR-10 with ResNet-18 architecture, we train the original model from scratch for 200 epochs using SGD with a Cosine Annealing learning rate schedule, starting from an initial learning rate of 0.1. We set the momentum to 0.9 and a batch size of 64. The RT model adopts the same training configuration. Other models are trained for the following durations: FT for 20 epochs, RL for 10 epochs, SalUn for 10 epochs, GA for 1 epoch (to avoid over-forgetting and significant RA degradation), NegGrad+ for 10 epochs (reduced to 2 epochs in class-wise scenarios), and SFRon for 10 epochs. All other hyperparameters match those of the original model.

For the ViT architecture, we initialize the original model by training a pretrained ViT model for 15 epochs on Tiny ImageNet. We start with a learning rate of 0.001, while other training parameters match those used for ResNet-18. We use SGD and set the momentum to 0.9 and a batch size of 64. The RT model follows the same training procedure as the original model. Other models are trained for the following durations: FT for 5 epochs, RL for 5 epochs, Salun for 5 epochs, GA for 1 epoch, NegGrad+ for 5 epochs, and SFRon for 5 epochs. All other hyperparameters are consistent with the original model's training.

For CIFAR-10/Tiny ImageNet, we randomly select $200/50$ data points per class ($2000/10000$ data points in total) as calibration data $\mathcal{D}_c$ and $\mathcal{D}'_c$, respectively. The calibration data $\mathcal{D}_c$ does not participate in the model training or unlearning processes and is only used for calibrating the threshold $\hat{q}$, while $\mathcal{D}'_c$ is used in the process of our unlearning framework to generate $\bar{q}$. All experiments are conducted on 1 Tesla V100-SXM2 GPU card with 32GB memory in a single node.

C MIA IMPLEMENTATION DETAILS

Following prior works Jia et al. (2023); Kurmanji et al. (2023); Zhao et al. (2024); Song et al. (2019); Yeom et al. (2018), we adopt a confidence-based membership inference attack to evaluate the privacy preservation of the unlearning model. Specifically, we construct an MIA predictor by training it on a balanced dataset sampled from the retain set $\mathcal{D}_r$ (labeled as members) and the test set $\mathcal{D}_{test}$ (labeled as non-members). The trained support vector classifier (SVC) is then applied to the unlearning model θ_u during evaluation.

To measure unlearning effectiveness, we compute the MIA success rate, which quantifies how many samples in the forget set $\mathcal{D}_f$ are still predicted as training members by the MIA predictor. Formally,

$$\text{MIA} = \frac{\text{TP}}{|\mathcal{D}_f|}, \tag{11}$$

where TP represents the count of forget samples still identified as training samples and $|\mathcal{D}_f|$ is the size of the forget data $\mathcal{D}_f$.

Intuitively, since the MIA score reflects the success rate of membership inference attacks on the forget data, a lower score indicates that less membership information about $\mathcal{D}_f$ is retained in θ_u, implying stronger privacy preservation and more effective unlearning.

D EVALUATING MU METHODS

D.1 MIS-LABEL NUMBER AND IN-SET RATIOS

Table 6: Mis-label number and in-set ratios of UA and MIA metrics.

Methods	10% Forgetting			50% Forgetting		
	Mis-label ↑	In-set ↓	Ratio ↓	Mis-label ↑	In-set ↓	Ratio ↓
Mis-label and In-set Ratio of UA						
RT	431	132	30.6%	2,745	1,573	57.3%
FT	192	112	58.3%	647	431	66.6%
RL	380	173	45.5%	2,625	1,795	68.4%
GA	30	2	6.7%	150	9	6.0%
Teacher	40	4	10%	400	37	9.3%
SSD	25	2	8.0%	116	9	7.8%
NegGrad+	435	115	26.4%	711	249	35.5%
Salun	185	117	63.2%	1,065	695	65.3%
SFRon	240	125	52.1%	1,000	610	61.0%
Mis-label and In-set Ratio of MIA						
RT	654	209	32.0%	4,303	1,391	32.3%
FT	400	216	54.0%	1,769	813	46.0%
RL	1,289	1,011	78.4%	9,713	8,295	85.4%
GA	60	10	16.7%	284	31	10.9%
Teacher	638	586	91.8%	1,689	895	53.0%
SSD	61	11	18.0%	282	24	8.5%
NegGrad+	486	106	21.8%	1,545	415	26.9%
Salun	2,121	1,848	87.1%	10,221	9,121	89.2%
SFRon	423	121	28.6%	1,871	433	23.1%

Conformal prediction is applied to UA and MIA predictions to determine the number of misclassified data points (mis-label) and the number of these points that fall within the conformal prediction set (in-set). We evaluate both the UA and MIA metrics by counting the misclassified data points and calculating how many of them are included in the conformal prediction set. The detailed results are presented in Table 6, which is the extended results of Table 2.

D.2 DISTRIBUTION COMPARISON OF FORGOTTEN DATA ON UA AND CR

As shown in Figures 7-10, we further analyze the probability and loss distributions of ground truth labels for data identified as truly forgotten by CR (i.e., out-set) and UA (i.e., mis-label), respectively. The distribution curves are fitted using KDE for clearer visualization. The softmax outputs for 'out-set' are consistently near 0 compared to 'mis-label', which strongly suggests that 'out-set' more rigorously captures real forgotten data. In the cross-entropy loss distribution, forgotten data identified by CR consistently show higher cross-entropy loss than UA. Higher loss indicates better forgetting quality, which further validates that CR better removes fake forgetting data.

D.3 CR METRIC

Tables 11 and 12 show the unlearning performance on CIFAR-10 with ResNet-18 in 10% and 50% random data forgetting scenarios, while Table 13 is the results in class-wise forgetting scenario. Tables 14 and 15 present the unlearning performance on Tiny ImageNet with ResNet-18 in the random data forgetting scenario, while Table 16 details the unlearning performance in the class-wise forgetting scenario. For class-wise forgetting scenario, we note $\mathcal{D}_{test} = \mathcal{D}_{tf} \cup \mathcal{D}_{tr}$. $\mathcal{D}_{tf}$ corresponds to the test-forget data exclusively containing the forget class, while $\mathcal{D}_{tr}$ represents the test-retain data within the test data $\mathcal{D}_{test}$.

For all unlearning methods, as α level increases, it results in reduced Coverage and smaller Set Size. This happens because a higher α loosens the conformal threshold $\hat{q}$, allowing fewer predictions to be included within the prediction set for each data point. On the contrary, the CR tends to increase with increasing α. Although both Coverage and Set Size may decrease, Set Size often decreases more

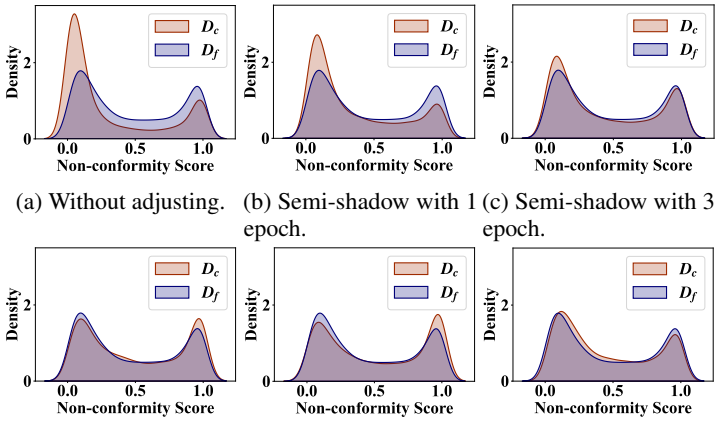
significantly. Consequently, the CR value of $\mathcal{D}_f$ generally becomes larger as α increases. It is natural that the adjustment of α affects both Coverage and Set Size. However, the final CR value really depends on the model's performance itself. For a strict evaluation, we encourage setting α to 0.5.

When α is set to 0.2, most methods show a value of Set Size less than 1 in both Table 11, 12, 14, 15. The intuition behind it is that conformal prediction, as a static predictor, is intrinsically tied to the model's base prediction performance and accuracy. When the model's accuracy is significantly higher than the confidence level, conformal prediction can achieve the required coverage with ease. In fact, it can generate partial empty prediction sets for some data points while still meeting the target coverage. Thus, the choice of α is crucial. Overly high α values may skew evaluation results by failing to let CR accurately reflect model performance. Therefore, we emphasize that a small α is generally appropriate for most unlearning scenarios.

Notably, the insights gained from the random data forgetting scenario can also be extended to the class-wise forgetting scenario. Additionally, in the class-wise scenario, some unlearning methods like RT and RL with UA = 100% and CR approaching 0% indicate they are truly effective at forgetting the specified class.

D.4 MIACR METRIC

Table 17 presents the performance of 9 machine unlearning methods on CIFAR-10 in ResNet-18, evaluated with the MIACR metric. In addition to the settings discussed in Section 4, we include results for $\alpha \in [0.1, 0.15, 0.2]$ in Table 17.



(a) Without adjusting. (b) Semi-shadow with 1 epoch. (c) Semi-shadow with 3 epoch.

(d) Semi-shadow with 4 epoch. (e) Semi-shadow with 5 epoch. (f) Shadow with 5 epoch.

Figure 4: Distribution shifting processing with different strategies. The distribution of calibration data gradually converges with that of forget data.

D.5 MEASURING FORGETTING UNDER DISTRIBUTION SHIFTS

RL and Salun are unlearning methods that employ label corruption in their unlearning strategy, which can cause distribution shifts. Here, we introduce how to better measure forgetting under these circumstances. Figure 4(a) shows the non-conformity score distribution of calibration data $\mathcal{D}_c$ and forget data $\mathcal{D}_f$ in the unlearning model θ_u obtained by the RL method in Tiny ImageNet with ViT. It looks like there is a significant discrepancy between the distribution of the forget data and the calibration data.

To align the distribution of $\mathcal{D}_c$ with that of $\mathcal{D}_f$ and minimize the differences between them, we design a shadow model. To make the explanation clearer and more intuitive, we take RL as an example. In the RL unlearning method, the forget data is assigned random labels. Therefore, we apply the same random labeling process to the calibration data and train a shadow model accordingly. We designed two methods:

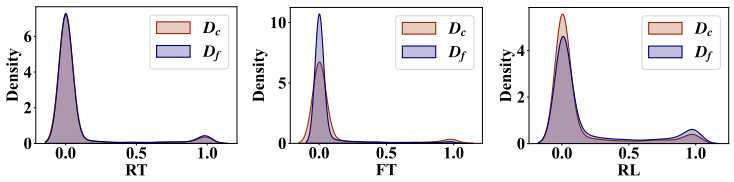


Figure 5: Non-conformity density of calibration data $\mathcal{D}_c$ and forget data $\mathcal{D}_f$ **without our unlearning framework** in CIFAR-10 with ResNet-18 under 10% random data forgetting scenario.

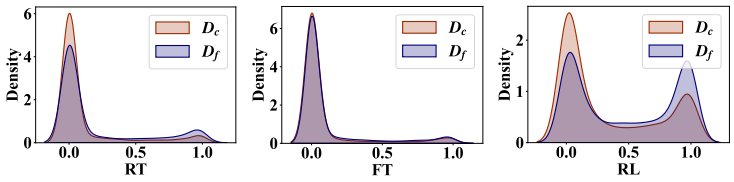


Figure 6: Non-conformity score density of calibration data $\mathcal{D}_c$ and forget data $\mathcal{D}_f$ **with our unlearning framework** in CIFAR-10 with ResNet-18 under 10% random data forgetting scenario. Our unlearning framework shifts the distribution of the forget data to the right, demonstrating improved forgetting quality.

1. **Shadow model**. A shadow model replicates the behavior of forget data $\mathcal{D}_f$ throughout the unlearning process. A shadow model is a two-step approach: (1) it firstly trains a shadow original model θ_o' using train data $\mathcal{D}_{train}$ and clean calibration data $\mathcal{D}_c$ with the same epoch number as the original model θ_o; (2) subsequently, we finetune the θ_o' using the random labeled calibration data.
2. **Semi-shadow model**. The semi-shadow model only adopts the second step in the shadow model. It finetunes the original model θ_o with random-labeled calibration data.

The results are presented in Figure 4, where (b)-(e) present the results of the semi-shadow model with different epochs and (f) illustrates the shadow model's result. Under the semi-shadow model, as the number of epochs increases, the distribution of calibration data gradually moves to the right until it becomes consistent with the distribution of forget data. It also shows that the shadow model demonstrates the best ability to handle distribution shifts compared to the semi-shadow model. However, this comes at the cost of higher computational overhead. Overall, the semi-shadow model offers a balanced trade-off between handling distribution shifts effectively and maintaining lower computational costs.

E Performance of Our Unlearning Framework

E.1 Unlearning Performance

Table 18 presents the performance of our unlearning framework, including $\alpha \in [0.05, 0.1, 0.15, 0.2]$. We explored the impact of varying λ within the range $[0, 0.2, 0.5, 0.1]$, where $\lambda = 0$ serves as the baseline without applying our framework, which can be found in Tables 11 and 14. The results reveal a clear trend: as λ increases, the UA improves significantly across all methods, accompanied by a substantial reduction in $\text{CR}_{\mathcal{D}_f}$. Interestingly, the RA, TA, and CR$\mathcal{D}test$ metrics remain relatively stable. These results underscore the effectiveness of our unlearning framework in achieving substantial improvements in forgetting quality while preserving the stability of the model's predictive performance.

Furthermore, we conduct an ablation study and analyze the impact of using our unlearning framework. As illustrated in Figures 5 and 6, we compare the density distributions of non-conformity scores for calibration data $\mathcal{D}_c$ and forget data $\mathcal{D}_f$ under the RT, FT, and RL unlearning methods. We set λ to 1. Clearly, a higher non-conformity score for $\mathcal{D}_f$ indicates that it is less likely to be included in the conformal prediction set, reflecting more effective forgetting.

Comparing Figures 5 and 6, after applying our unlearning framework, we observe a significant rightward shift in the non-conformity score distribution of forget data, which is a promising signal

according to evaluation criterion ❷. Furthermore, the FT distribution in Figure 6 exhibits substantial overlap with the calibration data, nearly matching the distribution observed in RT. Based on evaluation criterion ❶, since calibration data represents unseen examples, the similarity between forget data and calibration data distributions provides strong evidence of effective forgetting. Overall, the results evaluated on both evaluation criteria ❶ and ❷ consistently confirm the efficacy of our framework in enhancing forgetting quality.

Table 7: Training time comparison (in minutes) with and without our CPU loss.

Methods	w/o CPU	w/ CPU
CIFAR-10 with ResNet18		
RT	70.1	72.1
FT	6.3	6.8
RL	6.3	6.8
Tiny ImageNet with ViT		
RT	60.75	62.85
FT	20.2	22.1
RL	21.3	23.4

E.2 Time Comparison

We compare the training time with and without our unlearning calibration process on CIFAR-10 and Tiny ImageNet under the 10% random data forgetting scenario. As shown in Table 7, the training times with and without CPU support differ only marginally, confirming that our CPU loss computation introduces negligible overhead.

F Other Conformal Prediction Methods

Table 8: CR performance with different conformal prediction methods. The performance gap relative to the RT method is represented in (•).

Methods	LAC		EntmaxScore		APS	
	CR($\mathcal{D}_f$) ↓	CR($\mathcal{D}_f$) ↑	CR($\mathcal{D}_f$) ↓	CR($\mathcal{D}_f$) ↑	CR($\mathcal{D}_f$) ↓	CR($\mathcal{D}_f$) ↑
RT	0.862(0.000)	0.876(0.000)	0.863(0.000)	0.877(0.000)	0.805(0.000)	0.836(0.000)
FT	0.901(0.039)	0.846(0.030)	0.901(0.038)	0.848(0.029)	0.808(0.004)	0.784(0.052)
RL	0.676(0.186)	0.752(0.124)	0.883(0.020)	0.838(0.039)	0.573(0.232)	0.670(0.166)
GA	0.995(0.133)	0.931(0.055)	0.995(0.132)	0.930(0.054)	0.985(0.180)	0.875(0.038)
Teacher	0.988(0.127)	0.915(0.039)	0.987(0.125)	0.917(0.040)	0.511(0.293)	0.536(0.300)
SSD	0.995(0.133)	0.933(0.057)	0.994(0.131)	0.930(0.054)	0.985(0.181)	0.876(0.039)
NegGrad+	0.865(0.003)	0.863(0.013)	0.869(0.006)	0.870(0.006)	0.860(0.056)	0.856(0.020)
Salun	0.881(0.019)	0.839(0.037)	0.878(0.015)	0.839(0.038)	0.407(0.398)	0.430(0.407)
SFRon	0.893(0.031)	0.838(0.038)	0.893(0.030)	0.838(0.039)	0.815(0.010)	0.769(0.067)

While we adopt vanilla split-conformal as the default due to its simplicity and reproducibility, our framework is not limited to this variant. Here, we report the results using other conformal prediction methods, LAC Sadinle et al. (2019), EntmaxScore Campos et al. (2025), and ASP Romano et al. (2020b) on CIFAR-10 with ResNet18 under 10% random data forgetting.

As shown in the Table 8, the CR results of LAC and EntmaxScore are similar to those obtained using SCP in Table 3. This suggests that the results are stable under conformal prediction methods that offer formal coverage guarantees. However, APS produces different CR values compared to LAC, SCP, and EntmaxScore. This discrepancy is expected and is due to the inherent characteristics of APS, which make it unsuitable for evaluating unlearning metrics. APS generally produces loose prediction sets and is highly sensitive to noisy probability estimates in the lower-ranked classes Angelopoulos et al. (2020), which introduces randomness in the ordering of unlikely classes and leads to unreliable set construction. Our findings indicate that not all conformal prediction methods are inherently suitable for evaluating forgetting quality. And the reliability of such evaluation depends critically on whether the resulting prediction sets faithfully capture the model's uncertainty.

Table 9: Unlearning performance on **CIFAR-10** with **ResNet-18** in 10% **worst-case data forgetting** scenario. The results are reported in the format a±b, where a is the mean and b is the standard deviation from 3 independent trials. The performance gap relative to the RT method is represented in (•).

Methods	Existing Metrics			Coverage		Set Size		CR	
	UA ↑	RA ↑	TA ↑	$\mathcal{D}_f$ ↓	$\mathcal{D}_{test}$ ↑	$\mathcal{D}_f$ ↑	$\mathcal{D}_{test}$ ↓	$\mathcal{D}_f$ ↓	$\mathcal{D}_{test}$ ↑
RT	0.0%(0.0)	99.2%(0.0)	91.5%(0.0)	1.000(0.000)	0.948(0.000)	1.000(0.000)	1.116(0.000)	1.000(0.000)	0.850(0.000)
FT	0.0%(0.0)	99.8%(0.6)	94.1%(2.6)	1.000(0.000)	0.938(0.010)	1.000(0.000)	0.992(0.124)	1.000(0.000)	0.945(0.095)
RL	21.3%(21.3)	97.4%(1.7)	88.5%(3.0)	0.976(0.024)	0.955(0.007)	6.753(5.753)	2.192(1.076)	0.146(0.854)	0.441(0.409)
GA	0.3%(0.3)	96.9%(2.2)	91.3%(0.2)	0.999(0.001)	0.954(0.006)	1.029(0.029)	1.179(0.063)	0.971(0.029)	0.810(0.040)
Teacher	15.8%(15.8)	97.9%(1.2)	90.6%(0.9)	0.850(0.150)	0.946(0.002)	1.177(0.177)	1.249(0.133)	0.745(0.255)	0.760(0.090)
SSD	0.0%(0.0)	99.7%(0.5)	94.0%(2.6)	1.000(0.000)	0.954(0.006)	1.000(0.000)	1.037(0.079)	1.000(0.000)	0.920(0.070)
NegGrad+	0.0%(0.0)	99.8%(0.6)	94.2%(2.7)	1.000(0.000)	0.947(0.001)	1.000(0.000)	1.012(0.104)	1.000(0.000)	0.936(0.086)
SalUn	13.0%(13.0)	97.6%(1.6)	90.0%(1.5)	0.962(0.038)	0.947(0.001)	3.991(2.991)	1.567(0.451)	0.246(0.754)	0.606(0.244)
SFRon	0.0%(0.0)	99.5%(0.3)	93.8%(2.4)	1.000(0.000)	0.956(0.008)	1.000(0.000)	1.053(0.063)	1.000(0.000)	0.908(0.058)

Overall, conformal prediction serves as a component within our uncertainty quantification-based evaluation framework. The simplest and most straightforward conformal prediction methods, especially SCP, are often the most suitable tools. While many recent conformal prediction variants improve upon different issues, e.g., by modifying the nonconformity scores or explicitly penalizing low-probability classes Angelopoulos et al. (2020); Huang et al. (2023), these techniques often distort the nonconformity values across some classes. Since our goal is to use conformal prediction as a tool for designing fair metrics and evaluating forgetting quality, we intentionally avoid such modifications. Introducing these more complex methods could result in additional noise, thereby compromising the fairness and interpretability of our evaluation.

G Other Forgetting Scenario

Worst-case Forgetting scenario Random data forgetting may affect unlearning models differently, introducing variance and bias that make it a relatively weak evaluation setting. To more rigorously assess the effectiveness of our proposed metrics, we further evaluate them using worst-case forget sets Fan et al. (2024a). As shown in Table 9, the results are consistent with our previous analysis.

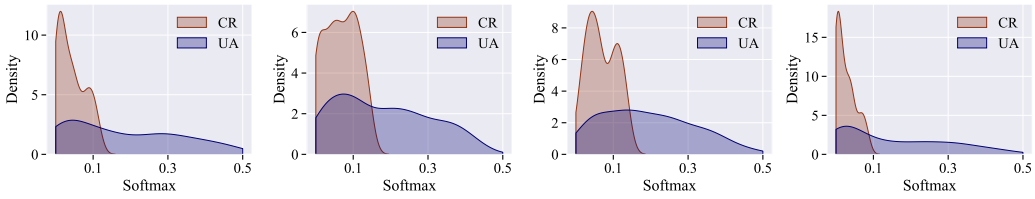
Table 10: Unlearning performance on **CIFAR-20** with **ResNet18** in **subclass-wise forgetting** scenario.

Methods	Existing Metrics				Coverage			Set Size			CR		
	UA ↑	UA_f ↑	RA ↑	TA ↑	$\mathcal{D}_f$ ↓	$\mathcal{D}_{tf}$ ↓	$\mathcal{D}_{tr}$ ↑	$\mathcal{D}_f$ ↑	$\mathcal{D}_{tf}$ ↑	$\mathcal{D}_{tr}$ ↓	$\mathcal{D}_f$ ↓	$\mathcal{D}_{tf}$ ↓	$\mathcal{D}_{tr}$ ↑
RT	97.6%(0.0)	94.0%(0.0)	99.9%(0.0)	84.5%(0.0)	1.000(0.000)	1.000(0.000)	0.953(0.000)	20.000(0.000)	20.000(0.000)	1.713(0.000)	0.050(0.000)	0.050(0.000)	0.556(0.000)
FT	70.9%(26.7)	74.7%(19.3)	95.7%(4.1)	76.0%(8.6)	0.994(0.006)	0.987(0.013)	0.952(0.001)	17.637(2.363)	16.893(3.107)	3.091(1.377)	0.057(0.007)	0.059(0.009)	0.312(0.245)
RL	99.5%(1.9)	94.7%(0.7)	98.2%(1.7)	76.7%(7.9)	0.931(0.069)	1.000(0.000)	0.955(0.001)	18.807(1.193)	19.527(0.473)	3.300(1.586)	0.050(0.000)	0.051(0.001)	0.289(0.267)
GA	40.7%(56.9)	60.7%(33.3)	99.0%(0.8)	82.2%(2.3)	0.999(0.001)	0.993(0.007)	0.954(0.001)	18.305(1.695)	17.553(2.447)	2.409(0.695)	0.055(0.005)	0.057(0.007)	0.397(0.159)
Teacher	90.6%(7.0)	97.3%(3.3)	98.6%(1.3)	81.3%(3.2)	0.989(0.011)	0.933(0.067)	0.948(0.005)	19.871(0.129)	18.840(1.160)	2.747(1.034)	0.050(0.000)	0.050(0.000)	0.350(0.206)
SSD	73.6%(24.0)	80.0%(14.0)	99.8%(0.0)	84.5%(0.1)	0.997(0.003)	0.980(0.020)	0.955(0.001)	19.206(0.794)	17.740(2.260)	2.407(0.694)	0.052(0.002)	0.055(0.005)	0.423(0.133)
NegGrad+	98.9%(1.3)	100.0%(6.0)	97.0%(2.8)	80.9%(3.7)	1.000(0.000)	1.000(0.000)	0.950(0.003)	20.000(0.000)	20.000(0.000)	2.761(1.048)	0.050(0.000)	0.050(0.000)	0.372(0.184)
Salun	99.9%(2.3)	96.0%(2.0)	98.8%(1.0)	78.9%(5.6)	0.955(0.045)	0.993(0.007)	0.951(0.002)	19.235(0.765)	19.707(0.293)	2.737(1.023)	0.050(0.000)	0.050(0.000)	0.348(0.208)
SFRon	99.9%(2.3)	100.0%(6.0)	91.9%(7.9)	79.7%(4.9)	1.000(0.000)	1.000(0.000)	0.951(0.003)	20.000(0.000)	20.000(0.000)	2.587(0.874)	0.050(0.000)	0.050(0.000)	0.370(0.186)

Subclass-wise Forgetting Scenario To further verify our metrics in other forgetting scenarios, we report subclass-wise forgetting results on CIFAR-20 (derived from CIFAR-100) using ResNet-18, following the setting proposed in Foster et al. (2024). As shown in the Table 10, the findings align well with our prior analysis.

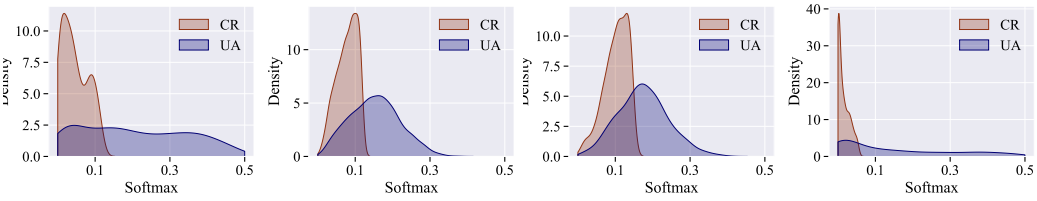
H Large Language Models Usage Statement

We used a large language model (LLM) to polish the language and improve the clarity of the paper. All content, including the core ideas, methodology, and experimental results, was originally created by the authors. The LLM was used exclusively as an editing tool to enhance readability and grammatical correctness, without generating any substantive or technical content.



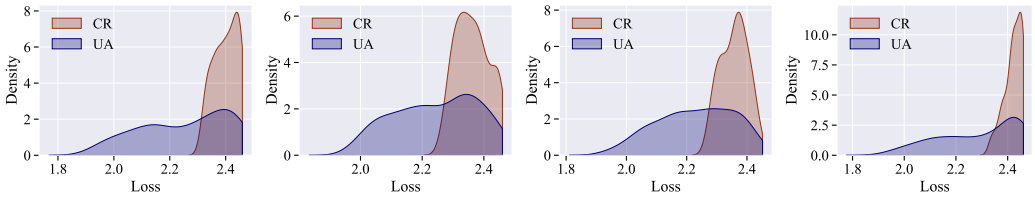
(a) FT (b) RL (c) Salun (d) SFRon

Figure 7: **Softmax distribution** in 10% **random data forgetting** scenario. We analyze the softmax distributions of true labels for data identified as truly forgotten by CR and UA, respectively. The distribution curves are fitted using KDE for clearer visualization. The results illustrate the softmax distributions of CR consistently closer to 0 when compared to UA, providing strong evidence that CR is better than UA in accurately capturing and measuring 'real forgetting'.



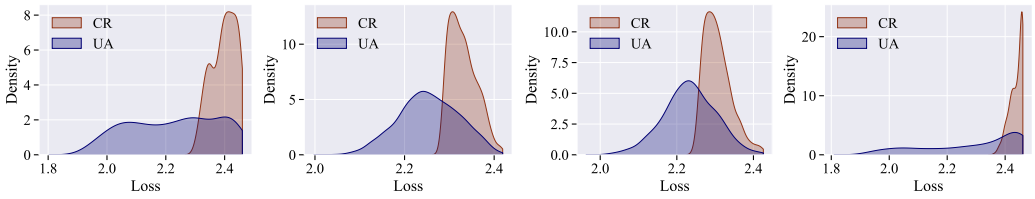
(a) FT (b) RL (c) Salun (d) SFRon

Figure 8: **Softmax distribution** in 50% **random data forgetting** scenario.



(a) FT (b) RL (c) Salun (d) SFRon

Figure 9: **Loss distribution** in 10% **random data forgetting** scenario. We analyze the cross-entropy loss distributions of true labels for data identified as truly forgotten by CR and UA, respectively. Forgotten data identified by CR consistently show higher cross-entropy loss than UA. Higher loss indicates better forgetting quality, which further validates that CR better captures 'real forgetting'.



(a) FT (b) RL (c) Salun (d) SFRon

Figure 10: **Loss distribution** in 50% **random data forgetting** scenario.

Table 11: Unlearning performance of 9 unlearning methods on **CIFAR-10** with **ResNet-18** in 10% **random data forgetting** scenario. The results are reported in the format a±b, where a is the mean and b is the standard deviation from 3 independent trials. The performance gap relative to the RT method is represented in (•).

Methods	α	Coverage		Set Size		CR		
		$\mathcal{D}_f \downarrow$	$\mathcal{D}_{test} \uparrow$	$\mathcal{D}_f \uparrow$	$\mathcal{D}_{test} \downarrow$	$\mathcal{D}_f \downarrow$	$\mathcal{D}_{test} \uparrow$	$\hat{q}$
RT **UA**8.6%, **RA**99.7%, **TA**91.8%	0.05	$0.941_{\pm 0.002}(0.000)$	$0.944_{\pm 0.005}(0.000)$	$1.089_{\pm 0.002}(0.000)$	$1.074_{\pm 0.011}(0.000)$	$0.864_{\pm 0.004}(0.000)$	$0.879_{\pm 0.004}(0.000)$	$0.883_{\pm 0.007}$
	0.1	$0.881_{\pm 0.000}(0.000)$	$0.895_{\pm 0.010}(0.000)$	$0.934_{\pm 0.004}(0.000)$	$0.947_{\pm 0.008}(0.000)$	$0.943_{\pm 0.011}(0.000)$	$0.945_{\pm 0.001}(0.000)$	$0.192_{\pm 0.001}$
	0.15	$0.820_{\pm 0.002}(0.000)$	$0.839_{\pm 0.008}(0.000)$	$0.841_{\pm 0.009}(0.000)$	$0.867_{\pm 0.009}(0.000)$	$0.975_{\pm 0.001}(0.000)$	$0.968_{\pm 0.003}(0.000)$	$0.015_{\pm 0.011}$
	0.2	$0.780_{\pm 0.007}(0.000)$	$0.808_{\pm 0.004}(0.000)$	$0.789_{\pm 0.002}(0.000)$	$0.824_{\pm 0.009}(0.000)$	$0.988_{\pm 0.006}(0.000)$	$0.981_{\pm 0.007}(0.000)$	$0.003_{\pm 0.002}$
FT **UA**3.8%, **RA**98.1%, **TA**91.6%	0.05	$0.994_{\pm 0.001}(0.053)$	$0.951_{\pm 0.004}(0.007)$	$1.008_{\pm 0.003}(0.081)$	$1.026_{\pm 0.008}(0.048)$	$0.986_{\pm 0.003}(0.122)$	$0.927_{\pm 0.004}(0.048)$	$0.721_{\pm 0.045}$
	0.1	$0.968_{\pm 0.001}(0.087)$	$0.899_{\pm 0.005}(0.004)$	$0.969_{\pm 0.001}(0.035)$	$0.924_{\pm 0.008}(0.023)$	$0.998_{\pm 0.001}(0.055)$	$0.972_{\pm 0.003}(0.027)$	$0.079_{\pm 0.013}$
	0.15	$0.915_{\pm 0.003}(0.095)$	$0.848_{\pm 0.002}(0.009)$	$0.916_{\pm 0.003}(0.075)$	$0.860_{\pm 0.001}(0.007)$	$1.000_{\pm 0.000}(0.025)$	$0.986_{\pm 0.002}(0.018)$	$0.008_{\pm 0.000}$
	0.2	$0.861_{\pm 0.010}(0.081)$	$0.806_{\pm 0.008}(0.002)$	$0.861_{\pm 0.010}(0.072)$	$0.811_{\pm 0.009}(0.013)$	$1.000_{\pm 0.000}(0.012)$	$0.993_{\pm 0.001}(0.012)$	$0.002_{\pm 0.000}$
RL **UA**7.6%, **RA**97.4%, **TA**90.6%	0.05	$0.970_{\pm 0.006}(0.029)$	$0.949_{\pm 0.005}(0.005)$	$1.242_{\pm 0.151}(0.153)$	$1.197_{\pm 0.098}(0.123)$	$0.788_{\pm 0.089}(0.076)$	$0.796_{\pm 0.061}(0.083)$	$0.877_{\pm 0.057}$
	0.1	$0.913_{\pm 0.010}(0.032)$	$0.897_{\pm 0.007}(0.002)$	$0.975_{\pm 0.028}(0.041)$	$0.980_{\pm 0.025}(0.033)$	$0.936_{\pm 0.022}(0.007)$	$0.916_{\pm 0.019}(0.029)$	$0.572_{\pm 0.059}$
	0.15	$0.825_{\pm 0.006}(0.005)$	$0.843_{\pm 0.009}(0.004)$	$0.854_{\pm 0.010}(0.013)$	$0.888_{\pm 0.017}(0.021)$	$0.966_{\pm 0.006}(0.009)$	$0.949_{\pm 0.009}(0.019)$	$0.329_{\pm 0.021}$
	0.2	$0.755_{\pm 0.021}(0.025)$	$0.798_{\pm 0.005}(0.010)$	$0.774_{\pm 0.020}(0.015)$	$0.832_{\pm 0.009}(0.008)$	$0.976_{\pm 0.002}(0.012)$	$0.959_{\pm 0.005}(0.022)$	$0.234_{\pm 0.028}$
GA **UA**0.6%, **RA**99.5%, **TA**94.1%	0.05	$0.994_{\pm 0.003}(0.053)$	$0.945_{\pm 0.008}(0.001)$	$1.002_{\pm 0.010}(0.087)$	$1.009_{\pm 0.010}(0.065)$	$0.994_{\pm 0.016}(0.130)$	$0.936_{\pm 0.011}(0.057)$	$0.621_{\pm 0.015}$
	0.1	$0.990_{\pm 0.005}(0.109)$	$0.905_{\pm 0.019}(0.010)$	$0.990_{\pm 0.014}(0.056)$	$0.928_{\pm 0.005}(0.019)$	$0.998_{\pm 0.002}(0.055)$	$0.973_{\pm 0.012}(0.028)$	$0.062_{\pm 0.016}$
	0.15	$0.969_{\pm 0.012}(0.149)$	$0.848_{\pm 0.004}(0.009)$	$0.969_{\pm 0.014}(0.128)$	$0.858_{\pm 0.019}(0.009)$	$1.000_{\pm 0.014}(0.025)$	$0.986_{\pm 0.008}(0.018)$	$0.006_{\pm 0.009}$
	0.2	$0.925_{\pm 0.012}(0.145)$	$0.805_{\pm 0.022}(0.003)$	$0.924_{\pm 0.007}(0.135)$	$0.811_{\pm 0.013}(0.013)$	$0.998_{\pm 0.013}(0.010)$	$0.992_{\pm 0.012}(0.011)$	$0.003_{\pm 0.005}$
Teacher **UA**0.8%, **RA**99.4%, **TA**93.5%	0.05	$0.991_{\pm 0.022}(0.050)$	$0.941_{\pm 0.001}(0.003)$	$1.003_{\pm 0.012}(0.086)$	$1.021_{\pm 0.009}(0.053)$	$0.993_{\pm 0.021}(0.129)$	$0.922_{\pm 0.015}(0.043)$	$0.744_{\pm 0.015}$
	0.1	$0.967_{\pm 0.000}(0.086)$	$0.898_{\pm 0.007}(0.003)$	$0.963_{\pm 0.007}(0.029)$	$0.929_{\pm 0.018}(0.018)$	$0.998_{\pm 0.000}(0.055)$	$0.969_{\pm 0.013}(0.024)$	$0.591_{\pm 0.005}$
	0.15	$0.913_{\pm 0.006}(0.093)$	$0.845_{\pm 0.007}(0.006)$	$0.912_{\pm 0.014}(0.071)$	$0.859_{\pm 0.005}(0.008)$	$0.996_{\pm 0.018}(0.021)$	$0.983_{\pm 0.015}(0.015)$	$0.481_{\pm 0.009}$
	0.2	$0.865_{\pm 0.009}(0.085)$	$0.806_{\pm 0.021}(0.002)$	$0.866_{\pm 0.009}(0.077)$	$0.816_{\pm 0.012}(0.008)$	$0.998_{\pm 0.008}(0.010)$	$0.988_{\pm 0.016}(0.007)$	$0.426_{\pm 0.007}$
SSD **UA**0.5%, **RA**99.5%, **TA**94.2%	0.05	$0.996_{\pm 0.004}(0.055)$	$0.945_{\pm 0.002}(0.001)$	$0.999_{\pm 0.019}(0.090)$	$1.008_{\pm 0.011}(0.066)$	$0.994_{\pm 0.006}(0.130)$	$0.936_{\pm 0.014}(0.057)$	$0.622_{\pm 0.019}$
	0.1	$0.987_{\pm 0.003}(0.106)$	$0.902_{\pm 0.010}(0.007)$	$0.990_{\pm 0.003}(0.056)$	$0.926_{\pm 0.017}(0.021)$	$0.998_{\pm 0.020}(0.055)$	$0.973_{\pm 0.002}(0.028)$	$0.063_{\pm 0.022}$
	0.15	$0.967_{\pm 0.016}(0.147)$	$0.849_{\pm 0.009}(0.010)$	$0.965_{\pm 0.000}(0.124)$	$0.862_{\pm 0.012}(0.005)$	$1.002_{\pm 0.019}(0.027)$	$0.990_{\pm 0.002}(0.022)$	$0.007_{\pm 0.007}$
	0.2	$0.922_{\pm 0.006}(0.142)$	$0.803_{\pm 0.000}(0.005)$	$0.923_{\pm 0.009}(0.134)$	$0.811_{\pm 0.005}(0.013)$	$1.002_{\pm 0.020}(0.014)$	$0.992_{\pm 0.009}(0.011)$	$0.001_{\pm 0.005}$
NegGrad+ **UA**8.7%, **RA**98.8%, **TA**92.2%	0.05	$0.934_{\pm 0.007}(0.007)$	$0.948_{\pm 0.007}(0.004)$	$1.068_{\pm 0.017}(0.021)$	$1.086_{\pm 0.022}(0.012)$	$0.875_{\pm 0.008}(0.011)$	$0.873_{\pm 0.011}(0.006)$	$0.989_{\pm 0.013}$
	0.1	$0.895_{\pm 0.004}(0.014)$	$0.898_{\pm 0.008}(0.003)$	$0.964_{\pm 0.008}(0.030)$	$0.950_{\pm 0.013}(0.003)$	$0.928_{\pm 0.005}(0.015)$	$0.946_{\pm 0.005}(0.001)$	$0.044_{\pm 0.041}$
	0.15	$0.851_{\pm 0.013}(0.031)$	$0.851_{\pm 0.016}(0.012)$	$0.896_{\pm 0.016}(0.055)$	$0.876_{\pm 0.019}(0.009)$	$0.950_{\pm 0.003}(0.025)$	$0.971_{\pm 0.003}(0.003)$	$0.000_{\pm 0.000}$
	0.2	$0.800_{\pm 0.006}(0.020)$	$0.799_{\pm 0.001}(0.009)$	$0.832_{\pm 0.006}(0.043)$	$0.813_{\pm 0.001}(0.011)$	$0.961_{\pm 0.002}(0.027)$	$0.983_{\pm 0.001}(0.002)$	$0.000_{\pm 0.000}$
Salun **UA**3.7%, **RA**98.9%, **TA**91.8%	0.05	$0.987_{\pm 0.002}(0.046)$	$0.950_{\pm 0.001}(0.006)$	$1.132_{\pm 0.007}(0.043)$	$1.143_{\pm 0.002}(0.069)$	$0.872_{\pm 0.006}(0.008)$	$0.832_{\pm 0.003}(0.047)$	$0.867_{\pm 0.001}$
	0.1	$0.936_{\pm 0.010}(0.055)$	$0.896_{\pm 0.008}(0.001)$	$0.956_{\pm 0.012}(0.022)$	$0.954_{\pm 0.011}(0.007)$	$0.979_{\pm 0.003}(0.036)$	$0.939_{\pm 0.003}(0.006)$	$0.489_{\pm 0.029}$
	0.15	$0.871_{\pm 0.005}(0.051)$	$0.849_{\pm 0.008}(0.010)$	$0.881_{\pm 0.006}(0.040)$	$0.886_{\pm 0.016}(0.019)$	$0.989_{\pm 0.002}(0.014)$	$0.958_{\pm 0.002}(0.010)$	$0.314_{\pm 0.020}$
	0.2	$0.788_{\pm 0.010}(0.008)$	$0.794_{\pm 0.001}(0.014)$	$0.794_{\pm 0.010}(0.005)$	$0.821_{\pm 0.004}(0.003)$	$0.992_{\pm 0.001}(0.004)$	$0.966_{\pm 0.003}(0.015)$	$0.221_{\pm 0.005}$
SFRon **UA**4.8%, **RA**97.4%, **TA**91.4%	0.05	$0.977_{\pm 0.003}(0.036)$	$0.953_{\pm 0.004}(0.009)$	$1.100_{\pm 0.023}(0.011)$	$1.143_{\pm 0.021}(0.069)$	$0.889_{\pm 0.015}(0.025)$	$0.834_{\pm 0.012}(0.045)$	$0.926_{\pm 0.018}$
	0.1	$0.945_{\pm 0.004}(0.064)$	$0.905_{\pm 0.005}(0.010)$	$0.986_{\pm 0.005}(0.052)$	$0.977_{\pm 0.008}(0.030)$	$0.958_{\pm 0.001}(0.015)$	$0.927_{\pm 0.003}(0.018)$	$0.435_{\pm 0.043}$
	0.15	$0.895_{\pm 0.002}(0.075)$	$0.847_{\pm 0.002}(0.008)$	$0.912_{\pm 0.004}(0.071)$	$0.879_{\pm 0.001}(0.012)$	$0.982_{\pm 0.002}(0.007)$	$0.963_{\pm 0.003}(0.005)$	$0.082_{\pm 0.007}$
	0.2	$0.857_{\pm 0.008}(0.077)$	$0.808_{\pm 0.002}(0.000)$	$0.868_{\pm 0.007}(0.079)$	$0.826_{\pm 0.005}(0.002)$	$0.988_{\pm 0.002}(0.000)$	$0.978_{\pm 0.004}(0.003)$	$0.025_{\pm 0.005}$

Table 12: Unlearning performance of 9 unlearning methods on **CIFAR-10** with **ResNet18** in 50% **random data forgetting** scenario.

Methods	α	Coverage		Set Size		CR		
		$\mathcal{D}_f \downarrow$	$\mathcal{D}_{test} \uparrow$	$\mathcal{D}_f \uparrow$	$\mathcal{D}_{test} \downarrow$	$\mathcal{D}_f \downarrow$	$\mathcal{D}_{test} \uparrow$	$\hat{q}$
RT **UA**11.0%, **RA**99.8%, **TA**89.2%	0.05	$0.955_{\pm 0.004}(0.000)$	$0.947_{\pm 0.005}(0.000)$	$1.287_{\pm 0.001}(0.000)$	$1.214_{\pm 0.010}(0.000)$	$0.742_{\pm 0.005}(0.000)$	$0.780_{\pm 0.006}(0.000)$	$0.984_{\pm 0.002}$
	0.1	$0.898_{\pm 0.011}(0.000)$	$0.904_{\pm 0.010}(0.000)$	$1.023_{\pm 0.005}(0.000)$	$1.021_{\pm 0.003}(0.000)$	$0.878_{\pm 0.003}(0.000)$	$0.886_{\pm 0.003}(0.000)$	$0.650_{\pm 0.004}$
	0.15	$0.833_{\pm 0.007}(0.000)$	$0.847_{\pm 0.005}(0.000)$	$0.883_{\pm 0.002}(0.000)$	$0.906_{\pm 0.003}(0.000)$	$0.943_{\pm 0.010}(0.000)$	$0.934_{\pm 0.005}(0.000)$	$0.090_{\pm 0.004}$
	0.2	$0.782_{\pm 0.005}(0.000)$	$0.814_{\pm 0.004}(0.000)$	$0.812_{\pm 0.010}(0.000)$	$0.850_{\pm 0.009}(0.000)$	$0.964_{\pm 0.005}(0.000)$	$0.958_{\pm 0.003}(0.000)$	$0.018_{\pm 0.006}$
FT **UA**2.6%, **RA**99.1%, **TA**91.8%	0.05	$0.996_{\pm 0.000}(0.041)$	$0.952_{\pm 0.002}(0.005)$	$1.007_{\pm 0.000}(0.280)$	$1.029_{\pm 0.004}(0.185)$	$0.989_{\pm 0.001}(0.247)$	$0.925_{\pm 0.002}(0.145)$	$0.738_{\pm 0.014}$
	0.1	$0.975_{\pm 0.006}(0.077)$	$0.896_{\pm 0.013}(0.008)$	$0.976_{\pm 0.006}(0.047)$	$0.921_{\pm 0.017}(0.100)$	$0.999_{\pm 0.000}(0.121)$	$0.972_{\pm 0.004}(0.086)$	$0.081_{\pm 0.033}$
	0.15	$0.936_{\pm 0.004}(0.103)$	$0.854_{\pm 0.004}(0.007)$	$0.936_{\pm 0.004}(0.053)$	$0.867_{\pm 0.006}(0.039)$	$1.000_{\pm 0.000}(0.057)$	$0.985_{\pm 0.002}(0.051)$	$0.011_{\pm 0.002}$
	0.2	$0.859_{\pm 0.010}(0.077)$	$0.790_{\pm 0.010}(0.024)$	$0.859_{\pm 0.010}(0.047)$	$0.795_{\pm 0.011}(0.055)$	$1.000_{\pm 0.000}(0.036)$	$0.993_{\pm 0.001}(0.035)$	$0.001_{\pm 0.000}$
RL **UA**10.5%, **RA**93.9%, **TA**85.8%	0.05	$0.976_{\pm 0.001}(0.022)$	$0.949_{\pm 0.002}(0.002)$	$1.973_{\pm 0.396}(0.686)$	$1.971_{\pm 0.406}(0.757)$	$0.508_{\pm 0.100}(0.234)$	$0.495_{\pm 0.098}(0.285)$	$0.899_{\pm 0.012}$
	0.1	$0.942_{\pm 0.011}(0.043)$	$0.907_{\pm 0.009}(0.003)$	$1.227_{\pm 0.103}(0.204)$	$1.235_{\pm 0.107}(0.214)$	$0.771_{\pm 0.064}(0.107)$	$0.738_{\pm 0.064}(0.147)$	$0.837_{\pm 0.016}$
	0.15	$0.891_{\pm 0.013}(0.058)$	$0.856_{\pm 0.012}(0.009)$	$1.009_{\pm 0.047}(0.125)$	$1.011_{\pm 0.045}(0.105)$	$0.884_{\pm 0.039}(0.059)$	$0.847_{\pm 0.037}(0.087)$	$0.770_{\pm 0.022}$
	0.2	$0.834_{\pm 0.003}(0.051)$	$0.799_{\pm 0.005}(0.016)$	$0.897_{\pm 0.026}(0.086)$	$0.893_{\pm 0.025}(0.043)$	$0.929_{\pm 0.024}(0.034)$	$0.895_{\pm 0.022}(0.063)$	$0.713_{\pm 0.028}$
GA **UA**0.6%, **RA**99.5%, **TA**94.3%	0.05	$0.996_{\pm 0.000}(0.041)$	$0.945_{\pm 0.008}(0.002)$	$1.003_{\pm 0.007}(0.284)$	$1.005_{\pm 0.007}(0.209)$	$1.050_{\pm 0.007}(0.308)$	$0.945_{\pm 0.007}(0.165)$	$0.616_{\pm 0.008}$
	0.1	$0.985_{\pm 0.006}(0.087)$	$0.902_{\pm 0.009}(0.002)$	$0.989_{\pm 0.006}(0.034)$	$0.926_{\pm 0.006}(0.095)$	$1.095_{\pm 0.004}(0.217)$	$0.916_{\pm 0.006}(0.030)$	$0.057_{\pm 0.005}$
	0.15	$0.966_{\pm 0.006}(0.133)$	$0.848_{\pm 0.007}(0.001)$	$0.966_{\pm 0.002}(0.083)$	$0.857_{\pm 0.009}(0.049)$	$1.141_{\pm 0.001}(0.198)$	$0.879_{\pm 0.006}(0.055)$	$0.005_{\pm 0.007}$
	0.2	$0.929_{\pm 0.004}(0.147)$	$0.809_{\pm 0.007}(0.005)$	$0.932_{\pm 0.000}(0.120)$	$0.817_{\pm 0.005}(0.033)$	$1.150_{\pm 0.002}(0.186)$	$0.871_{\pm 0.001}(0.087)$	$0.001_{\pm 0.007}$
Teacher **UA**1.6%, **RA**98.3%, **TA**91.7%	0.05	$0.985_{\pm 0.015}(0.030)$	$0.944_{\pm 0.018}(0.003)$	$1.066_{\pm 0.003}(0.221)$	$1.143_{\pm 0.012}(0.071)$	$0.923_{\pm 0.010}(0.181)$	$0.823_{\pm 0.017}(0.043)$	$0.857_{\pm 0.013}$
	0.1	$0.949_{\pm 0.012}(0.051)$	$0.909_{\pm 0.016}(0.005)$	$0.970_{\pm 0.006}(0.053)$	$0.986_{\pm 0.014}(0.035)$	$0.980_{\pm 0.001}(0.102)$	$0.918_{\pm 0.009}(0.032)$	$0.834_{\pm 0.005}$
	0.15	$0.885_{\pm 0.010}(0.052)$	$0.849_{\pm 0.018}(0.002)$	$0.894_{\pm 0.017}(0.011)$	$0.893_{\pm 0.010}(0.013)$	$0.992_{\pm 0.002}(0.049)$	$0.950_{\pm 0.013}(0.016)$	$0.813_{\pm 0.013}$
	0.2	$0.818_{\pm 0.014}(0.036)$	$0.798_{\pm 0.014}(0.016)$	$0.823_{\pm 0.009}(0.011)$	$0.826_{\pm 0.002}(0.024)$	$0.997_{\pm 0.015}(0.033)$	$0.971_{\pm 0.007}(0.013)$	$0.793_{\pm 0.012}$
SSD **UA**0.5%, **RA**99.5%, **TA**94.3%	0.05	$0.993_{\pm 0.005}(0.038)$	$0.944_{\pm 0.011}(0.003)$	$0.999_{\pm 0.007}(0.288)$	$1.001_{\pm 0.009}(0.213)$	$0.995_{\pm 0.009}(0.253)$	$0.941_{\pm 0.013}(0.161)$	$0.585_{\pm 0.014}$
	0.1	$0.991_{\pm 0.015}(0.093)$	$0.904_{\pm 0.014}(0.000)$	$0.991_{\pm 0.001}(0.032)$	$0.929_{\pm 0.011}(0.092)$	$1.000_{\pm 0.011}(0.122)$	$0.975_{\pm 0.010}(0.089)$	$0.060_{\pm 0.011}$
	0.15	$0.964_{\pm 0.016}(0.131)$	$0.850_{\pm 0.011}(0.003)$	$0.967_{\pm 0.009}(0.084)$	$0.860_{\pm 0.014}(0.046)$	$1.000_{\pm 0.001}(0.057)$	$0.988_{\pm 0.003}(0.054)$	$0.005_{\pm 0.010}$
	0.2	$0.930_{\pm 0.018}(0.148)$	$0.807_{\pm 0.002}(0.007)$	$0.929_{\pm 0.002}(0.117)$	$0.814_{\pm 0.017}(0.036)$	$1.000_{\pm 0.003}(0.036)$	$0.992_{\pm 0.001}(0.034)$	$0.002_{\pm 0.005}$
NegGrad+ **UA**2.8%, **RA**99.6%, **TA**92.9%	0.05	$0.986_{\pm 0.000}(0.031)$	$0.949_{\pm 0.001}(0.001)$	$1.039_{\pm 0.008}(0.248)$	$1.062_{\pm 0.011}(0.152)$	$0.949_{\pm 0.008}(0.207)$	$0.893_{\pm 0.008}(0.113)$	$0.855_{\pm 0.028}$
	0.1	$0.951_{\pm 0.005}(0.053)$	$0.903_{\pm 0.004}(0.001)$	$0.964_{\pm 0.008}(0.059)$	$0.944_{\pm 0.010}(0.076)$	$0.987_{\pm 0.003}(0.109)$	$0.956_{\pm 0.007}(0.070)$	$0.177_{\pm 0.055}$
	0.15	$0.889_{\pm 0.004}(0.056)$	$0.845_{\pm 0.003}(0.002)$	$0.892_{\pm 0.004}(0.009)$	$0.861_{\pm 0.003}(0.045)$	$0.996_{\pm 0.000}(0.053)$	$0.981_{\pm 0.001}(0.047)$	$0.012_{\pm 0.002}$
	0.2	$0.825_{\pm 0.003}(0.043)$	$0.796_{\pm 0.004}(0.018)$	$0.827_{\pm 0.003}(0.015)$	$0.805_{\pm 0.004}(0.045)$	$0.999_{\pm 0.000}(0.035)$	$0.989_{\pm 0.000}(0.032)$	$0.002_{\pm 0.000}$
Salun **UA**4.3%, **RA**97.7%, **TA**89.4%	0.05	$0.988_{\pm 0.001}(0.034)$	$0.951_{\pm 0.003}(0.004)$	$1.314_{\pm 0.113}(0.027)$	$1.381_{\pm 0.121}(0.167)$	$0.756_{\pm 0.064}(0.014)$	$0.692_{\pm 0.058}(0.088)$	$0.871_{\pm 0.013}$
	0.1	$0.956_{\pm 0.003}(0.058)$	$0.897_{\pm 0.005}(0.007)$	$1.015_{\pm 0.003}(0.008)$	$1.021_{\pm 0.001}(0.001)$	$0.941_{\pm 0.006}(0.064)$	$0.878_{\pm 0.004}(0.007)$	$0.776_{\pm 0.002}$
	0.15	$0.910_{\pm 0.005}(0.078)$	$0.847_{\pm 0.006}(0.000)$	$0.937_{\pm 0.009}(0.054)$	$0.916_{\pm 0.008}(0.010)$	$0.972_{\pm 0.004}(0.029)$	$0.924_{\pm 0.003}(0.010)$	$0.714_{\pm 0.010}$
	0.2	$0.856_{\pm 0.008}(0.074)$	$0.796_{\pm 0.010}(0.019)$	$0.872_{\pm 0.008}(0.060)$	$0.844_{\pm 0.008}(0.006)$	$0.982_{\pm 0.003}(0.019)$	$0.943_{\pm 0.004}(0.015)$	$0.669_{\pm 0.008}$
SFRon **UA**4.0%, **RA**97.3%, **TA**91.6%	0.05	$0.977_{\pm 0.003}(0.022)$	$0.953_{\pm 0.004}(0.006)$	$1.100_{\pm 0.023}(0.188)$	$1.143_{\pm 0.021}(0.071)$	$0.889_{\pm 0.015}(0.147)$	$0.834_{\pm 0.012}(0.054)$	$0.926_{\pm 0.018}$
	0.1	$0.945_{\pm 0.004}(0.047)$	$0.905_{\pm 0.005}(0.001)$	$0.986_{\pm 0.005}(0.037)$	$0.977_{\pm 0.008}(0.044)$	$0.958_{\pm 0.001}(0.081)$	$0.927_{\pm 0.003}(0.042)$	$0.435_{\pm 0.043}$
	0.15	$0.895_{\pm 0.002}(0.062)$	$0.847_{\pm 0.002}(0.000)$	$0.912_{\pm 0.004}(0.029)$	$0.879_{\pm 0.001}(0.027)$	$0.982_{\pm 0.002}(0.039)$	$0.963_{\pm 0.003}(0.029)$	$0.082_{\pm 0.007}$
	0.2	$0.857_{\pm 0.008}(0.075)$	$0.808_{\pm 0.002}(0.006)$	$0.868_{\pm 0.007}(0.056)$	$0.826_{\pm 0.005}(0.024)$	$0.988_{\pm 0.002}(0.024)$	$0.978_{\pm 0.004}(0.020)$	$0.025_{\pm 0.005}$

Table 13: Unlearning performance of 9 unlearning methods on **CIFAR-10** with **ResNet18** in **class-wise forgetting** scenario.

Methods	α	Coverage $\mathcal{D}_f \downarrow$	Coverage $\mathcal{D}_f \downarrow$	Coverage $\mathcal{D}_r \uparrow$	Set Size $\mathcal{D}_f \uparrow$	Set Size $\mathcal{D}_f \uparrow$	Set Size $\mathcal{D}_r \downarrow$	CR $\mathcal{D}_f \downarrow$	CR $\mathcal{D}_f \downarrow$	CR $\mathcal{D}_r \uparrow$	$\hat{q}_f$	$\hat{q}_{test}$
RT	0.05	$1.000_{\pm 0.001}(0.000)$	$1.000_{\pm 0.001}(0.000)$	$0.964_{\pm 0.006}(0.000)$	$10.000_{\pm 0.000}(0.000)$	$10.000_{\pm 0.000}(0.000)$	$1.148_{\pm 0.013}(0.000)$	$0.100_{\pm 0.000}(0.000)$	$0.100_{\pm 0.000}(0.000)$	$0.840_{\pm 0.002}(0.000)$	$1.000_{\pm 0.000}$	$0.982_{\pm 0.003}$
UA100%, $\mathbf{UA}_f$100%,	0.1	$1.000_{\pm 0.000}(0.000)$	$1.000_{\pm 0.001}(0.000)$	$0.882_{\pm 0.011}(0.000)$	$10.000_{\pm 0.000}(0.000)$	$10.000_{\pm 0.000}(0.000)$	$0.922_{\pm 0.009}(0.000)$	$0.100_{\pm 0.000}(0.000)$	$0.100_{\pm 0.001}(0.000)$	$0.956_{\pm 0.007}(0.000)$	$1.000_{\pm 0.001}$	$0.080_{\pm 0.003}$
RA99.9%, TA92.4%	0.15	$1.000_{\pm 0.000}(0.000)$	$1.000_{\pm 0.000}(0.000)$	$0.856_{\pm 0.012}(0.000)$	$10.000_{\pm 0.000}(0.000)$	$10.000_{\pm 0.000}(0.000)$	$0.882_{\pm 0.007}(0.000)$	$0.100_{\pm 0.001}(0.000)$	$0.100_{\pm 0.001}(0.000)$	$0.970_{\pm 0.004}(0.000)$	$1.000_{\pm 0.000}$	$0.018_{\pm 0.010}$
	0.2	$1.000_{\pm 0.000}(0.000)$	$1.000_{\pm 0.000}(0.000)$	$0.814_{\pm 0.010}(0.000)$	$10.000_{\pm 0.000}(0.000)$	$10.000_{\pm 0.000}(0.000)$	$0.830_{\pm 0.001}(0.000)$	$0.100_{\pm 0.001}(0.000)$	$0.100_{\pm 0.001}(0.000)$	$0.981_{\pm 0.002}(0.000)$	$1.000_{\pm 0.000}$	$0.003_{\pm 0.001}$
FT	0.05	$0.994_{\pm 0.003}(0.006)$	$0.962_{\pm 0.022}(0.038)$	$0.944_{\pm 0.011}(0.020)$	$9.854_{\pm 0.127}(0.146)$	$9.403_{\pm 0.501}(0.597)$	$1.045_{\pm 0.040}(0.103)$	$0.101_{\pm 0.001}(0.001)$	$0.102_{\pm 0.003}(0.002)$	$0.904_{\pm 0.028}(0.065)$	$1.000_{\pm 0.000}$	$0.731_{\pm 0.166}$
UA100%, $\mathbf{UA}_f$100%,	0.1	$0.969_{\pm 0.011}(0.031)$	$0.882_{\pm 0.020}(0.118)$	$0.908_{\pm 0.016}(0.026)$	$9.495_{\pm 0.255}(0.505)$	$8.528_{\pm 0.571}(1.472)$	$0.956_{\pm 0.006}(0.034)$	$0.102_{\pm 0.002}(0.002)$	$0.104_{\pm 0.005}(0.004)$	$0.950_{\pm 0.007}(0.006)$	$1.000_{\pm 0.000}$	$0.314_{\pm 0.010}$
RA96.7%, TA90.8%	0.15	$0.951_{\pm 0.014}(0.049)$	$0.840_{\pm 0.011}(0.160)$	$0.840_{\pm 0.033}(0.005)$	$9.265_{\pm 0.279}(0.735)$	$8.131_{\pm 0.523}(1.869)$	$0.872_{\pm 0.039}(0.010)$	$0.103_{\pm 0.003}(0.003)$	$0.103_{\pm 0.007}(0.003)$	$0.976_{\pm 0.009}(0.006)$	$1.000_{\pm 0.000}$	$0.073_{\pm 0.054}$
	0.2	$0.942_{\pm 0.014}(0.058)$	$0.818_{\pm 0.072}(0.182)$	$0.838_{\pm 0.016}(0.023)$	$9.163_{\pm 0.245}(0.837)$	$7.934_{\pm 0.533}(2.066)$	$0.854_{\pm 0.019}(0.024)$	$0.103_{\pm 0.003}(0.003)$	$0.103_{\pm 0.010}(0.003)$	$0.981_{\pm 0.005}(0.000)$	$1.000_{\pm 0.000}$	$0.039_{\pm 0.017}$
RL	0.05	$0.995_{\pm 0.002}(0.005)$	$0.954_{\pm 0.009}(0.046)$	$0.959_{\pm 0.015}(0.005)$	$9.993_{\pm 0.003}(0.007)$	$9.900_{\pm 0.011}(0.100)$	$1.170_{\pm 0.155}(0.022)$	$0.100_{\pm 0.000}(0.000)$	$0.096_{\pm 0.001}(0.004)$	$0.828_{\pm 0.097}(0.012)$	$1.000_{\pm 0.000}$	$0.870_{\pm 0.145}$
UA100%, $\mathbf{UA}_f$100%,	0.1	$0.984_{\pm 0.003}(0.016)$	$0.907_{\pm 0.015}(0.093)$	$0.918_{\pm 0.021}(0.036)$	$9.978_{\pm 0.004}(0.022)$	$9.800_{\pm 0.015}(0.200)$	$0.982_{\pm 0.036}(0.059)$	$0.099_{\pm 0.000}(0.001)$	$0.093_{\pm 0.002}(0.007)$	$0.936_{\pm 0.022}(0.021)$	$1.000_{\pm 0.000}$	$0.469_{\pm 0.250}$
RA98.0%, TA92.7%	0.15	$0.961_{\pm 0.009}(0.039)$	$0.859_{\pm 0.014}(0.141)$	$0.870_{\pm 0.019}(0.014)$	$9.950_{\pm 0.017}(0.050)$	$9.700_{\pm 0.066}(0.300)$	$0.904_{\pm 0.045}(0.021)$	$0.097_{\pm 0.001}(0.003)$	$0.089_{\pm 0.001}(0.011)$	$0.964_{\pm 0.027}(0.006)$	$1.000_{\pm 0.000}$	$0.144_{\pm 0.163}$
	0.2	$0.935_{\pm 0.027}(0.065)$	$0.815_{\pm 0.012}(0.185)$	$0.804_{\pm 0.016}(0.010)$	$9.919_{\pm 0.035}(0.081)$	$9.637_{\pm 0.076}(0.363)$	$0.820_{\pm 0.026}(0.010)$	$0.094_{\pm 0.002}(0.006)$	$0.085_{\pm 0.001}(0.015)$	$0.981_{\pm 0.012}(0.000)$	$0.999_{\pm 0.001}$	$0.014_{\pm 0.013}$
GA	0.05	$1.000_{\pm 0.003}(0.000)$	$1.000_{\pm 0.005}(0.000)$	$0.948_{\pm 0.006}(0.016)$	$10.000_{\pm 0.009}(0.000)$	$10.000_{\pm 0.003}(0.000)$	$1.204_{\pm 0.002}(0.056)$	$0.100_{\pm 0.007}(0.000)$	$0.100_{\pm 0.011}(0.000)$	$0.787_{\pm 0.011}(0.053)$	$1.000_{\pm 0.010}$	$0.988_{\pm 0.000}$
	0.1	$1.000_{\pm 0.003}(0.000)$	$1.000_{\pm 0.010}(0.000)$	$0.899_{\pm 0.006}(0.017)$	$10.000_{\pm 0.005}(0.000)$	$10.000_{\pm 0.006}(0.000)$	$1.005_{\pm 0.013}(0.083)$	$0.100_{\pm 0.015}(0.000)$	$0.100_{\pm 0.006}(0.000)$	$0.894_{\pm 0.002}(0.062)$	$1.000_{\pm 0.000}$	$0.562_{\pm 0.003}$
UA84.6%, $\mathbf{UA}_f$82.5%,	0.15	$1.000_{\pm 0.006}(0.000)$	$1.000_{\pm 0.001}(0.000)$	$0.843_{\pm 0.011}(0.013)$	$10.000_{\pm 0.005}(0.000)$	$10.000_{\pm 0.000}(0.000)$	$0.893_{\pm 0.010}(0.011)$	$0.100_{\pm 0.004}(0.000)$	$0.100_{\pm 0.008}(0.000)$	$0.944_{\pm 0.007}(0.026)$	$1.000_{\pm 0.001}$	$0.051_{\pm 0.002}$
RA96.4%, TA89.6%	0.2	$0.828_{\pm 0.003}(0.172)$	$0.782_{\pm 0.011}(0.218)$	$0.838_{\pm 0.010}(0.024)$	$9.550_{\pm 0.007}(0.450)$	$9.366_{\pm 0.002}(0.634)$	$0.884_{\pm 0.000}(0.054)$	$0.087_{\pm 0.008}(0.013)$	$0.084_{\pm 0.005}(0.016)$	$0.948_{\pm 0.010}(0.033)$	$1.000_{\pm 0.002}$	$0.038_{\pm 0.003}$
Teacher	0.05	$0.994_{\pm 0.003}(0.006)$	$0.959_{\pm 0.002}(0.041)$	$0.939_{\pm 0.003}(0.025)$	$9.877_{\pm 0.000}(0.123)$	$9.502_{\pm 0.003}(0.498)$	$1.000_{\pm 0.004}(0.148)$	$0.101_{\pm 0.004}(0.001)$	$0.101_{\pm 0.004}(0.001)$	$0.939_{\pm 0.001}(0.099)$	$0.955_{\pm 0.005}$	$0.588_{\pm 0.004}$
UA90.1%, $\mathbf{UA}_f$86.5%,	0.1	$0.931_{\pm 0.000}(0.069)$	$0.904_{\pm 0.001}(0.096)$	$0.890_{\pm 0.001}(0.008)$	$9.199_{\pm 0.002}(0.801)$	$8.604_{\pm 0.004}(1.396)$	$0.914_{\pm 0.004}(0.008)$	$0.101_{\pm 0.004}(0.001)$	$0.105_{\pm 0.004}(0.005)$	$0.974_{\pm 0.003}(0.018)$	$0.926_{\pm 0.004}$	$0.116_{\pm 0.005}$
RA99.5%, TA94.0%	0.15	$0.879_{\pm 0.004}(0.121)$	$0.881_{\pm 0.001}(0.119)$	$0.834_{\pm 0.001}(0.022)$	$8.730_{\pm 0.002}(1.270)$	$8.081_{\pm 0.001}(1.919)$	$0.845_{\pm 0.005}(0.037)$	$0.101_{\pm 0.004}(0.001)$	$0.109_{\pm 0.002}(0.009)$	$0.986_{\pm 0.004}(0.016)$	$0.921_{\pm 0.001}$	$0.017_{\pm 0.002}$
	0.2	$0.809_{\pm 0.004}(0.191)$	$0.841_{\pm 0.004}(0.159)$	$0.816_{\pm 0.000}(0.002)$	$8.141_{\pm 0.003}(1.859)$	$7.525_{\pm 0.003}(2.475)$	$0.824_{\pm 0.003}(0.006)$	$0.099_{\pm 0.002}(0.001)$	$0.112_{\pm 0.003}(0.012)$	$0.990_{\pm 0.002}(0.009)$	$0.916_{\pm 0.005}$	$0.010_{\pm 0.003}$
SSD	0.05	$0.995_{\pm 0.014}(0.005)$	$0.935_{\pm 0.013}(0.065)$	$0.940_{\pm 0.007}(0.024)$	$1.030_{\pm 0.014}(8.970)$	$1.067_{\pm 0.013}(8.933)$	$0.991_{\pm 0.011}(0.157)$	$0.966_{\pm 0.010}(0.866)$	$0.876_{\pm 0.007}(0.776)$	$0.949_{\pm 0.010}(0.109)$	$0.804_{\pm 0.015}$	$0.447_{\pm 0.007}$
UA1.16%, $\mathbf{UA}_f$7.75%,	0.1	$0.984_{\pm 0.021}(0.016)$	$0.910_{\pm 0.009}(0.090)$	$0.880_{\pm 0.001}(0.002)$	$0.992_{\pm 0.011}(9.008)$	$0.982_{\pm 0.005}(9.018)$	$0.896_{\pm 0.003}(0.026)$	$0.992_{\pm 0.004}(0.892)$	$0.926_{\pm 0.017}(0.826)$	$0.981_{\pm 0.012}(0.025)$	$0.434_{\pm 0.007}$	$0.022_{\pm 0.005}$
RA99.5%, TA94.3%	0.15	$0.960_{\pm 0.012}(0.040)$	$0.876_{\pm 0.011}(0.124)$	$0.847_{\pm 0.007}(0.009)$	$0.962_{\pm 0.007}(9.038)$	$0.931_{\pm 0.006}(9.069)$	$0.857_{\pm 0.013}(0.025)$	$0.998_{\pm 0.016}(0.898)$	$0.941_{\pm 0.002}(0.841)$	$0.989_{\pm 0.002}(0.019)$	$0.215_{\pm 0.007}$	$0.005_{\pm 0.017}$
	0.2	$0.895_{\pm 0.020}(0.105)$	$0.816_{\pm 0.010}(0.184)$	$0.823_{\pm 0.015}(0.009)$	$0.895_{\pm 0.014}(9.105)$	$0.850_{\pm 0.004}(9.150)$	$0.831_{\pm 0.002}(0.001)$	$0.999_{\pm 0.001}(0.899)$	$0.960_{\pm 0.014}(0.860)$	$0.991_{\pm 0.003}(0.010)$	$0.078_{\pm 0.003}$	$0.002_{\pm 0.009}$
NegGrad+	0.05	$0.989_{\pm 0.016}(0.011)$	$0.961_{\pm 0.056}(0.039)$	$0.945_{\pm 0.027}(0.019)$	$9.432_{\pm 0.803}(0.568)$	$9.038_{\pm 1.360}(0.962)$	$1.053_{\pm 0.020}(0.096)$	$0.105_{\pm 0.007}(0.005)$	$0.107_{\pm 0.010}(0.007)$	$0.897_{\pm 0.008}(0.058)$	$1.000_{\pm 0.000}$	$0.835_{\pm 0.085}$
UA96.2%, $\mathbf{UA}_f$95.2%,	0.1	$0.980_{\pm 0.029}(0.020)$	$0.954_{\pm 0.065}(0.046)$	$0.881_{\pm 0.028}(0.001)$	$9.250_{\pm 1.061}(0.750)$	$8.836_{\pm 1.647}(1.164)$	$0.913_{\pm 0.018}(0.009)$	$0.106_{\pm 0.009}(0.006)$	$0.109_{\pm 0.013}(0.009)$	$0.965_{\pm 0.012}(0.009)$	$1.000_{\pm 0.000}$	$0.057_{\pm 0.021}$
RA97.6%, TA92.8%	0.15	$0.952_{\pm 0.068}(0.048)$	$0.908_{\pm 0.130}(0.092)$	$0.849_{\pm 0.026}(0.007)$	$8.600_{\pm 1.980}(1.400)$	$8.077_{\pm 2.719}(1.923)$	$0.868_{\pm 0.016}(0.014)$	$0.113_{\pm 0.018}(0.013)$	$0.116_{\pm 0.023}(0.016)$	$0.977_{\pm 0.012}(0.007)$	$1.000_{\pm 0.000}$	$0.012_{\pm 0.003}$
	0.2	$0.958_{\pm 0.060}(0.042)$	$0.921_{\pm 0.111}(0.079)$	$0.814_{\pm 0.007}(0.001)$	$8.673_{\pm 1.327}(1.327)$	$8.219_{\pm 2.516}(1.781)$	$0.828_{\pm 0.026}(0.002)$	$0.112_{\pm 0.017}(0.012)$	$0.115_{\pm 0.022}(0.015)$	$0.983_{\pm 0.015}(0.001)$	$1.000_{\pm 0.000}$	$0.004_{\pm 0.003}$
Salun	0.05	$0.996_{\pm 0.001}(0.004)$	$0.941_{\pm 0.008}(0.059)$	$0.952_{\pm 0.001}(0.012)$	$9.996_{\pm 0.002}(0.004)$	$9.892_{\pm 0.003}(0.108)$	$1.028_{\pm 0.008}(0.121)$	$0.100_{\pm 0.000}(0.000)$	$0.095_{\pm 0.001}(0.005)$	$0.926_{\pm 0.008}(0.087)$	$1.000_{\pm 0.000}$	$0.785_{\pm 0.049}$
UA100%, $\mathbf{UA}_f$100%,	0.1	$0.988_{\pm 0.004}(0.012)$	$0.906_{\pm 0.011}(0.094)$	$0.901_{\pm 0.002}(0.020)$	$9.985_{\pm 0.003}(0.015)$	$9.817_{\pm 0.045}(0.183)$	$0.928_{\pm 0.006}(0.006)$	$0.099_{\pm 0.000}(0.001)$	$0.092_{\pm 0.001}(0.008)$	$0.971_{\pm 0.004}(0.015)$	$1.000_{\pm 0.000}$	$0.042_{\pm 0.011}$
RA99.6%, TA94.3%	0.15	$0.960_{\pm 0.003}(0.040)$	$0.851_{\pm 0.005}(0.149)$	$0.878_{\pm 0.006}(0.022)$	$9.952_{\pm 0.000}(0.048)$	$9.677_{\pm 0.088}(0.323)$	$0.896_{\pm 0.005}(0.013)$	$0.096_{\pm 0.000}(0.004)$	$0.088_{\pm 0.000}(0.012)$	$0.980_{\pm 0.001}(0.010)$	$1.000_{\pm 0.000}$	$0.009_{\pm 0.001}$
	0.2	$0.915_{\pm 0.019}(0.085)$	$0.807_{\pm 0.038}(0.193)$	$0.820_{\pm 0.035}(0.005)$	$9.893_{\pm 0.024}(0.107)$	$9.511_{\pm 0.192}(0.489)$	$0.828_{\pm 0.039}(0.002)$	$0.092_{\pm 0.002}(0.008)$	$0.085_{\pm 0.002}(0.015)$	$0.990_{\pm 0.004}(0.009)$	$1.000_{\pm 0.000}$	$0.001_{\pm 0.001}$
SFRon	0.05	$1.000_{\pm 0.000}(0.000)$	$1.000_{\pm 0.000}(0.000)$	$0.952_{\pm 0.005}(0.013)$	$10.000_{\pm 0.000}(0.000)$	$10.000_{\pm 0.000}(0.000)$	$1.022_{\pm 0.030}(0.127)$	$0.100_{\pm 0.000}(0.000)$	$0.100_{\pm 0.000}(0.000)$	$0.932_{\pm 0.024}(0.092)$	$1.000_{\pm 0.000}$	$0.677_{\pm 0.206}$
UA100%, $\mathbf{UA}_f$100%,	0.1	$1.000_{\pm 0.000}(0.000)$	$1.000_{\pm 0.000}(0.000)$	$0.908_{\pm 0.013}(0.026)$	$10.000_{\pm 0.000}(0.000)$	$10.000_{\pm 0.000}(0.000)$	$0.937_{\pm 0.028}(0.014)$	$0.100_{\pm 0.000}(0.000)$	$0.100_{\pm 0.000}(0.000)$	$0.970_{\pm 0.015}(0.014)$	$1.000_{\pm 0.000}$	$0.089_{\pm 0.092}$
RA99.3%, TA94.4%	0.15	$1.000_{\pm 0.000}(0.000)$	$1.000_{\pm 0.000}(0.000)$	$0.840_{\pm 0.026}(0.016)$	$10.000_{\pm 0.000}(0.000)$	$10.000_{\pm 0.000}(0.000)$	$0.849_{\pm 0.026}(0.033)$	$0.100_{\pm 0.000}(0.000)$	$0.100_{\pm 0.000}(0.000)$	$0.989_{\pm 0.003}(0.019)$	$1.000_{\pm 0.000}$	$0.002_{\pm 0.001}$
	0.2	$1.000_{\pm 0.000}(0.000)$	$1.000_{\pm 0.000}(0.000)$	$0.807_{\pm 0.024}(0.008)$	$10.000_{\pm 0.000}(0.000)$	$10.000_{\pm 0.000}(0.000)$	$0.813_{\pm 0.025}(0.017)$	$0.100_{\pm 0.000}(0.000)$	$0.100_{\pm 0.000}(0.000)$	$0.992_{\pm 0.003}(0.010)$	$1.000_{\pm 0.000}$	$0.001_{\pm 0.001}$

Table 14: Unlearning performance of 9 unlearning methods on **Tiny ImageNet** with **ViT** in 10% **random data forgetting** scenario.

Methods	α	Coverage $\mathcal{D}_f \downarrow$	Coverage $\mathcal{D}_{test} \uparrow$	Set Size $\mathcal{D}_f \uparrow$	Set Size $\mathcal{D}_{test} \downarrow$	CR $\mathcal{D}_f \downarrow$	CR $\mathcal{D}_{test} \uparrow$	$\hat{q}$
RT	0.05	$0.944_{\pm 0.006}(0.000)$	$0.949_{\pm 0.026}(0.000)$	$1.876_{\pm 0.009}(0.000)$	$1.840_{\pm 0.014}(0.000)$	$0.503_{\pm 0.018}(0.000)$	$0.516_{\pm 0.018}(0.000)$	$0.984_{\pm 0.002}$
	0.1	$0.892_{\pm 0.006}(0.000)$	$0.900_{\pm 0.025}(0.000)$	$1.151_{\pm 0.002}(0.000)$	$1.144_{\pm 0.018}(0.000)$	$0.775_{\pm 0.016}(0.000)$	$0.786_{\pm 0.026}(0.000)$	$0.853_{\pm 0.003}$
UA14.7%, RA98.8%, TA86.0%	0.15	$0.841_{\pm 0.024}(0.000)$	$0.850_{\pm 0.017}(0.000)$	$0.956_{\pm 0.014}(0.000)$	$0.956_{\pm 0.017}(0.000)$	$0.880_{\pm 0.014}(0.000)$	$0.889_{\pm 0.019}(0.000)$	$0.539_{\pm 0.001}$
	0.2	$0.790_{\pm 0.015}(0.000)$	$0.799_{\pm 0.023}(0.000)$	$0.846_{\pm 0.004}(0.000)$	$0.854_{\pm 0.014}(0.000)$	$0.934_{\pm 0.012}(0.000)$	$0.935_{\pm 0.015}(0.000)$	$0.238_{\pm 0.012}$
FT	0.05	$0.994_{\pm 0.005}(0.050)$	$0.950_{\pm 0.019}(0.001)$	$2.133_{\pm 0.008}(0.257)$	$2.440_{\pm 0.011}(0.600)$	$0.466_{\pm 0.009}(0.037)$	$0.389_{\pm 0.016}(0.127)$	$0.994_{\pm 0.020}$
	0.1	$0.978_{\pm 0.007}(0.086)$	$0.903_{\pm 0.003}(0.003)$	$1.234_{\pm 0.010}(0.083)$	$1.317_{\pm 0.001}(0.173)$	$0.792_{\pm 0.018}(0.017)$	$0.685_{\pm 0.001}(0.101)$	$0.935_{\pm 0.012}$
UA6.9%, RA97.9%, TA84.1%	0.15	$0.938_{\pm 0.001}(0.097)$	$0.851_{\pm 0.010}(0.001)$	$1.014_{\pm 0.005}(0.058)$	$1.017_{\pm 0.016}(0.061)$	$0.925_{\pm 0.007}(0.045)$	$0.836_{\pm 0.016}(0.053)$	$0.681_{\pm 0.003}$
	0.2	$0.888_{\pm 0.009}(0.098)$	$0.801_{\pm 0.012}(0.002)$	$0.915_{\pm 0.006}(0.069)$	$0.885_{\pm 0.000}(0.031)$	$0.970_{\pm 0.020}(0.036)$	$0.905_{\pm 0.005}(0.030)$	$0.326_{\pm 0.011}$
RL	0.05	$0.969_{\pm 0.021}(0.025)$	$0.952_{\pm 0.008}(0.003)$	$17.890_{\pm 0.003}(16.014)$	$8.572_{\pm 0.010}(6.732)$	$0.054_{\pm 0.013}(0.449)$	$0.111_{\pm 0.002}(0.405)$	$0.996_{\pm 0.019}$
	0.1	$0.892_{\pm 0.017}(0.000)$	$0.902_{\pm 0.013}(0.002)$	$2.639_{\pm 0.017}(1.488)$	$1.843_{\pm 0.019}(0.699)$	$0.338_{\pm 0.022}(0.437)$	$0.489_{\pm 0.013}(0.297)$	$0.971_{\pm 0.014}$
UA26.9%, RA96.0%, TA81.4%	0.15	$0.793_{\pm 0.021}(0.048)$	$0.855_{\pm 0.008}(0.005)$	$1.225_{\pm 0.013}(0.269)$	$1.164_{\pm 0.000}(0.208)$	$0.648_{\pm 0.002}(0.232)$	$0.734_{\pm 0.000}(0.155)$	$0.894_{\pm 0.022}$
	0.2	$0.681_{\pm 0.010}(0.109)$	$0.803_{\pm 0.003}(0.004)$	$0.831_{\pm 0.006}(0.015)$	$0.946_{\pm 0.011}(0.092)$	$0.820_{\pm 0.022}(0.114)$	$0.849_{\pm 0.006}(0.086)$	$0.715_{\pm 0.013}$
GA	0.05	$0.996_{\pm 0.003}(0.052)$	$0.947_{\pm 0.002}(0.002)$	$1.539_{\pm 0.004}(0.337)$	$2.018_{\pm 0.007}(0.178)$	$0.647_{\pm 0.003}(0.144)$	$0.469_{\pm 0.002}(0.047)$	$0.988_{\pm 0.004}$
	0.1	$0.986_{\pm 0.006}(0.094)$	$0.900_{\pm 0.000}(0.000)$	$1.104_{\pm 0.006}(0.047)$	$1.224_{\pm 0.005}(0.080)$	$0.894_{\pm 0.003}(0.119)$	$0.736_{\pm 0.006}(0.050)$	$0.899_{\pm 0.001}$
UA3.2%, RA97.4%, TA84.9%	0.15	$0.967_{\pm 0.002}(0.126)$	$0.852_{\pm 0.005}(0.002)$	$1.003_{\pm 0.008}(0.047)$	$0.993_{\pm 0.004}(0.037)$	$0.964_{\pm 0.005}(0.084)$	$0.859_{\pm 0.006}(0.030)$	$0.632_{\pm 0.009}$
	0.2	$0.934_{\pm 0.001}(0.144)$	$0.800_{\pm 0.007}(0.001)$	$0.946_{\pm 0.008}(0.100)$	$0.871_{\pm 0.008}(0.017)$	$0.987_{\pm 0.008}(0.053)$	$0.919_{\pm 0.005}(0.016)$	$0.296_{\pm 0.009}$
Teacher	0.05	$0.977_{\pm 0.004}(0.033)$	$0.956_{\pm 0.003}(0.007)$	$5.473_{\pm 0.006}(3.597)$	$5.080_{\pm 0.004}(3.240)$	$0.179_{\pm 0.008}(0.324)$	$0.188_{\pm 0.002}(0.328)$	$0.987_{\pm 0.008}$
	0.1	$0.930_{\pm 0.003}(0.038)$	$0.902_{\pm 0.008}(0.002)$	$1.991_{\pm 0.004}(0.840)$	$1.959_{\pm 0.002}(0.815)$	$0.467_{\pm 0.004}(0.308)$	$0.460_{\pm 0.002}(0.326)$	$0.971_{\pm 0.007}$
UA17.3%, RA86.7%, TA79.0%	0.15	$0.873_{\pm 0.003}(0.032)$	$0.850_{\pm 0.009}(0.000)$	$1.295_{\pm 0.006}(0.339)$	$1.319_{\pm 0.005}(0.363)$	$0.674_{\pm 0.007}(0.206)$	$0.645_{\pm 0.003}(0.244)$	$0.944_{\pm 0.006}$
	0.2	$0.816_{\pm 0.007}(0.026)$	$0.803_{\pm 0.009}(0.004)$	$1.020_{\pm 0.006}(0.174)$	$1.058_{\pm 0.004}(0.204)$	$0.800_{\pm 0.005}(0.134)$	$0.758_{\pm 0.005}(0.177)$	$0.910_{\pm 0.006}$
SSD	0.05	$0.998_{\pm 0.004}(0.054)$	$0.950_{\pm 0.006}(0.001)$	$1.354_{\pm 0.008}(0.522)$	$1.827_{\pm 0.002}(0.013)$	$0.737_{\pm 0.008}(0.234)$	$0.520_{\pm 0.008}(0.004)$	$0.985_{\pm 0.005}$
	0.1	$0.993_{\pm 0.008}(0.101)$	$0.897_{\pm 0.008}(0.003)$	$1.039_{\pm 0.002}(0.112)$	$1.134_{\pm 0.008}(0.010)$	$0.956_{\pm 0.007}(0.181)$	$0.791_{\pm 0.002}(0.005)$	$0.852_{\pm 0.001}$
UA1.5%, RA98.5%, TA86.1%	0.15	$0.981_{\pm 0.005}(0.140)$	$0.853_{\pm 0.001}(0.003)$	$0.993_{\pm 0.001}(0.037)$	$0.962_{\pm 0.005}(0.006)$	$0.988_{\pm 0.004}(0.108)$	$0.887_{\pm 0.004}(0.002)$	$0.542_{\pm 0.007}$
	0.2	$0.956_{\pm 0.002}(0.166)$	$0.805_{\pm 0.003}(0.006)$	$0.960_{\pm 0.003}(0.114)$	$0.864_{\pm 0.009}(0.010)$	$0.996_{\pm 0.005}(0.062)$	$0.932_{\pm 0.002}(0.003)$	$0.249_{\pm 0.006}$
NegGrad+	0.05	$0.999_{\pm 0.000}(0.146)$	$0.890_{\pm 0.002}(0.030)$	$0.949_{\pm 0.002}(0.005)$	$1.614_{\pm 0.023}(0.665)$	$1.052_{\pm 0.002}(0.823)$	$0.552_{\pm 0.007}(1.289)$	$0.995_{\pm 0.000}$
	0.1	$0.995_{\pm 0.001}(0.144)$	$0.848_{\pm 0.000}(0.013)$	$0.898_{\pm 0.000}(0.005)$	$1.093_{\pm 0.005}(0.193)$	$1.109_{\pm 0.001}(0.042)$	$0.775_{\pm 0.003}(0.369)$	$0.933_{\pm 0.002}$
UA19.4%, RA98.3%, TA84.0%	0.15	$0.987_{\pm 0.000}(0.137)$	$0.814_{\pm 0.001}(0.047)$	$0.850_{\pm 0.001}(0.009)$	$1.009_{\pm 0.000}(0.159)$	$1.161_{\pm 0.001}(0.206)$	$0.807_{\pm 0.002}(0.149)$	$0.685_{\pm 0.002}$
	0.2	$0.966_{\pm 0.001}(0.108)$	$0.783_{\pm 0.003}(0.078)$	$0.802_{\pm 0.002}(0.012)$	$0.972_{\pm 0.000}(0.173)$	$1.205_{\pm 0.004}(0.359)$	$0.805_{\pm 0.003}(0.049)$	$0.320_{\pm 0.001}$
Salun	0.05	$0.995_{\pm 0.003}(0.142)$	$0.964_{\pm 0.026}(0.103)$	$2.803_{\pm 1.607}(1.859)$	$2.726_{\pm 0.727}(1.777)$	$0.528_{\pm 0.454}(1.347)$	$0.376_{\pm 0.129}(1.464)$	$0.988_{\pm 0.001}$
	0.1	$0.977_{\pm 0.014}(0.126)$	$0.924_{\pm 0.040}(0.064)$	$1.229_{\pm 0.286}(0.337)$	$1.281_{\pm 0.120}(0.381)$	$0.831_{\pm 0.237}(0.319)$	$0.728_{\pm 0.104}(0.417)$	$0.939_{\pm 0.005}$
UA9.2%, RA97.7%, TA83.6%	0.15	$0.936_{\pm 0.041}(0.086)$	$0.874_{\pm 0.041}(0.013)$	$0.972_{\pm 0.103}(0.131)$	$1.032_{\pm 0.005}(0.182)$	$0.974_{\pm 0.155}(0.018)$	$0.847_{\pm 0.044}(0.109)$	$0.819_{\pm 0.003}$
	0.2	$0.870_{\pm 0.081}(0.012)$	$0.810_{\pm 0.017}(0.051)$	$0.845_{\pm 0.036}(0.055)$	$0.925_{\pm 0.046}(0.126)$	$1.034_{\pm 0.143}(0.188)$	$0.876_{\pm 0.025}(0.022)$	$0.630_{\pm 0.003}$
SFRon	0.05	$0.989_{\pm 0.001}(0.045)$	$0.948_{\pm 0.001}(0.001)$	$2.000_{\pm 0.059}(0.124)$	$2.208_{\pm 0.037}(0.368)$	$0.495_{\pm 0.014}(0.008)$	$0.429_{\pm 0.007}(0.086)$	$0.986_{\pm 0.000}$
	0.1	$0.960_{\pm 0.003}(0.068)$	$0.899_{\pm 0.002}(0.001)$	$1.227_{\pm 0.017}(0.076)$	$1.268_{\pm 0.007}(0.123)$	$0.783_{\pm 0.010}(0.008)$	$0.709_{\pm 0.003}(0.077)$	$0.902_{\pm 0.003}$
UA9.3%, RA97.0%, TA83.9%	0.15	$0.917_{\pm 0.002}(0.076)$	$0.849_{\pm 0.002}(0.001)$	$1.024_{\pm 0.006}(0.068)$	$1.015_{\pm 0.005}(0.059)$	$0.896_{\pm 0.007}(0.016)$	$0.837_{\pm 0.004}(0.053)$	$0.689_{\pm 0.012}$
	0.2	$0.866_{\pm 0.006}(0.076)$	$0.802_{\pm 0.003}(0.003)$	$0.916_{\pm 0.004}(0.070)$	$0.892_{\pm 0.005}(0.037)$	$0.946_{\pm 0.002}(0.012)$	$0.899_{\pm 0.003}(0.036)$	$0.426_{\pm 0.018}$

Table 15: Unlearning performance of 9 unlearning methods on **Tiny ImageNet** with **ViT** in **50%** **random data forgetting** scenario.

Methods	α	Coverage		Set Size		CR		
		$\mathcal{D}_f \downarrow$	$\mathcal{D}_{test} \uparrow$	$\mathcal{D}_f \uparrow$	$\mathcal{D}_{test} \downarrow$	$\mathcal{D}_f \downarrow$	$\mathcal{D}_{test} \uparrow$	$\hat{q}$
RT **UA**16.0%, **RA**98.8%, **TA**84.9%	0.05	$0.946_{\pm 0.001}(0.000)$	$0.948_{\pm 0.003}(0.000)$	$2.146_{\pm 0.006}(0.000)$	$2.106_{\pm 0.002}(0.000)$	$0.441_{\pm 0.004}(0.000)$	$0.450_{\pm 0.005}(0.000)$	$0.987_{\pm 0.004}$
	0.1	$0.892_{\pm 0.007}(0.000)$	$0.899_{\pm 0.008}(0.000)$	$1.222_{\pm 0.002}(0.000)$	$1.211_{\pm 0.007}(0.000)$	$0.730_{\pm 0.004}(0.000)$	$0.742_{\pm 0.002}(0.000)$	$0.889_{\pm 0.009}$
	0.15	$0.838_{\pm 0.004}(0.000)$	$0.847_{\pm 0.001}(0.000)$	$0.977_{\pm 0.002}(0.000)$	$0.977_{\pm 0.006}(0.000)$	$0.858_{\pm 0.008}(0.000)$	$0.868_{\pm 0.006}(0.000)$	$0.607_{\pm 0.001}$
	0.2	$0.786_{\pm 0.005}(0.000)$	$0.796_{\pm 0.002}(0.000)$	$0.856_{\pm 0.007}(0.000)$	$0.863_{\pm 0.001}(0.000)$	$0.918_{\pm 0.007}(0.000)$	$0.922_{\pm 0.008}(0.000)$	$0.304_{\pm 0.008}$
FT **UA**5.4%, **RA**97.1%, **TA**84.4%	0.05	$0.995_{\pm 0.013}(0.051)$	$0.949_{\pm 0.024}(0.000)$	$1.879_{\pm 0.014}(0.003)$	$2.216_{\pm 0.003}(0.376)$	$0.527_{\pm 0.028}(0.024)$	$0.428_{\pm 0.020}(0.088)$	$0.992_{\pm 0.019}$
	0.1	$0.979_{\pm 0.021}(0.087)$	$0.901_{\pm 0.014}(0.001)$	$1.183_{\pm 0.018}(0.032)$	$1.281_{\pm 0.020}(0.137)$	$0.828_{\pm 0.029}(0.053)$	$0.701_{\pm 0.010}(0.085)$	$0.926_{\pm 0.025}$
	0.15	$0.953_{\pm 0.024}(0.112)$	$0.850_{\pm 0.022}(0.000)$	$1.014_{\pm 0.011}(0.058)$	$1.017_{\pm 0.026}(0.061)$	$0.940_{\pm 0.027}(0.060)$	$0.839_{\pm 0.004}(0.050)$	$0.681_{\pm 0.020}$
	0.2	$0.910_{\pm 0.029}(0.120)$	$0.806_{\pm 0.024}(0.007)$	$0.937_{\pm 0.018}(0.091)$	$0.895_{\pm 0.001}(0.041)$	$0.977_{\pm 0.029}(0.043)$	$0.902_{\pm 0.007}(0.033)$	$0.345_{\pm 0.016}$
RL **UA**22.5%, **RA**93.5%, **TA**77.1%	0.05	$0.974_{\pm 0.011}(0.028)$	$0.953_{\pm 0.001}(0.005)$	$26.032_{\pm 0.007}(23.886)$	$23.369_{\pm 0.008}(21.263)$	$0.038_{\pm 0.015}(0.403)$	$0.038_{\pm 0.016}(0.412)$	$0.994_{\pm 0.010}$
	0.1	$0.930_{\pm 0.016}(0.038)$	$0.902_{\pm 0.013}(0.003)$	$5.277_{\pm 0.001}(4.055)$	$4.621_{\pm 0.007}(3.410)$	$0.178_{\pm 0.011}(0.552)$	$0.197_{\pm 0.001}(0.545)$	$0.987_{\pm 0.008}$
	0.15	$0.875_{\pm 0.011}(0.037)$	$0.856_{\pm 0.008}(0.009)$	$1.758_{\pm 0.004}(0.781)$	$1.657_{\pm 0.005}(0.680)$	$0.496_{\pm 0.006}(0.362)$	$0.516_{\pm 0.009}(0.352)$	$0.970_{\pm 0.017}$
	0.2	$0.810_{\pm 0.006}(0.024)$	$0.805_{\pm 0.013}(0.009)$	$1.147_{\pm 0.005}(0.291)$	$1.144_{\pm 0.005}(0.281)$	$0.707_{\pm 0.004}(0.211)$	$0.707_{\pm 0.013}(0.215)$	$0.945_{\pm 0.005}$
GA **UA**3.9%, **RA**96.1%, **TA**84.2%	0.05	$0.998_{\pm 0.007}(0.052)$	$0.949_{\pm 0.001}(0.001)$	$1.807_{\pm 0.001}(0.339)$	$2.338_{\pm 0.001}(0.232)$	$0.552_{\pm 0.006}(0.111)$	$0.407_{\pm 0.006}(0.043)$	$0.992_{\pm 0.006}$
	0.1	$0.986_{\pm 0.009}(0.094)$	$0.896_{\pm 0.007}(0.003)$	$1.147_{\pm 0.003}(0.075)$	$1.278_{\pm 0.007}(0.067)$	$0.863_{\pm 0.008}(0.133)$	$0.703_{\pm 0.002}(0.039)$	$0.918_{\pm 0.010}$
	0.15	$0.968_{\pm 0.008}(0.130)$	$0.850_{\pm 0.002}(0.003)$	$1.015_{\pm 0.008}(0.038)$	$1.020_{\pm 0.002}(0.043)$	$0.954_{\pm 0.009}(0.096)$	$0.835_{\pm 0.002}(0.033)$	$0.696_{\pm 0.009}$
	0.2	$0.931_{\pm 0.011}(0.145)$	$0.804_{\pm 0.004}(0.008)$	$0.948_{\pm 0.000}(0.092)$	$0.893_{\pm 0.003}(0.030)$	$0.983_{\pm 0.006}(0.065)$	$0.900_{\pm 0.004}(0.022)$	$0.363_{\pm 0.002}$
Teacher **UA**22.1%, **RA**85.7%, **TA**76.2%	0.05	$0.967_{\pm 0.013}(0.021)$	$0.950_{\pm 0.017}(0.002)$	$6.465_{\pm 0.007}(4.319)$	$6.233_{\pm 0.004}(4.127)$	$0.151_{\pm 0.002}(0.290)$	$0.151_{\pm 0.006}(0.299)$	$0.990_{\pm 0.014}$
	0.1	$0.922_{\pm 0.008}(0.030)$	$0.899_{\pm 0.002}(0.000)$	$2.202_{\pm 0.012}(0.980)$	$2.167_{\pm 0.005}(0.956)$	$0.418_{\pm 0.009}(0.312)$	$0.419_{\pm 0.024}(0.323)$	$0.977_{\pm 0.001}$
	0.15	$0.869_{\pm 0.025}(0.031)$	$0.852_{\pm 0.002}(0.005)$	$1.467_{\pm 0.015}(0.490)$	$1.459_{\pm 0.004}(0.482)$	$0.591_{\pm 0.005}(0.267)$	$0.581_{\pm 0.001}(0.287)$	$0.958_{\pm 0.021}$
	0.2	$0.814_{\pm 0.020}(0.028)$	$0.801_{\pm 0.017}(0.005)$	$1.125_{\pm 0.005}(0.269)$	$1.138_{\pm 0.001}(0.275)$	$0.718_{\pm 0.017}(0.200)$	$0.704_{\pm 0.009}(0.218)$	$0.927_{\pm 0.017}$
SSD **UA**1.3%, **RA**98.4%, **TA**86.1%	0.05	$0.999_{\pm 0.001}(0.053)$	$0.952_{\pm 0.001}(0.004)$	$1.346_{\pm 0.001}(0.800)$	$1.824_{\pm 0.000}(0.282)$	$0.742_{\pm 0.000}(0.301)$	$0.522_{\pm 0.001}(0.072)$	$0.986_{\pm 0.001}$
	0.1	$0.995_{\pm 0.001}(0.103)$	$0.897_{\pm 0.000}(0.002)$	$1.033_{\pm 0.001}(0.189)$	$1.135_{\pm 0.001}(0.076)$	$0.959_{\pm 0.000}(0.229)$	$0.790_{\pm 0.000}(0.048)$	$0.847_{\pm 0.001}$
	0.15	$0.982_{\pm 0.001}(0.144)$	$0.847_{\pm 0.000}(0.000)$	$0.987_{\pm 0.000}(0.010)$	$0.956_{\pm 0.000}(0.021)$	$0.989_{\pm 0.001}(0.131)$	$0.890_{\pm 0.001}(0.022)$	$0.517_{\pm 0.001}$
	0.2	$0.959_{\pm 0.001}(0.173)$	$0.804_{\pm 0.001}(0.008)$	$0.961_{\pm 0.000}(0.105)$	$0.862_{\pm 0.000}(0.001)$	$0.995_{\pm 0.001}(0.077)$	$0.932_{\pm 0.001}(0.010)$	$0.243_{\pm 0.001}$
NegGrad+ **UA**11.5%, **RA**98.7%, **TA**83.8%	0.05	$0.999_{\pm 0.000}(0.053)$	$0.979_{\pm 0.001}(0.031)$	$0.946_{\pm 0.002}(1.200)$	$1.443_{\pm 0.028}(0.663)$	$1.056_{\pm 0.002}(0.615)$	$0.678_{\pm 0.012}(0.228)$	$0.992_{\pm 0.001}$
	0.1	$0.996_{\pm 0.000}(0.104)$	$0.946_{\pm 0.002}(0.047)$	$0.900_{\pm 0.003}(0.322)$	$1.078_{\pm 0.006}(0.134)$	$1.107_{\pm 0.003}(0.377)$	$0.877_{\pm 0.003}(0.135)$	$0.933_{\pm 0.003}$
	0.15	$0.990_{\pm 0.000}(0.152)$	$0.900_{\pm 0.003}(0.052)$	$0.853_{\pm 0.004}(0.124)$	$1.008_{\pm 0.002}(0.031)$	$1.161_{\pm 0.005}(0.303)$	$0.892_{\pm 0.001}(0.025)$	$0.712_{\pm 0.015}$
	0.2	$0.977_{\pm 0.000}(0.191)$	$0.848_{\pm 0.003}(0.052)$	$0.805_{\pm 0.002}(0.052)$	$0.982_{\pm 0.000}(0.119)$	$1.214_{\pm 0.003}(0.296)$	$0.863_{\pm 0.003}(0.058)$	$0.381_{\pm 0.009}$
Salun **UA**9.2%, **RA**95.7%, **TA**81.9%	0.05	$0.993_{\pm 0.003}(0.047)$	$0.962_{\pm 0.026}(0.014)$	$3.284_{\pm 2.048}(1.138)$	$4.112_{\pm 0.813}(2.007)$	$0.500_{\pm 0.478}(0.059)$	$0.241_{\pm 0.057}(0.209)$	$0.989_{\pm 0.001}$
	0.1	$0.976_{\pm 0.011}(0.084)$	$0.924_{\pm 0.039}(0.026)$	$1.386_{\pm 0.423}(0.164)$	$1.579_{\pm 0.130}(0.368)$	$0.764_{\pm 0.292}(0.034)$	$0.590_{\pm 0.077}(0.152)$	$0.973_{\pm 0.002}$
	0.15	$0.944_{\pm 0.024}(0.106)$	$0.876_{\pm 0.046}(0.029)$	$1.051_{\pm 0.175}(0.074)$	$1.139_{\pm 0.017}(0.162)$	$0.920_{\pm 0.195}(0.062)$	$0.770_{\pm 0.051}(0.098)$	$0.942_{\pm 0.002}$
	0.2	$0.900_{\pm 0.044}(0.114)$	$0.825_{\pm 0.049}(0.029)$	$0.910_{\pm 0.097}(0.054)$	$0.969_{\pm 0.037}(0.105)$	$1.000_{\pm 0.164}(0.082)$	$0.851_{\pm 0.020}(0.071)$	$0.893_{\pm 0.002}$
SFRon **UA**6.3%, **RA**96.8%, **TA**82.9%	0.05	$0.994_{\pm 0.001}(0.048)$	$0.947_{\pm 0.003}(0.001)$	$2.010_{\pm 0.188}(0.136)$	$2.327_{\pm 0.087}(0.222)$	$0.497_{\pm 0.045}(0.057)$	$0.407_{\pm 0.016}(0.043)$	$0.983_{\pm 0.002}$
	0.1	$0.980_{\pm 0.006}(0.087)$	$0.900_{\pm 0.003}(0.001)$	$1.245_{\pm 0.060}(0.023)$	$1.338_{\pm 0.039}(0.126)$	$0.788_{\pm 0.041}(0.058)$	$0.673_{\pm 0.020}(0.069)$	$0.909_{\pm 0.003}$
	0.15	$0.951_{\pm 0.011}(0.113)$	$0.849_{\pm 0.003}(0.001)$	$1.041_{\pm 0.020}(0.065)$	$1.044_{\pm 0.023}(0.067)$	$0.913_{\pm 0.028}(0.055)$	$0.813_{\pm 0.016}(0.055)$	$0.738_{\pm 0.029}$
	0.2	$0.910_{\pm 0.011}(0.125)$	$0.803_{\pm 0.003}(0.008)$	$0.947_{\pm 0.006}(0.091)$	$0.910_{\pm 0.022}(0.046)$	$0.961_{\pm 0.017}(0.044)$	$0.884_{\pm 0.017}(0.038)$	$0.523_{\pm 0.068}$

Table 16: Unlearning performance of 9 unlearning methods on **Tiny ImageNet** with **ViT** in **class-wise forgetting** scenario.

Methods	α	Coverage			Set Size			CR				
		$\mathcal{D}_f \downarrow$	$\mathcal{D}_{tf} \downarrow$	$\mathcal{D}_{tr} \uparrow$	$\mathcal{D}_f \uparrow$	$\mathcal{D}_{tf} \uparrow$	$\mathcal{D}_{tr} \downarrow$	$\mathcal{D}_f \downarrow$	$\mathcal{D}_{tf} \downarrow$	$\mathcal{D}_{tr} \uparrow$	$\hat{q}_{tf}$	$\hat{q}_{test}$
RT **UA**100%, **UA**$_f$100%, **RA**98.7%, **TA**86.4%	0.05	$1.000_{\pm 0.000}(0.000)$	$1.000_{\pm 0.000}(0.000)$	$0.950_{\pm 0.003}(0.000)$	$200.000_{\pm 0.000}(0.000)$	$200.000_{\pm 0.000}(0.000)$	$1.785_{\pm 0.056}(0.000)$	$0.005_{\pm 0.000}(0.000)$	$0.005_{\pm 0.000}(0.000)$	$0.532_{\pm 0.009}(0.000)$	$1.000_{\pm 0.000}$	$0.984_{\pm 0.002}$
	0.1	$0.936_{\pm 0.011}(0.000)$	$0.960_{\pm 0.016}(0.000)$	$0.903_{\pm 0.009}(0.000)$	$192.882_{\pm 0.912}(0.000)$	$193.340_{\pm 2.620}(0.000)$	$1.146_{\pm 0.002}(0.000)$	$0.005_{\pm 0.000}(0.000)$	$0.005_{\pm 0.000}(0.000)$	$0.788_{\pm 0.008}(0.000)$	$1.000_{\pm 0.000}$	$0.859_{\pm 0.004}$
	0.15	$0.904_{\pm 0.039}(0.000)$	$0.960_{\pm 0.046}(0.000)$	$0.853_{\pm 0.005}(0.000)$	$186.791_{\pm 2.173}(0.000)$	$188.880_{\pm 1.802}(0.000)$	$0.957_{\pm 0.010}(0.000)$	$0.005_{\pm 0.000}(0.000)$	$0.005_{\pm 0.000}(0.000)$	$0.892_{\pm 0.003}(0.000)$	$1.000_{\pm 0.000}$	$0.535_{\pm 0.002}$
	0.2	$0.787_{\pm 0.061}(0.000)$	$0.860_{\pm 0.024}(0.000)$	$0.805_{\pm 0.003}(0.000)$	$171.051_{\pm 3.183}(0.000)$	$174.480_{\pm 2.311}(0.000)$	$0.860_{\pm 0.010}(0.000)$	$0.005_{\pm 0.000}(0.000)$	$0.005_{\pm 0.000}(0.000)$	$0.936_{\pm 0.002}(0.000)$	$1.000_{\pm 0.000}$	$0.232_{\pm 0.001}$
FT **UA**13.8%, **UA**$_f$22.0%, **RA**97.5%, **TA**84.1%	0.05	$0.993_{\pm 0.006}(0.007)$	$0.960_{\pm 0.009}(0.040)$	$0.952_{\pm 0.006}(0.002)$	$8.360_{\pm 0.007}(191.640)$	$8.280_{\pm 0.006}(191.720)$	$2.442_{\pm 0.011}(0.657)$	$0.119_{\pm 0.018}(0.114)$	$0.116_{\pm 0.001}(0.111)$	$0.390_{\pm 0.023}(0.142)$	$0.999_{\pm 0.006}$	$0.993_{\pm 0.005}$
	0.1	$0.984_{\pm 0.009}(0.048)$	$0.860_{\pm 0.013}(0.100)$	$0.898_{\pm 0.005}(0.005)$	$1.802_{\pm 0.009}(191.080)$	$1.660_{\pm 0.018}(191.680)$	$1.287_{\pm 0.009}(0.141)$	$0.546_{\pm 0.008}(0.541)$	$0.518_{\pm 0.004}(0.513)$	$0.698_{\pm 0.019}(0.090)$	$0.971_{\pm 0.019}$	$0.924_{\pm 0.016}$
	0.15	$0.902_{\pm 0.019}(0.002)$	$0.800_{\pm 0.004}(0.160)$	$0.852_{\pm 0.017}(0.001)$	$1.120_{\pm 0.021}(185.671)$	$1.040_{\pm 0.006}(187.840)$	$1.021_{\pm 0.017}(0.064)$	$0.806_{\pm 0.012}(0.801)$	$0.769_{\pm 0.013}(0.764)$	$0.835_{\pm 0.022}(0.057)$	$0.809_{\pm 0.010}$	$0.686_{\pm 0.004}$
	0.2	$0.860_{\pm 0.022}(0.073)$	$0.760_{\pm 0.003}(0.100)$	$0.809_{\pm 0.018}(0.005)$	$0.969_{\pm 0.002}(170.082)$	$0.960_{\pm 0.003}(173.520)$	$0.882_{\pm 0.010}(0.022)$	$0.888_{\pm 0.005}(0.883)$	$0.792_{\pm 0.002}(0.787)$	$0.907_{\pm 0.006}(0.029)$	$0.595_{\pm 0.002}$	$0.338_{\pm 0.019}$
RL **UA**100%, **UA**$_f$100%, **RA**98.2%, **TA**84.6%	0.05	$0.998_{\pm 0.006}(0.002)$	$0.980_{\pm 0.003}(0.020)$	$0.952_{\pm 0.049}(0.002)$	$199.489_{\pm 0.512}(0.511)$	$195.220_{\pm 1.903}(4.780)$	$2.317_{\pm 0.069}(0.532)$	$0.005_{\pm 0.000}(0.000)$	$0.005_{\pm 0.000}(0.000)$	$0.411_{\pm 0.000}(0.121)$	$1.000_{\pm 0.000}$	$0.995_{\pm 0.032}$
	0.1	$0.971_{\pm 0.013}(0.035)$	$0.900_{\pm 0.017}(0.060)$	$0.900_{\pm 0.002}(0.003)$	$180.442_{\pm 0.716}(12.440)$	$170.960_{\pm 0.948}(22.380)$	$1.237_{\pm 0.050}(0.991)$	$0.005_{\pm 0.000}(0.000)$	$0.005_{\pm 0.000}(0.000)$	$0.727_{\pm 0.016}(0.061)$	$1.000_{\pm 0.000}$	$0.925_{\pm 0.024}$
	0.15	$0.922_{\pm 0.011}(0.018)$	$0.900_{\pm 0.011}(0.060)$	$0.852_{\pm 0.015}(0.001)$	$165.884_{\pm 2.037}(20.907)$	$159.980_{\pm 1.012}(28.900)$	$1.001_{\pm 0.003}(0.044)$	$0.006_{\pm 0.001}(0.001)$	$0.006_{\pm 0.000}(0.001)$	$0.851_{\pm 0.029}(0.041)$	$1.000_{\pm 0.000}$	$0.641_{\pm 0.035}$
	0.2	$0.882_{\pm 0.007}(0.095)$	$0.860_{\pm 0.007}(0.000)$	$0.807_{\pm 0.007}(0.002)$	$154.896_{\pm 2.028}(16.155)$	$149.280_{\pm 3.013}(25.200)$	$0.886_{\pm 0.032}(0.026)$	$0.006_{\pm 0.000}(0.001)$	$0.006_{\pm 0.001}(0.001)$	$0.912_{\pm 0.013}(0.024)$	$1.000_{\pm 0.000}$	$0.262_{\pm 0.022}$
GA **UA**9.1%, **UA**$_f$20.0%, **RA**98.6%, **TA**86.1%	0.05	$1.000_{\pm 0.001}(0.000)$	$0.980_{\pm 0.002}(0.020)$	$0.948_{\pm 0.026}(0.002)$	$22.836_{\pm 0.045}(177.164)$	$20.600_{\pm 0.011}(179.400)$	$1.781_{\pm 0.017}(0.004)$	$0.044_{\pm 0.017}(0.019)$	$0.048_{\pm 0.028}(0.043)$	$0.532_{\pm 0.013}(0.000)$	$1.000_{\pm 0.000}$	$0.984_{\pm 0.033}$
	0.1	$0.991_{\pm 0.022}(0.055)$	$0.900_{\pm 0.014}(0.060)$	$0.897_{\pm 0.016}(0.006)$	$1.631_{\pm 0.031}(191.251)$	$1.720_{\pm 0.005}(191.620)$	$1.133_{\pm 0.014}(0.013)$	$0.608_{\pm 0.006}(0.603)$	$0.523_{\pm 0.007}(0.518)$	$0.792_{\pm 0.037}(0.004)$	$0.972_{\pm 0.033}$	$0.849_{\pm 0.039}$
	0.15	$0.958_{\pm 0.002}(0.054)$	$0.820_{\pm 0.010}(0.140)$	$0.850_{\pm 0.006}(0.003)$	$1.151_{\pm 0.036}(185.640)$	$1.140_{\pm 0.043}(187.740)$	$0.958_{\pm 0.026}(0.001)$	$0.822_{\pm 0.005}(0.827)$	$0.719_{\pm 0.021}(0.714)$	$0.887_{\pm 0.044}(0.005)$	$0.868_{\pm 0.023}$	$0.535_{\pm 0.011}$
	0.2	$0.880_{\pm 0.047}(0.093)$	$0.800_{\pm 0.051}(0.060)$	$0.803_{\pm 0.025}(0.002)$	$0.929_{\pm 0.002}(170.122)$	$0.900_{\pm 0.009}(173.580)$	$0.861_{\pm 0.006}(0.001)$	$0.947_{\pm 0.036}(0.942)$	$0.889_{\pm 0.029}(0.884)$	$0.933_{\pm 0.027}(0.003)$	$0.473_{\pm 0.016}$	$0.238_{\pm 0.000}$
Teacher **UA**100%, **UA**$_f$100%, **RA**88.8%, **TA**78.6%	0.05	$0.982_{\pm 0.014}(0.018)$	$1.000_{\pm 0.007}(0.000)$	$0.952_{\pm 0.025}(0.002)$	$199.971_{\pm 0.009}(0.029)$	$200.000_{\pm 0.000}(0.000)$	$5.095_{\pm 0.020}(3.310)$	$0.005_{\pm 0.000}(0.000)$	$0.005_{\pm 0.000}(0.000)$	$0.187_{\pm 0.008}(0.345)$	$1.000_{\pm 0.000}$	$0.989_{\pm 0.001}$
	0.1	$0.909_{\pm 0.013}(0.027)$	$0.940_{\pm 0.015}(0.020)$	$0.903_{\pm 0.032}(0.000)$	$199.813_{\pm 0.009}(6.931)$	$199.900_{\pm 0.013}(6.560)$	$2.033_{\pm 0.031}(0.887)$	$0.005_{\pm 0.000}(0.000)$	$0.005_{\pm 0.000}(0.000)$	$0.444_{\pm 0.006}(0.344)$	$1.000_{\pm 0.000}$	$0.965_{\pm 0.003}$
	0.15	$0.887_{\pm 0.014}(0.017)$	$0.880_{\pm 0.011}(0.080)$	$0.854_{\pm 0.003}(0.001)$	$199.667_{\pm 0.030}(12.876)$	$199.760_{\pm 0.026}(10.880)$	$1.331_{\pm 0.012}(0.374)$	$0.004_{\pm 0.000}(0.001)$	$0.004_{\pm 0.001}(0.001)$	$0.641_{\pm 0.010}(0.251)$	$1.000_{\pm 0.000}$	$0.919_{\pm 0.001}$
	0.2	$0.838_{\pm 0.022}(0.051)$	$0.840_{\pm 0.002}(0.020)$	$0.799_{\pm 0.017}(0.006)$	$199.413_{\pm 0.024}(28.362)$	$199.620_{\pm 0.030}(25.140)$	$1.022_{\pm 0.017}(0.162)$	$0.004_{\pm 0.001}(0.001)$	$0.004_{\pm 0.001}(0.001)$	$0.781_{\pm 0.019}(0.155)$	$1.000_{\pm 0.000}$	$0.825_{\pm 0.002}$
SSD **UA**100%, **UA**$_f$100%, **RA**98.4%, **TA**86.1%	0.05	$1.000_{\pm 0.000}(0.000)$	$1.000_{\pm 0.000}(0.000)$	$0.950_{\pm 0.017}(0.000)$	$198.769_{\pm 0.052}(1.231)$	$197.320_{\pm 1.010}(2.680)$	$1.866_{\pm 0.019}(0.081)$	$0.005_{\pm 0.000}(0.000)$	$0.005_{\pm 0.000}(0.000)$	$0.509_{\pm 0.013}(0.023)$	$1.000_{\pm 0.000}$	$0.986_{\pm 0.006}$
	0.1	$0.949_{\pm 0.012}(0.013)$	$0.900_{\pm 0.012}(0.060)$	$0.897_{\pm 0.007}(0.006)$	$171.073_{\pm 0.209}(21.809)$	$169.360_{\pm 2.002}(23.980)$	$1.141_{\pm 0.014}(0.005)$	$0.006_{\pm 0.000}(0.001)$	$0.005_{\pm 0.000}(0.000)$	$0.786_{\pm 0.021}(0.002)$	$1.000_{\pm 0.000}$	$0.854_{\pm 0.006}$
	0.15	$0.913_{\pm 0.007}(0.009)$	$0.880_{\pm 0.020}(0.080)$	$0.852_{\pm 0.015}(0.001)$	$157.140_{\pm 1.209}(29.651)$	$154.960_{\pm 0.907}(33.920)$	$0.959_{\pm 0.011}(0.002)$	$0.006_{\pm 0.001}(0.001)$	$0.006_{\pm 0.000}(0.001)$	$0.888_{\pm 0.012}(0.004)$	$1.000_{\pm 0.000}$	$0.538_{\pm 0.007}$
	0.2	$0.833_{\pm 0.007}(0.046)$	$0.800_{\pm 0.013}(0.060)$	$0.806_{\pm 0.022}(0.001)$	$136.502_{\pm 3.022}(34.549)$	$136.420_{\pm 2.422}(38.060)$	$0.864_{\pm 0.002}(0.004)$	$0.006_{\pm 0.000}(0.001)$	$0.006_{\pm 0.000}(0.001)$	$0.932_{\pm 0.015}(0.004)$	$1.000_{\pm 0.000}$	$0.254_{\pm 0.005}$
NegGrad+ **UA**100%, **UA**$_f$100%, **RA**99.0%, **TA**85.8%	0.05	$1.000_{\pm 0.000}(0.000)$	$1.000_{\pm 0.000}(0.000)$	$0.947_{\pm 0.002}(0.003)$	$200.000_{\pm 0.000}(0.000)$	$200.000_{\pm 0.000}(0.000)$	$1.850_{\pm 0.036}(0.065)$	$0.005_{\pm 0.000}(0.000)$	$0.005_{\pm 0.000}(0.000)$	$0.512_{\pm 0.009}(0.020)$	$1.000_{\pm 0.000}$	$0.987_{\pm 0.001}$
	0.1	$0.927_{\pm 0.104}(0.009)$	$0.950_{\pm 0.071}(0.010)$	$0.894_{\pm 0.001}(0.009)$	$193.994_{\pm 8.493}(1.112)$	$197.490_{\pm 3.550}(4.150)$	$1.140_{\pm 0.007}(0.006)$	$0.005_{\pm 0.000}(0.000)$	$0.005_{\pm 0.000}(0.000)$	$0.784_{\pm 0.004}(0.004)$	$1.000_{\pm 0.000}$	$0.859_{\pm 0.003}$
	0.15	$0.862_{\pm 0.013}(0.042)$	$0.870_{\pm 0.042}(0.090)$	$0.849_{\pm 0.000}(0.004)$	$188.686_{\pm 0.954}(1.894)$	$195.590_{\pm 0.863}(6.710)$	$0.961_{\pm 0.001}(0.004)$	$0.004_{\pm 0.000}(0.000)$	$0.004_{\pm 0.000}(0.001)$	$0.884_{\pm 0.000}(0.008)$	$1.000_{\pm 0.000}$	$0.537_{\pm 0.003}$
	0.2	$0.830_{\pm 0.027}(0.043)$	$0.840_{\pm 0.085}(0.020)$	$0.802_{\pm 0.002}(0.003)$	$187.219_{\pm 0.064}(16.168)$	$194.310_{\pm 0.948}(19.830)$	$0.861_{\pm 0.001}(0.002)$	$0.004_{\pm 0.000}(0.000)$	$0.004_{\pm 0.000}(0.001)$	$0.931_{\pm 0.001}(0.005)$	$1.000_{\pm 0.000}$	$0.220_{\pm 0.002}$
Salun **UA**100%, **UA**$_f$100%, **RA**98.4%, **TA**86.1%	0.05	$0.997_{\pm 0.003}(0.003)$	$0.993_{\pm 0.012}(0.007)$	$0.949_{\pm 0.001}(0.001)$	$199.599_{\pm 0.207}(0.401)$	$197.440_{\pm 1.244}(2.560)$	$1.980_{\pm 0.010}(0.196)$	$0.005_{\pm 0.000}(0.000)$	$0.005_{\pm 0.000}(0.000)$	$0.479_{\pm 0.012}(0.053)$	$1.000_{\pm 0.000}$	$0.989_{\pm 0.001}$
	0.1	$0.975_{\pm 0.019}(0.039)$	$0.927_{\pm 0.023}(0.033)$	$0.899_{\pm 0.001}(0.003)$	$191.973_{\pm 1.616}(0.910)$	$185.220_{\pm 0.918}(8.120)$	$1.169_{\pm 0.002}(0.023)$	$0.005_{\pm 0.000}(0.000)$	$0.005_{\pm 0.000}(0.000)$	$0.769_{\pm 0.001}(0.019)$	$1.000_{\pm 0.000}$	$0.884_{\pm 0.001}$
	0.15	$0.961_{\pm 0.022}(0.057)$	$0.860_{\pm 0.040}(0.100)$	$0.850_{\pm 0.001}(0.004)$	$187.825_{\pm 3.461}(1.034)$	$180.307_{\pm 2.908}(8.573)$	$0.969_{\pm 0.002}(0.012)$	$0.005_{\pm 0.000}(0.000)$	$0.005_{\pm 0.000}(0.000)$	$0.877_{\pm 0.002}(0.015)$	$1.000_{\pm 0.000}$	$0.562_{\pm 0.003}$
	0.2	$0.960_{\pm 0.015}(0.173)$	$0.840_{\pm 0.020}(0.020)$	$0.801_{\pm 0.001}(0.004)$	$184.838_{\pm 3.478}(13.787)$	$177.647_{\pm 2.627}(3.167)$	$0.863_{\pm 0.004}(0.003)$	$0.005_{\pm 0.000}(0.001)$	$0.005_{\pm 0.000}(0.000)$	$0.928_{\pm 0.003}(0.008)$	$1.000_{\pm 0.000}$	$0.230_{\pm 0.009}$
SFRon **UA**100%, **UA**$_f$100%, **RA**96.1%, **TA**84.3%	0.05	$1.000_{\pm 0.000}(0.000)$	$1.000_{\pm 0.000}(0.000)$	$0.948_{\pm 0.001}(0.002)$	$200.000_{\pm 0.000}(0.000)$	$200.000_{\pm 0.000}(0.000)$	$2.264_{\pm 0.254}(0.479)$	$0.005_{\pm 0.000}(0.000)$	$0.005_{\pm 0.000}(0.000)$	$0.423_{\pm 0.050}(0.110)$	$1.000_{\pm 0.000}$	$0.990_{\pm 0.003}$
	0.1	$1.000_{\pm 0.000}(0.064)$	$1.000_{\pm 0.000}(0.040)$	$0.901_{\pm 0.001}(0.003)$	$200.000_{\pm 0.000}(7.118)$	$200.000_{\pm 0.000}(6.660)$	$1.266_{\pm 0.044}(0.120)$	$0.005_{\pm 0.000}(0.000)$	$0.005_{\pm 0.000}(0.000)$	$0.711_{\pm 0.026}(0.077)$	$1.000_{\pm 0.000}$	$0.912_{\pm 0.017}$
	0.15	$1.000_{\pm 0.000}(0.096)$	$1.000_{\pm 0.000}(0.040)$	$0.850_{\pm 0.002}(0.003)$	$200.000_{\pm 0.000}(13.209)$	$200.000_{\pm 0.000}(11.120)$	$1.009_{\pm 0.012}(0.051)$	$0.005_{\pm 0.000}(0.000)$	$0.005_{\pm 0.000}(0.000)$	$0.843_{\pm 0.011}(0.049)$	$1.000_{\pm 0.000}$	$0.668_{\pm 0.029}$
	0.2	$1.000_{\pm 0.000}(0.213)$	$1.000_{\pm 0.000}(0.140)$	$0.802_{\pm 0.003}(0.003)$	$200.000_{\pm 0.000}(28.949)$	$200.000_{\pm 0.000}(25.520)$	$0.886_{\pm 0.006}(0.026)$	$0.005_{\pm 0.000}(0.000)$	$0.005_{\pm 0.000}(0.000)$	$0.905_{\pm 0.007}(0.031)$	$1.000_{\pm 0.000}$	$0.358_{\pm 0.017}$

Table 17: **MIACR** performance on **CIFAR-10** with **ResNet-18**.

Methods	α	10% Forgetting MIACR ↑	10% Forgetting $\hat{q}$	50% Forgetting MIACR ↑	50% Forgetting $\hat{q}$
RT **MIA**86.92% (**10% Forgetting**) **MIA**82.79% (**50% Forgetting**)	0.05 0.1 0.15 0.2	$0.089_{\pm 0.001}$ (0.000) $0.147_{\pm 0.000}$ (0.000) $0.203_{\pm 0.010}$ (0.000) $0.246_{\pm 0.000}$ (0.000)	$0.877_{\pm 0.004}$ $0.589_{\pm 0.008}$ $0.485_{\pm 0.005}$ $0.473_{\pm 0.001}$	$0.117_{\pm 0.010}$ (0.000) $0.201_{\pm 0.011}$ (0.000) $0.272_{\pm 0.011}$ (0.000) $0.318_{\pm 0.006}$ (0.000)	$0.899_{\pm 0.007}$ $0.570_{\pm 0.001}$ $0.472_{\pm 0.009}$ $0.459_{\pm 0.003}$
FT **MIA**92.00% (**10% Forgetting**) **MIA**92.92% (**50% Forgetting**)	0.05 0.1 0.15 0.2	$0.037_{\pm 0.011}$ (0.052) $0.077_{\pm 0.008}$ (0.070) $0.128_{\pm 0.007}$ (0.075) $0.196_{\pm 0.003}$ (0.050)	$0.745_{\pm 0.013}$ $0.627_{\pm 0.000}$ $0.517_{\pm 0.008}$ $0.483_{\pm 0.003}$	$0.038_{\pm 0.001}$ (0.079) $0.103_{\pm 0.011}$ (0.098) $0.159_{\pm 0.011}$ (0.113) $0.244_{\pm 0.010}$ (0.074)	$0.780_{\pm 0.011}$ $0.558_{\pm 0.012}$ $0.494_{\pm 0.011}$ $0.476_{\pm 0.004}$
RL **MIA**74.21% (**10% Forgetting**) **MIA**61.15% (**50% Forgetting**)	0.05 0.1 0.15 0.2	$0.056_{\pm 0.010}$ (0.033) $0.178_{\pm 0.027}$ (0.031) $0.272_{\pm 0.006}$ (0.069) $0.320_{\pm 0.025}$ (0.074)	$0.627_{\pm 0.011}$ $0.572_{\pm 0.005}$ $0.492_{\pm 0.015}$ $0.485_{\pm 0.011}$	$0.057_{\pm 0.016}$ (0.060) $0.137_{\pm 0.030}$ (0.064) $0.194_{\pm 0.031}$ (0.078) $0.261_{\pm 0.001}$ (0.057)	$0.547_{\pm 0.000}$ $0.547_{\pm 0.001}$ $0.547_{\pm 0.001}$ $0.546_{\pm 0.000}$
GA **MIA**98.80% (**10% Forgetting**) **MIA**98.86% (**50% Forgetting**)	0.05 0.1 0.15 0.2	$0.010_{\pm 0.002}$ (0.079) $0.032_{\pm 0.003}$ (0.115) $0.076_{\pm 0.000}$ (0.127) $0.146_{\pm 0.016}$ (0.100)	$0.862_{\pm 0.016}$ $0.502_{\pm 0.016}$ $0.477_{\pm 0.007}$ $0.476_{\pm 0.019}$	$0.010_{\pm 0.019}$ (0.107) $0.055_{\pm 0.003}$ (0.146) $0.107_{\pm 0.016}$ (0.165) $0.164_{\pm 0.016}$ (0.154)	$0.771_{\pm 0.008}$ $0.486_{\pm 0.005}$ $0.474_{\pm 0.015}$ $0.473_{\pm 0.011}$
Teacher **MIA**87.24% (**10% Forgetting**) **MIA**93.24% (**50% Forgetting**)	0.05 0.1 0.15 0.2	$0.011_{\pm 0.006}$ (0.078) $0.038_{\pm 0.023}$ (0.109) $0.072_{\pm 0.013}$ (0.131) $0.113_{\pm 0.008}$ (0.133)	$0.750_{\pm 0.014}$ $0.672_{\pm 0.028}$ $0.625_{\pm 0.029}$ $0.588_{\pm 0.019}$	$0.031_{\pm 0.003}$ (0.086) $0.065_{\pm 0.021}$ (0.136) $0.110_{\pm 0.017}$ (0.162) $0.159_{\pm 0.017}$ (0.159)	$0.635_{\pm 0.018}$ $0.582_{\pm 0.013}$ $0.548_{\pm 0.007}$ $0.532_{\pm 0.006}$
SSD **MIA**98.78% (**10% Forgetting**) **MIA**98.87% (**50% Forgetting**)	0.05 0.1 0.15 0.2	$0.010_{\pm 0.011}$ (0.079) $0.031_{\pm 0.010}$ (0.116) $0.077_{\pm 0.005}$ (0.126) $0.139_{\pm 0.011}$ (0.107)	$0.861_{\pm 0.012}$ $0.511_{\pm 0.011}$ $0.480_{\pm 0.013}$ $0.475_{\pm 0.013}$	$0.011_{\pm 0.002}$ (0.106) $0.051_{\pm 0.005}$ (0.150) $0.104_{\pm 0.006}$ (0.168) $0.168_{\pm 0.012}$ (0.150)	$0.748_{\pm 0.011}$ $0.488_{\pm 0.001}$ $0.477_{\pm 0.015}$ $0.477_{\pm 0.006}$
NegGrad+ **MIA**90.30% (**10% Forgetting**) **MIA**93.82% (**50% Forgetting**)	0.05 0.1 0.15 0.2	$0.076_{\pm 0.025}$ (0.013) $0.128_{\pm 0.018}$ (0.019) $0.174_{\pm 0.022}$ (0.029) $0.213_{\pm 0.012}$ (0.033)	$0.844_{\pm 0.024}$ $0.481_{\pm 0.009}$ $0.480_{\pm 0.005}$ $0.480_{\pm 0.004}$	$0.045_{\pm 0.008}$ (0.072) $0.109_{\pm 0.007}$ (0.092) $0.167_{\pm 0.017}$ (0.105) $0.230_{\pm 0.014}$ (0.088)	$0.863_{\pm 0.025}$ $0.511_{\pm 0.008}$ $0.477_{\pm 0.010}$ $0.472_{\pm 0.008}$
Salun **MIA**57.58% (**10% Forgetting**) **MIA**59.12% (**50% Forgetting**)	0.05 0.1 0.15 0.2	$0.055_{\pm 0.014}$ (0.034) $0.113_{\pm 0.009}$ (0.034) $0.198_{\pm 0.006}$ (0.005) $0.267_{\pm 0.009}$ (0.021)	$0.691_{\pm 0.011}$ $0.681_{\pm 0.013}$ $0.642_{\pm 0.015}$ $0.608_{\pm 0.011}$	$0.044_{\pm 0.001}$ (0.073) $0.115_{\pm 0.009}$ (0.086) $0.170_{\pm 0.009}$ (0.102) $0.220_{\pm 0.005}$ (0.098)	$0.670_{\pm 0.008}$ $0.630_{\pm 0.009}$ $0.610_{\pm 0.003}$ $0.586_{\pm 0.005}$
SFRon **MIA**91.55% (**10% Forgetting**) **MIA**92.52% (**50% Forgetting**)	0.05 0.1 0.15 0.2	$0.060_{\pm 0.001}$ (0.029) $0.040_{\pm 0.004}$ (0.107) $0.113_{\pm 0.003}$ (0.090) $0.184_{\pm 0.002}$ (0.062)	$0.711_{\pm 0.009}$ $0.626_{\pm 0.025}$ $0.517_{\pm 0.003}$ $0.487_{\pm 0.002}$	$0.058_{\pm 0.002}$ (0.059) $0.046_{\pm 0.002}$ (0.155) $0.134_{\pm 0.013}$ (0.138) $0.206_{\pm 0.014}$ (0.112)	$0.715_{\pm 0.008}$ $0.562_{\pm 0.013}$ $0.498_{\pm 0.003}$ $0.483_{\pm 0.002}$

Table 18: Performance of our unlearning framework. We show the unlearning performance on **CIFAR-10** with **ResNet-18** and **Tiny ImageNet** with **ViT** in 10% **random data forgetting** scenario.

Methods	α	$\lambda = 0.2$ UA ↑	RA ↑	TA ↑	$CR_{\mathcal{D}_f}$ ↓	$CR_{\mathcal{D}_{test}}$ ↑	$\lambda = 0.5$ UA ↑	RA ↑	TA ↑	$CR_{\mathcal{D}_f}$ ↓	$CR_{\mathcal{D}_{test}}$ ↑	$\lambda = 1$ UA ↑	RA ↑	TA ↑	$CR_{\mathcal{D}_f}$ ↓	$CR_{\mathcal{D}_{test}}$ ↑
CIFAR-10 with ResNet-18																
RT	0.05	10.8%(2.2)	98.3%(1.4)	91.0%(0.8)	0.788(0.076)	0.824(0.055)	14.0%(5.4)	97.8%(1.9)	90.4%(0.4)	0.763(0.101)	0.825(0.054)	17.7%(9.1)	96.8%(2.9)	90.5%(1.3)	0.719(0.145)	0.820(0.059)
	0.1				0.914(0.029)	0.924(0.021)				0.879(0.064)	0.912(0.033)				0.838(0.105)	0.911(0.034)
	0.15				0.956(0.019)	0.959(0.009)				0.936(0.039)	0.954(0.014)				0.906(0.069)	0.951(0.017)
	0.2				0.977(0.011)	0.976(0.005)				0.963(0.025)	0.966(0.015)				0.932(0.056)	0.965(0.016)
FT	0.05	6.8%(1.8)	97.0%(2.7)	90.8%(1.0)	0.844(0.020)	0.829(0.050)	7.9%(0.7)	96.9%(2.8)	90.9%(0.9)	0.853(0.011)	0.843(0.036)	9.2%(0.6)	97.9%(1.8)	91.2%(0.6)	0.835(0.029)	0.854(0.025)
	0.1				0.948(0.005)	0.924(0.021)				0.940(0.003)	0.927(0.018)				0.938(0.005)	0.936(0.009)
	0.15				0.983(0.008)	0.959(0.009)				0.975(0.000)	0.961(0.007)				0.976(0.001)	0.970(0.002)
	0.2				0.989(0.001)	0.974(0.007)				0.983(0.005)	0.975(0.006)				0.986(0.002)	0.984(0.003)
RL	0.05	9.7%(1.1)	96.6%(3.1)	89.4%(2.4)	0.709(0.155)	0.736(0.143)	9.9%(1.3)	96.9%(2.8)	89.7%(2.1)	0.708(0.156)	0.731(0.148)	12.6%(4.0)	95.3%(4.4)	88.1%(3.7)	0.629(0.235)	0.669(0.210)
	0.1				0.896(0.047)	0.887(0.058)				0.902(0.041)	0.896(0.049)				0.845(0.098)	0.858(0.087)
	0.15				0.946(0.029)	0.931(0.037)				0.939(0.036)	0.932(0.036)				0.911(0.064)	0.913(0.055)
	0.2				0.964(0.024)	0.949(0.032)				0.959(0.029)	0.950(0.031)				0.936(0.052)	0.938(0.043)
Tiny ImageNet with ViT																
RT	0.05	19.3%(4.6)	98.8%(0.0)	86.0%(0.0)	0.458(0.045)	0.516(0.000)	26.4%(11.7)	98.7%(0.1)	85.8%(0.2)	0.396(0.107)	0.489(0.027)	35.7%(21.0)	98.6%(0.2)	85.2%(0.8)	0.346(0.157)	0.481(0.035)
	0.1				0.729(0.046)	0.786(0.000)				0.649(0.126)	0.765(0.021)				0.549(0.226)	0.739(0.047)
	0.15				0.841(0.039)	0.889(0.000)				0.768(0.112)	0.880(0.009)				0.658(0.222)	0.861(0.028)
	0.2				0.898(0.036)	0.932(0.003)				0.839(0.095)	0.929(0.006)				0.743(0.191)	0.918(0.017)
FT	0.05	9.8%(4.9)	97.4%(1.4)	83.6%(2.4)	0.441(0.062)	0.399(0.117)	13.6%(0.9)	97.2%(1.6)	83.6%(2.4)	0.413(0.090)	0.401(0.115)	20.0%(5.3)	96.4%(2.4)	82.9%(3.1)	0.342(0.161)	0.363(0.153)
	0.1				0.753(0.022)	0.683(0.103)				0.718(0.057)	0.683(0.103)				0.627(0.148)	0.652(0.134)
	0.15				0.884(0.004)	0.823(0.066)				0.848(0.032)	0.819(0.070)				0.772(0.108)	0.802(0.087)
	0.2				0.942(0.008)	0.893(0.042)				0.914(0.020)	0.890(0.045)				0.856(0.078)	0.877(0.058)
RL	0.05	31.8%(17.1)	95.3%(17.9)	80.9%(5.1)	0.051(0.452)	0.111(0.405)	36.2%(21.5)	95.3%(3.5)	80.4%(5.6)	0.051(0.452)	0.121(0.395)	40.2%(25.5)	94.5%(4.3)	79.5%(6.5)	0.048(0.455)	0.119(0.397)
	0.1				0.278(0.497)	0.451(0.335)				0.254(0.521)	0.449(0.337)				0.236(0.539)	0.436(0.350)
	0.15				0.579(0.301)	0.710(0.179)				0.541(0.339)	0.708(0.181)				0.480(0.400)	0.673(0.216)
	0.2				0.752(0.182)	0.825(0.110)				0.718(0.216)	0.827(0.108)				0.642(0.292)	0.793(0.142)