Redefining Machine Unlearning: A Conformal Prediction-Motivated Approach

Anonymous Authors[1]

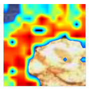
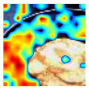
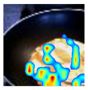
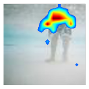
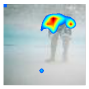
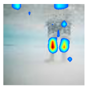
Abstract

Machine unlearning seeks to systematically remove specified data from a trained model, effectively achieving a state as though the data had never been encountered during training. While metrics such as Unlearning Accuracy (UA) and Membership Inference Attack (MIA) provide a baseline for assessing unlearning performance, they fall short of evaluating the completeness and reliability of forgetting. This is because the ground truth labels remain potential candidates within the scope of uncertainty quantification, leaving gaps in the evaluation of true forgetting. In this paper, we identify critical limitations in existing unlearning metrics and propose enhanced evaluation metrics inspired by conformal prediction. Our metrics can effectively capture the extent to which ground truth labels are excluded from the prediction set. Furthermore, we observe that many existing machine unlearning methods do not achieve satisfactory forgetting performance when evaluated with our new metrics. To address this, we propose an unlearning framework that integrates conformal prediction insights into Carlini & Wagner adversarial attack loss. Extensive experiments on the image classification task demonstrate that our enhanced metrics offer deeper insights into unlearning effectiveness, and that our unlearning framework significantly improves the forgetting quality of unlearning methods.

1. Introduction

Machine unlearning has become crucial in advancing data privacy, especially under legal requirements like the General Data Protection Regulation (GDPR) (Bourtoule et al., 2021). These regulations emphasize the right for individuals to have their data removed or forgotten, creating a demand for machine unlearning methods that can enable machine learning models to behave as if specific forget data were never used in training stage. The existing post hoc machine unlearning methods can be categorized into training-based (Graves et al., 2021; Warnecke et al., 2021; Thudi et al., 2022; Tarun et al., 2023) and training-free (Guo et al., 2019; Sekhari et al., 2021; Nguyen et al., 2020; Golatkar et al., 2020b; 2021; Foster et al., 2024) approaches, depending on whether they require any traditional model training steps during the unlearning process (Foster et al., 2024).

Table 1: Grad-CAM maps of one original model in CIFAR-10 with ResNet18 and two corresponding unlearning models. The **Prediction** 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. Although the Finetune unlearning method, can misclassify the forget data, Grad-CAM can still highlight key features of the object under this model since the true label is included in the prediction set. In contrast, our unlearning method removes the true label from the set, with activation regions shifting significantly away from the object's key features. This confirms that the forgetting quality is better if the true label can be excluded from the prediction set.

Class Name	Forget Data	Original Model	Finetune Method	Our Method
Wok				
Swimming Trunks				
Prediction	–	✓	✗	✗
In Set	–	✓	✓	✗

To measure the forgetting quality and predictive performance of the model after the unlearning process (i.e., unlearning model), several unlearning metrics have been proposed (Kashef, 2021; Shokri et al., 2017; Chen et al., 2021; Brophy & Lowd, 2021; Cao & Yang, 2015). However, existing unlearning metrics, such as unlearning accuracy (UA) (Brophy & Lowd, 2021; Foster et al., 2024) and membership inference attack (MIA) (Shokri et al., 2017; Chen et al., 2021), fall short in fully evaluating forgetting quality and reliability. These metrics primarily focus on whether mod-

[1] Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

els can predict forget data accurately without sufficiently addressing how well the model forgets. In a nutshell, misclassifying forget data does not mean that the model has completely forgotten it to some extent.

To verify this view, we apply conformal prediction (Papadopoulos et al., 2002; Lei & Wasserman, 2014) to the unlearning models. Through experiments, we find that although the model misclassifies some forget data, a significant portion of these misclassified instances still appear in the conformal prediction set. Pictures in Table 1, which visualize the important feature of models' prediction by using Grad-CAM (Selvaraju et al., 2017), further explain this phenomenon. Despite the Finetune method incorrectly predicting the forget data, the Grad-CAM maps still focus heavily on the important features of the object itself.

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

- We identify pivotal limitations in current unlearning metrics, as they overlook misclassified data where ground truth labels remain potential candidates under uncertainty quantification.
- We design two novel metrics to address the limitations motivated by conformal prediction.
- We propose a general unlearning framework for training-based machine unlearning methods motivated by conformal prediction and C&W loss.
- Extensive experiments demonstrate the effectiveness of novel metrics and our unlearning framework.

2. Enhancing Metrics for Machine Unlearning Based on Conformal Prediction

2.1. Preliminaries and Notations

Machine Unlearning. Machine unlearning is the targeted removal of certain training data effects from a machine learning model. In our work, two different forgetting scenarios are considered: (i) *random data forgetting* focuses on randomly forgetting specific instances within the training data, and (ii) *class-wise forgetting* aims to remove all information associated with an entire class. 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 $\mathcal{D}_r = \mathcal{D}_{train} \setminus \mathcal{D}_f$. In random data forgetting, $\mathcal{D}_{test}$ represents test data. In class-wise forgetting, $\mathcal{D}_{tf}$ corresponds to the test-forget data, exclusively containing the forget class, while $\mathcal{D}_{tr}$ represents the test-retain data within the test data $\mathcal{D}_{test}$. Let θ_u denote the model after the unlearning process, where the influence of forget data $\mathcal{D}_f$ has been removed. The purpose of machine unlearning is to enable θ_u to forget $\mathcal{D}_f$ (thus no longer recognizing $\mathcal{D}_{tf}$ in the class-wise forgetting scenario), while maintaining accuracy on the retain data $\mathcal{D}_r$ and avoiding a significant drop in accuracy on $\mathcal{D}_{test}$ in the random data forgetting scenario or $\mathcal{D}_{tr}$ in the class-wise setting.

Conformal Prediction. Conformal prediction is proposed to quantify uncertainty in machine learning models, providing prediction sets that contain the true label with a guaranteed probability (Angelopoulos & Bates, 2021). Among the various types of conformal prediction, this work specifically focuses on split conformal prediction (SCP)[1]. To construct the conformal prediction set, it involves four steps:

1. *Calibration Data.* SCP first chooses unseen data as calibration data. The number of calibration data points should be enough to evaluate the model's uncertainty.
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 output (through a softmax function) for the class y_i.
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 on the calibration data $\mathcal{D}_c$.
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:
$$\mathbb{C}(\boldsymbol{x}) = \{y_i : S(\boldsymbol{x}, y_i) \leq \hat{q}\}, \tag{2}$$
where y_i represents different classes.

2.2. Identifying Limitations in Current Unlearning Metrics

Evaluating the effectiveness of machine unlearning involves a trade-off between how well the model forgets unwanted

[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.

Table 2: Unlearning performance measured by existing metrics across RT, FT and RL methods. We implement **CIFAR-10** with **ResNet-18** in $\mathbf{10\%}$ and $\mathbf{50\%}$ **random data forgetting** scenarios. All values in percent (%). The sign ↑ (↓) represents the greater (smaller) is better.

	10% Forgetting				**50% Forgetting**			
Methods	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

Table 3: Mis-label (mis-classification) count and in-set ratio of UA and MIA metrics for RT, FT and RL on **CIFAT-10** with **ResNet-18** under $\mathbf{10\%}$ and $\mathbf{50\%}$ **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 B.

	10% Forgetting			**50% Forgetting**		
Methods	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%

data (forgetting quality) and how much it retains useful knowledge (predictive performance). The first key question we pose is as follows:

> **(Q1)** *Are existing commonly used metrics sufficient to evaluate unlearning performance?*

We apply the most commonly used metrics, unlearning accuracy (**UA**), retain accuracy (**RA**), test accuracy (**TA**) and **MIA**, to evaluate three classic machine unlearning methods: Retrain (RT), Finetune (FT) (Warnecke et al., 2021), and Random Label (RL) (Graves et al., 2021). These methods are implemented on a ResNet-18 model trained with the CIFAR-10 dataset in a random data forgetting scenario. In Table 2, compared with RA on $\mathcal{D}_r$, the UA of the three methods proves that the model indeed forgets some forget data. The MIA column also indicates that many data points seem to be forgotten. These metrics appear to provide valuable insights into the extent of forgetting achieved. However, higher unlearning accuracy cannot ensure that these forget data points are no longer represented in any form within the model's predictions.

To validate our hypothesis, we employ the process of conformal prediction to examine whether the true labels of the forget data appear within the prediction set. We set the confidence level to 95%, i.e., $\alpha = 0.05$, and set the calibration set size to $2,000$. In Table 3, we count both UA metric and MIA metric's misclassified data points (mis-label) and, after applying conformal prediction, determined how many of these points fall within the conformal prediction set (in-set). Table 3 demonstrates that even though the model misclassifies part of the forget data, on average 54.6% of these misclassified instances are still present in the conformal prediction set. This indicates that a high UA does not mean the model has truly forgotten the data. Relying solely on UA to evaluate the forgetting quality is insufficient.

In MIA, the prediction of '0' indicates a data point is predicted as "truly forget", while '1' signifies "still retain". And "mis-label" refers to the number of the model predicts the forget data points as '0'. In conformal prediction of MIA, there are three potential prediction set results, i.e., $\{0\}, \{1\}, \{0, 1\}$. The "in-set" here refers to the number of mislabeled data points whose conformal prediction set includes 1. Thus, the in-set ratio indicates that, although the MIA fails to identify an average of 18.33% of the forget data points, conformal prediction can still recognize more than half of these forget data points within its prediction set.

Overall, high in-set ratios in conformal prediction highlight that a considerable portion of the forget data remains identifiable through conformal prediction. This implies these unlearning methods may not be fully effective in eliminating traces of the forget data from another perspective.

2.3. Designing Metrics Motivated by Conformal Prediction

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

> **(Q2)** *Can we develop metrics to address the shortcomings 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 limitations of UA, we introduce a novel metric, Conformal Ratio (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.

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

$$\text{Coverage} := \frac{1}{|\mathcal{D}|} \sum_{(\boldsymbol{x}, y_t) \in \mathcal{D}} \mathbb{I}(y_t \in \mathbb{C}(\boldsymbol{x})), \tag{3}$$

where y_t is the true label of data point $\boldsymbol{x}$. Coverage reflects the probability that the true label falls within the prediction set $\mathbb{C}(\boldsymbol{x})$. For $\mathcal{D} = \mathcal{D}_f$, high coverage indicates that the model retains significant information about forget data, suggesting incomplete unlearning. Conversely, higher coverage is better for test data $\mathcal{D}_{test}$.

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

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

where $|\mathbb{C}(\boldsymbol{x})|$ represents the prediction set size of a specific instance. When $\mathcal{D} = \mathcal{D}_f$, a smaller set size implies that the model cannot forget specific information about $\mathcal{D}_f$ well. In contrast, when $\mathcal{D} = \mathcal{D}_{test}$, a small set size may indicate that the model maintains a strong performance after unlearning.

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

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

CR balances the information captured by coverage and set size. A lower CR value implies successful unlearning on forget data $\mathcal{D}_f$, but a significant drop in the model's utility for $\mathcal{D}_{test}$. CR is inspired by conformal prediction which is proposed to assess the model's behavior on new, unseen data (e.g., validation or test 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). To address the limitations of the current MIA metric, we propose a new metric called MIACR. For three potential prediction sets $\{0\}$, $\{1\}$ and $\{0, 1\}$, set $\{0\}$ is the only ideal case for MIA prediction since '1' represents the data point is part of the training data. Therefore, we introduce a new metric to measure the probability of the set $\{0\}$ occurring, referred to as **MIACR**:

$$\text{MIACR} := \frac{1}{|\mathcal{D}_f|} \sum_{(\boldsymbol{x}, y_t) \in \mathcal{D}_f} \mathbb{I}(\mathbb{C}(\boldsymbol{x}) = \{0\}), \tag{6}$$

where $\mathbb{C}(\boldsymbol{x}) = \{0\}$ denotes prediction set is $\{0\}$. MIACR only calculates the average of $\mathbb{C}(\boldsymbol{x}) = \{0\}$, and a higher MIACR value represents better forgetting for $\mathcal{D}_f$.

Superiority of Our Metrics. Unlike existing metrics, both the CR and MIACR metrics incorporate the fact that the true labels of some forgotten data points may remain within the prediction set, even when the predictions deviate from the true labels. This allows for a more robust assessment of the model's forgetting quality and predictive performance.

Evaluation Criteria. Similar to existing unlearning metrics, we consider two different criteria to measure unlearning performance with our metrics. ❶ *Gap to RT Criterion*: A lower gap to RT method is better for both CR and MIACR metrics. The performance gap relative to the RT method is represented in (•). ❷ *Limit-Based Criterion*: For the CR metric, a lower CR value of forget data $\mathcal{D}_f$ indicates better unlearning performance while a higher CR value of $\mathcal{D}_{test}$ is desirable. For the MIACR metric, a higher MIACR value for forget data ($\mathcal{D}_f$) reflects improved unlearning effectiveness.

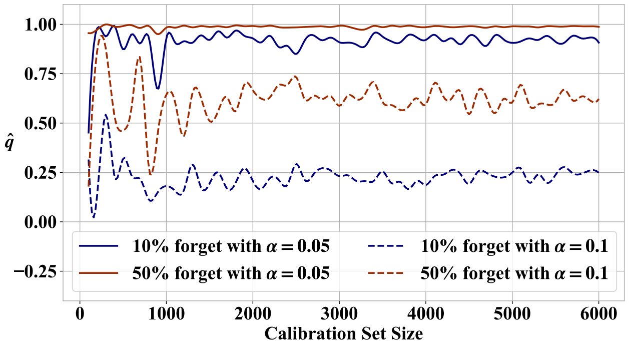


Figure 1: The stability of $\hat{q}$ in different calibration set sizes (results of RT method on CIFAR-10 with ResNet-18). When the calibration set size is greater than $2,000$, the fluctuation of $\hat{q}$ can be kept within a certain range.

2.3.2. Confidence Level and Calibration Data

In conformal prediction, miscoverage rate α and calibration set size are two crucial parameters for determining prediction set coverage and reliability. We next discuss the suitable settings for these two parameters and the rationale behind our chosen configuration.

In terms of α, a higher confidence level (which is $1-\alpha$) provides more reliable coverage but typically results in larger prediction sets. Common settings of confidence level are 0.95 and 0.9, corresponding to α values of 0.05 and 0.1, respectively. To make our work more broadly applicable and provide a comprehensive reference, we expand our experiments by setting α to a wider range of values: 0.05, 0.1, 0.15, and 0.2. Notably, **our default and recommended value for α is 0.1 and subsequent experimental analysis is based on $\alpha = 0.1$**. Higher values of α would result in smaller, potentially more precise prediction sets, but this reduction comes at the cost of decreased confidence, which may not be acceptable in many applications.

As for calibration data, it should be independent of train and test data. One requirement for the choice of the calibration set size is that the distribution of the calibration set should match the distribution of the test data as closely as possible. That means a small sample size may lead to distributional dissimilarity and unstable coverage estimation. Therefore, a sufficient sample size of calibration data must be ensured to obtain stable estimates. Figure 1 illustrates the stability of $\hat{q}$ across varying calibration set sizes. We implement them on CIFAR-10 with ResNet-18 in 10% and 50% random data forgetting scenarios. The results shown in Figure 1 are smoothed using B-spline. It reveals that $2,000$ is a sufficient calibration set size for different settings on CIFAR-10 with ResNet-18 to get a stable threshold $\hat{q}$. We also analyze

Table 4: Unlearning performance of 9 unlearning methods on **CIFAR-10** with **ResNet-18** in 10% **random data forgetting** scenario. The sign ↑ represents the greater is better, while ↓ denotes ideally small. The results are average values from 3 independent trials and the standard deviation values are reported in Appendix B. The performance gap relative to the RT method is represented in (•) in results tables. It shows the unlearning methods that excel under traditional metrics (UA, RA and TA) do not necessarily perform well under the CR metric.

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
UA8.6%(0.0), **RA**99.7%(0.0), **TA**91.8%(0.0)
	0.1
	0.2
		0.881(0.000)
		0.780(0.000)
			0.895(0.000)
			0.808(0.000)
				0.934(0.000)
				0.789(0.000)
					0.947(0.000)
					0.824(0.000)
						0.943(0.000)
						0.988(0.000)
							0.945(0.000)
							0.981(0.000)
								0.192
								0.003

FT
UA3.8%(4.8), **RA**98.1%(1.6), **TA**91.6%(0.2)
	0.1
	0.2
		0.968(0.087)
		0.861(0.081)
			0.899(0.004)
			0.806(0.002)
				0.969(0.035)
				0.861(0.072)
					0.924(0.023)
					0.811(0.013)
						0.998(0.055)
						1.000(0.012)
							0.972(0.027)
							0.993(0.012)
								0.079
								0.002

RL
UA7.6%(1.0), **RA**97.4%(2.3), **TA**90.6%(1.2)
	0.1
	0.2
		0.913(0.032)
		0.755(0.025)
			0.897(0.002)
			0.798(0.010)
				0.975(0.041)
				0.774(0.015)
					0.980(0.033)
					0.832(0.008)
						0.936(0.007)
						0.976(0.012)
							0.916(0.029)
							0.959(0.022)
								0.572
								0.234

GA
UA0.6%(8.0), **RA**99.5%(0.2), **TA**94.1%(2.3)
	0.1
	0.2
		0.990(0.109)
		0.925(0.145)
			0.905(0.010)
			0.805(0.003)
				0.990(0.056)
				0.924(0.135)
					0.928(0.019)
					0.811(0.013)
						0.998(0.055)
						0.998(0.010)
							0.973(0.028)
							0.992(0.011)
								0.062
								0.003

Teacher
UA0.8%(7.8), **RA**99.4%(0.3), **TA**93.5%(1.7)
	0.1
	0.2
		0.967(0.086)
		0.865(0.085)
			0.898(0.003)
			0.806(0.002)
				0.963(0.029)
				0.866(0.077)
					0.929(0.018)
					0.816(0.008)
						0.998(0.055)
						0.998(0.010)
							0.969(0.024)
							0.988(0.007)
								0.591
								0.426

FF
UA59.9%(51.3), **RA**40.1%(59.6), **TA**41.1%(50.7)
	0.1
	0.2
		0.933(0.052)
		0.835(0.055)
			0.899(0.004)
			0.794(0.014)
				7.129(6.195)
				5.750(4.961)
					6.566(5.619)
					5.219(4.395)
						0.131(0.812)
						0.145(0.843)
							0.137(0.808)
							0.153(0.828)
								0.998
								0.993

SSD
UA0.5%(8.1), **RA**99.5%(0.2), **TA**94.2%(2.4)
	0.1
	0.2
		0.987(0.106)
		0.922(0.142)
			0.902(0.007)
			0.803(0.005)
				0.990(0.056)
				0.923(0.134)
					0.926(0.021)
					0.811(0.013)
						0.998(0.055)
						1.002(0.014)
							0.973(0.028)
							0.992(0.011)
								0.063
								0.001

NegGrad+
UA8.7%(0.1), **RA**98.8%(0.9), **TA**92.2%(0.4)
	0.1
	0.2
		0.895(0.014)
		0.800(0.020)
			0.898(0.003)
			0.799(0.009)
				0.964(0.030)
				0.832(0.043)
					0.950(0.003)
					0.813(0.011)
						0.928(0.015)
						0.961(0.027)
							0.946(0.001)
							0.983(0.002)
								0.044
								0.000

Salun
UA3.7%(4.9), **RA**98.9%(0.8), **TA**91.8%(0.0))
	0.1
	0.2
		0.936(0.055)
		0.788(0.008)
			0.896(0.001)
			0.794(0.014)
				0.956(0.022)
				0.794(0.005)
					0.954(0.007)
					0.821(0.003)
						0.979(0.036)
						0.992(0.004)
							0.939(0.006)
							0.966(0.015)
								0.489
								0.221

the calibration set size choice of the class-wise forgetting scenario and find that 50 data points per class as a calibration set is enough. The reason is that, the model's prediction distribution becomes more complex in a dataset with more classes, requiring a larger calibration set to stabilize the estimation.

2.4. Evaluating Unlearning Methods with New Metrics

Having established the new metrics CR and MIACR, a natural question to ask is:

> **(Q3)** *Do existing machine unlearning methods truly achieve forgetting, and how do they perform under our new evaluation metrics?*

In this section, we assess the performance of various unlearning methods using the newly introduced metrics, specifically CR, together with coverage and set size. We employ 9 different unlearning methods, **RT**, **FT** (Warnecke et al., 2021), **RL** (Graves et al., 2021), **Gradient Ascent (GA)** (Thudi et al., 2022), **Teacher** (Tarun et al., 2023), **FisherForgetting (FF)** (Golatkar et al., 2020a), **SSD** (Foster et al., 2024), **NegGrad+** (Kurmanji et al., 2024) and **Salun** (Fan et al., 2023).

The experimental results are presented in Table 4, which summarizes the unlearning performance under 10% random data forgetting scenario on CIFAR-10 with ResNet-18. For more experimental results on CR and MIACR, please check Section 4 and Appendix B. The table 4 includes coverage percentages, set size, CR for both $\mathcal{D}_f$ and $\mathcal{D}_{test}$, as well as $\hat{q}$ in different settings of α. As observed, it shows a decreasing coverage and set size as the alpha increases in all methods since the threshold $\hat{q}$ is relaxed due to the reduced confidence level.

Based on evaluation criterion ❶, while NegGrad+ outperforms RL under the traditional UA metric, RL exhibits superior performance under the CR metric. This observation suggests that methods excelling in 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 limitation of UA. Based on evaluation criterion ❷, NegGrad+ and RL exhibit the best trade-off between forgetting quality and model performance. FF achieves the lowest CR, indicating the highest forgetting quality, but it exhibits the worst prediction performance on $\mathcal{D}_{test}$. On the other hand, GA, Teacher, SSD, FT and Salun maintain relatively high CR of the forget data $\mathcal{D}_f$ across different alpha settings, reflecting their poor forgetting quality.

3. Enhancing Machine Unlearning via Conformal Prediction

Our analysis indicates that machine unlearning methods generally do not exhibit strong performance under our newly developed evaluation metrics. Even methods with relatively low CR values demonstrate significant compromises in other metrics. This observation raises an important question:

> **(Q4)** *Can we explore advanced learning techniques via conformal prediction to optimize unlearning model's performance under our evaluation metrics?*

A training-based machine unlearning method is typically optimized for prediction loss but cannot efficiently support the improvement of forgetting quality in our new metrics. Therefore, we propose a novel and general unlearning framework for training-based machine unlearning methods to enhance machine unlearning. The key innovation of this framework lies in leveraging conformal prediction principles combined with a loss function inspired by the C&W attack (Carlini & Wagner, 2017) to improve forgetting quality.

The original C&W loss function is designed to find adversarial examples that cause misclassification. Here we extend it to an unlearning loss. For forget data $\mathcal{D}_f$, the objective of the unlearning loss function is to decrease the model's confidence in the true labels of $\mathcal{D}_f$. Based on this, the C&W-inspired unlearning loss can be defined as:

$$\mathcal{L}_{\text{cw}}(\boldsymbol{x}, y_t) = \max\left(p_t(\boldsymbol{x}) - \max_{i \neq t}\{p_i(\boldsymbol{x})\},\ -\Delta\right), \quad (7)$$

where $(\boldsymbol{x}, y_t) \in \mathcal{D}_f$. $p_i(\boldsymbol{x})$ is the possibility (after softmax function) of class y_i, while $z_t(\boldsymbol{x})$ is the true label y_t's possibility output. We denote $\max_{i \neq t}\{p_i(\boldsymbol{x})\}$ as the highest possibility value of the incorrect class. Δ is the margin parameter (also known as the confidence parameter) that controls the attack strength.

This loss is minimized when $p_t(\boldsymbol{x}) - \max_{i \neq t}\{p_i(\boldsymbol{x})\} \leq -\Delta$ to make $\max_{i \neq t}\{p_i(\boldsymbol{x})\} - p_t(\boldsymbol{x}) \geq \Delta$. It means that the loss maximizes the difference between the highest possibility value for class y_i $(i \neq t)$ and the possibility value for the true label y_t, which makes it more difficult for the model to include the true label in the prediction set during conformal prediction process. Additionally, increasing the value of Δ further reduces the confidence of the true label y_t being included in the prediction set. In our work, we set $\Delta = 0.01$.

Based on the insight discussed in Section 2, 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. Therefore, we reserve a portion of calibration data $\mathcal{D}'_c$ for the unlearning phase, which is kept separate from the calibration data $\mathcal{D}_c$ used in the evaluation phase.

According to Eq. 1, calibration data $\mathcal{D}'_c$ is used to calculate non-conformity scores and generate a threshold $\bar{q}$ based on an α. Then, we calculate the non-conformity scores of $\mathcal{D}_f$ to obtain the corresponding prediction set. Then, by revising C&W-inspired unlearning loss with this calibration step, a general unlearning loss function is defined as follows:

$$\mathcal{L}_{\text{unlearn}}(\boldsymbol{x}, y_t) = \max\left(S(\boldsymbol{x}, y_t) - \bar{q},\ -\Delta\right), \quad (8)$$

where $(\boldsymbol{x}, y_t) \in \mathcal{D}_f$. We replace $p_t(\boldsymbol{x})$ in Eq. 7 with the non-conformity score $S(\boldsymbol{x}, y_t)$ of true label y_t, and replace $\max_{i \neq t}\{p_i(\boldsymbol{x})\}$ with $\bar{q}$ which will be updated before each epoch. It ensures that $\mathcal{L}_{\text{unlearn}}$ adheres to the principle of $\mathcal{L}_{\text{cw}}$, which encourages $\bar{q} - S(\boldsymbol{x}, y_t) \geq \Delta$.

It is a general framework applicable across various training-based machine unlearning methods. To preserve the effectiveness of specific machine unlearning methods, we retain their original loss $\mathcal{L}_{\text{original}}$. Consequently, we sum these terms to form the final objective loss function as follows:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{original}} + \lambda \cdot \mathcal{L}_{\text{unlearn}}, \quad (9)$$

where λ is a hyperparameter that controls the forgetting degree.

4. Experiment

4.1. Experimental setting

Datasets and Models. We report experiments on CIFAR-10 (Krizhevsky, 2009) with ResNet-18 (He et al., 2016) and Tiny ImageNet (Le & Yang, 2015) with ViT (Dosovitskiy et al., 2021).

Implementation Details. For CIFAR-10/Tiny ImageNet, we randomly select 200/50 data points per class ($2,000/10,000$ 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 $\bar{q}$.

For the hyperparameter in our work, miscoverage rate $\alpha \in [0.05, 0.1, 0.15, 0.2]$, margin parameter $\Delta = 0.01$, unlearning loss weight $\lambda \in [0, 0.2, 0.5, 1]$. Due to space limitations, **more training details and experimental results under settings of $\boldsymbol{\alpha \in [0.05, 0.15]}$, $\boldsymbol{\lambda = 0.2}$ and class-wise forgetting scenario can be found in Appendix A-C**.

4.2. Measure Machine Unlearning Method via New Metrics

CR Metric. For the CR metric, we already show the results of CIFAR-10 with ResNet-18 in Tabel 4. Here, we further evaluate the unlearning performance in Tiny ImageNet with ViT under 10% random data forgetting scenario by CR metric. The results of Tiny ImageNet with ViT and CIFAR-10 with ReNet18 in 50% random data forgetting and class-wise forgetting scenarios can be found in Appendix B. We remark that it is computationally intractable for the FF method in Tiny ImageNet with ViT, so we omit its results. The results of Tiny ImageNet with ViT in 10% random data forgetting scenario are summarized in Table 5. Our findings reveal the following insights:

For all machine 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

Table 5: Performance of 8 unlearning methods on **Tiny ImageNet** with **ViT** in 10% **random data forgetting** scenario.

Methods	α	Coverage		Set Size		CR		$\hat{q}$
		$\mathcal{D}_f \downarrow$	$\mathcal{D}_{test} \uparrow$	$\mathcal{D}_f \uparrow$	$\mathcal{D}_{test} \downarrow$	$\mathcal{D}_f \downarrow$	$\mathcal{D}_{test} \uparrow$	
RT	0.1	0.892(0.000)	0.900(0.000)	1.151(0.000)	1.144(0.000)	0.775(0.000)	0.786(0.000)	0.853
UA14.7%(0.0), **RA**98.8%(0.0), **TA**86.0%(0.0)	0.2	0.790(0.000)	0.799(0.000)	0.846(0.000)	0.854(0.000)	0.934(0.000)	0.935(0.000)	0.238
FT	0.1	0.978(0.086)	0.903(0.003)	1.234(0.083)	1.317(0.173)	0.792(0.017)	0.685(0.101)	0.935
UA6.9%(7.8), **RA**97.9%(0.9), **TA**84.1%(1.9)	0.2	0.888(0.098)	0.801(0.002)	0.915(0.069)	0.885(0.031)	0.970(0.036)	0.905(0.030)	0.326
RL	0.1	0.892(0.000)	0.902(0.002)	2.639(1.488)	1.843(0.699)	0.338(0.437)	0.489(0.297)	0.971
UA26.9%(12.2), **RA**96.0%(2.8), **TA**81.4%(4.6)	0.2	0.681(0.109)	0.803(0.004)	0.831(0.015)	0.946(0.092)	0.820(0.114)	0.849(0.086)	0.715
GA	0.1	0.986(0.094)	0.900(0.000)	1.104(0.047)	1.224(0.080)	0.894(0.119)	0.736(0.050)	0.899
UA3.2%(11.5), **RA**97.4%(1.4), **TA**84.9%(1.1)	0.2	0.934(0.144)	0.800(0.001)	0.946(0.100)	0.871(0.017)	0.987(0.053)	0.919(0.016)	0.296
Teacher	0.1	0.930(0.038)	0.902(0.002)	1.991(0.840)	1.959(0.815)	0.467(0.308)	0.460(0.326)	0.971
UA17.3%(2.6), **RA**86.7%(12.1), **TA**79.0%(7.0)	0.2	0.816(0.026)	0.803(0.004)	1.020(0.174)	1.058(0.204)	0.800(0.134)	0.758(0.177)	0.910
SSD	0.1	0.993(0.101)	0.897(0.003)	1.039(0.112)	1.134(0.010)	0.956(0.181)	0.791(0.005)	0.852
UA1.5%(13.2), **RA**98.5%(0.3), **TA**86.1%(0.1)	0.2	0.956(0.166)	0.805(0.006)	0.960(0.114)	0.864(0.010)	0.996(0.062)	0.932(0.003)	0.249
NegGrad+	0.1	0.995(0.103)	0.848(0.052)	0.898(0.253)	1.093(0.051)	1.225(0.450)	1.287(0.501)	0.933
UA19.4%(4.7), **RA**98.3%(0.5), **TA**84.0%(2.0)	0.2	0.966(0.176)	0.783(0.016)	0.802(0.044)	0.972(0.118)	0.922(0.012)	0.891(0.043)	0.320
Salun	0.1	0.977(0.085)	0.924(0.024)	1.229(0.078)	1.281(0.137)	0.918(0.143)	0.884(0.097)	0.939
UA9.2%(5.5), **RA**97.7%(1.1), **TA**83.6%(2.4)	0.2	0.870(0.080)	0.810(0.011)	0.845(0.001)	0.925(0.071)	0.924(0.009)	0.894(0.041)	0.630

Table 6: **MIACR** performance on **CIFAR-10** with **ResNet-18**. We show the MIA value in **10% random data forgetting** scenario, while the MIA results in 50% scenario are presented in Appendix B.

Methods	α	10% Forgetting		50% Forgetting	
		MIACR $\uparrow$	$\hat{q}$	MIACR $\uparrow$	$\hat{q}$
RT	0.1	0.147(0.000)	0.589	0.201(0.000)	0.570
MIA86.9%(0.0)	0.2	0.246(0.000)	0.473	0.318(0.000)	0.459
FT	0.1	0.077(0.070)	0.627	0.103(0.098)	0.558
MIA92.0%(5.1)	0.2	0.196(0.050)	0.483	0.244(0.074)	0.476
RL	0.1	0.178(0.031)	0.572	0.137(0.064)	0.547
MIA74.2%(12.7)	0.2	0.320(0.074)	0.485	0.261(0.057)	0.546
GA	0.1	0.032(0.115)	0.502	0.055(0.146)	0.486
MIA98.8%(11.9)	0.2	0.146(0.100)	0.476	0.164(0.154)	0.473
Teacher	0.1	0.038(0.109)	0.672	0.065(0.136)	0.582
MIA87.2%(0.3)	0.2	0.113(0.133)	0.588	0.159(0.159)	0.532
FF	0.1	0.051(0.096)	0.486	0.089(0.112)	0.509
MIA71.5%(15.4)	0.2	0.109(0.137)	0.473	0.168(0.150)	0.499
SSD	0.1	0.031(0.116)	0.511	0.051(0.150)	0.488
MIA98.8%(11.9)	0.2	0.139(0.107)	0.475	0.168(0.150)	0.477
NegGrad+	0.1	0.128(0.019)	0.481	0.109(0.092)	0.511
MIA90.3%(3.4)	0.2	0.213(0.033)	0.480	0.230(0.088)	0.472
Salun	0.1	0.113(0.034)	0.681	0.115(0.086)	0.630
MIA57.6%(19.3)	0.2	0.267(0.021)	0.608	0.220(0.098)	0.586

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 generally becomes larger as α increases.

When α is set to 0.2, most methods show a value of Set Size less than 1 in both Table 4 and 5. 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 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.

Notably, when $\alpha = 0.2$, RL's coverage on the forget data $\mathcal{D}_f$ decreases significantly compared to other methods. This results from the use of label corruption during RL's unlearning process, which enlarges the distribution difference between calibration data and forget data. We can mitigate this gap if we know the details of specific unlearning methods, and a potential solution to this issue is discussed in Appendix B.

Regarding the forgetting quality of specific machine unlearning methods, the Teacher method achieves the strongest performance on the traditional UA metric under both evaluation criteria ❶ and ❷. However, its forgetting quality is significantly weaker when measured by the CR metric. This 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. By explicitly addressing this phenomenon, CR ensures a more robust and reliable evaluation for unlearning quality.

MIACR Metric. In Table 6, we show the MIACR results on CIFAR-10 under both 10% and 50% random data forgetting scenarios. Compared to MIACR, traditional MIA remains a privacy concern. Albeit MIA cannot predict some forget data points as training data, it still has high confidence in appearing in the prediction set. For example, under evaluation criterion ❶, the teacher's MIA is the method closest to the RT method, but it performs poorly in the CR metric. Similarly, FF and Salun are the top two unlearning methods in the traditional MIA metric (ideally small) under evaluation criterion ❷, but FF has low MIACR (ideally greater) values in both 10% and 50% forgetting scenarios.

4.3. Performance of Our Unlearning Framework

In this experiment, we apply the RT, FT, and RL methods to our framework. Table 7 presents the results for CIFAR-10 with ResNet-18 and Tiny ImageNet with ViT in 10% random data forgetting scenario.

Table 7: Performance of our unlearning framework. We show the performance on **CIFAR-10** with **ResNet-18** and **Tiny ImageNet** with **ViT** in **10% random data forgetting** scenario. $\lambda = 0$ represents the baseline without our framework applied. It shows our framework significantly improves the forgetting quality, while preserving stable predictive performance.

Methods	α	$\boldsymbol{\lambda = 0}$					$\boldsymbol{\lambda = 0.2}$					$\boldsymbol{\lambda = 0.5}$				
		UA ↑	RA ↑	TA ↑	$\mathrm{CR}_{\mathcal{D}_f}$ ↓	$\mathrm{CR}_{\mathcal{D}_{test}}$ ↑	UA ↑	RA ↑	TA ↑	$\mathrm{CR}_{\mathcal{D}_f}$ ↓	$\mathrm{CR}_{\mathcal{D}_{test}}$ ↑	UA ↑	RA ↑	TA ↑	$\mathrm{CR}_{\mathcal{D}_f}$ ↓	$\mathrm{CR}_{\mathcal{D}_{test}}$ ↑
CIFAR-10 with ResNet-18																
RT	0.1 0.2	8.6%(0.0)	99.7%(0.0)	91.8%(0.0)	0.943(0.000) 0.988(0.000)	0.945(0.000) 0.981(0.000)	10.8%(2.2)	98.3%(1.4)	91.0%(0.8)	0.914(0.029) 0.977(0.011)	0.924(0.021) 0.976(0.005)	14.0%(5.4)	97.8%(1.9)	90.4%(0.4)	0.879(0.064) 0.963(0.025)	0.912(0.033) 0.966(0.015)
FT	0.1 0.2	3.8%(4.8)	98.1%(1.6)	91.6%(0.2)	0.998(0.055) 1.000(0.012)	0.972(0.027) 0.993(0.012)	6.8%(1.8)	97.0%(2.7)	90.8%(1.0)	0.948(0.005) 0.989(0.001)	0.924(0.021) 0.974(0.007)	7.9%(0.7)	96.9%(2.8)	90.9%(0.9)	0.940(0.003) 0.983(0.005)	0.927(0.018) 0.975(0.006)
RL	0.1 0.2	7.6%(1.0)	97.4%(2.3)	90.6%(1.2)	0.936(0.007) 0.976(0.012)	0.916(0.029) 0.959(0.022)	9.7%(1.1)	96.6%(3.1)	89.4%(2.4)	0.896(0.047) 0.964(0.024)	0.887(0.058) 0.949(0.032)	9.9%(1.3)	96.9%(2.8)	89.7%(2.1)	0.902(0.041) 0.959(0.029)	0.896(0.049) 0.950(0.031)
Tiny ImageNet with ViT																
RT	0.1 0.2	14.7%(0.0)	98.8%(0.0)	86.0%(0.0)	0.775(0.000) 0.934(0.000)	0.786(0.000) 0.935(0.000)	19.3%(4.6)	98.8%(0.0)	86.0%(0.0)	0.729(0.163) 0.898(0.108)	0.786(0.114) 0.932(0.133)	26.4%(11.7)	98.7%(0.1)	85.8%(0.2)	0.649(0.243) 0.839(0.049)	0.765(0.135) 0.929(0.130)
FT	0.1 0.2	6.9%(7.8)	97.9%(0.9)	84.1%(1.9)	0.791(0.087) 0.969(0.005)	0.685(0.201) 0.905(0.053)	9.8%(4.9)	97.4%(1.4)	83.6%(2.4)	0.753(0.139) 0.942(0.152)	0.683(0.217) 0.893(0.094)	13.6%(0.9)	97.2%(1.6)	83.6%(2.4)	0.718(0.174) 0.914(0.124)	0.683(0.217) 0.890(0.091)
RL	0.1 0.2	26.9%(12.2)	96.0%(2.8)	81.4%(4.6)	0.338(0.540) 0.821(0.143)	0.488(0.398) 0.848(0.110)	31.8%(17.1)	95.3%(17.9)	80.9%(5.1)	0.278(0.614) 0.752(0.038)	0.451(0.449) 0.825(0.026)	36.2%(21.5)	95.3%(3.5)	80.4%(5.6)	0.254(0.638) 0.718(0.072)	0.449(0.451) 0.827(0.028)

We vary λ in the range $[0, 0.2, 0.5]$, where $\lambda = 0$ represents the baseline without our framework applied. The experimental results demonstrate a significant improvement in UA across all methods and a marked decrease in $\mathrm{CR}_{\mathcal{D}_f}$, reflecting improved forgetting quality as λ increases. Notably, the RA, TA, and $\mathrm{CR}_{\mathcal{D}_{test}}$ values remain relatively stable, indicating that the substantial improvement in forgetting quality does not compromise the model's predictive performance.

From the perspective of evaluation criterion ❶, the FT method achieves an average gap of 1.47 across UA, RA, and TA on ResNet-18 when $\lambda = 0.5$, compared to an average gap of 2.2 when $\lambda = 0$. Similarly, on the ViT model, FT reduces the average gap from 3.53 to 1.63 when $\lambda = 0.5$. If UA of an unlearning method is already lower than that of RT, e.g., RL, our framework cannot further improve their performance under 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, while TA decreases only slightly by 1.0% and 0.57% on ResNet-18 and ViT respectively. As similarly shown in CR metric, the value of $\mathrm{CR}_{\mathcal{D}_{test}}$ remains nearly unchanged compared to the baseline ($\lambda = 0$), while $\mathrm{CR}_{\mathcal{D}_f}$ shows a significant reduction. It can be found that this advantage is more obvious when $\lambda = 1$ as shown in Table 16. In summary, the experimental results demonstrate that our framework notably enhances the forgetting quality while maintaining stable 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; Warnecke et al., 2021; Thudi et al., 2022; Tarun et al., 2023) and training-free approaches (Guo et al., 2019; Sekhari et al., 2021; Nguyen et al., 2020; Golatkar et al., 2020b; 2021; Foster et al., 2024).

To evaluate these methods, several unlearning metrics have been proposed, including UA (Brophy & Lowd, 2021; Foster et al., 2024) and MIA (Shokri et al., 2017; Chen et al., 2021). 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 (Papadopoulos et al., 2002; Lei & Wasserman, 2014; Romano et al., 2020; Lei et al., 2018).

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 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 addiction, 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.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

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

Becker, A. and Liebig, T. Evaluating machine unlearning via epistemic uncertainty. *arXiv preprint arXiv:2208.10836*, 2022.

Bourtoule, L., Chandrasekaran, V., Choquette-Choo, C. A., Jia, H., Travers, A., Zhang, B., Lie, D., and Papernot, N. Machine unlearning. In *2021 IEEE Symposium on Security and Privacy (SP)*, pp. 141–159. IEEE, 2021.

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

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

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

Chen, M., Zhang, Z., Wang, T., Backes, M., Humbert, M., and Zhang, Y. When machine unlearning jeopardizes privacy. In *Proceedings of the 2021 ACM SIGSAC conference on computer and communications security*, pp. 896–911, 2021.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. URL https://arxiv.org/abs/2010.11929.

Fan, C., Liu, J., Zhang, Y., Wong, E., Wei, D., and Liu, S. Salun: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation. *arXiv preprint arXiv:2310.12508*, 2023.

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

Golatkar, A., Achille, A., and Soatto, S. Eternal sunshine of the spotless net: Selective forgetting in deep networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9304–9312, 2020a.

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

Golatkar, A., Achille, A., Ravichandran, A., Polito, M., and Soatto, S. Mixed-privacy forgetting in deep networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 792–801, 2021.

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

Guo, C., Goldstein, T., Hannun, A., and Van Der Maaten, L. Certified data removal from machine learning models. *arXiv preprint arXiv:1911.03030*, 2019.

He, K., Zhang, X., Ren, S., and Sun, J. Identity mappings in deep residual networks. In Leibe, B., Matas, J., Sebe, N., and Welling, M. (eds.), *Computer Vision – ECCV 2016*, pp. 630–645, Cham, 2016. Springer International Publishing. ISBN 978-3-319-46493-0.

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

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

Kurmanji, M., Triantafillou, E., and Triantafillou, P. Machine unlearning in learned databases: An experimental analysis. *Proc. ACM Manag. Data*, 2(1), March 2024. doi: 10.1145/3639304. URL https://doi.org/10.1145/3639304.

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

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

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., and Wasserman, L. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018.

Nguyen, Q. P., Low, B. K. H., and Jaillet, P. Variational bayesian unlearning. *Advances in Neural Information Processing Systems*, 33:16025–16036, 2020.

Papadopoulos, H., Proedrou, K., Vovk, V., and Gammerman, A. 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.

Romano, Y., Sesia, M., and Candes, E. Classification with valid and adaptive coverage. *Advances in Neural Information Processing Systems*, 33:3581–3591, 2020.

Sekhari, A., Acharya, J., Kamath, G., and Suresh, A. T. Remember what you want to forget: Algorithms for machine unlearning. *Advances in Neural Information Processing Systems*, 34:18075–18086, 2021.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. 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.

Shokri, R., Stronati, M., Song, C., and Shmatikov, V. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pp. 3–18. IEEE, 2017.

Tarun, A. K., Chundawat, V. S., Mandal, M., and Kankanhalli, M. Deep regression unlearning. In *International Conference on Machine Learning*, pp. 33921–33939. PMLR, 2023.

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

Warnecke, A., Pirch, L., Wressnegger, C., and Rieck, K. Machine unlearning of features and labels. *arXiv preprint arXiv:2108.11577*, 2021.

A. Setting Details

For CIFAR-10 with a 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 momentum to 0.9 and a batch size of 64. 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. In class-wise forgetting, in each dataset, we designate the same class as the forget data $\mathcal{D}_f$ across all unlearning methods, with the remaining classes used as the retain data $\mathcal{D}_r$.

B. Evaluating MU methods

B.1. Mis-label Number and In-set Ratios

Table 8: Mis-label number and in-set ratios of UA and MIA metrics for 10 unlearning methods.

	10% Forgetting			**50% Forgetting**		
Methods	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%
FF	2,995	2,751	91.9%	15,083	14,061	93.2%
SSD	25	2	8.0%	116	9	7.8%
NegGrad+	435	115	26.4%	711	249	35.5%
Salun	185	117	63.2%	1065	695	65.3%
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%
FF	1,424	1,424	100%	5,996	4,850	80.9%
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%

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) across 9 unlearning methods. 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 8.

B.2. CR Metric

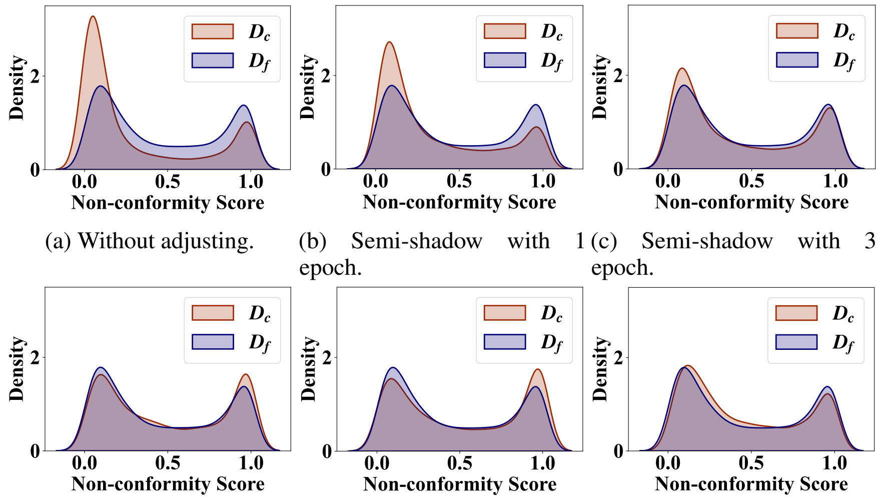
Tables 9 and 10 show the unlearning performance of seven unlearning methods on CIFAR-10 with ResNet-18 in 10% and 50% random data forgetting scenarios, while Table 11 is the results in class-wise forgetting scenario. Tables 12 and 13 present the unlearning performance on Tiny ImageNet with ResNet-18 in random data forgetting scenario, while Table 11 details the unlearning performance in class-wise forgetting scenario.

For all machine unlearning methods, increasing the α level leads to lower Coverage and smaller Set Size. Conversely, the CR tends to grow as α increases. Most methods yield a Set Size of less than 1 when α is set too low, since conformal prediction is fundamentally linked to the model's baseline prediction performance (accuracy). If the model's accuracy significantly exceeds the confidence level, conformal prediction can easily achieve the required coverage. In fact, it may produce empty prediction sets for certain data points while still satisfying the coverage target. Excessively high α values can distort evaluation results, preventing the CR metric from accurately reflecting the model's performance. Therefore, the choice of α should be carefully considered in relation to the model's inherent performance. The more experimental results for random data forgetting scenarios have already been analyzed in detail in Section 4 and will not be reiterated here.

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.

B.3. MIACR Metric

Table 15 presents the performance of seven 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 $\alpha \in [0.05, 0.15]$ in Table 15.



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

B.4. How can we better measure forgetting under distribution shifts

Recall that the coverage value of RL has a relatively large deviation in Table 5 since it employs label corruption in its unlearning strategy which can cause distribution shifts. Here, we introduce how to better measure forgetting under these circumstances.

Figure 2(a) shows the non-conformity score distribution of calibration data $\mathcal{D}_c$ and forget data $\mathcal{D}_f$ in unlearning model θ_u obtained by 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:

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 2, where (b)-(e) present the results of the semi-shadow model with different epochs and

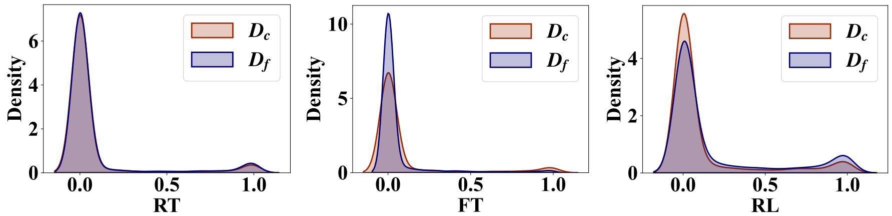


Figure 3: Non-conformity density of calibration data $\mathcal{D}_c$ and forget data $\mathcal{D}_f$ **without our unlearning framework** in CIFAR-10 with ResNet-18 under 10% random data forgetting scenario.

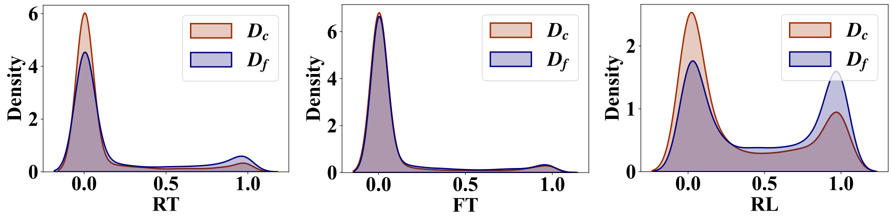


Figure 4: Non-conformity score density of calibration data $\mathcal{D}_c$ and forget data $\mathcal{D}_f$ **with our unlearning framework** in CIFAR-10 with ResNet-18 under 10% random data forgetting scenario. Our unlearning framework shifts the distribution of the forget data to the right, demonstrating improved forgetting quality.

(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.

C. Performance of Our Unlearning Framework

Table 16 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 9 and 12. 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 3 and 4, 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 3 and 4, it is evident that after applying our unlearning framework, the distribution of non-conformity scores for forget data shifts noticeably to the right. This demonstrates the effectiveness of our framework in enhancing forgetting quality.

Table 9: Unlearning performance of 10 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 Retrain 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)	$0.426_{\pm 0.007}$
FF **UA**59.9%, **RA**40.1%, **TA**41.1%	0.05	$0.973_{\pm 0.009}$(0.032)	$0.949_{\pm 0.001}$(0.005)	$7.966_{\pm 0.212}$(6.877)	$7.408_{\pm 0.000}$(6.334)	$0.122_{\pm 0.002}$(0.742)	$0.128_{\pm 0.000}$(0.751)	$0.999_{\pm 0.000}$
	0.1	$0.933_{\pm 0.020}$(0.052)	$0.899_{\pm 0.001}$(0.004)	$7.129_{\pm 0.148}$(6.195)	$6.566_{\pm 0.166}$(5.619)	$0.131_{\pm 0.000}$(0.812)	$0.137_{\pm 0.004}$(0.808)	$0.998_{\pm 0.001}$
	0.15	$0.888_{\pm 0.029}$(0.068)	$0.852_{\pm 0.008}$(0.013)	$6.431_{\pm 0.078}$(5.590)	$5.903_{\pm 0.262}$(5.036)	$0.138_{\pm 0.003}$(0.837)	$0.144_{\pm 0.008}$(0.824)	$0.996_{\pm 0.002}$
	0.2	$0.835_{\pm 0.048}$(0.055)	$0.794_{\pm 0.012}$(0.014)	$5.750_{\pm 0.034}$(4.961)	$5.219_{\pm 0.368}$(4.395)	$0.145_{\pm 0.007}$(0.843)	$0.153_{\pm 0.013}$(0.828)	$0.993_{\pm 0.003}$
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)	$0.886_{\pm 0.010}$(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 10: Unlearning performance of 10 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)	$0.793_{\pm 0.012}$
FF **UA**60.0%, **RA**40.1%, **TA**40.6%	0.05	$0.972_{\pm 0.006}$(0.017)	$0.954_{\pm 0.002}$(0.006)	$8.023_{\pm 0.189}$(6.736)	$7.461_{\pm 0.019}$(6.247)	$0.121_{\pm 0.002}$(0.621)	$0.128_{\pm 0.000}$(0.652)	$0.999_{\pm 0.000}$
	0.1	$0.930_{\pm 0.020}$(0.032)	$0.897_{\pm 0.003}$(0.007)	$7.091_{\pm 0.151}$(6.068)	$6.521_{\pm 0.171}$(5.501)	$0.131_{\pm 0.000}$(0.747)	$0.138_{\pm 0.004}$(0.748)	$0.998_{\pm 0.001}$
	0.15	$0.887_{\pm 0.024}$(0.054)	$0.852_{\pm 0.006}$(0.005)	$6.402_{\pm 0.034}$(5.519)	$5.837_{\pm 0.438}$(4.931)	$0.139_{\pm 0.004}$(0.804)	$0.146_{\pm 0.010}$(0.788)	$0.996_{\pm 0.001}$
	0.2	$0.840_{\pm 0.046}$(0.058)	$0.803_{\pm 0.015}$(0.012)	$5.805_{\pm 0.042}$(4.994)	$5.253_{\pm 0.376}$(4.403)	$0.145_{\pm 0.007}$(0.819)	$0.153_{\pm 0.014}$(0.804)	$0.993_{\pm 0.003}$
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 11: Unlearning performance of 10 unlearning methods on **CIFAR-10** with **ResNet18** in **class-wise forgetting** scenario.

Methods	α	Coverage			Set Size			CR				
		$\mathcal{D}_f \downarrow$	$\mathcal{D}_{tf} \downarrow$	$\mathcal{D}_{tr} \uparrow$	$\mathcal{D}_f \uparrow$	$\mathcal{D}_{tf} \uparrow$	$\mathcal{D}_{tr} \downarrow$	$\mathcal{D}_f \downarrow$	$\mathcal{D}_{tf} \downarrow$	$\mathcal{D}_{tr} \uparrow$	$\hat{q}_f$	$\hat{q}_{test}$
RT **UA**100%, **UA**$_{tf}$100%, **RA**99.9%, **TA**92.4%	0.05	$1.000_{\pm0.001}$(0.000)	$1.000_{\pm0.001}$(0.000)	$0.964_{\pm0.008}$(0.000)	$10.000_{\pm0.000}$(0.000)	$10.000_{\pm0.000}$(0.000)	$1.148_{\pm0.013}$(0.000)	$0.100_{\pm0.000}$(0.000)	$0.100_{\pm0.000}$(0.000)	$0.840_{\pm0.002}$(0.000)	$1.000_{\pm0.000}$	$0.982_{\pm0.003}$
	0.1	$1.000_{\pm0.000}$(0.000)	$1.000_{\pm0.001}$(0.000)	$0.882_{\pm0.011}$(0.000)	$10.000_{\pm0.000}$(0.000)	$10.000_{\pm0.000}$(0.000)	$0.922_{\pm0.009}$(0.000)	$0.100_{\pm0.000}$(0.000)	$0.100_{\pm0.001}$(0.000)	$0.956_{\pm0.007}$(0.000)	$1.000_{\pm0.001}$	$0.080_{\pm0.003}$
	0.15	$1.000_{\pm0.000}$(0.000)	$1.000_{\pm0.000}$(0.000)	$0.856_{\pm0.012}$(0.000)	$10.000_{\pm0.000}$(0.000)	$10.000_{\pm0.000}$(0.000)	$0.882_{\pm0.007}$(0.000)	$0.100_{\pm0.001}$(0.000)	$0.100_{\pm0.001}$(0.000)	$0.970_{\pm0.004}$(0.000)	$1.000_{\pm0.000}$	$0.018_{\pm0.010}$
	0.2	$1.000_{\pm0.000}$(0.000)	$1.000_{\pm0.000}$(0.000)	$0.814_{\pm0.010}$(0.000)	$10.000_{\pm0.000}$(0.000)	$10.000_{\pm0.000}$(0.000)	$0.830_{\pm0.001}$(0.000)	$0.100_{\pm0.001}$(0.000)	$0.100_{\pm0.001}$(0.000)	$0.981_{\pm0.002}$(0.000)	$1.000_{\pm0.000}$	$0.003_{\pm0.001}$
FT **UA**100%, **UA**$_{tf}$100%, **RA**96.7%, **TA**90.8%	0.05	$0.994_{\pm0.003}$(0.006)	$0.962_{\pm0.022}$(0.038)	$0.944_{\pm0.011}$(0.020)	$9.854_{\pm0.127}$(0.146)	$9.403_{\pm0.501}$(0.597)	$1.045_{\pm0.040}$(0.103)	$0.101_{\pm0.001}$(0.001)	$0.102_{\pm0.003}$(0.002)	$0.904_{\pm0.028}$(0.065)	$1.000_{\pm0.000}$	$0.731_{\pm0.166}$
	0.1	$0.969_{\pm0.011}$(0.031)	$0.882_{\pm0.020}$(0.118)	$0.908_{\pm0.010}$(0.026)	$9.495_{\pm0.255}$(0.505)	$8.528_{\pm0.571}$(1.472)	$0.956_{\pm0.006}$(0.034)	$0.102_{\pm0.002}$(0.002)	$0.104_{\pm0.005}$(0.004)	$0.950_{\pm0.007}$(0.006)	$1.000_{\pm0.000}$	$0.314_{\pm0.010}$
	0.15	$0.951_{\pm0.014}$(0.049)	$0.840_{\pm0.011}$(0.160)	$0.851_{\pm0.031}$(0.005)	$9.265_{\pm0.279}$(0.735)	$8.131_{\pm0.523}$(1.869)	$0.872_{\pm0.039}$(0.010)	$0.103_{\pm0.003}$(0.003)	$0.103_{\pm0.007}$(0.003)	$0.976_{\pm0.009}$(0.006)	$1.000_{\pm0.000}$	$0.073_{\pm0.054}$
	0.2	$0.942_{\pm0.014}$(0.058)	$0.818_{\pm0.072}$(0.182)	$0.838_{\pm0.016}$(0.023)	$9.163_{\pm0.245}$(0.837)	$7.934_{\pm0.533}$(2.066)	$0.854_{\pm0.019}$(0.024)	$0.103_{\pm0.003}$(0.003)	$0.103_{\pm0.010}$(0.003)	$0.981_{\pm0.005}$(0.000)	$1.000_{\pm0.000}$	$0.039_{\pm0.017}$
RL **UA**100%, **UA**$_{tf}$100%, **RA**98.0%, **TA**92.7%	0.05	$0.995_{\pm0.002}$(0.005)	$0.954_{\pm0.009}$(0.046)	$0.959_{\pm0.015}$(0.005)	$9.993_{\pm0.003}$(0.007)	$9.900_{\pm0.011}$(0.100)	$1.170_{\pm0.155}$(0.022)	$0.100_{\pm0.000}$(0.000)	$0.096_{\pm0.001}$(0.004)	$0.828_{\pm0.097}$(0.012)	$1.000_{\pm0.000}$	$0.870_{\pm0.145}$
	0.1	$0.984_{\pm0.003}$(0.016)	$0.907_{\pm0.015}$(0.093)	$0.918_{\pm0.021}$(0.036)	$9.978_{\pm0.004}$(0.022)	$9.800_{\pm0.019}$(0.200)	$0.982_{\pm0.036}$(0.059)	$0.099_{\pm0.000}$(0.001)	$0.093_{\pm0.002}$(0.007)	$0.936_{\pm0.022}$(0.021)	$1.000_{\pm0.000}$	$0.469_{\pm0.250}$
	0.15	$0.961_{\pm0.009}$(0.039)	$0.859_{\pm0.014}$(0.141)	$0.870_{\pm0.019}$(0.014)	$9.950_{\pm0.017}$(0.050)	$9.700_{\pm0.066}$(0.300)	$0.904_{\pm0.045}$(0.021)	$0.097_{\pm0.001}$(0.003)	$0.089_{\pm0.001}$(0.011)	$0.964_{\pm0.027}$(0.006)	$1.000_{\pm0.000}$	$0.144_{\pm0.163}$
	0.2	$0.935_{\pm0.027}$(0.065)	$0.815_{\pm0.012}$(0.185)	$0.804_{\pm0.016}$(0.010)	$9.919_{\pm0.035}$(0.081)	$9.637_{\pm0.076}$(0.363)	$0.820_{\pm0.026}$(0.010)	$0.094_{\pm0.002}$(0.006)	$0.085_{\pm0.001}$(0.015)	$0.981_{\pm0.012}$(0.000)	$0.999_{\pm0.001}$	$0.014_{\pm0.013}$
GA **UA**84.6%, **UA**$_{tf}$82.5%, **RA**96.4%, **TA**89.6%	0.05	$1.000_{\pm0.003}$(0.000)	$1.000_{\pm0.005}$(0.000)	$0.948_{\pm0.004}$(0.016)	$10.000_{\pm0.009}$(0.000)	$10.000_{\pm0.003}$(0.000)	$1.204_{\pm0.002}$(0.056)	$0.100_{\pm0.007}$(0.000)	$0.100_{\pm0.011}$(0.000)	$0.787_{\pm0.011}$(0.053)	$1.000_{\pm0.010}$	$0.988_{\pm0.000}$
	0.1	$1.000_{\pm0.003}$(0.000)	$1.000_{\pm0.010}$(0.000)	$0.899_{\pm0.008}$(0.017)	$10.000_{\pm0.005}$(0.000)	$10.000_{\pm0.006}$(0.000)	$1.005_{\pm0.003}$(0.083)	$0.100_{\pm0.012}$(0.000)	$0.100_{\pm0.006}$(0.000)	$0.894_{\pm0.002}$(0.062)	$1.000_{\pm0.000}$	$0.562_{\pm0.003}$
	0.15	$1.000_{\pm0.006}$(0.000)	$1.000_{\pm0.001}$(0.000)	$0.843_{\pm0.011}$(0.013)	$10.000_{\pm0.005}$(0.000)	$10.000_{\pm0.006}$(0.000)	$0.893_{\pm0.010}$(0.011)	$0.100_{\pm0.004}$(0.000)	$0.100_{\pm0.008}$(0.000)	$0.944_{\pm0.007}$(0.026)	$1.000_{\pm0.001}$	$0.051_{\pm0.002}$
	0.2	$0.828_{\pm0.003}$(0.172)	$0.782_{\pm0.011}$(0.218)	$0.838_{\pm0.010}$(0.024)	$9.550_{\pm0.007}$(0.450)	$9.366_{\pm0.002}$(0.634)	$0.884_{\pm0.000}$(0.054)	$0.087_{\pm0.008}$(0.013)	$0.084_{\pm0.005}$(0.016)	$0.948_{\pm0.010}$(0.033)	$1.000_{\pm0.002}$	$0.038_{\pm0.003}$
Teacher **UA**90.1%, **UA**$_{tf}$86.5%, **RA**99.5%, **TA**94.0%	0.05	$0.994_{\pm0.003}$(0.006)	$0.959_{\pm0.002}$(0.041)	$0.939_{\pm0.003}$(0.025)	$9.877_{\pm0.000}$(0.123)	$9.502_{\pm0.003}$(0.498)	$1.000_{\pm0.004}$(0.148)	$0.101_{\pm0.004}$(0.001)	$0.101_{\pm0.004}$(0.001)	$0.939_{\pm0.001}$(0.099)	$0.955_{\pm0.005}$	$0.588_{\pm0.004}$
	0.1	$0.931_{\pm0.000}$(0.069)	$0.904_{\pm0.001}$(0.096)	$0.890_{\pm0.001}$(0.008)	$9.199_{\pm0.002}$(0.801)	$8.604_{\pm0.004}$(1.396)	$0.914_{\pm0.004}$(0.008)	$0.101_{\pm0.004}$(0.001)	$0.105_{\pm0.004}$(0.005)	$0.974_{\pm0.003}$(0.018)	$0.926_{\pm0.004}$	$0.116_{\pm0.005}$
	0.15	$0.879_{\pm0.004}$(0.121)	$0.881_{\pm0.001}$(0.119)	$0.834_{\pm0.001}$(0.022)	$8.730_{\pm0.002}$(1.270)	$8.081_{\pm0.001}$(1.919)	$0.845_{\pm0.005}$(0.037)	$0.101_{\pm0.004}$(0.001)	$0.109_{\pm0.002}$(0.009)	$0.986_{\pm0.004}$(0.016)	$0.921_{\pm0.001}$	$0.017_{\pm0.002}$
	0.2	$0.809_{\pm0.004}$(0.191)	$0.841_{\pm0.004}$(0.159)	$0.816_{\pm0.000}$(0.002)	$8.141_{\pm0.003}$(1.859)	$7.525_{\pm0.003}$(2.475)	$0.824_{\pm0.003}$(0.006)	$0.099_{\pm0.002}$(0.001)	$0.112_{\pm0.003}$(0.012)	$0.990_{\pm0.002}$(0.009)	$0.916_{\pm0.005}$	$0.010_{\pm0.003}$
FF **UA**100%, **UA**$_{tf}$100%, **RA**34.3%, **TA**36.1%	0.05	$0.984_{\pm0.005}$(0.016)	$0.962_{\pm0.004}$(0.038)	$0.950_{\pm0.002}$(0.015)	$7.760_{\pm1.386}$(2.240)	$7.479_{\pm1.223}$(2.521)	$7.627_{\pm0.058}$(6.479)	$0.129_{\pm0.022}$(0.029)	$0.130_{\pm0.021}$(0.030)	$0.125_{\pm0.001}$(0.715)	$0.999_{\pm0.000}$	$0.997_{\pm0.003}$
	0.1	$0.964_{\pm0.013}$(0.036)	$0.917_{\pm0.037}$(0.083)	$0.909_{\pm0.006}$(0.027)	$7.360_{\pm0.999}$(2.640)	$6.988_{\pm0.749}$(3.012)	$6.853_{\pm0.390}$(5.930)	$0.132_{\pm0.020}$(0.032)	$0.132_{\pm0.019}$(0.032)	$0.133_{\pm0.007}$(0.823)	$0.998_{\pm0.000}$	$0.996_{\pm0.003}$
	0.15	$0.922_{\pm0.006}$(0.078)	$0.852_{\pm0.015}$(0.148)	$0.863_{\pm0.017}$(0.007)	$6.941_{\pm1.088}$(3.059)	$6.562_{\pm0.856}$(3.438)	$6.278_{\pm0.232}$(5.395)	$0.134_{\pm0.020}$(0.034)	$0.131_{\pm0.015}$(0.031)	$0.138_{\pm0.008}$(0.832)	$0.997_{\pm0.001}$	$0.994_{\pm0.006}$
	0.2	$0.897_{\pm0.001}$(0.103)	$0.818_{\pm0.020}$(0.182)	$0.789_{\pm0.015}$(0.025)	$6.757_{\pm0.978}$(3.243)	$6.352_{\pm0.757}$(3.648)	$5.307_{\pm0.527}$(4.477)	$0.134_{\pm0.019}$(0.034)	$0.130_{\pm0.012}$(0.030)	$0.150_{\pm0.018}$(0.832)	$0.993_{\pm0.002}$	$0.993_{\pm0.006}$
SSD **UA**1.16%, **UA**$_{tf}$7.75%, **RA**99.5%, **TA**94.3%	0.05	$0.995_{\pm0.014}$(0.005)	$0.935_{\pm0.013}$(0.065)	$0.940_{\pm0.007}$(0.024)	$1.030_{\pm0.014}$(8.970)	$1.067_{\pm0.013}$(8.933)	$0.991_{\pm0.011}$(0.157)	$0.966_{\pm0.010}$(0.866)	$0.876_{\pm0.007}$(0.776)	$0.949_{\pm0.010}$(0.109)	$0.804_{\pm0.015}$	$0.447_{\pm0.007}$
	0.1	$0.984_{\pm0.021}$(0.016)	$0.910_{\pm0.009}$(0.090)	$0.880_{\pm0.001}$(0.002)	$0.992_{\pm0.011}$(9.008)	$0.982_{\pm0.005}$(9.018)	$0.896_{\pm0.003}$(0.026)	$0.992_{\pm0.003}$(0.892)	$0.926_{\pm0.017}$(0.826)	$0.981_{\pm0.012}$(0.025)	$0.434_{\pm0.007}$	$0.022_{\pm0.005}$
	0.15	$0.960_{\pm0.012}$(0.040)	$0.876_{\pm0.011}$(0.124)	$0.847_{\pm0.007}$(0.009)	$0.962_{\pm0.007}$(9.038)	$0.931_{\pm0.006}$(9.069)	$0.857_{\pm0.013}$(0.025)	$0.998_{\pm0.016}$(0.898)	$0.941_{\pm0.002}$(0.841)	$0.989_{\pm0.002}$(0.019)	$0.215_{\pm0.007}$	$0.005_{\pm0.017}$
	0.2	$0.895_{\pm0.020}$(0.105)	$0.816_{\pm0.010}$(0.184)	$0.823_{\pm0.015}$(0.009)	$0.895_{\pm0.014}$(9.105)	$0.850_{\pm0.004}$(9.150)	$0.831_{\pm0.002}$(0.001)	$0.999_{\pm0.001}$(0.899)	$0.960_{\pm0.014}$(0.860)	$0.991_{\pm0.003}$(0.010)	$0.078_{\pm0.003}$	$0.002_{\pm0.009}$
NegGrad+ **UA**96.2%, **UA**$_{tf}$95.2%, **RA**97.6%, **TA**92.8%	0.05	$0.989_{\pm0.016}$(0.011)	$0.961_{\pm0.056}$(0.039)	$0.945_{\pm0.027}$(0.019)	$9.432_{\pm0.803}$(0.568)	$9.038_{\pm1.360}$(0.962)	$1.053_{\pm0.020}$(0.096)	$0.105_{\pm0.007}$(0.005)	$0.107_{\pm0.010}$(0.007)	$0.897_{\pm0.008}$(0.058)	$1.000_{\pm0.000}$	$0.835_{\pm0.085}$
	0.1	$0.980_{\pm0.029}$(0.020)	$0.954_{\pm0.065}$(0.046)	$0.881_{\pm0.028}$(0.001)	$9.250_{\pm1.061}$(0.750)	$8.836_{\pm1.647}$(1.164)	$0.913_{\pm0.018}$(0.009)	$0.106_{\pm0.009}$(0.006)	$0.109_{\pm0.013}$(0.009)	$0.965_{\pm0.012}$(0.009)	$1.000_{\pm0.000}$	$0.057_{\pm0.021}$
	0.15	$0.952_{\pm0.068}$(0.048)	$0.908_{\pm0.130}$(0.092)	$0.849_{\pm0.026}$(0.007)	$8.600_{\pm1.980}$(1.400)	$8.077_{\pm2.719}$(1.923)	$0.868_{\pm0.016}$(0.014)	$0.113_{\pm0.018}$(0.013)	$0.116_{\pm0.023}$(0.016)	$0.977_{\pm0.012}$(0.007)	$1.000_{\pm0.000}$	$0.012_{\pm0.003}$
	0.2	$0.958_{\pm0.060}$(0.042)	$0.921_{\pm0.111}$(0.079)	$0.814_{\pm0.007}$(0.001)	$8.673_{\pm1.876}$(1.327)	$8.219_{\pm2.519}$(1.781)	$0.828_{\pm0.020}$(0.002)	$0.112_{\pm0.017}$(0.012)	$0.115_{\pm0.022}$(0.015)	$0.983_{\pm0.015}$(0.001)	$1.000_{\pm0.000}$	$0.004_{\pm0.003}$
Salun **UA**100%, **UA**$_{tf}$100%, **RA**99.6%, **TA**94.3%	0.05	$0.996_{\pm0.001}$(0.004)	$0.941_{\pm0.008}$(0.059)	$0.952_{\pm0.001}$(0.012)	$9.996_{\pm0.002}$(0.004)	$9.892_{\pm0.003}$(0.108)	$1.028_{\pm0.008}$(0.121)	$0.100_{\pm0.000}$(0.000)	$0.095_{\pm0.001}$(0.005)	$0.926_{\pm0.008}$(0.087)	$1.000_{\pm0.000}$	$0.785_{\pm0.049}$
	0.1	$0.988_{\pm0.004}$(0.012)	$0.906_{\pm0.011}$(0.094)	$0.901_{\pm0.002}$(0.020)	$9.985_{\pm0.003}$(0.015)	$9.817_{\pm0.045}$(0.183)	$0.928_{\pm0.006}$(0.006)	$0.099_{\pm0.000}$(0.001)	$0.092_{\pm0.001}$(0.008)	$0.971_{\pm0.004}$(0.015)	$1.000_{\pm0.000}$	$0.042_{\pm0.011}$
	0.15	$0.960_{\pm0.003}$(0.040)	$0.851_{\pm0.005}$(0.149)	$0.878_{\pm0.006}$(0.022)	$9.952_{\pm0.000}$(0.048)	$9.677_{\pm0.088}$(0.323)	$0.896_{\pm0.005}$(0.013)	$0.096_{\pm0.000}$(0.004)	$0.088_{\pm0.000}$(0.012)	$0.980_{\pm0.001}$(0.010)	$1.000_{\pm0.000}$	$0.009_{\pm0.001}$
	0.2	$0.915_{\pm0.019}$(0.085)	$0.807_{\pm0.038}$(0.193)	$0.820_{\pm0.035}$(0.005)	$9.893_{\pm0.024}$(0.107)	$9.511_{\pm0.192}$(0.489)	$0.828_{\pm0.039}$(0.002)	$0.092_{\pm0.002}$(0.008)	$0.085_{\pm0.002}$(0.015)	$0.990_{\pm0.004}$(0.009)	$1.000_{\pm0.000}$	$0.001_{\pm0.001}$
SFRon **UA**100%, **UA**$_{tf}$100%, **RA**99.3%, **TA**94.4%	0.05	$1.000_{\pm0.000}$(0.000)	$1.000_{\pm0.000}$(0.000)	$0.952_{\pm0.005}$(0.013)	$10.000_{\pm0.000}$(0.000)	$10.000_{\pm0.000}$(0.000)	$1.022_{\pm0.030}$(0.127)	$0.100_{\pm0.000}$(0.000)	$0.100_{\pm0.000}$(0.000)	$0.932_{\pm0.024}$(0.092)	$1.000_{\pm0.000}$	$0.677_{\pm0.206}$
	0.1	$1.000_{\pm0.000}$(0.000)	$1.000_{\pm0.000}$(0.000)	$0.908_{\pm0.013}$(0.026)	$10.000_{\pm0.000}$(0.000)	$10.000_{\pm0.000}$(0.000)	$0.937_{\pm0.028}$(0.014)	$0.100_{\pm0.000}$(0.000)	$0.100_{\pm0.000}$(0.000)	$0.970_{\pm0.015}$(0.014)	$1.000_{\pm0.000}$	$0.089_{\pm0.092}$
	0.15	$1.000_{\pm0.000}$(0.000)	$1.000_{\pm0.000}$(0.000)	$0.840_{\pm0.026}$(0.016)	$10.000_{\pm0.000}$(0.000)	$10.000_{\pm0.000}$(0.000)	$0.849_{\pm0.026}$(0.033)	$0.100_{\pm0.000}$(0.000)	$0.100_{\pm0.000}$(0.000)	$0.989_{\pm0.003}$(0.019)	$1.000_{\pm0.000}$	$0.002_{\pm0.001}$
	0.2	$1.000_{\pm0.000}$(0.000)	$1.000_{\pm0.000}$(0.000)	$0.807_{\pm0.024}$(0.008)	$10.000_{\pm0.000}$(0.000)	$10.000_{\pm0.000}$(0.000)	$0.813_{\pm0.025}$(0.017)	$0.100_{\pm0.000}$(0.000)	$0.100_{\pm0.000}$(0.000)	$0.992_{\pm0.003}$(0.010)	$1.000_{\pm0.000}$	$0.001_{\pm0.001}$

Table 12: Unlearning performance of 9 unlearning methods on **Tiny ImageNet** with **ViT** in 10% **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**14.7%, **RA**98.8%, **TA**86.0%	0.05	$0.944_{\pm0.006}$(0.000)	$0.949_{\pm0.026}$(0.000)	$1.876_{\pm0.009}$(0.000)	$1.840_{\pm0.014}$(0.000)	$0.503_{\pm0.018}$(0.000)	$0.516_{\pm0.018}$(0.000)	$0.984_{\pm0.002}$
	0.1	$0.892_{\pm0.006}$(0.000)	$0.900_{\pm0.025}$(0.000)	$1.151_{\pm0.002}$(0.000)	$1.144_{\pm0.018}$(0.000)	$0.775_{\pm0.016}$(0.000)	$0.786_{\pm0.026}$(0.000)	$0.853_{\pm0.003}$
	0.15	$0.841_{\pm0.024}$(0.000)	$0.850_{\pm0.017}$(0.000)	$0.956_{\pm0.014}$(0.000)	$0.956_{\pm0.017}$(0.000)	$0.880_{\pm0.014}$(0.000)	$0.889_{\pm0.019}$(0.000)	$0.539_{\pm0.001}$
	0.2	$0.790_{\pm0.015}$(0.000)	$0.799_{\pm0.023}$(0.000)	$0.846_{\pm0.004}$(0.000)	$0.854_{\pm0.014}$(0.000)	$0.934_{\pm0.012}$(0.000)	$0.935_{\pm0.015}$(0.000)	$0.238_{\pm0.012}$
FT **UA**6.9%, **RA**97.9%, **TA**84.1%	0.05	$0.994_{\pm0.005}$(0.050)	$0.950_{\pm0.019}$(0.001)	$2.133_{\pm0.008}$(0.257)	$2.440_{\pm0.011}$(0.600)	$0.466_{\pm0.009}$(0.037)	$0.389_{\pm0.016}$(0.127)	$0.994_{\pm0.020}$
	0.1	$0.978_{\pm0.007}$(0.086)	$0.903_{\pm0.003}$(0.003)	$1.234_{\pm0.010}$(0.083)	$1.317_{\pm0.001}$(0.173)	$0.792_{\pm0.018}$(0.017)	$0.685_{\pm0.001}$(0.101)	$0.935_{\pm0.012}$
	0.15	$0.938_{\pm0.001}$(0.097)	$0.851_{\pm0.010}$(0.001)	$1.014_{\pm0.005}$(0.058)	$1.017_{\pm0.016}$(0.061)	$0.925_{\pm0.007}$(0.045)	$0.836_{\pm0.016}$(0.053)	$0.681_{\pm0.003}$
	0.2	$0.888_{\pm0.009}$(0.098)	$0.801_{\pm0.012}$(0.002)	$0.915_{\pm0.006}$(0.069)	$0.885_{\pm0.000}$(0.031)	$0.970_{\pm0.020}$(0.036)	$0.905_{\pm0.005}$(0.030)	$0.326_{\pm0.011}$
RL **UA**26.9%, **RA**96.0%, **TA**81.4%	0.05	$0.969_{\pm0.021}$(0.025)	$0.952_{\pm0.008}$(0.003)	$17.890_{\pm0.003}$(16.014)	$8.572_{\pm0.010}$(6.732)	$0.054_{\pm0.013}$(0.449)	$0.111_{\pm0.002}$(0.405)	$0.996_{\pm0.019}$
	0.1	$0.892_{\pm0.017}$(0.000)	$0.902_{\pm0.013}$(0.002)	$2.639_{\pm0.017}$(1.488)	$1.843_{\pm0.019}$(0.699)	$0.338_{\pm0.022}$(0.437)	$0.489_{\pm0.013}$(0.297)	$0.971_{\pm0.014}$
	0.15	$0.793_{\pm0.021}$(0.048)	$0.855_{\pm0.008}$(0.005)	$1.225_{\pm0.013}$(0.269)	$1.164_{\pm0.000}$(0.208)	$0.648_{\pm0.002}$(0.232)	$0.734_{\pm0.000}$(0.155)	$0.894_{\pm0.022}$
	0.2	$0.681_{\pm0.010}$(0.109)	$0.803_{\pm0.003}$(0.004)	$0.831_{\pm0.006}$(0.015)	$0.946_{\pm0.011}$(0.092)	$0.820_{\pm0.022}$(0.114)	$0.849_{\pm0.006}$(0.086)	$0.715_{\pm0.013}$
GA **UA**3.2%, **RA**97.4%, **TA**84.9%	0.05	$0.996_{\pm0.003}$(0.052)	$0.947_{\pm0.002}$(0.002)	$1.539_{\pm0.004}$(0.337)	$2.018_{\pm0.007}$(0.178)	$0.647_{\pm0.003}$(0.144)	$0.469_{\pm0.002}$(0.047)	$0.988_{\pm0.004}$
	0.1	$0.986_{\pm0.006}$(0.094)	$0.900_{\pm0.000}$(0.000)	$1.104_{\pm0.006}$(0.047)	$1.224_{\pm0.005}$(0.080)	$0.894_{\pm0.003}$(0.119)	$0.736_{\pm0.006}$(0.050)	$0.899_{\pm0.001}$
	0.15	$0.967_{\pm0.002}$(0.126)	$0.852_{\pm0.005}$(0.002)	$1.003_{\pm0.008}$(0.047)	$0.993_{\pm0.004}$(0.037)	$0.964_{\pm0.005}$(0.084)	$0.859_{\pm0.006}$(0.030)	$0.632_{\pm0.009}$
	0.2	$0.934_{\pm0.001}$(0.144)	$0.800_{\pm0.007}$(0.001)	$0.946_{\pm0.008}$(0.100)	$0.871_{\pm0.008}$(0.017)	$0.987_{\pm0.008}$(0.053)	$0.919_{\pm0.005}$(0.016)	$0.296_{\pm0.009}$
Teacher **UA**17.3%, **RA**86.7%, **TA**79.0%	0.05	$0.977_{\pm0.004}$(0.033)	$0.956_{\pm0.003}$(0.007)	$5.473_{\pm0.006}$(3.597)	$5.080_{\pm0.004}$(3.240)	$0.179_{\pm0.008}$(0.324)	$0.188_{\pm0.002}$(0.328)	$0.987_{\pm0.008}$
	0.1	$0.930_{\pm0.003}$(0.038)	$0.902_{\pm0.008}$(0.002)	$1.991_{\pm0.004}$(0.840)	$1.959_{\pm0.002}$(0.815)	$0.467_{\pm0.004}$(0.308)	$0.460_{\pm0.002}$(0.326)	$0.971_{\pm0.007}$
	0.15	$0.873_{\pm0.003}$(0.032)	$0.850_{\pm0.009}$(0.000)	$1.295_{\pm0.006}$(0.339)	$1.319_{\pm0.005}$(0.363)	$0.674_{\pm0.007}$(0.206)	$0.645_{\pm0.003}$(0.244)	$0.944_{\pm0.006}$
	0.2	$0.816_{\pm0.007}$(0.026)	$0.803_{\pm0.009}$(0.004)	$1.020_{\pm0.006}$(0.174)	$1.058_{\pm0.004}$(0.204)	$0.800_{\pm0.005}$(0.134)	$0.758_{\pm0.005}$(0.177)	$0.910_{\pm0.006}$
SSD **UA**1.5%, **RA**98.5%, **TA**86.1%	0.05	$0.998_{\pm0.004}$(0.054)	$0.950_{\pm0.006}$(0.001)	$1.354_{\pm0.008}$(0.522)	$1.827_{\pm0.002}$(0.013)	$0.737_{\pm0.008}$(0.234)	$0.520_{\pm0.008}$(0.004)	$0.985_{\pm0.005}$
	0.1	$0.993_{\pm0.008}$(0.101)	$0.897_{\pm0.008}$(0.003)	$1.039_{\pm0.002}$(0.112)	$1.134_{\pm0.008}$(0.010)	$0.956_{\pm0.007}$(0.181)	$0.791_{\pm0.002}$(0.005)	$0.852_{\pm0.001}$
	0.15	$0.981_{\pm0.005}$(0.140)	$0.853_{\pm0.001}$(0.003)	$0.993_{\pm0.001}$(0.037)	$0.962_{\pm0.005}$(0.006)	$0.988_{\pm0.004}$(0.108)	$0.887_{\pm0.004}$(0.002)	$0.542_{\pm0.007}$
	0.2	$0.956_{\pm0.002}$(0.166)	$0.805_{\pm0.003}$(0.006)	$0.960_{\pm0.003}$(0.114)	$0.864_{\pm0.009}$(0.010)	$0.996_{\pm0.005}$(0.062)	$0.932_{\pm0.002}$(0.003)	$0.249_{\pm0.006}$
NegGrad+ **UA**19.4%, **RA**98.3%, **TA**84.0%	0.05	$0.999_{\pm0.000}$(0.055)	$0.890_{\pm0.002}$(0.059)	$0.949_{\pm0.002}$(0.927)	$1.614_{\pm0.023}$(0.227)	$2.184_{\pm0.052}$(1.681)	$2.499_{\pm0.059}$(1.984)	$0.995_{\pm0.000}$
	0.1	$0.995_{\pm0.001}$(0.103)	$0.848_{\pm0.000}$(0.052)	$0.898_{\pm0.000}$(0.253)	$1.093_{\pm0.005}$(0.051)	$1.225_{\pm0.007}$(0.450)	$1.287_{\pm0.003}$(0.501)	$0.933_{\pm0.002}$
	0.15	$0.987_{\pm0.000}$(0.146)	$0.814_{\pm0.001}$(0.036)	$0.850_{\pm0.001}$(0.106)	$1.009_{\pm0.000}$(0.053)	$1.017_{\pm0.002}$(0.137)	$1.023_{\pm0.003}$(0.133)	$0.685_{\pm0.002}$
	0.2	$0.966_{\pm0.001}$(0.176)	$0.783_{\pm0.003}$(0.016)	$0.802_{\pm0.002}$(0.044)	$0.972_{\pm0.000}$(0.118)	$0.922_{\pm0.004}$(0.012)	$0.891_{\pm0.001}$(0.043)	$0.320_{\pm0.001}$
Salun **UA**9.2%, **RA**97.7%, **TA**83.6%	0.05	$0.995_{\pm0.003}$(0.051)	$0.964_{\pm0.026}$(0.015)	$2.803_{\pm1.607}$(0.927)	$2.726_{\pm0.727}$(0.886)	$1.311_{\pm1.810}$(0.808)	$1.157_{\pm1.481}$(0.641)	$0.988_{\pm0.001}$
	0.1	$0.977_{\pm0.014}$(0.085)	$0.924_{\pm0.040}$(0.024)	$1.229_{\pm0.286}$(0.078)	$1.281_{\pm0.120}$(0.137)	$0.918_{\pm0.387}$(0.143)	$0.884_{\pm0.374}$(0.097)	$0.939_{\pm0.005}$
	0.15	$0.936_{\pm0.041}$(0.095)	$0.874_{\pm0.041}$(0.024)	$0.972_{\pm0.103}$(0.016)	$1.032_{\pm0.005}$(0.076)	$0.935_{\pm0.087}$(0.055)	$0.893_{\pm0.124}$(0.004)	$0.819_{\pm0.003}$
	0.2	$0.870_{\pm0.081}$(0.080)	$0.810_{\pm0.017}$(0.011)	$0.845_{\pm0.036}$(0.001)	$0.925_{\pm0.046}$(0.071)	$0.924_{\pm0.047}$(0.009)	$0.894_{\pm0.006}$(0.041)	$0.630_{\pm0.003}$
SFRon **UA**9.3%, **RA**97.0%, **TA**83.9%	0.05	$0.989_{\pm0.001}$(0.045)	$0.948_{\pm0.001}$(0.001)	$2.000_{\pm0.059}$(0.124)	$2.208_{\pm0.037}$(0.368)	$0.495_{\pm0.014}$(0.008)	$0.429_{\pm0.007}$(0.086)	$0.986_{\pm0.000}$
	0.1	$0.960_{\pm0.003}$(0.068)	$0.899_{\pm0.002}$(0.001)	$1.227_{\pm0.017}$(0.076)	$1.268_{\pm0.007}$(0.123)	$0.783_{\pm0.010}$(0.008)	$0.709_{\pm0.003}$(0.077)	$0.902_{\pm0.003}$
	0.15	$0.917_{\pm0.002}$(0.076)	$0.849_{\pm0.002}$(0.001)	$1.024_{\pm0.006}$(0.068)	$1.015_{\pm0.005}$(0.059)	$0.896_{\pm0.007}$(0.016)	$0.837_{\pm0.004}$(0.053)	$0.689_{\pm0.012}$
	0.2	$0.866_{\pm0.006}$(0.076)	$0.802_{\pm0.003}$(0.003)	$0.916_{\pm0.004}$(0.070)	$0.892_{\pm0.005}$(0.037)	$0.946_{\pm0.002}$(0.012)	$0.899_{\pm0.003}$(0.036)	$0.426_{\pm0.018}$

Table 13: 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	$0.946_{\pm0.001}$(0.000)	$0.948_{\pm0.003}$(0.000)	$2.146_{\pm0.006}$(0.000)	$2.106_{\pm0.002}$(0.000)	$0.441_{\pm0.004}$(0.000)	$0.450_{\pm0.005}$(0.000)	$0.987_{\pm0.004}$
	0.1	$0.892_{\pm0.007}$(0.000)	$0.899_{\pm0.008}$(0.000)	$1.222_{\pm0.002}$(0.000)	$1.211_{\pm0.007}$(0.000)	$0.730_{\pm0.004}$(0.000)	$0.742_{\pm0.002}$(0.000)	$0.889_{\pm0.009}$
	0.15	$0.838_{\pm0.004}$(0.000)	$0.847_{\pm0.001}$(0.000)	$0.977_{\pm0.002}$(0.000)	$0.977_{\pm0.006}$(0.000)	$0.858_{\pm0.008}$(0.000)	$0.868_{\pm0.006}$(0.000)	$0.607_{\pm0.001}$
	0.2	$0.786_{\pm0.005}$(0.000)	$0.796_{\pm0.002}$(0.000)	$0.856_{\pm0.007}$(0.000)	$0.863_{\pm0.001}$(0.000)	$0.918_{\pm0.007}$(0.000)	$0.922_{\pm0.008}$(0.000)	$0.304_{\pm0.008}$
FT **UA**5.4%, **RA**97.1%, **TA**84.4%	0.05	$0.995_{\pm0.013}$(0.051)	$0.949_{\pm0.024}$(0.000)	$1.879_{\pm0.014}$(0.003)	$2.216_{\pm0.003}$(0.376)	$0.527_{\pm0.028}$(0.024)	$0.428_{\pm0.020}$(0.088)	$0.992_{\pm0.019}$
	0.1	$0.979_{\pm0.021}$(0.087)	$0.901_{\pm0.014}$(0.001)	$1.183_{\pm0.018}$(0.032)	$1.281_{\pm0.020}$(0.137)	$0.828_{\pm0.029}$(0.053)	$0.701_{\pm0.010}$(0.085)	$0.926_{\pm0.025}$
	0.15	$0.953_{\pm0.024}$(0.112)	$0.850_{\pm0.022}$(0.000)	$1.014_{\pm0.011}$(0.058)	$1.017_{\pm0.026}$(0.061)	$0.940_{\pm0.027}$(0.060)	$0.839_{\pm0.004}$(0.050)	$0.681_{\pm0.020}$
	0.2	$0.910_{\pm0.029}$(0.120)	$0.806_{\pm0.024}$(0.007)	$0.937_{\pm0.018}$(0.091)	$0.895_{\pm0.001}$(0.041)	$0.977_{\pm0.029}$(0.043)	$0.902_{\pm0.007}$(0.033)	$0.345_{\pm0.016}$
RL **UA**22.5%, **RA**93.5%, **TA**77.1%	0.05	$0.974_{\pm0.011}$(0.028)	$0.953_{\pm0.001}$(0.005)	$26.032_{\pm0.007}$(23.886)	$23.369_{\pm0.008}$(21.263)	$0.038_{\pm0.015}$(0.403)	$0.038_{\pm0.016}$(0.412)	$0.994_{\pm0.010}$
	0.1	$0.930_{\pm0.016}$(0.038)	$0.902_{\pm0.013}$(0.003)	$5.277_{\pm0.001}$(4.055)	$4.621_{\pm0.007}$(3.410)	$0.178_{\pm0.011}$(0.552)	$0.197_{\pm0.001}$(0.545)	$0.987_{\pm0.008}$
	0.15	$0.875_{\pm0.011}$(0.037)	$0.856_{\pm0.008}$(0.009)	$1.758_{\pm0.004}$(0.781)	$1.657_{\pm0.005}$(0.680)	$0.496_{\pm0.006}$(0.362)	$0.516_{\pm0.009}$(0.352)	$0.970_{\pm0.017}$
	0.2	$0.810_{\pm0.006}$(0.024)	$0.805_{\pm0.013}$(0.009)	$1.147_{\pm0.005}$(0.291)	$1.144_{\pm0.005}$(0.281)	$0.707_{\pm0.004}$(0.211)	$0.707_{\pm0.013}$(0.215)	$0.945_{\pm0.005}$
GA **UA**3.9%, **RA**96.1%, **TA**84.2%	0.05	$0.998_{\pm0.007}$(0.052)	$0.949_{\pm0.001}$(0.001)	$1.807_{\pm0.001}$(0.339)	$2.338_{\pm0.001}$(0.232)	$0.552_{\pm0.006}$(0.111)	$0.407_{\pm0.006}$(0.043)	$0.992_{\pm0.006}$
	0.1	$0.986_{\pm0.009}$(0.094)	$0.896_{\pm0.007}$(0.003)	$1.147_{\pm0.003}$(0.075)	$1.278_{\pm0.007}$(0.067)	$0.863_{\pm0.008}$(0.133)	$0.703_{\pm0.002}$(0.039)	$0.918_{\pm0.010}$
	0.15	$0.968_{\pm0.008}$(0.130)	$0.850_{\pm0.002}$(0.003)	$1.015_{\pm0.008}$(0.038)	$1.020_{\pm0.002}$(0.043)	$0.954_{\pm0.009}$(0.096)	$0.835_{\pm0.002}$(0.033)	$0.696_{\pm0.009}$
	0.2	$0.931_{\pm0.011}$(0.145)	$0.804_{\pm0.004}$(0.008)	$0.948_{\pm0.000}$(0.092)	$0.893_{\pm0.003}$(0.030)	$0.983_{\pm0.006}$(0.065)	$0.900_{\pm0.004}$(0.022)	$0.363_{\pm0.002}$
Teacher **UA**22.1%, **RA**85.7%, **TA**76.2%	0.05	$0.967_{\pm0.013}$(0.021)	$0.950_{\pm0.017}$(0.002)	$6.465_{\pm0.007}$(4.319)	$6.233_{\pm0.004}$(4.127)	$0.151_{\pm0.002}$(0.290)	$0.151_{\pm0.006}$(0.299)	$0.990_{\pm0.014}$
	0.1	$0.922_{\pm0.008}$(0.030)	$0.899_{\pm0.002}$(0.000)	$2.202_{\pm0.012}$(0.980)	$2.167_{\pm0.005}$(0.956)	$0.418_{\pm0.009}$(0.312)	$0.419_{\pm0.024}$(0.323)	$0.977_{\pm0.001}$
	0.15	$0.869_{\pm0.025}$(0.031)	$0.852_{\pm0.002}$(0.005)	$1.467_{\pm0.015}$(0.490)	$1.459_{\pm0.004}$(0.482)	$0.591_{\pm0.005}$(0.267)	$0.581_{\pm0.001}$(0.287)	$0.958_{\pm0.021}$
	0.2	$0.814_{\pm0.020}$(0.028)	$0.801_{\pm0.017}$(0.005)	$1.125_{\pm0.005}$(0.269)	$1.138_{\pm0.001}$(0.275)	$0.718_{\pm0.017}$(0.200)	$0.704_{\pm0.009}$(0.218)	$0.927_{\pm0.017}$
SSD **UA**1.3%, **RA**98.4%, **TA**86.1%	0.05	$0.999_{\pm0.001}$(0.053)	$0.952_{\pm0.001}$(0.004)	$1.346_{\pm0.001}$(0.800)	$1.824_{\pm0.000}$(0.282)	$0.742_{\pm0.000}$(0.301)	$0.522_{\pm0.001}$(0.072)	$0.986_{\pm0.001}$
	0.1	$0.995_{\pm0.001}$(0.103)	$0.897_{\pm0.000}$(0.002)	$1.033_{\pm0.001}$(0.189)	$1.135_{\pm0.001}$(0.076)	$0.959_{\pm0.000}$(0.229)	$0.790_{\pm0.000}$(0.048)	$0.847_{\pm0.001}$
	0.15	$0.982_{\pm0.001}$(0.144)	$0.847_{\pm0.000}$(0.000)	$0.987_{\pm0.000}$(0.010)	$0.956_{\pm0.000}$(0.021)	$0.989_{\pm0.001}$(0.131)	$0.890_{\pm0.001}$(0.022)	$0.517_{\pm0.001}$
	0.2	$0.959_{\pm0.001}$(0.173)	$0.804_{\pm0.001}$(0.008)	$0.961_{\pm0.000}$(0.105)	$0.862_{\pm0.000}$(0.001)	$0.995_{\pm0.001}$(0.077)	$0.932_{\pm0.001}$(0.010)	$0.243_{\pm0.001}$
NegGrad+ **UA**11.5%, **RA**98.7%, **TA**83.8%	0.05	$0.999_{\pm0.000}$(0.053)	$0.979_{\pm0.001}$(0.031)	$0.946_{\pm0.002}$(1.200)	$1.443_{\pm0.028}$(0.663)	$2.248_{\pm0.063}$(1.807)	$2.358_{\pm0.095}$(1.908)	$0.992_{\pm0.001}$
	0.1	$0.996_{\pm0.000}$(0.104)	$0.946_{\pm0.002}$(0.047)	$0.900_{\pm0.003}$(0.322)	$1.078_{\pm0.006}$(0.134)	$1.295_{\pm0.010}$(0.565)	$1.332_{\pm0.008}$(0.590)	$0.933_{\pm0.003}$
	0.15	$0.990_{\pm0.000}$(0.152)	$0.900_{\pm0.003}$(0.052)	$0.853_{\pm0.004}$(0.124)	$1.008_{\pm0.002}$(0.031)	$1.032_{\pm0.010}$(0.174)	$1.033_{\pm0.011}$(0.165)	$0.712_{\pm0.015}$
	0.2	$0.977_{\pm0.000}$(0.191)	$0.848_{\pm0.003}$(0.052)	$0.805_{\pm0.002}$(0.052)	$0.982_{\pm0.000}$(0.119)	$0.909_{\pm0.004}$(0.009)	$0.898_{\pm0.007}$(0.024)	$0.381_{\pm0.009}$
Salun **UA**9.2%, **RA**95.7%, **TA**81.9%	0.05	$0.993_{\pm0.003}$(0.047)	$0.962_{\pm0.026}$(0.014)	$3.284_{\pm2.048}$(1.138)	$4.112_{\pm0.813}$(2.007)	$1.546_{\pm2.290}$(1.105)	$1.558_{\pm2.336}$(1.108)	$0.989_{\pm0.001}$
	0.1	$0.976_{\pm0.011}$(0.084)	$0.924_{\pm0.039}$(0.026)	$1.386_{\pm0.423}$(0.164)	$1.579_{\pm0.130}$(0.368)	$0.922_{\pm0.566}$(0.192)	$0.896_{\pm0.607}$(0.154)	$0.973_{\pm0.002}$
	0.15	$0.944_{\pm0.024}$(0.106)	$0.876_{\pm0.046}$(0.029)	$1.051_{\pm0.175}$(0.074)	$1.139_{\pm0.017}$(0.162)	$0.919_{\pm0.194}$(0.061)	$0.871_{\pm0.226}$(0.003)	$0.942_{\pm0.002}$
	0.2	$0.900_{\pm0.044}$(0.114)	$0.825_{\pm0.049}$(0.029)	$0.910_{\pm0.097}$(0.054)	$0.969_{\pm0.037}$(0.105)	$0.928_{\pm0.040}$(0.011)	$0.876_{\pm0.063}$(0.045)	$0.893_{\pm0.002}$
SFRon **UA**6.3%, **RA**96.8%, **TA**82.9%	0.05	$0.994_{\pm0.001}$(0.048)	$0.947_{\pm0.003}$(0.001)	$2.010_{\pm0.188}$(0.136)	$2.327_{\pm0.087}$(0.222)	$0.497_{\pm0.045}$(0.057)	$0.407_{\pm0.016}$(0.043)	$0.983_{\pm0.002}$
	0.1	$0.980_{\pm0.006}$(0.087)	$0.900_{\pm0.003}$(0.001)	$1.245_{\pm0.060}$(0.023)	$1.338_{\pm0.039}$(0.126)	$0.788_{\pm0.041}$(0.058)	$0.673_{\pm0.020}$(0.069)	$0.909_{\pm0.003}$
	0.15	$0.951_{\pm0.011}$(0.113)	$0.849_{\pm0.003}$(0.001)	$1.041_{\pm0.020}$(0.065)	$1.044_{\pm0.023}$(0.067)	$0.913_{\pm0.028}$(0.055)	$0.813_{\pm0.016}$(0.055)	$0.738_{\pm0.029}$
	0.2	$0.910_{\pm0.011}$(0.125)	$0.803_{\pm0.003}$(0.008)	$0.947_{\pm0.006}$(0.091)	$0.910_{\pm0.022}$(0.046)	$0.961_{\pm0.017}$(0.044)	$0.884_{\pm0.017}$(0.038)	$0.523_{\pm0.068}$

Table 14: Unlearning performance of 9 unlearning methods on **Tiny ImageNet** with **ViT** in **class-wise forgetting** scenario.

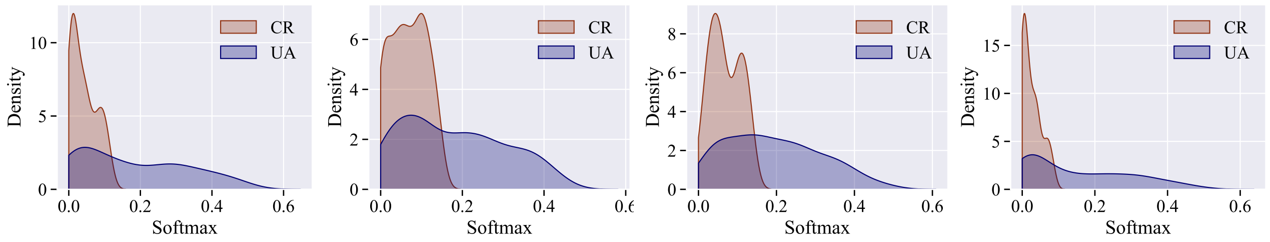
Methods	α	Coverage			Set Size			CR				
		$\mathcal{D}_f \downarrow$	$\mathcal{D}_{tf} \downarrow$	$\mathcal{D}_{tr} \uparrow$	$\mathcal{D}_f \uparrow$	$\mathcal{D}_{tf} \uparrow$	$\mathcal{D}_{tr} \downarrow$	$\mathcal{D}_f \downarrow$	$\mathcal{D}_{tf} \downarrow$	$\mathcal{D}_{tr} \uparrow$	$\hat{q}_f$	$\hat{q}_{test}$
RT **UA**100%, **UA**$_{tf}$100%, **RA**98.7%, **TA**86.4%	0.05	$1.000_{\pm0.000}$(0.000)	$1.000_{\pm0.000}$(0.000)	$0.950_{\pm0.003}$(0.000)	$200.000_{\pm0.000}$(0.000)	$200.000_{\pm0.000}$(0.000)	$1.785_{\pm0.056}$(0.000)	$0.005_{\pm0.000}$(0.000)	$0.005_{\pm0.000}$(0.000)	$0.532_{\pm0.009}$(0.000)	$1.000_{\pm0.000}$	$0.984_{\pm0.002}$
	0.1	$0.936_{\pm0.011}$(0.000)	$0.960_{\pm0.016}$(0.000)	$0.903_{\pm0.009}$(0.000)	$192.882_{\pm0.912}$(0.000)	$193.340_{\pm2.620}$(0.000)	$1.146_{\pm0.002}$(0.000)	$0.005_{\pm0.000}$(0.000)	$0.005_{\pm0.000}$(0.000)	$0.788_{\pm0.008}$(0.000)	$1.000_{\pm0.000}$	$0.859_{\pm0.004}$
	0.15	$0.904_{\pm0.039}$(0.000)	$0.960_{\pm0.046}$(0.000)	$0.853_{\pm0.005}$(0.000)	$186.791_{\pm2.173}$(0.000)	$188.880_{\pm1.802}$(0.000)	$0.957_{\pm0.010}$(0.000)	$0.005_{\pm0.000}$(0.000)	$0.005_{\pm0.000}$(0.000)	$0.892_{\pm0.003}$(0.000)	$1.000_{\pm0.000}$	$0.535_{\pm0.002}$
	0.2	$0.787_{\pm0.061}$(0.000)	$0.860_{\pm0.024}$(0.000)	$0.805_{\pm0.003}$(0.000)	$171.051_{\pm3.183}$(0.000)	$174.480_{\pm2.311}$(0.000)	$0.860_{\pm0.010}$(0.000)	$0.005_{\pm0.000}$(0.000)	$0.005_{\pm0.000}$(0.000)	$0.936_{\pm0.002}$(0.000)	$1.000_{\pm0.000}$	$0.232_{\pm0.001}$
FT **UA**13.8%, **UA**$_{tf}$22.0%, **RA**97.5%, **TA**84.1%	0.05	$0.993_{\pm0.006}$(0.007)	$0.960_{\pm0.009}$(0.040)	$0.952_{\pm0.006}$(0.002)	$8.360_{\pm0.007}$(191.640)	$8.280_{\pm0.006}$(191.720)	$2.442_{\pm0.011}$(0.657)	$0.119_{\pm0.018}$(0.114)	$0.116_{\pm0.001}$(0.111)	$0.390_{\pm0.023}$(0.142)	$0.999_{\pm0.006}$	$0.993_{\pm0.005}$
	0.1	$0.984_{\pm0.009}$(0.048)	$0.860_{\pm0.013}$(0.100)	$0.898_{\pm0.005}$(0.005)	$1.802_{\pm0.009}$(191.080)	$1.660_{\pm0.018}$(191.680)	$1.287_{\pm0.009}$(0.141)	$0.546_{\pm0.008}$(0.541)	$0.518_{\pm0.004}$(0.513)	$0.698_{\pm0.019}$(0.090)	$0.971_{\pm0.019}$	$0.924_{\pm0.016}$
	0.15	$0.902_{\pm0.019}$(0.002)	$0.800_{\pm0.004}$(0.160)	$0.852_{\pm0.017}$(0.001)	$1.120_{\pm0.021}$(185.671)	$1.040_{\pm0.006}$(187.840)	$1.021_{\pm0.017}$(0.064)	$0.806_{\pm0.012}$(0.801)	$0.769_{\pm0.013}$(0.764)	$0.835_{\pm0.022}$(0.057)	$0.809_{\pm0.010}$	$0.686_{\pm0.004}$
	0.2	$0.860_{\pm0.021}$(0.073)	$0.760_{\pm0.003}$(0.100)	$0.800_{\pm0.018}$(0.005)	$0.969_{\pm0.002}$(170.082)	$0.960_{\pm0.003}$(173.520)	$0.882_{\pm0.010}$(0.022)	$0.888_{\pm0.005}$(0.883)	$0.792_{\pm0.002}$(0.787)	$0.907_{\pm0.006}$(0.029)	$0.595_{\pm0.002}$	$0.338_{\pm0.019}$
RL **UA**100%, **UA**$_{tf}$100%, **RA**98.2%, **TA**84.6%	0.05	$0.998_{\pm0.005}$(0.002)	$0.980_{\pm0.003}$(0.020)	$0.952_{\pm0.049}$(0.002)	$199.489_{\pm0.512}$(0.511)	$195.220_{\pm1.003}$(4.780)	$2.317_{\pm0.009}$(0.532)	$0.005_{\pm0.000}$(0.000)	$0.005_{\pm0.000}$(0.000)	$0.411_{\pm0.000}$(0.121)	$1.000_{\pm0.000}$	$0.995_{\pm0.032}$
	0.1	$0.971_{\pm0.013}$(0.035)	$0.900_{\pm0.017}$(0.060)	$0.900_{\pm0.002}$(0.003)	$180.442_{\pm0.710}$(12.440)	$170.960_{\pm0.948}$(22.380)	$1.237_{\pm0.050}$(0.991)	$0.005_{\pm0.000}$(0.000)	$0.005_{\pm0.000}$(0.000)	$0.727_{\pm0.016}$(0.061)	$1.000_{\pm0.000}$	$0.925_{\pm0.024}$
	0.15	$0.922_{\pm0.011}$(0.018)	$0.900_{\pm0.011}$(0.060)	$0.852_{\pm0.015}$(0.001)	$165.884_{\pm2.037}$(20.907)	$159.980_{\pm1.012}$(28.900)	$1.001_{\pm0.003}$(0.044)	$0.006_{\pm0.001}$(0.001)	$0.006_{\pm0.000}$(0.001)	$0.851_{\pm0.023}$(0.041)	$1.000_{\pm0.000}$	$0.641_{\pm0.035}$
	0.2	$0.882_{\pm0.007}$(0.095)	$0.860_{\pm0.007}$(0.000)	$0.807_{\pm0.007}$(0.002)	$154.896_{\pm2.028}$(16.155)	$149.280_{\pm3.013}$(25.200)	$0.886_{\pm0.032}$(0.026)	$0.006_{\pm0.000}$(0.001)	$0.006_{\pm0.001}$(0.001)	$0.912_{\pm0.013}$(0.024)	$1.000_{\pm0.000}$	$0.262_{\pm0.022}$
GA **UA**9.1%, **UA**$_{tf}$20.0%, **RA**98.6%, **TA**86.1%	0.05	$1.000_{\pm0.001}$(0.000)	$0.980_{\pm0.002}$(0.020)	$0.948_{\pm0.026}$(0.002)	$22.836_{\pm0.045}$(177.164)	$20.600_{\pm0.011}$(179.400)	$1.781_{\pm0.017}$(0.004)	$0.044_{\pm0.017}$(0.019)	$0.048_{\pm0.028}$(0.043)	$0.532_{\pm0.013}$(0.000)	$1.000_{\pm0.000}$	$0.984_{\pm0.033}$
	0.1	$0.991_{\pm0.022}$(0.055)	$0.900_{\pm0.014}$(0.060)	$0.897_{\pm0.016}$(0.006)	$1.631_{\pm0.031}$(191.251)	$1.720_{\pm0.005}$(191.620)	$1.133_{\pm0.044}$(0.013)	$0.608_{\pm0.006}$(0.603)	$0.523_{\pm0.007}$(0.518)	$0.792_{\pm0.037}$(0.004)	$0.972_{\pm0.033}$	$0.849_{\pm0.039}$
	0.15	$0.958_{\pm0.002}$(0.054)	$0.820_{\pm0.010}$(0.140)	$0.850_{\pm0.006}$(0.003)	$1.151_{\pm0.039}$(185.640)	$1.140_{\pm0.042}$(187.740)	$0.958_{\pm0.026}$(0.001)	$0.832_{\pm0.003}$(0.827)	$0.719_{\pm0.021}$(0.714)	$0.887_{\pm0.044}$(0.005)	$0.868_{\pm0.023}$	$0.535_{\pm0.011}$
	0.2	$0.880_{\pm0.047}$(0.093)	$0.800_{\pm0.051}$(0.060)	$0.803_{\pm0.025}$(0.002)	$0.929_{\pm0.002}$(170.122)	$0.900_{\pm0.009}$(173.580)	$0.861_{\pm0.006}$(0.001)	$0.947_{\pm0.036}$(0.942)	$0.889_{\pm0.029}$(0.884)	$0.933_{\pm0.027}$(0.003)	$0.473_{\pm0.016}$	$0.238_{\pm0.000}$
Teacher **UA**100%, **UA**$_{tf}$100%, **RA**88.8%, **TA**78.6%	0.05	$0.982_{\pm0.014}$(0.018)	$1.000_{\pm0.007}$(0.000)	$0.952_{\pm0.025}$(0.002)	$199.971_{\pm0.009}$(0.029)	$200.000_{\pm0.000}$(0.000)	$5.095_{\pm0.020}$(3.310)	$0.005_{\pm0.000}$(0.000)	$0.005_{\pm0.000}$(0.000)	$0.187_{\pm0.008}$(0.345)	$1.000_{\pm0.000}$	$0.989_{\pm0.001}$
	0.1	$0.909_{\pm0.013}$(0.027)	$0.940_{\pm0.015}$(0.020)	$0.903_{\pm0.032}$(0.000)	$199.813_{\pm0.009}$(6.931)	$199.900_{\pm0.013}$(6.560)	$2.033_{\pm0.031}$(0.887)	$0.005_{\pm0.000}$(0.000)	$0.005_{\pm0.000}$(0.000)	$0.444_{\pm0.006}$(0.344)	$1.000_{\pm0.000}$	$0.965_{\pm0.003}$
	0.15	$0.887_{\pm0.014}$(0.017)	$0.880_{\pm0.011}$(0.080)	$0.854_{\pm0.003}$(0.001)	$199.667_{\pm0.030}$(12.876)	$199.760_{\pm0.026}$(10.880)	$1.331_{\pm0.012}$(0.374)	$0.004_{\pm0.000}$(0.001)	$0.004_{\pm0.001}$(0.001)	$0.641_{\pm0.010}$(0.251)	$1.000_{\pm0.000}$	$0.919_{\pm0.001}$
	0.2	$0.838_{\pm0.022}$(0.051)	$0.840_{\pm0.002}$(0.020)	$0.799_{\pm0.017}$(0.006)	$199.413_{\pm0.024}$(28.362)	$199.620_{\pm0.030}$(25.140)	$1.022_{\pm0.017}$(0.162)	$0.004_{\pm0.001}$(0.001)	$0.004_{\pm0.001}$(0.001)	$0.781_{\pm0.019}$(0.155)	$1.000_{\pm0.000}$	$0.825_{\pm0.002}$
SSD **UA**100%, **UA**$_{tf}$100%, **RA**98.4%, **TA**86.1%	0.05	$1.000_{\pm0.000}$(0.000)	$1.000_{\pm0.000}$(0.000)	$0.950_{\pm0.017}$(0.000)	$198.769_{\pm0.052}$(1.231)	$197.320_{\pm1.010}$(2.680)	$1.866_{\pm0.019}$(0.081)	$0.005_{\pm0.000}$(0.000)	$0.005_{\pm0.000}$(0.000)	$0.509_{\pm0.013}$(0.023)	$1.000_{\pm0.000}$	$0.986_{\pm0.006}$
	0.1	$0.949_{\pm0.017}$(0.013)	$0.900_{\pm0.012}$(0.060)	$0.897_{\pm0.007}$(0.006)	$171.073_{\pm0.209}$(21.809)	$169.360_{\pm2.002}$(23.980)	$1.141_{\pm0.014}$(0.005)	$0.006_{\pm0.000}$(0.001)	$0.005_{\pm0.000}$(0.000)	$0.786_{\pm0.021}$(0.002)	$1.000_{\pm0.000}$	$0.854_{\pm0.006}$
	0.15	$0.913_{\pm0.007}$(0.009)	$0.880_{\pm0.020}$(0.080)	$0.852_{\pm0.015}$(0.001)	$157.140_{\pm1.209}$(29.651)	$154.960_{\pm0.907}$(33.920)	$0.959_{\pm0.011}$(0.002)	$0.006_{\pm0.001}$(0.001)	$0.006_{\pm0.000}$(0.001)	$0.888_{\pm0.012}$(0.004)	$1.000_{\pm0.000}$	$0.538_{\pm0.007}$
	0.2	$0.833_{\pm0.007}$(0.046)	$0.800_{\pm0.013}$(0.060)	$0.806_{\pm0.022}$(0.001)	$136.502_{\pm3.022}$(34.549)	$136.420_{\pm2.422}$(38.060)	$0.864_{\pm0.002}$(0.004)	$0.006_{\pm0.000}$(0.001)	$0.006_{\pm0.000}$(0.001)	$0.932_{\pm0.015}$(0.004)	$1.000_{\pm0.000}$	$0.254_{\pm0.005}$
NegGrad+ **UA**100%, **UA**$_{tf}$100%, **RA**99.0%, **TA**85.8%	0.05	$1.000_{\pm0.000}$(0.000)	$1.000_{\pm0.000}$(0.000)	$0.947_{\pm0.002}$(0.003)	$200.000_{\pm0.000}$(0.000)	$200.000_{\pm0.000}$(0.000)	$1.850_{\pm0.036}$(0.065)	$0.005_{\pm0.000}$(0.000)	$0.005_{\pm0.000}$(0.000)	$0.512_{\pm0.009}$(0.020)	$1.000_{\pm0.000}$	$0.987_{\pm0.001}$
	0.1	$0.927_{\pm0.104}$(0.009)	$0.950_{\pm0.071}$(0.010)	$0.894_{\pm0.001}$(0.009)	$193.994_{\pm8.493}$(1.112)	$197.490_{\pm3.550}$(4.150)	$1.140_{\pm0.007}$(0.006)	$0.005_{\pm0.000}$(0.000)	$0.005_{\pm0.000}$(0.000)	$0.784_{\pm0.004}$(0.004)	$1.000_{\pm0.000}$	$0.859_{\pm0.003}$
	0.15	$0.862_{\pm0.013}$(0.042)	$0.870_{\pm0.042}$(0.090)	$0.849_{\pm0.000}$(0.004)	$188.686_{\pm0.954}$(1.894)	$195.590_{\pm0.863}$(6.710)	$0.961_{\pm0.001}$(0.004)	$0.005_{\pm0.000}$(0.000)	$0.004_{\pm0.000}$(0.001)	$0.884_{\pm0.000}$(0.008)	$1.000_{\pm0.000}$	$0.537_{\pm0.003}$
	0.2	$0.830_{\pm0.027}$(0.043)	$0.840_{\pm0.085}$(0.020)	$0.802_{\pm0.002}$(0.003)	$187.219_{\pm0.064}$(16.168)	$194.310_{\pm0.948}$(19.830)	$0.861_{\pm0.001}$(0.002)	$0.004_{\pm0.000}$(0.000)	$0.004_{\pm0.000}$(0.001)	$0.931_{\pm0.001}$(0.005)	$1.000_{\pm0.000}$	$0.220_{\pm0.002}$
Salun **UA**100%, **UA**$_{tf}$100%, **RA**98.4%, **TA**86.1%	0.05	$0.997_{\pm0.003}$(0.003)	$0.993_{\pm0.012}$(0.007)	$0.949_{\pm0.001}$(0.001)	$199.599_{\pm0.207}$(0.401)	$197.440_{\pm1.244}$(2.560)	$1.980_{\pm0.050}$(0.196)	$0.005_{\pm0.000}$(0.000)	$0.005_{\pm0.000}$(0.000)	$0.479_{\pm0.012}$(0.053)	$1.000_{\pm0.000}$	$0.989_{\pm0.001}$
	0.1	$0.975_{\pm0.019}$(0.039)	$0.927_{\pm0.023}$(0.033)	$0.899_{\pm0.001}$(0.003)	$191.973_{\pm1.616}$(0.910)	$185.220_{\pm0.918}$(8.120)	$1.169_{\pm0.002}$(0.023)	$0.005_{\pm0.000}$(0.000)	$0.005_{\pm0.000}$(0.000)	$0.769_{\pm0.001}$(0.019)	$1.000_{\pm0.000}$	$0.884_{\pm0.001}$
	0.15	$0.961_{\pm0.022}$(0.057)	$0.860_{\pm0.040}$(0.100)	$0.850_{\pm0.001}$(0.004)	$187.825_{\pm3.461}$(1.034)	$180.307_{\pm2.908}$(8.573)	$0.969_{\pm0.002}$(0.012)	$0.005_{\pm0.000}$(0.000)	$0.005_{\pm0.000}$(0.000)	$0.877_{\pm0.002}$(0.015)	$1.000_{\pm0.000}$	$0.562_{\pm0.003}$
	0.2	$0.960_{\pm0.015}$(0.173)	$0.840_{\pm0.020}$(0.020)	$0.801_{\pm0.001}$(0.004)	$184.838_{\pm3.478}$(13.787)	$177.647_{\pm2.627}$(3.167)	$0.863_{\pm0.004}$(0.003)	$0.005_{\pm0.000}$(0.001)	$0.005_{\pm0.000}$(0.000)	$0.928_{\pm0.003}$(0.008)	$1.000_{\pm0.000}$	$0.230_{\pm0.009}$
SFRon **UA**100%, **UA**$_{tf}$100%, **RA**96.1%, **TA**84.3%	0.05	$1.000_{\pm0.000}$(0.000)	$1.000_{\pm0.000}$(0.000)	$0.948_{\pm0.001}$(0.002)	$200.000_{\pm0.000}$(0.000)	$200.000_{\pm0.000}$(0.000)	$2.264_{\pm0.254}$(0.479)	$0.005_{\pm0.000}$(0.000)	$0.005_{\pm0.000}$(0.000)	$0.423_{\pm0.050}$(0.110)	$1.000_{\pm0.000}$	$0.990_{\pm0.003}$
	0.1	$1.000_{\pm0.000}$(0.064)	$1.000_{\pm0.000}$(0.040)	$0.900_{\pm0.001}$(0.003)	$200.000_{\pm0.000}$(7.118)	$200.000_{\pm0.000}$(6.660)	$1.266_{\pm0.044}$(0.120)	$0.005_{\pm0.000}$(0.000)	$0.005_{\pm0.000}$(0.000)	$0.711_{\pm0.026}$(0.077)	$1.000_{\pm0.000}$	$0.912_{\pm0.017}$
	0.15	$1.000_{\pm0.000}$(0.096)	$1.000_{\pm0.000}$(0.040)	$0.850_{\pm0.002}$(0.003)	$200.000_{\pm0.000}$(13.209)	$200.000_{\pm0.000}$(11.120)	$1.009_{\pm0.012}$(0.051)	$0.005_{\pm0.000}$(0.000)	$0.005_{\pm0.000}$(0.000)	$0.843_{\pm0.011}$(0.049)	$1.000_{\pm0.000}$	$0.668_{\pm0.029}$
	0.2	$1.000_{\pm0.000}$(0.213)	$1.000_{\pm0.000}$(0.140)	$0.802_{\pm0.003}$(0.003)	$200.000_{\pm0.000}$(28.949)	$200.000_{\pm0.000}$(25.520)	$0.886_{\pm0.006}$(0.026)	$0.005_{\pm0.000}$(0.000)	$0.005_{\pm0.000}$(0.000)	$0.905_{\pm0.007}$(0.031)	$1.000_{\pm0.000}$	$0.358_{\pm0.017}$

Table 15: **MIACR** performance on **CIFAR-10** with **ResNet-18**.

Methods	α	**10% Forgetting**		**50% Forgetting**	
		MIACR ↑	$\hat{q}$	**MIACR** ↑	$\hat{q}$
	0.05	$0.091_{\pm0.001}$(0.000)	$0.877_{\pm0.004}$	$0.117_{\pm0.010}$(0.000)	$0.899_{\pm0.007}$
RT	0.1	$0.147_{\pm0.000}$(0.000)	$0.589_{\pm0.008}$	$0.201_{\pm0.011}$(0.000)	$0.570_{\pm0.001}$
MIA86.92% (**10% Forgetting**)	0.15	$0.203_{\pm0.010}$(0.000)	$0.485_{\pm0.005}$	$0.272_{\pm0.011}$(0.000)	$0.472_{\pm0.009}$
MIA82.79% (**50% Forgetting**)	0.2	$0.246_{\pm0.000}$(0.000)	$0.473_{\pm0.001}$	$0.318_{\pm0.006}$(0.000)	$0.459_{\pm0.003}$
	0.05	$0.039_{\pm0.011}$(0.052)	$0.745_{\pm0.013}$	$0.036_{\pm0.001}$(0.081)	$0.780_{\pm0.011}$
FT	0.1	$0.077_{\pm0.008}$(0.070)	$0.627_{\pm0.000}$	$0.103_{\pm0.011}$(0.098)	$0.558_{\pm0.012}$
MIA92.00% (**10% Forgetting**)	0.15	$0.128_{\pm0.007}$(0.075)	$0.517_{\pm0.008}$	$0.159_{\pm0.011}$(0.113)	$0.494_{\pm0.011}$
MIA92.92% (**50% Forgetting**)	0.2	$0.196_{\pm0.003}$(0.050)	$0.483_{\pm0.003}$	$0.244_{\pm0.010}$(0.074)	$0.476_{\pm0.004}$
	0.05	$0.083_{\pm0.010}$(0.008)	$0.627_{\pm0.011}$	$0.050_{\pm0.016}$(0.067)	$0.547_{\pm0.000}$
RL	0.1	$0.178_{\pm0.027}$(0.031)	$0.572_{\pm0.005}$	$0.137_{\pm0.030}$(0.064)	$0.547_{\pm0.001}$
MIA74.21% (**10% Forgetting**)	0.15	$0.272_{\pm0.006}$(0.069)	$0.492_{\pm0.015}$	$0.194_{\pm0.031}$(0.078)	$0.547_{\pm0.001}$
MIA61.15% (**50% Forgetting**)	0.2	$0.320_{\pm0.025}$(0.074)	$0.485_{\pm0.011}$	$0.261_{\pm0.001}$(0.057)	$0.546_{\pm0.000}$
	0.05	$0.012_{\pm0.002}$(0.079)	$0.862_{\pm0.016}$	$0.012_{\pm0.019}$(0.105)	$0.771_{\pm0.008}$
GA	0.1	$0.032_{\pm0.003}$(0.115)	$0.502_{\pm0.016}$	$0.055_{\pm0.003}$(0.146)	$0.486_{\pm0.005}$
MIA98.80% (**10% Forgetting**)	0.15	$0.076_{\pm0.000}$(0.127)	$0.477_{\pm0.007}$	$0.107_{\pm0.016}$(0.165)	$0.474_{\pm0.015}$
MIA98.86% (**50% Forgetting**)	0.2	$0.146_{\pm0.016}$(0.100)	$0.476_{\pm0.019}$	$0.164_{\pm0.016}$(0.154)	$0.473_{\pm0.011}$
	0.05	$0.013_{\pm0.006}$(0.078)	$0.750_{\pm0.014}$	$0.031_{\pm0.003}$(0.086)	$0.635_{\pm0.018}$
Teacher	0.1	$0.038_{\pm0.023}$(0.109)	$0.672_{\pm0.028}$	$0.065_{\pm0.021}$(0.136)	$0.582_{\pm0.013}$
MIA87.24% (**10% Forgetting**)	0.15	$0.072_{\pm0.013}$(0.131)	$0.625_{\pm0.029}$	$0.110_{\pm0.017}$(0.162)	$0.548_{\pm0.007}$
MIA93.24% (**50% Forgetting**)	0.2	$0.113_{\pm0.008}$(0.133)	$0.588_{\pm0.019}$	$0.159_{\pm0.017}$(0.159)	$0.532_{\pm0.006}$
	0.05	$0.038_{\pm0.009}$(0.053)	$0.500_{\pm0.003}$	$0.043_{\pm0.003}$(0.074)	$0.508_{\pm0.003}$
FF	0.1	$0.051_{\pm0.017}$(0.096)	$0.486_{\pm0.018}$	$0.089_{\pm0.001}$(0.112)	$0.509_{\pm0.013}$
MIA71.52% (**10% Forgetting**)	0.15	$0.080_{\pm0.015}$(0.123)	$0.474_{\pm0.013}$	$0.130_{\pm0.017}$(0.142)	$0.506_{\pm0.007}$
MIA76.02% (**50% Forgetting**)	0.2	$0.109_{\pm0.004}$(0.137)	$0.473_{\pm0.002}$	$0.168_{\pm0.010}$(0.150)	$0.499_{\pm0.006}$
	0.05	$0.011_{\pm0.011}$(0.080)	$0.861_{\pm0.012}$	$0.012_{\pm0.002}$(0.105)	$0.748_{\pm0.011}$
SSD	0.1	$0.031_{\pm0.010}$(0.116)	$0.511_{\pm0.011}$	$0.051_{\pm0.005}$(0.150)	$0.488_{\pm0.001}$
MIA98.78% (**10% Forgetting**)	0.15	$0.077_{\pm0.005}$(0.126)	$0.480_{\pm0.013}$	$0.104_{\pm0.006}$(0.168)	$0.477_{\pm0.015}$
MIA98.87% (**50% Forgetting**)	0.2	$0.139_{\pm0.011}$(0.107)	$0.475_{\pm0.013}$	$0.168_{\pm0.012}$(0.150)	$0.477_{\pm0.006}$
	0.05	$0.076_{\pm0.025}$(0.015)	$0.844_{\pm0.024}$	$0.045_{\pm0.008}$(0.072)	$0.863_{\pm0.025}$
NegGrad+	0.1	$0.128_{\pm0.018}$(0.019)	$0.481_{\pm0.009}$	$0.109_{\pm0.007}$(0.092)	$0.511_{\pm0.008}$
MIA90.30% (**10% Forgetting**)	0.15	$0.174_{\pm0.022}$(0.029)	$0.480_{\pm0.005}$	$0.167_{\pm0.017}$(0.105)	$0.477_{\pm0.010}$
MIA93.82% (**50% Forgetting**)	0.2	$0.213_{\pm0.012}$(0.033)	$0.480_{\pm0.004}$	$0.230_{\pm0.014}$(0.088)	$0.472_{\pm0.008}$
	0.05	$0.055_{\pm0.014}$(0.036)	$0.691_{\pm0.011}$	$0.044_{\pm0.001}$(0.073)	$0.670_{\pm0.008}$
Salun	0.1	$0.113_{\pm0.009}$(0.034)	$0.681_{\pm0.013}$	$0.115_{\pm0.009}$(0.086)	$0.630_{\pm0.009}$
MIA57.58% (**10% Forgetting**)	0.15	$0.198_{\pm0.006}$(0.005)	$0.642_{\pm0.015}$	$0.170_{\pm0.009}$(0.102)	$0.610_{\pm0.003}$
MIA59.12% (**50% Forgetting**)	0.2	$0.267_{\pm0.009}$(0.021)	$0.608_{\pm0.011}$	$0.220_{\pm0.005}$(0.098)	$0.586_{\pm0.005}$
	0.05	$0.017_{\pm0.001}$(0.074)	$0.711_{\pm0.009}$	$0.017_{\pm0.002}$(0.100)	$0.715_{\pm0.008}$
SFRon	0.1	$0.040_{\pm0.004}$(0.107)	$0.626_{\pm0.025}$	$0.046_{\pm0.002}$(0.155)	$0.562_{\pm0.013}$
MIA91.55% (**10% Forgetting**)	0.15	$0.113_{\pm0.003}$(0.090)	$0.517_{\pm0.003}$	$0.134_{\pm0.013}$(0.138)	$0.498_{\pm0.003}$
MIA92.52% (**50% Forgetting**)	0.2	$0.184_{\pm0.002}$(0.062)	$0.487_{\pm0.002}$	$0.206_{\pm0.014}$(0.112)	$0.483_{\pm0.002}$

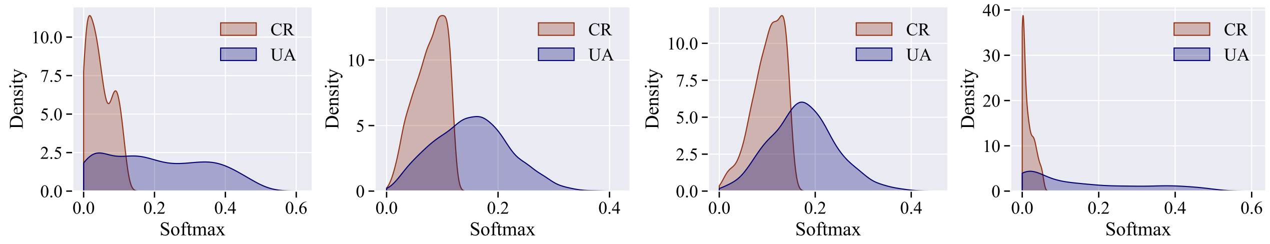
Table 16: Performance of our unlearning framework. We show the unlearning performance on **CIFAR-10** with **ResNet-18** and **Tiny ImageNet** with **ViT** in 10% **random data forgetting** scenario.

Methods	α			$\boldsymbol{\lambda = 0.2}$					$\boldsymbol{\lambda = 0.5}$					$\boldsymbol{\lambda = 1}$		
		UA ↑	RA ↑	TA ↑	CR$_{\mathcal{D}_f}$ ↓	CR$_{\mathcal{D}_{test}}$ ↑	UA ↑	RA ↑	TA ↑	CR$_{\mathcal{D}_f}$ ↓	CR$_{\mathcal{D}_{test}}$ ↑	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(0.486)	0.516(0.433)	26.4%(11.7)	98.7%(0.1)	85.8%(0.2)	0.396(0.548)	0.489(0.460)	35.7%(21.0)	98.6%(0.2)	85.2%(0.8)	0.346(0.598)	0.481(0.468)
	0.1				0.729(0.163)	0.786(0.114)				0.649(0.243)	0.765(0.135)				0.549(0.343)	0.739(0.161)
	0.15				0.841(0.000)	0.889(0.039)				0.768(0.073)	0.88(0.030)				0.658(0.183)	0.861(0.011)
	0.2				0.898(0.108)	0.932(0.133)				0.839(0.049)	0.929(0.130)				0.743(0.047)	0.918(0.119)
FT	0.05	9.8%(4.9)	97.4%(1.4)	83.6%(2.4)	0.441(0.503)	0.399(0.550)	13.6%(0.9)	97.2%(1.6)	83.6%(2.4)	0.413(0.531)	0.401(0.548)	20.0%(5.3)	96.4%(2.4)	82.9%(3.1)	0.342(0.602)	0.363(0.586)
	0.1				0.753(0.139)	0.683(0.217)				0.718(0.174)	0.683(0.217)				0.627(0.265)	0.652(0.248)
	0.15				0.884(0.043)	0.823(0.027)				0.848(0.007)	0.819(0.031)				0.772(0.069)	0.802(0.048)
	0.2				0.942(0.152)	0.893(0.094)				0.914(0.124)	0.890(0.091)				0.856(0.066)	0.877(0.078)
RL	0.05	31.8%(17.1)	95.3%(17.9)	80.9%(5.1)	0.051(0.893)	0.111(0.838)	36.2%(21.5)	95.3%(3.5)	80.4%(5.6)	0.051(0.893)	0.121(0.828)	40.2%(25.5)	94.5%(4.3)	79.5%(6.5)	0.048(0.896)	0.119(0.830)
	0.1				0.278(0.614)	0.451(0.449)				0.254(0.638)	0.449(0.451)				0.236(0.656)	0.436(0.464)
	0.15				0.579(0.262)	0.710(0.140)				0.541(0.300)	0.708(0.142)				0.480(0.361)	0.673(0.177)
	0.2				0.752(0.038)	0.825(0.026)				0.718(0.072)	0.827(0.028)				0.642(0.148)	0.793(0.006)



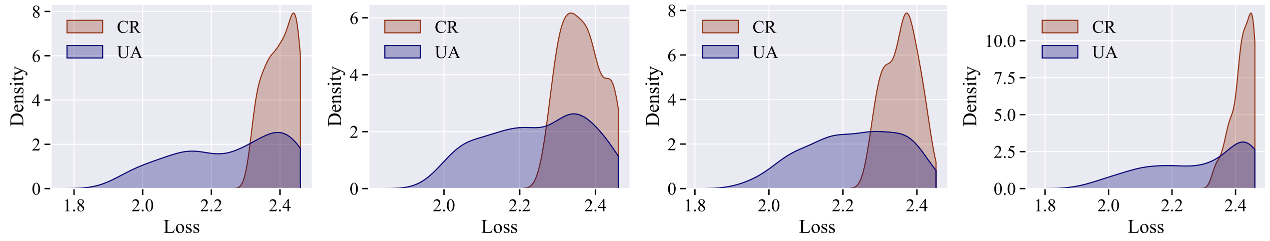
(a) FT (b) RL (c) Salun (d) SFRon

Figure 5: Softmax distribution in 10% **random data forgetting** scenario. We analyze the softmax distributions of ground truth 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".



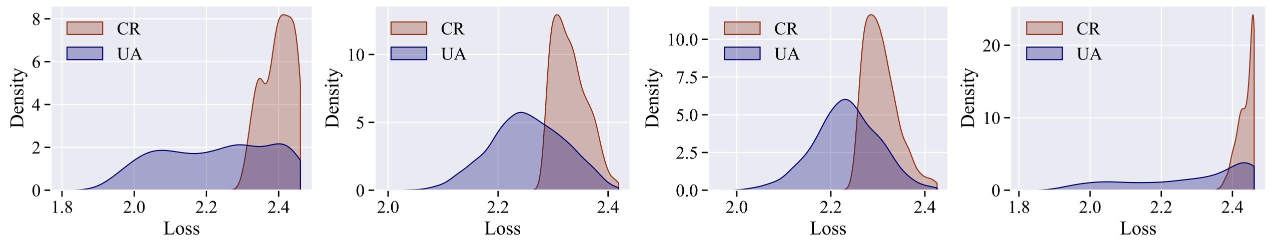
(a) FT (b) RL (c) Salun (d) SFRon

Figure 6: Softmax distribution in 50% **random data forgetting** scenario.



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

Figure 7: Loss distribution in 10% **random data forgetting** scenario. We analyze the cross entropy loss distributions of ground truth 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 8: Loss distribution in 50% **random data forgetting** scenario.