OMS: One More Step Noise Searching to Enhance Membership Inference Attacks for Diffusion Models

Xiaomeng Fu[1,3], Xi Wang[*2,4], Qiao Li[1,3], Jin Liu[1,3], Jiao Dai[*1], Jizhong Han[1], Xingyu Gao[2,4]
[1]Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
[2]Institute of Microelectronics, Chinese Academy of Sciences, Beijing, China
[3]School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China
[4]University of Chinese Academy of Sciences, Beijing, China
fuxiaomeng@iie.ac.cn, wangxiboss@163.com, {liqiao,liujin,daijiao,hanjizhong}@iie.ac.cn,
gxy9910@gmail.com

Abstract

The data-intensive nature of Diffusion models amplifies the risks of privacy infringements and copyright disputes, particularly when training on extensive unauthorized data scraped from the Internet. Membership Inference Attacks (MIA) aim to determine whether a data sample has been utilized by the target model during training, thereby serving as a pivotal tool for privacy preservation. Current MIA employs the prediction loss to distinguish between training member samples and non-members. These methods assume that, compared to non-members, members, having been encountered by the model during training result in a smaller prediction loss. However, this assumption proves ineffective in diffusion models due to the random noise sampled during the training process. Rather than estimating the loss, our approach examines this random noise and reformulate the MIA as a noise search problem, assuming that members are more feasible to find the noise used in the training process. We formulate this noise search process as an optimization problem and employ the fixed-point iteration to solve it. We analyze current MIA methods through the lens of the noise search framework and reveal that they rely on the first residual as the discriminative metric to differentiate members and non-members. Inspired by this observation, we introduce **OMS**, which augments existing MIA methods by iterating **O**ne **M**ore fixed-point **S**tep to include a further residual, i.e., the second residual. We integrate our method into various MIA methods across different diffusion models. The experimental results validate the efficacy of our proposed approach.

1 Introduction

Recently, diffusion models [Ho *et al.*, 2020; Song *et al.*, 2020b] have been widely recognized for their unparalleled capability to generate images of exceptional quality, which are increasingly becoming indistinguishable from their real-world counterparts. Due to the high quality of images generated by diffusion models, an increasing number of AI companies are developing generative tools predicated on diffusion models for commercial art design.

Nonetheless, these advancements are accompanied by inherent challenges [Brittain, 2023; Liu *et al.*, 2021]. The data-intensive nature of diffusion models has amplified the risk of privacy infringements and copyright disputes. Trained on extensive unauthorized data scraped from the Internet, these methods overlook the copyrights and privacy of the original owners. A case in point is the recent lawsuit filed by Getty Images against Stability AI, alleging unauthorized use of 12 million of Getty's images for model training. Thus, it is imperative to develop tools to detect diffusion models' privacy infringements.

To audit these privacy risks, Membership Inference Attacks (MIA) [Shokri *et al.*, 2017] have emerged as a potential solution. The objective of MIA is to ascertain whether a data sample has been utilized in the training process of a machine learning model. Existing MIA methods [Sablayrolles *et al.*, 2019; Salem *et al.*, 2019; Song and Mittal, 2021] typically operate under the assumption that member records tend to exhibit lower prediction losses compared to non-member records. Consequently, these methodologies compute the prediction losses and utilize this metric to differentiate between member and non-member records.

Although the utilization of prediction loss to differentiate between member and non-member records has been empirically validated for numerous deterministic models, such as classification models and Generative Adversarial Networks (GANs) [Chen *et al.*, 2020; Hayes *et al.*, 2019; Hilprecht *et al.*, 2019; Choquette-Choo *et al.*, 2021; Hanzlik *et al.*, 2021], its efficacy is diminished when applied to diffusion models due to the intractability of the training loss. More precisely, during the training process of the diffusion model, a random noise is sampled, serving not only as a component of the model's input but also as the target in the training loss. However, during the execution of the membership inference, it is virtually impossible to replicate the exact noise sampled during the training phase. The discrepancy between the noise used during training and membership inference contributes

[*] Corresponding author

to the inaccuracy of the loss estimation.

Instead of the loss assumption, we propose an alternative hypothesis: **it is more feasible for members to find the noise counterpart used in the training process**. This assumption aligns more closely with the inherent stochastic characteristics of diffusion model training. Based on this assumption, we introduce a novel MIA framework for diffusion models, leveraging a noise searching mechanism. We formalize the noise searching process as an optimization problem with the training loss as the optimization objective.

Moreover, we propose to utilize the fixed-point iteration to solve the optimization problem. By iteratively applying a function to the initial guess, we strive to facilitate the convergence to the noise encountered by the members during the training stage. We begin with an empirical analysis focuses on the convergence properties of the fixed-point iteration. We discern a distinct attribute where member samples exhibit faster convergence rate compared to non-member samples. This observation implies that the convergence rate can be employed as a discriminative feature to differentiate between member and non-member samples.

This attribute also provides further insights into current MIAs for diffusion models. Specifically, from the perspective of the fixed-point iteration, we reinterpret current MIA methods as assessments of convergence rate, primarily through the first residual. To refine this measurement and capture the nuances of convergence dynamics more comprehensively, we introduce an augmentation to the iteration process, incorporating an additional step that considers the second residual which is termed as the "**O**ne **M**ore **S**tep" (**OMS**) approach.

We conduct experiments across various diffusion models, spanning CNN-based and Transformer-based architectures, along with various datasets and MIA methods. Notably, to the best of our knowledge, **we are the first** to evaluate MIA performance on Transformer-based diffusion models. The experimental results demonstrate the effectiveness of the proposed **OMS** approach and the noise searching MIA framework. In summary, our paper makes the following contributions:

- We reveal the noise inconsistency issues in current MIA methods for diffusion models. To address this, we devise a novel framework in the perspective of noise searching. Formally, we conceptualize the noise searching process as an optimization problem.
- We propose the fixed-point iteration to solve the noise searching optimization problem. Moreover, we investigate its convergence properties in practice and find that members exhibit faster convergence rate compared to non-members.
- We analyze existing MIA methods through the proposed framework, revealing that the efficacy of existing methods is linked to the convergence rate, particularly as characterized by the first residual. Motivated by this, we introduce a refinement strategy by iterating **O**ne **M**ore **S**tep (**OMS**) to include the second residual.
- We conduct experiments on various diffusion models, encompassing CNN-based and Transformer-based architectures, using various datasets. The results not only confirm the validity of our MIA framework but also underscore the efficacy of the OMS refinements.

2 Related Work

Membership Inference Attack (MIA). The goal of MIA is to predict the presence of a specific data record in the training set of a given model. The effectiveness of MIA fundamentally relies on the hypothesis that machine learning models exhibit differential responses to member records versus unfamiliar non-member records. Given the manner to exploit model's reactions, existing methods can be divided into two categories, **model-based methods** and **metric-based methods**. In the realm of model-based methods [Shokri *et al.*, 2017; Salem *et al.*, 2019; Long *et al.*, 2020; Chen *et al.*, 2020; Truex *et al.*, 2019], a shadow model is trained to mimic the responses of the target machine learning model. Subsequently, attack algorithms are formulated, predicated on the reactions of the shadow model, with the ultimate objective of achieving generalization to the target model. Despite the significant advancements in the field, model-based methods are characterized by their computational intensity and exhibit susceptibility to model's architecture. Methods grounded in metrics [Sablayrolles *et al.*, 2019; Yeom *et al.*, 2018; Salem *et al.*, 2020; Bentley *et al.*, 2020] primarily employ a metric (typically the loss value) as a representative measure of the model's response to each sample. The membership of a specific sample is subsequently determined based on the numerical values of the selected metric.

MIA for diffusion models. Given the computational intensity of training a shadow model with comparable parameters, model-based methods are deemed unsuitable for diffusion models. As a result, most current MIA tailored for diffusion models [Duan *et al.*, 2023; Matsumoto *et al.*, 2023; Kong *et al.*, 2023] are metric-based methods. These methods assume that the loss value for members is smaller than that for non-members. Despite achieving substantial performance, these methods still suffer from the inaccuracy of loss estimation. During the training phase of the diffusion model, the loss value is dictated by the training target (a random Gaussian noise). However, when executing the MIA, replicating the exact noise sampled during the training process is virtually unattainable. Instead, we propose a novel MIA framework predicated on noise searching. This innovative approach promises to enhance the overall performance of MIA and provide insight into the principles of MIA methods for diffusion models.

3 Method

Given a data record x_0, the goal of MIA is to identify whether x_0 is in the training set of the target diffusion model ϵ_θ. Existing MIAs [Duan *et al.*, 2023; Matsumoto *et al.*, 2023; Kong *et al.*, 2023] tailored for diffusion models predominantly assume the members' loss values are lower than those for non-members. However, these loss-based approaches are susceptible to inaccuracies in loss estimation, which are caused by noise inconsistency between the training and inference stages (Section 3.2). In contrast, we propose a novel MIA framework that employs noise searching, an approach we believe aligns more closely with the stochastic nature of the model's training process. We formulate the noise searching process as an optimization problem and use the fixed-point iteration to solve this problem (Section 3.3). We posit that members can retrieve the training noise with less effort compared to non-members. Subsequently, we analyze the convergence properties of the fixed-point iteration and further validate that it is more feasible for members to search the noise than non-members (Section 3.4). Motivated by this, we reinterpret the underlying mechanisms contributing to the efficacy of existing MIA methods and propose an enhancement by incorporating **O**ne **M**ore iteration **S**tep (**OMS**) (Section 3.5).

3.1 Background and Notations

We begin with a brief introduction of the background and notations of the diffusion models. Denoising Diffusion Probabilistic Models (DDPM) [Ho *et al.*, 2020; Song *et al.*, 2020a] consist of a forward and a reverse process. The forward process, also named as the diffusion process, gradually adds Gaussian noise to the input image x_0 in T time steps according to a predefined variance schedule $\beta_1, ..., \beta_T$:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t\mathbf{I}) \tag{1}$$

Let $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$, this process can be simplified to:

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)\mathbf{I}) \tag{2}$$

When t is large enough, the $\bar{\alpha}$ is approaching 0, making x_t an isotropic Gaussian noise. The reverse process aims to recover the data distribution from the Gaussian noise. The reverse process in one step can be represented as:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_t) \tag{3}$$

where Σ_t is a constant depending on the variance schedule β_t and $\mu_\theta(x_t, t)$ is determined by a neural network:

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t, t)) \tag{4}$$

By recursively leveraging the reverse step, Gaussian noise can be recovered to the original image. To train the DDPM, an image x_0, a timestep t and a random noise $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ are first sampled. A noisy image x_t is then obtained by using the forward process (Equation 2). We then input both the noisy image x_t and the timestep t into a U-Net [Ronneberger *et al.*, 2015] ϵ_θ to predict the noise within x_t. The optimization objective for the denoising U-Net can be written as:

$$\mathcal{L} = \mathbb{E}_{t,x_0,\epsilon}[||\epsilon - \epsilon_\theta(x_0, t, \epsilon)||_2^2] \tag{5}$$

3.2 Noise Inconsistency Between Training and Inference

The diffusion model's training procedure can be described by Equation 5. To elaborate, given the input image x_0 and a specific timestep t, a random noise ϵ_{train} is sampled from the standard normal distribution. This noise is then utilized to perturb x_0 into a corrupted version x_t, following the schedule predefined in Equation 2. Subsequently, the diffusion model, parameterized by θ generates a prediction of the noise within x_t (denoted as $\epsilon_\theta(x_0, t, \epsilon_{train})$). The training loss for the diffusion model is computed as the distance between the predicted noise $\epsilon_\theta(x_0, t, \epsilon_{train})$ and the actual sampled noise ϵ_{train}. During the inference phase, due to the infeasibility of the training noise ϵ_{train}, an alternate noise ϵ_{inf} is sampled to estimate the loss value. However, it is important to note that there exists no guarantee that ϵ_{inf} is identical or approximately similar to the noise ϵ_{train}. This inconsistency in noise significantly impacts the accuracy of the loss values, consequently affecting the effectiveness of existing MIA targeting diffusion models.

3.3 MIA by Noise Searching

In contrast to existing MIA that rely on the randomly sampled inference noise as a surrogate for the training noise ϵ_{train} to approximate the loss, thereby encountering the noise inconsistency issue, we introduce a novel MIA framework for diffusion models, leveraging a noise search strategy. Our approach aims to reconstruct, for a given record x_0, the corresponding training noise ϵ_{train} that minimizes the training loss as defined in Equation 5. We assume that **it is more feasible for members to obtain the training noise** ϵ_{train}. We formulate the process of noise searching as an optimization problem:

$$\begin{aligned} &\min_{\epsilon} ||\epsilon - \epsilon_\theta(x_0, t, \epsilon)||_p \\ &s.t. \quad \epsilon \sim \mathcal{N}(0, \mathbf{I}) \end{aligned} \tag{6}$$

This optimization framework directly addresses the noise inconsistency issue by focusing on identifying the true training noise ϵ_{train}. Compared with loss-based methods, the proposed approach emphasizes the identification of ϵ_{train}. We argue this approach aligns more coherently with the inherent stochastic nature of the diffusion model's training process.

Fixed-point iteration. We address the aforementioned optimization problem by the fixed-point iteration. For a given record x_0 and timestep t, the predicted noise $\epsilon_\theta(x_0, t, \epsilon)$ is solely dependent on ϵ. This dependency can be represented as an implicit function $\epsilon = f(\epsilon)$. The optimal noise, which pairs with the record x_0 during the training process, is also identified as the solution to the implicit function f. To address this, we employ the fixed-point iteration [Smart, 1980]. The iterative process can be represented as follows:

$$\epsilon^n = f(\epsilon^{n-1}), \quad n = 1, 2, ... \tag{7}$$

We assume that the fixed-point iteration process essentially satisfies the constraints embedded within the optimization problem, given the fact that the model ϵ_θ is trained to generate noises adhering to the distribution. Consequently, we hypothesize that the model's outputs also conform to the standard normal distribution. Note that we do not use more advanced methods for solving implicit functions such as Newton-Raphson or Conjugate Gradient [Nocedal and Wright, 1999]. This is because the Newton-Raphson method needs to compute gradients while the Conjugate Gradients need to search high dimension gradients, both of which are computationally intensive and potentially intractable.

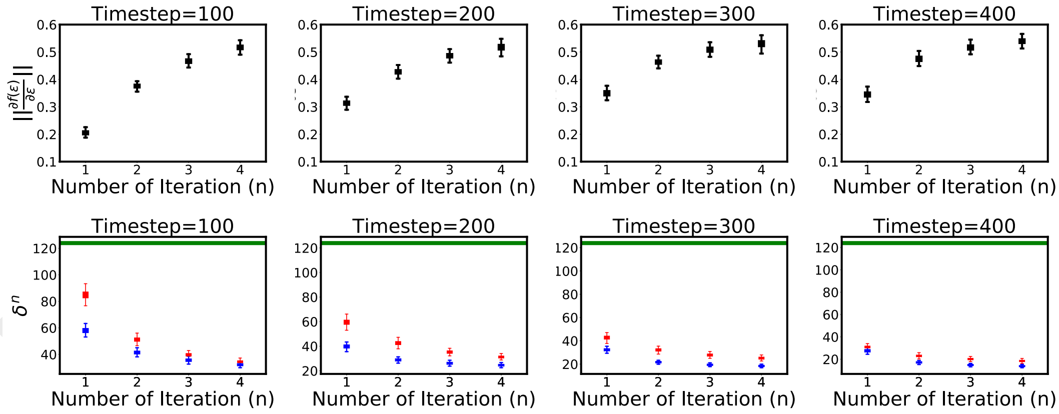


Figure 1: The top row is the Jacobian norm in different timesteps and iterations. The bottom row shows the residuals δ^n, with the blue and red boxes representing the members and non-members. The green line is the theoretical distance between two random Gaussian noise. We report the results calculated over 1000 images of Cifar10 dataset (500 members and 500 non-members). The diffusion model is also trained on the Cifar10 dataset.

3.4 Convergence of the Fixed-point Iteration

The primary concern about the fixed-point iteration lies in its practical convergence properties. We present an empirical analysis to address this concern. Given the initial ϵ^0, our objective is to demonstrate that the sequence $\{\epsilon^n\}, n \to \infty$ generated by Equation 7 converges. To achieve this, we aim to prove that the residual $\delta^n = \epsilon^n - \epsilon^{n-1}$ converges, which would imply that $\{\epsilon^n\}$ is the Cauchy sequence. The residual can be expressed as follows:

$$||\delta^{n+1}|| = ||\epsilon^{n+1} - \epsilon^n|| = ||f(\epsilon^n) - f(\epsilon^{n-1})|| \quad (8)$$

Through Taylor Expansion, we obtain:

$$\begin{aligned} & ||f(\epsilon^n) - f(\epsilon^{n-1})|| \\ =& ||f(\epsilon^{n-1}) + \frac{\partial f(\epsilon)}{\partial \epsilon}|_{\epsilon=\epsilon^{n-1}} \cdot \delta^{n-1} + \mathcal{O}(||\delta^{n-1}||^2) - f(\epsilon^{n-1})|| \quad (9) \\ \leq& ||\frac{\partial f(\epsilon)}{\partial \epsilon}|_{\epsilon=\epsilon^{n-1}}|| \cdot ||\delta^n|| + \mathcal{O}(||\delta^n||^2) \end{aligned}$$

In a sufficiently confined domain, the term $||\mathcal{O}(\delta^2)||$ can be considered negligible, and the convergence dynamics are primarily governed by the Jacobian norm $||\frac{\partial f(\epsilon)}{\partial \epsilon}||$. If the Jacobian norm is below 1, it indicates that the implicit function f is contractive, leading to an exponential decay in the residuals, thereby affirming the convergence of the fixed-point iteration. We visualize this Jacobian norm (the top row) along with the residuals (the bottom row) across various iterations and timesteps in Figure 1. Notably, the Jacobian norm consistently remains below the threshold of 1, thereby empirically validating the convergence of the fixed-point iteration. In the residual plots, the theoretical distance (i.e., the distance between two random Gaussian noise) is depicted in green, whereas the residuals for member and non-member sets are depicted in blue and red, respectively. Moreover, it is observed that the residuals for both member and non-member sets exhibit rapid convergence, with the member set residuals smaller than those of non-member set, further indicating that it is more feasible for members to obtain the training noise. It is also important to highlight that the first residual δ^1, representing the divergence between ϵ^0 and ϵ^1, is frequently employed as an estimation of the training loss in current MIA methods. We also provide validation about the convergence property and convergence speed from the lens of contraction mapping theorem [Berinde and Takens, 2007].

3.5 Existing MIAs and One More Step (OMS)

We re-assess the efficacy of prevailing MIA methods through the lens of the fixed-point iteration. Current MIA methods approximate the training loss of diffusion models by measuring the divergence between the initial noise ϵ^0 and the noise after the first iteration ϵ^1. In the context of the fixed-point iteration, this divergence is equivalently characterized as the first residual δ^1. As illustrated in the bottom row of Figure 1, this residual effectively discriminates between members and non-members. However, it is also discernible that the second residual δ^2 for members is smaller than non-members, indicating its substantial potential for enhancing membership discrimination. Motivated by this observation, we take **O**ne **M**ore **S**tep (**OMS**) beyond ϵ^1 to obtain ϵ^2 and utilize the distance between ϵ^0 and ϵ^2 as the discriminative metric. This metric can be interpreted as an ensemble of the first and second residuals (δ^1 and δ^2). This approach not only preserves the discriminative capability inherent in the traditional loss-based approach (i.e., δ^1), but also incorporating the extra information δ^2 to augment the performance. The relationship is mathematically expressed as:

$$||\epsilon^0 - \epsilon^2|| = ||\underbrace{(\epsilon^0 - \epsilon^1)}_{loss\,term} + \underbrace{(\epsilon^1 - \epsilon^2)}_{extra\,term}|| = ||\delta^1 + \delta^2|| \quad (10)$$

Note that we do not leverage further residuals such as δ^3 and δ^4, though they also seem potential metric to distinguish the member and non-member records. This is because the marginal gain is decreased. We also provide experiments utilizing different residuals in Section 4.5.

Method	Cifar10		Cifar100		LFW		LSUN-Cat		Ave	
	ASR	AUC	ASR	AUC	ASR	AUC	ASR	AUC	ASR	AUC
NA	71.86	78.28	75.97	82.18	72.40	79.54	62.60	67.02	70.71	76.75
+OMS	78.21	85.71	81.49	88.27	83.02	90.73	68.89	75.10	77.90	84.95
$\Delta\uparrow$	**+6.35**	**+7.43**	**+5.52**	**+6.09**	**+10.62**	**+11.20**	**+6.29**	**+8.09**	**+7.19**	**+8.20**
SecMI	84.01	90.68	79.83	86.77	61.49	64.90	73.13	79.36	74.62	80.43
+OMS	86.48	92.82	83.32	90.62	77.80	85.67	84.59	91.20	83.05	90.08
$\Delta\uparrow$	**+2.47**	**+2.14**	**+3.49**	**+3.85**	**+16.31**	**+20.77**	**+11.46**	**+11.84**	**+8.43**	**+9.65**
PIA	88.75	94.89	85.20	92.21	82.10	90.17	77.61	84.87	83.42	90.54
+OMS	91.78	97.26	89.68	96.08	84.49	92.27	82.58	89.65	87.13	93.82
$\Delta\uparrow$	**+3.03**	**+2.37**	**+4.49**	**+3.86**	**+2.39**	**+2.11**	**+4.96**	**+4.78**	**+3.72**	**+3.28**

Table 1: The ASR and AUC metrics for existing MIA methods on DDPM, both with and without the integration of the One More Step (OMS). The symbol Δ is employed to denote the improvement in performance resulting from the integration of the OMS procedure.

Method	TPR@1%FPR				TPR@0.1%FPR			
	Cifar10	Cifar100	LFW	LSUN-Cat	Cifar10	Cifar100	LFW	LSUN-Cat
NA	6.42	3.66	10.86	3.40	0.88	0.23	1.12	0.36
+OMS	12.12	8.03	30.01	6.37	2.34	0.66	5.66	0.79
$\Delta\uparrow$	**+5.70**	**+4.37**	**+19.15**	**+2.98**	**+1.46**	**+0.44**	**+4.54**	**+0.43**
SecMI	9.15	7.19	3.65	3.10	0.49	0.22	0.42	0.12
+OMS	15.87	17.33	28.05	11.24	0.99	1.33	7.64	0.56
$\Delta\uparrow$	**+6.72**	**+10.14**	**+24.41**	**+8.14**	**+0.50**	**+1.11**	**+7.21**	**+0.44**
PIA	28.86	19.41	25.74	8.90	1.05	2.31	7.00	0.42
+OMS	60.11	48.30	29.72	13.59	13.24	10.66	9.22	0.94
$\Delta\uparrow$	**+31.25**	**+28.89**	**+3.99**	**+4.69**	**+12.19**	**+8.35**	**+2.22**	**+0.52**

Table 2: The TPR at extremely low FPR for existing MIA methods on DDPM, both with and without the integration of the One More Step (OMS). The symbol Δ is employed to denote the improvement in performance resulting from the integration of the OMS procedure.

4 Experiment

4.1 Experimental Setup

Diffusion Models and Datasets. We evaluate our proposed method across diverse diffusion models, specifically DDPM [Ho *et al.*, 2020], Stable Diffusion [Rombach *et al.*, 2022] and U-ViT [Bao *et al.*, 2023]. DDPM represents a foundational approach in the realm of diffusion models, which employs convolutional neural networks as the backbone. We train DDPM on four datasets: Cifar10, Cifar100 [Krizhevsky *et al.*, 2009], LFW [Huang *et al.*, 2008] and Lsun-Cat [Yu *et al.*, 2015]. The Stable Diffusion models, which are prominently recognized for their text-to-image synthesis capabilities, have undergone numerous iterations. We selectively adopt SD1.5 and SD2.1 due to their widespread usage and recognition within the research community. U-ViT, a recently introduced diffusion model, incorporates transformers as its core architecture. Our investigation leverages the open-source implementation of U-ViT which has been trained on the Cifar10 datasets.

Evaluation Metrics. To evaluate the performance of our proposed method, we adopte established metrics in previous works [Carlini *et al.*, 2022; Carlini *et al.*, 2023; Duan *et al.*, 2023; Kong *et al.*, 2023] including Attack Success Rate (ASR), AUC and the True Positive Rate (TPR) at extremely low False Positive Rate (FPR). Specifically, TPR@1%FPR and TPR@0.1%FPR refer to the True Positive Rate (TPR) when the False Positive Rate (FPR) is constrained to 1% and 0.1%, respectively.

Implementation Details. To evaluate the effectiveness of the proposed OMS, we conduct a series of experiments, aligning our benchmarks with state-of-the-art MIA methods designed for diffusion models, which include the Naive Attack (NA) [Matsumoto *et al.*, 2023], SecMI [Duan *et al.*, 2023], PIA [Kong *et al.*, 2023], GSA [Pang *et al.*, 2023] and Quantile [Bertran *et al.*, 2024; Tang *et al.*, 2023]. We strictly follow the prescribed settings of these methods, and exclusively introduce a further fix-point iteration. Notably, our approach not only seamlessly integrates with these established MIA methods but also augments their performance.

4.2 Evaluation Results

Performance on DDPM. The comparative results on DDPM, with and without OMS, are presented in Table 1. It can be observed that the OMS confers substantial improvements in performance, with increases of 8.20%, 9.65% and 3.28% in the Average AUC across the four datasets, compared to those baselines (NA, SecMI, PIA) without OMS. The improvements demonstrate the advantage of executing multiple fixed-point iterations over the conventional single-iteration approaches. We also observe that our method is particularly effective for weak attackers: an AUC increase from 64.90 to 85.67 for SecMI. Besides, we also note our method can further boost strong attackers with an average 3.72% AUC improvement for PIA. There results demonstrate the broad applicability of our proposed OMS. Furthermore, we provide the results of TPR at extremely low FPR in Table 2. These results demonstrate that the OMS notably enhances the

Method	SD1.5				SD2.1			
	ASR	AUC	TPR@1%	TPR@0.1%	ASR	AUC	TPR@1%	TPR@0.1%
NA	71.04	76.67	19.64	4.62	69.58	74.99	18.76	4.42
+OMS	73.34	79.00	24.40	8.89	71.45	77.40	23.14	8.47
$\Delta\uparrow$	**+2.30**	**+2.33**	**+4.76**	**+4.26**	**+1.86**	**+2.41**	**+4.38**	**+4.04**
SecMI	57.20	57.60	6.29	2.04	57.24	56.61	3.98	0.78
+OMS	60.62	61.38	13.21	5.47	61.84	62.29	10.97	3.24
$\Delta\uparrow$	**+3.42**	**+3.77**	**+6.93**	**+3.42**	**+4.60**	**+5.69**	**+6.99**	**+2.46**
PIA	63.17	67.59	12.71	3.76	71.15	78.38	18.56	3.36
+OMS	72.08	78.89	25.33	4.26	77.59	85.45	30.47	9.61
$\Delta\uparrow$	**+8.90**	**+11.30**	**+12.61**	**+0.50**	**+6.44**	**+7.07**	**+11.91**	**+6.25**

Table 3: Performance of existing MIA methods on text-to-image diffusion models, both with and without the integration of the One More Step (OMS). The symbol Δ is employed to denote the improvement in performance resulting from the integration of the OMS procedure.

Method	Without OMS			With OMS		
	ASR	AUC	TPR@1%FPR	ASR(Δ)	AUC(Δ)	TPR@1%FPR(Δ)
NA	61.47	63.66	2.62	68.51(**+7.04**)	74.31(**+10.65**)	8.65(**+6.03**)
SecMI	68.21	74.44	12.88	74.95(**+6.74**)	82.31(**+7.87**)	24.35(**+11.47**)
PIA	54.60	52.91	1.80	-	-	-
PIAN	59.36	61.13	3.42	69.11(**+9.75**)	75.42(**+14.29**)	9.86(**+6.44**)

Table 4: Performance of existing MIA methods on U-ViT, both with and without the integration of the One More Step (OMS). The symbol Δ is employed to denote the improvement in performance resulting from the integration of the OMS procedure.

prediction confidence, thereby amplifying the practical applicability in scenarios requiring high prediction certainty.

Performance on text-to-image diffusion models. Distinct from unconditional diffusion models, text-to-image diffusion models require dual inputs: the image itself and an accompanying text. However, in real-world scenario, images are seldom annotated by texts. It is a common case that users do not have access to the text employed during the training phase. To replicate this real-world scenario, we leverage BLIP [Li *et al.*, 2022] to generate text captions for the input images. The results on text-to-image diffusion models are detailed in Table 3. These evaluation further corroborate the substantial performance improvements that can be achieved by incorporating the OMS into current MIA methods.

Performance on Transformer-based diffusion models. The traditional diffusion models predominantly leverage CNNs as their backbone. However, recent advancements have seen an increasing trend towards the adoption of Transformers as the foundational architecture [Bao *et al.*, 2023; Chen *et al.*, 2024; Peebles and Xie, 2023; Esser *et al.*, 2024]. To assess the efficacy of existing MIA methods on Transformer-based diffusion models and substantiate the effectiveness of our approach, we conduct experiments utilizing the U-ViT model, a continuous time diffusion model based on Transformers. Notably, the majority of existing MIA methods are specifically designed for discrete time diffusion models, posing a challenge for direct application to the U-ViT model. To address this, we implement a simple mapping strategy, converting the discrete timestep within the range [0, 1000] to the continuous range [0, 1]. Additionally, we observe that the performance of PIA approximates random guessing. To mitigate this issue, we leverage a regularization technique [Kong *et al.*, 2023], hereby referred to as PIAN. The results, presented in Table 4, reveal that the incorporation of an additional fixed-point iteration, as proposed in our method, led to performance improvements in existing methods, suggesting the robustness and efficacy of OMS approach across diffusion models with diverse architectures.

Method	DDPM-Cifar10			U-ViT-Cifar10		
	@5%	@1%	@0.1%	@5%	@1%	@0.1%
QR (t-error)	27.76	6.10	0.38	17.15	3.13	0.54
QR (t-error+OMS)	44.66	17.14	1.38	31.97	8.62	1.20
$\Delta\uparrow$	**+16.90**	**+11.04**	**+1.00**	**+14.82**	**+5.49**	**+0.66**

Table 5: Performance of OMS in Quantile Regression (QR). @5%, @1%, @0.1% is short for the TPR value when the FPR is set to 5%, 1% and 0.1% separately.

4.3 Integration with Quantile Regression

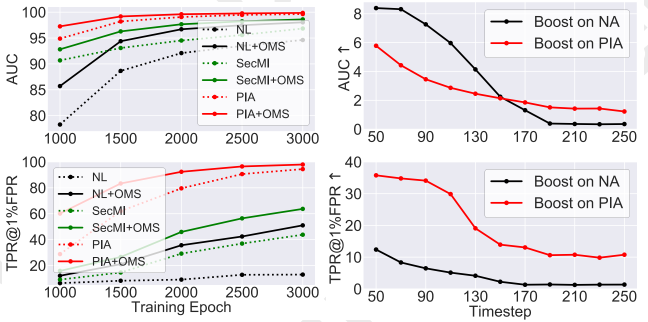
Quantile Regression [Tang *et al.*, 2023] incorporates the t-error metric (proposed by SecMI [Duan *et al.*, 2023]) to learn a quantile regression model that predicts the α-quantile of the t-error for each individual sample. This approach enables the estimation of a sample-specific α-quantile as a refined per-sample threshold for identifying membership status. While t-error serves as a fundamental confidence metric for quantile regression, we have shown that the t-error can be augmented through OMS (Table 1- 3). Similarly, we refine quantile regression by incorporating an additional fixed-point iteration to current confidence metric (t-error). This refinement leads to improved performance, as evidenced in Table 5.

4.4 Integration with Gradient-based Method

GSA [Pang *et al.*, 2023] constitutes a gradient-based MIA method which posits that the gradients inherently convey a

Model	Method	ASR	AUC	TPR@1%	TPR@0.1%
SD1.5	GSA	87.56	94.19	55.33	21.85
	GSA+OMS	88.12	94.53	56.17	22.18
	Δ ↑	**+0.56**	**+0.34**	**+0.84**	**+0.33**
SD2.1	GSA	87.94	94.50	55.86	21.37
	GSA+OMS	88.64	95.11	58.82	22.53
	Δ ↑	**+0.70**	**+0.61**	**+2.96**	**+1.16**

Table 6: Performance of OMS in Gradient-Based Method (GSA).



(a) The results of OMS for different training epochs.

(b) The results of OMS for different timesteps.

Figure 2: The results of AUC and TPR@1%FPR metrics of OMS for different training epochs and different timesteps on Cifar10.

more direct indication of how the target model responds to member and non-member samples. As a white-box attacker, GSA demonstrates significant efficacy against diffusion models compared to other attackers. We concentrate on the back-propagation GSA, which harnesses the backward pass of gradient computation during loss optimization to execute MIA. We refine this back-propagation GSA by backwarding the loss after OMS (Equation 7). Notably, as evidenced in Table 6, the OMS is also capable of enhancing the efficacy of gradient-based methods.

4.5 Ablation Study

The Training Steps. Previous researches [Leino and Fredrikson, 2020; Salem *et al.*, 2019] have highlighted the tendency of machine learning models to memorize training data as the training procedure progresses. Based on these insights, we conduct an evaluation of our method throughout the training process. The results are presented in Figure 2(a). We observe that all these examined MIA methods exhibit enhanced performance as the training epochs increase, which corroborates the phenomenon of model memorization. Another notable observation is the consistent efficiency of our method throughout the entire training process. Furthermore, we identify that PIA begins to saturate in terms of AUC after 1500 training epochs. Our method further boosts PIA's performance on TPR@1%FPR, providing compelling evidence for the superiority and robustness of our approach.

The Timesteps. The timestep serves as a crucial parameter for the level of noise incorporated into the input of the denoising U-Net, significantly impacting the performance of the diffusion models. Consequently, we execute MIA across a range of timesteps, specifically from 50 to 250. The performance enhancements attributed to OMS are presented in Figure 2(b).

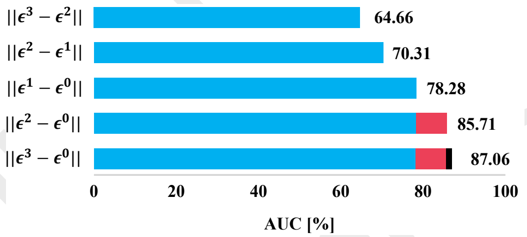


Figure 3: The results (AUC) utilizing different number of fixed-point iteration in Cifar10 dataset.

Our observations indicate that the OMS exhibits robust performance across timesteps but the improvements diminishes as the timestep increases. This observed decline can be attributed to the increasing prominence of noise in the model's input. Specifically, as the timestep increases, the noise component becomes the dominant factor, potentially disrupting the stability of the fixed-point iteration process.

The number of iteration. We incorporate an additional fixed-point iteration for computational efficiency, which is also validated by previous experimental results. In this experiment, we explore varying fixed-point iterations and harness the distance to execute MIA. The results are depicted in Figure 3. Specifically, $||\epsilon^1 - \epsilon^0||$ represents the NA approach, whereas $||\epsilon^2 - \epsilon^0||$ represents NA with OMS in previous experiments. It is evident that increasing the number of iterations leads to improved performance, with $||\epsilon^3 - \epsilon^0||$ demonstrating the optimal results. While residuals ($||\epsilon^2 - \epsilon^1||$ and $||\epsilon^3 - \epsilon^2||$) exhibit some level of effectiveness, their performance diminishes as the iteration count increases. It is also noteworthy that while additional fixed-point iterations hold the potential for superior performance, the marginal gains diminish progressively.

5 Conclusion

In this paper, we explore the MIA for diffusion models in a novel perspective, i.e., the noise searching. We first analyze the noise inconsistency issue between the training and membership inference stage. To address this issue, we introduce a noise searching framework that formulates the search for optimal training noise as an optimization problem. Utilizing the fixed-point iteration, we solve the optimization problem and conduct a thorough examination of its convergence properties, revealing distinct convergence rates between member and non-member data. Based on this observation, we rethink the effectiveness of current MIA methods and propose an enhancement through one more iteration step, resulting in a substantial performance boost for existing MIA methods. In conclusion, the proposed noise searching framework provides a unique and unified perspective for comprehending the fundamental principles of MIA tasks for diffusion models. We anticipate that our contributions will foster further research into the privacy risks associated with diffusion models and contribute to the ongoing research in this field.

Ethical Statement

The primary objective of our research is to devise a method capable of discerning whether a particular sample was included in the training dataset. The proposed method offers a multitude of beneficial applications, encompassing the detection of privacy violations and the assessment of model privacy. While acknowledging the potential for malevolent entities to misuse our method for privacy attacks, we underscore the capacity of privacy protection techniques, such as differential privacy, to counteract such threats. It is crucial to note that the development of these techniques is not intended to facilitate malicious activities, but rather to advance the field of privacy protection. We trust that our contributions will be used responsibly to enhance privacy protection measures and promote ethical practices in machine learning research.

Acknowledgments

This work was supported in part by Science and Technology Innovation (STI) 2030—Major Projects under Grant 2022ZD0208700, and National Natural Science Foundation of China under Grant 62376264.

References

[Bao *et al.*, 2023] Fan Bao, Shen Nie, Kaiwen Xue, Yue Cao, Chongxuan Li, Hang Su, and Jun Zhu. All are worth words: A vit backbone for diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22669–22679, 2023.

[Bentley *et al.*, 2020] Jason W Bentley, Daniel Gibney, Gary Hoppenworth, and Sumit Kumar Jha. Quantifying membership inference vulnerability via generalization gap and other model metrics. *arXiv preprint arXiv:2009.05669*, 2020.

[Berinde and Takens, 2007] Vasile Berinde and F Takens. *Iterative approximation of fixed points*, volume 1912. Springer, 2007.

[Bertran *et al.*, 2024] Martin Bertran, Shuai Tang, Aaron Roth, Michael Kearns, Jamie H Morgenstern, and Steven Z Wu. Scalable membership inference attacks via quantile regression. *Advances in Neural Information Processing Systems*, 36, 2024.

[Brittain, 2023] Blake Brittain. Getty images lawsuit says stability ai misused photos to train ai. *Reuters, Feb 6th*, 2023.

[Carlini *et al.*, 2022] Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. Membership inference attacks from first principles. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 1897–1914. IEEE, 2022.

[Carlini *et al.*, 2023] Nicolas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramer, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 5253–5270, 2023.

[Chen *et al.*, 2020] Dingfan Chen, Ning Yu, Yang Zhang, and Mario Fritz. Gan-leaks: A taxonomy of membership inference attacks against generative models. In *Proceedings of the 2020 ACM SIGSAC conference on computer and communications security*, pages 343–362, 2020.

[Chen *et al.*, 2024] Junsong Chen, YU Jincheng, GE Chongjian, Lewei Yao, Enze Xie, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-alpha: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *The Twelfth International Conference on Learning Representations*, 2024.

[Choquette-Choo *et al.*, 2021] Christopher A Choquette-Choo, Florian Tramer, Nicholas Carlini, and Nicolas Papernot. Label-only membership inference attacks. In *International conference on machine learning*, pages 1964–1974. PMLR, 2021.

[Duan *et al.*, 2023] Jinhao Duan, Fei Kong, Shiqi Wang, Xiaoshuang Shi, and Kaidi Xu. Are diffusion models vulnerable to membership inference attacks? *International Conference on Machine Learning*, 2023.

[Esser *et al.*, 2024] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024.

[Hanzlik *et al.*, 2021] Lucjan Hanzlik, Yang Zhang, Kathrin Grosse, Ahmed Salem, Maximilian Augustin, Michael Backes, and Mario Fritz. Mlcapsule: Guarded offline deployment of machine learning as a service. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3300–3309, 2021.

[Hayes *et al.*, 2019] J Hayes, L Melis, G Danezis, and E De Cristofaro. Logan: Membership inference attacks against generative models. In *Proceedings on Privacy Enhancing Technologies (PoPETs)*, pages 133–152. De Gruyter, 2019.

[Hilprecht *et al.*, 2019] Benjamin Hilprecht, Martin Härterich, and Daniel Bernau. Monte carlo and reconstruction membership inference attacks against generative models. *Proc. Priv. Enhancing Technol.*, 2019(4):232–249, 2019.

[Ho *et al.*, 2020] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[Huang *et al.*, 2008] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008.

[Kong *et al.*, 2023] Fei Kong, Jinhao Duan, RuiPeng Ma, Hengtao Shen, Xiaofeng Zhu, Xiaoshuang Shi, and Kaidi

Xu. An efficient membership inference attack for the diffusion model by proximal initialization. *arXiv preprint arXiv:2305.18355*, 2023.

[Krizhevsky *et al.*, 2009] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. Technical report, University of Toronto, Toronto, ON, Canada, 2009.

[Leino and Fredrikson, 2020] Klas Leino and Matt Fredrikson. Stolen memories: Leveraging model memorization for calibrated {White-Box} membership inference. In *29th USENIX security symposium (USENIX Security 20)*, pages 1605–1622, 2020.

[Li *et al.*, 2022] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pretraining for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022.

[Liu *et al.*, 2021] Changxin Liu, Zhenan Fan, Zirui Zhou, Yang Shi, Jian Pei, Lingyang Chu, and Yong Zhang. Achieving model fairness in vertical federated learning. *arXiv preprint arXiv:2109.08344*, 2021.

[Long *et al.*, 2020] Yunhui Long, Lei Wang, Diyue Bu, Vincent Bindschaedler, Xiaofeng Wang, Haixu Tang, Carl A Gunter, and Kai Chen. A pragmatic approach to membership inferences on machine learning models. In *2020 IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 521–534. IEEE, 2020.

[Matsumoto *et al.*, 2023] Tomoya Matsumoto, Takayuki Miura, and Naoto Yanai. Membership inference attacks against diffusion models. *arXiv preprint arXiv:2302.03262*, 2023.

[Nocedal and Wright, 1999] Jorge Nocedal and Stephen J Wright. *Numerical optimization*. Springer, 1999.

[Pang *et al.*, 2023] Yan Pang, Tianhao Wang, Xuhui Kang, Mengdi Huai, and Yang Zhang. White-box membership inference attacks against diffusion models. *arXiv preprint arXiv:2308.06405*, 2023.

[Peebles and Xie, 2023] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023.

[Rombach *et al.*, 2022] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[Ronneberger *et al.*, 2015] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

[Sablayrolles *et al.*, 2019] Alexandre Sablayrolles, Matthijs Douze, Cordelia Schmid, Yann Ollivier, and Hervé Jégou. White-box vs black-box: Bayes optimal strategies for membership inference. In *International Conference on Machine Learning*, pages 5558–5567. PMLR, 2019.

[Salem *et al.*, 2019] Ahmed Salem, Yang Zhang, Mathias Humbert, Mario Fritz, and Michael Backes. Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models. In *Network and Distributed Systems Security Symposium 2019*. Internet Society, 2019.

[Salem *et al.*, 2020] Ahmed Salem, Apratim Bhattacharya, Michael Backes, Mario Fritz, and Yang Zhang. {Updates-Leak}: Data set inference and reconstruction attacks in online learning. In *29th USENIX security symposium (USENIX Security 20)*, pages 1291–1308, 2020.

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

[Smart, 1980] David Roger Smart. *Fixed point theorems*, volume 66. Cup Archive, 1980.

[Song and Mittal, 2021] Liwei Song and Prateek Mittal. Systematic evaluation of privacy risks of machine learning models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2615–2632, 2021.

[Song *et al.*, 2020a] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2020.

[Song *et al.*, 2020b] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2020.

[Tang *et al.*, 2023] Shuai Tang, Zhiwei Steven Wu, Sergul Aydore, Michael Kearns, and Aaron Roth. Membership inference attacks on diffusion models via quantile regression. *arXiv preprint arXiv:2312.05140*, 2023.

[Truex *et al.*, 2019] Stacey Truex, Ling Liu, Mehmet Emre Gursoy, Lei Yu, and Wenqi Wei. Demystifying membership inference attacks in machine learning as a service. *IEEE Transactions on Services Computing*, 14(6):2073–2089, 2019.

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

[Yu *et al.*, 2015] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.