Are Diffusion Models Vulnerable to Membership Inference Attacks?

Jinhao Duan [1] **Fei Kong** [2] **Shiqi Wang** [3] **Xiaoshuang Shi** [2] **Kaidi Xu** [1]

Abstract

Diffusion-based generative models have shown great potential for image synthesis, but there is a lack of research on the security and privacy risks they may pose. In this paper, we investigate the vulnerability of diffusion models to Membership Inference Attacks (MIAs), a common privacy concern. Our results indicate that existing MIAs designed for GANs or VAE are largely ineffective on diffusion models, either due to inapplicable scenarios (e.g., requiring the discriminator of GANs) or inappropriate assumptions (e.g., closer distances between synthetic samples and member samples). To address this gap, we propose *Step-wise Error Comparing Membership Inference* (SecMI), a query-based MIA that infers memberships by assessing the matching of forward process posterior estimation at each timestep. SecMI follows the common overfitting assumption in MIA where member samples normally have smaller estimation errors, compared with hold-out samples. We consider both the standard diffusion models, e.g., DDPM, and the text-to-image diffusion models, e.g., Latent Diffusion Models and Stable Diffusion. Experimental results demonstrate that our methods precisely infer the membership with high confidence on both of the two scenarios across multiple different datasets. Code is available at `https://github.com/jinhaoduan/SecMI`.

1. Introduction

Recently, diffusion models (Song & Ermon, 2019; Song et al., 2020b; Ho et al., 2020) have dominated the image generation fields. Large-scale diffusion models, such as Stable Diffusion (Rombach et al., 2022), DALLE-2 (Ramesh et al., 2022), Imagen (Saharia et al., 2022), achieve remarkable fidelity and utility in text-to-image generation. Images generated by these models are full of diversity and creativity, which may profoundly change the way of human creation.

However, tremendous privacy risks (Bommasani et al., 2021) and copyright disputes (Ihalainen, 2018) are emerging, with the deployment of these generative models. For example, previous works demonstrate that GANs (Goodfellow et al., 2020)/VAEs (Kingma & Welling, 2013)-based generative models suffer from privacy leaking (Hayes et al., 2017) and data reconstruction (Zhang et al., 2020) issues, yet we are still unaware of whether diffusion models have similar concerns. On the other side, artists are getting unions against commercial large-scale generative models recently (BBC, 2022; CNN, 2022; WashingtonPost, 2022), since the creations of human artists may be exploited unauthorizedly. Lots of security assessments and risk screening need to be addressed before releasing these models.

Membership Inference Attacks (MIAs) (Shokri et al., 2016) is one of the most common privacy risks that associated with various privacy concerns. For a given pre-trained model, MIAs aim to identify the membership of a single sample as *member sample* that comes from the member set (training set), or as *hold-out sample* that comes from the hold-out set. Although MIAs have been well explored in the classic classification models and conventional generative models, related works on diffusion models are still missing in the literature. In this paper, we intend to answer the following question:

Are diffusion-based generative models vulnerable to MIAs?

We first summarize existing MIAs designed for conventional generative models and evaluate their performances in the diffusion model setting. Specifically, we summarize five MIAs and make them available to the diffusion models by either training shadow models (Shokri et al., 2017) or providing the required information in an alternative manner. Our results indicate that these methods are largely ineffective for diffusion models due to various potential reasons, such as more standard and stronger evaluations with larger member sets, limited exploitation of the specific properties of diffusion models, and inappropriate assumptions, i.e., synthetic images are closer to member images (Hu & Pang, 2021; Chen et al., 2020; Mukherjee et al., 2021). The lack-

[1]Drexel University [2]University of Electronic Science and Technology of China [3]AWS AI Lab. Correspondence to: Kaidi Xu <kx46@drexel.edu>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

ing of effective MIAs may cause a false sense of security for diffusion models.

To mitigate this gap, we consider designing MIAs by leveraging the specific properties of diffusion models. Our motivation derives from the learning objectives where diffusion models are trained to match the forward process posterior distribution at each timestep with a parameterized model ϵ_θ (Ho et al., 2020). Recall that it is commonly known that membership privacy leaking benefits from the overfitting issue (Shokri et al., 2017; Yeom et al., 2018a): member samples normally are memorized "better" than hold-out samples. Therefore a natural assumption regarding the membership exposure to the diffusion models is that member samples may have smaller posterior estimation errors compared with hold-out samples.

Based on that, we propose Step-wise Error Comparing Membership Inference (SecMI) to investigate the privacy leaking of diffusion models. SecMI is a query-based MIA that only relies on the inference results and can be applied to various diffusion models. We study three popular diffusion models: DDPM (Ho et al., 2020), Latent Diffusion Models (LDMs) (Rombach et al., 2022), and Stable Diffusion, across multiple popular datasets, including CIFAR-10/100 (Krizhevsky et al., 2009), STL10-Unlabeled (STL10-U) (Coates et al., 2011), Tiny-ImageNet (Tiny-IN) for DDPM, Pokemon (Pinkney, 2022) and COCO2017-val (Lin et al., 2014) for LDMs, and Laion (Schuhmann et al., 2022) for Stable Diffusion. Our contributions can be summarized as the following:

- To the best of our knowledge, this is the first work that investigates the vulnerability of diffusion models on MIAs. We summarize conventional MIAs and evaluate their performances on diffusion models. Results show that they are largely ineffective.
- We propose SecMI, a query-based MIA relying on the error comparison of the forward process posterior estimation. We apply SecMI on both the standard diffusion model, e.g., DDPM, and the state-of-the-art text-to-image diffusion model, e.g., Stable Diffusion.
- We evaluate SecMI across multiple datasets and report attack performance, including the True-Positive Rate (TPR) at low False-Positive Rate (FPR) (Carlini et al., 2022). Experimental results show that SecMI precisely infers the membership across all the experiment settings (≥ 0.80 avg. Attack Success Rate (ASR) and ≥ 0.85 avg. Area Under Receiver Operating Characteristic (AUC)).

2. Related Works

Generative Diffusion Models Different from Generative Adversarial Networks (GAN) (Goodfellow et al., 2020; Yuan & Moghaddam, 2020; Yuan et al., 2023), diffusion models refer to specific latent variable models that approximate the real data distribution by matching a diffusion process with a parameterized reverse process (Sohl-Dickstein et al., 2015; Ho et al., 2020). The diffusion process and reverse process can be either the continuous Langevin dynamics (Song & Ermon, 2019) or the discrete Markov chains (Ho et al., 2020), which are proven to be equivalent to the variance-preserving (VP) SDE in (Song et al., 2020b). Various diffusion models are proposed for speeding up inference (Song et al., 2020a; Salimans & Ho, 2022; Dockhorn et al., 2022; Xiao et al., 2022; Watson et al., 2022; Rombach et al., 2022; Meng et al., 2021), conditional generation (Dhariwal & Nichol, 2021; Ho & Salimans, 2022; Meng et al., 2021), and multi-modality generative tasks, such as audio synthesis (Kong et al., 2020).

Membership Inference Privacy Membership Inference Attack (Shokri et al., 2016) has been widely explored for classification models. MIAs can be recognized as black-box attacks (Shokri et al., 2017; Salem et al., 2019; Yeom et al., 2018b; Sablayrolles et al., 2019; Song & Mittal, 2020; Choquette-Choo et al., 2021; Hui et al., 2021; Truex et al., 2019; Salem et al., 2018; Pyrgelis et al., 2017) and white-box attacks (Nasr et al., 2019; Rezaei & Liu, 2020), depending on the accessibility to the target models. (Choquette-Choo et al., 2021) shows that logits are not necessary for MIA and proposes a label-only attack. (Carlini et al., 2022) reveals that MIA should be evaluated with strict metrics due to the fact that correctly and incorrectly inferring a membership are not equally important. (Sablayrolles et al., 2019) proves that black-box MIA can approximate white-box MIA performance under certain assumptions on the model weights distribution. In this paper, we mainly consider the query-based settings, i.e., only access to the query results of diffusion models.

Similarly, in terms of generative models, (Hayes et al., 2017) reveals that memberships can be effectively identified by the logits of the discriminator from GANs. (Hilprecht et al., 2019) proposes the Monte Carlo score for black-box MIA, and the Reconstruction Attack for VAE by leveraging the reconstruction loss item. The Monte Carlo score measures the distance between synthetic samples and member samples, which is further adopted by several works (Hu & Pang, 2021; Chen et al., 2020; Mukherjee et al., 2021).

Concurrently, (Hu & Pang, 2023; Wu et al., 2022; Carlini et al., 2023) also investigate the MIA issues in diffusion models. (Wu et al., 2022) assumes that the member set and the hold-out set come from different distributions, which makes their MIAs much easier. Our paper follows the common "MI security game protocol" (Hu & Pang, 2023; Carlini et al., 2023) where the member set and the hold-out set are in the same distribution. (Hu & Pang, 2023) infers mem-

Table 1: Taxonomy of MIAs against generative models over the previous works and our work. All the methods are evaluated on DDPM trained with a 50% training split of the CIFAR-10 dataset as the member set and take the rest of the training split as the hold-out set. DMs stands for Diffusion Models.

Attack Type	Method	Discriminator	Generator	Synthetic	Applicable to DMs	ASR to DMs ↑
-	Random Guess	✗	✗	✗	YES	0.500
White-box	GAN-Leaks (white-box) (Chen et al., 2020)	✗	□	✗	NO	0.615†
Black-box	Shadow Model + LOGAN	✗	✗	✔	YES	0.544
	Shadow Model + TVD	✗	✗	✔	YES	0.089‡
	Over-Representation (Hu & Pang, 2021)	✗	✗	✔	YES	0.532
	Monte-Carlo Set (Hilprecht et al., 2019)	✗	✗	✔	YES	0.505
	GAN-Leaks (black-box) (Chen et al., 2020)	✗	✗	✔	YES	0.507
Query-based	LOGAN (Hayes et al., 2017)	■	✗	✗	NO	-
	TVD (Mukherjee et al., 2021)	■	✗	✗	NO	-
	SecMI_{stat} (ours)	✗	■	✗	YES	0.811
	SecMI_{NNs} (ours)	✗	■	✗	YES	0.888

✔: with access ✗: without access □: white-box access ■: black-box access

†: GAN-leaks (white-box) is computationally infeasible for diffusion models. Here is the theoretical performance upper-bound of GAN-leaks, by generating the precise latent code of generated data through DDIM.

‡: (Mukherjee et al., 2021) measures the attack efficiency by calculating the upper bound Total Variation Distance (TVD $\in [0, 1]$, a higher value means better attack performance).

bership by comparing the loss values of member samples and hold-out samples. (Carlini et al., 2023) discloses that training data extraction and MIAs are both feasible for diffusion models. They share a similar loss-based MIA idea with (Hu & Pang, 2023) and further incorporate with the powerful LiRA (Carlini et al., 2022) to improve the attack performance. Different from that, our method identifies membership by assessing the posterior estimation with deterministic sampling and reversing, which achieves much stronger performances with the simple threshold inference strategy.

3. Preliminary Analysis

In this section, we formally define the membership inference problem and investigate whether existing MIAs designed for GANs or VAE work for diffusion models.

3.1. Problem Statement

Membership Inference (MI) aims to predict whether or not a specific sample is used as a training sample. Given a model f_θ parameterized by weights θ and dataset $D = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n\}$ drawn from data distribution q_{data}, we follow the common assumption (Sablayrolles et al., 2019; Carlini et al., 2022) where D is split into two subsets, D_M and D_H, and $D = D_M \cup D_H$. f_θ is solely trained on D_M. In this case, D_M is the *member set* of f_θ and D_H is the *hold-out set*. Each sample $\boldsymbol{x}_i$ is equipped with a membership identifier m_i, where $m_i = 1$ if $\boldsymbol{x}_i \sim D_M$; otherwise $m_i = 0$. The attacker only has access to D while having no knowledge about D_M and D_H. An attack algorithm $\mathcal{M}$ is designed to predict whether or not $\boldsymbol{x}_i$ is in D_M:

$$\mathcal{M}(\boldsymbol{x}_i, \theta) = \mathbb{1}\left[\mathbb{P}(m_i = 1|\theta, \boldsymbol{x}_i) \geq \tau\right] \quad (1)$$

where $\mathcal{M}(\boldsymbol{x}_i, \theta) = 1$ means x_i comes from D_M, $\mathbb{1}[A] = 1$ if A is true, and τ is the threshold. For the generative model scenario, we reuse θ as the weights of generator G and let $p_\theta(\boldsymbol{x})$ denote the generative distribution where generated sample $\boldsymbol{x} \sim p_\theta(\boldsymbol{x}|\boldsymbol{z})$ given latent code $\boldsymbol{z}$.

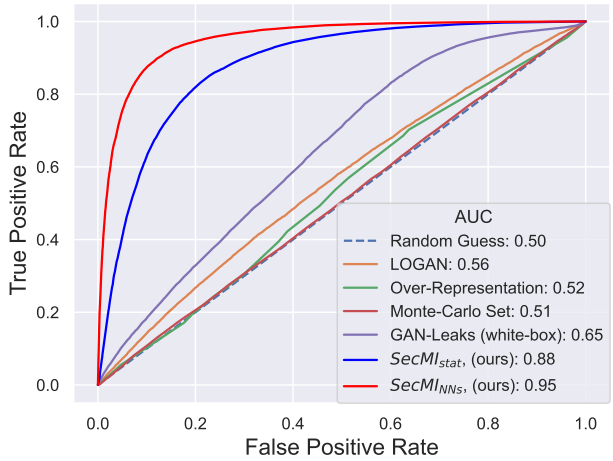


Figure 1: Comparing the TPR v.s. FPR of prior MIAs designed for generative models. Evaluations are conducted on DDPM with half of the CIFAR-10 training split as the member set and the other half as the hold-out set. Prior MIAs are largely ineffective on DDPM.

Evaluation Metrics. Following the most convincing met-

rics used in MIAs (Carlini et al., 2022; Choquette-Choo et al., 2021), we measure the performance of MIAs with Attack Success Rate (ASR), Area Under Receiver Operating Characteristic (AUC), and True-Positive Rate (TPR) at extremely low False-Positive Rate (FPR), e.g., TPR@1%/0.1% FPR.

3.2. Evaluating Existing MIAs on Diffusion Models

As summarized in Table 1, we consider five different MIAs designed for generative models: LOGAN (Hayes et al., 2017), TVD (Mukherjee et al., 2021), Over-Representation (Hu & Pang, 2021), Monte-Carlo Set (Hilprecht et al., 2019), and GAN-Leaks (Chen et al., 2020), as summarized in Table 1. Since LOGAN and TVD require access to the discriminator of GANs, we train Shadow Models (Shokri et al., 2017) to make them available to diffusion models. All the methods are evaluated on DDPM (Ho et al., 2020) trained over CIFAR-10 (Krizhevsky et al., 2009) with 50% training split as the member set and the rest as the hold-out set. Also, we choose StyleGAN2 (Karras et al., 2020) as the Shadow Model.

The ROC curves of these methods are presented in Figure 1. For black-box MIAs that only require the query results of the generator, only LOGAN shows marginal effectiveness on diffusion models while other MIAs are largely ineffective. The original white-box GAN-leaks requires the full gradient to optimize the latent code, which is computationally infeasible for diffusion models. Here we calculate its theoretical upper bound by providing the latent codes generated by DDIM (Song et al., 2020a). In this setting, the white-box GAN-leaks shows certain effectiveness.

3.3. Analytical Insights

We summarize the potential reasons for the limited success of existing methods in diffusion models:

Stronger Evaluations. Previous works (Hayes et al., 2017; Chen et al., 2020; Hilprecht et al., 2019; Hu & Pang, 2023) adopt very limited member set size (e.g., $\leq 10\%$ of the training split of D) while we employ a size of 50%. It is known that MIAs benefit from overfitting (Shokri et al., 2017; Yeom et al., 2018a). Smaller member set may exacerbate overfitting, which amplifies the effect of these methods.

Diffusion Models Generalize Better. Prior works mainly assume that a sample that "occurred" in a higher frequency when sampling from the generator, is more likely to be in the member set:

$$\mathbb{P}(m_i = 1|\theta_G, \boldsymbol{x}_i) \propto \mathbb{P}(\boldsymbol{x}_i|\theta_G). \tag{2}$$

However, this holds when the generative distribution overfits the member set, i.e., $d(p_\theta, p_{D_M}) < d(p_\theta, p_{D_H})$ where $d(\cdot,\cdot)$ is a distance measurement and $p_\theta, p_{D_M}, p_{D_H}$ are generative distribution, member distribution, and hold-out distribution, respectively. To measure it, we estimate $d(p_\theta, p_{D_M})$ and $d(p_\theta, p_{D_H})$ by calculating the FIDs (Heusel et al., 2017) between 25,000 synthetic images and 25,000 member/hold-out samples, and we get 9.66 v.s. 9.85, which shows diffusion models have no distinct bias toward the member samples.

Limited Exploitation of Diffusion Models. Existing MIAs were primarily designed for GANs or VAEs, and thus do not take into account the specific properties of diffusion models.

4. Methodology

In this section, we provide the first MIA design for diffusion models exploiting the step-wise forward process posterior estimation.

4.1. Notations

We follow the common notations of diffusion models (Ho et al., 2020) where we denote by $q(\boldsymbol{x}_0)$ the real data distribution and $p_\theta(\boldsymbol{x}_0)$ the latent variable model approximating $q(\boldsymbol{x}_0)$ with noise-prediction model ϵ_θ parameterized by weights θ. Diffusion models consist of the T-step diffusion process $q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})$ and the denoising process $p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t), (1 \leq t \leq T)$, with the following transitions:

$$\begin{aligned} q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) &= \mathcal{N}(\boldsymbol{x}_t; \sqrt{1-\beta_t}\boldsymbol{x}_{t-1}, \beta_t \mathbf{I}) \\ p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) &= \mathcal{N}(\boldsymbol{x}_{t-1}; \mu_\theta(\boldsymbol{x}_t, t), \Sigma_\theta(\boldsymbol{x}_t, t)) \end{aligned} \tag{3}$$

where $\beta_1, \cdots, \beta_T$ is a variance schedule. The forward sampling at arbitrary time step t can be obtained by

$$q(\boldsymbol{x}_t|\boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{x}_t; \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0, (1-\bar{\alpha}_t)\mathbf{I}), \tag{4}$$

where $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$.

4.2. Exposing Membership via Step-Wise Error Comparison

(Sablayrolles et al., 2019) demonstrates that the Bayes optimal performance of MIA can be approximated as

$$\mathcal{M}_{opt}(\boldsymbol{x}, \theta) = \mathbb{1}\left[\ell(\theta, \boldsymbol{x}) \leq \tau\right], \tag{5}$$

under a mild assumption on the model weights distribution, where τ is the threshold and ℓ is the loss function, i.e., the cross-entropy loss for classification tasks. Eq. (5) reveals that membership can be exposed by measuring how well θ is learned on data point $\boldsymbol{x}$.

For diffusion models, a similar metric can also be derived from the learning objective. Recall that diffusion models are trained to optimize the variational bound $p_\theta(\boldsymbol{x}_0)$ by matching the forward process posteriors at each step t:

$$\ell_t = \mathbb{E}_q\left[\frac{1}{2\sigma_t^2}||\tilde{\mu}_t(\boldsymbol{x}_t, \boldsymbol{x}_0) - \mu_\theta(\boldsymbol{x}_t, t)||^2\right], \tag{6}$$

where $\tilde{\mu}_t(\boldsymbol{x}_t, \boldsymbol{x}_0)$ is the mean of the posterior distribution $q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)$ and $\mu_\theta(\boldsymbol{x}_t, t)$ refers to the estimation. $||\cdot||^2$ refers to the mean squared error (MSE). Eq. (6) indicates that the local estimation error of single data point $\boldsymbol{x}_0$ at timestep t is

$$\ell_{t,\boldsymbol{x}_0} = ||\hat{\boldsymbol{x}}_{t-1} - \boldsymbol{x}_{t-1}||^2, \tag{7}$$

where $\boldsymbol{x}_{t-1} \sim q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)$ and $\hat{\boldsymbol{x}}_{t-1} \sim p_\theta(\hat{\boldsymbol{x}}_{t-1}|\boldsymbol{x}_t)$ (we omit $\frac{1}{2\sigma_t^2}$ since it is constant). A natural assumption regarding the membership exposure is that samples from the member set D_M may have smaller estimation errors at step t, compared with samples from the hold-out set D_H:

$$\ell_{t,\boldsymbol{x}_m} \leq \ell_{t,\boldsymbol{x}_h}, \ 1 \leq t \leq T, \tag{8}$$

where $\boldsymbol{x}_m \sim D_M$ and $\boldsymbol{x}_h \sim D_H$.

However, the above quantity is intractable since it involves $q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)$ and $p_\theta(\hat{\boldsymbol{x}}_{t-1}|\boldsymbol{x}_t)$. Due to the essence of non-deterministic diffusion and denoising processes, i.e., the Markov Chain, it is intractable to calculate their analytical solutions. Although estimating it with Monte Carlo sampling is possible, it is time-consuming for diffusion models requiring multi-step denoising, e.g., T = 1,000. Inspired by recent works on deterministic reversing and sampling from diffusion models (Song et al., 2020a; Kim et al., 2022; Song et al., 2020b), we consider approximate Eq. (7) with deterministic processes:

$$\begin{aligned} \boldsymbol{x}_{t+1} &= \phi_\theta(\boldsymbol{x}_t, t) \\ &= \sqrt{\bar{\alpha}_{t+1}} f_\theta(\boldsymbol{x}_t, t) + \sqrt{1 - \bar{\alpha}_{t+1}} \epsilon_\theta(\boldsymbol{x}_t, t), \end{aligned} \tag{9}$$

$$\begin{aligned} \boldsymbol{x}_{t-1} &= \psi_\theta(\boldsymbol{x}_t, t) \\ &= \sqrt{\bar{\alpha}_{t-1}} f_\theta(\boldsymbol{x}_t, t) + \sqrt{1 - \bar{\alpha}_{t-1}} \epsilon_\theta(\boldsymbol{x}_t, t), \end{aligned} \tag{10}$$

where

$$f_\theta(\boldsymbol{x}_t, t) = \frac{\boldsymbol{x}_t - \sqrt{1 - \bar{\alpha}_t} \epsilon_\theta(\boldsymbol{x}_t, t)}{\sqrt{\bar{\alpha}_t}}. \tag{11}$$

We denote by $\Phi_\theta(x_s, t)$ the deterministic reverse, i.e., from x_s to x_t ($s < t$), and $\Psi_\theta(x_t, s)$ the deterministic denoise process, i.e., from x_t to x_s:

$$\begin{aligned} \boldsymbol{x}_t &= \Phi_\theta(\boldsymbol{x}_s, t) = \phi_\theta(\cdots \phi_\theta(\phi_\theta(\boldsymbol{x}_s, s), s+1), t-1) \\ \boldsymbol{x}_s &= \Psi_\theta(\boldsymbol{x}_t, s) = \psi_\theta(\cdots \psi_\theta(\psi_\theta(\boldsymbol{x}_t, t), t-1), s+1) \end{aligned} \tag{12}$$

Then, we define *t-error* as the approximated posterior estimation error at step t.

Definition 4.1 (*t-error*)**.** For given sample $\boldsymbol{x}_0 \sim D$ and the deterministic reverse result $\tilde{\boldsymbol{x}}_t = \Phi_\theta(\boldsymbol{x}_0, t)$ at timestep t, the approximated posterior estimation error at step t is defined as *t-error*:

$$\tilde{\ell}_{t,\boldsymbol{x}_0} = ||\psi_\theta(\phi_\theta(\tilde{\boldsymbol{x}}_t, t), t) - \tilde{\boldsymbol{x}}_t||^2. \tag{13}$$

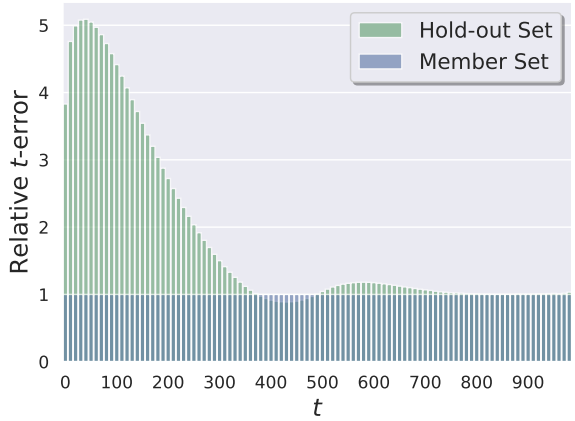


Figure 2: Comparisons of *t-errors* for samples from the member and hold-out sets. Since the magnitudes are different at each step t, we set the *t-errors* of member set as 1 at every timestep and report the relative sizes of *t-errors* of Hold-out Set. It is shown that samples from the hold-out set have higher *t-errors* compared with samples from the member set, proving that *t-error* is an effective metric for identifying memberships.

We prove that $\tilde{\ell}_{t,\boldsymbol{x}_0}$ is converged to $\ell_{t,\boldsymbol{x}_0}$ when the diffusion model is converged to the learning objective, i.e., $\tilde{\ell}_{t,\boldsymbol{x}_0} \to \ell_{t,\boldsymbol{x}_0}$, when $||\epsilon_\theta(\boldsymbol{x}_t, t) - \epsilon||^2 \to 0$ at any timestep t, where $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ (the "noise-prediction" loss in (Ho et al., 2020):

$$\begin{aligned} \Delta_{t,\boldsymbol{x}_0} &= \ell_{t,\boldsymbol{x}_0} - \tilde{\ell}_{t,\boldsymbol{x}_0} \\ &= ||\epsilon_\theta(\boldsymbol{x}_t, t) - \epsilon||^2 \\ &- ||\sqrt{1 - \bar{\alpha}_t}(\epsilon_\theta(\tilde{\boldsymbol{x}}_t, t) - \epsilon_\theta(\phi_\theta(\tilde{\boldsymbol{x}}_t, t), t+1))||^2. \end{aligned} \tag{14}$$

It is shown that $\Delta_{t,\boldsymbol{x}_0} \to 0$ when $||\epsilon_\theta(\boldsymbol{x}_t, t) - \epsilon||^2 \to 0$ at any timestep t.

Figure 2 presents the relative scales of *t-error* at each timestep for D_M and D_H. We show that samples from the hold-out set normally have larger *t-errors* compared with member samples, which verifies the membership exposure assumption in Eq. (8). Besides, this phenomenon is getting distinct as t approaches 0. One of the possible reasons is that $\boldsymbol{x}_T$ is an approximated Gaussian Noise that contains almost zero information about the member sample while $\boldsymbol{x}_0$ is the member sample. A smaller t triggers more "memory" about member samples, which results in serious privacy leaking.

4.3. Error Comparing Based Membership Inference

Our analysis indicates that comparing the approximated step-wise posterior estimation error, i.e., *t-error*, is effective for identifying memberships. In this section, we design two strategies to infer membership by leveraging *t-error*: the statistic-based inference, SecMI$_{stat}$, and the neu-

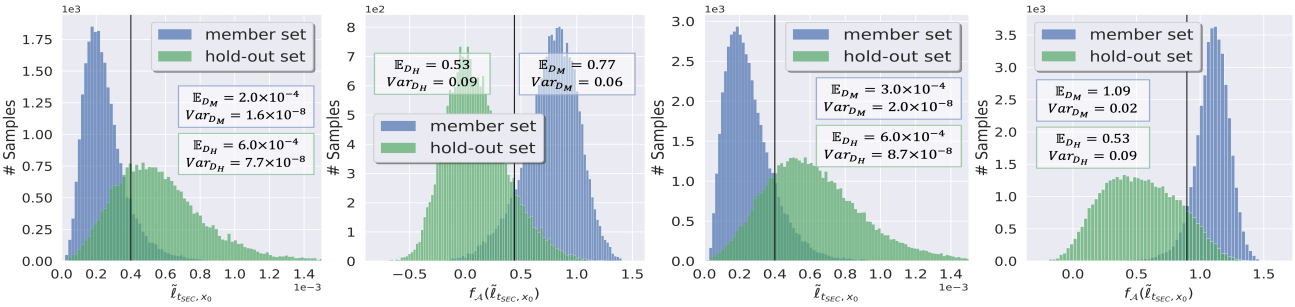
Table 2: Performance of SecMI on DDPM across four datasets. It is shown that both of the two variants achieve significant performances in all the settings.

Method	θ	# Query	CIFAR-10		CIFAR-100		STL10-U		Tiny-IN		Average	
			ASR↑	AUC↑	ASR↑	AUC↑	ASR↑	AUC↑	ASR↑	AUC↑	ASR↑	AUC↑
GAN-Leaks†	□	$\geq 1000 \times 2$	0.615	0.646	0.513	0.459	0.566	0.535	0.545	0.457	0.560	0.524
SecMI$_{stat}$	■	$10+2$	0.811	0.881	0.798	0.868	0.809	0.881	0.821	0.894	0.810	0.881
SecMI$_{NNs}$ ‡	■	$10+2$	0.888	0.951	0.872	0.940	0.892	0.950	0.903	0.956	0.889	0.949

□: white-box access ■: black-box access

†: GAN-Leaks is computationally infeasible for diffusion models. Here are the theoretical results by providing the exact latent codes.

‡: We are aware that it is unfair to directly compare SecMI$_{NNs}$ with other methods since it is evaluated on a slightly smaller member set and hold-out set. Here just to show the effectiveness of our method.



(a) SecMI$_{stat}$ on CIFAR10. (b) SecMI$_{NNs}$ on CIFAR10. (c) SecMI$_{stat}$ on Tiny-IN. (d) SecMI$_{NNs}$ on Tiny-IN.

Figure 3: The $\tilde{\ell}_{t_{\text{SEC}},x_0}$ and $f_{\mathcal{A}}(\tilde{\ell}_{t_{\text{SEC}},x_0})$ distributions for samples from member set and hold-out set. It is clear that *t-error* is a desirable indicator for membership identification. The vertical black line refers to the selected threshold τ for each figure.

ral networks-based inference, SecMI$_{NNs}$.

We denote by t_{SEC} the selected timestep for error comparing. For each sample $\boldsymbol{x}_0 \sim D$, we calculate the corresponding *t-error*, $\tilde{\ell}_{t_{\text{SEC}},\boldsymbol{x}_0}$, based on Eq. (13). For SecMI$_{stat}$, we predict membership as the following:

$$\mathcal{M}(x_0, \theta) = \mathbb{1}\left[\tilde{\ell}_{t_{\text{SEC}},\boldsymbol{x}_0} \leq \tau\right], \tag{15}$$

where τ is the threshold. For SecMI$_{NNs}$, we adopt an attack model $f_{\mathcal{A}}$ to infer membership. Specifically, $f_{\mathcal{A}}$ takes the pixel-wise absolute value of estimation error as the input and predicts the confidence of being a member sample. We use a 1-output classification model as the backbone of $f_{\mathcal{A}}$ and train it in a binary classification manner. We randomly sample a small subset of D_M and D_H as its training data. The membership is predicted as the following:

$$\mathcal{M}(\boldsymbol{x}_0, \theta) = \mathbb{1}\left[f_{\mathcal{A}}(|\psi_\theta(\phi_\theta(\tilde{\boldsymbol{x}}_t, t), t) - \tilde{\boldsymbol{x}}_t|) \leq \tau\right]. \tag{16}$$

It is worth noting that although we only adopt the estimation error at a single timestep when predicting membership, experimental results show that this is already effective. We may leave how to fuse multi-step errors for better membership inference in the future.

4.4. Generalization to Various Diffusion Models

SecMI can be easily adapted to other popular diffusion models. To prove that, we provide the adaption to Latent Diffusion Models. We leave the implementation details in Appendix A.

5. Experiments

In this section, we evaluate the performance of SecMI across various datasets and settings.

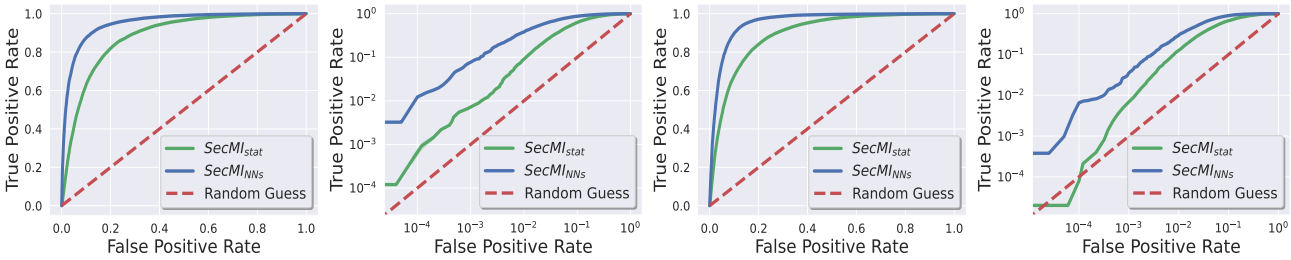
5.1. Experimental Setup

Datasets and Diffusion Models. For all the datasets, we randomly select 50% of the training samples as D_M and use the rest of the training samples as D_H. For example, CIFAR-10 contains 50,000 images in the training set, so we have 25,000 images for D_M and another 25,000 images for D_H. We summarize the adopted diffusion models and datasets in Appendix B. In addition to the text-to-image settings, we adopt the image captions provided by the dataset organizers as the prompts.

Implementation Details. We train DDPM from scratch with the default hyper-parameters, except for the data split-

Table 3: The TPR at low FPR of SecMI on DDPM over four datasets.

Methods	TPR @ 1% FPR (%) ↑				TPR @ 0.1% FPR (%) ↑			
	CIFAR-10	CIFAR-100	STL10-U	Tiny-IN	CIFAR-10	CIFAR-100	STL10-U	Tiny-IN
GAN-Leaks	2.80	1.85	1.17	1.01	0.29	0.23	0.24	0.13
SecMI_{stat}	9.11	9.26	10.87	12.67	0.66	0.46	0.73	0.69
SecMI_{NNs}	37.98	30.17	26.66	29.77	7.59	5.09	3.76	3.50



(a) ROC on CIFAR10. (b) Log-scaled ROC on CIFAR10. (c) ROC on Tiny-IN. (d) Log-scaled ROC on Tiny-IN.

Figure 4: ROC curves of SecMI on CIFAR10 and Tiny-IN datasets. The overall ROC curves show that our methods are largely effective on diffusion models. The log-scaled ROC curves indicate that our methods are capable of generating high-confidence predictions.

tings. For text-to-image experiments, we adopt the HuggingFace pre-trained Stable Diffusion as the victim diffusion models. More implementation details can be found in Section 5.5 and Section 5.6. We also adopt DDIM (k) to speed up the inference, where k refers to the inference interval, e.g., DDIM (10) consisting of 100 denoising steps when T=1,000. We set t_{SEC} to 100 for all the experiments. For attack model $f_{\mathcal{A}}$, we choose ResNet-18 as the backbone and adopt 20% of D_M and D_H as its training samples. $f_{\mathcal{A}}$ is trained in 15 epochs with a learning rate of 0.001 and batch size of 128.

5.2. Comparison to Baselines

We take GAN-leaks as the baseline method. We train DDPM on four datasets, including CIFAR-10/100, STL10-U, Tiny-IN, and summarize AUCs and ASRs in Table 2. It is shown that SecMI accurately infers most of the memberships for samples from these datasets, i.e., 81.0% and 88.9% average ASRs for SecMI_{stat} and SecMI_{NNs}. Compared with naive statistical inference, training a neural network as the inference strategy can significantly improve performance (SecMI_{NNs} outperforms SecMI_{stat} by over 7%.). In Figure 3, we show that *t-error* is a qualified extractor to distinguish member samples and hold-out samples.

As pointed out by (Carlini et al., 2022), the risk of correctly inferring membership is greater than that of being inferred incorrectly for some scenarios, e.g., medical data. Therefore, we also consider the TPR at very low FPR, e.g., 1% FPR and 0.1% FPR in Table 3. It is shown that SecMI achieves notable TPR in both of the two evaluations, which proves its effectiveness. We present the overall ROC curves and the log-scaled ROC curves in Figure 4.

5.3. Ablation Study

We study how hyper-parameters, such as t_{SEC}, DDIM inference interval k, and the distance measurement $d(\cdot,\cdot)$ used in $\tilde{\ell}_{t,\boldsymbol{x}}$, affect the performance of SecMI. Besides, since Weights Averaging (WA) is one of the most common techniques in training diffusion models, we also analyze how WA affects privacy leaking. For each experiment, we run 5 trials with random data splittings. Results are summarized in Figure 5 and Figure 6.

Generally, we show that our algorithm is stable with limited variance, e.g., ≤ 0.05 for both AUC and ASR over all trials. For timestep t_{SEC}, although we determine t_{SEC} empirically, we show that the attack performances are not sensitive to specific timesteps. As long as $50 \leq t_{\text{SEC}} \leq 150$, the attack will be effective. This property also generalizes well to other experiments. k affects the query efficiency and we show that SecMI still achieves remarkable adversary even with only 3 queries. In terms of WA, an interesting phenomenon is that WA brings a certain degree of privacy leaking, which alarms the community to carefully select the training protocols. For distance measurement, the sensitivity to the estimation error also affects the performance of our methods.

5.4. SecMI v.s. Defense

Recall that MIAs primarily benefit from overfitting (Shokri et al., 2016; Yeom et al., 2018b; Salem et al., 2018). We

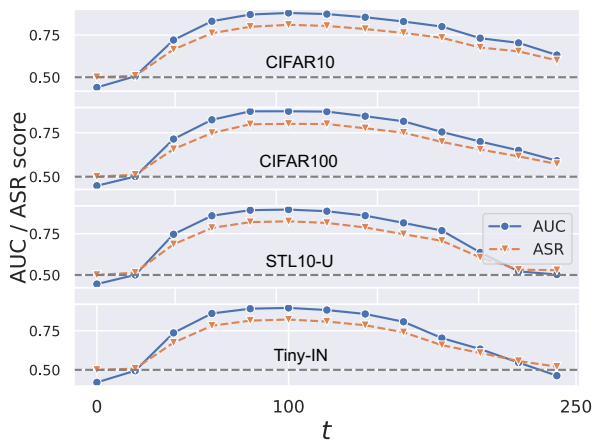


Figure 5: AUC and ASR of SecMI$_{stat}$ v.s. timestep, among four datasets. The attack performances are stable and not sensitive to the selection of timestep t_{SEC}.

Figure 6: Ablation studies of hyper-parameters and the sensitivity to the randomness. Models are evaluated by SecMI$_{stat}$ on the CIFAR-10 dataset.

Table 4: Evaluations in resisting data augmentations.

Method	SecMI$_{stat}$		SecMI$_{NNs}$	
	ASR↑	AUC↑	ASR↑	AUC↑
No Augmentation	0.964	0.912	0.972	0.928
w/ RandomHorizontalFlip†	0.811	0.881	0.888	0.951
w/ Cutout (DeVries & Taylor, 2017)	0.961	0.908	0.992	0.973
w/ RandAugment (Cubuk et al., 2020)‡	0.533	0.530	-	-

† RandomHorizontalFlip is the default augmentation for diffusion models.
‡ RandAugment makes diffusion models fail to converge.

investigate how data augmentation, one of the most popular methods in resisting overfitting, affects the performance of SecMI. Specifically, we study Cutout (DeVries & Taylor, 2017), RandomHorizontalFlip. Results are summarized in Table 4. It is shown that the attack performance increased significantly when no augmentation is applied to the diffusion model. After applying mild data augmentation, e.g., Cutout and RandomHorizontalFlip, the ASR and AUC get decreased to a certain degree.

We also try to examine SecMI with stronger privacy-preserving methods and training tricks, including DP-SGD (Abadi et al., 2016), ℓ_2 regularization, and stronger data augmentation such as RandAugment (Cubuk et al., 2020). We follow the same training settings as before but solely apply these techniques during model training. Experimental results show that DDPM training with these strong defense methods even failed to converge. For instance, with ℓ_2 regularization, the trained DDPM can only generate random and meaningless information. Similarly, applying RandAugment will make the generated data full of distortions, which makes it valueless to evaluate our methods. Some generated images under the above settings are provided in Appendix C.

5.5. Evaluations on Latent Diffusion Models (LDMs)

We conduct experiments on text-to-image diffusion models in this section. Specifically, we adopt the Huggingface pre-trained stable-diffusion-v1-4 as the backbone and fine-tune 15,000 and 150,000 steps over Pokemon and COCO2017-Val, respectively. Both two datasets are not considered during the Stable Diffusion pre-training phase. We study the sensitivity to ground-truth prompts, i.e., whether adopting empty prompts or prompts generated from other sources (e.g., BLIP (Li et al., 2022)) during the attack will affect performances.

Results are summarized in Table 5. It is shown that our method achieves superior attack performances on both two datasets, which proves that SecMI generalizes well to the sophisticated diffusion models. In terms of the sensitivity to ground-truth prompts, we show that the actual influences vary among different tasks. For the Pokemon dataset, only a marginal drop (around 0.03) can be observed when equipped with empty prompts. We believe this is because the Pokrmon dataset has highly centralized images and text (they are all the characters of PokemonGo with a very uniform style), which makes the generation less sensitive to the prompt.

In contrast, we show that empty prompts will largely affect (around 0.15) the attack performance over the COCO2017-val set. However, with pseudo-prompts generated by BLIP, this drop can be almost mitigated (around 0.05). We believe it is because of the diversity and variety of the COCO2017-val images (i.e., diverse backgrounds, objects, and events), which makes the generation more sensitive to the prompt. This indicates that our SecMI is applicable when the attacker can't get access to the ground-truth prompts.

We also provide visualization results in Figure 7, by directly

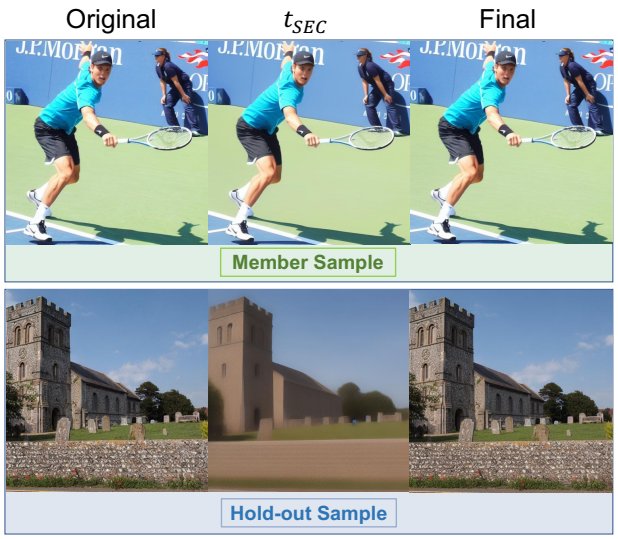


Figure 7: Visualization results of image reconstruction from a member sample and a hold-out sample at the t_{SEC}-th step and the final step.

reconstructing the reverse results at the t_{SEC}-th step and the final step, to show the qualitative differences between member samples and hold-out samples.

5.6. Evaluations on Stable Diffusion

We conduct experiments on the original Stable Diffusion, i.e., stable-diffusion-v1-4 and v1-5 provided by Huggingface, without further fine-tuning or other modifications. Specifically, we randomly sample 2500 images from Laion-aesthetic images with aesthetic scores greater than 5 (Laion-aesthetic-5plus) as the member set and randomly sample 2500 images from COCO2017-val as the hold-out set.

As shown in Table 6, our method achieves notable attack performances on both stable-diffusion-v1-4 and v1-5. It indicates that SecMI is still effective for large-scale pre-training scenarios, which sheds light on applying our method in the real world.

6. Conclusion

In this paper, we investigate whether diffusion models are vulnerable to MIAs. Specifically, we first summarize existing MIAs and show that most of them are largely ineffective for diffusion models. Then, to mitigate this gap, we propose Step-wise Error Comparing Membership Inference (SecMI), based on the step-wise posterior matching of diffusion models. We evaluate SecMI on both standard diffusion models, e.g., DDPM, and state-of-the-art text-to-image models, e.g., Stable Diffusion. Experimental results across multiple datasets demonstrate the effectiveness of our method.

Table 5: Evaluations on LDMs. Models are evaluated by SecMI$_{stat}$.

Dataset	Prompt	ASR↑	AUC↑	TPR@1%FPR (%) ↑
Pokemon	Ground-truth	0.821	0.891	7.20
	Empty	0.782	0.860	11.06
COCO2017-val	Ground-truth	0.803	0.875	13.98
	Empty	0.663	0.720	6.04
	BLIP (Li et al., 2022)	0.750	0.820	9.40

Table 6: Evaluations on Stable Diffusion. Models are evaluated by SecMI$_{stat}$.

Victim Model	AUC	ASR	TPR@1%FPR (%)
stable-diffusion-v1-4†	0.707	0.664	18.47
stable-diffusion-v1-5	0.701	0.661	18.58

† stable-diffusion-v1-4 is trained in less than 1 epoch over the Laion-aesthetic-5plus dataset. The real attack performance should be scaled.

Our research reveals that current diffusion models suffer from serious privacy issues. We hope that our work can inspire the community and encourage more privacy considerations.

Limitations

Despite SecMI showing great attack performances on various diffusion models and datasets, there are some limitations when applied in the physical world: 1) diffusion models are normally provided as black-box API services and it is less possible the attacker will access the intermediate results of the victim diffusion models, while SecMI requires the access to these results; 2) SecMI is only evaluated on public data, the effectiveness and sensitivity to a given demographic (or a subgroup of a dataset) is not investigated yet; 3) although MIA is one of the most common privacy concern in academia, the scope of MIA is limited in the real world.

Ethics and Broader Impacts

This paper proposes a membership inference attack algorithm, which is a threat to privacy for current diffusion models. To mitigate any possible abuse caused by this paper, all the experiments are conducted on public datasets and common model architectures. Figures presented in this paper are licensed under the Creative Commons 4.0 License, which is allowed to be distributed.

Despite there are limitations when applying our method in the physical world, it is still possible that our paper will cause privacy risks. However, we believe our paper acts more like an alert to the generative model community. We encourage more privacy and security considerations before releasing diffusion models to the public.

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pp. 308–318, 2016.

BBC. "Art is dead Dude" - the rise of the AI artists stirs debate. 2022. URL https://www.bbc.com/news/technology-62788725.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Carlini, N., Chien, S., Nasr, M., Song, S., Terzis, A., and Tramer, F. Membership inference attacks from first principles. In *2022 IEEE Symposium on Security and Privacy (SP)*, pp. 1897–1914. IEEE, 2022.

Carlini, N., Hayes, J., Nasr, M., Jagielski, M., Sehwag, V., Tramèr, F., Balle, B., Ippolito, D., and Wallace, E. Extracting training data from diffusion models. *arXiv preprint arXiv:2301.13188*, 2023.

Chen, D., Yu, N., Zhang, Y., and Fritz, M. Gan-leaks: A taxonomy of membership inference attacks against generative models. In *Proceedings of the 2020 ACM SIGSAC conference on computer and communications security*, pp. 343–362, 2020.

Choquette-Choo, C. A., Tramer, F., Carlini, N., and Papernot, N. Label-only membership inference attacks. In *International conference on machine learning*, pp. 1964–1974. PMLR, 2021.

CNN. AI won an art contest, and artists are furious. 2022. URL https://www.cnn.com/2022/09/03/tech/ai-art-fair-winner-controversy/index.html.

Coates, A., Ng, A., and Lee, H. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 215–223. JMLR Workshop and Conference Proceedings, 2011.

Cubuk, E. D., Zoph, B., Shlens, J., and Le, Q. V. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 702–703, 2020.

DeVries, T. and Taylor, G. W. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.

Dhariwal, P. and Nichol, A. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.

Dockhorn, T., Vahdat, A., and Kreis, K. Score-Based Generative Modeling with Critically-Damped Langevin Diffusion. In *International Conference on Learning Representations*, 2022.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

Hayes, J., Melis, L., Danezis, G., and De Cristofaro, E. Logan: Membership inference attacks against generative models. *arXiv preprint arXiv:1705.07663*, 2017.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Hilprecht, B., Härterich, M., and Bernau, D. Monte carlo and reconstruction membership inference attacks against generative models. *Proc. Priv. Enhancing Technol.*, 2019 (4):232–249, 2019.

Ho, J. and Salimans, T. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Hu, H. and Pang, J. Membership inference attacks against gans by leveraging over-representation regions. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, pp. 2387–2389, 2021.

Hu, H. and Pang, J. Membership inference of diffusion models. *arXiv preprint arXiv:2301.09956*, 2023.

Hui, B., Yang, Y., Yuan, H., Burlina, P., Gong, N. Z., and Cao, Y. Practical blind membership inference attack via differential comparisons. *arXiv preprint arXiv:2101.01341*, 2021.

Ihalainen, J. Computer creativity: Artificial intelligence and copyright. *Journal of Intellectual Property Law & Practice*, 13(9):724–728, 2018.

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., and Aila, T. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8110–8119, 2020.

Kim, G., Kwon, T., and Ye, J. C. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2426–2435, 2022.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Kong, Z., Ping, W., Huang, J., Zhao, K., and Catanzaro, B. Diffwave: A versatile diffusion model for audio synthesis. *arXiv preprint arXiv:2009.09761*, 2020.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

Li, J., Li, D., Xiong, C., and Hoi, S. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pp. 12888–12900. PMLR, 2022.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, 2014.

Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.-Y., and Ermon, S. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2021.

Mukherjee, S., Xu, Y., Trivedi, A., Patowary, N., and Ferres, J. L. privgan: Protecting gans from membership inference attacks at low cost to utility. *Proc. Priv. Enhancing Technol.*, 2021(3):142–163, 2021.

Nasr, M., Shokri, R., and Houmansadr, A. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In *IEEE symposium on security and privacy (SP)*. IEEE, 2019.

Pinkney, J. N. M. Pokemon blip captions. https://huggingface.co/datasets/lambdalabs/pokemon-blip-captions/, 2022.

Pyrgelis, A., Troncoso, C., and De Cristofaro, E. Knock knock, who's there? membership inference on aggregate location data. *arXiv preprint arXiv:1708.06145*, 2017.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

Rezaei, S. and Liu, X. Towards the infeasibility of membership inference on deep models. *arXiv preprint arXiv:2005.13702*, 2020.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Sablayrolles, A., Douze, M., Schmid, C., Ollivier, Y., and Jégou, H. White-box vs black-box: Bayes optimal strategies for membership inference. In *International Conference on Machine Learning*, pp. 5558–5567. PMLR, 2019.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S. K. S., Ayan, B. K., Mahdavi, S. S., Lopes, R. G., et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022.

Salem, A., Zhang, Y., Humbert, M., Berrang, P., Fritz, M., and Backes, M. Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models, 2018.

Salem, A., Zhang, Y., Humbert, M., Fritz, M., and Backes, M. Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models. In *Annual Network and Distributed System Security Symposium (NDSS)*, 2019.

Salimans, T. and Ho, J. Progressive Distillation for Fast Sampling of Diffusion Models. In *International Conference on Learning Representations*, 2022.

Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*, 2022.

Shokri, R., Stronati, M., Song, C., and Shmatikov, V. Membership inference attacks against machine learning models, 2016.

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.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pp. 2256–2265. PMLR, 2015.

Song, J., Meng, C., and Ermon, S. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020a.

Song, L. and Mittal, P. Systematic evaluation of privacy risks of machine learning models. *arXiv preprint arXiv:2003.10595*, 2020.

Song, Y. and Ermon, S. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020b.

Truex, S., Liu, L., Gursoy, M. E., Yu, L., and Wei, W. Demystifying membership inference attacks in machine learning as a service. *IEEE Transactions on Services Computing*, 2019.

WashingtonPost. He made a children's book using AI. Then came the rage. 2022. URL https://www.washingtonpost.com/technology/2023/01/19/ai-childrens-\book-controversy-chatgpt-midjourney/.

Watson, D., Chan, W., Ho, J., and Norouzi, M. Learning Fast Samplers for Diffusion Models by Differentiating Through Sample Quality. In *International Conference on Learning Representations*, 2022.

Wu, Y., Yu, N., Li, Z., Backes, M., and Zhang, Y. Membership inference attacks against text-to-image generation models. *arXiv preprint arXiv:2210.00968*, 2022.

Xiao, Z., Kreis, K., and Vahdat, A. Tackling the Generative Learning Trilemma with Denoising Diffusion GANs. In *International Conference on Learning Representations*, 2022.

Yeom, S., Giacomelli, I., Fredrikson, M., and Jha, S. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*, pp. 268–282. IEEE, 2018a.

Yeom, S., Giacomelli, I., Fredrikson, M., and Jha, S. Privacy risk in machine learning: Analyzing the connection to overfitting. In *IEEE 31st Computer Security Foundations Symposium (CSF)*. IEEE, 2018b.

Yuan, C. and Moghaddam, M. Attribute-aware generative design with generative adversarial networks. *Ieee Access*, 8:190710–190721, 2020.

Yuan, C., Marion, T., and Moghaddam, M. Dde-gan: Integrating a data-driven design evaluator into generative adversarial networks for desirable and diverse concept generation. *Journal of Mechanical Design*, 145(4):041407, 2023.

Zhang, Y., Jia, R., Pei, H., Wang, W., Li, B., and Song, D. The secret revealer: Generative model-inversion attacks against deep neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 253–261, 2020.

A. Generalization to Latent Diffusion Model (LDM)

For the Latent Diffusion Model (LDM), the calculation of t-*error* is similar to DDPM, except the intermediate latent variables are in the latent space and the reverse process is conditioned by text embeddings. Specifically, we denote by V the Variational Autoencoders (VAEs) utilized to encode the original image into the latent space, i.e., $\boldsymbol{v}_0 = V(\boldsymbol{x}_0), \boldsymbol{x}_0 \sim D$, and denote by C the text condition. The diffusion process and the denoising process can be derived as:

$$\begin{aligned} q(\boldsymbol{v}_t|\boldsymbol{v}_{t-1}) &= \mathcal{N}(\boldsymbol{v}_t; \sqrt{1-\beta_t}\boldsymbol{v}_{t-1}, \beta_t \mathbf{I}) \\ p_{\boldsymbol{\theta}}(\boldsymbol{v}_{t-1}|\boldsymbol{v}_t) &= \mathcal{N}(\boldsymbol{v}_{t-1}; \mu_\theta(\boldsymbol{v}_t, t, C), \Sigma_\theta(\boldsymbol{v}_t, t)). \end{aligned} \tag{17}$$

Then t-*error* can be rewrite as:

$$\tilde{\ell}_{t,\boldsymbol{v}_0} = ||\psi_{\boldsymbol{\theta}}(\phi_{\boldsymbol{\theta}}(\tilde{\boldsymbol{v}}_t, t, C), t, C) - \tilde{\boldsymbol{v}}_t||^2, \tag{18}$$

where we reuse the symbols $\phi_{\boldsymbol{\theta}}$ and $\psi_{\boldsymbol{\theta}}$ as the deterministic reverse and sampling regarding $\boldsymbol{v}_t$:

$$\begin{aligned} \boldsymbol{v}_{t+1} &= \phi_{\boldsymbol{\theta}}(\boldsymbol{v}_t, t, C) \\ &= \sqrt{\bar{\alpha}_{t+1}} f_{\boldsymbol{\theta}}(\boldsymbol{v}_t, t, C) + \sqrt{1-\bar{\alpha}_{t+1}}\epsilon_{\boldsymbol{\theta}}(\boldsymbol{v}_t, t, C), \end{aligned} \tag{19}$$

$$\begin{aligned} \boldsymbol{v}_{t-1} &= \psi_{\boldsymbol{\theta}}(\boldsymbol{v}_t, t, C) \\ &= \sqrt{\bar{\alpha}_{t-1}} f_{\boldsymbol{\theta}}(\boldsymbol{v}_t, t, C) + \sqrt{1-\bar{\alpha}_{t-1}}\epsilon_{\boldsymbol{\theta}}(\boldsymbol{v}_t, t, C), \end{aligned} \tag{20}$$

where

$$f_{\boldsymbol{\theta}}(\boldsymbol{v}_t, t, C) = \frac{\boldsymbol{v}_t - \sqrt{1-\bar{\alpha}_t}\epsilon_{\boldsymbol{\theta}}(\boldsymbol{v}_t, t, C)}{\sqrt{\bar{\alpha}_t}}. \tag{21}$$

B. Adopted Diffusion Models and Datasets

Adopted diffusion models and datasets, along with the data splittings, are summarized in Table 7.

Table 7: Adopted diffusion models and datasets.

Model	Dataset	Resolution	# Member	# Hold-out	Cond.
DDPM	CIFAR-10	32	25,000	25,000	-
	CIFAR-100	32	25,000	25,000	-
	STL10-U	32	50,000	50,000	-
	Tiny-ImageNet	32	50,000	50,000	-
Latent Diffusion Model	Pokemon	512	416	417	text
	COCO2017-Val	512	2,500	2,500	text
Stable Diffusion V1-4/5	Laion	512	2500	2500	text

C. Failed Defensive Training Results

We show some sampling images of defensive training results from Figure 8 and 9. The generated results are vague and unrealistic due to the aggressive regularization or data argumentation applied to the training process. Subsequently, images from whatever member or hold-out set will be reconstructed at a low quality, rendering them unworthy to conduct MIA.

Figure 8: Sampling results at the 800,000th step from DDPM trained with ℓ_2 regularization.

Figure 9: Sampling results at the 800,000th step from DDPM trained with RandAugment.