Membership Inference on Text-to-image Diffusion Models via Conditional Likelihood Discrepancy

Shengfang Zhai[1,2], **Huanran Chen**[3,6], **Yinpeng Dong**[3,6*], **Jiajun Li**[1,2], **Qingni Shen**[1,2*], **Yansong Gao**[4], **Hang Su**[3,5], **Yang Liu**[7]
[1]School of Software and Microelectronics, Peking University
[2]PKU-OCTA Laboratory for Blockchain and Privacy Computing, Peking University
[3]Dept. of Comp. Sci. and Tech., Institute for AI, BNRist Center, THBI Lab, Tsinghua University
[4]The University of Western Australia [5]Zhongguancun Laboratory, Beijing, China
[6]RealAI [7]Nanyang Technological University
{zhaisf, jiajun.lee}@stu.pku.edu.cn huanran.chen@outlook.com
{dongyinpeng, suhangss}@tsinghua.edu.cn
qingnishen@ss.pku.edu.cn garrison.gao@uwa.edu.au
yangliu@ntu.edu.sg

Abstract

Text-to-image diffusion models have achieved tremendous success in the field of controllable image generation, while also coming along with issues of privacy leakage and data copyrights. Membership inference arises in these contexts as a potential auditing method for detecting unauthorized data usage. While some efforts have been made on diffusion models, they are not applicable to text-to-image diffusion models due to the high computation overhead and enhanced generalization capabilities. In this paper, we first identify a conditional overfitting phenomenon in text-to-image diffusion models, indicating that these models tend to overfit the conditional distribution of images given the corresponding text rather than the marginal distribution of images only. Based on this observation, we derive an analytical indicator, namely **C**onditional **Li**kelihood **D**iscrepancy (**CLiD**), to perform membership inference, which reduces the stochasticity in estimating memorization of individual samples. Experimental results demonstrate that our method significantly outperforms previous methods across various data distributions and dataset scales. Additionally, our method shows superior resistance to overfitting mitigation strategies, such as early stopping and data augmentation.

1 Introduction

Text-to-image diffusion models have achieved remarkable success in the guided generation of diverse, high-quality images based on text prompts, such as Stable Diffusion [42, 46], DALLE-2 [43], Imagen [49], and DeepFloyd-IF [31]. These models are increasingly adopted by users to create photorealistic images that align with desired semantics. Moreover, they can generate images of specific concepts [32] or styles [61] when fine-tuned on relevant datasets. However, the impressive generative capabilities of these models depend heavily on high-quality image-text datasets, which involve collecting image-text data from the web. This practice raises significant privacy and copyright concerns in the community [5, 18]. The pretraining and fine-tuning processes of text-to-image diffusion models can cause copyright infringement, as they utilize unauthorized datasets published by human artists or stock-image websites [2, 10, 44, 45, 58].

*Corresponding authors.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

https://doi.org/10.52202/079017-2358

Membership inference (also known as the membership inference attack) is widely used for auditing privacy leakage of training data [4, 53], defined as determining whether a given data point has been used to train the target model. Dataset owners can thus leverage membership inference to determine if their data is being used without authorization [14, 39].

Previous works [5, 15–17, 28, 38] have attempted membership inference on diffusion models. Carlini et al. [5] employ LiRA (Likelihood Ratio Attack) [4] to perform membership inference on diffusion models. LiRA requires training multiple shadow models to estimate the likelihood ratios of a data point from different models, which incurs high training overhead (e.g., 16 shadow models for DDPM [22] on CIFAR-10 [30]), making it neither scalable nor applicable to text-to-image diffusion models. Other query-based membership inference methods [15, 17, 28, 38] design and compute indicators to evaluate whether a given data point belongs to the member set. These methods require only a few or even a single shadow model, making them scalable to larger text-to-image diffusion models. However, these methods mainly estimate model memorization for data points and do not fully utilize the conditional distribution of image-text pairs. Consequently, they achieve limited success only under excessively high training steps and fail under real steps or common data augmentation methods (Tab. 2), which do not reflect real training scenarios. Text-to-image diffusion models have demonstrated excellent performance in zero-shot image generation [1, 42, 46], indicating their strong generalization, which makes it difficult to distinguish membership by directly measuring overfitting to data points. And due to the stochasticity of diffusion training loss [22, 46], this kind of measuring becomes more challenging.

To address the challenges, we firstly identify a **Conditional Overfitting** phenomenon of text-to-image diffusion models with empirical validation, where the models exhibit more significant overfitting to the conditional distribution of the images given the corresponding text than the marginal distribution of the images only. It inspires the revealing of membership by leveraging the overfitting difference. Based on it, we propose to perform membership inference on text-to-image diffusion models via **C**onditional **Li**kelihood **D**iscrepancy (**CLiD**). Specifically, CLiD quantifies overfitting difference analytically by utilizing Kullback-Leibler (KL) divergence as the distance metric and derives a membership inference indicator that estimates the discrepancy between the conditional likelihood of image-text pairs and the likelihood of images only. We approximate the likelihoods by employing Monte Carlo sampling on their ELBOs (Evidence Lower Bounds), and design two membership inference methods: a threshold-based method CLiD$_{th}$ and a feature vector-based method CLiD$_{vec}$.

We conduct extensive experiments on three text-to-image datasets [32, 35, 66] with various data distributions and dataset scales, using the mainstream open-sourced text-to-image diffusion models [11, 47] under both fine-tuning and pretraining settings. First, our methods consistently outperform existing baselines across various data distributions and training scenarios, including fine-tuning settings and the pretraining setting. Second, our experiments on fine-tuning settings with different training steps (Sec. 4.2) reveal that excessively high step/image ratios cause overfitting, leading to hallucination success; and we develop a more realistic pretraining setting following [13, 16], where our experiments reveal the insufficient effect of existing membership inference works [15, 17, 28, 38]. Third, our comparison experiment with varying training steps (Sec. 4.3) indicates that the effectiveness of membership inference grows with higher step/image ratios and should be evaluated under reasonable settings for realistic results. Next, ablation studies further demonstrate the effect of our CLiD indicator, even with fewer query count, our method still outperforms baseline methods (Fig. 3). Last, experiments show that our methods exhibit stronger resistance to data augmentation, and exhibit resistance to even adaptive defenses.

2 Diffusion Model Preliminaries

Denoising Diffusion Probabilistic Model (DDPM) [22] learns the data distribution $\mathbf{x}_0 \sim q(\mathbf{x})$ by reversing the forward noise-adding process. For the forward process, DDPM defines a Markov process of adding Gaussian noise step by step:

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) := \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}), \tag{1}$$

where $\beta_t \in (0,1)$ is the hyperparameter controlling the variance. For the reverse process, DDPM defines a learnable Markov chain starting at $p(\mathbf{x}_T) = \mathcal{N}(\mathbf{x}_T; \mathbf{0}, \mathbf{I})$ to generate $\mathbf{x}_0$:

$$p_\theta(\mathbf{x}_0) = \int_{\mathbf{x}_{1:T}} p(\mathbf{x}_T) \prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)\, \mathrm{d}\mathbf{x}_{1:T}, \qquad p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \sigma_t^2), \tag{2}$$

https://doi.org/10.52202/079017-2358

where σ_t^2 is the untrained time-dependent constant. θ represents the trainable parameters. To maximize $p_\theta(\mathbf{x}_0)$, DDPM optimizes the Evidence Lower Bound (ELBO) of the log-likelihood [22, 33]:

$$\log p_\theta(\mathbf{x}_0) \geq \mathbb{E}_{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}\left[\log \frac{p_\theta(\mathbf{x}_{0:T})}{q(\mathbf{x}_{1:T} \mid \mathbf{x}_0)}\right] = -\mathbb{E}_{\epsilon,t}\left[||\epsilon_\theta(\mathbf{x}_t, t) - \epsilon||^2\right] + C, \tag{3}$$

where $\epsilon \sim \mathcal{N}(0, \mathcal{I})$, $t \sim \text{Uniform}(1, ..., T)$ and C is a constant. $\mathbf{x}_t$ is obtained from Eq. (1), and ϵ_θ is a function approximator intended to predict the noise ϵ from $\mathbf{x}_t$. Omitting the untrainable constant in Eq. (3) and taking its negative yields the loss function of training DDPM.

Conditional diffusion models [21, 40, 46]. To achieve controllable generation ability, text-to-image diffusion models incorporate the conditioning mechanism into the model, which are also known as conditional diffusion models, enabling them to learn conditional probability as:

$$p_\theta(\mathbf{x}_0|\mathbf{c}) = \int_{\mathbf{x}_{1:T}} p(\mathbf{x}_T) \prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{c})\, \mathrm{d}\mathbf{x}_{1:T}, \tag{4}$$

where $\mathbf{c}$ denotes the embedding of condition. For text-to-image synthesis, $\mathbf{c} := \mathcal{T}(\mathbf{y})$, where $\mathbf{y}$ and $\mathcal{T}$ denote the text input and the text encoder, respectively. Similar to Eq. (3), through derivation [33], we can obtain the ELBO of the conditional log-likelihood:

$$\log p_\theta(\mathbf{x}_0|\mathbf{c}) \geq -\mathbb{E}_{\epsilon,t}\left[||\epsilon_\theta(\mathbf{x}_t, t, \mathbf{c}) - \epsilon||^2\right] + C. \tag{5}$$

3 Methodology

In this section, we detail the proposed **C**onditional **Li**kelihood **D**iscrepancy (**CLiD**) method. We first introduce the threat model of query-based membership inference in Sec. 3.1. We then identify the conditional overfitting phenomenon with experimental validation in Sec. 3.2. We further drive the membership inference indicator based on CLiD in Sec. 3.3 and design two practical membership inference methods in Sec. 3.4. We finally provide the implementation details in Sec. 3.5.

3.1 Threat Model

We use the standard security game of membership inference on image-text data following previous work [4, 5, 38, 48]. We define a challenger $\mathcal{C}$ and an adversary $\mathcal{A}$ who performs membership inference. $\mathcal{C}$ samples a member set $D_{\text{mem}} \leftarrow \mathbb{D}$ and trains or fine-tunes a text-to-image diffusion model f_θ (i.e., target model) with D_{mem}. The rest of $\mathbb{D}$ is denoted by hold-out set $D_{\text{out}} = \mathbb{D} \setminus D_{\text{mem}}$. For a given data point $(\mathbf{x}, \mathbf{c}) \in \mathbb{D}$, $\mathcal{A}$ designs an algorithm $\mathcal{M}$ to yield a membership prediction:

$$\mathcal{M}(\mathbf{x}, \mathbf{c}, f_\theta) = \mathbb{1}\left[\mathcal{M}'(\mathbf{x}, \mathbf{c}, f_\theta) > \tau\right], \tag{6}$$

where $\mathcal{M}'$ denotes an indicator function that reflects membership information, and τ denotes a tunable decision threshold of query-based membership inference [15, 17, 28, 38].

We consider a grey-box setting [2] consistent with previous query-based methods [15, 17, 28, 38]. This setting assumes that $\mathcal{A}$ has access to the intermediate outputs of models without knowledge of specific model parameters. For the given image-text data point $(\mathbf{x}, \mathbf{c})$, we assume that $\mathbf{x}$ and $\mathbf{c}$ always correspond within the dataset $\mathbb{D}$. This assumption is evident in scenarios where dataset copyright owners perform membership inference to audit usage. And we also consider a weaker assumption of conducting membership inference without the groundtruth text in Sec. 4.6.

Conversely, challenger $\mathcal{C}$ can mitigate the effectiveness of membership inference during training by utilizing data augmentation or even adaptive defense methods, which we discuss in Sec. 4.5. Our work primarily focuses on fine-tuning scenarios because the weights of pretrained models are readily available, making this scenario more prone to copyright risks [41, 56]. Numerous projects are implemented by fine-tuning open-source models on specific datasets [3, 24, 60, 64]. We also conduct experiments on pretrained text-to-image diffusion models (Tab. 3) to demonstrate the effectiveness of our method even in pretraining scenarios.

[2] Note that in most real-world scenarios, the requirements for $\mathcal{A}$ in gray-box and white-box settings are nearly identical. We use this terminology here for consistency with previous works [15, 28].

https://doi.org/10.52202/079017-2358

3.2 Conditional Overfitting Phenomenon

The rationale behind previous studies primarily hinges on the overfitting of diffusion models to training data (usually image data $\mathbf{x}$) [7, 8, 15, 28, 38]. This overfitting tends to result in lower estimation errors for images in the member set (training data) compared to those in the hold-out set during the diffusion process. Various indicators [15, 28, 38] are designed based on this to expose membership information. Specifically, let q_{mem} and q_{out} represent the image distributions of the member set and the hold-out set, respectively. p represents the diffusion models' estimated distribution, and D denotes a distance metric (which will be specified later). This rationale can be formulated as:

$$D(q_{\text{mem}}(\mathbf{x}), p(\mathbf{x})) \leq D(q_{\text{out}}(\mathbf{x}), p(\mathbf{x})). \tag{7}$$

However, if considering the membership inference on text-to-image diffusion models with image-text data $(\mathbf{x}, \mathbf{c})$, we emphasize the following assumption:

Assumption 3.1 (Conditional overfitting phenomenon)**.** *The overfitting of text-to-image diffusion models to the conditional distribution of* $(\mathbf{x}, \mathbf{c})$ *is more salient than to the marginal distribution of* $\mathbf{x}$*:*

$$\underbrace{\mathbb{E}_{\mathbf{c}}[D(q_{out}(\mathbf{x}|\mathbf{c}), p(\mathbf{x}|\mathbf{c})) - D(q_{mem}(\mathbf{x}|\mathbf{c}), p(\mathbf{x}|\mathbf{c}))]}_{\textit{overfitting to conditional distribution}} \geq \underbrace{D(q_{out}(\mathbf{x}), p(\mathbf{x})) - D(q_{mem}(\mathbf{x}), p(\mathbf{x}))}_{\textit{overfitting to marginal distribution}}. \tag{8}$$

Empirically, we validate this assumption by using Fréchet Inception Distance (FID) [20] as the metric D, i.e., D_{FID}. We calculate $D_{FID}(q(\mathbf{x}|\mathbf{c}), p(\mathbf{x}|\mathbf{c}))$ using the MS-COCO [35] dataset on a fine-tuned Stable Diffusion [46] model. Then by gradually truncating the original condition text to $\{2/3, 1/3, Null\}$ to obtain $\mathbf{c}^*$, we calculate $D_{FID}(q(\mathbf{x}|\mathbf{c}^*), p(\mathbf{x}|\mathbf{c}^*))$ as a stepwise approximation of $D_{FID}(q(\mathbf{x}), p(\mathbf{x}))$. In Fig. 1, we report the FID scores of synthetic images under different conditions of member set and hold-out set. A smaller FID value indicates a closer match between model distributions and dataset distributions. From Fig. 1 (a), it can be observed that for the full condition, the FID difference between the member set and the hold-out set is consistently higher than that for the truncated conditions, which validates our assumptions. We also demonstrate the validation utilizing other metrics in Appendix A.

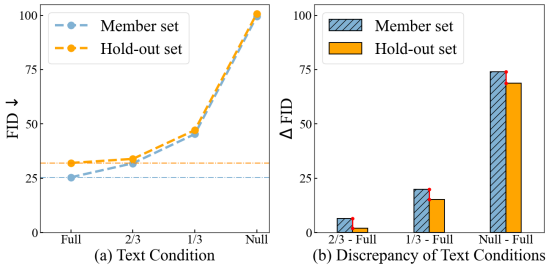

Figure 1: FID values and the FID differences of synthetic images ($2500/2500$ samples for member/hold-out set) under different conditions of member set and hold-out set.

We further compute the change in FID after truncating the condition and observe that the change in FID of the member set is consistently greater than that of the hold-out set (Fig. 1 (b)), which inspires revealing membership by this overfitting discrepancy. Recalling the aim of text-to-image diffusion model is to fit a latent space mapping from text to image, image data augmentation is commonly used to enhance the model generalization. For instance, the official fine-tuning script of Hugging-Face [24] employs Random-Crop and Random-Flip as the default augmentation [25]. However, few trainers disturb the text condition as it is discrete and such disturbance would result in a decline of model utility (Sec. 4.5). Therefore, we believe that leveraging this phenomenon contributes to addressing the challenges of the strong generalization of text-to-image diffusion models with the resistance to data augmentation.

3.3 Condition Likelihood Discrepancy

In this section, we derive a membership inference indicator for a given individual sample based on Assumption 3.1. Calculating FID requires sampling lots of images from the p distribution, which is impractical under membership inference scenarios. Instead, we employ Kullback-Leibler (KL) divergence as the distance metric, which is widely used and computationally convenient (the usage of other metrics is discussed in Appendix C). Then we have the following theorem:

Theorem 3.2. *(Proof in Appendix B) When using* $D = D_{KL}$ *as distance metric, Assumption 3.1 is equivalent to:*

$$\mathbb{E}_{q_{mem}(\mathbf{x},\mathbf{c})}[\log p(\mathbf{x}|\mathbf{c}) - \log p(\mathbf{x})] \geq \mathbb{E}_{q_{out}(\mathbf{x},\mathbf{c})}[\log p(\mathbf{x}|\mathbf{c}) - \log p(\mathbf{x})] + \delta_H, \tag{9}$$

https://doi.org/10.52202/079017-2358

where

$$\delta_H = H(q_{out}(\mathbf{x})) + \mathbb{E}_{\mathbf{c}}[H(q_{mem}(\mathbf{x}|\mathbf{c}))] - H(q_{mem}(\mathbf{x})) - \mathbb{E}_{\mathbf{c}}[H(q_{out}(\mathbf{x}|\mathbf{c}))]. \tag{10}$$

Let us define:

$$\mathbb{I}(\mathbf{x}, \mathbf{c}) = \log p(\mathbf{x}|\mathbf{c}) - \log p(\mathbf{x}). \tag{11}$$

If δ_H is negligible, then according to Eq. (9), it holds that $\mathbb{E}_{q_{\text{mem}}}[\mathbb{I}(\mathbf{x})] \geq \tau \geq \mathbb{E}_{q_{\text{out}}}[\mathbb{I}(\mathbf{x})]$, where τ is a constant intermediate between the left-hand side and right-hand side. Membership inference is then posed as follows: given an input instance $(\mathbf{x}, \mathbf{c})$, measuring $\mathbb{I}(\mathbf{x}, \mathbf{c})$ to predict how probable it is that the input is a sample from q_{mem} rather than q_{out}. Intuitively, if $\mathbb{I}(\mathbf{x})$ exceeds a threshold τ, the instance is likely from q_{mem}; otherwise, it belongs to q_{out}. In the community of membership inference methods [4–6, 15, 17, 28, 38, 65] , setting such a threshold τ is a standard practice to differentiate between the two distributions. Therefore, we can utilize the indicator $\mathbb{I}(\mathbf{x}, \mathbf{c})$ for membership inference. Since Eq. (11) actually involves measuring the likelihood discrepancy under different conditions of diffusion models, we call it **C**onditional **Li**kelihood **D**iscrepancy (**CLiD**).

In order to calculate the likelihoods in Eq. (11) for a given data point $(\mathbf{x}, \mathbf{c})$, we utilize the ELBOs in Eq. (3) and Eq. (5) as an approximation of the log-likelihoods:

$$\mathbb{I}(\mathbf{x}, \mathbf{c}) = \mathbb{E}_{t,\epsilon}\left[||\epsilon_\theta(\mathbf{x}_t, t, \mathbf{c}_{\text{null}}) - \epsilon||^2\right] - \mathbb{E}_{t,\epsilon}\left[||\epsilon_\theta(\mathbf{x}_t, t, \mathbf{c}) - \epsilon||^2\right], \tag{12}$$

where $\mathbf{c}_{\text{null}}$ denotes an empty text condition input used to estimate the approximation of $\log p_\theta(\mathbf{x})$.

3.4 Implementation of CLiD-MI

In practice, calculating Eq. (12) needs a Monte Carlo estimate for data point by sampling N times using (t_i, ϵ_i) pairs, with $\epsilon_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $t_i \sim [1, 1000]$. Performing two Monte Carlo estimations independently incurs high computational costs, resulting in $2 \times N$ query count, where N is typically a large number to ensure accurate estimation. To simplify computation, we instead perform Monte Carlo estimation on the difference of the ELBOs inspired by [33]:

$$\mathbb{I}(\mathbf{x}, \mathbf{c}) = \mathbb{E}_{t,\epsilon}\left[||\epsilon_\theta(\mathbf{x}_t, t, \mathbf{c}_{\text{null}}) - \epsilon||^2 - ||\epsilon_\theta(\mathbf{x}_t, t, \mathbf{c}) - \epsilon||^2\right]. \tag{13}$$

In experiments, to further mitigate randomness, we also consider diverse reduced conditions along with $\mathbf{c}_{\text{null}}$, forming the reduced condition set $\mathbb{C} = \{\mathbf{c}_1^*, \mathbf{c}_2^* ..., \mathbf{c}_k^*\}$, where we set $\mathbf{c}_k^* = \mathbf{c}_{\text{null}}$. Then we compute multiple condition likelihood discrepancies:

$$\mathcal{D}_{\mathbf{x},\mathbf{c},\mathbf{c}_i^*} = \mathbb{E}_{t,\epsilon}\left[||\epsilon_\theta(\mathbf{x}_t, t, \mathbf{c}_i^*) - \epsilon||^2 - ||\epsilon_\theta(\mathbf{x}_t, t, \mathbf{c}) - \epsilon||^2\right], \tag{14}$$

where $\mathbf{c}_i^* \in \mathbb{C}$. In subsequent parts, we employ their mean or treat them as feature vectors to reveal membership information. We will introduce how to obtain $\mathbb{C}$ in Sec. 3.5.

Combining $p_\theta(\mathbf{x}|\mathbf{c})$ for further enhancement. Recall that the practical significance of sample likelihood is the probability that a data point originates from the model distribution, which essentially can also be used to assess membership. Due to the monotonicity of the log function, we can also use ELBO of Eq. (5) to estimate $p_\theta(\mathbf{x}|\mathbf{c})$:

$$\mathcal{L}_{\mathbf{x},\mathbf{c}} = -\mathbb{E}_{t,\epsilon}\left[||\epsilon_\theta(\mathbf{x}_t, t, \mathbf{c}) - \epsilon||^2\right]. \tag{15}$$

Additionally, this estimation can reuse results from estimating Eq. (14), thus obviating any additional query counts. Next, we consider two strategies to combine Eq. (14) and Eq. (15) to construct the final membership inference method.

Threshold-based attack–CLiD$_{th}$. First, we normalize the two indicators to the same feature scale. Due to the outliers in the data, we use Robust-Scaler: $\mathcal{S}(a_i) = (a_i - \tilde{a})/IQR$, where a_i denotes the i-th value, $\tilde{a}$ denotes the mean and IQR (interquartile range) is defined as the difference between the third quartile (Q3) and the first quartile (Q1) of the feature. Then we have:

$$\mathcal{M}_{\text{CLiD}_{th}}(\mathbf{x}, \mathbf{c}) = \mathbb{1}\left[\alpha \cdot \mathcal{S}(\frac{1}{k}\sum_{i}^{k}\mathcal{D}_{\mathbf{x},\mathbf{c},\mathbf{c}_i^*}) + (1-\alpha)\cdot\mathcal{S}(\mathcal{L}_{\mathbf{x},\mathbf{c}}) > \tau\right], \tag{16}$$

where k denotes the total number of reduced $\mathbf{c}^*$ (i.e., $k = |\mathbb{C}|$), and α is a weight parameter.

Vector-based attack–CLiD$_{vec}$. We combine the estimated values of Eq. (14) and Eq. (15) to obtain the feature vectors corresponding to each data point:

$$\mathbf{V} = \left(\mathcal{D}_{\mathbf{x},\mathbf{c},\mathbf{c}_1^*}, \mathcal{D}_{\mathbf{x},\mathbf{c},\mathbf{c}_2^*} \ldots \mathcal{D}_{\mathbf{x},\mathbf{c},\mathbf{c}_k^*}, \mathcal{L}_{\mathbf{x},\mathbf{c}}\right). \tag{17}$$

https://doi.org/10.52202/079017-2358

We use a simple classifier to distinguish feature vectors in order to determine the membership of the samples:

$$\mathcal{M}_{\text{CLiD}_{vec}}(\mathbf{x}, \mathbf{c}) = \mathbb{1}\left[\mathcal{F}_{\mathcal{M}}(\mathbf{V}) > \tau\right], \tag{18}$$

where $\mathcal{F}_{\mathcal{M}}$ denotes the predict confidence of the classifier.

3.5 Practical Considerations

Reducing conditions to obtain $\mathbf{c}^*$. We consider three methods for diverse reduction: (1) Simply taking the first, middle, and last thirds of the sentences as text inputs. (2) Randomly adding Gaussian noises with various scales to the text embeddings. (3) Calculating the importance of words in the text [55, 57] and replacing them with "pad" tokens by varying proportions in descending order. For all three methods, we additionally use the null text input as $\mathbf{c}_k^*$. These methods are all effective and we use (3) with $k = 4$ in subsequent experiments (details in Appendix D).

Monte Carlo sampling. Let M and N denote the Monte Carlo sampling numbers of estimating $\mathcal{L}(\mathbf{x}, \mathbf{c})$ and $\mathcal{D}_{\mathbf{x},\mathbf{c},\mathbf{c}_i^*}$, respectively. We set $M = N$ to achieve result reuse between Eq. (14) and Eq. (15), reducing the number of Monte Carlo sampling. Hence the overall query count of one data point is $M + K \cdot N$. Significant effects can be observed even when $M, N = 1$ (Fig. 3).

Classifiers of CLiD$_{vec}$. Due to the simplicity of the feature vectors, we do not need a neural network as the classifier [15]. Simpler classifiers help to prevent overfitting. In our experiments, we utilize XGBoost [9] and utilize its predict confidence.

4 Experiments

4.1 Setups

Datasets and models. For the fine-tuning setting, we select $416/417$ samples on Pokémon [32], $2500/2500$ samples on MS-COCO [35] and $10,000/10,000$ samples on Flickr [66] as the member/hold-out dataset, respectively. These three datasets involve diverse data distributions and dataset scales. We use the most widely used text-to-image diffusion model, Stable Diffusion v1-4[3] [11], as the target model to fine-tune it on these three datasets. For the pretraining setting, we conduct experiments on Stable Diffusion v1-5[4] [47] using the processed LAION dataset [51] (detailed in Sec. 4.2) to minimize distribution shift [13, 16].

Fine-tuning setups. For fine-tuning, previous membership inference on text-to-image diffusion models usually relies on strong overfitting settings. To evaluate the performance more realistically, we consider the two following setups: (1) *Over-training.* Following the previous works [15, 17, 28], we fine-tune 15,000 steps on Pokemon datasets, and 150,000 steps on MS-COCO and Flickr (with only $2500/2500$ dataset size). (2) *Real-world training.* Considering that trainers typically do not train for such high steps, we recalibrate the steps based on the training steps/dataset size ratio (approximately 20) of official fine-tuning scripts on Huggingface [24]. Thus, we train 7,500 steps, 50,000 steps and 200,000 steps for the Pokémon, MS-COCO and Flickr datasets, respectively. Additionally, we employ the default data augmentation (Random-Crop and Random-Flip [25]) in training codes [25] to simulate real-world scenarios.

Baselines. We broadly consider existing member inference methods applicable to text-to-image diffusion models as our baselines: Loss-based inference [38], SecMI$_{\text{stats}}$ (SecMI) [15], PIA [28], PFAMI$_{\text{Met}}$ (PFAMI) [17] and an additional method of directly conducting Monte Carlo estimation (M. C.) on Eq. (15) for comparison. For all baselines, we use the parameters recommended in their papers. We omit some membership inference methods for generative models [6, 19, 36], as they have been shown ineffective for diffusion models in previous works [15, 17].

Evaluation metrics. We follow the widely used metrics of previous works [4, 5, 15, 17, 28], including ASR (i.e., the accuracy of membership inference), AUC and the True Positive Rate (TPR) when the False Positive Rate (FPR) is 1% (i.e., TPR@1%FPR).

Implementation details. Our evaluation follows the setup of representative membership inference works [4, 5]. It is important to note that some implementations [26, 29] of previous works assume

[3] https://huggingface.co/CompVis/stable-diffusion-v1-4

[4] https://huggingface.co/runwayml/stable-diffusion-v1-5

Table 1: Results under *Over-training* setting. We mark the best and second-best results for each metric in **bold** and underline, respectively. Additionally, the best results from baselines are marked in blue for comparison.

Method	MS-COCO			Flickr			Pokemon			Query
	ASR	AUC	TPR@1%FPR	ASR	AUC	TPR@1%FPR	ASR	AUC	TPR@1%FPR	
Loss	81.92	89.98	32.28	81.90	90.34	40.80	83.76	91.79	25.77	1
PIA	68.56	75.12	5.08	68.56	75.12	5.08	83.37	90.95	13.31	2
M. C.	82.04	89.77	36.04	83.32	91.37	41.20	79.35	86.78	23.74	3
SecMI	83.00	90.81	50.64	62.96[†]	89.29	48.52	80.49	90.64	9.36	12
PFAMI	94.48	98.60	78.00	90.64	96.78	50.96	89.86	95.70	65.35	20
CLiD$_{th}$	99.08	**99.94**	**99.12**	91.42	97.39	**74.00**	**97.96**	99.28	**97.84**	15
CLiD$_{vec}$	**99.74**	99.31	95.20	**91.78**	**97.52**	73.88	97.36	**99.46**	96.88	15

[†] When conducting SecMI [15], we observe that the thresholds obtained on the shadow model sometimes do not transfer well to the target model.

Table 2: Results under *Real-world training* setting. We also highlight key results according to Tab. 1.

Method	MS-COCO			Flickr			Pokemon			Query
	ASR	AUC	TPR@1%FPR	ASR	AUC	TPR@1%FPR	ASR	AUC	TPR@1%FPR	
Loss	56.28	61.89	1.92	54.91	56.60	1.83	61.03	65.96	2.82	1
PIA	54.10	55.52	1.76	51.96	52.73	1.28	58.34	59.95	2.64	2
M. C.	57.98	61.97	2.64	54.92	56.78	2.16	61.10	66.48	3.84	3
SecMI	60.94	65.40	3.92	55.60	63.85	2.76	61.28	65.56	0.84	12
PFAMI	57.36	60.39	2.72	54.68	56.13	1.80	58.94	63.53	5.76	20
CLiD$_{th}$	88.88	96.13	**67.52**	87.12	94.74	53.56	**86.79**	**93.28**	**61.39**	15
CLiD$_{vec}$	**89.52**	**96.30**	66.36	**88.86**	**95.33**	**53.92**	85.47	92.61	59.95	15

access to a portion of the exact member set and the hold-out set to obtain a threshold for calculating ASR or to train a classification network [26]. This assumption does not align with real-world scenarios. Therefore, we strictly adhere to the fundamental assumption of membership inference [4, 17]: knowing only the overall dataset without any knowledge of the member/hold-out split. Hence, we first train a shadow model to obtain the α for Eq. (16), classifiers for Eq. (18) and the threshold τ for calculating ASR with auxiliary datasets of the same distribution. Then we perform the test on the target models. Other implementation details are provided in Appendix D.

4.2 Main Results

Over-training setting (fine-tuning). In Tab. 1, models are trained for excessive steps on all three datasets, resulting in significant overfitting. We observe that under this over-training scenario, both of our methods nearly achieve ideal binary classification effectiveness. For instance, CLiD$_{th}$ achieves over 99% ASR, AUC and TPR@1%FPR value on the MS-COCO dataset [35]. With this training setup, the metrics for different baselines are very similar. Even the simplest loss-based method [38] (with the query count of 1) also yields satisfactory results compared with other high query count methods. Therefore, we emphasize: *This unrealistic over-training setting fails to adequately reflect the effectiveness differences among various membership inference methods*.

Real-world training setting (fine-tuning). In Tab. 2, we adjust the training steps simulating real-word training scenario [24] and utilize default data augmentation [25]. The best value of ASR and AUC of baseline methods decreases to around 65%, and the best value of TPR@1%FPR decreases to around 5%, indicating insufficient effectiveness of previous member inference methods in real-world training scenarios of text-to-image diffusion models. In contrast, our methods maintain ASR above 86% and AUC above 93%, exceeding the best baseline values by about 30%. The TPR@1%FPR of our methods exceeds the best baseline values by 50%~60%. The results demonstrate the effectiveness of our methods across various data distributions and scales in real-world training scenarios.

Pretraining setting. For the pretraining setting, we adopt a stringent and realistic membership inference setting based on previous works [13, 16]. (1) To ensure the distribution consistency between the member and hold-out set, we respectively select 2500 samples from the LAION-Aesthetics v2 5+ and LAION-2B MultiTranslated [51] as member/hold-out set following [16]; (2) We filter out samples containing non-English characters to ensure there are no other "distinguishable tails" [13] in

https://doi.org/10.52202/079017-2358

Method	LAION			Query
	ASR	AUC	TPR@1%FPR	
Loss	51.78	50.90	1.75	1
PIA	52.13	52.42	1.25	2
M. C.	53.18	53.96	1.25	3
SecMI	57.43	58.59	2.45	12
PFAMI	59.08	61.11	1.45	20
CLiD$_{th}$	**64.53**	**67.82**	**5.01**	15

Table 3: The performance of membership inference methods on Stable Diffusion v1-5 [47] in pretraining setting. We utilize the processed LAION dataset to ensure the distribution consistency between member / hold-out sets [13, 16]. The best results are highlighted in **bold**.

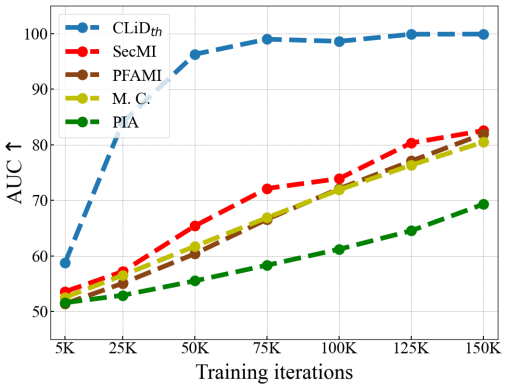

Figure 2: Effectiveness trajectory on various training steps.

the dataset[5]. We conduct membership inference on Stable Diffusion v1-5 [46]. As shown in Tab. 3, our method consistently outperforms the baselines across all three metrics.

4.3 Performance on Various Training Steps

From Tab. 1 and Tab. 2, we find that the training steps greatly influence the effectiveness of membership inference. All membership inference methods tend to exhibit satisfactory performance when the model is trained for an excessive number of steps that conflicts with real-world scenarios. Therefore, we emphasize that the ***effectiveness trajectory*** of membership inference across varying training steps should also be utilized to evaluate different methods. Better membership inference methods should reveal membership information earlier as training progresses.

To explore this, we fine-tune Stable Diffusion models with the MS-COCO dataset for varying training steps under *real-world training* setting and report the AUC values of different membership inference methods in Fig. 2. It can be observed that as the training progresses, CLiD$_{th}$ exhibits a significantly faster increase in effectiveness trajectory. By $25,000$ steps, CLiD$_{th}$ effectively exposes membership information, whereas other baselines achieve similar results only at around $150,000$ steps. This demonstrates that our method can effectively reveal membership information when the overfitting degree of the text-to-image diffusion model is much weaker.

4.4 Ablation Study

To conduct an ablation study, we vary the Monte Carlo sampling count in Eq. (14) and Eq. (15), perform CLiD$_{th}$ with MS-COCO dataset under *real-world training* setting and report the AUC values in Fig. 3. To further compare the effects of Eq. (14) and Eq. (15), we discard each term in Eq. (16) and denote it as $M/N = 0$. We also include the result of the best baseline, SecMI [15], as a comparison.

Effect of $\mathcal{D}_{\mathbf{x},\mathbf{c},\mathbf{c}^*}$. In Fig. 3, results of "M=1, N=0" and "M=1, N=1" show a significant improvement of membership inference by including $\mathcal{D}_{\mathbf{x},\mathbf{c},\mathbf{c}^*}$. Results of "M=5, N=0" and "M=1, N=1" further show that the method utilizing $\mathcal{D}_{\mathbf{x},\mathbf{c},\mathbf{c}^*}$ performs much better under the same sampling numbers. Additionally, the results of "M=0, N=1" and "M=1, N=1" indicates that only considering both Eq. (14) and Eq. (15) achieves the optimal performance.

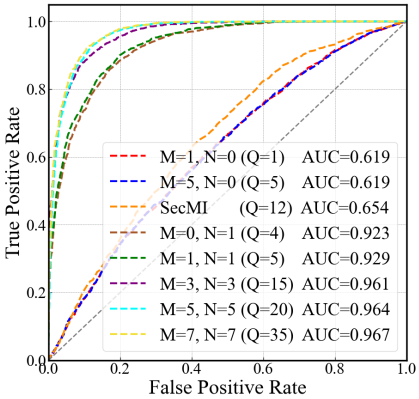

Figure 3: Performance of CLiD$_{th}$ and SecMI under various Monte Carlo sampling numbers (i.e., query count). The legend labels are sorted in ascending order by AUC values.

[5]Das et al. [13] indicates that MultiTranslated-LAION dataset contains fewer non-English characters than the LAION dataset due to the use of the translation model.

https://doi.org/10.52202/079017-2358

Table 4: The performance of different methods under no augmentation and default augmentation.

Method	No Augmentation			Defaut Augmentation		
	ASR	AUC	TPR@1%FPR	ASR (Δ)	AUC (Δ)	TPR@1%FPR (Δ)
Loss	66.54	72.73	7.72	56.28 (-10.26)	61.89 (-10.84)	1.92 (-5.80)
PIA[†]	56.56	59.28	2.00	54.10 (-2.46)	55.52 (-3.76)	1.76 (-0.24)
SecMI	72.02	81.07	13.72	60.94 (-11.08)	65.40 (-15.08)	3.92 (-9.80)
PFAMI	79.20	87.05	18.44	57.36 (-21.84)	60.39 (-26.66)	2.72 (-15.72)
CLiD$_{th}$	96.76	99.47	91.72	88.88 **(-7.88)**	96.13 **(-3.34)**	67.52 (-24.20)[‡]

[†]We omit the discussion of PIA as it shows no effectiveness at this training steps, with the metrics consistently approximating random guessing.

[‡]The TPR@1%FPR value changes significantly here because its ROC curve is very sharp when FPR close to 0.

Monte Carlo sampling numbers. In Fig. 3, we observe that when setting $M = N$, the performance improves as the number of Monte Carlo sampling increases. And the performance is improved slightly when $M, N > 3$. Hence, we set $M, N = 3$ to ensure the balance between a low query count and satisfied performance. Moreover, the experiment results of "M=1, N=1" and "SecMI" also demonstrate: CLiD$_{th}$ outperforms previous works even with a much fewer query count.

4.5 Resistance to Defense

Since data augmentation is commonly used in training and can mitigate the effectiveness of membership inference [15], we use it to evaluate the performance of methods under defense. As the baseline methods already exhibit weak performance under *real-world training* setting, we opt not to incorporate additional data augmentation. Instead, we remove the default data augmentation from training scripts [25] to observe the effectiveness change of different methods. We fine-tune Stable Diffusion models for 50,000 steps with MS-COCO, report the metrics, and calculate the metrics changes in Tab. 4. We observe that the effectiveness of all membership inference methods declines after data augmentation is introduced during training. Note that PFAMI [17] exhibits the highest sensitivity to data augmentation since it infers membership by probability fluctuation after images are perturbed, which also explains its significant performance decline between Tab. 1 and Tab. 2. Compared to the baselines, our method exhibits the smallest decrease, which indicates its stronger resistance to data augmentation.

Table 5: Effectiveness of CLiD$_{th}$ in adaptive defense. We calculate the FID [20] with 10,000 unseen MS-COCO samples to assess the model utility.

Defense	CLiD$_{th}$ on MS-COCO			FID ↓ / Δ
	ASR	AUC	TPR@1%FPR	
None	88.88	96.13	67.52	13.17
Reph	85.32	93.83	55.67	13.58 / +0.41
Del-1	86.40	93.59	59.52	13.18 / -0.01
Del-3	83.91	91.52	52.03	12.92 / -0.25
Shuffle	65.89	67.37	0.15	18.26 / +5.09[†]

[†]Compared to other methods, the increase in FID caused by shuffling is unacceptable for generative models.

Adaptive defense. We further consider adaptive defense: assuming the trainers are aware of our methods and perturb the text of image-text datasets before training. We consider the following adaptive defense methods: (1) rephrasing the original text[6], (2) randomly deleting 10%, 30% words in text, and (3) shuffling 50% of the image-text mappings in the dataset. In Tab. 5, we observe that except for *shuffling*, the other adaptive defense methods have almost no effect on CLiD$_{th}$. And *shuffling* damages the model utility (too high FID values), rendering this defense meaningless.

4.6 Weaker Assumption

Although in Sec. 3.1 we assume that the adversary can access the entire image-text pairs based on the real-world data usage auditing scenario, we also consider a weaker assumption: the adversary can only access the image without the corresponding text.

In this scenario, we first generate pseudo-text corresponding to the images using an image captioning model (BLIP [34] in our experiments), and then conduct CLiD-MI based on the image-pseudo_text pairs. In Tab. 6, we observe that our method still broadly outperforms baselines. We believe this is because the pseudo-text preserves the image's key semantics, keeping our methods effective.

[6]We utilize ChatGPT-3.5 with the following prompt: "Please rewrite the following sentences while keeping the key semantics."

https://doi.org/10.52202/079017-2358

Table 6: Results without access to the corresponding text under *Over-training* setting and *Real-world training* setting. We fine-tune MS-COCO on SDv1-4. Key results are highlighted as Tab. 1.

Method	*Over-training* (Pseudo-Text)			*Real-world training* (Pseudo-Text)			Query
	ASR	AUC	TPR@1%FPR	ASR	AUC	TPR@1%FPR	
Loss	73.80	81.01	9.71	56.08	58.47	1.60	1
PIA	61.40	65.75	1.20	53.44	54.38	1.52	2
M. C.	74.36	81.55	11.28	56.68	60.00	1.28	3
SecMI	82.04	88.97	40.80	60.48	64.04	3.28	12
PFAMI	91.56	95.16	68.16	58.12	59.77	2.64	20
CLiD$_{th}$	92.84	95.43	**72.36**	76.16	83.27	**19.76**	15
CLiD$_{vec}$	**93.26**	**96.59**	71.73	**77.76**	**84.48**	18.06	15

5 Related Works

Copyright protection in text-to-image synthesis. To protect the copyright of text-to-image models, several works [67, 68] propose inserting backdoors to embed watermarks in text-to-image models. To protect the copyright of image-text datasets, some works [50, 52, 69] incorporate imperceptible perturbations to render the released datasets unusable. Other works [12, 56] utilize the backdoor or watermark to track the usage of image-text datasets. In contrast, our method indicates the possibility of auditing the unauthorized usage of individual image-text data points utilizing membership inference.

Membership inference on diffusion models. In the grey-box or white-box setting, Carlini et al. [5] firstly conduct membership inference on unconditional diffusion models by conducting LiRA (Likelihood Ratio Attack) [4], with the requirement of training multiple shadow models. Matsumoto et al. [38] make the first step by utilizing diffusion loss to conduct query-based membership inference. Some works [15, 28] leverage the DDIM [54] deterministic forward process [27] to access the posterior estimation errors of diffusion models. And Fu et al. [17] leverage the probability fluctuations by perturbing image samples. Few works consider the black-box settings [41, 62]. However, these studies either assume partial knowledge of member set data [62] or assume extensive fine-tuning steps [41] ($100 \sim 500$ epochs), both of which do not align with real-world scenarios.

Memorization detection in text-to-image models. A similar work [59] detects token memorization by inspecting the magnitude of text-conditional predictions, but differs from ours by lacking in-depth rationale analysis and a rigorous membership inference setup with randomly selected member/hold-out sets.

6 Conclusion

In this paper, we identify the phenomenon of conditional overfitting in text-to-image models and propose **CLiD-MI**, the membership inference framework on text-to-image diffusion models utilizing the derived indicator, conditional likelihood discrepancy. Experimental results demonstrate the superiority of our method and its resistance against early stopping and data augmentation. Our method aims to inspire a new direction for the community regarding unauthorized usage auditing.

Limitations: Due to the limited availability of open-source text-to-image diffusion models, evaluations under the pretraining setting are not sufficient. Considering fine-tuning setting involves a multi step/image ratio, we acknowledge that the superiority of **CLiD-MI** over the baselines in the pretraining setting is not as evident compared to fine-tuning setting. We emphasize our experiments under pretraining setting (Tab. 3) reveal the hallucination success of existing works and encourage future research to focus on this more challenging and practical scenario.

Acknowledgments

We thank anonymous reviewers for their valuable feedback. In addition, we thank Xin Zhang for his editorial comments. This work is supported by the National Key R&D Program of China (No.2022YFB2703301), NSFC Projects (Nos. 92370124, 62076147). Y. Dong is also supported by the China National Postdoctoral Program for Innovative Talents and Shuimu Tsinghua Scholar Program.

References

[1] Fan Bao, Shen Nie, Kaiwen Xue, Chongxuan Li, Shi Pu, Yaole Wang, Gang Yue, Yue Cao, Hang Su, and Jun Zhu. One transformer fits all distributions in multi-modal diffusion at scale. In *International Conference on Machine Learning*, pages 1692–1717. PMLR, 2023.

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

[3] BIGWILLY. *Heart of Apple XL*, 2024. https://civitai.com/models/272440/heart-of-apple-xl-love.

[4] 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.

[5] 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.

[6] 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.

[7] Huanran Chen, Yinpeng Dong, Zhengyi Wang, Xiao Yang, Chengqi Duan, Hang Su, and Jun Zhu. Robust classification via a single diffusion model. *arXiv preprint arXiv:2305.15241*, 2023.

[8] Huanran Chen, Yinpeng Dong, Shitong Shao, Zhongkai Hao, Xiao Yang, Hang Su, and Jun Zhu. Your diffusion model is secretly a certifiably robust classifier. *arXiv preprint arXiv:2402.02316*, 2024.

[9] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[10] 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.

[11] CompVis. Stable-Diffusion-v1-4. 2024. URL https://huggingface.co/CompVis/stable-diffusion-v1-4.

[12] Yingqian Cui, Jie Ren, Yuping Lin, Han Xu, Pengfei He, Yue Xing, Wenqi Fan, Hui Liu, and Jiliang Tang. Ft-shield: A watermark against unauthorized fine-tuning in text-to-image diffusion models. *arXiv preprint arXiv:2310.02401*, 2023.

[13] Debeshee Das, Jie Zhang, and Florian Tramèr. Blind baselines beat membership inference attacks for foundation models. *arXiv preprint arXiv:2406.16201*, 2024.

[14] Daniel DeAlcala, Aythami Morales, Gonzalo Mancera, Julian Fierrez, Ruben Tolosana, and Javier Ortega-Garcia. Is my data in your ai model? membership inference test with application to face images. *arXiv preprint arXiv:2402.09225*, 2024.

[15] Jinhao Duan, Fei Kong, Shiqi Wang, Xiaoshuang Shi, and Kaidi Xu. Are diffusion models vulnerable to membership inference attacks? In *International Conference on Machine Learning*, pages 8717–8730. PMLR, 2023.

[16] Jan Dubiński, Antoni Kowalczuk, Stanisław Pawlak, Przemyslaw Rokita, Tomasz Trzciński, and Paweł Morawiecki. Towards more realistic membership inference attacks on large diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4860–4869, 2024.

[17] Wenjie Fu, Huandong Wang, Chen Gao, Guanghua Liu, Yong Li, and Tao Jiang. A probabilistic fluctuation based membership inference attack for generative models. *arXiv preprint arXiv:2308.12143*, 2023.

[18] Juliana Neelbauer Gil Appel and David A. Schweidel. Generative AI Has an Intellectual Property Problem. 2023. URL https://hbr.org/2023/04/generative-ai-has-an-intellectual-property-problem.

[19] Jamie Hayes, Luca Melis, George Danezis, and Emiliano De Cristofaro. Logan: Membership inference attacks against generative models. *Proceedings on Privacy Enhancing Technologies*, 2019.

 https://doi.org/10.52202/079017-2358

[20] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[21] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

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

[23] Emiel Hoogeboom, Jonathan Heek, and Tim Salimans. simple diffusion: End-to-end diffusion for high resolution images. In *International Conference on Machine Learning*, pages 13213–13232. PMLR, 2023.

[24] Huggingface. *The training script of stable-diffusion*, 2024. URL https://huggingface.co/docs/diffusers/training/text2image#launch-the-script. Accessed: May 22, 2024.

[25] Huggingface. *The python code of fine-tuning stable-diffusion*, 2024. https://github.com/huggingface/diffusers/blob/main/examples/text_to_image/train_text_to_image.py.

[26] Jinhaoduan. Secmi, 2023. URL https://github.com/jinhaoduan/SecMI/blob/main/mia_evals/secmia.py.

[27] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2426–2435, 2022.

[28] Fei Kong, Jinhao Duan, RuiPeng Ma, Heng Tao Shen, Xiaofeng Zhu, Xiaoshuang Shi, and Kaidi Xu. An efficient membership inference attack for the diffusion model by proximal initialization. In *The Twelfth International Conference on Learning Representations*, 2024.

[29] Kong13661. Pia, 2023. URL https://github.com/kong13661/PIA.

[30] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[31] DeepFloyd Lab. Deepfloyd if. https://github.com/deep-floyd/IF, 2023.

[32] Lambda. Pokemon-blip-captions. 2023. URL https://huggingface.co/datasets/lambdalabs/pokemon-blip-captions.

[33] Alexander C Li, Mihir Prabhudesai, Shivam Duggal, Ellis Brown, and Deepak Pathak. Your diffusion model is secretly a zero-shot classifier. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2206–2217, 2023.

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

[35] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[36] Kin Sum Liu, Chaowei Xiao, Bo Li, and Jie Gao. Performing co-membership attacks against deep generative models. In *2019 IEEE International Conference on Data Mining (ICDM)*, pages 459–467. IEEE, 2019.

[37] David Lopez-Paz and Maxime Oquab. Revisiting classifier two-sample tests. *arXiv preprint arXiv:1610.06545*, 2016.

[38] Tomoya Matsumoto, Takayuki Miura, and Naoto Yanai. Membership inference attacks against diffusion models. In *2023 IEEE Security and Privacy Workshops (SPW)*, pages 77–83. IEEE, 2023.

[39] Yuantian Miao, Minhui Xue, Chao Chen, Lei Pan, Jun Zhang, Benjamin Zi Hao Zhao, Dali Kaafar, and Yang Xiang. The audio auditor: User-level membership inference in internet of things voice services. *Proceedings on Privacy Enhancing Technologies*, 2021.

[40] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

[41] Yan Pang and Tianhao Wang. Black-box membership inference attacks against fine-tuned diffusion models. *arXiv preprint arXiv:2312.08207*, 2023.

[42] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

[43] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

[44] Reuters. Lawsuits accuse AI content creators of misusing copyrighted work. 2023. URL https://www.reuters.com/legal/transactional/lawsuits-accuse-ai-content-creators-misusing-copyrighted-work-2023-01-17/.

[45] Reuters. Getty images lawsuit says stability ai misused photos to train ai. 2023. URL https://www.reuters.com/legal/getty-images-lawsuit-says-stability-ai-misused-photos-train-ai-2023-02-06/.

[46] 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.

[47] Runwayml. Stable-Diffusion-v1-5. 2024. URL https://huggingface.co/runwayml/stable-diffusion-v1-5.

[48] 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.

[49] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

[50] Hadi Salman, Alaa Khaddaj, Guillaume Leclerc, Andrew Ilyas, and Aleksander Madry. Raising the cost of malicious ai-powered image editing. In *International Conference on Machine Learning*, pages 29894–29918. PMLR, 2023.

[51] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*, 2022.

[52] Shawn Shan, Jenna Cryan, Emily Wenger, Haitao Zheng, Rana Hanocka, and Ben Y Zhao. Glaze: Protecting artists from style mimicry by {Text-to-Image} models. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 2187–2204, 2023.

[53] 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.

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

[55] Raphael Tang, Linqing Liu, Akshat Pandey, Zhiying Jiang, Gefei Yang, Karun Kumar, Pontus Stenetorp, Jimmy Lin, and Ferhan Türe. What the daam: Interpreting stable diffusion using cross attention. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5644–5659, 2023.

[56] Zhenting Wang, Chen Chen, Lingjuan Lyu, Dimitris N Metaxas, and Shiqing Ma. Diagnosis: Detecting unauthorized data usages in text-to-image diffusion models. In *The Twelfth International Conference on Learning Representations*, 2023.

[57] Zhijie Wang, Yuheng Huang, Da Song, Lei Ma, and Tianyi Zhang. Promptcharm: Text-to-image generation through multi-modal prompting and refinement. *arXiv preprint arXiv:2403.04014*, 2024.

[58] 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/.

 https://doi.org/10.52202/079017-2358

[59] Yuxin Wen, Yuchen Liu, Chen Chen, and Lingjuan Lyu. Detecting, explaining, and mitigating memorization in diffusion models. In *The Twelfth International Conference on Learning Representations*, 2024.

[60] Jonathan Whitaker. *Fine-tuning a CLOOB-Conditioned Latent Diffusion Model on WikiArt*, 2024. https://johnowhitaker.dev/dsc/2022-04-12-fine-tuning-a-cloob-conditioned-latent-diffusion-model-on-wikiart.html.

[61] WikiArt. WikiArt. 2024. URL https://www.wikiart.org/.

[62] Yixin Wu, Ning Yu, Zheng Li, Michael Backes, and Yang Zhang. Membership inference attacks against text-to-image generation models. *arXiv preprint arXiv:2210.00968*, 2022.

[63] Qiantong Xu, Gao Huang, Yang Yuan, Chuan Guo, Yu Sun, Felix Wu, and Kilian Weinberger. An empirical study on evaluation metrics of generative adversarial networks. *arXiv preprint arXiv:1806.07755*, 2018.

[64] Shih-Ying Yeh, Yu-Guan Hsieh, Zhidong Gao, Bernard BW Yang, Giyeong Oh, and Yanmin Gong. Navigating text-to-image customization: From lycoris fine-tuning to model evaluation. In *The Twelfth International Conference on Learning Representations*, 2023.

[65] 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.

[66] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.

[67] Shengfang Zhai, Yinpeng Dong, Qingni Shen, Shi Pu, Yuejian Fang, and Hang Su. Text-to-image diffusion models can be easily backdoored through multimodal data poisoning. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 1577–1587, 2023.

[68] Yunqing Zhao, Tianyu Pang, Chao Du, Xiao Yang, Ngai-Man Cheung, and Min Lin. A recipe for watermarking diffusion models. *arXiv preprint arXiv:2303.10137*, 2023.

[69] Zhengyue Zhao, Jinhao Duan, Xing Hu, Kaidi Xu, Chenan Wang, Rui Zhang, Zidong Du, Qi Guo, and Yunji Chen. Unlearnable examples for diffusion models: Protect data from unauthorized exploitation. *arXiv preprint arXiv:2306.01902*, 2023.

A Validation of Assumption 3.1

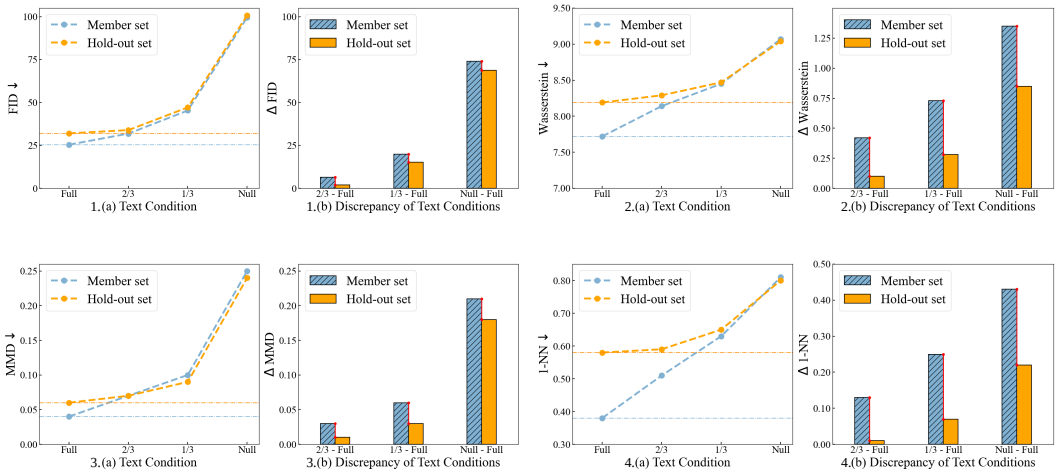

Figure A.1: Metric values and the metric differences of synthetic images, with the same setting as Sec. 3.2.

To extensively validate the effectiveness of Assumption 3.1, we utilize additional metrics as the distances metric D in Eq. (8), including *Wasserstein Distance* [63], *Kernel MMD (Maximum Mean Discrepancy)* [63] and *1-Nearest Neighbor Classifier (1-NN)* [37], in addition to FID [20]. As observed in Fig. A.1, regardless of the metric used for D, Assumption 3.1 consistently holds, thereby confirming the broad applicability of **Conditional Overfitting** phenomenon.

B Proof of Theorem 3.2

Proof. Eq. (8) is equivalent to:

$$\begin{aligned} &\mathbb{E}_{\mathbf{c}}[D(q_{\text{out}}(\mathbf{x}|\mathbf{c}), p(\mathbf{x}|\mathbf{c}))] - D(q_{\text{out}}(\mathbf{x}), p(\mathbf{x})) \\ \geq &\mathbb{E}_{\mathbf{c}}[D(q_{\text{mem}}(\mathbf{x}|\mathbf{c}), p(\mathbf{x}|\mathbf{c}))] - D(q_{\text{mem}}(\mathbf{x}), p(\mathbf{x})). \end{aligned} \tag{B.1}$$

Given that both the member set and the hold-out set are mixtures of Dirac distributions:

$$q(\mathbf{x}) = \frac{1}{|D_{set}|} \sum_{\mathbf{x}_i \in D_{set}} \delta(\mathbf{x} - \mathbf{x}_i), \tag{B.2}$$

where D_{set} denotes the set of images in the corresponding dataset. We can derive the analytical form for the Kullback-Leibler (KL) KL divergence when using D_{KL} as the distance metric:

$$\begin{aligned} D_{KL}(q(\mathbf{x}), p(\mathbf{x})) &= \int q(\mathbf{x}) \log \frac{q(\mathbf{x})}{p(\mathbf{x})} d\mathbf{x} \\ &= -\int (\frac{1}{|D_{set}|} \sum_{\mathbf{x}_i \in D_{set}} \delta(\mathbf{x} - \mathbf{x}_i)) \log p(\mathbf{x}) dx + H(q(\mathbf{x})) \\ &= -\frac{1}{|D_{set}|} \sum_{\mathbf{x}_i \in D_{set}} \int \delta(\mathbf{x} - \mathbf{x}_i) \log p(\mathbf{x}) dx + H(q(\mathbf{x})) \\ &= -\frac{1}{|D_{set}|} \sum_{x_i \in D_{set}} \log p(x_i) + H(q(\mathbf{x})). \end{aligned} \tag{B.3}$$

where H is the entropy functional. Therefore, we have:

$$\begin{aligned} &\mathbb{E}_{\mathbf{c}}[D_{KL}(q(\mathbf{x}|\mathbf{c}), p(\mathbf{x}|\mathbf{c}))] - D_{KL}(q(\mathbf{x}), p(\mathbf{x})) \\ = &-\frac{1}{|D_{set}|} \sum_{\mathbf{x},\mathbf{c} \in D_{set}} [\log p(\mathbf{x}|\mathbf{c}) - \log p(\mathbf{x})] + \mathbb{E}_{\mathbf{c}}[H(q(\mathbf{x}|\mathbf{c}))] - H(q(\mathbf{x})), \end{aligned} \tag{B.4}$$

https://doi.org/10.52202/079017-2358

where D_{set} is the corresponding dataset (member set or hold-out set). Substituting Eq. (B.4) into Eq. (B.1), we can get:

$$\begin{aligned} &-\frac{1}{|D_{\text{out}}|}\sum_{\mathbf{x},\mathbf{c}\in D_{\text{out}}}[\log p(\mathbf{x}|\mathbf{c})-\log p(\mathbf{x})]+\mathbb{E}_{\mathbf{c}}[H(q_{\text{out}}(\mathbf{x}|\mathbf{c}))]-H(q_{\text{out}}(\mathbf{x})) \\ \geq &-\frac{1}{|D_{\text{mem}}|}\sum_{\mathbf{x},\mathbf{c}\in D_{\text{mem}}}[\log p(\mathbf{x}|\mathbf{c})-\log p(\mathbf{x})]+\mathbb{E}_{\mathbf{c}}[H(q_{\text{mem}}(\mathbf{x}|\mathbf{c}))]-H(q_{\text{mem}}(\mathbf{x})). \end{aligned} \tag{B.5}$$

Eq. (B.5) is equivalent to:

$$\begin{aligned} &-\mathbb{E}_{q_{\text{out}}(\mathbf{x},\mathbf{c})}[\log p(\mathbf{x}|\mathbf{c})-\log p(\mathbf{x})]+\mathbb{E}_{\mathbf{c}}[H(q_{\text{out}}(\mathbf{x}|\mathbf{c}))]-H(q_{\text{out}}(\mathbf{x})) \\ \geq &-\mathbb{E}_{q_{\text{mem}}(\mathbf{x},\mathbf{c})}[\log p(\mathbf{x}|\mathbf{c})-\log p(\mathbf{x})]+\mathbb{E}_{\mathbf{c}}[H(q_{\text{mem}}(\mathbf{x}|\mathbf{c}))]-H(q_{\text{mem}}(\mathbf{x})). \end{aligned} \tag{B.6}$$

Finally, we can get:

$$\mathbb{E}_{q_{\text{mem}}(\mathbf{x},\mathbf{c})}[\log p(\mathbf{x}|\mathbf{c})-\log p(\mathbf{x})]\geq\mathbb{E}_{q_{\text{out}}(\mathbf{x})}[\log p(\mathbf{x}|\mathbf{c})-\log p(\mathbf{x})]+\delta_H, \tag{B.7}$$

where

$$\delta_H=H(q_{\text{out}}(\mathbf{x}))+\mathbb{E}_{\mathbf{c}}[H(q_{\text{mem}}(\mathbf{x}|\mathbf{c}))]-H(q_{\text{mem}}(\mathbf{x}))-\mathbb{E}_{\mathbf{c}}[H(q_{\text{out}}(\mathbf{x}|\mathbf{c}))]. \tag{B.8}$$

□

C Metrics Discussion of Sec. 3.3

In the derivation of Sec. 3.3, we also consider other metrics besides KL divergence. The results show that KL divergence yields the most easily computable analytical form. For instance, we briefly discuss Jensen–Shannon (JS) divergence as follows:

Recall the expression for Jensen-Shannon divergence:

$$D_{JS}(q,p)=D_{KL}(q,\frac{1}{2}(q+p))+D_{KL}(p,\frac{1}{2}(q+p). \tag{C.1}$$

The first parameter of the KL divergence should be a simple distribution that is easy to compute; otherwise, deriving the analytical form for such divergence is typically difficult. In Eq (8), JS divergence cannot be efficiently computed because it includes $D_{KL}(p,\frac{1}{2}(q+p))$, where p denotes the model distribution. It needs to use the Monte Carlo method, which involves sampling images from both q and p to make an approximation. As a result, this process is extremely time-consuming.

D Experiment Details

D.1 Monte Carlo Sampling

In our method, the key to accurate membership inference lies in estimating ELBO with fewer sampling steps for better precision. To achieve this, firstly, we reduce the number of Monte Carlo samples by directly estimating the ELBO difference (Eq. (13)). Secondly, recalling Monte Carlo sampling using (t_i,ϵ_i) pairs with $\epsilon_i\sim\mathcal{N}(0,\mathbf{I})$ and $t_i\sim[1,1000]$, we explore the effect of the sampling time t_i. We conduct a single Monte Carlo sampling test using MS-COCO on *real-word training* setting and report the AUC values in Fig. D.1.

In Fig. D.1, we observe that the single Monte Carlo estimation achieves optimal accuracy when $t_i\in[400,500]$. Similar results are shown in [33]. Therefore, consistent with [33], we sample at intervals of 10 centered around the timestep 450. In our experiments, M,N in Eq (14) are both uniformly set to 3 (i.e., the estimation number is 3), and we use the time list of $[440,450,460]$, resulting in the query count of 15. Note that [5, 15] indicate that for DDPM of Cifar-10 [30], the best estimation timestep is around 100. This difference may arise from the different signal-to-noise ratios of images with various resolution [23]. This finding suggests that the Monte Carlo sampling timestep should be designed differently for diffusion models of different scales.

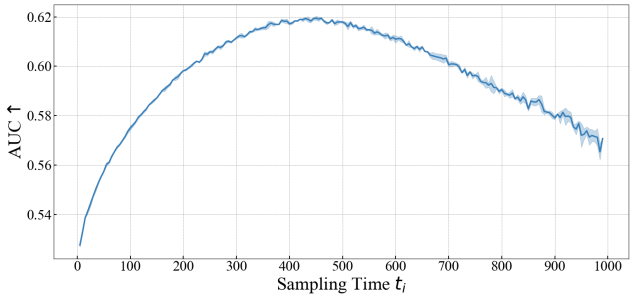

Figure D.1: Effectiveness of single Monte Carlo estimation of various timesteps. Small t_i corresponds to less noise added, and large t_i corresponds to significant noise. AUC value is highest when the timestep is around 450.

Table D.1: The membership inference performance with different reduction methods. "Null" denotes employing null text solely to compute Eq. (14) without reduction methods.

Reduction Methods	CLiD$_{th}$ on MS-COCO			Query
	ASR	AUC	TPR@1%FPR	
Null (K=1)	85.10	93.60	42.96	6
Simply Clipping (K=4)	88.02	95.90	66.53	15
Gaussian Noise (K=4)	86.58	94.79	56.78	15
Importance Clipping (K=4)	**88.88**	**96.13**	**67.52**	15

D.2 Reduction Methods

In implementation, we actually diversely reduce the condition $\mathbf{c}$ to $\mathbf{c}^*$ and calculate $p_\theta(\mathbf{x}|\mathbf{c}^*)$ to approximate $p_\theta(\mathbf{x})$. In this part, we evaluate the effectiveness of different reduction methods. We consider three methods in Sec. 3.5: (1) Simply Clipping. We simply use the first, middle, and last thirds of the sentences as text inputs. (2) Gaussian Noise. We add Gaussian noises with the scales of $50\%, 70\%, 90\%$ to the overall text embeddings. (3) Importance Clipping. We calculate the importance of words in the text[7] [55, 57] and replace them with "pad" tokens in descending order by varying proportions of $30\%, 50\%, 70\%$. For all three methods, we additionally use the null text as a $\mathbf{c}^*$. The experiments are conducted on the *real-world training* setting with MS-COCO dataset. And we also employ null text solely to compute Eq. (14) without reduction methods for comparison.

In Tab. D.1, we observe that *Importance Clipping* achieves the best results due to its more deterministic reduction. So we adopt it as the reduction method used in our experiments. Additionally, we note that all three reduction methods exhibit satisfactory results, demonstrating the general applicability of our method. Comparing the results without the usage of reduction methods, the results validate the effectiveness of reduction methods in Sec. 3.5.

E Compute Overhead and Resources

Computational Overhead. As a query-based member inference method, the computational efficiency of our method primarily depends on the number of queries. A lower query count signifies a more efficient member inference method. Our method significantly outperforms the baselines when the query count are about the same (such as SecMI and PFAMI in Sec. 4.2). Furthermore, even with a much lower query count such as $M = 1, N = 1 (Q = 5)$ (Fig. 3), our method exhibits a noticeable improvement compared to the baselines.

Compute Resources. Our experiments are divided into two main parts: training (fine-tuning) and inference, both conducted on a single RTX A6000 GPU. The time of execution in the training phase depends on the training steps. For example, we perform $7,500$, $50,000$, and $200,000$ steps for Pokemon [32], MS-COCO [35] and Flickr [66] dataset, which take about 2 hours, 12 hours, and 48

[7] https://github.com/ma-labo/PromptCharm

https://doi.org/10.52202/079017-2358

hours, respectively. The time of execution in inference time depends on the methods' query count. For example, with the query count of 15, our membership inference method on a dataset of size $2500/2500$ takes approximately 80 minutes per run for all data points. Typically, we perform this inference once on the shadow model and once on the target model, resulting in a total time cost of 160 minutes.

F Ethics Statements

Although the current threat models for membership inference methods include privacy attack scenarios and data auditing scenarios, we emphasize that for text-to-image diffusion models, the potential application of membership inference lies more in unauthorized data usage auditing than in data privacy leakage. This is because most training data is obtained by scraping open-source image-text pairs, which are more likely to pose copyright threats rather than privacy violations. So we emphasize that our method can make a positive societal impact for inspiring unauthorized usage auditing technologies of text-image datasets in the community.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: In the abstract and introduction, we state that we focus on the membership inference on text-to-image diffusion models in the scenario of auditing unauthorized usage.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: See Sec. 6

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

https://doi.org/10.52202/079017-2358

Justification: Our Assumption is provided in Sec. 3.2. The Proof is provided in Appendix B.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide all the implementation details and the hyper-parameters. We also provide the code. All the models and utilized datasets are open-soured which make it easy to reproduce.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: All the models and utilized datasets are open-soured in this paper. And we provide the code in supplemental material.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: See Sec. 4.1. When training models, we utilize the open-sourced official training scripts by Huggingface with the default hyperparameters, type of optimizer, etc.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: In our main experiments, due to the **high computational cost of repeatedly fine-tuning diffusion models**, we fine-tune only once under different settings for the MS-COCO and Flickr datasets and then evaluate various member inference methods. Given the large dataset sizes, the variance between multiple runs is minimal, and even single-run training introduces negligible errors, thereby reducing carbon emissions. For the smaller Pokemon dataset, we conducted experiments three times and reported the average results in Tab. 1 and Tab. 2

Guidelines:

- The answer NA means that the paper does not include experiments.

https://doi.org/10.52202/079017-2358

- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide extensive discussion in Appendix E.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: See Appendix F.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: See Appendix F.

    Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: We utilize only open-source models and datasets. Additionally, the current text-to-image diffusion models are also trained on open-source datasets. Therefore, our method is solely applicable for detecting unauthorized dataset usage and does not pose a high risk for misuse.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We correctly cite the utilized models and datasets in the paper.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.

https://doi.org/10.52202/079017-2358

- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

https://doi.org/10.52202/079017-2358