Noise as a Probe: Membership Inference Attacks on Diffusion Models Leveraging Initial Noise

Puwei Lian
220255555@seu.edu.cn
Southeast University
Nanjing, China

Yujun Cai
yujun.cai@uq.edu.au
The University of Queensland
Brisbane, Australia

Songze Li*
songzeli@seu.edu.cn
Southeast University
Nanjing, China
Engineering Research Center of Blockchain Application, Supervision and Management (Southeast University), Ministry of Education
Nanjing, China

Bingkun Bao
bingkunbao@hfut.edu.cn
Hefei University of Technology
Heifei, China

arXiv:2601.21628v2 [cs.CR] 13 Aug 2026

Abstract

Diffusion models have achieved remarkable progress in image generation, but their increasing deployment raises serious concerns about privacy and copyright. In particular, fine-tuned models are highly vulnerable, as they are often fine-tuned on small and private datasets. Membership inference attacks (MIAs) are used to assess privacy risks by determining whether a specific sample was part of a model's training data. Existing MIAs against diffusion models either assume obtaining the intermediate results or require training the shadow model on auxiliary datasets. In this work, we utilized a critical yet overlooked vulnerability: the widely used noise schedules fail to fully eliminate semantic information in the images, resulting in residual semantic signals even at the maximum noise step. We empirically demonstrate that the fine-tuned diffusion model captures hidden correlations between the residual semantics in initial noise and the original images. Building on this insight, we propose a simple yet effective membership inference attack, which injects semantic information into the initial noise and infers membership by analyzing the model's generation result. Extensive experiments demonstrate that the semantic initial noise can strongly reveal membership information, highlighting the vulnerability of diffusion models to MIAs. Code is available at https://github.com/S3IC-Lab/NoiseMIA.

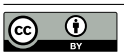
CCS Concepts

• **Security and privacy** → **Social aspects of security and privacy**; • **Computing methodologies** → *Computer vision.*

Keywords

Diffusion Model; Membership Inference Attacks; Privacy

*Corresponding author.



This work is licensed under a Creative Commons Attribution 4.0 International License.
MM '26, Rio de Janeiro, Brazil
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2213-4/2026/11
https://doi.org/10.1145/3767308.3836139

ACM Reference Format:
Puwei Lian, Yujun Cai, Songze Li, and Bingkun Bao. 2026. Noise as a Probe: Membership Inference Attacks on Diffusion Models Leveraging Initial Noise. In *Proceedings of the 34th ACM International Conference on Multimedia (MM '26), November 10–14, 2026, Rio de Janeiro, Brazil.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3767308.3836139

1 Introduction

Diffusion models have shown outstanding performance in generating high-quality images. With the widespread release of large-scale pre-trained models, users can easily download and fine-tune them on downstream datasets. However, this convenience raises serious concerns about the copyright and privacy of training data [33, 45]. Membership inference attacks (MIAs) are crucial methods for assessing the privacy risks. These attacks typically exploit the model's overfitting behavior, leveraging the differences in its responses to training versus non-training samples to infer membership status [48]. In the context of diffusion models, MIAs aim to determine whether an image was used for training [38]. In particular, fine-tuning is widely regarded as the stage most prone to privacy leakage, since the datasets used are relatively small and often private (e.g., personal photos, proprietary artwork) [20, 30]. Consequently, studying MIAs on fine-tuned diffusion models is significant for understanding and mitigating potential privacy risks.

Most existing studies [8, 15, 19, 27, 50] assumed the adversary can manipulate the intermediate denoising network (denoted as intermediate result attacks). However, this assumption is unrealistic in real-world diffusion systems, which typically expose only end-to-end generation interfaces. Previous end-to-end attacks [30, 47] assumed the availability of auxiliary data drawn from the same distribution as the fine-tuning data, which is used to train shadow models and classifiers. However, such attacks depend heavily on the quality of the auxiliary data, and training shadow models and classifiers incurs substantial computational overhead.

We identify a fundamental vulnerability that enables end-to-end attacks without training shadow models or classifiers: standard noise schedules fail to fully eliminate semantic information from



Figure 1: Visualization of generation results. In our method, the generated images of members are clearly closer to their original images, while the non-members differ significantly from their originals.

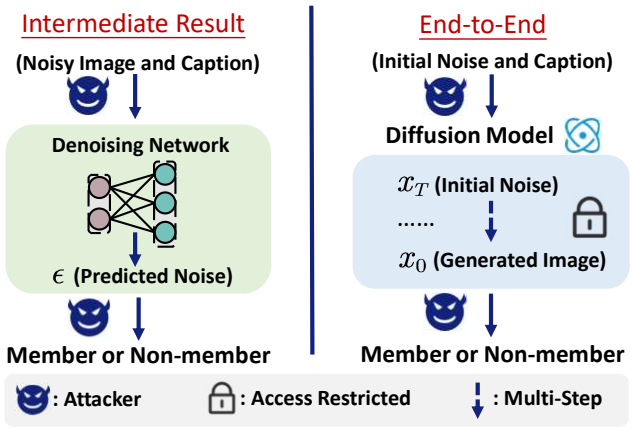


Figure 2: Left: the intermediate result attacks, where the adversary supplies inputs to the denoising network and performs attacks based on its predictions. *Note: the diffusion model includes the denoising network, scheduler, and other components.* Right: the end-to-end attacks, where the adversary provides inputs to the diffusion model and performs attacks based on the final generation.

images. Analyzing the widely used schedules, we find the signal-to-noise ratio (SNR) at the maximum noise timestep T remains non-zero (Tab. 2), leaving residual semantic signals **(Observation 1)**. More critically, via DDIM inversion [6], we find that diffusion models inadvertently learn to exploit these residual signals during training, establishing hidden correlations between initial noise and training data, as evidenced by reconstruction fidelity (Tab. 3) and cross-attention analysis (Fig. 4) **(Observation 2)**.

This presents an exploitable attack opportunity: if we can inject an image's semantic information into the initial noise, the model's generation behavior may reveal whether the image belongs to the training set. The key challenge is that adversaries cannot access the target model's denoising network to perform inversion. Fortunately, we observe that fine-tuned models preserve the semantic space of their pre-trained counterparts (Fig. 4 and Tab. 4), enabling us to use publicly available pre-trained models for semantic injection via DDIM inversion **(Observation 3)**.

Building on these insights, we propose leveraging a pre-trained model to inject semantic information into the initial noise through a DDIM inversion procedure and determine membership by examining the model's generation from semantic initial noise. When the model generates images from this semantic noise, members produce outputs significantly closer to their originals than non-members (Fig. 1). This method is an end-to-end attack that requires no access to the target model's parameters or intermediate denoising network. We compare the conditions of our method with those of previous intermediate result attacks, as illustrated in Fig. 2. Furthermore, our method does not need to train shadow models or classifiers. We summarize our contributions as follows:

- To the best of our knowledge, this is the first study to explore the role of initial noise in MIAs against diffusion models. Our key observations indicate that the diffusion model inadvertently captures hidden correlations between the initial noise and the training data, which serves as a crucial indicator for revealing membership information.
- Building on this insight, we propose a simple yet effective membership inference attack that leverages the inversion procedure to obtain semantic initial noise. The attack analyzes the model's generation from initial noise containing the original image semantics, with membership determined by the similarity between the generated image and the original image.
- Extensive experiments across various datasets validate the effectiveness of our method, with an Area Under the Curve (AUC) of 90.46% and a True Positive Rate at 1% False Positive Rate (T@F=1%) of 21.80%. It demonstrates that the initial noise can strongly expose membership information, revealing the vulnerability of diffusion models to MIAs.

2 Related Work

Membership Inference Attacks. Shokri et al. [38] proposed membership inference attacks (MIAs), which primarily targeted classification models in machine learning. The attacks aim to determine whether a specific piece of data has been included in the training set of the target model. MIAs are typically categorized based on the adversary's access level to the target model. In the white-box setting, the attacker is assumed to have full access to the model parameters [17, 28, 36, 48]. In contrast, black-box attacks assume no access to model parameters. Among them, some methods utilize confidence scores or logits provided by the model [2, 37, 38]. Other researchers have proposed attacks under a more restrictive assumption where only the final predicted labels are available [4, 21, 46].

Membership Inference Attacks on Diffusion Models. The inherent generative mechanism of diffusion models substantially increases the difficulty of membership inference and renders previous approaches largely ineffective [8]. Recently, MIAs on diffusion

models have garnered increasing attention. Pang et al. [31] proposed executing an attack by utilizing gradient information. Recent works [8, 15, 19, 22, 27, 50] assumed that the adversary has access to the model's intermediate denoising process and is allowed to modify the inputs of the denoising networks. This assumption enables queries on the denoising networks' prediction to infer membership information. Some works [30, 47] relied on an auxiliary dataset drawn from the same distribution to train shadow models, and trained a classifier based on the behavioral differences of the shadow models for member and non-member samples. Zhang et al. [51] assumed access to non-member data and constructed member samples through extensive queries for the attack.

Denoising Diffusion Implicit Model (DDIM) Inversion. DDIM Inversion is a technique that utilizes the reverse process of the diffusion model to obtain the initial state of the generated image [6]. In addition, it can be viewed as a noise addition process that integrates semantic information [1, 54]. It adds the predicted noise to the original image, resulting in an initial noise that retains the semantics of the original image. In contrast, naively adding random noise to the image is not connected to the model's understanding of the original image's semantics [52]. Due to the unique advantages of DDIM inversion, it shows great potential in applications such as image quality optimization [1, 54] and image editing [7, 10, 52].

3 Threat Model

MIAs aim to determine whether a specific sample was used during the model training. Formally, let G_θ be a fine-tuned diffusion model with parameters θ. Let D be a dataset drawn from the data distribution q_{data} and each sample x_i in D has a caption $\mathbf{c}_i$. Following established conventions [2, 8, 36], we split D into two subsets D_M and D_N, where D_M denotes the member set used to fine-tune the diffusion model G_θ and D_N denotes the non-member set, so that $D = D_M \cup D_N$ and $D_M \cap D_N = \varnothing$. Each image sample x_i is associated with a membership label m_i, where $m_i = 1$ if $x_i \in D_M$ and $m_i = 0$ otherwise. The adversary has access to the dataset D, but does not know the partition.

Adversary's Goal. The adversary's goal is to design a membership inference attack algorithm $\mathcal{A}$ that, for any given sample x_i, predicts its corresponding membership label:

$$\mathcal{A}(x_i, \theta) = \mathbb{1}\left[\mathbb{P}(m_i = 1 \mid \theta, x_i) \geq \tau\right], \tag{1}$$

where $\mathcal{A}(x_i, \theta) = 1$ means x_i comes from D_M, $\mathbb{1}[A] = 1$ if A is true, and τ is the threshold.

Adversary's Capabilities. In this paper, the adversary is limited to performing end-to-end generation using the diffusion model. Specifically, the adversary can modify the model's initial noise and prompt, and can only observe the final generated image, but has no access to any model parameters or intermediate denoising steps. In the diffusers library [43], pipeline interfaces can accept an initial noise input, which can be adjusted to guide the generation process. Moreover, tasks such as image editing [42, 44, 54], and noise engineering [10, 11, 26] also rely on interfaces that allow the modification of the initial noise. The fine-tuned model is obtained by fine-tuning a pre-trained model on a downstream dataset. Following prior membership inference settings for fine-tuned models [8, 9, 30, 50], we assume that the adversary has access to the pre-trained version of the fine-tuned model.[1] We compare the capabilities of adversaries across different attack algorithms in Tab. 1.

Table 1: Adversary capabilities for different attacks. The top half does not apply to end-to-end generation; the bottom half is feasible. Pm: access to model parameters; Inter: control the inputs to the denoising network at intermediate timesteps; Shadow: train shadow models; Arch: known fine-tuning architecture version; cls: train classifiers. Aux: access to auxiliary dataset. *Symbols*: ✓= required, ✗= not required.

Method	Pm	Inter	Shadow	Arch	Cls	Aux
GSA [31]	✓	✓	✓	✓	✓	✓
Loss [27]	✗	✓	✗	✓	✗	✓
SecMI [8]	✗	✓	✗	✓	✗	✓
PIA [15]	✗	✓	✗	✓	✗	✓
CLiD [50]	✗	✓	✓	✓	✗	✓
NA-P [47]	✗	✗	✓	✓	✓	✓
Feature-T [30]	✗	✗	✓	✗	✗	✓
Feature-C [30]	✗	✗	✓	✓	✓	✓
Feature-D [30]	✗	✗	✓	✓	✓	✓
Ours	✗	✗	✗	✓	✗	✓

4 Methodology

In this section, we present our three key observations systematically. Based on these observations, we propose a membership inference attack that leverages initial noise, which is performed in two steps, as shown in Fig. 3.

4.1 Diffusion Models

Given an original image x_0, the forward diffusion process gradually introduces noise over T timesteps, transforming x_0 into a nearly Gaussian-distributed x_T. The forward diffusion process at each timestep t is described as follows:

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon, \tag{2}$$

where $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$ and $(\alpha_1, \ldots, \alpha_T)$ are the predefined noise schedules, $\epsilon \sim \mathcal{N}(0, I)$ represents Gaussian noise at each step. As t approaches T, $\bar{\alpha}_t$ gradually diminishes, making $x_T \approx \epsilon$ closely resemble pure Gaussian noise. In this paper, we use ϵ_θ to denote the prediction of the denoising network.

4.2 The Semantics in Initial Noise

Observation 1. The widely adopted noise schedules fail to eliminate the semantic information in the original image, even at the maximum noise step.

Explanation 1. We begin with the forward noise addition process in diffusion models. We follow the definition of the signal-to-noise ratio (SNR) in the diffusion model training process from prior work [3]. For a given timestep t, SNR can be characterized as follows:

$$\text{SNR}(t) := \bar{\alpha}_t / (1 - \bar{\alpha}_t). \tag{3}$$

[1] In the following sections, we demonstrate that our method remains effective even when the pre-trained version is unknown.

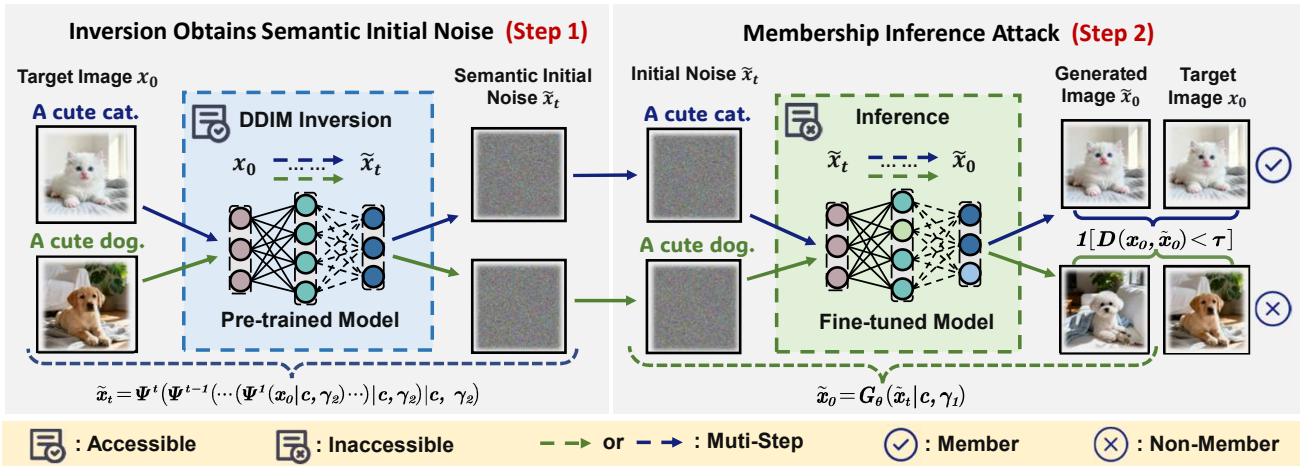


Figure 3: Overview of our method. Step 1: Use a pre-trained model for DDIM inversion to obtain initial noise with semantics. Step 2: Generate images using the noise and determine membership based on the generation results.

We compared the SNR of images at step T across different schedules in Tab. 2. Consistent with the analyses in prior work [23, 44], noise injection at step T cannot eliminate the original signal.

Table 2: Comparison of different noise schedules in the final signal-to-noise ratio SNR(T) and the corresponding $\sqrt{\bar{\alpha}_T}$. The results show that, despite large differences across schedules (Linear [39], Cosine [29], and Stable Diffusion [34]), residual signals consistently remain at the last step.

Schedule	SNR(T)	$\sqrt{\bar{\alpha}_T}$
Linear	4.04e-05	0.006353
Cosine	4.24e-09	0.00004928
Stable Diffusion	4.68e-03	0.068265

The diffusion model's training process can be interpreted as learning a transformation from a Gaussian noise distribution to an image distribution. Previous studies on noise engineering [1, 42, 44, 54] have shown that the initial noise contains semantic information that influences the generation process. Based on the evidence above, we speculate that the semantics of the initial noise may be linked to the training data, potentially revealing membership information. Our subsequent observation further supports this speculation.

Observation 2. The diffusion model inadvertently learns to exploit residual information in the initial noise, thereby establishing a hidden connection between the initial noise and the training data during training.

Explanation 2(a). As discussed in Sec. 2, DDIM inversion can be regarded as a noise injection process that embeds the semantics of the original image into the initial noise. The inversion process Ψ can be expressed as follows:

$$\begin{aligned}\tilde{x}_t = \Psi^t(\tilde{x}_{t-1}|\mathbf{c},\gamma_2) = \sqrt{\bar{\alpha}_t}\tilde{f}_\theta(\tilde{x}_{t-1},t-1) \\ + \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(\tilde{x}_{t-1},t-1),\end{aligned} \tag{4}$$

where $\tilde{f}_\theta(\tilde{x}_{t-1},t-1)$ can be expressed as:

$$\tilde{f}_\theta(\tilde{x}_{t-1},t-1) = \frac{\tilde{x}_{t-1} - \sqrt{1-\bar{\alpha}_{t-1}}\epsilon_\theta(\tilde{x}_{t-1},t-1)}{\sqrt{\bar{\alpha}_{t-1}}}, \tag{5}$$

where $\epsilon_\theta(\tilde{x}_{t-1},t-1) = (1+\gamma_2)\epsilon_\theta(\tilde{x}_{t-1},\mathbf{c},t-1) - \gamma_2\epsilon_\theta(\tilde{x}_{t-1},\varnothing,t-1)$, $\mathbf{c}$ is the text prompt, $\varnothing$ represents the null prompt and γ_2 is the inversion guidance scale. Analyzing the diffusion model's training process and the characteristics of DDIM inversion, we attempt to use the target model for DDIM inversion to obtain semantic initial noise, which can be expressed as:

$$\begin{aligned}\tilde{x}_t &= Inv_\theta^t(x_0|\mathbf{c},\gamma_2) \\ &= \Psi^t\Big(\Psi^{t-1}\big(\cdots(\Psi^1(x_0 \mid \mathbf{c},\gamma_2)\cdots) \mid \mathbf{c},\gamma_2\big) \mid \mathbf{c},\gamma_2\Big).\end{aligned} \tag{6}$$

Then, we use the obtained noise as the starting point for image generation, which can be formulated as:

$$\tilde{x}_0 = G_\theta(\tilde{x}_t|\mathbf{c},\gamma_1). \tag{7}$$

where γ_1 is the guidance scale during generation. We compared the normalized ℓ_2 distance between the original images and those generated from random noise or semantic noise (obtained via inversion). Tab. 3 reports the statistical results for both member and non-member samples across different datasets. The results demonstrate that semantic initial noise leads to higher fidelity reconstructions of the training data. This confirms that diffusion models capture residual information from images, and when conditioned on initial noise containing member semantics, they generate outputs that are closer to the original member images.

Explanation 2(b). Previously, we observed that DDIM inversion enables better reconstruction of member data. We attribute this to the model having learned correlations between residual semantics and the original images during training. To further validate this hypothesis, we analyze the cross-attention of the denoising network during the generation process. Specifically, we visualize the cross-attention heatmaps at the first denoising step, which compute the attention of the token corresponding to the main object in the image. As shown in Fig. 4, when using initial noise obtained via

Table 3: Statistics of normalized ℓ_2 distance across different datasets. *Random*: generate from random noise, *Inversion*: generate from semantic noise, $\Delta = Inversion - Random$.

Method	Pokémon		MS-COCO		Flickr	
	Mem	Non-Mem	Mem	Non-Mem	Mem	Non-Mem
Random	0.3839	0.3961	0.4034	0.4963	0.3543	0.3972
Inversion	0.2779	0.5061	0.2469	0.5394	0.2722	0.4464
Δ	-0.1060	+0.1100	-0.1565	+0.0431	-0.0821	+0.0492

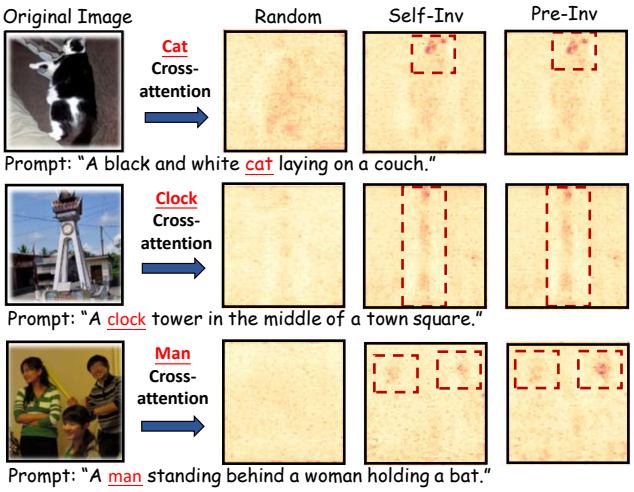


Figure 4: Visualization of cross-attention heatmaps. Heatmaps display the local contributions of the second attention modules in the third upsampling block. Random: generation using random noise; Self-Inv: generation using semantic noise obtained via inversion of the target model; Pre-Inv: generation using semantic noise obtained via inversion of the pre-trained model. The red boxes highlight regions with high attention, which precisely correspond to the locations of the main objects in the original images.

DDIM inversion, the denoising network immediately attends to the semantic regions of the image. In contrast, when initialized with random noise, the network exhibits no clear focus on specific regions. This phenomenon further supports that the model has inadvertently established a hidden connection between the initial noise and the original image.

Although inversion provides semantic initial noise for both member and non-member samples, members can be reconstructed much more faithfully because the model has learned correlations between such semantic cues and the original training images. As a result, the semantic information contained in the initial noise allows the model to better recover the original images of members, making it a discriminative signal for membership inference.

4.3 Semantic Injection via Pre-trained Model

One major challenge is that the adversary cannot directly access the parameters of the target model, making it infeasible to perform DDIM inversion on the target itself. Despite these constraints, we observe a crucial aspect of fine-tuning in the context of membership inference attacks.

Observation 3. Models fine-tuned from pre-trained ones essentially preserve the original semantic space and representational capabilities.

Explanation 3. Previous works [32, 53] suggested that fine-tuning does not substantially alter the original distribution. Based on this characteristic, we hypothesize that the model's understanding and representation of semantic information have not undergone significant changes after fine-tuning. To verify this hypothesis, we perform inversion with the pre-trained version of the target model to obtain semantic initial noise. Following the same procedure as before, we then feed this noise into the target model and examine its cross-attention heatmaps. As shown in Fig. 4, the fine-tuned model still captures the semantics in the initial noise, with high cross-attention similarity to that obtained from self-inversion[2].

Table 4: Similarity of cross-attention heatmaps. Across different datasets and fine-tuning epochs, the cross-attention heatmaps show a remarkably high similarity between the noises obtained from self-inversion and pre-inversion.

Epoch	50	100	150	200	250	300
Pokémon	0.968	0.942	0.926	0.897	0.895	0.892
T-to-I	0.970	0.954	0.924	0.899	0.893	0.893
MS-COCO	0.973	0.953	0.937	0.897	0.895	0.895
Flickr	0.978	0.953	0.934	0.904	0.902	0.903

To further confirm this similarity, we computed the cosine similarity of cross-attention heatmaps between the initial noise obtained from self-inversion and pre-inversion. As shown in Tab. 4, across different fine-tuning epochs, the cross-attention heatmaps of the noise from the two inversions exhibit a high degree of similarity. All the findings provide evidence that the fine-tuned model retains the semantic space of its pre-trained version, making it possible to obtain semantic initial noise using a pre-trained model.

4.4 MIAs Leveraging Initial Noise

Based on the above experiments and observations, we propose a simple yet effective membership inference attack that exploits the correlations between the semantic initial noise and the training data. This attack proceeds in two main steps: obtaining semantic initial noise and conducting the membership inference attack. *We provide the detailed algorithmic procedure in algorithm 1.*

Step 1: Given a target image x_0 and its corresponding text prompt $\mathbf{c}$, we first employ a pre-trained diffusion model to perform DDIM inversion, thereby obtaining a semantic initial noise $\tilde{x}_t$.

Step 2: The adversary feeds the semantic initial noise $\tilde{x}_t$ and the same prompt $\mathbf{c}$ into the target model to generate a candidate image $\tilde{x}_0$. If the target image x_0 was part of the fine-tuning dataset, the generated candidate image tends to preserve its structural and semantic consistency, resulting in a smaller perceptual distance.

[2]For clarity, DDIM inversion with the target model is termed self-inversion, and that with the pre-trained model is termed pre-inversion.

Table 5: AUC and T@F=1% on different datasets. The compared baselines are divided into two categories: intermediate result attacks (*Upper part*) and end-to-end attacks (*Lower part*). In each column, the best performance in the end-to-end attacks is displayed in bold, while the best performance across all attacks is underlined. Our method achieves the best performance in end-to-end attacks and demonstrates performance comparable to that of intermediate result attacks. Notably, unlike intermediate result attacks, our method does not require access to the denoising network's inputs or outputs.

Method	Pokémon		T-to-I		MS-COCO		Flickr		Average	
	AUC	T@F=1%	AUC	T@F=1%	AUC	T@F=1%	AUC	T@F=1%	AUC	T@F=1%
SecMI	83.26	12.88	88.26	26.66	89.37	16.79	76.42	14.40	84.33	17.68
PIA	76.82	7.85	84.80	11.78	71.38	5.20	72.59	7.20	76.40	8.01
NA-P	59.37	4.80	70.77	6.60	52.41	2.20	56.51	4.00	59.77	4.40
GD	52.67	1.20	59.67	5.20	51.22	1.00	53.41	2.00	54.24	2.35
Feature-T	56.60	3.00	70.40	7.00	57.70	3.20	58.20	3.00	60.73	4.05
Feature-C	60.67	7.33	83.77	17.40	73.08	14.60	63.80	5.20	70.33	11.13
Feature-D	55.10	2.80	65.00	6.00	58.00	3.00	57.20	3.00	58.83	3.70
Ours	**82.44**	**14.00**	**89.24**	**21.60**	**90.46**	**21.80**	**76.23**	**16.00**	**84.59**	**18.35**

Algorithm 1 MIAs Leveraging Initial Noise

Input: Target model G_θ, pre-trained model $G_{\theta_{\text{pre-trained}}}$, threshold τ, target data $(x, \mathbf{c})$, distance metric $D(\cdot,\cdot)$, inference guidance scale γ_1, inversion guidance scale γ_2.

// Inversion obtains semantic initial noise. (Step 1)
Perform inversion $\tilde{x}_t = Inv^t_{\theta_{\text{pre-trained}}}(x|\mathbf{c}, \gamma_2)$.
// Membership inference attack. (Step 2)
Generate the image $\tilde{x} = G_\theta(\tilde{x}_t|\mathbf{c}, \gamma_1)$.
Compute the membership score: $Score = D(x, \tilde{x})$.
if $Score < \tau$ **then**
    Conclude that $\mathbb{1} = 1$, i.e., x is a member data.
else
    Conclude that $\mathbb{1} = 0$, i.e., x is a non-member data.
end if

Output: Membership status $\mathbb{1} \in \{0, 1\}$.

Conversely, the non-member samples will yield larger deviations in the generated outputs. The membership inference decision is made by comparing the reconstruction distance using a metric $D(\cdot,\cdot)$. In summary, our method can be summarized as follows:

$$\begin{cases} \tilde{x}_t = Inv^t_{\theta_{pre-trained}}(x_0|\mathbf{c}, \gamma_2), \ (Step\ 1) \\ \mathcal{A}(x_i, \theta) = \mathbb{1}\left[D(x_0, G_\theta(\tilde{x}_t|\mathbf{c}, \gamma_1)) \le \tau\right]. \ (Step\ 2) \end{cases} \tag{8}$$

where $D(\cdot,\cdot)$ represents the distance metric.

5 Experiments

5.1 Experiment Setup

Datasets and Models. We follow the previous stringent assumption that both member and non-member data are drawn from the same distribution [8, 15, 22, 50]. We construct member/non-member datasets using 416/417 samples from Pokémon [16], 500/500 samples from text-to-image-2M (T-to-I) [14], 2500/2500 samples from MS-COCO [24], and 1000/1000 samples from Flickr [49]. We fine-tune Stable Diffusion-v1-4 (SD-v1-4) [5] using the official fine-tuning scripts from the Hugging-Face Diffusers library [13].

Evaluation Metrics. We adopt the evaluation metrics commonly used in prior membership inference attacks for large models [8, 9, 12, 15, 30]. Specifically, we report the Area Under the Curve (denoted as **AUC**), which reflects the average success of membership inference attacks. In addition, we measure the True Positive Rate (TPR) at 1% False Positive Rate (FPR) (**denoted as T@F=1%**), which assesses attack efficacy under a strict decision threshold, emphasizing performance at an extremely low FPR.

Baselines. We compare our method with existing end-to-end attacks, including NA-P [47], Feature-T [30], Feature-D [30], Feature-C [30], and GD [51]. We also evaluate intermediate result attacks, including SecMI [8] and PIA [15].

Implementation Details. All fine-tuning and inference experiments were conducted on a single RTX 4090 GPU (24 GB). During the DDIM inversion step, we set the guidance scale $\gamma_2 = 1.0$ and the number of steps $i_{\text{step}} = 100$. For the membership inference step, we set the guidance scale $\gamma_1 = 3.5$ and the number of inference steps to 50. We use the ℓ_2 distance as the $D(\cdot,\cdot)$ by default.

5.2 Main Result

Overall Attack Performance. We report comprehensive attack results and compare the performance of our method against all baselines. As shown in Tab. 5, our method consistently achieves superior performance across different datasets, delivering significant improvements over existing end-to-end attacks. In particular, compared to the state-of-the-art (SOTA) end-to-end attack, Feature-C, our method can yield improvements of up to 21.77% in AUC and 11.80% in TPR@1%FPR. Notably, Feature-C requires auxiliary data to train shadow models and classifiers, resulting in significant training overhead. In contrast, our method performs the attack solely through threshold setting, achieving superior performance with a lower computational cost. Moreover, our method remains highly competitive compared to intermediate result attacks, e.g., SecMI and PIA, and outperforms them on the MS-COCO dataset, all while not requiring access to the intermediate results.

Visual Analysis of Attack Performance. To further illustrate the effectiveness of our method, we compare it against Feature-T,

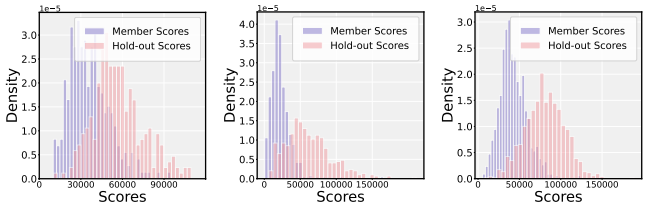
Table 6: Hyperparameter analysis of i_{step} and γ_2. The results show that our method achieves consistently high performance across a wide range of hyperparameter values, demonstrating its robustness.

i_{step}/γ_2	$\gamma_2 = 0.0$		$\gamma_2 = 1.0$		$\gamma_2 = 3.5$		$\gamma_2 = 4.5$		$\gamma_2 = 7.5$	
	AUC	T@F=1%	AUC	T@F=1%	AUC	T@F=1%	AUC	T@F=1%	AUC	T@F=1%
$i_{step} = 25$	89.20	21.40	89.16	21.40	87.36	21.00	87.46	21.20	87.25	21.00
$i_{step} = 50$	87.36	21.20	87.24	21.00	89.14	21.20	89.36	21.60	89.33	21.40
$i_{step} = 100$	88.48	21.00	90.46	21.80	88.56	21.00	88.61	21.00	90.12	21.80
$i_{step} = 200$	89.12	21.20	89.12	21.40	91.00	22.00	91.04	22.00	89.31	21.20

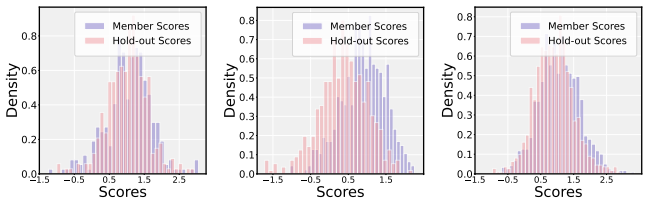
Table 7: Ablation study on different datasets, showing the impact of semantic initial noise on attack performance.

Method	Pokémon		T-to-I		MS-COCO		Flickr	
	AUC	T@F=1%	AUC	T@F=1%	AUC	T@F=1%	AUC	T@F=1%
Naive	55.66	6.50	72.76	10.50	62.24	7.00	57.29	5.00
Ours	82.44	14.00	89.24	21.60	90.46	21.80	76.23	16.00
Gain	**+26.78**	**+7.50**	**+16.48**	**+11.10**	**+28.22**	**+14.80**	**+18.94**	**+11.00**

the previous most potent threshold-based end-to-end attack. As shown in Fig. 5, we visualize the membership score distributions of member and non-member data. The separation between the two distributions achieved by our method is substantially larger than that of Feature-T. This visualization provides intuitive evidence of the enhanced distinguishability achieved by our method.



Membership score distribution of our method.



Membership score distribution of Feature-T.

Figure 5: Membership score distribution of member and non-member data in the Pokémon, T-to-I, and MS-COCO dataset, arranged from left to right. The score distribution gap between member data and hold-out data is significantly larger in our method.

5.3 Ablation Study

Contribution of Semantic Initial Noise. In this section, we define the Naive method, which generates images solely from the given caption using randomly initialized noise. The distance between the generated image and the target image is then measured to determine membership. The ablation results are reported in Tab. 7. On average, the AUC improved by 21.57% across different datasets, and T@F=1% increased by 10.63%. It clearly shows that injecting semantics into the initial noise substantially improves the attack performance. These results validate the effectiveness of our approach and underscore the importance of carefully managing and protecting initial noise.

Ablation on Hyperparameter. We validate the impact of hyperparameters i_{step} and γ_2 on the performance of our method. As shown in Tab. 6, our method demonstrates very low sensitivity to i_{step} and γ_2, with the best and worst AUC being 91.04% and 87.25%, respectively (a variation of only 3.79%). It provides strong confirmation of the robustness of our method.

5.4 A More Knowledge-Restricted Adversary

Without Access to the Model Architecture. In real-world scenarios, model publishers may deliberately withhold the architecture version of the pre-trained model used during fine-tuning, thereby increasing the difficulty of attacks. To evaluate performance under this condition, we use SD-v1-5 [35], SD-2-1 [41], SDXL-turbo [40], and Dreamshaper-XL [25] to perform inversion and obtain initial noise, where the architectural discrepancy from the SD-v1-4 gradually increases. And these models are used as shadow models for the baseline attacks. As shown in Tab. 8, our method remains effective even when the architecture version of the target model is unknown. Although all methods exhibit performance degradation as the architectural gap widens, our method achieves the best performance even using SDXL-turbo and Dreamshaper-XL, whose architectures differ substantially from the target model. This finding aligns with the observations in [44], which indicate that the semantics of the initial noise can be transferred across models due to the shared distributions learned during large-scale pre-training. This further validates that semantic initial noise can be leveraged to reveal membership information.

Table 8: Attack performance using semantic initial noise obtained from different models, demonstrating that our method remains effective when the target model's architecture version is unknown. Intermediate result attacks are not applicable in this scenario.

Method	SD-v1-5		SD-2-1		SDXL-turbo		Dreamshaper	
	AUC	T@F=1%	AUC	T@F=1%	AUC	T@F=1%	AUC	T@F=1%
NA-P	70.53	6.60	69.21	6.60	55.53	2.20	56.02	2.40
Feature-C	83.41	17.40	81.28	16.70	68.41	6.20	67.42	6.00
Feature-D	65.00	6.00	64.59	6.00	58.00	4.00	57.66	4.20
Ours	**89.03**	**20.20**	**88.04**	**19.60**	**76.91**	**8.40**	**76.87**	**8.20**

Table 9: Attack performance under defense mechanisms. All attacks experience varying degrees of performance degradation, while our method consistently achieves the best performance.

SS_{e_i}	DataAug	SecMI		PIA		NA-P		GD		Feature-T		Feature-C		Feature-D		**Ours**	
		AUC	T@F=1%	AUC	T@F=1%	AUC	T@F=1%	AUC	T@F=1%	AUC	T@F=1%	AUC	T@F=1%	AUC	T@F=1%	**AUC**	**T@F=1%**
×	×	89.43	16.99	74.36	6.40	64.77	5.57	51.31	1.00	59.43	4.00	73.26	14.60	59.01	3.60	**91.12**	**22.90**
×	✓	89.37	16.79	71.38	5.20	63.41	5.20	51.22	1.00	57.70	3.20	73.08	14.60	58.00	3.20	**90.46**	**21.80**
✓	×	52.11	2.10	58.93	3.20	62.98	4.00	51.05	1.00	57.53	3.20	71.99	13.40	57.20	3.00	**87.68**	**18.00**
✓	✓	51.21	1.90	54.41	1.50	62.56	4.00	51.03	1.00	57.20	3.20	71.02	13.20	57.50	2.60	**86.74**	**17.90**

Lacking Access to Image Captions. In reality, attackers may not have the image captions used for fine-tuning. Therefore, we also evaluate the attack when the image captions are unavailable. To address this, we employ BLIP [18] to generate captions for the images and use these generated captions to conduct the attack. As shown in Tab. 10, using the MS-COCO dataset as an example, our method outperforms the strongest competitor, Feature-C, with an AUC improvement of 10.66% and a T@F=1% improvement of 2.20%. The experiments reveal that end-to-end attacks are vulnerable to the absence of captions, leading to a noticeable degradation in performance. This is likely because end-to-end attacks involve a multi-step prediction process during image generation, making them more sensitive to changes in the caption. Although our method is also affected under this setting and exhibits a performance drop, it still achieves the best results in end-to-end attacks.

Table 10: Attack performance using BLIP-generated captions. Our method achieves the best performance.

Method	T-to-I		MS-COCO		Flickr	
	AUC	T@F=1%	AUC	T@F=1%	AUC	T@F=1%
NA-P	54.03	1.00	53.00	1.00	52.01	1.00
GD	52.23	1.00	51.03	1.00	50.71	1.00
Feature-T	55.80	1.00	50.20	1.00	50.10	1.00
Feature-C	63.20	4.00	52.88	3.40	50.98	1.00
Feature-D	53.00	2.00	50.20	1.00	50.20	1.00
Ours	**71.47**	**6.80**	**63.54**	**5.60**	**57.82**	**4.80**

5.5 Impact of Defense

To assess the robustness of our method, we investigate the impact of the SOTA defense method SS_{e_i} [45]. This defense dynamically evaluates the model's memorization during training and adjusts the training process accordingly. Following this method, we set the SS_{e_i} threshold to 4. Additionally, data augmentation techniques are commonly employed to mitigate MIAs. During the fine-tuning process of the diffusion model, Random-Crop and Random-Flip are applied by default [13]. Following previous work [8, 31, 50], we also investigate the impact of data augmentation on the performance of attacks. We conduct an in-depth analysis of the performance changes of various attacks before and after applying defenses on the MS-COCO dataset, as shown in Tab. 9. Experimental results show that our method maintains excellent performance even against the advanced defense strategies. In the presence of both defense mechanisms, our method achieves substantially better performance than all the other methods. Specifically, the intermediate result attacks show a significant performance drop under defense. In our method, AUC and TPR@1%FPR decrease by only 4.38% and 5.00%, respectively, compared to performance without any defense.

5.6 Visualization of Generated Results.

We present the generation results for member and non-member samples on the Flickr dataset in Fig. 6. The first three columns display the member images alongside their generated counterparts, obtained using random initial noise and semantic initial noise, respectively. The last three columns present non-member images and their corresponding generated results. In Naive, the generated results, whether from member or non-member samples, significantly deviate from the originals. In contrast, our method generates member images that closely resemble their originals, and the generated non-member images differ noticeably. This clear separation between member and non-member reconstruction quality is the primary reason for the attack performance achieved by our method.

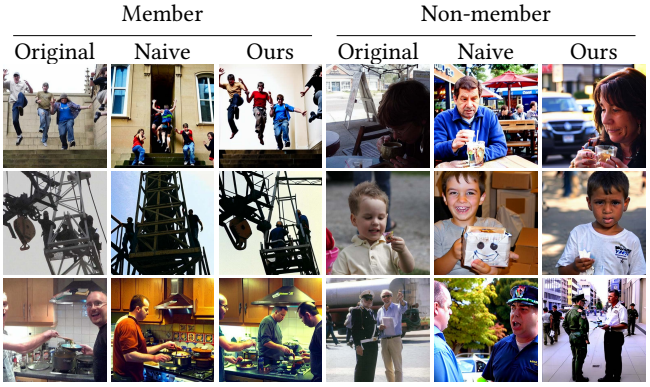


Figure 6: Visualization of generation. In our method, the generated images of members are clearly closer to their originals. In Naive, both the generated member and non-member data differ significantly from their originals.

6 Conclusion

In this paper, we reveal that standard noise schedules in diffusion models retain residual semantic information in the initial noise, leading the model to inadvertently learn correlations between the initial noise and the training images. Leveraging this vulnerability, we propose a simple yet effective membership inference attack that uses DDIM inversion to inject semantics into the initial noise and analyzes the generations. Our experiments confirm that these semantic residuals pose significant privacy risks in fine-tuned models.

7 Acknowledgments

This work is supported in part by New Generation Artificial Intelligence-National Science and Technology Major Project (2025ZD0123504), and in part by the Fundamental Research Funds for the Central Universities (Grant No. 2242026K30047).

References

[1] Lichen Bai, Shitong Shao, Zipeng Qi, Haoyi Xiong, Zeke Xie, et al. 2025. Zigzag Diffusion Sampling: Diffusion Models Can Self-Improve via Self-Reflection. In *The Thirteenth International Conference on Learning Representations*.

[2] Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. 2022. Membership inference attacks from first principles. In *2022 IEEE symposium on security and privacy (SP)*. IEEE, 1897–1914.

[3] Jooyoung Choi, Jungbeom Lee, Chaehun Shin, Sungwon Kim, Hyunwoo Kim, and Sungroh Yoon. 2022. Perception prioritized training of diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 11472–11481.

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

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

[6] Prafulla Dhariwal and Alexander Nichol. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems* 34 (2021), 8780–8794.

[7] Wenkai Dong, Song Xue, Xiaoyue Duan, and Shumin Han. 2023. Prompt tuning inversion for text-driven image editing using diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 7430–7440.

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

[9] Wenjie Fu, Huandong Wang, Chen Gao, Guanghua Liu, Yong Li, and Tao Jiang. 2024. Membership inference attacks against fine-tuned large language models via self-prompt calibration. *Advances in Neural Information Processing Systems* 37 (2024), 134981–135010.

[10] Daniel Garibi, Or Patashnik, Andrey Voynov, Hadar Averbuch-Elor, and Daniel Cohen-Or. 2024. Renoise: Real image inversion through iterative noising. In *European Conference on Computer Vision*. Springer, 395–413.

[11] Xiefan Guo, Jinlin Liu, Miaomiao Cui, Jiankai Li, Hongyu Yang, and Di Huang. 2024. Initno: Boosting text-to-image diffusion models via initial noise optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9380–9389.

[12] Yu He, Boheng Li, Liu Liu, Zhongjie Ba, Wei Dong, Yiming Li, Zhan Qin, Kui Ren, and Chun Chen. 2025. Towards label-only membership inference attack against pre-trained large language models. In *USENIX Security*.

[13] Hugging Face. 2024. Diffusers Text-to-Image Training Script. https://github.com/huggingface/diffusers/tree/main/examples/text_to_image.

[14] jackyhate. 2024. text-to-image-2M. https://huggingface.co/datasets/jackyhate/text-to-image-2M.

[15] Fei Kong, Jinhao Duan, RuiPeng Ma, Heng Tao Shen, Xiaoshuang Shi, Xiaofeng Zhu, and Kaidi Xu. 2024. An Efficient Membership Inference Attack for the Diffusion Model by Proximal Initialization. In *The Twelfth International Conference on Learning Representations*.

[16] Lambda Labs. 2023. Pokémon BLIP Captions. https://huggingface.co/datasets/lambdalabs/pokemon-blip-captions.

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

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

[19] Qiao Li, Xiaomeng Fu, Xi Wang, Jin Liu, Xingyu Gao, Jiao Dai, and Jizhong Han. 2024. Unveiling structural memorization: Structural membership inference attack for text-to-image diffusion models. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 10554–10562.

[20] Zhangheng Li, Junyuan Hong, Bo Li, and Zhangyang Wang. 2024. Shake to leak: Fine-tuning diffusion models can amplify the generative privacy risk. In *2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*. IEEE, 18–32.

[21] Zheng Li and Yang Zhang. 2021. Membership leakage in label-only exposures. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*. 880–895.

[22] Puwei Lian, Yujun Cai, and Songze Li. 2025. Unveiling Impact of Frequency Components on Membership Inference Attacks for Diffusion Models. *arXiv preprint arXiv:2505.20955* (2025).

[23] Shanchuan Lin, Bingchen Liu, Jiashi Li, and Xiao Yang. 2024. Common diffusion noise schedules and sample steps are flawed. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. 5404–5411.

[24] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*. Springer, 740–755.

[25] Lykon. 2024. dreamshaper-xl-v2-turbo. https://huggingface.co/Lykon/dreamshaper-xl-v2-turbo.

[26] Jiafeng Mao, Xueting Wang, and Kiyoharu Aizawa. 2023. Guided image synthesis via initial image editing in diffusion model. In *Proceedings of the 31st ACM International Conference on Multimedia*. 5321–5329.

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

[28] Milad Nasr, Reza Shokri, and Amir Houmansadr. 2019. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In *2019 IEEE symposium on security and privacy (SP)*. IEEE, 739–753.

[29] Alexander Quinn Nichol and Prafulla Dhariwal. 2021. Improved denoising diffusion probabilistic models. In *International conference on machine learning*. PMLR, 8162–8171.

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

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

[32] Evani Radiya-Dixit and Xin Wang. 2020. How fine can fine-tuning be? learning efficient language models. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 2435–2443.

[33] Jie Ren, Yaxin Li, Shenglai Zeng, Han Xu, Lingjuan Lyu, Yue Xing, and Jiliang Tang. 2024. Unveiling and mitigating memorization in text-to-image diffusion models through cross attention. In *European Conference on Computer Vision*. Springer, 340–356.

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

[35] RunwayML. 2024. Stable Diffusion v1-5. https://huggingface.co/stable-diffusion-v1-5/stable-diffusion-v1-5.

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

[37] Ahmed Salem, Yang Zhang, Mathias Humbert, Pascal Berrang, Mario Fritz, and Michael Backes. 2018. Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models. *arXiv preprint arXiv:1806.01246* (2018).

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

[39] Jiaming Song, Chenlin Meng, and Stefano Ermon. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502* (2020).

[40] stabilityai. 2024. sdxl-turbo. https://huggingface.co/stabilityai/sdxl-turbo.

[41] stabilityai. 2024. stable-diffusion-2-1. https://huggingface.co/sd2-community/stable-diffusion-2-1.

[42] Wenqiang Sun, Teng Li, Zehong Lin, and Jun Zhang. 2024. Spatial-aware latent initialization for controllable image generation. *arXiv preprint arXiv:2401.16157* (2024).

[43] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, Dhruv Nair, Sayak Paul, William Berman, Yiyi Xu, Steven Liu, and Thomas Wolf. 2022. Diffusers: State-of-the-art diffusion models.

[44] Ruoyu Wang, Huayang Huang, Ye Zhu, Olga Russakovsky, and Yu Wu. 2024. The silent prompt: Initial noise as implicit guidance for goal-driven image generation. *arXiv e-prints* (2024), arXiv–2412.

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

[46] Yutong Wu, Han Qiu, Shangwei Guo, Jiwei Li, and Tianwei Zhang. 2024. You only query once: An efficient label-only membership inference attack. In *The Twelfth International Conference on Learning Representations*.

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

[48] Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. 2018. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE*

31st computer security foundations symposium (CSF). IEEE, 268–282.

[49] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the association for computational linguistics* 2 (2014), 67–78.

[50] Shengfang Zhai, Huanran Chen, Yinpeng Dong, Jiajun Li, Qingni Shen, Yansong Gao, Hang Su, and Yang Liu. 2024. Membership inference on text-to-image diffusion models via conditional likelihood discrepancy. *Advances in Neural Information Processing Systems* 37 (2024), 74122–74146.

[51] Minxing Zhang, Ning Yu, Rui Wen, Michael Backes, and Yang Zhang. 2024. Generated distributions are all you need for membership inference attacks against generative models. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. 4839–4849.

[52] Yuxin Zhang, Nisha Huang, Fan Tang, Haibin Huang, Chongyang Ma, Weiming Dong, and Changsheng Xu. 2023. Inversion-based style transfer with diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10146–10156.

[53] Yichu Zhou and Vivek Srikumar. 2022. A Closer Look at How Fine-tuning Changes BERT. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1046–1061.

[54] Zikai Zhou, Shitong Shao, Lichen Bai, Shufei Zhang, Zhiqiang Xu, Bo Han, and Zeke Xie. 2024. Golden noise for diffusion models: A learning framework. *arXiv preprint arXiv:2411.09502* (2024).