Noise Aggregation Analysis Driven by Small-Noise Injection: Efficient Membership Inference for Diffusion Models

Guo Li
South China University of Technology
Guangzhou, China
csliguo@mail.scut.edu.cn

Weihong Chen
South China University of Technology
Guangzhou, China
cswilliam@mail.scut.edu.cn

Yongfu Fan
University of Electronic Science and Technology of China
yf.fan@std.uestc.edu.cn

***Abstract*—Diffusion models have demonstrated powerful performance in generating high-quality images. A typical example is text-to-image generator like Stable Diffusion. However, their widespread use also poses potential privacy risks. A key concern is membership inference attacks, which attempt to determine whether a particular data sample was used in the model training process. Existing membership inference attacks against diffusion models either directly exploit sample loss differences or rely on image-level reconstruction differences. Both approaches commonly ignore the consistency characteristics of noise prediction during the diffusion process, resulting in either low inference accuracy or high computational costs. To address these shortcomings, we propose a membership inference method based on noise aggregation analysis, and introduce a single-step, low-intensity noise injection diffusion strategy to amplify differences between member and non-member samples. Our proposed approach substantially reduces model query requirements while delivering more efficient and accurate membership inference.**

***Index Terms*—Diffusion models, membership inference attack, privacy, machine learning security**

I. Introduction

With the rapid development of generative models, high-quality image synthesis based on diffusion models has achieved significant success [1]–[6]. Compared to GANs and VAEs, diffusion models demonstrate superior generalization and semantic consistency due to their unique training schemes [7], [8]. However, this has also raised significant privacy concerns, specifically regarding membership inference attacks (MIA), which aim to determine whether a specific data sample was used during training [9]–[13]. For example, in medical scenarios, MIAs may result in the severe leakage of sensitive patient data; in commercial scenarios, competitors can exploit MIAs to identify proprietary training data. Consequently, investigating MIAs against diffusion models is particularly important for revealing internal mechanisms and facilitating the adoption of stronger privacy-preserving strategies.

Existing MIA methods for diffusion models can be broadly categorized into three types. The first type leverages sample loss differences [14], [15], assuming that member samples yield lower losses. But such methods introduce randomly sampled noise during loss computation. The randomness of the noise leads to instability in the attack results, which ultimately affects the overall attack performance. The second type relies on image-level reconstruction and comparison to determine membership [16], [17], typically employing multi-step denoising strategies of DDIM deterministic sampling to mitigate the influence of random noise. However, this approach significantly increases the number of model queries, resulting in high computational cost. The third category focuses on text-to-image diffusion models [18], [19], utilizing text features that limit their applicability to other types of diffusion architectures.

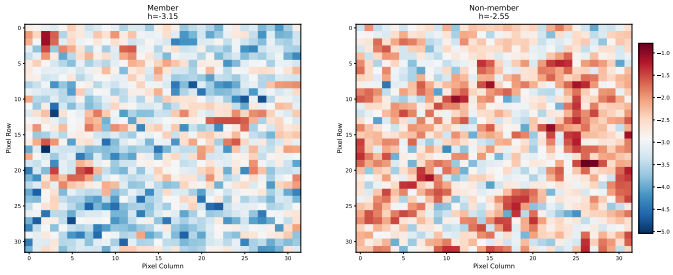


Fig. 1: Conditional differential entropy maps of sample pairs. The left shows a member sample, and the right shows a non-member sample. The non-member sample exhibits noticeably redder regions, indicating substantially higher pixel-level uncertainty.

To overcome these limitations, we revisit diffusion models from a more fundamental perspective. Diffusion models can be naturally viewed as noise-prediction networks trained to estimate the injected noise at each diffusion timestep. From this perspective, overfitting in diffusion models is closely tied to overfitting in the underlying noise predictor, which may lead to memorization of training data. This observation motivates us to analyze the outputs of the noise-prediction network, providing a principled basis for designing effective membership inference attacks against diffusion models. By comparing the pixel-level conditional differential entropy heatmaps of member and non-member samples within a local neighborhood of specific timesteps (as shown in Figure 1),

we observe that non-member samples exhibit substantially higher uncertainty. To further quantify this uncertainty and achieve more accurate membership inference, we introduce a metric method based on noise aggregation that measures the consistency of predicted noise vectors across a sequence of denoising steps around a specific timestep. Our intuition is that, since member samples participate in the training process, The noise-prediction network generates more accurate noise estimates for member samples, and consequently shows stronger consistency in its predictions across the neighborhood of specific timesteps. In contrast, predictions for non-member samples exhibit stronger randomness due to their absence from the training process.

We additionally observe that high-intensity noise injection disrupts semantic information in images (as shown in Figure 2), significantly degrading inference performance. Conversely, low-intensity noise better preserves structural content and amplifies inter-class discrepancies between member and non-member samples. Building on this insight, we adopt a single-step small-noise injection strategy to approximate one diffusion step toward the target timestep, which not only drastically reduces the number of model queries but also improves membership inference accuracy.

In this way, we propose a membership inference method based on noise aggregation analysis and small-noise injection strategy. To validate the effectiveness of our method, we conducted extensive experiments. The results demonstrate that our approach achieves comprehensive improvements on DDPM benchmarks and also delivers strong performance on text-to-image models.

The main contributions of this paper are summarized as follows:

- To the best of our knowledge, we are the first to explicitly leverage noise aggregation analysis for membership inference attack against diffusion models. We argue that noise sampling in diffusion models is inherently stochastic, and analyzing noise aggregation can effectively mitigate this randomness, enabling more accurate membership inference.
- We introduce an efficient diffusion-based sampling strategy for membership inference attacks based on one-time small-noise injection. This strategy substantially reduces sampling cost while amplifying the inter-class differences between members and non-members, leading to improved attack efficiency and effectiveness.
- We conduct extensive experiments on multiple datasets to validate the effectiveness of the proposed method. Through ablation studies and sample distribution analysis, we identify the key factors that are critical to the attack performance and provide insights into the memorization behavior of diffusion models.

II. Related Work

A. Diffusion Models

Diffusion models [20]–[22] have emerged as the leading approach for high-quality image generation, offering superior stability and quality compared to GANs [23] and VAEs [24] through their iterative denoising process. Beyond image synthesis, their flexible conditional control has revolutionized diverse fields, including video [25], [26], speech [27]–[29], and molecular design [30], [31]. At the same time, researchers have proposed many improvements to increase efficiency and quality of the diffusion models, such as accelerated sampling (DDIM [32]), latent space diffusion (LDM [6]), and architectural optimizations that combine attention mechanisms with large-scale pre-training [5], [6], [33]. These advances have greatly expanded the application scope of diffusion models and raised broad discussions on privacy, security, and copyright.

B. Membership Inference Attacks

Membership Inference Attacks (MIA) aim to determine whether a specific sample was used to train a target model. Initially focused on classification tasks, early works established foundational strategies using shadow models and loss-based metrics [9]–[13]. As generative models gained popularity, research expanded to GANs and VAEs, where scholars developed white-box and black-box attacks based on generator outputs, Monte Carlo integration, and reconstruction errors [34]–[37]. These studies revealed fundamental privacy vulnerabilities in generative tasks, laying the groundwork for investigating more complex architectures.

With the widespread adoption of diffusion models, researchers have adapted MIA strategies to this domain, employing techniques such as posterior estimation, pixel error analysis, and likelihood comparison [14]–[16], [18], [19]. However, existing methods exhibit notable limitations: loss-based approaches often fail to account for sampling stochasticity, limiting accuracy [14], [15]; reconstruction-based methods incur high computational costs due to extensive model queries [16], [17]; and text-conditioned strategies often struggle to generalize across architectures [18], [19]. To address these challenges, we propose a method that leverages the consistency of noise predictions across adjacent timesteps to analyze the aggregation of predicted noise, effectively mitigating the instability caused by random noise sampling. Meanwhile, we introduce a single-step small-noise injection strategy that significantly reduces the number of model queries, thereby improving the overall efficiency of the attack.

III. Preliminaries

Denoising Diffusion Probabilistic Models (DDPM [20]) define a stochastic Markov chain that gradually transforms an image into noise. The forward process, denoted as q, iteratively adds Gaussian noise to the data at each timestep. This process can be written as:

$$q(x_{1:T}|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1}) \tag{1}$$

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I) \tag{2}$$

where β_t is a variance schedule that controls the noise level at step t. As t increases, $\bar{\alpha}_t$ approaches zero, making x_t isotropic

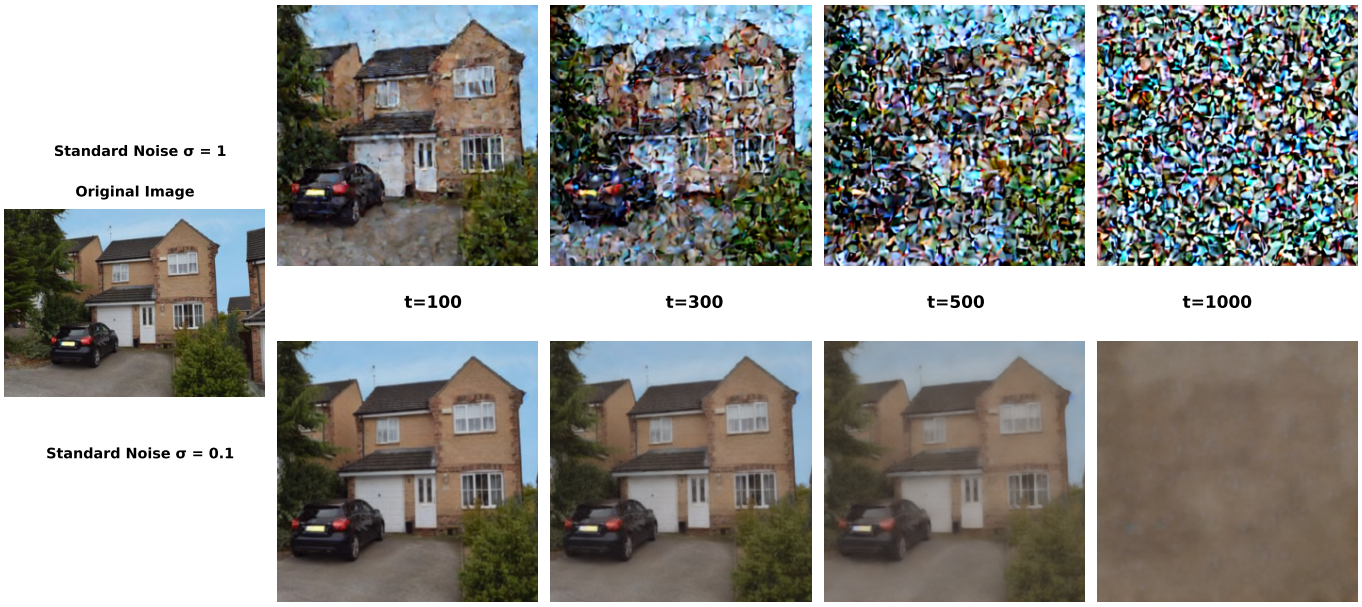


Fig. 2: Comparison of diffusion effects under different noise intensities. The first row shows the diffusion process after injecting noise with standard deviation $\sigma = 1$, while the second row shows the process with standard deviation $\sigma = 0.1$. The results indicate that low-intensity noise better preserves the overall image structure and is more favorable for subsequent membership inference.

Gaussian noise. The reverse process aims to reconstruct the data distribution from Gaussian noise. This process can be expressed as:

$$p(x_{0:T}) = p(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t) \tag{3}$$

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)) \tag{4}$$

where $\Sigma_\theta(x_t, t)$ is a constant depending on the variance schedule β_t, and $\mu_\theta(x_t, t)$ is determined by a neural network. Through recursive application of reverse steps, Gaussian noise is converted back to the original image.

Training DDPM requires sampling image x_0, timestep t, and random noise $\epsilon \sim \mathcal{N}(0, I)$. The forward process generates noisy image x_t. A U-Net ϵ_θ predicts the noise in x_t. The loss function for the denoising U-Net is:

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

To control the randomness of the reverse process, DDIM [32] modifies the noise added at each step:

$$x_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\hat{x}_0 + \sqrt{1 - \bar{\alpha}_{t-1} - \sigma_t^2}\epsilon_\theta(x_t, t) + \sigma_t\epsilon_t \tag{6}$$

where $\hat{x}_0$ is the estimated initial data, with expressions:

$$\hat{x}_0 = \frac{x_t - \sqrt{1 - \bar{\alpha}_t}\epsilon_\theta(x_t, t)}{\sqrt{\bar{\alpha}_t}} \tag{7}$$

IV. Methodology

This section provides a detailed exposition of our proposed membership inference attack. As illustrated in Fig. 3, our approach consists of the following three stages: small-scale noise injection, iterative denoising prediction, and noise aggregation degree quantification analysis. Overall, we infer the membership by injecting small-noise into the image and then quantifying the spatial aggregation degree of the predicted noise.

A. Small-Scale Noise Injection Strategy

The forward process of diffusion models is inherently a stochastic Markov chain that gradually adds Gaussian noise to the data. Based on the reparameterization trick, the cumulative noise injection from the initial state x_0 to any timestep t can be mathematically equivalent to a single-step transformation:

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon \tag{8}$$

where $\epsilon \sim \mathcal{N}(0, I)$ is standard Gaussian noise. This property allows us to directly obtain noisy images at target timesteps without iterative computation.

However, directly sampling with standard large variance poses limitations for membership inference. As shown in the first row of Fig. 2, standard noise ($\sigma = 1$) introduces severe high-frequency corruption. Even at early timesteps, the semantic structure of the original image is rapidly obliterated and transformed into unrecognizable random noise. This destruction pushes the sample entirely out of the model's local memorization scope, compressing the statistical distance between members and non-members.

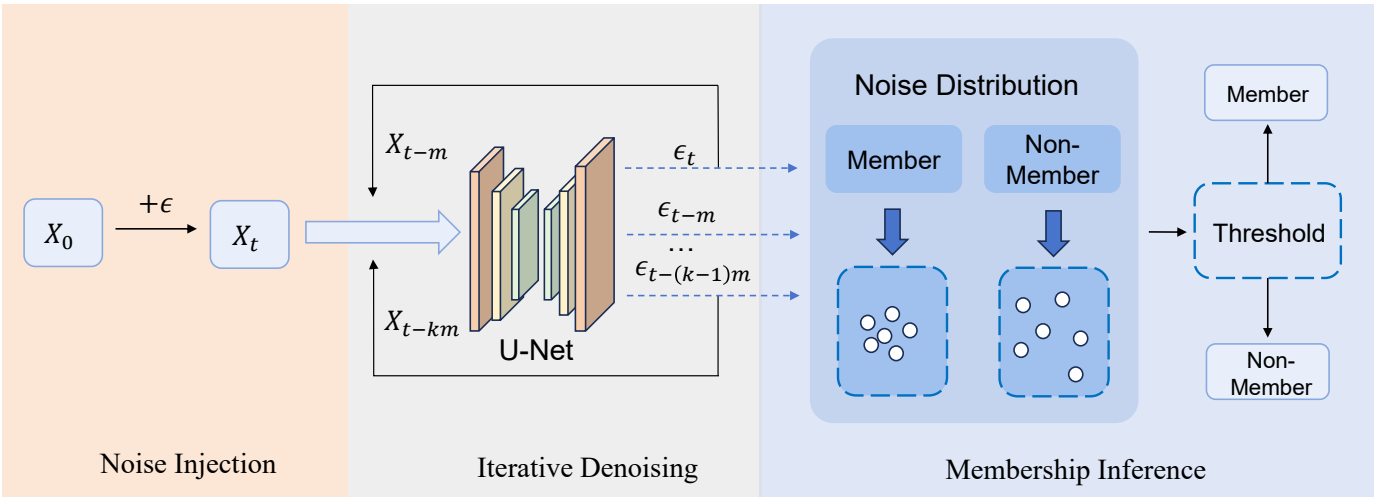


Fig. 3: Overview of our proposed membership inference attack pipeline. The approach injects small noise ϵ into test images X_0, predicts noise at selected timesteps t, and evaluates the aggregation degree of predictions to determine the membership of the sample.

To address this, we propose a strategy based on small-scale noise injection. Our motivation draws on the geometry of the loss landscape, particularly the distinction between sharp and flat minima [38]. We hypothesize that member samples, having been overfitted, typically reside in sharp minima (steep basins of attraction), whereas non-members are located in flatter regions. As demonstrated in the second row of Fig. 2, applying a limited noise variance (e.g., $\sigma = 0.1$) acts as a structure-preserving perturbation. Unlike the standard process, the image content remains visually recognizable, indicating the sample is still retained within its potential attraction domain. For member samples, the steep gradients characteristic of sharp minima exert a consistent "restoration force" on these preserved structures, pulling the prediction back to the origin. Conversely, the recovery of non-members exhibits higher uncertainty. This approach amplifies the distinguishability of members while maintaining high computational efficiency.

B. Iterative Denoising Prediction Stage

After obtaining the noisy image x_t, we acquire noise predictions at different timesteps through an iterative denoising process to construct feature vectors for membership analysis. This process is based on the intuition that training set members, having been repeatedly optimized during model training, exhibit higher consistency and concentration in the model's noise predictions.

First, we input the noisy image x_t into the target U-Net denoising network to obtain the predicted noise at timestep t:

$$\hat{\epsilon}_t = \epsilon_\theta(x_t, t) \tag{9}$$

Based on the predicted noise, we can estimate the original image:

$$\hat{x}_0 = \frac{x_t - \sqrt{1-\bar{\alpha}_t}\hat{\epsilon}_t}{\sqrt{\bar{\alpha}_t}} \tag{10}$$

Based on Eq. (6), we set $\sigma_t = 0$, the process of denoising becomes deterministic. Then we can estimate the original image at timestep $t - m$:

$$x_{t-m} = \sqrt{\bar{\alpha}_{t-m}}\hat{x}_0 + \sqrt{1-\bar{\alpha}_{t-m}}\hat{\epsilon}_t \tag{11}$$

where m is the stride between sampled timesteps. By repeating the above denoising process, we sequentially obtain predicted noise sequences for k consecutive timesteps $\{\hat{\epsilon}_t, \hat{\epsilon}_{t-m}, \hat{\epsilon}_{t-2m}, \ldots, \hat{\epsilon}_{t-(k-1)m}\}$. This noise sequence constitutes the fundamental feature representation for subsequent membership analysis.

C. Noise Aggregation Degree Quantification Analysis

From an information-theoretic perspective, member samples are repeatedly optimized during training, resulting in more concentrated predicted noise and lower entropy. In contrast, non-member samples exhibit greater uncertainty in their predicted noise, leading to higher entropy and more dispersed predictions. Let $(H(\epsilon|x))$ denote the conditional differential entropy of the predicted noise given a sample (x). The above hypothesis can therefore be expressed as

$$H(\epsilon|x_{member}) < H(\epsilon|x_{non-member}) \tag{12}$$

As shown in Fig. 1, we visualize the differential entropy of member and non-member samples using 2D heatmaps. It can be observed that the heatmaps of non-member samples appear redder, indicating higher entropy values and greater uncertainty compared to member samples. To further quantify the uncertainty between member and non-member samples, we introduce the noise aggregation degree to characterize the consistency of the predicted noise.

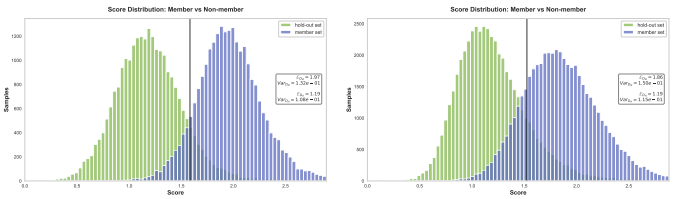
Let the noise sequence be denoted as $\mathcal{E} = \{\hat{\epsilon}_{t-im}\}_{i=0}^{k-1}$, and define the noise aggregation degree as $C(\mathcal{E})$. We measure

TABLE I: Comparison of ASR and AUC across different datasets. **Bold** values denote the best results, and blue values denote the second-best results. The ↑ symbol means higher is better.

		CIFAR-10		CIFAR-100		Tiny-IN		Average	
Method	Query	ASR↑	AUC↑	ASR↑	AUC↑	ASR↑	AUC↑	ASR↑	AUC↑
GAN-Leaks	1000	0.615	0.646	0.513	0.459	0.545	0.457	0.558	0.521
NaiveLoss	1	0.663	0.718	0.654	0.709	0.646	0.700	0.654	0.709
SecMI	12	0.811	0.881	0.798	0.868	0.821	0.894	0.810	0.881
Ours	5	**0.901**	**0.957**	**0.839**	**0.903**	**0.842**	**0.912**	**0.861**	**0.924**

TABLE II: Comparison of TPR @ 1% FPR (%) and TPR @ 0.1% FPR (%) across different datasets. **Bold black** values denote the best results, and blue values denote the second-best results. The ↑ symbol means higher is better.

	CIFAR-10		CIFAR-100		Tiny-IN	
Method	TPR@1%↑	TPR@0.1%↑	TPR@1%↑	TPR@0.1%↑	TPR@1%↑	TPR@0.1%↑
GAN-Leaks	2.80	0.29	1.85	0.23	1.01	0.13
NaiveLoss	3.35	0.40	4.79	0.76	4.44	0.44
SecMI	9.11	0.66	9.26	0.46	12.67	0.96
Ours	**28.7**	**1.22**	**9.65**	**0.78**	**14.58**	**1.03**



(a) CIFAR10 (b) Tiny-IN

Fig. 4: Distribution histogram of member samples and non-member samples

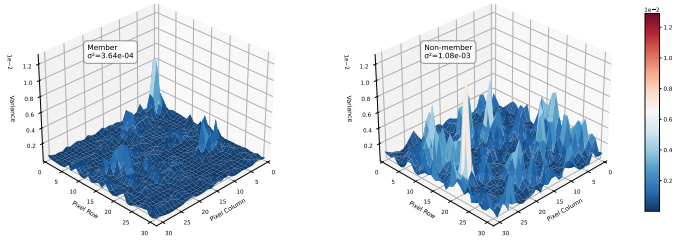


Fig. 5: 3D heat map of variance between member samples and non-member samples

TABLE III: Attack performance across denoising steps

K	ASR↑	AUC↑	TPR@1%FPR↑	TPR@0.1%FPR↑
2	0.888	0.944	25.5	1.19
3	0.892	0.952	26.9	1.18
4	0.897	0.955	29.6	1.20
5	0.901	0.956	29.7	1.28
6	0.900	0.956	26.9	1.01
7	0.893	0.952	22.0	0.74
8	0.875	0.938	17.7	0.63

TABLE IV: Attack performance across initial noise intensity

std_dev	ASR↑	AUC↑	TPR@1%FPR↑	TPR@0.1%FPR↑
0.05	0.884	0.944	16.20	0.864
0.10	0.901	0.957	30.10	0.980
0.15	0.896	0.955	31.20	2.170
0.20	0.882	0.945	24.80	1.880
0.30	0.862	0.927	14.50	0.696
0.50	0.837	0.903	7.73	0.452
1.00	0.742	0.806	4.12	0.292

$C(\mathcal{E})$ by computing the average pairwise Euclidean distance among all predicted noises, formulated as:

$$C(\mathcal{E}) = \frac{1}{|\mathcal{E}|^2} \sum_{i,j} \|\hat{\epsilon}_i - \hat{\epsilon}_j\|_2 \tag{13}$$

To enhance numerical stability, the final membership score is defined as:

$$S_m = -\log(C(\mathcal{E}) + \delta) \tag{14}$$

where δ is a numerical stability constant. Then the membership inference function is defined as:

$$\mathcal{A}(x, g_\phi) = \mathbb{1}[S_m \geq \tau] \tag{15}$$

where $\mathbb{1}$ denotes the indicator function, which outputs 1 if the membership score exceeds the threshold τ, indicating that the sample is inferred as a member;

V. Experiments

A. Experimental Setup

Dataset Configuration: We conducted comprehensive evaluations on three widely used visual datasets: CIFAR-10, CIFAR-100, and Tiny-ImageNet (Tiny-IN). We employed a random partitioning strategy, allocating 50% of samples from each dataset as the training set and the remaining 50% for testing. Specifically, the training and test sets for CIFAR-10 and CIFAR-100 each contain 25,000 samples, while Tiny-IN contains 50,000 samples each.

For text-to-image model evaluation, as they trained on LAION datasets, we randomly sampled 1,000 images from it as the member set. For the non-member set, we randomly sampled 1,000 images from the COCO2017-Val dataset, which is commonly used for generative model evaluation.

Implementation Details: The training procedures and hyperparameter settings for diffusion models remain consistent with [16] to ensure fair comparison. Key parameter settings include: attack diffusion steps $t = 80$, noise prediction sampling count $k = 5$, and DDIM accelerated sampling denoising steps $m = 10$. For text-to-image models, we directly employ pre-trained Stable Diffusion v1.4 and v1.5 models from HuggingFace as attack targets.

Evaluation Metric: We adopt standard evaluation metrics widely used in previous membership inference research, including Attack Success Rate (ASR), Area Under the ROC Curve (AUC), and True Positive Rate at fixed false positive rates.

B. Baseline Method Comparison

We compare against a suite of representative baselines, including GAN-Leaks [37], NaiveLoss [14], and SecMI [16]. To ensure a fair evaluation, we adopt a unified protocol based on a threshold decision rule without training on the test set. And we follow the implementation details reported in each method's original paper. We trained DDPM models on CIFAR-10, CIFAR-100, and Tiny-IN for comprehensive performance comparison.

As shown in Table I and Table II, our method outperforms all existing approaches on all metrics. Moreover, compared with SecMI, which achieves the best overall performance among existing approaches, our method significantly reduces the number of model queries, thereby greatly improving the efficiency of membership inference attacks.

C. Comparative Analysis of Member and Non-member Samples

As illustrated in Figure 4, we plot the histogram of membership scores obtained using our proposed method. It can be observed that the score distributions of member and non-member samples differ significantly. Member samples generally exhibit higher scores, which is the key factor enabling the effectiveness of our approach.

To further provide an intuitive visualization of the differences between member and non-member samples, we also plot their corresponding 3D heatmaps, which display the variance of each pixel's predicted noise within the neighborhood of a specific timestep. As shown in Figure 5, the member samples appear darker in color, indicating smaller pixel-wise variations and a higher degree of noise aggregation. This demonstrates that member samples exhibit stronger consistency in their predicted noise during the generative process.

D. Ablation Studies

1) Denoising Steps Impact: Our method relies on analyzing the aggregation of predicted noises across multiple consecutive timesteps, making the choice of the number of noise samples crucial. To investigate this, we conducted experiments on the CIFAR-10 dataset, with results shown in Table III. As the number of denoising steps K increases, the performance of our method exhibits a "rise-then-fall" trend. We think this is because when the number of denoising steps is small, the inherent randomness of the sampling noise dominates, preventing the aggregation analysis from fully demonstrating its advantage. And when the number of denoising steps is large, the denoising process introduces noticeable differences between images, which amplifies the differences in predicted noise and leads to a performance drop. Notably, even with $K = 2$, our method outperforms the current state-of-the-art method [16], further demonstrating that our method achieves efficient inference while significantly reducing the cost of the attack.

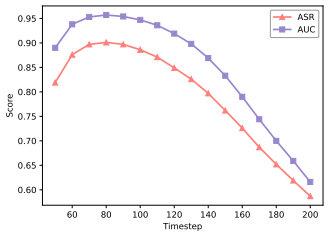
2) Initial Noise Intensity Impact Analysis: Our method enhances attack efficiency by injecting low-intensity noise and performing a single-step diffusion, which greatly reduces the number of model queries. As shown in Table IV, we further investigate the impact of different noise standard deviations on attack performance. The attack achieves optimal performance when the standard deviation of the injected noise is around 0.1. When the noise variance is too small or too large, the attack effectiveness decreases. We think this is because when the noise level is too low, the image maintains a high signal-to-noise ratio, resulting in a lack of clear distinction between member and non-member samples. When the noise level is too high, it destroys the semantic information of the image, making it difficult even for member samples to accurately predict the noise, which weakens the distinction between member and non-member samples.

3) Aggregation Degree Metrics Comparative Analysis: To evaluate the noise aggregation central to our method, we initially employed the L2 mean distance. To verify robustness, we explored four alternative metrics: the L1 average distance for outlier resistance, the centroid distance for computational simplicity and intuitive concentration measurement, the average density for assessing local compactness, and the convex hull volume for capturing global geometric dispersion. Experimental results (Table V) indicate that while the L2 mean distance achieves the best performance, the performance gaps compared to density and centroid-based metrics are negligible, confirming that our approach remains robust to the choice of aggregation metric.

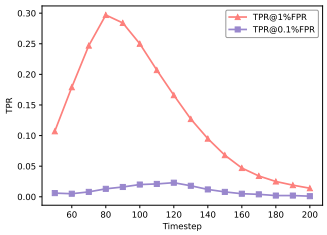
4) Timestep Parameter Impact Analysis: We systematically examined the impact of different timesteps on attack perfor-

TABLE V: Impact of different aggregation degree metrics

Aggregation Metrics	ASR↑	AUC↑	TPR@1%FPR↑ (%)
Distance to centroid	0.899	0.956	28.5
Convex hull volume	0.771	0.839	5.9
Average density	0.901	0.957	28.2
L1 average distance	0.885	0.942	12.9
L2 average distance	0.901	0.957	29.7

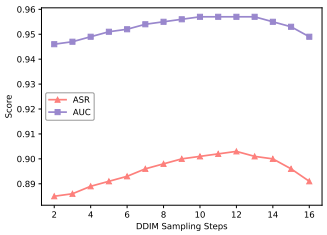


(a) ASR and AUC

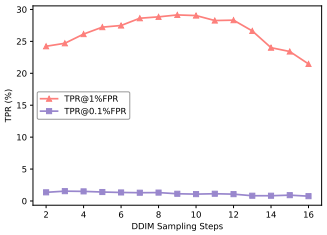


(b) TPR

Fig. 6: Attack performance across timesteps



(a) ASR and AUC



(b) TPR

Fig. 7: Attack performance across DDIM sampling intervals

mance. As illustrated in Figure 6, empirical findings indicate that our method performs well within the $T \in [50, 140]$ range. We think it is because within this range, images suffer relatively minimal noise interference and retain sufficient structural information, enabling diffusion models to effectively predict original noise based on preserved structural features. In corresponding neighborhoods, predicted noise distributions for member images are more concentrated compared to non-member images, which enables effective member differentiation through noise aggregation.

5) DDIM Sampling Intervals Impact Analysis: We further analyzed the impact of DDIM sampling intervals on overall performance. As illustrated in Figure 7, experiments shows that increasing sampling intervals can enhance attack performance, but performance stabilizes after reaching a certain range. We believe the reason is that if the sampling intervals is too short, the images will be too similar to identify. Both member and non-member images will predict similar noise within adjacent denoising steps, making it difficult to distance them. But if the sampling intervals is too long, the difference between the image and its denoised neighboring images will be too large. Even adjacent images denoised by the member image may be considered different images, resulting in large differences in noise predictions, which reduced the accuracy of the attack model. Ultimately, we select sampling interval $m = 10$ to balance performance and computational efficiency.

E. Large-Scale Text-to-Image Model Evaluation

To validate the effectiveness of our method in practical application scenarios, we conducted attack testing on mainstream text-to-image generation models Stable Diffusion v1.4 and v1.5 provided by HuggingFace. We randomly sampled 1,000 images from the LAION-aesthetic-5plus dataset as the member set and sampled 1,000 images from the COCO2017-Val dataset as the non-member set.

As shown in Table VI, our method outperforms existing approaches in terms of ASR and AUC, but the attack effectiveness decreases noticeably compared with DDPM. We attribute this to the VAE-based latent diffusion process, where encoding images into a lower-dimensional space reduces information diversity and weakens our method's ability to distinguish member samples. Moreover, the large-scale training dataset enhances the model's generalization capability, making it more resistant to MIA. In addition, our method performs significantly worse than the baseline NaiveLoss [14] on TPR@1%FPR. We attribute this to the fact that the method uses the loss value as the decision criterion, making it less affected by model components such as the VAE encoder. In particular, when the loss differences are minimal, the method can confidently infer that the discrepancies arise from training overfitting, thereby achieving high-confidence inference at low false positive rates.

VI. Conclusion

Existing membership inference attack methods for diffusion models typically rely on either the loss function or image-level reconstruction comparisons, but they struggle to achieve both high accuracy and high efficiency. In this work, we propose a membership inference approach based on the aggregation analysis of predicted noises, which effectively mitigates the impact of stochastic noise sampling and enables rapid diffusion of samples through single-step low-intensity noise injection. Our method fully leverages the consistency of predicted noises in local timesteps of diffusion models, allowing it to accurately distinguish member samples from non-member samples and substantially reducing the number of model queries. Empirical findings indicate that our method outperforms existing methods on both standard diffusion models, e.g., DDPM and text-to-image models, e.g., Stable Diffusion. In addition, our analysis of noise intensity and sampling steps provides further insights into the memorization behavior of diffusion models.

References

[1] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," *Advances in neural information processing systems*, vol. 34, pp. 8780–8794, 2021.

[2] X. Liu, D. H. Park, S. Azadi, G. Zhang, A. Chopikyan, Y. Hu, H. Shi, A. Rohrbach, and T. Darrell, "More control for free! image synthesis with semantic diffusion guidance," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2023, pp. 289–299.

[3] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv:2204.06125*, vol. 1, no. 2, p. 3, 2022.

[4] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 22 500–22 510.

[5] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans *et al.*, "Photorealistic text-to-image diffusion models with deep language understanding," *Advances in neural information processing systems*, vol. 35, pp. 36 479–36 494, 2022.

TABLE VI: Performance on Stable Diffusion Models. **Bold** values denote the best results, and blue values denote the second-best results.

Methods	SD1.4			SD1.5		
	ASR↑	AUC↑	TPR@1%FPR↑	ASR↑	AUC↑	TPR@1%FPR↑
GAN-Leaks	0.533	0.468	1.73	0.541	0.472	1.64
NaiveLoss	0.630	0.638	**23.7**	0.631	0.639	**23.7**
SecMI	0.602	0.605	15.3	0.602	0.606	15.3
Ours	**0.701**	**0.652**	8.0	**0.696**	**0.661**	8.3

[6] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[7] J. Ho and T. Salimans, "Classifier-free diffusion guidance," *arXiv preprint arXiv:2207.12598*, 2022. [Online]. Available: https://arxiv.org/abs/2207.12598

[8] J. Ma, T. Hu, W. Wang, and J. Sun, "Elucidating the design space of classifier-guided diffusion generation," in *NeurIPS 2023 Workshop on Score-Based Methods*, 2023. [Online]. Available: https://arxiv.org/abs/2310.11311

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

[10] S. Yeom, I. Giacomelli, M. Fredrikson, and S. Jha, "Privacy risk in machine learning: Analyzing the connection to overfitting," in *2018 IEEE 31st computer security foundations symposium (CSF)*. IEEE, 2018, pp. 268–282.

[11] A. Salem, Y. Zhang, M. Humbert, P. Berrang, M. Fritz, and M. Backes, "Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models," *arXiv preprint arXiv:1806.01246*, 2018.

[12] Y. Long, V. Bindschaedler, L. Wang, D. Bu, X. Wang, H. Tang, C. A. Gunter, and K. Chen, "Understanding membership inferences on well-generalized learning models," *arXiv preprint arXiv:1802.04889*, 2018.

[13] Y. Long, L. Wang, D. Bu, V. Bindschaedler, X. Wang, H. Tang, C. A. Gunter, and K. Chen, "A pragmatic approach to membership inferences on machine learning models," in *2020 IEEE European Symposium on Security and Privacy (EuroS&P)*. IEEE, 2020, pp. 521–534.

[14] T. Matsumoto, T. Miura, and N. Yanai, "Membership inference attacks against diffusion models," in *2023 IEEE Security and Privacy Workshops (SPW)*. IEEE, 2023, pp. 77–83.

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

[16] J. Duan, F. Kong, S. Wang, X. Shi, and K. Xu, "Are diffusion models vulnerable to membership inference attacks?" in *International Conference on Machine Learning*. PMLR, 2023, pp. 8717–8730.

[17] X. Fu, X. Wang, Q. Li, J. Liu, J. Dai, J. Han, and X. Gao, "Unlocking generative priors: A new membership inference framework for diffusion models," *IEEE Transactions on Information Forensics and Security*, 2025.

[18] S. Zhai, H. Chen, Y. Dong, J. Li, Q. Shen, Y. Gao, H. Su, and Y. Liu, "Membership inference on text-to-image diffusion models via conditional likelihood discrepancy," *Advances in Neural Information Processing Systems*, vol. 37, pp. 74 122–74 146, 2024.

[19] Q. Li, X. Fu, X. Wang, J. Liu, X. Gao, J. Dai, and J. Han, "Unveiling structural memorization: Structural membership inference attack for text-to-image diffusion models," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 10 554–10 562.

[20] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[21] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *International conference on machine learning*. pmlr, 2015, pp. 2256–2265.

[22] A. Q. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," in *Proceedings of the 38th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 139. PMLR, 2021, pp. 8162–8171. [Online]. Available: http://proceedings.mlr.press/v139/nichol21a.html

[23] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

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

[25] O. Bar-Tal, H. Chefer, O. Tov, C. Herrmann, R. Paiss, S. Zada, A. Ephrat, J. Hur, G. Liu, A. Raj *et al.*, "Lumiere: A space-time diffusion model for video generation," in *SIGGRAPH Asia 2024 Conference Papers*, 2024, pp. 1–11.

[26] X. Ma, Y. Wang, G. Jia, X. Chen, Z. Liu, Y.-F. Li, C. Chen, and Y. Qiao, "Latte: Latent diffusion transformer for video generation," *arXiv preprint arXiv:2401.03048*, 2024.

[27] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, and M. Kudinov, "Grad-tts: A diffusion probabilistic model for text-to-speech," in *International conference on machine learning*. PMLR, 2021, pp. 8599–8608.

[28] M. Jeong, H. Kim, S. J. Cheon, B. J. Choi, and N. S. Kim, "Diff-tts: A denoising diffusion model for text-to-speech," *arXiv preprint arXiv:2104.01409*, 2021.

[29] Y. Long, K. Yang, Y. Ma, and Y. Yang, "Mixdiff-tts: Mixture alignment and diffusion model for text-to-speech," *Applied Sciences*, vol. 15, no. 9, p. 4810, 2025.

[30] L. Huang, T. Xu, Y. Yu, P. Zhao, X. Chen, J. Han, Z. Xie, H. Li, W. Zhong, K.-C. Wong *et al.*, "A dual diffusion model enables 3d molecule generation and lead optimization based on target pockets," *Nature Communications*, vol. 15, no. 1, p. 2657, 2024.

[31] M. Oestreich, E. Merdivan, M. Lee, J. L. Schultze, M. Piraud, and M. Becker, "Drugdiff: small molecule diffusion model with flexible guidance towards molecular properties," *Journal of cheminformatics*, vol. 17, no. 1, p. 23, 2025.

[32] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," *arXiv preprint arXiv:2010.02502*, 2020.

[33] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Dall· e 2: Hierarchical text-conditional image generation with clip latents," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

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

[35] B. Hilprecht, M. Härterich, and D. Bernau, "Monte carlo and reconstruction membership inference attacks against generative models," *Proceedings on Privacy Enhancing Technologies*, 2019.

[36] K. S. Liu, C. Xiao, B. Li, and J. Gao, "Performing co-membership attacks against deep generative models," in *2019 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2019, pp. 459–467.

[37] D. Chen, N. Yu, Y. Zhang, and M. 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*, 2020, pp. 343–362.

[38] N. S. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. T. P. Tang, "On large-batch training for deep learning: Generalization gap and sharp minima," in *International Conference on Learning Representations (ICLR)*, 2017.