Exploring Privacy Risks in Diffusion Models through Membership Inference Attacks

Keerthana Ravichandran

Student ID: 2866714

MSc Data Science

Supervised by Professor Huiping Chen

School of Computer Science

College of Engineering and Physical Sciences

University of Birmingham

2024-25

Honour Code

I hereby declare that this thesis is my own original work, carried out in fulfillment of the requirements for the Master's degree at University of Birmingham. All sources of information used have been appropriately acknowledged through citation.

In line with the university's academic integrity policy, I disclose the use of generative AI tools during this project. Specifically, OpenAI's ChatGPT was used to assist in improving clarity of expression, ensuring consistency in academic tone, and providing debugging support for code. The research problem, methodology design, experiments, data analysis, and interpretation of results are entirely my own. All AI-assisted outputs were critically reviewed, revised, and integrated by me to ensure correctness, originality, and compliance with academic standards.

I certify that, to the best of my knowledge, this thesis represents my own independent work and complies fully with the standards of academic integrity and honesty.

Acknowledgements

I would like to extend my sincere gratitude to my supervisor, Professor Huiping Chen, for her remarkable guidance and unwavering support throughout this dissertation. Her insightful feedback and constructive criticism were instrumental in shaping the direction of this work. Her expertise in privacy risk analysis and diffusion models was invaluable, and her sustained interest in these topics provided both motivation and clarity at every stage of the research.

I am also grateful to the faculty and staff of the University of Birmingham for fostering an academic environment that encouraged inquiry and supported my learning. I wish to thank my colleagues and peers, whose discussions and collaboration contributed significantly to refining my ideas and methodology.

Finally, I owe a profound debt of gratitude to my family and friends for their patience, encouragement, and unwavering support during this journey. Their belief in me has been a constant source of motivation and strength in completing this thesis.

Abstract

Diffusion models have recently emerged as state-of-the-art generative techniques, with applications in image synthesis, data augmentation, and privacy-preserving data release. However, their adoption in sensitive domains requires careful evaluation of privacy risks, particularly with respect to membership inference attacks (MIAs), which aim to determine whether specific records were included in training.

This thesis investigates privacy leakage in two representative models: *DP-Promise*, which incorporates formal differential privacy guarantees, and *DCTDiff*, a non-private baseline. The methodology combines generative quality assessment using Fréchet Inception Distance (FID) and Inception Score (IS) with systematic MIA evaluation. Black-box attacks were applied to both models, leveraging semantic and perceptual features (CLIP and LPIPS), while white-box analysis was conducted only for DP-Promise using loss-based reconstruction signals across diffusion timesteps.

Experimental results demonstrate that (i) membership inference remains feasible in both private and non-private diffusion models, (ii) no consistent monotonic trend in leakage was observed with respect to privacy budget (ϵ) or sampler choice, (iii) white-box signals provide substantially stronger leakage than black-box features, and (iv) outcomes are sensitive to random seeds, underscoring the importance of reporting medians across runs.

The findings highlight that while differential privacy mechanisms support generative training under constraints, they do not eliminate leakage. Generative fidelity gains, whether from higher ϵ or optimized samplers, do not reduce vulnerability. These results motivate the need for standardized empirical privacy audits, careful reporting practices, and conservative deployment of diffusion models. Recommendations and directions for future research are outlined to bridge the gap between theoretical guarantees and practical robustness.

Table of Contents

1. Introduction

1.1 Background

Generative models have revolutionized artificial intelligence (AI) by enabling the synthesis of highly realistic images, text, audio, and more. Early frameworks such as Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) laid the foundation for data-driven creativity, but it is the emergence of diffusion models that has recently catalyzed a new wave of innovation in generative AI. Diffusion models surpass earlier approaches in producing high-fidelity and diverse outputs, accelerating their adoption across both academia and industry. Their applications now range from creative art tools and content generation (e.g., DALL·E, Stable Diffusion) to sensitive domains such as medical imaging and biometrics.

However, the widespread deployment of generative models introduces serious privacy concerns. Unlike traditional statistical models, modern generative architectures may not only learn general patterns but also inadvertently memorize specific training examples. This memorization can lead to privacy leakage, where information about individuals in the training data may be revealed through model outputs. Responsible development of generative AI therefore requires careful study of these privacy risks, particularly in diffusion models, whose rapid adoption has in some cases, outpaced systematic understanding of their vulnerabilities [14].

1.2 Privacy Risks in Generative AI

A central privacy concern in generative modeling is the threat of *Membership Inference Attacks (MIAs)*. In an MIA, an adversary seeks to determine whether a specific data sample was included in the training dataset of a model. A successful attack indicates that the model has memorized aspects of its training data, thereby compromising dataset confidentiality and potentially revealing sensitive information about individuals. In generative settings, leakage may surface either through direct reproduction or through systematic output shifts correlated with training inclusion.

Much of the early research on MIAs has focused on classification models, with subsequent extensions to generative architectures such as GANs and VAEs. These studies have shown that generative models can also leak membership information, albeit through mechanisms different from discriminative models. However, systematic evaluation of MIAs in state-of-the-art diffusion models remains limited. Given the rapid adoption of diffusion models in high-stakes domains, investigating their vulnerability to MIAs is both timely and necessary [1].

1.3 Motivation

The motivation for this research arises from two pressing gaps. First, diffusion models are increasingly deployed in sensitive applications such as biometric analysis and medical imaging, where privacy protection is critical. Yet, their vulnerability to membership inference attacks remains underexplored compared to earlier generative frameworks like GANs and VAEs. Second, while Differential Privacy (DP) has emerged as a principled defense mechanism that formally limits information leakage, its practical effectiveness in diffusion models has not been evaluated with standardized, comparable protocols.

This work is motivated by the need to rigorously assess whether diffusion models, particularly variants trained with DP mechanisms such as DP-Promise [22] can resist membership inference attacks in black-box and white-box settings. Addressing this question is not only theoretically significant but also vital for ensuring the safe deployment of generative AI in real-world, privacy-sensitive domains.

1.4 Problem Statement

Building on these concerns, this work focuses on evaluating the vulnerability of diffusion models to membership inference attacks (MIAs). The study considers two representative frameworks: **DP-Promise**, which incorporates Differential Privacy into the training process, and **DCTDiff**, which explores generative modeling in the Discrete Cosine Transform (DCT) domain.

The central research question guiding this work is:

> *To what extent are diffusion models—across different privacy budgets, sampling strategies, and dataset resolutions—susceptible to membership inference attacks, in black-box settings for both DP-Promise and DCTDiff, and in white-box settings for DP-Promise?*

1.5 Contributions

This study makes the following key contributions:

- **Attack frameworks:** Implementation of both black-box and white-box membership inference attack frameworks tailored to diffusion models.
- **Model analysis:** Systematic evaluation of privacy leakage across two representative frameworks DP-Promise (with differential privacy) and DCTDiff (without differential privacy).
- **Black-box setting:** Comprehensive experiments on DP-Promise with the DDIM sampler across three privacy budgets (ϵ values: 1, 5 and 10) and two resolutions (32×32 and 64×64), and on DCTDiff with three sampling strategies (DPM-Solver, Euler Maruyama ODE, Euler Maruyama SDE).
- **White-box setting:** In-depth analysis of DP-Promise under direct access to its denoising network, evaluating memorization signals via noise prediction loss and reconstruction fidelity across multiple ϵ values and resolutions.

- **Quantitative and qualitative benchmarks:** Use of ROC-AUC scores, loss distribution histograms, and reconstruction visualizations to reveal privacy-utility trade-offs and highlight memorization effects.

Together, these contributions provide one of the first systematic studies of membership inference risks in diffusion models, offering insights into both theoretical vulnerabilities and practical implications.

1.6 Thesis Organization

The remainder of this thesis is structured as follows:

- Chapter 2 provides background on diffusion models, differential privacy, and membership inference attacks, and reviews related work.
- Chapter 3 describes the methodology, including datasets, model architectures, attack designs, and evaluation metrics.
- Chapter 4 presents experimental results for black-box and white-box attacks.
- Chapter 5 discusses the key findings, implications, and limitations of the study.
- Chapter 6 outlines directions for future work, such as extending to larger datasets, testing stronger privacy-preserving mechanisms, and studying multimodal diffusion models.
- Chapter 7 concludes with a summary of contributions and key takeaways.

2. Background and Literature Review

2.1 Generative Diffusion Models

Generative models aim to synthesize new data by learning the underlying distribution of real-world samples. Early paradigms such as Generative Adversarial Networks (GANs) [6] and Variational Autoencoders (VAEs) [10] laid the foundation for modern generative modeling, but they often faced challenges such as mode collapse, training instability, and limited output diversity.

Diffusion probabilistic models (DPMs) [9] have recently emerged as a state-of-the-art alternative, demonstrating the ability to generate highly realistic and diverse images. Their success has driven rapid adoption in domains ranging from creative content generation to biometrics and medical imaging.

The operation of diffusion models consists of two complementary phases:

- **Forward (diffusion) process:** Gaussian noise is gradually added to a clean image over T steps until it becomes indistinguishable from random noise, forming a Markov chain from data space to noise space:

$$q(x_{1:T} \mid x_0) = \prod_{t=1}^{T} q(x_t \mid x_{t-1}), \quad q(x_t \mid x_{t-1}) = \mathcal{N}\left(x_t; \sqrt{1-\beta_t}\, x_{t-1}, \beta_t I\right) \tag{2.1}$$

- **Reverse (denoising) process:** A neural network is trained to invert this corruption by iteratively denoising. At each step, the network predicts the noise and reconstructs the clean sample:

$$p_\theta(x_{t-1} \mid x_t) = \mathcal{N}\big(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)\big) \tag{2.2}$$

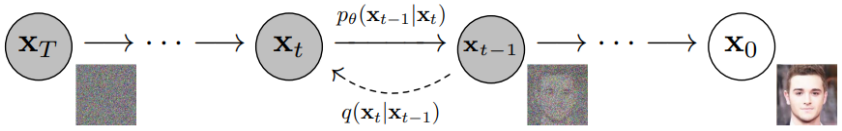


Figure 2.1: Forward (diffusion) and reverse (denoising) processes in diffusion models. Adapted from Ho et al. (2020) [9].

Since the reverse process can require thousands of steps, several sampling strategies have been proposed to accelerate inference while preserving fidelity. For instance, DDIM (Denoising Diffusion Implicit Models) [21] reduces the number of sampling steps through a deterministic non-Markovian formulation, enabling faster generation with minimal loss of quality. DPM-Solver [13] employs high-order ordinary differential equation (ODE) solvers to substantially shorten sampling time while maintaining fidelity. Euler methods (ODE and SDE variants) [11] provide deterministic and stochastic integration schemes, offering a flexible balance between precision and diversity.

These sampling strategies influence not only computational efficiency and visual quality but also the way information from the training distribution is propagated through the generation process.

Consequently, they are directly relevant for understanding the privacy risks associated with diffusion models.

2.2 Differential Privacy in Generative Models

Differential Privacy (DP) is a rigorous framework that provides formal guarantees against information leakage about individuals in a dataset. A randomized mechanism $\mathcal{M}$ is said to satisfy (ϵ, δ)-differential privacy if, for any two neighboring datasets D and D' differing by a single record, and for all measurable subsets S of the output space:

$$\Pr[\mathcal{M}(D) \in S] \leq e^{\epsilon} \cdot \Pr[\mathcal{M}(D') \in S] + \delta. \tag{2.3}$$

Here, ϵ is the *privacy budget* (smaller values indicate stronger privacy), and δ accounts for a small probability of failure. Intuitively, this ensures that the presence or absence of any single datapoint has only a negligible influence on the model's output [4].

2.2.1 DP in Generative Modeling

When applied to generative models, DP constrains the extent to which training data can be memorized and reproduced in generated samples. However, ensuring DP in high-dimensional image data is challenging due to the tendency of models to overfit. The most common approach is Differentially Private Stochastic Gradient Descent (DP-SGD) [1], which clips per-sample gradients and adds Gaussian noise before parameter updates. While effective in theory, DP-SGD often leads to non-trivial utility loss, which can manifest as degraded image quality in generative tasks.

2.2.2 DP-Promise Framework

DP-Promise [22] is a recently proposed framework that adapts differential privacy to diffusion probabilistic models. Its key novelty lies in leveraging the intrinsic Gaussian noise of the diffusion process itself as part of the privacy mechanism. The framework splits the training process into two phases:

- **Phase I: Non-private training with inherent diffusion noise.** During timesteps $[S, T]$, noisy images are generated according to the forward diffusion process (cf. Equation 2.1). This step can be interpreted as applying the Gaussian mechanism for DP, since each training sample is perturbed with Gaussian noise:

  $$x^{(t)} = \sqrt{\alpha_t}\, x^{(0)} + \sqrt{1 - \alpha_t}\, z, \quad z \sim \mathcal{N}(0, I). \tag{2.4}$$

  By the post-processing property of DP, all subsequent training computations in Phase I remain privacy-preserving without additional noise injection.

- **Phase II: Private training with DP-SGD.** For earlier timesteps $[1, S-1]$, where diffusion noise is insufficient to guarantee privacy, DP-SGD is applied. Gradients are clipped and perturbed with Gaussian noise to ensure that each update satisfies (ϵ, δ)-DP.

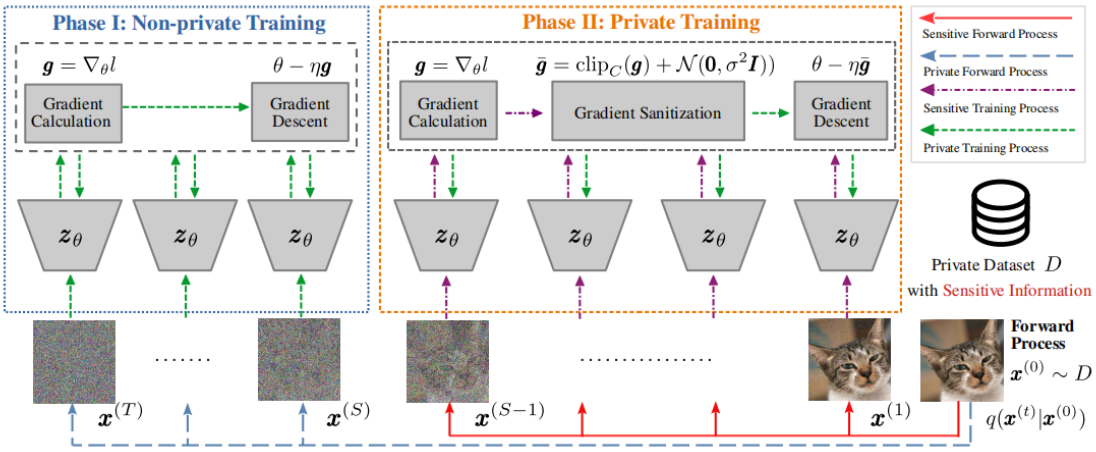


Figure 2.2: Framework of DP-Promise, which integrates differential privacy into diffusion model training by combining inherent diffusion noise (Phase I) and DP-SGD (Phase II). Adapted from Wang et al. (2022) [22].

By combining these two phases, DP-Promise ensures formal privacy guarantees across the entire diffusion training process while preserving utility. Phase I avoids unnecessary privacy budget consumption, while Phase II enforces rigorous DP guarantees where needed. This two-phase design strikes a balance between privacy and fidelity, allowing DP-Promise to outperform naive DP-SGD-only training in terms of image quality. Given these properties, DP-Promise serves as a strong baseline for evaluating the vulnerability of differentially private diffusion models to membership inference attacks in this thesis.

2.3 Alternative Diffusion Architectures: DCTDiff

While most diffusion models operate directly in the pixel domain, DCTDiff [16] introduces a novel approach by performing the diffusion process in the frequency domain using the Discrete Cosine Transform (DCT). This design is motivated by the observation that natural images exhibit significant redundancy in the pixel space but can be more compactly represented in the frequency spectrum. By operating in the DCT space, DCTDiff aims to improve generative efficiency while preserving fidelity.

2.3.1 Framework Overview

The overall architecture of DCTDiff is shown in Figure 2.3, where the image is first transformed to the frequency domain, perturbed via the forward diffusion process, denoised through a learned reverse process, and finally mapped back to the pixel space.

2.3.2 Significance

Operating in the DCT space enables DCTDiff to capture global structures more compactly, and empirically the model demonstrates strong image quality and diversity. However, unlike DP-Promise, DCTDiff does not incorporate any explicit privacy-preserving mechanism. This makes it a valuable

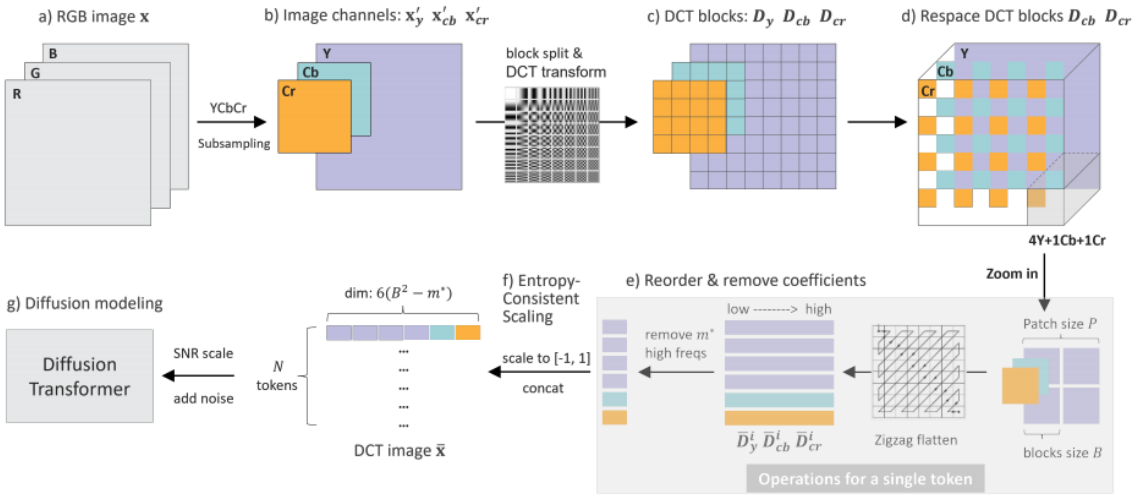


Figure 2.3: Overview of the DCTDiff framework, where diffusion occurs in the frequency domain rather than the pixel space. Adapted from Ning et al. [16].

comparator for our study: by subjecting DCTDiff to membership inference attacks, we can assess privacy risks in a diffusion model that is optimized for efficiency and fidelity but not for privacy.

2.4 Membership Inference Attacks (MIAs)

Membership Inference Attacks (MIAs) are among the most widely studied threats in machine learning privacy. The objective of an MIA is to determine whether a particular data point was used in training a model. A successful attack indicates that the model has retained traces of its training data, thereby undermining privacy guarantees and potentially exposing sensitive information.

2.4.1 General Idea

MIAs are typically categorized into two threat models:

- **Black-box MIAs:** The adversary interacts with the model only through queries, observing generated samples, probabilities, or confidence scores. In this setting, the attacker has no access to internal parameters but can still exploit systematic differences in outputs, as shown by Shokri et al. [19].
- **White-box MIAs:** The adversary has full access to the model's parameters, gradients, or intermediate representations. This stronger setting often yields more effective attacks, since memorization signals can be directly extracted from internal computations, as demonstrated by Nasr et al. [15].

2.4.2 MIAs in Generative Models

While early work focused on classifiers [19], subsequent research revealed that generative models such as GANs and VAEs are also at risk. For instance, Hayes et al. introduced LOGAN [7], one

of the first frameworks to attack GANs using MIA techniques, while Chen et al. proposed GAN-Leaks [3], which systematically categorized attack strategies against generative models. These works demonstrated that generative models trained on sensitive datasets (e.g., medical or facial data) can inadvertently encode individual examples, allowing adversaries to detect training-set inclusion. This highlighted the broader need to study privacy leakage in generative systems beyond discriminative models.

2.4.3 MIAs in Diffusion Models

Until recently, diffusion models had received little attention with respect to MIAs. A notable advancement is the **Step-wise Error Comparing Membership Inference (SecMI)** framework proposed by Luo et al. [14]. SecMI introduced a white-box attack that leverages the model's noise prediction loss as a proxy for reconstruction fidelity. By injecting noise at different timesteps and comparing reconstruction errors between training and non-training samples, the attack revealed substantial privacy leakage in diffusion models. These findings strongly motivate the present study, which evaluates the vulnerability of both differentially private and non-private diffusion models under black-box and white-box threat models.

2.5 Related Evaluation Techniques

Evaluating membership inference attacks in generative models often requires metrics that go beyond pixel-level similarity. Recent works have therefore turned to perceptual and semantic representations to capture richer signals of data influence.

CLIP (Contrastive Language–Image Pretraining) [17] provides joint embeddings of images and text, offering a semantic space in which similarities can be measured. Likewise, **LPIPS** (Learned Perceptual Image Patch Similarity) [23] has emerged as a perceptual similarity metric, designed to align more closely with human visual perception.

These techniques have become standard in evaluating privacy leakage in generative models. While their detailed application is discussed in Chapter 3 (Methodology), it is important to note here that they represent the shift from raw pixel comparisons to semantically and perceptually grounded measures in modern attack pipelines.

2.6 Gap Analysis

Despite significant progress in generative modeling and privacy-preserving machine learning, several critical gaps remain in our understanding of privacy risks in diffusion models:

- **Limited focus on diffusion models:** While Membership Inference Attacks (MIAs) have been extensively studied in classifiers [19, 15] and explored in GANs and VAEs [7, 3], systematic investigations in state-of-the-art diffusion models are scarce. Only recent work such as SecMI

[14] has begun to address this gap, revealing strong white-box vulnerabilities through noise prediction loss.

- **Absence of comparative evaluation across architectures:** Existing works typically analyze a single diffusion model in isolation. There has been little systematic comparison of privacy risks in different architectural paradigms, such as *DP-Promise* [22], which integrates Differential Privacy into training, and *DCTDiff* [16], which shifts diffusion to the frequency domain.
- **Insufficient analysis of sampling strategies:** The role of samplers such as DDIM [21], DPM-Solver [13], and Euler methods [11] has been primarily studied from the perspective of generation quality and efficiency. Their influence on privacy leakage remains unexplored, leaving open questions about whether vulnerabilities are sampler-dependent or sampler-agnostic.
- **Unclear effect of Differential Privacy budgets:** Although DP is a widely accepted defense [5], its practical impact on diffusion models is underexplored. In particular, how different privacy budgets (ϵ values) affect membership leakage in models like DP-Promise has not been systematically quantified.
- **Neglect of dataset resolution and scaling:** Most prior studies rely on small-scale benchmarks or low-resolution datasets, leaving unresolved whether privacy risks scale with resolution (e.g., 32×32 vs. 64×64) or dataset size. This is particularly relevant given the growing use of high-resolution diffusion models in practice.
- **Lack of holistic evaluation frameworks:** Existing studies typically rely on a narrow set of metrics, such as pixel-level similarity or noise-prediction loss. There remains a gap in comprehensive benchmarks that integrate both quantitative metrics (e.g., ROC-AUC, loss distributions, histograms) in black-box settings and combined quantitative–qualitative analysis (including reconstructions) in white-box settings.

Summary: The literature demonstrates that diffusion models hold significant promise for generative tasks, but remain underexplored from a privacy perspective. Comparative studies systematically spanning architectures, samplers, privacy budgets, and dataset resolutions are notably lacking, and prior work has only partially addressed the integration of black-box and white-box threat models. To the best of our knowledge, this thesis provides one of the first comprehensive empirical benchmarks of membership inference attacks in diffusion models. Specifically, we evaluate black-box attacks on both differentially private (DP-Promise)[22] and non-private (DCTDiff)[16] frameworks, while white-box analysis is conducted in detail for DP-Promise. By examining the role of samplers, ϵ-values, and image resolution, this study advances the understanding of privacy risks in state-of-the-art diffusion architectures and highlights the trade-offs between utility and privacy.

3. Methodology

This chapter outlines the methodology designed to evaluate the vulnerability of diffusion models to membership inference attacks (MIAs). The experimental design reflects both practical and theoretical considerations, enabling a comparative analysis across differentially private and non-private settings, multiple sampling strategies, and varying dataset resolutions.

The study relies on the CelebA dataset as a benchmark, chosen for its scale and diversity. Two diffusion architectures are investigated: **DP-Promise**[22], which incorporates formal differential privacy guarantees, and **DCTDiff**[16], which explores frequency-domain generative modeling without explicit privacy mechanisms. To probe privacy risks, we design two types of attacks:

- **Black-box attacks**, evaluated on both DP-Promise and DCTDiff, using feature-based comparisons of generated outputs across different samplers, privacy budgets, and resolutions.
- **White-box attacks**, evaluated exclusively on DP-Promise, leveraging direct access to the denoising network to analyze reconstruction fidelity through noise prediction loss.

Evaluation metrics are selected to capture both quantitative and qualitative evidence of leakage. In black-box settings, we employ ROC-AUC scores and histogram analysis of feature-based similarities. In the white-box setting, we extend this with reconstruction visualizations across timesteps, highlighting memorization effects.

Together, these design choices establish a rigorous methodology for assessing privacy risks in diffusion models, balancing theoretical guarantees with empirical evaluation.

3.1 Datasets

All experiments in this thesis are conducted on the CelebA dataset [12], a widely used benchmark in computer vision and generative modeling. CelebA consists of over 200,000 celebrity face images annotated with rich attribute labels. Its scale and diversity make it a suitable testbed for studying privacy risks in diffusion models, as the dataset captures significant intra-class variation while retaining recognizable identity-specific features.

3.1.1 Preprocessing and Resolutions

To align with the training requirements of DP-Promise and DCTDiff, two different image resolutions are considered:

- **CelebA-32:** Images are center-cropped and downsampled to 32×32 pixels. This resolution allows for computationally efficient experimentation and serves as a baseline for privacy analysis.
- **CelebA-64:** Images are resized to 64×64 pixels, providing higher fidelity while increasing both computational cost and the potential for memorization. This resolution is particularly relevant

for evaluating the privacy–utility trade-off in DP-Promise under stronger training settings.

3.1.2 Dataset Usage Across Models

The two diffusion frameworks under study differ in their use of resolution:

- **DP-Promise:** Trained and evaluated on both CelebA-32 and CelebA-64. This dual resolution setup enables investigation into how image resolution impacts membership inference leakage under different privacy budgets.
- **DCTDiff:** Trained and evaluated only on CelebA-64. The authors of DCTDiff focused their experiments on this resolution, and we retain consistency with their setup for reproducibility.

3.1.3 Relevance for Privacy Analysis

The choice of CelebA is motivated by its combination of large-scale availability and inherent privacy sensitivity. As the dataset contains identifiable human faces, successful membership inference attacks represent a clear privacy breach with real-world implications. The variation across resolutions further allows us to probe whether leakage scales with the richness of image detail, an important consideration for the deployment of high-resolution diffusion models in practice.

3.2 Diffusion Model Architectures

Two diffusion architectures are evaluated in this study: **DP-Promise**, which integrates differential privacy guarantees into diffusion training, and **DCTDiff**, which operates in the frequency domain without explicit privacy mechanisms. Whereas Chapter 2 introduced these models conceptually, this section provides the exact configurations and adaptations used in our experiments.

3.2.1 DP-Promise

DP-Promise [22] was evaluated on both **CelebA-32** and **CelebA-64**. The official implementation supports these datasets; however, the provided configuration files referenced ImageNet vanilla checkpoints as initialization. To maintain dataset consistency, we first performed a **vanilla diffusion pretraining step** on CelebA-32 and CelebA-64, and used the resulting checkpoints for DP-Promise training. This ensured that both pretraining and privacy-preserving phases were aligned with the CelebA data.

The DP-Promise training followed its two-phase design:

- **Phase I (non-private):** Later timesteps $[S, T]$ used the standard forward diffusion noise as a Gaussian mechanism (cf. Equation 2.4 in Chapter 2, Section 2.2.2).
- **Phase II (private):** Earlier timesteps $[1, S-1]$ applied DP-SGD with max_grad_norm = 0.01 and Gaussian noise calibrated by σ.

All DP-Promise experiments were conducted with three privacy budgets, $\epsilon \in \{1, 5, 10\}$ ($\delta = 10^{-6}$).

Training used a learning rate of 3×10^{-4}, an EMA decay of = 0.9999, and batch sizes of 32 (CelebA-32) or 16 (CelebA-64). Sampling was performed with the **DDIM sampler**, using 200 denoising steps.

The key settings were:

- **CelebA-32:** $\epsilon = 1$ ($\sigma = 3.66$), $\epsilon = 5$ ($\sigma = 1.01$), $\epsilon = 10$ ($\sigma = 0.73$); epochs = (2,30) for Phases I and II.
- **CelebA-64:** $\epsilon = 1$ ($\sigma = 2.83$), $\epsilon = 5$ ($\sigma = 0.825$), $\epsilon = 10$ ($\sigma = 0.632$); epochs = (1,15) for Phases I and II.

Here, the noise multiplier σ was scaled inversely with ϵ, reflecting the privacy–utility trade-off inherent in DP-SGD: stronger privacy (lower ϵ) requires higher noise injection, which can degrade fidelity. These controlled variations allow us to systematically study how different privacy budgets affect both generative quality and membership leakage.

3.2.2 DCTDiff

DCTDiff [16] was evaluated on the CelebA-64 dataset, following the official implementation without modification. The architecture is based on a U-ViT backbone with 12 transformer layers, embedding dimension 512, and 8 attention heads. Training was conducted for 400,000 steps using AdamW optimizer (learning rate 2×10^{-4}, weight decay 0.03, $\beta = (0.99, 0.99)$) and a customized learning-rate warmup schedule of 5000 steps.

Since DCTDiff does not incorporate differential privacy mechanisms, our evaluation focused on the effect of different samplers on potential membership leakage. To this end, experiments were conducted under three sampling strategies, each offering a distinct balance between fidelity, efficiency, and diversity. Specifically, DPM-Solver [13] provided efficient sampling in as few as 10–20 steps while maintaining fidelity, Euler-Maruyama ODE [11] served as a deterministic baseline with stable reconstructions, and Euler-Maruyama SDE [11] introduced stochastic variation to enhance diversity. For each sampler, 10,000 synthetic images were generated to ensure statistically robust evaluation of membership inference performance.

Unlike DP-Promise, DCTDiff does not enforce explicit privacy guarantees, making it a valuable non-private baseline. Its evaluation isolates whether membership leakage arises from the diffusion process itself or is influenced by differential privacy mechanisms, while also highlighting the role of different sampling strategies in shaping potential risks.

3.2.3 Summary of Configurations

To facilitate clarity, Table 3.1 summarizes the complete set of experimental configurations used in this study. For each model–dataset pair, the table specifies the privacy regime (if applicable), the training schedule, the sampling strategy, and the number of sampling steps. Clearly documenting these settings is essential, since even small differences in training schedules, privacy budgets, or sampling choices can substantially influence both generative quality and privacy leakage. The structured summary in

Table 3.1 therefore provides a concise reference point for understanding how DP-Promise and DCTDiff were instantiated in our evaluation.

Table 3.1: Summary of Experimental Configurations

Model	Dataset	Privacy	Training	Sampler	Sample Steps
DP-Promise	CelebA-32	$\epsilon = 1$	epochs (2, 30)	DDIM	200
		$\epsilon = 5$	epochs (2, 30)	DDIM	200
		$\epsilon = 10$	epochs (2, 30)	DDIM	200
DP-Promise	CelebA-64	$\epsilon = 1$	epochs (1, 15)	DDIM	200
		$\epsilon = 5$	epochs (1, 15)	DDIM	200
		$\epsilon = 10$	epochs (1, 15)	DDIM	200
DCTDiff	CelebA-64	Non-private	400K steps	DPM, Euler ODE/SDE	100

3.3 Attack Frameworks

This section presents the attack methodologies employed to evaluate privacy risks in diffusion models. Two complementary threat models are considered: (i) black-box attacks, where the adversary interacts with the model only through its outputs, and (ii) white-box attacks, where the adversary has access to the model's internal parameters and intermediate computations. By combining both perspectives, our study captures realistic and worst-case scenarios of membership inference attacks (MIAs).

3.3.1 Black-box Attacks

Motivation

The black-box setting represents the most practical threat model, as it assumes minimal adversarial capabilities. In this scenario, the attacker can only query the model (or access synthetic samples it produces) but has no knowledge of internal weights or gradients. Despite this limitation, prior work has shown that black-box MIAs can succeed by exploiting differences in how a model reproduces its training versus unseen samples. Thus, evaluating black-box vulnerability is critical for understanding the risks of publicly releasing diffusion-generated datasets.

Attack Pipeline

Our black-box attack pipeline follows four stages:

1. **Sample generation:** For DP-Promise, we generated 60,000 synthetic samples per configuration (across privacy budgets $\epsilon \in \{1, 5, 10\}$ and resolutions 32×32 and 64×64). For DCTDiff, 10,000 samples were generated separately under each of the three samplers (DPM-Solver, Euler ODE, Euler SDE).
2. **Reference sets:** The real dataset was partitioned into two disjoint subsets: members (training images) and non-members (held-out test images).

3. **Feature extraction:** Generated samples were compared against both members and non-members using three similarity metrics: raw pixel cosine similarity, CLIP-based semantic embeddings, and LPIPS perceptual distances.
4. **Score aggregation and classification:** Descriptive statistics (minimum, maximum, mean, and median) of the similarities were aggregated into attack features, which were then used to train a logistic regression classifier. Attack performance was measured using ROC-AUC and histogram analysis of score distributions.

Raw Pixel Similarity (Baseline)

As a baseline, we first evaluated similarity in the raw pixel space. Generated and real images were flattened into vectors, normalized to $[-1, 1]$, and compared using cosine similarity, the standard metric for measuring closeness in high-dimensional vector spaces:

$$s_{\text{pixel}}(x_g, x_r) = \frac{\langle x_g, x_r \rangle}{\|x_g\| \cdot \|x_r\|} \tag{3.1}$$

This equation is not introduced as a novel contribution, but rather presented here to explicitly describe the universal cosine similarity metric used in our implementation.

For each generated image x_g, the maximum similarity to the member set and non-member set was recorded. Although conceptually straightforward, this approach proved ineffective: ROC–AUC scores ranged between 0.49–0.53, essentially equivalent to random guessing. This confirmed that raw pixel matching fails to capture higher-level structure or semantics relevant for membership inference. Similarity-based MIAs of this style were initially explored in early work by Shokri et al. [19], though rarely applied in generative contexts.

CLIP-based Similarity

To overcome the limitations of raw pixel space, we employed **Contrastive Language–Image Pre-training (CLIP)** [17], a vision–language model trained on 400M image–text pairs to align images and text in a shared embedding space. The CLIP ViT-B/32 encoder was used to extract 512-dimensional semantic embeddings $f_{\text{CLIP}}(x)$ for all images. These embeddings are optimized such that semantically related images lie closer together in the embedding space. This property makes CLIP particularly suitable for detecting membership leakage: if a diffusion model has memorized specific facial attributes or identities, generated samples corresponding to training data will exhibit systematically higher embedding similarity to member images than to non-members.

To operationalize this, we compute cosine similarity between a generated image x_g and a reference image x_r:

$$s_{\text{CLIP}}(x_g, x_r) = \frac{\langle f_{\text{CLIP}}(x_g), f_{\text{CLIP}}(x_r) \rangle}{\|f_{\text{CLIP}}(x_g)\| \cdot \|f_{\text{CLIP}}(x_r)\|} \tag{3.2}$$

This formulation is not explicitly given in the CLIP paper, but reflects the standard practice of applying

cosine similarity in the learned embedding space for downstream evaluation. We include it here for clarity rather than as a novel contribution.

For each generated sample, descriptive statistics (minimum, maximum, mean, median similarities) against both members and non-members were aggregated to construct feature vectors for the attack classifier. By leveraging high-level semantic features such as expressions, attributes, and identity cues, CLIP provides a much stronger signal of membership leakage than low-level pixel metrics.

LPIPS-based Similarity

In addition to semantic similarity, we evaluated perceptual similarity using the **Learned Perceptual Image Patch Similarity (LPIPS)** metric [23]. Unlike cosine similarity in embedding space, LPIPS was motivated by psychophysical evidence that human judgments of visual similarity are better approximated by distances in the feature space of deep neural networks rather than raw pixels. LPIPS has since become a standard benchmark for assessing the fidelity and diversity of generative models, making it a natural candidate for probing potential memorization effects in diffusion sampling.

Following Zhang et al. [23], the LPIPS distance between two images x and x_0 is defined as:

$$d_{\text{LPIPS}}(x, x_0) = \sum_l \frac{1}{H_l W_l} \sum_{h,w} \left\| w_l \odot \left(\hat{y}^l_{hw}(x) - \hat{y}^l_{hw}(x_0) \right) \right\|_2^2, \tag{3.3}$$

where $\hat{y}^l$ are channel-normalized features at layer l, $H_l \times W_l$ denote the spatial dimensions, and w_l are learned channel-wise weights. In our implementation, we adopted the AlexNet-based variant with fixed $w_l = 1$, corresponding to the original LPIPS formulation without perceptual calibration. Lower distances indicate stronger perceptual similarity, making LPIPS particularly well-suited for detecting subtle memorization effects in diffusion models.

For each generated image, we compute statistics with respect to both sets (members and non-members) and both metrics (CLIP cosine similarity, LPIPS distance). Concretely, we collect *(min, max, mean, median)* of CLIP similarities to members and to non-members, and *(min, max, mean, median)* of LPIPS distances to members and to non-members, yielding a 16-dimensional feature vector per generated image. These features serve as the input representation for the final classification stage.

Classifier and Evaluation

The aggregated feature vectors were cast as a binary classification problem, where the goal was to distinguish members from non-members. We employed a logistic regression classifier (`sklearn`, max_iter=500) trained on a stratified 70/30 train–test split with a fixed random seed to ensure reproducibility. By default, the classifier applies ℓ_2 regularization, but no explicit hyperparameter tuning was performed, as the focus was on relative trends across models and settings rather than absolute optimization.

Attack performance was primarily quantified using the area under the receiver operating characteristic curve (ROC–AUC), which measures the probability that the classifier assigns a higher score to a

randomly chosen member than to a non-member. Values near 0.5 indicate chance performance, while higher values indicate successful membership inference. To complement ROC–AUC, histograms of attack scores were also plotted to visualize distributional separation between classes, providing intuitive diagnostics beyond a single scalar metric.

This black-box methodology was applied systematically to both DP-Promise (across privacy budgets $\epsilon \in \{1, 5, 10\}$ and resolutions) and DCTDiff (across sampling strategies), enabling a comprehensive cross-model comparison of vulnerability. A high-level overview of this pipeline is shown in Figure 3.1.

Summary

In summary, the black-box attack framework combines semantic (CLIP) and perceptual (LPIPS) similarity signals to augment weak pixel-level cues. After aggregating descriptive statistics across members and non-members, even a simple linear classifier is sufficient to detect leakage where present. This design highlights that diffusion models can be vulnerable to practical, low-capability adversaries.

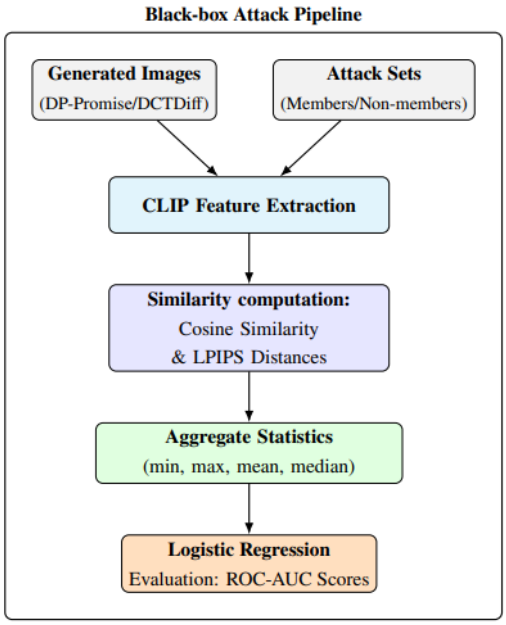


Figure 3.1: Black-box attack pipeline. Generated images are compared with member and non-member sets using CLIP and LPIPS, and aggregated features are classified with logistic regression.

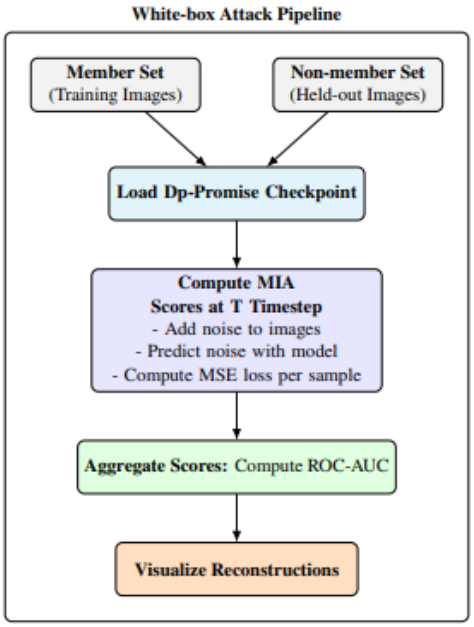


Figure 3.2: White-box attack pipeline (DP-Promise). Per-sample noise prediction losses are computed across timesteps and visualized with reconstructions.

3.3.2 White-box Attacks

Motivation

The white-box setting assumes the strongest adversary, with unrestricted access to the model's parameters, gradients, and internal activations. While such a threat model is less common in practice, it

provides an upper bound on privacy leakage and highlights worst-case vulnerabilities. Recent work, such as the Step-wise Error Comparing Membership Inference (SecMI) framework [14], demonstrated that diffusion models can be probed by analyzing their denoising behavior across timesteps. Building on this idea, we design a white-box attack tailored to the DP-Promise model, exploiting its internal noise prediction loss as a discriminative signal.

Attack Pipeline

Our white-box attack follows four main stages:

1. **Input preparation:** Member and non-member images are selected from the training and held-out test sets, respectively. Each image is normalized and transformed into the model's input space.
2. **Noise injection:** For a given timestep $t \in [0, T]$, Gaussian noise is sampled and added to the clean image x_0, producing a noisy sample x_t according to the forward diffusion process.
3. **Noise prediction:** The DP-Promise denoising network predicts the added noise $\hat{\epsilon}_\theta(x_t, t)$ from the noisy input. This prediction is compared against the ground-truth sampled noise ϵ.
4. **Scoring:** The prediction error is used as the membership inference signal, with lower errors expected for members than for non-members.

Loss-based Scoring

In the white-box setting, we adopt the model's denoising loss as the membership signal. Diffusion models are trained to predict the Gaussian noise added at each timestep, optimized via a mean squared error (MSE) objective [9]. Following the SecMI framework [14], we compute this prediction error on a per-sample basis and use it directly as the attack score. The intuition is that if an image was present during training, the model should predict its noise more accurately, resulting in a lower error compared to non-member images. By evaluating this loss across timesteps, we obtain a quantitative signal of memorization that distinguishes members from non-members.

Timestep Analysis and Score Aggregation

To capture leakage across the generative process, the denoising loss is evaluated at multiple timesteps $t \in \{0, 100, 200, \ldots, 999\}$. This produces distributions of scores for members and non-members at each t, enabling both pointwise discrimination and temporal analysis. For each timestep, we compute ROC–AUC and visualize histograms of loss distributions. Aggregating results across timesteps reveals where leakage is most pronounced.

Reconstruction-based Diagnostics

Beyond scalar loss values, we also assess memorization qualitatively by reconstructing clean images from their noisy versions using the model's predicted noise. This process follows the reverse denoising formulation of diffusion models [9], and was adapted for membership inference analysis in the SecMI

framework [14]. In practice, reconstructions of member images tend to preserve sharper facial details and identity cues, while reconstructions of non-members degrade more quickly under noise. These visualizations provide intuitive, human-interpretable evidence of privacy leakage in the white-box setting.

Classifier and Evaluation

In contrast to the black-box setting, no separate classifier is required. Instead, the per-sample denoising loss is directly used as the attack score. Evaluation follows the same protocol as before, with ROC–AUC as the primary quantitative metric and histograms as complementary diagnostics. The white-box attack was applied exclusively to DP-Promise, as its differential privacy mechanisms made it the most relevant case for worst-case analysis. A schematic of this pipeline is shown in Figure 3.2.

Summary

The white-box attack leverages the model's internal denoising loss as a direct membership signal. By evaluating prediction errors across timesteps and complementing them with qualitative reconstructions, this framework provides both quantitative and visual insights into potential memorization. Unlike the black-box setting, it assumes a stronger adversary with full access to model internals, serving as an upper bound on privacy leakage in diffusion models.

Taken together, the black-box and white-box attacks provide complementary perspectives: the black-box case highlights practical risks under limited access, while the white-box case reveals worst-case leakage when full model knowledge is assumed. We next describe the evaluation metrics that provide a common basis for systematically comparing the black-box and white-box attack results (Section 3.4).

3.4 Evaluation Metrics

To systematically compare the effectiveness of membership inference attacks (MIAs), we employ a set of quantitative and qualitative evaluation metrics. These metrics provide scalar measures of discriminative performance as well as distributional and visual diagnostics. Together, they form a comprehensive basis for evaluating privacy risks without relying solely on single-number summaries.

3.4.1 ROC–AUC (Primary Metric)

The **area under the receiver operating characteristic curve (ROC–AUC)** is our primary metric. Originally introduced in classification and widely adopted in security and privacy research, ROC–AUC has become the standard measure of MIA effectiveness [19, 20, 24]. It quantifies the probability that an attack assigns a higher score to a randomly chosen member than to a randomly chosen non-member.

Formally, ROC–AUC evaluates the trade-off between true positive rate (TPR) and false positive rate (FPR) across all decision thresholds. A value of 0.5 corresponds to chance-level discrimina-

tion, while values significantly greater than 0.5 indicate that the attack features contain information about membership. This property makes ROC–AUC particularly suitable for MIAs, where threshold calibration is not straightforward and the focus is on separability of distributions rather than fixed accuracy. We report ROC–AUC consistently across all experiments, ensuring comparability between attack settings (black-box vs. white-box), datasets, and privacy budgets.

3.4.2 Histogram Analysis

While ROC–AUC condenses performance into a scalar, it does not convey the shape of score distributions. Following prior work on membership inference [19, 24], we therefore plot histograms of attack scores for members and non-members. These visualizations provide distributional diagnostics that complement ROC–AUC, revealing whether separability arises from global shifts, heavier tails, or other structural differences. Histograms are included for both black-box similarity-based scores and white-box loss-based scores, allowing intuitive inspection of how member and non-member distributions diverge.

3.4.3 Reconstruction-based Visualization (White-box Only)

In the white-box setting, we supplement scalar metrics with qualitative reconstructions. Specifically, starting from a noisy image x_t at timestep t, the model predicts the injected noise $\hat{\epsilon}_\theta(x_t, t)$. Subtracting this predicted noise from the noisy sample yields an approximate reconstruction of the clean image. This process follows the reverse denoising formulation of diffusion models [9], and was adapted for membership inference in the SecMI framework [14]. By inspecting reconstructions across timesteps, one can assess how effectively the model recovers identity-preserving details from noisy inputs. Reconstructions thus serve as a complementary, human-interpretable diagnostic of potential memorization, without replacing quantitative metrics.

3.4.4 Statistical Robustness

To account for stochastic variation in both training and evaluation, blackbox experiments were repeated across multiple random seeds. Results are reported as *medians* over seeds, following established practice in membership inference evaluation [19, 24]. For whitebox experiments, which are largely deterministic given fixed checkpoints and timesteps, robustness was assessed by evaluating a large cohort of member and nonmember images across the full diffusion trajectory.

Confidence intervals and interquartile ranges were not computed due to runtime constraints, but reproducibility is supported by documenting all seeds, configurations, and code (see Appendix).

4. Experimental Results

This chapter presents the empirical evaluation of the diffusion models. Generative quality is first assessed using Inception Score (IS) and Fréchet Inception Distance (FID), which provide a baseline measure of model utility. Subsequently, privacy risks are examined through membership inference attacks (MIAs) under black-box and white-box settings, offering a comprehensive perspective on the balance between generation fidelity and privacy leakage.

4.1 Generated Image Quality

The quality and diversity of generated samples are evaluated using two widely adopted metrics in generative modeling: the **Inception Score (IS)** [18], which reflects image diversity and semantic consistency, and the **Fréchet Inception Distance (FID)** [8], which measures the distributional similarity between generated and real images. Higher IS and lower FID values indicate better generative performance.

DP-Promise

Table 4.1 reports IS and FID scores for DP-Promise across different privacy budgets ($\epsilon \in \{1, 5, 10\}$) and image resolutions (32×32, 64×64). Results indicate that relaxing the privacy budget generally improves generative quality (lower FID, slightly higher IS), with the effect being more pronounced at lower resolution.

Table 4.1: IS and FID scores for DP-Promise (CelebA 32×32 and 64×64) across privacy levels.

Model	**Metric**	**CelebA 32×32**			**CelebA 64×64**		
		$\epsilon = 1$	$\epsilon = 5$	$\epsilon = 10$	$\epsilon = 1$	$\epsilon = 5$	$\epsilon = 10$
DP-Promise	IS	2.58	2.61	**2.61**	2.32	2.43	**2.48**
	FID	17.87	14.60	**13.59**	43.49	34.34	**32.24**

DCTDiff

For DCTDiff, we evaluate FID scores across different samplers (DPM-Solver, Euler Maruyama ODE, Euler Maruyama SDE) and varying numbers of function evaluations (NFE). Results are shown in Table 4.2. FID improves consistently with increasing NFE for all samplers, with DPM-Solver and Euler ODE clearly outperforming Euler SDE. Based on these results, we adopt $NFE = 100$ for subsequent privacy analysis, as it balances generative quality with computational efficiency.

Taken together, these results confirm that while DP-Promise sacrifices some generative quality due to differential privacy, it produces samples of sufficient fidelity to enable privacy analysis. For DCTDiff, increasing NFE consistently improves sample quality, and we use $NFE = 100$ in subsequent

Table 4.2: FID scores for DCTDiff (CelebA 64×64) across samplers and NFE.

Model	NFE	DPM-Solver	Euler ODE	Euler SDE
DCTDiff	10	13.77	64.45	359.36
	20	6.26	7.73	304.09
	50	5.92	5.24	107.13
	100	**5.89**	**5.28**	**45.21**

experiments as it provides competitive performance without incurring excessive cost. Both models therefore provide a sound basis for evaluating membership inference attacks.

4.2 Black-box Attack Results

Black-box membership inference results are presented next, assuming an adversary who has access only to generated samples. Two approaches are evaluated: (i) raw pixel similarity, used as a baseline, and (ii) a combined feature-based approach leveraging CLIP semantic embeddings and LPIPS perceptual distances. The former tests whether low-level pixel alignment suffices for membership inference, while the latter captures higher-level semantic and perceptual cues that are more relevant for generative models.

4.2.1 Raw Pixel Similarity (DP-Promise only)

As a baseline, we evaluated membership inference under raw pixel similarity. Following the black-box pipeline described in Section 3.3, generated images were compared against member and non-member reference sets using both cosine similarity and Euclidean distance in flattened pixel space. Table 4.3 reports ROC–AUC scores for DP-Promise across privacy budgets ($\epsilon \in \{1, 5, 10\}$) and image resolutions (32×32, 64×64).

Table 4.3: ROC–AUC scores of raw pixel similarity (cosine and Euclidean) for DP-Promise. Scores near 0.5 indicate chance-level discrimination.

Dataset	Cosine Similarity			Euclidean Distance		
	$\epsilon = 1$	$\epsilon = 5$	$\epsilon = 10$	$\epsilon = 1$	$\epsilon = 5$	$\epsilon = 10$
CelebA 32×32	0.5003	0.5029	0.5037	0.5012	0.4995	0.4997
CelebA 64×64	0.5242	0.5145	0.5117	0.5307	0.5171	0.5127

The results show that raw pixel similarity fails to distinguish members from non-members, with ROC–AUC values consistently close to 0.5 across all settings. This confirms that pixel-level comparisons capture neither the semantic structure nor the perceptual cues needed for effective membership inference. Consequently, raw pixel evaluation is reported only for DP-Promise and not extended to DCTDiff, as the metric does not provide meaningful discrimination. This finding aligns with prior work (e.g., Shokri et al. [19]), where raw features alone proved insufficient for MIAs in high-dimensional domains.

Given these limitations, subsequent analyses focus on CLIP-based semantic similarity and LPIPS perceptual distance, which together provide stronger and more interpretable signals of privacy leakage.

4.2.2 CLIP and LPIPS Similarity

Black-box membership inference is evaluated using combined CLIP semantic embeddings and LPIPS perceptual distances (Section 3.3). Unlike raw pixel similarity, these features capture higher-level semantic and perceptual cues, and are therefore expected to provide stronger discrimination between members and non-members. Both DP-Promise and DCTDiff are analyzed under this setting.

Feature-Based Attack on DP-Promise

Table 4.4 reports ROC–AUC scores for initial runs under seeds 42 and 1234. Experiments were conducted on CelebA 32×32 and 64×64 across different privacy budgets ($\epsilon \in \{1, 5, 10\}$) with varying numbers of generated images and sampled members/non-members.

Table 4.4: ROC–AUC scores for DP-Promise black-box attacks with CLIP+LPIPS features, under seeds 42 and 1234. Settings vary by resolution, number of generated images (Gen), and number of member/non-member samples (M/NM).

Dataset	**Gen (M/NM)**	**Seed 1234**			**Seed 42**		
		$\epsilon = 1$	$\epsilon = 5$	$\epsilon = 10$	$\epsilon = 1$	$\epsilon = 5$	$\epsilon = 10$
CelebA 32	1k (500/500)	0.6271	0.5979	0.5914	0.9745	0.9758	0.9760
	10k (100/100)	0.7752	0.7725	0.7718	0.9043	0.9131	0.9149
	10k (200/200)	0.7185	0.6978	0.6935	0.6631	0.6626	0.6583
	60k (100/100)	0.7736	0.7741	0.7717	0.9095	0.9179	0.9190
CelebA 64	1k (500/500)	0.6810	0.6961	0.6578	0.8895	0.8187	0.8198
	10k (100/100)	0.9575	0.9493	0.9473	0.7545	0.7258	0.7290
	10k (200/200)	0.8676	0.8613	0.8532	0.7509	0.7393	0.7209
	60k (100/100)	0.9600	0.9530	0.9509	0.7618	0.7331	0.7341

Beyond smaller-scale evaluations (1k and 10k generated images), testing was extended to 60k samples against a fixed 100/100 member/non-member split to examine whether larger comparisons clarified trends across privacy budgets. However, even at this scale, no consistent monotonic trend with respect to ϵ emerged (revisited in Chapter 5). Moreover, fluctuations across seeds persisted, underscoring the role of randomness in member/non-member selection. To mitigate seed-specific variability, experiments were repeated across ten random seeds [10, 21, 37, 42, 73, 85, 1234, 2024, 4096, 9999], each using 100 members and 100 non-members compared against 1000 generated samples per run. Median ROC–AUC values are reported in Table 4.5, with full per-seed results in Appendix A.2. Following common practice in MIA evaluation [20, 2], reporting medians across seeds reduces randomness from sample selection and yields a more stable estimate of leakage. Overall, results confirm consistently above-chance inference, though with variability depending on the random seed.

Table 4.5: Median ROC–AUC scores for DP-Promise across 10 random seeds (100 member, 100 non-member, 1000 generated).

Dataset	**CelebA 32**			**CelebA 64**		
	$\epsilon = 1$	$\epsilon = 5$	$\epsilon = 10$	$\epsilon = 1$	$\epsilon = 5$	$\epsilon = 10$
ROC–AUC (Median)	0.8425	0.8445	0.8416	0.8960	0.8474	0.8340

ROC–AUC Curves and Histograms (DP-Promise): Figure 4.1 shows cumulative ROC–AUC curves across the ten random seeds, with the median run highlighted. The x-axis denotes the false positive rate (FPR) and the y-axis the true positive rate (TPR). Curves near the diagonal correspond to chance-level discrimination, while those approaching the top-left corner indicate stronger separation. For illustration, results are shown for a representative case (CelebA 32, $\epsilon = 1$), chosen because it provides a balanced view of leakage without extreme separation. Across seeds, ROC–AUC values ranged from 0.59 to 0.94 (Appendix A.2), with the median stabilizing near 0.84 (Table 4.5). This variability underscores the stochasticity of subset sampling, even when attack features are semantically rich.

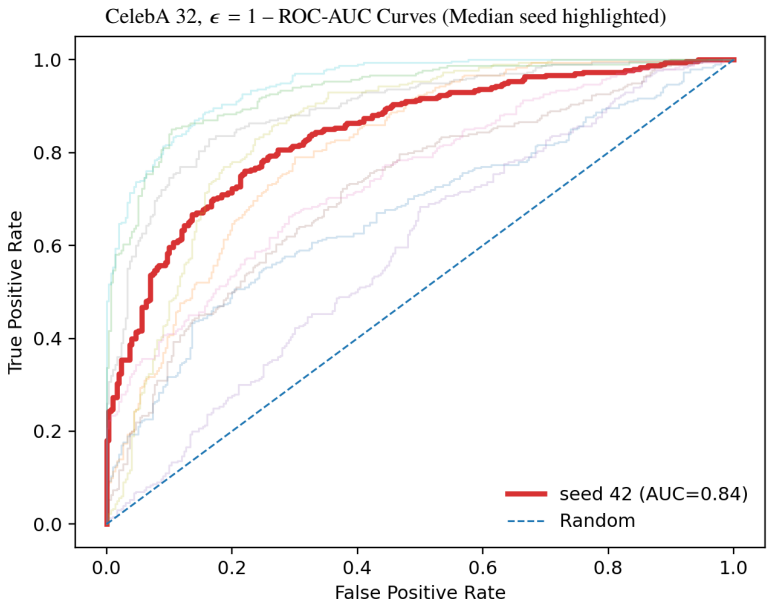


Figure 4.1: Cumulative ROC–AUC curves for DP-Promise (CelebA 32, $\epsilon = 1$) across 10 random seeds. The highlighted curve corresponds to the median run.

Histograms of attack scores provide complementary distributional insight. Figure 4.2 shows member vs. non-member score distributions for the same representative case (CelebA 32, $\epsilon = 1$), aggregated across seeds in a grid layout. In some seeds, distributions are well separated, consistent with higher ROC–AUC values (e.g., ≈0.93), whereas in others, overlaps dominate, corresponding to weaker scores (≈0.59). This pattern is consistent with the spread observed across seeds (Appendix A.2), while the aggregated median stabilizes near 0.84 as reported in Table 4.5. Comprehensive histograms for all privacy budgets and resolutions are included in Appendix A.2.

In summary, CLIP+LPIPS features yield substantially stronger membership inference than raw pixels, with ROC–AUC well above 0.5 across both resolutions. However, leakage varies by seed and shows

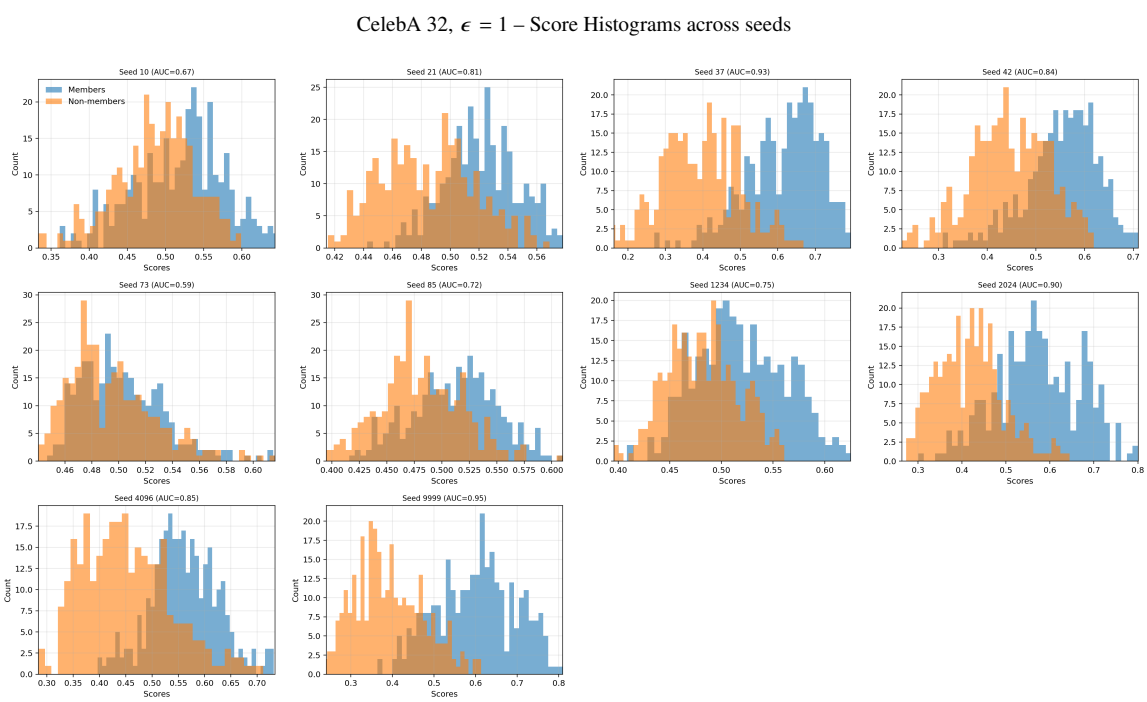


Figure 4.2: Histograms of member vs. non-member attack scores for DP-Promise (CelebA 32, $\epsilon = 1$) across 10 random seeds.

no consistent dependence on ϵ, suggesting that while the attack is effective, its behavior under different privacy budgets is more nuanced and warrants further discussion (see Chapter 5).

Feature-Based Attack on DCTDiff

Black-box membership inference on DCTDiff is evaluated using CLIP+LPIPS features. Unlike DP-Promise, DCTDiff is a non-private model, and testing was conducted exclusively on CelebA 64×64, using three samplers: DPM-Solver, Euler Maruyama ODE, and Euler Maruyama SDE. Initial experiments were performed with seeds 42 and 1234 across multiple evaluation scales (1k generated with 500/500 member/non-member samples, 10k with 100/100 and 200/200 splits). Results are reported in Table 4.6.

Table 4.6: ROC–AUC scores for DCTDiff black-box attacks with CLIP+LPIPS features, under seeds 42 and 1234. Abbreviations: DPM = DPM-Solver, ODE = Euler Maruyama ODE, SDE = Euler Maruyama SDE; Gen = number of generated images, M/NM = members/non-members.

Dataset	**Gen (M/NM)**	**Seed 1234**			**Seed 42**		
		DPM	ODE	SDE	DPM	ODE	SDE
CelebA 64	1k (500/500)	0.7847	0.7740	0.9112	0.6745	0.7455	0.6021
	10k (100/100)	0.8477	0.8685	0.9053	0.5863	0.5895	0.7914
	10k (200/200)	0.8136	0.8143	0.7841	0.7548	0.7371	0.9108

As with DP-Promise, results exhibited fluctuations across seeds, reflecting randomness in sample selection. To mitigate this, evaluation was extended to the same ten random seeds [10, 21, 37, 42, 73, 85, 1234, 2024, 4096, 9999] used previously. Each run compared 100 members and 100 non-

members against 1000 generated samples. Median ROC–AUC values are reported in Table 4.7, with full per-seed results provided in Appendix A.2. Following common practice in MIA evaluation [20, 2], reporting medians across seeds reduces variability and yields a more reliable estimate of leakage.

Table 4.7: Median ROC–AUC scores for DCTDiff across 10 random seeds (100 member, 100 non-member, 1000 generated).

Dataset	**CelebA 64**		
	DPM-Solver	**Euler ODE**	**Euler SDE**
ROC–AUC (Median)	0.7991	0.7935	0.7675

ROC–AUC Curves and Histograms (DCTDiff): Figure 4.3 shows cumulative ROC–AUC curves across ten random seeds, with the median run highlighted. Here, the focus is on the DPM-Solver sampler, which achieved the most stable median performance (≈0.80; Table 4.7). Unlike DP-Promise, this section does not repeat the ROC–AUC formalism; instead, the analysis emphasizes that most runs remained consistently above chance, though variability across seeds was evident (minimum ≈0.55, maximum ≈0.88). Full ROC-AUC curves for Euler ODE and Euler SDE are provided in Appendix A.2.

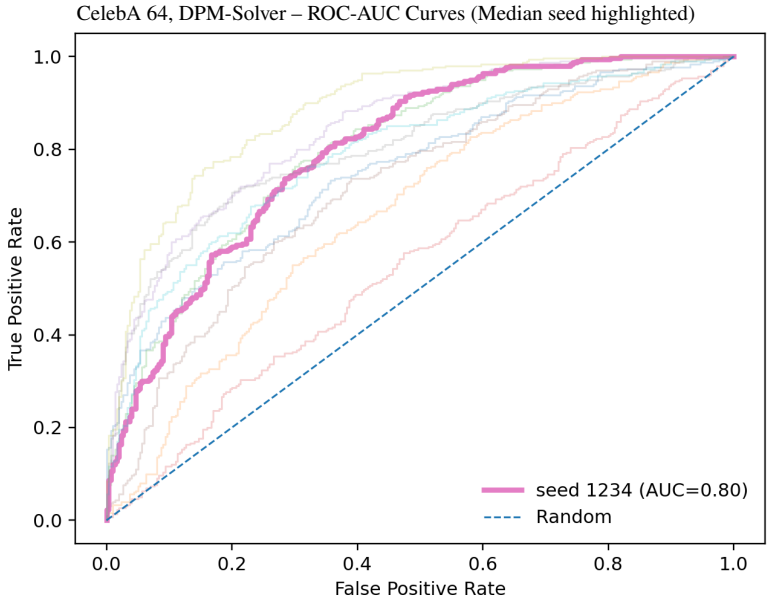


Figure 4.3: Cumulative ROC–AUC curves for DCTDiff (CelebA 64, DPM-Solver) across 10 random seeds. The highlighted curve corresponds to the median run.

Histograms of member vs. non-member scores, shown in Figure 4.4, provide parallel distributional evidence. In the median seed, members tend to concentrate at higher similarity scores than non-members, consistent with a median ROC–AUC of ≈ 0.80 for DPM-Solver (Table 4.7). Across seeds, however, separation varies substantially, with ROC–AUC spanning roughly 0.55–0.88 for DPM-Solver (Appendix A.2). This variability echoes the DP-Promise results and further motivates reporting medians over multiple seeds. Comprehensive histograms for all three samplers are included in Appendix A.2.

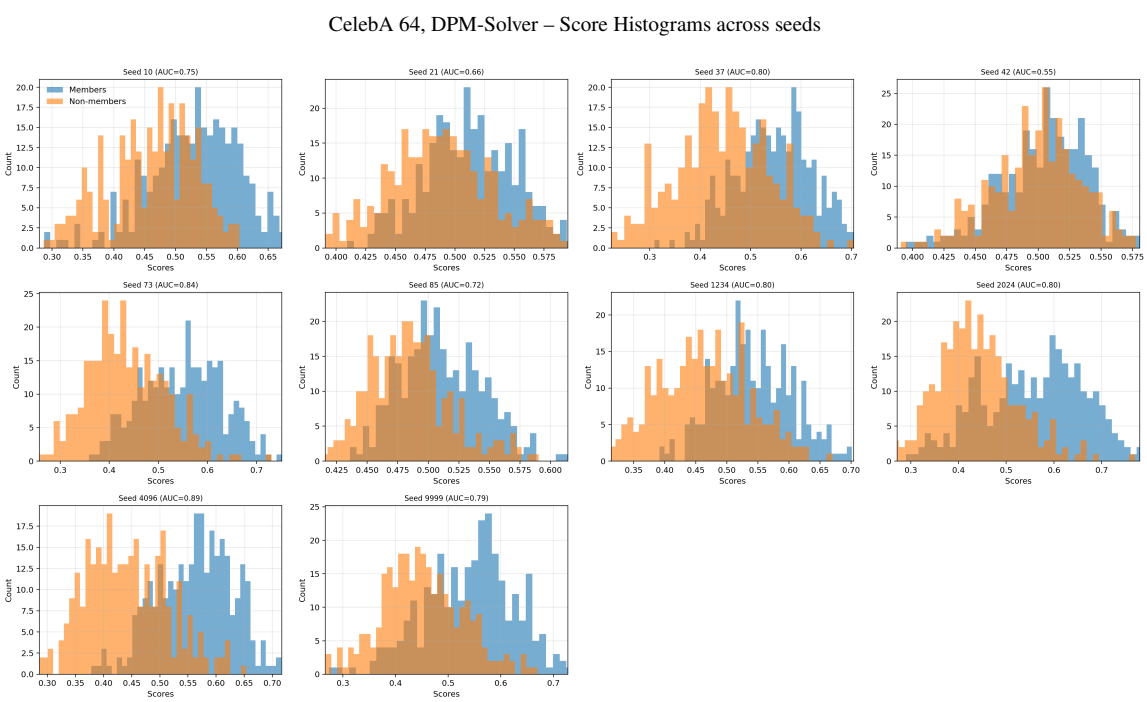


Figure 4.4: Histograms of member vs. non-member attack scores for DCTDiff (CelebA 64, DPM-Solver) across 10 random seeds.

In summary, CLIP+LPIPS features reveal above-chance membership inference in DCTDiff across all samplers, with median ROC–AUC values around 0.77–0.80. No sampler consistently outperformed the others, and observed variability suggests that, under the present experimental scope, all three are comparably vulnerable. These results establish that DCTDiff, despite lacking differential privacy mechanisms, exhibits leakage patterns similar in magnitude to DP-Promise. Broader implications of these findings are discussed in Chapter 5.

4.2.3 Cross-Model Comparison (DP-Promise vs. DCTDiff)

This subsection provides a comparative view of the two models under feature-based black-box attacks. Although DP-Promise incorporates formal differential privacy guarantees and DCTDiff serves as a non-private baseline, both demonstrate above-chance membership inference. Median ROC–AUC values for DP-Promise ranged from 0.83 to 0.90 across resolutions and privacy budgets (Table 4.5), while DCTDiff scores stabilized around 0.77–0.80 across samplers (Table 4.7).

Two common patterns emerge. First, leakage magnitudes for the two models are of comparable scale, with neither model consistently outperforming the other in terms of resistance to attack. Second, both settings show marked variability across random seeds, highlighting the sensitivity of attack outcomes to sample selection. Unlike DP-Promise, DCTDiff does not vary with ϵ, yet neither model exhibits strong dependence on its respective parameters (ϵ for DP-Promise, sampler choice for DCTDiff).

Together, these observations establish a consistent empirical baseline: feature-based membership inference remains feasible in both private and non-private diffusion models. The broader implications for differential privacy and model robustness are examined further in Chapter 5.

4.3 White-box Attack Results (DP-Promise)

This section evaluates membership inference in a white-box setting, where an attacker can access internal model signals. Following Section 3.3.2, a *loss-based* score is computed from the model's per-step denoising predictions: for a noisy sample x_t, the network predicts $\hat{\epsilon}_\theta(x_t, t)$, and the attack score is the negative per-sample loss (lower loss $\Rightarrow$ higher membership likelihood). Checkpoints are probed across the diffusion trajectory at timesteps $t \in \{0, 100, \ldots, 900, 999\}$. *Unless otherwise noted, each configuration was evaluated over 40,000 members and 40,000 non-members to obtain per-timestep score distributions.*

4.3.1 Loss-based ROC–AUC Across Timesteps

Table 4.8 reports ROC–AUC for CelebA 32×32 and 64×64 under $\epsilon \in \{1, 5, 10\}$ across eleven timesteps. A consistent pattern emerges: discrimination is strong at early denoising (t=100–200, ROC–AUC $\approx$ 0.87–0.92), weakens at intermediate steps ($t\approx$300–600, $\approx$ 0.58–0.65), and strengthens again at late steps ($t\geq$700, up to $\approx$ 1.0). This non-monotonic trajectory indicates that membership signals are most pronounced when the model's predictions align with stable structural cues (early) or re-enter the low-noise regime (late), while mid-trajectory noise confounds the loss signal.

Table 4.8: White-box ROC–AUC for DP-Promise across timesteps (t) using loss-based scores. Left: CelebA 32×32; Right: CelebA 64×64.

Timestep$_t$	**CelebA 32×32**			**CelebA 64×64**		
	$\epsilon = 1$	$\epsilon = 5$	$\epsilon = 10$	$\epsilon = 1$	$\epsilon = 5$	$\epsilon = 10$
0	0.5855	0.5727	0.5678	0.6568	0.6448	0.6389
100	0.8772	0.8745	0.8702	0.9181	0.9083	0.9060
200	0.7659	0.7768	0.7771	0.8555	0.8441	0.8409
300	0.6448	0.6606	0.6628	0.7365	0.7266	0.7226
400	0.5790	0.5915	0.5936	0.6489	0.6474	0.6426
500	0.5584	0.5699	0.5721	0.6140	0.6179	0.6120
600	0.5852	0.5941	0.5937	0.6272	0.6244	0.6156
700	0.7064	0.7106	0.6983	0.7186	0.7239	0.7126
800	0.9003	0.9080	0.8913	0.9337	0.9354	0.9309
900	0.9927	0.9987	0.9986	0.9988	0.9997	0.9997
999	1.0000	1.0000	1.0000	0.9998	1.0000	1.0000

ROC-AUC Curves Across Timesteps

Figure 4.5 overlays ROC-AUC curves for all eleven timesteps (one color per t) for a representative configuration (CelebA 64, $\epsilon = 1$).[1] The x-axis is false positive rate (FPR), the y-axis true positive rate (TPR). Curves move away from the diagonal at t=100–200, drift closer at mid-trajectory ($t\approx$300–600),

[1] 64×64 is shown for visual clarity; full results for 32×32 and all privacy budgets appear in Appendix A.2.

and then tighten toward the top-left again for $t \geq 700$, with t=900 and t=999 nearly saturating the plot.

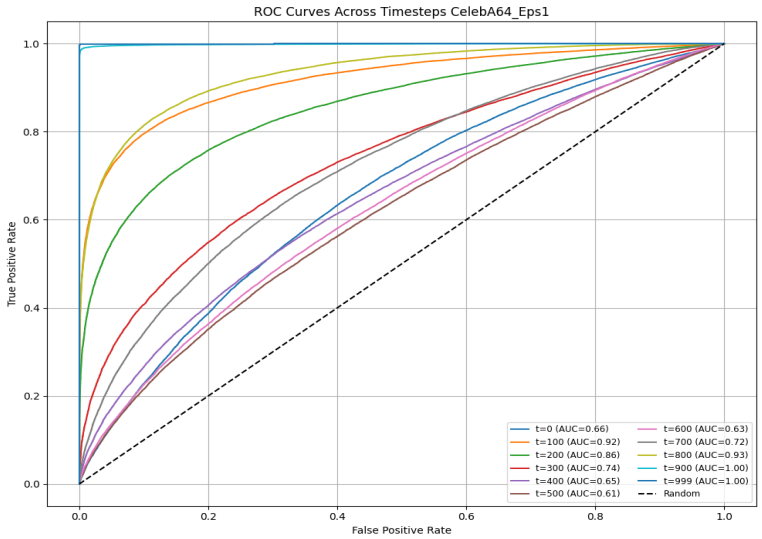


Figure 4.5: White-box ROC-AUC overlays across timesteps ($t \in \{0, \ldots, 999\}$) for DP-Promise (CelebA 64, $\epsilon = 1$). Early (t=100–200) and late ($t \geq 700$) steps show stronger separability; mid-range ($t \approx 300$–600) weakens.

Score Histograms Across Timesteps

Figure 4.6 shows member vs. non-member score histograms for the same setting (CelebA 64, $\epsilon = 1$). Panels are ordered left-to-right and top-to-bottom by timestep: the first panel corresponds to t=0, followed by t=100, t=200, and so on; the final panels cover t=900 and t=999.

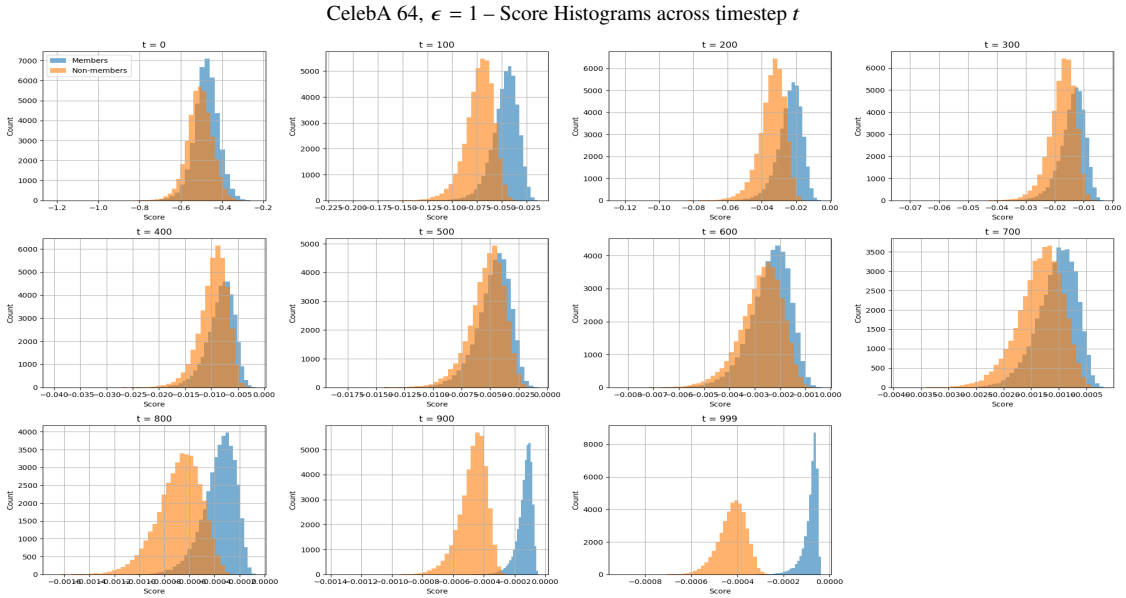


Figure 4.6: Member vs. non-member loss-score histograms across timesteps (DP-Promise, CelebA 64, $\epsilon = 1$). Panels proceed by timestep left-to-right, top-to-bottom. Separation is strongest at t=100–200 and $t \geq 700$.

Member distributions shift toward higher scores at t=100–200, contract toward non-members at mid-trajectory, and separate again at $t \geq 700$, matching Table 4.8 and the ROC-AUC overlays. Full-resolution grids for all ϵ and both resolutions are included in Appendix A.2.

4.3.2 Reconstruction-based Diagnostics

To visualize how signals manifest, reconstructions are obtained by subtracting predicted noise $\hat{\epsilon}_\theta(x_t, t)$ along the reverse trajectory. As illustrated in Figure 4.7, side-by-side trajectories for a member (left) and a non-member (right) at CelebA 64, $\epsilon = 1$ reveal the contrast that drives the white-box results: in late steps ($t \geq 900$), reconstructed *members* retain coarse facial structure despite heavy corruption, whereas *non-members* often remain noise-like, consistent with the near-1.0 ROC–AUC. At early steps (t=100–200), member identities also become discernible sooner than non-members, aligning with the early AUC peak and the score-separation patterns observed in the histograms.

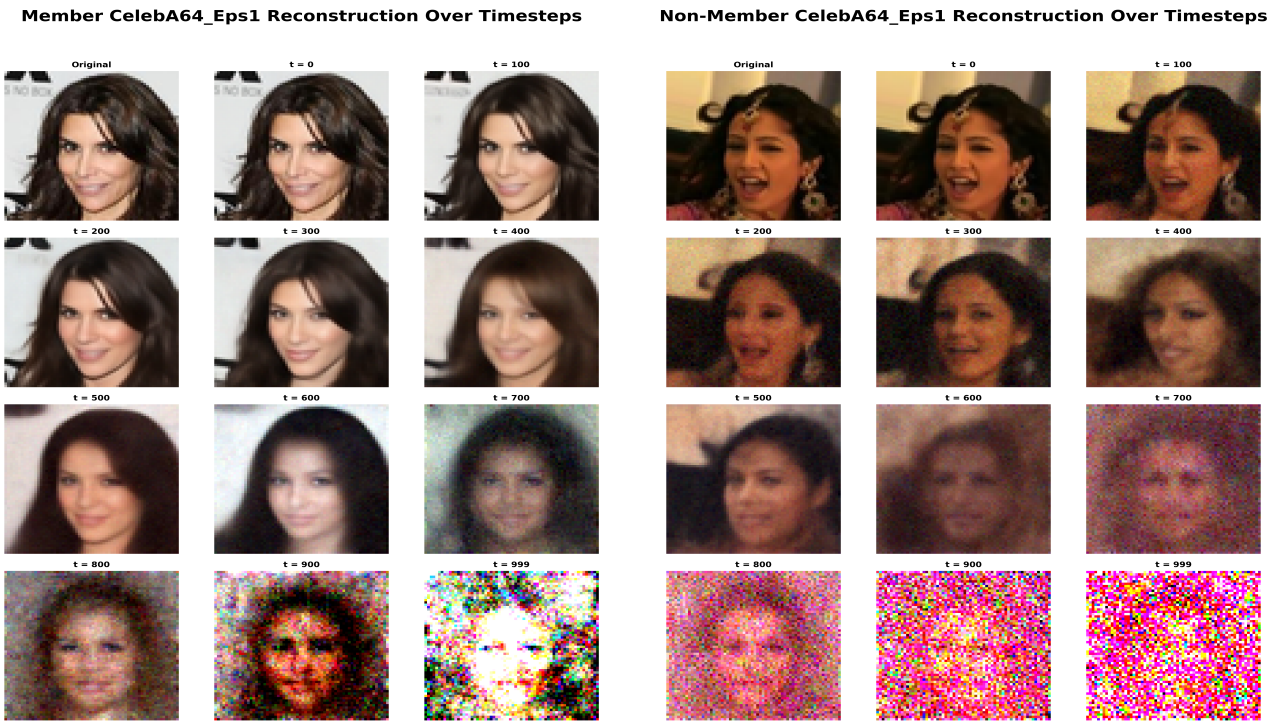


Figure 4.7: Reconstruction trajectories across timesteps t for a *member* (left) and a *non-member* (right) under DP-Promise (CelebA 64, $\epsilon = 1$). Members preserve facial structure earlier and more reliably, while non-members remain noise-like at late t, aligning with the ROC/histogram trends. Additional examples for all settings are provided in Appendix A.2.

White-box analysis thus reveals strong leakage signals at specific timesteps, with near-perfect separability for $t \geq 900$ across resolutions and privacy budgets, and a secondary peak at t=100–200. These results represent an *upper bound* on practical risk, as full internal access is uncommon in deployment, but they are consistent with SecMI-style white-box methodologies [14] and point to late-step (and early-step) memorization effects that warrant targeted defenses. Broader implications are discussed in Chapter 5.

Taken together, the experiments show that feature-based black-box MIAs succeed against both private and non-private diffusion models, while loss-based white-box signals expose near-perfect separability at late timesteps and a secondary early-step peak. Chapter 5 interprets these findings, their implications for privacy budgets and sampler choices, and the limits of current evaluation practice.

5. Discussion

This chapter interprets the experimental findings from Chapter 4, highlighting what was observed, what could not be confirmed, and how these results connect to broader debates on the privacy of diffusion models. The discussion is organized into four parts: key findings, implications, limitations, and broader context.

5.1 Key Findings

Membership inference remains feasible in both models

Feature-based black-box attacks (CLIP+LPIPS) yielded above-chance ROC–AUC in both DP-Promise ($\approx$ 0.83–0.90) and DCTDiff ($\approx$ 0.77–0.80). Despite DP-Promise incorporating formal differential privacy, leakage remained comparable in magnitude to the non-private baseline. This does not contradict DP theory, but highlights that empirical leakage is measurable even when privacy guarantees are in place.

No consistent monotonicity with privacy budgets or samplers

Across a broad set of runs—including 1k, 10k, and 60k generated samples, varying member/non-member splits, and ten random seeds—no monotonic trend emerged with respect to ϵ for DP-Promise. Similarly, leakage in DCTDiff showed little dependence on sampler choice, despite clear differences in generative quality (FID). These results cannot prove that leakage is inherently non-monotonic, but within the scope of tested conditions, no trend was identified. More extensive sweeps of privacy budgets, samplers, and training schedules are required to draw stronger conclusions.

White-box access reveals stronger signals

Loss-based, white-box evaluations on DP-Promise exposed near-perfect separability at late timesteps ($t \geq 900$) and a secondary peak at early steps (t=100–200). These signals were substantially stronger than black-box feature-based attacks, emphasizing the heightened risk when internal model states or per-sample losses are accessible.

Leakage is sensitive to random seeds

Across both models, results varied significantly between seeds, with ROC–AUC spanning from near chance ($\approx$ 0.55) to high separation ($\approx$ 0.95). Reporting medians across ten seeds provided a stable estimate, consistent with best practices in the membership inference literature.

5.2 Implications

Generative quality does not preclude leakage

For DP-Promise, CelebA 64×64 achieved FID scores of 43, 34, and 32 for ϵ=1, 5, 10, while CelebA 32×32 obtained 17, 14, and 13 under the same budgets. Although these values reflect substantial differences in generative quality across resolutions, both settings produced leakage of comparable magnitude. Similarly, DCTDiff samplers exhibited large variation in FID yet yielded similar ROC–AUC ranges. In particular, sampling compute was varied (NFE=10, 20, 50, 100), with NFE=100 selected for downstream attacks based on its superior FID. Despite this improvement in sample quality, leakage magnitudes remained within a similar range across samplers, reinforcing that higher fidelity alone does not mitigate membership inference.

Privacy budgets alone may not suffice

Since no consistent monotonic trend in leakage was found across ϵ, relying solely on the privacy budget as a measure of practical protection is problematic. While DP guarantees remain valid at the theoretical level, empirical risk requires supplementary audits. In practice, conservative ϵ values should be paired with post-hoc leakage evaluation across multiple seeds and sample sizes.

Access boundaries are critical

White-box leakage underscores the importance of restricting access to internal model signals. Even if black-box leakage is moderate, exposing per-sample losses or intermediate denoising states could lead to near-perfect inference. Practical deployments should enforce strong API boundaries to limit this risk.

5.3 Limitations

Dataset scope

All experiments were conducted on CelebA at 32×32 and 64×64 resolutions. Results may differ for higher resolutions or datasets with different characteristics, such as medical or text-to-image data.

Training and evaluation constraints

DP-Promise training followed the repository's two-phase schedule, but with resolution-specific settings: CelebA 32×32 used 2 epochs in Phase 1 and 30 epochs in Phase 2, whereas CelebA 64×64 used 1 epoch and 15 epochs, respectively. The relatively higher FID at 64×64 (versus 32×32) may partly reflect this schedule and optimization stability constraints. For DCTDiff, the model was trained for 400K steps. Sampler effectiveness was benchmarked across $NFE \in \{10, 20, 50, 100\}$; NFE=100

yielded the best FID and was adopted for the privacy analyses. Evaluation budgets required limiting member/non-member sizes (e.g., 100/100 vs. 200/200) and the seed count to ten; larger-scale runs could further tighten leakage estimates and confidence intervals.

Attack surface

Only membership inference was analyzed. Other privacy risks such as attribute inference, inversion, or data extraction were not examined. Likewise, black-box evaluation relied on CLIP+LPIPS features, and white-box analysis focused on loss-based scores; stronger attacks may yield different outcomes.

5.4 Broader Context

Position in the privacy literature

The findings echo prior work showing that (i) MIAs remain effective across modalities [19, 20], (ii) empirical leakage under DP does not always align with the theoretical budget [2], and (iii) white-box signals are especially diagnostic [14]. This study adds evidence specific to diffusion models under both DP and non-DP training.

Practical considerations

For practitioners, two recommendations emerge: (i) treat synthetic data release as carrying residual risk, even under DP; and (ii) design deployment interfaces to avoid exposing internal states.

Standardizing empirical privacy audits

The results suggest several practical conventions for diffusion models: (i) report medians across multiple random seeds; (ii) evaluate across multiple sample sizes and member/non-member splits to probe instability; (iii) include both black-box (feature-based) and white-box (loss-based) perspectives when feasible; and (iv) document training schedules, sampler settings, and NFE alongside FID/IS to enable fair comparisons. Such reporting complements formal DP guarantees with reproducible evidence of empirical risk.

In summary, the experiments demonstrate that membership inference remains feasible in both private and non-private diffusion models. Empirical leakage did not align predictably with ϵ or with sampler choice, though substantially stronger signals were observed under white-box access. These findings highlight the importance of careful empirical auditing, transparent reporting of evaluation settings, and conservative deployment practices. Building on these observations, Chapter 6 outlines directions for future work aimed at broadening experimental scope and investigating defenses against the identified leakage modes.

6. Future Work

The experiments in this thesis establish that diffusion models, whether trained with or without differential privacy, are vulnerable to membership inference. While these findings provide a strong empirical baseline, several avenues remain for future work to broaden scope, strengthen evaluation, and inform defenses.

Broader experimental scope

This study focused on CelebA at 32×32 and 64×64 resolutions. Future evaluations should extend to higher resolutions and diverse domains such as medical imaging or text-to-image, where privacy risks may manifest differently. Equally, leakage in DP-Promise did not exhibit monotonic dependence on ϵ within the tested range. Further sweeps of privacy budgets, training schedules, and optimizer settings could clarify whether this reflects a deeper property or an artifact of the experimental setup. For DCTDiff, sampler choice was varied in terms of NFE, but other hyperparameters warrant exploration. Together, these extensions would test the generality of observed leakage patterns and sharpen understanding of how training and evaluation parameters influence privacy risk.

Expanding attack surfaces

Only membership inference was studied here. Future work should extend to complementary threats such as attribute inference, inversion, and data extraction, which may exploit different vulnerabilities. Such attacks would provide a more complete picture of privacy risks and clarify whether differentially private training mitigates some classes of threats more effectively than others.

Standardization and defenses

Results highlighted the importance of reporting leakage medians across random seeds and documenting training and evaluation settings. Future work should formalize these practices into standardized benchmarks for generative privacy research, enabling fairer comparisons and more reproducible evidence of risk. In parallel, defenses must move beyond measurement toward mitigation, including regularization to reduce memorization, alternative training objectives, or hybrid approaches combining DP with post-training audits. Balancing privacy, fidelity, and computational cost remains a key challenge for practical deployment.

Looking ahead, future work should broaden datasets, attacks, and parameter sweeps while moving toward standardized benchmarks and practical defenses. Such steps are necessary to ensure that diffusion models can be deployed with greater confidence in their privacy guarantees.

7. Conclusion

This thesis examined privacy risks in diffusion models through the lens of membership inference attacks (MIAs). While diffusion models have emerged as state-of-the-art generative techniques, their adoption in sensitive domains depends critically on understanding whether privacy guarantees such as differential privacy (DP) are sufficient in practice. To this end, two representative models were analyzed: DP-Promise, which incorporates formal DP guarantees, and DCTDiff, a non-private baseline.

The evaluation combined generative quality metrics (FID, IS) with privacy risk assessments under black-box and white-box MIAs. Black-box experiments demonstrated that feature-based attacks (CLIP+LPIPS) achieved above-chance discrimination in both private and non-private models, with leakage magnitudes broadly comparable across settings. No consistent monotonic dependence on privacy budgets or samplers was observed, and results varied considerably across seeds, reinforcing the importance of reporting medians over multiple runs. White-box experiments on DP-Promise, in contrast, revealed substantially stronger signals, particularly at early and late denoising timesteps, highlighting the elevated risks when internal states are exposed.

Taken together, the results show that while DP mechanisms supported generative training under privacy constraints, relaxing ϵ improved fidelity but did not prevent membership leakage. At the same time, generative fidelity gains (e.g., through sampler choice or ϵ relaxation) did not correspond to reduced vulnerability. These findings stress the need for careful auditing of diffusion models, transparent reporting of experimental settings, and conservative deployment practices when synthetic data are intended to protect individual privacy.

Looking forward, broader evaluation across datasets, parameter ranges, and attack modalities is necessary to refine understanding of privacy leakage in diffusion models. Such extensions, together with standardization of empirical privacy audits, will be essential for aligning theoretical guarantees with empirical evidence of privacy risk in practice. The future work outlined in Chapter 6 provides a roadmap for these directions.

A. Appendix

A.1 GitHub Repository

The code and experimental pipelines developed for this thesis are available in the GitHub repository:

https://git.cs.bham.ac.uk/projects-2024-25/kxr414

The repository is organized into a set of Jupyter notebooks, each corresponding to a stage of the workflow:

- **DP-Promise_Model_Eval.ipynb** – trains the DP-Promise model and generates synthetic images.
- **DCTDiff_Model_Eval.ipynb** – trains the non-private DCTDiff model and generates synthetic images.
- **DP-Promise_Black_Box_Attack.ipynb** – implements black-box membership inference attacks on DP-Promise.
- **DCTDiff_Black_Box_Attack.ipynb** – implements black-box membership inference attacks on DCTDiff.
- **DP-Promise_White_Box_Attack.ipynb** – implements white-box membership inference attacks on DP-Promise.

This structure reflects the experimental workflow: model training and synthesis, followed by black-box attacks on both models and white-box analysis for DP-Promise. A detailed `README.md` file is included in the repository to guide environment setup and execution.

A.2 Additional Figures and Tables

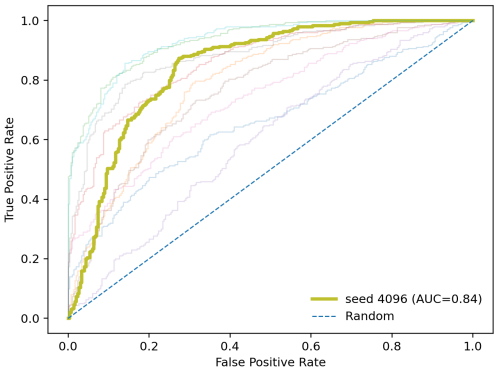
Black-box Results: DP-Promise and DCTDiff

Table A.1: ROC–AUC scores for DP-Promise and DCTDiff across random seeds. DP-Promise scores shown for CelebA 32 and 64 at $\epsilon = \{1, 5, 10\}$. DCTDiff scores shown for CelebA 64 with samplers: DPM (DPM-Solver), ODE (Euler Maruyama ODE), and SDE (Euler Maruyama SDE).

Random Seed	**DP-Promise CelebA 32**			**DP-Promise CelebA 64**			**DCTDiff CelebA 64**		
	$\epsilon = 1$	$\epsilon = 5$	$\epsilon = 10$	$\epsilon = 1$	$\epsilon = 5$	$\epsilon = 10$	DPM	ODE	SDE
10	0.6710	0.6583	0.6537	0.8465	0.7934	0.8018	0.7522	0.7550	0.6828
21	0.8085	0.7936	0.7880	0.7280	0.7685	0.7581	0.6623	0.7034	0.7179
37	0.9312	0.9315	0.9280	0.9278	0.8480	0.8439	0.8016	0.7935	0.6300
42	0.8425	0.8596	0.8593	0.6958	0.6490	0.6456	0.5502	0.5679	0.7801
73	0.5902	0.6001	0.6035	0.8130	0.7697	0.7737	0.8440	0.8384	0.7675
85	0.7207	0.7622	0.7708	0.5931	0.5843	0.5907	0.7190	0.7034	0.6201
1234	0.7463	0.7387	0.7302	0.9308	0.9033	0.9127	0.7991	0.8178	0.8588
2024	0.8954	0.8829	0.8763	0.8960	0.8474	0.8340	0.8035	0.8373	0.7058
4096	0.8516	0.8445	0.8416	0.9202	0.8934	0.9113	0.8853	0.8881	0.8466
9999	0.9475	0.9316	0.9272	0.9035	0.8765	0.8647	0.7892	0.7915	0.7924
Mean	0.8005	0.8003	0.7979	0.8255	0.7933	0.7937	0.7606	0.7696	0.7402
Std Dev	0.1166	0.1117	0.1103	0.1167	0.1057	0.1065	0.0966	0.0921	0.0830
Median	0.8425	0.8445	0.8416	0.8960	0.8474	0.8340	0.7991	0.7935	0.7675

CelebA 32 - $\epsilon = 5$

CelebA 32, $\epsilon = 5$ – ROC-AUC Curves (Median seed highlighted)



CelebA 32, $\epsilon = 5$ – Histogram across seeds

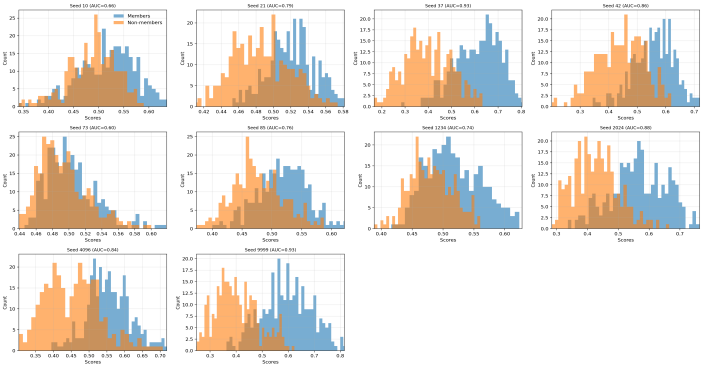
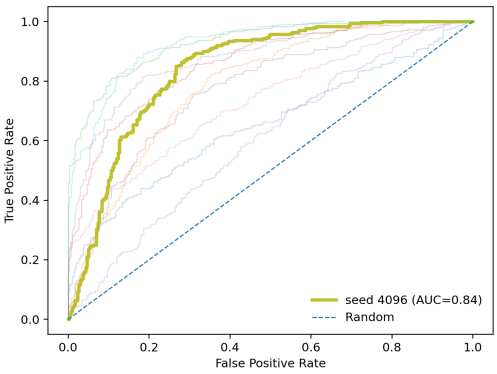


Figure A.1: Black-Box: ROC-AUC curves and histograms across seeds (DP-Promise, CelebA 32×32, $\epsilon = 5$

CelebA 32 - $\epsilon = 10$

CelebA 32, $\epsilon = 10$ – ROC-AUC Curves (Median seed highlighted)



CelebA 32, $\epsilon = 10$ – Histogram across seeds

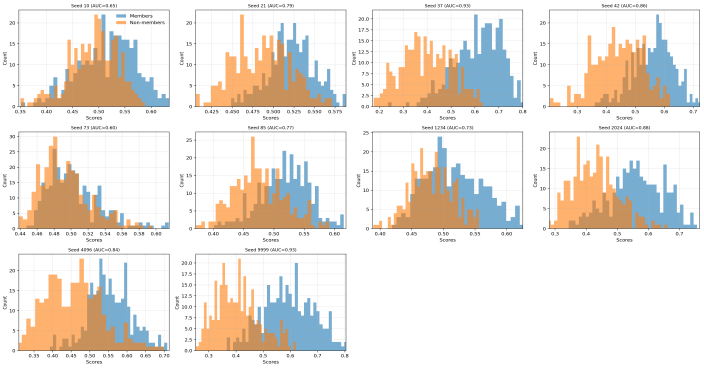
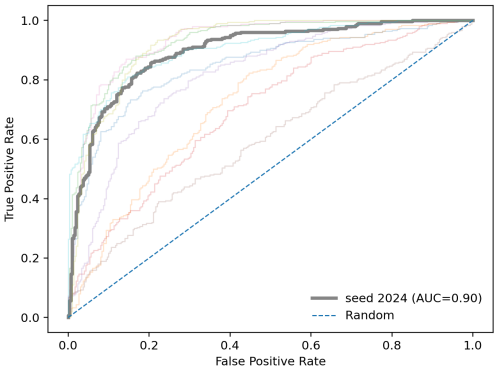


Figure A.2: Black-Box: ROC-AUC curves and histograms across seeds (DP-Promise, CelebA 32×32, $\epsilon = 10$

CelebA 64 - $\epsilon = 1$

CelebA 64, $\epsilon = 1$ – ROC-AUC Curves (Median seed highlighted)



CelebA 64, $\epsilon = 1$ – Histogram across seeds

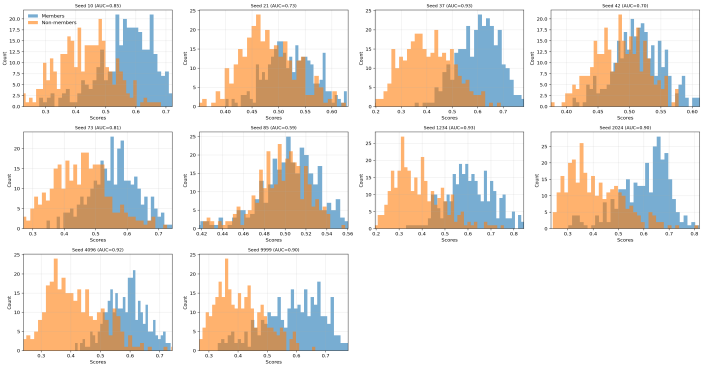
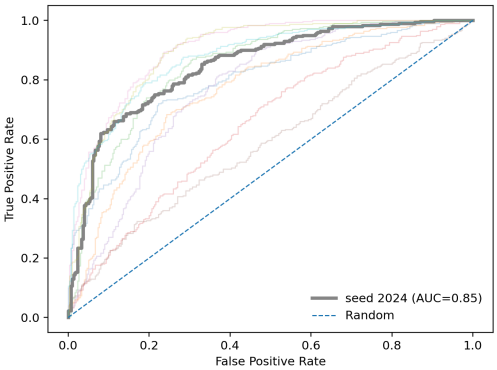


Figure A.3: Black-Box: ROC-AUC curves and histograms across seeds (DP-Promise, CelebA 64×64, $\epsilon = 1$

CelebA 64 - $\epsilon = 5$

CelebA 64, $\epsilon = 5$ – ROC-AUC Curves (Median seed highlighted)



CelebA 64, $\epsilon = 5$ – Histogram across seeds

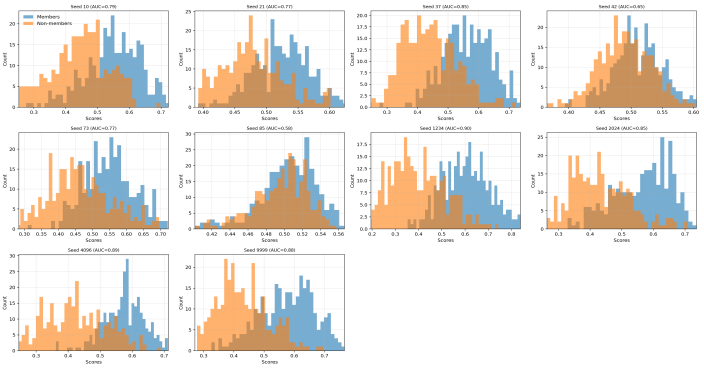
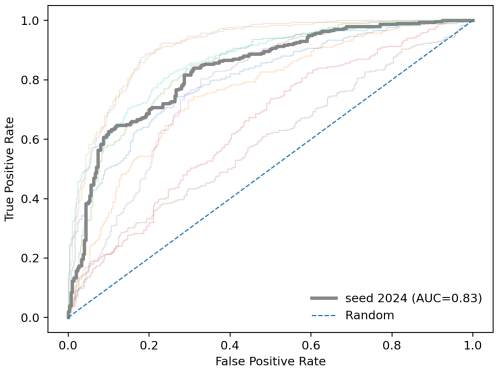


Figure A.4: Black-Box: ROC-AUC curves and histograms across seeds (DP-Promise, CelebA 64×64, $\epsilon = 5$

CelebA 64 - $\epsilon = 10$

CelebA 64, $\epsilon = 10$ – ROC-AUC Curves (Median seed highlighted)



CelebA 64, $\epsilon = 10$ – Histogram across seeds

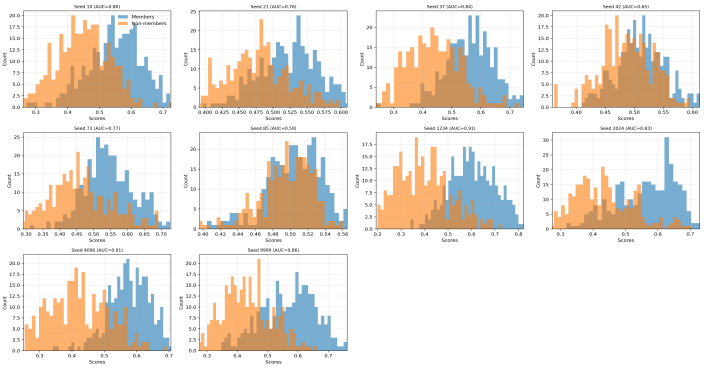
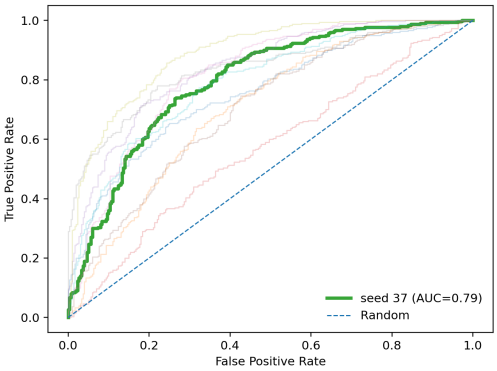


Figure A.5: Black-Box: ROC-AUC curves and histograms across seeds (DP-Promise, CelebA 64×64, $\epsilon = 10$

CelebA 64 - Euler ODE

CelebA 64, ODE – ROC-AUC Curves (Median seed highlighted)



CelebA 64, Euler ODE – Histogram across seeds

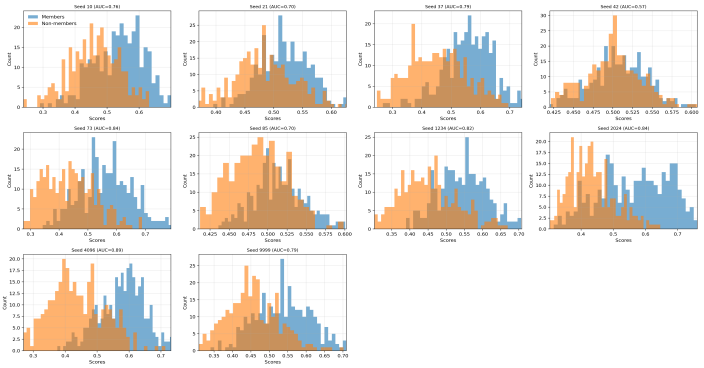
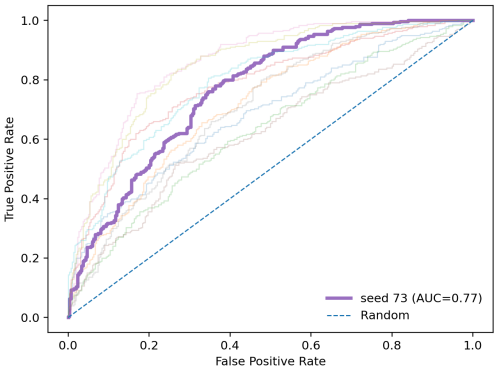


Figure A.6: Black-Box: ROC-AUC curves and histograms across seeds (DCTDiff, CelebA 64×64, Euler ODE

CelebA 64 - Euler SDE

CelebA 64, SDE – ROC-AUC Curves (Median seed highlighted)



CelebA 64, Euler SDE – Histogram across seeds

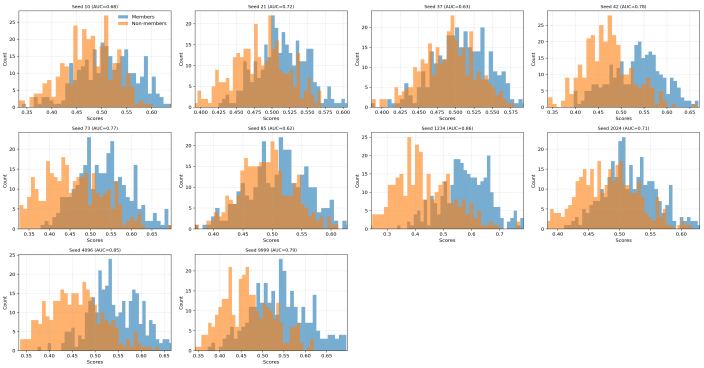
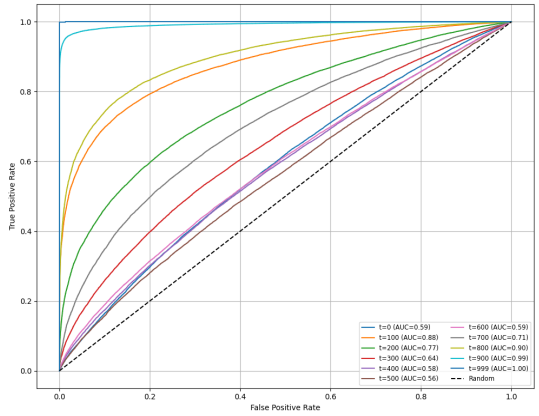


Figure A.7: Black-Box: ROC-AUC curves and histograms across seeds (DCTDiff, CelebA 64×64, Euler SDE

White-box Results: DP-Promise

ROC-AUC Curve (CelebA 32 - $\epsilon \in \{1, 5, 10\}$)

CelebA 32, $\epsilon = 1$ – ROC-AUC Curves across timesteps



CelebA 32, $\epsilon = 1$ – Histogram across timesteps

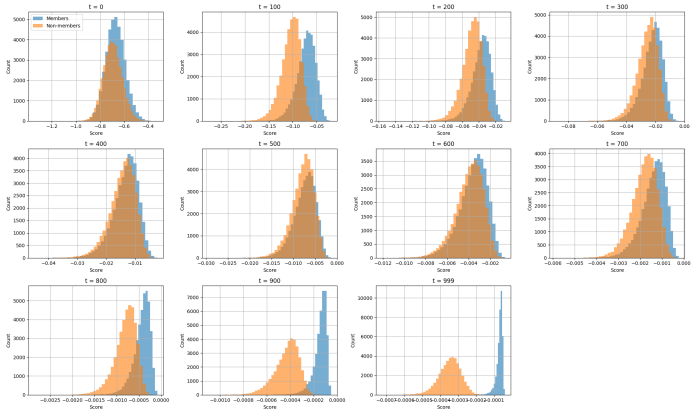
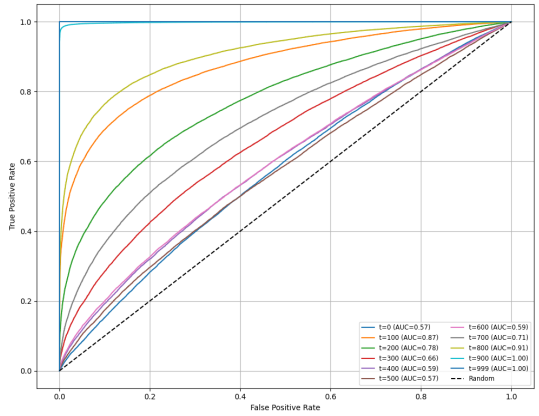


Figure A.8: White-Box: ROC-AUC curves and histograms across t timesteps (CelebA 32×32, $\epsilon = 1$)

CelebA 32, $\epsilon = 5$ – ROC-AUC Curves across timesteps



CelebA 32, $\epsilon = 5$ – Histogram across timesteps

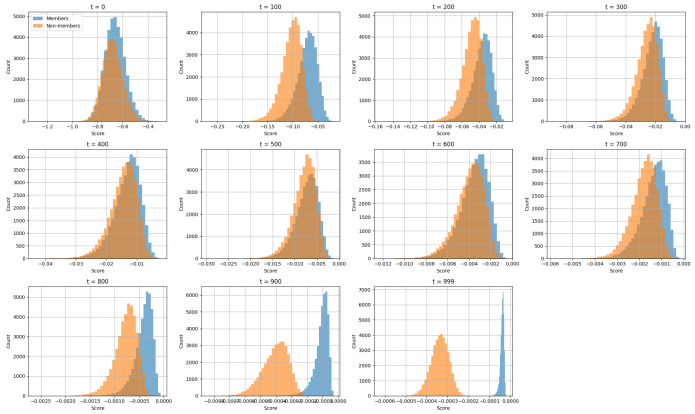
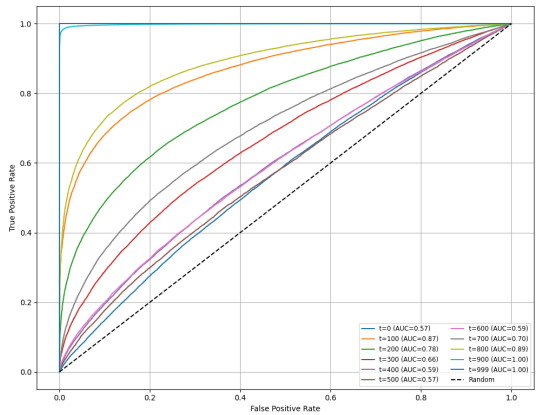


Figure A.9: White-Box: ROC-AUC curves and histograms across t timesteps (CelebA 32×32, $\epsilon = 5$)

CelebA 32, $\epsilon = 10$ – ROC-AUC Curves across timesteps



CelebA 32, $\epsilon = 10$ – Histogram across timesteps

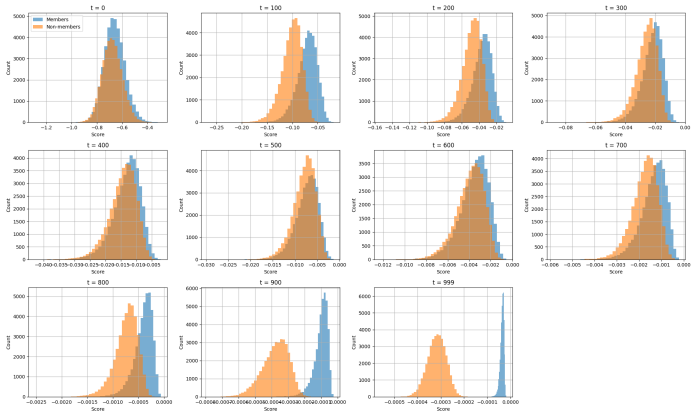


Figure A.10: White-Box: ROC-AUC curves and histograms across t timesteps (CelebA 32×32, $\epsilon = 10$)

Reconstruction Visual (CelebA 32 - $\epsilon \in \{1, 5, 10\}$**)**

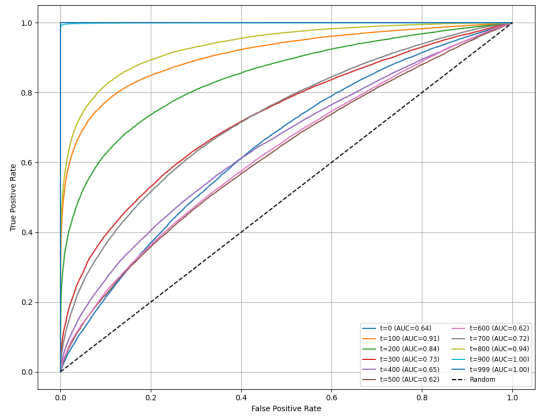
Figure A.11: Reconstruction trajectories across timesteps t for a *member* (left) and a *non-member* (right) under DP-Promise (CelebA 32, $\epsilon = 1$).

Figure A.12: Reconstruction trajectories across timesteps t for a *member* (left) and a *non-member* (right) under DP-Promise (CelebA 32, $\epsilon = 5$).

Figure A.13: Reconstruction trajectories across timesteps t for a *member* (left) and a *non-member* (right) under DP-Promise (CelebA 32, $\epsilon = 10$).

CelebA 64 - $\epsilon \in \{5, 10\}$

CelebA 64, $\epsilon = 5$ – ROC-AUC Curves across timesteps



CelebA 64, $\epsilon = 5$ – Histogram across timesteps

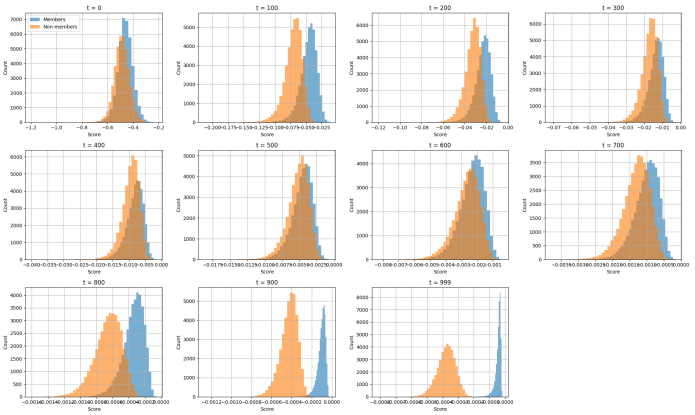
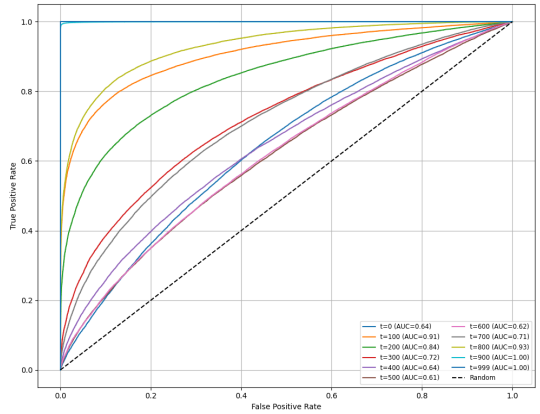


Figure A.14: White-Box: ROC-AUC curves and histograms across t timesteps (CelebA 64×64, $\epsilon = 5$)

CelebA 64, $\epsilon = 10$ – ROC-AUC Curves across timesteps



CelebA 64, $\epsilon = 10$ – Histogram across timesteps

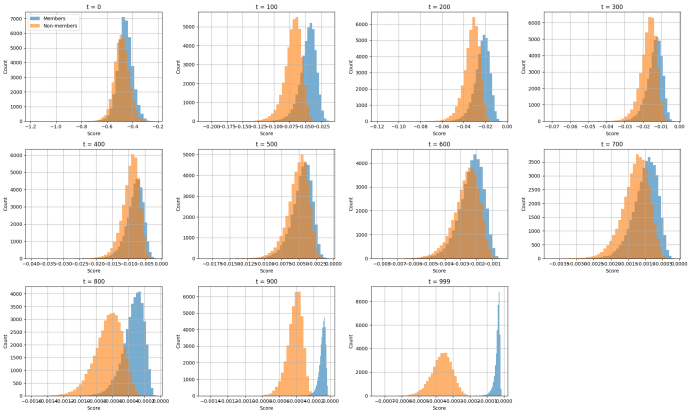


Figure A.15: White-Box: ROC-AUC curves and histograms across t timesteps (CelebA 64×64, $\epsilon = 10$)

Figure A.16: Reconstruction trajectories across timesteps t for a *member* (left) and a *non-member* (right) under DP-Promise (CelebA 64, $\epsilon = 5$).

Figure A.17: Reconstruction trajectories across timesteps t for a *member* (left) and a *non-member* (right) under DP-Promise (CelebA 64, $\epsilon = 10$).

References

[1] Martin Abadi et al. “Deep Learning with Differential Privacy”. In: *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (CCS)*. ACM. 2016, pp. 308–318.

[2] Nicholas Carlini et al. “Membership Inference Attacks From First Principles”. In: *2022 IEEE Symposium on Security and Privacy (SP)*. 2022, pp. 1897–1914. DOI: 10.1109/SP46214.2022.9833869.

[3] Danial Chen, Lingjiao Chen, and Neil Zhenqiang Gong. “GAN-Leaks: A Taxonomy of Membership Inference Attacks against Generative Models”. In: *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS)*. 2020, pp. 343–362. DOI: 10.1145/3372297.3417238.

[4] Cynthia Dwork and Aaron Roth. *The Algorithmic Foundations of Differential Privacy*. Vol. 9. 3–4. Foundations and Trends in Theoretical Computer Science, 2014, pp. 211–407. DOI: 10.1561/0400000042.

[5] Cynthia Dwork et al. “Calibrating Noise to Sensitivity in Private Data Analysis”. In: *Theory of Cryptography Conference (TCC)*. 2006.

[6] Ian Goodfellow et al. “Generative Adversarial Nets”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2014.

[7] Jamie Hayes et al. “LOGAN: Membership Inference Attacks Against Generative Models”. In: *Proceedings on Privacy Enhancing Technologies (PoPETs)*. Vol. 2019. 1. 2019, pp. 133–152. DOI: 10.2478/popets-2019-0008.

[8] Martin Heusel et al. “GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2017.

[9] Jonathan Ho, Ajay Jain, and Pieter Abbeel. “Denoising Diffusion Probabilistic Models”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2020.

[10] Diederik P Kingma and Max Welling. “Auto-Encoding Variational Bayes”. In: *arXiv preprint arXiv:1312.6114* (2013). Available: https://arxiv.org/abs/1312.6114.

[11] Peter E. Kloeden and Eckhard Platen. “Stochastic Differential Equations and Applications”. In: *Springer* (1992).

[12] Ziwei Liu et al. “Deep Learning Face Attributes in the Wild”. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2015, pp. 3730–3738.

[13] Cheng Lu et al. “DPM-Solver: A Fast ODE Solver for Diffusion Probabilistic Model Sampling in Around 10 Steps”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2022.

[14] Ziang Luo et al. “Are Diffusion Models Vulnerable to Membership Inference Attacks?” In: *arXiv preprint arXiv:2302.01316* (2023). Available: https://arxiv.org/abs/2302.01316.

[15] Milad Nasr, Reza Shokri, and Amir Houmansadr. “Comprehensive Privacy Analysis of Deep Learning: Passive and Active White-box Inference Attacks against Centralized and Federated Learning”. In: *Proceedings of the IEEE Symposium on Security and Privacy (S&P)*. 2019, pp. 739–753. DOI: 10.1109/SP.2019.00065.

[16] Mang Ning et al. “DCTDiff: Intriguing Properties of Image Generative Modeling in the DCT Space”. In: *arXiv preprint arXiv:2505.04512* (2025). Available: https://arxiv.org/abs/2505.04512.

[17] Alec Radford et al. “Learning Transferable Visual Models From Natural Language Supervision”. In: *International Conference on Machine Learning (ICML)*. 2021.

[18] Tim Salimans et al. “Improved Techniques for Training GANs”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2016.

[19] Reza Shokri et al. “Membership Inference Attacks Against Machine Learning Models”. In: *Proceedings of the IEEE Symposium on Security and Privacy (S&P)*. 2017, pp. 3–18. DOI: 10.1109/SP.2017.41.

[20] Congzheng Song, Thomas Ristenpart, and Vitaly Shmatikov. “Auditing Data Provenance in Text-Generation Models”. In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD)*. ACM, 2019, pp. 196–206. DOI: 10.1145/3292500.3330885.

[21] Jiaming Song, Chenlin Meng, and Stefano Ermon. “Denoising Diffusion Implicit Models”. In: *International Conference on Learning Representations (ICLR)*. 2021.

[22] Haichen Wang et al. “DP-Promise: Differentially Private Diffusion Probabilistic Models for Image Synthesis”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2022.

[23] Richard Zhang et al. “The Unreasonable Effectiveness of Deep Features as a Perceptual Metric”. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018.

[24] Yang Zhang et al. “Membership Inference Attacks Against Generative Models”. In: *Proceedings of the IEEE Symposium on Security and Privacy (S&P)*. IEEE, 2021, pp. 1210–1227. DOI: 10.1109/SP40001.2021.00074.