

面向去噪扩散隐式模型的模型提取方法

杨雨川, 李 男, 朱汝权, 钱雯珺, 牛子浚, 赵贝宁

(信息工程大学, 河南 郑州 450001)

摘要: 作为扩散模型的典型代表, 去噪扩散隐式模型(DDIM)凭借确定性采样机制实现了高效图像生成, 但其可逆的生成路径也引发了隐私泄露风险, 易受模型提取攻击等恶意行为的威胁。提出一种基于确定性可逆路径的 DDIM 模型提取方法。该方法利用 DDIM 生成过程的确定性与可逆性, 通过有限次查询, 重构模型推理过程中的中间隐含状态及其对应的预测噪声, 进而构建代理数据集, 最终实现对 DDIM 模型的高保真功能复制。实验结果表明, 该方法仅需少量查询即可训练出高保真代理模型且保持良好的生成图像多样性, 揭示 DDIM 架构潜在的安全脆弱性。

关键词: 模型提取攻击; 实用灰盒攻击; 隐私安全; 去噪扩散隐式模型

中图分类号: TP309

文献标识码: A

文章编号: 1671-0673(2026)02-0231-08

Model Extraction Method for Denoising Diffusion Implicit Models

YANG Yuchuan, LI Nan, ZHU Ruquan, QIAN Wenjun, NIU Zijun, ZHAO Beining

(Information Engineering University, Zhengzhou 450001, China)

Abstract: As a representative of diffusion models, denoising diffusion implicit models (DDIM) enable efficient image generation through a deterministic sampling mechanism. However, the invertible nature of their generative trajectory introduces privacy risks, rendering them vulnerable to malicious model extraction attacks. A DDIM model extraction method based on the deterministic and invertible sampling path is proposed in this paper. By exploiting the determinism and reversibility inherent in the DDIM generation process, intermediate latent states and their corresponding predicted noise are reconstructed from a limited number of queries. A surrogate dataset is then constructed using these reconstructed pairs, facilitating high-fidelity functional replication of the target DDIM. Experimental results demonstrate that the extracted surrogate model, trained with only a small number of queries, achieves high generation fidelity while preserving diversity, thereby revealing a critical security vulnerability in the DDIM architecture.

Key words: model extraction attacks; practical gray-box attacks; privacy and security; denoising diffusion implicit models

近年来,生成式人工智能在图像、音频、文本等生成任务中取得了突破性进展。其中,扩散模型(Diffusion Models)因其较好的生成质量与训练稳定性,迅速成为生成模型领域的主流范式。

其中,去噪扩散概率模型(Denoising Diffusion

Probabilistic Models, DDPM)^[1]通过马尔可夫链结构实现数据生成,在图像合成任务中因展现出优于 GAN(Generative Adversarial Network)的稳定性和细节表现力而被广泛关注。但 DDPM 的生成过程依赖大量迭代采样,每一步均为随机过程,推理效

收稿日期: 2025-11-12; 修回日期: 2026-01-26

基金项目: 国家自然科学基金(62472447)

作者简介: 杨雨川(2001—),男,硕士生,主要研究方向为人工智能安全。E-mail: xsgcdxyyc@163.com

通信作者: 李 男(1977—),男,副教授,博士,主要研究方向为高性能计算、大数据挖掘、大数据安全。E-mail: linan_happy@126.com

率低下。为解决该问题,去噪扩散隐式模型^[2](Denoising Diffusion Implicit Models, DDIM)被提出,通过引入非马尔可夫生成过程与确定性生成路径,实现了生成过程的效率优化。

生成模型的优势使其在开放环境中被大规模部署,其知识产权与隐私安全风险日益受到关注。其中,模型提取攻击(Model Extraction Attack, MEA)通过向目标模型发送有限查询并利用其响应,构建功能高度近似的代理模型,不仅能绕过授权机制非法复现模型能力,还可为对抗攻击、成员推理等高阶威胁提供跳板^[3-5]。

当前普遍的模型提取方法为构建输入—输出对训练代理模型。然而将这种方法直接应用于扩散模型时,表现不佳,代理模型往往难以复现目标模型生成图像的语义细节与分布特性,生成结果模糊或失真严重。究其原因,扩散模型并非一步完成从噪声到数据的映射,而是通过多步迭代去噪,每一步仅需预测当前时间步的噪声。这一过程将复杂的生成任务分解为一系列局部、模型易学习的子问题。而上述提取方法忽略了反向轨迹中的中间状态及其对应的噪声,代理模型无法利用目标模型实际依赖的中间监督信号,导致其难以有效逼近目标模型的生成行为,最终生成结果失真或语义不一致。

因此,要实现扩散模型的有效提取,关键在于还原其完整的加噪—去噪路径。为此,本文聚焦于探究 DDIM 受到模型提取攻击的可能性,研究其隐私泄露风险,提出一种实用灰盒场景下的模型提取攻击方法。该方法提取的代理模型在生成质量、语义一致性与图像相似度等指标上均表现出良好的性能,揭示了 DDIM 在高效性与安全性之间的内在冲突。本文主要工作有如下 3 点。

1) 揭示了 DDIM 模型的信息泄露风险。通过研究非马尔可夫生成机制与中间状态可重构性,分析了对其进行模型提取的可行性,揭示了其高效推理模式下隐含的模型信息泄露风险。

2) 设计并实现了实用灰盒场景下的基于确定性可逆路径的 DDIM 模型提取算法 RDPE-DDIM (Reversible Deterministic Path Extraction for DDIM), 该算法利用 DDIM 生成过程的确定性与可逆性,重构模型推理过程中的中间隐含状态及其对应的预测噪声,进而构建代理数据集并完成代理模型训练,实现了对 DDIM 模型的高保真功能复制。

3) 验证了 RDPE-DDIM 算法在 DDIM-CIFAR-

10-32 等 5 个扩散模型上的有效性和泛化能力。实验表明,传统模型提取方法在扩散模型提取任务中无法实现有效提取,而本文方法仅需少量查询即可训练出高保真代理模型且保持良好的生成图像多样性。此外,实验验证了该算法在推理步数、调度方式及参数等不同条件变换下的实验效果。

1 背景知识

1.1 模型提取攻击

模型提取攻击指攻击者通过反复查询目标模型获取输入—输出响应,进而推断模型精确参数或重建一个功能上与原始模型高度相似的替代模型。Tramèr 等^[6]首次提出,通过方程求解攻击从黑盒 API 中恢复逻辑回归和朴素贝叶斯模型的权重参数,随后进一步利用决策树寻径攻击和梯度符号分析实现了对模型内部结构的部分还原。

生成模型提取攻击方面,Szyller 等^[7]以图像转换模型为目标,在通过查询获得的图像对上训练代理模型,窃取了用于神经风格迁移和超分辨率任务的生成对抗网络功能。Hu 等^[8]针对高保真度和高精度提取两种场景,在目标生成对抗网络生成的图像上训练了替代生成对抗网络。Liu^[9]通过在 PixelCNN、VAE、GAN 等 3 类代表性生成模型上构建跨架构攻击实验,揭示了生成模型在黑盒查询下的可提取性。Chent 等^[10]对通过查询受害模型获得的文本—图像对进行语义相似度筛选,构建高质量的替代训练集,实现了 81% 的保真度。Sha 等^[11]提出了 Cont-Steal 攻击,通过对比学习原理训练替代编码器,成功实现模型提取且有效规避多数防御机制。然而,同样在图像生成领域广泛应用的扩散模型^[12-14],尚缺乏模型提取的相关研究。

1.2 去噪扩散隐式模型

DDIM 的核心思想是使用非马尔可夫链去除推理图像状态间的依赖性,生成一条完全确定性的可跳步的生成路径,在保持高质量图像输出的同时显著提升推理效率。

其前向过程是一个固定的马尔可夫链,通过逐步向真实数据 x_0 添加高斯噪声,在 T 步后将其转化为纯噪声 x_T 。其加噪过程如式(1)所示:

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon_t \quad (1)$$

式中: x_t 为第 t 步(当前步)生成的新图像; x_0 为初始清晰图像; $t = 1, 2, \dots, T$; $\bar{\alpha}_t$ 是从第 1 步到第 t 步的累积保留系数,计算方式由具体噪声调度方式确定; $\epsilon_t \sim \mathcal{N}(0, I)$ 是标准高斯噪声, I 为单位矩阵。

1.2.1 DDIM 确定性生成机制

DDIM 的核心创新是构建了一条确定性生成路径,即给定初始噪声 x_T ,整个去噪过程可通过一系列可逆的确定性变换逐步还原出清晰图像 x_0 ,而无需引入额外的随机噪声。其广义反向采样公式记为

$$x_{t-1} = \sqrt{\bar{\alpha}_{t-1}} \cdot \left(\frac{x_t - \sqrt{1 - \bar{\alpha}_t} \cdot \epsilon_\theta(x_t, t)}{\sqrt{\bar{\alpha}_t}} \right) + \sqrt{1 - \bar{\alpha}_{t-1} - \sigma_t^2} \epsilon_\theta(x_t, t) + \sigma_t z \quad (2)$$

式中: $\bar{\alpha}_t$ 表示累积噪声系数; $\epsilon_\theta(x_t, t)$ 为神经网络 ϵ_θ 预测的噪声; σ_t 为反向过程中的噪声标准差,用于在生成时引入随机性,决定生成过程的随机程度; $z \sim \mathcal{N}(0, I)$ 。

设 β_t 为中间过程 t 时间步对应的噪声方差, $\bar{\alpha}_t$ 和 β_t 具体的计算方式由模型采用的调度方式决定,目前主流的调度方式为余弦调度和线性调度,余弦调度及线性调度相关参数计算方式分别如式(3)、式(4)所示:

$$\bar{\alpha}_t = \frac{\cos^2((t/T + s) \times \pi/2)}{\cos^2(s \times \pi/2)} \quad (3)$$

$$\begin{cases} \beta_t = \beta_1 + \frac{t-1}{T-1} (\beta_T - \beta_1); \\ \bar{\alpha}_t = \prod_{k=1}^t (1 - \beta_k). \end{cases} \quad (4)$$

式中: t 为当前所在的时间步(去噪步数); β_1 和 β_T 分别为起始和结束的噪声方差。

当 $\sigma_t = \sqrt{(1 - \bar{\alpha}_{t-1})/(1 - \bar{\alpha}_t)} \times \sqrt{1 - \bar{\alpha}_t/\alpha_{t-1}}$ 对任意 t 都成立时,该过程转变为 DDPM 的反向过程,而噪声方差 $\sigma_t = 0$ 时,上述过程退化为完全确定性路径,即

$$x_{t-1} = \sqrt{\bar{\alpha}_{t-1}} \cdot \left(\frac{x_t - \sqrt{1 - \bar{\alpha}_t} \cdot \epsilon_\theta(x_t, t)}{\sqrt{\bar{\alpha}_t}} \right) + \sqrt{1 - \bar{\alpha}_{t-1}} \cdot \epsilon_\theta(x_t, t) \quad (5)$$

式(5)去除了式(2)中的最后一项,此时 $\epsilon_\theta(x_t, t)$ 是一个确定性函数(固定权重的神经网络),表明反向去噪过程已不再包含任何未知随机项。每一步的 x_{t-1} 可由当前图像状态 x_t 和神经网络输出 $\epsilon_\theta(x_t, t)$ 及累积噪声系数 $\bar{\alpha}_t, \bar{\alpha}_{t-1}$ 计算得出。整个反向轨迹 $x_T \rightarrow x_{T-1} \rightarrow \dots \rightarrow x_0$ 中的每一个状态都可以依据式(5)计算得出唯一值。

综上,路径还原的条件已具备,为本文提出的模型提取攻击提供了理论基础,攻击者能够通过查询获取的 (x_T, x_0) 对恢复唯一确定的去噪路径。

1.2.2 非马尔可夫跳步采样

基于式(5), DDIM 在反向过程中支持跳步采样。具体而言,设 $\tau = (t_1, t_2, \dots, t_s)$ 为从原始扩散时间序列 $[1, \dots, T]$ 中选取的递增子序列,并满足 $1 \leq t_1 < t_2 < \dots < t_s = T$, 其中, $t_s = T$ 。对于子序列 τ 中相邻的两个时间步 t_{i-1} 和 t_i , 其条件边缘分布可表示为

$$q_\sigma(x_{t_{i-1}} | x_{t_i}, x_0) = \mathcal{N} \left(\sqrt{\bar{\alpha}_{t_{i-1}}} x_0 + \sqrt{1 - \bar{\alpha}_{t_{i-1}} - \sigma_{t_i}^2} \cdot \frac{x_{t_i} - \sqrt{\bar{\alpha}_{t_i}} x_0}{\sqrt{1 - \bar{\alpha}_{t_i}}}, \sigma_{t_i}^2, I \right) \quad (6)$$

令 $\sigma_{t_i} = 0$ 时,该分布退化为确定性映射,其采样结果可直接表示为

$$x_{t_{i-1}} = \sqrt{\bar{\alpha}_{t_{i-1}}} x_0 + \sqrt{1 - \bar{\alpha}_{t_{i-1}}} \cdot \frac{x_{t_i} - \sqrt{\bar{\alpha}_{t_i}} x_0}{\sqrt{1 - \bar{\alpha}_{t_i}}} \quad (7)$$

由式(7)可知,在 DDIM 的反向推理过程中,任意中间状态 $x_{t_{i-1}}$ 均可由更晚时刻的状态 x_{t_i} 与原始数据 x_0 唯一确定,而无需逐步遍历所有中间时间步。因此,通过预先选取一个稀疏的时间步子序列 τ , DDIM 可以在远少于原始 T 步的迭代次数下完成生成,并保持生成过程的确定性。这一性质不仅显著提高了推理效率,也为后续基于生成路径重构的模型提取方法提供了理论基础。

2 本文算法的提出及设计与实现

2.1 传统模型提取方法的局限性

当前主流的模型提取攻击方法普遍通过向目标模型提交大量输入,获取对应的输出响应,进而构建代理数据集,训练功能近似的代理模型。

然而,当此类方法直接迁移至扩散模型时,其提取性能显著下降(见 3.4.1 节)。其原因在于扩散模型训练的核心思想是将复杂的数据分布建模分解为一系列细粒度的噪声扰动与恢复步骤,而传统提取方法通过查询获取大量端到端的输入—输出对 (x_T, x_0) , 组成形式为 $N \times (x_T, T, \epsilon_T)$ 的代理数据集,训练代理模型将完全噪声一步转变为清晰图像的能力,忽略了反向轨迹中的中间状态及其对应的噪声。代理模型无法利用目标模型实际依赖的中间监督信号,导致其难以有效逼近目标模型的生成行为。传统模型提取方法难以有效迁移至扩散模型,要实现扩散模型的提取攻击,关键在于还原完整中间状态及对应噪声的加噪路径。

2.2 DDIM 加噪路径还原

攻击者视角下, DDIM 的确定性生成机制确保了从初始噪声 x_T 到最终图像 x_0 的映射是唯一且可复现的。生成路径的确定性对应了扩散路径的确定性, 因此, 构建代理模型训练数据集的关键就是重构扩散路径。还原整条生成路径即中间状态序列及对应步数的预测噪声, 从而构造代理数据集是实现模型提取的关键。通过查询已知 (x_T, x_0) , 同时根据 DDIM 跳步采样机制, 即式(7), 能够还原任意时间步的中间状态。具体地, 固定式(7)中的 $\tau_i = T$, 令 $\tau_{i-1} = t$, 并利用 DDPM 前向过程的定义可得:

$$\epsilon_{\tau_i} = \epsilon_T = \frac{x_T - \sqrt{\bar{\alpha}_T} x_0}{\sqrt{1 - \bar{\alpha}_T}} \quad (8)$$

同时, 由式(1)可反解出:

$$x_0 = \frac{x_T - \sqrt{1 - \bar{\alpha}_T} \epsilon_T}{\sqrt{\bar{\alpha}_T}} \quad (9)$$

将式(8)、式(9)带入式(7), 经整理可得任意时间步 t 的状态表达式:

$$x_t = \sqrt{\frac{\bar{\alpha}_t}{\bar{\alpha}_T}} \left(x_T - \sqrt{1 - \bar{\alpha}_T} \cdot \epsilon_T \right) + \sqrt{1 - \bar{\alpha}_t} \cdot \epsilon_T \quad (10)$$

式中: ϵ_T 为从 x_0 到 x_T 的全局噪声, 可直接由单次查询结果计算得出; 累积系数 $\bar{\alpha}_t$ 预设的噪声调度策略(如线性、余弦等)及参数 t 唯一确定。

因此, 仅通过一次合法查询得到 (x_T, x_0) 即可精确重构整条去噪轨迹上的任意时间步的图像状态及各时间步对应的预测噪声 ϵ_t , 从而构建 T 个包含 (x_t, t, ϵ_t) 配对的高质量监督数据集, 用以训练代理模型。基于此, 本文提出了实用灰盒下的基于确定性可逆路径的 DDIM 模型提取算法 RDPE-DDIM。

2.3 RDPE-DDIM 算法的设计与实现

针对 DDIM 实用灰盒场景下的模型提取攻击方法研究, 本文将实用灰盒场景设置为: 目标模型的网络结构及参数设置为完全不可见, 推理步数 T 、调度方式及对应调度参数设置为已知, 用户可以调用该模型接口, 输入噪声后仅可得到推理 T 步数后生成的图像, 无法获取中间推理步骤图像状态。本文方法流程如图1所示。

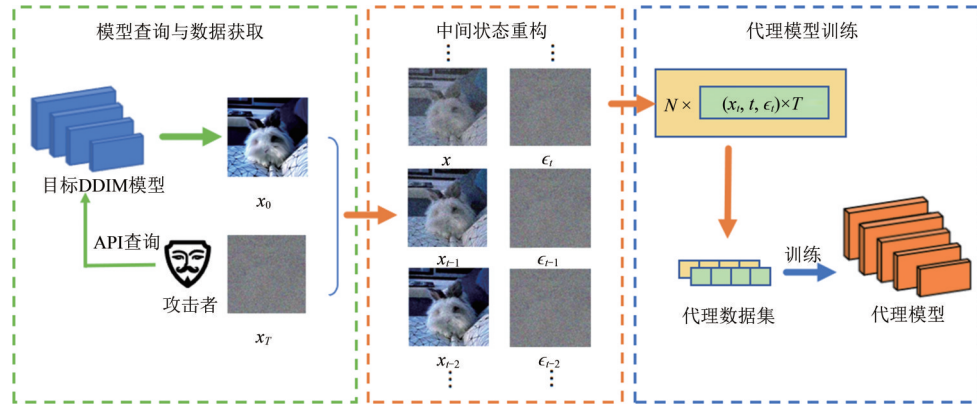


图1 DDIM 模型提取方法流程示意图

整个方法流程可分为以下3个关键阶段。

1) 模型查询与数据获取。生成随机初始噪声, 向目标 DDIM 模型发起 N 次查询, 噪声图像 x_T 作为输入, 并调用目标模型的生成接口。由于 DDIM 的生成过程是确定性的, 模型会输出一个唯一的、高质量的最终生成图像 x_0 , (x_T, x_0) 对构成攻击的起点。

2) 中间状态重构。攻击者利用 DDIM 噪声调度参数和获取到的 (x_T, x_0) 对, 通过数学推导逆向计算出生成路径上的所有中间状态。然后, 根据式(10)一次查询即可推导出完整的去噪轨迹 $\{x_T, x_{T-1}, \dots, x_1, x_0\}$ 。

3) 代理模型训练。利用上一步重构出的大量中间状态 (x_t, t) 及其对应的预测噪声 $\epsilon_\theta(x_t, t)$, 攻击

者构建一个高质量的代理训练数据集。攻击者使用这个数据集训练一个结构相似的代理神经网络, 使其学习并逼近目标模型的噪声预测函数 ϵ_θ 。

RDPE-DDIM 算法具体如下所示。

算法1: RDPE-DDIM 算法

输入: 目标 DDIM 模型 M , 推理步数 T , 查询次数 N , 噪声调度参数 s (以余弦调度为例)

输出: 训练完成的代理模型 f_θ

1. 初始化代理数据集 $D \leftarrow \emptyset$
2. for $i = 1$ to N do:
3. 设置随机种子, 生成初始噪声图像 $x_T \sim \mathcal{N}(0, I)$
4. 调用目标模型 M , 输入 x_T , 得到输出图像 $x_0 = M(x_T)$
5. 计算全局噪声估计值 ϵ_T
6. for $t = T$ to 1 do:

7. 计算噪声累积系数 $\bar{\alpha}_i$
8. 计算中间状态 x_i
9. 计算 x_i 对应噪声 ϵ_i
10. 将样本 (x_i, ϵ_i) 添加至数据集 D
11. 使用数据集 D 训练代理模型 f_θ
12. 返回训练完成的代理模型 f_θ

3 实验结果与分析

3.1 实验环境

实验在配备 Intel Core i5-9400F CPU (2.90 GHz)、32 GB 内存及 NVIDIA RTX 2080 Ti GPU 的台式机上完成,以确保计算资源的充足且一致。实验参数设置详见表1。

表1 实验参数设置

参数	数值
查询次数	10
加噪步数	1 000
调度方式	余弦调度
平滑因子 s	0.008
总样本数	10 000

3.2 数据集与目标模型

本文选取 DDIM-butterflies-128、DDIM-LSUN-bedroom、DDIM-CIFAR-100-32、DDIM-CIFAR-10-32、DDIM-MNIST-32 等5个模型作为目标模型。

代理模型采用 2D U-Net 架构,包含编码器—解码器结构与跳跃连接。网络使用组归一化(Group Normalization)和 Swish 激活函数,输出层通道数与目标模型输入一致。基础特征通道数根据输入分辨率调整:低分辨率(32×32)代理模型通道数设为64,高分辨率(128×128)代理模型通道数设为128,且加噪步数、调度方式与目标模型相同,调度方式设置为余弦调度,平滑因子 s 为固定值0.008。

3.3 评估指标

评估指标分为生成质量评估和功能一致性评估两方面。本文采用如下3个评估指标。

1) FID(Fréchet Inception Distance)值。FID 距离定义为 $\mathcal{N}(\mu_r, \Sigma_r)$ 和 $\mathcal{N}(\mu_g, \Sigma_g)$ 这两个多元高斯分布之间的 Fréchet 距离:

$$D_{\text{FID}} = \left\| \mu_r - \mu_g \right\|^2 + \text{Tr}(\Sigma_r + \Sigma_g - 2\sqrt{\Sigma_r \Sigma_g}) \quad (11)$$

式中: μ_r 、 μ_g 为特征均值; Σ_r 、 Σ_g 为特征协方差矩阵;Tr表示矩阵的迹。 D_{FID} 值越低,表明两组图像在深度特征空间具有越高的一致性。一般认为, $D_{\text{FID}} < 20$ 为优秀,生成图像的质量非常高,通常与真

实图像难以区分,细节丰富,多样性好; $20 \leq D_{\text{FID}} < 50$ 为良好,此时生成图像质量较好,大部分图像清晰,但可能在细节、真实感或多样性上略逊于真实数据; $50 \leq D_{\text{FID}} < 100$ 为一般,生成图像存在明显缺陷,如模糊、失真或模式崩溃,但整体结构和内容尚可辨认。本文实验均使用目标模型和代理模型生成的图像计算 FID 距离,两组图像数量均为 500。对于多类别模型,FID 距离基于分类器预测的类别标签计算;对于单类别模型,由于无跨类别多样性,FID 距离改用 Inception-v3 特征在第一主成分上的投影值计算。

2) K-S 距离。K-S 距离用于衡量两个一维概率分布之间的最大差异。在本实验中将目标模型和代理模型生成的所有图像的像素值分别视为两个样本集,计算其经验累积分布函数 $F_r(x)$ 和 $F_g(x)$ 。K-S 距离定义为这两个联合经验累积分布函数之间上确界(最大垂直距离):

$$D_{\text{K-S}} = \sup_x \left| F_r(x) - F_g(x) \right| \quad (12)$$

较低的 K-S 距离表明两个模型的输出图像在全局像素强度分布上高度相似。为简化计算,分别对3个颜色通道的像素值进行归一化后合并,计算其联合经验累积分布函数。一般认为, $D_{\text{K-S}} < 0.1$,高度相似,两个分布非常接近,表明生成图像在全局统计特性(如亮度、对比度)上与真实图像高度一致; $0.1 \leq D_{\text{K-S}} < 0.2$,分布有一定差异,但整体统计规律相似; $0.2 \leq D_{\text{K-S}} < 0.3$,分布存在较明显差异,生成图像可能在色调、明暗等方面与真实图像有偏差。本文实验均使用随机噪声输入目标模型和代理模型生成的对应图像的像素值计算 K-S 距离,两组图像数量均为 500。对于多类别模型,K-S 距离基于分类器预测的类别标签计算;对于单类别模型,由于无跨类别多样性,K-S 距离改用 Inception-v3 特征在第一主成分上的投影值计算。

3) K-L 散度。K-L 散度用于衡量两个概率分布之间的信息差异程度。K-L 散度定义为

$$D_{\text{K-L}}(P \parallel Q) = \sum_{c=1}^C P(c) \ln \frac{P(c)}{Q(c)} \quad (13)$$

式中: $P(c)$ 表示真实数据集或目标模型生成样本的类别分布; $Q(c)$ 表示代理模型生成样本的类别分布; C 为总类别数; $\log$ 通常取自然对数。

较低的 K-L 散度表明代理模型生成的图像类别分布与目标模型高度一致。一般认为, $D_{\text{K-L}} < 0.1$,表明两个分布高度相似; $0.1 \leq D_{\text{K-L}} < 0.2$,分布存在一定差异但整体规律接近; $0.2 \leq D_{\text{K-L}} < 0.3$,分布存

在较明显差异,代理模型可能在某些类别上出现偏移或缺失。本文实验均使用随机噪声输入目标模型和代理模型生成的对应图像,分别统计其类别分布后计算 K-L 散度,两组图像数量均为 500。

3.4 实验结果与分析

3.4.1 RDPE-DDIM 算法有效性验证

本组实验目的是验证 RDPE-DDIM 算法的有效性。目标模型选取 DDIM-CIFAR-10-32,代理模型超参数设置为学习率 0.001,优化器 Adam,训练轮次 30。使用本文算法对目标模型进行模型提取,实验设置见 3.1 节,实验效果如图 2 所示。对照组采用标准黑盒查询策略对目标模型进行攻击:独立采样标准高斯分布中的 50 000 个初始噪声,然后输入目标模型,执行完整的反向去噪过程以获得对应的清晰图像 x_0 。由此构建一个包含 50 000 对 (x_T, T, ϵ_T) 数据的代理数据集进行代理模型训练,传统模型实验效果如图 3 所示。

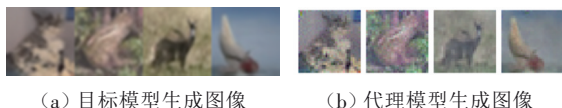


图 2 RDPE-DDIM 算法提取实验效果

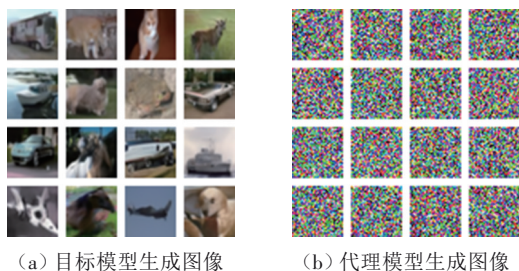


图 3 传统模型提取方法实验结果

如图 2、图 3 所示,本文算法代理模型的生成图片与目标模型的生成图片视觉效果十分接近,而传统模型提取方法代理模型生成的图像均为彩色斑点图像,无法识别生成图片的类别。

本组实验中,本文算法 FID 距离为 12.5,表明图像生成质量与目标模型十分接近;K-S 距离为 0.098,表明代理模型输出图像与目标模型高度一致。对比传统模型提取方法实验结果(FID 距离为 219.4、K-S 距离 0.61),本文算法的实验效果显著提高,且查询次数由 50 000 降至 10,查询成本大幅下降,验证了 RDPE-DDIM 算法的有效性。

3.4.2 泛化性验证

本组实验考察本文算法的泛化性。通过增加在 DDIM-butterflies-128 等多个不同目标模型上的实验,旨在验证所提算法在其他扩散模型上的提

取能力与适用范围。实验相关参数设置与前文相同,实验结果如图 4~图 7 所示,结果对比见表 2。



图 4 DDIM-butterflies-128 模型提取实验结果

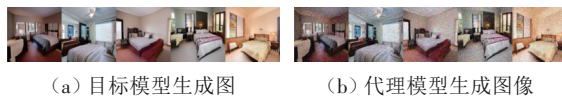


图 5 DDIM-LSUN-bedroom 模型提取实验结果



图 6 DDIM-CIFAR-100-32 模型提取实验效果



图 7 DDIM-MNIST-10-32 模型提取实验效果

表 2 实验结果对比

模型	FID 距离	K-S 距离
DDIM-butterflies-128	24.2	0.087
DDIM-LSUN-bedroom	26.8	0.131
DDIM-CIFAR-100-32	19.3	0.120
DDIM-MNIST-10-32	53.9	0.215

如表 2 所示,代理模型在 DDIM-Butterflies-128、DDIM-LSUN-bedroom 和 DDIM-CIFAR-100-32 上均取得良好性能:FID 值均位于区间[0,30]内,K-S 距离均接近经验阈值 0.1,表明生成质量接近目标模型,输出分布高度相似。

相比之下,在 DDIM-MNIST-10-32 任务上,FID 距离为 53.9>50,K-S 距离达 0.215>0.2,表明代理模型生成图像质量一般、与目标模型相似度有限。该组实验性能下降的原因为目标模型在真实数据集上进行训练,而代理模型则基于目标模型生成的图像进行训练。由于生成图像与真实图像之间存在一定的分布偏差,且 MNIST 图像为结构简单的黑白图像,对生成误差尤为敏感,导致代理模型最终生成的图像质量较差。

综上可得,RDPE-DDIM 能有效适应不同数据分布和任务场景,尤其在处理包含丰富纹理和色彩信息的复杂图像时优势明显,而在低维、强对比度、低噪声敏感度的数据集上,其性能会受到限制。

3.4.3 代理模型生成多样性的效果

本组实验目的是对代理模型与目标模型在图像生成类别多样性方面进行对比。分别使用两组不同的随机数种子生成随机噪声作为输入,令目标模型和代理模型独立生成500张图像,生成图像如图8所示(20张示例)。



(a) 目标模型生成图像 (b) 代理模型生成图像

图8 多样性验证实验结果(示例)

实验使用在CIFAR-10训练集上训练的ResNet-18分类器(测试准确率95.2%)对全部1 000张生成图像进行类别预测。对每张图像取置信度最高的类别作为标签,并统计各模型生成图像在10个类别上的频次分布。

表3为目标模型和代理模型生成图像的类别分布,对应的K-L散度 $D_{K-L}(P_{\text{target}} \| P_{\text{proxy}}) = 0.087$,表明两者的生成分布一致性较高。综上,本文方法在实现高保真模型复制的同时,有效维护了目标模型的生成多样性。

表3 目标模型与代理模型生成图像类别分布

类别	目标模型类别数量	代理模型类别数量
airplane	58	55
automobile	45	48
bird	42	34
cat	55	63
deer	40	32
dog	60	65
frog	50	49
horse	48	46
ship	52	53
truck	50	55

3.4.4 不同查询次数对模型提取效果的影响

本组实验验证本文方法在不同查询次数下的性能表现。为确保与前文实验的可比性,实验在DDIM-CIFAR-10-32模型上进行。代理模型结构与训练超参数均与前文一致,仅改变查询次数 N ,变化范围如下: $\{1, 2, 5, 10, 20\}$ 。每组实验重复3次,取FID距离与K-S距离的平均值作为最终结果。图9直观显示了FID距离、K-S值随查询次数变化的趋势。

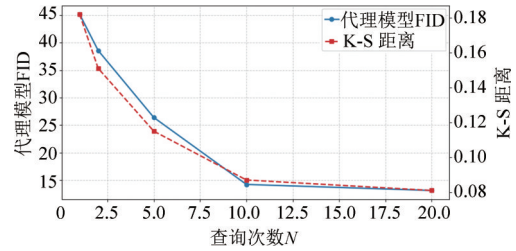


图9 查询次数对DDIM-CIFAR-10-32模型提取效果的影响

图9显示,模型提取效果随查询次数增加而提升。也反映查询次数对性能增益的影响,查询次数从10次增加到20次时,查询成本翻倍,而性能增益十分有限,模型提取效果在 $N = 10$ 后趋于稳定,且随查询次数增加并不能显著提升提取效果。

3.4.5 推理步数、噪声调度参数对实验效果的影响

本组实验验证本文所提方法在不同推理步数 T 或噪声调度参数时的性能表现,以评估其在实用灰盒设定下的可行性。实验在DDIM-CIFAR-10-32模型上进行,在模型提取攻击阶段引入不同类型的参数偏差,包括推理步数,调度类型以及调度关键参数,实验结果如图10、表4所示。表中, β_1 和 β_T 分别为起始和结束的噪声方差。

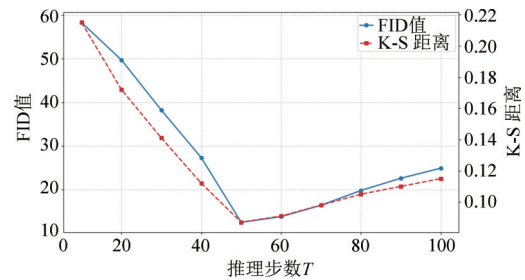


图10 不同推理步数下实验结果

表4 不同调度方式及参数下的模型提取性能

调度类型	关键参数设置	代理模型 FID距离	K-S距离
余弦调度	$s=0.002$	62.1	0.241
	$s=0.004$	51.3	0.192
	$s=0.006$	32.7	0.128
	$s=0.008$	12.5	0.087
	$s=0.010$	17.2	0.102
	$s=0.012$	29.8	0.156
线性调度	$\beta_1=0.0001, \beta_T=0.02$	78.4	0.310
	$\beta_1=0.001, \beta_T=0.02$	82.6	0.325

实验结果表明, RDPE-DDIM的提取效果对目标模型的噪声调度参数高度敏感。如图10所示,当代理模型使用的推理步数 T 与目标模型一致($T=50$)时,取得最优性能($D_{\text{FID}}=12.5, D_{\text{K-S}}=0.087$),而 T 偏离真实值后,性能显著下降。除推理步数外,噪

声调度参数的偏差同样会显著影响模型提取效果。如表4所示,在余弦调度下,当平滑因子 s 与目标模型真实参数一致,即 $s = 0.008$ 时,代理模型性能最优;当 s 偏离真实值时, D_{FID} 与 $D_{\text{K-S}}$ 距离均明显上升,当 $s = 0.006$ 小于真实值时, D_{FID} 与 $D_{\text{K-S}}$ 距离分别增大至32.7和0.128;而当 $s = 0.002$ 时, D_{FID} 与 $D_{\text{K-S}}$ 进一步增大至62.1和0.241;当 $s = 0.010$ 大于真实值时, D_{FID} 与 $D_{\text{K-S}}$ 增大至17.2和0.102,若 s 进一步增大,则性能退化更加明显, D_{FID} 与 $D_{\text{K-S}}$ 进一步增大至29.8和0.156。该结果表明,平滑因子的微小偏差尤其在设置的 s 小于真实值时即可对代理模型的生成质量和分布一致性造成明显影响。

4 结束语

本文提出了一种面向DDIM架构的实用灰盒模型提取方法,其核心思想是利用DDIM生成过程的确定性与可逆性,通过少量API查询,反向推导出完整的去噪轨迹。该方法不仅在butterflies-128和LSUN-bedroom等复杂数据集上实现了有效提取,同时在CIFAR-10、CIFAR-100等典型生成任务上实现了高保真的功能复制,揭示了DDIM生成模型面临的知识产权泄露风险。

参考文献:

- [1] HO J, JAIN A, ABBEEL P. Denoising diffusion probabilistic models[DB/OL]. (2020-06-19)[2026-01-26]. <https://arxiv.org/abs/2006.11239>.
- [2] SONG J M, MENG C L, ERMON S. Denoising diffusion implicit models[DB/OL]. (2020-10-06)[2026-01-26]. <https://arxiv.org/abs/2010.02502>.
- [3] 陈帅,李威.基于案例复盘的舆情传播控制模型研究:以新浪微博为例[J].信息工程大学学报,2020,21(4):495-500.
- [4] 刘文钊,杨奎武,陈越,等.元微调对抗训练:面向多类型对抗攻击的对抗训练改进方法[J].信息工程大学学报,2025,26(4):470-477.
- [5] CINÀ A E, GROSSE K, DEMONTIS A, et al. Wild patterns reloaded: a survey of machine learning security against training data poisoning[J]. ACM Computing Surveys, 2023,55(13s):No.294.
- [6] TRAMÈR F, ZHANG F, JUELS A, et al. Stealing machine learning models via prediction APIs[C]// Proceedings of the 25th USENIX security symposium. Berkeley, USA: USENIX Association, 2016:601-618.
- [7] SZYLLER S, DUDDU V, GRÖNDAHL T, et al. Good artists copy, great artists steal: model extraction attacks against image translation models[DB/OL]. (2021-04-26)[2026-01-26]. <https://arxiv.org/abs/2104.12623>.
- [8] HU H L, PANG J. Stealing machine learning models: attacks and countermeasures for generative adversarial networks[C]// Proceedings of the 37th Annual Computer Security Applications Conference. New York, USA: ACM, 2021:1-16.
- [9] LIU S Y. Model extraction attack and defense on deep generative models[J]. Journal of Physics: Conference Series, 2022,2189(1):No.012024.
- [10] CHENT Y, GUO W H, GE P Y, et al. Model extraction attacks on text-to-image generative adversarial networks[C]// Proceedings of the 2024 IEEE Cyber Science and Technology Congress. Piscataway, USA: IEEE, 2024:272-279.
- [11] SHA Z Y, HE X L, YU N, et al. Can't steal? cont-steal! contrastive stealing attacks against image encoders[C]// Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE, 2023:16373-16383.
- [12] NICHOL A Q, DHARIWAL P. Improved denoising diffusion probabilistic models[C]// Proceedings of Machine Learning Research. Cambridge, USA: PMLR, 2021:8162-8171.
- [13] SONG Y, ERMON S. Generative modeling by estimating gradients of the data distribution[C]// Proceedings of the 33rd International Conference on Neural Information Processing Systems. New York, USA: Curran Associates Inc., 2019:11918-11930.
- [14] VAHDAT A, KREIS K, KAUTZ J. Score-based generative modeling in latent space[DB/OL]. (2021-06-10)[2026-01-26]. <https://doi.org/10.48550/arxiv.2106.05931>.

(编辑:邵 华)