arXiv:2602.01645v1 [cs.SD] 2 Feb 2026

基于生成流形扰动的音乐扩散模型成员推断攻击

Yuxuan Liu、*Peihong Zhang*、*Rui Sang*、*Zhixin Li*、*Yizhou Tan*、*Yiqiang Cai*、*Shengchen Li*

中国苏州，西交利物浦大学

摘要

成员推断攻击（MIAs）用于检测特定音频片段是否曾用于训练模型，因此是审计生成音乐模型版权合规性的关键工具。然而，基于损失的信号（例如重建误差）在实践中与人类感知的关联较弱，导致其在取证所要求的低假阳性率（FPR）下区分能力不佳。我们提出潜变量稳定性对抗探针（LSA-Probe），这是一种白盒方法，用于测量反向扩散过程的一种几何属性：在中间扩散状态下，跨越固定感知退化阈值所需的最小时间归一化扰动预算。我们发现，训练成员位于更稳定的区域，因此表现出显著更高的退化代价。在计算量匹配且阈值固定为（τ=第 95 百分位数）的条件下，与最佳基线相比，LSA-Probe 在 DiffWave 和 MusicLDM 上将 FPR 为 1% 时的 TPR 提高了 3–8 个百分点。这些结果表明，局部生成稳定性可为音频扩散模型提供可靠的成员身份信号。

索引词— 成员身份推理、扩散模型、对抗鲁棒性。

1. 引言

扩散模型正在变革音乐生成，合成出具有高度感知真实感的高保真音频 [1, 2]。然而，这些模型的卓越性能有赖于使用海量数据集进行训练，而这些数据集的来源和授权情况往往不明确 [3]。这种依赖使艺术家和权利持有人面临版权侵权和隐私侵犯的重大风险。因此，亟需可靠的审计工具，以确定特定音乐片段是否曾被用于训练给定模型。

成员推断攻击（MIA）旨在利用模型对其训练数据的记忆，以区分成员（训练集样本）与非成员（留出样本） [4, 5]。在计算机视觉领域，此类攻击基于一个核心假设：成员样本产生的重建损失低于非成员样本 [6]。然而，越来越多的近期研究表明，当这一基本假设应用于音频领域时，会受到严重挑战。例如，Rossi 等 [7] 证明，标准成员推断攻击在序列模型上表现不佳，因为它们未能充分建模数据内部复杂的相关性。Kong 等 [8] 报告称，此类端点式仅损失准则在音频扩散中会受到混淆且不稳定：它们对内容复杂度和音色因素敏感，并在不同噪声种子和时间步之间表现出高方差。

这些发现共同推动了成员推断攻击研究的范式转变：从依赖单一、静态的端点信号（如重建损失），转向利用信息更丰富的动态、基于过程的信号 [9, 10]。这一转变的动机来自与公认的“平坦极小值”原则的类比，即损失景观中更平滑的区域与更好的泛化和鲁棒性密切相关 [9]。Matsumoto 等人 [11] 表明，在加入相同的对抗扰动时，成员比非成员更具鲁棒性。

受此启发，我们提出了潜变量稳定性对抗探针（LSA-Probe），这是一种新颖的白盒框架，可直接探测生成过程的几何稳定性，从而推断成员身份。我们的核心假设是，模型会在其训练成员的局部邻域中学习到更平滑、更稳定的生成映射。我们通过量化对抗代价来落实这一假设：即对反向扩散过程的中间潜变量状态施加扰动时，为造成固定程度的感知退化所需的最小扰动预算。由于成员位于生成流形中更稳定的区域，要使其退化所需的对抗代价显著高于非成员，因此可为我们的攻击提供可靠信号。

所提出的方法在两类音乐扩散模型上进行了评估：(i) 波形去噪扩散概率模型（DDPM） [12], , 具体实现为 DiffWave [13]; 以及 (ii) 潜在扩散模型（LDM），具体实现为 MusicLDM [1]——评估所用数据集包括 MAESTRO v3 [14] 钢琴演奏语料库和自由音乐档案馆（FMA）大型子集 [15]。我们的实验表明，该方法能够稳健地识别成员身份，在取证至关重要的低假阳性率区间，其表现优于传统基线方法。本研究的主要贡献如下：

1. 据我们所知，这是首次对音乐扩散模型的成员推断攻击进行系统性研究。2. 我们提出了 LSA-Probe，这是一种白盒成员推断攻击，通过探测反向扩散轨迹上的时间归一化潜在稳定性，并根据达到固定感知退化所需的对抗代价为样本评分。3. 我们通过一阶分析将局部生成稳定性与成员身份联系起来，从而得到一种实用的评分方法，无需似然或影子模型。

2. 所提方法

我们研究一种白盒的开发者侧设置：对手可以读取模型参数、反向采样器，以及中间扩散状态下相对于输入的梯度。目标是仅使用经过审计的模型（不使用影子模型或训练标签），判断给定波形 x_0 是否属于成员。我们的双循环

项目演示可访问：https://github.com/kaslim/LSA-Probe

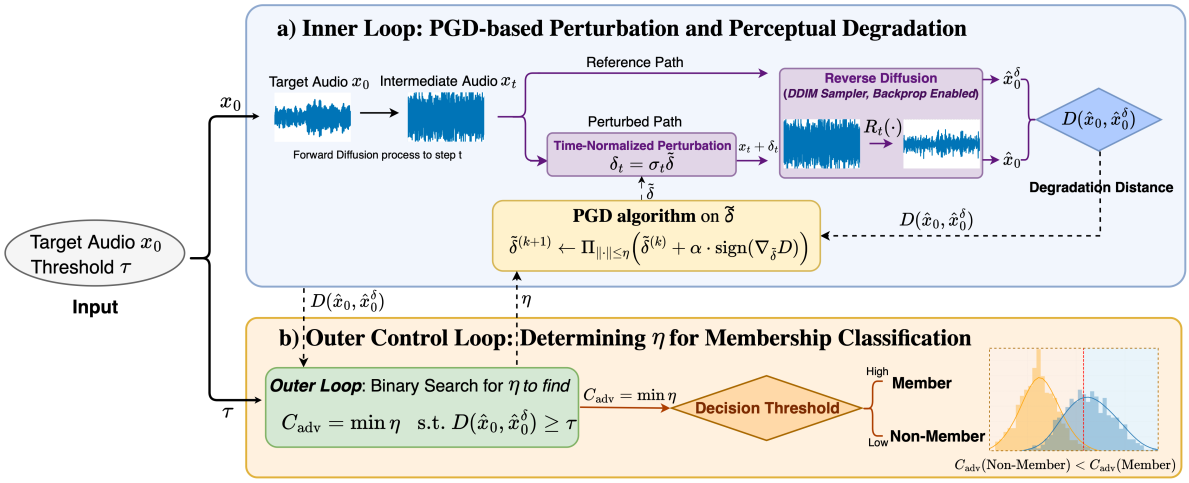


图 1：LSA-Probe 概览（时间步 t处的双循环流程）。(a) 外层控制循环。我们通过二分搜索预算 η，以找到满足 $C_{\text{adv}}(x_0; t, \tau) = \min\{\eta: D(\hat{x}_0, \hat{x}_0^\delta) \ge \tau\}$ 的对抗代价；随后将 C_{adv}用作决策分数（$\Rightarrow$越高，越可能是成员）；虚线表示用于 TPR@FPR 分析的决策阈值。(b) 内层循环（在 $\tilde{\delta}$上进行投影梯度下降）。给定目标波形 x_0，我们通过前向过程构造x_t，并注入时间归一化潜变量扰动 $\delta_t = \sigma_t\tilde{\delta}$，其$\sigma_t = \sqrt{1-\bar{\alpha}_t}$为 $\sigma_t = \sqrt{1-\bar{\alpha}_t}$。我们通过 ℓ_p球上的投影，对 $\tilde{\delta}$ 进行 K 步更新 $\|\delta\|_p \le \eta$（对 ℓ_∞使用符号步长，或对 ℓ_2使用归一化梯度），同时通过确定性逆向算子 $R_t(\cdot; \theta)$以及所选的可微距离 D（CDPAM、MR-STFT、对数梅尔均方误差或波形均方误差）进行反向传播。对于当前的 η，内层循环输出退化量 $D(\hat{x}_0, \hat{x}_0^\delta)$，外层循环据此调整 η。对于每个(x_0, t)，我们固定前向噪声 ϵ（通过随机种子设定），以隔离 δ_t的影响；为清晰起见，图中省略了重启次数和动量。

流程（外层通过二分搜索确定预算，内层对经过时间归一化的扰动执行投影梯度下降）如图 1 所示，并将在本节中多次提及。

2.1. 预备知识：前向过程/噪声参数化与反向算子

令 $x_0 \in \mathbb{R}^n$ 表示一个波形。在离散步骤 $t \in \{1, \dots, T\}$ 中，前向加噪遵循标准的 DDPM 参数化

$$x_t = \sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{1-\bar{\alpha}_t}\, \epsilon, \qquad \epsilon \sim \mathcal{N}(0, I), \tag{1}$$

，其调度表为 $\{\bar{\alpha}_t\}_{t=1}^T$。我们采用确定性的 DDIM 风格反向算子，且 $\sigma_t = 0$。给定噪声预测器 $\epsilon_\theta(\cdot, t)$，定义逐步去噪估计

$$\hat{x}_0(x_t, t) = \frac{x_t - \sqrt{1-\bar{\alpha}_t}\, \epsilon_\theta(x_t, t)}{\sqrt{\bar{\alpha}_t}}. \tag{2}$$

单个反向步骤通过以下方式将 $x_t \mapsto x_{t-1}$ 映射为

$$x_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\, \hat{x}_0(x_t, t) + \sqrt{1-\bar{\alpha}_{t-1}}\, \epsilon_\theta(x_t, t). \tag{3}$$

将式 (3) 从索引 t 到 1 依次复合，可得到可微映射 $R_t(\cdot; \theta)$：$x_t \mapsto \hat{x}_0 \in \mathbb{R}^n$：

$$\hat{x}_0 = R_t(x_t; \theta). \tag{4}$$

我们既可以用整数 t，也可以用比例 $t_{\text{ratio}} \in (0, 1]$，映射为 $t = \lfloor t_{\text{ratio}}(T-1) \rfloor + 1$）来表示时间步。

2.2. 时间归一化潜在扰动

不同 t 下的扰动无法直接比较，因为正向过程的噪声方差为 $1-\bar{\alpha}_t$。因此，我们在 x_t处注入时间归一化扰动：

$$\delta_t = \sigma_t\, \tilde{\delta}, \qquad \sigma_t = \sqrt{1-\bar{\alpha}_t}, \quad \|\tilde{\delta}\|_p \le \eta, \tag{5}$$

其中 $p \in \{2, \infty\}$，全局预算为 $\eta > 0$。辅助变量 $\tilde{\delta}$ 按样本独立同分布地从高斯分布方向中采样，并作为我们攻击中的优化变量。[1] 选择 $\delta_t = \sqrt{1-\bar{\alpha}_t}\, \tilde{\delta}$与正向过程的方差相匹配，从而使不同时间步的信噪比保持一致，并让预算具有可比性。对于每一对 (x_0, t)，我们通过样本专用种子固定正向过程的 ϵ，以确保配对评估能够分离出 δ_t的影响；在多次重复实验中，我们对 K 个随机种子取平均。从扰动潜变量开始的第二次反向过程得到

$$\hat{x}_0^\delta = R_t(x_t + \delta_t; \theta). \tag{6}$$

我们通过在波形上计算的可微距离 $D(\cdot, \cdot)$ 来量化退化程度。除非另有说明，我们使用以下度量：*CDPAM*、多分辨率短时傅里叶变换（MR-STFT）距离、对数梅尔均方误差和波形均方误差。主要度量的梯度仅用于攻击优化内部；在推理阶段报告所有度量，以评估鲁棒性。内循环投影梯度下降及其与 R_t 和 D 的交互如图 1(b)所示。

2.3. 目标与阈值校准

我们研究两个互补的目标。为简洁起见，令

$$\hat{x}_0 = R_t(x_t; \theta), \qquad \hat{x}_0^{\tilde{\delta}} = R_t(x_t + \sigma_t\tilde{\delta}; \theta). \tag{7}$$

（O1）固定预算最大退化

$$\max_{\|\tilde{\delta}\|_p \le \eta} D\big(\hat{x}_0,\ \hat{x}_0^{\tilde{\delta}}\big). \tag{8}$$

（O2）对抗代价

$$C_{\text{adv}}(x_0; t, \tau) = \inf\Big\{\eta \ge 0 \,\Big|\, \exists\, \tilde{\delta},\ \|\tilde{\delta}\|_p \le \eta,\ D\big(\hat{x}_0, \hat{x}_0^{\tilde{\delta}}\big) \ge \tau\Big\}. \tag{9}$$

[1]拉德马赫方向在试运行中呈现出相似趋势；为确保可复现性，我们固定使用高斯分布。

算法 1 对抗代价 C_{adv} 在时间步 t（DDIM，$p \in \{2, \infty\}$）

要求： x_0，调度表 $\{\bar{\alpha}_s\}$、t，反向过程 R_t，度量 D，范数 p，最大预算 $\eta_{\max}$，步数 K，重启次数 r，步长尺度 β，随机种子 s

通过随机种子 s 固定前向 ϵ，并通过（1）形成 x_t；设置 $\sigma_t \leftarrow \sqrt{1-\bar{\alpha}_t}$

初始化 $[l, u] \leftarrow [0, \eta_{\max}]$

对 $b \leftarrow 1$至 10**执行** ▷ 二分法（图 1a）

  $(l+u)/2$; $D^\star \leftarrow 0$

  对于 $j \leftarrow 1$到 r **执行** ▷投影梯度下降内层循环（图 1b）

    采样 $\tilde{\delta}^{(0)} \sim \mathcal{N}(0, I)$ 并投影到 $\|\cdot\|_p \le \eta$

    对于 $k \leftarrow 0$ 到 $K-1$ **执行**

      $\hat{x}_0 \leftarrow R_t(x_t)$; $\hat{x}_0^{\delta} \leftarrow R_t(x_t + \sigma_t \tilde{\delta}^{(k)})$

      $g \leftarrow \nabla_{\tilde{\delta}} D(\hat{x}_0, \hat{x}_0^{\delta})$; 更新动量

      $\tilde{\delta}^{(k+1)} \leftarrow \Pi_{\|\cdot\|_p \le \eta}\big(\tilde{\delta}^{(k)} + \alpha 步长(g)\big)$, $\alpha \leftarrow \beta\,\eta / K$

    结束循环

    $D^\star \leftarrow$最大值$\{D^\star,\ D\big(R_t(x_t),\ R_t(x_t + \sigma_t \tilde{\delta}^{(K)})\big)\}$

  结束 for

  如果 $D^\star \ge \tau$ **则**

    $u \leftarrow \eta$

  else

    $l \leftarrow \eta$

  结束 if

结束 for

返回 $C_{\mathrm{adv}} \approx u$

校准 τ。我们在一个不相交的、无标签的开发集非成员 集合 $\mathcal{D}_{\mathrm{dev}}$上预先注册 τ。对于每个（x_0，t）$\in \mathcal{D}_{\mathrm{dev}}$ 以及 L 个随机单位高斯方向 u，我们评估

$$D\Big(\hat{x}_0,\ R_t\big(x_t + \sigma_t\, \eta_{\mathrm{ref}}\, u; \theta\big)\Big), \tag{10}$$

其中前向 ϵ 按（x_0，t）固定。我们将 τ 设为所有样本与方向上的第 95 百分位数（默认 η_{ref}=0.05）。我们在全部实验中将 τ 固定为第 95 百分位数。产生 C_{adv} 的外层预算搜索以及随后的成员资格判定如图 1(a) 所示。

2.4. 优化与计算对等

~我们通过在 δ 上的投影梯度下降（投影梯度下降）求解（O1）；（O2）在 $\eta \in [0, \eta_{\max}]$ 上采用外层二分法，内层为投影梯度下降。内层投影梯度下降采用动量 0.9、重启次数 r，以及与当前预算绑定的步长，

$$\alpha \;=\; \beta\,\frac{\eta}{K}, \quad \beta \in [0.2, 0.3],$$

共 K个步数。我们通过 $\Pi_{\|\cdot\|_2 \le \eta}(\mathbf{z}) = \eta\mathbf{z}/\max(\eta, \|\mathbf{z}\|_2)$ 对 p=2 进行投影，并对 p=∞ 进行逐元素裁剪。当 $\Delta D/D < 1\%$ 连续出现 3 个步数或 $\|\nabla\|_2 < 10^{-6}$ 时触发提前停止。外层二分搜索运行 10 个步数。梯度流经 R_t 以及所选的 D；我们在各逆向步骤上启用梯度检查点。为确保公平比较，我们报告 UNet 调用次数（前向/反向）、墙钟时间以及估计的浮点运算次数。在投影梯度下降（K）与二分法（B）下，该攻击每个样本大约使用（K+2）B 次反向传递以及（K+1）B次指标评估；基线方法通过重复执行或多重t 来匹配总计算量，误差控制在 ±5% 以内。

2.5. 与基于轨迹的成员推断攻击的关系

轨迹重建攻击（例如近端初始化风格）会从（x_0，x_t）中恢复一个“真实”DDIM轨迹点 $x_{t-t'}$，并将其与模型预测的 $x'_{t-t'}$ 进行比较，使用 ℓ_p 距离度量。

相较之下，我们的 LSA-Probe 会在 x_t 处注入经时间归一化的扰动，并通过投影梯度下降搜索最坏情况方向。在白盒威胁模型下，这两类方法互为补充；我们重新实现了轨迹方法 [8]，将其作为基线，在计算量匹配的条件下进行直接比较。

2.6. 对潜在扩散（MusicLDM）的适用性

该探测方法可以直接应用于潜在扩散模型，只需将操作对象从 VAE 潜变量 z_0。对于每个波形 x_0，我们计算 $z_0 = \mathrm{Enc}(x_0)$，向前加噪至 z_t，并在潜在空间中注入经时间归一化的扰动：$z_t \mapsto z_t + \delta_t$。确定性的反向过程 $R_t(\cdot;\theta)$ 在潜在空间中运行；通过 冻结的 解码器获得由 D 评分的波形：

$$\hat{x}_0 = \mathrm{Dec}\big(R_t(z_t;\theta)\big), \qquad \hat{x}_0^{\tilde{\delta}} = \mathrm{Dec}\big(R_t(z_t + \sigma_t\tilde{\delta};\theta)\big). \tag{11}$$

梯度会通过 R_t 和固定的解码器进行反向传播；使用条件时，在配对评估中保持文本提示和引导不变。计算量统计同时包括 UNet 和解码器的调用。

3. 实验

我们评估了两个音乐扩散模型系列：(i) 以 **DiffWave** [13] 实例化的波形 DDPM，以及 (ii) 以 **MusicLDM** 实例化的潜在扩散模型（在 VAE 潜变量中运行的无条件/弱条件 LDM）[1]。数据集包括 **MAESTRO v3**（钢琴独奏）[14] 和 **FMA-Large**（多种音乐流派）[15]。我们采用确定性的 DDIM 采样、时间步比例 t_{ratio}=0.6、范数 p=2、最大预算 $\eta_{\max}$=0.8，主要距离度量为 D CDPAM；此外还报告 MR-STFT、对数梅尔均方误差和波形均方误差，以评估鲁棒性。音频以 22.05 kHz 进行训练，并通过计算图内重采样在 16 kHz 下计算各项度量。我们的威胁模型为白盒、开发者侧攻击，并具有梯度访问权限。根据先前研究 [16–18],，决策阈值固定为不相交开发集非成员样本上的 τ = 第 95 百分位数。

3.1. 评估指标与协议

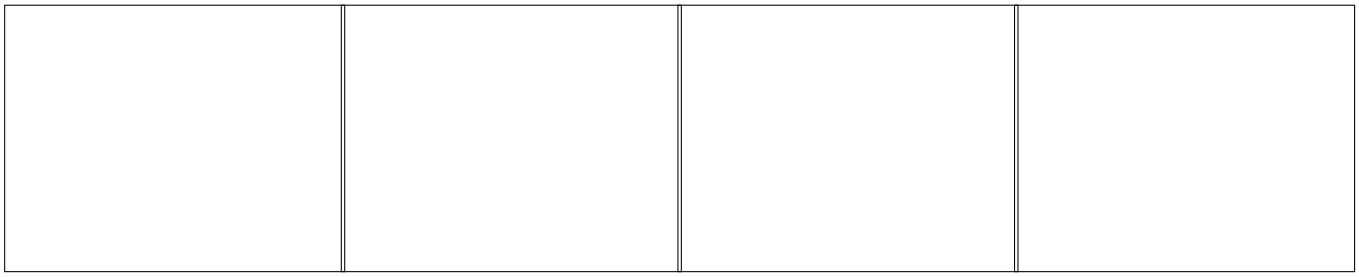
我们报告 AUC-ROC、TPR@FPR=1%、TPR@FPR=0.1%。AUC 置信区间使用*DeLong* [19]计算；其他端点使用样本级自助法（重采样 10,000 次）。对于不同时间步和度量，我们通过 Holm–Bonferroni [20]控制家族错误率。对于成员身份标签，成员池由训练音频片段构成；非成员池从留出的验证集/测试集中采样，并包含近域/跨域对照样本。我们使用 Chromaprint+局部敏感哈希去除重复曲目/翻唱版本，并进行人工抽查。

3.2. 模型、训练与数据划分

DiffWave 遵循官方 UNet 架构（训练 200k 步，批量大小为 32，使用 AdamW 和余弦学习率）。MusicLDM 在 VAE 潜空间中进行无条件训练（stride ×4）；攻击期间，解码器被冻结，梯度通过潜变量反向算子传播，而距离则在解码后的波形上计算。我们将音频切分为 4 秒的音频片段，并按作品/艺术家（MAESTRO）或曲目/艺术家（FMA）分层，以避免数据泄漏。除非另有说明，采样器均为确定性采样器（DDIM，σ_t=0）。

表1：匹配计算量下的主要结果（DDIM，t_{ratio}=0.6、p=2、$\eta_{\max}$=0.8）。我们报告真阳性率@1% FPR和AUC-ROC（均值±95%置信区间）。“最佳基线”是损失/轨迹/SecMI中的最佳结果。 Δ 为我们的方法减去最佳基线。

模型	数据集	最佳基线（真阳性率@1% / 曲线下面积）	我们的方法（真阳性率@1% / 曲线下面积）	Δ （我们的方法 −最佳）	
		–	–	真阳性率@1%	AUC
MusicLDM	MAESTRO	$0.10\ (0.07–0.12)/0.58 \pm 0.02$	$\mathbf{0.13\ (0.10–0.15)/0.61 \pm 0.02}$	$+0.03$	$+0.03$
MusicLDM	FMA-Large	$0.08\ (0.05–0.10)/0.56 \pm 0.01$	$\mathbf{0.14\ (0.10–0.16)/0.59 \pm 0.02}$	$+0.06$	$+0.03$
DiffWave	MAESTRO	$0.12\ (0.09–0.15)/0.63 \pm 0.02$	$\mathbf{0.20\ (0.16–0.24)/0.67 \pm 0.02}$	$+0.08$	$+0.04$
DiffWave	FMA-Large	$0.11\ (0.08–0.14)/0.62 \pm 0.02$	$\mathbf{0.18\ (0.14–0.22)/0.66 \pm 0.02}$	$+0.07$	$+0.04$



(a) 受试者工作特征 @ t_{ratio}=0.6(DiffWave / MusicLDM）。 （b） $t_{\mathrm{ratio}} \in \{0.2, 0.4, 0.6, 0.8\}$。 （c）预算 $\eta \in [0.05, 0.8]$。 （d） $D \in$ {CDPAM、MR-STFT、对数-梅尔均方误差、波形均方误差}。

图 2：**固定 τ=第 95 百分位数** 下的关键分析（DDIM，p=2，$\eta_{\max}$=0.8）。（a）具有 95% 置信区间的受试者工作特征曲线；我们的 LSA-Probe 改善了低 FPR 区域的表现。（b）时间步消融：轨迹中段的时间步带来更强的可分性。（c）预算消融：较大的预算有所帮助，直至效果略微趋于饱和。（d）度量鲁棒性：低 FPR 下，感知指标（CDPAM/MR-STFT）优于与训练对齐的均方误差。

3.3. 基线方法与计算量核算

基线方法包括：损失对齐方法（在 t 处及终点处的去噪/重建损失） [21], 轨迹方法（PIA/PIAN） [8], SecMI [22]。为确保公平，我们在 ±5% 的样本上进行匹配计算，并使用 A100-80GB 测量每个样本的实际运行时间。我们的对抗代价 C_{adv} 采用外层二分搜索（B=10），内层使用投影梯度下降（K=12 步，重启次数 r=5），每个样本总计约需（K+2）B 次反向传递。

3.4. 主要结果

表1报告了主要终点（TPR@1% FPR）和计算量匹配条件下的曲线下面积-受试者工作特征（DDIM，t_{ratio}=0.6，p=2，$\eta_{\max}$=0.8）。在两个数据集和模型系列中，与最强基线方法相比，LSA-Probe 始终改善低 FPR 工作点：对于 MusicLDM（LDM），MAESTRO 和 FMA-Large 上的 TPR@1% 分别从 $0.10 \to 0.13$ 和 $0.08 \to 0.14$ 提升（绝对增益分别为 $+0.03$ 和 $+0.06$），两个数据集上的曲线下面积均提升 $+0.03$；对于 Dif-fWave（DDPM），MAESTRO 和 FMA-Large 上的 TPR@1% 分别从 $0.12 \to 0.20$ 和 $0.11 \to 0.18$ 提升（增益分别为 $+0.08$ 和 $+0.07$），且两个数据集上的曲线下面积提升幅度更大，均为 $+0.04$。重要的是，尽管取得了这些提升，LDM 相较于 DDPM 仍对成员推断攻击具有更强的鲁棒性。

3.5. 关键分析：受试者工作特征、时间步、预算和度量

图2将信息量最大的四项分析整合到一个 4×1 面板中。**(a) 受试者工作特征：** 在 t_{ratio}=0.6下，LSA-Probe 提升了各模型在低 FPR 区域的表现，与表1一致。**(b) 时间步：** 扫描 $t_{\mathrm{ratio}} \in \{0.2, 0.4, 0.6, 0.8\}$ 后发现，中间轨迹位置（0.6）的可分性最高，这与以下直觉一致：反向路径从粗略的全局布局逐渐过渡到更精细的细节。**(c) 预算：** 增加 η 可提升真阳性率@1%，直至接近 $\eta \approx 0.6$–0.8，此时饱和率仍然较低。**(d) 度量：** 感知度量（CDPAM、MR-STFT）在低 FPR 下具有更强的区分能力，而与训练对齐的均方误差虽不及前者，但仍可作为有用的补充视角。

3.6. 实现细节

我们使用样本专用种子固定每个 (x_0, t) 的前向噪声，从而使配对评估能够隔离 δ_t 的影响；在多次重复实验中，我们对 S 个随机种子取平均。对抗内循环采用投影梯度下降，使用 K=12 步、动量 0.9、r=5 次重启，以及步长为 $\beta\,\eta/K$ $\beta \in [0.2, 0.3]$ $\Delta D/D$<1 with ；当连续三步达到 % 或 $\|\nabla\|_2 < 10^{-6}$ 时触发提前停止。

4. 结论

我们提出了 **LSA-Probe**，这是一种针对音乐扩散模型的白盒成员身份推断攻击，通过测量反向轨迹上的时间归一化潜在稳定性，并将超过固定感知退化阈值所需的对抗预算 作为得分。在两个模型系列（DiffWave 和 MusicLDM）以及两个语料库（MAESTRO v3、FMA-Large）上，在计算量匹配的条件下，LSA-Probe 始终能够相较于强大的基线方法提升低 FPR 检测性能。消融实验表明，中段轨迹时间步和适度预算所提供的信息最为丰富，并且感知度量（CDPAM/MR-STFT）比基于均方误差的距离更可靠地揭示稳定性差异。

5. 致谢

本研究获得江苏省基础研究计划项目BG2024027以及西交利物浦大学研究发展基金的资助，

发展基金（项目编号：RDF-22-02-046）。

6. 参考文献

[1] Ke Chen、Yusong Wu、Haohe Liu、Marianna Nezhurina、Taylor Berg-Kirkpatrick和Shlomo Dubnov，“MusicLDM：使用节拍同步混合策略增强文本生成音乐中的新颖性”，载于*ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*。IEEE，2024年。[2] Sanjoy Chowdhury、Sayan Nag、KJ Joseph、Balaji Vasan Srinivasan和Dinesh Manocha，“Melfusion：使用扩散模型根据图像和语言线索合成音乐”，载于*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*，2024年，第26826–26835页。[3] Tim W Dornis和Sebastian Stober，“生成式人工智能训练与版权法”，*arXivpreprint arXiv:2502.15858*，2024年。[4] Reza Shokri、Marco Stronati、Congzheng Song和Vitaly Shmatikov，“针对机器学习模型的成员推断攻击”，载于*2017 IEEE Symposium on Security and Privacy (SP)*。IEEE，2017年，第3–18页。[5] Kushal Tirumala、Aram Markosyan、Luke Zettlemoyer和Armen Aghajanyan，“无需过拟合的记忆：分析大语言模型的训练动态”，*Advances in Neural Information Processing Systems*，第35卷，第38274–38290页，2022年。[6] Benjamin Hilprecht、Martin H¨arterich和Daniel Bernau，“针对生成模型的蒙特卡洛与重构成员推断攻击”，载于*Proceedings on Privacy Enhancing Technologies*。Sciendo，2019年，第2019卷，第232–249页。[7] Lorenzo Rossi、Michael Aerni、Jie Zhang和Florian Tram`er，“针对序列模型的成员推断攻击”，*arXiv preprintarXiv:2506.05126*，2025年。[8] Fei Kong、Jinhao Duan、RuiPeng Ma、Heng Tao Shen、Xiaoshuang Shi、Xiaofeng Zhu和Kaidi Xu，“一种通过近端初始化实现的高效扩散模型成员推断攻击”，载于*TheTwelfth International Conference on Learning Representations (ICLR)*，2024年。[9] Nitish Shirish Keskar、Dheevatsa Mudigere、Jorge Nocedal、Mikhail Smelyanskiy和Ping Tak Peter Tang，“关于深度学习的大批量训练：泛化差距与尖锐极小值”，载于*International Conference on Learning Representations (I CLR)*，2017年。[10] Yiyong Liu、Zhengyu Zhao、Michael Backes和Yang Zhang，“通过利用损失轨迹实施成员推断攻击”，载于*Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*，2022年，第2085–2098页。[11]Tomoya Matsumoto、Takayuki Miura和Naoto Yanai，“针对扩散模型的成员推断攻击”，载于*2023 IEEE Security and Privacy Workshops (SPW)*。IEEE，2023年，第310–321页。[12] Jonathan Ho、Ajay Jain和Pieter Abbeel，“去噪扩散概率模型”，载于*Advances in Neural Information Processing Systems*，2020年，第33卷，第6840–6851页。[13] Zhifeng Kong、Wei Ping、Jiaji Huang、Kexin Zhao和Bryan Catanzaro，“DiffWave：一种用于音频合成的通用扩散模型”，载于*International Conference on Learning Representations (ICLR)*，2021年。

[14] Curtis Hawthorne、Andriy Stasyuk、Adam Roberts、Ian Simon、Cheng-Zhi Fu、Sanders Green、Sander Dieleman、Erich Hentschel 和 Douglas Eck，《借助 MAESTRO 数据集实现钢琴音乐的因子化建模与生成》，载于国际学习表征会议 *(ICLR)*，2019年。[15] Michael Defferrard、Kirell Benzi、Pierre Vandergheynst 和 Xavier Bresson，《FMA：一个用于音乐分析的数据集》，载于第 *18* 届国际音乐信息检索学会会议 *(ISMIR)*，2017年，第 316–323 页。[16] Nicholas Carlini 等，《从扩散模型中提取训练数据》，载于*USENIX* 安全会议，2023年。[17]对抗鲁棒性工具箱开发者，《仅标签成员推断：使用最高 t 百分位数校准距离阈值》，https: // adversarial-robustness-toolbox. readthedocs.io/en/latest/modules/attacks/inference/membership_inference.html，2024年。[18] Zheng Li 和 Yang Zhang，《机器学习模型仅标签暴露中的成员信息泄露》，载于*2021*年 *ACM SIGSAC* 计算机与通信安全会议 *(CCS)*论文集，2021年。[19]Elizabeth R. DeLong、David M. DeLong 和 Daniel L. Clarke-Pearson，《比较两个或多个相关受试者工作特征曲线下面积：一种非参数方法》，*Biometrics*，第 44 卷，第 3 期，第 837–845 页，1988年。[20] Sture Holm，《一种简单的序贯拒绝多重检验程序》，*Scandinavian Journal of Statistics*，第 6 卷，第 2 期，第 65–70 页，1979年。[21] Yan Pang、Tianhao Wang、Xuhui Kang、Mengdi Huai 和 Yang Zhang，《针对扩散模型的白盒成员推断攻击》，隐私增强技术论文集，2025年。[22] Jingwei Li、Jing Dong、Tianxing He 和 Jingzhao Zhang，《面向扩散模型的黑盒成员推断攻击》，载于*ICLR 2025*机器学习深度生成模型研讨会：理论、原理与效能。