arXiv:2611.32023v1 [cs.SD] 24 Nov 2026

IncrementDET：通过注册实现持续式人工智能生成音乐检测

杜润宝 芝加哥大学
runbao@uchicago.edu

Wei Li*复旦大学
weili-fudan@fudan.edu.cn

何启乐*复旦大学
26113050261@m.fudan.edu.cn

摘要

人工智能生成音乐在音乐流媒体平台上已十分普遍，听众往往无法将其与人类创作的音乐区分开来。因此，有必要进行 *AI* 音乐检测，但在固定生成器集合上训练的检测器通常无法应对未见生成器。由于新生成器频繁出现，更新一个涵盖所有生成器的通用检测器通常需要大量训练数据和计算资源，同时还可能导致其在此前已学习的生成器上的性能下降。为此，我们提出 *IncrementDET*，这是一种通过称为 注册 的过程学习每个新生成器的检测器。对于每个生成器，*IncrementDET* 训练一个线性分类器，我们称之为“针”，以将该生成器的样本与固定的人类创作参考音轨区分开来。注册一个新生成器只需增加一根针，而此前构建的针保持不变。随着针的增加，*Sid´ak* 校正可使整体假阳性率维持在目标水平。为了在这些条件下比较检测器，我们定义了持续式人工智能生成音乐检测 任务，并引入了一个生成器逐一到达的基准测试。该基准测试衡量检测器在每次更新前后的准确率以及每次更新的成本。针对这一基准测试，我们还构建了 *EvoMusicFake-17K*，这是一个包含来自六个近期生成器和四个人类音乐平台的器乐音乐数据集。与 *SpecTTTra*（一种最先进的通用检测器）相比，*IncrementDET* 仅使用少量样本就能更准确地学习新生成器，对未见生成器的泛化能力更强，并且能更好地保持在此前已学习生成器上的检测性能，同时更新成本大幅降低。*IncrementDET* 的结构同样可以应用于生成图像和视频的检测，因为新生成器的发布频率同样很高。

1. 引言

流媒体平台上人工智能生成音乐的迅速泛滥，引发了人们对内容真实性的日益担忧

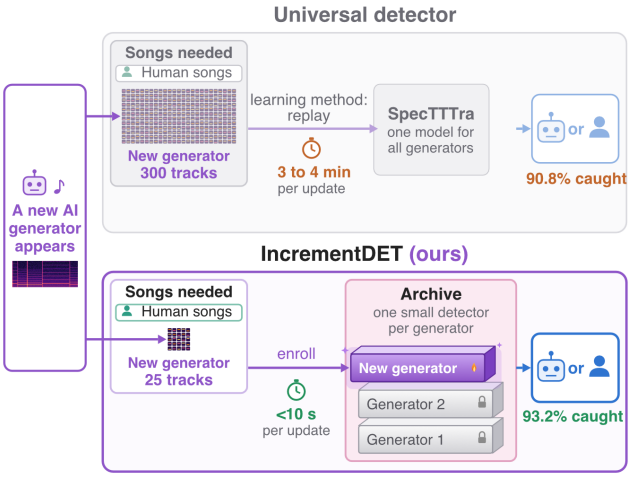


图 1。**我们工作的示意图**。上：通用检测器 SpecTTTra 通过经验回放，使用300条新生成器的音轨进行更新。下：IncrementDET 仅将25条新生成器的音轨输入冻结编码器，并添加一个线性分类器（一个针），同时保持早期的针不变。每一行显示一次更新所需的时间，以及在新生成器上的 5% 假正率下的真正率，该数值取十二个生成器的平均值（表 5），因此 IncrementDET 使用更少的音轨和更少的计算量，却能取得更高的准确率。

以及对人类艺术家的保护。Deezer 报告称，2026年6月的某些日子里，完全由人工智能生成的音轨占其每日投稿量的一半以上；在一项有 9，000 名参与者的盲听测试中，97% 的人无法区分完全由人工智能生成的音乐与人类创作音乐 [14]。这些发展凸显了可靠的 AI 生成音乐检测的必要性，以确保内容透明度并支持负责任的平台治理。

然而，由于对未见生成器的泛化能力有限（第 2节）以及新生成器的快速涌现，维护一个可靠的通用检测器仍然面临挑战。2023 年至 2025 年间，至少有十二个文本到音乐生成器公开发布， [8, 12, 17, 19, 25, 28, 30, 33, 36, 39, 47, 52], 因此，每当出现新生成器时，通用检测器都必须进行更新，而此时通常只能获得新生成器的少量样本。

然而，对于音乐，尚无先前研究探讨

*共同通信作者。何启乐是第一通信作者。

检测器应如何适应不断出现的新生成器。现有基准测试（如 SONICS [43], FakeMusicCaps [11]和 MoM [5] ）会在固定的生成器划分上对检测器进行一次评估，其中一些基准测试会留出生成器来测试泛化能力，但没有任何基准测试衡量检测器学习新生成器的能力。因此，我们定义了一项新任务——持续式人工智能生成音乐检测，并为此引入了一个基准测试，同时构建了一个新的生成音乐和人类创作音乐数据集。在该基准测试中，十二个生成器依次到来；在每一步，每个检测器都会接收到新生成器的相同样本以及相同的人类音轨。我们会衡量每个检测器在更新前后识别新生成器的准确度、对早期生成器的遗忘程度，以及每次更新所需的计算成本。

现有检测器难以在不遗忘早期生成器的情况下，高效适应新生成器。更大的模型和更多训练数据能够提升泛化能力，但在未见生成器上的检测准确率仍显著较低（第 2 节）。仅使用人类创作音乐的方法（例如 MusicDET）[21],对人类创作音乐进行建模，并将低似然音轨标记为人工智能生成。然而，密度模型可能会为分布外输入分配较高的似然 [38], ，从而可能漏检新生成器；与此同时，其仅使用人工数据进行训练的范式无法利用新获得的生成数据进行适应。通用检测器则可以从新生成器中学习，但需要大量新生成的训练数据，而这些数据可能难以获取。此外，微调存在遗忘早期生成器的风险 [34], ，而完全重新训练则需要保留历史样本，并会产生高昂的计算成本。这些局限促使我们设计一种检测器：能够仅凭少量样本准确学习新生成器，同时将计算开销降至最低，并且不损害此前学习到的检测能力。

因此，我们提出 IncrementDET——一种通过廉价更新来学习每个新生成器、从而满足这些要求的检测器；我们将这种更新称为 注册。该检测器使用冻结的音乐编码器（ MuQ）提取特征，并在时间维度上进行池化，以表示每条音轨。对于每个生成器，它训练一个逻辑回归分类器，我们称之为针，用于将该生成器的样本与固定的人类创作参考音轨集合区分开来。每根针都有自己的阈值，该阈值在单独的人类创作校准音轨集合上设定；此外，Šid´ák 校正会设定每根针的假阳性率，使整个检测器的假阳性率保持在其目标值。一条音轨只要有任意一根针给出的分数超过其阈值，就会被标记为人工智能生成。因此，注册一个新生成器只需使用少量样本在几秒内训练一根针，并且不会修改任何现有的针（图 1）。

我们将 IncrementDET 与 SpecTTTra [43], 进行比较。SpecTTTra 是一种先进的通用检测器，我们采用多种策略对其进行更新，直至使用所有存储样本进行完全重新训练。在标准的跨生成器评估中，IncrementDET 在全部 12 个留出生成器上的等错误率均更低；在我们的持续学习基准中，无论是在学习每个新生成器之前还是之后，表现也都更出色。在登记新生成器之前，IncrementDET 对其检测的平均真正率为 57.2%（假正率为 5%）；相比之下，SpecTTTra 在任何更新策略下的平均真正率最高仅为 27.4%。在序列的每一步中，对于迄今为止学到的生成器，IncrementDET 的检测准确度都高于采用任何更新策略的 SpecTTTra。完成最后一步后，IncrementDET 对第一个生成器的检测仍达到平均真正率 96.4%（假正率为 5%），而采用完全重新训练的 SpecTTTra 为 79.4%。仅使用新生成器的 25 条音轨，IncrementDET 对其检测的平均真正率就达到 93.2%，同时对其他生成器的检测率保持在 99.2%；相比之下，微调后的 SpecTTTra 达到 84.6%，且检测率降至 43.1%。

总而言之，我们作出以下贡献。

- 我们定义了持续人工智能生成音乐检测任务，并介绍了一个基准测试和 EvoMusicFake-17K——一个包含生成音乐和人工数据器乐的数据集。
- 我们提出了 IncrementDET，这是一种基于注册的检测器，通过独立的线性分类器（针）学习新的生成器，并采用 Sid´ak 校正，在生成器不断增加时控制整体假阳性率。
- 大量实验表明，在标准跨生成器评估和我们的持续检测基准测试中，IncrementDET 的表现均优于最先进的通用检测器，准确率更高、知识保留效果更好，更新成本也更低。
- 我们表明，通用检测器难以适应新生成器，需要大量训练数据并进行成本高昂的重新训练，才能避免遗忘早期生成器；相比之下，注册只需少量样本、以较低的计算成本即可学习新生成器，同时保留先前的性能。

2. 相关工作

音频深度伪造检测模型。大多数生成音频检测器都是判别式分类器，经过训练以区分人工制作的录音和人工智能生成的录音。最初的设计源自语音反欺骗。RawNet2 [48]直接从原始波形中学习特征，而 AA-SIST [26]则对频谱和时间特征应用图注意力。后来的检测器使用预训练的自监督编码器（例如 wav2vec 2.0 [4] 或 XLS-R [3], ）从音频中提取特征，并在这些特征上训练一个小型分类器，例如 AASIST [49], 。许多人工智能生成音乐检测器直接采用这些语音架构，例如将音乐编码器 MERT [32]与 AASIST 分类器 [21]结合。其他检测器则专为音乐设计。SONICS [43] 推出了 SpecTTTra，这是一种 Transformer，可将频谱图拆分为独立的频谱词元和时间词元，从而高效处理长曲目。其他方法采用双流对比学习 [5], 、音乐内在特征 [22], 和取证残差——

—— [40], 或基于 CLAP [51] 嵌入 [13]的分类器。本文使用 SpecTTTra 作为通用检测器，因为它是人工智能生成音乐检测中的常用基线 [5, 13, 21]，而且其公开的训练代码使我们能够在每个步骤中使用与 IncrementDET 相同的数据对其进行更新。

跨生成器泛化。先前研究表明，人工智能生成音乐检测器在未见过的生成器上表现不佳。在 MusicGen 上训练并在 Mustango 上测试的检测器，其等错误率分别为：SpecTTTra-α 为 38.9%，wav2vec2-AASIST 为 48.7%，MERT-AASIST [21] 为 46.4%。在 SONICS 上，所有经过评估的架构对训练中见过的生成器的检测效果都优于未见过的生成器，并且 SpecTTTra-α 的 F_1分数在后续的两个生成器 SeedMusic 和 SkyMusic 上分别降至 0.56 和 0.25 [43]。在 SONICS 上训练的 SpecTTTra 在 SONICS 上达到 0.972 的 F_1 分数，但在 Riffusion 和 YuE 上分别仅为 0.535 和 0.688 [5]。因此，无法指望仅在一组固定生成器上训练一次的检测器，在新生成器发布后仍能保持准确。

持续学习。随着新的生成器不断出现而更新检测器属于持续学习问题，在这一过程中，模型一次学习一个任务序列，同时必须保持在早期任务上的性能。这之所以困难，是因为仅使用新数据进行训练会导致灾难性遗忘 [18, 34]。经验回放 [44] 和弹性权重固化 [27] 可以减轻这种遗忘；而专为语音深度伪造检测器设计的方法，例如 RAWM [53]和 RegO [10], ，则通过限制检测器权重在每次更新期间的变化来减少遗忘。持续检测已针对图像 [31] 和语音深度伪造 [54], 进行了基准测试，但尚未应用于音乐。E3 [2] 在有限数据下为每个新的图像生成器训练一个专家模型，但会在每次更新时使用记忆缓冲区重新训练一个网络，以融合这些专家模型。IncrementDET 的设计最接近类别增量方法，这类方法会冻结预训练模型，例如 SimpleCIL [55]和 RanPAC [35], ；其中，每个新类别都由一个原型表示，因此不会遗忘任何早期类别。然而，这些方法只能在已知类别之间进行分类，无法控制假阳性率。

注册。我们对注册 一词的使用源自说话人验证：在该任务中，通过对新说话人的少量话语的嵌入取平均值来添加该说话人，只有当测试话语的嵌入与这一平均值足够接近时，才会接受该测试话语 [50]。与 SimpleCIL 和 RanPAC 一样，IncrementDET 利用少量样本添加每个生成器，而无需重新训练；与说话人验证类似，它会针对每个生成器进行阈值判定。此外，它还针对每个生成器，以人类创作音乐为对象训练线性分类器，并在人类数据上校准其阈值。

3. 持续性 AI 音乐检测基准

3.1. 评估协议

在持续式人工智能生成音乐检测中，生成器g_1、.....、 g_T 按照固定顺序逐一到来，检测器必须将它们的输出与人类创作音乐区分开来，我们用 H 表示人类创作音乐的分布。检测系统是一个序列 D_0、 D_1.....、 D_T，其中， D_{t-1} 在 g_t到来之前确定，而 D_t 则是系统使用 g_t 到来时提供的样本后得到的检测器。我们测量两个量：

$$\textsc{Pre}(t) = \text{performance of } D_{t-1} \text{ on } g_t, \quad (1)$$

$$\textsc{Post}(t) = \text{performance of } D_t \text{ on } g_1, \ldots, g_t \text{ and } H. \quad (2)$$

D_0 尚未见过生成器，因此报告 $t \geq 2$的 PRE。PRE 衡量检测器在不进行更新时所达到的效果，而 POST 衡量更新是否能够恢复对新生成器的检测能力，同时不降低对早期生成器的检测能力。

每次新数据到达时，每个检测器都会接收相同的 N 标记样本和相同的人类音轨，并记录其更新所需的 GPU 时间。

3.2. EvoMusicFake-17K 数据集

我们为该基准测试构建了一个数据集 EvoMusicFake-17K。该数据集包含 6,000 条音轨（115 小时），由六个近期生成器生成；还包含 11,111 条人工音轨（651 小时），收集自四个平台。除 EvoMusicFake-17K 外，该基准还使用了 FakeMusicCaps 中五个生成器生成的音轨 [11], Suno [47] 音轨（由我们生成但不予再分发），以及来自公开数据集 FMA [15] 和 MagnaTagATune [29]的人工音轨。所有音轨均为器乐曲，第 6 节将讨论这一选择。人工音轨比生成音轨更长，但每个检测器只对每条音轨中的一个 5 秒窗口进行评分。我们将发布 EvoMusicFake-17K 的提示词、生成设置、曲目标识符和收集脚本。

AI 生成器。我们使用 $T = 12$ 文本生成音乐生成器，其首个公开版本于 2023 至 2025 年间发布。这些生成器家族涵盖了当前音乐生成器的主要设计方案（表 1）。其中五种生成器的音轨取自 FakeMusicCaps，该数据集提供由 MusicCaps 描述文本生成的 10 秒片段 [1]。其余七种生成器的音轨由我们自行生成。EvoMusicFake-17K 中的六种生成器使用相同的 1,000 条器乐提示词，每种音乐流派各 125 条。Suno 使用其中 500 条提示词，每条生成两个输出。

人类创作音乐。我们从四个平台收集了人类音轨：Jamendo [24], Pixabay [42], ccMixter [7]和 Internet Archive 网标合集 [23]，并使用平台标签筛选纯器乐曲目，同时排除 Pixabay 上标记为人声的音轨。由于部分平台接受生成音乐上传内容，我们对两个大型来源进行了筛查，以排除

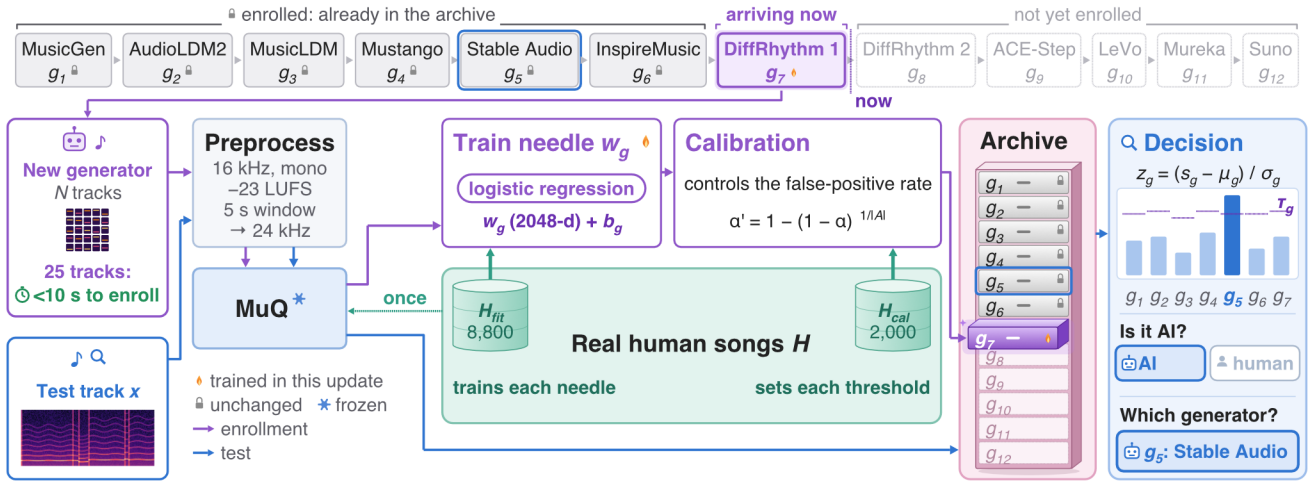


图 2. **IncrementDET 概览**。上：生成器按到达顺序排列，DiffRhythm1 到达时，较早的生成器已完成登记。紫色箭头表示注册。新生成器的 N 音轨经预处理后，由冻结的 MuQ 编码器编码；随后，针对来自 $\mathcal{H}_{\text{fit}}$ 的人类创作参考音轨训练一个针，并在 $\mathcal{H}_{\text{cal}}$ 上进行校准，再将其加入存档，而不更改较早的针。蓝色箭头表示测试。测试音轨由存档中的每个针进行评分；如果任一针的分数超过其阈值，决策面板就会将其标记为人工智能生成，并指出产生最高标准化分数的针所对应的生成器。

表 1. 生成器及其到达顺序。LM：语言模型。FM：流匹配。Transf.：Transformer。

生成器	版本	系列
MusicGen [12]	中等	自回归语言模型
AudioLDM2 [33]	音乐	潜在扩散
MusicLDM [8]	–	潜在扩散
Mustango [36]	–	潜在扩散
Stable Audio Open [17]	1.0	扩散 Transformer
InspireMusic [52]	1.5B-Long	词元 + 流匹配
DiffRhythm 1 [39]	v1.2	流匹配
DiffRhythm 2 [25]	–	流匹配
ACE-Step [19, 20]	v1.5	语言模型 + 渲染
LeVo [30]	大型	语言模型 + 渲染
Mureka [28]	mureka-9	未公开
Suno [47]	v5、v4。5+	未公开

生成音乐。在 Jamendo 上，74% 的音轨于 2023 年之前发布；我们排除了上传行为表明其包含生成音乐的账户，例如在短时间内上传大量内容的账户。在 Pixabay 上，我们排除了被单独训练的检测器判定为可能属于生成音乐的音轨。对于 FMA，我们保留了流派标签为“器乐”的音轨；对于 MagnaTagATune，则保留了众包标签为“无人声”的片段。ccMixter 和 Internet Archive网络厂牌合集构成一组留出来源，从未用于训练或校准。时长短于 10 秒的音轨将被排除。补充材料列出了所有来源和许可证。

数据划分。所有划分互不重叠，并根据音轨标识符的固定哈希值进行分配。每个生成器包含 800 条拟合音轨和 200 条测试音轨。Ja-mendo、FMA、Pixabay 和 MagnaTagATune 的人类音轨被划分为一个包含 $\mathcal{H}_{\text{fit}}$ 8，800 条音轨的拟合集、一个模型选择集（包含 2，200 条音轨）、一个包含 $\mathcal{H}_{\text{cal}}$ 2，000 条音轨的校准集，以及一个包含 $\mathcal{H}_{\text{test}}$ 2，000 条音轨的测试集。留出集 $\mathcal{H}_{\text{new}}$ 由来自 ccMixter 和 Internet Archive网络厂牌合集的 815 条音轨组成。生成音轨没有校准划分，因为阈值仅根据人类创作音乐进行设置。模型选择集仅用于为 SpecTTTra 选择检查点和超参数。补充材料列出了每个来源在各个划分中的音轨数量。

预处理与捷径移除。所有音频均裁剪为至多居中的 240秒，重采样为 16 千赫兹单声道，并将响度归一化至 −23LUFS，且所有检测器均对每条音轨的相同 5秒窗口进行评分。这减少了由平台存储音频的方式而非音乐本身产生的线索，例如采样率、比特率、编解码器、响度和声道布局。在现有数据集上，仅凭采样率、比特率或开头的静音就足以进行检测 [6, 13, 37, 46]。该5秒窗口还移除了时长线索（第 5.3节）。

3.3. 评估指标

我们报告等错误率（EER），以及在假正率（FPR）为 5% 和 1% 时的真正率（TPR）。较低的假正率之所以重要，是因为当检测器设置阈值以使少量人类音轨被误标记时，即使其等错误率较低，仍可能漏检大多数生成轨迹。所有检测器都在同一个人类校准集 $\mathcal{H}_{\text{cal}}$ 上设置阈值。随后，我们在 $\mathcal{H}_{\text{test}}$ 上测量其假正率；$\mathcal{H}_{\text{test}}$ 来自与 $\mathcal{H}_{\text{cal}}$ 相同的来源；同时还在 $\mathcal{H}_{\text{new}}$ 上进行测量，$\mathcal{H}_{\text{new}}$ 来自所有检测器都未见过的来源，用于测试目标率是否适用于新来源。我们还报告归因准确率，即被检测出的生成轨迹中被归因于正确生成器的比例。遗忘通过以下指标进行测量：

每个后续步骤完成后，第一个生成器的真正率（TPR）。

4. 方法

图 2 概述了 IncrementDET。每个已登记的生成器都有一个针测器，即一个基于冻结音乐特征的线性分类器，用于将该生成器与人工数据参考音轨区分开来（第 4.1 和 4.2 节）；每个针测器还根据人工数据校准音轨设置了一个带 Sid´ak 校正的阈值（第4.3 节）。登记新生成器会新增一个针测器并更新阈值（第 4.4 节）；在测试时，如果某个针测器超过其阈值，则将该音轨标记出来（第4.5 节）。

4.1. 特征提取

¯令 f 为一个冻结的自监督音频编码器，$f(x) \in \mathbb{R}^d$ 为音轨 x的池化表示f 。在所有实验中， [56] 均为 MuQ [56] (检查点 `MuQ-large-msd-iter`）。将第 3.2 节中的 5 秒窗口重采样为 24 千赫兹，对一个编码器层的隐状态沿时间维度进行池化，并将其均值和标准差拼接，因此得到$d = 2048$。冻结编码器可使每个针测器保持在同一特征空间中，因此人工数据参考集的特征只需提取一次。

该层根据固定规则从编码器的 13 层中选出。对于每一层以及每个可用于训练的生成器，在该生成器的 80% 拟合音轨上训练一个针，并在剩余的 20% 音轨上进行评估；每次评估都使用一组单独的人类拟合音轨作为对照，最终选用平均等错误率最低的层。在留一生成器比较中，每一轮都会分别应用该规则，因此留出生成器不会影响选择。在持续比较中，该规则仅在第一步应用一次，并且只使用第一个生成器的拟合音轨。在所有情况下，该规则都会选择第 0 层或第 1 层。

4.2. 针

对于已登记生成器集合 $\mathcal{A}$ 中的每个生成器 g，IncrementDET 都会存储一个针（w_g，b_g），其中包含 $w_g \in \mathbb{R}^d$。给定该生成器的 N 样本 $\mathcal{X}_g$，该针是正则化逻辑回归的解:

$$(w_g, b_g) = \arg\min_{w,\,b} \sum_{x \in \mathcal{X}_g \cup \mathcal{H}_{\text{ref}}} \ell\big(y_x,\ \langle w, \bar{f}(x)\rangle + b\big) + \lambda \|w\|^2, \quad (3)$$

其中 $y_x = 1$ 生成数据取0，人工音轨取该值，其中$\mathcal{H}_{\text{ref}}$ 是来自 $\mathcal{H}_{\text{fit}}$（第 5.1 节）的固定人工音轨集合。特征在针的训练集上进行标准化，ℓ 是类别平衡逻辑损失，且 λ 对每个生成器都是固定的（逆正则化强度为 $C = 1$）。

我们针对每个生成器分别使用一个线性分类器，而不是使用一个共享分类器，原因有三点。第一，每根针仅使用其所属生成器的样本和人工参考集进行训练，因此，注册一个生成器无需存储早期生成器的样本，也不会影响它们的针。第二，在严格的假阳性率下，即使新生成器的样本很少，独立的针也比一个共享分类器检测出更多生成轨迹，并将假阳性率保持在其目标值（表 4）。第三，响应最强的针可以识别生成器，因此检测与归因来自同一个模型。

4.3. 校准与决策

每根针 g 都会为音轨 x 分配分数 $s_g(x) = \langle w_g, f(x)\rangle + b_g$；如果任意一根针的分数超过该针的阈值 τ_g，则将该音轨标记为人工智能生成。因此，每登记一根针，都会增加在人类创作音乐上发生误报的机会。我们将针对不断扩大的针集合所做的决策视为多重比较问题，其中测试次数随每次注册而增加。据我们所知，此前没有人工智能生成内容检测器能够控制这种增长。如果 $|\mathcal{A}|$ 根已登记的针各自的假阳性率为 α'，且它们在人类创作音乐上的错误彼此独立，那么总体假阳性率将为$1-(1-\alpha')^{|\mathcal{A}|}$。为使其维持在 α，IncrementDET 在每次注册时都会通过经典的 Sid´ak 校正重新计算 α'˜ [45]，

$$\alpha' = 1 - (1 - \alpha)^{1/|\mathcal{A}|}, \quad (4)$$

使总体假阳性率为 α。每根针的阈值τ_g 是其分数的 ($1 - \alpha'$) 分位数，分数取自$\mathcal{H}_{\text{cal}}$。有十二根针且 $\alpha = 5\%$ 时，$\alpha' \approx 0.43\%$，因此每个阈值大约由得分最高的九条 2，000 条校准音轨确定。因此，该校正需要一个包含数千条音轨的校准集。不同针在人类创作音乐上的分数呈正相关，因为所有针读取的是相同的特征。因此，对于校准集中所代表的人类来源，该校正在实践中略显保守；而对于新的来源，测得的比率可能会超过目标值（第 5.3 节）。

固定假正率下的真正率使用每个针的这一决策。等错误率要求每条音轨对应一个分数，因此使用最大标准化分数 $z^{\max}(x) = \max_{g \in \mathcal{A}} z_g(x)$，其中$z_g(x) = (s_g(x) - \mu_g)/\sigma_g$，$\mu_g$ 和σ_g 分别是该针在 $\mathcal{H}_{\text{cal}}$上的分数均值和标准差。归因返回 $\operatorname{argmax}_g z_g(x)$，即相较于其在人类创作音乐上的分数，该生成器的针响应最强。

4.4. 注册

当新的生成器 g 携带 N 条音轨到来时，IncrementDET 会执行四个步骤。它提取 N条音轨的特征，训练针（w_g，b_g），计算该针在 $\mathcal{H}_{\text{cal}}$ 上的分数以获得 μ_g 和 σ_g，并根据存储的校准分数重新计算 α' 和所有阈值。

ˇ注册存在一项成本。由于 Sid´ak 校正会随着登记的生成器增多而收紧每个阈值，因此，尽管早期针的方向未发生变化，其真正率仍可能略有下降。这项成本几乎完全由最弱的针承担（第 5.3 节）。

4.5. 推理与存储

在测试时，一条音轨只需通过编码器一次，所有针会同时应用。每根针的标准化只会平移和缩放每个特征，因此可以将其吸收到该针的权重和偏置中。随后将这些针堆叠起来，即可通过一次矩阵-向量乘法得到所有分数，

$$\mathbf{s}(x) = W\bar{f}(x) + \mathbf{b}, \qquad W \in \mathbb{R}^{|\mathcal{A}|\times d},\ \mathbf{b} \in \mathbb{R}^{|\mathcal{A}|}. \quad (5)$$

因此，每个针的音轨评分需要进行一次编码器推理，并计算一次长度为 d 的点积。在消费级 GPU 上，每个窗口的编码器推理耗时约为 7.4 毫秒；相比之下，十二个针的点积计算耗时几乎可以忽略，因此增加针几乎不会改变推理时间。

5. 实验

5.1. 实验设置

基线方法。 我们将 IncrementDET 与 SpecTT- Tra [43], 进行比较，后者充当通用检测器。在持续学习比较中，我们采用五种策略更新 SpecTTTra，即 不更新、在截至目前见过的所有生成器上进行完全重新训练、在新生成器和人工数据音轨上进行微调、经验回放（为每个更早的生成器添加 50 条存储音轨），以及弹性权重固化 （EWC） [27]。为了测试 IncrementDET 的优势究竟来自 MuQ 编码器还是注册，我们还在相同的冻结 MuQ 特征上更新一个逻辑回归分类头，分别采用重新训练、微调或带回放的微调方式，并使用如 SimpleCIL 中的类均值分类器 [55]。

实现细节。 IncrementDET 的总体假阳性目标 α 为 5% 或 1%。SpecTTTra-α 使用其官方代码和 5 秒配置。完全重新训练进行 50 个训练轮次，其他策略进行 10 个。所有运行均使用一块 RTX 3060，成本按实际运行时间计算，不包括评估。人类参考集 $\mathcal{H}_{\text{ref}}$ 在留一生成器比较中包含全部 $\mathcal{H}_{\text{fit}}$，在持续比较中则是从其音轨中固定选取的 2,000 条；持续比较对每个到达的生成器使用 $N = 300$ 条拟合轨迹。阈值在 $\mathcal{H}_{\text{cal}}$上设定，95% 置信区间由所有测试集和校准集的 1,000 次自助法重采样得出。

5.2. 与通用检测器的比较

5.2.1. 留一生成器比较

依次留出十二个生成器中的每一个，并使用其余十一个生成器训练两个检测器。IncrementDET 使用其登记的十一个针检测留出生成器（第 4.3 节），因此本实验衡量登记的针迁移到新生成器的效果。

表 2报告全部十二轮的结果。Increment- DET 的等错误率在全部十二轮中均低于 SpecTTTra（平均为 13.2% 对 27.4%），除 Mureka 外，差异均显著；其在 5% 假正率下的真正率在 11 轮中更高，其中 9 轮的差异显著。DiffRhythm 1、DiffRhythm 2、Mureka 和 Suno 这四种生成器对两种检测器而言都很难检测。对于它们，IncrementDET 在 5% 假正率下的平均真正率为 15.1%，SpecTTTra 则为 13.6%。

表 2. 不进行更新时的检测（留一生成器法）。5% 和 1% 假正率下的等错误率和真正率，单位为百分比。

	SpecTTTra			IncrementDET		
留出	EER	@5	@1	EER	@5	@1
MusicGen	23.0	58.0	41.5	2.5	99.0	95.5
AudioLDM2	12.0	80.0	63.5	1.1	99.0	99.0
MusicLDM	35.5	15.5	1.5	6.5	90.0	66.0
Mustango	24.0	42.0	22.5	0.5	100.0	99.5
Stable Audio Open	11.6	75.5	49.5	2.2	98.5	96.5
InspireMusic	36.5	13.0	3.5	9.0	80.5	60.5
DiffRhythm 1	47.6	4.5	1.0	31.9	14.5	6.5
DiffRhythm 2	47.6	5.0	1.0	40.6	9.0	2.5
ACE-Step	9.0	83.0	53.5	5.0	93.5	68.0
LeVo	16.0	63.5	40.5	4.6	94.5	82.5
Mureka	35.1	10.5	1.5	33.1	11.5	0.0
Suno	30.5	34.5	11.0	21.9	25.5	9.5
Mean	27.4	40.4	24.2	**13.2**	**68.0**	**57.2**

架构族内的迁移。 当一个留出生成器所属的架构族中，已有一个以相同方式生成音频的生成器（表1）完成登记时，对该留出生成器的检测效果最佳。十二个生成器中有七个在其余十一个生成器中拥有同族成员。对于这七个生成器中的五个，同族成员的针在 60.6–98.5% 的检测音轨上给出了最高分。例外是 DiffRhythm 1 和 DiffRhythm 2，这两个流匹配模型的生成过程有所不同。DiffRhythm 1 联合生成完整歌曲 [39],，而 DiffRhythm 2 则分块连续生成歌曲 [25]。没有已登记同族成员的生成器仍然可以被检测出来。留出的 MusicGen、Stable Audio Open 和 In- spireMusic 达到 2.2–9.0% 等错误率。

针方向的几何结构。针的方向解释了第5.2.1节中的家族迁移。对于 MuQ、MERT [32], wav2vec 2。0 [4], XLS-R [3] 和 EnCodec， [16],我们在人工数据拟合轨迹上对特征进行标准化，针对每个生成器和层训练一根针，并计算每对针之间的余弦相似度。在四种自监督编码器中，同一家族的针比不同家族的针更相似，家族内的平均余弦相似度为 0。22–0。46，而家族间为0。10–0。20；这一差异超过了保持家族规模不变的99。7% 的5，000 种随机重新标记结果。在84 组“家族与其外部生成器”的配对中，相似度越高，迁移等错误率越低，二者呈秩相关。这里的迁移等错误率，是指在某一家族的所有成员上拟合的检测器在该家族外部生成器上的等错误率，斯皮尔曼相关系数为

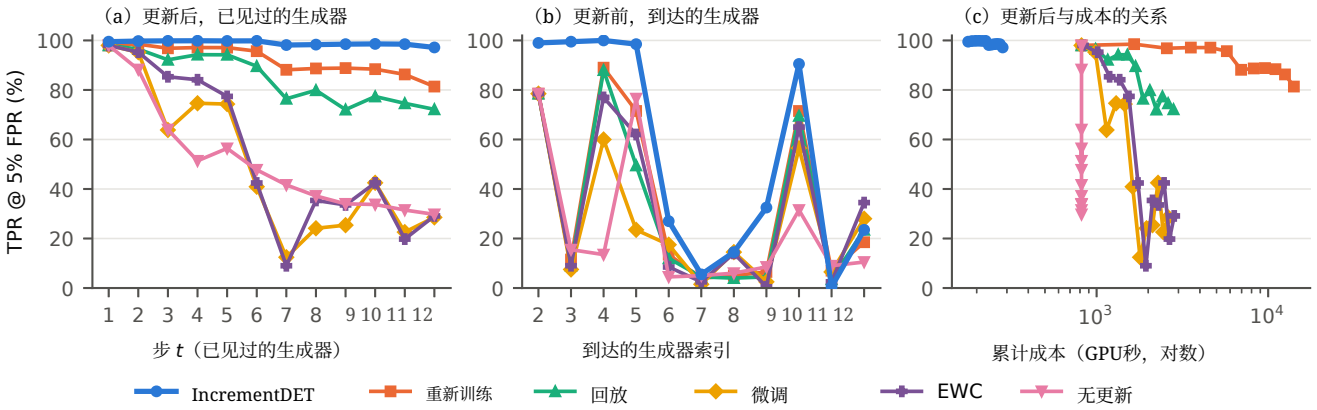


图3。持续比较 ，按表1中的顺序，TPR 为5% 假正率。（a）每次更新后，对截至目前见过的所有生成器取均值。（b） ）每次更新前，对即将到来的生成器。（c）将面板（a）与累积更新成本进行比较。

表 3。持续比较，最后一步，五种顺序上的均值（±标准差）。@5、@1：在 5%、1% 假正率下，POST(T) 作为真正率。Pre：在 $t = 2$、……、T 上，PRE(t) 的均值，作为 5% 假正率下的真正率。存储：保留较早生成器的音轨以进行更新。

方法	存储	@5	@1	Pre
SpecTTTra				
不更新	0	24.0 ± 5.2	14.5	17.3
微调	0	30.0 ± 3.0	17.3	17.8
EWC	0	37.8 ± 6.4	20.5	20.1
回放	550	74.5 ± 2.1	53.1	24.8
重新训练	3，300	81.4 ± 1.0	56.0	27.4
IncrementDET	0	**97.2** ± 0.0	**90.8**	**57.2**

从 -0.77 to -0.92。对于 EnCodec，它只有一层编解码器特征，差异并不显著，相关系数为 -0.47（见补充材料）。因此，纳入某一家族中的一个生成器，也能提升对该家族后续成员的检测能力。

5.2.2. 持续比较

十二个生成器按照表 1 （图 3）中的顺序，以及四种随机顺序（表 3）。

检测与遗忘。按照表1（图3a）的顺序，IncrementDET 在每次注册后都能以 97.2–99.9% 的准确率检测出截至目前已见的生成器。每一步中，它都能将 97.1–100% 的已检测音轨归因于正确的生成器。完全重新训练是 SpecTTTra 的最佳更新策略，但经过十二个步骤后，它只能检测出 81.4% 的生成音轨，而回放、EWC 和微调的检测率分别为 74.5%、37.8% 和 30.0%（表3）。在最后一步之后，IncrementDET 对第一个生成器的检测率在五种顺序上的平均值仍达到 96.4%，而完全重新训练仅为 79.4 %。微调能够学习每个到达的生成器，但会遗忘较早的生成器；EWC 只能延缓这种遗忘。在表1所示顺序的最后一步，IncrementDET 在每个生成器上的等错误率最高为 10.1%，而重新训练在其中五个生成器上的等错误率超过 11%

表4。在相同冻结的 MuQ 特征上使用 IncrementDET 和一个分类器。@5、@1、假正率：持续比较的最后一步，取五种顺序的均值，其中假正率是在 $\mathcal{H}_{\text{test}}$ 目标为 5% 时测量的。新、旧：当一个新生成器携带 25 条音轨到达时，在该生成器和其他十一个生成器上的 5% 假正率对应的真正率，取十二个生成器的均值。存储：为更新而保留的早期生成器的音轨。属性：将检测到的音轨归因于其生成器。

方法	存储	@5	@1	FPR	New	Old	属性
一个分类器							
重新训练	3，300	95.7	77.7	5.2	82.3	97.1	×
回放	550	92.9	73.9	6.0	80.0	96.3	×
微调	0	84.4	66.0	5.6	86.3	96.0	×
类别均值	0	9.3	4.8	5.1	5.0	5.7	✓
IncrementDET	0	**97.2**	**90.8**	**3.5**	**93.2**	**99.2**	✓

十二个。在更新之前，nodetector 达到的真正率为 IncrementDET 注册后达到的真正率（图 3b）。

编码器还是注册？表 4 比较了 IncrementDET 与在相同冻结的 MuQ 特征上使用的一个分类器。无论采用哪种三种更新策略，IncrementDET 的表现都优于该分类器。即使分类器在每一步都使用全部 3,300 条存储音轨重新训练，IncrementDET 在 1% 假正率下检测到的生成音轨仍更多（ 90.8%，而分类器为 77.7 %；差异的 95% 置信区间为 $[6.0, 19.8]$），并且其假阳性率始终低于 5% 的目标值，而每个分类器的假阳性率都超过了该目标。当新生成器仅提供 25 条音轨时，IncrementDET 的检测率达到 93.2%，而分类器为 80.0–86.3%，同时对其他生成器仍保持 99.2% 的检测率。它还会将每条检测到的音轨归因于其生成器，这是一个分类器无法做到的，并且不会存储早期生成器的任何音轨。使用类别均值的分类器（如 SimpleCIL）检测率仅为 9.3%，而 RAWM [53] 和 RegO [10] 的表现并不优于微调（见补充材料）。

成本。 IncrementDET 每次注册耗时 9.2–13.8 秒，其中约 9 秒用于提取 300 条新音轨的特征，不到 1 秒用于训练针，另有一次性成本

为 165秒，用于人类参考集。SpecTT-Tra 的一次更新在重新训练策略下耗时 838–1,585秒，在其他策略下耗时 157–201 秒。经过 12 个步骤，IncrementDET 的累计成本为 284 秒，而重新训练需要 14,118秒，约为其 50 倍（图3c）。

5.3. 消融与捷径分析

注册样本数量。从留一生成器比较的每一轮开始，留出生成器会带有 $N \in \{25, 50, 100, 300, 700\}$条音轨，其中每个较小集合都包含在所有较大的集合中。IncrementDET 使用一个针来注册该生成器；SpecTT-Tra 则在 N 条音轨和 2,000 条人类音轨上进行微调，训练 10 个训练轮次，或通过回放进行更新。所有模型都在持续比较的测试集上进行测试。

表 5. 使用 N 个样本进行注册。在5% 假正率下的真正率（百分比），取十二个留出生成器的均值。New：留出生成器。Old：其余十一个生成器。 $N = 0$ 表示更新前的模型。对于 SpecTTTra，该模型来自单独的推理运行，与表2略有不同。

	IncrementDET		微调		回放	
N	New	Old	New	Old	New	Old
0	**68.0**	**99.4**	40.6	94.8	40.6	94.8
25	**93.2**	**99.2**	84.6	43.1	58.5	91.3
50	**96.7**	**99.2**	89.7	45.6	71.4	92.0
100	**97.3**	**99.2**	93.7	48.4	81.4	91.0
300	**98.4**	**99.2**	96.2	50.6	90.8	90.7
700	**98.8**	**99.2**	97.2	47.5	94.1	89.5

如表5所示，IncrementDET 使用25首音轨时对新生成器的检测率为93.2%，使用700首时为98.8%；对于其他十一个生成器，在每个 N下的检测率均保持在99.2%。十二个生成器中，有十个在使用25首音轨时的检测率至少达到97.5%；例外是 DiffRhythm 1（68.5%）和Suno（57.0%）。对SpecTTTra进行微调也能很好地检测新生成器，但其对其他十一个生成器的检测率从94.8%降至43.1–50.6%，且不会随着N 增大而恢复。回放使对其他生成器的检测率保持在89.5–92.0%，但需要300首音轨才能使对新生成器的检测率达到90.8%。

ˇSid´ak 校正的有效性。我们再次进行持续比较，将每根针的阈值设为目标率，但不使用 ˇSid´ak 校正。在 5% 的目标率下，假阳性率会随针的数量增加：在 $\mathcal{H}_{\text{test}}$ 上从 3.8% 升至 33.6%，在 $\mathcal{H}_{\text{new}}$ 上从 5.5% 升至 41.0%。使用校正后，在 $\mathcal{H}_{\text{test}}$ 上假阳性率维持在 3.2–4.3%，在 $\mathcal{H}_{\text{new}}$ 上维持在 5.5– 7.9%。校正仅使最终均值真正率从 99.8% 降至 97.2%，而几乎全部损失都集中在表现最弱的两根针上：Suno（从 99.5% 降至 82.5%）和 DiffRhythm 1（从 98.5% 降至 85.5%）；其他生成器的降幅最多为 0.5 个百分点。在 MP3编码、重采样和均衡化条件下，IncrementDET 仍优于 SpecTTTra 的所有更新策略（阈值是在劣化人类音轨上设定的；参见补充材料）。

无效针。为检验针检测的是生成行为，还是仅仅检测来源之间的差异，我们将每个人类来源都视为一个生成器。Jamendo、FMA、Pixabay、ccMixter和Internet Archive网络厂牌的无效针等错误率为24.8–41.4%，而生成器针的平均等错误率为0.6%。只有MagnaTagATune因其低比特率MP3编码而被区分出来（等错误率为0.15%），因此人工校准集应包含预期的编码格式。

时长捷径。 五种 FakeMusicCaps 生成器生成的片段时长为 10 秒，而其他所有来源都包含更长的音轨。对于这五种生成器，在五个留出轮次中，仅凭音轨时长就能在输入时长为 120 秒时达到 0% 等错误率；将人类音轨截短至 10 秒并进行零填充后，SpecTTTra 会将其中 100% 标记为生成数据，而未经修改的同一批音轨中这一比例为 0.5%。因此，所有检测器都只对每条音轨未填充的一个 5 秒窗口进行评分。

6. 讨论与局限性

器乐。 我们将数据集限定为器乐，以确保检测不依赖歌声，而歌声深度伪造检测器几乎完全依赖歌声 [9]。因此，对于 Suno 和 Mureka 等商业生成器默认生成的带人声歌曲，其检测效果仍有待测试。无需改变结构，IncrementDET 即可扩展至此类歌曲：将每条音轨分离为人声和伴奏，并在每个部分分别注册一个针；或者使用也能表征歌声的编码器。补充材料还讨论了其他局限性及潜在危害。

扩展至视觉领域。 IncrementDET 只有编码器针对音频进行了专门设计，因此我们的发现也可为生成图像和视频的检测提供参考，因为新生成器也在频繁发布。首先，对于冻结特征，每个生成器配备一个线性分类器，与在所有存储数据上重新训练一个分类器相比，在严格的假阳性率要求下能得到更准确的检测器（表4）。因此，使用冻结特征的图像检测器 [41]可以通过注册而非重新训练来学习新生成器。其次，当检测器为每个生成器添加一个分类器时，除非进行校正，否则其整体假阳性率会随生成器数量增加而上升（第5.3节）。控制该比率的ˇSidak 校正不依赖于模态。第三，我们的基准测试协议会在每次更新前后评估检测器，并记录每次更新的成本，因此也适用于其他模态。

7. 结论

我们定义了持续人工智能生成音乐检测任务，为其引入了基准测试，构建了 EvoMusicFake-17K 数据集，并提出了 IncrementDET，

它将每个新生成器注册为一根针。IncrementDET 仅用 25 条音轨即可学习一个新生成器，在 5% 假正率下达到 93.2% 真正率，同时在其他生成器上的表现保持在 99.2%。在注册十二个生成器后，它对迄今为止见过的生成器达到 97.2% 真正率，而重新训练通用检测器的表现为 81.4%，更新计算量约为后者的五十分之一。因此，检测器无需重新训练，也能跟上新生成器的出现。注册的准确率更高、所需样本更少、成本更低，而且其结构也可应用于其他模态。

大语言模型使用声明

在本研究的整个过程中，我们使用了 Claude（Anthropic）这一大语言模型。在写作方面，它修改了各章节中的段落，这些内容均由作者起草并批准。在研究方面，它帮助检索相关文献、提出了一些分析建议，并审阅了数据处理、特征提取、实验、统计分析和图表相关的大部分代码。作者负责设计研究、决定开展哪些实验，并核实所有结果、参考文献和论述。

参考文献

[1] Andrea Agostinelli、Timo I. Denk、Zal´an Borsos、Jesse En- gel、Mauro Verzetti、Antoine Caillon、Qingqing Huang、Aren Jansen、Adam Roberts、Marco Tagliasacchi、Matt Sharifi、Neil Zeghidour 和 Christian Frank。MusicLM：从文本生成音乐。arXiv 预印本，2023 年。arXiv:2301.11325。3[2] Aref Azizpour、Tai D. Nguyen、Manil Shrestha、Kaidi Xu、Edward Kim 和 Matthew C. Stamm。E3：专家嵌入器集成，利用有限数据使合成图像检测器适应新生成器。发表于 IEEE/CVF计算机视觉与模式识别会议研讨会（CVPRW），第 4334–4344 页，2024 年。3[3] Arun Babu、Changhan Wang、Andros Tjandra、Kushal Lakho- tia、Qiantong Xu、Naman Goyal、Kritika Singh、Patrick von Platen、Yatharth Saraf、Juan Pino、Alexei Baevski、Alexis Conneau 和 Michael Auli。XLS-R：大规模自监督跨语言语音表征学习。发表于 Interspeech，2022 年。arXiv:2111.09296。2、6、14[4] Alexei Baevski、Yuhao Zhou、Abdelrahman Mohamed 和 Michael Auli。wav2vec 2.0：语音表征自监督学习框架。发表于神经信息处理系统进展（NeurIPS），2020 年。2、6、14[5] Arnesh Batra、Dev Sharma、Krish Thukral、Ruhani Bhatia、Naman Batra 和 Aditya Gautam。旋律还是机器：通过双流对比学习检测合成音乐。《机器学习研究汇刊》（TMLR），2025 年。arXiv:2512.00621。2、3[6] Stefano Borz`ı、Oliver Giudice、Filippo Stanco 和 Dario Allegra。合成语音检测研究是否走在正确的方向上？发表于 IEEE/CVF计算机视觉与模式识别会议研讨会（CVPRW），第 71–80 页，2022 年。4

[7] ccMixter。ccMixter。https://ccmixter.org，2025。访问于2026年7月。3, 12[8] Ke Chen、Yusong Wu、Haohe Liu、Marianna Nezhurina、Taylor Berg-Kirkpatrick 和 Shlomo Dubnov。《MusicLDM：利用节拍同步混合策略提升文本生成音乐的新颖性》。载于IEEE声学、语音与信号处理国际会议（ICASSP），第1206–1210页，2024。arXiv:2308.01546。1, 4, 13[9] Xuanjun Chen、Chia-Yu Hu、I-Ming Lin、Yi-Cheng Lin、I-Hsiang Chiu、You Zhang、Sung-Feng Huang、Yi-Hsuan Yang、Haibin Wu 和 Jyh-Shing Roger Jang。《器乐如何助力SingFake检测？》载于 IEEE声学、语音与信号处理国际会议（ICASSP），第19072–19076页，2026。arXiv:2509.14675。8[10] Yujie Chen、Jiangyan Yi、Cunhang Fan、Jianhua Tao、Yong Ren、Siding Zeng、Chu Yuan Zhang、Xinrui Yan、Hao Gu、Jun Xue、Chenglong Wang、Zhao Lv 和 Xiaohui Zhang。《音频深度伪造检测持续学习中的基于区域的优化》。载于 AAAI人工智能会议（AAAI），第23651–23659页，2025。3, 7, 15[11] Luca Comanducci、Paolo Bestagini 和 Stefano Tubaro。《FakeMusicCaps：用于检测和归因文本生成音乐模型所生成合成音乐的数据集》。《成像学杂志》，11(7):242，2025。arXiv:2409.10684。2, 3, 12[12] Jade Copet、Felix Kreuk、Itai Gat、Tal Remez、David Kant、Gabriel Synnaeve、Yossi Adi 和 Alexandre D´efossez。《简单且可控的音乐生成》。载于 神经信息处理系统进展（NeurIPS），2023。arXiv:2306.05284。1, 4, 13[13] Laura Cros Vila、Bob L. T. Sturm、Luca Casini 和 David Dalmazzo。《人工智能音乐军备竞赛：关于人工智能生成音乐的检测》。《国际音乐信息检索学会汇刊（TISMIR）》，8(1):179–194，2025。doi:10.5334/tismir.254。3, 4[14] Deezer。《Deezer：人工智能音乐首次占新上传音乐的50%以上》。https://newsroom-deezer.com/2026/07/ai-music-exceeds-50-percent-daily-uploads-deezer/，2026年7月21日。新闻稿，2026年7月。[15] Micha¨el Defferrard、Kirell Benzi、Pierre Vandergheynst 和 Xavier Bresson。《FMA：用于音乐分析的数据集》。载于 国际音乐信息检索学会会议（ISMIR），2017。arXiv:1612.01840。3, 12[16] Alexandre D´efossez、Jade Copet、Gabriel Synnaeve 和 Yossi Adi。《高保真神经音频压缩》。《机器学习研究汇刊（TMLR）》，2023。arXiv:2210.13438。6, 14[17] Zach Evans、Julian D. Parker、CJ Carr、Zack Zukowski、Josiah Taylor 和 Jordi Pons。《Stable Audio Open》。载于 IEEE声学、语音与信号处理国际会议（ICASSP），2025。arXiv:2407.14358。1, 4, 13[18] Robert M. French。《联结主义网络中的灾难性遗忘》。《认知科学趋势》，3(4):128–135，1999。3[19] Junmin Gong、Sean Zhao、Sen Wang、Shengyuan Xu 和 Joe Guo。《ACE-Step：迈向音乐生成基础模型》。arXiv预印本，2025。arXiv:2506.00045。1, 4, 13

[20] Junmin Gong、Yulin Song、Wenxiao Zhao、Sen Wang、Shengyuan Xu、Jing Guo 和 Xuerui Yang。《ACE-Step 1.5：拓展开源音乐生成的边界》。arXiv预印本，2026。arXiv:2602.00744。4, 13[21] Chaolei Han、Hongsong Wang 和 Jie Gui。《MusicDET：零样本人工智能生成音乐检测》。载于国际机器学习会议（ICML），2026。arXiv:2605.18072。2, 3[22] Yan Han、Zhibin Wen、Yuan Wang、Shuangrun Shao、Xiaobing Li、Yang Xu 和 Wei Li。《超越伪迹：通过音乐内在特征实现可泛化的合成歌曲检测》。arXiv预印本，2026。arXiv:2606.16612。2[23] Internet Archive。Netlabels。https://archive.org/details/netlabels，2025。3, 12[24] Jamendo S.A.。Jamendo。https://www.jamendo.com，2025。访问于2026年7月。3, 12[25] Yuepeng Jiang、Huakang Chen、Ziqian Ning、Jixun Yao、Zerui Han、Di Wu、Meng Meng、Jian Luan、Zhonghua Fu 和 Lei Xie。《DiffRhythm 2：通过块流匹配实现高效、高保真的歌曲生成》。arXiv预印本，2025。arXiv:2510.22950。1, 4, 6, 13[26] Jee-weon Jung、Hee-Soo Heo、Hemlata Tak、Hye-jin Shim、Joon Son Chung、Bong-Jin Lee、Ha-Jin Yu 和 Nicholas Evans。《AASIST：利用集成频谱—时序图注意力网络进行音频反欺骗》。载于 IEEE 声学、语音与信号处理国际会议（ICASSP），2022。arXiv:2110.01200。2[27] James Kirkpatrick、Razvan Pascanu、Neil Rabinowitz、Joel Veness、Guillaume Desjardins、Andrei A. Rusu、Kieran Milan、John Quan、Tiago Ramalho、Agnieszka Grabska-Barwinska、Demis Hassabis、Claudia Clopath、Dharshan Kumaran 和 Raia Hadsell。《克服神经网络中的灾难性遗忘》。《美国国家科学院院刊》，114(13):3521–3526，2017。3, 6[28] Kunlun Tech（Skywork AI）。Mureka。https://www.mureka.ai，2026。模型 mureka-9，通过官方API于2026年7月访问。1, 4, 13[29] Edith Law、Kris West、Michael Mandel、Mert Bay 和 J. Stephen Downie。《利用游戏评估算法：以音乐标注为例》。载于国际音乐信息检索学会会议（ISMIR），2009。3, 12[30] Shun Lei、Yaoxun Xu、Zhiwei Lin、Huaicheng Zhang、Wei Tan、Hangting Chen、Yixuan Zhang、Chenyu Yang、Haina Zhu、Shuai Wang、Zhiyong Wu 和 Dong Yu。《LeVo：通过多偏好对齐实现高质量歌曲生成》。载于神经信息处理系统进展（NeurIPS），2025。arXiv:2506.07520。1, 4, 13[31] Chuqiao Li、Zhiwu Huang、Danda Pani Paudel、Yabin Wang、Mohamad Shahbazi、Xiaopeng Hong 和 Luc Van Gool。《持续深度伪造检测基准：数据集、方法与要点》。载于 IEEE/CVF冬季计算机视觉应用会议（WACV），第1339–1349页，2023。3[32] Yizhi Li、Ruibin Yuan、Ge Zhang、Yinghao Ma、Xingran Chen、Hanzhi Yin、Chenghao Xiao、Chenghua Lin、Anton Ragni、Emmanouil Benetos、Norbert Gyenge、Roger Dannenberg、Ruibo Liu、Wenhu Chen、Gus Xia、Yemin Shi、Wenhao Huang、Zili Wang、Yike Guo 和 Jie Fu。MERT：采用大规模自监督训练的声学音乐理解模型。载于国际学习表征会议（ICLR），2024年。arXiv:2306.00107。2、6、14[33] Haohe Liu、Yi Yuan、Xubo Liu、Xinhao Mei、Qiuqiang Kong、Qiao Tian、Yuping Wang、Wenwu Wang、Yuxuan Wang 和 Mark D. Plumbley。AudioLDM 2：通过自监督预训练学习整体音频生成。IEEE/ACM音频、语音与语言处理汇刊，32：2871–2883，2024年。arXiv:2308.05734。1、4、13[34] Michael McCloskey 和 Neal J. Cohen。联结主义网络中的灾难性干扰：序列学习问题。载于《学习与动机心理学》，第109–165页。Academic Press，1989年。2、3[35] Mark D. McDonnell、Dong Gong、Amin Parvaneh、Ehsan Abbasnejad 和 Anton van den Hengel。RanPAC：用于持续学习的随机投影与预训练模型。载于神经信息处理系统进展（NeurIPS），第12022–12053页，2023年。3[36] Jan Melechovsky、Zixun Guo、Deepanway Ghosal、Navonil Majumder、Dorien Herremans 和 Soujanya Poria。Mustango：迈向可控的文本生成音乐。载于北美计算语言学协会会议（NAACL），第8293–8316页，2024年。arXiv:2311.08355。1、4、13[37] Nicolas M. M¨uller、Franziska Dieckmann、Pavel Czempin、Roman Canals、Konstantin B¨ottinger 和 Jennifer Williams。语音是银，静音是金：经过ASVspoof训练的模型究竟学到了什么？载于2021年自动说话人验证与欺骗对策挑战赛（ASVspoof）论文集，第55–60页，2021年。arXiv:2106.12914。4[38] Eric Nalisnick、Akihiro Matsukawa、Yee Whye Teh、Dilan Gorur 和 Balaji Lakshminarayanan。深度生成模型知道自己不知道什么吗？载于国际学习表征会议（ICLR），2019年。2[39] Ziqian Ning、Huakang Chen、Yuepeng Jiang、Chunbo Hao、Guobin Ma、Shuai Wang、Jixun Yao 和 Lei Xie。DiffRhythm：利用潜在扩散实现极快且极其简单的端到端全长歌曲生成。arXiv预印本，2025年。arXiv:2503.01183。1、4、6、13[40] Heewon Oh。ArtifactNet：通过取证残差物理学检测人工智能生成的音乐。arXiv预印本，2026年。arXiv:2604.16254。3[41] Utkarsh Ojha、Yuheng Li 和 Yong Jae Lee。迈向能够跨生成模型泛化的通用伪造图像检测器。载于IEEE/CVF计算机视觉与模式识别会议（CVPR），第24480–24489页，2023年。8[42] Pixabay。Pixabay Music。https://pixabay.com/music，2025年。访问日期：2026年7月。3、12[43]Md Awsafur Rahman、Zaber Ibn Abdul Hakim、Najibul Haque Sarker、Bishmoy Paul 和 Shaikh Anowarul Fattah。SONICS：合成的还是非合成的——识别伪造歌曲。载于国际学习表征会议（ICLR），2025年。arXiv:2408.14080。2、3、6、12[44] David Rolnick、Arun Ahuja、Jonathan Schwarz、Timothy Lillicrap 和 Gregory Wayne。用于持续

学习的经验回放。载于神经信息处理系统进展（NeurIPS），2019年。arXiv:1811.11682。3[45] Zbynˇek ˇSid´ak。多元正态分布均值的矩形置信区域。美国统计学会期刊，62(318)：626–633，1967年。5[46] Stefan Smeu、Dragos-Alexandru Boldisor、Dan Oneata 和 Elisabeta Oneata。利用无监督学习规避视听深度伪造检测数据集中的捷径。载于IEEE/CVF计算机视觉与模式识别会议（CVPR），第18815–18825页，2025年。arXiv:2412.00175。4[47] Suno, Inc.。Suno。https://suno.com，2025年。访问日期：2026年5月。1、3、4、13[48] Hemlata Tak、Jose Patino、Massimiliano Todisco、Andreas Nautsch、Nicholas Evans 和 Anthony Larcher。采用RawNet2的端到端反欺骗。载于IEEE声学、语音与信号处理国际会议（ICASSP），2021年。arXiv:2011.01108。2[49] Hemlata Tak、Massimiliano Todisco、Xin Wang、Jee-weon Jung、Junichi Yamagishi 和 Nicholas Evans。使用wav2vec 2.0和数据增强进行自动说话人验证欺骗与深度伪造检测。载于说话人与语言识别研讨会（Odyssey），2022年。arXiv:2202.12233。2[50] Ehsan Variani、Xin Lei、Erik McDermott、Ignacio Lopez Moreno 和 Javier Gonzalez-Dominguez。用于小型化文本相关说话人验证的深度神经网络。载于IEEE声学、语音与信号处理国际会议（ICASSP），第4052–4056页，2014年。3[51] Yusong Wu、Ke Chen、Tianyu Zhang、Yuchen Hui、Taylor Berg-Kirkpatrick 和 Shlomo Dubnov。大规模对比式语言—音频预训练：结合特征融合与关键词到字幕增强。载于IEEE声学、语音与信号处理国际会议（ICASSP），2023年。arXiv:2211.06687。3[52] Chong Zhang、Yukun Ma、Qian Chen、Wen Wang、Shengkui Zhao、Zexu Pan、Hao Wang、Chongjia Ni、Trung Hieu Nguyen、Kun Zhou、Yidi Jiang、Chaohong Tan、Zhifu Gao、Zhihao Du 和 Bin Ma。InspireMusic：融合超分辨率与大语言模型，实现高保真长篇音乐生成。arXiv预印本，2025年。arXiv:2503.00084。1、4、13[53] Xiaohui Zhang、Jiangyan Yi、Jianhua Tao、Chenglong Wang 和 Chu Yuan Zhang。你还记得吗？克服伪造音频检测中的灾难性遗忘。载于国际机器学习会议（ICML），第41819–41831页，2023年。3、7、15[54] Xiaohui Zhang、Jiangyan Yi 和 Jianhua Tao。迈向稳健的音频深度伪造检测：用于持续学习的演进基准。arXiv预印本，2024年。arXiv:2405.08596。3[55] Da-Wei Zhou、Zi-Wen Cai、Han-Jia Ye、De-Chuan Zhan 和 Ziwei Liu。重新审视采用预训练模型的类别增量学习：泛化能力与适应性就是你所需要的一切。国际计算机视觉期刊，133(3)：1012–1032，2025年。arXiv:2303.07338。3、6[56] Haina Zhu、Yizhi Zhou、Hangting Chen、Jianwei Yu、Ziyang Ma、Rongzhi Gu、Yi Luo、Wei Tan 和 Xie Chen。MuQ：采用梅尔残差向量量化的自监督音乐表征学习。*IEEE* 音频、语音与语言处理汇刊，33:3653–3664，2025。arXiv:2501.01108。5，14

补充材料

本补充材料通过进一步的细节和结果，支持主论文中的论断。第A 节记录了数据，并说明我们如何对其进行审计，以从人工数据中移除人工智能生成音轨，并防止模型根据音轨时长进行捷径学习。第B 节表明，IncrementDET 在未见生成器上的优势对于两种数据来源均成立，且不依赖于其配置。第C 节通过针的方向解释了这一优势。第D节表明，IncrementDET 在每一种生成器顺序下均处于领先地位，并且优于用于语音深度伪造检测的持续学习方法。第E 节表明，所报告的优势具有统计显著性，并进一步报告了消融实验和对退化的鲁棒性。第F节讨论了局限性与未来工作。

A. 数据

本节记录第 3.2 节数据的来源与许可证，并介绍对数据进行的两项审计。第一项从人类数据中移除人工智能生成音轨，第二项则防止模型根据音轨时长进行捷径学习。

A.1. 生成器

表 S1 列出了各生成器的版本、其音轨的来源、该版本的发布日期以及许可证或使用条款。MusicGen、AudioLDM2、MusicLDM、Mustango 和 Stable AudioOpen 的音轨取自 FakeMusic- Caps [11], , 该数据集依据知识共享署名-非商业性使用 4.0 许可协议发布。其余七个生成器的音轨由我们自行生成。DiffRhythm 2 没有官方的纯器乐模式，因此其歌词文件仅包含结构标签，不含歌词行，风格提示词以“纯器乐，无人声”结尾。我们试听了部分输出，确认其中没有人声。

A.2. 人类创作音乐

Jamendo [24] 音轨通过 Jamendo API 获取，每首音轨均采用其艺术家选择的知识共享许可协议进行分发。FMA 音轨来自FMA 大型FMA 数据集的子集 [15], , 该子集的元数据依据知识共享署名 4.0 进行分发，音频则采用每位艺术家选择的许可证进行分发。MagnaTa-gATune 片段 [29] 来自 Magnatune，并依据知识共享署名-非商业性使用-相同方式共享 3.0 许可证进行分发。Pixabay [42]音轨依据 Pixabay 内容许可协议进行分发。 $\mathcal{H}_{\text{new}}$ 音轨来自 ccMixter [7] 和 Internet Archive 网络厂牌合集 [23], , 我们仅使用采用知识共享许可协议的内容。表S2 列出了各来源的音轨数量。

一些平台接受生成音乐上传内容，因此人类来源中可能包含人工智能生成音轨，而这些音轨会抬高针对人类创作音乐设定的阈值。在一次早期审计中，我们的各生成器针标记出一组异常的 Jamendo 音轨。通过音轨的发布日期和模板化标题、上传账户的注册日期，以及同一账户上传的其他内容，我们识别出一组上传账户，其音轨很可能是生成的。在该次审计中，按账户移除其中 32 首音轨后，Suno 针在 5% 假阳性率下的真正率从 47.3% 提高至 84.9%。这种影响集中在人类数据得分分布的上尾部，也就是设置严格阈值的区域，因此仅凭等错误率无法体现这一影响。最终数据集除主论文中所述的筛选外，还排除了本次审计中识别出的 53 首音轨。因此，筛选人类数据对于实现低假阳性率下的检测是必要的。

A.3. 时长偏差与固定长度评估

AI 生成数据与人工数据在音轨时长分布上的差异可能引入基于时长的捷径，使检测器无需学习有意义的声学特征即可区分这两个类别。这种偏差在输入窗口较长时可能尤为明显，因为较短的音轨会以静音进行填充，从而间接暴露其原始时长。

为量化这一影响，我们使用 $\min(L,\ 120\text{s})$ 训练仅时长分类器，其中 L 表示原始音轨时长。在留一生成器评估中，该分类器在 FakeMusicCaps 的五轮测试集上均达到 0% 等错误率，并且在十二轮中的八轮里等错误率最高仅为 7。3%，尽管它完全无法访问音频内容。这些结果表明，仅凭时长就能提供强有力的判别线索，可能导致检测性能虚高。

为消除这一捷径，我们在固定长度的5秒窗口上评估所有检测器，且不进行填充（第 5.3节）。根据官方 SpecTTTra 协议 [43], , 测试窗口起始于 $\lfloor 3\ (L-5\text{秒})\ /4 \rfloor$。该设置不会使 SpecTTTra 处于不利地位：当留出 Suno 时，其等错误率从使用38。1% 的 120秒输入降至使用30。5% 的 5秒输入。因此，固定长度评估消除了时长线索，同时能够更公平地比较声学检测能力。

A.4. 数据和人类参与者的使用

我们仅使用所有音频来计算特征，并训练和评估用于非商业研究的检测器。我们仅重新分发自己生成且生成器条款允许分发的音频，不分发任何人工数据音频。本研究未招募人类参与者，我们也不将艺术家姓名或其他个人元数据作为模型输入。

表 S1. 按到达顺序排列的生成器。发布：我们所用版本的公开发布时间（年和月），可能晚于该生成器的首次发布。FMC：FakeMusicCaps。SAI CL：Stability AI 社区许可证。

生成器	版本	来源	发布	许可证或条款
MusicGen [12]	中型	FMC	2023-06	代码 MIT，权重 CC BY-NC 4.0
AudioLDM2 [33]	音乐检查点	FMC	2023-08	CC BY-NC-SA 4.0
MusicLDM [8]	–	FMC	2023-08	CC BY-NC-SA 4.0
Mustango [36]	–	FMC	2023-11	代码 MIT，权重 Apache 2.0
Stable Audio Open [17]	1.0	FMC	2024-06	Stability AI 社区许可证
InspireMusic [52]	1.5B-Long	Ours	2025-01	代码 Apache 2.0
DiffRhythm 1 [39]	v1.2	Ours	2025-05	Apache 2.0；变分自编码器：SAI CL
DiffRhythm 2 [25]	–	Ours	2025年10月	Apache 2.0
ACE-Step [19, 20]	v1.5	Ours	2026年1月	MIT
LeVo [30]	SongGeneration-large	Ours	2025年10月	仅限研究和教育用途
Mureka [28]	mureka-9	Ours	2026年4月	Mureka API服务条款
Suno [47]	v5 和 v4.5+	Ours	2025年9月、2025年7月	Suno服务条款

表 S2。数据划分。生成数据：每个生成器的音轨数。人工数据：每个来源的音轨数。Jam.：Jamendo；Pix.：Pixabay；MTAT：MagnaTa-gATune。留出集 $\mathcal{H}_{\text{new}}$ 来自 ccMixter（365）和 Internet Archive 网标合集（450）。

	生成数据	人工数据				
划分	每个生成数据	Jam。	FMA	Pix.	MTAT	总计
拟合（$\mathcal{H}_{\text{fit}}$）	800	4,125	2,440	2,000	235	8,800
模型选择	–	1,159	686	296	59	2,200
校准（$\mathcal{H}_{\text{cal}}$）	–	1,053	623	270	54	2,000
测试集（$\mathcal{H}_{\text{test}}$）	200	933	553	460	54	2,000
留出（$\mathcal{H}_{\text{new}}$）	–	其他来源				815

B. 关于未见生成器的更多结果

本节扩展了第5.2.1节中的留一生成器比较。

B.1. 两个数据来源上的结果

无论是在取自 FakeMusicCaps 的生成器上，还是在我们自行生成音轨的生成器上，IncrementDET 的平均等错误率都低于 SpecTTTra。在 FakeMusicCaps 的五轮实验中，IncrementDET 的平均等错误率为 2.6%，SpecTTTra 为 21.2%。在另外七轮中，二者分别为 20.9% 和 31.8%。因此，IncrementDET 的优势在这两组中均成立，并不限于 FakeMusicCaps 的生成器。

B.2. 训练数据量

检测未见生成器也只需要少量训练数据。使用所有拟合音轨时，IncrementDET 在十二轮中的平均等错误率达到 13.2%，在假正率为 5% 时真正率达到 68.0%。即使每个生成器仅使用 50 条音轨和 2,000 条人类音轨，这些针仍达到 14.1% 的等错误率和 65.8% 的真正率。

B.3. 针配置

我们以两种方式调整 IncrementDET 的配置，发现针的最简单配置已经足够。将每根针的训练数据从每条轨道一个窗口增加到五个窗口，等错误率为 14.3%，真正率为 67.4%；留出的人类来源上的假阳性率则从 5.6% 上升到 8.6%。添加一个非典型人类音乐检测器，用于标记经 PCA 白化后远离人类拟合音轨的音轨，在维度为 32 至 256 时，等错误率为 14.6–14.9%。

B.4. 新的人类来源上的假阳性

IncrementDET 和 SpecTTTra 在 $\mathcal{H}_{\text{new}}$ 上的假阳性率都超过了 5% 的目标值。不过，IncrementDET 的超出幅度较小，平均假阳性率为 5.6%，而 SpecTTTra 为 7.4%。在 $\mathcal{H}_{\text{test}}$ 上，IncrementDET 的假阳性率在每一轮中均为 4.0–4.7%，因此对于 $\mathcal{H}_{\text{cal}}$ 的来源而言，该目标得以满足。$\mathcal{H}_{\text{new}}$ 的超出则来自其两个来源中的一个。IncrementDET 在 ccMixter 上的假阳性率为 4.5%，在 Internet Archive网络厂牌上为 6.6%；SpecTTTra 的模型在持续比较中也呈现出相同的模式。因此，超出是由不同于 $\mathcal{H}_{\text{cal}}$ 的来源造成的，这意味着平台应使用其自身目录中的音轨来校准检测器。

C. 生成器族内的迁移

本节给出了第 5.2.1 节中关于针方向几何结构段落背后的测量结果。同一架构族的生成器的针指向相似的方向，而这种相似性可以预测检测器迁移到未见过的生成器时的效果。这一发现从实证角度解释了 IncrementDET 的跨生成器泛化能力，并表明了其潜力：纳入某个家族中的一个生成器，也能让检测器为该家族后续成员做好准备。

表 S3。针方向的几何特性：家族内和家族间的余弦相似度（各层均值），以及相似度与迁移等错误率之间的 Spearman ρ。

编码器	在……之内	跨越	Gap	Spearman ρ
MuQ [56]	0.460	0.202	+0.26	−0.92
MERT [32]	0.294	0.109	+0.19	−0.88
wav2vec 2.0 [4]	0.220	0.098	+0.12	−0.77
XLS-R [3]	0.243	0.117	+0.13	−0.92
EnCodec [16]	0.123	0.079	+0.04	−0.47

C.1. 针方向的相似性

相似的针方向可以预测四种自监督编码器中的迁移效果。对于五种编码器，表 S3 报告了家族内部和家族之间针方向的余弦相似度，以及这种相似度与迁移等错误率之间的斯皮尔曼相关系数；后者是在 84 对家族与其外部生成器之间的组合上计算的。在四种自监督编码器中，家族内部的相似度比家族之间的相似度高出 0。12–0。26；相似度越高，迁移等错误率越低（ρ 从 -0.77 到 -0.92）。这一差距超过了保持家族规模不变的 99。7% 的 5,000 次生成器随机重新标记结果。对于 EnCodec 这一只有一个特征层的编解码器，差距为0。04 且不显著，并且 ρ 为 -0.47。因此，这一发现并非 MuQ 所特有，而是适用于自监督编码器。

C.2. 生成器聚类

我们还仅根据十二个生成器的针方向进行聚类，不使用其家族标签。对于每个编码器，我们对各层的余弦相似度取平均值，并基于 1 减去该相似度所得的距离，采用平均连接法划分为八组，每个家族对应一组；随后通过调整兰德指数（ARI）比较这些分组与各家族。MERT 和 wav2vec 2.0 的分组与家族高度一致（ARI 0.78, $p < 0.001$），MuQ 的一致性较弱（0.33, $p = 0.043$），而 XLS-R 和 EnCodec 的结果则未达到显著水平。这里的 p 表示在 5,000 次随机重新标记中，ARI 至少同样大的比例。一些分组会在不同编码器中重复出现。ACE-Step 和 LeVo 在 MERT、wav2vec 2.0 和 EnCodec 中属于同一组。AudioLDM2、MusicLDM 和 Mustango 在 MERT 和 wav2vec 2.0 中构成一组，在 MuQ 中则属于同一个更大的组。在 MuQ、wav2vec 2.0 和 XLS-R 中，Suno 与 Mureka 彼此是最近的方向；在全部五种编码器中，Suno 都是 Mureka 最近的方向。DiffRhythm 1 和 DiffRhythm 2 从未被划入同一组，这与留一生成器比较中二者迁移能力较弱的结果一致。

D. 关于持续学习基准的更多结果

本节扩展了第5.2.2节中的持续比较。

D.1. 生成器顺序

IncrementDET 在每种生成器顺序下都表现最佳。在持续比较中，十二个生成器依次逐个加入，主论文采用表1中的顺序。由于更新策略的结果可能取决于哪些生成器先加入，我们使用另外四种加入顺序重复了持续比较。每种顺序都是对十二个生成器进行随机打乱，并使用固定的随机种子生成，以确保结果可复现。所有顺序使用相同的拟合、回放和人类音轨，因此只有加入顺序不同。主论文的表3报告了这五种顺序的均值，表S4 则给出了每种顺序最后一步后的真正率。在每种顺序下，IncrementDET 的真正率都最高，其次是重新训练和回放。
IncrementDET 的结果在所有顺序中都相同，因为每根针都是独立于其他针进行训练的。微调、RegO、EWC 和不更新的排名则会因顺序而异。

表 S4. 所有十二个生成器在最后一步后的真正率，假正率为5%，按生成器顺序排列。R：表 1中的顺序。1–4：随机顺序。

检测器	R	1	2	3	4	Mean
SpecTTTra，无更新	29.8	25.6	14.8	22.5	27.5	24.0
SpecTTTra，微调	28.5	27.7	35.9	29.7	28.2	30.0
SpecTTTra，RegO	29.5	27.3	31.6	34.0	26.0	29.7
SpecTTTra，EWC	29.1	38.5	47.2	41.6	32.5	37.8
SpecTTTra，回放	72.2	76.5	76.5	75.4	71.9	74.5
SpecTTTra，重新训练	81.4	81.7	79.9	81.2	83.0	81.4
IncrementDET	**97.2**	**97.2**	**97.2**	**97.2**	**97.2**	**97.2**

D.2. 更新前检测

没有任何更新策略能让 SpecTTTra 为下一个生成器做好准备。PRE 衡量检测器在尚未使用某个生成器的任何音轨进行更新之前，对即将到来的生成器进行检测的能力（第 3.1 节）。因此，较高的 PRE 意味着检测器已经能够迁移到新生成器。对五种顺序取均值后，IncrementDET 的 PRE 为 57.2%；相比之下，SpecTTTra 在重新训练下为 27.4%，在回放下为 24.8%，在 EWC 下为 20.1%，在微调下为 17.8%，而不进行更新时为 17.3%。因此，在更多生成器上更新 SpecTTTra，并不能让它为下一个生成器做好准备。主论文图3b按表1中的顺序展示了 PRE。对于 SpecTTTra 而言，这一顺序在五种顺序中最有利；对于 IncrementDET 而言则最不利。即使在这一顺序下，SpecTTTra 在重新训练下也仅达到 34.2%，在 EWC 下达到 32.2%，在回放下达到 31.5%，在微调下达到 27.0%，不进行更新时达到 23.5%；而 IncrementDET 达到 53.7%，这是它在五种顺序中的最低值。

D.3. 与持续学习方法的比较

IncrementDET 的表现优于 RAWM [53]和 RegO [10]，这两种专为语音深度伪造检测设计的持续学习方法。此前尚未针对生成音乐开展持续检测研究（第 2 节），因此这些方法是与我们任务最接近的早期工作。这两种方法都只限制检测器权重在每次更新期间的变化方式，因此不依赖语音，可直接用于生成音乐检测器，无需任何改动。持续比较的最后一步结束后，基于 MuQ 特征的 RegO 在假正率为 5% 时，真正率达到 83.9%；在假正率为 1% 时，达到 64.3%。RAWM 的相应数值为 44.7% 和 31.8%，而 IncrementDET 为 97.2% 和 90.8%（表 S5）。我们将这两种方法应用于 MuQ 特征上的线性分类器，并将 RegO 也应用于 SpecTTTra。我们依据谱范数而非弗罗贝尼乌斯范数对 RAWM 的投影矩阵进行归一化，因为使用官方代码中的弗罗贝尼乌斯范数时，投影在每一步都会缩小约 45 倍。在我们的基准测试中，这两种方法的表现都不如直接微调。RegO 在两种检测器上的表现与微调相当。RAWM 能将第一个生成器的识别率保持在 99% 或以上，但对新生成器的学习效果不佳；这可能是因为它的投影抑制了由人工音轨张成的方向，而这些方向在每一步都相同。

表S5。用于语音深度伪造检测的持续学习方法，持续比较的最后一步，五个顺序上的均值（± 标准差）。没有任何方法存储早期生成器的音轨。

检测器	更新	@5	@1	Pre
MuQ +线性分类器 微调		85.6 ± 0.4	67.4	59.7
	RegO	83.9 ± 1.0	64.3	58.6
	RAWM	44.7 ± 3.5	31.8	40.4
SpecTTTra	微调	30.0 ± 3.0	17.3	17.8
	RegO	29.7 ± 2.9	17.1	17.3
IncrementDET	–	$\mathbf{97.2} \pm 0.0$	**90.8**	57.2

E. 进一步消融与鲁棒性

本节首先表明，主论文报告的 IncrementDET 优势具有统计显著性，随后扩展了第5.3节中的消融实验。

E.1. 统计显著性

IncrementDET 的优势具有统计显著性。我们使用配对自助法估计主要比较结果的不确定性。每次重采样都针对每个生成器的 200 条测试音轨（$\mathcal{H}_{\text{cal}}$、$\mathcal{H}_{\text{test}}$ 和 $\mathcal{H}_{\text{new}}$）独立进行有放回抽样，并在重采样的 $\mathcal{H}_{\text{cal}}$上重新计算阈值，因此区间包含阈值变化所带来的影响。所有检测器都在相同的重采样数据上进行评估，因此两个检测器之间的差异是在相同音轨上测得的。在留一生成器比较中，每一轮的人工数据音轨都相同，并在每次重采样时抽取一次。我们使用 1,000 次重采样，并将2.5 百分位数和97.5 百分位数报告为95% 置信区间。当区间不包含零时，差异具有显著性。表 S6 列出了这些差异。在持续比较、留一生成器比较以及使用 25 条注册音轨时，IncrementDET 相较于 SpecTTTra 的优势具有显著性。在留一生成器比较中，IncrementDET 的平均等错误率低14.1 个百分点（95% CI $[12.9, 15.3]$）。按轮次来看，除 Mureka 外，每一轮的等错误率差异都具有显著性（-2.0，$[-5.6, 3.0]$），并且在 5% 假正率下，除 DiffRhythm 2、Mureka 和 Suno 外，每一轮的真正率差异都具有显著性。与一个在 MuQ 特征上重新训练的分类器相比，在 1% 假正率下的优势为 13.0 个百分点，同样具有显著性。

表 S6. IncrementDET 与其他检测器之间的差异（95% 自助法置信区间，单位为百分点）。持续：表 1所示顺序的最后一步。留出：留一生成器比较十二轮的均值。 N：新生成器登记 N 条音轨后的十二轮均值（表 5）。ST：SpecTTTra。MuQ：基于 MuQ 特征的单个线性分类器（表4）。

实验与对比对象		TPR@5	TPR@1
持续	ST再训练	$+15.8$ $[13.2, 18.2]$	$+33.3$ $[27.3, 42.3]$
	ST回放	$+25.0$ $[22.1, 28.4]$	$+41.8$ $[34.4, 54.4]$
	MuQ再训练	$+1.5$ $[-0.2, 2.9]$	$+13.0$ $[6.0, 19.8]$
留出	ST	$+27.5$ $[24.6, 30.5]$	$+33.0$ $[28.1, 37.5]$
$N = 25$	微调	$+8.7$ $[6.0, 11.1]$	–
$N = 700$	微调	$+1.6$ $[-0.3, 2.4]$	–

E.2. 假阳性控制

每登记一个针，都会引入额外的假阳性，因此控制整体假正率至关重要。若不进行校正，在5%的目标下，经过十二个生成器， $\mathcal{H}_{\text{test}}$ 上的假阳性率会从3.8%升至33.6%（第5.3节）。采用 Sid´ak 校正后，持续比较最后一步结束时， $\mathcal{H}_{\text{test}}$上的假阳性率在5%的目标下为3.45%，在1%的目标下为0.95%，因此该校正在 $\mathcal{H}_{\text{cal}}$的来源上满足这两个目标。我们还将该校正与另一种规则进行比较：在第4.3节的 $\mathcal{H}_{\text{cal}}$上，取最大标准化分数 $z^{\max}(x)$ 的 $(1-\alpha)$ 分位数作为单一阈值。这个单一阈值也满足两个目标（4.5%和0.55%）。我们倾向于采用 Sid´ak 校正，因为它在更严格的1%目标下取得了更高的真正率（90.8%对88.3%），而在5%目标下性能相近（97.2%对97.7%）。此外，它支持模块化注册：每个针的阈值都可以仅利用其存储的校准分数和针的总数进行更新，无需重新计算联合分数分布。

E.3. 人类参考数据的多样性

IncrementDET 的拟合集和校准集都应混合多个不同的人类来源。我们使用来自单一来源（Jamendo 或 FMA）而非混合数据集的人类音轨，训练并校准持续比较最后一步的针（表 S7）。要使检测器达到假阳性目标，校准集必须混合不同来源。如果校准集仅来自单一来源，$\mathcal{H}_{\text{test}}$上的假阳性率会超过5%的目标（6.9%和11.2%），在 MagnaTagATune 上则达到70.4–74.1%。使用混合校准集时，$\mathcal{H}_{\text{test}}$ 上的假阳性率保持在目标范围内（4.3%和3.8%），而5%假正率下的平均真正率变化不大（97.1%和96.0%，相比之下为97.2%）。混合拟合集还能进一步保护原本可能缺少相关来源的情况。在54条 MagnaTagATune 测试音轨上，使用单一来源拟合会使假阳性率从0.0%升至16.7%和24.1%。因此，IncrementDET 在两个数据集中都使用来自四个来源的人类音轨。

表 S7. 用于训练和校准针的数据（持续比较的最后一步）。TPR 和等错误率为十二个生成器的均值。在 $\mathcal{H}_{\text{test}}$（所有来源和仅 MagnaTagATune）以及 $\mathcal{H}_{\text{new}}$ 上，以 5% 为目标的假正率。所有数值均以百分比表示。

拟合	校准	@5	@1	等错误率	假正率	MTAT	留出
混合	混合	97.2	90.8	1.9	3.5	0.0	6.9
Jamendo	混合	97.1	89.9	2.5	4.3	16.7	5.8
FMA	混合	96.0	87.5	2.8	3.8	24.1	2.9
Jamendo	Jamendo	96.4	89.4	2.8	6.9	70.4	7.7
FMA	FMA	98.9	93.5	2.8	11.2	74.1	8.7

E.4. 无效针

为测试一根针能否区分任意两组音轨，我们将每个人类来源视为一个生成器（第 5.3 节中的无效针）。我们使用某一来源的 300 条音轨与其他来源的人类负样本进行拟合，在 $\mathcal{H}_{\text{cal}}$（不包含该来源）的数据上进行校准，并使用该来源的留出音轨与 $\mathcal{H}_{\text{test}}$ 其他来源的音轨进行测试。Jamendo、FMA 和 Pixabay 的无效针达到的等错误率分别为 41.4%、36.0% 和 24.8%（五次抽样的均值）；而两个留出来源 ccMixter 和 Internet Archive网络厂牌的无效针（每个来源 200 条拟合音轨）达到37.6% 和 30.8%。在相同协议下，十二个生成器针的等错误率最高仅为 4.1 %。因此，与区分不同人类来源相比，一根针能更好地区分生成音乐与人类创作音乐。

E.5. 对退化的鲁棒性

音频退化不会消除生成数据音乐与人工数据音乐之间的区分，但会使基于干净音频设定的阈值失效。在劣化人类音轨上重新校准后，在每种退化条件下，IncrementDET 的真正率均高于 SpecTTTra 的每一种更新策略（表 S9）。我们对每个测试窗口中的生成数据和人工数据都施加四种退化处理：128 kbit/s（以及 64 kbit/s）的 MP3编码、重采样至 11.025 千赫兹后再重采样回来、在 200 Hz 处增益为 +6分贝的低架滤波器，以及在 4 kHz 处增益为 −6 分贝的高架滤波器。表 S8 报告了基于干净校准集设定阈值时的等错误率和假阳性率。均衡化对每个检测器的等错误率影响最多为 1.5 个百分点。MP3编码和重采样使 IncrementDET 的等错误率分别从 1.9% 升至 9.2% 和 14.3%，而重新训练的等错误率则分别从 9.3% 升至 21.5% 和 24.5%，因此在每种退化条件下，IncrementDET 始终保持最低的等错误率。它们还会改变人类音轨的得分，使得基于干净数据设定的阈值不再适用，导致 IncrementDET 和重新训练的 SpecTTTra 的假阳性率升至 25.6–33.9%。如果改为在劣化人类测试音轨的一半上设定阈值，并在另一半上进行评估，假阳性率会恢复至 4.5–5.4%，且在每种更新策略下，IncrementDET 的真正率都高于 SpecTTTra。与重新训练相比，在 MP3编码条件下，其结果为 93.2% 对 49.4%，在重采样条件下为 84.1% 对 44.8%。因此，校准集应包含预计在部署中使用的处理流程。对于 IncrementDET，只需使用经过相同处理的人类音轨进行重新校准，无需重新训练。

表 S8. 持续比较最后一步的模型鲁棒性。上：等错误率（EER），取十二个生成器的均值。下：在 5% 目标下，人工数据测试音轨的假阳性率，阈值设定于干净校准集上。所有数值均以百分比表示。

		MP3（千比特/秒）		重采样	搁架（分贝）	
	干净	128	64	11 千赫兹	+6	−6
EER						
IncrementDET	**1.9**	**9.2**	**9.2**	**14.3**	**2.1**	**2.0**
重新训练	9.3	21.5	21.5	24.5	9.8	9.7
回放	13.5	26.0	25.8	29.6	14.1	14.1
EWC	29.7	30.4	30.4	37.3	30.8	30.3
微调	35.5	38.6	38.6	38.7	36.3	35.3
无更新	31.0	37.9	37.8	36.2	32.2	32.5
人工数据测试音轨上的假正率，干净阈值						
IncrementDET	3.5	33.9	33.3	30.6	4.7	4.0
重新训练	3.8	26.4	25.8	25.6	4.2	3.5
回放	4.4	18.7	18.4	50.0	5.6	3.8
EWC	5.1	0.4	0.4	37.8	6.0	4.2
微调	4.9	0.6	0.6	46.8	6.4	4.2
无更新	4.7	10.4	10.6	27.1	7.4	4.8

F. 局限性与未来工作

本节扩展了第6节的讨论。

F.1. 预处理权衡

在我们的基准测试中，所有音频都会重采样为16千赫兹单声道并进行响度归一化，这会消除平台处理音频的方式所带来的线索

表 S9. 持续比较最后一步模型的鲁棒性。在十二个生成器上的平均真正率（假正率为 5%），阈值根据劣化人类音轨设定。MP3，128 千比特/秒。所有数值均为百分比。

	干净	MP3	重采样	低 +6	高 −6
IncrementDET	**96.6**	**93.2**	**84.1**	**96.6**	**96.8**
重新训练	84.1	49.4	44.8	83.5	82.8
回放	74.2	34.0	32.1	70.5	72.4
EWC	28.8	26.4	18.3	28.2	28.4
微调	28.7	23.3	22.7	28.0	29.0
无更新	30.5	14.0	18.8	27.6	27.4

（第3.2节）。此预处理还会移除高于8千赫兹的内容，因此我们的基准测试不适用于分析全带宽残留信号的取证方法。我们接受这一成本，因为经过一次重采样操作后消失的线索并不是可靠的检测信号。未来的工作可以在基准测试中加入全带宽音轨，其中所有来源都以一种通用格式存储。

F.2. 模型更新与私有模型

对于 IncrementDET，针用于描述某个生成器在特定产品版本中的采样情况。如果商业服务未作通知便更改其模型，针可能会过时，此时必须使用新样本重新注册该生成器。我们尚未测量针过时的速度。未来的研究需要探讨检测器能否自动察觉已登记生成器的输出发生了变化，并判断何时需要重新注册该生成器。由于一次注册只需几秒钟，因此一旦发现变化，重新注册本身成本很低。此外，私有模型仅通过迁移纳入考察。IncrementDET 根据生成器输出的样本为其注册。对于从未发布的私有微调模型，只有在识别出其部分输出后才能进行注册；在此之前，只能依靠现有针向这些模型的迁移能力来检测它们。第C 节的家族迁移结果表明，已登记生成器的针可能迁移至该生成器的微调变体，未来的研究可以对此进行测试。

F.3. 人工数据的维护

每个依赖人类参考数据的检测器，都需要一个由人类创作音乐组成的数据集。正如第 A 节所示，这类数据集会逐渐过时，也可能包含来自开放平台的生成音乐上传内容。足够新、能够匹配当前音乐风格的数据集，往往也是最容易受到近期生成音乐上传内容影响的数据集。因此，维护人类数据是一项需要反复进行的工作，在严格阈值下会产生可衡量的影响。第 A 节的审计表明，针本身就能标记人类数据中的生成音乐上传内容，而将这一筛查流程自动化是未来的工作。

F.4. 潜在危害

假阳性会将人类音乐家的作品标记为生成数据，而采取此类标记的平台可能会影响该音乐家的收入和声誉。因此，平台应以较低的假阳性率运行检测器，这也是我们报告 5% 和 1% （第 3.3 节）假阳性率下真正率的原因。IncrementDET 允许平台直接设置总体假阳性目标 α，在1% 目标下，它仍能检测出十二个登记生成器所生成音轨的 90.8% 。平台应在自有目录上校准检测器，并在采取针对音乐家的行动之前，将阳性输出与进一步证据结合起来。