SYNCTRACK：多轨音乐生成中的节奏稳定性与同步

Hongrui Wang[1]*、FanZhang[1]*, †、Zhiyuan Yu[2], Ziya Zhou[3], Xi Chen[3], Can Yang[1], [4]†、Yang Wang[5] †

[1]香港科技大学数学系 [2]浙江大学 CAD&CG 国家重点实验室 [3]香港科技大学跨学科研究学院 [4]香港科技大学神经系统疾病国家重点实验室 [5]香港大学{hwangfb, zzhoucp, xchengx}@connect.ust.hk, yang.wang@hku.hk{macyang, mafzhang}@ust.hk, zhiyuan.yu@zju.edu.cnhttps://synctrack-v1.github.io

arXiv:2603.01101v1 [cs.SD] 1 Mar 2026

摘要

多轨音乐生成因其精确的混音与再混音能力而受到广泛研究关注。然而，现有模型往往忽视节奏稳定性与同步等关键属性，从而更关注音轨之间的差异，而非其固有特性。本文提出 SyncTrack，一种旨在捕捉多轨音乐独特特征的同步多轨波形音乐生成模型。SyncTrack 采用新颖架构，包含用于在所有音轨上建立共同节奏的轨道共享模块，以及用于适应多样音色与音高范围的轨道专属模块。每个轨道共享模块采用两种跨轨道注意力机制以同步节奏信息，而每个轨道专属模块则利用可学习的乐器先验，以更好地表征音色及其他独特特征。此外，我们通过三项新指标引入节奏一致性，以增强对多轨音乐质量的评估：轨内节奏稳定性（IRS）、跨轨节拍同步（CBS）和跨轨道节拍离散度（CBD）。实验表明，SyncTrack 通过增强节奏一致性，显著提升了多轨音乐质量。

1 引言

原始音频音乐生成领域的进展，使系统能够根据用户指定的流派、情绪、速度和乐器编制等条件生成音乐。当前音乐生成方法所产生的混合波形往往缺乏专业编辑所需的灵活性，因而限制了其对音乐人的实用价值（Cano 等人，2018年）。在音乐制作中，操控单个乐器轨道十分重要，包括混音、重新编排或添加新乐器。近年来，许多研究（Mariani等，2023；Karchkhadze 等人，2025年；Parker 等人，2024年；Yao 等人，2025b）聚焦于多轨音频生成。遗憾的是，多轨音乐生成及其质量评估均未考虑节拍稳定性和同步性。节奏作为音乐中组织时间的主要元素，为旋律与和声的构建提供了基本框架（Abrassart 与 Doras，2022年）。不规则节拍或轨道间失同步造成的时间预测误差，会持续违背听觉预期，从而引发听者的不适（Mas-Herrero 等人，2018年）。

To date，现有方法无法生成具有连贯且稳的令人满意的音乐
定的节奏，这是由于其模型架构与多轨音乐的内在特性不匹配所致 ck

*同等贡献。
†通讯作者。

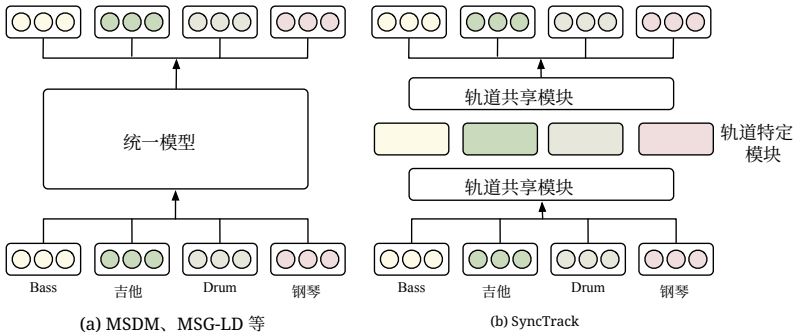


图 1：(a) 先前的方法（Mariani等，2023；Karchkhadze 等人，2025年）利用统一模型学习多轨音频分轨的联合分布。(b) 而我们提出的 SyncTrack 同时包含轨道共享模块和轨道专属模块，用于捕捉音轨之间的共性与特定信息。

音乐。如图 1(a) 所示，MSDM（Mariani 等，2023）和 MSG-LD（Karchkhadze 等人，2025）旨在学习多轨音频分轨的联合分布。它们将这一任务视为多变量时间序列（MTS）或视频生成，其中变量或通道之间存在复杂的交互和巨大的差异（Wang 等人，2022年；Qiu 等人，2025年）。与多变量时间序列和视频生成不同，多轨音乐生成面临着独特的挑战。在多轨音乐生成任务中，我们需要同时为多个音轨生成音乐，每个音轨代表乐曲中独立的层或通道，例如贝斯、鼓等。这一过程受两个核心方面支配：1）节奏指音乐事件在时间上的有序安排。在单个音轨内，节奏必须保持稳定，也就是说，音符起始和节拍必须始终遵循规则的节拍网格。在多个音轨之间，节奏则需要同步，确保不同乐器的事件彼此精准对齐，并与共享的基础脉搏保持一致。2）音色指声音独特的感知特质，使一种乐器有别于另一种乐器，例如贝斯温暖的音质或钢琴明亮的起音。这两类信息有所不同：1）良好的音乐节奏具有同步性，听起来更加和谐，因此节奏信息在各音轨间共享；2）另一方面，音色是每个音轨保留的独特信息，各音轨彼此独立。MSDM 和 MSG-LD 过分强调音轨间的差异，忽略了共有的节奏信息。因此，这两个模型在主观音频质量和节奏一致性方面都未能取得令人满意的结果。

为解决上述挑战，我们提出了一种同步多轨波形音乐生成模型，名为 SyncTrack，其采用适合捕捉数据特征的新颖架构。具体而言，为处理多条音轨的共性与特定信息，SyncTrack 采用统一架构，同时包含轨道共享模块和轨道专属模块，如图 1(b) 所示。在轨道共享模块中，我们设计了两类跨轨道注意力子模块，以进一步捕捉跨轨道节奏稳定性与同步：1）全局跨轨道注意力 对于建模全局稳定性至关重要，可确保所有音轨在整首乐曲中保持一致的节奏框架；2）时间特定跨轨道注意力 对于细粒度同步至关重要，可使各音轨在同一时间位置上的音乐事件对齐。在轨道专属模块中，我们为每条音轨构建可学习的乐器先验。我们将该先验融入轨道专属模块中每条音轨的潜在表示。由此，音色及其他轨道特定特征能够更好地捕捉音轨之间的差异。此外，轨道共享模块与轨道专属模块之间的分离在结构上较为简单，并可嵌入其他潜在音频扩散系统。

此外，我们创新性地提出在多轨音乐质量评估中引入节奏一致性。当前研究仅使用 Fr´echet 音频距离（FAD）（Kilgour等人，2018年）评估多轨音频生成的质量，但该指标无法量化稳定性和同步。FAD通过计算生成音频样本与参考音频样本之间的分布距离来衡量二者的相似性。然而，音频数据具有顺序性且高度结构化的特性，这给准确计算FAD带来了挑战（Yang和Lerch，2020年）。因此，FAD会将音频文件压缩为VGGish嵌入（Kilgour等人，2018年），而这种压缩过度丢失了时序信息，限制了对稳定性和同步的评估能力。本文中，我们

提出了三个新颖的指标：轨内节奏稳定性（IRS）、跨轨道节拍同步（CBS）和跨轨道节拍离散度（CBD），用于评估这些属性。具体而言，***IRS***根据单条音轨节拍间隔的方差，评估其节奏稳定性。***CBS*** 采用滑动容差窗口法，量化节奏对齐的节拍所占比例。***CBD*** 计算对齐节拍之间的时序误差，从而更精细地衡量节拍同步。将这三个指标与FAD结合起来，可以更全面、准确地评估多轨音乐生成的质量。

总而言之，我们的贡献如下：1）我们提出了一种同步多轨波形音乐生成模型，命名为SyncTrack。通过结合轨道共享模块和轨道特定模块，SyncTrack 能够有效处理通用信息和轨道特定信息；2）我们设计了两种跨轨注意力子模块，分别是用于实现全局稳定性的 全局跨轨道注意力 和用于实现细粒度同步的 时间特定跨轨道注意力；3）我们创新性地提出在多轨音乐质量评估中引入节奏一致性。三项指标 ***IRS***、***CBS*** 和***CBD***，可定量评估单条音轨的节奏稳定性以及跨音轨的节奏同步对齐；4）实证结果证实，SyncTrack 的多轨音乐生成性能优于现有最先进的方法。

2 相关工作

2.1 节奏稳定性与同步

音乐节奏作为音乐表达的核心要素，长期以来一直是音乐学和认知科学关注的焦点（Mirka，2004）。节奏稳定性与同步构成了音乐组织的时间骨干，也是多轨音乐创作中有效协作的基础（Abrassart 与 Doras，2022年；Hennig，2014）。听众对节奏稳定性与同步十分敏感；即使是未经训练的听众，也会对音乐中的时间波动或偏差高度敏感（Large 与 Palmer，2002）。不规则节拍或轨道间失同步造成的时间预测误差，会持续违背听觉预期，引发听众不适（Mas-Herrero 等，2018年）。因此，节奏不一致会扰乱听觉预期，降低整体音乐体验，使维持连贯的节奏成为计算音乐研究中的核心挑战。

2.2 音频生成与多轨波形音乐生成

音频生成。音频生成已取得快速进展，从早期的自回归模型（如 WaveNet（Van Den Oord 等，2016）和 SampleRNN（Mehri 等，2016））逐个生成音频采样点，发展到更高层级的离散词元模型和先进的生成方法。VQ-VAE（Razavi 等，2019）等向量量化方法将音频编码为紧凑的代码序列，使 Jukebox 和 MusicGen 等模型能够捕捉长程依赖关系，并实现高效解码。近期的进展采用了扩散模型——例如 DiffWave（Kong 等人，2020b）、WaveGrad（Chen 等，2020），以及 AudioLDM（Liu 等，2023）和 AudioLDM2（Liu 等，2024）等潜空间扩散框架——进一步提升音频保真度、可扩展性和可控性。此后，这些技术已被改编用于音乐领域，例如 Riffusion（Forsgren & Martiros，2022）、Moˆusai（Schneider 等，2024）、MusicLDM（Chen等，2024）和 Stable Audio（Evans 等，2024a；2025；2024b）。

多轨波形音乐生成。 尽管取得了这些进展，大多数模型仍只生成单一混合波形，而无法获取各个乐器轨道。这一局限阻碍了专业音乐人的下游应用，例如混音、自适应编曲和逐轨编辑（Cano 等人，2018年）。多轨音乐生成旨在通过为不同音乐元素生成独立分轨来弥补这一差距。近期方法将多轨音乐视为结构化、相互依存的组成部分：例如，StemGen（Parker 等人，2024年）和 Jen-1 Composer（Yao 等人，2025b）使用 Transformer 和潜在扩散模型，在提示词条件下生成多个标准分轨。其他方法，如多源扩散模型（MSDM）（Mariani等，2023）和 MSG-LD（Karchkhadze 等人，2025年），则将扩散模型扩展为在统一框架内联合建模固定或可变的音轨集合。

然而，当前模型架构与多轨音乐的固有特性并不相容，因为它们没有分别处理轨道特定信息和共有信息。因此，模型往往关注轨道间差异，而忽视共有节奏模式。

2.3 音频相关领域的同步生成

视频到音频生成。这项任务旨在根据视频内容生成音频，同时实现语义一致性和时间同步。近期的视频到音频同步生成方法主要通过节奏和分层视觉条件来建模对齐。Diff-BGM（Li 等人，2024b）和 LORIS（Yu 等人，2023）等模型明确利用动态运动特征来控制生成音频的节奏和时机，确保节拍与屏幕上的动作同步。另一种趋势是利用分层特征将视觉语义与节奏解耦。例如，VidMusician（Li 等人，2024a）使用全局特征表征语义风格，并使用局部特征提取节奏线索；VidMuse（Tian 等人，2025）则采用长短期建模。另一种方法 Mel-QCD（Wang 等人，2025）将音频分解为更易从视频中预测的组成部分，从而使预训练生成器能够实现精确的同步控制。

在这一领域，对生成音频的客观评估主要依据音频质量、跨模态对齐和节奏对齐（Li 等人，2024a）。为研究节奏对齐（又称时间同步），一些研究（Yu 等人，2023年；Li 等人，2024a）使用节拍覆盖分数（BCS）和节拍命中分数（BHS）进行评估。这些指标通常以秒为粒度，粗略估算音频中的节拍。虽然它们适用于简单的音频形式，但无法捕捉多轨音乐生成所必需的精细节奏互动。

伴奏生成。在这一领域，SingSong（Donahue 等人，2023年）等模型会让节奏与输入人声对齐；然而，评估往往依赖主观聆听测试，而非客观指标。这类主观评估需要投入大量人力和时间，因此难以用于迭代开发期间快速、一致地比较模型。

2.4 多轨音乐生成的客观评估指标

在**多轨音乐生成**领域，客观评估指标仍然有限。波形音乐生成中最常用的指标是弗雷歇音频距离（FAD）（Kilgour等人，2018年），用于衡量生成音频样本与参考音频样本之间的分布距离。尽管 FAD 能有效捕捉整体音频质量和相似度，但却无法体现节奏稳定性和跨轨同步等重要音乐特征。这一局限源于 FAD 依赖 VG-Gish 提取的压缩嵌入（Kilgour等人，2018年），而这种嵌入会显著降低时间分辨率。音乐具有顺序性强、结构严谨的特点，因此难以准确计算 FAD（Yang和Lerch，2020年）：音乐感知所必需的时间和节奏信息会在嵌入压缩过程中丢失。KAD（Chung等人，2025年）通过采用无分布假设的无偏估计量改进了 FAD，提供了更高效且与感知对齐的指标。SongEval（Yao 等人，2025年a）通过纳入专业音乐家的多维评分，弥补了客观指标的不足，能够捕捉更广泛的音乐感知维度。然而，这些新提出的指标仍无法充分捕捉多轨音乐中的细粒度节奏问题。

在**符号音乐生成**领域，研究者曾尝试使用专门的指标评估对齐情况（Yu 等人，2022年；Dong 等人，2018年）。然而，这些指标是针对符号表示设计的，由于数据模态不同且原始音频信号更为复杂，因此无法直接应用于多轨波形音频生成。

多轨音乐需要细粒度评估，不仅要评估每条音轨内部的节奏稳定性，还要评估不同音轨之间的同步与互动。为弥补这些不足，我们提出了三项专为评估多轨音乐生成而设计的新指标：IRS、CBS 和 CBD。将这些互补指标与 FAD 结合使用，可以更全面、更符合音乐特性的评估生成质量，同时兼顾音频保真度以及对多轨作品至关重要的复杂节奏关系。

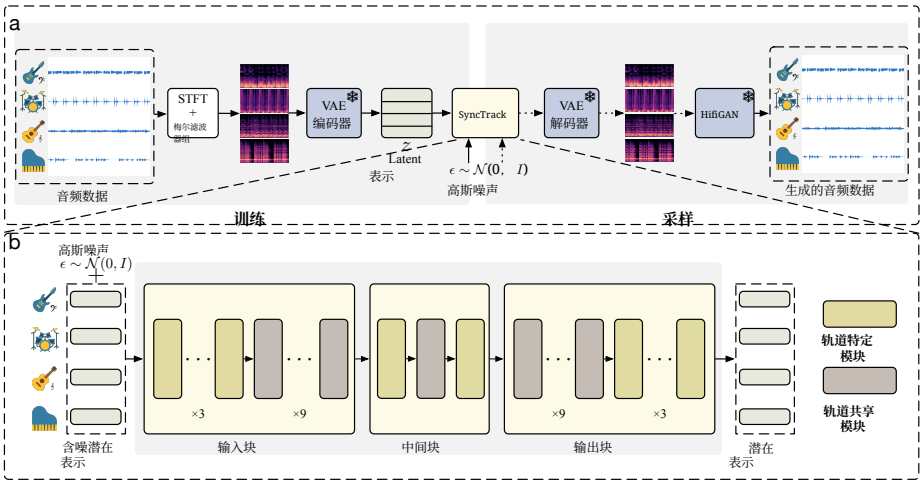


图 2：**a.** SyncTrack 的整体流程。训练流程：我们训练一个四轨潜在扩散模型。每条轨道根据第 l 个信噪比进行扰动。模型经过优化后，可预测添加的噪声 $\epsilon \in \mathcal{N}(0, I)$。更多细节见第 3.1 节。推理流程：测试时，生成四轨潜在表示，然后将其解码为音频数据。**b.** SyncTrack 由输入块、中间块和输出块组成，其中包含轨道专属模块和轨道共享模块。

3 SYNCTRACK：同步多轨音乐生成模式

L

在这项工作中，我们提出了一种基于潜在扩散模型（LDM）框架（Rombach 等人，2022年；Ho 等人，2020年）的新型多轨音乐生成模型。通过学习给定数据集的底层概率分布，SyncTrack 可以生成近似服从同一底层分布的多轨音乐样本。

整体框架。 如图 2a 所示，长度为 T' 的音频数据 $\{x^s\}_{s=1}^S \in \mathbb{R}^{T'}$ 首先经过两个步骤编码为潜在表示：1）使用短时傅里叶变换（STFT）和梅尔滤波器组将音频数据转换为梅尔频谱图；2）使用预训练的变分自编码器（VAE）（Kingma 和 Welling，2013）将梅尔频谱图转换为潜在表示，其形式如下：

$$z^s := \mathrm{VAE}_{\mathrm{enc}}(\mathrm{STFT\&MelFB}(x^s)) \in \mathbb{R}^{C \times T \times F}, \tag{1}$$

其中，T 和 F 分别表示压缩后的时间维度和频率维度，C 表示隐藏维度。

SyncTrack 旨在学习 z^s 的分布，而从易于采样的分布转换到该分布十分困难。受 DDPM（Ho 等人，2020年）的启发，我们设定了一系列由 L 个信噪比等级扰动的数据分布。SyncTrack 通过迭代将扰动数据分布的信噪比从 l 提升至 $l-1$，最终在信噪比达到 1 时生成多轨音乐。提升信噪比的方法是预测扰动数据上添加的噪声。具体而言，给定潜在表示 $\{z^s\}_{s=1}^S$，我们通过将噪声 ϵ 添加到信噪比为第 l 级的 $\{z^s\}_{s=1}^S$ 上，得到扰动数据 $\{z_l^s\}_{s=1}^S$。SyncTrack ϵ_θ 学习预测添加的噪声：

$$\mathcal{L}(\theta) = \mathbb{E}_{\epsilon \sim \mathcal{N}(0,I), \{z^s\}_{s=1}^S, l} \left\| \epsilon - \epsilon_\theta(\{z_l^s\}_{s=1}^S, l) \right\|^2. \tag{2}$$

其中，l 从 $\{1, \cdots, L\}$ 中均匀采样。

在采样阶段，如图 2a 所示，我们可以利用训练好的 SyncTrack 来近似 z^s 的分布。然后，借助 VAE 解码器和 HiFi-GAN 声码器（Kong 等人，2020a），将采样得到的 $\hat{z}^s$ 解码为第 s 轨的音频数据：

$$\hat{x}^s = \mathrm{HiFiGAN}\left(\mathrm{VAE}_{\mathrm{dec}}(\hat{z}^s)\right). \tag{3}$$

如图 2b 所示，针对音轨专属信息和音轨之间的共有信息，SyncTrack 被设计为一种结合了轨道共享模块和轨道专属模块的统一架构。接下来，我们将详细介绍这两类模块。

轨道共享模块。 如图 3a 所示，每个轨道共享模块由残差块（He 等，2016）、轨内注意力、全局跨轨道注意力和时间特定跨轨道注意力组成。需要注意的是，轨内注意力保留了二维 U-Net（Ronneberger 等，2015）中常用的相同架构，并采用相同的参数分别处理每条音轨的内部信息。

1) 全局跨轨道注意力。如图 3c(i) 所示，为了使所有乐器轨道保持全局一致的速度，我们引入了全局跨轨道注意力子模块。我们在时间维度的索引 t 处以及频率维度的索引 f 处对表示 $z^s \in \mathbb{R}^{C\times T\times F}$ 进行切片，得到 $z^s_{t,\ f} \in \mathbb{R}^C$。然后，我们沿时间维度和频率维度收集所有音轨的表示，记为 $z^{1:S}_{1:T,\ 1:F}$，并按如下方式聚合这些信息以得到 $z^s_{t,\ f}$：

$$\mathrm{Attn}_{\text{global_cross}}(z^s_{t,f}) = \mathrm{Attn}(W^{Q_1} z^s_{t,f}, W^{K_1} z^{1:S}_{1:T,1:F}, W^{V_1} z^{1:S}_{1:T,1:F}), \tag{4}$$

其中，W^{Q_1}、W^{K_1}、$W^{V_1} \in \mathbb{R}^C$ 是全局跨轨道注意力中的可学习参数，Attn(·) 是原始 Transformer 的编码器层。通过允许每个音轨参考所有音轨在时间和频率维度上的信息，这种方法有助于使所有乐器轨道保持一致的速度。

2) 时间特定跨轨道注意力。全局跨轨道注意力能够使整首乐曲保持一致的速度，但要实现音轨之间更细粒度的节奏同步，则需要局部时间上下文来实现适当的对齐。为此，如图 3c(ii) 所示，我们引入了时间特定跨轨道注意力子模块。以表示 $z^s_{t,\ f} \in \mathbb{R}^C$ 为例。在同一 t，我们沿频率维度收集所有音轨的表示，记为 $z^{1:S}_{t\ ,\ 1:F}$，并按如下方式聚合这些信息以得到 $z^s_{t,\ f} \in \mathbb{R}^C$：

$$\mathrm{Attn}_{\text{time_cross}}(z^s_{t,f}) = \mathrm{Attn}(W^{Q_2} z^s_{t,f}, W^{K_2} z^{1:S}_{t,1:F}, W^{V_2} z^{1:S}_{t,1:F}). \tag{5}$$

其中，W^{Q_2}、W^{K_2}和 $W^{V_2} \in \mathbb{R}^C$是时间特定跨轨道注意力中的可学习参数。通过在同一 t 上关注其他音轨，时间特定跨轨道注意力促使乐器在时间上对齐其音乐事件，从而形成紧密同步的起音模式和和弦结构。

轨道专用模块。在 SyncTrack 中，我们利用轨道专用模块捕捉轨道特定信息，例如不同的音色和音域。如图 3b 所示，我们设计了一种可学习的乐器先验。首先，我们利用独热向量 V表示不同的音轨。随后，这些向量 V通过位置编码（Mildenhall 等，2021）进行编码，并由两层神经网络进行转换。最后，将 V 的嵌入添加到 n的时间嵌入中。然后，我们将可学习的乐器先验添加到第一个残差块的输出中。最终的轨道特定表示由第二个残差块生成。

4 节奏稳定性与同步指标

为了解决多轨音乐生成中的节奏稳定性和同步问题，我们提供了三种不同的指标。这些指标直接衡量所有样本中 S 所有音轨的节拍是否同步 N，为评估不同方法下的多轨音乐生成结果提供了一种可复现且易于解释的方式。

轨内节奏稳定性（IRS）。对于节奏稳定的音乐，每条音轨内的节拍间隔应保持一致。IRS 通过计算每条音轨 s在所有样本中的节拍间隔（Dannenberg，1987年；Robertson，2012年）标准差的平均值，来量化时间一致性：

$$\mathrm{IRS} = \mathbb{E}_{s,n}\left[std(I^s_n)\right], \tag{6}$$

where I^s_n 表示样本 n中音轨 s 的节拍间隔，其中每个元素定义为两次 e diff相邻节拍之间的时间差。

跨轨节拍同步（CBS）。CBS 衡量多条音轨之间的节奏同步程度。受节拍跟踪中容差窗口概念的启发（Dixon，2001年），我们将时间轴划分为多个时间窗口，并计算每个窗口内至少包含一个节拍的音轨所占比例。仅考虑包含内容的音轨。CBS 定义如下：

$$\mathrm{CBS} = \mathbb{E}_n\left[\frac{\sum_{i=1}^{T} r_{i,n}}{\sum_{i=1}^{T} \mathbb{I}(r_{i,n} > 0)}\right], \tag{7}$$

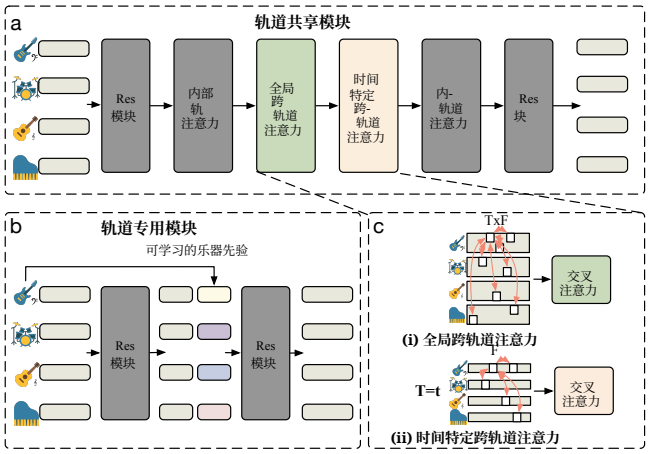


图 3： (a) 轨道共享模块和 (b) 轨道专用模块示意图。在 (a) 中，我们利用轨内注意力捕捉轨内节奏稳定性，并设计 (c) 两个跨轨道注意力子模块，以捕捉跨轨道节奏稳定性与同步。在 (b) 中，我们构建可学习的乐器先验，以捕捉音色及其他轨道特定特征。

其中 $r_{i,\ n}$ 表示第 i 个窗口内至少包含一个节拍的音轨比例，而我们使用$\mathbb{I}(r_{i,\ n} > 0)$排除所有音轨中均未出现节拍的窗口。

跨轨道节拍离散度（CBD）。CBD 指标通过测量所有音轨对之间节拍对齐的离散程度，量化多轨音乐中的节奏同步。受 Goto 方法（Goto & Muraoka，1997）的启发，该方法使用节拍误差序列评估估计节拍与参考节拍之间的对齐误差；CBD 将这一概念扩展到了多轨场景。

我们依次将每条音轨选作参考，并计算其与所有其他音轨之间的节拍误差序列。对于样本 n 中的音轨 s，令 $b^s_{n,t}$ 表示参考音轨中的第 t 个节拍。对于每个 $b^s_{n,t}$，我们在其他音轨中找到匹配节拍，并提取误差序列，记为 $e(b^s_{n,t})$。CBD 指标定义为节拍误差序列的均值或其他统计量：

$$\text{CBD(mean)} = \mathbb{E}_{s,n,t}\left[e(b^s_{n,t})\right]. \tag{8}$$

请注意，为消除速度变化的影响，我们使用节拍间隔对 $e(b^s_{n,t})$ 进行归一化。由于 $b^s_{n,t}$ 的匹配节拍位于两个间隔 $\left[-I^s_{n,t-1}/2 + b^s_{n,t}, b^s_{n,t}\right]$ 和 $\left[b^s_{n,t}, b^s_{n,t} + I^s_{n,t}/2\right]$ 之内，因此我们将 $e(b^s_{n,t})$ 除以相应的间隔长度 $I^s_{n,t-1}/2$ 或 $I^s_{n,t}/2$。

5 实验

本节通过实验回答以下研究问题：

RQ1 SyncTrack 是否优于最先进的多轨音乐生成方法？*RQ2* SyncTrack 在轨内稳定性和跨轨同步方面表现是否良好？研究问题*3* 各个关键模块分别对我们的方法作出了什么贡献？*RQ4* 所提出的指标是否稳健，能否反映人类主观偏好？

5.1 实验设置。

客观实验。 我们在 Slakh2100 数据集（Manilow 等，2019）上进行客观实验，并采用包含四种音轨的常用子集（Mariani等，2023）：贝斯、鼓、吉他和钢琴。所有音频文件均重采样至 16 kHz，并切分为 10.24 秒的片段。音频片段转换为梅尔频谱图用于训练。模型使用 MusicLDM（Chen等，2024）的预训练权重进行初始化，并以批次大小 16 训练 320K 次迭代。推理时使用 200 步的 DDIM采样器。更多详情请参见附录 A.5。

主观实验。 我们开展基于网页的评估，分为两部分：混合音乐评估和单轨质量评估。对于混合音，我们设计了四个评分组，每组由参与者在五分制下对三个音频样本进行评分，比较来自 Slakh2100 的真实样本与由 SyncTrack 和 MSG-LD 生成的样本。对于单独音轨，我们选取两种具有代表性的乐器：鼓和钢琴；每种乐器设置两个评分组，每组提供三个片段（分别来自 Slakh2100、SyncTrack 和 MSG-LD），并采用三分制评分。更多详情请参见附录 A.6。

基线方法。 我们选择 MSDM（Mariani等，2023年）、STEMGEN（Parker 等人，2024年）、JEN-1 Composer（Yao 等人，2025b）和 MSG-LD（Karchkhadze 等人，2025年）作为基线方法。我们的基线选择严格限定于多轨波形音乐生成任务。面向立体声混合音频或符号音乐设计的模型不在本比较范围内。

5.2 生成的多轨音乐质量（RQ1）

在多轨音乐生成中，用于评估整体音频质量的最广泛使用的指标，是在所有音轨的混合音频上计算的 FAD。我们在 Slakh 测试数据集上，以无条件模式评估 SyncTrack 在整体音乐生成任务中的表现。随后，我们计算生成的混合音频与测试集混合音频之间的 FAD。如表 1 所示，SyncTrack 显著优于 MSDM，使 FAD分数从 6.55 大幅降低至 1.26。与其他基线方法相比，SyncTrack 的 FAD 分别降低了 70.70%、68.81% 和 3.82%。这一显著改进证实了 SyncTrack 能够生成高质量且连贯的混合音频。

表 1：混合音乐的 FAD↓ 分数。

指标	MSDM	STEMGEN	JEN-1 Composer	MSG-LD	SyncTrack
FAD	6.55	4.3	4.04	1.31	**1.26**

由于 MSDM 和 MSG-LD 提供了开源代码，我们进一步使用它们作为基线方法，评估每条生成音轨的 FAD 分数。如表 2 所示，SyncTrack 在所有音轨上均稳定优于基线方法；与 MSG-LD 相比，贝斯、鼓、吉他和钢琴的 FAD 分别降低了 32.38%、27.55%、20.77% 和 45.59%。

表2：分轨FAD↓ 分数。

方法	Bass	Drum	吉他	钢琴
SyncTrack	**0.710**	**0.710**	**1.450**	**1.110**
MSG-LD	1.050	0.980	1.830	2.040
MSDM	6.304	6.721	4.259	5.563

贝斯和鼓等节奏轨的 FAD分数最低，它们呈现出最鲜明的打击乐模式，表明我们的模型擅长捕捉这些律动。对于钢琴等旋律轨，由于其音域宽广且频谱复杂，先前的模型通常表现不佳。钢琴轨的 FAD 提升幅度高达 45.59%，凸显了我们的模型对具有挑战性的轨道的复杂分布进行建模的能力有所增强。

表3：主观评估结果。

	混合音				Drum		吉他	
方法	1	2	3	4	1	2	1	2
真实值	4.2±0.9	4.5±0.6	4.7±0.5	4.6±0.6	3.0±0.2	2.6±0.7	2.9±0.3	3.0±0.2
SyncTrack	3.3±1.0	3.5±0.8	3.0±0.9	3.9±0.9	1.9±0.3	2.1±0.5	1.9±0.4	1.8±0.5
MSG-LD	1.5±0.6	1.3±0.5	1.8±0.9	1.7±0.8	1.2±0.5	1.3±0.6	1.2±0.5	1.2±0.4

主观实验的结果见表3。各方法取得的平均分如下：真实值（4.48）、MSG-LD（1.57）和 SyncTrack（3.42），这表明听众对 SyncTrack 生成的音乐评价更高。

5.3 节奏稳定性与同步（RQ2）

本节将探究质量提升是否源于节奏稳定性与同步能力的增强。我们评估了 SyncTrack 和基线方法生成音乐的节奏稳定性（IRS）以及跨轨道节奏同步（CBD、CBS）。此处仅提供统计数值。更多可视化结果见附录 A.3。

表4：分轨 IRS↓ 分数。

方法	Bass	Drum	吉他	钢琴
真实值	**0.015**	**0.005**	**0.016**	**0.015**
SyncTrack	**0.021**	**0.011**	**0.024**	**0.023**
MSG-LD	0.041	0.040	0.039	0.039
MSDM	0.050	0.036	0.034	0.046

轨道内稳定性分析。我们使用提出的 IRS 指标评估节奏稳定性，该指标量化了拍间间隔的方差。稳定的节拍对于高质量音乐至关重要，尤其是对于鼓等打击乐器。如表4所示，真实轨道的 IRS 数值低于所有生成样本，且打击乐器的得分最低，这与预期一致。在所有音轨上，SyncTrack 的 IRS 表现均优于 MSG-LD 和 MSDM，其中鼓的提升幅度最大，凸显了其在建模节奏模式方面的卓越能力。此外，尽管 FAD 无法反映质量的方方面面，我们仍能观察到，较低的 IRS 始终与较低的 FAD 相关，这表明生成音乐与真实音乐之间的差距有所缩小。

跨轨同步分析。为了分析多轨音乐生成中音轨之间的节奏同步情况，我们采用跨轨节拍同步（CBS）和跨轨道节拍离散度（CBD）作为评估指标。CBS 越高，表示同步节拍越多；CBD 越低，表示离散程度越小；两者均意味着同步性越强。

如表5所示，SyncTrack在CBS和CBD两项指标上均取得最佳表现，表明各音轨之间的节奏同步更为紧密。例如，SyncTrack的CBS达到0.5206，比MSG-LD（0.3861）高出34.8%，同时也优于MSDM（0.4694）。在CBD（均值）方面，SyncTrack达到0.2681，相较MSG-LD（0.3714）降低了27.8%。SyncTrack的CBD（标准差）和CBD（中位数）也最低，表明其跨音轨节奏同步效果最佳。

表5：跨轨道同步指标得分。

指标	真实值	SyncTrack	MSG-LD	MSDM
CBS ↑	0.5740	**0.5206**	0.3861	0.4694
CBD（均值）↓	0.2412	**0.2681**	0.3714	0.3127
CBD（标准差）↓	0.1578	**0.2131**	0.2642	0.2217
CBD（中位数）↓	0.2066	**0.2258**	0.3545	0.2811

5.4 消融研究（研究问题3）

SyncTrack 基于骨干网络模型构建，并融入了三个关键模块：a◯ 轨道专用模块、b◯ 全局跨轨道注意力、c◯ 时间特定跨轨道注意力。请注意，b◯ 和 c◯在每个轨道共享模块中按顺序出现。我们在此考察六种消融变体：(1) 骨干网络；(2) 骨干网络 + a◯；(3) 骨干网络 + a◯+b◯；(4) 骨干网络 + a◯+ c◯；(5) SyncTrack-reorder，它将这两个注意力模块 b◯ 和 c◯的顺序颠倒；(6) SyncTrack-alternate，它在每个轨道共享模块中交替仅使用 b◯ 或 c◯ 。如表6所示，我们报告了所有模型的 FAD分数，以及 SyncTrack 最终混合音乐相较于其消融变体的 FAD 提升（提升）。我们的实验旨在回答以下问题：

1. 这三个模块是否都有效？是。SyncTrack 相较于骨干网络取得了显著提升（50%），相较于各变体也有显著提升（11%-27%）

2. 这三个模块是否各自发挥不同作用？消融研究证实，三个模块各自发挥着不同且互补的作用。模块 a◯ 捕捉音色及其他音轨特征，使单个音轨的音频质量显著提升（提升幅度为 56.18% 至 84.41%）。模块 b◯ 增强了各音轨的稳定性。因此，与仅使用 a◯+ 的骨干网络相比，使用 a◯+ b◯ 的骨干网络在单轨音乐质量上有显著提升（6.3%-22.55%）。模块 c◯ 实现细粒度同步，使不同音轨上的音乐事件在相同的时间位置对齐。与使用 a◯ 的骨干网络相比，使用 a◯+ c◯ 的骨干网络在多轨音乐整体质量上显著提升（17.97%）。

3. 各模块的整合是否合理？是的。SyncTrack 的表现优于 SyncTrack-reorder 和 SyncTrack-alternate 两种变体，支持了按所提出的顺序同时纳入 b◯ 与 c◯的设计选择。由于 c◯ 能够在 b◯的基础上实现更细粒度的同步，先 b◯后 c◯ 的顺序是有充分依据的。

表6：轨道专用模块与两类跨轨注意力的消融研究。

模型	Bass	Drum	吉他	钢琴	混合音	提升
骨干网络	5.234	3.081	6.012	6.170	2.570	50.97%
骨干网络 w/ a◯	0.816	0.809	2.634	1.695	1.742	27.67%
骨干网络搭配 a◯+ b◯	**0.632**	0.758	2.367	1.359	1.627	22.56%
骨干网络搭配 a◯+ c◯	0.892	0.889	2.680	1.547	1.429	11.83%
SyncTrack-alternate	0.900	0.897	2.663	1.757	1.586	20.55%
SyncTrack-reorder	0.957	0.943	2.887	1.877	1.681	25.04%
SyncTrack	0.710	**0.710**	**1.450**	**1.110**	**1.260**	-

5.5 评估指标的鲁棒性与主观验证（RQ4）

我们选取客观指标分数（IRS、CBS 和 CBD）各不相同的样本，并将其呈现给人类听众进行主观评估。结果显示出明确的对应关系：IRS 较低的样本获得了较高的主观稳定性评分，而 CBS 较高且 CBD 较低的样本，其节奏同步评分较低，这表明用户认为这些样本的节奏同步性较差。实验详情和问卷设置见附录 A.6。

我们还通过改变节拍跟踪超参数来进一步验证指标鲁棒性；在所有设置下，模型排名均保持一致，证实了我们指标的稳定性（见附录 A.1 和 A.2）。

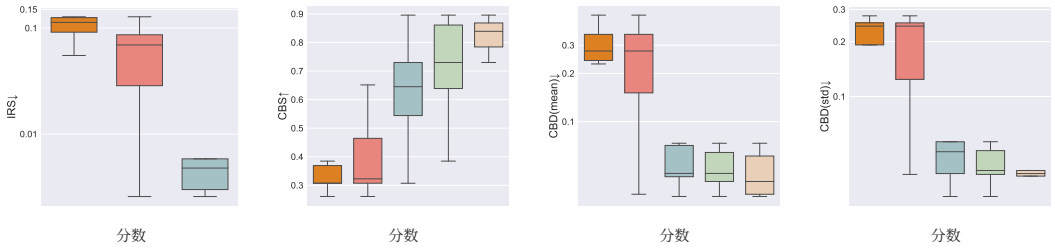


图 4：主观评分与客观指标分数的比较。

6 结论

总之，我们提出了 SyncTrack，有效解决了多轨音乐生成中经常被忽视的节奏稳定性与同步问题。SyncTrack 集成了包含跨轨道注意力子模块的轨道共享模块，以及融入可学习的乐器先验的轨道专用模块，因此能够有效实现节奏一致性，并捕捉每个轨道的独特特征。我们提出的指标——IRS、CBS 和 CBD——为评估节奏特征提供了全面的框架。实验表明，我们的模型在生成音乐的音乐与节奏质量方面取得了显著提升。未来，我们计划扩展 SyncTrack，以生成更长篇幅的多轨音乐，并拓宽其应用范围。

致谢

本研究部分获得中国香港特别行政区研究资助局卓越领域项目（AoE/E-601/24-N）和主题研究项目（T32-615/24-R）的资助。我们也感谢香港创新科技署提供的资助（ITCPD/17-9）。我们谨此感谢所有审稿人提出的宝贵建议，帮助改进本文。

伦理声明

本研究提出了一种机器学习领域的新方法。我们已考虑研究可能产生的社会影响。尽管我们的首要目标是推动知识进步，但我们也承认，任何强大的技术都可能遭到滥用。我们没有具体理由认为本研究会带来重大且迫在眉睫的风险，并鼓励负责任地使用本文开发的技术。本文所有作者声明，与本研究相关的财务或非财务利益均不存在冲突。

可复现性声明

为确保可复现性，我们在附录 A.5 中提供了全面的实验细节。音频样本以及模型和评估指标的源代码均可在我们的演示页面上获取。

参考文献

Mathilde Abrassart 和 Guillaume Doras。如果两个音乐版本没有共同的旋律、和声、节奏或歌词，会怎样？arXiv 预印本 arXiv:2210.01256，2022年。

Josh Achiam、Steven Adler、Sandhini Agarwal、Lama Ahmad、Ilge Akkaya、Florencia Leoni Aleman、Diogo Almeida、Janko Altenschmidt、Sam Altman、Shyamal Anadkat 等。GPT-4 技术报告。arXiv 预印本 arXiv:2303.08774，2023年。

Sebastian B¨ock、Florian Krebs 和 Gerhard Widmer。使用循环神经网络联合跟踪节拍和强拍。载于 ISMIR，第 255–261 页。纽约市，2016年。

Estefania Cano、Derry FitzGerald、Antoine Liutkus、Mark D Plumbley 和 Fabian-Robert St¨oter。音乐源分离：简介。《IEEE 信号处理杂志》，36(1):31–40，2018年。

Ke Chen、Yusong Wu、Haohe Liu、Marianna Nezhurina、Taylor Berg-Kirkpatrick 与 Shlomo Dubnov。MusicLDM：利用节拍同步混合策略提升文本到音乐生成中的新颖性。载于 ICASSP 2024-2024 IEEE 国际声学、语音与信号处理会议（ICASSP），第 1206–1210 页。IEEE，2024年。

Nanxin Chen、Yu Zhang、Heiga Zen、Ron J Weiss、Mohammad Norouzi 与 William Chan。WaveGrad：估计波形生成的梯度。arXiv 预印本 arXiv:2009.00713，2020年。

Yoonjin Chung、Pilsun Eu、Junwon Lee、Keunwoo Choi、Juhan Nam 与 Ben Sangbae Chon。KAD：告别 FAD！一种有效且高效的音频生成评估指标。arXiv 预印本 arXiv:2502.15602，2025年。

Roger B Dannenberg。实时跟随即兴演奏。ICMC 会议论文集，1987年，1987年。

Simon Dixon。从富有表现力的演奏中自动提取速度和节拍。《新音乐研究杂志》，30(1):39–58，2001年。

Chris Donahue、Antoine Caillon、Adam Roberts、Ethan Manilow、Philippe Esling、Andrea Agostinelli、Mauro Verzetti、Ian Simon、Olivier Pietquin、Neil Zeghidour 等。《SingSong：根据歌唱生成音乐伴奏》。arXiv 预印本 arXiv:2301.12662，2023年。

Hao-Wen Dong、Wen-Yi Hsiao、Li-Chia Yang 和 Yi-Hsuan Yang。《Musegan：用于符号音乐生成与伴奏的多轨序列生成对抗网络》。载于《AAAI人工智能会议论文集》，第32卷，2018年。Zach Evans、CJ Carr、Josiah Taylor、Scott H Hawley 和 Jordi Pons。《快速时序条件潜在音频扩散》。载于第四十一届国际机器学习会议，2024a年。Zach Evans、Julian D Parker、CJ Carr、Zack Zukowski、Josiah Taylor 和 Jordi Pons。《利用潜在扩散生成长篇音乐》。arXiv 预印本 arXiv:2404.10301，2024b年。Zach Evans、Julian D Parker、CJ Carr、Zack Zukowski、Josiah Taylor 和 Jordi Pons。《Stable Audio Open》。载于《ICASSP 2025—2025年IEEE 国际声学、语音与信号处理会议（ICASSP）》，第1–5页。IEEE，2025年。Seth Forsgren 和 Hayk Martiros。《Riffusion——用于实时音乐生成的稳定扩散》。网址：https://riffusion.com，2022年。Masataka Goto 和 Yoichi Muraoka。《节拍跟踪系统评估中的问题》。载于《IJCAI-97人工智能与音乐评估问题研讨会工作纪要》，第9–16页，1997年。Kaiming He、Xiangyu Zhang、Shaoqing Ren 和 Jian Sun。《用于图像识别的深度残差学习》。载于《IEEE计算机视觉与模式识别会议论文集》，第770–778页，2016年。Holger Hennig。《人类音乐节奏与相互作用的复杂系统中的同步》。美国国家科学院院刊，111(36)：12974–12979，2014年。Jonathan Ho、Ajay Jain 和 Pieter Abbeel。《去噪扩散概率模型》。神经信息处理系统进展，33：6840–6851，2020年。Tornike Karchkhadze、Mohammad Rasool Izadi 和 Shlomo Dubnov。《利用多轨潜在扩散模型同时进行音乐分离与生成》。载于《ICASSP 2025—2025年IEEE 国际声学、语音与信号处理会议（ICASSP）》，第1–5页。IEEE，2025年。Kevin Kilgour、Mauricio Zuluaga、Dominik Roblek 和 Matthew Sharifi。《Fr\'echet 音频距离：一种评估音乐增强算法的度量》。arXiv 预印本 arXiv:1812.08466，2018年。Diederik P Kingma 和 Max Welling。《自编码变分贝叶斯》。arXiv 预印本 arXiv:1312.6114，2013年。Jungil Kong、Jaehyeon Kim 和 Jaekyoung Bae。《HiFi-GAN：用于高效、高保真语音合成的生成对抗网络》。神经信息处理系统进展，33：17022–17033，2020a年。Zhifeng Kong、Wei Ping、Jiaji Huang、Kexin Zhao 和 Bryan Catanzaro。《DiffWave：一种多用途音频合成扩散模型》。arXiv 预印本 arXiv:2009.09761，2020b年。Edward W Large 和 Caroline Palmer。《感知音乐中的时间规律性》。认知科学，26(1)：1–37，2002年。Sifei Li、Binxin Yang、Chunji Yin、Chong Sun、Yuxin Zhang、Weiming Dong 和 Chen Li。《VidMusician：通过层次化视觉特征实现语义节奏对齐的视频转音乐生成》。arXiv 预印本 arXiv:2412.06296，2024a年。Sizhe Li、Yiming Qin、Minghang Zheng、Xin Jin 和 Yang Liu。《Diff-BGM：一种用于视频背景音乐生成的扩散模型》。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第27348–27357页，2024b年。Haohe Liu、Zehua Chen、Yi Yuan、Xinhao Mei、Xubo Liu、Danilo Mandic、Wenwu Wang 和 Mark D Plumbley。《AudioLDM：利用潜在扩散模型进行文本转音频生成》。arXiv 预印本 arXiv:2301.12503，2023年。

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、语音和语言处理汇刊》，32:2871–2883，2024年。

Ethan Manilow、Gordon Wichern、Prem Seetharaman和Jonathan Le Roux。给音乐源分离留点Slakh：一个用于研究训练数据质量与数量影响的数据集。见2019年IEEE音频与声学信号处理应用研讨会（WASPAA），第45–49页。IEEE，2019年。

Giorgio Mariani、Irene Tallini、Emilian Postolache、Michele Mancusi、Luca Cosmo和Emanuele Rodol`a。用于同时进行音乐生成与分离的多源扩散模型。arXiv预印本 arXiv:2302.02257，2023年。

Ernest Mas-Herrero、Alain Dagher 和 Robert J Zatorre。利用经颅磁刺激增强或减弱音乐奖赏敏感性。《自然·人类行为》，2(1):27–32，2018年。

Soroush Mehri、Kundan Kumar、Ishaan Gulrajani、Rithesh Kumar、Shubham Jain、Jose Sotelo、Aaron Courville 和 Yoshua Bengio。SampleRNN：一种无条件端到端神经音频生成模型。arXiv预印本 arXiv:1612.07837，2016。

Ben Mildenhall、Pratul P Srinivasan、Matthew Tancik、Jonathan T Barron、Ravi Ramamoorthi 和 Ren Ng。NeRF：将场景表示为用于视图合成的神经辐射场。《ACM通讯》，65(1):99–106，2021。

Danuta Mirka。《聆听时间：音乐节拍的心理学层面》，2004年。

Julian D Parker、Janne Spijkervet、Katerina Kosta、Furkan Yesiler、Boris Kuznetsov、Ju-Chiang Wang、Matt Avent、Jitong Chen 和 Duc Le。Stemgen：一种能够聆听的音乐生成模型。发表于 ICASSP 2024-2024 IEEE 国际声学、语音与信号处理会议 (ICASSP)，第 1116–1120 页。IEEE，2024年。

XiangfeiQiu、Hanyin Cheng、Xingjian Wu、Jilin Hu、Chenjuan Guo 和 Bin Yang。多元时间序列预测深度学习综述：通道策略视角。arXiv 预印本 arXiv:2502.10721，2025年。

Zafar Rafii、Antoine Liutkus、Fabian-Robert St¨oter、Stylianos Ioannis Mimilakis 和 Rachel Bittner。用于音乐分离的 musdb18 语料库。2017年。

Ali Razavi、Aaron Van den Oord 和 Oriol Vinyals。使用 vq-vae-2 生成多样且高保真的图像。《神经信息处理系统进展》，第 32 卷，2019年。

Andrew Robertson。根据节拍标注解码音乐录音中的速度和时序变化。载于 ISMIR，第 475–480 页，2012年。

Robin Rombach、Andreas Blattmann、Dominik Lorenz、Patrick Esser 和 Bj¨orn Ommer。使用潜在扩散模型进行高分辨率图像合成。载于 IEEE/CVF 计算机视觉与模式识别会议论文集，第 10684–10695 页，2022年。

Olaf Ronneberger、Philipp Fischer 和 Thomas Brox。U-Net：用于生物医学图像分割的卷积网络。载于医学图像计算与计算机辅助干预国际会议，第 234–241 页。施普林格，2015年。

Flavio Schneider、Ojasv Kamal、Zhijing Jin 和 Bernhard Sch¨olkopf。《Moˆusai：高效的文本转音乐扩散模型》。载于《计算语言学协会年会》第62届会议论文集（第1卷：长篇论文），第8050–8068页，2024年。Zeyue Tian、Zhaoyang Liu、Ruibin Yuan、Jiahao Pan、Qifeng Liu、Xu Tan、Qifeng Chen、Wei Xue 和 Yike Guo。《Vidmuse：一种采用长短期建模的简易视频生成音乐框架》。载于《计算机视觉与模式识别会议》论文集，第18782–18793页，2025年。

Aaron Van Den Oord、Sander Dieleman、Heiga Zen、Karen Simonyan、Oriol Vinyals、Alex Graves、Nal Kalchbrenner、Andrew Senior、Koray Kavukcuoglu 等。《WaveNet：一种原始音频生成模型》。arXiv预印本 arXiv:1609.03499，12:1，2016年。Juncheng Wang、Chao Xu、Cheng Yu、Lei Shang、Zhe Hu、Shujun Wang 和 Liefeng Bo。《基于梅尔量化-连续分解的同步视频到音频生成》。载于《计算机视觉与模式识别会议》论文集，第3111–3120页，2025年。Yunbo Wang、Haixu Wu、Jianjin Zhang、Zhifeng Gao、Jianmin Wang、Philip S Yu 和 Mingsheng Long。《PredRNN：用于时空预测学习的循环神经网络》。IEEE模式分析与机器智能汇刊，45(2):2208–2225，2022年。Li-Chia Yang 和 Alexander Lerch。《音乐生成模型的评估》。神经计算与应用，32(9):4773–4784，2020年。Jixun Yao、Guobin Ma、Huixin Xue、Huakang Chen、Chunbo Hao、Yuepeng Jiang、Haohe Liu、Ruibin Yuan、Jin Xu、Wei Xue 等。《SongEval：歌曲美学评价基准数据集》。arXiv预印本 arXiv:2505.10793，2025a。Yao Yao、Peike Li、Boyu Chen 和 Alex Wang。《Jen-1 Composer：高保真多轨音乐生成的统一框架》。载于《AAAI人工智能会议》论文集，第39卷，第14459–14467页，2025b。Botao Yu、Peiling Lu、Rui Wang、Wei Hu、Xu Tan、Wei Ye、Shikun Zhang、Tao Qin 和 Tie-Yan Liu。《Museformer：采用细粒度与粗粒度注意力的音乐生成Transformer》。神经信息处理系统进展，35:1376–1388，2022年。Jiashuo Yu、Yaohui Wang、Xinyuan Chen、Xiao Sun 和 Yu Qiao。《长期节奏视频音轨生成器》。载于《国际机器学习会议》，第40339–40353页。PMLR，2023年。Fan Zhang、Malu Zhang、Wenyu Chen、Siying Wang、Xin Zhang、Jiayin Li 和 Yang Yang。《利用数据稀疏性抑制离线强化学习中的误差加剧》。IEEE神经网络与学习系统汇刊，第1–15页，2025年。DOI：10.1109/TNNLS.2025.3615982。

A 附录

A.1 节拍跟踪实现细节

工具与默认设置。 我们在 Slakh2100 数据集（Manilow等，2019）上进行客观实验。需要注意的是，我们没有选择 MUSDB18，因为数据稀疏性会导致较大的近似误差（Zhang等，2025年）。在全部实验中，我们使用 madmom（B¨ock等，2016年）库中的 *RNNDownBeatProcessor* 和 *DBNDown-BeatTrackingProcessor* 进行节拍提取。除非另有说明，我们将帧率（fps）设为 150Hz，并对所有主要结果使用 madmom 的默认转移 Lambda（时间限制）数值（B¨ock等，2016年）。

关键参数。 节拍检测准确率会直接影响与节奏相关的指标。为检验我们指标的稳健性，我们系统地调整节拍跟踪中的两个关键参数：

• 每秒帧数（*fps*）：更高的帧率可提供更精细的时间分辨率，从而实现更精确的节拍定位——这对多轨或节奏稀疏的内容尤为重要。较低的帧率会降低计算成本，但可能降低检测准确率。

• 转移 *Lambda (tl)*：此参数控制节拍序列推断过程中的时间平滑。较高的 tl 会使速度估计更加平滑、一致，但可能掩盖生成音轨中真实存在的节奏不稳定性。过高的数值可能因掩盖局部不规则性而人为抬高节奏稳定性得分。我们发现，默认 tl 在敏感性与平滑效果之间取得了良好平衡，能够可靠地反映真实的节奏质量。

A.2 参数敏感性分析

为严格评估我们评估指标的稳健性，我们在一系列节拍跟踪配置下开展实验，同时改变帧率和 tl。图 A1 和图 A2 以及表 A1 总结了这些参数对三个指标的影响：轨内节奏稳定性（IRS）、跨轨节拍同步（CBS）和跨轨道节拍离散度（CBD）。

在所有设置下，模型的相对排名保持稳定，表明我们的指标对节拍跟踪超参数的变化具有稳健性。

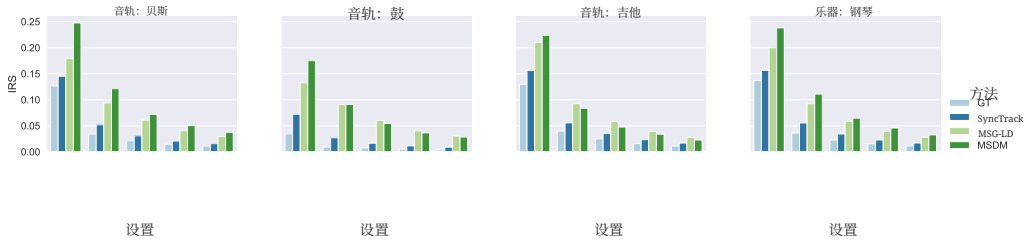


图 A1：不同超参数设置下 IRS 的比较，

轨内节奏稳定性（IRS）。如图 A1 所示，随着节拍跟踪参数变化，真实值、SyncTrack、MSG-LD 和 MSDM 之间的相对排序保持不变。这表明，我们的节奏稳定性指标对节拍跟踪超参数具有稳健性，无论采用何种具体配置，模型比较结果都可靠。

此外，提高帧率或时间限制都会使所有方法的 IRS 数值持续下降。这一趋势反映了节拍跟踪超参数的影响：更高的帧率带来更精细的时间分辨率，从而实现更精确、稳定的节拍定位；而更高的时间限制则会加强时间平滑，使速度估计更规则。这两种效应都会降低节拍间隔的测量方差，进而降低 IRS 分数。

跨轨节拍同步（CBS）。如图 A2 所示，CBS 始终受到节拍跟踪超参数和所选窗口大小的影响。改变窗口大小会产生明显且直观的影响：窗口越小，将节拍视为同步的标准越严格，因此 CBS 分数越低；窗口越大，该标准越宽松，CBS 数值也越高。这种关系在所有方法和参数设置下始终一致。

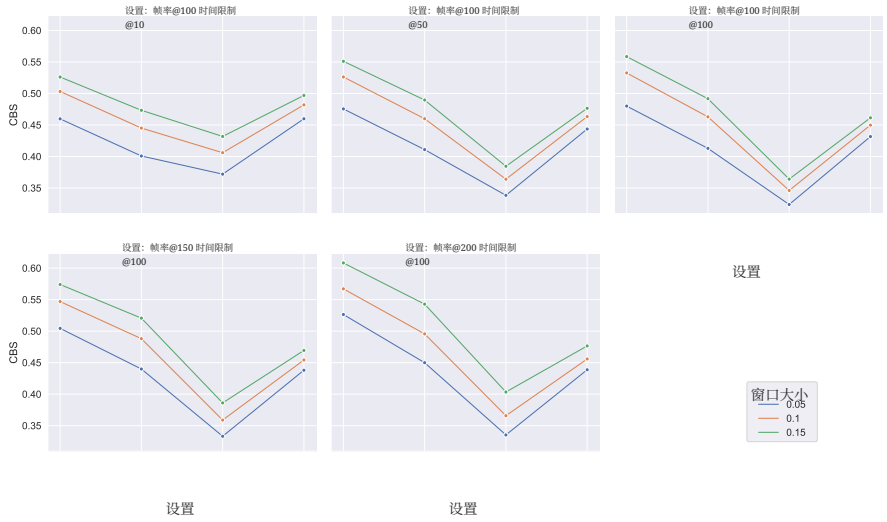


图 A2：不同超参数设置下 CBS 的比较。

尽管 CBS 数值绝对值存在这些变化，但真实值、SyncTrack、MSG-LD 和 MSDM 之间的相对排名在所有配置下均保持不变。真实值始终取得最高的 CBS，且无论选择何种超参数或窗口大小，模型之间的排序都保持不变。这种稳定性表明，CBS 作为比较指标，在多种评估设置下评估模型性能时具有稳健性。

跨轨道节拍离散度（CBD）。表 A1 展示了 CBD 均值、标准差和中位数的详细统计数据。此外，尽管 CBD 指标的绝对值会随帧率和时间限制略有变化，但模型的相对排名几乎不受影响。这进一步证实了我们的评估框架对节拍跟踪超参数不敏感。

表 A1：不同超参数设置下 CBD 的比较。

设置	指标	真实值	SyncTrack	MSG-LD	MSDM
帧率@100 时间限制 @50	CBD（均值）↓	0.2206	**0.2589**	0.3644	0.2923
	CBD（标准差）↓	0.1693	**0.2241**	0.2822	0.2370
	CBD（中位数）↓	0.1769	**0.2058**	0.3350	0.2424
帧率@100 时间限制 @100	CBD（均值）↓	0.2143	**0.2522**	0.3829	0.3205
	CBD（标准差）↓	0.1556	**0.2101**	0.2708	0.2329
	CBD（中位数）↓	0.1762	**0.2060**	0.3679	0.2870
帧率@150 时间限制 @100	CBD（均值）↓	0.2412	**0.2681**	0.3714	0.3127
	CBD（标准差）↓	0.1578	**0.2131**	0.2642	0.2217
	CBD（中位数）↓	0.2066	**0.2258**	0.3545	0.2811
帧率@200 时间限制 @100	CBD（均值）↓	0.2642	**0.2926**	0.3590	0.3109
	CBD（标准差）↓	0.1639	0.2203	0.2534	**0.2134**
	CBD（中位数）↓	0.2316	**0.2545**	0.3407	0.2807

A.3 SyncTrack 节奏稳定性与同步性的提升 N

我们考察 SyncTrack 的节奏稳定性与同步性提升在整个 Slakh2100 数据集中的一致性。

节奏稳定性。 我们通过计算贝斯、鼓、吉他和钢琴的 IRS 分数来考察节奏稳定性。从图 A3 可以看出，SyncTrack 的 IRS 低于 MSDM 和 MSG-LD，表明 SyncTrack 生成的音乐与真实音乐更为接近，具有更高的节奏稳定性。

节奏同步。 为了分析多轨音乐生成中各音轨之间的节奏同步情况，我们采用 CBS 和 CBD 作为评估指标。CBS 衡量多个音轨同时出现节拍的时间窗口占比；CBS 越高，表示同步性越强。CBD 量化各音轨节拍时序的离散程度；CBD 越低，表示节奏同步性越强。

如正文表4所示，SyncTrack在所有生成方法中取得了最佳性能。我们进一步检验了这一改进在节奏同步方面的一致性。如图A4所示，可以看到SyncTrack的生成音乐总体上优于MSDM和MSG-LD。SyncTrack确保了多条音轨之间更优的节奏同步。

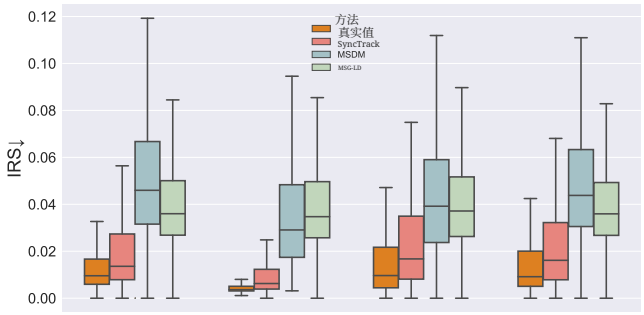
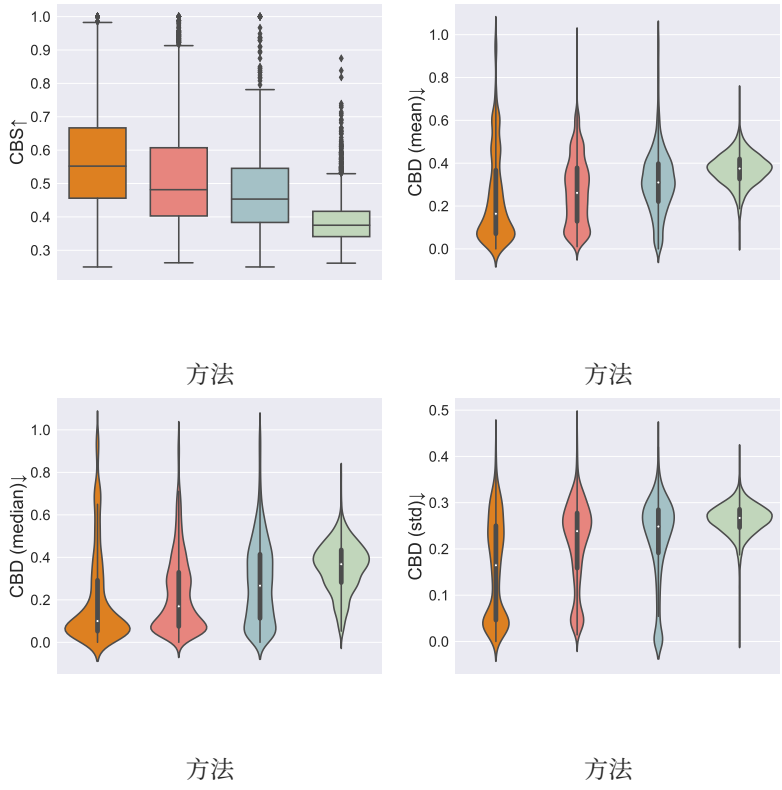


图 A3：逐轨道 IRS 分数。



图A4：跨轨道同步指标分数。

A.4 节奏评估指标案例研究

为进一步验证我们所提出的节奏评估指标的有效性与可解释性，我们给出一项案例研究，比较来自 Slakh2100 数据集的真实值（真实值）样本与来自基线模型的生成样本。我们重点关注三项关键指标：IRS、CBS和CBD。

轨内节奏稳定性。 图 A5 展示了真实值数据集中两个具有代表性的鼓轨（左）以及基线生成样本中的两个鼓轨（右）。对于每个鼓轨，我们展示了频谱图、节拍和强拍激活，以及提取的节拍序列。真实值鼓轨的节拍间隔高度规律且稳定，这一点从可视化的节拍网格以及极低的 IRS 数值（0.0049 和 0.0051）中均可看出。相比之下，基线生成的音轨呈现出不规则的节拍序列，节拍间隔随时间波动，IRS 数值也显著更高（0.1859 和 0.1494）。这些结果表明，IRS 能有效捕捉单条音轨内节奏模式的稳定性，其评估结果与人类的直观判断一致。

多轨节拍同步。 我们进一步分析了多轨同步情况，选取了两个真实值示例和两个基线生成示例，如图 A6 和表 A2 所示。真实值样本表现出较强的跨轨同步：不同乐器（贝斯、鼓、吉他和钢琴）的节拍对齐良好，这一点从各音轨节拍标注的垂直对齐情况中可以看出。相应地，CBS 和 CBD 指标均表明，不同音轨节奏之间具有较高的同步性和较低的离散程度。

相比之下，生成样本在各音轨之间呈现出混乱的节拍，某些情况下甚至整条音轨都是空的。这体现在较低的 CBS 数值和较高的 CBD 数值上，表明跨轨对齐效果更差，且音轨之间的节奏离散度更大。

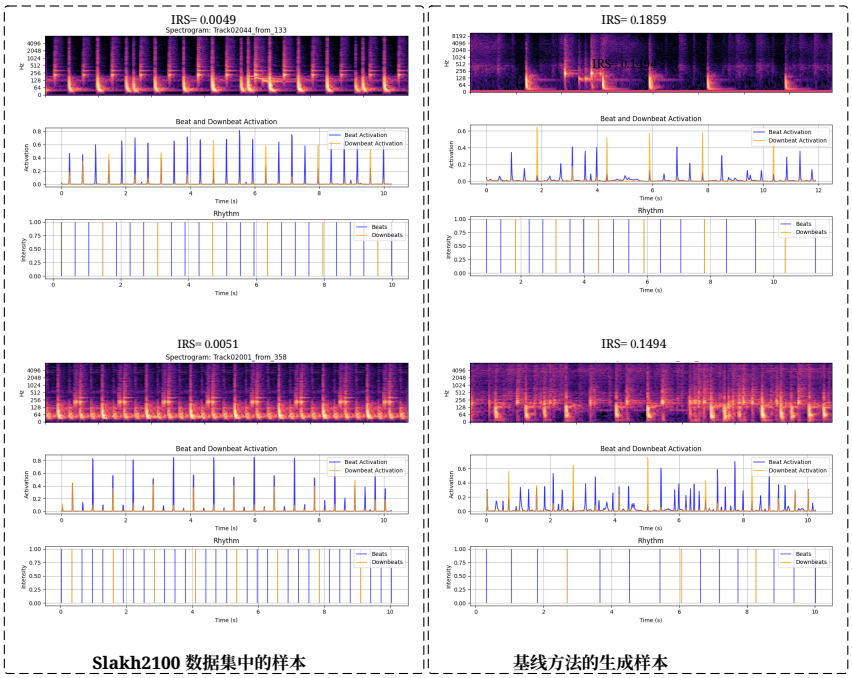


图 A5：真实值与生成样本的 IRS 案例研究。

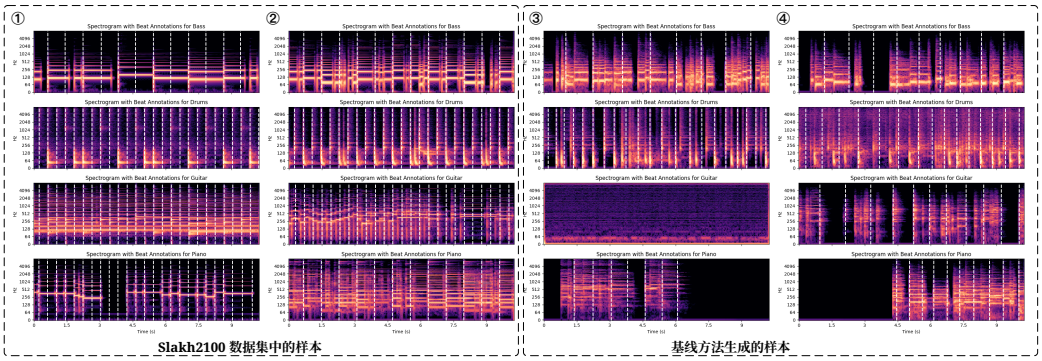


图 A6：真实值与生成样本中的跨轨同步可视化

es.

表 A2：真实值与生成样本的 CBS 和 CBD 对比。

	样本	CBD(均值)	CBD(标准差)	CBD(中位数)	CBS
Slakh2100	①	0.0702	0.0481	0.0709	0.5286
	②	0.0628	0.0677	0.0553	0.7857
由基线方法生成	③	0.2619	0.2512	0.1522	0.3629
	④	0.3975	0.2387	0.3815	0.3041

A.5 训练配置与模型详情。

训练期间，音频片段使用窗口大小为 1024、跳长为 160 个采样点的参数转换为梅尔频谱图。每个片段的混合音频通过对各个音轨求和获得。使用学习率为 $3e-5$ 的 Adam优化器。

SyncTrack包含2.41亿个可训练参数和1.28亿个不可训练参数。架构的更多详情汇总于表A3中。在A6000 GPU上以16为批次大小进行训练时，每个训练轮次约需11分钟。完整训练过程耗时不到3.5小时（3:07:37），共完成21个训练轮次。

表 A3：SyncTrack 的网络架构

层（类型）	深度索引	输入形状	输出形状	参数数量
SyncTrack	-	[1, 4, 8, 64, 64]	-	128,122,516
- 扩散包装器	1-1	[1, 4, 8, 64, 64]	[1, 4, 8, 64, 64]	-
- UNet模型	2-1	[4, 16, 64, 64]	[4, 8, 64, 64]	16
- 序列（时间嵌入）	3-1	[4, 128]	[4, 512]	-
- 线性层	4-1	[4, 128]	[4, 512]	66,048
- SiLU	4-2	[4, 512]	[4, 512]	-
- 线性层	4-3	[4, 512]	[4, 512]	262,656
- 顺序模块（标签嵌入）	3-2	[4, 8]	[4, 512]	-
- 线性层	4-4	[4, 8]	[4, 128]	1,152
- SiLU	4-5	[4, 128]	[4, 128]	-
- 线性层	4-6	[4, 128]	[4, 512]	66,048
- 模块列表（输入块）	3-3	-	-	-
- 轨道专用模块	4-7	[4, 8, 64, 64]	[4, 128, 64, 64]	9,344
- 轨道专用模块	4-8	[4, 128, 64, 64]	[4, 128, 64, 64]	426,880
- 轨道专用模块	4-9	[4, 128, 64, 64]	[4, 128, 64, 64]	426,880
- 轨道共享模块	4-10	[4, 128, 64, 64]	[4, 128, 32, 32]	147,584
- 轨道共享模块	4-11	[4, 128, 32, 32]	[4, 256, 32, 32]	3,681,280
- 轨道共享模块	4-12	[4, 256, 32, 32]	[4, 256, 32, 32]	3,943,424
- 轨道共享模块	4-13	[4, 256, 32, 32]	[4, 256, 16, 16]	590,080
- 轨道共享模块	4-14	[4, 256, 16, 16]	[4, 384, 16, 16]	8,323,712
- 轨道共享模块	4-15	[4, 384, 16, 16]	[4, 384, 16, 16]	8,667,648
- 轨道共享模块	4-16	[4, 384, 16, 16]	[4, 384, 8, 8]	1,327,488
- 轨道共享模块	4-17	[4, 384, 8, 8]	[4, 640, 8, 8]	22,392,448
- 轨道共享模块	4-18	[4, 640, 8, 8]	[4, 640, 8, 8]	23,621,120
- 音轨共享模块（中间）	3-4	[4, 640, 8, 8]	[4, 640, 8, 8]	-
- 轨道专用模块	4-19	[4, 640, 8, 8]	[4, 640, 8, 8]	8,032,640
- 轨道共享模块	4-20	[4, 640, 8, 8]	[4, 640, 8, 8]	15,588,480
- 轨道专用模块	4-21	[4, 640, 8, 8]	[4, 640, 8, 8]	8,032,640
- 模块列表（输出块）	3-5	-	-	-
- 轨道共享模块	4-22	[4, 1280, 8, 8]	[4, 640, 8, 8]	28,128,640
- 轨道共享模块	4-23	[4, 1280, 8, 8]	[4, 640, 8, 8]	28,128,640
- 轨道共享模块	4-24	[4, 1024, 8, 8]	[4, 640, 16, 16]	30,176,768
- 轨道共享模块	4-25	[4, 1024, 16, 16]	[4, 384, 16, 16]	11,274,368
- 轨道共享模块	4-26	[4, 768, 16, 16]	[4, 384, 16, 16]	10,290,816
- 轨道共享模块	4-27	[4, 640, 16, 16]	[4, 384, 32, 32]	11,126,528
- 轨道共享模块	4-28	[4, 640, 32, 32]	[4, 256, 32, 32]	4,993,024
- 轨道共享模块	4-29	[4, 512, 32, 32]	[4, 256, 32, 32]	4,665,088
- 轨道共享模块	4-30	[4, 384, 32, 32]	[4, 256, 64, 64]	4,927,232
- 轨道专用模块	4-31	[4, 384, 64, 64]	[4, 128, 64, 64]	771,584
- 轨道专用模块	4-32	[4, 256, 64, 64]	[4, 128, 64, 64]	607,488
- 轨道专用模块	4-33	[4, 256, 64, 64]	[4, 128, 64, 64]	607,488
- 顺序模块（卷积输出）	3-6	[4, 128, 64, 64]	[4, 8, 64, 64]	-
- GroupNorm32	4-34	[4, 128, 64, 64]	[4, 128, 64, 64]	256
- SiLU	4-35	[4, 128, 64, 64]	[4, 128, 64, 64]	-
- 二维卷积	4-36	[4, 128, 64, 64]	[4, 8, 64, 64]	9,224

A.6 主观评估。

为全面验证所提出指标的有效性，我们旨在确定这些客观指标是否能够反映人类主观感知。为此，我们开展了一项基于网络的主观评估，其中包含两个不同的实验。我们招募了23名参与者参加问卷调查。

实验 1：节奏同步。 本实验旨在评估所有音轨之间的节奏同步情况。我们设置了四个评分组，每组包含三个分别来自 Slakh2100、SyncTrack 和 MSG-LD、且 CBS/CBD 数值各不相同的音频片段。这些多轨片段经过混音后，以随机顺序呈现给参与者。参与者使用五分制对乐器之间的节奏同步程度进行评分。问卷详细说明了同步的定义，以及 1 到 5 分各评分的含义。

实验 2：节奏稳定性。 本实验研究所提出的 IRS 指标与人类对单条音轨节奏稳定性的感知之间的相关性。我们选取两种具有代表性的乐器：鼓（节奏乐器）和钢琴（旋律乐器）。我们从真实数据、我们模型生成的音乐以及 MSG-LD 中抽取 IRS 数值各不相同的单轨片段，即鼓轨或钢琴轨片段。参与者参照实验 1，使用三分制评分量表评估每段单轨片段的轨内节奏稳定性。

A.7 MUSDB18 上提出的指标。

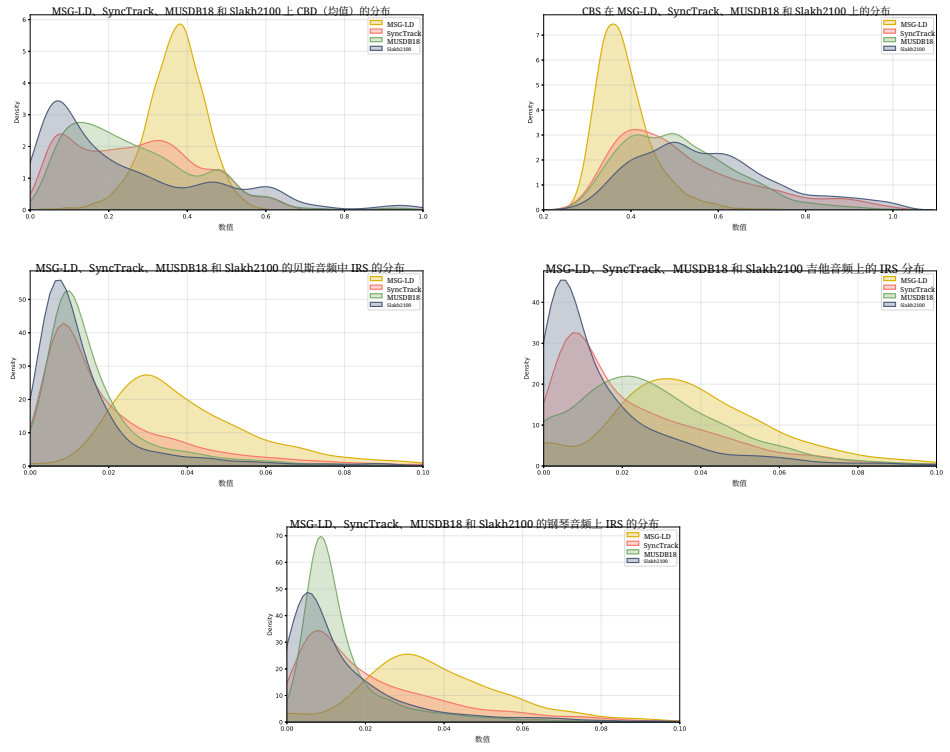


图 A7：MSG-LD、SyncTrack、MUSDB18 和 Slakh210 上三项指标的分布 0.

我们使用 MUSDB18（Rafii 等，2017）进一步说明所提出指标的有效性。MUSDB18 包含真实的录音室录音，涵盖广泛的音乐风格。然而，这类真实世界数据中固有的节奏变化和录音细节可能导致节拍检测的稳定性较低，进而使其在节奏同步指标上的表现逊于完美对齐的 Slakh2100 数据集。表 A4 也表明，MUSDB18 上的 CBS、CBD 和 IRS 分数略低于 Slakh2100 上的分数。不过，

真实世界的音乐仍优于生成音乐，我们的结果完全符合这一直觉。此外，我们在图 A7 中展示了 Slakh2100 、 MUSDB18 中的音乐以及模型生成的音乐（ MSDM 、 MSG-LD 和 SyncTrack ）在三项所提出指标上的分布。可以观察到，Slakh2100 中音乐在三项所提出指标上的分布优于 MUSDB18。尽管如此，MUSDB18 在整体和声和节奏稳定性方面仍优于 MSG-LD 和 SyncTrack 生成的音乐。

表 A4：MUSDB18 和 Slakh2100 在三个指标上的得分

	CBS↑	CBD(均值)↓	CBD(标准差)↓	CBD(中位数)↓	IRS(B)↓	IRS(D)↓	IRS(G)↓	IRS(P)↓
MSG-LD	0.386	0.371	0.264	0.355	0.041	0.040	0.039	0.039
MSDM	0.469	0.313	0.222	0.281	0.050	0.036	0.034	0.046
SyncTrack	0.521	0.268	0.213	0.226	0.021	0.011	0.024	0.023
MUSDB18	0.512	0.264	0.197	0.223	0.017	0.007	0.029	0.015
Slakh2100	0.574	0.241	0.158	0.207	0.015	0.005	0.016	0.015

A.8 条件生成的拟议指标

CBS/CBD 仍可应用于条件生成。条件生成指的是给定一部分已知音轨，生成其余音轨，从而完成连贯的多轨编曲。对于四音轨（贝斯、鼓、吉他和钢琴）音乐生成任务，我们测试了四种情形：(1) B：给定其他三条音轨生成贝斯；(2) BP：给定其他两条音轨生成贝斯和钢琴；(3) BGP：生成贝斯、吉他和钢琴；(4) BDGP：同时生成全部四条音轨。

如表 A5 所示，所有这些情形均可计算 CBS 和 CBD 指标。此外，指标结果符合直觉。根据任务难度，节奏同步应呈现 BGDP>BGP>BP>B 的顺序。根据我们的测试，所得 CBS 和 CBD 分数符合这一顺序。

表 A5：不同生成情形下的 CBS 和 CBD 指标。

	GT	B	BP	BGP	BDGP
CBS↑	0.5740	0.5620	0.5576	0.5259	0.5206
CBD(标准差)↓	0.1578	0.1902	0.1940	0.2118	0.2131
CBD(中位数)↓	0.2066	0.2192	0.2188	0.2304	0.2258
CBD(均值)↓	0.2412	0.2591	0.2579	0.2718	0.2681

A.9 不同片段长度下的拟议指标

我们考察了这些指标在长度为 10秒、30秒、60秒的片段以及完整歌曲（261.83±59.96秒）上的有效性。如表 A6 所示，我们可以看到，我们的指标在所有情况下都.....此外，随着片段长度增加，节奏同步更容易出现性能退化。根据我们的测试，得到的 CBS 和 CBD 分数符合这一排序。

此外，我们在图 A8 中展示了不同长度音乐片段的三项指标分布。直观来看，对于 10秒这类较短的音乐片段，大多数样本的指标表现良好，但仍有一些样本表现较差。对于较长的音乐片段，即完整歌曲，各样本的指标则相对稳定。

表 A6：Slakh2100 在不同音频片段长度下的分数

片段长度	CBS↑	CBD↓			IRS↓			
		mean	std	中位数	Bass	Drum	吉他	钢琴
10s	0.5740	0.2412	0.1578	0.2066	0.015	0.005	0.016	0.015
30s	0.5554	0.2344	0.1742	0.1922	0.0203	0.007	0.0298	0.0254
60s	0.5349	0.2471	0.1960	0.2018	0.0258	0.0121	0.0425	0.0366
完整歌曲	0.4844	0.2660	0.2466	0.1995	0.0489	0.0273	0.0671	0.0710

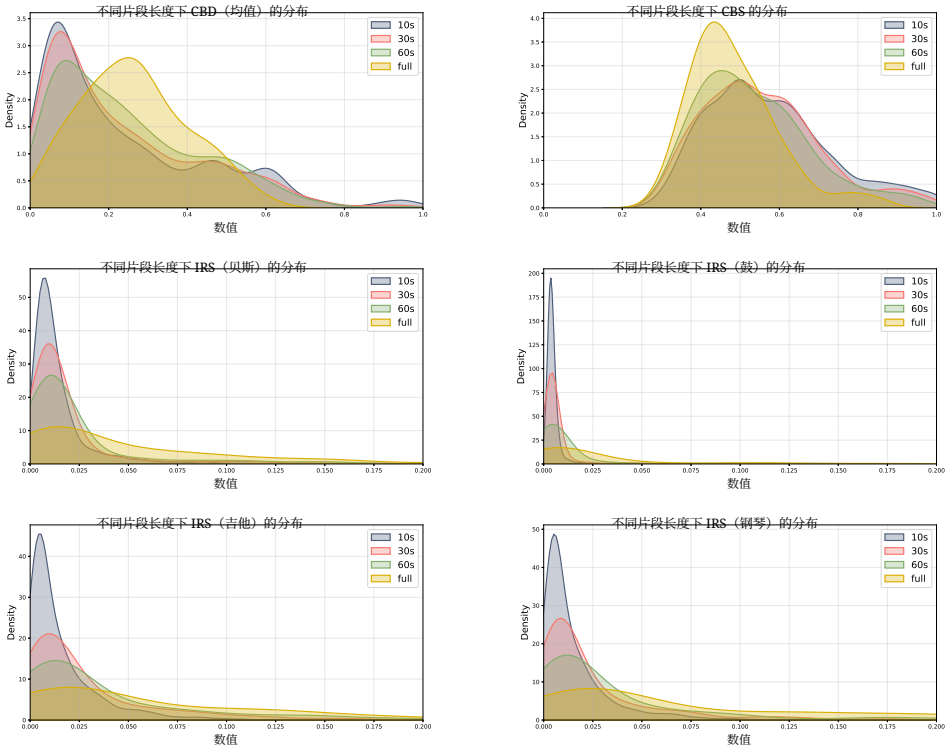


图 A8：不同片段长度下三个指标的分布。

A.10 大语言模型（LLM）的使用

在准备本作品的过程中，作者使用 GPT-4（Achiam 等人，2023年）润色英文文本并提升其清晰度。作者对所有生成内容都进行了仔细审阅、编辑和批判性评估。论文的核心思想、实验设计、数据分析、结果解读以及最终内容均由作者全权负责。大语言模型仅作为写作辅助工具，并未取代作者的批判性思维或学术判断。