DARC：具备细粒度节奏控制的鼓伴奏生成

Trey Brosnan卡内基梅隆大学
rbrosnan@andrew.cmu.edu

arXiv:2601.02357v1 [cs.SD] 5 Jan 2026

摘要

在音乐创作中，快速原型开发对于探索和完善创意至关重要，但现有生成工具往往难以同时满足用户对结构控制和风格灵活性的需求。以分轨为条件生成的既有方法可以根据其他音乐分轨生成内容，却无法充分控制节奏；音色迁移方法则允许用户指定特定节奏，却无法利用音乐上下文作为条件。我们提出 DARC，这是一种生成式鼓伴奏模型，既以其他分轨中的音乐上下文为条件，也接受明确的节奏提示，例如口技打节奏或敲击音轨。我们采用参数高效微调，在 STAGE [1], 这一先进的鼓分轨生成器上实现细粒度节奏控制，同时保持对音乐上下文的感知能力。

1. 引言

近年来，许多研究 [1–7] 实现了高质量且音乐连贯的伴奏生成。然而，这些方法往往缺乏对时变特征的细粒度控制。在音乐原型设计中，这种控制通常十分重要，因为创作者希望在投入大量时间之前，快速评估一个早期音乐创意。在这项工作中，我们聚焦于 Tap2Drum 任务：用户可以录制节奏提示（例如口技打节奏或敲击节奏），再由生成模型将其呈现为鼓声。Tap2Drum 的先进方法主要关注音色迁移，即用户提供音色提示，明确指定期望的鼓音色。例如，[8] 要求用户提供鼓音频作为音色提示；这可能会降低迭代速度，因为不同歌曲需要不同的鼓组音色，而用户必须寻找与期望音色相符的现有音频样本。音乐编辑领域的其他研究 [9] 提供了文本控制，但用文字描述鼓音色可能很困难；此外，这些方法往往会受到音色泄漏 [8]的影响。一些研究在 Tap2Drum [10,11] 和伴奏生成 [1, 12], 中都提供了基于起音的节奏控制，但这种控制粒度过粗，无法捕捉节奏提示中隐含的音色类别。

我们提出了 DARC，这是一种鼓伴奏生成模型，以音乐上下文和节奏提示为输入。我们基于非负矩阵分解（NMF）构建的节奏特征表示，通过将每个起音分类为一个音色类别，比基于起音的方法提供了更高的粒度。DARC 是对 STAGE[1], 这一当前最先进的鼓伴奏模型进行微调后的版本。我们选择从音乐上下文推断音色，而非使用音色提示，原因有二：首先，鼓很少作为独奏乐器出现，也就是说，鼓轨的最终目标通常是为混音伴奏；其次，免去用户提供音色提示的要求，可以缩短其迭代周期，使他们能够探索更多创意。对于数据集，我们使用 Demucs [14, 15]从 FMA 数据集 [13]中提取鼓声分轨。在微调过程中，我们采用了 [2]提出的参数高效方法。

我们的贡献主要有两点：

- 我们提出了一种生成式鼓模型，它可以根据音乐上下文和特定节奏进行条件生成，并支持音色类别。
- 我们评估了模型与输入混音的音乐连贯性，以及其对节奏提示中起音和音色类别的遵循程度，并揭示了现有评估指标的局限性。

2. 相关工作

2.1 伴奏生成

近期有一系列研究探索了音乐伴奏生成 [1–3,5–7, 16], , 这些方法能够生成一条或多条音轨，为给定的音乐混音提供伴奏。需要注意的是，其中许多模型支持文本条件控制，实际上是对文本生成音乐模型 MusicGen [17]进行微调得到的。虽然这些分轨到分轨生成模型可以以混音中的其他音轨作为条件，但它们并非为细粒度节奏控制而设计。一些方法允许以起音 [1, 12]作为条件。然而，这些方法提供的节奏控制相当宽松；模型并不会保留起音，而是将其作为指导，生成经过润色的鼓轨。此外，仅凭起音时序无法捕捉隐含的音色类别，例如某个起音究竟来自底鼓还是军鼓。我们的工作旨在提供更严格的节奏控制，并能够保留音色类别。

其他工作则侧重于鼓声生成中更专业的方面。例如， [18] 可实时生成鼓伴奏，而 [19]使用双向

© R. Brosnan.《DARC：具有细粒度节奏控制的鼓伴奏生成》，作为 *CMU 15-798*的期末项目提交。

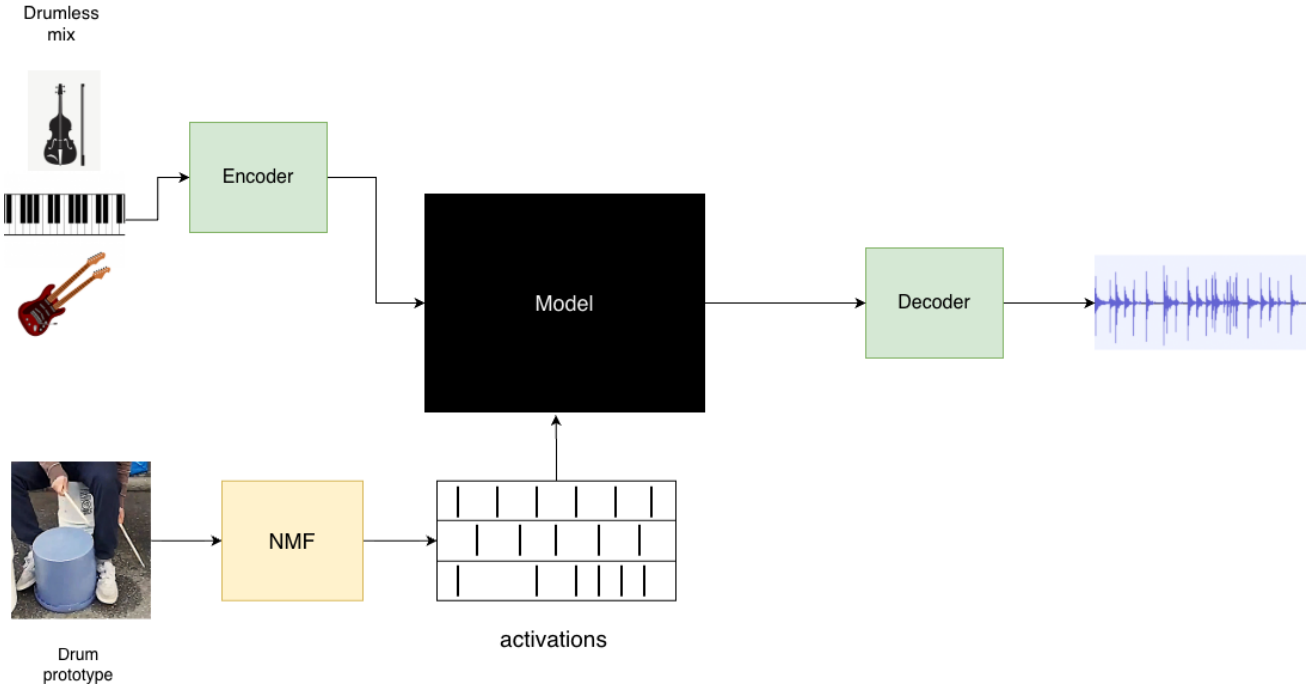

图 1。所提出的节奏条件音乐生成模型的架构。音乐上下文和节奏提示以音频输入的形式提供。将经过令牌化的音乐上下文置于输入序列的前端，并使用非负矩阵分解（NMF）将节奏提示转换为（起音时间、音色类别）对。节奏嵌入通过跳跃式微调和自适应内部注意力 [2]传递至自注意力层。模型输出 EnCodec 音频令牌，经解码后得到最终波形。

语言模型生成鼓点填充。我们将方法适配于实时生成或加花留待未来工作。

2.2 Tap2Drum 生成

另一类研究探索了 Tap2Drum 任务，该任务以敲击或口技打节奏为输入，旨在生成节奏相同的鼓轨。Tap2Drum 最初在 [10], 中提出，该方法以起音时间为输入，并将鼓声生成为 MIDI [1]。其他工作（如 TRIA [8]）则进行音色迁移，直接将节奏提示音频转换为高保真鼓声。除了节奏提示，这类方法还需要以音频形式提供音色提示，因此用户必须提供具有所需精确音色的音频样本。后续研究探索了非零样本音色迁移 [20–24],，这种方法需要针对每种目标音色重新训练模型。我们的模型 DARC 会根据给定的输入混音生成合适的音色，无需针对特定音色提供提示或进行训练。此外，我们的节奏特征除编码起音时间外，还编码音色类别，因此比现有音色迁移方法具有更高的细粒度。

[1]该成果以 Ableton Live 插件的形式发布：https://magenta.withgoogle.com/studio

3. 方法

3.1 概述

我们的模型接收两种音频形式的输入：作为音乐上下文的无鼓混音，以及节奏提示，例如口技打节奏或敲击音轨。我们的目标是生成鼓声分轨，忠实保留节奏提示中每种音色类别的起音，同时与输入混音保持高度的音乐连贯性。我们对 STAGE [1], 进行微调；这是一种近期推出的开源模型，可生成单独的伴奏分轨。STAGE 本身是对 MusicGen [17],进行微调的产物，训练时使用基于前缀的条件控制，同时以无鼓混音和节拍器式脉冲轨为条件。STAGE 包含约 6.2 亿个参数；按照 [2], ，我们采用参数高效微调技术，将可训练参数数量减少一个数量级。请注意，鼓声分轨和贝斯分轨分别使用独立的 STAGE 模型进行训练；本研究仅考虑鼓声模型。

3.2 节奏特征表示

Tap2Drum 任务中的一个关键挑战是音色泄漏：生成的音轨应紧密遵循节奏提示，但其音色应独立于节奏提示。为此，我们使用非负矩阵分解（NMF）提取节奏特征。NMF 将节奏提示的幅度频谱图 S 分解为多个矩阵的乘积， $S = WH$. 基矩阵 W 编码音色信息，而激活矩阵 H 编码时间信息。具体而言， W 的行索引和 H 的列索引分别对应

不同的音色类别。为了提取节奏特征，我们忽略矩阵 W，从而得到一个矩阵 H，其中包含每个音色类别的激活时间。因此，这种节奏特征表示类似 MIDI：对于节拍口技音轨，它会包含每个音符的起始时间和音色类别索引，但不包含底层人声音色的信息。关键的是，我们按照分量能量总和从高到低对音色类别进行排序，前三类大致对应底鼓、军鼓和踩镲。这样一来，模型无需了解音色信息矩阵 W，也能识别音色类别

3.3 微调

我们的基础模型 STAGE 是一个 MusicGen-Small 模型，经过微调后可根据无鼓混音生成鼓声分轨。训练期间，作者在输入前添加无鼓混音的音频令牌，随后添加一个分隔符令牌。因此，在推理阶段，鼓声分轨生成被设定为续写任务，并以输入混音作为提示。作者在其研究中发现，这种基于前缀的条件设定方法优于交叉注意力 [1]。我们保留了这种以无鼓混音为条件的机制，并采用另一种方法为 STAGE 增添细粒度节奏控制。

微调期间，我们冻结 STAGE 中约 80% 的参数。首先，冻结文本编码器和音频令牌嵌入模块。然后，我们采用 [2] 中提出的两种微调策略：跳跃式微调和自适应内部注意力。在跳跃式微调中，每个解码器块仅微调第一个自注意力层，其余三个层则被冻结。在自适应内部注意力中，条件信号会在每个块的第一层重新引入；该机制应用于前 75% 的块。例如，对于包含 48 个自注意力层的解码器，我们有 12 个自注意力块。除第 $0,4,8,12,\ldots,44$ 层外的所有层都会被冻结，并在第 $4,8,12,\ldots,32$. 层重新应用节奏条件。

对于我们的数据集，我们使用 FMA Small [13], ，并通过 Demucs [14, 15]提取鼓声分轨。我们对音乐上下文和节奏提示都进行了数据增强，包括速度和音高移调、高斯噪声以及带通滤波，每种增强都以.25 的概率独立应用。在训练过程中，对无鼓混音进行的任何增强也会应用于真实鼓声分轨，以鼓励分轨与混音之间保持一致。我们使用对数均匀采样，从音频中截取随机 10–30 秒片段进行训练，以偏向较短的片段。由此得到的平均输入长度为 18.2 秒，对应整个训练集的预期时长约为 6 小时。训练使用 A100 GPU，批次大小为 4，共训练 7 轮，耗时 2 小时。

4. 实验设置

我们将模型与 STAGE [1]和 TRIA [8],进行比较，评估音频质量、音乐连贯性和节奏提示遵循度，既考察整体表现，也考察特定音色类别中的表现。

我们使用 MUSDB18 数据集 [25] 评估音乐连贯性，并使用 AVP Beatbox 数据集 [26] 评估节奏遵循度。

4.1 音频质量

我们以主观方式对音频质量进行评估。总体而言，我们认为音频质量相当差，存在频繁的瑕疵，背景中也能听到非鼓类乐器的声音。我们怀疑这些问题源于创建数据集时的音轨分离步骤。众所周知，音轨分离错误会导致串音和瑕疵 [27], ，这与我们的观察一致。未来，我们希望尝试其他音轨分离模型，以及包含真实标注音轨的数据集，以验证这一判断。

4.2 节奏提示遵循度

我们将节奏提示遵循度分为时序准确率（以起音 F1 衡量）和音色类别准确率（以底鼓和军鼓 F1 衡量）。对于起音，我们采用 70 毫秒的容差，并使用 Beat-This [28]对生成的音轨和真实标注音轨进行起音检测。对于音色类别遵循度，我们使用 FrameRNN [29]转录生成的鼓声分轨，并计算底鼓和军鼓起音的 F1 分数，分别采用标准的 30 毫秒和 100 毫秒容差 [8], 。请注意，虽然我们尝试转录 AVP 中的真实标注口技打节奏音轨，但准确率极低，因此改用数据集提供的真实标注。

由于上文 4.1 节讨论的音频质量问题，起音检测和鼓声转录模型在 DARC 的输出上均表现出较低的准确率。因此，我们对音频进行了后处理：对高频进行门控以降低噪声和串音、增强瞬态，并施加轻度压缩和归一化。为确保公平比较，我们对真实标注节奏提示和所有参与比较的模型应用了相同的后处理。节奏提示被截短至 9 秒，检测到的起音少于 2 个的节奏提示则被忽略（在 AVP 中发现了 4 个此类文件）。

4.3 音乐连贯性

为评估音乐连贯性，我们计算每个鼓声分轨与无鼓输入混音之间的 COCOLA 分数 [30]。我们从 MUSDB18 中随机抽取 50 个样本的 10 秒片段作为评估集。作为基线，我们计算真实鼓音轨与无鼓混音之间的 COCOLA 分数。为评估 STAGE，我们按照原论文 [1] 所述进行节奏条件控制：检测节奏提示中的节拍，并将相应的节拍轨与音乐上下文相加，再将结果作为 STAGE 的输入。对于我们的模型，则如 3.2 节所述，直接以非负矩阵分解节奏特征作为条件。

5. 结果与讨论

5.1 节奏提示遵循度

表 1 展示了我们的节奏遵循度结果。我们发现，在全部三项指标上，DARC 的表现都不及 TRIA 和 STAGE。如上文 4.1 节所述，我们的模型音频质量很差，而评估模型对此并不稳健。即使使用 AVP 数据集中的真实节奏提示，这些模型的表现也很差，正如 4.2 节所讨论的那样。此外，尽管我们的后处理在定性上似乎提升了评估模型的表现，但这远非完美的解决方案。尤其是，我们预计，在其他条件不变的情况下，如果提升 DARC 的音频质量，其实验结果将显著改善。因此，提升输出音频保真度是未来工作的重要方向；我们推测，训练时采用生成对抗网络，或调整数据集（例如改用其他源分离模型，或采用包含真实鼓声分轨的数据集，如 MoisesDB [31] ），都可能是有效的方法。

5.2 音乐连贯性

表 2 展示了我们的音乐连贯性结果。与 STAGE 和真实数据相比，我们观察到 DARC 的 COCOLA 分数显著较低。我们再次推测，较低的音频保真度（见第 4.1 节）可能对这些结果产生了影响。有趣的是，在我们的实验中，STAGE 的表现略胜于真实数据。这令人惊讶；尽管 STAGE 确实有可能生成了比真实数据更连贯的鼓声分轨，但我们认为，这反映的其实是 COCOLA 模型本身的局限性。从定性角度看，我们观察到 STAGE 的输出往往比真实数据中的鼓轨更加华丽，因此音符总数也多得多。我们推测，COCOLA 会因 STAGE 输出中每个与音乐语境在节奏上连贯的音符而给予奖励，即使人类听众可能会认为这些修饰过于繁复。这表明，未来有必要开展人工聆听研究来评估音乐连贯性，并设计对音频保真度和与人类偏好的吻合度具有更强鲁棒性的音乐连贯性指标。

6. 结论

我们提出了 DARC，这是一种鼓伴奏生成模型，除了音乐上下文之外，还可以根据节奏提示进行条件控制。我们基于 NMF 的节奏特征能够保留音色类别，同时避免音色泄漏。从定性角度看，我们的模型似乎能够较好地遵循节奏提示，但由于输出的音频保真度较差，定量结果并不理想。这揭示了我们的模型以及现有节奏相似度和音乐连贯性评估指标的一项关键局限。未来的工作可以探索如何提升 DARC 的音频质量；我们建议在创建数据集时使用其他源分离模型，或者完全避免提取过程，转而使用包含真实鼓声分轨的数据集。此外，在训练过程中使用生成对抗网络，或许能够提供一种提升音频质量的机制 [32]。对于评估指标，我们鼓励未来的工作探索稳健的节奏遵循度和音乐连贯性评估指标，使其能够适应不同程度的音频保真度。最后，在改进模型后，我们鼓励未来的工作为 DARC 或其他旨在辅助音乐创作流程的模型开发面向用户的工具。通过观察人类音乐创作者如何与这项技术互动，我们可以更清楚地了解此类模型的现实应用性，并深入认识其广泛影响及有待改进之处。

7. 广泛影响

我们的模型 DARC 旨在与人类创作者协同创作，使其能够精细控制生成输出的节奏轮廓。不过，我们注意到，生成的鼓声分轨的音色由 DARC 根据音乐上下文决定；与音色迁移方法相比，这体现了便利性与控制力之间的权衡。同时，与先前的音轨生成模型（如 STAGE、MusiCon-Gen、StemGen 或 MusicGen-Stem）相比，DARC 接受的节奏输入详细得多；这些模型要么不进行节奏条件控制，要么将每分钟节拍数（BPM）或节拍轨作为节奏输入。因此，就用户控制而言，DARC 介于现有的音色迁移和音轨生成工作之间。

总的来说，鼓生成模型有可能取代人类鼓手。长此以往，这可能导致学习演奏实体架子鼓的人减少，使音乐文化逐渐远离人类鼓手。我们注意到，DARC 的设计初衷是用于协同创作和快速原型开发，但现实世界中的使用情况可能与最初的意图不同。如上所述，未来开展的人机交互研究能够为现实世界使用场景提供见解，也是分析 DARC 及其他模型广泛影响的重要工具。

尤其是当模型未使用足够多样化的数据集进行训练时，可能会表现出对某些音乐风格或声音的偏差，从而加剧音乐同质化。值得注意的是，我们的数据集 FMA Small 中，8 种不同音乐流派的占比均衡 [13], ，有助于促进多样性。然而，大多数音频样本都来自西方音乐。将 DARC 及其他音乐 AI 研究拓展至非西方音乐，是未来一项重要的研究方向，但由于数据稀缺，这可能颇具挑战性。

8. 参考文献

[1]G. Strano、C. Ballanti、D. Crisostomi、M. Man- cusi、L. Cosmo 和 E. Rodolà，“STAGE：通过基于前缀的条件控制进行音轨伴奏生成”，2025年。[在线]。可访问：https://arxiv.org/abs/2504.05690

模型	起音 F1 ↑	底鼓 F1 ↑	军鼓 F1 ↑
STAGE	0.270	0.056	0.134
TRIA	0.347	0.180	0.382
DARC	0.188	0.053	0.111

表 1。AVP[26]数据集上的节奏提示遵循度结果。起音检测由 Beat-This[28], 执行，鼓声转录由 FrameRNN[29]执行。起音、底鼓和军鼓的 F1 分数分别采用 70 毫秒、30 毫秒和 100 毫秒的容差计算。

模型	COCOLA 分数 ↑
STAGE	63.9816
DARC	53.5908
真实标注	63.7227

表 2。在 MUSDB18 [25] 的 50 条随机选取的音轨上评估生成的鼓声分轨与输入音乐上下文之间的音乐连贯性。

[2]Y.-H. Lan、W.-Y. Hsiao、H.-C. Cheng 和 Y.-H. Yang，“MusiConGen：用于基于 Transformer 的文本生成音乐的节奏与和弦控制”，2024。[在线]。可访问：https://arxiv.org/abs/2407.15060

[3]J. D. Parker、J. Spijkervet、K. Kosta、F. Yesiler、B. Kuznetsov、J.-C. Wang、M. Avent、J. Chen 和 D. Le，“Stemgen：一种能够聆听的音乐生成模型”，2024。 [在线]。可访问：https: //arxiv.org/abs/2312.08723

[4] L. Lin、G. Xia、J. Jiang 和 Y. Zhang，“音乐大语言模型建模的基于内容的控制”，2024年。[在线]。可访问：https://arxiv.org/abs/2310.17162

[5] Y.-K. Wu、C.-Y. Chiu 和Y.-H. Yang，“Jukedrummer：通过 Transformer VQ-VAE 在条件节拍感知音频域中生成鼓伴奏”，2022。[在线]。可访问：https://arxiv.org/abs/2210.06007

[6]S. Rouard、R. S. Roman、Y. Adi 和 A. Roebel，“MusicGen-Stem：通过自回归建模实现多声部音乐生成与编辑”，2025年。 [在线]。可访问：https://arxiv.org/abs/2501.01757

[7]C. Donahue、A. Caillon、A. Roberts、E. Manilow、P. Esling、A. Agostinelli、M. Verzetti、I. Simon、O. Pietquin、N. Zeghidour 和 J. Engel，“Singsong：根据歌唱生成音乐伴奏”，2023年。 [在线]。可访问：https://arxiv.org/abs/2301.12662

[8]P. O’Reilly、J. Barnett、H. F. García、A. Chu、N. Pruyne、P. Seetharaman 和 B. Pardo，“万物之节奏：利用掩码语言建模进行音频提示鼓声生成”，2025年。 [在线]。可访问：https://arxiv.org/abs/2509.15625

[9] G. L. Lan, B. Shi, Z. Ni, S. Srinivasan, A. Kumar, B. Ellis, D. Kant, V. Nagaraja, E. Chang, W.-N.Hsu, Y. Shi 和 V. Chandra，“通过单阶段流匹配实现高保真文本引导音乐编辑”，2024年。[在线]。可访问：https://arxiv.org/abs/2407.03648[10]J. Gillick、A. Roberts、J. Engel、D. Eck 和 D. Bamman，“利用逆序列变换学习律动”，2019年。 [在线]。可访问：https://arxiv.org/abs/1905.06118[11] D. Flores García、H. Flores García 和 M. Riondato，“Clavenet：通过数据增强生成非裔古巴鼓乐节奏型”，载于《第*19*届国际 *Audio Mostly* 会议：声音文化探索》，AM ’24 系列。美国纽约州纽约市：计算机协会，2024年，第355–361页。 [在线]。可访问：https://doi.org/10.1145/ 3678299.3678335

[12] S.-L. Wu、C. Donahue、S. Watanabe 和 N. J. Bryan，“音乐 ControlNet：用于音乐生成的多种时变控制”，2023年。 [在线]。可访问：https://arxiv.org/abs/2311.07069

[13]M. Defferrard、K. Benzi、P. Vandergheynst 和 X. Bresson，“FMA：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。 [在线]。可访问：https://arxiv.org/abs/1612.01840

[14] S. Rouard、F. Massa 和 A. Défossez，《用于音乐声源分离的 Hybridtrans-formers》，发表于*ICASSP23*，2023年。

[15]A. Défossez，《混合频谱图与波形的源分离》，发表于*ISMIR 2021* 音乐源分离研讨会，2021年。

[16] I. Villa-Renteria、M. L. Wang、Z. Shah、Z. Li、S. Kim、N. Ramachandran 和 M. Pilanci，《使用潜在扩散模型进行音乐分轨插入的减法训练》，2025年。 [在线]。获取地址：https://arxiv.org/ abs/2406.19328

[17]J. Copet、F. Kreuk、I. Gat、T. Remez、D. Kant、G. Synnaeve、Y. Adi 和 A. Défossez，《简单且可控的音乐生成》，2024。 [在线]。获取地址：https://arxiv.org/abs/2306.05284

[18]B. Haki、M. Nieto、T. Pelinski 和 S. Jordà，《使用 Transformer 架构进行实时鼓乐伴奏》，载于第三届人工智能与音乐创造国际会议论文集。AIMC，2022年9月。 [在线]。可访问：https://doi.org/10.5281/zenodo.7088343

[19]R. Dahale、V. Talwadker、P. Rao 和 P. Verma，《生成连贯的鼓伴奏、加花与即兴演奏》，2022年。 [在线]。可访问：https://arxiv.org/abs/2209.00291

[20]N. Demerlé、P. Esling、G. Doras 和 D. Genova，“结合音频控制与风格迁移的潜在扩散方法”，载于第 *25* 届国际音乐信息检索学会会议（*ISMIR*）论文集，2024 年。

[21]N. Demerlé、P. Esling、G. Doras 和 D. Genova，“结合音频控制与风格迁移的潜在扩散方法”，2024 年。 [在线]。可访问：https: //arxiv.org/abs/2408.00196

[22] M. Mancusi、Y. Halychanskyi、K. W. Cheuk、E. Mo-liner、C.-H. Lai、S. Uhlich、J. Koo、M. A. Martínez-Ramírez、W.-H. Liao、G. Fabbro 和 Y. Mitsufuji，《用于无监督音乐音频音色迁移的潜在扩散桥接》，2025年。 [在线]。可访问：https://arxiv.org/abs/2409.06096

[23]J. Engel、L. H. Hantrakul、C. Gu 和 A. Roberts，《Ddsp：可微分数字信号处理》，载于国际学习表征会议，2020年。 [在线]。可访问：https://openreview.net/forum?id=B1x1ma4tDr[24] A. C. Santos 和 A. Cardoso，《从拍击到鼓：音频到音频打击乐风格迁移》，载于第*25*届国际音乐信息检索学会会议（*ISMIR*）晚期成果演示环节扩展摘要，2023年。[25] Z. Rafii、A. Liutkus、F.-R. Stöter、S. I. Mimilakis 和 R. Bittner，《用于音乐分离的 MUSDB18 语料库》，2017年12月。 [在线]。可访问：https: //doi.org/10.5281/zenodo.1117372[26] A. Delgado、S. McDonald、N. Xu 和 M. Sandler，《用于业余人声打击乐分析的新数据集》，系列 AM ’19。美国纽约州纽约市：计算机协会，2019年，第17–23页。 [在线]。可访问：https://doi.org/10.1145/3356590.3356844[27] S. Vardhan、P. R. Acharya、S. S. Rao、O. R. Jasthi 和 S. Natarajan，《音乐声源分离的集成方法：传统音轨分离与分层音轨分离的比较分析》，载于音乐、声音、艺术与设计中的人工智能，P. Machado、C. J ohnson 和 I. Santos 编。瑞士：Springer Nature Switzerland，2025年，第186–201页。[28] F. Foscarin、J. Schlüter 和 G. Widmer，《节拍追踪！无需动态贝叶斯网络后处理的精准节拍跟踪》，载于第*25*届国际音乐信息检索学会会议（*ISMIR*）论文集，美国加利福尼亚州旧金山，2024年11月。[29]M. Zehren、M. Alunno 和 P. Bientinesi，《基于众包数据的高质量、可复现自动鼓声转录》，《信号》，第4卷第4期，第768–787页，2023年。 [在线]。可访问：https: //www.mdpi.com/2624-6120/4/4/42[30] R. Ciranni、G. Mariani、M. Mancusi、E. Postolache、G. Fabbro、E. Rodolà 和 L. Cosmo，《COCOLA：面向一致性的音乐音频表征对比学习》，2025年。 [在线]。可访问：https://arxiv.org/abs/2404.16969

[31] I. Pereira、F. Araújo、F. Korzeniowski 和 R. Vogl，《MoisesDB：一个超越四声部的源分离数据集》，2023年。 [在线]。可访问：https://arxiv.org/abs/2307.15913[32]C. Donahue、J. McAuley 和 M. Puckette，《对抗式音频合成》，2019年。 [在线]。可访问：https://arxiv.org/abs/1802.04208