arXiv:2603.01984v1 [cs.SD] 2 Mar 2026

VITEX：基于离散扩散模型的多轨符号音乐生成视觉纹理控制

Xiaoyu Yi[⋆†] *Qi He*[†] *Gus Xia*[†] *Ziyu Wang*[#†]

[⋆] 北京大学电子工程与计算机科学学院

[†]Music X 实验室，MBZUAI[#] 纽约大学库朗数学科学研究所

摘要

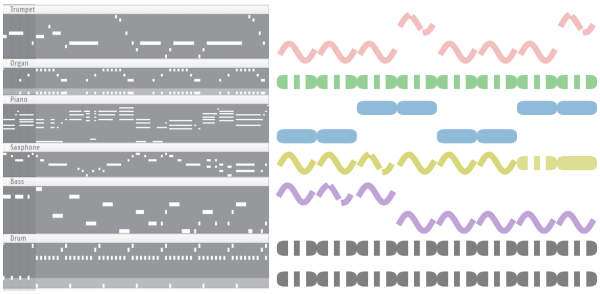
在自动音乐生成中，一个核心挑战是设计能够实现有意义的人机交互的控制方式。现有系统通常依赖文本提示或元数据等外部输入，无法让人直接塑造作曲。尽管先前研究探索了和弦或层级结构等内在控制方式，但这些方法主要针对钢琴或声乐伴奏场景，多轨符号音乐在很大程度上仍未得到充分探索。我们将乐器编制——即乐器的选择及其作用——视为多轨作曲中一种自然的控制维度，并提出 ViTex，一种器乐织体的视觉表征。在 ViTex 中，颜色编码乐器选择，空间位置表示音高和时间，笔画属性则捕捉局部织体。基于这一表征，我们开发了一种以 ViTex 和和弦进行为条件的离散扩散模型，用于生成 8 小节多轨符号音乐，从而实现显式的织体层面控制，同时保持出色的无条件生成质量。演示页面和代码可在 https://vitex2025.github.io/ 获取。

索引词— 音乐生成、扩散、可控生成、音乐信息检索

1. 引言

在自动音乐生成中，一个关键挑战是设计控制，以实现有意义的人机交互。许多现有系统依赖文本提示或元数据（如作曲家身份、音域）等外部输入，这些输入无法让人类直接参与创作过程 [1, 2]。为探索更内在的控制方式，先前研究 [3, 4, 5]探讨了和弦进行、伴奏织体和层次化抽象。然而，这些方法主要聚焦于仅使用钢琴或声乐伴奏的场景，使更复杂的多轨符号音乐生成领域在很大程度上尚未得到充分探索。

当音乐从钢琴创作扩展到多轨创作时，一种新的控制维度随之出现：乐器编制，即对乐器及其角色的选择，这为音乐。这对应于本文所关注的复调织体：确定哪些乐器在何时演奏、采用何种音域，以及它们的织体如何结合成一个协调的整体。主要挑战在于，纹理控制本质上具有高维度，因此难以用直观且易于解释的方式来表示。现有方法要么将这种控制简化为粗略的乐器标签 [1, 2], ，要么将织体编码为难以解释和操控的潜变量 [6, 7]。还有一些方法通过轨道条件生成或补全 [8, 9], ，隐式地处理乐器编制，却未能提供对复调织体的自上而下控制。



支持的乐器：钢琴、钢片琴、吉他、低音提琴、小提琴、合奏，小号、萨克斯风、长笛、合成器音效、鼓

图 1：ViTex 的可视化。左侧：数字音频工作站（DAW）中各轨道的钢琴卷帘。右侧：对应的 ViTex 可视化。不同的字形表示局部织体。横轴表示时间；每个轨道内的垂直位置表示音高。

为弥补这一不足，我们提出 ViTex，这是一种专为可控多轨音乐生成设计的器乐织体视觉表征。图1展示了一个示例。在 ViTex 中，颜色编码乐器选择，空间位置表示音高和时间，笔触属性则捕捉局部织体。当前版本支持多种乐器，包括在织体建模中常被忽略的鼓。这种视觉抽象提供了一种直观而富有表现力的方式来指定多轨织体；它与 Hyperscore [10], 理念相通。Hyperscore 是一种早期界面，曾让用户在深度生成模型出现之前通过绘画进行作曲，并展现了此类表征在实现直观的人机控制方面的潜力。同时，我们的设计也易于转换为适合用于条件控制现代生成模型的机器友好型数据。

基于 ViTex，我们开发了一种以 ViTex 表征和和弦进行为条件的离散扩散模型，用于生成 8小节多轨符号音乐。模型经过训练，可从吸收态对多轨钢琴卷帘进行去噪，并将 ViTex 表征和和弦作为条件。两种条件均采用无分类器引导进行训练 [11], ，使模型能够灵活地在条件生成与无条件生成之间切换。实验表明，我们的模型具有很强的可控性，生成质量与基线相当。

2. 相关工作

2.1. 符号音乐生成中的控制

符号音乐生成中的控制涉及多个方面。在钢琴音乐领域，许多研究采用变分自编码器提取潜在表征，以指导生成，例如用于节奏的 EC^2-VAE [12]、用于和弦与纹理的 Polydis [3] ，以及 Music SketchNet

[13] 用于音高与节奏。然而，在多轨设置中，控制通常较为简化，例如通过标记 [1, 2, 14] 对乐器集合或全局属性进行编码，或提供部分音乐进行图像修复 [9, 8]。不过，直接且可解释的轨道级纹理控制仍未得到探索，因为现有的基于潜变量的方法 [3] 难以解释。同时，Draw and Listen [15] 和 Hyperscore [10]等系统通过将草图和曲线映射为抽象的音乐理念，凸显了直观可视化与交互的价值。基于这些见解，我们提出 ViTex，这是一种补充现有方法的表征，可视化多轨复调纹理，同时支持对生成任务进行直观控制。

2.2. 多轨符号音乐生成

近期的多轨符号音乐生成方法大多采用Transformer架构 [16] 来建模音符级标记序列，既可以直接对序列本身进行建模 [9]，也可以将音符标记与抽象简单音乐条件的控制标记 [1, 2, 14] 联合建模。相比之下，GETMusic [8], 采用基于扩散的方法，利用吸收态离散扩散 [17]，从任意源轨道生成任意目标轨道。一些工作聚焦于编曲 [18, 6, 7, 19], ，通常以包含音高信息的输入（例如旋律或和弦）为基础，并可选择性地纳入其他轨道的信息来指导编曲。然而，这些方法均无法在轨道级复调纹理条件下生成音乐；至多只能控制乐器是否出现 [1, 2, 8, 19]，或乐器的简单统计特征 [7, 6]。在实践中，作曲者往往更希望获得精确控制，例如指定低音在低音区演奏一条旋律线条，而不仅仅是使其出现。借助我们提出的ViTex捕捉分轨纹理，并结合和弦进行，我们的离散扩散模型能够同时控制纹理与和声，从而弥补现有方法的不足。

2.3. 扩散技术

扩散模型 [20] 已在各类生成任务中展现出卓越能力，并近期被扩展至符号音乐生成 [21]。借助无分类器引导[11] 实现条件控制，Polyffusion [4]能够操控和弦与纹理，而整曲生成 [5] 则采用级联扩散，在外部与内部音乐条件下生成完整作品。GETMusic [8] 进一步使用离散扩散 [17] 模型来处理多轨音乐。这些工作凸显了扩散在多轨音乐建模方面的潜力及其强大的可控性。受其启发，我们采用吸收态离散扩散框架 [17] 处理多轨符号音乐，并利用无分类器引导 [11] 融入我们的ViTex与基于和弦的控制。

3. 方法

3.1. 数据表示

我们将一段多轨符号音乐表示为张量，其中每个条目表示给定乐器在特定时间步和音高上的状态。每种状态取静音、起音、延音或MASK四个值之一。我们采用十六分音符分辨率，考虑包含128个音高、跨越11个乐器类别的8小节片段（32拍）。整体上，该乐曲可表示为张量 $x \in \{0,1,2,3\}^{128\times128\times11}$. 由于鼓是无固定音高的，鼓轨由形状为 128×1 的独立单通道钢琴卷帘表示。

ViTex 为每个轨道提供直观的表征，该表征源自一种基于规则的方法；这种方法还会生成机器可读的表征——对于每个轨道，我们通过汇总钢琴卷帘中互不重叠的 16×16 区域，构建低分辨率空间特征图；每个区域在时间轴上对应一个节拍，在音高轴上对应 16 个半音。

每个区域的纹理由对其中音符进行的基于规则的分析确定。受 [22], 启发，我们从三个感知维度刻画音乐织体：旋律、和声和延音。为区分旋律纹理与和声纹理，我们统计每个区域内同时发声的音符数量：包含同时发声音符的区域标记为和声纹理，其余区域标记为旋律纹理。延音则通过计算每个区域内的非静音比例来确定，该比例定义为至少有一个音符发声的时长占区域时长（一拍）的比例。根据经验观察，若该比例低于 0.3，则将该区域标记为延音纹理。

因此，每条轨道都会生成一张 $8\times8\times3$ 纹理特征图，并沿轨道轴拼接 11 条轨道，得到 $y_{\text{ins}} \in \{0,1\}^{8\times8\times33}$，从而编码多轨乐曲中各轨道的乐器配置。

这种设计兼顾直观的可视化效果和紧凑的机器可读表示。为了便于可视化，每条轨道在整个时间轴上都采用一致的颜色。旋律区域以曲线呈现，和声区域则以矩形呈现。非延音区域用虚线符号标示，以区别于延音纹理。图1展示了一个可视化示例。第3至第4小节的红色小号轨道展示了非延音与延音旋律纹理之间的对比。相较之下，绿色风琴轨道和蓝色钢琴轨道则凸显了和声纹理中相应的区别。除了用于可视化，得到的 ViTex 特征图也用作条件输入，以引导生成模型。

我们用一个包含 12 个音级和 3 个通道的 12×3 特征向量表示和弦，这 3 个通道分别编码根音、音级色彩和低音。以逐拍分辨率提取后，可得到和弦表示 $y_{\text{chd}} \in \{0,1\}^{32\times12\times3}$。

鼓轨的处理方式与有音高乐器轨不同。根据固定且经经验预先定义的映射关系，每个鼓音高都会被映射到三个频段之一：低频、中频或高频。对于每个频段，我们通过描述是否出现鼓点（用二元指示变量编码）以及其活动是否随时间持续，来考察其节奏纹理。后者直接沿用上文定义的持续判据；在鼓轨的语境中，它对应于密集或稀疏的节奏模式。由此得到鼓轨的 $8\times3\times2$ ViTex 特征图。

3.2. 离散扩散模型

遵循 GETMusic [8], ，我们采用吸收态离散扩散 [17]。对于具有 K 种可能状态的分类变量，我们用 x 表示其取值，并用列向量 $v(x)$表示其独热表示。前向过程由转移矩阵 $Q_t \in \mathbb{R}^{K\times K}$定义，其中每个元素 $[Q_t]_{mn} = q(x_t = m \mid x_{t-1} = n)$表示在步骤 t中从状态 n 转移到状态 m 的转移概率。我们采用吸收态扩散，在每个时间步 t，每个元素要么以概率 $1-\beta_t$ 保持不变，要么转移至 MASK。β_t 是破坏率，采用余弦调度。

边缘分布和后验分布为

$$q(x_t \mid x_0) = v(x_t)^\top \hat{Q}_t\, v(x_0), \quad \text{where } \hat{Q}_t = \prod_{i=1}^{t} Q_i. \tag{1}$$

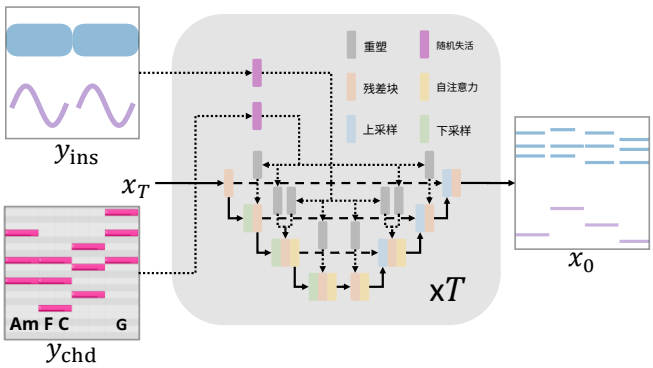


图 2：展示了 UNet 架构，以及 ViTex 和和弦进行提供的可选控制。

$$q(x_{t-1} \mid x_t, x_0) = \frac{q(x_t \mid x_{t-1}, x_0)\, q(x_{t-1} \mid x_0)}{q(x_t \mid x_0)} = \frac{\big(v(x_t)^\top Q_t v(x_{t-1})\big)\big(v(x_{t-1})^\top \hat{Q}_{t-1} v(x_0)\big)}{v(x_t)^\top \hat{Q}_t\, v(x_0)}. \tag{2}$$

反向转移的参数化形式为

$$p_\theta(x_{t-1} \mid x_t) = \sum_{\hat{x}_0} q(x_{t-1} \mid x_t, \hat{x}_0)\, p_\theta(\hat{x}_0 \mid x_t). \tag{3}$$

我们使用神经网络 $\epsilon_\theta(x_t, t)$ 来预测分布 $p_\theta(\hat{x}_0 \mid x_t)$ 的对数几率，并直接优化 x_0 预测损失

$$\mathcal{L}_{\text{pred}} = -\mathbb{E}_{t\sim \mathrm{U}(1,T),\, q(x_0)q(x_t|x_0)}[\log p_\theta(x_0 \mid x_t)]. \tag{4}$$

3.3. 引导机制

我们采用无分类器引导 [11] 来支持基于 ViTex 和和弦的条件生成，以及无条件生成，这需要对上述公式进行修改。我们的模型接收两个额外输入，得到 $\epsilon_\theta(x_t, t, y_1, y_2)$，其中 y_1 和 y_2 分别对应于 y_{ins} 和 y_{chd}，训练期间以 0.5 的概率将其替换为空值。因此，通过将 $p_\theta(x_0 \mid x_t)$ 替换为 $p_\theta(x_0 \mid x_t, y_1, y_2)$，对公式 4 中的训练目标进行相应调整。

推理期间，我们在条件预测和无条件预测之间进行插值。具体而言，给定乐器编制 y_{ins}和和弦进行 y_{chd}，以及相应的引导强度λ_{ins} 和 λ_{chd}，引导按如下方式应用[1]

$$\begin{aligned}\hat{\epsilon}_\theta(y_{\text{ins}}, y_{\text{chd}}) = \epsilon_\theta(\varnothing, \varnothing) &+ \lambda_{\text{ins}}\Big(\epsilon_\theta(y_{\text{ins}}, \varnothing) - \epsilon_\theta(\varnothing, \varnothing)\Big) \\ &+ \lambda_{\text{chd}}\Big(\epsilon_\theta(\varnothing, y_{\text{chd}}) - \epsilon_\theta(\varnothing, \varnothing)\Big) \\ &+ \lambda_{\text{ins}}\lambda_{\text{chd}}\Big(\epsilon_\theta(y_{\text{ins}}, y_{\text{chd}}) - \epsilon_\theta(y_{\text{ins}}, \varnothing) \\ &\quad - \epsilon_\theta(\varnothing, y_{\text{chd}}) + \epsilon_\theta(\varnothing, \varnothing)\Big).\end{aligned} \tag{5}$$

给定 x_t，我们首先使用公式 5 中的引导公式计算预测分布$\hat{x}_0$ 的对数几率，并采样得到一个$\hat{x}_0$。如果应用图像修复，[23], 则会将采样得到的$\hat{x}_0$ 部分替换为掩码区域中的已知值。最后，从公式 2 中的后验分布采样得到 x_{t-1}，完成公式 3 所述的单个反向扩散步骤。

[1]为简洁起见，我们在符号表示中省略 x_t 和 t。

3.4. UNet 架构

为了对 ϵ_θ（x_t，t，y_1，y_2）进行建模，我们采用标准的 UNet [24]，并使用自注意力 [16]处理 x_t和 t。条件输入通过逐元素相加注入：将由 y_1 和 y_2 得到的特征图，与每个 UNet 层中对应的潜在特征图相加。由于 y_1 和 y_2 的特征图最初与 UNet 潜在特征图的形状不匹配，我们首先将它们输入一个形状调整模块，其中宽度维度通过直接拉伸进行对齐，而高度和通道维度则通过一个小型多层感知机变换，以匹配对应的 UNet 特征图。完成对齐后，每一层的条件特征图即可与 UNet 潜在特征图逐元素相加。我们的 UNet 有四个层级，其中 y_1 注入底部两个层级，y_2 注入中间两个层级。图 2 展示了模型的架构。

4. 实验

本节将 ViTex 与现有方法进行比较，以突出其两项关键优势：（1）与现有控制方法相比，它能更精细地控制音乐织体；（2）其生成质量可媲美最先进的多轨模型。

4.1. 数据集与训练详情

实验采用经过筛选的 Lakh MIDI [25]和 Meta MIDI [26]数据集子集。我们保留四四拍、速度在 110 至 130 每分钟节拍数之间、没有调性变化、长度至少为 40 小节、至少有五条活跃轨道且涵盖三种以上乐器类别、音符数超过 50，并包含鼓以及钢琴、吉他或低音中至少一种的乐曲，同时剔除零时值音符过多的文件。最终得到 7,175 首乐曲，并按乐曲划分为 90% 的训练集和 10% 的测试集。数据使用 MusPy [27]处理。模型采用 AdamW 进行训练，学习率为 3e-4，批次大小为 100，并在 4 个 H100 GPU 上使用余弦衰减计划。

4.2. 基线

我们使用 Q&A [6] 作为条件生成的基线。Q&A 以复调乐谱作为输入，使我们能够提供和弦进行，并为每个轨道提供音高分布和音符密度，这些信息被编码为音轨纹理的潜在表示。为了使我们的 ViTex 与 Q&A 兼容并实现逐音轨纹理控制，我们设计了两种转换策略：

1. Q&A-1（严格）：直接将和弦用作复调乐谱；从基于 ViTex 和和弦进行生成的模拟多音轨乐曲中提取逐音轨的音高分布和音符密度。2. Q&A-2（宽松）：以 Q&A-1 为基础，添加随机的非和弦音，并对所得的音高分布和音符密度进行平滑处理。

这两个变体构成了我们的条件生成基线。我们认为，这种选择和改编是合理的，因为我们的模型支持非常精细的逐音轨纹理控制；据我们所知，Q&A 是在控制粒度方面最接近的可用基线。

我们使用预期音乐变换器（AMT）[9] 和多音轨音乐变换器（MMT）[1] 作为无条件生成和续写任务的基线。此外，MMT 支持指定应包含哪些轨道，因此适合作为条件生成任务的基线，并与 Q&A 一同用于主观聆听实验。

表1：条件生成的客观评估。

模型	乐器控制			和弦控制		质量
	配器准确度↑	音符时值分布重叠面积↑	起音间隔分布重叠面积↑	和弦准确度↑	音高分布重叠面积↑	编曲程度↑
问答-1[6]	0.584	0.135	0.451	0.607	0.450	0.188
问答-2[6]	0.299	0.082	0.110	0.043	0.253	0.111
Ours	**0.600**	**0.626**	**0.494**	**0.875**	**0.731**	**0.296**

表2：无条件生成的客观评估。*越接近真实值越好。对于 PCE 和 GPS，+ 和 - 表示偏差。

模型	PCE*	GPS*	编曲程度 ↑
真实值	1.741	0.804	0.303
MMT[1]	**+0.103**	+0.080	0.171
AMT[9]	-0.317	+0.174	0.278
Ours	-0.174	**+0.050**	**0.307**

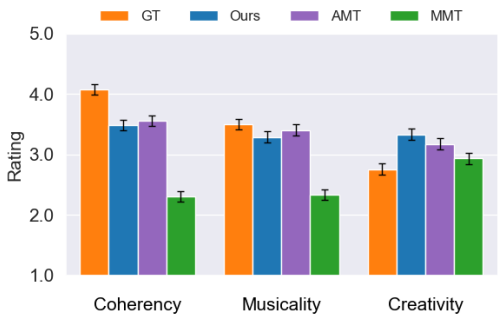
4.3. 客观评估

我们通过条件生成来评估可控性和质量。使用随机抽取的 1,000 个测试片段及其 ViTex 表示和和弦进行，我们分别通过模型、Q&A-1 和 Q&A-2 [6]生成结果。对于每个样本，我们计算生成结果与真实值的 ViTex 特征图之间的余弦相似度，以评估配器准确度（IA）；并计算生成结果与真实值的和弦进行之间的余弦相似度，以评估和弦准确度（CA）。此外，我们参照以往研究 [19]，评估音符时值（OAD）、起音间隔（OAIOI）和音高（OAP）分布的平均重叠面积。OAD、OAIOI 和 OAP 的值越大，表示生成样本与真实值在相应特征上的分布重叠程度越高，说明可控性有所提升。对于生成质量，我们采用文献 [7], 提出的编曲程度（DOA）指标；DOA 越高，表示编曲质量越好。如表1所示，我们的模型在可控性和质量两方面均优于基线，证明了其既能生成高质量音乐，又能有效遵循指定的控制条件。

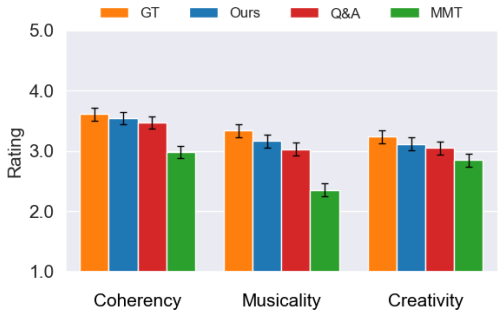
除条件生成外，我们还进行了无条件生成实验，以进一步评估生成样本的质量。对于每个模型，我们从头生成1,000个片段，并将其与从测试集中随机抽取的1,000个片段进行比较。我们计算DOA，其中数值越高表示编曲质量越好；此外，遵循先前的工作 [1, 28], , 我们还计算音级熵（PCE）以评估调性清晰度，并计算Grooving Pattern Similarity（GPS）以衡量节奏感，其中越接近真实值的数值越好。DOA报告绝对值，而PCE和GPS报告相对于真实值的偏差。如表2所示，我们的模型在GPS和DOA上优于基线模型，在PCE上略逊一筹；总体而言，该模型展现出从头生成高质量音乐的强大能力，表现可与现有方法相媲美。

4.4. 主观评估

我们邀请了 31 名参与者进行听力测试，以评估主观质量。研究包含两项任务：根据提示续写音乐，以及根据乐器生成音乐。在这两项任务中，参与者聆听了我们的模型、基线模型和真实值生成的输出，并按照以下标准对每个样本进行 5 点李克特量表评分：连贯性，即续写内容在多大程度上遵循提示的风格



（a）给定提示时音乐续写的结果。



（b）给定乐器编制时的音乐生成结果。

图3：主观评估结果，包括平均评分和被试内置信区间。

；音乐性，即音乐的整体质量和结构清晰度；以及样本的创造性。在续写任务中，MMT[1]和 AMT[9]作为基线；我们随机选取了测试集中的六首作品，并将每首作品的前两个小节用作提示，供各模型生成续写。在根据乐器生成音乐的任务中，MMT[1] 和 Q&A[6] 作为基线；我们随机选取了测试集中的四首作品，并向我们的模型和 Q&A 提供 ViTex 表示及和弦进行，而 MMT 仅接收乐器标签。

图3展示了结果，其中柱形高度代表平均评分，误差线表示通过无重复双因素方差分析计算得出的置信区间。在音乐续写任务中，我们的模型在连贯性和音乐性评分上略低于 AMT，但创造性评分更高，且所有指标均优于 MMT。这是合理的，因为 AMT 专门针对补全任务进行训练。我们的模型表现与 AMT 相当，说明在 ViTex 的引导下，它能够有效遵循提示，表明 ViTex 成功捕捉了乐器编制信息。在音乐生成任务中，我们的模型在所有评估指标上均优于所有基线模型，凸显了生成音乐的质量和可控性。

5. 结论

我们提出 ViTex，这是一种能够可视化分轨纹理并实现乐器编制新颖细粒度控制的表征。通过在离散扩散框架中将 ViTex 与和弦进行条件相结合，我们的方法支持条件式和无条件的多轨符号音乐生成，并能精确控制纹理。客观与主观评估表明，我们的模型在保持与最先进基线相当的生成质量的同时，具有很强的可控性。

6. 参考文献

[1] Hao-Wen Dong、Ke Chen、Shlomo Dubnov、Julian J. McAuley 和 Taylor Berg-Kirkpatrick，“多轨音乐变换器”，载于 *ICASSP 2023*，罗得岛，希腊，*2023*年*6*月*4*日至*10*日。2023年，第1–5页，IEEE。[2] Dimitri von R¨utte、Luca Biggio、Yannic Kilcher 和 Thomas Hofmann，“FIGARO：利用学习到的特征和专家特征实现可控音乐生成”，载于 *ICLR 2023*，基加利，卢旺达，*2023*年*5*月*1*日至*5*日。2023年，OpenReview.net。[3] Ziyu Wang、Dingsu Wang、Yixiao Zhang 和 Gus Xia，“学习用于可控复调音乐生成的可解释表征”，载于 *ISMIR 2020*，蒙特利尔，加拿大，*2020*年*10*月*11*日至*16*日，2020年，第662–669页。[4] Lejun Min、Junyan Jiang、Gus Xia 和 Jingwei Zhao，“Polyf-fusion：一种通过内部和外部控制生成复调乐谱的扩散模型”，载于 *ISMIR 2023*，米兰，意大利，*2023*年*11*月*5*日至*9*日，2023年，第231–238页。[5] Ziyu Wang、Lejun Min 和 Gus Xia，“使用级联扩散模型进行整曲层次化符号音乐生成”，载于 *ICLR 2024*，维也纳，奥地利，*2024*年*5*月*7*日至*11*日。2024年，OpenReview.net。[6] Jingwei Zhao、Gus Xia 和 Ye Wang，“问答：用于多轨符号音乐重新编曲的基于查询的表征学习”，载于 *IJCAI 2023*，*2023*年*8*月*19*日至*25*日，澳门，特别行政区，中国。2023年，第5878–5886页，ijcai.org。[7] Jingwei Zhao 和 Gus Xia，“Accomontage：通过乐句选择和风格迁移进行伴奏编配”，载于 *ISMIR 2021*，线上，*2021*年*11*月*7*日至*12*日，2021年，第833–840页。[8] Ang Lv、Xu Tan、Peiling Lu、Wei Ye、Shikun Zhang、Jiang Bian 和 Rui Yan，“GETMusic：利用统一表征和扩散框架生成任意音乐轨道”，*CoRR*，第abs/2305.10841卷，2023年。[9] John Thickstun、David Leo Wright Hall、Chris Donahue 和 Percy Liang，“预期音乐变换器”，*Trans. Mach. Learn. Res.*，第2024卷，2024年。[10] Morwaread Mary Farbood、Henry Kaufman 和 Kevin Jen-nings，“使用Hyperscore进行作曲：一种可视化音乐结构的直观界面”，载于 *ICMC 2007*，哥本哈根，丹麦，*2007*年*8*月*27*日至*31*日。2007年，Michigan Publishing。[11] Jonathan Ho 和 Tim Salimans，“无分类器扩散引导”，*CoRR*，第abs/2207.12598卷，2022年。[12] Ruihan Yang、Dingsu Wang、Ziyu Wang、Tianyao Chen、Jun-yan Jiang 和 Gus Xia，“通过潜在表征解耦实现深度音乐类比”，载于*ISMIR 2019*，代尔夫特，荷兰，*2019*年*11*月*4*日至*8*日，2019年，第596–603页。[13] Ke Chen、Cheng-i Wang、Taylor Berg-Kirkpatrick 和 Shlomo Dubnov，“Music SketchNet：通过音高与节奏的因子化表征实现可控音乐生成”，载于 *ISMIR 2020*，蒙特利尔，加拿大，*2020*年*10*月*11*日至*16*日，2020年，第77–84页。[14] Peiling Lu、Xin Xu、Chenfei Kang、Botao Yu、Chengyi Xing、Xu Tan 和 Jiang Bian，“Musecoco：从文本生成符号音乐”，*CoRR*，第abs/2306.00110卷，2023年。[15] Christodoulos Benetatos 和 Zhiyao Duan，“画出来，听起来！一种基于草图的音乐补全系统”，*Trans. Int. Soc. Music. Inf. Retr.*，第5卷，第1期，第141–155页，2022年。

[16] Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N. Gomez、Lukasz Kaiser 和 Illia Polosukhin，《注意力即你所需》，载于神经信息处理系统进展 *30*：*2017*年神经信息处理系统年度大会，*2017*年*12*月*9*日2017，5998–6008页美国加利福尼亚州长滩，*12*月。[17] Jacob Austin、Daniel D. Johnson、Jonathan Ho、Daniel Tarlow 和 Rianne van den Berg，《离散状态空间中的结构化去噪扩散模型》，载于*NeurIPS 2021*，*2021*年*12*月*6*日至*14*日，线上，2021年，第17981–17993页。[18] Hongyuan Zhu、Qi Liu、Nicholas Jing Yuan、Chuan Qin、Jiawei Li、Kun Zhang、Guang Zhou、Furu Wei、Yuanchun Xu 和 Enhong Chen，《小冰乐队：流行音乐旋律与编曲生成框架》，载于第*24*届 *ACM SIGKDD*知识发现与数据挖掘国际会议论文集，*KDD 2018*，英国伦敦，*2018*年*8*月*19*日至*23*日。2018年，第2837–2846页，ACM。[19] Yi Ren、Jinzheng He、Xu Tan、Tao Qin、Zhou Zhao 和 Tie-Yan Liu，《Popmag：流行音乐伴奏生成》，载于*MM ’20*：第*28*届 *ACM* 国际多媒体会议，线上活动／美国华盛顿州西雅图，*2020*年*10*月*12*日至*16*日。2020年，第1198–1206页，ACM。[20] Jonathan Ho、Ajay Jain 和 Pieter Abbeel，《去噪扩散概率模型》，载于神经信息处理系统进展 *33*：*2020*年神经信息处理系统年度大会（*NeurIPS 2020*），*2020*年*12*月*6*日至*12*日，线上，2020年。[21] Gautam Mittal、Jesse H. Engel、Curtis Hawthorne 和 Ian Simon，《使用扩散模型生成符号音乐》，载于*ISMIR 2021*，线上，*2021*年*11*月*7*日至*12*日，2021年，第468–475页。[22] Louis Couturier，用于计算机辅助分析与作曲的符号音乐纹理建模，博士论文，UPJV，2024年12月。[23] Andreas Lugmayr、Martin Danelljan、Andr´es Romero、Fisher Yu、Radu Timofte 和 Luc Van Gool，《RePaint：使用去噪扩散概率模型进行图像修复》，载于*CVPR 2022*，美国路易斯安那州新奥尔良，*2022*年*6*月*18*日至*24*日。2022年，第11451–11461页，IEEE。[24] 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——*MICCAI 2015*，第*18*届国际会议，德国慕尼黑，*2015*年*10*月*5*日至*9*日，论文集，第三部分。2015年，第9351卷计算机科学讲义，第234–241页，施普林格。[25] Colin Raffel，基于学习的序列比较方法及其在音频到 *MIDI* 对齐与匹配中的应用，博士论文，哥伦比亚大学，美国，2016年。[26] Jeffrey Ens 和 Philippe Pasquier，《构建 MetaMIDI数据集：关联符号音乐数据与音频音乐数据》，载于*ISMIR 2021*，线上，*2021*年*11*月*7*日至*12*日，2021年，第182–188页。[27] Hao-Wen Dong、Ke Chen、Julian J. McAuley 和 Taylor Berg-Kirkpatrick，《MusPy：符号音乐生成工具包》，载于 *ISMIR 2020*，*2020*年，2020年，*11*101–108加拿大蒙特利尔，*10*月，第101–108页。[28] Shih-Lun Wu 和 Yi-Hsuan Yang，《爵士变换器走在前沿：通过定量指标探索人工智能作曲的不足》，载于*ISMIR 2020*，加拿大蒙特利尔，*2020*年*10*月*11*日至*16*日，2020年，第142–149页。