arXiv:2602.03523v1 [cs.SD] 3 Feb 2026

D3PIA：一种基于主旋律谱生成钢琴伴奏的离散去噪扩散模型

Eunjin Choi、*Hounsu Kim*、*Hayeon Bang*、*Taegyun Kwon*、*Juhan Nam*

韩国科学技术院文化技术研究生院，大韩民国

摘要

在符号音乐领域生成钢琴伴奏是一项颇具挑战性的任务，需要根据给定的旋律和和弦约束（例如主旋律谱中提供的约束）创作完整的钢琴乐曲。本文提出一种基于离散扩散的钢琴伴奏生成模型 D3PIA，利用钢琴卷帘表示中主旋律谱与伴奏之间的局部对齐。D3PIA 采用邻域注意力（NA）对主旋律谱进行编码，并以其为条件预测钢琴伴奏中的音符状态。这种设计通过高效关注邻近的旋律和和弦条件，增强了局部上下文建模能力。我们使用广泛应用于钢琴伴奏生成基准测试的 POP909 数据集对模型进行评估。客观评估结果表明，与基于连续扩散和 Transformer 的基线模型相比，D3PIA 能更忠实地保留和弦条件。此外，主观聆听测试表明，与对比模型相比，D3PIA 生成的伴奏在音乐上更加连贯。

索引词— 钢琴伴奏生成、离散扩散、音乐信息检索

1. 引言

伴奏编曲是一项重要的音乐创作过程，需要根据给定的旋律创作伴奏。这一过程需要深入理解音乐理论，并兼顾风格和结构连贯性。在各种伴奏编曲中，钢琴伴奏在西方音乐中具有悠久的历史意义（例如，李斯特为艺术歌曲创作的钢琴改编曲），且由于钢琴拥有宽广的88个复音键音域，也带来了技术挑战。本文探讨在符号音乐领域中，如何根据给定旋律和和弦的主旋律谱自动生成钢琴伴奏。

过去十年间，符号音乐生成主要受到自然语言处理领域方法的推动。在这些方法中，音乐被表示为由音符事件构成的标记化序列，并以语言建模（LM）的方式训练模型。这一范式使基于 Transformer 的架构在众多研究中脱颖而出 [1, 2, 3]。尽管这些基于 LM 的方法在生成长序列方面表现出色，但它们需要复杂的符号音乐标记化过程。此外，推理阶段的自回归生成可能导致长序列中的误差累积，并限制对生成样本的灵活控制。

近年来，扩散模型已成为多个领域的强大工具，包括图像合成 [4], 音频 [5], 和音乐生成 [6], ，展现出捕捉复杂分布并生成高质量输出的卓越能力。

在符号音乐生成中，扩散模型已被直接应用于钢琴卷帘图像 [7, 8, 9, 10] ，或应用于变分自编码器（VAE）的潜在空间 [11, 12]。特别是，在扩散模型中采用钢琴卷帘，相较于基于语言模型的方法，能够提供更直观、更简单的表示。此前使用扩散模型生成钢琴音乐的研究采用了对钢琴卷帘进行连续扩散的方法 [7, 8, 10], ；而近期转录研究表明，离散扩散能够有效处理钢琴卷帘内在的离散（二元开/关）特性 [13]。由于扩散过程通过迭代添加或移除事件，离散形式天然支持编配操作（例如插入、删除或润色音符/声部）。这促使我们探究，相较于连续方法，离散扩散是否更适用于符号音乐。

本文提出 D3PIA，一种专为钢琴伴奏生成任务设计的基于离散扩散的模型。该模型由主旋律谱编码器和离散去噪解码器组成，后者直接在钢琴卷帘的离散音符状态上运行，并采用扩张 NA [14, 15]。基于 NA 的模型架构能够有效捕捉局部对齐的主旋律谱特征及其对应伴奏，从而增强依据和弦进行对钢琴伴奏的建模。通过与连续扩散及基于 Transformer 的基线进行全面比较，我们表明 D3PIA 对和声约束的遵循更强，并在主观评估中表现更优，尽管所需模型规模显著更小。消融研究进一步凸显了离散扩散与我们的模型设计对和弦连贯性和节奏一致性的贡献。模型训练代码与生成样本已公开提供 [1]。

2. 相关工作

2.1. 钢琴伴奏生成模型

自 POP909 [16], （一个中国流行歌曲钢琴伴奏数据集）发布以来，已有大量研究致力于从主旋律谱生成钢琴伴奏 [17, 18, 7, 8, 10]。其中，一种基于变分自编码器的方法 [17]通过解耦和弦与织体成分来生成伴奏，从而能够使用预训练的和弦与织体嵌入来调节生成模型 [7, 18, 8]。Accomontage 模型 [18] 采用基于模板的检索来生成伴奏。

Polyffusion [7]和 WholeSongGen [8], 都是基于钢琴卷帘连续扩散的无条件符号音乐生成模型，也可以根据主旋律谱生成钢琴伴奏。Polyffusion 利用预训练的和弦编码器 [17] ，结合无分类器引导来融入和弦条件。WholeSongGen

[1]https://jech2.github.io/D3PIA/

采用 Polyffusion 作为骨干网络，构建了一个四阶段级联扩散框架，从歌曲结构生成逐步推进到钢琴伴奏生成。FGG [10]，同样基于连续扩散，在钢琴卷帘中将和弦和起始信息作为无分类器引导，以实现细粒度条件控制。

除基于 POP909 的研究外，基于Transformer的方法也已应用于钢琴音乐生成 [1, 3]。其中，Compose & Embellish (C&E) 通过采用两阶段建模展现出显著的性能 [3]：先生成主旋律谱，再根据生成的主旋律谱完成钢琴伴奏。该模型采用类似 REMI 的表示 [2], ，在小节级别交错排列主旋律谱与伴奏。

尽管这些方法已成功应用于钢琴伴奏生成，但在保持与和弦条件的一致性、结构一致性及总体音乐质量方面仍面临挑战。本文采用离散扩散模型进行钢琴伴奏生成，并将其与基于连续扩散和基于Transformer的模型进行比较。

2.2. 离散扩散模型

离散扩散最初在 [19, 20, 21] 中提出，其通过转移矩阵将添加和移除噪声的连续过程改造为离散框架。该方法对系统状态进行操作，每次状态变化都由特定概率决定。已有少数研究将离散扩散用于符号音乐生成 [22, 23, 12]。这些研究将离散扩散应用于 VQ-VAE 的潜空间 [12], 、每个时间步的多热音乐事件标记向量 [22], ，以及每个音符的音高和时值复合标记 [23]。要将离散扩散应用于钢琴卷帘表示，可以将钢琴卷帘视为图像并在 VQ-VAE 潜空间中进行操作 [12] ，也可以为钢琴卷帘的每个像素定义音符状态并逐一去噪 [13]。后一种方法名为 D3RM，已证明将扩散过程应用于钢琴卷帘中的离散音符状态，对钢琴转录有效。该方法在去噪模块中使用堆叠式NA，从而能够高效处理高分辨率钢琴卷帘，同时将局部音频特征与转录出的音符对齐。本文采用 D3RM 作为骨干模型，利用其在局部性感知注意力和离散状态预测方面已验证的有效性。

3. 方法

3.1. 数据表示

钢琴卷帘表示能够有效捕捉垂直维度（和声）和水平维度（节奏）上的局部关系。遵循先前的研究 [7, 8, 10], ，我们采用十六分音符时间单位表示钢琴卷帘。对于音高，我们使用 88 个值，分别对应从 21 到 108 的 MIDI音符。我们将钢琴卷帘中的每个值标记为四种状态之一：起始、关闭、延音和*MASK*。对于主旋律谱，我们将旋律和和弦标签合并到单个钢琴卷帘通道中：将所有和弦标签转换为从 C4 开始的 MIDI音高，并向下移两个八度，以避免与旋律重叠。

3.2. 模型架构

我们提出的模型 D3PIA 由一个编码器和一个去噪解码器组成，如图1所示。编码器接收主旋律谱的钢琴卷帘作为输入，其输出特征会与带噪伴奏钢琴卷帘沿通道维度拼接，然后

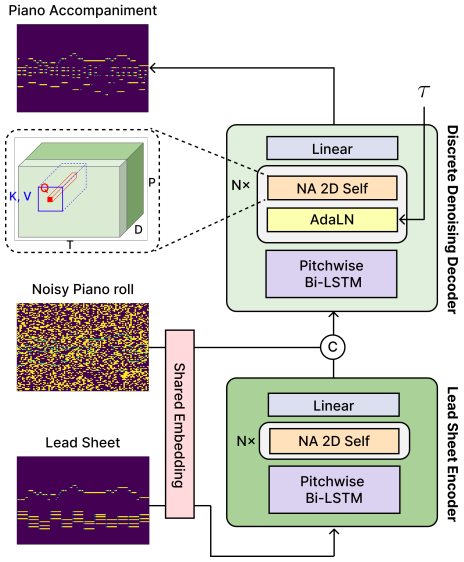


图 1：D3PIA 模型结构。编码器通过注意力处理主旋律谱信息，随后将编码器输出（深绿色部分）与带噪声的钢琴卷帘拼接（记为 C），并通过 NA 在去噪解码器中进一步进行注意力处理。N 表示 NA层的数量。在 NA二维自注意力模块中，T、D 和 P 分别表示时间步、嵌入维度和音高；Q、K 和 V 分别表示注意力计算中使用的查询、键和值。 τ 表示扩散时间步。

输入去噪解码器。编码器和解码器都包含一个逐音高双向LSTM层，用于捕捉各音高内部的时间变化；此外，还包含 NA 二维自注意力块，其后接一个线性层。两者的主要区别在于，解码器加入了自适应层归一化（AdaLN）层，用于调节扩散时间步。去噪解码器的嵌入层与编码器共享，从而利用带噪输入和主旋律谱条件具有相同的钢琴卷帘结构这一特点。使用 NA 二维编码器处理主旋律谱条件，可以有效捕捉和弦进行与旋律之间的局部垂直关系，同时保留其钢琴卷帘形状。在训练过程中，模型同时接收带噪伴奏输入和主旋律谱条件；在推理时，仅提供主旋律谱条件，模型通过迭代去噪生成伴奏。

3.3. 钢琴伴奏的离散扩散

我们将钢琴卷帘中的每个元素转换为独热向量，得到$\mathbf{y} \in \mathbb{R}^{T\times 88\times 4}$，其中 T 是帧长度，88 表示 MIDI音符的数量，4 表示上述音符状态的数量。前向过程通过将以下转移矩阵 $[\mathbf{Q}_\tau]_{ij} = q(y_\tau^{tp} = j|y_{\tau-1}^{tp} = i) \in \mathbb{R}^{4\times 4}$ 乘至最后一个维度，逐步对钢琴卷帘中每个元素的状态进行掩码处理或扰动，

$$\mathbf{Q}_\tau = \begin{bmatrix} \alpha_\tau + \beta_\tau & \beta_\tau & \beta_\tau & 0 \\ \beta_\tau & \alpha_\tau + \beta_\tau & \beta_\tau & 0 \\ \beta_\tau & \beta_\tau & \alpha_\tau + \beta_\tau & 0 \\ \gamma_\tau & \gamma_\tau & \gamma_\tau & 1 \end{bmatrix} \tag{1}$$

其中 y_τ^{tp}表示标签$\mathbf{y}$ 在时间位置 $t \in\{0, 1, ..., T-1\}$、音高 $p \in \{0, 1, ..., 87\}$ 以及扩散时间步的状态

$\tau \in \{0, 1, ..., \mathcal{T}-1\}$； α_τ、 β_τ和 γ_τ分别是保持、扰动和掩蔽该状态的概率。反向过程是预测反向概率分布$p_\theta(\mathbf{y}_{\tau-1}|\mathbf{y}_\tau, \mathbf{x})$，其中$\mathbf{x}$是表示为相同形状的钢琴主旋律谱，$\mathbf{x} \in \mathbb{R}^{T\times 88\times 4}$。训练目标描述如下：

$$\begin{aligned}\mathcal{L}_{vlb} &= D_{KL}\left[q(\mathbf{y}_\mathcal{T}|\mathbf{y}_0)||p(\mathbf{y}_\mathcal{T})\right] \\ &+ \sum_{\tau=1}^{\mathcal{T}}\left[D_{KL}[q(\mathbf{y}_{\tau-1}|\mathbf{y}_\tau, \mathbf{y}_0)||p_\theta(\mathbf{y}_{\tau-1}|\mathbf{y}_\tau, \mathbf{x})]\right]\end{aligned} \tag{2}$$

我们额外实现了来自 [13], 的吸收态（AS）采样方法，该方法已被证明能够增强离散扩散模型在钢琴转录任务中的细化能力——这是一项要求极高精度的判别式任务。模型按照第4.2节所述的调度进行训练，以逆转扰动和掩码；在推理过程中，它在所有扩散时间步中均以 $\beta_t = 0$ 生成样本。

4. 实验

4.1. 数据集

我们使用了POP909 [16], ，这是一个广泛使用的钢琴伴奏生成基准，包含909个MIDI文件。我们将数据集按8:1:1的训练集:验证集:测试集比例进行划分，为方便起见将测试集称为POP909-test。在训练时，我们将片段随机裁剪为8小节长度，并应用-5到6个半音的随机音高移调进行数据增强。在评估时，我们使用了包含86首乐曲的完整POP909-test。为确保我们的基于扩散的模型（能够生成8小节片段）与基于Transformer的模型之间的公平比较，我们将POP909-test切分为8小节片段，共得到905个片段。

4.2. 模型配置

每个钢琴卷帘状态都表示为一个 4 维嵌入向量。与最初为钢琴转录任务设计的 D3RM 模型不同，D3PIA 是一种生成模型，需要更强的模型容量，能够感知更大的输入上下文，并在更大的数据集上进行训练。因此，我们通过增加解码器的层规模，以及每层扩张 NA 的窗口大小来扩大模型，从而构建更大的感受野。根据初步研究，我们将层数设为 10，并将扩张窗口大小设为 5 [1, 2, 4, 8, , 16, 1, 2, 4, 8, 16] 8 扩张率设为 。模型使用批量大小为 的配置，在单个 NVIDIA A6000 GPU 上进行训练。训练进行了 200k 步，此后验证集上的音符级F1分数趋于稳定。扩散过程中，我们使用了 $T = 100$ 个时间步。每一步中，α_t、 γ_t 和 β_t 按照 [24, 13] 进行调度。辅助损失权重设为 $\lambda = 5.0 \times 10^{-4}$。使用 AdamW 进行优化，参数为 β =(0.9,0.96)。当扩散损失连续 25k 步未改善时，初始学习率 1.0×10^{-3} 将乘以 0.8。

4.3. 对比模型

我们将模型性能与基于扩散的模型，以及一种同时接受旋律和和弦条件的基于Transformer的模型进行了比较。对于基于扩散的模型，我们选取了WholeSongGen（WSG-4th）的Polyffusion[7], 4阶段模型 [8], 以及FGG [10] 进行比较。对于基于Transformer的模型，我们选取了C&E中Embellish部分（C&E-E）。由于Accomontage [18] 是在不同的POP909数据集划分上训练的，且没有可用的官方训练代码，因此未纳入比较。此外，由于GETMusic [23], 是一个无法获取检查点的大规模模型，我们也无法将其纳入比较。

模型	Base	模型输入
Polyffusion [7]	C	旋律、和弦（编码的色度向量 [17]）
WSG-4th [8]	C	旋律、和弦（钢琴卷帘）、歌曲结构、简化主旋律谱
FGG [10]	C	旋律、和弦（钢琴卷帘；扩展至完整八度）、起始
C&E-E [3]	T	旋律、和弦（标签作为词元）
D3PIA（我们的方法）	D	旋律、和弦（钢琴卷帘）

表 1：各方法在模型架构和实验所用输入方面的比较。C、D 分别表示连续扩散和离散扩散，T 表示 Transformer。

对于 C&E-E，我们复现了该模型，并在不使用速度标记的情况下对其进行训练。对于其他所有模型，我们使用官方代码，仅做了最少的修改，以尽可能确保比较公平；对于 FGG，我们将原始的 4 小节输入长度改为 8 小节。我们将 Polyffusion 和 C&E-E 视为直接对比模型，并指出，由于部分模型除旋律和和弦外还需要额外输入，因此无法将我们的模型与所有对比模型进行公平比较，如表 1 所示。

由于 POP909 数据集未提供官方划分，现有模型采用了不同的随机划分。因此，我们使用自己的划分重新训练了所有模型。所有模型均通过随机音高移调进行数据增强。根据评估指标，我们确认，若提供官方预训练检查点，重新训练的模型所达到的生成质量与其相近。

4.4. 评估指标

4.4.1. 客观评估

如第1节所述，伴奏的首要目标是在保持节奏模式一致的同时，跟随给定的旋律和和弦条件。鉴于此，我们评估了生成伴奏的和声连贯性和节奏一致性。为评估和声连贯性，我们采用调外音比例（**OOK**） [10], 、和弦准确率（**CA**） [25, 10], 和和弦相似度（**CS**） [10]。OOK 的计算方法是根据 POP909 的调号标签统计调外音的数量，并对统计结果进行归一化。对于 CA 和 CS，我们首先使用 Dai 等人提出的基于规则的方法从生成样本中提取和弦 [26], 再与 POP909-test 的参考数据进行比较：CA 统计每拍完全匹配的情况，而 CS 则衡量预训练和弦编码器 [17], 生成的双小节和弦嵌入之间的余弦相似度，从而将和弦音的重叠也计入评价。由于 CA 无法捕捉这类重叠，我们还额外报告了 CS。对于节奏一致性，我们采用律动模式相似度（**GS**） [27], ，该指标衡量一个片段内各小节二值起音向量两两之间的相似度。

4.4.2. 主观评估

我们进行了聆听测试以评估音乐质量。我们从 POP909-test 中随机选取了 10 个 8 小节片段，并根据相应的主旋律谱，由每个模型生成伴奏。所有刺激均经过匿名处理，并以随机顺序呈现。十三名参与者（每人至少拥有钢琴或作曲学士学位）在 5 点李克特量表上给出了平均意见分。聆听标准如下：

连贯性与给定的旋律连贯吗?

模型	参数量	推理时间（秒）	和声			节奏
			OOK (%)	和弦准确率 (%)	和弦相似度 (%)	律动模式相似度 (%)
GT	-	-	**0.0**	91.6	95.7	82.7
Polyffusion	41.1M	21.4	**0.0**	37.5	54.0	79.9
C&E-E	66.0M	18.7	14.8	58.1	70.6	80.8
D3PIA	**2.2M**	1.7	**0.0**	80.1	93.6	**82.1**
WSG-4th*	41.6M	79.0	2.4	**87.6**	**94.6**	75.4
FGG*	36.7M	**0.4**	**0.0**	62.0	77.3	78.9

表 2：POP909-test 上的客观结果。OOK（调外音符比例，越低越好）、和弦准确率（越高越好）、和弦相似度（越高越好）、GS（律动模式相似度，越接近真实值越好）、OA（重叠面积，越高越好）。最佳结果以**粗体**显示，次佳结果加下划线。* 表示该模型融合了额外输入，如表 1 所示。推理时间以一个 8 小节样本为单位报告。

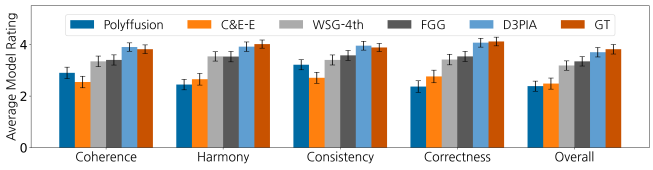


图 2：主观评估结果，置信度为 95% 程。

和声是否遵循和弦进行；是否缺少和弦音？**一致性**节奏一致性是否得到保持？**正确性**是否存在不和谐音、不自然的节奏或别扭的乐句？**总体**总体音乐质量如何?

5. 结果与讨论

5.1. 比较结果

表 2 展示了在 POP909-test 上的客观结果。D3PIA 和 WSG-4 取得了最高的和弦准确率和和弦相似度指标，且其和弦相似度值与真实值相当。对于 C&E-E，其和弦条件一致性较低，但与其他基于扩散的模型相当，这表明基于Transformer的模型在和弦标签信息极少的情况下，仍能在一定程度上预测和弦音高。然而，它们频繁产生调外音。在节奏方面，D3PIA 取得了高于其他模型的律动模式相似度，表明其节奏型比其他模型更为一致。在以往的音乐生成研究中， [27, 3], 更接近参考集的律动模式相似度值被认为更好。C&E-E 取得了次优的律动模式相似度值，这在意料之中，因为基于Transformer的模型具有较长的感受野范围，有助于节奏连贯性。

在主观聆听测试中，如图 2 所示，D3PIA 的表现优于所有对比模型，包括主要对比模型 Polyffusion 和 C&E-E。值得注意的是，WSG-4th 的表现尤为突出：其和弦准确率（CA）和和弦相似度（CS）得分高于 D3PIA，但在主观评估的和声指标上并未胜出。进一步分析聆听测试样本发现，尽管 CS 值非常高（>0.95），仍有若干 WSG-4th 输出的和声评分约为 3 [2]。我们认为，这种差异源于 WSG-4th 生成的不协和音符所带来的听感，这与表 2 中的调外音结果一致。我们的主观评估同时考量了连贯性和正确性，因此此类调外错误可能会对得分产生显著影响。

[2]我们邀请读者在补充网页中查看这些示例：https://jech2.github.io/D3PIA/

模型	和声		节奏
	和弦准确率 (%)	和弦相似度 (%)	律动模式相似度 (%)
原始D3RM解码器	79.9	78.0	74.9
带升调的D3RM解码器	79.8	78.0	77.1
不使用编码器中的和弦的D3PIA	35.9	59.2	83.9
不使用AS采样的D3PIA	76.8	77.2	79.6
D3PIA	**80.1**	**93.6**	**82.1**

表3：D3PIA模型的消融研究。所有模型的OOK均为0%。

此外，我们在表 2 中报告了生成一个 8 小节样本的推理时间。D3PIA 的推理速度显著快于其他模型（1.7 秒），能够实现实时生成，但仍慢于采用 DDIM采样的 FGG。

5.2. 消融研究

当我们聆听生成样本时，我们观察到D3PIA产生不协和音符的频率低于其他模型。我们推测这一改进源于使用了基于NA的主旋律谱编码器。另一种可能的解释是离散扩散模型的音符精炼能力，如 [13]中所讨论的。为了探究这些因素，我们进行了消融研究（表 3）。

首先，我们比较了在不使用编码器的情况下将主旋律谱直接拼接到原始D3RM解码器的模型。结果表明，使用原始D3RM解码器（仅477K参数）在无调外音的情况下取得了相对较高的和弦准确率和和弦相似度，数值约为80%，仍优于Polyffusion和C&E-E。这表明离散扩散架构的有效性。此外，在将模型扩展至D3PIA配置后，观察到额外的律动模式相似度增益。然而，这些模型在和声连贯性和节奏一致性方面的表现不如D3PIA，凸显了所提出的基于NA的主旋律谱编码器的重要性。接下来，从编码器中排除和弦导致和弦相似度大幅下降（至59.2%），证明了编码器中和弦输入条件化的有效性。此外，移除AS采样使和弦相似度相较完整D3PIA降低了约16.4%。这意味着离散扩散模型中的精炼过程（通过AS采样）会修正音符状态，从而在给定主旋律谱的情况下生成更连贯的伴奏。

6. 结论

与通常优先考虑创造性和多样性的一般音乐生成任务不同，伴奏生成的关键在于可靠地遵循旋律和和弦进行。在本研究中，我们强调了两点主要启示。第一，在离散音符状态层面进行建模，对符号音乐生成非常有效。第二，根据给定的旋律和和弦进行生成伴奏，需要具备强大的局部依赖能力，而这应当在模型设计中得到明确体现。严格遵循和弦音可能会降低多样性，但能够避免听感上不协和的音符，并增强可控性；这些对于实用的伴奏系统至关重要。

D3PIA 的一个主要局限在于其离散化的音符表示方式，无法对力度进行建模。未来的研究可以扩展音符状态空间，或纳入力度预测模型 [28]。另一个值得探索的方向是在遵循给定条件的同时，生成更多样且更具艺术性的伴奏。这可以通过扩大模型规模和输入长度，并利用预训练纹理嵌入实现风格变化，从而增强可控性 [17]。

7. 致谢

本研究获得了韩国国家研究基金会（NRF）资助，资助来源为韩国政府（科学技术信息通信部），项目编号为 RS-2023-NR077289 和 RS-2024-00358448。

8. 参考文献

[1] C.-Z. A. Huang、A. Vaswani、J. Uszkoreit、N. Shazeer、I. Simon、C. Hawthorne、A. M. Dai、M. D. Hoffman、M. Dinculescu 和 D. Eck，《音乐 Transformer》，发表于国际学习表征会议论文集，2018年。[2] Y.-S. Huang 和 Y.-H. Yang，《流行音乐 Transformer：基于节拍的表现力流行钢琴作品建模与生成》，发表于第*28*届 *ACM* 国际多媒体会议论文集，2020年，第 1180–1188 页。[3] S.-L. Wu 和 Y.-H. Yang，《Compose & Embellish：通过两阶段方法生成结构良好的钢琴演奏》，发表于国际声学、语音与信号处理会议论文集，2023年。[4] R. Rombach、A. Blattmann、D. Lorenz、P. Esser 和 B. Ommer，《利用潜在扩散模型进行高分辨率图像合成》，发表于*IEEE/CVF* 计算机视觉与模式识别会议论文集，2022年。[5] H. Liu、Y. Yuan、X. Liu、X. Mei、Q. Kong、Q. Tian、Y. Wang、W. Wang、Y. Wang 和 M. D. Plumbley，《 AudioLDM 2：通过自监督预训练学习整体音频生成》，发表于*IEEE/ACM* 音频、语音与语言处理汇刊，第 32 卷，第 2871–2883 页，2024年。[6] Z. Evans、J. D. Parker、C. J. Carr、Z. Zukowski、J. Taylor 和 J. Pons，《利用潜在扩散进行长篇音乐生成》，发表于第*25*届国际音乐信息检索学会会议论文集，2024年。[7] L. Min、J. Jiang、G. Xia 和 J. Zhao，《Polyffusion：一种带有内部和外部控制的复调乐谱生成扩散模型》，发表于第*24*届国际音乐信息检索学会会议论文集，2023年。[8] Z. Wang、L. Min 和 G. Xia，《使用级联扩散模型进行整首歌曲的层级符号音乐生成》，发表于第*12*届国际学习表征会议，2023年。[9] Q. Huang、D. S. Park、T. Wang、T. I. Denk、A. Ly、N. Chen、Z. Zhang、Z. Zhang、J. Yu、C. Frank、J. Engel、Q. V. Le、W. Chan、Z. Chen 和 W. Han，《Noise2Music：文本条件扩散模型音乐生成》，*arXiv* 预印本：*2302.03917*，2023年。[10] T. Zhu、H. Liu、Z. Wang、Z. Jiang 和 Z. Zheng，《用于基于扩散模型的符号音乐生成的高效细粒度引导》，发表于国际机器学习会议，*PMLR*，2025年。[11] G. Mittal、J. Engel、C. Hawthorne 和 I. Simon，《基于扩散模型的符号音乐生成》，发表于第*22*届国际音乐信息检索学会会议论文集，2021年。[12] J. Zhang、G. Fazekas 和 C. Saitis，《使用向量量化离散扩散模型进行作曲家风格特定的符号音乐生成》，发表于*2024*年 *IEEE* 第*34*届信号处理机器学习国际研讨会论文集，2024年。[13] H. Kim、T. Kwon 和 J. Nam，《D3RM：用于钢琴转录的离散去噪扩散精炼模型》，发表于国际声学、语音与信号处理会议论文集，2025年。[14] A. Hassani、S. Walton、J. Li、S. Li 和 H. Shi，《邻域注意力变换器》，发表于*IEEE/CVF* 计算机视觉与模式识别会议论文集，2023年。[15] A. Hassani 和 H. Shi，《膨胀邻域注意力变换器》，*arXiv* 预印本：*2209.15001*，2022年。[16] Z.* Wang、K.* Chen、J. Jiang、Y. Zhang、M. Xu、S. Dai、G. Bin 和 G. Xia，《POP909：用于音乐编配生成的流行歌曲数据集》，发表于第*21*届国际音乐信息检索学会会议论文集，2020年。[17] Z. Wang、D. Wang、Y. Zhang 和 G. Xia，《学习用于可控复调音乐生成的可解释表示》，发表于第*21*届国际音乐信息检索学会会议论文集，2020年。[18] J. Zhao 和 G. Xia，《Accomontage：通过乐句选择和风格迁移进行伴奏编配》，发表于第*22*届国际音乐信息检索学会会议论文集，2021年。[19] J. Sohl-Dickstein、E. A. Weiss、N. Maheswaranathan 和 S. Ganguli，《利用非平衡热力学进行深度无监督学习》，发表于第*32*届国际机器学习会议论文集，2015年。[20] E. Hoogeboom、D. Nielsen、P. Jaini、P. Forr´e 和 M. Welling，《Argmax流与多项式扩散：学习类别分布》，发表于神经信息处理系统进展，2021年。[21] J. Austin、D. D. Johnson、J. Ho、D. Tarlow 和 R. Van Den Berg，《离散状态空间中的结构化去噪扩散模型》，发表于第*35*届神经信息处理系统会议论文集，2021年。[22] M. Plasser、S. Peter 和 G. Widmer，《用于符号音乐生成的离散扩散概率模型》，发表于第三十二届国际人工智能联合会议论文集，2023年。[23] A. Lv、X. Tan、P. Lu、W. Ye、S. Zhang、J. Bian 和 R. Yan，《GETMusic：利用统一表示和扩散框架生成任意音乐音轨》，*arXiv* 预印本：*2305.10841*，2023年。[24] S. Gu、D. Chen、J. Bao、F. Wen、B. Zhang、D. Chen、L. Yuan 和 B. Guo，《用于文本到图像合成的向量量化扩散模型》，发表于*IEEE/CVF* 计算机视觉与模式识别会议论文集，2022年。[25] Yi Ren、Jinzheng He、Xu Tan、Tao Qin、Zhou Zhao 和 Tie-Yan Liu，《PopMAG：流行音乐伴奏生成》，发表于第*28*届 *ACM* 国际多媒体会议论文集，2020年。[26] S. Dai、H. Zhang 和 R. B. Dannenberg，《流行音乐中层级结构对旋律、节奏与和声的自动分析及影响》，发表于*2020*年人工智能音乐创作联合会议（*CSMC-MuMe 2020*）论文集，2020年。[27] S.-L. Wu 和 Y.-H. Yang，《走在前沿的爵士变换器：通过定量测量探索人工智能创作音乐的不足》，发表于第*21*届国际音乐信息检索学会会议论文集，2020年。[28] H. Kim 和 X. Serra，《利用带注意力和 FiLM 的 U-Net 进行钢琴演奏 MIDI 力度估计的方法》，发表于第*25*届国际音乐信息检索学会会议论文集，2024年。