arXiv:2601.15872v1 [cs.SD] 22 Jan 2026

PF-D2M：用于通用舞蹈到音乐生成的无姿态扩散模型

Jaekwon Im[1], *Natalia Polouliakh*[2],*Taketo Akama*[2]

[1] KAIST 文化技术研究生院，韩国[2] 索尼计算机科学实验室，日本东京jakeoneijk@kaist.ac.kr, nata@csl.sony.co.jp, taketo.akama@gmail.com

摘要

舞蹈到音乐生成旨在生成与舞蹈动作相匹配的音乐。现有方法通常依赖于从单人类舞者身上提取的身体运动特征，以及有限的舞蹈到音乐数据集，这限制了它们在涉及多名舞者和非人类舞者的现实场景中的性能与适用性。本文提出 PF-D2M，这是一种通用的、基于扩散模型的舞蹈到音乐生成模型，可利用从舞蹈视频中提取的视觉特征。PF-D2M 采用渐进式训练策略进行训练，有效应对数据稀缺和泛化方面的挑战。客观和主观评估均表明，PF-D2M 在舞蹈-音乐对齐和音乐质量方面达到了最先进性能。PF-D2M 的示例可在 https://jakeoneijk.github.io/pfd2m_project 查看

索引词：舞蹈到音乐生成、视频到音频生成、音乐生成、扩散模型、生成模型

1. 引言

舞蹈到音乐生成旨在合成与相应舞蹈动作同步的音乐。这项技术可用于支持编舞者、表演者和数字内容创作者的创作工作流，尤其是在短视频内容兴起的背景下。不同于使用时间缩放和时间扭曲来对齐给定音乐与舞蹈运动的传统方法 [1]，基于深度学习的方法可以根据舞蹈动作生成原创音乐。

为提取时间信息，以往的舞蹈到音乐生成方法主要利用从单人类舞者身上提取的运动特征，例如基于 3D Skinned Multi-Person Linear 模型（SMPL） [2] 的特征或二维身体关键点。一些方法 [3, 4, 5] 专注于生成以舞蹈序列为条件的符号表示，例如 MIDI。然而，符号音乐表示在表达真实舞蹈音乐的表现力和动态变化方面存在局限。为了解决这一局限，其他方法 [6]旨在生成包含细粒度声学信息的音频表示。近年来，扩散模型 [7] 已被广泛用于生成高质量音乐 [8, 9]。

尽管舞蹈到音乐生成领域近期取得了进展，但仍存在两个重要挑战。首先，从单人类舞者中提取的运动特征通常不足以表示舞蹈视频中的时间信息。尤其是在涉及多名舞者的场景中，编舞是依据多名表演者的动作而非单个表演者的动作与音乐进行对齐的。即使在只有一名舞者的情况下，现有大多数方法也主要专注于提取粗粒度的节奏特征 [8, 9]。此外，姿态估计模块可能会导致姿态序列抖动，并且在非人类舞者（如二维动画角色）上的表现有限。其次，公开的高质量舞蹈到音乐数据集十分稀缺。AIST++ [10] 被大多数现有工作用作主要训练数据集，但仅提供 60 首不同的歌曲，且其中的舞蹈视频具有相对简单的背景。这种数据集的稀缺性成为构建能够生成多样化高质量音乐的舞蹈到音乐模型的主要瓶颈。然而，由于版权问题，通过爬取数据来扩充训练数据仍然充满挑战。此外，野外舞蹈视频通常由于摄像机录制的声音而音频质量较差，或者包含与舞蹈无关的场景。

为克服这些局限，我们提出了 PF-D2M，一种用于通用舞蹈到音乐生成的扩散模型。PF-D2M 能够为一般的舞蹈视频有效生成多样且高质量的音乐，适用范围超越了仅包含单人类舞者的场景。PF-D2M 不依赖于从单人类舞者提取的运动特征，而是利用 Synchformer [11] 从视频中提取视觉特征。此外，为解决数据稀缺问题，我们提出了一套训练方案，通过逐步训练模型来拓展其能力。我们的贡献总结如下：

- 我们提出了 PF-D2M，这是一种通用的舞蹈到音乐生成模型，能够为一般的舞蹈视频生成音乐，包括有多名舞者和非人类舞者的场景。
- 我们提出了一种渐进式训练策略，能够有效缓解数据稀缺导致的过拟合，并提升模型的生成能力。
- 客观评估和主观评估均表明，与以往方法相比，PF-D2M 达到了最先进性能。

2. 方法

2.1. 概述

图 1 展示了 PF-D M 的整体架构。令 $\mathbf{v} \in \mathbb{R}^{(T \times f_v \times H \times W \times 3}$ 表示无声舞蹈视频，T 其中，f_v,H, 和 W 分别表示时长、视频帧率（FPS）、高度和宽度，而 3 对应 RGB通道。音频 $\mathbf{a} \in \mathbb{R}^{2 \times (T \times f_s)}$ 是与$\mathbf{v}$在时间上对齐的立体声音乐，其中 f_s 表示音频采样率。文本描述 $\mathbf{c}$ 描述音乐，例如流派、乐器和情绪。PF-D2M 以$\mathbf{v}$ 和 $\mathbf{c}$作为输入条件，生成$\mathbf{a}$。我们在第 2.2 节介绍 PF-D2M 的架构。在第 2.3 节，我们介绍由两个阶段组成的训练方案。

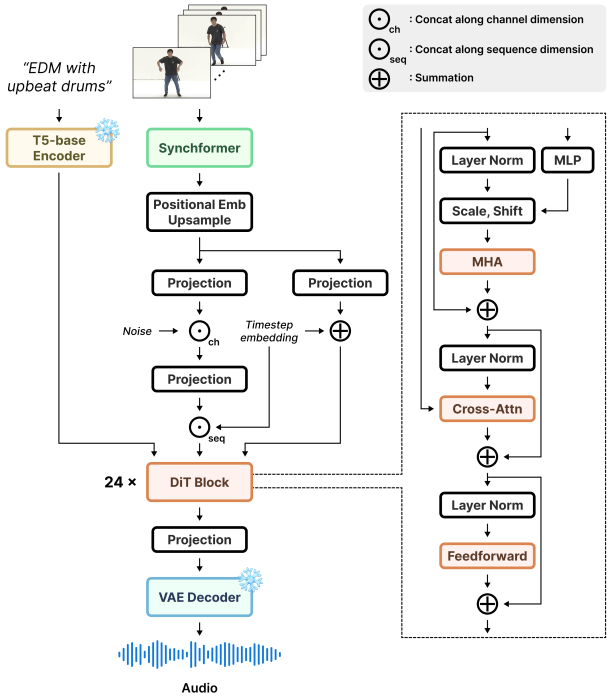


图 1： *PF-D2M* 概述。

2.2. 模型架构

我们采用 DiT架构以及来自 [12] 的预训练变分自编码器。变分自编码器编码器将 $\mathbf{a}$ 压缩为潜在表示$\mathbf{z} \in \mathbb{R}^{64\times(T\times f_s)/2048}$。DiT 根据三种条件特征生成$\mathbf{z}$ 。首先，使用预训练的 T5-base编码器 [13] 从$\mathbf{c}$ 提取文本嵌入，并通过交叉注意力输入 DiT。其次，将扩散时间步 [7] 的正弦嵌入添加到 DiT 的输入前端。第三，使用 Synchformer 的视觉编码器 [11] 从 $\mathbf{v}$ 中提取视觉特征，并按照 [14] 的方法将其与可学习的位置嵌入相加。为匹配$\mathbf{z}$的时间维度，通过最近邻插值对视觉特征进行上采样。随后，使用一维卷积层投影这些上采样后的视觉特征，使其与$\mathbf{z}$ 的通道维度相匹配，然后沿通道轴将其与 DiT 的输入拼接。我们发现，仅通过拼接提供视觉特征会导致收敛缓慢，因此还采用了另一种条件注入方法。将上采样后的视觉特征通过线性层投影，使其与 DiT 的隐藏维度相匹配，然后与时间步正弦嵌入相加。这些特征通过自适应层归一化（AdaLN） [15] 层中的逐帧缩放和偏置，在 DiT 的每一层中进行条件调制，方法与 [14] 类似。请注意，首帧不参与 AdaLN 调制，因为它们被视为前置的全局特征。

DiT 使用速度预测目标 [16]进行训练。视觉特征和文本嵌入均以 10% 的概率丢弃，从而实现无分类器引导 [17]。在推理时，我们采用 DPM-Solver++ [18] ，进行 100 步推理，并将无分类器引导系数设为 5.0。

2.3. 渐进式训练策略

我们发现，在 AIST++ 数据集 [10]上训练 PF-D2M 时存在两个问题。首先，模型对训练数据集中的音乐严重过拟合。由于 AIST++提供的音乐数据量有限，模型倾向于记忆训练音乐，缺乏生成新音乐的能力。其次，由于 AIST++舞蹈视频的摄像机角度和背景多样性有限，模型在现实场景中的泛化能力较差，而现实场景通常包含复杂的背景、光照条件和多样的视角。为了解决这些问题，我们引入了一种由两个阶段组成的渐进式训练策略。

2.3.1. 阶段 *0*：文本到音频预训练初始化

训练能够生成高质量、多样化音乐的模型需要大量计算资源。在实践中，我们使用 Stable Audio Open [12] 的权重初始化 PF-D2M 的权重。Stable Audio Open 是一种文本到音频模型，在音乐和音效方面都展现出强大的生成能力。为保留其音频生成能力，仅在 PF-D2M 中使用的模块权重会初始化为零。采用这种初始化策略后，PF-D2M 的表现与 Stable Audio Open 完全相同，但不使用时间条件，且处理不同长度的音频序列。

2.3.2. 阶段*1*：视频到音频对齐训练

在使用音乐数据训练模型之前，我们首先在 VGGSound [19] 上进行训练，使其学习视听同步。VGGSound 提供了约 500 小时的视频数据，涵盖多种类别。这种多样性使模型能够在真实场景中学习视听同步。对于文本描述，我们使用由 Qwen-Audio [20], 生成、由 AudioSetCaps [21]提供的描述。

在 VGGSound 提供的 309 个音频类别中，只有少数类别（例如踢踏舞 和人们拖着脚步走路）与舞蹈到音乐相关。它们所占比例很小（不足 10 小时），且这些类别中的许多样本都存在问题，包括现场录音导致音乐质量不佳、音乐仅由打击乐声组成，以及视频标注错误。此外，动作模式和音乐流派的多样性都十分有限。

尽管如此，我们观察到，经过第一阶段训练的模型能够为各种舞蹈视频生成多种流派的音乐，这得益于文本到音频预训练初始化，以及在多样化视频数据上进行的视频到音频对齐训练。然而，该模型存在两个主要局限。首先，它不会生成连贯的音乐，而是倾向于生成随视觉变化而频繁突变的音乐，最终形成一连串短音乐片段，而非一首连贯统一的曲目。其次，生成的音频往往会反映视频中存在的房间或环境特征。由于我们的目标是生成不受舞者所在位置影响的录音棚品质音乐，这种倾向与我们的目标并不一致。

2.3.3. 第二阶段：舞蹈到音乐微调

为解决第一阶段模型的局限性，我们在 AIST++ [10] 上对其进行微调。AIST 是 LORIS 基准 [8] 中提供的舞蹈到音乐数据集。受视频到音频任务中先前工作的启发[14, 22] ，我们在多模态数据集上训练模型，以提升音乐生成能力，并缓解模型在 AIST++ 上的过拟合。在第二阶段，模型在

舞蹈到音乐数据集（AIST++）、文本到音乐数据集（FMA [23], MoisesDB [24]）以及视频到音频数据集（VGGSound）上进行训练。由于各数据集的音频总时长不同，且为了让模型更专注于音乐生成而非音效，我们分别按照 2:4:1 的比例从舞蹈到音乐、文本到音乐和视频到音频数据集中采样训练批次。对于文本到音乐数据集，视觉特征使用可学习的空嵌入。

为舞蹈到音乐和文本到音乐数据集生成文本描述时，我们使用 Qwen2-Audio [25]。我们使用来自 [26] 的音乐标注提示词来标注流派、乐器和情绪标签。当 Qwen2-Audio 无法生成合适的标签时，我们会使用各数据集提供的元数据。训练期间，我们根据标注的标签生成文本提示词。具体来说，我们通过随机组合标签与预定义的文本模板来构建描述。当没有可用的有效标签时，我们会使用通用提示词，例如 “一段器乐曲目”。这种随机生成描述的方式提升了泛化能力，使模型无论文本描述的详细程度如何，都能生成高质量音乐。

3. 实验

3.1. 数据集

3.1.1. 文本到音乐数据集筛选

PF-D2M 专注于生成适合跳舞且不含人声内容的音乐。为此，我们使用 MoisesDB 中约 14 小时的音乐，这些音乐由不含人声的器乐音轨混合而成。对于 FMA，我们首先排除四种难以用于跳舞或可能包含歌声的流派：实验音乐、民谣音乐、旧时音乐和口语音乐。为进一步剔除低带宽音频，我们计算每首曲目的归一化频谱滚降频率 [27]，并丢弃数值低于 0.6 的样本。随后，我们筛除包含歌声的音乐。具体而言，我们首先使用音乐源分离模型 *htdemucs ft*[28]，分离人声成分，随后应用 Silero [29], 语音活动检测器来判断是否存在歌声。筛选后，所得数据集包含约 191 小时的音频。

3.1.2. 预处理

视频和音频分别重采样至 25 fps 和 44.1 kHz。训练时，所有视频和音频都被随机切分为时长 7.98 秒的片段。视频中的每个 RGB 帧都会调整大小并进行中心裁剪至 224×224。

3.2. 实现细节

对于第一阶段，PF-D2M 使用 1×10^{-5} 的基础学习率进行 200,000 步训练，并在前 1,000 步采用线性预热计划。训练 30,000 步后，学习率降至 1×10^{-6}。对于第二阶段，PF-D2M 使用固定学习率 1×10^{-6} 训练 1,500 步。所有阶段均采用批大小为 128 的设置，并使用 AdamW优化器，参数为 $\beta_1 = 0.9$ 和 $\beta_2 = 0.999$。

3.3. 评估

3.3.1. 客观评估

我们将 PF-D2M 与最先进的模型 CDCD [30], 、LORIS [8], 和 Text-Inv [9] 进行比较。所有对比模型均采用其官方实现。

我们在 LORIS基准 [8]下，使用 AIST++[10] 评估我们的方法和基线模型。按照以往研究的做法，我们保留数据集的 5% 用于测试。然而，与采用随机划分的先前方法不同，我们明确选取了特定的音乐曲目作为测试集：*mBR0*、*mMH0*、*mLO2*和*mJB5*。这主要基于两点考虑。首先，这确保测试集仅包含训练期间未见过的音乐曲目。其次，测试集应涵盖多样的音乐流派和速度。

沿用先前研究，我们使用五项客观指标 [8]评估生成音乐与真实音乐之间的节奏对应性。具体而言，我们采用节拍覆盖得分（BCS）、覆盖标准差（CSD）、节拍命中得分（BHS）、命中标准差（HSD）和 F1分数。

3.3.2. 主观评估

为评估真实场景，我们收集了具有不同舞者人数、舞蹈风格、舞者类型、背景、光照条件和摄像机角度的高难度舞蹈视频。我们定义了四个类别，并为每个类别收集五个样本，共计 20 个样本：单人类舞者、单个非人类舞者、多人类舞者和多个非人类舞者。我们将非人类舞者定义为非真实人类的表演者，包括二维和三维角色。单人类舞者 和单个非人类舞者 类别均指只有一名舞者表演的视频，但可能有其他人作为观众出现。相比之下，多人类舞者 和多个非人类舞者 类别指多名舞者共同表演的视频，舞者人数从 2 人到 91 人不等。

在主观评估中，我们将 PF-D2M 与 LORIS [8]和 Text-Inv [9] 进行比较。为了获取 LORIS 和 Text-Inv 的二维骨架，我们采用了 HRNet [31]，该模型来自 MMPose [32], ，也是这两项工作中使用的同一姿态估计器。为确保比较公平，所有生成的音频都被切分为时长 5.12 秒的片段，对应于 Text-Inv 的生成时长；在所比较的方法中，Text-Inv 的生成时长最短。

我们通过一项有 20 名参与者参加的听音测试进行了主观评估。参与者需要评估舞蹈-音乐对齐和音乐质量。舞蹈-音乐对齐衡量生成的音乐与舞蹈的匹配程度，不考虑音乐质量。音乐质量则评估生成音乐的感知质量，包括音频保真度和音乐创作，并将其独立于伴随的舞蹈进行考量。对于每个问题，参与者都要评估各个对比方法生成的音乐，并按 1 至 5 分的量表对每项标准打分。

4. 结果

4.1. 客观评估

表 1 展示了客观评估结果。除 BCS 外，PF-D2M 在所有指标上均达到了最先进性能。这些结果证明了融入 Synchformer 视觉特征的有效性；与依赖单人类舞者粗略节奏特征的方法相比，这些特征能够实现更好的舞蹈与音乐同步。

除 BCS 外，第二阶段模型在所有指标上均优于第一阶段模型。第一阶段模型缺乏音乐结构的一致性，而第二阶段模型生成的音乐

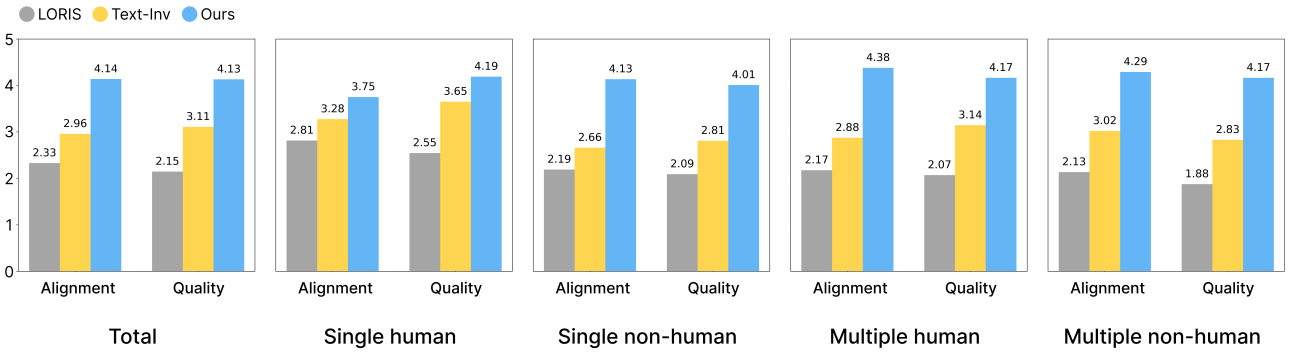


图 2： 野外数据集上的主观评估结果。

表 1： *AIST++* 测试集上的客观评估结果。*S1* 表示阶段 *1* 训练，*S2* 表示阶段 *2* 训练。每项指标的最佳结果以粗体显示。

方法	BCS↑	CSD↓	BHS↑	HSD↓	F1↑
CDCD [30]	89.2	9.0	93.8	10.0	91.5
LORIS [8]	89.9	8.9	95.3	8.9	92.5
文本反演 [9]	**90.6**	11.1	80.9	28.3	85.5
PF-D2M（阶段 1）	90.5	13.1	91.2	18.6	90.9
PF-D2M（阶段 2）	89.4	**8.1**	**99.8**	**1.9**	**94.3**

具有连贯的音乐结构，同时仍能反映细粒度的舞蹈动作。尽管节奏评估指标无法衡量音乐质量，但值得注意的是，无论舞者身处何处，第二阶段模型都能生成高质量音乐，带来更好的聆听体验；而第一阶段模型生成的音乐往往听起来像是在现场录制的音频。

与 Text-Inv 和 LORIS 相比，PF-D2M 在 BCS 上的表现略差。我们观察到，PF-D2M 经常生成节奏结构细分的音乐，例如密集的踩镲节奏型，尤其是在霹雳舞和嘻哈流派中。尽管这种节奏结构适合这些流派，但检测到的节拍数量增加可能会对 BCS 指标产生负面影响。

我们观察到，客观指标 BCS、CSD、BHS、HSD 和 F1 与感知质量并不总是高度相关。对于给定的舞蹈，并不存在唯一正确的音乐曲目，但这些指标通过将生成的音乐与真实音乐进行比较来评估其表现。因此，即使输出音乐与舞蹈配合良好，只要其节奏结构与真实音乐不同，也可能得分较低。此外，这些指标并非用于评估音频保真度或音乐结构。尽管一些先前的工作 [33] 采用弗雷歇音频距离（FAD） [33] 评估生成音乐的质量，但我们并未使用该指标，因为测试集过小，无法得出可靠的 FAD 结果。因此，我们更加重视主观评估，并将开发更合适的客观数据集和舞蹈到音乐生成评估指标留待未来工作。

4.2. 主观评估

图 2 展示了主观评估的结果。PF-D2M 在所有指标和测试案例中均显著优于对比方法，证明了其在野外场景中的有效性。我们发现，LORIS 生成的音乐往往缺乏丰富的音乐内容，导致器乐作品和编曲单调；而 Text-Inv 生成的音频通常音质较差。此外，LORIS 和 Text-Inv 往往都无法捕捉细致的舞者动作，舞蹈与生成音乐之间的对齐较弱。相比之下，PF-D2M 能够生成反映详细舞蹈动作的高质量音乐。值得注意的是，PF-D2M 经常生成突出特定动作的编曲，使舞蹈内容更具动感和表现力。

与基线方法相比，在单人类舞者设定之外的场景中，例如非人类舞者和多舞者场景，PF-D2M 进一步扩大了性能差距。此外，我们发现，PF-D2M 在复杂视觉条件下表现稳健，包括非典型摄像机视角和多镜头剪辑视频；相比之下，对比方法往往无法实现恰当的舞蹈-音乐对齐。这些结果凸显了 PF-D2M 的多功能性，以及它在现实场景中的更强适用性。

5. 结论

本文提出了 PF-D2M，一种用于通用舞蹈到音乐生成的扩散模型。通过融入从舞蹈视频中提取的视觉特征，PF-D2M 无需依赖单一的人体姿态表示，并支持更广泛的场景，包括多名舞者和非人类舞者。我们进一步提出了一种渐进式训练策略，结合文本到音频初始化、视频到音频对齐训练和舞蹈到音乐微调，有效缓解了数据稀缺导致的过拟合问题。客观和主观评估均表明，PF-D2M 在舞蹈-音乐对齐和感知音乐质量方面达到了最先进性能。

尽管 PF-D2M 在各种场景中表现出色，但仍存在若干有待未来工作解决的局限。首先，与实际应用场景相比，PF-D2M 生成的音乐时长相对较短。未来工作可以探索生成时长更长、且与不同舞蹈进程良好匹配的音乐。其次，缺乏用于舞蹈到音乐生成的评估数据集，阻碍了客观评估，这凸显了建立包含多样化音乐和舞蹈视频的基准的必要性。

6. 参考文献

[1] H.-C. Lee, I.-K. Lee 等人，“计算机动画中背景音乐与运动的自动同步”，载于《Computer Graphics Forum》，第 24 卷，第 3 期。阿姆斯特丹：北荷兰，1982-，2005 年，第 353–362 页。[2] M. Loper、N. Mahmood、J. Romero、G. Pons-Moll 和 M. J. Black，“SMPL：一种蒙皮多人线性模型”，载于《Seminal Graphics Papers: Pushing the Boundaries, Volume 2》，2023 年，第 851–866 页。[3] G. Aggarwal 和 D. Parikh，“Dance2Music：自动舞蹈驱动音乐生成”，arXiv 预印本 arXiv:2107 06252，2021 年。[4] B. Han、Y. Li、Y. Shen、Y. Ren 和 F. Han，“Dance2MIDI：舞蹈驱动的多乐器音乐生成”，《Computational Visual Media》，第 10 卷，第 4 期，第 791–802 页，2024 年。[5] X. Liang、W. Li、L. Huang 和 C. Gao，“DanceComposer：使用渐进式条件音乐生成器进行舞蹈到音乐生成”，《IEEE Transactions on Multimedia》，2024 年。[6] Y. Zhu、K. Olszewski、Y. Wu、P. Achlioptas、M. Chai、Y. Yan 和 S. Tulyakov，“从舞蹈视频生成复杂音乐的量化生成对抗网络”，载于《欧洲计算机视觉会议》。Springer，2022 年，第 182–199 页。[7] J. Ho、A. Jain 和 P. Abbeel，“去噪扩散概率模型”，载于 NeurIPS，第 33 卷，2020 年，第 6840–6851 页。[8] J. Yu、Y. Wang、X. Chen、X. Sun 和 Y. Qiao，“长时程节奏视频音轨器”，载于《国际机器学习会议》。PMLR，2023 年，第 40 339–40 353 页。[9] S. Li、W. Dong、Y. Zhang、F. Tang、C. Ma、O. Deussen、T.-Y. Lee 和 C. Xu，“使用基于编码器的文本反演进行舞蹈到音乐生成”，载于《SIGGRAPH Asia 2024 Conference Papers》，2024 年，第 1–11 页。[10] R. Li、S. Yang、D. A. Ross 和 A. Kanazawa，“AI Choreographer：音乐条件控制的 3D 舞蹈生成，基于 AIST++”，载于 IEEE/CVF计算机视觉国际会议论文集，2021 年，第 13 401–13 412 页。[11] V. Iashin、W. Xie、E. Rahtu 和 A. Zisserman，“Synchformer：利用稀疏线索实现高效同步”，载于 ICASSP 2024-2024 IEEE国际声学、语音与信号处理会议（ICASSP）。IEEE，2024 年，第 5325–5329 页。[12] Z. Evans、J. D. Parker、C. Carr、Z. Zukowski、J. Taylor 和 J. Pons，“Stable Audio Open”，载于 ICASSP。IEEE，2025 年，第 1–5 页。[13] C. Raffel、N. Shazeer、A. Roberts、K. Lee、S. Narang、M. Matena、Y. Zhou、W. Li 和 P. J. Liu，“探索统一文本到文本转换器的迁移学习极限”，《JMLR》，第 21 卷，第 140 期，第 1–67 页，2020 年。[14] H. K. Cheng、M. Ishii、A. Hayakawa、T. Shibuya、A. Schwing 和 Y. Mitsufuji，“MMAudio：驾驭多模态联合训练，实现高质量的视频到音频合成”，载于《计算机视觉与模式识别会议论文集》，2025 年，第 28 901–28 911 页。[15] E. Perez、F. Strub、H. De Vries、V. Dumoulin 和 A. Courville，“FiLM：借助通用条件层进行视觉推理”，载于《AAAI人工智能会议论文集》，第 32 卷，第 1 期，2018 年。[16] T. Salimans 和 J. Ho，“用于扩散模型快速采样的渐进式蒸馏”，载于 ICLR，2022 年。[17] J. Ho 和 T. Salimans，“无分类器扩散引导”，载于 NeurIPS 2021 深度生成模型及其下游应用研讨会，2021 年。[18] C. Lu、Y. Zhou、F. Bao、J. Chen、C. Li 和 J. Zhu，“DPM-Solver++：用于引导采样扩散概率模型的快速求解器”，《Machine Intelligence Research》，第 1–22 页，2025 年。[19] H. Chen、W. Xie、A. Vedaldi 和 A. Zisserman，“VGGSound：大规模视听数据集”，载于 ICASSP 2020-2020 IEEE国际声学、语音与信号处理会议（ICASSP）。IEEE，2020 年，第 721–725 页。[20] Y. Chu、J. Xu、X. Zhou、Q. Yang、S. Zhang、Z. Yan、C. Zhou 和 J. Zhou，“Qwen-audio：通过统一的大规模音频语言模型推进通用音频理解”，*arXiv* 预印本 *arXiv:2311.07919*，2023 年。[21] J. Bai、H. Liu、M. Wang、D. Shi、W. Wang、M. D. Plumbley、W.- S. Gan 和 J. Chen，“Audiosetcaps：利用大型音频语言模型生成增强型音频描述数据集”，发表于*Audio Imagination*：*NeurIPS 2024* 研讨会 人工智能驱动的语音、音乐和声音生成，2024 年。[22] Z. Chen、P. Seetharaman、B. Russell、O. Nieto、D. Bourgin、A. Owens 和 J. Salamon，“利用多模态控制实现视频引导的拟音生成”，发表于计算机视觉与模式识别会议论文集，2025 年，第 18 770–18781 页。[23] M. Defferrard、K. Benzi、P. Vandergheynst 和 X. Bresson，“FMA：一个音乐分析数据集”，发表于*ISMIR*，2017 年，第 316– 323 页。[24] I. Pereira、F. Ara´ujo、F. Korzeniowski 和 R. Vogl，“Moisesdb：一个超越 4 轨的源分离数据集”，发表于*ISMIR*，2023 年，第 619–626 页。[25] Y. Chu、J. Xu、Q. Yang、H. Wei、X. Wei、Z. Guo、Y. Leng、Y. Lv、J. He、J. Lin 等，“Qwen2-audio 技术报告”，*arXiv* 预印本 *arXiv:2407.10759*，2024 年。

[26] R. Yuan, H. Lin, Y. Liang, Y. Wu, Z. Dong, Y. Liu, Y. Guo, K. Zhu, J. Liang, W. Ma, X. Du, S. Liu, Y. Yue, 和 J. Fu，“Yue：扩展开放基础模型以实现长篇音乐生成”，*arXiv* 预印本 *arXiv:2503.08638*，2025。

[27] J. Im 和 J. Nam，“Saga-sr：语义与声学引导的音频超分辨率”，*arXiv* 预印本 *arXiv:2509.24924*，2025。[28] S. Rouard、F. Massa 和 A. D´efossez，“用于音乐源分离的混合变换器”，载于 *ICASSP23*，2023。[29] S. Team，“Silero vad：预训练的企业级语音活动检测器（vad）、数字检测器和语言分类器”，https://github.com/snakers4/silero-vad，2024。[30] Y. Zhu、Y. Wu、K. Olszewski、J. Ren、S. Tulyakov 和 Y. Yan，“用于跨模态音乐与图像生成的离散对比扩散”，载于国际学习表征会议（*ICLR*），2023。[31] K. Sun、B. Xiao、D. Liu 和 J. Wang，“用于人体姿态估计的深度高分辨率表征学习”，载于*IEEE/CVF*计算机视觉与模式识别会议论文集，2019，第5693–5703页。[32] M. Contributors，“OpenMMLab姿态估计工具箱和基准测试”，https://github.com/open-mmlab/mmpose，2020。[33] K. Kilgour、M. Zuluaga、D. Roblek 和 M. Sharifi，“Fr´echet 音频距离：一种用于评估音乐增强算法的无参考指标”，载于 *Interspeech 2019*，2019，第2350–2354页。