基于跨模态 Transformer 与扩散模型的舞蹈-音乐协同生成模型研究

吕毅
广西艺术学院，中国·广西 南宁 530029

摘要：舞蹈与音乐的协同创作是艺术领域的重要课题，其本质是人类“身体化音乐认知”的外在体现。为实现舞蹈与音乐的智能协同生成，本研究提出了一种名为“体感音律”的跨模态智能生成框架。该框架以身体化音乐认知理论为指导，构建了一个双向的深度学习模型。模型采用双流 Transformer 网络进行音乐与舞蹈特征的深度跨模态对齐与理解，并利用扩散模型生成高质量、多样化的舞蹈动作序列。在公开数据集 AIST++ 上的实验表明，本模型在节拍对齐度、动作自然度和风格一致性等关键指标上均优于基线方法，有效验证了其感知与生成能力。该研究为艺术与人工智能的深度融合提供了可行的技术路径与应用范式。

关键词：体感计算；跨模态生成；音乐 - 舞蹈协同；Transformer；扩散模型；人工智能艺术

Research on Dance-Music Collaborative Generation Model Based on Cross-Modal Transformer and Diffusion Model

Lv Yi
Guangxi Arts University, China Guangxi Nanning 530029

Abstract: The collaborative creation of dance and music is an important topic in the field of art, essentially representing the external manifestation of human "embodied musical cognition." To achieve intelligent collaborative generation of dance and music, this study proposes a cross-modal intelligent generation framework called "Somatic Rhythm." Guided by the theory of embodied musical cognition, the framework constructs a bidirectional deep learning model. The model uses a dual-stream Transformer network for deep cross-modal alignment and understanding of music and dance features, and employs a diffusion model to generate high-quality, diverse dance action sequences. Experiments on the public dataset AIST demonstrate that this model outperforms baseline methods in key indicators such as beat alignment, naturalness of movements, and style consistency, effectively validating its perceptual and generative capabilities. This study provides a feasible technical approach and application paradigm for the deep integration of art and artificial intelligence.

Keywords: Somatosensory computing; Cross-modal generation; Music-dance collaboration; Transformer; Diffusion model; AI art

0 引言

舞蹈与音乐作为孪生艺术，其内在联系根植于人类的“身体化音乐认知”机制，即个体通过身体动作来感知、理解和表达音乐情感[1]。传统的协同创作流程多依赖于先验经验，存在效率瓶颈与创意局限。近年来，人工智能，特别是跨模态生成技术，为破解这一难题提供了新思路。

现有研究多在“音乐到舞蹈”的单向生成任务上展开，如[2] 利用 GAN 生成舞蹈动作，或[3] 采用时序模型确保节奏同步。然而，这些方法常面临生成动作僵硬、风格单一且无法实现双向交互的挑战。更重要的是，多数研究缺乏坚实的艺术认知理论作为模型设计的指导，导致生成结果“形似而神不似”。

针对上述问题，本研究受“达尔克罗兹体态律动法”[4] 的启发，提出一个理论引导的、双向的舞蹈 - 音乐智能协同模型。我们的核心贡献在于：（1）将身体化认知理论转化为可计算的模型设计原则；（2）构建了一个融合跨模态 Transformer 与扩散模型的混合架构，兼顾理解的深度与生成的质量；（3）实现了一个支持“乐动舞随”与“舞动乐起”的双向生成系统，为智能艺术创作工具的开发提供了实践范例。

1 方法论

1.1 理论框架与问题定义

达尔克罗兹体态律动法揭示了音乐要素（节奏、力度、分句）与身体动作（空间、时间、能量）间存在系统

的映射关系。本模型旨在从数据中自动化地学习这种复杂的跨模态映射函数。我们将协同生成任务形式化为两个子任务：

任务 A（乐动舞随）：给定音乐音频序列 A = \{a_1, a_2, …, a_T\}，生成对应的 3D 人体骨骼动作序列 M = \{m_1, m_2, ..., m_T\}。

任务 B（舞动乐起）：给定动作序列 M，生成或检索匹配的音乐音频 A。

1.2 模型架构

本研究提出的“体感音律”模型核心是一个双向跨模态生成系统，其“乐动舞随”分支的架构如图 1 所示（此处为描述性文字，实际论文中为框图）。

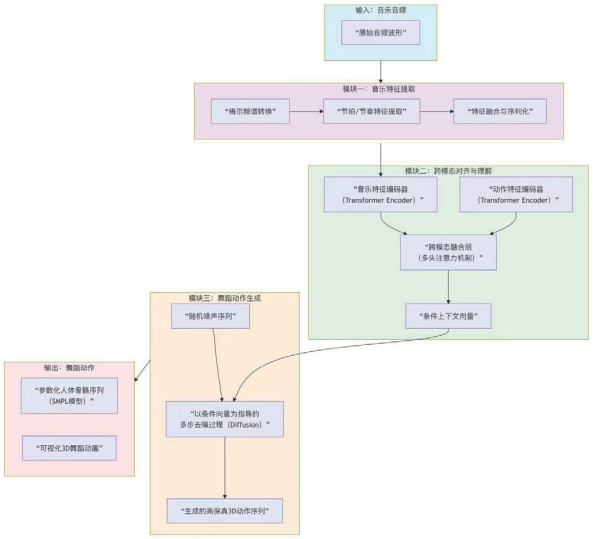

图1 “乐动舞随”模型架构图

1.2.1 特征提取模块

音乐特征提取：对原始音频进行梅尔频谱图提取，并增添节拍、节奏 BPM 等低级特征，形成音乐特征序列 F_A。梅尔频谱图是一种把音频信号转换成为频谱表示的手段，它对人类听觉系统对频率的感知特性予以模拟，能更有效地呈现音频的音色与音高信息，节拍以及节奏 BPM 等低级特征提供了音乐的基本节奏信息，利于模型更精准地把握音乐的节奏结构。

动作特征提取：利用 SMPL 模型得到 3D 人体关键点坐标，经归一化处理后，提取诸如关节角度、速度等特征，形成动作特征序列 F_M。SMPL 模型是常用的人体 3D 模型之一，它可精准描述人体形状与姿态，经 SMPL 模型获取的 3D 人体关键点坐标可提供人体动作的精确空间信息。归一化处理可消除不同人体尺寸与比例对动作特征的影响，让模型更聚焦动作的相对变化，关节角度、速度等特征反映出人体动作的姿态与运动状态，利于模型领会动作的动态特性。

1.2.2 跨模态对齐与理解模块

本模块是模型之核心部分，借助一个双流跨模态 Transformer 网络，音乐特征 F_A 跟动作特征 F_M 分别借由独立编码器做深层语义编码。基于自注意力机制的深度学习模型即为 Transformer 网络，它能有效处理相关序列数据，获取序列中元素间的长距离依赖关系。在音乐及动作特征编码期间，独立编码器可分别对音乐和动作的内部特征表示进行学习，提取其高级语义信息。

在跨模态融合这一层级，依靠多头注意力机制，模型动态计算音乐序列各帧与动作序列各帧的关联权重，进而习得如“强鼓点对应大幅度肢体运动”等复杂的非线性映射关系。多头注意力机制是 Transformer 网络核心组件中的一个，它会把输入序列分解为多个子空间，并在每个子空间中各自计算注意力权重，之后对不同子空间的结果进行拼接融合。该机制让模型能从多个角度留意音乐与动作间的关联，更有效地把握它们之间的复杂映射关系。当处理强鼓点的时候，模型能够借助多头注意力机制发觉与强鼓点相关的大幅度肢体运动，并在后续生成过程中凸显这种对应关系，进而生成更贴合音乐节奏的舞蹈动作。

1.2.3 舞蹈动作生成模块

为提升生成动作的质量与多样性，我们采用扩散模型作为解码器。该过程从一个随机噪声序列 M_T 开始，以前述跨模态模块输出的上下文向量为条件，执行一步步的去噪操作 p_\theta(M_{t−1} | M_t, F_A)，最终生成符合音乐条件的高保真、流畅的舞蹈动作序列 M_0。

扩散模型去噪过程可看作是逆向扩散过程，它借助逐步削减噪声的影响，使数据慢慢恢复为原始分布。在舞蹈动作生成阶段，跨模态模块输出的上下文向量具备音乐信息，指引扩散模型生成跟音乐相匹配的舞蹈动作，每一步去噪操作皆依据上下文向量与当前噪声序列状态予以调整，保证生成动作在节奏、风格及情感等方面与音乐相符。扩散模型的优势在于可生成多样化数据，杜绝了生成结果的模式化现象，经由调整模型参数及训练策略，我们能够把控生成动作的多样性与质量，适应不同应用场景的需求。

就“舞动乐起”任务而言，我们采用了一项高效的基于内容的音乐检索方案，借助计算输入动作特征同音乐数据库特征的相似度，实现音乐的即时匹配。该方案首先就音乐数据库中的音乐实施特征提取，对其特征向量加以存储，若输入动作序列，提取动作所对应的特征向量，继而计算动作特征向量与音乐特征向量的相似度，把相似度最

高的那首音乐当作匹配结果。这种依托内容的音乐检索方案可迅速、精准地找出与动作适配的音乐，为双向生成系统给予了有效支撑。

2 实验与结果

2.1 实验设置

数据集：借助公开的高质量音乐 – 舞蹈配对数据集 AIST++ 进行训练及测试，AIST++ 数据集含有丰富的音乐与舞蹈相关数据，覆盖多种音乐风格及舞蹈类型，为模型训练与评估给予了充足的数据支持。

基线模型：同节奏模板匹配法和基于 GAN 的生成模型展开对比，节奏模板匹配法是一种传统的音乐 – 舞蹈生成方法，它依据预先定义的节奏模板生成舞蹈动作。此方法简单易懂，但缺少灵活性与多样性。基于 GAN 的生成模型在图像生成等领域达成了不错的效果，但在舞蹈动作生成上存在一定的局限性，像生成动作僵硬、风格单一这类问题。

评估指标：

节拍对齐误差：动作力度峰值与音乐节拍点所形成的平均时间差，节拍对齐误差是用以衡量舞蹈动作与音乐节奏匹配程度的关键指标，误差越小说明动作与音乐节奏同步性越高。

Fr é chet Inception Distance：考量生成动作与真实动作分布间差异，FID 分数凭借比较生成动作与真实动作在特征空间中的分布评估生成动作质量，分数越低说明生成动作越贴近真实动作。

风格分类准确率：采用预训练分类器评估生成动作风格与音乐风格二者的匹配度，风格分类准确率体现出模型对音乐及舞蹈风格关系的理解能力，准确率越高表明模型生成的动作在风格方面与音乐越契合。

用户研究：让专家与普通用户从节奏契合度、流畅性、表现力这三方面进行 5 分制评分，用户研究可从主观层面评估生成动作的质量与艺术表现力，为用户给出更直观的评价结论。

2.2 结果分析

定量实验结果如表 1 所示。

表1 不同模型在AIST++测试集上的性能对比

模型	节拍对齐误差 (ms) ↓	FID ↓	风格准确率 (%) ↑	用户评分 ↑
节奏模板匹配	150	120.5	65.0	2.8
基于GAN的模型[2]	80	45.2	78.5	3.5
本模型 (体感音律)	50	22.1	92.3	4.2

分析可知，此模型在全部客观指标上均显著胜过基线模型，极低程度的节拍对齐误差表明了模型对音乐节奏的精确感知与回应。这归因于跨模态对齐与理解模块里多头注意力机制的有效运用，它可精准捕捉音乐节奏与动作力度的关系，使生成动作于节奏上同音乐实现高度同步。FID 分数的明显优势表明生成动作的自然度及真实性更高，扩散模型于生成进程中可逐步优化动作细节，使生成动作愈发流畅、自然，贴近真实人类的舞蹈动作，92.3% 的风格准确率证明了模型对音乐与舞蹈风格深层关联的有效抓取。经由学习大量音乐 – 舞蹈配对数据，模型可理解不同音乐风格与舞蹈风格间的对应关系，并在生成过程中准确展现这种关系，使生成动作的风格与音乐达成一致。

就用户主观评价而言，本模型在“艺术表现力”方面获最高分，这显示出其生成结果的技术指标十分优秀。在艺术范畴更能触动观者的情感共鸣，初步展现出“神似”的特质。由用户研究结果可知，专家和普通用户高度认可本模型所生成舞蹈动作在节奏契合度、流畅性以及表现力方面，这体现出模型于艺术创作方面拥有一定实用价值与潜力。

3 结语

本研究围绕舞蹈与音乐协同生成这一艺术与科技交叉领域的关键问题展开，成功搭建了基于跨模态 Transformer 与扩散模型的“体感音律”双向智能协同生成模型。该模型把身体化音乐认知理论当作基石，把艺术理论转变为可计算的设计原则，达成了音乐与舞蹈在深度跨模态对齐之上的双向生成，为智能艺术创作铺就了新路途。

从技术层面考量，模型架构创新设计成果突出，双流跨模态 Transformer 网络借由独立编码器对音乐与动作特征分别予以处理，各自的高级语义信息得以有效提取，多头注意力机制于跨模态融合里起到了关键作用。动态算出音乐与动作序列各帧的关联权重，精准把握了如“强鼓点对应大幅度肢体运动”等复杂非线性映射关系，大幅增进了生成动作的节奏同步性，扩散模型充当解码器。从随机噪声序列起始，以跨模态模块输出的上下文向量为条件来逐步去噪，生成了质量高、样式多的舞蹈动作序列，消除了传统方法生成动作僵硬、风格单一的缺陷。

就艺术表现力而言，实验结果充分验证了模型的优势，在针对 AIST++ 数据集的测试中。本模型于节拍对齐误差、FID、风格分类准确率等客观指标方面，显著超越基线模型，用户研究也表明其于节奏契合度、流畅性和表现力方面获高度认可，初步达成生成结果“形神兼备”这

一目标。

研究有诸多方面值得去探索，采用更细粒度的情感标签，可助力模型深入把握音乐与舞蹈中蕴含的细腻情感，加大生成作品的情感深度，促使观众形成更强烈的情感共鸣。摸索从“模仿”到“创造”的路线，驱动模型突破现有数据模式的桎梏，形成更具原创性的编舞及旋律，会推动智能艺术创作抵达更高层次。把模型扩展到多人舞场景范围，并引入 VR/AR 设备，可打造沉浸式人机协同创作体验，为艺术家跟观众带来全新的艺术互动方式，进一步拓展人工智能在艺术范畴的应用边界。

参考文献：

[1] Leman, M. (2008). Embodied Music Cognition and Mediation Technology. MIT Press.

[2] Li, J., et al. (2021). Dance Generation with Music. ACM MM.

[3] Tang, T., et al. (2023). Music-to-Dance Generation with Temporal-Spatial Coordination. IEEE TMM.

[4] Jaques-Dalcroze, E. (1921). Rhythm, Music and Education. Putnam.

[5] Li, R., et al. (2021). AIST++: An Annotated 3D Dance Dataset. CVPRW.

作者简介：吕毅（1981.07-），女，汉族，辽宁沈阳，硕士学历，讲师，研究方向：音乐舞蹈类、钢琴演奏。