arXiv:2512.18804v1 [cs.CV] 21 Dec 2025

节奏速度作为稳定线索：用于音乐到三维舞蹈生成的节奏速度与节拍专家的层次化混合

Guangtao Lyu[1], Chenghao Xu[2], Qi Liu[1],Jiexi Yan[3],Muli Yang[4],Fen Fang[4], Cheng Deng[1*1] 西安电子科技大学电子工程学院，中国， [2] 河海大学，中国，[3] 西安电子科技大学计算机科学与技术学院，中国，[4] 新加坡科技研究局资讯通信研究院（I²R），新加坡，{guangtaolyu,qiliu}@stu.xidian.edu.cn, fang fen@a-star.edu.sg,{jxyan1995,muliyang.xd,chdeng.xd}@gmail.com

摘要

音乐到三维舞蹈生成旨在根据音乐合成逼真且节奏同步的人体舞蹈。虽然现有方法常依赖额外的流派标签来进一步提升舞蹈生成效果，但这类标签通常存在噪声、粒度粗糙、难以获取或不足以刻画现实世界音乐的多样性，因而可能导致节奏错位或风格漂移。相比之下，我们观察到，速度作为反映音乐节奏与快慢的核心属性，在不同数据集和流派中相对稳定，通常介于每分钟 *60* 至 *200* 拍之间。基于这一发现，我们提出 *TempoMoE*，这是一种分层的速度感知混合专家模块，可增强扩散模型及其节奏感知能力。*TempoMoE* 将动作专家按不同速度范围组织为节奏速度结构化分组，并利用多尺度节拍专家捕捉短程与长程节奏动态。分层节奏自适应路由根据音乐特征动态选择并融合专家，无需人工提供流派标签，即可实现灵活且与节奏对齐的生成。大量实验表明，*TempoMoE* 在舞蹈质量和节奏对齐方面取得了当前最先进的成果。

1. 引言

音乐到三维舞蹈生成旨在根据给定的音乐序列，自动合成逼真、多样且节奏协调的人体舞蹈。这项任务对于虚拟人、数字编舞和视听内容创作等应用至关重要 [37, 47]，也有助于深入了解听觉与运动认知之间的神经耦合 [4]。

近期深度生成建模领域的进展，涵盖

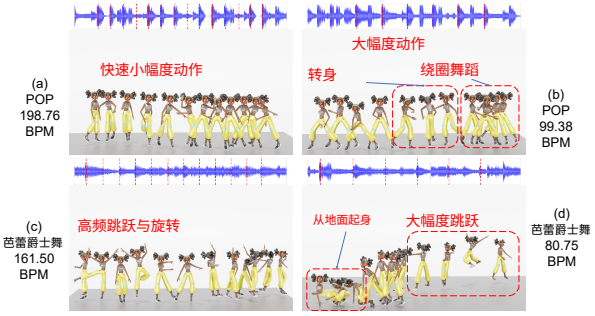


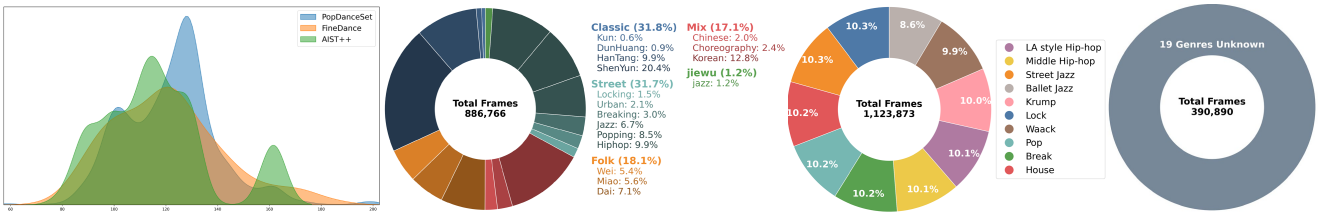
图1. 同一流派下不同速度的舞蹈可视化。即使在同一流派中，不同的每分钟节拍数也会带来不同的动作模式：每分钟节拍数较高时，每拍所占时间更短，因此动作更快、更局部（例如快速的手臂摆动、旋转）；每分钟节拍数较低时，每拍所占时间更长，因而能够完成更长、更复杂的动作姿态（例如身体旋转、全身转换）。

生成对抗网络（生成对抗网络）[23, 61]，自回归模型 [26, 48, 59]，和扩散模型 [10, 16,27, 34, 52]，，极大地提升了合成舞蹈的真实感、流畅度和多样性。

除了改进生成架构，近期研究也利用辅助信息来提升舞蹈生成质量，其中最常见的是通过流派条件控制。早期方法使用独热流派向量表示舞蹈风格 [23, 61]；后续研究学习连续的流派嵌入，以捕捉更丰富的风格变化 [27, 48]；近期研究则采用自然语言提示，灵活控制流派和风格 [45, 53]。

然而，将单热流派标签用作舞蹈生成的辅助线索存在若干局限：（1）节奏粒度粗糙：同一流派中的舞蹈，其速度（每分钟节拍数）、强度和动作风格可能存在很大差异。每分钟节拍数较高时，动作会更快、更紧凑；每分钟节拍数较低时，则可以呈现更慢、更舒展、更精细的动作姿态。因此，单一流派词元不足以捕捉流派内部如此丰富的节奏多样性 [26]；

*通讯作者



(a) 各数据集的每分钟节拍数分布　(b) FineDance 中的流派分布　(c) AIST 中的流派分布++　(d) PopDanceSet

图 2. (a) AIST++、FineDance 和 PopDanceSet 的每分钟节拍数分布表明，音乐速度主要集中在 60–200 BPM 的共同范围内，反映出一种共同的底层节奏结构。(b–d) 相比之下，流派分布高度不均衡，且因数据集而异：FineDance 和 AIST++ 采用不同的流派分类体系，而 PopDanceSet 未提供明确的流派标注。这些发现促使我们利用每分钟节拍数作为比流派标签更可靠的线索。

另请参阅图1。（2）可扩展性有限：音乐与舞蹈不断演变，催生出新的风格、子流派和混合风格 [2, 21, 22]。固定的流派词元无法适应这种开放且不断变化的风格空间。（3）监督信号不可靠：现实世界的数据集往往包含稀疏、模糊或不一致的流派标注。例如，PopDanceSet [34]完全缺少流派标签（图2（d）），使得流派条件生成难以可靠实现。这些问题限制了模型准确捕捉节奏并生成高质量、多样化舞蹈动作的能力。

这些局限促使我们探索能够可靠提升节奏舞蹈生成效果的替代线索。我们发现，速度是反映音乐节奏的核心属性，通常以每分钟节拍数衡量，并且在不同数据集和流派之间保持稳定 [20, 32, 39]。如图 2所示，尽管舞蹈风格丰富多样，每分钟节拍数仍聚集在一致的范围内（60–200 BPM），揭示了一种普适的节奏结构，可作为稳定先验，增强生成模型对节奏与拍点的感知。

受这些观察结果的启发，我们利用稳定的速度线索来改进扩散模型的架构，并增强其节奏感知动作建模。然而，由于速度分布广泛且连续，生成速度感知动作仍然充满挑战（图2）。高速度要求快速而复杂的动作，需要精确建模短时、细粒度的动态变化；而低速度通常表现为在更长时段内演变的平滑、连续动作姿态（图1），因此需要对全局动作结构进行建模 [5, 9]。因此，生成模型难以同时捕捉多样的速度特征，往往导致动作质量下降或节奏不一致。

为了有效解耦这些节奏变化，我们需要一种有效的架构，能够根据不同的速度条件自适应地分配建模能力。为此，专家混合（MoE）[17, 19] 提供了一种自然的解决方案。MoE 将模型分解为多个专门的子网络，称为专家，每个专家都能处理特定的输入模式。通过学习得到的路由网络会根据输入特征动态选择并组合专家，从而使模型能够高效扩展容量，并使专家能够针对不同条件进行专门化。这种将路由与专家处理明确分离的做法，已被证明在视觉、语音和多模态任务中十分有效 [30, 31]。

基于这些认识，我们提出了**TempoMoE**，一种节奏感知混合专家框架，将速度和节拍先验融入扩散模型，使专家能够专门处理不同的每分钟节拍数范围，并生成节奏对齐且结构连贯的动作。具体而言，TempoMoE包含两个关键组件。第一个组件是节奏结构化专家组。我们将节奏范围离散为八个每分钟节拍数区间，每个区间跨度为20每分钟节拍数，因为20每分钟节拍数与人类感知阈值以及先前研究中报道的常见舞蹈实践相一致 [13, 14]。每个速度区间对应一个专用专家组，其时间感受野经过设计，可与该速度层级下的动作节奏相匹配。在每个组内，我们定义了三个专门化专家，分别以不同的节拍尺度运行：四分之一拍、半拍和整拍，从细粒度重音到乐句级转换捕捉节奏结构。

第二个关键组件是一个两阶段路由机制，称为分层节奏自适应路由。第一阶段，我们使用轻量级门控网络 TempoGateNet 执行硬速度级别组选择。该网络根据全局音乐表示预测最相关的速度组。我们激活排名前 2 的组，以便在速度边界之间灵活混合。第二阶段，我们使用 BeatGateNet 执行软节拍尺度专家路由，为三个组内专家分配软门控权重，从而实现不同节拍尺度间的自适应融合。这种分层路由策略使模型能够沿速度和节奏分辨率两个维度动态地进行专门化，而无需依赖特定流派的标注或固定的流派标签，从而实现更准确、更具表现力的节奏对齐舞蹈生成。

总之，我们的主要贡献如下：

- 我们发现，速度是一种稳定的辅助线索，可用于提升舞蹈生成效果，并且在不同数据集和舞种中始终适用，无需人工标注。
- 我们提出 TempoMOE，将速度和节拍先验融入扩散模型结构，以增强节奏同步性和舞蹈连贯性。

•我们引入采用多尺度节拍建模的节奏结构化专家组，并结合分层节奏引导路由机制，使模型能够自适应地生成适应不同速度和节奏复杂度的动作动态。

•我们在 AIST++、Pop- DanceSet 和 FineDance 上进行了大量实验，结果表明，我们的方法在舞蹈动作质量和节奏对齐方面达到了最先进的性能。

2. 相关工作

专家混合。 专家混合 [17, 19] 是一种条件计算框架，可将每个输入路由至一部分专用子网络（专家），具有两大主要优势：(1) 扩展模型容量而无需按比例增加计算量，从而支持超大规模模型 [8, 24] (2) 动态专家专门化：门控函数可自适应地选择适合每个输入的专家，这一策略已广泛应用于视觉、语音和多模态建模领域 [30, 31,36]。专家混合中的路由通常分为两种：硬路由会激活得分最高的 k 个专家，并屏蔽其余专家，通常采用 TOP-1 或 TOP-2 选择，以兼顾效率与专门化程度 [11, 46]。软路由则根据门控分布对所有专家进行加权组合 [40, 43]。

音乐驱动舞蹈生成。该任务旨在根据音乐 [1, 6, 15, 16, 25, 28, 35, 54–56]生成时间连贯、风格一致的舞蹈动作。早期方法依赖动作图和基于检索的技术 [41, 58], , 但节奏对齐效果较差，泛化能力有限。随着深度学习的发展，研究者使用卷积神经网络、循环神经网络和图卷积网络，直接根据音乐回归生成动作 [12, 50, 61], ; 随后，基于生成对抗网络的模型进一步提升了动作的真实感和风格还原度 [23]。近期进展聚焦于生成式建模：自回归方法 [10, 49, 59] 利用学习得到的音乐风格嵌入增强时间一致性，而一些扩散模型 [45, 53]则通过自然语言提示或流派提示进一步提升合成质量。

同期工作 MEGADance [57] 采用专家混合框架，并通过基于流派的门控支持多风格舞蹈生成。然而，正如引言所述，以流派为条件的建模存在节奏粒度较粗、可扩展性有限以及监督信号不可靠等问题。相比之下，我们采用速度感知的专家混合框架，根据节奏信号进行路由，并利用具有不同时间范围和先验的专家，从而能够对多种风格进行多样化的节奏建模。

3. 预备知识

问题定义。音乐到舞蹈生成旨在合成多样且富有表现力的三维人体动作序列，使其在节奏上与输入音乐信号对齐。遵循先前的研究 [29], , 我们使用 Librosa [38], 提取帧级音乐特征 $c \in \mathbb{R}^{L\times 35}$ ，包括能量包络、梅尔频率倒谱系数、色度特征、起音和节拍，并按照标准惯例表示动作序列$\mathbf{x} \in \mathbb{R}^{L\times d}$ ，包括三维根节点平移、六维关节旋转和二值足部接触信号。

扩散模型。 扩散模型由两个核心部分组成：前向加噪过程和反向去噪过程。令$\mathbf{x}_0 \sim p(\mathbf{x}_0)$ 表示一个干净动作样本。前向过程通过在 T 个时间步中逐渐添加高斯噪声来破坏$\mathbf{x}_0$ ，从而得到一个序列 $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T\}$。该过程定义如下：

$$q(\mathbf{x}_t \mid \mathbf{x}_0) = \mathcal{N}\big(\mathbf{x}_t;\, \alpha_t \mathbf{x}_0, \sigma_t^2 \mathbf{I}\big), \tag{1}$$

$$\mathbf{x}_t = \alpha_t \mathbf{x}_0 + \sigma_t \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{2}$$

其中 $\alpha_t = \sqrt{1-\sigma_t^2}$， 噪声尺度 σ_t 随时间步 t递增。反向过程旨在从带噪输入中恢复原始样本$\mathbf{x}_0$ 。该过程通过高斯转移建模：

$$p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t) = \mathcal{N}\big(\mathbf{x}_{t-1};\, \hat{\boldsymbol{\mu}}_\theta(\mathbf{x}_t, t), \hat{\boldsymbol{\Sigma}}_\theta(\mathbf{x}_t, t)\big), \tag{3}$$

ˆ其中ˆ$\boldsymbol{\mu}_\theta$ 和 $\boldsymbol{\Sigma}_\theta$ 由神经网络估计。

为了训练去噪模型，我们采用干净样本预测目标 [44], 因其与下游约束兼容，并遵循先前的方法 [51, 52]:

$$\mathcal{L}_{\text{sample}} = \mathbb{E}_{\mathbf{x}_0, t}\left[\|\mathbf{x}_0 - \hat{\mathbf{x}}_\theta(\mathbf{z}_t, t, \mathbf{c})\|_2^2\right], \tag{4}$$

其中$\mathbf{z}_t$ 是时间步 t 处的带噪输入，$\mathbf{c}$表示条件信号。

4. 方法

4.1. 框架概述

我们采用去噪扩散变换器（DIT） [42] 架构进行动作生成。从干净的动作序列$\mathbf{x}_0$, 出发，我们在 T个时间步中迭代添加高斯噪声，以生成带噪样本$\mathbf{x}_t$，并训练模型通过学习到的反向扩散过程恢复$\mathbf{x}_0$ 。每个去噪模块包含三个组成部分：（1）用于建模动作内依赖关系的自注意力模块；（2）用于整合节奏感知音乐特征 c的交叉注意力模块；（3）用于全局条件控制的 AdaLN-Zero [42] 调制前馈网络。

为增强节奏感知能力和节奏自适应性，我们用速度感知专家混合（TempoMoE）模块替换前馈网络。TempoMoE 在层次化节奏自适应路由器的引导下，动态地将特征路由至一组按速度构建的专家组。这种设计既能精确地对节拍层面进行建模，又能灵活地实现速度泛化，从而提升不同速度下与音乐结构的对齐效果。

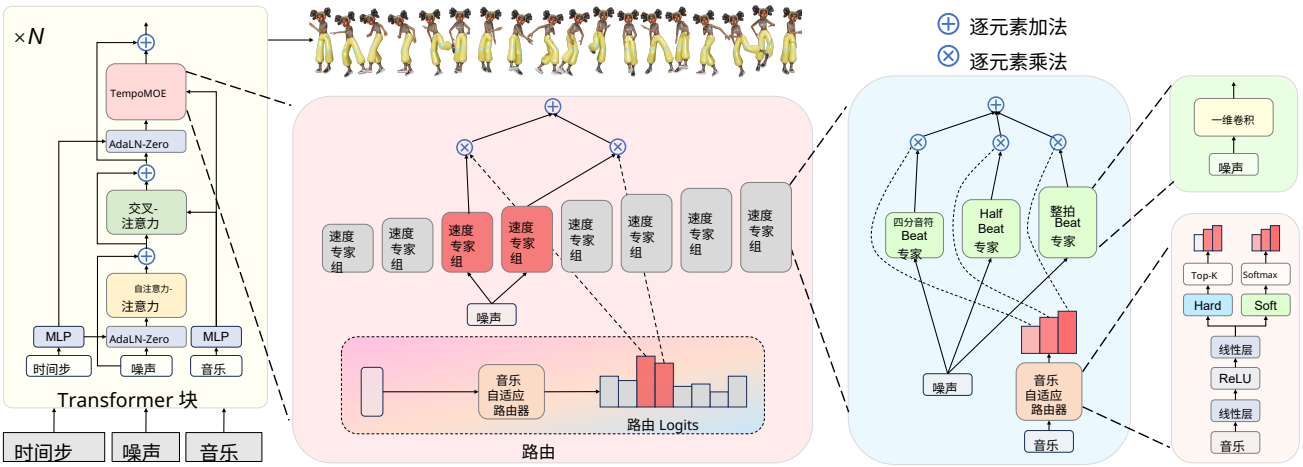


图3。TempoMoE框架概览：这是我们的节奏感知舞蹈生成扩散模型，采用 N Transformer块。除了通过交叉注意力融合音乐特征外，我们还用TempoMoE替换了原始前馈网络，使模型能够自适应地激活特定于节奏的专家组。这些专家组包含多尺度节拍专家，用于合成连贯且节奏对齐的三维舞蹈动作。

4.2. 节奏结构化专家组

人体舞蹈动作通常与音乐中的层次化节奏模式相契合，例如节拍细分和乐句。为体现这一结构，我们引入了速度结构化专家组，专门用于处理多个速度层级和节拍尺度上的动作特征。

我们将节奏范围离散为八个频段，各频段以锚定BPM为中心，间隔为20每分钟节拍数 [14]:

$$\{60, 80, 100, 120, 140, 160, 180, 200\}\ \text{BPM}. \tag{5}$$

每个速度区间定义一个节奏层级专家组，其中包含三个节拍尺度专家，分别以1 4-、1 2-和 1 拍的分辨率运行。这些专家分别捕捉细粒度的节拍重音和长程乐句。

每个专家旨在捕捉特定节拍尺度下的动作，并通过一维时间卷积层实现。通过调整卷积核大小，我们可以控制每个专家的时间感受野。给定帧率 f 和速度（BPM），我们计算每拍的预期帧数，以指导卷积核设计：

$$F_b = \frac{60 \times f}{\text{BPM}}, \quad k = \lceil r \times F_b \rceil_{\text{odd}}, \tag{6}$$

其中，r 表示节拍细分（例如1 4），$\lceil \cdot \rceil_{\text{odd}}$ 用于确保对称对齐。

总之，节奏结构化专家组将动作建模分解到不同的节奏层级和节拍尺度，使每个专家都能专门处理特定的时间分辨率。这种分层设计既能捕捉细粒度的节拍级重音，也能表现长程乐句；而不同速度区间之间的平滑过渡，则使模型能够稳健地适应不同的 BPM。

4.3. 分层节奏自适应路由

为充分发挥节奏结构化专家组的专长，我们提出了一种名为分层节奏自适应路由的两阶段路由策略，该策略根据全局速度上下文和局部节奏变化，动态选择并融合专家。它由两个关键组件构成：用于选择速度层级专家组的 TempoGateNet，以及用于组内节拍尺度路由的 BeatGateNet。

硬速度级专家组路由。 为了选择最相关的速度层级专家组，我们首先根据音乐特征$\mathbf{c} \in \mathbb{R}^d$计算所有 G 个组的分数向量：

$$\mathbf{s} = \text{TempoGateNet}(\mathbf{c}) \in \mathbb{R}^G, \tag{7}$$

其中，s_g 表示第 g个速度层级组的激活分数。然后，我们选取得分最高的 K 个组

$$\mathcal{G}_{\text{sel}} = \text{TopK}(\mathbf{s}, K) \tag{8}$$

用于激活。这种硬选择提高了计算效率，并确保每个活跃组都专注于其指定的 BPM范围，拥有量身定制的时间感受野。经验表明，$K = 2$ 在灵活性与路由清晰度之间实现了良好的平衡。

软节拍尺度专家路由。 在每个选定的组 $g \in \mathcal{G}_{\text{sel}}$ 中，我们使用 BeatGateNet 对其三个子专家进行软注意力处理（分别对应 1 4、1 2 和 1 拍分辨率）。注意力权重计算如下：

$$\boldsymbol{\gamma}^{(g)} = \text{Softmax}(\text{BeatGateNet}(\mathbf{c})) \in \mathbb{R}^3, \tag{9}$$

组 g 的融合输出为：

$$\mathbf{y}^{(g)} = \sum_{k=1}^{3} \gamma_k^{(g)} \cdot f_k^{(g)}(\mathbf{h}), \tag{10}$$

其中，$f_k^{(g)}(\mathbf{h})$ 表示组 g 中第 k 个节拍尺度专家在给定动作特征 $\mathbf{h}$ 时的输出。不同节拍分辨率之间的软融合使模型能够在微节拍细节与宏观乐句动态之间进行动态插值，并适应局部节奏纹理。

最终融合。 TempoMoE 模块的最终输出通过聚合选定组的输出计算得出：

$$\mathbf{h}' = \sum_{g \in \mathcal{G}_{\text{sel}}} \mathbf{y}^{(g)}. \tag{11}$$

这种分层设计能够以符合音乐规律的方式，控制时间跨度和节奏粒度。

4.4. 训练目标

我们联合优化扩散重建和运动学一致性，以提升动作的真实感和节奏保真度。

运动学损失。 为确保动作符合物理规律，我们遵循先前的研究 [29, 34, 52]，采用基于正向运动学（FK）的关节级运动学损失。该损失会惩罚不自然的关节旋转并施加生物力学约束，从而提升动作的真实感和时间连贯性：

$$\mathbf{x}_j^i = FK(\mathbf{x}_0^i), \quad \mathbf{x}_v^i = \text{Vel}(\mathbf{x}_0^i), \quad \mathbf{x}_a^i = \text{Acc}(\mathbf{x}_0^i), \tag{12}$$

$$\mathcal{L}_{\text{joint}} = \frac{1}{L} \sum_{i=1}^{L} \|\mathbf{x}_j^i - \hat{\mathbf{x}}_j^i\|_2^2, \tag{13}$$

$$\mathcal{L}_{\text{vel}} = \frac{1}{L} \sum_{i=1}^{L} \|\mathbf{x}_v^i - \hat{\mathbf{x}}_v^i\|_2^2, \tag{14}$$

$$\mathcal{L}_{\text{acc}} = \frac{1}{L} \sum_{i=1}^{L} \|\mathbf{x}_a^i - \hat{\mathbf{x}}_a^i\|_2^2, \tag{15}$$

$$\mathcal{L}_{\text{contact}} = \frac{1}{n-1} \sum_{j=1}^{n-1} \left\| (\mathbf{x}_j^{i+1} - \mathbf{x}_j^i) \cdot \hat{b}^{(i)} \right\|_2^2, \tag{16}$$

其中，$\hat{b}^{(i)}$ 表示每帧姿态中二值身体接触标签所占的部分 i。总运动学损失：

$$\mathcal{L}_{\text{kin}} = \lambda_{\text{joint}}\mathcal{L}_{\text{joint}} + \lambda_{\text{vel}}\mathcal{L}_{\text{vel}} + \lambda_{\text{contact}}\mathcal{L}_{\text{contact}} + \lambda_{\text{acc}}\mathcal{L}_{\text{acc}}. \tag{17}$$

总体损失。 完整的训练目标为：

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{simple}} + \mathcal{L}_{\text{kin}}. \tag{18}$$

5. 实验

数据集。 **AIST++** [26]包含 1,408 段高质量动作序列，涵盖 10 种街舞类型，每种舞种均包含多种速度。许多序列呈现相同的编舞，但伴奏的每分钟节拍数不同。**FineDance** [27]收录了通过光学动作捕捉采集的 7.7 小时以上专业舞蹈数据，涵盖 22 种舞种。序列平均长度超过 150 秒，可用于评估长距离编舞。**PopDanceSet** [34] 是从哔哩哔哩收集的野外数据集，包含 19 种多样的风格，但没有舞种标签。它是评估面向年轻群体的舞蹈生成的一个具有挑战性的基准。

评估指标。 遵循先前的研究 [34, 48], , 我们使用 FID_k 和 FID_g 衡量动作质量，使用 DIV_k 和 DIV_g 评估动作多样性，并使用节拍对齐分数（BAS）评估节奏同步。

实现细节。 扩散模型包含 8 个模块，潜在维度为 512，使用 Adam 训练 500 个 epoch（lr 1e-4，100 步预热，批量大小为 128）。无类别引导在训练时采用 10% 的 dropout，在推理时引导尺度为 2.5。采样采用 10 步 DPM-Solver++ [33]。遵循先前的研究 [52], , 损失权重为 $\lambda_{\text{joint}} = 0.646$、$\lambda_{\text{vel}} = 2.964$、$\lambda_{\text{contact}} = 10.942$、$\lambda_{\text{acc}} = 1$。所有实验均在 NVIDIA A6000 GPU 上运行。

5.1. 性能比较

定量结果。 我们在三个基准数据集上评估了我们的方法：AIST++、FineDance 和 PopDanceSet。表1 展示了与现有方法的比较，包括 FACT [26], Bailando [48], EDGE [52], 和 Lodge [29]。我们的方法在所有数据集和指标上均持续优于这些方法。值得注意的是，我们在 BAS 指标上取得了显著改进，凸显了我们速度感知专家路由的优势。更高的多样性指标（DIV_k、DIV_g）也表明，我们的模型能够生成更加多样且富有表现力的舞蹈，同时不牺牲连贯性，从而验证了其在不同速度和风格下的有效性。

定性结果。 图?? 展示了不同音乐流派下舞蹈生成效果的定性对比。与现有方法相比，后者往往会出现动作漂移、节奏不合或动作重复等问题，而我们的方法实现了更好的时间连贯性、更丰富的动作多样性，以及更具表现力、适应音乐流派的表演。生成的序列视觉上自然、风格一致，并与音频节拍结构高度契合。通过有效捕捉依赖节奏的动作模式和多样的风格细微差异，这些结果证明了我们具备节奏感知能力的专家混合设计能够生成节奏精准且富有表现力的舞蹈。

表1。定量　　结果在三个数据集上的表现。最佳和次佳结果分别以**粗体**和下划线标记 _____ .

数据集	方法	动作质量		动作多样性		节奏性	用户研究
		$FID_k \downarrow$	$FID_g \downarrow$	$Div_k \uparrow$	$Div_g \uparrow$	BAS ↑	我们获胜
AIST++	真实值	–	–	8.19	7.45	0.2374	–
	FACT	35.35	22.11	5.94	6.18	0.2209	91.2%
	Bailando	28.16	**9.62**	7.83	6.34	0.2332	81.6%
	EDGE	42.16	22.12	3.96	4.61	0.2334	81.2%
	Lodge	37.09	18.79	5.58	4.85	0.2423	86.5%
	TempoMOE	**25.13**	10.96	**7.98**	**6.86**	**0.2446**	–
FineDance	真实值	–	–	9.73	7.44	0.2120	–
	FACT	113.38	97.05	3.36	6.37	0.1831	93.4%
	Bailando	82.81	28.17	7.74	6.25	0.2029	81.5%
	EDGE	94.34	50.38	8.13	6.45	0.2116	76.3%
	Lodge	50.00	35.52	5.67	4.96	0.2269	69.1%
	TempoMoE	**38.42**	**25.62**	**8.57**	**6.83**	**0.2316**	–
PopDanceSet	真实值	–	–	8.32	7.68	0.2603	–
	FACT	37.62	26.32	5.63	6.13	0.2162	86.3%
	Bailando	29.56	22.47	5.92	6.29	0.2253	82.3%
	EDGE	34.58	23.72	6.13	6.48	0.2334	80.9%
	POPDG	27.13	21.41	6.52	6.37	0.2403	78.4%
	TempoMOE	**23.42**	**16.18**	**7.54**	**7.12**	**0.2482**	–

表2。模型架构与效率对比。报告的运行时间仅包括动作生成，不包括音乐特征提取。若将特征提取纳入其中，我们的效率优势会更加明显，因为轻量级 Librosa 特征的获取速度远快于 Jukebox 嵌入，而后者还需要额外的模型。

方法	音乐特征	#模型参数（M）		步数	时间（秒）
EDGE	Jukebox 4800 维度	单一	49.91	50	1.3
POPDG	Jukebox 4800 维度	单一	101.41	50	2.8
Lodge	Librosa 35 维度	Two	108.23	100	0.8
Ours	Librosa 35 维度	单一	70.13	10	0.6

表3. 专家组上的消融结果。结果证明了我们的异构多尺度设计在捕捉多样且节奏对齐的动作方面的有效性。

方法	$FID_k \downarrow$	$Div_k \uparrow$	BAS ↑
同质 同尺度	34.52	6.23	0.2126
同质多尺度	29.73	6.92	0.2249
异质多尺度（我们的方法）	**25.13**	**7.98**	**0.2446**

用户研究。 为了评估实际应用中的用户感知，我们邀请了20位具有舞蹈经验的参与者开展用户研究。每位参与者观看20组视频对，对比我们的方法与现有方法，并根据视觉质量、节奏同步性以及与音乐的风格匹配度，选择表现更好的舞蹈。如表1所示，我们的方法始终获得更高的偏好评价，凸显了其在真实场景中生成节奏感强、深受用户喜爱的舞蹈的能力。

表4。多尺度节拍专家的消融结果。结果验证了在每个速度级别组中使用多尺度节拍专家的有效性。

方法	$FID_k \downarrow$	$Div_k \uparrow$	BAS ↑
仅四分音符	34.52	6.16	0.2249
仅二分音符	31.62	6.42	0.2213
仅全音符	28.92	6.86	0.2142
混合（我们的方法）	**25.13**	**7.98**	**0.2446**

表5. 路由特征消融结果。由于缺少标签，PopDanceSet 中不包含基于流派的路由。

数据集	方法	$FID_k \downarrow$	$Div_k \uparrow$	BAS ↑
AIST++	音乐	**25.13**	**7.98**	**0.2446**
	舞蹈	31.26	6.68	0.2136
	流派	29.24	6.92	0.2268
FineDance	音乐	**38.42**	**8.57**	**0.2316**
	舞蹈	44.28	7.12	0.2172
	流派	52.48	6.12	0.2018
PopDanceSet	音乐	**23.42**	**7.54**	**0.2482**
	舞蹈	26.48	6.72	0.2272
	流派	–	–	–

模型效率。 表 2比较了模型架构和运行时效率。与一些需要多个模型或高维特征的先前方法不同——

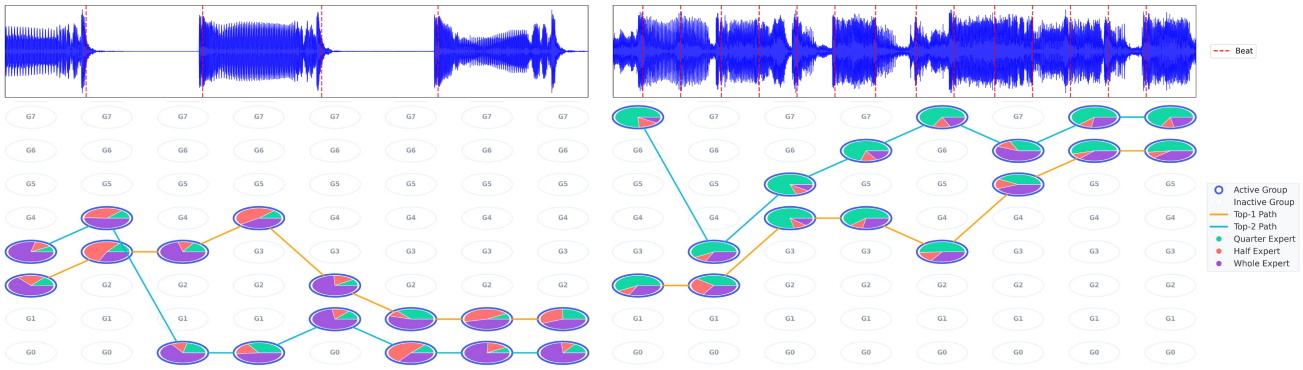


图4. 慢速（64.09 每分钟节拍数）和快速（184.75 每分钟节拍数）样本的逐样本路由。慢节奏会启用低每分钟节拍数组，并从四分之一拍专家切换至整拍专家，以捕捉长程动作；而快节奏会激活高每分钟节拍数组，并依靠四分之一拍专家实现快速、细粒度的动作。

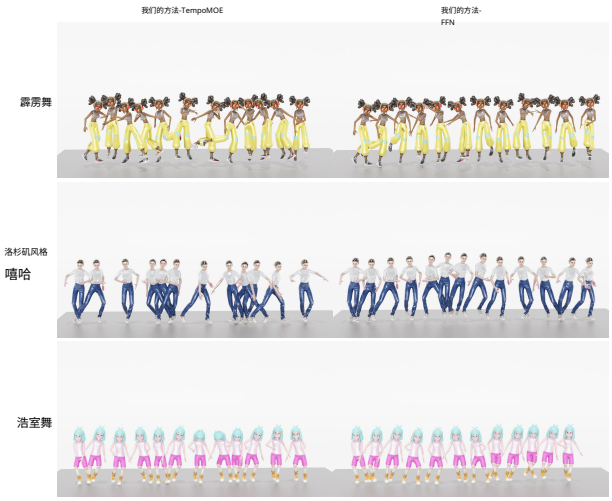


图5. TempoMoE 与标准前馈网络（FFN）基线在三种音乐流派上的定性比较。TempoMoE 生成的动作更加多样、富有表现力且节奏更加连贯，凸显了节奏感知专家路由的优势。详情请参阅补充视频。

例如，Jukebox 嵌入需要额外的模型进行特征提取，而且速度慢得多；我们的方法仅使用单个模型和轻量级的 35 维 Librosa 特征，计算速度很快。尽管模型复杂度降低，我们仍取得了有竞争力的性能，同时保持了快速推理，仅用 0.6 秒便生成 1024 个动作帧。此运行时间仅统计动作生成，不包括特征提取；在实际使用中，若计入特征计算，我们的效率优势将更加显著。这表明，我们的框架高效且实用，适用于实时或大规模应用。

5.2. 消融研究

专家组。我们从两个维度研究专家设计的影响：结构同质性（同质与异质）和时间分辨率（单尺度与多尺度）。具体而言，我们比较了：（1）同质单尺度：所有组均使用相同的专家，固定为 1/2 拍尺度；（2）同质多尺度：所有组共享一套滤波器，尺度为 1/4、1/2 和 1 拍；（3）异质多尺度（我们的方法）：每组使用适配其目标 BPM 范围的专属滤波器。如表 3所示，我们的设计取得了最佳结果，凸显了组内尺度多样性和组间专门化在建模节奏连贯、速度自适应舞蹈方面的优势。

多个节拍尺度专家。我们考察在每个专家组中使用多个节拍尺度是否能提升性能。我们比较了四种设置：(1) 仅四分音符，(2) 仅二分音符，(3) 仅全音符，其中所有专家都使用单一固定尺度；(4) 混合（我们的方法），其中每个组都包含全部三种节拍尺度专家。如表 4所示，混合设置取得了最高分，证实了对不同节奏分辨率下的动作动态进行建模有助于提升节奏自适应性。

专家组数量。如表8所示，使用 8 个专家组可获得最佳的整体性能。当组数过多（例如 16 或 32）时，专家之间的每分钟节拍数间隔会过小，导致节奏范围重叠，并增加训练难度。相反，组数过少（例如 4）则会限制模型捕捉多样化节奏变化的能力。8 组配置实现了良好的平衡，因为其每分钟节拍数间隔（20 每分钟节拍数）与先前研究 [13, 14]以及实际的舞蹈感知阈值相吻合：约 20 每分钟节拍数的节奏差异在感知上可以区分。

表6. 组间路由的消融结果。激活相邻的前2个速度组（TOP-2）取得了最佳性能，表明有限的硬路由配合互补的组重叠能够提升对BPM变化的鲁棒性。

方法	$FID_k \downarrow$	$Div_k \uparrow$	BAS ↑
TOP-1	27.36	7.16	0.2326
TOP-2（我们的方法）	**25.13**	**7.98**	**0.2446**
Soft	33.53	6.32	0.2136
平均	31.28	6.12	0.2142

表7. 组内路由消融结果。在每个组内对多尺度节拍专家进行软融合，可提升动作质量和节奏对齐效果。

方法	$FID_k \downarrow$	$Div_k \uparrow$	BAS ↑
TOP-1	29.53	6.59	0.2176
TOP-2	27.32	7.16	0.2243
软（我们的方法）	**25.13**	**7.98**	**0.2446**
平均	33.16	6.12	0.2156

路由特征。如表5所示，我们评估了三种用于专家路由的特征：（1）舞蹈：动作特征；（2）流派：流派嵌入特征；（3）音乐（我们的方法）：音乐特征。流派路由在 FineDance 上表现不佳，该数据集包含 19 个分布高度不均衡的长尾子流派。舞蹈路由的表现则在 PopDanceSet 上尤为逊色，因为其中的动作序列往往受到噪声影响。相比之下，音乐特征提供了稳定且具有前瞻性的信号，使模型能够在不同数据集上实现更准确、泛化能力更强的路由。

组间路由。为评估专家组选择策略，我们比较了以下方法：（1）TOP-1：仅激活排名第一的组；（2）TOP-2（我们的方法）：激活排名前两位的组；（3）软：对所有组进行加权求和；（4）平均：为所有组分配相同权重。表6显示，TOP-2表现最佳。激活两个相邻组可提升对BPM变化的鲁棒性，并避免使用软路由或均匀路由时观察到的专业化能力损失。这证实了有限硬路由结合互补组重叠的价值。

组内路由。我们探索了每组内的融合策略：(1) TOP-1、(2) TOP-2、(3) 软路由（我们的方法）和 (4) 平均。如表7所示，软路由取得了最佳性能，因为它能够自适应地整合四分之一拍、半拍和整拍专家，实现细粒度节奏控制。相比之下，硬选择（TOP-1/2）和均匀平均可能会导致信息损失。

与FFN基线的视觉比较。我们将基线中的FFN替换为TempoMoE，以评估专家路由的影响。如图5所示，FFN基线生成的动作重复性高、表现力较弱，也难以捕捉节奏变化或流派特有的模式。相比之下，TempoMoE生成的序列流畅、多样且与节奏同步得更好，能够更充分地体现音乐结构和风格细微差异。这表明，节奏结构化专家组和节奏引导路由能够有效利用速度线索，比标准前馈网络更准确地建模时间动态。

表8。Tempo专家组数量的消融结果。使用8组在捕捉多样的节奏变化与维持训练稳定性之间取得了最佳平衡，因为20 BPM间隔与人类感知阈值相符。

组数	$FID_k \downarrow$	$Div_k \uparrow$	BAS ↑
4	27.64	7.08	0.2328
8	**25.13**	**7.98**	**0.2446**
16	33.53	6.32	0.2146
32	38.28	6.03	0.2092

5.3. TempoMOE 与路由分析

样本级动态。 图4 展示了两个具有代表性的样本的路由行为：一个慢速样本（64.09 每分钟节拍数）和一个快速样本（184.75 每分钟节拍数）。在低速度样本中，模型逐渐从四分之一拍专家转向整拍专家，并激活低每分钟节拍数专家组，以捕捉较长范围的结构。相比之下，快速样本始终选择高每分钟节拍数专家组，并更多地依赖四分之一拍专家来建模快速、细粒度的动作。这些模式表明，路由器能够根据输入速度动态调整路由策略。我们进一步在附录A.6中进行了跨多个数据集的路由行为数据集级统计分析。

更多结果和视频请参阅补充材料。

6. 结论

我们提出了 TempoMoE，这是一种速度感知的舞蹈生成框架，它将速度结构化专家组与节奏引导路由相结合，以合成与节拍对齐且富有表现力的三维舞蹈动作序列。通过将生成过程与节奏结构进行分层对齐，并将输入动态路由至专门处理不同速度的专家，我们的方法能够有效适应多样的速度和复杂的音乐模式。大量实验结果证实，TempoMoE 在舞蹈动作质量和生成效率方面均表现出色。

参考文献

[1] Ho Yin Au、Jie Chen、Junkun Jiang 和 Yike Guo。Choreograph：一种在风格和速度一致的动态图上进行音乐条件自动编舞的方法。载于*ACMMM*，第 3917–3925 页，2022 年。3

[2] Dmytro Bazela、Tetiana Pavliuk、Olha Babych、Nataliya Batieieva 和 Petro Kyzim。21世纪初当代全球编舞艺术发展趋势。《SPORT》，第 60–60 页，2022年。2[3] Trevor Boffone。《叛逆者：从 Dubsmash 到 TikTok 的数字舞蹈文化》。牛津大学出版社，2021年。3[4] Steven Brown 和 Lawrence M Parsons。舞蹈神经科学。《科学美国人》，299(1):78–83，2008年。1[5] Birgitta Burger 和 Petri Toiviainen。舞蹈动作中的时间与同步。《表演时间：音乐与舞蹈中的同步与时间流动》，第 217 页，2023年。2[6] Kang Chen、Zhipeng Tan、Jin Lei、Song-Hai Zhang、Yuan-Chen Guo、Weidong Zhang 和 Shi-Min Hu。Choreomaster：编舞导向的音乐驱动舞蹈合成。《TOG》，40(4):1–13，2021年。3[7] Zeyuan Chen、Hongyi Xu、Guoxian Song、You Xie、Chenxu Zhang、Xin Chen、Chao Wang、Di Chang 和 Linjie Luo。X-dancer：富有表现力的音乐到人体舞蹈视频生成。ICCV，2025年。3[8]Damai Dai、Chengqi Deng、Chenggang Zhao、Rx Xu、Huazuo Gao、Deli Chen、Jiashi Li、Wangding Zeng、Xingkai Yu、Y Wu 等。Deepseekmoe：迈向混合专家语言模型中的终极专家专精。载于计算语言学协会年会，第 1280–1297 页，2024年。3[9] Simon Dixon、Elias Pampalk 和 Gerhard Widmer。基于周期模式的舞曲分类。国际音乐信息检索年会，2003年。2[10] Congyi Fan、Jian Guan、Xuanjia Zhao、Dongli Xu、Youtian Lin、Tong Ye、Pengming Feng 和 Haiwei Pan。对齐你的节奏：利用门控增强的节奏感知特征表示生成高度对齐的舞蹈姿势。arXiv 预印本 arXiv:2503.17340，2025年。1, 3[11] William Fedus、Barret Zoph 和 Noam Shazeer。Switch Transformer：以简单高效的稀疏性扩展至万亿参数模型。《机器学习研究期刊》，23(120):1–39，2022年。3[12] Joao P Ferreira、Thiago M Coutinho、Thiago L Gomes、Jos´e F Neto、Rafael Azevedo、Renato Martins 和 Erickson R Nascimento。学习舞蹈：一种利用音频生成逼真舞蹈动作的图卷积对抗网络。《计算机与图形学》，第 11–21 页，2021年 3 –。[13] Fabien Gouyon 和 Simon Dixon。舞曲分类：一种基于速度的方法。载于国际音乐信息检索年会，第 501–504 页，2004年 2, 7 –。[14] David Hammerschmidt、Clemens W¨ollner、Justin London 和 Birgitta Burger。迪斯科时间：音乐中感知时长与速度之间的关系。《音乐与科学》，4: 2059204320986384，2021年。2, 4, 7[15] Ruozi Huang、Huang Hu、Wei Wu、Kei Sawada、Mi Zhang 和 Daxin Jiang。舞蹈革命：利用课程学习生成与音乐相伴的长时舞蹈。载于 ICLR，2020年。3[16] Zikai Huang、Xuemiao Xu、Cheng Xu、Huaidong Zhang、Chenxi Zheng、Jing Qin 和 Shengfeng He。Beat-it：节拍同步的多条件三维舞蹈生成。载于欧洲计算机视觉会议，第 273–290 页。施普林格，2024年。1, 3[17] Robert A Jacobs、Michael I Jordan、Steven J Nowlan 和 Geoffrey E Hinton。自适应局部专家混合。《神经计算》，3(1):79–87，1991年。2, 3[18] Ali Johnson。TikTok舞蹈的版权：互联网时代的编舞。《华盛顿大学法律评论》，96:1225，2021年。3[19] Michael I Jordan 和 Robert A Jacobs。分层专家混合与期望最大化算法。《神经计算》，6(2):181–214，1994年。2, 3[20] Costas I Karageorghis、Leighton Jones 和 Daniel C Low。运动心率与音乐速度偏好之间的关系。《运动与体育研究季刊》，77(2):240–250，2006年。2[21] Gayle Kassing。《舞蹈史：互动艺术方法》。Human Kinetics，2007年。2[22] Gayle Kassing。《探索舞蹈》。Human Kinetics，2024年。2[23] Hsin-Ying Lee、Xiaodong Yang、Ming-Yu Liu、Ting-Chun Wang、Yu-Ding Lu、Ming-Hsuan Yang 和 Jan Kautz。随音乐起舞。NIPS，32，2019年。1, 3[24] Dmitry Lepikhin、HyoukJoong Lee、Yuanzhong Xu、Dehao Chen、Orhan Firat、Yanping Huang、Maxim Krikun、Noam Shazeer 和 Zhifeng Chen。GShard：通过条件计算和自动分片扩展巨型模型。arXiv预印本 arXiv:2006.16668，2020年。3[25] Jiaman Li、Yihang Yin、Hang Chu、Yi Zhou、Tingwu Wang、Sanja Fidler 和 Hao Li。使用Transformer学习生成多样化舞蹈动作。arXiv预印本 arXiv:2008 08171，2020年。3。[26] Ruilong Li、Shan Yang、David A Ross 和 Angjoo Kanazawa。AI编舞师：基于音乐条件的3D舞蹈生成，使用AIST++。载于ICCV，第13401–13412页，2021年。1, 5[27] Ronghui Li、Junfan Zhao、Yachao Zhang、Mingyang Su、Zeping Ren、Han Zhang、Yansong Tang 和 Xiu Li。FineDance：用于3D全身舞蹈生成的细粒度编舞数据集。载于IEEE/CVF国际计算机视觉会议论文集，第10234–10243页，2023年。1, 5[28]Ronghui Li、Hongwen Zhang、Yachao Zhang、Yuxiang Zhang、Youliang Zhang、Jie Guo、Yan Zhang、Xiu Li 和 Yebin Liu。Lodge++：借助生动的编舞模式实现高质量长序列舞蹈生成。arXiv预印本 arXiv:2410.20389，2024年。3[29] Ronghui Li、YuXiang Zhang、Yachao Zhang、Hongwen Zhang、Jie Guo、Yan Zhang、Yebin Liu 和 Xiu Li。Lodge：一种由特征舞蹈基元引导的粗到细扩散网络，用于长序列舞蹈生成。载于 CVPR，第1524–1534页，2024年。3, 5[30] Yunxin Li、Shenyuan Jiang、Baotian Hu、Longyue Wang、Wanqi Zhong、Wenhan Luo、Lin Ma 和 Min Zhang。Uni-MoE：通过专家混合扩展统一多模态大语言模型。TPAMI，2025年。2, 3[31] Bin Lin、Zhenyu Tang、Yang Ye、Jiaxi Cui、Bin Zhu、Peng Jin、Jinfa Huang、Junwu Zhang、Yatian Pang、Munan Ning 等。MoE-LLaVA：面向大型视觉语言模型的专家混合。arXiv预印本 arXiv:2401.15947，2024年。2, 3

[32] Justin London。什么是音乐速度。《表演时间：音乐与舞蹈中的同步与时间流动》，第85–97页，2023年。2[33] Cheng Lu、Yuhao Zhou、Fan Bao、Jianfei Chen、Chongx-uan Li 和 Jun Zhu。DPM-Solver++：用于扩散概率模型引导采样的快速求解器。arXiv预印本 arXiv:2211.01095，2022年。5[34] Zhenye Luo、Min Ren、Xuecai Hu、Yongzhen Huang 和 Li Yao。POPDG：借助PopDanceSet生成流行的3D舞蹈。载于CVPR，第26984–26993页，2024年。1, 2, 5[35] Guangtao Lyu、Chenghao Xu、Jiexi Yan、Muli Yang 和 Cheng Deng。通过稀疏可解释表征迈向统一的人体动作与语言理解。载于ICLR，2025年。3[36] Yueen Ma、Yuzheng Zhuang、Jianye Hao 和 Irwin King。3D-MoE：一种通过整流流实现3D视觉与姿态扩散的多模态专家混合大语言模型。arXiv预印本 arXiv:2501.16698，2025年。3[37] Guy Madison、Fabien Gouyon、Fredrik Ull´en 和 Kalle H¨ornstr¨om。对音乐诱发人体动作倾向的建模：不同音乐流派中与低级音频描述符的初步相关性。《JEPHPP》，37(1):1578，2011年。5[38] Brian McFee、Colin Raffel、Dawen Liang、Daniel PW Ellis、Matt McVicar、Eric Battenberg 和 Oriol Nieto。Librosa：使用Python进行音频与音乐信号分析。SciPy，2015:18–24，2015年。3[39] Dirk Moelants。炒作与自然速度：舞曲速度的长期研究。载于ICMPC，2008年。2[40] Mohammed Muqeeth、Haokun Liu 和 Colin Raffel。通过自适应路由实现专家软合并。arXiv预印本 arXiv:2306.03745，2023年。3[41] Ferda Ofli、Engin Erzin、Y¨ucel Yemez 和 A Murat Tekalp。Learn2Dance：学习音乐到舞蹈的统计映射以合成编舞。TMM，14(3):747–759，2011年。3[42] William Peebles 和 Saining Xie。使用Transformer的可扩展扩散模型。载于ICCV，第4195–4205页，2023年。3[43] Joan Puigcerver、Carlos Riquelme、Basil Mustafa 和 Neil Houlsby。从稀疏专家混合到软专家混合。arXiv预印本 arXiv:2308.00951，2023年。3[44] Aditya Ramesh、Prafulla Dhariwal、Alex Nichol、Casey Chu 和 Mark Chen。基于CLIP潜变量的分层文本条件图像生成。arXiv预印本 arXiv:2204.06125，1(2):3，2022年。3[45] Foram Niravbhai Shah、Parshwa Shah、Muhammad Us-ama Saleem、Ekkasit Pinyoanuntapong、Pu Wang、Hongfei Xue 和 Ahmed Helmy。DanceMosaic：具有多模态可编辑性的高保真舞蹈生成。arXiv预印本 arXiv:2504.04634，2025年。1, 3[46] Noam Shazeer、Azalia Mirhoseini、Krzysztof Maziarz、Andy Davis、Quoc Le、Geoffrey Hinton 和 Jeff Dean。规模惊人的大型神经网络：稀疏门控专家混合层。arXiv预印本 arXiv:1701.06538，2017年。3[47] Takaaki Shiratori、Atsushi Nakazawa 和 Katsushi Ikeuchi。随音乐起舞的角色动画。载于《计算机图形论坛》，第449–458页。Wiley Online Library，2006年。1

[48] Li Siyao、Weijiang Yu、Tianpei Gu、Chunze Lin、Quan Wang、Chen Qian、Chen Change Loy 和 Ziwei Liu。Bailando：采用演员-评论家GPT和编舞记忆的3D舞蹈生成。载于CVPR，第11050–11059页，2022年。1, 5[49] Li Siyao、Weijiang Yu、Tianpei Gu、Chunze Lin、Quan Wang、Chen Qian、Chen Change Loy 和 Ziwei Liu。Bailando++：具备编舞记忆的3D舞蹈GPT。TPAMI，45(12)：14192–14207，2023年。3[50] Taoran Tang、Jia Jia 和 Hanyang Mao。Dance with melody：一种面向音乐的舞蹈合成LSTM自编码器方法。载于ACMMM，第1598–1606页，2018年。3[51] Guy Tevet、Sigal Raab、Brian Gordon、Yonatan Shafir、Daniel Cohen-Or 和 Amit H. Bermano。人体动作扩散模型，2022年。3[52] Jonathan Tseng、Rodrigo Castellon 和 Karen Liu。EDGE：基于音乐的可编辑舞蹈生成。载于CVPR，第448–458页，2023年。1, 3, 5[53] Hongsong Wang、Yin Zhu 和 Xin Geng。借助风格描述提示实现灵活的音乐条件舞蹈生成。arXiv预印本 arXiv:2406.07871，2024年。1, 3[54] Qing Wang、Xiaohang Yang、Yilan Dong、Naveen Raj Govindaraj、Gregory Slabaugh 和 Shanxin Yuan。DanceChat：由大型语言模型引导的音乐到舞蹈生成。arXiv预印本 arXiv:2506.10574，2025年。3[55] Zilin Wang、Haolin Zhuang、Lu Li、Yinmin Zhang、Jun-jie Zhong、Jun Chen、Yu Yang、Boshi Tang 和 Zhiyong Wu。通过自动排序的演示和奖励模型探索3D舞蹈生成。载于AAAI，第301–309页，2024年。[56] Chenghao Xu、Lyu Guangtao、Yan Jiexi、Yang Muli 和 Cheng Deng。大型语言模型也懂肢体语言：将语音转换为人体手势。载于计算语言学协会年会，第14734–14751页，2024年。3[57] Kaixing Yang、Xulong Tang、Ziqiao Peng、Yuxuan Hu、Jun He 和 Hongyan Liu。MEGADance：用于流派感知的3D舞蹈生成的专家混合架构。arXiv预印本 arXiv:2505.17543，2025年。3[58] Zijie Ye、Haozhe Wu、Jia Jia、Yaohua Bu、Wei Chen、Fanbo Meng 和 Yanfeng Wang。ChoreoNet：探索借助编舞动作单元实现音乐到舞蹈的合成。载于ACMMM，第744–752页，2020年。3[59] Canyu Zhang、Youbao Tang、Ning Zhang、Ruei-Sung Lin、Mei Han、Jing Xiao 和 Song Wang。用于舞蹈生成的双向自回归扩散模型。载于CVPR，第687–696页，2024年。1, 3[60] Ye Zhu、Kyle Olszewski、Yu Wu、Panos Achlioptas、Menglei Chai、Yan Yan 和 Sergey Tulyakov。通过舞蹈视频实现复杂音乐生成的量化生成对抗网络。载于欧洲计算机视觉会议，第182–199页。施普林格，2022年。3[61] Wenlin Zhuang、Congyi Wang、Jinxiang Chai、Yangang Wang、Ming Shao 和 Siyu Xia。Music2Dance：用于音乐驱动舞蹈生成的DanceNet。TOMM，18(2)：21，2022年。1, 3。

节奏速度作为稳定线索：用于音乐到三维舞蹈生成的节奏速度与节拍专家的层次化混合

补充材料

A. 附录

A.1. 各流派节奏速度分布详情

为进一步考察音乐流派与节奏速度之间的关系，我们使用小提琴图可视化 AIST 中各流派的节奏速度分布 ++ 。如图 6所示，即使在同一流派内，节奏速度也存在显著差异，通常分布在每分钟 60–200 拍的范围内。这一观察凸显了流派内节奏速度方差较大，说明仅凭流派不足以刻画节奏属性。相比之下，节奏速度为节奏感知建模提供了更直接、更精细的线索。这进一步证明了我们基于节奏速度而非流派构建专家组和路由机制的合理性。

A.2. 节奏结构化专家组详情

层次化时序建模。 为了有效捕捉音乐与舞蹈中固有的多尺度节奏结构，我们构建了节奏结构化专家组，使其能够处理不同范围的节奏速度和节拍粒度。这种层次化设计确保每个专家都能处理与音乐学上有意义的时间尺度相匹配的动作特征，涵盖从快速的节拍级重音到较慢的乐句级动作姿态。

节奏速度离散化。我们将每分钟 60 至 200 个节拍的目标节奏范围划分为八个相互重叠的速度区间，每个区间以一个锚点值为中心：

$$60, 80, 100, 120, 140, 160, 180, 200 \text{ BPM}. \tag{19}$$

每个锚点对应一个专门的专家组，负责处理与该速度区间的节奏动态相匹配的动作。

节拍尺度专家层级。在每个速度组内，我们定义了三个对应不同节奏粒度的专家：

- **四分之一拍专家**：$_{1\,4}$ 拍分辨率
- **二分之一拍专家**：$_{1\,2}$ 拍分辨率
- **整拍专家**：1 拍分辨率

这些专家以一维逐通道时序卷积实现，其卷积核大小会根据节奏自适应，并经过计算以匹配各节奏单位的时长。

卷积核大小计算。给定视频帧率 f（单位为每秒帧数）、目标速度（BPM）和节拍分辨率 $r \in \frac{1}{4}$，$_{1\,2}$，1，我们首先计算每拍的帧数：

$$F_b = \frac{60 \times f}{\text{BPM}}. \tag{20}$$

卷积核大小 k 的计算公式为：

$$k = \lceil r \times F_b \rceil_{\text{odd}}, \tag{21}$$

其中 $\lceil \cdot \rceil_{\text{odd}}$ 表示向上取整到最近的奇数，以确保对称感受野。

示例。 假设标准视频帧率为 $f = 30$帧/秒，速度为 $= 120$ 每分钟节拍数。那么每拍的帧数为：

$$F_b = \frac{60 \times 30}{120} = 15. \tag{22}$$

对于每种节拍分辨率，卷积核大小为：

$$k_{\frac{1}{4}} = \left\lceil \frac{1}{4} \times 15 \right\rceil_{\text{odd}} = \lceil 3.75 \rceil_{\text{odd}} = 5, \tag{23}$$

$$k_{\frac{1}{2}} = \left\lceil \frac{1}{2} \times 15 \right\rceil_{\text{odd}} = \lceil 7.5 \rceil_{\text{odd}} = 9, \tag{24}$$

$$k_1 = \lceil 1 \times 15 \rceil_{\text{odd}} = \lceil 15 \rceil = 15. \tag{25}$$

这确保专家卷积核与音乐所决定的时间分辨率上的节奏保持对齐。

总结。 所有速度组的卷积核大小汇总于表9。这种节奏感知构造具有三大优势：(1) 节奏自适应性，使每个专家的感受野与音乐速度相匹配；(2) 多尺度建模，在每组中纳入节拍细分；(3) 音乐归纳偏置，将与速度对齐的时间先验融入模型架构。

表9. 速度结构化专家组中各专家的卷积核大小。

速度（每分钟节拍数）	四分之一拍专家	二分之一拍专家	整拍专家
60	9	15	31
80	7	11	23
100	5	9	19
120	5	9	15
140	5	7	13
160	3	7	11
180	3	5	11
200	3	5	9

A.3. 模型效率的详细比较

所有实验均在配备 NVIDIA A6000 GPU 的同一台机器上进行。表10 中报告的推理时间反映了生成 1024 帧长期舞蹈动作所需的平均运行时间，不包括音乐特征提取所需的时间。值得注意的是，依赖高维音频表征的方法，例如

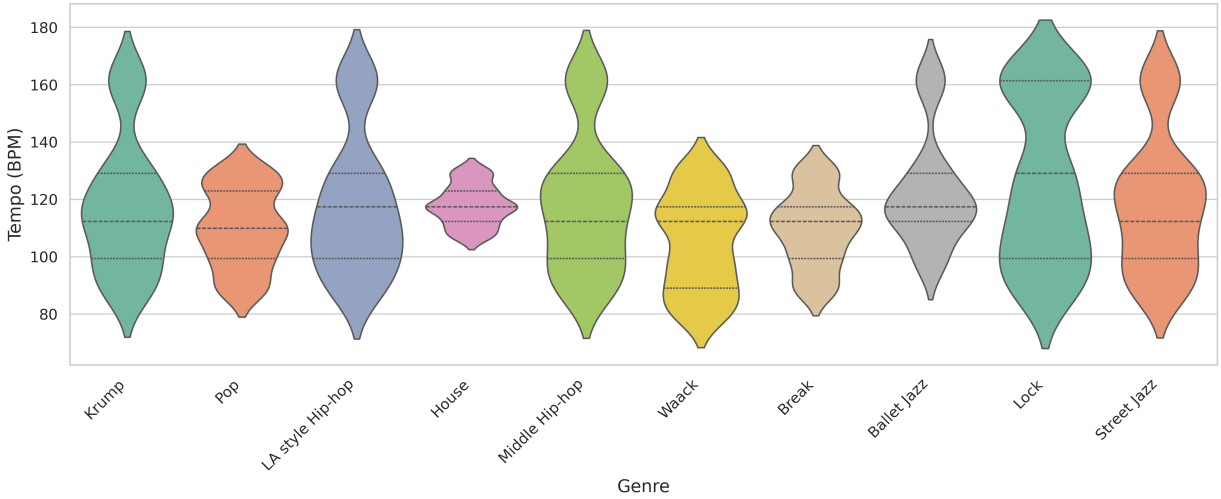


图6. AIST++中不同舞种的速度分布，以小提琴图呈现。尽管归属于同一舞种，速度值仍有很大差异，通常在60至200每分钟节拍数之间，这表明每种舞蹈内部都具有显著的节奏多样性。因此，我们基于速度而非粗略或含糊的流派标签对节奏模式进行建模。

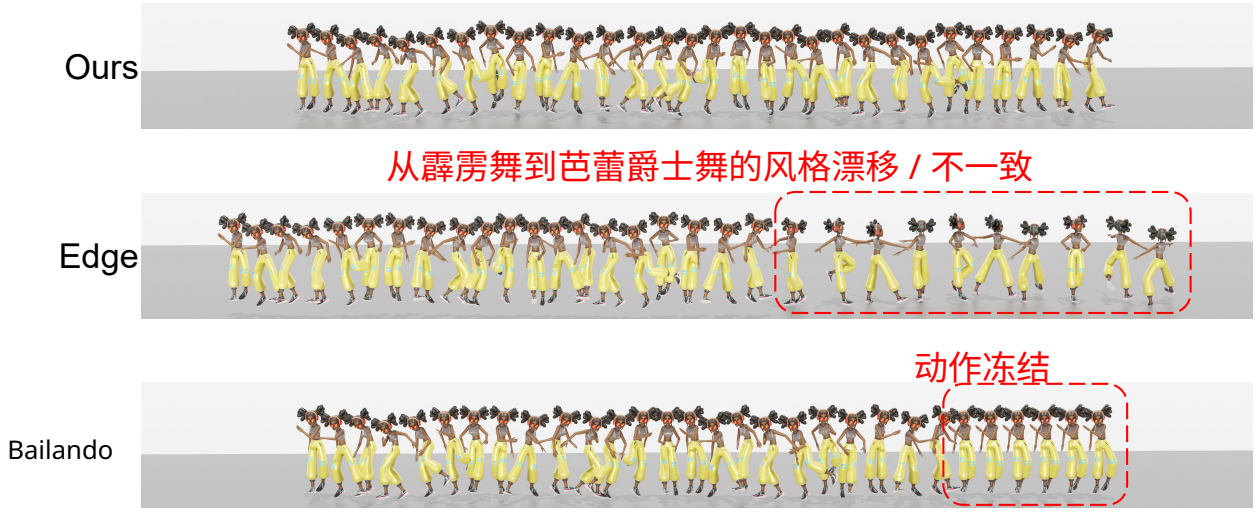


图7. 霹雳舞流派的定性比较。红色方框标出了多种动作问题，包括动作冻结、风格漂移和穿插伪影。我们的方法在整个序列中保持了风格和动作质量的一致性。

Jukebox，与使用轻量级 Librosa 特征的方法相比，预处理成本要高得多。因此，在实际部署中，其实际端到端延迟还会更高。

除了运行时间更快之外，我们的模型在推理时还不受输入序列长度限制。相比之下，现有基线方法通常受固定长度训练输入的限制，并倾向于生成时长相匹配的序列。此外，我们的架构无需进行输入截断或填充，从而增强了在实际应用中的鲁棒性和实用性。

此外，我们的方法通过采用单一模型，并且仅需 10 个去噪步数即可进行舞蹈动作生成，从而实现了显著的效率提升。这一设计带来了 2×–5× 的推理加速，同时生成了多样、质量更高且节奏更加协调的舞蹈序列。相比之下，以往的方法通常依赖 50–100 个迭代步数，在某些情况下还需要多个模型组件。

表10。模型架构、音乐特征和推理效率对比。推理时间不包括音乐特征提取。我们的方法通过采用单一模型，并且只需 10 个去噪步数即可生成动作，实现了显著的效率提升，推理加速达到 2×–5×。

方法	音乐特征	#模型	#参数 (M)	步数	时间 (s)
EDGE	Jukebox 4800 维度	单一	49.91	50	1.3
POPDG	Jukebox 4800 维度	单一	101.41	50	2.8
Lodge	Librosa 35 维度	Two	108.23	100	0.8
Ours	Librosa 35 维度	单一	70.13	10	0.6

A.4. Break舞蹈类型的详细示例

如图7所示，我们展示了一个具有代表性的生成霹雳舞序列示例，以说明不同模型之间的定性差异。

EDGE 展现出流派不一致的问题：尽管生成的动作起初呈现出类似霹雳舞的特征，却逐渐偏向更像芭蕾爵士舞的动作。这种退化主要源于 EDGE 的节奏感知能力有限——其生成过程未对节拍对齐或速度变化进行显式建模，因此容易产生风格不兼容的片段，在霹雳舞等节奏密集型流派中尤为明显。红框标出了由此产生的风格突变和节奏错位的过渡。

另一方面，Bailando 在序列后半段出现了动作停滞和动力学退化的问题。这一问题源于累积解码误差，以及其冻结的 VQ-VAE 动作表示缺乏灵活性。随着生成以自回归方式进行，细微的预测误差会不断累积，而固定的标记空间限制了模型根据不断变化的音乐提示进行修正或调整的灵活性。因此，解码器难以维持多样且能结合上下文的动作模式，导致序列末尾的动作多样性降低并趋于重复。

相比之下，我们的方法在整个序列中始终保持强烈的节奏感知能力和风格一致性。通过利用速度感知专家和路由机制，该方法生成流畅的霹雳舞动作，并使其与底层音乐保持良好的同步。

A.5. 不同速度下动作差异的可视化

图 10 展示了两种音乐流派——流行舞和爵士芭蕾——下生成舞蹈的定性比较。每种音乐流派均采用两种不同的速度设置进行呈现：快速（198 或 160 BPM）和慢速（99 或 80 BPM），分别记为 (a)–(d)。此分析在保持流派不变的同时，单独考察速度调制的影响，从而让我们能够观察模型如何根据节奏步调调整动作模式。

对于 (a) 和 (b) 中的流行舞，我们观察到动作节奏和空间范围均存在明显差异。在 198 BPM 下（a），生成的序列以高频、局部的动作为特点，例如快速摆臂、快速点脚和上半身分离动作。编舞与快节奏的节拍紧密契合，因而动作紧凑且与节奏同步。在 99 BPM 下（b），较慢的节奏使动作更加舒展、流畅，包括身体旋转、全身转换和幅度更大的步伐。这表明模型能够根据速度调整动作粒度和动作表现，同时保留流行舞的风格特征。

对于 (c) 和 (d) 中的爵士芭蕾，由于流派特征不同，速度适应呈现出不同的表现。在 160 BPM (c) 下，模型会生成紧凑的垂直跳跃、迅捷的躯干动作和频繁的旋转，体现了该流派动感而精致的动作语汇。相比之下，80 BPM 版本 (d) 展现出由低到高的转换（例如从地面起身）、挥扫式手臂动作以及大幅度的跳跃。这些动作体现出更大的空间延展性和结构表现力，这得益于较慢的节拍。值得注意的是，模型在调整动作密度和时机的同时，仍保持了符合流派特点的优雅。

总之，这些结果证实，我们的方法能够在多种音乐条件下实现精准且富有表现力的编舞控制。

A.6. 路由分析的数据集级统计。

图8 汇总了各数据集上的总体路由情况。AIST++ 在 100 和 160 BPM 附近呈现出显著的激活；PopDanceSet 的激活主要集中在 120–140 BPM，并倾向于使用四分之一拍专家；FineDance 的 BPM 分布范围更广，且更多依赖整拍专家，这反映出其侧重富有表现力、以乐句为单位的编舞。在所有数据集中，较低层优先关注更精细的节奏细节，而较高层则越来越多地关注长期结构——这表明模型能够有效地进行分层节奏建模。

A.7. 与前馈网络基线的视觉比较

我们通过将 TempoMoE 与一个采用前馈网络而非专家路由的基线进行比较，评估 TempoMoE。如图9所示，前馈网络基线往往会生成重复性强、表现力较弱的动作，而且经常无法准确契合节奏变化或流派特征。相比之下，以 TempoMoE 替换前馈网络后，生成的动作更加流畅、多样，并与输入音乐的节奏更加同步。TempoMoE 生成的序列能够忠实捕捉细粒度的音乐结构（如节拍级重音）以及更广泛的风格细微差异（包括适应音乐流派的动作模式）。这些比较清楚地证明了节奏结构化专家组和分层节奏引导路由的有效性。TempoMoE 通过动态利用速度线索，将不同的动作基元路由至专门的专家，从而能够比扁平的前馈网络基线更准确、灵活地建模时间动态，生成质量更高、音乐连贯性更强的舞蹈动作。

A.8. 最新分布外 TikTok 样本上的失败案例

补充材料包含若干使用热门 TikTok 舞蹈音乐的测试案例。这些片段往往混音繁多且由用户精选，具有不规则的节奏、突变和非典型节拍结构 [3, 7,18, 60], ，这给音乐到舞蹈生成带来了巨大挑战，也凸显了在真实世界的用户生成内容场景中采用更稳健建模策略的必要性

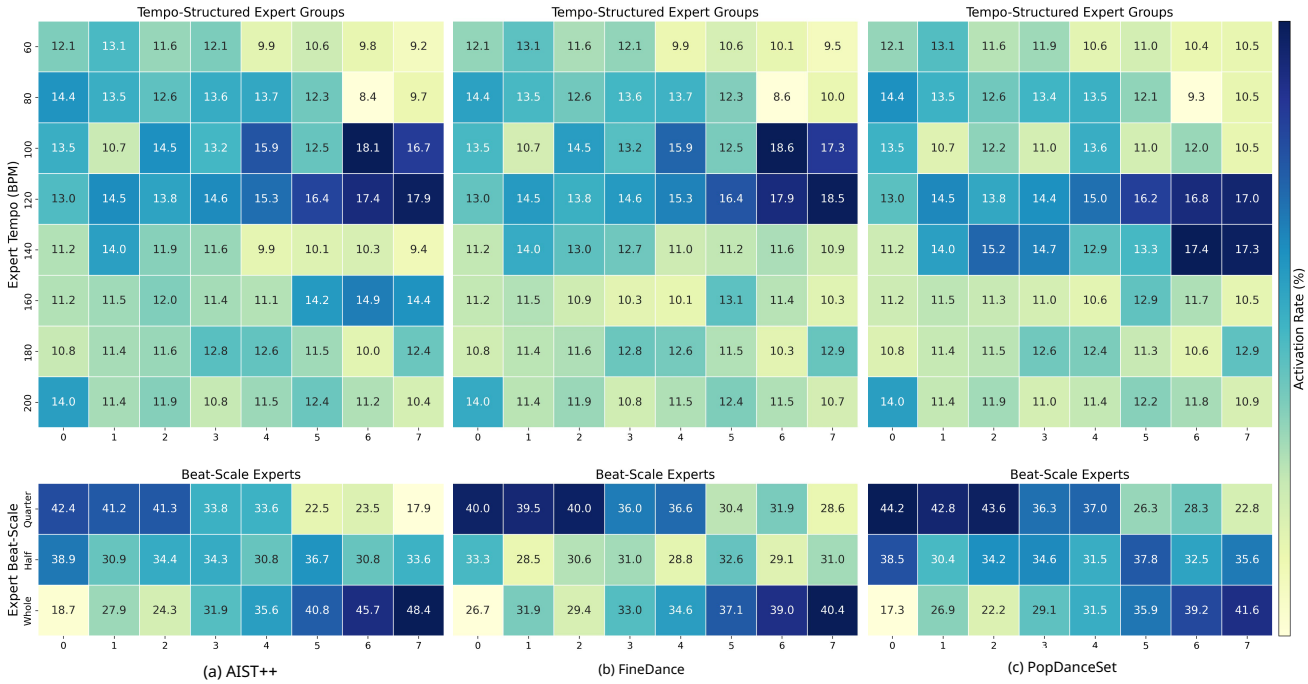


图 8. AIST++、PopDanceSet 和 FineDance 上的数据集级路由统计。TempoMoE 针对每个数据集呈现出不同的 BPM 和节拍尺度激活模式：AIST++ 在接近 100 和 160 BPM 处达到峰值，PopDanceSet 的速度集中在 120–140 BPM 左右，并倾向于选择四分之一拍专家；FineDance 的 BPM 覆盖范围更广，且更多依赖整拍专家。在所有数据集中，较低层捕捉细粒度的节奏细节，而较高层则侧重于长期动作结构。这些模式表明，TempoMoE 能够有效适应不同的速度分布和层次化节奏结构，从而支持节奏准确且富有表现力的舞蹈生成。

。如演示视频所示，EDGE 和我们的方法在这些条件下都表现不佳。EDGE 依赖高层 Jukebox 嵌入，往往无法捕捉有意义的节奏模式，生成静止或不合拍的动作。相比之下，我们的方法表现更好，这得益于结合使用低层 Librosa 特征和速度感知路由，它们能在多样化的音乐输入中维持更稳定的分布。这使节奏感知更加一致，动作质量也略有提升

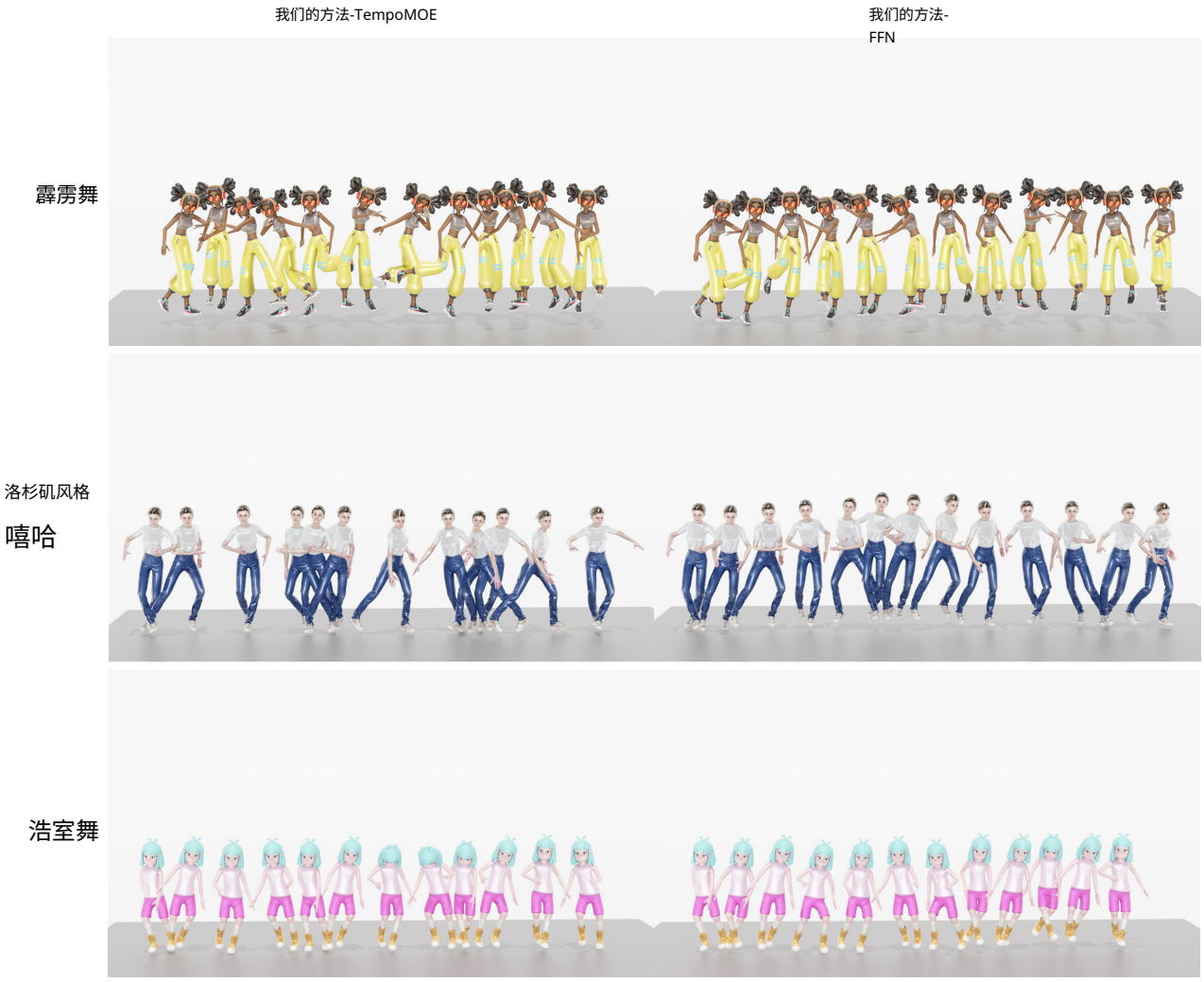


图 9. 我们的 TempoMoE 模型与基于标准前馈网络（FFN）的基线模型在三种音乐流派上的定性比较。TempoMoE 生成的动作更加多样、富有表现力且节奏连贯性更强，能够更好地与音乐乐句和风格先验相契合。这凸显了我们的节奏感知混合专家架构在捕捉特定流派及节奏敏感的编舞模式方面的优势。详情请参阅补充视频。

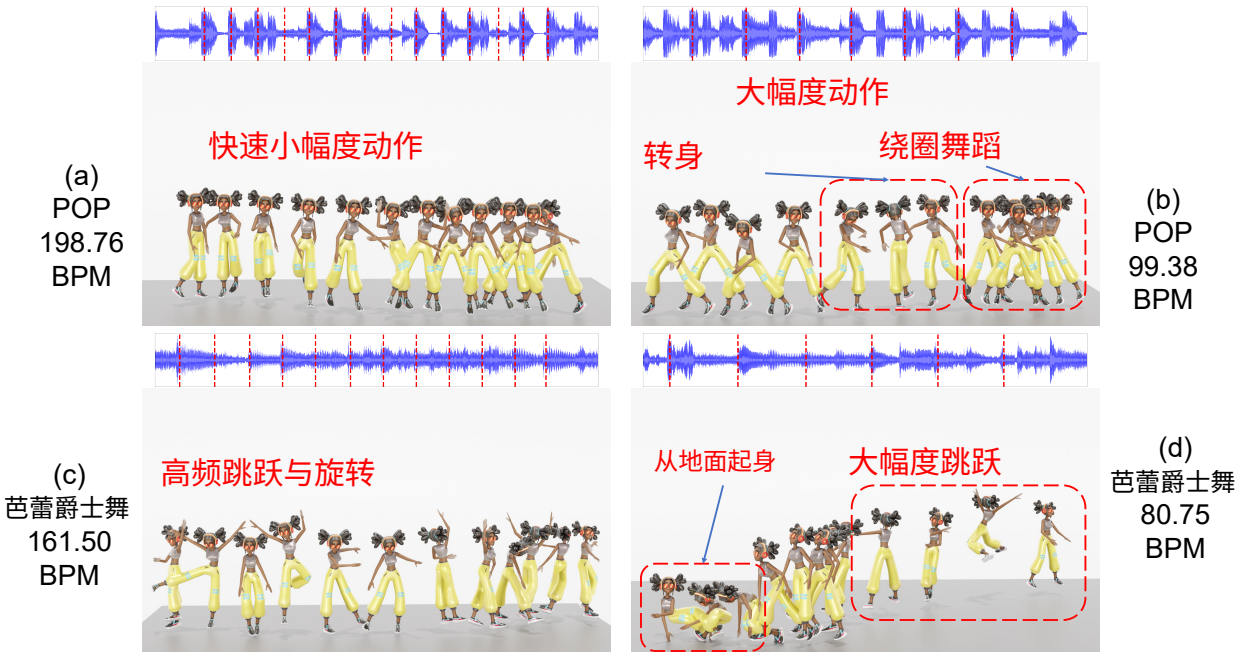


图10. 不同音乐流派和速度设置下生成舞蹈的定性可视化结果。(a) 流行舞，198 BPM：快节奏会产生密集且局部化的动作，包括快速的手臂摆动和高频分离动作。(b) 流行舞，99 BPM：较慢节奏允许幅度较大的动作，例如身体旋转和全身转换。(c) 爵士芭蕾，160 BPM：快节奏会产生紧凑但充满活力的动作，其特点是频繁旋转和小幅度垂直跳跃。(d) 爵士芭蕾，80 BPM：慢节奏能够实现舒展且富有表现力的编舞，包括从地面到站立的转换和大幅度跳跃。这些结果表明，我们的模型能够解耦速度与流派，从而实现可控且风格一致的舞蹈生成。