arXiv:2603.08023v1 [cs.CV] 9 Mar 2026

不拘泥于 Transformer：摒弃节拍表示，使用基于 Mamba 的扩散模型生成舞蹈

Sangjune Park[1] InhyeokChoi[1] Donghyeon Soon[2] Youngwoo Jeon[1] Kyungdon Joo[1*]蔚山科学技术院，韩国[2]大邱庆北科学技术院，韩国{psj9116, inhyeok.choi, youngwoo.jeon, kyungdon}@unist.ac.kr dhsoon@dgist.ac.kr

摘要

舞蹈是一种人体运动形式，以情感表达和交流为特征，在音乐、虚拟现实和内容创作等多个领域发挥着作用。现有的舞蹈生成方法往往无法充分捕捉舞蹈固有的序列性、节奏性以及与音乐同步的特征。本文提出MambaDance，一种利用基于 *Mamba* 的扩散模型进行舞蹈生成的新方法。*Mamba* 擅长处理长序列和自回归序列，我们将其融入两阶段扩散架构，以取代现成的 *Transformer*。此外，考虑到音乐节拍在舞蹈编排中的关键作用，我们提出一种基于高斯的节拍表示，显式引导舞蹈序列的解码。在 *AIST++*和 *FineDance* 数据集上针对不同序列长度进行的实验表明，与以往方法相比，我们提出的方法能够有效生成合理的舞蹈动作，并在从短舞蹈到长舞蹈的各种序列长度下始终体现其关键特征。更多定性结果和演示视频请访问 https://vision3d-lab.github.io/mambadance。

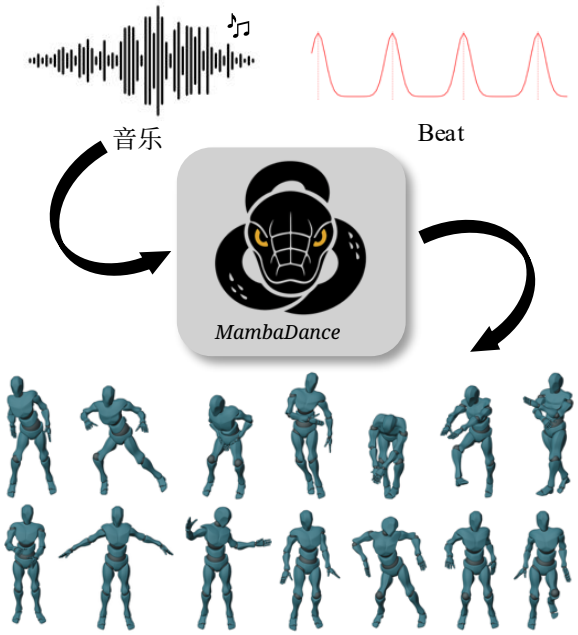


图 1。我们提出了 *MambaDance*，这是一种基于 Mamba 的两阶段扩散框架，采用信息丰富的高斯节拍表示。该方法能够在 AIST++ [13]和 FineDance [14]上生成不同长度的连贯且与节拍同步的动作。

1. 引言

生成与音乐同步且富有表现力、逼真的舞蹈动作是一项长期存在的挑战。舞蹈作为具身交流和艺术表达的媒介，以其结构化、富有节奏的特性为特征。传统上，编舞需要专家手动完成或逐帧制作动画，这两种方式都成本高昂且耗时。近年来，生成模型的发展推动了音乐驱动的舞蹈自动生成 [12, 13, 15, 17, 24, 25, 29] ，其应用涵盖娱乐、内容创作、游戏和虚拟现实。这些发展为手动编舞提供了可扩展的替代方案，但也带来了建模舞蹈中固有的时间复杂性和节奏性的技术挑战。

基于 Transformer 的架构 [12, 13, 15, 17, 24, 25, 29] 因能够建模全局时间依赖关系，而被广泛应用于当前的三维舞蹈生成方法。然而，舞蹈序列所需的不只是对上下文的一般理解，还需要强时间因果性，以及在较长时间范围内保持一致的自回归结构。Transformer [30] 模型从根本上缺乏对序列推进的归纳偏置，在生成长动作序列时往往效率低下且不连贯，

*通讯作者。

这使得生成兼具节奏对齐性和时间连贯性的动作颇具挑战。

除架构局限之外，音乐节拍的表示在组织舞蹈动作方面也起着关键作用。例如，在教学中，编舞首先被切分为节拍长度乐句（例如 8 拍），学习者先跟随这些锚点练习，再配合音乐完成整段曲目的表演。许多现有方法 [13, 15, 17, 24, 25, 29] 在音乐特征向量中纳入简单的一维节拍特征。尽管这类表示能够提供粗略对齐，但它们并未显式建模节拍如何影响动作序列。Beat-It [10] 将节拍线索与音频特征解耦，以加强节拍与舞蹈对齐，同时引入带有辅助对齐损失的节拍距离估计器。然而，该方法将节拍影响隐式编码在网络激活中，而不是建模一个随时间塑造动作的显式时间先验。这两个挑战，即长序列与自回归序列建模以及有效的节拍条件控制，看似相互独立，实则密切相关。不充分的序列建模会削弱节奏对齐，而不充分的节拍表示则会限制模型生成结构化且富有表现力的动作的能力。因此，需要一种统一方法同时应对这两个方面，从而生成既在时间上一致、又在节奏上同步的舞蹈。

如图 1所示，我们提出了*MambaDance*，这是一种基于 Mamba 的两阶段扩散架构构建的音乐到舞蹈生成框架。与以往的 Transformer 或混合设计不同，我们的模型完全以状态空间模块取代注意力，从而高效捕捉长程动态和自回归动态。我们遵循 Lodge [15]，整体的两阶段范式，但扩散去噪网络中的解码器块包含：(i) 处理动作特征的单模态 Mamba，以及 (ii) 将动作与音乐融合的跨模态 Mamba，随后接入自适应线性调制。为强化节奏结构，我们提出了一种基于高斯衰减的新型节拍表示，可围绕音乐节拍提供平滑且易于解释的时间先验。该表示使模型能够生成遵循音乐内在节奏乐句的动作。我们在 AIST++ [13] 和 FineDance [14] 数据集上使用不同的序列长度进行评估，以系统验证模型在多种序列长度下的优越性能，并展示其长度鲁棒合成能力。实验结果表明，*MambaDance*始终能够生成比现成方法在物理上更合理、且与节奏更加吻合的舞蹈动作。

综上所述，我们的贡献如下：

- 我们提出了一种新颖的框架，*MambaDance*，一种完全基于 Mamba 的扩散模型，专为自回归音乐到舞蹈生成而设计。
- 我们提出了一种基于高斯的节拍表示，它显式编码节奏结构以指导运动解码，并考虑音乐节拍的关键特性。
- 我们进行了全面的实验和消融研究，结果表明我们的方法在保真度和节奏对齐指标上始终优于以往的基线方法。

2. 相关工作

2.1. 人体运动生成

人体运动生成对于逼真动画和交互系统至关重要，引起了视觉、图形学和机器人学领域的关注。早期方法侧重于动作检索与插值，对于行走等简单动作只能得到合理但有限的结果。

深度生成模型的最新进展使神经网络能够生成更加灵活、富有表现力的人体运动。基于 GAN 的方法 [1]采用对抗训练和多尺度判别器来提升真实感并建模过渡动态表现。针对特定任务的损失函数进一步有助于捕捉动作不确定性。基于自编码器的方法将动作压缩到潜在空间中，通常使用循环模型或 Transformer 模型 [11, 22]来对齐动作与语言。近期工作通过身体部位感知向量量化 [23, 27]提升了表现力。扩散模型已成为一种有力的替代方案，兼具真实感与可控性。文本条件模型 [28, 34]将动作视为去噪过程，而潜在空间变体 [2] 则提升了效率和结构性。

然而，高效处理长自回归序列仍然颇具挑战，尤其是受可扩展性问题影响的基于 Transformer 的扩散模型。MotionMamba 等近期工作 [35] 通过采用状态空间模型（SSM）应对这一问题，该模型结合分层时间建模与双向空间建模，从而同时提升长期一致性和生成速度。尽管如此，如何有效建模长序列并融入时间结构或节奏线索，仍是有待解决的重要挑战。

2.2. 舞蹈生成

早期关于音乐驱动的舞蹈生成的研究将该任务视为自回归建模问题，并采用了传统机器学习技术。尽管这些方法捕捉到了基本的动作-音乐对齐，但它们缺乏对多样化音乐输入的适应性，并且难以生成长而多样的动作序列。

基于 Transformer 的模型（例如 FACT [13], ）在建模时间动态方面展现出强大的能力。此外，结合 VQ-

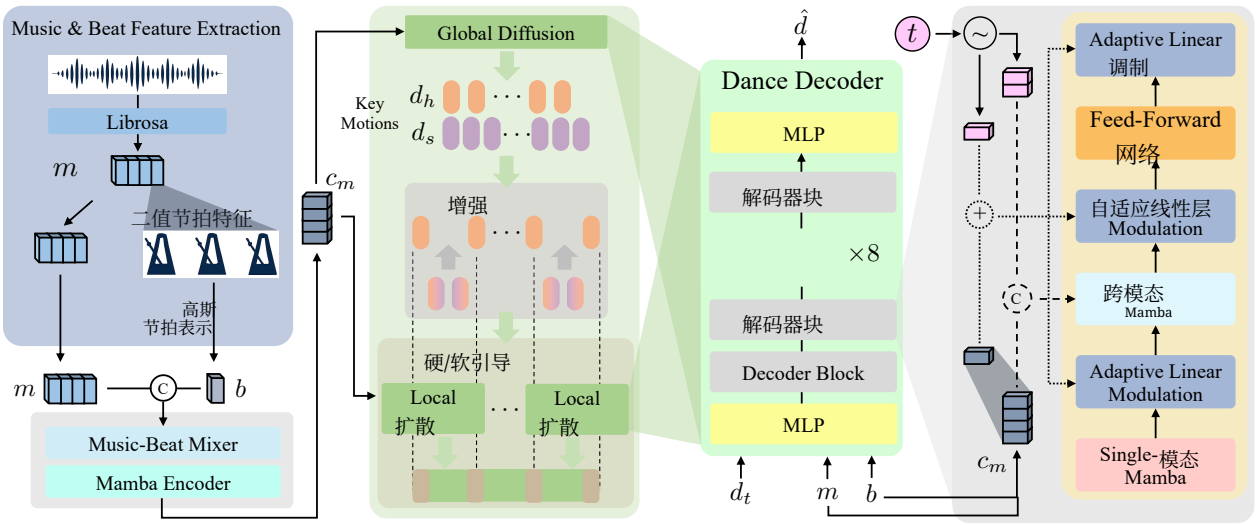


图 2. **MambaDance**的整体架构。 我们提取音乐特征 m，并从该特征的节拍二值掩码中提取一种新颖的节拍表示 b（蓝色方框）。两阶段扩散架构使我们的方法能够在单次推理中实现长度无关生成（绿色方框）。扩散模型的解码器由我们提出的基于 Mamba 的模块构成，例如单模态Mamba（SMM）、跨模态Mamba（CMM）和自适应线性调制（AdaLM）（灰色方框）。

变分自编码器与 GPT 的方法（以 Bailando [24]，为代表），以及包括 EDGE [29]、Lodge [15]，和 POPDG [17]，在内的扩散框架，进一步提升了生成舞蹈动作的真实感、多样性和可编辑性。然而，由于这些方法基于 Transformer 架构，因此在对舞蹈数据进行建模时面临挑战，因为这类数据通常序列较长，并具有自回归依赖关系。近期，若干研究 [32, 33] 采用 Mamba 结构 [3, 6]，以有效捕捉舞蹈数据的自回归特性。然而，这些研究大多采用混合架构，仅将 Mamba 与 Transformer 部分结合。虽然此类设计提升了局部连续性，但仍存在注意力机制的二次复杂度和不连续性问题。

2.3. 用于运动与舞蹈的状态空间模型

近期的舞蹈生成方法强调了长序列建模以及与音乐同步的重要性。基于Transformer的模型因能够捕捉全局依赖而得到广泛应用，但在生成长序列时，往往难以兼顾自回归一致性与计算效率。状态空间模型（SSM）近期已成为序列建模的一种有前景的替代方案。具体而言，Mamba [3, 6] 已被开发出来，可高效建模人体运动和舞蹈等序列数据中的长程自回归依赖。借助时变参数，Mamba比Transformer更有效地捕捉长期依赖，同时保持线性时间复杂度。此外，其结构天然蕴含序列归纳偏置，这有助于持平滑的时间连续性。

近期的多项研究将Mamba结构用于人体运动与舞蹈生成。MotionMamba [35]利用分层状态空间模型，增强人体运动生成中的时间连贯性和空间运动建模。AlignYourRhythm [5]集成了基于Mamba的节奏感知模块，以改善音乐-运动对齐。MegaDance [33] 和 MatchDance [32]采用Mamba-Transformer混合架构，利用Mamba建模局部依赖，并利用Transformer捕捉全局上下文。尽管这些研究成功发挥了Mamba的优势，但它们采用混合架构，将Mamba与Transformer部分结合。

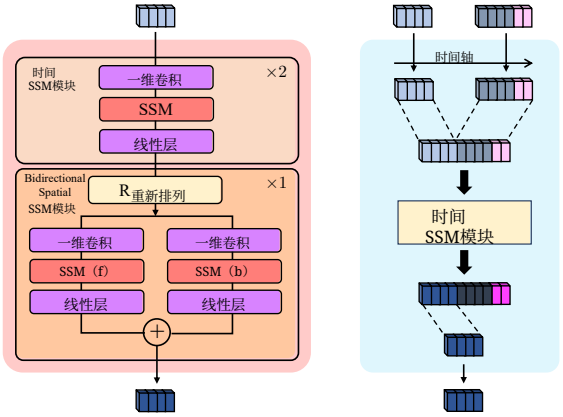
相比之下，据我们所知，我们的工作首次在舞蹈生成中完全以基于 Mamba 的模块替换基于 Transformer 的模块，包括单模态和跨模态组件。我们实现了结构化、节奏对齐的三维舞蹈生成，具备时间一致性和更高的计算效率，同时充分利用 Mamba 的自回归归纳偏置进行长序列建模。

3. 方法

在本研究中，我们提出了 *MambaDance*，这是一种两阶段扩散框架，利用 Mamba [3, 6] 高效的长序列建模和自回归序列建模能力，实现音乐驱动的三维舞蹈生成。该模型在利用状态空间模型对长序列和自回归三维动作序列进行建模的归纳偏置的同时，生成合理且自然的舞蹈动作。鉴于

音乐节拍在编舞中组织并锚定动作的重要作用，我们提出了一种新的节拍表示，能够显式增强节拍的影响。该表示提供了直观的控制信号，并在舞蹈生成过程中改善节奏引导。

如图 2所示，*MambaDance*以从单段原始音乐中提取的音乐特征和节拍特征为输入，并输出三维舞蹈序列。为了通过单次推理生成一段长舞蹈序列，全局扩散会生成关键动作。通过镜像和连接相邻动作来增强关键动作。随后，局部扩散基于这些基本动作，在硬/软引导下生成细节动作。这两个扩散模型共用相同的舞蹈解码器结构，该结构由两个多层感知机和 8 个解码器模块组成。每个解码器模块主要包含：(i) 单模态 Mamba（SMM），(ii) 跨模态Mamba（CMM），以及 (iii) 自适应线性调制（AdaLM）；该结构完全基于 Mamba 设计，并替代了所有注意力模块。



单模态Mamba（SMM）　　跨模态Mamba（CMM）

图 3. **单模态 Mamba（左）和跨模态 Mamba（右）**。对于输入跨模态 Mamba 的序列，浅蓝色、深蓝色和粉色方块分别对应动作、条件和时间步令牌。

3.1. 用于舞蹈生成的 Mamba

遵循先前的工作 [15, 29], ，我们将切片后的舞蹈动作表示为一系列 SMPL [16, 20] 姿态 $d \in \mathbb{R}^{l \times D_{\text{motion}}}$。这里，$l \in \{N, n\}$ 表示从总长度为 L 的动作或音乐序列中截取的片段长度，其中 N 和 n 分别用于全局扩散和局部扩散。给定音乐特征 $m \in \mathbb{R}^{l \times D_{\text{music}}}$ 和节拍表示 $b \in \mathbb{R}^{l \times 1}$，基于多层感知机的音乐-节拍混合器与 Mamba 编码器会生成节拍突出条件 $c_m \in \mathbb{R}^{l \times E}$，在解码过程中注入其他模态的信息。在基于 Mamba 的舞蹈解码器中，输入多层感知机将时间步 t 的带噪动作序列 d_t 映射为潜变量 $z_d \in \mathbb{R}^{l \times E}$，一组解码器模块在条件信息的作用下处理该潜变量，输出多层感知机再将其映射回 $\hat{d} \in \mathbb{R}^{l \times D_{\text{motion}}}$。每个解码器模块包含一个 单模态*Mamba*（*SMM*）、一个跨模态*Mamba*（*CMM*）、一个轻量级前馈网络（FFN），以及自适应线性调制（*AdaLM*）。以 SSM 替代自注意力、以 CMM 替代交叉注意力，可实现线性时间序列处理，并与基于注意力的设计相比，获得稳定的长时域生成 [3, 6]。

单模态Mamba（SMM）仅处理动作潜变量，而跨模态 Mamba（CMM）则将动作潜变量与音乐和扩散时间步令牌融合（图 3）。SMM 通过两个时间 SSM 模块和一个双向空间 SSM 模块转换输入的噪声动作潜变量。时间 SSM 模块遵循 [

3, 6], 沿长度轴传播序列 $a \in \mathbb{R}^{l \times E}$ l。空间 SSM 模块会重新排列序列，使其 $a' \in \mathbb{R}^{E \times l}$，并以双向方式沿潜在通道轴 E 传播，以促进通道间的协调。相比之下，CMM 通过拼接动作潜变量 z_d、音乐条件（音乐与节拍信息的混合） c_m 以及扩散时间步令牌 $e_t \in \mathbb{R}^{2 \times E}$ 作为 $[z_d, c_m, e_t]$，在拼接后的序列上应用时间 SSM，并截取输出中的动作部分。这些模块是注意力模块的有效替代方案，尤其适用于具有长序列、复杂且时间连续特征的三维舞蹈数据。

我们提出自适应线性调制（AdaLM），作为 FiLM [21], 的替代方法。FiLM 是一种简单的基于归一化的调制器，用于调节 1D 输入序列。与二维视觉中的自适应归一化技术 [4, 9]类似，我们对组归一化后的潜变量应用逐特征仿射调制：

$$\text{AdaLM}(z_d, \gamma, \beta) = (1 + \gamma)\,\text{GroupNorm}(z_d) + \beta, \quad (1)$$

其中，参数 γ, $\beta \in \mathbb{R}^E$ 通过对均值池化条件向量 $c_{\text{mod}} \in \mathbb{R}^E$进行线性投影得到，该向量由 c_m 和 e_t计算得出。

3.2. 节拍表示

先前的工作已经发现，音乐节拍与舞蹈动作节拍之间具有很强的一致性 [10, 13, 15, 17,24, 29]。捕捉这种节奏联系对于连贯且富有表现力的生成至关重要。在许多现有方法中 [13, 15, 17, 24, 29], , 节拍信息被纳入音乐特征。例如，Librosa [18] 常被用于提取音乐特征 $m_{\text{total}} \in \mathbb{R}^{L \times D_{\text{music}}}$，其中 $D_{\text{music}} = 35$ 的最后一个通道维度对应于二值节拍信号 $b_{\text{raw}} \in \mathbb{R}^{L \times 1}$。Beat-It [10]通过对每一帧编码其到最近节拍的时间距离来缓解稀疏性，但所得信号仍然是单调的，并未显式建模非节拍帧的衰减影响。

在实践中，节拍将编舞划分为乐句，并锚定高动能动作。因此，一种有效的表示应满足两个关键属性：(i) 更接近节拍的帧携带更强的信号，并且(ii) 该强度随时间距离迅速而平滑地衰减。我们提出高斯节拍表示，一种基于高斯衰减函数的直观节拍表示 $b_{\text{total}} \in \mathbb{R}^{L\times 1}$，它同时满足这两项标准，并为解码提供可解释的、时间局部化的线索。为简洁起见，我们在本小节中省略下标“total”。

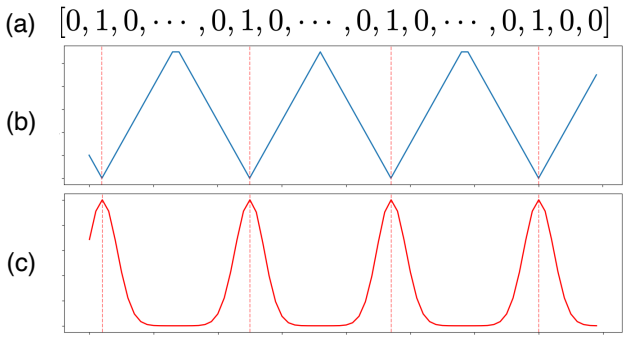


图 4. **原始节拍 (a)、最近节拍距离（NBD）(b) 和我们提出的高斯节拍表示 (c) 的可视化结果**。水平轴表示序列的帧索引（时间步），垂直轴表示信号。NBD 和所提出表示的信号范围分别为 $[0, 11]$ 和 $[0, 1]$,。

如图 4所示，令 b_{raw} 表示二值节拍序列 $b_{\text{raw}}(i) \in \{0, 1\}$ 在帧 $i \in \{0, \ldots, L-1\}$处的值。借鉴 Beat-It 的思路，我们将最近节拍距离（NBD）定义为每一帧到其前后最近节拍帧的最小距离：

$$\text{NBD}(i) = \min\left(\text{dist}^{-}(i), \text{dist}^{+}(i)\right), \tag{2}$$

其中，若存在前一个节拍，则 $\text{dist}^{-}(i) = i - \max(j \le i | b_j = 1)$，否则为 $+\infty$；若存在后一个节拍，则 $\text{dist}^{+}(i) = \min(j \ge i | b_j = 1) - i$，否则为 $+\infty$。这里，j 表示序列中的帧索引。设节拍帧为 $\tau_0 < \tau_1 < \cdots < \tau_{M-1}$。为覆盖完整序列并获得局部节奏度量，我们在边界处补入 $\tau_{-1} = 0$和 $\tau_M = L - 1$，并将拍间间隔 $l(i) = \tau_{k+1} - \tau_k$定义为 $\tau_k \le i \le \tau_{k+1}$。为满足两个要求，即节拍附近的信号更强，且信号远离节拍时平滑、快速衰减，同时对不同节奏进行归一化，我们采用平滑因子为$\alpha \in$ （0，1）的高斯衰减来构建节拍表示：

$$b(i) = \exp\left(-\frac{\text{NBD}(i)^2}{2\left(\alpha \cdot l(i)\right)^2}\right). \tag{3}$$

高斯函数在节拍帧周围提供钟形、平滑且局部化的强调，而自适应节奏的带宽 $\alpha \cdot l(i)$ 则能在不同节拍间隔下提供一致的引导。由此得到的信号是一种明确且易于解释的节奏先验，与舞蹈动作的乐句结构相契合。

3.3. 训练与推理

训练方案与损失。遵循 Lodge [15],，我们独立训练全局与局部扩散模型，并仅在推理时使用全局关键动作（硬线索 d_h 与软线索d_s）引导局部模型。仅通过扩散修复直接控制少数边界帧，可能导致每个局部窗口内的过渡不连贯，因此我们在训练期间将噪声输入 d_t 的首尾 $L_{\text{key}}/2$ 帧替换为对应的真实帧 d_0，以微调局部扩散。

我们基本上对*MambaDance*的舞蹈解码器 f_θ 采用扩散模型的标准重建损失 [8]，定义为：

$$\mathcal{L}_{\text{simple}} = \mathbb{E}_{d,t}\left[\|d - f_\theta(d_t, t, m, b)\|_2^2\right]. \tag{4}$$

额外的辅助损失使训练更稳定并提升物理合理性，包括位置损失 $\mathcal{L}_{\text{pos}}$、速度损失 $\mathcal{L}_{\text{vel}}$、加速度损失 $\mathcal{L}_{\text{acc}}$以及足部接触一致性损失 $\mathcal{L}_{\text{foot}}$。位置损失衡量真实值与预测舞蹈动作之间关节位置的相似度：

$$\mathcal{L}_{\text{pos}} = \frac{1}{l}\sum_{i=1}^{l}\left\|\text{FK}(d^i) - \text{FK}(\hat{d}^i)\right\|_2^2, \tag{5}$$

其中，FK(·) 表示将关节角度转换为关节位置的正向运动学函数。类似地，接触一致性损失可确保足地接触准确：

$$\mathcal{L}_{\text{foot}} = \frac{1}{l-1}\sum_{i=1}^{l-1}\left\|(\text{FK}_{\text{foot}}(\hat{d}^{i+1}) - \text{FK}_{\text{foot}}(\hat{d}^i))\cdot\hat{y}^i\right\|_2^2, \tag{6}$$

其中，$\text{FK}_{\text{foot}}(\cdot)$ 仅对足部关节执行正向运动学函数，$\hat{}$ y 表示预测的二元足部接触标签。速度损失和加速度损失用于评估关节速度和加速度的相似性：

$$\begin{aligned}\mathcal{L}_{\text{vel}} &= \frac{1}{l-1}\sum_{i=1}^{l-1}\left\|(d^{i+1} - d^i) - (\hat{d}^{i+1} - \hat{d}^i)\right\|_2^2,\\ \mathcal{L}_{\text{acc}} &= \frac{1}{l-2}\sum_{i=1}^{l-2}\left\|(v^{i+1} - v^i) - (\hat{v}^{i+1} - \hat{v}^i)\right\|_2^2.\end{aligned} \tag{7}$$

请注意，关节加速度可以通过关节速度 $v^i = d^{i+1} - d^i$ 计算得到。总损失定义为将所有损失按权重组合：

$$\begin{aligned}\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{simple}} &+ \lambda_{\text{pos}}\mathcal{L}_{\text{pos}} + \lambda_{\text{vel}}\mathcal{L}_{\text{vel}}\\ &+ \lambda_{\text{acc}}\mathcal{L}_{\text{acc}} + \lambda_{\text{foot}}\mathcal{L}_{\text{foot}}.\end{aligned} \tag{8}$$

数据集	模型	保真度			Beat	多样性		胜场数（↑）
		FID_k （↓）	FID_g (↓)	PFC (↓)	BAS (↑)	Div_k (→)	Div_g (→)	
FineDance	GT	-	-	0.1852	-	10.9924	7.7424	-
	EDGE	$179.01^{\pm 3.10}$	$1234.28^{\pm 236.95}$	$0.3994^{\pm 0.0177}$	$0.2261^{\pm 0.0013}$	$\mathbf{10.34}^{\pm 0.36}$	$31.74^{\pm 2.76}$	3.5%
	POPDG	$190.23^{\pm 1.18}$	$1479.14^{\pm 6.44}$	$0.3765^{\pm 0.0124}$	$0.2361^{\pm 0.0033}$	$7.22^{\pm 0.14}$	$14.53^{\pm 0.12}$	5.5%
	Lodge	$\underline{84.99}^{\pm 2.07}$	$\underline{64.57}^{\pm 10.74}$	$\underline{0.0585}^{\pm 0.014}$	$\underline{0.2410}^{\pm 0.0063}$	$\underline{7.98}^{\pm 0.20}$	$\mathbf{7.67}^{\pm 0.68}$	$\underline{35.0\%}$
	Ours	$\mathbf{51.36}^{\pm 0.67}$	$\mathbf{43.11}^{\pm 10.54}$	$\mathbf{0.0119}^{\pm 0.0008}$	$\mathbf{0.2441}^{\pm 0.0044}$	$6.38^{\pm 0.17}$	$\underline{6.44}^{\pm 0.88}$	**56.0%**
AIST++	GT	-	-	1.2544	-	9.61	7.78	-
	EDGE	$125.99^{\pm 128.69}$	$\underline{28.72}^{\pm 4.29}$	$3.1883^{\pm 0.5318}$	$\underline{0.2572}^{\pm 0.0112}$	$\mathbf{11.45}^{\pm 3.25}$	$4.91^{\pm 0.56}$	10.5%
	POPDG	$777.32^{\pm 711.65}$	$60.08^{\pm 5.98}$	$4.8615^{\pm 0.6010}$	$0.2318^{\pm 0.0129}$	$24.08^{\pm 7.40}$	$\mathbf{7.87}^{\pm 0.59}$	9.0%
	Lodge	$\underline{67.13}^{\pm 2.79}$	$28.93^{\pm 0.47}$	$\underline{1.4087}^{\pm 0.1296}$	$0.2397^{\pm 0.0158}$	$3.34^{\pm 0.32}$	$3.54^{\pm 0.14}$	$\underline{32.0\%}$
	Ours	$\mathbf{65.86}^{\pm 3.11}$	$\mathbf{26.58}^{\pm 1.02}$	$\mathbf{1.0622}^{\pm 0.2343}$	$\mathbf{0.2701}^{\pm 0.0116}$	$\underline{3.57}^{\pm 0.37}$	$\underline{4.98}^{\pm 0.43}$	**48.5%**

表 1. **FineDance 和 AIST 上的定量结果++ 数据集**。 真实动作用作参考。 ↓表示越低越好， → 表示越接近真实动作越好。最佳和次佳结果分别以粗体和下划线标出。

为进一步抑制根部不必要的漂移，我们在微调局部模型时加入根平移项 $\lambda_{\text{trans}}\mathcal{L}_{\text{trans}}$ （公式（7）中的根部位置）。

采用两阶段扩散模型进行并行推理。 受用于长舞蹈生成的两阶段扩散模型启发 [15], ，我们将推理流程扩展到可处理可变长度。给定总音乐特征 $m_{\text{total}} \in \mathbb{R}^{L\times D_{\text{music}}}$ 和高斯节拍表示 $b_{\text{total}} \in \mathbb{R}^{L\times 1}$，我们将其划分为 k 长度为 N的不重叠片段。回顾一下，切片后的动作和音乐序列长度为 $l \in \{N, n\}$，其中 N 用于全局扩散， n 用于局部扩散。先前的方法仅针对长序列（N， n） $=$ （1024， 256），而我们的推理支持 $N \in \{128, 256, 1024\}$ 和 $n \in \{64, 256\}$，可在保留由粗到细的设计的同时，以多种时间分辨率进行生成。

具体而言，在舞蹈解码器中，使用完整的音乐条件混合$c_{m,\ g} \in \mathbb{R}^{N\times E}$，全局扩散会预测具有代表性的关键动作$m_{\text{key}}$，这些动作以较高的动能捕捉高层次的编舞模式。关键动作 $m_{\text{key}} \in \mathbb{R}^{L_{\text{key}}\times D_{\text{motion}}}$ 包含硬线索 d_h （用于锚定窗口边界）和软线索 d_s （用于塑造窗口内的动态表现）。为利用双侧对称性，我们对 d_s 进行镜像，并将 d_s 置于长度为 n的序列中，以实现软引导。对于跨越多个片段的序列，我们会在提取中间区域基元之前，将片段 i 的最后 L_{key} 帧复制到片段 $i+1$ 的前 L_{key} 帧，以确保连续性。

每个片段 $c^i_{m,\ g}$ 会进一步划分为多个窗口$\{c^j_{m,l}\}_{j=1}^{N//n}$，其中 $c_{m,\ l} \in \mathbb{R}^{n\times E}$。我们在每个窗口的开头和结尾通过扩散修复注入 d_h ，以固定边界姿态，确保可靠拼接；同时，我们将 d_s 用作 $>(1000\cdot s)$ 去噪步的早期步引导，其中软线索引导尺度 s控制引导强度。由于边界已固定，各窗口可独立解码并并行生成，随后将生成的片段拼接起来，形成完整的舞蹈序列。

4. 实验

4.1. 实验设计

我们在 AIST++ [13] 和 FineDance [14]上评估 *MambaDance*。AIST++包含 1,408 段高质量的短三维舞蹈序列，由专业舞者演绎，涵盖 10 种风格。我们遵循 EDGE [29] ，采用切片式动作参数化 $D_{\text{motion}} = 151$ ，并以序列长度 $N = 128$进行训练。FineDance 提供从在线视频中收集的长形式舞蹈序列，涵盖多种舞蹈风格和时长，涉及 16 种风格。我们采用 Lodge [15]中的 139 维表示，序列长度为 $N = 1024$ 。这些数据集共同覆盖了互补的场景——30 FPS 下的短片段和长片段——使我们能够评估模型在不同时间分辨率和数据分布下的鲁棒性。

我们将本方法与近期用于音乐条件三维舞蹈生成的基于Transformer的扩散方法进行比较：以下基线方法通过在音乐条件三维舞蹈生成中采用基于Transformer的扩散架构，展现了近期的进展。

• **EDGE** [29]: 首个将基于Transformer的扩散模型用于三维舞蹈生成的方法，并采用了来自Jukebox的丰富音乐表示。

• **POPDG** [17]: 一种后续方法，利用基于额外对齐模块和空间增强算法改进的扩散模型（iDDPM [19]）。

• **Lodge** [15]: 一种基于 EDGE 的两阶段扩散框架，包含足部细化模块和多流派判别器，专注于长期舞蹈生成。

模型	保真度			Beat	多样性	
	FID_k (↓)	FID_g (↓)	PFC (↓)	BAS (↑)	部门$_k$ (→)	部门$_g$ (→)
GT	-	-	0.1852	-	10.9924	7.7424
Lodge	$84.99^{\pm 2.07}$	$64.57^{\pm 10.74}$	$0.0585^{\pm 0.014}$	$0.2410^{\pm 0.0063}$	$\mathbf{7.98}^{\pm 0.20}$	$\mathbf{7.67}^{\pm 0.68}$
仅 Mamba	$\underline{60.94}^{\pm 2.01}$	$43.46^{\pm 7.29}$	$\underline{0.0180}^{\pm 0.0017}$	$0.2402^{\pm 0.0044}$	$\underline{6.48}^{\pm 0.40}$	$\underline{7.21}^{\pm 0.79}$
Mamba+NBD	$83.55^{\pm 1.35}$	$\mathbf{38.43}^{\pm 0.71}$	$0.0687^{\pm 0.0046}$	$\mathbf{0.2478}^{\pm 0.0054}$	$5.3958^{\pm 0.17}$	$5.78^{\pm 0.30}$
Ours	$\mathbf{51.36}^{\pm 0.67}$	$\underline{43.11}^{\pm 10.54}$	$\mathbf{0.0119}^{\pm 0.0008}$	$\underline{0.2441}^{\pm 0.0044}$	$6.38^{\pm 0.17}$	$6.44^{\pm 0.88}$

表 2。 **Mamba 和节拍表示的消融实验（FineDance）**。 与基于 Transformer 的 Lodge 相比， [15], “仅 Mamba”表示不带显式节拍先验的完全基于 Mamba 的模型。“Mamba+NBD”使用最近节拍距离 [10]作为节拍表示。

4.2. 评估指标

我们采用 EDGE 和 Lodge 中的标准评估指标来评估生成的舞蹈。为衡量动作真实感，我们报告 Fr´echet Inception 距离，分别基于运动学特征（FID_k）和几何特征（FID_g）计算。物理合理性通过物理足部接触分数（PFC）进行评估。节拍对齐分数（BAS）用于衡量生成动作节拍（即动作能量峰值出现的时刻，例如关节速度或动能的局部最大值）与音乐节拍的对齐程度。动作多样性通过多样性指标进行量化，该指标计算运动学特征（Div_k）和几何特征（Div_g）上的平均成对距离。由于这些评估指标无法完全反映人类感知，我们开展了用户研究，并报告我们的方法相较于基线方法的胜率（胜场数）。

不同于以往工作 [15, 17, 29], ，我们在完整长度序列上计算所有评估指标，而不是将序列切分为长度为 N 的片段，从而更好地反映全局时间连贯性，这与为整段音乐生成完整舞蹈的目标一致。此外，我们报告 10 次独立运行的均值和标准差，以体现统计可靠性。有关评估指标和用户研究方案的更多细节，请参见补充材料。

4.3. 比较

我们在 AIST++和 FineDance 上，将*MambaDance*与近期的基于 Transformer的扩散模型、EDGE、POPDG 和 Lodge 进行评估。AIST 的全局/局部长度为（N，n） =（128，64），而 FineDance 的全局/局部长度为（1024，256）。

在两个数据集上，我们的方法在动作保真度和节拍对齐方面均取得最佳分数（FID_k、FID_g、PFC、BAS）（表1）。在 FineDance 上，我们的模型将 FID_k 降至 51.36（Lodge 为 84.99），FID_g 降至 43.11（Lodge 为 64.57），并将 PFC 降至 0.0119（Lodge 为 0.0585），表明动态表现更加逼真，足地交互也显著更加稳定。BAS 略微升至 0.2441（Lodge 为 0.2410），体现出节奏与动作对齐方面持续且显著的提升。在 AIST++ 上，*Mam-baDance* 的 FID_k 达到 65.86（Lodge 为 67.13），26.58

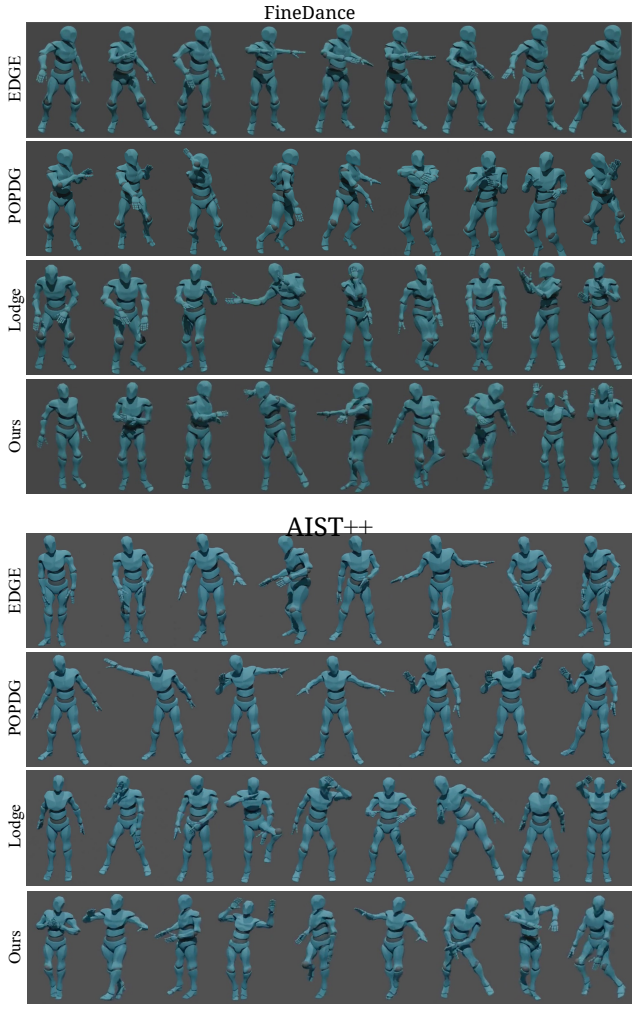


图 5. **FineDance（上）和 AIST++ （下）数据集上的定性比较**。每一行展示一组从完整序列中采样的帧，这些帧以固定间隔截取。

FID_g （EDGE 为 28.72）、PFC 为 1.0622（Lodge 为 1.4087），且 BAS 最高，达到 0.2701（EDGE 为 0.2572）。值得注意的是，我们在各项评估指标上的标准差都很小，说明模型在多次运行中的表现稳定，而

消融实验		评估指标				
AdaLM	CMM	FID_k (↓)	PFC (↓)	BAS (↑)	除以$_k$ (→)	除以$_g$ (→)
GT		-	-	0.1852	10.99	7.74
✗	✗	$\underline{61.28}^{\pm 1.72}$	$\underline{0.6316}^{\pm 0.0066}$	$\underline{0.2465}^{\pm 0.0044}$	$\underline{8.59}^{\pm 0.16}$	$5.77^{\pm 0.41}$
✗	✓	$62.93^{\pm 1.80}$	$0.7878^{\pm 0.0070}$	$\mathbf{0.2479}^{\pm 0.0057}$	$\mathbf{9.93}^{\pm 0.18}$	$\underline{6.11}^{\pm 0.60}$
✓	✓	$\mathbf{51.36}^{\pm 0.67}$	$\mathbf{0.0119}^{\pm 0.0008}$	$0.2441^{\pm 0.0044}$	$6.38^{\pm 0.17}$	$\mathbf{6.44}^{\pm 0.88}$

Ta表 3. **不同解码器结构的消融研究**。 最后一行对应我们的 *MambaDance*。

一些基线方法的方差则非常大，例如 EDGE 和 POPDG 在 AIST++ 上的 FID_k 以及在 FineDance 上的 FID_g 。

在多样性方面，Lodge 的 Div_g 7.67 最接近真实值（7.74），我们的方法以 6.44 的结果位居第二，在 FineDance 上表现如此。在 AIST++ 上，POPDG 的 Div_g 7.87 最接近真实值（7.78），我们的方法以 4.98 的结果位居第二。EDGE 和 POPDG 通常报告最高的多样性，但这些情况往往伴随着较高的 FID/PFC，表明额外的变化性部分源于脚滑等伪影。总体而言，我们的方法更重视物理合理性和节拍对齐，其多样性数值具有竞争力，但相较于真实值略显保守。

结果支持两个结论：（i）以状态空间解码替代注意力机制，能够提升自回归动作保真度和物理合理性，这一点体现在 PFC 的大幅改善上；（ii）显式的节拍先验能够持续带来显著的 BAS 增益。多样性处于平衡状态，而非被最大化，需要在保真度和物理合理性之间进行权衡。此外，正如用户研究中的胜率指标所反映的那样，我们的模型生成的舞蹈序列始终比基线方法生成的序列更受人工评估者青睐。值得注意的是，我们的方法在短序列和长序列上都能保持稳定的性能，而 EDGE 和 POPDG 在生成更长视频时往往会出现质量下降。

如图 5所示，尽管多样性指标可能表明变化有所减少，但定性结果显示，动作充满动态感且与节奏对齐。EDGE 和 POPDG 经常出现足部滑动等伪影，这与较高的 FID 和 Div 指标相吻合；此外，正如补充视频所示，它们还会不自然地一直盯着一侧。我们强烈建议观看补充视频，以便进行全面、详细的比较。

4.4. 消融研究

我们在 FineDance 数据集上分析了所提出的节拍表示和模型架构的影响。

我们在 FineDance [14] 上使用了三种变体：(i) Lodge [15]，一种基于 Transformer 的两阶段扩散模型；(ii)“仅 Mamba”，将 Transformer [30] 完全替换为 Mamba [3, 6] ，且不使用任何节拍表示；(iii)“Mamba+NBD”，使用最近节拍距离 [10] 代替高斯节拍表示，并结合我们的音乐节拍混合器。与 Lodge 相比，“仅 Mamba”显著提升了动作保真度和物理合理性（FID_k、FID_g、PFC 更低），而 BAS 仅略有变化，这表明状态空间模型能让舞蹈解码器生成更真实的动态表现和更稳定的足地交互。引入 NBD 先验后，BAS 最高、FID_g 最低，但运动学真实感和稳定性均大幅下降（例如，PFC 甚至高于 Lodge），多样性也有所降低（Div_k 和 Div_g 与真实值的差距更大）。我们的高斯节拍表示则实现了平衡：BAS 仍然较高（排名第二，且接近 NBD），同时 FID_k 和 PFC 达到最佳值，多样性也有所提升。总而言之，这些结果表明，Mamba 是保真度提升的主要驱动因素，而我们的高斯节拍先验则在不牺牲自然度的前提下保留了节奏对齐。

我们对跨模态 Mamba（CMM）和自适应线性调制（AdaLM）进行消融，其中 CMM 用于替代交叉注意力，AdaLM 则是基于组归一化的线性调制。与交叉注意力模块相比，采用 CMM 的模型具有更高的 BAS 和多样性（动作更加丰富，音频-动作耦合也更紧密）。在 CMM 的基础上加入 AdaLM，可恢复稳定性，并带来整体上最佳的保真度和合理性，同时 BAS 保持相近，多样性则略有下降。这一趋势表明，CMM 能有效融合音乐与动作，而 AdaLM 则对融合后的状态进行正则化，起到轻量级稳定器的作用，有助于改善接触表现和动态表现。

5. 结论

本文提出了*MambaDance*，一种用于音乐到三维舞蹈生成的新方法，完全以Mamba替代Transformer。此外，我们引入了一种基于高斯衰减的信息丰富的节拍表示，以考虑音乐节拍的重要特性。在两个具有不同序列长度的数据集上的实验结果表明，我们的方法相较于基线方法具有稳健性与优越性。尽管在应用方面仍有提升空间，但我们的工作聚焦于以音乐为条件生成人体运动，而未涉及制作流程的下游阶段，例如渲染。一个自然的扩展可以是一个端到端系统

将音乐映射为跳舞的三维虚拟形象，并包含用于摄像机、材质和灯光的渲染栈。

致谢

本研究获得韩国政府（MSIT）资助的信息通信技术规划评估院（IITP）项目支持（No.RS- 2020-II201336, 人工智能研究生院项目（UNIST）；No.RS-2022-II220612，面向具身人工智能的几何与物理常识推理型行为智能；No.RS-2025-25442149, 物理人工智能领域领先的大规模生成式人工智能模型人才培养项目 LG AI STAR；No.RS-2025-25442824, AI Star 奖学金计划（UNIST）），以及科学与ICT部 InnoCORE 计划（25-InnoCORE-01）的支持。

不拘泥于 Transformer：在基于 Mamba 的扩散模型中舍弃节拍表示，实现舞蹈生成

补充材料

6. 预备知识

选择性状态空间模型。 状态空间模型（SSM），尤其是结构化状态空间模型（S4 [7]）和 Mamba [3, 6], 在建模序列数据的长程依赖方面展现出了卓越能力。这些模型将输入序列 $x_t \in \mathbb{R}^T$ 映射为经过变换的输出序列 $y_t \in \mathbb{R}^T$，并通过隐状态 $h_t \in \mathbb{R}^N$进行处理。SSM 可通过步长 Δ离散化，如下所示:

$$\begin{aligned} h_t &= Ah_{t-1} + Bx_t \\ y_t &= C^\top h_t, \end{aligned} \tag{9}$$

其中， $A \in \mathbb{R}^{N\times N}$、 $B \in \mathbb{R}^{N\times 1}$和 $C \in \mathbb{R}^{N\times 1}$ 分别是由状态维度 N定义的状态矩阵、输入矩阵和输出矩阵。该系统可以通过结构化卷积核 K 进行全局卷积表示（注意， x 表示此处的一般序列输入）：

$$\begin{aligned} \bar{K} &= (C^\top \bar{B}, C^\top \bar{A}\bar{B}, \ldots, C\bar{A}^{L-1}\bar{B}) \\ y &= x * \bar{K}. \end{aligned} \tag{10}$$

为摆脱线性时不变性（LTI），Mamba1 [6]引入了具有时变参数的选择性扫描，并通过关联扫描克服计算方面的挑战。Mamba2 [3] 进一步通过在概念上建立 SSM 与注意力机制之间的联系来提升效率，从而在保持与 Transformer [30]相竞争的性能的同时实现更快的计算。

扩散模型。 我们采用DDPM [8]公式，其由潜变量 $\{z_t\}_{t=1}^T$ 的前向加噪过程定义：

$$q(z_t|x) \sim \mathcal{N}(\sqrt{\bar{\alpha}_t}x, (1-\bar{\alpha}_t)\mathbf{I}), \tag{11}$$

其中 $x \sim p(x)$和 $\bar{\alpha}_t \in (0,1)$是遵循单调递减调度的常数。给定来自音乐特征 m 和节拍表示b的音乐条件 c_m，扩散模型反转前向扩散过程，以估计所有时间步 t的$\hat{x}_\theta(z_t, t, m, b) \approx x$，其中$\theta$ 表示模型参数。

我们采用扩散模型的标准重建损失，定义为:

$$\mathcal{L}_{\text{simple}} = \mathbb{E}_{x,t}\left[\|x - \hat{x}_\theta(z_t, t, m, b)\|_2^2\right]. \tag{12}$$

7. 实现细节

我们以（FineDance,AIST++）的形式报告超参数。全局序列长度为 N = (1024,128)，局部窗口大小为 n = (256,64)；音乐与动作均以30帧/秒处理。全局扩散阶段每个序列输出13个关键动作m_{key} ——5个硬提示 d_h 和8个软提示 d_s。关键动作长度为 L_{key} = (8,4)，相对于 N选取。经过编舞增强后，每个d_s 被镜像以生成16个软提示实例，并放置在节拍对齐位置。全局模型与局部模型均使用Adan [31] 以 2×10^{-4}的学习率进行优化。在推理时，我们使用DDIM [26]采样，推理步数为50。各项的损失权重为： λ_{pos} =(1, 0.636)， λ_{vel} =(2.964, 2.964)， λ_{acc} = (2.964,2.964)，λ_{foot} = (20,10.942)，以及 λ_{trans} = (0.5,0.5)。

另一方面，两个扩散模型具有相同的舞蹈解码器结构，仅SMM的细节有所不同。我们省略了全局扩散的空间 SSM 模块，因为输入序列长度与输出序列长度不同（回想一下，空间SMM块处理 $a' \in \mathbb{R}^{l\times E}$）。

8. 评估指标

为了定量评估所生成舞蹈动作的质量，我们采用了先前工作中几种常用的评估指标。我们使用的序列长度为128，与原始基线设置的150略有不同，并针对整个整合舞蹈计算了所有评估指标，因此评估指标的数值可能与先前工作中报告的数值有所不同。

动作质量。 为了评估生成动作的质量，我们计算生成动作序列与真实动作序列的动作特征之间的 Fr´echet Inception Distance（FID）。对于每个动作，我们提取运动学特征和几何特征，分别用于捕捉物理自然度和整体舞蹈编排。

物理足部接触分数。 为评估足部运动对舞蹈动作的响应在物理上的合理性，我们采用 EDGE [29] 提出的物理足部接触分数（PFC）。这一受物理启发的指标无需进行显式物理建模，即可评估足地交互是否真实。它会评估质心（COM）在水平面和垂直轴上的加速度。PFC 分数越低，表示动作在物理上越合理。

身体接触评分。 受 POPDG 启发 [17],PBC 通过分析肢体间及上半身的接触，识别不合理的相互穿透或不自然姿势，从而衡量全身动作整体的物理可行性。

动作多样性。 为评估生成动作的多样性，我们计算生成动作与真实动作的平均特征距离。遵循 Bailando [24], ，我们同时考虑运动学和几何特征，分别记为 Div_k 和 Div_g。数值越高，表示动作模式的变化越丰富。

节拍对齐分数。 为评估生成舞蹈与音乐之间的节拍一致性，我们遵循 Bailando [24]，计算每个音乐节拍与其最近动作节拍之间的平均时间距离。BAS 数值越高，表示动作与音乐节奏的同步效果越好。

用户研究（胜场数） 在用户研究中，我们招募了 20 名参与者。每位参与者评估了由 2 个数据集、10 首音乐曲目和 4 个模型（*Mam-baDance*、EDGE [29], POPDG [17], 和 Lodge [15]）生成的舞蹈视频。总计，每位参与者观看了 $2 \times 10 \times 4$ 段舞蹈视频；每组视频包含四个模型针对同一段音乐生成的序列。

参与者被要求根据以下标准选出每组中最好的视频：

- 哪个视频展现的舞蹈动作最自然?
- 哪个视频在节拍和节奏同步方面与音乐最契合?
- 哪个视频展现的动作最多样、最富动感?

为减轻位置偏差，每组中的四个视频都经过随机排序。

参考文献

[1] Emad Barsoum、John Kender 和 Zicheng Liu。HP-GAN：通过生成对抗网络进行概率三维人体动作预测。收录于 *CVPRW*，2018 年。 2[2] Xin Chen、Biao Jiang、Wen Liu、Zilong Huang、Bin Fu、Tao Chen 和 Gang Yu。通过潜在空间中的动作扩散执行命令。收录于*CVPR*，2023 年。 2[3] Tri Dao 和 Albert Gu。Transformer 即 SSM：通过结构化状态空间对偶实现泛化模型与高效算法。收录于 *ICML*，2024 年。 3，4，8，1[4] Prafulla Dhariwal 和 Alexander Nichol。扩散模型在图像合成方面胜过生成对抗网络。收录于*NeurIPS*，2021 年。 4[5] Congyi Fan、Jian Guan、Xuanjia Zhao、Dongli Xu、Youtian Lin、Tong Ye、Pengming Feng 和 Haiwei Pan。调整你的 rhythm：通过门控增强的节奏感知特征表示，生成高度对齐的舞蹈姿势。载于 ICCV，2025。3[6] Albert Gu 和 Tri Dao。Mamba：利用选择性状态空间进行线性时间序列建模，2023。arXiv 预印本 arXiv:2312.00752。3、4、8、1[7] Albert Gu、Karan Goel 和 Christopher R´e。利用结构化状态空间高效建模长序列。载于 ICLR，2022。1[8] Jonathan Ho、Ajay Jain 和 Pieter Abbeel。去噪扩散概率模型。载于 NeurIPS，2020。5、1[9] Xun Huang 和 Serge Belongie。利用自适应实例归一化实现实时任意风格迁移。载于 ICCV，2017。4[10] 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 ECCV，2024。2、4、7、8[11] Boeun Kim、Jungho Kim、Hyung Jin Chang 和 Jin Young Choi。MOST：在多样化动作内容之间进行动作风格迁移的 Transformer。载于 CVPR，2024。2[12] Jiaman Li、Yiha ng Yin、Hang Chu、Yi Zhou、Tingwu Wang、Sanja Fidler 和 Hao Li。学习使用Transformer生成多样化舞蹈动作，2020。arXiv 预印本 arXiv:2008.08171。1[13] Ruilong Li、Shan Yang、David A Ross 和 Angjoo Kanazawa。AI Choreographer：音乐条件控制的三维舞蹈生成，基于 AIST++。载于 ICCV，2021。1、2、4、6[14] Ronghui Li、Junfan Zhao、Yachao Zhang、Mingyang Su、Zeping Ren、Han Zhang、Yansong Tang 和 Xiu Li。FineDance：用于三维全身舞蹈生成的细粒度编舞数据集。载于 ICCV，2023。1、2、6、8[15] 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、用于长舞蹈生成的由粗到细扩散网络。载于 CVPR，2024。1、2、3、4、5、6、7、8[16] Matthew Loper、Naureen Mahmood、Javier Ro mero、Gerard Pons-Moll 和 Michael J Black。SMPL：一种蒙皮多人线性模型。ACM TOG，2015。4[17] Zhenye Luo、Min Ren、Xuecai Hu、Yongzhen Huang 和 Li Yao。POPDG：利用 PopDanceSet 进行流行三维舞蹈生成。载于 CVPR，2024。1、2、3、4、6、7[18] Brian McFee、Colin Raffel、Dawen Liang、Daniel P.W. Ellis、Matt McVicar、Eric Battenberg 和 Oriol Nieto。Librosa：Python 中的音频与音乐信号分析。载于第 14 届 Python科学会议论文集，2015。4[19] Alexander Quinn Nichol 和 Prafulla Dhariwal。改进的去噪扩散概率模型。载于 ICML，2021。6[20] Georgios Pavlakos、Vasileios Choutas、Nima Ghorbani、Timo Bolkart、Ahmed A. A. Osman、Dimitrios Tzionas 和 Michael J. Black。表现性人体捕捉：从单张图像中捕捉三维手部、面部和身体。载于 CVPR，2019。4[21] Ethan Perez、Florian Strub、Harm De Vries、Vincent Du-moulin 和 Aaron Courville。FiLM：利用通用条件层进行视觉推理。载于 AAAI，2018。4

[22] Mathis Petrovich、Michael J. Black 和 G¨ul Varol。基于动作条件控制的三维人体运动合成，采用Transformer变分自编码器。载于 ICCV，2021。2[23] Mathis Petrovich、Michael J. Black 和 G¨ul Varol。TEMOS：根据文本描述生成多样化的人体运动。载于 ECCV，2022。2[24] Li Siyao、Weijiang Yu、Tianpei Gu、Chunze Lin、Quan Wang、Chen Qian、Chen Change Loy 和 Ziwei Liu。Bailando：通过带有编舞记忆的演员-评论家 GPT 实现三维舞蹈生成。载于 CVPR，2022。1、2、3、4[25] Li Siyao、Weijiang Yu、Tianpei Gu、Chunze Lin、Quan Wang、Chen Qian、Chen Change Loy 和 Ziwei Liu。Bailando++：带有编舞记忆的三维舞蹈 GPT。IEEE 模式分析与机器智能汇刊，2023。1、2[26] Jiaming Song、Chenlin Meng 和 Stefano Ermon。去噪扩散隐式模型。载于 ICLR，2021。1[27] Guy Tevet、Brian Gordon、Amir Hertz、Amit H Bermano 和 Daniel Cohen-Or。MotionCLIP：将人体运动生成拓展至 CLIP 空间。载于 ECCV，2022。2[28] Guy Tevet、Jonathan Gordon、Amir Hertz 等。人体运动扩散模型。载于 ICLR，2023。2[29] Jonathan Tseng、Rodrigo Castellon 和 Karen Liu。EDGE：从音乐生成可编辑舞蹈。载于 CVPR，2023。1、2、3、4、6、7[30] Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszko-reit、Llion Jones、Aidan N Gomez、Łukasz Kaiser 和 Illia Polosukhin。Attention is all you need。NeurIPS，2017。1、8[31] Xingyu Xie、Pan Zhou、Huan Li、Zhouchen Lin 和 Shuicheng YAN。Adan：用于加速深度模型优化的自适应 Nesterov 动量算法。IEEE 模式分析与机器智能汇刊，2024。1[32] Kaixing Yang、Xulong Tang、Yuxuan Hu、Jiahao Yang、Hongyan Liu、Qinnan Zhang、Jun He 和 Zhaoxin Fan。MatchDance：用于高质量三维舞蹈合成的协作式 Mamba-Transformer 架构匹配，2025。arXiv 预印本 arXiv:2505.14222。3[33] Kaixing Yang、Xulong Tang、Ziqiao Peng、Yuxuan Hu、Jun He 和 Hongyan Liu。MegaDance：用于类型感知三维舞蹈生成的混合专家架构，2025。arXiv 预印本 arXiv:2505.17543。3[34] Wen Zhang、Xiaojie Peng、Yebin Ma 等。MotionDiffuse：通过扩散模型进行文本驱动的人体运动生成。IEEE 模式分析与机器智能汇刊，2024。2[35] Zeyu Zhang、Akide Liu、Ian Reid、Richard Hartley、Bohan Zhuang 和 Hao Tang。Motion Mamba：高效的长序列人体运动生成。载于 ECCV，2024。2、3