arXiv:2304.02419v2 [cs.CV] 1 Oct 2023

TM2D：通过音乐-文本融合驱动双模态3D舞蹈生成

Kehong Gong[1*] Dongze Lian[1*] Heng Chang[2] Chuan Guo[2]
Zihang Jiang[1] Xinxin Zuo[2] Michael Bi Mi[2] Xinchao Wang[1†]
[1] 新加坡国立大学 [2] 华为技术有限公司

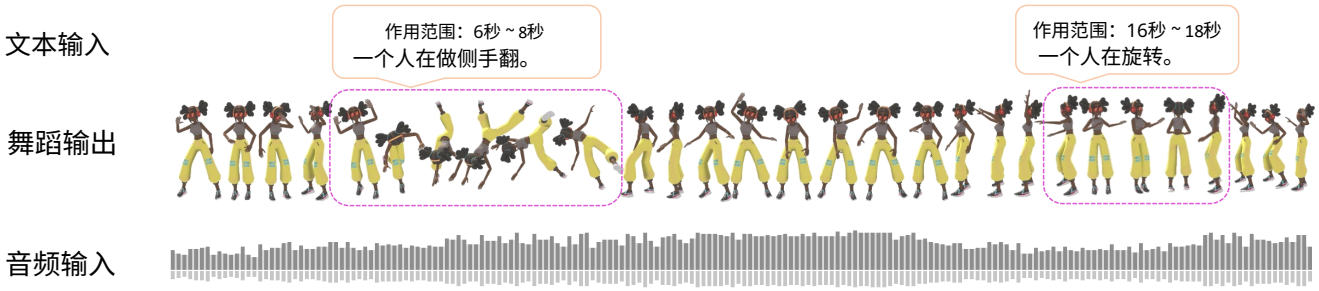


图 1：使用我们的方法生成的、以音乐和文本为条件的 3D舞蹈示例。给定具有特定起始时间和时长的音频输入和文本输入，我们的方法能够生成一系列与音乐和文本指令相契合的舞蹈动作。角色来自 Mixamo [1]。

摘要

我们提出一项新任务：生成同时融合文本和音乐两种模态的三维舞蹈动作。不同于现有使用音乐等单一模态生成舞蹈动作的方法，我们旨在利用文本提供的指导信息，生成内容更丰富的舞蹈动作。然而，缺乏同时配对音乐和文本模态的动作数据，限制了融合两者生成舞蹈动作的能力。为应对这一挑战，我们提出使用*3D*人体动作*VQ-VAE*，将两个数据集中的动作映射到由量化向量构成的潜在空间，从而有效混合来自不同分布的两个数据集的动作词元，用于训练。此外，我们提出一种跨模态*Transformer*，将文本指令融入动作生成架构，以生成三维舞蹈动作，同时不降低音乐条件舞蹈生成的性能。为更好地评估生成动作的质量，我们引入两项新指标：动作预测距离（*MPD*）和冻结分数（*FS*），分别用于衡量生成动作的连贯性和冻结比例。大量实验表明，我们的方法能够生成以文本和音乐为条件、真实且连贯的舞蹈动作，同时保持与两种单模态方法相当的性能。代码见 https://garfield-kh.github.io/TM2D/。

1. 引言

近年来，音乐条件舞蹈生成已成为备受关注的研究课题。生成与音乐同步的舞蹈动作具有广泛的应用，例如行为理解、仿真，并有助于舞者和音乐家社群的发展 [26, 9, 30, 47]。虽然音乐已被用作生成舞蹈动作的引导信号，但另一种重要的模态线索——文本（或语言）——能够提供更丰富的动作和更灵活的动作引导，并已应用于图像分类等任务 [41], 检测[16], 分割 [53], 以及文本驱动图像生成[45], ，但在舞蹈生成中尚未得到充分探索。为此，我们首次提出一项新任务：生成同时融入文本和音乐两种模态的三维舞蹈动作，使生成的人物能够根据音乐和文本表演丰富的舞蹈动作。

为这一双模态驱动的3D舞蹈生成任务设计系统流程并非易事。需要考虑两

*共同贡献：gongkehong@u.nus.edu, dongze@nus.edu.sg
†通讯作者：xinchao@nus.edu.sg

项重大挑战：i）现有数据集仅适用于音乐驱动（music2dance）[49, 5, 60, 30]或文本驱动（text2motion）的 [39, 17] 人体运动生成，尚无兼顾音乐和文本的配对式三维舞蹈生成数据集。虽然可以基于音乐和文本构建新的大规模配对式三维舞蹈数据集，但要对三维人体运动进行全面标注会耗费大量时间 [30]；ii）将文本融入音乐条件舞蹈生成需要合适的架构。然而，现有利用音乐驱动舞蹈动作生成的方法可能会产生时间冻结帧，或无法泛化至真实场景 [30, 47]。因此，将文本简单地融入现有的音乐条件架构，可能会降低这一新任务的舞蹈生成质量。

为应对第一个挑战，我们利用现有的音乐舞蹈和文本动作数据集来完成这项新任务。然而，由于这两个数据集中的动作处于完全不同的运动空间，直接混合它们的动作会导致性能下降。为克服这一问题，我们提出采用 VQ-VAE 架构，将动作投影到一致且共享的潜在空间中。具体而言，我们为训练集中的所有动作构建一个共享码本，使来自两个数据集的动作都表示为离散词元，并受到隐式约束，落入共享的潜在空间。针对第二个挑战，我们提出采用跨模态 Transformer 架构，将 music2dance 和 text2motion 都建模为序列到序列的转换任务。该架构直接将音频和文本特征转换为动作词元，并通过在潜在空间中引入融合策略，以及为两项任务共用一个动作解码器，实现由双模态驱动的生成能力。借助共享解码器，音频和文本信息可以在推理过程中高效融合。整个跨模态 Transformer 架构兼具有效性与效率，能够整合文本指令，生成连贯的 3D舞蹈动作，如图1所示。

为了更好地评估我们任务中生成的舞蹈的连贯性，我们提出了一种新的评估指标——运动预测距离（MPD），用于衡量融合文本时预测的动作与真实值之间的距离，从而更准确地评估帧间连贯性。此外，我们还引入了冻结分数（FS），用于量化舞蹈生成过程中时间冻结帧所占的比例；这是音乐条件舞蹈生成中的常见问题。为了更好地评估我们的方法在真实场景中的表现，我们还收集了一些真实场景音乐数据用于评估。我们的方法能够同时根据文本和音乐生成舞蹈，同时在单模态任务（music2dance、text2motion）上的表现与其他最先进的方法相当。

总而言之，我们的贡献如下：i) 我们提出了一项有趣的任务，即同时利用音乐和文本进行3D舞蹈生成，并为此提出了名为TM2D（文本-音乐到舞蹈）的流程。ii) 我们没有收集新的训练集，而是有效结合现有的music2dance和text2motion数据集，并采用VQ-VAE框架，将所有训练集中的动作编码到共享特征空间中。iii) 我们提出了跨模态Transformer以及双模态特征融合策略，用于编码音频和文本特征，兼具有效性和高效性。iv) 我们提出了两个新指标MPD和FS，可高效反映生成动作的质量。v) 我们成功地根据音乐和文本指令生成逼真、连贯的舞蹈，同时在单模态任务（music2dance、text2motion）上保持了相当的性能。

2. 相关工作

2.1. 音乐到舞蹈

Music2Dance 通常分为 2D 和 3D 舞蹈生成，相关研究已开展多年。近期方法从网络架构的角度对 [27] 3D 舞蹈生成进行建模。例如，Lee 等人 [26]在因果膨胀设置中探索卷积神经网络。Lee 等人 [44] 提出了一种基于变分自编码器（VAE）的舞蹈单元。Ren 等人 [12]采用循环神经网络（RNN），而 [57, 3, 25] 中则引入了图卷积网络。对于 3D 舞蹈生成，[57, 3, 25] 采用卷积神经网络，[48, 14]则将对抗学习应用于生成的舞蹈。[5, 49, 54, 61, 24]采用长短期记忆网络（LSTM），[9] 则通过匹配方法，利用学习得到的音乐/舞蹈嵌入构建运动图 [10]。近年来，Transformer 已被应用于舞蹈生成 [29, 30, 47, 28]。然而，以往的方法在生成长序列时通常会产生时间冻结帧，或者难以泛化到野外音乐，因此直接用于双模态驱动的 3D 舞蹈生成时效果并不理想。本文设计了一种跨模态 Transformer，能够有效且高效地融合文本指令，以生成 3D 舞蹈。

2.2. 文本到动作

除了音乐驱动的运动生成外，文本也被用作生成运动的指令。Text2motion可分为基于动作标签的运动生成和基于语言描述的运动生成。基于动作标签的方法 [19, 37] 根据动作标签生成运动。Action2Motion[19] 实现GRU-VAE来迭代生成下一个

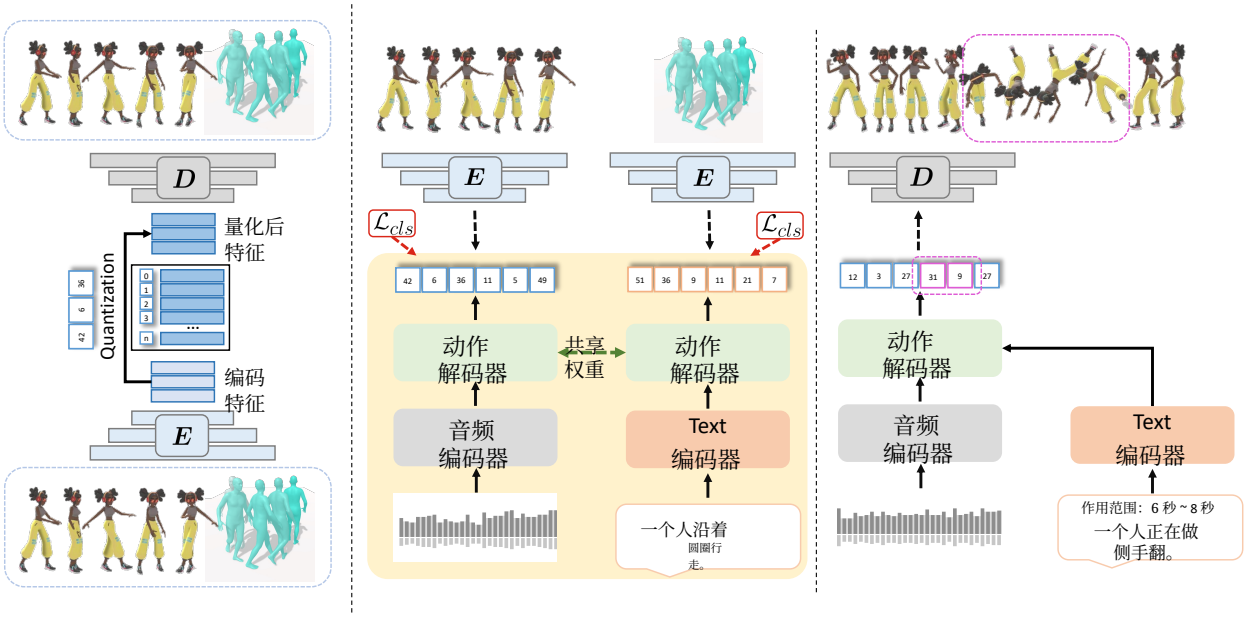


图2：我们提出的音乐与文本条件3D舞蹈生成流程。从左至右分为三个阶段：3D人体动作VQ-VAE、跨模态Transformer训练阶段，以及我们流程的推理阶段。第一阶段使用来自music2dance和text2motion数据集的人体运动训练VQ-VAE，随后用它对所有动作进行词元化。第二阶段采用双路径跨模态Transformer执行序列到序列翻译任务（即，音频到动作词元、文本到动作词元），并共享一个动作解码器。第三阶段给定音频和文本输入后，音频和文本编码器首先提取相应特征，然后在动作解码器中进行融合（后期融合），从而生成同时受音乐和文本条件控制的舞蹈。

基于动作标签和先前帧的帧。AC- TOR [37] 实现Transformer-VAE来一次性编码和解码整个姿态序列。由于基于动作标签的方法受限于一小套动作标签，因此提出了基于语言描述的方法以实现更灵活的运动生成。[40, 32, 2, 17, 18] 将文本到运动任务表述为机器翻译问题，其他 [55, 4, 13, 38] 学习文本和运动的联合嵌入空间。其中， [40, 32, 2, 55, 4, 13] 使用RNN编码器-解码器来学习文本和运动之间的映射。 [18, 17] 使用Transformer编码器和RNN解码器，而TM2T [18]进一步引入运动到文本作为文本到运动的逆向对齐。TEMOS [38] 和TEACH [6] 对编码器和解码器都使用Transformer进行一次性生成。与这些text2motion方法不同，我们专注于如何将两种模态（即，音乐和文本）结合在一起进行3D舞蹈生成。

3. 方法

3.1. 系统流程

整体流程如图2所示，由三个阶段组成。第一阶段采用三维人体动作VQ-VAE 将 music2dance 和 text2motion 数据集中的动作编码到由多个向量构成的共享码本中，使每个舞蹈动作都能表示为一个离散动作标记。随后，将音乐、文本和动作标记的组合输入跨模态 Transformer 进行训练，使其有效学习如何根据先前的动作标记以及音乐和文本预测一系列动作标记。在第三阶段，生成一个随机的起始动作标记，并将其与给定的音乐和文本一同输入跨模态 Transformer，以预测一系列动作标记。随后，第一阶段预训练的 3D人体动作VQ-VAE 对这些标记进行解码，生成 3D舞蹈。

3.2. 3D人体动作VQ-VAE

由于不存在同时以音乐和文本为条件的配对3D舞蹈动作，我们尝试利用现有的music2dance和text2motion数据集来解决这一任务。我们没有直接混合两个数据集中的动作进行训练，而是采用VQ-VAE将动作映射到一致且共享的潜在空间中，并用码本表示，如图2所示。具体而言，给定一个三维人体动作 $M \in \mathbb{R}^{T \times d_m}$，其中 T 是时间长度

而 d_m 是人体运动的维度，VQ-VAE的编码器由若干个一维卷积层组成，可将 M映射为潜在向量 $z \in \mathbb{R}^{T' \times d}$，其中 $T' = \frac{T}{t}$和t 是下采样的时间间隔， d 是潜在向量的维度。可学习的码本 $e \in \mathbb{R}^{K \times d}$描述了整个数据集的所有潜变量特征，其中 K 和 d 分别是码本的长度和维度。量化后的潜在向量 $z_q \in \mathbb{R}^{T' \times d}$将记录码本 e 中与之最近的向量，具体如下:

$$z_{q,i} = \arg\min_{e_j \in e} \|z_i - e_j\| \in \mathbb{R}^d, \tag{1}$$

而动作词元 $t_m \in \mathbb{R}^{T' \times 1}$ 存储最接近向量的索引

$$t_{m,i} = \arg\min_{j} \|z_i - e_j\| \in \mathbb{R}^1. \tag{2}$$

量化后的 z_q 将通过堆叠的卷积层解码，以重建人体运动 M，而 t_m将用于跨模态 Transformer 的训练阶段。

对于 3D人体动作VQ-VAE 的训练，我们遵循 [51] 中的策略，总损失包含人体运动回归的重建损失、用于字典学习的码本损失，以及用于稳定训练过程的承诺损失:

$$\mathcal{L}_{vq} = \|\hat{M} - M\|_1 + \|\text{sg}[e] - e_q\|_2^2 + \beta\|e - \text{sg}[e_q]\|_2^2, \tag{3}$$

其中，sg[·] 表示“停止梯度”， β 是用于调整承诺损失权重的因子项。反向传播时，还采用了直通估计器，将梯度从解码器传递到编码器。

正如我们将在第 4节所示，由 3D人体动作VQ-VAE 对两个数据集进行编码后得到的动作词元几乎落在共享的潜在空间中，这表明可以使用分别以音乐和文本为条件的人体运动数据集来完成音乐和文本条件下的 3D舞蹈生成任务。

3.3. 跨模态Transformer

跨模态Transformer包含一个音频编码器、一个文本编码器和一个动作解码器。由于我们的任务使用两个独立的数据集，跨模态Transformer被分为两个分支，一个用于music2dance，另一个用于text2motion。它以音频特征、文本特征以及由3D人体动作VQ-VAE编码的动作标记 t_m 作为输入，并执行序列到序列的转换任务，以生成未来的动作标记。

注意力。 注意力最初被引入 Transformer [52] 用于自然语言处理，随后应用于许多领域 [31, 8, 21, 42, 11, 56, 33, 58, 43] 一个 L层 Transformer 通常由 L Transformer模块组成，其中包含多头自注意力（MSA）和前馈网络（FFN）。给定输入 $x \in \mathbb{R}^{N \times c}$，其中$N$ 是序列长度， c是输入维度，Transformer模块首先通过线性投影将其映射为键 $K \in \mathbb{R}^{N \times c}$、查询 $Q \in \mathbb{R}^{N \times c}$、值 $V \in \mathbb{R}^{N \times c}$，然后执行自注意力运算

$$\text{Attention}(Q, K, V) = \text{Softmax}(\frac{QK^T}{\sqrt{c}})V. \tag{4}$$

输出后接归一化层和前馈网络。**音频编码器**。给定一个音乐序列，音频编码器首先按照 Bailando 提取原始音频特征 [47], ，其中使用了一个公开的音频处理工具箱 Librosa [23], 来获取梅尔频率倒谱系数（*MFCC*）、*MFCC*差分、恒定*Q*色度图、节奏图和起始强度。随后，通过一个嵌入层和若干包含自注意力操作的Transformer模块（例如，公式（4））来生成处理后的音频特征 $f_a \in \mathbb{R}^{T' \times d}$。

文本编码器。 给定一条文本指令，我们首先使用 GloVe 提取文本词元 [36], ，然后通过嵌入层和若干 Transformer 模块，获得处理后的文本特征 $f_t \in \mathbb{R}^{n \times d}$，其中 n 表示文本特征的长度。

动作解码器。 动作解码器以 3D人体动作VQ-VAE 编码的动作词元为输入，并输出未来的动作词元，其思路与序列到序列翻译任务相同。由于时间 t 之后的信息在时刻 t尚未知晓，因此首先采用掩码多头自注意力来交互处理不同时刻的动作。掩码的形状为简单的上三角矩阵，其定义见式（4）。对于时刻 t，只有时刻 t 之前的动作才能执行自注意力操作。动作解码器还通过音频编码器和文本编码器，将提取的音频特征 f_a 和文本特征 f_t 分别映射为 K 和 Q，以便与动作词元执行跨模态注意力。为实现音乐-文本条件舞蹈生成（即特征融合，我们采用共享参数），并使用共享参数的动作解码器以提高模型效率，如图2所示。

3.4. 音乐文本融合的详细信息

详细架构见图 3。给定音频特征和文本特征后，动作解码器首先通过自注意力层处理过去的动作词元，随后依次执行加法、归一化、分别与音频和文本特征进行交叉注意力、加法、归一化以及前馈层。重复此过程 $L(L = 6)$ 次，以构建典型的 L 层 Transformer。然后，我们采用后期融合策略，对特定时刻的音频和文本特征进行加权和（即，作用范围）。权重曲线如图3所示，

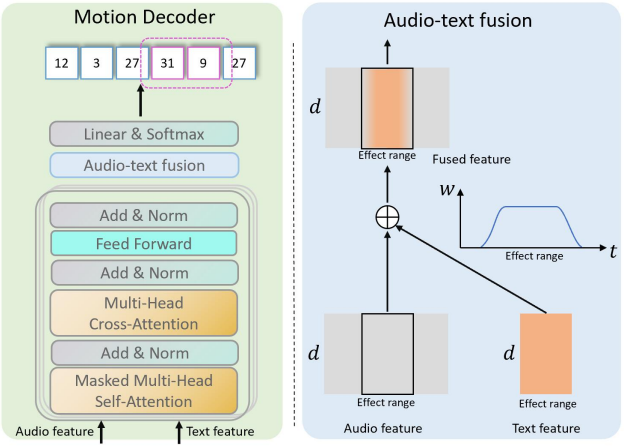


图3：音乐-文本特征融合的细节：给定编码器输出的音频和文本特征，解码器在前几层中分别处理它们。随后，在作用范围内通过加权和与权重曲线应用音频-文本融合层。最后，应用线性投影层和 softmax 操作，预测以音乐-文本为条件的动作词元。

在作用范围开始时（作用范围时长的 20% 处），我们通过半余弦曲线逐渐增大文本特征的权重，直至峰值 0.8；在结束时（作用范围时长的 20% 处），则通过半余弦曲线逐渐减小该权重。音频特征的权重 W_{audio} 为 1 - W_{text}，以确保特征保持相同的尺度。利用融合后的特征，通过线性投影层和 softmax 运算预测由音乐和文本共同条件化的动作词元，再由人体运动 VQ-VAE 的解码器对其进行解码，获得由音乐和文本共同条件化的 3D舞蹈序列。

3.5. 训练与推理

训练。 为了训练3D人体动作VQ-VAE，我们使用滑动窗口从原始动作序列中裁剪64帧作为输入，即， $T = 64$，用于music2dance 和 text2motion 两个数据集。我们使用三层编码器和解码器，使下采样的时间间隔 t 为8。我们从两个数据集中随机采样动作，并采用 Adam 优化器，批量大小为128，学习率为 $1e^{-4}$，以训练我们的3D人体动作VQ-VAE。

为训练跨模态Transformer，我们将两个分支拆分为两个流，在两个数据集上分别训练music2dance和text2motion任务。对于music2dance任务，我们使用与动作词元相同的频率对音乐向量进行采样，使音乐特征与动作词元具有相同的时间长度。我们使用动作解码器进行序列到序列的转换，获得以音乐输入为条件的动作词元预测。对于text2motion任务，由于文本与动作词元的长度不同，我们将文本的最大长度设为84，并采用填充策略进行text2motion转换。我们为音频编码器、文本编码器和动作解码器各使用6层自注意力层，隐藏维度为512，注意力头数为8。music2dance和text2motion任务均采用交叉熵损失

$$\mathcal{L}_{cls} = -\frac{1}{m}\sum_{i=0}^{m-1}\sum_{j=0}^{C-1} y_{ij}\log \hat{y_{ij}}, \tag{5}$$

其中，$\hat{y}$ 是动作词元的预测值， y 是真实值。 m 是动作词元的长度， C 是动作词元的类别数，即3D人体动作VQ-VAE的码本长度。两个任务同时进行优化，批量大小为64，学习率为 $1e^{-4}$。

推理。我们的目标是生成融合音乐与文本的3D舞蹈。因此，在推理阶段，我们首先将音频编码器和文本编码器提取的音乐特征与文本特征进行加权和融合。具体而言，给定时间长度为 T' 的音频特征和长度为 n的文本特征，我们首先将它们输入动作解码器，以预测未来的动作词元。随后，我们采用后期融合策略，在融合文本对应的时长内，对根据音频特征和文本特征生成的动作词元进行加权和融合，再经过线性投影层和softmax操作，得到融合后的动作词元。最后，融合后的动作词元经过3D人体动作VQ-VAE模型的解码器，解码为3D舞蹈序列。

4. 实验

4.1. 实验设置

数据集。 由于我们的任务没有同时以音乐和文本为条件的配对3D舞蹈动作，因此我们结合了现有的music2dance和text2motion数据集。对于music2dance数据集，我们采用 AIST++ 数据集 [30]。对于text2motion数据集，我们采用 HumanML3D [17]数据集。此外，我们还收集了一个新的真实场景音乐数据集，用于评估我们方法的泛化能力，其中包含来自YouTube的82段野外音乐片段。该数据集总时长为53分钟（规模是AIST++ 测试集的8倍），涵盖了多种音乐风格和内容，其分布确实位于AIST++之外（详见补充材料）。在我们的数据集上进行评估，能更好地反映模型在真实场景中的泛化能力。

基线方法。 由于我们是首个提出这种由双模态驱动的3D舞蹈生成任务的研究，目前尚无现有方法可供比较。我们实现了一种传统的动作编辑算法——球面线性插值 [46], ，该算法将舞蹈最后一帧与文本生成动作的第一帧对齐，并

使用四元数球面插值填补两者之间的过渡部分。为了验证我们的方法在music2dance任务上的有效性，我们将其与以往的music2dance研究进行比较 [29, 60, 22, 30, 47]。为了验证我们的方法在text2motion任务上的有效性，我们将其与以往的text2motion研究进行比较 [32, 4, 7, 13, 50, 26, 18]。

评估指标。我们采用 FACT [30] 和 Bailando [47] 建议的相同评估设置来评估舞蹈生成质量，包括弗雷歇 Inception 距离（FID）[20], 、多样性和节拍对齐分数。对于 text2motion 质量，我们采用 TM2T [18], 建议的相同评估设置，包括 R精度、多模态距离、FID、多样性和多模态性。

此外，我们还提出了两种新的评估指标：动作预测距离（MPD）和冻结分数。前者反映融合文本时帧之间的连贯性，后者反映生成的舞蹈中时间冻结帧所占的比例。对于冻结分数，我们引入冻结帧百分比（PFF）和 AUC$_f$，其中 PFF 通过以下两项标准测量冻结帧所占的比例：1）最大关节速度低于阈值（0.015m/s）；2）持续时间超过一定时长（3s）。AUC$_f$ 定义为 PFF 在 0 至 0.03 的阈值范围内曲线下的面积。至于 MPD，其定义为

$$\text{MPD} = \min_i \| f_i(M_{t_0 \to t_1}) - M_{t_1 \to t_2} \|_2, \qquad (6)$$

其中，M 是舞蹈动作，f 是动作预测模型，f_i（$M_{t_0 \to t_1}$）是第 i 个预测的可能未来动作，t_0、t_1和 t_2 是时间戳。这意味着模型会根据从 t_0 到 t_1的动作，预测从 t_1 到 t_2 的各种潜在动作。如果该距离较小（即，生成的舞蹈$M_{t_1 \to t_2}$ 处于可能的未来范围内），那么在整合文本时，生成的舞蹈动作 $M_{t_1 \to t_2}$ 在时间上会更加连贯。在此评估指标中，我们采用 DLow [59] 动作预测模型，因为该模型旨在预测多样化的潜在未来动作。

感知评估。除上述指标测量外，我们还在亚马逊土耳其机器人（AMT）上进行了大量用户研究，从感知层面评估我们生成的 3D 舞蹈结果的视觉效果。具体而言，对于从我们的方法和其他方法中采样、配有相同音乐片段的每一对舞蹈动作，我们邀请 AMT 上 3 名不同的用户表达他们对音乐舞蹈对齐、动作真实感和机动性的偏好。我们还设定了参与用户的门槛：只有获得硕士级认可、完成任务超过 1,000 项且认可率超过 98% 的用户才会被纳入考虑。总共有 55 名用户参与了我们的用户研究，他们来自不同地区，年龄、种族和性别各异。用户研究的结果更能体现生成的舞蹈在实践中的效果。

4.2. 音乐文本条件舞蹈生成评估

我们在 AIST++ 测试集和我们自己的数据集上验证了音乐文本条件舞蹈生成的结果，如表2所示。此处我们主要使用 MPD 指标评估音乐与文本衔接时生成舞蹈的连贯性，即文本开始生效的时间点。我们以纯 music2dance 生成的连贯性作为基线，命名为 a2d，以反映有无文本指令的影响。我们使用过去 25 帧动作来预测未来 30 帧，并分别计算从未来帧（ft）= 10 到 ft = 30的 MPD。我们的方法在两个数据集上均持续优于球面线性插值 [46] 和 a2d 基线，且相较于表2所示的真实值，取得了相近的结果，这验证了我们生成动作的自然性。更重要的是，用户研究实验表明，相较于仅以音乐为条件生成的舞蹈，我们的方法生成的 3D舞蹈更为逼真；即使与表2所示的真实值相比，胜率也达到 70.0%。用户研究结果表明，参与者对音乐文本条件舞蹈生成的评分较高，凸显了在评估舞蹈质量时考虑观众偏好的重要性。

4.3. 音乐条件下的舞蹈生成评估

为验证我们架构的有效性，我们将我们的方法与音乐条件下3D舞蹈生成领域的先进方法进行了定量比较。AIST++测试集上的结果如表1所示。可以发现，在运动多样性（Div$_k$ 和 Div$_g$）方面，我们的方法优于以往方法；但在不使用真实值中的种子运动的情况下，与 FACT [30]和 Bailando [47]相比，我们的方法在运动质量（FID$_g$）和节拍对齐分数方面表现较差。

现有的评估指标不足以反映生成的舞蹈在实际应用中的质量，因此我们提出了 PFF 和 AUC$_f$，用于评估冻结帧的比例。值得注意的是，在 PFF 和 AUC$_f$方面，我们的架构优于 FACT [30] 和 Bailando [47]，这表明我们的方法几乎不会生成冻结帧。同时，Li 等. [29] 和 DanceNet [60]在 PFF 和 AUC$_f$上的冻结现象也几乎为零。这是因为 Li 等. [29]生成的舞蹈抖动严重，导致速度变化极大；[30, 47], 也报告了这一点，并指出这会造成不冻结的问题。至于 DanceNet [60], ，它以重复的动作模式生成舞蹈，几乎没有冻结现象，但多样性较低。此外，用户研究的结果表明，与其他方法相比，生成的 3D 舞蹈在视觉上更加逼真

Data	方法	运动质量		运动多样性		冻结		节拍对齐分数 ↑	用户研究
		FID_k ↓	$FID_g^\dagger$ ↓	Div_k ↑	$Div_g^\dagger$ ↑	PFF↓	AUC_f ↓		我们的方法胜出
AIST ++	真实值	17.10	10.60	8.19	7.45	0.00	0.00	0.2374	41.9%
	Li 等。 [29]	86.43	43.46	6.85*	3.32	**0.00**	**0.00**	0.1607	98.3%
	DanceNet [60]	69.18	25.49	2.86	2.85	**0.00**	0.98	0.1430	87.2%
	DanceRevolution [22]	73.42	25.92	3.52	4.87	11.01	12.22	0.1950	70.5%
	FACT [30]	35.35	22.11	5.94	6.18	25.29	21.59	0.2209	83.3%
	Bailando [47]	28.16	9.62	7.83	6.34	14.91	13.25	**0.2332**	65.0%
	我们的方法（仅使用舞蹈数据）	23.94	**9.53**	7.69	4.53	**0.00**	**0.00**	0.2127	-
	ours	**19.01**	20.09	**9.45**	**6.36**	**0.00**	**0.00**	0.2049	–
Wild 音频	FACT [30]	70.36	20.23	7.33	**6.34**	32.64	27.21	0.2211	89.6%
	Bailando [47]	50.56	22.55	3.80	6.04	17.55	14.14	0.2166	70.8%
	我们的方法（仅使用舞蹈数据）	43.85	**13.08**	**8.52**	4.76	0.78	0.76	0.1998	–
	ours	**27.65**	20.34	7.88	5.27	**0.12**	**0.11**	**0.2290**	–

表 1：音乐条件舞蹈生成：在 AIST++ 和 Wild Audio测试集上的定量结果。最佳值和次佳值分别以粗体和下划线标示。在对比方法中，“Li 等”、DanceNet 和 FACT 复用了 AIST++基准测试的相同结果 [29], , 而 DanceRevolution [22]沿用了 Bailando [47]的结果。 † FID_k 和 DIV_k取自 [30] , 而 FID_g 和 DIV_g 取自 [47]。

Data	方法	动作预测距离			用户研究
		英尺=10	英尺=20	英尺=30	我们的方法胜出
AIST ++	GT	0.048	0.072	0.088	70.0%
	a2d	0.052	0.084	0.108	66.6%
	球面线性插值 [46]	0.088	0.122	0.135	73.3%
	ours	**0.049**	**0.080**	**0.102**	-
Wild 音频	a2d	0.052	0.088	0.113	60.0%
	球面线性插值 [46]	0.104	0.148	0.171	70.0%
	ours	**0.048**	**0.079**	**0.107**	-

表 2：音乐文本条件舞蹈生成：AIST++和 Wild Audio测试集上的定量结果。最佳值和次佳值分别以粗体和下划线标示。

。即使与真实值相比，我们生成的舞蹈平均有 41.9% 被评为更好。我们还报告了仅使用舞蹈数据训练的结果，其表现与之相当。

除了在 AIST 测试集上进行评估外++，我们还在表1中展示了在野外数据集上的实验结果。我们的方法在几乎所有指标上都优于 FACT [30] 和 Bailando [47], FID_g Div_g除外。可以观察到，FACT [30] 和 Bailando [47]的 FID_k均大幅下降，而我们的方法变化较小。这是因为 FACT [30]和 Bailando [47]需要种子动作，但真实场景中没有真实值。使用随机采样的种子动作或词元时，它们的方法无法很好地适用于真实场景。

在真实场景中表现良好。还可以注意到，我们的方法在野外数据集上存在冻结现象，但在 AIST 测试集上的冻结帧比例为零++，这是由 AIST++与我们的数据集分布不同所致。这也表明我们的数据集更具挑战性。我们还报告了仅使用舞蹈数据训练的结果，展示了我们的混合训练策略在真实场景中的优势。

4.4. 文本条件动作生成的评估。

我们在两种不同场景下评估 text2motion 方法：仅使用文本进行推理（t2m），以及融合文本和音乐特征进行推理。表 3 展示了结果。在仅使用文本的设置下，我们的方法取得了与 TM2T 基线（T） [18], 相当的表现，这表明混合数据/任务训练不会影响 text2motion 生成的质量。随后，我们采用后期融合方法，随机抽取一段音乐片段（来自整个 AIST++测试集）和一段文本（来自整个 t2m 测试集），并按照与 TM2T 相同的协议评估生成的舞蹈片段 [18]。如表3 所示，当后期融合率为 0.8 时，文本与舞蹈之间的一致性仍然可以接受。有关后期融合率与 text2motion 结果之间关系的更多详情，请参阅补充材料。

4.5. 定性结果

我们还在图中展示了我们的方法在音乐-文本条件下进行 3D舞蹈生成的定性结果：4、图5 和图6。具体而言，图4展示了生成的舞蹈——

方法	R精度↑		FID↓	多模态距离↓	多样性→	多模态性↑
真实动作	0.511	0.002	2.974	9.503	-	
Seq2Seq [32]	0.180	11.75	5.529	6.223	-	
Language2Pose [4]	0.246	11.02	5.296	7.676	-	
Text2Gesture [7]	0.165	5.012	6.030	6.409	-	
Hier [13]	0.301	6.532	5.012	8.332	-	
MoCoGAN [50]	0.037	94.41	9.643	0.462	0.019	
Dance2Music [26]	0.033	66.98	8.116	0.725	0.043	
TM2T 基线(T) [18]	0.351	1.669	4.046	9.632	4.352	
TM2T [18]	**0.424**	1.501	**3.467**	8.589	2.424	
TM2D(t2m)	0.319	**1.021**	4.098	**9.513**	4.139	
TM2D(LFR 0.8)	0.300	1.105	4.307	8.887	**4.443**	

表 3：在 Hu- manML3D 数据集上对 text2motion 的评估，结果为 20 次运行的平均值。

在相同音乐序列下，使用不同文本指令生成的3D舞蹈。图5 展示了在相同文本指令下、采用不同起始时间生成的3D舞蹈。图6 展示了在相同文本指令下、采用不同持续时间生成的3D舞蹈。可以看出，在这三种情况下，我们的方法都能根据文本指令生成合理的视觉效果，这证明了我们的方法具有更高的灵活性。

4.6. 混合数据的影响

我们采用两个数据集（AIST++ 和 HumanML3D）训练 3D人体动作VQ-VAE，使动作词元落入相同的潜在空间。为通过实验证明这一点，我们进行了实验，统计共享动作词元的数量。对于使用 AIST++和 HumanML3D 训练的 3D人体动作 VQ-VAE，AIST++ 和 HumanML3D 中分别有 855 个和 912 个码本向量用于构成动作词元。码本共包含 1024 个向量。其中，846 个向量（占 AIST++ 的 98.9% 和 HumanML3D 的 92.8%）可共享用于生成动作词元，这表明使用彼此独立的音乐条件和文本条件动作数据集来完成音乐与文本条件下的 3D舞蹈生成任务是可行的。我们进一步验证了 VQ-VAE 对动作潜在空间的表示，并在图7中对音频到舞蹈动作和文本到动作的原始动作分布（VQ-VAE 处理前）及其经 VQ-VAE 编码后的特征分布进行了 t-SNE 可视化。两个动作数据集成功实现了混合，为融合双模态舞蹈生成提供了可能。更多详情请参阅补充材料。

4.7. 模型效率

我们还在表4中展示了架构的效率。在推理阶段，我们比较了完整模型参数量以及生成 1 秒音乐的推理时间。我们的架

该架构仅需约一半的模型参数，并且

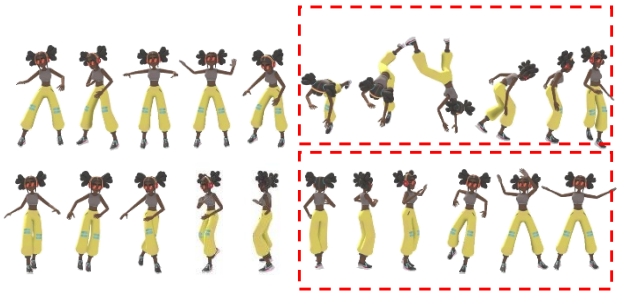


图 4：**相同的音乐，不同的动作**。这两段生成的舞蹈使用相同的音乐和作用范围（从 6 秒到 8 秒），但文本指令不同：“一个人在做侧手翻”（上），“一个人张开双臂旋转”（下）。

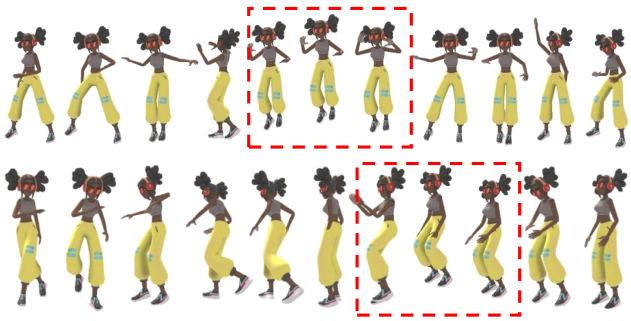


图 5：**相同的动作，不同的开始时间点**。这两段生成的舞蹈使用相同的音乐和文本指令（“一个人上下跳跃”），但作用范围不同：从 6 秒到 8 秒（上），从 7 秒到 9 秒（下）。

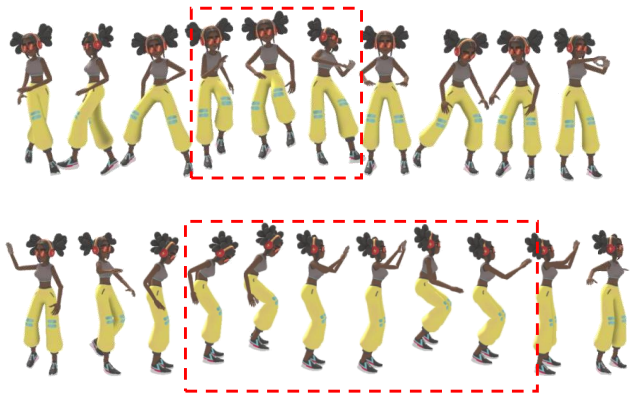
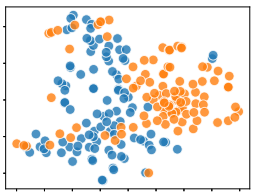
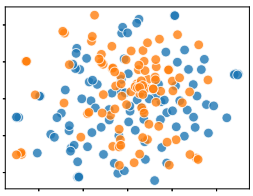


图 6：**相同的动作，不同的持续时间**。这两段生成的舞蹈使用相同的音乐和文本指令（“一个人一直在跳跃”），且动作开始时间点相同（即，6 秒），但作用持续时间不同：2 秒（上），3 秒（下，即，跳了两次）。

与 Bailando 相比的推理时间 [47], 哪些属性



(a) VQ-VAE 之前



(b) VQ-VAE 之后

图 7：VQ-VAE 前后动作的 t-SNE 可视化（橙色：AIST++ 中的动作，蓝色：HumanML3D 中的动作）。

方法	参数	推理时间
FACT [30]	120M	8.300秒/(1秒音乐)
Bailando [47]	152M	0.236秒/(1秒音乐)
ours	72M	0.143秒/(1秒音乐)

表 4：参数和推理时间的比较。

因此，我们无需为上半身和下半身分别编码特征。与 FACT [30], 相比，我们的架构显著缩短了推理时间。主要原因是 FACT 为音乐和舞蹈动作使用了两个 Transformer，并使用一个跨模态Transformer；而我们将其建模为标准的序列到序列翻译任务，只使用一个 Transformer，这体现了我们架构的优势。

5. 结论

本文提出了一项新任务，将音乐和文本指令同时结合，用于3D舞蹈生成。由于缺乏同时以音乐和文本为条件的配对3D舞蹈动作，我们采用了两个现有数据集，即即., music2dance和text2motion，来完成这项任务，并使用3D人体动作VQ-VAE将两个数据集中的动作投影到共享的潜在空间中，从而有效混合分布不同的两个数据集进行训练。此外，我们还提出了一种跨模态Transformer架构，用于生成3D舞蹈，同时不降低音乐条件舞蹈生成的性能。我们提出了两个新的评估指标MPD和FS，用于反映生成动作的质量。大量实验表明，我们的方法能够生成与音乐和文本均相匹配、真实且连贯的舞蹈动作，同时在两种单模态上的性能也具有可比性。

致谢。 本项目获得新加坡教育部学术研究基金第二级（资助编号：MOE-T2EP20122-0006）的支持，以及新加坡国家研究基金会的支持（“CogniVision——面向分布式实时视觉的能量自给型常开认知与注意摄像头，功耗为毫瓦级”项目，资助编号：NRF-CRP20-2017-0003）——www.green-ic.org/CogniVision。

参考文献

[1] mixamo. https://www.mixamo.com/. 访问日期：2023-03-07。 1[2] Hyemin Ahn、Timothy Ha、Yunho Choi、Hwiyeon Yoo 和 Songhwai Oh。Text2action：从语言到动作的生成对抗合成。发表于 2018 年 IEEE 国际机器人与自动化会议（ICRA），第 5915–5920 页。IEEE，2018。 3[3] Hyemin Ahn、Jaehun Kim、Kihyun Kim 和 Songhwai Oh。基于音乐合成三维舞蹈动作的生成式自回归网络。IEEE 机器人与自动化快报，5(2)：3501–3508，2020。 2[4] Chaitanya Ahuja 和 Louis-Philippe Morency。Language2pose：以自然语言为基础的姿态预测。发表于 2019 年国际三维视觉会议（3DV），第 719–728 页。IEEE，2019。 3, 6, 8[5]Omid Alemi、Jules Franc¸oise 和 Philippe Pasquier。Groovenet：使用人工神经网络实时生成音乐驱动的舞蹈动作。发表于第 23 届 ACM SIGKDD 知识发现与数据挖掘会议的创造力机器学习研讨会，第 8 卷，第 26 页，2017。 2[6] Nikos Athanasiou、Mathis Petrovich、Michael J. Black 和 G¨ul Varol。TEACH：面向三维人体的时序动作组合。发表于国际三维视觉会议（3DV），2022。 3[7] Uttaran Bhattacharya、Nicholas Rewkowski、Abhishek Banerjee、Pooja Guhan、Aniket Bera 和 Dinesh Manocha。Text2Gestures：一种用于为虚拟智能体生成富有情感的身体手势的 Transformer 网络。发表于 IEEE 虚拟现实与三维用户界面会议（VR），第 1–10 页。IEEE，2021。 6, 8[8] Hanting Chen、Yunhe Wang、Jianyuan Guo 和 Dacheng Tao。VanillaNet：深度学习中的极简主义力量。arXiv 预印本 arXiv:2305.12972，2023。 4[9] Kang Chen、Zhipeng Tan、Jin Lei、Song-Hai Zhang、Yuan-Chen Guo、Weidong Zhang 和 Shi-Min Hu。ChoreoMaster：面向编舞的音乐驱动舞蹈合成。ACM 图形学汇刊（TOG），40(4)：1–13，2021。 1, 2[10] Rukun Fan、Songhua Xu 和 Weidong Geng。基于示例的自动音乐驱动传统舞蹈动作合成。IEEE 可视化与计算机图形学汇刊，18(3)：501–515，2011。 2[11] Gongfan Fang、Xinyin Ma 和 Xinchao Wang。扩散模型的结构化剪枝。arXiv 预印本 arXiv:2305.10924，2023。 4[12] Joao P Ferreira、Thiago M Coutinho、Thiago L Gomes、Jos´e F Neto 等。学习舞蹈：一种从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》，94：11–21，2021。 2

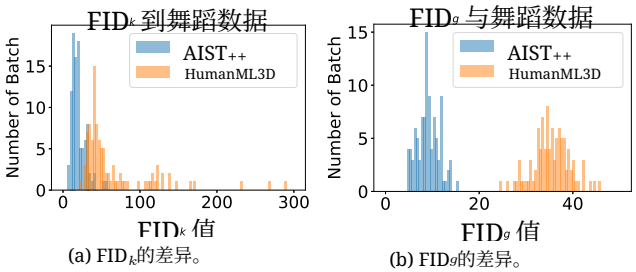
[13] Anindita Ghosh、Noshaba Cheema、Cennet Oguz、Christian Theobalt 和 Philipp Slusallek。从文本描述合成组合动画。发表于 IEEE/CVF国际计算机视觉会议论文集，第 1396–1406 页，2021。 3, 6, 8[14] Shiry Ginosar 、 Amir Bar、Gefen Kohavi、Caroline Chan、Andrew Owens 和 Jitendra Malik。学习对话手势的个体风格。发表于 IEEE/CVF计算机视觉与模式识别会议论文集，第 3497 3506, 2019 2 – 页。Deepak Gopinath 和 [15] Jungdam Won。fairmotion：用于加载、处理和可视化动作捕捉数据的工具。GitHub，2020。 12[16] Xiuye Gu、Tsung-Yi Lin、Weicheng Kuo 和 Yin Cui。通过视觉与语言知识蒸馏实现开放词汇目标检测。发表于国际学习表征会议，2022。 1[17] Chuan Guo、Shihao Zou、Xinxin Zuo、Sen Wang、Wei Ji、Xingyu Li 和 Li Cheng。从文本生成多样且自然的三维人体动作。发表于 IEEE/CVF计算机视觉与模式识别会议论文集，第 5152 5161, 2022 2, 3, 5, 12 – 页。Chuan Guo、[18] Xinxin Zuo、Sen Wang 和 Li Cheng。TM2T：三维人体动作与文本相互生成的随机化令牌建模。发表于欧洲计算机视觉会议，第 580–597 页，2022。 3, 6, 7, 8, 12[19] Chuan Guo、Xinxin Zuo、Sen Wang、Shihao Zou、Qingyao Sun、Annan Deng、Minglun Gong 和 Li Cheng。Action2motion：条件式三维人体动作生成。发表于第 28 届 ACM 国际多媒体会议论文集，第 2021–2029 页，2020。 2[20]Martin Heusel、Hubert Ramsauer、Thomas Unterthiner、Bernhard Nessler 和 Sepp Hochreiter。通过双时间尺度更新规则训练的 GAN 收敛至局部纳什均衡。发表于神经信息处理系统进展，第 30 卷，2017。 6, 12[21] Duc N.M Hoang、Shiwei Liu、Radu Marculescu 和 Zhangyang Wang。通过 Ramanujan 图的视角重新审视初始化时的剪枝。发表于国际学习表征会议，2023。 4[22] Ruozi Huang、Huang Hu、Wei Wu、Kei Sawada、Mi Zhang 和 Daxin Jiang。Dance Revolution：通过课程学习实现音乐驱动的长期舞蹈生成。发表于国际学习表征会议，2021。 6, 7[23] Yanghua Jin、Jiakai Zhang、Minjun Li、Yingtao Tian、Huachun Zhu 和 Zhihao Fang。利用生成对抗网络实现动漫角色的自动创建。arXiv 预印本 arXiv:1708.05509，2017。 4[24] Hsuan-Kai Kao 和 Li Su。时间引导的音乐到身体动作生成。发表于第 28 届 ACM国际多媒体会议论文集，第 147 155 页，2020 2 – 。[25] Kosmas Kritsis、Aggelos Gkiokas、Aggelos Pikrakis 和 Vassilis Katsouros。DanceConv：使用卷积网络生成舞蹈动作。IEEE Access，10：44982–45000，2022。 2[26] 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。《神经信息处理系统进展》，第32卷，2019年。1、2、6、8[27] Juheon Lee、Seohyun Kim 和 Kyogu Lee。Listen to dance：使用自回归编码器-解码器网络进行音乐驱动的编舞生成。arXiv 预印本 arXiv:1811 00818，2018年 2 . . [28] Buyu Li、Yongchi Zhao、Shi Zhelun 和 Lu Sheng。Dance-former：使用参数化动作 Transformer 进行音乐条件控制的3D舞蹈生成。载于《AAAI人工智能会议论文集》，第36卷，第1272–1279页，2022年。2[29] 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 2、6、7 . . [30] Ruilong Li、Shan Yang、David A Ross 和 Angjoo Kanazawa。AI Choreographer：使用 AIST 进行音乐条件控制的3D舞蹈生成。载于《IEEE/CVF国际计算机视觉会议论文集》，第 13401–13412页，2021年。1、2、5、6、7、9、12[31] Dongze Lian、Daquan Zhou、Jiashi Feng 和 Xinchao Wang。缩放与平移特征：一种高效模型调优的新基线。载于《神经信息处理系统进展》（NeurIPS），2022年。4[32] Angela S. Lin、Lemeng Wu、Rodolfo Corona、Kevin W. H. Tai、Qixing Huang 和 Raymond J. Mooney。根据自然语言描述生成人类活动的动画视频。2018年。3、6、8 [33] Xinyin Ma、Gongfan Fang 和 Xinchao Wang。LLM-Pruner：关于大型语言模型的结构化剪枝。arXiv 预印本 arXiv:2305.11627，2023年。4[34] Meinard M¨uller、Tido R¨oder 和 Michael Clausen。高效的基于内容的动作捕捉数据检索。载于 SIGGRAPH，第677–685页。2005年。12[35] Kensuke Onuma、Christos Faloutsos 和 Jessica K Hodgins。FMDistance：一种快速有效的动作捕捉数据距离函数。载于 Eurographics，第83–86页，2008年。12[36] Jeffrey Pennington、Richard Socher 和 Christopher D Man-ning。GloVe：词表示全局向量。载于《2014年自然语言处理实证方法会议（EMNLP）论文集》，第1532–1543页，2014年 4 – .[37] Mathis Petrovich、Michael J Black 和 G¨ul Varol。使用 Transformer VAE 进行动作条件控制的三维人体动作合成。载于《IEEE/CVF国际计算机视觉会议论文集》，第10985–10995页，2021年。2、3[38] Mathis Petrovich、Michael J Black 和 G¨ul Varol。TEMOS：根据文本描述生成多样化人体动作。载于《欧洲计算机视觉会议》，第480–497页，2022年 3 – . [39] Matthias Plappert、Christian Mandery 和 Tamim Asfour。KIT动作-语言数据集。《大数据》，第4卷，第252页，2016年 4 236 2 ( ): – .[40] Matthias Plappert、Christian Mandery 和 Tamim Asfour。使用深度循环神经网络学习人体全身动作与自然语言之间的双向映射。

循环神经网络。《机器人学与自主系统》，第109卷，第26页，2018年 13 3 : – . [41]Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sastry、Amanda Askell、Pamela Mishkin、Jack Clark 等。学习从自然语言监督中迁移视觉模型。载于《国际机器学习会议》，第8748–8763页。PMLR，2021年。1[42] Sucheng Ren、Xingyi Yang、Songhua Liu 和 Xinchao Wang。SG-Former：具有演进式令牌重分配的自引导 Transformer。载于 IEEE 国际计算机视觉会议，2023年。4[43] Sucheng Ren、Daquan Zhou、Shengfeng He、Jiashi Feng 和 Xinchao Wang。通过多尺度令牌聚合实现分流自注意力。载于 IEEE/CVF计算机视觉与模式识别会议，2022年。4[44] Xuanchi Ren、Haoran Li、Zijian Huang 和 Qifeng Chen。以音乐为条件的自监督舞蹈视频合成。载于第28届 ACM国际多媒体会议论文集，第46–54页，2020年。2[45]Robin Rombach、Andreas Blattmann、Dominik Lorenz、Patrick Esser 和 Bj¨orn Ommer。使用潜在扩散模型进行高分辨率图像合成。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第 10684 10695, 2022 1 –页。Ken [46] Shoemake。使用四元数曲线制作旋转动画。载于第12届年度计算机图形学与交互技术会议论文集，第245–254页，1985年。5、6、7、14[47] Li Siyao、Weijiang Yu、Tianpei Gu、Chunze Lin、Quan Wang、Chen Qian、Chen Change Loy 和 Ziwei Liu。Bailando：通过具有编舞记忆的演员-评论家 GPT 进行3D舞蹈生成。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第11050–11059页，2022年。1、2、4、6、7、8、9、12、14[48] Guofei Sun、Yongkang Wong、Zhiyong Cheng、Mohan S Kankanhalli、Weidong Geng 和 Xiangdong Li。Deepdance：通过对抗学习进行音乐到舞蹈的动作编舞。IEEE多媒体汇刊，第23卷，第497–509页，2020年。2[49] Taoran Tang、Jia Jia 和 Hanyang Mao。随旋律起舞：一种基于 LSTM 自编码器的音乐导向舞蹈合成方法。载于第26届 ACM国际多媒体会议论文集，第1598–1606页，2018年。2[50] Sergey Tulyakov、Ming-Yu Liu、Xiaodong Yang 和 Jan Kautz。MoCoGAN：分解视频生成中的动作与内容。载于《IEEE计算机视觉与模式识别会议论文集》，第1526–1535页，2018年 6、8 – . [51] Aaron Van Den Oord、Oriol Vinyals 等。神经离散表示学习。《神经信息处理系统进展》，第30卷，2017年。4[52] Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N Gomez、Łukasz Kaiser 和 Illia Polosukhin。注意力就是你所需要的一切。《神经信息处理系统进展》，第30卷，2017年。4[53] Jiarui Xu、Shalini De Mello、Sifei Liu、Wonmin Byeon、Thomas Breuel、Jan Kautz 和 Xiaolong Wang。GroupViT：语义分割源自文本监督。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第18134–18144页，2022年。1 [54] Nelson Yalta、Shinji Watanabe、Kazuhiro Nakadai和Tetsuya Ogata。用于基础舞步生成的弱监督深度循环神经网络。载于《2019年国际神经网络联合会议（IJCNN）论文集》，第1–8页。IEEE，2019年。2[55] Tatsuro Yamada、Hiroyuki Matsunaga和Tetsuya Ogata。用于机器人动作与语言描述之间双向翻译的成对循环自编码器。《IEEE机器人学与自动化快报》，3(4):3441–3448，2018年。3[56] Xingyi Yang、Daquan Zhou、Songhua Liu、Jingwen Ye和Xinchao Wang。深度模型重组。载于《神经信息处理系统进展》，2022年。4[57] Zijie Ye、Haozhe Wu、Jia Jia、Yaohua Bu、Wei Chen、Fanbo Meng和Yanfeng Wang。Choreonet：基于编舞动作单元的音乐到舞蹈合成。载于《第28届ACM多媒体国际会议论文集》，第744–752页，2020年。2[58] Weihao Yu、Chenyang Si、Pan Zhou、Mi Luo、Yichen Zhou、Jiashi Feng、Shuicheng Yan和Xinchao Wang。视觉MetaFormer基线。arXiv预印本，arXiv:2210 13452，2022年4。 . [59] Ye Yuan和Kris Kitani。DLow：通过多样化潜在流实现多样化的人体运动预测。载于《欧洲计算机视觉会议论文集》，第346–364页。Springer，2020年。6 [60] 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。《ACM多媒体计算、通信与应用汇刊》（TOMM），18(2):1–21，2022年。2、6、7[61] Wenlin Zhuang、Yangang Wang、Joseph Robinson、Congyi Wang、Ming Shao、Yun Fu和Siyu Xia。迈向三维舞蹈动作合成与控制。arXiv预印本，arXiv:2006.05743，2020年。2

附录：

摘要

本补充材料进一步介绍了本研究以下方面的详细信息：*i*）我们使用的数据集；*ii*）采用的评估指标；*iii*）混合数据对共享动作词元空间的影响；*iv*）收集到的数据分布；*v*）音乐-文本融合权重的作用；*vi*）我们的舞蹈定格问题较少的原因；*vii*）更多结果可视化。



(a) FID_k的差异。 (b) FID_g的差异。

图 8：实验 A 中不同批次的 FID_k 和 FID_g 。

A. 数据集详情

对于 music2dance 数据集，我们采用了 AIST++ 数据集 [30], ，该数据集包含 30 名受试者和 10 种舞蹈类型。其中有 992 段三维人体姿态序列，952 段用于训练，其余用于评估。

对于 text2motion 数据集，我们采用了 HumanML3D [17]数据集，这是一个大规模三维人体动作数据集，涵盖了广泛的人体动作，例如移动、体育运动和舞蹈。该数据集包含 14,616 段动作和 44,970 条文本描述。每段动作片段至少配有 3 条描述。为联合训练这两个数据集，我们以每秒 60 帧（FPS）的速率对动作进行采样，以保持与 AIST++数据集的时间一致性，因此时长范围为 2 至 10 秒。

为了评估我们方法的泛化能力，我们还从 YouTube 收集了一个新的音乐片段数据集，其中的音乐片段未包含在 AIST++中。该数据集包含 82 段片段，总时长为 53 分钟，是 AIST++ 测试集的八倍。片段涵盖了各种音乐风格和内容，这些都不属于 AIST++的数据分布。具体而言，我们的数据是从 YouTube 收集的流行音乐，涵盖 Glitch hop、Electro house、摇滚、future bass、独立流行和节奏布鲁斯等多种风格。相比之下，AIST++ 使用的纯舞蹈音乐来自 Old School（Break、Pop、Lock 和 Waack）与 New School（中式嘻哈、洛杉矶风格嘻哈、House、Krump、街头爵士和芭蕾爵士）等流派。此外，我们根据流行度选取了诸如“Faded”“Beat It”“Coincidence”“Baby”“Poker Face”“Despacito”“Panama”“Love Story”等野外音乐作品，这些作品的播放量均达数百万。除此之外，我们还提供了 t-SNE 特征分布图（图11），以展示与 AIST++相比，我们数据集的多样性和独特性。

B. 评估指标

我们遵循 FACT [30]和Bailando [47]以定量衡量生成的舞蹈的质量、动作的多样性以及音乐与生成的节拍对齐-

生成的动作。具体而言，对于舞蹈质量，我们计算——生成的 3D舞蹈与 AIST++ 数据集中的所有动作在动觉特征 [35] (记为‘*k*’)和几何特征 [34](记为‘*g*’)上的 Fr´echet Inception 距离（FID） [20]，这些特征由 [15] 提取，用于评估生成的舞蹈质量。我们还遵循 [30] 计算生成动作的平均特征距离，以衡量动作的多样性。音乐节拍与其最近舞蹈节拍之间的平均距离定义为节拍对齐分数，如下所示

$$\frac{1}{|B^m|}\sum_{b^m\in B^m}\exp\left\{-\frac{\min_{b^d\in B^d}\|b^d-b^m\|^2}{2\sigma^2}\right\}, \qquad (7)$$

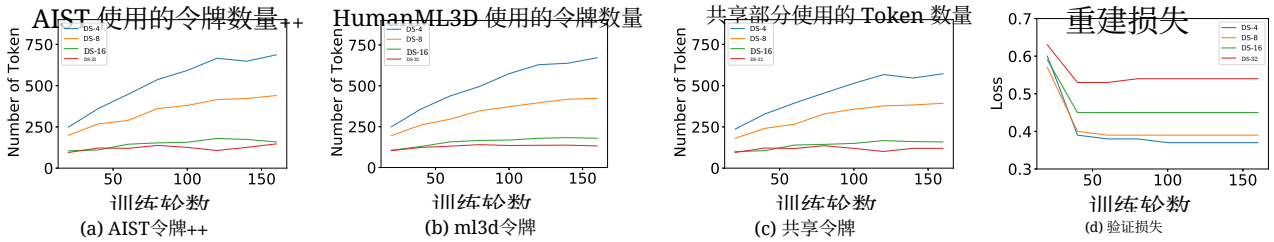
其中 B^d 和 B^m 分别是舞蹈节拍和音乐节拍。 σ 是一个归一化参数，我们在实验中将其设置为$\sigma = 3$。

对于 text2motion 质量，我们遵循 TM2T [18] 建议的相同设置：R精度和 Multimodal-Dist 量化生成运动与输入提示之间的相关性；FID 计算生成分布与真实值分布（在潜在空间中）之间的距离；多样性评估生成运动的变化；多模态性估计单个提示的方差

我们还引入了两个新的评估指标：冻结帧百分比 (PFF) 和运动预测距离 (MPD)。PFF 衡量生成的舞蹈中的冻结程度，而 MPD 评估当文本被整合时帧的连贯性。

C. 混合数据的影响

如正文所述，直接在运动空间中结合 music2dance (AIST++ [30]) 和 text2motion (HumanML3D [17]) 进行训练可能并非最优，因为这两个数据集中的动作分布在完全不同的空间中。相比之下，我们采用人体运动VQ-VAE架构，将动作投影到一致且共享的潜在空间中。为了定量展示所提出方法的有效性，我们设计了



(a) AIST令牌++ (b) ml3d令牌 (c) 共享令牌 (d) 验证损失

图 9：在实验 B 中，采用人体运动VQ-VAE架构的共享词元（潜在空间）。

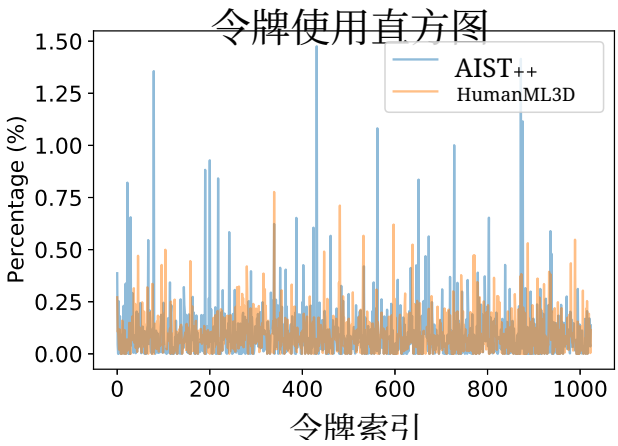


图 10：词元使用直方图，按各数据集的总帧数进行归一化。

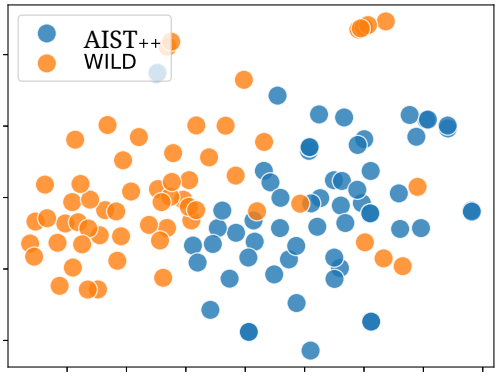


图 11：数据集的音频-t-SNE（橙色：AIST++，蓝色：我们的数据集）。

以下两项实验。

- 实验 A：我们从两个数据集中随机抽取 100 个数据批次（大小与 AIST++测试集相同），并测量随机批次与完整舞蹈数据之间的 FID。

- 实验 B：我们从两个数据集中各抽取原始数据的 30%，并使用不同的下采样率（4、8、16、32）的人体运动VQ-VAE进行训练。

在实验 A 中，图 8 展示了两个数据集的 FID 结果分布。从图 8可以看出，两个数据集在几何特征上存在明显差异，而在运动学特征上则略有重叠。

在实验 B 中，从图9中，我们有以下三个发现：i) 图 9 (a) 和图 9 (b) 显示每个数据集使用的令牌数量将随着训练轮次增加。ii) 在图 9 (c) 中，共享令牌数量也随着两个数据集一起增加。iii) 下采样率越低，使用的令牌数量和共享令牌数量越高，重建损失（验证损失）越小。考虑到下采样率越低

率越低，我们管道第二步中 Transformer 的标记化序列越长。我们选择下采样率为 8，（相对较小的验证损失、丰富的共享令牌数量和相对较短的标记化序列长度）。

从图 10可以看出，当使用 VQ-VAE 对动作进行编码时，两个数据集几乎共用一个码本。具体而言，码本共包含 1024 个向量，其中分别有 855 个和 912 个向量用于构建 AIST++和 Hu- manML3D 中的动作。846 个向量（在 AIST++中占 98.9%，在 HumanML3D 中占 92.8%）被共同用于生成动作标记，这比图 8中的特征距离结果好得多。

D. 收集数据集的分析

为验证源音乐与野生音乐之间的领域差异，我们对通过 Librosa（用于框架训练）提取的音乐特征进行采样，并在图 11中绘制 t-SNE 图。两个音乐数据集分布在不同的分布上，仅有少量重叠，这体现了我们方法的泛化能力。与混合训练相比，表1 (主

文)中的结果较差，这表明混合训练具有更好的泛化性能。

E. 融合权重与 Text2motion 结果

我们进一步探究了后期融合率（LFR）的影响，如图12所示。随着 LFR 增大，多模态距离和 Top 1 精度都会变差。为平衡特征内容，我们选择将后期融合率设为 0.8。

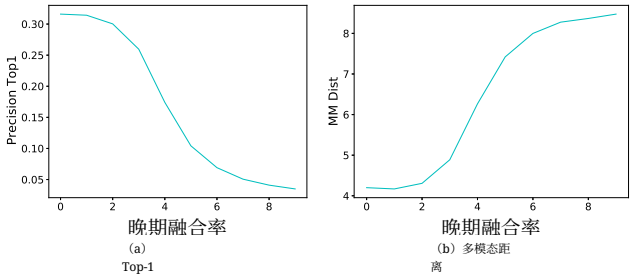


图 12：LFR 对 t2m 结果的影响。

训练	AIST++		混合数据	
test	AIST++	wild	AIST++	wild
1	4.08 / 4.08	3.76 / 3.40	5.67 / 4.88	1.21 / 0.87
10	1.31 / 1.38	0.61 / 0.63	2.58 / 2.10	0.10 / 0.12
100	0.00 / 0.00	0.78 / 0.75	0.00 / 0.00	0.12 / 0.11

表 5：PFF/AUC$_f$ 与前 k=1、10、100。

F. 冻结改进分析。

由于我们的方法在定格问题上取得了更好的结果，我们推测这种改进源于架构设计和混合训练方法。表中报告了 PFF 5。在架构方面，我们从概率最高的前 k 个词元中采样，而不是像 Bailando 那样选择概率最高的词元 [47]，从而降低了 PFF。借助额外的 HumanML3D 数据，共享动作解码器能够学习更多动作序列的统计特征。因此，PFF 进一步提升。由此可见，架构和额外数据混合训练都能改善 PFF（AUC 相同）。

G. 更多结果可视化

我们还在附件“demo.mp4”文件中展示了更多结果可视化，其中包含以下内容。

- 在 AIST++ 测试集和我们的野外数据集上与其他音乐到舞蹈方法的比较。
- 相同音乐、不同动作／时间／时长下的结果。
- 与球面线性插值的比较 [46]，用于音乐-文本条件舞蹈生成。

从这些视频中可以看出，我们的结果优于其他方法，而且更加逼真。