计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
2025,34(3):40−50 [doi: 10.15888/j.cnki.csa.009830] [CSTR: 32024.14.csa.009830]
©中国科学院软件研究所版权所有.
E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于 UnifiedGesture 改进模型的三维人体动画生成①

郭二伟, 朱欣娟, 高全力

(西安工程大学 计算机科学学院, 西安 710600)
通信作者: 朱欣娟, E-mail: 19930402@xpu.edn.cn

摘　要: 为了提升音频驱动人体动画生成的真实性, 对 UnifiedGesture 模型进行了改进研究. 首先, 通过引入编码器-解码器架构, 从音频中提取面部特征, 以弥补原模型在面部表情生成方面的不足. 其次, 结合交叉局部注意力机制和基于 Transformer-XL 的多头注意力机制, 以增强长序列中的时序依赖性. 同时, 利用变分量化自动编码器 (vector quantized variational autoencoder, VQVAE), 融合生成全身运动序列, 以提升生成动作的多样性和完整性. 最后, 在 BEAT 数据集上进行实验, 通过定量和定性分析结果表明, 改进后的 UnifiedGesture-F 模型在音频与人体动作同步性和整体真实感方面相比原模型有显著提升.

关键词: 音频驱动; 人体动画生成技术; UnifiedGesture 模型; VQVAE

引用格式: 郭二伟,朱欣娟,高全力.基于 UnifiedGesture 改进模型的三维人体动画生成.计算机系统应用,2025,34(3):40–50. http://www.c-s-a.org.cn/1003-3254/9830.html

3D Human Animation Generation Based on Improved UnifiedGesture Model

GUO Er-Wei, ZHU Xin-Juan, GAO Quan-Li

(School of Computer Science, Xi'an Polytechnic University, Xi'an 710600, China)

Abstract: This study researches improving the UnifiedGesture model to enhance the realism of audio-driven human body animation generation. Firstly, an encoder-decoder architecture is introduced to extract facial features from audio, compensating for the deficiencies of the original model in facial expression generation. Secondly, the cross-local attention mechanism and the multi-head attention mechanism based on Transformer-XL are combined to enhance the temporal dependency within long sequences. Simultaneously, the vector quantized variational autoencoder (VQVAE) is utilized to integrate and generate full-body motion sequences, enhancing the diversity and integrity of the generated motions. Finally, experiments are conducted on the BEAT dataset. The quantitative and qualitative analysis results demonstrate that the improved UnifiedGesture-F model achieves a significant improvement in the synchronicity between audio and human body movements as well as in the overall realism compared to the original model.

Key words: audio-driven; human animation generation technique; UnifiedGesture model; vector quantized variational autoencoder (VQVAE)

面部表情和肢体动作在人类交流中扮演着至关重要的角色. 它们不仅是语言的补充, 还能传达一系列的情绪、意图和细微差别. 这些非言语线索与口头语言相辅相成, 增强了交流的复杂性和丰富性. 研究表明, 面部表情和肢体动作与口头交流的整合显著影响人类互动的效率和有效性[1]. 例如, 伴随肢体动作的问题比

① 基金项目: 陕西省科技厅重点研发计划 (2024GX-YBXM-548)
收稿时间: 2024-08-11; 修改时间: 2024-09-24, 2024-11-07; 采用时间: 2024-11-19; csa 在线出版时间: 2025-01-21
CNKI 网络首发时间: 2025-01-22

© 中国科学院软件研究所 http://www.c-s-a.org.cn

没有肢体动作的问题能够引起更快速的反应, 表明肢体动作可能通过提供额外线索, 促进信息的处理, 帮助更准确地预测说话者的意图或对话走向. 此外, 通过适当的肢体动作和面部表情的表达, 可以增强交流者的可信度和参与感. 它们不仅传达信息的内容, 还展示交流者的投入和态度, 从而加深人际之间的理解和联系. 进一步地, 自动协同音频运动生成技术被认为是一种前沿使能技术, 它能够在电影、游戏和虚拟社交空间中创造出逼真的三维化身[2]. 这项技术不仅增强虚拟角色的真实感, 还使得交互能力更加自然和富有吸引力.

人体运动可以使用不同生成方法创建, 分为基于规则的过程和数据驱动的过程. 虽然基于规则的方法在一定程度上可以提供可解释性和可控性, 但随着深度学习方法的发展和普及, 数据驱动的方法通常更受青睐, 因为它们能够从数据中学习复杂的模式和规律, 更好适应不同输入和条件[3]. 而最近的运动生成方法[4,5]可以直接生成以音频为条件的人体运动. 尽管这些方法主要采用基于生成对抗网络 (generative adversarial network, GAN)[6]、变分自动编码器 (variational auto-encoder, VAE)[7]或基于 Flow 流 (normalizing flow) 模型[8]的技术, 但仍然存在一些限制. 基于 GAN 的合成方法存在模式崩溃的问题, 这可能会导致在未见过的训练数据上合成出低质量的姿态. 使用 VAE 和 Flow 的方法需要在生成质量和多样性之间做出权衡. 近期, 扩散模型 (diffusion model) 作为一种新的生成方法, 因其具备生成高质量和多样性的特性而备受关注. 例如图像生成[9], 视频生成[10]和文本生成[11]. 这些研究展示了扩散模型学习真实数据分布的能力, 同时提供了不同的采样和操作方式, 如编辑和插值. 但是, 这些研究并没有建立时间依赖序列来解决像音频驱动运动这样的时间对齐问题, 而且它们在计算资源上是密集型的.

为解决这些问题, Yang 等人[12]提出了一种音频驱动运动生成框架 UnifiedGesture. 该框架首先使用一个骨骼感知的重定位网络, 将不同骨骼标准的姿态数据统一映射到一个基准骨骼的潜在表示空间, 从而扩大数据集的规模. 然后设计一个基于扩散模型的音频驱动运动生成网络, 通过交叉局部注意力 (cross-local attention) 和自注意力机制 (self-attention), 捕捉音频和身体运动之间的时序相关性. 最后, 通过强化学习机制 (reinforcement learning, RL) 和物理引导来生成真实且符合物理规律的人体动画. 实验结果表明, UnifiedGesture 在 Trinity[13]和 ZEGGS[14]数据集上生成的身体动作更为自然, 优于先前的模型. 然而, UnifiedGesture 模型主要关注身体动作动画, 未涉及面部表情. 在人类交互中, 面部表情和身体动作相辅相成, 仅生成身体动作无法全面反映角色的情感和意图. 此外, 音频与身体动作之间的弱相关性使得传统自注意力机制在处理长序列数据时容易出现记忆力不足的问题, 难以捕捉长时间依赖关系, 从而影响生成身体动作的整体连贯性.

基于此, 本研究旨在改进 UnifiedGesture 模型, 通过结合面部表情生成和增强长时间依赖关系捕捉机制, 以实现更全面和完整的人体动画生成. 本研究的主要贡献如下所示.

(1) 面部与全身动作的协同生成: 在现有 UnifiedGesture 模型的基础上, 首次通过引入编码器-解码器架构, 将面部表情生成与全身动作生成相结合. 通过这种方式, 本研究有效弥补了原模型在面部表情生成方面的不足, 使得生成的三维人体动画不仅在全身动作上表现自然流畅, 同时也能够准确反映音频驱动下的面部情感表达.

(2) 增强的长序列依赖性处理: 原模型 UnifiedGesture 的人体运动生成方法, 虽然可以生成细粒度的动作, 但在长时间依赖处理上依然存在一定的局限性. 本研究通过结合交叉局部注意力机制和基于 Transformer-XL 的多头注意力机制, 来提升模型在捕捉长序列依赖问题上的能力.

(3) VQVAE 技术的创新: 相对于原模型, 首先, 本研究将 VQVAE 技术用于面部表情与全身动作的联合生成, 通过共享码本机制, 实现面部和身体动作的有效融合, 提升整体动画生成的质量与一致性. 其次, 在 VQVAE 的基础上引入对比损失机制, 进一步增强面部动作和全身动作特征之间的关联性, 解决面部和全身动作之间可能出现的不协调问题.

1 相关工作

1.1 协同音频运动生成

人体运动生成是一项复杂的任务, 需要理解音频、身体动作及其关系. 数据驱动的方法主要从人类动作的范式中学习相关的技巧. 研究中通常考虑 4 种主要模式: 文本[15]、音频、身体运动以及说话人身份[16]. 每种模式在人体运动生成过程中发挥着独特作用: 文本模式提供语义信息, 帮助模型理解所生成运动的语

© 中国科学院软件研究所 http://www.c-s-a.org.cn

义背景; 音频模式通过语音信号捕捉讲话者的情感、语调和韵律; 身体运动模式负责生成自然的人体姿态和肢体动作; 说话人身份模式则用于生成特定人物风格的动作和姿态. 这些模式通过协同作用, 实现了更为自然的运动生成. Habibie 等人[17]提出首个从音频输入自动生成 3D 对话中的身体运动、面部表情和头部动画的系统. 他们引入卷积神经网络 (convolutional neural network, CNN) 和生成对抗网络, 通过共享编码器学习面部表情、身体和手部动作之间的内在联系, 从而实现多模态的动作生成. Yi 等人[18]的研究引入一种全新的音频驱动全息 3D 人体运动生成技术. 该技术结合了自编码器 (autoencoder, AE) 和基于 VQVAE 的框架, 能够生成一个可组合的离散运动空间, 并通过跨条件的自回归模型生成多样且连贯的身体和手部动作. 这种方法在复杂运动的生成中表现出良好的效果. Xie 等人[19]提出一种基于离散扩散模型的非自回归生成方案, 该方案通过 VQVAE 框架将运动序列量化为离散的潜在代码序列, 并使用 CodeUnet 架构将文本转化为运动序列. 该方法在生成签名运动序列方面展示出强大的生成能力和多样性. Stan 等人[20]介绍了 FaceDiffuser 方法, 其核心是在训练过程中使用预训练的 HuBERT 音频编码器, 通过扩散模型生成高质量、多样化的面部动画序列. 该方法通过结合音频和运动特征, 能够有效地合成复杂的表情与动作动画.

尽管现有方法在生成多样化和自然的运动方面取得了一定成果, 但在面部运动生成中, 数据稀缺性仍是一个主要挑战. 为了解决这一问题, 本文研究采用了编码器-解码器架构, 通过有效的特征提取和数据增强, 缓解数据稀缺对生成质量的影响.

1.2 基于扩散模型的运动生成

扩散模型在模拟复杂数据分布和生成自然、流畅的运动序列方面表现出色, 能够实现从起点到目标的过渡. 许多研究人员将基于扩散理论的生成式模型应用于运动领域, 通过精心设计, 将无分类器的扩散生成模型网络结构适配于人体运动域. 例如, Kim 等人[21]提出一种名为 FLAME 的扩散基运动合成和编辑模型, 该模型可以生成与给定文本高度对齐的高保真运动, 并且可以无需微调地编辑运动的各个部分. Tevet 等人[22]介绍了一种用于人体运动领域的运动扩散模型 (motion diffusion model, MDM), 该模型在每个扩散步骤中预测样本而不是噪声, 这有助于使用已建立的几何损失函数对运动的位置和速度进行训练. Ren 等人[23]采用去噪扩散概率模型, 结合无分类器引导策略将文本嵌入模型训练中, 优化变分下界进行高效训练. 这些模型均基于传统的 Transformer[24]架构, 通过不同的设计优化生成运动的真实性和流畅性. 然而, 这些方法在生成长时间依赖的运动序列时仍存在不足之处.

此外, Li 等人[25]提出一个名为 Bailando 的音乐到舞蹈框架, 该框架包括一个用于将 3D 运动序列转换为量化编码序列的编舞记忆模块, 及一个通过引入交叉条件因果注意力层来增强动作生成一致性的基于演员-评论家生成预训练转换器 (GPT) 框架. Chang 等人[26]将运动生成问题形式化为一个序列到序列 (sequence to sequence, Seq2Seq) 的转换任务, 采用 Tacotron2 架构作为基础, 引入局部约束注意力机制来引导解码器从邻近潜在特征中学习依赖关系, 使生成的运动分布更接近自然运动的分布. Zhang 等人[27]提出 MotionDiffuse, 这是一种基于扩散模型的文本驱动人体动作生成框架, 通过使用交叉注意力机制 (cross-attention) 的概率映射、逼真合成和多层次操作, 实现对多样化和细粒度动作生成的优越性能. 这些模型均结合不同的注意力机制模块来提升自然运动的表现.

在本文研究中, 采用基于 routing transform[28]的交叉局部注意力机制来捕捉人体动作和音频的局部细节信息. routing transform 通过动态选择和调整注意力路径, 能够有效处理局部相关性问题, 减少信息冗余, 提高细节捕捉精度. 同时, 利用基于 Transformer-XL[29]的多头注意力机制 (multi-head attention) 捕捉全局上下文信息. Transformer-XL 通过引入相对位置编码 (relative positional encoding, RPE) 和片段级 (segment-level) 递归机制, 能够有效处理长序列数据, 捕捉长时间依赖关系. 这种结合方式有助于生成更自然和连贯的动作序列, 使其与音频更好地匹配.

2 改进的框架设计

给定一个音频片段, 本文目标是生成与之匹配的全身运动序列. 为此, 对现有框架进行改进, 提出了一个名为 UnifiedGesture-F 的框架. 如图 1 所示, 网络框架主要由两部分组成: (1) 基于编码器-解码器架构的面部生成器, 能够从音频输入中提取面部特征; (2) 基于扩散模型的身体生成器, 通过逐步地添加噪声和去噪的方式生成逼真的身体运动. 最终的全身运动序列通

© 中国科学院软件研究所 http://www.c-s-a.org.cn

过 VQVAE 网络中的联合解码器融合面部运动序列和身体运动序列得出.

2.1 面部生成器

由于面部与音频信号高度相关, 本文采用编码器-解码器架构生成面部动作, 如图 1(a) 所示. 具体流程如图 2 所示, 首先使用预训练的 wav2vec 2.0 模型[30]对音频信号$A_f=\{a_i\}_{i=0}^{T_d}$进行编码, 提取与面部相关的特征序列$F_a=\left(a_0,a_1,\cdots,a_{T_d}\right)$. 为了生成特定风格的面部动作, 将风格样式进行独热编码 (one-hot encoding), 得到编码向量E_s. 并将其E_s拼接到F_a的每个时间步上, 得到新的特征序列F'_a:

$$F'_a=\{concat\left(a_i,E_s\right)\}_{i=0}^{T_d} \tag{1}$$

其中, T_d是总时间步长. 独热编码是一种将分类变量转换为二进制向量的方式, 使模型能够识别不同的风格类别. 在解码器部分, 引入多头注意力机制来进一步处理F'_a. 多头注意力机制通过多个注意力头并行计算每个特征的注意力权重, 捕捉输入特征的长短期依赖关系, 从而提升模型的表现力和精度. 然后, 使用时间卷积网络 (temporal convolution network, TCN) 对特征序列进行建模. TCN 通过膨胀卷积扩展感受野, 捕捉更长时间范围内的依赖, 并具有残差连接以保证梯度顺畅传播. 最后, 通过一个全连接层 (fully connected layer, FC) 输出面部动作. 本文采用均方误差 (mean squared error, MSE) 损失函数训练编码器和解码器, 以最小化生成动作与真实动作之间的误差.

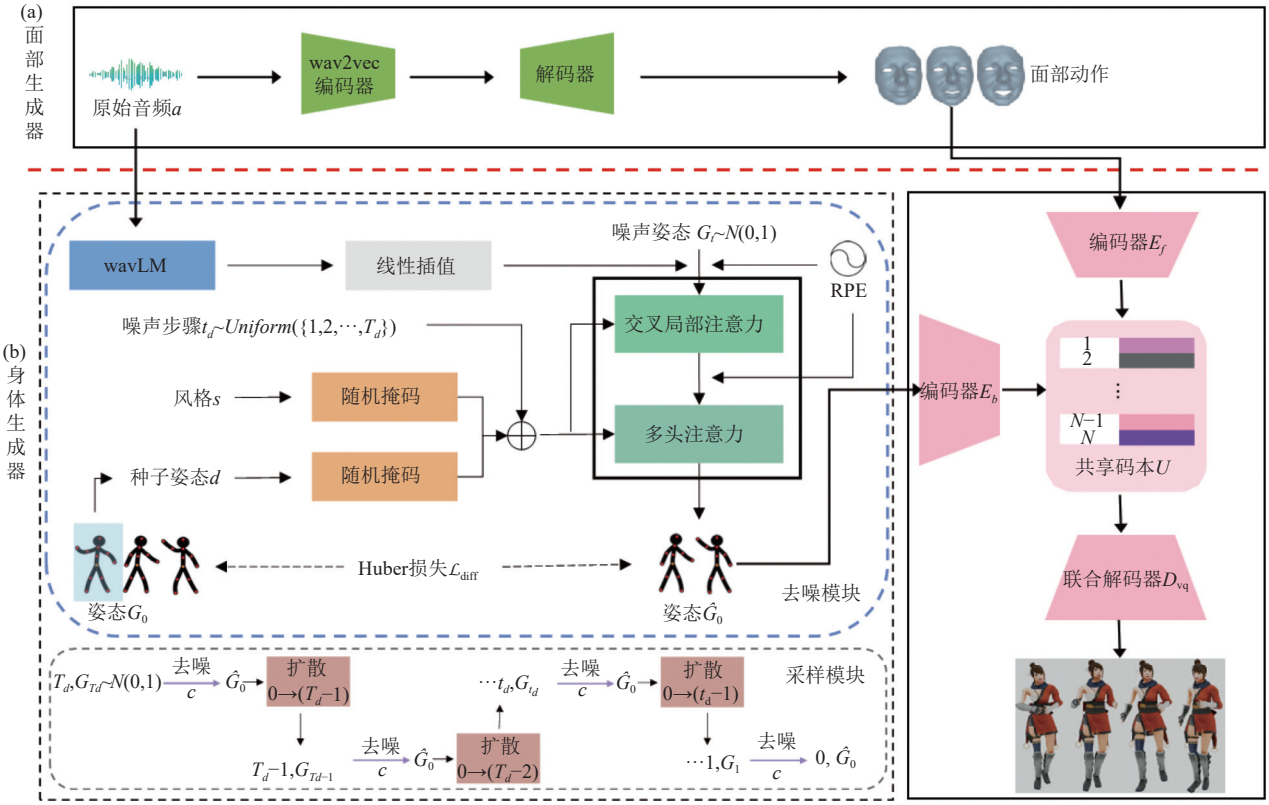

图 1 UnifiedGesture-F 模型的整体框架 (图中实线框为改进部分)

2.2 身体生成器

对于身体生成器, 通过扩散模型架构来生成音频驱动的身体动作, 如图 1(b) 所示. 扩散模型由正向过程 (扩散过程) q和反向过程 (去噪过程) p_θ两部分组成. 本文将扩散过程q中生成的身体姿态表示为G, 它与观测数据$G_0\sim q(G_0)$具有相同的维度, $q(G_0)$表示真实数据的分布. 根据方差时间表$\beta_1,\beta_2,\cdots,\beta_{T_d}$ $(0<\beta_1<\beta_2<\cdots<\beta_{T_d}<1)$, ($T_d$为总时间步长), 加入高斯噪声:

$$q\left(G_{t_d}|G_{t_d-1}\right)=N\left(G_{t_d};\sqrt{1-\beta_{t_d}}G_{t_d-1},\beta_{t_d}I\right) \tag{2}$$

去噪过程p_θ是一个通过神经网络学习参数θ的过程. 假设在时间t_d处的噪声G_{t_d}学习到的参数为μ_θ, Σ_θ, 则:

$$p_\theta\left(G_{t_d-1}|G_{t_d}\right)=N\left(G_{t_d-1};\mu_\theta\left(G_{t_d},t_d\right),\Sigma_\theta\left(G_{t_d},t_d\right)\right) \tag{3}$$

© 中国科学院软件研究所 http://www.c-s-a.org.cn

2.2.1 去噪模块

使用基于相对位置编码 (RPE)[31]的注意力机制实现去噪. 主要思路是在给定噪声步骤t_d、噪声姿态G_{t_d}和条件c (包括种子姿态d、风格s和音频a) 的情况下合成一个长度为N的姿态$G^{1:N}$:

$$\hat{G}_0 = Denoise\left(G_{t_d}, t_d, c\right) \tag{4}$$

其中, 种子姿态d和风格s使用伯努利掩码随机屏蔽 10% 的样本, 用于无分类器学习. 然后, 结合预测的条件模型$Denoise\left(G_{t_d}, t_d, c_1\right)$, $c_1 = [d, s, a]$和无条件模型$Denoise\left(G_{t_d}, t_d, c_2\right)$, $c_2 = [\varnothing, \varnothing, a]$, 通过$\gamma$插值实现姿态生成的无分类器引导以及不同样式的控制:

$$\begin{aligned}\hat{G}_{0\gamma,c_1,c_2} =& \gamma Denoise\left(G_{t_d}, t_d, c_1\right) \\ &+ (1-\gamma) Denoise\left(G_{t_d}, t_d, c_2\right)\end{aligned} \tag{5}$$

最后, 通过优化生成的姿态G_0与真实姿态$\hat{G}_0$之间的 Huber 损失[32]来训练去噪模块:

$$\mathcal{L}_{\text{diff}} = E_{G_0 \sim q(G_0|c), t_d \sim [1, T_d]}\left[HuberLoss\left(G_0 - \hat{G}_0\right)\right] \tag{6}$$

在训练过程中, 对每个特征的处理过程如下.

(1) 噪声步骤t_d从均匀分布$t_d \sim Uniform(\{1, 2, \cdots, T_d\})$中采样. 并通过多层感知机 (multilayer perceptron, MLP) 映射到 256 维的空间T.

(2) 噪声姿态G_{t_d}与从标准正态分布$N(0, 1)$中采样得到的姿态G_0具有相同的维数. 随后, 通过线性层调整到 256 维的空间G.

(3) 音频特征由 wavLM large[33]的预训练模型生成, 并通过线性插值将 wavLM 特征和姿态G_0在时间维度上对齐, 以确保特征的一致性. 然后, 使用线性层将维度降低到 64 维, 形成最终的音频特征A_b.

(4) 风格s用独热向量 (one-hot vector) 表示, 选定的样式中只有一个元素是非零的, 并通过线性层映射到 64 维的空间S.

(5) 种子姿态d从真实姿态数据中提取的一个初始片段获得, 其第 1 帧N_{seed}作为种子姿态d, 其余N帧作为真实姿态G_0来计算损失, 以实现连续合成之间的平滑过渡. 之后, 通过线性层映射到 192 维的空间D.

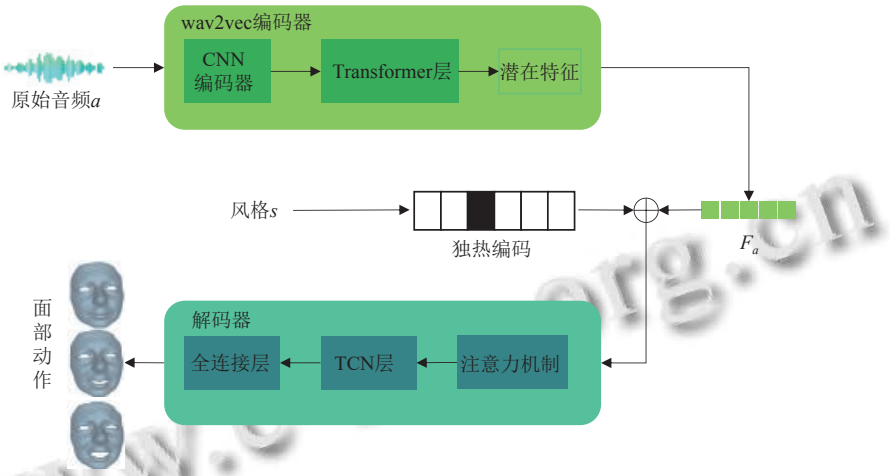

图 2 面部动作生成概述

接下来, 将种子手势D和风格S拼接成一个 256 维的向量, 并添加噪声步骤T的信息形成向量Z. 然后, 将向量Z复制 N 次, 与音频特征A_b和姿态特征G合并, 以便网络更好地考虑长时间范围内的上下文信息. 接着, 将合并后的特征输入交叉局部注意网络, 以在局部区域内聚焦注意力, 捕捉关键信息. 之后, 将交叉局部注意的输出与向量Z连接, 输入多头注意网络, 增强网络对输入数据的理解和处理能力, 生成更准确的共语言姿态. 最后, 多头注意网络的输出经过一个线性层后被映射回与G_0相同的维度.

2.2.2 采样模块

最终的身体姿态是通过拼接一些时长为T_c, 帧长度为N的片段得到的. 第 1 个片段的种子姿态可以通过从数据集中随机选择一个姿态或者将其设定为平均姿态来生成. 然后其他片段的种子姿态是上一个片段生成的姿态的末端N_{seed}帧. 在每个片段的去噪步骤t_d中, 都会预测出干净的姿态$G_0 = Denoise\left(G_{t_d}, t_d, c\right)$, 并使用式 (1) 将噪声添加到去噪步骤$G_{t_d-1}$中. 这个过程从$t_d = T_d$重复到$t_d = 0$为止.

© 中国科学院软件研究所 http://www.c-s-a.org.cn

2.2.3 运动融合框架

本文设计了一种新的基于 VQVAE[34]的框架, 旨在实现面部运动和身体运动的融合生成. VQVAE 的量化过程不仅有助于减少运动冻结并保留运动细节, 同时对生成多样性也起着重要作用. 通过训练 VQVAE, 可以学习到一个紧凑而有意义的运动空间. 如图 3 所示, VQVAE 框架包括两个编码器、一个共享码本和一个联合解码器.

首先, 面部运动F_0和身体运动B_0分别通过一维时间卷积网络编码成潜在特征z_f和z_b:

$$z_f = E_f(F_0) \tag{7}$$

$$z_b = E_b(B_0) \tag{8}$$

其中, $z_f, z_b \in \mathbb{R}^{\frac{T}{d_{\text{vq}}} \times C''}$, d_{vq}为下采样率, C''是量化特征的通道维数. 然后, 利用可学习的共享码本对编码后的潜在特征进行量化. 设共享码本为$U = \{u_i\}_{i=1}^{N}$, 其中N是码本的大小. 通过计算编码特征与码本中所有代码的欧氏距离, 选择最接近的代码u_i作为量化结果:

$$z_{q_f} = \arg\min_{u_i \in U} \left\| z_f - u_i \right\|_2 \tag{9}$$

$$z_{q_b} = \arg\min_{u_i \in U} \left\| z_b - u_i \right\|_2 \tag{10}$$

最后, 将量化后的特征z_{q_f}和z_{q_b}传递给联合解码器D_{vq}, 以生成最终的全身运动序列$\hat{M}_0$:

$$\hat{M}_0 = D_{\text{vq}}\left(z_{q_f}, z_{q_b}\right) \tag{11}$$

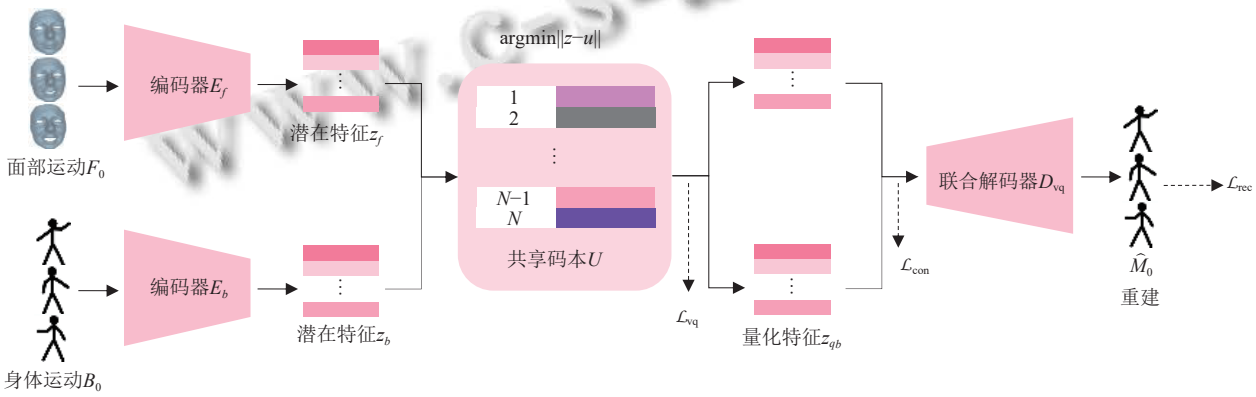

图 3 VQVAE 框架概述

VQVAE 的训练过程包括重构损失$\mathcal{L}_{\text{rec}}$、向量量化损失$\mathcal{L}_{\text{vq}}$和对比损失$\mathcal{L}_{\text{con}}$. 重构损失$\mathcal{L}_{\text{rec}}$用于精确地恢复运动的位置、速度和加速度:

$$\mathcal{L}_{\text{rec}} = \left\| \hat{M}_0 - M_0 \right\|_1 + \alpha_1 \left\| \hat{M}_0' - M_0' \right\|_1 + \alpha_2 \left\| \hat{M}_0'' - M_0'' \right\|_1 \tag{12}$$

其中, M_0'和M_0''分别为运动表示M_0的一阶和二阶偏导数, α_1和α_2分别为对应项的平衡权值. 向量量化损失$\mathcal{L}_{\text{vq}}$确保网络学习到的数据关键特征能够通过离散的向量进行有效表示:

$$\mathcal{L}_{\text{vq}_f} = \left\| sg\left[z_f\right] - z_{q_f} \right\|_2 + \beta_f \left\| z_f - sg\left[z_{q_f}\right] \right\|_2 \tag{13}$$

$$\mathcal{L}_{\text{vq}_b} = \left\| sg\left[z_b\right] - z_{q_b} \right\|_2 + \beta_b \left\| z_b - sg\left[z_{q_b}\right] \right\|_2 \tag{14}$$

其中, $sg[\cdot]$是停止梯度函数, 第 1 项是码本损失, 第 2 项是承诺损失, 权重分别为β_f和β_b. 为增强面部运动和身体运动特征之间的关联性, 并促进多样化的运动表示, 引入了对比损失$\mathcal{L}_{\text{con}}$:

$$\mathcal{L}_{\text{con}} = \lambda_{\text{pos}} \mathcal{L}_{\text{con}}^{+} + \lambda_{\text{neg}} \mathcal{L}_{\text{con}}^{-} \tag{15}$$

其中, λ_{pos}和λ_{neg}是正负样本对比损失的权重系数. 正负样本对比损失分别定义为:

$$\mathcal{L}_{\text{con}}^{+} = \frac{1}{N_{\text{pos}}} \sum_{j=1}^{N_{\text{pos}}} \left\| z_{q_f,j} - z_{q_b,j} \right\|_2^2 \tag{16}$$

$$\mathcal{L}_{\text{con}}^{-} = \frac{1}{N_{\text{neg}}} \sum_{j=1}^{N_{\text{neg}}} \max\left(0, m - \left\| z_{q_f,j} - z_{q_b,j} \right\|_2\right)^2 \tag{17}$$

其中, N_{pos}和N_{neg}分别是正负样本对的数量, $z_{q_f,j}$和$z_{q_b,j}$分别是第j个样本的面部和身体的特征向量, m是控制负样本对间隔的超参数. 对于每个样本, 将其面部特征向量z_{q_f}和身体特征向量z_{q_b}视为正样本对, 通过最小化这两个特征向量之间的距离来增强它们的相似性. 同时, 从批次中的其他样本中随机选择一个不同的样本作为负样本对, 通过最大化这两个特征向量之间的距离来增强它们的差异性. 总损失函数如下:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{vq}_f} + \mathcal{L}_{\text{vq}_b} + \mathcal{L}_{\text{con}} \tag{18}$$

© 中国科学院软件研究所 http://www.c-s-a.org.cn

VQVAE 通过最小化$\mathcal{L}_{\text{total}}$促进模型学习并生成逼真、多样且细致的全身运动序列.

3 实验

3.1 实验准备

3.1.1 实验设置

本研究在 GeForce RTX 3090 Ti GPU 上进行实验, 采用 BEAT 数据集[35]进行训练和评估. BEAT 数据集包含来自 4 种不同语言的 30 位演讲者的演讲和对话数据, 涵盖 8 种不同情绪 (自然、愤怒、快乐、恐惧、厌恶、悲伤、蔑视和惊讶) 下的谈话动作, 总长度为 76 h. 本文选择英语演讲者的高质量数据, 总计约 32 h, 并将数据按照 8:1:1 的比例分为训练集、验证集和测试集. 动作数据下采样至 30 FPS, 音频数据下采样到 16 kHz.

面部生成器的编码器参数使用预训练的 wav2vec 2.0 权重进行初始化. 多头注意力网络使用 8 个注意力头, 每个头维度为 64. TCN 设置为 4 层, 每层 TCN 包含 64 个卷积核, 卷积核大小为 3. 优化器采用 Adam, 学习率为 1E–4, $\beta_1=0.9$, $\beta_2=0.999$. 在扩散模型中, 交叉局部注意网络使用 8 个头, 32 个注意通道, 256 个通道, 窗口大小为 6, 前后窗口差值为 0.1. 多头注意网络由 8 层、8 个头、32 个注意通道和 256 个通道组成, 差值为 0.1. 采用 AdamW 优化器, 学习率设为 3E–5, 批次大小为 256, 噪声步骤 T=1000. 对于 VQVAE 模型, 码本 U 的大小为 512, 维度为 512. 损失函数中, 下采样率 $d_{\text{vq}}=2$、$\alpha_1=\alpha_2=1$、$\beta_f=\beta_b=0.1$、$\lambda_{\text{pos}}=\lambda_{\text{neg}}=0.5$、$N_{\text{pos}}=N_{\text{neg}}=1024$. 优化器采用 Adam, 学习率设为 1E–4、$\beta_1=0.5$、$\beta_2=0.98$, 批次大小为 256, 训练 200 个 epoch.

3.1.2 评价指标

为验证改进后的模型 UnifiedGesture-F 的性能, 本文将其与 3 个最相关的模型进行对比.

(1) ZeroEGGS[14]: 最佳基线模型, 是一个音频驱动的姿态生成模型, 通过一个短例子的动作片段来控制风格, 并能够泛化到训练数据之外的风格.

(2) UnifiedGesture[12]: 一种统一的姿态合成模型, 用于多种骨架, 通过使用基于扩散的模型进行音频驱动的姿态生成, 并引入强化学习以提高生成运动与音频的匹配度.

(3) TalkShow[18]: 从人类语音中生成三维全身运动的方法, 通过利用两个学习过的身体和手的码本来生成确定性的面部运动和不同的身体和手的运动.

评估标准涵盖了面部和身体运动的真实感、多样性及同步性, 具体包括以下指标.

(1) *LVD* (lip-sync value distance)[36]: 用于评估音频与面部动画同步性的指标, 主要考察嘴唇动作与发音的匹配程度. *LVD* 值越低, 嘴唇动作与音频同步性越好.

$$LVD=\frac{1}{N}\sum_{i=1}^{N}\|y_i-\hat{y}_i\| \tag{19}$$

其中, y_i和$\hat{y}_i$分别表示第i帧的真实和生成的嘴唇特征, N是总帧数.

(2) *L2* 距离 (Euclidean distance)[37]: 用于衡量生成的面部表情特征点与真实面部表情特征点之间的差异. 平均 *L2* 距离值越小, 说明生成的面部动画与真实面部的差异越小, 真实感越高.

$$L2\ distance=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2} \tag{20}$$

其中, x_i和y_i是生成和真实面部表情特征点的坐标.

(3) *CCA* (canonical correlation analysis)[38]: 用于测量音频特征与身体运动之间的相关性. 较高的 *CCA* 值表明音频特征与身体运动之间有较强的统计关联.

$$CCA=\max_{\alpha,\beta}\frac{\alpha^{\mathrm{T}}\sum_{xy}\beta}{\sqrt{\alpha^{\mathrm{T}}\sum_{xx}\alpha}\sqrt{\beta^{\mathrm{T}}\sum_{yy}\beta}} \tag{21}$$

其中, $\sum_{xy}$是音频特征和身体运动特征的协方差矩阵, $\sum_{xx}$和$\sum_{yy}$分别是音频特征和身体运动特征的协方差矩阵, α和β是线性组合系数向量.

(4) *FGD* (Fréchet gesture distance)[39]: 基于 Fréchet 距离概念, 用于评估生成的动作数据与真实人类动作数据之间的相似度. *FGD* 值越小, 表示生成的身体动作与真实动作在统计意义上越接近, 真实感和多样性越好.

$$FGD=\left\|\mu_r-\mu_g\right\|^2+\mathrm{tr}\left(\sum_r+\sum_g-2\left(\sum_r\sum_g\right)^{\frac{1}{2}}\right) \tag{22}$$

其中, μ_r和μ_g分别是真实和生成动作数据的均值向量, $\sum_r$和$\sum_g$分别是其协方差矩阵, tr表示迹运算.

3.2 比较结果

3.2.1 客观评价

针对身体动作生成的性能, 定量结果如表 1 所示. 可以看出, 本文提出的模型在全局 *CCA* 上优于其他所

© 中国科学院软件研究所 http://www.c-s-a.org.cn

有方法. 这表明该模型能有效地捕捉并复制身体动作的基本结构和特征, 使生成的运动具有高度的保真度和连贯性. 在每个序列的 *CCA* 方面也表现出相对稳定的性能, 这意味着该模型能够在不同的序列中保持较高的一致性. 此外, 在比较 *FGD* 时, 本文的方法相较于最佳基线模型 ZeroEGGS 提高了 65.4%, 相较于原模型 UnifiedGesture 提高了 5.8%, 进一步突显了生成动作的高质量.

表 1 基于测试集上身体运动的定量结果

模型	全局*CCA*	每个序列的*CCA*	*FGD*↓
ZeroEGGS	0.905	**0.96±0.01**	11.48
UnifiedGesture	0.978	0.94±0.01	4.213
UnifiedGesture-F (本文)	**0.984**	0.95±0.01	**3.968**

注: 加粗字体表示最佳度量

对于面部动作生成的定量比较, 结果如表 2 所示. 对比显示, 本文的面部生成器能够生成与输入音频高度一致的面部运动, 证明了其优越性能.

表 2 基于测试集上面部运动的定量结果

模型	*LVD*↓	*L2*↓
TalkShow	0.012 41	0.073 28
UnifiedGesture-F (本文)	**0.012 40**	**0.063 02**

注: 加粗字体表示最佳度量

3.2.2 用户研究

为对所提出的方法进行定性评估, 本文通过用户研究分析生成的运动性能. 研究设计 10 个样本, 每个样本包含 2 个问题. 在这些样本中, 视频动作基于相同的音频输入, 分别由 UnifiedGesture-F、UnifiedGesture[12] 和 TalkShow[18]生成. 参与者需要从两个方面对不同方法的结果进行排名: (1) 生成动作的真实感 (不考虑言语的影响); (2) 生成动作与输入音频的匹配度 (考虑言语的节奏与语义). 此次研究收集了 20 名参与者的回答. 其中男性 17 人, 女性 3 人. 年龄分布为 20–30 岁 18 人, 40–60 岁 2 人. 本文从不同的角度评估了每种方法被认为排名第 1 的百分比. 图 4 显示了 3 种方法在两个指标上的统计结果. 结果显示, 超过 60% 的用户认为本文方法产生的结果更真实, 超过 55% 的用户认为本文方法产生的结果更符合音频节奏. 这些结果表明, 引入的面部生成器、多头注意力机制和 VQVAE 技术在提升动画生成效果方面是有效的.

3.3 消融实验

此外, 本文进行了消融实验, 以评估框架中不同模块对生成性能的影响. 包括: (1) 共享码本, (2) 长时间依赖注意力机制, 以及 (3) VQVAE. 实验结果总结在表 3 中. *LVD* 指标显示, 去除共享码本后, 面部和身体动作的协同性和同步性显著降低, 表明生成动作的分布偏离真实动作. 从 *CCA* 和 *L2* 距离指标可以看出, 长时间依赖注意力机制显著提升了动作生成的时间一致性和真实感, 去除该模块后, 动作生成的时间连贯性和分布相似性减弱. 去除 VQVAE 模块时, 所有指标均出现显著退化, 尤其是 *FGD* 和 *CCA*, 表明 VQVAE 通过离散特征量化有效增强了生成动作的多样性和真实感. 这些结果验证了各模块设计对生成性能的有效性和必要性.

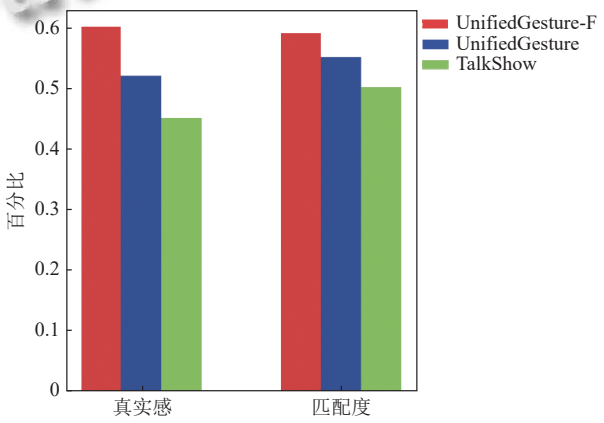

图 4 统计结果

表 3 消融实验结果

消融实验设置	*LVD*↓	*L2*↓	*FGD*↓	全局*CCA*↑
UnifiedGesture-F	**0.012 40**	**0.063 02**	**3.968**	**0.984**
–共享码本	0.015 31	0.075 03	4.116	0.962
–长时间依赖的注意力机制	0.012 75	0.069 12	4.103	0.975
–VQVAE	0.015 90	0.076 48	4.210	0.953

注: “–”表示未使用的模块, 加粗表示最佳度量

3.4 人体动画生成效果展示

本文提出的方法能够生成高度逼真的音频驱动人体动画, 尤其注重面部表情的生成效果. 为了展示生成结果的多样性, 本文对比了不同模型在相同音频输入下的动画效果, 这些音频涵盖“快乐”“愤怒”“悲伤”“惊讶”“自然”和“厌恶”6 种情绪, 并用虚线框标示同一帧下的动作进行对照. 图 5 展示了 TalkShow 和 UnifiedGesture-F 模型在“快乐”“愤怒”和“悲伤”3 种情绪下的面部表情生成效果. 可以看出, UnifiedGesture-F 模型在嘴型的弧度和自然度方面表现更好, 而 TalkShow 模型的生成效果较为僵硬. 图 6 展示了 UnifiedGesture 和

© 中国科学院软件研究所 http://www.c-s-a.org.cn

Unidentified-F 模型在“惊讶”“自然”和“厌恶” 3 种情绪下的身体动作生成效果. 结果表明, UnifiedGesture-F 模型生成的胳膊和手臂动作更协调, 腿部和脚的运动也更加自然和细腻, 而 UnifiedGesture 模型的动作略显机械和不连贯. 这些对比清晰展示了本文方法在不同情绪下生成真实人体动画的优越效果.

图 5 TalkShow 生成视频和 UnifiedGesture-F 生成视频连续 10 帧面部运动对比效果

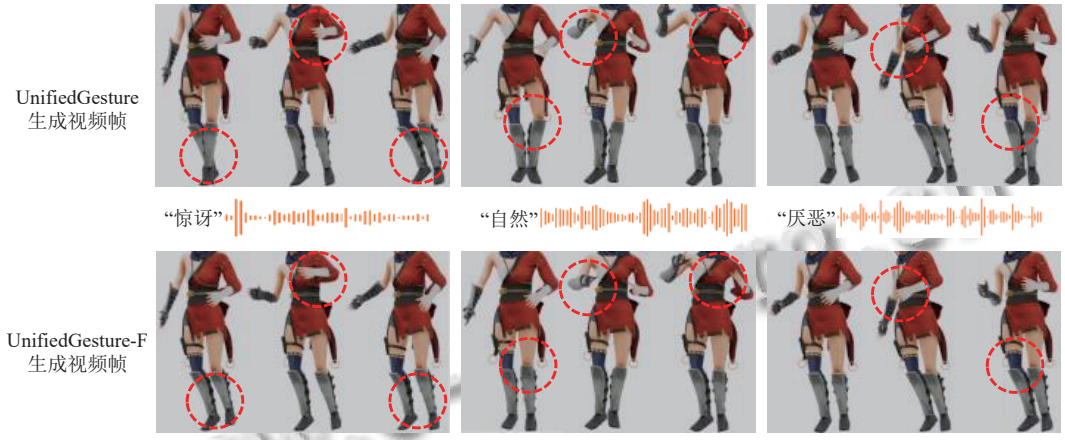

图 6 UnifiedGesture 生成视频和 UnifiedGesture-F 生成视频连续 10 帧身体运动对比效果

4 结论与展望

本文提出了一种基于 UnifiedGesture 改进模型的三维人体动画生成方法: UnifiedGesture-F. 该方法使用简洁高效的编码器-解码器架构生成真实自然的面部表情. 通过引入多头注意力机制和 VQVAE 技术, 增强了生成动画的多样性和连贯性, 能够生成高质量、音频匹配且风格可控的姿态. 尽管该方法在生成逼真动画方面取得显著进展, 但仍面临一些挑战, 解决扩散模型在实时系统中采样步骤过多、耗时较长的问题, 将是未来值得探索的研究方向.

参考文献

1 Holler J, Kendrick KH, Levinson SC. Processing language in face-to-face conversation: Questions with gestures get faster responses. Psychonomic Bulletin & Review, 2018, 25(5): 1900–1908.

2 Nyatsanga S, Kucherenko T, Ahuja C, *et al*. A comprehensive review of data-driven co-speech gesture generation. Computer Graphics Forum, 2023, 42(2): 569–596. [doi: 10.1111/cgf.14776]

3 Li ZH, Liu JZ, Zhang ZS, *et al*. CLIFF: Carrying location information in full frames into human pose and shape

© 中国科学院软件研究所 http://www.c-s-a.org.cn

estimation. Proceedings of the 17th European Conference on Computer Vision. Tel Aviv: Springer, 2022. 590–606.
4 Harz L, Voß H, Kopp S. FEIN-Z: Autoregressive behavior cloning for speech-driven gesture generation. Proceedings of the 25th International Conference on Multimodal Interaction. Paris: ACM, 2023. 763–771.
5 Kucherenko T, Jonell P, Yoon Y, *et al*. A large, crowdsourced evaluation of gesture generation systems on common data: The GENEA challenge 2020. Proceedings of the 26th International Conference on Intelligent User Interfaces. College Station: ACM, 2021. 11–21.
6 Lu ZM, Xiao Y, Sun ZJ, *et al*. Adversarial learning for implicit semantic-aware communications. Proceedings of the 2023 IEEE International Conference on Communications. Rome: IEEE, 2023. 4063–4069.
7 Kucherenko T, Hasegawa D, Henter GE, *et al*. Analyzing input and output representations for speech-driven gesture generation. Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents. Paris: ACM, 2019. 97–104.
8 Alexanderson S, Henter GE, Kucherenko T, *et al*. Style-controllable speech-driven gesture synthesis using normalising flows. Computer Graphics Forum, 2020, 39(2): 487–496. [doi: 10.1111/cgf.13946]
9 Ramesh A, Dhariwal P, Nichol A, *et al*. Hierarchical text-conditional image generation with CLIP latents. arXiv:2204.06125, 2022.
10 Mei KF, Patel V. VIDM: Video implicit diffusion models. Proceedings of the 37th AAAI Conference on Artificial Intelligence. Washington: AAAI, 2022. 9117–9125.
11 Lin ZH, Gong YY, Shen YL, *et al*. Text generation with diffusion language models: A pre-training approach with continuous paragraph denoise. Proceedings of the 40th International Conference on Machine Learning. Honolulu: PMLR, 2023. 21051–21064.
12 Yang SC, Wang ZL, Wu ZY, *et al*. UnifiedGesture: A unified gesture synthesis model for multiple skeletons. Proceedings of the 31st ACM International Conference on Multimedia. Ottawa: ACM, 2023. 1033–1044.
13 Ferstl Y, McDonnell R. Investigating the use of recurrent motion modelling for speech gesture generation. Proceedings of the 18th International Conference on Intelligent Virtual Agents. Sydney: ACM, 2018. 93–98.
14 Ghorbani S, Ferstl Y, Holden D, *et al*. ZeroEGGS: Zero-shot example-based gesture generation from speech. Computer Graphics Forum, 2023, 42(1): 206–216. [doi: 10.1111/cgf.14734]
15 Alexanderson S, Székely É, Henter GE, *et al*. Generating coherent spontaneous speech and gesture from text. Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents. ACM, 2020. 1.
16 Liu X, Wu QY, Zhou H, *et al*. Learning hierarchical cross-modal association for co-speech gesture generation. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 10452–10462.
17 Habibie I, Xu WP, Mehta D, *et al*. Learning speech-driven 3D conversational gestures from video. Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents. ACM, 2021. 101–108.
18 Yi HW, Liang HL, Liu YF, *et al*. Generating holistic 3D human motion from speech. Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023. 469–480.
19 Xie P, Zhang QP, Li ZX, *et al*. G2P-DDM: Generating sign pose sequence from gloss sequence with discrete diffusion model. Proceedings of the 38th AAAI Conference on Artificial Intelligence. Washington: AAAI, 2024. 6234–6242.
20 Stan S, Haque KI, Yumak Z. FaceDiffuser: Speech-driven 3D facial animation synthesis using diffusion. Proceedings of the 16th ACM SIGGRAPH Conference on Motion, Interaction and Games. Rennes: ACM, 2023. 13.
21 Kim J, Kim J, Choi S. FLAME: Free-form language-based motion synthesis & editing. Proceedings of the 37th AAAI Conference on Artificial Intelligence. Washington: AAAI, 2023. 8255–8263.
22 Tevet G, Raab S, Gordon B, *et al*. Human motion diffusion model. arXiv:2209.14916, 2022.
23 Ren ZY, Pan ZH, Zhou X, *et al*. Diffusion motion: Generate text-guided 3D human motion by diffusion model. Proceedings of the 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Rhodes Island: IEEE, 2023. 1–5.
24 Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 6000–6010.
25 Li SY, Yu WJ, Gu TP, *et al*. Bailando: 3D dance generation by actor-critic GPT with choreographic memory. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 11040–11049.

© 中国科学院软件研究所 http://www.c-s-a.org.cn

26 Chang CJ, Zhang S, Kapadia M. The IVI lab entry to the GENEA challenge 2022-A Tacotron2 based method for co-speech gesture generation with locality-constraint attention mechanism. Proceedings of the 2022 International Conference on Multimodal Interaction. Bengaluru: ACM, 2022. 784–789.

27 Zhang MY, Cai ZA, Pan L, *et al*. MotionDiffuse: Text-driven human motion generation with diffusion model. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(6): 4115–4128. [doi: 10.1109/TPAMI.2024.3355414]

28 Aurko R, Mohammad S, Ashish V, *et al*. Efficient content-based sparse attention with routing Transformers. Transactions of the Association for Computational Linguistics, 2021, 9: 53–68.

29 Dai ZH, Yang ZL, Yang YM, *et al*. Transformer-XL: Attentive language models beyond a fixed-length context. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: ACL, 2019. 2978–2988.

30 Baevski A, Zhou H, Mohamed A, *et al*. wav2vec 2.0: A framework for self-supervised learning of speech representations. Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2020. 1044.

31 Kitaev N, Kaiser L, Levskaya A. Reformer: The efficient Transformer. Proceedings of the 8th International Conference on Learning Representations. OpenReview.net, 2020.

32 Huber PJ. Robust estimation of a location parameter. In: Kotz S, Johnson NL, eds. Breakthroughs in Statistics: Methodology and Distribution. New York: Springer, 1992. 492–518.

33 Chen SY, Wang CY, Chen ZY, *et al*. WavLM: Large-scale self-supervised pre-training for full stack speech processing. IEEE Journal of Selected Topics in Signal Processing, 2022, 16(6): 1505–1518. [doi: 10.1109/JSTSP.2022.3188113]

34 Van Den Oord A, Vinyals O, Kavukcuoglu K. Neural discrete representation learning. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 6309–6318.

35 Liu HY, Zhu ZH, Iwamoto N, *et al*. BEAT: A large-scale semantic and emotional multi-modal dataset for conversational gestures synthesis. Proceedings of the 17th European Conference on Computer Vision. Tel Aviv: Springer, 2022. 612–630.

36 Zhou Y, Han XT, Shechtman E, *et al*. MakeltTalk: Speaker-aware talking-head animation. ACM Transactions on Graphics (TOG), 2020, 39(6): 221.

37 Ng E, Joo H, Hu LW, *et al*. Learning to listen: Modeling non-deterministic dyadic facial motion. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 20363–20373.

38 Sadoughi N, Busso C. Speech-driven animation with meaningful behaviors. Speech Communication, 2019, 110: 90–100. [doi: 10.1016/j.specom.2019.04.005]

39 Yoon Y, Cha B, Lee JH, *et al*. Speech gesture generation from the trimodal context of text, audio, and speaker identity. ACM Transactions on Graphics (TOG), 2020, 39(6): 222.

(校对责编: 王欣欣)

© 中国科学院软件研究所 http://www.c-s-a.org.cn