基于细粒度时空建模的语音驱动手势生成模型

万浩聪[1]，**刘长红**[1*]，**杨海**[1]，**江爱文**[1,2]，**王明文**[1,2]
[1]江西师范大学，计算机信息工程学院，江西，南昌，330022
[2]江西师范大学，数字产业学院，江西，上饶，334000
Email: {202341600104, liuch, 202341600091, jiangaiwen, mwwang }@jxnu.edu.cn

摘要

语音驱动手势生成技术根据输入的语音自动生成丰富的虚拟角色动作，在数字动画、虚拟现实和人机交互等领域具有广泛的应用前景。虽然现有方法在时序连贯性方面取得一定进展，但由于缺乏对关节间局部交互的显式建模，生成的肢体动作往往存在机械感且缺乏自然性。针对这一问题，提出一种基于细粒度时空注意力的扩散模型，从细粒度层面建模骨架关节点间的动态依赖关系。具体而言，设计了一种时空Transformer，其中空间注意力层显式建模了关节间的空间结构关系，而时序注意力层捕获手势运动的动态性。此外，通过自适应实例归一化技术AdaIN引入说话者身份控制，实现个性化手势生成。在BEAT、BEAT2和SHOW数据集上验证了所提模型的有效性。

关键词： 语音驱动手势生成；扩散模型；时空Transformer；AdaIN

Fine-grained Spatio-temporal Modeling for Speech-driven Gesture Generation Model

HaoCong Wan[1]，**ChangHong Liu**[1*]，**Hai Yang**[1]，
AiWen Jiang[1,2]，**Mingwen Wang**[1,2]
[1]School of Computer and Information Engineering, Jiangxi Normal University, Nanchang, Jiangxi , 330022
[2]School of Digital Industry, Jiangxi Normal University, Shangrao, Jiangxi, 334000
Email: {202341600104, liuch, 202341600091, jiangaiwen, mwwang }@jxnu.edu.cn

Abstract

Speech-driven gesture generation technology has broad application prospects in digital animation, virtual reality, and human-computer interaction, enabling more natural and expressive body movements for virtual characters. Although existing methods have made some progress in temporal coherence, the generated movements are inflexible and unnatural due to the lack of explicit modeling of local interactions between key joints. To address this issue, a fine-grained spatio-temporal attention-based diffusion model, FineGesture, is proposed, which models the dynamic dependencies between joints at a fine-grained level. Specifically, a spatio-temporal transformer is designed, in which the spatial attention layer explicitly models the spatial structural relationships between joints, while the temporal attention layer captures the dynamics of movements. Additionally, speaker identity is introduced to achieve personalized gesture generation through the Adaptive Instance Normalization (AdaIN) technique. The effectiveness of the proposed model is validated on the BEAT, BEAT2 and SHOW datasets.

Keywords: Speech-driven gesture generation , Diffusion model , Spatio-temporal transformer , Adaptive instance normalization

第二十四届中国计算语言学大会论文集，第684页-第695页，济南，中国，2025年8月11日至14日。

1 引言

在人类的日常交流中，手势作为一种非语言信息的载体，与语言表达紧密相连，共同构成丰富的沟通方式(De Ruiter et al., 2012)。手势不仅能够增强语言的表达力，还能传递情感、态度和意图，使交流更加生动和自然(Burgoon et al., 1990)。随着人工智能技术的飞速发展，虚拟角色和具身智能体等概念逐渐从科幻走向现实(Nyatsanga et al., 2023)。语音驱动手势生成技术能够根据输入的语音自动合成逼真的肢体动作，是这些应用中关键挑战之一，对于提升交互的真实感和自然度至关重要(Yang et al., 2022)。

许多近期的工作(Chen et al., 2024a; Liu et al., 2025a; Xu et al., 2024)通常将人体分为脸部，上半身，下半身和手四个部分，采用连续或离散的矢量量化表示人体姿势(Liu et al., 2023; Yi et al., 2023; Yang et al., 2023b)，能够生成局部合理的身体姿势，但由于采用统一的运动特征学习过程，忽略了不同关节点之间细微的互动，在整体动作的协调性和自然度方面仍存在不足(Yang et al., 2023c)。近年来，扩散模型（Diffusion Models）(Ho et al., 2020; Nichol and Dhariwal, 2021; Cheng et al., 2025)因其强大的生成能力被引入手势生成领域。DiffGesture(Zhu et al., 2023)提出了一种基于扩散模型的语音驱动手势生成框架，通过时序Transformer(Zhang et al., 2022)建模动作的长期依赖关系，显著提升了生成质量。GestureLSM(Liu et al., 2025b)采用潜在空间建模和区域划分策略提升生成效率，但其基于预定义身体区域的注意力机制难以捕捉跨区域的关节协同，如：手部转动带动手腕微调。这类方法通常将身体局部作为整体处理，忽略了全身关节间的细粒度空间关系，导致生成的局部动作（如：手指运动）不够自然。

针对上述问题，提出一种基于细粒度时空建模的语音驱动手势生成模型（**Fine**-grained Spatio-temporal Modeling for Speech-driven **Gesture** Generation Model），称作FineGesture。该模型分为两个阶段，第一个阶段采用矢量量化变分自编码器(Vector Quantized Variational Autoencoder,VQVAE)(Van Den Oord et al., 2017)框架构建离散运动空间，以学习特定的手势动作表示，并且通过整合速度损失和加速度损失增强VQVAE，学习运动先验知识；第二阶段利用VQVAE学习到的运动先验来优化动作生成；在生成过程中，以潜在扩散模型(Rombach et al., 2022; Ji et al., 2023)作为手势生成器，设计空间和时间注意力机制捕获手势动作的空间结构和时序动态性，实现关节点间更自然的交互，并通过自适应实例归一化（Adaptive Instance Normalization,AdaIN）(Huang and Belongie, 2017)机制引入说话人身份ID属性，实现个性化手势生成。本文的主要贡献包括：

- 提出一种关节级时空注意力机制。设计细粒度的时空Transformer层，显式建模关节点间的生物力学约束，从而更精准地捕捉局部动作细节。
- 通过AdaIN模块融合说话人身份特征，在统一框架下实现身份感知的个性化手势生成。
- 大量的实验与分析验证了所提模型FineGesture的有效性。

2 相关工作

2.1 语音驱动手势生成

语音驱动手势生成是一个复杂的问题，需要全面理解语音和手势之间的关系。基于规则的手势生成通过手动设置的规则将语言特征（如：语义和语调）与预定义的手势进行匹配(Cassell et al., 1994; Cassell et al., 2001; Kopp et al., 2006)，生成相应的手势序列。这种方法逻辑清晰，易于定制，但灵活性较差，需要手动定义大量规则，难以应对复杂多变的语言场景。深度学习的发展使神经网络能够直接从原始多模态数据中捕捉语音和手势之间的复杂关系(Habibie et al., 2021; Liu et al., 2022b; Kucherenko et al., 2020)，基于数据驱动的方法至此逐渐成为主流(Yoon et al., 2020; Cassell et al., 2001; Yi et al., 2023; Liu et al., 2022a; Nyatsanga et al., 2023)。Ginosar等人(Ginosar et al., 2019)提出的Speech2Gesture首次证明了从原始语音直接生成手势的可行性；Yoon等人(Yoon et al., 2020)提出的Trimodal引入文本转录和说话人身份信息，显著提升了生成手势的语义相关性；Liu等人(Liu et al., 2022b)的HA2G通过分层音频特征提取器捕获语音中的不同粒度信息，进一步改善了动作节奏的匹配度；DiffGesture(Zhu et

al., 2023)通过扩散音频-手势Transformer捕捉音频与手势的长期依赖关系，并利用扩散手势稳定器优化动作流畅性；DiffSHEG(Chen et al., 2024b)提出一种基于Transformer的扩散方法，实现任意长度的语音生成手势动作。近年来，基于多阶段训练和分离-组合策略实现全身手势生成展现出一种趋势(Liu et al., 2023; Chen et al., 2024a; Yi et al., 2023)，首先通过多个预训练VQVAE离散表示学习人体各部分运动，再组合各部分离散表示训练全局运动生成器，从而增强人体姿势的多样性和真实性。

2.2 基于扩散模型的手势生成

扩散模型在动作生成领域的应用最初集中在文本到动作任务上。MDM(Tevet et al., 2022)首次验证了扩散模型在3D人体动作生成中的潜力，其通过逐步去噪过程实现了高质量动作合成。随后，DiffGesture(Zhu et al., 2023)将该框架扩展到语音驱动手势场景，设计了一个基于Transformer的音频-手势扩散模型。由于音频中较弱的语义信号通常会导致动作与输入音频的语义内容不一致，EMoG(Yin et al., 2023)引入自适应层归一化(Huang and Belongie, 2017)，将情感条件注入到手势生成过程，DiffuseStyleGesture(Yang et al., 2023a)将情绪融入手势生成过程，增强模型的跨模态关联能力。最近的工作开始探索扩散模型的效率优化(Jiang et al., 2023; Zhang et al., 2023; Liu et al., 2025b; Liu et al., 2025a)，通过引入注意力蒸馏技术加速推理或潜在空间扩散降低计算成本。

3 方法

所提出模型FineGesture基于潜在扩散模型(Rombach et al., 2022)，以历史手势、语音音频和说话人ID作为输入，在潜在空间中通过扩散去噪过程，实现自然且个性化的手势合成。整个实现过程采用两阶段训练方式：首先，基于VQVAE框架训练了一个手势编解码器，通过量化编码技术捕获手势运动的关键特征，为后续生成任务奠定基础；然后，基于潜在扩散模型构建多模态条件生成网络，生成逼真且风格化的手势。

3.1 多模态特征表示

(1)手势特征

基于VQVAE构建手势特征学习网络，学习紧凑的运动表示。该网络为编码器-解码器结构，编码器对手势特征进行编码，得到潜在特征表示，由4个一维卷积网络组成，解码器重构手势动作，由2个残差块和4个一维卷积网络组成，如图1所示。

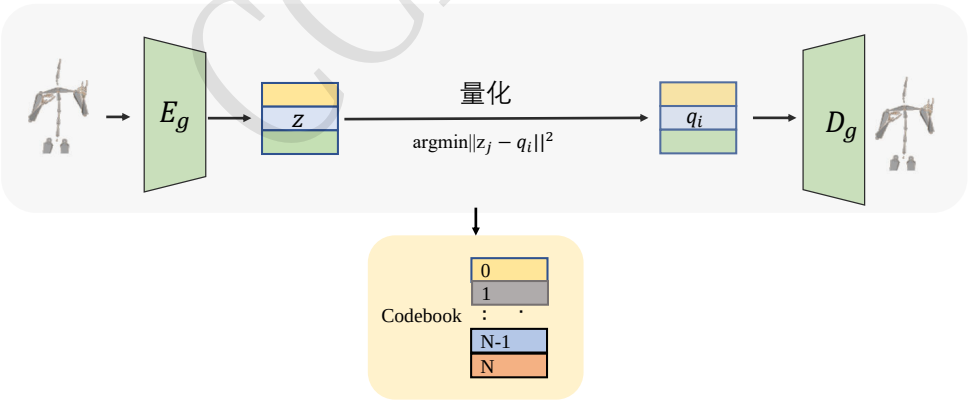

图 1: 手势特征学习网络结构图

给定一个手势序列$g = \{g_i\}_{i=1}^{N}$，其中$g_i \in R^{3J}$(J和N分别为关节点数量和手势序列的总长度)，手势编码器E_g将手势序列g编码成潜在向量$Z = \{z_j\}_{j=1}^{N}$，其中$z_j \in R^d$(d表示潜在特征维度)，描述为：

$$z = E_g(g), \tag{1}$$

对于每个潜在向量z_j，通过最近邻搜索在码本$q = \{q_i\}_{i=1}^{L}$（L为码本大小）中找到对应的量化向量q_i，量化过程表示为：

$$q_i = \text{argmin}||\text{z}_j - q_i||^2, \tag{2}$$

通过将高维度的连续手势运动转化为低维度的离散表征，既保留了运动的关键特征，又显著提升了生成效率，为后续的条件生成任务奠定基础。

手势解码器D_g根据量化后的向量重构原始手势序列，表示为：

$$\widehat{g} = D_g\left(q_i\right). \tag{3}$$

为了更充分地学习手势运动信息，结合手势动作的速度和加速度先验知识，通过联合优化重构损失、速度损失和加速度损失训练VQVAE网络:

$$\mathcal{L}_{\text{VQVAE}} = \mathcal{L}_{\text{rec}}(g, \hat{g}) + \mathcal{L}_{\text{vel}}(g', \hat{g}') + \mathcal{L}_{\text{acc}}(g'', \hat{g}'') + \|\text{sg}[z] - q\|^2 + \|z - \text{sg}[q]\|^2, \tag{4}$$

其中重构损失$\mathcal{L}_{\text{rec}}$、速度损失$\mathcal{L}_{\text{vel}}$和加速度损失$\mathcal{L}_{\text{acc}}$均采用L_1损失函数，sg表示梯度运算结束，g'和g''分别表示手势序列在时序上的一阶差分和二阶差分。

(2)音频特征

音频被下采样到15fps，每帧的音频特征维度为1067，通过音频编码器WavEncoder编码得到音频潜在特征$\mathbf{Z}_A \in R^{128}$。音频编码器由一个12层时间卷积网络（Temporal Convolutional Network, TCN）和2层多层感知机（Multilayer Perceptron, MLP）(Van Den Oord et al., 2017)组成。

(3)说话人身份ID特征

说话人身份ID首先初始化为One-hot表示，然后通过嵌入层(Yoon et al., 2020)映射为低维特征$\mathbf{Z}_{ID} \in R^8$。

3.2 FineGesture模型

FineGesture基于潜在扩散模型(Rombach et al., 2022)，在潜在特征空间中应用正向和反向扩散过程实现手势生成。扩散过程被建模为马尔可夫噪声过程，正向扩散过程逐渐对提取的手势序列潜在特征$\mathbf{Z}_0$添加高斯噪声，直到其分布接近$N(0, I)$，其分布演变描述为:

$$q(\boldsymbol{Z}_t|\boldsymbol{Z}_{t-1}) = \mathcal{N}\left(\sqrt{\alpha_t}\boldsymbol{Z}_{t-1}, (1-\alpha_t)\boldsymbol{I}\right), \tag{5}$$

其中Z_t表示第n步扩散采样的手势序列潜在特征，$t \in \{1, \ldots, T\}$，a_t由方差调度策略确定。相反，逆向扩散(去噪)过程对含噪手势序列潜在特征$\boldsymbol{Z}_T \sim \mathcal{N}(0, \boldsymbol{I})$逐步去除噪声，通过去噪网络$E_\theta$预测噪声$E_n$，恢复原始手势序列潜在特征$Z_0$。该去噪网络结构见图2所示。$E_\theta$可表述为：

$$\boldsymbol{E_n} = \boldsymbol{E_\theta}(\boldsymbol{Z_n}, \boldsymbol{t}, \boldsymbol{A}, \boldsymbol{ID}) \tag{6}$$

其中E_n是去噪网络预测的噪声，t是扩散步数，A是语音音频，ID是说话人身份ID。

给定音频序列$A \in \mathbb{R}^{N \times H}$和历史手势序列$P = \{P_m\}_{m=1}^{4}$，首先分别通过音频编码器WavEncoder和手势编码器E_g得到音频特征$\mathbf{F_a}$和历史手势特征$\mathbf{F_p}$，描述为：

$$F_a = WavEncoder\left(g\right), \tag{7}$$

$$F_p = E_g\left(P\right), \tag{8}$$

其中N为手势序列的帧数，H为单帧音频的维度。

然后将历史手势特征$\mathbf{F_p}$与音频特征$\mathbf{F_a}$及时间步长t沿着特征维度拼接，得到上下文c，接着将加噪后的手势序列潜在特征Z_t与上下文c拼接得到$\widehat{Z}_t$，随后将$\widehat{Z}_t$充当一个单独的令牌输入时空Transformer。

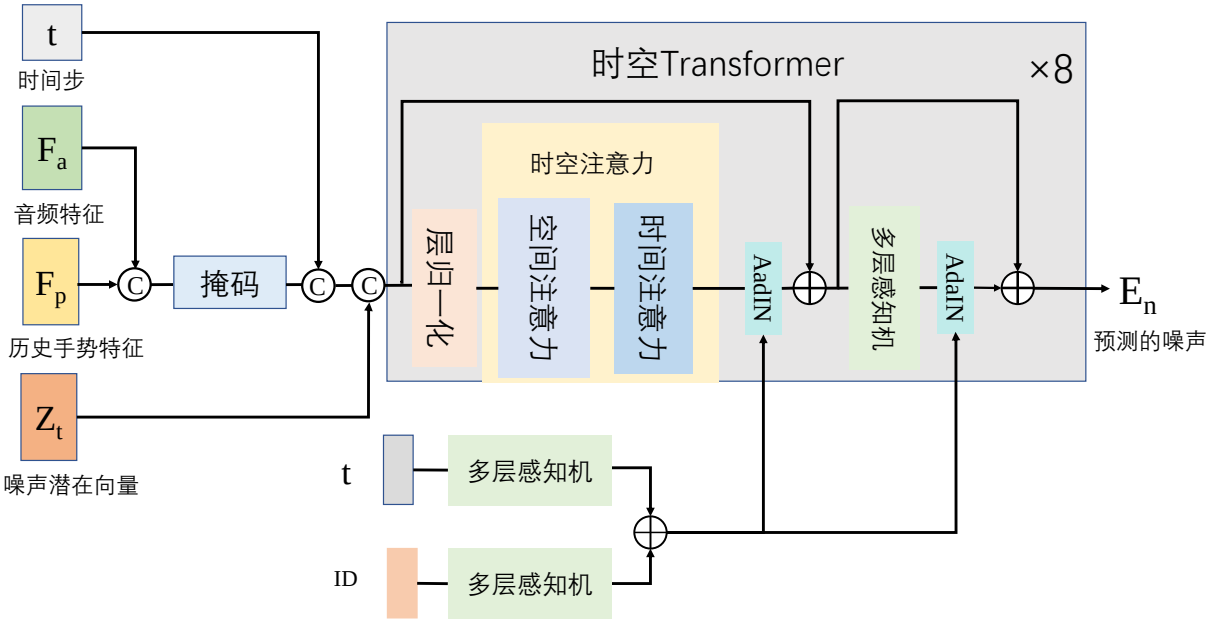

图 2: FineGesture模型去噪网络结构图

3.3 时空Transformer

为了捕捉手势的时空依赖关系并实现个性化手势生成，设计了时空Transformer模块，其中时空注意力模块分别通过空间和时间注意力机制捕获手势动作的空间结构和时序动态性，实现关节点间更自然的交互，并通过自适应实例归一化机制引入说话人身份ID属性和时间步t，实现个性化手势生成，具体结构见图2所示。

空间注意力捕获每一帧中不同关节点之间的空间依赖关系，确保生成的手势在空间上的协调性。将加噪后的手势序列潜在特征Z_t与上下文c拼接得到的$\widehat{Z}_t$分别作为查询、键和值计算空间注意力，表示为:

$$Attetion(Q_s, K_s, K_s)_s = soft\max\left(\frac{Q_s \cdot K_s^T}{\sqrt{d}}\right)V_s, \tag{9}$$

其中$Q_s, K_s, V_s \in \mathbb{R}^{d\times L}$，d代表特征维度大小，L代表手势序列长度。

时间注意力捕获手势序列中不同时间帧之间的依赖关系，确保生成的手势在时序上的连贯性。具体地，将空间注意力的输出作为时间注意力的查询、键和值输入，计算时间注意力，表示为:

$$Attetion(Q_t, K_t, K_t)_t = softmax\left(\frac{Q_t \cdot K_t^T}{\sqrt{d}}\right)V_t, \tag{10}$$

其中$Q_t, K_t, V_t \in \mathbb{R}^{L\times d}$，d代表特征维度大小，L代表手势序列长度。

时空注意力模块通过空间注意力和时间注意力分别建模手势的依赖关系。空间注意力计算单帧内各关节点的交互，确保生成手势在空间上自然协调；时间注意力则分析帧间运动关联，保证动作在时间上的流畅过渡。两者协同工作，使手势既符合人体运动学约束，又保持动态连贯性。

进一步地，将说话人ID作为风格条件和扩散步数t分别通过线性投影后，相加得到融合的全局条件向量，在时间注意力和MLP层之后通过AdaIN引入全局条件，控制去噪过程生成风格化的手势动作。

3.4 模型训练

遵循条件去噪扩散模型的训练流程(Rombach et al., 2022)，利用无分类器引导训练模型FineGesture，使扩散模型在训练过程中能够在条件分布和非条件分布下进行学习。对于输入

的手势动作序列，采用第一阶段预训练的手势编码器进行编码，得到手势序列潜在特征Z_0，随后向Z_0逐步添加随机高斯噪声ϵ，加噪t步后的噪声手势序列潜在特征为Z_t，通过优化以下损失函数训练去噪网络E_θ。

$$L_{\text{noise}} = \|\epsilon - \boldsymbol{E}_\theta(\boldsymbol{Z}_t, \boldsymbol{t}, \boldsymbol{A}, \boldsymbol{ID})\|_2^2, \tag{11}$$

4 实验

为了评价模型的有效性，在三个公开的基准数据集BEAT(Cassell et al., 2001)、BEAT2(Liu et al., 2023)和SHOW(Yi et al., 2023)上进行了对比实验，采用FGD(Yoon et al., 2020)、SRGR(Cassell et al., 2001)、BeatAlign(Li et al., 2021b)、BC(Li et al., 2021b)、Diversity(Li et al., 2021a)和PCM(Chen et al., 2024b)作为评价指标。

4.1 实验细节

所有对比方法采用统一训练参数设置，并在NVIDIA A4000 GPU（16GB显存）和CUDA Toolkit 11.8环境下完成。FineGesture模型使用ADAM优化器(Xue et al., 2020)，去噪网络基于一个8层、512特征维度、8个注意力头的Transformer，扩散步骤为500步。在构建VQVAE模型时，将码本特征长度大小设为64，码本大小固定为256。该码本仅在训练第一阶段更新，在第二阶段训练期间冻结。VQVAE模型共训练500个epoch，采用3×10^{-4}的学习率；第二阶段模型训练100个epoch,训练批次大小为32，采用2×10^{-4}的学习率。

4.2 数据集

BEAT(Cassell et al., 2001)是一个大规模、多模态的对话手势数据集，包含身体动作、面部表情、音频和文本四种模态，总时长76小时，来自30位说话者，涵盖四种语言和八种情感。遵循基准，使用4位说话人(“Socct”，“Lawrence”，“Carla，”and “Catherine”)

BEAT2是由EMAGE(Liu et al., 2023)提出的大规模多模态人体手势数据集，包含文本转录、语义与情感标签以及60小时的运动数据。(Liu et al., 2023)将其划分为BEAT2-Standard（27小时）和BEAT2-Additional（33小时）两个子集。遵循了文献(Liu et al., 2023)的实验设置，在BEAT2-Standard的Speaker-2子集上按85%、7.5%和7.5%的比例划分训练集、验证集和测试集。

SHOW(Yi et al., 2023)是一个音频-视觉数据集，其中包括从30fps的视频中重建的4个人的SMPLX(Pavlakos et al., 2019)参数，以及相对应的22K采样率的同步音频。

4.3 评价指标

Frechet Gesture Distance (FGD)：用于评估生成手势$\hat{\mathbf{m}}$与真实手势$\mathbf{m}$之间的分布距离。它通过预训练的网络提取手势的潜在特征，并计算这些特征之间Frechet Inception Distance (FID):

$$\text{FGD}(\mathbf{m}, \hat{\mathbf{m}}) = \|\mu_r - \mu_g\|^2 + \text{Tr}(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{1/2}), \tag{12}$$

其中μ_r和Σ_r是真实手势$\mathbf{m}$ 的潜在特征的均值和协方差，μ_g和Σ_g是生成手势$\hat{\mathbf{m}}$的潜在特征的均值和协方差。

Semantic Relevance Gesture Recall (SRGR)：用于评估生成手势的语义相关性。它通过将语义得分作为权重，计算生成手势与真实手势之间Probability of Correct Keypoint (PCK)。PCK 是指在给定阈值内成功匹配的关节数量。

Beat Alignment Score(BeatAlign)：通过计算手势节拍和音频节拍之间的Chamfer Distance来评估手势和音频节拍之间的相似性。

$$\text{BeatAlign} = \frac{1}{n}\sum_{i=1}^{n} \exp\left(-\frac{\min_{\forall b_j^a \in B^a} \|\mathbf{b}_i^m - \mathbf{b}_j^a\|_2^2}{2\sigma^2}\right), \tag{13}$$

其中$B^m = \{b_i^m\}$和$B^a = \{b_j^a\}$分别表示手势节拍和音频节拍的集合，σ 是根据帧率调整的参数。

Beat Constancy(BC)：用来评估语音和动作的一致性。

第二十四届中国计算语言学大会论文集，第684页-第695页，济南，中国，2025年8月11日至14日。

Diversity：用于评估生成模型输出结果多样化程度的指标，其核心目标是衡量模型能否生成动态丰富且差异显著的动作或样本。具体而言，该指标通过计算生成样本之间的特征距离(如欧氏距离、余弦距离等)来量化多样性。

Percent of Correct Motion paramter(PCM): 是一种基于运动参数（而非关键点）计算的性能指标，用于评估运动估计的准确率。

4.4 定量实验

在BEAT、BEAT2和SHOW数据集上对所提模型FineGesture与最近先进方法进行了定量对比实验，评估所提模型的性能，见表1、表2和表3所示，表中加粗字体表示指标在所有方法中的最优性能，下划线表示该指标的次优性能。

表 1: 在BEAT数据集上定量对比结果

方法	FGD↓	SRGR↑	BeatAlign↑
CaMN(Cassell et al., 2001)	<u>173.27</u>	<u>0.216</u>	0.725
HA2G(Liu et al., 2022b)	303.17	0.180	0.904
DiffGesture(Zhu et al., 2023)	365.99	0.079	**0.922**
EMAGE(Liu et al., 2023)	246.93	0.211	0.793
DiffSHEG(Chen et al., 2024b)	439.93	**0.251**	<u>0.914</u>
FineGesture	**90.29**	0.195	0.910

BEAT数据集是一个多模态数据集，重点关注语音与手势的同步性和多样性，因此选择了FGD、SRGR和BeatAlign这些能够全面评估生成手势质量和同步性的指标。表1展示了BEAT数据集上的定量对比结果，从表中可以看出，所提出模型FineGesture在FGD指标上明显优于所有基线方法，表明生成的手势分布与真实数据最为接近。具体而言，FineGesture在训练过程中更注重生成手势的全局分布和时序一致性，而DiffSHEG和CaMN更注重手势的语义相关性，所以FineGesture的SRGR指标低于DiffSHEG和CaMN。值得注意的是，虽然DiffGesture在BeatAlign指标上表现略优，但其FGD和SRGR指标表现较差。综合来看，FineGesture在生成手势的整体质量、语义相关性和同步性上实现了较好的平衡。

表 2: 在BEAT2数据集上定量对比结果

方法	FGD↓ $\times 10^{-1}$	BC↑ $\times 10^{-1}$	Diversity↑
CaMN(Cassell et al., 2001)	6.644	6.769	10.86
HA2G(Liu et al., 2022b)	12.32	6.779	8.686
TalkSHOW(Yi et al., 2023)	6.209	6.947	**13.47**
EMAGE(Liu et al., 2023)	**5.512**	<u>7.724</u>	13.06
FineGesture	<u>6.133</u>	**8.101**	<u>13.13</u>

BEAT2数据集是BEAT的改进版本，增加了对节奏稳定性和全局运动的关注，因此引入了BC指标来补充BeatAlign的不足，更全面地评估手势的节奏表现。表2展示了BEAT2数据集上的定量对比结果。从表中可以看出，所提出模型FineGesture在BC指标上获得最优性能，显著优于其他对比方法，这表明生成的手势与音频具有更好的同步性，而且在FGD和Diversity指标上，同样获得了较优的性能，分别略低于EMAGE和TalkSHOW，这也验证了FineGesture能够有效地建模关节点间的局部交互关系，使生成的手势动作具有较好的真实性和多样性。特别值得注意的是，在所有三个指标上都保持排名在前两位，展现了优异的综合性能。

表3展示了SHOW数据集上的定量对比结果。该数据集相比较于BEAT和BEAT2数据集，提供了更全面的性能分析视角，采用了PCM指标对生成的手势运动进行评估。由于其他方法没有SHOW数据集上的实验结果，所以仅与SHOW数据集上已有对比的方法TalkSHOW(Yi et al., 2023)和DiffSHEG(Chen et al., 2024b)方法进行了对比分析。从表中可以看出，所提出的FineGesture模型在FGD指标上同样取得了最优性能，在与音频节奏同步性指标BeatAlign上同样获得了较好的性能，接近于最优的方法。由于FineGesture通过时空注意力机制捕获关节点

表 3: 在SHOW数据集上定量对比结果

方法	FGD↓	BeatAlign↑	PCM↑
TalkSHOW(Yi et al., 2023)	0.00323	0.869	0.902
DiffSHEG(Chen et al., 2024b)	0.00271	**0.902**	**0.912**
FineGesture	**0.00268**	0.896	0.898

间的交互关系，更侧重于提升整体手势动作的真实性和多样性，而DiffSHEG侧重优化运动参数，在PCM指标上低于DiffSHEG。总体而言，这些结果验证了FineGesture在跨数据集场景下的稳定表现，同时明确了后续优化的方向。

4.5 定性实验

为了验证手势生成的效果，进一步对不同方法在BEAT数据集上生成的手势动作序列进行了可视化比较，如图3所示，其中“CaMN”表示CaMN(Cassell et al., 2001)生成的手势序列，“HA2G”为“HA2G”(Liu et al., 2022b)生成的手势序列，“DiffGesture”为DiffGesture(Zhu et al., 2023)生成的手势序列，“EMAGE”为EMAGE(Liu et al., 2023)生成的手势序列，”DiffSHE“为DiffSHE(Chen et al., 2024b)生成的手势序列。

从图3中可以发现，HA2G和DiffGesture常产生违背生物力学的异常姿势，如：出现前臂非自然外翻（红色实框线），CaMN、EMAGE和DiffSHEG虽能保持关节合理性，但生成手势幅度受限且节奏变化平缓，相比之下，FineGesture生成的手势更为连贯、自然且逼真。

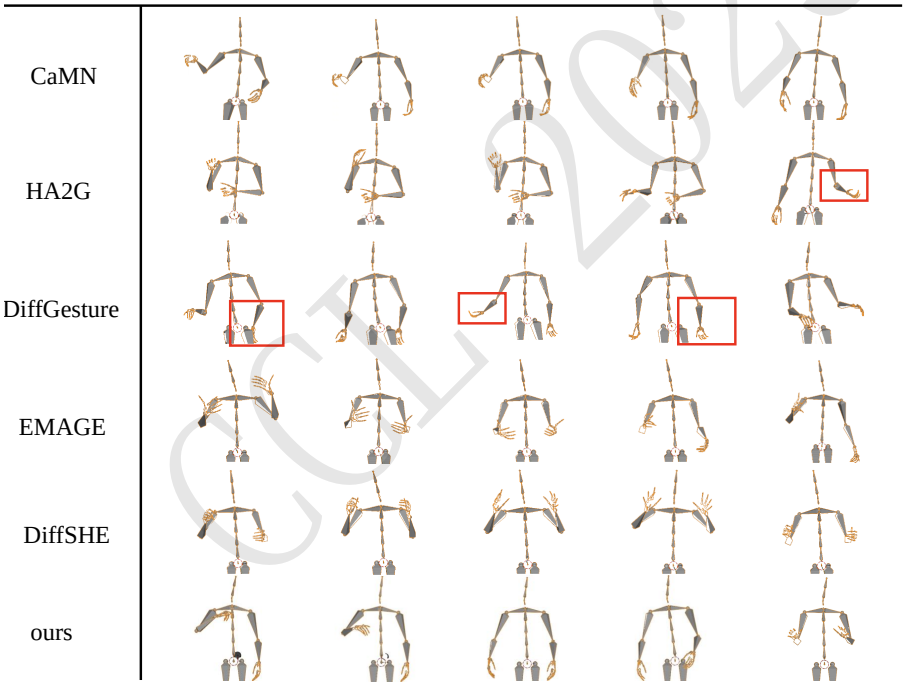

图 3: 各方法在BEAT数据集上的可视化对比结果

4.6 消融实验

在BEAT数据集上进行了三组消融实验。实验包括：模型架构的有效性验证、VQVAE模块的有效性验证、VQVAE码本特征长度的消融实验。通过对比不同实验设置，逐步分析并确定最佳实验配置。

(1)模块消融实验

为了验证FineGesture各个模块的有效性，对比分析了以下架构变体:(1)无空间注意力：时空间注意力模块中仅使用时间注意力。(2)无时间注意力机制：时空注意力模块中仅使用空间

注意力。实验结果见表4。从表中可观察到，移除空间注意力导致FGD和SRGR指标性能明显下降，移除时间注意力虽对节奏同步影响较小但仍会明显损害手势运动的连贯性，而增加了时空注意力在所有评估指标上都取得了最优性能。这是由于空间注意力捕获了手势动作在空间结构上的几何关系且时间注意力优化了运动动态性，通过时空注意力协同使语义匹配度较单注意力变体有所提升，这也充分表明了时空注意力机制在模型架构中的有效性。

表 4: 不同模块的消融实验结果

模型	FGD↓	SRGR↑	BeatAlign↑
w/o spatital	157.85	0.180	0.908
w/o temporal	122.57	0.154	0.902
FineGesture	**90.29**	**0.195**	**0.910**

(2)基于VQVAE手势编码模块的消融实验

FineGesture采用两阶段训练的方式实现手势生成，为了验证第一阶段基于VQVAE的手势编码的有效性，对比分析了使用和未使用VQVAE的FineGesture，实验结果见表5所示，其中"w/o VQVAE"表示未使用VQVA模块对手势进行编解码，直接利用FineGesture进行手势生成。"FineGesture"表示文中所提出模型，使用VQVAE模块对手势进行编解码。

表 5: 基于VQVAE手势编码模块的消融实验结果

	FGD↓	SRGR↑	BeatAlign↑
w/o VQVAE	257.34	0.157	0.905
FineGesture	**90.29**	**0.195**	**0.910**

从表5结果可以看出，引入VQVAE模块显著提高了模型的整体性能，在各项指标上均取得了显著提升。此外，可视化结果图4进一步表明了VQVAE模块在FineGesture模型的手势生成中的有效性，未使用VQVAE编解码手势动作时，生成的手势动作抖动严重，动作幅度过大（蓝色实框线），极其不自然，相比之下，加入VQVAE模块后生成的手势动作平滑连贯。这表明VQVAE能提供紧凑的动作表示，每个码本代表一个独特的手势单元，因此随机抖动问题得到有效缓解。

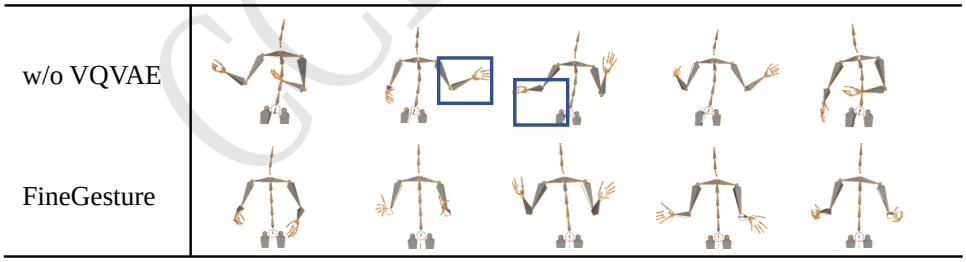

图 4: 有无VQVAE模块生成手势的可视化对比结果

(3)VQVAE码本特征长度的参数分析

码本特征长度大小直接决定了模型的特征提取能力和隐空间表示的容量，进而影响最终手势生成的效果，因此在BEAT数据集上对VQVAE模块中不同码本特征长度参数进行了对比分析，见表6显示。其中"VQVAE-32"、"VQVAE-64"、"VQVAE-128"和"VQVAE-256"分别表示VQVAE模块码本特征长度为32、64、128和256。

从对比结果可以看出，VQVAE-64在生成质量上和语义相关性表现最好。随着码本特征长度的增加，VQVAE-128和VQVAE-256在节奏对齐上有所提升，但在生成质量和语义相关性略有下降，这是因为过大的特征空间会分散对手势整体语义的关注。

表 6: 不同VQVAE码本特征长度对手势生成性能的影响

码本特征长度	FGD↓	SRGR↑	BeatAlign↑
VQVAE-32	108.12	0.157	0.902
VQVAE-64	**90.29**	**0.195**	0.910
VQVAE-128	102.57	0.194	0.911
VQVAE-256	107.34	0.191	**0.919**

5 结论

本文提出了一个基于细粒度时空建模的语音驱动手势生成模型FineGesture。该模型通过时空Transformer架构显式建模了骨架关节间的几何依赖与时序连贯性，并结合AdaIN实现了个性化手势生成，有效解决了现有语音驱动手势生成方法因局部交互建模不足导致的机械感问题。虽然当前方法取得了显著进展，但仍存在实时性不足、多模态信息利用有限等局限性，未来可通过轻量化设计、多模态融合以及跨语言泛化等方向进一步优化，这些改进将推动该技术在虚拟数字人、元宇宙交互、智能娱乐等领域的更广泛应用，为人机自然交互提供更强大的技术支持。

参考文献

Judee K Burgoon, Thomas Birk, and Michael Pfau. 1990. Nonverbal behaviors, persuasion, and credibility. *Human communication research*.

Justine Cassell, Catherine Pelachaud, Norman Badler, Mark Steedman, Brett Achorn, Tripp Becket, Brett Douville, Scott Prevost, and Matthew Stone. 1994. Animated conversation: rule-based generation of facial expression, gesture & spoken intonation for multiple conversational agents. In *Proceedings of the 21st annual conference on Computer graphics and interactive techniques*, pages 413–420.

Justine Cassell, Hannes Högni Vilhjálmsson, and Timothy Bickmore. 2001. Beat: the behavior expression animation toolkit. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 477–486.

Bohong Chen, Yumeng Li, Yao-Xiang Ding, Tianjia Shao, and Kun Zhou. 2024a. Enabling synergistic full-body control in prompt-based co-speech motion generation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, page 6774–6783.

Junming Chen, Yunfei Liu, Jianan Wang, Ailing Zeng, Yu Li, and Qifeng Chen. 2024b. Diffsheg: A diffusion-based approach for real-time speech-driven holistic 3d expression and gesture generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7352–7361.

Yongkang Cheng, Mingjiang Liang, Shaoli Huang, Gaoge Han, Jifeng Ning, and Wei Liu. 2025. Conditional gan for enhancing diffusion models in efficient and authentic global gesture generation from audios. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2164–2173.

Jan P De Ruiter, Adrian Bangerter, and Paula Dings. 2012. The interplay between gesture and speech in the production of referring expressions: Investigating the tradeoff hypothesis. *Topics in cognitive science*.

Shiry Ginosar, Amir Bar, Gefen Kohavi, Caroline Chan, Andrew Owens, and Jitendra Malik. 2019. Learning individual styles of conversational gesture. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3497–3506.

Ikhsanul Habibie, Weipeng Xu, Dushyant Mehta, Lingjie Liu, Hans-Peter Seidel, Gerard Pons-Moll, Mohamed Elgharib, and Christian Theobalt. 2021. Learning speech-driven 3d conversational gestures from video. In *Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents*, pages 101–108.

第二十四届中国计算语言学大会论文集，第684页-第695页，济南，中国，2025年8月11日至14日。

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, pages 6840–6851.

Xun Huang and Serge Belongie. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pages 1501–1510.

Longbin Ji, Pengfei Wei, Yi Ren, Jinglin Liu, Chen Zhang, and Xiang Yin. 2023. C2g2: Controllable co-speech gesture generation with latent diffusion model. *arXiv preprint arXiv:2308.15016*.

Chiyu Jiang, Andre Cornman, Cheolho Park, Benjamin Sapp, Yin Zhou, Dragomir Anguelov, et al. 2023. Motiondiffuser: Controllable multi-agent motion prediction using diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9644–9653.

Stefan Kopp, Brigitte Krenn, Stacy Marsella, Andrew N Marshall, Catherine Pelachaud, Hannes Pirker, Kristinn R Thórisson, and Hannes Vilhjálmsson. 2006. Towards a common framework for multimodal generation: The behavior markup language. In *Intelligent Virtual Agents: 6th International Conference, IVA 2006, Marina Del Rey, CA, USA, August 21-23, 2006. Proceedings 6*, pages 205–217. Springer.

Taras Kucherenko, Patrik Jonell, Sanne Van Waveren, Gustav Eje Henter, Simon Alexandersson, Iolanda Leite, and Hedvig Kjellström. 2020. Gesticulator: A framework for semantically-aware speech-driven gesture generation. In *Proceedings of the 2020 international conference on multimodal interaction*, pages 242–250.

Jing Li, Di Kang, Wenjie Pei, Xuefei Zhe, Ying Zhang, Zhenyu He, and Linchao Bao. 2021a. Audio2gestures: Generating diverse gestures from speech audio with conditional variational autoencoders. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11293–11302.

Ruilong Li, Shan Yang, David A Ross, and Angjoo Kanazawa. 2021b. Ai choreographer: Music conditioned 3d dance generation with aist++. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13401–13412.

Haiyang Liu, Naoya Iwamoto, Zihao Zhu, Zhengqing Li, You Zhou, Elif Bozkurt, and Bo Zheng. 2022a. Disco: Disentangled implicit content and rhythm learning for diverse co-speech gestures synthesis. In *Proceedings of the 30th ACM international conference on multimedia*, pages 3764–3773.

Xian Liu, Qianyi Wu, Hang Zhou, Yinghao Xu, Rui Qian, Xinyi Lin, Xiaowei Zhou, Wayne Wu, Bo Dai, and Bolei Zhou. 2022b. Learning hierarchical cross-modal association for co-speech gesture generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10462–10472.

Haiyang Liu, Zihao Zhu, Giorgio Becherini, Yichen Peng, Mingyang Su, You Zhou, Xuefei Zhe, Naoya Iwamoto, Bo Zheng, and Michael J Black. 2023. Emage:towards unified holistic co-speech gesture generation via expressive masked audio gesture modeling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages arXiv–2401.

Binjie Liu, Lina Liu, Sanyi Zhang, Songen Gu, Yihao Zhi, Tianyi Zhu, Lei Yang, and Long Ye. 2025a. Mag: Multi-modal aligned autoregressive co-speech gesture generation without vector quantization. *Topics in cognitive science*.

Pinxin Liu, Luchuan Song, Junhua Huang, Haiyang Liu, and Chenliang Xu. 2025b. Gesturelsm: Latent shortcut based co-speech gesture generation with spatial-temporal modeling. *arXiv preprint arXiv:2501.18898*.

Alexander Quinn Nichol and Prafulla Dhariwal. 2021. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pages 8162–8171. PMLR.

Simbarashe Nyatsanga, Taras Kucherenko, Chaitanya Ahuja, Gustav Eje Henter, and Michael Neff. 2023. A comprehensive review of data-driven co-speech gesture generation. In *Computer Graphics Forum*, pages 569–596.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. 2019. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10975–10985.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695.

Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. 2022. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*.

Aaron Van Den Oord, Oriol Vinyals, et al. 2017. Neural discrete representation learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6309–6318.

Zunnan Xu, Yukang Lin, Haonan Han, Sicheng Yang, Ronghui Li, Yachao Zhang, and Xiu Li. 2024. Mambatalk: Efficient holistic gesture synthesis with selective state space models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

Sicheng Yang, Zhiyong Wu, Minglei Li, Mengchen Zhao, Jiuxin Lin, Liyang Chen, and Weihong Bao. 2022. The reprgesture entry to the genea challenge 2022. In *Proceedings of the 2022 International Conference on Multimodal Interaction*, pages 758–763.

Sicheng Yang, Zhiyong Wu, Minglei Li, Zhensong Zhang, Lei Hao, Weihong Bao, Ming Cheng, and Long Xiao. 2023a. Diffusestyleggsture: Stylized audio-driven co-speech gesture generation with diffusion models. *arXiv preprint arXiv:2305.04919*.

Sicheng Yang, Zhiyong Wu, Minglei Li, Zhensong Zhang, Lei Hao, Weihong Bao, and Haolin Zhuang. 2023b. Qpgesture: Quantization-based and phase-guided motion matching for natural speech-driven gesture generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2321–2330.

Sicheng Yang, Haiwei Xue, Zhensong Zhang, Minglei Li, Zhiyong Wu, Xiaofei Wu, Songcen Xu, and Zonghong Dai. 2023c. The diffusestylegesture+ entry to the genea challenge 2023. In *Proceedings of the 25th International Conference on Multimodal Interaction*, pages 779–785.

Hongwei Yi, Hualin Liang, Yifei Liu, Qiong Cao, Yandong Wen, Timo Bolkart, Dacheng Tao, and Michael J Black. 2023. Generating holistic 3d human motion from speech. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 469–480.

Lianying Yin, Yijun Wang, Tianyu He, Jinming Liu, Wei Zhao, Bohan Li, Xin Jin, and Jianxin Lin. 2023. Emog: Synthesizing emotive co-speech 3d gesture with diffusion model. *arXiv preprint arXiv:2306.11496*.

Youngwoo Yoon, Bok Cha, Joo-Haeng Lee, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. 2020. Speech gesture generation from the trimodal context of text, audio, and speaker identity. *ACM Transactions on Graphics (TOG)*.

Mingao Zhang, Changhong Liu, Yong Chen, Zhenchun Lei, and Mingwen Wang. 2022. Music-to-dance generation with multiple conformer. In *Proceedings of the 2022 International Conference on Multimedia Retrieval*, pages 34–38.

Fan Zhang, Naye Ji, Fuxing Gao, and Yongping Li. 2023. Diffmotion: Speech-driven gesture synthesis using denoising diffusion model. In *Proceedings of the International Conference on Multimedia Modeling*, pages 231–242.

Lingting Zhu, Xian Liu, Xuanyu Liu, Rui Qian, Ziwei Liu, and Lequan Yu. 2023. Taming diffusion models for audio-driven co-speech gesture generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10544–10553.