个性化语音驱动的三维面部动画生成方法

李　伟[1,2], 宋玉璞[1,2], 刘亚志[1,2], 安　逸[1,2]

(1.华北理工大学 人工智能学院,河北 唐山 063210;2.华北理工大学 河北省工业智能感知重点实验室,河北 唐山 063210)

摘　要: 为解决语音驱动的三维面部动画生成中语音与动作对齐困难、身份特征易丢失以及个性化动态表现不足的问题,提出一种基于条件扩散模型的生成方法。该方法设计双路风格编码结构,分别提取层次化的身份特征与动态运动特征,并通过双向注意力机制实现语音特征与加噪运动特征的深度融合。在此基础上,引入风格条件引导的改进 Transformer 解码器,以合成高质量运动序列。在 BIWI、VOCASET 和 3DMEAD 数据集上的实验结果表明,所提方法在平均顶点误差(*MVE*)、唇部顶点误差(*LVE*)和面部动态偏差(*FDD*)指标上均取得最优性能。与对应指标上的最佳基线方法相比,在 BIWI 上的 *MVE*、*LVE* 和 *FDD* 分别降低 4.8%、15.4%和 13.4%;在 VOCASET 上的 *LVE* 降低 14.9%;在 3DMEAD 上的 *MVE* 和 *FDD* 分别降低 10.2%和 13.7%。主观评测结果进一步验证了所提方法在视觉自然度与真实感方面的优势。所提方法为三维面部动画的高保真生成、身份保持与个性化建模提供了新的技术路径。

关键词: 语音驱动动画; 三维面部动画; 深度学习; 扩散模型; 个性化

中图分类号: TP391.41; TN912.3　　**文献标志码:** A　　**doi:** 10.13705/j.issn.1671-6833.2026.04.023

语音驱动的三维面部动画技术旨在根据输入语音合成高质量的面部运动,在影视特效、虚拟现实与游戏特效等领域应用广泛。该任务的核心挑战源于语音与面部运动之间复杂的非线性映射,即相同的语音可能对应多种合理的运动模式,且不同个体的发音口型与表情亦存在显著差异。因此,生成高质量动画需要兼顾口型同步、表情自然度与风格个性化。围绕这一挑战,早期研究从时序卷积增强声面部关联[1]等角度展开,但生成的运动多集中于嘴部。为生成更丰富的全脸动画,后续研究引入 Transformer 以建模长程音频依赖[2],并借助 VQ-VAE 学习离散运动编码[3]。尽管模型架构不断演进,但大多数方法仍以确定性生成为主,难以充分刻画语音与面部运动间固有的随机性与多样性,从而限制了生成动画在自然度与真实感上的提升。相较之下,扩散模型[4-5]通过前向加噪与反向去噪的迭代学习机制,在图像生成领域已展现出兼顾生成质量与多样性的能力[6-7]。在三维运动合成领域,Tevet 等[8]率先提出基于文本条件的运动扩散模型,实现了多样化的人体动作生成。近年来,该技术范式进一步延伸至三维面部动画生成,其中 Stan 等[9]首次将扩散模型应用于该任务,Lin 等[10]则在图量化潜空间中引入扩散模型,有效提升了语音与面部网格之间的跨模态对齐质量。

尽管上述方法在动画生成质量上不断提升,但在捕捉个体特有的说话风格与动态运动特征方面仍存在局限。为此,本文提出一种基于条件扩散模型的个性化三维面部动画生成方法。该方法设计了双路风格编码结构,一方面通过静态身份编码器提取层次化的静态身份特征;另一方面通过动态风格编码器从参考运动中提取动态运动特征。该生成模型以音频特征与上述两类特征作为多模态条件,借助扩散模型在概率分布建模与多样性生成方面的优势,有效处理语音到运动之间复杂的“一对多”映射

收稿日期: 2026-03-17;**修订日期:** 2026-04-12

基金项目: 河北省高等学校科学技术研究项目(ZD2022102)

作者简介: 李伟(1979—),男,河北平山人,华北理工大学高级工程师,主要从事图像与计算机视觉、计算机网络技术研究,E-mail:lw@ncst.edu.cn。

通信作者: 安逸(1982—),女,河北唐山人,华北理工大学高级工程师,主要从事检测与控制技术及智能装置、计算机视觉研究,E-mail:beyond@ncst.edu.cn。

关系。本文主要贡献如下。

(1)提出了一种基于双路风格编码的语音驱动的面部动画生成方法。该方法通过静态身份编码器提取层次化的静态身份特征,结合动态风格编码器方法提取动态运动特征,实现了身份一致性与运动自然度的有效平衡。

(2)设计了融合多尺度时序建模的动态风格编码器。通过时序卷积网络、双向门控循环单元与Transformer的级联架构,实现了对动态运动特征的细粒度提取与表征。

(3)构建了双向多模态特征融合的扩散生成模型。通过音频与运动特征的双向交叉注意力机制,实现了跨模态特征的深度对齐;结合双路风格条件与时间步编码,利用扩散模型处理一对多映射问题,生成高质量个性化面部动画。

1 本文方法

本文提出一种条件基于扩散模型的语音驱动的三维面部动画生成方法,如图1所示。该方法以音频特征与风格条件为引导,通过迭代去噪重建高质量三维顶点运动。

1.1 问题定义与扩散模型框架

设一段长度为 T 的原始面部运动序列为 $\boldsymbol{M}^0=(\boldsymbol{m}_1^0,\boldsymbol{m}_2^0,\ldots,\boldsymbol{m}_T^0)$,其中第 t 帧 $\boldsymbol{m}_t^0\in\mathbf{R}^{3V}$ 表示 V 个顶点相对于中性面部网格模板 $\boldsymbol{f}\in\mathbf{R}^{3V}$ 在三维空间中的位移量。相应的语音片段记为 $\boldsymbol{A}=(\boldsymbol{a}_1,\boldsymbol{a}_2,\ldots,\boldsymbol{a}_T)$,其中 $\boldsymbol{a}_t\in\mathbf{R}^D$ 为包含 D 个样本点的音频帧。本文的目标是在给定语音片段 $\boldsymbol{A}$ 的条件下,合成与之同步的面部运动序列。在获得预测的运动序列后,将其与中性面部网格模板 $\boldsymbol{f}$ 逐帧叠加,即可得到最终的三维面部动画序列 $\boldsymbol{O}=(\hat{\boldsymbol{m}}_1^0+\boldsymbol{f},\hat{\boldsymbol{m}}_2^0+\boldsymbol{f},\ldots,\hat{\boldsymbol{m}}_T^0+\boldsymbol{f})$。

为实现上述目标,本文采用扩散模型进行建模。在训练阶段,通过预设噪声调度对原始面部运动序列 $\boldsymbol{M}^0$ 进行前向加噪,时间步索引记为 $n=1,\ldots,N$,使其逐步演变为近似高斯分布[5]:

$$q(\boldsymbol{M}^n\mid\boldsymbol{M}^{n-1})=\mathcal{N}(\sqrt{1-\alpha_n}\boldsymbol{M}^{n-1},\alpha_n\boldsymbol{I})。\quad(1)$$

式中:$\boldsymbol{M}^n$ 为经过 n 步扩散后的运动序列;$\alpha_n\in(0,1)$ 为预设的噪声方差,控制第 n 步引入的噪声强度;$\boldsymbol{I}$ 为单位矩阵。反向去噪过程则通过端到端的去噪网络从 $\boldsymbol{M}^n$ 中恢复 $\boldsymbol{M}^0$。在预测目标上,本文选择直接回归去噪后的运动序列[11],以减少因间接预测噪声带来的误差累积[5],从而提升生成质量。

1.2 音频编码器

音频编码器采用预训练的HuBERT模型[12]将原始语音片段 $\boldsymbol{A}$ 编码为驱动面部动画的音频特征 $\boldsymbol{F}_{\mathrm{a}}$。处理流程包括3个环节:首先、通过时序卷积网络提取局部声学特征;其次、利用Transformer编码器构建长程上下文依赖;最后、通过线性插值实现与动画序列的帧率对齐。在参数优化上,本文冻结HuBERT底层参数以保留通用声学表征能力,仅对上层Transformer及投影层进行微调,使其适配语音到面部的映射任务。

1.3 风格编码器

1.3.1 动态风格编码器

动态风格编码器旨在从输入的参考面部运动序列中提取动态运动特征。输入运动序列经投影层映射至潜在表示 $\boldsymbol{X}\in\mathbf{R}^{B\cdot T\cdot C}$,随后由多尺度时序卷积

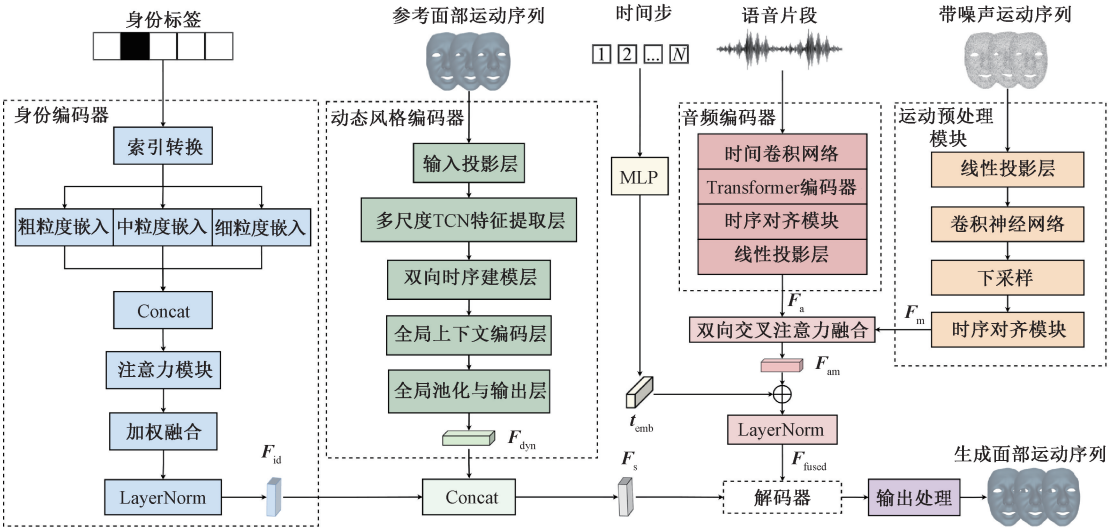

图1 本文网络总体框架图

Figure 1 Overall framework of the proposed network

网络(temporal convolutional network, TCN)提取局部时空特征[13]。该模块采用残差连接与膨胀卷积结构,并行使用5个膨胀率 $d \in \{1,2,4,8,16\}$ 的卷积分支,以同步捕获从肌肉微运动到完整表情动态的多样化面部运动模式。将膨胀率为 d 的TCN块记为函数 $T_d(\cdot)$,即 $\boldsymbol{F}_{\mathrm{TCN}}^{\mathrm{d}} = T_d(\boldsymbol{X})$。各分支输出的多尺度特征通过均值池化进行融合:

$$\boldsymbol{F}_{\mathrm{TCN}} = \frac{1}{5}\sum_{d\in\{1,2,4,8,16\}} \boldsymbol{F}_{\mathrm{TCN}}^{d}。\tag{2}$$

为进一步提取时序依赖特征,编码器在TCN之后引入双向门控循环单元(bidirectional gated recurrent unit, Bi-GRU)[14]。该网络沿时间轴前向与后向处理序列,融合每一时刻的上下文信息。对于第 t 帧,其前向与后向隐藏状态的更新过程可分别表示为

$$\overrightarrow{\boldsymbol{h}}_t = \mathrm{GRU}(\boldsymbol{F}_{\mathrm{TCN},t}, \overrightarrow{\boldsymbol{h}}_{t-1});\tag{3}$$

$$\overleftarrow{\boldsymbol{h}}_t = \mathrm{GRU}(\boldsymbol{F}_{\mathrm{TCN},t}, \overleftarrow{\boldsymbol{h}}_{t+1})。\tag{4}$$

式中:$\boldsymbol{F}_{\mathrm{TCN},t}$ 为 t 时刻的多尺度TCN特征;$\overrightarrow{\boldsymbol{h}}_{t-1}$ 和 $\overleftarrow{\boldsymbol{h}}_{t+1}$ 分别为前向和后向的相邻隐藏状态。所有帧位置上的双向隐藏状态拼接形成完整的序列级特征 $\boldsymbol{F}_{\mathrm{GRU}}$:

$$\boldsymbol{F}_{\mathrm{GRU}} = [\overrightarrow{\boldsymbol{h}}_1;\overleftarrow{\boldsymbol{h}}_1;\ldots;\overrightarrow{\boldsymbol{h}}_T;\overleftarrow{\boldsymbol{h}}_T]。\tag{5}$$

式中:T 为序列长度。得益于Bi-GRU的双向结构,该特征能够同时融合历史与未来上下文信息[15]。

将Bi-GRU输出的特征 $\boldsymbol{F}_{\mathrm{GRU}}$ 与可学习位置编码 $\boldsymbol{P}$ 逐元素相加后,输入至Transformer编码器。令 $\mathrm{TransEnc}(\cdot)$ 表示由 L 层多头自注意力与前馈子层按Pre-Norm方式堆叠而成的编码器,本文中 $L=2$,则上下文感知特征 $\boldsymbol{F}_{\mathrm{attn}}$ 计算如下所示:

$$\boldsymbol{F}_{\mathrm{attn}} = \mathrm{TransEnc}(\boldsymbol{F}_{\mathrm{GRU}} + \boldsymbol{P})。\tag{6}$$

为将整个序列的时序信息聚合为一个固定维度的风格表示,采用一种可学习的注意力池化机制:

$$\gamma_t = \mathrm{Softmax}(\boldsymbol{w}_a^{\mathrm{T}}\tanh(\boldsymbol{W}_1\boldsymbol{F}_{\mathrm{attn}} + \boldsymbol{b}_1));\tag{7}$$

$$\boldsymbol{s} = \sum_{t=1}^{T}\gamma_t\boldsymbol{F}_{\mathrm{attn}}。\tag{8}$$

式中:$\boldsymbol{w}_a$ 为一个可学习的注意力投影向量,其转置 $\boldsymbol{w}_a^{\mathrm{T}}$ 与 $\tanh(\cdot)$ 的输出进行内积运算;$\boldsymbol{W}_1$ 和 $\boldsymbol{b}_1$ 为可学习参数;γ_t 为第 t 帧的注意力权重;$\boldsymbol{s}$ 为加权求和后的全局风格向量。最后,将 $\boldsymbol{s}$ 输入多层感知机进行非线性映射得到最终风格嵌入 $\boldsymbol{F}_{\mathrm{dyn}}$:

$$\boldsymbol{F}_{\mathrm{dyn}} = \mathrm{MLP}(\boldsymbol{s})。\tag{9}$$

1.3.2 静态身份编码器

静态身份编码器从one-hot身份标签中提取个体固有特征。为增强细节建模能力,借鉴隐空间尺度分组的思想[16],本文设计分层嵌入机制,将身份信息分解为粗、中、细3个抽象层次。将one-hot身份标签通过argmax操作转换为身份索引 i。以 i 为索引,从3个独立的嵌入矩阵 $\boldsymbol{E}_{\mathrm{c}},\boldsymbol{E}_{\mathrm{m}},\boldsymbol{E}_{\mathrm{f}}$ 中分别获取对应粗、中、细粒度的嵌入向量 $\boldsymbol{e}_{\mathrm{c}},\boldsymbol{e}_{\mathrm{m}},\boldsymbol{e}_{\mathrm{f}}$。这些嵌入矩阵在训练过程中通过反向传播优化,最终收敛到能够有效区分不同身份的特征表示。获得3个粒度的嵌入向量 $\boldsymbol{e}_{\mathrm{c}},\boldsymbol{e}_{\mathrm{m}},\boldsymbol{e}_{\mathrm{f}}$ 后,将其在通道维度拼接为组合向量 $\boldsymbol{e}_{\mathrm{comb}} = [\boldsymbol{e}_{\mathrm{c}};\boldsymbol{e}_{\mathrm{m}};\boldsymbol{e}_{\mathrm{f}}]$,通过一个轻量级的注意力模块计算各粒度的权重系数:

$$[\beta_{\mathrm{c}},\beta_{\mathrm{m}},\beta_{\mathrm{f}}] = \mathrm{Softmax}(\boldsymbol{W}_2\boldsymbol{e}_{\mathrm{comb}} + \boldsymbol{b}_2)。\tag{10}$$

式中:$\boldsymbol{W}_2$ 和 $\boldsymbol{b}_2$ 均为可学习参数;β_{c}、β_{m} 和 β_{f} 分别为粗、中、细粒度嵌入向量的注意力权重,满足 $\beta_{\mathrm{c}}+\beta_{\mathrm{m}}+\beta_{\mathrm{f}}=1$。对各粒度嵌入向量进行加权融合,得到最终身份嵌入:

$$\boldsymbol{F}_{\mathrm{id}} = \mathrm{LN}([\beta_{\mathrm{c}}\boldsymbol{e}_{\mathrm{c}};\beta_{\mathrm{m}}\boldsymbol{e}_{\mathrm{m}};\beta_{\mathrm{f}}\boldsymbol{e}_{\mathrm{f}}])。\tag{11}$$

式中:$\mathrm{LN}(\cdot)$ 为层归一化。这种分层加权的设计不仅能够增强身份特征的判别性,还能够提高模型对不同身份特征分布的适应性,为后续的多模态融合提供更具表达力的身份条件信息。

1.4 基于条件扩散的动画生成

为建立语音与运动模态间的细粒度对齐,本文设计了一种双向交叉注意力机制,其结构如图2所示。在扩散去噪过程中,音频特征 $\boldsymbol{F}_{\mathrm{a}}$ 与当前步的带噪运动特征 $\boldsymbol{F}_{\mathrm{m}}$ 进行双向交互。一方面,以音频特征 $\boldsymbol{F}_{\mathrm{a}}$ 为查询、以运动特征 $\boldsymbol{F}_{\mathrm{m}}$ 为键和值,从运动模态中提取与语音相关的上下文信息,以引导去噪生成;另一方面,以运动特征 $\boldsymbol{F}_{\mathrm{m}}$ 为查询,以增强后的音频特征 $\boldsymbol{F}'_{\mathrm{a}}$ 为键和值,进一步调制运动表示,从而实现两种模态的深度融合。将双向交互后增强的两路特征相加,得到融合特征 $\boldsymbol{F}_{\mathrm{am}}$:

$$\boldsymbol{F}_{\mathrm{am}} = \boldsymbol{F}'_{\mathrm{a}} + \boldsymbol{F}'_{\mathrm{m}}。\tag{12}$$

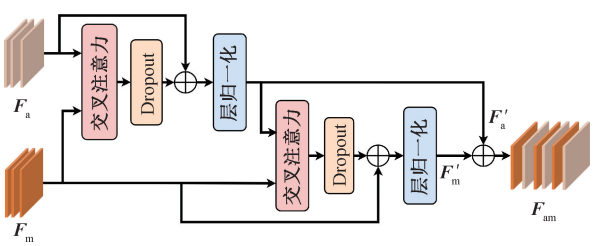

图2 双向交叉注意力机制

Figure 2 Bidirectional cross-attention mechanism

如图1总体框架所示,上述融合特征随后与当前时间步嵌入 $\boldsymbol{t}_{\mathrm{emb}}$ 相加,经层归一化后得到解码器的融合输入 $\boldsymbol{F}_{\mathrm{fused}}$:

$$\boldsymbol{F}_{\mathrm{fused}} = \mathrm{LayerNorm}(\boldsymbol{F}_{\mathrm{am}} + \boldsymbol{t}_{\mathrm{emb}})。\tag{13}$$

将静态身份编码器提取的身份嵌入 $\boldsymbol{F}_{\mathrm{id}}$ 与动态

风格编码器提取的风格嵌入 $\boldsymbol{F}_{\text{dyn}}$ 沿特征维度拼接,得到统一的风格条件向量 $\boldsymbol{F}_{\text{s}}$:

$$\boldsymbol{F}_{\text{s}} = \text{Concat}(\boldsymbol{F}_{\text{id}}, \boldsymbol{F}_{\text{dyn}})。 \tag{14}$$

解码器采用基于 Transformer 的模块化设计,其核心结构如图 3 所示,包含时序自注意力模块与风格条件跨注意力模块。时序自注意力模块引入旋转位置编码[17],以有效刻画序列内部的长期依赖关系。风格条件跨注意力模块则以风格条件向量 $\boldsymbol{F}_{\text{s}}$ 作为条件,实现特征的风格调制。解码器的输出 $\boldsymbol{F}_{\text{dec}}$ 经线性投影映射至顶点空间,与中性面部运动模板叠加得到最终的三维动画序列。

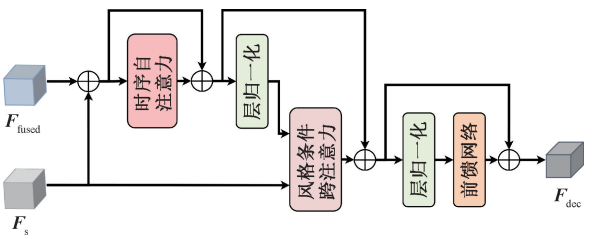

图 3　解码器网络结构

Figure 3　Decoder network architecture

1.5　训练与推理

本文采用多任务损失函数综合优化模型性能,总损失函数由扩散重建损失、速度一致性损失、加速度一致性损失和面部动态损失 4 部分构成:

$$L_{\text{total}} = L_{\text{recon}} + \lambda_{\text{vel}} L_{\text{vel}} + \lambda_{\text{acc}} L_{\text{acc}} + \lambda_{\text{face}} L_{\text{face}}。 \tag{15}$$

式中:λ_{vel}、λ_{acc}、λ_{face} 表示对应损失函数的权重系数。

扩散重建损失使用 Huber 损失[8]进行计算,以衡量模型从带噪数据中恢复原始面部运动序列的能力:

$$L_{\text{recon}} = \| \boldsymbol{M}^0 - \hat{\boldsymbol{M}}^0 \|_{\text{H}}。 \tag{16}$$

速度一致性损失用于约束生成运动序列的时序平滑性,通过最小化相邻帧间顶点速度的预测误差来实现[1]:

$$L_{\text{vel}} = \| (\boldsymbol{m}^0_{t+1} - \boldsymbol{m}^0_t) - (\hat{\boldsymbol{m}}^0_{t+1} - \hat{\boldsymbol{m}}^0_t) \|_{\text{H}}。 \tag{17}$$

加速度一致性损失进一步约束运动的速度变化率,以确保生成的运动具有自然的动力学特性[18]。当序列长度 $T<3$ 时,该损失自动置零。

$$L_{\text{acc}} = \| (\boldsymbol{m}_{t+2} - 2\boldsymbol{m}_{t+1} + \boldsymbol{m}_t) - (\hat{\boldsymbol{m}}_{t+2} - 2\hat{\boldsymbol{m}}_{t+1} + \hat{\boldsymbol{m}}_t) \|_{\text{H}}。 \tag{18}$$

面部动态损失在上面部区域内计算预测与真实顶点位移幅度标准差的平均绝对误差,以约束生成动画的动态自然度:

$$L_{\text{face}} = E\left(\frac{1}{|R|} \sum_R | \sigma - \hat{\sigma} | \right)。 \tag{19}$$

式中:σ 和 $\hat{\sigma}$ 分别为真实与预测运动序列在时间维度上的位移幅度标准差;R 为上面部区域的顶点集合,该区域划分遵循 CodeTalker[3]的设置。

基于上述损失函数,本文的训练与推理流程如算法 1 和算法 2 所示,其中反向去噪过程参照 Tevet 等[8]的设计。

算法 1　训练流程

输入:语音片段 $\boldsymbol{A}$、原始面部运动序列 $\boldsymbol{M}^0$、参考运动序列 $\boldsymbol{M}_{\text{ref}}$、身份标签 $\boldsymbol{C}$、中性面部运动模板 $\boldsymbol{f}$,总时间步数 N、总轮数 E;

输出:优化后的模型参数 θ。

① for $e=1$ to E do
② 　从训练集中采样一个批次 $(\boldsymbol{A}, \boldsymbol{M}^0, \boldsymbol{M}_{\text{ref}}, \boldsymbol{C}, \boldsymbol{f})$;
③ 　从 1 到 N 中均匀随机采样一个时间步索引 n,并根据式(1)对 $\boldsymbol{M}^0$ 进行前向加噪,得到 $\boldsymbol{M}^n$;
④ 　从 $\boldsymbol{A}$、$\boldsymbol{M}_{\text{ref}}$、$\boldsymbol{C}$ 中提取条件特征 $\boldsymbol{F}_{\text{a}}$、$\boldsymbol{F}_{\text{dyn}}$、$\boldsymbol{F}_{\text{id}}$;
⑤ 　将 $\boldsymbol{M}^n$ 输入运动预处理模块,得到 $\boldsymbol{F}_{\text{m}}$;
⑥ 　将 $\boldsymbol{F}_{\text{a}}$、$\boldsymbol{F}_{\text{m}}$、$\boldsymbol{F}_{\text{dyn}}$、$\boldsymbol{F}_{\text{id}}$、$\boldsymbol{f}$、$n$ 输入生成网络,得到预测 $\hat{\boldsymbol{M}}^0$;
⑦ 　按式(15)计算总损失,并沿梯度更新参数 θ;
⑧ end for
⑨ return θ

算法 2　推理流程

输入:语音片段 $\boldsymbol{A}$、参考面部运动序列 $\boldsymbol{M}_{\text{ref}}$、身份标签 $\boldsymbol{C}$、中性面部运动模板 $\boldsymbol{f}$、总时间步数 N;

输出:生成的运动序列 $\hat{\boldsymbol{M}}^0$。

① $\boldsymbol{M}^N \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$;
② 从 $\boldsymbol{A}$、$\boldsymbol{M}_{\text{ref}}$、$\boldsymbol{C}$ 中提取条件特征 $\boldsymbol{F}_{\text{a}}$、$\boldsymbol{F}_{\text{dyn}}$、$\boldsymbol{F}_{\text{id}}$;
③ for $n = N$ downto 1 do
④ 　将 $\boldsymbol{M}^n$ 输入运动预处理模块,得到 $\boldsymbol{F}_{\text{m}}$;
⑤ 　将 $\boldsymbol{F}_{\text{a}}$、$\boldsymbol{F}_{\text{m}}$、$\boldsymbol{F}_{\text{dyn}}$、$\boldsymbol{F}_{\text{id}}$、$\boldsymbol{f}$、$n$ 输入模型,得到预测 $\hat{\boldsymbol{M}}^0$;
⑥ 　若 $n>1$ 则随机采样噪声 $\boldsymbol{z}$,否则 $\boldsymbol{z}=0$;
⑦ 　由 $\boldsymbol{M}^n$ 和 $\hat{\boldsymbol{M}}^0$ 根据反向去噪过程计算 $\boldsymbol{M}^{n-1}$;
⑧ end for
⑨ return $\hat{\boldsymbol{M}}^0$

2　实验

2.1　数据集

为全面评估方法性能,本文选用 BIWI[19]、VOCASET[1]与 3DMEAD[20]数据集进行实验。BIWI 数据集包含 14 名说话者,每名说话者朗读 40 条英文句子,每条句子均以中性与情感化两种方式录制。3D 面部几何以 25 帧/s 捕获,每帧含 23 370 个顶点。实验采用其中的情感表达子集,数据划分遵循 CodeTalker[3]的设置。VOCASET 数据集包含 12 名

受试者共480个4D扫描序列,总时长约29 min,采样率为60帧/s。所有网格均配准到FLAME拓扑,顶点数为5 023。为确保对比的公平性,本文沿用CodeTalker[3]的数据划分方式,将数据集划分为训练集、验证集和测试集。3DMEAD数据集是在二维MEAD情感视频数据集[21]的基础上,利用DECA[22]和MICA[23]对原始视频序列逐帧进行三维重建得到,包含连续的三维面部几何与表情参数,覆盖多种面部身份以及从中性到强烈情感状态的变化。

2.2 实验设置

所有实验均在一台配置为Intel Xeon Silver 4214R CPU(2.40 GHz,24核)与NVIDIA GeForce RTX 3090 GPU(24 GB显存)的服务器上完成。操作系统采用Ubuntu 20.04,基于Python 3.8环境,使用PyTorch 2.7.0(CUDA 12.6)进行模型训练与测试。数据预处理阶段,所有音频重采样至16 kHz,并使用预训练的HuBERT模型提取音频特征。将VOCASET数据集的运动序列帧率从60帧/s降采样至30帧/s;BIWI与3DMEAD数据集的帧率设置为25帧/s。本文方法共训练50个epoch,批量大小设为1,采用AdamW优化器,初始学习率为1×10^{-4},权重衰减系数为0.05。扩散步数设为1 000,噪声调度采用余弦调度。损失函数的权重系数$\lambda_{vel}=0.003$,$\lambda_{acc}=0.003$,$\lambda_{face}=0.000\ 5$。

2.3 评估指标

本文从顶点精度与运动动态两个维度评估生成质量,采用三维面部动画领域广泛使用的3个指标。平均顶点误差(*MVE*)[9]衡量整体面部几何误差,对每一帧全脸顶点计算L2误差,并取该帧的最大值,最后在序列维度上求平均,以评估模型保持全局面部几何一致性的能力。唇部顶点误差(*LVE*)[2]衡量唇部区域的口型同步误差。对每一帧唇部顶点计算L2误差,并取该帧的最大值,最后在序列维度上求平均。面部动态偏差(*FDD*)[3]衡量面部动态真实性。该指标比较生成序列与真实序列在上面部区域运动统计量上的差异,用于反映生成动作与真实动作分布的一致程度。

2.4 定量评估

本节通过定量指标对本文方法与现有方法进行对比评估。如表1所示,在BIWI数据集上,本文方法在各项指标上均取得最优结果。其中,*MVE*较最佳基线方法FaceDiffuser降低4.8%,表明本文方法全局面部重建更为准确;*LVE*较FaceDiffuser降低15.4%,说明本文方法唇部同步精度更高;*FDD*较该指标上的最优基线SelfTalk降低13.4%,表明本文方法生成序列的动态模式更接近真实分布。

表1 BIWI数据集的客观评估结果

Table 1 Objective evaluation results on BIWI dataset

方法	*MVE*/10^{-3} mm	*LVE*/10^{-4} mm	*FDD*/10^{-5} mm
VOCA[1]	8.360 6	6.715 5	7.532 0
FaceFormer[2]	7.176 7	5.178 0	5.029 6
CodeTalker[3]	7.398 0	4.791 4	4.117 0
FaceDiffuser[9]	6.845 5	4.501 6	3.857 0
SelfTalk[24]	7.130 3	4.582 4	3.594 5
TalkingStyle[25]	7.128 4	4.687 6	4.416 6
本文方法	**6.519 4**	**3.808 8**	**3.111 5**

为验证方法的泛化能力,本文在3DMEAD和VOCASET数据集上进一步进行了评估,结果如表2所示。在3DMEAD上,本文方法同样在3项指标上均取得最优结果,其中*MVE*、*LVE*和*FDD*分别较FaceDiffuser降低10.2%、5.8%和13.7%,反映出生成结果在全局面部几何、唇部同步及上面部动态方面均更接近真实数据。对于主要关注口型运动的VOCASET数据集,仅采用*LVE*进行评估,本文方法较FaceDiffuser降低14.9%,进一步验证了其在不同数据分布下的唇形同步能力与泛化性能。

表2 3DMEAD与VOCASET数据集的客观评估结果

Table 2 Objective evaluation results on 3DMEAD and VOCASET datasets

方法	3DMEAD			VOCASET
	MVE/10^{-3} mm	*LVE*/10^{-4} mm	*FDD*/10^{-7} mm	*LVE*/10^{-5} mm
CodeTalker[3]	2.015 4	3.035 6	18.040	6.848 6
FaceDiffuser[9]	1.759 6	1.679 6	7.188 2	6.793 5
本文方法	**1.579 5**	**1.581 6**	**6.205 2**	**5.779 0**

2.5 定性评估

为直观评估本文方法在唇形同步与面部动态表达上的性能,本文在VOCASET、BIWI以及3DMEAD数据集上选取了多个具有代表性的发音帧进行可视化对比。如图4所示,与CodeTalker和FaceDiffuser相比,本文方法在开口元音、双元音以及辅音起始音节等多种发音状态下生成的唇形均与真实唇形更为接近。

在涵盖多种情感状态的3DMEAD数据集上,本文方法亦表现出良好的适应能力。如图5所示,在发音“high”时,本文方法能够生成下颌明显下沉、口型充分张开的姿态;而在发音“make”中的/m/时,则能够还原出明显的双唇闭合特征。同时,本文方

法在悲伤、惊讶等不同情感状态下均能生成符合情感语义的面部表情。视觉对比结果表明,本文方法在多个数据集、多种情感条件下均能够生成更加准确且自然的唇部运动与面部表情。

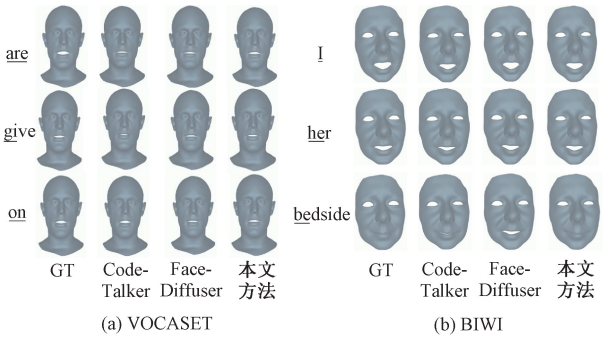

图 4　VOCASET 与 BIWI 数据集的定性对比

Figure 4　Qualitative comparison on VOCASET and BIWI dataset

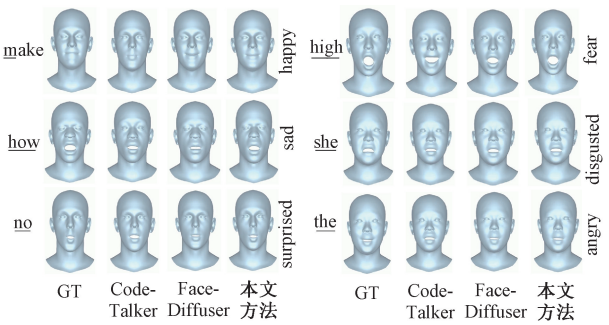

图 5　3DMEAD 数据集的定性对比

Figure 5　Qualitative comparison on 3DMEAD dataset

为评估所生成面部运动的动态多样性与真实性,本文基于 VOCASET 和 BIWI 数据集,通过计算各顶点在整个序列上的均值和标准差,对面部运动动力学进行了热图可视化分析,结果如图 6 所示。图 6 中,均值反映运动的整体强度,标准差刻画运动在时序上的起伏变化,二者均以暖色调表示较高数值,分别对应区域运动的剧烈程度与动态变化的丰富性。与其他方法相比,本文方法生成的面部运动在整体运动幅度及其局部时序波动模式上,均与真实数据的分布更为接近。

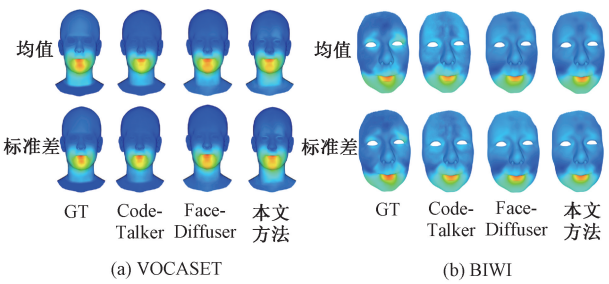

图 6　不同数据集的热图对比

Figure 6　Heatmap comparison on different datasets

2.6　消融实验

为评估所提出方法中各个组件的作用,本文在 BIWI 数据集上进行了 3 组消融实验,结果如表 3 所示。除任一模块均会导致生成性能不同程度的下降。动态风格编码器的移除使 *MVE*、*LVE* 和 *FDD* 较完整模型分别上升 12.3%、24.3%与 27.6%,表明该模块提取的细粒度动态特征是生成自然生动动画的关键。

表 3　消融实验结果

Table 3　Ablation experiment results

方法	*MVE*/ 10^{-3} mm	*LVE*/ 10^{-4} mm	*FDD*/ 10^{-5} mm
移除动态风格编码器	7.323 5	4.735 5	3.971 6
移除静态身份编码器	6.560 0	4.130 0	4.237 9
移除双向交叉注意力	7.065 0	5.172 3	5.123 7
本文方法	**6.519 4**	**3.808 8**	**3.111 5**

静态身份编码器的移除主要影响模型对身份相关几何细节的保真能力。*LVE* 相较完整模型上升 8.4%,说明在缺乏明确身份先验时,模型对唇部等身份敏感区域的还原精度会下降,进一步印证了分层身份编码在维持生成结果身份一致性方面的重要作用。

在针对双向交叉注意力机制的消融实验中,移除该模块后,模型仅将音频特征与带噪声的运动特征及时间步嵌入直接相加,导致 *LVE* 和 *FDD* 分别上升 35.8%和 64.7%。结果表明,该机制对于建立语音与面部运动的细粒度对齐、生成符合运动规律的面部表情至关重要。缺失该机制后,语音信息难以有效调制运动生成过程。

2.7　用户实验

为从主观感知层面评估生成动画的质量,本文在 BIWI 与 3DMEAD 数据集上开展了用户实验。实验采用两两强制选择的偏好测试方法,并以在线问卷形式进行。参与者观看由同一语音驱动、不同方法生成的三维面部动画视频对,并从中选择感知上更自然的结果。实验样本随机呈现,且隐藏方法标识,以尽量减少主观偏差并保证实验的客观性。

最终共回收 120 份有效问卷,统计了每个对比组中本文方法被选中的比例。如表 4 所示,本文方法在主观评估中优于现有方法。与真实样本相比,本文方法在 BIWI 和 3DMEAD 上仍分别获得了 43.33%和 45.83%的选择率,表明其生成结果在视觉上已接近真实样本。上述结果从主观感知层面验证了本文方法在提升动画自然度与视觉逼真度方面的有效性。

表 4　用户实验结果

Table 4　User study results

对比组	本文方法被选中的比例/%	
	BIWI	3DMEAD
本文方法和 CodeTalker	74.17	75.83
本文方法和 FaceDiffuser	65.83	67.50
本文方法和 Ground Truth	43.33	45.83

3　结论

本文构建了一种基于条件扩散模型的语音驱动的三维面部动画生成方法。该方法采用双路风格编码结构,其中分层静态身份编码器用于提取身份特征,多尺度动态风格编码器用于从参考运动序列中捕获时序动态模式,并引入双向交叉注意力机制,实现音频与运动特征的深度交互与细粒度对齐。在此基础上,通过扩散模型有效建模语音到运动的复杂映射,并结合改进的 Transformer 解码器生成高质量面部运动序列。实验结果表明,本文方法在多个数据集和评估指标上均取得了优越性能。未来工作将进一步提升高分辨率顶点序列的生成能力,探索更丰富的情感与风格控制机制,并优化推理效率,以满足实时应用需求。

参考文献:

[1] Cudeiro D, Bolkart T, Laidlaw C, et al. Capture, learning, and synthesis of 3D speaking styles[C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2019: 10093-10103.

[2] Fan Yingruo, Lin Zhaojiang, Saito J, et al. FaceFormer: speech-driven 3D facial animation with Transformers[C]//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 18749-18758.

[3] Xing Jinbo, Xia Menghan, Zhang Yuechen, et al. CodeTalker: speech-driven 3D facial animation with discrete motion prior[C]//Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2023: 12780-12790.

[4] Sohl-Dickstein J, Weiss E A, Maheswaranathan N, et al. Deep unsupervised learning using nonequilibrium thermodynamics[C]//Proceedings of the 32nd International Conference on Machine Learning. New York: ACM, 2015: 2256-2265.

[5] Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models[C]// 34th Conference on Neural Information Processing Systems. Vancouver: NeurIPS, 2020: 6840-6851.

[6] Bigioi D, Basak S, Stypułkowski M, et al. Speech driven video editing via an audio-conditioned diffusion model[J]. Image and Vision Computing, 2024, 142: 104911.

[7] Tang Ying, Liu Yazhi, Li Xiong, et al. Adaptive diffusion landmark dynamic rendering for realistic talking face video generation[J]. The Visual Computer, 2025, 41(11): 8935-8945.

[8] Tevet G, Raab S, Gordon B, et al. Human motion diffusion model[PP/OL]. V2. arXiv (2022-10-03)[2026-03-01]. https://arxiv.org/abs/2209.14916.

[9] Stan S, Haque K I, Yumak Z. FaceDiffuser: speech-driven 3D facial animation synthesis using diffusion[C]// Proceedings of the 16th ACM SIGGRAPH Conference on Motion, Interaction and Games. New York: ACM, 2023: 1-11.

[10] Lin Yihong, Fan Zhaoxin, Wu Xianjia, et al. GLDiTalker: speech-driven 3D facial animation with graph latent diffusion transformer[PP/OL]. V5. arXiv (2025-12-05)[2026-03-01]. https://arxiv.org/abs/2408.01826.

[11] Ramesh A, Dhariwal P, Nichol A, et al. Hierarchical text-conditional image generation with CLIP latents[PP/OL]. arXiv (2022-04-13)[2026-03-01]. https://arxiv.org/abs/2204.06125.

[12] Hsu W N, Bolte B, Tsai Y H, et al. HuBERT: self-supervised speech representation learning by masked prediction of hidden units[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 3451-3460.

[13] Bai Shaojie, J. Kolter Z, Koltun V. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling[PP/OL]. V2. arXiv (2018-04-19)[2026-03-01]. https://arxiv.org/abs/1803.01271.

[14] Cho K, Van Merriënboer B, Gulcehre C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Stroudsburg: ACL, 2014: 1724-1734.

[15] Rao Zhuang, Ding Dazhao, Wang Yijing. Human activity recognition method based on CSI principal component segmentation[J]. Journal of Zhengzhou University (Engineering Science), 2025, 46(6): 49-57. [饶壮, 丁大钊, 王依菁. 基于 CSI 主成分分割的人体动作识别方法[J]. 郑州大学学报(工学版), 2025, 46(6): 49-57.]

[16] Yang Jie, Fan Jiarou, Wang Yiru, et al. Hierarchical feature embedding for attribute recognition[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway:

IEEE, 2020: 13052-13061.

[17] Su Jianlin, Ahmed M, Lu Yu, et al. RoFormer: enhanced transformer with rotary position embedding[J]. Neurocomputing, 2024, 568: 127063.

[18] Sun Zhiyao, Luy Tian, Ye Sheng, et al. DiffPoseTalk: speech-driven stylistic 3D facial animation and head pose generation via diffusion models[J]. ACM Transactions on Graphics, 2024, 43(4): 1-9.

[19] Fanelli G, Gall J, Romsdorfer H, et al. A 3-D audio-visual corpus of affective communication [J]. IEEE Transactions on Multimedia, 2010, 12(6): 591-598.

[20] Daněček R, Chhatre K, Tripathi S, et al. Emotional speech-driven animation with content-emotion disentanglement[C]//SIGGRAPH Asia 2023. New York: ACM, 2023: 1-13.

[21] Wang Kaisiyuan, Wu Qianyi, Song Linsen, et al. MEAD: a large-scale audio-visual dataset for emotional talking-face generation [C] // European Conference on Computer Vision. Cham: Springer, 2020: 700-717.

[22] Feng Yao, Feng Haiwen, Black M J, et al. Learning an animatable detailed 3D face model from in-the-wild images[J]. ACM Transactions on Graphics, 2021, 40(4): 1-13.

[23] Zielonka W, Bolkart T, Thies J. Towards metrical reconstruction of human faces[C]//European Conference on Computer Vision. Cham: Springer, 2022: 250-269.

[24] Peng Ziqiao, Luo Yihao, Shi Yue, et al. SelfTalk: a self-supervised commutative training diagram to comprehend 3D talking faces[C]//Proceedings of the 31st ACM International Conference on Multimedia. New York: ACM, 2023: 5292-5301.

[25] Song Wenfeng, Wang Xuan, Zheng Shi, et al. TalkingStyle: personalized speech-driven 3D facial animation with style preservation[J]. IEEE Transactions on Visualization and Computer Graphics, 2025, 31(9): 4682-4694.

A Method for Personalized Speech-Driven 3D Facial Animation Generation

LI Wei[1,2], SONG Yupu[1,2], LIU Yazhi[1,2], AN Yi[1,2]

(1. College of Artificial Intelligence, North China University of Science and Technology, Tangshan 063210, China; 2. Hebei Key Laboratory of Industrial Intelligent Perception, North China University of Science and Technology, Tangshan 063210, China)

Abstract: To address the challenges of speech-driven 3D facial animation, including difficult alignment between speech and motion, loss of identity features, and limited personalized dynamic expression, a conditional diffusion-based generation framework was proposed. The framework used a dual-path style encoding structure to extract hierarchical identity features and dynamic motion features, and then applied a bidirectional attention mechanism to deeply fuse speech features with noisy motion features. Based on this design, an improved Transformer decoder guided by style conditions was introduced to generate high-quality motion sequences. Experiments on the BIWI, VOCASET, and 3DMEAD datasets showed that the proposed method achieved the best results in average vertex error (*MVE*), lip vertex error (*LVE*), and facial dynamic deviation (*FDD*). Compared with the best baseline method on each metric, *MVE*, *LVE*, and *FDD* were reduced by 4.8%, 15.4%, and 13.4% respectively on BIWI, *LVE* was reduced by 14.9% on VOCASET, and *MVE* and *FDD* were reduced by 10.2% and 13.7% respectively on 3DMEAD. Subjective evaluation results further confirmed its advantages in visual naturalness and realism. The proposed method provided a new technical approach for high-fidelity generation, identity preservation, and personalized modeling of 3D facial animation.

Keywords: speech-driven animation; 3D facial animation; deep learning; diffusion model; personalization