arXiv:2601.10000v1 [cs.MM] 15 Jan 2026

EditEmoTalk：通过连续表情编辑实现可控的语音驱动三维面部动画

Diqiong Jiang[1], Kai Zhu[1], Dan Song[2],Jian Chang[3],Chenglizhao Chen[1*],
Zhenyu Wu[41]中国石油大学, [2]天津大学, [3]伯恩茅斯大学,[4]西南交通大学

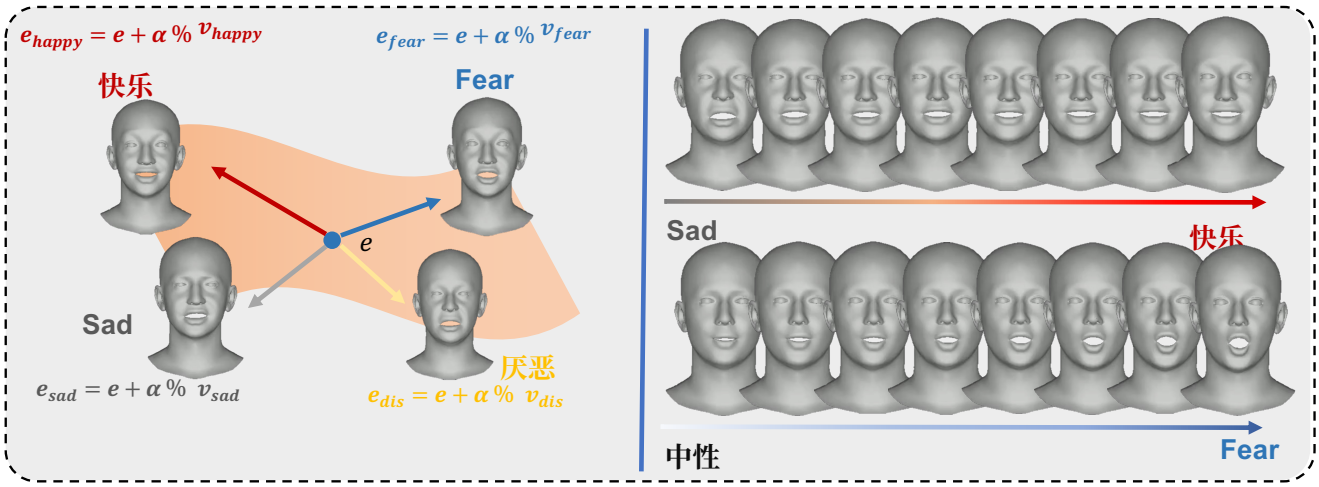


图 1. **EditEmoTalk：面向语音驱动三维面部动画的连续情感编辑。**（左）在学习到的表情流形中进行情感编辑。每个点代表一个表情嵌入，在流形上进行平滑插值可实现自然的情感过渡。（右）EditEmoTalk 生成的连续三维面部网格动画结果。顶行展示了从悲伤 → 快乐的平滑过渡，底行展示了从中性 → 恐惧的过渡，体现了在保持自然唇部同步的同时实现细粒度情感控制。

摘要

语音驱动的三维面部动画旨在直接从音频生成逼真且富有表现力的面部运动。尽管近期方法实现了高质量的唇部同步，但它们往往依赖离散情感类别，限制了连续且细粒度的情感控制。我们提出 *EditEmoTalk*，这是一种支持连续情感编辑的可控语音驱动三维面部动画框架。其核心思想是一种边界感知语义嵌入，通过学习不同情感决策边界的法向量，构建连续的表情流形，从而实现平滑的情感操控。此外，我们引入一种情感一致性损失，通过映射网络确保生成的运动动态与目标情感嵌入在语义上对齐，以确保情感表达忠实准确。大量实验表明，*EditEmoTalk* 在可控性、表现力和泛化能力方面表现出色，同时保持精准的唇部同步。代码和预训练模型将会发布。

1. 引言

根据语音生成自然且富有表现力的三维面部动画，一直是计算机图形学、计算机视觉和人机交互领域长期面临的挑战。语音驱动动画为人工制作提供了一种经济高效的替代方案，并能在虚拟助手、游戏、电影制作和沉浸式通信等应用中实现数字人的规模化部署。尽管基于深度学习的方法取得了显著进展，

法 [4, 9, 17, 37, 40], 当前方法在捕捉和控制情感表达方面仍然有限。

大多数现有技术 [14, 15, 32] 直接将语音映射为面部运动，往往产生僵硬且情感中性的表情。表情条件方法 [6, 10, 22, 29] 会融入情感线索，但通常依赖离散情感类别（例如“快乐”“悲伤”），这限制了平滑插值和连续控制。因此，这些方法难以生成细腻的情感过渡或精细的强度调节，而这正是实现可信且具有交互性的数字人的关键要素。

为克服这些局限，我们提出**EditEmoTalk**，一种支持连续情感编辑的可控语音驱动三维面部动画框架。我们的方法不依赖类别情感嵌入，而是学习情感间决策边界的法向量，从而围绕具有感知意义的情感过渡构建连续的表情流形 。沿这些边界感知方向进行编辑，可确保生成的轨迹始终处于语义有效且情感鲜明的区域内，从而实现平滑、可解释且情感一致的面部编辑，同时保持高保真度的语音同步。

为进一步确保编辑后的情感嵌入与生成的动作动态之间的语义对齐，我们引入了情感一致性损失。该目标通过学习到的映射网络，强制编辑后的情感向量与生成的面部运动相对应，从而确保输出能够忠实反映预期的情感语义。此外，我们采用双重训练策略，在训练过程中随机交替使用语音驱动输入和情感编辑输入。该设计使模型在未进行编辑时能够稳健地保持自然的唇形同步，同时在引入编辑后实现精确且连续的情感操控。

通过将情绪控制重新表述为在学习到的表情流形上的连续遍历，**EditEmoTalk**在单一框架内统一了离散情感分类与连续情感编辑。

我们的主要贡献如下：

- **EditEmoTalk：** 一种可控的语音驱动三维面部动画框架，实现了连续且细粒度的情感控制，弥合了离散情感识别与连续表情编辑之间的鸿沟。
- **边界感知语义嵌入：** 一种新颖的表示方法，通过学习情感间边界法向量来构建连续的表情流形，从而实现平滑且可解释的情感过渡。
- **情感一致性损失和双重训练策略：** 一种协同设计，使运动动态与编辑后的情感嵌入保持一致，同时保留自然的唇部同步与语义一致性。

2. 相关工作

2.1. 语音驱动的三维面部动画

语音驱动面部动画长期以来一直是计算机视觉和计算机图形学领域的核心问题。早期方法依赖语言学分析，利用手工制定的协同发音规则将音素映射到视素 [36]。随后，人们引入了隐马尔可夫模型（HMM） [2]和高斯过程潜变量模型（GPLVM） [7, 8] 等统计方法，以捕捉运动轨迹中的时间依赖关系。

随着深度学习的兴起，基于卷积神经网络和循环神经网络的模型开始直接根据音频预测面部运动 [19, 34, 35], , 其唇部同步效果比传统方法更加准确。近期基于 Transformer 和扩散的方法 [6, 9, 29, 37] 进一步增强了时间连贯性和表现力。与基于图像的方法相比，基于网格的表示 [24, 38, 41] 具有更高的几何保真度，能够实现更加逼真的面部运动。尽管取得了这些进展，现有方法大多主要关注下半脸运动和精确的唇部同步，往往忽视眉毛和眼部运动等上半脸动态，而这些对于传达情感至关重要。此外，情感通常以离散类别或全局风格代码表示，限制了平滑插值和细粒度控制。

2.2. 面部表情编辑与控制

面部表情的编辑与控制已在二维和三维领域得到广泛探索。在图像和视频中，基于生成对抗网络和扩散的方法通过潜在空间遍历或语义条件控制实现逼真的编辑 [13, 16, 26, 27]。在三维建模中，表情控制通常通过混合形状插值、潜在因子解耦或将身份与表情因子分离的参数化模型来实现 [1, 18, 21], , 从而支持插值、强度缩放和跨主体迁移。

近期一些研究尝试将情感融入语音驱动动画 [6, 10, 11, 29], , 但它们大多依赖离散情感类别或低维风格嵌入。这限制了细粒度控制，也阻碍了情绪状态之间的连续过渡，导致动画的表现力有限。

相比之下，我们的工作提出了一种用于语音驱动三维面部动画的连续情感感知控制统一框架。通过借助边界感知语义嵌入构建连续表情流形，并通过动作对齐确保时间情感一致性，EditEmoTalk 能够对下半脸和上半脸动作进行细粒度、可平滑插值的情感控制，从而弥合真实的

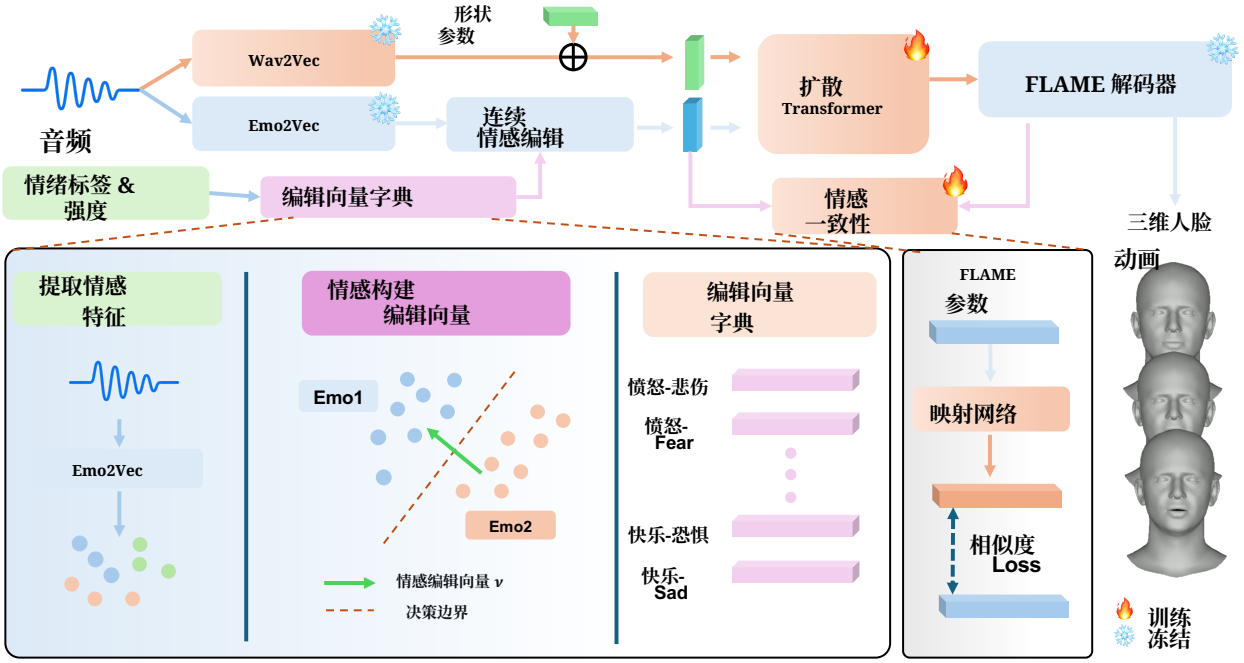


图2。**可控三维面部动画框架概览**。我们的模型能够根据语音音频生成三维面部运动，并支持连续情感编辑。该流程首先提取语音特征（HuBERT）和情感线索（emo2vec）。连续编辑的关键在于对扩散过程进行表情条件控制：将边界感知编辑向量字典中的表情向量与音频特征及 FLAME 参数相结合，形成扩散 Transformer 的条件信号。该条件信号引导生成式去噪过程，从而精确控制输出表情——直接调整表情向量即可调节生成网格序列的情感风格，同时保持唇部同步。随后，网格解码器生成最终的三维面部动画。

语音驱动动态与灵活、富有表现力的动画之间的差距。

3. 方法

如图 2所示，**EditEmoTalk** 由四个主要组件构成，这些组件协同生成可控且富有情感表现力的三维面部动画。该流程首先通过语音特征编码器从原始音频中提取具有判别性的声学表示，并将其与从**编辑向量字典**中检索到的表情嵌入融合，以提供情感条件。最后，**连续表情编辑模块**在连续情绪流形的引导下调节表情强度和风格，同时通过情感一致性损失，确保情绪嵌入与运动动态之间的语义对齐。

3.1. 语音特征编码器

为捕捉语音的语言和情感方面，我们采用双分支编码器，将原始波形分解为内容和情感成分。

• **内容编码器：** 预训练HuBERT [12] 模型用作内容编码器 $\mathcal{E}_c$，从输入波形中 X提取音素级表示 $A = \mathcal{E}_c(X)$。

• **情绪编码器：** 我们采用 emotion2vec [25] 作为情绪编码器 $\mathcal{E}_e$，以获得能够捕捉话语整体情绪状态的紧凑嵌入$e = \mathcal{E}_e(X)$。该嵌入为我们框架中的连续情感编辑提供了基础。

3.2. 连续表情编辑模块

为实现细粒度的连续情绪控制，我们引入了一个在学习得到的情绪特征空间中运作的连续表情编辑模块。与基于离散标签的方法不同，该模块对情绪空间的几何结构进行建模，从而能够实现平滑且易于理解的表情操控。

情绪空间。 设 $E = \{e_1, e_2, \ldots, e_N\}$ 表示从带标签数据集中提取的情感嵌入，其中每个 $e_i \in \mathbb{R}^d$ 都对应一个带有标签$y_i \in \{1, \ldots, K\}$的音频片段。这些嵌入形成多个簇，反映人类情绪的语义拓扑。我们的目标

是揭示簇之间的几何关系，以实现连续编辑。

边界感知语义嵌入。我们的编辑策略的核心洞见是，情感过渡中语义意义最丰富的方向编码在情绪类别之间的决策边界中。直观而言，决策边界的法向量指向远离某一情绪类别、朝向另一情绪类别的最陡上升方向。因此，我们通过在情绪嵌入 E 上训练多类线性分类器，为每种情感 k 估计一个局部分离超平面。该超平面表示为：

$$f_k(e) = \mathbf{w}_k^T e + b_k = 0,$$

其中，$\mathbf{w}_k$ 是决策边界的法向方向。这些边界并非用于分类，而是刻画相邻情绪之间的语义过渡，形成编码相对情绪变化的几何感知表示。

连续情绪流形。归一化后的边界法向量定义了主要编辑方向：

$$\mathbf{v}_k = \frac{\mathbf{w}_k}{\|\mathbf{w}_k\|_2},$$

每个方向都对应情绪流形上的一个切向方向。这些方向共同张成一个局部线性子空间，用以近似连续情绪流形，从而实现情绪状态之间的平滑过渡。

连续控制。给定输入嵌入 e_{orig}，通过一个标量 α调整情绪强度：

$$e_{edited} = e_{orig} + \alpha \cdot \mathbf{v}_k,$$

或者，对于多情绪编辑，

$$e_{edited} = e_{orig} + \sum_{k=1}^{K} \alpha_k \cdot \mathbf{v}_k.$$

这支持直观的连续过渡，例如从中性过渡到微笑，或从平静过渡到兴奋。随后，编辑后的嵌入 e_{edited} 被馈送为条件信号，输入我们扩散 Transformer 的交叉注意力层，从而引导生成过程输出与目标表情一致的网格。

实现。 决策边界离线计算，归一化的方向向量 $\{\mathbf{v}_1$、……、$\mathbf{v}_K\}$存储在**编辑向量字典** 中，以供实时推理使用。该设计：(1) 只需类别标签，无需强度标注；(2) 提供具有语义可解释性且几何一致的编辑；(3) 仅修改与情感相关的特征，从而保留语音内容相关性。

3.3. 扩散 Transformer 与网格解码器

我们采用条件扩散模型，根据语音生成三维面部运动。我们的扩散 Transformer（DiT）骨干网络 [28] 对 FLAME 参数序列进行去噪 $\mathbf{\Theta}_t \in \mathbb{R}^{T\times D}$，条件包括来自 HuBERT 的语音特征（$A$）、来自我们的编辑向量字典的情绪嵌入（$e$）以及身份系数（$\beta$）。去噪过程 $\mathbf{\Theta}_{pred}$ =DiTBlock（$\mathbf{\Theta}_t$、t、A、e、β）确保情感一致性，同时保持唇部同步。

最终去噪后的参数 $\mathbf{\Theta}_0 = \{\beta, \psi, \theta\}$ 通过 FLAME 模型 $\mathcal{M} = \mathcal{G}_{FLAME}(\beta,\ \psi,\ \theta)$进行解码，以生成输出的三维网格序列，同时保持时间连贯性和网格质量。

3.4. 训练目标

我们采用多层级目标函数，以确保高质量的三维面部动画生成，涵盖参数重建、几何保真度、时间连贯性和情感一致性。

3.4.1. 重建损失

采用掩码均方误差（MSE）损失，确保预测的 FLAME 动作与真实动作之间的参数级重建精度：

$$\mathcal{L}_{recon} = \frac{1}{TD} \sum_{t=1}^{T} \|\mathbf{P}_t - \mathbf{P}_t^{gt}\|_2^2,$$

其中，$\mathbf{P} \in \mathbb{R}^{T\times D}$ 表示预测的动作参数，$\mathbf{P}^{gt}$ 表示对应的真实值，T 表示序列长度，D 表示参数维度。

3.4.2. 网格级几何损失

为确保生成的三维网格具有良好的几何保真度，我们在解码后的 FLAME 输出上定义顶点级几何损失：

• **网格重建损失：** 通过最小化逐顶点欧氏距离，确保整体几何精度：

$$\mathcal{L}_{mesh} = \frac{1}{TV} \sum_{t=1}^{T} \|\mathbf{V}_t - \mathbf{V}_t^{gt}\|_2^2,$$

其中$\mathbf{V}_t \in \mathbb{R}^{V\times 3}$ 和$\mathbf{V}_t^{gt}$ 分别表示预测的顶点位置和真实顶点位置，V 表示顶点数量。

•**表面法线损失：** 通过惩罚顶点法线的偏差来保持局部几何平滑性：

$$\mathcal{L}_{normal} = \frac{1}{TV} \sum_{t=1}^{T} \|\mathbf{N}_t - \mathbf{N}_t^{gt}\|_2^2,$$

其中$\mathbf{N}_t$ 和$\mathbf{N}_t^{gt}$ 分别为预测的顶点法线和真实顶点法线。

3.4.3. 时间连贯性损失

通过一阶和二阶平滑约束来增强时间一致性：

• **网格速度损失：**保持一阶时间连续性：

$$\mathcal{L}_{vel} = \frac{1}{(T-1)V} \sum_{t=1}^{T-1} \|\Delta \mathbf{V}_t - \Delta \mathbf{V}_t^{gt}\|_2^2,$$

其中 $\Delta \mathbf{V}_t = \mathbf{V}_{t+1} - \mathbf{V}_t$。

•**网格加速度损失：**约束二阶平滑性：

$$\mathcal{L}_{acc} = \frac{1}{(T-2)V} \sum_{t=1}^{T-2} \|\Delta^2 \mathbf{V}_t - \Delta^2 \mathbf{V}_t^{gt}\|_2^2,$$

其中 $\Delta^2 \mathbf{V}_t = \mathbf{V}_{t+2} - 2\mathbf{V}_{t+1} + \mathbf{V}_t$。

3.4.4. 情绪一致性损失

我们引入了一个轻量级**映射网络**，由两个全连接层和 GELU 激活函数组成，用于将预测的表情参数映射到**情感一致性损失**所处的空间，从而确保编辑后的嵌入与生成的动作保持一致。**情感相似度损失**定义为余弦距离：

$$\mathcal{L}_{emo} = 1 - \frac{\mathbf{e} \cdot \mathbf{e}^{gt}}{\|\mathbf{e}\|_2 \, \|\mathbf{e}^{gt}\|_2},$$

其中，$\mathbf{e}$和$\mathbf{e}^{gt}$ 分别是预测的情感嵌入和目标情感嵌入。

3.4.5. 总体目标与训练策略

最终目标函数以经过经验调优的权重整合了所有组成部分：

$$\mathcal{L}_{total} = \lambda_{recon}\mathcal{L}_{recon} + \lambda_{mesh}\mathcal{L}_{mesh} + \lambda_{normal}\mathcal{L}_{normal} + \lambda_{vel}\mathcal{L}_{vel} + \lambda_{acc}\mathcal{L}_{acc} + \lambda_{emo}\mathcal{L}_{emo}.$$

为了有效训练模型，使其既能准确地根据语音生成动作，又能进行连续情感编辑，我们采用随机训练策略：以 50% 的概率，将从音频中提取的情感嵌入作为输入，并使用完整目标函数 $\mathcal{L}_{total}$ 对模型进行优化；在其余 50% 的情况下，则使用来自连续表情流形的编辑后情感嵌入，并仅应用情感一致性损失

$$\mathcal{L}_{edit} = \lambda_{emo}\mathcal{L}_{emo}$$

。该策略有助于实现忠实的语音驱动动画以及精准、连续的情绪控制。

这种混合策略确保模型学会生成逼真且时间连贯的面部动画，同时能够精准、连续地操控情感表达。

4. 实验

我们对所提出的 **EditE-moTalk** 框架进行了全面评估，以验证其在生成真实感、唇形同步准确度和表情可控性方面的有效性。此外，我们还开展了消融研究，以证明关键设计选择的合理性。

4.1. 数据集与预处理

我们的模型在四个公开可用的视听数据集上进行训练和评估，以确保说话人、语言和情感表达的多样性：

MEAD [39]：一个高质量、多视角的情感视听数据集，包含 60 位演员在严格受控的条件下，通过七台同步摄像机录制的表演；演员演绎了八种情感，每种情感包含三个强度等级。该数据集是我们进行可控情感建模的主要数据集。 **CREMA-D** [3]：包含 91 位背景各异的演员演绎 12 个句子的 7,442 段片段，涵盖六种情感和四个强度等级。该数据集还包含大规模众包情感评分，可用于稳健地评估情感泛化能力。**RAVDESS** [23]：包含 24 位专业演员录制的 7,356 段音频，内容为语音和歌曲表演，涵盖八种情感和两个强度等级；数据提供纯音频、纯视频和视听三种模态，支持多模态情感分析。**HDTF** [42]：由从 YouTube 收集的约 16 小时高分辨率说话人脸部视频组成。作为一个自然环境数据集，它包含多样的姿态、光照和背景，是评估真实场景泛化能力的挑战性基准。

预处理： 将音频重采样至 16 kHz。我们采用现成的三维人脸重建方法 [30]获取作为训练目标的真实三维网格序列，并将面部运动表示为相对于中性帧的顶点位移。

4.2. 评估指标

我们使用定量和定性指标评估 EditEmoTalk。定量方面，顶点误差 *(VE)*衡量预测顶点与真实顶点之间的平均 L2 距离（单位：mm），反映整体重建精度。唇部顶点误差 *(LVE)* [31] 评估唇形同步精度，嘴部张开偏差 *(MOD)* [5]衡量发音幅度，而面部动态偏差 *(FDD)* [40]则通过比较顶点速度来评估时间维度上的运动平滑度。

此外，我们引入 ΔCH，用于通过 Calinski–Harabasz（CH）指数评估生成动作中情绪聚类结构的保留情况。CH 指数衡量*FLAME* 表情参数空间中簇间离散度与簇内离散度之比。 ΔCH 值越小，表明生成的动作所维持的情绪簇内部越紧凑，

且与其他情感的区分越明显，从而与真实情感流形高度吻合。

不同于以往工作，我们**不报告情感分类准确率**。更高的分类分数并不一定意味着更具表现力或更逼真的动画；相反，强制采用离散情感类别通常会降低**时间连贯性**并增加 LVE，因为僵化的类别边界与**人类情感的连续动态**相冲突。

在定性评估方面，一项用户研究采用 5 分制，从自然度、唇形同步和情感真实性三个方面报告平均意见分（*MOS*），并将 EditEmoTalk 与基线方法进行比较。

4.3. 基线方法

EditEmoTalk 与多种最先进的三维说话头生成方法进行了比较。**DiffPoseTalk** [33] 采用姿态级扩散，以生成具有时间平滑性的表现力动作。**Emote** [6]采用离散情感嵌入，以实现可控表情。**EmoTalk** [29] 引入情感解耦编码器（EDE），通过交叉重建将情感与内容分离，在保持唇形同步的同时实现可控的表情合成。**DeepTalk** [20]训练了一个动态情感嵌入（DEE）模块，并结合概率对比学习对语音和面部运动进行联合建模，从而捕捉细腻的情绪变化。

4.4. 实现细节

我们的模型使用 PyTorch 实现，并通过 PyTorch Lightning 进行训练，采用 AdamW优化器（lr $= 1 \times 10^{-4}$，权重衰减 1×10^{-5}）和余弦退火调度器。在单张 RTX 4090 GPU 上以批量大小 8 训练 200 个训练轮次。模型包含 446M 个参数。训练期间会随机启用情绪编辑，以促进解耦控制。在推理时，情感向量 $\mathbf{e}$（独热编码或连续值）可实现灵活的情感操控。**推理速度**。 模型每帧（转换为 FLAME参数）耗时 0.035ms，达到实时性能。

4.5. 结果与分析

4.5.1. 定量结果

跨数据集分析。 表 1 展示了全面的跨数据集比较，并纳入了所提出的 Δ_{CH}指标以评估情感一致性。在所有数据集上，EditEmoTalk 的 VE 和 LVE 均为最低，表明其几何精度和唇形同步精度更优。同时，它的 MOD 和 FDD 数值也具有竞争力，反映出稳定的发音幅度和平滑的时序动作。

除了动作质量之外，EditEmoTalk 在 CREMA-D、RAVDESS 和 MEAD 上的 ΔCH 值也最低，这表明其生成的面部运动保留了最接近真实值的情感聚类结构。相比之下，DeepTalk 和 DiffPoseTalk 等方法的 ΔCH 值更高，意味着其情绪分离效果较弱，情感表征也不够连贯。值得注意的是，尽管 Emote 生成的动作在视觉上富有表现力，其 ΔCH 值却极高。我们认为，这是因为它依赖离散的情绪嵌入来强制划分僵化的类别边界，这与情感表达的连续性相悖，导致其聚类模式显得不自然，并偏离了自然的情感流形。

4.5.2. 定性结果

我们进一步进行了定性评估，以直观地评估生成面部动画的感知质量。本节重点关注两个关键方面：**表情丰富度**和**可控性分析**。第一方面考察不同方法再现与情感语音相对应的多样且自然的面部表情的能力，第二方面则评估模型能否以一致且逼真的方式控制和编辑情感强度或类别。

4.5.3. 表情丰富度

为了进一步评估生成动画的表现力，我们在图3中展示了不同方法的结果。输入语音为“现在是十一点”，重音自然落在音节“一”上，并分别搭配五种不同的情感：愤怒、快乐、中性、恐惧和厌恶。我们提出的**EditEmoTalk**展现出更出色的表情丰富度和面部动态，能够生成细腻且时间连贯的表情，并与语音节奏和情感线索高度契合。尽管**DiffPoseTalk** 采用了与我们的方法类似的基于扩散的架构，但缺少**EditEmoTalk**所采用的双通道结构和情感一致性损失。因此，其输出中的情感变化有限，往往接近中性或表现力较弱的状态。**DeepTalk** 在愤怒和厌恶情感中表现出明显的强度；然而，不同情感下生成的结果往往呈现相似的面部紧张模式，导致情感区分度较低、情绪多样性不足。相比之下，我们的方法能够生成类别间区分更明显、变化更自然的情感表达，从而有效提升语音驱动的三维面部动画的情绪真实感和视觉吸引力。

4.5.4. 可控性分析

EditEmoTalk 的一项关键优势在于其明确且细粒度的情绪可控性。通过操控表情控制向量$\mathbf{e}$，我们的模型无需重新训练，即可将同一语音输入渲染为不同的情绪状态。

我们进一步与

表 1. 跨数据集泛化性能比较。由于 HDTF 缺少聚类评估所需的一致情绪标签，因此省略其 ΔCH。由于 EmoTalk 未采用基于 FLAME 的面部拓扑，某些指标（例如 FDD 和 **ΔCH**）无法直接比较，因此予以省略。

数据集	方法	VE（毫米）↓	LVE (毫米) ↓	MOD (毫米) ↓	FDD（10^{-5} 毫米/秒）↓	Δ通道 ↓
CREMA-D	DeepTalk	2.0733	11.4756	1.7794	42.8523	4.7572
	DiffPoseTalk	1.8965	12.3115	2.4100	31.2141	0.8483
	EmoTalk	2.4353	12.4324	1.8323	–	0.9110
	Emote	1.8902	9.7072	1.7654	64.4193	62.8581
	Ours	**0.5553**	**4.8682**	**0.9318**	**30.8475**	**0.0865**
RAVDESS	DeepTalk	2.2248	13.2099	2.3545	33.4368	2.2636
	DiffPoseTalk	1.6545	9.4647	1.5972	39.3294	0.8540
	EmoTalk	2.3432	10.3224	2.3324	–	0.8539
	Emote	2.0747	10.9277	2.3453	55.2231	17.708
	Ours	**0.6940**	**6.4486**	**1.1563**	**33.3030**	**0.0942**
MEAD	DeepTalk	2.0430	11.5721	2.0534	28.0587	0.6194
	DiffPoseTalk	1.6939	10.6261	1.7888	31.6314	0.9655
	EmoTalk	2.5432	11.2343	1.7943	–	0.9888
	Emote	1.8046	8.6625	1.5026	49.3305	35.855
	Ours	**0.5340**	**4.7738**	**0.9179**	**25.3608**	**0.0623**
HDTF	DeepTalk	2.0551	13.6581	2.6189	69.7752	–
	DiffPoseTalk	1.5078	8.5308	1.1930	87.2130	–
	EmoTalk	1.9344	11.2342	2.0892	–	–
	Emote	1.6248	9.8394	2.2399	89.8650	–
	Ours	**0.8084**	**7.8892**	**1.7177**	**86.8194**	–

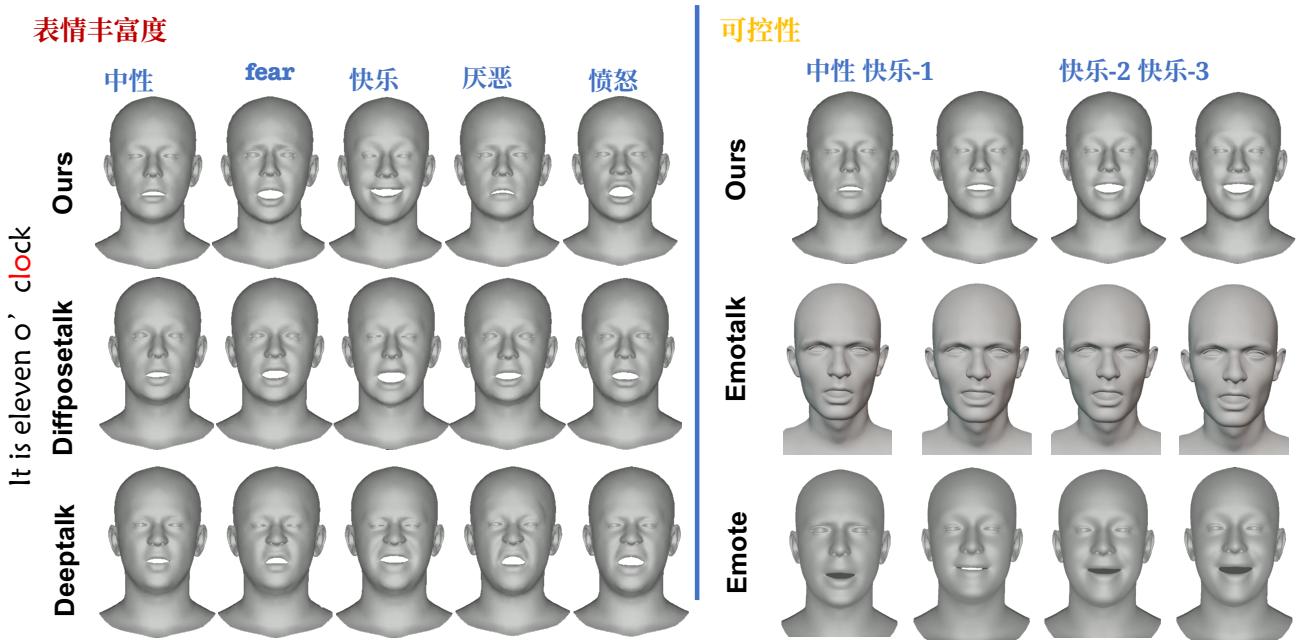


图3。**左：**不同方法在表情丰富度方面的比较。每种方法都根据相同的语句“现在是十一点”生成面部动画，并在音节“点”上强调情感。我们的方法能够产生更丰富的情感表现力和更流畅的面部动态，尤其是在刻画愤怒、快乐、中性、恐惧和厌恶等细腻表情方面。**右：**可控性分析，展示了自然表情以及三个等级的微笑强度编辑结果。

EmoTalk 和**Emote** 进行定性比较，以评估可控性和视觉表现力。如图3所示，EmoTalk 表现出有限的编辑能力——在不同情感设置下生成的表情几乎没有变化。

表2. 用户研究结果（平均意见得分，1–5）。我们的 **EditEmoTalk** 在所有评估方面均取得了最高的感知质量。

方法	自然度 ↑	唇形同步 ↑	情感 ↑
DiffPoseTalk	3.82	**4.15**	3.05
DeepTalk	3.56	3.47	3.29
EmoTalk	3.21	3.88	3.67
Emote	4.07	3.74	3.93
Ours	**4.46**	4.07	**4.49**

表3. 在 MEAD 验证集上的消融研究。双重训练策略和情感一致性损失都有助于提升面部动画质量。LVE 和 MOD 的单位为 mm，FDD 的单位为 10^{-5} mm/s。

模型变体	LVE ↓	MOD ↓	FDD ↓
完整模型	**4.7738**	**0.9179**	**28.3608**
无双重训练	5.1024	0.9451	29.6286
无 $\mathcal{L}_{emo}$	4.9261	0.9324	30.7245

Emote 在不同情绪状态之间呈现出明显差异；然而，情感强度和面部动态不够突出，有时也不一致。相比之下，我们的 **EditEmoTalk** 能够生成清晰可辨的情感表情，并实现平滑过渡和一致的唇部同步，有效地将与情感相关的面部形变与语音驱动的动作解耦。

这种可控合成展示了**EditEmoTalk** 在富有表现力且可交互的数字人应用中的潜力，例如虚拟主持人、数字助理和情感感知型虚拟形象。

4.5.5. 用户研究

为进一步评估生成面部动画的感知质量和情感保真度，我们开展了一项全面的用户研究，重点考察三个方面：自然度、唇部同步准确性和情感真实性。我们邀请了 30 名参与者（15 名男性和 15 名女性，年龄为 20 至 35 岁），对五种方法生成的渲染视频进行评估：DiffPoseTalk、DeepTalk、EmoTalk、Emote 和我们提出的 EditEmoTalk。每位参与者观看了从不同情感类别中随机选取的 20 个样本；为消除潜在偏差，方法和情感类型的呈现顺序均经过随机化。参与者使用 5 分制平均意见分（MOS）量表（1 = 很差， 5 =极好）对每段视频进行评分，分别评估整体真实感和运动平滑度（自然度）、唇部运动与语音之间的同步程度（唇形同步准确性），以及所表达情感的清晰度和真实感（情感真实性）。

如表 2所示，EditEmoTalk 在所有感知维度上的整体平均意见得分（MOS）均最高。尽管基于扩散的方法**DiffPoseTalk** 表现出出色的唇部同步性能，这得益于其姿态级建模，但其表达多样性不足，导致面部动态单调。**DeepTalk**生成的动作连贯但相对僵硬，而**EmoTalk**虽提供情绪控制，却无法维持平滑的时间过渡。尽管**Emote** 呈现出视觉细节丰富的形变，但其情感一致性在不同帧之间波动，降低了感知真实感。相比之下，我们的**EditEmoTalk** 能够生成时间稳定且情感一致的动画，在自然动作、同步发音和真实表达之间实现有效平衡。这凸显了我们的双通道情感引导架构和情感一致性损失在生成生动逼真的三维说话头部方面的优势。

4.6. 消融研究

为评估我们框架中关键组件的作用，我们在 MEAD数据集上进行了消融实验，重点考察（1）双重训练策略和（2）情感一致性损失 $\mathcal{L}_{emo}$。 **双重训练策略**。移除双重训练策略后，模型在训练期间无法充分利用原始语音驱动嵌入和编辑后的情感嵌入。这会降低连续情感编辑的灵活性，并略微削弱面部运动保真度，表明双重训练能够有效平衡忠实的语音驱动生成与可控的情感操控。 **情感一致性损失**。排除 $\mathcal{L}_{emo}$ 会导致面部运动抖动，以及不同帧之间的情感强度不一致。加入此损失可确保时间连贯性和情感稳定性，从而生成更平滑、更自然的面部动态。表3 报告了定量结果。两种组件均显著提升了性能，而采用双重训练策略并加入 $\mathcal{L}_{emo}$的完整模型在顶点误差（VE）、唇部顶点误差（LVE）、嘴部张开偏差（MOD）和面部动态偏差（FDD）各项指标上均取得了最佳表现。

5. 结论

我们提出了 EditEmoTalk，这是一种可控的语音驱动三维面部动画框架，可实现连续且细粒度的情感编辑。该框架利用边界感知语义嵌入对情绪间边界进行建模，构建连续表情流形，从而实现平滑且可解释的情感操控。通过轻量级映射网络实现的情感一致性损失，确保生成的动作忠实反映目标情绪。实验表明，EditEmoTalk 在表现力、可控性和跨数据集泛化方面均表现出色，同时保持了精准的唇部同步。

参考文献

[1] V Blanz 和 T Vetter。用于合成三维人脸的可变形模型。载于第26届计算机图形学与交互技术年度会议（SIGGRAPH 1999），第187–194页。ACM出版社，1999年。2[2] Matthew Brand。语音木偶操控。载于第26届计算机图形学与交互技术年度会议论文集，第21–28页，1999年。2[3] Houwei Cao、David G Cooper、Michael K Keutmann、Ruben C Gur、Ani Nenkova 和 Ragini Verma。CREMA-D：众包情感多模态演员数据集。IEEE《情感计算汇刊》，5(4)：377–390，2014年。5[4] Yujin Chai、Tianjia Shao、Yanlin Weng 和 Kun Zhou。通过风格-内容解耦实现个性化音频驱动三维面部动画。IEEE《可视化与计算机图形学汇刊》，30(3)：1803–1820，2022年。2[5] Daniel Cudeiro、Timo Bolkart、Cassidy Laidlaw、Anurag Ranjan 和 Michael J Black。三维说话风格的捕捉、学习与合成。载于 IEEE/CVF计算机视觉与模式识别会议论文集，第10101–10111页，2019年。5[6] Radek Danˇeˇcek、Kiran Chhatre、Shashank Tripathi、Yandong Wen、Michael Black 和 Timo Bolkart。基于内容-情感解耦的情感语音驱动动画。载于 SIGGRAPH Asia 2023 会议论文集。2, 6[7] Salil Deena 和 Aphrodite Galata。使用共享高斯过程潜变量模型的语音驱动面部动画。载于国际视觉计算研讨会，第89–100页。Springer，2009年。2[8] Salil Deena、Shaobo Hou 和 Aphrodite Galata。使用变阶切换共享高斯过程动力学模型的视觉语音合成。IEEE《多媒体汇刊》，15(8)：1755–1768，2013年。2[9] Yingruo Fan、Zhaojiang Lin、Jun Saito、Wenping Wang 和 Taku Komura。Faceformer：使用 Transformer 的语音驱动三维面部动画。载于 IEEE/CVF计算机视觉与模式识别会议论文集，第18770–18780页，2022年。2[10] Guanwen Feng、Zhiyuan Ma、Yunan Li、Junwei Jing、Jiahao Yang 和 Qiguang Miao。FaceEditTalker：通过面部属性编辑实现交互式说话头生成。arXiv 预印本 arXiv:2505.22141，2025年。2[11] Tianshun Han、Benjia Zhou、Ajian Liu、Yanyan Liang、Du Zhang、Zhen Lei 和 Jun Wan。PeSTalk：具有个性化情感风格的语音驱动三维面部动画。载于第33届 ACM 多媒体国际会议论文集，第7893–7901页，2025年。2[12] Wei-Ning Hsu、Benjamin Bolte、Yao-Hung Hubert Tsai、Kushal Lakhotia、Ruslan Salakhutdinov 和 Abdelrahman Mohamed。HuBERT：通过对隐藏单元进行掩码预测实现的自监督语音表征学习。IEEE/ACM《音频、语音与语言处理汇刊》，29：3451–3460，2021年。3[13] Ziqi Huang、Kelvin CK Chan、Yuming Jiang 和 Ziwei Liu。用于多模态人脸生成与编辑的协同扩散。载于 IEEE/CVF计算机视觉与模式识别会议论文集，页码6080–6090，2023年2。[14] Farzaneh Jafari、Stefano Berretti 和 Anup Basu。Jambatalk：基于混合 Transformer-Mamba 模型的语音驱动三维说话头生成。arXiv预印本 arXiv:2408.01627，2024年。2[15] Xiaozhong Ji、Chuming Lin、Zhonggan Ding、Ying Tai、Junwei Zhu、Xiaobin Hu、Donghao Luo、Yanhao Ge 和 Chengjie Wang。Realtalk：利用三维面部先验引导的身份对齐网络，实现实时逼真的音频驱动人脸生成。arXiv预印本 arXiv:2406.18284，2024年。2[16] Diqiong Jiang、Dan Song、Ruofeng Tong 和 Min Tang。Styleipsb：用于高保真人脸交换的 StyleGAN 身份保持语义基。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，页码352–361，2023年。2[17] Diqiong Jiang、Jian Chang、Lihua You、Shaojun Bian、Robert Kosk 和 Greg Maguire。深度学习驱动的面部动画：综述。《Information》，15(11):675，2024年。2[18] Zi-Hang Jiang、Qianyi Wu、Keyu Chen 和 Juyong Zhang。三维人脸形状的解耦表征学习。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，页码11957–11966，2019年。2[19] Tero Karras、Timo Aila、Samuli Laine、Antti Herva 和 Jaakko Lehtinen。通过姿态和情感的端到端联合学习实现音频驱动的面部动画。《ACM图形学汇刊》（ToG），36(4):1–12，2017年。2[20] Jisoo Kim、Jungbin Cho、Joonho Park、Soonmin Hwang、Da Eun Kim、Geon Kim 和 Youngjae Yu。DeepTalk：用于概率式语音驱动三维人脸动画的动态情感嵌入。载于《AAAI人工智能会议论文集》，页码4275–4283，2025年。6[21] Tianye Li、Timo Bolkart、Michael J Black、Hao Li 和 Javier Romero。从四维扫描中学习面部形状与表情模型。《ACM图形学汇刊》，36(6):194–1，2017年。2[22] Xiangyu Liang、Wenlin Zhuang、Tianyong Wang、Guangxing Geng、Guangyue Geng、Haifeng Xia 和 Siyu Xia。Cstalk：相关性监督的语音驱动三维情感面部动画生成。载于《2024年IEEE第18届自动人脸与手势识别国际会议（FG）》，页码1–5。IEEE，2024年。2[23]Steven R Livingstone 和 Frank A Russo。Ryerson情感语音与歌曲视听数据库（RAVDESS）：北美英语中面部和声音表情的动态多模态集合。《PloS one》，13(5):e0196391，2018年。5[24] Yuanxun Lu、Jinxiang Chai 和 Xun Cao。实时语音肖像：实时逼真的说话头动画。《ACM图形学汇刊》（ToG），40(6):1–17，2021年。2[25] Ziyang Ma、Zhisheng Zheng、Jiaxin Ye、Jinchao Li、Zhifu Gao、Shiliang Zhang 和 Xie Chen。emotion2vec：语音情绪表征的自监督预训练。arXiv预印本 arXiv:2312.15185，2023年。3[26] Andrew Melnik、Maksim Miasayedzenkau、Dzianis Makaravets、Dzianis Pirshtuk、Eren Akbulut、Dennis Holzmann、Tarek Renusch、Gustav Reichert 和 Helge Ritter。人脸生

成与编辑：StyleGAN综述。《IEEE模式分析与机器智能汇刊》，46(5):3557–3576，2024年。2[27] Or Patashnik、Zongze Wu、Eli Shechtman、Daniel Cohen-Or 和 Dani Lischinski。StyleCLIP：文本驱动的StyleGAN图像操控。载于《IEEE/CVF国际计算机视觉大会论文集》，页码2085–2094，2021年2。[28] William Peebles 和 Saining Xie。基于Transformer的可扩展扩散模型。载于《IEEE/CVF国际计算机视觉大会论文集》，页码4195–4205，2023年4。[29] Ziqiao Peng、Haoyu Wu、Zhenbo Song、Hao Xu、Xiangyu Zhu、Jun He、Hongyan Liu 和 Zhaoxin Fan。EmoTalk：用于三维面部动画的语音驱动情感解耦。载于《IEEE/CVF国际计算机视觉大会论文集》，页码20687–20697，2023年。2，6[30] George Retsinas、Panagiotis P Filntisis、Radek Danecek、Victoria F Abrevaya、Anastasios Roussos、Timo Bolkart 和 Petros Maragos。通过分析驱动的神经合成生成3D面部表情。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，页码2490–2501，2024年。5[31] Alexander Richard、Michael Zollh¨ofer、Yandong Wen、Fernando De la Torre 和 Yaser Sheikh。MeshTalk：通过跨模态解耦实现语音驱动的三维人脸动画。载于《IEEE/CVF国际计算机视觉大会论文集》，页码1173–1182，2021年。5[32] Stefan Stan、Kazi Injamamul Haque 和 Zerrin Yumak。FaceDiffuser：使用扩散模型合成语音驱动的三维面部动画。载于《第16届ACM SIGGRAPH运动、交互与游戏大会论文集》，页码1–11，2023年2。[33] Zhiyao Sun、Tian Lv、Sheng Ye、Matthieu Lin、Jenny Sheng、Yu-Hui Wen、Minjing Yu 和 Yong-jin Liu。DiffPoseTalk：通过扩散模型生成语音驱动的风格化三维面部动画与头部姿态。《ACM图形学汇刊》（TOG），43(4):1–9，2024年。6[34] Supasorn Suwajanakorn、Steven M Seitz 和 Ira Kemelmacher-Shlizerman。合成奥巴马：从音频中学习唇形同步。《ACM图形学汇刊》（ToG），36(4):1–13，2017年。2[35] Sarah Taylor、Taehwan Kim、Yisong Yue、Moshe Mahler、James Krahe、Anastasio Garcia Rodriguez、Jessica Hodgins 和 Iain Matthews。通用语音动画的深度学习方法。《ACM图形学汇刊》（TOG），36(4):1–11，2017年。2[36] Nadia Magnenat Thalmann 和 Daniel Thalmann。计算机动画中的模型与技术。Springer，1993年。2[37] Balamurugan Thambiraja、Ikhsanul Habibie、Sadegh Aliakbarian、Darren Cosker、Christian Theobalt 和 Justus Thies。Imitator：个性化语音驱动的三维面部动画。载于《IEEE/CVF国际计算机视觉大会论文集》，页码20621–20631，2023年。2[38] Linrui Tian、Qi Wang、Bang Zhang 和 Liefeng Bo。Emo：Emote肖像动画生成——在弱条件下利用音频转视频扩散模型生成富有表现力的肖像视频。载于

欧洲计算机视觉会议，第244–260页。Springer，2024年。2[39] Kaisiyuan Wang 、 Qianyi Wu 、 Linsen Song 、 Zhuoqian Yang、Wayne Wu、Chen Qian、Ran He、Yu Qiao 和 Chen Change Loy。MEAD：用于情感说话人脸生成的大规模音视听数据集。发表于 *ECCV*，2020年。 5[40] Jinbo Xing、Menghan Xia、Yuechen Zhang、Xiaodong Cun、Jue Wang 和 Tien-Tsin Wong。Codetalker：基于离散运动先验的语音驱动三维面部动画。发表于 *IEEE/CVF*计算机视觉与模式识别会议论文集，第12780–12790页，2023年。 2，5[41] Sicheng Xu、Guojun Chen、Yu-Xiao Guo、Jiaolong Yang、Chong Li、Zhenyu Zang、Yizhong Zhang、Xin Tong 和 Baining Guo。Vasa-1：实时生成的逼真音频驱动说话人脸。神经信息处理系统进展，第37卷，660–684页，2024年。 2[42] Zhimeng Zhang、Lincheng Li、Yu Ding 和 Changjie Fan。基于高分辨率音视听数据集的流引导一次性说话人脸生成。发表于*IEEE/CVF*计算机视觉与模式识别会议论文集，第3661–3670页，2021年。 5