中图法分类号： 文献标识码：A 文章编号：1006-8961(XXXX)XX-0001-15

论文引用格式：Chen Sheng, Sun Qiang, Zhu Xiatian. XXXX. Emotion-controllable 3D Talking Face Generation with Hierarchical Disentanglement-guided VQ-VAE. Journal of Image and Graphics, XX(XX):0001-0015(陈胜, 孙强, 朱霞天. XXXX. 分层解耦引导的情感可控VQ-VAE3D说话人脸生成方法. 中国图象图形学报, XX(XX):0001-0015)[DOI:10.11834/jig.250451]

分层解耦引导的情感可控VQ-VAE3D说话人脸生成方法

陈胜[1],孙强[1*],朱霞天[2]

1. 西安理工大学自动化与信息工程学院,西安 710048; 2. 英国萨里大学以人为本人工智能研究所&视觉、语音和信号处理中心,吉尔福德 GU2 7XH,英国

摘 要：目的 当前,基于语音驱动的3D说话人脸生成技术已经较为成熟,尤其是基于向量量化变分自编码器(VQ-VAE)的模型。但现有方法情感引导方式较为简单,未能充分利用多模态信息;此外,现有的基于VQ-VAE的人脸生成模型重建的人脸细节有限。**方法** 本文在VQ-VAE框架下对人脸特征进行分层解耦,将其分为顶层和底层特征,并引入身份向量和描述文本作为外部条件解耦顶层特征,随后顶层特征再作为内部条件,与外部条件一起解耦底层特征,实现特征的分层解耦,以提升人脸的重建质量并增强情感表达的稳定性。在生成过程中,语音和文本共同引导人脸生成,从而充分利用多模态信息。**结果** 在3DMEAD和TA-MEAD数据集上的定量实验表明,本模型在生成人脸的平均顶点误差(MVE)、平均估计误差(MEE)和覆盖误差(CE)等关键指标上均达到最优,唇部顶点误差(LVE)为次优。可视化分析显示生成的表情整体上更生动逼真。消融实验验证了外部条件与内部条件对模型效果的关键作用。**结论** 本文提出的方法有效提升了人脸重建的准确性与表情生成的稳定性,生成结果在整体上更接近真实人脸。本文代码已在GitHub开源并于ScienceDB存档,访问链接为:https://github.com/XAUT-VcLab/2022CS-Project1和https://doi.org/10.57760/sciencedb.j00240.00095。

关键词： 3D人脸生成;VQ-VAE;Transformer;特征量化;情感可控

Emotion-controllable 3D Talking Face Generation with Hierarchical Disentanglement-guided VQ-VAE

Chen Sheng[1], Sun Qiang[1*], Zhu Xiatian[2]

1. School of Automation and Information Engineering, Xi'an University of Technology, Xi'an 710048, China; 2. The Surrey Institute for People-Centred Artificial Intelligence and the Centre for Vision, Speech and Signal Processing (CVSSP) at the University of Surrey, Guildford GU2 7XH, UK

Abstract: Objective Currently, speech-driven 3D talking face generation technology has matured considerably, particularly with recent models based on Vector Quantized Variational Autoencoders (VQ-VAE), which achieve more vivid facial expression generation by quantizing facial motion features. However, existing methods still suffer from limitations in facial reconstruction accuracy and emotional control stability. On the one hand, the facial detail reconstruction capability of VQ-VAE models requires further improvement; on the other hand, emotion guidance strategies in existing models remain rela-

收稿日期：2025-09-17;**修回日期：**2026-01-13

*通信作者：孙强 qsun@xaut.edu.cn

基金项目：西安理工大学国际科技合作促进项目(2024GHCJ014)

Supported by: International Science and Technology Cooperation Promotion Project of Xi'an University of Technology(2024GHCJ014)

tively simplistic, failing to fully utilize multimodal information beyond speech, resulting in generated facial expressions that lack realism. **Method** Inspired by the structures of Conditional Variational Autoencoders (CVAE) and VQ-VAE-2, this paper proposes a hierarchical disentanglement method within the VQ-VAE framework. Facial features are decoupled into high-level and low-level features. Identity vectors and descriptive text are introduced as external conditions to disentangle the high-level features. The disentangled high-level features are then used as internal conditions, working together with the two external conditions (identity vector and text) to further disentangle the low-level features. This hierarchical disentanglement mechanism aims to improve facial reconstruction quality and enhance the stability of emotional expression. During generation, speech and text information jointly guide facial generation in different ways: the text is encoded by the CLIP text encoder to leverage its rich semantic information for holistic expression control, while the speech is encoded by Wav2Vec 2.0 to precisely control lip movements and other subtle expressions. By integrating multimodal information at different levels, the model generates more accurate facial motions. **Results** Experiments were conducted on the 3DMEAD and TA-MEAD datasets and included quantitative evaluation, visual analysis, and ablation studies. The proposed model was compared with FaceFormer, CodeTalker, FaceDiffuser, and ProbeTalk3D. In terms of overall facial reconstruction accuracy, the proposed model achieved a 5.9% lower Mean Vertex Error (MVE) than ProbeTalk3D—which also incorporates emotional input—demonstrating that explicit emotion modeling effectively enhances geometric consistency and preserves facial structural accuracy, particularly for complex emotions. For lip motion accuracy (LVE), the proposed model ranked second, with a 16.7% gap compared to the top-performing FaceDiffuser, while still outperforming FaceFormer and ProbeTalk3D. The limited improvement may stem from the hierarchical disentanglement approach, which, while decoupling emotional conditions, offers restricted decoupling for the mouth region, affecting detailed performance. In upper-face dynamics deviation (FDD), although inferior to ProbeTalk3D, the proposed model significantly surpassed traditional methods, reflecting its ability to express emotional intensity. More importantly, the model showed notable advantages in distribution-based quality metrics: it achieved optimal Mean Estimation Error (MEE) and Coverage Error (CE), with MEE 4.1% lower than ProbeTalk3D and 58.4% lower than CodeTalker, and CE 8.2% lower than ProbeTalk3D and 61.6% lower than CodeTalker, indicating that hierarchical disentanglement enhances distribution concentration and coverage of real samples, making generated facial motions closer to real faces. However, the model had the lowest Diversity score, highlighting a trade-off where hierarchical disentanglement ensures semantic consistency but constrains variation—beneficial for emotion accuracy yet potentially limiting naturalness. While traditional methods showed orders-of-magnitude differences in MVE and FDD due to lacking emotional input, and ProbeTalk3D excelled in FDD and Diversity but lagged in accuracy, the proposed model balances high-precision reconstruction with superior distribution estimation through hierarchical disentanglement and multimodal modeling. **Conclusion** Experimental results demonstrate that the proposed hierarchical disentanglement-based method produces more accurate and stable facial reconstructions, excelling in maintaining facial geometric consistency and expressing complex emotions. The model's key innovation lies in its layered feature disentanglement within the VQ-VAE framework, using identity vectors and text as external conditions and high-level features as an internal condition, which significantly enhances reconstruction quality and emotional controllability. This offers practical value for HCI, digital humans, and animation. Quantitative evaluation shows superior performance in MVE, LVE, and FDD, alongside stable MEE and CE scores, confirming the model's ability to effectively integrate multimodal information for generating realistic expressions. Ablation studies validate the critical roles of both the hierarchical feature disentanglement and the synergy between internal and external conditions. However, limitations remain: generated lip movements are less pronounced compared to real speech, the current two-level feature disentanglement is relatively simplistic, and the diversity and subtlety of emotional expressions require further improvement. Future work will focus on improving lip-sync quality, exploring finer-grained disentanglement dimensions for differentiated guidance, and enhancing expression diversity while preserving accuracy. The source code is available on GitHub and has been archived on ScienceDB athttps://github.com/XAUT-VcLab/2022CS-Project1andhttps://doi.org/10.57760/sciencedb.j00240.00095, respectively.

Key words: 3D face generation; VQ-VAE; Transformer; feature quantization; emotionally controllable

0 引 言

基于语音驱动的3D说话人脸生成技术是一种基于语音内容生成对应3D说话人脸视频的技术，广泛应用于虚拟现实、人机交互、电影制作、数字娱乐等领域(Gao等，2024；Zhen等，2023)。在人机交互中，该技术可增强互动的沉浸感和真实性；在数字内容创作中，该技术能够显著降低人工制作的成本和时间，极大地提升生产效率。近年来，随着深度学习技术的发展和硬件算力的提升，3D说话人脸生成技术取得了显著进展，基于深度学习的3D说话人脸生成模型逐渐取代了传统的基于规则的生成方法，展现出更强的通用性和生成质量(Song等，2023)。

目前，基于深度学习的3D说话人脸生成方法通过从大量数据中学习语音与人脸数据之间的复杂映射关系，使生成的人脸在准确性和真实性上有了显著的提升(Jiang等，2022)。基于3D人脸数据的生成模型可以通过Transformer(Vaswani等，2017)等模型结构直接学习语音和人脸数据的映射关系，得到较高的合成精度。此外，现有模型通过如wav2vec 2.0(Baevski等，2020)或HuBERT(Hsu等，2021)等预训练的语音处理模型预先对输入语音进行编码，可以使编码后的语音特征能够利用预训练模型的先验知识，学习与人脸数据的内在联系，提高生成质量。不只是音频模态，3D说话人脸生成模型也可以借助如CLIP(Radford等，2021)等多模态预训练模型，利用其潜在空间的丰富语义信息，通过多模态信息生成表情更加真实的人脸。但是，尽管目前的3D说话人脸生成模型在嘴部动作方面表现出色，其生成的人脸在面部表情的准确性和稳定性方面仍有提升空间。

在众多生成模型中，向量量化变分自编码器(VQ-VAE)模型(Van等，2017)因其对面部动作特征的量化处理而被广泛应用于3D说话人脸生成。Kingma等人(2013)提出的VAE(变分自编码器)模型奠定了生成模型“编码-采样-解码”的基础框架，但其连续潜在空间导致生成结果模糊，而VQ-VAE通过引入向量量化将潜在空间离散化，实现了高质量、清晰的重建。通过量化层，基于VQ-VAE的3D说话人脸生成模型能够生成更自然的面部表情和动作，提高了生成的真实感。然而，当前基于VQ-VAE的3D说话人脸生成模型在实际应用中仍面临一些问题。现有的基于VQ-VAE的3D说话人脸生成模型未能充分利用多模态输入中的情感信息，导致生成人脸的面部表情质量有限，难以反映语音中的丰富情感。此外，VQ-VAE模型生成人脸的稳定性和准确性也有待提高。

为了解决以上问题，本文从条件变分自编码器(CVAE)模型(Sohn等，2015)和VQ-VAE-2模型(Razavi等，2019)的结构中受到启发，提出一种基于分层解耦的情感可控VQ-VAE 3D说话人脸生成方法。CVAE模型通过为VAE注入条件信息解耦输入特征，实现了可控生成，VQ-VAE-2模型则在VQ-VAE基础上引入了分层结构，对输入特征内部进行解耦，显著提升了生成结果的准确性与一致性，本文将这两种方法相结合，用于3D说话人脸生成中，以提高面部重建的质量，增强生成人脸和情感表达的稳定性。一方面，在VQ-VAE框架中引入身份向量和文本描述作为外部条件解耦人脸特征，以丰富模型对多样化人脸特征的捕捉能力；另一方面将人脸特征分为顶层和底层，进行分层解耦处理，顶层特征作为内部条件进行控制，实现对面部表情的多层次控制。相较于CVAE仅对单层特征解耦，本文方案实现了多层解耦，相较于VQ-VAE-2仅在特征内部解耦，本文方案将外部条件也参与进解耦中，通过外部条件的辅助提升解耦效果。

本文的主要贡献包括以下几个方面：

1)提出了一种基于分层解耦的情感可控VQ-VAE 3D说话人脸生成方法：通过对人脸特征的顶层和底层进行分层处理，在外部条件解耦和内部条件解耦的基础上，实现对人脸特征的多层次控制，提升了人脸重建的准确性。

2)稳定VQ-VAE的生成结果：通过在VQ-VAE框架的基础上添加身份向量和文本描述作为外部条件，并将顶层特征作为内部条件，实现了对人脸的身份特征和情感特征的更好捕捉和学习，增强VQ-VAE模型生成人脸的稳定性。

3)增强模型的可解释性：通过分层解耦结构和条件引导方法，本模型能够较为清晰地分析不同模态的条件对面部表情生成的影响，为3D说话人脸生成模型的应用和改进提供了透明化的理论依据。

4)通过基于3DMEAD和TA-MEAD数据集的大量实验表明，本文方法生成人脸的准确性和稳定性

优于经典的3D说话人脸生成方法。

1 相关工作

1.1 基于语音驱动的3D说话人脸生成

基于语音驱动的说话人脸生成技术旨在从语音信号中提取特征，进而生成相应的人脸动作，使生成的人脸视频能够自然、准确地展现对应的面部表情和口型变化（Pan等，2025）。现有的说话人脸生成方法按输出维度可分为基于2D的方法和基于3D的方法。基于2D的方法通常利用语义映射，通过将语音特征映射到人脸特征点上，以生成说话者的平面表情（Ji等，2022；Gan等，2023；Liang等，2022）。这些方法通常可以较好地处理嘴部动作，但生成的面部动作缺乏深度信息，难以实现真实的3D效果。而基于3D的方法则通过如D3DFR（Deng等，2019）、3DMesh（Richard等，2021）、blendshape（Cudeiro等，2019）、FLAME（Li等，2017）等3D人脸重建方法得到的3D人脸参数，来生成具备深度信息的说话人脸。通过这些参数，模型可以得到精确的3D人脸顶点，生成的面部表情不仅在视觉上更具真实感，还可以有效控制面部的多维度运动，能更好地适应多样化的应用场景（Liu等，2024）。

在基于语音驱动的3D说话人脸生成中，根据生成方式的不同，可以分为自回归模型和回归模型。自回归模型通过循环生成机制，将当前帧的面部特征输入作为下一帧面部特征生成的依据，并在语音特征的引导下依次生成后续帧（Chu等，2024）。FaceFormer模型（Fan等，2022）是一种典型的自回归模型，其人脸生成部分基于Transformer解码器结构，在每一时刻的面部特征生成中，交叉注意力层将语音特征作为外部输入，生成下一帧的人脸特征。然而，自回归模型的训练过程需要较长的时间，且对输入序列的变化较为敏感，稳定性不足。相较之下，回归模型则直接将语音输入映射为面部特征序列，通常具有较高的灵活性和更快的生成速度（Ma等，2023）。PMMTalk模型（Han等，2023）是一种典型的回归模型，通过多种编码器分别提取输入语音的不同特征，并在时序和语义上进行对齐，最终通过解码器生成复杂的面部动作。这类模型可以通过在语音输入的基础上添加其他特征（如情感标签或个人风格）进一步丰富面部生成效果，因此具有较高的实用性。

为了进一步提升基于语音驱动的3D说话人脸的生成效果，近年来一些方法加入了VAE结构，先通过VAE学习面部特征的先验知识，以更好地引导语音生成说话人脸的过程。例如，EMOTE模型（Daněček等，2023）在训练生成模型前，先通过基于Transformer编码器结构的VAE模型学习FLAME面部参数的特征，再将VAE模型的解码器部分作为生成模型的解码器，为人脸生成提供先验。目前的研究进一步采用VQ-VAE框架进行重建学习（Chen等，2023；Kim等，2024）。在3D说话人脸生成过程中，VQ-VAE模型通过在编码过程中引入离散的量化层，将连续面部动作特征转换为离散表示，显著提升了生成面部表情的真实性。CodeTalker（Xing等，2023）是首个将VQ-VAE用于人脸生成中的模型，其通过VQ-VAE将每一帧3D顶点坐标量化，再通过自回归结构生成连续的人脸特征序列，使生成的面部动作更加流畅自然。另一种方法VividTalker（Zhao等，2023）则通过在时间维度上分别量化面部的不同特征，利用窗口Transformer模型结合语音内容逐段重建面部特征，有效提升了生成效果。

尽管VQ-VAE模型在3D说话人脸生成方面表现出色，但目前仍存在一些局限。VQ-VAE模型重建人脸的精度和稳定性有限，且目前基于VQ-VAE的3D说话人脸生成模型未能充分利用多模态信息，生成的人脸表情准确度有限。为解决这些问题，本文对VQ-VAE模型进行了改进，通过引入身份向量和文本描述作为条件约束，优化多层次的人脸特征解耦，以增强模型对不同情感的控制能力，提高生成过程的稳定性与准确性。

1.2 情感引导的3D说话人脸生成

情感引导的3D说话人脸生成技术旨在语音驱动的基础上，通过额外的情感输入生成更加丰富的面部表情，以增强生成人脸情感表达的准确性和多样性。此类方法通常通过如情感标签（Sun等，2024；Ye等，2022）、描述文本（Sun等，2024）、面部表情图像（Xu等，2023；Wang等，2023）等多种形式的情感信息作为额外输入，使模型能够更加细致地控制面部表情的生成。情感引导可以显著提升3D说话人脸生成的真实性和多样性，适用于角色动画、社交机器人等需表达多样化情绪的应用场景。

一些方法选择将人脸情感特征单独提取并进行

预测。比如，Chen等人(2023)将情感标记与基于语音的去中心化情感先验相结合，作为额外的情感输入，再利用独立的情感增强网络生成面部表情，并将其直接添加到生成的人脸上，以实现情感引导的面部表情增强效果。另一类方法将情感特征作为外部引导输入，以更全面地控制人脸的情感表达。EmoTalk模型(Peng等，2023)便是此类方法中的代表。该模型通过一个情绪解耦编码器分离语音中

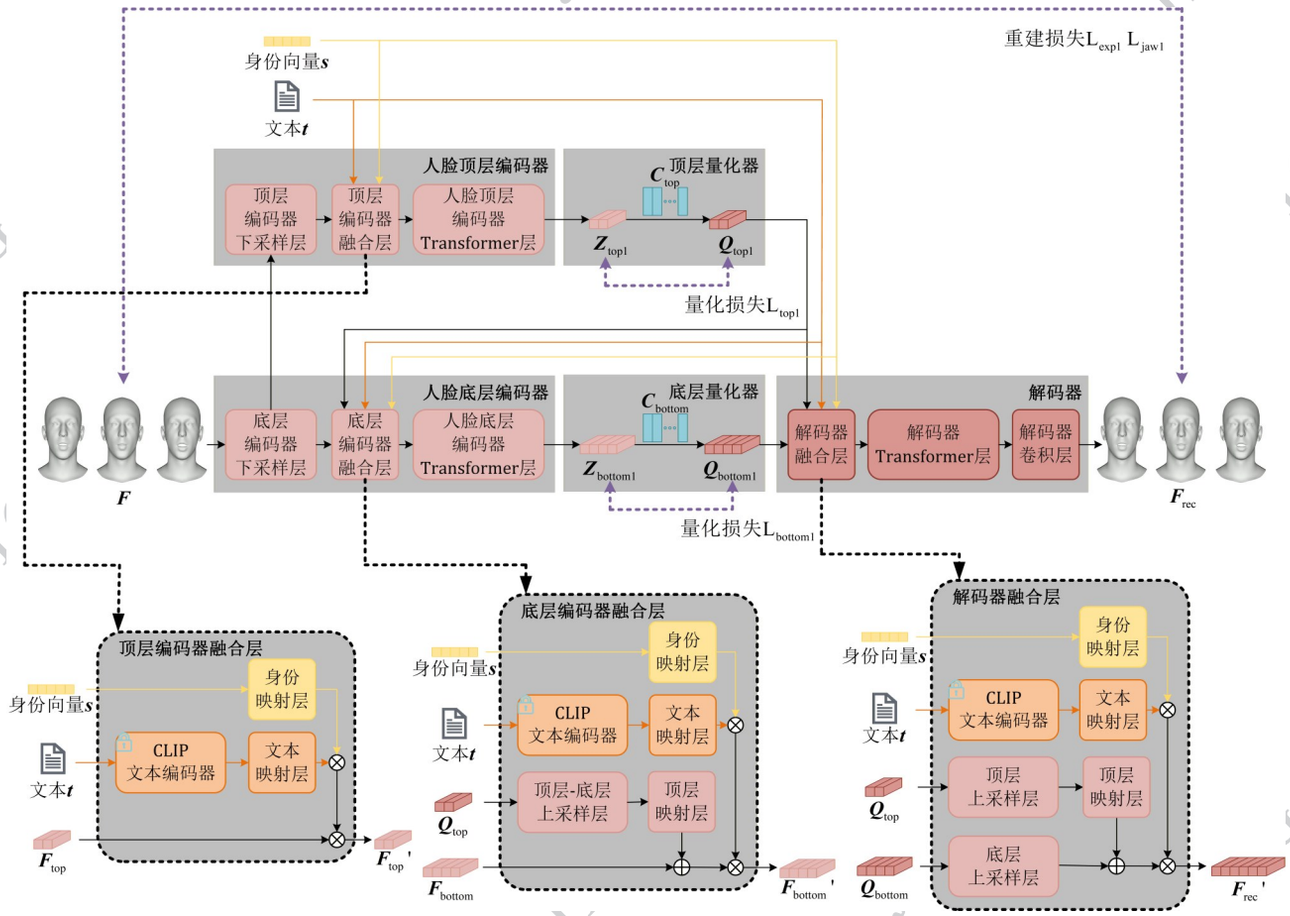

图1　第一阶段模型结构和训练流程

Fig. 1　The model architecture and training pipeline on stage 1

的情绪和内容，得到的语音情感特征一方面与语音内容特征、个人风格和情绪控制特征一起输入Transformer解码器层，另一方面也作为交叉注意力层的外部输入来引导人脸特征，从而生成情感丰富的面部表情。ExpCLIP模型(Zhong等，2024)则利用CLIP模型的丰富语义空间，将文本描述和面部表情图像分别作为情感输入，通过预训练的CLIP文本编码器和CLIP图像编码器，以及对应的映射网络处理后，得到情感特征，再通过交叉注意力层引导情感的生成，使生成的面部表情与输入的文本或表情图片情感保持一致。还有一些方法将情感特征与语音特征组合输入模型，以直接生成带有情感的面部表情。比如，EMOTE模型(Daněček等，2023)通过不同编码器分别提取的语音特征和时序的情感标记相加后直接输入解码器，以预测人脸的运动。ProbTalk3D(Wu等，2024)则通过将描述整体情感的情感向量与语音特征相乘的方式实现引导，以增强生成表情的情感表现力。

尽管上述情感引导方法在情感特征处理上有所创新，但大多仅是将情感特征与人脸特征简单组合，对人脸特征本身的处理仍较为有限，因此在生成情感的准确性方面仍存在一定的局限。此外，现有的基于VAE或VQ-VAE的方法，对如何利用VAE结构的特点进行情感解耦的探索并不充分，只是简单将其作为条件加入生成过程中。针对上述不足，本文提出了一种改进的情感引导方法，通过在模型中引入分层解耦和条件引导，以进一步提升情感生成的可控性，使得生成的面部表情不仅更加生动细腻，同

时具有更高的多样性和准确性。

2 方 法

2.1 问题描述

与现有的基于VQ-VAE的3D说话人脸生成方法类似,本模型的总体训练过程同样分为两个阶段:基于VQ-VAE的条件解耦阶段和基于条件引导的3D说话人脸生成阶段。

在第一阶段,训练一个VQ-VAE模型,从而在第二阶段的生成过程中为人脸特征提供先验。现有的基于VQ-VAE的3D说话人脸生成方法,其人脸重建效果有限,也未充分利用多模态条件。本方法与现有方法相比,并非只是对人脸特征整体进行单层量化或对不同人脸特征分别量化,对于其他模态的输入,也不只是在处理后直接与人脸特征组合,而是让其深度参与到人脸特征的解耦与重建中。本方法的VQ-VAE模型参考了VQ-VAE-2模型和CVAE模型的结构,结合身份向量$\boldsymbol{s}$和文本$\boldsymbol{t}$,以分层解耦的方式学习人脸特征表示。人脸特征$\boldsymbol{F}$是通过FLAME方法得到的3D人脸参数,由表情参数$\boldsymbol{\psi}$以及下颚参数$\boldsymbol{\theta}_{\text{jaw}}$组成,其在解耦后再经过解码器重建人脸特征$\boldsymbol{F}_{\text{rec}}$。身份向量$\boldsymbol{s}$为0和1组成的一维

one-hot向量,长度为对应数据集的人数,1在one-hot向量中的位置代表对应的某个人。文本$\boldsymbol{t}$基于视频标签生成,描述了对应数据中的人脸表情与面部动作。将这两个条件合称为外部条件,用于从外部解耦人脸特征。模型通过逐层解耦人脸特征$\boldsymbol{F}$,以实现比现有方法更精细的人脸重建效果。用VQVAE2(·)表示VQ-VAE-2模型,第一阶段的重建过程用可公式表示为:

$$\boldsymbol{F}_{\text{rec}}=\text{VQVAE2}(\boldsymbol{s},\boldsymbol{t},\boldsymbol{F}) \tag{1}$$

第二阶段与其他基于VQ-VAE的3D说话人脸生成方法一样,在固定第一阶段模型解码器部分权重的基础上,使用人脸特征$\boldsymbol{F}$对应的语音$\boldsymbol{a}$作为模型的输入,结合身份向量$\boldsymbol{s}$和文本$\boldsymbol{t}$,训练一个新的编码器部分的模型,从而共同引导生成3D说话人脸特征$\boldsymbol{F}_{\text{gen}}$。用Generation(·)表示人脸生成过程,第二阶段的3D说话人脸生成过程可用公式表示为:

$$\boldsymbol{F}_{\text{gen}}=\text{Generation}(\boldsymbol{s},\boldsymbol{t},\boldsymbol{a}) \tag{2}$$

2.2 基于VQVAE的条件解耦

为了提升VQ-VAE模型重建人脸特征的效果,第一阶段的重建过程将人脸特征分为顶层和底层两个层次,并分别使用外部条件和外部条件结合内部条件实现分层解耦。第一阶段的整个模型由人脸顶层编码器、人脸底层编码器以及对应的顶层量化器、底层量化器,加上最后的解码器部分组成。第一阶段的模型结构和训练流程如图1所示。

在编码器部分,输入的人脸特征$\boldsymbol{F}$首先通过人脸底层编码器的底层编码器下采样层,压缩时间维度的长度至原长度的1/8,得到人脸底层特征$\boldsymbol{F}_{\text{bottom}}$,$\boldsymbol{F}_{\text{bottom}}$再次通过人脸顶层编码器的顶层编码器下采样层,压缩时间维度的长度至原长度的1/16,得到人脸顶层特征$\boldsymbol{F}_{\text{top}}$。通过顶层编码器融合层,人脸顶层特征$\boldsymbol{F}_{\text{top}}$加上了外部条件$\boldsymbol{s}$和$\boldsymbol{t}$作为约束,再通过人脸顶层编码器Transformer层学习人脸顶层特征,得到$\boldsymbol{Z}_{\text{top1}}$。$\boldsymbol{Z}_{\text{top1}}$输入顶层量化器,经由顶层码本$\boldsymbol{C}_{\text{top}}$量化得到$\boldsymbol{Q}_{\text{top1}}$,实现与外部条件的解耦。$\boldsymbol{Q}_{\text{top1}}$也会作为内部条件,参与人脸底层特征的解耦。人脸底层特征$\boldsymbol{F}_{\text{bottom}}$经过底层编码器融合层,给特征加上外部条件$\boldsymbol{s}$、$\boldsymbol{t}$,与内部条件$\boldsymbol{Q}_{\text{top1}}$形成约束,再通过人脸底层编码器Transformer层学习人脸底层特征,得到$\boldsymbol{Z}_{\text{bottom1}}$。$\boldsymbol{Z}_{\text{bottom1}}$输入底层量化器,经由底层码本$\boldsymbol{C}_{\text{bottom}}$量化得到$\boldsymbol{Q}_{\text{bottom1}}$,实现与外部和内部条件的解耦。

在解码器部分,首先解码器融合层将身份向量$\boldsymbol{s}$、文本$\boldsymbol{t}$、量化后的人脸顶层特征$\boldsymbol{Q}_{\text{top1}}$与量化后的人脸底层特征$\boldsymbol{Q}_{\text{bottom1}}$这4个条件融合到一起,并将时间维度的长度扩展回输入人脸特征$\boldsymbol{F}$的长度,得到融合后的人脸特征$\boldsymbol{F}_{\text{rec}}'$。$\boldsymbol{F}_{\text{rec}}'$再经过解码器Transformer层重建人脸特征,最后通过解码器卷积层输出重建后的人脸特征$\boldsymbol{F}_{\text{rec}}$。$\boldsymbol{F}_{\text{rec}}$可以拆分为表情参数$\boldsymbol{\psi}'$以及下颚参数$\boldsymbol{\theta}_{\text{jaw}}'$,用于计算重建损失。

第一阶段模型的整体重建过程可以用下面的公式表示,式中Down(·)表示下采样操作,Fu(·)表示融合层,T(·)表示Transformer层,Quant(·)表示量化操作,$\text{Conv}_{\text{dec}}(\cdot)$表示解码器卷积层:

$$\begin{aligned}\boldsymbol{F}_{\text{bottom}}&=\text{Down}_{\text{bottom}}(\boldsymbol{F})\\ \boldsymbol{F}_{\text{top}}&=\text{Down}_{\text{top}}(\boldsymbol{F}_{\text{bottom}})\end{aligned} \tag{3}$$

$$\boldsymbol{Q}_{\text{top1}}=\text{Quant}_{\text{top}}(\text{T}_{\text{top1}}(\text{Fu}_{\text{top}}(\boldsymbol{s},\boldsymbol{t},\boldsymbol{F}_{\text{top}})))$$
$$\boldsymbol{Q}_{\text{bottom1}}=\text{Quant}_{\text{bottom}}(\text{T}_{\text{bottom1}}(\text{Fu}_{\text{bottom}}(\boldsymbol{s},\boldsymbol{t},\boldsymbol{Q}_{\text{top1}},\boldsymbol{F}_{\text{bottom}}))) \tag{4}$$

$$\boldsymbol{F}_{\text{rec}}=\text{Conv}_{\text{dec}}(\text{T}_{\text{dec}}(\text{Fu}_{\text{dec}}(\boldsymbol{s},\boldsymbol{t},\boldsymbol{Q}_{\text{top1}},\boldsymbol{Q}_{\text{bottom1}}))) \tag{5}$$

融合层位于人脸顶层编码器和人脸底层编码器的下采样层后，以及解码器部分的输入后，作用是将输入的条件与人脸特征融合在一起，以便后续Transformer层学习特征。通过对多种不同融合策略的训练效果进行比较，本模型最终选择采用相加与相乘组合的方式实现融合。

在各融合层的内部，身份向量$\boldsymbol{s}$通过一个基于前馈网络(FFN)的身份映射层实现编码，得到编码后的身份特征$\boldsymbol{S}$。文本$\boldsymbol{t}$则通过固定权重的CLIP文本编码器加上一个基于FFN的文本映射层实现编码，以充分利用CLIP潜在空间的丰富语义信息，得到编码后的文本特征$\boldsymbol{T}$。最后再将两者与对应融合

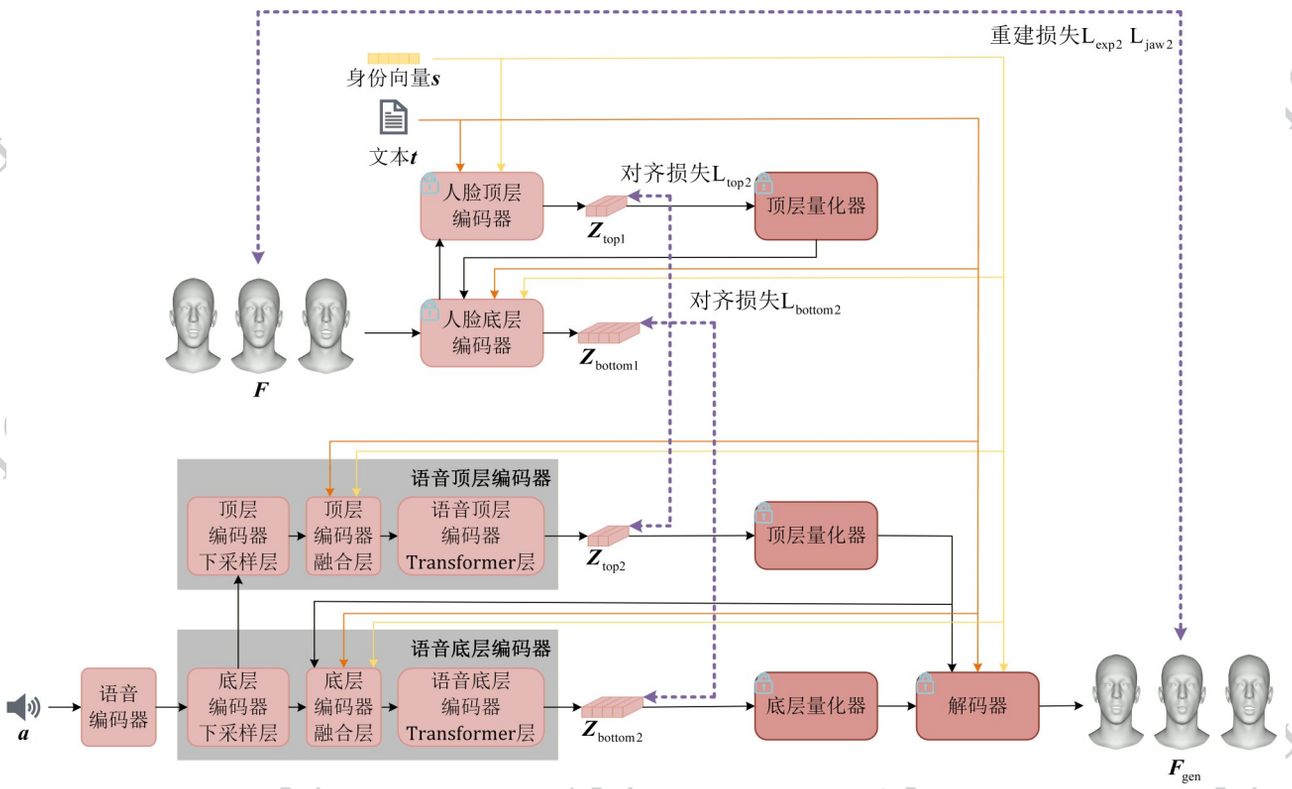

图2　第二阶段模型结构和训练流程

Fig. 2　The model architecture and training pipeline on stage 2

层的人脸特征相乘，实现不同特征之间的融合。在底层编码器融合层中，内部条件$\boldsymbol{Q}_{\text{top1}}$先通过顶层-底层上采样层扩展时间维度的长度到与人脸底层特征$\boldsymbol{F}_{\text{bottom}}$时间维度的长度一致，再通过一个基于FFN的顶层映射层编码，最后将两部分相加。在解码器融合层中，$\boldsymbol{Q}_{\text{top1}}$和$\boldsymbol{Q}_{\text{bottom1}}$先分别通过顶层上采样层和底层上采样层扩展时间维度的长度到与输入的人脸特征$\boldsymbol{F}$时间维度的长度一致，$\boldsymbol{Q}_{\text{top1}}$再通过顶层映射层编码，最后将两部分相加。外部条件的编码和各融合层的融合过程可以用下面的公式表示，式中FFN(·)表示映射层，$\text{Enc}_{\text{clip}}(\cdot)$表示CLIP文本编码器，Up(·)表示上采样操作：

$$\begin{aligned}\boldsymbol{S} &= \text{FFN}_{\text{s}}(\boldsymbol{s})\\ \boldsymbol{T} &= \text{FFN}_{\text{t}}(\text{Enc}_{\text{clip}}(\boldsymbol{t}))\end{aligned} \tag{6}$$

$$\begin{aligned}\boldsymbol{F}_{\text{top}}' &= \boldsymbol{S}\times\boldsymbol{T}\times\boldsymbol{F}_{\text{top}}\\ \boldsymbol{F}_{\text{bottom}}' &= \boldsymbol{S}\times\boldsymbol{T}\times(\text{FFN}_{\text{top}}(\text{Up}_{\text{top-bottom}}(\boldsymbol{Q}_{\text{top1}}))+\boldsymbol{F}_{\text{bottom}})\\ \boldsymbol{F}_{\text{rec}}' &= \boldsymbol{S}\times\boldsymbol{T}\times(\text{FFN}_{\text{top}}(\text{Up}_{\text{top}}(\boldsymbol{Q}_{\text{top1}}))+\text{Up}_{\text{bottom}}(\boldsymbol{Q}_{\text{bottom1}}))\end{aligned} \tag{7}$$

在第一阶段模型的训练过程中，整体的损失L_{stage1}包括四个部分：顶层量化器的量化损失L_{top1}、底层量化器的量化损失$\text{L}_{\text{bottom1}}$、表情参数$\boldsymbol{\psi}$的重建损失$\text{L}_{\text{exp1}}$以及下颚参数$\boldsymbol{\theta}_{\text{jaw}}$的重建损失$\text{L}_{\text{jaw1}}$。

量化损失用公式表示为：

$$\begin{aligned}\text{L}_{\text{top1}} &= \|\text{sg}(\boldsymbol{Z}_{\text{top1}})-\boldsymbol{Q}_{\text{top1}}\|+\beta\|\boldsymbol{Z}_{\text{top1}}-\text{sg}(\boldsymbol{Q}_{\text{top1}})\|\\ \text{L}_{\text{bottom1}} &= \|\text{sg}(\boldsymbol{Z}_{\text{bottom1}})-\boldsymbol{Q}_{\text{bottom1}}\|+\beta\|\boldsymbol{Z}_{\text{bottom1}}-\text{sg}(\boldsymbol{Q}_{\text{bottom1}})\|\end{aligned} \tag{8}$$

式中，$\|\cdot\|$表示均方损失，sg(·)表示停止梯度操作，$\beta=0.25$为控制代码簿更新稳定性的权重。

重建损失的表达式分别为：

$$L_{exp1} = \|\boldsymbol{\psi} - \boldsymbol{\psi}'\|$$
$$L_{jaw1} = \|\boldsymbol{\theta}_{jaw} - \boldsymbol{\theta}_{jaw}'\| \tag{9}$$

整体损失表示为：

$$L_{stage1} = \lambda_1 L_{top1} + \lambda_2 L_{bottom1} + \lambda_3 L_{exp1} + \lambda_4 L_{jaw1} \tag{10}$$

式中，损失函数的权重经验性地设定为：$\lambda_1 = 2$，$\lambda_2 = 1$，$\lambda_3 = 6$，$\lambda_4 = 8$。

2.3 条件引导的语音驱动3D说话人脸生成

在前面第一阶段的模型解耦人脸特征的基础上，第二阶段的模型使用语音 $\boldsymbol{a}$ 作为输入。语音顶层编码器、语音底层编码器与对应的人脸顶层编码器、人脸底层编码器的结构类似，区别仅在于修改了各Transformer层的层数与注意力头数。此外，在语音底层编码器前增加了包含HUBERT预训练模型在内的语音编码器来编码输入语音。编码后的语音特征 $\boldsymbol{A}$ 同样利用外部条件和内部条件解耦，生成语音顶层特征 $\boldsymbol{A}_{top}$ 和语音底层特征 $\boldsymbol{A}_{bottom}$，再与外部条件一起输入第一阶段训练好并固定权重的量化器和解码器，实现说话人脸特征 $\boldsymbol{F}_{gen}$ 的生成。第二阶段的模型结构和训练流程如图2所示。

为了使训练过程更加稳定，在第二阶段模型训练的同时，生成第一阶段模型固定权重后得到的对应人脸顶层特征 $\boldsymbol{Z}_{top1}$ 和人脸底层特征 $\boldsymbol{Z}_{bottom1}$，并将这两个特征与第二阶段生成的对应语音顶层特征 $\boldsymbol{Z}_{top2}$ 和语音底层特征 $\boldsymbol{Z}_{bottom2}$ 进行对比，结果作为对齐损失项加入到损失中。两个语音编码器输出的 $\boldsymbol{Z}_{top2}$ 和 $\boldsymbol{Z}_{bottom2}$ 使用第一阶段训练的两个码本 $\boldsymbol{C}_{top}$ 和 $\boldsymbol{C}_{bottom}$ 量化，得到量化后的 $\boldsymbol{Q}_{top2}$ 和 $\boldsymbol{Q}_{bottom2}$，再输入固定权重的解码器得到生成的说话人脸特征 $\boldsymbol{F}_{gen}$。$\boldsymbol{F}_{gen}$ 可以拆分为表情参数 $\boldsymbol{\psi}''$ 以及下颚参数 $\boldsymbol{\theta}_{jaw}''$，用于生成人脸视频和计算重建损失。

第二阶段模型的整体训练过程和3D人脸生成过程可以用下面的公式表示，式中interpolate(·)表示线性插值层，HUBERT(·)表示HUBERT模型：

$$\boldsymbol{A} = \mathrm{FFN}_a(\mathrm{interpolate}(\mathrm{HUBERT}(\mathrm{Re}_a(\boldsymbol{a}))))$$
$$\boldsymbol{A}_{bottom} = \mathrm{Down}_{bottom}(\boldsymbol{A})$$
$$\boldsymbol{A}_{top} = \mathrm{Down}_{top}(\boldsymbol{A}_{bottom}) \tag{11}$$

$$\boldsymbol{Q}_{top2} = \mathrm{Quant}_{top}(\mathrm{T}_{top2}(\mathrm{Fu}_{top}(\boldsymbol{s},\boldsymbol{t},\boldsymbol{A}_{top})))$$
$$\boldsymbol{Q}_{bottom2} = \mathrm{Quant}_{bottom}(\mathrm{T}_{bottom2}(\mathrm{Fu}_{bottom}(\boldsymbol{s},\boldsymbol{t},\boldsymbol{Q}_{top2},\boldsymbol{A}_{bottom}))) \tag{12}$$

$$\boldsymbol{F}_{gen} = \mathrm{Conv}_{dec}(\mathrm{T}_{dec}(\mathrm{Fu}_{dec}(\boldsymbol{s},\boldsymbol{t},\boldsymbol{Q}_{top2},\boldsymbol{Q}_{bottom2}))) \tag{13}$$

在第二阶段模型的训练过程中，整体损失 L_{stage2} 由四部分组成：语音顶层特征的对齐损失 L_{top2}、语音底层特征的对齐损失 $L_{bottom2}$，以及 $\boldsymbol{\psi}$ 的重建损失 L_{exp2} 和 $\boldsymbol{\theta}_{jaw}$ 的重建损失 L_{jaw2}。用公式表示为：

$$L_{top2} = \|\boldsymbol{Z}_{top1} - \boldsymbol{Z}_{top2}\|$$
$$L_{bottom2} = \|\boldsymbol{Z}_{bottom1} - \boldsymbol{Z}_{bottom2}\| \tag{14}$$

$$L_{exp2} = \|\boldsymbol{\psi} - \boldsymbol{\psi}''\|$$
$$L_{jaw2} = \|\boldsymbol{\theta}_{jaw} - \boldsymbol{\theta}_{jaw}''\| \tag{15}$$

整体损失表示为：

$$L_{stage2} = \lambda_5 L_{top2} + \lambda_6 L_{bottom2} + \lambda_7 L_{exp2} + \lambda_8 L_{jaw2} \tag{16}$$

式中，损失函数的权重经验性地设定为：$\lambda_5 = 0.2$，$\lambda_6 = 0.1$，$\lambda_7 = 6$，$\lambda_8 = 8$。

3 实 验

3.1 数据集

本研究使用两个主要数据集：3DMEAD和TA-MEAD，以支持模型的训练和评估。

3DMEAD：3DMEAD数据集来自EMOTE论文，基于2D音视频数据集MEAD（Wang等，2020）构建，通过FLAME方法重建得到3D人脸参数。MEAD数据集包含48位不同种族和性别的受试者的英语语音和视频数据，涵盖7种情感类别（快乐、悲伤、惊讶、恐惧、厌恶、愤怒、蔑视），每种情感分为3个强度等级，以及1种中性情感。3DMEAD数据集中的人脸参数帧 $\{\boldsymbol{\beta},\boldsymbol{\psi},\boldsymbol{\theta}\} \in \mathbb{R}^{406}$ 为25帧/秒，包含面部形状参数 $\boldsymbol{\beta} \in \mathbb{R}^{300}$、表情参数 $\boldsymbol{\psi} \in \mathbb{R}^{100}$、下颚参数 $\boldsymbol{\theta}_{jaw} \in \mathbb{R}^3$ 和全局头部参数 $\boldsymbol{\theta}_{global} \in \mathbb{R}^3$。本研究选用其中的表情参数 $\boldsymbol{\psi}$ 和下颚参数 $\boldsymbol{\theta}_{jaw}$ 训练模型，生成面部动作序列，再结合面部形状参数β和全局头部参数 $\boldsymbol{\theta}_{global}$ 生成完整的人脸视频。其中将全局头部参数 $\boldsymbol{\theta}_{global}$ 设置为0，以简化实验条件。

TA-MEAD：TA-MEAD数据集来自TalkCLIP论文（Ma等，2023），是一个用于描述MEAD视频中人脸表情与动作的文本数据集。TA-MEAD为每个视频提供一句随机组合成的描述文本，文本格式为[<SUB> <EMOTION> speaks with <AU>]，其中<SUB>表示描述对象，<EMOTION>表示情感类别和强度，<AU>表示面部动作信息。由于<AU>部分生成的面部动作信息内容较为随机，会影响模型学习与生成过程的稳定性，本实验中将<AU>部分去除，使文本专注于描述情感，以实现情感与人脸动作的解耦。

3.2 实验设置

本研究将TA-MEAD数据集中的文本描述，与

3DMEAD数据集的3D人脸参数、MEAD数据集的语音数据共同用于模型的训练。为优化模型训练和评估效果，本文对数据集做了预处理，将各数据集的文本、语音和3D人脸参数一一对应，并将语音和3D人脸参数按时间对齐，最终处理后的数据集包含46位受试者的19，371个数据。处理完成后，数据集按比例随机划分为训练集和测试集，其中80%用于模型训练，剩余20%用于测试与评估，以确保模型的泛化能力并提升性能评估的可靠性。

本文中的模型使用单张NVIDIA 4090 GPU进行训练，第一阶段共训练800轮，耗时约16小时。第二阶段共训练200轮，耗时约18小时。为验证提出方法的有效性，本模型与近年来发表的其他主要3D人脸生成模型进行了对比实验，从生成3D人脸的重建精度、情感表达的准确性及多样性等方面对模型进行定量和定性评估。为了评估本模型不同组件对生成效果的影响，本文进行了对应的消融实验。

表1　定量分析结果

Table 1　The results for quantitative analysis

模型	MVE↓ ($\times10^{-3}mm$)	LVE↓ ($\times10^{-4}mm$)	FDD↓ ($\times10^{-4}mm$)	MEE↓ ($\times10^{-4}mm$)	CE↓ ($\times10^{-4}mm$)	Diversity↑ ($\times10^{-3}mm$)
FaceFormer	N/A	0.616 8	N/A	N/A	N/A	N/A
CodeTalker	N/A	1.624 3	N/A	1.291 8	1.271 3	**1.290 1**
FaceDiffuser	N/A	**0.488 7**	N/A	N/A	N/A	N/A
ProbeTalk3D	<u>0.715 7</u>	0.624 5	**0.005 1**	<u>0.560 5</u>	<u>0.532 5</u>	<u>0.357 6</u>
本模型	**0.673 8**	<u>0.586 7</u>	<u>0.155 4</u>	**0.537 4**	**0.488 6**	0.254 7

注：加粗字体为最优值，下划线字体为次优值。其中，N/A表示FaceFormer、CodeTalker、FaceDiffuser由于没有情感输入，生成的人脸没有明显表情，因而MVE和FDD指标与本模型相差巨大，不作比较。

表2　消融实验结果

Table 2　The results for the ablation study

模型	MVE↓ ($\times10^{-3}mm$) ($\times 10^{-3}$mm)	LVE↓ ($\times10^{-4}mm$)	FDD↓ ($\times10^{-4}mm$)	MEE↓ ($\times10^{-4}mm$)	CE↓ ($\times10^{-4}mm$)	Diversity↑ ($\times10^{-3}mm$)
完整模型	**0.673 8**	**0.586 7**	**0.155 4**	**0.537 4**	**0.488 6**	0.254 7
无s、t	1.030 5	1.060 0	0.846 5	1.000 9	0.867 6	**0.341 1**
无Q_{top}	0.711 8	0.670 7	0.202 1	0.643 0	0.605 4	0.131 5
无s、t、Q_{top}	1.041 7	1.079 6	0.577 5	1.048 9	0.970 5	0.160 0

注：加粗字体为最优值。

3.3　对比实验

3.3.1　定量分析

在定量分析中，采用多组指标定量评估模型的重建质量和情感生成效果。平均顶点误差、嘴唇顶点误差和上脸动态偏差基于单个样本输出计算，用于评估确定性模型的指标。平均估计误差、覆盖率误差和多样性在多个样本输出上计算，用于评估非确定性模型的指标。

平均顶点误差(Mean Vertex Error，MVE)计算

预测帧的人脸顶点与真实值之间的平均欧氏距离，以评估面部重建的整体精度。N表示所有测试集的总帧数，$\boldsymbol{x}_i$为第i帧的真实值，$\hat{\boldsymbol{x}}_i$为预测帧。MVE的计算公式为：

$$\mathrm{MVE}=\frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}}\|\boldsymbol{x}_i-\hat{\boldsymbol{x}}_i\| \tag{17}$$

唇部顶点误差(Lip Vertex Error，LVE)计算预测帧的唇部区域顶点与真实值之间的最大L2误差，

然后对所有帧求平均,以评估生成口型动作的精度。N表示所有测试集的总帧数,$\boldsymbol{x}_{\text{lip}}^{\text{i}}$表示真实帧的唇部区域顶点,$\hat{\boldsymbol{x}}_{\text{lip}}^{\text{i}}$表示预测帧的唇部区域顶点。LVE的计算公式为:

$$\text{LVE}=\frac{1}{\text{N}}\sum_{\text{i}=1}^{\text{N}}\max\|\boldsymbol{x}_{\text{lip}}^{\text{i}}-\hat{\boldsymbol{x}}_{\text{lip}}^{\text{i}}\|_2 \tag{18}$$

上脸动态偏差(Upper Face Dynamic Deviation, FDD)计算上脸区域的面部动态变化与真实值的偏差,以评估其接近程度。A表示测试集的总数据量,$\boldsymbol{M}_{\text{upper}}$表示真实帧的上脸区域顶点运动,$\hat{\boldsymbol{M}}_{\text{upper}}$表示预测帧的上脸区域顶点运动,人脸模板由数据集的平均人脸顶点得到,dyn(·)表示沿时间的L2误差。FDD的计算公式为:

$$\text{FDD}=\frac{1}{\text{A}}\sum_{\text{i}=1}^{\text{A}}\|\text{dyn}(\boldsymbol{M}_{\text{upper}})-\text{dyn}(\hat{\boldsymbol{M}}_{\text{upper}})\| \tag{19}$$

平均估计误差(Mean Estimate Error, MEE)(Yang等,2024)评估模型的采样分布的均值与真实值的接近程度。对每个数据生成10个样本并计算这些样本的均值$\text{E}(\hat{\boldsymbol{x}})$,再对所有测试数据的MEE取平均。MEE越小表示模型能更有效地生成接近真实的唇部运动。MEE的计算公式为:

$$\text{MEE}=\text{LVE}(\boldsymbol{x},\text{E}(\hat{\boldsymbol{x}})) \tag{20}$$

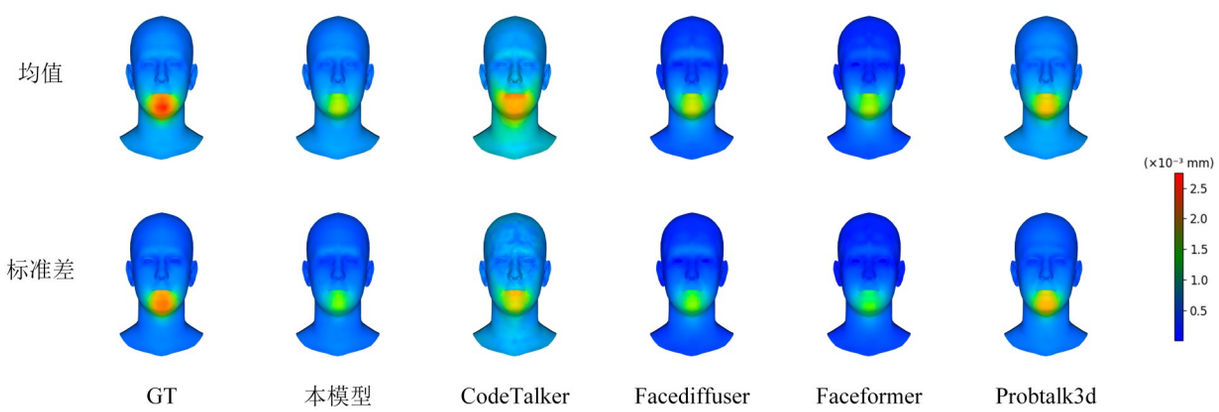

图3 平均运动比较

Fig. 3 The comparisons on mean motions

覆盖误差(Coverage Error, CE)评估模型的采样分布与真实值的接近程度。通过与MEE类似的方式,生成包含10个样本的$\boldsymbol{S}$,并在真实值和生成样本之间取LVE计算的最小值,再对所有测试数据的CE取平均。CE越小表示模型的预测更好地覆盖了真实样本的唇部运动。CE的计算公式为:

$$\text{CE}=\min_{\hat{\boldsymbol{x}}\in\boldsymbol{S}}\text{LVE}(\boldsymbol{x},\text{E}(\hat{\boldsymbol{x}})) \tag{21}$$

多样性(Diversity)(Ren等,2023)评估人脸生成的多样性。与MEE和CE类似,对每个测试数据生成10个样本,对于第i个数据,将生成的10个样本平分为$\boldsymbol{S}_1$和$\boldsymbol{S}_2$两个子集,计算各子集第j个样本间的平均欧氏距离,再对所有测试数据的Diversity取平均。Diversity的计算公式为:

$$\text{Diversity}=\frac{1}{\text{A}\times\text{B}}\sum_{\text{i}=1}^{\text{A}}\sum_{\text{j}=1}^{\text{B}}\|(\hat{\boldsymbol{x}}_{\text{i,j}}\in\boldsymbol{S}_1)-(\hat{\boldsymbol{x}}_{\text{i,j}}\in\boldsymbol{S}_2)\|_2 \tag{22}$$

将本模型与FaceFormer、CodeTalker、FaceDiffuser(Stan等,2023)、ProbeTalk3D进行对比。实验结果如表1所示。

本模型的MVE相较于同样引入情感输入的ProbeTalk3D降低了5.9%。这表明情感特征的显式建模能有效提升面部重建的几何一致性,特别是在表达复杂情感时保持面部结构的准确性。

在唇部相关指标上,虽然本模型LVE为次优,但较最优的FaceDiffuser存在16.7%的差距,相较于FaceFormer提升了4.9%,相较于ProbeTalk3D提升了6.1%,提升也有限。这可能是由于分层解耦的方法虽然对情感条件进行了解耦,但同时也导致对嘴部区域的解耦效果有限,使嘴部区域在细节表现上有所不足。

本模型的FDD指标远不及ProbeTalk3D,差异主要源于情感生成任务中眉毛、眼睑等区域的运动幅度,与标准模板的中性表情产生动态偏差,反映了情

感表达的强度特征。原因可能是由于文本条件中去掉了<AU>信息，使模型无法学习嘴部区域以外的更多面部动作细节。

本模型在MEE和CE指标上均达到最优，相较于ProbeTalk3D，MEE指标降低了4.1%，CE指标降低了8.2%，相较于CodeTalker，MEE指标降低了58.4%，CE指标降低了61.6%。这表明通过引入分层解耦的方法，模型能有效提升生成样本嘴部动作和人脸整体动作的分布集中度和对真实样本的覆盖能力，从而使生成的人脸动作更加接近真实人脸。

本模型的多样性得分最低，相较于ProbeTalk3D和CodeTalker都存在巨大差距。反映出分层解耦在生成过程中保证语义一致性的同时，也对生成人脸的多样性产生了一定的约束。这种准确性与多样性之间的权衡符合情感驱动动画的应用需求，即在保持情感表达准确性的前提下允许适度的动作变化。一方面，情感的准确表达能够使生成结果更加可控，但在另一方面，多样性过低也会使生成的人脸动画显得较为生硬。

从LVE、多样性等指标效果有限以及MVE、CE等指标的优秀表现来看，本模型使用的条件引导与分层解耦方法在提升生成结果的稳定性的同时，也会对生成的具体细节和生成的多样性方面有负面影响。为此，或许可以通过优化解耦方法，将其与条

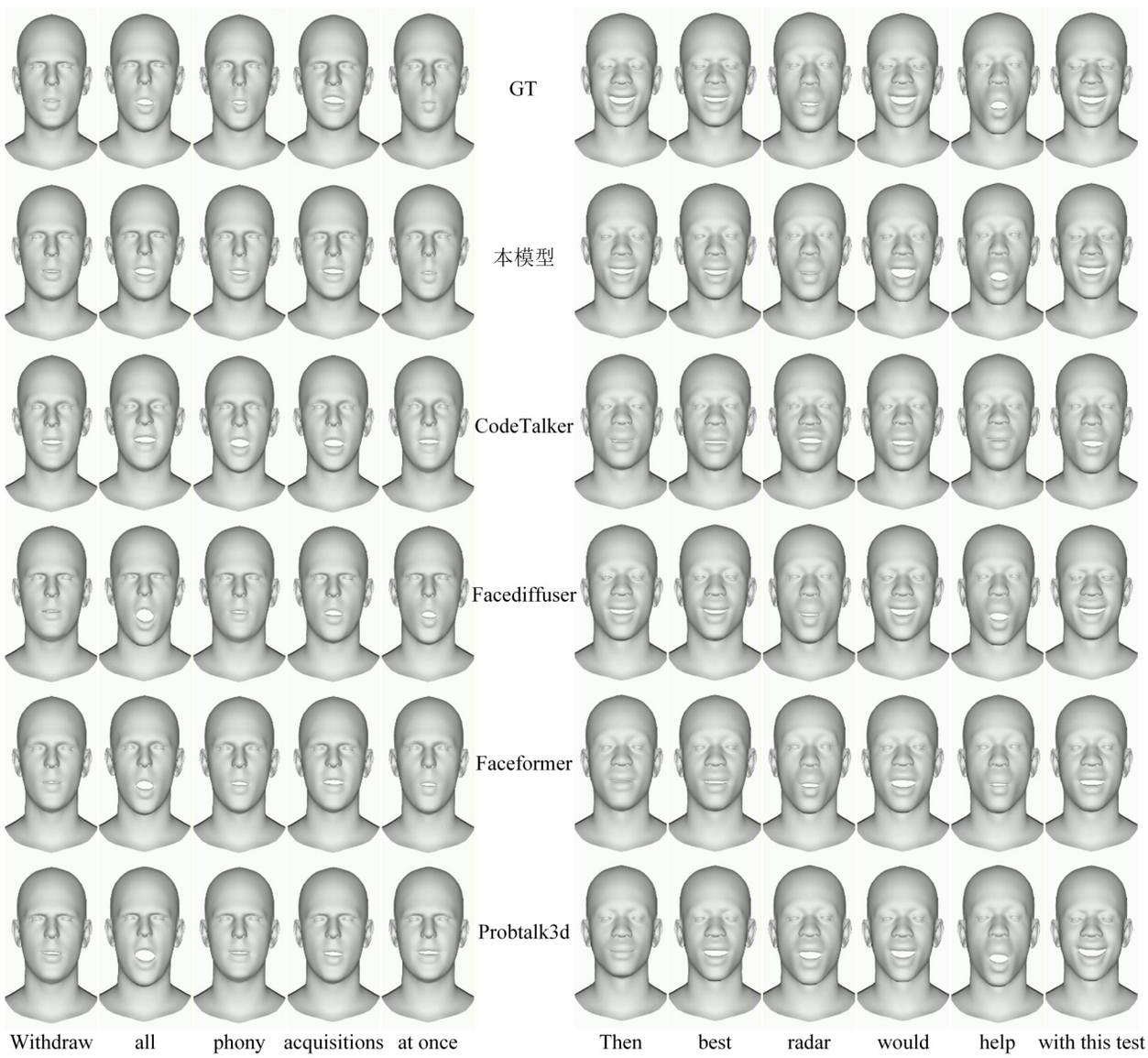

图4　人脸视频帧的视觉比较

Fig. 4　The visual comparison of facial video frames

件引导有机结合来缓解这个问题，平衡生成结果的准确性与多样性，实现更加灵活的控制与生成。

传统方法由于缺乏情感输入，在MVE、FDD等指标上出现数量级差异，表明情感特征对3D人脸重建的重要性。ProbeTalk3D虽在FDD和Diversity指标上表现优异，但在生成人脸的准确性上相较于本模型有所欠缺。本模型通过引入分层解耦和多模态协同建模，在保持高精度重建的同时，实现了更优的分布估计能力。

3.3.2　可视化分析

在可视化分析中，将不同模型生成的面部动作与真实的面部动作（GT）进行比较。通过热图（heatmap）比较不同模型生成的动作与真实动作整体上的差异，并通过视频截图直接对比不同模型生成的视频与真实视频的差异。

通过热图可视化不同模型生成的3D人脸顶点的平均运动，并与真实值进行比较。给定音频序列，内容为“Who authorized the unlimited expensive account?”（谁授权了无限费用账户？），情感为愤怒，生成的平均运动对比如图3所示。图3中展示了各模型生成的3D人脸顶点热图的均值和标准差，并与真实值生成的热图进行对比，深蓝色表示平均运动较少，亮红色表示平均运动较多。

从图3中可以看到，FaceDiffuser和FaceFormer由于缺乏情感输入，面部动作较小，只在嘴部有明显动作。CodeTalker可能是由于缺乏情感输入，生成的面部动作一直在抽动。而本模型与ProbeTalk3D相比，虽然嘴部动作没有ProbeTalk3D模型明显，但整体动作上更接近真实人脸。

图4展示不同模型生成的人脸视频与真实人脸视频对应视频帧的可视化比较，其中给定音频序列内容为“Withdraw all phony acquisitions at once”（立即撤回所有虚假收购）和“Then best radar would help with this test”（那么最好的雷达将有助于这项测试），情感分别为愤怒和快乐。每一行展示了特定模型生成视频对应的一些关键视频帧。

从图4中可以看到，CodeTalker、FaceDiffuser和FaceFormer这三个模型生成的视频缺少真实的面部表情，只在嘴部区域有明显动作，而CodeTalker的嘴部区域一直在抽搐。本模型生成的人脸视频在整体面部表情上相较于其他模型生成的人脸视频更加真实和生动，而与FaceDiffuser、FaceFormer和ProbeTalk3D的嘴部动作相比，也更加贴近真实人脸的嘴部动作，但缺点在于嘴部动作幅度较小，导致不够真实。

3.4　消融实验

为了评估模型中不同组件对生成效果的具体贡献，本文通过消融实验，逐步移除模型中的关键模块来分析各模块对性能的影响。消融实验主要对外部条件（身份向量s和文本描述t）以及内部条件（顶层特征Q_{top}）进行消融，以测试它们对人脸重建和生成效果的作用。消融实验包含以下几种模型设置：

完整模型：包含外部条件以及内部条件，用于基准对比，作为模型最佳性能的参考。

无外部条件：移除身份向量s和文本描述t，以评估外部条件在生成多样性和情感控制上的作用。

无内部条件：移除顶层特征Q_{top}，以观察顶层特征在提升人脸特征重建精度和可控性方面的贡献。

无外部条件和内部条件：同时移除外部和内部条件，仅依赖底层特征Q_{bottom}，以评估条件解耦的作用。

通过这些设置，分析了各组件对模型性能的影响，实验结果如表2所示。

移除身份向量和文本描述后，MVE指标上升了52.9%，LVE指标上升了80.7%，表明模型的重建精度降低，可知身份信息与文本描述的约束对保持面部重建精度具有关键作用。另一方面FDD上升54.7%，表明生成人脸的表情动作也有了失真，证明了外部条件能有效控制上脸区域的动态情感表达。此外，Diversity增加33.9%，说明移除约束后生成的自由度有了提升。同时MEE增加86.2%，CE增加77.6%，反映出人脸生成的准确性大幅下降。

移除顶层特征后，MVE指标上升5.6%，LVE指标上升14.3%，说明模型的细节重建能力有了一定下降，证明顶层特征的加入提升了人脸的生成精度。FDD上升30.1%，表明顶层特征对表情动作的生成同样具有调节作用。Diversity骤降了48.4%，多样性严重衰减，揭示分层解耦对生成多样性的重要作用。

当同时移除外部条件和顶层特征时，所有指标均达最差水平，说明模型性能严重下降。较完整模型MEE上升95.2%，CE上升98.6%，证实底层特征单独使用时准确性大幅下降。FDD显著高于仅移除

顶层特征的场景，说明外部条件与内部条件共同构建了对抗无效面部运动的双重约束机制。

通过消融实验，明确了模型的各部分对人脸生成能力的不同作用。外部条件确保了面部表情与动作的准确性，顶层特征则强化细节重建与动态控制，二者通过互补机制共同实现了稳定的情感表达。此外，Diversity指标与MEE和CE指标的负相关表明，生成人脸的多样性与准确性不可能同时实现，需通过多方面机制实现动态调节。

4 结 论

本研究提出了一种基于分层解耦的情感可控3D说话人脸生成方法，通过在VQ-VAE框架下将人脸特征分解为顶层和底层特征，实现了人脸重建质量的提升与情感控制能力的增强。模型创新性地引入身份向量和描述文本作为外部条件，顶层人脸特征作为内部条件，逐层解耦人脸特征，从而在面部重建的准确性与情感表达的稳定性上取得显著效果。本方法的优势在于在人脸生成过程中的情感可控，使用简单方便，生成结果准确且稳定，在人机交互、数字人、动画制作等方面有较好的应用价值。

实验结果表明，所提出的方法在MVE、LVE、FDD等重建精度指标上表现优异，且在MEE、CE和Diversity等指标中表现出较高的稳定性，验证了模型在多模态信息融合下生成面部表情动作的能力。通过消融实验，进一步验证了各模块的有效性，表明顶层与底层特征的分层解耦，以及外部与内部条件的协同作用，对提升情感表达的准确性和稳定性具有关键作用。

尽管本研究在情感可控3D说话人脸生成方面取得了良好的效果，但仍存在一些有待解决的问题。最主要的是，本模型生成的人脸在嘴部动作上仍有缺陷，动作幅度相较于真实人脸的嘴部动作不够明显。其次，尽管当前模型实现了对人脸特征的分层解耦，但仅限于简单地分为顶层和底层特征，未来可以进一步探索如何在解耦的基础上更清晰地实现特征维度的差异化引导。最后，在情感表达方面，情感特征的多样性与细腻度有待进一步提升。

参考文献（References）

Baevski A, Zhou Y, Mohamed A and Auli M. 2020. Wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in Neural Information Processing Systems, 33: 12449-12460 [DOI: 10.5555/3495724.3495884].

Chen L, Bao W, Lei S, Fu J, Liu S and Li X. 2023. AdaMesh: Personalized facial expressions and head poses for speech-driven 3D facial animation [EB/OL]. [2023-10-12]. https://arxiv.org/abs/2310.07236.

Chen Y, Zhao J and Zhang W Q. 2023. Expressive speech-driven facial animation with controllable emotions. 2023 IEEE International Conference on Multimedia and Expo Workshops (ICMEW), Melbourne, Australia: 387-392 [DOI: 10.1109/ICMEW57084.2023.00101].

Chu Z, Guo K, Xing X, Cao C and Zhang X. 2024. CorrTalk: Correlation between hierarchical speech and facial activity variances for 3D animation. IEEE Transactions on Circuits and Systems for Video Technology, 34(9): 8953-8965 [DOI: 10.1109/TCSVT.2023.3327584].

Cudeiro D, Bolkart T, Laidlaw C, Ranjan A and Black M J. 2019. Capture, learning, and synthesis of 3D speaking styles. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA: 10101-10111 [DOI: 10.1109/CVPR.2019.01035].

Daněček R, Chhatre K, Tripathi S, Gupta R, Wang J and Wang H. 2023. Emotional speech-driven animation with content-emotion disentanglement. SIGGRAPH Asia 2023 Conference Papers (SA '23), Sydney, Australia: 41 [DOI: 10.1145/3610547.3618171].

Deng Y, Yang J L, Xu S C, Chen D, Jia Y D and Tong X. 2019. Accurate 3D face reconstruction with weakly-supervised learning: From single image to image set. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, Long Beach, CA, USA [DOI: 10.1109/cvprw.2019.00038].

Fan Y, Lin Z, Saito J, Lucey P, Wu Y and Song L. 2022. Faceformer: Speech-driven 3D facial animation with transformers. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA: 18770-18780 [DOI: 10.1109/CVPR52688.2022.01821].

Gan Y, Yang Z, Yue X, Ma C, Zhang Y and Zhou H. 2023. Efficient emotional adaptation for audio-driven talking-head generation. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France: 22634-22645 [DOI: 10.1109/ICCV.2023.02093].

Gao X, Liu D Y and Zhang J Y. 2024. Multi-modal digital human modeling, synthesis, and driving: a survey. Journal of Image and Graph-

ics, 29(09): 2494-2512 (高玄, 刘东宇, 张举勇. 2024. 多模态数字人建模、合成与驱动综述. 中国图象图形学报, 29(09): 2494-2512) [DOI: 10.11834/jig.230649]

Han T, Gui S, Huang Y, Luo X and Zhang W. 2023. PMMTalk: Speech-driven 3D facial animation from complementary pseudo multi-modal features [EB/OL]. [2023-12-05]. https://arxiv.org/abs/2312.02781.

Hsu W N, Bolte B, Tsai Y H H, Lakhotia K, Salakhutdinov R and Mohamed A. 2021. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29: 3451-3460 [DOI: 10.1109/TASLP.2021.3122291]

Ji X, Zhou H, Wang K, Wang S and Li Z. 2022. EAMM: One-shot emotional talking face via audio-based emotion-aware motion model. ACM SIGGRAPH 2022 Conference Proceedings (SIGGRAPH '22), Vancouver, Canada: 61 [DOI: 10.1145/3528233.3530727].

Jiang L, Yu Z, Wang P F, Zhou D S and Hou Y Q. 2022. Survey of audio-driven cross-modal visual generation algorithms. Journal of Graphics, 43(2): 181-188 (姜莱, 于震, 王鹏飞, 周东生, 侯亚庆. 2022. 音频驱动跨模态视觉生成算法综述. 图学学报, 43(2): 181-188) [DOI: 10.11834/jig.20220203].

Kim J, Cho J, Park J, Lee H, Choi J and Kim S. 2024. DEEPTalk: Dynamic emotion embedding for probabilistic speech-driven 3D face animation [EB/OL]. [2024-08-06]. https://arxiv.org/abs/2408.06010.

Kingma D P and Welling M. 2013. Auto-encoding variational bayes [EB/OL]. [2013-12-20]. https://arxiv.org/abs/1312.6114.

Li T, Bolkart T, Black M J, Li H and Romero J. 2017. Learning a model of facial shape and expression from 4D scans. ACM Transactions on Graphics, 36(6): 194 [DOI: 10.1145/3130800.3131041].

Liang B, Pan Y, Guo Z, Yang J, Wang J and Zhou H. 2022. Expressive talking head generation with granular audio-visual control. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA: 3387-3396 [DOI: 10.1109/CVPR52688.2022.00339].

Liu F, Zhang K B, Yang Q, Zhou S B, Wang Y L and Sun Z N. 2024. 3D face imaging and reconstruction technology: a review. Journal of Image and Graphics, 29(09): 2441-2470 (刘菲, 张堃博, 杨青, 周树波, 王云龙, 孙哲南. 2024. 三维人脸成像及重建技术综述. 中国图象图形学报, 29(09): 2441-2470) [DOI: 10.11834/jig.230697]

Ma Y, Wang S, Hu Z, Zhang J and Li B. 2023. StyleTalk: One-shot talking head generation with controllable speaking styles. Proceedings of the AAAI Conference on Artificial Intelligence, 37(2): 1896-1904 [DOI: 10.1609/aaai.v37i2.25239].

Ma Y, Wang S, Ding Y, Liu Y and Chen X. 2023. TalkCLIP: Talking head generation with text-guided expressive speaking styles [EB/OL]. [2023-04-01]. https://arxiv.org/abs/2304.00334.

Pan Y, Li S X, Tan S, Wei J J, Zhai G T and Yang X K. 2025. Advancements in digital character stylization, multimodal animation, and interaction. Journal of Image and Graphics, 30(02): 0334-0360 (潘烨, 李韶旭, 谭帅, 韦俊杰, 翟广涛, 杨小康. 2025. 数字人风格化、多模态驱动与交互进展. 中国图象图形学报, 30(02): 0334-0360) [DOI: 10.11834/jig.230639]

Peng Z Q, Wu H Y, Song Z B, Xu H, Zhu X Y, He J, Liu H Y and Fan Z X. 2023. EmoTalk: Speech-Driven Emotional Disentanglement for 3D Face Animation. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France: 20687-20697 [DOI: 10.1109/ICCV51070.2023.01894]

Radford A, Kim J W, Hallacy C, Ramesh A and Goh G. 2021. Learning transferable visual models from natural language supervision. International Conference on Machine Learning, PMLR: 8748-8763 [DOI: 10.5555/3524938.3525037].

Razavi A, Van Den Oord A and Vinyals O. 2019. Generating diverse high-fidelity images with VQ-VAE-2. Advances in Neural Information Processing Systems, 32: 14866-14876 [DOI: 10.5555/3454287.3455027].

Ren Z, Pan Z, Zhou X, Wang H, Zhang M and Zhao L. 2023. Diffusion motion: Generate text-guided 3D human motion by diffusion model. ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP): 1-5 [DOI: 10.1109/ICASSP49357.2023.10094791].

Richard A, Zollhöfer M, Wen Y, Engelhardt T and Theobalt C. 2021. MeshTalk: 3D face animation from speech using cross-modality disentanglement. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, Canada: 1173-1182 [DOI: 10.1109/ICCV48922.2021.00122].

Sohn K, Lee H and Yan X. 2015. Learning structured output representation using deep conditional generative models. Advances in Neural Information Processing Systems, 28: 3483-3491 [DOI: 10.5555/2969239.2969426].

Song Y F, Zhang W, Chen Z N, Zhao N and Li H. 2023. Survey of digital talking-head video generation. Journal of Computer-Aided Design and Computer Graphics, 35(10): 1457-1468 (宋一飞, 张炜, 陈智能, 赵宁, 李华. 2023. 数字说话人视频生成综述. 计算机辅助设计与图形学学报, 35(10): 1457-1468) [DOI: 10.3724/SP.J.1089.2023.26579].

Stan S, Haque K I and Yumak Z. 2023. FaceDiffuser: Speech-Driven 3D Facial Animation Synthesis Using Diffusion. Proceedings of the 16th ACM SIGGRAPH Conference on Motion, Interaction and Games (MIG '23). New York, NY, USA: 1-11 [DOI: 10.1145/3623264.3624447]

Sun Y, Chu W, Zhou H, Yang C, Zhang H and Xu W. 2024. AVI-Talking: Learning audio-visual instructions for expressive 3D talking face generation. IEEE Access, 12: 57288-57301 [DOI: 10.

1109/ACCESS.2024.3391741].

Sun Z, Wen Y H, Lv T, Liu S and Zhou M. 2024. Continuously controllable facial expression editing in talking face videos. IEEE Transactions on Affective Computing, 15(3): 1400-1413 [DOI: 10.1109/TAFFC.2022.3225674].

Van Den Oord A, Vinyals O and Kavukcuoglu K. 2017. Neural discrete representation learning. Advances in Neural Information Processing Systems, 30: 6306-6315 [DOI: 10.5555/3295222.3295309].

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser Ł and Polosukhin I. 2017. Attention is all you need. Advances in Neural Information Processing Systems, 30: 5998-6008 [DOI: 10.5555/3295222.3295349].

Wang D, Deng Y, Yin Z, Zhou H, Lin Y and Zhu C. 2023. Progressive disentangled representation learning for fine-grained controllable talking head synthesis. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, Canada: 17979-17989 [DOI: 10.1109/CVPR52729.2023.01755].

Wang K, Wu Q, Song L, Zhu Y and Zhu C. 2020. MEAD: A large-scale audio-visual dataset for emotional talking-face generation. European Conference on Computer Vision, Glasgow, UK: Springer International Publishing: 700-717 [DOI: 10.1007/978-3-030-58583-9_42].

Wu S, Haque K I and Yumak Z. 2024. ProbTalk3D: Non-deterministic emotion controllable speech-driven 3D facial animation synthesis using VQ-VAE [EB/OL]. [2024-09-10]. https://arxiv.org/abs/2409.07966.

Xing J, Xia M, Zhang Y, Zhou H and Wu X. 2023. Codetalker: Speech-driven 3D facial animation with discrete motion prior. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, Canada: 12780-12790 [DOI: 10.1109/CVPR52729.2023.01238].

Xu C, Zhu J, Zhang J, Liu Y and Wang H. 2023. High-fidelity generalized emotional talking face generation with multi-modal emotion space learning. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, Canada: 6609-6619 [DOI: 10.1109/CVPR52729.2023.00645].

Yang K D, Ranjan A, Chang J H R, Liu L, Wang K and Black M J. 2024. Probabilistic speech-driven 3D facial motion synthesis: New benchmarks, methods, and applications. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA: 27294-27303 [DOI: 10.1109/CVPR59830.2024.02749].

Ye Z, Sun Z, Wen Y H, Zhou S and Zhang F. 2022. Dynamic neural textures: Generating talking-face videos with continuously controllable expressions [EB/OL]. [2022-04-13]. https://arxiv.org/abs/2204.06180.

Zhao W, Wang Y, He T, Liang J and Liu L. 2023. Breathing life into faces: Speech-driven 3D facial animation with natural head pose and detailed shape [EB/OL]. [2023-10-28]. https://arxiv.org/abs/2310.20240.

Zhen R, Song W, He Q, Gao Z and Liu H. 2023. Human-computer interaction system: A survey of talking-head generation. Electronics, 12(1): 218 [DOI: 10.3390/electronics12010218].

Zhong Y, Wei H, Yang P, Zhang S and Lu X. 2024. ExpCLIP: Bridging text and facial expressions via semantic alignment. Proceedings of the AAAI Conference on Artificial Intelligence, 38(7): 7614-7622 [DOI: 10.1609/aaai.v38i7.29549].

作者简介

陈胜，男，硕士研究生，主要研究方向为情感生成。E-mail: 1936580290@qq.com

孙强，通信作者，男，副教授，主要研究方向为情感计算与智能交互。E-mail：qsun@xaut.edu.cn

朱霞天，男，高级讲师，主要研究方向是可扩展机器学习、计算机视觉、多模态GenAI。E-mail：xiatian.zhu@surrey.ac.uk