计 算 机 研 究 与 发 展
Journal of Computer Research and Development
https://crad.ict.ac.cn
63(6): 1451−1477, 2026

基于深度学习的 2D 虚拟人驱动技术综述

杨璐嘉　周溪石　王俊超　张金山　金　涛　尹建伟
（浙江大学软件学院　浙江宁波　315106）
（yanglujia@zju.edu.cn）

Survey of 2D Virtual Human Driving Technology Based on Deep Learning

Yang Lujia, Zhou Xishi, Wang Junchao, Zhang Jinshan, Jin Tao, and Yin Jianwei
（*School of Software Technology*, *Zhejiang University*, *Ningbo*, *Zhejiang* 315106）

Abstract　In recent years, the rapid advancement of deep learning and multimodal information fusion has significantly propelled the development of 2D virtual human driving technologies. Existing methods can generate highly realistic facial expressions and body movements for specified characters by leveraging diverse input data such as video, audio, facial cues, and poses. This remarkable progress offers substantial convenience for applications in virtual entertainment, online education, and intelligent interaction. This survey provides a systematic review of the current state and evolutionary trajectory of deep learning-based 2D virtual human driving methodologies. It thoroughly outlines the underlying principles and fundamental models commonly employed in this domain. The driving techniques are categorized and summarized according to both the target driving components (e.g., face, body) and the corresponding architectural frameworks, with a focused comparison and detailed elaboration of prominent methods developed in recent years. The research status and development prospects of holistic driving systems are also introduced. Furthermore, this survey delves into the critical technical challenges currently faced by the field, including generation realism, the quality of datasets, and the real-time performance of models. The aim of this survey is to offer a comprehensive technological overview and perspective for researchers in related fields, thereby contributing to the further advancement of this domain.

Key words　virtual digital human; talking head generation; multi-modal fusion; video generation; deep learning

摘　要　近年来，随着深度学习和多模态信息融合技术的快速发展，现有 2D 虚拟人驱动技术能够根据输入的视频、语音、表情和姿态等信息，驱动指定人物生成逼真的表情和动作，为虚拟娱乐、在线教育和智能交互等场景带来了极大的便利。综述系统地梳理了基于深度学习的 2D 虚拟人驱动技术的发展现状，概括了其原理与所采用的基础模型。同时，根据驱动部位及实现架构的不同，对驱动技术进行分类与概述，重点阐述并对比了近年来的主要方法，介绍了集成驱动的研究状况与发展前景。此外，还深入探讨了当前研究面临的生成逼真性、数据集质量、模型实时性等方面的技术挑战，旨在为相关领域研究人员提供较为全面的技术梳理与展望，帮助推动该领域的进一步发展。

关键词　虚拟数字人；说话人生成；多模态融合；视频生成；深度学习

收稿日期：2025-07-30；**修回日期**：2025-12-24

基金项目：国家重点研发计划项目（2022YFF0902005）；浙江省尖兵重点研发计划项目（2023C01002）；杭州市重大研发计划项目（2022AIZD0140）；甬江引才项目（2022A-236-G）

This work was supported by the National Key Research and Development Program of China (2022YFF0902005), the Key Research and Development Jianbing Program of Zhejiang Province (2023C01002), the Hangzhou Major Project and Development Program (2022AIZD0140), and the Yongjiang Talent Introduction Program (2022A-236-G).

通信作者：张金山（zhangjinshan@zju.edu.cn）

中图法分类号 TP391.41

DOI: 10.7544/issn1000-1239.202550560 CSTR: 32373.14.issn1000-1239.202550560

近年来,随着深度学习(deep learning)的快速发展,生成式人工智能(generative artificial intelligence)领域取得了突破性的进展,在文本生成、语音生成、图像生成以及视频生成等领域上都展现出了优秀的生成效果,为生活带来了翻天覆地的变化。虚拟人生成技术(virtual human generation technology)是指利用计算机技术和人工智能算法创建数字化、拟人化的虚拟角色,使其具备与人类相似的外貌、行为、语言和情感等特征的生成技术。这一技术结合了多个学科的研究成果,旨在生成具备丰富表情、生动姿态与细腻运动细节的自然虚拟人,进而提升用户体验,同时满足各类应用场景的需求。根据表现形式的不同,虚拟数字人可以分为3D和2D虚拟人。其中2D虚拟人是指以2D图像或视频形式呈现,用于人机交互、虚拟内容生成等场景的拟人化数字角色。与3D虚拟人不同,2D虚拟人不以完整的3D网格建模为基础,通常更侧重于生成视觉上自然的表情、姿态和动作动画,具有建模简单、渲染成本低等优势。值得注意的是,近年来随着对生成质量和可控性要求的提高,越来越多的2D虚拟人驱动方法引入了3D建模的信息表示,如3D可变形模型(3D morphable model, 3DMM)、面部关键点、深度图等。这些3D信息作为中间特征在视频生成流程中发挥了关键作用,有效提升了动作一致性、细节保真度与身份保持能力,已逐步成为当前2D虚拟人驱动的主流技术路径。由于不需要复杂的3D建模,2D虚拟人制作成本往往低于3D虚拟人,多用于动画短片、独立游戏和虚拟主播的制作。2D虚拟人的外观生成经历了从基于传统图形学[1]、基于生成对抗网络(generative adversarial networks, GANs)[2]到基于扩散模型(diffusion model)[3-4]的技术发展,在效果上已经实现了高质量、多风格的虚拟人像生成。当下2D虚拟人的研究重点也更多地聚焦于如何驱动人物,生成自然的面部表情、头部运动和身体姿态上来[5-6]。2D虚拟人驱动任务示意图如图1所示,可以分为面部驱动技术和姿态驱动技术。

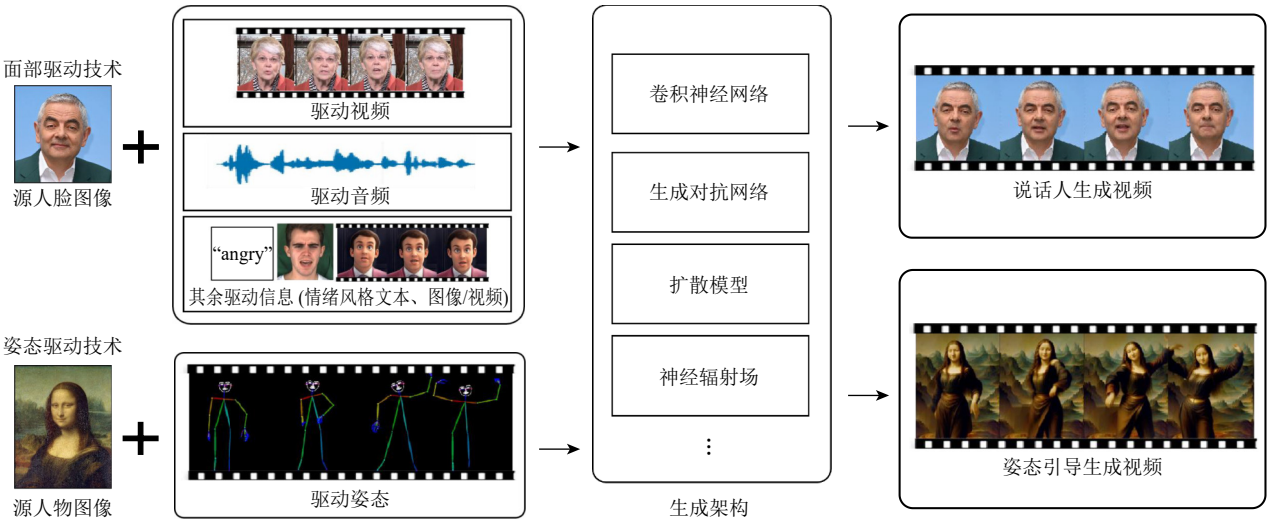


Fig. 1 2D virtual human-driven task

图1 2D虚拟人驱动任务

2D虚拟人面部驱动技术的质量对于实现虚拟人的生动表现至关重要,是相关领域的热门研究任务,也被称为说话人生成任务(talking head generation task)。这一任务要求在给定目标人物图像和驱动信息(一段音频、视频、文本或者情绪提示等)的情况下,从中提取特征,整合生成与驱动数据相匹配的目标人物视频。本质上是多模态的提取、扩充、对齐和融合任务。该任务的难点是生成与给定源人物身份背景一致、与驱动数据信息一致、自然连贯无伪影的高质量说话人视频。而根据驱动部位与驱动信号之间的相关度,2D虚拟人面部驱动又可以细分为强相关的唇形驱动,以及弱相关的面部表情与头部姿势驱动。2D虚拟人身体运动的驱动也称为姿态引导生成任务(pose-guided generation task),根据驱动信

息可以分为图像到图像、视频到视频、语音到视频和文本辅助4类驱动，目前大多数的姿态引导生成任务都采用骨骼图作为姿态表征，基于扩散模型生成视频序列。

本文综合回顾了近年来2D虚拟人驱动技术的发展，对不同方法进行了分类和概述，旨在清晰展现该领域的技术脉络。

1 虚拟人驱动原理与基础模型

虚拟人驱动生成的核心思想是将静态的外观信息（如2D图像或3D模型）动画化，使其动作动态与输入的驱动信号源（如视频、音频或其他条件）相匹配。这一过程通常建模为条件生成任务，通过深度学习模型（如卷积神经网络、Transformer[7]、生成对抗网络[2]或扩散模型[3-4]等）生成连续的动作序列，并进一步还原为视频。训练时常以真实视频为监督，通过重建真实的运动轨迹进行优化。整个生成过程可形式化为

$$V = G(I, D, c),$$

其中I表示外观特征，D为驱动信号，c为可选控制条件，G为生成函数，V则是生成的图像帧或视频。

具体而言，虚拟人驱动生成通常包含以下4个关键步骤。

1）特征提取。首先对输入数据进行特征提取，以将多模态信息编码成模型可以理解的形式：一方面，使用深度学习模型从静态的外观信息中学习并提取其外观特征作为虚拟人的身份表示；另一方面，从驱动信号中提取时序特征，捕捉驱动信号中的空间结构、深层语义、情感状态等控制信息。

2）动作生成。在特征空间中，模型将驱动特征与目标动作进行条件映射，建模两者之间的对应关系，根据驱动信号及其他控制条件，生成连续的动作参数序列以描述身体各部位运动的变化过程。

3）特征扭曲或渲染。根据驱动信号映射得到的动作参数序列，按照所使用方法的不同，将预测的动作序列与外观特征编码进行结合，对外观特征进行空间扭曲；或使用动作参数对人物进行重新渲染，驱动人物的对应关键区域发生变化以得到目标特征。

4）视频生成。根据所使用的方法，将扭曲后的特征输入解码器进行图像生成，或通过投影的方式将渲染的人物模型还原到像素空间，按照时序拼接为视频序列，实现从静态外观到动态虚拟人的完整生成。

上述流程是虚拟人生成的通用范式，在这一基础流程上，研究者们提出了诸多改进策略，例如引入注意力机制聚焦于面部关键区域以提升表达细节，使用中间表示提升特征解耦能力，引入参考帧稳定生成结果，或设计多尺度编码与解码器以增强时空建模能力等。同时，不同的网络结构及训练策略（如自监督学习、对抗训练、感知损失）也会根据任务需求进行灵活组合与替换，使虚拟人生成系统在准确性、实时性与可控性之间取得更优权衡。

1.1 驱动信号源的区别与本质

视频驱动的说话人生成技术是一类利用驱动视频中提供的详细视觉线索，如面部表情、姿态、口型变化等驱动目标人物以生成自然说话人视频的技术。视频驱动的难点在于从驱动视频中解耦出身份无关的运动信息，从源图片或视频解耦出运动无关的身份信息，并将其整合映射到目标视频中。

音频驱动说话人生成是指给定驱动音频，生成具有与音频一致的说话唇形、表情和头部运动的目标人物视频的技术。相较视频驱动，音频驱动的难度大大增加。除了中间表征的选择和图像重建的优化之外，音频驱动在生成时还需要重点考虑3个问题：

1）跨模态的驱动信息生成。音频中不包含视觉信息，跨模态的生成更为困难，相较视频驱动需要更多高质量数据集来训练学习音频到视觉运动信息的映射。

2）音频中运动信息的解耦。生成说话人视频往往需要唇部运动、面部表情和头部运动等多个运动信息，但多个运动之间的关系较弱，与音频之间的映射关系也概不相同，通常需要采用多个不同的编码模块从音频中分别解耦出不同的运动信息指导生成，增加了训练的成本和难度。

3）1对多的映射问题。音频到视觉运动信息的映射并非简单的1对1映射关系，即单个音频可能有多个对应的正确运动信息，如果采用普通的自回归训练学习映射很容易会导致运动信息的过度平滑[8]，生成的面部运动容易大同小异。

这3个问题使得音频驱动技术更具挑战性，也更受研究者关注。

条件驱动的说话人生成技术是一类利用文本或结构化条件信息（如说话文本、情感标签、动作指令等）作为驱动信号，生成符合条件语义的说话人视频的技术。相较于视频驱动和音频驱动，条件驱动方式不依赖连续的感知模态输入，而是通过更高层次、更抽象的语义信息指导生成过程。这类方法的核心

挑战在于如何将语言或结构化信息高效地编码成可驱动图像生成的运动表示,并实现语义一致性与视觉真实感的统一。在技术实现上,条件驱动方法往往依赖多模态融合架构,通过将条件信息与目标身份图像或潜在特征结合,引导生成符合特定说话内容、情绪状态或交互动作的视频输出。相比其他模态驱动,条件驱动存在以下 3 个关键难点:

1)语义到运动的跨模态映射问题。文本等条件信息并不直接携带具体的运动特征,生成系统需要学习从语义空间到视觉运动空间的复杂映射关系,这一过程高度依赖大规模对齐数据和精细建模。

2)条件信息的粒度控制与一致性建模。条件驱动不仅要实现说话唇形与语义一致,还要同步驱动面部表情、头部姿态乃至身体动作,这要求模型在理解条件语义的同时,具备细粒度的表情与姿态建模能力。

3)生成多样性与控制性的权衡。由于文本驱动天然存在 1 对多的生成可能,系统既要保证生成动作与语义的合理对应,又需避免模式坍缩带来的重复与僵硬。

从驱动方式的核心差异来看,视频、音频与条件驱动在信号利用、技术难点及适用场景上各有侧重,优劣特征显著:视频驱动依托详细视觉线索,运动信息直观可解,无需突破跨模态映射障碍,生成稳定性更强,但需参考视频,核心难点集中在身份与运动信息的解耦,更适用于有真人参考的虚拟人复刻场景,如真人数字分身;音频驱动虽摆脱了视觉输入依赖,交互灵活性更高,但因音频无直接视觉信息,核心难点集中在跨模态生成、多运动信息生成及“1 对多”映射导致的运动平滑问题,技术复杂度高于视频驱动,更适配语音交互类场景,如电话客服虚拟人;条件驱动则通过抽象语义指令来实现虚拟人的精细化语义控制,无需连续感知模态输入,个性化更高,但需解决语义到运动的跨模态映射、表情姿态多维度一致性建模及多样性和可控性之间的权衡,数据依赖与技术复杂度最高,更适合脚本驱动、指令交互等语义可控场景。

在虚拟人驱动生成的整体流程中,虽然音频、视频等信号源在模态形式上各不相同,但在实际建模过程中,它们的作用本质上是一致的。作为驱动信息的来源,这些原始信号通常会先经过特征提取模块,转化为具有语义或运动信息的高维驱动特征。随后,这些统一格式的驱动特征作为条件输入,与目标人物的外观特征表示共同输入生成网络,从而实现符合语义、身份一致、动态自然的虚拟人生成。因此,不同模态的驱动信号在系统中的角色是趋同的,核心在于如何高效而准确地从中提取出可用于生成的动态控制信息。为此,本综述主要探讨所使用的具体模型技术方法的不同,对驱动信号不进行主要区分。

1.2 基础深度学习模型

在虚拟人生成领域,基础神经网络模型作为底层构建单元,承载着从多模态输入中提取语义、建模时空依赖以及合成高质量输出的关键功能。理解这些基础模型的结构原理与建模特性,是分析虚拟人生成方法体系、把握发展趋势的基础。

1.2.1 卷积神经网络

卷积神经网络(convolutional neural network, CNN)是一种专门设计用于处理具有类似网格结构数据(如图像)的深度神经网络,其核心在于:通过卷积操作提取局部区域的空间特征并实现参数共享,并使用多层网络堆叠学习高级抽象表示。CNN 最早应用于图像分类和物体识别等计算机视觉任务,其主要优势在于其对图像局部纹理和面部结构的建模能力,特别适用于静态图像编码、面部区域细节合成和生成图像的去模糊等环节。在虚拟人生成任务中,CNN 被用于提取源参考图像的外观特征,并将运动驱动信号融合用于帧级图像重建。相比传统图形渲染方法,CNN 可以在端到端的网络中学习图像到图像的映射规则,显著提升虚拟人生成的效率与真实性。然而,CNN 本质上是局部感知结构,受限于卷积核大小和层级堆叠深度,难以捕捉长时间序列中的上下文关系,也不擅长跨模态输入的对齐任务,通常需要与其他时序模型如 Transformer[7] 配合使用。

1.2.2 长短时记忆网络

长短时记忆(long short-term memory, LSTM)网络是一种经典的循环神经网络(recurrent neural network, RNN)变体,专为解决普通 RNN 在长序列建模中易出现的梯度消失或爆炸问题而设计。其引入了“记忆单元”和“门控机制”,使得网络能够选择性地记住或遗忘历史信息。LSTM 最初被广泛应用于语音识别、语言建模、时间序列预测等任务,在处理音频和语言等强时序依赖性数据方面展现出优秀能力。在虚拟人说话面部生成中,LSTM 通常用于建模音频与面部运动之间的时间动态关系,尤其适合唇形同步和表情随语音变化的连续生成,能够利用其时序记忆机制,生成与语音同步的面部动作,提升说话动画的自然度。然而,LSTM 计算效率低,难以并行化;

此外，尽管引入门控机制缓解了信息遗失问题，但仍难以建模特别长距离的全局依赖；同时，其在多模态融合与复杂语义理解方面能力有限，难以胜任情感表达或多角色对话等高维控制任务。因此，在近期的研究中，LSTM已逐步被Transformer等更强的时序建模结构所取代。但得益于结构简单、易于训练的特点，仍在轻量化和嵌入式场景中具有重要应用价值。

1.2.3 变分自编码器

自编码器（autoencoder）是一种通过将输入压缩到低维潜空间并重建输出的无监督神经网络架构，其核心思想是学习紧凑的表示形式。变分自编码器（variational autoencoders, VAE）[9]在此基础上引入概率建模，使得潜变量空间更具连续性和生成能力。这类模型最初主要用于图像压缩、去噪和生成建模等任务。在虚拟人生成中，VAE被广泛用于身份嵌入、风格迁移、表情参数控制等任务中。通过在潜空间中对面部特征或动画参数进行表示，模型可以实现个性化驱动、风格可控的表情生成。同时，VAE在潜空间中可实现性较强，适合进行线性插值、类别过渡等操作，特别适合多情绪或跨语言说话头像生成的调节需求。但其主要缺点在于生成结果通常较模糊，尤其在高保真图像重建任务中表现逊色，此外其KL散度（Kullback-Leibler divergence）约束可能导致学习效果不稳定，常与GANs结合使用以弥补这一问题。

1.2.4 Transformer

Transformer[7]是一种基于注意力（attention）机制的神经网络架构，最初用于机器翻译任务，其核心在于完全移除了递归结构，通过位置编码与全连接注意力机制实现对全序列的建模与时序关系捕捉。该模型已成为自然语言处理领域的主流架构，并广泛扩展到语音理解、图像生成和多模态建模等领域。在虚拟人说话头像生成中，Transformer的最大优势在于其强大的长距离依赖建模能力与跨模态对齐能力。具体而言，Transformer能够从完整的音频序列中捕捉语义、韵律等多层信息，并将其与图像、表情或运动序列进行跨模态的全局对齐，适用于情感表达同步、语义驱动动画等高复杂度任务。然而，由于其参数量大、计算复杂度高，对训练数据和资源要求较高，同时对时序建模的泛化能力存在过拟合风险，需要在设计中结合稀疏注意力或结构剪枝等策略加以优化。且其上下文长度的限制导致其无法无限外推，在长音频驱动情况下无法发挥全部性能。

1.2.5 GANs

GANs[2]是一种通过博弈机制进行训练的生成模型，由生成器和判别器组成，其中生成器用于生成仿真的图像，判别器用于辨别输入的图像是否真实，通过对抗训练逐步提升生成图像的质量和真实性。GANs最初被设计用于图像生成和图像修复任务，后迅速扩展至图像风格迁移、语音合成等多模态生成场景。在虚拟人说话头像生成中，GANs主要用于提升图像输出质量和增强生成结果的自然性，其最大优势在于其强大的图像生成能力，尤其擅长纹理恢复、细节增强等任务，是高质量合成的保障之一。然而，GANs存在训练不稳定、模式崩溃（mode collapse）等典型问题，难以对生成内容进行精确控制，尤其在视频连续性和多模态一致性方面表现不足，限制了其在长时序虚拟人生成任务中的单独应用。

1.2.6 扩散模型

扩散模型[3-4]是一种基于马尔可夫链的生成式模型，其基本原理是通过逐步向数据添加噪声来模拟热扩散，随后通过逐步去噪学习数据的反向生成过程。最初用于图像生成任务，后逐渐被拓展到视频生成领域。在虚拟人生成中，扩散模型因其稳定训练过程和优异的生成质量而获得广泛关注。其优势在于能够捕捉高维数据的复杂分布，生成连续且自然的说话动作序列，尤其适合建模音频与表情、动作之间的高阶非线性映射。扩散模型具备良好的可控性与多模态融合能力，适合生成复杂、带有情绪风格的动态动画。但其主要缺点是推理速度慢，生成过程需要迭代多个去噪步骤，导致实时性较差。其核心网络结构通常采用U-Net[10]，但标准U-Net不具备时间信息建模能力，需要引入专门的时间建模模块。Peebles等人[11]将扩散模型与Transformer结合，提出了DiT（diffusion Transformer），通过融合Transformer的时序建模优势，提升了扩散模型在动态序列生成中的表现。

1.2.7 模型组合

在复杂的虚拟人生成任务中，单一模型往往难以同时兼顾时序建模、语义控制与高质量合成，因此目前普遍采用多模型协同的策略，将不同类型的基础模型按功能模块化组合，以充分发挥各自结构优势。典型做法是以CNN作为前端编码器，提取面部图像或姿态序列中的局部空间特征；再结合Transformer或LSTM模型对驱动信号进行动态建模，其中Transformer凭借其全局注意力机制，尤其适合处理长时间跨度的语音到表情映射与模态对齐问题，

而 LSTM 则常用于对低延迟与短期动作响应有严格要求的场景。在生成阶段，为增强潜空间的结构性与控制性，VAE 往往被引入作为生成器的核心结构，其连续、可微的潜空间表示可与 Transformer 有效融合，从而实现语义可控、风格一致的表情或姿态序列生成。而在提升生成质量方面，扩散模型近年来被广泛整合入音频驱动虚拟人系统中，结合条件编码器实现对生成图像序列的高保真建模，尤其在口型精度、情绪还原与跨模态一致性方面展现出显著优势。为了进一步提升最终输出的真实感，GANs 判别器模块通常作为辅助网络嵌入，以提供感知层面的质量约束，指导生成器生成更具自然感与细节表现力的图像或视频内容。综上，这类多模型融合架构不仅实现了从语音信号到面部视频的端到端高质量合成，同时也为情感建模、说话风格迁移、多角色泛化等下游任务提供了可扩展的技术基础，正成为当前主流研究趋势。

1.3 驱动目标部位

在虚拟人动画生成中，驱动部位的划分是实现精细化控制与自然生成的关键因素。驱动技术可以根据生成重点、技术差异和产业需求的不同分为 4 个层次：唇形驱动、面部与头部驱动、姿态驱动和集成驱动。其中，唇形驱动主要关注音频与口型之间的时序一致性，是实现语音驱动虚拟人最基本的模块，这一领域需求明确、技术体系独立，因此单独列出。面部与头部驱动从唇形驱动扩展而来，在此基础上进一步生成自然的面部表情和头部运动来增强虚拟人情感表达与交互真实感。由于头部动作生成方式与其他细微表情相似，因此将其与面部表情合并归类。姿态驱动主要集中于躯干、四肢等部位的动作驱动，更侧重于表达特定的行为动作，如舞蹈、运动等。集成驱动是指综合上述多种驱动部位，通常通过骨骼图生成，整合多种技术实现虚拟人在复杂语境下的全身协调动作。不同层级的驱动部位相互协同，构成了自然、连贯且具备互动能力的虚拟人动画系统。本文的核心内容和分类情况汇总在图 2 中，后文将根据所划分的层次，按照驱动部位展开介绍。

2 唇形驱动

唇部驱动生成技术专注于模拟说话者唇部的运动，也被称为唇形合成（lip-syncing）和对口型（dubbing）任务。在这个任务中，音频大多被看作整体，多采用短时傅里叶变换（short-time Fourier transform, STFT）、提取梅尔频谱图（Mel spectrogram）和梅尔频率倒谱系数（Mel-scale frequency cepstral coefficients, MFCC）或是基于网络编码处理后输入深度学习网络中进行进一步训练模态对齐，1 对 1 地映射到唇部运动中。代表性唇形驱动技术汇总如表 1 所示。

2.1 基于简单机器学习架构的技术

早期唇部驱动生成技术大多基于简单机器学习架构，如 CNN 和 LSTM 等，学习音频序列到唇形特征的映射。Suwajanakorn 等人[12]于 2017 年提出了

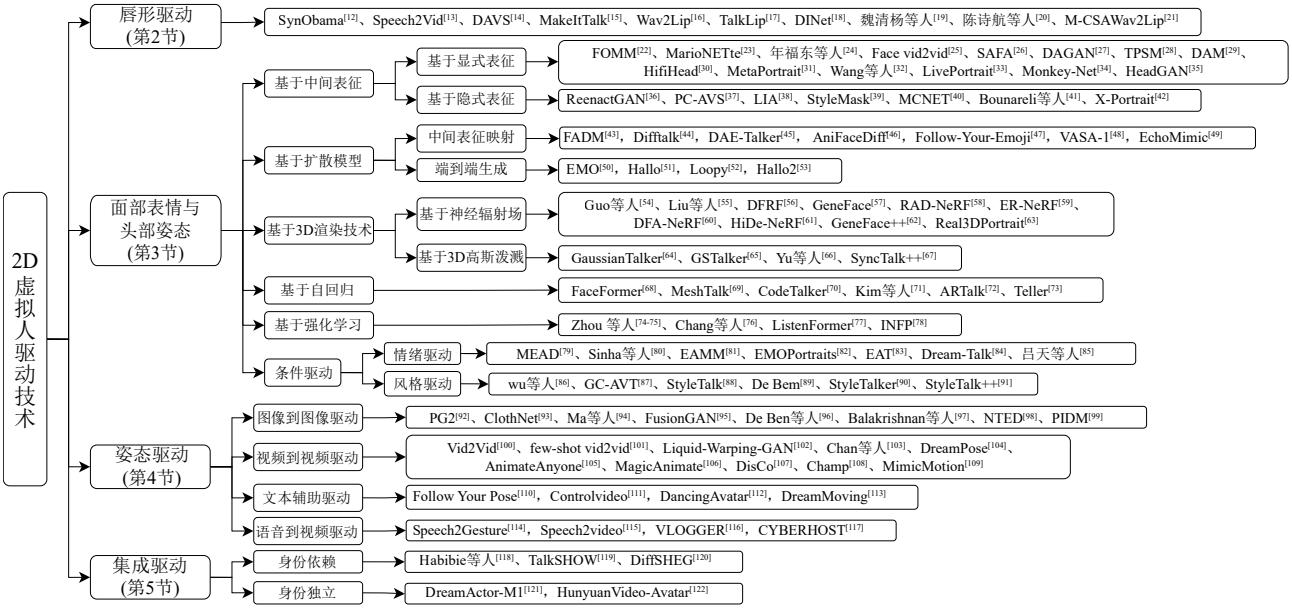


Fig. 2 Classification of 2D virtual human-driven technologies

图 2 2D 虚拟人驱动技术分类

Table 1 Summary of Lip-Driven Talking Head Generation Technologies

表 1 唇形驱动说话人生成技术汇总

年份	模型	驱动对象	基础模型	中间表征	唇部运动	面部表情	头部运动
2017	SynObama[12]	特定	LSTM	landmarks	√		
2017	Speech2Vid[13]	任意	CNN		√		
2019	DAVS[14]	任意	GANs		√		
2020	MakeItTalk[15]	任意	LSTM	landmarks	√	√	
2020	Wav2Lip[16]	任意	GANs		√		
2023	TalkLip[17]	任意	GANs		√		
2024	DINet[18]	任意	GANs		√		
2024	文献 [19]	任意	CNN+GANs	landmarks	√		
2024	文献 [20]	任意	GANs		√		
2025	M-CSAWav2Lip[21]	任意	GANs		√		

SynObama, 针对奥巴马这一特定人物, 采用了单向 LSTM 网络学习音频特征到人工标注的唇形标准关键点(landmarks)的映射, 实现了音频驱动说话人视频合成, 在唇音同步上取得了较好的效果, 但这一模型只针对特定人物, 模型的泛化性较差。Jamaludin 等人[13] 提出了 Speech2Vid, 基于 CNN 将音频映射拓展到任意面孔, 主要包含对音频 MFCC 特征进行编码的音频编码器、从参考图像中提取身份信息的身份编码器和融合身份与音频特征向上卷积生成视频帧图像的图像解码器 3 个核心部分。Speech2Vid 实现了单阶段音频驱动, 编码器和解码器都采用 CNN 架构实现, 使用了后处理去模糊模块来提升生成的图像质量。但其网络结构较为简单, 生成视频质量较低, 没有考虑到生成视频帧之间的时间连续性, 在序列的一致性上有待改进。MakeItTalk[15] 相比之下更进一步, 基于单张图像进行唇形合成, 在基本唇形动作的基础上还实现了基本的面部表情合成。提出了语音转换(voice conversion)网络将语音内容信息和说话人身份特性解耦, 基于 LSTM 和简单神经网络架构分别学习身份特征和音频特征对应的面部标准关键点位移, 生成具有说话人特征的生动语音驱动视频。基于简单机器学习架构的唇形合成方法的优势在于实现相对容易、计算开销较低、能够满足一定的生成质量需求, 但对音频和视频之间复杂动态关系的捕捉能力较弱, 难以保证生成视频的逼真度和连贯性, 特别是在高保真和细节生成上容易出现不自然的情况。

2.2 基于 GANs 和 Transformer 的技术

随着 GANs 的发展, 越来越多的方法基于 GANs[2] 和 Transformer[7] 架构进行唇形合成, 提高生成图像的质量。GANs 一般通过生成器和判别器的博弈学习, 将生成器生成的图像与判别器识别的真实图像进行对比, 以提高生成图像的真实性。在唇形合成中, 生成器负责从音频特征生成唇部动作, 而判别器负责检测生成唇形与输入音频的同步性, 通过不断优化生成器, 提升生成图像的质量和同步性。Zhou 等人[14] 提出了解耦的视听系统(disentangled audio-visual system, DAVS), 以区别于其他说话人生成模型直接将音频信息映射到面部特征的方法, DAVS 采用 2 个编码器网络分别提取音频和视觉特征, 并采用对抗训练的方式使得 2 个特征空间独立生成不同的信息, 可以将任意视频和音频进行配对。Wav2Lip 是 Prajwal 等人[16] 于 2020 年提出的一种基于 GANs 的经典音频驱动唇形生成方法, 技术架构如图 3 所示, 该技术在语音驱动的口型同步生成中引入了预训练的专家唇形同步(lip-sync expert)鉴别器, 用于指导 GANs, 实现了优秀的唇形同步效果。

TalkLip[17] 将更多关注放在唇语清晰度上, 基于 Transformer[7] 架构编码语音, 采用唇读专家(lip-reading expert)自监督惩罚视频生成中不正确的唇读, 并提出了一种新的基于唇读专家的评估策略, 进一步提升了视频唇语的精确度和清晰度。Zhang 等人[18] 针对高分辨率的唇形合成提出了一个新的形变修复网络(deformation inpainting network, DINet), 采用形变网络在高维空间中编码和对齐驱动音频和源人像, 采用修补网络基于高维空间特征补全源人像嘴部区域, 相比其他方法保留了更多细节, 在高分辨率视频上唇形合成效果更好。在 DINet 的基础上, 魏清杨等人[19] 引入标准关键点与草图(sketches)作为中间表示, 首先将音频特征映射至面部标准关键点, 再根据关

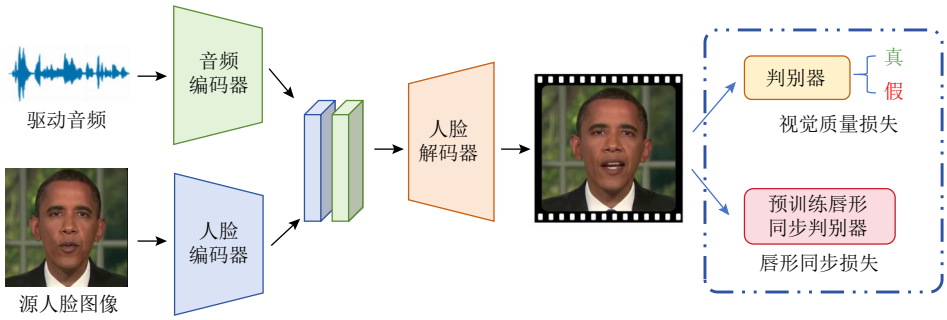


Fig. 3 Lip-driven generation technology based on GANs[16]

图 3 基于 GANs 的唇部驱动生成技术[16]

键点生成草图，并与参考图像进行语义对齐，从而缓解音频与图像的跨模态偏差，在保证唇形同步的同时，提升面部区域的整体协调性。陈诗航等人[20]通过仿射变换实现音频与视觉特征在解码阶段的深度融合，提升语音对人脸生成的控制能力。模型引入通道与空间注意力机制增强局部特征响应，采用双判别器结构实现对唇部运动与图像真实感的约束，并结合 StyleGAN[123] 跳跃连接与谱归一化技术进一步提升生成稳定性与图像清晰度。M-CSAWav2Lip[21]重点关注中文音频输入下的唇音同步问题，使用 MFCC 和梅尔频谱图分别提取音频的时间与频率特征，同样也利用通道与空间注意力机制增强模型对细微动作的识别能力，最终提高了中文语境下的唇音同步性。

唇部驱动生成方法已经在唇形同步度、唇形清晰度和生成精确度上取得了相当不错的效果，但该类技术局限在嘴部区域的驱动，生成的说话人视频缺乏头部运动和面部表情等信息，视频生动性有待改进，应用场景受限。因此，部分研究开始探寻如何从音频中提取头部运动和面部表情等多种信息，驱动生成生动的说话人视频。

3 面部与头部驱动

面部与头部运动生成是虚拟人生成中的核心组成部分，直接决定了生成视频的真实感、情感表达力与交互自然性。该任务旨在保持人物身份一致性的前提下，根据参考图像和多模态驱动信号，生成符合语义内容的说话头视频，包含丰富、细腻的面部表情与符合语境的头部运动。这一过程不仅涉及多模态信息的深度融合，还需解决模态间对齐困难、表情歧义性高、运动连续性与稳定性难以保证等技术难点。面对跨模态映射与身份保持的双重挑战，研究者不断引入更强的时序建模能力、注意力机制以及条件生成框架，以提升生成效果的自然性与个性化。面部驱动包含多种实现方式，在本节将分别做详细介绍。其中基于中间表征与扩散模型的方法是目前研究最广泛、技术最成熟的 2 类方法，其面部生成技术汇总如表 2 所示，致力于使用 2D 模型生成单目说话头像。

3.1 基于中间表征的生成技术

基于中间表征(intermediate representations, IR)的生成技术如图 4 所示，一般包括以下 4 个步骤：1)参考图像或视频的身份特征提取；2)根据驱动信号提取或映射唇部特征、表情和头部姿态等运动表征信息；3)中间表征生成，驱动信号的运动信息到参考图像或视频的映射；4)基于生成模型对目标人物重建与生成。根据生成过程中人脸建模用到的参数或者中间特征，可以将基于中间表征的 2D 说话人生成技术分为基于预定义的显式表征的技术和基于隐式表征的技术。

3.1.1 基于显式表征的技术

基于显式表征的视频驱动方法通常将有明确定义或特征的参数(如标准关键点、模型自学习关键点和 3D 参数等)作为中间特征来描述虚拟角色的动作和表情，采用形变网络(warping net)来从中提取运动流并映射到目标人物，最终基于生成模型重建视频帧。

基于标准关键点的显式表征方法依赖于预定义的结构点进行人脸建模，数据处理时采用预训练的面部检测模型标记面部特征点(如眼睛、鼻子、嘴巴、面部轮廓等)。MarioNETte[23] 是一个小样本面部重演框架，在常规的视频驱动编码器解码器结构上增添了标准关键点转换器、图像注意力和目标特征对齐 3 个新模块。标准关键点转换器解耦了驱动视频身份和面部运动信息，图像注意力和目标特征对齐更好地对齐和融合了驱动信息和目标信息，较好地保留了目标人物的身份。年福东等人[24]采用并行多

Table 2 Summary of Monocular Talking Head Generation Technologies Based on 2D Models

表 2 基于 2D 模型的单目说话头生成技术汇总

年份	模型	驱动信号	驱动对象	基础模型	显式表征	隐式表征	中间表征
2018	ReenactGAN[36]	视频	特定	GANs		√	
2019	FOMM[22]	视频	任意	CNN	√		2D 关键点
2020	MarioNETte[23]	视频	任意	Transformer	√		landmarks
2021	文献 [24]	音频	任意	GANs	√		landmarks
2021	Face vid2vid[25]	视频	任意	CNN	√		3D 关键点
2021	SAFA[26]	视频	任意	CNN	√		3DMM
2021	PC-AVS[37]	音频	任意	GANs		√	
2022	DAGAN[27]	视频	任意	GANs	√		2D 关键点
2022	TPSM[28]	视频	任意	CNN	√		2D 关键点
2022	DAM[29]	视频	任意	CNN	√		2D 关键点
2022	HifiHead[30]	视频	特定	GANs	√		3DMM
2022	LIA[38]	视频	特定	GANs		√	
2023	MetaPortrait[31]	视频	任意	GANs	√		landmarks
2023	Wang 等人[32]	音频	任意	CNN	√		landmarks
2023	StyleMask[39]	视频	特定	GANs		√	
2023	MCNET[40]	视频	任意	CNN		√	
2023	FADM[43]	视频	任意	Diffusion	√		3D 关键点
2023	Difftalk[44]	音频	任意	Diffusion	√		landmarks
2023	DAE-Talker[45]	音频	特定	Diffusion		√	
2023	Diffused Heads[124]	音频	任意	Diffusion			端到端
2024	LivePortrait[33]	视频	任意	CNN	√		3D 关键点
2024	DiffusionAct[41]	视频	任意	Diffusion		√	
2024	X-Portrait[42]	视频	任意	Diffusion		√	
2024	AniFaceDiff[46]	视频	任意	Diffusion	√		3D 关键点
2024	Follow-Your-Emoji[47]	视频	任意	Diffusion	√		3D 关键点
2024	VASA-1[48]	音频	任意	Diffusion		√	
2024	EchoMimic[49]	音频	任意	Diffusion	√		landmarks
2024	EMO[50]	音频	任意	Diffusion			端到端
2024	Hallo[51]	音频	任意	Diffusion			端到端
2024	Loopy[52]	音频	任意	Diffusion			端到端
2024	Hallo2[53]	音频	任意	Diffusion			端到端

分支网络分别建模语音到唇部与头部关键点的映射关系，以实现对唇动与头动的解耦建模。生成过程中，模型以人脸轮廓和唇部关键点序列为中间表示，结合模板图像，通过注意力机制融合局部运动特征，最终输出具有语音同步性和面部协调性的视频序列。MetaPortrait[31] 则是提出了一个新的个性化适应框架，使用元学习方法来快速实现高效的个性化微调，将驱动和源人脸都编码到密集的标准关键点空间，增加生成精度的同时便于进行身份融合。Wang 等人[32] 设计了一个内存共享情感特征提取器用于从音频中提取情感特征，采用基于 U-Net 的注意力增强翻译器来学习标准关键点到视频帧的映射，提高了说话人视频中情感的真实性。

基于模型自学习 2D 关键点的显式表征方法区别于预定义的面部标准关键点，它们是由模型自己学习得到的、不一定可解释的非固定关键点。这些点主要负责捕捉全局和局部的形状变化和运动特征，由网络决定图像中哪些位置最能表达运动信息，并从图像中自适应地提取 2D 关键点，通过这些点来捕捉源图像与目标图像之间的运动关系。Monkey-Net[34]

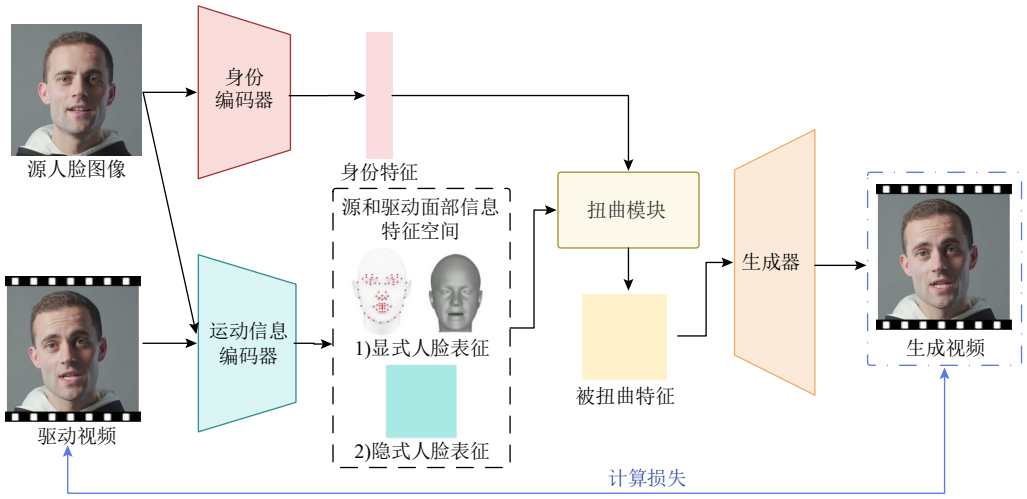


Fig. 4 Facial generation techniques based on intermediate representations

图 4 基于中间表征的面部生成技术

采用关键点作为中间特征，分别设计了 3 个基于 U-Net 模型的核心模块。其中，无监督训练的关键点检测器用于提取源图片的关键点信息，运动预测网络根据关键点生成稠密运动热力图(dense motion heatmap)，运动转换网络将运动信息和外观信息整合得到最终生成视频。由于架构较简单，Monkey-Net 的效果较差，生成的细粒度也较低。Siarohin 等人[22]对其进行改进，提出了 1 阶运动模型(first order motion model, FOMM)，基于自监督学习和 1 阶局部仿射变换对关键点周围运动进行建模，将模型应用范围拓宽到复杂运动上，并采用图像修复(image inpainting)技术针对遮挡问题补充生成。Hong 等人[27]的深度对抗生成网络(depth-aware generative adversarial network, DAGAN)进一步在说话人头像生成中引入深度感知，通过估计深度图来捕捉脸部的 3D 结构，以更好地生成关键点，在捕捉头部转动和面部表情变化上表现更好。不同于之前工作[22,125]大多采用局部仿射变换处理关键点，TPSM(thin-plate spline motion model)[28]增加了多组 TPS(thin-plate spline)变换，进行源到驱动的变换，缓解了源和驱动图像姿态差距较大带来的对齐问题，适应了更复杂的运动和形变，设计了额外的背景和形变损失约束，提高了运动流的质量，防止背景跟着人一起运动。DAM(deformable anchor model)[29]受可变形部件模型(deformable part model)的启发，在关键点表达中额外增加锚点来表达全局运动，利用关键点和根锚点的先验关系更好地提高关键点的准确性，有效地捕获运动中的结构信息。

基于 3D 参数的显式表征方法通常采用 3DMM 和融合变形(blendshape)中的 3D 关键点等信息作为中间特征，辅助描述 2D 视频中的人像形状和面部动作等信息，来实现高质量和高可控性的人物动画生成。HeadGAN[35]将驱动视频分为视觉图像部分和音频部分，从视觉部分构造 3D 人脸表示，解耦身份的运动，采用音频增强嘴部运动，将优化后的驱动 3D 面部表示适应源人像的面部集合形状，以实现进一步的编辑和生成。SAFA(structure aware face animation)[26]在 FOMM[22]基础上添加了 3DMM 模块进行面部模拟，同时应用多次仿射变换来渲染头发和胡须等前景元素，改善了细节生成。HifiHead[30]提出了一种高保真说话人合成方法，基于 StyleGAN[123]设计了包含细节对齐和细化的生成器模块，结合 3DMM 信息和图像生成技术来重建说话人，能够生成高分辨率、真实感极强的头像，有效保持源图像身份的同时提供了 3D 可控性。还有部分工作提取 3D 空间中具有语义价值的部分重要关键点作为中间特征，加快生成速度，更关注关键部位的生成。Face vid2vid[25]采用 3D 关键点来表达驱动视频的表情和头部运动，采用无监督学习方法训练模型提取关键点。Face vid2vid 在生成时可以对提取的 3D 关键点做有意义的编辑，实现可控调整说话人运动状态。LivePortrait[33]在 Face vid2vid[25]的框架上进行改进，引入标准关键点信息，结合级联感知和局部 GANs 等损失，混合图像和视频进行训练，提升了 3D 隐式关键点提取网络架构，增加了拼接模块和重定位模块，使得模型在不同头像大小的视频之间的映射更自然，眼唇开合线性可控，在生成上取得优异的效果。

得益于显式表征与人物面部和运动信息的直接关联性，基于显式表征的视频驱动技术具有较高的

可解释性和生成可控性，可以通过对中间特征进行有意义的语义编辑来实现对应生成内容的控制。但生成效果较大程度上取决于提取特征的准确性，如果特征提取有误，生成效果也会受到较大影响，需要在训练时充分学习特征的提取，生成质量依赖数据集的大小和质量。此外，由于显式表征大多为预定义的信息，缺乏一定灵活性，难以捕捉复杂运动和细微变化。

3.1.2 基于隐式表征的技术

基于隐式表征的视频驱动技术将深度学习模型的语义空间或编码空间作为人脸和运动的表征空间，在潜在空间中学习源到驱动的映射和变换，这类方法主要基于GANs和扩散模型实现。Wu等人[36]提出的ReenactGAN将驱动视频的人脸角度、表情和光照角度等多种信息在潜空间中编码，采用转换模块将驱动和源信息进行对齐，并用特定的编码器重建视频帧得到目标人物说话视频。Zhou等人[37]提出了一种语音驱动可控头部姿态的说话人生成模型PC-AVS(pose-controllable audio-visual system)，在潜空间中定义了12个维度的人脸姿态表征，使用GANs和解纠缠表示学习，实现了较好鲁棒性的带头部姿态的说话人视频，但本质上是唇形合成技术的简单扩展，视频质量仍然较差，没有丰富的面部表情。StyleMask[39]针对驱动视频人物姿态变化幅度大时源图像的身份特征保留较差这一问题，提出了身份和头部姿态解耦框架，在StyleGAN2[126]的风格空间解耦身份和姿态，引入掩码机制分离出姿态、表情和身份等因素，实现了更具可控性和高质量的面部再现效果。LIA(latent image animator)[38]从源和驱动信息之间的变换入手，不同于传统方法在图像层面或者结构层面的变换，提出了一种在潜在空间中学习源图像到驱动图像运动变化轨迹的方法，通过学习潜在空间内的正交运动方向并将其线性组合来实现动画，避免了结构表示的复杂处理。记忆补偿网络(memory compensation network, MCNET)[40]通过隐式身份表示来提取源图像身份信息，设计身份记忆模块和面部记忆补偿机制来增强目标视频中的身份信息，构建了类似FOMM[22]的架构进行驱动视频运动流的提取，在动态视频生成中有效保持了身份特征。Bounareli等人[41]将DAE(diffusion autoencoder)[127]的语义空间作为面部运动潜空间，通过条件语义编码器实现外观和身份的编辑，无需对特定主体微调，能够实现与GANs相当甚至更好的面部重演效果。X-Portrait[42]提出了一种基于分层注意力可精细控制姿态和面部表情的动画人物肖像生成架构。采用预训练扩散模型提取的先验信息作为渲染支撑，基于ControlNet[128]进行头部姿态和表情细粒度控制，在动画肖像视频生成上取得了较好的视觉效果。

基于隐式表征的视频驱动技术能够通过深度学习模型自动学习复杂的运动变换，不需要人为定义显式特征，在处理复杂的面部表情和细节变化时表现更加出色。但基于隐式表征的方法可解释性较差、生成可控性较低，且训练和推理时通常都需要更大的数据量和计算资源。

3.2 基于扩散模型的生成技术

近2年来，扩散模型[3-4]发展迅速，在图像生成、图像编辑和视频生成等领域展现出了优秀的生成和编辑能力[129-130]，也有部分工作开始尝试基于扩散模型驱动2D图像。现有方法根据去噪生成的内容可以分为2类：一类是中间表征映射方法，利用扩散模型根据输入驱动信号去噪生成中间表征，实现跨模态的特征映射；另一类是端到端生成方法，直接将驱动信号与时间建模条件输入扩散模型，通过去噪直接生成符合时序条件的目标图像帧。

3.2.1 中间表征映射

基于扩散模型的视频驱动技术为了修正GANs等生成方法导致的不自然伪影和画面失真，Zeng等人[43]提出了属性(表情和头部姿态)引导的扩散模型说话人生成框架FADM(face animation diffusion model)，基于DECA[131]对驱动和源图像进行3D重建以获取表情和头部运动参数等信息，设计粗略动画生成模块用于获得初步的动画效果，构建属性引导条件网络整合上述信息引导生成，改善了生成视频中的面部细节。AniFaceDiff[46]基于Stable Diffusion[132]构造了一个2阶段式面部重演架构，从源和驱动中分别提取出3D面部形状，对齐后经头部姿态编码器与噪声一起传入去噪网络中进行生成，并采用表情适配器提取驱动视频中的表情运动作为额外条件引导生成。Follow-Your-Emoji[47]则是从驱动视频3D关键点中投影来获得情绪感知标准关键点序列作为引导条件，设计面部细粒度损失函数，计算面部表情感知的空间距离，使得模型更专注于捕捉细微的表情变化和身份细节外观。

基于扩散模型的音频驱动技术也在生成效果上取得了显著成就。Diffused Heads[124]是一篇将扩散模型用于语音驱动说话人生成任务的工作。Diffused Heads采用扩散模型重建视频帧，在生成中引入了当前帧的前2帧作为运动帧来保持视频帧之间的连续

性以及采用了随机采样的身份帧来保持身份一致性,将预训练音频编码器提取的音频特征注入到扩散模型的交叉注意力中来保持音频和生成帧之间的一致性。之后基于扩散模型的说话人生成技术逐渐开始发展,Difftalk[44]将参考图像和标准关键点作为额外的驱动信息辅助表情和头部姿态的建模,使说话人生成过程更可控的同时使得模型可以跨身份泛化。DAE-Talker[45]基于DAE构造了一个人脸视频潜空间,比起基于关键点的方法能够生成更精准的稠密流畅,训练模块学习音频到人脸视频潜空间的映射。作为扩散模型在说话人生成任务上的初期探索工作,DAE-Talker生成的视频在驱动上展现出了较好的细粒度,但是仍然无法捕捉到生动的面部表情和头部姿态,生成效果有待改进。后续工作在细节运动信息的生成和控制以及视频的稳定性上进行了深入探索。EchoMimic[49]在扩散模型去噪的输入中增加了可选的面部标准关键点信息,通过音频和关键点的双重训练,使得模型支持通过语音和结构点分别或是组合驱动说话人生成,增加了生成的准确性和可控性。VASA-1[48]基于DiT学习音频到面部和头部运动潜在空间的映射,并在生成过程中引入了一组可选的条件信号,如说话人注视方向、头部位移和情绪程度,极大地提升了说话人的自然性和真实性。

3.2.2 端到端生成

Tian等人[50]提出了一种基于扩散模型的音频到视频的端到端说话人生成框架EMO,设计了速度控制器(speed controller)和面部区域控制器(face region controller),引入了弱控制信号来控制生成说话人头部运动和面部生成位置,增强了生成过程的稳定性。Hallo[51]采用了与EMO相似的架构,设计了分层视听交叉注意力(hierarchical audio-visual cross attention)模块,从音频信息中解耦唇形、表情和头部姿态运动,并借助掩码和零卷积自适应分层处理多个条件,同时采用时间对齐模块增强了视频的一致性,提高了唇同步精度和运动多样性。Cui等人[53]在此基础上针对长视频生成进行优化,采用图像块随机丢弃数据增强技术随机丢弃运动帧部分图像块,确保参考图像身份在长视频生成中的有效引导,扩展了VQGAN,引入时间对齐机制提升长视频的平滑性和质量,在保持生成视频质量的前提下,将视频时长扩展到了10 min。此外,Loopy[52]设计了片段间(inter-clip)和片段内(intra-clip)的时间模块,分别保证片段间的连贯性和片段内运动驱动的有效性,利用数据中的长期运动信息来学习自然运动模式,提高了生成的稳定性。这类端到端的生成方法无需任何中间特征就能够捕捉到丰富的面部表情和自然的头部运动,生成的视频具有较强的表现力,动画十分逼真。

基于扩散模型的虚拟人生成技术凭借其逐步去噪的生成机制,具备较强的建模能力与细节还原能力,在生成真实自然的表情与动作方面表现突出,尤其适合处理复杂的动态和多模态信息。然而,由于其生成过程涉及多步推理,推理开销大,难以满足实时性要求;此外,模型训练与调控过程较为复杂,难以精确控制生成内容,语义可控性相对较弱,对高质量数据和精细对齐也有较高依赖。

3.3 基于3D渲染的生成技术

基于3D渲染的生成技术通过显式或隐式方式对单张图像或视频中人脸进行3D重建,从音频中解耦运动信息并整合到3D几何模型中,最终根据图像质量和唇形一致性等损失训练模型,渲染生成2D图像序列。这类典型方法包括基于神经辐射场(neural radiance field, NeRF)[133]和3D高斯溅射(3D Gaussian splatting, 3DGS)[134]的技术。

3.3.1 基于神经辐射场的技术

基于神经辐射场的方法利用NeRF模型对目标人物的表情和动态变化进行高维编码,以生成高保真的2D影像。核心原理是通过学习3D空间中的光场密度和颜色信息,将复杂的面部、头部特征和动态表情映射为3D隐式表示,再通过体积渲染生成逼真的2D图像序列。Guo等人[54]将NeRF引入到2D虚拟人语音驱动任务中,通过多层感知机实现音频特征到神经辐射场的映射;生成时将人体肖像场景的神经辐射场分解为头部和躯干2个分支分别进行建模,生成的视频具有高保真度,能够实现说话人视频编辑、背景替换和姿势操控等任务,具有强大的生成控制能力。Liu等人[55]针对头像与躯干分离生成所导致的衔接部分不自然问题,提出了语义感知的说话人生成架构SSP-NeRF(semantic-aware speaking portrait NeRF),使用一个统一的神经辐射场来生成说话人像,设计了语义感知的动态光线采样模块(semantic-aware dynamic ray sampling module),根据与音频的相关性对不同部位的光线采样设置了不同程度的关注,在生成框架中添加了整体躯干变形模块,通过学习头部信息来预测躯干的位移,既保证了局部细节信息的生成,又缓解了躯干与头部运动的分离。

受3D渲染方法本身缺点的影响,对驱动身份较为敏感,每个新身份都需要重新训练,跨身份驱动效果较差。针对泛化性缺陷,DFRF[56]和GeneFace[57]被

相继提出。Shen等人[56]提出的动态面部辐射场DFRF(dynamic audio-driven facial radiance field)通过参考机制和可微面部形变模块，将音频信号和2D参考图像结合起来，通过2层卷积网络计算图像特征，将其与音频特征一起作为附加条件来训练面部辐射场，使其能够从中学习到通用的先验信息，便于快速将模型推广到新的身份，实现了少样本的说话头合成。GeneFace[57]在大规模唇读数据集上学习音频到运动的映射，结合唇形同步专家Wav2Lip[16]增强了模型对各种域外音频的泛化能力。受相关工作WaveNet[135]的启发，针对1对多的映射问题，GeneFace还引入了基于VAE的变分运动生成器，将编码器和解码器扩展为全卷积网络，并使用高斯流作为先验分布，实现了多样化的3D头部运动生成。

RAD-NeRF[58]和ER-NeRF[59]从提高实时性角度出发，分别从改进模型架构和压缩表征入手，减少了模型的规模，提高了训练推理的效率。RAD-NeRF将高维音频引导的人脸表示分解为2个低维特征网格，提高了模型的效率和渲染质量。对于躯干部分，使用轻量化的伪3D可变形模块进行建模，通过2D特征网格和头部姿势来预测躯干的变形，保持与头部运动同步的同时减少了计算成本。利用快速建模技术与分解模块设计的结合，RAD-NeRF能够在GPU上实现实时推理速度，大幅提升了基于神经辐射场渲染生成的效率。ER-NeRF针对RAD-NeRF渲染和训练收敛中存在的哈希碰撞的问题，设计了一种3平面(tri-plane)哈希表征，将3D空间分解为3个正交的2D哈希网格，将碰撞局限在2D平面上，减少了碰撞，进一步提高了渲染质量和收敛速度。同时采用区域注意力框架有意识地用音频特征影响关键部位，实现了高效的局部运动建模。

其他工作则是着眼于生成质量和稳定性的提高，对模型进行创新，设计了多种改进技术，引入了多模块协同工作的架构，以提高生成效果和性能。Yao等人[60]考虑到语音特征与非唇部运动的弱相关性，提出了一种解耦面部特征的神经辐射场生成架构DFA-NeRF(disentangled face attributes NeRF)，将音频驱动的说话头生成问题分解为唇部运动和个性化属性2个解耦的表示，通过对比学习和概率模型分别处理，实现了更精确的唇形同步和个性化属性生成。HiDe-NeRF(high-fidelity deformable NeRF)[61]采用可变形NeRF(deformable NeRF)来表示3D动态场景，通过学习多尺度体积特征和轻量级变形模块，实现了高保真和自由视角的说话头合成。GeneFace++[62]基于之前的工作[57]改进，将关注点放在更加稳定和实时的生成效果上，引入音高感知(pitch-aware)和标准关键点局部线性嵌入(locally linear embedding, LLE)等技术提高了长音频与唇部运动的同步质量和模型渲染中对异常数据的鲁棒性，采用基于网格的嵌入器和可变形切片表面改进了渲染器，实现了快速训练和实时推理。特别地，随着3D重建技术的发展，使得从单张图片中重建3D表征变得可能[136]，Real3D-Portrait[63]提出了一种从单张图片生成说话人视频的模型，对于说话人视频的头部、躯干和背景3个部分构建不同的生成分支，实现有针对性的渲染和优化，可以同时支持视频和语音2种驱动。Real3D-Portrait基于预训练3D人脸生成模型构建了3D先验知识数据集，训练了一个I2P(image-to-plane)前馈网络以提升人脸重建能力，并设计了视频和音频到PNCC(projected normalized coordinate code)[137]的中间运动表征映射模块，使得眼唇张合幅度可控，能够同时实现准确的3D头像重建和动画，生成了具有自然躯干运动和可切换背景的真实感视频。

3.3.2 基于3D高斯溅射的技术

基于3D高斯溅射的方法以高斯分布表示3D点云，通过点云的密度和分布生成人脸，适合动态场景。相较NeRF而言，3DGS采用了显式建模表征改进了渲染和生成速度，实现了实时的辐射场渲染。具体来说，在重建方面，3DGS采用3D高斯作为场景表示，从相机校准的稀疏点出发，通过优化其位置、协方差、不透明度等高斯参数，结合自适应密度控制创建了精确的场景表示。在渲染方面，设计了基于分块的光栅化(tiled-based rasterizer)快速渲染器，利用GPU排序算法实现各向异性溅射(anisotropic splatting)和近似α混合(alpha blending)，支持高效的反向传播，从而在多种场景下实现了与当前最佳方法相当的视觉质量，同时具备更快的训练速度和实时渲染能力，有效解决了传统方法在速度与质量平衡上的难题。

3D高斯溅射技术发展以来，部分工作[138]探索了基于3DGS技术在单张图像和视频驱动下改变目标人物面部运动的可行性，为基于3DGS的音频驱动说话人技术的发展[64-67]提供了技术参考。由于3DGS空间中点之间的独立性和庞大的数量，导致音频到3D高斯形变空间的映射变得十分困难。Cho等人[64]重建了头部3DGS表征，将3D高斯属性编码到共享隐式特征表示中，通过特征体积表示来增强相邻高斯之间的空间一致性，设计了多分辨率3平面(multi-resolution triplane)提取3D高斯位置的低维特征来描

述 3D 头部; 采用空间到音频的注意力模块与音频特征融合以操纵每个高斯属性, 在面部保真度、嘴唇同步精度和渲染速度方面都取得了较好的效果。类似地, Chen 等人[65] 采用多层神经网络模块融合 3 平面哈希特征和时间平滑处理后的音频特征来预测高斯点的空间偏移, 通过静态高斯初始化和自适应密度控制策略提高了模型的训练和优化效率。Yu 等人[66] 针对特定人物的语音驱动准确性和细节纹理的生成问题, 提出了基于 FLAME[139] 的 3DGS 音频驱动说话头合成新框架, 包括特定说话人运动转换器和动态高斯渲染器 2 个模块, 特定说话人运动转换器通过通用音频编码器、个性化运动解码器和唇读约束机制, 将音频信号转换为特定说话者的 FLAME 参数序列, 结合渲染器实现了对不同说话者的音频特征提取和个性化运动生成。

总的来说, 这些基于 3D 渲染的生成方相较于基于 2D 模型的方法在生成速度和可控性上有较大的优势, 但仍存在一些明显的局限性。首先, 大多数基于 3D 渲染的方法针对特定人物进行训练, 面对新的身份时往往需要重新训练模型, 泛化能力较为有限, 难以直接应用于新人物。此外, 3D 模型的构建和渲染对场景细节及动态精确性的高度依赖, 导致其对输入信号和误差敏感, 生成结果容易失真。这些问题都限制了相关技术的进一步广泛应用。未来可以通过提出更高效的渲染方式与前沿的重建技术在泛化性和实时性方面实现突破, 推动相关领域进一步发展。

3.4 基于自回归的生成技术

基于自回归(auto-regressive, AR)的生成技术是一种依赖时间序列建模的虚拟人驱动方法, 其核心思想是将视频帧或运动系数的生成过程建模为一个条件概率序列, 即每一帧的生成依赖于此前已生成的帧以及当前时间步的驱动信息, 通过逐帧递推的方式实现连续生成。基于自回归的方法大致流程如图 5 所示。该方法通过显式建模帧间的时间顺序, 有效提升了唇形、表情和头部动作的时间一致性, 特别适用于对时序连续性要求较高的说话视频合成任务。

Fan 等人[68] 最早将自回归的思路用于说话头部生成, 提出了一种基于 Transformer 的语音驱动 3D 面部动画生成方法 FaceFormer, 利用全局注意力机制建模音频与面部运动之间的复杂时序关系, 摒弃传统中间表征, 直接从音频生成顶点级别的动态人脸网格。模型采用编码与解码结构, 并引入局部窗口注

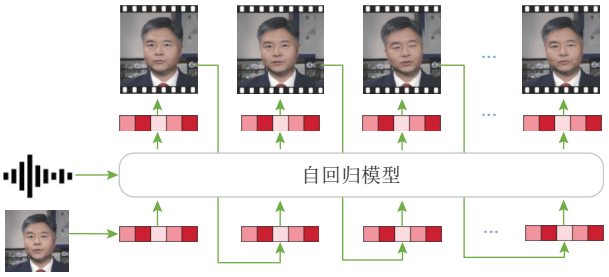


Fig. 5 Autoregressive talking head generation techniques

图 5 自回归的说话头部生成技术

意力与位置编码增强时间建模能力, 自回归地预测连续的面部运动, 在提升表达自然性与时序一致性的同时具备良好的跨身份泛化能力。Richard 等人[69] 提出的 MeshTalk 使用自回归模型在离散潜在空间中生成与语音对应的表情序列, 该模型能够在缺乏表情输入的情况下, 仅依据当前和过去的音频信息, 以及先前生成的潜在标签, 逐步采样出合理的表情表示, 以补充无法根据音频推断的缺少信息。Xing 等人[70] 提出的 CodeTalker 将音频驱动的面部动画生成建模为有限码本空间的查询任务, 使用矢量量化自动编码器(variational quantized autoencoder, VQ-VAE)对真实面部运动进行重建以学习码本。基于码本先验, CodeTalker[70] 采用自回归的码查询模型, 将语音信号逐步映射为一系列码本索引, 再通过解码器还原成面部运动序列。Kim 等人[71] 延续了 CodeTalker[70] 的离散码本与自回归生成思路, 并通过引入多语言风格嵌入和大规模多语种数据集, 有效提升了语音驱动 3D 头像在多语言环境下的口型同步与表达准确性。

基于预定义的特征编码简化了跨模态映射, 但也限制了动作生成的泛化能力。ARTalk[72] 提出了一种多尺度的双层自回归模型, 首先使用时序多尺度 VQ-VAE 按从粗到细的尺度顺序, 将输入的连续面部运动序列编码为多个分辨率下的离散码, 并使用时序因果推理机制保持时间一致性。随后, 使用多尺度双层自回归架构, 选用先前时间窗口与当前窗口更粗尺度的离散码进行迭代, 同时在时间窗口与尺度层级之间进行自回归的面部表示生成, 从而将语音信息映射到离散运动空间。Teller[73] 提出了一种实时语音驱动人像动画生成的流式自回归生成框架, 使用残差向量量化器(residual vector quantization, RVQ)将连续的面部运动编码成离散表示单元(token), 随后利用自回归 Transformer 根据过去的运动表示和音频嵌入逐步生成运动表示, 并采用运动表示对(token pair)并行生成策略提高自回归推理的效率, 实现语

音到面部运动的流式生成。

基于自回归的说话头部生成技术通过条件概率链式建模当前帧的运动参数，具备较强的时序建模能力，能够细致捕捉语音驱动下的动态面部变化，实现高度同步的口型与表情生成。但其迭代生成的策略及有限的参考帧选取，也容易导致错误累积、面部运动过于平滑、实时性较低等缺点。针对上述问题，未来可通过对抗训练与身份自检查机制减少以上序列生成中的偏移问题，引入更合适的上下文感知方法提升语境的理解与控制能力，并通过模型结构优化或并行生成策略以缓解递推生成所带来的速度瓶颈。

3.5 基于强化学习的二元对话生成技术

基于强化学习的二元对话生成技术是一种面向交互式虚拟人应用的新兴方法，其核心目标是在多轮人机对话中，动态生成自然且符合情境的面部表达与响应行为。该方法不再局限于传统的单向语音驱动生成，而是建模对话过程中说话者与聆听者(listener)的动态交互关系，通过强化学习框架自适应地优化生成策略，使虚拟人能够根据语音内容、情感状态和交互反馈生成更加真实、沉浸的面部表现。基于强化学习的方法大致流程如图 6 所示。该技术通常被建模为一个马尔可夫决策过程(Markov decision process, MDP)，将虚拟角色(说话者或聆听者)视为智能体，面部动作或图像帧的生成作为行动，语音、文本、情感和对方表情等构成状态输入。系统通过接收反馈评估生成结果在自然性、协调性和交互性方面的表现，并通过强化学习算法不断优化策略，从而促进智能体在每一轮交互后做出更自然、情境捕捉的面部响应生成。

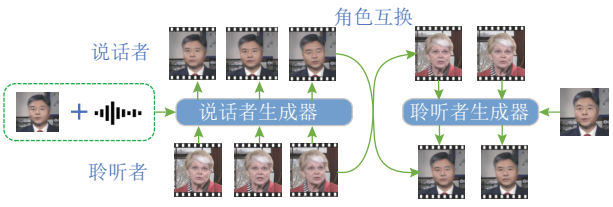


Fig. 6 Facial generation techniques for dyadic dialogues based on reinforcement learning

图 6 基于强化学习的二元对话面部生成技术

Zhou 等人[74-75]最早开始交互式对话头像生成的研究，系统地建模了对话中说话者与聆听者的双角色行为。他们将该问题视为一个视频到视频的翻译任务，分别构建说话者与聆听者的生成模块：其中聆听者模块根据说话者的视频和自身身份图像生成符合说话者语义和非语言信号理解的自然反馈(如点头、微笑)，并通过引入 LSTM 序列模型以及解耦的运动表示，学习独立于身份特征的一般性聆听行为模式；说话者生成则在传统音频驱动基础上，进一步融入聆听者的非语言信号反馈作为控制条件，以增强交互感。最终，对聆听者和说话者通过加权求和的方式进行融合，构造出可动态身份切换、具备理解与表达能力的虚拟交互体，开创了面向真实二元对话的头像生成新范式。Chang 等人[76]在 ViCo[75]的基础上，对音频编码器进行优化，旨在生成能够分层多模态反馈的聆听头。其中使用层级音频编码器提取多个级别的语义特征，并通过对比学习将其与音频中识别的文本特征对齐，通过预训练语言模型 BERT 捕获文本的内容语义，并提取原始语音中的梅尔频率倒谱系数特征，最后将上述不同类型的音频特征进行融合，通过集成多个级别的语音特征，进一步丰富聆听者的理解与表达能力。

上述方法通过顺序型时序建模，使得聆听者的动作生成依赖于历史输出。为缓解自回归生成带来的误差累积问题，ListenFormer[77]提出了一种非自回归的聆听者反馈动作生成方法，该方法采用 2 阶段训练策略：首先在预训练阶段，利用教师强制方案(teacher-forcing scheme)，通过交叉注意力机制融合说话者的音频与运动特征，训练 Transformer 式编码-解码结构对聆听者的视频进行还原；随后在微调阶段，将聆听者的外观图像在时间维度上复制来作为条件输入，在相同的网络结构下进行非自回归推理，以生成稳定自然的聆听动作。这一设计不仅避免了时序误差扩散，还提升了头部姿态和表情的一致性。

此外，该类技术还可扩展为多智能体学习架构，同时训练说话者与聆听者 2 种角色的面部生成策略，从而实现更真实、沉浸式的虚拟二元互动效果。INFP[78]采用 2 阶段学习的方式，第 1 阶段将对话视频的面部动作投影到低维潜在空间，并使用潜在动作驱动静态图像生成视频。第 2 阶段利用语言与非语言记忆库和交叉注意力机制，根据对话双方的语音特征，动态构建驱动虚拟角色说话或倾听状态的交互式运动特征。随后通过一个轻量级条件 DiT[11]，将交互式运动特征映射到多帧运动潜在空间，结合时间建模和跨窗口条件，生成自然流畅的说话或聆听动作。这种设计不再需要手动分配角色，而是根据输入的双人对话音频动态切换身份。

与传统的监督学习方法相比，基于强化学习的生成方式不再依赖严格的 1 对 1 配对监督信号，而是

通过长期交互反馈自主学习生成策略，因此更适合建模开放域、多轮对话等复杂语境下的人脸响应生成任务。但与此同时，该方法通常需要大量的交互数据与训练轮次以实现策略收敛，计算成本较高，训练过程容易出现不稳定或策略陷入局部最优的问题。因此，尽管强化学习在构建具备长期规划与自适应能力的虚拟人系统中展现出广阔前景，但其在实际部署中的复杂性与高资源需求仍是急需解决的关键问题。

3.6 条件驱动的生成技术

视频与语音驱动的说话人生成技术在生成效果上展现了不错的能力，但对于表情、姿态和风格的细粒度生成还存在改进空间。部分工作通过引入表情描述、姿态引导和风格视频参考等新型驱动信息，在既有音频与视频驱动技术的基础上，提出新的驱动方式和架构，相关条件驱动技术的汇总如表 3 所示。这些创新进一步丰富了说话人的表情、姿态与个性化风格表现，使得生成的视频变得更加生动和可控。

3.6.1 情绪驱动说话人生成技术

情绪驱动说话人生成技术一般通过图片、视频和文本标签的形式在生成过程中引入情绪信息，以更准确地控制和生成说话人在特定情境下的面部表情、唇形变化及情感表现，使得说话人在视频中显得更加生动和富有表现力。

将表情图片和视频作为情绪驱动信息的工作大多在生成架构中构建额外的表情编码器来提取情绪面部运动，并将其与编码语音获得的运动信息相结合进行人脸生成。Wang 等人[79]采用音频到标准关键点模块将音频映射到嘴唇运动，再采用情感转换器结合驱动情感图像，在目标人脸上合成了强度可控的所需情感，并构建了情感视听数据集 MEAD。EAMM[81]在驱动输入信息中引入情绪类型图片视频，采用隐式情感位移学习器将驱动信息中情绪合成到该面部同态中，生成了高质量、情感细腻的符合驱动音频和情绪的说话人视频。类似的，Dream-Talk[84]提出了基于 DiT[11]的表情提取模块 EMODIff，将音频信息和表情风格参考图作为驱动条件来控制生成说话人的表情风格，并把唇部单独分离出来处理，减少表情对唇形同步的影响。EMOPortraits[82]基于视频驱动技术 MegaPortraits[140]改进，用原架构的身份和运动编码器提取源身份信息和驱动情绪信息，增添语音编码器以生成与语音相匹配的运动信息，在情绪迁移任务上表现出优秀的性能。

使用文本标签作为情绪驱动信息的技术一般采用卷积神经网络编码标签来获得文本特征，并将其与语音特征合并来学习到运动信息的映射。吕天等人[85]通过情感编辑的形式为说话头增加情感，利用基于 Basel 人脸模型条件的人脸几何编辑网络实现多目标人物的几何形状修改，同时通过结合纹理生成对抗网络与人物分类器引导的隐向量编辑，解决多人任务中情感编辑方向不一致的问题。Sinha 等人[80]构建了基于语音内容特征的图卷积神经网络，经音频编码器和表情编码器处理音频和文本表情驱动信息后，在面部几何感知标准关键点表示上产生对应运动，以保留面部结构和几何形状以进行情感渲染。EAT[83]基于给定的表情文本学习深度可用的

Table 3 Summary of Talking Head Generation Technologies Driven by Conditions

表 3 条件驱动说话人生成技术汇总

年份	模型	驱动信息	基础模型	中间表达	表情驱动	头部驱动	姿态驱动
2020	MEAD[79]	音频、表情图像	GANs	landmarks	√		
2021	Wu 等人[86]	音频、风格视频	GANs	3DMM	√	√	
2022	Sinha 等人[80]	音频、表情标签	GANs	landmarks	√		
2022	EAMM[81]	音频、情绪视频	LSTM	隐向量	√		
2022	GC-AVT[87]	音频、视频	Transformer	隐向量	√	√	
2023	EMOPortraits[82]	音频、表情图像	CNN	隐向量	√	√	√
2023	EAT[83]	音频、情绪标签	GANs	3D 隐式关键点	√		
2023	Dream-Talk[84]	音频、情绪标签	Diffusion	blendshapes	√		
2023	文献 [85]	音频、情绪标签	GANs		√		
2023	StyleTalk[88]	音频、视频	GANs	3DMM	√		
2024	SAAS[89]	音频、视频	GANs	3DMM	√	√	
2024	Style2Talker[90]	音频、情绪文本、艺术图像	Diffusion，GANs	3DMM	√		
2024	StyleTalk++[91]	音频、视频	GANs	3DMM	√	√	

prompt, 利用轻量级情感变形网络(emotional deformation network, EDN)将表情形变添加到人脸编码模块获得的增强3D关键点中, 借助即插即用的情感适应模块(emotional adaptation module, EAM)进一步提升视觉效果。通过轻量化表情引导生成模块和可学习的提示词, EAT把预训练的语音驱动模型迁移到表情驱动上, 在一定程度上提高了任务的效率。

3.6.2 风格驱动说话人生成技术

风格驱动说话人生成技术引入了风格视频或图片信息, 从中学习包括头部姿态、表情变化以及艺术风格等多维度的特征。这些技术使得生成的说话人不仅能在情感上表现得更自然, 还能够在视觉风格上展现更多元化的个性化表现, 能够制作出更加有表现力的说话人视频。

Wu等人[86]提出了一个可以模仿任意说话风格的音频驱动生成框架, 收集了包含834个谈话风格的视频数据集Ted HD。设计了潜在风格融合(LSF)模型, 从音频中学习与运动相关的信息, 将从参考视频中提取的风格代码与潜在音频表示逐帧融合, 生成具有目标说话风格的视频。GC-AVT[87]将参考图像分别处理成用于姿态学习的头部蒙版、用于口型对齐的唇周部分和用于表情学习的脸部上部分。核心模型采用了基于Transformer的跨模态转换结构, 协同不同部分组件, 使得生成的说话人头不仅口型准确, 而且在表情、头部姿态等方面都能够进行细粒度的控制。SAAS[89]采用3DMM作为中间特征, 结合头部运动生成器, 将3D信息渲染解码以输出最终的说话人视频, 采用了离散表示学习和动态权重调整技术, 生成了生动的风格一致的说话人视频。大多数包含视频信息驱动的方法都只是在静态帧层面复刻面部运动特征, 而忽略了时间序列上的风格特征。Wang等人[91]基于他们之前的StyleTalk[88]方法, 提出了StyleTalk++, 设计了一个通用风格编码器来模拟任意参考风格视频中面部表情和头部姿势的运动模式, 从参考风格视频的连续3DMM表情或头姿中提取潜在的序列表情和姿态代码, 将这些代码与音频信息和身份信息相结合, 最终渲染输出视频。其相较StyleTalk[88]拓展了头部姿态, 生成了更生动的视频。Style2Talker[90]将注意力放在艺术风格上, 将文字描述的情感风格和图片展示的艺术风格作为新的驱动信息, 利用CLIP[141]模型获取情绪表征, 采用基于扩散模型的端到端方式生成包含情绪和音频信息的3DMM运动序列, 基于StyleGAN[123]进一步融合身份和艺术风格信息生成最终视频。

4 姿态引导的2D虚拟人驱动技术

由于面部信息的丰富性和生成的复杂性, 2D虚拟人领域目前更多地聚焦于面部的驱动和生成, 但也有部分工作关注着面部之外的动作生成, 采用姿态引导生成, 为2D虚拟人添加各种躯体动作。

姿态引导2D虚拟人驱动技术是一种通过输入参考音视频、骨骼关键点或动作序列, 驱动虚拟人图像视频生成动态动画的技术。该技术通过深度学习模型或基于关键点的运动匹配算法, 使虚拟人动作与输入姿态高度一致, 其核心是实现输入姿态与虚拟人图像之间的精准映射和动态协调, 包括身体、手部、头部等部位的动作控制, 被广泛应用于舞蹈生成、游戏角色控制、实时直播等场景。

4.1 图像到图像的姿态驱动技术

图像到图像的姿态驱动技术通过输入一张静态图像以及目标姿态信息, 基于条件GANs或VAE生成具有目标姿态的图像, 任务和方法都比较简单, 适用于生成静态图像的任务, 如虚拟试衣、艺术绘制。Ma等人[92]于2017年首次提出了姿态引导图像生成任务, 利用条件GANs进行端到端生成, 奠定了姿态驱动图像生成的基础, 但生成结果在复杂背景下的细节保真度较低。Lassner等人[93]针对服饰细节的准确生成提出了ClothNet, 结合了人体姿态和服饰特征, 实现了自然感较强的换装效果, 但对于复杂服饰纹理的生成仍存在一定误差。后续Ma等人[94]和Joo等人[95]分别采用姿态身份解耦和特征分离融合等技术, 实现了对姿态和人物身份的独立控制。De Bem等人[96]针对姿态大变形情况下的生成问题, 引入了变形卷积模块和最邻近损失, 提升了生成效果。Balakrishnan等人[97]则是采用跨姿态对齐模块, 有效减少了源姿态和目标姿态之间的结构差异, 重点解决了从未见过的目标姿态生成问题。此外, 还有一些其他工作[98-99,142], 针对生成的质量、背景一致性和细节纹理特征等方面进行了优化。这些方法总体上推动了姿态驱动虚拟人生成技术的发展, 但生成效果和应用场景都较为有限。

4.2 视频到视频的姿态驱动技术

视频到视频的姿态驱动技术通过输入一个源视频和动作视频, 从中构造流场或提取目标姿态序列, 生成目标姿态驱动的输出视频。这类技术需要保证生成视频的时序一致性和运动自然性, 生成难

度远大于姿态引导的单图生成。早期工作主要采用构造运动流场的方式并基于 GANs 实现视频帧生成, 如 Vid2Vid[100], few-shot vid2vid[101], Liquid-Warping-GAN[102] 等。基于流场的姿态视频生成技术大多不仅限于人体姿态的生成, 可以应用于人脸或者物体运动迁移, 在前面视频驱动说话人生成技术中提到的 Monkey-Net, FOMM, TPSM 等也都属于这一类方法。

目前大多姿态生成技术都采用骨骼图作为中间特征, 结合骨骼图提取技术(如 OpenPose[143], DensePose[144], DWPose[145], SMPL[146] 等)从视频中获得动作序列, 用以生成或重构目标人物在不同姿态下的图像序列。Chan 等人[103] 利用骨骼图作为中间特征, 引入额外的 GANs 关注脸部的生成, 实现了较复杂的运动迁移, 但生成质量有限。随着高性能大规模文生图模型 Stable Diffusion 的提出, 姿态生成领域也涌现出了许多基于此模型的新技术。DreamPose[104] 采用 DensePose 序列作为驱动信息, 将 Stable Diffusion 原有的文本编码器替换为自定义的条件适配器, 结合 CLIP-VAE 编码器的编码信息增强身份一致性, 实现了简单序列参考图像的替换。AnimateAnyone[105] 使用 DWPose 和 OpenPose 生成骨架图, 采用 ReferenceNet 编码身份信息, 设计了轻量级姿态引导器, 在 U-Net 中添加时间层实现姿态引导与时间平滑, 在保持身份一致性方面表现突出, 其网络结构如图 7 所示。MagicAnimate[106] 架构将 2D U-Net 扩展到 3D, 并对 DensePose 序列进行微调, 从而改善了人体视频生成的平滑性。DisCo[107] 借鉴 ControlNet[128] 的思想, 解耦背景和姿态控制, 生成的人物运动精确, 在新视角和结构上表现较好, 但细节表现欠佳。Champ[108] 和 MimicMotion[109] 针对姿态对齐和优化问题进行了进一步探索。Champ[108] 基于 SMPL 提取多种 3D 人体参数, 将其渲染到骨骼图、深度图、语义图等信息, 通过多层运动融合模块提升了对复杂运动的建模能力, 但计算成本较高。MimicMotion[109] 针对长视频和手部生成问题提出了创新性解决方案, 通过置信度感知机制优化手部关键点生成, 同时利用时域渐进融合增强时间连续性和细节一致性, 提高了长时间动态场景的生成质量。

总的来说, 基于骨骼图的视频到视频的姿态驱动技术能够精准表达运动姿态, 适用于姿态迁移、动画制作等任务。通过先进架构和时间平滑策略的提出, 生成序列较为流畅自然。但生成需要多阶段处理, 如基于预训练模型[143-146] 的骨骼提取、基于 ControlNet 的背景分离和条件引入、基于 Stable Diffusion 的扩散生成, 导致计算资源需求较高, 难以满足实时性要求; 且在复杂场景(如手部、衣物细节)或高动态姿态下, 生成细节和精度仍有提升空间。同时, 尽管有泛化能力的优化, 但大多方法都针对特定人物, 许多方法在新人物或未见场景下的表现仍不够理想, 需要进一步的探索和优化, 以提高视频到视频的姿态

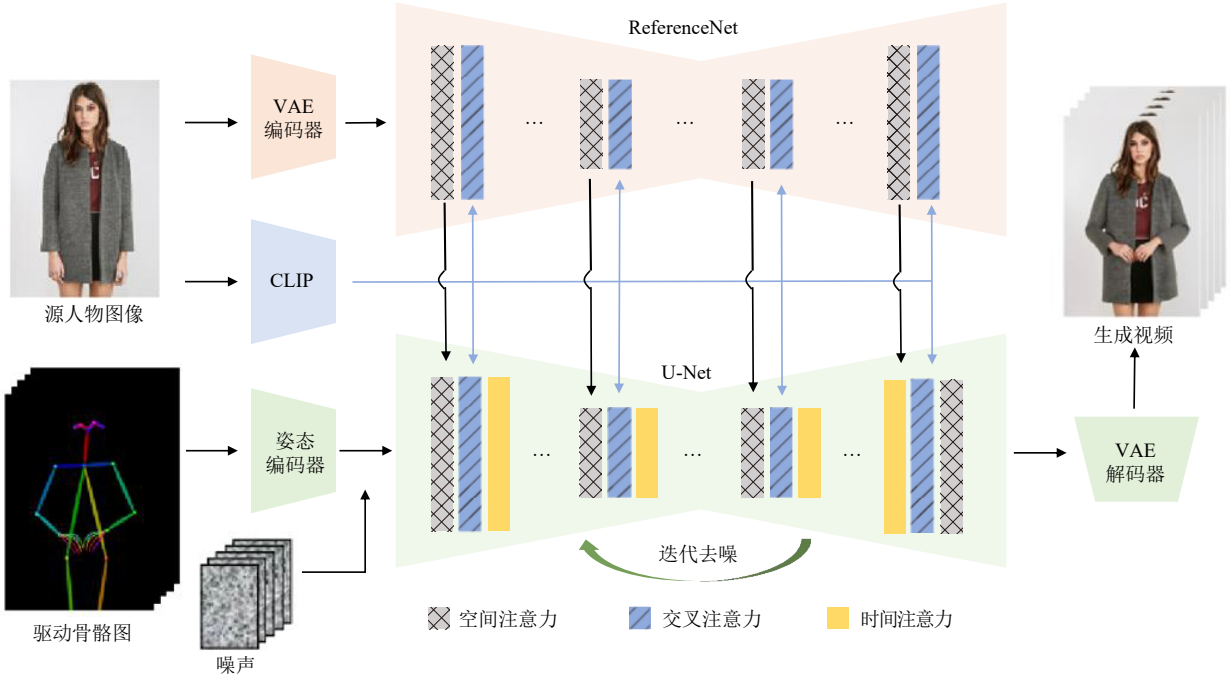


Fig. 7 Video-to-video pose-guided generation technology based on diffusion model[105]

图 7 基于扩散模型的视频到视频姿态引导生成技术[105]

驱动技术的实时性、细节性和泛化性。

4.3 文本辅助的姿态驱动技术

扩散模型的发展也促进了文本生成视频领域的突破[147]，除了采用视觉图像和视频信息作为身份和背景来源之外，现有技术[110-113]还使用文本提示信息作为人物和背景视频生成条件，结合动作序列生成更加个性化的目标人物视频。甚至有些技术直接一步到位，采用文本描述人物运动，直接生成符合要求的视频[113]。Follow Your Pose[110]有效利用图像姿态对和无姿态视频数据集，将训练分为侧重于姿态对齐和时间一致性2个部分，有效引入了姿态控制，提高了视频的时间一致性。DancingAvatar[112]利用ChatGPT理解服饰和人脸文本提示，结合预训练文生图模型生成对应的目标人物形象，设计了帧内、帧间、背景对齐模块，结合分割和图像修复技术保证背景一致性，基于自回归方式增强帧间细节一致性提高了视频的整体连贯性。目前相关领域仍然在发展中，生成质量与直接视觉信息引导的生成相比，稳定性更差，帧间抖动也更明显。随着文本生成视频技术的发展，相信未来在生成质量和长视频生成方面都会有所进展。

4.4 语音到视频的姿态驱动技术

语音到视频的姿态驱动技术在给定一段驱动语音的情况下，从中生成运动序列，驱动目标人物生成对应姿态运动视频。由于语音信息和姿态信息之间的弱相关性，技术训练难度较高且数据集稀缺，相关技术也较少。早期的研究[114-115]只能针对特定人物进行驱动，其中Speech2Gesture[114]收集了144 h的10个特定说话人手势视频，发现语音和手势之间存在弱相关性，提出一种语音生成手势模型。VLOGGER[116]和CYBERHOST[117]分别基于扩散模型和最新的语音驱动说话人生成架构，为语音到视频的姿态驱动技术带来了新的可能。VLOGGER[116]设计了2个扩散模型，一个用于学习音频到面部和上半身运动的映射，另一个用于扩散生成视频，采用了时间外延(temporal outpainting)技术来基于前面帧迭代生成后面帧，优化了视频的时间一致性，实现了从单张图像与音频生成半身虚拟人说话视频。CYBERHOST[117]构建了单阶段音频驱动的人体动画框架，设计了区域码本注意力机制和一套针对音频驱动场景的人类先验引导训练策略，包括身体运动图、手部清晰度分数、姿态对齐参考特征和局部增强监督等，借助新的监督和先验信息提高了合成结果的质量和稳定性。

目前，语音到视频的姿态驱动技术仍存在生成质量不稳定的问题，语音和姿态之间的弱相关性也导致模型训练和推理的困难，限制了相关技术在复杂场景的应用前景。未来可以考虑结合先验信息和其他模态信息来辅助学习语音到姿态的映射，生成更真实的语音驱动姿态视频。

5 虚拟人集成驱动系统与现实应用

集成式(holistic)虚拟人生成方法是一类融合多模态、多维度运动控制的综合生成技术，旨在同时建模并生成虚拟人的嘴型同步、面部表情、头部运动以及身体姿态变化等多个层次的表现，从而实现高保真、高一致性和高沉浸感的虚拟人驱动系统。与传统只关注表情或姿态的单部位生成方法不同，集成方法强调各个模态间的整体协调性、时序一致性与语义一致性，在语音驱动的基础上实现完整的虚拟人行为建模。根据生成模型对特定人物身份的依赖程度，集成式虚拟人生成方法可分为身份依赖(identity-dependent)和身份独立(identity-independent)2类技术。身份依赖方法通常基于特定人物的数据进行建模与训练，能够在目标人物上实现更为还原的生成效果，但由于泛化能力有限，这类模型难以迁移至训练数据之外的身份，因而主要适用于个性化定制等封闭式场景。相比之下，身份独立方法通过身份信息分离、驱动信号映射与动作解耦生成等机制，学习训练数据中蕴含的普适性生成规律，该类模型可在任意人物身份上复用，具备更强的泛化能力和通用性，更适用于开放域场景下的虚拟人构建。

5.1 身份依赖的虚拟人生成技术

Habibie等人[118]最早开始面部表情与身体姿势联合生成的研究，提出了一种基于CNN的方法，通过共享的音频编码器捕捉面部、躯干和手部之间的内在关联，随后使用3个独立的解码器分别生成各部位的关键点序列，并在训练过程中使用判别器进行对抗学习。TalkSHOW[119]将不同模态的动作建模任务进行解耦：其中面部动作通过一个条件自动编码器生成，该编码器以音频特征和身份信息作为输入，实现个性化的面部表情生成；而身体动作和手势则通过分区域训练的组合VQ-VAE进行建模，学习离散的表示空间以捕捉复杂的运动结构。在动作生成阶段，TalkSHOW采用交叉条件自回归模型，在融合语音MFCC特征与身份信息的条件下，按时间顺序预测身体与手部的离散码索引，通过交叉条件建模实现多部位动作的同步性。最终，通过查表和解

码方式恢复高质量的3D身体与手部动作。

TalkSHOW[119]通过引入码本机制对音频驱动信号进行离散化处理,借助向量量化增强建模效率,有效降低了音频到动作映射过程中的复杂度。然而,这种基于固定码本的离散表示却在一定程度上牺牲了生成动作的多样性与灵活性,难以充分捕捉更复杂的姿态变化。DiffSHEG[120]观察到面部表情与身体姿势的单向条件依赖关系,在基于扩散模型的统一框架中引入了从表情到姿势的单向信息流,以表情作为姿势生成的必要条件,从而捕捉两者的联合分布关系,并统一语音同步的表情和姿势生成。此外,DiffSHEG提出一种基于快速外绘机制的局部自回归采样策略,使模型具备任意长度视频生成能力,并支持高效的实时流式推理。

上述方法在泛化能力上仍较为有限,通常需要针对特定人物进行个性化微调,难以直接适配开放域的新角色。对此,最近的研究逐渐针对通用型虚拟人生成来进行。

5.2 身份独立的虚拟人生成技术

DreamActor-M1[121]与HunyuanVideo-Avatar[122]是近年来集成驱动虚拟人生成技术的代表性成果,均实现了面部与全身动作的统一建模,具备高度可定制性与商业适应性,展现出强大的实用价值与落地潜力。

DreamActor-M1[121]是一种基于DiT的混合引导框架,有效提升人体图像动画在多尺度驱动适应性、长时间一致性和细粒度整体可控性方面的表现。具体而言,该方法引入集成了隐式面部表示、3D头部球体与身体骨架的混合控制信号,实现对面部和身体动作的多信号鲁棒控制。同时引入互补外观引导(complementary appearance guidance)机制,利用目标运动中采样的多姿态参考帧来填补参考图像中不可视区域的缺失信息,以增强长时间生成的稳定性。通过对于不同尺度数据的渐进式训练策略,模型可适应肖像交谈至全身舞蹈等多尺度场景,显著提升了人像动画的整体性和表现力。最终实现具有混合引导的整体、富有表现力和鲁棒性的人体图像动画。

HunyuanVideo-Avatar[122]提出了一种用于生成维持身份动态一致、情感可控、多角色人像动画的高保真音频驱动方法,解决了传统方案在角色一致性、情感对齐和多角色场景中的关键难题。该方法首先引入角色图像注入模块,替代以往加法形式条件机制,将人体图像特征转换为更易于模型学习的表示,沿通道维度注入图像特征,显著提升角色外观一致性与动作连贯性。其次,引入音频情感模块(audio emotion module, AEM),利用情感参考图像引导实现对音频中情绪的精准提取与迁移,实现对人物动画的细粒度情感风格控制。最后,设计面部感知音频适配器(face-aware audio adapter, FAA),通过潜在空间中面部掩码的处理与交叉注意力机制,实现多角色场景下不同角色间的独立音频驱动,有效支持多角色高质量对话动画生成。

随着深度学习和多模态融合建模能力的不断演进,集成驱动虚拟人生成技术正朝着高质量生成、多角色协同、精细化控制与低成本部署4个方向快速发展。一方面,提升虚拟人的视觉真实感与运动自然性成为评价系统能力的重要标准;另一方面,多人物并发驱动与风格迁移、情绪表达控制等可定制功能,也正在成为新一代虚拟人系统不可或缺的核心能力。同时,低资源环境下的轻量化建模、边缘部署与高效渲染也逐步进入研究与产业化视野,为大规模应用奠定基础。

总体而言,集成驱动方法凭借其在多模态建模能力、交互复杂度适应性与内容生成一致性等方面的综合优势,已成为满足高真实感、多功能性与用户可控性等关键需求的技术路径。它不仅能够支撑当前虚拟人技术在社交互动、内容创作、虚拟直播与智能客服等多场景中的实际应用,更具备良好的扩展性和商业化前景[148]。尽管如此,该类方法在跨模态特征对齐、统一空间建模、生成稳定性与实时推理效率等方面仍存在较大改进空间。未来可以进一步探索轻量化统一模型架构、多尺度协同生成机制及稳健的风格迁移策略,以实现更强的泛化能力与更低的系统资源开销,从而推动集成驱动虚拟人技术在更广泛的开放场景与互动式应用中的落地与持续发展。

6 挑战与展望

尽管2D虚拟人驱动技术在近年来取得了显著进展,但在技术效果和应用实现等方面仍面临诸多挑战。这些挑战主要集中在改进生成效果、提高时间一致性、构造高质量数据集、提高模型实时性和降低计算资源等方面。

1)改进生成效果。2D虚拟人驱动技术的核心目标是生成自然、逼真且具有丰富面部和身体运动的虚拟人。然而,现有技术在捕捉或根据驱动信息生成面部细节运动上仍然存在不足。在表情方面,现

有技术在处理复杂情绪表达时往往表现不佳，无法完全搬运驱动视频的表情，难以从音频中生成具有丰富细节的表情，解决这一问题需要更细致的表情数据标注和多层次的表情控制技术的发展。在头部运动方面，现有模型的头部运动常常显得僵硬或不协调，快速的头部转动或细小的姿态调整可能因生成过程不连续而失真，未来需要进一步研究头部动作平滑等技术，缓解相关方面的不自然感。在唇形同步方面也需要进一步优化唇形生成的细粒度控制和动态响应能力，将唇形同步精确到帧甚至更细的粒度，以适应快速语速、多语种和多主题的音频。

2）提高时间一致性。时间一致性是2D虚拟人生成技术中的一个关键挑战。长时间视频生成需要模型对序列信息有较强的记忆和关联能力，但现有效果较好的生成模型通常只能关注较短时间内的图像生成。对于较长的时间序列，由于模型缺乏对上下文的持续记忆和关联，生成的虚拟人表情、动作在长时段内可能出现突兀的跳跃或不一致的变化。虽然现有技术通过引入运动帧和时间模块试图解决这一问题，但模型仍然难以生成具有完整运动表达的长视频，且随着视频时间的增加生成效果显著下降。未来工作可以探索通过时序一致性约束、时序生成模型、多模态对齐与噪声控制等方法来改进生成视频的时间一致性，提升用户体验。

3）构建高质量面部和姿态数据集。2D虚拟人驱动技术的训练效果依赖于高质量的数据集。然而，在现实条件下，同时满足大规模、高分辨率、高一致性和强多样性的高质量数据集难以获取。大多数数据集的片段较短，且身份数量和场景有限。这导致模型在训练过程中缺乏持久性和多样性特征，难以生成时间连续且运动丰富的虚拟人表现。现有优秀工作大部分都是企业发布，它们在自收集的数据集上进行深度训练得到模型，相关领域缺乏公开的高质量数据集，后续应当更加注重高质量开源面部和姿态数据集的构建，为更多的生成模型提供强有力的数据集支持。

4）训练支持多语种的生成模型。在2D虚拟人生成技术中，支持多语种生成的模型具有极大的应用潜力，能够满足全球用户的需求，能够广泛应用于多语种的虚拟助手、跨语言教育、国际化直播等场景。但由于数据集的欠缺，导致训练多语种模型变得困难；且由于不同语种之间发音和情绪表达的不同，限制了迁移学习在该项任务上的表现。未来可以利用更多先验信息，考虑设计通用的多语言口型编码器，引入细粒度的语言差异，构造一种支持多语种的生成模型。

5）降低计算资源，提高模型的实时性。在虚拟人生成系统中，降低计算资源消耗、提升模型的实时性已成为制约其大规模落地的关键问题。当前主流方法多依赖于大规模的Diffusion[3]，Transformer[7]，StyleGAN[123]架构进行高质量视频生成，这类模型虽然具备出色的表现能力，但往往伴随着巨大的计算和存储开销，经典的StyleGAN2模型参数量约为36×10^6，在NVIDIA GeForce GTX 1080Ti显卡上生成512×512分辨率虚拟人图像时，推理帧率仅为5~8 FPS[57]；基于Diffusion架构的Stable Diffusion 1.5参数量达860×10^6，采用原始50步DDPM采样时，在RTX 3090显卡上生成单帧512×512虚拟人图像需1~2 s，换算为视频流时帧率不足2 FPS，难以适应移动设备或边缘平台的部署需求。在实际应用中，尤其是在虚拟助理、智能客服和虚拟主播等要求及时响应的应用场景中，2D虚拟人生成模型的实时性通常是影响用户体验的关键。因此，如何在有限的硬件资源下保证生成质量与推理效率之间的平衡，成为研究的重点方向。未来的研究可以重点从结构优化与模型压缩两大方向着手，提升虚拟人生成系统的实时性与计算效率。结构优化方面，可通过引入轻量化网络架构（如MobileNets[149]，EfficientNet[150]等），结合模块化设计与动态计算路径，减少冗余计算开销；同时，设计高效且延迟可控的推理流程有助于平衡生成质量与响应速度。模型压缩技术则包括知识蒸馏，通过教师-学生模型迁移丰富特征表达能力，模型剪枝与权重量化则在保证精度的前提下显著压缩模型规模，降低计算资源需求。此外，针对扩散模型采样步骤冗长的瓶颈，已有DDIM[4]，DPM-Solver[151]等快速采样方法被提出，极大缩短了生成时间，提高了实用性。综上，实现高效、低资源消耗的虚拟人生成系统，不仅需要在建模层面对网络结构进行精简与重构，还应结合目标硬件平台特性，优化推理路径与执行策略，从而显著提升整体系统的实时响应能力，推动其在智能交互、虚拟直播等实时应用中的广泛部署。

7 总　　结

本研究聚焦于2D虚拟人驱动技术中基于深度学习的方法，首先介绍了虚拟人驱动的基本原理及常用的深度学习基础模型，为后续方法的分类与分析奠定理论基础。在此之上，围绕驱动部位的划分

视角,系统梳理了口型与面部表情、头部姿态、身体动作等关键驱动单元的建模方式与主流驱动方法,深入探讨了各类模型架构的设计特点与适用场景,并注重不同方法间的对比分析。最后汇总并分析了面向整体虚拟人生成的集成驱动策略,旨在揭示当前虚拟人生成系统向统一建模框架演进的技术趋势与集成挑战。

本综述旨在梳理2D虚拟人驱动技术的研究现状与发展脉络,为该技术的系统化理解提供技术参考,围绕驱动单元与建模架构2个核心视角展开全面梳理,构建了较为完整的知识框架。通过分部位与系统集成方法的逐层展开,不仅揭示了当前主流方法在模型设计、数据依赖与生成效果等方面的差异,也展现了多模态融合驱动在实现高一致性、高真实度虚拟人中的关键作用。然而,尽管现有方法已在生成效果上取得显著进展,但在实时性能、细节复现及跨域泛化能力方面仍存在显著挑战。未来工作可进一步探索轻量化与流式模型架构、多模态条件建模策略,以及高效的数据自动标注与自监督学习机制,以推动2D虚拟人驱动技术在真实交互与个性化生成中的实用性与适应性发展,为虚拟互动和娱乐体验带来新的可能。

作者贡献声明:杨璐嘉负责文献调研并撰写驱动部分;周溪石负责撰写基础模型和集成部分;王俊超负责部分文献整理和论文勘误;张金山提出指导意见并修改论文;金涛对论文提出修改意见;尹建伟提出总体指导意见。

参 考 文 献

[1] Wei Liyi, Lefebvre S, Kwatra V, et al. State of the art in example-based texture synthesis[J]. Eurographics 2009, State of the Art Reports, 2009: 93−117

[2] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial networks[C]//Proc of the 27th Int Conf on Neural Information Processing Systems. Red Hook, NY: Curran Associates Inc, 2014: 2672−2680

[3] Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models[J]. Advances in Neural Information Processing Systems, 2020, 33: 6840−6851

[4] Song Jiaming, Meng Chenlin, Ermon S. Denoising diffusion implicit models[J]. arXiv preprint, arXiv: 2010.02502, 2020

[5] Song Yifei, Zhang Wei, Chen Zhineng, et al. A survey on talking head generation[J]. Journal of Computer-Aided Design & Computer Graphics, 2023, 35(10): 1457−1468 (in Chinese)
(宋一飞, 张炜, 陈智能, 等. 数字说话人视频生成综述[J]. 计算机辅助设计与图形学学报, 2023, 35(10): 1457−1468)

[6] Bao Yihua, Tu Ziqi, Luo Le, et al. Digital human facial expression transfer: A review from face representation to expression transfer[J]. Journal of Computer-Aided Design & Computer Graphics, 2025, 37(8): 1283−1299 (in Chinese)
(包仪华, 涂子奇, 骆乐, 等. 数字人面部表情迁移: 从人脸表示到表情迁移综述[J]. 计算机辅助设计与图形学学报, 2025, 37(8): 1283−1299)

[7] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Proc of the 31st Int Conf on Neural Information Processing Systems. Red Hook, NY: Curran Associates Inc, 2017: 6000−6010

[8] Ren Yi, Liu Jinglin, Zhao Zhou. Portaspeech: Portable and high-quality generative text-to-speech[J]. Advances in Neural Information Processing Systems, 2021, 34: 13963−13974

[9] Kingma D P, Welling M. Auto-encoding variational Bayes[C/OL]// Proc of the 2nd Int Conf on Learning Representations. 2014[2025-07-04]. https://openreview.net/forum?id=33X9fd2-9FyZd

[10] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation[C]//Proc of the 18th Int Conf on Medical Image Computing and Computer-Assisted Intervention. Berlin: Springer, 2015: 234−241

[11] Peebles W, Xie S. Scalable diffusion models with transformers[C]// Proc of the IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2023: 4195−4205

[12] Suwajanakorn S, Seitz S M, Kemelmacher-Shlizerman I. Synthesizing Obama: Learning lip sync from audio[J]. ACM Transactions on Graphics, 2017, 36(4): 95: 1−95: 13

[13] Jamaludin A, Chung J S, Zisserman A. You said that?: Synthesising talking faces from audio[J]. International Journal of Computer Vision, 2019, 127(11): 1767−1779

[14] Zhou Hang, Liu Yu, Liu Ziwei, et al. Talking face generation by adversarially disentangled audio-visual representation[C]//Proc of the 33rd AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2019: 9299−9306

[15] Zhou Yang, Han Xintong, Shechtman E, et al. MakeitTalk: Speaker-aware talking-head animation[J]. ACM Transactions on Graphics, 2020, 39(6): 1−15

[16] Prajwal K, Mukhopadhyay R, Namboodiri V P, et al. A lip sync expert is all you need for speech to lip generation in the wild[C]//Proc of the 28th ACM Int Conf on Multimedia. New York: ACM, 2020: 484−492

[17] Wang Jian, Qian Xinyuan, Zhang Malu, et al. Seeing what you said: Talking face generation guided by a lip reading expert[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2023: 14653−14662

[18] Zhang Zhimeng, Hu Zhipeng, Deng Wenjin, et al. DINet: Deformation inpainting network for realistic face visually dubbing on high resolution video[C]//Proc of the 37th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2023: 3543−3551

[19] Wei Qingyang, Xu Shugong. Research on few-shot speaker video generation algorithm guided by sketches[J]. Computer Measurement & Control, 2024, 32(10): 236−242, 249 (in Chinese)
(魏清杨, 徐树公. 基于草图引导的少样本说话人视频生成算法研

究[J]. 计算机测量与控制, 2024, 32(10): 236−242, 249)

[20] Chen Shihang, Sun Yubao. Voice-controlled face video adversarial generation based on modal affine fusion[J/OL]. Computer Engineering, [2025-07-04]. https://www.ecice06.com/CN/10.19678/j.issn.1000-3428.0069992(in Chinese)
(陈诗航, 孙玉宝. 基于模态仿射融合的语音控制说话人脸视频对抗生成[J/OL]. 计算机工程[2025-07-04]. https://www.ecice06.com/CN/10.19678/j.issn.1000-3428.0069992)

[21] Guo Xingxing, Xiao Yannan, Wen Peizhi, et al. Audio driven digital facial video generation method based on attention mechanism[J/OL]. Computer Science, 2025: 1−11. [2025-07-04]. http://kns.cnki.net/kcms/detail/50.1075.TP.20250317.1902.040 (in Chinese)
(郭星星, 肖雁南, 温佩芝, 等. 基于注意力机制的音频驱动数字人脸视频生成方法[J/OL]. 计算机科学, 2025: 1−11. [2025-07-04]. http://kns.cnki.net/kcms/detail/50.1075.TP.20250317.1902.040)

[22] Siarohin A, Lathuilière S, Tulyakov S, et al. First order motion model for image animation[J/OL]. Advances in Neural Information Processing Systems, 2019, 32[2025-07-04]. https://proceedings.neurips.cc/paper/2019/hash/31c0b36aef265d9221af80872ceb62f9-Abstract.html

[23] Ha S, Kersner M, Kim B, et al. MarioNETte: Few-shot face reenactment preserving identity of unseen targets[C]//Proc of the 34th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2020: 10893−10900

[24] Nian Fudong, Wang Wentao, Wang Yan, et al. Speech driven talking face video generation via landmarks representation[J]. Pattern Recognition and Artificial Intelligence, 2021, 34(6): 572−580 (in Chinese)
(年福东, 王文涛, 王妍, 等. 基于关键点表示的语音驱动说话人脸视频生成[J]. 模式识别与人工智能, 2021, 34(6): 572−580)

[25] Wang Tingchun, Mallya A, Liu Mingyu. One-shot free-view neural talking-head synthesis for video conferencing[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 10039−10049

[26] Wang Qiulin, Zhang Lu, Li Bo. SAFA: Structure aware face animation[C]//Proc of the 2021 Int Conf on 3D Vision. Piscataway, NJ: IEEE, 2021: 679−688

[27] Hong Fating, Zhang Longhao, Shen Li, et al. Depth-aware generative adversarial network for talking head video generation[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 3397−3406

[28] Zhao Jian, Zhang Hui. Thin-plate spline motion model for image animation[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 3657−3666

[29] Tao Jiale, Wang Biao, Xu Borun, et al. Structure-aware motion transfer with deformable anchor model[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 3637−3646

[30] Zhu Feida, Zhu Junwei, Chu Wenqing, et al. HifiHead: One-shot high fidelity neural head synthesis with 3D control[C/OL]//Proc of the 31st Int Joint Conf on Artificial Intelligence. 2022: 1750−1756. [2025-07-04]. https://doi.org/10.24963/ijcai.2022/244

[31] Zhang Bowen, Qi Chenyang, Zhang Pan, et al. MetaPortrait: Identity-preserving talking head generation with fast personalized adaptation[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2023: 22096−22105

[32] Wang Jianrong, Zhao Yaxin, Liu Li, et al. Emotional talking head generation based on memory sharing and attention-augmented networks[J]. arXiv preprint, arXiv: 2306.03594, 2023

[33] Guo Jianzhu, Zhang Dingyun, Liu Xiaoqing, et al. LivePortrait: Efficient portrait animation with stitching and retargeting control[J]. arXiv preprint, arXiv: 2407.03168, 2024

[34] Siarohin A, Lathuilière S, Tulyakov S, et al. Animating arbitrary objects via deep motion transfer[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 2377−2386

[35] Doukas M C, Zafeiriou S, Sharmanska V. HeadGAN: One-shot neural head synthesis and editing[C]//Proc of the IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2021: 14398−14407

[36] Wu Wayne, Zhang Yunxuan, Li Cheng, et al. ReenactGAN: Learning to reenact faces via boundary transfer[C]//Proc of the 15th European Conf on Computer Vision. Berlin: Springer, 2018: 603−619

[37] Zhou Hang, Sun Yasheng, Wu Wayne, et al. Pose-controllable talking face generation by implicitly modularized audio-visual representation[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 4176−4186

[38] Wang Yaohui, Yang Di, Bremond F, et al. Latent image animator: Learning to animate images via latent space navigation[C/OL]//Proc of the Int Conf on Learning Representations. 2022[2025-07-04]. https://inria.hal.science/hal-03714584/document

[39] Bounareli S, Tzelepis C, Argyriou V, et al. StyleMask: Disentangling the style space of styleGAN2 for neural face reenactment[C]//Proc of the 17th IEEE Int Conf on Automatic Face and Gesture Recognition. Piscataway, NJ: IEEE, 2023: 1−8

[40] Hong Fating, Xu Dan. Implicit identity representation conditioned memory compensation network for talking head video generation[C]//Proc of the IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2023: 23062−23072

[41] Bounareli S, Tzelepis C, Argyriou V, et al. DiffusionAct: Controllable diffusion autoencoder for one-shot face reenactment[J]. arXiv preprint, arXiv: 2403.17217, 2024

[42] Xie You, Xu Hongyi, Song Guoxian, et al. X-Portrait: Expressive portrait animation with hierarchical motion attention[C]//Proc of the ACM SIGGRAPH 2024 Conf. New York: ACM, 2024: 1−11

[43] Zeng Bohan, Liu Xuhui, Gao Sicheng, et al. Face animation with an attribute-guided diffusion model[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2023: 628−637

[44] Shen Shuai, Zhao Wenliang, Meng Zibin, et al. Difftalk: Crafting diffusion models for generalized audio-driven portraits animation[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2023: 1982−1991

[45] Du Chenpeng, Chen Qi, He Tianyu, et al. DAE-Talker: High fidelity speech-driven talking face generation with diffusion autoencoder[C]//Proc of the 31st ACM Int Conf on Multimedia. New York: ACM, 2023: 4281−4289

[46] Chen Ken, Seneviratne S, Wang Wei, et al. AniFaceDiff: High-fidelity face reenactment via facial parametric conditioned diffusion models[J]. arXiv preprint, arXiv: 2406.13272, 2024

[47] Ma Yue, Liu Hongyue, Wang Hongfa, et al. Follow-Your-Emoji: Fine-controllable and expressive freestyle portrait animation[J]. arXiv preprint, arXiv: 2406.01900, 2024

[48] Xu Sicheng, Chen Guojun, Guo Yuxiao, et al. VASA-1: Lifelike audio-driven talking faces generated in real time[J]. arXiv preprint, arXiv: 2404.10667, 2024

[49] Chen Zhiyaun, Cao Jiaojiong, Chen Zhiquan, et al. EchoMimic: Lifelike audio-driven portrait animations through editable landmark conditions[J]. arXiv preprint, arXiv: 2407.08136, 2024

[50] Tian Linrui, Wang Qi, Zhang Bang, et al. EMO: Emote portrait alive generating expressive portrait videos with audio2video diffusion model under weak conditions[C]//Proc of the European Conf on Computer Vision. Berlin: Springer, 2024: 244−260

[51] Xu Mingwang, Li Hui, Su Qingkun, et al. Hallo: Hierarchical audio-driven visual synthesis for portrait image animation[J]. arXiv preprint, arXiv: 2406.08801, 2024

[52] Jiang Jianwei, Liang Chao, Yang Jiaqi, et al. Loopy: Taming audio-driven portrait avatar with long-term motion dependency[J]. arXiv preprint, arXiv: 2409.02634, 2024

[53] Cui Jiahao, Li Hui, Yao Yao, et al. Hallo2: Long-duration and high-resolution audio-driven portrait image animation[J]. arXiv preprint, arXiv: 2410.07718, 2024

[54] Guo Yudong, Chen Keyu, Liang Sen, et al. Ad-NeRF: Audio driven neural radiance fields for talking head synthesis[C]//Proc of the IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2021: 5784−5794

[55] Liu Xian, Xu Yinghao, Wu Qianyi, et al. Semantic-aware implicit neural audio-driven video portrait generation[J]. arXiv preprint, arXiv: 2201.07786, 2022

[56] Shen Shuai, Li Wanhua, Zhu Zheng, et al. Learning dynamic facial radiance fields for few-shot talking head synthesis[C]//Proc of the 17th European Conf on Computer Vision. Berlin: Springer, 2022: 666−682

[57] Ye Zhenhui, Jiang Ziyue, Ren Yi, et al. GeneFace: Generalized and high-fidelity audio-driven 3D talking face synthesis[J]. arXiv preprint, arXiv: 2301.13430, 2023

[58] Tang Jianxing, Wang Kaisiyuan, Zhou Hang, et al. Real-time neural radiance talking portrait synthesis via audio-spatial decomposition[J]. arXiv preprint, arXiv: 2211.12368, 2022

[59] Li Jianhe, Zhang Jiawei, Bai Xiao, et al. Efficient region-aware neural radiance fields for high-fidelity talking portrait synthesis[C]//Proc of the IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2023: 7568−7578

[60] Yao Shunyu, Zhong Ruizhe, Yan Yichao, et al. Dfa-nerf: Personalized talking head generation via disentangled face attributes neural rendering[J]. arXiv preprint, arXiv: 2201.00791, 2022

[61] Li Weichuang, Zhang Longhao, Wang Dong, et al. One-shot high-fidelity talking-head synthesis with deformable neural radiance field[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2023: 17969−17978

[62] Ye Zhenhui, He Jinzheng, Jiang Ziyue, et al. GeneFace++: Generalized and stable real-time audio driven 3D talking face generation[J]. arXiv preprint, arXiv: 2305.00787, 2023

[63] Ye Zhenhui, Zhong Tianyun, Ren Yi, et al. Real3d-Portrait: One-shot realistic 3D talking portrait synthesis[J]. arXiv preprint, arXiv: 2401.08503, 2024

[64] Cho K, Lee J, Yoon H, et al. GaussianTalker: Real-time high-fidelity talking head synthesis with audio-driven 3D Gaussian splatting[J]. arXiv preprint, arXiv: 2404.16012, 2024

[65] Chen Bo, Hu Shoukang, Chen Qi, et al. GSTalker: Real-time audio-driven talking face generation via deformable Gaussian splatting[J]. arXiv preprint, arXiv: 2404.19040, 2024

[66] Yu Hongyun, Qu Zhan, Yu Qihang, et al. GaussianTalker: Speaker-specific talking head synthesis via 3D Gaussian splatting[C]//Proc of the 32nd ACM Int Conf on Multimedia. New York: ACM, 2024: 3548−3557

[67] Peng Ziqiao, Hu Wentao, Ma Junyuan, et al. SyncTalk++: High-fidelity and efficient synchronized talking heads synthesis using Gaussian splatting[J]. arXiv preprint, arXiv: 2506.14742, 2025

[68] Fan Yingruo, Gao Chenliang, Fu Ying, et al. FaceFormer: Speech-driven 3D facial animation with transformers[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 18749−18758

[69] Richard A, Nguyen P, Corcoran P, et al. MeshTalk: 3D face animation from speech using cross-modality disentanglement[C]//Proc of the IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2021: 1153−1162

[70] Xing Jinbo, Lei Xiao, Zhang Wei, et al. CodeTalker: Speech-driven 3D facial animation with discrete motion prior[C/OL]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2023[2025-07-04]. https://doi.org/10.48550/arXiv.2301.02379

[71] Kim S, Lee C, Son G, et al. MultiTalk: Enhancing 3D talking head generation across languages with multilingual video dataset[J]. arXiv preprint, arXiv: 2406.14272, 2024

[72] Chu Xuangeng, Goswami N, Cui Ziteng, et al. ARTalk: Speech-driven 3D head animation via autoregressive model[J]. arXiv preprint, arXiv: 2502.20323, 2025

[73] Zhen Dingcheng, Yin Shunshun, Qin Shiyang, et al. Teller: Real-time streaming audio-driven portrait animation with autoregressive motion generation[J]. arXiv preprint, arXiv: 2503.18429, 2025

[74] Zhou Mohan, Bai Yalong, Zhang Wei, et al. Responsive listening head generation: A benchmark dataset and baseline[C]//Proc of the 17th European Conf on Computer Vision. Berlin: Springer, 2022: 124−142

[75] Zhou Mohan, Bai Yalong, Zhang Wei, et al. Interactive conversational head generation[J]. arXiv preprint, arXiv: 2307.02090, 2023

[76] Chang Zhigang, Wang Bin, Liu Feng, et al. Hierarchical semantic perceptual listener head video generation: A high-performance pipeline[C]//Proc of the 31st ACM Int Conf on Multimedia. New York: ACM, 2023: 9581−9585

[77] Miao Liu, Jing Wang, Xinyuan Qian, et al. ListenFormer: Responsive listening head generation with non-autoregressive transformers[C]//Proc of the 32nd ACM Int Conf on Multimedia.

New York: ACM, 2024: 7094−7103

[78] Zhu Yongming, Zhang Longhao, Rong Zhengkun, et al. INFP: Audio-driven interactive head generation in dyadic conversations[J]. arXiv preprint, arXiv: 2412.04037, 2024

[79] Wang Kaisiyuan, Wu Qianyi, Song Linsen, et al. MEAD: A large-scale audio-visual dataset for emotional talking-face generation[C]// Proc of the 16th European Conf on Computer Vision. Berlin: Springer, 2020: 700−717

[80] Sinha S, Biswas S, Yadav R, et al. Emotion-controllable generalized talking face generation[J]. arXiv preprint, arXiv: 2205.01155, 2022

[81] Ji Xinya, Zhou Hang, Wang Kaisiyuan, et al. EAMM: One-shot emotional talking face via audio-based emotion-aware motion model[C]//Proc of the 2022 ACM SIGGRAPH Conf. New York: ACM, 2022: 1−10

[82] Drobyshev N, Casademunt A B, Vougioukas K, et al. EMOPortraits: Emotion-enhanced multimodal one-shot head avatars[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2024: 8498−8507

[83] Gan Yuan, Yang Zongxin, Yue Xihang, et al. Efficient emotional adaptation for audio-driven talking-head generation[C]//Proc of the IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2023: 22634−22645

[84] Zhang Chenxu, Wang Chao, Zhang Jianfeng, et al. Dream-Talk: Diffusion-based realistic emotional audio-driven method for single image talking face generation[J]. arXiv preprint, arXiv: 2312.13578, 2023

[85] Lv Tian, Wen Yuhui, Sun Zhiyao, et al. A speaker video continuous emotion editing model based on texture and geometry decoupling[J]. SCIENTIA SINICA Informations, 2023, 53(12): 2423−2439 (in Chinese)
(吕天, 温玉辉, 孙志尧, 等. 基于纹理与几何解耦的说话人视频连续情感编辑模型[J]. 中国科学：信息科学, 2023, 53(12): 2423−2439)

[86] Wu Haozhe, Jia Jia, Wang Haoyu, et al. Imitating arbitrary talking style for realistic audio driven talking face synthesis[C]//Proc of the 29th ACM Int Conf on Multimedia. New York: ACM, 2021: 1478−1486

[87] Liang Borong, Pan Yan, Guo Zhizhi, et al. Expressive talking head generation with granular audio-visual control[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 3387−3396

[88] Ma Yifeng, Wang Suzhen, Hu Zhipeng, et al. StyleTalk: One-shot talking head generation with controllable speaking styles[C]//Proc of the 37th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2023: 1896−1904

[89] Tan Shuai, Ji Bin, Ding Yu, et al. Say anything with any style[C]//Proc of the 38th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2024: 5088−5096

[90] Tan Shuai, Ji Bin, Pan Ye. Style2Talker: High-resolution talking head generation with emotion style and art style[C]//Proc of the 38th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2024: 5079−5087

[91] Wang Suzhzen, Ma Yifeng, Ding Yu, et al. StyleTalk++: A unified framework for controlling the speaking styles of talking heads[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(6): 4331−4347

[92] Ma Liqian, Jia Xu, Sun Qianru, et al. Pose guided person image generation[C]//Proc of the 31st Int Conf on Neural Information Processing Systems. Red Hook, NY: Curran Associates Inc, 2017: 406−416

[93] Lassner C, Pons-Moll G, Gehler P V. A generative model of people in clothing[C]//Proc of the IEEE Int Conf on Computer vision. Piscataway, NJ: IEEE, 2017: 853−862

[94] Ma Liqian, Sun Qianru, Georgoulis S, et al. Disentangled person image generation[C]//Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 99−108

[95] Joo D, Kim D, Kim J. Generating a fusion image: One's identity and another's shape[C]//Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 1635−1643

[96] De Bem R, Ghosh A, Boukhayma A, et al. A conditional deep generative model of people in natural images[C]//Proc of the 2019 IEEE Winter Conf on Applications of Computer Vision. Piscataway, NJ: IEEE, 2019: 1449−1458

[97] Balakrishnan G, Zhao A, Dalca A V, et al. Synthesizing images of humans in unseen poses[C]//Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 8340−8348

[98] Ren Yurui, Fan Xiaoqing, Li Ge, et al. Neural texture extraction and distribution for controllable person image synthesis[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 13535−13544

[99] Bhunia A K, Khan S, Cholakkal H, et al. Person image synthesis via denoising diffusion model[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2023: 5968−5976

[100] Wang Tingchun, Liu Mingyu, Zhu Junyan, et al. Video-to-video synthesis[J]. arXiv preprint, arXiv: 1808.06601, 2018

[101] Wang Tingchun, Liu Mingyu, Tao A, et al. Few-shot video-to-video synthesis[J]. arXiv preprint, arXiv: 1910.12713, 2019

[102] Liu Wen, Piao Zhixin, Tu Zhi, et al. Liquid-Warping-GAN with attention: A unified framework for human image synthesis[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 44(9): 5114−5132

[103] Chan C, Ginosar S, Zhou Tinghui, et al. Everybody dance now[C]//Proc of the IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2019: 5933−5942

[104] Karras J, Holynski A, Wang Tingchun, et al. DreamPose: Fashion image-to-video syn thesis via stable diffusion[C]//Proc of the IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2023: 22623−22633

[105] Hu Li, Gao Xin, Zhang Peng, et al. AnimateAnyone: Consistent and controllable image-to-video synthesis for character animation[J]. arXiv preprint, arXiv: 2311.17117, 2023

[106] Xu Zhongcong, Zhang Jianfeng, Liew Jun Hao, et al. MagicAnimate: Temporally consistent human image animation using diffusion

model[J]. arXiv preprint, arXiv: 2311.16498, 2023

[107] Wang Tan, Li Linjie, Lin K, et al. DisCo: Disentangled control for realistic human dance generation[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2024: 9326−9336

[108] Zhu Shenhao, Chen J L, Dai Zuozhou, et al. Champ: Controllable and consistent human image animation with 3D parametric guidance[J]. arXiv preprint, arXiv: 2403.14781, 2024

[109] Zhang Yuang, Gu Jiaxi, Wang Liwen, et al. MimicMotion: High-quality human motion video generation with confidence-aware pose guidance[J]. arXiv preprint, arXiv: 2406.19680, 2024

[110] Ma Yue, He Yingqing, Cun Xiaodong, et al. Follow Your Pose: Pose-guided text-to-video generation using pose-free videos[C]//Proc of the 38th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2024: 4117−4125

[111] Zhang Yabo, Wei Yuxiang, Jiang Dongsheng, et al. Controlvideo: Training-free controllable text-to-video generation[J]. arXiv preprint, arXiv: 2305.13077, 2023

[112] Qin Bosheng, Ye Wentao, Yu Qifan, et al. DancingAvatar: Pose and text-guided human motion videos synthesis with image diffusion model[J]. arXiv preprint, arXiv: 2308.07749, 2023

[113] Feng Mengyang, Liu Jinlin, Yu Kai, et al. DreaMoving: A human dance video generation framework based on diffusion models[J]. arXiv preprint, arXiv: 2312.05107, 2023

[114] Ginosar S, Bar A, Kohavi G, et al. Learning individual styles of conversational gesture[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 3497−3506

[115] Liao Miao, Zhang Sibo, Wang Peng, et al. Speech2video synthesis with 3D skeleton regularization and expressive body poses[C/OL]// Proc of the 15th Asian Conf on Computer Vision. 2020: 308−323. [2025-07-04]. https://dl.acm.org/doi/abs/10.1007/978-3-030-69541-5_19

[116] Corona E, Zanfir A, Bazavan E G, et al. VLOGGER: Multimodal diffusion for embodied avatar synthesis[J]. arXiv preprint, arXiv: 2403.08764, 2024

[117] Lin Gaojie, Jiang Jianwen, Liang Chao, et al. CYBERHOST: Taming audio-driven avatar diffusion model with region codebook attention[J]. arXiv preprint, arXiv: 2409.01876, 2024

[118] Habibie I, Holden D, Tzionas D, et al. Learning speech-driven 3D conversational gestures from video[C]//Proc of the 21st ACM Int Conf on Intelligent Virtual Agents. New York: ACM, 2021: 101–108

[119] Yi Hongwei, Li Xueting, Yu Chen, et al. Generating holistic 3D human motion from speech[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2023: 469−480

[120] Chen Junming, Wu Yinghao, Ma Zhaofan, et al. DiffSHEG: A diffusion-based approach for real-time speech-driven holistic 3D expression and gesture generation[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2024: 7352−7361

[121] Luo Yuxuan, Rong Zhengkun, Wang Lizhen, et al. DreamActor-M1: Holistic, expressive and robust human image animation with hybrid guidance[J]. arXiv preprint, arXiv: 2504.01724, 2025

[122] Chen Yi, Liang Sen, Zhou Zixiang, et al. HunyuanVideo-Avatar: High-fidelity audio-driven human animation for multiple characters[J]. arXiv preprint, arXiv: 2505.20156, 2025

[123] Karras T, Laine S, Aila T. A style-based generator architecture for generative adversarial networks[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 4401−4410

[124] Stypułkowski M, Vougioukas K, He Sen, et al. Diffused Heads: Diffusion models beat GANs on talking-face generation[C]//Proc of the IEEE/CVF Winter Conf on Applications of Computer Vision. Piscataway, NJ: IEEE, 2024: 5091−5100

[125] Siarohin A, Woodford O J, Ren J, et al. Motion representations for articulated animation[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 13653−13662

[126] Karras T, Laine S, Aittala M, et al. Analyzing and improving the image quality of styleGAN[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 8110−8119

[127] Preechakul K, Chatthee N, Wizadwongsa S, et al. Diffusion autoencoders: Toward a meaningful and decodable representation[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 10619−10629

[128] Zhang Lvmin, Rao Anyi, Agrawala M. Adding conditional control to text-to-image diffusion models[C]//Proc of the IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2023: 3836−3847

[129] Liu Zerun, Yin Yufei, Xue Wenhao, et al. A review of conditional image generation based on diffusion models[J]. Journal of Zhejiang University: Science Edition, 2023, 50(6): 651−667 (in Chinese)
(刘泽润, 尹宇飞, 薛文灏, 等. 基于扩散模型的条件引导图像生成综述[J]. 浙江大学学报: 理学版, 2023, 50(6): 651−667)

[130] Yan Zhihao, Zhou Changbing, Li Xiaocui. Survey on generative diffusion model[J]. Computer Science, 2024, 51(1): 273−283 (in Chinese)
(闫志浩, 周长兵, 李小翠. 生成扩散模型研究综述[J]. 计算机科学, 2024, 51(1): 273−283)

[131] 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

[132] Rombach R, Blattmann A, Lorenz D, et al. High-resolution image synthesis with latent diffusion models[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 10684−10695

[133] Mildenhall B, Srinivasan P P, Tancik M, et al. Nerf: Representing scenes as neural radiance fields for view synthesis[J]. Communications of the ACM, 2021, 65(1): 99−106

[134] Kerbl B, Kopanas G, Leimkühler T, et al. 3D Gaussian splatting for real-time radiance field rendering[J]. ACM Transactions on Graphics, 2023, 42(4): 139: 1−139: 14

[135] VanDenOord A, Dieleman S, Zen Heiga, et al. WaveNet: A generative model for raw audio[C/OL]//Proc of the 33rd Int Conf on

Machine Learning. 2016[2025-07-04]. https://doi.org/10.48550/arXiv.1609.03499

[136] Chan E, Lin C, Chan M, et al. Efficient geometry-aware 3D generative adversarial networks[C]//Proc of the IEEE/CVF Conf on computer vision and pattern recognition. Piscataway, NJ: IEEE, 2022: 16123−16133

[137] Zhu Xiangyu, Lei Zhen, Liu Xiaoming, et al. Face alignment across large poses: A 3D solution[C]//Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 146−155

[138] Xu Yuelang, Chen Benwang, Li Zhe, et al. Gaussian head avatar: Ultra high-fidelity head avatar via dynamic gaussians[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2024: 1931−1941

[139] Li T, Bolkart T, Black M J, et al. Learning a model of facial shape and expression from 4D scans[J]. ACM Transactions on Graphics, 2017, 36(6): 194: 1−194: 17

[140] Drobyshev N, Chelishev J, Khakhulin T, et al. MegaPortraits: One-shot megapixel neural head avatars[C]//Proc of the 30th ACM Int Conf on Multimedia. New York: ACM, 2022: 2663−2671

[141] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision[C]//Proc of the 38th Int Conf on Machine Learning. Red Hook, NY: Curran Associates Inc, 2021: 8748−8763

[142] De Bem R, Ghosh A, Ajanthan T, et al. A semi-supervised deep generative model for human body analysis[C]//Proc of the European Conf on Computer Vision Workshops. Berlin: Springer, 2018: 500–517

[143] Cao Zhe, Simon T, Wei S E, et al. Realtime multi-person 2D pose estimation using part affinity fields[C]//Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2017: 7291−7299

[144] Güler R A, Neverova N, Kokkinos I. DensePose: Dense human pose estimation in the wild[C]//Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 7297−7306

[145] Yang Zhendong, Zeng Ailing, Yuan Chun, et al. Effective whole-body pose estimation with two-stages distillation[C]//Proc of the IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2023: 4210−4220

[146] Loper M, Mahmood N, Romero J, et al. SMPL: A skinned multi-person linear model[J]. ACM Transactions on Graphics, 2015, 34(6): 1−16

[147] Deng Zijian, He Xiangteng, Peng Yuxin. Text-to-video generation: Research status, progress and challenges[J]. Journal of Electronics & Information Technology, 2024, 46(5): 1632−1644 (in Chinese)
(邓梓焌, 何相腾, 彭宇新. 文本到视频生成: 研究现状、进展和挑战[J]. 电子与信息学报, 2024, 46(5): 1632−1644)

[148] Gong Zhixiang, Yu Xiuping. Research on the production process and application scenarios of virtual digital human[J]. TV Technology, 2024, 48(12): 178−181 (in Chinese)
(巩志祥, 于秀平. 虚拟数字人制作流程及应用场景研究[J]. 电视技术, 2024, 48(12): 178−181)

[149] Howard A G, Zhu Menglong, Chen Bo, et al. MobileNets: Efficient convolutional neural networks for mobile vision applications[J]. arXiv preprint, arXiv: 1704.04861, 2017

[150] Tan Mingxing, Le Q. EfficientNet: Rethinking model scaling for convolutional neural networks[C/OL]//Proc of the 36th Int Conf on Machine Learning. 2019: 6105−6114. [2025-07-04]. https://proceedings.mlr.press/v97/tan19a.html

[151] Lu Cheng, Zhou Yuhao, Bao Fan, et al. DPM-Solver: A fast ODE solver for diffusion probabilistic model sampling in around 10 steps[C]//Proc of the 36th Int Conf on Neural Information Processing Systems. Red Hook, NY: Curran Associates Inc, 2022: 5775−5787



Yang Lujia, born in 2001. Master candidate. Her main research interest includes talking head generation.

杨璐嘉, 2001 年生。硕士研究生。主要研究方向为说话人生成。



Zhou Xishi, born in 2003. Maser candidate. His main research interest includes generative artificial intelligence.

周溪石, 2003 年生。硕士研究生。主要研究方向为生成式人工智能。



Wang Junchao, born in 2000. Master candidate. His main research interest includes talking head generation.

王俊超, 2000 年生。硕士研究生。主要研究方向为说话人生成。



Zhang Jinshan, born in 1983. PhD, associate professor. His main research interest includes generative artificial intelligence.

张金山, 1983 年生。博士, 副教授。主要研究方向为生成式人工智能。



Jin Tao, born in 1996. PhD, assistant professor. His main research interest includes multimodal understanding and generation.

金　涛, 1996 年生。博士, 助理研究员。主要研究方向为多模态理解与生成。



Yin Jianwei, born in 1974. PhD, professor. His main research interests include metaverse and digital services, data science, and artificial intelligence.

尹建伟, 1974 年生。博士, 教授。主要研究方向为元宇宙与数字服务、数据科学、人工智能。