中图法分类号： **文献标识码：** **文章编号：** 1006-8961(XXXX)XX-0001-22

论文引用格式： Li Weibin, Gao Jiafeng, Xu Bing, Hou Biao, Jiao Licheng. Progress in Audio-Driven Digital Human Technologies[J/OL]. Journal of Image and Graphics, XXXX:1-22. DOI: 10.11834/jig.260103. (李卫斌, 高佳峰, 徐兵, 侯彪, 焦李成. 音频驱动的数字人技术进展[J/OL]. 中国图象图形学报,XXXX:1-22. DOI: 10.11834/jig.260103.)[DOI:10.11834/jig.260103]

音频驱动的数字人技术进展

李卫斌[1,2],高佳峰[1],徐兵[3],侯彪[2],焦李成[2]

1. 西安电子科技大学杭州研究院,杭州 311231; 2. 西安电子科技大学人工智能学院,西安 710126; 3. 中国航空工业集团民机试飞中心,西安 710026

摘 要： 随着元宇宙与沉浸式人机交互技术的飞速发展,音频驱动的数字人生成(Audio-Driven Talking Head Generation)已成为数字人领域的研究热点。该技术旨在建立从一维语音信号到三维视觉流的跨模态映射,其核心挑战在于在保证唇形精准同步的同时,实现高保真的视觉外观与实时渲染。本文提出了音频驱动数字人通用技术框架,并从技术演进的视角系统梳理了该领域的最新进展。本文回顾了早期的二维(two-dimensional,2D)图像生成方法,分析了其在三维一致性与大姿态驱动上的局限性;进而深入探讨了基于神经辐射场(neural radiance fields,NeRF)的方法,阐述了基于隐式空间建模解决视角一致性的技术方案,并总结该方法面临的推理效率瓶颈;随后,重点综述了当前两大前沿范式:3D高斯溅射(3DGS)范式,其利用显式几何原语突破实时渲染的算力限制;扩散模型(diffusion models)范式,该范式可显著提升生成的细节纹理与动作表现力。此外,本文还归纳了主流的音频驱动数字人数据集与客观评价体系(如唇形同步网络(SyncNet)、Fréchet inception distance(FID)等)。最后,对跨身份泛化、情感与全身交互、以及端侧轻量化部署等开放性挑战进行了深入分析与展望。

关键词： 数字人;音频驱动;说话头生成;神经辐射场;3D高斯溅射;扩散模型

Progress in Audio-Driven Digital Human Technologies

Li Weibin[1,2], Gao Jiafeng[1], Xu Bing[3], Hou Biao[2], Jiao Licheng[2]

1. Hangzhou Institute of Technology, Xidian University, Hangzhou 311231, China; 2. School of Artificial Intelligence, Xidian University, Xi'an 710126, China; 3. AVIC Civil Aircraft Flight Test Center, Xi'an 710026, China

Abstract: With the exponential growth of the Metaverse, immersive virtual reality (VR), and advanced human-computer interaction (HCI) paradigms, audio-driven digital human generation has rapidly emerged as a paramount research frontier in both computer vision (CV) and computer graphics (CG). This transformative technology aims to establish a robust and seamless cross-modal mapping from one-dimensional (1D) speech signals to three-dimensional (3D) visual video streams. Unlike traditional motion-capture-based approaches, which are constrained by expensive hardware and rigid physical setups, audio-driven methodologies offer a highly scalable, cost-effective, and interactive solution. From a cognitive psychology perspective, human speech perception is inherently a multimodal process. The classic McGurk effect demonstrates that visual cues, such as precise lip movements, play a decisive role in deciphering auditory information, especially in noisy environments. Consequently, synthesizing a lifelike digital human requires not only ultra-high-definition visual rendering but also millimeter-level, time-synchronized alignment between auditory signals and visual lip dynamics to prevent the notorious "uncanny valley" effect, where subtle unnaturalness provokes discomfort in human observers. Despite the proliferation of extensive research and rapid algorithmic advancements, generating photorealistic and real-time interactive digital avatars continues to encounter three formidable challenges. First, bridging the immense semantic gap between distinct

收稿日期： 2026-02-24;**修回日期：** 2026-04-23

modalities—specifically, translating semantic and acoustic features into complex spatial facial muscle deformations—remains fundamentally difficult. Second, the trade-off between high-fidelity rendering and real-time inference latency poses a critical bottleneck. Achieving pore-level skin details and realistic hair rendering typically demands immense computational resources, conflicting with the strict low-latency requirements (often under 100 milliseconds) necessary for seamless live streaming or real-time conversational agents. Third, zero-shot cross-identity generalization and the expansion of driving scope—from isolated talking heads to full-body expressive agents encompassing co-speech gestures—demand robust structural priors that current isolated models struggle to provide. To systematically deconstruct this complex domain, this comprehensive review first proposes a unified "encoding-mapping-rendering" technical framework that encapsulates the majority of audio-driven digital human architectures. The framework initiates with an audio feature extractor, utilizing state-of-the-art self-supervised foundational models (such as WavLM and Whisper) to capture rich phonetic, semantic, and emotional representations from raw audio waveforms. These acoustic features are then fed into visual geometric representations, transitioning from sparse two-dimensional (2D) landmarks to sophisticated three-dimensional morphable models (3DMM), which provide an explicit facial motion space. Finally, a cross-modal mapping and rendering network translates these spatial configurations into final visual video frames. Following this structural foundation, this paper meticulously reviews the evolutionary trajectory of audio-driven rendering technologies, categorizing them into four dominant paradigms. First, early 2D image-based generation methods, predominantly driven by generative adversarial networks (GAN), are examined. While these models excel in achieving rapid inference speeds and fundamental lip synchronization, they inherently lack topological constraints, suffering from severe structural degradation and texture stretching when handling large-pose head rotations. Second, the integration of neural radiance fields (NeRF) introduced a paradigm shift by leveraging implicit volumetric representations. NeRF-based approaches successfully resolved multi-view consistency issues; however, the heavy computational burden of volumetric ray marching restricts their real-time application, while their subject-specific optimization nature limits broad generalization. Third, the recent breakthrough of 3D Gaussian splatting (3DGS) has revolutionized the field by combining explicit geometric primitives with highly optimized rasterization pipelines. 3DGS-based models break the real-time rendering constraints, easily surpassing 100 frames per second (FPS), though they still grapple with dynamic consistency during extreme facial deformations due to their discrete point-cloud nature. Lastly, diffusion models represent the current pinnacle of visual fidelity and generative diversity. By iteratively denoising latent spaces, these models synthesize unprecedented skin textures, hair dynamics, and micro-expressions, although their inherent Markov chain sampling mechanisms introduce significant inference latency and temporal jitter challenges in long-video generation. To facilitate an objective assessment of these diverse methodologies, this paper synthesizes mainstream datasets and rigorous evaluation metrics. We detail the utilization of large-scale in-the-wild datasets, such as VoxCeleb, alongside high-resolution benchmarks like HDTF and emotional datasets like MEAD. The evaluation criteria are multi-dimensional, encompassing lip-sync accuracy through SyncNet metrics (lip sync error-distance and lip sync error-confidence), visual quality via Fréchet inception distance (FID) and learned perceptual image patch similarity (LPIPS), temporal video consistency utilizing Fréchet video distance (FVD), and identity preservation measured by cosine similarity (CSIM). Finally, this paper provides an in-depth analysis of open challenges and outlines prospective future research trajectories. We hypothesize that the integration of multimodal large language models (MLLMs) will fundamentally redefine cross-identity generalization, serving as powerful world models to decode audio into structured motion commands. Furthermore, transitioning from purely correlation-based learning to causal modeling will be vital for decoupling authentic emotional expressions from entangled acoustic noise. To address the theoretical bottlenecks of long-sequence temporal consistency, we propose exploring state space models, such as Mamba architectures, and flow matching trajectories. Concurrently, as these generative capabilities approach indistinguishable realism, the ethical implications, particularly concerning deepfake misuse, necessitate the parallel development of robust forensic detection and defense mechanisms. Ultimately, this review aims to serve as a definitive guide for researchers, navigating the transition of digital humans from mere visual replicas to fully embodied, emotionally intelligent interactive agents in the digital era.

Key words: Digital Human; Audio-Driven; Talking Head Generation; Neural Radiance Fields (NeRF); 3D Gaussian Splatting; Diffusion Models

0 引 言

数字人(digital human)，或称虚拟人，是指利用计算机图形学、深度学习等人工智能技术，创建出的具有人体几何形态、外观纹理以及行为模式的数字化虚拟形象。它既可以是对真实世界中特定人物的数字化复刻，也可以是完全虚构的原生创造。作为物理世界中人类在数字空间的映射，数字人正迅速成为连接现实与虚拟世界的关键媒介。从虚拟偶像、智能助手到元宇宙中的虚拟化身，其应用场景日益广阔，为社交娱乐、影视制作、在线教育等领域带来了颠覆性的变革。数字人技术按照对数字人描述的维度，可分为二维数字人、三维数字人及多维数字人。其中三维数字人最为常用。三维数字人技术，利用人工智能模型自动创建和驱动逼真的三维数字角色，其终极目标是实现大规模、低成本、高效率的个性化数字人内容生产，将数字人的创造从少数专家的领域解放出来，赋能于广大用户。

如今数字人已在虚拟主播、远程会议、在线教育及元宇宙社交等场景中展现出巨大的应用潜力(Duan 等，2025)。特别是音频驱动的说话人生成(audio-driven talking head generation)技术，旨在通过一段语音信号驱动任意人脸图像，使其产生与语音内容同步的口型、自然的表情以及协调的头部运动，相关问题也可被视为多模态数字人驱动的重要组成部分(高玄等，2024)。相比于基于动作捕捉的驱动方式，音频驱动具有低成本、易部署和交互自然的优势，是实现智能化、拟人化人机交互的核心技术引擎。

从认知心理学的角度来看，人类的言语感知本质上是一个多模态过程。经典的麦格克效应(McGurk Effect)(McGurk 等，1976)研究表明，视觉线索(如唇形变化)对听觉语音的理解起着决定性的辅助作用，尤其是在噪声环境下，视觉信息的缺失会导致语义理解效率显著下降。因此，生成的数字人视频不仅需要画质清晰，更需要保证视觉唇形与听觉信号在时间上的毫秒级对齐，以维持用户的沉浸感和信任感。

尽管近年来相关研究层出不穷，但要生成以假乱真且能实时交互的数字人，仍面临以下三大核心挑战：

1. 跨模态对齐与恐怖谷效应：音频(一维时序信号)与视频(三维时空信号)存在巨大的模态鸿沟。生成模型若无法精准建立语音音素与面部肌肉运动之间的映射，产生的微小伪影或僵硬表情极易触发恐怖谷效应(uncanny valley)(Mori 等，2012)，即当非人与人的相似度达到特定临界点时，任何微小的不完美都会引发观察者的强烈反感。

2. 高保真与实时性的权衡：这是当前最具技术难度的挑战。一方面，为了追求发丝级、毛孔级的渲染细节，基于NeRF和Diffusion的方法往往计算开销巨大；另一方面，在直播或实时对话场景中，系统端到端延迟通常要求低于100ms(Albert 等，2017)。正如Zhen 等人(2025)在最新的Teller系统中所指出的，现有的流式(streaming)生成架构仍难以在维持高帧率(FPS)的同时，保证长序列动作的自然度与连贯性。

3. 泛化性与驱动范围：传统的模型往往需要针对特定目标人物进行长时间训练，缺乏跨身份的泛化能力。此外，如何从仅驱动面部扩展到包含手势、肢体的全身协同驱动，也是迈向下一代数字人的关键。

随着数字人技术从静态形象生成走向动态交互，研究重点逐渐由外观重建扩展至语音、表情与动作的协同建模。音频驱动数字人正是在这一背景下发展起来，其核心价值在于以低成本方式实现更自然的多模态交互，并推动数字人从“可视化呈现”走向“可交流、可响应”的智能化形态。

回顾音频驱动技术的发展历程，大体可分为三个阶段：早期的2D图像变形与生成阶段(利用GAN实现快速但低维的合成)、中期的神经辐射场(NeRF)重建阶段(引入隐式三维表达以提升多视角一致性)、以及当前的3D高斯溅射(3DGS)与扩散模型(Diffusion)爆发阶段(分别突破了实时渲染速度与生成画质的上限)。

本文将系统梳理上述技术路线的演进逻辑。第一部分提出了音频驱动的通用技术框架；第二部分至第五部分分别对基于2D、NeRF、3DGS及Diffusion的方法进行深入综述；最后总结主流数据集与评价指标，并探讨当前的局限性并展望未来发展方向。

1 音频驱动数字人的通用框架

尽管音频驱动数字人的具体实现技术路线(如2D GAN、NeRF、3DGS等)各不相同,但其核心逻辑均遵循一个通用的"编码-映射-渲染"范式。从任务属性上看,音频驱动说话头生成与面部重演密切相关,其核心都涉及在保持身份信息的同时,根据驱动信号生成相应的嘴型、表情与姿态变化(刘锦等,2022)。如图1所示,该框架通常包含三个核心模块:音频特征提取器(负责理解语音内容与韵律)、视觉几何表征

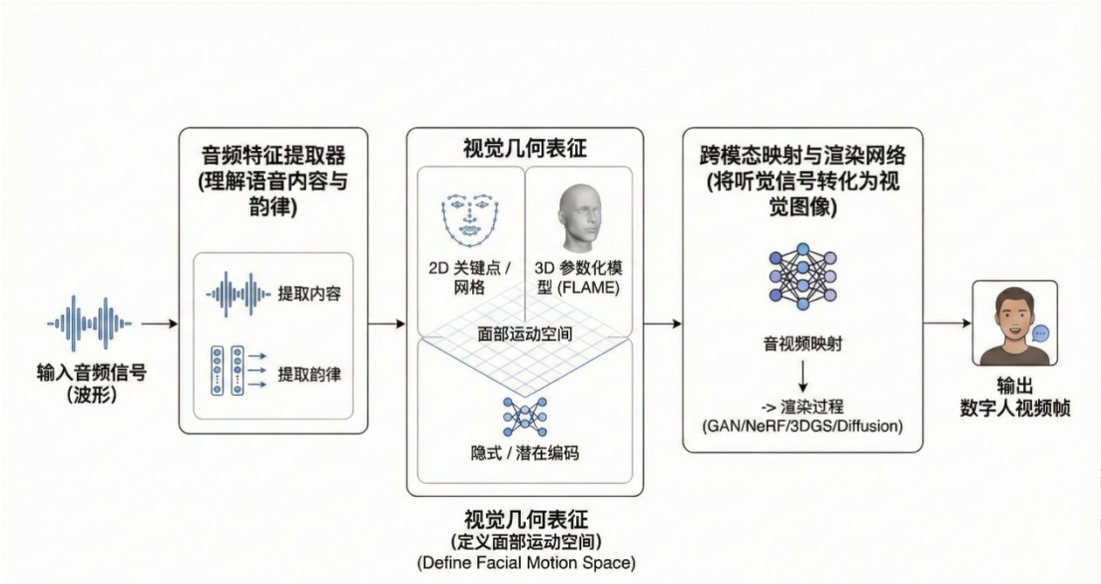

图1 音频驱动数字人的通用框架

Figure 1 A general framework for audio-driven digital humans

(负责定义面部的运动空间)以及跨模态映射与渲染网络(负责将听觉信号转化为视觉图像)。

1.1 音频特征提取与语义编码

驱动数字人的首要任务是从原始波形中提取具有判别性的声学特征。这一过程经历了从手工特征向自监督预训练特征的演进。

手工声学特征(hand-crafted features):早期的2D方法通常使用梅尔频率倒谱系数(mel-frequency cepstral coefficients,MFCC)或线性预测编码(linear predictive coding,LPC)。这些特征计算量小,但主要包含低层次的信号属性,缺乏对语音语义和情感的高层理解,导致生成的口型往往机械且缺乏韵律感。

自监督预训练特征(self-supervised features):为了解决泛化性问题,现代方法(特别是NeRF和Diffusion类)普遍采用在大规模语音数据集上预训练的模型,如Wav2Vec 2.0(Baevski 等,2020)和HuBERT(Hsu 等,2021)。这些模型通过掩码预测任务学习到了丰富的上下文信息,不仅能精准捕捉音素(phoneme)级别的发音特征,还能隐式地编码语调和情感,显著提升了数字人在复杂语境下的驱动效果。

随着基础多模态模型的发展,音频特征提取器正向着多任务、富情感与高泛化方向演进。为了应对复杂野外环境下的噪声干扰,WavLM(Chen 等,2022)通过引入去噪任务与掩码隐层预测,在保留发音音素信息的同时,大幅增强了对说话人身份与背景环境的鲁棒性,目前已被广泛应用于追求高稳定性的数字人模型中。针对多语种与跨语境驱动需求,Whisper(Radford 等,2023)凭借其在大规模弱监督预训练中获得的强大语义理解能力,能够为跨模态网络提供极为精准的音素级与词级对齐特征。此外,在追求高保真情感表达的最新研究中,诸如emotion2vec(Ma 等,2023)等专门针对情感表征优化的音频基础模型逐渐脱颖而出。这类编码器能够从一维语音信号中显式或隐式地解耦出细粒度的情感与韵律信息,为下游生成具备细腻微表情(如悲伤时的皱眉、激动时的张口)的交互式数字人提供了关键的特征支持。

1.2 视觉几何表征与参数化模型

如何用数学方式描述人脸的形状和运动，直接决定了模型还原真实世界的上限。不同的表示方法在精度、稳定性和表现力上各有侧重：从丢失深度的2D关键点，到具备物理一致性的三维可变形模型（three-dimensional morphable model，3DMM），再到高信息承载力的隐式编码，建模维度的每一次提升都抬高了生成的质量天花板。

2D关键点与网格（Landmarks & 2D Mesh）：早期方法利用稀疏的2D面部关键点（通常为68点）作为中间表示。这种表示计算简单，但丢失了深度信息，在大角度头部旋转时容易导致面部结构崩坏。

3D参数化模型（3DMM & FLAME）：为了实现三维一致性，Blanz和Vetter提出的3DMM（3D Morphable Model）（2023）及其衍生模型FLAME（Faces Learned with Articulated Model and Expressions）（Li等，2017）成为了当前的主流选择。FLAME模型将人脸解耦为形状、表情和姿态三个独立参数。这使得音频驱动模型可以专注于预测表情参数和下颌姿态，而保持形状参数（即身份）不变。这一机制是3DGS和NeRF实现可控生成的几何基石。

隐式与潜在编码（implicit & latent codes）：在扩散模型中，视觉表征往往不是显式的几何结构，而是压缩在潜在空间中的特征向量。这种表示虽然缺乏物理可解释性，但具有极强的信息承载能力，能够生成发丝、口腔内部等精细纹理。

除了作为几何先验被用于NeRF或3DGS等渲染框架外，基于3DMM与FLAME等参数化模型的显式驱动路线近年来也取得了显著进展。FaceFormer（Richard等，2022）率先引入Transformer架构，利用自回归方式将语音特征映射为3D网格的顶点动画，有效捕捉了较长时序的语音上下文信息，解决了传统MLP容易导致的平滑过度问题。EMOTE（Daněček等，2023）则进一步在FLAME参数化空间中实现了富有表现力的语音驱动面部动画，强调了内容（speech）与情感（emotion）特征的协同驱动，使得生成的面部动作不再局限于机械的唇形匹配。而在生成式范式下，FaceDiffuser（Stan等，2023）创造性地将扩散模型引入三维面部动画生成，通过在网格顶点特征空间中执行去噪过程，显著提升了语音驱动三维网格动画的非确定性与动态细节。这些基于显式参数化模型的工作，不仅独立成派，更往往作为强大的几何先验（geometry priors），为后续结合NeRF或3DGS的混合渲染方案提供了坚实的骨架支撑。

1.3 跨模态映射与驱动策略

该模块的核心挑战在于解决音频与视频之间的一对多映射难题（即同一句语音可能对应多种合理的表情和头部动作）。

确定性映射与概率性映射：确定性（regression）：早期方法通过简单的多层感知机（multilayer perceptron，MLP）或长短期记忆网络（long short-term memory，LSTM）直接回归面部参数。这种方法容易产生平均脸效应，导致表情平淡；概率性（generative）：为了增强表现力，现代方法引入了变分自编码器（variational autoencoder，VAE）或扩散模型（Diffusion）。通过学习运动分布而非单一数值，模型能够生成眨眼、微表情等具有随机性的自然动作。

头颈肩协同与解耦（head-torso decoupling）：正如第四部分和第五部分将重点讨论的，头部运动（高频、非刚性）与躯干运动（低频、刚性）具有不同的物理特性。因此，分而治之成为了一种通用策略：通常利用两个独立的网络分别预测头部姿态和躯干变形，最后在渲染阶段进行融合，以避免头动身不动的僵硬感或身体跟着脸一起变形的伪影。

1.4 渲染技术演进

渲染模块负责将几何表示转化为最终视频帧，其技术路径的更迭直接定义了数字人技术的代际跨越。各主流方案在渲染质量、推理效率与空间一致性之间进行了不同的取舍：

2D图像生成（生成对抗网络（generative adversarial networks，GANs））：为早期主流，GANs通过纹理合成快速生成图像。虽然其推理速度极快，但由于缺乏显式的三维几何约束，在处理头部大转动时极易出现面部扭曲。

神经体渲染（volumetric rendering）：以NeRF为代表的方法通过沿射线采样进行积分，实现了极佳的三维一致性。它解决了GANs的结构崩坏问题，但其繁重的采样计算导致渲染成本高昂，难以达到实时交互的要求。

光栅化渲染（rasterization）：以3DGS为核心的最新技术，将三维高斯球直接投影到2D平面。它在保留三维结构的同时，利用硬件加速实现了媲美GANs的渲染速度，是当前实时高质量数字人的最优解之一。

去噪生成(denoising generation):基于 Diffusion 的方法通过迭代去噪从噪声中还原图像。虽然推理延迟最高,但其对发丝、皮肤毛孔等精细纹理的还原度最高,代表了当前视觉生成的上限。

1.5 多维分类体系与混合架构

尽管本文后续章节主要沿“视觉渲染表示”(2D、NeRF、3DGS、Diffusion)这一主线对现有技术进行演进梳理,但为了更立体、全面地理解音频驱动数字人领域的技术版图,我们还需引入以下三个关键的分类维度:

1. 按泛化能力与训练范式分类(generalization & training paradigm):

特定人优化模型(person-specific):以早期的 AD-NeRF 及部分绑定式 3DGS 方法为代表。这类方法需要目标人物数分钟的音视频数据进行专属拟合优化。其优势在于能极高保真地还原目标人物的特有微表情与精细纹理,但缺乏灵活性,无法直接驱动未见过的身份。

跨身份通用模型(person-agnostic/zero-shot):以 EMO、Hallo 及 Wav2Lip 为代表。这类模型通常在大规模野外数据集(如 VoxCeleb)上进行通用预训练(universal pre-training),推理时仅需单张或少量参考图像(one-shot/few-shot)即可驱动任意身份。这是当前迈向消费级应用的主流趋势。

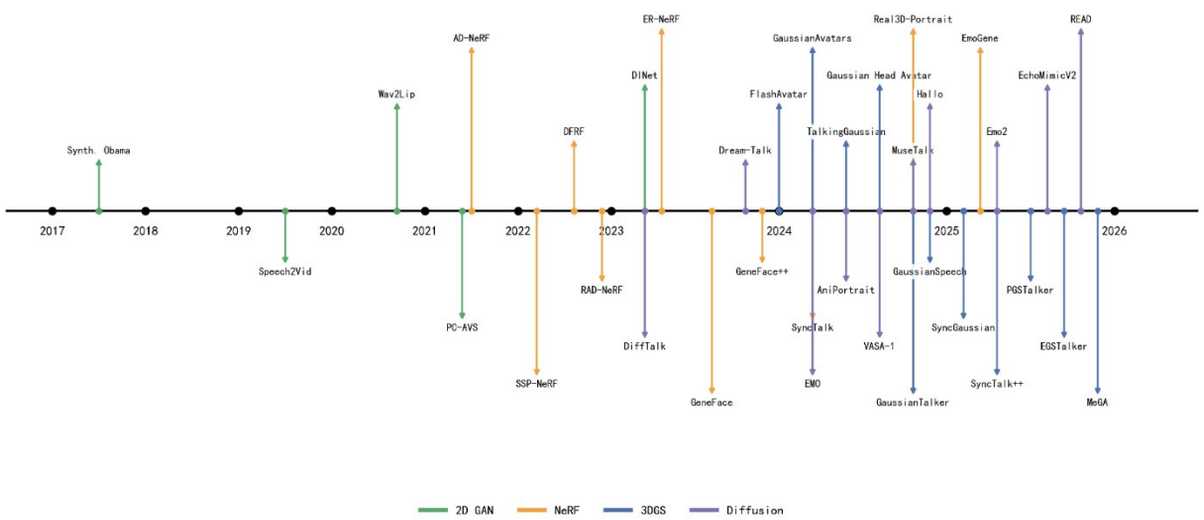

图2 音频驱动数字人技术发展历程概览

Figure 2 Overview of the development history of audio-driven digital human technology

2 驱动粒度分类(driving granularity)

数字人的表现力取决于驱动空间的维度。现有技术可分为:局部唇形同步(仅重绘嘴部区域,如 MuseTalk,主打极致的实时性与低开销)、全脸表情驱动(包含面部肌肉与眨眼)、头颈肩协同驱动(解耦头部高频运动与躯干低频形变,如 PC-AVS)、以及向全身肢体动作生成(co-speech gestures)演进的高阶形态。

3 跨界融合的混合架构(hybrid architectures & boundary cases)

值得注意的是,随着生成式 AI 的爆发,不同技术流派之间的边界正逐渐模糊,单一分类标准已难以涵盖所有前沿工作。研究者开始组合不同范式的优势,催生了大量跨界的混合架构。例如,在运动生成阶段,利用 Diffusion 模型强大的分布拟合能力来生成富有表现力的动作时序潜变量或形变场;而在最终渲染阶段,则将这些动作先验注入到 3DGS 或显式网格(mesh)中进行光栅化,从而巧妙结合了“扩散模型的高泛化运动先验”与“3DGS 的高效实时渲染能力”。此外,如前文述及的 MeGA 模型,采用网

格与高斯溅射的混合表示(hybrid mesh-gaussian)，也证明了打破单一渲染表示边界是突破高保真与实时性瓶颈的重要途径。

2 基于2D图像的音频驱动数字人技术

基于2D图像的生成方法是音频驱动数字人技术的先驱。其核心思想是直接在像素空间或特征空间对源人脸图像进行变形(warping)或重绘(inpainting)，以匹配输入的音频信号。早期的探索主要关注唇形的同步性，随后逐渐发展为对头部姿态、眨眼等非语言行为的综合建模。

2.1 基于图像拼接与纹理合成的方法

早期方法主要依赖图像拼接与纹理检索，如MikeTalk（Ezzat 等，2002）和 Synthesizing Obama（Suwajanakorn 等，2017），虽能实现初步视听同步，但高度依赖特定人物数据库与离散动作合成，缺乏泛化能力，随后被深度生成模型迅速取代。

2.2 端到端对抗生成网络

随着生成对抗网络的发展，研究者开始采用端到端框架直接将音频特征映射为面部图像，从而摆脱对显式纹理检索与拼接的依赖。代表性工作如Temporal GAN（Vougioukas 等，2020）和 Speech2Vid（Jamaludin 等，2019）证明了模型能够从音频中学习唇形运动及部分面部动态信息，而Wav2Lip（Prajwal 等，2020）则进一步通过引入唇形同步判别器，显著提升了野外场景下的口型对齐精度，成为该路线中最具代表性的基准方法之一。

2.3 基于中间表示与注意力机制的方法

相比端到端像素生成，中间表示降低了跨模态映射难度，使模型更容易学习音素与嘴部运动之间的对应关系，同时也更容易引入韵律、情感等高层控制信号。为了提高生成的准确性和可控性，研究者引入了地标或注意力图作为中间媒介，将复杂的生成任务分解。Chen 等人（2019）提出了一种层级跨模态生成框架，利用动态像素级损失和注意力机制引导网络关注嘴部区域的精细生成，解决了全脸生成中细节丢失的问题。MakeItTalk（Zhou 等，2020）利用长短期记忆网络（LSTM）从音频预测面部地标的位移，再根据地标生成图像。Chen 等人（2020）进一步挖掘音频中的韵律信息，不仅预测嘴部动作，还利用ATVG-Net根据音频韵律生成自然的头部晃动，增强了数字人的表现力。这类方法的价值在于“可控性”，但由于最终仍在2D平面生成，面对大姿态时仍难以保持结构一致性。

2.4 姿态解耦与高保真修复

2D方法的后期研究主要集中在解决大姿态头部运动和高分辨率画质这两个瓶颈上。X2Face（Wiles 等，2018）利用姿态码驱动嵌入网络，实现了在给定音频和姿态下的面部重演。PC-AVS（pose-controllable audio-visual system）（Zhou 等，2021）通过隐式模块化网络，彻底解耦了语音内容、头部姿态和身份信息，突破了传统2D方法头不动的限制。Audio2Head（Wang 等，2021）进一步实现了One-shot场景下的自然头部运动生成，通过运动场预测网络让静态图片动起来。针对生成图像模糊的问题，DINet（Zhang 等，2023）利用形变修复网络（deformation inpainting network），结合参考图像的特征对嘴部区域进行高频细节修复，成功实现了高分辨率视频的视觉配音。

但这些方法仍然无法解决根本问题，主要在于大姿态控制和高保真修复的困难，具体而言，2D缺少深度与遮挡建模，旋转导致不可见区域需要补全，此外，嘴部区域属于高频纹理+强形变区域，GAN/ U形网络（U-Net，UNet）修复容易产生闪烁。

2.5 本节方法对比分析

总体来看，基于2D图像的音频驱动方法在发展过程中逐步完成了从“纹理拼接”到“端到端生成”，再到“中间表示建模与局部高保真修复”的演进，其核心优势在于实现成本低、推理速度快、部署门槛低。以 Wav2Lip 为代表的方法推动了唇形同步性能的显著提升，以 DINet 为代表的方法则提升了高分辨率嘴部细节的表现上限。然而，这一路线的根本瓶颈仍在于缺乏显式三维几何与遮挡约束，因此在大姿态、多视角和长序列场景下容易出现结构不稳定与时序伪影。

进一步而言，2D方法的核心瓶颈并非网络结构复杂度不足，而是缺乏显式的三维几何与遮挡建模约束：在大姿态驱动过程中，面部自遮挡区域需要依赖隐式补全，易产生纹理拉伸、结构漂移与局部崩坏；在长序列生成中，像素级生成误差会逐帧累积，导致时间一致性下降并出现抖动与闪烁伪影。相比之下，后续基于NeRF的隐式三维辐射场方法与基

于3D高斯溅射的显式几何渲染方法通过引入三维一致性约束,从根本上缓解了大姿态与多视角生成的结构稳定性问题。因此,从技术路线取舍的角度来看,2D方法更适合作为音频驱动数字人系统中的轻量化模块(如快速唇形同步、局部口腔区域修复或低延迟预览),而难以作为追求高保真三维一致性与大幅度姿态驱动的最终解决方案。总体而言,2D路线在工程部署与实时交互场景中仍将长期存在,但其未来的发展趋势更可能是与3DGS或扩散模型范式形成互补:即由2D模块承担高效同步与局部细节增强,由三维或生成式模型负责全局结构一致性与高保真外观表达。

3 基于神经辐射场的音频驱动数字人技术

基于2D图像的生成方法虽然在画质上取得了长足进步,但由于缺乏三维几何的显式约束,生成的视频在大角度头部运动下往往会出现纹理扭曲和透视错误。Mildenhall等人(Mildenhall 等,2020)提出的NeRF通过MLP隐式地表示场景的体积密度和颜色,实现了照片级逼真的新视图合成,为解决上述问题提供了新的范式。本节将探讨将静态的NeRF扩展为音频驱动的动态NeRF,并分别从实时渲染、泛化能力及多模态交互三个维度综述相关进展。

3.1 动态辐射场的构建与早期探索

将NeRF应用于说话人合成的核心挑战在于建立音频信号与辐射场变化之间的映射关系。Guo等人(Guo 等,2021)提出的AD-NeRF是该领域的开山之作。该方法将音频特征与头部姿态参数作为条件输入连接到两个独立的NeRF网络中,分别负责渲染头部和躯干。AD-NeRF首次实现了音频驱动的高保真三维头部合成,解决了2D方法中的头部漂移问题。尽管AD-NeRF效果惊艳,但其面临两个严重瓶颈:一是推理速度极慢,基于体渲染(volume rendering)的机制导致每帧渲染需要查询数百万个点,难以实时运行;二是头颈连接处的伪影,由于头部和躯干分开建模,两者的运动不协调常导致颈部区域出现撕裂或模糊。该阶段的NeRF方法以质量优先,牺牲实时性,由于头身分离是工程上必然策略,因此也是伪影源头。

3.2 空间分解与实时渲染优化

为了解决NeRF推理效率低下的问题,后续研究引入了空间分解和哈希编码(hash encoding)技术,通过用空间换时间的策略显著提升了渲染速度。RAD-NeRF(Tang 等,2022)引入了基于网格的音频编码策略。该模型将三维空间分解为多个低维平面,并利用音频特征调制这些平面的属性,避免了深层MLP的繁重计算,成功实现了实时推理。ER-NeRF(Li 等,2023)则进一步优化了空间建模策略。该研究指出,面部不同区域对音频的响应程度不同(如嘴部活跃,额头静止),因此提出了区域感知(region-aware)的条件机制,并结合三平面哈希表示(tri-plane hash representation),在保证高保真画质的同时,实现了高效的训练与推理。针对头身分离导致的伪影,SSP-NeRF(Liu 等,2022)引入了语义感知机制,利用语义解析图引导辐射场的形变,使得复杂背景和身体部分的渲染更加清晰稳定。

3.3 泛化性与少样本生成

传统的NeRF方法通常需要针对特定目标人物的一段长视频(通常数分钟)进行训练,缺乏跨身份的泛化能力。如何实现少样本(few-shot)甚至单样本(one-shot)驱动是近两年的研究热点。

DFRF(Shen 等,2022)引入了标准空间与变形场的概念。模型学习一个通用的面部先验,通过预测音频驱动的非线性变形来实现少样本条件下的快速适应。GeneFace(Ye 等,2023)旨在解决训练数据(合成数据)与测试数据(真实视频)之间的域差异,该方法设计了一个变分运动生成器,在大规模音频-唇形数据集上预训练,再迁移到目标NeRF中,显著提高了在未见过的音频上的表现。随后,GeneFace++(Ye 等,2023)中引入了音素级基准以消除音频中的非面部信息干扰,进一步提升了生成的稳定性。而在One-shot场景下,Real3D-Portrait(Ye 等,2024)利用预训练的3D先验,实现了仅凭一张照片即可生成多视角的说话人视频,极大地降低了NeRF方法的应用门槛。

3.4 高保真同步与情感控制

当NeRF路线通过空间分解与哈希编码缓解实时性瓶颈后,研究重点开始从“能否生成”转向“是否自然可信”,其中最关键的两个方向是唇形同步稳定性与情感表达。

由于NeRF具备连续可微的三维表征,情感控

制可以通过调制表情空间或局部辐射场实现更平滑的过渡，相比2D像素生成更不易产生突变伪影。Peng等人(Peng 等，2024)指出现有方法在长序列生成中容易出现身份漂移和唇形不同步的问题。他们提出的SyncTalk引入了一个面部同步控制器，显式地约束面部地标与音频的对齐关系，特别是优化了高频唇部运动的准确性。

此外，最新的研究开始关注非语言信息。EmoGene(Wang 等，2025)将情感强度作为额外的控制变量引入NeRF，使得生成的数字人不仅能对口型，还能根据音频的情感色彩表现出愤怒、快乐等微表情，推动了数字人从读稿机器向情感代理的进化。

3.5 本节方法对比分析

为进一步明确不同NeRF方法在实时性、泛化能力与自然度控制等维度上的取舍关系，表1对本节涉及的代表性工作进行了归纳对比。可以看出，NeRF路线的研究主线可概括为三类：第一类以AD-NeRF为代表，强调高保真三维一致性但牺牲推理效率；第二类以RAD-NeRF、ER-NeRF为代表，通过空间分解与显式表示加速实现准实时渲染，是当前可交互NeRF数字人的主流方向；第三类以GeneFace++、Real3D-Portrait等为代表，聚焦少样本与泛化能力，旨在降低身份建模成本并提升跨身份适配能力。与此同时，SyncTalk与EmoGene表明，当渲染效率瓶颈逐步缓解后，研究重点开始转向同步稳定性与情感表达等更高层语义控制，从而提升生成结果的可信度与表现力。NeRF路线在泛化能力上的理论瓶颈在于其优化目标。由于MLP网络倾向于将特定身份的高频面部细节直接隐式编码至网络权重中，导致身份特征(Identity)与运动特征(Motion)在三维流形空间中高度纠缠。当引入域外(Out-of-domain)的新音频或新身份驱动时，运动形变场难以实现完美的解耦控制，极易引发辐射场密度分布的混沌，造成面部模糊或伪影。

从技术路线取舍角度来看，NeRF方法尽管在三维一致性方面具有优势，但其核心瓶颈仍集中在推理延迟与部署成本：即便在哈希编码与空间分解策略加持下，NeRF仍需执行射线积分过程，在端侧或低算力场景下难以达到稳定高帧率；同时，NeRF通常需要针对特定身份进行辐射场拟合，其训练过程本质上属于场景级优化，因此跨身份泛化能力仍然有限。上述限制也为后续基于显式几何原语与光栅化管线的3D高斯溅射(3DGS)方法(Kerbl 等，2023)的兴起奠定了基础。总体而言，NeRF路线更适合用于追求多视角一致性与影视级画质的离线或准实时生成任务，并在未来可能作为3DGS与扩散模型体系中的三维先验、身份建模或高保真细节补全模块持续发挥作用。

4 基于3D高斯溅射的音频驱动数字人技术

虽然基于NeRF的方法在视图合成质量上取得了突破，但其依赖于昂贵的体积射线采样(volumetric ray marching)和庞大的多层感知机(MLP)查询，导致推理速度难以满足实时交互需求。针对这一瓶颈，Kerbl等人(2023)于2023年提出了3D高斯溅射(3D gaussian splatting，3DGS)。该技术采用各向异性的3D高斯球作为显式几何原语，结合基于图块的光栅化管线，实现了100FPS以上的实时渲染速度和照片级画质。与NeRF的黑盒隐式表示不同，3DGS的显式特性使其更易于进行几何变形和控制。因此，如何将音频信号映射到高斯原语的属性(位置、旋转、缩放、不透明度)变化上，成为了继NeRF之后数字人领域最新的研究热点。

4.1 动态高斯骨架与几何驱动基础

在实现音频驱动之前，研究者首先探索了如何基于视觉信号驱动3DGS，为说话人生成提供了关键的几何先验和骨架支撑。

为了解决动态人脸的建模难题，主流方法通常将3D高斯与FLAME或3DMM等参数化网格绑定。FlashAvatar(Xiang 等，2024)通过将高斯嵌入到网格表面(gaussian embedding)，利用网格的形变带动高斯的运动；GaussianAvatars(Qian 等，2024)则引入了骨骼绑定高斯(rigged gaussians)概念，实现了对面部姿态的显式控制。

针对单目视频重建的难点，MonoGaussianAvatar(Chen 等，2024)和PSAvatar(Zhao 等，2024)分别通过基于点的高斯溅射和点基形变模型，提高了在稀疏视角下的几何准确性。Deformable 3DGS(Yang 等，2024)和Rig3DGS(Rivero 等，2025)则进一步优化了非刚性形变的物理合理性，使得数字人能够处理复杂的面部表情。

此外，HUGS(Kocabas 等，2024)将高斯溅射扩

表1　基于神经辐射场路线典型方法对比

Table 1　Comparison of representative methods in the NeRF paradigm

方法	核心建模策略	实时性/效率	泛化能力	同步性/自然度控制	主要局限
AD-NeRF	头躯干双NeRF条件驱动	低	低	同步性较强，整体自然度较高	渲染慢；头颈伪影；训练成本高
RAD-NeRF	音频-空间分解驱动	中-高	低-中	同步性稳定，细节较AD-NeRF略弱	画质上限受限；复杂表情不足
ER-NeRF	区域感知+Tri-plane Hash	中-高	中	同步性与细节表现更优，稳定性更好	训练复杂；依赖高质量对齐
DFRF	标准空间+动态形变场	中	中	自然度较好，跨姿态更稳定	泛化受限；复杂运动易漂移
GeneFace	预训练运动先验+NeRF渲染	中	中-高	表情与头动更丰富，整体更自然	域差异敏感；高保真细节不足
GeneFace++	音素约束+稳定运动先验	中-高	高	同步稳定性更强，长序列更平滑	极端情感与夸张表情的表达有限
Real3D-Portrait	预训练3D先验的One-shot驱动	中	高	自然度较好，跨视角一致性较强	单图细节受限；身份特征易平滑
SyncTalk	显式同步控制器增强对齐	中	中	同步性提升，长序列漂移减少	依赖对齐质量；情感与头动控制有限
EmoGene	情感条件控制NeRF表达	中	中	情感表达更强，神态更丰富	情感一致性不足；可能夸张或不稳定

展至全身人体建模，虽然主要关注身体运动，但其处理关节大变形的策略为未来全身音频驱动提供了借鉴。

4.2　音频驱动的端到端生成网络

在上述几何骨架的基础上，近两年的研究重点转向了构建端到端的音频-高斯映射网络，核心挑战在于如何在保持三维结构稳定的同时，实现精准的唇形同步和高效的动态变形。

早期的直接驱动尝试往往会导致面部结构在运动中崩坏。TalkingGaussian（Li 等，2024）指出了3DGS在动态序列中容易出现结构不一致的问题，提出了一种结构持久性（structure-persistent）策略，将面部分解为静态和动态两个分支，显著提升了说话时的面部稳定性。GaussianTalker（Cho 等，2024）构建了一个实时推理框架，通过音频特征直接预测高斯属性的残差，在保证高保真画质的同时实现了极低的延迟。Gaussian Head Avatar（Xu 等，2024）利用动态高斯网络处理高频细节，进一步提升了发丝和皮肤纹理的真实感。

针对生成视频中常见的唇形抖动和不同步现象，研究者引入了更精细的监督机制。SyncGaussian（Liu 等，2025）设计了一种基于判别性语音特征的同步模块，通过增强音频特征与嘴部运动的互对齐，解决了复杂语境下的口型匹配问题。SyncTalk++（Peng 等，2025）则在同步性的基础上优化了效率，通过轻量级的高斯变形场，在维持高FPS的同时实现了SOTA级别的唇形同步精度。SynGauss（Zhou 等，2025）明确了面部表情与嘴部运动的解耦策略，减少了语音内容对上半脸的干扰，使得眨眼和头部晃动更加自然。

为了进一步压缩计算成本并提升边缘质量：EGSTalker（Zhu 等，2025）引入了高效高斯变形模块（efficient gaussian deformation），大幅减少了冗余高斯的计算量。针对3DGS在稀疏区域容易产生雾状伪影的问题，Zhu 等人的PGSTalker（Zhu 等，2025）提出了像素感知密度控制（pixel-aware density control），有效锐化了牙齿和舌头等精细区域的边缘轮廓。

4.3　泛化性与混合架构的演进

传统的3DGS方法通常需要针对特定人物训练，最新的趋势是向跨身份泛化和混合渲染架构发展。

GaussianSpeech（Aneja 等，2024）和 GGTalker（Hu 等，2025）利用预训练的通用高斯先验（generalizable gaussian priors），实现了仅需少量数据甚至单张图片即可驱动任意身份的数字人，极大地降低了应用门槛。

为了兼顾显式网格的几何稳定性与高斯溅射的极致渲染真实感，MeGA（Wang 等，2025）提出了一种前沿的混合架构。该模型将面部主体核心区域绑定于稳定的拓扑网格上以防止大表情下的结构崩坏，同时在毛发、口腔内部等非刚性且难以网格化的区域采用3D高斯进行细节增强 。这种“稳定骨架+高频细节”的混合建模范式，从底层规避了纯3DGS架构易产生高斯漂移的缺陷，被视为通向下一代超写实数字人的最优路径之一。

4.4 本节方法对比分析

基于3D高斯溅射（3DGS）的音频驱动数字人方法通过显式几何原语与光栅化渲染管线实现了高效的三维表达，相比NeRF依赖射线积分的体渲染方式，3DGS在推理阶段能够显著降低渲染开销，从而在保持多视角一致性与高保真外观的同时更容易实现实时交互。与NeRF路线以隐式场景函数拟合为核心的建模范式不同，3DGS更强调显式结构表达与可微渲染效率，其技术路线的关键突破主要体现在两个方面：一方面，通过将人物外观与几何以高斯集合的形式显式存储并进行快速光栅化，显著提升了渲染速度与系统可部署性；另一方面，通过引入动态高斯更新、形变场或显式绑定机制，使得三维数字人能够在保持结构一致性的前提下实现可控的表情与头动变化，为音频驱动任务提供了更稳定的三维载体。

尽管3DGS突破了渲染速度，但其在复杂表情（如大幅度张口）下的动态一致性问题，根源在于其点云基元的离散显式表达特性。缺乏类似网格（mesh）的底层连通性拓扑约束，使得高斯球在受音频特征驱动进行位置与协方差更新时，容易在强形变区域（如唇齿交界处）发生过度分裂或不规则聚合，从而在视觉上产生令人不适的撕裂感与雾状伪影。

为进一步明确3DGS路线在动态建模、驱动稳定性与实时性等维度的取舍关系，表2对本章涉及的代表性工作进行了归纳对比。可以看出，当前3DGS音频驱动方法大体可归纳为三类：第一类以静态3DGS重建为基础，通过额外的驱动网络或局部形变实现表情变化，具有实现简单、渲染效率高的特点，但动态一致性与长序列稳定性相对有限；第二类通过显式绑定（如网格/骨架/形变场）或可控形变参数实现动态3DGS建模，在大姿态与长序列场景下具有更好的结构稳定性，是当前高质量实时数字人的主流方向；第三类进一步引入高层语义控制信号（如韵律、情感或风格参数），在保证同步性的同时提升了神态表达能力，使研究从“结构稳定”逐步迈向“表现自然”。总体而言，3DGS路线在实时交互与多视角一致性方面具有明显优势，尤其适用于虚拟直播、数字人对话与实时驱动等场景。

然而，从技术路线取舍角度来看，3DGS方法尽管具备实时渲染优势，但其核心挑战仍集中在动态一致性与可编辑性：由于高斯原语属于显式点基元表示，动态更新过程中容易出现高斯漂移、密度分布不稳定或局部撕裂等问题；同时，在缺乏显式拓扑约束的情况下，复杂表情（如大幅度张口、牙齿与口腔内部显露）仍难以稳定表达。此外，与NeRF相比，3DGS的外观表达更依赖高斯分布的覆盖与优化质量，面对极端光照变化或稀疏视角输入时可能出现细节缺失。总体而言，3DGS路线更适合作为面向实时交互的三维数字人核心表示，而NeRF路线在离线高保真重建与连续表征方面仍具有优势。未来更可能的趋势是两者形成互补：以3DGS承担实时渲染与交互，以NeRF或扩散模型提供高质量先验与细节补全，从而构建兼具效率与真实感的音频驱动数字人系统。

5 基于扩散模型的数字人技术

在生成对抗网络（GAN）和NeRF主导数字人生成的时代，模型往往难以兼顾生成的多样性与高保真度。GAN容易陷入模式崩溃（mode collapse），而NeRF在处理非刚性形变（如发丝、衣物）时常出现模糊。去噪扩散概率模型（DDPM）（Ho 等，2020）通过模拟热力学扩散过程，先在前向过程中逐步添加高斯噪声破坏数据，再通过反向过程学习去噪以恢复数据分布。这种基于似然的生成范式在图像合成质量上超越了GAN，为数字人视频生成提供了新的技术路径。然而，原始的扩散模型推理速度较慢。流匹配（flow matching）（Lipman 等，2024）理论为加速

表2　3DGS模型路线典型方法对比

Table 2　Comparison of representative methods in the 3DGS paradigm

方法	核心建模策略	动态建模方式	实时性/效率	驱动稳定性	主要局限
3DGS静态重建基线	显式高斯集合表示+光栅化渲染	无	高	无	不能表达动态；仅用于重建与渲染
静态3DGS+驱动网络	静态3DGS外观为主，外接驱动器	预测局部形变/参数	高	中	动态一致性有限；长序列易漂移
动态3DGS（隐式更新）	直接学习随时间变化的高斯参数	高斯中心/尺度/颜色动态更新	中-高	中	易产生高斯漂移；训练不稳定
动态3DGS（显式形变场）	连续形变场约束高斯运动	形变场驱动高斯位置更新	中	高	训练复杂；对数据覆盖要求高
绑定式3DGS	以网格/骨架作为拓扑约束	绑定权重+蒙皮变形	中-高	高	拓扑约束强；细节需要额外补偿
语义/区域感知3DGS	对嘴部/眼部等区域强化建模	区域分支+局部高斯密度调制	中	高	需要分割/关键点；实现复杂
情感/风格可控3DGS	引入韵律/情感等高层控制	控制向量调制形变或外观	中	中-高	语义一致性难；可能引入不稳定
3DGS+细节补全/修复	3DGS提供结构，外接细节增强模块	2D/3D细节补全	中	高	系统复杂；训练/推理链路更长

生成过程提供了新的数学框架，通过构建最优传输路径，显著减少了采样步数，成为MuseTalk等后续实时数字人模型的重要理论基础。

5.1　音频驱动的视频生成范式

早期的扩散模型尝试如DiffTalk（Shen等，2023），主要关注于将潜在扩散模型（LDM）应用于人脸重演，但仍局限于头部姿态的简单控制。随着技术的演进，研究重心转向了Audio2Video的端到端生成，旨在赋予数字人更强的表现力。

Tian等人提出的EMO（Tian等，2024）是该领域的里程碑式工作。它摒弃了传统的3D中间表示（如3DMM或关键点），直接学习音频到视频帧的映射。EMO引入了帧间注意力机制和速度控制器，使得数字人能够根据音频节奏产生大幅度的头部晃动、歌唱甚至说唱动作，极大地突破了传统方法呆板的印象。针对长视频生成中的一致性问题，针对长视频生成中的一致性难题，Hallo系列进行了持续迭代。早期的Hallo引入了分层音频驱动视觉合成模块以强化局部同步，随后Hallo2通过时间对齐增强策略将生成分辨率提升至4K并支持长时长输出。而最新提出的Hallo3（Cui等，2025）则实现了架构跃迁，彻底引入了视频扩散Transformer（video diffusion transformer，VDT）。该架构通过时空注意力机制的全局感受野，显著提升了数字人在剧烈运动（如大幅度转头、舞蹈）下的动态保真度与拓扑连贯性，有效抑制了传统U-Net架构在长序列中易出现的局部闪烁与形变伪影。

AniPortrait（Wei等，2024）则采用基于ReferenceNet的架构，利用CLIP图像编码器提取参考图像的纹理特征，结合Motion Adapter捕捉动作，在保持身份一致性的同时实现了逼真的面部动画。

5.2　推理效率的极致优化

尽管生成质量极高，扩散模型的迭代去噪机制导致推理延迟大，难以应用于实时交互。2024-2025年的研究重点在于打破这一效率瓶颈。

MuseTalk（Zhang等，2024）通过修改潜在空间修复机制，实现了30fps以上的实时推理。该方法仅对嘴部区域的Latent Feature进行重绘，而非全图生成，在保证高同步率（SyncNet得分SOTA）的同时大幅降低了计算开销。VASA-1（Xu等，2024）通过在潜在空间解耦外观、3D姿态和身份特征，实现了低延迟的交互式生成。Wang等人提出的READ（2025）则进一步创新，设计了异步扩散架构，允许音频特征提取与视频生成过程异步进行，显著提升了

端到端的吞吐量。

5.3 情感、半身与交互

最新的研究趋势不再局限于头部生成，而是向半身动作和情感交互扩展。

Dream-Talk（Zhang 等，2023）针对扩散模型表情僵硬的问题，引入了情感风格控制器，使得生成的说话头能够根据语音语调表现出开心、悲伤等细腻情绪。

EchoMimicV2（Meng 等，2025）率先打破了“仅有头部运动”的局限，将驱动范围无缝扩展至半身（semi-body）交互 。该方法创新性地引入了音频-姿态联合引导机制（audio-pose guided mechanism），不仅能根据语音韵律自动生成自然的手部协同手势（co-speech gestures），还能确保手势与面部表情在时空流形上的语义一致性 ，为构建虚拟主播等高度沉浸式的全身交互场景奠定了生成学基础。Emo2（Tian 等，2025）则引入了末端执行器引导机制，允许用户精确控制数字人的手部轨迹和头部朝向，增强了数字人在虚拟主播等场景下的交互能力。

5.4 本节方法对比分析

基于扩散模型的音频驱动数字人方法通过强大的生成先验显著提升了人脸纹理细节、局部真实感与生成多样性，为解决GAN类方法易出现的模式崩溃与局部伪影问题提供了新的技术路径。与NeRF或3DGS等三维表示路线强调几何一致性不同，扩散模型更侧重于学习高维数据分布并在采样过程中逐步恢复高频细节，因此在复杂表情、发丝、皮肤纹理与风格一致性方面表现突出。

然而，尽管扩散模型在生成质量上取得显著优势，其迭代去噪机制导致推理延迟大，成为制约实时交互应用的核心瓶颈。围绕推理效率优化，有通过潜在空间修复机制仅对嘴部区域进行重绘，实现30fps以上的实时推理，并在保持较高同步率的同时显著降低计算开销；有通过在潜在空间解耦外观、三维姿态与身份特征实现低延迟交互式生成。可以看出，扩散模型路线的关键突破正在从更高保真转向更高效率，并逐步形成以局部重绘、潜空间解耦与异步推理为代表的加速范式。

扩散模型在生成长视频时易出现时序抖动，其核心失效机制归因于独立帧间噪声调度的随机性。现有的帧间注意力机制（temporal attention）虽然能在一定程度上平滑相邻帧，但在深层隐空间中，音频条件引导（audio guidance）对去噪轨迹的微小扰动会被放大，导致生成的面部微表情在时序上缺乏严格的物理连续性，这也是扩散模型从图像跨向长视频生成时亟待解决的数学底层难题。

在高层语义控制与交互方面，扩散模型正在从说话头生成扩展到半身动作与情感交互，推动数字人生成从低层唇形同步向更高层的语义表现与交互控制演进。

综合来看，2D、NeRF、3DGS 与 Diffusion 四条路线本质上是在“同步精度—视觉保真—三维一致性—实时性—泛化能力”之间进行不同取舍。扩散模型在音频驱动任务中的主要瓶颈仍集中在推理效率与时序一致性：即便采用局部重绘、潜空间建模与异步扩散等策略，其生成过程仍依赖多步采样，难以在端侧或低算力条件下实现稳定实时；同时，扩散模型在缺乏显式三维几何约束的情况下，跨视角一致性与遮挡建模能力仍相对不足，在大姿态变化或长序列生成中仍可能出现局部漂移与闪烁伪影。总体而言，扩散模型路线更适合用于追求高保真视觉质量与风格表现力的离线或准实时生成任务，并在未来更可能作为音频驱动数字人系统中的高质量生成先验，与3DGS等实时三维表示形成互补：即由3D表示负责全局结构一致性与实时交互，由扩散模型负责局部细节增强、纹理补全与风格一致性，从而构建兼具效率、真实感与可控性的数字人生成框架。

6 数据集与评价体系

6.1 主流数据集

数据集的质量、规模与多样性直接决定了数字人生成模型的性能上限。根据任务需求的不同（如泛化能力训练、高保真重建或情感控制），现有的主流数据集可分为以下三类：

1. 大规模野外数据集（in-the-wild datasets）这类数据集采集自真实的网络环境（如 YouTube 视频），涵盖了极其丰富的头部姿态、光照条件和背景干扰，主要用于训练模型的鲁棒性和泛化能力，特别是用于预训练音频-唇形同步判别器（如SyncNet）。

VoxCeleb 系列：VoxCeleb1（Nagrani 等，2017）和VoxCeleb2（Chung 等，2018）是该领域最具影响力的数据集。尽管其分辨率相对较低且面部较小，但其庞大的数据量使其成为训练跨身份通用模型的首选

表3　扩散模型路线典型方法对比

Table 3　Comparison of representative methods in the diffusion model paradigm

方法	核心策略	输出形式	推理效率	时序一致性	表现力/可控性	主要局限
DiffTalk	潜在扩散人脸重演框架	逐帧生成视频	低	低-中	中	姿态控制能力有限;长序列稳定性不足
EMO	端到端 Audio2Video 扩散	直接生成视频帧序列	低	中	高	计算开销大;高分辨率与长序列生成成本高
Hallo	分层音频驱动合成	直接生成视频	低	中-高	高	推理较慢;显存需求高
Hallo2	时间对齐增强生成	直接生成长视频	低	高	高	训练推理成本高;实时性不足
Hallo3	视频扩散 Transformer 生成	直接生成视频	低	高	高	模型规模大;推理延迟高;部署复杂
AniPortrait	ReferenceNet + Motion Adapter	参考图像驱动视频	中-低	中	高	对参考图像质量敏感;大姿态下会出现身份漂移
MuseTalk	潜空间嘴部局部修复	局部重绘视频	高	中	中	主要聚焦嘴部;全局表情与头动表现受限
VASA-1	潜空间外观-姿态-身份解耦	交互式生成视频	中-高	中-高	高	系统复杂;高保真细节仍受限于推理预算
READ	异步扩散生成架构	视频生成(异步推理)	中-高	中	中-高	工程实现复杂;仍存在扩散采样延迟
Dream-Talk	情感风格可控扩散	情感可控视频	低-中	中	高(情感)	情感一致性不足;易夸张或不稳定
EchoMimicV2	音频-姿态联合引导	半身视频生成	低	中	高(动作)	时序稳定性与肢体一致性难;推理开销大
Emo2	末端执行器引导控制	半身交互生成	低	中	高(交互)	控制与自然度平衡困难;实时性不足

基石。

LRS3-TED:针对视觉语音识别(visual speech recognition)任务,LRS3-TED(Afouras 等,2018)数据集。由于其相比 VoxCeleb 具有更正面的视角和更清晰的口型,LRS3 目前被广泛用于评估生成视频的唇形同步准确性。

表4　主流音频驱动数字人数据集对比

Table 4　Comparison of mainstream audio-driven digital human datasets

数据集	数据规模	分辨率/视角	标注	主要用途
VoxCeleb	10万+ utterances;1251 身份	YouTube 野外视频,分辨率不统一	身份;时间戳;说话片段	泛化训练;鲁棒性学习;预训练
VoxCeleb 2	100万+ utterances;6112 身份	YouTube 野外视频,分辨率不统一	身份;时间戳;说话片段	泛化训练;鲁棒性学习;预训练
LRS3-TED	400+小时;5594段演讲	224×224, 25 fps	转写;词级对齐;人脸轨迹	唇形同步评测;视听对齐;VSR
HDTF	362段;15.8小时;300+ 主体	720P/1080P;常用 512×512	时间戳;分辨率;裁剪信息	高保真生成;NeRF/3DGS 训练与评测
MEAD	281400片段;60名演员	1920×1080, 7 视角, 30 fps	情感类别;强度;多视角同步	情感驱动;表情建模;多视角生成
CelebV-HQ	35666片段;15653身份;约65小时	至少 512×512	外观;动作;情感属性	高保真生成;扩散训练;属性编辑

2. 高分辨率重建数据集(high-resolution Datasets)随着NeRF和3DGS等神经渲染技术的兴起，传统的低分辨率数据集已无法满足发丝级细节的建模需求。

HDTF(high-definition talking face)(Zhang等，2021)：Zhang等人在提出流导向生成模型时构建了HDTF数据集。由于其高画质特性，HDTF已成为当前NeRF和3DGS类方法(如ER-NeRF，TalkingGaussian)进行高保真训练的标准基准。

3. 情感与多属性数据集(emotional & attribute datasets)为了赋予数字人更强的表现力和细腻的情感，研究重心逐渐向带有情感标注的数据集转移。

MEAD(Wang等，2020)：作为高质量的情感音视频数据集，其在严格控制的演播室环境下录制并提供了多视角数据，是目前研究情感驱动说话头的核心数据来源。

CelebV-HQ(Zhu等，2022)：针对大规模高保真生成的需求而提出。其高画质与丰富的面部属性(外观、动作、情感)标注相结合，使其成为训练基于扩散模型(diffusion models)的视频生成算法的理想选择。

6.2 评价指标

为了客观量化生成数字人的质量，学术界建立了一套包含唇形同步、视觉质量及身份一致性的综合评价体系。

在唇形同步性方面，常用唇形同步网络(SyncNet)分数。目前通用的标准是利用SyncNet(Chung等，2016)模型进行评估。该模型通过计算音频特征与视频嘴部特征的时间偏移来衡量同步性。常用指标为唇形同步误差置信度(lip sync error-confidence，LSE-C)和唇形同步误差距离(lip sync error-distance，LSE-D)。其中，LSE-C表示音视频对应关系的置信度，数值越高越好；LSE-D表示音频与口型特征之间的距离，数值越低越好。这两项指标能够较直接反映语音与嘴部运动之间的对齐程度，是当前说话头生成任务中最常用的同步性评价指标。

在视觉质量方面，学习感知图像块相似度(learned perceptual image patch similarity，LPIPS)(Zhang等，2018)通过计算生成图像与真实图像在深度特征空间的距离，来衡量感知上的相似度，比像素级指标更符合人类视觉体验。Fréchet inception distance(FID)(Heusel等，2017)也是目前较常见的指标，用于衡量生成图像与真实图像在特征分布上的差异，数值越低表示生成结果越接近真实分布。相比传统的峰值信噪比(peak signal-to-noise ratio，PSNR)、结构相似性(structural similarity，SSIM)等像素级指标，FID更能反映生成结果的整体真实感，因此在近年的音频驱动数字人研究中被广泛采用。

表5 代表性方法在HDTF/MEAD-Neutral上的身份保持与同步对比

Table 5 Comparison of identity preservation and synchronization of representative methods on HDTF/MEAD-Neutral

方法	驱动方式	HDTF FID↓	HDTF FVD↓	HDTF Sync-C↑	HDTF Sync-D↓	MEAD FID↓	MEAD FVD↓	MEAD Sync-C↑	MEAD Sync-D↓
SadTalker	A	22.34	589.63	7.75	7.36	36.88	132.27	6.46	8.07
AniTalker	A	51.66	583.70	7.73	7.43	68.01	941.49	6.76	7.64
AniPortrait	A	17.71	676.30	3.75	10.63	42.43	379.08	2.30	12.38
Hallo	A	17.15	276.31	7.99	7.50	52.07	210.56	7.45	7.47
EmotiveTalk	A	16.64	140.96	8.24	7.09	53.21	207.67	6.82	7.43
PD-FGC	A+V	67.97	464.90	7.30	7.72	121.46	353.75	5.15	8.77
StyleTalk	A+V	29.65	184.60	4.34	10.35	118.48	197.18	3.86	10.74
DreamTalk	A+V	29.37	263.78	6.80	8.03	105.92	204.48	5.64	8.69
EmotiveTalk	A+V	16.09	120.70	8.41	7.11	50.84	153.71	6.79	7.58

在视频级评估方面，Fréchet video distance(FVD)(Unterthiner等，2018)常用于衡量生成视频的时序一致性与整体自然度；在身份保持方面，余弦相似度(cosine similarity，CSIM)通常利用ArcFace

(Deng 等,2019)人脸识别模型提取生成帧与源图像的特征向量,并计算两者之间的余弦相似度。用于衡量生成结果与源人物之间的身份相似性,数值越高说明身份保持越好。由于不同方法在同步精度、画质、时序稳定性和身份一致性之间往往存在权衡,因此通常需要结合多种指标进行综合分析,而不能仅依据单一指标判断模型优劣。

需要指出的是,由于不同方法在分辨率设置、人脸裁剪策略、测试集划分及预处理流程等方面存在差异,本文所列量化结果主要用于文献层面的横向比较与趋势分析,而不构成统一实验协议下的严格公平评测。

表6　代表性方法在 LRS3 benchmark 上的量化结果

Table 6　Quantitative results of representative methods on the LRS3 benchmark

方法	HDTF FID ↓	HDTF CSIM ↑	HDTF LSE-C ↑	MEAD-Neutral FID ↓	MEAD-Neutral CSIM ↑	MEAD-Neutral LSE-C ↑
Wav2Lip	11.21	0.8184	7.46	24.11	0.8432	6.74
VideoReTalking	10.93	0.7989	7.70	26.82	0.8380	5.82
TalkLip	17.61	0.7898	2.29	45.28	0.8509	2.07
DI-Net	7.27	0.8154	6.17	26.34	0.8274	4.40
MuseTalk	6.43	0.8225	6.53	13.42	0.8662	5.15

7　挑战与未来展望

7.1　跨身份泛化与少样本生成

目前的音频驱动模型大多依赖于针对特定目标人物的长视频进行训练,这限制了其在大规模应用中的灵活性。

如今,NeRFFaceSpeech(Kim 等,2024)利用生成先验,在仅有一张参考图像的情况下实现了3D说话头的合成,迈出了One-shot应用的重要一步。GAIA(He 等,2023)则进一步构建了大规模通用的说话人生成框架,展示了Zero-shot驱动的巨大潜力。未来的核心挑战在于如何实现零样本(zero-shot)或少样本(one-shot)的高保真驱动。

此外,突破跨身份泛化瓶颈的技术路径,正不可避免地与多模态大语言模型(multimodal large language models,MLLMs)的演进相交汇。与早期轻量级项目中仅依赖判别式嵌入(embedding)模型进行音视特征隐式对齐不同,未来的泛化范式将被MLLM 彻底重塑。具体而言,MLLM 有望作为统一的“世界模型先验”,直接在海量多模态数据中学习人类发音动作的通用物理规律。未来的技术路径建议探索“大模型语义推理 + 轻量级渲染解码”的非对称架构:即由参数量庞大的 MLLM 负责将音频解码为结构化的 3D 运动指令或潜变量,再由端侧的 NeRF 或 3DGS 模块进行高效的渲染还原。这种范式有望从根本上解决 Zero-shot 场景下的身份漂移难题。

7.2　情感理解与风格化交互

现有的数字人大多表现为冷漠的读稿机器,缺乏对语音中细腻情感的理解和表达。真正的交互式数字人应当具备情感感知能力。

EmoTalker(Shen 等,2024)不仅关注唇形同步,还引入了情感解耦机制,使得生成的面部表情能够根据音频的情感色彩发生变化。CodeTalker(Xing 等,2023)则通过离散动作先验(discrete motion prior),解决了传统方法中表情过度平滑的问题,使得微表情更加生动。

当前情感驱动面临的关键问题在于,模型过度依赖于训练集中的统计相关性而非真正的因果关系。例如,模型往往将“音量增大”错误地直接绑定于“眉头紧锁”。未来的具体技术攻关路径应从表征学习转向因果建模:通过引入结构因果模型对音频特征、文本语义、情感动机与面部肌肉运动进行显式的反事实推理,剔除环境噪声和身份特异性偏置。这不仅能增强微表情生成的合理性,还能赋予数字人根据上下文语境进行风格化情绪过渡的能力。

7.3　从说话头向全身智能体演进

目前的综述和研究主要集中在头部和肩部的生成,但在元宇宙和虚拟社交场景中,自然的肢体语言

(co-speech gestures)是实现沉浸式交互的关键。

如今TalkSHOW(Yi 等,2023)通过自回归网络从语音生成全身的三维动作,初步实现了语音内容与手势的协同。最新的EchoMimicV2等工作也开始尝试将手部动作纳入生成范围。

这里的难点在于"头-手-躯干"的协同建模。未来需要探索如何利用扩散模型生成具有物理合理性的全身动作,解决脚滑(foot sliding)和肢体穿模问题,实现真正的全身虚拟人。

7.4 边缘端部署与轻量化

无论是庞大的NeRF网络还是计算密集的Diffusion模型,目前的高保真数字人大多依赖高端GPU服务器,难以在移动端或VR头显上实时运行。

虽然3D高斯溅射技术的出现缓解了渲染瓶颈,但其存储开销依然巨大。LightGaussian(Fan 等,2024)通过剪枝和蒸馏技术,将3DGS的体积压缩了15倍以上,为移动端部署提供了可能。

未来针对数字人的模型压缩(model compression)、神经渲染烘焙(baking)以及端云协同渲染将是实现数字人普及化的必经之路。

7.5 长序列时序一致性

现有的扩散模型与自回归生成范式在处理长视频(分钟级以上)时,普遍面临误差累积与时序崩坏的理论瓶颈。传统的滑动窗口注意力机制难以维持长程的上下文连贯性。针对这一问题,未来的技术路径可聚焦于两个维度:其一,在网络架构层面,引入具有线性时间复杂度且具备无限上下文记忆容量的状态空间模型(state space models,如 Mamba 架构),以替代或增强现有的时序 Transformer,从而在维持高帧率渲染的同时锁定长期动作一致性;其二,在生成理论层面,探索流匹配(flow matching)框架下的最优传输轨迹约束,通过在底层常微分方程(ODE)的求解过程中引入帧间物理动力学正则化,从数学底层消除生成视频的随机闪烁与高频抖动伪影。

7.6 伦理风险与防御机制(Deepfake检测)

随着基于3DGS与扩散模型等前沿技术的普及,音频驱动数字人生成的视觉逼真度与时序连贯性已逐渐跨越"恐怖谷",达到人眼难以辨识的临界点。这种高度拟真性在赋能虚拟陪伴、数字分身与元宇宙交互等积极应用的同时,也催生了严峻的伦理争议与信息安全风险。恶意使用者可能利用该技术实施身份盗用、伪造公众人物的虚假发言以操纵舆论,或进行基于"熟人面孔与声音"的精准电信金融诈骗(即Deepfake滥用)。这些行为不仅严重侵犯了个人的肖像权与名誉权,更对现有的社会信任体系与公共安全构成了实质性威胁。

因此,数字人技术走向成熟的必由之路,是同步构建坚实的防御与检测防御体系。在被动鉴伪(passive detection)方面,未来的研究需跳出单一的二维图像伪影识别,向多模态与物理规律检测演进:例如,开发基于视听跨模态时空一致性的鉴伪算法,专门捕捉语音韵律与面部微表情之间的语义错位;或引入生物医学信号先验,通过检测生成视频中缺乏真实人类特有的远程光电容积脉搏波(rPPG,如微小的血液流动与心率变化)来识别伪造内容。在主动防御(active defense)方面,从数据源头与生成管线进行干预是关键趋势:这包括在模型生成阶段强制注入抗擦除的频域不可见数字水印、引入基于区块链的内容溯源确权技术,以及建立严格的声纹与面部特征双重授权校验机制。

总体而言,音频驱动数字人技术的未来发展不能仅局限于"渲染保真度"的单向突破,而必须构建起"生成-鉴伪-追溯"的完整技术闭环。在追求极致表现力的同时,坚守技术向善的底线,实现技术发展与安全治理的双轮驱动,是该领域可持续发展的核心保障。

8 结论与总结

本文系统回顾了音频驱动三维数字人视频生成技术的发展历程,从早期的2D图像变形与生成,到基于神经辐射场(NeRF)的隐式三维重建,再到当前基于3D高斯溅射(3DGS)与扩散模型(diffusion models)的前沿范式。通过对各阶段核心算法、主流数据集及评价体系的深入梳理,可以清晰地看到该领域的技术演进脉络:驱动方式从确定性回归迈向生成式概率建模,视觉表征从低维参数化模型迈向高保真神经渲染,而应用场景正从离线视频合成迈向实时交互体验。

当前,3DGS以其显式的几何表达突破了实时渲染的算力瓶颈,而扩散模型则以其强大的分布拟合能力确立了生成画质的天花板。然而,要实现真正如影随形、如人随行的数字人,仍需在跨身份泛化、

全身协同驱动以及边缘端轻量化部署等方面取得进一步突破。

展望未来，随着多模态大语言模型与生成式人工智能的深度融合，音频驱动技术将不再局限于对唇形与面部动作的机械模拟，而是向具备语义理解、情感共鸣及个性化表达的具身智能体演进。数字人将不仅是元宇宙中的视觉载体，更将成为下一代人机交互中连接物理世界与数字世界的核心枢纽，为社交娱乐、虚拟陪伴及远程协作带来革命性的体验。

参考文献

Afouras T, Chung J S and Zisserman A. 2018. LRS3-TED: a large-scale dataset for visual speech recognition[EB/OL].[2026-04-02]. https://arxiv.org/pdf/1809.00496.pdf [DOI: 10.48550/arXiv.1809.00496]

Albert R, Patney A, Luebke D and Kim J. 2017. Latency Requirements for Foveated Rendering in Virtual Reality. ACM Transactions on Applied Perception, 14(4): 1-13 [DOI: 10.1145/3127589]

Aneja S, Sevastopolsky A, Kirschstein T, Thies J, Dai A and Niessner M. 2025. GaussianSpeech: Audio-Driven Personalized 3D Gaussian Avatars//Proceedings of the IEEE/CVF International Conference on Computer Vision. 13065-13075. https://openaccess.thecvf.com/content/ICCV2025/html/Aneja_GaussianSpeech_Audio-Driven_Personalized_3D_Gaussian_Avatars_ICCV_2025_paper.html

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. https://proceedings.neurips.cc/paper_files/paper/2020/hash/92d1e1eb1cd6f9fba3227870bb6d7f07-Abstract.html

Blanz V and Vetter T. 1999. A morphable model for the synthesis of 3D faces//Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques. ACM Press: 187-194 [DOI: 10.1145/311535.311556]

Chen G, Guo B, Guo Y X, Li C, Tong X, Xu S, et al. 2024. VASA-1: Lifelike Audio-Driven Talking Faces Generated in Real Time//Advances in Neural Information Processing Systems, 37: 660-684 [DOI: 10.52202/079017-0021]

Chen L, Cui G, Liu C, Li Z, Kou Z, Xu Y, et al. 2020. Talking-Head Generation with Rhythmic Head Motion//VedaldiA, BischofH, BroxT, FrahmJ M. Computer Vision – ECCV 2020: Vol. 12354. Cham: Springer International Publishing: 35-51 [DOI: 10.1007/978-3-030-58545-7_3]

Chen L, Maddox R K, Duan Z and Xu C. 2019. Hierarchical Cross-Modal Talking Face Generation With Dynamic Pixel-Wise Loss//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, CA, USA: IEEE: 7824-7833 [DOI: 10.1109/CVPR.2019.00802]

Chen S, Wang C, Chen Z, Wu Y, Liu S, Chen Z, et al. 2022. WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing. IEEE Journal of Selected Topics in Signal Processing, 16(6): 1505-1518 [DOI: 10.1109/JSTSP.2022.3188113]

Chen Y, Wang L, Li Q, Xiao H, Zhang S, Yao H, et al. 2024. MonoGaussianAvatar: Monocular Gaussian Point-based Head Avatar//ACM SIGGRAPH 2024 Conference Papers. Denver CO USA: ACM: 1-9 [DOI: 10.1145/3641519.3657499]

Cho K, Lee J, Yoon H, Hong Y, Ko J, Ahn S, et al. 2024. GaussianTalker: Real-Time High-Fidelity Talking Head Synthesis with Audio-Driven 3D Gaussian Splatting[EB/OL].[2026-04-02]. https://arxiv.org/pdf/2404.16012.pdf [DOI: 10.48550/arXiv.2404.16012]

Chung J S, Nagrani A and Zisserman A. 2018. VoxCeleb2: Deep Speaker Recognition//Interspeech 2018. ISCA: 1086-1090 [DOI: 10.21437/Interspeech.2018-1929]

Chung J S and Zisserman A. 2017. Out of Time: Automated Lip Sync in the Wild//Chen C S, Lu J, Ma K K. Computer Vision – ACCV 2016 Workshops: 10117. Cham: Springer International Publishing: 251-263 [DOI: 10.1007/978-3-319-54427-4_19]

Cui J, Li H, Zhan Y, Shang H, Cheng K, Ma Y, et al. 2025. Hallo3: Highly Dynamic and Realistic Portrait Image Animation with Video Diffusion Transformer//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, TN, USA: IEEE: 21086-21095 [DOI: 10.1109/CVPR52734.2025.01964]

Daněček R, Chhatre K, Tripathi S, Wen Y, Black M and Bolkart T. 2023. Emotional Speech-Driven Animation with Content-Emotion Disentanglement//SIGGRAPH Asia 2023 Conference Papers. Sydney NSW Australia: ACM: 1-13 [DOI: 10.1145/3610548.3618183]

Deng J, Guo J, Xue N and Zafeiriou S. 2019. ArcFace: Additive Angular Margin Loss for Deep Face Recognition//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, CA, USA: IEEE: 4685-4694 [DOI: 10.1109/CVPR.2019.00482]

Duan H, Li J, Fan S, Lin Z, Wu X and Cai W. 2021. Metaverse for Social Good: A University Campus Prototype//Proceedings of the 29th ACM International Conference on Multimedia. Virtual Event China: ACM: 153-161 [DOI: 10.1145/3474085.3479238]

Ezzat T, Geiger G and Poggio T. 2002. Trainable videorealistic speech animation. ACM Transactions on Graphics, 21(3): 388-398 [DOI: 10.1145/566654.566594]

Fan Y, Lin Z, Saito J, Wang W and Komura T. 2022. FaceFormer: Speech-Driven 3D Facial Animation with Transformers//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, LA, USA: IEEE: 18749-18758 [DOI:

10.1109/CVPR52688.2022.01821]

Fan Z, Wang K, Wang Z, Wen K, Xu D and Zhu Z. 2024. LightGaussian: Unbounded 3D Gaussian Compression with 15x Reduction and 200+ FPS//Advances in Neural Information Processing Systems, 37: 140138-140158 [DOI: 10.52202/079017-4447]

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

Guo Y, Chen K, Liang S, Liu Y J, Bao H and Zhang J. 2021. AD-NeRF: Audio Driven Neural Radiance Fields for Talking Head Synthesis//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, QC, Canada: IEEE: 5764-5774 [DOI: 10.1109/ICCV48922.2021.00573]

He T, Guo J, Yu R, Wang Y, Zhu J, An K, et al. 2024. GAIA: Zero-Shot Talking Avatar Generation//The Twelfth International Conference on Learning Representations. https://openreview.net/forum?id=ATEawsFUj4

Ho J, Jain A and Abbeel P. 2020. Denoising Diffusion Probabilistic Models//Advances in Neural Information Processing Systems, 33: 6840-6851. https://proceedings.neurips.cc/paper_files/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html

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]

Hu W, Li S, Peng Z, Zhang H, Shi F, Liu X, et al. 2025. GGTalker: Talking Head Synthesis with Generalizable Gaussian Priors and Identity-Specific Adaptation[EB/OL].[2026-04-02]. https://arxiv.org/pdf/2506.21513.pdf [DOI: 10.48550/arXiv.2506.21513]

Jamaludin A, Chung J S and Zisserman A. 2019. You Said That?: Synthesising Talking Faces from Audio. International Journal of Computer Vision, 127(11-12): 1767-1779 [DOI: 10.1007/s11263-019-01150-y]

Kerbl B, Kopanas G, Leimkuehler T and Drettakis G. 2023. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. ACM Transactions on Graphics, 42(4): 1-14 [DOI: 10.1145/3592433]

Kim G, Seo K, Cha S and Noh J. 2024. NeRFFaceSpeech: One-shot Audio-Driven 3D Talking Head Synthesis via Generative Prior[EB/OL].[2026-04-02]. https://arxiv.org/pdf/2405.05749.pdf [DOI: 10.48550/arXiv.2405.05749]

Kocabas M, Chang J H R, Gabriel J, Tuzel O and Ranjan A. 2024. HUGS: Human Gaussian Splats//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, WA, USA: IEEE: 505-515 [DOI: 10.1109/CVPR52733.2024.00055]

Li J, Zhang J, Bai X, Zheng J, Ning X, Zhou J, et al. 2025. TalkingGaussian: Structure-Persistent 3D Talking Head Synthesis via Gaussian Splatting//Leonardis A, Ricci E, Roth S, Russakovsky O, Sattler T, Varol G. Computer Vision – ECCV 2024: Vol. 15068. Cham: Springer Nature Switzerland: 127-145 [DOI: 10.1007/978-3-031-72684-2_8]

Li J, Zhang J, Bai X, Zhou J and Gu L. 2023. Efficient Region-Aware Neural Radiance Fields for High-Fidelity Talking Portrait Synthesis//Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 7534-7544 [DOI: 10.1109/ICCV51070.2023.00696]

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): 1-17 [DOI: 10.1145/3130800.3130813]

Lipman Y, Chen R T Q, Ben-Hamu H, Nickel M and Le M. 2023. Flow Matching for Generative Modeling//The Eleventh International Conference on Learning Representations. https://openreview.net/forum?id=PqvMRDCJT9t

Liu J, Chen P, Wang X, Fu X M, Dai J and Han J Z. 2022. Critical review of human face reenactment methods. Journal of Image and Graphics, 27(09): 2629-2651 (刘锦, 陈鹏, 王茜, 付晓蒙, 戴娇, 韩冀中. 2022. 人类面部重演方法综述. 中国图象图形学报, 27(09): 2629-2651) [DOI: 10.11834/jig.211243]

Liu K, Wei J, He S, Ma Z, Zhang C, Xie N, et al. 2025. SyncGaussian: Stable 3D Gaussian-Based Talking Head Generation with Enhanced Lip Sync via Discriminative Speech Features//Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence. Montreal, Canada: International Joint Conferences on Artificial Intelligence Organization: 1576-1584 [DOI: 10.24963/ijcai.2025/176]

Liu X, Xu Y, Wu Q, Zhou H, Wu W and Zhou B. 2022. Semantic-Aware Implicit Neural Audio-Driven Video Portrait Generation//AvidanS, BrostowG, CisséM, FarinellaG M, HassnerT. Computer Vision – ECCV 2022: Vol. 13697. Cham: Springer Nature Switzerland: 106-125 [DOI: 10.1007/978-3-031-19836-6_7]

Ma Z, Zheng Z, Ye J, Li J, Gao Z, Zhang S, et al. 2024. emotion2vec: Self-Supervised Pre-Training for Speech Emotion Representation//Findings of the Association for Computational Linguistics ACL 2024. Bangkok, Thailand and virtual meeting: Association for Computational Linguistics: 15747-15760 [DOI: 10.18653/v1/2024.findings-acl.931]

McGurk H and MacDonald J. 1976. Hearing lips and seeing voices. Nature, 264(5588): 746-748 [DOI: 10.1038/264746a0]

Meng R, Zhang X, Li Y and Ma C. 2025. EchoMimicV2: Towards Striking, Simplified, and Semi-Body Human Animation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, TN, USA: IEEE: 5489-5498 [DOI: 10.1109/

CVPR52734.2025.00516]

Mildenhall B, Srinivasan P P, Tancik M, Barron J T, Ramamoorthi R and Ng R. 2020. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis//VedaldiA, BischofH, BroxT, FrahmJ M. Computer Vision – ECCV 2020: Vol. 12346. Cham: Springer International Publishing: 405-421 [DOI: 10.1007/978-3-030-58452-8_24]

Mori M, MacDorman K F and Kageki N. 2012. The Uncanny Valley [From the Field]. IEEE Robotics & Automation Magazine, 19(2): 98-100 [DOI: 10.1109/MRA.2012.2192811]

Nagrani A, Chung J S and Zisserman A. 2017. VoxCeleb: A Large-Scale Speaker Identification Dataset//Interspeech 2017. ISCA: 2616-2620 [DOI: 10.21437/Interspeech.2017-950]

Peng Z, Hu W, Ma J, Zhu X, Zhang X, Zhao H, et al. 2025. SyncTalk++: High-Fidelity and Efficient Synchronized Talking Heads Synthesis Using Gaussian Splatting. IEEE Transactions on Pattern Analysis and Machine Intelligence: 1-18 [DOI: 10.1109/TPAMI.2025.3630057]

Peng Z, Hu W, Shi Y, Zhu X, Zhang X, Zhao H, et al. 2024. SyncTalk: The Devil is in the Synchronization for Talking Head Synthesis//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, WA, USA: IEEE: 666-676 [DOI: 10.1109/CVPR52733.2024.00070]

Prajwal K R, Mukhopadhyay R, Namboodiri V P and Jawahar C V. 2020. A Lip Sync Expert Is All You Need for Speech to Lip Generation In the Wild//Proceedings of the 28th ACM International Conference on Multimedia. Seattle WA USA: ACM: 484-492 [DOI: 10.1145/3394171.3413532]

Qian S, Kirschstein T, Schoneveld L, Davoli D, Giebenhain S and Nießner M. 2024. GaussianAvatars: Photorealistic Head Avatars with Rigged 3D Gaussians//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, WA, USA: IEEE: 20299-20309 [DOI: 10.1109/CVPR52733.2024.01919]

Radford A, Kim J W, Xu T, Brockman G, McLeavey C and Sutskever I. 2023. Robust Speech Recognition via Large-Scale Weak Supervision//Proceedings of the 40th International Conference on Machine Learning. PMLR, 202: 28492-28518. https://proceedings.mlr.press/v202/radford23a.html

Rivero A, Athar S, Shu Z and Samaras D. 2025. Rig3DGS: Creating Controllable Portraits From Casual Monocular Videos//2025 International Conference on 3D Vision (3DV). Singapore, Singapore: IEEE: 1541-1550 [DOI: 10.1109/3DV66043.2025.00144]

Shen S, Li W, Zhu Z, Duan Y, Zhou J and Lu J. 2022. Learning Dynamic Facial Radiance Fields for Few-Shot Talking Head Synthesis//AvidanS, BrostowG, CisséM, FarinellaG M, HassnerT. Computer Vision – ECCV 2022: Vol. 13672. Cham: Springer Nature Switzerland: 666-682 [DOI: 10.1007/978-3-031-19775-8_39]

Shen S, Zhao W, Meng Z, Li W, Zhu Z, Zhou J, et al. 2023. DiffTalk: Crafting Diffusion Models for Generalized Audio-Driven Portraits Animation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, BC, Canada: IEEE: 1982-1991 [DOI: 10.1109/CVPR52729.2023.00197]

Shen X, Khan F F and Elhoseiny M. 2025. EmoTalker: Audio Driven Emotion Aware Talking Head Generation//Cho M, Laptev I, Tran D, Yao A, Zha H. Computer Vision – ACCV 2024: Vol. 15476. Singapore: Springer Nature Singapore: 131-147 [DOI: 10.1007/978-981-96-0917-8_8]

Stan S, Haque K I and Yumak Z. 2023. FaceDiffuser: Speech-Driven 3D Facial Animation Synthesis Using Diffusion//ACM SIGGRAPH Conference on Motion Interaction and Games. Rennes France: ACM: 1-11 [DOI: 10.1145/3623264.3624447]

Suwajanakorn S, Seitz S M and Kemelmacher-Shlizerman I. 2017. Synthesizing Obama: learning lip sync from audio. ACM Transactions on Graphics, 36(4): 1-13 [DOI: 10.1145/3072959.3073640]

Tang J, Wang K, Zhou H, Chen X, He D, Hu T, et al. 2025. Real-Time Neural Radiance Talking Portrait Synthesis via Audio-Spatial Decomposition. International Journal of Computer Vision, 133(9): 6362-6373 [DOI: 10.1007/s11263-025-02481-9]

Tian L, Hu S, Wang Q, Zhang B and Bo L. 2025. EMO2: End-Effector Guided Audio-Driven Avatar Video Generation[EB/OL].[2026-04-02]. https://arxiv.org/pdf/2501.10687.pdf [DOI: 10.48550/arXiv.2501.10687]

Tian L, Wang Q, Zhang B and Bo L. 2025. EMO: Emote Portrait Alive Generating Expressive Portrait Videos with Audio2Video Diffusion Model Under Weak Conditions//Leonardis A, Ricci E, Roth S, Russakovsky O, Sattler T, Varol G. Computer Vision – ECCV 2024: Vol. 15141. Cham: Springer Nature Switzerland: 244-260 [DOI: 10.1007/978-3-031-73010-8_15]

Vougioukas K, Petridis S and Pantic M. 2020. Realistic Speech-Driven Facial Animation with GANs. International Journal of Computer Vision, 128(5): 1398-1413 [DOI: 10.1007/s11263-019-01251-8]

Wang C, Kang D, Sun H, Qian S, Wang Z, Bao L, et al. 2025. MeGA: Hybrid Mesh-Gaussian Head Avatar for High-Fidelity Rendering and Head Editing//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, TN, USA: IEEE: 26274-26284 [DOI: 10.1109/CVPR52734.2025.02447]

Wang H, Weng Y, Du J, Xu H, Wu X, He S, et al. 2025. READ: Real-Time and Efficient Asynchronous Diffusion for Audio-Driven Talking Head Generation[EB/OL].[2026-04-02]. https://arxiv.org/pdf/2508.03457.pdf [DOI: 10.48550/arXiv.2508.03457]

Wang K, Wu Q, Song L, Yang Z, Wu W, Qian C, et al. 2020. MEAD: A Large-Scale Audio-Visual Dataset for Emotional Talking-Face Generation//VedaldiA, BischofH, BroxT, FrahmJ M. Computer

Vision – ECCV 2020: Vol. 12366. Cham: Springer International Publishing: 700-717 [DOI: 10.1007/978-3-030-58589-1_42]

Wang S, Li L, Ding Y, Fan C and Yu X. 2021. Audio2Head: Audio-Driven One-shot Talking-Head Generation with Natural Head Motion//Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence. Montreal, Canada: International Joint Conferences on Artificial Intelligence Organization: 1098-1105 [DOI: 10.24963/ijcai.2021/152]

Wang W and Fu Y. 2025. EmoGene: Audio-Driven Emotional 3D Talking-Head Generation//2025 IEEE 19th International Conference on Automatic Face and Gesture Recognition (FG). Tampa/Clearwater, FL, USA: IEEE: 1-10 [DOI: 10.1109/FG61629.2025.11099460]

Wei H, Yang Z and Wang Z. 2024. AniPortrait: Audio-Driven Synthesis of Photorealistic Portrait Animation[EB/OL].[2026-04-02]. https://arxiv.org/pdf/2403.17694.pdf [DOI: 10.48550/arXiv.2403.17694]

Wiles O, Koepke A S and Zisserman A. 2018. X2Face: A Network for Controlling Face Generation Using Images, Audio, and Pose Codes//FerrariV, HebertM, SminchisescuC, WeissY. Computer Vision – ECCV 2018: Vol. 11217. Cham: Springer International Publishing: 690-706 [DOI: 10.1007/978-3-030-01261-8_41]

Xiang J, Gao X, Guo Y and Zhang J. 2024. FlashAvatar: High-Fidelity Head Avatar with Efficient Gaussian Embedding//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, WA, USA: IEEE: 1802-1812 [DOI: 10.1109/CVPR52733.2024.00177]

Xing J, Xia M, Zhang Y, Cun X, Wang J and Wong T T. 2023. CodeTalker: Speech-Driven 3D Facial Animation with Discrete Motion Prior//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, BC, Canada: IEEE: 12780-12790 [DOI: 10.1109/CVPR52729.2023.01229]

Xu M, Li H, Su Q, Shang H, Zhang L, Liu C, et al. 2024. Hallo: Hierarchical Audio-Driven Visual Synthesis for Portrait Image Animation[EB/OL].[2026-04-02]. https://arxiv.org/pdf/2406.08801.pdf [DOI: 10.48550/arXiv.2406.08801]

Xu Y, Chen B, Li Z, Zhang H, Wang L, Zheng Z, et al. 2024. Gaussian Head Avatar: Ultra High-Fidelity Head Avatar via Dynamic Gaussians//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, WA, USA: IEEE: 1931-1941 [DOI: 10.1109/CVPR52733.2024.00189]

Yang Z, Gao X, Zhou W, Jiao S, Zhang Y and Jin X. 2024. Deformable 3D Gaussians for High-Fidelity Monocular Dynamic Scene Reconstruction//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, WA, USA: IEEE: 20331-20341 [DOI: 10.1109/CVPR52733.2024.01922]

Ye Z, He J, Jiang Z, Huang R, Huang J, Liu J, et al. 2023. GeneFace++: Generalized and Stable Real-Time Audio-Driven 3D Talking Face Generation[EB/OL].[2026-04-02]. https://arxiv.org/pdf/2305.00787.pdf [DOI: 10.48550/arXiv.2305.00787]

Ye Z, Jiang Z, Ren Y, Liu J, He J and Zhao Z. 2023. GeneFace: Generalized and High-Fidelity Audio-Driven 3D Talking Face Synthesis//The Eleventh International Conference on Learning Representations. https://openreview.net/forum?id=YfwMIDhPccD

Ye Z, Zhong T, Ren Y, Yang J, Li W, Huang J, et al. 2024. Real3D-Portrait: One-Shot Realistic 3D Talking Portrait Synthesis//The Twelfth International Conference on Learning Representations. https://openreview.net/forum?id=7ERQPyR2eb

Yi H, Liang H, Liu Y, Cao Q, Wen Y, Bolkart T, et al. 2023. Generating Holistic 3D Human Motion from Speech//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, BC, Canada: IEEE: 469-480 [DOI: 10.1109/CVPR52729.2023.00053]

Zhang C, Wang C, Zhang J, Xu H, Song G, Xie Y, et al. 2023. DREAM-Talk: Diffusion-based Realistic Emotional Audio-Driven Method for Single Image Talking Face Generation[EB/OL].[2026-04-02]. https://arxiv.org/pdf/2312.13578.pdf [DOI: 10.48550/arXiv.2312.13578]

Zhang R, Isola P, Efros A A, Shechtman E and Wang O. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT: IEEE: 586-595 [DOI: 10.1109/CVPR.2018.00068]

Zhang Y, Zhong Z, Liu M, Chen Z, Wu B, Zeng Y, et al. 2024. MuseTalk: Real-Time High-Fidelity Video Dubbing via Spatio-Temporal Sampling[EB/OL].[2026-04-02]. https://arxiv.org/pdf/2410.10122.pdf [DOI: 10.48550/arXiv.2410.10122]

Zhang Z, Hu Z, Deng W, Fan C, Lv T and Ding Y. 2023. DINet: Deformation Inpainting Network for Realistic Face Visually Dubbing on High Resolution Video. Proceedings of the AAAI Conference on Artificial Intelligence, 37(3): 3543-3551 [DOI: 10.1609/aaai.v37i3.25464]

Zhang Z, Li L, Ding Y and Fan C. 2021. Flow-guided One-shot Talking Face Generation with a High-resolution Audio-visual Dataset//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, TN, USA: IEEE: 3660-3669 [DOI: 10.1109/CVPR46437.2021.00366]

Zhao Z, Bao Z, Li Q, Qiu G and Liu K. 2026. PSAvatar: A Point-Based Shape Model for Real-Time Head Avatar Animation With 3D Gaussian Splatting. IEEE Transactions on Visualization and Computer Graphics: 1-15 [DOI: 10.1109/TVCG.2026.3676544]

Zhen D, Yin S, Qin S, Yi H, Zhang Z, Liu S, et al. 2025. Teller: Real-Time Streaming Audio-Driven Portrait Animation with Autoregressive Motion Generation//Proceedings of the IEEE/CVF Confer-

ence on Computer Vision and Pattern Recognition. Nashville, TN, USA: IEEE: 21075-21085 [DOI: 10.1109/CVPR52734.2025.01963]

Zhou H, Sun Y, Wu W, Loy C C, Wang X and Liu Z. 2021. Pose-Controllable Talking Face Generation by Implicitly Modularized Audio-Visual Representation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, TN, USA: IEEE: 4174-4184 [DOI: 10.1109/CVPR46437.2021.00416]

Zhou Y, Han X, Shechtman E, Echevarria J, Kalogerakis E and Li D. 2020. MakeItTalk: Speaker-Aware talking-head animation. ACM Transactions on Graphics, 39(6): 1-15 [DOI: 10.1145/3414685.3417774]

Zhou Z, Feng Q and Li H. 2025. SynGauss: Real-Time 3D Gaussian Splatting for Audio-Driven Talking Head Synthesis. IEEE Access, 13: 42167-42177 [DOI: 10.1109/ACCESS.2025.3548015]

Zhu H, Wu W, Zhu W, Jiang L, Tang S, Zhang L, et al. 2022. CelebV-HQ: A Large-Scale Video Facial Attributes Dataset//AvidanS, BrostowG, CisséM, FarinellaG M, HassnerT. Computer Vision – ECCV 2022: Vol. 13667. Cham: Springer Nature Switzerland: 650-667 [DOI: 10.1007/978-3-031-20071-7_38]

Zhu T, Yu Y, Wang L, Sun F and Zheng W. 2025. EGSTalker: Real-Time Audio-Driven Talking Head Generation with Efficient Gaussian Deformation//2025 IEEE International Conference on Systems, Man, and Cybernetics (SMC). Vienna, Austria: IEEE: 7542-7547 [DOI: 10.1109/SMC58881.2025.11343147]

Zhu T, Yu Y, Wang L, Sun F and Zheng W. 2026. PGSTalker: Real-Time Audio-Driven Talking Head Generation via 3D Gaussian Splatting with Pixel-Aware Density Control//TaniguchiT, LeungC S A, KozunoT, YoshimotoJ, MahmudM, DoborjehM, et al. Neural Information Processing: Vol. 16312. Singapore: Springer Nature Singapore: 112-126 [DOI: 10.1007/978-981-95-4384-7_9]

作者简介

李卫斌,通信作者,男,教授,主要研究方向为AIGC大语言模型、工业智能与数字人。E-mail:weibinli@xidian.edu.cn

高佳峰,男,硕士,主要研究方向为数字人技术。E-mail:25241215082@stu.xidian.edu.cn

徐兵,男,高级工程师,主要研究方向为民机试飞与人工智能的融合应用。E-mail:

侯彪,男,教授,主要研究方向为遥感图像解译与目标识别、无人系统协同感知、人工智能芯片及系统。E-mail:avcodec@163.com

焦李成,男,教授,主要研究方向为图像理解与目标识别、智能感知与计算、深度学习与类脑计算。E-mail:lchjiao@mail.xidian.edu.cn