中图法分类号:TP18;TP37　　文献标识码:A　　文章编号:1006-8961(2026)10-3545-27

论文引用格式:Du R Q, Yang B A, Zhou F B, Qu W and Li T. 2026. A survey on multimodal real-time interactive digital humans. Journal of Image and Graphics, 31(10):3545-3571(杜瑞麒, 杨柏蔼, 周丰波, 屈薇, 李涛. 2026. 多模态实时交互式数字人综述. 中国图象图形学报, 31(10):3545-3571)[DOI:10.11834/jig.250511]

多模态实时交互式数字人综述

杜瑞麒[1],杨柏蔼[2],周丰波[1],屈薇[1],李涛[1*]

1. 湖南科技大学计算机科学与工程学院,湘潭　411201; 2. 湖南大学设计艺术学院,长沙　410082

摘　要: 多模态实时交互式数字人作为新一代人机交互的核心载体,正随着多模态大模型、AR(augmented reality)/VR(virtual reality),以及5G/6G等技术的快速发展,逐步从传统的单模态输入方式,演进为融合语音、视觉、动作乃至情感信号的多模态自然交互形式。数字人本身也经历了从非交互式虚拟形象到具备语义理解、情感感知与主动响应能力的智能体的转变。本文从发展脉络、关键特征与技术体系3个方面系统梳理了这一进程:首先回顾了数字人由静态展示向多模态交互演进的轨迹,突出其在沉浸感、实时响应与情感共鸣能力上的提升;随后重点剖析了建模、实时驱动与渲染三大核心技术,涵盖3D高斯溅射(3D Gaussian splatting)、神经辐射场(neural radiance field,NeRF)隐式表征、多模态融合驱动及神经渲染等前沿手段,揭示了高保真视觉呈现与低延迟交互之间的技术权衡;进一步提出了多模态数字人的通用系统框架,划分为感知、融合、生成与拓展4个层次,并总结了语言生成、情感语音合成与表情驱动等关键模块间的协同机制。未来,数字人技术的发展将更加注重轻量化部署、跨模态一致性保障与情感共生智能的实现,有望在教育培训、医疗健康、文化娱乐及人机协作等场景中,提供更加自然、可信且富有温度的人机交互体验。

关键词: 数字人;多模态交互;多模态融合;实时驱动;人机交互

A survey on multimodal real-time interactive digital humans

Du Ruiqi[1], Yang Boai[2], Zhou Fengbo[1], Qu Wei[1], Li Tao[1*]

1. School of Computer Science and Engineering, Hunan University of Science and Technology, Xiangtan　411201, China;
2. School of Design, Hunan University, Changsha　410082, China

Abstract: With the continued evolution of artificial intelligence, 5G/6G communication, and multimodal interaction technologies, digital humans are rapidly transforming from static virtual avatars into intelligent agents capable of real-time, natural, and emotionally resonant communication. This transition marks a broader shift in human-computer interaction, moving beyond scripted responses toward context-aware, adaptive systems that can perceive, reason, and respond with human-like fluidity across multiple sensory channels. This paper presents a comprehensive review of real-time multimodal interactive digital humans, focusing on their conceptual evolution, technical framework, and key challenges. We trace the historical trajectory from early non-interactive virtual characters on platforms such as second life to today's AI-driven agents that integrate language understanding, emotional intelligence, and physical embodiment in immersive environments. The survey systematically examines the core technologies underlying modeling, real-time driving, and rendering, and provides an integrated view of how multimodal data—including speech, facial expression, gaze, gestures, and physiological signals—can be perceived, fused, and generated to support coherent and lifelike human-machine interaction. From the perspective of modeling, we analyze the transition from traditional 3D parametric models to neural implicit repre-

收稿日期:2025-10-28;修回日期:2026-02-14;预印本日期:2026-02-21

*通信作者:李涛　tlee@hnust.edu.cn

sentations such as NeRF and 3D Gaussian splatting, which enable high-fidelity reconstruction and photorealistic rendering while still supporting real-time performance. Recent advances in 3D Gaussian splatting, including reported rendering speeds of more than 300 frame/s on consumer hardware, have helped bridge the longstanding gap between visual fidelity and interactive responsiveness that previously constrained practical deployment. In terms of motion and expression control, the survey highlights the role of multimodal fusion and diffusion-based generation in synchronizing voice, facial animation, and body motion, thereby improving the naturalness and temporal consistency of interactive behavior. We critically examine the evolution from rule-based animation systems to end-to-end neural approaches that jointly model cross-modal dependencies, reducing the uncanny-valley effect through stronger temporal coherence and greater emotional authenticity. Rendering technologies are reviewed from both physically based and neural perspectives, with emphasis on balancing visual realism and latency in real-time deployment. Emerging hybrid approaches that combine explicit geometric representations with neural detail enhancement appear particularly promising for achieving both physical plausibility and computational efficiency in resource-constrained settings. Furthermore, the paper proposes a unified system architecture for digital humans consisting of four hierarchical layers: perception, fusion, generation, and extension, each responsible for multimodal sensing, semantic integration, expressive generation, and cross-domain adaptability, respectively. The perception layer integrates heterogeneous inputs ranging from microphone arrays and RGB-D cameras to emerging physiological sensors such as EEG and EDA, thereby establishing a rich basis for contextual understanding. Within this architecture, the integration of large multimodal models (LMMs) and large language models (LLMs) enables digital humans to perceive complex environmental and emotional cues, reason about user intent, and produce personalized, empathetic responses. Modern foundation models also support in-context adaptation, allowing digital humans to adjust their communication style dynamically according to interaction history, user characteristics, and situational context without explicit retraining. Representative technologies such as neural radiance fields, real-time diffusion-based speech synthesis, and emotion-driven dialogue generation are analyzed to illustrate their contributions to interactivity and immersion. Particular attention is given to latency-critical components, for which end-to-end response times below 200 ms are often essential to preserving conversational flow and user trust in real-world applications. Based on this synthesis, the paper identifies current challenges including latency optimization, multimodal alignment, semantic coherence, and emotional authenticity. We highlight the fundamental tension between computational complexity and real-time constraints, especially when deploying high-fidelity models on edge devices with limited power and thermal budgets. Addressing these issues will require advances in lightweight model compression, edge-native deployment, and cross-modal consistency learning. Promising directions include neural architecture search for modality-specific subnetworks and dynamic computation scaling based on interaction complexity. Looking ahead, digital-human research is expected to move toward more deeply integrated affective computing, multimodal reasoning, and socially adaptive intelligence, ultimately enabling human-machine interaction that is perceptually natural, emotionally grounded, and ethically trustworthy. At the same time, crucial issues such as digital identity, consent-based data usage, and robust watermarking mechanisms must be addressed to ensure responsible large-scale deployment. By summarizing state-of-the-art methods and mapping emerging trends, this study provides both theoretical insight and practical guidance for the development of the next generation of real-time multimodal interactive digital humans, with broad applications in education, healthcare, entertainment, and collaborative robotics.

Key words: digital human; multimodal interaction; multimodal fusion; real-time driving; human-computer interaction

0 引 言

自1946年第一台电子计算机诞生至今，人机交互方式经历了3次重大演进：20世纪60—70年代，命令行接口以文本命令和键盘输入为核心，操作精准但学习门槛高，仅面向专业人员；80—90年代，图形用户界面引入窗口、图标、菜单与鼠标范式，实现“所见即所得”，大幅降低学习成本，推动个人电脑普及；移动互联网时期，多点触控交互依赖电容屏与手势识别，支持自然、直观的双手操作，使移动设备成为日常生活入口。近年来，随着多模态大模型、AR（augmented reality）/ VR（virtual reality）、5G/6G低延迟网络、实时渲染管线以及机器人技术的快速发展，

人机交互正迈向第 4 次重大演进，即基于语音、表情、手势、眼动乃至生理信号的“多模态自然交互”时代（Sinha 等，2024）。大模型具备跨模态理解与生成能力，能够融合语言、视觉与动作信息，实现智能响应；AR/VR 构建沉浸式 3D 交互空间；高速网络保障实时反馈；机器人则赋予系统物理交互能力。这些技术的协同突破，使得人机交互不再依赖单一输入方式，而是趋向于模拟人类自然交流的多通道融合模式。图 1 概述了从早期的命令行交互到现代的多模态自然交互时代，技术演进的关键节点。它不仅展示了每个阶段的核心技术进展，还体现了人机交互从单一输入方式到多模态、智能化交互模式的转变。

在此背景下，数字人作为多模态交互的核心载体，正从概念走向规模化应用。数字人是以计算机图形学、人工智能、传感控制与认知科学为基础构建的虚拟或实体拟人化代理，具备类人外貌、动作、语言表达及一定程度的认知交互能力。其发展可追溯至 20 世纪末的虚拟形象（如 Second Life 或游戏中的卡通化角色），这些早期形象依赖预设动画与脚本驱动，交互能力有限且缺乏智能性。需要明确的是，数字人与游戏中固定程序设定的非玩家角色（non-player character，NPC）不同，后者遵循确定性脚本，缺乏自主适应能力，而数字人具备上下文感知的独立交流能力。其演进脉络可概括为：从需人工操控的预设虚拟角色，到具备基础反应能力的半自主 NPC，再到利用人工智能（artificial intelligence，AI）实现主动、适应性交互的全智能数字人。随后，在客服自动化、远程教学以及虚拟主播等场景中，国内外虚拟偶像（如初音未来、洛天依）推动了二维（2D）/三维（3D）虚拟助手的视觉表现提升，但仍受限于单向输出与低层次语义理解能力。

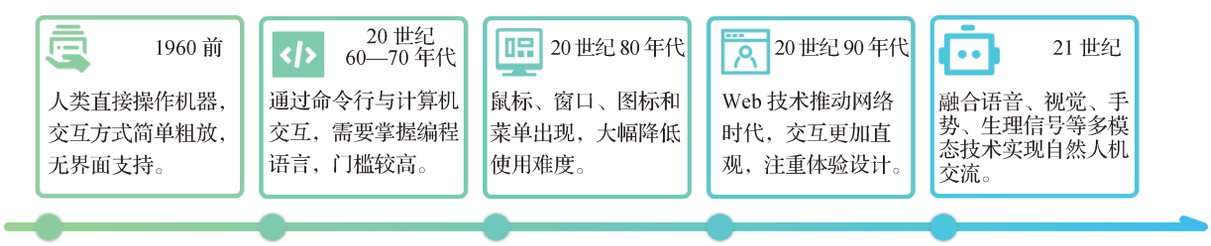


图 1　人机交互发展历程

Fig. 1　The evolution of human-computer interaction

与早期虚拟形象相比，新一代数字人呈现出三大典型特征：

1）高沉浸感。早期虚拟角色普遍存在动作僵硬、口型不同步以及情感表达单一等问题，严重削弱用户体验的真实感。为解决多模态表达失调问题，研究者提出融合语音语义与情感状态的表情生成机制，构建了基于 ResNet18（residual network）与 GPT（generative pre-trained Transformer）的复合表情生成框架（Yang 等，2025），通过“语音语义解析—多模态情感识别—AU（action unit）级 3D 面部动作单元控制”流程，提升了表情动态的自然度，用户体验测评显示，其自然度评分获得可观提升（表情自然度从 3.54/5 提升至 3.94/5）。此外，Meta 团队提出结合向量量化（vector quantization，VQ）与扩散模型的方法，实现了从语音信号到多样化全身手势动作的生成，增强了非语言行为的表现力与多样性（Ng 等，2024）。

2）高实时性。对于交互式数字人而言，极低的端到端响应延迟是实现自然、流畅人机对话的核心保障。较明显的延迟不仅会破坏对话节奏，还易引发用户的认知中断与交互断裂感，严重削弱沉浸体验与信任感。为此，现代数字人系统广泛采用边缘计算架构，将感知理解与生成推理任务下沉至近用户侧，降低云端往返时延（Carvalho 等，2022）；同时结合模型轻量化技术（如知识蒸馏、量化压缩、TinyML 等），实现大模型在终端或边缘设备上的高效部署。在传输层面，依托 WebRTC（web real-time communication）、5G 超可靠低延迟通信（ultra-reliable low-latency communication，URLLC）等协议，保障音视频流与控制信号的高并发、低抖动传输。

3）高共情能力。情感理解是实现深度人机共情的关键。传统虚拟助手往往基于固定规则或简单分类模型进行情绪识别，难以捕捉复杂语境下的细微情感变化，更无法生成具有情感适配性的回应，导致

交互缺乏温度与人性化。为突破这一局限，HumanSense（Qin等，2026）构建了一个以人为中心基准的评估框架，专注于多模态感知、隐式上下文理解以及交互场景中的响应策略。该框架通过涵盖情绪识别、意图推断和共情响应等维度，提升了对人类中心化交互能力的细粒度评估水平，并为模型优化（如基于多阶段强化学习的推理能力增强）提供明确方向和量化依据。

在上述技术演进与特征深化的推动下，学术界与产业界对数字人系统的研究正迅速拓展，其核心目标已从单一模态的生成优化转向多模态融合与实时交互性能的系统提升。基于这一趋势，学术界围绕数字人技术开辟了一系列新兴研究方向，并在若干关键领域取得显著进展。例如：在数字人建模方面，扩散模型与神经辐射场（neural radiance field，NeRF）实现了动态光照下的高保真建模；在动作生成方面，结合动作捕捉数据与生成对抗网络（generative adversarial network，GAN）或扩散模型，可生成多样化、情感化的人体姿态；在驱动机制方面，端到端语音到表情映射模型（如Audio2Face）显著提升了口型同步的自然性与时序一致性；在语义与认知层面，大语言模型（large language model，LLM）增强了语义理解与上下文推理能力，使数字人具备更高层次的交互智能。

相关研究在各模态内部持续突破，然而当前数字人系统仍普遍存在“重生成质量—轻交互效率”的倾向。在端到端建模与优化、动作生成管线设计以及交互响应时延的量化评估等方面尚缺乏系统性研究。这种局限使得现有系统难以满足远程实时协作、沉浸式VR/AR及即时人机对话等高实时性场景对低延迟与高同步性的要求，严重制约了数字人从“可看”向“可交互”的实质性跨越。因此，本文将围绕“多模态实时交互式数字人”的前沿进展展开系统综述。

1 数字人的发展与演进

数字人的发展历程展现了人机交互技术的持续演进。从最初以信息传递和拟人化展示为主的非交互式数字人，到能够实时响应用户输入的交互式数字人，其形态与功能经历了重要转变。本文将从非交互式数字人入手，梳理单模态与多模态交互式数字人的演进逻辑，以揭示数字人技术在交互性、智能化与沉浸感方面的逐步提升。

1.1 非交互式数字人

非交互式数字人是拟数字人发展的早期形态，其核心特征在于缺乏与用户的实时互动，通常通过预设脚本、录制影像或语音合成的方式呈现内容。这类数字人最初的使命主要是信息传递与拟人化展示，以满足数字化传播和人机界面友好化的需求。

早期虚拟人多依赖2D或3D建模与语音合成技术相结合，能够实现持续、低成本的内容输出，但在交互性与情境适应性方面存在明显不足。在产业化应用初期，非交互式数字人主要活跃于新闻播报、品牌代言和虚拟客服等场景，其价值在于替代部分重复性人力，实现信息传递的拟人化与高效化。然而，由于缺乏多轮交互能力，其在用户体验和黏性培养方面受到限制。

从学术研究角度看，非交互式数字人被视为探讨人机关系的重要参照对象。Arima等人（2025）通过虚拟现实实验发现，非交互式数字人的出现虽然能诱发个体的社会性反应（如一致性行为），但由于缺乏真实的互动协调，难以有效激发合作或复杂的社会交往，揭示了其在社会存在感营造方面的局限性。

另外，在在线消费情境中，Liew等人（2017）的实证研究表明，电商网站嵌入的非交互式会说话数字人可在一定程度上提升消费者的社会临场感，并对男性群体的信任与重复访问意图产生积极作用。然而，对于女性群体，因缺乏自然交互与语音拟真度不足，可能导致信息可信度下降，甚至削弱购买意愿。这一结果说明，非交互式数字人虽然能在感知层面营造“有人陪伴”的氛围，但其影响存在群体差异，其社会价值需结合用户特征进行理解。

总体而言，非交互式数字人作为虚拟人技术的重要过渡阶段，凭借低成本、易部署和突出的拟人化表达优势，为虚拟人产业的普及与社会认知奠定了基础。然而，其在互动性、可信度与个性化体验上的不足，也促使研究与产业逐步迈向可交互、实时响应的数字人形态，以更好地满足人类在信息传递、社会交往与学习陪伴等方面的复杂需求。

1.2 交互式数字人

随着人机交互技术的不断发展，交互式数字人已从依赖单一模态的早期形态，逐步演进为融合多

模态信息的复杂体系。为系统呈现这一发展路径，本节将从单模态与多模态两个维度进行梳理与比较。

1.2.1 基于单模态的交互式数字人

基于单模态的交互式数字人，通常依赖单一的输入或输出通道（如文本、语音或视觉）实现与用户的交流，在人机交互的发展历程中扮演了重要的角色。最早的文本对话体强调“回应相关性”，当系统能在语义上与用户需求保持高度贴合时，会提升用户对社会临场感与拟人性的感知，即使没有外显的物理形体或数字形象，也能凭借语言本身营造出一定的社会性互动氛围。进一步地，文本风格的可控化使人格塑造成为可能，例如通过操控措辞稳定地注入“宜人性”等人格特质，并且用户对高宜人性的文本代理表现出更强偏好，提示人格匹配是提升单模态交互体验的有效路径（Völkel 和 Kaya，2021）。在视觉单模态方向上，生成与重建技术则推动了外观表现的跃升，已有方法能够仅凭单幅图像结合文本提示重建出高保真的3D人体网格，准确保留面部特征，为后续叠加语音驱动或表情控制提供了坚实外观基座。与此同时，工程实践也形成了模块化与可插拔的范式，即便在单模态条件下，识别和理解等环节依然保持分层设计，便于系统扩展与迭代。总体而言，单模态交互的优势在于成本低、部署快、可解释性强，并能通过人格化或逼真外观实现初步的沉浸感与信任感。然而，其局限亦十分突出：语言或语音虽然可提升临场感，但在复杂语境下难以替代表情与手势等更丰富的信号；人格表达的粒度有限，实现情境化自适应仍具挑战；合成语音的不自然与节奏失衡亦可能削弱互动的真实感；而视觉重建虽然能达到高保真，但缺乏语义与情感的闭环，单凭外观仍难以实现自然交流（Sonlu 等，2025）。

1.2.2 基于多模态的交互式数字人

多模态交互式数字人是人工智能与人机交互融合发展的重要方向，其核心特征在于能够综合运用语音、视觉、文本和姿态等多种模态，实现自然、真实且沉浸式的交流体验。传统的数字人多依赖人工建模与渲染，制作周期长、交互性有限，而随着深度学习与生成模型的突破，研究者开始探索通过多模态融合实现更高效地生成与更自然的交互（Zhou 等，2023）。在此背景下，学界提出了“数字孪生角色”的理念，强调不仅要再现外在的外貌与语音，还要捕捉人物的性格特征和行为习惯（Xuanyuan 等，2026），从而实现更具一致性和真实性的互动体验。这一思路伴随着大规模多模态语料的构建而快速发展，为个性化与情境化的数字人建模奠定了数据基础。与此同时，3D 建模与编辑的需求日益增长，多模态输入（如文本描述、参考图像与草图）被引入到数字人生成中，使非专业用户也能通过自然语言或直观操作完成 3D 人物的构建和修改（Hu 等，2024）。这种方式显著降低了使用门槛，并扩展了数字人在虚拟服饰、娱乐和沉浸式媒体中的应用空间。随着交互复杂度的提高，研究逐渐从“生成真实感”转向“实现智能化”。其中，强化学习结合人类反馈的方法被用于优化数字人的多模态交互能力，使其在语言对答、动作执行和环境适应上更接近人类表现。这不仅提升了系统的鲁棒性，也推动数字人从被动的展示者向主动的交互体演进（Abramson 等，2022）。近年来，实时性和沉浸感成为新的研究重点，学者们尝试通过多模态条件控制与高效的视频生成技术，使数字人在连续对话和即时场景中能够保持表情、动作和语音的协调一致，从而满足实时交流的需求（Chen 等，2025a）。更进一步，多感官交互（Sheremetieva 等，2023）的探索逐渐兴起，与传统的多模态交互主要关注视觉、听觉等感知方式不同，多感官交互强调通过整合触觉、嗅觉等多种感官体验，提升沉浸感。例如，将手势识别与触觉反馈结合，突破了传统的听觉与视觉局限，使用户能够在公共空间或虚拟环境中获得更加直观和沉浸的体验。

总体来看，单模态交互式数字人奠定了早期人机交流的技术基础，但其在表达力与适应性上存在明显局限；多模态交互的出现则为数字人赋予了更真实的感知与表达能力，使交互更趋自然和沉浸。随着技术不断进步，研究重点正逐步从外观与感官的逼真转向智能化与个性化，这不仅拓展了数字人的应用场景，也标志着其正朝着“具备社会性与主动性”的方向演进。

2 交互式数字人的关键技术

本节聚焦于多模态实时交互式数字人的核心技术，围绕建模、实时驱动与实时渲染 3 个关键环节展开分析。通过系统梳理 2D 与 3D 建模方法、文本/语音/视频/多模态驱动机制以及基于物理与神经网络

的渲染方案，旨在揭示数字人从外观构建到动态呈现的完整技术链条，为后续研究与应用提供理论支撑与实践参考。

2.1 建模技术

建模构成了虚拟角色生成过程的核心环节，其目标在于依据给定的输入数据或预设的设计参数，高精度地还原虚拟人物的视觉外观与动态特性。当前的主流方法可大致划分为2D与3D两大类别。表1总结了代表性建模方法的对比。

2.1.1 3D建模技术

3D建模是指利用计算机图形学与数学方法，在3D空间中构建具有几何结构和外观属性的虚拟对象或角色的过程。它不仅关注形体的精确重建，还涉及纹理、光照与动态特性的表现，为数字人提供逼真的视觉与交互基础。

1）在头部建模方面。Bao等人（2022）提出两阶段帧选择与可微渲染优化的3D可变形人脸模型（3D morphable face model，3DMM）拟合方法，该方法能够利用多视图数据恢复更精细的几何结构，捕捉腮帮轮廓、毛发与皱纹等细粒度特征。典型的3DMM参数化模型可表示为

$$s = \bar{s} + S\alpha, \quad t = \bar{t} + T\beta \tag{1}$$

式中，s为人脸几何，$\bar{s}$为均值形状，S为形状基，α为形状参数；t为纹理，$\bar{t}$为均值纹理，T为纹理基，β为纹理参数。

表1 数字人建模技术分类对比

Table 1 Taxonomy and comparison of digital-human modeling techniques

类型	方法/技术	核心思想	优势	局限
3D建模	3DMM参数化模型（Bao等，2022）	基于均值形状 + 形状/纹理基线性组合，拟合人脸几何与纹理	能恢复细粒度特征（腮帮、皱纹等）；较成熟	依赖多视角或RGB-D数据，非端到端
	NeRF/隐式表征（Hong等，2022）	用连续函数表示场景（体积积分）	能捕捉复杂细节（牙缝、胡须）	训练与推理开销大
	高斯溅射/混合方法（Shao等，2024）	显式网格运动控制 + 隐式高斯细节渲染	> 300 帧/s；移动端可运行	存在时间一致性等挑战
	多视图姿态建模（Ye等，2022）	将多视图3D体素投影至2D平面，用卷积神经网络推理	多人姿态实时恢复，鲁棒性高	依赖多视图
	单图像生成式管道（Kim等，2025）	单图像生成UV贴图 + SMPL-X骨架	实现端到端，支持Unity实时驱动	姿态复杂时几何一致性不足
	跨模态建模（Xue等，2021）	利用毫米波信号重建动态网格	弱光/隐私保护优势	分辨率受限
	服装与细节建模（Zielonka等，2025）	高斯原语嵌入四面体笼结构，驱动服装层分解	细节与动画兼容性优越	方法复杂
2D建模	文化定制Live2D（Safitri和Wibawa，2022）	face-tracking + Live2D参数化	支持直播互动与文化定制	表达力有限
	GAN（StyleGAN）（Melnik等，2024）	解耦潜在空间 + 风格迁移	高质量语义编辑，身份保持	训练成本高
	扩散模型 + 文本生成（He等，2025a）	根据自然语言生成可动画人物角色	可控性强，支持复杂属性描述	缺乏真实3D几何

这类基于参数模型的方案虽然并非端到端，但在细节刻画上优于传统单目方法。然而这类方法依旧受限于“线性子空间 + 显式网格”的范式：一方面，其高精度拟合依赖多视角或RGB-D（red，green，blue-depth）数据，难以直接迁移到仅有单目RGB的移动端场景；另一方面，3DMM对训练数据分布较为敏感。相比基于显式网格的3DMM，NeRF及其变体提供了纯隐式建模的路径。隐式表征是指通过连续函数表示场景或物体，而非直接生成显式网格或点云。典型代表包括NeRF、SDF（signed distance field）以及近年来的高斯溅射。例如，Hong等人（2022）提出的HeadNeRF将NeRF融合到参数化头部模型中，用以

替代纹理网格进行高保真渲染，并支持身份、表情和光照的语义解耦。该端到端的方式在捕捉细节（如牙缝与胡须）方面表现突出。但从驱动和系统集成角度看，HeadNeRF这类隐式辐射场的优势与短板也较为突出：与3DMM相比，它在视角自由度和光照建模上更具优势，却缺乏天然的显式拓扑结构，若要与现有面部骨骼/BlendShape管线对接，仍需额外网格提取或场到骨骼的绑定步骤。此外，NeRF体渲染本身的体积积分开销较大，训练与推理的能耗不易满足大规模在线服务的需求，这也促使近期研究向更高效的隐式—显式混合设计演进。SplattingAvatar（Shao等，2024）正是这一趋势下的代表性工作，它结合了显式网格的运动控制优势与隐式高斯溅射的细节渲染能力，如图2所示。

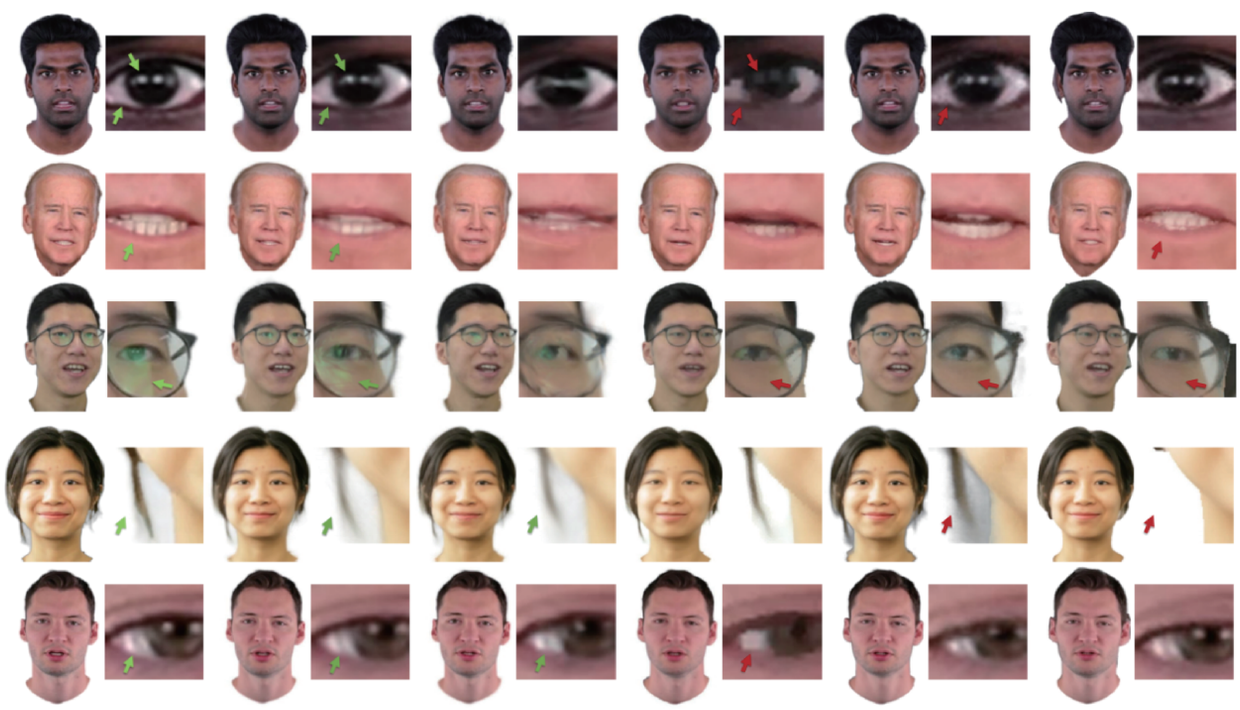


图2　头部化身的定性比较（Shao等，2024）

Fig. 2　Qualitative comparison on head avatars（Shao et al.，2024）

与依赖多视角数据的3DMM方法不同，SplattingAvatar仅需单目视频即可训练；与计算开销巨大的纯NeRF方法相比，它通过将高斯原语（以3D高斯函数建模的可渲染基本单元，包含位置、协方差、颜色和不透明度等属性）“嵌入”到驱动网格上，实现了超过300帧/s的实时渲染性能（RTX 3090）和移动端30帧/s的流畅体验。值得注意的是，NeRF渲染的核心思想在于对光线上采样点进行体积积分，具体计算为

$$C(\boldsymbol{r})=\int_{t_n}^{t_f}T(t)\sigma\big(\boldsymbol{r}(t)\big)c(\boldsymbol{r}(t),d)\mathrm{d}t$$

$$T(t)=\exp\left(-\int_{t_n}^{t}\sigma\big(\boldsymbol{r}(s)\big)\mathrm{d}s\right)\tag{2}$$

式中，$\boldsymbol{r}(t)$是光线采样点，σ表示体密度，c表示颜色，T为透射率。在高斯溅射中，单个高斯原语可写为

$$G(\boldsymbol{x})=\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)\tag{3}$$

式中，$\boldsymbol{\mu}$为高斯中心，$\boldsymbol{\Sigma}$为协方差矩阵，$\boldsymbol{x}$为空间点。式（2）给出了NeRF中基于体积渲染的经典表达：沿光线对空间中连续采样的点进行积分，以合成最终颜色。该方法虽然能实现高质量的视图合成，但其依赖密集采样和多层感知机（multilayer perceptron，MLP）隐式建模，计算开销大，且在高频几何细节（如发丝、睫毛或镜面反射）上的表达能力受限。相比之下，式（3）描述了高斯溅射中单个高斯原语的显式表示。该表达以显式、局部化的方式直接建模几何与外观，无需沿光线积分，显著提升了渲染效率；同时，其对局部曲率和尺度的灵活控制能力，使其能更精确地刻画高频细节。

2)在身体姿态建模方面。传统方法依赖多视图几何与深度融合,虽然具有较高的几何精度,但数据采集与处理复杂度高。Ye等人(2022)提出的Faster VoxelPose是典型的多视图融合管道:该方法将多视图3D体素特征通过正交投影映射到2D平面,并利用高效的2D-CNN(2D convolutional neural network)进行姿态推理,从而恢复出完整的3D姿态。其关键监督机制可以通过投影热图损失函数表述为

$$L_{2\mathrm{d}}=\sum_{i=1}^{L}\sum_{j=1}^{W}\left\|\boldsymbol{H}_{i,j}^{(xy)}-\hat{\boldsymbol{H}}_{i,j}^{(xy)}\right\|^{2} \tag{4}$$

式中,$\boldsymbol{H}^{(xy)}$为真实的2D热图,$\hat{\boldsymbol{H}}^{(xy)}$为预测结果,用于约束3D姿态在$xy$平面的投影一致性。该方法在多人姿态估计上具备实时优势,但依赖多视图输入。相比之下,基于单图像输入的生成式AI管道则属于端到端方法,此类方法降低了对硬件部署和数据采集的门槛,它通过合成额外视图和SMPL-X(skinned multi-person linear model expressive)兼容的UV贴图,实现解剖精确的光真实重建,并可在Unity中结合BlazePose(Kim等,2025)实现实时运动同步。然而,此类方法在复杂姿态变化下仍存在几何一致性不足的问题。

3)在运动与动态建模方面。跨模态感知成为新突破口。Xue等人(2021)提出的mmMesh利用毫米波信号重建人体动态网格,平均顶点误差仅为2.47 cm,能够实现实时建模。与视觉方法相比,这类基于传感的方案在弱光环境和隐私保护上具有优势,但受限于毫米波的分辨率和抗噪性能。

4)在服装细节建模方面。Drivable 3D Gaussian Avatars(Zielonka等,2025)将高斯原语嵌入四面体笼结构,解耦低维驱动姿态与高维细节参数,实现了面部表情与身体/服装的独立控制,并支持服装层分解,在逼真度和动画兼容性上,该方法均优于现有方案。

总体而言,无论是两阶段参数优化、端到端生成管道,还是显隐式混合方法,与传统手工建模方法相比,这些研究工作都显著降低了数字人建模的门槛。但仍面临挑战:隐式表示的计算开销、跨模态融合的泛化能力以及动态实时交互场景下的时间一致性等。未来研究方向可聚焦于神经—几何混合框架,以平衡效率与保真度,并探索多模态先验(如毫米波+RGB)驱动的自监督学习,以推动交互式数字人向更普适的实时应用演进。

2.1.2 2D建模技术

相较于复杂的3D建模,2D建模因其计算效率高、实现便捷,在实时交互式数字人中仍然扮演着重要补充角色。早期研究多依赖镜头帧的重排序,或利用配音演员构建的3D模型来合成训练集中未出现的语音对应的嘴型。这类方法在一定程度上奠定了研究基础,但由于音频与面部关键点(landmarks)之间的跨模态映射高度复杂,往往导致背景模糊、面部特征不自然等问题(Pham等,2024)。为克服上述局限,近期研究提出了通过解耦音频风格与视觉特征的新方法,从而能够基于单一静态肖像和任意音频生成高保真2D动画(Pham等,2024)。

同时,开源工具的广泛应用推动了2D建模的发展。研究者通常以Krita(Hietamies,2024)等栅格绘图软件为起点,在借助OpenToonz、Blender或Godot等平台中进一步开展骨骼绑定与动画测试,从而高效完成角色原型的验证。

Live2D技术作为一种轻量、高效、低开销的解决方案,因其对移动端与实时交互场景的高度适配而备受关注。其核心思想是通过分层2D图形模拟类3D的几何变形,从而在不依赖完整3D重建的情况下实现逼真动态。近期研究引入3D人脸建模中的"形状基"思想,He等人(2025b)首次将3DMM的线性组合思想迁移至Live2D域,提出适用于2.5D角色的混合形状(blendshape)参数化框架。图3为Live2D面部纹理绑定,具体而言,他们为每个语义组件(眉、眼、鼻、口)独立定义3条正交控制轴:水平位移x、垂直位移y和缩放scale,参数区间统一归一化至[-30,30],从而构成一组可微的局部仿射基。整体面部配置被建模为基模与线性加权和,具体为

$$\boldsymbol{F}=\boldsymbol{F}_0+\boldsymbol{\Sigma}_i\boldsymbol{\omega}_i\boldsymbol{B}_i,\ \ \boldsymbol{\omega}_i=\left[\omega_i^x,\omega_i^y,\omega_i^z\right] \tag{5}$$

式中,$\boldsymbol{F}_0$表示中性模板,$\boldsymbol{B}_i$为组件i的基,$\boldsymbol{\omega}_i$为推断权重。为了从单幅肖像快速估计$\boldsymbol{\omega}$,该研究构建了一个包含10万对〈渲染图,参数〉的合成数据集,并利用4层MLP在面部真值与混合形状系数之间学习确定性映射。推理阶段仅需检测2D关键点,即可在约30 s内端到端地生成身份一致、动态无伪影的Live2D模型,生成速度显著优于以小时计的传统手工rigging或prompt驱动流程。

在应用层面,Saraswati等人(2023)验证了基于face-tracking的Live2D驱动链路:通过VTube Studio

捕捉摄像头面部动作，并实时映射至2D网格变形，实现低延迟直播互动；Safitri和Wibawa（2022）则成功地将西巴布亚文化元素（如Cenderawasih鸟纹样、Melanesoid肤色和传统身体彩绘）融入了基于Live2D Cubism的参数化虚拟YouTuber角色（Erum Jaelynn）设计中，为非主流文化的虚拟形象创作提供了一个详尽且可行的技术范例。

GAN的引入进一步扩展了2D建模的能力。StyleGAN通过解耦潜在空间实现了高质量的语义编辑。其在FFHQ或CelebA-HQ等大规模人脸数据集的StyleGAN变体不仅支持风格混合与身份保持，还能结合对比语言—图像预训练（contrastive language-image pretraining，CLIP）实现文本驱动的人脸编辑，在图像增强与修复等任务中展现出可观的潜力（Melnik等，2024）。最近，基于扩散模型的文本条件生成逐渐成为新的研究方向。该类方法能够根据自然语言提示生成可动画化的2D角色，并通过解析复杂文本输入提取，如发型、眼睛颜色、服饰等属性，再结合轮廓引导的图像生成完成细节补全。进一步地，研究者将其与Live2D及ARKit（He等，2025b）混合形状结合，使其在低资源平台上也能实现高效、实时的数字人建模。



图3　Live2D面部纹理绑定（He等，2025b）

Fig. 3　Binding of Live2D facial texture（He et al.，2025b）（(a)facial feature texture and binding effect；(b)lower level facial effect）

此外，2D与3D建模技术在数字人建模中的互补性日益凸显。尽管2D建模在计算效率和实时交互上具有显著优势，但其缺乏3D几何信息使其在复杂姿态和视角变化下的表现有限。而3D建模能够提供更细致的几何结构与动态效果，尤其适用于高质量的图形渲染。未来的发展可能会倾向于将两者结合，通过互补优势解决各自的局限性。2D建模可以在低资源环境下快速渲染并进行实时交互；而3D建模则能够提供更精确的细节和真实感。这样的混合方案有望在效率与保真度之间找到平衡，推动数字人技术向更广泛的实时应用领域发展。

2.2　实时驱动技术

在多模态实时交互数字人系统中，实时驱动是实现自然交互的核心，其本质是在严格时间约束下，将多源输入信号映射为数字人的多模态表现（潘烨等，2025）。该过程不仅需要保证动作、表情与姿态在感知上的连续性，还要求计算架构在低延迟与高吞吐量之间实现平衡，从而支持沉浸式与高响应性的交互体验。为便于比较，表2对不同驱动路径的优劣与适配场景进行了横向比较。

本节将综述基于单模态与多模态的实时驱动方法，分析其技术路径、优势与局限，并探讨未来发展方向。

2.2.1　基于单模态的实时驱动

单模态实时驱动因架构简洁、实现清晰，在数字人早期研究中占据主导地位。其典型输入包括文本、音频或视觉，通过模态内映射生成相应的动作、表情或姿态。

1）文本驱动。文本驱动方法的挑战主要在于语义解耦、跨模态对齐与实时生成。例如，Liu等人（2025b）提出的VividTalker框架利用轻量化模型与跨模态提示缓解语义漂移，但在开放域语义泛化方面仍受限。相比之下，ChatPLUG（Tian等，2023）借助互联网增强的指令微调，将知识检索与人格建模融入对话生成，实现了上下文感知的个性化回复，但在模型规模与响应延迟之间仍存在权衡。在非语言行为生成方面，GestureCLR（Ali等，2025）通过对比学习将文本与手势映射至统一语义空间，提升了语义一致性，但个性化与风格化控制仍不足。另一类代表性工作TeCH（Huang等，2024）则结合单幅图像与文本提示，实现高保真3D人体建模与动态驱动，为基于文本的个性化外观编辑与表现开辟了新路径。

2）音频驱动。音频驱动方法利用语音的韵律特征（如音高、强度、时长）生成唇形同步与面部/全身动作，在实时交互中具有重要地位。早期方法多依赖规则驱动，例如，Zoric等人（2011）基于隐马尔可夫（hidden Markov model，HMM）和规则模型生成眨眼、点头等表情。该类方法计算开销低、支持实时渲

染，但难以表现复杂情感与个体风格。深度学习方法的引入提升了自然度，例如，Li等人（2022）提出的FAAN模型在唇形一致性与帧真实感方面表现优异，其生成视频在像素峰值信噪比（peak signal-to-noise ratio，PSNR）、结构相似度（structural similarity index measure，SSIM）、整体分布真实感（Frechet inception distance，FID）以及动态流畅性（光流图）等多维度指标上均优于传统方法，但此类方法多聚焦于2D面部，难以满足沉浸式交互对3D的需求。

在3D方法中，研究重点转向为情感动态与实时集成。Pan等人（2025）提出的VASA-Rig框架通过Latents2Rig模型将2D输入映射至MetaHuman rig参数，实现了40帧/s的高保真面部动画，并增强了眼部与眉部的表现力。Song等人（2025）的TalkingStyle框架则通过解耦风格、语音与运动，保留个体化说话习惯，进一步推动了数字人的个性化定制。在全身动画生成方面，Li等人（2025b）的InfinityHuman采用粗到细的生成策略与手部奖励机制，实现了长序列（> 60 s）的高质量唇同步与手势对齐，适用于持续交互场景。

近年来，情感与实时性的平衡成为研究重点，Liu等人（2024）的EmoFace将情感标签与音频联合建模，实现了3D表情的精确控制，并通过结构优化提升推理效率。与此同时，Lin等人（2025）的CyberHost则利用单阶段扩散框架与区域注意力机制，有效缓解了全身动画生成中的细节丢失与运动不稳定问题。

表2　数字人实时驱动技术对比

Table 2　Comparison of real-time digital human driving technologies

类型	参考文献	输入模态	技术特点	优势	局限
文本驱动	Liu等人（2025b） Tian等人（2023） Ali等人（2025） Huang等人（2024）	文本（+单图像）	文本到语音/图像/动作的映射，跨模态提示，个性化对话生成	架构简洁，支持语义驱动，易于集成语言模型	语义泛化有限，开放域表现不足，个性化控制弱
音频驱动	Zoric等人（2011） Li等人（2022） Pan等人（2025） Song等人（2025） Li等人（2025a） Liu等人（2024） Lin等人（2025）	语音（音高、韵律、情感）	从规则映射到深度学习、扩散与Transformer，支持面部与全身动画	自然度高，能反映语音情感，适合连续交互	计算复杂，跨文化情感泛化不足，3D实时性仍受限
视觉驱动	Chan等人（2019） Wu等人（2024a） Siarohin等人（2019） Liu等人（2023） Adiya等人（2024） Wiles等人（2018）	视频	关键点中间表示，时序一致性建模，自监督关键点学习	跨身份动作迁移效果好，泛化能力强	对输入视频质量依赖高，复杂交互中稳定性不足
多模态驱动	Zhou等人（2023） Saffaryazdi等人（2026） Chen等人（2025a） Li等人（2025a） Li等人（2024）	文本、语音、图像、生理信号…	多模态融合建模，扩散/自回归生成，大模型in-context推理	融合多信号，表现自然，支持毫秒级响应，具备预测与主动性	架构复杂，计算开销大，轻量化与可部署性仍是瓶颈

总体来看，音频驱动方法已从规则映射发展至基于扩散与Transformer的生成范式，但在实时性优化、跨文化情感泛化与轻量化部署方面仍面临挑战。

在视觉驱动方面。基于视觉的实时驱动通过提取视频流中的动作、表情或姿态信息，驱动数字人实现自然交互体验。在视觉驱动的实时方法中，Chan等人（2019）提出利用关键点作为中间表示，将复杂舞蹈动作迁移到目标人物身上，并通过预测相邻帧来提升空间—时序一致性，方法简洁且效果显著。Wu等人（2024a）则通过Gromov-Wasserstein损失缓解训练数据稀缺的限制，并引入记忆模块以增强生成细节和整体质量。尽管上述基于关键点的方法取得了进展，但在大幅度姿态变化下往往难以保持全局结构。为此，Siarohin等人（2019）采用自监督关键

点学习与局部仿射变换，提升了对任意对象的泛化能力。在此基础上，Liu 等人（2023）提出 Human MotionFormer，结合 Transformer 的全局依赖建模与卷积的局部感知，有效解决了大范围动作匹配不足的问题。

在时间一致性建模方面。Adiya 等人（2024）双向时序扩散机制缓解了单向生成常见的纹理漂移，使生成序列在长时间范围内更加稳定。在人脸驱动方向，Wiles 等人（2018）的 X2Face 则提供了轻量化的自监督方案，并可扩展至音频或姿态编码等模态，展现了跨模态拓展的潜力。整体来看，视觉驱动方法在跨身份动作迁移与时序一致性上取得了相关进展，但仍依赖输入视频质量与动作幅度，在复杂交互场景中存在局限。

尽管单模态实时驱动系统因其架构简洁、实现清晰，在数字人早期研究与实践中占据主导地位，但其也存在显著的局限性。依赖单一输入模态（如文本、音频或图像）容易导致上下文理解的不足，且在复杂场景下表现欠佳。例如，用户输入的语义模糊或缺乏细节的内容会严重影响语言模型，而现有音频驱动的方法跨文化跨语言的内容与情感多样性处理上仍存在困难，此外大幅度姿态变化或长时间动作序列仍然会影响视觉驱动系统对一致性的保持。这些局限性表明，单模态驱动系统无法满足更复杂、动态和个性化交互的需求，从而为多模态系统的发展铺平了道路。

2.2.2　基于多模态的实时驱动

随着人工智能与大规模生成模型的发展，如何在多模态输入下实现低延迟、强语境感知的自然驱动成为关键挑战。现有工作大致沿着两条路线演进：一类通过工程化的多模型流水线，将文本、语音和图像等模态在系统层面“拼接”起来；另一类则尝试在统一潜空间内进行端到端建模，以减小模态间切换开销并提升时序一致性。总体来看，如何在延迟、表现力、可控性与部署成本之间取得平衡，是当前多模态实时驱动面临的核心问题。

在多模态驱动中，统一建模输入信号是首要问题。Zhou 等人（2023）提出的多模态融合数字人系统通过文本、语音与图像的联合建模，实现了从预处理到驱动再到后处理的全流程。该系统采用公式化的建模方式，例如语音合成过程可表示为

$$\boldsymbol{S} = G_s(\boldsymbol{R}) \| G_c(\boldsymbol{R}, \boldsymbol{F}) \tag{6}$$

式中，G_s 表示文本到语音转换（text-to-speech，TTS），G_c 表示基于目标声纹特征 $\boldsymbol{F}$ 的语音克隆，$\boldsymbol{R}$ 为语言模型的输出文本。该方案的优势在于工程上易落地：各模块可以独立替换升级，便于在现有产业链中快速集成。与传统依赖人工建模的数字人流程相比，其多源信号并行融合方法显著降低了开发与生成成本。但从“统一建模”的角度看，这一系统仍主要在特征级进行模态拼接，缺乏在单一潜空间内对跨模态相关性的联合建模，也未对端到端时延和复杂场景鲁棒性给出系统性量化，因此在严格的毫秒级交互场景下，仍然更多属于模块化流水线方案而非真正的统一多模态模型。

另一方面，Saffaryazdi 等人（2026）的研究强调引入神经与生理信号作为感知模态，以提升实时情感识别的准确性。他们提出利用 EEG（electroencephalogram）、EDA（electrodermal activity）与 PPG（photoplethysmography）等低延迟生理特征，并通过加权决策融合。具体为

$$p_o^x = a \cdot p_{\mathrm{EEG}}^x + b \cdot p_{\mathrm{EDA}}^x + c \cdot p_{\mathrm{PPG}}^x \tag{7}$$

式中，$x \in \{Arousal, Valence\}$，$x$ 为待识别的情绪维度，*Arousal* 为唤醒度，反映由平静到兴奋的情绪激活强度，*Valence* 为效价，反映情绪由消极到积极的愉悦程度，a，b，c 根据模态可靠性设定。这种跨模态加权机制保证了实时情绪驱动的鲁棒性，为数字人在交互中展现出“共情”反应提供了可能。

在生成端，实现毫秒级推理是保证实时性的关键。Chen 等人（2025a）提出的 MIDAS（multimodal interactive digital human synthesis）框架将大语言模型与自回归视频生成结合，通过深度压缩自编码器实现 64×的空间压缩，有效缓解了长时序生成的延迟。其生成过程遵循自回归条件概率建模，具体为

$$p(\boldsymbol{C}, \boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N) = \prod_{i=1}^{N} p\left(\boldsymbol{x}_i \middle| \boldsymbol{C}, \boldsymbol{x}_1, \cdots, \boldsymbol{x}_{i-1}\right) \tag{8}$$

式中，$\boldsymbol{C}$ 表示多模态条件（如音频、姿态、文本），$\boldsymbol{x}_i$ 为第 i 帧视频的潜在表示。结合扩散解码头，该框架在保证高质量合成的同时支持流式推理，使数字人可实现毫秒级响应。相比之下，Li 等人（2025a）提出的 MoDA 框架利用多模态扩散和粗到细的融合策略，有效解决了语音、身份与表情之间的不一致问题，如图 4 所示，MoDA 框架能够根据一幅参考图像、一段音频及其他控制信号，实时生成高保真的会说话头

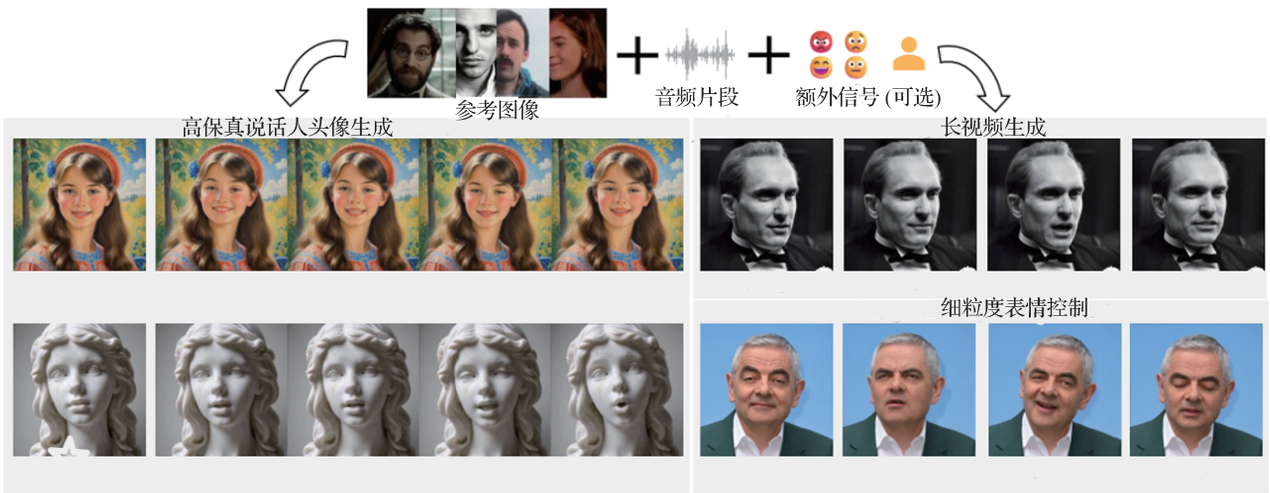


图4　MoDA(Li等,2025a)框架概览

Fig. 4　The overview of the MoDA framework(Li et al., 2025a)

像视频,不仅能精确同步唇形,还能生成丰富多样的面部表情和自然的头部运动,并支持细粒度的表情控制和长时间视频合成。其联合参数空间与整合式Transformer融合机制,使得系统能够在实时对话场景下保持唇形精确与面部动态自然。

近年来,多模态大模型成为实时驱动的核心。Sun等人(2024)提出的Emu2模型展现出强大的in-context学习能力,能够在少样本场景中快速适配新的任务。其统一的自回归训练目标为

$$L=-\sum_{t=1}^{T}\log P\left(\boldsymbol{y}_t\left|\boldsymbol{y}_{<t},\boldsymbol{x}\right.\right) \tag{9}$$

式中,$\boldsymbol{x}$为输入的多模态序列,$\boldsymbol{y}_t$为下一个预测元素(文本或视觉token)。该机制保证了模型可在对话中根据上下文进行实时调整,使数字人表现出类人般的即兴反应。同时,OpenAI的GPT-4o(generative pre-trained transformer 4 omni)在实际部署中展现了232 ms的语音响应延迟,接近人类对话反应时间。这表明大模型在处理多模态输入与生成时已具备支持实时数字人交互的能力。更重要的是,GPT-4o的单一模型架构能同时处理文本、音频与视频,避免了多模型切换带来的延迟瓶颈。

与此同时,为了让数字人自身的交互更加流畅自然,RITA(real-time interactive talking avatar)框架(Cheng等,2024)提供了一种集成化的实时动画生成路径。它结合生成模型与LLM(用于生成对话内容),通过两阶段的"基帧生成+动态帧匹配"并辅以实时插帧技术,显著降低了语音驱动面部动画的延迟,同时保证了视觉流畅性。实验证明,其视频生成延迟可控在亚秒级范围内,使数字人能够与用户进行真正的实时对话。

除了交互的流畅自然性,预测用户的后续动作也是多模态驱动的重要研究方向。不同于常规的模态驱动系统,Li等人(2024)提出的OmniActions系统利用LLM对现实世界的多模态感知进行结构化推理,并预测用户可能的后续操作。OmniActions引入了一种新的思路,通过链式推理,实时分析用户在增强现实或其他交互环境中的行为模式,生成可能的动作(如搜索、保存、分享等)。这一系统不仅使数字人在交互过程中能够迅速响应,还能主动预测并理解用户需求,从而提供更加个性化和智能化的辅助支持。

综上所述,基于多模态的实时驱动技术正在经历从单一模态驱动到跨模态深度融合,从离线生成到低延迟流式生成,从简单回应到上下文感知与预测行动的转变。未来的发展趋势将集中于以下3个方向:1)通过轻量化表示与自回归架构进一步优化延迟;2)利用大模型的上下文推理能力实现更自然的对话与动作预测;3)构建具备情感与环境感知的数字人,以支撑医疗、教育和娱乐等高实时性场景。

2.3　实时渲染技术

交互式数字人的实时渲染是多模态交互系统的核心技术之一,目标是在低延迟下呈现高真实感的虚拟角色视觉效果。同时,实时渲染需要在图像质量、计算效率与交互性之间取得平衡。本节将重点探讨两种主要实时渲染方法:基于物理的渲染和基于神经网络的渲染。

2.3.1　基于物理的渲染方法

基于物理的渲染(physically-based rendering, PBR)在多模态交互式数字人研究中被视为实现高真实感与一致性的重要技术路径。然而,不同方法在几何重建、材质建模和实时性之间仍存在明显的权衡。Hong 等人(2025)提出的 BEAM(bridging physically-based rendering and gaussian modeling for relightable volumetric video)框架通过结合四维(4D)高斯表示与 PBR,有效解决了动态体积视频的几何一致性和材质属性解耦问题,并利用高斯光线追踪实现环境光遮蔽与基色的高效估计。该方法在真实感与可重光照能力上取得突破,但流程依赖复杂的多阶段优化,整体计算成本较高,限制了其在资源受限环境中的应用。相比之下,Qin 等人(2025)提出的 GauFace 表示及其生成器 TransGS 试图在传统计算机图形学面部资产与高斯点渲染之间建立桥梁,从而在移动端实现接近离线渲染质量的实时表现。然而,该方法高度依赖预训练的高质量面部资产与环境光照数据,其在非标准输入条件下仍需进一步验证。

在硬件层面,Guo 等人(2025b)提出的 Mobile-PBR 采用定制化的 28 nm 芯片架构,将逆向渲染与物理光线追踪结合,并采用背景聚类与混合精度处理单元,降低了移动端的计算与存储负担。该方案提升了能效比与实时性,但依赖特定硬件实现,灵活性不足,难以适配通用平台。在数字人资产生成方面,Wang 等人(2025)提出的 PBRGAN 模型则摆脱了对 Light Stage 的依赖,可直接基于自然场景图像生成包含反照率、粗糙度和法线的多通道 PBR 纹理,并引入文本引导以提升可控性。该方法提升了生成效率与多样性,但在细节一致性和高频结构的物理准确性上仍存在不足,尤其在复杂光照场景下容易产生伪影。另外,Rossoni 等人(2024)将 PBR 引入动态点云渲染,通过在点云格式中扩展法线、粗糙度和金属性等材质参数并结合 360° 环境贴图,使点云在不同光照和背景下能够产生与周围场景一致的高光、阴影和环境反射,相比仅依赖采集时颜色的传统渲染方式,在 XR(extended reality)场景中的视觉融合度和主观真实感更优。

2.3.2　基于神经网络的渲染方法

随着神经网络渲染技术的发展,其在数字人生成中的应用提升了渲染的真实感和质量(晏轶超等,2023)。相较于基于物理的渲染方法,神经渲染无须依赖复杂的物理建模,能够从多视角图像中自动学习场景的几何与外观,并在任意视角与光照条件下合成真实感图像。

Mildenhall 等人(2022)提出的 NeRF 将场景建模为连续的五维(5D)辐射场函数,以空间位置和视角方向为输入,输出密度与颜色,并通过体渲染实现新视角合成。该方法首次证明,仅依赖图像监督即可恢复高质量的 3D 几何与外观,避免了显式 3D 重建。然而,其基于 MLP 的网络结构对高频细节表达能力有限,且每次渲染需对每条光线进行数百次网络查询,导致渲染效率低下,难以满足实时交互需求。

为实现动态人体建模,Peng 等人(2021)提出 Neural Body,将 SMPL 参数化人体模型引入 NeRF 框架,并将潜变量锚定于骨骼节点,通过观测空间到规范空间的映射实现了运动的显式控制,有效解决了动态一致性问题,为可控数字人渲染提供了范式。但其仍受限于体积采样机制,渲染延迟较高,无法实现实时性。

为进一步提升效率,研究者开始探索结构化表示与混合建模策略。Habermann 等人(2023)提出的 HDHumans 首次将显式变形网格与神经辐射场紧耦合:通过变形网格引导体采样以减少无效点,同时利用神经场反向优化网格几何,实现双向增强。该方法在 4 K 分辨率下可生成高质量图像,但渲染一帧仍需数秒,未达实时标准。Liu 等人(2021)提出的 Neural Actor 则将纹理图作为局部姿态特征,驱动规范空间中的 NeRF,支持符合 SMPL 拓扑结构的衣物建模,但其体积采样机制依然制约实时性能。

真正推动神经渲染迈向实时的是基于 3D 高斯溅射的方法。Kerbl 等人(2023)提出的 3D-GS(3D Gaussian splatting)使用显式高斯椭球表示场景,并结合可微分光栅化,实现了百帧每秒的渲染速度。在此基础上,Pang 等人(2024)提出 ASH(animatable Gaussian splats),首次将高斯溅射应用于可驱动人体建模,如图 5 所示。ASH 将高斯参数嵌入变形模板的 UV 空间,通过 2D 卷积网络学习姿态到高斯属性的映射,实现 30 帧/s 以上的实时渲染。在新视角与新姿态合成任务中的实时性和渲染质量显著优于 DDC 和 HDHumans 等方法。此外,其 UV 参数化策略支持宽松衣物的建模,成为当前实时神经人体渲染的 SOTA(state-of-the-art)方法。

除姿态与视角控制外,光照可控性亦是交互式

数字人系统的核心需求。传统NeRF方法将光照与材质耦合于辐射场中，难以实现重打光。Chen和Liu(2022)提出的Relighting4D首次从普通视频中实现动态人体的新视角重打光，该方法将神经场分解为法线、遮挡、漫反射与高光4个分量，并结合微表面BRDF模型，实现物理可解释的光照编辑。通过引入最小熵稀疏先验和平滑正则项，该方法保证了数字人在不同光照环境下的真实感与一致性。

整体而言，基于物理的实时渲染方法强调几何一致性和材质精确建模，在光照可控性与物理真实性方面具有显著优势，但通常依赖复杂的计算流程和专用硬件，难以兼顾通用性与效率。相比之下，基于神经网络的渲染方法通过学习多视角图像的几何与外观特征，实现了跨视角和跨光照条件下的高保真合成，具备更强的自动化与适应性，但在实时性和可控性方面仍存在挑战。随着高斯溅射等新兴技术的出现，两类方法在真实感与效率上的差距正逐渐缩小。未来的融合与互补发展，有望为多模态交互式数字人提供既具物理一致性又能高效生成的渲染路径。



图5　ASH(Pang等，2024)方法概览

Fig. 5　ASH method overview(Pang et al.，2024)

3　多模态实时交互式数字人架构

3.1　通用框架

通过对多模态交互领域的大量文献调研和相关技术的深入分析，本文总结出一种实时交互式多模态数字人的通用框架，包含感知层、融合层、生成层和拓展层，如图6所示。该分层设计基于对交互系统复杂性的解构，旨在系统性地应对实时多模态输入处理、跨模态语义整合、自然响应生成以及高效数据管理等关键挑战。

3.1.1　感知层

在多模态交互式数字人的体系结构中，感知层承担着对外部环境与人类行为信号的采集与解析任务。它是实现自然交互的前提条件，也是信息流进入融合与生成环节的起点。本节将其细分为语音感知、面部识别与表情感知、手势识别以及脑机接口4个模块来展开介绍。

1)语音感知模块。该模块在设计中不仅关注识别准确率，更强调实时处理能力。早期的语音活动检测(voice activity detection，VAD)方法由Atal和Rabiner(1976)提出，利用能量比和零交叉率等声学特征实现语音端点检测，但在噪声环境下鲁棒性较差。随后研究转向统计建模方法，如HMM、高斯混合模型(Gaussian mixture model，GMM)等，以提升在多样化场景下的稳定性。近年来，随着深度学习的发展，Zuo等人(2023)提出基于DFSMN(deep feed-forward sequential memory network)的VAD系统，并进一步发展为多任务语义VAD。为实现低延迟流式处理，他们采用RWKV模型结构，有效控制模型延迟。在特征提取方面，梅尔频率倒谱系数(Mel-frequency cepstral coefficients，MFCC)技术(Basak等，2023)自20世纪末期以来得到广泛应用，至今仍是实时语音识别的重要前端。图7展示了MFCC技术的处理流程，其中包含了语音信号的预处理、分帧、窗函数应用、高频分量的傅里叶变换以及对数运算等步骤，从而提取出语音特征。同时，麦克风阵列、降噪芯片以及专用的VAD前端(Chen等，2024)(如Silero VAD)已应用于服务机器人和工业数字人，以提升噪声环境下的鲁棒性。

进一步地，近年来国内外研究者普遍采用端到端的神经网络模型，如基于CNN、长短期记忆网络(long short-term memory，LSTM)或Transformer的流式架构，这些方法能够在几十毫秒级别的时间尺度

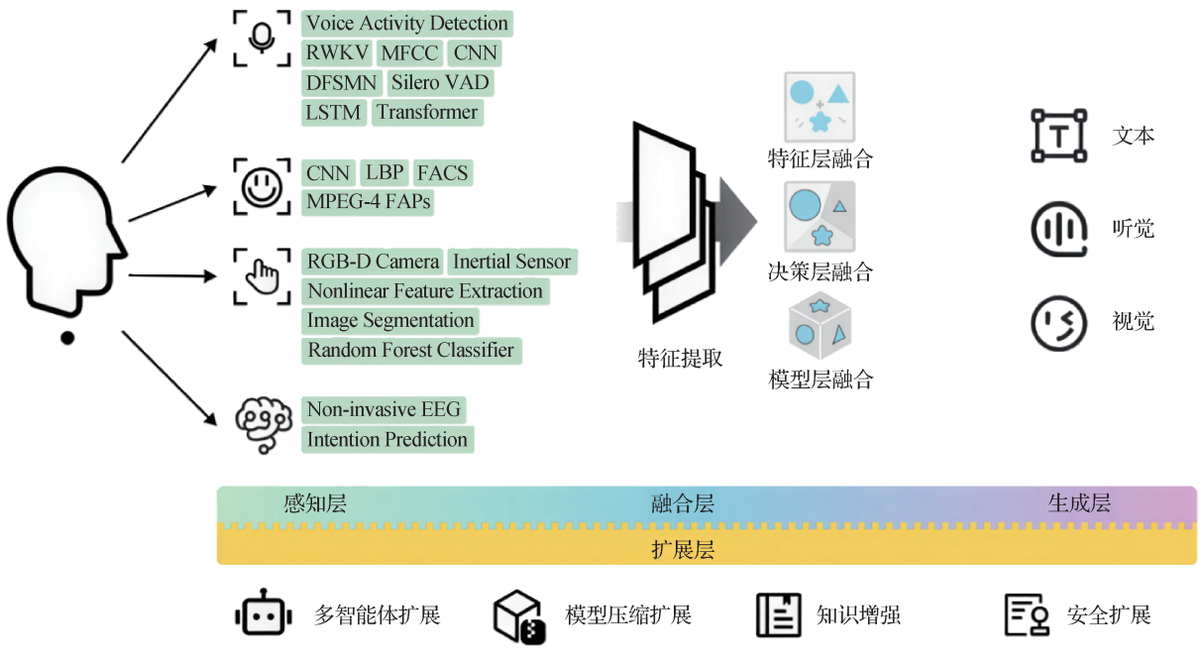


图 6　多模态实时交互式数字人框架

Fig. 6　Framework of multimodal real-time interactive digital humans

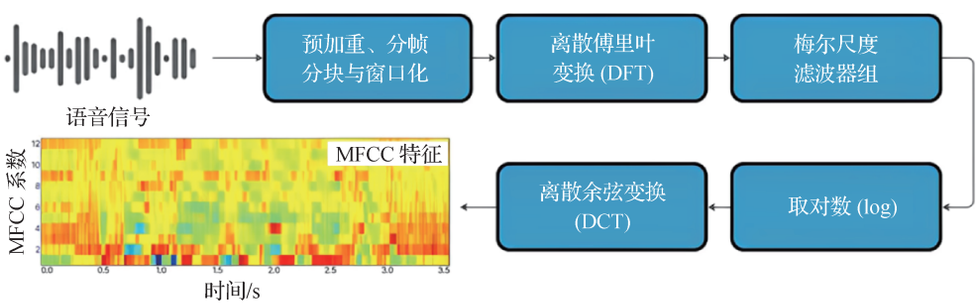


图 7　MFCC 技术流程（Basak 等，2023）

Fig. 7　MFCC technical process（Basak et al. ，2023）

内完成语音片段的转写，实现低延迟的在线识别（Yu 等，2025）。

2）面部识别与表情感知模块。该模块利用高分辨率摄像头和先进的计算机视觉技术，实时检测人脸区域并提取关键特征点。面部识别方面，3D 人脸识别通过深度相机同步采集 RGB 与深度信息，并结合 CNN 与局部二值模式（local binary patterns）特征（Dong，2022），实现复杂环境下的人脸即时建模与识别，从而在动态交互场景中保持高精度与稳定性。与此同时，表情感知通过检测面部动作单元（action units），结合 FACS（facial action coding system）或 MPEG-4 FAPs 等标准（Yang 等，2021），将用户微表情实时转化为情绪状态。在边缘计算框架下，这些计算过程被高效分布到本地设备与云端协同执行，使得表情捕捉和识别能够实时完成。

3）手势识别模块。该模块通过实时捕捉和理解人体的手势动作，实现用户与数字人的交互。例如，Fiorini 等人（2021）提出通过融合 RGB-D 相机与可穿戴惯性传感器识别由多个基本手势组成的日常生活场景。该研究证明，多模态融合能够有效提升在不同视角（如机器人位于用户侧面）下，对连续手势及其过渡过程的识别精度。与此同时，Mo 等人（2023）基于数字孪生的人机界面传感器提出了非接触式指尖运动捕捉方法，通过计算机视觉和自适应图像分割实现手指动态的实时编码与控制信号映射，在无需佩戴设备的条件下也能完成高精度手势识别。进一步地，Xue 等人（2022）的工作展示了结合指尖轨迹、接触力和肌电信号的多模态融合方案，其通过非线性特征提取与随机森林分类器，实现了对复杂手内操作的 93% 的识别率，有效支撑了高精度的动作

理解。

4)脑机接口模块。该模块为多模态数据采集提供了一种直接通向人脑活动的渠道。例如,非侵入式EEG设备可以在毫秒级时间分辨率下解析脑电信号(Zhu等,2024),借助实时信号处理和机器学习分类器,系统能够在用户尚未作出外显动作之前预测其意图或识别其情绪状态,这为数字人交互提供了超低延迟的感知通路。进一步地,结合信号数字化、特征提取与分类(Khan等,2022)的实时处理流程,脑机接口能够将神经活动映射为可操作的指令,实现与外部多媒体或交互系统的即时通信。

总的来说,该模块不仅在数字人感知层中承担着高带宽、低延迟的信息入口,也为后续融合层的多模态理解与生成层的自然交互奠定了关键基础。

3.1.2 融合层

在多模态交互式数字人系统中,融合层是实现不同模态数据整合与协同工作的关键组成部分。该层负责处理来自文本、图像、音频、视频数据和传感器数据等不同源的信息,并采用多种融合策略和技术以确保数据的一致性和互补性。

为了实现多模态信息的高效整合与协同运作,研究者通常从不同的融合层次和策略入手。根据信息交互发生的阶段与深度,现有多模态融合方法大体可分为特征级融合、决策级融合和模型级融合3类。以下将分别对这3类方法的基本原理与典型实践进行系统阐述。

1)特征级融合。是最直接的多模态融合方式,它在多模态数据输入到模型之前,将不同模态的原始数据或已提取的特征融合在一起,形成一个综合的表示作为模型的输入。常见的融合方法包括拼接、加法、乘法和双线性融合等。

(1)拼接。将多个特征向量简单拼接成一个更长的向量,保留了多模态的原始信息,但可能导致维度灾难,增加模型复杂度。

(2)加法。通过线性加权求和融合多个模态的特征,能够体现各模态的重要性,但容易造成语义信息丢失。

(3)乘法。将单模态特征向量相乘融合,能够弥补加法的语义信息丢失,但对特征的维度和尺度要求较高。

(4)双线性融合。通过张量外积和展平操作捕捉不同模态之间的交互信息,保留空间结构和语义关联。

2)决策级融合。将每个模态的独立决策结果进行数学公式规定或赋予不同结果不同的权重,得出最终的决策结果(Baltrušaitis等,2019)。常见的决策级融合策略包括投票法、加权平均法和多数投票法等。

(1)投票法。将多个模态的独立决策结果进行投票统计,选择获得最高票数的类别或结果作为最终决策。

(2)加权平均法。将不同模态的决策结果按照权重进行加权平均,得到综合的决策结果。

(3)多数投票法。根据各个模态的决策结果中出现频率最高的类别或结果进行决策。

3)模型级融合。通过在模型级别上将不同模态的特征信息进行融合,实现跨模态的信息交互和整合。基于深度学习的模型级融合方法应用广泛,主要包括多核学习方法和基于神经网络的融合方法(张虎成 等,2024)。

(1)多核学习方法。通过学习一组预定义的基本核的线性或非线性组合,将不同模态的特征映射到一个共享的语义空间中。该方法能够更好地融合异构数据,但需要大量的内存资源,且在测试阶段需要重新计算核函数的权重。

(2)基于神经网络的融合方法。利用神经网络的强大学习能力(Gao等,2020),将不同模态的数据进行融合。常见的方法包括GAN和注意力机制融合方法。GAN可以生成高质量的融合图像或视频,而注意力机制则能够动态地评估不同模态的重要性,提取互补信息。

3.1.3 生成层

生成层是实时交互式多模态数字人系统的核心组成部分,其主要功能是将融合层输出的多模态信息转化为具体的数字人表现,并确保这些表现具备实时性、自然性与沉浸感。在该层中,文本生成模块作为基础环节,承担语义表达与信息传递的关键任务;语音合成、表情/动作生成等模块则分别从听觉与视觉层面提升表达的自然度与感染力。麦拉宾(Mehrabian)的研究表明,在人际交流中,个体对他人情感与态度的理解不仅依赖语言内容,更受到语音语调与面部表情等非语言因素的影响。具体而言,语言文字所承载的信息约占7%,语音特征(如语调、节奏、音色等)占38%,而视觉表现(如面部表情、

肢体动作等）则高达 55%（Mehrabian 和 Diamond，1971）。这表明在数字人系统中，非语言模态的实时与高保真呈现对于提升交互质量具有重要意义。

1）自然语言生成模块。自然语言生成（natural language generation）模块是实现语义表达与交互逻辑的关键环节。该模块依托 LLM 的生成能力，能够在实时对话中生成语义连贯、上下文敏感的自然语言文本。Lan 等人（2023）最早提出将 ChatGPT 融入动画数字人生成，使数字人具备基于自然语言的动态对话与交互能力，突破了传统依赖模板和预设脚本的局限。这一方法提升了交互的灵活性，使数字人在面对用户输入时能够生成多样化的语言响应。在此基础上，Brito 等人（2025）进一步提出利用零样本、少样本以及微调等多种生成策略，使大模型在自然语言生成过程中能够保持一致的人格特征。这种方法不仅改善了文本生成的上下文连贯性，还使数字人在长时交互中展现出稳定的个性化语言风格。Coll 等人（2025）提出基于人类数字孪生的生成架构，将记忆检索与上下文生成机制相结合，使数字人的语言输出不仅具备连贯性，还能体现个性化的交互风格。与此同时，AIGC（artificial intelligence generated content）在自然语言生成中的应用不断拓展，使数字人能够进行更复杂的多模态对话，融合情感识别与反馈（Wang 和 Zhang，2024），从而提升语言交互的自然性与沉浸感。总体来看，自然语言生成模块正从单一的文本应答转向多维度、情境化和个性化的智能生成（Xu 和 Zhang，2024），成为推动数字人走向高拟真与高交互的关键技术引擎。

2）语音生成模块（text-to-speech，TTS）。语音生成模块是生成层中实现听觉输出的关键环节，其核心任务是将自然语言生成模块输出的文本或语义特征转化为自然、流畅且具有情感感染力的语音信号，从而实现数字人与用户之间的实时语音交互。该模块的性能直接决定了系统在语音输出层面的自然度、响应速度与沉浸感。随着深度生成模型与 LLM 的发展，语音合成技术正从基于拼接与参数化的传统方法，演进为以概率建模为核心的端到端高保真生成框架。近年来，以扩散模型（diffusion model）和流匹配（flow matching）为代表的生成机制在 TTS 中取得了进展。Diff-TTS 模型（Jeong 等，2021）通过去噪扩散过程实现从噪声到梅尔频谱的连续映射，提升了语音自然度与可控性；RapFlow-TTS（Park 等，2025）则利用一致性流匹配（consistency flow matching）在少步推理下生成高质量语音，实现了语音质量与推理速度的平衡。这类模型有效缓解了传统扩散模型推理慢、时延高的问题，为实时语音合成提供了理论与工程基础。

在实时交互场景下，低延迟与流式生成成为重要研究方向。SyncSpeech（Sheng 等，2025）使用了基于时序掩码 Transformer 的双流 TTS 架构，实现了文本输入与语音输出的同步生成，在接收极少量文本 token 后即可启动语音输出，极大降低了首包延迟（first-packet delay），使语音生成能够与上游语言模型实现真正意义上的“同步交互”。此外，为提升语音生成的泛化性与自然度，CosyVoice3（Du 等，2025）通过百万小时级多语言语料的预训练与多任务语音分词器设计，将语音识别、情感识别和语言识别等任务融入统一训练框架，实现了在多语言、跨音色条件下的零样本语音生成，并通过可微分奖励优化（differentiable reward optimization，DiffRO）机制提升了语音自然度与人类偏好一致性。

语音生成模块的技术演进呈现出“高保真、低时延、多语种”3D 融合趋势：一方面，通过扩散与流匹配机制提升语音合成的质量与稳定性；另一方面，依托大模型与多任务学习实现语义驱动的情感表达与跨语言泛化。为了更清晰地对比当前主流 TTS 模型的技术特点与性能差异，表 3 总结了近年来代表性模型的核心方法、实时因子（real-time factor，RTF）以及主要优势与局限。

3）表情/动作生成模块。在生成层中，表情/动作生成模块承担着将感知与融合层的输入信号转化为自然、真实且具有情感共鸣的数字人动态视觉表现的核心任务。第 2 节已介绍过数字人的实时驱动的表情/动作生成机制，以及低延迟对交互自然度的关键作用，因此本节将从生成层视角系统梳理技术演进与方法挑战。

已有的视觉生成方法常依赖于对数字人表情与动作的参数化与离散化描述。其思想可追溯至 MPEG-4 FAP 或 Ekman 的 FACS 等标准，将复杂表情分解为若干基础动作单元。在实现上，Blendshape（Condegni 等，2023）线性模型是此类思想的一种主流工程化方案，其基本形式为

$$\boldsymbol{f} = \boldsymbol{s}_0 + \sum_{k=1}^{n} w_k \left(\boldsymbol{s}_k - \boldsymbol{s}_0 \right) \tag{10}$$

表3 代表性语音合成模型技术对比

Table 3 Technical comparison of representative speech synthesis models

模型	方法/技术	RTF(处理时间/音频时长)	优势	局限
ConvNeXt-TTS (Okamoto等,2024)	基于ConvNeXt的Transformer-free端到端TTS与VC模型,结合WaveNeXt声码器,实现卷积结构的高效时序建模	0.05	速度快,比Transformer快约3倍;合成质量优于HiFi-GAN与JETS;模型轻量、结构统一	未引入语言建模或上下文增强;在多说话人物与情感泛化上略弱
FireRedTTS-1S (Guo等,2025a)	大规模语言模型式TTS,采用text→semantic→acoustic两阶段流匹配或多流语言模型,支持实时流式合成	0.1(流匹配)~0.3(多流语言模型)	可进行高保真实时合成;支持中英双语、50万小时级数据;支持零样本语者克隆	GPU资源需求高
NaturalSpeech (Tan等,2024)	基于变分自编码器(variational autoencoder)的端到端TTS模型,提出可微分时长预测与记忆机制,首次实现与人类语音无统计显著差异	0.013	合成质量接近人类语音;端到端消除级联误差;提升韵律与表达力	需完整输入文本,无法逐字生成
RapFlow-TTS (Park等,2025)	基于Consistency Flow Matching的快速高保真TTS,引入速度一致性约束与时间调度策略	0.031	推理速度快,仅需2步生成;自然度接近扩散模型;支持少步合成与质量平衡	模型依赖高质量对齐
SyncSpeech (Sheng等,2025)	双流低延迟TTS,基于Temporal Masked Transformer,同步生成流式语音(边接收边输出)	0.05~0.08	首包延迟极低;与大模型流式接口兼容;中英双语表现优良	依赖精确的文本—语音token对齐数据;对时长预测误差较敏感
Speaker Transfer TTS (RoBERTa + VITS) (Zhang等,2024)	基于VITS(variational inference text-to-speech)并融合RoBERTa文本特征的说话人迁移TTS,提升低资源说话人音色自适应能力	0.021	样本说话人迁移效果显著;RoBERTa增强文本韵律特征	仅支持中文
Diff-TTS (Jeong等,2021)	扩散模型式TTS,采用似然优化与加速采样,实现稳定非自回归生成	0.035	扩散模型首次应用于TTS;高稳定性与高保真;可控制音高与节奏	依赖外部对齐工具;推理仍需多次迭代
F5-TTS (Chen等,2025b)	基于Flow Matching与ConvNeXt V2的非自回归TTS,结合Diffusion Transformer与Sway Sampling策略	0.15	训练与推理均高效;自然度和流畅度高;多语言、多说话人适应性强	依赖梅尔谱表示导致序列较长;缺乏对情感等副语言特征的显式控制

式中,s_0为中性表情,s_k为目标表情,w_k为权重。此类方法结构简单,但难以表现复杂肌肉动态

随后,3D可变形人脸模型(3DMM)及其演进形态(如FLAME)成为数字人面部表情迁移的主流技术基础。通过利用大规模、高质量的数据集(如FaceWarehouse、FaceScape),这些模型实现了身份与表情的双维度参数化建模,从而提升了生成表情的几何保真度和自然度。在此基础上,研究进一步转向表情迁移,即将源主体的表情映射至目标数字人,方法包括基于几何相似的网格变换、跨维度(2D→3D)映射以及隐式神经网络驱动(Bao等,2022),其中卷积隐式表达和高斯场建模能够在保持细节的同时实现动态表情的高效合成。近年来,深度学习方法(特别是卷积神经网络CNN和循环神经网络如GRU(gated recurrent unit))(Vilchis等,2022)被广泛应用于表情生成框架中,主要用于从2D图像/视频中识别面部动作单元,并将结果实时重定向到3D角色模型上,以驱动其面部动画生成(Ou等,2023)。同时,低成本与个性化趋势逐渐凸显,如结合手机端RGB摄像与惯性测量单元(inertial measurement unit)传感器的可穿戴系统,可实时捕捉个体化表情、凝视与关节动作,实现个性化数字孪生。

表情/动作生成模块已形成从参数化模型→数字建模→表情/动作迁移→隐式神经生成的演进脉络。当前挑战主要集中在表情与身份的解耦、跨主体泛化、自然度提升与“恐怖谷效应”缓解。

3.1.4 拓展层

在多模态实时交互式数字人的总体架构中，拓展层旨在对感知、融合与生成层进行补充与延展，使数字人具备“可演化、可溯源、可协作”的高级智能。近年来，学界和业界对这一层提出了多种探索，其中多智能体系统（multi-agent system，MAS）、模型压缩（model compression）、检索增强生成（retrieval-augmented generation，RAG）与数字水印（digital watermarking）是较为典型的代表。

MAS的引入为数字人赋予了更强的协作、规划与自适应交互能力。在虚拟现实、教育培训和应急响应等复杂动态环境中，MAS可以通过多角色分工与协同决策实现任务的分布式感知与自主规划（Ospina-Bohórquez 等，2021）。这不仅使数字人从“单一交互体”演变为“协同智能体群”，还能模拟人群行为、组织多轮对话、进行动态任务分配，从而支撑更自然的群体交互和场景适应。例如，在虚拟仿真和智能培训中，多智能体的引入可生成逼真的人群互动、自动化的任务演练和多角色协同决策（Ospina-Bohórquez 等，2021），为用户提供更具沉浸感和可操作性的体验。

与此同时，模型压缩为拓展层的长期演化与普适部署提供了基础支撑。随着数字人模型日益庞大，如何在低功耗、低时延的设备上稳定运行成为关键挑战。常见方法包括剪枝、蒸馏、低秩分解与混合精度量化（黄震华 等，2022），进一步的研究还发展了轻量化架构（如 MobileNet、ShuffleNet、EfficientNet）以及神经架构搜索（neural architecture search，NAS），能够结合特定硬件特性自动搜索精度与效率的最佳折中点，从而为包括数字人在内的边缘智能应用提供有力支撑（Li 等，2023）。

除此之外，RAG通过将大模型的参数化记忆与外部非参数知识库相结合，提高了数字人回答的真实性、时效性与专业性。与传统纯生成模型相比，RAG可在生成过程中动态检索与问题高度相关的外部文档，并基于检索内容进行推理和回答，减少“幻觉”现象，提高事实一致性和可追溯性（Lewis 等，2020）。其非参数记忆模块还能灵活替换和更新，使数字人能快速吸收最新知识、适应新领域场景，尤其适用于知识快速迭代的垂直专业场景（Kim 等，2024）。此外，自适应RAG框架还可根据问题复杂度在“无检索—单步检索—多步增强”之间灵活切换，为交互系统提供兼顾效率与准确性的知识增强机制。

数字水印则从内容治理的角度保障了数字人的“可溯源”。通过在语音、图像、视频乃至3D虚拟形象中嵌入不可感知标识，有助于版权保护与篡改检测（Malanowska 等，2024）。进一步的发展还包括3D水印技术，能够在NeRF与3D高斯溅射等新型建模渲染框架下保护数字人资产（Wu 等，2024b），确保跨平台分发中的真实性与可追溯性。

总体而言，RAG、MAS、模型量化和数字水印分别从知识增强、群体协作、模型可部署性与内容可信性4个维度展现出代表性价值。它们共同推动数字人从“被动响应”向“主动演化、可信溯源与协同智能”的方向演进，为教育、医疗、工业与元宇宙等场景的落地应用奠定了多层次的技术基础。

3.2 小结

本节系统梳理了多模态实时交互式数字人的总体架构及关键组成模块，构建了一个由感知层、融合层、生成层和拓展层组成的通用框架。从技术流程上看，感知层承担了用户交互信息的高精度采集与信号解析任务，包括语音感知、表情与手势识别以及脑机接口等多模态输入模块，为交互系统提供了高带宽、低延迟的数据入口；融合层通过特征级、决策级与模型级的多层融合策略，实现了异构模态信息的统一表征与协同建模，有效提升了跨模态理解与语义一致性；生成层则是实现自然交互的核心环节，涵盖自然语言生成、语音合成、表情与动作生成等模块，依托大模型与生成式算法的引入，使数字人具备了实时对话、情感表达与多模态反馈的综合能力；拓展层作为系统的外延与生态接口，承担了模型部署、资源管理、隐私安全及跨平台交互的关键任务。

本节所构建的体系框架不仅揭示了多模态实时交互的内在逻辑，为多种应用场景下特定数字人系统的构建提供了技术选型指导，也为后续的整体—局部性能优化与应用拓展奠定了理论基础。通过对感知—融合—生成的分层设计，系统实现了从多源信息采集到多模态响应输出的闭环流程，体现了人机交互由“响应式”向“主动智能化”转变的趋势。该框架的提出为数字人技术在教育、医疗、娱乐及虚拟

社交等高实时性场景中的落地提供了可行路径，并为未来构建具备共情能力与环境感知能力的智能数字人提供了参考方向。

4 多模态实时交互式数字人应用场景

多模态实时交互式数字人的应用正逐步拓展至教育培训、医疗健康和零售服务等多个领域，在促进人机交互的智能化与沉浸化方面发挥着关键作用。

在教育场景中，数字人可充当虚拟教师、实验导师与实训教练，相较于传统的录播教学，基于数字人的教学模式在交互性、情感表达和学习黏性方面更具优势（Liu等，2025a）。感知层接入语音识别、表情识别和语义理解支持实时的多模态输入信号采集，生成层主要接入语音输出接口以辅助学员开展学习任务，扩展层接入多元化知识以增强个性化教学。这类数字人更加聚焦于根据学习者的语调、表情和行为变化自适应调整讲解内容与语气，以提升学习专注度与知识掌握率为核心指标。教学数字人已在职业技术教育、企业内训及远程实验指导等方向实现初步落地。

在医疗健康场景中，多模态实时交互式数字人可担任虚拟问诊助手、慢病管理教练、心理陪伴员或康复指导员。其感知层集成语音识别、面部表情分析、可穿戴设备生理信号乃至脑电接口，实时捕捉患者的语言、情绪与生理状态（Ayata等，2020）；融合层通过特征级或模型级融合，实现对疼痛、焦虑和认知障碍等健康风险的跨模态识别；生成层则基于大语言模型与情感语音合成，以自然、共情的方式输出个性化健康建议、用药提醒或心理干预话术，并通过逼真的表情与动作增强信任感。拓展层引入RAG机制，动态接入临床指南与电子病历，确保医学准确性；同时结合模型压缩技术实现边缘端低延迟部署，辅以数字水印保障数据隐私。目前，该类数字人已在老年慢病随访、术后康复以及青少年心理筛查等场景初步落地，有效提升了患者依从性、服务可及性与情感支持水平。

在零售与客户服务领域，多模态实时交互式数字人可作为智能导购、品牌虚拟客服或个性化推荐助手，显著提升用户体验与转化效率。感知层通过语音识别、人脸属性分析（如年龄、性别、情绪）、视线追踪及手势交互，实时捕获顾客意图与兴趣状态；生成层则驱动数字人以自然语言、情感化语音和品牌调性一致的表情动作进行响应，提供商品讲解、穿搭建议或促销引导。拓展层通过数字水印保护品牌资产与交互内容安全。

在VR社交场景中，多模态实时交互式数字人作为用户的虚拟化身（avatar）或社交陪伴体，支撑沉浸式、高拟真的远程人际互动。感知层通过VR头显内置眼动/表情追踪、手柄或手势识别、语音输入及空间音频，实时捕获用户言语、微表情、凝视方向与身体姿态；生成层则驱动虚拟化身以低延迟同步复现用户表情、口型、手势及语音，同时可基于大模型生成自然对话响应或非语言反馈（如点头、微笑），增强共情与临场感。

5 当前挑战与未来展望

尽管多模态实时交互式数字人技术发展迅速，但要构建具备真实感与情感理解的数字人仍面临多重挑战，主要集中在实时性能、数据对齐与融合的鲁棒性、情感与语境理解、个性与一致性的权衡和伦理与隐私安全等方面。本节将系统分析这些关键问题，并探讨其未来发展方向。

1）实时性能与可扩展性的瓶颈。多模态实时交互式数字人往往集成语音识别与合成、表情与口型驱动、动作生成、语言理解与规划以及渲染等多个高计算负载模块，如果端到端时延过高或抖动过大，用户会明显感到“卡顿”，交互自然度与沉浸感随之下降。

在边缘设备或移动终端上，实现低延迟、高保真的多模态生成尤为困难，而在多用户或群体交互场景下，算力、带宽与能耗压力进一步放大。根本原因在于模型规模与推理复杂度持续提升，高保真驱动和渲染方法对算力要求很高，而现有“云端推理 + 终端渲染”架构对网络波动和任务优先级的自适应能力有限。特别是在实时互动中，语言生成、音频合成与表情驱动需要在百毫秒级内保持严格同步，一旦任意模块发生抖动，用户即可明显察觉交互节奏被破坏，这是数字人相比一般多模态任务更严格的体验约束。

目前，模型压缩、蒸馏以及轻量化Transformer等技术为缓解算力压力提供了思路，边缘计算与云端协同架构也成为重要方向，通过任务分布式调度与

动态资源分配，在保证关键环节低时延的前提下提升系统可扩展性。随着AI专用芯片（Talati，2021）与异构计算硬件的普及，多模态数字人的实时生成与渲染能力将进一步增强。未来可围绕“延迟—能耗—保真度”的三者权衡设计专用轻量模型与边缘—云协同编排策略，为不同终端形态和应用场景自动选择合适的模型规模与运行模式。

2）多模态数据对齐与融合的鲁棒性。多模态数据的对齐与融合是数字人逼真交互的基础。语音与唇部动作的不匹配、手势与语义表达的脱节，以及背景噪声或遮挡造成的信息丢失，都会严重破坏交互的真实感。由于多模态数据通常来自不同的传感器或输入源，噪声和缺失的出现是不可避免的，这使得其在融合过程中面临巨大的挑战。噪声和缺失数据会破坏各模态之间的协调性，甚至可能导致不一致的输出。例如，语音识别可能会受到环境噪声的影响，导致错误的文本转录，进而影响后续的面部与肢体动作生成；而图像模态中的遮挡或低质量数据则可能导致面部表情或手势的错误捕捉。目前，针对这些问题的解决方案主要集中在噪声抑制、缺失数据填充和鲁棒的多模态融合方法上。例如，跨模态学习方法和GAN被用于填补丢失的数据，或者通过自监督学习从不完整数据中提取有用的信息。然而，现有的技术在动态环境下仍难以有效应对各种复杂场景中的噪声与缺失问题。未来的发展趋势是自适应融合模型的构建，即根据场景动态调整模态权重，实现模态冗余与信息互补。此外，大规模多模态预训练模型（如CLIP、Gemini、Kosmos系列）的应用将进一步提升跨模态理解与生成的精确度，使数字人具备更自然、协调的交互能力（高玄 等，2024）。

3）情感智能与上下文理解的局限。情感交互是数字人“拟人化”的关键，但在聊天陪伴、医疗陪护或心理咨询等高共情场景中，多模态实时交互式数字人在情绪识别和表达方面仍显不足，往往难以准确把握用户在长时程、多轮对话中的真实情绪与语境（赵思成 等，2025）。这类不足的根源不仅在于情绪本身具有模糊性和文化差异，还在于不同模态的情绪信号在时间动态与表达强度上的不一致，例如语音情绪变化快、表情具有迟滞性，而肢体动作则多呈低频变化，使跨模态情绪线索难以在统一语义空间中对齐。近期，情感计算（affective computing）（Pei 等，2024）与大语言模型的结合，为在语义上下文中融合视觉和语音线索、识别多层次情感状态提供了新的技术路径。然而，在实时数字人应用中，情绪识别与情绪生成需要同时保持跨模态一致性，例如语气、表情与动作必须保持同一情绪意图，否则用户会立即感受到“情绪不协调”，从而降低信任度与沉浸感。未来有望通过自适应情感学习与个性化情感建模，使数字人在长期交互中逐步学习用户的情绪模式与表达偏好，从而实现更贴合个体的响应。例如，构建基于语言大模型的多层情绪推理链路、引入用户情绪画像（emotion profile）用于捕捉跨轮次情绪依赖，以及采用情绪—动作解耦的生成机制以避免“语调悲伤但表情兴奋”等跨模态冲突，都将成为改善情感一致性的重要方向。

4）个性与一致性的权衡。在多模态实时交互式数字人中，鲜明而可感知的“个性”是提升拟人化程度与用户黏性的关键。然而，如果数字人在不同场景、不同时间段给出的言行风格前后矛盾，或在价值观、态度上表现出明显摇摆，就会造成“人设崩塌”，削弱用户对其身份与角色的信任感。尤其在教育辅导、医疗陪护和心理支持等高信任场景中，数字人一方面需要根据用户年龄、情绪状态和场景任务进行个性化表达，另一方面又必须在安全边界、价值取向和基本行为准则上保持高度一致。现有对话系统通常通过显式的人设描述来控制模型的说话风格与行为倾向，但静态、有限的人设文本在开放场景下容易出现覆盖不足的问题，导致生成内容偏离既定人设甚至自相矛盾。已有研究通过扩展与检索可用的人设信息，并在解码阶段对与当前语境一致的人设进行加权约束，从而在保证回复多样性的同时提升多轮对话中的人格连贯性和风格一致性。这类方法为多模态数字人的个性管理提供了思路：可以在语言、语音、表情与动作等多模态生成模块之间共享统一的人设表示，并通过长时记忆和一致性约束机制，在跨轮次、跨场景交互中维持整体人格的稳定呈现。对于多模态实时交互式数字人而言，个性与一致性的权衡不仅体现在文本内容层面，还体现在跨模态的一致对齐上：语音音色与语调应与设定性格相吻合，表情与肢体动作应与语言情绪保持协调。未来可以结合结构化人设知识图谱与层次化记忆机制，将“稳定的人格核心”（如价值观、基本态度）与“可变的情境策略”（如语气强弱、幽默程度）分层建模，使数字人在保持整体人设不变的前提下实现适度的个

性化自适应，从工程上更可控地平衡“有个性”、“不跑偏”和“可信赖”三者之间的关系。

5）伦理与隐私的风险挑战。多模态数字人依赖的语音、面部、动作和视线轨迹等数据具有高度敏感性，多模态实时交互又意味着这些数据在长时间尺度上被持续采集与处理，从而放大了隐私泄露与滥用风险。一方面，数据收集与训练过程中集中存储和跨域标注可能带来潜在泄露；另一方面，高拟真数字人也可能被用于制造“深度伪造”内容，引发身份冒用与社会信任危机。此外，由于实时数字人系统通常需要在推理链路中保留部分中间特征（如隐式身份向量、动作嵌入等）以实现连续交互，这些“可重构特征”本身也具有被逆向恢复的风险，使隐私暴露面进一步扩大。问题的深层矛盾在于，高质量数字人对大规模多模态数据的“数据饥渴”与现实中对“最小必要采集”、“用户可控”的隐私要求之间存在张力，生成模型的黑盒特性也使内容来源与责任划分模糊。应对之策包括在采集与训练阶段引入联邦学习、差分隐私等技术，尽量将敏感数据留在本地；在系统中构建透明的数据使用机制与一定程度可解释的决策过程，并配合深伪检测、AI溯源机制与数字水印，对多模态内容进行可追踪与可验证标记。此外，还需从制度层面建立多模态AI的伦理标准与监管框架，在算法与系统设计阶段融入“可追踪”、“可问责”的原则，明确数字人形象的所有权与使用边界。

6 结 语

本文系统综述了多模态实时交互式数字人的研究进展，围绕发展演进、关键技术与系统架构三大层面进行了深入分析与归纳。首先，本文梳理了数字人从非交互式到多模态交互式的发展脉络，揭示了其由静态展示向具备感知、理解与主动交互能力的智能体转变；其次，从建模、实时驱动与实时渲染3个方面系统阐述了多模态数字人技术的核心要素，重点讨论了NeRF、高斯溅射、扩散模型以及多模态驱动架构等前沿方法在提升真实感、降低时延与增强沉浸体验方面的关键作用。在此基础上，论文提出了感知层、融合层与生成层相结合的系统化框架，为构建具备情感理解、环境适应与社会交互能力的数字人提供了理论参考与技术支撑。

尽管当前研究在视觉逼真度与交互自然性方面取得显著进展，但在低延迟推理、跨模态协同学习及情感一致性建模等方面仍存在不足。现有系统普遍面临计算开销高、端到端优化不足以及轻量化部署受限等问题。此外，数字人伦理规范、隐私保护及人机共情评价体系尚未完善。未来研究应聚焦高效神经—几何混合表示以平衡保真度与实时性，进一步结合大模型的上下文推理能力，实现情感驱动与主动交互，最终构建可信、可解释且具社会适应性的多模态实时交互式数字人体系。

参考文献（References）

Abramson J, Ahuja A, Carnevale F, Georgiev P, Goldin A, Hung A, et al. 2022. Improving multimodal interactive agents with reinforcement learning from human feedback [EB/OL]. [2025-10-15]. https://doi.org/pdf/2211.11602.pdf

Adiya T, Yoon J S, Lee J, Kim S and Lim H. 2024. Bidirectional temporal diffusion model for temporally consistent human animation//Proceedings of the 12th International Conference on Learning Representations. Vienna, Austria: OpenReview.net

Ali G, Kim W, Anwar M S, Hwang J I and Choi A. 2025. Expanding multilingual co-speech interaction: the impact of enhanced gesture units in text-to-gesture synthesis for digital humans. IEEE Access, 13: 145144-145157 [DOI: 10.1109/ACCESS.2025.3596328]

Arima Y, Harada Y and Okada M. 2025. Classifying interpersonal interaction in virtual reality: sensor-based analysis of human interaction with pre-recorded avatars. Frontiers in Virtual Reality, 6: #1623764 [DOI: 10.3389/frvir.2025.1623764].

Atal B and Rabiner L. 1976. A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition. IEEE Transactions on Acoustics, Speech, and Signal Processing, 24(3): 201-212 [DOI: 10.1109/TASSP.1976.1162800]

Ayata D, Yaslan Y and Kamasak M E. 2020. Emotion recognition from multimodal physiological signals for emotion aware healthcare systems. Journal of Medical and Biological Engineering, 40(2): 149-157 [DOI: 10.1007/s40846-019-00505-7]

Baltrušaitis T, Ahuja C and Morency L P. 2019. Multimodal machine learning: a survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(2): 423-443 [DOI: 10.1109/TPAMI.2018.2798607].

Bao L C, Lin X K, Chen Y J, Zhang H X, Wang S, Zhe X F, et al. 2022. High-fidelity 3D digital human head creation from RGB-D selfies. ACM Transactions on Graphics (TOG), 41(1): #3 [DOI: 10.1145/3472954]

Basak S, Agrawal H, Jena S, Gite S, Bachute M, Pradhan B, et al. 2023. Challenges and limitations in speech recognition technology:

a critical review of speech signal processing algorithms, tools and systems. Computer Modeling in Engineering and Sciences, 135(2): 1053-1089 [DOI: 10.32604/cmes.2022.021755]

Brito I A, Dollis J S, Färber F B, Ribeiro P S F B, Sousa R T and Filho A R G. 2025. Integrating personality into digital humans: a review of LLM-driven approaches for virtual reality [EB/OL]. [2025-10-15]. https://doi.org/pdf/2503.16457.pdf

Carvalho A, Correia L M, Grilo A and Dinis R. 2022. Analysis of strategies for minimising end-to-end latency in 5G networks//Proceedings of 2022 International Conference on Broadband Communications for Next Generation Networks and Multimedia Applications (CoBCom). Graz, Austria: IEEE: 1-6 [DOI: 10.1109/CoBCom55489.2022.9880722]

Chan C, Ginosar S, Zhou T H and Efros A. 2019. Everybody dance now//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South): IEEE: 5932-5941 [DOI: 10.1109/ICCV.2019.00603]

Chen M, Cui L Y, Zhang W Y, Zhang H X, Zhou Y, Li X H, et al. 2025a. MIDAS: multimodal interactive digital-human synthesis via real-time autoregressive video generation [EB/OL]. [2025-10-15]. https://doi.org/pdf/2508.19320.pdf

Chen X H, Luo K, Gee T and Nejati M. 2024. Does ChatGPT and Whisper make humanoid robots more relatable? [EB/OL]. [2025-10-15]. https://doi.org/pdf/2402.07095.pdf

Chen Y S, Niu Z K, Ma Z Y, Deng K Q, Wang C H, Zhao J, et al. 2025b. F5-TTS: a Fairytaler that fakes fluent and faithful speech with flow matching//Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vienna, Austria: Association for Computational Linguistics: 6255-6271 [DOI: 10.18653/v1/2025.acl-long.313]

Chen Z X and Liu Z W. 2022. Relighting4D: neural relightable human from videos//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer: 606-623 [DOI: 10.1007/978-3-031-19781-9_35]

Cheng W X L, Wan C, Cao Y P and Chen S H. 2024. RITA: a real-time interactive talking avatars framework [EB/OL]. [2025-10-15]. https://doi.org/pdf/2406.13093.pdf

Coll L C, Lauer-Schmaltz M W, Cash P, Hansen J P and Maiel A. 2025. Towards the "digital me": a vision of authentic conversational agents powered by personal human digital twins [EB/OL]. [2025-10-15]. https://doi.org/pdf/2506.23826.pdf

Condegni A, Wang W T and Li R. 2023. A digital human system with realistic facial expressions for friendly human-machine interaction//Proceedings of the 19th International Conference on Advanced Intelligent Computing Technology and Applications. Zhengzhou, China: Springer: 787-798 [DOI: 10.1007/978-981-99-4755-3_68]

Dong Y M. 2022. 3D face recognition neural network for digital human resource management. Scientific Programming, 2022(1): #6544282 [DOI: 10.1155/2022/6544282]

Du Z H, Gao C F, Wang Y X, Yu F, Zhao T Y, Wang H, et al. 2025. CosyVoice 3: towards in-the-wild speech generation via scaling-up and post-training [EB/OL]. [2025-10-15]. https://arxiv.org/pdf/2505.17589.pdf

Fiorini L, Loizzo F G C, Sorrentino A, Kim J, Rovini E, Di Nuovo A, et al. 2021. Daily gesture recognition during human-robot interaction combining vision and wearable systems. IEEE Sensors Journal, 21(20): 23568-23577 [DOI: 10.1109/JSEN.2021.3108011]

Gao J, Li P, Chen Z K and Zhang J N. 2020. A survey on deep learning for multimodal data fusion. Neural Computation, 32(5): 829-864 [DOI: 10.1162/neco_a_01273]

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(9): 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(9): 2494-2512) [DOI: 10.11834/jig.230649]

Guo H H, Hu Y, Shen F Y, Tang X, Wu Y C, Xie F L, et al. 2025a. FireRedTTS-1S: an upgraded streamable foundation text-to-speech system [EB/OL]. [2025-10-15]. https://arxiv.org/pdf/2503.20499.pdf

Guo S Y, Ju Y H, Chen X, Sapatnekar S S and Gu J. 2025b. MobilePBR: a 28-nm energy-efficient rendering processor for photorealistic augmented reality with inverse rendering and background clustering. IEEE Journal of Solid-State Circuits, 60(1): 125-135 [DOI: 10.1109/JSSC.2024.3484212]

Habermann M, Liu L J, Xu W P, Pons-Moll G, Zollhoefer M and Theobalt C. 2023. HDHumans: a hybrid approach for high-fidelity digital humans. Proceedings of the ACM on Computer Graphics and Interactive Techniques, 6(3): #36 [DOI: 10.1145/3606927]

He C, Ren J Q, Dong Y, Xiang J J, Shen X J, Yuan W H, et al. 2025a. Textoon: generating vivid 2D cartoon characters from text descriptions [EB/OL]. [2025-10-16]. https://doi.org/pdf/2501.10020.pdf

He C, Ren J Q, Xiang J J and Shen X J. 2025b. CartoonAlive: towards expressive Live2D modeling from single portraits [EB/OL]. [2025-10-16]. https://doi.org/pdf/2507.17327.pdf

Hietamies A. 2024. Avoimen Lähdekoodin Ohjelmien Käyttö 2D-Pelin Grafiikoihin. Oulu, Finland: Oulun ammattikorkeakoulu

Hong Y, Peng B, Xiao H Y, Liu L G and Zhang J Y. 2022. HeadNeRF: a realtime NeRF-based parametric head model//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 20342-20352 [DOI: 10.1109/CVPR52688.2022.01973]

Hong Y, Wu Y Z, Shen Z H, Guo C C, Jiang Y H, Zhang Y L, et al. 2025. BEAM: bridging physically-based rendering and Gaussian modeling for relightable volumetric video//Proceedings of the 33rd ACM International Conference on Multimedia. Dublin, Ireland: Association for Computing Machinery: 7968-7977 [DOI: 10.1145/3746027.3755233]

Hu T, Hong F Z, Chen Z X and Liu Z W. 2024. FashionEngine: interactive 3D human generation and editing via multimodal controls [EB/OL] [2025-10-16]. https://doi.org/pdf/2404.01655.pdf

Huang Y Y, Yi H W, Xiu Y L, Liao T T, Tang J X, Cai D, et al. 2024. TeCH: text-guided reconstruction of lifelike clothed humans//Proceedings of 2024 International Conference on 3D Vision (3DV). Davos, Switzerland: IEEE: 1531-1542 [DOI: 10.1109/3DV62453.2024.00152]

Huang Z H, Yang S Z, Lin W, Ni J, Sun S L, Chen Y W, et al. 2022. Knowledge distillation: a survey. Chinese Journal of Computers, 45(3): 624-653 (黄震华, 杨顺志, 林威, 倪娟, 孙圣力, 陈运文, 等. 2022. 知识蒸馏研究综述. 计算机学报, 45(3): 624-653) [DOI: 10.11897/SP.J.1016.2022.00624]

Jeong M, Kim H, Cheon S J, Choi B J and Kim N S. 2021. Diff-TTS: a denoising diffusion model for text-to-speech//Proceedings of the 22nd Annual Conference of the International Speech Communication Association. Brno, Czechia: ISCA: 3605-3609

Kerbl B, Kopanas G, Leimkuehler T and Drettakis G. 2023. 3D Gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics (TOG), 42(4): #139 [DOI: 10.1145/3592433]

Khan A A, Laghari A A, Shaikh A A, Dootio M A, Estrela V V and Lopes R T. 2022. A blockchain security module for brain-computer interface (BCI) with multimedia life cycle framework (MLCF). Neuroscience Informatics, 2(1): #100030 [DOI: 10.1016/j.neuri.2021.100030]

Kim M, Kim D, Park Y and Jeong D. 2024. Development of an expert Chatbot for digital forensics using RAG model implementation//Proceedings of 2024 International Conference on Platform Technology and Service (PlatCon). Jeju, Korea (South): IEEE: 182-187 [DOI: 10.1109/PlatCon63925.2024.10830748]

Kim M, Kim T and Lee K T. 2025. 3D digital human generation from a single image using generative AI with real-time motion synchronization. Electronics, 14(4): #777 [DOI: 10.3390/electronics14040777]

Lan C, Wang Y S, Wang C Z, Song S R and Gong Z. 2023. Application of ChatGPT-based digital human in animation creation. Future Internet, 15(9): #300 [DOI: 10.3390/fi15090300]

Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, et al. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks//Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: #793 [DOI: 10.5555/3495724.3496517]

Li J N, Xu Y, Grossman T, Santosa S and Li M. 2024. OmniActions: predicting digital actions in response to real-world multimodal sensory inputs with LLMs//Proceedings of 2024 CHI Conference on Human Factors in Computing Systems. Honolulu, USA: Association for Computing Machinery: #8 [DOI: 10.1145/3613904.3642068]

Li X, Zhang J Y and Liu Y Z. 2022. Speech driven facial animation generation based on GAN. Displays, 74: #102260 [DOI: 10.1016/j.displa.2022.102260]

Li X D, Xie P, Ren Y, Gan Q J, Zhang C, Kong F Y, et al. 2025b. InfinityHuman: towards long-term audio-driven human [EB/OL]. [2025-10-16]. https://arxiv.org/pdf/2508.20210.pdf

Li X Y, Li G, Lin Z H, Qian Y C, Yao G X, Jia W N, et al. 2025a. MoDA: multi-modal diffusion architecture for talking head generation [EB/OL]. [2025-10-16]. https://arxiv.org/pdf/2507.03256.pdf

Li Z, Li H Y and Meng L. 2023. Model compression for deep neural networks: a survey. Computers, 12(3): #60 [DOI: 10.3390/computers12030060]

Liew T W, Tan S M and Ismail H. 2017. Exploring the effects of a non-interactive talking avatar on social presence, credibility, trust, and patronage intention in an e-commerce website. Human-centric Computing and Information Sciences, 7(1): #42 [DOI: 10.1186/s13673-017-0123-4]

Lin G J, Jiang J W, Liang C, Zhong T Y, Yang J Q, Zheng Z R, et al. 2025. CyberHost: a one-stage diffusion framework for audio-driven talking body generation//Proceedings of the 13th International Conference on Learning Representations. Singapore, Singapore: OpenReview.net

Liu C, Lin Q F, Zeng Z J and Pan Y. 2024. EmoFace: audio-driven emotional 3D face animation//Proceedings of 2024 IEEE Conference Virtual Reality and 3D User Interfaces. Orlando, USA: IEEE: 387-397 [DOI: 10.1109/VR58804.2024.00060].

Liu H Y, Han X T, Jin C B, Qian L H, Wei H W, Lin Z, et al. 2023. Human MotionFormer: transferring human motions with vision transformers//Proceedings of the 11th International Conference on Learning Representations. Kigali, Rwanda: OpenReview.net

Liu L J, Habermann M, Rudnev V, Sarkar K, Gu J T and Theobalt C. 2021. Neural actor: neural free-view synthesis of human actors with pose control. ACM Transactions on Graphics (TOG), 40(6): #219 [DOI: 10.1145/3478513.3480528]

Liu Q, Sha Y H, Zhang K, Huang Z Y, Zhu L B, Lu J Y, et al. 2025a. Advancements in digital humans for recorded courses: enhancing learning experiences via personalized interaction. Frontiers of Digital Education, 2(4): #35 [DOI: 10.1007/s44366-025-0072-9]

Liu X Y, Liu X Y, Yang P, Wang Z Q and Liu F Y. 2025b. An approach to optimizing semantic consistency for text-to-digital human generation. Engineering Applications of Artificial Intelligence, 160: #111909 [DOI: 10.1016/j.engappai.2025.111909]

Liu Y F, Wei W, Liu J Y, Mao X L, Fang R and Chen D Y. 2022. Improving personality consistency in conversation by persona extending//Proceedings of the 31st ACM International Conference on Information and Knowledge Management. Atlanta, USA: Association for Computing Machinery: 1350-1359 [DOI: 10.1145/3511808.3557359]

Malanowska A, Mazurczyk W, Araghi T K, Megías D and Kuribayashi M. 2024. Digital watermarking——A meta-survey and techniques for fake news detection. IEEE Access, 12: 36311-36345 [DOI:

10.1109/ACCESS.2024.3374201]

Mehrabian A and Diamond S G. 1971. Seating arrangement and conversation. Sociometry, 34(2): 281-289 [DOI: 10.2307/2786417]

Melnik A, Miasayedzenkau M, Makaravets D, Pirshtuk D, Akbulut E, Holzmann D, et al. 2024. Face generation and editing with StyleGAN: a survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 46(5): 3557-3576 [DOI: 10.1109/TPAMI.2024.3350004]

Mildenhall B, Srinivasan P P, Tancik M, Barron J T, Ramamoorthi R and Ng R. 2022. NeRF: representing scenes as neural radiance fields for view synthesis. Communications of the ACM, 65(1): 99-106 [DOI: 10.1145/3503250]

Mo D H, Tien C L, Yeh Y L, Guo Y R, Lin C S, Chen C C, et al. 2023. Design of digital-twin human-machine interface sensor with intelligent finger gesture recognition. Sensors, 23(7): #3509 [DOI: 10.3390/s23073509]

Ng E, Romero J, Bagautdinov T, Bai S J, Darrell T, Kanazawa A, et al. 2024. From audio to photoreal embodiment: synthesizing humans in conversations//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 1001-1010 [DOI: 10.1109/CVPR52733.2024.00101]

Okamoto T, Ohtani Y, Toda T and Kawai H. 2024. ConvNeXt-TTS and ConvNeXt-VC: ConvNeXt-based fast end-to-end sequence-to-sequence text-to-speech and voice conversion//Proceedings of 2024 IEEE International Conference on Acoustics, Speech and Signal Processing. Seoul, Korea (South): IEEE: 12456-12460 [DOI: 10.1109/ICASSP48485.2024.10446890]

Ospina-Bohórquez A, Rodríguez-González S and Vergara-Rodríguez D. 2021. On the synergy between virtual reality and multi-agent systems. Sustainability, 13(8): #4326 [DOI: 10.3390/su13084326]

Ou H P, Yue P, Duan Q S, Mo S W, Zhao Z and Qu X D. 2023. Development of a low-cost and user-friendly system to create personalized human digital twin//Proceedings of the 45th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). Sydney, Australia: IEEE: 1-4 [DOI: 10.1109/EMBC40787.2023.10340461]

Pan Y, Li S X, Tan S, Wei J J, Zhai G T and Yang X K. 2025. Advancements in digital character stylization, multimodal animation, and interaction. Journal of Image and Graphics, 30(2): 334-360 (潘烨, 李韶旭, 谭帅, 韦俊杰, 翟广涛, 杨小康. 2025. 数字人风格化、多模态驱动与交互进展. 中国图象图形学报, 30(2): 334-360) [DOI: 10.11834/jig.230639]

Pan Y, Liu C, Xu S C, Tan S and Yang J L. 2025. VASA-Rig: audio-driven 3D facial animation with 'Live' mood dynamics in virtual reality. IEEE Transactions on Visualization and Computer Graphics, 31(5): 2416-2425 [DOI: 10.1109/TVCG.2025.3549168]

Pang H K, Zhu H M, Kortylewski A, Theobalt C and Habermann M. 2024. ASH: animatable Gaussian splats for efficient and photoreal human rendering//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 1165-1175 [DOI: 10.1109/CVPR52733.2024.00117]

Park H J, Liu J, Kim J S, Yang J Y, Han S W and Song E. 2025. RapFlow-TTS: rapid and high-fidelity text-to-speech with improved consistency flow matching//Proceedings of the 26th Annual Conference of the International Speech Communication Association. Rotterdam, the Netherlands: ISCA

Pei G X, Li H Y, Lu Y D, Wang Y L, Hua S Z and Li T H. 2024. Affective computing: recent advances, challenges, and future trends. Intelligent Computing, 3: #0076 [DOI: 10.34133/icomputing.0076]

Peng S D, Zhang Y Q, Xu Y H, Wang Q Q, Shuai Q, Bao H J, et al. 2021. Neural body: implicit neural representations with structured latent codes for novel view synthesis of dynamic humans//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 9050-9059 [DOI: 10.1109/CVPR46437.2021.00894]

Pham T T, Do T, Le N, Le N, Nguyen H, Tjiputra E, et al. 2024. Style transfer for 2D talking head generation//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 7500-7509 [DOI: 10.1109/CVPRW63382.2024.00745]

Qin D F, Lin H Y, Zhang Q X, Qiao K C, Zhang L W, Saito J, et al. 2025. Instant Gaussian splatting generation for high-quality and real-time facial asset rendering. IEEE Transactions on Pattern Analysis and Machine Intelligence: 1-15 [DOI: 10.1109/TPAMI.2025.3550195]

Qin Z, Zheng R B, Wang Y B, Li T Q, Yuan Y, Chen J D, et al. 2026. HumanSense: from multimodal perception to empathetic context-aware responses through reasoning MLLMs//Proceedings of 2026 AAAI Conference on Artificial Intelligence. Singapore, Singapore: AAAI Press: 24973-24981 [DOI: 10.1609/aaai.v40i30.39685]

Rossoni M, Pozzi M, Colombo G, Gribaudo M and Piazzolla P. 2024. Physically based rendering of animated point clouds for extended reality. Journal of Computing and Information Science in Engineering, 24(5): #054501 [DOI: 10.1115/1.4063559]

Saffaryazdi N, Gunasekaran T S, Loveys K, Broadbent E and Billinghurst M. 2026. Empathetic conversational agents: utilizing neural and physiological signals for enhanced empathetic interactions. International Journal of Human-Computer Interaction, 42(6): 4555-4579 [DOI: 10.1080/10447318.2025.2540500]

Safitri D R and Wibawa M. 2022. West papua culture-based virtual Youtuber avatar design with animated rigging on live2D cubism//Proceedings of the 3rd International Conference on Information Technology and Security. Malang City, Indonesia: LPPM: 183-203

Saraswati A M, Setiawan I R and Asriyanik. 2023. Pemanfaatan tools Live2D terhadap animasi menggunakan metode face tracking. Jurnal Ilmiah Komputer Grafis, 16(1): 186-195

Shao Z J, Wang Z L, Li Z, Wang D T, Lin X R, Zhang Y, et al. 2024.

SplattingAvatar: realistic real-time human avatars with mesh-embedded Gaussian splatting//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 1606-1616 [DOI: 10.1109/CVPR52733.2024.00159]

Sheng Z Y, Du Z H, Zhang S L, Yan Z J and Chen L P. 2025. SyncSpeech: efficient and low-latency text-to-speech based on temporal masked transformer [EB/OL]. [2025-10-15]. https://doi.org/pdf/2502.11094.pdf

Sheremetieva A, Romanovych T, Frish S, Maksymenko M and Georgiou O. 2023. What's my future: a multisensory and multimodal digital human agent interactive experience//Proceedings of 2023 ACM International Conference on Interactive Media Experiences. Nantes, France: ACM: 40-46 [DOI: 10.1145/3573381.3596161]

Siarohin A, Lathuilière S, Tulyakov S, Ricci E and Sebe N. 2019. First order motion model for image animation//Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: #7641 [DOI: 10.5555/3454287.3454928]

Sinha A K, Kulkarni C and Olwal A. 2024. Levels of multimodal interaction//Companion Proceedings of the 26th International Conference on Multimodal Interaction. San Jose, Costa Rica: ACM: 51-55 [DOI: 10.1145/3686215.3690153]

Song W F, Wang X, Zheng S, Li S, Hao A M and Hou X. 2025. TalkingStyle: personalized speech-driven 3D facial animation with style preservation. IEEE Transactions on Visualization and Computer Graphics, 31(9): 4682-4694 [DOI: 10.1109/TVCG.2024.3409568]

Sonlu S, Bendiksen B, Durupinar F and Güdükbay U. 2025. Effects of embodiment and personality in LLM-based conversational agents//Proceedings of 2025 IEEE Conference Virtual Reality and 3D User Interfaces. Saint Malo, France: IEEE: 718-728 [DOI: 10.1109/VR59515.2025.00094]

Sun Q, Cui Y F, Zhang X S, Zhang F, Yu Q Y, Luo Z X, et al. 2024. Generative multimodal models are in-context learners//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: #13286 [DOI: 10.48550/arXiv.2312.13286]

Talati D V. 2021. Silicon minds: the rise of AI-powered chips. International Journal of Science and Research Archive, 1(2): 97-108 [DOI: 10.30574/ijsra.2021.1.2.0019]

Tan X, Chen J W, Liu H H, Cong J, Zhang C and Liu Y Q. 2024. NaturalSpeech: end-to-end text-to-speech synthesis with human-level quality. IEEE Transactions on Pattern Analysis and Machine Intelligence, 46(6): 4234-4245 [DOI: 10.1109/TPAMI.2024.3356232]

Tian J F, Chen H H, Xu G H, Yan M, Gao X, Zhang J H, et al. 2023. ChatPLUG: open-domain generative dialogue system with internet-augmented instruction tuning for digital human [EB/OL]. [2025-10-15]. https://doi.org/pdf/2304.07849.pdf

Vilchis C, Gonzalez-Mendoza M, Chang L, Navarro-Tuch S A, Ruiz G O and Rudomin I. 2022. A study of the frameworks for digital humans: analyzing facial tracking evolution and new research directions with AI//Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications-Volume 2. Online Streaming: SciTePress: 154-162 [DOI: 10.5220/0010823600003124]

Völkel S T and Kaya L. 2021. Examining user preference for agreeableness in Chatbots//Proceedings of the 3rd Conference on Conversational User Interfaces. Bilbao, Spain: ACM: #38 [DOI: 10.1145/3469595.3469633]

Wang C, Huang J M, Zhang R, Wang Q, Yang H T, Wan P F, et al. 2025. Physically based facial texture generation in the wild. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1-12 [DOI: 10.1109/TPAMI.2025.3580953]

Wang N and Zhang H Y. 2024. Integration of AI generated content (AIGC) in the development of digital human in the metaverse//Proceedings of 2024 International Conference on Language Technology and Digital Humanities. Bhubaneswar, India: IEEE: 92-96 [DOI: 10.1109/LTDH64262.2024.00027]

Wiles O, Koepke A S and Zisserman A. 2018. X2Face: a network for controlling face generation using images, audio, and pose codes//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer: 690-706 [DOI: 10.1007/978-3-030-01261-8_41]

Wu S F, Liu Z G, Zhang B B, Zimmermann R, Ba Z J, Zhang X S, et al. 2024a. Do as I do: pose guided human motion copy. IEEE Transactions on Dependable and Secure Computing, 21(6): 5293-5307 [DOI: 10.1109/TDSC.2024.3371530]

Wu X, Guan H, Huang Y, Song C H, Niu B N, Zhang S W, et al. 2024b. A comprehensive review of three-dimensional watermarking algorithms//Proceedings of the 10th International Conference on Communication and Information Processing. Lingshui, China: ACM: 469-475 [DOI: 10.1145/3708657.3708734]

Xu B T and Zhang X L. 2024. The technological evolution of digital humans and their application in Chinese media//Proceedings of 2024 International Conference on Artificial Intelligence, Digital Media Technology and Interaction Design. Tianjin, China: Association for Computing Machinery: 62-66 [DOI: 10.1145/3726010.3726019]

Xuanyuan M D, Wang Y W, Guo H L, Qu H S, Zhang K, Li Z M, et al. 2026. Creating multimodal interactive digital twin characters from videos: a dataset and baseline. IEEE Transactions on Pattern Analysis and Machine Intelligence, 48(1): 92-108 [DOI: 10.1109/TPAMI.2025.3603653]

Xue H F, Ju Y, Miao C L, Wang Y J, Wang S Y, Zhang A D, et al. 2021. mmMesh: towards 3D real-time dynamic human mesh construction using millimeter-wave//Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services. ACM: 269-282 [DOI: 10.1145/3458864.3467679]

Xue Y X, Yu Y D, Yin K Y, Li P F, Xie S X and Ju Z J. 2022. Human

in-hand motion recognition based on multi-modal perception Information Fusion. IEEE Sensors Journal, 22(7): 6793-6805 [DOI: 10.1109/JSEN.2022.3148992]

Yan Y C, Cheng Y H, Chen Z, Peng Y C, Wu S J, Zhang W T, et al. 2023. A survey on generative 3D digital humans based on neural networks: representation, rendering, and learning. Scientia Sinica Informationis, 53(10): 1858-1891 (晏轶超, 程宇豪, 陈琢, 彭乙骢, 吴思婧, 张维天, 等. 2023. 基于神经网络的生成式三维数字人研究综述: 表示, 渲染与学习. 中国科学: 信息科学), 53(10): 1858-1891 [DOI: 10.1360/SSI-2022-0319]

Yang F, Fang L, Suo R, Zhang J and Whang M. 2025. Development of an interactive digital human with context-sensitive facial expressions. Sensors, 25(16): #5117 [DOI: 10.3390/s25165117]

Yang J N, Qian T T, Zhang F and Khan S U. 2021. Real-time facial expression recognition based on edge computing. IEEE Access, 9: 76178-76190 [DOI: 10.1109/ACCESS.2021.3082641]

Ye H, Zhu W T, Wang C Y, Wu R J and Wang Y Z. 2022. Faster VoxelPose: real-time 3D human pose estimation by orthographic projection//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer: 142-159 [DOI: 10.1007/978-3-031-20068-7_9]

Yu S D, Huang Y X, Huang Y T, Lin S Y, Tan L T and Chen K L. 2025. Design and implementation of a high-real-time Chinese digital human dialogue platform//Proceedings of the 4th International Conference on Computer, Artificial Intelligence and Control Engineering. Hefei, China: ACM: 996-1001 [DOI: 10.1145/3727648.3727812]

Zhang H C, Li L X and Liu D J. 2024. Survey of multimodal data fusion research. Journal of Frontiers of Computer Science and Technology, 18(10): 2501-2520 (张虎成, 李雷孝, 刘东江. 2024. 多模态数据融合研究综述. 计算机科学与探索, 18(10): 2501-2520) [DOI: 10.3778/j.issn.1673-9418.2403083]

Zhang H M, Lu Z Y and Liu K J. 2024. A text-to-speech synthesis method based on speaker transfer//Proceedings of the 24th IEEE International Conference on Communication Technology. Chengdu, China: IEEE: 1822-1826 [DOI: 10.1109/ICCT62411.2024.10946354]

Zhao S C, Feng Y F, Zhang Z C, Sun B, Zhang S P, Gao Y, et al. 2025. Research advancements on emotionally and intellectually integrated digital humans and robotics. Journal of Image and Graphics, 30(6): 2139-2160 (赵思成, 丰一帆, 张知诚, 孙斌, 张盛平, 高跃, 等. 2025. 情智兼备数字人与机器人研究进展. 中国图象图形学报, 30(6): 2139-2160) [DOI: 10.11834/jig.240780]

Zhou Y J, Chen Y D, Bi K Y, Xiong L and Liu H. 2023. An implementation of multimodal fusion system for intelligent digital human generation [EB/OL]. [2025-10-15]. https://doi.org/pdf/2310.20251.pdf

Zhu H Y, Hieu N Q, Hoang D T, Nguyen D N and Lin C T. 2024. A human-centric Metaverse enabled by brain-computer interface: a survey. IEEE Communications Surveys and Tutorials, 26(3): 2120-2145 [DOI: 10.1109/COMST.2024.3387124]

Zielonka W, Bagautdinov T, Saito S, Zollhöfer M, Thies J and Romero J. 2025. Drivable 3D Gaussian avatars//Proceedings of 2025 International Conference on 3D Vision (3DV). Singapore, Singapore: IEEE: 979-990 [DOI: 10.1109/3DV66043.2025.00095]

Zoric G, Forchheimer R and Pandzic I S. 2011. On creating multimodal virtual humans—real time speech driven facial gesturing. Multimedia Tools and Applications, 54(1): 165-179 [DOI: 10.1007/s11042-010-0526-y]

Zuo L Y, An K Y, Zhang S L and Yan Z J. 2023. Advancing VAD systems based on multi-task learning with improved model structures [EB/OL]. [2025-10-15]. https://doi.org/pdf/2312.14860.pdf

作者简介

杜瑞麒，男，硕士研究生，主要研究方向为多模态交互与数字人集成。E-mail: 23010502021@mail.hnust.edu.cn

李涛，通信作者，男，教授，主要研究方向为大语言模型、大数据与边缘计算。E-mail: tlee@hnust.edu.cn

杨柏蔼，男，博士研究生，主要研究方向为智能设计与人机交互。E-mail: yangboai@hnu.edu.cn

周丰波，男，本科生，主要研究方向为情感语音合成与实时人机对话系统。E-mail: 2305010716@mail.hnust.edu.cn

屈薇，女，讲师，主要研究方向为大语言模型与多模态数据融合。E-mail: weiqu23@mail.hnust.edu.cn