计 算 机 学 报
CHINESE JOURNAL OF COMPUTERS

数字人评测综述：现状、技术与展望

蒋一笔[1)] 秦川[1)] 钱振兴[2)] 张新鹏[2)]

[1)](上海理工大学 光电信息与计算机工程学院 上海 200093)

[2)](复旦大学 计算与智能创新学院 上海 200433)

摘 要 近年来，随着计算机视觉、自然语言处理等人工智能技术的进步，数字人的综合性能显著提高，其在教育、医疗和娱乐等行业得到了广泛的应用。然而，种类繁多的数字人功能各异，如何准确且有效地评测成为一个亟待解决的问题。由于现有标准与指标缺少严谨、全面的评测流程，因此结果的客观性与科学性有待提升。同时，用户对数字人拟人度、交互性等要求的日益提高，对评测方法也提出了更高的要求。本文聚焦于数字人评测，首先介绍数字人的技术架构，然后对行业标准和相关指标进行了综述，最后为数字人综合评测方法提供了参考框架。具体地，在行业标准方面，对比了国际电信联盟、IEEE 计算机学会标准活动委员会和世界超高清视频产业联盟提出的相关标准，进而明确数字人评测的主要维度。在相关指标方面，介绍了数字人质量和内容性能指标的研究现状，对性能评测框架以及相关数据集进行了探讨。基于对当前评测方法的系统性分析，本文从质量与内容两个维度展望了数字人综合评测方法的发展趋势，以期为相关领域的研究与实践提供参考。

关键词 数字人；评测方法；大语言模型；多模态学习；人机交互

中图法分类号 TP18

A Survey of Digital Human Evaluation: Present, Technology, and Prospect

JIANG Yi-Bi[1)] QIN Chuan[1)] QIAN Zhen-Xing[2)] ZHANG Xin-Peng[2)]

[1)](School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093)

[2)](College of Computer Science and Artificial Intelligence, Fudan University, Shanghai 200433)

Abstract Recent advances in computer vision, natural language processing, and generative artificial intelligence have significantly enhanced the capabilities of digital humans, enabling their widespread adoption across education, healthcare, entertainment, e-commerce, and customer service. With the Chinese AI digital human core market projected to grow from 33.9 billion RMB in 2025 to 93.6 billion RMB by 2030, and the global market expected to reach 1.92 trillion USD by 2035, the rapid proliferation of digital human systems has outpaced the development of rigorous evaluation methodologies. Existing industry standards and academic metrics suffer from critical limitations, including ambiguous attribute weighting, insufficient fine-grained quantitative indicators, over-reliance on subjective evaluations that introduce individual biases and lack reproducibility, poor adaptability to emerging technologies such as large language models, and the absence of unified frameworks integrating multimodal quality and content assessment. This paper presents a systematic survey of digital human evaluation, covering technical architectures, industry standards, quantitative metrics, evaluation frameworks, datasets, and future research directions. We first delineate the four-layer technical architecture of digital humans(perception, decision, expression, and extension), establishing a theoretical foundation for targeted assessment. The perception layer receives and processes multimodal user inputs

本课题得到国家自然科学基金面上项目（No.62571333、No.62172280）、国家重点研发计划项目 (No. 2023YFF0905000)的资助。蒋一笔，博士研究生，主要研究领域为多模态学习和AIGC检测。秦川(**通信作者**)，博士，教授，博士生导师，计算机学会（CCF）高级会员（会员号36102M），主要研究领域为多媒体信息安全、AI安全。钱振兴，博士，教授，主要研究领域为信息隐藏、AI安全。张新鹏，博士，教授，主要研究领域为多媒体信息安全、AI安全。

including speech, video, and other sensory data. The decision layer fuses multimodal information and generates semantic understanding, increasingly leveraging large language models for reasoning and response generation. The expression layer converts decision outputs into natural speech through text-to-speech technologies and generates visual representations via 3D modeling, binding, driving, and rendering pipelines. The extension layer augments basic functionalities with retrieval-augmented generation, digital watermarking for copyright protection, and multi-agent collaboration capabilities. We then critically compare major industrial standards, including those from the International Telecommunication Union, the IEEE Computer Society Standards Activities Committee, and the UHD World Association. Our analysis identifies persistent limitations across these standards: unclear dimension weights preventing comprehensive scoring, insufficient quantification for cross-product comparison, inherent biases in subjective evaluation including personal preferences and question framing effects, high labor costs limiting scalability, and over-simplified objective metrics that fail to capture authentic user experience. Subsequently, we review objective evaluation metrics from two complementary perspectives—quality and content. For quality assessment, we examine video quality indicators including full-reference metrics, reduced-reference methods, and no-reference deep learning approaches, as well as geometry-aware methods specifically designed for 3D digital humans. For audio quality, we cover traditional signal-based metrics, perceptual measurement standards, and emerging no-reference methods leveraging self-supervised and multimodal models. For content assessment, we analyze appearance fidelity metrics measuring similarity between generated and reference identities, lip synchronization quantification for audio-visual alignment, prosody evaluation capturing pitch and intonation accuracy, motion naturalness indicators for head and eyebrow dynamics, action consistency measures aligning generated movements with driving semantics, and multi-view consistency metrics for 3D digital humans. For interaction assessment, we cover automatic speech recognition accuracy, intent understanding capabilities, and safety evaluation. We also present prominent evaluation frameworks that move beyond single-metric approaches. These include zero-shot no-reference quality assessment frameworks combining semantic, spatial, and geometric features, intrinsic faithfulness benchmarks evaluating physical plausibility and commonsense reasoning, disentangled frameworks separating aesthetic and technical quality dimensions, and user perception-based models incorporating functional, emotional, and social values. We survey relevant datasets including those focused on head quality, talking head video quality, full-body digital human quality with subjective ratings, and various video quality assessment datasets, while identifying critical gaps such as predominant focus on heads rather than full bodies, artificial distortions failing to represent authentic generative model artifacts, and limited audio-visual paired data. Finally, we propose future directions for comprehensive digital human evaluation from two core dimensions—quality and content. For quality assessment, we advocate multimodal large language model-based architectures integrating video, audio, and textual modalities through unified encoding, followed by multi-stage training strategies incorporating human preference alignment to produce objective scores that correlate strongly with subjective perception. For content assessment, we highlight the need for dedicated anomaly detection to identify generative artifacts including anatomical inconsistencies, identity and clothing temporal instability, as well as specialized metrics for motion plausibility and speech naturalness. We emphasize that unified evaluation frameworks should move beyond simple aggregation of single-modality metrics toward holistic single-model assessment that captures cross-modal synergies and outputs human-aligned comprehensive scores. Collectively, this survey provides a structured reference for researchers and practitioners, identifies critical research gaps including limited full-body datasets, insufficient generative distortion coverage, lack of explainability in evaluation results, and the absence of unified industry standards. It outlines pathways toward objective, human-aligned, and comprehensive digital human evaluation systems to support technological iteration, informed user selection, and sustainable industry development in this rapidly evolving field.

Keywords digital human; evaluation method; large language model; multimodal learning; human-computer interaction

1 引言

数字人是指基于现实人类特征进行建模，通过计算机图形技术生成，并依靠真人动作捕捉或算法驱动实现动态表现，最终在多媒体设备上展示的数字化虚拟人物形象[1]。依据不同的分类标准，数字人可被分为不同的种类[2]。根据图形维度，虚拟数字人可划分为二维与三维两类。二维虚拟数字人以平面图像为表现载体，其形象可通过传统手绘或数字绘图软件构建；而三维虚拟数字人则需借助三维建模技术呈现，具备更强的立体感与真实感。从交互模态的角度，数字人可分为单模态与多模态两种类型。基于单一模态的数字人仅依赖某一种输入或输出通道（例如文本、语音或视频）与用户进行交流。相比之下，多模态数字人的核心在于能够协同利用语音、视频和文本等多种模态，以实现更自然、真实的交流体验。

根据国际数据公司发布的《2025 年中国数字人产业发展新洞察》[3]的数据，2025 年中国 AI 数字人核心市场规模约 339.2 亿元，预计 2030 年将达到 935.6 亿元。中国互联网协会发布的《中国数字人发展报告（2024）》[4]显示，2024 年的中国数字人相关企业总量超 4500 家，近五年每年新增数字人企业数量总体呈现上升趋势，其中 2024 年新注册企业数量超 1200 家。随着人工智能技术的突破，应用领域的不断拓展，数字人具有巨大的发展潜力[5]。根据 Business Research Insight 的报告[6]，全球智能虚拟数字人市场有望大幅增长，2026 年规模为 974.6 亿美元，预计到 2035 年将达到 19207.6 亿美元，2026 年至 2035 年复合年增长率为 34.73%。

以深度学习和生成式人工智能为代表的技术，极大地增强了数字人在图像生成、语音合成、内容生成等领域的能力。例如：多模态大模型、情感计算技术极大地增强了数字人的交互能力；深度学习、生成模型等技术的发展赋予数字人更加生动鲜活的形象。在新技术的加持下，数字人的各方面性能远超过去，也展现出更多新特性。随着数字人功能的拓展与类别的增加，当前行业尚未形成统一的评估指标体系。这不仅导致用户的选择缺乏客观、标准化的决策依据，同时也在很大程度上制约了数字人的技术迭代与产业发展的可持续性。

目前，数字人领域相关标准数量有限，因为制定时间较早，无法有效反映当前数字人的综合性能。首先，缺乏针对性指标。数字人产品虽然种类繁多，但在基础架构上基本一致，传统评测方法难以区分数字人产品在关键特性上的优劣。随着技术的进步，数字人会出现更多特性，新的评测方法还应具有一定的可扩展性。其次，存在指标公平的问题。评测指标应尽可能客观、公平地反映不同数字人的性能，其结果应具有稳定性和可重复性。许多标准和指标都基于统计评测人员的评分得到，因此人员的选择会很大程度改变最终结果。并且，个人主观成分会不可避免地影响性能指标，进而削弱公平性。此外，在指标设计时对不同功能的重视程度也会导致结果公平性不足的问题。再次，目前尚未提出统一的评测标准。目前的行业标准对评测内容的描述不够准确，许多评测项目只提出了最低标准。而相关指标在单一方面能反映性能差异，由于指标计算方法的差异和缺少系统性的规划，指标的简单叠加获取总性能指标往往无法取得理想的效果。上述只是数字人评测的部分问题，和其他模型评测指标类似，更多维度和更深层的考察也是数字人评测发展的方向。因此，在数字人技术和应用进步的同时，相关评测方法也要与之同步。

本文以数字人评测的方法与指标为研究对象，在系统梳理现有评测指标的基础上，提出了数字人评测框架的综合展望。全文章节安排如下：第二章对数字人的技术架构进行说明，为后续介绍评测奠定理论与技术基础；第三章对数字人领域已有的行业标准及进行回顾与总结；第四章从质量与内容两方面分析了数字人单一评测指标，同时对性能评测框架及相关数据集展开介绍，剖析当前研究的局限性；第五章展望了数字人的综合评测方法，以期提供参考；第六章对全文进行总结。

2 数字人技术架构

根据现有对现有数字人研究的总结分析，本文对数字人技术架构的总结如图 1 所示。

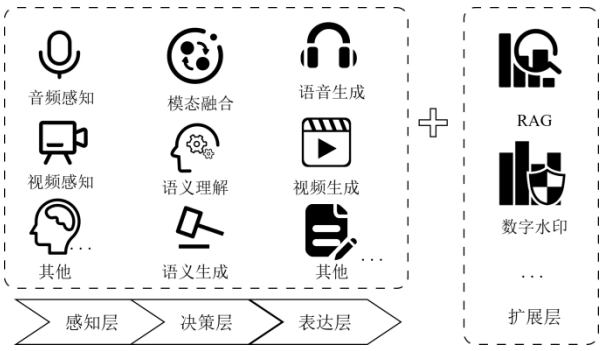

图 1 数字人技术架构

2.1 感知层

在数字人系统中，感知层是实现数字人与用户实时交互的关键组成部分，接收用户各种模态的信息。该模块主要包含语音感知、视频感知和其他感知。语音感知通过 ASR（Automatic Speech Recognition）技术，将用户语音输入转化为文本。近年来，随着深度学习的发展，Dahl 等[7]首次将深度神经网络替代传统方法，应用于隐马尔可夫声学模型训练，证明了深度学习在语音特征提取的优势。以 Wav2vec 2.0[8]为代表的自监督学习和以 Whisper[9]为代表的弱监督大规模预训练，极大地降低了对人工标注数据的依赖，提升了语音识别技术的鲁棒性。而目前主流方法全面转向基于 LLM（Large Language Model）的端到端架构，Kai 等[10]提出了 Dolphin 模型，通过离散化视觉编码和全局-局部注意力机制，在保证分离质量上的同时，大幅提高了计算效率。

视觉感知主要集中在用户的表情和动作识别。表情识别不仅要求捕捉人脸位置，还要能检测面部表情实现情绪分析。TransFace++[11]在 TransFace[12]的基础上，创新性地直接使用图像文件的原始字节作为模型输入，从而保护了用户隐私。表情检测方面，Lv 等[13]基于变分推理建模表情的潜在概率分布，提出 VICH（Variational Inference-based Classification Head）与双分支多尺度融合架构，有效缓解表情标签模糊与类别重叠问题，显著提升复杂场景下人脸表情识别的精度与鲁棒性。ExpLLM[14]将 LLM 链式思维专门用于面部表情识别的框架，解决了传统方法只输出标签、缺乏推理依据、微表情和复合表情识别弱的痛点。动作识别主要分为基于 RGB 视觉的方法、3D 骨架识别方法和多模态动作识别方法。InternVideo2[15]采用三阶段渐进式预训练范式，通过掩码自监督与多模态对齐预训练实现高效时空特征建模，是当前基于 RGB 纯视觉动作识别领域的领先方法。Yang 等[16]通过表达性细粒度关键点与渐进式骨骼演进策略，在提升骨架动作识别精度的同时大幅降低计算量。Feng 等[17]通过主动模态选择与双向互知识蒸馏，在少样本动作识别任务中实现 RGB、骨架等多模态信息的高效协同，显著提升遮挡与模态缺失场景下的识别精度与泛化能力。其他感知则采集其他模态的数据，例如脑电图和眼动图等。Mahnoosh 等[18]设计了跨被试、跨数据集域适应策略，对齐不同被试的尺度不变特征分布，从根源缓解 EEG（Electroencephalogram）个体差异导致的泛化瓶颈。Wang 等[19]利用脑电图和眼动图作为互补信息，提出了受限对抗网络，实现了跨模态情感识别的先进性能。

2.2 决策层

决策层负责融合、理解感知层的多模态交互信息并生成语义，其主要组成为多模态融合层和语义生成层。多模态融合旨在通过联合建模视觉、听觉等多模态信息，利用不同模态间的互补性与一致性，从而提升情绪识别与用户感知的精度与泛化能力。多模态融合的策略可分为三类：特征级融合、模型级融合和决策级融合。特征级融合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初级特征进行融合，形成一个综合的表示作为模型的输入。常见的特征级融合方法包括拼接、加法、乘法和双线性融合等。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，但要求特征的维度和尺度匹配。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，能够较好地捕捉到不同模态之间的交互信息，并且保留了一定的空间结构和语义关联。特征级融合的特征选择还可通过预处理方法筛选，例如主成分分析[20]和贝叶斯估计[21]等。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[22]。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。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公共的语义空间中。基于神经网络的融合方法利用神经网络将不同模态的数据进行融合。常见的方法包括基于生成对抗网络和基于注意力的方法。Chahi 等[23]基于重建驱动与语义引导自监督，提出了 R2GAN（Reconstruction-guided Generative Adversarial Network），解决多模态融合无标注数据

的难题。TouchFormer[24]采用模态-自适应门控机制以及模态内和模态间注意力机制，自适应整合跨模态特征，增强模型的稳健性。模型级融合可以减少模态间的不匹配问题和维度差异问题，提高模型的鲁棒性和泛化能力。然而，该方法需要更多的计算资源和存储空间，每个模态都需要单独的网络结构和参数，对多个模型的联合优化存在困难。决策级融合将每个模态单独的决策结果进行聚合，得出最终的决策结果[25]。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。投票法将多个模态的独立决策结果进行投票，选择获得最高票数的选项作为最终的决策。Wang 等[26]提出了一种无需训练的参数自由方法，通过检索增强投票直接从视觉-触觉数据中构建跨模态知识。在无需训练的情况下，达到与大规模跨模态模型相当的性能；且数据质量越高，性能越好，计算成本为零增长。加权平均法将不同模态的决策结果按照一定权重进行加权平均，得到最终的决策结果。此外，根据不同的应用场景灵活地组合使用上述融合方式，以更好地利用不同模态信息，从而提高了融合结果的综合性和鲁棒性。

语义生成层是实现语义表达与实时交互的关键环节。目前，依托 LLM 的各项能力，数字人可以具备情感识别、实时交互、知识推理与指令生成等功能。Lan 等[27]首次提出将 ChatGPT 应用于数字人语义生成，使其具备独立学习和实时对话的能力，彻底改变了传统数字人依靠预设脚本的生成方式。Lai 等[28]提出了多模态大模型架构，联合处理语音内容、声学特征和说话人情绪，并利用 VQ-VAE（Vector Quantized Variational AutoEncoder）框架中的实例归一化和自适应归一化实现身份解耦，生成高保真的聆听头部反应，支持灵活的身份控制。Lin 等[29]构建了基于 LLM 的三维人类任务助手系统，能够自主选择、应用和解释多种专用工具（如 3D 姿态估计、形状重建、接触检测等），利用检索增强生成和学术出版物指导 LLM 理解工具用法，在多项三维人类任务上超越现有模型。

2.3 表达层

表达层旨在通过数字人的多种模态表达决策层的信息。语音合成和视觉生成模块分别从听觉与视觉层面提升表达的自然度与感染力。语音生成模块采用 TTS（Text To Speech）技术将决策层输出的语义信息转化为真实、自然的语音，从而实现数字人与用户的实时语音交互。近年来，以扩散模型和流匹配为代表的生成机制在 TTS 中取得了进展，流程对比如图 2。StyleTTS2[30]结合了风格扩散、对抗训练和端到端波形生成，达到人类级自然度，但在实时交互场景中仍存在明显时延瓶颈。Diff-TTS[31]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。该方法采用 DDIM（Denoising Diffusion Implicit Models）加速和限制扩散步数，使得推理时延 RTF（Real-Time Factor）降低到 0.035，但音频质量有待提升。基于扩散模型的 TTS 技术具有高稳定性、情感可控与高保真等优点，但是推理需多次迭代，时延较高。流匹配是通过学习从先验分布到目标数据分布的连续向量场，能够在极少推理步数下保持高生成质量，从根本上缓解扩散模型固有的步数与性能权衡问题。因此，目前更多研究转向了基于流匹配的 TTS。Matcha-TTS[32]是一种基于 OT-CFM（Optimal Transport Conditional Flow Matching）的非自回归文本到语音声学模型，以梅尔频谱为生成目标，通过轻量化编码器-解码器架构单调对齐策略实现文本与语音序列的对齐，并利用最优传输流匹配学习从高斯先验到目标语音分布的变换向量场。Chen 等[33]提出了 F5-TTS（Fairytaler that Fakes Fluent and Faithful Speech with Flow Matching），一种基于全非自回归流匹配（Flow Matching）和 DiT（Diffusion Transformer）的 TTS 系统。该方法使用 ConvNeXt[34]细化文本表示，以易于与语音对齐，并且通过推断时间摇摆抽样策略显著提升了模型的性能和效率，实现了 0.15 的 RTF。在实际应用时，扩散 TTS 凭借细腻的韵律与风格表达能力仍具优势；而在实时交互、端侧部署、流式合成等对延迟敏感的实际应用中，流匹配 TTS 凭借更低时延、更快推理速度与相当的生成质量，也是当前数字人语音生成的主流技术路线。

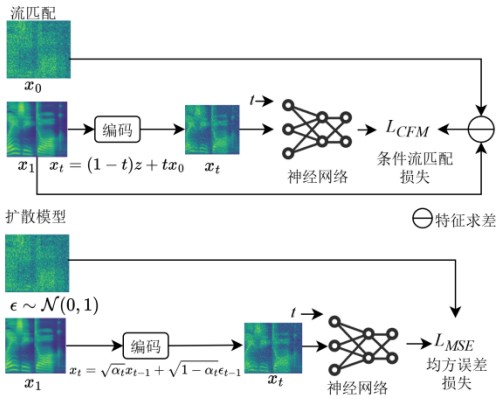

图 2 扩散与流匹配 TTS 对比

视觉生成模块的功能是将决策层的语义信息通过数字人的视觉模态传达。本文将从数字人建模、驱动与渲染三个关键环节进行介绍。数字人根据视觉形式可分为二维数字人和三维数字人，在建模方式上也存在差异。二维数字人具有建模高效、交互时延低等优点。目前，GAN（Generative Adversarial Network）和 Stable Diffusion 等深度学习方法逐渐成为主流，例如：StyleGAN[35]、ShowMaker[36]和 ClipFaceFusion[37]。三维数字人是指在计算机三维空间中，通过数字化技术构建、具备人体外形特征和可实现运动与表情，并能在虚拟环境中呈现与交互的拟人化虚拟角色。如图 3，构建三维数字人的步骤包括四个阶段：首先通过显式或隐式表征构建数字人三维几何形态；运动学绑定为静态几何模型构建运动控制架构的过程，核心任务是建立底层几何变形与高层控制参数之间的可微映射关系；驱动将外部控制信号转换为绑定系统中运动参数；最后根据几何模型、材质属性和运动参数与虚拟光照，通过求解渲染方程生成二维图像序列的过程。

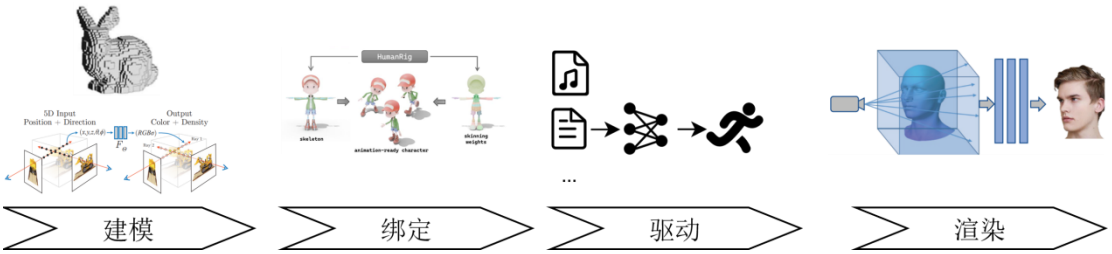

图 3 三维数字人构建流程[38-41]

三维数字人建模方法根据几何表示的形式分为显式、隐式和融合式，表 1 对不同类别的三维建模技术进行了对比。显式模型即物体的三维结构被直接给出或通过参数映射的方式表示，如：点云、多边形网络和体素网格等。3DGS（3D Gaussian Splatting）[50]是一种显式的 3D 场景表示方法，将场景建模为数十万至数百万个可学习的高斯椭球，通过可微分光栅化实现实时渲染，具有实时性和可编辑性等优点。目前主流的显式表示三维建模方法还有：SplatTouch[42]、MixedGaussianAvatar[51]和 Speedy-splat[44]等。隐式模型通过连续函数约束三维空间实现外观塑造，如：Nerf（Neural Radiance Fields）[38]、NeuS2（High-fidelity Neural Surface 2）[45]和 SDFusion[46]等。Nerf[41]将场景的几何结构与表面信息编码为一个连续的 5D 辐射场函数，通过沿采样投影射线采样 5D 坐标（位置和观察方向）进行体积渲染，并施加时序一致性约束，最终生成高保真、时序连贯的动态视频序列。Hi-NeuS[52]是一种无需多视图对象掩码的神经隐式表面重建框架，通过累积多视图渲染权重分布生成自监督信号，并结合全局几何约束优化 SDF（Signed Distance Field），实现更紧凑、更精确的三维表面重建，显著降低表面噪点并提升细节保真度。目前，融合显式直观性与隐式鲁棒性的三维建模方法正成为主要发展方向。NeuralGS[53]结合了 3DGS[50]与 NeRF[38]的混合压缩与表示框架的核心是用紧凑神经网络替代大量离散高斯参数，增强了 3DGS[50]的压缩、泛化与可控性，同时保留实时渲染优势。D3-Human[47]基于 SMPL-X（Skinned Multi-Person Linear Model eXpressive）显式身体建模与 hmSDF（Human Manifold Signed Distance Field）隐式衣物解耦建模的动态数字人方法，可从单目视频中重建出时序一致、可独立动画与换装的人体与衣物网格。

三维数字人绑定是为静态的三维几何模型建立参数化的运动控制与形变映射机制，使模型能够依据驱动信号产生符合预期的动态形变与运动。三维数字人绑定领域从传统方法转向基于自回归生成、隐式表征学习和自监督微调的框架。ControlFace[54]通过双分支 U 形架构在无需逐个体微调的前提下，实现对输入人脸图像的姿态、表情和光照等属性的灵活高保真编辑并保持身份细节。Guo 等[55]提出了 Make-It-Animatable 架构，能够快速为任意姿态和形状的 3D 人形模型生成高质量的骨骼、蒙皮权重及姿态变换，支持非标准骨骼结构。三维数字人驱动技术总体分为单模态驱动和多模态驱动。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。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。SimAvatar[56]是一种文本生成、可物理仿真的穿衣三维高斯数字

人，衣物褶皱、发丝飘动真实，兼容标准物理引擎。此外，相关工作还有：生成合符角色人设的专属动作风格的 PersonaBooth[57]、提升复杂动作生成的 StickMotion[58]和实时渲染的 DreamWaltz-G[59]等。

类别	显式方法	隐式方法	融合方法
主要思想	以点、体素、面片等显式结构定义物体	用连续函数或神经网络间接描述三维空间	显式结构与隐式函数结合
拓扑表达	拓扑固定，变形、拓扑变化困难	支持任意拓扑变化，细节难以精确控制	兼顾宏观结构精度与微观细节连续性
布尔运算	布尔操作直观，复杂场景鲁棒性差	仅场函数运算，鲁棒稳定	在保证结构可控的同时提升布尔稳定性
渲染效率	光栅化硬件友好，渲染快	需光线步进或球追踪，渲染较慢	显式部分快速渲染，隐式部分按需采样
内存占用	高细节模型体积大，面片数量多	存储紧凑，参数远少于面片数	存储适中，兼顾结构参数与场参数
代表方法	SplatTouch[42]、Matrix3d[43]和 Speedy-splat[44]等	Nerf[38]、NeuS2[45]和 SDFusion[46]等	D3-Human[47]、H3R[48]和 GEOcc[49]等

表 1 三维建模技术对比

音频驱动即利用语音驱动数字人生成面部表情和动作。Audio2Mesh[60]是一种端到端的音频驱动模型，利用深度神经网络将音频映射到面部网格的 3D 顶点坐标，同时利用潜在情感编码以消除音频无法体现的表情变化。Audio2Lip[61]利用大量音视频数据学习语音与面部运动表征，基于语音编码驱动的 2D 肖像动画重建 3D 面部动画，降低了对高精度扫描数据的依赖，提升了模型对中英多语言的泛化能力。目前，研究者从多个方面优化音频驱动三维模型，例如：基于扩散模型的 StreamingTalker[62]、融合动作及表情的 UniTalker[63]和 FaceDiffuser[64]等。基于视觉的实时驱动方法利用视频流中提取的动作、表情与姿态信息，驱动数字人实现自然交互。Lu 等[65]提出截断时间步调度和掩码混合训练统一姿态，将视频二维的关键点姿态优化为自然的 SMPL-X 全身姿态，解决驱动中的漂移与不自然问题。由于鲁棒性、自然度、时序一致性与情感表达能力的优势，多模态融合驱动逐渐成为研究方向。

渲染是指将构建好的三维数字人模型，通过计算机图形学算法，生成具有真实感或特定艺术风格的二维图像或视频序列的过程。由于数字人可编辑性和实时性的要求，渲染技术逐渐转向以 3DGS[50]为核心的显式表征。3DGUT[66]通过用无迹变换替代传统的 EWA（Elliptical Weighted Average） Splatting 算法，在保持实时渲染效率的同时，首次原生支持鱼眼相机、滚动快门等非线性失真相机模型以及反射/折射等二次光线追踪效果。Gao 等[67]通过将场景元素表示为位置(三维)、时间(一维)和视角方向(三维)的七维高斯分布，首次在单一框架内统一建模空间几何、时间动态和视角依赖外观。

2.4 扩展层

在数字人的总体架构中，拓展层是对数字人基本功能的补充与拓展。由于该层的方法种类众多，本文仅对其中部分技术介绍。RAG（Retrieval Augmented Generation）技术是一种将外部知识检索机制与大语言生成模型相融合的混合架构，旨在解决数字人系统在开放域交互中面临的知识固化与事实性幻觉问题。随着多模态 RAG 框架的引入，数字人可在视觉问答任务中有效平衡外部知识利用与推理鲁棒性[68]。数字水印则从安全层面对数字人进行了保障。例如，在数字人建模阶段进行鲁棒水印的嵌入，即使经历 3D/2D 失真的攻击，也能保证版权信息可从 3D 高斯参数和 2D 渲染图像的可靠提取[69]。在渲染输出阶段加入水印则可以直接保护生成内容的版权，只需相关图片或视频即可提取，无需原模型[70]。总之，以上仅是扩展层的部分功能，扩展层还可以从多智能体协作、模型规模和应用场景优化等方面设计以增强数字人的功能。

3 数字人行业标准

随着生成式人工智能技术的发展，数字人在智能化、多样化、仿真化得到了很大进步，也因此更多地参与到生产生活中。全面的评测能更好了解数字人的真实性能，有利用户选择，明确了数字人的发展方向。因此，相关组织为了保证数字人行业健康发展和产品的基本质量，提出了行业标准，规定了数字人的评测维度。下面将介绍主流的行业标准，并分析现有标准的不足。

3.1 ITU-T 数字人标准

ITU-T F.748.14[71]（《非交互式二维真人数字人

应用系统的要求和评估方法》，以下简称“非交互标准”）与 ITU-T F.748.15[72]（《数字人应用系统的框架和指标》，以下简称“系统标准”），均为国际电信联盟电信标准化部门（International Telecommunication Union-Telecommunication Standardization Sector, ITU-T）针对数字人系统制定的标准规范。该两项标准从图像、语音、动画及多模式输入/输出等多个维度，构建了数字人系统的客观评估指标与主观评估指标体系。相较于 ITU-T F.748.14，ITU-T F.748.15 作为适用于所有数字人系统的评测标准，进一步纳入了针对三维数字人的评估指标以及交互处理相关指标。此外，中国通信标准化协会于 2023 年发布了行业标准 YD/T 4393.1—2023《虚拟数字人指标要求和评估方法第 1 部分：参考框架》与 YD/T 4393.2-2023《虚拟数字人指标要求和评估方法第 2 部分：2D 虚拟真人形象类产品》，其核心内容与 ITU-T 制定的上述标准基本一致。在视觉质量层面，“非交互标准”明确了二维数字人需规避的问题类型，具体包括失真、马赛克、明显跳帧及其他画面破坏现象。针对三维数字人，该标准则围绕三维模型的角度提出了需关注的问题，涵盖严重变形、严重穿模、未连接的点、被破坏的表面以及其他破坏情况。“系统标准”进一步补充了可选评估指标，即精细度与画面舒适度。其中，精细度为针对三维模型的评估指标，其评定需依据模型面数、面部细节、面部布线结构及人像分辨率等参数展开对比分析；画面舒适度则属于主观评估指标，用于表征用户对数字人画面的主观舒适感受程度。该指标的具体测评方法为收集用户通过李克特量表完成的主观评分，评分内容为用户对“你喜欢这个画面的设计吗？”等问题的打分结果（评分范围为 1 至 5 分，分别对应从最差到最好的评价等级），详见表 2。

ITU-T 的两标准在语音方面的指标分为：语音正确率、韵律正确率和语音舒适度。语音正确率的是指数字人的文本合成语音的性能。语音错误包括发音缺失、发音过度、音素错误和语调错误。单词发音正确率可通过如下公式计算：

$$R_{pwc}=(1-\frac{E_w}{N_w})\times 100\% \tag{1}$$

其中，R_{pwc} 表示单词发音正确率，E_w 表示发音错误的单词数量，N_w 单词总数。R_{psc} 表示单词发音正确率，E_s 表示发音错误的单词数量，N_s 单词总数，语句发音正确率可通过如下公式计算：

$$R_{psc}=(1-\frac{E_s}{N_s})\times 100\% \tag{2}$$

韵律包括停顿句、音高、音长和音量等。韵律正确率表明系统语音合成的性能，计算公式如下：

$$A_r=\frac{N_c}{N}\times 100\% \tag{3}$$

其中 A_r 是韵律正确率，N_c 是正确停顿的样本数量，N 是测试样本的总数。

语音舒适度指数字人合成语音给用户带来的生理舒适感受程度。该指标同样属于主观指标，因此其获取方法与画面舒适度基本一致，后续主观指标的获取方法亦与此相同，不再另行阐述。

动画维度的指标分为动作契合度与主观层面的动作舒适度。其中，动作契合度用于衡量数字人动作与当前语境的匹配程度，其评估需综合考察嘴唇、眉毛、眼皮及眼球等部位的动作表现（例如，嘴唇动作需同时与语音信息及所表达的情绪相匹配）。

在交互指标方面，由于“系统标准”的内容更为全面且已涵盖“非交互标准”的相关要求，故下文仅针对“系统标准”进行介绍。交互处理部分明确了数字人多模态输入的类型限定，具体包括文本、语音、图像及触觉。如图 4 是交互指标的分类，其中主要指标是语音识别正确率，分为单词和语句识别正确率，两者计算方式类似。单词正确率计算如下：

表 2 图像舒适度主观评分规则[72]

评价维度	问题描述	5	4	3	2	1
舒适度	你喜欢这张图的设计吗？	非常喜欢	比较喜欢	一般	比较不喜欢	完全不喜欢
自然度	这张图片自然吗？	非常自然	比较自然	基本自然	比较不自然	完全不自然
视觉感受	你愿意接受这个画面的服务吗？	非常愿意	比较愿意	一般	比较不愿意	完全不愿意

$$R_{wrc}=(1-\frac{E_w}{N_w})\times 100\% \tag{4}$$

其中 R_{wrc} 为单词识别正确率，E_w 为错误识别的单词数量，N_w 为单词总数。多模态输出的基础指标包括视频帧率、画面正确率、音视频匹配度及输出类型。其中，视频帧率用于表示画面的流畅程度；画面正确率指固定帧率视频中画面的准确程度，即画面出现跳帧、卡顿等不准确现象的程度；音视频匹配度用于衡量数字人输出的画面内容（如人物口型）与音频内容的匹配程度；输出类型指输出支持的硬件，则涵盖手机、投影及 VR（Virtual Reality）等。此外，多模态输出指标还包含可选指标实时系数，其具体定义为视频合成时间与输出视频时间的比值。

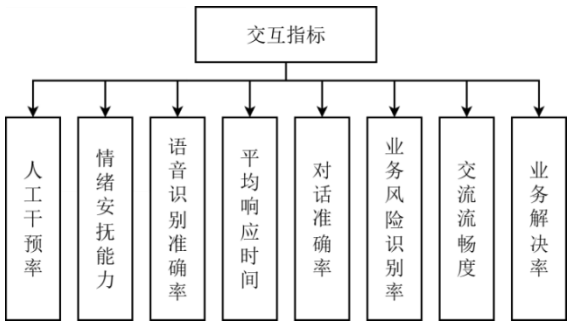

图 4 交互指标的组成

3.2 IEEE数字人标准

IEEE 标准活动委员会（IEEE Standards Activities Board Standards Committee）于 2023 年发布了《IEEE 标准：评估数字人类质量的框架》（IEEE Standard for a Framework for Evaluating the Quality of Digital Humans）[73]，该标准主要对数字人的评估方法做出规定，其评估框架采用主观评估与客观评估相结合的方式。

主观评测通过招募一组评测人员体验数字人服务，并采用多种方法对数字人服务质量进行评估，具体方法包括问卷法、排名法、配对比较法、分级法及评论法。其中，问卷法的题目设计围绕数字人的预期表现或属性表现展开，由评测人员完成作答；用于评估数字人质量的问卷分为简单问卷与评分问卷两类，简单问卷的题目仅设置“是”或“否”两个选项，评分问卷则采用更细化的评分形式，如 1-5 分制或 0%-20%-40%-60%-80%-100%梯度百分比制。排名法要求评测人员基于对数字人的实际体验，对若干维度的指标进行优劣排序。配对比较法用于体现两个数字人模型在质量表现上的差异。排名法包含三种评分方式：其一为标签评分，即根据预设标签分配对应分值，例如 A：95、B：85、C：75、D：65、F：0；其二为成功/失败评分，该方式类似标签评分，但仅设“成功”与“失败”两个标签并对应分配分值，例如“成功”计 100 分、“失败”计 0 分；其三为分数评分，由评测人员在 0 至 100 分的区间内直接为评估对象分配分数。评论法指评测人员对数字人内容质量进行描述性评价，并提供相应的示例与证据支持。

客观评测基于图像质量指标构建评估方法，所采用的核心指标包括结构相似性指数（Structure Similarity Index Measure, SSIM[74]）与弗雷歇初始距离（Fréchet Inception Distance, FID[75]）。客观评测不仅针对数字人画面的视觉质量展开考察，同时也对数字人的形象、动作等维度进行评估。

IEEE 标准明确了数字人评估流程的具体步骤：在测试实施前，需明确应用场景、评估目的、数字人内容等基础信息；随后结合主观与客观评估方法设计评估框架，并为两类评估方法分配相应权重。测试流程分为预测试与正式测试两个阶段：预测试通过开展小规模简化评测任务，收集评测人员的意见反馈，用于修正模糊问题并优化评测流程；正式测试阶段需招募涵盖不同年龄、性别及文化背景的评测人员，并对其进行评测培训。最终测试结果综合主观评估与客观评估数据形成，同时生成包含问题分析及优化建议的评估报告。

3.3 UWA 数字人标准

《3D 数字人质量分级技术要求》[76]是世界超高清视频产业联盟（UHD World Association, UWA）于 2023 年发布的关于三维真人形象数字人视觉与交互效果的分级规范。该文件以用户体验为核心导向，将三维数字人的评价体系划分为角色效果、识别感知、互动决策三个维度的指标参数。其中，角色效果指用户可感知的数字人形象特征，主要涵盖视觉与听觉两大范畴；识别感知体现数字人识别用户及处理外部环境信息的能力，具体可通过语音转文字准确率、人脸识别率、情绪识别准确率等指标衡量；互动决策反映数字人与用户进行“自主”互动的能力，相关评估指标包括对话交互完成率、表情反馈正确率和肢体反馈正确率等。图 5 展示了三维数字人评价维度及具体指标。

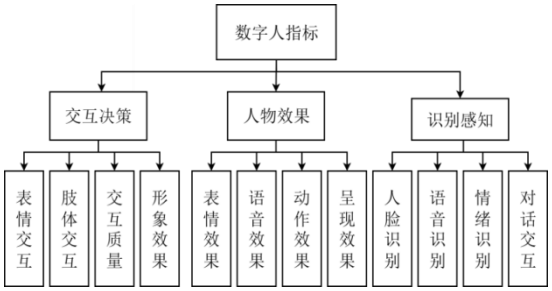

图 5 UWA 三维数字人质量分级技术要求的评价维度及指标

3.4 现有标准的不足

随着数字人技术的进步，现有评测标准仅能作为产品的基础要求，已难以全面且有效地反映不同数字人产品之间的性能差异。这一现状导致在数字人应用落地与技术迭代改进过程中，均缺乏客观依据支撑。以下对现有行业标准存在的问题进行简要分析。

首先，前述标准虽对数字人的多项属性提出了要求，但存在两方面核心问题：其一，未明确界定各属性在最终综合评价中的权重占比，无法体现不同属性对数字人整体性能的影响；其二，缺乏细化的量化指标及统一的计算方法，导致不同数字人产品之间难以开展横向对比分析。

其次，现有标准普遍采用主观评价方法，其中 IEEE 标准尤为侧重评测人员的主观意见，这直接导致评估结果的公平性与客观性存疑。尽管数字人的服务对象为人类，主观意见具有重要的参考价值，但主观意见的采集过程存在明显的不确定性，易受内外部多种因素干扰。例如，评测人员的个人偏好会直接影响其对数字人的主观感受，进而为评测结果引入系统性偏差；评测问题的表述内容与排列顺序可能对评测人员形成引导，干扰其独立判断，最终影响评测结果的真实性。此外，主观评价的实施需招募大量评测人员参与实验，耗时费力且成本较乏，本文将围绕数字人产品的基础特性，梳理并筛选相关评价方法及相关数据集，以期为数字人产品的评估实践提供参考依据。

最后，尽管部分标准提出了客观指标以量化评测结果，但相关指标的计算方法过于简化，难以真实反映用户的实际使用体验。同时，前述标准在视频、音频等专业领域的评价方法上，未能充分借鉴该领域的专业知识与现成方案，导致评测结果缺乏足够的专业支撑与可解释性。不仅如此，客观指标的准确计算依赖于大量的输入样本，但现有标准未对样本的数量规模与分布做出明确要求，进一步影响了客观评测结果的可靠性与可比性。

4 数字人技术指标与评测框架

鉴于行业标准在数字人产品评价中存在若干不足，且当前针对数字人评测方法的专项研究较为匮乏，本文将围绕数字人产品的基础特性，梳理并筛选相关评价方法及相关数据集，以期为数字人产品的评估实践提供参考依据。

4.1 数字人质量评价指标

4.1.1 视频指标

视觉作为数字人信息输出的主要渠道，流畅且准确的视觉质量是提升用户基础体验的关键，因此本文首先对数字人的视觉性能指标展开阐述。视频质量指标可划分为主观指标与客观指标两大类。

主观指标通过评测人员打分获取主观平均得分（Mean Opinion Score, MOS）或平均主观得分差异（Differential Mean Opinion Score, DMOS）。由于该方法需耗费大量人力与时间成本，目前其获取方式多依赖于主观评价数据集。Fine-VQ[77]数据集包含 6104 段来自 B 站点播与直播的真实无合成失真短视频，覆盖多分辨率与全种类场景，拥有超过 80 万主观评分、色彩/噪声/伪影/模糊/时序/整体六维细粒度 MOS 标注及 12 类真实失真标签，相比传统仅单一总分数据集，可支持视频质量归因、失真定位及细粒度无参考 VQA 算法的训练与评测。

客观指标指依据预设特征或方法计算得出的指标，其可靠性验证取决于与主观评分的一致性程度。常用的一致性验证评价指标包括皮尔森线性相关系数（Pearson Linear Correlation Coefficient, PLCC）、斯皮尔曼等级相关系数（Spearman Rank-order Correlation Coefficient, SRCC）、肯德尔等级相关系数（Kendall's Rank Correlation Coefficient, KRCC）及均方误差（Root Mean Square Error, RMSE）。其中，PLCC 用于计算预测结果的线性相关性；SRCC 反映预测结果的单调性；KRCC 着重关注预测结果的顺序一致性；RMSE 则是计算预测结果的平均差异。它们的计算方式如下：

$$PLCC=\frac{\sum_{i=1}^{N}(s_i-\bar{s})(p_i-\bar{p})}{\sqrt{\sum_{i=1}^{N}(s_i-\bar{s})^2\sum_{i=1}^{N}(p_i-\bar{p})^2}} \tag{5}$$

$$SRCC=1-\frac{6\sum_{i=1}^{N}d_i^2}{N(N^2-1)} \tag{6}$$

$$KRCC=\frac{2(N_C-N_D)}{n(n-1)} \tag{7}$$

$$RMSE=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(s_i-p_i)^2} \tag{8}$$

式(5)，(6)，(8)中，s_i 和 p_i 分别表示第 i 个视频的主观质量分数及客观质量分数，$\overline{s}$ 和 $\overline{p}$ 分别表示主观质量分数平均值和客观质量分数平均值；N 表示样本数量；d_i 表示第 i 个视频主观质量分数排名与客观质量分数排名的差值。式(7)中，N_D、N_C 分别表示一致对和不一致对的数量，N 为样本量。

客观指标依据其对原始视频的依赖程度[78]，可划分为全参考指标（Full-Reference, FR）、半参考指标（Reduced-Reference, RR）与无参考指标（No-Reference, NR）。其中，半参考模型则从样本的时频域特征进行评价，代表方法有 SpEED-QA[79]等。

结构相似性指数（SSIM）是典型的全参考指标，其从亮度、对比度及结构三个维度对两幅图像的相似性展开评估。该指标既可用于衡量单幅图像的失真程度，亦可用于比较两幅图像间的相似程度。峰值信噪比（Peak Signal-to-Noise Ratio, PSNR[80]）则通过计算两幅图像像素值的差异来衡量失真程度，同样可应用于图像相似性的比较。上述两种指标均基于图像的视觉参数计算相似性，具备可解释性强的优势，可被应用于数字人视觉质量评估领域。但二者亦存在一定局限性，例如：SSIM 无法适应位移、缩放、旋转等非结构性失真，而 PSNR 的评估结果与人眼的真实主观感受存在偏差。

鉴于此，研究者结合深度学习模型提出了一系列更符合人类视觉感知的指标。FID 是其中之一，用于评估生成图像与真实图像之间的差异。与传统指标不同，FID 并非直接计算图像样本本身的差异，而是通过模型提取的特征来衡量差异，通常采用预训练的网络实现，具体过程为：

$$d^2((\mu_r,\Sigma_r),(\mu_g,\Sigma_g))=\left\|\mu_r-\mu_g\right\|^2+ Tr(\Sigma_r+\Sigma_g-2\sqrt{\Sigma_r\Sigma_g}) \tag{9}$$

其中 μ_r 和 μ_g 分别是真实样本和生成样本特征的均值，Σ_r 和 Σ_g 是方差，Tr 表示矩阵的迹，$\|\cdot\|$ 表示向量的二范数。FID 是图像的评价指标，而 FVD（Fréchet Video Distance）[81]是基于同样方法设计的视频评价指标，将特征提取网络替换为视频模型。

与 FID 的设计思路相似，LPIPS（Learned Perceptual Image Patch Similarity）[82]同样是基于特征提取的全参考指标。由于更侧重于贴合人类实际的视觉感知体验，LPIPS 采用 VGG[83]、AlexNet[84]等浅层模型来提取图像的多层特征。在指标计算阶段，需先对每层特征进行归一化处理，再计算各层特征的 L2 距离，最后通过对各层距离取平均值得到最终的 LPIPS 指标值。

相较于 SSIM 等传统指标，FID、FVD 及 LPIPS 这类基于深度学习模型的指标在多个维度均实现了性能提升。其中，最显著的优势在于更符合人类的感知规律——深度学习模型能够从样本中提取更高维度的特征，从而更贴近人类的视觉特性。另一方面，此类模型对图像的结构失真更为敏感，而这恰恰是传统指标难以有效捕捉的。此外，基于深度学习的模型具备更强的泛化能力，针对不同的应用场景，通过对模型进行微调即可使其更好地适配场景需求，进而提升评估结果的准确性。Shi[85]等人则基于人类视觉的方向选择机制，提出将每帧图像中基于方向选择性的视觉模式直方图作为空间特征，并提取连续帧差值的离散余弦变换系数（Discrete Cosine Transform, DCT）以表征时间的变化。得益于对时间特征与空间特征的优化设计，该模型的计算效率得到有效提升。无参考方法无需参考视频，基于待评价视频自身预测其质量。而无论是三维还是二维数字人都难以获取参考模版。因此，无参考方法是更适合于评价数字人视频质量的方法。目前，无参考分为基于知识驱动方法和基于深度学习方法。

知识驱动的视频质量评估（Video Quality Assessment, VQA）依赖手工设计特征，适用于压缩、模糊等传统失真类型。此类方法的早期模型（如 CPBDM[86]）于 21 世纪初提出，2010 年后出现的 TLVQM（Two-Level Video Quality Model）[87]等模型则结合统计特征进行了优化。知识驱动方法具有计算效率高、可解释性强等优势，但由于其所采用的边缘强度、统计矩等特征较为简单，导致其无法处理混合失真问题，泛化能力较弱，且与人类主观感知的拟合程度较低。随着深度学习技术的发展，神经网络被逐步应用于视频质量评估领域，其与人类主观评分的相关性显著优于知识驱动的视频质量评估方法。因此，下文将重点介绍基于深度学习的视频质量评估方法。

表 3 列举了部分典型基于深度学习的无参考视频质量评估（VQA）方法。2019 年，Li 等人[88]首次将门控循环单元（Gated Recurrent Units, GRU）网络引入该领域，用于模拟人类视觉的“感知滞后效应”，并通过非线性映射捕捉视频帧间的时间依

赖关系。2020 年，Chen 等人[89]采用预训练网络与空间金字塔池化技术提取单帧视觉特征。此后，诸多 VQA 方法的设计均借鉴了主流网络结构，例如：SIONR[90]将预训练网络提取的语义特征与低层级统计特征相结合；STDAM[92]引入长短期记忆网络对时空特征进行聚合。

早期方法对视频时序信息的提取存在不足，而 Transformer 结构的提出推动了 VQA 领域性能的提升。Xing 等人[98]首次注意力机制引入该任务，提出 StarVQA 方法，通过时空补丁编码与向量回归损失函数，优化了长时序视频的特征提取效果。此外，VQA 研究不仅需关注性能指标，还需提升评估效率。FAST-VQA[103]采用网格小块采样策略，使计算量减少 97.6%，同时通过门控相对位置偏置模块建模时空关系，实现了精度与效率的平衡。

当前，多模态模型已成为 VQA 领域的主流方向，其通过语义理解增强对视频场景的评估能力，性能已接近人类主观评分水平。MaxVQA[106]将对比语言-图像预训练（Contrastive Language-Image Pre-training, CLIP）[97]模型提取的语义特征与 FAST-VQA[103]提取的纹理特征相融合，在相同数据集下取得了领先性能。COVER[110]则在多模态基础上进一步引入美学特征，分别通过 Swin Transformer 提取技术特征、卷积神经网络提取美学特征和 CLIP 提取语义特征，当前已实现先进性能，这也印证了多模态融合在视频质量评估中的重要价值。

鉴于视频是数字人常用的输出形式，视频质量评估（VQA）方法应用于数字人质量评价具有有效性。然而，传统 VQA 方法受限于视频拍摄视角，且对三维模型的几何失真不敏感。为此，Zhang 等人[114]提出一种融合几何感知的无参考 VQA 方法。该方法从数字人的几何网格中提取二面角、高斯曲率等几何特征并估计统计参数，同时从渲染视频中提取空间特征（通过 2D-CNN 提取）与时间特征（通过 3D-CNN 提取），将上述特征融合后通过回归得到质量评估值。如图 6 所示，该方法基于三维数字人模型的二面角与高斯曲率，通过统计参数估计获取基础统计参数（均值、方差、熵）及分布模型参数（广义高斯分布参数、广义非对称高斯分布参数、伽马分布参数），以此构建几何特征；视频特征则分为时间特征与空间特征，其中时间特征用于捕捉数字人运动失真（如运动不自然、模型裁剪），通过预训练的 Slow-Fast R50[115]提取，空间特征用于识别视频常见失真（如模糊、噪声），通过 ResNet 提取多尺度特征；质量回归阶段采用非线性全连接层将融合后的特征映射为质量评分。该方法通过引入几何特征提升了三维数字人质量评估的性能，但仅依赖几何统计参数描述三维特征显然不够全面，且手工设计特征存在较大局限性。而利用神经网络[116]从多维度提取三维模型的几何特征，能够增强评测的多元性，从而更好地适配多种失真类型。

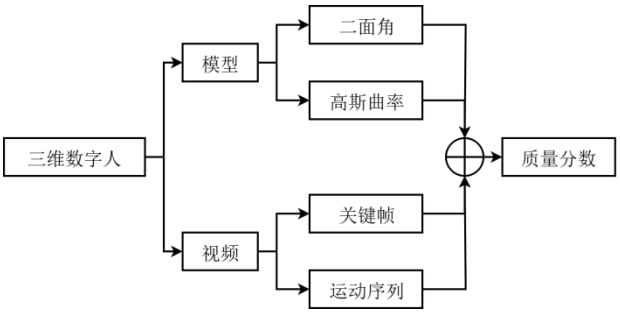

图 6 三维数字人融合几何模型和视频的无参考评价框架

4.1.2 音频指标

音频作为数字人传递信息的另一重要媒介，其质量直接影响数字人的播报效果与交互功能实现，因此音频质量评估具有重要意义。与视觉指标类似，音频质量指标同样划分为主观指标与客观指标。其中，主观指标的定义与视频主观指标一致，通过评测人员打分获取。音频客观指标可分为全参考指标与无参考指标两大类。信噪比（Signal-to-Noise Ratio, SNR）是一种传统的音频客观指标，用于表征信号中有效部分与背景噪声的比例关系。SNR 数值越高，表明信号中的有用信息相对于噪声的占比越大，音质清晰度越高。信噪比通常以分贝（dB）为单位，其计算公式如下：

$$SNR(dB) = 10\log_{10}\frac{P_S}{P_N} \tag{10}$$

其中 P_S 表示信号的有效功率，P_N 表示噪声的有效功率。

表 3 基于深度学习的无参考 VQA（性能指标为 SRCC/PLCC）

发布年份	模型名称	技术特点	测试数据集	性能指标
2021	SIONR[90]	2D CNN+低维特征+时间池化	LIVE-VQC[91]	0.7361/0.7821

发布年份	模型名称	技术特点	测试数据集	性能指标
2021	STDAM[92]	图卷积+空间注意力+双向长短记忆	LIVE-VQC[91]	0.7610/0.8320
2021	AB-VQA[93]	双流时空特征融合	YouTube-UGC[94]	0.7710/0.7900
2021	Patch-VQ[95]	2D/3D 特征融合+ROIPool	KoNViD-1k[96]	0.8031/0.8175
2021	CONVIQT[97]	自监督对比学习+双流架构	LIVE-VQC[91]	0.8093/0.8478
2021	StarVQA[98]	时空交替注意力+向量回归损失	T2VQA-DB[99]	0.7173/0.7295
2022	SWDF-DF-VQA[100]	加权深度特征+多模型决策融合	YouTube-UGC[94]	0.3621/0.3949
2022	2BiVQA[101]	CNN 特征+双层双向记忆	KoNViD-1k[96]	0.8150/0.8350
2022	BVQA[102]	模拟人眼视觉+无参考评估	T2VQA-DB[99]	0.7390/0.7486
2022	FAST-VQA[103]	质量无损采样	YouTube-UGC[94]	0.8617/0.8669
2022	DisCoVQA[104]	时空失真提取+时序内容注意力	LIVE-VQC[91]	0.8211/0.8359
2023	Q-Align[105]	LLM 匹配人类评分	LIVE-VQC[91]	0.7730/0.8300
2024	Zoom-VQA[106]	CLIP[107]+三级时空集成	LIVE-VQC[91]	0.8141/0.8327
2024	SAMA[108]	多粒度金字塔采样+时空掩码	GAIA[109]	0.2361/0.2432
2024	COVER[110]	技术+美学+语义三分支评估	YouTube-UGC[94]	0.9143/0.9165
2025	CAMPVQA[111]	显式质量感知+CLIP[107]图文编码	CVD2014[112]	0.966/0.964
2025	LMMVQA[113]	时空特征编码+LLM 推理	KoNViD-1k[96]	0.901/0.902

但是，这些传统的音频指标同样存在与主观评价不一致的问题。而 PEAQ（Perceptual Evaluation of Audio Quality）[117]的提出解决了这一问题。PEAQ 由国际标准（ITU-R BS.1387）定义，是目前唯一音频质量客观评价的国际标准，在国际上得到了广泛的应用。PEAQ 核心算法结构如图 7，通过模拟人耳对声音的感知过程来计算音频失真。

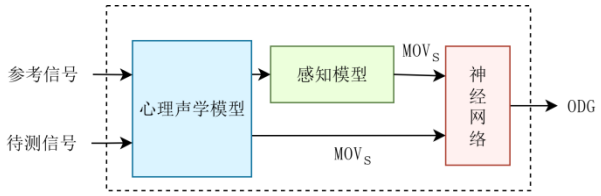

图 7 PEAQ 算法框架[117]

PEAQ 方法的实施流程如下：首先，将参考信号与待测信号输入心理声学模型，依据心理声学理论对输入信号进行时频域扩散处理；随后，将心理声学模型的输出部分输入感知模型，基于人耳的听觉特性进一步提取特征；最后，将提取得到的多维模型输出变量（Model Output Variables, MOV）输入神经网络进行信息融合，并通过测试计算失真系数（Distortion Index, DI）。最终，将失真系数 DI 值转化为客观评价分数（Objective Difference Grade, ODG）。其中，ODG 与 DI 分别对应小失真和中等失真场景下的评价结果，二者的转换公式如下：

$$ODG = \mathrm{b}_{\min} + (\mathrm{b}_{\max} - \mathrm{b}_{\min}) \times sig(DI) \quad (11)$$

其中 $\mathrm{b}_{\min}$ 和 $\mathrm{b}_{\max}$ 为预设的权重参数，DI 为失真系数，$sig(\cdot)$ 为阈值系数。

然而，PEAQ 算法在实际应用中仍存在一定局限性，难以有效适配更高码率及超宽带音频信号的评估需求。为此，国际电信联盟电信标准化部门（ITU-T）推出了 P.863 POLQA（Perceptual Objective Listening Quality Assessment）[118]语音质量评估算法。该算法适用于 50~1400Hz 的音频范围及更高的语音编码速率，能够提供更精准的语音质量评估结果。

POLQA[118]算法基于感知与认知模型构建，通过模拟人类听觉系统的工作机制，将音频信号划分为一系列短时帧，并对每一帧进行频谱分析。其具体步骤如图 8 所示：POLQA 首先通过滤波、预校准及粗校准等步骤实现音频信号的时间对齐，以便计算宏帧时延；随后根据时延结果判断是否需对输入音频进行重采样，以确保两路音频的采样率一致。POLQA 的核心模型包含感知与认知两部分，分别承担语音的客观感知描述与认知评分任务：感知模型将声音响度映射为符合人耳听觉机理的巴克域响度谱[119]；认知模型则基于感知模型输出的频响指标、噪声指标等参数计算 POLQA 值，并将其映射为最终的质量指标。然而，针对 POLQA 与 PEAQ 算法的改进，仍无法克服全参考音频质量评价方法的固有缺陷，即其必须依赖参考音频的支撑。鉴于数字人参考音频的获取存在显著局限性，无参考音频质量评价方法遂成为数字人音频评测的更优选择。表 4 对典型的无参考音频评估方法进行了阐述。无参考音频评估方法的发展历程呈现出清晰的演进轨迹，即从早期基于信号处理的规则驱动模型，逐步过渡至基于自监督学习的预测模型。

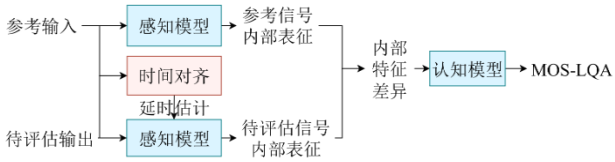

图 8 POLQA 算法框架[118]

早期由信号处理驱动的无参考评估方法，需预先设定失真类型（如噪声干扰、信号中断等）或语音分布模型（如隐马尔可夫模型 HMM/高斯混合模型 GMM），通过量化待测语音与“正常”语音之间的偏差程度实现质量评估。国际电信联盟电信标准化部门（ITU-T）提出的 P.563 算法[120]，便是一种适用于窄带电话系统的单端语音质量评估方法。该算法针对公用电话网络中常见的多种失真场景（包括强噪声干扰、信号中断、静音异常等），通过预处理、特征参数估计及感知映射模型等特定算法流程，最终输出预测的人类主观评分。在 P.563 算法失真特征提取思路的基础上，ANIQUE+算法[121]采用临界带滤波器与调制滤波器组等技术进行特征提取，并对音频失真展开针对性建模，其在多组数据集上的测试性能均显著优于 P.563 算法。此后，众多研究者提出将机器学习模型与 P.563 特征相结合的音频质量评估方法，其性能均得到有效提升，这一结果充分印证了非线性建模在该领域的重要价值。鉴于此，具备非线性拟合能力的神经网络在音频质量评估领域得到了广泛应用。深度学习在该领域的典型应用代表为 MOSNet 模型[122]，该模型通过卷积神经网络（Convolutional Neural Network, CNN）捕捉语音信号的时序特征与局部声学结构，并借助长短期记忆网络[134]整合语音的长期时序信息。MOSNet 不仅与人类主观感知具有高度相关性，其泛化能力及多场景适应性亦优于传统评估方法。值得注意的是，基于机器学习预测人类主观评估分数的语音质量评估方法，因依赖监督学习范式，需耗费大量成本进行大规模人工标注。而 SpeechLMScore[128]通过将语音信号映射为离散令牌，并计算其平均对数概率实现语音质量评估，无需人工标注环节，从而有效规避了高昂的标注成本问题。

表 4 无参考音频评测方法

发布年份	模型名称	技术特点	测试数据集	性能指标
2004	P.563[120]	多项式拟合人类评价	ITU-T Supplement 23	PLCC: 0.88
2007	ANIQUE+[121]	HMM+对数似然度	多个数据集	PLCC: 0.939
2008	回归树[122]	P.563 特征+回归树	2008 Blizzard Challenge[122]	PLCC: 0.69
2015	SVM 与非线性感知正则化[123]	SVM+MFCC	2012 Blizzard Challenge[123]	PLCC: 0.9
2019	MOSNet[124]	CNN+时序建模	VCC 2018[125]	SRCC: 0.978
2022	SSL-MOS[126]	微调 Wav2vec2.0[8]	BVCC[127]	SRCC: 0.944
2023	SpeechLMScore[128]	自监督预训练+域适应	VoiceMOS[129]	SRCC: 0.572
2025	QAMRO[130]	自适应排序损失+加权	AudioMOS Challenge 2025[131]	PLCC: 0.972
2025	MMMOS[132]	多编码器融合+多维度任务	AES-natural[133]	SRCC: 0.847

无监督学习虽解决了样本标注的难题，但仍需大量训练样本及较长的训练时间。相比之下，预训练多模态模型蕴含丰富的先验知识，经微调后即可完成特定任务，在提升模型性能的同时显著提高训练效率。Deshmukh 等[135]提出的 PAM（Prompting Audio-Language Models）模型，是一种基于音频语言模型（Audio-Language Models, ALM）的无参考音频质量评估指标。如图 9 该模型利用 ALM 在大规模音频-文本对上习得的预训练信息，通过“声音清晰干净”与“声音嘈杂有杂音”这两个对立提示词，计算音频嵌入与文本嵌入之间的余弦相似度，进而实现音频质量评估。其中，ALM 作为一种预训练模型，通过在海量音频-文本数据对上进行训练，能够学习并建立音频与文本之间的关联关系。与传统将质量相关提示及待评估音频作为输入的方式不同，PAM 模型引入一对对立提示词用于预测，将音频质量评估问题转化为分类问题。该方法不仅解决了传统无参考模型对大量训练样本的依赖问题，在跨任务音频评估中亦展现出优异性能。

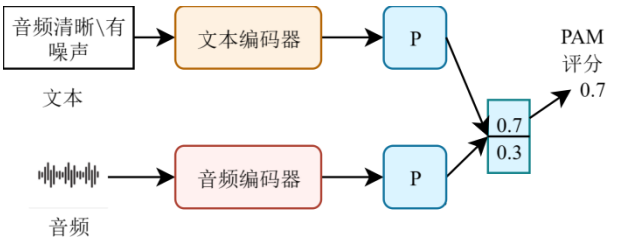

图 9 PAM 算法框架[135]

此外，无参考（NR）模型可利用参考信息提升评估稳定性，而参考信息亦可通过损失函数增强无

参考模型的泛化能力，因此融合二者优势的音频质量评估方法具备实际应用潜力。Manocha 等[136]提出一种名为 CORN（Co-trained Full-Reference and No-reference audio metrics）的新型框架，通过联合训练全参考模型与无参考模型实现语音质量评估。该框架中的无参考模型在训练过程中能够获取参考录音，其性能优于独立训练的基线无参考模型；框架中的全参考模型虽采用与无参考模型相同的训练数据及网络架构，但由于在训练过程中融入了无参考损失，有效避免了对训练内容的过度泛化，性能同样超越独立训练的全参考模型。如图 10 所示，全参考分支通过基础模型将测试信号 x_i 与参考信号 r_j 分别转换为嵌入特征 e_i 与 e_j，并将两个特征进行拼接后输入浅层线性输出头，进而预测评分 f_{ij}；无参考分支仅输入测试信号，x_i 其特征提取方式及评分预测流程与全参考分支一致。

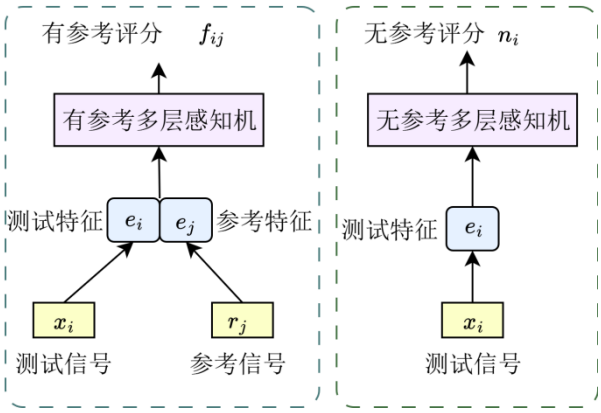

图 10 CORN 算法框架[136]

4.2 数字人内容评价指标

相较于数字人质量评测，数字人内容评测的维度更丰富，其核心是在对语义进行理解的基础上，对内容的准确度展开评估。本文对数字人内容评测的介绍将从形象与交互两个维度展开：其中，形象指标用于衡量数字人形象的仿真与程度，主要从视觉与语音两个层面进行考察；交互指标则聚焦于数字人在交互过程中的动态表现。

4.2.1 形象指标

数字人的形象指标用于评估数字人人物外形与预设形象的相似度。目前已有若干视频指标可用于判别人像的失真程度，因此形象指标属于有参考指标，评测过程中需以参考模板为基准。在无预设参考模板的情况下，亦可选取样本中失真程度最小的一帧画面作为临时模板。

余弦相似度（Cosine Similarity, CSIM）可以用于判别人像相似度的指标，其通过 ArcFace[137]模型提取画面中人像的特征，并以余弦相似度算法计算特征向量间的距离，进而实现相似度评估。ArcFace[137]是 Deng 提出的一种用于人脸识别的方法，通过加性角度优化测地距离，提升特征判别力并稳定训练，其性能优于 SphereFace[138]与 CosFace[139]等先进方法。ArcFace 的计算流程如图 11，首先对输入特征对输入特征向量 x_i 进行 $L2$ 归一化，固定其模长为常数 s，使特征分布在半径为 s 的超球面上。然后，计算特征与权重的点积，即夹角余弦值 $W_j^T x_i = \cos\theta_j$，通过反余弦函数极端角度 $\theta_{y_i} = \arccos(\cos\theta_{y_i})$。接着，在目标角度 θ_{y_i} 上添加固定角度边际 m，得到 $\theta_{y_i} + m$，对应超球面上的弧长边际，得到新的目标对数 N 几率

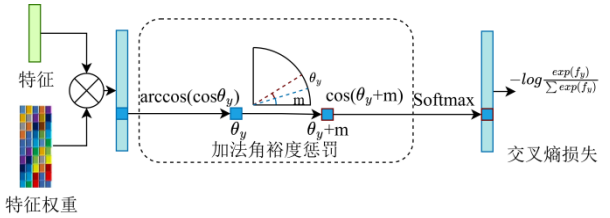

图 11 利用 ArcFace 损失训练 DCNN 网络[137]

$\cos(\theta_{y_i} + m)$。最后，设 N 为样本数，基于 Softmax 函数计算损失函数 L：

$$L = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{s(\cos(\theta_{y_i}+m))}}{e^{s(\cos(\theta_{y_i}+m))}+\sum_{j=1,j\neq y_i}^{n}e^{s\cos\theta_{y_i}}} \quad (12)$$

CSIM 则是基于 ArcFace[137]提取的身份向量计算得到的余弦相似度，公式为：

$$CSIM = \frac{\mathbf{x}\cdot\mathbf{y}}{\|\mathbf{x}\|\|\mathbf{y}\|} \quad (13)$$

式中 x 和 y 分别为原始人物形象特征和待评估人物形象特征。此外，LPIPS 也用于测量生成的人脸和真实人脸之间的特征级相似性。Zhang 等人[82]将 LPIPS 用于测量生成的人脸和真实人脸之间的特征相似性。

此外对于数字人面部的评测，仅从视觉质量考察[140]已无法有效反映各类数字人差异。Xu 等人[140]从医学、美学和表情协调两方面进行针对性的评测，能更有效地与人类感知对齐。首个针对纹理网格数字人类的半参考（RR）质量评估指标，该指标通过提取 3D 几何网格的四个曲率相关特征（顶点缺陷、二面角、离散高斯曲率、离散平均曲率）和 2D 纹理的两个质量指标（灰度照明图、梯度幅度图）。

数字人虽主要依托语音传递信息，但视觉与听觉同步能有效提升信息传递效率——例如音画不

同步会对用户造成干扰，进而影响用户体验。唇同步度正是用于衡量数字人唇形动作与音频匹配程度的核心指标。Chung 等人[141]提出一种双流卷积神经网络架构 SyncNet[142]，该架构能够从无标签数据中学习声音与嘴部图像的联合嵌入特征，用于判定视频中嘴部动作与语音的音视频同步性，并在标准基准数据集上取得了当前最优结果。其中，音频流将梅尔频率倒谱系数（Mel Frequency Cepstral Coefficient, MFCC）值编码为热图，视觉流则采用嘴部区域的灰度图像序列，模型通过对比损失函数完成训练。如图 12 所示，该模型采用双流架构设计：音频流的输入为 MFCC 值，MFCC 基于人耳对声音频率的非线性感知特性（即梅尔频率标度）计算得出，能够有效捕捉语音信号的频谱包络特征；视频流的输入为 5 帧嘴部区域灰度图像序列。在用于唇同步检测时，该模型通过计算 5 帧视频特征与 ±1 秒范围内所有音频特征之间的余弦距离，将距离的最小值定义为唇同步度。

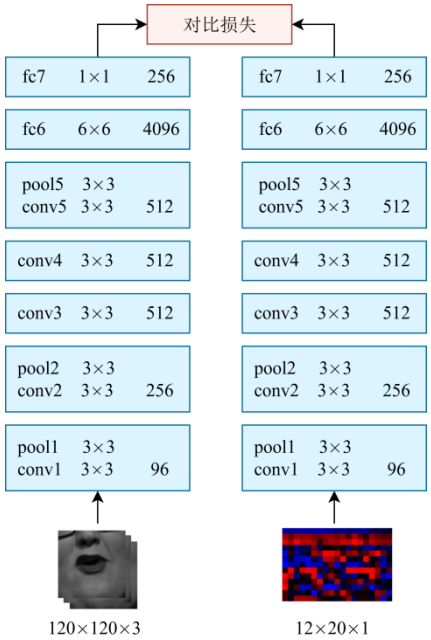

图 12 SyncNet 架构[142]

在 SyncNet[142]模型的基础上，Prajwal 等人[143]提出了 LSE-C（Lip Sync Error-Confidence）与 LSE-D（Lip Sync Error-Distance）两种指标。其中，LSE-C 通过余弦相似度计算音频-视频对的同步概率，其计算公式为：

$$LSEC=\frac{\mathrm{v}\cdot\mathrm{s}}{\max(\|\mathrm{v}\|_2\cdot\|\mathrm{s}\|_2,\varepsilon)} \tag{14}$$

式中 $\mathrm{v}\cdot\mathrm{s}$ 表示视频向量和音频向量的点积。使用预训练的 SyncNet[142]模型，分别对生成视频的唇部区域和对应音频段进行特征编码，得到视频嵌入向量和音频嵌入向量，并计算两者的 $L2$ 距离，即：

$$LSED=\frac{1}{N}\sum_{i=1}^{N}\|\mathrm{v}_i-\mathrm{s}_i\|_s \tag{15}$$

其中 v_i 为第 i 个视频片段的唇动特征嵌入，s_i 表示对应音频的特征嵌入。对比两指标的计算逻辑可知，LSE-C 侧重于评估可信度，反映的是模型对当前音视频唇同步状态判断的可靠程度；LSE-D 则聚焦于量化误差，衡量视频中唇部运动特征与对应音频特征之间的实际错位程度。因此，综合考虑 LSE-C 和 LSE-D 才能有效评估模型的音视频同步能力。

LMD（Landmark Distance）[144]是 Chen 等人提出的用于评估生成唇动视频与真实唇动同步精度的量化指标。首先使用基于方向梯度直方图（Histogram of oriented gradient, HOG）面部检测器[145]，检测生成视频和真实视频的唇动坐标，分别记为 LF 和 LR，每帧视频记录 20 个坐标点。LMD 根据每帧对应坐标的欧氏距离，再按时间长度和地标点数量归一化，公式为：

$$LMD=\frac{1}{T}\times\frac{1}{P}\sum_{t=1}^{T}\sum_{p=1}^{P}\left\|LR_{t,p}-LF_{t,p}\right\|_2 \tag{16}$$

其中 T 为视频时间长度，$P=20$ 为坐标点数量。

语音指标也是评测数字人形象的重要维度。梅尔频率倒谱系数 MFCC 是一种基于人类听觉感知特性的语音特征提取方法，其核心思想在于通过梅尔频率标度与临界带滤波对语音信号的频谱进行变换，从而获取更贴合人耳感知规律的特征参数。

MFCC 的计算过程主要分为预处理与特征提取两大阶段。预处理阶段包括预加重、分帧及加窗操作，旨在增强语音信号的高频成分并保留其时序信息。特征提取阶段的具体流程如下：首先，通过快速傅里叶变换将时序域的语音信号转换为频域特征；随后，将所得频谱输入梅尔滤波器组，该滤波器组包含 M 个三角滤波器，覆盖人类听觉敏感的频率范围，且在低频区域滤波器间隔较密集，在高频区域间隔较稀疏；接着，对滤波后的结果进行对数运算，以模拟人类听觉系统对声音强度的对数响应特性；最后，通过离散余弦变换生成倒谱系数，从而去除各系数之间的相关性。梅尔倒谱系数的具体计算如下：

$$MC_x(i,k)=\sum_{n=1}^{20}X_{kn}\cdot\cos\left[\frac{\pi i(n-0.5)}{20}\right] \tag{17}$$

式中 $MC_x(i,k)$ 为输入语音第 k 帧的第 i 个梅尔倒

谱系数。X_{kn}是第n个临界带滤波器的对数功率输出，计算公式如下：

$$X_{k,n}=\log_{10}\{\sum_{m}|X(k,m)|^2\cdot\omega_n(m)\} \quad (18)$$

式中$X_{k,n}$是第k个输入语音帧的第m个频率段的傅里叶变换，$\omega_n(m)$是第n个临界带滤波器。

基于梅尔频率倒谱系数与临界带滤波，Kubichek 等人提出了梅尔倒谱距离（Mel-Cepstral Distance, MCD）[146]，用于语音质量的客观评价。作为对倒谱距离（Cepstral Distance measure, CD）的改进方法，MCD 融入了听觉系统的非线性频率相关效应：一方面强化了临界带滤波器在匹配主观评价中的重要性，另一方面在保留 CD 计算简便这一优势的同时，提升了评价方法的有效性。假设$MC_X(i,k)$与$MC_Y(i,k)$分别代表输出音频与输入音频第i帧的第k个梅尔倒谱系数，则 MCD 的计算公式如下：

$$MCD(k)=\sqrt{\sum_{i=1}^{M}[MC_X(i,k)-MC_Y(i,k)]^2} \quad (19)$$

尽管梅尔倒谱距离在描述数字人语音的音色特征方面表现突出，但该指标并未考量语音语义的准确性。在中文语音场景中，语义的有效传递高度依赖音调变化，因此数字人对音调的表达能力也需要相应的评估指标进行量化。

基频（Fundamental Frequency, F0）是音频信号中频率最低的基音成分，同时也是人类听觉感知音高的核心依据。语音基频提取技术在语音信号处理领域有着广泛应用，例如语音分离、语音合成等任务[147]。针对基频提取算法的综合性能评估，Nakatani 等人[148]提出了浊音判断误差（Voicing Decision Error, VDE）与粗基音误差（Gross Pitch Error, GPE）两项指标，其计算公式如下：

$$VDE=\frac{N_V^{Verr}+N_U^{Verr}}{N_{V/U}}\times 100 \quad (20)$$

其中N_V^{Verr}和N_U^{Verr}分别是清、浊音的音频帧被错误分类的数量，$N_{V/U}$是清、浊音的帧总数。

GPE 则是评估发声状态（有无发声）判断的准确性，衡量算法误判发声帧和非发声帧的比例，公式如下：

$$GPE=\frac{\sum_t 1_{\mathrm{A}}[|p_t-p_t^{'}|>0.2p_t]1_{\mathrm{A}}[v_t]1_{\mathrm{A}}[v_t^{'}]}{\sum_t 1_{\mathrm{A}}[v_t]1_{\mathrm{A}}[v_t^{'}]} \quad (21)$$

其中p_t，$p_t^{'}$是参考和预测音频的音调信号，v_t，$v_t^{'}$由参考和预测音频的发声决定，1_{A}是指示函数。GPE 指标表示预测音高与参考音高偏差超过 20%的有声帧的百分比。

基于上述两指标，Chu 等人[149]提出了 FFE(F0 Frame Error)，该指标综合了 GPE 和 VDE，更全面地评估 F0 提取的性能。其计算公式为：

$$\begin{aligned}FFE&=\frac{N_{F0E}}{N}\times 100\%+\frac{N_{U\to V}+N_{V\to U}}{N}\times 100\%\\&=\frac{N_{VV}}{N}\times GPE+VDE\end{aligned} \quad (22)$$

式中N是音频的总帧数，N_{VV}是提取的 F0 值和真实值都判断为浊音的帧数，$N_{V\to U}$为清音帧被判断为浊音的帧数，$N_{U\to V}$为浊音帧被判断为清音的帧数。N_{F0E}是满足 GPE 音高偏差条件的帧数：

$$\left|\frac{F0_{i,\text{estimated}}}{F0_{i,\text{reference}}}-1\right|>\delta\% \quad (23)$$

其中i是帧数，δ是阈值，通常为 20。FFE 整合了 GPE（音高偏差）和 VDE（发声准确度）两大关键维度，具有多维度综合评估韵律准确性、与人类对韵律具有高度关联和跨说话者韵律迁移等优点。

视觉是用户交互过程中的主要模态，而数字人通过表情和动作表达情绪、传递信息。因此，人物视觉表现也是重要的评测维度，本文将从头部姿态合成自然度、动作一致性和多视角一致性进行介绍。THEval[150]是一个数字人评估框架，围绕质量、自然度、同步性三大维度设计 8 项细粒度指标。针对数字人自然度，提出了唇部动态、头部运动动态和眉毛动态三方面的指标。具体地，唇部动态是量化嘴唇开合和形状变化丰富度的指标。首先，通过 MediaPipe Face Mesh[151]提取每帧视频的 40 个唇关键点。其次，对关键点分别计算欧氏距离，得到每帧的唇形特征向量s_j，计算全视频标准差θ_m。最后，对所有标准差取平均，得到指标L_D，即：

$$L_D=\frac{1}{M}\sum_{m=1}^{M}\sigma_m \quad (24)$$

$$\sigma_m=\sqrt{\frac{1}{N-1}\sum_{j=1}^{N}(d_{j,m}-\bar{d}_m)^2} \quad (25)$$

式中N是总帧数，$d_{j,m}$是第j帧第m组唇距离，$\bar{d}_m$是第m组距离的均值。

头部运动动态是量化头部旋转与平移的自然运动幅度的指标。首先，通过 FaceXFormer[152]估计每一帧头部姿态，包括旋转（俯仰、偏航、翻滚）和平移（人脸中心在画面中的坐标）。其次，计算头部旋转三个角度的标准差均值$\overline{\sigma_{angle}}$，并得到对应的一阶差分的方差均值$\overline{V_{\Delta angle}}$。接着，计算人脸中心平移方差均值$\overline{V_{trans}}$。最后，根据公式(26)融合得到头部运动动态$D_H$：

$$D_H = \sqrt{(\overline{\sigma_{angle}} \cdot \overline{V_{\Delta angle}}) + \overline{V_{trans}}} \tag{26}$$

眉毛动态是衡量眉毛运动的情绪表达能力的指标。该指标通过提取每一帧眉毛到眼睛的垂直距离，并计算全视频的标准差作为指标D_E：

$$D_E = \sigma_{eb} = \sqrt{\frac{1}{N-1}\sum_{j=1}^{N}(d'_{eb,j} - \overline{d'_{eb}})^2} \tag{27}$$

式中$d'_{eb,j}$和$\overline{d'_{eb,j}}$分别表示每一帧眉毛到眼睛的垂直距离和距离均值。动作一致性是评价数字人的动作表现是否与驱动语义匹配的指标。CLaM[153]是针对文本驱动动作生成设计的对比预训练评估器，采用文本提取器和动作提取器双编码器架构，加入 LLM 同义词用于数据增强与 InfoNCE（Information Noise Contrastive Estimation）损失函数辅助对比学习。具体地，文本提取器负责把自然语言描述转化为全局语义特征，是一个结合 LLM 数据增强的因果注意力语义编码器。动作提取器把 3D 人体动作序列转化为时序语义特征，采用预训练的基于动作自编码器。该方法的核心是混合对比损失训练，强制文本、动作特征空间对齐，损失函数由对比损失L_{Con}和 InfoNCE 损失L_{NCE}加权融合。随着多模态大模型的提出，其强大的语义对齐能力也被用于多模态评测。Video-Bench[154]是基于多模态大模型的文本生成视频人类对齐评估基准，从视频条件对齐与视频质量两个维度进行评测，此处对视频条件对齐进行介绍。视频条件对齐从对象类别、动作、颜色、场景和视频-文本整体的一致性进行打分。该方法通过多模态大模型将视频转为文字描述，再生成针对性的问题链要求大模型回答并核对，最后给出精准分数。然而，上述仅是文本驱动数字人的动作一致性评测，而对视频驱动数字人的相关研究存在空白，是未来的研究方向之一。

三维数字人评测不仅需要关注二维画面的表现，更需要衡量不同三维视角下的一致性。多视角一致性指标弥补了传统图像指标只关注像素相似、忽略几何正确性的缺陷，为单视图到多视图生成提供了科学的评估标准。RE@SG（Reprojection Error under Super Glue estimation）[155]是一种衡量 3D 几何自洽性的指标，其特点在于无需真实标签，仅通过生成的多视角图像与已知相机姿态即可计算，从而评估不同视角之间是否遵循真实三维空间的几何规则。首先，通过 SuperPoint[156]对多视角图像进行稀疏特征关键点（角点、边缘交点、纹理丰富点）检测，得到对应坐标。其次，将不同视角的关键点进行鲁棒匹配，得到跨视角的匹配点对。再次，根据匹配点对和已知相机，使用 DLT 三角化（Direct Linear Transformation），恢复出对应的 3D 空间点。从次，将三角化得到的 3D 空间点，依据相机投影模型重新投影至各视角的 2D 图像平面，得到重投影 2D 点。最后，计算原始 2D 关键点与重投影 2D 点间的$L2$距离，其平均值即$RE@SG$：

$$RE@SG = \frac{1}{N}\sum_{i=1}^{N}\frac{\left\| p_i - \pi(P_i) \right\|_2}{S} \tag{28}$$

其中，p_i是第i个原始 2D 关键点，$\pi(P_i)$是 3D 点P_i重投影的 2D 点，$\|\cdot\|_2$是$L2$距离，S是用于归一化的图像分辨率，N是匹配关键点总数。

此外，三维数字人视角一致性的评测还有基于重建模型一致性的 MRC[157]、无真值和已知相机姿态的评测的 MEt3R[158]和几何-结构-语义跨视角稳定性的 Eval3D[159]等。

4.2.2 交互指标

前文已对数字人的形象指标进行了阐述，本章将聚焦数字人的交互指标，该类指标用于评估数字人实时处理输入信息与生成输出响应的能力。交互指标不仅适用于交互式数字人的评价，其大部分内容对非交互式数字人同样有效。

数字人处理输入信息的能力可依据输入模态划分为音频输入处理与视频输入处理两类。自动语音识别技术能够将人类语音转换为文本，是数字人实现模态转换的核心技术之一。单词错误率（Word Error Rate, WER）是评估 ASR 性能的常用指标，通过计算错误识别的单词数量与总处理单词数量的比例得出；字符错误率（Character Error Rate, CER）的计算法则与WER一致。在给定长度的文本范围内，设I（插入错误数）、D（删除错误数）、S（替换错误数）、H（命中正确数）、N（输入单词总数），则WER的计算公式如下：

$$\begin{aligned} WER &= \frac{S+D+I}{N} \\ &= \frac{S+D+I}{H+S+D} \end{aligned} \tag{29}$$

此外，ASR 的性能评估指标不仅包含识别正确率，还需满足处理效率与内容理解深度等方面的要求。例如，实时因子（Real-time factor，RTF）是衡量 ASR 系统识别效率的核心指标，用于评估其处理时间成本，具体定义为处理一秒钟语音信号所需的平均时间。在完成语音内容识别的基础上，ASR 系统还应具备情感识别、意图分类等更深层次的语义理解功能。这类功能的评估通常依托特定数据集展开，以模型在数据集上的分类正确率作为核心评价指标。常用的情感识别数据集包括 SWEA[160]、IEMOCAP[161]及 RECOLA[162]等。

部分数字人系统还会通过视频模态与用户进行交互：一方面借助 ASR 技术处理用户的语音输入；另一方面从视频画面中提取用户的关键信息，例如人脸定位、动作识别、情感识别等。

对人体的有效定位，不仅有助于精准识别表情与动作，更能提升交互体验。由于该功能本质上属于特定目标的检测任务，其评估指标与通用目标检测指标保持一致。目前主流的目标识别指标包括交并比（Intersection over Union, IoU）、平均精度（Average Precision at IoU, AP）、均值平均精度（Average Precision at IoU=0.5, mAP）等。交并比（IoU）用于衡量模型预测框与真实目标框的重合程度，计算方式为两者交集面积与并集面积的比值，取值范围为 0-1，数值越大表明预测框与真实框的匹配度越高。鉴于数字人交互场景中需识别的目标仅为“人类”这一单一类别，平均精度（AP）成为反映其识别性能的有效指标。AP 是目标检测中评估单个类别检测效果的核心指标，综合体现了模型对该类目标的“检测准确性”与“检测完整性”。AP 的数值通过计算“精确率-召回率（Precision-Recall）曲线”下的面积得到。精确率-召回率曲线以召回率为横轴、精确率为纵轴绘制而成。由于召回率通常呈现离散分布，为获得更准确的 AP 值，需先对召回率进行插值处理，再通过积分计算曲线下面积，具体计算公式如下：

$$P = \frac{TP}{TP + FP} \tag{30}$$

$$R = \frac{TP}{TP + FN} \tag{31}$$

$$AP = \int_0^1 P(r)dr \tag{32}$$

式中 P 和 R 分别为精确率和召回率，TP 表示真正例（正类预测为正类），FP 表示假正例（负类预测为正类），FN 表示假负例（正类预测为负类）。式 32 表示 AP 是 Precision-Recall 曲线与横轴和纵轴的面积，通过积分得到。

在人体定位的基础上，对用户动作与表情的识别能够辅助数字人更精准地理解用户诉求，因此可作为数字人性能评估的附加指标。数字人在动作与表情识别任务上的性能，可通过识别准确率进行量化：测试时输入包含不同动作或表情的视频样本集，计算模型输出预测结果的准确率作为基础指标。为更全面地反映模型性能（兼顾精确率与召回率），目前多采用 F1-score 作为核心评估指标。公开的人脸表情视频数据集包括 CK+[163]、AFEW[164]、CASMII[165]等;动作识别数据集则有 HMDB-51[166]、UCF-101[167]、The Hollywood Dataset2[168]等。此外，情感识别领域已发展出多模态技术，通过融合文本、语音、图像等多种模态数据，实现更全面的情感识别与分析。

意图理解模块基于用户的多模态输入，先将其转换为自然语言形式，再通过信息检索、自然语言生成等技术准确捕捉用户意图。目前数字人系统多采用 LLM 完成语义理解，并生成对应的回复与指令。随着 LLM 的快速发展，相关评测基准不仅需关注模型的基础理解能力，还要评估其在多场景、多任务下的综合性能[169]。其一为理解能力评估，自然语言理解的评测形式通常为问答任务：要求模型在指定语境下回答问题，通过计算模型答案与参考答案的匹配程度量化性能。C-Eval[170]是一个全面的中文基础模型评估套件，包含 13948 道多项选择题，覆盖 52 个不同学科（含人文科学、理工科、社会科学等）及 4 个难度级别（中学、高中、大学、专家）。CMMLU[171]是一个涵盖多元学科（自然科学、社会科学、工程学及人文科学等）的综合性中文基准测试，能有效反映 LLM 的综合性能差异。当前，多数评测基准通过在知识问答、数学推理与代码生成等维度的综合表现来评估模型能力，代表性工作包括 GEM[172]、HLE[173]和 LiveBench[174]等。其二为安全性能评估，SAFETYPROMPTS 基准[175]从 8 类典型安全场景与 6 种对抗性指令攻击场景出发，全面评估模型在安全场景及对抗性攻击下的表现，进而助力提升模型安全性能。针对不同安全场景，可采用不同的评估框架，例如：在权利保护与公平性方面可参考 HELM[176]对其鲁棒性与偏见进行评估；而在防范违法犯罪内容生成方面，则可借助 JADE[177]等工具进行专项检测。

4.3 数字人性能评测框架

随着生成模型的发展，数字人评测方法也同步引入了新的评测技术与评价维度。与传统单一指标

的评测模式不同，当前的评测框架在评价维度上更全面：不仅关注数字人的基础质量表现，还着重考察其内在一致性——例如生成视频是否符合现实世界的物理规律等核心问题。此外，数字人性能评测还需要从用户实际体验的角度进行考察。因此，本文还介绍了一种基于用户感知的评测方法。该方法通过对大量主观评价数据的分析，在一定程度上抵消了个体差异带来的偏差，从而能够直接、准确地反映用户对数字人的真实使用感受。这类新型评价方法同样能够对数字人产品的综合性能做出评判，为相关技术的迭代优化与持续发展提供有力支撑。

4.3.1 基于多维质量的数字人评测框架

为了解决数字人质量评测的问题，Zhang 等[178]提出了零样本、无参考的数字人质量评估方法 DHQA（Digital Human Quality Assessment），随之还提出了首个大规模全身数字人主观质量评估数据库 SJTU-H3D。该方法融合了基于文本提示的语义亲和力、空间自然度及几何损失三类特征。文本提示语义亲和力 Q_A 的核心思路是：高质量数字人的投影应与描述高质量的正向文本具有较高的语义亲和力，而与描述低质量的负向文本具有较低的亲和力。具体实现时，利用 CLIP[97]模型分别计算数字人投影与正向、负向文本对的余弦相似度。空间自然度 Q_N 旨在捕捉数字人投影中存在的低层次失真（例如高斯噪声、模糊、JPEG 压缩伪影等），以补充语义评估在技术细节方面的不足。该方法衡量原始分辨率图像与高质量图像多元高斯分布模型之间的距离，来量化图像的自然度损失。几何损失 Q_G 直接从数字人的网格结构中提取特征，用于量化形状失真。通过计算数字人的网格二面角分布，平均池化得到几何损失评分。在指标融合阶段，由于上述三个指标均已通过 Sigmoid 函数归一化至[0, 1]区间，因此最终质量指数 Q 可直接由三者求和得到，即 $Q = Q_A + Q_N + Q_G$，无需引入额外的权重微调。

4.3.2 基于内容保真度的评测框架

VBench-2.0[179]是一款专为评估视频生成模型内在保真度而设计的基准套件，其评估维度涵盖人类逼真度、可控性、创造力、物理及常识五大核心方面，并进一步细化为18 个细粒度能力维度，具体划分如图 13 所示。该套件通过融合大型语言模型、视觉语言模型等通用模型，以及异常检测等专用模型开展评估工作。

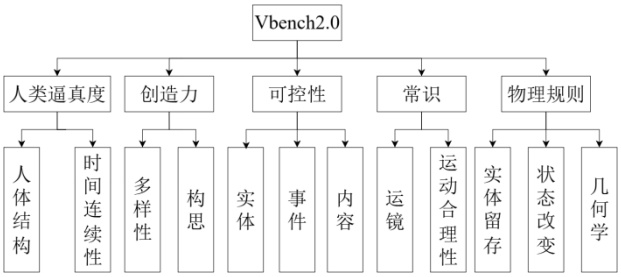

图 13 VBench2.0 的评测维度

VBench-2.0[179]针对上述五大维度分别设计了差异化评估方案，以确保评测的全面性与专业性。针对语义推理和结构化判断任务，该套件采用 GPT-4o、LLaVA-video-7B 等先进的视觉语言模型与大型语言模型；针对特定细分任务，则通过训练专用异常检测模型（如人类解剖异常检测模型）及集成特征匹配算法（如 SIFT[180]、FLANN[181]等），提升细分领域的评估精度。

对于动态属性、力学状态变化等通用概念的视觉理解任务，其评估流程如下：首先构造互补或冗余的测试问题；随后由视觉语言模型根据输入视频直接对相关问题进行作答；最后依据模型回答与标准答案的一致性进行打分。如图 14 所示，针对视频力学变化的示例问题为“易拉罐是否会随着空气排出而被挤压？”，若视频中易拉罐未出现挤压现象，评分为 0；若出现挤压现象，评分为 1。针对多步情节、人类交互等复杂语义场景的理解任务，该方法首先引导视觉语言模型生成视频字幕，随后由大型语言模型对生成的字幕与参考文本进行比对，并依据二者的相似度输出评估分数。对于专业维度的评估，则通过训练基于视觉 Transformer（Vision Transformer, ViT）的模型[182]，实现对手部、面部等关键部位不自然变形的检测。通过对主流视频生成模型的评测发现，各模型的短板主要集中于复杂情节建模、动态控制能力及常识推理三个方面。这些问题在数字人产品中同样存在——无论是三维数字人还是二维数字人，其最终呈现的内容不仅需满足基础视觉指标要求，更需符合现实世界的基本规律，这一要求也对数字人相关技术的发展提出了更高标准。

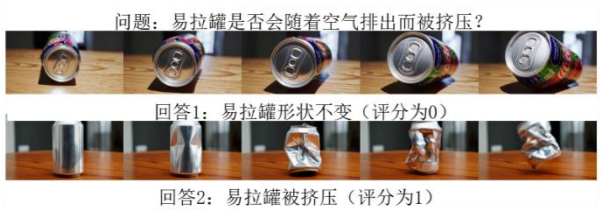

图 14 VBench2.0 力学评价示例[179]

4.3.3 基于美学和技术双分支的评测框架

传统的 VQA 研究以技术层面为核心关注点，其目标在于量化视频中的失真现象（如模糊、伪影等）及其对视频质量的影响，进而为不同技术方案的比较与技术改进提供依据。另一方面，近期有研究提出[183]，非技术层面的语义因素（如内容主题、画面构图等）偏好同样会影响人类对视频质量的主观评估。然而，审美偏好具体如何作用于人类对视频的质量评价，这一问题仍存在争议，有待进一步验证。

鉴于此，Wu 等人[184]提出了一种双分支模型 DOVER（Disentangled Objective Video Quality Evaluator）。该模型通过视图分解策略分别设计美学分支与技术分支，并结合基于主观认知启发的融合策略，明确了 VQA 质量评估的关键影响因素，有效提升了视频质量评估的性能。

为探究美学与技术两个维度对视频质量感知的具体影响，研究者构建了 DIVIDE-3k（Disentangled Video Quality Database）数据集。该数据集通过实验采集了受试者对视频的美学评分、技术评分、综合评分及评分依据。对数据集的分析结果显示，美学评分与技术评分的加权和相比单一维度评分更接近综合质量评分。这一发现表明，融合美学与技术两个维度的评估视角能更准确地反映视频的真实质量感知，同时也为 DOVER 双分支模型的构建提供了理论基础。

技术分支则专注于评估视频的技术质量，核心针对视频中的低级别视觉失真（如模糊、噪声、伪影等），同时尽可能弱化美学因素对技术质量评估的干扰。为实现这一目标，该分支引入由随机裁剪的视频碎片拼接而成的片段作为“技术视图”：通过对视频每一帧进行画面打乱并随机重组处理，拼接后的视图不仅丢失了视频帧的大部分原始信息、破坏了原有的构图关系，还大幅降低了美学因素对评估的影响。为保留视频中的时间域失真（如帧间卡顿、动态模糊等），技术分支采用连续帧采样策略。尽管技术视图丢弃了大部分内容信息，但该分支仍能保留微弱的全局语义作为背景信息，用于区分视频中的真实纹理（如沙粒纹理）与失真现象（如噪声干扰）。技术分支的骨干网络采用专为计算机视觉任务设计的 Video Swin Transformer Tiny 模型[185]。该模型通过移位窗口机制构建层次化特征表示，且其计算复杂度与输入图像大小呈线性关系，兼顾了特征提取的有效性与计算效率。通过对数据集的研究，总体主观平均得分可以近似为美学平均得分（MOSA）和技术评分得分（MOST）的加权和。为了保证每个分支的预测尽可能主要取决于其相应的角度，提出了损失函数 L_{LVBS}，该损失将每个分支的预测之间的相对损失降至最低，计算如下：

$$\mathrm{L}_{LBVS}=\mathrm{L}_{Rel}(Q_{pred,A},MOS)+\mathrm{L}_{Rel}(Q_{pred,T},MOS)+\lambda_{CR}\mathrm{L}_{CR} \tag{33}$$

式中 $Q_{pred,A}$ 和 $Q_{pred,T}$ 分别是美学分支和技术分支的预评测分，λ_{CR} 为超参数。

基于 DOVER 模型的分支结构，作者进一步增强分支的独立性，提出了 DOVER++。对各个分支设计了单独监督的损失函数 L_{DS}，以增强单分支的评估能力，计算如下：

$$\mathrm{L}_{DS}=\mathrm{L}_{Rel}(Q_{pred,A},MOS_A)+\mathrm{L}_{Rel}(Q_{pred,T},MOS_T) \tag{34}$$

因此，所提出的 DOVER++通过融合这两个分支驱动，以联合学习更准确的整体质量以及每个分支的质量预测，损失函数如下：

$$\mathrm{L}_{DOVER++}=\mathrm{L}_{DS}+\lambda_{LVBS}\mathrm{L}_{LVBS} \tag{35}$$

得益于 DOVER++的改进，模型能够更有效地解耦两个分支，使美学分支的预测 $Q_{pred,A}$ 与美学主观意见 MOSA 具有更高相关度，技术分支也如此。因此，单一美学或技术分支也能进行可靠的质量评估。相比 DOVER 仅用整体意见监督，DOVER++在分支解耦上更具优势，提升了单视角评估的准确性。

4.3.4 基于用户感知的评测框架

Zhu 等人[186]以 Zaithanml 的顾客感知价值理论为基础，融合功能价值、情感价值、社会价值等核心维度，构建了数字人评测的理论框架。如图 15 所示，该研究进一步扩展提出“五维评测模型”，涵盖感知技术性、功能性、交互性、情感性、社会性五个维度，旨在从用户视角对数字人的价值进行量化评估。其具体实施流程如下：首先，通过网络爬虫技术采集哔哩哔哩平台上 5 个典型数字人的 11 万余条视频评论；其次，对收集到的评论文本数据进行清洗处理，剔除无效信息；最后，利用 TextRank 算法提取文本中的高频关键词，并通过 snowNLP 工具分析评论的情感倾向，再通过对各维度情感得分进行加权平均，最终得到数字人在感知技术性、功能性等五个维度上的具体评分。

该方法借助大规模数据与文本处理工具，在一定程度上弥补了主观评价客观性不足的缺陷，能够

较为真实地反映用户对数字人的综合评价。然而，该方法仍存在局限性：一方面，数据分布的显著不均衡导致结论的普适性难以保障——不同数字人的评论数量差异悬殊（如洛天依的评论量超 10 万条，而 AI 手语主播的评论量仅 2000 余条）；另一方面，该方法亦未能突破主观评价难以普遍推广的问题。

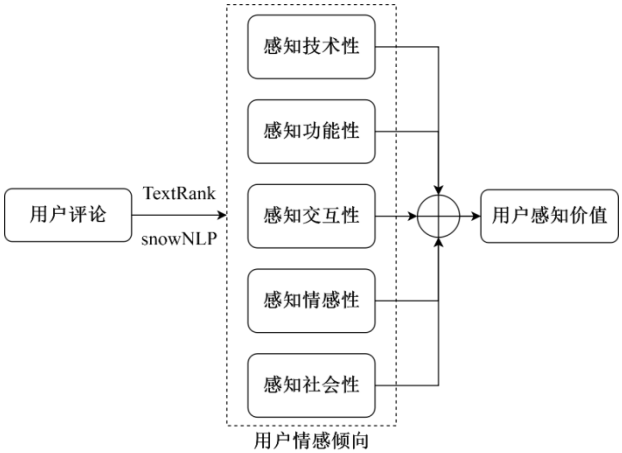

图 15 用户感知视角下数字人评测模型[186]

4.4 数字人评测相关数据集

由于数字人的评测基于模型对人类偏好的学习，因此依赖大规模数字人样本及相应高质量人工评价。然而，数字人评测相关数据集目前仍有很大发展空间，在种类和数量上都存在不足。目前，数字人数据集样本类型主要集中在人物头部，一般采用生成模型构建或真实人物扫描。DHH[187]是一个 3D 扫描数字人头部数据集，包含 55 个参考数字人头部和 1540 个失真样本以及主观感知评分，覆盖 4 个等级的 7 种失真。THQA[188]包含 800 个由 8 种不同语音驱动方法生成的说话人头部视频，其构建基于 StyleGAN[189]，并通过主观实验获得平均意见得分。THQA-10K[190]是当前规模最大、最全面的 AI 生成说话头像质量评估数据集，选用了 12 个主流文生图模型结合 14 种主流语音驱动方法，生成了 10457 个样本。此外，数据集还为每个样本配套了 25 条主观评分，建立了评测质量的人类感知基准，为客观评估方法的设计和验证提供依据。上述数据集虽然已具有一定的规模，但都局限于头部特征，对数字人身体的构建还存在不足。SJTU-H3D[179]数据集包含 40 个高质量参考数字人，并通过对这些参考样本施加 7 类失真处理，生成了 1120 个标注失真样本。通过组织 40 名受试者进行主观评分实验，共收集 44800 条有效评分，并以此计算出每个样本的平均意见得分。该数据集创新地将整个数字人作为样本，但人为添加的失真不能涵盖生成模型的失真，因此存在局限性。

为了增强评测方法对数字人质量的鉴别能力，在模型初期训练时可采用 VQA 数据集，以增加数据多样性。早期 VQA 数据集大多来源于电视和专业拍摄视频，并基于实验需求进行人工扭曲处理，例如 KoNViD-150k[191]、LIVE-VQC[91]、BVI-VFI[192]和 LIVE-YT-HFR[193]等。相比之下，非专业视频在采集、处理和传输过程中会产生多种自然失真，时空失真类型更为复杂且可能相互混合，形成难以命名的新型失真。为此，近年来出现了多个基于用户生成内容的数据集，如 DIVIDE-3k[194]、KVQ[195]、TaoLive[196]和 Maxwell[197]等。随着 AI 生成内容的普及，相关主观质量评估数据集也相继提出，覆盖人类偏好、感知质量、文本对齐及压缩感知等多个维度，代表性工作包括 LGVQ[198]、Gaia[199]、T2VQA-DB[99]和 VBench[200]等。上述数据集为评估生成式模型的视觉质量提供了重要支持。

而对于音频质量评价数据集，其音频类型应涵盖合成语音、真实语音、音乐等多种形式，标注内容通常包括质量评分、声学特征描述等多个维度，以适应不同的音频质量评价任务。例如，QualiSpeech[201]包含中英文语音样本，覆盖合成语音与真实语音，其中约 20%为合成语音与噪声的混合样本，且涵盖多种失真类型，如背景噪声、回声及声音模糊等。AES-Natural[133]包含约 9.7 万个音频样本，来源于多个开源数据集，覆盖语音、音效和音乐三种主要音频类型，能够全面支持不同类型音频的评估任务。该数据集还标注了音频录制与后期处理的工艺水准，可作为音频质量评价的参考依据。为增强模型对不同质量与种类音频的识别能力，训练过程还可引入多个音频数据集，如 DEMAND[202]、FSDnoisy18k[203]、DAPS[204]及 ODAQ[205]等。

4.5 当前研究的不足

随着通用大语言模型的发展，现有数字人评测方法及基准逐渐表现出局限性。这不仅使得实际应用中的模型选型面临困境，也为技术开发阶段的模型改进带来了巨大挑战。下文将对现有评测体系的不足进行分析。

首先，针对数字人核心特性指标的缺失。尽管数字人形象在美学风格上呈现出多样性，但对其拟人化能力等关键属性的评估仍具有重要意义。由于数字人产品以人类形象为核心载体，其音视频等输出内容均带有人类特征，因此该领域的评估往往更依赖主观评价方式。然而，主观评价存在固有缺陷：一方面，大规模主观评价数据的采集成本较高；另

一方面，个体认知与偏好的差异导致评价结果的客观性难以保障。与此同时，现有客观评价方法亦存在明显局限：部分方法仅聚焦于人像的静态相似度评测，未能对数字人的动态表现（如肢体动作连贯性、面部表情自然度）及音频特征（如音色匹配度、语调合理性）进行多维度综合考量。在交互性能评估层面，当前方法也无法实现对多模态信息处理、用户意图理解等单一功能的精准量化评测。此外，数字人技术与功能的迭代速度不断加快，现有评测方法难以完全跟上其发展节奏，导致评测滞后于技术实践。因此，数字人评测基准的设计需从系统层面进行统筹规划：既要针对各项细分功能（如形象生成、交互响应、语义理解等）设计专项评测方案，又要构建能够综合反映数字人整体性能的评估框架。

其次，针对数字人的多模态评测指标体系存在缺失。数字人输出形式通常为文字、语音、视觉等多模态的融合结果，但当前适用于数字人的多模态评测体系尚未发展成熟。多模态的核心并非多个单模态的简单叠加，而是各模态协同作用以实现整体效果。一款用户体验优异的数字人，不仅要求各单模态自身质量达标，还需满足多模态间的输出同步（如唇形与语音匹配）、形象协调（如肢体动作与表情一致）、内容相关（如文字回复与语音表达语义统一）等关键要求。相较于二维数字人，三维数字人还需额外对模型的几何特征（如模型精度、拓扑结构合理性）进行评测。现有评测方法多源自计算机视觉、音频处理等单一模态任务，其评测维度与标准难以覆盖多模态融合场景，导致评测结果与用户实际体验存在偏差。从评测体系构建而言，数字人多模态评测方法需具备两大核心功能：一是对文字、语音、视觉等单模态的独立质量评测；二是对多模态融合效果的综合评估。在结构设计上，还需重点考虑不同模态的输入适配、模态间的时序与语义对齐，以及跨模态内容的一致性对比等关键模块。

再次，现有诸多数字人评测方法存在可解释性缺失的问题。在上述提及的数字人评测方法中，评测结果往往难以提供明确的解释依据，这一问题在主观指标评测中尤为突出。现有评测基准通常仅以单一数字指标（如准确率、相似度分数等）呈现最终结果，既缺乏对评测方法原理的阐释，也未包含对结果成因的分析。而主观评测方法尽管能在一定程度上拟合用户的真实感受，却受限于方法本身的特性，无法从输出的评分中反推“为何获得该评分”的可解释性信息。这种缺失导致此类评测方法存在明显局限：虽能通过指标差异区分不同数字人的性能优劣，却难以阐明性能差异产生的具体原因。开发者也无法精准发现数字人的短板，进而难以针对性地开展技术改进与优化工作。目前，部分先进评测方法已开始尝试突破这一局限：它们以大语言模型为基础，借助模型内置的先验知识与逻辑推理能力，在输出评测结果的同时，同步提供部分评测依据，从而在一定程度上增强了评测结果的可解释性。

最后，数字人领域缺乏统一的评测标准，导致不同数字人之间难以开展横向性能比较，不利于行业形成共同的发展方向。当前各类评测指标的计算逻辑各不相同，结果的数值范围也存在差异（如部分指标取值为 0-1 和 0-100），如何将这些异质指标聚合为一个综合评价结果，成为一大难题。同时，不同属性的指标对数字人最终体验的影响程度存在显著差异，这需要通过额外实验量化各属性对实际体验的贡献权重。目前常用的权重确定方法是采集评测人员的主观评分，通过模型拟合主观评分与各项指标的关联，进而得到指标在综合评分中的权重占比[179]。但该方法不仅需要大规模有效用户评价数据支撑，采集成本较高，对数据的总体分布和组成也有一定要求。

此外相关数据集仍有存在空白，构建全面的评测标准面临数据集不足的挑战。目前的数字人评测数据集存在种类单一与数量少等问题。SJTU-H3D 数据集虽然弥补了样本数量和数字人完整样本的不足，但其未能还原生成模型的失真。因此，数字人评测数据集仍需在样本多样性、失真还原度等方面进一步发展。

5 数字人综合评测方法展望

前文所述的评测方法往往仅从单一维度考察数字人性能，且缺乏对数字人特性的综合设计。为此，本章对数字人综合评测方法进行了展望，以期为未来研究提供可能的发展方向。本章将从质量与内容两个维度展开分析，如图 16 所示。

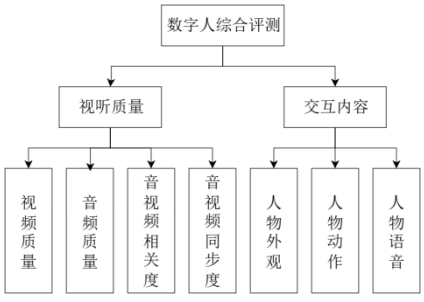

图 16 数字人综合评测方法的展望

5.1 视听质量评测

目前对于数字人质量评测多类似于视频质量评测，往往针对单一的视频模态进行研究，在音视频融合评测和生成式失真检测方面存在不足。而多模态大语言模型（Multimodal Large Language Models, MLLM）在语义对齐、基础常识、端到端应用等优势，评测方法正从自动评估指标（FVD 等）转向多模态大模型。因此，多模态大模型在数字人视听质量评测任务中同样展现出良好的应用前景。如图 17 所示，MLLM 的基础架构可分为三个核心模块，即多模态编码器、大语言模型以及连接上述模块的模态交互接口。多模态编码器的主要作用在于将视频、音频等原始数据编码为维度紧凑、表征高效的特征向量，以满足下游模型的处理需求。针对数字人评测场景主要以视频和音频为核心输出模态的特点，因此模态编码器可采用视频编码器与音频编码器。常见的视频编码器包括 UMT-L[206]、UniFormerV2[207]、VideoCLIP[208]和 ViViT（Video Vision Transformer）[209]等。其中，VideoCLIP 基于 CLIP 架构扩展而来，通过对比学习实现视频与文本的语义对齐，主要适用于短视频与简短文本的跨模态表示，对长视频的处理能力有限。VideoPrism 则基于 ViT[182]架构，采用 ViViT[209]的时空分解注意力机制，能够适配多源视频数据，并支持分类、定位、检索与生成等多种任务。典型的音频编码器包括 CLAP（Contrastive Language-Audio Pretraining）[210]、Wav2vec 2.0[8]、BEATs[211]、AudioMAE[212]与 MMS[213]等。CLAP[211]在大规模音频-文本对上进行预训练，其核心思想与视频编码器相似。Wav2vec 2.0[8]利用无监督预训练直接从原始音频中学习表征，因而能够有效利用大量未标注语音数据，已成为音频编码的基础模型之一。此外，除单一模态编码器外，目前也出现了可同时对齐多个模态语义的全模态编码器，例如 Align-Anything[214]、ImageBind[215]与 Unified-IO 2[216]。这类编码器为大模型的语义对齐提供了完整解决方案，推动了全模态大模型与人类意图的对齐。考虑到数字人多模态的特性，以及建立统一评估体系的实际需求，全模态编码器是更好的选择，也是数字人评测研究的重要基础之一。

预训练大语言模型通过在通用网络语料库上进行大规模训练，具备丰富的世界知识与较强的逻辑推理能力，因而成为多模态信息理解与推理的核心组件。其主要作用是对编码器提取的多模态特征进行深层语义解析与综合判断。在开源大语言模型的选型中，DeepSeek R1、GLM 4.5V、Qwen3 及 Gemma2 等模型被广泛应用。上述模型在语言理解、逻辑推理、交互对话、安全伦理等方面均表现出色，且输出稳定性与可靠性持续提升，因此适用于视听质量的综合评测任务。模态接口作为连接多模态编码器与大语言模型的关键枢纽，其核心作用是将多模态特征映射到统一的语义空间。

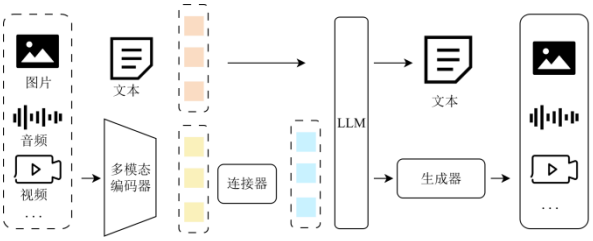

图 17 MLLM 架构[217]

为使多模态模型适应数字人评测任务，一般采用“预训练-指令微调-对齐微调”三级训练策略。其中，预训练阶段的核心目标是实现跨模态特征对齐与多模态基础能力的构建，为后续任务优化提供基础。指令微调阶段旨在引导模型理解特定任务指令并完成相应推理，从而提升其在特定评测任务中的零样本泛化能力。该阶段的训练数据通常采用三元组 (I,M,R) 的形式构建，其中 I 表示任务指令（例如“从清晰度与流畅度维度评价该数字人视频质量”），M 为多模态输入数据（如数字人视频或音频），R 为对应的目标响应（如参考评分或评价结果）。在微调过程中，通过构建与音视频质量评估相关的多样化指令集，并对模型输出的评价结果范围（如评分区间、评价维度表述等）进行约束，从而增强模型响应的规范性与针对性。对齐微调的核心目标在于使模型的输出更贴合人类主观偏好，具体体现为减少幻觉现象、提升评价结果的可信度与合理性。当前该阶段主要采用 RLHF（Reinforcement Learning from Human Feedback）[218]进行优化。RLHF 的核心流程通常分为监督微调、奖励建模和强化学习优化三步，逐步使模型对齐人类偏好。监督微调（Supervised Fine-Tuning, SFT）基于高质量的人类标注数据对预训练模型进行进一步调优。奖励建模（Reward Modeling, RM）旨在将人类的主观偏好转化为可量化的奖励信号。强化学习优化通常采用 PPO（Proximal Policy Optimization），其目标是通过最大化累积奖励来优化模型策略。与 PPO 不同，DPO（Direct Preference Optimization）则直接利用偏好数据进行对比学习，无需显式构建奖励模型，简化了训练流程。具体到

数字人评测任务中，“预训练-指令微调-对齐微调”的策略则分别对应训练模型失真识别能力、赋予评价能力和对齐人类偏好。因此，基于多模态大模型的数字人评测研究一方面要构建丰富多样的数据集，另一方面在训练策略上需要侧重融合评测的能力。

5.2 交互内容评测

交互是数字人的重要功能之一，需要从形象、语音、动作等多个维度考察。通用多模态大模型的训练数据多来源于真实世界视频，而数字人评测则需要识别生成式视频特有的异常现象，例如人体解剖结构异常、实体突然融合或消失、“假动作”缺乏物理效果等。由于通用大模型缺乏对这类生成式异常的识别经验，在细粒度视觉分析和帧间时序异常检测上表现不足。因此，对于数字人交互内容的评估，不仅需要依赖多模态大模型，还需引入专用模型，以提取细粒度特征进行判别。

数字人外观异常的评估可以从人体结构、人物身份及服装三个方面开展。人体结构异常通常出现在肢体、手部、面部等部位，表现为形态畸变、位置错乱、结构缺陷或数量异常等。人体结构异常检测可通过目标检测模型与异常检测模型相结合的方式实现：目标检测模型用于提取数字人的人体区域片段，确保异常检测聚焦于特定部位，排除背景干扰，如 YOLO-World[219]、DETR[220]与 GLIP[221]等；异常检测模型则负责判断局部区域是否存在异常，通过微调使其能够有效识别人体结构异常，如 Vit-base[222]、Swin Transformer[223]和 ConvNext[224]等图像分类模型。人物身份一致性评估需要考虑数字人各帧中人脸的一致程度。计算各帧与基准的人脸相似度，并基于阈值判断身份是否一致。具体地，提取数字人动态序列中的人脸特征并映射为高维向量，压缩向量并用于相似度计算。此处以 Candide3[225]为例，该模型能够在不同光照和姿态条件下实现单样本人脸识别与特征提取。此外，针对各类场景可采用不同的人脸检测模型，例如：RetinaFace[226]对视频中小尺寸人脸、遮挡及姿态变化具有较强鲁棒性；FSA-Net[227]可处理大角度侧脸及严重遮挡情况，并支持姿态校正；YOLOv9-Face[228]在保持高精度的同时具备极快的检测速度。

服装作为数字人外观的另一重要组成部分，也应纳入评测体系。类似地，可通过提取各帧中的服装片段并与基准进行比较，以评估服装一致性。为充分捕捉服装从全局到局部的多尺度特征，本章以 MCFM（Multiscale CNN Feature Model）[229]为例说明提取流程。该方法的计算流程如图 18 所示。首先，使用 YOLOv3[230]算法对图像进行区域检测与提取，识别出全局区域、服装主体区域及款式部件区域。随后，根据各区域类别将其分别输入至对应的网络分支中。最后，对全局、主体及部件分支输出的特征向量进行级联融合，形成全面表征服装语义的特征表示用于相似度计算。数字人外观异常的评测从人体结构、人物身份及服装三个方面分别进行，缺少对内在关联度与整体性的考量，因此未来的研究可将数字人视为整体评测外观异常，用单一指标反映全面质量性能。

动作指标用于量化数字人的动作表现能力，主要涵盖多样性与合理性两个维度。动作合理性关注数字人生成动作是否符合现实物理规律与人类行为常识。数字人的动作不仅应丰富多样，更需遵循客观约束——若出现手臂旋转角度超限、肢体拉伸违背生理结构等异常动作，将显著降低用户体验。动作骨骼提取模型可提取人物骨骼关键点的时间序列，并用于统计异常关节出现的比例。在动作合理的前提下，动作多样性是衡量其动作生成能力的另一关键指标。同样基于动作骨骼提取模型，可以构建动作多样性量化指标。目前，人体动作序列的提取模型主要包括 DMS-GCN[231]、AQ-GCN[232]和 TU-GCN[233]等。上述方法适用于简单的动作提取，若需处理复杂动作序列，可结合微调多模态大模型进行更细粒度的动作提取，以实现更严格的检测。

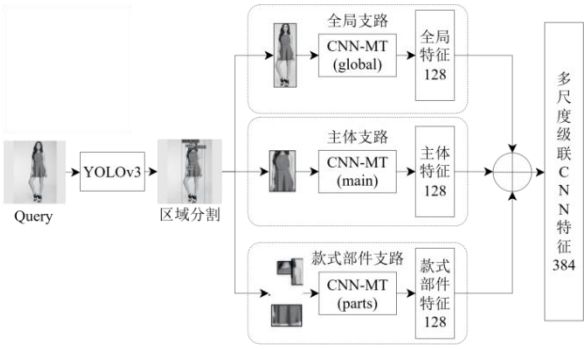

图 18 MCFM 的计算流程[229]

在语音表达方面，评测重点为数字人语音的还原度，即生成语音与基准语音的相似程度。通常基于双流特征提取架构计算基准与生成语音的差异。如图 19 所示，使用两个并行的神经网络提取语音特征，训练目标为准确判别两个语音样本是否一致。该方法本质上是一种基于对比学习的相似度计算框架。针对不同任务，也可选用其他适合的相似

度计算方法，例如：ECAPA-TDNN[234]通过多分支结构分别捕捉细粒度与粗粒度特征，具有较强的抗干扰能力；WavLM[235]采用自监督与微调结合的框架，在噪声和失真条件下仍能保持较好的区分性能；SVSNet[236]则是一种端到端的音频相似度计算模型，具备良好的泛化能力。上述方法从语音相似度考察了数字人的交互能力，语音自然度、多语言等也是不可或缺的能力，值得进一步研究。此外，在语言理解、任务完成等能力的评测更应纳入考量，如何借助大语言模型的综合能力进行全面评测也需要更多的研究投入。

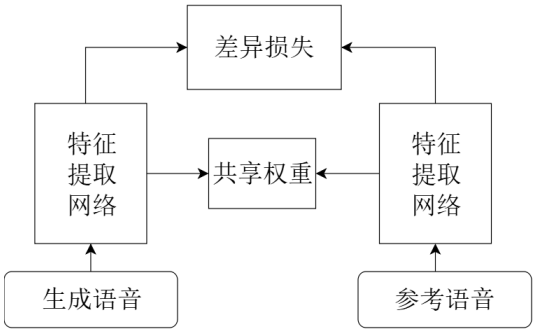

图 19 语音评测的双流架构[142]

上述框架针对数字人的单一方面分别进行评测，为确保评估结果在评测维度上与人类偏好吻合，需要采集人工标注对评测模型进行校正。目前生成模型的评测主要采用 RLHF 方法实现感知对齐，用收集的人类反馈数据训练模型学会模仿人类的判断标准，最终输出预测人类偏好分数。然而 RLHF 方法需要高质量、多样化、无偏见的数据，因此目前许多评测方法采用已有数据集进行拟合，通过加权公式将子指标与人类判断对齐。目前，已有评测框架通过收集人类评估分数来提供参考基准，例如：视觉内容生成评价基准 VBench2.0[179]、长视频理解与评估基准 LOVE[237]，以及视频生成质量与内容评估工具集 Evalcrafter[238]等。目前多采用加权平均的方法实现，且依赖大量主观样本，如何高效融合各项指标是各种评测方法共同的问题。在数字人评测的研究中，单一模型实现全面评测目前还难以实现，但是基于大模型的多维度评测是具有前景且可行的研究方向。

本章基于当前技术现状对数字人评测方法进行了展望。由于单一模型难以同时捕捉数字人多维特征与多粒度语义信息，现有方法通常采用融合多个模型输出结果的策略。然而，这种策略主要在结果层面进行融合，未能有效建模各模态之间的内在协同关系。数字人依靠音频、视频等多模态信号实现交互，仅将各模态指标简单叠加难以准确反映其整体表现。此外，数字人传递内容的逻辑性、真实性也是重要的测评对象，反映其在语言处理上的性能。未来的统一评测架构应立足于单一模型设计，从单模态特征提取与模态间协同建模两方面出发，实现对数字人表现的整体评估，并输出与人类主观评价一致的综合指标。

6 总结

本文围绕数字人评测构建，从数字人技术架构、行业标准、已有相关指标和展望评测方法四个维度展开阐述，并进行总结分析。首先，本文明确界定了数字人的核心技术架构，该架构由感知层、决策层、表达层和扩展层构成。由于数字人的全面评测需覆盖各功能模块的检验，因此清晰的框架划分可为针对性评测方法的设计提供明确依据与逻辑支撑。其次，本文系统梳理了数字人领域的行业标准及评测指标。随着数字人技术的快速迭代，早期制定的行业标准已难以适配当前技术形态与应用场景的需求，为此，本文进一步介绍了近年的新型评测指标及对应的评估方法。同时指出，单一维度的评测方法存在评估视角局限、难以反映数字人综合性能的问题，因此多维度融合的综合评价方法可作为数字人评测体系构建的重要参考方向。最后，针对上述问题，本文从“质量”与“内容”两大核心维度出发，展望了未来数字人的评测方法。

参 考 文 献

[1] National Radio, film and television Standardization Technical Committee Technical requirements for digital virtual human: GY/T 411-2024. State Administration of radio and television, 2023: 2-3.
(全国广播电影电视标准化技术委员会. 数字虚拟人技术要求：GY/T 411—2024. 国家广播电视总局, 2023: 2-3.)

[2] Chen S, The development status and risk countermeasures of virtual digital human Radio and television network, 2025, 32 (06): 34-38.
(陈实. 虚拟数字人发展现状及风险对策分析. 广播电视网络, 2025, 32(06): 34-38.)

[3] iiMedia Research. New insights into the development of China's digital human industry in 2025. https://www.iimedia.cn/c880/107807.html.
(艾媒咨询. 2025 年中国数字人产业发展新洞察. https://www.iimedia.cn/c880/107807.html.)

[4] Internet Society of China, China academy of information and communications technology, communication university of China. China digital human development report. China Internet association, 2024.
(中国互联网协会, 中国信息通信研究院, 中国传媒大学. 中国数字人发展报告. 北京：中国互联网协会, 2024.)

[5] Sun S C，Xue Z Y. Analysis on the development status of virtual digital human. Advanced Television Engineering, 2024(02): 12-18.
(孙苏川，薛子育. 虚拟数字人发展现状分析. 现代电视技术，2024(02): 12-18.)

[6] INSIGHTS B R. Intelligent virtual digital human market size, share, growth and industry analysis. https://www.businessresearchinsights.com/zh/market-reports/intelligent-virtual-digital-human-market-122012.

[7] Dahl G E, Yu D, Deng L, et al. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2012, 20(1): 30-42.

[8] Baevski A, Zhou Y, Mohamed A, Et Al. Wav2vec 2.0: a framework for self-supervised learning of speech representations. Advances in neural information processing systems, 2020, 33: 12449-12460.

[9] Radford A, Kim J W, Xu T, er al. Robust speech recognition via large-scale weak supervision//Proceedings of the 40th International Conference on Machine Learning. Honolulu, USA, 2023: Article 1182.

[10] Li K, Gao K, Hu X. Efficient audio-visual speech separation with discrete lip semantics and multi-scale global-local attention. arXiv preprint arXiv:2509.23610, 2025.

[11] Dan J, Liu Y, Sun B, et al. Transface++: Rethinking the face recognition paradigm with a focus on accuracy, efficiency, and security. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2026: 48(2): 1243-1261.

[12] Dan J, Liu Y, Xie H, et al. Transface: calibrating transformer training for face recognition from a data-centric perspective//2023 IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France, 2023: 20585-20596.

[13] Lv G, Zhang J, Tsoi C. Facial expression recognition via variational inference. Scientific Reports 16, 2026: Article 7323.

[14] Lan X, Xue J, Qi J, et al. Expllm: Towards chain of thought for facial expression recognition. IEEE Transactions on Multimedia, 2025, 27: 3069-3081.

[15] Chen L H, Lu S, Zeng A, et al. Motionllm: Understanding human behaviors from human motions and videos. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2025, doi: 10.1109/TPAMI.2025.3627546.

[16] Yang Y, Zhang J, Zhang J, et al. Expressive keypoints for skeleton-based action recognition via progressive skeleton evolution. IEEE Transactions on Image Processing, 2025, 34: 7585-7599.

[17] Feng W, Zhu Y, Zhang R, et al. Active multimodal distillation for few-shot action recognition//Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence. Montreal, Canada, 2025: 999-1007.

[18] Tajmirriahi M, Rabbani H. Eeg coupled scale-invariant dynamics for emotion recognition: a domain adaptation approach. IEEE Transactions on Affective Computing, 2025, 16(4): 3584-3595.

[19] Wang Y, Liu J W, Lu B L, et al. From eeg to eye movements: Cross-modal emotion recognition using constrained adversarial network with dual attention. IEEE Transactions on Affective Computing, 2025, 16(3): 1543-1556.

[20] Mahapatra A, Patra P K. A unified and scalable machine learning framework for feature fusion in object classification using weighted PCA with adaptive concatenation and dynamic scaling. Discover Computing, 2025, 28(1): 114.

[21] Möderl J, Westerkam A M, Venus A, et al. A block-sparse bayesian learning algorithm with dictionary parameter estimation for multi-sensor data fusion//2025 28th International Conference on Information Fusion (FUSION). Rio de Janeiro, Brazil, 2025: 1-8.

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

[23] Chahi A, Kas M, Kajo I, et al. R2GAN: Enhancing unseen image fusion with reconstruction-guided generative adversarial network. Applied Intelligence, 2025, 55(11): 821.

[24] Lyu K, Xiao L, Zeng J, et al. Touchformer: A robust transformer-based framework for multimodal material perception //Proceedings of the AAAI Conference on Artificial Intelligence: vol. 40: 22. Singapore, 2026: 18496-18504.

[25] Baltruaitis T, Ahuja C, Morency L P. Multimodal machine learning: A survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(2): 423-443.

[26] Wang J, Ji Y, Yang H. Rav: Retrieval-augmented voting for tactile descriptions without training//Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. Suzhou, China, 2025: 6198-6205.

[27] Lan C, Wang Y, Wang C, Et Al. Application of ChatGPT-based digital human in animation creation. Future Internet, 2023, 15(9): 300.

[28] Lai P, Zhong W, Qin Y, Et Al. Llm-driven multimodal and multi-identity listening head generation//Proceedings of the Computer Vision and Pattern Recognition Conference. Nashville, USA, 2025: 10656-10666.

[29] Lin J, Feng Y, Liu W, et al. Chathuman: Chatting about 3d humans with tools//Proceedings of the Computer Vision and Pattern Recognition Conference. Nashville, USA, 2025: 8150-8161.

[30] Li Y A, Han C, Raghavan V, et al. Styletts 2: Towards human-level text-to-speech through style diffusion and adversarial training with large speech language models. Advances in neural information processing systems, 2023, 36: 19594-19621.

[31] Jeong M, Kim H, Cheon S J, et al. Diff-tts: A denoising diffusion model for text-to-speech. arXiv preprint arXiv:2104.01409, 2021.

[32] Mehta S, Tu R, Beskow J, et al. Matcha-TTS: A fast TTS architecture with conditional flow matching//ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Seoul, South Korea, 2024: 11341-11345.

[33] Chen Y, Niu Z, Ma Z, et al. 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, 2025: 6255-6271.

[34] Liu Z, Mao H, Wu C Y, et al. A convnet for the 2020s// Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. New Orleans, USA, 2022: 11976-11986.

[35] Karras T, Laine S, Aila T. A style-based generator architecture for generative adversarial networks//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Long Beach, USA, 2019: 4401-4410.

[36] Yang Q, Guan J, Wang K, et al. Showmaker: Creating high-fidelity 2d human video via fine-grained diffusion modeling. Advances in Neural Information Processing Systems, 2024, 37: 51039-51062.

[37] Jiang X, Ding Y. Clipfacefusion multi modal diffusion for high fidelity facial generation and modification. Scientific Reports, 2025.

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

[39] Gezawa A S, Zhang Y, Wang Q, et al. A review on deep learning approaches for 3D data representations in retrieval and classifications. IEEE access, 2020, 8: 57566-57593.

[40] Chu Z, Xiong F, Liu M, et al. Humanrig: Learning automatic rigging for humanoid character in a large scale dataset//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA, 2025: 304-313.

[41] Yan Y C, Cheng Y H, Chen Z, et al. A survey of research on generative 3D digital human based on neural network: representation, rendering and learning. Science China Information Sciences, 2023, 53(10): 1858-1891.
(晏轶超, 程宇豪, 陈琢, 等. 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, 2023, 53(10): 1858-1891.)

[42] Wang Y, Han Q, Habermann M, et al. Neus2: Fast learning of neural implicit surfaces for multi-view reconstruction//Proceedings of the IEEE/CVF international conference on computer vision. Paris, France, 2023: 3295-3306.

[43] Lu Y, Zhang J, Fang T, et al. Matrix3d: Large photogrammetry model all-in-one//Proceedings of the Computer Vision and Pattern Recognition Conference. Nashville, USA, 2025: 11250-11263.

[44] Hanson A, Tu A, Lin G, et al. Speedy-splat: Fast 3d gaussian splatting with sparse pixels and sparse primitives//Proceedings of the Computer Vision and Pattern Recognition Conference. Nashville, USA, 2025: 21537-21546.

[45] Wang Y, Han Q, Habermann M, et al. Neus2: Fast learning of neural implicit surfaces for multi-view reconstruction//Proceedings of the IEEE/CVF international conference on computer vision. Paris, France, 2023: 3295-3306.

[46] Cheng Y C, Lee H Y, Tulyakov S, et al. Sdfusion: Multimodal 3d shape completion, reconstruction, and generation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada, 2023: 4456-4465.

[47] Chen H, Peng B, Tao Y, et al. Dˆ3-human: Dynamic disentangled digital human from monocular video//Proceedings of the Computer Vision and Pattern Recognition Conference. Nashville, USA, 2025: 10836-10846.

[48] Jia H, Zhu L, Zhao N. H3r: Hybrid multi-view correspondence for generalizable 3d reconstruction//Proceedings of the IEEE/CVF International Conference on Computer Vision. Shenzhen, China, 2025: 7655-7665.

[49] Tan X, Wu W, Zhang Z, et al. Geocc: Geometrically enhanced 3d occupancy network with implicit-explicit depth fusion and contextual self-supervision. IEEE Transactions on Intelligent Transportation Systems, 2025, 12(3): 1543-1556.

[50] Kerbl B, Kopanas G, Leimkuehler T, et al. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. ACM Transactions on Graphics, 2023, 42(4): 139:1-139:14.

[51] Chen P, Wei X, Wu Q, et al. Mixedgaussianavatar: Realistically and geometrically accurate head avatar via mixed 2d-3d gaussians//Proceedings of the 33rd ACM International Conference on Multimedia. Dublin, Ireland, 2025: 945-954.

[52] Wang Y, Zhang X, Zhan K, et al. Hineus: High-fidelity neural surface mitigating low-texture and reflective ambiguity//Proceedings of the IEEE/CVF International Conference on Computer Vision. Shenzhen, China, 2025: 25746-25755.

[53] Tang Z, Feng C, Cheng X, et al. Neuralgs: Bridging neural fields and 3d gaussian splatting for compact 3d representations//Proceedings of the AAAI Conference on Artificial Intelligence: vol. 40: 11. Singapore, 2026: 9493-9501.

[54] Jang W, Hong Y, Cha G, et al. Controlface: Harnessing facial parametric control for face rigging//Proceedings of the Computer Vision and Pattern Recognition Conference. Nashville, USA, 2025: 5614-5624.

[55] Guo Z, Xiang J, Ma K, et al. Make-it-animatable: An efficient framework for authoring animation-ready 3d characters//Proceedings of the Computer Vision and Pattern Recognition Conference. Nashville, USA, 2025: 10783-10792.

[56] Li X, Yuan Y, De Mello S, et al. Simavatar: Simulation ready avatars with layered hair and clothing//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA, 2025: 26320-26330.

[57] Kim B, Jeong H I, Sung J, et al. Personabooth: Personalized text-to-motion generation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA, 2025: 22756-22765.

[58] Wang T, Wu Z, He Q, et al. Stickmotion: Generating 3d human motions by drawing a stickman//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA, 2025: 12370-12379.

[59] Huang Y, Wang J, Zeng A, et al. Dreamwaltz-g: Expressive 3d gaussian avatars from skeleton-guided 2d diffusion. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2025, doi: 10.1109/TPAMI.2025.3586284.

[60] Karras T, Aila T, Laine S, et al. Audio-driven facial animation by joint end-to-end learning of pose and emotion. ACM Transactions on Graphics, 2017, 36(4): Article 94.

[61] Wang Z, Chen B, Liang Z, et al. Audio2lip: Speech-driven 3d face animation based on facial motion prediction and face reconstruction//2025 IEEE 6th International Seminar on Artificial Intelligence, Networking and Information Technology (AINIT). Shenzhen, China, 2025: 1-10.

[62] Yang Y, Cen Z, Peng S, et al. Streamingtalker: Audio-driven 3d facial animation with autoregressive diffusion model//Proceedings of the

AAAI Conference on Artificial Intelligence. Singapore, 2026, 40(14): 11766-11774.

[63] Fan X, Li J, Lin Z, et al. Unitalker: Scaling up audio-driven 3d facial animation through a unified model//European Conference on Computer Vision. Milan, Italy, 2024: 204-221.

[64] Stan S, Haque K I, Yumak Z. Facediffuser: Speech-driven 3d facial animation synthesis using diffusion//Proceedings of the 16th ACM SIGGRAPH Conference on Motion, Interaction and Games. New York, USA, 2023: 1-11.

[65] Lu J, Lin J, Dou H, et al. DPoser-X: Diffusion model as robust 3D whole-body human pose prior//Proceedings of the IEEE/CVF International Conference on Computer Vision. Shenzhen, China, 2025: 9988-9997.

[66] Wu Q, Esturo J M, Mirzaei A, et al. 3dgut: Enabling distorted cameras and secondary rays in gaussian splatting//Proceedings of the Computer Vision and Pattern Recognition Conference. Nashville, USA, 2025: 26036-26046.

[67] Gao Z, Planche B, Zheng M, et al. 7DGS: Unified spatial-temporal-angular Gaussian splatting//Proceedings of the IEEE/CVF International Conference on Computer Vision. Shenzhen, China, 2025: 26316-26325.

[68] Krayem I, Ghourabi M, AL Assaad M. Mma-rag: Multi-modal agents for insurance document processing with retrieval-augmented generation// Proceedings of the Intelligent Systems Conference. Amsterdam, The Netherlands, 2025: 196-209.

[69] Huang X, Li R, Cheung Y M, et al. Gaussianmarker: Uncertainty-aware copyright protection of 3d gaussian splatting. Advances in Neural Information Processing Systems, 2024, 37: 33037-33060.

[70] Yoo I, Chang H, Luo X, et al. Deep 3D-to-2D watermarking: Embedding messages in 3D meshes and extracting them from 2D renderings//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Virtual, 2022: 10031-10040.

[71] ITU-T. Requirements and evaluation methods of non-interactive 2D real-person digital human application systems. 2022. https://www.itu.int/rec/T-REC-F.748.14-202203-I/en.

[72] ITU-T. Framework and metrics for digital human application systems. 2022. https://www.itu.int/rec/T-REC-F.748.15-202203-I/en.

[73] IEEE Standard for a Framework for Evaluating the Quality of Digital Humans. IEEE Standard 3079. 32023, 1-53.

[74] Zhou W, Bovik A C, Sheikh h R, et al. Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing, 2004, 13(4): 600-612.

[75] Heusel M, Ramsauer H, Unterthiner T, et al. Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems. Long Beach, USA, 2017: 6629–6640.

[76] UHD World Association. Technical requirements for 3D digital human quality grading. Beijing: UHD World Association, 2023.
(世界超高清视频产业联盟. 3D 数字人质量分级技术要求. 北京: 世界超高清视频产业联盟, 2023.)

[77] Duan H, Hu Q, Wang J, et al. Finevq: Fine-grained user generated content video quality assessment//Proceedings of the Computer Vision and Pattern Recognition Conference. Nashville, USA, 2025: 3206-3217.

[78] Yan J B, Fang Y M, Liu X L, et al. A review of research on video quality evaluation. Journal of Computer Science and Technology, 2023, 46(10): 2196-2224.
(鄢杰斌, 方玉明, 刘学林, 等. 视频质量评价研究综述. 计算机学报, 2023, 46(10): 2196-2224.)

[79] Bampis C G, Gupta P, Soundararajan R, et al. Speed-qa: Spatial efficient entropic differencing for image and video quality. IEEE Signal Processing Letters, 2017, 24(9): 1333-1337.

[80] Zhang R, Isola P, Efros A A, et al. The unreasonable effectiveness of deep features as a perceptual metric//Proceedings of the IEEE conference on computer vision and pattern recognition. Salt Lake City, USA, 2018: 586-595.

[81] Unterthiner T, Van Steenkiste S, Kurach K, et al. Towards accurate generative models of video: A new metric & challenges. arXiv preprint arXiv:1812.01717, 2018.

[82] Zhang R, Isola P, Efros A A, et al. The unreasonable effectiveness of deep features as a perceptual metric// Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 586-595.

[83] Liu S, Deng W. Very deep convolutional neural network based image classification using small training sample size// Proceedings of the 2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR). Kuala Lumpur, Malaysia, 2015: 730-734.

[84] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks. Communication ACM, 2017, 60(6): 84-90.

[85] Shi W, Sun Y, Li S, et al. Spatial and temporal feature-based reduced reference quality assessment for rate-varying videos in wireless networks. International Journal of Pattern Recognition and Artificial Intelligence, 2019, 33(13): 1950021.

[86] Narvekar N D, Karam L J. A no-reference image blur metric based on the cumulative probability of blur detection (cpbd). IEEE Transactions on Image Processing, 2011, 20(9): 2678-2683.

[87] Korhonen J. Two-level approach for no-reference consumer video quality assessment. IEEE Transactions on Image Processing, 2019, 28(12): 5923-5938.

[88] Li D, Jiang T, Jiang M. Quality assessment of in-the-wild videos //Proceedings of the 27th ACM international conference on multimedia. Nice, France, 2019: 2351-2359.

[89] Jiang S, Sang Q, Hu Z, et al. Self-supervised representation learning for video quality assessment. IEEE Transactions on Broadcasting, 2022, 69(1): 118-129.

[90] Wu W, Li Q, Chen Z, et al. Semantic information oriented no-reference video quality assessment. IEEE Signal Processing Letters, 2021, 28: 204-208.

[91] Sinno Z, Bovik A C. Large-scale study of perceptual video quality. IEEE Transactions on Image Processing, 2018, 28(2): 612-627.

[92] Xu J, Li J, Zhou X, et al. Perceptual quality assessment of internet videos//Proceedings of the 29th ACM International Conference on Multimedia. Chengdu, China, 2021: 1248-1257.

[93] Yi F, Chen M, Sun W, et al. Attention based network for no-reference ugc video quality assessment//Proceedings of the 2021 IEEE International Conference on Image Processing (ICIP). Anchorage, USA, 2021: 1414-1418.

[94] Wang Y, Inguva S, Adsumilli B. YouTube UGC dataset for video compression research//Proceedings of the 2019 IEEE 21st international workshop on multimedia signal processing (MMSP). Kuala Lumpur, Malaysia, 2019: 1-5.

[95] Ying Z, mandal M, Ghadiyaram D, et al. Patch-vq: ’patching up’ the video quality problem//Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Virtual, 2021: 14014-14024.

[96] Hosu V, Hahn F, Jenadeleh M, et al. The konstanz natural video database (KoNViD-1k)//2017 Ninth international conference on quality of multimedia experience (QoMEX). Erfurt, Germany, 2017: 1-6.

[97] Wang Y, Ke J, TalebI H, et al. Rich features for perceptual quality assessment of UGC videos//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Virtual, 2021: 13430-13439.

[98] Xing F, Wang Y G, Wang H, et al. Starvqa: Space-time attention for video quality assessment//2022 IEEE International Conference on Image Processing (ICIP). Bordeaux, France, 2022: 2326-2330.

[99] Kou T, Liu X, Zhang Z, et al. Subjective-aligned dataset and metric for text-to-video quality assessment//Proceedings of the 32nd ACM International Conference on Multimedia. Melbourne, Australia, 2024: 7793-7802.

[100] Varga D. No-reference video quality assessment using multi-pooled, saliency weighted deep features and decision fusion. Sensors, 2022, 22(6): 2209.

[101] Telili A, Fezza S A, Hamidouche W, et al. 2bivqa: Double bi-lstm-based video quality assessment of ugc videos. ACM Transactions on Multimedia Computing, Communications, and Applications, 2023, 20(4): Article 100.

[102] Li B, Zhang W, Tian M, et al. Blindly assess quality of in-the-wild videos via quality-aware pre-training and motion perception. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(9): 5944-5958.

[103] Wu H, Chen C, Hou J, et al. Fast-vqa: Efficient end-to-end video quality assessment with fragment sampling//Proceedings of the European conference on computer vision. Tel Aviv, Israel, 2022: 538-554.

[104] Wu H, Chen C, Liao L, et al. Discovqa: Temporal distortion content transformers for video quality assessment. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33(9): 4840-4854.

[105] Wu H, Zhang Z, Zhang W, et al. Q-ALIGN: teaching LMMs for visual scoring via discrete text-defined levels//Proceedings of the 41st International Conference on Machine Learning. Vienna, Austria, 2024: 54015-54029.

[106] Zhao K, Yuan K, Sun M, et al. Zoom-vqa: Patches, frames and clips integration for video quality assessment//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada, 2023: 1302-1310.

[107] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision//Proceedings of the International conference on machine learning. Virtual, 2021: 8748-8763.

[108] Liu Y, Quan Y, Xiao G, et al. Scaling and masking: a new paradigm of data sampling for image and video quality assessment// Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence. Vancouver, Canada, 2024, 38(4): 3792-3801.

[109] Zavras A, Michail D, Zhu X X, et al. GAIA: A global, multi-modal, multiscale vision-language dataset for remote sensing image analysis. IEEE Geoscience and Remote Sensing Magazine, 2026, 14(2), 36-63.

[110] He C, Zheng Q, Zhu R, et al. COVER: A comprehensive video quality evaluator//Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Seattle, USA, 2024: 5799-5809.

[111] Wang X, Katsenou A, Shen J, et al. Camp-vqa: Caption embedded multimodal perception for no-reference quality assessment of compressed video//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Denver, USA, 2026: 2042-2051.

[112] Nuutinen M, Virtanen T, Vaahteranoksa M, et al. CVD2014-A database for evaluating no-reference video quality assessment algorithms. IEEE Transactions on Image Processing, 2016, 25(7): 3073-3086.

[113] Ge Q, Sun W, Zhang Y, et al. Lmm-vqa: Advancing video quality assessment with large multimodal models. IEEE Transactions on Circuits and Systems for Video Technology, 2025, 35(11): 11083-11096.

[114] Zhang Z, Zhou Y, Sun W, et al. Geometry-aware video quality assessment for dynamic digital human// Proceedings of the 2023 IEEE International Conference on Image Processing (ICIP). Kuala Lumpur, Malaysia, 2023: 1365-1369.

[115] Feichtenhofer C, Fan H, Malik J, et al. Slowfast networks for video recognition//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, South Korea, 2019: 6201-6210.

[116] Duggal S, Hu Y, Michel O, et al. Eval3d: Interpretable and fine-grained evaluation for 3d generation//Proceedings of the Computer Vision and Pattern Recognition Conference. Nashville, USA, 2025: 13326-13336.

[117] M. Torcoli, T. Kastner and J. Herre. Objective measures of perceptual audio quality reviewed: An evaluation of their application domain dependence. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 1530-1541.

[118] Torcoli M, Kastner T, Herre J. Perceptual objective listening quality assessment (polqa), the third generation itu-t standard for end-to-end speech quality measurement part i-temporal alignment. Journal of the audio engineering society, 2017, 61(9): 1333-1337.

[119] Eamdeelerd C, Songwatana K. Audio noise classification using bark scale features and knn technique//Proceedings of the 2008 International Symposium on Communications and Information Technologies. Vientiane, Laos, 2008: 131-134.

[120] Malfait L, Berger J, Kastner M. P.563-The itu-t standard for single-ended speech quality assessment. IEEE Transactions on Audio, Speech, and Language Processing, 2006, 14(6): 1924-1934.

[121] Kim D S, Tarraf A. ANIQUE+: A new American national standard for non-intrusive estimation of narrowband speech quality. Bell Labs Technical Journal, 2007, 12(1): 221-236.

[122] Falk T H, Möller S, Karaiskos V, Et Al. Improving instrumental quality prediction performance for the blizzard challenge//Blizzard Challenge. The University of Edinburgh, UK, 2008: 1-6.

[123] Norrenbrock C, Hinterleitner F, Heute U, et al. Towards perceptual quality modeling of synthesized audiobooks-blizzard challenge 2012//The Blizzard Challenge, Portland, USA, 2012: 59-64.

[124] Lo C C, Fu S W, Huang W C, et al. Mosnet: Deep learning-based objective assessment for voice conversion//Proceedings of the International Symposium on Computer Architecture. Phoenix, USA, 2019: 1541-1545.

[125] Lorenzo-Trueba J, Yamagishi J, Toda T, et al. The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods. arXiv preprint arXiv: 1804.04262, 2018.

[126] Cooper E, Huang W C, Toda T, et al. Generalization ability of MOS prediction networks//ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Singapore, 2022: 8442-8446.

[127] Cooper E, Yamagishi J. How do voices from past speech synthesis challenges compare today? arXiv preprint arXiv:2105.02373, 2021.

[128] Maiti S, Peng Y, Saeki T, et al. Speechlmscore: evaluating speech generation using speech language model//ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Rhodes, Greece, 2023: 1-5.

[129] Huang W C, Fu S W, Cooper E, et al. The Voice MOS challenge 2024: Beyond speech quality prediction//2024 IEEE Spoken Language Technology Workshop (SLT). Macao, China, 2024: 803-810.

[130] Wang C C, Huang K T, Yang C Y, et al. Qamro: Quality-aware adaptive margin ranking optimization for human-aligned assessment of audio generation systems//2025 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). Honolulu, USA, 2025: 1-4.

[131] Huang W C, Wang H, Liu C, et al. The audiomos challenge 2025//2025 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). Honolulu, USA, 2025: 1-8.

[132] Lin Y C, Chen J H, Lee H Y. Mmmos: Multi-domain multi-axis audio quality assessment//2025 IEEE Automatic Speech Recognition and Understanding Workshop. Honolulu, USA, 2025: 1-8.

[133] Tjandra A, Wu Y C, Guo B, et al. Meta audiobox aesthetics: Unified automatic assessment for speech, music and sound//2025 IEEE Automatic Speech Recognition and Understanding Workshop. Honolulu, USA, 2025: 125128 vol.1.

[134] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation, 1997, 9(8): 1735-1780.

[135] Deshmukh S, Alharthi D, Elizalde B, et al. Pam: prompting audio-language models for audio quality assessment. arXiv preprint arXiv: 2402.00282, 2024.

[136] Manocha P, Williamson D, Finkelstein A, et al. Corn: Co-trained full and no-reference speech quality assessment//49th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). Seoul, Korea, 2024: 376-380.

[137] Deng J, Guo J, Xue N, et al. Arcface: Additive angular margin loss for deep face recognition//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA, 2019: 4685-4694.

[138] Liu W, Wen Y, Yu Z, et al. Sphereface: deep hypersphere embedding for face recognition//Proceedings of the IEEE conference on computer vision and pattern recognition. Hawaii, USA, 2017: 212-220.

[139] Wang H, Wang Y, Zhou Z, et al. Cosface: large margin cosine loss for deep face recognition//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 5265-5274.

[140] Zhang Z, Zhou Y, Sun W, et al. Perceptual quality assessment for digital human heads//ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Rhodes, Greece, 2023: 1-5.

[141] Xu L, Zhou Y, Liu S, et al. Facial quality assessment of digital humans: A dual-branch framework integrating morphological harmony and expressive coordination. Displays, 2026, 91: 103221.

[142] Chung J S, Zisserman A. Out of time: Automated lip sync in the wild//Lecture Notes in Computer Science: 13th Asian Conference on Computer Vision (ACCV): vol. 10117. Taipei, Taiwan, China, 2016: 251-263.

[143] Prajwal K, Mukhopadhyay R, Namboodiri V P, et al. A lip sync expert is all you need for speech to lip generation in the wild//Proceedings of the 28th ACM international conference on multimedia. Virtually and in Seattle, Washington, USA, 2020: 484-492.

[144] Chen L L, Li Z H, Maddox R K, et al. Lip movements generation at a glance//Lecture Notes in Computer Science: 15th European Conference on Computer Vision (ECCV): vol. 11211. Munich, Germany, 2018: 538-553.

[145] Felzenszwalb P F, Girshick R B, Mcallester D, et al. Object detection with discriminatively trained part-based models. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, 32(9): 1627-1645.

[146] Kubichek R. Mel-cepstral distance measure for objective speech quality assessment//Proceedings of IEEE Pacific Rim Conference on Communications Computers and Signal Processing: vol. 1. Victoria, Canada,1993: 125-128.

[147] Zhang J, Long Z Y, Zhang B, et al. A survey of pitch extraction algorithms in speech signal processing. Journal of Electronic Science and Technology, 2010, 39(S1): 99-102+126.
(张杰, 龙子夜, 张博, 等. 语音信号处理中基频提取算法综述. 电子科技大学学报, 2010, 39(S1): 99-102+126.)

[148] Nakatani T, Amano S, Irino T, et al. A method for fundamental frequency estimation and voicing decision: Application to infant utterances recorded in real acoustical environments. Speech Communication, 2008, 50(3): 203-214.

[149] Chu W, Alwan A, IEEE. Reducing f0 frame error of f0 tracking algorithms under noisy conditions with an unvoiced/voiced classification frontend//IEEE International Conference on Acoustics, Speech and Signal Processing. Taipei, China, 2009: 3969-3972.

[150] Quignon N, Chopin B, Wang Y, et al. THEval. Evaluation Framework for Talking Head Video Generation. arXiv preprint arXiv: 2511.04520, 2025.

[151] Lugaresi C, Tang J, Nash H, et al. Mediapipe: A framework for perceiving and processing reality//Third workshop on computer vision for AR/VR at IEEE Computer Vision and Pattern Recognition (CVPR). Long Beach, USA, 2019: 2.

[152] Narayan K, Vs V, Chellappa R, et al. Facexformer: A unified transformer for facial analysis//Proceedings of the IEEE/CVF International Conference on Computer Vision. Xi'an, China, 2025: 11369-11382.

[153] Chen X, He K, Liu W, et al. Clam: An open-source library for performance evaluation of text-driven human motion generation//Proceedings of the 32nd ACM International Conference on Multimedia. Melbourne, Australia, 2024: 11194-11197.

[154] Han H, Li S, Chen J, et al. Video-bench: Human-aligned video generation benchmark//Proceedings of the Computer Vision and Pattern Recognition Conference. Nashville, USA, 2025: 18858-18868.

[155] Kant Y, Weber E, Kim J K, et al. Pippo: High-resolution multiview humans from a single image//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA, 2025: 16418-16429.

[156] Detone D, Malisiewicz T, Rabinovich A. Superpoint: Self-supervised interest point detection and description//Proceedings of the IEEE conference on computer vision and pattern recognition workshops. Salt Lake City, USA, 2018: 224-236.

[157] Xie D, Li J, Tan H, et al. Carve3d: improving multi-view reconstruction consistency for diffusion models with rl-finetuning//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2024: 6369-6379.

[158] Asim M, Wewer C, Wimmer T, et al. Met3r: measuring multi-view consistency in generated images//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA, 2025: 6034-6044.

[159] Duggal S, Hu Y, Michel O, et al. Eval3d: Interpretable and fine-grained evaluation for 3d generation//Proceedings of the Computer Vision and Pattern Recognition Conference. Nashville, USA, 2025: 13326-13336.

[160] Kossaifi J, Walecki R, Panagakis Y, et al. Sewadb: A rich database for audio-visual emotion and sentiment research in the wild. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(3): 1022-1040.

[161] Busso C, Bulut M, Lee C C, et al. IEMOCAP: interactive emotional dyadic motion capture database. Language Resources and Evaluation, 2008, 42(4): 335-359.

[162] Ringeval F, Sonderegger A, Sauer J, et al. Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions// Proceedings of the 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition. Shanghai, China, 2013: 1-8.

[163] Lucey P, Cohn J F, Kanade T, et al. The Extended CohnKanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression// Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Francisco, USA, 2010: 94-101.

[164] Dhall A, Goecke R, Lucey S, et al. Acted facial expressions in the wild database. Australian National University, Canberra, Australia, Technical Report TR-CS-11, 2011, 2(1): 1-13.

[165] Yan W J, Li X, Wang S J, et al. CASME II: an improved spontaneous micro-expression database and the baseline evaluation. PLoS One, 2014, 9(1): e86041.

[166] Kuehne H, Jhuang H, Garrote E, et al. HMDB: A large video database for human motion recognition// Proceedings of the 2011 International Conference on Computer Vision. Barcelona, Spain, 2011: 2556-2563.

[167] Soomro K, Zamir A R, Shah M. Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402, 2012.

[168] Marszalek M, Laptev I, Schmid C. Actions in context// Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami Beach, USA, 2009: 2929-2936.

[169] Luo W, Wang H F. Review of large language model evaluation. Journal of Chinese Information Processing, 2024, 38(01): 1-23. (罗文, 王厚峰. 大语言模型评测综述. 中文信息学报, 2024, 38(01): 1-23.)

[170] Huang Y, Bai Y, Zhu Z, et al. C-Eval: A multi-level multidiscipline Chinese evaluation suite for foundation models//Advances in Neural Information Processing Systems. New Orleans, USA, 2023: 62991-63010.

[171] Li H, Zhang Y, Koto F, et al. CMMLU: Measuring massive multitask language understanding in Chinese//Findings of the Association for Computational Linguistics: ACL 2024. Bangkok, Thailand, 2024: 11260-11285.

[172] Arabzadeh N, Clarke C L. Benchmarking LLM-based relevance judgment methods//Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval. Padua, Italy, 2025: 3194-3204.

[173] Phan L, Gatti A, Han Z, et al. Humanity's last exam. arXiv preprint arXiv: 2501.14249, 2025.

[174] White C, Dooley S, Roberts M, et al. LiveBench: a challenging, contamination-free LLM benchmark// Proceedings of the Thirteenth International Conference on Learning Representations. Singapore, 2025: 1-37.

[175] Sun H, Zhang Z, Deng J, et al. Safety assessment of Chinese large language models[J]. arXiv preprint arXiv:2304.10436, 2023.

[176] Bommasani R, Liang P, Lee T. Holistic evaluation of language models. Annals of the New York Academy of Sciences, 2023, 1525(1): 140-146.

[177] Zhang J, Sanderson A C. JADE: adaptive differential evolution with optional external archive. IEEE Transactions on Evolutionary Computation, 2009, 13(5): 945-958.

[178] Zhang Z, Sun W, Zhou Y, et al. Advancing zero-shot digital human quality assessment through text-prompted evaluation. IEEE Transactions on Image Processing, 2025, 34: 3503-3517.

[179] Zheng D, Huang Z, Liu H, et al. Vbench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. arXiv preprint arXiv:2503.21755, 2025.

[180] Zhu D. SIFT algorithm analysis and optimization// Proceedings of the 2010 International Conference on Image Analysis and Signal Processing. Xiamen, China, 2010: 415-419.

[181] Suju D A, Jose H. FLANN: Fast approximate nearest neighbor search algorithm for elucidating human-wildlife conflicts in forest areas// Proceedings of the 2017 Fourth International Conference on Signal Processing, Communication and Networking (ICSCN). Chennai, India, 2017: 1-6.

[182] Dosovitskiy A, Beyer L, Kolesnikov a, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale// Proceedings of the International Conference on Learning Representations. Virtual, 2021.

[183] Li D, Jiang T, Jiang M. Quality assessment of in-the-wild videos//Proceedings of the 27th ACM international conference on multimedia. Nice, France, 2019: 2351-2359.

[184] Wu H, Zhang E, Liao L, et al. Exploring video quality assessment on user generated contents from aesthetic and technical perspectives// Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France, 2023: 20087-20097.

[185] Liu Z, Ning J, Cao Y, et al. Video swin transformer// Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada, 3192-3201.

[186] Zhu Y F, Xu X, Zhang X P. Breaking my image: Construction and application of a 'digital human' evaluation model. Library Triune, 2023, 43(02): 132-140.
(朱奕帆, 许鑫, 张煦频. 勘破我相：“数字人”测评模型构建与应用. 图书馆论坛, 2023, 43(02): 132-140.)

[187] Zhang Z, Zhou Y, Sun W, et al. Perceptual quality assessment for digital human heads//ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Rhodes, Greece, 2023: 1-5.

[188] Zhou Y, Zhang Z, Sun W, et al. Thqa: A Perceptual Quality Assessment Database for Talking Heads//Proceedings of the 2024 IEEE International Conference on Image Processing (ICIP). Abu Dhabi, United Arab Emirates, 2024: 15-21.

[189] Karras T, Laine S, Aittala M, et al. Analyzing and improving the image quality of style gan//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Virtual, 2020: 8110-8119.

[190] Zhou Y, Cao J, Zhang Z, et al. Who is a better talker: subjective and objective quality assessment for AI-generated talking heads//Proceedings of the IEEE/CVF International Conference on Computer Vision. Shenzhen, Greater Bay Area, China, 2025: 12201-12211.

[191] Götz-HahN F, Hosu V, Lin H, et al. Konvid-150k: A dataset for no-reference video quality assessment of videos in-the-wild. IEEE Access, 2021, 9: 72139-72160.

[192] Danier D, Zhang F, Bull D R. BVI-VFI: A video quality database for video frame interpolation. IEEE Transactions on Image Processing, 2023, 32: 6004-6019.

[193] Madhusudana P C, Yu X, Birkbeck N, et al. Subjective and objective quality assessment of high frame rate videos. IEEE Access, 2021, 9: 108069-108082.

[194] Wu H, Zhang E, Liao L, et al. Exploring video quality assessment on user generated contents from aesthetic and technical perspectives// Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France, 2023: 20087-20097.

[195] Lu Y, Li X, Pei Y, et al. Kvq: Kwai video quality assessment for short-form videos//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA, 2024: 25963-25973.

[196] Zhang Z, Wu W, Sun W, et al. Md-vqa: Multi-dimensional quality assessment for ugc live videos// Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada, 2023: 1746-1755.

[197] Wu H, Zhang E, Liao L, et al. Towards explainable in-the-wild video quality assessment: A database and a language-prompted approach//MM'23: Proceedings of the 31st ACM International Conference on Multimedia. Ottawa, Canada, 2023: 1045-1054.

[198] Liang G, Zhang B, Wang Y, et al. Lg-vq: Language-guided codebook learning. arXiv preprint arXiv:2405.14206, 2024.

[199] Chen Z, Sun W, Tian Y, et al. Gaia: rethinking action quality assessment for ai-generated videos. Advances in Neural Information Processing Systems, 2024, 37: 40111-40144.

[200] Huang Z, He Y, Yu J, et al. Vbench: comprehensive benchmark suite for video generative models//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2024: 21807-21818.

[201] Wang S, Yu W, Chen X, et al. Qualispeech: a speech quality assessment dataset with natural language reasoning and descriptions//Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vienna, Austria, 2025: 23588-23609.

[202] Thiemann J, Ito N, Vincent E. The diverse environments multi-channel acoustic noise database (demand): A database of multichannel environmental noise recordings. Proceedings of Meetings on Acoustics, 2013, 19(1): 35-81.

[203] Fonseca E, Plakal M, Ellis D P W, et al. Learning sound event classifiers from web audio with noisy labels// Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton, UK, 2019: 21-25.

[204] Mysore G J. Can we automatically transform speech recorded on common consumer devices in real-world environments into professional production quality speech?-a dataset, insights, and challenges. IEEE Signal Processing Letters, 2015, 22(8): 1006-1010.

[205] Torcoli M, Wu C W, Dick S, et al. Odaq: Open dataset of audio quality//ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Seoul, Korea, 2024: 836-840.

[206] Li K, Wang Y, Li Y, et al. Unmasked teacher: towards training efficient video foundation models//Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris, France, 2023: 19948-19960.

[207] Zhao L, Gundavarapu N B, Yuan L, et al. Videoprism: A foundational visual encoder for video understanding. arXiv preprint arXiv: 2402.13217.

[208] Xu H, Ghosh G, Huang P Y, et al. Videoclip: Contrastive pretraining for zero-shot video-text understanding//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Online and Punta Cana, Dominican Republic, 2021: 6787-6800.

[209] Arnab A, Dehghani M, Heigold G, et al. Vivit: A video vision transformer// Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada, 2021: 6836-6846.

[210] Huang J, Zhang C, Dolby J. CLAP: recording local executions to reproduce concurrency failures//Proceedings of the 34th ACM SIGPLAN Conference on Programming Language Design and Implementation. New York, USA, 2013: 141-152.

[211] Bharadwaj S, Cornell S, Choi K, et al. Openbeats: A fully open-source general-purpose audio encoder//Proceedings of the 2025 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA). Lake Tahoe, USA, 2025: 1-5.

[212] Huang P Y, Xu H, Li J, et al. Masked autoencoders that listen. Advances in neural information processing systems. New Orleans, USA, 2022: 28708-28720.

[213] Pratap V, Tjandra A, Shi B W, et al. Scaling speech technology to 1,000+languages. Journal of Machine Learning Research, 2024, 25(97): 1-52.

[214] Ji J, Zhou J, Lou H, et al. Align anything: training all-modality models to follow instructions with language feedback. arXiv preprint arXiv:2412.15838, 2024.

[215] Girdhar R, El-Nouby A, Liu Z, et al. Imagebind one embedding space to bind them all//Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada, 2023: 15180-15190.

[216] Lu J, Clark C, Lee S, et al. Unified-io 2: Scaling autoregressive multimodal models with vision language audio and action//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2024: 26439-26455.

[217] Wu J, Gan W, Chen Z, et al. Multimodal large language models: A survey// Proceedings of the 2023 IEEE International Conference on Big Data (BigData). 2023: 2247-2256.

[218] Ziegler D M, Stiennon N, Wu J, et al. Fine-tuning language models from human preferences. arXiv preprint arXiv:1909.08593, 2019.

[219] Cheng T, Song L, Ge Y, et al. Yolo-world: real-time open vocabulary object detection//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2024: 16901-16911.

[220] Carion N, Massa F, Synnaeve G, et al. End-to-end object detection with transformers// Proceedings of the European Conference on Computer Vision. Glasgow, UK, 2020: 213-229.

[221] Li L H, Zhang P, Zhang H, et al. Grounded language-image pre-training//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: 2022: 10965-10975.

[222] Huang P Y, Xu H, Li J, et al. Masked autoencoders that listen. Advances in Neural Information Processing Systems, 2022, 35: 28708-28720.

[223] Liu Z, Lin Y, Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows// Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada, 2021: 9992-10002.

[224] Todi A, Narula N, Sharma M, et al. Convnext: A contemporary architecture for convolutional neural networks for image classification// Proceedings of the 2023 3rd International Conference on Innovative Sustainable Computational Technologies (CISCT). Dehradun, India, 2023: 1-6.

[225] Hu F S, Zhang M J, Zou B J, et al. Single sample face recognition with variable illumination and pose based on HMM. Journal of Computer Science and Technology, 2009, 32(07): 1424-1433.
(胡峰松, 张茂军, 邹北骥, 等. 基于 HMM 的单样本可变光照、姿态人脸识别. 计算机学报, 2009, 32(07): 1424-1433.)

[226] Deng J, Guo J, Ververas E, et al. Retinaface: Single-shot multi-level face localization in the wild// Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA, 2020: 5202-5211.

[227] Yang T Y, Chen Y T, Lin Y Y, et al. Fsa-net: Learning fine-grained structure aggregation for head pose estimation from a single image// Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA, 2019: 1087-1096.

[228] Perikamana Narayanan S, Sabarimalai Manikandan M, Cenkeramaddi L R. Yolov9-based human face detection and counting under human-animal faces, complex imaging environments, and image qualities. IEEE Access, 2025, 13: 129600-129637.

[229] Wang Z W, Pu Y Y, Wang X, et al. Accurate retrieval of multi-scale clothing image based on multi feature fusion. Journal of Computer Science and Technology, 2020, 43(4): 740-754.
(王志伟, 普园媛, 王鑫, 等. 基于多特征融合的多尺度服装图像精准化检索. 计算机学报, 2020, 43(4): 740-754.)

[230] Redmon J, Farhadi A. Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767, 2018.

[231] Li J, Zhou Y, Chu H, et al. Dual multi-scale gcn with deformable temporal kernel for skeleton-based action recognition//ICASSP 2025 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Hyderabad, India, 2025: 1-5.

[232] Zhou R, Wang J. Dynamic adaptive graph convolution with attention for skeleton-based action recognition// Proceedings of the 2025 2nd International Conference on Informatics Education and Computer Technology Applications (IECA). Kuala Lumpur, Malaysia, 2025: 79-82.

[233] Li T, Chen X, Huang S, et al. Topology unshared graph convolutional networks for skeleton-based action recognition// Proceedings of the 2024 9th International Conference on Automation, Control and Robotics Engineering (CACRE). Jeju Island, Korea 2024: 270-274.

[234] Desplanques B, Thienpondt J, Demuynck K. Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification// Proceedings of the Interspeech 2020. Shanghai, China, 2020: 3830-3834.

[235] Chen S, Wang C, Chen Z, et al. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. IEEE Journal of Selected Topics in Signal Processing, 2022, 16(6): 1505-1518.

[236] Hu C H, Peng Y H, Yamagishi J, et al. SVSNet: an end-to-end speaker voice similarity assessment model. IEEE Signal Processing Letters, 2022, 29: 767-771.

[237] Wang J, Duan H, Zhai G, et al. Aigv-assessor: benchmarking and evaluating the perceptual quality of text-to-video generation with lmm//Proceedings of the Computer Vision and Pattern Recognition Conference. Seattle, USA, 2025: 18869-18880.

[238] Liu Y, Cun X, Liu X, et al. Evalcrafter: benchmarking and evaluating large video generation models//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2024: 22139-22149.

JIANG Yi-bi, Ph.D. candidate. His research interests include multimodal learning and AIGC detection.

QIN Chuan, Ph.D., professor, Ph.D. supervisor. His research interests include multimedia security and AI security

QIAN Zhen-Xing, Ph.D., professor, Ph.D. supervisor. His research interests include information hiding and AI security.

ZHANG Xin-Peng, Ph.D., professor, Ph.D. supervisor. His research interests include multimedia security and AI security.