ICS 35.240
CCS

T/CAPT

团　　体　　标　　准

T/CAPT 018—2025

多模态交互式数智人应用技术要求

Multimodality Interaction (Digital Human) Application Technical Requirements

2025-12-03 发布　　2025-12-03 实施

中国新闻技术工作者联合会 发布

目　录

前　言

本文件按照GB/T 1.1—2020《标准化工作导则 第1部分：标准化文件的结构和起草规则》的规定起草。

请注意本文件的某些内容可能涉及专利。本文件的发布机构不承担识别这些专利的责任。

本文件由中国新闻技术工作者联合会提出并归口。

本文件起草单位：湖北长江云新媒体集团有限公司、华为云计算技术有限公司、新华通讯社通信技术局、新华网股份有限公司、新华社国家重点实验室生物感知智能应用研究部、新华网融媒体未来研究院、中国互联网新闻中心、湖北省科学技术厅、华中科技大学、武汉大学、武汉理工大学、华中师范大学、中南民族大学、湖北经济学院、南京航空航天大学、武汉广播电视台、鹤峰县融媒体中心、巴东县融媒体中心、湖北省科技信息研究院、江苏天闳低空数字技术研究院有限公司、中联超清（北京）科技有限公司、中广核（湖北）新能源投资有限公司、楚天龙股份有限公司、长江云数创（湖北）科技有限公司、空间视创（重庆）科技股份有限公司、武汉东湖大数据科技股份有限公司。

本文件主要起草人：朱昊、代卓、李瑛、蒋大可、李俊杰、李康、叶紫薇、万静、余艳君、熊溢平、路海燕、张芮宁、杨溟、杨昊、王建平、王岳、傅丽枫、刘云、杨铀、刘琼、彭蓉、周俊伟、孙建文、李启磊、帖军、郑禄、张潇、黎桦、张洪海、刘光国、傅静琴、周金平、向锋、邓坤烘、徐大凯、王付生、赵海亮、田祥、张超、刘芳、王晶、杜登伟。

引　言

近年来，元宇宙和人工智能技术快速发展，成为推动数字经济发展的核心驱动力。2022年，工业和信息化部、教育部、文化和旅游部、国家广播电视总局、国家体育总局五部门联合发布《虚拟现实与行业应用融合发展行动计划（2022—2026年）》，明确提出加速虚拟现实、人工智能等技术的融合创新，构建元宇宙产业生态。多模态交互式数智人作为元宇宙与人工智能技术的重要载体，正在重塑人机交互方式，并在传媒、金融、教育、文旅等领域展现出广阔的应用前景。

为规范多模态交互式数智人的技术研发与应用，提升其在元宇宙环境下的交互能力、智能水平及安全性，中国新闻技术工作者联合会经广泛调研，依据国家元宇宙和人工智能相关产业政策，特组织编制本技术要求。

本文件围绕多模态交互式数智人的基础架构、交互技术、数据合规、应用场景等核心内容，提出系统化的技术要求与评估体系，旨在推动数智人技术在元宇宙生态中的标准化落地。通过优化多模态融合、强化数据隐私保护、提升用户体验，本文件将为元宇宙和人工智能产业的健康发展提供技术支撑，助力构建更加智能、沉浸、安全的数字交互新范式。

多模态交互式数智人应用技术要求

1 范围

本文件规定了多模态交互式数智人在技术架构、交互能力、安全合规、性能评估及应用实施等方面的技术要求。

本文件适用于通讯社、报社、广播电台、电视台、杂志社、网络媒体等多种媒体机构在传媒、金融、教育、文旅等领域运用多模态交互式数智人技术有关的研发、部署与应用等。

2 规范性引用文件

下列文件对于本文件的应用是必不可少的。凡是注日期的引用文件，仅注日期的版本适用于本文件。凡是不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

GY/T 411-2024 数字虚拟人技术要求

3 术语和定义

下列术语和定义适用于本文件。

3.1 数字虚拟人 digital human

基于现实世界设计，通过计算机生成，再借助真人或计算驱动，在多模态输出设备呈现的虚拟人物。

注：简称为数字人或虚拟人。

[来源：GY/T 411-2024，3.1]

3.2 多模态交互式数智人 multimodality interactive digital intelligence human

是指能够融合视觉、听觉、触觉等多种感知方式，实现与用户进行多模态智能交互的人工智能系统。

注：是从 “数字虚拟人” 演进而来，具备更全面的感知和交互能力，能够处理文本、语音、图像、视频、手势等多种输入，并通过语音合成、面部表情、肢体动作、情感表达等多种输出方式与用户进行互动。

3.3 跨模态协同 cross-modality collaboration

是指多模态交互式数智人系统中不同模态之间的协调配合机制，包括同步、互补与冲突解决等，以实现更高效、更自然的交互。

3.4 情绪感知 emotional perception

是指通过多模态生理/行为信号（如皮肤电、脑电、肌电、心电、图像等）的量化分析，利用机器学习模型客观推断情绪状态的技术，核心流程包括信号采集、特征提取与模式识别。

4 缩略语

下列缩略语适用于本文件。

DDoS：分布式拒绝服务（攻击） Distributed Denial of Service

DevOps：开发运维（一体化） Development and Operations

FPS：帧率/帧每秒 Frames Per Second

GMV：商品交易总额 Gross Merchandise Volume

RPO：恢复点目标 Recovery Point Objective

PII：个人身份信息 Personally Identifiable Information

PTP：精确时间协议 Precision Time Protocol

RTO：恢复时间目标 Recovery Time Objective

SSD：固态硬盘 Solid State Drive

WebRTC：网页实时通信 Web Real-Time Communication

5 基础框架

5.1 技术分类

多模态交互式数智人基于不同的技术特点，可分为：

a) 实时交互型数智人：支持即时响应用户输入，适用于智能客服、实时对话等场景，对交互流畅性和响应速度要求较高。

b) 预生成内容型数智人：主要用于展示预先生成的内容，如固定节目播报，交互性较弱，注重内容呈现效果。

c) 混合交互型数智人：结合实时交互与预生成内容，适用于电商直播等场景，既能实时互动，又能按预设流程执行任务。

5.2 逻辑架构

多模态交互式数智人系统的逻辑架构包括表现层、感知层、认知层和支撑层，如图1所示：

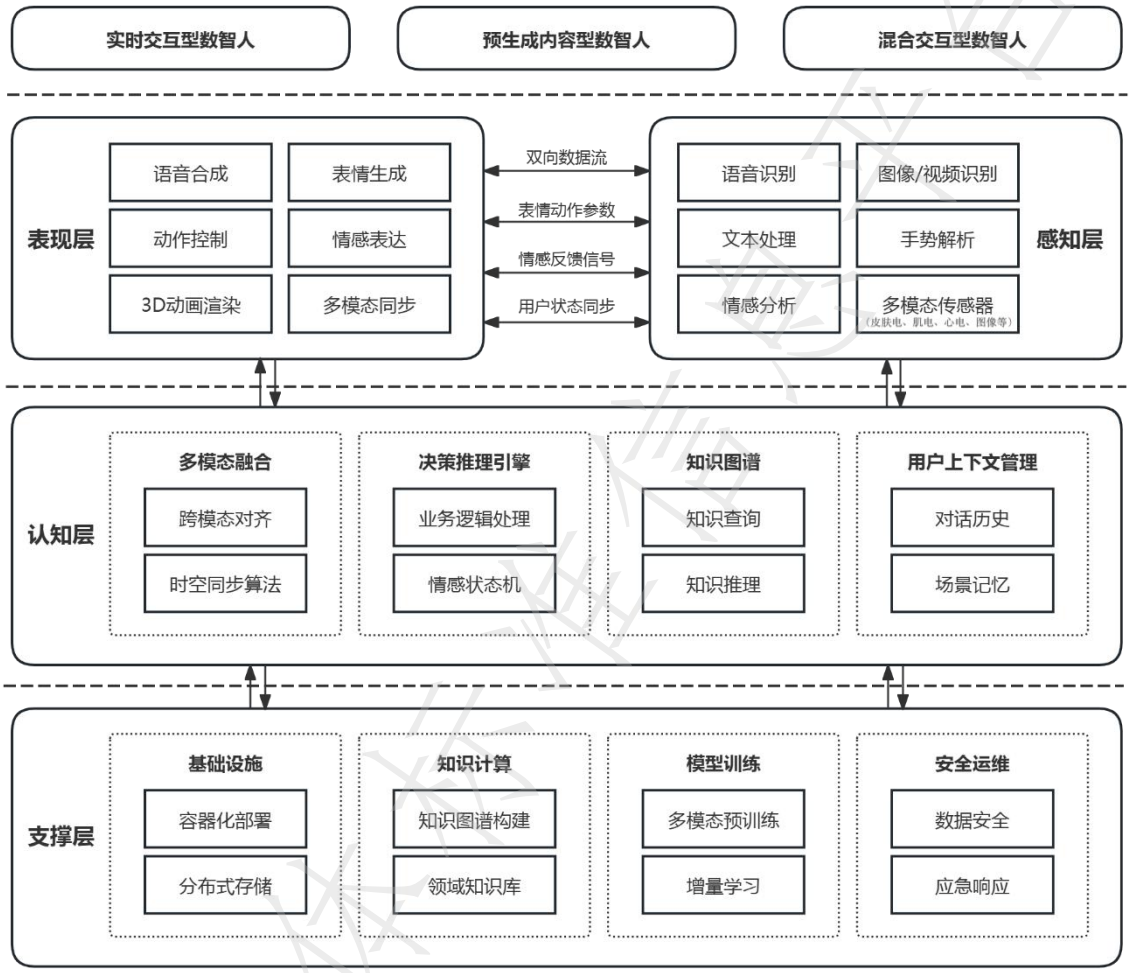

图1 多模态交互式数智人系统逻辑架构图

5.2.1 表现层

将认知层的决策结果转化为多模态输出；确保语音、表情、动作的时空一致性；根据交互场景调整表现风格；实现自然流畅的人机交互体验。

5.2.2 感知层

实时采集用户的多模态输入信息（语音、图像、文本、手势等）；对输入信息进行预处理和特征提取；将处理后的结构化数据传递给认知层；确保在复杂环境（如噪声、光照变化等）下的稳定识别。

5.2.3 认知层

对感知层传递的信息进行深度语义理解；结合知识图谱进行推理决策；根据用户历史交互数据实现个性化响应；维护多轮对话的上下文一致性。

5.2.4 支撑层

提供高性能计算资源；保障数据安全存储与快速检索；支持算法模型的训练与部署；实现系统的可扩展性与高可用性。

6 多模态交互技术要求

6.1 输入模态

6.1.1 文本输入

输入模态的文本输入技术，应满足如下要求：

a) 支持UTF-8、GB18030等主流编码格式，确保常用的中文（简体字、繁体字）、英文（美式、英式）等多语言字符正确解析。

b) 单次输入文本长度支持10—5000字符范围，超出时可自动分段处理。

c) 具备文本纠错功能，可有效识别和提示拼写及语法错误。

d) 支持富文本解析，包括段落、列表、超链接等结构化内容。

6.1.2 语音输入

输入模态的语音输入技术，应满足如下要求：

a) 在不同环境条件下保持较高的语音识别准确率，在极端嘈杂环境下可支持模态切换（如结合文本输入辅助）。

b) 支持中文普通话识别，可扩展支持粤语等主要方言和英语等其他语种。

c) 音频采集可支持PCM/WAV等常见格式和MP3/OPUS等压缩编码。

d) 采用WebRTC等实时传输协议，确保语音交互的流畅性。

6.1.3 图像/视频输入

输入模态的图像/视频输入技术，应满足如下要求：

a) 静态图像分辨率宜≥1280×720dpi，支持JPEG/PNG/BMP等常见格式。

b) 动态视频帧率宜≥25fps，支持H.264/MP4/AVI等常见格式。

c) 具备较高的人脸与物体识别准确率，能够满足常见应用场景的识别需求。

d) 具备图像增强功能，包括去模糊、超分辨率重建等。

6.1.4 手势输入

输入模态的手势输入技术，应满足如下要求：

a) 支持多种常见静态与动态手势的准确识别，满足自然交互的基本需求。

b) 具备良好的空间识别范围，识别角度宜≥±45°，识别距离宜在0.5—5米之间。

c) 数据传输可采用蓝牙、Wi-Fi5/6等无线协议，确保实时性和稳定性。

6.1.5 动作捕捉设备输入

输入模态的动作捕捉设备输入技术，应满足如下要求：

a) 支持光学、惯性、视觉等主流动作捕捉系统，可提供高精度骨骼数据流。

b) 具备降噪与滤波等功能，可有效抑制抖动和异常值，保证动作平滑自然。

c) 提供校准与重定位功能，支持快速恢复跟踪并适应不同表演者体型。

6.1.6 情绪输入

输入模态的情绪输入技术，应满足如下要求：

a) 提供适宜的数据采集环境，控制环境噪音干扰，确保情绪数据采集质量。

b) 情绪数据采样频率宜≥20Hz，确保情绪变化的细腻捕捉和实时分析。

c) 情绪数据格式可包括实时采集生理数据、时间戳、设备编号等。

d) 情绪数据传输可采用蓝牙、射频、Wi-Fi5/6等无线协议，确保实时性和稳定性。

6.2 输出模态

6.2.1 语音合成

输出模态的语音合成技术，应满足如下要求：

a) 支持HiFi-GAN、DurIAN等主流语音合成模型，语音MOS评分宜≥4分。

b) 普通话合成自然度达到专业播音员水平，英语合成具备高清晰度与可理解性。

c) 实现多情感语音合成，至少支持6种基础情感类型（喜/怒/哀/惧/惊/平）。

d) 语速调节范围宜为基准语速的0.8—1.5倍（基准语速180字/分钟）。

e) 支持PCM/WAV等常见音频格式和MP3/OPUS等压缩编码。

f) 具备实时音频流驱动与处理能力，可支持超低延迟的语音合成与输出，确保实时交互场景的流畅性。

6.2.2 面部表情

输出模态的面部表情技术，应满足如下要求：

a) 至少支持6种基础表情（喜/怒/哀/惧/惊/平），符合FACS编码标准。

b) 表情过渡自然流畅，无明显延迟或跳跃感，确保表情变化的连贯性。

c) 表情丰富度评分宜≥4分（详见“7.1.2 表2 表情丰富度主观评分规则”），与语音情感输出保持同步。

d) 在4K等高分辨率显示环境下，确保面部渲染的流畅性与细节表现力，避免出现卡顿或失真。

6.2.3 肢体动作

输出模态的肢体动作技术，应满足如下要求：

a) 具备高自由度的关节控制能力，符合生物力学标准，确保动作的灵活性与真实性。

b) 具备丰富的基础动作库，支持常见动作及其组合，满足多样化交互需求。

c) 动作自然度评分宜≥4 分（详见“7.1.3 表 3 动作自然度主观评分规则”）。

d) 支持物理模拟渲染，包括头发、衣物等动态效果。

6.2.4 情感表达

输出模态的情感表达技术，应满足如下要求：

a) 至少支持6种基础情感类型（喜/怒/哀/惧/惊/平）表达，每种情感可实现3级强度（轻微/中等/强烈）的可控调节。

b) 支持多种基础情感的混合表达功能，确保跨模态情感表达协调一致。

c) 支持专业播报、亲切服务等5种以上个性化表达风格模板。

d) 情感表达持续时间宜设定在300—5000ms范围，支持自适应调整。

e) 特殊场景可支持人工干预模式，确保情感表达的准确性。

6.3 跨模态协同

6.3.1 同步机制

跨模态协同的同步机制技术，应满足如下要求：

a) 语音与口型保持高度同步，支持实时动态修正，确保视听一致性。

b) 表情变化与语音输出协调一致，保障情感表达的完整性与自然度。

c) 肢体动作与语音节奏相匹配，维持交互过程的流畅性与和谐性。

d) 动作表现与场景变化实时适配，避免出现延迟或脱离情境的差异。

e) 多流同步可采用PTP等协议，减少时序偏差。

f) 具备自动校准能力，快速恢复异常同步状态。

6.3.2 互补机制

跨模态协同的互补机制技术，应满足如下要求：

a) 当单一模态输入置信度不足时，支持自动启用其他模态进行辅助增强，提升交互的完整性与鲁棒性。

b) 在关键指令或高风险场景中，可采用多模态冗余确认机制，通过多种输入方式交叉验证，确保交互的准确性与可靠性。

c) 互补过程中保持上下文连贯性，避免因模态切换导致交互逻辑中断或用户体验不一致。

d) 模态互补与切换过程保持流畅自然，无明显延迟或响应断层，保障交互过程的实时性与连续性。

6.3.3 冲突解决机制

跨模态协同的冲突解决机制技术，应满足如下要求：

a) 支持建立模态优先级体系（如语音＞视觉＞触觉），指导冲突决策。

b) 冲突决策过程迅速高效，避免对用户体验造成明显影响。

c) 对于非紧急冲突，支持启动二次确认流程确保准确性。

d) 具备记录冲突处置日志功能，支持事后分析优化。

e) 支持基于用户反馈持续优化冲突解决策略，提升系统自适应能力。

6.3.4 数据延时补偿机制

跨模态协同的数据延时补偿机制技术，应满足如下要求：

a) 支持建立模态优先级体系（如语音＞视觉＞触觉），优先保障高优先级模态延时补偿。

b) 延时补偿响应快速有效，避免影响整体交互流畅度和用户体验。

c) 对于非紧急延时场景（如触觉模态短时延时），支持启动缓存调度补偿，确保交互连续性。

d) 具备记录延时补偿日志功能，包括延时模态、补偿策略、处理结果等，支持事后分析优化。

e) 具备基于用户反馈的自适应优化能力，持续提升延时补偿策略的适配性和有效性。

7 性能评估指标

7.1 形象质量指标

7.1.1 外观拟真度

形象质量指标中的外观拟真度为主观性评估指标，用户根据看到的数智人形象质量，在李克特量表中给出一个主观评分评价质量优劣，具体评分规则见表1所示。

表1 外观拟真度主观评分规则

评价维度	5分 （非常拟真）	4分 （比较拟真）	3分 （一般）	2分 （较不拟真）	1分 （完全不拟真）
面部真实感	与真人几乎无差异，皮肤纹理、毛孔等细节完美呈现	接近真人，细节略有简化但整体自然	基本真实但能看出数字痕迹	明显数字化特征，细节缺失	完全不像真人，严重失真
皮肤质感	皮肤光影反射完全自然，可见细腻的纹理和皮下散射效果	皮肤质感良好，主要光影效果准确	基本皮肤质感具备，但缺乏细节	塑料感明显，缺乏真实皮肤特性	完全不符合真实皮肤特性
身体比例	完全符合人体解剖学标准，各部位比例协调完美	基本符合解剖学，比例协调良好	大体比例正确但局部略有失调	明显比例失调（如头身比不当）	严重违反人体比例标准
服饰细节	衣物纹理清晰，物理特性（褶皱、垂感等）完全拟真	衣物细节丰富，主要物理特性准确	基本衣物表现，但细节简化	衣物表现简单，缺乏真实感	衣物完全不符合物理规律
动态自然度	动作表情完全自然流畅，无任何机械感	动作表情自然，偶有微小不连贯	基本自然但能察觉数字痕迹	明显机械感，动作生硬	动作完全违反自然规律

7.1.2 表情丰富度

形象质量指标中的表情丰富度为主观性评估指标，用户根据看到的数智人形象质量，在李克特量表中给出一个主观评分评价质量优劣，具体评分规则见表2所示。

表2 表情丰富度主观评分规则

评价维度	5分（非常丰富）	4分（比较丰富）	3分（一般）	2分（较单一）	1分（非常单一）
基础表情种类	能自然呈现6种以上基础表情（喜/怒/哀/惧/惊/平）	能呈现5种基础表情	能呈现4种基础表情	仅能呈现3种基础表情	只能呈现2种及以下表情
复合表情表现	能流畅展现多种复合表情（如悲喜交加）	能表现简单复合表情	基础表情组合略显生硬	复合表情转换不自然	无法表现复合表情
表情细腻度	微表情丰富自然，细节完美呈现	微表情控制良好，细节较丰富	有基本微表情表现	微表情简单粗糙	无有效微表情
表情动态过渡	表情切换极其流畅，无任何痕迹	过渡自然，偶有轻微不连贯	基本流畅但有可察觉过渡	明显卡顿感	表情切换生硬断裂
情境适配度	表情与情境完美匹配，感染力强	表情与情境匹配良好	基本匹配但偶有不协调	时有错位不匹配	完全不符合情境

7.1.3 动作自然度

形象质量指标中的动作自然度为主观性评估指标，用户根据看到的数智人形象质量，在李克特量表中给出一个主观评分评价质量优劣，具体评分规则见表3所示。

表3 动作自然度主观评分规则

评价维度	5分（非常自然）	4分（比较自然）	3分（一般自然）	2分（不太自然）	1分（非常不自然）
肢体协调性	动作完全符合人体工学，各部位配合完美	动作协调性好，偶尔有微小不协调	基本协调，但能看出机械感	明显不协调，动作生硬	严重不协调，违反人体运动规律
动作流畅度	动作过渡极其平滑，无任何卡顿	过渡自然，偶有轻微不连贯	基本流畅，但能察觉过渡痕迹	明显卡顿，动作不连贯	严重卡顿，动作断裂感强
力度表现	力度控制精准，完全符合物理规律	力度表现良好，接近真实	力度表现基本合理	力度控制不稳定，时轻时重	力度表现完全失真
节奏感	动作节奏完全符合情境需求，张弛有度	节奏控制良好，基本符合情境	节奏基本合理，偶有快慢不当	节奏控制较差，经常不合时宜	节奏完全混乱，与情境严重不符
情感表达	动作完美传达情感，感染力强	动作较好传达情感，有一定感染力	动作基本能表达情感	动作情感表达模糊	动作与情感表达完全脱节

7.2 交互能力指标

7.2.1 响应延迟

交互能力指标中的响应延迟技术应满足表4要求。

表4 响应延迟技术要求

技术指标	具体要求
文本交互延迟	a) 短文本输入处理延迟宜≤300ms，实现快速响应。 b) 长文本输入可支持分段处理，首段响应延迟宜≤300ms。
语音交互延迟	a) 安静环境下语音输入响应延迟宜≤800ms，确保对话流畅自然。 b) 嘈杂环境下语音输入响应延迟宜≤1200ms，支持自动启用降噪优化。 c) 语音合成输出延迟宜≤300ms，保证语音输出的即时性。
视觉交互延迟	a) 静态图像识别延迟宜≤300ms，确保快速图像解析。 b) 动态视频处理延迟宜≤200ms/帧，保证视频交互流畅度。
手势交互延迟	a) 静态手势识别延迟宜≤300ms，提供即时反馈。 b) 动态手势跟踪延迟宜≤500ms，确保动作连贯性。

7.2.2 多轮对话

交互能力指标中的多轮对话技术应满足表5要求。

表5 多轮对话技术要求

技术指标	具体要求
上下文保持能力	a) 支持≥10轮次对话的上下文记忆。 b) 具备良好的话题聚焦能力，有效维持对话主题的一致性，避免偏离核心主题。 c) 确保关键信息的准确提取与保持，避免在多轮对话中出现重要内容丢失或误解。
指代消解能力	a) 支持代词精准解析，威诺格拉德测试准确率宜≥85%。 b) 支持跨轮次实体关联，正确理解“这个”“那个”等指代。 c) 具备模糊指代澄清功能，可自动补充缺失信息。
意图延续能力	a) 支持复杂任务自动分解，可正确拆解多步骤需求。 b) 支持子任务进度跟踪，确保分步执行不遗漏。 c) 具备对话中断恢复功能，支持随时继续未完成对话。
知识衔接能力	a) 支持跨领域知识关联，实现多话题自然过渡。 b) 支持历史对话知识复用，避免重复询问。 c) 具备时效知识快速更新功能，延迟宜≤24小时，保证信息的时效性。

7.2.3 多模态理解

交互能力指标中的多模态理解技术应满足表6要求。

表6 多模态理解技术要求

技术指标	具体要求
跨模态融合能力	a) 实现语音-视觉模态的深度融合，确保视听信息协同理解的准确性与一致性。
	b) 支持文本-手势模态的精准对齐，达成自然流畅的人机交互体验。 c) 具备多模态联合理解能力，可有效处理语音、视觉、文本等复杂交互场景。 d) 跨模态特征对齐准确，保障多源信息融合的完整性和有效性。

语义理解深度	a) 能够准确识别多模态联合意图，确保对用户真实需求的精准把握。 b) 至少支持6种基础情感类型（喜/怒/哀/惧/惊/平）的精准分析，实现情感化交互体验。 c) 具备解析隐喻、反讽等复杂语言现象的能力。 d) 支持理解和执行包含多个条件的复合指令。
系统鲁棒性	a) 在部分模态失效时，可支持通过剩余模态维持基本交互功能。 b) 在噪声环境下可保持稳定的识别和理解性能。 c) 具备不同光照条件的适应能力，确保视觉信息的可靠获取。 d) 支持跨语言场景下的混合输入理解。
特殊场景处理	a) 支持建立紧急指令的优先处理机制，确保关键操作可靠性。 b) 敏感内容识别准确率≥99%，符合内容安全要求。 c) 支持快速匹配用户画像，提供个性化的交互体验。 d) 可准确理解各领域的专业术语和特定表达。

7.2.4 异常处理

交互能力指标中的异常处理技术应满足表7要求。

表7 异常处理技术要求

技术指标	具体要求
输入异常	a) 支持语音中断自动补偿功能，确保语音输入的连续性和完整性。 b) 实现视觉遮挡场景下的多模态切换，切换时间宜≤800ms。 c) 具备传感器失效检测能力，可及时发现硬件异常并自动告警。
交互异常	a) 支持多模态指令冲突仲裁，确保交互决策的及时性与准确性。 b) 具备语义歧义自动识别与澄清功能，提升交互理解的精确度。 c) 支持交互超时自动恢复，重建对话上下文环境，恢复时间宜≤800ms。
多轮对话异常	a) 支持话题丢失快速恢复，恢复时间宜≤3轮对话。 b) 可实现歧义自动澄清，通过精准提问减少用户重复解释，平均询问次数宜≤2次/会话。 c) 具备冲突检测功能，可及时发现并解决理解分歧。
响应延迟异常	a) 建立紧急指令优先处理机制，确保关键操作的及时响应与执行。 b) 具备网络异常快速恢复能力，通过本地缓存机制维持交互连续性。 c) 可支持动态降级机制，在资源不足时优先保障核心功能。
系统异常	a) 可实现网络异常自动容错，通过离线模式维持基本服务。 b) 支持系统负载动态调节，可根据资源状况自动调整服务质量。 c) 具备安全威胁实时拦截能力，防范恶意攻击行为。

7.3 智能水平指标

7.3.1 知识覆盖度

智能水平指标中的知识覆盖度技术应满足表8要求。

表8 知识覆盖度技术要求

技术指标	具体要求
领域知识覆盖	a) 建立完整的领域知识体系，可覆盖核心概念和关键知识点，确保专业领域问题能够得到准确解答。 b) 实现知识图谱的自动化扩展功能，可支持新知识的自动识别、分类和整合，

	保持知识库的动态更新。 c) 开发跨领域知识关联算法，可实现不同领域概念间的智能关联和推理，提升综合问题解决能力。
常识知识理解	a) 具备常识推理与判断能力，威诺格拉德测试准确率宜≥90%。 b) 支持常见生活场景的理解和应用，可覆盖多种典型场景。 c) 具备基础物理规律认知能力，确保常识推理的合理性。
时效知识更新	a) 建立热点事件监测和快速更新机制，确保重要时事信息能够及时纳入知识体系。 b) 支持政策法规实时同步，可实现相关政策法规的自动识别、解析和更新。 c) 具备知识时效性验证功能，可通过多重验证机制识别和过滤过期、失效信息。
知识检索性能	a) 保证知识检索的响应速度，单知识点检索时间宜≤300ms。 b) 支持高并发的知识查询，处理能力宜≥1000次/秒。 c) 实现多模态知识检索，支持通过文本、语音、图像等多种方式获取相关知识。
知识验证机制	a) 建立多层次的知识可信度评估体系，可对知识来源、准确性和时效性进行综合评价。 b) 实现矛盾知识自动检测，能够识别知识库中的不一致信息并给出处理建议。 c) 具备知识溯源功能，可记录每个知识点的来源、更新历史和使用情况。

7.3.2 问题解决

智能水平指标中的问题解决能力技术应满足表9要求。

表9 问题解决能力技术要求

技术指标	具体要求
分析决策能力	a) 支持复杂问题多维度拆解，可通过语义理解和知识图谱技术准确识别问题中的关键要素及其相互关系。 b) 实现基于因果推理的问题归因分析功能，可采用贝叶斯网络等方法准确定位问题根源，识别潜在影响因素。 c) 具备多方案智能生成能力，可结合案例推理和约束满足技术，为每个问题提供多个可行性方案并评估优劣。
逻辑推理能力	a) 通过标准三段论测试验证系统的基础逻辑能力，确保能够正确处理各类演绎推理问题。 b) 支持归纳和演绎相结合的混合推理模式，能够处理包含不确定性的复杂推理场景。 c) 具备基于反事实推理的假设分析能力，可支持“如果...那么...”式的条件推理和情景模拟。
执行验证能力	a) 实现基于任务分解的多步骤规划功能，可通过层次化任务网络等方法确保执行计划的完整性。 b) 支持执行过程的动态监测和实时优化，方案优化响应时间宜≤500ms。 c) 建立多层级的结果验证机制，可通过预设条件和实际输出的比对及时发现执行偏差。
创新优化能力	a) 支持基于发散思维的非标准方案生成，可通过类比推理和跨领域知识迁移产生创新性解决方案。 b) 实现解决方案的持续进化机制，可采用遗传算法等优化方法不断提升执行效率。 c) 具备跨领域方案智能迁移能力，可通过深度语义匹配实现有效的知识复用。

7.3.3 个性化适配

智能水平指标中的个性化适配技术应满足表10要求。

表10 个性化适配技术要求

技术指标	具体要求
用户画像建模	a) 支持多维度用户特征提取，构建涵盖用户基本属性、行为偏好、兴趣特征等多维度的精准画像模型。 b) 建立实时动态更新机制，可基于用户最新的交互行为和反馈数据持续优化画像模型，确保画像的时效性和准确性。 c) 具备跨平台数据整合功能，可实现不同场景和渠道下用户行为数据的有效关联与统一分析。
交互风格适配	a) 提供至少5种预设交互模式，支持根据用户画像智能匹配最适合的交互方式，包括语音风格、响应速度等维度。 b) 支持语音语调个性化调整，表情丰富度评分宜≥4分（详见“7.1.2 表2 表情丰富度主观评分规则”）。 c) 实现响应速度动态调节，动作自然度评分宜≥4分（详见“7.1.3 表3 动作自然度主观评分规则”）。
内容推荐系统	a) 建立精准的推荐算法引擎，可基于用户画像和行为数据分析，提供个性化的内容推荐服务。 b) 建立实时的兴趣捕捉机制，可通过持续监测用户交互行为，动态更新兴趣模型和推荐策略。 c) 完善用户反馈系统，支持用户对推荐内容进行评价和调整，实现推荐质量的持续优化。
界面呈现适配	a) 实现布局自动优化，可根据不同终端设备的特性和用户偏好，智能调整界面元素和排版方式。 b) 提供丰富的个性化定制选项，用户可根据个人喜好调整界面颜色、字体等视觉元素。 c) 加强无障碍访问功能设计，充分考虑特殊用户群体的使用需求，提供多样化的辅助功能。

7.4 用户体验指标

7.4.1 易用性

用户体验指标中的易用性为主观性评估指标，用户根据数智人的体验效果，在李克特量表中给出一个主观评分评价质量优劣，具体评分规则见表11所示。

表11 易用性主观评分规则

评价维度	5分 （非常满意）	4分 （比较满意）	3分 （一般）	2分 （不太满意）	1分 （非常不满）
操作直观性	所有功能一目了然，无需指导即可使用	主要功能容易发现，少量提示后能掌握	基本功能可用，但需要多次尝试	经常找不到所需功能	界面混乱，无法正常操作
学习成本	首次使用30秒内即可掌握核心功能	1分钟内能理解基本操作	需要3-5分钟学习才能使用	需要专门培训才能操作	即使培训也难以使用

交互效率	完成任务比预期快50%以上	完成任务时间符合预期	完成任务比预期慢20%	操作明显迟缓	操作过程令人烦躁
错误恢复	系统自动预防错误，几乎不会误操作	错误提示清晰，1步即可恢复	需要2-3步才能纠正错误	错误恢复流程复杂	经常无法恢复操作错误
帮助系统	智能帮助精准预测需求，无需主动查询	帮助信息完整易懂，1次查询即解决	需要多次查询才能找到答案	帮助信息不完整	帮助系统完全无效

7.4.2 满意度

用户体验指标中的满意度为主观性评估指标，用户根据数智人的体验效果，在李克特量表中给出一个主观评分评价质量优劣，具体评分规则见表12所示。

表12 满意度主观评分规则

评价维度	5分（非常满意）	4分（满意）	3分（一般）	2分（不满意）	1分（非常不满）
整体满意度	远超预期，强烈推荐使用	达到预期，愿意推荐	基本满足需求	未达预期，有待改进	完全不符合需求
交互自然度	如同真人交流般自然流畅	交互顺畅偶有机械感	基本可完成交互	明显机械生硬	完全无法自然交互
功能实用性	功能设计完美满足所有需求	主要需求都能满足	基本功能可用	关键功能缺失	完全无法满足需求
情感体验	带来愉悦的情感共鸣	交互过程令人舒适	情感体验中性	偶有负面情绪	产生强烈负面情绪
推荐意愿	会主动向多人推荐	被询问时会推荐	可接受使用	不推荐使用	会劝阻他人使用

7.4.3 情感共鸣度

用户体验指标中的情感共鸣度为主观性评估指标，用户根据数智人的体验效果，在李克特量表中给出一个主观评分评价质量优劣，具体评分规则见表13所示。

表13 情感共鸣度主观评分规则

评价维度	5分（强烈共鸣）	4分（明显共鸣）	3分（中等共鸣）	2分（微弱共鸣）	1分（毫无共鸣）
情绪感染力	能深刻引发我的情感波动	能有效调动我的情绪	能感知到情绪传递	情绪表达较平淡	完全感受不到情感
表情生动性	表情变化极其丰富自然	表情变化较为生动	有基本表情变化	表情略显僵硬	表情完全机械化
语音情感度	语音充满情感张力	语音情感表达良好	语音情感表达适中	语音情感较单调	语音完全无情感
肢体表现力	肢体语言极具表现力	肢体语言较为丰富	有基本肢体动作	肢体动作较生硬	无有效肢体语言
共情理解力	完全理解我的情绪并恰当回应	能较好理解我的情绪	能基本感知我的情绪	偶尔误解我的情绪	完全无法理解我的情绪

8 安全与合规要求

8.1 数据隐私保护

数据隐私保护应满足如下要求：

a) 遵循《中华人民共和国个人信息保护法》《中华人民共和国数据安全法》等法律法规，同时符合ISO/IEC 27701:2025标准中对个人身份信息（PII）全生命周期保护的要求，确保用户数据采集、存储、使用的合法性与安全性。

b) 采用加密传输（如TLS 1.2+）和存储（如AES-256）技术，对PII进行全程加密保护，防止数据泄露。

c) 提供数据访问权限管理，确保仅授权人员可访问敏感数据。

d) 支持数据生命周期管理，包括定期清理和合规销毁。生物特征数据保留周期宜≤30 天，到期后采用 NIST800 合规擦除标准销毁；非敏感数据保留宜≤1 年，销毁日志需留存宜≥3 年，支持监管审计。

8.2 内容安全审核

内容安全审核应满足如下要求：

a) 建立完善的三级审核机制，确保生成内容符合法律法规及公序良俗。

b) 支持关键词过滤、敏感内容识别及实时拦截功能。

c) 提供内容审核日志，支持追溯与责任认定。

8.3 知识产权合规

知识产权合规应满足如下要求：

a) 确保训练数据来源合法，并提供版权授权证明。

b) 明确AI生成内容的权属规则，支持在用户协议中约定权利归属。

c) 支持数智人形象的版权登记与侵权监测。

8.4 系统安全防护

系统安全防护应满足如下要求：

a) 基础服务场景支持等保2.0二级；涉及金融、医疗、教育等敏感领域，满足等保2.0三级及以上，核心模块需通过渗透测试（高危漏洞为0）。

b) 提供防DDoS攻击、防数据篡改等安全防护措施。

c) 支持联邦学习或差分隐私技术，保障数据脱敏与隐私安全。

8.5 合规认证与审计

合规认证与审计应满足如下要求：

a) 支持第三方安全合规认证（如ISO27001）。

b) 提供完整的操作日志与审计功能，满足监管审查要求。

c) 定期开展数据安全风险评估，确保符合最新法律法规。

8.6 用户知情与授权

用户知情与授权应满足如下要求：

a) 建立用户授权记录机制，完整记录用户同意的时间、内容、形式等信息，形成可追溯的文件化证据，留存期限应符合相关法律法规及ISO/IEC 27701:2025对文件化信息的控制要求。

b) 提供清晰的数据采集与使用告知，内容需明确包含PII处理的目的、方式、范围、存储期限、安全保护措施、用户权利及行使方式等信息，确保用户知情权。

c) 支持用户数据删除与撤回同意功能，用户撤回同意后，应停止对相关PII的进一步处理（法律法规另有规定的除外），并按要求删除或匿名化处理相关数据。

d) 在涉及生物特征数据（如语音、面部识别）时，单独获取用户明示同意。

8.7 不良应用应急处置与责任追溯

为防范多模态交互式数智人技术被用于欺诈、虚假新闻、虚假宣传等不良场景，需建立全链路风险管控机制，具体要求如下：

a) 应急处置机制

1) 建立7×24小时实时监测系统，重点识别欺诈诱导、虚假信息等风险内容。

2) 实施分级处置：轻度违规内容屏蔽并警示，中度违规暂停服务并人工复核，重度违规立即终止服务。

3) 处置响应时间宜≤30分钟，并建立跨平台协同处置通道实现多渠道同步管控。

b) 日志溯源机制

1) 留存数智人全生命周期操作日志，包括开发、运行、变更等各环节。

2) 日志包含时间戳（毫秒级）、操作主体、数据流向等要素，且不可篡改。

3) 日志存储期限宜≥3年，支持多维度快速检索，单条查询响应时间宜≤3秒。

c) 法律追诉与后果处置机制

1) 明确技术提供方、应用方、用户的权责边界，约定法律责任归属。

2) 留存违规主体身份信息及完整证据链，支持向监管和司法部门提供证据。

3) 建立用户损失评估与补偿协助机制，定期优化风险管控策略（每季度至少1次）。

9 应用实施要求

9.1 系统集成与部署最佳实践

9.1.1 硬件配置

硬件配置应满足如下要求：

a) 服务器：支持配置多核CPU（如Intel Xeon Gold）、GPU（如NVIDIA A100）及高速SSD存储，存储容量宜≥1TB。

b) 输入设备：支持4K摄像头、阵列麦克风及动作捕捉设备，摄像头帧率宜≥30fps，麦克风拾音距离宜≥5米。

c) 网络：支持千兆带宽，网络延迟宜≤50ms。

d) 边缘节点：支持分布式部署，覆盖全国主要区域，节点间数据同步延迟宜≤200ms。

e) 容灾备份：支持异地双活架构，RPO=0，RTO≤5分钟，备份数据保存时间宜≥30天。

9.1.2 软件集成

软件集成应满足如下要求：

a) 提供标准化API与SDK，支持Java/Python/C++等语言调用。

b) 支持容器化部署，实现资源弹性扩展。

c) 兼容主流数据库（如MySQL、MongoDB）及中间件（如Redis、RabbitMQ）。

d) 提供CI/CD流水线，支持自动化测试与部署。

e) 集成DevOps监控体系，实现全链路可观测性。

f) 支持与第三方认证系统集成（如OAuth2.0、SAML），保障访问安全。

g) 满足跨平台兼容性要求，兼容Windows、macOS、Linux等桌面操作系统，以及Android 10.0+、iOS 14.0+等移动操作系统，确保不同平台核心功能的一致性，避免功能缺失或异常现象。

h) 具备移动端优化功能，支持自适应屏幕分辨率，移动端应用启动时间宜≤3秒，页面加载时间宜≤2秒，确保移动端交互流畅性。

9.2 运维管理与持续优化

9.2.1 运维管理

运维管理应满足如下要求：

a) 监控告警：支持实时监测系统负载、API成功率等指标，出现异常自动告警，告警延迟时间宜≤1分钟。

b) 日志管理：保留全链路操作日志宜180天以上，支持快速溯源。

c) 容灾备份：支持每日增量备份，RTO宜≤30分钟，RPO宜≤5分钟。

d) 变更管理：支持灰度发布，确保影响范围可控。

e) 权限管理：采用最小权限原则，支持多角色权限配置与审计，权限变更记录保存时间宜≥1年。

9.2.2 持续优化

持续优化应满足如下要求：

a) 支持定期（如每月）更新AI模型，优化识别准确率与交互体验。

b) 支持定期（如每季度）扩展虚拟人形象库与场景模板。

c) 提供用户反馈通道，驱动功能优化。

d) 建立技术雷达，持续评估新兴技术。

e) 定期进行安全漏洞扫描与渗透测试，每季度至少1次，高危漏洞修复时间宜≤72小时。

附录 A

（资料性）

典型应用场景技术适配方案

A.1 新闻播报场景

新闻播报场景中多模态交互式数智人的应用，应满足如下要求：

a) 支持高精度语音合成与口型同步技术，确保播报自然流畅，同步误差满足“6.3.1 同步机制”要求。

b) 提供多语种、多方言适配能力，至少涵盖20种以上主流语言及10种以上方言，满足不同地区需求。

c) 支持实时新闻内容接入与动态播报风格调整，可根据新闻类型自动切换庄重、活泼等风格。

d) 支持突发新闻插播功能，插播时不影响原有节目连贯性。

e) 支持多平台同步分发，包括电视、移动端和社交媒体等。

f) 具备新闻内容审核能力，对敏感信息识别准确率≥99%。

g) 支持虚拟主播形象自定义，可根据新闻主题更换服饰、背景等元素。

h) 支持坐姿、站姿等多种播报姿势，并具备丰富的专业肢体动作库（如手势引导、点头示意、身体转向），确保播报姿态与新闻内容的严肃性或互动性相匹配，且动作切换自然流畅。

A.2 电商直播场景

电商直播场景中多模态交互式数智人的应用，应满足如下要求：

a) 支持实时弹幕互动与商品3D展示功能，提升用户参与感和体验流畅性。

b) 支持多模态交互（如语音、手势、表情），识别准确率满足“6.1 输入模态”要求，提升用户参与感。

c) 支持虚拟试穿/试用功能，有效促进购买转化。

d) 提供销售数据分析看板，实时监控GMV、客单价、转化率等关键指标，数据更新时间宜≤1分钟。

e) 具备商品信息实时更新能力，库存、价格变动响应时间宜≤10秒。

f) 提供智能推荐算法，可根据用户行为动态调整直播内容。

A.3 教育辅导场景

教育辅导场景中多模态交互式数智人的应用，应满足如下要求：

a) 支持文字、语音等多模态交互，确保信息识别的准确性。

b) 支持虚拟教师形象自定义，高精度个性化模拟教师形象与声音。

c) 支持个性化教学内容生成与动态难度调整。

d) 提供学生状态监测（如注意力分析）与互动引导功能。

e) 集成知识图谱，实现智能答疑与个性化学习路径规划功能。

f) 提供学情分析报告，包括知识点掌握度、学习时长、进步趋势等多维度分析内容。

g) 支持多学科内容覆盖，包括语文、数学、英语等主流学科及拓展课程。

A.4 政务服务场景

政务服务场景中多模态交互式数智人的应用，应满足如下要求：

a) 支持多轮对话管理与政策条文精准解读，对话轮次宜≥10 轮，政策内容解读准确率满足“7.2.2 多轮对话”要求。

b) 支持与政务服务平台（如一网通办、行政审批系统）无缝对接，实现办事流程闭环。

c) 具备7×24小时不间断服务能力，确保系统的高可用性和快速响应。

d) 支持复杂事项自动转接人工坐席机制，实现服务流程的无缝衔接与信息同步。

e) 提供政务知识库实时更新功能，政策法规变动后宜在24小时内完成知识库更新。

f) 支持多渠道政务服务接入，包括政务APP、小程序、网站、自助终端等，确保服务体验一致。

g) 具备多语种服务能力，支持主要外语及方言，满足多样化服务需求。

A.5 医疗问诊场景

医疗问诊场景中多模态交互式数智人的应用，应满足如下要求：

a) 支持患者症状多模态输入（如文字描述、症状图片、语音阐述），确保信息采集的全面性。

b) 集成权威医学知识库，可提供初步病情分析与建议。

c) 支持预约挂号、检查报告解读功能，可有效提升医疗服务效率。

d) 建立完善的隐私保护机制，对患者信息进行加密存储和严格权限管理。

e) 支持紧急情况快速对接急诊通道，确保及时响应。

A.6 文旅导览场景

文旅导览场景中多模态交互式数智人的应用，应满足如下要求：

a) 支持景区全景虚拟游览与路线规划，全景画面可支持4K等高分辨率。

b) 提供多语种景点讲解功能，至少涵盖5种以上语言，讲解内容准确生动。

c) 支持实时天气、人流密度查询与预警功能，数据更新频率宜≤30分钟。

d) 具备文化知识拓展功能，可介绍景点相关历史典故、民俗风情等。

A.7 影视剧评测场景

影视剧评测场景中多模态交互式数智人的应用，应满足如下要求：

a) 建立演员或现场参加评测人员生命体征监测体系，支持实时采集与展示心率、皮电、肌电等生理传感器数据，监测压力与情绪反应。

b) 提供多模态数据融合监控看板，支持同时展示拍摄画面、生理数据、环境参数等多种信息流。

c) 支持表演质量评估功能，基于生理数据与画面内容对演员或现场参加评测人员的表演效果进行实时分析与量化评分。

d) 支持评测数据的统计、分析与播报，提供标准化接口供后期制作数智人等环节使用。

A.8 金融服务场景

金融服务场景中多模态交互式数智人的应用，应满足如下要求：

a) 支持金融业务全流程身份认证，集成人脸识别、声纹验证、数字证书等多因素认证方式，确保业务办理的安全性与合规性。

b) 具备金融数据实时处理与分析能力，支持理财产品、信贷业务、保险产品等多类型金融服务的精准推荐与风险提示。

c) 提供财富管理可视化服务，支持资产配置分析、收益模拟、风险评估等多维度数据的直观呈现。

d) 符合金融行业监管要求，实现业务办理全过程可追溯，数据留存期限宜≥5年，满足审计与监管要求。

e) 嵌入智能风控引擎，覆盖客户识别、交易监控、异常预警等全业务流程，确保风险识别的准确性。

f) 支持银行、保险、证券、信托等多元金融业务的智能化服务，确保跨业务领域的服务一致性。

附录 B

（资料性）

名词解释

B.1 李克特量表（Likert scale）

是一种通过让受访者在预先设定的、具有等级顺序的选项中进行选择，从而将其主观态度、意见或感受进行量化的测量工具。它是行为科学和市场研究中最基础、最重要的数据收集工具之一。

B.2 威诺格拉德测试（Winograd Schema Challenge）

是一种专门用于评估机器常识推理能力的自然语言处理任务，其通过设计需要依赖上下文常识才能正确解析的指代歧义问题，来检验模型是否具备人类般的深层语义理解能力。

B.3 平均意见得分（Mean Opinion Score，MOS）

是语音通信领域评估人类交流质量的主观测量方法。国际电信联盟（ITU-T）在P.800标准中将MOS评测规范化为绝对等级评分（ACR）体系，采用五分制评估语音质量。

参考文献

[1] ISO/IEC 27701:2025 EN Information security, cybersecurity and privacy protection - Privacy information management systems - Requirements and guidance

[2] GY/T 411-2024 数字虚拟人技术要求

[3] 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

[4] YD/T 4393.2-2023 虚拟数字人指标要求和评估方法 第 2 部分：2D 真人形象类合成技术

[5] T/BIA 17-2024 数字人指标要求及评估方法 第 1 部分：平台基础能力

[6] T/AIIA 001-2021 支持语音和视觉交互的虚拟数字人技术规范