ICS 35.240
CCS L67

团 体 标 准

T/BIA 30-2025

基于大模型的数字人系统技术要求

Technical requirements of large model enhanced digital human system

2025 年 9 月 1 日 发布　　2025 年 9 月 1 日 实施

北京信息化协会

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》的规则起草。

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。

本文件由北京信息化协会提出并归口。

本文件起草单位：中国信息通信研究院，北京信息化协会元宇宙创新发展工作委员会，北京聆心智能科技有限公司，世优（北京）科技股份有限公司，憨猴科技集团有限公司，北京渲光科技有限公司，中图云创智能科技(北京)有限公司，北京广益集思智能科技有限公司，凌云光技术股份有限公司，陕西元景数创科技有限责任公司，北京神州数码云计算有限公司，云知声智能科技股份有限公司，北京灵动天地文化发展有限公司，国术科技(北京)有限公司，马上消费金融股份有限公司，北京智谱华章科技股份有限公司，北京神州数码云计算有限公司，明芒（北京）科技有限公司，北京艾力泰尔信息技术股份有限公司，北京中科金财科技股份有限公司，宁波菊风系统软件有限公司，南中轴（北京）国际文化科技发展有限公司，北京华盛智数科技有限公司，中移（苏州）软件技术有限公司，北京飞天云动科技有限公司，金辰宇（天津）科技有限公司，成都明途科技有限公司，爱化身科技（北京）有限公司。

本文件主要起草人：王思雨，颜媚，石霖，和婕，凌玲，纪菁，龚任娇子，盛琳子，郑叔亮，彭立彪，万大振，黄永康，纪智辉，王新国，李睿，邓先才，张倩，庞建青，何鹏，方顺，崔铭，邵丹，孙明，陈曦，李世尊，闫强，曾义，熊伟，金哲楠，万千，傅航，李刚，李盛，黄伟，梁家恩，刘征，刘晶，袁国术，彭小国，冯月，郝征鹏，杜冀中，袁永强，李刚，李盛，金辉，尹宪文，李玉奎，朱烨东，王姣杰，钱晓炯，蒋莹凯，王伟，李胜国，严锋，齐骥，刘婷，裴熬，那一牧，周世晟，王立波，朱建州，朱泽圣，严帅，郭林，陈伯昆。

基于大模型的数字人系统技术要求

1 范围

本文件规定了基于大模型的数字人系统在建模、驱动、交互等关键开发环节的技术要求和相应的评估方法。

本文件适用于基于大模型的数字人开发系统的研发、测试、评估和验收等。

2 规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中，注日期的引用文件，仅该日期对应的版本适用于本文件；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。

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

3 术语和定义

YD/T 4393.1-2023界定的以及下列术语和定义适用于本文件。

3.1

虚拟数字人 digital human

简称数字人或虚拟人，是指基于现实世界设计、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。

[来源：YD/T 4393.1-2023，3.1.1]

3.2

数字人系统 digital human system

以数字人技术为核心的软硬件集合,可实现音视频内容制作或视听双通道多模态人机交互的系统。

3.3

大模型 large model

一种具有大规模参数和复杂计算结构的机器学习模型，通过自监督或者无监督技术从海量的通用数据中训练得到基础模型，并结合下游具体任务对其进行微调，最终被训练成具有逻辑推理和分析能力的人工智能模型。

4 缩略语

下列缩略语适用于本文件。

[a] 2D [b] 二维 [c] 2-Dimensional

[d] 3D	[e] 三维	[f] 3-Dimensional
[g] NeRF	[h] 神经辐射场	[i] Neural Radiance Fields
[j] AI	[k] 人工智能	[l] Artificial Intelligence
[m] API	[n] 应用程序接口	[o] Application Programming Interface
[p] MOS	[q] 平均主观得分	[r] Mean Opinion Score

5 数字人系统应用大模型总体视图

数字人系统包括建模、渲染、驱动、交互等环节，其中部分环节如建模、驱动、交互等都可以调用大模型系统来完成任务。大模型数字人应用支撑系统包括大模型、应用领域知识库、应用领域工具库、记忆存储库、智能体、输入内容安全组件、输出内容安全组件等模块，如图1所示：

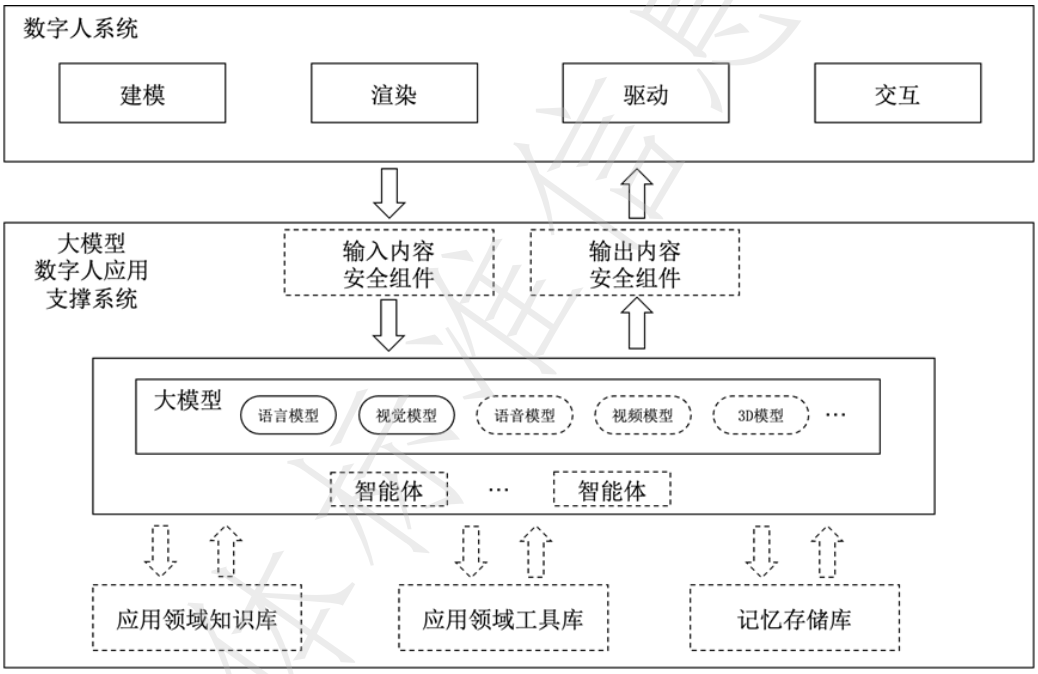

图 1 数字人系统应用大模型关系视图

•大模型：泛指参数量较大的人工智能模型，支撑系统的核心部分，根据应用任务不同，可能会包括语言大模型、视觉大模型、语音大模型、视频大模型、3D大模型等不同模态或者可支持多模态的大模型。主要负责对数字人输入的任务进行分析，并判断是否需要调用外部应用领域知识库或工具库获取专业知识或者执行专业操作等，并在综合处理后得到返回给数字人系统的内容。

•应用领域知识库：可以提供高时效、专业、可信和丰富的应用领域专业知识，来补足大模型在应用领域专业性上的不足。

•应用领域工具库：通过API接口对外提供应用专业工具服务能力的工具集合。

•记忆存储库：可以提供稳定、高效、安全的数据存储和管理服务。

•智能体：可与模型交互，自主的对复杂任务进行分解、规划、执行，并可通过不断学习和总结经验进行持续优化的一类工具。

•输入内容安全组件：对于数字人系统提出的服务请求内容进行分析，并判断服务请求是否存在安全合规风险，如存在安全风险，可以对请求进行拦截。

•输出内容安全组件：对于大模型生成的待返回给请求方的内容进行分析，并判断待输出内容是否存在安全合规风险，如存在安全风险，可以对输入内容进行安全改写，或者进行

拦截。

6 基于大模型的数字人系统技术框架

在数字人系统建模、渲染、驱动、交互等关键开发环节，大模型赋能实现总体视图如图2所示：

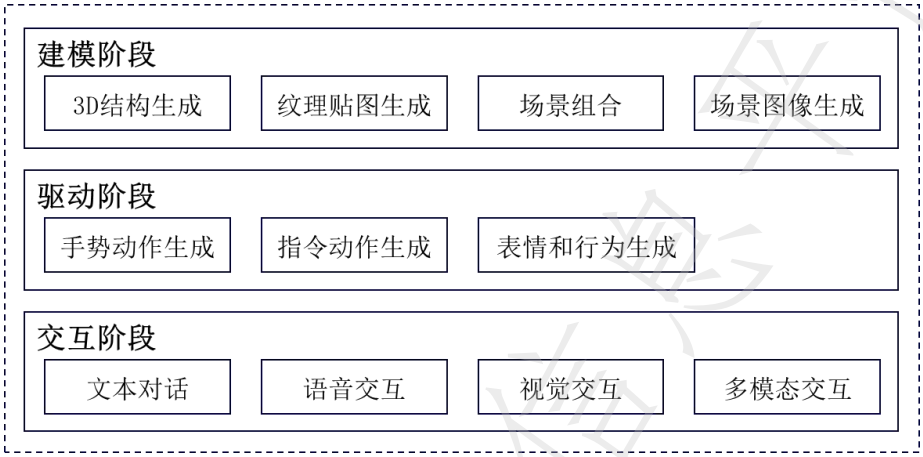

图 2 基于大模型的数字人系统技术框架

（1）建模阶段

在数字人建模阶段，大模型技术应用于三维结构生成、纹理映射生成、场景组合、场景图像生成等。

•2D结构生成：生成可被计算驱动的2D数字人。

•3D结构生成：3D大模型，输入图片或文本prompt，端到端生成3D mesh，生成内容包括数字人、数字人附着资产、数字人场景物品等。

•纹理贴图生成：通过大模型生成纹理贴图，包括漫反射、法相、材质等信息。

•场景组合：基于3D素材库，或过程化3D生成能力，通过大模型理解用户需求，并“组合”出符合需求的3D场景。

•场景图生成（HDRi/Skybox）：通过大模型生成3D场景的2D展开图片（HDRi或Skybox）。

（2）驱动阶段

在数字人驱动阶段，大模型技术可以应用于2D数字人和3D数字人的口型同步生成、手势动作生成、指令动作生成、情感和动作生成等。

•口型同步生成：输入语音或文本，生成2D或3D数字人口型同步动画。

•手势动作生成：输入数字人播报内容，生成2D或3D数字人讲话手势。

•指令动作生成：输入数字人行为指令，生成2D或3D数字人行为动作。

•表情和行为生成：当使用文本大模型生成数字人播报内容时，可以将情绪标签与动作标签作为结构化文本输出，下游动作驱动系统识别标签并生成对应表情或动作。

（3）交互阶段

在数字人机交互阶段，大模型技术可以应用于文本对话、语音交互、视觉交互、多模态交互等。

•文本对话：通过文本大模型实现对话能力。

•语音交互：通过语音大模型实现语音识别、语音合成能力。

•视觉交互：通过视觉大模型识别用户姿态或手势，理解用户意图，控制数字人响应。

•多模态交互：通过多模态大模型实现多模态交互。

7 技术要求

7.1 建模

7.1.1 2D 结构生成

2D数字人生成指利用大模型根据输入信息生成和编辑人物形象并保存为2D图像或视频的系统。

应支持文本提示词生成2D形象，文本提示词包括但不限于背景描述，服装描述，发型描述，配饰描述和性别描述，可通过多次提示达成目标，MOS评分达到4.0以上。

应支持基于大模型对生成的2D形象素材进行编辑，包括但不限于更换背景，扩图，宜具备高级编辑功能，如配饰的添加删除，服装的更换。

宜支持2D数字人风格迁移功能，根据选定的风格模版对数字人形象进行风格匹配。

7.1.2 3D 结构生成

利用AI文生模型、图生模型技术端到端生成3D网格模型。生成的内容包括数字人、附加到数字人的资产以及数字人场景中的项目。

对于几何结构，应支持高精度3D模型细节生成，包括但不限于面部细节、躯干、服装等，头部模型含面部、口腔、上下牙、舌头、左右眼球等，多边形复杂度至少为50,000，几何误差控制在5毫米以内，身体模型顶点数量至少为20,000，身体骨骼节点数量至少为200个，服装模型的顶点数量至少为20,000；对于文生3D模型，render-FID低于4，clip(I-T)大于0.24。

宜支持利用神经辐射场(NeRF)技术合成3D的场景表达。

7.1.3 纹理贴图生成

纹理贴图生成指利用Substance Painter、Substance Designer、Photoshop、大模型等工具生成纹理贴图，包括漫反射、法线和粗糙度信息。

应支持Physically Based Rendering（PBR）材质生成，包括但不限于漫反射、法线、粗糙度、金属度，指标达到纹理分辨率不低于2K。

宜具备纹理生成过程中的错误检测和自动修正功能，错误检测和修正的准确率达到95%以上。

7.1.4 场景组合

场景组合指通过利用3D素材库和大模型，了解用户需求并组成满足这些需求的3D场景。

应支持智能场景组合技术，包括但不限于3D素材库的自动检索、场景元素生成、用户需求的语义理解以及场景元素的逻辑布局，支持快速响应用户指令和高度符合用户预期的场景构建，通过用户输入场景描述，提供选择组合不少于10种，整体美观度评分（MOS）≥4.0，宜提供多轮指令交互完善3D场景优化布局能力。

应支持场景素材的导入或提供素材模板库供选择，每个场景模板都应设计有贴近真实的背景、道具模型以及与流程匹配的交互逻辑。

7.1.5 场景图像生成

场景图像生成指利用大模型生成3D场景的2D展开图像。

应支持高级场景图像生成技术，包括但不限于HDRi环境贴图的动态渲染和360度全景Skybox的生成，指标达到至少8K分辨率的图像质量和逼真的光照效果。

7.2 驱动

7.2.1 口唇同步生成

口唇同步生成指根据输入语音、文本生成2D或3D数字人口唇同步动画，使用户可轻松创建准确、自然和流畅的说话表演。

应支持对多国语言及多地方言实现清晰精准唇形同步，唇形和语音同步的准确率超过90%。

宜具备根据输入语音的变化，包括但不限于音量、语调、语速、情绪等，自动调整口型参数，参数和语音同步的准确率超过80%。

7.2.2 手势动作生成

手势动作生成指根据输入内容生成2D或3D数字人说话手势，用于数字人叙述。

应支持根据输入内容生成自然流畅的手势动作，包括但不限于挥手、摆手、点头等基本手势，支持的手势动作不少于10种。

应支持基于视频动作模版的手势动作生成，包括2D数字人的手势动作视频生成，3D模型的手势动画生成，手部动作自然度MOS评分指标达到4.0以上。

宜具备文本生成手势动作功能，包括但不限于基于对文本内容的意图识别，播放预置的手势动作视频或动画，手势动作的衔接切换连贯，自然度MOS评分指标达到4.0以上。

宜具备根据输入的音乐生成匹配的舞蹈动作，包括但不限于风格、节奏、鼓点等自动匹配，舞蹈动作与音乐节奏的匹配度超过80%。

7.2.3 指令动作生成

指令动作生成是指输入行为指令生成2D或3D数字人相应动作。

应支持多样化的行为动作生成，包括但不限于站立、坐下、行走、指向等动作，指令响应时间小于200毫秒，准确执行预设指令的动作准确率超过98%。

宜具备根据行为指令与场景或者场景中物体进行互动的能力。

宜具备自动分析复杂指令，并执行一系列合成动作响应指令，整体动作连贯、自然。

7.2.4 表情和行为生成

表情和行为生成指在利用大语言模型生成数字人叙事内容时，表情和行为标签可以作为结构化文本输出。驱动系统识别这些标签，并产生相应的表情或行为。

应支持丰富的表情和行为标签生成，包括但不限于高兴、悲伤、愤怒、平静等情感表情，以及点头、挥手、拍手等行为动作，支持的表情标签数量不少于5种，表情与标签匹配度超过95%，支持的行为标签数量不少于10种，行为与标签匹配度超过95%。

宜支持自动同步表情和行为动作标签数据库，动态更新扩容支持的标签数量。

7.3 交互

7.3.1 文本对话

文本对话指通过大语言模型实现会话能力。

应支持多场景对话技术，包括但不限于基于知识库的特定领域问答，开放领域问答，复杂场景多轮对话，人设一致性和稳定性，敏感内容检测和过滤，以及快速回复响应。对话准确率不低于80%，相关性不低于80%，自然度不低于80%，首句延迟不超过3s。

应支持自然语言理解，包括但不限于意图识别、上下文管理、情感识别与回应、知识问答等，指标达到意图识别准确率≥90%，情感识别准确率≥85%，知识问答正确率≥80%。

宜具备个性化对话能力，包括但不限于基于用户偏好的对话内容调整，以及主动学习机

制以优化对话质量，对话自然度评分（MOS）≥4.5（满分5分）。

宜具备多语言对话能力，语种支持中文、英文以及常见方言等；具备主动学习能力，能够在开放场景交互中主动学习和持续学习；高度拟人化能力，包括人设、人格、情感、观点等维度，在多轮对话中保持高度一致性；具有外部信息获取能力，实时获取最新数据。

7.3.2 语音交互

语音交互指通过大语音模型实现语音识别和语音合成功能。

应支持多语种的语音识别、克隆、翻译、合成等功能。语种至少支持中文普通话、英文，并支持但不限于以下业务并达到相应的指标：

应支持高精度语音识别和自然流畅的语音合成，包括但不限于在各种环境噪声下的准确识别，以及接近人类自然语音的回复生成，指标达到语音识别准确率≥90%，语音合成自然度评分（MOS）≥4.0。

应支持将用户的语音转化为可处理的文本信息，支持结合前后的语音内容进行综合分析，支持自动进行纠错，提高对特定语义的理解准确度，指标达到准确率≥90%。

应支持小样本声音克隆功能，用户需提供的音频素材小于20句，训练时间小于30分钟，声音质量评分（MOS）≥3.5。

宜支持多样性的音色供用户选择，包括但不限于不同性别、年龄段以及特定角色属性（如，客服、讲解员、故事讲述者等）的音色，满足不同应用场景和用户偏好需求。

宜支持多语种音色合成，一键翻译数字人视频。

宜支持开放场景语音识别，单字准确率≥90%。

宜支持多语种语音克隆与合成，包括但不限于英语、汉语、法语、德语等，克隆目标语音长度不超过10s，端到端时延不超过300ms、WER不超过0.04、克隆SIM不低于0.7。

宜具备对不同地区的口音的自动检测与识别，包括但不限于四川口音、广东口音等，识别准确率不低于80%。

宜支持数字人跨语种翻译，即数字人接收到语音后，保持音色、韵律、情绪等情况下，用指定语种讲述翻译后的语音，端到端时延不超过1000ms、WER不超过0.06、SIM不低于0.67。

宜具备对语音特征进行分析，包括但不限于语调的高低、语速的快慢、音量的大小以及语音的韵律等，实现对说话者情绪状态的判断，包括但不限于准确判断喜怒哀乐等基本情绪，指标达到情绪判断准确率不低于80%。

宜具备声纹识别功能，包括但不限于用户身份验证和个性化语音交互，以及情感语音合成能力，以丰富语音交互的情感表达。

7.3.3 视觉交互

视觉交互指利用大型视觉模型来识别用户的姿势或手势，理解用户的意图，并控制数字人的反应。

应支持多维度的肢体姿势识别，包括但不限于全身姿势和表情，指标端到端时延不超过50ms，响应成功率高于90%。宜具备用户身份识别能力，包括但不限于通过面部特征进行精准识别，指标达到身份识别准确率不低于90%。

宜具备高度自然的人机交互能力，包括但不限于实时反馈用户肢体动作和手势变化，以及根据这些输入判断用户意图，并调整数字人的行为和反应，指标达到意图识别准确度不低于80%，实现用户数字人的眼神接触率高于70%，满意度评分高于80%。

宜具备复杂场景适应能力，包括但不限于多人场景中的目标用户识别，以及视线追踪功能，以提升视觉交互的自然性和精准度。

7.3.4 多模态交互

多模态交互指在多模态大模型增强的数字人交互背景下，数字人类可以通过视觉、文本、语音等多种感官模式与用户进行复杂的交互。这种交互不局限于单一模式，而是集成了多种输入输出方式，提供了更丰富、更自然的交互体验。

应支持跨模态融合，支持多种输入方式，包括但不限于文本、语音、视觉等信息的有效整合，实现多模态指令的准确解析和执行，多模态指令解析正确率≥90%。

应支持数字人的表情、动作、手势的表达，包括但不限于常见的喜怒哀乐表情、常见的引导动作和示意手势。

宜具备持续学习能力，通过不断收集和分析多模态交互数据，优化交互模型，提升用户体验。

宜具备在数字人播报的同时，以结构化的卡片形式呈现信息，包括但不限于关键数据、选项列表和示意图等。

宜具备对话填充能力，包括但不限于根据等待时间长短进行不同程度的安抚。

宜具备对话打断能力，包括但不限于快速响应并准确判断用户意图。

参 考 文 献

[1] ITU/T F.748.15 数字人应用系统基础框架和评测指标(Framework and metrics for digital human application systems)