GY

中华人民共和国广播电视和网络视听行业标准

GY/T XXX—XXXX

数字虚拟人技术要求

Technical requirements for digital human

（报批稿）

XXXX－XX－XX 发布　　XXXX－XX－XX 实施

国家广播电视总局　　发布

目　　次

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。

本文件由全国广播电影电视标准化技术委员会（SAC/TC 239）归口。

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。

本文件起草单位：中广电广播电影电视设计研究院有限公司、国家广播电视总局广播电视科学研究院、国家广播电视总局广播电视规划院、腾讯云计算（北京）有限责任公司、中国传媒大学、拓尔思信息技术股份有限公司、北京中科深智科技有限公司、北京七维视觉科技有限公司、湖南芒果无际科技有限公司、四川广播电视台、山东广播电视台、成都索贝数码科技股份有限公司、北京中科睿鉴科技有限公司。

本文件主要起草人：王嘉、郭晓强、宁金辉、姚琼、谢东霖、徐永太、赵天晓、严明、金启棣、姚高远、张小雨、张博文、梁妙纵、任家萱、肖婧、王宇、程辉、宋健、徐立、常帅、谷燕京、魏忠书、陈智、郑涛、孙琳、陈磊、刘晶、李洋、李晶晶、徐超、罗志文、艾斌、韩庆秋、欧翔、陈尧森。

数字虚拟人技术要求

1 范围

本文件规定了广播电视和网络视听行业数字虚拟人的技术要求，对于数字虚拟人分类、应用场景、形象、驱动技术、平台能力、安全能力提出规范要求。

本文件适用于广播电视和网络视听行业数字虚拟人的系统建设、创作和应用。

2 规范性引用文件

本文件没有规范性引用文件。

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：YD/T 4393.1—2023，3.1.1]

3.2

视频合成实时率 video synthesis real-time rate

视频合成耗时与输出视频时长比值。

3.3

数字虚拟人应用主体 digital human operation entity

使用数字虚拟人服务制作、发布、传播、交互信息的组织或个人。

3.4

蒙皮 skin

在3D数字虚拟人制作中，将模型绑定在骨骼上，通过骨骼驱动虚拟人模型的技术。

4 缩略语

下列缩略语适用于本文件。

AI 人工智能（Artificial Intelligence）

APaaS 应用平台即服务（Application Platform as a Service）

APP 应用软件（Application software）

ASR 自动语音识别技术（Automatic Speech Recognition）

DurIAN 基于告知时长信息注意力网络的多模态语音合成模型（Duration Informed Attention Network For Multimodal Synthesis）

FPS 每秒帧数（Frames Per Second）

H5 超文本标记语言5（Hyper Text Markup Language 5）

HiFi-GAN 基于对抗学习网络的高效高保真语音合成模型（Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis）

HTTPS 超文本传输安全协议（Hypertext Transfer Protocol Secure）
JSON JS对象简谱（JavaScript Object Notation）
NLP 自然语言处理（Natural Language Processing）
PC 个人计算机（Personal Computer）
SSML 语音合成标记语言（Speech Synthesis Markup Language）
TTS 从文本到语音（Text To Speech）
UE 虚幻引擎（Unreal Engine）
VAE 变分自编码器模型（Variational Autoencoder）
2D 二维（Two Dimensional）
3D 三维（Three Dimensional）

5 总体架构

5.1 数字虚拟人分类及应用场景

5.1.1 数字虚拟人分类

数字虚拟人按照人物形象分类，分为2D数字虚拟人和3D数字虚拟人。
数字虚拟人按照交互模式分类，分为非交互式数字虚拟人和交互式数字虚拟人。
数字虚拟人按照驱动模式分类，分为算法驱动型数字虚拟人和真人驱动型数字虚拟人。

5.1.2 数字虚拟人应用场景

数字虚拟人主要应用场景分为内容播报、交互客服、虚拟演播和内容创作等。其中内容播报包含新闻资讯播报/手语播报、电影/电视/专题片/纪录片介绍和直播带货等；交互客服包含虚拟客服、智能助手和交互问答等；虚拟演播包含综艺主持、虚拟演唱会、文娱活动和用户代理虚拟分身等；内容创作包含影视创作、视频创作、广告创作和游戏创作等。

5.2 数字虚拟人总体技术架构

数字虚拟人总体技术架构包括数字虚拟人形象、算法驱动能力、真人驱动能力、平台能力和安全能力等内容，总体架构见图1。

数字虚拟人形象包括2D真人、2D卡通、3D写实、3D卡通和建模技术。

数字虚拟人算法驱动包括驱动能力、合成能力和多模态能力。其中，驱动能力又分为文本驱动能力、语音驱动能力和视频驱动能力；合成能力包含语音合成能力和视频合成能力；多模态能力包含语音识别能力和自然语言处理能力。

数字虚拟人真人驱动包含身体动作捕捉、表情捕捉和捕捉数据。

数字虚拟人平台能力，应支持数字虚拟人的制作和生成，支持数字虚拟人的维护配置。平台服务能力可选择云服务或者本地服务。

数字虚拟人安全能力，应为数字虚拟人应用提供安全保障，覆盖数据及算法安全和个人信息保护等。

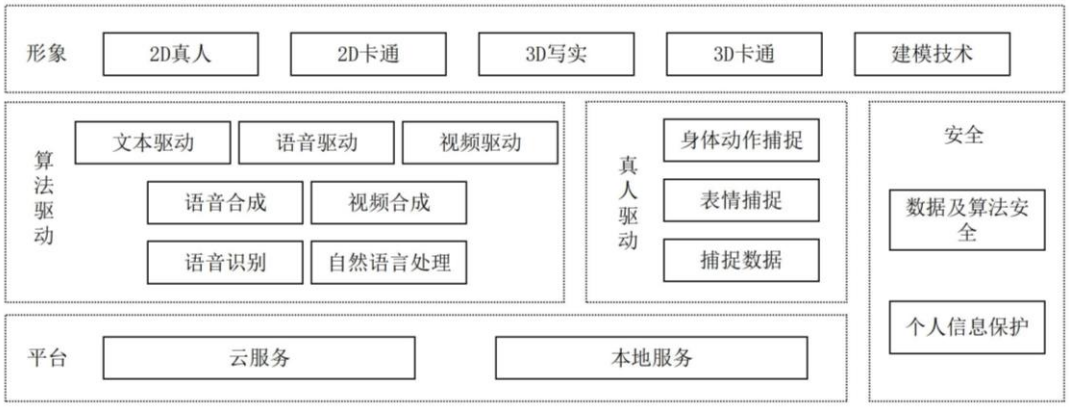

图1　数字虚拟人总体技术架构

6　形象要求

6.1　总体要求

数字虚拟人形象，应满足如下要求：

a）　符合场景的任务设定，在人物形象、表情、服饰等方面得体、美观；

b）　支持全身、大半身、半身不同景别姿态；

c）　形象完好，不存在扭曲、马赛克、跳帧、破损、音视频延时、口唇不一致等情况；

d）　支持装扮、服饰的更换；

e）　不存在侵犯第三方权利及法律法规禁止的其他情形。

6.2　2D 数字虚拟人的形象要求

6.2.1　2D 真人数字虚拟人的形象要求

2D真人数字虚拟人形象，应满足如下要求：

a）　支持真人形象复刻，形象逼真自然，语音自然流畅；

b）　支持通过照片、视频等方式生成形象，保证面部五官、肤色、牙齿、明暗等准确还原。

6.2.2　2D 卡通数字虚拟人的形象要求

2D卡通数字虚拟人形象，应满足如下要求：

a）　支持 2D 卡通形象绘制，对特有的卡通形象进行建模；

b）　支持不同景别、姿态，形象灵动活泼，动作自然舒展。

6.3　3D 数字虚拟人的形象要求

6.3.1　3D 写实数字虚拟人的形象要求

3D写实数字虚拟人形象，应满足如下要求。

a）　支持通过 3D 建模或真人扫描等方式刻画形象，头部模型覆盖面部、口腔、上下牙、舌头、独立左右眼球、眼睑、泪腺等；头部、面部、身体纹理有效区域面积高；毛发系统，如头发、睫毛、面部绒毛等纹理清晰。

b）　支持对形象的美型、加工和风格化等。

c）　支持不同角度、景别、姿态的灵活转换。

d) 支持丰富的动作类型。
e) 支持光照效果的处理，如光影、折射、反射等效果。
f) 支持人形骨骼、蒙皮建模。
g) 支持按 1:1 的比例，对真人进行复刻。

6.3.2 3D 卡通数字虚拟人的形象要求

3D卡通数字虚拟人形象，应满足如下要求：
a) 支持 3D 卡通形象绘制等方式，对特有的卡通形象进行建模；
b) 支持不同景别、角度、姿态，形象灵动活泼，动作自然舒展；
c) 支持丰富的动作类型。

6.4 建模技术要求

数字虚拟人建模能力，满足如下要求：
a) 模型网格线应均匀、平滑，不应出现尖刺现象；
b) 人体网格应闭合，部位之间不应出现缝隙和破洞；
c) 应具有易于应用的拓扑结构，能够适应常态化的数据存储和传输需求；
d) 骨骼系统设计应符合生物生理及运动结构，实现自然、真实的动作效果；
e) 骨骼控制器应易于操作，确保动画制作的精度和效率；
f) 蒙皮权重应适用于不同关节和组织部件，实现平滑过渡；
g) 建模流程应包括原画设计、3D 网格模型搭建与雕刻、贴图和材质绘制；
h) 宜支持头发、衣服等配件的物理模拟渲染和交互；
i) 网格拓扑应能够兼容不同的软件和工具。

7 算法驱动能力要求

7.1 驱动能力技术要求

7.1.1 基本要求

数字虚拟人算法驱动能力，满足如下要求：
a) 应支持多种驱动方式，包括文本驱动、语音驱动、视频驱动等；
b) 应支持单一技术驱动和混合技术驱动的方式；
c) 应支持的驱动范围包括口型、面部表情、肢体等；
d) 数字虚拟人展示应支持端侧渲染，宜兼容多操作系统。

7.1.2 文本驱动能力要求

数字虚拟人文本驱动能力，应满足如下要求：
a) 支持不同的 TTS 模型和实现框架，例如 HiFi-GAN、DurIAN 等；
b) 支持文本的处理，支持区分文本中的数字和英文字母；
c) 支持从文本中提取信息，判断文本对应的情感情绪、重音位置、常见多音字，基于多模态交互系统，驱动生成数字虚拟人的语音、动作、表情；
d) 支持中文、英文的文本驱动。

7.1.3 语音驱动能力要求

数字虚拟人语音驱动能力，应满足如下要求：

a) 支持通过“语音驱动”及人像合成的处理流程，驱动生成数字虚拟人的语音、动作、表情、口型；
b) 至少支持中文普通话的语音输入；
c) 支持排除语音中的背景噪声；
d) 具备动态语音活性检测能力，在传统语音活性检测基础上，增加对长语音场景下停顿的检测，动态调整静音门限，识别其中的有效文本。

7.1.4 视频驱动能力要求

数字虚拟人视频驱动能力，应满足如下要求：

a) 支持计算机视觉算法，基于记录面部表情和肢体动作的视频，驱动生成数字虚拟人的语音、动作、表情、口型；
b) 支持的视频包括通过摄像头记录人体面部表情、肢体动作的视频；
c) 支持实时或离线的驱动方式。

7.2 合成能力技术要求

7.2.1 语音合成能力要求

数字虚拟人语音合成能力，应满足如下要求：

a) 支持端到端语音合成模型，支持 HiFi-GAN、VAE、Diffusion（扩散模型）、Glow（流生成模型）、DurIAN 等多种语音合成模型；
b) 语音合成效果自然，音质音效贴近真人；
c) 实现字词级别的音量、时长的细粒度控制，实现音量、语速的调节；
d) 实现多情感高表现力的可控语音合成效果，根据文本内容自动切换合成不同情感的语音；
e) 支持文本语音合成、真人语音录制、真人语音变声等多种形式；
f) 支持针对应用场景（包括播报、解说、诗歌、阅读、客服等），生成多种语音合成风格；
g) 支持从分钟级到小时级不同语料规模快速个性化定制；
h) 支持流式合成的方式。

7.2.2 视频合成能力要求

数字虚拟人视频合成能力，应满足如下要求：

a) 支持多种渲染引擎技术对数字虚拟人形象进行渲染，包括 UE、Unity 等；
b) 支持图像增强技术，改善视频质量和用户体验；
c) 支持视频离线合成或实时渲染合成；
d) 支持通过人脸的图像或视频内容进行视频合成；
e) 支持不同分辨率、码率的视频合成；
f) 在 1080P 分辨率条件下，视频合成实时率不高于 1；
g) 合成后的视频流畅，支持帧率不小于 25FPS。

7.3 多模态能力技术要求

7.3.1 总体要求

数字虚拟人多模态能力，应满足如下要求。

a) 发音准确，不存在漏音吞音、多余发音、音素错误、音调错误等情况；语速、停顿断句、音高、音长、音量、重音等符合自然语言发音规律；语音语调舒适；常见多音字发音正确。
b) 口型、唇形自然，与发音同步，符合发音的规律，具备饱满度和表现力。
c) 动作精准、自然，与交互语境契合，动作包括但不限于头部、肢体、全身等部位。
d) 支持实时渲染技术，支持基于物理光照和实际环境光源、相机位置、材质参数等实时计算，完成图像渲染。
e) 在交互客服场景下，支持多轮对话能力，根据上下文内容或用户的问询，进行判断选择，完成用户交互；支持通过对话树等方式，完成不同业务场景下的多轮对话流程跳转及应答。

7.3.2 语音识别要求

数字虚拟人语音识别能力，应满足如下要求：
a) 支持对语音进行预处理；
b) 支持将语音信号转化成对应的表示向量；
c) 支持基于深度学习的语音识别模型；
d) 支持区分并识别不同的说话人；
e) 支持识别语音中所表达出的情感信息。

7.3.3 自然语言处理要求

数字虚拟人自然语言处理能力，应满足如下要求：
a) 支持对文本进行词法分析；
b) 支持理解文本的语法结构；
c) 支持理解文本的含义；
d) 支持对文本进行分类；
e) 支持将一种语言的文本转化成另一种语言的文本；
f) 支持回答用户提出的问题，包括基于检索的问答系统和基于生成的问答系统。

8 真人驱动能力要求

8.1 身体动作捕捉要求

8.1.1 基本要求

数字虚拟人动作捕捉能力，应满足如下要求：
a) 支持对采集到的运动数据进行去噪、对齐、插值、滤波等处理，得到准确、平滑、连续的运动数据；
b) 支持将运动数据应用于虚拟人物角色的动画制作、影视记录、直播等领域中，可以实现逼真、流畅、自然的运动效果；
c) 支持数据时间同步和坐标系转换。

8.1.2 光学动捕技术要求

数字虚拟人光学动捕能力，满足如下要求：
a) 应支持使用红外 LED 灯或激光灯作为光源，确保灯光强度和均匀性，同时避免干扰和反射；
b) 宜支持采用自发光或高反光率的球体、贴片作为标记点，标记点数量和布局应根据运动物体的特点和运动需求进行合理设置；

c） 摄像机数量和布局应根据运动物体的特点和运动需求进行合理设置，确保覆盖面积和视角的完整性和重叠度；
d） 应支持使用特殊的光学摄像机和数据采集软件，将运动物体的标记点运动轨迹记录下来，形成运动数据。

8.1.3 惯性动捕技术要求

数字虚拟人惯性动捕能力，应满足如下要求：
a） 支持使用惯性传感器进行运动捕捉，传感器数量和布局根据捕捉物体的特点和运动需求进行合理设置，通常包括加速度计、陀螺仪、磁力计等多种传感器；
b） 支持将传感器采集到的运动数据记录下来，形成运动数据，在数据采集过程中需要确保传感器的稳定性和准确性。

8.1.4 视觉动捕技术要求

数字虚拟人视觉动捕能力，应满足如下要求：
a） 支持使用摄像机进行运动捕捉，摄像机性能、数量和布局根据捕捉物体的特点和运动需求进行合理设置；
b） 支持在摄像机采集画面上添加标记点，标记点数量和布局根据运动物体的特点和运动需求进行合理设置；
c） 支持将摄像机拍摄到的标记点运动轨迹进行记录，形成运动数据。

8.2 表情捕捉要求

数字虚拟人表情捕捉能力，应满足如下要求：
a） 支持使用摄像机、红外雷达等传感器或标记点的方式对动捕演员面部进行捕捉；
b） 支持使用传感器采集面部标记点的位置、形状等数据形成面部表情的运动数据；
c） 支持对采集到的面部运动数据进行处理，提取面部表情的关键特征，如面部肌肉的收缩程度、面部皱纹的形状等，形成面部表情的模型；
d） 支持将面部表情模型映射到虚拟人物的面部模型上，实现虚拟人物面部表情的变化；
e） 支持根据实际应用需求对面部表情的映射效果进行校准和调整，使虚拟人物的面部表情更加逼真、自然、流畅。

注：动捕演员又称为“中之人”，是指将自身动作、表情、语音等信息实时映射到数字虚拟人身上的演员。

8.3 捕捉数据要求

数字虚拟人捕捉数据能力，应满足如下要求：
a） 能够兼容捕捉数据文件的格式，以便进行数据解析和处理；
b） 支持实时捕捉数据，以便快速响应并展示出相应的动作效果；
c） 支持精确还原真实人体的捕捉效果，保证捕捉数据的采样率和精度满足实际使用；
d） 支持对动作控制器的参数进行调整，以便实现各种不同的动作效果。

9 平台能力要求

9.1 平台基本要求

数字虚拟人平台，应满足如下要求：
a） 支持针对内容播报、交互客服、虚拟演播、内容创作等应用场景；

b） 支持创作不同类型的数字虚拟人；
c） 支持数字虚拟人形象的资产管理、业务服务配置及内容生产服务；
d） 支持数字虚拟人形象租赁；
e） 支持数字虚拟人形象选型、音色配置、背景空间管理、发音及动作配置、会话管理、流程管理等功能；
f） 支持多种 AI 模型和算法；
g） 平台生成的数字虚拟人具备多模态交互能力；
h） 具备丰富的语音及动作库；
i） 支持离线、实时的数字虚拟人生成方式；
j） 真人驱动型平台技术支持真人驱动的模式，真人驱动可以和算法驱动混合使用，相互接管。

9.2 平台部署要求

数字虚拟人平台部署，满足如下要求：
a） 应支持公有云部署、私有云部署或本地化部署方式；
b） 应支持多类型前端接入能力，包括但不限于 PC、移动终端、大屏设备等终端接入设备，以及网页、APP、小程序、H5 等应用形式，满足系统的前端兼容性；
c） 宜支持运用微服务、集群的部署方式；
d） 宜采用负载均衡、分布式数据库等技术。

9.3 平台服务要求

数字虚拟人平台服务，满足如下要求。
a） 应支持数字虚拟人形象实时生成和动态配置，支持实时视频推流服务，支持云渲染。
b） 应支持可视化的数字虚拟人配置，包括形象、语音、服饰、姿态、位置等。
c） 应支持可视化的合成和驱动参数配置，包括 ASR、NLP、TTS 等参数设置；支持 HTTPS、JSON 等协议进行参数的调用和配置。
d） 应支持用于生成或配置数字虚拟人的素材输入，包括图片、视频、文本等。
e） 应支持数字虚拟人形象、音色、服饰、姿态、动作的定制。
f） 应支持 SSML 文本和数字虚拟人进行视频制作。
g） 应支持视频背景、内嵌字幕的配置。
h） 应支持视频制作进度、数字虚拟人资源的查询。
i） 应支持对接第三方 3D 形象模型等数字资产，适配对应的模型参数，完成渲染及驱动。
j） 宜配置大屏、音响、麦克风、摄像头等数字虚拟人展示及交互设备，保障数字虚拟人多模态识别及交互效果。
k） 云端部署方式应支持 APaaS 的接口调用方式，具备权限、公共参数、签名等访问控制机制。
l） 本地部署方式应支持数字虚拟人形象与 ASR、NLP、TTS 等底层技术能力的解耦。
m） 本地部署方式应支持多种通信协议与第三方音视频系统对接。

10 安全能力要求

10.1 数据及算法安全

数字虚拟人应用主体对其所处理的数据及算法安全负责，满足如下要求：
a） 应在法律、行政法规规定的目的和范围内收集、使用数据；
b） 应采取相应的技术措施和其他必要措施，保障数据安全；

c） 应根据业务需求，配置数据存储和使用的安全策略，为用户配置合理的权限，具备相应的访问控制机制；
d） 在数据采集过程中应采用必要的技术手段，具备数据准确性、安全性的保障能力；
e） 在数据传输过程中，对于需要进行加密处理并传输的业务数据，应部署相应的加密措施；
f） 不准许制作、复制、发布、传播虚假内容。

10.2 个人信息保护

数字虚拟人应用主体对其所处理的个人信息安全负责，满足如下要求：
a） 处理个人信息应遵循合法、正当、必要和诚信原则，不应通过误导、欺诈、胁迫等方式处理个人信息；
b） 处理个人信息应具有明确、合理的目的，并应与处理目的直接相关，采取对个人权益影响最小的方式；
c） 处理个人信息前，应以显著方式、清晰易懂的语言真实、准确、完整地向个人告知个人信息的处理目的、方式、范围；
d） 当对真实人脸、人声等生物识别信息进行编辑时，应告知被编辑的个人，并取得其单独同意。

参 考 文 献

[1] 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
[2] 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
[3] 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
[4] ITU-T F.748.15 Framework and metrics for digital human application systems
[5] ITU-T F.748.14 Requirements and evaluation methods of non-interactive 2D real-person digital human application systems