

ICS 号: xxxx-xxxx
中国标准文献分类号

世界超高清视频产业联盟标准

T/UWA xxxx-xxxx

3D 真人形象数字人视觉与交互效果分级的技术要求

Technical requirements for grading the visual and interactive effects of

3-Dimensional real person digital human

(V1.0)

xxxx-xx-xx 发布

xxxx-xx-xx 实施

世界超高清视频产业联盟

目 次

前 言	III
1 范围	4
2 规范性引用文件	4
3 术语和定义	4
3.1 3D 真人形象数字人 (3-Dimensional real person digital human)	4
4 缩略语	4
5 指标及规范性描述	4
5.1 人物效果	4
5.2 识别感知	5
5.3 交互决策	5
6 3D 真人形象数字人分级	6
6.1 3D 真人形象数字人分级原则	6
6.2 3D 真人形象数字人等级技术要求	6
6.3 3D 真人形象数字人分级标准	9
6.4 3D 真人形象数字人分级标准	10

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》给出的规则起草。

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。

本文件由世界超高清视频产业联盟提出并归口。

本文件主要起草单位：。

本文件主要起草人：。

3D 真人形象数字人视觉与交互效果分级的技术要求

1 范围

本文件规定了 3D 真人形象数字人视觉和交互效果的分级参考框架，并描述了人物特征、识别感知、交互决策三个维度的评估指标。

本文件适用于对 3D 真人形象数字人应用效果作出分级，用于为供需双方根据场景需求选择数字人产品提供参考。

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。

GB/T 21023-2007 中文语音识别系统通用技术规范

GB/T 36464.4-2018 信息技术 智能语音交互系统 第 4 部分:移动终端

GY/T 307-2017 超高清清晰度电视系统节目制作和交换参数值

T/UWA 012.6-2022 “百城千屏”超高清视音频传播系统网络传输技术要求

ISO/IEC 14496-10:2022 Information technology — Coding of audio-visual objects — Part 10: Advanced video coding 信息技术.视听对象的编码.第 10 部分：高级视频编码

ITU-T P.800.1 Mean opinion score (MOS) terminology 平均意见得分（MOS）术语

3 术语和定义

下列术语以及定义适用于本文件。

3.1 3D 真人形象数字人 (3-Dimensional real person digital human)

3D 真人形象数字人是指基于数字技术，在数字世界具备拟人化外形的三维立体的数字人。

4 缩略语

下列缩略语适用于本文件。

3D 三维 (3-Dimensional)

5 指标及规范性描述

5.1 人物效果

1. 面部拟真度

面部拟真度指3D真人形象数字人面部拟人化的程度。

2. 视觉精细度

视觉精细度指的是3D真人形象数字人形象的精细程度，包括3D真人形象数字人模型的毛发、牙齿、皮肤等细节呈现程度。

3. 基础表情数量

指3D真人形象数字人可呈现的基础表情数量。

4. 唇动效果

指3D真人形象数字人说话时嘴唇动态效果。

5. 文字转语音准确率

指3D真人形象数字人语音合成并播放时的准确率。

文字转语音准确率=符合条件的测定值个数/总测定值个数*100%。

6. 语音自然度

指3D真人形象数字人语音合成或真人发音的感知自然度。

7. 基础肢体动作数量

3D真人形象数字人基础动作数量。

8. 肢体动作自然度

3D真人形象数字人肢体动作的自然程度。

9. 帧率

3D真人形象数字人以帧称为单位的位图图像连续出现在显示器上的频率。

10. 分辨率

3D真人形象数字人在显示器上呈现的像素。

5.2 识别感知

1. 语音转文字准确率

用于评价数字人识别用户语音并转为文本的准确率。

2. 人脸识别率

3D真人形象数字人在进行人脸识别的过程中正确识别人脸的概率。

3. 情绪识别准确率

在积极、消极、中立三类情绪识别中，正确识别用户对话情绪样本数占总样本数的概率。

5.3 交互决策

1. 对话交互完成率

3D真人形象数字人在与用户进行对话交互时能够正确理解用户意图并能给出相应回答的比率。

2. 表情反馈正确率

3D真人形象数字人在与用户进行表情交互时能够正确理解用户表情并反馈正确表情的比率。

3. 肢体反馈正确率

用户对3D真人形象数字人肢体交互时数字人能够正确理解用户动作并反馈合适动作的比率。

4. 对话、表情、肢体行动的一致性

用于考核数字人固定时长（单位：s）视频中音、容、行三项的匹配度，若出现音频提前、表情多余、缺失，肢体动作延迟、肢体动作错位等均视为不匹配。在标准评测时间内，EN代表出现音频提前、表情多余、缺失，肢体动作延迟、肢体动作错位等错误数所占时长，T代表总时长， F代表总分值100。则：

$$S=F*（1-EN/T*100%）$$

5. 平均卡顿时长

与数字人交互过程中，发生卡顿的平均时长，包括视频画面卡顿、语音动作不匹配的感知卡顿、语音卡顿等。在标准评测时间内，获取单次卡顿时长（记为ti，即本次开始出现卡顿到本次卡顿结束的时间差）和卡顿次数。则：

$$平均卡顿时长 = \Sigma (ti) / 卡顿总次数$$

6 3D 真人形象数字人分级

6.1 3D 真人形象数字人分级原则

基于3D真人形象数字人在人物效果、识别感知、交互决策三个一级维度，对细分指标提供测算方式及参数说明。

6.2 3D 真人形象数字人等级技术要求

表 1 数字人等级技术要求

一级指标	二级指标	细分指标	测算方式和参数说明
人物效果	形象效果	面部拟真度	抽取 N 张数字人图片
			①使用经过 LFW 人脸数据训练的 StyleGan 生成对抗网络中的判别器对数字人人脸进行评分使用 1000 张示例作为分箱标准，对最终图片的人脸拟真度进行分箱
			②给 k(k>=10)位有专业背景知识的测试人员进行分箱判断
			③主客观一致性比率=算法与测试人员分箱结果相同的数字人图片数量 / N*100%
		视觉精细度	抽取 N 张数字人图片
			①通过图像清晰度算法（如：Tenengrad 梯度方法）计算数字人的不同位置图片的精细程度，对不同位置的锐度根据视觉的重要程度进行加权计算作为最终的视觉精细度，对最终加权的精细度得分进行分箱得出分值
			②给 k(k>=10)位有专业背景知识的测试人员进行分箱判断
			③主客观一致性比率=算法与测试人员

			分箱结果相同的数字人图片数量 / $N \times 100\%$
	表情效果	基础表情数量	参考 ISO/IEC 14496-10:2022 Information technology — Coding of audio-visual objects — Part 10: Advanced video coding
		唇动效果	对数字人说话的视频片段进行评测, 通过 click 的方式进行点击, 唇动效果不好的字记为 fp: +1, 准确率=总字数-fp/总字数
	语音效果	文字转语音准确率	对采用语音合成技术的数字人视频片段进行评测, 通过 click 的方式进行点击, 发音不准确、发音错误的字记为 fp: +1, 准确率=总字数-fp/总字数*100%
		语音自然度	参考 GB/T 36464.4-2018 《信息技术智能语音交互系统 第4部分:移动终端》
	动作效果	基础肢体动作数量	参考 ISO/IEC 14496-10:2022 Information technology — Coding of audio-visual objects — Part 10: Advanced video coding
		肢体动作自然度	抽取 N 张数字人图片
			①使用 3D Human Pose Estimation, 基于 Human3.6M 动作库, 对 3D 动态数字人进行识别, 跟踪其动作, 并观察穿模的数量点对识别后的动作骨骼进行计算并分析各数字人活动的自由度数量使用 1000 张示例作为分箱标准, 对自由度数量进行分箱得出分值
			②给 k(k≥10)位有专业背景知识的测试人员进行分箱判断
	呈现效果	帧率	③主客观一致性比率=算法与测试人员分箱结果相同的数字人图片数量 / $N \times 100\%$
		分辨率	GY/T 307-2017 超高清晰度电视系统节目制作和交换参数值 T/UWA 012.6-2022 “百城千屏”超高清视音频传播系统网络传输技术要求。
识别感知	人脸识别	人脸识别率	基于基准的 LFW 人脸数据集供数字人识别, 判断其识别的准确率使用 1000 张示例作为分箱标准, 对识别百分数进行分箱
	语音识别	语音转文字准确率	参考 GB/T 21023-2007 《中文语音识别系统通用技术规范》 参考 ISO 9241-154:2013: Ergonomics of human-system interaction — Part 154: Interactive voice response (IVR)

交互决策	情绪识别	情绪识别准确率	applications ① 基于 Multi-lingual Sentiment Analysis (MLSA) 算法和 Universal Dependency Treebanks 数据库，通过语音转文字后对用户话语进行情绪识别，积极、消极、中立三类 ② 给 $k(k \geq 10)$ 位有专业背景知识的测试人员进一步结合上下文语境对回答情绪进一步判别 ③ 主客观一致性比率 = 算法与测试人员判别结果相同的文本数量 / 数据库语料总数 * 100%
	对话交互	对话交互完成率	① 通过算法如剑桥提出的 Dialog State Tracking Challenge11 种评测指标和 3 种评测时机 (schedule) 基于语料库测试出基础值并分箱 ② 给 $k(k \geq 10)$ 位有专业背景知识的测试人员进一步分箱判别 ③ 主客观一致性比率 = 算法与测试人员判别结果相同的对话轮次数量 / 总对话轮数 * 100%
	表情交互	表情反馈正确率	① 通过 VGG19 和 Resnet18 给出表情识别与分类 ② 给 $k(k \geq 10)$ 位有专业背景知识的测试人员进一步判别 ③ 主客观一致性比率 = 算法与测试人员判别结果相同的表情数量 / 总反馈数量 * 100%
	肢体交互	肢体反馈正确率	① 通过微软 smooth model 给出判别分箱 ② 给 $k(k \geq 10)$ 位有专业背景知识的测试人员进一步分箱判别 ③ 主客观一致性比率 = 算法与测试人员判别结果相同的肢体反馈数量 / 总反馈数量 * 100%
	交互质量	对话、表情、肢体反馈的一致性	用于考核数字人固定时长 (单位: s) 视频中音、容、行三项的匹配度，若出现音频提前、表情多余、缺失，肢体动作延迟、肢体动作错位均视为不匹配该指标总得分为 100，计算公式为：$S = F * (1 - EN / T * 100\%)$ (EN 代表出现音频提前、表情多余、缺失，肢体动作延迟、肢体动作错位等错误数所占时长，T 代表总时长，

			T>600s, F 代表总分值 100
		平均卡顿时长	交互过程中, 发生卡顿的平均时长

6.3 3D 真人形象数字人细分指标分级标准

1. 3D真人形象数字人人物效果分级要求

a) 1 级人物效果, 面部拟真度<40%, 视觉精细度<40%, 基础表情数量<20, 唇动效果<6, 文字转语音准确率<40%, 语音自然度为 1, 基础肢体动作数量<76, 肢体动作自然度<40%, 帧率<25FPS, 分辨率<2K。

b) 2 级人物效果, 面部拟真度≥40%, 视觉精细度≥40%, 基础表情数量≥20, 唇动效果≥6, 文字转语音准确率≥40%, 语音自然度为 2, 基础肢体动作数量≥76, 肢体动作自然度≥40%, 帧率≥25FPS, 分辨率≥2K。

c) 3 级人物效果, 面部拟真度≥60%, 视觉精细度≥60%, 基础表情数量≥30, 唇动效果≥9, 文字转语音准确率≥60%, 语音自然度为 3, 基础肢体动作数量≥116, 肢体动作自然度≥60%, 帧率≥50FPS, 分辨率≥4K。

d) 4 级人物效果, 面部拟真度≥80%, 视觉精细度≥80%, 基础表情数量≥52, , 唇动效果≥12, , 文字转语音准确率≥80%, , 语音自然度为 4, 基础肢体动作数量≥156, 肢体动作自然度≥80%, 帧率≥50FPS, 分辨率 8K。

e) 5 级人物效果, 面部拟真度≥90%, 视觉精细度≥90%, 基础表情数量≥86, 唇动效果≥15, 文字转语音准确率≥99%, 语音自然度为 5, 基础肢体动作数量≥196, 肢体动作自然度≥90%, 帧率≥120 FPS, 分辨率≥8K。

表 2 数字人人物效果分级要求

一级指标	二级指标	分级				
		1 级	2 级	3 级	4 级	5 级
人物效果	面部拟真度	<40%	≥40%	≥60%	≥80%	≥90%
	视觉精细度	<40%	≥40%	≥60%	≥80%	≥90%
	基础表情数量	<52	≥52	≥100	≥200	≥600
	唇动效果	<6	≥6	≥9	≥12	≥15
	文字转语音准确率	<40%	≥40%	≥60%	≥80%	≥99%
	语音自然度	1	2	3	4	5
	基础肢体动作数量	<76	≥76	≥116	≥156	≥196
	肢体动作自然度	<40%	≥40%	≥60%	≥80%	≥90%
	帧率	<25FPS	≥25FPS	≥50FPS	≥50FPS	≥120FPS
	分辨率	<1920*1080 /2048* 1080,	≥1920*1080 /2048* 1080	≥4K	≥8K	≥8K

2. 3D真人形象数字人识别感知分级要求

- a) 1级识别感知, 人脸识别误识率<40%, 语音转文字准确率<40%, 情绪识别准确率<40%。
- b) 2级识别感知, 人脸识别误识率 $\geq 40\%$, 语音转文字准确率 $\geq 40\%$, 情绪识别准确率 $\geq 40\%$ 。
- c) 3级识别感知, 人脸识别误识率 $\geq 60\%$, 语音转文字准确率 $\geq 60\%$, 情绪识别准确率 $\geq 60\%$ 。
- d) 4级识别感知, 人脸识别误识率 $\geq 80\%$, 语音转文字准确率 $\geq 80\%$, 情绪识别准确率 $\geq 80\%$ 。
- e) 5级识别感知, 人脸识别误识率 $\geq 99\%$, 语音转文字准确率 $\geq 99\%$, 情绪识别准确率 $\geq 99\%$ 。

表3 数字人识别感知分级要求

一级指标	二级指标	分级				
		1级	2级	3级	4级	5级
识别感知	人脸识别误识率	<40%	$\geq 40\%$	$\geq 60\%$	$\geq 80\%$	$\geq 99\%$
	语音转文字准确率	<40%	$\geq 40\%$	$\geq 60\%$	$\geq 80\%$	$\geq 99\%$
	情绪识别准确率	<40%	$\geq 40\%$	$\geq 60\%$	$\geq 80\%$	$\geq 99\%$

3. 3D真人形象数字人交互决策分级要求

- a) 1级交互决策, 对话交互完成率<40%, 表情反馈正确率<40%, 肢体反馈正确率<40%, 对话、表情、肢体反馈的一致性<40%, 平均卡顿时长>10s。
- b) 2级交互决策, 对话交互完成率 $\geq 40\%$, 表情反馈正确率 $\geq 40\%$, 肢体反馈正确率 $\geq 40\%$, 对话、表情、肢体反馈的一致性 $\geq 40\%$, 平均卡顿时长 $\leq 10s$ 。
- c) 3级交互决策, 对话交互完成率 $\geq 60\%$, 表情反馈正确率 $\geq 60\%$, 肢体反馈正确率 $\geq 60\%$, 对话、表情、肢体反馈的一致性 $\geq 60\%$, 平均卡顿时长 $\leq 6s$ 。
- d) 4级交互决策, 对话交互完成率 $\geq 80\%$, 表情反馈正确率 $\geq 80\%$, 肢体反馈正确率 $\geq 80\%$, 对话、表情、肢体反馈的一致性 $\geq 80\%$, 平均卡顿时长 $\leq 3s$ 。
- e) 5级交互决策, 对话交互完成率, 表情反馈正确率, 肢体反馈正确率, 对话、表情、肢体反馈的一致性 $\geq 90\%$, 平均卡顿时长 $\leq 1s$ 。

表4 数字人交互决策分级要求

一级指标	二级指标	分级				
		1级	2级	3级	4级	5级
交互决策	对话交互完成率	<40%	$\geq 40\%$	$\geq 60\%$	$\geq 80\%$	$\geq 99\%$
	表情反馈正确率	<40%	$\geq 40\%$	$\geq 60\%$	$\geq 80\%$	$\geq 99\%$
	肢体反馈正确率	<40%	$\geq 40\%$	$\geq 60\%$	$\geq 80\%$	$\geq 99\%$
	对话、表情、肢体反馈的一致性	<40%	$\geq 40\%$	$\geq 60\%$	$\geq 80\%$	$\geq 90\%$
	平均卡顿时长	>10	≤ 10	≤ 6	≤ 3	≤ 1

6.4 3D真人形象数字人总体分级标准

数字人系统按照其应用场景和具体需求, 其包含的评测指标也会有差异。具体指标的要求如下。

表 5 数字人分级标准

一级指标	二级指标	总体分级				
		1 级	2 级	3 级	4 级	5 级
人物效果	面部拟真度	●	●	●	●	●
	视觉精细度	●	●	●	●	●
	基础表情数量	●	●	●	●	●
	唇动效果	●	●	●	●	●
	文字转语音准确率	●	●	●	●	●
	语音自然度	●	●	●	●	●
	基础肢体动作数量	●	●	●	●	●
	肢体动作自然度	●	●	●	●	●
	帧率	●	●	●	●	●
	分辨率	●	●	●	●	●
识别感知	人脸识别误识率	○	○	○	●	●
	语音转文字准确率	○	○	○	●	●
	情绪识别准确率	○	○	○	●	●
交互决策	对话交互完成率	○	○	○	●	●
	表情反馈正确率	○	○	○	●	●
	肢体反馈正确率	○	○	○	●	●
	对话、表情、肢体反馈的一致性	○	○	○	●	●
	平均卡顿时长	○	○	○	●	●

注：“●”表示必要指标项，“○”表示可选指标项

1. 1级3D真人形象数字人

应满足：

- 3D真人形象数字人形象效果一般，面部拟真度一般。
- 3D真人形象数字人表情效果一般，表情数量少，唇动效果简单开合，语音效果一般。
- 3D真人形象数字人动作效果一般，基础动作少，肢体动作一般。
- 3D真人形象数字人呈现效果一般，画面流畅。

2. 2级3D真人形象数字人

应满足：

- 3D真人形象数字人形象效果较好，具有写实的面部拟真度和视觉精细度。
- 3D真人形象数字人表情效果较好，表情较为丰富，唇动效果较好，语音效果较好。
- 3D真人形象数字人动作效果较好，基础动作较丰富，肢体动作较较好。
- 3D真人形象数字人呈现效果较好，画面清晰。

3. 3级3D真人形象数字人

应满足：

- a) 3D真人形象数字人形象效果良好，具有超写实的面部拟真度和视觉精细度。
- b) 3D真人形象数字人表情效果良好，表情丰富，唇动效果良好，语音效果良好。
- c) 3D真人形象数字人动作效果良好，基础动作丰富，肢体动作良好。
- d) 3D真人形象数字人呈现效果良好，画面高清。

4. 4级3D真人形象数字人

应满足：

- a) 3D真人形象数字人形象效果优秀，具有超写实的面部拟真度和视觉精细度。
- b) 3D真人形象数字人表情效果优秀，表情丰富，唇动效果优秀，语音效果优秀。
- c) 3D真人形象数字人动作效果优秀，基础动作丰富，肢体动作优秀。
- d) 3D真人形象数字人呈现效果优秀，画面高清。
- e) 3D真人形象数字人具有人脸识别、语音识别、情绪识别功能中的一种或多种，其识别准确率优秀。
- f) 3D真人形象数字人具有对话交互、表情反馈、肢体反馈功能中的一种或多种，正确率优秀，对话、表情、肢体反馈一致性优秀，卡顿低。

5. 5级3D真人形象数字人

应满足：

- a) 3D真人形象数字人形象效果优异，具有超写实的面部拟真度和视觉精细度。
- b) 3D真人形象数字人表情效果优异，表情丰富，唇动效果优异，语音效果准确优异。
- c) 3D真人形象数字人动作效果优异，基础动作丰富，肢体动作优异。
- d) 3D真人形象数字人呈现效果优异，画面超高清。
- e) 3D真人形象数字人具有人脸识别、语音识别、情绪识别功能中的一种或多种，其识别准确率优异。
- f) 3D真人形象数字人具有对话交互、表情反馈、肢体反馈功能中的一种或多种，正确率优异，对话、表情、肢体反馈一致性优异，无卡顿。

（注：在本节中，相关文字描述中与MOS主观描述规则一致，“一般”=1级，“较好”=2级，“良好”=3级，“优秀”=4级，“优异”=5级）