ICS：33.160
CCS：M61

UHD World Association
世界超高清视频产业联盟

# 世 界 超 高 清 视 频 产 业 联 盟 标 准

T/UWA xxxx-xxxx

# 3D 数字人质量分级技术要求

Technical requirements for grading the quality of 3-Dimensional digital human

（ 报批稿 ）

xxxx-xx-xx 发布　　xxxx-xx-xx 实施

世界超高清视频产业联盟 发布

# 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 数字人（3-Dimensional digital human） ..... 4
3.2 拟真度（fidelity） ..... 4
3.3 精细度（detail） ..... 4
3.4 唇动效果（lip sync） ..... 4
3.5 自然度（naturalness） ..... 4
3.6 帧率（frame rate） ..... 5
3.7 分辨率（resolution） ..... 5
4 缩略语 ..... 5
5 分级参数 ..... 5
5.1 人物效果 ..... 6
5.2 识别和感知 ..... 7
5.3 交互和决策 ..... 7
6 3D 真人形象数字人分级 ..... 8
6.1 3D 真人形象数字人分级原则 ..... 8
6.2 3D 真人形象数字人细分指标分级标准 ..... 8
6.3 3D 真人形象数字人总体分级标准 ..... 10
附录 A ..... 11
3D 真人形象数字人分级参数计算方法建议 ..... 11

# 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：中国移动通信集团有限公司、咪咕文化科技有限公司、北京清博智能科技有限公司、深圳思谋信息科技有限公司、中国电子技术标准化研究院、中国信息通信研究院、北京百度网讯科技有限公司、上海数字电视国家工程研究中心有限公司、OPPO 广东移动通信有限公司、深圳市洲明科技股份有限公司、京东方科技集团股份有限公司、聚好看科技股份有限公司、中兴通讯股份有限公司、中央广播电视总台、凌云光技术股份有限公司、华为技术有限公司、深圳市腾讯计算机系统有限公司、深圳市奥拓电子股份有限公司、北京三星通信技术研究有限公司、山东浪潮超高清智能科技有限公司。

本文件主要起草人：李琳、单华琦、毕蕾、高山、王雷、朱泓、李锦枝、向安玲、李亭竹、张亚男、赵轶、刘志杰、李婧欣、耿一丹、傅蓉蓉、刘毓伟、许闻苑、查丽、殷惠清、史梦蕾、来航曼、康峰、白莹洁、谭胜淋、陈于思、杨智远、朱家林、李秋婷、黄成、王子建、谭阳、范晓轩、李丹、熊伟、曾义、陈曦、胡颖、吴未、孙信中、王立众、吴越、王培元、王宗增。

## 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方法。

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

## 3 术语和定义

下列术语以及定义适用于本文件。

### 3.1 3D数字人（3-Dimensional digital human）

利用3D技术在数字世界创建的具有人类外貌、语音、动作和交互能力的数字人物。

注1：3D数字人可以通过语音、姿态、面部表情等多种方式与用户进行交互，具有智能问答、情感识别、语音合成等功能，能够模拟真实人类的行为和反应。

注2：3D数字人可以在虚拟现实、增强现实、视频游戏、娱乐和教育等领域中应用。

注3：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。

注4：3D数字人的应用领域越来越广泛，可以用于虚拟演员、虚拟营销、虚拟导游、虚拟医生、虚拟教师等多个领域。

### 3.2 拟真度（fidelity）

用于评价虚拟世界中的计算机图像、3D模型、动画或游戏的真实感和逼真程度的度量指标。它反映了数字世界与现实世界之间的相似程度。

### 3.3 精细度（detail）

指的是图像、模型或场景中细节的数量和质量。在计算机图形学和计算机视觉领域，精细度是衡量渲染质量和模型准确性的重要指标之一。

### 3.4 唇动效果（lip sync）

是指在动画和特效制作中，通过模拟和表现人物的嘴唇运动和发音过程，使其与配音内容同步，增强人物形象的真实感和表达能力。

### 3.5 自然度（naturalness）

自然度是指一个言语或行为的表现是否符合自然、真实、真诚的程度。在人工智能领域，自然度可以用来评估机器人或语音助手的交流能力和人机交互的质量。一个具有高自然度的智能机器人能够以流畅、自然的方式与用户交流，使用适当的语言和表情回答问题，给予用户满意的体验。自然度不仅仅涉及语言和语音的准确性，还包括机器人的情感表达能力、语调和语速的调节、上下文的理解和适当的回

应等方面。

### 3.6 帧率（frame rate）

是指在一秒钟内显示的图像帧数,是衡量图像流畅度和流畅度的重要指标。

### 3.7 分辨率（resolution）

分辨率是指图像或视频中可显示的细节数量，通常以像素为单位表示。

### 3.8 面部动画参数 (FAP)

是指控制面部的关键特征点，这些特征点用于生成动画视位和面部表情，以及头部和眼睛的运动。这些特征点是 MPEG-4 定义的面部定义参数的一部分[1]。FAP 表示特征点相对于中性面部位置的 66 个位移和旋转，其定义为嘴巴闭合、眼睑与虹膜相切、视线和头部方向正前方、牙齿接触、舌头接触牙齿。这些 FAP 被设计为与人类面部肌肉运动 密切相关。

### 3.9 身体动画参数 (BAP)

是指控制身体关节的关键特征点。这些特征点是 MPEG-4 定义的面部定义参数的一部分[1]。MPEG-4 定义了 168 个身体动画参数，描述几乎所有可能的身体姿势，其中 12 个参数描述了每个手臂的运动，而 29 个参数描述了每个手的运动。

## 4 缩略语

下列缩略语适用于本文件。

3D 三维（3-Dimensional）

FAP 面部动画参数Face Animation Parameter

BAP 身体动画参数Body Animation Parameter

FPS 画面每秒传输帧数Frame Per Second

## 5 分级参数

随着计算机图形和动画技术的发展，3D 数字人类已经广泛应用于电影、游戏和虚拟现实等行业。然而，数字人类的呈现和交互质量存在较大差异，目前缺乏普遍的评估标准来评估其体验效果。因此，有必要开发综合评估数字人类效果的方法,以确保其高质量和可用性,在促进数字人类产业发展的同时，增强各种应用中的用户体验。

[1] MPEG-4 Facial Animation: The Standard, Implementation and Applications. Wiley. pp. 17 – 55. ISBN 978-0-470-84465-6.


图 1 数字人基于用户体验质量的三个维度

基于用户体验，可以分为三个维度的指标参数。角色效果：用户能察觉的数字人形特征，目前主要包括视觉和听觉方面；识别感知：数字人形识别和察觉用户和外部环境输入信息的能力，例如语音转文字准确率、人脸识别率、情绪识别准确率等；互动决策：数字人形与用户“自主”互动的能力，例如对话交互完成率、表情反馈正确率、肢体反馈正确率等指标。

## 5.1 人物效果

### 1. 面部拟真度

用来表征3D真人形象数字人面部拟人化的程度。指3D数字人或虚拟角色面部表情、结构、动作等人形特征与真实人类相似的程度。范围是1-100%，数值越大相似程度越高。

### 2. 视觉精细度

用来表征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。范围是1-100%，数值越大，视觉内容越丰富，精细度越高。

### 3. 面目动态效果

用来表征3D真人形象数字人可呈现的表情丰富度，用面部动画参数个数量化，个数越多，可以表达的表情越丰富，数值为是0~66个。

### 4. 唇动效果

用来表征3D数字人说话时嘴唇动态效果。着重分析嘴唇运动和发音过程，使其与配音内容同步，增强人物形象的真实感和表达能力。范围是1-100%，数值越大，唇动效果越好。

### 5. 文字转语音准确率

用来表征3D真人形象数字人语音合成并播放时的准确率。范围是1-100%，数值越大，准确度越高。计算方法如公式（1）所示：

$$文字转语音准确率 = \frac{符合条件的测定值个数}{总测定值个数} \times 100\% \quad \cdots\cdots\cdots\cdots\cdots\cdots（1）$$

**6. 语音自然度**

用来表征3D数字人语音合成或真人发音的感知自然度。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》的5.2.3章节内容进行评测，评测方法是主观评测，取值是1~5，其中5是最优。

**7. 肢体动作效果**

用来表征3D真数字人动作丰富度，用身体动画参数个数来度量，个数越多则动作丰富度越高。取值范围为0~168。

**8. 组合肢体动作自然度**

用来表征3D数字人组合肢体动作的自然流畅度。范围是1-100%，数值越大，唇动效果越好。

**9. 帧率**

用来表征3D真人形象数字呈现图像的流畅度。以每秒帧数（FPS）为单位来表示。

**10. 分辨率**

用来表征3D真人形象数字呈现图像的细节梳理。以像素为单位表示,图像宽度和高度的像素点个数。

5.2 识别和感知

**1. 语音识别准确性**

指3D数字人对用户进行语音识别的性能表现。其性能表现依据系统中语音识别用途的不同，分别按照 GB/T 21023-2007 中的5.2.1、5.2.2、5.2.3进行评估。

**2. 人脸识别率**

指3D数字人对用户进行语音识别的性能表现。用3D数字人在进行人脸识别的过程中正确识别人脸的概率表示，范围为0-100%。

**3. 情绪识别准确率**

用来表征3D数字人对用户在积极、消极、中立三类情绪识别的准确率，计算方法为正确识别用户对话情绪样本数占总样本数的概率，范围为0-100%。

5.3 交互和决策

**1. 对话交互完成率**

用来表征3D数字人和用户对话的能力。计算方法为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，范围0-100%。

**2. 表情反馈正确率**

用来表征3D数字人和用户表情交互过程中能正确理解用户表情并反馈正确表情的能力。计算方法为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,范围0-100%。

**3. 动作反馈正确率**

用来表征3D数字人和用户表情交互过程中能正确理解用户动作并反馈正确肢体动作的能力。计算方法为用户对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，范围0-100%。

**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。计算方法如公式（2）所示：

$$S = F \times (1 - \frac{EN}{T}) \times 100\% \quad \text{.....（2）}$$

**5. 平均卡顿时长**

用来考核3D数字人和用户交互的流畅程度，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、语音卡顿等认为不流畅。计算方法为在标准评测时间内，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。计算方法如公式（3）所示，单位为秒/次：

$$\text{平均卡均卡顿} = \frac{\sum(ti)}{\text{卡顿总次数}} \quad \text{.....（3）}$$

## 6 3D 数字人分级

### 6.1 3D 数字人分级原则

随着数字人技术的不断发展，越来越多的应用场景需要数字人具备高度的真实感和交互性。因此，对3D数字人视觉与交互效果的评估成为了数字人技术发展的重要一环。通过建立统一的评估标准和方法，可以规范数字人行业，提高数字人质量，推动数字人技术在虚拟现实、游戏、教育、医疗、娱乐等领域的广泛应用。

本章基于前面 3D 数字人指标的详细说明，以及附录提供的计算方法建议，对 3D 数字人进行分级。在从总分级之前，先完成 3D 数字人细分指标人物效果、识别感知、交互决策三个维度的分级，这三个维度分别涵盖了数字人的不同方面，可以提供更详细和具体的评估结果。通过将数字人分成这三个维度并进行分级，可以更准确地评估数字人在不同方面的表现和能力，帮助用户选择适合自己需求的数字人，并提供有针对性的反馈和建议，提高用户体验。

### 6.2 3D 数字人细分指标分级标准

**1. 3D数字人人物效果分级要求**

人物效果主要评估的是数字人的外观和表现。这包括外貌、语音、动作等方面，以及与真实人物的相似度和真实感。人物效果的好坏会影响用户对数字人的接受程度。主要是应用在影视制作、广告宣传、数字人播报等非交互场景数字人的评级。分级具体要求见表 1：

表 1 数字人人物效果分级要求

| 一级指标 | 二级指标 | 分级（范围） | | | | |
|---|---|---|---|---|---|---|
| | | 1 级 | 2 级 | 3 级 | 4 级 | 5 级 |
| 人物效果 | 面部拟真度 | 0-39% | 40%-59% | 60%-79% | 80%-89% | 90%-100% |
| | 视觉精细度 | 0-39% | 40%-59% | 60%-79% | 80%-89% | 90%-100% |
| | 基础表情数量 | 0-29 | 30-39 | 40-49 | 50-65 | 66 |
| | 唇动效果 | 0-5 | 6-8 | 9-11 | 12-14 | 15 |
| | 文字转语音准确率 | 0-39% | 40%-59% | 60%-79% | 80%-98% | 99%-100% |
| | 语音自然度 | 1 | 2 | 3 | 4 | 5 |
| | 基础肢体动作数量 | 0-75 | 76-115 | 116-155 | 156-195 | 196 |
| | 肢体动作自然度 | 0-39% | 40%-59% | 60%-79% | 80%-89% | 90%-100% |
| | 帧率（FPS） | 0-24 | 25-39 | 40-49 | 50-119 | 120 以上 |
| | 分辨率 | 不足 1920*1080/2048* 1080, | 1920*1080 /2048* 1080 | 4K | 8K | 8K 以上 |

**2. 3D真人形象数字人识别感知分级要求**

识别感知是指数字人对周围环境和用户的感知能力。它包括语音识别、图像识别、情感识别等技术，在与用户交互的过程中，能够准确地理解用户的需求和指令。主要是应用在基于环境识别感知场景的数字人、如智能输入助手、智能家居等。分级具体要求见表 2：

表 2　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 |
| 识别和感知 | 人脸识别误识率 | 0-39% | 40%-59% | 60%-79% | 80%-98% | 99%-100% |
| | 语音转文字准确率 | 0-39% | 40%-59% | 60%-79% | 80%-98% | 99%-100% |
| | 情绪识别准确率 | 0-39% | 40%-59% | 60%-79% | 80%-98% | 99%-100% |

**3. 3D真人形象数字人交互决策分级要求**

交互决策是指数字人在特定情境下根据用户需求做出的反应和决策能力。这包括从用户提供的信息中进行分析和推理，做出适当的回应或提供合适的建议等，主要用户智能客服、数字员工等场景。分级具体要求见表 3：

表 3　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 |
| 交互和决策 | 对话交互完成率 | 0-39% | 40%-59% | 60%-79% | 80%-98% | 99%-100% |

| | | | | | | |
|---|---|---|---|---|---|---|
| | 表情反馈正确率 | 0-39% | 40%-59% | 60%-79% | 80%-98% | 99%-100% |
| | 肢体反馈正确率 | 0-39% | 40%-59% | 60%-79% | 80%-98% | 99%-100% |
| | 对话、表情、肢体反馈的一致性 | 0-39% | 40%-59% | 60%-79% | 80%-89% | 90%-100% |
| | 平均卡顿时长（秒） | 大于 10 | 7-10 | **4**-6 | 2-3 | 0-1 |

## 6.3 3D 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 4 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 | ● | ● | ● | ● | ● |
| 识别和感知 | 人脸识别误识率 | ○ | ○ | ○ | ● | ● |
| | 语音转文字准确率 | ○ | ○ | ○ | ● | ● |
| | 情绪识别准确率 | ○ | ○ | ○ | ● | ● |
| 交互和决策 | 对话交互完成率 | ○ | ○ | ○ | ● | ● |
| | 表情反馈正确率 | ○ | ○ | ○ | ● | ● |
| | 动作反馈正确率 | ○ | ○ | ○ | ● | ● |
| | 对话、表情、动作反馈的一致性 | ○ | ○ | ○ | ● | ● |
| | 平均卡顿时长 | ○ | ○ | ○ | ● | ● |

注：“●”表示必要指标项，“○”表示可选指标项

## 附录 A

（资料性附录）

## 3D 数字人分级参数计算方法建议

数字人分级指标计算方法建议见下表：

附表 A.1　数字人分级技术要求

| 一级指标 | 二级指标 | 细分指标 | 测算方式和参数说明（做到附录） |
|---|---|---|---|
| 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[2] 人脸数据训练的 StyleGan[3] 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*100% |
| | 表情效果 | 面部动态丰富度 | 参考 MPEG-4 Facial Animation: The Standard, Implementation and Applications. Wiley. pp. 17–55. ISBN 978-0-470-84465-6. |
| | | 唇动效果 | 对数字人说话的视频片段进行评测，用打点计数的方式来统计，N 初始值为 0，唇动效果不好的字则 N=N+1，根据统计结果计算准确率=(总字数-N)/总字数*100% |
| | 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，点计数的方式来统计，N 初始值为 0，发音不准确、发音错误的字记为 N=N+1，准确率=(总字数-N)/总字数 |

[2] http://vis-www.cs.umass.edu/lfw/
[3] https://github.com/NVlabs/stylegan

| | | | |
|---|---|---|---|
| | | | *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 | 肢体动作丰富度 | 参考 MPEG-4 Facial Animation: The Standard, Implementation and Applications. Wiley. pp. 17–55. ISBN 978-0-470-84465-6. |
| | | 肢体动作自然度 | 抽取 N 张数字人图片 |
| | | | ①使用 3D 人体估计方法，基于 Human3.6M[4]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 |
| | | | 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% |
| | 呈现效果 | 帧率 | GY/T 307-2017 超高清晰度电视系统节目制作和交换参数值中的 3.超高清晰度电视节目制作基本参数<br>T/UWA 012.6-2022 “百城千屏”超高清视音频传播系统节目播出技术要求中的 6.超高清播出节目视频信号技术要求 |
| | | 分辨率 | |
| 识别和感知 | 人脸识别 | 人脸识别率 | 基于基准的 LFW[2] 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 | **语音识别准确性** | 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) applications |
| | 情绪识别 | 情绪识别准确率 | ①基于多语言情感分析算法和 Universal Dependency Treebanks 数据库[5]，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 |

[4] http://vision.imar.ro/human3.6m/description.php
[5] https://universaldependencies.org

| | | | |
|---|---|---|---|
| | | | ②给 k(k>=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<br>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% |
| 交互和决策 | 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 Challenge[6] 11 种评测指标和 3 种评测时机，基于语料库测试出基础值并分箱<br>② 给 k(k>=10)位有专业背景知识的测试人员进一步分箱判别<br>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通过基于面部动作编码系统，例如 FACS[7]给出表情识别与分类<br>②给 k(k>=10)位有专业背景知识的测试人员进一步判别<br>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 | 动作反馈正确率 | ①通过微软提出的单个深度图像中的实时人体姿势识别[8]方法给出判别分箱<br>②给 k(k>=10)位有专业背景知识的测试人员进一步分箱判别<br>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%)，其中 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] https://paperswithcode.com/dataset/dialogue-state-tracking-challenge
[7] https://web.cs.wpi.edu/~matt/courses/cs563/talks/face_anim/ekman.html
[8] https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/BodyPartRecognition.pdf