G/SZKJ 0004 - 2024

《新媒体智能视频生成平台评估导则》

Guidelines for Evaluating Intelligent Video Generation Platforms of New Media

2024-11-13 发布

工业和信息化部工业文化发展中心

目 次

前 言

伴随 5G 移动通信技术的普及，视频迅速成为风靡全球的新媒体型态和主流传播方式。随着人工智能技术在媒体领域应用的不断深化，能够自动化地生成高质量视频内容、满足用户个性化和创意表达等需求的新媒体智能视频生成平台也应运而生。由于技术方向和成本等原因，当前市场上的新媒体智能视频生成平台良莠不齐，同时，作为新生事物，也缺乏相关技术和应用标准规范来对视频生成平台进行评估。为了规范行业发展，提升用户体验，促进文化繁荣，本文件旨在制定一套全面的新媒体智能视频生成平台评价的行业应用标准。

《导则》全面覆盖了新媒体智能视频生成平台从使用到产出的各个环节，重点评估了平台的易用性、安全性、视频画面与音效质量、内容逻辑性、多模态输入输出以及交互体验等关键维度。这些评估标准不仅有助于推动平台的持续优化升级，还能为政策制定者、企业决策者和技术开发者提供宝贵的参考和指导。

《导则》特别强调评估工作的基础性和指导原则，确保评估过程的公正性、有效性和实用性。其核心原则包括科学性、引导性、实效性、可操作性、可扩展性、持续性和公正性，旨在为评估活动奠定坚实基础，推动新媒体智能视频平台的健康发展，并引领技术革新与实践应用朝着满足社会需求、追求经济效益以及遵循伦理道德的方向迈进。

通过实施《导则》，期望能够有效约束并引导新媒体智能视频生成平台的技术进步与产品力提升，为行业的健康、有序发展提供坚实保障，推动科技与社会的和谐共赢。

《新媒体智能视频生成平台评估导则》

1 范围

本导则规定了新媒体智能视频生成平台的参考框架，并描述了视频、声音、时空逻辑一致性、多模态输入输出、交互、安全等维度的评测指标。

本导则适用于指导第三方测评机构对新媒体智能视频生成平台服务功能的评估、验收等工作。

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 21024-2007 中文语音合成系统通用技术规范

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、定义和缩略语

下列术语、定义和缩略语适用于本文件。

3.1 视频生成

指基于用户输入内容，利用计算机完成声音、图像的统一制作并输出的工作流程。

3.2 韵律

一般也称为超音段特征（suprasegmental feature），是将各种语言学单位组织成话语或话语中关联组块的系统组织。从物理角度，指基频、时长和强度等声学参数；从语言学的角度，指音段层次以上的音系组织。

注：韵律的实现涉及语音的音段和超音段特征，它不但能够传递语言学信息，而且能够传递副语言学和非语言学信息。

[来源：GB/T21024-2007，定义 3.18]

4 新媒体智能视频生成平台系统参考框架

声音模块：提供视频的声音。包括但不限于符合视频逻辑的人声、物体声、背景音乐等。

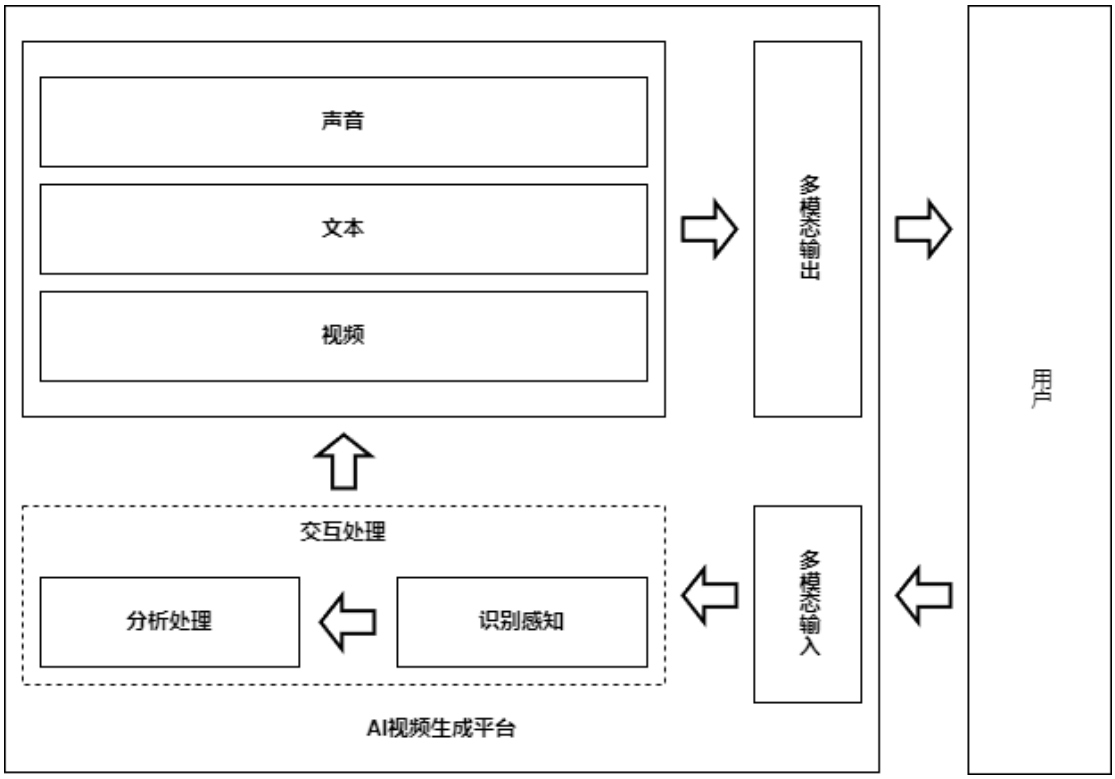
文本模块：提供视频的文本。包括但不限于视频画面物体中的文本、字幕等。

视频模块：生成符合用户要求的视频。包括但不限于符合空间逻辑与时间逻辑的连贯视频。

交互处理模块：通过语音语义图像识别等技术识别用户输入信息，理解用户意图，指导视频生产。

多模态输入：用于接收用户的输入信息，包括但不限于文本、语音、图片、视频等。

多模态输出：用于将输出的结果呈现给用户。



5 指标及规范性描述

5.1 视频

5.1.1 视频分辨率

指系统生成视频的像素大小，决定了视频的清晰度。行业常用的视频分辨率由低到高分别有：480P（640×480）、720P（1280×720）、1080P（1920×1080）、4K（3840x2160）、8K（7680x4320）。

5.1.2 视频流畅性

指系统生成视频的连贯性和流畅度，主要通过视频帧率，即 FPS 值（单位：帧/秒）来评估。帧率越高，视频的流畅性就越好。标准的视频帧率通常为 24fps、25fps、30fps、60fps、120fps。

5.1.3 画面准确度

用于考核系统生成固定帧数视频时画面的准确率，若出现跳帧、卡顿等错误均视为画面不准确。

5.1.4 剪辑与转场

指系统对生成视频的处理效果，剪辑指进行剪切、调整，以达到最佳的表现效果。转场指视频场景之间过渡的艺术处理。合理的剪辑和转场能大幅提升视频表现力及观感。该指标为主观性评估指标。

用户根据观看视频的剪辑及转场质量，在李克特量表中，给出一个主观评分评价质量优劣，1 最差，5 最优，具体评分规则见下表。

评测维度	描述	5	4	3	2	1
转场流畅	场景切换流畅自然，不会让观众感到突兀或跳跃。	非常流畅	比较流畅	基本流畅	不太流畅	完全不流畅
节奏合理	根据视频内容和情绪氛围，控制镜头的长短、切换速度等，形成良好的视觉节奏。	非常合理	比较合理	基本合理	不太合理	完全不合理
音乐匹配	镜头切换与音乐、音效等声音元素的配合程度	非常好	比较好	一般	不太好	差
创意效果	使用不同的转场效果、剪辑技巧或创意性的镜头组合增加视频的创造力和艺术感。	非常有创意	比较有创意	一般	不太有创意	完全没创意

5.1.5 文本准确率

指系统生成视频画面内文本总数中正确文本的占比，用以衡量视频生成对文字图像的生成能力。文本错误包括但不限于未能生成正确字、前后不搭、整句错误等。其指标的计算方法如公式（1）所示。

$$R_t = \frac{N_t}{N} \times 100\% \quad \text{.....（1）}$$

式中

N_t ——正确文本数量；

N ——文本总数；

R_t ——文本准确率。

5.2 时间、空间、逻辑一致性

5.2.1 时间一致性

指系统生成视频的视频帧之间的连贯性和平滑性。该指标为客观评价指标，包括时间合理性指标、帧率稳定性指标和符合现实性指标。

时间合理性指系统生成的视频画面内容按照合理的时间顺序发生，没有明显的跳跃或倒置，其指标的计算方法如公式（2）所示。

$$R_t = 1 - \frac{f_t}{F} \quad \text{.....（2）}$$

式中

R_t ——视频中时间顺序合理帧数的占比；

f_t ——视频中时间顺序错误帧数；

F ——视频的总帧数。

帧率稳定性指系统生成的画面内容帧率稳定，避免出现卡顿或不自然的运动，其指标的计算方法如公式（3）所示。

$$T = 1 - \frac{\sigma f}{f_{avg}} \quad \text{.....（3）}$$

式中

T ——时间合理性，范围为 [0, 1]，越接近 1 表示时间顺序越合理；

σf ——帧率的标准差，反映帧率的波动情况；

f_{avg} ——视频的平均帧率。

符合现实指物体运动的速率符合现实规律，没有不自然地加速或减速，其指标的计算方法如公式（4）所示。

$$R_t = 1 - \frac{f_t}{F} \quad \text{.....(4)}$$

式中

R_t ——视频中物体运动速率合理帧数的占比；

f_t ——视频中物体运动速率失常帧数；

F ——视频的总帧数。

5.2.2 空间一致性

指系统生成视频的每个帧的空间结构、物理规律和细节的一致性，为主观评价指标。

用户根据观察视频的空间一致性表现，在李克特量表中，给出一个主观评分评价质量优劣，1 最差，5 最优，具体评分规则见下表。

评测维度	描述	5	4	3	2	1
位置合理	物体、场景和人物在空间上的位置和关系合理，没有突兀地出现或消失。	非常合理	比较合理	基本合理	不太合理	完全不合理
比例相符	物体的大小、距离和比例与现实相符，避免出现不合理的缩放或变形。	非常相符	比较相符	基本相符	不太相符	完全不相符
细节稳定	场景内各物体的边缘和纹理稳定，物体边界清晰。	非常稳定	比较稳定	基本稳定	不太稳定	完全不稳定
模糊合理	背景与前景之间的模糊度合理，有景深的变化。	非常合理	比较合理	基本合理	不太合理	完全不合理
阴影一致	光源位置、强度与视频中物体的阴影一致，尤其是变化光源条件下。	非常一致	比较一致	基本一致	不太一致	完全不一致
色彩平衡	没有突然的色温变化或色彩饱和度跳变。	非常平衡	比较平衡	基本平衡	不太平衡	完全不平衡

视角一致	对于具有多个视角的视频，不同视角下的一致性和互兼容性。	非常一致	比较一致	基本一致	不太一致	完全不一致
遮挡合理	物体之间的遮挡、透视关系合理。	非常合理	比较合理	基本合理	不太合理	完全不合理

5.2.3 逻辑一致性

指系统生成视频的逻辑性和合理性。该指标为主观性评估指标。

用户根据观察视频的逻辑一致性表现，在李克特量表中，给出一个主观评分评价质量优劣，1 最差，5 最优，具体评分规则见下表。

评测维度	描述	5	4	3	2	1
符合逻辑	情节和故事发展符合逻辑，没有矛盾或不合理的情节转折。	非常符合	比较符合	基本符合	不太符合	完全不符合
适应情境	角色的行为模式、语言与故事情境相适应。	非常适应	比较适应	基本适应	不太适应	完全不适应
符合常识	事件的因果关系符合常识和逻辑推理。	非常符合	比较符合	基本符合	不太符合	完全不符合
符合规律	符合现实世界的逻辑规律，例如重力、速度、碰撞反应等。	非常符合	比较符合	基本符合	不太符合	完全不符合

5.3 声音

5.3.1 总体质量

指系统生成视频的声音整体质量。若出现下列情况中任意一种或几种视为有瑕疵。

——音质失真

——音量过大、过小或不稳定

——音质不清晰

——背景噪音

——语音合成质量不佳

5.3.2 音效、环境音

指系统生成视频的物体音效、环境音等质量。该指标为主观性评估指标。

用户聆听物体音效、环境音的感受，在李克特量表中，给出一个主观评分评价质量优劣，1 最差，5 最优，具体评分规则见下表。

评测维度	描述	5	4	3	2	1
音质清晰	音乐音质清晰高质量，无杂音、无失真。	非常清晰	比较清晰	一般清晰	不太清晰	完全不清晰
混音效果	物体音效、人声之间处理和谐，不存在互相干扰或覆盖的情况。	非常和谐	比较和谐	基本和谐	不太和谐	完全不和谐
生成创意	视频音效有创意，具有新颖性、原创性。	非常有创意	比较有创意	一般	不太有创意	完全没创意

声画匹配	物体音、环境音与视频画面相匹配。	非常匹配	比较匹配	基本匹配	不太匹配	完全不匹配

5.3.3 背景音

指系统生成视频背景音质量。该指标为主观性评估指标。

用户根据聆听视频背景音感受，在李克特量表中，给出一个主观评分评价质量优劣，1 最差，5 最优，具体评分规则见下表。

评测维度	描述	5	4	3	2	1
声画协调	音乐的节奏、风格、情感等方面与视频主题和画面相协调。	非常适配	比较适配	基本适配	不太适配	完全不适配
连贯自然	音乐的转换自然流畅，避免突然切换造成观众的不适。	非常自然	比较自然	基本自然	不太自然	完全不自然
艺术表现	背景音乐能够增强视频的艺术表现力，使视频更加吸睛。	非常好	比较好	一般	不太好	不好

5.3.4 发音准确度

指视频生成中人声的发音准确程度。发音不准确的表现包括漏音吞音、多余发音、音素错误、音调错误等，相应的性能指标包括发音字准确率和发音句准确率，计算方法如公式（5）、（6）所示。

$$R_{wc} = (1 - \frac{N_{ew}}{N_w}) \times 100\% \quad \text{.....（5）}$$

式中：

N_w ——文本总字数，单位为个；

N_{ew} ——发音错误字数（多种发音错误字数之和），单位为个；

R_{wc} ——发音字准确率。

$$R_{sc} = (1 - \frac{N_{es}}{N_s}) \times 100\% \quad \text{.....（6）}$$

式中：

N_s ——文本总句数，单位为个；

N_{es} ——发音错误句数，单位为个；

R_{sc} ——发音句准确率。

5.3.5 韵律准确度

指视频生成中人声的韵律准确程度。韵律包括停顿断句、音高、音长、音量、重音位置、焦点位置等因素，对应了焦点发音、问句语调、感叹句语调等自然发音规律，此处只考察停顿断句，具体可参考 ACL“黄金”标准分词文件。其指标的计算方法如公式（7）所示。

$$R_{pc} = \frac{N_{pc}}{N} \times 100\% \quad \text{.....（7）}$$

式中：

N_{pc} ——停顿正确用例数，单位为个；

N ——总用例数，单位为个；

R_{pc} ——韵律准确率。

5.3.6 拟真度

指视频生成中人声的真实程度。该评价指标为主观评价，采用 MOS（Mean Opinion Score，平均得分法）对人声拟真度进行评价，分值为 1-5 分，分值越高，人声拟真度和自然度也越高。如果平均主观评价值 MOS 是 4 或者更高，被认为是比较好的语音质量，而若平均 MOS 低于 3.6，则表示大部分接听者对这个语音质量不满。

等级	分数	听感
优	4.0~5.0	很好，十分真实自然。
良	3.5~4.0	稍差，较为真实自然。
中	3.0~3.5	还可以，听得过去。
差	1.5~3.0	勉强，不太真实自然。
劣	0~1.5	极差，一点也不真实。

5.4 多模态输入

用于考核系统支持的输入方式种类，包括但不限于语音、文字、图像、触控等。

5.5 多模态输出

5.5.1 视频合成实时率

指系统的视频合成实时率，即视频合成耗时与视频输出时长的比值。当视频帧数、分辨率等参数不同时，视频合成实时率会产生较大变化，在评价时应按照同等输出格式进行比较，计算方法如公式（8）所示。

$$R_v = \frac{T_m}{T} \times 100\% \quad \text{.....(8)}$$

式中：

T_m ——视频合成耗时；

T ——视频输出时长；

R_v ——视频合成实时率。

5.5.2 音视频匹配度

用于考核系统生成固定时长（单位：秒）视频时音视频的匹配度，若出现画面多余、缺失、音频提前、延迟等错位均视为音视频不匹配。

5.5.3 多模态输出方式

用于考核系统支持的输出方式种类，包括但不限于手机、电视、投影、LED 显示、裸眼立体、VR、AR 显示等。

5.6 交互

5.6.1 一致性

用于考核系统视觉、功能、术语的一致性，若平台视觉和布局设计不一致、相同或相似的功能呈现方式不同、界面术语不清晰给用户带来操作或理解上的错误则认为该系统一致性不够。

5.6.2 易用性

用于考核系统的易用性，若新用户在短时间内基本掌握系统功能则被认为易用性达标。

5.6.3 用户体验

用于考核系统是否能为用户提供流畅、直观且满足用户需求的体验，若系统未能提供流畅、

直观且满足用户需求的体验则认为用户体验不友好。

5.6.4 可访问性

用于考核系统是否为不同能力的用户提供了使用上的便利，若系统未能为不同能力的用户提供使用上的便利则视为可访问性较差。

5.7 安全

5.7.1 数据隐私

用于评价系统的数据隐私是否规范，包括以下几个方面。

数据收集：评估软件收集了哪些用户数据，以及收集数据的目的和方式是否透明；

数据使用：分析软件如何使用收集的数据，是否符合用户的预期和同意；

数据共享：检查软件是否与第三方共享用户数据，共享的条件和用户的控制权；

数据安全：评价软件采取了哪些措施来保护用户数据不被未授权访问或泄露；

用户控制权：用户能否容易地管理自己的数据，例如访问、更正、删除或转移自己的数据；

合规性：系统是否遵守相关的数据保护法规，包括但不限于《数据安全法》《个人信息保护法》《信息安全技术—个人信息安全规范 GB/T 35273-2020》等。

5.7.2 内容偏见

用于考核系统生成视频内容的数据偏见，若系统生成的内容中出现包括但不限于种族、性别、外貌等歧视性内容则被视为内容偏见。

5.7.3 系统安全

用于考核系统整体安全情况，若出现以下情况则视为系统安全不合格：

——未提供有效的用户身份验证机制和访问控制；

——不能抵御拒绝服务攻击，确保服务可靠性；

——未能提供足够的审计和日志记录，以防止用户否认其操作；

——安全事故监测和响应计划不完善。

5.8 合规性

5.8.1 内容合规

用于考核系统生成视频内容的合规性，若系统生成的内容中出现违反国家法律、行业法规、市场政策以及虚假、血腥、暴力、色情等内容则视为内容不合规。

5.8.2 版权合规

用于考核平台生成内容所使用的数据、素材等是否具有合法授权，若系统生成内容时候所使用的数据、素材出现侵权情况或版权争议则视为版权不合规。

5.8.3 肖像权合规

在平台生成的内容中使用真人形象或模拟的真人肖像时，必须获得被拍摄者或其法定代表人的明确授权，避免未经同意而擅自使用他人肖像。

5.8.4 价值导向合规

用于考核平台生成的内容是否符合社会主义核心价值观导向，避免生成和传播与党和国家方针政策不一致的内容，助力构建积极健康的舆论环境。

参 考 文 献

[1] GB/T 21024-2007 中文语音合成系统通用技术规范

[2] 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] 网信办 工业和信息化部 公安部令〔2022〕12 号《互联网信息服务深度合成管理规定》

[4] 国家互联网信息办公室 中华人民共和国国家发展和改革委员会 中华人民共和国教育部 中华人民共和国科学技术部 中华人民共和国工业和信息化部 中华人民共和国公安部 国家广播电视总局令〔2023〕第 15 号 《生成式人工智能服务管理暂行办法》