ICS 点击此处添加 ICS 号
点击此处添加中国标准文献分类号

DB11

北　京　市　地　方　标　准

DB 11/ XXXXX—XXXX

信息安全 虚拟数字人安全技术要求

Security technical requirements for virtual digital human

（征求意见稿）

XXXX－XX－XX 发布　　XXXX－XX－XX 实施

北京市市场监督管理局　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。

本文件由北京市公安局提出并归口。

本文件由北京市公安局组织实施。

本文件起草单位：

本文件主要起草人：

信息安全 虚拟数字人安全技术要求

1 范围

本文件规定了虚拟数字人在制作安全、应用安全等方面的技术要求。

本文件适用于虚拟数字人系统的安全。

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 35273 信息安全技术 个人信息安全规范

GB/T 45654-2025 网络安全技术 生成式人工智能服务安全基本要求

3 术语和定义

下列术语和定义适用于本文件。

3.1

虚拟数字人 virtual digital human

通过计算机图形学、计算机视觉和语音交互及人工智能生成内容（AIGC）等技术，进行形象、声音、动作等模型训练后，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：GB/T 46483—2025，3.1]

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，可实现视听双通道多模态人机交互的系统。

3.3

显式标识 explicit label

在人工智能生成合成内容或交互场景界面中添加的，以文字、声音、图形等方式呈现并可被用户明显感知到的标识。

[来源：GB/T 45438—2025，3.3]

3.4

隐式标识 implicit label

采取技术措施在人工智能生成合成内容文件数据中添加的，不易被用户明显感知到的标识。

[来源：GB/T 45438—2025，3.4]

4 框架要求

数字人安全技术要求整体框架如图1所示，主要包括两个方面，一是虚拟数字人制作安全，包括数据安全、模型安全要求、形象制作安全三部分内容，二是虚拟数字人应用安全，主要包括交互内容安全、系统安全以及场景安全三部分内容，如图1所示。

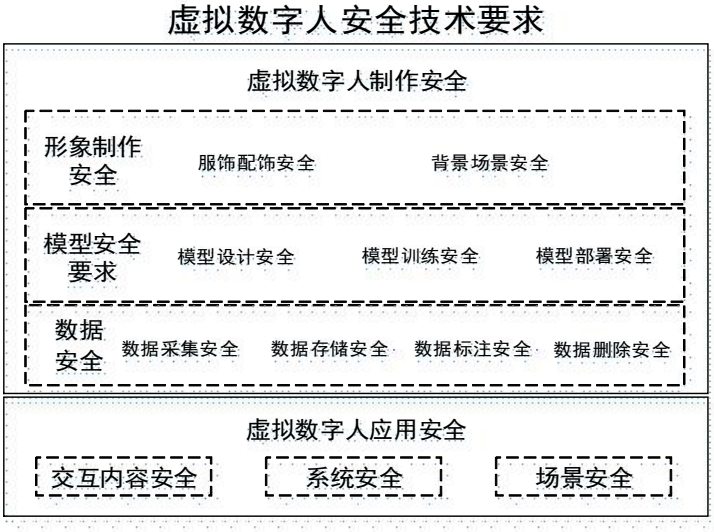

图 1 虚拟数字人安全技术要求整体框架

5 安全技术要求

5.1 虚拟数字人制作安全

5.1.1 数据安全

5.1.1.1 数据采集安全

虚拟数字人制作过程中：

a) 应在数据采集时明确虚拟数字人制作数据采集的目的、范围和方式，并告知被采集人；

b) 应在使用开源数据时遵循该数据来源的开源许可协议或取得相关授权文件；

c) 应遵循最小必要原则，只采集实现虚拟数字人功能所必需的数据；

d) 应通过数字证书、数字水印或元数据等技术手段对渲染生成的图像、视频等文件进行检测，防

止数据被窃取、篡改或产生侵权行为；

e) 应在采集用户面部、声音、动作等生物特征信息时，获得用户的明确同意，并遵守GB/T 35273-2020的要求；

f) 采集的数据不应包含GB/T 45654—2025附录 A 中A.1~A.4列出的违法不良信息。

5.1.1.2 数据存储安全

虚拟数字人使用数据：

a) 应采取加密、访问控制等技术措施，保护数据的保密性；

b) 应建立完善的数据备份和恢复机制；

c) 应根据数据的敏感程度，采取不同的存储策略；

d) 应采用加密技术对训练数据进行存储加密，保护数据在磁盘、数据库等存储介质中的静态安全；

e) 应支持数据脱敏处理，对姓名、身份证号等敏感字段进行去标识化或匿名化处理；

f) 应支持数据更新与版本管理，在数据定期更新和优化的过程中，做好新数据与原有数据的兼容性和一致性，防止因数据更新导致模型性能下降或出现安全漏洞。

5.1.1.3 数据标注安全

虚拟数字人制作过程：

a) 标注人员应具备生成合成内容、数字人、个人信息保护等相关法律法规知识、标注规则理解能力、标注平台或工具使用能力、安全风险判定能力、数据安全管理能力等；

b) 应对标注人员进行专业岗前培训，应掌握数据的标注目标、数据格式、标注方法、质量指标等内容；

c) 应对每一批标注数据进行人工检测，发现标注准确性未达到预期的，应对该批次数据重新标注；发现标注内容中包含违法不良信息的，该批次标注数据应作废。

5.1.1.4 数据删除安全

虚拟数字人在停止服务后，其相关的数据删除操作：

a) 应全面删除虚拟数字人相关数据，包括训练数据、标注数据、配置数据等；

b) 应采取“不可恢复”的方式进行删除，彻底删除数据；

c) 应记录删除过程，包括删除时间、内容、操作人、删除手段等，形成记录表；

d) 应通过专业检测工具验证数据删除效果，确认无数据残留、无法通过技术手段恢复。

5.1.2 虚拟数字人形象制作安全

5.1.2.1 服饰配饰安全

虚拟数字人服装配饰在制作过程中：

a) 服装配饰的素材来源应合法合规，服装配饰素材应具有合法的版权和质量保证；

b) 服装配饰应符合社会伦理，避免涉及种族歧视、性别歧视、宗教敏感等元素，包括但不限于虚拟服装、饰品、道具；

c) 数字虚拟人的服饰应符合其应用场景的任务设定，避免出现与场景不符的低俗或不当设计。

5.1.2.2 背景场景安全

虚拟数字人背景场景，包括但不限于各类虚拟场景的3D模型、贴图、光照参数：

a) 素材的版权应法律法规的要求，如禁止包含敏感地理标识、政治符号等；
b) 应对场景模型中的纹理贴图、关键结构等信息进行加密存储和访问控制；
c) 应记录场景素材的访问、修改、导出操作，包括操作人员、时间、操作内容等信息，支持日志检索和安全事件回溯；
d) 在背景场景更新和维护过程中，应保证数据的安全性和一致性，对更新过程进行严格的管理和监控，防止出现数据错误或安全漏洞。

5.1.3 模型安全要求

5.1.3.1 模型设计安全

虚拟数字人涉及的模型在设计期间：

a) 应进行模型安全风险评估，识别潜在的安全风险；
b) 应采用安全可靠的模型设计方法，防止模型被恶意利用。

5.1.3.2 模型训练安全

虚拟数字人涉及的模型在训练期间：

a) 应使用安全可靠的训练数据，防止恶意数据污染模型；
b) 应采取完整性校验、对抗样本防御等技术措施，防止模型被攻击或篡改；
c) 应定期进行模型的优化训练，提升模型的准确性和可靠性；
d) 应监控模型的性能，及时发现和处理异常情况；
e) 应对渲染工具、模型进行安全加固，防止恶意代码注入或漏洞利用。

5.1.3.3 模型部署安全

虚拟数字人涉及的模型在部署期间：

a) 应采取访问控制措施，防止未经授权的访问；
b) 应进行安全漏洞扫描和渗透测试，及时发现和修复安全漏洞；
c) 应监控模型的运行状态，及时发现和处理异常情况；
d) 应建立模型安全事件应急响应机制，采取一键熔断或紧急冻结等功能及时处理安全事件。

5.2 虚拟数字人应用安全

5.2.1 交互内容安全

虚拟数字人的交互内容：

a) 应建立完善的内容审核机制，交互内容应符合法律法规、伦理道德和社会规范；
b) 应采用技术手段，对交互内容过滤虚假新闻、违法不良信息，包括但不限于政治、宗教、民族

等内容；

c) 应对交互数据进行匿名化处理，去除用户的个人身份信息，保护用户的隐私；

d) 应定期更新敏感词库和内容过滤规则；

e) 应在虚拟数字人输出内容中添加清晰、明显的显式标识，如在文本首尾添加 “AI 生成”符号，在视频开场嵌入动态水印，为虚拟主播标注 “数字人” 身份等，让用户能够直观地识别出内容是由数字人生成；

f) 应保证显式标识的位置、大小、颜色等属性应清晰可见，且不影响内容的正常观看和使用；

g) 应在虚拟数字人输出内容中添加隐式标识，包括三维模型或输出内容中嵌入不可见的特征或信息，用于内容的溯源、防伪、版权保护等目的；

h) 应保证添加的隐式标识应在压缩、裁剪、噪声干扰后仍能成功提取。

5.2.2 系统安全

虚拟数字人系统在应用过程中：

a) 应支持会话 ID 动态生成与校验，防止会话劫持攻击；

b) 在系统全生命周期中，当系统被对抗样本攻击时，应及时发现并阻断针对抗样本攻击，并及时阻断攻击或进行服务降级。

5.2.3 场景安全

5.2.3.1 政务服务场景

虚拟数字人在政务服务场景中：

a) 应明确政务服务数字人所属单位、运维主体、服务范围、算法与模型来源，并完成备案；

b) 应定期审定政务数字人的知识库、政策文件与办事指南，核验生成信息；

c) 政务数字人应对政务审批、行政处罚、资金发放等高风险事项仅提供咨询指导，不应做出实质审批；

d) 应对公民个人信息、政务业务数据全程加密，后台管理实行分级权限、双人操作，交互数据记录保存6个月以上。

5.2.3.2 金融服务场景

虚拟数字人在金融服务场景中：

a) 应与真实服务人员、金融机构唯一绑定，采用多模态生物特征验证，操作权限分级管控，核心权限双人复核；

b) 应在虚拟数字人执行查询、交易引导等操作时，具备操作确认机制，操作人员和操作日志全程留存，明确操作责任，防范恶意操作；

c) 应规范虚拟数字人形象、语言，符合金融行业服务礼仪。

5.2.3.3 教育教学场景

虚拟数字人在教育教学场景中：

a) 虚拟数字人输出的教学内容应符合国家教育方针及法律法规；
b) 应严格保护学生个人信息，设置访问权限分级管理，仅授权人员可访问；
c) 教学应用系统应稳定运行，定期备份教学数据；
d) 应针对未成年人使用的虚拟数字人应用设置未成年人模式，防范不良信息接触。

5.2.3.4 社交娱乐场景

虚拟数字人在社交娱乐场景中：

a) 虚拟数字人在社交过程中生成的输出内容应经过审核；
b) 虚拟数字人的账号应进行实名认证，防范虚假账号、恶意账号；
c) 应建立社交风险监测机制，针对未成年人使用的社交应用，设置未成年人模式。

5.2.3.5 医疗辅助场景

虚拟数字人在医疗辅助场景中应：

a) 虚拟数字人输出的医疗信息应准确、权威，符合医疗行业规范，医疗内容应经过专业医护人员审核确认，语音交互清晰、规范；
b) 应对用户健康信息、就医记录等医疗数据严格保密，加密存储，符合医疗行业数据安全标准，留存期限应符合医疗行业监管要求；
c) 应将虚拟数字人身份与医疗机构、医护人员绑定，操作人员应具备相应资质，采用多因素认证，操作日志全程留存。

参考文献

[1] GB/T 25069 信息安全技术 术语

[2] GB/T 28449—2018 信息安全技术 网络安全等级保护测评过程指南

[3] GB/T 35273-2020 信息安全技术 个人信息安全规范

[4] GB 45438-2025 网络安全技术 人工智能生成合成内容标识方法

[5] GB/T 45654-2025 网络安全技术 生成式人工智能服务安全基本要求

[6] GB/T 46483-2025 信息技术 客服型虚拟数字人通用技术要求

[7] 生成式人工智能服务管理暂行办法(2023年7月10日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第15号公布)