ICS 35.020
CCS L80

团　　体　　标　　准

T/CHI XX—2026

生成式人工智能多模态合成数据鉴别的安全技术规范

Security technical specification for the identification of generative AI multimodal synthetic data

（征求意见稿）

提交反馈意见时，请将您知道的专利连同支持性文件一并附上。

2026-X-X 发布　　2026-X-X 实施

中国高技术产业发展促进会　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。

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专利的责任。

本文件由内蒙古科技大学提出。

本文件由中国高技术产业发展促进会归口。

本文件起草单位：内蒙古科技大学、太原科技大学、新疆大学、重庆理工大学、桂林电子科技大学、中原工学院、重庆文理学院、集美大学、河南科技大学。

本文件主要起草人：

生成式人工智能多模态合成数据鉴别的安全技术规范

1 范围

本文件规定了多模态合成数据鉴别的总体原则、技术要求、鉴别方法及结果判定等内容，涵盖文本、图像、音频、视频及其跨模态组合形式的合成数据鉴别技术。

本文件适用于对多模态合成数据进行安全鉴别的相关活动，包括但不限于内容审核、身份认证、信息溯源、风险评估及安全治理等应用场景，可为多模态合成数据鉴别系统的设计、开发、测试、评估和应用提供技术参考。本文件不涉及多模态合成数据的生成方法及生成模型的具体实现要求。

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。凡是注日期的引用文件，仅该日期对应的版本适用于本文件；凡是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 45652—2025 网络安全技术 生成式人工智能预训练和优化训练数据安全规范

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

GB/T 45674—2025 网络安全技术 生成式人工智能数据标注安全规范

3 术语及定义

3.1

多模态合成数据 multimodal synthetic data

通过人工智能模型或自动化算法生成，且涉及两种及以上模态信息的数据，包括但不限于文本、图像、音频、视频及其跨模态组合形式。

3.2

合成数据鉴别 synthetic data detection

通过分析数据的内容特征、统计特性、时序一致性或跨模态关联关系，对目标数据是否为人工智能模型生成进行判定的过程。

3.3

身份伪造 identity spoofing

利用合成数据或篡改手段，模拟或冒充特定自然人或组织身份的行为，包括但不限于伪造人脸、声音、语言风格或行为特征等。

3.4

鉴别置信度 detection confidence

用于表示合成数据鉴别结果可信程度的量化指标，通常以概率值、评分或等级形式给出。

3.5

鉴别结果 detection result

对合成数据鉴别的判定结论，通常包括对结论的推理分析报告及最终结论，结论包括“合成数据”“非合成数据”或“不确定”等类型。

3.6

跨模态一致性 cross-modal consistency

不同模态数据在语义、时序或逻辑关系上的一致性程度。

4 多模态合成数据概述

4.1 多模态合成数据的定义与鉴别

见本文件第 3.1 条。鉴别目标是基于生成式模型合成的多模态数据，同时数据拟合真实世界的复杂多维分布，其鉴别数据、特征、方法和主流生成式模型如图 1 所示：

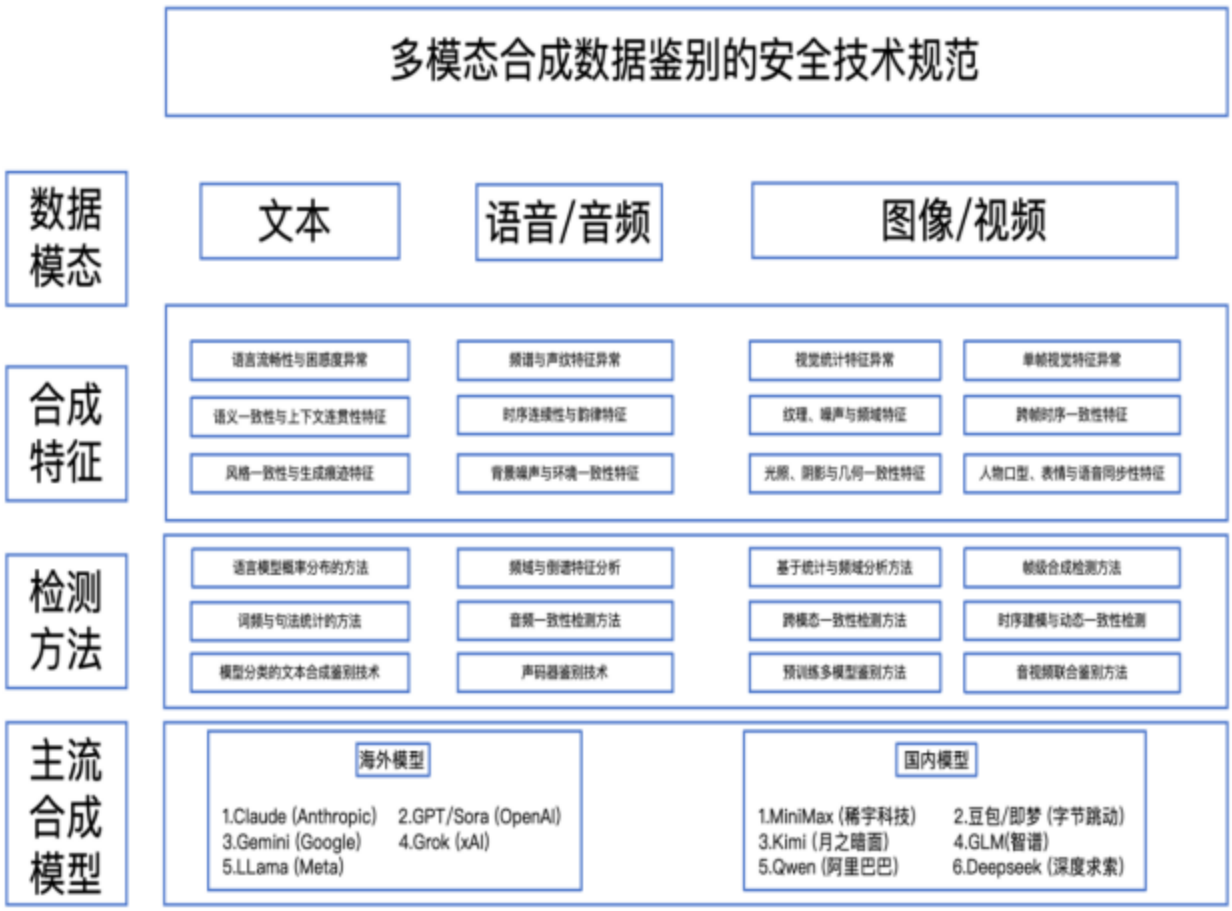

图 1 多模态合成数据鉴别的安全技术框架

4.2 多模态合成数据的类型划分

4.2.1 文本合成数据

指基于预训练语言模型或规则系统生成的文本内容，包括但不限于文章、代码、对话、摘要、翻译及指令性文本。该类数据通常在语义连贯性、风格一致性及事实准确性等方面呈现出与人工创作不同的统计特征。

4.2.2 语音合成数据

指以人类语音为目标的合成音频数据，包括文本转语音（Text-to-Speech, TTS）、语音转换/克隆（Voice Conversion, VC）以及带有特定情绪、语调或说话人特征的合成语音。

4.2.3 音频合成数据

指非语音类的可听音频合成数据，包括合成音乐、环境音效及其他非语言声源（如机械声、自然声等），其主要特征不依赖于语言内容，而侧重节奏、音色、频谱结构等听觉属性。

4.2.4 图像合成数据

指通过生成模型或编辑算法产生的 2D 或者 3D 静态视觉内容，包括全合成图像、局部编辑或修复图像、风格迁移图像，以及针对特定生物特征（如人脸、指纹等）的替换或重构图像。

4.2.5 视频合成数据

指通过生成、编辑或重构技术产生的时序视觉数据，包括数字人视频生成、视频翻译与口型适配、时序插帧、内容重定向以及深度伪造（Deepfake）视频。

4.2.6 多模态合成数据

多模态合成数据是在前述文本、语音、音频、图像、视频等单模态合成技术基础上的系统性集成，指由两种及以上模态协同生成，且各模态之间存在明确时序一致性或语义一致性的合成数据。例如，文本生成视频（Text-to-Video）、图像与同步语音生成的视频内容，以及多模态联合生成的交互式内容等。

4.3 多模态合成数据的安全风险概述

4.3.1 身份伪造与欺骗风险

指利用合成文本、语音、图像或视频仿冒自然人的生物特征、行为特征或表达方式，造成身份冒用、虚假背书或社会工程攻击等风险，可能对个人财产安全、社会信任体系及公共秩序产生不良影响。

4.3.2 内容篡改与虚假信息传播风险

指通过合成或编辑技术生成虚假或误导性文本、音频、图像或视频内容，用于歪曲事实、伪造事件或操纵舆论，可能导致公众认知偏差、信息污染及社会治理风险。

4.3.3 数据滥用与隐私侵害风险

指在未获得合法授权的情况下，使用受版权保护的数据或包含个人敏感信息的数据进行合成，或在合成内容中泄露、推断他人隐私信息，可能引发隐私侵权、版权纠纷及合规风险。

5 文本合成数据鉴别技术

5.1 文本合成数据特征分析

5.1.1 语言流畅性与困惑度异常

文本合成数据在语言表达、语义组织及风格特征方面，通常与人类自然书写文本存在一定差异。通过对文本内容特征进行分析，可为合成文本鉴别提供基础依据。合成文本通常由模型根据概率分布预测，本项要求包括：

a) 测量困惑度（Perplexity），合成数据通常显著低于自然语言；
b) 统计语法结构和词汇衔接规律，合成数据通常具有较高一致性；
c) 词汇平滑性及随机性统计，合成数据缺乏自然语言中的随机性。

5.1.2 语义一致性与上下文连贯性特征

针对长文本，通过分析文本的语义一致性和上下文连贯性特征，可辅助识别合成痕迹。本项要求包括：

a）统计语义前后逻辑矛盾，包括逻辑自洽但语境不合理等问题的情况
b）事实性错误，包括出现语义跳跃、主题偏移

5.1.3 风格一致性与生成痕迹特征

合成文本在文体、语气及用词习惯上，可能呈现出高度稳定或模板化特征，因此本项要求包括（参见附录 A）：

a) 检测文本中是否存在模型特有的高频词汇偏好；
b) 检查是否存在固定的句式结构或特殊；
c) 检查文本标点符号使用习惯是否符合模型特有规律；
d) 检查文本风格一致性，是否存在表述风格的不连续性。

5.2 基于规则与统计分析的文本合成鉴别技术

5.2.1 基于语言模型概率分布的方法

本项要求包括：

a) 利用已知模型计算待测文本的对数似然值；
b) 分析其是否处于高概率生成区间（如 GLTR 方法）。

5.2.2 基于词频与句法统计的方法

通过与真人语料库对比识别异常，本项要求包括：

a) 统计文本中的词频分布，句法结构特征及标点使用规律，检测合成文本与真实文本在统计特性上的差异；
b) 统计用词的深度分布，如统计文本中虚词、实词的比率及句法树。
该类方法可用于初步筛查或与其他鉴别方法结合使用。

5.3 基于深度学习的文本合成鉴别技术

5.3.1 基于预训练语言模型的鉴别方法

该方法宜结合多样化训练样本以提升泛化能力，本项要求包括：
a）利用预训练语言模型对文本进行特征编码， 并通过模型输出分类鉴别结果；
b）基于预训练语言模型的鉴别方法，通过微调 (Fine-tuning) 构建二分类器，作为分类鉴别模型。

5.3.2 对抗生成文本检测方法

本项要求包括：
a）针对生成模型不断演进的特点，采用对抗训练或鲁棒性增强方法鉴别
b）针对经过混淆处理的合成文本，利用对抗训练提升模型对同义词替换、语序重组等攻击的识别鲁棒性。

5.4 文本合成数据鉴别流程与结果判定

文本合成数据鉴别流程宜包括数据预处理、特征提取、模型判定及结果输出等环节。鉴别结果应给出明确判定结论，并宜同时提供鉴别置信度或风险等级，用于辅助决策。
流程应包括：文本清洗（去除噪声标签）→ 分词与向量化 → 深度特征提取 → 逻辑一致性核查 → 判定输出。

6 图像合成数据鉴别技术

6.1 图像合成数据特征分析

6.1.1 视觉统计特征异常

图像合成数据在视觉统计特性、纹理细节及空间一致性等方面，可能存在可识别的异常特征。合成图像在颜色分布、亮度直方图等统计特征上，可能与真实图像存在差异，并常在颜色饱和度分布、像素间关联性上存在非自然规律。本项要求包括：
a) 分析图像像素值的二阶或高阶统计量。
b) 检测合成图像在颜色直方图上的不连续跳变、饱和度通道（如 HSV 空间）的非自然聚集。
c) 核查像素间预测残差的分布，检测图像是否因量化或生成算法影响表现出分布偏移。

6.1.2 纹理、噪声与频域特征

通过分析图像的纹理连续性、噪声模式及频域分布特征，可识别潜在的合成痕迹。生成算法在图像细节处（如皮肤、毛发）常表现出过度平滑或重复伪影，且在频域中存在由于上采样引起的“格点效应”。
本项要求包括：

a) 利用局部三值模式（LTP）等算子量化纹理平滑度。
b) 针对频域特征，检测离散余弦变换（DCT）或傅里叶变换（FFT）后的能量分布。
c) 识别生成对抗网络（GAN）或扩散模型（Diffusion Model）特有的上采样周期性伪影（Grid Artifacts）。

6.1.3 光照、阴影与几何一致性特征

合成图像在光照方向、阴影关系及几何结构一致性方面，可能存在不符合真实物理规律的情况。因此需检测多光源方向不一、投影形状不符合几何透视、反射高光缺失等物理不一致现象。本项要求包括：

a) 利用球面对衬度或形状恢复（SFS）技术重建场景光效。
b) 核查人物面部高光点与环境光源位置的几何对应关系。
c) 检测人脸图像中左右瞳孔反射光的一致性，以及五官各关键点在欧几里得空间下的投影几何约束。

6.2 基于传统方法的图像合成鉴别技术

6.2.1 基于统计分析的方法

本项要求包括：

a） 基于图像统计特征（如局部导数特征、像素共生矩阵或马尔可夫随机场特征），检测图像合成或编辑过程引入的整体统计分布异常。

b） 基于局部二值模式（Local Binary Pattern, LBP）、差分直方图等纹理统计方法，判定图像像素是否经历合成生成或二次编辑操作。

6.2.2 基于频域分析的方法

本项通过分析图像在空间域或频域中的分布异常，实现对合成痕迹的识别。本项要求包括：

a) 基于图像频域变换（如傅里叶变换、离散余弦变换），分析高频与低频成分的能量分布及结构特征。

b) 利用傅里叶域中的高频响应特征，检测生成模型或合成算法引入的周期性或规则性痕迹。

c) 通过分析 DCT 系数的拉普拉斯分布偏移或统计异常，识别与特定压缩或生成过程相关的合成痕迹。

d) 利用双谱分析（Bispectral Analysis）捕捉频域中的非线性相位耦合特征，以检测高阶统计特性异常。

6.3 基于深度学习的图像合成鉴别技术

6.3.1 基于卷积神经网络的鉴别方法

本项要求包括：

a） 利用卷积神经网络（CNN）对图像的局部与全局特征进行联合建模，捕获合成图像中常见的局部伪影（如人脸关键区域拼接或纹理异常）；

b） 通过约束卷积（Constrained Convolution）或高通滤波结构提取图像中的高频残差成分，以识别生成模型在微观层面引入的规律性噪声特征。

6.3.2 基于 Transformer 结构的鉴别方法

基于 Transformer 的方法能够建模图像中长距离依赖关系，适用于识别复杂语义层面的合成异常。本项要求包括：

a） 利用自注意力（Self-Attention）机制分析图像全局构图特征，识别语义层面的不一致性；

b) 采用视觉 Transformer（Vision Transformer, ViT）对图像补丁（Patch）间的全局关联关系进行建模；

c) 结合注意力权重或热力图分析图像中是否存在语义逻辑异常，如人体结构畸变、物体遮挡关系不合理等。

6.3.3 多特征融合鉴别方法

本项要求包括：

a） 综合利用空间域像素特征、频域噪声特征及残差特征，对多种模型输出结果进行融合分析。

b) 构建双流或多流神经网络结构，分别提取图像的视觉语义信息与微观频域指纹信息。

c) 采用特征级或决策级融合策略，并可引入对比学习等机制以提升跨生成方式的泛化能力。

6.4 图像合成数据鉴别流程与结果判定

图像合成数据的鉴别流程应包括图像预处理、特征分析、模型判定及结果输出等环节，并结合置信度信息对高风险样本进行提示。

典型流程包括：

图像预处理（如统一分辨率与色彩空间）→ 特征图生成 → 置信度预测 → 伪造区域定位与结果判定。

7 音频合成数据鉴别技术

7.1 语音合成数据特征分析

7.1.1 频谱与声纹特征异常

语音包含显著的人类生物特征。通过分析频谱结构与声纹特征，可识别合成语音在生成过程中产生的异常模式。其中，合成语音在频谱分布及声纹稳定性方面可能与真实语音存在差异。本项要求包括：

a） 检测语音在高频频段（如 8 kHz 以上）是否存在非自然的能量衰减、截断或异常谐波分布；

b) 分析声纹嵌入向量在高维特征空间中的分布与聚类特性，识别是否存在偏离正常人类发声机理的频谱或声纹指纹。

7.1.2 时序连续性与韵律特征

通过分析语音的时序连续性、语速及韵律特征，可辅助识别合成语音。本项要求包括：

a） 分析基频（F0）轨迹的连续性与平滑性，检测不符合生理发声特性的瞬时跳变或异常波动；

b) 针对流式语音合成场景，统计音节或词组间停顿时长分布，检测声码器生成可能引入的相位或节奏不连续现象。

7.1.3 背景噪声与环境一致性特征

合成语音在背景噪声及声学环境一致性方面可能呈现不自然特征。本项要求包括：

a） 分析人声与背景噪声的频谱耦合关系，识别背景噪声在时域或频域上的异常重复或规律性模式；

b) 检测不同语句或语段间的环境脉冲响应（Room Impulse Response, RIR）一致性，以判断是否存在不合理的环境特征复用。

7.2 基于信号处理的音频合成鉴别技术

7.2.1 频域与倒谱特征分析

通过对音频信号进行频域或倒谱分析，可提取用于合成鉴别的关键特征。本项要求包括：

a） 提取梅尔频率倒谱系数（MFCC）、常数 Q 变换（CQT）特征或线性预测倒谱系数（LPCC）等表征音频结构的特征。

b）利用高阶统计量分析音频信号的非线性特性，检测传统合成或拼接过程引入的相位或结构异常。

7.2.2 音频一致性检测方法

通过分析音频特征在时间或频率维度上的一致性，识别潜在的合成痕迹。本项要求包括：

a） 利用预训练音频或说话人识别模型提取固定维度的音频嵌入特征（如 X-Vector）。

b) 计算不同音频片段间的特征相似度（如余弦相似度或欧氏距离），判定音频整体或局部特征的一致性。

7.3 基于深度学习的音频合成鉴别技术

7.3.1 语音合成检测模型

利用深度学习模型对语音信号进行端到端建模，实现对合成语音的自动检测。本项要求包括：

a） 采用端到端语音鉴别模型（如基于原始波形建模的结构）学习声码器生成语音的统计特性。

b) 结合注意力机制或时间加权策略，重点关注合成语音中对伪造判定敏感的关键帧或频段。

7.3.2 伪造语音检测方法

针对语音模仿、语音克隆等场景，采用专门的说话人伪造检测方法。本项要求包括：

a） 在包含多种合成算法的大规模数据集上引入对比学习机制，构建区分真实语音与合成语音的判别模型，提升对未知生成方式的泛化能力。

b） 利用多尺度判别结构，对不同频率子带或时间尺度上的细微失真进行联合分析。

7.4 音频合成数据鉴别流程与结果判定

音频合成数据的鉴别流程应包括音频预处理、特征提取、模型判定及结果输出等环节，并结合具体应用场景给出相应的风险提示。

典型流程包括：

音频流截取 → 时频表示生成 → 深度特征推断 → 置信度得分计算与结果判定。

8 视频合成数据鉴别技术

8.1 视频合成数据特征分析

8.1.1 单帧视觉特征异常

合成视频在空间特征、时间连续性及多模态同步性方面，可能存在异常。合成视频的单帧画面可能表现出与真实图像相似的合成特征。检测关键帧内的人脸边缘模糊、背景扭曲或光影突变。本项要求包括：

a）检测视频关键帧中人脸边界的掩码痕迹（Mask Artifacts）、瞳孔反射光的不对称性以及牙齿等复杂纹理处的重复伪影。

b）识别背景纹理与主体运动边缘的逻辑失真。

8.1.2 跨帧时序一致性特征

通过分析视频帧间的时序一致性，可识别动态合成痕迹。本项要求包括：
检测视频帧间生理信号（如心率引起的微弱色差）的非自然跳变，或眨眼、点头等动作的连贯性缺陷。

a) 分析视频帧间的像素残差，检测面部关键点在时间维度上的运动噪声。
b) 基于远程光电容积脉搏波（rPPG）技术提取人物血流脉动特征，识别人物生理特征的缺失或异常。

8.1.3 人物口型、表情与语音同步性特征

在涉及人物的视频中，口型、表情与语音之间的同步关系是重要的鉴别特征。本项要求包括：

a) 计算唇部运动轨迹与音轨语义的匹配度（AV-Consistency）。
b) 评估口型闭合状态与音轨中辅音/元音出现的时序匹配度。
c) 检测面部表情肌肉的联动关系（如笑时眼部肌肉的收缩）是否符合生理真实性。

8.2 基于图像与时序分析的视频合成鉴别技术

8.2.1 帧级合成检测方法

本项要求包括：

a) 逐帧分析视频帧，检测局部或阶段性合成内容。
b) 抽样视频帧，利用高精度图像鉴别器识别局部篡改。
c) 分层抽样视频流，利用 CNN 或 Vision Transformer 架构对每一帧图像进行空间域鉴别，快速定位伪造关键帧。

8.2.2 时序建模与动态一致性检测

本项要求包括：

a) 利用时序建模方法分析视频整体动态特征。
b) 使用 3D-CNN 或卷积长短期记忆网络（ConvLSTM）分析动作的流形连续性。
c) 采用 3D 卷积（C3D）或循环神经网络（RNN）捕获视频的时序依赖关系。

d) 分析帧间光流场（Optical Flow）的平滑度，识别合成帧引起的运动轨迹中断。

8.3 基于多模态融合的视频合成鉴别技术

8.3.1 音视频联合鉴别方法

通过联合分析音频与视频特征，可提高对复杂合成视频的检测能力。本项要求包括：

a) 提取音视频的双流特征，通过注意力机制融合，判断是否属于同一生成过程。
b) 构建音视频对齐模型，学习音轨与唇部动作的联合嵌入空间，识别语义层面的不匹配。
c) 利用交叉注意力机制融合视觉轨迹与音频节律特征，检测潜在时序矛盾。

8.3.2 跨模态一致性检测方法

分析不同模态之间的一致性关系，有助于识别跨模态合成异常。本项要求包括：

a) 校验文本字幕、语音语调与视频场景的一致性，识别“画音不符”的合成痕迹。
b) 校验视觉场景中的光影动态与语音传达的情感强度的一致性。
c) 利用文本、音频、视频三位一体的校验机制，识别场景逻辑上的低级逻辑错误（如背景天气与语音描述不符）。

8.4 视频合成数据鉴别流程与结果判定

流程应包括：视频流解复用 → 关键帧与音轨提取 → 跨模态特征校验 → 综合逻辑判定。
对鉴别结果存在不确定性的情况，宜结合多种方法进行综合判定。方法应包括：

a) 引入熵值或方差度量判定结果的确定性，对处于模糊阈值区间的样本启动多算法集成表决或人工复核。
b) 对单帧伪影、时序一致性及音视频同步性等多维指标设定权重，构建综合置信度计算模型。
c) 监测视频全时段的置信度波动，针对局部区间的异常跌落触发精细化伪造风险预警。
d) 结合物理规律校验与元数据溯源信息。
e) 建立模态间一致性校验机制。

9 多模态合成数据综合鉴别方法

9.1 跨模态特征一致性检测

本项要求包括：

a) 对齐分析图像、音频、文本等不同模态特征，识别多模态合成过程中产生的跨模态失配现象。
b) 利用多模态对齐技术校验文本描述与图像内容的语义符合度。
c) 重点检测视频中语音情感特征与面部微表情（Action Units）的映射一致性。

9.2 多模态联合判定机制

构建多模态联合判定机制，对各模态鉴别结果进行融合分析，通过规则约束或模型加权方式输出综合鉴别结论，降低单一模态误判风险。本项要求包括：

a) 建立基于证据理论（如 D-S 证据理论）或逻辑回归的融合模型。
b) 根据输入数据质量与应用场景动态调整各模态权重。

9.3 不确定性与置信度评估方法

针对多模态鉴别结果的不确定性，引入置信度评估机制，对模型输出结果进行可信度量化，为风险分级与人工复核提供依据。本项要求包括：

a) 引入蒙特卡洛采样（Monte Carlo Dropout）或保序回归（Isotonic Regression）对模型原始得分进行校准。
b) 划分判定结果等级与量化的概率分布区间。

10 鉴别工具与评估方法

10.1 合成数据鉴别工具要求

合成数据鉴别工具宜具备多模态数据处理能力，支持自动化分析、结果可解释输出及批量检测，以满足不同应用场景下的使用需求。本项要求包括：

a) 工具应支持主流音视频容器格式的解析与解码。
b) 工具应具备神经网络及其它深度学习模型推理能力。
c) 输出结果除分类结果外，还应包括置信度，可解释性、特征可视化组件，如热力图（Heatmap）定位伪造区域、频谱异常点标记等
d) 工具具备满足实时流检测的吞吐量能力。

10.2 鉴别性能评估方法

通过构建包含真实数据与多类型合成数据的评测集，对鉴别工具的检测性能进行系统评估，验证其在不同模态、不同生成方式及不同失真条件下的有效性与稳健性。本项要求包括：

a） 文本数据集

应包括常见字符编码格式，如 UTF-8，中文场景应覆盖 GB2312/GBK 等编码形式；同时包含真实文本与由不同文本生成模型生成的合成文本，覆盖不同文本长度、主题与风格，以评估在多样语义与格式条件下的鉴别能力。

b） 图像数据集

应包含真实图像（可来源于公开数据集或用户生成内容）与合成图像，支持多种图像格式（如 JPEG、PNG、BMP、TIFF 等）。针对压缩图像，应覆盖有损压缩（如不同质量因子的 JPEG）与无损压缩（如 PNG）。对于包含中文文本的图像，应考虑嵌入式字体与字符编码差异（如支持 GB2312 或 UTF-8 的 OCR 场景）。

c） 语音数据集

应包含真实语音与合成语音，覆盖不同性别、语速、情感及语言/口音特征；音频格式可包括 WAV、MP3、AAC 等，并考虑不同采样率、比特率及常见失真条件（如背景噪声、混响、编码压缩），以评估在复杂声学环境下的鉴别性能。

d） 视频数据集

应包含真实视频与合成视频，覆盖不同编码方式与压缩率（如 H.264/H.265，不同位速率从低比特率约 500 kbps 至高比特率约 10 Mbps）、不同分辨率（如 360p、720p、1080p、4K），并包含典型扰动或攻击样本（如雨雾、抖动、模糊、音频干扰等），用于评估在多种退化条件下的鲁棒性。评估过程需采用交叉验证法，确保算法在非公开数据集上的鲁棒性。

10.3 鉴别指标体系

10.3.1 准确率、召回率与误报率

本项要求包括：

a) 采用准确率、召回率及误报率等指标，量化评估鉴别结果的正确性与风险控制能力。
b) 引入 F1-Score 指标鉴别合成性能。

10.3.2 鲁棒性与泛化能力评估

通过跨模型、跨数据分布及抗干扰测试，评估鉴别方法在未知合成算法及复杂应用环境下的鲁棒性与泛化能力。本项要求包括：

a) 对抗性攻击（Adversarial Attacks）的防御测试。
b) 模拟真实互联网传播环境中的缩放、裁剪、高斯模糊及多轮有损压缩，测定算法性能的衰减曲线。

附 录 A
（资料性）
常见合成数据特征

表 A.1 常见合成数据特征

序号	模态	特征类别	典型表现	检测方法建议
A.1	文本	高频词偏好	重复短语或模式	统计词频分布，分析 n-gram 重复率>20%相似度阈值
A.2		特殊符号异常	出现重复有规律特征的符号、格式统一过度	与人工写作习惯不符，特殊符号频率>30%
A.3		机器幻觉	引用不存在的事实、文献、法规条款	事实核查工具验证内容真实性，溯源信息编号特征及内容。
A.4		逻辑一致性异常	前后观点矛盾但语法通顺	局部连贯但整体不一致知识图谱验证
A.5		情绪表达异常	长文本中语气、情绪高度一致	缺乏人类写作常识及技巧
A.6	语音/音频	性别特征异常	声纹与说话内容的性别特征不匹配	比较基频（F0）一致性
A.7		语调停顿异常	语速均匀、停顿位置固定，不符合自然语言节律	时序分析停顿间隔，标准差＞自然语音 2 倍
A.8		合成噪声	出现电音、哑音、金属感	频谱分析可识别高频噪声峰 SNR<30dB
A.9		呼吸特征缺失	长时间连续发声无呼吸声，非自然发声特征	检测低频呼吸信号缺失 MFCC 特征
A.10		情感特征异常	情绪变化与语义不匹配	情感曲线平滑，语调与内容不匹配>50%
A.11	图像/视频	局部模糊	面部、文字、边缘区域模糊不一致	分辨率区域异常，Laplacian 方差＜阈值
A.12		字符异常	不存在或结构错误的字母、汉字	OCR 无法识别
A.13		科学规律违背	光影、物理结构不符合现实	阴影/反射异常检测
A.14		音视频不同步	口型与语音不同步	时间轴错位，唇音对齐误差>100ms
A.15		人物说话不自然	面部肌肉运动不协调	微表情异常，唇动模型检，LipNet 评分<0.8
A.16		运动帧异常	动作跳帧、肢体穿模，帧间连续性异常	光流分析帧间连续性向量异常>10%
A.17		物体形变异常	物品边缘漂移、结构变化	几何一致性破坏，边缘检测畸变率>5%