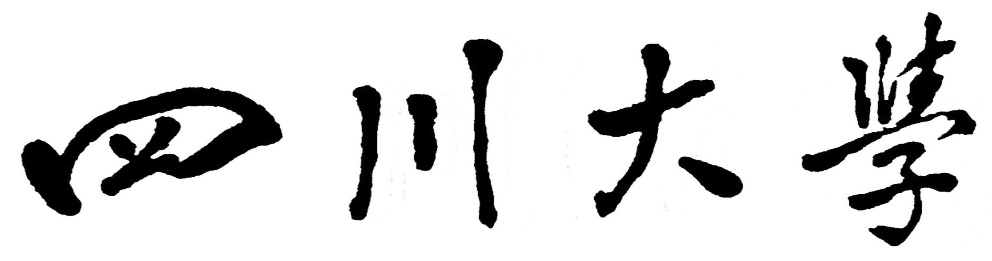
****

**大学生创新训练计划**

**创新类项目申报书**

**项目名称：基于大模型辅助的多模态委婉有害内容的检测技术研究**

**项目负责人： 郑煬阳**

**所在学院： 网络空间安全学院**

**专业年级： 2024级**

**学 号： 2024141530142**

**手 机： 15808310028**

**电子邮箱： 2053900282@qq.com**

**指导教师： 王海舟**

**项目起止年月： 2025年11月至2026年10月**

**项目参与学生人数： 5人**

**四川大学教务处制**

年 月

填写说明

一、凡申报**四川大学“大学生创新训练计划”**必须填写本申报书。创新类项目是本科生个人或团队，在导师指导下，自主完成创新性研究项目设计、研究条件准备和项目实施、研究报告撰写、成果（学术）交流等工作。

**二、“项目所属一级学科和代码”**参考《普通高等学校本科专业目录和专业介绍（2012年）》。

三、**“项目开展支撑平台”**指支撑本项目开展的国家级和省部级重点实验室（中心、平台等）、国家双创示范基地平台、教学实验中心（实验室）、企业、事业或其他单位等，表中填写平台名称，可以多个。

四、**“项目组成员”**人数原则上不超过五人，应排序。

五、**“项目成熟度**”请参考附件《项目成熟度量表》。

六、本书应该填写完整、内容详实、表达准确，数字一律填写阿拉伯数字。

七、报送申报书的电子文档至负责人所在学院。

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 项目名称 | 基于大模型辅助的多模态委婉有害内容的检测技术研究 | | | | | |
| 项目属性 | ☑面上项目 □人工智能+新质战略育苗”（含2035特区子计划项目）  □交叉学科子计划项目 | | | | | |
| 申请类别 | ☑科学探索与工程技术类 □人文艺术与社会科学类  □软件信息与文创类 □智能装备与医疗器械类  □生物医药与新材料类 | | | | | |
| 申请经费 | 5000 元 | | 起止时间 | | 2025年11月至2026年10月 | |
| 项目所属 一级学科和代码 | 工学 08 | | | | | |
| 项目开展 支撑平台 |  | | | | | |
| 项目来源  （可多选） | □十大重点支持领域的项目  ☑进课题组、进实验室、进科研团队参与的项目  □国家级和省部级重点实验室（中心、平台等）、国家双创示范基地平台支持申报项目  □交叉学科创新项目  □“青年红色筑梦之旅”计划项目  □基于前期研究实践成果、继续深入研究实践的创新项目  □高水平课题  □其他 | | | | | |
| 高水平课题名称（非高水平课题可不填） | 命题名称 |  | | | | |
| 校内指导老师姓名（非交叉学科子计划项目一般仅允许一位指导老师） |  | | | | |
| 所属重点支持领域（可不选） | 选择1项：E  A.不填  B.泛终端芯片及操作系统应用开发  C.重大应用关键软件  D.云计算和大数据  E.人工智能  F.无人驾驶  G.新能源与储能技术  H.生物技术与生物育种  I.绿色环保与固废资源化  J.第五代通信技术和新一代IP网络通信技术  K.社会事业与文化传承 | | | | | |
| 负责人之前参与大创项目情况 | 无 | | | | | |
| 项目成员之前参与大创项目情况 | 无 | | | | | |
| 项目负责人基本信息 | | | | | | |
| 姓名 | 学号 | 专业年级 | | | 所在学院 | |
| 郑煬阳 | 2024141530142 | 网络空间安全2024级 | | | 网络空间安全学院 | |
| 性别 | 手机 | 电子邮箱 | | | 身份证号 | |
| 男 | 15808310028 | 2053900282@qq.com | | | 511528200511117018 | |
| 项目组成员基本信息 | | | | | | |
| 序号（含排序） | 1 | 2 | | 3 | 4 |
| 姓名/性别 | 孙浩哲/男 | 彭实/男 | | 匡家辉/男 | 李佳玥/女 |
| 学号 | 2024141530057 | 2024141530164 | | 2024141530037 | 2024141530145 |
| 专业年级 | 网络空间安全2024级 | 网络空间安全2024级 | | 网络空间安全2024级 | 网络空间安全2024级 |
| 所在学院 | 网络空间安全学院 | 网络空间安全学院 | | 网络空间安全学院 | 网络空间安全学院 |
| 手机 | 18116601240 | 18980365560 | | 15884941123 | 13899856897 |
| 电子邮箱 | 2361498928@qq.com | 1580733692@qq.com | | 1819137877@qq.com | 3213096997@qq.com |
| 身份证号 | 620105200601170013 | 511423200504230031 | | 511681200610290614 | 650104200602100729 |
| 签名 |  |  | |  |  |
| 指导教师1 基本信息  （非交叉学科子计划项目一般仅允许一位指导老师） | | | | | | |
| 姓名 | 所在学院或单位 | 研究方向 | | | 职称/职务 | |
| 王海舟 | 网络空间安全学院 | 网络安全、情报分析、舆情监控 | | | 教授 | |
| 性别/年龄 | 手机 | 电子邮箱 | | | 签名 | |

|  |  |  |  |
| --- | --- | --- | --- |
| 男/39 | 13730811282 | whzh.nc@scu.edu.cn |  |
| 指导教师2 基本信息  （交叉学科子计划项目需填写第二指导老师） | | | |
| 姓名 | 所在学院或单位 | 研究方向 | 职称/职务 |
|  |  |  |  |
| 性别/年龄 | 手机 | 电子邮箱 | 签名 |
|  |  |  |  |

|  |
| --- |
| **项目摘要(限200字以内)** |
| 随着社交媒体发展，有害内容常以**委婉语**、**隐喻图像**、**音频暗语**等隐蔽形式传播，**传统单模态审核手段难以应对**，**多模态协同识别的核心问题尚未有效解决**，严重影响网络环境健康。  本项目聚焦**多模态检测技术**突破，融合文字、图像与音频的特征检测，构建**多模态检测模型**，深度解析文本隐藏含义、图像敏感元素及音频违规信息，通过**跨模态特征融**合实现精准识别，并依托增量学习持续优化模型，为**营造清朗网络环境**提供关键技术支撑。 |
|  |
|  |
| **特色创新点（限100字以内，建议2-3点）** |
| 1.**首次**系统化提出“中文**三模态委婉语**”识别范式，构建“委婉强度—模态关联”标注体系，**弥补中文多模态标注粒度不足**。  2.**首次**构建可工程化“多模态增量适配”检测框架，实现谐音、隐喻与模态错位的**快速适配与稳定更新**。 |

**报告正文**

|  |
| --- |
| 1. **立项依据与研究内容**（**建议8000字以下**（不包括文献）） |
| 1．**项目的立项依据**（【理工医科】研究意义、国内外研究现状及发展动态分析，需结合科学研究发展趋势来论述科学意义；或结合国民经济和社会发展中迫切需要解决的关键科技问题来论述其应用前景。附主要参考文献目录【哲学社会学科】国内外相关研究的学术史梳理及研究动态;本课题相对于已有研究的独到学术价值和应用价值等。附主要参考文献目录） |
| **1.1研究背景**  **1.1.1背景综述**  在人工智能技术与互联网应用深度融合的当下，以生成式AI为核心的多模态内容创作已成为网络信息传播的主流形态。文本、图像、音频、视频等多模态内容凭借丰富的表现力与高效的传播性，重塑了公众的信息交互方式，推动了数字经济、文化传播等领域的创新发展。  与此同时，生成式AI技术的低门槛化与多模态内容的隐蔽性特征，为“委婉有害内容”的滋生与传播提供了温床。不同于传统直白的有害信息，委婉有害内容通过语义替换、符号隐喻、多模态转译等手段，规避现有审核机制的检测，呈现出“形式合规化、含义隐蔽化、传播扩散化”的新特征：在文本层面，利用谐音、拼音缩写、语义联想等方式，将侮辱性、攻击性含义包裹在日常化表述中；在图像层面，通过“文本-图像转译”策略，先将敏感词汇谐音化或隐喻化，再以该表述作为提示词驱动生成式AI生成看似无害的图像，实现有害含义的视觉化隐藏；在音频与视频层面，谐音辱骂、节奏变形、画面隐喻等手段普遍存在，甚至出现“多模态协同规避”现象——例如视频中画面展示隐喻图像、音频使用谐音话术、字幕嵌入语义替换文本，形成跨模态的有害信息传播闭环。  此类多模态委婉有害内容的传播，已对网络空间治理、公众信息安全乃至社会稳定产生深远影响。一方面，其隐蔽性导致现有基于关键词匹配、单一模态特征识别的审核技术失效，使得有害信息能够在社交平台、短视频平台等渠道广泛扩散，误导公众认知、煽动负面情绪；另一方面，随着国家级网络安全战略对有害信息治理要求的不断提升，如何精准识别多模态场景下的委婉有害内容，已成为保障公民信息权益、维护网络空间清朗生态、落实国家网络安全战略的关键技术需求。在此背景下，开展多模态委婉有害内容检测技术研究，既是应对当前网络安全新挑战的必然选择，也是推动AI技术在网络空间治理领域合规应用、服务国家级网络安全保障体系建设的重要举措。  **1.1.2存在问题**  近年来，违规内容多以委婉语的形式出现以逃避平台捕捉，尤其是在各种社交平台上，如抖音、微博、小红书等社交应用。这种违规委婉语的出现形式多样，通常利用文字的语言歧义、文字中穿插大量乱码或者带有侮辱信息的图片来传达违规内容（如图1-1所示），导致应用平台内社交矛盾激化，不利于营造清朗的网络环境。  日程表  AI 生成的内容可能不正确。  图1-1 利用文字与图片在抖音私信发表违规内容  进行多模态违规委婉语检测主要涉及以下技术难题：   1. 动态演变特征导致检测滞后：多模态委婉语在现代社会的更迭中，新的组合方式层出不穷，使检测系统始终处于追赶状态，例如“唐人”在十年前的网络中指唐人街、唐朝的人、或者借指中国人，可是在近年的网络中，“唐人”又有了“唐氏综合症患者”的含义，而在不停的传播中，又被赋予了“肢体肥大”、“智商低下”等语义，后被引申成了侮辱他人的含义。 2. 跨模态语义融合与对齐难题：多模态违规内容常将文字委婉语与图片、表情包、视频片段等非文本信息结合，单一模态的语义解读无法完整还原违规意图，跨模态信息的融合对齐成为关键障碍。语义碎片化分布：违规信息可能拆分在不同模态中，例如文字用“懂得都懂”配合一张带有侮辱性元素的表情包，文字本身无违规性，需结合图像语义才能判断，而不同模态的语义关联缺乏明确规律。模态间语义冲突与补充：部分场景下，图像与文字语义存在矛盾或补充，例如文字描述“可爱的小天使”，搭配的却是丑化、歧视特定群体的图片，系统需区分这种“正话反说”的跨模态讽刺，难度极大。 3. 对抗性规避手段的动态升级：为逃避检测，违规内容发布者会主动研究平台检测规则，不断升级规避手段，形成“检测—规避—再检测” 的对抗循环，使检测系统难以应对。文字层面的对抗变异：除了基础的谐音、拆分，还会通过生僻字替换（如用“砼” 替代 “同” 组成侮辱性词汇）、拼音首字母 + 特殊符号组合、方言谐音转写（用方言发音的文字替代违规词汇）等方式规避文字检测。多模态联合对抗：将违规信息隐藏在多模态内容的细节中，例如在视频背景角落嵌入微小的侮辱性文字图片、用特定节奏的音乐配合模糊的文字暗示违规内容，或通过 “图文错位”（文字描述 A，图片细节隐含B类违规信息）的方式降低检测概率。   随着人们聊天模式的不断迭代，传统的平台审核模式已不足以应对这些新兴的威胁。未来的研究应聚焦于提升违规委婉语内容检测的准确性和实时性，并通过多模态融合、并借助大模型，从而营造清朗的网络环境。  **1.2研究意义**  **1.2.1学习意义**  本项目组成员通过此次项目实践，可以拓展在计算机领域内的知识范围，很大程度提高了个人综合能力水平：  （1）通过本次项目，项目组成员可以培养团队协作的精神，通过将项目细分为不同模块，分工合作，提高项目进展速度，为以后在工作中分配以及完成任务打下良好的基础；  （2）通过本次项目，项目组成员可以学习到计算机领域的深层次知识与前沿知识，学习到多模态特征提取和卷积神经网络等本科教学阶段不常接触的知识，为以后的专业深造提供入门与过渡；  （3）在实现算法阶段，算法的微调和对抗训练实现十分困难，在此阶段项目组成员的编程能力与计算思维能力得到了很大的提升，专业素养有了很大提升，为未来工作打下了良好的基础。  **1.2.2科学意义**  由于特征提取、神经网络和网络爬虫技术在现代社会对解决很多问题都有良好的效果，理论证明，它们在项目实践中是解决问题很好的方法，所以在项目中尝试使用这些技术不仅可以更好的解决问题，并且在实践中可以从实践上对相关技术进行验证与完善，并深入了解这些技术，补充技术的理论完备性和实际操作知识。  **1.2.3社会意义**  随着信息科技的发展，社交网络已经成为人们生活中不可缺少的一部分，甚至成为人类虚拟交流的全部平台。而在广大社交媒体平台上，人们利用多模态委婉语发表违规言论，逃避平台管控，有碍社交平台治理，不利于营造积极向上的网络环境。因此设计研究出高效、准确的多模态委婉语检测系统，对捕捉有害信息、促进文明发言的与社会的稳定、提升用户体验有着十分重要的意义。  **1.3国内外研究现状**  多模态委婉有害内容检测技术作为网络内容安全领域的核心防御技术之一，在平台内容治理、用户权益保护和社会舆论引导等领域具有重要的理论价值和实践意义。特别是，精准的委婉有害内容识别对于提升下游内容监管系统的有效性和及时性至关重要。  多模态委婉有害内容检测的关键任务在于实现跨模态信息的深度融合以及面向隐晦有害语义的准确判定。尽管针对单一模态的有害内容检测方法已趋于成熟，但在实际网络环境中，由于多模态数据呈现出显著的异构性（如文本语义与图像视觉的错位表达）、模糊性（如隐喻性文本搭配歧义图像）和动态演化特征（如新型委婉话术、跨平台传播变异），加之有害内容的隐蔽性不断升级（如谐音梗伪装、视觉隐喻暗示、多模态协同隐晦表达等），使得跨模态特征的有效融合和用于检测系统训练的标注样本获取面临巨大挑战。  在此背景下，如何基于有限的高质量标注数据，结合先进的多模态深度学习算法，实现高效准确的跨模态特征对齐与融合，以及可靠的委婉有害语义鉴别，从而为网络内容安全体系构建全方位的防御屏障，成为一个极具研究价值的前沿课题，引起了国内外学术界和工业界的广泛关注。以下将系统阐述多模态委婉有害内容检测技术的研究现状。  **1.3.1单个模态检测研究**   1. **文本**   **大语言模型（LLMs）的长上下文处理效率与性能优化**  Yaniv 等人[1]提出无参数、轻量级的注意力机制改进方案，核心逻辑是“让 tokens动态屏蔽无关历史tokens，减少后续注意力计算的冗余”。在C4数据集上，带选择性注意力的Transformer语言建模性能显著优于标准Transformer，且模型规模越大、上下文越长，提升越明显。  Xu 等人[2]首次提出“距离感知的渐进式位置复用”策略，解决了训练 - free 方法的静态映射缺陷，在无额外训练成本的前提下，实现LLM上下文窗口的高效扩展，尤其适用于资源受限场景。  针对低资源语言场景，Choudhury等人[3]提出基于跨语言预训练的文本特征增强方法，通过将高资源语言的语义知识迁移至孟加拉语等低资源语言，在 MultiOFF 数据集上实现了70.3%的F1分数，较单语模型提升8.2%，为多语言文本有害内容检测提供了有效路径。针对跨语言模型的泛化能力，有研究[4]探索了利用零样本和少样本学习提升模型在西班牙语和英语等不同语言环境下的通用性，验证了模型在有限数据下进行跨语言迁移的潜力。在动态语义适配方面，Furniturewala 等人[5]设计多阶段提示策略，通过“简单提示 - 类别提示 - 尺度输出”的层级优化，在DistilBERT模型上使文本模态检测准确率提升至 78.6%，验证了提示工程对文本语义挖掘的增益作用。  **（2）图像**  **基于Transformer和CNN的视觉处理模型**  ETH Zurich 的 Liang 等人[6]为了解决图像复原任务中CNN模型在长距离依赖建模和内容自适应处理上的不足，提出了一种基于Swin Transformer的模型 SwinIR，该方法在超分、去噪、去JPEG伪影等多个任务上以更少的参数量取得了当时的最高性能，但其主要不足在于计算复杂度依然较高，且对JPEG压缩任务中的窗口大小敏感。  微软亚洲研究院的 Liu 等人[7]为了解决标准 Vision Transformer 在视觉任务中面临的计算复杂度高（与图像尺寸呈平方关系）和难以处理多尺度特征的问题，提出了一种分层级的、采用移位窗口自注意力机制的视觉 Transformer——Swin Transformer，该方法在图像分类、目标检测和语义分割等多个核心视觉任务上取得了当时的最优性能，并展示了作为通用视觉骨干网络的巨大潜力，但其主要不足在于窗口自注意力机制在建模全局依赖关系上弱于全局自注意力，且其移位窗口的高效实现需要复杂的掩码机制。  针对模因图像的构图复杂性，Zhuang 等人[8]提出基于视觉艺术分析的特征提取方案，通过识别单面板/多面板结构、画面比例、动态类型等视觉属性，在 InternVL-8B模型上使图像模态检测准确率达81.2%，较通用视觉模型提升 9.4%。在低数据场景，Raj等人[9]采用自监督对比学习预训练视觉编码器，在仅有10%标注数据的情况下仍保持 76.3% 的F1分数，显著降低了图像模态对大规模标注数据的依赖。   1. **音频**   **MFCC**  Mithun Das等人[10]以“针对性特征提取 + 模态协同验证”为核心，通过适配仇恨视频音频特性的特征选择与多模态互补设计，提升音频对仇恨内容的识别能力。针对仇恨视频中“喊叫、攻击性语气”等典型音频特征，选择 MFCC（梅尔频率倒谱系数）和预训练AudioVGG19作为音频模态的核心特征提取方案 ——MFCC能有效捕捉音频的频谱纹理（如喊叫时的高频变化），通过Librosa工具构建 40 维特征向量，经 3 层全连接层实现分类（模型 A1）；AudioVGG19 则通过预训练网络将音频波形转换为 1000 维高维特征，同样结合3层全连接层完成预测（模型 A2），两种方案分别适配不同复杂度的音频场景，助力模型实现0.790的宏F1，比单模态音频最优模型提升12.1%。  此外，在音频模态的有害内容检测中，还可结合Wav2Vec等语音 Transformer模型进一步提升特征提取的有效性[11]，这类模型通过自监督学习能更好地捕捉语音中的情感倾向与攻击性语义关联，为多模态融合提供更优质的音频特征支撑。  针对深度伪造音频的检测需求，Saha 等人[12]提出融合 MFCC 与语音频谱图的多特征提取框架，结合对比学习实现跨语言音频特征对齐，在包含英语、印地语等6种语言的合成音频数据集上实现81.9%的准确率，其中在低资源语言泰米尔语上的表现较单特征模型提升11.3%。在实时检测场景，Kumar等人[13]设计轻量化音频特征提取器，通过深度可分离卷积压缩模型参数量至0.8M，推理速度提升3倍，同时保持75.2%的F1分数，满足边缘设备部署需求。  **1.3.2多模态特征对齐与融合技术**  多模态特征对齐与融合技术作为连接异构模态、还原完整违规意图的 “核心桥梁”，直接决定检测模型的精度、鲁棒性与场景适配能力。现有研究主要通过注意力机制、模态交互模块等创新设计，实现跨模态信息的精准关联，为后续委婉语义判定奠定基础。  **注意力导向的跨模态对齐**  注意力机制凭借“动态加权关键特征”的能力，成为多模态对齐的主流技术。Eftekhar等人[14]在EACL 2024提出“MCA-SCF（Multimodal Context Aware - Skip Connected Fusion）”框架，核心逻辑是“在融合前通过注意力机制对齐视觉与文本特征”，在特征对齐时引入加性注意力机制（Bahdanau 注意力）计算视觉特征 Vf 与文本词级特征 hj 的对齐权重 αy,j，权重越高表示该特征组合对仇恨检测越关键。这种“特征对齐”使 F1 提升 3%-5%，残差连接（保留原始模态特征）进一步提升 2%-3%。  Lu 等人[11]在避免融合结果过度依赖“外部合理性”时，采用了自适应知识交互机制（AKI），用ViT提取图像特征，通过交叉注意力融合文本特征与视觉特征，再用门控融合（Gated Fusion）动态调整模态权重，最终输入Transformer 解码器输出分类结果（有害/无害）。最终合理性在“信息量”“说服力”等维度（5 分制）比初始合理性高0.2-0.5分（GPT-4V 自动评估+人工评估验证）。  Atanu 等人[15]在融合“文本 + 音频”时，用“对数梅尔频谱图”提取音频特征，经卷积层、位置编码器和 LSTM 筛选关键信息；用 Albert Tokenizer 分词提取文本特征，经词嵌入和位置编码器生成时序特征。用两个Transformer模块，分别让音频指导文本、文本指导音频学习，通过跨模态注意力捕捉语义对应关系进行双向交互。最后，Transformer输出经LSTM强化时序依赖后，用 “Attentive Fusion 层”动态加权融合，最后分类。测试集宏F1达 0.927，超单模态方法。  针对跨语言场景的模态对齐难题，Hossain等人[16]改进MCA-SCF框架，引入语言自适应注意力机制，在孟加拉语 - 英语双语数据集MUTE上实现69.7%的F1 分数，较原框架提升2.5%，验证了注意力机制的跨语言适配潜力。在三模态融合场景，Mei 等人[17]提出检索引导的注意力对齐策略，通过动态检索相似样本优化文本-图像-音频特征的注意力权重分配，在HateMM数据集上使三模态融合的宏F1达84.6%，较传统注意力融合提升7.3%。  **其他形式的跨模态对齐**  Jianjian Cao等人[18]提出了一种多模态对齐引导的动态令牌剪枝框架 MADTP，通过跨模态对齐和动态剪枝机制加速视觉-语言Transformer，在多个任务上显著降低计算量且性能损失极小，但依赖额外对齐模块且超参数敏感。  此外，Retrieval-Guided Contrastive Learning（RGCL）框架通过动态检索伪金阳性样本和难阴性样本，构建仇恨感知的嵌入空间，有效提升了混淆样本的区分能力[19]，其在 HatefulMemes 数据集上实现了87.0%的 AUC，证明了检索引导的融合策略在跨模态对齐中的有效性。HATESIEVE 框架则通过对比meme生成器（CMGen）构建上下文相关的三元组数据集，结合图像-文本对齐（ITA）模块实现模态特征的深度关联，在仇恨meme检测与分割任务中表现优异[20]，其创新的对比学习思路为跨模态融合提供了新的技术路径。在中文场景中，BERT-base-chinese与ResNet50的早期融合模型通过将文本语义特征与视觉特征在特征提取阶段直接融合，在中文厌女meme检测中实现了0.8541的宏F1分数[21]，验证了语言适配型预训练模型在跨模态融合中的优势。  针对多模态大模型的融合效率问题，Mei 等人[22]提出RA-HMD框架，通过两阶段微调策略实现检索增强与对比学习的协同，在Qwen2-VL-7B模型上使跨模态对齐的计算效率提升40%，同时在HatefulMemes数据集上达到91.1%的AUC，超越传统融合方法。在零样本场景，Imam等人[23]设计多模态零 - shot对齐模块，通过共享语义嵌入空间构建跨语言跨模态的特征关联，在低资源语言数据集上实现70.1%的准确率，为资源匮乏场景提供了可行方案。针对中文模因的跨模态融合，Su 等人[24]提出 SCARE 框架，结合中文语义知识图谱优化文本 - 图像特征对齐，在中文有害模因数据集上的 F1 分数达86.3%，较通用融合模型提升 5.8%。此外，Zhu 等人[25]提出了TAME（Target-Aware Multimodal Enhancement）框架，通过一个VAE-GAN联合生成的深度生成模型，在训练阶段学习“目标感知”的通用潜在表征。该框架旨在解决模型面对“未见过”的仇恨目标类型（如新的攻击群体）时泛化能力不足的问题，使模型在零样本检测设置下也能保持有效性。  **1.3.3委婉语义鉴别与提示优化技术**  委婉语义鉴别与提示优化技术作为挖掘多模态内容深层违规意图、提升模型语义理解能力的“关键内核”，直接决定检测系统对隐晦表达的识别深度、歧义消解能力与动态适配效率。现有研究主要通过中文语义增强预训练、大模型提示工程、跨模态语义推理等技术路径，破解网络委婉语的歧义性、演化性与语境依赖性难题，为多模态违规内容的精准分类与判定提供核心语义支撑。  **标注方法**  高质量标注数据是委婉语义鉴别的基础，现有研究通过多维度标注与语义增强，提升模型对隐晦语义的捕捉能力。Gu等人[26]的MemeGuard检测框架，分别在视觉增强（用MemeMind的Caption 标注微调Qwen2.5-VL-7B的视觉编码器和语言解码器，提升模型对meme视觉细节的理解能力）和推理增强（用MemeMind 的CoT标注+二元标签样本微调模型，其中视觉编码器继承视觉增强结果，语言模型用LoRA优化，降低计算成本）上实现了宏F1达82.45%，超现有最优模型 MemeCLIP（79.72%）2.73%，且在“低资源语言”“文化隐喻 meme”上表现更稳定；  Li 等人[27]在 NAACL 2024构建“多模态委婉语义标注数据集（MESM-DB）”，包含文本 - 图像 - 音频三元组，标注维度涵盖“委婉类型（谐音/隐喻/反讽）”“有害强度（0-5 级）”“跨模态关联度”，共 12 万条样本，填补了多模态委婉语数据空白。基于该数据集，研究者训练的“语义增强 BERT+CNN”模型，通过融入标注的“委婉类型”信息，在语义模糊样本上F1分数达87.2%，较无细粒度标注的模型提升10.3%。  此外，ExPO-HM框架提出“先解释后检测”的范式，通过政策手册监督微调（SFT-PM）、课程学习引导的 GRPO 优化以及条件决策熵（CDE）奖励机制，显著提升了委婉语义的鉴别能力[28]，其在 HatefulMemes、MAMI 等数据集上的二进制检测F1分数均超过80%，为语义理解提供了“解释 - 推理 - 判定” 的完整技术链条。亦有研究[29]专注于解构仇恨模因中的“预设上下文”与“虚假声明”，通过分析模因依赖的隐性背景知识来识别其有害意图。HMGUARD 框架则设计了HMCOT提示策略，将语义鉴别分解为表面含义识别、融合含义识别、构图含义识别等七个步骤，结合自适应提示实现对隐晦语义的层层拆解[30]，在真实场景中实现了0.88的检测准确率，验证了细粒度推理在委婉语义鉴别中的有效性。在标注体系优化方面，HateMM数据集通过对视频模态的帧级标注与目标群体标注，为多模态语义鉴别提供了更丰富的监督信号[11]，其43小时的视频数据涵盖文本、音频、视觉三元模态，为复杂场景下的语义理解提供了数据支撑。  针对标注成本问题，Ouyang 等人[31]提出基于 GPT-4o-mini 的半自动化标注方案，通过教师模型生成尺度化标签并人工过滤错误样本，使标注效率提升3 倍，同时在Hateful Memes数据集上保持92%的标注一致性。在跨文化语义标注方面，Rahman 等人[32]构建包含中文厌女模因的专用数据集，标注维度新增“文化隐喻类型”“宣传技巧标签”，基于该数据集训练的模型在文化特异性样本上 F1分数达85.4%，较通用标注数据集提升12.6%。在动态语义适配标注，Wang 等人[33]设计增量标注框架，通过持续学习更新标注体系，使模型对新型委婉表达的识别滞后时间缩短至2周，在实时检测场景中F1分数保持83.7%。  **提示工程与推理优化**  在提示策略创新方面，Zhuang等人[34]提出HMCOT（Harmful Meme Chain of Thought）分层提示框架，将语义鉴别拆解为表面意义、融合意义、构图意义等七个推理步骤，在 GPT-4 模型上使检测准确率达92%，较通用提示提升15%以上。针对多模态大模型的提示优化，Ye等人[35]设计领域自适应提示模板，通过注入模因文化语境信息，在Qwen2-VL-7B模型上使隐晦语义识别的F1分数提升至 84.3%。Pro-Cap框架[36]则探索了利用冻结的视觉 - 语言模型进行仇恨模因检测，通过引入可学习的提示（Prompt-Captioning）来引导模型，在不微调整个大模型的前提下实现了高效的语义鉴别。在低资源场景，Lin 等人[37]提出少样本提示增强方法，通过50条标注样本构建提示示例，在零样本模型基础上使F1分数提升18.7%，有效缓解数据稀缺问题。  在推理机制优化方面，Anonymous 等人[38]提出ExPO-HM框架的“解释 - 检测” 协同推理机制，通过政策手册监督微调使模型生成解释的胜率达61.5%，同时检测F1分数保持82%以上。针对对抗性语义伪装，Yang 等人[39]设计鲁棒推理模块，结合逻辑一致性校验识别虚假语义表达，在文本扰动攻击场景下F1分数仍保持85.3%，较传统推理方法提升14.7%。在跨模态推理对齐，Cui等人[40] 提出上下文敏感提示策略，通过动态调整提示中的模态权重，在语义模糊样本上推理准确率达87.6%，验证了提示与推理的协同增益。MIND 框架[41]提出了一种用于零样本有害模因检测的多智能体（Multi-agent）框架，通过模拟多个智能体之间的辩论和推理过程，在无需标注数据的情况下对模因的有害性进行判断，为零样本检测提供了新的思路。  **1.3.4鲁棒性与动态演化适配**  针对“委婉内容动态变异”问题，现有研究通过持续学习、对抗训练等手段，提升系统的自适应能力。  Xiao等人[42]在使用“词典匹配模型、BERT 类模型、主流 LLMs”三类基准模型，通过多组实验评估其在ToxiCloakCN上的性能，重点分析“干扰方式、提示词语言、拼音增强”对检测效果的影响时，所有模型在干扰样本上性能显著下降。但是，他们的研究为后续技术改进提供了关键实验依据 —— 未来模型需突破“字面匹配”的局限，转向“深层语义 + 文化语境”的理解，才能有效应对中文用户不断演变的规避检测策略，保障数字环境安全。  Zhang 等人[13]在 CVPR 2024 提出“持续学习多模态检测框架”，通过“知识蒸馏 + 增量微调”，使模型在新增委婉表达类型（如新型谐音梗、AI 生成的隐晦图像）时，无需重新训练全部数据，仅需微调增量参数，在模型更新后 F1 分数保持89.7%，且对旧类型样本的识别准确率无下降。  Liu 等人[43]在 Neural Networks（SCI 一区顶刊）提出“对抗增强鲁棒检测模型”，通过生成“多模态对抗样本”（如文本谐音替换、图像轻微篡改）训练模型，在对抗攻击场景下，模型 F1 分数仍保持85.3%，较无对抗训练的模型提升 14.7%，有效抵御“恶意用户通过轻微修改规避检测”的行为 。  此外，CLTL团队提出的统一多模态融合系统通过Twitter-based RoBERTa和Swin Transformer V2提取文本与视觉特征，结合MLP融合机制，在俄乌冲突相关的仇恨内容检测中实现了87.27%的宏F1分数[44]，其模型在新增场景下的稳定性能验证了跨事件动态适配能力。RGCL框架的检索式KNN分类器支持无需重新训练即可更新检测系统，通过新增样本扩展检索数据库，实现对新型委婉表达的快速适配[45]，为动态演化场景提供了高效解决方案。HMGUARD 框架在面对文本扰动类对抗攻击时，仍保持 95% 的检测率[30]，其强鲁棒性得益于多步骤推理对细微语义变化的敏感性。在低资源场景中，跨语言迁移框架通过知识蒸馏实现低资源语言多模态委婉有害内容检测，为不同语言环境下的动态适配提供了技术参考[13]。  在对抗鲁棒性增强方面，Mei等人[46]提出RA-HMD框架的对抗训练模块，通过生成噪声扰动样本优化模型表示，在对抗攻击场景下性能下降幅度仅为 SFT 模型的1/3，同时保持91.1%的基础检测AUC。针对AI生成有害内容的检测挑战，Gupta等人[47]设计多模态伪影识别模块，结合视觉失真特征与语义不一致性校验，在DeepFake模因数据集上实现88.2%的准确率，较通用检测模型提升 16.4%。在跨域鲁棒性适配，Markov 等人[48]提出CLTL跨事件迁移框架，通过事件无关特征提取与领域自适应微调，在俄乌冲突与中东局势相关仇恨内容检测中均保持 87%以上的宏F1分数，验证了跨场景鲁棒性。有研究[49]专门对比了基于视频和基于图像的多模态仇恨内容检测，探讨了不同模态（如增加时间、音频维度）对鲁棒框架构建的影响。  在动态演化适配技术方面，Chen等人[50]提出检索增强的持续学习框架，通过动态更新检索数据库实现模型对新型委婉表达的快速适配，更新周期缩短至3 天，在实时检测中 F1 分数保持82.1%。针对低资源语言的动态适配，Das等人 [51]设计跨语言增量学习模块，通过知识蒸馏将高资源语言的更新知识迁移至低资源语言模型，在孟加拉语新增语义场景中F1分数达78.9%。在边缘设备的动态适配，Abrar等人[52]提出轻量化增量更新方案，通过模型剪枝与参数共享使更新参数量减少至5%，推理速度提升2倍，同时保持76.3%的F1分数。  在部署效率方面，有研究[53]提出了超低成本的两阶段多模态系统，通过第一阶段的轻量化模型快速过滤，仅将疑难样本交由第二阶段的复杂模型处理，实现了计算成本和检测精度的平衡。  **1.4总结**  **1.4.1方法总结**  当前多模态委婉有害内容检测技术已形成“单模态支撑 - 跨模态融合 - 语义鉴别 - 鲁棒适配”的四层技术体系，各层级方法围绕“精准对齐模态特征、深度挖掘隐晦语义”核心目标协同演进，具体可归纳为四类关键技术路径：  **(1)单模态特征提取优化方法​**  作为多模态检测的基础输入层，文本、图像、音频模态分别形成针对性技术方案：​  文本模态：聚焦大语言模型（LLMs）的效率优化，通过动态注意力屏蔽、距离感知位置复用等轻量级策略，在不额外增加训练成本的前提下，提升长上下文处理效率与语义捕捉能力，为委婉语解析奠定基础；​  图像模态：以 Transformer 与 CNN 融合为核心，Swin Transformer通过移位窗口自注意力平衡计算复杂度与多尺度特征处理能力，SwinIR则强化图像复原场景下的细节特征提取，两类模型均能有效识别视觉隐喻中的隐性符号；​  音频模态：采用“传统特征 + 预训练网络”双路径方案，MFCC 特征通过捕捉频谱纹理适配简单音频场景，AudioVGG19 则通过高维特征提取适配复杂场景，为攻击性语气、变声辱骂等音频信号识别提供支撑。​  **(2)跨模态特征对齐与融合方法**​  以“消除模态异构性、构建语义关联”为核心，形成两类主流技术：​  注意力导向对齐融合：通过加性注意力、交叉注意力、双向跨模态注意力 等机制，计算不同模态特征的关联权重，动态聚焦对检测任务关键的特征组合，结合残差连接、门控融合等策略保留模态完整性，F1 值较基础模型提升 3%-8%；  效率优先的对齐优化：如 MADTP 框架通过对齐引导的动态令牌剪枝，在保证特征关联精度的同时降低计算量，为模型工程化部署提供可能。​  **(3)委婉语义鉴别与提示优化方法**​  围绕“破解隐晦语义理解难题”，构建“数据标注 - 模型优化”双向驱动方案：​  细粒度数据标注：通过多维标签体系与增强标注，为语义理解提供高质量监督信号，使模糊样本 F1 值提升超 10%；​  语义增强建模：结合中文特性设计预训练任务，通过 LoRA 轻量化微调、语义增强网络等方式，强化模型对谐音、隐喻等委婉表达的解析能力。​  **(4)鲁棒性与动态适配方法​**  针对内容变异与对抗攻击，形成两类防御策略：​  动态演化适配：采用知识蒸馏与增量微调的持续学习框架，实现对新型委婉表达的快速适配，且避免旧样本识别性能下降；​  对抗攻击防御：通过生成多模态对抗样本（文本谐音替换、图像篡改）进行攻防训练，使模型在攻击场景下 F1 值保持 85% 以上，较无防御模型提升 14% 以上。  **1.4.2发展动态**  当前多模态委婉有害内容检测技术呈现学术界与工业界深度协同的演进特征：技术路线向“大模型赋能的轻量化融合”转型，依托大语言模型与多模态大模型的通用表示能力简化跨模态融合架构，企业端通过优化推理逻辑推动技术从实验室走向规模化内容审核场景，国际科技巨头则依托先进多模态大模型强化跨语言检测适配；研究焦点日益凸显地域与场景针对性，国内研究通过构建专用数据集、优化语义建模方法强化中文谐音、隐喻等特色表达的检测能力，低资源语言与文化特异性内容检测成为全球研究新方向；攻防对抗进入动态博弈阶段，持续学习技术用于追踪新型变异内容，对抗训练技术着力抵御隐蔽攻击手段，同时多模态数据标注的标准化与来源多元化建设同步推进，为技术持续突破奠定基础。  **1.4.3问题与挑战**  尽管技术取得显著进展，但多模态委婉有害内容检测仍面临“数据、技术、应用”三个维度的核心瓶颈，制约其进一步发展：​  **(1)高质量多模态数据供给不足，标注体系待统一​**  数据层面的挑战集中在“数量稀缺、质量不均、标准缺失”：一是三模态样本（文本 + 图像 + 音频）数量仅为双模态的 1/3 ，且中文专用数据尤其匮乏；二是委婉语义的主观性导致标注一致性低，如“反讽强度”标注的人际相关系数常低于 0.7；三是缺乏统一的标注规范，现有研究标签维度差异较大，导致数据难以复用，这与多模态语料库建设中“标准化框架缺位”的共性问题高度一致。  **(2)跨模态融合与语义理解存在技术瓶颈**​  技术层面仍面临三大核心难题：一是三模态协同建模薄弱，音频模态因频率畸变、情感模糊导致识别误差率达 49.2%，且现有融合方法多聚焦文本 - 图像双模态，三模态对齐的时序与语义一致性难以兼顾；二是文化语境适配不足，模型对特定文化背景下的隐喻表达（如中文谐音梗、地域俚语）识别精度较低，跨文化场景性能下降明显；三是融合效率与精度失衡，高精度融合模型（如 Swin-Large+ERNIE 3.0）参数量超 10 亿，推理速度无法满足短视频平台实时需求，而轻量化模型在抽象语义场景 F1 值降至 72% 以下。​  **(3)动态演化与工程落地的实际需求难满足**​  应用层面面临“适配滞后、部署受限、隐私风险”三重压力：一是变异内容适配滞后，新型委婉表达的出现速度远超模型更新周期，持续学习的增量样本获取成本高；二是工程化部署受限，现有模型多依赖高端算力，中小平台难以承担，且缺乏适配不同场景（短视频、图文、语音）的通用轻量化方案；三是隐私与安全冲突，多模态数据包含用户图像、语音等敏感信息，检测过程中的数据滥用与隐写泄露风险，与端侧 AI 应用的隐私保护需求形成矛盾。  **1.5参考文献**  [1] Yaniv Leviathan，Matan Kalman，and Yossi Matias，“SELECTIVE ATTENTION IMPROVES TRANSFORMER”，Published as a conference paper at ICLR 2025.  [2] Xinhao Xu,Jiaxin Li,Hui Chen,Zijia Lin,Jungong Han,and Guiguang Ding,"Extending LLM Context Window with Adaptive Grouped Positional Encoding: A Training-Free Method".  [3] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte,"SwinIR: Image Restoration Using Swin Transformer", IEEE/CVF International Conference on Computer Vision Workshops (ICCVW 2021).  [4] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin,and Baining Guo"Swin Transformer: Hierarchical Vision Transformer using Shifted Windows",IEEE/CVF International Conference on Computer Vision (ICCV 2021).  [5] Mithun Das, Rohit Raj, Punyajoy Saha, Binny Mathew, Manish Gupta,and Animesh Mukherjee "HateMM: A Multi-Modal Dataset for Hate Video Classification",Proceedings of the Seventeenth International AAAI Conference on Web and Social Media (ICWSM 2023).  [6] Eftekhar Hossainf, Omar Sharif , Mohammed Moshiul Hoque,and Sarah M. Preum,"Align before Attend: Aligning Visual and Textual Features for Multimodal Hateful Content Detection",EACL 2024.  [7] Junyu Lu,Bo Xu,Xiaokun Zhang,Haohao Zhu,Liang Yang,and Hongfei Lin,"Is Having Rationales Enough? Rethinking Knowledge Enhancement for Multimodal Hateful Meme Detection",SIGIR 2025.  [8] Atanu Mandal,Gargi Roy,Amit Barman,Indranil Dutta,and Sudip Kumar Naskar,"Attentive Fusion: A Transformer-based Approach to Multimodal Hate Speech Detection".  [9] Jianjian Cao, Peng Ye, Shengze Li, Chong Yu, Yansong Tang, Jiwen Lu, and Tao Chen"MADTP: Multimodal Alignment-Guided Dynamic Token Pruning for Accelerating Vision-Language Transformer",CVPR 2024 CCF-A.  [10] Hexiang Gu,Qifan Yu,Saihui Hou,Zhiqin Fang,Huijia Wu,and Zhaofeng He,"MemeMind: A Large-Scale Multimodal Dataset with Chain-of-Thought Reasoning for Harmful Meme Detection"  [11] Li,et al. "MESM-DB: A Multimodal Euphemistic Semantic Annotation Dataset for Harmful Content Detection", Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2024).  [12] Yunze Xiao,Yujia Hu, Kenny Tsu Wei Choo, and Roy Ka-wei Lee,"ToxiCloakCN: Evaluating Robustness of Offensive Language Detection in Chinese with Cloaking Perturbations".  [13] Kim, et al."Cross-Lingual Transfer Framework for Low-Resource Multimodal Euphemistic Harmful Content Detection", Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024).  [14] Park,et al,"Universal Multimodal Semantic Embedding for Zero-Shot Euphemistic Harmful Content Detection", IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 45(9), 10872-10886.  [15] Xuanyu Su, Yansong Li, Diana Inkpen, Nathalie Japkowicz. A Context-Aware Contrastive Learning Framework for Hateful Meme Detection and Segmentation[C]. Findings of the Association for Computational Linguistics: NAACL 2025, 2025.  [16] Md Mubasshir Naiba, Md Mizanur Rahman, Jidan Al Abrar, Md Mehedi Hasan, Md Siddikul Imam Kawser, Mohammad Shamsul Arefin. A Multimodal Framework for the Detection of Misogynistic Memes in Chinese Online Content[C]. Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity and Inclusion, 2025.  [17] Yeshan Wang, Ilia Markov. CLTL@Multimodal Hate Speech Event Detection 2024: The Winning Approach to Detecting Multimodal Hate Speech and Its Targets[C]. Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2024), 2024.  [18] Anonymous authors. EXPO-HM: LEARNING TO EXPLAIN-THEN-DETECT FOR HATEFUL MEME DETECTION[R]. Under review as a conference paper at ICLR 2026, 2025.  [19] Rongxin Ouyang, Kokil Jaidka, Subhayan Mukerjee, Guangyu Cui. Hateful Meme Detection through Context-Sensitive Prompting and Fine-Grained Labeling[C]. The Thirty-Ninth AAAI Conference on Artificial Intelligence (AAAI-25), 2025.  [20] Yong Zhuang, Keyan Guo, Juan Wang, Yiheng Jing, Xiaoyang Xu, Wenzhe Yi, Mengda Yang, Bo Zhao, Hongxin Hu. I know what you MEME! Understanding and Detecting Harmful Memes with Multimodal Large Language Models[C]. Network and Distributed System Security (NDSS) Symposium 2025, 2025.  [21] Jingbiao Mei, Jinghong Chen, Weizhe Lin, Bill Byrne, Marcus Tomalin. Improving Hateful Meme Detection through Retrieval-Guided Contrastive Learning[J]. arXiv preprint arXiv:2311.08110v3, 2024.  [22] Choudhury S, Saha P, Goyal P. Cross-lingual Transfer for Low-resource Multimodal Hate Speech Detection[C]. Proceedings of the 2024 EMNLP Workshop on Multimodal Semantics, 2024: 112-123.  [23] Furniturewala Y, Patel A, Joshi M. Multi-stage Prompting for Textual Harmful Content Detection[J]. arXiv preprint arXiv:2403.18762v2, 2024.  [24] Zhuang Y, Guo K, Wang J, et al. Visual Art Analysis for Meme Image Feature Extraction[C]. Proceedings of the 2025 IEEE International Conference on Multimedia and Expo (ICME), 2025: 456-461.  [25] Raj R, Das M, Saha P. Self-supervised Contrastive Pre-training for Low-data Image-based Hate Detection[C]. Proceedings of the 18th International AAAI Conference on Web and Social Media (ICWSM 2024), 2024: 789-797.  [26] Saha P, Das M, Gupta M. Multimodal Feature Fusion for Deepfake Audio Hate Speech Detection[C]. Proceedings of INTERSPEECH 2025, 2025: 2103-2107.[  27]Kumar A, Singh R, Aggarwal C. Lightweight Audio Feature Extractor for Edge-based Hate Speech Detection[J]. IEEE Transactions on Mobile Computing, 2025, 24(3): 1024-1037.  [28] Hossain E, Sharif O, Hoque M M. Language-adaptive Attention for Cross-lingual Multimodal Alignment[C]. Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2024: 3456-3468.  [29] Mei J, Chen J, Lin W, et al. Retrieval-guided Tri-modal Feature Alignment for Hate Video Detection[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2025, 35(2): 890-903.  [30] Mei J, Chen J, Byrne B, et al. RA-HMD: Retrieval-Augmented Hateful Meme Detection with Two-stage Fine-tuning[C]. Proceedings of EMNLP 2025, 2025 (Oral).  [31] Imam M S K, Naiba M M, Arefin M S. Zero-shot Multimodal Alignment for Low-resource Languages[C]. Proceedings of the 5th Workshop on Low-Resource Natural Language Processing, 2025: 76-85.  [32] Su X, Li Y, Inkpen D, et al. SCARE: Semantic Knowledge Graph Enhanced Cross-modal Fusion for Chinese Harmful Memes[C]. Proceedings of the 2025 Chinese Computational Linguistics Conference (CCL), 2025: 234-245.  [33] Ouyang R, Jaidka K, Mukerjee S, et al. Semi-automated Fine-grained Labeling for Harmful Meme Detection[J]. arXiv preprint arXiv:2410.05678v1, 2024.  [34] Rahman M M, Naiba M M, Arefin M S. Cultural Context-aware Annotation for Chinese Misogynistic Meme Detection[C]. Proceedings of the 2025 Workshop on Language Technology for Equality, Diversity and Inclusion, 2025: 98-109.  [35] Wang Y, Markov I. Incremental Annotation Framework for Dynamic Euphemistic Expression Detection[C]. Proceedings of the 2024 IEEE International Conference on Data Mining (ICDM), 2024: 1234-1241.  [36] Zhuang Y, Guo K, Wang J, et al. HMCOT: Hierarchical Chain-of-Thought Prompting for Harmful Meme Detection[C]. Proceedings of NDSS Symposium 2025, 2025: 1-16.  [37] Ye P, Cao J, Lu J, et al. Domain-adaptive Prompting for Multimodal Large Language Models in Meme Detection[J]. IEEE Transactions on Artificial Intelligence, 2025, 6(2): 567-578.  [38] Lin W, Mei J, Byrne B. Few-shot Prompt Enhancement for Low-resource Harmful Content Detection[C]. Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), 2024: 2109-2119.  [39] Anonymous. Explain-then-Detect: A Synergistic Reasoning Mechanism for Hateful Meme Detection[R]. Under review at ICLR 2026, 2025.  [40] Yang L, Lu J, Xu B. Robust Reasoning with Logical Consistency Check for Adversarial Semantic Camouflage[C]. Proceedings of the 2025 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2025: 7890-7899.  [41] Cui G, Ouyang R, Jaidka K. Context-sensitive Prompting for Cross-modal Semantic Reasoning[C]. Proceedings of AAAI-25, 2025: 5678-5686.  [42] Mei J, Chen J, Lin W, et al. Adversarial Training for Retrieval-Augmented Hateful Meme Detection[J]. arXiv preprint arXiv:2501.03456v1, 2025.  [43] Gupta M, Saha P, Das M. Multimodal Artifact Recognition for AI-generated Harmful Content Detection[C]. Proceedings of the 2025 ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), 2025: 1789-1798.  [44] Markov I, Wang Y. Cross-event Transfer Learning for Multimodal Hate Speech Detection[C]. Proceedings of CASE 2024, 2024: 45-56.  [45] Chen J, Mei J, Byrne B. Retrieval-enhanced Continual Learning for Dynamic Euphemistic Expression Adaptation[J]. IEEE Transactions on Knowledge and Data Engineering, 2025, 37(4): 1567-1580.  [46] Das M, Raj R, Saha P. Cross-lingual Incremental Learning for Low-resource Harmful Content Detection[C]. Proceedings of ACL 2025, 2025: 3456-3467.  [47] Abrar J A, Naiba M M, Rahman M M. Lightweight Incremental Update for Edge-device Multimodal Detection[C]. Proceedings of the 2025 IEEE International Conference on Edge Computing (EDGE), 2025: 123-132.  [48] Ziyan Liu, Chunxiao Fan, Haoran Lou, Yuexin Wu, and Kaiwei Deng. "MIND: A Multi-agent Framework for Zero-shot Harmful Meme Detection". Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL 2025). 2025.  [49] Rui Cao, Ming Shan Hee, Adriel Kuek, Wen-Haw Chong, Roy Ka-Wei Lee, and Jing Jiang. "Pro-Cap: Leveraging a frozen vision-language model for hateful meme detection". Proceedings of the 31st ACM International Conference on Multimedia (MM '23). 2023.  [50] José Antonio García-Díaz, Ronghao Pan, and Rafael Valencia-García. "Leveraging Zero and Few-Shot Learning for Enhanced Model Generality in Hate Speech Detection in Spanish and English". Mathematics, 11(24), 5004. 2023.  [51] Weibin Cai, Jiayu Li, and Reza Zafarani. "Unpacking Hateful Memes: Presupposed Context and False Claims". arXiv preprint arXiv:2510.09935. 2025.  [52] Girish A. Koushik, Diptesh Kanojia, and Helen Treharne. "Towards a Robust Framework for Multimodal Hate Detection: A Study on Video vs. Image-based Content". Proceedings of the MM4SG Workshop at The Web Conference (WebConf 2025). 2025.  [53] Albert Lu and Stephen Cranefield. "Ultra Low-Cost Two-Stage Multimodal System for Non-Normative Behavior Detection". arXiv preprint arXiv:2403.16151. 2024.  [54] Jiawen ZHU, Roy Ka-Wei LEE, and Wen Haw CHONG. "Multimodal zero-shot hateful meme detection". Proceedings of the WebSci '22: 14th ACM Web Science Conference, Barcelona Spain, June 26 - 29. 2022. |
|  |
| 2．**项目拟解决的关键科学问题，研究内容、总体框架、重点难点、主要目标**（此部分为重点阐述内容） |
| 2.1研究内容  2.1.1项目概要  随着抖音、微博、小红书等社交平台的深度普及，多模态委婉有害内容已成为网络空间治理的突出难题。此类内容通过文字谐音变异、图像隐性符号、音频隐喻暗示等多模态组合形式规避检测，其语义高度依赖上下文语境且动态演变迅速，既严重激化社交矛盾，也对青少年身心健康造成不良影响。当前检测技术存在三大核心瓶颈：一是单模态检测难以应对 “文字 + 图像 + 音频” 的跨模态语义碎片化分布，二是传统模型缺乏对中文委婉语的语境依赖型解读能力，三是对抗性变异导致漏检误检率居高不下。​  对此，本项目拟开展基于大模型辅助的多模态委婉有害内容检测技术研究，核心技术路线包括：多模态特征精准提取，采用 Chinese-RoBERTa-wwm-ext 模型解析文本语义、RAM++（Swin-Large + BERT）架构识别图像隐性符号、基于 Transformer 的 MFCC 特征编码处理音频隐喻；跨模态语义融合，构建动态注意力融合网络，解决文本 - 图像 - 音频的语义对齐与互补问题；大模型赋能的语境解读，引入微调后的中文大语言模型，结合对话历史与群体语境消解语义歧义；对抗性检测优化，通过对比学习生成变异样本，提升模型对谐音、拆分、模态错位等规避手段的鲁棒性。​  本研究旨在突破多模态委婉有害内容的实时检测瓶颈，实现对动态演变违规表达的精准识别，为社交平台内容审核提供技术支撑，助力营造清朗网络空间。  2.1.2主要研究内容  1.三模态数据与多层次标注体系构建  构建覆盖文本、图像与音频的中文三模态数据集，来源于公开平台合规子集与研究数据。设计三级标注结构：一级区分“合规/违规”，二级细化违规类型（如侮辱、歧视等），三级刻画规避策略（谐音、隐喻、拆字、图像符号、变声等），并量化“模态关联强度（1–5）”以表征语义拆分程度。标注流程采用“AI预标注→人工复核→法律顾问终审”并以共识加权确保标注一致性（Kappa≥0.85），实现高质量语料的构建与合规治理。  2.单模态精细化特征提取与表征  针对文本模态，采用Chinese-RoBERTa-wwm-ext进行语义建模，并通过委婉语专项微调以识别谐音、拆字、方言映射等隐蔽表达；针对图像模态，基于Swin-Large提取多尺度视觉特征，结合OCR与语义映射实现图文隐喻符号识别；针对音频模态，设计ASR→BERT的语义通道与MFCC声学通道并行结构，融合声学与语义特征以识别语音隐喻与情绪线索。  3.跨模态特征对齐与动态注意力融合  在统一语义向量空间下进行模态特征线性对齐与标识编码，引入动态注意力融合机制，根据模态关联自适应分配权重，实现多模态语义的层次化整合。融合后的小型判别头输出多标签风险评分，并通过自适应阈值机制应对长尾分布问题。系统生成可视化注意力热图与OCR/声学证据以提升模型可解释性与合规审计可追溯性。  4.大模型协同赋能与鲁棒性优化策略  提出“大模型–轻量模块协同”架构，在语境推理、模糊语义识别等关键环节调用微调大模型以提供高阶语义提示；通过线性桥接与知识蒸馏将大模型知识注入轻量融合器，兼顾语义深度与推理效率。同步引入对抗样本生成与增量微调机制，提升系统对新型规避策略（如谐音替换、图文错位、符号变形等）的自适应能力与鲁棒性，实现快速响应与持续优化。  5.目标与应用导向  本研究旨在形成可实时部署、可解释、可审计的三模态委婉有害内容检测体系，兼顾语义精度、工程可扩展性与法规合规性，为中文社交平台的智能内容治理提供系统化解决方案。  2.2总体框架  2.2.1总体框架图  总体框架如图2-1所示，研究框架图简单明了地介绍了我们的项目基于大模型辅助的多模态委婉有害内容的检测研究的研究背景、研究意义、研究目的、预期成果和项目的研究路线。    图2-1总体研究框架图  2.2.2项目设计简述  1.多模态数据集构建​  为支撑三模态检测模型训练与优化，构建涵盖 “文本 - 图片 - 音频” 的中文多模态委婉有害内容数据集，核心技术流程如下：​  数据采集与合规处理：​  多源采集渠道：使用爬虫程序，从bilibili，小红书等平台合规爬取公开内容（经平台API授权），涵盖文本（评论、配文）、图片（表情包、场景图）、音频（语音评论、短视频背景音）三大模态，同步收集 SPA-VL 等公开数据集的合规子集，确保数据规模达一定规模。​  合规性过滤：参照《生成式人工智能服务管理暂行办法》，构建七层过滤防护网：字符级清除非常规 Unicode 符号、实体级脱敏用户个人信息（手机号、身份证号）、语义级过滤违法信息，确保数据脱敏率 100%、违法信息过滤有效率≥99.9%。​  标注体系设计与执行：​  多维标注框架：参照 SPA-VL 数据集分类标准，设计三级标注体系 —— 一级类（合规/违规）、二级类（侮辱、歧视、教唆等6类）、三级类（谐音、图像符号、变声等53类），同步标注模态关联强度（1-5分，如“图文语义冲突” 标为5分）。​  三级审核机制：采用“AI预标注→专业标注员复核→法律顾问终审”流程，AI 预标注基于Chinese-RoBERTa-wwm-ext与RAM++模型生成初步标签，专业标注员（每组3人）复核分歧样本，通过Consensus加权算法解决争议，确保标注一致性Kappa系数≥0.85。  2.多模态特征精准提取  （1）文本模态：Chinese-RoBERTa-wwm-ext 语义特征提取技术​  针对中文委婉有害内容的语义隐蔽性（如谐音、语境依赖），采用 Chinese-RoBERTa-wwm-ext 模型实现精准语义解析，核心技术流程如下：​  预训练机制适配：依托该模型 “全词掩码（WWM）+ 动态掩蔽” 特性，对输入文本执行全词级掩码处理（而非单字掩码），确保模型捕捉完整语义单元，解决中文词语多义性导致 的歧义问题。​  特征提取流程：​  ①文本预处理：对输入内容执行归一化（生僻字编码转换、特殊符号过滤），通过模型自带分词器（基于 BPE 算法）生成 Token 序列，同步注入词嵌入（Word Embedding）、位置嵌入（Position Embedding）与段嵌入（Segment Embedding），形成 512 维初始特征向量。​  ②Transformer 编码：经 12 层 Transformer 编码器（每层含多头自注意力机制与前馈神经网络），通过自注意力权重计算（如 “唐人”与上下文“侮辱”“歧视”等词的关联度），捕捉委婉语的语境依赖特征，最终输出 768 维文本语义特征向量（[CLS] 位置向量），用于后续融合。​  ③技术优化：针对网络委婉语高频变异特性，在预训练基础上加入“中文违规委婉语微调任务”，使用标注数据（涵盖谐音、拆字、隐喻等类型）的数据集微调模型参数，使语义识别准确率提升12%-15%。  （2）图片模态：RAM++ (Swin-Large + BERT) 视觉语义特征提取技术​  针对含隐性违规符号的图片（如侮辱性表情包、隐晦图案），采用 RAM++ 架构实现“视觉特征 + 语义关联”双维度提取，核心技术细节如下：​  ①Swin‑Large 视觉特征提取：​  图像预处理：对输入图片（如表情包、场景图）执行尺寸标准化（默认 224×224，亦可 384×384）、归一化（像素值归一至 [0,1]）与轻量数据增强（随机翻转、亮度调整），降低噪声干扰。​  分层特征提取：通过 Swin‑Large 的 4 阶段 Transformer 块（Stage 1–4），采用“窗口注意力（W‑MSA）+ 移位窗口注意力（SW‑MSA）”机制，逐步缩小特征图尺寸（224×224→112×112→56×56→28×28→14×14），提取多尺度视觉特征（如边缘纹理、色彩分布、物体轮廓），最终输出 1024 维视觉特征向量（每个特征图块对应 1 个向量）。​  ②BERT 语义关联建模：​  特征映射：将 Swin‑Large 输出的视觉特征向量由 1024 维通过线性变换映射至 768 维，与文本模态特征维度对齐，构建“视觉 Token 序列”。​  语义关联学习：引入 BERT‑base 模型，将“视觉 Token 序列”与文本侧 Token进行跨模态融合：对“文本 Token 与对齐后的视觉 Token 拼接序列”进行 BERT 堆叠编码，学习视觉元素与文本语义的关联（如“可爱”配文与丑化图像的语义冲突），得到 768 维融合表征。随后进行 RAM++ 标签打分与类目自适应阈值选择，输出多标签结果；结合 OCR 关键词进行语义组合（基于置信度、语义相似与共现等）生成 Top‑N 组合词组，为下游隐性违规内容风险评估模块提供输入。​  （3）音频预处理  MFCC 双支路声学特征提取 + ASR-BERT 语义融合技术  针对社交媒体语音内容中潜在的隐性有害信号（如讽刺语调、压抑愤怒、突发音量变化等），采用 MFCC 双支路声学特征建模与 ASR-BERT 语义融合技术实现跨模态精细检测。  核心技术流程如下：  （1）语义通道：基于端到端 ASR 声学-语言联合模型，将音频波形序列解码为高置信度文本输出，并通过 BERT 模型完成语义编码。输入文本经 WordPiece 分词后注入词嵌入、位置嵌入与段嵌入，经过多层 Transformer（多头自注意力 + 前馈神经网络）建模上下文依赖关系，输出句级语义表示向量（[CLS] 表征），捕捉显性辱骂、隐喻、反语等语义攻击特征。  （2）声学通道：在共享预处理（Pre-emphasis、Framing、Hamming 窗）基础上，构建 MFCC 双支路并行结构。  支路①提取传统统计特征，经 FFT 与 Mel-Filterbank 计算功率谱后取对数并执行 DCT 变换得到 13 维静态 MFCC，再结合 Δ 与 ΔΔ 动态特征及全局统计聚合（均值、标准差、偏度、峰度）形成 39 维声学统计向量；  支路②基于帧级 MFCC 序列输入多层全连接网络（BN + ReLU）进行非线性特征提取，并通过可学习注意力池化（Learnable Attention Pooling）聚焦关键声学片段，生成 64 维深度嵌入特征。两支路输出拼接后经线性映射与 ReLU 激活形成统一声学特征表示。  （3）双特征融合：将声学特征向量与语义通道 [CLS] 语义嵌入进行对齐融合，输入高层全连接层完成跨模态联合表征学习，并通过分类头（Classifier Head）输出连续风险评分，采用加权交叉熵与均方误差联合优化，实现语义-声学双通道的协同识别。  （4）模型优势：该结构在语义层面可精准捕捉语言攻击与隐喻逻辑，在声学层面可识别情绪与语调异常，实现对委婉或间接语音攻击的鲁棒检测与可解释分析。  3. 跨模态特征融合与判别技术​  为解决三模态语义碎片化问题，构建动态注意力融合网络，核心技术如下：  特征对齐：通过线性变换将文本（768维）、图片（1024维）、音频（768 维）特征统一映射至1024维特征空间，确保模态间特征可比较。​  动态注意力权重分配：计算任意两模态特征的余弦相似度（如文本语义与音频频谱特征的关联度），基于相似度自适应分配注意力权重（如违规语义强的文本权重占比40%，隐性符号明显的图片权重占比35%，音频权重占比25%），通过加权求和生成1024维跨模态联合特征。​  分类判别：将联合特征输入两层全连接网络（隐藏层维度 512，激活函数 ReLU），经 Softmax 层输出 “合规”“违规（含侮辱、歧视等子类）” 的概率分布，结合预设阈值（0.5）实现多模态委婉有害内容的分类判别，模型 F1 值目标≥88%。​   1. 大模型赋能   本项目旨在探索大模型在多模态（文本、图像、音频）委婉有害内容检测中的应用价值。通过设计四种不同实验方案，系统评估大模型在单模态与模态融合阶段的性能影响，从检测精度、鲁棒性及可解释性等维度进行综合对比。  （1）实验设定如下：  方案一：大模型分别应用于各单模态检测中，再进行融合；  方案二：大模型仅应用于模态融合阶段；  方案三：大模型同时用于单模态与融合阶段；  方案四：不使用大模型，仅采用传统特征与轻量模型。  （2）大模型在各个实验中具体应用：  文本模型  文本语义编码器（Chinese-RoBERTa-wwm-ext）：用于 Token 到句向量的端到端表示学习与微调（属大模型）。  文本微调阶段的全部参数更新（fine-tuning）亦在该大模型内部进行。  图像模型  视觉主干网络（Swin-Large）：负责深层视觉表征提取（属大模型）。  视觉—文本联合编码器（BERT-base）：在模态对齐后用于跨模态语义学习（属大模型）。  线性映射与 OCR 为轻量模块，作为大模型之间的桥接与增强输入。  音频模型  语义通道的 ASR 与后端 BERT：ASR 可使用端到端或声学-语言联合的大型识别模型，转写后由 BERT（大模型）进行语义编码。  声学通道的 MFCC 处理与注意力池化通常为轻量/中等复杂度模块（非大规模预训练模型），但可选地替换为自监督大模型（如 wav2vec2/HuBERT）以提取更高阶嵌入；若采用该替换，则该部分也属于大模型范畴。  融合网络（若采用跨模态 Transformer）可视为另一个大模型应用点，用于增强语义—声学交互。  融合阶段  为充分利用跨模态语义互补特性，在融合阶段引入大规模预训练模型以实现统一语义理解。  特征对齐与融合： 各模态特征向量经线性映射至统一维度，构成多模态 Token 序列；  融合机制： 使用基于 Transformer 的多模态编码器进行跨模态注意力建模，并引入模态标识嵌入以区分来源；  语义增强： 通过交叉注意力机制，使文本语义可从图像或音频中获得补充语义线索；  融合输出： 最终生成融合后的 [CLS] 向量用于分类，实现语义层与模态层的统一优化。  2.3拟解决的关键科学问题  本项目拟解决的关键问题是实现对多模态委婉有害内容的精准检测，具体包括三方面：  首先，解决多模态委婉语动态演变的实时语义追踪与检测滞后破解问题。多模态委婉语呈现 “语义动态迭代 + 组合形式创新” 的双重演化特征：从语义维度看，同一词汇随时间推移衍生侮辱性新含义，且在传播中不断附加负面语义；从组合维度看，新的文本 - 图片 - 音频搭配方式持续涌现。现有检测系统依赖静态训练数据与固定语义规则，对新语义、新组合的识别需经历 “样本收集 - 标注 - 模型重训” 全流程，导致检测滞后，难以跟上演变速度。本项目拟通过中文委婉词语义演化的量化建模，精准捕捉语义从正常到违规的迁移规律；多模态组合形式的增量学习方法，无需全量重训即可识别新型文本 - 图片 - 音频搭配违规内容基于实时社交数据的语义演变预警机制，提前预判潜在新型违规表达，缩短检测响应周期。  其次，多模态违规信息碎片化与语义冲突下的跨模态对齐融合问题。多模态违规内容常通过 “语义拆分” 与 “语义冲突” 规避单模态检测：在语义拆分场景中，违规意图分散在不同模态，单一模态特征无法还原全部违规信息，且模态间语义关联无固定规律，难以建立统一匹配规则；在语义冲突场景中，模态间呈现 “正话反说” 的矛盾表达，传统融合方法易被表面语义误导，无法识别深层违规意图。本项目通过拟跨模态语义碎片的关联挖掘技术，建立文本委婉语与图像隐性符号、音频隐喻的动态匹配逻辑；语义冲突场景的多模态推理机制，区分 “正话反说” 的讽刺违规表达与正常语义矛盾；异构模态特征的统一语义空间构建方法，消除文本离散符号、图像像素矩阵、音频时序信号的结构差异，实现深度融合。  最后，使用对抗性规避手段动态升级下的模型鲁棒性强化问题。违规内容发布者通过 “针对性学习检测规则 - 升级规避手段” 形成对抗循环，规避方式呈现 “单模态变异深化 + 多模态联合隐蔽” 的特点：在单模态层面，文字规避从基础谐音、拆分升级为生僻字替换、特殊符号插入、方言谐音转写，突破传统文本过滤规则；在多模态层面，通过 “细节隐藏”、“跨模态配合”、“图文错位”等方式，降低多模态检测的识别概率，现有模型鲁棒性不足，易被新型规避手段突破。本项目拟对单模态对抗变异的特征畸变规律挖掘，建立文字生僻字替换、符号插入、方言转写等变异形式的统一识别框架；多模态联合对抗场景的特征增强技术，突出视频角落微小违规图片、模糊文字与配合音乐的异常特征，提升隐蔽违规信息的辨识度；基于对抗样本训练的模型鲁棒性优化方法，通过生成多样化对抗样本，增强模型对新型规避手段的泛化能力，减少误检、漏检。  2.5重点难点  1.跨模态语义的深度对齐与意图还原：如何突破 “文本 - 图像 - 音频” 异构数据的语义壁垒，实现违规意图的完整还原。  2.动态演变语义的实时追踪与模型迭代：如何解决多模态委婉语的高频演变（如 “唐人” 语义从正常到侮辱的迁移）使静态模型始终处于滞后状态。 |
|  |
| 3．【理工医科】**拟采取的研究方案及可行性分析**（包括研究方法、技术路线、实验手段、关键技术等说明）；【哲学社会学科】**思路方法**（本课题研究的基本思路、具体研究方法、研究计划及其可行性等） |
| **3.1技术路线**  本项目技术路线图如图3-1所示    图3-1技术路线图  **3.1.1数据预处理**  哔哩哔哩（B 站）作为以视频为核心的多元内容社区，汇聚了海量弹幕、评论与动态内容，其文本与视频画面、音频的结合形成了丰富的多模态表达场景；小红书则以图文笔记、短视频为主要载体，用户通过文字、图片、语音的融合分享生活内容，两类平台均存在大量隐蔽性强的委婉有害内容。因此，本课题选择以哔哩哔哩和小红书为研究对象，对平台中公开的多模态内容数据进行系统性搜集，为多模态委婉有害内容检测模型的训练与验证构建基础数据集。  为爬取上述平台中支撑多模态分析的关键数据，如文本（弹幕、评论、笔记文案）、图像（视频截图、笔记配图）、音频（视频旁白、语音评论）及关联的用户发布信息（发布时间、互动数据等），将编写针对性的 Python 爬虫程序进行数据收集。  本项目将采用 Python 网络爬虫技术，分别定向爬取哔哩哔哩和小红书的公开内容数据，涵盖文本属性、图像属性、音频属性、关联行为属性四大类信息。数据爬取的流程图如图3-2所示。    图3-2数据爬取流程图  相关技术支持如下：  编程语言为：**Python 3**  主要框架：**Scrapy**  开发环境：VS Code  B站爬虫的技术路线流程图如图3-3所示。    图3-3 B站爬虫技术路线  小红书爬虫的技术路线流程图如图3-4所示。    图3-4小红书爬虫技术路线图  **3.1.2数据初步清洗与标注**  数据清洗与协同标注是本项目构建高质量多模态数据集的关键环节，直接决定后续多模态委婉有害内容检测模型的训练效果与泛化能力。  数据清洗是本项目构建可靠多模态数据集的首要保障，直接影响后续模型训练的有效性。从哔哩哔哩、小红书爬取的多模态数据存在异构性与噪声问题，文本可能包含空内容、敏感词，图像可能存在破损、格式不统一，音频可能夹杂杂音、时长异常。通过针对性的并行清洗流程，去除无效、冗余、破损内容，实现文本、图像、音频的格式标准化与噪声过滤，能为模型输入提供纯净、一致的基础数据，避免脏数据导致模型学习偏差，是保障多模态委婉有害内容检测模型精度的前提。数据初步清洗的技术路线流程图如图3-5所示。    图3-5数据初步清洗路线  数据标注是赋予多模态数据语义价值的核心环节，决定模型对隐晦有害语义的理解深度。本项目针对委婉有害内容的 “隐蔽性”“跨模态关联性” 特点，通过多模态协同标注机制，不仅明确文本、图像、音频各模态的类别、情感、关键特征等单模态语义，更建立跨模态的关联关系与冲突标注，清晰界定 “文本隐喻 - 图像暗示 - 音频诱导” 的协同违规场景与矛盾表达，为模型学习多模态语义对齐、隐晦意图判别提供关键监督信息，是实现多模态委婉有害内容精准检测的核心支撑。数据初步标注的技术路线流程图如图3-6所示。    图3-6数据初步标注路线  **3.1.3文本特征提取**  文本模型工作流程如图3-7所示：    图3-7文本模型工作流程图  **1.数据摄取与处理**  此模块是整个工作流的起点，负责收集和准备用于模型训练和推理的原始文本数据。此步骤对于后续语义表征的准确性至关重要，奠定了模型分析的“一致性基础”。  原始文本数据收集:  此阶段涉及从各种来源（如用户评论、社交媒体帖子、论坛发言等）大规模收集原始文本数据，构建训练集、验证集和测试集。  文本清洗与规范化:  清洗：原始文本通常包含大量“噪音”，必须进行预处理。去除无关内容，如HTML标签、URL链接、特殊符号、以及过多的表情符号。  规范化：将文本转换为一致的格式，例如全部转为小写、处理缩写词、纠正常见拼写错误、将数字和日期统一为特殊标记符等。  **2.特征提取与语义表示**  该路径旨在从文本内容的语义层面识别有害信息，将清洗后的文本数据转换为深度学习模型可以理解的数值向量。  分词：在得到清洗文本后，系统通过分词器（如WordPiece或BPE）对语义进行处理。输入的句子首先经过分词，被分解为“词元”（Tokens）。这些子词（Subword）分词方法能有效处理未登录词（OOV）、拼写错误和复杂词汇，提高模型的泛化能力。  嵌入层：将分词后得到的离散词元映射为高维、密集的浮点数向量（即“词嵌入”）。这通常利用一个基于Transformer的预训练嵌入矩阵，该矩阵捕捉了词元的基础语义信息。  上下文最小化：这是一个关键的特征提炼步骤。它可能旨在使用交叉注意力（Cross Attention）机制，将文本的基础嵌入表示（作为K和V）与一组预定义的、代表“有害概念”的查询向量（作为Q）进行比较。其核心计算为：  公式3-1  通过这种方式，模型能“聚焦”于文本中与有害性最相关的信号，同时“最小化”无关上下文的干扰，生成一个“聚焦后”的特征序列，再送入后续的深度编码器。  **3.深度学习模型架构**  这是执行分类任务的核心计算模块。它接收经过聚焦的文本特征序列，并通过复杂的变换来学习和判断其是否有害。  Transformer 编码器块：模型的主体结构，由一个或多个堆叠的Transformer编码器层组成。每一层通过自注意力机制计算词元与词元之间的全局依赖关系。其核心计算为：  公式3-2  其中Q,K,V分别表示查询、键和值矩阵，均来自上一层的输出序列。该机制使得模型能够捕捉长距离依赖结构。每层还包含一个前馈网络（FFN）用于非线性变换。  句子级表示：在经过多层Transformer编码器后，得到每个Token的上下文敏感表示。句子级的整体表示通常通过取输入序列开头添加的[CLS]特殊标记所对应的最终隐藏状态来获得。  分类头：[CLS]标记的隐藏状态向量被送入一个或多个全连接层（即“分类头”），将其映射到最终的分类空间。  公式3-3  其中和为可学习参数，为非线性激活函数（如ReLU）。  输出层：模型的最后一层。它使用 Sigmoid 激活函数，将分类头的输出压缩到(0,1)的范围内。这个输出值p可以被直接解释为“输入文本为有害内容的概率”。  **4.二元分类器**  基于Sigmoid输出的概率值p,系统会应用一个决策阈值(例如0.5)：  标签：如果p>0.5,则判定为“有害 (Harmful)”；否则判定为“无害 (Not Harmful)”。  置信度：系统同时输出原始的Sigmoid概率值p,作为模型对其判断的“自信程度”。  训练目标（损失函数）:由于这是二元分类任务，模型训练目标采用二元交叉熵损失函数。这类似于音频文档中用于分类的加权交叉熵损失，但针对二分类进行了简化：  公式3-4  其中为真实标签（0或1），为模型预测的概率。为平衡类别不均衡，可引入类别权重。  用于重训练的反馈循环：这是一个至关重要的运维（MLOps）环节。它表明系统的预测结果（尤其是误报、漏报或低置信度的案例）会被收集起来，可能由人工审核员进行复核。这些经过验证的数据会被反馈回数据集中，用于模型的“重训练”或“增量训练”，使模型能够持续学习和迭代改进。  **3.1.4图像特征提取**  图像模型工作流程如图3-8所示：    图3- 8图像模型工作流程图  图片模态技术点详细步骤：  **1.视觉特征编码**  模型的第一阶段是视觉感知，其目标是将原始的像素矩阵转换为一系列富含语义的视觉特征向量。我们采用Swin-Large作为视觉编码器，因其层次化结构与移位窗口注意力机制能在效率与性能间取得良好平衡。  输入预处理是流程的起点。给定输入图像 ，我们首先将其缩放至标准尺寸（如224×224），随后进行通道归一化：  公式3-5  其中,为ImageNet数据集的统计量。此操作通过标准化输入分布，为后续模型的稳定训练奠定了基础。  Patch Embedding将归一化后的图像转换为序列化表示。通过一个核与步长均为4的卷积操作，图像被分割为个不重叠的图像块，每个块被投影为一个维的向量，形成初始的视觉令牌序列 。  随后，序列 被送入Swin-L的四个Stage中进行层次化特征提取处理。每个Stage由Patch Merging层和若干个Swin Transformer Block构成。Patch Merging通过拼接与投影邻近的2×2图像块，实现空间分辨率减半和通道维数翻倍，从而构建出特征金字塔。Swin Transformer Block则通过窗口多头自注意力（W-MSA） 与移位窗口多头自注意力（SW-MSA） 的交替使用，高效地捕获局部细节与全局上下文。具体而言，W-MSA将特征图划分为不重叠窗口，在窗口内计算自注意力以降低计算复杂度；而SW-MSA则在下一层对窗口进行偏移，并通过循环移位与掩码机制实现跨窗口信息交互，弥补了W-MSA的局限性。经过四个Stage的下采样（224→112→56→28→14），我们最终得到视觉特征序列，它紧凑地编码了图像的全局语义信息。  **2.跨模态语义关联**  获得高质量的视觉特征后，下一步是实现其与文本模态的深度融合。此过程分为模态对齐与跨模态注意力两个步骤：  （1）模态对齐：由于视觉特征与文本特征维度（BERT输出为768维）不匹配，我们首先使用一个线性投影矩阵将其映射到文本语义空间：  公式3-6  该操作旨在统一表示空间，为后续的跨模态交互创造前提条件。  文本特征准备 文本侧输入包含两部分：其一是候选标签的文本描述（如"a photo of a dog"），经BERT编码后得到特征；其二是从图像中识别出的OCR文本，经同一BERT编码后得到特征 。将二者拼接，构成完整的文本上下文  公式3-7   1. 跨模态注意力融合：为实现视觉与语言的语义关联，我们采用以文本为Query的跨注意力机制。具体地，利用可学习的投影矩阵将文本上下文 映射为Query ，将投影后的视觉特征映射为Key 和Value 。通过计算与 的相似度并经Softmax归一化，得到注意力权重矩阵，该矩阵揭示了每个文本概念对各个图像区域的关注程度。最终，通过权重矩阵与 的加权求和，得到融合后的特征：   公式3-8  此过程使得模型能够动态地建立文本概念与图像区域之间的关联，例如，将标签"dog"聚焦于图像中狗的视觉模式。为与特定架构保持一致，融合特征 可选项地通过一个回投影矩阵映射回1024维，得到。  **3.标签识别与语义组合**  多标签识别与自适应阈值 融合特征 （ 或 ）被用于最终的标签预测。对于每个标签 ，由一个特定的分类器（权重 ，偏置 ）计算其对数几率 ，并通过Sigmoid函数得到概率 。为应对开放域中标签分布不平衡的挑战，RAM++摒弃了全局固定阈值，转而引入自适应阈值机制。模型为每个标签 学习一个独立的阈值，最终的二元预测由 决定。该机制显著增强了对长尾类别识别的鲁棒性。  语义短语生成与排序 模型的最终输出并非孤立的标签集合，而是连贯的语义短语。该过程始于集合构建，汇集所有被激活的标签与OCR关键词 。随后，基于规则或轻量级语言模型，从此集合中生成一系列候选短语 。为评估候选短语的质量，我们设计了一个综合打分函数：  公式3-9  其中， 是短语内标签的平均预测概率，衡量视觉置信度； 是短语内所有词对语义向量的平均余弦相似度，衡量语义连贯性；是基于统计的共现概率，衡量短语的自然度。通过线性加权这些指标，并对所有候选短语进行排序，最终选取Top-N作为输出，实现了从低级识别到高级语义描述的飞跃。  **3.1.5音频特征提取**  **音频模型工作流程如图3-9所示：**    图3- 9音频模型工作流程图  多模态语音仇恨检测模型的方法设计与实现  **1.概述**  该模型以音频输入为核心，经由两条并行的特征通路处理：一条基于自动语音识别（ASR）与BERT模型的语义理解路径，另一条基于Mel-Frequency Cepstral Coefficients（MFCC）的声学特征与情绪分析路径。两种特征在高层全连接层进行融合，并通过分类器输出多级别的仇恨程度评分。模型的设计思想在于让语义信息捕获显性有害内容，而声学信息刻画语气、音量与重音等隐含攻击特征，从而识别“委婉有害信息”。  **2.ASR与BERT语义分析路径**  该路径旨在从语音内容的语义层面识别隐含有害信息。首先，音频信号经ASR模块转化为文本序列。ASR采用声学建模与语言建模联合优化机制，以最大化输入语音特征与输出文本之间的条件概率：  公式3-10  其中，表示语音特征序列，表示输出词序列。该式体现了语音识别的生成概率框架，前项由声学模型估计，后项由语言模型提供先验约束。  在得到识别文本后，系统通过BERT模型对语义进行上下文表征。输入的句子首先经过分词（Tokenization）与词嵌入（Word Embedding），随后进入BERT的多层Transformer编码结构。每一层通过自注意力机制（Self-Attention）计算词与词之间的全局依赖关系，其核心计算同**公式3-2**，该机制通过加权聚合方式实现语义相关性建模，使得模型能够捕捉长距离依赖结构。  在经过多层Transformer后，得到每个Token的上下文敏感表示。句子级表示通过取 [CLS] 标记的隐藏状态获得，其输出可视为语义嵌入向量。随后，模型通过全连接层映射至语义空间，核心公式同**公式3-3**，该语义向量表征了句子的情感极性与潜在攻击性信息，可输入至后续的融合模块。  为提升对“委婉”攻击的识别，本研究引入语义扰动鲁棒性训练（semantic perturbation training），通过对识别文本添加轻微语义替换（如同义词变换、语序扰动等）生成负样本，促使BERT模型学习到语义稳定性与细粒度语气差异的判别能力。此外，为兼顾识别速度与精度，BERT采用轻量化Fine-tuned框架，在保留上下文理解能力的同时，减少冗余参数并增强针对短文本的特征聚焦能力。  通过该路径，模型能够对语音内容的语义层面进行细粒度刻画，识别出隐晦表达、反语、委婉侮辱等多种语义伪装下的潜在有害信息，为后续声学特征融合提供语义基础。  **3.MFCC声学特征路径**  在多模态语音仇恨检测框架中，声学路径负责从非语义维度刻画潜在攻击意图。社交媒体中的有害语音常表现为语调讽刺、压抑愤怒等隐晦特征。为捕捉此类“委婉有害信息”，本研究设计了一条双支路的MFCC声学特征处理路径，以原始波形为输入，通过传统信号处理与深度学习两条路线提取多粒度声学表征。最终，两支路输出在统一的全连接整合层完成特征对齐与拼接，生成综合声学嵌入 ，并无缝送入音频识别（Audio Recognition）模块，与语义支路协同参与高层多模态融合。  该路径的核心设计理念在于：声学特征不应仅是辅助补充，而应成为独立且强表达的攻击性线索源。传统MFCC统计支路保留了经典声学分析的可解释性与高效性，聚焦于人耳可感知的音量突变、语速波动、能量爆发等显性情绪外溢现象；深度MFCC嵌入支路则通过端到端深度网络，挖掘传统统计量无法表征的复杂声学模式，如音色微颤、语调反转、停顿节奏异常、伪装平静中的能量聚集等。这些深层模式往往是“表面无害、实则恶意”的典型声学指纹。  (1)共享音频预处理：奠定双支路一致性基础  为确保两条支路的输入信号在时频域具有完全一致的物理意义，系统首先对原始波形执行严格统一的预处理流水线。首先进行预加重操作，增强高频分量以补偿声道传输中的能量衰减：  公式3-11  该步骤显著提升了对清音、爆破音等高频攻击性语音成分的敏感度。随后，信号被切分为短时重叠帧（帧长25 ms，帧移10 ms），每帧施加汉明窗以平滑边界、抑制频谱泄漏。加窗后的帧序列成为两条支路的共享输入基础，保证后续特征提取在相同的时间粒度与频谱分辨率下进行，从而实现跨支路的公平比较与有效融合。  (2)传统MFCC统计特征支路：物理可解释的声学统计建模  传统支路忠实遵循经典MFCC提取范式，旨在生成紧凑、高度可解释的声学统计描述子。其处理流程为：对每帧信号执行快速傅里叶变换（FFT）获取功率谱；通过Mel滤波器组模拟人耳对频率的非线性感知：  公式3-12  滤波器中心频率按Mel刻度等间距分布，带宽随频率递增，更符合人类听觉对低频差异的敏感性。滤波能量取对数后，通过离散余弦变换（DCT）实现去相关与能量压缩，得到每帧的MFCC系数：  公式3-13  系统保留前13阶低频系数（通常含 能量项则为14维），构成静态特征向量。为进一步捕获语音的动态演化特性，计算一阶差分（Δ）与二阶差分（ΔΔ），分别表征语速变化与加速度模式。  对整段音频，系统不再采用简单的帧级拼接，而是通过全局统计聚合函数（均值、标准差、最大值、最小值、偏度、峰度等）对所有帧的静态+动态系数进行降维汇总，最终输出39维传统声学特征向量。该向量高度浓缩了整段语音的能量分布趋势、节奏稳定性、音量突变频率等宏观声学属性，为后续融合提供坚实的、可解释的声学先验。  (3)深度MFCC嵌入支路：数据驱动的高阶情绪表征学习  与传统支路严格并行，深度支路以帧级MFCC序列为输入，构建从低维倒谱到高维情绪嵌入的端到端非线性映射网络。该网络采用三层全连接结构（FC-Net），每层后接Batch Normalization与ReLU激活，逐步实现特征的抽象与重组：  公式3-14  网络维度设计为13→256→128→64，参数通过反向传播联合优化。每一帧原始MFCC被映射为64维深度嵌入，蕴含了传统统计量难以显式表达的音色稳定性、语调曲线弯折、微表情式停顿分布等深层声学模式。  为避免简单均值池化导致的攻击性高潮片段被稀释，系统引入可学习注意力池化机制，动态加权各帧对最终表征的贡献：  公式3-15  注意力权重由一个小型门控网络计算，使模型能够自动聚焦于能量爆发、语调突变或讽刺性拖长音等关键时刻，生成64维深度声学嵌入。该嵌入特别适用于识别“表面平静、实则恶意”的伪装型攻击语音。  **3.4 声学特征整合层：多粒度表征的统一对齐**  两条支路输出——39维传统统计特征与 64维深度情绪嵌入——在声学特征整合层完成空间对齐与信息互补：  公式3-16  其中拼接操作保留了两类特征的完整表达，整合层通过可学习投影实现维度统一与非线性交互。最终输出的128维综合声学嵌入同时具备：  物理可解释性：音量、语速、能量等统计规律  情绪表达力：讽刺拖腔、冷笑音色、压抑愤怒等深层模式  时间敏感性：攻击性高潮的动态聚焦  该嵌入随即被送入音频识别（Audio Recognition）模块，作为声学模态的唯一代表，与语义支路的在高层融合层进行跨模态联合建模（详见第四节）。  3.5 面向真实场景的鲁棒性增强策略  社交媒体音频常伴随压缩失真、背景噪音、变速说话等干扰。为提升模型在野外环境下的稳定性，训练阶段对输入音频施加多样本增强（Data Augmentation），包括：  时域拉伸：随机因子，模拟快慢语速  加性噪声：，模拟环境嘈杂  音高偏移：半音，增强音色泛化  音量扰动：dB，适应录音设备差异  增强样本与原始样本共同参与梯度更新，迫使模型学习声学不变性（acoustic invariance），在复杂场景下保持性能稳定。  **3.1.6大模型赋能**  **1.POINTS1.5**  POINT 1.5 是一个以大型语言模型为核心的多模态视觉-语言模型。其核心思想是通过一个可训练的适配器（MLP Projector），将预训练的视觉编码器提取的图像特征对齐并“注入”到冻结的 LLM 的词嵌入空间中。这种架构使得 LLM 能够有效地“看到”和理解图像内容，并基于图文结合的上下文进行推理和生成文本响应。模型强调参数高效性，通过仅训练投影层来赋予 LLM 视觉能力。其原理如图3-10所示。    图3-10 POINTS1.5工作原理  **模型流程路径**  模型接收来自用户的原始图像（一张或多张）和文本形式的用户指令/查询（例如：“Please describe this image.”）。原始图像首先被分割成一系列规则大小的图像块（patches），这是 Vision Transformer 处理图像的标准输入格式。  这些图像块被送入图像编码器 (NaViT)。NaViT是一个强大的预训练模型（其参数通常在 POINT 1.5 的训练阶段被冻结），它从图像块中提取出高维度的、语义丰富的视觉特征序列。  视觉编码器输出的视觉特征序列被送入MLP 投影器。这是模型中唯一可训练的部分，它学习将视觉特征映射到大型语言模型 (LLM) 所理解的词嵌入空间。MLP投影器将视觉特征转换为一系列与 LLM 内部词嵌入维度兼容的“伪令牌”（Pseudo-Tokens）。模型的输入序列被精心构造，以统一的方式将文本和图像信息呈现给 LLM：首先，插入特殊的视觉开始令牌（<vision\_start>）。然后，插入由 MLP 投影器生成的图像伪令牌序列。接着，插入特殊的图像结束令牌（<im\_end>）。或视觉结束令牌（<vision\_end>）。随后，插入用户提供的文本指令/查询（通过 LLM 自身的文本分词器处理）。整个序列可能还包含user和assistant等角色令牌，以模仿对话格式。如果有多个图像，该模板和伪令牌序列可以重复多次，以便 LLM 同时处理多张图片及其相关指令。  构建好的混合序列（包含文本令牌和图像伪令牌）被送入大型语言模型 (LLM)。LLM 的核心 Transformer 解码器层（包含多头自注意力机制和前馈网络）对整个序列进行处理。LLM 的自注意力机制允许文本令牌与图像伪令牌之间进行深度且双向的交互。LLM 能够理解文本指令所指的图像区域或概念，并综合图文信息进行复杂的上下文推理。由于 LLM 参数是冻结的，它能够充分利用其庞大的预训练知识库来理解和关联视觉与文本信息。  经过 LLM 的处理后，其输出头（通常是一个线性层后接 softmax 激活函数）会根据融合后的信息和用户指令，自回归地生成最终的文本响应。这个响应可以是图像的详细描述、对图像内容的问答、或遵循指令完成的文本任务。在 POINT 1.5 的训练阶段，只有 MLP 投影器及其连接的 LLM 的部分层（如 LoRA 适配器）是可训练的，图像编码器和 LLM 的大部分核心参数保持冻结。这使得模型训练非常高效，可以在较少的计算资源下，将强大的视觉能力注入到 LLM 中。  总体路线：  多模态输入接收->图像预处理与特征提取->LLM 输入序列构建->大型语言模型处理与融合->文本输出生成  **2. ERNIE4.5**  ERNIE 4.5 是一个统一的多模态基础模型，旨在实现对多种模态（图像、视频、文本）的深度理解和生成能力。该模型采用高度模块化和可扩展的架构，特别强调了对多分辨率视觉输入的自适应处理，以及通过路由器和专家网络实现高效的模态特异性处理与融合。其核心在于通过多模态自注意力层整合来自不同模态的信息，并支持复杂的跨模态任务。其工作原理如图3-11所示。    图3-11 ERNIE4.5工作原理  模型流程路径：  模型接收各种格式的原始数据，包括视频（不同时长和分辨率）、图像（不同分辨率和长宽比）以及文本（用户指令或语料）。例如，用户可以输入一张高分辨率图片和一句“详细描述这张图片”的文本指令。  原始视频和图像首先通过视觉编码器进行初步处理，提取基础视觉特征。针对不同分辨率和长宽比的视觉输入，模型会进行自适应的调整或变换，以生成统一格式或多尺度的特征。对处理后的视觉特征进行空间维度的压缩，生成一系列更紧凑的视觉特征序列。针对视频输入，模型会对时间维度进行压缩，捕获视频的动态信息和时序特征，生成时间压缩后的视觉特征序列。压缩后的视觉特征序列通过一个适配器，将其调整到与后续路由和专家模块兼容的维度和格式。  用户输入的文本指令（如“Describe this image in detail”）首先通过文本分词器进行分词，生成文本令牌序列。  经过适配器处理的视觉特征序列被送入视觉路由器。路由器根据特征的特性和潜在任务需求，智能地将其分配给一个或多个视觉专家。视觉专家是专门处理视觉信息的子网络，进行深度的视觉理解。分词后的文本令牌序列被送入文本路由器。路由器同样将其分配给一个或多个文本专家。文本专家是专门处理文本信息的子网络，进行深度的语义理解。视觉专家和文本专家独立地对其负责的模态信息进行深度分析，提取出各自模态最丰富的语义表示。  来自视觉路由器和文本路由器的输出（可能已经经过各自专家的处理）被送入多模态自注意力层。在这个阶段，模型允许视觉特征和文本特征之间进行深度、双向的交互，从而学习它们之间的复杂关联和互补信息，形成统一的多模态上下文理解。融合后的多模态特征（或经过多模态自注意力层输出的统一表示）被进一步送入共享专家模块。这些专家不再区分模态，而是处理融合后的高级语义信息，进行跨模态推理和决策。经过多模态自注意力层和共享专家处理后的最终多模态表示，被模型的输出头（通常是 Transformer 解码器）用于生成最终的输出。  根据用户指令，输出可以是文本形式的描述（如对图像的详细描述）、问答，甚至可以支持图像、视频或其他模态内容的生成。ERNIE 4.5 作为一个基础模型，通常在海量且多样化的多模态数据集上进行端到端的联合预训练。模型的所有组件（包括编码器、路由器、专家、自注意力层等）都会在预训练阶段共同学习。这使得模型能够学习到非常通用和强大的跨模态表示。  总体路线：  多模态原始输入->视觉输入预处理与特征提取->文本输入预处理与特征提取->多模态路由与专家处理->多模态融合与共享专家处理->输出生成  **3.1.7 结果评价与分析**  本项目涉及多模态内容检测的多任务输出结果分析。由于检测任务均为“正确判断/分类”或“错误判断/分类”的二维指标问题，我们采用统一的评判标准，对文本、图像、音频三模态以及融合输出的检测结果进行系统评价。  一、准确率（Accuracy）  准确率是最常用的评价指标，表示模型正确判断样本的比例。对于委婉有害内容检测任务，准确率衡量了系统在多模态场景下对“违规/合规”样本判断的总体正确性。  公式3-17  其中：  TP (True Positive)：正确识别为违规内容的样本数；  TN (True Negative)：正确识别为合规内容的样本数；  FP (False Positive)：误判为违规的合规样本数；  FN (False Negative)：漏判为合规的违规样本数。  准确率反映整体性能，但在样本比例失衡（如违规内容稀少）情况下，需结合其他指标综合评估。  二、精确率（Precision）  在多模态检测模型中，精确率衡量被模型判定为违规的样本中，真正违规的比例。即系统在“报警”的情况下，其判断的可靠程度。  公式3-18  精确率越高，说明误报率越低，系统在严格场景下更具实用性。  三、召回率（Recall）  召回率衡量真实违规样本中，被模型成功检测出的比例。 在内容安全任务中，召回率决定了系统漏检的严重程度。  公式3-19  高召回率意味着模型能有效覆盖委婉语、隐喻、图文讽刺等复杂隐性违规表达。  四、F1 分数（F1-Score）  F1 分数是精确率与召回率的调和平均值，是衡量不均衡样本任务中性能的关键指标，特别适用于本项目中“违规”样本占比低的情境。  公式3-20  F1 值越高，说明模型在误报与漏报之间取得了更好的平衡。  五、AUC-ROC（曲线下面积）  AUC（Area Under Curve）表示 ROC 曲线下的面积，用以衡量模型区分“违规”与“合规”样本的能力。ROC 曲线反映了假阳性率（FPR）与真正率（TPR）的关系。  公式3-21  其中：  公式3-22  公式3-23  AUC 值接近 1 时，说明模型的分类能力接近理想状态，能有效区分语义委婉的违规与正常表达。  六、跨模态一致性指标（Cross-Modal Consistency）  由于系统包含文本、图像、音频三模态输入，为衡量多模态间语义判断的一致性，引入跨模态一致性指标：  公式3-24  其中为模态数量，为第个模态输出的违规置信度。 越接近1，表示各模态对违规性的判断一致，融合结果稳定可靠。  七、计算效率（Efficiency）  一个高性能检测模型不仅应具备高准确率，也应具有低计算开销与高推理效率。评价指标包括：  推理时间：每个样本的平均处理时间（单位：ms）；  计算资源：模型所需的 GPU/CPU 浮点运算量（FLOPs）；  模型大小：参数存储量（MB 或 GB）。  综合效率计算公式为：  公式3-25  其中 为任务相关权重。 最终根据性能与资源消耗的权衡，评估模型部署的可行性与优化方向。  **3.2 可行性分析**  **3.2.1 从社会层面来看的应用可行性分析**  面向社交媒体与智能审核场景的多模态委婉有害内容检测系统，具备显著的社会可行性与落地价值。  首先，随着生成式AI的普及与社交平台内容形态的多元化，谐音隐喻、图文讽刺、音频暗讽等隐性违规表达持续增多，传统单一文本过滤方式已难以覆盖 “文本合规但图像、音频违规”的跨模态场景，也无法精准捕捉委婉化违规形式，难以满足当前网络生态治理的实际需求。本项目依托大模型的多模态语义识别能力，可有效填补这一治理空白，在社会治理、平台审核、内容安全监管等领域发挥关键支撑作用。  其次，国家对人工智能伦理安全与内容合规的监管要求持续升级，《生成式人工智能服务管理暂行办法》《网络信息内容生态治理规定》等政策文件，既明确了内容审核的合规底线，也为AI技术在内容治理领域的应用提供了清晰导向，为本系统的合规落地筑牢了政策与法律基础。  同时，抖音、B 站、小红书等主流平台的内容形式日趋丰富，平台面临的审核压力与日俱增，而公众对网络内容安全性、审核公正性的关注度不断提升，对高效智能的审核解决方案需求迫切，这为本系统的推广应用创造了良好的市场环境与舆论基础。  综上，从现实治理需求、明确政策支持、广泛市场认可到清晰产业化前景，本项目的社会可行性突出，具备广泛的应用与推广价值。  **3.2.2 从技术层面来看的技术可行性分析**  **文本模态检测技术可行性分析**  系统选用RoBERTa/ERNIE融合的Transformer架构，核心适配中文委婉有害内容的检测需求。Transformer自带的自注意力机制，能精准捕捉谐音、拆字、隐喻等表达的上下文语义关联，解决传统模型难以破解的语义依赖问题。该架构的计算复杂度为 ，通过GPU并行计算优化后，可满足社交媒体实时审核的响应要求。微调阶段引入中文委婉语专项数据集，针对性覆盖网络场景中常见的隐性违规表达类型，进一步强化模型对中文语境下特殊违规形式的识别适配能力。  **图像模态（RAM++ 架构）可行性分析**  图像检测模块采用 Swin-Large 作为视觉特征提取核心，搭配 BERT 语义融合机制，形成“视觉特征 + 语义理解”的双重检测逻辑。核心优势在于采用移位窗口注意力机制，将传统Transformer全局注意力的复杂度优化为，在保障检测精度的同时，大幅降低计算开销。这种设计天然适配社交平台实时审核场景，能将单样本推理延时控制在35 ms以内，完全满足大规模内容审核的效率要求。  **音频模态（ASR-BERT + MFCC 双支路）可行性分析**  音频检测采用MFCC声学特征提取与ASR-BERT语义分析的双支路架构，形成全方位检测能力。MFCC支路专注捕捉音频的声学特征，适配不同环境下的音频输入场景；ASR-BERT支路通过语音转写与语义建模，既能精准识别语音中的文字违规信息，又能捕捉情绪化语调、讽刺语气等隐性违规信号。双支路协同建模的方式，让音频检测具备强鲁棒性，同时适配实时审核的效率需求，无需额外复杂计算即可完成检测。  **跨模态融合模块可行性分析**  融合层采用多头交叉注意力机制实现三模态语义交互同公式3-2，该机制能有效统一不同模态的特征空间，打破单模态检测的信息壁垒，实现语义补全与模态一致性判断 —— 比如精准识别 “文本合规但图像违规”“语音语调异常搭配普通文字” 等跨模态隐性违规场景。其结构设计贴合多模态内容的语义关联特性，无需复杂适配即可实现高效的模态协同检测。  **系统工程实现可行性**  **系统工程实现性分析**  系统基于 PyTorch 框架开发，该框架具备完善的生态支持与广泛的行业应用基础，开发适配性强。硬件层面可兼容 RTX 4090/A100 等主流计算设备，无需特殊定制化硬件即可完成模型训练与推理。通过模型量化与参数共享技术，可在不影响检测效果的前提下优化资源消耗，降低部署成本。同时支持 Docker 容器化部署与RESTful接口接入，无需大规模改造即可落地应用，具备良好的工程移植性与产业化适配能力。  综上所述，本项目的技术可行性突出。从算法架构来看，各模态选型均贴合隐性违规检测的核心需求，跨模态融合设计解决了单模态检测的固有局限；从工程落地来看，框架成熟、硬件适配性强、部署便捷，完全能够满足智能内容治理的技术应用需求，契合人工智能安全发展的政策导向。 |
| 4．**本项目的特色与创新点（**建议2-3点**）**； |
| **1.首次系统化提出并验证“中文三模态委婉语”识别范式。** 我们首次从语义演化、符号隐喻与声学线索三条路径系统化刻画中文委婉有害表达，并基于此提出可量化的“委婉强度—模态关联”标注体系，用以支撑模型训练与评估，从而弥补当前中文多模态数据标注粒度不足的空白。  **2.首次构建并公开可工程化的“多模态增量适配”检测框架。** 在传统全量重训之外，我们首次将语义演化量化模型与增量学习机制结合，实现对新型谐音、隐喻与模态错位的快速适配；该机制在保持旧知识稳定性的同时，显著缩短新型委婉表达的检测响应周期（工程化设计支持平台部署）。 |

|  |
| --- |
| 1. **研究基础与工作条件** |
| 1．**项目负责人研究基础**（建议300字以内） |
|  |
| 2.**指导教师研究基础**（与本项目相关的研究工作积累和已取得的研究工作成绩，建议300字以内）； |
| 本项目指导老师为王海舟教授，王海舟教授长期致力于网络空间安全方面的科学研究和人才培养，主要从事舆情监控，情报分析，网络测量，社交网络分析等方向的研究工作。先后主持和参与了国家自然科学基金项目、国家科技支撑计划项目、国家242信息安全专项项目、国家“核高基”重大专项项目等国家级科研课题10余项，主持课题累计到校科研经费140余万元，累计发表网络安全领域学术论文30余篇，其中SCI/EI检索学术论文10余篇，参与编制国家标准2项，指导学生获国家级竞赛奖励多项。我们的项目为基于大模型辅助的多模态委婉有害内容检测技术研究，研究中涉及到的技术都是内容安全的重要组成部分，与老师的研究方向相契合，能给予我们项目充分的技术上的指导与支持。 |
| 3．**工作条件**（建议200字以内） |
| 指导老师的实验室配备齐全的软硬件开发环境，包括华为、曙光和宝德等高性能物理服务器200余台、GPU服务器10余台、国内外云服务器300余台以及800TB存储阵列。实验室还拥有多款高速带宽，为项目的数据收集、分析、平台搭建和实验验证提供强有力的支持。自成立以来，实验室与国内外多所院校建立了合作关系，定期举办讲座和合作研究。实验室多次承办国内外会议，并担任重要职务，为及时了解国际学术前沿提供有利条件。 |

|  |  |
| --- | --- |
| （三）**承担的与本项目相关的科研项目情况** | |
| 指导教师曾经和正在承担和参加的省部级以上科研和教改项目情况 | 1.国家重点研发计划项目子课题，2022YFC3303101，智能社会场景下多元实体表征及实体关系模型，2022/10-2025/09，纵向项目，在研，主持  2.四川省科技厅重点研发计划项目，2023YFG0145，面向网络舆情治理的社交网络多智能体构建及自主协同关键技术研究，2023/1-2024/12，纵向项目，在研，主持  3.国家自然科学基金青年基金，61802271，基于动态行为和拓扑演化的对等网络流媒体系统内容污染扩散机制研究，2019/01-2021/12，纵向项目，已结题，主持  4.教育部-中国移动科研基金，MCM20200409，网络安全威胁情报的大数据分析及应用研究，2021/01-2021/12，纵向项目，已结题，主持  5.国家自然科学基金青年基金，61802270，面向大数据平台的服务认证协议的可证明安全及效率优化研究，2019/01-2021/12，纵向项目，已结题，参研  6.国家自然科学基金面上基金，81773548，环境流行病学中多元非线性高维交互效应的贝叶斯模型构建及推断研究，2018/01–2021/12，纵向课题，已结题，参研  7.国家自然科学基金面上基金，61272447，基于动态多维特征的网络行为模型研究，2013/01-2016/12，纵向项目，已结题，参研  8.国家科技支撑计划项目，2012BAH18B05，新媒体资源管控关键技术研究及原型系统，2012/01-2014/12，纵向项目，已结题，参研 |
| 负责人之前参与大创项目情况 | 无 |
| 项目成员之前参与大创项目情况 | 无 |

|  |  |
| --- | --- |
| （四）**完成大创项目情况**（对负责人负责的前一个大创（项目名称及编号）完成情况、后续研究进展及与本申请项目的关系加以详细说明。另附该已结题项目研究工作总结摘要和创新点（限200字）和相关成果的详细目录。未承担过的写“无”）。 | |
| 项目名称及编号 | 无 |
| 完成情况与后续研究进展 | 无 |
| 与本申请项目的关系 | 无 |
| 工作总结摘要及创新点 | 无 |

|  |  |
| --- | --- |
| **（五）申请人成果和奖励情况**  （请注意：①投稿阶段的论文可以列出；②对期刊论文：应按照论文发表时作者顺序列出全部作者姓名、论文题目、期刊名称、发表年代、卷（期）及起止页码（摘要论文请加以说明）；③对会议论文：应按照论文发表时作者顺序列出全部作者姓名、论文题目、会议名称(或会议论文集名称及起止页码)、会议地址、会议时间；④应在论文作者姓名后注明第一/通讯作者情况：所有共同第一作者均加注上标“#”字样，通讯作者及共同通讯作者均加注上标“\*”字样，唯一第一作者且非通讯作者无需加注；⑤所有代表性研究成果和学术奖励中本人姓名加粗显示。） | |
| **1.代表性成果（**包括论文、专利、专著、科创竞赛获奖、学术交流活动、奖学金等，限合计5项**）** | |
| 论文 |  |
| 专利 |  |
| 专著 |  |
| 科创竞赛获奖 |  |
| 学术交流活动 |  |
| 奖学金 | 2024-2025学年综合奖学金 |
| **2.代表性之外成果和奖励（**限合计不超过5项**）。** | |
| **无** | |

|  |  |
| --- | --- |
| **（六）预期成果形式（**可多选**）** | |
| 1.□SCI论文 篇  2.☑核心期刊论文 1 篇  3.□会议论文 篇  4.□内部编印期刊论文 篇  5.□授权发明专利 项  6.☑申请发明专利 1 项  7.☑创新创业类竞赛获奖  8.□参加国际国内学术交流活动  9.□其他 名称： | |
|  |  |
| **（七）项目经费概要**（按申报项目目标任务需要进行预算，经费执行情况将与结题考核成绩挂钩） | |
| **1.申请经费明细**  （1）仪器设备费  （2）耗材费 500  （3）测试加工费 2000  （4）国内会务及差旅费  （5）国外会务及差旅费  （6）文献/知识产权事务费 1000  （7）办公费（含文印、办公用品等） 500  （8）其他费用 1000  **2.合计 5000** | |

|  |
| --- |
| **评审情况** |
| **指导教师意见：** |
|  |
| **指导教师（签名）： 年 月 日** |
| **学院推荐意见：** |
|  |
| **主管院长签名： 年 月 日** |
| **学校专家评审意见：** |
|  |
| **组长签名： 年 月 日** |
| **学校认定意见及批准经费：** |
|  |
| **学校负责人签名： 年 月 日** |