开源情报多模态智能处理系统设计与工程实现

董泽云[1]，甘莅豪[2]，薛楠[3]，陆泰廷[4]
1. 上海安啼特科技有限公司，上海 200235；
2. 华东师范大学传播学院，上海 200241；
3. 中国人民大学数学学院，北京 100872；
4. 宾夕法尼亚州州立大学电气工程与计算机科学学院，宾夕法尼亚州 16802

摘要
针对开源情报系统存在的模态割裂、结构化能力不足及用户交互性差等问题，提出一种融合计算机视觉、自然语言处理与文本转语音技术的智能信息处理系统。基于多源异构数据设计了涵盖数据采集、预处理、深度建模、智能决策与用户交互反馈的完整闭环流程，重点突破跨模态数据融合、情报内容结构化处理、语音播报与多媒体可视化呈现等关键技术。实验结果表明，系统在情报抽取准确率、响应时间及用户可解释反馈等关键指标上表现优异，具备模块化与可扩展性，适配政务安全、金融风控与公共舆情等场景。

关键词
开源情报；计算机视觉；自然语言处理；文本转语音；语音识别；多模态融合；大语言模型；人工智能

中图分类号：G350.7；TP391 文献标志码：A doi：10.11959/j.issn.2096-0271.2026009

Design and engineering implementation of an open-source intelligence multimodal intelligent processing system

Dong Zeyun[1], Gan Lihao[2], Xue Nan[3], Lu Taiting[4]
1. Shanghai Antite Technology Co., Ltd., Shanghai 200235, China
2. School of Communication, East China Normal University, Shanghai 200241, China
3. School of Mathematics, Renmin University of China, Beijing 100872, China
4. School of Electrical Engineering and Computer Science, Pennsylvania State University, Pennsylvania 16802, USA

Abstract
To address the issues of modal fragmentation, insufficient structural processing capabilities, and poor user interactivity in open-source intelligence systems, an intelligent information processing system integrating computer vision, natural language processing, and text-to-speech technologies has been proposed. Based on multi-source heterogeneous data, a complete closed-loop workflow was designed, covering data acquisition, preprocessing, deep modeling, intelligent decision-making, and user interaction feedback. The study focused on breakthroughs in key technologies such as cross-modal data fusion, structured intelligence content processing, voice broadcasting, and multimedia visualization. Experimental results showed that the system achieved excellent performance in key indicators such as intelligence

extraction accuracy, response time, and user-interpretable feedback. The system is modular and scalable, making it suitable for applications in government security, financial risk control, and public opinion monitoring.

Key words
open-source intelligence, computer vision, natural language processing, text-to-speech, automatic speech recognition, multimodal fusion, large language model, artificial intelligence

0 引言

在数字化浪潮下，互联网正以前所未有的速度吞吐着海量、异构、碎片化的数据。这种信息爆炸的格局，为开源情报（open-source intelligence，OSINT）带来了前所未有的土壤——资源广泛、更新迅捷、来源多元[1]。然而，数据虽多，情报却未必多，这背后隐藏着一个根本性的矛盾，即真正有价值的线索不易挖掘。所谓开源情报，是指从公开可获取的信息中提炼出的可用于决策支持的洞见，它成本低、风险小、收益却惊人[2]。数据显示，美国国防情报局的情报报告中约有80%是基于开源信息的[3]。但低门槛同时也带来质量良莠不齐的问题，新闻报道、社交媒体、论坛帖文等[2]平台的信息汹涌而来，问题也随之而来——信息冗余、模态割裂、人工分析缓不济急。市面上的大多数情报工具依旧停留在单模态处理的层面，面对图文音频交织的现状，它们无法有效应对。正是在这一背景下，人工智能，尤其是深度学习，作为一股智能化浪潮，开始重构情报生产的流程。机器学习模型不仅能在社交平台上自动聚类信息事件[4]，也能借助循环神经网络（recurrent neural network，RNN）、长短期记忆网络（long short-term memory，LSTM）精准识别恶意域名[5]。分类—提取—警报这条流水线逐渐成形[6]。但这真的是智能情报吗？还是仅仅是自动归档？事实是这些工具往往只能处理某一类任务，缺乏一个系统性、交互性、可解释的全流程架构，更不用提实时响应、闭环反馈这些高阶能力。

直到多模态技术出现。图像、视频、语音乃至多语言文本等，过去看似分立的信息碎片终于可在同一空间中映射共振[7-8]。这种跨模态的融合不仅丰富了情报的维度，也提升了模型对事件本质的捕捉力。例如，通过对社交媒体上的文字与图片进行联合分析，系统能够交叉验证事件真伪，并将图文不符的帖子识别为潜在谣言。跨模态深度学习的迅猛发展令智能系统具备了图文双解的能力，它们不仅能读懂语义，还能看懂场景，这在情报领域是一次质的飞跃，让人们得以重构对事件的认知。

基于此，本文构建了一套以多模态技术为核心的开源情报处理系统。在架构设计上，强调模块解耦与协同执行；在关键技术方面，涵盖多源信息融合、跨模态表示学习与动态优化机制；在系统能力上，致力于实现信息采集、处理、反馈的闭环智能。在此基础上，本文还设计了系统验证实验，评估其实用性能，并从实际落地场景出发，提出未来可能的演化路径。

1 相关研究综述

1.1 开源情报自动化处理研究现状

开源情报之所以成为显学，不仅因为

信息本身易得，更因为传统情报体系的结构性惯性，它需要一套足够廉价、效率高，还能在一定程度上看懂世界的系统，以对抗信息不对称的新问题。早在10余年前，美国便将OSINT纳入五大情报支柱之一，而后陆续在自动化处理能力上投入，力图将海量公开数据转化为可操作知识[9]。在学术界，从2019年前后开始，以人工智能（artificial intelligence，AI）赋能OSINT为关键词的研究数量显著上升[5]。但是大多数方案仍局限于局部环节，自动抓取、聚类分析、关键词过滤实则只是将“体力活”外包给机器，并未触及情报本质——推理、判断与矛盾处理[10]。更根本的问题在于，大多数OSINT系统仍建立在单一模态的假设上，在图像、音频和视频等多模态信息快速增长的背景下，这一假设显得愈发过时。而现有模型之间缺乏协同，信息之间难以对齐，导致系统最终只是多个孤立模块的拼合[6]。问题不在于是否使用了AI，而在于这些AI是否真正构建了有效的理解机制。

1.2 NLP、CV与TTS/ASR在情报系统中的典型应用

如果说自然语言处理（natural language processing，NLP）是当前情报系统的语言中枢，那么计算机视觉（computer vision，CV）、文本转语音（text to speech，TTS）和自动语音识别（automatic speech recognition，ASR）则是逐渐觉醒的感官系统和声带。在多数现有系统中，NLP仍扮演着主角，它帮助人们从社交媒体的海量信息中提炼出脉络，在政策文件中定位实体，甚至用来预测下一场舆论风暴。得益于多语种模型的成熟，它也在语言边界问题上取得了一定进展。但文本不是全部，图像和视频携带的情绪张力、视觉异常，甚至仅凭一帧画面即可触发的预警，都是传统NLP无力企及的盲区。CV的作用正在于此，其可从卫星图像中识别军事部署[11]，从抗议现场中辨识关键物体，甄别新闻图像的伪造与篡改。TTS/ASR则将系统从读懂情报的阶段推进到能叙述的阶段，尤其在突发事件的实时预警中，语音播报带来的响应优势尤为关键。

1.3 多模态融合系统的挑战与前沿探索

多模态融合技术在情报处理方面取得显著进展。然而技术跃进的背后仍潜藏诸多桎梏，值得人们警惕与深究。首先，异构数据的表征与关联是跨模态情报融合绕不过的门槛。视觉图像携带的空间信息与文本蕴含的语义特征如何才能精准对齐？目前已有研究提出特征融合机制与注意力对齐方法，如事件对抗神经网络（event adversarial neural networks，EANN），试图在图文之间搭建可泛化的桥梁[2]。其次是内容一致性与可信度问题的双重困扰。在多模态数据中，图像与文字常常呈现貌合神离特征，对此部分研究引入知识图谱机制，如面向中文开源情报的智能体（COSINT-Agent）尝试借助背景知识提升情报推理的逻辑完整性，但其可解释性与动态适应能力仍有待考验[12]。更现实的挑战则关乎实时性与可扩展性，面对高频流式数据与海量接入节点，系统若不能快速响应、弹性扩展，就不能在实际场景中部署。未来的研究必须进一步聚焦如何增强多语种适配、提高抗虚假信息干扰能力，以及实现系统与战术场景的无缝衔接[5]。归根结底，关键问题仍在于如何兼顾技术深度与系统实用性。

2 系统总体架构设计

2.1 多系统3层结构设计与模块划分

如**图1**所示，为应对多模态数据的复杂性与动态性，本文提出的开源情报处理系统采用3层架构布局。采集层、处理层到交互层各司其职，整个系统如同一条智能流水线，从采集、清洗到并行处理、深度融合，最终实现输出闭环。各模块间松耦合、高内聚，既可独立升级，又支持快速替换，这种架构设计为系统的演进提供了生长空间。

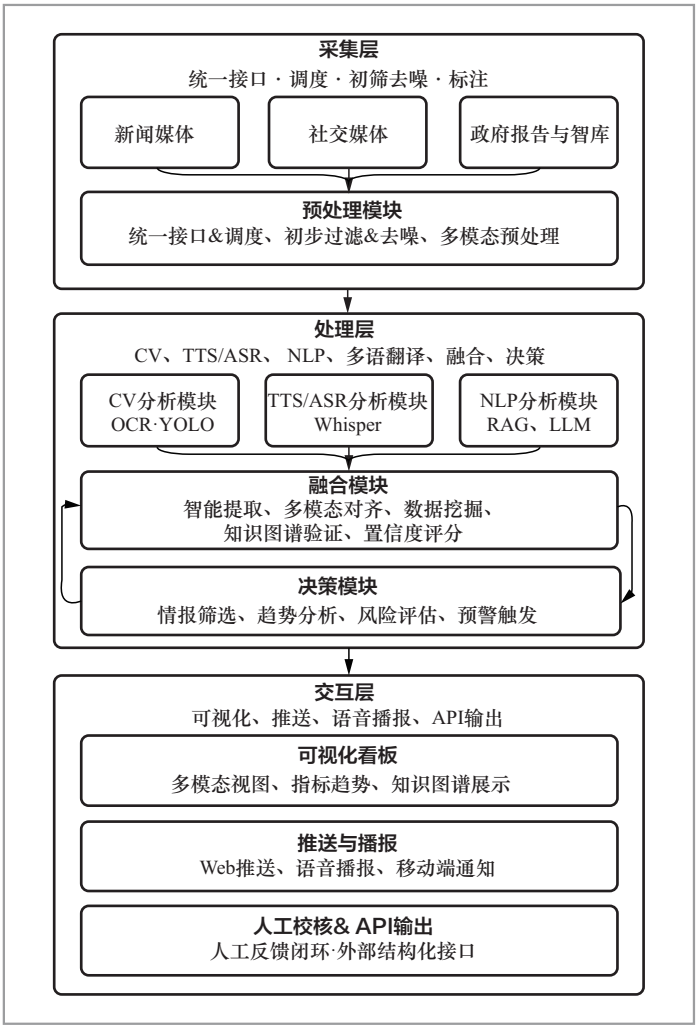


图1 3层架构与模块划分

采集层面向广泛的数据源，完成情报原料的收集。它通过统一接口将网络爬虫与应用程序编程接口（application programming interface，API）组件整合进调度体系，可定时或按需从新闻平台、社交媒体、政府公开数据库中抓取数据。此阶段不只收集数据，而且进行数据的初步过滤与去噪，完成格式转换，并为文本、图像、音视频3类数据分别打上时间戳与来源标签。随后，预处理模块对文本进行分词、停用词剔除、拼写校正；图像被压缩、色彩归一；视频则被转写为文本、提取关键帧，并统一语言为中文或中性语料，为后续模型分析打好基础。进入处理层，系统的认知引擎正式启动。CV模块分为目标检测与光学字符识别（optical character recognition，OCR）两个子系统，前者基于精调YOLO框架锁定图像中的人物、车辆与装备；后者识别图中文字，全部附带置信度评分，为后续融合模块提供参数参考。TTS/ASR模块则集成Whisper语音识别与文本合成等引擎，既能将视频语音实时转写为文字，也能将情报内容合成为语音播报，用于紧急预警。在多语种环境下，系统会自动识别语种并调用翻译服务，确保中英文混杂情境下处理连贯无误。NLP模块分为基础与深度两个层级，基础层负责分词、词性标注与实体识别，深度层则结合检索增强生成（retrieval-augmented generation，RAG）技术与大语言模型（large language model，LLM）执行语义理解、要点提取与摘要生成。其输出不仅是JSON格式的结果，还能转化为知识图谱三元组结构，为结构化推理奠定基础。

融合模块位于多模态信息交汇的核心，

它统一CV、TTS/ASR与NLP模块的输出至同一嵌入空间，借助领域知识图谱比对实体、验证信息可信度。当图文语义一致时，系统上调置信评分；若出现语义冲突，则打上“需人工核查”的标签，生成带可信度评估的增强型情报。最终，决策模块根据情报评分与用户预设策略筛选高价值信息，执行趋势分析与风险评估，若触发预警阈值，还会自动发出语音播报或列表推送。其展示方式可灵活切换，以适应不同场景的需求。

2.2 数据通路与服务部署结构

为了让复杂的智能系统落地，笔者设计了一套微服务导向的分布式架构，每一个功能模块都被拆解、封装、独立孵化，化整为零，再以轻量协议，如REST API或远程过程调用（remote procedure call，RPC）织网连接，构建出一个松耦合、高可用的服务体系。

如**图2**所示，系统的神经从采集数据流开始延展，该部分部署了一整套爬虫/API集群，它们在调度节点的指挥下并行执行。采集的信息不直接被处理，而是先进入消息队列，这一缓冲机制既解耦又保障系统韧性。信息流进入分析层后，笔者在此部署了多种能力模块（CV、TTS/ASR与NLP）。每一模块都非“孤岛”，而是由多个实例并行构成的服务池。这些模块接收到消息队列中的任务后，由智能调度器分配空闲实例进行应对，从而实现任务高并发处理。融合与决策服务模块是综合整合前序结果的核心模块，它们不只是逻辑中枢，更承担了合成视听语言的重任。融合模块需同步调取下游CV、TTS/ASR与NLP的结果，进行并行分析，最终做出全局判断。笔者在同一物理节点上部署融合和决策服务，使用分布式架构与边缘计算，以缩短服务间通信链路，最大限度降

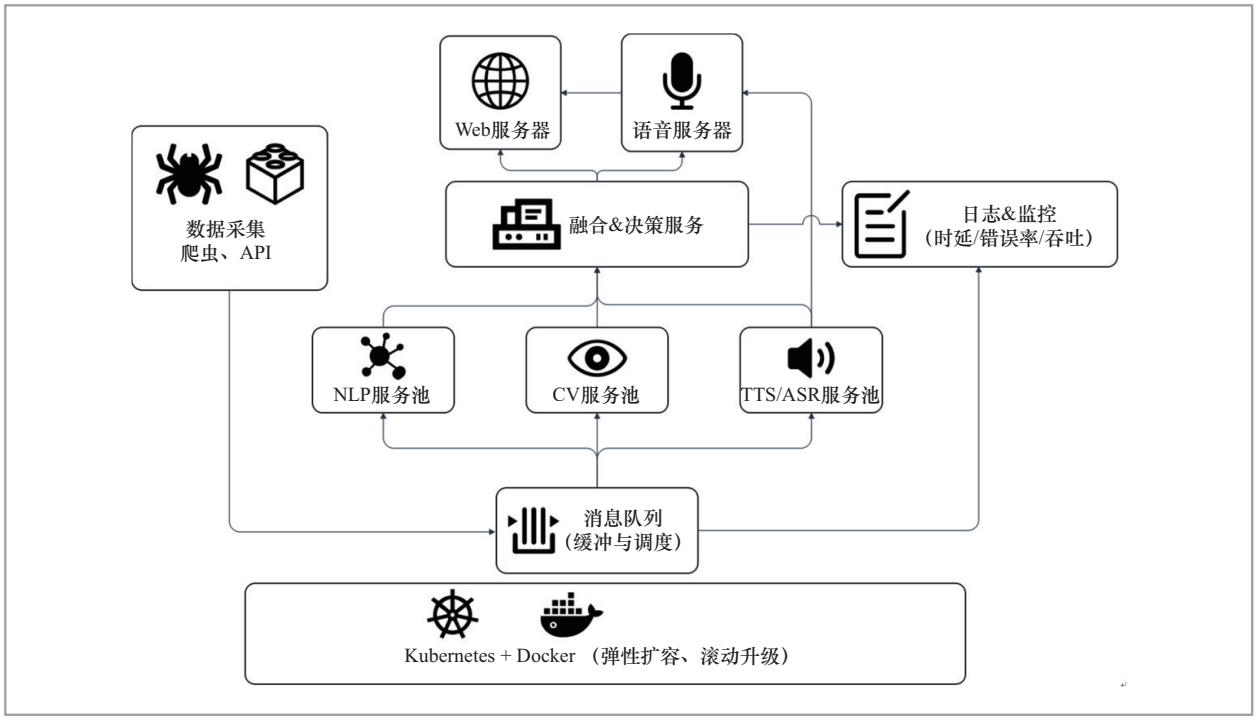


图2 数据通路与服务部署结构

低时延。更上层是面向用户的交互层，由Web服务器与语音服务器组成。前者托管可视化界面，是用户洞察全局的窗口，后者是最后的发声机器。当一条决策级预警产生时，它通过WebSocket协议立即推送前端刷新，同时触发语音模块播报。至此，一条完整的信息流从数据采集到用户通知形成闭环。

整个系统以Kubernetes集群与Docker容器为部署基石，支持服务弹性扩容，系统具备随需而变的神经反射能力。监控与日志模块负责记录每一跳的数据处理过程，对处理时延、服务错误率等指标进行实时监控。综上，该套架构为解耦与协同双重目标服务，既能在资源丰富的云集群上运行，也可在资源受限的边缘节点上精简运作。

3 关键技术模块详解

3.1 CV模块：OCR识别与图像线索分析

在CV模块中，图像处理流程从标准化预处理开始，涉及尺寸缩放、噪声抑制和平滑滤波、色彩空间归一化等操作，旨在降低原始图像中的冗余信息，提升后续模型对结构特征的提取准确性。随后，系统引入经微调的YOLO11目标检测模型，实现对人物、车辆等高频实体的自动识别。在特定任务场景下，如军事装备或企业Logo检测，则部署基于COCO数据集和PASCAL VOC数据集进行预训练，并经情报图像微调的专用模型，实现更高的分辨能力。识别结果包括目标边界框、类别标签及置信度分值，为后续图像理解提供结构化输入。在场景识别层面，系统集成Places365等主流场景识别（scene recognition）模型，识别图像所处环境（如灾害现场、游行示威等），并进一步通过目标空间分布与姿态分析模型建构对象间的语义关系链，例如推断“人群–围堵–警车”的动态场景构型。在文字识别方面，OCR子模块基于飞桨PaddleOCR引擎，可精准提取图像中嵌入的文本内容，覆盖交通标识、牌照、截图等多源文本；对于低分辨率或噪声严重的图像，系统采用图像超分辨率重建与二值化处理，增强字符识别能力。识别输出经语义校正后传递至NLP模块，进入文本语义层处理流程。CV模块的高级功能在于线索聚合分析，通过对关键目标与背景环境的语义比对，系统能够在多帧图像间自动建立实体关联关系，实现如“嫌疑人与事故车辆”的跨图配对，从而提升情报推理的完整性和连贯性。CV模块不仅完成视觉信息抽取，更支撑图像语义的逻辑重构。

3.2 NLP模块：信息抽取、实体识别与知识融合

NLP模块以文本的语言学结构解析为基础，首先进行分词与句法树构建，辅以专有词典以提升命名实体识别（named entity recognition，NER）在地名、机构名、人名等类目的准确率。系统采用5W1H（what、when、who、where、why、how）语义框架对文本进行主结构拆解，为下游信息抽取任务建立语义锚点。在信息抽取环节，系统调用BERT、Llama、DeepSeek与Qwen等大规模预训练语言模型，对原始文本生成主谓宾结构的三元组抽取结果，并同步识别实体间的关系类型（如隶属关系、参与关系），同时引入共指消解与语义上下文建模机制，以打通跨句、跨段落的实体与事件引用链条。例如，“爆炸事件–发生时间”与“爆炸事

件-地点”可整合为统一事件单元。针对复杂文本的语义分析，系统借助大语言模型执行事件类型判定、情感分析与紧急程度评估任务，文本分类与回归模型可进一步量化事件的潜在影响力。在长文本处理方面，模块基于Llama、Qwen等模型完成语义摘要生成，具备较强的信息压缩与关键信息提取能力，便于后续快速评阅与分析决策。此外，系统通过与基于RAG机制的知识库联动，实现对抽取实体属性的补全与验证，例如查询人物背景、地理坐标、组织关系等辅助信息，以增强情报准确性与交叉验证能力。最终输出包括命名实体集、事件关系网、摘要文本与情报在内的分类标签，为多模态融合提供基础语言结构支持。

3.3 TTS/ASR模块：多语言语音合成与反馈机制

TTS/ASR模块承担情报信息从视觉/语言模态向语音模态的跨域转换。首先，ASR系统利用Whisper语音识别模型将视频语料中的自然语言内容转录为结构化文本，后续语音合成过程则调用基于Transformer与梅尔频谱生成对抗网络(Mel-spectrogram generative adversarial network，MelGAN）结构的神经网络TTS模型，合成接近人类语感的普通话语音输出，并支持多语种适配。系统具备语言自动识别与语音模型选择能力，确保多语境情报播报的清晰度与一致性。在处理高优先级情报播报时，模块会对输入文本进行韵律重构，包括重音调整、停顿插入等语音韵律增强操作。对专有名词与数值信息，采用定制词典进行发音校正，以提升可听性与专业性。语音内容可通过Web端、移动App或电话网关等多终端分发，并支持语音指令交互功能。系统同步记录用户交互反馈，基于监听体验对语速、音高、节奏等参数进行迭代调优，形成闭环反馈机制，以持续提升播报体验质量。TTS/ASR模块不仅拓展了情报的传播媒介维度，更通过多语种、多终端覆盖机制增强了其跨文化与跨平台的情报影响力。

3.4 模型调用与多模态对齐策略

在多模态智能处理体系中，有效协调异构模型调用并实现跨模态对齐与融合是系统性能的关键。本系统采用按需异步调用策略，基于输入数据类型动态调度CV、NLP与TTS/ASR等子模块，结合消息队列机制与异步并发计算结构，显著提升资源利用效率。信息融合由专设模块执行，首先引入多模态语义嵌入策略，通过对比语言-图像预训练（contrastive language image pre-training，CLIP）等预训练模型将图像与文本特征映射至统一的语义向量空间[13]。在该空间内，不同模态的信息将以相似度为尺度进行语义比对，例如“车辆”一词的文本向量与图像中的视觉特征若呈高余弦相似度，则可推断该目标与文本语义存在强关联。为避免嵌入空间误配问题，系统融合基于规则的验证机制，包括实体精确匹配、时间一致性检查与地理元数据校对等手段，对图文一致性进行二次确认。对于存在显著语义冲突的结果，系统标记为“模态不一致”，并引入人工审阅流程。最终，融合模块将图像与文本向量拼接输入多层感知机（multilayer perceptron，MLP），以完成情报事件的分类判断、真假标记或影响等级评分任务。该机制通过聚焦于跨模态语义交集，提升了系统在复杂任务下的稳健性与决策可靠性。

4 数据闭环与系统优化机制

4.1 反馈驱动的模型优化

为提升系统性能与适应性，让模型有感知后果的能力，平台构建了以用户反馈为核心的数据闭环机制。系统记录了用户的浏览行为（点击、浏览时长等）、情报标注（如点赞、屏蔽等）、语音收听记录与查询轨迹，脱敏后形成结构化用户行为数据库，用于评估情报筛选效果与识别高价值信息，并据此优化交互设计与播报策略。反馈数据也用于模型训练，被标记为“无关”的内容作为负样本用于优化筛选与排序算法，重点情报则用于增强评分模型的准确性。TTS/ASR 模块对播报歧义词汇进行词典与生成逻辑调整；NLP 模块中 NER 识别错误将经专家审核后用于构建小样本集，提升特定领域识别精度。系统支持轻量级在线微调，并非一成不变，当检测到新兴术语（如流行词、组织简称）识别性能下降时，将自动构建训练子集并快速完成参数更新，实现语义模型的局部适应。每月定期进行离线重塑，基于用户反馈与专家标注对模型进行全量优化，并评估准确率、召回率与时延表现，仅在性能提升条件下部署更新。通过在线快调加离线迭代的多层次机制，系统实现了持续优化与稳定演进的统一，构建起具备自适应能力的闭环数据机制。

4.2 多模态一致性优化方法

在多模态融合场景中，信息之间的共存并不意味着共识。为解决图像与文本之间的语义不一致问题，系统构建了规则与学习模型相结合的判别信息冲突机制。当图文之间出现显著不协调时，系统通过实体匹配、情绪一致性等特征判定其为冲突，并交由二分类模型生成一致性标签。该标签不仅作为风险提示，也参与最终可信度评估。对于轻微不符，系统保留多值，并引导用户判别；存在严重矛盾时，则分析源头权重，若图像质量高且来源可信，文本将被部分修正。图文间的交叉验证机制也同步运行，文本验证视觉内容的语义边界，图像反向印证语言描述的准确性，彼此纠偏、互为补充。一致性模型也在不断学习，通过持续收集一致与冲突样本，优化判别精度，使其逐步掌握模态失衡的微妙边界。最终，系统将图文一致性评分作为融合结果的重要调节因子：一致者得高分，互斥者降权，严重冲突则触发审查。这不仅提升了结果的语义连贯性，也降低了误报率。

5 实验设计与性能评估

5.1 数据集与测试平台说明

为了全面验证系统在多模态情报处理任务中的适用性，笔者构建了一套覆盖文本、图像、语音的多源测试基座。在文本层面，选取多种中英文开源数据集，包括 CNN/REUTERS 新闻集（英文，评估摘要能力）、THUCNews/新浪新闻数据（中文，用于分类与实体识别），以及自建的社交媒体谣言集（含微博原帖与辟谣说明，检验可信度模型）。此外，采用 ACE2005 中文信息抽取语料作为 NER 与关系抽取的标准评测集。图像部分，除常用的 MS-COCO（微软常见物体上下文数据集）、Pascal VOC 与 ICDAR（国际文档分析与识别会议）印刷体文本测试集外，重点引入多个 Kaggle 假新闻图文数据集（如

Fakeddit），以测试图文一致性识别能力。同时构建了情报图像测试集，包含500组新闻照片与文字对，人工标注图文对应关系，用于跨模态对齐效果验证。TTS/ASR模块主要依赖主观评估，组织志愿听评团，对中英文播报样本各50段（10 s内）进行MOS打分，维度包括发音清晰度、自然度及情感贴合度。实验硬件平台为高性能本地服务器（Intel i7-14700K、NVIDIA A6000 GPU、128 GB内存），各模块通过Docker容器部署，实现模块级并行通信。为验证在资源受限场景下的部署弹性，系统还在一台搭载NVIDIA 4060Ti 16 GB的PC上运行了精简版本，运行稳定、功能完整。所有实验均重复运行，取平均值以保证数据的可靠性。

5.2 各模块性能对比

为量化系统在各关键子模块上的性能表现，将CV、NLP与TTS/ASR模块的核心指标进行整理。各项数据均来自标准测试集或实测评估，并与现有主流工具进行了横向对比，详见**表1**。

从结果可见，CV模块在目标检测与OCR任务上均显著优于主流开源方法，特别是在复杂社交媒体图像中表现出较强的稳健性。NLP模块依托大模型能力，在文本分类、实体识别与关系抽取方面取得了高于传统方法的F1与准确率，知识融合与多语种处理功能进一步扩展了实用性维度。TTS/ASR模块在平均意见得分（mean opinion score，MOS）主观测试中表现优异，生成语音自然流畅，播报时延低，支持自动语言切换，稳定性强，适用于连续播报场景。

整体来看，本系统各模块在精度、速度与可用性方面均达到了工程应用要求，构成了多模态智能情报系统的坚实技术底座。

表1 各模块性能对比

模块	子模块	数据集	系统表现	对比/优势说明
CV	目标检测	Pascal VOC	mAP 96.5%	优于YOLOv5(76%)
	OCR识别	ICDAR / 社交图	ICDAR字符识别 86.1% 社交图准确率 94%	远高于原始PaddleOCR(43%)
	线索聚合	自建情报图像集	线索聚合准确率 96%	人工核验一致性高
NLP	文本分类	THUCNews	准确率 96.3%	显著优于SVM(62.5%)
	实体识别	ACE2005	F1 = 88.7%	接近最新文献水平
	关系抽取	ACE2005	F1 = 62.4%	比基线系统高近15%
	谣言识别	微博谣言数据集	准确率约 85%	具备实用性
	知识融合	新闻+知识库案例	50条新闻中42条信息增强	有效补足上下文信息
	多语言支持	中英混合10篇	中英混合文本语言识别与切换准确	验证语义保持一致
TTS/ASR	MOS得分	TTS中英文听评样本	中文 4.6 / 英文 4.4(满分5)	优于商业TTS系统(4.3)
	播报速度		平均每条1.2 s	满足实时播报需求
	多语言发音	多语言情报语料	支持中英文自动发音切换	提升理解体验
	稳定性		连续播报100条无异常	稳定性强，零中断

5.3 系统端到端效果评估与响应分析

为验证系统的实际应用能力，笔者设计了一次端到端模拟测试，详细结果见**表2**。系统需处理10篇新闻、5条微博和3张配图，从抓取到播报完成共耗时约35 s，其中数据采集用时10 s，CV、NLP、TTS/ASR模块并行处理约20 s，决策与语音播报不超过5 s，响应速度达到秒级。在情报质量评估方面，3名分析人员对50份自动生成报告进行打分。90%被认为准确反映要点，70%体现出明显的信息增益（如背景补充、图文对齐等）。完整性和可读性平均得分分别为4.2与4.5，接近人工报告水准。与传统文本分析流程对比，使用本系统的团队平均用时减少83.8%，所提交报告在内容丰富度与准确性上均更优。这表明多模态融合与自动化调度机制在效率与质量两方面均具显著优势，具备落地潜力。

综上所述，实验评估证明了本系统在多模态情报提取融合方面的有效性和领先性。无论是各模块的性能指标，还是端到端的应用效果，本系统均表现出优异的性能。并且系统在确保准确性的同时，实现了情报处理流程的高度自动化和快速响应，在实际应用中有望极大减轻情报人员的负担，提升情报工作的智能化水平。

6 典型应用场景与落地

在舆情监控的图文联动识别场景中，本系统被持续部署于微博、论坛等社交高频场景。围绕突发事件、社会治安等主题，系统自动抓取图文内容，并以热度拐点为信号触发多模态分析流程。一旦事件的传播强度陡增，且内容经可信度模型判定为高置信，决策模块随即生成事件摘要，TTS/ASR模块同步启动，将警报通过语音迅速播报至值班席位。界面则联动展示图像、文本与热评，辅助人工直觉与机器判断融合，形成对舆情走势的即时洞察。实测中，系统往往能在事件爆发2~3 min内完

表2 端到端测试说明

评估维度	结果数据	测试说明
处理内容规模	10篇新闻+5条微博+3张图像	模拟突发事件多模态情报抓取
整体耗时	约35 s	系统整体从采集到播报时间
数据采集耗时	约10 s	含网络爬虫与任务调度
分析层耗时(CV、NLP、TTS/ASR)	约20 s	模块并行异步分析阶段
决策与播报耗时	约5 s	语音生成与提醒触发
预警响应时延	≤5 s	满足实时告警要求
情报报告准确率	90%(45/50)	由3位分析员主观评估
信息增益覆盖率	70%(35/50)	超出原始信息的深度补充
报告完整性评分	4.2 / 5	主观评估平均分
报告可读性评分	4.5 / 5	主观评估平均分
与人工用时对比	节省83.8%的分析时间	与传统单模态分析组对比
融合系统提升效果	准确率更高、内容更丰富	融合文本、图像、语音与知识库后结果增强

成全流程响应，为舆情控制争取黄金300 s。

金融风控情境则更具技术代表性。系统面向财经快讯、监管披露、舆情异动等多类信息源，搭建起一条精准采集与智能识别的处理链。文本经语义剖析后输入情报评分模型，结合实体引用、影响范围、波动特征等因子进行优先级判断。高风险情报将以语音形式在交易中枢播报，避免高频操作中因视觉忽视造成的判断滞后。尤其在行情突变、政策释放的窗口期，这种听得见的判断往往比屏幕上闪烁的数字更具先发优势。

政务安全场景则凸显了系统的全域整合能力。部署于封闭内网，平台集成社交媒体、新闻图文、监控截图、举报信件等异构数据源，构建起多模态线索采集体系。系统基于时间与空间的语义聚合逻辑，自动判定事件间的潜在关联，重构事件图谱，并通过地图加时间轴可视化情报脉络。遇警情上升，语音播报机制可直达决策中枢，实现分钟级闭环预警。在应急演练中，系统曾精准匹配求助帖与视频截图，协助现场定位与出警响应。更重要的是，其态势感知模块具备异源聚合、热点成形的能力，已被验证可服务于维稳、应急、反恐等高压任务场景。

7 结论与展望

本文设计并实现了一套融合多模态技术的开源情报智能处理系统，架构上采取模块化、分层式设计，打通了从数据采集、智能解析到反馈决策的全链条闭环。在技术融合方面，系统整合CV、NLP与TTS/ASR三大模块，有效解决了跨模态对齐、信息结构化表达以及语音可视化播报等关键难点，构建起一套具备实时响应与深度理解能力的情报处理机制。实验结果表明，该系统在情报抽取准确率与处理时延方面达到先进水准，多模态融合显著拓展了情报分析的边界。而在实际应用中，无论是舆情监测的快速预警、金融风控的即时提示，还是政务安全的线索整合与可视化展示，系统均展现出稳定高效的实用价值。

未来，系统将向轻量化版本拓展，适配低算力场景与边缘设备，真正做到端侧自治；语言处理模块将进一步多语种化，增强对文化语境差异的感知与容错；笔者计划引入更强大的GPT类语言模型，提升系统对复杂语境的推理能力，并优化多模态情报报告的生成质量。在深层语义推理方面，系统将逐步融合知识图谱与逻辑引擎，实现从数据到知识的跃迁，特别是在识别虚假信息与深度伪造方面，探索更具前瞻性的技术路线。最终，这一系统有望演进为支撑多行业、多语境、多源输入的智能情报平台。

参考文献：

[1] 袁唯淋，赵卫伟，胡振震，等. 智能情报融合综述：对抗视角下的开源情报融合分析[J]. 智能科学与技术学报，2024，6(3)：284-300.
Yuan W L，Zhao W W，Hu Z Z，et al. Summary of intelligent intelligence fusion：analysis of open source intelligence fusion from the perspective of confrontation[J]. Chinese Journal of Intelligent Science and Technology，2024，6(3)：284-300.

[2] 颜克冬，徐琳，宫小泽，等. 开源情报可信分析系统的关键模型与技术[J]. 指挥信息系统与技术，2023，14(1)：57-61，80.
Yan K D，Xu L，Gong X Z，et al. Key model and technology of open source in-

telligence trusted analysis system[J]. Command Information System and Technology, 2023, 14(1): 57-61, 80.

[3] U.S. Department of Defense. Department of defense open source intelligence strategy[Z]. 2025.

[4] Ech-Chammakhy Y, Motii A, Rabii A, et al. EventHunter: dynamic clustering and ranking of security events from hacker forum discussions[EB]. arXiv preprint, 2025, arXiv: 2507.09762.

[5] Josan G S, Kaur J. LSTM network based malicious domain name detection[J]. International Journal of Engineering and Advanced Technology, 2019, 8(6): 3187-3191.

[6] Browne T O, Abedin M, Chowd- Hury M J M. A systematic review on research utilising artificial intelligence for open source intelligence (OSINT) applications [J]. International Journal of Information Security, 2024, 23(4): 2911-2938.

[7] Zhang C, Yang Z C, He X D, et al. Multimodal intelligence: representation learning, information fusion, and applications[J]. IEEE Journal of Selected Topics in Signal Processing, 2020, 14 (3): 478-493.

[8] Chen J, Guo H, Yi K, et al. VisualGPT: data-efficient adaptation of pretrained language models for image captioning [C]//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2022: 18009-18019.

[9] Henrico S, Putter D. Intelligence collection disciplines: a systematic review[J]. Journal of Applied Security Research, 2025, 20(1): 46-70.

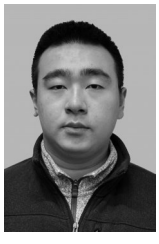
[10] Weir G R S. The limitations of automating OSINT: understanding the question, not the answer[J]. Automating Open Source Intelligence, 2016: 159-169.

[11] Imbert J, Dashyan G, Goupilleau A, et al. Improving performance of aircraft detection in satellite imagery while limiting the labelling effort: hybrid active learning[C]//Proceedings of the 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS. Piscataway: IEEE Press, 2021: 220-223.

[12] Li W T, Wang C C, Cui X X, et al. COSINT-agent: a knowledge-driven multimodal agent for Chinese open source intelligence[EB]. arXiv preprint, 2025, arXiv: 2503.03215.

[13] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision[C]// Proceedings of the 2021 International Conference on Machine Learning. [S.l.]: PMLR, 2021: 8748 - 8763.

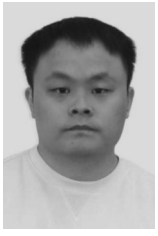
作者简介



董泽云（1997– ），男，上海安啼特科技有限公司创始人与执行董事，主要研究方向为人工智能、数据科学。

甘莅豪（1977– ），男，华东师范大学传播学院教授，主要研究方向为智能传播。

薛楠（1996– ），女，中国人民大学数学学院博士生，主要研究方向为数据科学、机器学习。



陆泰廷（1998– ），男，宾夕法尼亚州州立大学电气工程与计算机科学学院博士生，主要研究方向为人工智能、智能穿戴。

收稿日期: 2025-11-06

通信作者: 甘莅豪，lhgan@comm.ecnu.edu.cn；薛楠，xuenan@ruc.edu.cn