## 4 构建“3D–2T–MC”韧性治理协同框架

基于前述风险解构与国际比较，本文创新性地提出“三维动态定级 - 双轨技术融合 - 多元协同共治”（3D-2T-MC）韧性治理框架。该框架旨在通过动态适应、技术融合与多元协同，提升整个政务数据生态的韧性。

### 4.1 三维动态定级：实现安全合规的自适应演进

为解决静态分类分级的僵化问题，需构建一个能随数据属性、使用场景动态变化的智能定级系统。

构建政务数据敏感度评估指数（GDSPI）：设计一个量化模型，综合评估数据的敏感度，考量因素包括数据类型权重、精度权重和场景权重。通过算法计算得出综合敏感度得分，对应不同安全等级。

建立动态降密与销毁机制：制定实施指南，明确触发条件（如数据脱敏处理后、法定使用期限届满、场景变更导致风险降低）。创设数据销毁公证制度，要求责任单位通过国家认可的区块链存证平台提交销毁指令、执行日志与验证报告，实现全过程可追溯、不可篡改，增强公众信任。

### 4.2 双轨技术融合：打造自主可控的纵深防御体系

技术是实现安全的基石，需推动核心技术的深度融合与前瞻布局。

构建国家级 PETs“工具超市”与低代码平台：由国家牵头，整合多项核心 PETs 技术，形成一个开放、标准化的“工具超市”。同时，开发用户友好的低代码 / 无代码开发平台，使非专业技术人员也能快速部署基础隐私计算模块，降低技术应用门槛。建立 PETs 技术认证与评估体系，培育通过国家认证的服务商，保障技术自主可控与服务质量。

前瞻性布局抗量子加密（PQC）迁移：启动“量子安全迁移工程”，评估 NIST 标准化的 PQC 算法在政务场景的适用性。在涉及国计民生的关键领域率先试点部署。制定迁移路线图，明确完成核心系统改造的目标。建立量子 - 经典混合加密验证平台，确保新旧体系能平稳过渡，保障业务连续性。

### 4.3 多元协同共治：构建开放透明的治理生态

安全治理不能仅靠政府“单打独斗”，需构建政府、市场、社会共同参与的治理共同体。

强化顶层设计与跨部门协同：建议在中央层面成立最高协调机构，下设政策制定、技术标准、应急处置等专业委员会。建立关键部门轮流主持的“部际联席会议旋转主席”制度，确保决策的权威性与执行力。开发集成全国关键系统数据流动监测数据的“数据安全监管驾驶舱”，实现集中监控。

创新风险分担与市场激励机制：深化相关实践经验，系统性引入“数据安全责任险”。由大型保险公司设计标准化产品，保费根据数据规模、敏感等级、安全投入等因素分级，最高赔付额可达较高水平。此机制能有效转移财务风险，更重要的是，保险公司为控制赔付风险，会主动对投保单位进行安全审计与风险评估，形成强大的市场监督与激励机制，倒逼安全水平提升。

深化社会监督与公众参与：开发“政务数据透明度仪表盘”，实时、动态展示各部门数据调用日志、安全风险热力图、主要平台隐私保护合规评分等信息。建立强制性的数据保护影响评估（DPIA）公众听证制度，对涉及大规模人脸识别、健康码数据二次利用、跨部门数据融合等高风险项目，必须举行听证会，接受多方质询与监督，确保决策的民主性与正当性。

培育专业人才与“旋转门”机制：在专项计划中设立重点工程，培养首席数据安全官、数据合规审计师、网络安全运维工程师等专业人才。建立“政产学研旋转门”机制，促进高校专家与政府官员的知识、经验与人才双向流动。

## 5 结语

本研究系统分析了政务数据信息化进程中面临的技术、管理、法律复合型风险，揭示了现有治理模式的局限性。研究系统解构了各层面的风险耦合机制，通过国际比较提炼先进治理要素。在此基础上，创新性地提出了整合动态合规、技术融合与社会共治的“3D-2T-MC”韧性治理框架。该框架的核心价值在于其动态性、融合性和协同性。

未来研究与实践需在以下方向持续深化：

人工智能治理：生成式 AI 在政务场景的应用需建立严格的训练数据安全规范，防止模型泄露敏感信息，并建立算法偏见与决策可解释性的审计机制。

跨境数据流动：探索建立区域性数据安全合作框架，推动数据跨境流动的互认与安全标准对接，平衡数据开放与国家安全。

弹性防护体系：发展基于大数据分析与 AI 的攻击画像技术，构建具备预测、预警、自动响应与自愈能力的主动防御体系，全面提升政务系统的整体韧性。

唯有通过技术、制度、生态的协同创新，方能实现数据要素价值与安全屏障的动态平衡，为数字中国建设筑牢坚实根基。

## 参考文献

[1] 邓文宏. 大数据时代信息安全与隐私保护研究[J]. 中国新通信, 2017, 19(3): 56-58.

[2] 拖洪华. 大数据时代安全隐私保护技术探究[J]. 网络安全技术与应用, 2016,11(5): 88-89.

[3] 林璟锵, 任奎, 郑昉昱. 数字社会中的隐私计算技术进展[J]. 信息安全研究, 2025, 11(2): 112-120.

[4] 朱辉, 王伟. 政务数据共享中的隐私保护框架设计[J]. 计算机研究与发展, 2024, 61(4): 887-898.

 DOI: https://doi.org/10.12345/bdai.v6i6.34067

# Design of multimodal fusion intelligent defense system and analysis of adversarial attack defense effect

**Xinfei Dong**

Hainan International College, Communication University of China, Lingshui, Hainan, 572400, China

**Abstract**

This article focuses on the research of multimodal fusion intelligent defense systems. Firstly, a system design framework consisting of perception layer, fusion layer, decision layer, and feedback layer is constructed to confirm the three fusion strategies of data layer, feature layer, and decision layer, as well as key supporting technologies such as deep learning and robust feature extraction; Secondly, examine the three typical adversarial attack characteristics of intra modal, cross modal, and mixed modal, and design defense measures such as adversarial training and modal consistency verification; Furthermore, relying on relevant indicators such as accuracy and robustness, and relying on multimodal adversarial datasets to complete the evaluation of defense effectiveness; Finally, it points out the challenges faced by modal heterogeneity, adaptive attack defense, and looks forward to the direction of technological development.

**Keywords**

multimodal fusion; Intelligent defense system; Adversarial attacks; Defense effect

# 多模态融合的智能防御系统设计及对抗性攻击防御效果分析

董欣菲

中国传媒大学海南国际学院，中国 · 海南 陵水 572400

**摘　要**

本文以多模态融合的智能防御系统为研究焦点，首先构建由感知层、融合层、决策层及反馈层组成的系统设计框架，确认数据层、特征层、决策层三类融合策略及深度学习、鲁棒特征提取等关键支撑技术；其次审视模态内、跨模态、混合模态三种典型对抗攻击特征，设计对抗训练、模态一致性校验之类的防御手段；进而依靠准确率、鲁棒性等相关指标，依托多模态对抗数据集完成防御效果的测评；最后指明模态异构、自适应攻击防御等面临的挑战，并展望技术发展方向。

**关键词**

多模态融合；智能防御系统；对抗性攻击；防御效果

## 1 引言

随着信息技术的不断发展，网络攻击方式从单一向多手段、多方位、多元化方向发展，因此，将先进的人工智能技术应用到当下的网络安全防御系统中有助于全面提升网络安全防护能力。当前，智能系统已深度渗透至自动驾驶、医疗诊断、金融风控等关键领域，但其安全运行正面临多重挑战。多模态融合技术凭借整合视觉、听觉、文本、传感器等多渠道的数据，可达成信息互补及冗余去除，为增强智能系统抵御攻击的能力开辟新途径。然而，目前的研究在多模态防御系统架构设计、跨模态攻击防御机制以及防御效果标准化评估等方面依旧存在欠缺。因此，本文针对多模态融合的智能防御系统展开相关研究，有条理地梳理其设计框架、对抗性攻击的应对举措及防御效果评估方法，归纳当前研究碰到的瓶颈，进而展望未来方向，为各行业构建高鲁棒性的智能安全防护体系提供支撑力量。

【作者简介】董欣菲（2005–），女，中国山东日照人，在读本科，从事智能科学与技术研究。

## 2 多模态融合智能防御系统设计框架

### 2.1 系统整体架构

多模态融合智能防御系统的整体架构以“防御导向”为关键核心，分成感知、融合、决策、反馈四个彼此协同的层级。感知层的职责是进行多模态数据的采集与预处理，需依照不同模态的特性采用适配技术，比如针对视觉数据过滤噪声、针对文本数据进行语义清洗、针对传感器数据去除异常值，同时利用时间同步和空间校准达成多源数据的初步对齐，为后续的多源数据融合打下基础。融合层是架构里的核心要素，担起多模态信息的协同整合任务，必须确立模态间的关联逻辑，保证异构数据实现有效交互，规避信息存在冲突或冗余。决策层按照融合后的信息生成防御策略，需平衡

实时性和准确性，好比在工业控制情形中快速识别攻击信号并实施拦截动作 [1]。反馈层借助防御效果数据反向对各层级参数进行优化，打造“采集 - 融合 - 决策 - 优化”的闭环体系，提升系统长期防御的坚实性，各层级借助标准化接口达成衔接，保证数据传输与指令执行达到高效水平。

### 2.2 多模态融合策略设计

多模态融合策略需依照应用场景需求选择匹配方案，主要归类为数据层、特征层与决策层三类。数据层融合是针对原始数据做操作，先采用模态对齐技术对数据异构性进行消除，然后采用标准化做法统一数据格式，该策略可最大限度地保存原始信息，只是对数据质量的要求偏高，适合医疗影像等对精度要求高的场景。特征层融合聚焦抽象特征的彼此交互，采用 Cross-Attention 注意力机制增强关键模态特征的权重大小，也借助多模态 Transformer 达成跨模态特征深度融合目标，高效提取模态间的互补信息，不过计算复杂程度较高，更贴合算力充足的安防监控等应用场景。决策层融合基于各模态独立决策结果整合，多采用投票法、加权求和法或者 D-S 证据理论，采用动态调整模态权重的办法保障决策可靠性，虽说精度略低，但响应速度快，适合应用于自动驾驶等实时性要求高的场景。

### 2.3 防御系统关键技术支撑

防御系统实现高效运行依赖三类核心技术的支撑。深度学习模型为系统赋予基础处理能力，CNN-LSTM 混合模型可开展时空关联的多模态数据处理工作，多模态 Transformer 可捕获跨模态的语义关联，GAN 可生成对抗样本用于防御训练相关事宜，提升系统抵抗攻击的能力。作为防御核心的是鲁棒特征提取技术，依靠对抗训练引导模型学习不受攻击干扰的特征，减小扰动引发的不良影响，同时夯实模态的冗余特征，如文本语义跟图像内容的关联特点，保证某一模态遭到攻击的当口，其他模态可给出补充性信息。联邦学习与隐私保护技术处理多模态数据共享难题，采用分布式训练达成跨机构数据协同利用，规避原始数据泄露引起的攻击危机，尤其适用于金融风控、医疗诊断等数据敏感领域，为防御系统大范围的规模化应用提供安全保障 [2]。

## 3 多模态场景下的对抗性攻击类型与防御机制

### 3.1 多模态对抗性攻击的典型类型与特征

多模态场景下的对抗性攻击主要归为三类，且都围绕“依靠模态特性或关联打破防御”而展开。第一类是模态内攻击，针对单一模态开展精准干扰，如向图像数据中加入人眼难以分辨的细小噪声，引发视觉分类模型误判现象；又或者在文本数据里替换同义干扰词语，打乱语义理解模型的决策逻辑。这类攻击的特点是针对性十分强，仅仅影响某一单一模态，而扰动幅度一般都较小，不容易被人察觉。第二类是跨模态攻击，借助多模态之间的关联性开展间接攻击，就如通过篡改商品的相关文本描述，诱导图像分类模型把危险品错认成普通物品；或者采用伪造语音指令，干扰视频监控系统对人物行为的甄判。其核心特性是“以一扰多”，凭借模态间信息传递关系去扩大攻击影响，防御的难度远比模态内攻击要高。第三类是混合模态攻击，同时针对多个模态施加协同性干扰，如在自动驾驶场景中，既篡改由摄像头捕捉的交通标志图像，又对毫米波雷达的距离检测数据进行扰乱。这类攻击呈现出攻击强度大、成功率高的特性，可同时毁坏多模态系统的信息源，非常容易导致系统整体失效。

### 3.2 多模态融合导向的防御机制设计

针对多模态对抗性攻击的特性，防御机制应依托“融合优势”构筑多层防护体系。首先是对抗训练的防御机制，在模型训练阶段添加多模态对抗样本，比如往图像 - 文本数据集当中加入跨模态攻击样本，让模型摸索攻击扰动的规律，增进对各类攻击的适应水平。该机制可从根源上提升模型的鲁棒性，尤其对于模态内攻击的防御效果极为显著。其次是模态一致性校验机制，凭借多模态信息的内在关联设计校验准则，比如文本描述和图像内容不一致、语音指令与行为动作不匹配时，自动触发守护预警。此机制能切实识别跨模态攻击，利用排除矛盾信息降低误判率 [3]。最后是动态权重融合防御机制，实时监测各模态的完整程度与可靠程度，若检测到某一模态受到了攻击，自动把该模态的决策权重降低，同时增强其他未受波及模态的权重。该机制能灵活地对混合模态攻击作出应对，保障系统在部分模态失效时依旧能维持基础防御能力，兼顾防御效果和系统可用性。

## 4 多模态智能防御系统对抗性攻击的效果评估

### 4.1 评估指标体系

评估指标应兼顾防御的有效性与系统的实用性，创建多维度量化体系。性能指标是核心，包含防御后系统相关的准确率、鲁棒性与误报率：准确率反映出正常及攻击场景下决策的正确状态，鲁棒性依靠攻击扰动下准确率的降幅进行衡量，说明其抗干扰能力越出色，要把误报率控制到较低水平，防止影响系统正常运行。效率指标要与实际应用需求相匹配，着重留意防御响应时长与计算资源的消耗，如自动驾驶场景需把时间控制在 100ms 以内，工业控制场景需要降低 GPU 显存的占用以及 CPU 使用率，防止复杂算法造成系统卡顿。泛化性指标审定防御能力的通用水平，包括对不同攻击类型防御的契合度，以及在医疗、安防、交通等不同行业场景上的迁移实际成效，保证指标全面覆盖防御系统的核心价值要素 [4]。

### 4.2 评估数据集与实验设计

评估要借助具有针对性的数据集与科学实验设计，保障结果的可靠程度。优先选择标注完备的多模态对抗数据集作为数据集，CMU-MOSEI 对抗版包括文本、语音、视频三模态内容，含有跨模态攻击的样本；MSCOCO 对抗版把焦点放在图像 - 文本模态，给出模态内的噪声攻击样本；