面向不可靠网络证据的认知驱动型多模态事实核查方法

王植琴[1),2)] 王 蕊[1),2)] 荆丽桦[1),2)] 刘俐君[1),2)] 吕飞霄[1),2)]

[1)](中国科学院信息工程研究所 北京 100093)

[2)](中国科学院大学网络空间安全学院 北京 100049)

摘 要 在当代社会,社交媒体极大地提高了信息传播的效率,但也为虚假信息的快速扩散提供了温床。虚假信息的传播不仅削弱了媒体的公信力,损害公众的知情权,更会引发舆论失序、社会信任危机等深远影响,严重威胁网络信息生态的健康发展。当前,虚假信息制作技术日趋复杂,传播形式已由简单文本升级为图文拼贴的多模态形式,其隐蔽性和欺骗性显著增强。这种演变使得虚假信息检测方法面临严峻考验,新型检测技术不仅需破解多模态内容间的语义关联陷阱,更要结合不同来源的共现信息进行动态事实核查。现有的多模态事实核查方法虽然能够通过自动收集和比对网络证据来完成这一耗时且需要复杂推理的过程,但在实际应用中存在三个核心局限性:(1)面对网络证据固有的不可靠性,现有方法缺乏有效的质量感知与筛选机制,难以应对证据质量参差带来的干扰;(2)在证据核查层面,现有技术框架未能融入人类认知逻辑,导致多角度比对与推理能力不足;(3)在决策阶段,现有方法忽视了多重验证线索间的内在关联性,使得整个判定过程缺乏可解释性,严重制约了核查结果的可靠性。这些局限性共同导致了现有方法在处理复杂多模态网络证据时的性能瓶颈。本文创新性地提出了一种面向不可靠网络证据的认知驱动型多模态事实核查框架。该框架的核心是模拟人类认知过程的"关注-比对-判定"三级推理机制:在关注阶段,通过设计的双通道注意力网络(相关性注意力和有效性注意力)实现证据的多维度质量感知,从证据相关性和有效性两个关键维度进行智能筛选;在比对阶段,采用全局-局部协同的多模态特征分析与比对策略,实现待核查信息与证据的多粒度深度匹配;在判定阶段,通过创新的自赋权多分类器集成机制,有效解决多核查分支的判定模糊性与冲突问题,提升决策可靠性。在多个基准数据集上的实验表明,本方法相较现有最优技术取得显著提升(性能提高超1.5%),尤其在处理噪声数据和复杂多模态证据时展现出卓越的鲁棒性和场景适应能力。

关键词 多模态事实核查;虚假信息检测;谣言检测;多模态理解;注意力机制

中图法分类号 TP18 **DOI号** 10.11897/SP.J.1016.2026.01209

A Cognitive-Driven Multimodal Fact-Checking Method for Unreliable Web Evidence

WANG Zhi-Shen[1),2)] WANG Rui[1),2)] JING Li-Hua[1),2)] LIU Li-Jun[1),2)] LV Fei-Xiao[1),2)]

[1)](Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093)

[2)](School of Cyber Security, University of Chinese Academy of Sciences, Beijing 100049)

Abstract In contemporary society, social media has greatly enhanced the efficiency of information dissemination, but it has also become a breeding ground for the rapid spread of misinformation. The proliferation of misinformation not only undermines the credibility of the media and infringes upon the public's right to know, but also leads to disordered public opinion

收稿日期:2025-07-14;在线发布日期:2026-01-29。本课题得到国家自然科学基金面上项目(No. 62176253)资助。**王植琴**,博士研究生,主要研究领域为计算机视觉以及多模态模型。E-mail:wangzhishen@iie.ac.cn。**王 蕊**(通信作者),博士,研究员,中国计算机学会(CCF)会员,主要研究领域为计算机视觉以及深度学习。E-mail:wangrui@iie.ac.cn。**荆丽桦**,博士,助理研究员,主要研究领域为计算机视觉以及深度学习。**刘俐君**,博士,主要研究领域为计算机视觉以及知识蒸馏。**吕飞霄**,博士,主要研究领域为计算机视觉以及图像描述生成。

and crises of social trust, posing a serious threat to the healthy development of the online information ecosystem. Today, misinformation is becoming increasingly sophisticated in its creation, evolving from simple textual content to multimodal forms that combine both text and images, significantly enhancing its concealment and deceptiveness. Unlike unimodal fabrication, multimodal misinformation increasingly relies on semantic manipulation—such as pairing authentic, unaltered images with conflicting or fabricated captions—to construct a false sense of authenticity that is cognitively difficult to reject. For instance, the strategic use of context-displaced imagery capitalizes on the inherent persuasiveness and perceived objectivity of visual media, creating a "seeing is believing" trap that is harder to debunk. This cognitive bias effectively lowers the psychological threshold for believing falsehoods, making users less likely to critically verify information when it is accompanied by multimodal proof. This evolution also presents a severe challenge to misinformation detection technologies. New detection methods must not only bridge the semantic gaps between modalities but also dynamically verify facts using co-occurring information from diverse sources. Although existing multimodal fact-checking methods can automate the labor-intensive and reasoning-intensive process of evidence retrieval and comparison, they face three fundamental limitations in practice: (1) In the face of inherently unreliable web evidence, current methods lack effective mechanisms for quality perception and filtering, making them vulnerable to interference from inconsistent or low-quality evidence; (2) At the analysis level, these methods fail to incorporate human-like cognitive logic, resulting in inadequate multi-perspective reasoning and comparison capabilities; (3) By ignoring the internal correlations among multiple verification clues, they suffer from a lack of explainability in the decision-making process, thereby undermining the overall reliability of fact-checking results. These limitations together contribute to the performance bottlenecks of current approaches when dealing with complex multimodal web evidence. To address these challenges, we propose a novel human-cognitive-driven multimodal fact-checking framework for unreliable web evidence. At its core, the framework simulates the human cognitive process through a three-stage reasoning paradigm: Attend, Compare, and Determine (ACD). In the attend stage, we design a dual-channel attention module—comprising relevance attention and validity attention—to enable multidimensional quality perception and intelligent evidence screening based on both relevance and reliability. In the compare stage, we employ a global-local collaborative strategy to perform fine-grained feature analysis and matching between the query and the retrieved evidence across modalities. In the determine stage, we introduce an innovative self-weighted multi-classifier ensemble, which dynamically integrates signals from multiple verification branches to mitigate decision ambiguity and enhance the robustness of the final judgment. Extensive experiments on multiple benchmark datasets demonstrate that our method achieves significant performance improvements over state-of-the-art baselines (with over 1.5% performance gain), and exhibits superior robustness and adaptability, particularly when handling noisy or complex multimodal evidence.

Keywords multi-modal fact-checking; misinformation detection; rumor detection; multi-modal understanding; attention mechanism

1 引 言

在当代社会，网络社交媒体已成为信息分享、获取和汇聚的主要平台。这些平台提供了快速获取大量信息的便捷途径。然而，同时它们也促进了虚假信息的快速传播，尤其在一些重大事件中[1]，这种传播对社会产生了显著的负面影响，媒体公信力、公众

知情权和网络信息的活力均遭受巨大损害。伴随deepfake技术[2-3]和生成式人工智能[4]的发展，虚假信息的制造变得更加隐蔽，且数量激增。随着大型语言模型的出现，情况变得更加复杂，因为这些模型可能被故意滥用来生成错误信息，或因幻觉问题而错误地传播错误信息[5]。在此背景下，自动识别网络信息真实性已成为紧迫课题，研究者们积极探索有效的虚假信息检测方法[6-9]。

早期的虚假信息检测研究主要聚焦于对单一模态内容（如纯文本或图像数据）的独立分析，一般通过验证输入内容的逻辑结构、语义特性和篡改痕迹来实现检测目标[10-14]。这类方法通常依赖于识别特定模态中的异常特征，例如文本中的语气特性、语义矛盾，或是图像中的人工处理痕迹（如PS痕迹、生成痕迹等）。随着虚假信息制作技术的演进和传播形式的复杂化，现代虚假信息往往采用多模态融合的表现形式，将文本、图像、视频等多种媒体元素有机结合以增强欺骗性。这种转变使得传统单模态检测方法的局限性日益凸显，而基于多模态分析的检测技术则展现出显著优势[15-19]。新兴的多模态检测方法不仅分析文本语义特征，还整合了图像内容理解、视频时序分析、网络传播元数据等多种维度的信息，通过挖掘跨模态间的关联性和一致性特征，构建起更全面的虚假信息识别框架。

随着虚假信息检测技术的不断发展，传播者也在持续调整其对抗策略，催生出一类隐蔽性更强的新型虚假信息形态。这类信息往往借助真实素材的“局部可信性”掩盖其整体欺骗意图，对现有检测方法构成了新的挑战。如图1所示案例，此类信息的典型特征在于：其构成元素（如图像、文本）单独审视可能完全真实，甚至具备良好的跨模态一致性，但经过刻意组合后，整体传达出与原始语境截然不同的语义内涵。具体而言，传播者可能从不同真实来源中提取可信的图片与文本，通过重新拼接构建出具有误导性的新语境。这种“真实元素＋虚假关联”的构造方式，使传统检测方法面临双重困境：单模态分析难以识别独立真实元素中的篡改痕迹，而常规跨模态检测又易被表层一致性所迷惑。此类“脱离上下文的虚假信息”凸显出现有检测范式的局限性，其核心挑战在于必须超越对内容本身的孤立分析，构建基于多源交叉验证的事实核查方法。

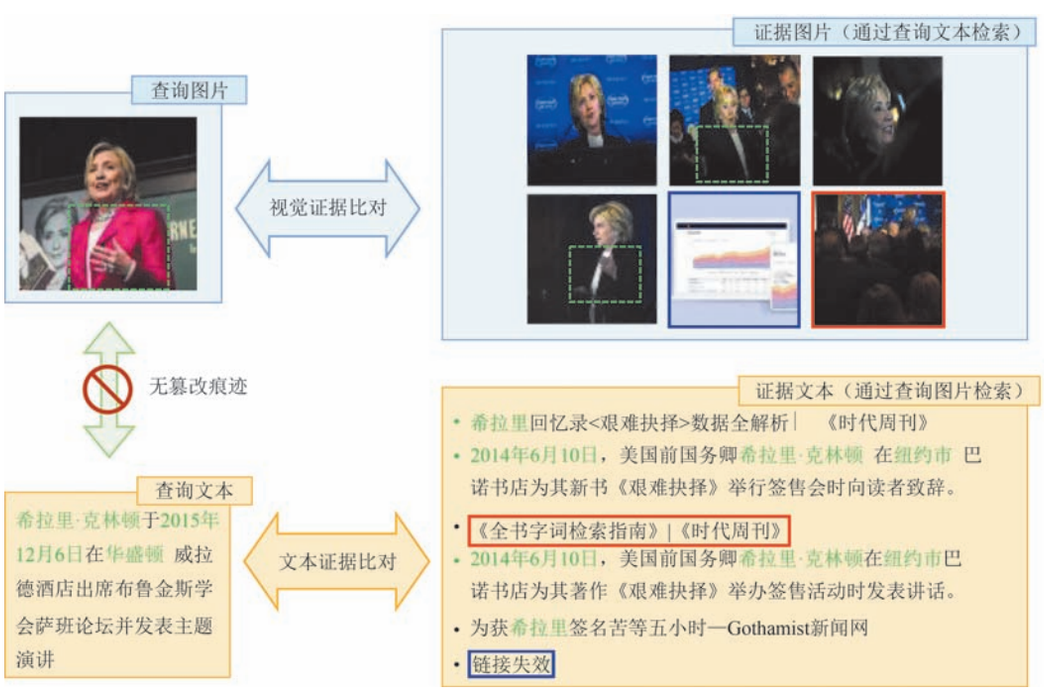


图1　事实核查流程示例

如图1所示，有效的事实核查需搜集与待验证信息相关的多个网络证据源，并综合判别其真实性。近年来，基于多源共现信息的自动化事实核查方法逐渐成为研究重点，并已取得一系列显著进展[20-23]。然而，现有研究仍存在若干关键局限性。首先，由于证据来自网络，具有固有的不可靠性，自然带来两方面的噪声：(1)无效证据造成的噪声，如无意义文本或图像（图1蓝色框所示）；(2)无关证据造成的噪声，即既不支持也不反驳查询内容的证据（图1红色框所示）。现有方法缺乏有效的质量感知

与筛选机制，难以应对证据质量参差不齐所带来的干扰。其次，虚假信息常对真实信息中的某些细节大做文章，如文本主张中的时间、人物、地点，图像中的关键区域等，这些事件特定细节（如图1中的绿色虚线框和绿色加粗文本）对检测极为关键，但以往研究通常只关注全局信息而忽视了这些细节，并且未形成类人认知逻辑的多角度证据比对与推理机制。第三，以往方法通常将不同验证分支的特征拼接后输入单一分类器，忽略了不同分支间多条验证线索的关联性和矛盾性，使得整个判定过程缺乏可解释性，严重制约了核查结果的可靠性。

为解决上述问题，本文创新性地提出了一种认知驱动型多模态事实核查框架，该框架的核心是模拟人类认知过程的“关注-比对-判定”（Attend，Compare，Determine；简称ACD）三级推理机制实现可靠的多模态事实核查。在“关注”阶段，我们基于有效性和相关性两个关键维度对证据进行质量评估，以实现精准核查。为此，本文创新性地提出了一种双通道注意力机制：首先，有效性注意力通过分析证据特征间的内在关联，自动识别并强化有效证据的表示；其次，相关性注意力模块通过计算待查询信息特征与各证据特征的交互权重，动态突出与当前待查询信息最相关的证据内容。通过这种双通道注意力协同工作机制，本框架能够从多个维度全面感知证据质量，实现智能化的证据筛选。在“比较”阶段，我们从全局和局部双重视角进行全面事实核查。在全局视角验证中，一方面关注待查询信息与证据的全局语义一致性比较，另一方面引入待查询信息内部跨模态内容比对分析，以应对证据缺失场景；在局部视角验证中，则重点聚焦待查询信息与证据之间的事件特定细节，实现宏观语义与微观细节的双重验证。这种多层次的验证机制既能把握整体语义关联，又能精准捕捉关键局部特征的匹配程度。在“判定”阶段，本文设计了一种自赋权多分类器集成机制，用于动态调节不同验证分支的贡献，有效解决多核查线索融合的模糊性问题。

本文贡献总结如下：

（1）引入双通道证据感知机制：针对网络检索证据中存在的大量无效或无关信息，我们从“有效性”与“相关性”两个认知维度出发，引导模型关注高质量证据，有效降低证据信息噪声，提高核查的准确性与鲁棒性。

（2）提出双视角验证机制：本方法从人类认知推理的角度出发，设计了一种结合全局视角与局部视角的验证方式，能够对待查询信息与证据之间进行多层次、多粒度的比对，从而实现更加全面和深入的事实核查。

（3）构建自赋权多分类器结构：为解决不同验证分支间信息融合不足与解释性弱的问题，我们设计了一种自赋权机制，能够自主评估和调节各类核查线索的贡献程度，显著增强核查过程的可解释性与判定可靠性。

（4）通过系统性实验验证方法有效性：在多个公开多模态事实核查数据集上进行的实证研究表明，本文所提出方法在准确率、鲁棒性与解释性等方面均优于现有最先进技术，展示了显著的性能提升和广泛的适应能力。

2 相关工作

2.1 基于内容分析的虚假信息检测方法

基于内容的多模态虚假信息检测方法是虚假信息检测领域研究最为广泛的方向。这类方法基于以下假设：虚假信息往往通过特定的表现形式来说服人们相信其真实性，那么从机器学习的角度来看，虚假信息的内容特征与真实信息存在显著差异。早期的虚假信息检测研究主要集中于单模态内容分析，包括纯文本[12,24]或纯图像[10]的数据分析。Ma等人[12]首次将RNN应用于社交媒体谣言检测当中，通过LSTM网络捕捉谣言的时序语言特征。Alkhodair等人[25]采用Word2Vec生成词向量并结合LSTM网络捕捉文本序列特征，引入注意力机制突出关键语义，来实现新闻谣言的检测。Cao等人[10]则开展了一项综合研究，涵盖图像取证特征、语义特征、统计特征和上下文特征等多个维度进行虚假信息检测。杨延杰等人[26]提出了一种融合门控的传播图卷积网络（GUCNH），通过门控机制动态筛选邻居节点信息，并结合多头自注意力建模全局传播关系，强化源帖特征以提升谣言检测的鲁棒性。苏兴等人[27]提出一种层次门控交互融合网络，通过多层级门控机制动态融合文本与传播结构特征，显著提升了谣言检测性能。

随着网络信息越来越多地采用图文结合的多模态呈现形式，单模态方法在检测跨模态关联和多模态信息一致性方面显得力不从心。相比之下，多模态虚假信息检测方法展现出更优越的性能。Yang等人[28]提出的TI-CNN模型利用CNN分别提取文本和图像特征，然后将其拼接为联合特征。此外，他

们还创建了首个多模态虚假信息数据集“Twitter”。Wang等人[19]开发了事件感知神经网络EANN，通过提取事件不变特征来增强模型的鲁棒性，使其能够检测与新事件相关的虚假信息。然而，这些方法仅简单拼接不同模态的特征，而忽视了模态间的交互融合。王友卫等人[29]构建了事件-词语-特征异质图，通过融合语义与传播异质信息，实现了微博谣言的高效检测。Jin等人[16]设计了一种带有注意力机制的RNN模型，能够有效融合视觉和文本模态的特征，并构建了首个中文多模态虚假信息数据集“微博”。Khattar等人[17]提出的MVAE模型利用变分自编码器学习多模态信息的共享表示，发现不同模态间的相关性。随着注意力机制[30]在多模态任务中的优异表现，更多研究尝试利用其生成更具交互性的多模态融合特征。Qian等人[31]提出的HMCAN模型采用自注意力建模单模态特征，并利用交叉注意力增强跨模态特征；Qi等人[18]开发的EM-FEND模型则将视觉实体和文本实体等更多特征融入注意力计算，以获得更精细的跨模态特征，同时考虑了跨模态不一致性问题。Zhou等人[32]提出了MMFN，通过结合粗粒度语义和细粒度跨模态交互，并在模态冲突情况下自适应调整融合权重，用于提高社交媒体上的多模态假新闻检测性能。Ma等人[33]提出的Event-Radar用事件级图结构建模多视角，通过多视角融合与事件图推理检测假新闻。Yin等人[34]提出的GAMC采用图自编码器、掩码增强和对比学习的无监督框架，从传播/交互图结构中学习判别表示，避免对大量带标签数据或大语言模型的依赖。Wu等人[35]提出的UEEI框架，侧重从评论/外部证据检索与多视图推理来发现新闻中的“可疑片段”，构建实体/关系层面的证据检索与多视图一致性推理。Yu等人[36]提出了RaCMC框架，通过多尺度残差补偿与多粒度约束增强跨模态特征融合并放大真假新闻差异，从而显著提升多模态假新闻检测性能。

2.2　基于证据比对的虚假信息事实核查

基于证据比对的虚假信息事实核查是一种新兴的检测方法，其核心在于整合多源共现信息进行交叉验证，通过外部收集的证据来验证信息真实性，更贴近人类判断信息真伪的认知过程。早期的研究主要集中在文本证据领域：Popat等人[37]开发了首个端到端的文本声明证据感知可信度评估框架；Augenstein等人[38]则通过聚合不同事实核查网站的文本证据和丰富元数据进行验证。随着多模态虚假信息的泛滥，研究者开始探索融合多模态特征的核查方法。Yao等人[39]通过收集文本和视觉证据并生成解释性验证描述来核实文本声明，但其验证目标仍局限于文本内容。Zlatkova等人[40]研究了图像-声明对的事实性，但仅依赖图像的标签或URL等文本信息，未能充分利用视觉特征。Factify系列研究[41]通过收集社交媒体虚假信息及对应证据文档，构建了首个多模态事实核查数据集，推动了该领域的重大进展[21,42-43]。Logically[42]提出了两种基线方法——单模态组件集成模型和用于建模声明-证据间图文交互的多模态注意力网络；INO[43]综合运用句长、词汇相似度、语义相似度和图像相似度特征，采用随机森林分类器进行评估；Triple-Check[21]则整合了高效嵌入、多模态融合、元数据表示和统一集成机制以提升性能。

随着虚假信息检测技术的发展，虚假信息传播者也在不断进化其对抗策略，催生出一类更具迷惑性的脱离上下文的虚假信息形态[44-46]，其一般表现为声明中的图文内容看似真实，但存在故意错配或篡改，其通过精心设计的“真实元素＋虚假关联”的构造手法，以部分真实内容隐藏整体的虚假本质。解决问题的关键在于超越对单一内容的孤立检验，建立融合多维度信息相互印证的综合性识别框架。为实现对脱离上下文式虚假信息的有效检测，Abdelnabi等人[20]开创性地提出了循环一致性检测范式。该范式的核心创新在于通过互联网交叉检索获取图文证据，以此辅助评估待查询信息的真实性。研究团队还扩展了NewsCLIPpings[44]数据集，新增网络证据构建了脱离上下文（Out-of-Context，OoC）虚假信息事实核查数据集，为合成式虚假信息检测领域提供了首个系统性研究框架和基准方法。为验证循环一致性范式在真实场景中的适用性，Hu等人[22]进一步开展了实证研究。他们从微博和推特平台采集真实传播的多模态虚假信息实例，通过交叉证据检索构建了包含相关证据链的MR2数据集。实验结果表明，循环一致性方法在真实场景中仍保持良好检测性能。MR2数据集的建立具有双重意义：一方面显著提升了合成式虚假信息的数据多样性，另一方面通过真实社交媒体环境下的样本采集，为事实核查技术的实际应用提供了重要的数据支撑和评估基准。

针对循环一致性检测范式，学界持续探索更高效的多模态事实核查方法。Zhang等人[23]提出了一种基于实体和场景检查的多模态事实核查框架ESCNet（实体和场景检查网络），该方法通过联合建模场景语义特征和实体知识表征来进行检测，首先提取文本、图像和实体特征，再利用场景变换器计算声

明与证据的跨模态一致性；其特色在于通过知识图谱检索实体语义路径，并采用符号推理构建知识级表征。为进一步提升模型可解释性，Zhang等人[47]提出了一种可解释的上下文增强网络（ECENet），该方法引入基于强化学习的解释生成模块，通过训练智能体选择关键证据句子并重写生成解释，无需额外标注。最近的一些研究开始尝试使用大模型来解决事实核查的问题，例如，Sniffer方法[48]尝试利用多模态大语言模型的视觉推理能力，设计两阶段指令微调，使MLLM可以适用于多模态事实核查任务。Lee等人[49]通过在开放多模态模型上引入跨数据集的知识迁移和利用大型语言模型生成的解释作为辅助监督，从而提升事实核查的准确性与泛化能力。但是基于大模型的事实核查方法不可避免地容易受到大模型幻觉的影响，因此这类方法仍处于初步探索阶段。我们在表1中列举了部分方法实现细节的比对，这些研究成果不仅为本研究奠定了坚实的理论基础，同时也为后续研究提供了重要的方向指引。

表1 相关工作的实现细节比对

方法	证据	视觉感知	文本感知	推理
MVAE	×	VGG	LSTM	VAE
EMFEND	×	VGG	BERT	Cross-Attention
MMFN	×	Swin+CLIP	BERT+CLIP	Cross-Attention
GAMC	×	-	BERT	GNN
RACMC	×	ResNet+CLIP	BERT+CLIP	Cross-Attention
Logically	√	ResNet	BigBird	Cross-attention
INO	√	ResNet+CLIP	SBERT+CLIP	Feature-fusion
3-Check	√	Swinv2	DeBERTa	Cross-Attention
OOC	√	ResNet	BERT	Memory-Network
ESCNet	√	ResNet	BERT	Cross-Attention
ECENet	√	ResNet+CLIP	BERT+CLIP	CFgAN
Sniffer	√	ViT	Vicuna	LLM
ACD(本文)	√	CLIP	CLIP	Self-Attention

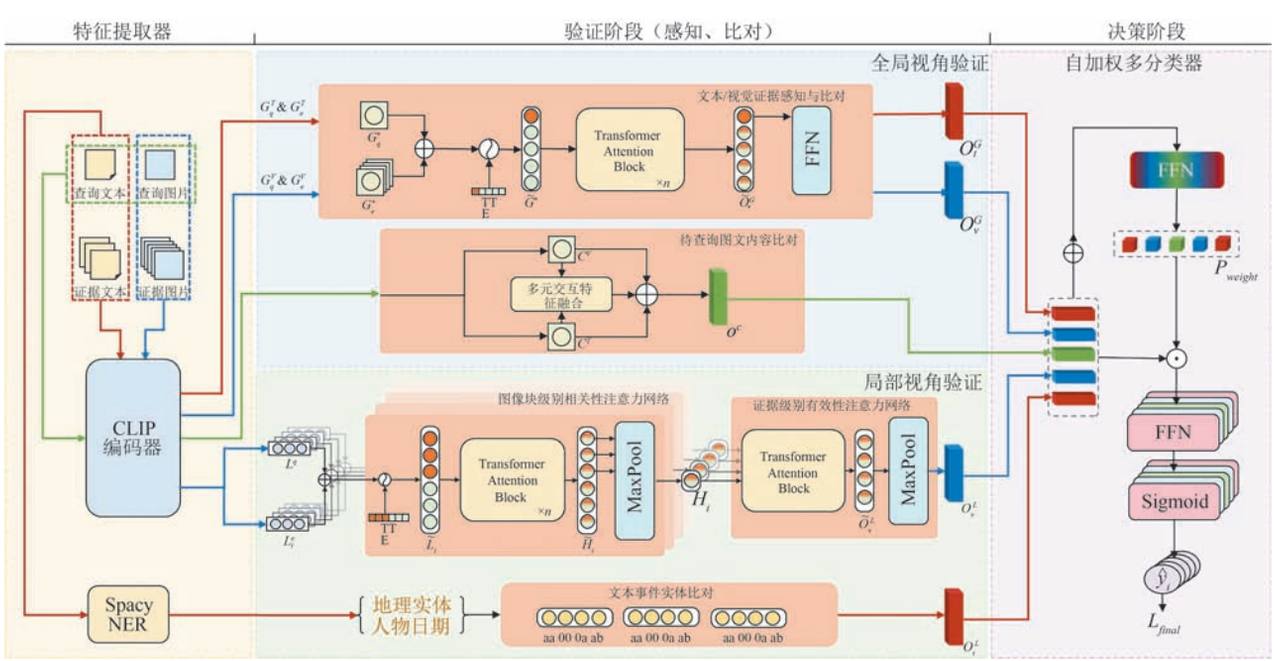


图2 本文提出的认知驱动型多模态事实核查方法ACD的框架图

3 方法论述

本节首先在3.1节中整体介绍多模态事实核查任务的问题定义以及本文所提出的ACD方法的模型框架和数据处理流程。然后在3.2节中介绍不可靠证据带来的噪声的特点以及本文提出的创新性双通道注意力证据感知机制；随后在3.3节和3.4节中

分别从全局视角和局部视角两个方面介绍我们的核查流程，并分别在其中介绍所提出的双通道注意力感知机制的具体实现方法；最后在3.5节中介绍自赋权多分类器的设计并对算法中涉及的损失函数进行梳理。

3.1　问题定义与模型框架总览

本文遵循的多模态事实核查任务的循环一致性检测范式，通过跨模态证据检索与双向交叉验证来评估多模态图文信息的真实性，具体而言，首先基于待查图像检索网络中的关联文本证据（如网页标题或描述），将其与待查文本内容进行相关性核查比对；同时，利用待查文本检索相关视觉证据（如其他关联图像），并与原图像进行视觉一致性核查。如图3所示，这种双向闭环的多模态循环一致性核查范式通过文本→视觉和图像→文本的双向证据链交叉印证，为真实性判断提供依据。为了参数化表述多模态虚假信息事实核查任务，本文将问题形式化为一个二分类问题，即每个待查询文本-图像对（为方便表述，后文简称为查询）可被分类为真实（$y=0$）或虚假（$y=1$）。每个查询文本-图像对$Q=(T^q, V^q)$对应数量不等的文本证据和视觉证据$E=(T^e, V^e)$，其中文本证据$T^e=\{T_1^e, T_2^e, \cdots, T_{n_t}^e\}$是利用Google反向图像搜索API在互联网中检索与查询图像相关的文本（如图片标题）获得。类似地，图像证据$V^e=\{V_1^e, V_2^e, \cdots, V_{n_v}^e\}$是利用Google可编程搜索引擎通过检索与查询文本相关的图像获取的。我们的目标是学习一个事实核查方法F：$F(Q,E)\to y$，其中$y\in\{0,1\}$表示查询文本-图像对的真实标签。

图3　循环一致性检测范式示意图

本文提出名为“关注-比较-判定”（ACD）的新型认知驱动多模态事实核查框架，通过模拟人类事实核查的三级推理机制，实现在面向证据不可靠场景下更全面、更有效的多模态虚假信息检测。如图2所示，本框架工作流程如下：在“关注”阶段，首先通过预训练多模态编码器提取多模态特征，然后设计了双通道注意力机制，分别从证据的有效性和相关性两个维度进行质量评估，显著降低了噪声干扰；在

“比较”阶段，采用全局与局部相结合的双视角验证机制，通过文本分支和视觉分支分别对待查询信息与证据进行多层次、多粒度的事实比对，并在全局视角中引入跨模态内容比对作为辅助判断分支，以应对证据缺失情况；在“判定”阶段，引入自赋权多分类器集成策略，通过可学习的权重生成网络综合分析各验证分支的预测确定性与特征区分度，动态融合多路验证线索，最终生成具有可解释性的加权决策。整个框架通过各个模块的有机协同，实现了对不可靠证据的智能感知、多角度证据比对，以及复杂判定条件下的稳健决策。该设计遵循“关注-比较-判定”的类人认知逻辑，确保事实核查过程的全面性和可靠性。

3.2 双通道注意力证据感知

由于证据是从互联网动态采集而来，其来源广泛且形式多样，导致证据质量参差不齐，从而在比对过程中引入大量非结构化噪声。我们将这些非预期噪声证据分为两个维度：(1) 无效证据导致的噪声，如图4(a)所示，这些无效证据一般形式为搜索引擎搜集到的无意义文本、网页广告插图像或格式错误数据；(2) 无关证据产生的噪声，即与查询主题关联较弱的、既无法支持也无法反驳的内容，例如图4(b)所示，证据图像虽与文本“竞选”相关，但是其无法证明待查询图文的真实性。这些证据中的噪声是使用自动化证据检索方法不可避免的，因此，我们设计从有效性和相关性两个维度对证据进行感知，从而降低噪声对比对验证过程的影响。具体而言，我们提出了一种双通道注意力感知框架，用于对多模态网络证据进行质量感知建模。该框架引入了两种互补的注意力机制：有效性注意力与相关性注意力，分别对应“有效性感知”和“相关性感知”两个信息通路，以模拟人类在事实判断中对信息的选择性关注机制。如图4所示，有效性注意力机制聚焦于证据内部的交互关系，通过自注意力机制对所有候选证据信息之间的上下文联系进行建模。通过计算证据片段之间的相似度关系，模型能够识别出内部逻辑一致、信息支持强的证据子集，从而提升对证据“是否有效”的感知能力。而相关性注意力机制的核心思想则是从待查询信息出发，计算其与所有候选证据的匹配程度。相关性注意力通过在待查询图/文信息和证据图/文间进行交叉注意力计算，引导模型优先关注与当前查询最相关的证据，从而实现证据筛选。这一机制可看作是对“是否相关”问题的建模。通过将这两种注意力机制协同整合，模型不仅能够从“外部对齐性”角度筛选出与查询最匹配的证据，也能从“内部一致性”角度提升对高质量证据的关注强度，为后续的多模态比对与判定提供了更加可靠的输入支持。

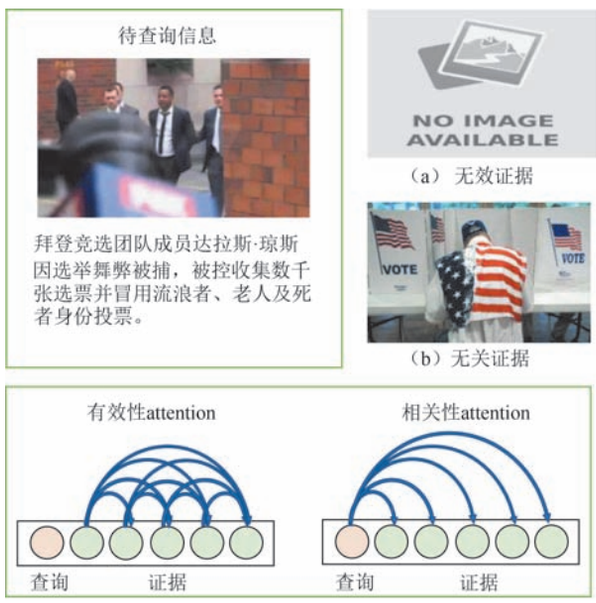


图4 无关和无效证据的示例以及本文设计的有效性attention和相关性attention示意图

3.3 全局视角验证

对于文本证据全局视角验证，我们使用预训练CLIP的文本编码器提取查询文本T_q和文本证据T_e的特征，分别得到全局文本语义特征$G_q^t \in \mathbb{R}^{d_t}$和$G_e^t = \{G_{e_1}^t, G_{e_2}^t, \cdots, G_{e_{n_t}}^t\}$(其中$G_{e_i}^t \in \mathbb{R}^{d_t}$)，这里$n_t$表示文本证据数量，$d_t$为文本特征维度。对于视觉证据比较，我们采用预训练CLIP的视觉编码器提取查询图像V^q和视觉证据V^e的全局视觉语义特征，表示为$G_q^v \in \mathbb{R}^{d_v}$和$G_e^v = \{G_{e_1}^v, G_{e_2}^v, \cdots, G_{e_{n_v}}^v\}$(其中$G_{e_i}^V \in \mathbb{R}^{d_v}$)，这里$n_v$表示视觉证据数量，$d_v$为视觉特征维度。

为实现第3.2节所提出的双通道注意力机制，我们摒弃了常规多模态融合方法中“先自注意力、后交叉注意力”的串行建模范式，创新性地设计了一种基于令牌类型区分的统一注意力机制。具体而言，我们首先在全局语义特征表示G_q^*与G_e^*中引入令牌类型嵌入(token-type embedding)，以显式标识每个标记的来源(查询或证据)。随后，我们将文本查询与候选证据拼接为统一的序列输入，以便在统一注意力图中同时建模跨源注意力与内源注意力：

$$\tilde{G}_q^* = G_q^* + token\_type\_embedding(0),$$
$$\tilde{G}_{e_i}^* = G_{e_i}^* + token\_type\_embedding(1),$$
$$\tilde{G}^* = [\tilde{G}_q^*, \tilde{G}_{e_1}^*, \tilde{G}_{e_2}^*, \cdots, \tilde{G}_{e_{n_e}}^*],$$

其中，$token_type_embedding(\cdot)$用于初始化可学习嵌入，符号$*$代表$t$（文本）或$v$（视觉）。基于令牌类型区分的统一注意力机制通过单次前向传播同步完成模态内与模态间信息交互，不仅避免了串行架构中多个注意力模块的依次计算，有效降低了计算复杂度与延迟，而且摆脱了固定处理顺序对信息流动的限制。统一注意力允许模型动态权衡来自查询和证据的信息，从而在提升交互效率的同时，实现了更灵活、更充分的深度融合。

随后，我们对序列$\tilde{G}^*$进行注意力计算，得到注意力输出$\tilde{O}_*^G$，其定义为：

$$\tilde{O}_*^G = Softmax\left(\frac{(\tilde{G}^* W^{Q_G})(\tilde{G}^* W^{K_G})}{\sqrt{d_k}}\right)(\tilde{G}^* W^{V_G}),$$

其中，投影矩阵$W^{Q_G}, W^{K_G}, W^{V_G} \in \mathbb{R}^{d_* \times d_k}$，$\tilde{O}_*^G \in \mathbb{R}^{(n_*+1)\times d_k}$，$d_k$为注意力模型的维度。

通过本文设计的基于令牌类型区分的统一注意力机制计算，我们既能精确捕捉证据间的有效性注意力，又能建立查询与证据间的相关性注意力，同时保留查询中与证据相关的信息而过滤泛化信息。

最终，我们从注意力输出$\tilde{O}_*^G$中提取对应查询的首个标记的隐藏状态，经包含线性层和$tanh$激活函数的前馈神经网络处理后，得到全局视角证据比较表示$O_*^G = FFN_1(O_*^G[0])$。至此，我们分别获得全局视角的文本证据比较表示O_t^G和视觉证据比较表示O_v^G。

除依赖证据比较进行验证外，本方法还增设了基于查询图像-文本自身多模态内容分析的多全局比对分支。该分支作为辅助判断依据，特别适用于证据缺失或质量低下导致比较方法失效的场景。我们使用CLIP提取的查询图像表示$C^v \in \mathbb{R}^{d_v}$和查询文本表示$C^t \in \mathbb{R}^{d_t}$（d_v和d_t为CLIP输出特征维度）。为了建模跨模态图文一致性和图文内源信息，我们提出了一种多元交互特征融合方法，通过四种特征运算机制实现跨模态图文一致性建模。具体而言，采用特征加法运算捕获图文特征向量的整体一致性，通过减法运算突显模态间的差异性信息，借助哈达玛积提取双模态共同激活的特征维度，最后利用点积运算度量两模态特征的跨模态一致性。最终，将这些融合后的交互特征与原始图文特征向量进行拼接，使模型能够同时整合图文内源特征信息以及模态间的关联、对比和耦合信息，从而显著增强模型的表示能力。基于这一多元交互特征融合框架，我们最终得到具有丰富语义表征的内容比较特征O_c^G：

$$O_c^G = [C^v + C^t : C^v - C^t : C^v * C^t : C^v \cdot C^t : C^v : C^t],$$

其中$(*)$表示哈达玛积运算，$(\cdot)$表示点积运算，$[:]$为拼接操作。该设计既保留了模态间的一致性信息，又维持了各模态的独立特征表达，可以作为在事实核查框架中基于内容分析判定的辅助分支。本文设计的融合机制具有双重优势：一方面能够有效保留跨模态间的一致性特征信息，另一方面又能完整维持各模态的独立表征特性。在事实核查框架中，该模块可以作为基于内容分析的虚假信息检测分支，为整体系统提供重要的辅助性判定支持。

3.4　局部视角验证

在局部视角验证中，我们聚焦于查询与证据中事件相关的具体细节（如人物、地点、时间等）的匹配性验证。针对局部视觉证据比较，我们采用预训练CLIP中的ViT模型提取图像块级局部特征。给定查询图像V^q和证据图像集V^e中的每个V_i^e，ViT将其转换为块嵌入序列$L^q = \{L_1^q, L_2^q, \cdots, L_{n_p}^q\}$和$L^{e_i} = \{L_1^{e_i}, L_2^{e_i}, \cdots, L_{n_p}^{e_i}\}$，其中$L_j^q \in \mathbb{R}^{d_p}$表示查询图像第$j$个块嵌入，$L_j^{e_i} \in \mathbb{R}^{d_p}$表示第$i$个证据图像第$j$个块嵌入，$d_p$为块嵌入维度，$n_p$为ViT划分的块数量。

由于查询与每条证据均表示为块级特征序列，若直接拼接所有块将导致序列过长，难以有效计算块级注意力。为此，我们采用查询引导的分阶段注意力机制：先进行查询与单条证据间的图像块级别相关性注意力计算，再执行证据级的有效性注意力聚合。

类比全局视角处理，我们为查询块和证据块添加不同的标记类型嵌入：

$$\tilde{L}_j^q = L_j^q + token\_type\_embedding(0),$$
$$\tilde{L}_j^{e_i} = L_j^{e_i} + token\_type\_embedding(1),$$
$$\tilde{L}_i = [\tilde{L}_1^q, \tilde{L}_2^q, \cdots, \tilde{L}_{n_p}^q, \tilde{L}_1^{e_i}, \tilde{L}_2^{e_i}, \cdots, \tilde{L}_{n_p}^{e_i}]。$$

其中，i为证据索引，j为块索引。该设计通过类型嵌入保持查询与证据的块身份标识，为后续分阶段注意力计算建立结构化输入。

接着，我们通过块级相关性注意力计算来聚焦与查询相关的图像块：

$$\tilde{H}_i = \mathrm{Softmax}\left(\frac{(\tilde{L}_i W^{Q_L^1})(\tilde{L}_i W^{K_L^1})}{\sqrt{d_k}}\right)(\tilde{L}_i W^{V_L^1}),$$

其中，投影矩阵 $W^{Q_L^1}, W^{K_L^1}, W^{V_L^1} \in \mathbb{R}^{d_p \times d_k}$，$\tilde{H}_i \in \mathbb{R}^{(2\times n_p)\times d_k}$。由于查询图像不仅对应注意力输出序列的首个标记，还包含所有查询图像块的对应标记，我们对这些标记的隐藏状态执行最大池化操作，得到查询关于第 i 条证据的局部视角增强特征：

$$H_i = Max\_Pooling(\tilde{H}_i[:n_p])。$$

为优化局部比较并实现证据级有效性关注，我们将所有 H_i 拼接为 $H=[H_1:H_2:\cdots:H_{n_e}]$ 后执行证据级有效性注意力计算：

$$\tilde{O}_v^L = \mathrm{Softmax}\left(\frac{(HW^{Q_L^2})(HW^{K_L^2})}{\sqrt{d_k}}\right)(HW^{V_L^2}),$$

此处投影矩阵 $W^{Q_L^2}, W^{K_L^2}, W^{V_L^2} \in \mathbb{R}^{d_k \times d_k}$。最终对 $\tilde{O}_v^L$ 进行均值池化并通过含 *tanh* 激活函数的前馈网络，得到局部视觉比较特征：

$$O_v^L = FFN_2(Mean\_Pooling(\tilde{O}_v^L))。$$

该处理流程通过分层注意力机制（块级→证据级）逐步提炼局部特征，既保留了细粒度视觉匹配信息，又通过证据级聚合抑制了噪声干扰。最大池化操作确保捕获最显著的查询相关特征，而均值池化则平衡各证据的贡献度，最终输出的 O_v^L 兼具判别力和鲁棒性。

在文本局部视角比对方面，我们采用命名实体识别算法 SpacyNER 分别从查询文本和文本证据中提取事件相关实体。特别地，我们针对与虚假信息事件最相关的三类实体（地理实体 GPE、人物 PERSON 和日期 DATE）提取并生成二进制特征。为每类实体设计了一个包含 4 比特的二进制特征：首比特表示查询与证据提取的同类别实体存在重叠；次比特表示双方当前类别实体均为空；第三比特表示查询或证据中当前类别实体有一方为空；末比特表示双方当前类别实体非空但无重叠。观察发现，从首位到末位四位，实体相关性依次递减。同理，我们使用相同方法为剩余两类实体提取 8 比特二进制特征，最终拼接成 12 比特的二进制特征向量 $O_t^L \in \mathbb{R}^{12}$，该向量能从事件捏造最易发生的三个角度检测虚假信息。

3.5 自赋权多分类器

在获得全局视角和局部视角的比对特征后，我们将各特征输入独立的分类器。为评估多分类器协同的有效性，我们提出赋权机制：通过包含两个线性层、批归一化层[50]和 *Tanh* 激活函数的前馈网络，将不同特征映射为对应的权重系数，记为：

$$P_{weight} = FFN_3\left(\left[O_v^G:O_t^G:O_c^G:O_v^L:O_t^L\right]\right),$$

其中，$P_{weight} \in \mathbb{R}^5$，并通过 Sigmoid 函数归一化至 $[0,1]$ 区间。

以全局视角文本比对特征 O_t^G 为例，其分类器为包含两个线性层、批归一化层和 RELU 激活函数的前馈神经网络。在输入 O_t^G 前，我们使用 P_{weight} 对应位置的权重系数对特征进行加权，再经 Sigmoid 函数得到预测结果 $\hat{y}_c$：

$$\hat{y}_t^G = \mathrm{Sigmoid}(FFN_4(P_{weight}[0]\cdot O_t^G))。$$

同理可获得五组验证特征对应的分类结果 $Y=\{\hat{y}_v^G, \hat{y}_t^G, \hat{y}_c^G, \hat{y}_v^L, \hat{y}_t^L\}$。通过最小化真实标签 y 与各 $y_i \in Y$ 之间的二元交叉熵损失，构建损失集合 $\mathcal{L}=\{\mathcal{L}_v^G, \mathcal{L}_t^G, \mathcal{L}_c^G, \mathcal{L}_v^L, \mathcal{L}_t^L\}$。整体目标函数为

$$\mathcal{L}_{final} = \sum_i \alpha_i \mathcal{L}_i,$$

经验性设置超参数 $\alpha_i = 1$ 以促进模型更有效地学习权重矩阵。详细完整的算法流程如算法 1 所示。

算法 1. ACD 事实核查框架流程

输入：查询图像 V^q、文本 T^q，证据图像 V^e、文本 T^e

输出：预测标签 y_{pred}

全局感知验证：

1. 使用 CLIP 提取全局特征：$G_q^t = f_t(T^q)$，$G_e^t = \{f_t(T_i^e)\}$，$G_q^v = f_v(V^q)$，$G_e^v = \{f_v(V_i^e)\}$
2. 基于令牌类型嵌入的统一注意力机制，得到：$O_t^G = \mathrm{UnifiedAttn}(G_q^t, G_e^t)$，$O_v^G = \mathrm{UnifiedAttn}(G_q^v, G_e^v)$
3. 构建图文多元交互特征：$O_c^G = [C^v + C^t : C^v - C^t : C^v * C^t : C^v \cdot C^t : C^v : C^t]$

局部感知验证：

4. 提取视觉查询与证据块嵌入：$L^q = \{L_j^q\}$，$L^{e_i} = \{L_j^{e_i}\}$
5. 对每个证据执行块级相关性注意力：$H_i = \mathrm{UnifiedAttn}(L^q, L^{e_i})$
6. 聚合证据级注意力，得到局部视觉表示：$O_v^L = \mathrm{SelfAttn}(\{H_i\}_{i=1}^{n_e})$
7. 使用 NER 抽取 {PERSON，GPE，DATE} 三类实体，生成 12-bit 二进制特征：$O_t^L = \mathrm{BinaryEntityFeature}(T^q, T^e)$

多分类器融合：

8. 拼接五类特征并计算权重系数：$F = [O_v^G, O_t^G, O_c^G, O_v^L, O_t^L]$，$P_{weight} = \sigma(\mathrm{FFN}_3(F))$
9. 加权并输入独立分类器：$y_i = \sigma(\mathrm{FFN}_i(P_{weight}[i]\cdot O_i))$
10. 聚合分类结果：$y_{pred} = \mathrm{Average}(\{y_i\})$

4 实　验

4.1 实验设置

4.1.1 数据集

首先介绍本文在实验中用到的两个数据集以及其交叉检索证据的具体流程与方法。

合成式虚假信息数据集OoC[20]是基于NewsCLIPpings[44]数据集通过证据扩充构建而成的。该数据集的原始正确配对图文来源于VisualNews[51]新闻语料库，涵盖了《卫报》、BBC、《今日美国》和《华盛顿邮报》四大权威新闻机构的新闻内容。在构建合成式虚假信息时，研究者采用了两种伪造策略：(1) 使用数据集中语义最相似的图像替换原始图像；(2) 采用语义最相似的文本替换原始文本描述。为了实现对NewsCLIPpings样本的多模态事实核查，Abdelnabi等人引入了交叉检索的证据收集方法，对原始数据集进行了系统性的证据扩充。最终构建的数据集包含71 072个训练样本、7024个验证样本和7264个测试样本，每个样本都包含真实或伪造的图文配对，并附有相应的多模态证据支持。

真实世界虚假信息数据集MR2[22]是通过收录来自社交媒体的真实虚假信息而得到的。其包含英文(MR2-E)和中文(MR2-C)两个子集。该数据集通过权威渠道(Google Fact Check Tools API和微博辟谣中心)收集了2017-2022年间的来自Twitter和微博的社交媒体内容。Twitter数据集包含1418条谣言、2318条非谣言、3240条未验证帖子；中文微博数据集包含1754条谣言、2609条非谣言、3361条未验证帖子。在我们的实验中，为了将两个数据集输出维度统一，我们只保留MR2数据集中标签谣言与非谣言的数据，将其统一为一个二分类任务。

OoC与MR2采用了相同的图文交叉证据检索框架。在文本证据采集方面，系统通过Google Vision API [52]对查询图像进行反向搜索，该API会返回与图像相关联的实体列表，同时还会返回图像URL及其所在页面的URL。此外，为了扩充文本证据，额外收集了图像标题(captions)；通过专门设计的网络爬虫访问页面后，利用图像URL或基于感知哈希(perceptual hashing)的图像内容匹配技术定位图像标签，系统性地提取〈figcaption〉标签及〈img〉标签的文本属性(包括alt、image-alt、caption、data-caption和title)。对于每个页面，收集所有非冗余文本片段(API最多返回20条搜索结果)。在视觉证据采集方面，系统以待核查描述作为查询词，通过Google自定义搜索API[53]检索相关图像(最多10条结果)，并记录其来源域名。值得注意的是，与反向图像搜索不同，正向图像搜索返回的结果可能与查询文本的语义匹配度较低，这可能导致视觉证据与查询图像的关联性较弱，从而为事实核查任务引入额外的噪声。

4.1.2 实现细节

实验中采用的预训练CLIP模型为“CLIP VIT L/14”，文本嵌入特征维度d_t设为768。本文将输入图像尺寸设置为224 × 224，ViT中的图像块数量为49，视觉嵌入特征维度d_t为768。训练阶段冻结所有预训练骨干网络的参数，批处理大小设为64。使用Adam优化器训练30个周期，采用最大值为6e-5的循环学习率策略[54]。为防止过拟合，对预训练骨干网络提取的特征设置0.05的dropout[55]率，对各个分支提取的判定特征设置0.25的dropout率。

4.2 对比实验

4.2.1 定量分析

我们将所提方法与基于内容的多模态虚假信息检测方法MVAE[17]和基于证据的多模态事实核查领域的代表性先进方法Logically[42]、INO[43]、TripleCheck[21]、CCN[20]和ESCNet[23]进行对比。如表2所示，ACD模型在MR2数据集、OoC数据集上分别取得89.4%和89.6%的平均准确率，达到最优性能，验证了其在信息真实性鉴别方面的可靠性。值得注意的是，本方法在精确率、召回率和F1分数上均超越现有方案，表明其在不同评估维度上具备更鲁棒均衡的检测能力。对比方法中，MVAE作为典型的多模态虚假信息检测方法，仅依赖跨模态特征进行真实性判断而未引入外部证据。本文发现，虽然MVAE在真实虚假数据集MR2上表现尚可(79.0%准确率)，但在更具挑战性的上下文脱节(OoC)数据集上性能显著下降(64.7%准确率)。这表明当面对图像-文本内容表面合理但语义错位的复杂篡改时，仅依赖多模态内在一致性的方法存在局限。相比之下，事实核查类方法(Logically、INO、TripleCheck、CCN和ESCNet)通过联合网络共现证据进行验证，获得了更稳定的性能。例如ESCNet在MR2(87.9%)和OoC(87.7%)数据集上的优异表现，印证了外部验证的重要性。我们的ACD框架通过全局语义分析与局部细节验证相结合的动态证据可靠性评估机制，进一步推进了该范式——即使在OoC

这类复杂虚假信息场景下，仍以89.6%的准确率超越所有基线方法，展现出显著的性能优势。

表2 在MR2和OoC数据集上的实验对比结果

数据集	方法	准确率	精确率	召回率	F1值
MR2	MVAE	0.790	0.791	0.780	0.783
	Logically	0.712	0.714	0.712	0.713
	INO	0.801	0.796	0.802	0.800
	TripleCheck	0.816	0.817	0.812	0.814
	CCN	0.856	0.856	0.850	0.853
	ESCNet	0.879	0.881	0.877	0.879
	ACD(本文)	**0.894**	**0.891**	**0.893**	**0.892**
OoC	MVAE	0.647	0.644	0.639	0.641
	Logically	0.786	0.791	0.786	0.788
	INO	0.823	0.834	0.823	0.828
	TripleCheck	0.848	0.850	0.851	0.851
	CCN	0.847	0.853	0.852	0.852
	ESCNet	0.877	0.879	0.872	0.875
	ACD(本文)	**0.896**	**0.895**	**0.903**	**0.899**

此外，为进一步评估模型的特征判别能力，我们采用t-SNE可视化技术对ACD方法与代表性基线方法生成的特征分布进行了分析。如图5所示，基于内容分析的MVAE方法在面对脱离上下文数据时，不同类别样本的特征存在显著重叠，反映出其在处理复杂虚假信息时的局限性。INO与CCN方法通过引入外部证据，在一定程度上改善了特征的分离度。相比之下，本文所提出的ACD方法通过模拟人类认知的推理过程，结合双通道注意力机制与双视角验证策略，学习到了判别性更强的特征表示，在嵌入空间中呈现出更清晰的类别边界与更紧凑的簇内结构，直观印证了本方法在复杂场景下判别能力。

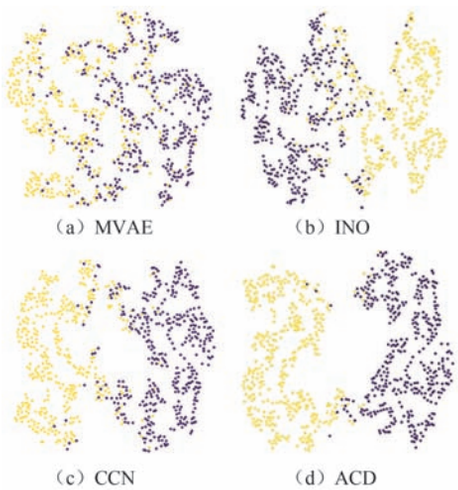


图5 OoC数据集下各方法特征的t-SNE可视化

4.2.2 定性分析

我们在图6中可视化展示了MVAE[17]、CCN[20]和本研究的ACD模型对某条虚假信息的检测结果。该案例呈现为图文组合形式的虚假信息，并附有多条网络证据。由于查询图文本身不存在明显篡改痕迹或内容矛盾，基于内容分析的MVAE方法以高置信度将其误判为真实信息，这揭示了传统内容分析方法在检测拼接型虚假信息时的固有局限。事实核查基线方法CCN虽能利用证据进行检测，并关注查询文本与文本证据的差异，但受限于单一分类器架构，当查询图像与视觉证据高度相似时，其难以有效处理这种混淆情境，最终输出接近阈值0.5的错误预测。我们提出的ACD框架不仅从有效性和相关性维度评估证据，通过全局与局部双重视角进行对比验证，还通过自赋权多分类器机制对不同验证线索进行权重分配，实现多验证分支的解耦分析。

图6 MVAE、CCN与ACD模型在事实核查数据集OoC上的虚假信息检测结果示例

ACD相较于现有方法的优势可归纳为三个方面：1）双视角验证机制。通过全局语义特征与局部细节特征的协同分析，实现更全面的事实核查。例如在文本比对时提取事件相关实体，有效捕捉虚假信息中常被操纵的细微语义差异；2）证据质量感知。针对网络证据质量参差不齐引入的噪声问题，在证据比对过程中从有效性和相关性维度进行筛选，提升验证可靠性；3）自赋权决策系统：在决策阶段集成多分类器与自赋权机制，自动评估不同验证线索的重要性并解耦多验证分支，有效解决多判定分支模糊情境下的决策难题，精准识别关键判别特征，提高决策系统的可解释性。

4.3　消融实验

为探究ACD模型中关键组件的影响，我们通过不同组件组合下的模型性能进行评估。每组实验中，我们选择性移除特定组件后重新训练模型，具体设置如下：

ACD w/o L：移除局部视角证据比对分支。

ACD w/o G：移除全局视角证据比对分支。

ACD w/o Q：移除基于内容的全局多模态查询比对分支。

ACD w/o T：移除全局与局部视角的文本查询-证据比对模块。

ACD w/o V：移除全局与局部视角的视觉查询-证据比对模块。

ACD w/o M：移除最终决策阶段自赋权多分类器模块，将所有比对特征拼接后输入单分类器。

ACD w/o TE：移除令牌类型嵌入，将查询与证据特征直接拼接进行后续的注意力计算。

4.3.1　验证分支消融分析

如表3所示，我们发现：1）完整ACD模型性能优于ACD w/o L和ACD w/o G变体，证明双视角验证机制能捕捉更隐蔽的虚假线索，实现更全面准确的事实核查。值得注意的是，ACD w/o G的性能衰减比ACD w/o L更显著，表明全局视角验证在系统中起主导作用；2）ACD相对ACD w/o Q的性能优势，验证了全局跨模态查询比对分支在证据缺失等特殊场景下的有效性；3）ACD w/o V和ACD w/o T的性能下降证实视觉与文本模态的不可替代性，其中文本比对模块的移除（ACD w/o T）导致更大幅度的性能衰减，这与虚假信息检测领域“文本特征更具判别性”的传统认知一致。4）ACD w/oTE性能下降约6%，这证明了令牌类别编码是实现有效统一注意力计算的前提条件。若缺乏令牌类型标识，模型在融合查询与证据的拼接序列时就难以有效区分不同来源和信息角色的令牌，从而导致语义混淆与交互效率降低。因此，令牌类别编码通过为模型提供结构化的来源先验，在维持统一计算框架的同时，保障了对多源信息的精准辨识与融合，是提升模型判别能力的关键设计。

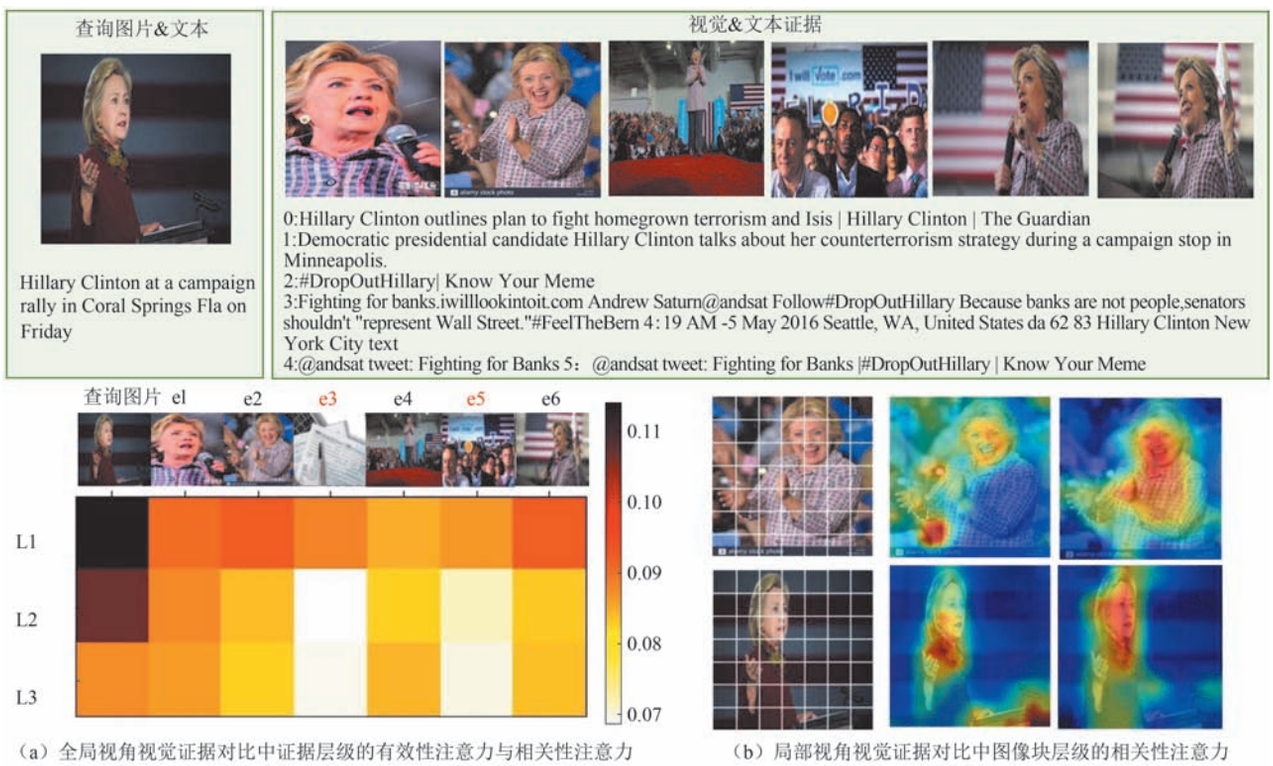
表3　OoC数据集上的消融实验对比结果

方法	全局视角			局部视角		分类器		准确率
	Evidence(t)	Evidence(v)	Query	Evidence(t)	Evidence(v)	single	multiple	
ACD	√	√	√	√	√		√	0.8963
ACD w/o L	√	√	√				√	0.8736
ACD w/o G			√	√	√		√	0.8518
ACD w/o Q	√	√		√	√		√	0.8498
ACD w/o T		√	√		√		√	0.7945
ACD w/o V	√		√	√			√	0.8352
ACD w/o M	√	√	√	√	√	√		0.8616
ACD w/o TE	√	√	√	√	√		√	0.8359
CCN+AC	√	√	√			√		0.8644

4.3.2　有效性关注与相关性关注机制评估

为验证我们提出的双通道注意力机制，将CCN原有记忆网络[56]替换为我们的关注和比对网络模块，构建CCN（+AC）对照模型。在保持其他设置相同的情况下，如表2所示，ACD仍展现出性能优势。这证明我们设计的有效性-相关性双重注意力机制优于CCN仅计算查询-证据相关性的方法，传统记忆网络忽视证据间内在关联的缺陷被我们的交互式比对机制有效克服。此外，我们通过可视化呈现了有效性关注与相关性关注机制的作用效果。以图像证据比对为例，图7（a）展示了全局视角下的验证，我们可视化展示了查询-证据序列上第1、3、5层注意力矩阵的变化趋势，发现初始层注意力分布近乎随机，但随着网络层数加深，系统

显著降低对无效/无关证据的关注度。图7(b)展示了局部视角下的验证，前两行分别对应查询图像与证据图像，每行最左侧为图像分块可视化，中间为ViT原始特征生成的块级注意力，最右侧为经我们相关性关注机制处理后的注意力分布。可见经过网络优化后，面部、服饰等关键区域获得显著关注，为局部证据比对中的不一致性检测提供了更精准的线索定位。



(a) 全局视角视觉证据对比中证据层级的有效性注意力与相关性注意力 (b) 局部视角视觉证据对比中图像块层级的相关性注意力

图7 有效性注意力与相关性注意力可视化。

4.3.3 证据噪声的影响分析

在实际网络环境中，用于事实核查的证据常包含大量无效或无关信息，这对模型的证据筛选与质量感知能力提出了严峻挑战。为验证本文所提双通道证据感知机制在处理不可靠网络证据时的创新价值，我们设计了可控的噪声注入实验。由于噪声数据难以通过自动化方法进行精确量化，我们采用人工判定的方式，从测试集中构建了包含不同噪声比例的数据子集。具体而言，我们在原始测试集中分别人工采样是否包含证据噪声的样本子集，并按预设比例组合，以此模拟真实网络环境中证据质量参差不齐的复杂场景。

实验结果如图8所示。随着噪声比例的逐步升高，所有基线模型的性能均出现明显下降，这表明噪声证据对传统方法构成了显著干扰。相比之下，本文提出的ACD方法在相同噪声条件下的性能下降幅度显著更小，展现出优异的稳定性。这一结果从定量角度证明，本文基于有效性感知与相关性感知的双通道注意力机制，能够有效识别并抑制噪声证据的干扰，提升模型在不可靠网络证据环境下的鲁棒性。

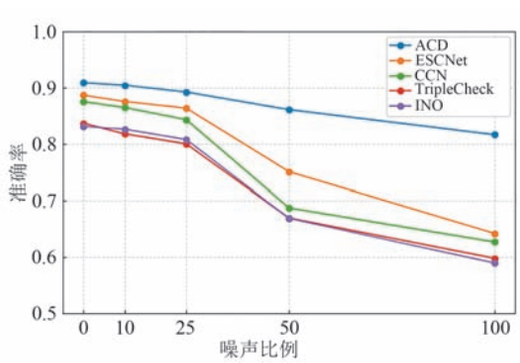


图8 各方法的证据噪声鲁棒性分析

4.3.4 自赋权多分类器的影响分析

为系统评估本文提出的自赋权多分类器机制的有效性，我们对完整ACD框架及其消融变体(ACD w/o M)进行了对比实验。表2的实验结果表明，两种配置之间存在显著的性能差异。采用新型自赋权多分类器的完整ACD模型，其性能明显优于仅简单

拼接特征并使用单一分类器的简化版本(ACD w/o M)。这一结果充分验证了我们提出的赋权集成策略的重要性，该策略通过动态评估各验证信号的相对重要性和可靠性，实现了多源验证信号的智能融合。我们的多分类器设计具有以下显著优势：(1)架构灵活性：通过将验证过程分解为多个独立的决策分支并采用自适应赋权机制，可以根据不同应用场景灵活调整决策架构，自由增删分支模块，而无需重新训练整个系统；(2)过程可解释性：模块化设计使得每个分类器分支对应特定的验证维度，这种透明性有助于追溯最终决策的主要影响因素，满足了实际应用中对于模型可解释性的需求(在某些场景下，可解释性与准确性同等重要)；(3)实验数据证实，相较于传统的单分类器或简单特征拼接方法，我们的自赋权融合机制能够有效解决多验证信号冲突导致的决策模糊问题，从而获得更加鲁棒和可靠的性能表现。

5 结 论

本文针对现有事实核查方法在证据质量感知、细粒度推理与决策融合方面的不足，创新性地提出了一个认知驱动型多模态事实核查框架。该框架的核心是模拟人类认知的“关注-比对-判定”(ACD)三级推理机制。在“关注”阶段，设计了双通道注意力机制，分别从有效性和相关性两个维度对证据进行质量评估，显著降低了噪声干扰；在“比较”阶段，采用全局与局部相结合的双视角验证机制，实现了从宏观语义到微观细节的多层次事实比对；在“判定”阶段，引入了自赋权多分类器集成策略，能够动态融合多路验证线索，有效提升了决策的可靠性与可解释性。实验结果表明，所提方法在OOC、MR2等多个基准数据集上均取得了超过1.5%的准确率提升，验证了其有效性与泛化能力。本工作为多模态事实核查提供了新的研究思路和技术路径，具有重要的理论价值和实践意义。

本文提出的认知驱动型多模态事实核查框架在未来具有广阔的拓展空间。尽管该框架在性能上表现优异，但其多层级认知结构仍会带来一定的计算开销，未来将探索轻量化模型设计与知识蒸馏等优化策略，在保持核心推理能力的同时提升计算效率。同时，可推动大语言模型与视觉基础模型与ACD推理机制的深度融合，以增强语义理解与零样本泛化能力。本框架内嵌的客观验证机制还展现出识别与缓解社会信息偏见的潜力，为构建更具纠偏能力的信息环境提供了技术路径。此外，面对深度伪造等新兴威胁，提升模型的对抗鲁棒性、发展更透明的可解释性工具，亦是推动该技术走向成熟应用的关键方向。

参 考 文 献

[1] Lazer D M J, Baum M A, Benkler Y, et al. The science of fake news. Science, 2018, 359(6380): 1094-1096

[2] Dack S. Deep fakes, fake news, and what comes next. The Henry M. Jackson School of International Studies, 2019

[3] Botha J, Pieterse H. Fake news and deepfakes: A dangerous threat for 21st century information security//Proceedings of the ICCWS 2020 15th International Conference on Cyber Warfare and Security. Academic Conferences and publishing limited. Norfolk, USA, 2020: 57

[4] Caramancion K M. News verifiers showdown: a comparative performance evaluation of ChatGPT 3.5, ChatGPT 4.0, bing ai, and bard in news fact-checking//Proceedings of the 2023 IEEE Future Networks World Forum (FNWF). Baltimore, USA, 2023: 1-6

[5] Zhuo T Y, Huang Y, Chen C, et al. Red teaming ChatGPT via jailbreaking: Bias, robustness, reliability and toxicity. arXiv preprint arXiv:2301.12867, 2023

[6] Islam M R, Liu S, Wang X, et al. Deep learning for misinformation detection on online social networks: a survey and new perspectives. Social Network Analysis and Mining, 2020, 10(1): 82

[7] Pelrine K, Danovitch J, Rabbany R. The surprising performance of simple baselines for misinformation detection//Proceedings of the Web Cconference 2021. Ljubljana, Slovenia, 2021: 3432-3441

[8] Wu L, Morstatter F, Carley K M, et al. Misinformation in social media: definition, manipulation, and detection. ACM SIGKDD explorations newsletter, 2019, 21(2): 80-90

[9] ZHANG Zhi-Yong, JING Jun-Chang, LI Fei, et al. Survey on fake information detection, propagation and control in online social networks from the perspective of artificial intelligence. Chinese Journal of Computers, 2021, 44(11): 2261-2282 (in Chinese)
(张志勇，荆军昌，李斐等. 人工智能视角下的在线社交网络虚假信息检测、传播与控制研究综述. 计算机学报，2021，44(11): 2261-2282)

[10] Cao J, Qi P, Sheng Q, et al. Exploring the role of visual content in fake news detection//Disinformation, Misinformation, and Fake News in Social Media: Emerging Research Challenges and Opportunities, Cham, Switzerland, 2020: 141-161

[11] Conroy N K, Rubin V L, Chen Y. Automatic deception detection: Methods for finding fake news Proceedings of the Association for Information Science and Technology, 2015, 52(1): 1-4

[12] Ma J, GaoW, Mitra P, et al. Detecting rumors from microblogs

with recurrent neural networks//Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI). New York City, USA, 2016: 3818-3824

[13] Pérez-Rosas V, Kleinberg B, Lefevre A, et al. Automatic detection of fake news//Proceedings of the 27th International Conference on Computational Linguistics. Santa Fe, USA, 2018: 3391-3401

[14] Qazvinian V, Rosengren E, Radev D, et al. Rumor has it: Identifying misinformation in microblogs//Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. Edinburgh, UK, 2011: 1589-1599

[15] Alam F, Cresci S, Chakraborty T, et al. A survey on multimodal disinformation detection//Proceedings of the 29th International Conference on Computational Linguistics. Gyeongju, Republic of Korea, 2022: 6625-6643

[16] Jin Z, Cao J, Guo H, et al. Multimodal fusion with recurrent neural networks for rumor detection on microblogs// Proceedings of the 25th ACM International Conference on Multimedia. Mountain View, USA, 2017: 795-816

[17] Khattar D, Goud J S, Gupta M, et al. Mvae: Multimodal variational autoencoder for fake news detection// Proceedings of the World Wide Web Conference. San Francisco, USA, 2019: 2915-2921

[18] Qi P, Cao J, Li X, et al. Improving fake news detection by using an entity-enhanced framework to fuse diverse multimodal clues//Proceedings of the 29th ACM International Conference on Multimedia. Chengdu, China, 2021: 1212-1220

[19] Wang Y, Ma F, Jin Z, et al. Eann: Event adversarial neural networks for multi-modal fake news detection//Proceedings of the 24th ACM Sigkdd International Conference on Knowledge Discovery & Data Mining. London, UK, 2018: 849-857

[20] Abdelnabi S, Hasan R, Fritz M. Open-domain, content-based, multi-modal fact-checking of out-of-context images via online resources//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 14940-14949

[21] Du W W, Wu H W, Wang W Y, et al. Team triple-check at factify 2: Parameter-efficient large foundation models with feature representations for multi-modal fact verification. arXiv preprint arXiv:2302.07740, 2023

[22] Hu X, Guo Z, Chen J, et al. Mr2: A benchmark for multimodal retrieval-augmented rumor detection in social media// Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. Taipei, China, 2023: 2901-2912

[23] Zhang F, Liu J, Xie J, et al. Escnet: Entity-enhanced and stance checking network for multi-modal fact-checking// Proceedings of the ACM Web Conference 2024. Singapore, 2024: 2429-2440

[24] LIU Ya-Hui, JIN Xiao-Long, SHEN Hua-Wei, et al. A survey on rumor identification over social media. Chinese Journal of Computers, 2018, 41(07): 1536-1558 (in Chinese)
(刘雅辉,靳小龙,沈华伟,等.社交媒体中的谣言识别研究综述.计算机学报,2018,41(07):1536-1558)

[25] Alkhodair S A, Ding S H H, Fung B C M, et al. Detecting breaking news rumors of emerging topics in social media. Information Processing & Management, 2020, 57(2): 102018

[26] Yang Yan-Jie, Wang Li, Wang Yu-Hang. Rumor detection based on source information and gating graph neural network. Journal of Computer Research and Development, 2021, 58(7): 1412-1424 (in Chinese)
(杨延杰,王莉,王宇航.融合源信息和门控图神经网络的谣言检测研究.计算机研究与发展,2021,58(7):1412-1424)

[27] Su Xing, Yu Ke, Wu Xiao-Fei. Gated interactive fusion network for rumor detection. Journal of Beijing University of Posts and Telecommunications, 2023, 46(4):97-102 (in Chinese)
(苏兴,禹可,吴晓非.基于层次门控交互融合网络的谣言检测方法.北京邮电大学学报,2023,46(4):97-102)

[28] Yang Y, Zheng L, Zhang J, et al. TI-CNN: Convolutional neural networks for fake news detection. arXiv preprint arXiv: 1806.00749, 2018

[29] Wang You-Wei, Feng Li-Zhou, Wang Wei-Qi, et al. Weibo rumor detection based on heterogeneous graph of event-word-feature. Journal of Chinese Information Processing, 2023, 37(9): 161-174 (in Chinese)
(王友卫,凤丽洲,王炜琦,等.基于事件-词语-特征异质图的微博谣言检测新方法.中文信息学报,2023,37(9):161-174)

[30] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need//Advances in neural information processing systems(2017. LongBeach, USA, 2017, 30: 5998-6008

[31] Qian S, Wang J, Hu J, et al. Hierarchical multi-modal contextual attention network for fake news detection// Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. Montreal, Canada, 2021: 153-162

[32] Zhou Y, Yang Y, Ying Q, et al. Multi-modal fake news detection on social media via multi-grained information fusion// Proceedings of the 2023 ACM International Conference on Multimedia Retrieval. Thessaloniki, Greece, 2023: 343-352

[33] Ma Z, Luo M, Guo H, et al. Event-radar: Event-driven multi-view learning for multimodal fake news detection//Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Bangkok, Thailand, 2024: 5809-5821

[34] Yin S, Zhu P, Wu L, et al. Gamc: an unsupervised method for fake news detection using graph autoencoder with masking// Proceedings of the AAAI conference on artificial intelligence. Vancouver, Canada, 2024, 38(1): 347-355

[35] Wu L, Wang L, Zhao Y. Unified evidence enhancement inference framework for fake news detection//Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24. Jeju, Republic of Korea, 2024: 6541-6549

[36] Yu X, Sheng Z, Lu W, et al. Racmc: Residual-aware compensation network with multi-granularity constraints for fake news detection//Proceedings of the AAAI Conference on Artificial Intelligence. Philadelphia, USA, 2025, 39(1): 986-994

[37] Popat K, Mukherjee S, Yates A, et al. Declare: Debunking fake news and false claims using evidence-aware deep learning. arXiv preprint arXiv:1809.06416, 2018

[38] Augenstein I, Lioma C, Wang D, et al. MultiFC: A real-world multi-domain dataset for evidence-based fact checking of claims. arXiv preprint arXiv:1909.03242, 2019

[39] Yao B M, Shah A, Sun L, et al. End-to-end multimodal fact-checking and explanation generation: A challenging dataset and models//Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. Taipei, China, 2023: 2733-2743

[40] Zlatkova D, Nakov P, Koychev I. Fact-checking meets fauxtography: Verifying claims about images. arXiv preprint arXiv:1908.11722, 2019

[41] Suryavardan S, Mishra S, Patwa P, et al. Factify 2: A multimodal fake news and satire news dataset. arXiv preprint arXiv:2304.03897, 2023

[42] Gao J, Hoffmann H F, Oikonomou S, et al. Logically at the factify 2022: Multimodal fact verification. arXiv preprint arXiv: 2112.09253, 2021

[43] Zhang Y, Tao Z, Wang X, et al. Ino at factify 2: Structure coherence based multi-modal fact verification. arXiv preprint arXiv:2303.01510, 2023

[44] Luo G, Darrell T, Rohrbach A. Newsclippings: Automatic generation of out-of-context multimodal media. arXiv preprint arXiv:2104.05893, 2021

[45] Jaiswal A, Sabir E, AbdAlmageed W, et al. Multimedia semantic integrity assessment using joint embedding of images and text//Proceedings of the 25th ACM International Conference on Multimedia. Mountain View, USA, 2017: 1465-1471

[46] Sabir E, AbdAlmageed W, Wu Y, et al. Deep multimodal image-repurposing detection//Proceedings of the 26th ACM International Conference on Multimedia. Seoul, Republic of Korea, 2018: 1337-1345

[47] Zhang F, Liu J, Zhang Q, et al. Ecenet: Explainable and context-enhanced network for muti-modal fact verification// Proceedings of the 31st ACM International Conference on Multimedia. Ottawa, Canada, 2023: 1231-1240

[48] Qi P, Yan Z, Hsu W, et al. Sniffer: Multimodal large language model for explainable out-of-context misinformation detection// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2024: 13052-13062

[49] Lee J, Lu X, Hessel J, et al. How to train your fact verifier: knowledge transfer with multimodal open models. arXiv preprint arXiv:2407.00369, 2024

[50] Ioffe S. Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv: 1502.03167, 2015

[51] Liu F, Wang Y, Wang T, et al. Visual news: Benchmark and challenges in news image captioning//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Punta Cana, Dominican Republic, 2021: 6761-6771

[52] Google Vision API. Detect web entities and pages. Available at: https://cloud.google.com/vision/docs/detecting-web

[53] DevelopersGoogle. Programmable search engine. Available at: https://developers.google.com/customsearch/v1/overview

[54] Smith L N. Cyclical learning rates for training neural networks// Proceedings of the 2017 IEEE Winter Conference on Applications of Computer Vision (WACV). Santa Rosa, USA, 2017: 464-472

[55] Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: a simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research, 2014, 15(1): 1929-1958

[56] Kumar A, Irsoy O, Ondruska P, et al. Ask me anything: Dynamic memory networks for natural language processing// Proceedings of the International Conference on Machine Learning. New York City, USA, 2016: 1378-1387

WANG Zhi-Shen, Ph. D. candidate. His main research interests include computer vision and multi-modal learning.

WANG Rui, Ph. D., professor. Her interests include deep learning and computer vision.

JING Li-Hua, Ph. D. assistant researcher. Her main research interests include deep learning and computer vision.

LIU Li-Jun, Ph. D. Her main research interests include computer vision and knowledge distillation.

LV Fei-Xiao, Ph. D. His main research interests include computer vision and image caption.

Background

In today's digital world, misinformation spreads rapidly on social media, undermining public discourse, politics, and global health. While early detection methods focused on analyzing single modalities like text or images, modern misinformation increasingly combines multiple media formats in sophisticated ways that evade traditional detection approaches. Recent advances in multimodal fact-checking have made progress by examining cross-modal consistency and leveraging external evidence, yet significant limitations remain in handling noisy web-sourced data and performing human-like reasoning. Current

systems struggle with three fundamental issues: they lack effective mechanisms to filter irrelevant or low-quality evidence, fail to perform hierarchical verification that considers both broad context and fine details, and provide insufficient transparency in their decision-making processes. These shortcomings become particularly apparent when dealing with advanced misinformation tactics that skillfully combine authentic elements from different sources to create deceptive narratives. The absence of cognitive-inspired verification frameworks in existing methods leaves them vulnerable to manipulation by increasingly sophisticated fake content.

Our work addresses these challenges by developing a novel cognitive-driven framework that mimics human fact-checking processes through an "Attend-Compare-Determine" (ACD) mechanism. The system first intelligently filters evidence through dual attention channels that assess both relevance and validity, then performs comprehensive verification at multiple levels of granularity, and finally integrates these signals through an adaptive weighting mechanism that enhances both accuracy and explainability. Experimental results demonstrate significant improvements over state-of-the-art methods, particularly in handling noisy data and complex multimodal deception scenarios. By bridging cognitive science principles with advanced machine learning techniques, this research provides a more robust and interpretable approach to combating modern misinformation while advancing the theoretical foundations of automated fact-checking systems. The framework's ability to combine evidence quality assessment with hierarchical verification and transparent decision-making represents an important step forward in developing reliable tools for maintaining information integrity in the digital age.

This work was supported in part by the National Natural Science Foundation of China Under Grants No. 62176253. The opinions, findings and conclusions expressed in this paper are those of authors and do not necessarily reflect the views of the funding agencies or the government.