中国科学 : 信息科学 2026 年 第 56 卷 第 1 期: 1–22
SCIENTIA SINICA Informationis
评述

大模型时代的深度伪造检测

彭春蕾[1,2], 李俊烨[1,2], 刘德成[1,2], 王楠楠[1*], 胡瑞敏[3], 高新波[1]

1. 西安电子科技大学空天地一体化综合业务网全国重点实验室, 西安 710071
2. 西安电子科技大学网络与信息安全学院, 西安 710071
3. 西安电子科技大学杭州研究院, 杭州 311231
* 通信作者. E-mail: nnwang@xidian.edu.cn

收稿日期: 2025–06–19; 修回日期: 2025–09–08; 接受日期: 2025–10–21; 网络出版日期: 2026–01–04

国家自然科学基金 (批准号: 62276198, U22A2035, U22A2096, 62571405) 资助项目

摘要 随着人工智能技术的不断演进, 深度伪造 (Deepfake) 正从单一模态合成发展为包含视、听、文多种媒介的复杂生成形式. 多模态大模型 (large multimodal models, LMMs) 的出现在显著增强伪造内容生成能力的同时, 也为伪造检测任务带来了前所未有的机遇与挑战. 本文归纳了在大模型背景下伪造检测技术的研究进展与技术演化, 回顾了近年来相关研究成果, 并总结了近期各类多模态伪造检测数据集. 在此基础上, 深入分析了多模态大模型在检测性能、幻觉现象、判断准确性与公平性等方面展现的潜力与挑战, 探讨了问题背后的原因并提出了未来的解决方案路径. 最后, 本文展望了大模型时代的伪造检测技术的发展趋势, 包括伪造信息的复杂化、传统检测方法的价值、判别可解释性和技术对抗性等方向.

关键词 多模态大模型, 深度伪造检测, 视觉 – 文本融合, 可解释性检测, 跨模态推理

1 引言

深度伪造 (Deepfake) 技术指的是, 利用机器学习或深度神经网络算法, 对视频、音频、图像及文本等内容进行高度逼真的合成与伪造. 深度伪造手段在大模型 (如大型语言模型与多模态模型) 的加持下日益精进, 并已在多个应用场景中落地, 如影视娱乐、文化宣传、云教育中替换影视作品中的劣迹艺人、修复老照片、AI 合成虚拟主播以及提供虚拟教师等. 但其泛滥使用也引发了大量社会问题, 网络诈骗、虚假新闻、隐私侵犯和政治操纵等现象频发, 成为全球范围内的严峻挑战.

Science 2024 年刊载的文章 [1] 指出, 33%~50% 的受访群体难以辨别深度伪造内容. 2024 年 7 月, 一张特朗普与泰勒 · 斯威夫特支持者的合影在网络上广泛传播, 虽然该图像后来被证明是通过 AI 技术伪造的, 但它仍对部分观众产生了误导. 同年 8 月, 特朗普与埃隆 · 马斯克的跳舞视频也被深度伪造, 试图制造两位公众人物关系亲密的假象. 2024 年 1~9 月, 中国检察机关办理的 159 件假新闻案件中, 深度伪造技术滥用占比达 37%. 在社交媒体、网络新闻等高频传播环境中, 深度伪造的表现形式已由早期单模态演变为复杂的多模态组合, 包括图像、音频与文本的协同生成与传播 [2~5]. 文献 [6] 指出, 具有图像的文章的平均转发次数比没有图像的文章高 11 倍,

引用格式: 彭春蕾, 李俊烨, 刘德成, 等. 大模型时代的深度伪造检测. 中国科学: 信息科学, 2026, 56: 1–22, doi: 10.1360/SSI-2025-0289
Peng C L, Li J Y, Liu D C, et al. Deepfake detection in the era of large models. Sci Sin Inform, 2026, 56: 1–22, doi: 10.1360/SSI-2025-0289

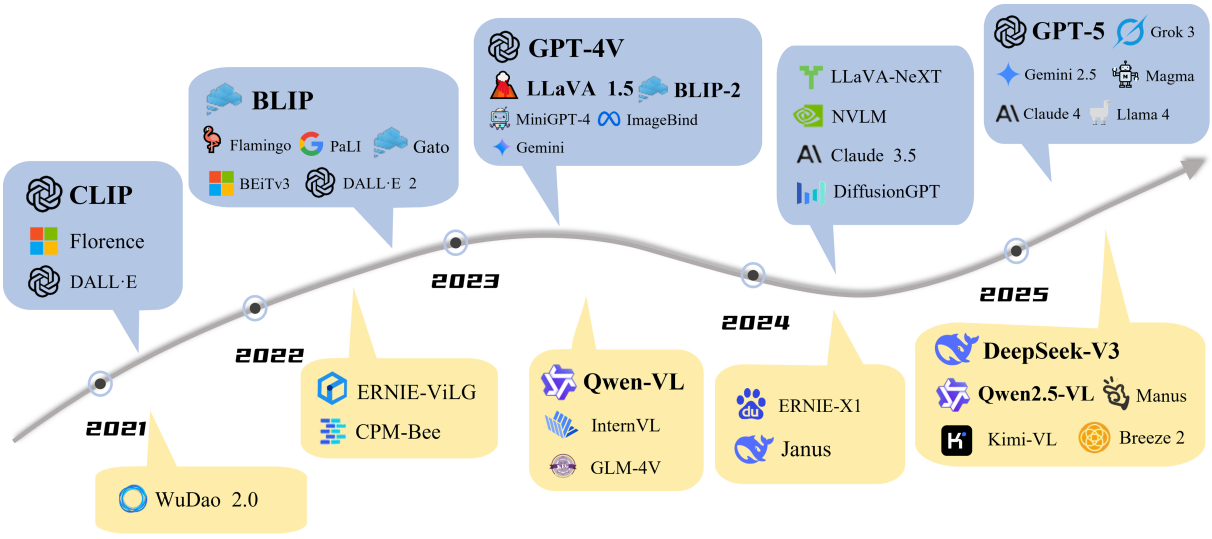

图 1　(网络版彩图) 国内外多模态大模型发展概略图.

Figure 1　(Color online) An overview of the development of multimodal large models at home and abroad.

由于多模态内容在信息传递中的表现力和覆盖面更大更广, 潜在误导性与社会影响也愈加显著 [2]. 因此, 伪造检测技术必须从传统的单一模态识别向多模态融合的智能感知范式跃迁.

多模态大模型 (large multimodal models, LMMs) 的发展使得文本模态在伪造检测中的作用逐渐凸显. 这些模型一方面显著提升伪造的图文关联内容的生成能力, 推动了更加复杂的跨模态伪造技术的发展; 另一方面, 也为伪造检测提供了更加丰富的跨模态语义分析和逻辑推理能力 [1]. 正如 NeurIPS 2024 年度会议的一项研究 [7] 指出的, 尽管多模态大模型在诸多下游任务中表现出色, 但其在伪造检测任务中的潜力尚未被充分挖掘.

使用大模型技术进行伪造判别、定位与解释已成为伪造检测技术研究领域的重要趋势. 相关研究总结大多未能有效覆盖这一新兴方向, 文献 [8~11] 初步探讨了多模态伪造技术的发展趋势, 但目前大多数研究仍主要集中于视听模态, 且较少探讨大模型尤其是大语言模型技术加持下的伪造检测. 同时, 另有研究 [12, 13] 聚焦于单一文本模态的伪造检测, 但缺乏对视觉语言伪造检测的整理. 本文首次系统性地聚焦于多模态大模型背景发展下伪造检测技术, 旨在系统性地回顾和评估截至 2025 年 10 月关于视觉 – 文本多模态大模型伪造检测的最新研究进展. 通过梳理相关技术发展、多模态数据集, 本文旨在揭示该领域的研究趋势与关键挑战, 为未来相关研究提供有益参考与理论支持.

2　背景

2.1　多模态大模型技术

LMMs 是指能够同时处理多种类型的输入和输出 (如文本、图像、音频等) 的大型人工智能模型. 这些模型的提出标志着人工智能在处理视觉与语言交织数据方面的重大进展 [14]. 近年来, 随着 Transformer 架构在视觉和语言领域的突破 [15~19] (如图 1 所示), 多个 LMMs 相继被提出, 推动了自然语言处理 (NLP) 和计算机视觉 (CV) 领域的融合与进步, 现将典型模型介绍如下.

CLIP (contrastive language-image pre-training) [19]: 由 OpenAI 于 2021 年推出, CLIP 通过学习海量的图文对数据, 实现了图像和文本的高效对比学习, 该模型具有强大的零样本图像分类和图像检索能力, 成为图文理解领域的基准模型之一.

DALL·E 2 [20]: 同样由 OpenAI 开发, 基于 CLIP 的技术, DALL·E 2 能够根据文本描述生成高质量、逼真的图像, 展示了文本到图像生成任务中的强大能力, 进一步推动了多模态生成技术的发展.

BLIP (bootstrapping language-image pre-training) [18]: 由 Salesforce 于 2022 年开发, BLIP 采用了全新的编码器 – 解码器多模态混合结构, 并引入了基于大规模野生数据清洗的预训练策略; 基于 BLIP 的 BLIP-2 进一步引入了轻量级的 Querying Transformer, 通过两个训练阶段将视觉编码器与大型语言模型对齐, 显著提升了多模态任务的性能.

GPT-4V (generative pre-trained transformer) [17, 21, 22]: GPT 模型经历了不断的迭代与发展, GPT-4 在视觉 – 文本多模态输入的处理上取得了显著进展, 不仅能完成文本回答生成, 还具备视觉分类能力. 基于 GPT-4 的 GPT-4o 在文本、语音及视觉方面进一步优化, 提升了其多模态处理能力.

MiniGPT-4 [23]: 该模型通过将一个冻结的视觉编码器 BLIP-2 与一个冻结的大型语言模型 Vicuna [24] 结合, 并通过投影层对其进行对齐, 从而实现了多模态信息的融合. MiniGPT-4 的设计展示了较为简洁的多模态模型架构, 具有高效的性能.

LLaVA (large language and vision assistant) [25, 26]: 使用简单的线性层将视觉特征映射为文本特征, 通过 CLIP 和 LLaMA 的指令微调, 构建了开源的多模态大模型 LLaVA, 能够处理视觉和语言的交互任务.

Gemini [27]: 由谷歌开发的多模态大型语言模型, 2025 年 3 月, 谷歌发布了 Gemini 2.5, 进一步提升了模型的推理和分析能力.

除了上述模型, 近年来还涌现了多个新兴的多模态大模型, 如 Llama 4 [28], InternVL [29], Qwen-VL [30], DeepSeek-V3 [31] 等, 限于篇幅和研究范畴, 这些模型的详细内容在此不再赘述. 大模型技术的持续发展为本文所聚焦的伪造检测等复杂下游任务提供了新的解决方案.

2.2　深度伪造检测技术

传统的伪造检测方法通常将该任务建模为二元分类问题, 即通过卷积神经网络 (CNN) 提取输入图像的高维特征, 并将其分类为 “真实” 或 “伪造” [32, 33]. 在此基础上, 后续研究尝试通过提升模型的泛化能力与鲁棒性来提高检测效果. 例如, 部分方法引入了照明和颜色匹配分析、频域分析以及人物目光分析等技术 [34~36], 以捕捉深度伪造中的微小伪造痕迹. 这些方法通过利用图像二次加工的信息辅助, 取得了较为显著的性能提升. 然而, 这些方法仍面临不少挑战.

首先, 传统的伪造检测方法普遍缺乏有效的跨模态语义一致性建模 [37]. 主流方法聚焦于视觉单模态检测, 大多数融合策略仅在特征层级进行拼接, 忽视了不同模态 (如图像、文本、音频等) 之间的逻辑与因果关系, 导致检测结果的解释性不足. 在当前的伪造检测任务中, 尤其在实际应用中, 跨模态的语义一致性尤为重要, 缺乏这一能力使得传统方法在多模态数据融合时无法准确地进行深层推理.

其次, 传统方法依然停留在二元分类层面 [37], 缺乏对伪造内容的细粒度分析. 具体而言, 传统方法无法有效地识别伪造的具体位置、篡改类型或伪造的意图等重要信息 [38], 传统方法的功能局限性使得其在处理复杂且多维度的数据时, 难以满足细致的判别需求, 这在司法取证、虚假新闻辟谣等领域的应用中显得尤为薄弱.

相比之下, 随着多模态大模型的发展, 伪造检测技术进入了一个新的阶段. 如图 2 所示, 大模型通过融合图像、文本、音频等多种模态数据, 能够从多个角度进行信息推理与交叉验证, 从而显著提高检测的准确性与鲁棒性. 具体而言, 多模态大模型在跨模态语义对齐、推理能力和解释能力方面展现了巨大的潜力. 这些模型能够高效整合不同模态的数据, 不仅提升了模型的泛化能力, 还能够提供可解释的推理过程和生成相应的自然语言描述, 进而增强检测结果的可理解性. 因此, 多模态大模型在伪造检测方面展现出更加广阔的发展前景. 特别是在跨模态理解、推理和解释性分析等方面, 大模型的优势更加明显, 这也是未来伪造检测技术发展的必然趋势.

3　多模态大模型伪造检测方法

本节讨论大模型时代的伪造检测方法. 根据研究所采用的模型架构, 我们将伪造检测方法归为三类. 第一类是基于视觉 Transformer 模型 (Transformer model, TM) 的方法, 进一步再根据使用模型不同, 细分为基于 vision transformer (ViT) 和基于 swin transformer (SwinT) 的两种方法. 第二类是基于视觉 – 语言大模型 (vision

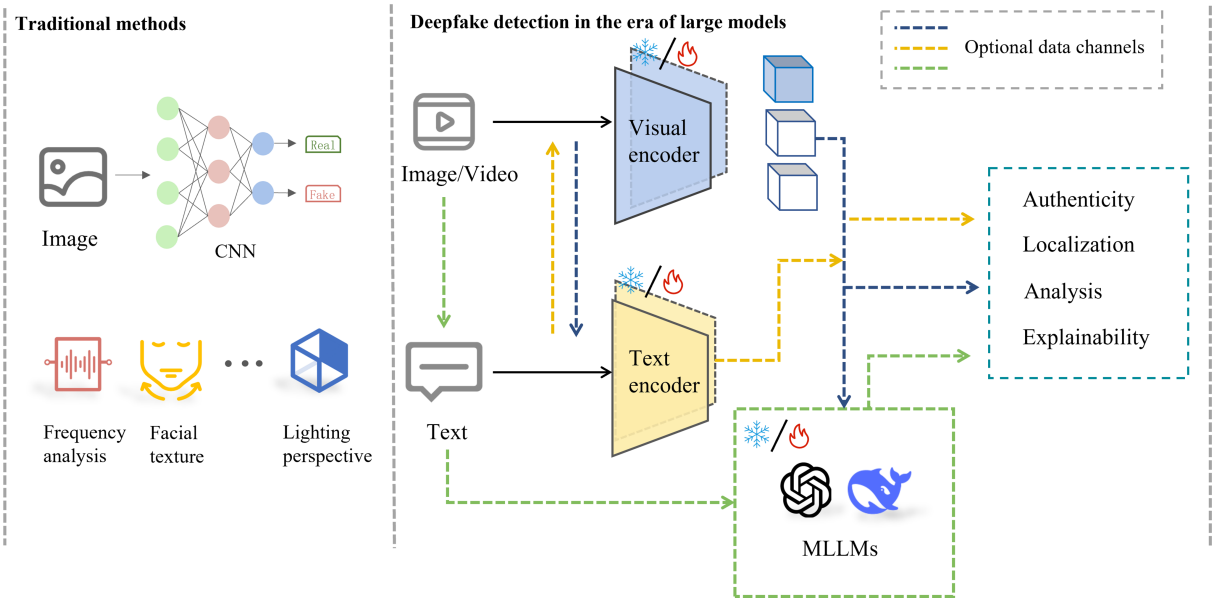

图 2　(网络版彩图) 传统伪造检测任务与多模态大模型伪造检测任务.

Figure 2　(Color online) Deepfake detection: traditional methods vs. large multimodal models.

language models, VLMs) 的方法, 并根据使用的模型不同, 分为基于 CLIP 和基于 BLIP 的两种方法. 最后, 第三类是基于多模态大语言模型 (multimodal large language models, MLLMs) 的方法, 由于多模态大语言模型 (如 GPT-4, LLaVA 等) 通常具有高度的通用性, 其应用范围和方法特征并不依赖于某一特定模型. 因此, 我们将该类方法按照研究角度细分为几个方向, 包括大语言模型伪造检测评估基准 (MLLMs-B, bench)、基于细粒度分析和可解释性推理的方法 (MLLMs-E, explainability)、基于跨模态语义不一致的检测方法 (MLLMs-CI, cross-modal inconsistency), 以及解决幻觉现象的策略 (MLLMs-H, hallucination). 我们在表 1 [2,5,7,14,37~85] 中对研究进行归类, 总结各类方法的典型代表, 并提供它们的开源项目地址[1)2)3)4)5)6)7)8)9)10)11)12)].

3.1　基于视觉大模型的伪造检测

在大语言模型迈向多模态之前, 单模态大语言模型已经在语言任务中展现出卓越的性能, 并逐渐吸引了伪造检测领域研究者的关注. ViT 和 SwinT 作为两种在视觉任务中取得巨大成功的视觉变换器模型, 凭借其创新的设计理念和强大的特征提取能力, 为多模态伪造检测领域提供了有力的支持.

一方面, 许多方法注重视觉与文本的模型架构融合, 结合 ViT 和 SwinT 等作为视觉编码器, 与 BERT [86] 和 RoBERTa [87] 等单模态大语言模型共同作用, 推动了多模态伪造检测技术的发展. 哈尔滨工业大学邵睿教授团队 [2] 提出了多模态层次化篡改推理模型 (HAMMER), 旨在检测和定位多模态媒体的操控. 该模型结合了 ViT-B/16 图像编码器和 BERT 文本编码器, 通过操控感知对比学习对图像和文本嵌入进行浅层对齐, 捕捉细微的操控迹象; 同时, 利用模态感知的跨注意力机制进行深层推理. 文献 [45] 提出的 MPFN (multimodal progressive

1) https://github.com/rshaojimmy/MultiModal-DeepFake.
2) https://github.com/Reality-Defender/Research-DD-VQA.git.
3) https://github.com/sfimediafutures/CLIPping-the-Deception.
4) https://github.com/chuangchuangtan/C2P-CLIP-DeepfakeDetection.
5) https://github.com/SparkleXFantasy/MM-Det.
6) https://github.com/NickyFot/HitchhikersGuide.
7) https://github.com/liuxuannan/MMFakeBench.
8) https://github.com/Forensics-Bench.
9) https://github.com/zhipeixu/FakeShield.
10) https://github.com/liuxuannan/FAK-Owl.
11) https://github.com/MischaQI/Sniffer.
12) https://github.com/skJack/VLFFD.

表 1　大模型时代伪造检测分类与典型方法.

Table 1　Classification and representative methods of Deepfake detection in the era of large models.

Category		Methods	Typical methods & code
TM	ViT	[2, 37, 39∼44]	HAMMER[1] [2], Common Sense[2] [44]
	SwinT	[45∼47]	
VLM	CLIP	[5, 48∼61, 69]	CLIPping[3] [48], C2P-CLIP[4] [49], MM-Det[5] [52]
	BLIP	[7, 38, 62, 63]	HitchhikersGuide[6] [7]
MLLMs	MLLMs-B	[64∼68]	MMFakeBench[7] [67], Forensics-bench[8] [68]
	MLLMs-E	[7, 14, 69∼79]	FakeShield[9] [14], FAK-Owl[10] [77]
	MLLMs-CI	[52, 74, 77, 80∼83]	Sniffer[11] [81]
	MLLMs-H	[47, 82, 84, 85]	VLFFD[12] [84]

fusion network), 结合基于 BERT 的文本编码器与基于 SwinT 和 VGG19 的视觉编码器, 通过图像频域信息的融合, 进一步提升了多模态伪造检测的效果. 山东理工大学高明亮教授团队 [39] 提出了一种统一的多模态伪造检测框架 ViKI (vision-language knowledge interaction), 该框架通过图文对齐与交互机制实现伪造内容的检测与定位, 利用 ViT 作为视觉编码器和 BERT 作为文本编码器, 采用对比学习与 MMD 距离约束将图文特征嵌入统一空间, 从而实现跨模态一致性的表示. 蚂蚁人工智能研究院 Qu 等 [46] 使用 SwinT 提取图像中的可疑区域作为初始伪造候选, 并以此引导多模态大模型 (如 GPT-4o) 生成对篡改内容的自然语言解释. 香港理工大学 Lap-pui Chau 教授团队 [47] 将 SwinT 与大语言模型 Qwen2.5 结合, 以综合时空伪造特征提取与多模态语言推理的方式, 提升了视频伪造检测的泛化能力和可解释性.

另一方面, 部分研究着力于多模态特征的信息融合, 嘉泉大学 Chang Choi 教授团队 [40] 根据少样本深度伪造场景, 设计了一种三模态多模态大模型 (TMiformer). 该模型通过使用 Word2Vec [88] 和 MFCC [89] 方法对文本和音频进行处理, 之后将视觉特征与音频、文本特征进行跨模态残差连接, 有效应对了伪造检测中的复杂多模态问题. 文献 [41] 使用正则化和交互式知识聚合来实现模态映射和对齐, 并通过 BERT 和 ViT 进行特征提取和融合. 中国科学技术大学刘斌教授团队 [42] 认为跨模态容易出现 “模态竞争” 问题, 即模型过于依赖某一模态的信息, 可能忽视其他模态, 从而导致不准确的结果, 提出了一种融合模态特异性与跨模态交互的多模态篡改检测与定位框架, 基于 ViT 和 RoBERTa 作为图像与文本编码器, 引入了隐式篡改查询模块和双分支交叉注意力机制, 提升了对伪造区域的定位能力. 中国科学技术大学于灵云等 [43] 为了解决现有方法中 “任务间竞争同一图像特征空间” 的问题, 将图文信息的语义不一致解耦为伪造特征 (forgery features), 如边缘不一致、纹理异常, 以及内容特征 (content features), 如人物、场景、语义信息, 分别用于不同的子任务, 避免互相干扰. Zhang 等 [44] 使用 ViT 和 BERT 提取视觉与语言信息, 将问题文本转化为上下文相关的表示向量, 传入多模态交互模块后与视觉词元进行跨模态对齐学习. 中国科学院雷震教授团队 [37] 提出了上下文语义一致性学习 (CSCL) 方法. 该方法通过使用两个解码器, 分别确保单模态内的上下文一致性和跨模态语义对齐, 最终通过一致性矩阵评估细粒度嵌入的连贯性, 从而有效提升了模型在多模态伪造检测任务上的精度.

3.2　基于视觉 – 语言大模型的伪造检测

3.2.1　基于 CLIP 模型的方法

CLIP [19] 作为一种将图像与文本对齐的视觉语言预训练模型, 近年来被广泛应用于深度伪造检测任务. 通过 CLIP 模型的跨模态对比学习能力, 研究者尝试利用其在语义空间中对图文关系的建模能力, 实现伪造内容的有效识别. 现有研究主要沿两条路径展开: 一是围绕 CLIP 模型自身机制进行结构优化与迁移学习, 二是结合伪造检测任务的具体需求进行定制化设计.

首先, 围绕 CLIP 模型机制的研究主要致力于提升其结构能力与泛化适应性. 例如, 卑尔根大学 Khan 等 [48] 采用了一种名为 “语境优化” (CoOp), 即向文本提示 (prompt) 的前部、中部或尾部注入可学习的向量, 探

索包括微调、线性探测、及时调整和训练适配器网络的转移学习策略. 北京交通大学 Tan 等 [49] 从 CLIP 原理出发, 提出了 C2P-CLIP 方法, 首次揭示了 CLIP 在伪造检测中并非直接理解 “真或假”, 而是借助对图像概念的识别实现间接分类. 基于这一发现, C2P-CLIP 向文本提示中注入 “类别通用提示 (category common prompt)”, 结合原始图像描述形成新的训练语义, 进一步增强 CLIP 编码器对真假图像的判别能力. 中国科学技术大学 Liu 等 [69] 在 CLIP 图像编码的基础上引入 “视觉词汇表” 构建模块, 通过图文对齐、掩码文本对齐与任务特定微调三阶段训练流程, 释放 CLIP 编码器的结构化能力与语义泛化能力. 中山大学卢伟教授团队 [50] 在其 RaCMC 的工作中, 以 CLIP 提取的特征为基础, 引入了掩码注意力机制和残差感知交互进行模态间融合, 以及通过最大均值差异和余弦相似度对特征进行优化. 合肥工业大学韦炎炎等 [51] 通过 CLIP 的语义检索能力挑选最相关的示例, 利用图传播与几何门控的策略对示例进行筛选, 从而无需微调即可提高多模态大语言模型在伪造检测任务中的泛化性与准确性.

其次, 任务导向的改进方法聚焦于添加特定机制与模块, 提升了 CLIP 对具体伪造类型的适应能力. 例如, 在视频伪造检测领域, 上海交通大学刘笑宏教授团队 [52] 提出了 MM-Det 框架, 将 CLIP 图像编码器与多模态大模型结合, 用于检测由扩散模型生成的视频内容伪造, 强调了时序不一致性与局部生成缺陷的联合建模. Liu 等 [53] 针对社交媒体短视频中多模态信息错配问题, 融合 CLIP 与 RoBERTa 编码器, 通过事件触发词提取与三模态一致性聚合模块, 识别视频语境与文本陈述之间的语义不一致性.

针对人脸伪造检测问题, 深圳大学李斌教授团队 [54] 利用 CLIP 的视觉编码器与可学习的视觉提示进行人脸伪造检测, 将面部嵌入信息与文本提示结合. 邓伟洪教授团队 [55] 提出了一种基于 LongClip 的文本编码器, 结合 ViT 图片编码器进行细粒度文本引导的面部攻击检测方法. 美国布法罗大学 Siwei Lyu 教授团队 [56] 提出的 SJEDD 方法, 同时结合数据集扩展、联合嵌入和动态调整损失权重, 成功实现了特定面部区域的细粒度分类, 并展现出优异的泛化能力.

针对扩散模型生成的图像检测, 摩德纳大学 Amoroso 等 [57] 提出了 “Parents and Children” 框架, 强调了对扩散模型生成图像进行语义聚类的特定化伪造检测任务, 并使用 CLIP/OpenCLIP 提取图像语义特征进行分析. 为提升模型对伪造生成图像的敏感性, 清华大学刘欢教授团队 [5] 基于 CLIP 预训练模型提出了 FatFormer 模型, 通过引入伪造感知适配器与语言引导对齐模块, 设计了增强型图文对比监督目标. 深圳大学沈琳琳教授团队 [58] 也针对扩散模型生成的人脸伪造检测问题, 提出了 MFCLIP (multi-modal fine-grained CLIP) 模型, 该模型在 CLIP 框架基础上引入了图像模态与噪声模态的细粒度伪造特征提取机制, 通过动态正负样本权重控制, 提升了模型跨模态对齐的鲁棒性和伪造辨识能力, 尤其在扩散图像检测与泛化性方面表现出显著优势.

针对检测模型的细粒度分类和解释能力, 上海交通大学刘笑宏教授团队 [59] 在 M2F2-Det 的工作中, 引入了伪造提示学习机制, 通过在文本提示中注入通用伪造提示和层级伪造标记, 并提出桥接适配器概念, 将图像编码器的中间特征与 LLM 结合, 整合图像特征和文本生成, 提升了伪造检测的准确性和解释能力. Zhao 等 [60] 认为传统伪造检测的层次化推理策略往往先进行浅层图文融合, 再依次传递到各子任务 (分类、定位), 其在研究中提出 CrUr 集中推理, 直接在统一的图文语义空间中处理所有伪造检测任务, 通过 Transformer 统一建模, 避免重复处理. 文献 [61] 发现不同伪造检测数据集直接合并训练会造成领域冲突, 通过引入 Dataset Embedding 机制与 Meta-Domain 优化策略, 结合图文对齐和掩码图像重建模块进行多模态融合.

3.2.2　基于 BLIP 模型的方法

BLIP 及其改进版本 (如 BLIP-2, InstructBLIP) 在视觉语言理解和生成任务中展现出出色的表现. 其 “视觉 – 语言双向融合机制” 成为伪造检测任务的有力工具. 密歇根州立大学 Zhang 等 [7] 构建了名为 BLIP-TI 的伪造检测方法, 引入了图像对与文本对比损失, 模型生成的解释不仅覆盖伪造部位, 还可提供基于常识的因果描述. 文献 [38] 提出的 ForgeryTalker 框架, 通过 Forgery Prompter 模块精确识别伪造区域并生成区域提示, 利用 InstructBLIP 模型生成关于伪造痕迹的解释性文本, 其在 CIDEr 分数和伪造区域 IoU 上均优于传统 BLIP 模型与 LISA-7B 等多模态模型. 德国亥姆霍兹信息安全中心张阳等 [62] 提出的 ZeroFake 方法将 BLIP 应用于 “零样本伪造图像检测” 场景中. 该方法针对由 Stable Diffusion, DALL·E 2 等模型生成的图像, 利用 BLIP 自动生成

图像描述提示词, 并基于 DDIM 扩散逆过程构造多个对抗性提示, 展示了 BLIP 在生成图像伪造理解中的结构引导与语义解释优势. 此外, BLIP 也被用于辅助检测全局语义不一致的伪造图像, 昆士兰大学 Gagandeep Singh 等 [63] 通过 BLIP 模型生成描述, 并利用 MiniLM 对前背景描述进行语义相似度计算, 显著提升了模型对 “人物真实、场景荒谬” 类伪造的检测能力.

3.3　基于多模态大语言模型的伪造检测

随着 MLLMs 在视觉问答、多模态推理与指令理解等任务中取得显著进展 [90], 其在伪造检测领域的应用也逐渐受到关注. MLLMs 结合了大型语言模型 (如 GPT-3, LLaMA-3 等) 的自然语言理解与生成能力, 以及对图像、文本、音频等多模态信息的联合处理能力, 具备在复杂语境中进行语义推理与跨模态一致性判断的潜力.

3.3.1　MLLMs 测试评估基准

近年来, 已有多项研究对 MLLM 在伪造检测任务中的基础能力进行了系统性评估, 并发现了将 MLLMs 直接应用于该任务的缺陷. 例如, 美国布法罗大学 Siwei Lyu 教授团队 [64] 提出了一种无需编程即可调用 GPT-4, Gemini 等主流 MLLM 的测试方法, 通过输入人脸图像与文本提示, 考察其在人脸伪造识别任务中的性能. 结果显示, 当前模型在真实性判断、推理链条构建与对抗样本鲁棒性等方面仍存在明显局限. 此外, 伊拉克巴格达大学 Omar 等 [65] 也对 GPT-4, Bard 与 Bing 等模型在深度伪造图像识别中的能力进行了对比, 进一步揭示了当前 MLLMs 在视觉伪造检测中的可行性与挑战性并存. 为统一评估 MLLMs 在多维伪造任务中的表现. 普渡大学胡暑教授团队 [66] 对 MLLMs 在真实政治语境下的伪造检测能力进行了分析. 该团队收集了来自 TikTok, X(Twitter), Facebook 等社交平台上真实流通的政治图像与视频伪造案例, 系统评估了包括 GPT-5, Claude 4.5, Gemini 2.5, LLaVA, CogVLM 等 10 个主流模型在图像与视频检测中的表现, 提出了学术界与政府开发的检测器性能不足、检测器在视频域的泛化能力差于图像域等问题. 北京邮电大学李佩佩教授团队 [67] 构建了 MMFakeBench 测评基准, 覆盖三类任务: 文本真实性检测、视觉真实性检测与图文语义一致性判断, 该基准共评估了 GPT-4V, Claude, LLaVA-1.5, MiniGPT-4 等 15 个主流模型, 系统分析了其在真实性判断、推理有效性方面的能力边界. 香港大学罗平教授团队 [68] 通过对包括 LLaVA, InternVL, GPT-4o, Gemini 在内的 25 个主流大模型进行了系统评估, 揭示了当前模型在复杂伪造类型和多模态场景下存在显著的识别偏差与推理瓶颈. 例如, GPT-4o 等闭源模型因回答更为保守, 检测性能反而弱于开源模型; 尽管多数模型在伪造二分类任务中表现尚可, 但在伪造区域定位 (SLS/SLD) 与时间定位 (TL) 等更复杂任务上普遍表现不佳. 此外, 在 “鲁棒性检测” 中发现, 当输入图像加入对比度或饱和度扰动后, 模型的伪造检测能力明显下降, 进一步暴露出其在真实环境中应用的稳定性问题.

3.3.2　基于细粒度分析与可解释性推理的方法

多模态大语言模型所具备的细粒度分类与逻辑推理能力, 与当前伪造检测任务日益增长的可解释性需求高度契合. 在深度伪造检测中, 模型不仅需输出 “真假” 判断结果, 更需要能够说明判断依据与决策逻辑, 以提升系统的透明性与用户信任度.

研究者逐渐关注如何通过优化提示词 (prompt) 设计, 引导 MLLM 给出更具逻辑性的推理过程与因果链条. 香港城市大学王诗淇教授团队 [70] 通过提示词设计针对多模态大模型伪造检测的因果研究和细粒度任务, 对比多种匹配策略并揭示在 AI 风险控制中重视解释性、推理性元素的必要性. 中国科学院万军教授团队 [14] 提出了一种新型的多属性思维链 (MA-COT) 范式, 提高了多模态人脸伪造检测任务中多模态大模型的鲁棒性和可解释性, 在此基础上, 团队在 Veritas [71] 的工作中进一步利用偏好标注数据和奖励函数, 引导模型学习更细粒度和更准确的推理模式. 中国科学技术大学 Liu 等 [69] 通过构建视觉词汇表与掩码感知机制, 将伪造区域掩码转化为语言提示 token, 引导 GPT-4 生成高一致性的解释性文本. 北京大学张健教授团队 [72] 提出了 FakeShield, 结合多模态大语言模型的可解释图像伪造检测与定位框架通过引入 GPT-4o 和分析篡改图像及其相应的二进制篡改掩码, 精确定位图像中被篡改的区域, 通过模型生成的文本描述进行引导. 启元实验室傅睿博教授等 [73] 以

Vicuna-7B 为语言模型, 结合 ImageBind [91] 作为多模态编码器, 并引入残差网络面部伪造特征提取器, 用于强化对人脸主导型伪造的识别, 构建多层提示学习模块引导大语言模型进行链式思维.

扩充细粒度标签乃至重构为问题与回答, 也是一种主流路径. 例如, 西安交通大学罗敏楠教授团队 [74] 提出的 MGCA (multi-granularity clue alignment) 模型, 包括图像伪造 (image fabrication)、图像无证据 (ImageNoE)、实体不一致 (EntityInc)、事件不一致 (EventInc) 与时间不一致 (TimeInc) 五类伪造归因标签. 卢森堡大学 Niki 等 [7] 将人脸伪造检测任务转换为视觉问题答案 (VQA) 任务, 并在细粒度多标签识别与自然语言解释层面进行了全面评估. 文献 [75] 提出了一种基于大型视觉语言模型的可解释深伪检测框架, 通过卷积网络计算图像特征与文本描述之间的相似度, 生成一致性热图与伪造分割图, 并编码为伪造提示, 与视觉/文本提示一同输入 PandaGPT 多轮对话解释. 清华大学杨文明教授团队 [76] 提出的 FFAA (face forgery analysis assistant) 是一个基于 LLaVA 精调而成的多模态伪造检测框架. 通过内置的 MIDS 模块, 模型进一步比较两个假设回答与图像内容的一致性, 自动选择更可信的解释作为最终判断.

此外, 还有其他角度提升模型的解释能力. 中国科学院黄怀波教授团队 [77] 使用双分支的跨注意力机制, 指导视觉和文本特征之间的交互, 结合伪造特定知识和语言模型, 增强了对图像和文本操控的推理能力. 新加坡管理大学 Wei Gao 教授团队 [78] 从数据集采样的角度出发, 提出了一种利用合成数据训练 LLaVA 的策略, 不依赖人工标注, 而是从多个大规模合成伪造数据集中筛选代表性的训练样本用于微调, 输出伪造解释与生成推理链条. 北京交通大学 Tan 等 [79] 提出的 AnomReasonor 模型进一步将可解释检测拓展至语义层面. 该模型基于 Qwen2.5 微调, 区别于依赖像素或纹理特征的传统伪造检测, 属于一类面向语义理解的高层次异常检测方法, 重点关注内容在常识、物理规律与逻辑一致性方面的细粒度分析与解释.

3.3.3　基于跨模态语义不一致问题的方法

跨模态语义不一致问题已成为伪造检测领域的重要切入点. 文献 [80] 指出, 图文信息的语义不一致 (semantic inconsistency) 是辨别虚假内容的一个关键线索. 为了帮助识别这种语义不一致, 研究将其分为脱离上下文 OOC (out-of-context) 和跨模态命名实体不一致 NEI (cross-modal named entity inconsistency) 两种类型. 此外, 文献 [74] 将跨模态的不一致性细化为实体、事件与时间三个层面, 并分别构建了文本与图像模态下的实体识别、事件理解与时间标注机制, 在语义一致性判断中, 利用 LLaVA, Google Lens 提取文本与图像的关键语义单元.

当前主流方法聚焦于多模态大模型架构设计与内外部知识融合, 例如, 新加坡国立大学 Qi 等 [81] 设计了 SNIFFER, 一种专为检测 OOC 虚假信息而设计的多模态大语言模型结合内外部验证机制, 同时输出判断结果和可解释理由. 中国科学院黄怀波教授团队 [77] 通过将伪造特定知识注入 LVLM, 以 Vicuna-7B 为语言基座, 结合 ImageBind [91] 作为多模态编码器, 增强了跨模态语义检测的鲁棒性. 上海人工智能实验室何聪辉等 [82] 设计了专门用于检测合成图像模型的多模态大模型 FakeVLM, 通过从预训练的 LMM 的最后一层中提取视觉特征, 并训练轻质线性分类器以确定多模态检测源的真实性. 德国达姆施塔特工业大学 Marcus Rohrbach 教授团队 [83] 结合多模态大语言模型与外部工具 (如网页搜索、图像搜索、反向图像搜索和地理定位), 从动态规划与多轮推理的角度出发, 构建了 DEFAME 框架. 该框架应对多模态事实核查中的跨模态语义不一致问题. 另外, 在特定的伪造检测子任务中, 如针对视频模态的伪造检测, 复旦大学陈静静教授团队 [52] 提出了将伪造定位任务纳入对视频跨帧不一致、重构残差与语义矛盾的协同分析, 并提出了多模态大模型 (LLaVA) + 时空特征建模 (ST Branch) + 动态融合机制.

3.3.4　基于幻觉问题的方法

幻觉问题 (hallucination) 是多模态大语言模型的典型缺陷之一, 指的是生成式 AI 模型在生成内容时, 产生与输入源内容不一致、荒谬或缺乏事实依据的信息, 即模型自信地提供错误或虚构的信息 [92]. 在伪造检测领域, 大模型的幻觉现象不仅影响检测的准确性, 还削弱了模型的可信度. 研究者也发现了这一问题, 部分研究尝试通过知识嵌入、掩码提示、多模型聚合等方式缓解多模态大模型的幻觉问题. 例如, 厦门大学孙晓帅教授团队 [84] 针对当前伪造检测注释中大模型存在的幻觉问题, 提出了面部伪造文本生成器 (FFTG), 利用伪造掩码的初始区

表 2　多模态伪造检测数据集统计.

Table 2　Multimodal forgery detection dataset statistics.

Dataset	Type	Year	Modalities	Description
DGM4 [2]	Textual forgery	2023	Image, Text	230000 image-text pairs: 152574 forged, 77426 real
IDForge [4]	Textual forgery	2024	Video, Text, Audio	169311 forged clips; 214438 real references; each clip 5∼7 s
M3A [3]	Textual forgery	2024	Video, Image, Text, Audio	708425 real samples, 6566386 multimodal fake samples
AMG [74]	Textual forgery	2024	Image, Text	5022 image-text pairs: 3018 real news, 2004 fake news
Forensics-Bench [68]	Textual forgery	2025	Video, Image, Text	63292 samples with multi-choice QA and tri-modal content
MMFakeBench [67]	Textual forgery	2025	Image, Text	11000 image-text pairs; includes 12 types of multimodal manipulations
SID-Set [93]	Explanation-based	2024	Image, Text	300000 images (100k real, 100k synthetic, 100k tampered), 3000 explanations
DD-VQA [44]	Explanation-based	2024	Image, Text	2968 images, 14782 QA pairs, each image with 3∼6 questions
LOKI [94]	Explanation-based	2024	Video, Image, Text, Audio, 3D	18000+ task samples including binary, multiple-choice, and explanation-based questions
FakeBench [70]	Explanation-based	2024	Image, Text	6000 images, 54000 QA pairs covering judgment, explanation, and reasoning tasks
ExDDV [95]	Explanation-based	2024	Video, Text	5369 videos, 21282 forgery annotations and explanations, covering methods like Face2Face and DF-VAE
FFA-VQA [76]	Explanation-based	2024	Image, Text	7436 images with hypotheses, explanations, forgery tags, and localized regions
VLForgery [85]	Explanation-based	2025	Image, Text	96500 real images, 446600 fake images (composite/partial), with EkCot-style reasoning chains
FakeClue [82]	Explanation-based	2025	Image, Text	100000+ images, each annotated with forged clue explanations by LMMs
MS-UFAD [55]	Explanation-based	2025	Image, Text, Video	5000 real videos, 260000 fake videos, 60000 images, all with clues and attribute descriptions
COCOFake [57]	Prompt-based generation	2023	Image, Text	113287 real images, each with 5 captions; 1200000 fake images generated via SD in a “1 real + 5 fake” structure
AutoSplice [96]	Prompt-based generation	2023	Image, Text	3621 fake and 2273 real images with pixel-level masks
DFLIP-3K [97]	Prompt-based generation	2024	Image, Text	300000 forged images and 190000 generation prompts
DiFF [98]	Prompt-based generation	2024	Image, Text	23661 real images, 537466 fake images, over 30000 associated prompts

域和类型识别来生成准确的文本描述. 深圳大学李斌教授团队 [85] 在 VLForgery 工作中引入低级视觉管道比较与文本描述模块, 构建了多模态微调与三任务联合推理框架 (检测、定位、归因), 有效提升了 MLLM 在扩散式伪造图像下的推理准确率与可解释能力. 另外, 上海人工智能实验室何聪辉等 [82] 通过使用多个大语言模型 (如 Qwen2-VL, InternVL, Deepseek) 的聚合注释策略以及外部标签提示引导模型关注图像的关键区域, 减少了合成图像检测中的幻觉问题. 与图像场景下的幻觉问题相比, 香港理工大学 Lap-pui Chau 教授团队 [47] 指出, 视频检测中面临更复杂的时序不一致、微表情无法被捕捉等动态幻觉干扰, 强调通过面部关键点坐标的跨帧动态、模糊度变化、颜色分布变化等指标作为 “硬约束”, 压制大模型的幻觉生成.

4　多模态伪造检测数据集

当前新兴多模态伪造检测数据集, 如表 2 [2∼4, 44, 55, 57, 67, 68, 70, 74, 76, 82, 85, 93∼98] 所示, 在任务设计与模态组织方面呈现出显著差异. 特别是文本模态在伪造检测流程中所处的位置, 可作为一种具有代表性的划分依据. 本文将现有代表性数据集划分为三类: (1) 文本作为输入模态; (2) 文本作为输出模态; (3) 文本作为其他伪造模态的生成提示.

4.1　含伪造文本的多模态数据集

本类数据集以文本作为伪造检测的输入模态之一, 其中文本本身可能为伪造内容的来源, 或与其他模态 (图像、音频、视频) 之间存在语义错配、身份不一致、立场矛盾等问题. 此类数据集主要服务于跨模态一致性判断、文本真实性评估与事实核查等任务, 具有极高的现实应用价值, 尤其广泛应用于社交媒体、新闻资讯与 AI 生成内容识别场景中.

(1) DGM4 [2] 数据集构建于经筛选后的 VisualNews [99] 数据集, 该数据集来源于现实世界的新闻报道, 包括 BBC, The Guardian 和 The Washington Post 等主流媒体. DGM4 主要聚焦于人物相关新闻, 针对图像与文本的输入模态设计了四种篡改类型: 人脸替换 (face swap)、人脸属性修改 (face attribute)、文本替换 (text swap) 与文本情感倾向篡改 (text attribute). 最终构建了包含约 23 万条图像 – 文本对的多模态伪造数据集, 并提供了丰富的注释信息, 支持对伪造类型的分类、伪造区域的定位及多模态一致性推理等任务.

(2) DGM4+ 数据集 [100] 是对 DGM4 数据集的扩展, 以弥补原始数据集中缺乏全局语义错配问题的不足, 引入了新的篡改类型: 背景不一致与文本情感篡改, 并使用 OCR 去除可识别文本, 防止 “捷径” 识别, 共计生成 5000 条合成图文样本.

(3) IDForge [4] 数据集是一个身份驱动的多模态伪造检测数据集, 涵盖视频、音频与文本三种模态. 该数据集特别引入了文本伪造, 包括基于 GPT-3.5 的文本生成与跨身份文本互换, 以模拟现实中语义伪造场景. IDForge 共收集了来自 YouTube 的约 40 万个时长在 5∼7 s 之间的视频镜头, 涵盖 54 位名人, 并为每位个体提供丰富的伪造样本与参考样本, 支持多模态身份感知伪造检测任务的深入研究.

(4) M3A [3] 数据集是一个面向多模态社交媒体、新闻报道真实性分析的大规模伪造数据集, 涵盖文本、图像、音频与视频四种模态, 共包含 708425 条真实样本和 6566386 条由多种方式生成的伪造样本, 支持包括语境错配检测、深度伪造识别、跨模态事实冲突判别与开放域泛化测试等多项任务.

(5) AMG (attribution multi-granularity) [74] 数据集是面向多模态虚假新闻检测与归因任务的大规模数据集, 数据集覆盖 Instagram, Facebook 与 X(Twitter) 三大社交平台, 涵盖 2016∼2024 年间的图文新闻样本, 强调图文之间的跨模态一致性与时序关系. AMG 不仅标注新闻的真假属性, 还引入了五种伪造类型的细粒度归因标签: 图像伪造 (image fabrication)、非证据图像 (non-evidential image)、实体不一致 (entity inconsistency)、事件不一致 (event inconsistency) 与时间不一致 (time inconsistency). 该数据集共包含 5022 条图文对, 其中真实新闻为 3018 条, 虚假新闻为 2004 条.

(6) MMFakeBench [67] 数据集是一个混合来源的大型图文虚假信息检测数据集. 该数据集共包含约 11000 个图文对, 覆盖文本真实性扭曲、视觉真实性扭曲、图文一致性冲突、真实数据伪造类别, 涵盖自然生成谣言、ChatGPT 虚假文本、AI 合成图像、语义错配等现实操控方式. 可应用于多模态模型的可解释性、安全性与推理评估任务.

(7) Forensics-Bench [68] 数据集是一个专门面向大模型检测评估的数据集, 伪造类型包括整图合成、人脸替换、属性编辑、文本篡改、拼接、删除、风格迁移、时序篡改等共 21 类复杂伪造操作. 该数据集共包含 63292 个样本, 每个样本中有统一格式的图像/视频/文本内容及对应的多项选择题, 用于评估模型在伪造识别、伪造区域定位与推理解释等方面的能力.

(8) COVID-VTS [53] 是一个针对短视频平台设计的多模态事实核查数据集. 数据来源自 TikTok, X(Twitter), Facebook 上关于 COVID-19 的短视频, 包含 10000 条样本 (文本 – 音频 – 视频), 以及 “真实” “文本伪造” “语音、视频伪造” 三类标注.

(9) NewsCLIPpings [101] 是一个大规模的图文错配检测数据集, 基于 VisualNews 中的 509730 条真实图文新闻样本采用多种语义、人物、场景维度的匹配方法生成伪造样本, 最终构建了约 988000 个图文对样本, 其中每条文本既用于真实图配对, 也参与伪造图配对, 旨在构建挑战性的 “真实图 + 真实文” 语义错配检测任务, 强调多模态一致性推理能力, 避免传统 “文本改写伪造” 带来的单模态偏差.

(10) MFND (multimodal fake news detection) 数据集 [102] 是一套面向多模态假新闻检测与定位的大规模基准数据集, 共使用 11 种图像/文本生成与编辑方法, 综合模拟真实社交媒体环境中的假新闻, 包含 125000 条多模态伪造新闻样本, 覆盖 11 种深度伪造技术, 包含新闻真假判断、图像/文本真伪检测、图像伪造区域定位等四类任务标注.

4.2　含解释文本的多模态数据集

伪造输出文本数据集强调在伪造检测过程中生成可解释的文本输出, 文本内容作为模型检测结果的一部分, 用于描述伪造类型、分析可疑区域、进行因果或链式推理等. 这类数据集多数以视觉问答 (VQA)、因果推理、线索挖掘与细粒度文本解释为核心任务, 广泛用于训练和评估多模态大模型在伪造理解任务中的推理与输出透明性.

(1) SID-Set [93] 数据集是一个面向社交媒体场景的图像伪造检测、定位与解释任务的数据集, 共涵盖 30 万张图像, 包括 AI 生成图像、图像篡改样本以及真实图像, 并配有 GPT-4o 注释信息. 真实图像来源于 OpenImages

V7, 而伪造图像则由生成模型 FLUX 在 Flickr30k [103] 和 COCO [104] 数据集基础上生成, 涵盖对象替换与部分区域篡改等复杂伪造形式, 具有较高的真实性与多样性.

(2) DD-VQA [44] 数据集是一个以 FaceForensics++ [105] 数据集 (FF++) [106] 为基础构建的图像伪造问答数据集, 主要用于研究基于视觉与语言的伪造检测与解释. 该数据集通过 Amazon Mechanical Turk 平台收集标注信息, 问题聚焦于图像整体及面部局部区域 (如眉毛、眼睛、鼻子、皮肤等) 是否存在伪造, 共包含 2968 张图像与 14782 个图文问答对, 每个问题同时提供真假判断与基于常识的解释文本, 适用于多模态可解释伪造检测模型的训练与评估.

(3) LOKI [94] 数据集是一个面向大模型评估的多模态合成数据检测基准数据集, 涵盖图像、视频、音频、文本、三维图像等模态, 共包含超过 18000 个任务样本, 细分为 26 个伪造类型类别. 该数据集设计了多种任务形式, 包括真假判断、多项选择、伪造细节识别以及自然语言解释, 全面考察模型对伪造信息的识别能力与解释能力. 每个样本均配有由多模态大模型生成的细粒度自然语言解释, 指示伪造痕迹.

(4) FakeBench [70] 数据集是一个面向多维度伪造解释任务的大规模多模态数据集, 涵盖 3000 张真实图像与 3000 张伪造图像, 共设计 54000 个问题, 包括检测判断、因果解释、细粒度伪造分析等任务, 支持多种自然语言提示 (如 Basic, In-context, Chain-of-thought, Free Prompting), 通过三个子集 (FakeClass, FakeClue, FakeQA) 全面评估 LMMs 对图像真实性的理解、解释与推理能力.

(5) ExDDV [95] 是面向可解释视频伪造检测的数据集, 整合自 4 个主流数据集: FaceForensics++ [105], DeeperForensics [107], DFDC [108], BioDeepAV [10], 共 5369 个视频 (4369 伪造 + 1000 真实), 含 21282 个伪造区域坐标与文字解释, 支持在视频维度的伪造识别 + 定位 + 解释文本生成三重任务.

(6) FFA-VQA [76] 数据集是一个面向开放世界人脸伪造分析任务的伪造视觉问答数据集. 该数据集聚合了 FF++ [105], DFDC [108], Celeb-DF [109] 等七个公开伪造数据集的图像样本, 共包含 7436 个图像样本, 并为每个图像提供注释, 引入 “真假假设提示” (如 “假设该图像为真实/伪造, 解释其依据”) 作为文本输入.

(7) VLForgery [85] 数据集是一个面向扩散模型人脸伪造检测的三任务多模态数据集, 共包含来自 12 个子集的 500000+ 张图像, 每个样本包含图像 (真实/伪造)、问题 (由 ChatGPT-4o 自动生成)、答案 (包含检测结果、伪造类型、伪造区域、伪造方法等结构化文本) 和 EkCot 推理链条描述 (融合视觉线索与伪造生成知识).

(8) FakeClue [82] 数据集是一个面向多模态伪造检测与伪造线索解释任务的大规模数据集, 共包含超过 100000 张图像, 涵盖动物、人类、物体、自然风景、卫星图像、文档和 Deepfake 图像等七大类. 该数据集采用多模型联合标注策略 (如 Qwen2-VL [30], InternVL [29], Deepseek [31] 等), 为每张图像生成细粒度的自然语言描述, 标注图像中可能存在的伪造伪迹 (artifact clues), 包括纹理异常、几何结构扭曲、光照不一致等, 是一个覆盖类别广、注释粒度细的伪造图像解释数据集.

(9) MS-UFAD [55] 数据集是一个大规模、面向真实场景的人脸攻击检测数据集, 首次为每个样本提供了细粒度的文本描述. 该数据集共包含来自 5000 名个体的 795000 个视频样本和 60000 张图像样本, 涵盖 52 种攻击类型, 包括物理攻击、对抗攻击与深度伪造. 文本注释由多模态大语言模型 miniCPM [110] 半自动生成, 内容涵盖样本属性 (如年龄、性别、光照条件) 以及伪造类型与线索信息, 为文本引导的伪造检测模型提供了有力的数据支持.

(10) MMTT [38] 是提供伪造图像、像素级掩码和人工撰写解释的大规模人脸伪造数据集, 涵盖三种主流伪造技术 (GAN, Transformer, Diffusion), 总共包含 128303 条图文样本, 伪造涉及的面部区域多达 21 类, 单图最多含 11 个被改区域, 平均每条解释为 26.9 个词.

4.3　含文本提示生成的多模态数据集

此类数据集强调文本作为控制模态参与伪造图像的生成过程, 通常通过文本描述 (prompt) 引导扩散模型、生成模型或图像编辑器合成具有语义伪造性质的图像. 相较于传统图像篡改方式, 这类数据集更贴合当前大模型主导的 AIGC 生成机制, 可有效模拟语义驱动伪造、局部内容操控、风格 – 语义错配等复杂伪造情景, 适用于评估模型对语义控制下图像真实性的判断与解释能力.

表 3　现有方法在 DGM4 数据集上的性能对比.

Table 3　Performance comparison of existing methods on the DGM4 dataset.

Method	Category	Binary classification			Multi-Label classification			Image grounding			Text grounding		
		AUC ↑	EER ↓	ACC ↑	mAP ↑	CF1 ↑	OF1 ↑	IoUmean ↑	IoU50 ↑	IoU75 ↑	Precision ↑	Recall ↑	F1 ↑
CLIP [2]	Baseline	83.22	24.61	76.40	66.00	59.52	62.31	49.51	50.03	38.79	58.12	22.11	32.03
VILT [2]	Baseline	85.16	22.88	78.38	72.37	66.14	66.00	59.32	65.18	48.10	66.48	49.88	57.00
CrUr (CLIP) [60]	CLIP	84.93	22.61	78.52	73.42	67.25	67.24	52.03	57.38	54.53	64.82	44.38	52.72
HAMMER [2]	ViT	93.19	14.10	86.39	86.22	79.37	80.37	76.45	83.75	76.06	75.01	68.02	71.35
Exploiting [42]	ViT	95.11	11.36	88.75	91.42	83.60	84.38	80.83	88.35	80.39	76.51	70.61	73.44
VLP-GF [41]	ViT	92.84	14.45	86.13	85.65	80.02	79.07	76.73	83.89	76.24	76.42	66.80	71.29
ViKI [39]	ViT	93.51	13.87	86.67	86.58	81.07	80.10	76.51	83.95	75.77	77.79	66.06	72.44
IDseq [43]	ViT	94.55	11.40	88.94	90.01	83.00	84.90	83.33	89.39	86.19	75.96	71.23	73.52
Unleashing [37]	ViT	96.34	9.88	90.32	92.48	86.19	86.92	84.07	90.48	87.17	75.33	77.95	76.62

(1) COCOFake [57] 数据集是一个大规模文本图像伪造检测数据集, 基于 COCO [104] 的 113287 张真实图像与每图对应的 5 条文本描述, 通过 Stable Diffusion [111] v1.4 和 v2.0 分别生成约 1200000 张伪造图像 (每版约 600000), 形成 “1 张真实图像 + 5 张伪造图像” 的语义集群, 支持低级特征与语义线索的对比检测与伪造风格 – 语义解耦研究.

(2) AutoSplice [96] 数据集是一项聚焦于文本引导的局部伪造检测数据集, 该数据集通过引入语义掩码与自然语言提示词对真实图像进行区域级伪造生成, 包含 3621 张操控图像与 2273 张真实图像, 并配套像素级文本伪造掩码.

(3) DFLIP-3K [97] 包含约 300000 张由 3000 多个生成模型生成的深度伪造图像, 并配套约 190000 条原始生成提示词, 支持伪造检测、源模型识别与提示词预测. 该数据集构建了基于真实图像 (如 LAION-5B [112]) 与生成伪造图像的对比标注体系.

(4) DiFF [98] 数据集是基于扩散模型生成的人脸伪造图像的大规模数据集, 涵盖 13 种 SOTA 扩散方法, 覆盖 Text-to-Image, Image-to-Image, Face Swapping, Face Editing 四种伪造形式, 总共生成 537466 张伪造图像, 并提供了来自 1070 个身份、23661 张真实图像与 30000 条高质量文本和视觉 prompt, 支持细粒度检测、归因分析与边缘特征正则化研究.

5　性能分析与方法差异

在大模型时代, 随着视觉与语言模型的不断发展, 伪造检测技术逐渐多样化. 不同的模型方法在训练成本、结构设计和检测效果方面存在显著差异, 适用于不同的应用场景. 本节对各类伪造检测方法进行综合分析, 重点探讨基于视觉大模型、视觉 – 语言大模型和多模态大语言模型的方法的优缺点.

5.1　基于视觉大模型的伪造检测

基于视觉大模型的方法, 如 vision transformer (ViT) 和 swin transformer (SwinT), 最初专注于优化视觉伪造检测任务. 随着检测任务的深入, 这些方法逐步扩展到多模态任务. 尽管 ViT 和 SwinT 在视觉任务中表现出色, 但它们并未特别针对跨模态数据设计, 因此在面对包含文本和视觉信息的多模态数据时, 通常需要与文本编码器 (如 BERT 或 RoBERTa) 结合, 以增强跨模态处理能力. 这些方法特别适用于视觉信息占主导地位或者跨模态语义检测要求不高的伪造检测任务. 例如, 表 3 [2,37,39,41~43,60] 中的数据展示了现有研究在 DGM4 数据集 [2] 上的性能指标. 其中文献 [37,42] 等基于视觉 Transformer 的方法 (TM), 在 AUC (曲线下面积) 和 EER (等错误率) 等指标上均表现优秀, 特别是在 Image Grounding (图像定位) 和 Text Grounding (文本定位) 等任务上, 取得了较为稳定的效果.

表 4　各方法跨数据集测试 AUC 得分汇总.

Table 4　Summary of AUC scores for different methods in cross-dataset test.

Method	Category	Training-set	Test-set					
		FF++	CDF	DFDC	DFDCP	DF-1.0	WDF	DFD
RECCE (Xception) [113]	Baseline	99.32	68.71	69.06	–	74.10	64.31	–
SBI (EfficientNet) [114]	Baseline	99.64	93.18	72.42	86.15	77.70	–	97.56
Standing [54]	CLIP	–	80.00	77.34	90.57	–	85.42	–
M2F2-Det [59]	CLIP	99.34	95.10	87.80	–	–	87.20	97.70
MFCLIP [58]	CLIP	99.63	83.46	86.08	–	78.99	–	–
SJEDD [56]	CLIP	99.86	92.22	84.14	–	93.21	–	–
Towards [84]	MLLMs-H	99.16	83.15	–	83.21	–	85.10	94.81
LVLM-DFD [75]	MLLMs-E	99.53	94.71	79.12	91.81	78.99	–	99.64

5.2　基于视觉 – 语言大模型的伪造检测

其次, 基于视觉 – 语言大模型的方法, 如 CLIP 和 BLIP, 则在近年来发展较为成熟, 尤其是 CLIP 方法, 其通过联合学习图像和文本的嵌入空间, 能够在图像和文本之间进行语义对齐. 这使得 CLIP 在伪造检测任务中展示了出色的性能. 例如, 根据表 4 [54, 56, 58, 59, 75, 84, 113, 114] 中数据, CLIP 分类的方法在 FF++ [105] 上进行训练, 在其他伪造检测数据集 [105, 107~109, 115, 116] 上测试, 获得了良好的泛化性能. 此外, CLIP 方法的优势还体现在其灵活性和扩展性上, 许多研究已将其应用于视频域 [52, 53]、伪造定位 [60] 和扩散模型生成的检测 [57] 等更为复杂的任务. 在这些任务中, CLIP 能够通过对比学习方法有效地检测伪造图像, 并在图像与文本一致性方面展现了突出的能力. 因此, 基于视觉 – 语言大模型的方法, 已经成为当前伪造检测领域的重要方向.

5.3　基于多模态大语言模型的伪造检测

最后, 基于 MLLMs 的方法发展较晚. 针对 MLLMs 在伪造检测任务中的应用评估, 如表 5 [67, 70, 105] 所示 (测试数据来源自文献 [7]), 这些模型在零样本 (zero-shot learning) 测试过程中, 即在没有接触过特定训练数据的情况下, 直接将模型迁移到伪造检测任务上进行测试, 部分模型已经接近甚至超过人类平均水平, 且远高于 CLIP 和 BLIP 等模型, 这展现出 MLLMs 在伪造检测任务中广阔的前景, 尤其是在推理能力、语义理解和记忆方面. 然而, 需要特别关注的是, 许多方法将多模态大语言模型定位为 “解释性分析” 工具, 而非单纯依赖其提升检测精度. 这反映了这些多模态大语言模型在科研领域尚未成熟的现状, 如大模型固有的 “幻觉” 问题, 尤其在高风险应用中可能导致错误判断, 训练成本高, 推理速度较慢, 且需要大量的计算资源. 由于这些问题的存在, MLLMs 在伪造检测领域的可靠性和应用仍受到一定制约.

总体来看, 基于 TM 大模型和基于 VLM 大模型的方法发展较为成熟, 部分研究之间有一致的评估体系和相近的性能. 基于视觉大模型的方法以 ViT 模型为主, 通常倾向于结合语言模型进行模态融合创新, 并强调其在跨模态能力方面的优势. 基于 VLM 大模型的方法则以 CLIP 模型为主, 不仅能处理跨模态识别任务, 还能够应对更复杂的任务, 如扩散模型的检测、视频域检测等. 基于 MLLMs 的方法, 在自行构建数据集进行实验测试时有较高的研究比例, 且倾向于提出对伪造检测的新要求 (解释、问答、外部事实验证). 但对标传统的单一模态、二分类伪造检测任务进行比较, 研究较少. 具体的挑战和解决措施将在第 6 节进一步讨论.

6　挑战、趋势与前瞻

6.1　现有问题

近年来, 多模态大模型在伪造检测领域展现出令人瞩目的能力, 特别是在语义理解、跨模态推理与可解释性分析等方面取得了显著进展, 但其在当前研究中仍面临专用性不足、幻觉现象、公平性等问题.

表 5　各模型在伪造检测任务中的零样本实验效果.

Table 5　Zero-shot performance of different large models in forgery detection tasks.

Zero-shot test-set	Model	Accuracy (%)
FF++ [105]	LLaVA-1.5	54.90
	BLIP-2	49.04
	InstructBLIP	42.22
	CLIP	34.85
FakeBench [70]	GPT-4V	78.03
	Human Evaluation	74.51
	GeminiPro	67.50
	InstructBLIP	57.73
	LLaVA-1.5	57.70
	Qwen-VL	56.42
	Claude3 Sonnet	55.12
MMFakeBench [67]	GPT-4V	54.00
	Human Evaluation	37.90
	BLIP2	32.80
	PandaGPT	30.00
	Qwen-VL	11.00
	MiniGPT4	9.00

6.1.1　专用性不足问题

首先, 相较于专用的伪造检测器, 通用型 LMMs 在准确性、鲁棒性以及任务适配性方面仍显逊色. 例如, FKA-Owl [77] 等工作指出, 现有的 LMMs 在多模态伪造检测任务 (MFND) 中的表现并不总是令人满意. 一方面, 这些模型并非针对 MFND 任务量身定制; 另一方面, LMMs 对局部空间细节缺乏敏感性 [117]. 例如, 在人脸交换伪造场景中, 编辑区域与原始背景之间的微妙图像差异往往难以被大模型感知. 多项研究 [44, 59, 64] 显示, 在伪造检测任务中, 通用型多模态大模型的鉴伪性能可能低于传统的专用检测器, 并且面对不稳定因素时, 如输入图像扰动, 或者操控面部特征过小、伪造图像分辨率较高等复杂情况, 模型的表现有一定程度下降 [2, 64]. 其可能的原因如下.

通用大模型在处理特定视觉伪造任务时, 缺乏针对伪造检测的知识支持. 文献 [64] 指出, 现有的深度伪造检测数据集和标签体系仍显不足, 缺乏细粒度的操控标签, 这使得大模型在多标签和细粒度的伪造检测任务中难以充分发挥作用. 尽管已有多个多模态数据集被介绍, 但当前多模态伪造检测所使用的数据集仍显杂乱、不统一, 解决这一问题也正是本综述的目的所在.

除了数据方面的不足, 针对伪造痕迹的建模能力是目前多模态大模型面临的一大挑战. 文献 [44] 指出, “通用描述能力” 不等于 “伪造感知能力”. 在多模态任务中, 模态偏见 (模态竞争) 也可能是导致通用大模型专用性差的一个原因, 即模型可能过度依赖某一模态信息, 忽视另一模态. 例如, 在视觉问答、图文对齐等任务中, 模型可能只关注图像, 忽略问题文本, 或者只依赖语言常识, 无视图像中的细节, 从而生成错误或者解释不准确的回答 [118]. 尽管已有研究 [2, 69, 70] 尝试通过低级视觉对齐、掩码注入、提示词设置等手段提升表现, 其实际效果仍有较大提升空间.

为解决这些问题, 我们主张从大模型与传统取证方法在检测机理上的差异入手. 已有研究验证了该思路的有效性, 如文献 [49], 通过词频分析得出结论: CLIP 模型在进行伪造检测时并不具备固有的真实或伪造语义, 而是通过识别和匹配相似概念来执行伪造检测. 因此, 保留文本编码器并设计适合伪造检测的提示词成为一种可行的方向. 进一步地, 受网络深度影响, 多模态大模型对局部空间细节缺乏敏感性 [77, 117], 难以感知编辑临近区

域的纹理细微变化. 研究 [40, 52] 均指出, 大模型在其深层网络中进行多次降采样操作, 导致细微的伪造痕迹被忽视或消失. 改善大模型这种 “语义强敏、微痕弱敏” 的特性, 可使用的策略有残差连接、伪造纹理分支、边缘损失等, 也可以将不同的类别 (semantic 相同) 聚到一起, 再执行每个子类别的判别 (如假狗 vs. 真狗), 大大减小判别复杂度, 并提升模型的泛化性 [119].

6.1.2　幻觉问题

其次, 另一个突出的问题是幻觉现象, 幻觉是指大模型生成的内容与提供的事实不符 [92]. 文献 [83] 指出, 幻觉是大语言模型固有的特性. 在伪造检测任务中, 幻觉现象表现为模型错误地将真实样本判定为伪造和对检测结果的不准确解释. 例如, 针对口腔位置的篡改, 模型却将其错误地解释为鼻子的对称性问题 [84]. 这种现象不仅削弱了伪造检测的可信度, 还可能带来潜在的实际应用风险. 文献 [82, 84, 85] 等中方法, 通过引入局部伪造掩码、外部信息验证、视觉提示词、伪造知识预学习等策略, 增强模型能力, 这些方法有助于在一定程度上遏制大模型产生幻觉.

有趣的是, 与此相对的, 有一些研究揭示了为了解决幻觉问题而导致的模型保守策略的负面作用. 例如, 研究 [7] 发现, 当图像中包含多个操控区域 (如嘴巴、眼睛的操控) 时, 若模型未能准确识别这些区域, 则模型可能只会给出一个笼统的回答, 如 “该图像的面部区域已被修饰”. 研究 [64] 指出, 闭源大模型 (如 GPT-4o) 比开源模型 (如 LLaVA) 表现出相对较弱的性能, 这主要是因为商业化的大模型倾向以更保守的答案作出回应, 承认他们无法得出结论. 因此, 在提升模型能力的同时, 如何平衡模型的谨慎性与准确性, 仍然是解决幻觉问题的关键挑战之一.

6.1.3　公平性问题

公平性问题在大模型与伪造检测领域中都非常值得关注. 大模型本身广泛存在性别、种族、职业等偏见. 2024 年 *Nature* 子刊 *Humanities & Social Sciences Communications* 的一项研究 [120] 指出, 大语言模型在行为识别任务中普遍存在性别偏见. 文献 [121] 的数据表明, 性别反转后, 模型对正确职业的预测概率平均变化了 3.09%. 在偏见最严重的前 50 个职业类别中, 这一变化高达 9.25%. 文献 [122] 指出, 大模型生成的内群体句子比外群体句子更倾向于表现出积极的情感的可能性高出 93%, 且随着规模增大, 性别偏见也随之增加.

在伪造检测领域, 尤其是在面部伪造检测任务中, 不同性别、年龄和肤色群体的识别精度存在显著差异. 目前, 公开的深度伪造检测单模态数据集 (如 FF++, Celeb-DF 等) 主要集中于欧美人群, 这些数据集在文化背景、肤色和年龄等方面缺乏多样性 [123]. 然而, 遗憾的是, 目前主流研究在多模态数据集构建中, 通常以伪造类型、模态、任务和样本分类为导向, 尚未充分考虑模型输出的公平性问题, 长期忽视这一问题可能会削弱模型在真实世界应用中的公平性和可信度. 为应对这一问题, 伪造检测领域可以借鉴其他领域在数据集构建和公平性评估方面的经验. 例如, 在人脸识别和医学影像领域, 已有的工作 [124, 125] 强调了数据集的多样性, 并引入公平性约束以减少偏见. 虽然我们不打算在本文中深入探讨这些领域的细节, 但通过参考这些同样面临大模型引发的公平性问题的下游任务领域的解决思路, 或许能为提升伪造检测的公平性提供有益启示.

6.2　发展趋势与研究展望

随着生成式人工智能和多模态大模型的不断进步, 伪造检测领域面临着前所未有的机遇与挑战. 新的内容生成方式推动了伪造检测技术与应用场景的持续变革. 为此, 本文结合当前研究热点, 分析了未来的发展趋势, 并展望了下一代伪造检测系统的研究方向.

6.2.1　伪造信息复杂化

在以大模型为核心的生成内容浪潮中, 虚假信息不仅具备更强的语言风格拟合能力, 还呈现出高度拟人化的交互特征. 例如, 基于大模型驱动的社交代理人能够精准模仿特定社群的语言习惯, 通过情感渲染、语义模糊与持续对话, 生成更具欺骗性的虚假内容. 扩散模型与图文生成大模型的普及, 使得跨模态合成成为主流的内容

操控手段. 在这一语境下, 伪造图像和伪造文本往往 “各自真实但语义错配”, 例如, 图像中展示一场示威活动, 文本却描述其为庆典. 此类 “语境错配式伪造” 极具迷惑性, 而传统依赖局部特征或模态单一的检测方法难以识别其背后的深层语义矛盾. 文献 [2∼4, 68] 同时涵盖多伪造模态与跨语义操控的数据集相继出现, 也从侧面反映出在生成大模型迅猛发展之下, 虚假信息的表现形式、传播机制与干扰手段日益复杂, 对伪造检测提出了前所未有的挑战.

伪造信息的变化直接推动了伪造检测任务形态的演进 —— 从早期的基于单模态与二分类的简单判别, 逐步发展为需要跨模态推理、局部定位与自然语言解释的复杂任务形式. 其最直观的体现是, 检测数据集从最初的 “图像 + 真伪标签” [106, 108, 109], 发展为 “图像 – 文本对 + 一致性判定” [2, 67, 101], 再到 “开放式问答型任务样本” [44, 76]. 面对这一趋势, 伪造检测模型不仅要能够判断 “是否为伪造”, 更需要回答 “为何是伪造” 和 “伪造位置在哪里” 以及 “依据何种视觉与语言特征作出判断”. 这一多层次、跨模态、可解释的检测需求, 正成为未来伪造检测系统演进的关键方向, 也对多模态大语言模型的语义理解深度、模态协同能力与推理稳定性提出更高挑战.

6.2.2　伪造线索深度挖掘

使用相对较小模型进行伪造检测的方法已完全落后于时代发展? 事实并非如此. 已有研究 [76] 指出, 在特定任务场景下, 相较于通用型多模态大模型, 专用的伪造检测器在准确性、鲁棒性和领域适配能力方面仍具部分优势. LMMs 通常接受用于语义级别的视觉对齐的训练, 缺乏细粒度的法医感知能力 [85]. Forensics-Bench [68] 实验显示, 大模型对面部细节伪影、图像边界模糊、遮挡纹理等 “低级伪造痕迹” 的感知弱. 单模态检测方法并非被替代, 而是与多模态大模型形成互补融合的协同机制. 近年来, 越来越多研究探索如何将传统检测方法中的显式结构化伪造特征 (如频域特征、人脸特征、几何一致性) 嵌入多模态大模型的视觉输入中. 此外, 多模态大模型也促使伪造检测研究从人脸伪造检测、假新闻、图文事实核查等小领域的相对独立, 走向领域交叉融合 [45]. 因此, 未来伪造检测的研究重点不仅仅局限于如何在伪造检测任务中迁移更强大的通用模型, 更在于对多模态信息特征的提取和融合, 在纹理性、结构性伪造感知与语义一致性建模之间建立桥梁, 并且更新已有伪造检测技术与思路.

6.2.3　回归人类鉴伪思维

在深度学习出现之前, 传统取证方法, 即未使用深度学习方法的几何、光线分析, 更贴近人类 “观察 – 对照 – 解释” 的鉴伪思维, 但精度有限; 深度学习方法显著提升了准确率, 却在一定程度上牺牲了可解释性与可核查性. 基于此, 我们提出一个核心追问: LMMs 能否在保持性能的同时, 引导检测流程回归人类鉴伪思维? 一方面, VLLM 具备开放式回答与推理能力, 能够围绕图像内容生成灵活的自然语言结论与理由, 并将 “发现伪迹 – 建立因果 – 形成判断” 的链条外显 [7]. 文献 [44] 研究亦表明, 先解释再判定比仅输出 “真/假” 模板更有利于检测效果与跨域泛化. 另一方面, 我们也需要承认部分伪造线索可能超出人类直觉与分辨能力. 文献 [81] 报告显示, 普通人对真假脸的识别仅约 65%, 专家与熟练者约 75%. 因此, 更可取的路线是采用 “解释先行 – 证据对齐 – 符合审查” 的范式: 先产出可核查的证据与推理过程, 再给出结论; 对不确定或证据不足的情形触发特殊复核规则, 以在可解释性与准确率之间取得兼顾与提升.

6.2.4　对抗性攻击

伪造技术不断进步, 对抗性攻击的出现对现有检测方法构成了前所未有的挑战. 对抗性攻击被广泛认为是深度学习算法中的已知威胁之一. 攻击者通过精心设计的输入, 能够诱使深度学习系统作出错误判断, 从而达到其有利目的 [126]. 这类攻击利用生成对抗算法 (如对抗样本和数据增强等技术), 或通过人工剪辑压缩、图像压缩等手段掩盖伪造痕迹, 误导检测模型, 使其错误地将伪造内容判定为真实, 或者将真实内容误判为伪造.

目前, 基于多模态大模型 (如 GPT-4, CLIP 等) 的深度伪造检测方法, 尽管在多个领域取得了显著进展, 但它们在应对对抗性攻击时仍存在明显的局限性. 首先, 这些大模型往往依赖于大规模数据的训练过程. 如果研究

采用预训练模型权重, 那么在应用伪造检测任务之前, 这些模型的对抗性弱点就已经潜伏在训练数据中, 难以避免. 其次, 现有的 “对抗性噪声” 技术隐蔽性强、攻击成功率高 [127, 128]. 最后, 跨模式场景中的大模型伪造检测也面临着安全风险. 攻击者能够同时利用对抗性攻击对文本和视觉等多模态输入进行攻击, 从而进一步削弱系统的判别能力. 尽管已有大量的研究针对伪造检测中的对抗性攻击与防御展开, 通用大模型的对抗性攻击和防御方面也有不少探索, 但目前尚未看到多模态大模型的伪造检测方法与对抗性攻击和防御研究的交集. 这一领域存在着明显的研究空白. 许多研究已经在多个大模型上开展了伪造检测任务的零样本实验, 因此针对应用在伪造检测任务中的大模型方法进行统一的对抗性扰动攻击实验, 为未来的研究建立新的对抗性测试评估基准.

6.2.5　开放世界伪造检测

在现实环境中, 新型伪造类型不断涌现, 文献 [129] 提出了开放世界深度伪造检测的挑战: 在伪造信息的开放语境中, 生成意图多样、生成模型家族迭代快, 风格各异, 且包括整图合成、局部插入、擦除、背景替换等精细化操作, 导致伪造内容越来越难以识别. 这些不断演化的伪造内容, 对检测模型提出了更高的要求. 面对这一挑战, 增量学习是一种有效解决方案. 增量学习又称持续学习、终身学习, 指模型在接收到新数据时, 不需要重新训练, 而是逐步更新已有知识 [130]. 大模型的增量学习已有一定研究进展 [131, 132], 但这些研究大多集中于从通用任务到特定任务的迁移 (如从多模态自然语言处理到伪造检测), 而针对新增伪造手段进行增量学习, 应对伪造技术本身的动态演化的任务研究较少. 少数相关工作 (如文献 [133, 134]) 已尝试探讨这一问题, 但使用方法多基于传统卷积网络或 ViT 架构, 未能充分利用多模态大模型的优势, 且评估范围局限于图像层面. 传统模型参数小. 而在大模型伪造检测技术兴起背景下, 考虑到训练成本和不断升级的伪造手段, 预计伪造增量学习策略的重要性将进一步凸显.

7　结束语

随着多模态大模型技术的快速演进, 伪造检测正从单一场景、单一模态、二分类判断的早期范式, 转变为一个涉及语言理解、视觉识别、跨模态融合、因果推理与社会公平的综合挑战. 构建一个准确、安全、公平的伪造检测体系, 不应仅通过通用大模型的直接迁移来应对伪造检测挑战, 更需要在伪造检测问题定义、伪造线索挖掘、评估指标、技术对抗性等层面开展重构性思考与前瞻性布局.

参考文献

1　Erduran S. Deepfakes and students' deep learning: a harmonious pair in science. Science, 2024, 385: eadr8354
2　Shao R, Wu T, Liu Z. Detecting and grounding multi-modal media manipulation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023. 6904–6913
3　Xu Q, Chen H, Du H, et al. M3A: A multimodal misinformation dataset for media authenticity analysis. Comput Vision Image Understand, 2024, 249: 104205
4　Xu J, Chen J, Song X, et al. Identity-driven multimedia forgery detection via reference assistance. In: Proceedings of the 32nd ACM International Conference on Multimedia, 2024. 3887–3896
5　Nie F, Ni J, Zhang J, et al. Frade: forgery-aware audio-distilled multimodal learning for Deepfake detection. In: Proceedings of the 32nd ACM International Conference on Multimedia, 2024. 6297–6306
6　Jin Z, Cao J, Zhang Y, et al. Novel visual and statistical image features for microblogs news verification. IEEE Trans Multimedia, 2016, 19: 598–608
7　Foteinopoulou N M, Ghorbel E, Aouada D. A hitchhiker's guide to fine-grained face forgery detection using common sense reasoning. Adv Neural Inform Process Syst, 2024, 37: 2943–2976
8　Comito C, Caroprese L, Zumpano E. Multimodal fake news detection on social media: a survey of deep learning techniques. Soc Netw Anal Min, 2023, 13: 101
9　Mubarak R, Alsboui T, Alshaikh O, et al. A survey on the detection and impacts of Deepfakes in visual, audio, and textual formats. IEEE Access, 2023, 11: 144497
10　Croitoru F A, Hiji A I, Hondru V, et al. Deepfake media generation and detection in the generative AI era: a survey and outlook. ArXiv:2411.19537

11 Liu P, Tao Q, Zhou J T. Evolving from single-modal to multi-modal facial Deepfake detection: progress and challenges. ArXiv:2406.06965

12 Pu J, Sarwar Z, Abdullah S M, et al. Deepfake text detection: limitations and opportunities. In: Proceedings of the 2023 IEEE Symposium on Security and Privacy (SP), 2023. 1613–1630

13 Adithya B, Ranjith J, Poojitha J, et al. A survey on detection of Deepfake text and sentiment analysis using machine learning models. In: Sentiment Analysis Unveiled. Boca Raton: CRC Press, 2025. 83–107

14 Shi Y, Gao Y, Lai Y, et al. SHIELD: an evaluation benchmark for face spoofing and forgery detection with multimodal large language models. Vis Intell, 2025, 3: 9

15 Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16×16 words: transformers for image recognition at scale. ArXiv:2010.11929

16 Liu Z, Lin Y, Cao Y, et al. Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021. 10012–10022

17 Radford A, Narasimhan K, Salimans T, et al. Improving language understanding by generative pre-training. 2018. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf

18 Li J, Li D, Savarese S, et al. Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In: Proceedings of International Conference on Machine Learning, 2023. 19730–19742

19 Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision. In: Proceedings of International Conference on Machine Learning, 2021. 8748–8763

20 Ramesh A, Dhariwal P, Nichol A, et al. Hierarchical text-conditional image generation with clip latents. ArXiv:2204.06125

21 Achiam J, Adler S, Agarwal S, et al. GPT-4 technical report. ArXiv:2303.08774

22 Hurst A, Lerer A, Goucher A P, et al. GPT-4o system card. ArXiv:2410.21276

23 Zhu D, Chen J, Shen X, et al. Minigpt-4: enhancing vision-language understanding with advanced large language models. ArXiv:2304.10592

24 Chiang W L, Li Z, Lin Z, et al. Vicuna: An open-source chatbot impressing GPT-4 with 90%* ChatGPT quality. 2023. https://vicuna.lmsys.org

25 Liu H, Li C, Wu Q, et al. Visual instruction tuning. Adv Neural Inform Process Syst, 2023, 36: 34892–34916

26 Liu H, Li C, Li Y, et al. Improved baselines with visual instruction tuning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024. 26296–26306

27 Team G, Anil R, Borgeaud S, et al. Gemini: a family of highly capable multimodal models. ArXiv:2312.11805

28 Meta A. The Llama 4 herd: the beginning of a new era of natively multimodal AI innovation. 2025. https://ai.meta.com/blog/llama-4-multimodal-intelligence/

29 Chen Z, Wu J, Wang W, et al. Interval: scaling up vision foundation models and aligning for generic visual-linguistic tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024. 24185–24198

30 Wang P, Bai S, Tan S, et al. Qwen2-Vl: enhancing vision-language model's perception of the world at any resolution. ArXiv:2409.12191

31 Liu A, Feng B, Xue B, et al. DeepSeek-V3 technical report. ArXiv:2412.19437

32 Yan Z, Zhang Y, Yuan X, et al. Deepfakebench: a comprehensive benchmark of Deepfake detection. In: Proceedings of Advances in Neural Information Processing Systems, 2023. 4534–4565.

33 Afchar D, Nozick V, Yamagishi J, et al. Mesonet: a compact facial video forgery detection network. In: Proceedings of the 2018 IEEE International Workshop on Information Forensics and Security (WIFS), 2018. 1–7

34 Corvi R, Cozzolino D, Poggi G, et al. Intriguing properties of synthetic images: from generative adversarial networks to diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023. 973–982

35 Peng C, Miao Z, Liu D, et al. Where Deepfakes gaze at? Spatial-temporal gaze inconsistency analysis for video face forgery detection. IEEE Trans Inform Forensics Security, 2024, 19: 4507–4517

36 Liu D, Chen T, Peng C, et al. Attention consistency refined masked frequency forgery representation for generalizing face forgery detection. IEEE Trans Inform Forensics Security, 2025, 25: 504–515

37 Li Y, Yang Y, Tan Z, et al. Unleashing the potential of consistency learning for detecting and grounding multi-modal media manipulation. In: Proceedings of the Computer Vision and Pattern Recognition Conference, 2025. 9242–9252

38 Lian J, Liu L, Wang Y, et al. A large-scale interpretable multi-modality benchmark for facial image forgery localization. ArXiv:2412.19685

39 Li Q, Gao M, Zhang G, et al. Towards multimodal disinformation detection by vision-language knowledge interaction. Inf Fusion, 2024, 102: 102037

40 Yoon J H, Panizo-LLedot A, Camacho D, et al. Triple-modality interaction for Deepfake detection on zero-shot identity. Inf Fusion, 2024, 109: 102424

41 Zhang G, Gao M, Li Q, et al. Multi-modal generative Deepfake detection via visual-language pretraining with gate fusion for cognitive computation. Cogn Comput, 2024, 16: 2953–2966

42 Wang J, Liu B, Miao C, et al. Exploiting modality-specific features for multi-modal manipulation detection and grounding. In: Proceedings of the ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2024. 4935–4939

43 Liu R, Xie T, Li J, et al. IDseq: decoupled and sequentially detecting and grounding multi-modal media manipulation. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2025. 496–504

44 Zhang Y, Colman B, Guo X, et al. Common sense reasoning for Deepfake detection. In: Proceedings of the European Conference on Computer Vision, 2024. 399–415

45 Jing J, Wu H, Sun J, et al. Multimodal fake news detection via progressive fusion networks. Inform Process Manag, 2023, 60: 103120

46 Qu C, Liu J, Chen H, et al. Explainable tampered text detection via multimodal large models. ArXiv:2412.14816

47 Sun H, Cai C, Zhuang H, et al. Edvd-llama: explainable Deepfake video detection via multimodal large language model reasoning. ArXiv:2510.16442

48 Khan S A, Dang-Nguyen D T. Clipping the deception: adapting vision-language models for universal Deepfake detection. In: Proceedings of the 2024 International Conference on Multimedia Retrieval, 2024. 1006–1015

49 Tan C, Tao R, Liu H, et al. C2P-CLIP: injecting category common prompt in clip to enhance generalization in Deepfake detection. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2025. 7184–7192

50 Yu X, Sheng Z, Lu W, et al. Racmc: residual-aware compensation network with multi-granularity constraints for fake news detection. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2025. 986–994

51 Liu Y, Wang F, Li K, et al. Training-free multimodal Deepfake detection via graph reasoning. ArXiv:2509.21774

52 Song X, Guo X, Zhang J, et al. On learning multi-modal forgery representation for diffusion generated video detection. Adv Neural Inform Process Syst, 2024, 37: 122054–122077

53 Liu F, Yacoob Y, Shrivastava A. COVID-VTS: fact extraction and verification on short video platforms. ArXiv:2302.07919

54 Lin K, Lin Y, Li W, et al. Standing on the shoulders of giants: reprogramming visual-language model for general Deepfake detection. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2025. 5262–5270

55 Jiang N, Zeng D, Gao L, et al. MS-UFAD: a large-scale dataset for real-world unified face attack detection with text descriptions. In: Proceedings of the ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2025. 1–5

56 Zou M, Yu B, Zhan Y, et al. Semantics-oriented multitask learning for Deepfake detection: a joint embedding approach. IEEE Trans Circuits Syst Video Tech, 2025, 35: 9950–9963

57 Amoroso R, Morelli D, Cornia M, et al. Parents and children: distinguishing multimodal Deepfakes from natural images. ACM Trans Multimed Comput Commun Appl, 2024, 21: 1–23

58 Zhang Y, Wang T, Yu Z, et al. MFCLIP: multi-modal fine-grained clip for generalizable diffusion face forgery detection. IEEE Trans Inform Forensics Security, 2025, 20: 5888–5903

59 Guo X, Song X, Zhang Y, et al. Rethinking vision-language model in face forensics: multi-modal interpretable forged face detector. In: Proceedings of the Computer Vision and Pattern Recognition Conference, 2025. 105–116

60 Zhao W, Lu Y, Jiao G, et al. Concentrated reasoning and unified reconstruction for multi-modal media manipulation. In: Proceedings of the ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2024. 8190–8194

61 Lai Y, Yu Z, Yang J, et al. GM-DF: generalized multi-scenario Deepfake detection. ArXiv:2406.20078

62 Sha Z, Tan Y, Li M, et al. Zerofake: zero-shot detection of fake images generated and edited by text-to-image generation models. In: Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security, 2024. 4852–4866

63 Singh G, Amarsinghe S, Thani U, et al. Sgs: segmentation-guided scoring for global scene inconsistencies. ArXiv:2509.26039

64 Jia S, Lyu R, Zhao K, et al. Can ChatGPT detect Deepfakes? A study of using multimodal large language models for media forensics. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024. 4324–4333

65 Al-Janabi O M, Alyasiri O M, Jebur E A. GPT-4 versus bard and bing: LLMs for fake image detection. In: Proceedings of the 2023 3rd International Conference on Intelligent Cybernetics Technology & Applications (ICICyTA), 2023. 249–254

66 Lin G, Lin L, Walker C P, et al. Fit for purpose? Deepfake detection in the real world. ArXiv:2510.16556

67 Liu X, Li Z, Li P, et al. Mmfakebench: a mixed-source multimodal misinformation detection benchmark for LVLMS. In: Proceedings of the International Conference on Learning Representations, 2025

68 Wang J, Lv C, Li X, et al. Forensics-bench: a comprehensive forgery detection benchmark suite for large vision language models. 2025. https://arxiv.org/abs/2503.15024

69 Liu J, Zhang F, Zhu J, et al. ForgeryGPT: multimodal large language model for explainable image forgery detection and

localization. ArXiv:2410.10238
70 Li Y, Liu X, Wang X, et al. FakeBench: probing explainable fake image detection via large multimodal models. IEEE Trans Inform Forensics Security, 2025, 20: 8730–8745
71 Tan H, Lan J, Tan Z, et al. Veritas: generalizable Deepfake detection via pattern-aware reasoning. ArXiv:2508.21048
72 Xu Z, Zhang X, Li R, et al. Fakeshield: explainable image forgery detection and localization via multi-modal large language models. In: Proceedings of International Conference on Learning Representations, 2025
73 Jin R, Fu R, Wen Z, et al. Fake news detection and manipulation reasoning via large vision-language models. ArXiv: 2407.02042
74 Guo H, Ma Z, Zeng Z, et al. Each fake news is fake in its own way: an attribution multi-granularity benchmark for multimodal fake news detection. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2025. 228–236
75 Yu P, Fei J, Gao H, et al. Unlocking the capabilities of vision-language models for generalizable and explainable Deepfake detection. ArXiv:2503.14853
76 Huang Z, Xia B, Lin Z, et al. FFAA: multimodal large language model based explainable open-world face forgery analysis assistant. ArXiv:2408.10072
77 Liu X, Li P, Huang H, et al. FKA-Owl: advancing multimodal fake news detection through knowledge-augmented LVLMS. In: Proceedings of the 32nd ACM International Conference on Multimedia, 2024. 10154–10163
78 Zeng F, Li W, Gao W, et al. Multimodal misinformation detection by learning from synthetic data with multimodal LLMs. ArXiv:2409.19656
79 Tan C, Ming X, Wang J, et al. Semantic visual anomaly detection and reasoning in AI-generated images. ArXiv:2510.10231
80 Papadopoulos S I, Koutlis C, Papadopoulos S, et al. Synthetic misinformers: generating and combating multimodal misinformation. In: Proceedings of the 2nd ACM International Workshop on Multimedia AI against Disinformation, 2023. 36–44
81 Qi P, Yan Z, Hsu W, et al. Sniffer: multimodal large language model for explainable out-of-context misinformation detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024. 13052–13062
82 Wen S, Ye J, Feng P, et al. Spot the fake: large multimodal model-based synthetic image detection with artifact explanation. ArXiv:2503.14905
83 Braun T, Rothermel M, Rohrbach M, et al. DEFAME: dynamic evidence-based fact-checking with multimodal experts. In: Proceedings of the 42nd International Conference on Machine Learning, 2025
84 Sun K, Chen S, Yao T, et al. Towards general visual-linguistic face forgery detection. In: Proceedings of the Computer Vision and Pattern Recognition Conference, 2025. 19576–19586
85 He X, Zhou Y, Fan B, et al. Vlforgery face triad: detection, localization and attribution via multimodal large language models. ArXiv:2503.06142
86 Devlin J, Chang M W, Lee K, et al. Bert: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1, 2019. 4171–4186
87 Liu Y, Ott M, Goyal N, et al. ROBERTa: a robustly optimized bert pretraining approach. ArXiv:1907.11692
88 Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space. ArXiv:1301.3781
89 Davis S, Mermelstein P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans Acoust Speech Signal Process, 1980, 28: 357–366
90 Zhang D, Yu Y, Dong J, et al. MM-LLMs: recent advances in multimodal large language models. ArXiv:2401.13601
91 Girdhar R, El-Nouby A, Liu Z, et al. Imagebind: one embedding space to bind them all. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023. 15180–15190
92 Huang L, Yu W, Ma W, et al. A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions. ACM Trans Inform Syst, 2025, 43: 1–55
93 Huang Z, Hu J, Li X, et al. Sida: social media image Deepfake detection, localization and explanation with large multimodal model. In: Proceedings of the Computer Vision and Pattern Recognition Conference, 2025. 28831–28841
94 Ye J, Zhou B, Huang Z, et al. LOKI: a comprehensive synthetic data detection benchmark using large multimodal models. In: Proceedings of International Conference on Learning Representations, 2025
95 Hondru V, Hogea E, Onchis D, et al. Exddv: a new dataset for explainable Deepfake detection in video. ArXiv:2503.14421
96 Jia S, Huang M, Zhou Z, et al. Autosplice: a text-prompt manipulated image dataset for media forensics. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023. 893–903
97 Wang Y, Huang Z, Ma Z, et al. Linguistic profiling of Deepfakes: an open database for next-generation Deepfake detection. ArXiv:2401.02335
98 Cheng H, Guo Y, Wang T, et al. Diffusion facial forgery detection. In: Proceedings of the 32nd ACM International Conference on Multimedia, 2024. 5939–5948

99 Liu F, Wang Y, Wang T, et al. Visual news: benchmark and challenges in news image captioning. ArXiv:2010.03743

100 Singh G, Amarsinghe S, Singh P, et al. DGM4+: dataset extension for global scene inconsistency. ArXiv:2509.26047

101 Luo G, Darrell T, Rohrbach A. Newsclippings: automatic generation of out-of-context multimodal media. ArXiv:2104. 05893

102 Zhu Y, Wang Y, Yu Z. Multimodal fake news detection: MFND dataset and shallow-deep multitask learning. ArXiv:2505.06796

103 Plummer B A, Wang L, Cervantes C M, et al. Flickr30k entities: collecting region-to-phrase correspondences for richer image-to-sentence models. In: Proceedings of the IEEE International Conference on Computer Vision, 2015. 2641–2649

104 Lin T Y, Maire M, Belongie S, et al. Microsoft COCO: common objects in context. In: Computer Vision–ECCV 2014. Lecture Notes in Computer Science, vol. 8693. Cham: Springer, 2014. 740–755

105 Rossler A, Cozzolino D, Verdoliva L, et al. Faceforensics++: learning to detect manipulated facial images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019. 1–11

106 Rössler A, Cozzolino D, Verdoliva L, et al. Faceforensics: a large-scale video dataset for forgery detection in human faces. ArXiv:1803.09179

107 Jiang L, Li R, Wu W, et al. Deeperforensics-1.0: a large-scale dataset for real-world face forgery detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020. 2889–2898

108 Dolhansky B, Bitton J, Pflaum B, et al. The Deepfake detection challenge (DFDC) dataset. ArXiv:2006.07397

109 Li Y, Yang X, Sun P, et al. Celeb-df: a large-scale challenging dataset for Deepfake forensics. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020. 3207–3216

110 Team M, Xiao C, Li Y, et al. MiniCPM4: ultra-efficient LLMs on end devices. ArXiv:2506.07900

111 Rombach R, Blattmann A, Lorenz D, et al. High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022. 10684–10695

112 Schuhmann C, Beaumont R, Vencu R, et al. Laion-5b: an open large-scale dataset for training next generation image-text models. Adv Neural Inform Process Syst, 2022, 35: 25278–25294

113 Cao J, Ma C, Yao T, et al. End-to-end reconstruction-classification learning for face forgery detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022. 4113–4122

114 Shiohara K, Yamasaki T. Detecting Deepfakes with self-blended images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022. 18720–18729

115 Zi B, Chang M, Chen J, et al. WildDeepfake: a challenging real-world dataset for Deepfake detection. In: Proceedings of the 28th ACM International Conference on Multimedia, 2020. 2382–2390

116 Dufour N, Gully A, Karlsson P, et al. Deepfakes detection dataset by Google & jigsaw. 2019. https://www.kaggle.com/datasets/sanikatiwarekar/deep-fake-detection-dfd-entire-original-dataset

117 Yuan Y, Li W, Liu J, et al. Osprey: pixel understanding with visual instruction tuning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024. 28202–28211

118 Zhang Z, Tang H, Sheng J, et al. Debiasing multimodal large language models via noise-aware preference optimization. In: Proceedings of the Computer Vision and Pattern Recognition Conference, 2025. 9423–9433

119 Yan Z, Wang J, Jin P, et al. Orthogonal subspace decomposition for generalizable AI-generated image detection. ArXiv: 2411.15633

120 Wu J, Song Y, Wu D C. Does ChatGPT show gender bias in behavior detection? Hum Soc Sci Commun, 2024, 11: 1–8

121 Xiao Y, Liu A, Cheng Q, et al. Genderbias-VL: benchmarking gender bias in vision-language models via counterfactual probing. ArXiv:2407.00600

122 Hu T, Kyrychenko Y, Rathje S, et al. Generative language models exhibit social identity biases. Nat Comput Sci, 2025, 5: 65–75

123 Ju Y, Hu S, Jia S, et al. Improving fairness in Deepfake detection. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2024. 4655–4665

124 Karkkainen K, Joo J. Fairface: face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021. 1548–1558

125 Tian Y, Shi M, Luo Y, et al. Fairseg: a large-scale medical image segmentation dataset for fairness learning using segment anything model with fair error-bound scaling. ArXiv:2311.02189

126 Shayegani E, Mamun M A A, Fu Y, et al. Survey of vulnerabilities in large language models revealed by adversarial attacks. ArXiv:2310.10844

127 Peng R, Tan S, Mo X, et al. Active adversarial noise suppression for image forgery localization. ArXiv:2506.12871

128 Meng X, Wang L, Guo S, et al. AVA: inconspicuous attribute variation-based adversarial attack bypassing Deepfake detection. In: Proceedings of the 2024 IEEE Symposium on Security and Privacy (SP), 2024. 74–90

129 Wang Y, Huang Z, Hong X. Opensdi: spotting diffusion-generated images in the open world. In: Proceedings of the Computer Vision and Pattern Recognition Conference, 2025. 4291–4301

130 Zhou D W, Wang Q W, Qi Z H, et al. Class-incremental learning: a survey. IEEE Trans Pattern Anal Mach Intell, 2024. 46:

9851–9873
131 Shi H, Xu Z, Wang H, et al. Continual learning of large language models: a comprehensive survey. ACM Comput Surveys, 2024, 58: 1–42
132 Zheng J, Shi C, Cai X, et al. Lifelong learning of large language model based agents: a roadmap. ArXiv:2501.07278
133 Li C, Huang Z, Paudel D P, et al. A continual Deepfake detection benchmark: dataset, methods, and essentials. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023. 1339–1349
134 Wang Y, Huang Z, Hong X. Benchmarking deepart detection. ArXiv:2302.14475

Deepfake detection in the era of large models

Chunlei PENG[1,2], Junye LI[1,2], Decheng LIU[1,2], Nannan WANG[1*], Ruimin HU[3] & Xinbo GAO[1]

1. *State Key Laboratory of Integrated Services Networks, Xidian University, Xi'an 710071, China*
2. *School of Cyber Engineering, Xidian University, Xi'an 710071, China*
3. *Hangzhou Institute of Technology, Xidian University, Hangzhou 311231, China*
* Corresponding author. E-mail: nnwang@xidian.edu.cn

Abstract With the continuous advancement of artificial intelligence, Deepfakes have evolved from single-modality synthesis into complex generative forms involving visual, auditory, and textual media. The emergence of large multimodal models (LMMs) has significantly enhanced the capability to generate forged content, while simultaneously bringing unprecedented opportunities and challenges to the task of forgery detection. This paper presents a comprehensive review of recent progress and technological evolution in forgery detection under the background of large models. It surveys relevant research outcomes over the past three years and summarizes recent multimodal forgery detection datasets. On this basis, we conduct an in-depth analysis of the potential and challenges of LMMs in terms of detection performance, hallucination, judgment accuracy, and fairness. We analyze the underlying causes and propose future solutions to address these issues. Finally, the paper explores future trends in forgery detection technologies, including the increasing complexity of forgery information, the value of traditional techniques, explainability, and technological adversarial dynamics.

Keywords large multimodal models, Deepfake detection, vision-language fusion, explainability in detection, cross-modal reasoning