E-FreeM^2: Efficient Training-自由 Multi-Scale 和跨模态新闻验证通过 M 大型语言模型

Van-Hoang Phan*
University of Science
Ho Chi Minh City, Vietnam
Vietnam National University
Ho Chi Minh City, Vietnam
21120459@student.hcmus.edu.vn

Long-Khanh Pham*
University of Science
Ho Chi Minh City, Vietnam
Vietnam National University
Ho Chi Minh City, Vietnam
FPT Software AI Center
Vietnam
21120479@student.hcmus.edu.vn

Dang Vu*
University of Science
Ho Chi Minh City, Vietnam
Vietnam National University
Ho Chi Minh City, Vietnam
FPT Software AI Center
Vietnam
23C15023@student.hcmus.edu.vn
dangvqm@fpt.com

Anh-Duy Tran
DistriNet, KU Leuven
Leuven, Belgium
anh-duy.tran@kuleuven.be

Minh-Son Dao
National Institute of
Information and
Communications Technology
Tokyo, Japan
dao@nict.go.jp

摘要

虚假信息在移动和无线网络中的迅速传播带来了关键的安全挑战。本研究介绍了一种无需训练的、基于检索的多模态事实验证系统，该系统利用预训练的视觉-语言模型和大型语言模型进行可信度评估。通过动态检索并交叉引用受信任的数据源，我们的方法减轻了传统基于训练模型（如对抗攻击和数据中毒）的脆弱性。此外，其轻量级设计使无缝集成边缘设备成为可能，而无需大量的设备端处理。在两个事实核查基准上的实验实现了 SOTA 结果，确认了它在虚假信息检测中的有效性及其对各种攻击向量的强大抵抗力，突显了它增强移动和无线通信环境安全的潜力。

*These authors contributed equally to this research.

SCID '25, Hanoi, Vietnam

ACM ISBN 979-8-4007-1417-7/2025/08
https://doi.org/10.1145/3709020.3734834

CCS Concepts

• **Security and privacy → Social aspects of security and privacy**; • **Computing methodologies → Knowledge representation and reasoning**; *Natural language processing*; *Multi-agent systems*.

Keywords

无训练、跨模态、检索系统、大型语言模型、提示工程

ACM Reference Format:

Van-Hoang Phan, Long-Khanh Pham, Dang Vu, Anh-Duy Tran, and Minh-Son Dao. 2025. E-FreeM^2: Efficient Training-自由 M ulti-Scale 和跨模态新闻验证通过 M 大型语言模型. In *The 2nd Workshop on Security-Centric Strategies for Combating Information Disorder (SCID '25), August 25–29, 2025, Hanoi, Vietnam.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3709020.3734834

1 介绍

近年来，媒体操纵技术（如深度造假）因其特质及其在传播虚假信息方面的作用而引起了广泛关注 [26, 28, 31, 39]。这些由人工智能进步推动的技术，包括视频、图像和语音合成，引发了人们对利用它们生成假新闻并促进金融欺诈的潜在滥用的担忧，从而促使监管机构和行业利益相关者采取措施以减轻其影响。然而，新闻核实所面临的最普遍但研究不足的一个挑战是脱离上下文（OOC）的错误信息，即真实且未被篡改的图像被重新用于带有误导性或虚假叙述的情境中 [5]。例如，在最近的以色列哈马斯冲突期间，社交媒体平台上出现了大量脱离上下文的错误信息案例，其中来自无关冲突的老照片甚至视频游戏画面被错误地呈现为当前事件 [25]。与操纵过的媒体不同，OOC 错误信息难以检测，因为视觉内容仍然是真实的，欺骗仅来源于伴随文本提供的误导性背景。

应对这一“信息流行病”需要多方面的策略，事实核查和揭穿虚假声明 [12, 15] 在此过程中起着关键作用。无意分享的错误信息和出于政治动机故意传播以欺骗他人的虚假信息都在数字生态系统中 [4] 得到了蓬勃发展。像 TikTok 这样的社交媒体平台已经成为了虚假信息传播的重要途径，特别是在年轻受众中，通过 Deepfakes（深度伪造）和不太复杂的 “Cheapfakes”（廉价伪造）等格式 [2]。随着这些平台扩大了错误叙述的覆盖范围，开发强大的新闻核查技术变得至关重要，以确保信息的完整性。本研究旨在解决检测和缓解 OOC 虚假信息及操纵媒体的关键挑战，利用人工智能的进步来提高新闻核查系统的准确性和可靠性。

最近在情境外（OOC）检测方面的进展——该领域主要关注确定给定声明中的真实图像是否被准确表示——越来越多地依赖于通过自定义搜索引擎如 Google API[17, 36, 37] 检索外部信息。这些方法通过整合额外的情境证据显示出了显著的改进。相比之下，早期缺乏外部数据集成 [3, 14, 18, 30] 的方法仅依赖于图像-声明对的内在特征，这限制了它们在验证声明方面的有效性。专业事实核查人员通常会评估图像的内部特征和外部情境信息以确定图像-声明关系的有效性。然而，有效地利用检索到的证据仍然是 OOC 检测中的一个重大挑战，特别是在识别图像与相关声明之间微妙不一致方面。为了解决这一问题，最近的研究 [1, 18, 21, 27] 已经探讨了将检索到的证据整合进分类器、预训练模型和大型视觉语言模型（LVLMs）中。虽然这些改进提高了准确性，但它们往往以增加架构复杂性和计算开销为代价。

若干先前的研究 [1, 18, 27, 29] 在多模态任务的事实验证方面做出了重要贡献。通过利用预训练的知识、推理能力和生成能力，这些方法能够有效地检测图像文本对中的事实不一致性，并生成连贯的自然语言解释。在此基础上，近期的工作 [21] 通过在应用指令调优进行训练之前，使用基于工具的方法检索外部信息来增强输入数据，从而改进了事实核查模型。然而，在策划的事实核查数据集上训练用于检测 OOC 虚假信息的模型面临多个挑战。依赖固定数据集的主要缺点是它们无法跟上不断演变的虚假信息趋势。由于这些数据集必须经过收集、预处理和发布才能使用，因此它们很快就会过时。此外，它们容易受到静态数据中毒攻击的影响，降低了其适应性、安全性和实时事实核查的有效性。另一个关键挑战是训练或调整多模态大型语言模型（MLLMs）的高计算成本。此外，

为了对抗快速演变的虚假信息保持有效性，需要频繁重新训练，这进一步增加了计算和存储开销。

为应对这些挑战，我们引入了 E-FreeM^2，这是一个专为实时 OOC 虚假信息检测设计的多模态大型语言模型（MLLM），具有高准确性。我们的方法通过利用跨模态、多尺度框架来检索和生成高质量候选信息，从而增强验证过程。E-FreeM^2 首先进行外部搜索以收集额外证据，采用跨模态策略对候选人选进行排名，并执行多阶段过滤以精炼检索到的数据。最终决策过程遵循两阶段的推理机制：第一阶段将检索到的候选信息与输入验证对比，而第二阶段则基于前一步骤汇总的判断和解释来检测 OOC 虚假信息。

我们的贡献可以概括如下：

- **跨模态数据管道**：我们设计了一个新型的跨模态数据管道，该管道在局部和全局尺度上运行，并借助一个滤波模块，将 OOC 图像-文本对转换为高质量的候选集。这些精选的候选集作为准确判断和连贯解释的基础。
- **无需训练的适应方法用于 MLLMs**：我们提出了 E-FreeM^2，这是一种实用的、无需训练的方法，用于通过两阶段思维链指令机制适应现有的多模态语言模型以检测情境外的虚假信息。我们的方法有效地利用了检索到的跨模态候选者，有效建模图像文本内部关系和声明证据外部连接，实现了同时的情境外检测和解释。
- **无需训练数据的顶级性能**：综合实验表明，E-FreeM^2 在检测精度上显著优于最先进的（SOTA）方法，同时无需训练数据或额外的计算资源。此外，它生成精确且有说服力的解释，增强了其决策的可解释性。

本文的其余部分结构如下。第 2 节回顾了多模态模型中上下文外检测和链式思维提示的相关工作。随后，第 3 节介绍了所提出的 E-FreeM^2 框架，并概述其证据检索管道和两阶段推理过程。第 4 节描述了实验设置，包括数据集、指标、结果和消融研究。最后，第 5 和 6 节讨论了限制性因素，提出了未来的研究方向，并总结了主要贡献。

2 相关工作

2.1 上下文外检测

随着多模态虚假信息在社交网络上的迅速蔓延，研究人员越来越多地专注于开发多模态事实核查的解决方案，特别是在检测脱离上下文（OOC）的虚假信息方面。一些方法 [6, 13] 利用知识丰富的预训练模型对给定的图文配对进行内部检查。尽管这些方法在 OOC 检测中表现出强大的性能，但它们往往忽视了与图文声明配对相关的外部知识的整合。因此，这类模型难以有效捕捉逻辑或事实上的不一致，导致其学习能力受限。

替代策略纳入外部资源进行验证。例如，先前的研究 [8, 9, 24] 利用包含未修改声明的参考数据集来近似现实世界知识，通过比较给定声明与检索到的参考内容以促进 OOC 检测。同样地，[1] 分别检索基于网络的文本和视觉输入证据，然后评估它们在多种模式下与声明的一致性。此外，Qi 等人 [21] 首次将 LVLM 模型与 Vicuna-13B 集成以解决生成式 AI 领域的 OOC 不实信息问题。相比之下，我们的方法通过采用无训练框架以及多尺度跨模态候选检索来强调效率。

2.2 多模态大语言模型中的思维链

多模态大语言模型（MLLMs）已开发用于将感知与语言模型对齐，允许在语言和多模态领域之间进行跨模态知识转移 [7]。引入了如 SPIQA 等数据集以实现多模态问答及链式思维（CoT）等评估策略，进一步提高了模型在解释科学论文中复杂信息方面的性能 [20]。此外，研究还探讨了在多模态大型语言模型（MLLM）中使用 CoT 提示系统进行图像质量评估（IQA），展示了上下文检索和细粒度评估 [34] 的有效性。另外，在体育理解、绿色建筑设计决策等领域，将链式思维方法整合到多模态推理任务中已显示出有希望的结果 [11, 35]。

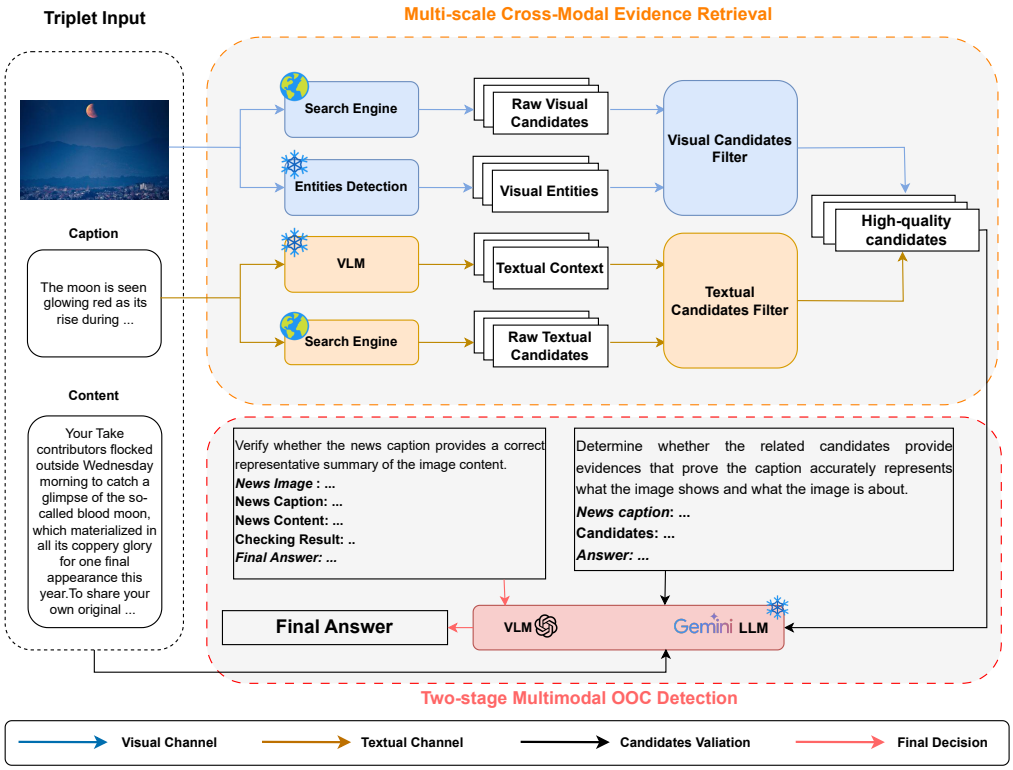

图 1: E-FreeM^2 框架概述。我们的高效无训练多尺度跨模态新闻验证系统通过两个主要组件运行：(1) 多尺度跨模态证据检索- 使用双文本和视觉检索管道，结合层次相似性过滤和以视觉为中心的排名，从外部来源获取上下文相关的证据；以及 (2) 两阶段多模态 OOC 检测- 第一阶段使用结构化推理进行基于证据引导的多模态验证，评估声明-证据一致性，而第二阶段通过对外部证据的全面综合和直接视觉分析进行最终的多模态决策推理。

3 通过 MLLMs 进行高效无训练多尺度和跨模态新闻验证的设计与实现

3.1 多尺度跨模态证据检索

为了有效检测情境不符（OOC）的虚假信息，我们提出了一种**多尺度跨模态证据检索**框架，该框架包括三个关键组件：**跨模态检索**，**基于相似性的证据精炼**和**视觉中心排名**。该框架系统地从文本和视觉模态中检索、过滤并排名外部证据。与仅依赖基于文本或基于图像匹配的传统单模态检索策略不同，我们的方法整合了两种模态以更有效地捕捉跨模态不一致性。这确保检索到的证据在语境上相关，从而提高事实验证和解释推理的能力。

跨模态检索. 我们的系统采用两条互补的检索管道来获取多样且语境相关的证据。

- **文本检索管道**：给定一个输入声明，文本管道使用基于声明的查询从外部来源中提取相关标题、文章片段和元数据。
- **可视检索管道**：为了评估视觉一致性，执行基于图像的反向搜索以检索视觉上相似的图像及其相关元数据，例如来源域、上下文标题和发布时间戳。

多阶段基于相似性的证据优化. 为了提高检索准确性，我们引入了一个模块化设计的**过滤模块**，旨在使用特定模式和跨模态相似性指标进行分层证据细化。这个两步过程确保仅保留高质量的相关证据用于下游推理。过滤包括两个主要阶段：**模态特定相似性过滤**和**上下文相关性过滤**。

在第一阶段，系统通过计算相似度分数来评估文本和视觉证据的相关性：**模态特定相似性过滤**如果候选者的相似度分数超过预定义的阈值：

- **文本相似性过滤**：输入声明与候选文本证据（例如，标题、片段）之间的语义相似度通过基于变压器的密集嵌入（all-MiniLM-L6-v2）和余弦相似度评分进行计算。
- **视觉相似性过滤**: 感知相似性通过使用视觉变压器（ViT）嵌入计算，然后与查询图像进行余弦相似性匹配。

则保留这些候选者，

$$S_{\text{similarity}} \geq \theta, \quad \theta = 0.7 \tag{1}$$

确保只有强烈相关的证据被保存以进行进一步细化。

第二阶段，**上下文相关性过滤**，应用额外的标准来消除无关、不可靠或冗余的信息。此过程涉及：

- **域过滤**：仅保留来自经过验证和信誉良好的来源（例如，值得信赖的新闻机构）的证据以确保可靠性。
- **语言过滤**：排除非英语证据以保持下游推理中的语言一致性，防止因翻译错误而导致的误解。
- **冗余减少**：通过域名标题去重启发式方法去除重复或近似重复条目，以防止来自冗余来源的偏差。

视觉中心排名. 虽然文本和视觉相似性都被计算在内，但排名机制将**视觉相似性**作为最终候选选择中的主导因素：

$$S_{\text{final}} = S_{\text{visual}} \tag{2}$$

此优先级排序基于这样一个观察：视觉上相似的图像倾向于保持现实世界的语境完整性，并且对操纵性的文本重新表述不太敏感。尽管语义相似性仍然是次要标准，但它用于解决边缘案例并细化在视觉上相似候选集中的排名。经过过滤和排名阶段后，精炼过的证据集被整合为一个统一的多模态表示，其中包括基于文本和基于图像的证据。这个结构化的集合随后由推理模块处理，该模块利用它进行 OOC 检测和解释性输出生成。

3.2 两阶段多模态 OOC 检测

为了实现健壮且高效的非上下文（OOC）虚假信息检测，我们提出了一种结合外部证据检索与结构化推理机制的两阶段多模态 OOC 检测框架。该方法受到了专业事实核查工作流程的启发，在这些流程中，检索到的知识与场景级图像理解相结合以评估声明的真实性。通过利用跨模态检索和逐步多模态推理，系统确保了准确、可解释且具备上下文感知的决策制定。

阶段 *1*: 证据引导的多模态验证。. 在第一阶段，我们专注于验证通过新闻标题表达的主张与从可信来源检索到的外部候选证据之间的一致性。一个多模态大型语言模型（Gemini）被用于评估标题文本内容与检索到的候选信息之间的对齐情况。每个候选者使用一种包含语义特征和相似度指标的结构化模式进行表示，以捕捉外部证据在多大程度上支持或挑战了标题叙述。该模型执行一个集成相似性分析与上下文推理的分层验证过程。它不仅考虑孤立的特征，而是共同考量图像-标题一致性、外部来源的可信度以及检索到的证据在多大程度上证实或挑战主张的情况。特别注意验证关键元素，如描绘的实体、时间空间引用以及新闻图片、标题和外部证据之间的整体连贯性。

阶段 *2*: 最终多模态决策推理。. 在第二阶段，系统对多模态见解进行全面综合，以最终确定声明的离题状态。由多模态大型语言模型（GPT-4o mini）驱动的专门决策模块将第一阶段的验证结果与新闻图像的直接视觉推理相结合。该模型结合外部证据的结果和场景级别的图像理解来进行结构化推理。它从图像中提取关键语义元素，如实体、地点、物体和情境线索，并将这些与标题中的声明和暗示进行交叉参考。此外，该

模型评估标题是否引入了误导性叙述或选择性框架，可能扭曲图像所传达的意图信息。

具体来说，我们概述了集成到最终检查提示中的*CoT*技术：.

- **逐步分析：**提示要求用户遵循一系列逻辑步骤，将任务分解为可管理的部分，如字幕匹配、证据验证和上下文对齐 [33]。这种任务的分解增强了清晰度并支持结构化推理。
- **情境推理：**提示强调了理解图像周围更广泛背景的重要性 [22]。这包括验证诸如地点、时间、人物以及图像中描绘的具体事件等细节。情境推理确保字幕不仅被检查直接匹配，还根据更广泛的环境因素进行评估。
- **证据验证：**关键的 CoT 技术之一是在提示中验证来自多个来源的证据。指令鼓励使用支持性证据（如文本描述、图像相似度分数、领域权威性）来证明字幕所作的声明。这种基于证据的推理增强了最终决策的准确性。
- **置信评分**：这些提示中包含一个信心评分系统（范围从 0 到 10），以反映决策的确定性 [19]。在这一阶段，CoT 技术显而易见，因为每一条证据（例如标题或图片的相似度）都逐步增加了整体的信心分数，从而促进了一个更细致的决策过程，并避免了幻觉。

总体而言，在多模态语言模型中融入链式思维方法显示出显著的潜力，能够增强模型性能并解决各领域的复杂推理任务。

4 实验

4.1 实验设置

4.1.1 **数据集**. 我们在 NewsCLIPpings 数据集上评估我们的框架，这是迄今为止最大的用于检测脱离上下文（OOC）的虚假信息的数据集。NewsCLIPpings 是从 VisualNews 构建的，后者是一个大规模语料库，包含来自四个主要新闻媒体的图像-标题配对：《卫报》、BBC、《今日美国》和《华盛顿邮报》。这些 OOC 实例是通过用从无关新闻事件中检索到的视觉上或语义上相似的图像替换原始图像而生成的，这在图像与标题之间产生了具有挑战性的不匹配。遵循先前的工作，我们仅在合并/平衡测试集（7,264 个样本）上进行评估，这确保了各种检索策略下上下文内和脱离上下文情况的均衡混合。采用无需训练的方法，我们报告总体准确率，以及脱离上下文（OOC）和非脱离上下文（NOOC）情况下分别的成绩。

4.1.2 **评估指标**. **准确性：**测量模型预测的整体正确性，通过评估正确分类样本的比例来实现。它在三个类别中报告：(1) **所有**，代表所有样本的整体准确性，(2) **上下文外 (OOC)**，测量检测假新闻的准确性，以及 (3)**NOOC（非脱离上下文）**，评估识别真实新闻的准确性。

4.2 性能比较

方法	所有	离群值	NOOC
SAFE [40]	52.8	54.8	52.0
EANN [32]	58.1	61.8	56.2
VisualBERT [10]	58.6	38.9	78.4
CLIP [23]	66.0	64.3	67.7
DT-Transformer [1]	77.1	74.8	75.6
CCN [16]	84.7	84.8	84.5
Neu-Sym detector [38]	68.2	-	-
SNIFFER	88.4	86.9	**91.8**
E-FreeM^2 (Ours)	**90.0**	**90.67**	89.49

表 1: 不同方法的准确性（%）比较。

4.2.1 *OOC* **检测比较**. 表 1 表明 E-FreeM^2 达到了最高的总体准确率（90.0%），超过了所有基线方法，包括 SNIFFER（88.8%）、CCN（84.7%）和 DT-Transformer（77.1%）。值得注意的是，E-FreeM^2 在上下文外（OOC）检测中表现出色，准确率达到 90.67%，比 SNIFFER（86.9%）高出显著的幅度，在非上下文外（NOOC）情况下也保持了竞争力的表现（89.49%）。这些结果表明 E-FreeM^2

能够有效区分上下文内和上下文外样本，验证了其在不同检索策略上的鲁棒性。

4.2.2 **效率比较**. 表 2 强调了我们的方法在可训练参数和推理时间方面与现有方法相比的效率。不同于 VisualBERT (110M)、CLIP (149M) 和 SNIFFER (99M)，这些需要大量的可训练参数，我们的方法以**零个可训练参数**运行，展示了其轻量级和无需训练的特点。此外，我们在对样本进行分类时实现了 **12.77 秒**的推理时间，强调了其计算效率，如表 1 所示。

方法	可训练参数
VisualBERT [10]	110M
CLIP [23]	149M
SNIFFER	99M
Ours	**0M**

表 2: 不同方法的可训练参数和推理时间。

4.3 消融研究

4.3.1 **排名和过滤策略**. 为了评估不同证据过滤策略的有效性，我们使用了三种排名方法进行了性能比较：仅相似度、仅领域和两者结合（将相似性和领域过滤相结合）。结果在图 2 中可视化，其中 x 轴表示排名位置（Top-1, Top-2 和 Top-3），y 轴表示检测出上下文外（OOC）不实信息的准确率（%）。结果显示，在所有排名位置上，结合基于相似性和领域过滤的方法都达到了最高的准确性，超过了仅依赖其中任何一种方法的方式。虽然基于相似性的排序提供了强大的上下文相关性，但融合领域的可靠性进一步提升了性能，这表明混合过滤策略对于有效检测 OOC 不实信息的重要性。

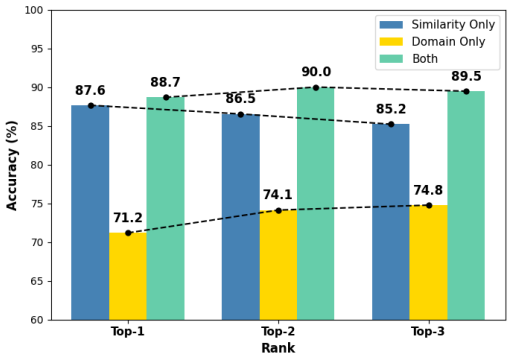

图 2: 滤波策略的消融研究结果

4.3.2 **训练数据集 利用率**. 为了比较我们的方法与 SNIFFER 的有效性，在图 3 中 y 轴绘制了准确率（%），x 轴绘制了使用的训练数据的比例，其中 SNIFFER 的性能随着更多数据而提升，而我们的方法（用红星标记）在不需要任何训练数据的情况下实现了更高的准确性。结果显示我们的方法显著优于 SNIFFER，突显其作为无需训练的解决方案在检测 OOC 不实信息方面的有效性。

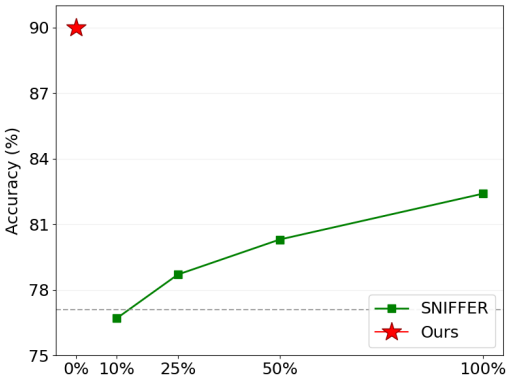

图 3: 使用训练集的百分比为 NewsCLIPpings[13]

Method	Accuracy
E-FreeM^2 (Ours)	**90.0**
- w/o Image Evidences	76.48
- w/o Text Evidences	77.05
- w/o Domain Filters	56.46

表 3: 不同证据过滤方法的评估结果。

4.3.3 **滤波方法的影响**. 表 3 显示了证据过滤对模型性能的影响。完整模型达到了最高的准确率（90.0%）。移除图像证据（准确率降至 76.48%）或文本证据（准确率降至 77.05%）都会显著降低性能，这表明两者都至关重要。相反，低质量的证据严重损害了性能：移除领域过滤器，这些过滤器消除了此类证据，导致准确率骤降到 56.46%。这强调了域过滤在通过去除有害输入来维持语境理解和高准确率方面的重要作用。

5 限制与讨论

尽管 E-FreeM² 表现出色，但仍需承认一些限制以供未来改进。首先，框架依赖外部搜索引擎进行证据检索，这导致了对外部资源可用性和质量的依赖。其次，当前实现主要侧重于英文内容，可能限制其在其他语言环境中的适用性。第三，虽然对 OOC 误导信息有效，但其对抗诸如深度伪造等其他复杂操纵的稳健性尚未得到充分评估。此外，以视觉为中心的排名虽然有益，但在文本线索比视觉相似性更重要的场景下可能会表现不佳。最后，多阶段过滤过程的效率还可以进一步优化，以减少实时应用中的延迟。

6 结论

本文提出了一种无需训练的跨模态事实验证新框架。脱离上下文的虚假信息检测任务是识别误导性的图文配对，这些配对曲解了视觉内容。我们的方法 E-FreeM² 利用多尺度检索管道和基于大型多模态语言模型的两阶段推理机制来分析图像、标题与外部检索证据之间的语境一致性。通过结合跨模态证据精炼和结构化的链式思考推理，我们的系统提供了高度准确且可解释的验证结果。

未来的研究方向包括扩展 E-FreeM² 的功能以处理多语言内容，并探索防御更复杂的操作，如深度伪造。我们还旨在进一步优化延迟和检索机制，以便在边缘设备上实现更快的事实核查。通过充分利用多模态推理和检索的潜力，我们希望构建更具适应性、安全性和鲁棒性的解决方案，以对抗动态媒体环境中存在的虚假信息。

Acknowledgments

我们想感谢 Vingroup 创新基金会（VINIF）部分资助了本研究，资助号为 VINIF.2021.JM01.N2。

References

[1] Sahar Abdelnabi, Rakibul Hasan, and Mario Fritz. 2022. Open-domain, content-based, multi-modal fact-checking of out-of-context images via online resources. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 14940–14949.

[2] Nadia Alonso-López, Pavel Sidorenko Bautista, and Fábio Giacomelli. 2021. Beyond Challenges and Viral Dance Moves: TikTok as a Vehicle for Disinformation and Fact-Checking in Spain, Portugal, Brazil, and the USA. *Anàlisi* 64 (06 2021), 65–84. doi:10.5565/rev/analisi.3411>

[3] Shivangi Aneja, Chris Bregler, and Matthias Nießner. 2023. COSMOS: catching out-of-context image misuse using self-supervised learning. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 37. 14084–14092.

[4] Sarah Evanega, Mark Lynas, Jordan Adams, and Karinne Smolenyak. 2020. Coronavirus misinformation: quantifying sources and themes in the COVID-19 'infodemic' (Preprint). doi:10.2196/preprints.25143

[5] Lisa Fazio. 2020. Out-of-context photos are a powerful low-tech form of misinformation. *The Conversation* 14, 1 (2020).

[6] Yimeng Gu, Mengqi Zhang, Ignacio Castro, Shu Wu, and Gareth Tyson. 2024. Learning Domain-Invariant Features for Out-of-Context News Detection. *arXiv preprint arXiv:2406.07430* (2024).

[7] Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, et al. 2023. Language is not all you need: Aligning perception with language models. *Advances in Neural Information Processing Systems* 36 (2023), 72096–72109.

[8] Ayush Jaiswal, Ekraam Sabir, Wael AbdAlmageed, and Premkumar Natarajan. 2017. Multimedia semantic integrity assessment using joint embedding of images and text. In *Proceedings of the 25th ACM international conference on Multimedia*. 1465–1471.

[9] Ayush Jaiswal, Yue Wu, Wael AbdAlmageed, Iacopo Masi, and Premkumar Natarajan. 2019. Aird: Adversarial learning framework for image repurposing detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11330–11339.

[10] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557* (2019).

[11] Yihui Li, Xiaoyue Yan, Hao Zhou, and Borong Lin. 2024. Question Answering for Decisionmaking in Green Building Design: A Multimodal Data Reasoning Method Driven by Large Language Models. *arXiv preprint arXiv:2412.04741* (2024).

[12] Piper Liu and Vincent Huang. 2020. Digital Disinformation About COVID-19 and the Third-Person Effect: Examining the Channel Differences and Negative Emotional Outcomes. *Cyberpsychology, Behavior, and Social Networking* 23 (07 2020). doi:10.1089/cyber.2020.0363

[13] Grace Luo, Trevor Darrell, and Anna Rohrbach. 2021. Newsclippings: Automatic generation of out-of-context multimodal media. *arXiv preprint arXiv:2104.05893* (2021).

[14] Michael Mu, Sreyasee Das Bhattacharjee, and Junsong Yuan. 2023. Self-supervised distilled learning for multi-modal misinformation identification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2819–2828.

[15] Alex Nikolov, Giovanni Da San Martino, Ivan Koychev, and Preslav Nakov. 2020. Team Alex at CLEF CheckThat! 2020: Identifying Check-Worthy Tweets With Transformer Models. arXiv:2009.02931 [cs.CL] https://arxiv.org/abs/2009.02931

[16] Stefanos-Iordanis Papadopoulos, Christos Koutlis, Symeon Papadopoulos, and Panagiotis Petrantonakis. 2023. Synthetic misinformers: Generating and combating multimodal misinformation. In *Proceedings of the 2nd ACM International Workshop on Multimedia AI against Disinformation*. 36–44.

[17] Stefanos-Iordanis Papadopoulos, Christos Koutlis, Symeon Papadopoulos, and Panagiotis C Petrantonakis. 2023. Red-dot: Multimodal fact-checking via relevant evidence detection. *arXiv preprint arXiv:2311.09939* (2023).

[18] Long-Khanh Pham, Hoa-Vien Vo-Hoang, and Anh-Duy Tran. 2024. A Generative Adaptive Context Learning Framework for Large Language Models in Cheapfake Detection *(ICMR '24)*. Association for Computing Machinery, New York, NY, USA, 1288 – 1293. doi:10.1145/3652583.3657597

[19] Gwenyth Portillo Wightman, Alexandra Delucia, and Mark Dredze. 2023. Strength in Numbers: Estimating Confidence of Large Language Models by Prompt Agreement. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, Anaelia Ovalle, Kai-Wei Chang, Ninareh Mehrabi, Yada Pruksachatkun, Aram Galystan, Jwala Dhamala, Apurv Verma, Trista Cao, Anoop Kumar, and Rahul Gupta (Eds.). Association for Computational Linguistics, Toronto, Canada, 326–362. doi:10.18653/v1/2023.trustnlp-1.28

[20] Shraman Pramanick, Rama Chellappa, and Subhashini Venugopalan. 2025. SPIQA: A Dataset for Multimodal Question Answering on Scientific Papers. arXiv:2407.09413 [cs.CL] https://arxiv.org/abs/2407.09413

[21] Peng Qi, Zehong Yan, Wynne Hsu, and Mong Li Lee. 2024. 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*. 13052–13062.

[22] Shuofei Qiao, Yixin Ou, Ningyu Zhang, Xiang Chen, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, and Huajun Chen. 2023. Reasoning with Language Model Prompting: A Survey. arXiv:2212.09597 [cs.CL] https://arxiv.org/abs/2212.09597

[23] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PmLR, 8748–8763.

[24] Ekraam Sabir, Wael AbdAlmageed, Yue Wu, and Prem Natarajan. 2018. Deep multimodal image-repurposing detection. In *Proceedings of the 26th ACM international conference on Multimedia*. 1337–1345.

[25] Ian Sample. 2020. What are deepfakes – and how can you spot them? Retrieved Mar 08, 2024 from https://www.reuters.com/world/middle-east/false-claims-israel-hamas-war-mushroom-online-put-focus-musks-x-2023-10-10/

[26] Rui Shao, Tianxing Wu, and Ziwei Liu. 2023. Detecting and grounding multi-modal media manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6904–6913.

[27] Sahar Tahmasebi, Eric Müller-Budack, and Ralph Ewerth. 2024. Multimodal misinformation detection using large vision-language models. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. 2189–2199.

[28] Ruben Tolosana, Ruben Vera-Rodriguez, Julian Fierrez, Aythami Morales, and Javier Ortega-Garcia. 2020. Deepfakes and beyond: A survey of face manipulation and fake detection. *Information Fusion* 64 (2020), 131–148.

[29] Jonathan Tonglet, Marie-Francine Moens, and Iryna Gurevych. 2024. "Image, Tell me your story!" Predicting the original meta-context of visual misinformation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 7845–7864. doi:10.18653/v1/2024.emnlp-main.448

[30] Hoa-Vien Vo-Hoang, Long-Khanh Pham, and Minh-Son Dao. 2024. Detecting Out-of-Context Media with LLaMa-Adapter V2 and RoBERTa: An Effective Method for Cheapfakes Detection. In *Proceedings of the 2024 International Conference on Multimedia Retrieval* (Phuket, Thailand) *(ICMR '24)*. Association for Computing Machinery, New York, NY, USA, 1282 – 1287. doi:10.1145/3652583.3657596

[31] Xueyu Wang, Jiajun Huang, Siqi Ma, Surya Nepal, and Chang Xu. 2022. Deepfake disrupter: The detector of deepfake is my friend. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14920–14929.

[32] Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. 2018. Eann: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining*. 849–857.

[33] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. arXiv:2201.11903 [cs.CL] https://arxiv.org/abs/2201.11903

[34] Tianhe Wu, Kede Ma, Jie Liang, Yujiu Yang, and Lei Zhang. 2024. A Comprehensive Study of Multimodal Large Language Models for Image Quality Assessment. arXiv:2403.10854 [cs.CV] https://arxiv.org/abs/2403.10854

[35] Haotian Xia, Zhengbang Yang, Junbo Zou, Rhys Tracy, Yuqing Wang, Chi Lu, Christopher Lai, Yanjun He, Xun Shao, Zhuoqing Xie, Yuan fang Wang, Weining Shen, and Hanjie Chen. 2024. SPORTU: A Comprehensive Sports Understanding Benchmark for Multimodal Large Language Models. arXiv:2410.08474 [cs.CV] https://arxiv.org/abs/2410.08474

[36] Xin Yuan, Jie Guo, Weidong Qiu, Zheng Huang, and Shujun Li. 2023. Support or refute: Analyzing the stance of evidence to detect out-of-context mis-and disinformation. *arXiv preprint arXiv:2311.01766* (2023).

[37] Fanrui Zhang, Jiawei Liu, Qiang Zhang, Esther Sun, Jingyi Xie, and Zheng-Jun Zha. 2023. Ecenet: explainable and context-enhanced network for muti-modal fact verification. In *Proceedings of the 31st ACM International Conference on Multimedia*. 1231–1240.

[38] Yizhou Zhang, Loc Trinh, Defu Cao, Zijun Cui, and Yan Liu. 2024. Interpretable Detection of Out-of-Context Misinformation with Neural-Symbolic-Enhanced Large Multimodal Model. arXiv:2304.07633 [cs.CL] https://arxiv.org/abs/2304.07633

[39] Hanqing Zhao, Wenbo Zhou, Dongdong Chen, Tianyi Wei, Weiming Zhang, and Nenghai Yu. 2021. Multi-attentional deepfake detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2185–2194.

[40] Xinyi Zhou, Jindi Wu, and Reza Zafarani. 2020. : Similarity-aware multimodal fake news detection. In *Pacific-Asia Conference on knowledge discovery and data mining*. Springer, 354–367.