arXiv:2303.01217v1 [cs.MM] 2 Mar 2023

合成虚假信息生成器：生成和应对多模态虚假信息

Stefanos-Iordanis Papadopoulos*[1,2]，Christos Koutlis[1]，Symeon Papadopoulos[1]，和 Panagiotis C. Petrantonakis[2]

[1]希腊研究与技术中心信息技术研究所。[2]塞萨洛尼基亚里士多德大学电气与计算机工程学院。*{stefpapad,ckoutlis,papadop}@iti.gr, ppetrant@ece.auth.gr*

摘要

随着社交媒体的发展和多媒体内容传播的日益增加，错误信息的扩散已成为一个重大问题。因此，有必要采用有效的多模态错误信息检测（MMD）策略，以检测图像及其配套文本的组合是否可能产生误导或传播错误信息。由于深度神经网络具有数据密集型特征，且人工标注过程十分耗时，研究人员一直在探索各种自动生成合成多模态错误信息的方法——我们将其称为合成虚假信息生成器——以用于训练 MMD 模型。然而，对现实世界中的错误信息进行的评估有限，且缺乏与其他合成虚假信息生成器的比较，使得研究领域的进展难以评估。为此，我们对现有的和新提出的合成虚假信息生成器开展了一项比较研究，其中涉及：（1）脱离语境（OOC）的图像-标题对；（2）跨模态命名实体不一致（NEI）；以及（3）混合方法，并使用 COSMOS 基准对它们进行了现实世界错误信息评估。比较研究表明，我们提出的基于 *CLIP* 的命名实体交换能够生成在多模态准确率方面超过其他 OOC 和 NEI误导信息生成器的 MMD 模型，而混合方法能够带来更高的检测准确率。然而，在从 COSMOS评估协议中消除信息泄漏后，较低的敏感度分数表明，该任务比先前研究所认为的更具挑战性。最后，我们的研究结果显示，基于 NEI 的合成虚假信息生成器往往存在单模态偏差，即纯文本多模态错误信息检测模型可以胜过多模态检测器。

***K*关键词** 错误信息检测、多模态学习、合成数据集、比较研究

1 引言

错误信息的泛滥正成为当今社会日益严峻的挑战，尤其是在社交媒体和互联网得到广泛使用的背景下。因此，自动检测错误信息已成为一项重要挑战，研究人员正在探索通过自然语言处理[1] 识别虚假主张，以及通过计算机视觉技术 [2]检测篡改图像（如深度伪造）的各种方法。上述挑战主要集中于单一模态，因此需要使用单模态检测模型。然而，研究表明，多媒体内容比纯文本内容 [3]更能吸引注意力，也更易于广泛传播。此外，图像的存在会使虚假陈述对人们更具说服力 [4], ，这凸显了多模态错误信息检测（MMD）的重要性。

如图 1 所示，多模态虚假信息检测模型经过训练，能够判断图像及其配套说明文字组合在一起后是真实的，还是具有误导性（会导致错误信息）。图 1 上方的图像展示了一场遍地垃圾的音乐节，而配套说明文字声称，事件发生于2022年6月，在环保主义者 Greta Thunberg 发表演讲之后。实际上，尽管图像描绘的是格拉斯顿伯里音乐节，但它并非拍摄于

*通讯作者

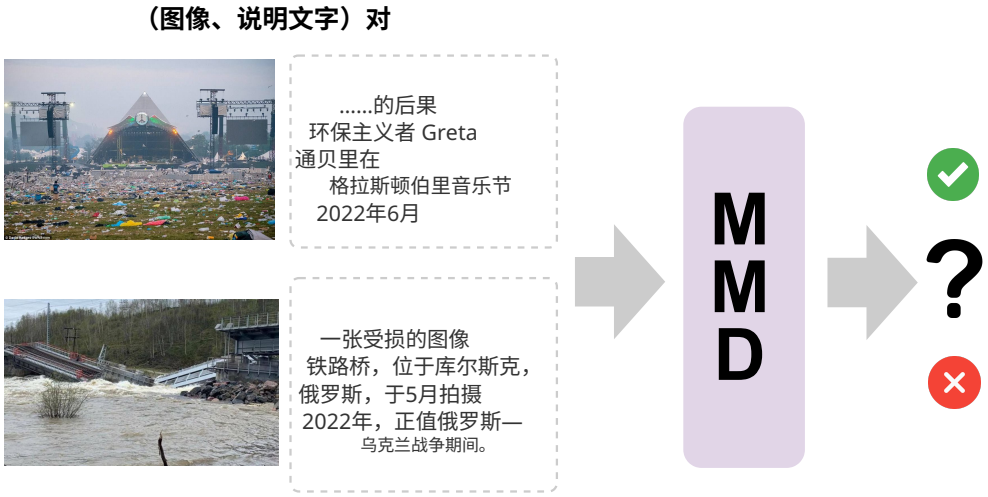


图 1：多模态错误信息检测（MMD）模型旨在判断一对（图像、说明文字）是真实还是具有误导性。图像和说明文字取自 reuters.com。

Greta Thunberg 演讲后的2022年6月，而是拍摄于2015年；那时 Greta Thunberg 还未成为公众人物[2]。此案例通过操纵命名实体（人物和日期），将矛头指向一位公众人物及其支持者，使他们受到负面关注。下方图像展示了一座坍塌入水中的铁路桥，说明文字声称事件发生在俄乌战争期间的俄罗斯库尔斯克。然而，库尔斯克当天确实有一座桥梁受损，但这张图像实际拍摄于2020年，地点是俄罗斯摩尔曼斯克[3]。此案例展示了如何使用脱离语境的图像，无论是出于失误，还是意在夸大或淡化事件的严重程度。从以上示例可以看出，多模态虚假信息形式多样，传播动机各异，而且往往包含难以辨别的细微线索。

研究人员一直在探索如何训练用于多模态错误信息检测的深度神经网络 [5]。考虑到训练用于多模态错误信息检测的深度神经网络需要大量数据，且人工标注既耗时又费力，研究人员一直在研究自动生成合成多模态虚假信息的方法，我们将这类方法称为合成虚假信息生成器。这些方法包括生成脱离语境（OOC）的图文对，或制造跨模态实体不一致（NEI）。OOC 是指将图像与不相符的说明文字配对 [6]，而 NEI 则是修改原本真实的说明文字中的命名实体 [7]。以往的研究采用随机抽样 [8, 6] 或基于特征的抽样方法 [9, 10] 来生成 OOC，并采用簇内随机采样 [11] 或基于规则的随机采样 [7] 来生成 NEI。图 2 展示了生成的 OOC 和 NEI 错误信息示例。

然而，尽管以往研究做出了种种努力，但这些研究仍存在局限：它们在由自身的合成虚假信息生成器生成的测试集上评估方法，而非在现实世界多模态错误信息上进行评估；唯一的例外是介绍 COSMOS 数据集的研究 [6]。不过，后一项研究的作者采用了存在问题的评估协议，该协议受到信息泄漏的影响，如图 3 所示（详情请参见第 3.1.4 节）。此外，先前的研究并未将其方法与其他合成虚假信息生成器进行比较。缺乏比较，也未能在真实世界数据上进行评估，妨碍了研究界衡量研究进展、确定当前多模态虚假信息检测领域最先进水平的能力。为此，我们复现了多个合成虚假信息生成器，在生成的数据上微调我们的基于 Transformer 的多模态虚假信息检测模型（称为 DT-Transformer），最后在涵盖现实世界多模态错误信息的 COSMOS 基准 [6] 上对其进行比较。在我们的比较研究中，我们考察 OOC 或 NEI 哪个更能代表现实世界中的错误信息。我们的假设是，尽管先前研究将二者视为独立任务，但 OOC 和 NEI 都是有效开展多模态虚假信息检测不可或缺的组成部分。因此，我们还研究了多种混合型合成虚假信息生成器。

我们工作的主要贡献可概括如下：

[2]https://www.reuters.com/article/factcheck-glastonbury-greta-idUSL1N2YE1JD[3] https://www.reuters.com/article/factcheck-destroyed-bridge-idUSL2N2WU1CM



图 2：由两种类型的合成虚假信息生成器生成的训练数据：一种生成脱离语境（OOC）的错误信息，另一种生成跨模态命名实体不一致（NEI）。给定一个真实的（I_a，C_a）图像-标题对，OOC 会对图像进行采样 I_x，生成（I_x，C_a）；而 NEI 会操纵 C_a 中的命名实体，以生成伪造的C_f。这些示例分别通过基于 CLIP 的采样（*CSt-alt*）和基于 CLIP 的命名实体交换（*CLIP-NESt-alt*）生成，用于 OOC 和 NEI（有关这些方法的更多详情，请参见第 3.2 节）。图像和标题取自 VisualNews 数据集 [12]。

- 我们首次对合成虚假信息生成器开展比较研究，全面评估了当前多模态虚假信息检测领域的最先进水平，并为该领域的未来研究提供指导。
- 我们提出了“基于 CLIP 的命名实体交换”，其多模态准确率高于其他 OOC 和 NEI 方法。此外，我们还证明，混合型合成虚假信息生成器可以进一步提高检测准确率。

- 我们的研究结果表明，尽管基于 NEI 的方法可能优于脱离上下文方法，但它们往往存在单模态偏差，即仅文本模型的表现可能优于多模态检测器。此外，“伪造”配对的敏感度分数较低，或其命中率较低，表明该问题比以往研究所指出的更具挑战性。我们就未来研究如何应对这些挑战提出了建议。

2 相关工作

关于多模态错误信息检测（MMD）的研究主要关注脱离语境的图像-语言对（OOC），或跨模态命名实体不一致（NEI）。多模态虚假信息的一种常见形式是去语境化：将真实图像与脱离语境的说明文字配对，从而造成误导性印象。因此，研究人员采用随机抽样（[8, 6]）和基于特征的抽样方法（[9, 10]）来生成 OOC。MAIM 数据集是通过对从 Flickr 收集的图像-标题对进行随机抽样而创建的 [8]。作者利用深度表征学习构建了联合嵌入，随后计算图像-标题的一致性。类似地，Aneja 等人[6] 通过从可信新闻网站收集真实的图像-标题对，再将标题与随机图像匹配以生成 OOC，创建了 COSMOS 训练数据集。作者采用自监督深度学习，并在由现实世界多模态错误信息构成的 COSMOS

基准测试上评估了他们的方法。然而，随机抽样无法确保图像-标题对之间存在任何关联，而且往往会生成容易识别、与现实中的多模态虚假信息（即能够欺骗人类的信息）并不相似的负样本。为此，Luo 等人[9] 利用大型跨模态 CLIP 模型 [13]，结合场景学习和人物匹配模型，创建了 NewsCLIPings 数据集，以生成难负样本。类似地，Twitter-COMMs 数据集通过结合并应用基于 CLIP 的采样（用于生成难负样本）和主题内随机采样（用于解决类别不平衡），在从 Twitter 收集的数据上构建而成[4]，；这些数据涉及三个主题：气候、COVID 和军用车辆 [10]。

另一方面，NEI 指的是合法图像配上经过篡改的说明文字，其中的命名实体（人物、地点、事件）与图像内容或语境不符。“多模态实体图像再利用”（MEIR）数据集通过聚类图像-标题对来构建，聚类依据是“相关性”——地点距离、文本相似度和图像相似度，分别通过 GPS 坐标、word2vec 和在 ImageNet 上预训练的 VGG19 来衡量。随后，他们在当前说明文字与取自同一簇的另一条说明文字之间，随机交换类型相同的命名实体 [11]。类似地，TamperedNews 数据集通过随机替换命名实体来构建，替换后的实体类型相同，并且被替换的人物性别和/或国籍相同、地点之间的地理距离较近、事件属于同一类别（例如体育赛事或自然灾害） [7]。

上述研究要么提供内部消融实验 [11, 7, 9, 10]，要么与简单基线进行比较 [11]，并未将其方法与其他合成虚假信息生成器进行比较。此外，使用上述部分数据集的研究大多侧重于渐进式的方法改进 [14],，例如在 MMD 模型中整合情感 [15]或证据 [16]。除 [6], 外，先前研究并未在现实世界中的错误信息上评估其方法，而是在各自的合成误导者生成的数据上进行评估。因此，研究界缺乏明确的方式来评估该领域的进展，包括 MMD 当前的最先进水平、生成 MMD 训练数据的最佳方式，以及 OOC 或 NEI（或两者）是否更能代表现实世界中的错误信息

错误信息。为弥补这一空白，我们对各种合成虚假信息生成器进行了比较研究——包括 OOC、NEI 和混合方法——并在现实世界多模态错误信息上对它们进行了评估。

最后，Luo 等人[9] 指出，采用命名实体操纵的方法可能会引入语言偏差。为探究这一点，作者训练了一个仅文本 BERT [17] 模型，其结果与 [7] 中使用的多模态模型相近。然而，后者从 ImageNet 预训练的 ResNet 中提取视觉特征，并从现成的 fastText [18]中提取文本特征。在撰写本文时（2021 [9]），BERT 类模型被视为文本任务的前沿模型之一，而 fastText 则是较早期的架构，因此两者的训练方案存在显著差异，无法直接比较。因此，仅凭这些结果无法对是否存在单模态偏差得出明确结论。为弥补这一不足，我们在受控的训练框架下重新考察 NEI 方法是否存在单模态偏差，并在现实世界中的错误信息上进行评估。



图 3：COSMOS 基准 [6]中的伪造（$I, C1, C2$）三元组示例。说明文字 $C2$ 要么为图像提供正确描述（左），要么解释 $C1$ 为何为假（右）。我们认为 $C2$ 属于信息泄漏，因此将其排除在评估协议之外。

[4]https://twitter.com

3 方法框架

3.1 问题表述

本研究比较了多种合成虚假信息生成器——用于生成合成多模态错误信息的方法——并在现实世界多模态错误信息上对其进行评估。该问题被定义为一个二分类任务，其中一个 (I,C) 图像-标题对要么真实，要么伪造。真实标题取自可信新闻来源，伪造标题则由合成虚假信息生成器生成。每个 (I,C) 都由视觉编码器 $E_V(\cdot)$ 和文本编码器 $E_T(\cdot)$ 进行编码，分别生成对应的图像 I 和标题 C 的向量表示$\mathbf{v}_I \in \mathbb{R}^{e\times 1}$ 和$\mathbf{t}_C \in \mathbb{R}^{e\times 1}$，其中 e 是编码器的嵌入维度。提取出的特征会被拼接起来，并输入多模态检测深度学习神经网络 D(·)——即检测器；其参数需要优化，超参数也需要调优。最后，将训练好的检测器的预测结果与由现实世界多模态错误信息组成的测试集进行评估。为了准确、公平地比较不同的错误信息生成器，它们需要采用共同的训练与评估框架。因此，在比较研究中，(1) 编码器、(2) 检测器、(3) 优化与超参数调优，以及 (4) 评估流程都应保持不变，只有错误信息生成器的方法有所不同。该框架确保性能变化是由错误信息生成器造成的，而非其他因素。图4展示了所提出工作流程的整体示意图。本节将分别讨论各个方面。

3.1.1 编码器

在使用合成误导者生成训练数据集后，我们使用编码器从所有图像和标题中提取视觉与文本特征，用于训练检测器。早期关于多模态虚假信息检测的研究主要依赖在 ImageNet 上预训练的卷积神经网络来提取图像特征（即 VGG-19 [8, 11] 和 ResNet50 [7, 6]），并使用词嵌入来提取标题特征（即 word2vec [8, 11] 和 fastText[7]）。近期的方法则采用了大规模多模态与跨模态模型，即 CLIP [9, 10, 19]、VisualBERT [9] 和 VinVL [19]，以提取视觉与文本特征。在上述研究中，CLIP[13] 在多模态虚假信息检测 [9, 10, 19]方面往往优于其他跨模态方法（VinVL 和 VisualBERT）。

对比式语言-图像预训练（简称 CLIP）是一种跨模态模型，经过训练后能够将最相关的文本与图像匹配起来。CLIP 由 Radford 等人[13]开发，并在大规模数据集上训练，包含约 4×10^8 个图文对。事实证明，CLIP 具有强大的零样本能力；也就是说，它在未经过明确训练的任务和领域中也能表现出色。在本研究中，我们首先使用 CLIP ViT-B/32 在 NewsCLIPings 数据集上进行实验，以便将我们的训练流程与 [9] 进行比较；同时，我们也在比较研究中采用了更新且改进后的[5] CLIP ViT-L/14 版本。CLIP ViT-B/32 生成的嵌入向量大小为 $e = 512$，而 ViT-L/14 生成的向量大小为 $e = 768$。我们直接使用预训练的 CLIP[6]，由于计算资源限制，不再对其进行微调。Luo 等人[9] 尝试对 CLIP-ResNet-50 的全部层或顶层进行微调，但结果喜忧参半；在所有情况下，微调都无法稳定地优于“冻结”的预训练 CLIP。已有研究提出了针对大规模跨模态神经网络的稳健微调方法 [20]，但这些方法不在本研究的讨论范围内；因为我们并不试图追求尽可能高的性能，而是主要致力于对各种合成虚假信息生成器进行公平的比较研究。

3.1.2 检测器

以往的研究要么计算跨模态相似度分数 [8, 7, 6, 14]，要么在顶层定义二分类器 [9, 15, 11, 10]。也有少数研究在最终分类层之前添加全连接层，以分析提取出的特征[15, 11]。相比之下，我们认为 Transformer 架构 [21] 更适合作为检测器。我们的基于Transformer的检测器（简称 DT-Transformer）首先将编码后的标题和图像串联起来，并将它们输入由 L层组成的 Transformer 架构；每层包含 h 个注意力头，嵌入维度为d。随后，其输出依次经过归一化层、Dropout 层、由 GELU 函数激活的全连接层、第二个 Dropout 层，以及最终的二分类层。

3.1.3 优化

鉴于我们将多模态虚假信息检测定义为一项二分类任务，因此基于二元交叉熵损失函数对检测器进行优化。由于分布、规模、复杂度等因素存在差异，我们假设检测器在不同训练数据集上要取得最佳表现，可能需要采用不同的超参数。因此，我们根据以下网格搜索来调整检测器的超参数： $L \in \{1, 4\}$ 个Transformer层，每层 $d \in \{128, 1024\}$

[5]https://huggingface.co/sentence-transformers/clip-ViT-L-14[6]
https://github.com/openai/CLIP

维度、 $h \in \{2, 8\}$ 个注意力头，学习率为 $lr \in \{1e-4, 5e-5\}$。丢弃率固定为0.1，批次大小固定为512。所选的超参数网格针对每个合成误导信息数据集共包含16项实验。这显然并未穷尽所有可能，但再增加选项会使所需时间和计算资源呈指数级增长。因此，我们旨在让每种方法都有机会达到足够且具有代表性的表现，即使这并非通过穷尽式优化所能实现的全局最优结果。使用ADAM优化器对检测器进行优化，最多训练30轮，并在10轮后早停。最后，我们提取验证准确率最高的检查点，并将其用于在测试集上的最终评估。

3.1.4 评估协议

除 Aneja 等人[6], 外，先前关于合成虚假信息生成器的研究都未在现实世界中的错误信息上评估其方法。相反，这些研究首先提出一种生成多模态虚假信息的方法，并将其应用于一批真实的图像-标题对。生成数据集后，作者将其划分为训练集、验证集和测试集，并报告在测试集上的最佳性能。这只能说明在合成数据集上训练的模型在通过相同过程生成的数据集上的表现，而不能说明它在“实际环境中”可能多么准确地检测错误信息。此外，先前的研究未与其他合成虚假信息生成器进行直接比较，因此我们无法评估该领域的进展或当前最先进水平。据我们所知，COSMOS[6] 是唯一公开可用的、经过人工标注的多模态虚假信息检测基准。COSMOS 基准——也用于 MMSys'21 的“廉价伪造品检测大挑战” [22] ——包含来自事实核查网站 SNOPES[8] 的 850 个图像-标题对，以及来自可信新闻来源的 850 个真实图像-标题对。因此，我们将使用 COSMOS 评估集开展研究。

然而，值得指出的是， [6]所采用的评估协议存在一些问题。在评估过程中，作者提供由一张图像和两条说明文字组成的三元组（I、 $C1$、 $C2$），并通过考察 $C1$-$C2$ 相似度及其与图像中物体的重合程度，作出基于阈值的判定。首先，这一协议并不能反映我们在现实生活中遇到错误信息的方式：图像通常配有一条说明文字或一小段文字（例如在 Twitter 上）。更重要的是， $C2$ 在伪造实例中，要么是解释 $C1$ 为何虚假，要么是图像的真实说明文字。在图 3 中展示的两个 COSMOS 基准示例里， $C2$的内容分别是“多伦多猛龙队刚刚推出了黑人的命也是命主题巴士，球迷们对此印象深刻”和“展示巴黎一座‘蹦床桥’的照片，这是一项建筑设计竞赛的概念设计”。这显然属于信息泄漏 ，并不能反映我们在现实世界中遇到错误信息的方式。事实核查人员通常不会面对两段彼此独立的文本，再判断哪一段是正确的。相反，他们通常会看到一张图像和一段文字，并需要判断文字是否真实，以及是否准确地匹配并描述了图像。值得注意的是，这些并非个别案例。我们已人工检查了 COSMOS 基准中的数百个实例，发现说明文字 2 也存在同样的问题。因此，我们认为作者报告的 88% 检测准确率并不具有代表性 [6]。因此，在本研究中，我们只使用 COSMOS 基准中的 (I,C1) 元组。

3.2 合成虚假信息生成器

我们定义了三种合成虚假信息生成器，即能够生成以下内容的方法：(1) 脱离语境（OOC）的图像-标题配对；(2) 跨模态实体不一致（NEI），即标题中的某些命名实体经过篡改，与图像内容不对应；(3) 混合方法，将脱离语境和 NEI 错误信息结合起来。表1展示了每种合成虚假信息生成器（OOC 或 NEI）生成的训练样本数量，以及真实配对的数量。

3.2.1 脱离语境的错误信息

为了创建脱离语境（OOC）的图像-标题对，我们首先需要一个真实配对的数据集（I_a， C_a），然后需要一种对 OOC 图像 I_x 或 OOC 说明文字 C_x进行采样的方法。在本研究中，我们使用 VisualNews 数据集[12] ，该数据集包含由四个可信来源（《华盛顿邮报》、《今日美国》、《卫报》和英国广播公司）围绕 159 个主题收集的 1,259,732 个真实（I_a, C_a）配对，主题包括：艺术与文化、世界、法律与犯罪、国际关系、科学与技术、体育、环境、选举及其他。我们使用 VisualNews 训练集生成训练数据，并使用 VisualNews 验证集生成验证数据，以避免样本重叠和信息泄漏。我们采用以下 OOC 合成误导方法进行实验：

[7]DARPA SemaFor 是另一个由现实世界中的错误信息构成的基准测试，并曾用于 [10, 19] ——其中包含 200 个伪造实例——但我们无法获得访问权限。[8]https://www.snopes.com/

表1：每　种类别的实例数，不同合成误导者生成的训练数据集　　s.

合成误导者	真实	OOC	NEI
VisualNews-Training	1,007,785	-	-
NC/Bal	35,536	35,536	-
NC/I-T	226,564	226,564	-
NC/T-T	258,036	258,036	-
随机抽样	1,007,744	1,007,744	-
基于 CLIP 的采样	1,007,744	1,007,744	-
MEIR	82,156	-	57,940
R-NESt	1,007,744	-	924,586
CLIP-NESt-C	1007744	-	835,537
CLIP-NESt-I	1007744	-	859,618
CLIP-NESt-alt	1007744	-	847,693
R-NESt +NC/I-T	226,564	226,564	226,564
R-NESt +CSt-alt	1,007,744	1,007,744	924,586
CLIP-NESt-alt +CSt-alt	1,007,744	982,645	847,693

• 随机字幕采样（**RS-C**）：对于每个实际（I_a，C_a）配对，从整个语料库中随机采样一条说明文字 C_x。该过程曾用于生成 COSMOS 训练集[6]，但我们将其应用于 VisualNews 数据集。

• 主题内随机字幕采样（**RSt-C**）：对于每个（I_a，C_a），随机抽取一条与其主题相同的说明文字 C_x（例如国际政治、选举、环境等）。使用来自同一主题的候选项可以提高相关性的概率。[10] 中也采用了类似流程，但仅用于缓解类别不平衡。我们还定义了主题内随机图像采样（**RSt-I**）：为一个实际的（I_a，C_a）随机抽取一张图像 I_x；以及主题内图像或字幕交替随机采样（**RSt-alt**）：随机选择抽取一张 I_x 还是一条 C_x（由随机选择函数决定）。

• 同样，我们通过字幕间相似度（**CSt-C**）、图像间相似度（**CSt-I**），或交替使用图像间相似度和字幕间相似度（**CSt-alt**）来实现基于 CLIP 的主题内采样：根据从 CLIP 提取的特征，计算最相似的项目（I_x 或 C_x）。我们将来自同一主题的（I_a, C_a）对作为候选项，并使用余弦相似度作为相似性度量。

• 最后，我们对 NewsCLIPings（NC）数据集的三个版本进行实验，分别是：**(1)** NewsCLIPings 语义 / CLIP 文本-图像（**(NC/T-I**）、**(2)** NewsCLIPings 语义 / CLIP 文本-文本（**NC/T-T**）以及 **(3)** NewsCLIPings 合并 / 均衡（**NC/Bal**），这些版本由作者提供[9] [9]。

3.2.2 跨模态命名实体不一致

为了生成存在跨模态命名实体不一致的图像-标题对，我们需要真实的（I_a, C_a）对，以及一种对 C_a 中的命名实体进行采样和替换的方法，从而创建伪造的 C_f。我们使用 VisualNews，并对以下方法进行实验：

• 主题内随机命名实体替换（**R-NESt**）：对于每个（I_a，C_a），识别出 C_a 中的所有实体，并将它们替换为随机抽取的同类型实体（人物、地点、组织、日期、事件等），且这些实体与 C_a属于同一主题。

• 作者提供的 MEIR数据集[10]，但我们使用 CLIP ViT L/14 提取特征，而非像原论文中那样使用 VGG19 [11]。

• 我们提出基于图像-图像相似度（**CLIP-NESt-I**）、标题间相似度（**CLIP-NESt-C**），或交替使用图像-图像相似度和标题间相似度（**CLIP-NESt-alt**）的基于 CLIP 的主题内命名实体替换（CLIP-NESt）：对于每个 (I_a, C_a) 对，根据从 CLIP 提取的特征，找出最相似的 (I_x, C_x) 对，并在 C_a 和 C_x 之间交换相同类型的实体，以生成 C_f。C_x至少应包含一个与 C_a 属于相同类型、但并非同一命名实体的实体（避免将命名实体与其自身交换）；否则，我们会选择下一个最相似的对。我们将以下内容定义为候选项

[9]https://github.com/g-luo/news_剪报[10]
https://github.com/Ekraam/MEIR

来自同一主题的图像-标题对，并使用余弦相似度作为相似度指标。我们使用 SpaCy命名实体识别器（NER），具体来说是 en_core_web_trf 模块，该模块在 NER 任务上的 F1分数为 0.90[11]。我们提出 CLIP-NESt 的理由是，基于 CLIP 的相似度能够检索到语义或主题相似的标题，因此其中的实体更可能在某些方面相关。由此，在相似标题之间交换实体，比随机抽样更有可能生成可信的错误信息。

3.2.3 混合方法

我们还尝试了结合 OOC 和 NEI 错误信息的方法，并将其称为混合方法。我们遵循相同的训练过程，但不进行二分类，而是使用交叉熵损失函数训练检测器执行多分类。检测器经过训练后，会将（I, C）对分为三个类别：真实、NEI 或 OOC。在 COSMOS 数据集上进行评估时，由于 COSMOS 是一个二分类数据集，因此将 NEI 和 OOC 的预测结果设为伪造对。在本研究中，我们结合了几种表现最佳的方法：(1) R-NESt +CSt-alt、(2) R-NESt +NC/I-T，以及 (3) CLIP-NESt-alt + CSt-alt。我们探索混合方法的动机在于，现实世界多模态错误信息可能无法仅由 OOC 或 NEI 充分表示，而是可能需要将两者结合起来。OOC 和 MEI 方法会生成平衡数据集，因为它们会为每个真实对创建一个伪造对。另一方面，一些混合方法呈现出类别不平衡，在这种情况下，我们采用随机下采样。

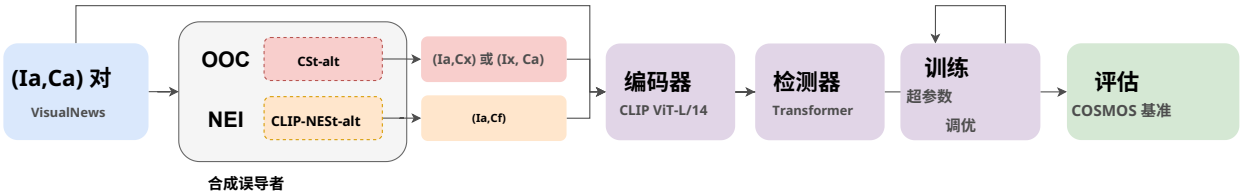


图4：所提出工作流程的高层概览。真实的（I_a，C_a）图像-标题对经合成误导者处理后，生成伪造对。真实与伪造配对由 CLIP ViT-L/14 编码，并用于训练检测器；使用一个 OOC 或 NEI 数据集时，检测器针对二分类进行优化；结合一种 OOC 方法和一种 NEI 方法时，则针对多分类进行优化。此处展示的是混合合成误导者 *CLIP-NESt-alt + CSt-alt*。

4 实验结果

在开始比较研究之前，我们想先考察训练流程——编码器和检测器的选择——是否合理，即能否超越或至少达到以往工作的水平。为此，我们将训练流程与 [9] 进行比较。在表2中可以看到，使用相同的编码器（CLIP ViT-B/32）时，我们的 DT-Transformer 在全部三个 NewsCLIPings 数据集上都能稳定地优于 [9] 。此外，使用 CLIP ViT-L/14 编码器能够显著优于 ViT-B/32。因此，我们在比较研究中采用 DT-Transformer 和 CLIP ViT-L/14 编码器。

表3展示了各种合成虚假信息生成器之间的比较研究。首先来看分布外误导信息生成器，我们发现，同主题候选项可以提升随机抽样的效果（提升+3%）。这是因为，与来自完全不同主题（例如选举和体育）的图像-标题对相比，来自同一主题（例如选举）的图像-标题对更有可能彼此相关。其次，我们发现，在（I_a, C_x）和（I_x, C_a）对之间交替抽样，能够提升*RSt*和*CSt*的性能。以往关于分布外误导信息生成器的研究只对 C_x 标题进行抽样。因此，数据集中每张图像都会出现两次，而标题则可能出现一次、两次或多次，具体取决于抽样方法。尽管影响较小，这一过程仍可能造成某些偏差和不平衡。同样，我们发现，在图像-图像相似度和标题间相似度之间交替使用，可以提升*CLIP-NESt*的性能；这可能是因为该方法生成了更多样化的图像-标题对。此外，结果表明，基于特征的负采样（包括*CSt* 和*NC*）能够胜过随机负采样，因为它们可以生成困难负样本分布外对，更准确地反映现实世界中的错误信息。最后，所有基于分布外数据的多模态合成误导信息生成器都优于其单模态对应方法（仅图像和仅文本），因此不会受到单模态偏差的影响。

转向NEI误导信息生成器后，我们看到*R-NESt* 和*CLIP-NESt* 都超过了*MEIR* [11]。该方法可能是影响因素之一，但我们也应考虑数据规模的差异。*MEIR* 仅包含82,156个真实实例和

[11]https: //spacy.io/models/en#en_core_web_trf-acc uracy

表2：在三个 NewsCLIPings 数据集上训练和评估的不同检测器与编码器的比较。我们报告真实与伪造的图像-标题对的二元准确率和每类召回率。我们提出的基于Transformer的检测器（DT-Transformer）始终优于原始 NewsCLIPings 检测器，使用 CLIP ViT-L/14 的特征还能进一步提升性能。我们分别报告总体准确率，以及各类别（真实与伪造配对）的命中率，或特异度和敏感度。

合成误导者	检测器	编码器	准确率	真实	伪造
NC/Bal	[9]	CLIP ViT-B/32	60.23	70.07	50.39
NC/Bal	DT-Transformer	CLIP ViT-B/32	65.67	73.73	57.60
NC/Bal	DT-Transformer	CLIP ViT-L/14	77.09	78.58	75.61
NC/T-T	[9]	CLIP ViT-B/32	69.39	74.09	64.69
NC/T-T	DT-Transformer	CLIP ViT-B/32	71.63	75.56	67.69
NC/T-T	DT-Transformer	CLIP ViT-L/14	84.12	87.45	80.79
NC/I-T	[9]	CLIP ViT-B/32	66.98	75.43	58.53
NC/I-T	DT-Transformer	CLIP ViT-B/32	69.12	67.51	70.73
NC/I-T	DT-Transformer	CLIP ViT-L/14	80.98	84.62	77.33

57,940个伪造实例，而其他错误信息生成器生成的数据中约有1,2M个伪造数据点，这可能不足以正确训练检测器。我们还观察到，所提出的*CLIP-NESt-alt* 超过了所有其他多模态NEI和OOC方法，达到了56.9%的准确率。此外，将R-NESt与CSt-alt结合也取得了相同的分数，但将NC/I-T与R-NESt结合时表现不佳，因为NC/I-T包含226,564个样本，并且需要对真实样本和R-NESt样本（各1M个）进行大幅下采样，从而得到明显更小的数据集。最后，我们观察到，我们的混合*CLIP-NESt-alt + CSt-alt* 方法能够取得最高的多模态准确率（58.1%）。

尽管如此，我们也认识到存在一些问题和局限性。首先，即使是表现最好的方法，也难以准确识别伪造对，其伪造命中率（敏感度）低于 50%，而真实命中率（特异度）则较高。这表明，多模态错误信息检测任务远比以往研究所认为的更具挑战性。例如，先前研究在 COSMOS 数据集上报告了高于 88% 的得分 [6, 14, 15] ，但使用的评估协议存在问题（详见第 3.1.4 节）。最后，我们发现，在本应是多模态的任务中，单模态纯文本方法（如 *CLIP-NESt-C* 和 *R-NESt*、*R-NESt + CSt-alt*）能够胜过多模态方法；后者的得分为 59.1%。这表明数据集中存在单模态偏差，需要在未来的研究中加以解决。

5 结论

在本研究中，我们探讨了多模态错误信息检测（MMD）任务。更具体而言，我们考察并比较了多种为 MMD 生成训练数据（合成虚假信息生成器）的方法，即生成脱离上下文的图文对（OOC）或跨模态命名实体不一致（NEI）的方法。我们开展了一项比较研究，并在由现实世界多模态错误信息构成的 COSMOS 基准上评估了所有合成虚假信息生成器。比较研究表明，NEI 方法在 COSMOS 基准上的平均表现往往优于 OOC 方法。此外，我们提出的*CLIP-NESt-alt* 方法在 NEI 和 OOC 方法中取得了最高的多模态准确率（56.9%）；相较于两类方法中各自表现次优的方法，分别具有 2.15% 和 2.7% 的优势。进一步而言，我们假设，现实世界中的错误信息并非仅由 OOC 或 NEI 实例单独涵盖，而是需要二者兼具。所提出的混合方法（*CLIP-NESt-alt +CSt-alt*）取得了最高的多模态准确率（58.1%），从而验证了这一假设；相较于*CLIP-NESt-alt*提升了 2.47%。

尽管如此，较低的敏感度分数（表3）表明，在修正后的评估协议下，多模态虚假信息检测比以往研究所认为的更具挑战性 [6] ，因此亟需开展大量后续研究。未来研究可考虑整合外部证据 [16, 23] 或知识图谱 [24] ，不仅提高检测准确率，还可开发新的合成虚假信息生成器，以生成更逼真的合成训练数据，从而训练出更优秀的检测器。此外，尝试不同的模态融合技术也能进一步提升性能 [25, 26]。此外，我们的实证结果表明，NEI误导信息生成器往往会引入单模态偏差，导致单模态检测器能够与多模态检测器相抗衡，甚至表现更佳。命名实体操纵可能会产生某些语言模式、偏差或捷径，从而降低视觉信息的重要性。未来研究可探索开发生成去偏差NEI的方法，或减少单模态偏差的学习策略 [27]。此外，任务特定的模态融合方法或许有助于缓解

表3：在 COSMOS 基准上评估的多种合成误导方法的比较研究。我们使用 DT-Transformer 和 CLIP ViT L/14编码器。我们报告单模态（仅图像、仅文本）和多模态检测器的准确率，以及各类别（真实与伪造配对）的多模态命中率，或特异度和敏感度。**粗体** 表示总体最高准确率，下划线表示最高多模态准确率。

Type	合成误导者	仅图像 准确率	仅文本 准确率	多模态 准确率	多模态 真实	多模态 伪造
OOC	RS-C	50.0	50.0	52.2	92.4	12.0
	RSt-C	50.0	50.0	53.8	93.5	14.0
	RSt-I	50.2	50.0	52.3	96.6	8.0
	RSt-alt	50.0	50.0	53.9	92.0	15.9
	CSt-C	50.0	48.5	53.1	70.7	35.5
	CSt-I	49.5	50.0	54.7	83.4	25.9
	CSt-alt	52.3	51.5	55.0	76.6	33.4
	NC/Bal [9]	50.0	50.0	54.1	52.8	55.3
	NC/T-T [9]	48.8	50.0	53.7	66.1	41.3
	NC/I-T [9]	49.2	50.0	55.4	84.6	26.2
NEI	MEIR [11]	49.4	50.5	53.9	90.5	17.3
	R-NESt	50.1	58.5	55.7	90.5	20.8
	CLIP-NESt-C	50.0	58.1	55.3	75.1	35.5
	CLIP-NESt-I	50.0	54.1	55.2	81.2	29.3
	CLIP-NESt-alt	50.0	56.7	56.9	81.7	32.2
混合	R-NESt +CSt-alt	51.8	**59.1**	56.9	89.3	24.5
	R-NESt +NC/I-T	51.5	57.2	54.4	90.5	18.1
	CLIP-NESt-alt +CSt-alt	51.3	52.8	<u>58.1</u>	74.4	41.9

这一挑战 [28]。最后，单模态偏差可能并非源于训练过程，而是来自评估数据集。COSMOS 基准的收集并未遵循旨在明确增加单模态架构难度的特定标准。未来研究可探索并制定相关规则和标准，以收集更可靠的真实世界多模态虚假信息基准。

6 致谢

本研究部分由项目“vera.ai：人工智能辅助验证”资助，资助协议编号为 101070093。

参考文献

[1] Muhammad F Mridha、Ashfia Jannat Keya、Md Abdul Hamid、Muhammad Mostafa Monowar 和 Md Saifur Rahman。基于深度学习的假新闻检测：全面综述。*IEEE Access*，9:156151–156170，2021年。[2] Md Shohel Rana、Mohammad Nur Nobi、Beddhu Murali 和 Andrew H Sung。深度伪造检测：系统性文献综述。*IEEEAccess*，2022年。[3] Yiyi Li 和 Ying Xie。一张图片胜过千言万语吗？图像内容与社交媒体互动的实证研究。*Journal of Marketing Research*，57(1):1–19，2020年。[4] Eryn J Newman、Maryanne Garry、Daniel M Bernstein、Justin Kantner 和 D Stephen Lindsay。无证明力的照片（或文字）会增强真实感。*PsychonomicBulletin & Review*，19:969–974，2012年。[5] Firoj Alam、Stefano Cresci、Tanmoy Chakraborty、Fabrizio Silvestri、Dimiter Dimitrov、Giovanni Da San Martino、Shaden Shaar、Hamed Firooz 和 Preslav Nakov。多模态虚假信息检测综述。*arXiv*预印本 *arXiv:2103.12541*，2021年。[6] Shivangi Aneja、Chris Bregler 和 Matthias Nießner。Cosmos：利用自监督学习捕捉脱离语境的错误信息。*arXiv*预印本 *arXiv:2101.06278*，2021年。

[7] Eric Müller-Budack、Jonas Theiner、Sebastian Diering、Maximilian Idahl 和 Ralph Ewerth。利用跨模态实体一致性指标分析真实场景中的新闻多模态数据。载于《2020年国际多媒体检索会议论文集》，第16–25页，2020年。[8] Ayush Jaiswal、Ekraam Sabir、Wael AbdAlmageed 和 Premkumar Natarajan。利用图像与文本的联合嵌入评估多媒体语义完整性。载于第25届ACM国际多媒体会议论文集，第1465–1471页，2017年。[9] Grace Luo、Trevor Darrell 和 Anna Rohrbach。Newsclippings：自动生成脱离语境的多模态媒体。arXiv预印本 arXiv:2104.05893，2021年。[10] Giscard Biamby、Grace Luo、Trevor Darrell 和 Anna Rohrbach。Twitter-comms：检测气候、COVID及军事领域的多模态错误信息。arXiv预印本 arXiv:2112.08594，2021年。[11] Ekraam Sabir、Wael AbdAlmageed、Yue Wu 和 Prem Natarajan。深度多模态图像再利用检测。载于第26届ACM国际多媒体会议论文集，第1337–1345页，2018年。[12] Fuxiao Liu、Yinghan Wang、Tianlu Wang 和 Vicente Ordonez。视觉新闻：新闻图像描述生成的基准测试与挑战。arXiv预印本 arXiv:2010.03743，2020年。[13] Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sastry、Amanda Askell、Pamela Mishkin、Jack Clark 等。通过自然语言监督学习可迁移视觉模型。载于《国际机器学习会议论文集》，第8748–8763页。PMLR，2021年。[14] Tankut Akgul、Tugce Erkilic Civelek、Deniz Ugur 和 Ali C Begen。强化版 Cosmos：用于廉价伪造的廉价检测器。载于第12届ACM多媒体系统会议论文集，第327–331页，2021年。[15] Muhannad Alkaddour、Abhinav Dhall、Usman Tariq、Hasan Al Nashash 和 Fares Al-Shargie。用于脱离语境的图像描述检测的情感感知分类器。载于第30届ACM国际多媒体会议论文集，第7180–7184页，2022年。[16] Sahar Abdelnabi、Rakibul Hasan 和 Mario Fritz。利用在线资源对脱离语境的图像进行开放领域、基于内容的多模态事实核查。载于IEEE/CVF计算机视觉与模式识别会议论文集，第14940–14949页，2022年。[17] Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova。BERT：用于语言理解的深度双向Transformer预训练。arXiv预印本 arXiv:1810.04805，2018年。[18] Piotr Bojanowski、Edouard Grave、Armand Joulin 和 Tomas Mikolov。利用子词信息丰富词向量。计算语言学协会会刊，5:135–146，2017年。[19] Mingzhen Huang、Shan Jia、Ming-Ching Chang 和 Siwei Lyu。利用视觉语言模型进行文本图像去语境化检测。载于《ICASSP 2022—2022年IEEE国际声学、语音与信号处理会议（ICASSP）论文集》，第8967–8971页。IEEE，2022年。[20] Mitchell Wortsman、Gabriel Ilharco、Jong Wook Kim、Mike Li、Simon Kornblith、Rebecca Roelofs、Raphael Gon-tijo Lopes、Hannaneh Hajishirzi、Ali Farhadi、Hongseok Namkoong 等。零样本模型的稳健微调。载于IEEE/CVF计算机视觉与模式识别会议论文集，第7959–7971页，2022年。[21] Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N Gomez、Łukasz Kaiser 和 Illia Polosukhin。Attention Is All You Need。《神经信息处理进展》，30，2017年。[22] Shivangi Aneja、Cise Midoglu、Duc-Tien Dang-Nguyen、Michael Alexander Riegler、Paal Halvorsen、Matthias Nießner、Balu Adsumilli 和 Chris Bregler。MMSys' 21廉价伪造品检测大挑战。arXiv预印本 arXiv:2107.05297，2021年。[23] Max Glockner、Yufang Hou 和 Iryna Gurevych。缺少反证使自然语言处理事实核查无法真实应对错误信息。arXiv预印本 arXiv:2210.13865，2022年。[24] Lihui Liu、Houxiang Ji、Jiejun Xu 和 Hanghang Tong。知识图谱事实核查的比较推理。载于《2022年IEEE国际大数据会议（Big Data）论文集》，第2309–2312页。IEEE，2022年。[25] Rina Kumari 和 Asif Ekbal。AMFB：用于多模态假新闻检测的基于注意力的多模态因子化双线性池化。Expert Systems with Applications，184:115412，2021年。[26] Chuanming Yu、Yinxue Ma、Lu An 和 Gang Li。BCMF：用于假新闻检测的双向跨模态融合模型。Information Processing & Management，59(5):103063，2022年。[27] Remi Cadene、Corentin Dancette、Matthieu Cord、Devi Parikh 等。RUBi：减少视觉问答中的单模态偏差。《神经信息处理进展》，32，2019年。[28] Peiguang Li、Xian Sun、Hongfeng Yu、Yu Tian、Fanglong Yao 和 Guangluan Xu。面向实体的多模态对齐与融合网络，用于假新闻检测。IEEE多媒体汇刊，24:3455–3468，2021年。