检测和巩固多模态媒体操纵

Rui Shao1,2*, Tianxing Wu2 , Ziwei Liu2 †
[1] 哈尔滨工业大学（深圳)计算机科学与技术学院
[2] 南洋理工大学 S-Lab
shaorui@hit.edu.cn, {tianxing.wu, ziwei.liu}@ntu.edu.sg
https://github.com/rshaojimmy/MultiModal-DeepFake

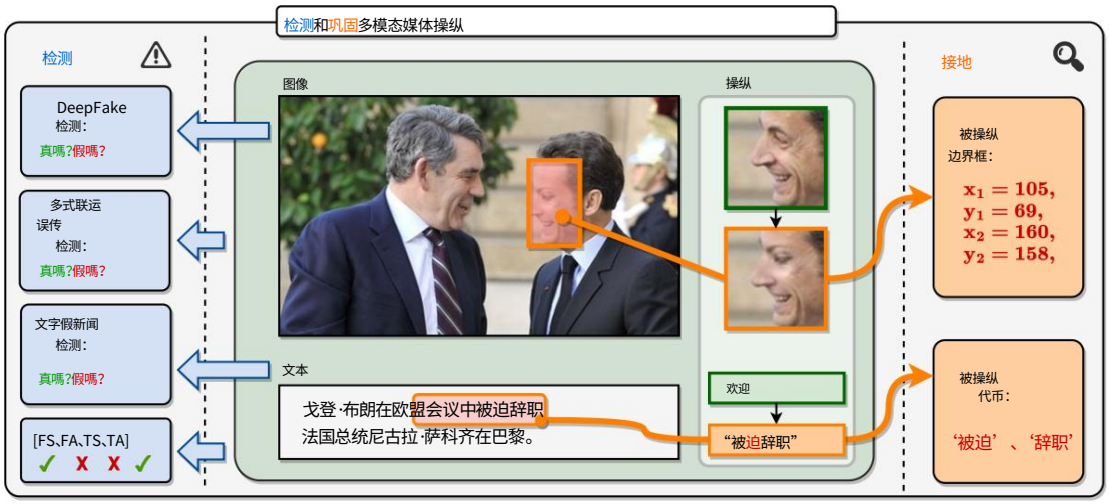

图 1. 与现有的单模态伪造检测任务不同， DGM4不仅对输入进行真/假分类
图像-文本对,而且还尝试检测更多细粒度的操作类型和地面操作的图像框和文本标记。
除了二元分类之外,它们还提供了有关操纵检测的更全面的解释和更深入的理解。
（FS:脸部交换操作、FA:脸部属性操作、TS:文本交换操作、TA:文本属性操作）

抽象的

虚假信息已成为一个紧迫问题。虚假媒体,无论是视觉形式还是文字形式,在互联网上广泛存在。网络。虽然各种深度伪造检测和文本假新闻目前已提出的检测方法仅针对基于二分类的单模态伪造,更不用说分析和推理细微伪造

跨不同模态的痕迹。在本文中,我们重点介绍了多模态假媒体的一个新研究问题,即检测和定位多模态媒体

操纵（DGM4 ） 。DGM4的目的不仅在于检测多模态媒体的真实性,同时也为操纵的内容（即图像边界框和文本标记)奠定基础,这需要对多模态媒体操纵进行更深入的推理。为了支持大规模调查，

我们构建了第一个DGM4数据集,其中图像-文本对可以通过各种方法进行操纵,具有丰富的注释

各种操作的结合。此外,我们提出了一个新颖的层次化多模态操控推理tTransformer （HAMMER)能够充分捕捉细粒度的不同模态之间的相互作用。HAMMER执行1)操纵感知对比学习，

两个单模态编码器作为浅层操作推理,以及2)多模态感知交叉注意聚合器作为深度操纵推理。集成了专用的操纵检测和接地头

基于多模态信息的交互,从浅层到深层地进行分类。最后,我们为这一新的研究问题建立了广泛的基准并建立了严格的评估指标。综合实验表明

我们的模型的优越性;一些有价值的观察结果也有助于未来多模式研究媒体操纵。

1. 简介

随着深度生成模型的最新进展,越来越多的超现实面部图像或视频可以被自动

*这项工作是在南洋理工大学 S-Lab 完成的
†通讯作者

表 1. 提出的DGM4与现有的图像和文本伪造检测相关任务的比较。

问题设定	图像伪造		文本伪造	多式联运
	检测	接地检测	接地伪造检测	
DeepFake 检测[30, 60]				
文本虚假新闻检测[53,58]				
多模态错误信息检测[1, 29]				
DGM4				

自动化生成,这会导致各种安全问题[40–46,48,57] ,例如严重的深度伪造问题[8,
15, 24, 39, 47]在视觉媒体上传播大量捏造的内容。这种威胁引起了计算机视觉领域的极大关注
社区和各种深度伪造检测方法
被提出。随着大型语言模型的出现，
例如 BERT [7]、GPT [36]、大量文本假新闻[53, 58]
很容易生成恶意的误导性信息,在文本媒体上传播。自然语言处理
(NLP) 领域非常关注这个问题,并且

提出了多种文本假新闻检测方法。

与单一模式相比,多模式媒体
（以图文对的形式)传播的信息范围更广,对我们的日常生活影响更大。因此,多模态伪造媒体往往更具危害性。为了应对

为应对这一新威胁,本文提出了一个更易于解释和解读的解决方案,即检测和接地多模态媒体

操作（DGM4 ）。如表1和图1所示，
DGM4带来了两个挑战： 1)虽然当前
deepfake 检测和文本假新闻检测方法
旨在检测单模态伪造品， DGM4
要求同时检测伪造的存在
2)除了二进制之外,还有图像和文本模式
像当前的单模伪造检测这样的分类，
DGM4进一步考虑了基础操作图像边界框 (bbox) 和文本标记。这

意味着现有的单模态方法无法
这一新颖的研究问题。更全面和
有必要对两种模态之间的操纵特征进行更深入的推理。请注意,一些多模态错误信息研究[1, 29]已经发展起来。但

它们只需要确定多模态媒体的二进制类,更不用说操作基础了。

为了促进DGM4的研究 ，本文贡献
第一个大规模DGM4数据集。在这个数据集中,我们研究
一种具有代表性的多模式媒体形式,以人为本
新闻。它通常涉及有关政客和名人的错误信息,造成严重的负面影响。
我们开发了两种不同的图像处理（即

面部交换/属性操作)和两种文本操作（即文本交换/属性操作)方法

形成多模态媒体操控场景。丰富
提供用于检测和基础的注释,包括二进制标签、细粒度操作类型、操作的图像框和操作的文本标记。

与原始图像-文本对相比,经过处理的
多模态媒体必然会留下操纵痕迹
在操纵的图像区域和文本标记中。所有
这些轨迹一起改变了跨模态相关性,并且
从而导致两种模态之间的语义不一致。因此,推理两种模态之间的语义相关性
图像和文本为检测提供了提示,并
多模态操控的基础。为此,受现有视觉语言表征学习的启发

文献[19, 20, 35],我们提出了一种新颖的层次化多模态操作推理转换器（HAMMER）
解决DGM4 。 为了充分捕捉
图像和文本， HAMMER 1)通过可感知操作的对比学习对齐图像和文本嵌入
两个单模编码器之间的浅层操作
推理和2)通过以下方式聚合多模态嵌入
多模态聚合器的模态感知交叉注意力
作为深度操纵推理。基于交互
不同层次的多模态嵌入、专用的操作检测和接地头被分层集成,以检测二进制类、细粒度的操作类型和接地操作图像框、操作文本标记。这种分层机制有助于

更细粒度和全面的操纵检测和定位。本文的主要贡献如下：

· 我们引入了一个新的研究问题,即检测和
基础多模式媒体操控（DGM4 ），
目的是检测并确定以人为本的新闻图像-文本对中的操纵行为。

· 我们贡献了包含样本的大规模DGM4数据集
通过两种图像处理和两种文本处理方法生成。为

检测并证实不同的操作。

· 我们提出了一种强大的HierArchical Multi-modal
操纵推理变压器（HAMMER） 。
综合基准是基于严格的评估协议和指标建立的。广泛的定量和

定性实验证明了它的优越性。

2.相关工作

DeepFake 检测。为了检测人脸伪造图像,当前的 DeepFake 检测方法都是基于空间和频域构建的。基于空间的 DeepFake 检测方法利用空间视觉线索,例如混合伪影[22]、纹理特征[5, 60, 62]、3D 信息[62]、块一致性[61]和噪声特征。

基于频率的深度伪造检测方法
检测频谱伪影,例如高频分量
从离散傅里叶变换（DFT） [11]分解而来，
离散余弦变换 (DCT) [34] 产生的细微频率差异,隐藏的上采样伪影

基于相位谱[26]和基于频率的度量学习[21]。上述大多数深度伪造检测方法

只对图像媒体进行二元分类,而不是
提及跨多模态的操作基础。
多模态错误信息检测。现有的一些研究成果研究了多模态错误信息的检测[1、2、16、18、29、54] 。其中一些研究的是小规模的人工生成的多模态假新闻[16、18、54] ，

而另一些人则处理断章取义的错误信息，
将真实图像与另一段交换的文本配对,无需进行图像和文本处理[1, 2, 29]。所有这些
方法仅基于简单的二元分类
图像与文本的相关性。相比之下， DGM4研究大规模机器生成的多媒体操作，

在实践中更接近于网络上广泛的错误信息。
此外， DGM4不仅需要对二元分类进行操作检测,还需要对多模态操作进行具有更多解释的操作基础分析。

3. 多模态媒体操作数据集

现有的大多数错误信息数据集集中于
单模态图像伪造[8, 15, 23, 39]或文本伪造
[49, 53, 58]. 一些多模态数据集已经建立,但是
它们通常含有少量人为产生的
假新闻[6,16]或脱离上下文的二元对[2,29]
伪造检测。为了更好地促进提出的新研究问题,我们提出了DGM4数据集,研究大规模机器生成的多模态媒体操纵。

DGM4数据集通过多种操作构建
图像和文本模态的技术。所有样本
带有丰富、细粒度的标签注释,可以同时实现
媒体操纵的检测与根除。

3.1. 源数据收集

在所有形式的多模式媒体中,我们特别
关注以人为本的新闻,考虑到其巨大的
公众影响力。因此,我们基于以下因素开发数据集：
VisualNews 数据集[25],收集了来自现实世界新闻来源的大量图像-文本对（《卫报》、
BBC、今日美国和华盛顿邮报）。为了构建一个以人为本、具有有意义背景的场景,我们
进一步对图像和文本模态进行数据过滤,只保留合适的对作为源

池 O = {po|po = (Io, To)} 用于操作。

3.2. 多模态媒体操控

我们对两者都采用了两种有害的操纵
图像和文本模式。“交换”类型旨在包括
相对全局的操作痕迹,而“属性”类型
引入了更细粒度的局部操作。然后将经过处理的图像和文本随机混合到

原始样品,共计 8 个假货和一个正品
操纵类别。操纵类别的分布
图2 （a)展示了部分样本。
换脸 (FS) 操作。在此操作类型中，
主角身份被互换攻击
将自己的脸与另一个人的脸进行对比。我们采用了两种有代表性的脸部交换方法,SimSwap [4]和 InfoSwap [12]。
对于每个原始图像Io,我们选择两种方法之一来交换最大的人脸 I f
这 随机来源
面对我名人 来自 CelebA-HQ 数据集[17],生成一张人脸
交换操作样本是。MTCNN bbox 的
然后将交换面ybox = {x1, y1, x2, y2}保存为接地注释。

人脸属性 (FA) 操纵。作为一种更细粒度的图像操纵场景,人脸属性操纵试图
在保留身份的同时操纵主角面部的情绪。例如,如果

原始面部表情是微笑的,我们故意将其编辑成相反的情绪,例如愤怒的面部表情。为了实现这一点,我们首先
预测对齐人脸的原始面部表情 这
使用基于 CNN 的网络,然后将脸部朝向
使用基于 GAN 的方法表达相反情绪,HFGI [52]
和 StyleCLIP [33]。在获得经过处理的脸部后
如果情绪摇滚， 我们将其重新渲染回原始图像Io以获得
还提供了操纵的样本Ia.Bboxybox 。
文本交换 (TS) 操作。在这种情况下,文本是
通过改变其整体语义来操纵,同时保留有关主角的单词。给定原始标题To,我们使用命名实体识别 (NER) 模型来

提取人员姓名作为查询“PER”。然后我们检索
不同的文本样本 T 这 包含相同的“PER”实体
来自源语料库 O. T 这 然后选择作为操纵
文本Ts。注意,我们使用 Sentence-BERT [37]计算每个文本的语义嵌入,并且只接受
电视这 与 To 具有较低的余弦相似度。这确保了
检索到的文本在语义上与 To 不一致,因此
所获取的
对pm = (Io, Ts)进行处理。之后,给定 M
Ts中的文本标记,我们用 M 维来注释它们
独热向量ytok = {yi}M 我=1, 其中yi ∈ {0, 1} 表示
Ts中的第 i 个 token 是否被操纵。
文本属性 (TA) 操作。虽然新闻是
相对客观的媒体形式,我们观察到相当一部分新闻样本po ∈ O 在文本 To 中仍然带有情感偏见,如图2 (d) 所示。

恶意操纵文本属性,尤其是其情感倾向,可能更具危害性,也更难

检测不到,因为它会导致较少的跨模态不一致性
比文本交换操作更容易。为了反映这种特定情况,我们首先使用
RoBERTa [27]模型来拆分 cap-

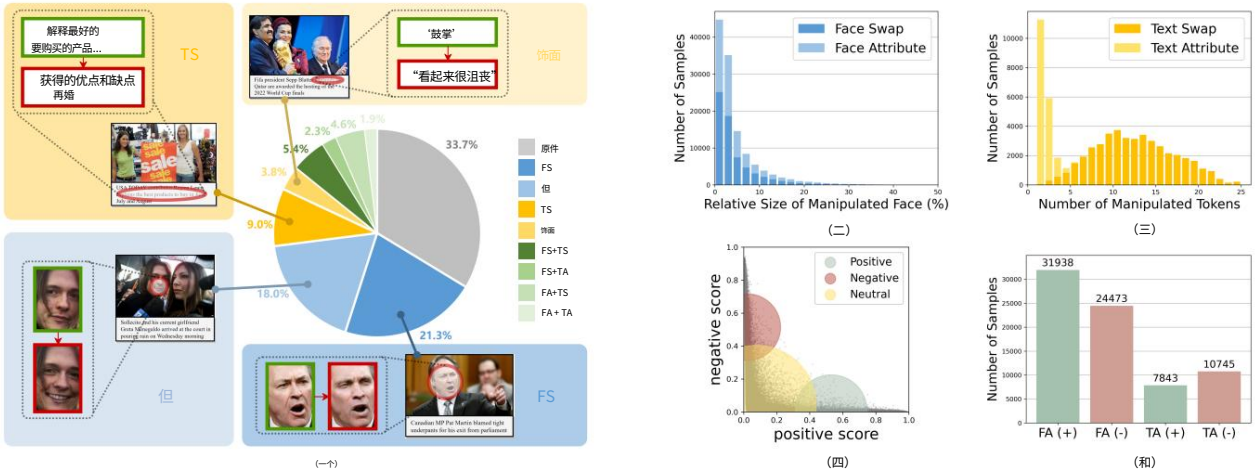

图 2. DGM4数据集的统计数据。(a) 操作类别的分布;(b) 大多数图像的操作区域尺寸较小，
尤其对于人脸属性操作;(c) 文本属性操作的操作标记少于文本交换操作;(d)
源池中文本情感分数的分布。(e)针对每个面部/文本属性方向操纵的样本数量。

分为积极、消极和中性情绪语料库：
{O+, O−, Oneu}。根据[50],我们用以下公式替换所有情绪
原文中的词语To用相反的情绪
在我们自己的语料库{O+, O−}上训练的B-GST模型生成的文本,得到Ta。与文本
交换操作类似,所有文本标记也都用真值注释

矢量ytok。
组合和扰动。一旦所有单模态
操作完成后,将得到的操作样本Is， Ia， Ts和Ta与原始样本合并

(Io, To)对。这形成了一个具有完整操纵类型的多模态操纵媒体池:P = {pm|pm
=
(Ix, Ty), x, y ∈ {o, s, a}}。池中的每一对pm都提供了一个二进制标签
ybin,这是一种细粒度的操作
类型注释ymul、上述注释ybox和
ytok. ybin描述图像-文本对pm是否真实
或假的,并且ymul = {yj}$_{j=1}^{4}$ 是一个四维向量,表示第 j 个操作类型（即
FS、FA、TS、
TA)出现在下午。为了更好地反映现实世界的情况
由于操作痕迹可能被噪音掩盖,我们
对 50% 的媒体采用随机图像扰动
池P,例如JPEG压缩,高斯模糊等。

3.3. 数据集统计

DGM4数据集的总体统计数据如下
图2 (a)。它总共包含230k 个新闻样本,包括 77,426 个原始图像-文
本对和 152,574 个经过处理的图像-文本对。经过处理的图像-文本
对包含 66,722 个换脸图像
操作,56,411 人脸属性操作,43,546
文本交换操作和 18,588 个文本属性操作。约 1/3 的操作图像和约 1/2 的

经过处理的文本组合在一起形成 32,693
混合操作对。由于图像和文本属性都可以朝着两个相反的情绪方向进行编辑,因此
我们故意保持平衡的比例来创建
情绪平衡的数据集,如图2 (e) 所示。
此外,从图2 (b)-(c)可以看出
大多数图像的操纵区域和数量
操纵的文本标记相对较小。这表明DGM4数据集提供了更具挑战性的

与现有的深度伪造和多模态错误信息数据集相比,伪造检测场景更加出色。

4. 锤子

解决DGM4 ，如图3 所示,我们提出了一个层次化的多模态
操作推理
tTransformer （HAMMER）,由两个单模编码器（即图像编
码器Ev、文本编码器Et)组成，
多模态聚合器 F,以及专用操作
检测和接地头（即二元分类器Cb，
多标签分类器Cm、 BBox 检测器Dv和 Token
检测器Dt）。所有这些单模编码器和多模聚合器都是基于基于转换器的架构构
建的[51]。如上所述,建模语义

相关性并捕捉语义不一致
两种方式可以促进检测和接地
多模态操控。然而,存在两个挑战1)如第3.3节所述,如图2 (b) 所示 -

（c)大部分多模态操作都是轻微的
并且很微妙,定位在一些小尺寸的脸部和一些单词
2)存在大量视觉和文本噪音[20]
网络上的多模态媒体。因此,一些语义
操纵造成的不一致性可以被忽略
或被噪声覆盖。这需要更细粒度的多模态相关性推理。为此,我们设计了

HAMMER进行分层操作推理,从浅层探索多模态交互

到深层次,以及分层操纵检测和基础。在浅层操纵推理中,我们在图像和文本之间
进行语义对齐

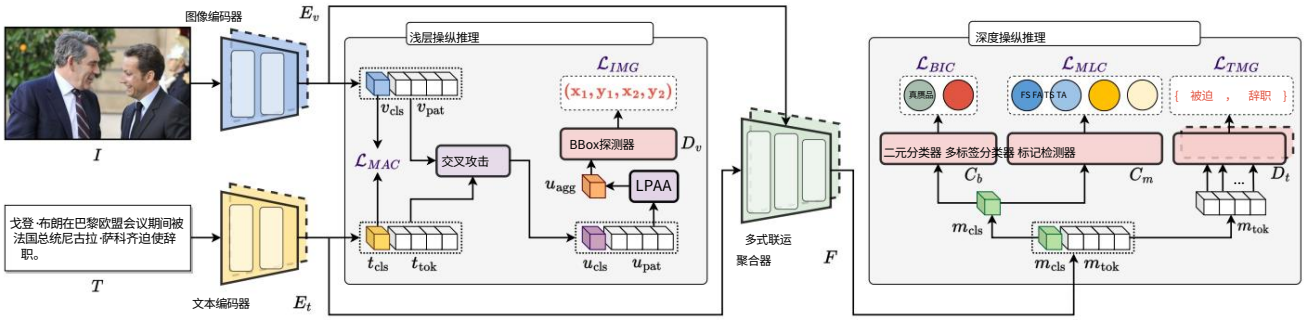

图 3. 所提出的HAMMER 概览。它1)在浅层操作推理中通过图像编码器Ev 和文本编码器Et之间的操作感知对比学习来对齐图像和文本嵌入； 2)在深度操作推理中通过多模态聚合器 F 的模态感知交叉注意进一步聚合多模态嵌入。基于不同层次的交互多模态嵌入,集成了各种操作检测和接地头（多标签分类器Cm、二元分类器Cb、 BBox 检测器Dv和 Token 检测器Dt)以分层执行其任务。带有虚线的模块表示它们分别是图像编码器、文本编码器、多模态聚合器和 Token 检测器的相应动量版本。

通过操作感知对比损失LMAC检测文本嵌入,并在图像操作基础损失LIMG 下进行操作的 bbox 基础处理。在深度操作推理中,基于多模态聚合器生成的更深层次的交互多模态信息,我们使用二元分类损失LBIC检测二元类,使用多标签分类损失LMLC检测细粒度操作类型,并通过文本操作基础损失 LTMG 基础操作文本标记。通过结合以上所有损失,操作推理以分层方式进行,从而形成以下联合优化框架：

L = LMAC + LIMG + LMLC + LBIC + LTMG (1)

4.1. 浅层操纵推理

给定一个图像-文本对 (I, T) P,我们通过 Image Encoder 中的自注意层和前馈网络将图像 I 拼接并编码为图像嵌入序列,即Ev(I) = {vcls, vpat},其中 vcls是 [CLS] token 的嵌入， vpat = {v1, ..., vN}是 N 个相应图像块的嵌入。Text Encoder 提取 T 的文本嵌入序列,即Et(T) = {tcls, ttok},其中tcls是 [CLS] token 的嵌入， ttok = {t1, ..., tM}是 M 个文本 token 的嵌入。

操作感知对比学习。为了帮助两个单模态编码器更好地利用图像和文本的语义相关性,我们通过跨模态对比学习对齐图像和文本嵌入。然而,一些微妙的多模态操作会导致两种模态之间出现轻微的语义不一致,而正常的对比学习很难发现这种不一致。为了强调操作造成的语义不一致， HAMMER提出了对图像和文本嵌入进行操作感知对比学习。与正常的跨模态对比学习不同,它将原始图像-文本对的嵌入拉近,而只将这些嵌入推近

除了不匹配对之外,操纵感知对比学习还会推开操纵对的嵌入,从而进一步强调它们产生的语义不一致性。根据 InfoNCE 损失[31],我们用以下公式来表示图像到文本的对比损失： exp(S(I, T +)/τ)

$$Lv2t(I, T+, T-) = -Ep(I,T) \log \frac{\exp(S(I, T+)/\tau)}{\sum_{k=1}^{K} \exp(S(I, T_k)/\tau)} \quad (2)$$

其中 τ 是温度超参数,T − {T −与 I 匹配,并且属于经过操作的图像-文本对。由{T1, ..., TK} 是一组不属于

于[CLS] token 作为整个图像和文本的语义表示,我们使用两个投影头hv和 ht将两种模态的 [CLS] token 映射到低维（256)嵌入空间进行相似度计算： S(I, T) = hv(vcls) Th t(tcls)。受 MoCo [14] 的启发,我们分别学习了动量单模态编码器 E E t （指数移动平均版本)和两种模态的动量投影头。两个队列用于存储最近的 K 个图像-文本对嵌入。这里tcls是来自文本动量编码器的 [CLS] 标记,ht(tcls)表示来自文本动量投影头的投影文本嵌入。类似地,文本到图像的对比损失如下： exp(S(T, I+)/τ)

$$Lt2v(T, I+, I-) = -Ep(I,T) \log \frac{\exp(S(T, I+)/\tau)}{\sum_{k=1}^{K} \exp(S(T, I_k)/\tau)} \quad (3)$$

I− 表示与 T 不匹配 {I −其中1, ..., I− K} 是 K 个最近负面消息的队列的图像样本以及属于被操纵的图像-文本对的图像样本。S(T, I) = ht(tcls) Th v(vcls)。受[56] 的启发,为了在每个单一模态中保持合理的语义关系,我们进一步在两种模态中开展模态内对比学习。我们结合所有损失,形成如下操纵感知对比损失： 1 LMAC = 4

$$-[Lv2t(I, T+, T-) + Lt2v(T, I+, I-) + Lv2v(I, I+, I-) + Lt2t(T, T+, T-)] \quad (4)$$

操纵图像边界框基础。如上所述,FS 或 FA 交换身份或编辑属性

图像中的人脸。这会改变它们与人名或情绪等相应文本的相关性。鉴于

因此,我们认为被操纵的图像区域可以
通过查找与文本嵌入不一致的局部补丁来定位。在这方面,我们执行

图像和文本嵌入之间的交叉注意力以获得包含图像-文本相关性的补丁嵌入。

注意函数[51]在规范化的查询上执行
（Q）、关键（K)和价值（V)特征如下：

注意力机制（Q,K,V)=Softmax（KT Q/√D)V（5）

在这里,我们将图像嵌入与文本嵌入进行交叉关注,将 Q 视为图像嵌入,将 K 和 V 视为
文本嵌入如下：

Uv(I) = Attention(Ev(I), Et(T), Et(T)) + Ev(I),其中Uv(I) (6)

= {ucls, upat}。upat = { u1, ..., uN}是 N
图像补丁嵌入与文本信息交互。
与 [CLS] token ucls 不同， patch token upat是通过位置编码生成的[51]。
这意味着它们具有更丰富的局部空间信息,因此更适合操纵图像 bbox grounding。基于此

分析,我们提出局部块注意力聚合
（LPAA)通过
注意机制。这种聚合是通过
使用upat交叉关注 [AGG] 令牌,如下所示：

uagg = 注意([AGG],四,四) (7)

与以前的工作[59]直接使用 [CLS] to-token 进行 bbox 预测不同,我们执行操纵的 bbox
基于注意力聚合嵌入uagg 的接地。具体来说,我们将uagg输入到 BBox 检测器中
Dv并计算图像处理接地损耗
结合正常损失和广义交集
并集（IoU)损失[38]如下：

LIMG =E(I,T) P [Sigmoid(Dv(uagg)) − ybox
+ LIoU(Sigmoid(Dv(uagg)) − ybox)] (8)

4.2. 深度操纵推理

操纵 token 接地比操纵 bbox 接地更艰巨,因为它需要更深入的分析和

图像与文本相关性推理。例如
例如,如图3 所示,我们能够检测到
仅操纵 T 中的标记,即“force”和“resign”
当我们意识到这些负面词语不匹配时
积极情绪（即笑脸）。此外,我们
需要总结多模态信息来检测细粒度的操作类型和二进制类。这要求在此阶段进行全面的信息总结。

为此,我们提出了深度操纵推理。
操纵文本标记基础。为了更深入地建模
多模态交互,如图3 所示,我们提出模态感知交叉注意力机制,进一步引导文本嵌入Et(T) 与图像嵌入Ev(I)进行交互

通过多模态聚合器 F 中的多个交叉注意层。这将生成聚合多模态嵌入F(Ev(I), Et(T)) = {mcls, mtok}。具体来说，

mtok = {m1, ..., mM}表示与 T 中每个 token 对应的更深层次的聚合嵌入。在此阶段，
T 中的每个 token 都经过了多次自注意力
Et中的层和 F 中的交叉注意层。这样，
mtok中的每个 token 嵌入不仅可以完全探索
文本的上下文信息,还与
图像特征,适合操纵的文本标记基础。此外,基础操纵的标记等于

将每个 token 标记为真或假。这类似于 NLP 中的序列标记任务。值得注意的是,与主要在文本模态中研究的现有序列标记任务不同,这里的操纵文本 token 基础可以被视为一种新颖的
多模态序列标记,因为每个标记都是相互交互的
包含两种模态信息。在本例中,我们使用一个 Token
检测器Dt预测mtok中每个 token 的标签,并
计算交叉熵损失如下：

Ltok = E(I,T) P H(Dt(talk), ytalk) (9)

其中H(·) 是交叉熵函数。如上所述，
网络上的新闻通常充斥着与
配对图像[20]。为了缓解对嘈杂文本的过度拟合，
如图3 所示,我们进一步学习动量版本
分别是多模式聚合器和令牌检测器，
表示为 F^ 和 D^吨。我们可以得到多模态
来自动量模块的铺垫,因为 F (E v(I), E t(T)) = {m cls, m tok}。基于此,动量令牌检测器
生成软伪标签来调制原始 token
预测,通过计算 KL 散度如下：

大号 妈妈 采取了 = E(I,T) P最 KL Dt(mtok)D t(mtok) (10)

终的文本操作基础损失是一个加权
组合如下：

LTMG = (1 − α)Ltok + αL 妈妈 采取了 (11)

细粒度操作类型检测和二进制
分类。与目前的伪造检测方法不同
主要执行真/假二进制分类,我们希望我们的模型能够为操纵检测提供更多的解释。如第3.2 节所述,两个图像

并介绍了两种文本处理方法
DGM4数据集。鉴于此,我们旨在进一步检测四个
细粒度的操作类型。不同的操作
类型可以同时出现在一个图像-文本对中，
我们将此任务视为特定的多模态多标签分类。由于 [CLS] token mcls聚合了多模态
经过模态感知交叉注意后的信息,可以
作为操纵特征的综合总结。因此,我们连接了一个多标签分类器

在其上方的Cm来计算多标签分类损失：

LMLC = E(I,T) P H(Cm(mcls), ymul) (12)

当然,我们也进行正常的二分类
根据mcls如下：

LBIC = E(I,T) P H(Cb(mcls), ybin) (13)

表2. DGM4的多模态学习方法比较。

类别 方法 AUC	二进制类		多标签分类	形象接地		文本基础	
EER ACC mAP CF1 OF1 CLIP [35] 76.40 59.52 62.31 ViLT [19] 78.38 66.14 66.00我们的 14.10				IoUmean IoU50 IoU75 准确率 召回率 50.03 38.79 58.12 22.11 65.18			F1
86.39 79.37 80.37	83.22 24.61 85.16		66.00	49.51	48.10 66.48 49.88 83.75 76.06 75.01 68.02		32.03
	22.88 93.19		72.37	59.32			57.00
			86.22	76.45			71.35

表 3. DGM4 的深度伪造检测方法比较。

类别 图像接地	二进制类		
方法 AUC EER ACC IoUmean IoU50 IoU75			
特许厅[30] 91.80 17.11 82.89 74.06		72.85	79.12
与[60] 91.31 17.65 82.36 74.70		72.88	78.98
我们的 94.40 13.18 86.80 75.65		75.69	82.93

表4. DGM4 序列标记方法比较。

类别 二进制 Cls 文本基础			
方法 AUC EER ACC 准确率 召回率 BERT [7] 80.82 28.02 68.98 41.39 63.85 LUKE [55]			F1
81.39 27.88 76.18 50.52 37.93我们的 93.44 13.83 87.39 70.90 73.30			50.23
			43.33
			72.08

表5.图像模态的消融研究。

类别	二进制类		形象接地		
方法 AUC EER ACC IoUmean IoU50 IoU75					
我们的图片 93.96我们的	13.83 86.13 13.18		75.58	82.44	75.80
94.40	86.80		75.69	82.93	75.65

表 6.文本模态的消融研究。

类别	二进制类		文本基础		
方法 AUC EER ACC 准确率 召回率					F1
我们的文本	75.67	32.46 72.17 13.83	42.99	33.68	37.77
我们的	93.44	87.39	70.90	73.30	72.08

5.实验

实施细节请参阅附录。
严格设定评估指标。

5.1. DGM4 基准

与多模态学习方法的比较。我们
将两种 SOTA 多模态学习方法应用于DGM4
设置进行比较。具体来说,CLIP [35]是
最流行的双流方法,其中两种模态在输入层面上不连接。为了适应,

我们让两个流的输出相互作用
通过交叉注意层。检测和接地
头部进一步集成在其上。此外,
ViLT [19]是一种代表性的单流方法,其中
跨模态交互层在图像和文本输入的连接上运行。我们还通过将检测和接地头连接到相应的

模型的输出。我们将比较结果列于
表2. 结果表明,所提出的方法在所有评估方面都明显优于两个基线
指标。这表明,分层操纵
推理更能准确、全面
建立图像和文本之间的相关性模型并捕捉
操纵造成的语义不一致,有助于更好地检测和确定操纵。

与深度伪造检测和序列标记的比较
ging 方法。我们将我们的方法与竞争方法进行了比较
两个单模态伪造数据分割中的单模态方法,
为了公平比较,除了原始的二元分类基本事实外,我们进一步将操纵基础头集成到单模态

具有相应接地注释的型号。对于
图像模态,我们列出了与两个 SOTA 的比较
表 3 中的 deepfake 检测方法。对于文本模态,
我们比较了两种广泛使用的序列标记方法
NLP 到地面操纵的 token 以及二元分类。我们在表4 中报告了比较结果。表3和表 4显示HAMMER 的表现优于

单模态伪造检测的单模态方法
大幅度。这表明我们用多模态媒体训练的方法在每种单一模态中也实现了良好的操作检测和接地性能。

5.2. 实验分析

两种模态的消融研究。为了验证多模态相关性对我们模型的重要性,我们进行了
通过仅保留与图像（我们的图像)相对应的输入和网络组件来进行消融研究,或者
文本（Ours-Text)模态。我们将结果列于表5
6 ,展示我们完整版本的性能
模型超越了其简化的部分,特别是在文本模态方面。这表明一旦其中一个
如果没有跨模态相互作用,则两种模态之间的关联性是缺失的。也就是说,通过利用
通过我们的模型比较两种模态,可以挖掘出它们之间更多互补的信息来促进我们的任务。
具体来说,这种相关性对于文本模态中的操作检测和基础来说更加重要。
损失消融研究。表 7 列出了每种情况下考虑的损失和相应的结果。从表 7 可以看出,删除任务通用

损失,即LMAC ,几乎所有的性能都会下降。
这意味着操纵感知对比学习对于我们的任务是必不可少的。在摆脱任何一项特定于任务的损失（即LMLC 、 LIMG和LTMG）后,不仅
相应任务的性能急剧下降,而且整体二分类性能也会下降

可能会变得更低。相比之下,我们的模型
完整的损失函数在
大多数情况下,表明所有损失的有效性和互补性。特别是表7的第一行
表示当前仅使用LBIC的多模态错误信息检测场景。我们的方法在二元分类上大大优于该基线,

表 7. 所提出方法中的损失消融研究。

损失				二进制类			多标签分类	形象接地			文本基础		
BIC	MLC	MAC	IMG	TMG	AUC	EER	ACC	mAP	CF1	OF1	IoUmean	IoU50	IoU75	精度	记起	F1
					91.04	16.91	83.81	20.79 33.84 33.48 27.22 30.81	4.81	0.33	0.00	15.95	78.70	26.53
					91.74	16.08	84.39	27.62 85.52 79.09 79.86	74.05	81.34	72.59	74.30	66.84	70.37
					92.77	14.53	86.01		75.98	83.37	75.25	77.82	61.83	68.91
					93.21	14.30	92.99	86.28	86.29 79.37 80.32 86.06 79.06	4.69	0.17	0.00	75.72	67.44	71.34
					14.62	93.19	14.10	86.15	79.93 86.22 79.37 80.37	76.51	83.73	76.05	13.93	58.87	22.53
								86.39		76.45	83.75	76.06	75.01	68.02	71.35

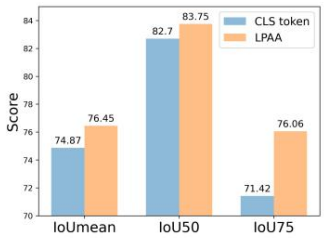

图4 局部贴剂疗效
注意力聚集（LPAA）。

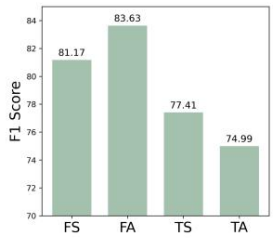

图 5. 性能
每种操作类型。

这意味着DGM4中更多的操作基础任务也有助于二元分类。

LPAA 的有效性。关于操纵的 bbox 接地，我们在图 4 中比较了
[CLS] token [59]的使用情况与提出的 LPAA。图4显示 LPAA 产生了更好的
所有指标下的表现，验证了其有效性。

操纵类型检测的细节。我们根据以下数据绘制了每种操纵类型的分类性能
图 5 中多标签分类器的输出结果
提供比文本操纵检测更多的解释
比图像模态更难，而TA是最难的情况。

操纵检测和接地的可视化。
我们在图 6 中提供了一些操作检测和基础的可视化结果。图6 (a)-(b) 展示了我们的方法
可以准确地确定操纵的框并检测 FA 和 FS 的正确操纵类型。此外，

TS 中的大部分被操纵的文本标记以及所有这些
在图6 (c)-(d)中，TA 中的接地成功。
它们可以直观地验证有效的操作检测，并且
可以通过HAMMER实现接地。

注意力图的可视化。我们提供 Grad-CAM
图 7显示了我们模型对操纵的文本标记的可视化。图7 (a) 显示了我们的模型关注
图片中人物的周围环境。这些周围环境
表示角色正在发表演讲，这在语义上与 TS 操纵的文本标记不同。至于
TA，图7 （b)显示了每个单词的可视化
到被操纵的单词（“哀悼”）。它暗示了我们的模型
关注图像中语义上
与操纵词（“哀悼”)表达的悲伤情绪不一致。这些样本证明我们的模型可以
确实捕捉到了图像之间的语义不一致
和文本来解决DGM4 。

(a)GT:假 FS， Pred:假 FS　　(b)GT:假 FA， Pred:假 FA

(c)GT:假 TS， Pred:假 TS　　(d)GT:假 TA， Pred:假 TA

图 6. 检测和接地结果的可视化。接地
真实注释为红色，预测结果为蓝色。

(a)TS 中的注意力图　　(b)TA 中的注意力图

图 7.Grad-CAM 对所操作的文本标记的可视化。

6. 结论

本文研究了一种新型DGM4问题，旨在检测和验证多模态操作。我们构建
第一个具有丰富注释的大规模DGM4数据集。
提出了强大的模型HAMMER ，并进行了大量实验来证明其有效性。

致谢

本研究得到教育部支持，
新加坡，根据其 MOE AcRF Tier 2（MOE-T2EP20221-0012)、NTU NAP 以及 RIE2020 产业协调基金 - 产业合作项目 (IAF-ICP)
资助计划以及现金和实物捐助
来自行业合作伙伴。

参考

[1] Sahar Abdelnabi、Rakibul Hasan 和 Mario Fritz。通过在线资源对脱离上下文的图像进行开放域、基于内容、多模式事实核查。CVPR,2022年。2、3

[2] Shivangi Aneja、Chris Bregler 和 Matthias Nießner。COS-MOS:利用自我监督学习捕捉脱离语境的错误信息。在 ArXiv 预印本 arXiv:2101.06278 中，2021. 3

[3] 伊曼纽尔·本-巴鲁克、塔尔·里德尼克、纳达夫·扎米尔、阿萨夫·诺伊，伊塔玛·弗里德曼、马坦·普罗特和利希·泽尔尼克·马诺。多标签分类的非对称损失。在 ICCV 中，2021. 11

[4] Renwang Chen, Xuanhong Chen, Bingbing Ni, and Yanhao Ge. Simswap：一个高效的高保真人脸识别框架交换。在 ACM MM,2020 年。3

[5] Shen Chen, Taiping Yao, Yang Chen, Shouhong Ding, Jilin Li 和 Rongrong Ji. 人脸伪造的局部关系学习检测。在 AAAI,2021 年。2

[6] 检测和可视化推特上的误导性内容。Boididou、christina 和 papadopoulos、symeon 和 zam-poglou、markos 和 apostolidis、lazaros 和 papadopoulou，奥尔加和科帕齐亚里斯，伊安尼斯。艾杰米尔， 2018。3

[7] Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova。Bert:用于语言理解的深度双向变换器的预训练。在 NAACL,2019 年。2、7，11

[8]布莱恩·多尔汉斯基、乔安娜·比顿、本·普夫劳姆、陆吉阔、拉斯 Howes、Menglin Wang 和 Cristian Canton Ferrer。deepfake 检测挑战 (dfdc) 数据集。arXiv 预印本 arXiv:2006.07397, 2020. 2, 3

[9] 阿列克谢·多索维茨基、卢卡斯·拜尔、亚历山大·科列斯尼科夫，德克·魏森伯恩、翟晓华、托马斯·恩特蒂纳、Mostafa Dehghani、Matthias Minderer、Georg Heigold、Syl-vain Gelly 等人。一张图片胜过 16x16 个单词:用于大规模图像识别的 Transformers。ICLR，2020 年。11

[10] Thibaut Durand、Nazanin Mehrasa 和 Greg Mori。学习深度卷积网络进行部分多标签分类标签。在 CVPR,2019 年。11

[11] Tarik Dzanic、Karan Shah 和 Freddie Witherden。Fourier 深度网络生成图像中的光谱差异。NeurIPS， 2020.3

[12] Gege Gao, Huaibo Huang, Chaoyou Fu, Zhaoyang Li, and 何冉。身份认同的信息瓶颈解构交换。在 CVPR,2021 年。3

[13] Qiqi Gu, Shen Chen, Taiping Yao, Yang Chen, Shouhong Ding 和 Ran Yi。利用细粒度人脸伪造线索通过渐进式增强学习。在 AAAI,2022 年。3

[14] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick。无监督视觉表征学习的动量对比。CVPR,2020 年。5

[15] Liming Jiang, Ren Li, Wayne Wu, Chen Qian, and Chen Change Loy。Deeperforensics-1.0:大规模用于真实世界人脸伪造检测的数据集。CVPR,2020 年。2、3

[16] Zhiwei Jin, Juan Cao, Han Guo, Yongdong Zhang, and Jiebo Luo. 使用循环神经网络进行多模态融合微博谣言检测。ACM MM， 2017。3

[17] Tero Karras、Timo Aila、Samuli Laine 和 Jaakko Lehtinen。逐步提高 gans 的质量、稳定性，和变化。在 ICLR,2018 年。3

[18] Dhruv Khattar、Jaipal Singh Goud、Manish Gupta 和 Va-sudeva Varma。 Mvae:多模态变分自动编码器用于检测虚假新闻。在 WWW,2019 年。3

[19] Wonjae Kim、Bokyung Son 和 Ildoo Kim。Vilt:无需卷积或区域监督的视觉和语言转换器。ICML,2021年。2、7

[20] 李俊楠,Ramprasaath Selvaraju,Akhilesh Gotmare，沙菲克·乔蒂、蔡明熊和史蒂文·朱洪海。融合前对齐:通过动量蒸馏进行视觉和语言表征学习。在 NeurIPS,2021 年。2、4、6

[21] Jiaming Li, Hongtao Xie, Jiahong Li, Zhongyuan Wang, and 张永东. 频率感知判别特征单中心损失监督学习用于人脸伪造检测。在 CVPR,2021 年。3

[22] Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong 陈、方文和郭百宁。更多详情请看面部X光通用人脸伪造检测。CVPR,2020. 2

[23] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-DF:用于 Deepfake 的大规模具有挑战性的数据集法医。在 CVPR,2020 年。 3

[24] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df:用于深度伪造取证的新数据集。在 CVPR 中，2020. 2

[25] 刘福晓、王英汉、王天璐、Vicente Or-donez。视觉新闻:新闻图像字幕的基准和挑战。在 EMNLP,2021 年。3

[26] Honggu Liu, Xiaodan Li, Wenbo Zhou, Yuefeng Chen, Yuan 何晖、薛辉、张伟明、余能海。空间相位浅层学习:重新思考人脸伪造检测频域。在 CVPR,2021 年。3

[27] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi、Danqi Chen、Omer Levy、Mike Lewis、Luke Zettle-moyer 和 Veselin Stoyanov。Roberta:稳健优化bert 预训练方法。arXiv 预印本 arXiv:1907.11692，2019. 3

[28] Ilya Loshchilov 和 Frank Hutter。解耦权重衰减正则化。arXiv 预印本 arXiv:1711.05101,2017 年11 月

[29] Grace Luo, Trevor Darrell 和 Anna Rohrbach. NewsCLIP-pings:自动生成脱离上下文的多模态媒体。在 EMNLP,2021. 2, 3

[30] Yuchen Luo, Yong Zhang, Junchi Yan, and Wei Liu. Gener-alizing face forgery detection with high-frequency features. 在 CVPR,2021. 2, 7, 11

[31] Aaron van den Oord、Yazhe Li 和 Oriol Vinyals。使用对比预测编码的表示学习。arXiv预印本 arXiv:1807.03748, 2018. 5

[32] 亚当·帕斯克、萨姆·格罗斯、苏米特·钦塔拉、格雷戈里 Chanan、Edward Yang、Zachary DeVito、Zeming Lin、Al-ban Desmaison、Luca Antiga 和 Adam Lerer。自动在 PyTorch 中区分。2017. 11

[33] Or Patashnik、Zongze Wu、Eli Shechtman、Daniel Cohen-Or、和 Dani Lischinski。Styleclip:文本驱动的stylegan 图像。在 ICCV,2021 年。3

[34] Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao。频率思考：通过挖掘频率感知线索进行人脸伪造检测。ECCV，2020 年。3

[35] 亚历克·雷德福、金钟旭、克里斯·哈拉西、阿迪亚 Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sastry、Amanda Askell、 Pamela Mishkin、Jack Clark 等人。从自然语言超级视觉中学习可迁移的视觉模型。在 ICML，2021 年。2、7

[36] 亚历克·拉德福德、Jeffrey Wu、Rewon Child、David Luan、Dario Amodei、Ilya Sutskever 等人。语言模型是无人监督的多任务学习者。OpenAI 博客，2019 年。2

[37] Nils Reimers 和 Iryna Gurevych。Sentence-bert：使用连体 bert 网络进行句子嵌入。在 EMNLP-IJCNLP 中，2019 年。3

[38] 哈米德·雷扎托菲吉、内森·蔡、JunYoung Gwak、阿米尔 Sadeghian、Ian Reid 和 Silvio Savarese。广义交集与并集：边界框的度量和损失 回归。在 CVPR，2019 年。6

[39] 安德烈亚斯·罗斯勒、达维德·科佐利诺、路易莎·韦尔多利瓦、克里斯蒂安·里斯、贾斯图斯·蒂斯和马蒂亚斯·尼埃纳。 Faceforen-sics++：学习检测被操纵的面部图像。在 ICCV, 2019. 2, 3

[40] Rui Shao, Xiangyuan Lan, Jiawei Li, and Pong C. Yuen. 多对抗判别深度域泛化 用于人脸呈现攻击检测。CVPR， 2019。2

[41] Rui Shao、Xiangyuan Lan 和 Pong C Yuen。深度卷积动态纹理学习与自适应通道可辨别性，用于 3D 面具人脸防欺骗。在 IJCB 中， 2017. 2

[42] Rui Shao, Xiangyuan Lan, 和 Pong C Yuen. 3D 面具脸深度动态纹理的联合判别学习 反欺骗。IEEE 信息取证交易 和安全，14(4):923–938，2018。2

[43] Rui Shao, Xiangyuan Lan, and Pong C. Yuen. Regularized 细粒度元人脸反欺骗。在 AAAI，2020 年。2

[44] Rui Shao、Pramuditha Perera、Pong C Yuen 和 Vishal M Patel。开放集对抗性防御。ECCV，2020 年。2

[45] Rui Shao、Pramuditha Perera、Pong C Yuen 和 Vishal M Patel。联合广义人脸呈现攻击检测。IEEE 神经网络与学习学报 系统， 2022.2

[46] Rui Shao、Pramuditha Perera、Pong C Yuen 和 Vishal M Patel。开放式对抗性防御与清洁对抗性 相互学习。国际计算机视觉杂志， 130(4):1070–1087, 2022.2

[47] Rui Shao, Tianxing Wu, and Ziwei Liu. Detecting and re-covering sequential deepfake manipulation. In ECCV, 2022. 2

[48] Rui Shao, Bochao Zhang, Pong C Yuen, and Vishal M Patel. 联合测试时自适应人脸呈现攻击检测与双阶段隐私保护。在 FG，2021 年。2

[49] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan 刘。社交媒体上的虚假新闻检测：数据挖掘 观点。ACM SIGKDD 探索通讯，2017 年。 3

[50] Akhilesh Sudhakar、Bhargav Upadhyay 和 Arjun Maheswaran。转变删除、检索、生成方法 用于受控文本样式转换。在 EMNLP-IJCNLP，2019 年。 4

[51] Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszko-Reit、Llion Jones、Aidan N Gomez、 ukasz Kaiser 和 Illia Polosukhin。注意力就是你所需要的一切。NeurIPS，2017年。4、6

[52] Tengfei Wang, Yong Zhang, Yanbo Fan, Jue Wang, and 陈奇峰. 图像属性的高保真 GAN 反演 编辑。在 CVPR，2022 年。3

[53] William Yang Wang. “liar, liar pants on fire”：一个新的假新闻检测基准数据集. In ACL, 2017. 2, 3

[54] Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha、Lu Su 和 Jing Gao。Eann：用于多模态假新闻检测的事件对抗神经网络。 和 KDD， 2018。3

[55] 山田郁也、浅井明里、新藤博之、秀明 Takeda、Yuji Matsumoto。LUKE：深度语境化 具有实体意识自我注意力的实体表示。在 EMNLP, 2020. 7

[56] Jinyu Yang, Jiali Duan, Son Tran, Yi Xu, Sampath Chanda, Liqun Chen, Belinda Zeng, Trishul Chilimbi, and Junzhou Huang. 基于三重对比的视觉语言预训练 学习。在 CVPR，2022 年。5

[57] Zitong Yu, Jun Wan, Yunxiao Qin, Xiaobai Li, Stan Z Li, and 赵国英。Nas-fas：静态-动态中心差异 网络搜索人脸反欺骗。IEEE 交易 模式分析与机器智能，43(9):3005–3023， 2020. 2

[58] 罗温·泽勒斯、阿里·霍尔兹曼、汉娜·拉什金、尤纳坦 Bisk、Ali Farhadi、Franziska Roesner 和 Yejin Choi。《防御神经假新闻》。NeurIPS，2019年。2、3

[59] Yan Zeng、Xinsong Zhang 和 Hang Li。多粒度视觉语言预训练：将文本与视觉概念对齐。在 ICML，2022年。6、8

[60] Hanqing Zhao, Wenbo Zhou, Dongdong Chen, Tianyi Wei, Weiming Zhang, and Nenghai Yu. Multi-attentional deep-fake detection. In CVPR, 2021. 2, 7, 11

[61] Tianchen Zhao, Xiang Xu, Mingze Xu, Hui Ding, Yuanjun Xiong 和 Wei Xia。学习 deepfake 的自洽性 检测。在 ICCV，2021 年。2

[62] Xiangyu Zhu, Hao Wang, Hongyan Fei, Zhen Lei, and Stan Z Li. 通过 3d 分解进行人脸伪造检测。 CVPR, 2021. 2

补充材料

A.实施细节。

我们的所有实验都是在 8 NVIDIA
带有 PyTorch 框架的 V100 GPU [32]。图像编码器
由 12 层的 ViT-B/16 [9]实现。文本
编码器和多模式聚合器基于
6 层 Transformer 由前 6 层和
BERTbase [7]的最后 6 层。二分类器、多标签分类器、BBox 检测器和 Token
检测器设置为两个多层感知器 (MLP) 层，输出维度分别为 2、4、4、2。我们设置
队列大小K = 65,536。采用AdamW [28]优化器
权重衰减为 0.02。学习率已预热
至 $1e^{-4}$ 在前 1000 步中，衰减至 $1e^{-5}$ 跟随-
制定余弦时间表。

B.评估指标。

评估提出的新研究问题 DGM4
我们制定了全面的严格评估协议
以及所有操作检测和接地任务的指标。

·二元分类：遵循当前的 deepfake 方法[30, 60]，我们采用准确率 (ACC)、面积
受试者工作特征曲线 (AUC)，以及
用于评估二元分类的等错误率（EER）。

·多标签分类：与现有的多标签分类方法[3,10]一样，我们使用平均准确率
（MAP）、平均每类 F1（CF1）和平均总体
F1（OF1）用于评估细粒度操作类型的检测。

·操纵图像边界框基础：为了检查预测操纵边界框的性能，
我们计算交集与并集的平均值
（IoUmean）所有测试样本的预测坐标与真实坐标之间的IoU。此外，我们设
置了两个阈值（0.5,0.75），并计算平均准确率
（如果 IoU 高于阈值，则正确接地，反之亦然
反之），分别记为IoU50和IoU75。

·操纵文本标记接地：考虑
被操纵的代币属于类别不平衡场景
比原始 token 少得多，我们采用 Precision、Re-call、F1 Score 作为指标。
这有助于更
对操纵文本标记进行公平合理的评估
接地。