FAKESHIELD: EXPLAINABLE IMAGE FORGERY DETECTION AND LOCALIZATION VIA MULTI-MODAL LARGE LANGUAGE MODELS

中文题目：FAKESHIELD：基于多模态大型语言模型的可解释图像伪造检测与定位

发布于：ICLR 2025

级别：CCF-A

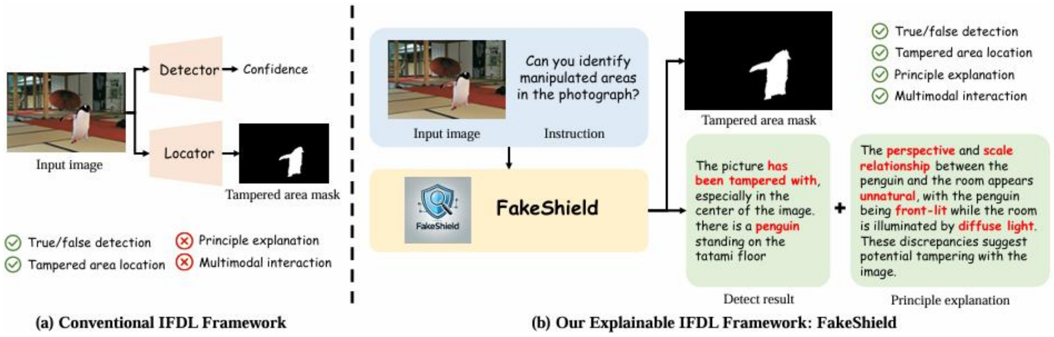
论文链接：2410.02761

论文类别：图像伪造检测与定位

标签：多模态大型语言模型，可解释性伪造检测

本文摘要：生成式人工智能的迅速发展是一把双刃剑，它不仅极大地促进了内容创作，也让图像篡改变得更加容易且更难以检测。尽管现有的图像伪造检测与定位（IFDL）方法总体上效果良好，但它们面临两个主要挑战：一是检测原理的不透明性（黑箱特性），二是在不同类型的篡改方式（如 Photoshop、DeepFake、AIGC 编辑等）中泛化能力有限。为了解决这些问题，我们提出了可解释的 IFDL 任务，并设计了一个名为 FakeShield 的多模态框架，该框架能够评估图像的真实性、生成被篡改区域的掩码，并基于像素级和图像级的篡改线索提供判断依据。此外，我们利用 GPT-4o 扩充了现有的 IFDL 数据集，构建了多模态篡改描述数据集（MMTD-Set），用于训练 FakeShield 的篡改分析能力。同时，我们引入了领域标签引导的可解释伪造检测模块（DTEFDM）和多模态伪造定位模块（MFLM），以应对多种篡改检测与解释任务，并通过详细的文本描述实现伪造定位。大量实验结果表明，FakeShield 能够有效检测并定位多种篡改技术，相较于以往的 IFDL 方法，提供了一种更具可解释性和更优性能的解决方案。代码已在 https://github.com/zhipeixu/FakeShield

本文聚焦的问题：


Input image
Detector
Confidence
Locator
Tampered area mask
True/false detection
Tampered area location
Principle explanation
Multimodal interaction
(a) Conventional IFDL Framework
Can you identify manipulated areas in the photograph?
Input image
Instruction
FakeShield
Tampered area mask
The picture has been tampered with, especially in the center of the image. there is a penguin standing on the tatami floor
The perspective and scale relationship between the penguin and the room appears unnatural, with the penguin being front-lit while the room is illuminated by diffuse light. These discrepancies suggest potential tampering with the image.
Detect result
Principle explanation
(b) Our Explainable IFDL Framework: FakeShield

1.缺乏可解释性（Explainability Gap）

现有图像伪造检测模型大多属于黑箱系统，只能输出“真/假”结果，却无法说明判断依据或具体篡改位置，导致结果难以被信任或用于取证。

2.泛化能力不足（Limited Generalization）

模型在不同篡改方式之间（如 Photoshop、DeepFake、AIGC 编辑）迁移性能较差，容易过拟合特定数据集，无法适应真实世界中多样化的伪造类型。

3.单模态问题（Lack of Multi-modal Understanding）

传统检测方法仅依赖视觉信号，缺少语言层面的推理能力，难以实现从“看到问题”到“解释问题”的跨模态理解。

4.训练数据有限且缺少语言标注（Data Limitation）

现有伪造检测数据集大多只有图像和掩码，没有自然语言描述，限制了多模态模型在可解释任务上的训练与评估。

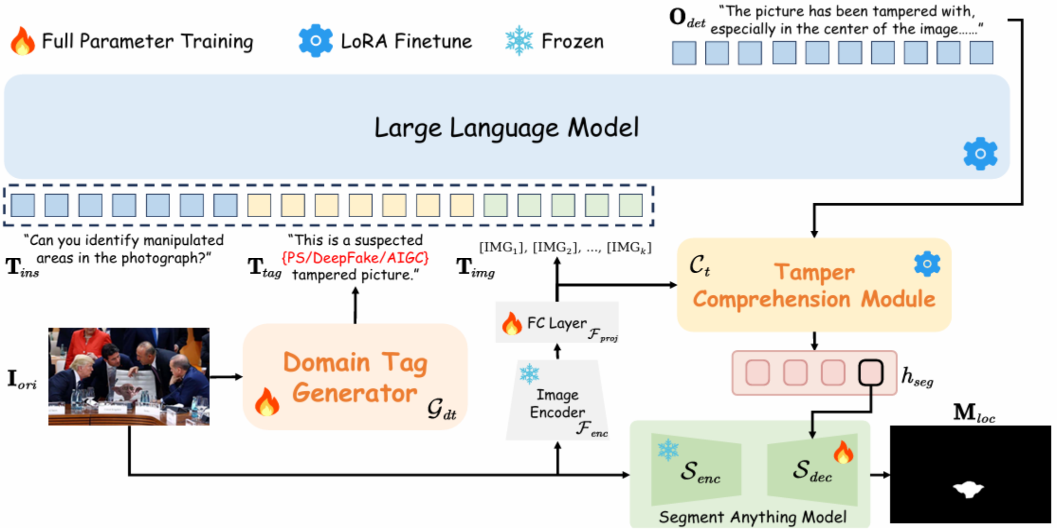
5.不同伪造领域的特征差异（Domain Discrepancy）

不同伪造类型存在显著的分布差异，导致统一检测模型难以兼顾各类篡改形式，亟需一种能够利用“领域标签”引导学习的机制。

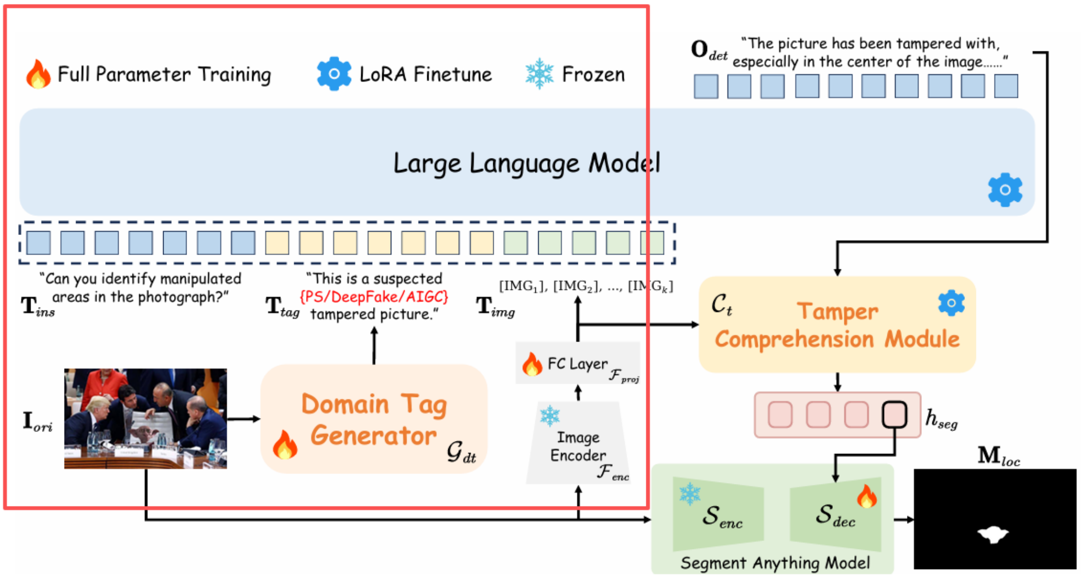
本文提出的方法：

网络架构

模型设计为包含两个解耦部分：Domain Tag - guided Explainable Forgery Detection Module(DTE - FDM)和 Multi-modal Forgery Localization Module(MFLM)。整体工作流程是将原始疑似图像和指令文本输入 DTE - FDM，通过域标签生成器（DTG）获得域标签，与图像编码后的文本和指令一同输入 LLM，预测检测结果和判断依据；再将和输入 MFLM，利用篡改理解模块（TCM）对齐视觉和文本特征，增强视觉基础模型对长描述的理解能力，最终通过 SAM 生成准确的篡改掩码。


Full Parameter Training
LoRA Finetune
Frozen
O_det "The picture has been tampered with, especially in the center of the image......"
Large Language Model
T_ins "Can you identify manipulated areas in the photograph?"
T_tag "This is a suspected {PS/DeepFake/AIGC} tampered picture."
T_img [IMG_1], [IMG_2], ..., [IMG_k]
C_t Tamper Comprehension Module
FC Layer F_proj
Image Encoder F_enc
h_seg
I_ori
Domain Tag Generator G_dt
M_loc
S_enc
S_dec
Segment Anything Model

下面将从模型设计的两部分 DTE - FDM、MFLM 依次讨论。首先是 DTE - FDM，图中红色圈出来的地方是其实现流程。


Full Parameter Training
LoRA Finetune
Frozen
O_det "The picture has been tampered with, especially in the center of the image......"
Large Language Model
T_ins "Can you identify manipulated areas in the photograph?"
T_tag "This is a suspected {PS/DeepFake/AIGC} tampered picture."
T_img [IMG_1], [IMG_2], ..., [IMG_k]
C_t Tamper Comprehension Module
FC Layer F_proj
Image Encoder F_enc
h_seg
I_ori
Domain Tag Generator G_dt
M_loc
S_enc
S_dec
Segment Anything Model

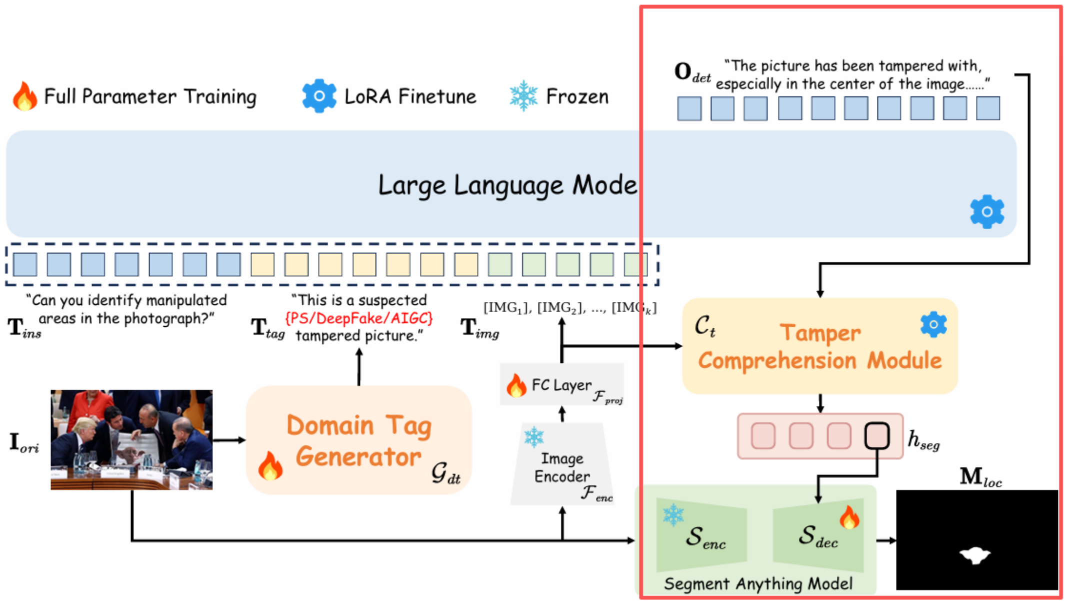
其目标是实现疑似图像检测结果的文本解释。图像输入分类器，将常见篡改类型分为 Photoshop - based editing、DeepFake 和 AIGC - based tampering 三类，使用特定模板文本（如 “This is a suspected {篡改类型}-tampered picture.”）作为标识符生成域标签。同时，经过图像编码器和线性投影层生成图像文本。最后，将、与指令（如 “Can you identify manipulated areas in the photograph?”）拼接后输入 LLM。由于 LLM 规模大且计算资源有限，采用冻结 LLM 并利用 LoRA 微调技术的策略，经过多次自回归预测，输出包含检测结果、篡改区域位置描述和检测解释性依据的文本描述。需要注意其中的分类器，本质就是三分类模型，而且

需要单独提前训练完成才能在此架构中使用。另外的编码器使用的是完全冻结的

iml-vit

IML-VIT 实战训练与测试：基于 Vision transform 实现图像伪造检测）。

而最后输入 LLM 的 token 指令集：与标签是固定版式、则是图片编码后的特征向量。


🔥 Full Parameter Training
LoRA Finetune
❄ Frozen
$\mathbf{O}_{det}$ "The picture has been tampered with, especially in the center of the image......"
Large Language Mode
"Can you identify manipulated areas in the photograph?"
$\mathbf{T}_{ins}$
"This is a suspected {PS/DeepFake/AIGC} tampered picture."
$\mathbf{T}_{tag}$
[IMG$_1$], [IMG$_2$], ..., [IMG$_k$]
$\mathbf{T}_{img}$
$\mathcal{C}_t$ Tamper Comprehension Module
FC Layer $\mathcal{F}_{proj}$
Image Encoder $\mathcal{F}_{enc}$
$\mathbf{I}_{ori}$
Domain Tag Generator $\mathcal{G}_{dt}$
h_{seg}
$\mathbf{M}_{loc}$
$\mathcal{S}_{enc}$
$\mathcal{S}_{dec}$
Segment Anything Model

篡改图像的检测结果文本描述与图像文本输入到 TCM（篡改理解模块），编码器 TCM 对齐长文本特征与视觉特征，增强 SAM 定位伪造区域的精度。通过引入特定令牌<SEG>，提取 TCM 最后一层嵌入并经 MLP 投影层转换为。之后，原始图像经 SAM 编码器和解码器处理，作为的提示指导生成掩码。在 MFLM 中也应用 LoRA 微调，目的是让 SAM 能更精确地定位伪造区域。在训练过程中模型两部分是分开训练的，并不是端到到的链路，首先是训练检测 DTE - FDM，它 loss 如下：

$$\ell_{det} = \ell_{ce}(\hat{\mathbf{O}}_{det}, \mathbf{O}_{txt}) + \lambda \cdot \ell_{ce}(\hat{\mathbf{T}}_{tag}, \mathbf{T}_{tag}),$$

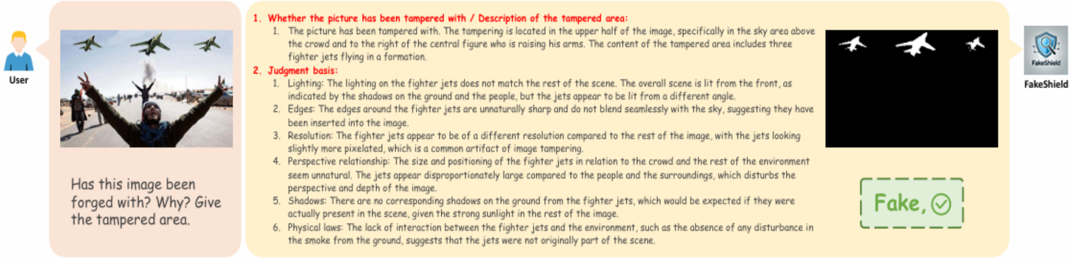
交叉熵损失训练标签 tag 以区分不同数据类型；LLM 的训练目标是自回归文本生成交叉熵损失，其中用于平衡损失权重。

之后微调 TCM 和 SAM，定位 MFLM 的 loss 如下：

$$\ell_{loc} = \ell_{ce}(\hat{\mathbf{y}}_{txt}, \mathbf{y}_{txt}) + \alpha \cdot \ell_{bce}(\hat{\mathbf{M}}_{loc}, \mathbf{M}_{loc}) + \beta \cdot \ell_{dice}(\hat{\mathbf{M}}_{loc}, \mathbf{M}_{loc}),$$

其中和为平衡权重，ytxt 表示 TCM 结果，Mloc 是图片结果，其中^表示预测，没有^的是人工标签。

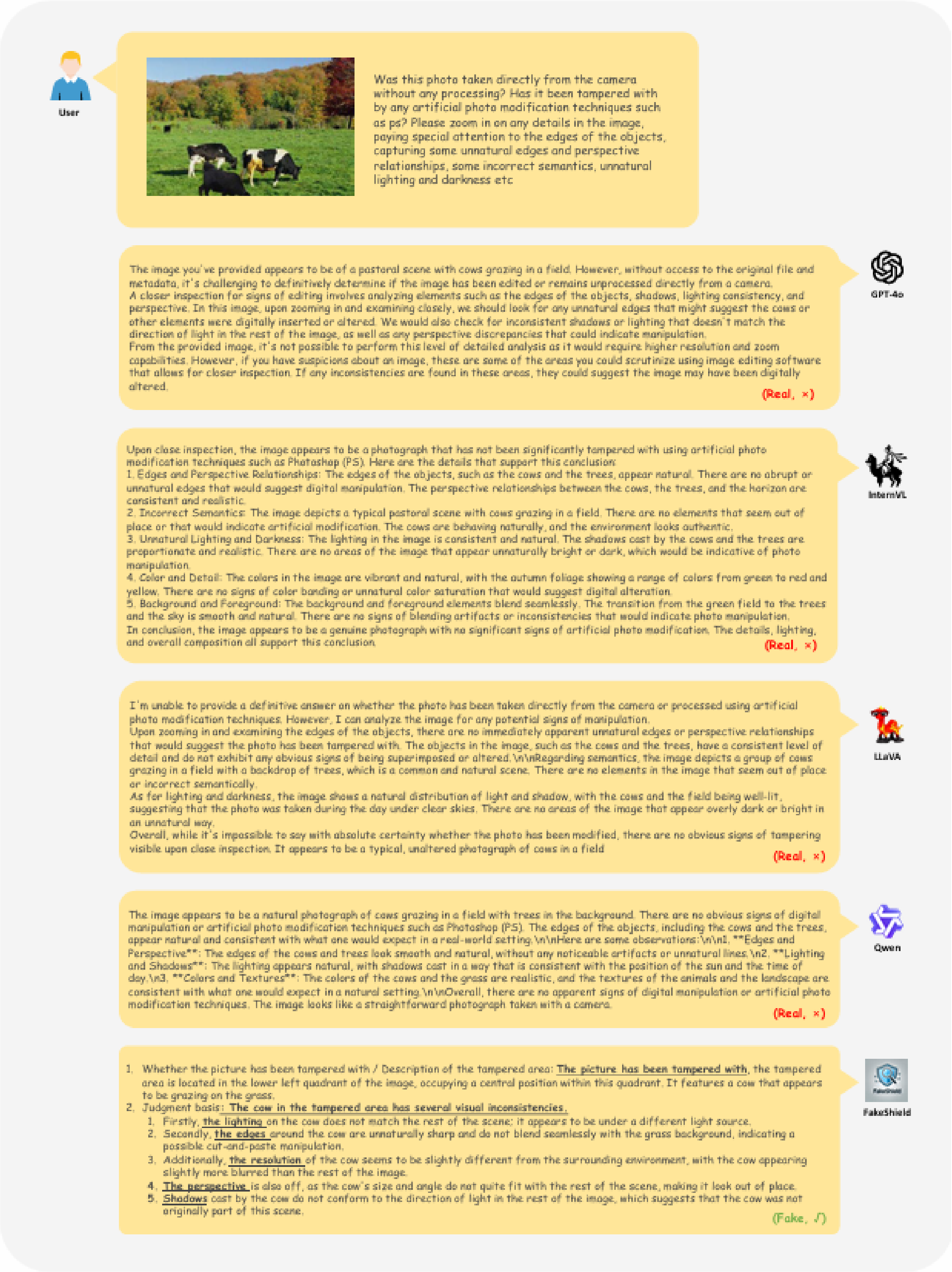
实验结果：



Has this image been forged with? Why? Give the tampered area.

1. **Whether the picture has been tampered with / Description of the tampered area:**
   1. The picture has been tampered with. The tampering is located in the upper half of the image, specifically in the sky area above the crowd and to the right of the central figure who is raising his arms. The content of the tampered area includes three fighter jets flying in a formation.
2. **Judgment basis:**
   1. Lighting: The lighting on the fighter jets does not match the rest of the scene. The overall scene is lit from the front, as indicated by the shadows on the ground and the people, but the jets appear to be lit from a different angle.
   2. Edges: The edges around the fighter jets are unnaturally sharp and do not blend seamlessly with the sky, suggesting they have been inserted into the image.
   3. Resolution: The fighter jets appear to be of a different resolution compared to the rest of the image, with the jets looking slightly more pixelated, which is a common artifact of image tampering.
   4. Perspective relationship: The size and positioning of the fighter jets in relation to the crowd and the rest of the environment seem unnatural. The jets appear disproportionately large compared to the people and the surroundings, which disturbs the perspective and depth of the image.
   5. Shadows: There are no corresponding shadows on the ground from the fighter jets, which would be expected if they were actually present in the scene, given the strong sunlight in the rest of the image.
   6. Physical laws: The lack of interaction between the fighter jets and the environment, such as the absence of any disturbance in the smoke from the ground, suggests that the jets were not originally part of the scene.

Figure 4: Detection, localization and explanation results of our FakeShield.



User

Was this photo taken directly from the camera without any processing? Has it been tampered with by any artificial photo modification techniques such as ps? Please zoom in on any details in the image, paying special attention to the edges of the objects, capturing some unnatural edges and perspective relationships, some incorrect semantics, unnatural lighting and darkness etc

GPT-4o

The image you've provided appears to be of a pastoral scene with cows grazing in a field. However, without access to the original file and metadata, it's challenging to definitively determine if the image has been edited or remains unprocessed directly from a camera.
A closer inspection for signs of editing involves analyzing elements such as the edges of the objects, shadows, lighting consistency, and perspective. In this image, upon zooming in and examining closely, we should look for any unnatural edges that might suggest the cows or other elements were digitally inserted or altered. We would also check for inconsistent shadows or lighting that doesn't match the direction of light in the rest of the image, as well as any perspective discrepancies that could indicate manipulation.
From the provided image, it's not possible to perform this level of detailed analysis as it would require higher resolution and zoom capabilities. However, if you have suspicions about an image, these are some of the areas you could scrutinize using image editing software that allows for closer inspection. If any inconsistencies are found in these areas, they could suggest the image may have been digitally altered.

(Real, ×)

InternVL

Upon close inspection, the image appears to be a photograph that has not been significantly tampered with using artificial photo modification techniques such as Photoshop (PS). Here are the details that support this conclusion:
1. Edges and Perspective Relationships: The edges of the objects, such as the cows and the trees, appear natural. There are no abrupt or unnatural edges that would suggest digital manipulation. The perspective relationships between the cows, the trees, and the horizon are consistent and realistic.
2. Incorrect Semantics: The image depicts a typical pastoral scene with cows grazing in a field. There are no elements that seem out of place or that would indicate artificial modification. The cows are behaving naturally, and the environment looks authentic.
3. Unnatural Lighting and Darkness: The lighting in the image is consistent and natural. The shadows cast by the cows and the trees are proportionate and realistic. There are no areas of the image that appear unnaturally bright or dark, which would be indicative of photo manipulation.
4. Color and Detail: The colors in the image are vibrant and natural, with the autumn foliage showing a range of colors from green to red and yellow. There are no signs of color banding or unnatural color saturation that would suggest digital alteration.
5. Background and Foreground: The background and foreground elements blend seamlessly. The transition from the green field to the trees and the sky is smooth and natural. There are no signs of blending artifacts or inconsistencies that would indicate photo manipulation.
In conclusion, the image appears to be a genuine photograph with no significant signs of artificial photo modification. The details, lighting, and overall composition all support this conclusion.

(Real, ×)

LLaVA

I'm unable to provide a definitive answer on whether the photo has been taken directly from the camera or processed using artificial photo modification techniques. However, I can analyze the image for any potential signs of manipulation.
Upon zooming in and examining the edges of the objects, there are no immediately apparent unnatural edges or perspective relationships that would suggest the photo has been tampered with. The objects in the image, such as the cows and the trees, have a consistent level of detail and do not exhibit any obvious signs of being superimposed or altered.\n\nRegarding semantics, the image depicts a group of cows grazing in a field with a backdrop of trees, which is a common and natural scene. There are no elements in the image that seem out of place or incorrect semantically.
As for lighting and darkness, the image shows a natural distribution of light and shadow, with the cows and the field being well-lit, suggesting that the photo was taken during the day under clear skies. There are no areas of the image that appear overly dark or bright in an unnatural way.
Overall, while it's impossible to say with absolute certainty whether the photo has been modified, there are no obvious signs of tampering visible upon close inspection. It appears to be a typical, unaltered photograph of cows in a field

(Real, ×)

Qwen

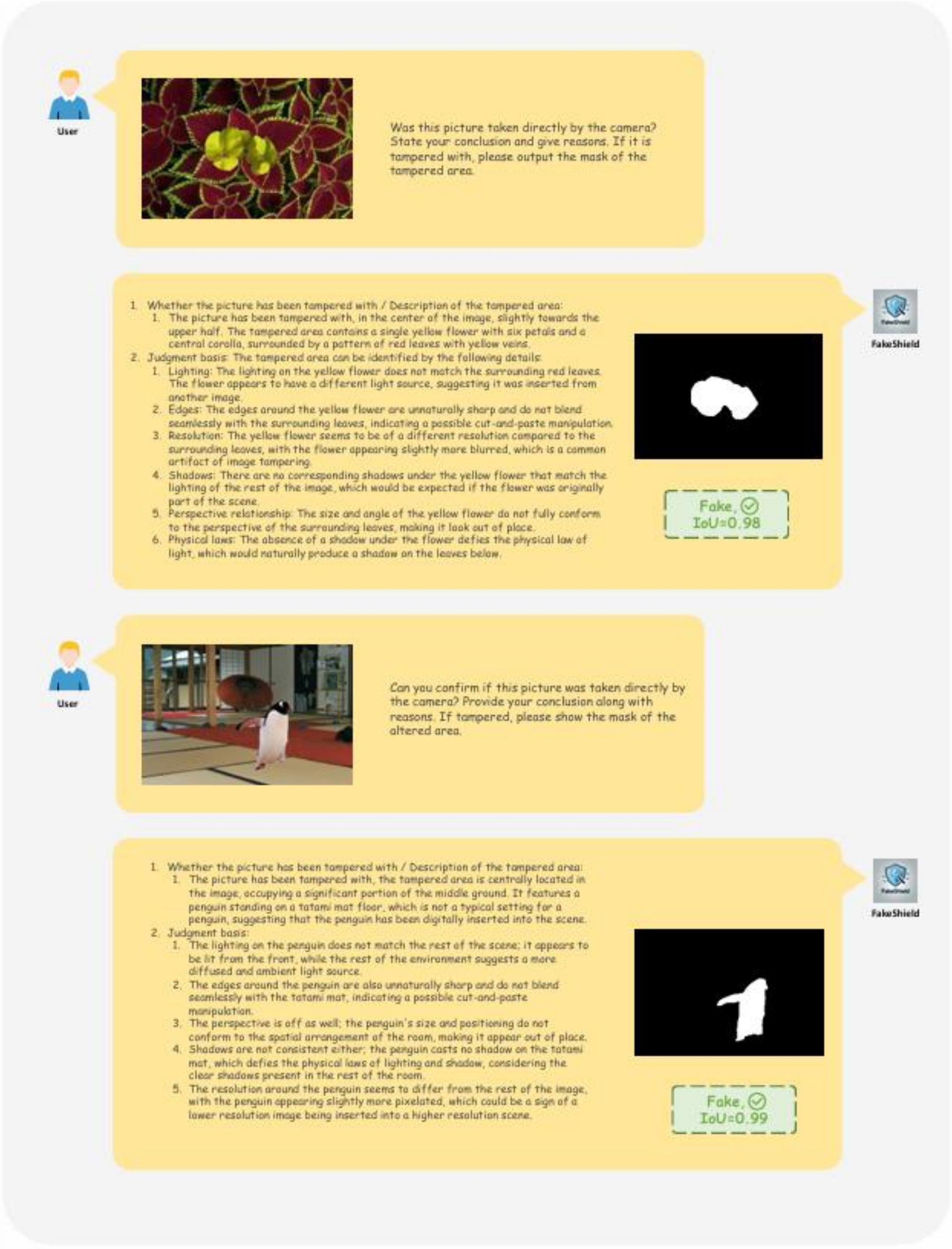
The image appears to be a natural photograph of cows grazing in a field with trees in the background. There are no obvious signs of digital manipulation or artificial photo modification techniques such as Photoshop (PS). The edges of the objects, including the cows and the trees, appear natural and consistent with what one would expect in a real-world setting.\n\nHere are some observations:\n\n1. **Edges and Perspective**: The edges of the cows and trees look smooth and natural, without any noticeable artifacts or unnatural lines.\n2. **Lighting and Shadows**: The lighting appears natural, with shadows cast in a way that is consistent with the position of the sun and the time of day.\n3. **Colors and Textures**: The colors of the cows and the grass are realistic, and the textures of the animals and the landscape are consistent with what one would expect in a natural setting.\n\nOverall, there are no apparent signs of digital manipulation or artificial photo modification techniques. The image looks like a straightforward photograph taken with a camera.

(Real, ×)

FakeShield

1. Whether the picture has been tampered with / Description of the tampered area: **The picture has been tampered with**, the tampered area is located in the lower left quadrant of the image, occupying a central position within this quadrant. It features a cow that appears to be grazing on the grass.
2. Judgment basis: **The cow in the tampered area has several visual inconsistencies.**
   1. Firstly, **the lighting** on the cow does not match the rest of the scene; it appears to be under a different light source.
   2. Secondly, **the edges** around the cow are unnaturally sharp and do not blend seamlessly with the grass background, indicating a possible cut-and-paste manipulation.
   3. Additionally, **the resolution** of the cow seems to be slightly different from the surrounding environment, with the cow appearing slightly more blurred than the rest of the image.
   4. **The perspective** is also off, as the cow's size and angle do not quite fit with the rest of the scene, making it look out of place.
   5. **Shadows** cast by the cow do not conform to the direction of light in the rest of the image, which suggests that the cow was not originally part of this scene.

(Fake, ✓)



Was this picture taken directly by the camera? State your conclusion and give reasons. If it is tampered with, please output the mask of the tampered area.

1. Whether the picture has been tampered with / Description of the tampered area:
    1. The picture has been tampered with, in the center of the image, slightly towards the upper half. The tampered area contains a single yellow flower with six petals and a central corolla, surrounded by a pattern of red leaves with yellow veins.
2. Judgment basis: The tampered area can be identified by the following details:
    1. Lighting: The lighting on the yellow flower does not match the surrounding red leaves. The flower appears to have a different light source, suggesting it was inserted from another image.
    2. Edges: The edges around the yellow flower are unnaturally sharp and do not blend seamlessly with the surrounding leaves, indicating a possible cut-and-paste manipulation.
    3. Resolution: The yellow flower seems to be of a different resolution compared to the surrounding leaves, with the flower appearing slightly more blurred, which is a common artifact of image tampering.
    4. Shadows: There are no corresponding shadows under the yellow flower that match the lighting of the rest of the image, which would be expected if the flower was originally part of the scene.
    5. Perspective relationship: The size and angle of the yellow flower do not fully conform to the perspective of the surrounding leaves, making it look out of place.
    6. Physical laws: The absence of a shadow under the flower defies the physical law of light, which would naturally produce a shadow on the leaves below.

Can you confirm if this picture was taken directly by the camera? Provide your conclusion along with reasons. If tampered, please show the mask of the altered area.

1. Whether the picture has been tampered with / Description of the tampered area:
    1. The picture has been tampered with, the tampered area is centrally located in the image, occupying a significant portion of the middle ground. It features a penguin standing on a tatami mat floor, which is not a typical setting for a penguin, suggesting that the penguin has been digitally inserted into the scene.
2. Judgment basis:
    1. The lighting on the penguin does not match the rest of the scene; it appears to be lit from the front, while the rest of the environment suggests a more diffused and ambient light source.
    2. The edges around the penguin are also unnaturally sharp and do not blend seamlessly with the tatami mat, indicating a possible cut-and-paste manipulation.
    3. The perspective is off as well; the penguin's size and positioning do not conform to the spatial arrangement of the room, making it appear out of place.
    4. Shadows are not consistent either; the penguin casts no shadow on the tatami mat, which defies the physical laws of lighting and shadow, considering the clear shadows present in the rest of the room.
    5. The resolution around the penguin seems to differ from the rest of the image, with the penguin appearing slightly more pixelated, which could be a sign of a lower resolution image being inserted into a higher resolution scene.

总结：

论文的图文对齐多模态让单一视觉或文本特征提取的欠缺因为模态的对齐可以起到互补的作用。在泛化性方面，则是将自然场景 ps、人脸 ps、生成式虚假图统一到一个模型。当单一的范式（图像）无法解决问题时可以寻求其他的解决方案（文本描述）。

缺点：

1.作者在对比主流 CNN 篡改检测方法时，并非直接使用公开模型，而是在新构建的 **MMTD-Set 数据集** 上重新训练这些传统方法，然后与 FakeShield 进行比较。这种做法在一定程度上提升了基线性能，但同时也可能导致结果倾向于验证 MMTD-Set 的特定特征分布，从而影响公正性。

2.在多模态大语言模型的对比中，论文将经过微调（fine-tuning）的 FakeShield 与未经过微调的其他 M-LLM 方法进行性能比较，这种设定本身并不对等。同时，FakeShield 的视觉模块采用 Transforme 架构，却未与近期基于 Transformer 的最新图像篡改检测模型或论文（如 ViT、Swin Transformer 等）进行比较，而是主要对比了三年前的主流 CNN 模型（如 CAT-Net、MVSS-Net），从研究严谨性上看对比选择略显滞后。