基于协同注意力解释的视觉语言预训练模型多模态对抗攻击方法

韩骐鸿[1)] 崔展齐[1)] 陈 翔[2)] 陈菁菁[1)] 李 莉[1)]

[1)](北京信息科技大学计算机学院 北京 100101)

[2)](南通大学信息科学技术学院 江苏 南通 226019)

摘 要 随着视觉语言预训练模型(Vision-Language Pre-training Models,VLPMs)在图像到文本检索(Image-to-Text Retrieval,TR)、文本到图像检索(Text-to-Image Retrieval,IR)、视觉定位(Visual Grounding,VG)和视觉蕴含(Visual Entailment,VE)等多模态任务中的广泛应用,其可靠性和安全性问题逐渐引发关注。对抗攻击是评估VLPMs鲁棒性的重要手段,但现有方法大多依赖于全局扰动策略,未能结合模型内部注意力机制生成针对性的扰动,导致攻击样本缺乏解释性、扰动区域不集中,难以有效干扰模型关键决策过程。为解决上述问题,本文提出了一种基于协同注意力解释的多模态对抗攻击方法CoAtt-attack。CoAtt-attack利用Co-Attention机制提取图文对齐中的注意力区域,引导图像模态中的全局扰动在扰动空间上聚焦于VLPM决策所依赖的关键区域,从而生成更具针对性的图像扰动;同时,文本模态基于BERT-Attack方法生成语义一致的文本扰动,协同生成自然性更强、干扰性更高的对抗样本。本文选取三种具有代表性的VLPM模型作为测试对象,包括融合型结构的ALBEF模型和TCL模型,以及对齐型结构的CLIP模型。测试ALBEF、TCL和CLIP模型的实验结果表明,CoAtt-attack在IR、TR、VG与VE等任务中的攻击成功率相较于现有方法(Co-attack、VLAttack和SSAP)提升了2.04%~53.58%,并显著降低了生成图像对抗样本的LPIPS值,有效提升了图像对抗样本的真实性和自然性。

关键词 视觉语言预训练模型;多模态;对抗攻击;注意力机制;协同注意力

中图法分类号 TP311 **DOI号** 10.11897/SP.J.1016.2026.00423

Co-Attention Interpretability Based Multimodal Adversarial Attack for Vision-Language Pre-training Models

HAN Qi-Hong[1)] CUI Zhan-Qi[1)] CHEN Xiang[2)] CHEN Jing-Jing[1)] LI Li[1)]

[1)](School of Computer Science and Technology, Beijing Information Science and Technology University, Beijing 100101)

[2)](School of Information Science and Technology, Nantong University, Nantong, Jiangsu 226019)

Abstract With the rapid progress of multimodal learning, Vision-Language Pre-training Models (VLPMs) have been widely adopted in tasks such as Image-to-Text Retrieval (TR), Text-to-Image Retrieval (IR), Visual Grounding (VG), and Visual Entailment (VE). These models have shown remarkable performance and have been increasingly deployed in real-world scenarios, including multimodal retrieval, autonomous driving, content moderation, and human-computer interaction. However, when dealing with complex cross-modal reasoning tasks, VLPMs remain highly sensitive to data bias and adversarial perturbations, which can substantially degrade their

收稿日期:2025-04-25;在线发布日期:2025-10-20。本课题得到江苏省前沿引领技术基础研究专项(BK20202001)、北京信息科技大学"勤信人才"培育计划项目(QXTCP B202406)资助。**韩骐鸿**,硕士研究生,中国计算机学会(CCF)学生会员,主要研究领域为可信人工智能。E-mail:qihong_han@bistu.edu.cn。**崔展齐**(通信作者),博士,教授,博士生导师,中国计算机学会(CCF)高级会员,主要研究领域为智能软件工程及可信人工智能。E-mail:czq@bistu.edu.cn。**陈 翔**,博士,副教授,中国计算机学会(CCF)高级会员,主要研究领域为软件测试、软件维护、经验软件工程和软件仓库挖掘。**陈菁菁**,博士,实验师,主要研究领域为光学无损检测及机器视觉。**李 莉**,博士,副教授,中国计算机学会(CCF)会员,主要研究领域为数据科学与智能系统、数据融合。

prediction accuracy and even trigger safety-critical failures. As a result, assessing the robustness of VLPMs has become an essential challenge in the field of multimodal artificial intelligence. Adversarial attacksare an effective approach to evaluate the robustness of VLPMs. Nonetheless, existing adversarial methods mainly rely on global perturbation strategies, without taking into account the model's internal attention mechanisms that determine cross-modal alignment. Consequently, the generated adversarial examples usually lack interpretability, exhibit scattered perturbation regions, and fail to accurately disrupt the model's core decision-making process. To overcome these limitations, this paper proposes a Co-Attention Interpretability Based Multimodal Adversarial Attack, referred to as CoAtt-attack. The proposed method introduces a Co-Attention mechanism to identify cross-modal semantic alignment regions that capture the interactions between visual features and textual tokens. These regions correspond to the image areas that are semantically correlated with key textual elements and represent the crucial basis for model alignment and reasoning. By leveraging these attention maps, CoAtt-attack guides perturbations to focus on critical visual regions, thereby generating targeted and semantically coherent adversarial samples. Unlike conventional global attacks, CoAtt-attack not only optimizes the perturbations based on the task loss, but also exploits the internal attention distribution of VLPMs to constrain the perturbation generation process. This design ensures that perturbations are spatially concentrated in the most influential areas of the model′s decision path, enhancing both attack efficiency and interpretability. To validate the effectiveness of the proposed approach, three representative VLPMs have been selected for evaluation, including two fusion-based models (ALBEF and TCL) and one alignment-based model (CLIP). Extensive experiments have been conducted on multiple benchmark datasets, and the results have demonstrated that CoAtt-attack has achieved an improvement of 2.04%—53.58% in attack success rates compared with state-of-the-art baselines such as Co-attack, VLAttack, and SSAP in IR, TR, VG, and VE tasks. In addition, CoAtt-attack has generated adversarial images with superior perceptual qualityover these baseline methods, as evidenced by significant improvements in three quantitative metrics: the average LPIPS has decreased by 0.0065, 0.5023, and 0.3649, while the average PSNR (measured in dB) has increased by 3.81 dB, 23.24 dB, and 17.97 dB, and the average SSIM has improved by 0.0037, 0.3128, and 0.2395, respectively. Furthermore, the adversarial texts produced by CoAtt-attack have achieved an average BERTScore of 0.8507, indicating strong semantic consistency in the text modality. In summary, CoAtt-attack enhances the specificity, interpretability, and perceptual realism of multimodal adversarial attacks while maintaining high attack success rates. The proposed method provides a reliable and explainable framework for robustness evaluation of vision-language models. Future work will extend CoAtt-attack to black-box and transfer-based attack settings, exploring lightweight attention estimation and feature alignment strategies to further promote the robustness and interpretability of multimodal learning.

Keywords vision-language pre-training models; multimodal; adversarial attacks; attention mechanism; co-attention

1 引 言

随着深度学习和自然语言处理技术的快速发展，视觉语言预训练模型(Vision-Language Pre-training Models，VLPMs)在图像到文本检索(Image-to-Text Tetrieval，TR)、文本到图像检索(Text-to-Image Retrieval，IR)、视觉定位(Visual Grounding，VG)、视觉蕴含(Visual Entailment，VE)等跨模态信息处理任务中展现出显著性能优势[1-7]。

这些模型在执行多模态任务时，通常依赖注意力机制（Attention Mechanism）来捕捉图像与文本之间的语义关联。通过自注意力与交叉注意力模块，VLPM能够在输入数据中识别出与当前任务高度相关的关键区域或重要词汇，从而增强语义对齐与推理能力[8-9]。在实际应用中，VLPMs首先通过在大规模未标记的图像-文本数据集上进行预训练来学习图像与文本之间的语义相关性，然后在不同的下游视觉语言任务上使用带标注的图像文本对进行微调，以适应不同的视觉语言任务[10-12]。

在实际部署中，VLPMs也存在一些困难和挑战。例如，模型在处理复杂的跨模态推理任务时，容易受到数据偏见和对抗性数据的影响，导致模型的预测性能显著下降[13]。此外，由于VLPMs使用庞大且复杂的训练数据，并且模型的内部结构复杂，它们表现出“黑箱”特性，导致模型决策过程难以解释[14]。这些问题不仅降低了模型的鲁棒性，还可能在实际应用中引发安全隐患。例如，在多模态医疗检索中，文本描述中的空间性词汇（如“left lung lesion”）若未被模型准确建模，可能导致返回图像与实际部位不符（如右肺区域的影像），误导医生诊断[15]；在自动驾驶系统中，视觉输入中的轻微干扰可能使模型误识交通标志或行人行为，造成安全风险[16]；在内容审核与问答系统中，模型对模态信息的理解偏差也可能引发误判，带来用户体验下降甚至法律责任[17]。因此，研究有效的VLPM测试方法，对于深入理解模型鲁棒性、提升模型安全性具有重要的理论意义和实际价值。

目前，主要通过单模态对抗攻击和多模态对抗攻击两类方法对VLPMs进行测试[18-22]。单模态对抗攻击方法仅在图像或文本单一模态上进行扰动，例如Yang等人[18]提出的SSAP（Single-Source Adversarial Perturbations）方法利用投影梯度下降（Projected Gradient Descent，PGD[19]）算法，通过最小化交叉熵损失在图像模态中添加微小扰动，从而误导VLPMs的判断；Li等人[20]提出的BERT-Attack方法利用BERT模型对文本中的单词进行语义相似的替换，以生成语义高度相似但能误导模型决策的对抗样本。由于这类方法忽视了图像与文本模态之间的关联关系，仅对单一模态进行扰动，容易被另一模态的信息所补偿，导致攻击效果受限。为了解决这一问题，研究者们提出多模态对抗攻击方法来测试VLPMs，例如Zhang等人[21]提出的Co-attack（Collaborative Multimodal Adversarial Attack）方法先通过文本模态的词替换确定图像扰动方向，再对图像扰动进行迭代优化，从而实现对图像和文本的扰动以攻击VLPMs；Yin等人[22]提出的VLAttack（Vision-Language Attack strategy）方法通过结合单模态攻击与多模态攻击的协同策略，逐步生成图像和文本的对抗样本以攻击VLPMs。这类方法多以全局扰动方式来影响模型输出，而未能结合模型的注意力分布信息来识别并集中干预对决策起关键作用的区域。此外，这些方法通常依赖于输出损失的变化来优化扰动，缺乏对扰动与模型决策机制之间关联的有效建模，从而难以实现更精确、更高效的攻击。例如，在自动驾驶场景中，即使通过在图像输入上添加全局扰动来成功欺骗了模型，也无法判断模型在识别行人或交通标志时究竟依赖了哪些图像特征，难以确定模型的脆弱性具体源于哪些输入特征或区域的误导。这种对抗扰动的不可解释性使得攻击效果难以复现或改进，也不能为模型的鲁棒性提升提供针对性的优化建议[23]。

为解决上述方法中添加的全局扰动分布离散且难以集中于模型决策的关键区域，导致削弱了对抗攻击的针对性与有效性的问题。本文提出了一种基于协同注意力解释的多模态对抗攻击方法CoAtt-attack（Co-Attention Interpretability Based Multimodal Adversarial Attack）。该方法通过引入能够表征图像与文本对齐关系的Co-Attention机制来捕捉模型在决策过程中关注的跨模态语义关联区域。这些区域是图像中与文本关键词存在高度语义相关的区域，是模型执行图文语义对齐的关键依据。CoAtt-attack基于这些区域生成更加针对性的扰动。与现有方法不同的是，CoAtt-attack不仅依赖于输出损失的优化，还借助模型内部的注意力分布信息来引导扰动生成，从而使扰动更集中于决策过程中最为重要的区域。通过这一机制，CoAtt-attack能够在提高攻击效率和图像真实性的同时，利用注意力解释提供对扰动效果的合理解释与结构化支持。实验结果表明，CoAtt-attack在IR、TR、VG与VE等任务中的攻击成功率相较于现有方法（Co-attack[21]、VLAttack[22]和SSAP[18]）提升了2.04%～53.58%，并显著降低了生成图像对抗样本的LPIPS（Learned Perceptual Image Patch Similarity）值，有效提升了图像对抗样本的真实性与自然性。

本文的主要贡献如下。

（1）提出一种结合协同注意力解释的VLPMs测试方法CoAtt-attack，该方法利用Co-Attention机

制深入分析视觉语言模型决策过程中关注的跨模态语义关联区域，并在此基础上生成有针对性且可解释性更强的图像和文本扰动，提升了攻击成功率和对抗样本图像的真实性。

（2）基于CoAtt-attack实现了原型工具，并选取ALBEF、TCL和CLIP三种视觉语言预训练模型作为攻击对象，分别针对IR、TR、VG和VE四种下游任务进行对比实验，以验证CoAtt-attack的有效性。

本文剩余内容结构安排如下：第2节介绍相关工作；第3节介绍本文方法的动机和相关示例；第4节介绍CoAtt-attack的框架和具体流程；第5节介绍实验设计；第6节对实验结果进行分析；第7节对本文的工作进行讨论；第8节对全文进行了总结，并对未来工作进行展望。

2 基础知识与相关工作

本节将介绍VLPMs的模型结构和典型下游任务，并对现有的VLPMs对抗攻击方法进行介绍和分析。

2.1 视觉语言预训练模型

VLPMs旨在通过对大规模图像-文本对进行预训练，以提升下游多模态任务的性能[24]。早期的大多数工作基于预训练的目标检测器，利用区域特征来学习视觉语言表示[25-28]。这类方法虽然取得了一定效果，但由于图像特征依赖于外部检测器，难以实现端到端的训练，限制了模型对图像整体语义的理解能力。随着视觉Transformer（Vision Transformer，ViT）的广泛应用[29-31]，一些研究开始将ViT作为图像编码器使用，将输入图像划分为若干图像块（patch），并以端到端的方式提取图像特征。这类方法摆脱了对目标检测器的依赖，使图像特征提取更加灵活统一，同时增强了模型对图像整体语义的建模能力，在多个下游任务中取得了显著性能提升。

根据模型结构的不同，VLPMs可以分为融合型和对齐型两种典型类型的VLPMs。融合型VLPMs（如ALBEF[32]、TCL[33]）首先通过独立的单模态编码器分别处理文本标记（token）和图像特征，获得各自模态的嵌入表示，再进一步利用一个多模态编码器对图像与文本的嵌入表示进行融合，得到跨模态联合表示。这种融合结构能够有效捕捉图文模态间的深层语义关联关系，更适合需要精细跨模态推理的下游任务。与此不同的是，对齐型VLP模型（如CLIP[34]）仅包含两个单模态编码器，分别独立地对图像和文本模态进行编码，将它们映射到一个统一的语义嵌入空间。在这一过程中，模型通过对比学习策略，使语义相关的图像与文本在嵌入空间中距离更近，而无关样本之间的距离更远。这种模型结构较为简单，适用于大规模跨模态检索任务。本文关注当前主流的融合型与对齐型视觉语言预训练模型，并对这些模型在典型下游任务中的鲁棒性进行系统测试与分析。

2.2 视觉语言下游任务

本文关注VLPMs的四类代表性下游任务，即TR、IR、VE以及VG，它们可用于评估VLPMs在图文对齐、跨模态推理与定位等方面的能力。以下将对这些任务及主流VLPMs处理这些任务的方式进行简要介绍。

图文检索任务包含TR和IR两类子任务。ALBEF和TCL对这两个任务的处理流程较为相似。具体而言，这类模型首先计算所有图像-文本对中图像与文本嵌入向量之间的相似度分数，并根据相似度对所有样本进行排序，从中筛选出最相关的Top-N图文对（即相似度最高的前N个图文对）。随后，这些Top-N对会被输入到图像-文本匹配模块中，通过计算得到的匹配得分进行排序，以确定最终的检索结果。相比之下，CLIP在执行TR和IR任务时更为简洁，直接使用图像和文本嵌入之间的相似度进行排序。

视觉蕴含是一种跨模态推理任务，目标是判断图像与文本之间的语义关系属于“蕴含（entailment）”、“中性（neutral）”还是“矛盾（contradiction）”。ALBEF和TCL模型在处理该任务时，会将其建模为一个三分类问题，从多模态编码器的融合表示中提取用于整体语义建模的特殊位置向量（如[CLS]标记），并以此作为输入，通过全连接层预测三个类别的概率分布。

视觉定位任务的目标是根据输入文本的描述，在图像中定位对应的区域。ALBEF扩展了Grad-CAM方法[35]，利用得到的注意力热图对候选检测框进行排序，从而实现对目标区域的定位[36]。TCL则通过多模态特征中的跨模态注意力得分，评估文本与图像中各区域之间的关联性，并根据注意力强度对候选区域进行排序，从而完成定位。

2.3 视觉语言预训练模型对抗攻击方法

VLPMs虽然在多个下游视觉语言任务中取得了优异表现，但已有研究表明[18]，这类模型对输入中的扰动仍然较为敏感，易受到对抗样本的干扰。因

此，大量研究关注如何通过构造对抗性输入样本来测试模型的鲁棒性。根据扰动施加的模态不同，现有的对抗攻击方法大致可分为单模态对抗攻击方法和多模态对抗攻击方法。

2.3.1　单模态对抗攻击方法

单模态对抗攻击方法仅针对VLPMs的一个模态（图像或文本）施加扰动，通常用于评估模型在面对部分输入变化时的鲁棒性。这类方法的优点在于实现较为简单，可控性强，便于对各模态输入对模型行为的影响进行定量分析。

对于图像模态，由于图像是连续的像素空间，模型对图像的预测输出对每个像素是可微的，因此可以直接利用梯度信息来生成对抗扰动[37]。其中，最常用的方法是PGD[19]，该方法通过计算图像输入对模型损失函数的梯度方向，在扰动范围受限的条件下对像素值进行迭代更新，从而生成感知上与原图像接近、但能误导模型的对抗图像。例如，Yang等人[18]提出的SSAP方法在图像模态中引入基于PGD的扰动，即使在文本输入保持不变的情况下，仍能显著削弱模型的图文匹配能力。这表明，单独对图像模态施加扰动即可有效误导模型，验证了图像模态在视觉语言模型决策过程中的关键作用。与图像模态不同，文本输入是由离散的词元组成，模型对其输入不可微，无法直接通过梯度进行扰动[38]。因此，文本模态的对抗攻击通常采用词级替换策略，通过修改输入中的部分关键单词，在保持语义合理与语法正确的前提下，诱导模型产生错误判断。其中，BERT-Attack[20]是最具代表性的方法之一。该方法利用BERT模型构建上下文语义表示，识别出对模型预测影响最大的词汇，并为其生成多个语义相似的替代项，最终选择最具攻击性的替换结果以生成文本对抗样本。这类方法在视觉语言任务中同样展现出较强的攻击能力，即使图像保持不变，仅通过轻微的文本修改也能对模型造成误导。

尽管单模态对抗攻击方法在实验设计中更具可控性，有助于观察模型对不同输入模态的响应差异，但这类方法往往忽视了图像与文本之间的跨模态信息融合过程，无法对模型的联合语义建模过程形成有效干扰。为了更全面地评估VLPMs在多模态任务中的鲁棒性，研究者们开始转向多模态对抗攻击方法。

2.3.2　多模态对抗攻击方法

与单模态攻击方法相比，多模态对抗攻击能够同时干扰VLPMs中的图像与文本输入，从而更有效地破坏其跨模态对齐机制，增强攻击强度。在图文检索、视觉定位等典型下游视觉语言任务中，模型通常需要融合图像与文本的深层语义特征才能完成准确推理。单一模态的扰动可能会被另一模态的信息所补偿，导致攻击效果受限。而多模态攻击通过在两个模态中联合生成干扰，能够削弱模型对跨模态语义的建模能力，因此成为逐渐视觉语言预训练模型鲁棒性测试的重要方法。

目前具有代表性的多模态攻击方法包括Zhang等人[21]提出的Co-attack和Yin等人[22]提出的VLAttack。Co-attack通过在文本模态中生成语义等价的词替换扰动，并基于该文本引导图像模态进行针对性扰动优化，从而增强两个模态扰动之间的一致性。该方法将扰动效果映射到多模态嵌入空间中，通过最大化扰动前后嵌入向量之间的距离来实施攻击。尽管Co-attack在提升攻击成功率方面表现优越，但其攻击流程依赖于固定的扰动顺序，导致图像扰动强依赖于已扰动文本，忽视了图像模态在模型决策中的独立作用。同时，该方法未考虑VLPMs在决策过程中对不同输入区域的差异化关注（如注意力机制），其扰动生成方式缺乏针对性，难以精确干扰模型在决策时依赖的关键特征区域。VLAttack方法则通过交替优化图像与文本模态的扰动，引入模态协同策略与语义保持约束，以提升对抗样本的自然性与一致性。该方法在两个模态间轮流更新扰动，引导扰动分别作用于两个输入通道，并通过迭代交叉搜索攻击策略（ICSA）动态调整图像与文本的扰动过程。其扰动生成仍以模型最终输出结果为唯一优化目标，未利用模型内部的注意力结构进行引导，导致扰动分布较为分散，不能集中于模型实际依赖的关键区域。由于干扰缺乏针对性，这类方法的干扰效果存在明显不足，特别是当模型在特定区域进行语义对齐或特征提取时，全局扰动难以有效削弱模型的判别能力。此外，这种缺乏针对性的全局扰动方式容易对与文本语义相关性不大的图像背景区域进行不必要的修改，导致显著降低图像对抗样本的真实性与感知质量。

从上述分析可知，现有多模态对抗攻击方法尽管通过联合扰动图像与文本模态来增强攻击效果，并在一定程度上提高了攻击成功率，但仍存在以下两个关键问题。（1）扰动生成未有效利用VLPMs的注意力机制。无论是Co-attack中的文本引导图像扰动，还是VLAttack中基于输出结果的交替优化，

这些方法均未利用VLPMs的注意力机制来生成扰动，无法精准定位决策过程中真正敏感的区域，导致扰动分布离散且难以集中于模型决策的关键区域，从而削弱了对抗攻击的针对性与有效性。(2)生成的对抗样本缺乏解释性。现有方法多通过全局扰动的方式干预输入空间，而未能识别并有针对性地干预模型决策中起关键作用的区域。这种扰动策略未有效利用模型的注意力分布信息，难以明确哪些扰动区域对VLPMs预测结果起到了关键作用，导致生成的对抗样本缺乏针对性与解释性。

为解决上述问题，本文提出了一种基于协同注意力解释的多模态对抗攻击方法CoAtt-attack。该方法在前向传播阶段引入模型自身的跨模态注意机制，以识别模型在决策过程中最为关注的图像区域与文本关键词，并据此引导扰动生成。与现有方法相比，CoAtt-attack更加关注模型内部的决策机制与图文语义对齐关系，旨在实现更具针对性、结构感知性与可解释性的对抗扰动生成策略，为视觉语言模型的鲁棒性测试提供更具分析能力与实用价值的技术手段。

3 动机示例

实际上，VLPMs在进行跨模态任务时，通常通过注意力机制捕捉图像与文本之间的关联信息，为理解VLPMs的决策过程提供了更加结构化的线索。在多模态架构中，VLPMs通过自注意力与交叉注意力模块计算不同模态特征之间的相似性，从而定位出与输入文本高度相关的图像区域[9]。这一特性为构建更具可解释性的对抗攻击方法提供了潜在的思路。为更好地理解模型的决策机制，我们基于Co-Attention机制对ALBEF模型在图文检索任务中的注意力进行了计算，并进行了可视化分析。如图1所示，注意力热力图展示了模型在输入图像与文本对齐过程中所关注的显著区域。图中左侧为原始图像与对应的文本描述，右侧为通过模型生成的注意力热力图。注意力热力图反映了模型在跨模态对齐过程中对图像中不同区域的关注程度，颜色越接近红色表示模型的注意力越集中。从图中可以看出，模型在对输入图像与文本进行语义匹配时，并非

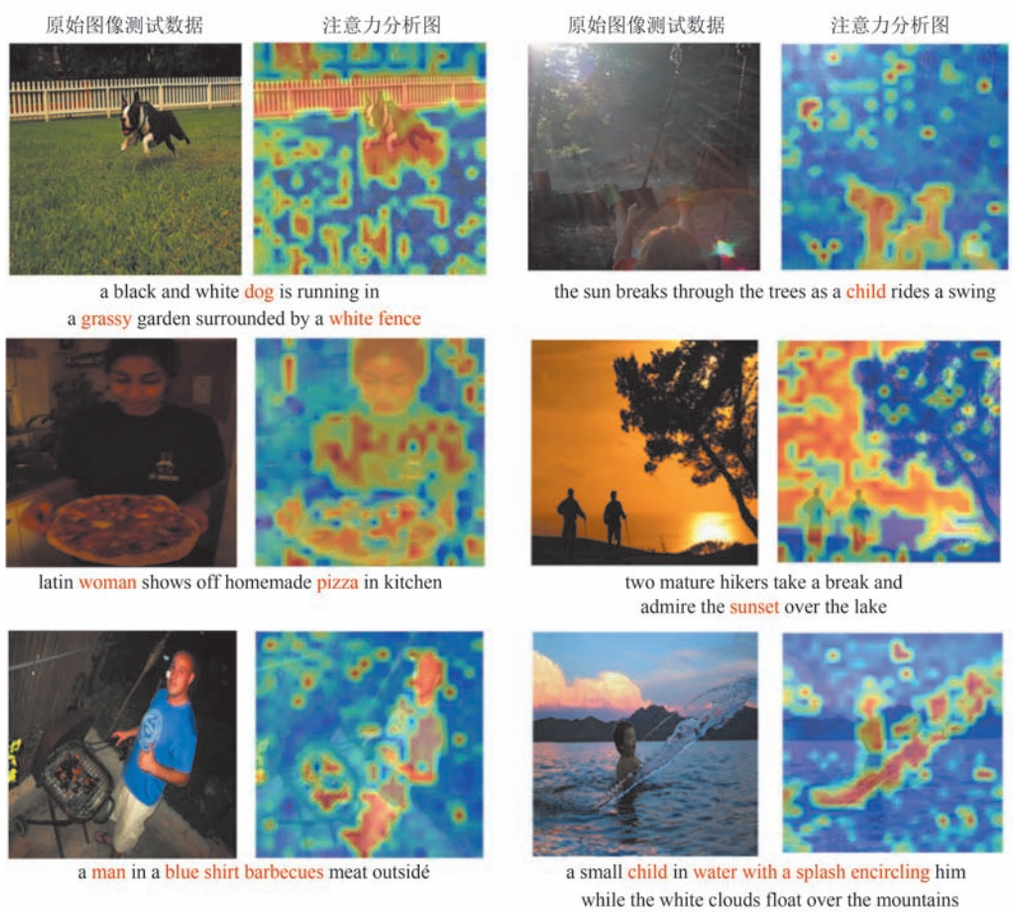

图1 ALBEF执行IR任务时的注意力可视化示例

关注整幅图像，而是聚焦于与文本描述中关键词相关的区域。例如，处理描述为“a black and white dog is running in a grassy garden surrounded by a white fence”的图像时，模型的注意力主要集中在图像中“dog”、“grassy garden”以及“white fence”等与文本内容高度相关的区域；处理描述为“a small child in water with a splash encircling him…”的图像时，模型的注意力显著集中在与“child”和“splash”等相关的区域。这表明模型在决策过程中将会根据输入文本中的关键词所对应的图像区域进行语义匹配和推理。

然而，现有的多模态对抗攻击方法未有效利用模型内部的注意力机制，通常会生成全局扰动，既无法集中作用于关键区域，也容易对无关区域产生不必要的干扰。这种无差别扰动不仅效率较低，且容易破坏原始图像的自然性与语义一致性，进而降低测试输入的真实性和有效性。此外，这种全局扰动的策略生成的对抗样本通常是不可解释的。如图2展示了Co-attack方法中，通过投影梯度下降(PGD)生成的全局扰动灰度图。可以看出，生成的噪声覆盖了整个图像，无论是与输入文本强相关的区域还是背景区域都受到了不同程度的扰动。这种全局化、无结构的扰动模式表明，现有方法无法有效捕捉模型在决策过程中真正依赖的关键区域，无法获取导致模型误判的图像区域。类似的问题在VLAttack方法中也同样存在，尽管引入了交替优化策略，但其扰动生成过程仍主要依赖于输出损失的优化，而未能对模型的注意力机制进行建模与利用。

基于上述分析，我们提出了一种基于协同注意力解释的对抗攻击方法CoAtt-attack。CoAtt-attack通过分析模型生成的注意力矩阵，识别出模型在决策过程中最为关注的区域，并据此设计针对性扰动。与传统的全局扰动策略相比，基于注意力解释的扰动生成方式具有以下优点：(1)能够精准定位并干扰模型依赖的关键区域；(2)通过减少对无关区域的干扰，保持原始图像的自然性和语义一致性，从而提高测试输入的真实性；(3)通过与模型的注意力机制相结合，使得攻击过程可解释性更强，有助于揭示模型在决策过程中存在的潜在薄弱点。

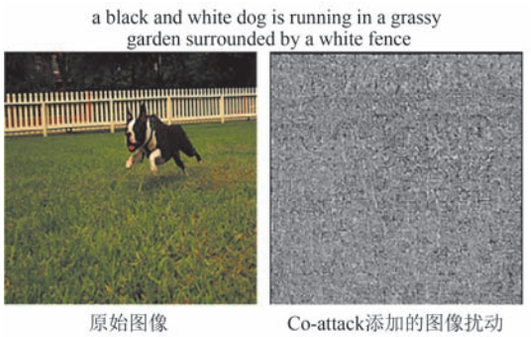

图2　Co-attack添加图像扰动示例

4　方法设计

针对现有方法中全局扰动随机性高、难以集中于关键区域、破坏图像真实性等问题，CoAtt-attack通过引入Co-Attention机制对视觉语言模型的决策过程进行解释分析，CoAtt-attack能够有效识别模型在图文对齐过程中最为关注的区域，并据此生成更加精确且具有语义一致的扰动。CoAtt-attack的框架如图3所示，主要由解释分析、图像扰动掩膜与文本扰动生成三个部分组成。

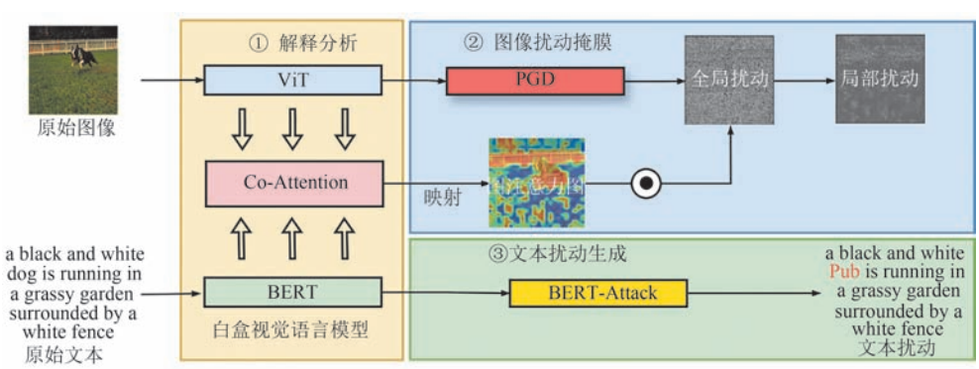

图3　CoAtt-attack框架图

4.1　解释分析

为解决现有对抗攻击方法中未有效利用模型决策机制的问题，CoAtt-attack在前向传播阶段引入Co-Attention机制，用于分析和提取模型在处理

图文输入时关注的区域，并以此引导生成扰动。与传统的自注意力（Self-Attention）或单向交叉注意力（Cross-Attention）不同，Co-Attention具备双向建模能力，能够同时反映“图像对文本”和“文本对图像”的相互关注关系，从而更全面地捕捉图文之间的语义关联关系[39]。相比其他注意力机制，Co-Attention机制更契合当前主流融合型视觉语言模型（如ALBEF、TCL）的特征融合方式，能够提供更丰富、更贴近模型判别依据的注意力分布。如图4所示，CoAtt-attack方法首先将输入图像和文本分别送入视觉编码器（ViT）和文本编码器（BERT）以提取图像和文本模态的嵌入特征。随后，这些特征通过Co-Attention机制生成图像和文本的注意力分布，并根据注意力分布引导后续扰动，使扰动与模型的实际决策行为相关，更具针对性。

具体而言，在VLPM处理下游任务的前向传播阶段，给定图像-文本对$(I,\ T)$，图像I首先被划分为n个图像块，经视觉Transformer编码为图像嵌入矩阵$V\in\mathbb{R}^{d\times n}$，文本$T$被分词器（Tokenizer）处理为$m$个词元，经Transformer编码为文本嵌入矩阵$Q\in\mathbb{R}^{d\times m}$，其中$d$为图像和文本嵌入矩阵的嵌入维度。Co-Attention机制会根据图像嵌入矩阵和文本嵌入矩阵来计算两种模态间的亲和矩阵（Affinity Matrix）C，具体计算方式如公式（1）所示：

$$C=\tanh\left(Q^{T}W_{b}V\right) \tag{1}$$

其中，$W_{b}\in\mathbb{R}^{d\times d}$。

亲和矩阵C被用来衡量图像和文本不同位置间或词元间的语义关联性，也可作为计算跨模态注意力分布的重要中间表示。该计算方式通过双线性投影将文本嵌入$Q=\{q_1,q_2,\cdots,q_i\}$与图像嵌入$V=\{v_1,v_2,\cdots,v_j\}$映射到共享语义空间中，其中W_{b}为可学习的跨模态投影参数。在共享空间中，如果词元q_i和某个图像块区域v_j语义上强相关，那么q_i和$W_{b}v_j$之间的投影内积就会较大；若它们语义无关则内积较小或为负。亲和矩阵C中的每个矩阵元素c_{ij}反映了词语q_i和图像区域v_j在嵌入空间中的语义对齐度。通过该亲和矩阵，模型在后续注意力分布计算中能够基于统一的跨模态关联信息，建立图像与文本之间的语义对齐关系。为根据C生成图像或文本的注意力分布，一种常用的策略是，对每个模态的位置，取其与另一模态中所有位置之间亲和度的最大值作为注意力得分。例如，图像中第n个图像块的注意力，可以由其与所有文本词元之间亲和度中的最大值来确定，表示该图像块在整个文本语境中最相关的匹配程度；同样地，文本中第m个词元的注意力得分也可由其与所有图像区域的最大亲和度决定，用于衡量该词元在图像中的语义相关性。然而，这种最大化亲和度的操作会因为模型只关注最大亲和度，忽略图像和文本之间那些较弱但仍然重要的关联，其过于依赖亲和矩阵中相关性最大的位置，无法捕捉到图像和文本之间复杂的非线性关系[39]。为更准确地反映图文之间的复杂非线性关系，我们将亲和矩阵C作为一种特征表示，并通过公式（2）和（3）计算图像和文本的注意力分布：

$$\begin{cases}H^{v}=\tanh\left(W_{v}V+(W_{q}Q)C\right)\\H^{q}=\tanh\left(W_{q}Q+(W_{v}V)C^{T}\right)\end{cases} \tag{2}$$

$$\begin{cases}a^{v}=\mathrm{softmax}\left(w_{hv}^{T}H^{v}\right)\\a^{q}=\mathrm{softmax}\left(w_{hq}^{T}H^{q}\right)\end{cases} \tag{3}$$

其中，W_{v}、$W_{q}\in\mathbb{R}^{d\times d}$，$w_{hv}$、$w_{hq}\in\mathbb{R}^{d}$为权重参数。$a^{v}\in\mathbb{R}^{n}$和$a^{q}\in\mathbb{R}^{m}$分别表示图像与文本模态的注意力分布向量。

图注意力分布向量a^{v}表示模型对图像模态中的n个图像块的关注程度。由于对图像的扰动是在像素级别进行的，为了将图像注意力分布与实际图像区域对应，CoAtt-attack需要将a^{v}映射至与原图尺寸一致的像素空间。具体过程如公式（4）所示：

$$A_{\text{pixel}}=Bilinear\left(\mathrm{reshape}\left(a^{v}\right),\ size=(h,w)\right) \tag{4}$$

其中，$A_{\text{pixel}}\in\mathbb{R}^{h\times w}$表示图像像素空间中的注意力分布矩阵，$A_{\text{pixel}}$中每个元素对应一个像素点的注意力权重，其大小反映了模型在判别过程中对该像素位置的关注程度。注意力权重值越高，表示该像素在当前下游任务中对模型决策的重要性越大。$h\times w$表示原图像的尺寸大小。

CoAtt-attack将a^{v}重构为二维形式，再使用双线性差值（Bilinear Interpolation）将其上采样为与原始图像大小一致的注意力图。图4中最右边的热力图是该图像注意力分布矩阵的可视化结果。从热力图中可以明显观察到模型的关注区域主要集中在与文本描述中关键词语义更相关的图像区域（图中红色区域）。这些区域揭示了模型在语义对齐与判断过程中所依赖的关键图像特征，在后续的扰动掩码阶段中将被视为重要区域。

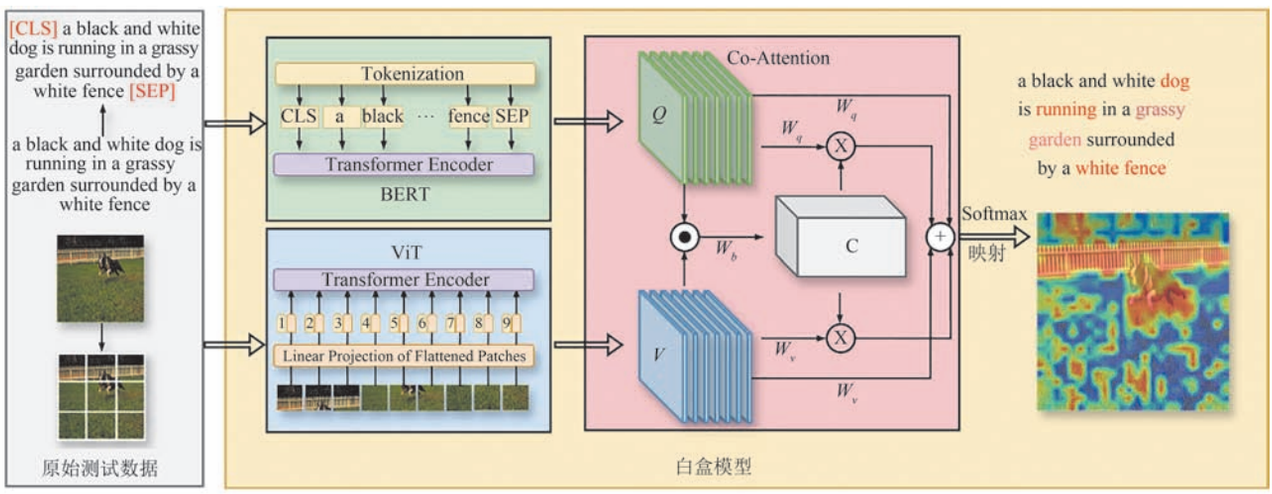

图4　Co-Attention解释分析示意图

4.2　图像扰动掩膜

Co-attack、VLAttack等扰动方法直接在整个图像输入上添加全局扰动，未考虑不同区域对模型判别过程的重要性差异，可能导致关键区域未被有效扰动，而无关区域则被不必要地修改，降低对抗样本的有效性和真实性。为解决这一问题，实现更具解释性和针对性的图像扰动，CoAtt-attack将像素级图注意力分布矩阵A_{pixel}引入图像模态扰动生成过程，并使用扰动掩膜操作从全局扰动中去除非重要区域扰动，使其更加集中于模型关注的核心区域。扰动掩膜的具体过程如公式(5)和公式(6)所示：

$$A_{\text{pixel}}{}'=\frac{A_{\text{pixel}}-\min\left(A_{\text{pixel}}\right)\times E}{\max\left(A_{\text{pixel}}\right)-\min\left(A_{\text{pixel}}\right)} \tag{5}$$

$$Mask\left(P,A_{\text{pixel}}{}'\right)=P\cdot A_{\text{pixel}}{}'=\begin{bmatrix}p_{11}a_{11} & \cdots & p_{1m}a_{1m}\\ \vdots & \ddots & \vdots\\ p_{n1}a_{n1} & \cdots & p_{nm}a_{nm}\end{bmatrix} \tag{6}$$

其中，E表示所有元素均为1且大小与A_{pixel}相同的矩阵，A_{pixel}中越重要区域的值越大。

CoAtt-attack首先采用PGD算法生成图像模态的全局扰动$P\in\mathbb{R}^{h\times w}$，并使用最大最小标准化将$A_{\text{pixel}}$中的元素归一化到0至1之间以得到归一化后的$A_{\text{pixel}}{}'$。然后将全局扰动$P$与$A_{\text{pixel}}{}'$进行点乘掩膜，其中$p_{ij}(1\leqslant i\leqslant n,1\leqslant j\leqslant m)\in P$和$a_{ij}(1\leqslant i\leqslant n,1\leqslant j\leqslant m)\in A_{\text{pixel}}{}'$，分别为原始图像模态测试数据中第$i$行第$j$列像素的针对性扰动值和重要程度。对全局扰动$P$进行扰动掩膜后得到局部扰动$Mask\left(P,A_{\text{pixel}}{}'\right)$。

通过这种掩膜操作，模型关注度较低的区域(如背景)中的扰动被显著抑制，而关键区域的扰动得以保留，从而避免了无差别扰动对图像整体的过多干扰。该策略使得生成的对抗图像在视觉上更加自然，并在感知上更接近原始图像。

4.3　文本扰动

对于图像模态，CoAtt-attack通过注意力机制和掩膜操作实现了对图像的针对性扰动，以提高图像扰动的有效性和真实性。对于文本模态，也希望尽可能生成语义一致但具有误导性的文本输入，以配合图像模态实现更有效的联合攻击。CoAtt-attack集成了基于BERT模型的文本对抗攻击方法BERT-Attack，该方法通过语义保持的单词替换策略，在不破坏文本可读性与语义一致性的前提下，生成可误导模型的对抗样本。其流程包括单词重要性评估、候选替换词筛选，以及对抗样本选择。

首先，BERT-Attack通过逐步遮掩原始文本中每一个单词，并将替换后的遮掩文本输入到视觉语言模型中，以评估每个单词对模型输出的重要性。具体过程如公式(7)所示：

$$Score(w_i)=KL\left(f(T)\,\|\,f(T_{[UNK]_i})\right) \tag{7}$$

其中，$f(\cdot)$表示VLPM对输入文本的输出表示，$T_{[UNK]_i}$表示第i个单词被遮掩的文本，$Score(w_i)$表示原始文本T中第i个单词的重要性得分。

对于由n个单词组成的原始文本$T=\{w_1,w_2,\cdots,w_n\}$，将单词w_i替换为特殊符号$[UNK]$，并获取其对应模型输出表示。通过计算替换前后模型输出之间的KL散度(Kullback-Leibler Divergence)来度量该单词的重要性得分。

然后，对于得分前k的单词，BERT-Attack利用

预训练BERT掩蔽语言模型(Masked Language Model, MLM)生成候选替换词集合S_i。根据替换词集合,得到替换后的候选文本$T_j'(j\in[1,k])$。为了确保语义一致性,对每个候选文本T_j'与原始文本T使用通用句子编码器(Universal Sentence Encoder)计算余弦相似度,计算方式如公式(8)所示:

$$\gamma_j = Cos\left(USE(T), USE\left(T_j'\right)\right) \tag{8}$$

其中,$USE(\cdot)$表示通用句子编码器对文本的编码。当$\gamma_j \geqslant \sigma_s$时,$T_j'$被保留作为语义相似性高的对抗候选文本,$\sigma_s$为设置的语义相似性阈值。

最后,在所有满足语义相似性的候选样本中,选择具有最大攻击效果的文本作为最终对抗样本。具体过程如公式(9)所示:

$$T_{\text{adv}} = \underset{T_j', \gamma_j \geqslant \sigma_s}{\arg\max}\ \mathcal{L}(T_j') \tag{9}$$

其中,$\mathcal{L}(T_j')$表示基于模型输出的攻击损失,用于衡量扰动后文本对模型判别能力的影响。从所有满足语义相似性约束的候选文本中,选择使攻击损失最大的文本作为文本模态的对抗样本T_{adv}。

通过单词重要性评估、候选替换词的筛选以及对抗样本的选择,CoAtt-attack在文本模态上构建了语义一致且具干扰性的文本扰动T_{adv}。该文本扰动将与图像扰动掩膜过程中生成的图像针对性扰动联合构成完整的多模态对抗样本,实现多模态协同扰动,从而削弱模型的语义对齐机制与判别能力。

5 实验设计

我们基于Python 3.8和Pytorch 2.1.0实现了CoAtt-attack原型工具,并在具有64GB RAM,AMD 7950X CPU和NVIDIA RTX 4090 GPU的计算机上进行实验。

5.1 研究问题

为评价CoAtt-attack的有效性,本文设置了以下三个研究问题。

RQ1:与基线方法相比,CoAtt-attack生成对抗攻击样本的攻击效果、图像对抗样本的真实性和文本对抗样本的质量如何?

为评估CoAtt-attack的有效性,我们将其与Co-attack[21]、VLAttack[22]和SSAP[18]三种基线方法就对抗攻击的性能、图像对抗样本的真实性和文本对抗样本的质量进行对比。

RQ2:使用Co-Attention机制作为解释方法如何影响CoAtt-attack对抗攻击的性能?

解释方法可用于解释VLPMs在做出决策时的行为。CoAtt-attack使用Co-Attention机制作为解释方法定位图像输入中的重要区域,通过仅在图像的重要区域添加扰动来降低对图像测试输入的扰动幅度,同时也保留了图像测试输入的对抗攻击性。为了分析使用Co-Attention机制如何影响CoAtt-attack对抗攻击的性能,实验在CoAtt-attack原型工具的基础上去除了Co-Attention机制,将其用于IR和TR任务,并对其对抗攻击的性能的变化情况进行了分析。

RQ3:不同模态的扰动对视觉语言预训练模型的鲁棒性影响如何?

为分析图像模态扰动、文本模态扰动及多模态联合扰动对VLPMs鲁棒性的影响,本文在IR和TR任务中分别对单模态扰动(图像扰动采用PGD方法,文本模态扰动采用BERT-Attack方法)和联合多模态扰动进行了对比,以探讨不同模态扰动对视觉语言预训练模型鲁棒性的影响。

以上研究问题中,RQ1的实验对象覆盖了四种典型视觉语言下游任务,以全面验证CoAtt-attack在多任务环境下的适应能力与通用性。RQ2和RQ3选取了IR与TR两个任务作为代表进行深入分析,这两个任务分别对应图像到文本与文本到图像的语义检索,能够反映VLPM在双向跨模态语义对齐中的推理能力。

5.2 数据集和实验对象

数据集:参考Co-attack[21]、VLAttack[22]等工作,实验选择Flickr30K、MSCOCO、RefCOCO+和SNLI-VE作为数据集。其中Flickr30K和MSCOCO常被用来评估TR和IR任务[40],RefCOCO+常被用来评估VG任务[41],SNLI-VE常被用来评估VE任务[42]。在VE任务中,由于我们只关注对抗攻击的效果,因此实验仅从SNLE-VE测试数据集中选择带有蕴意标签的图像-文本对,而舍弃带有中性或矛盾标签的图像-文本对。

视觉语言预训练模型:实验选择ALBEF[32]、TCL[33]和CLIP[34]三种VLPMs作为被测模型,它们都是编码器模型,均采用视觉Transformer(如ViT[29])和语言Transformer(如Bert[43])来分别处理图像特征和文本特征。但三种模型将图像和文本特征进行融合和对齐的方式存在显著差异。其中,ALBEF模型采用跨模态对齐机制,通过对比学习来实现图像与文本特征的语义对齐,这种方式促使

模型更精准地理解图文之间的语义关联，从而提高模型在下游任务中的视觉语言推理性能。TCL模型不仅关注跨模态之间的语义关联，还引入了跨模态与模态内部的自监督学习机制，以同时捕捉图像和文本模态各自的局部与结构信息，从而进一步增强模型对图像与文本内容的细粒度理解能力。CLIP模型则直接利用对比学习将图像与文本分别映射到一个统一的跨模态语义空间，以确保语义上相似的图像-文本对在该空间中具有更高的相似度，这种映射方式有助于模型在下游视觉语言任务中获得更高效且通用的表现能力。

实验中，我们将预训练的ALBEF、TCL和CLIP模型作为白盒模型，将用对应下游任务数据集微调后的ALBEF、TCL和CLIP模型作为黑盒模型，研究者无法访问微调后模型的参数。三种模型都可用于处理TR和IR任务，ALBEF和TCL模型还能用于处理VG和VE任务。这是因为ALBEF模型通过图像-文本匹配结合跨模态对齐机制，实现了对图像和文本的深层次语义关联理解，从而可适用于多种视觉任务场景。TCL模型则通过跨模态特征交互和内模态自监督进一步增强了图像和文本关系的理解，同样可适用于多种视觉语言任务。与ALBEF和TCL模型相比，CLIP模型将图像和文本嵌入到统一的向量空间以处理多模态任务的机制，缺乏更深入的多模态信息融合能力，因此主要适用于较为简单的TR和IR任务，不适用于VG和VE任务。具体情况如表1所示。

表1　论文中评估的所有数据集和任务的说明

Datasets	Task	Attack Model		
		ALBEF	TCL	CLIP
Flickr30k	IR/TR	√	√	√
MSCOCO	IR/TR	√	√	√
RefCOCO+	VG	√	√	--
SNLI-VE	VE	√	√	--

5.3　基线方法

实验将CoAtt-attack和以下三种最先进的视觉语言预训练模型对抗攻击方法进行对比。

Co-attack[21]是一种多模态白盒对抗攻击方法，该方法通过同时对图像和文本施加扰动来实现对抗攻击。具体而言，Co-attack首先利用文本模态中的词替换攻击确定对图像模态扰动的攻击方向，然后基于此方向对图像执行迭代式多步扰动优化，实现对视觉语言模型的攻击。

VLAttack[22]是一种多模态黑盒对抗攻击方法，该方法通过融合来自单模态和多模态层面的图像和文本扰动来生成对抗样本。在单模态层面，使用块级相似性攻击（Block-wise Similarity Attack，BSA）策略学习图像扰动，并使用BERT-Attack策略生成与图像模态攻击独立的文本扰动。在多模态层面，使用迭代交叉搜索攻击（ICSA，Iterative Cross-Search Attack）方法迭代更新对抗性图像-文本配对。

SSAP[18]是一种单模态白盒对抗攻击方法，通过对图像的单一模态进行扰动，从而欺骗视觉语言模型。该方法利用PGD算法，通过最小化交叉熵损失，逐步生成能够最大化模型预测错误的扰动。这些扰动被添加到原始图像中，导致输入图像的特征发生微小变化，进而引起模型误判。

5.4　参数设置

为将CoAtt-attack与其他基线方法进行公平对比，实验中所有方法均使用相同的原始测试数据集，并尽可能地保持攻击参数设置的一致性。实验中使用到的PGD方法和BERT-Attack方法的参数设置均与Co-attack[21]中的参数设置保持一致。具体而言，对于图像模态的对抗攻击，CoAtt-attack、Co-attck和SSAP均采用PGD方法，并将最大扰动值设置为2/255，步长设置为1.25，迭代次数设置为10。对于文本模态的对抗性攻击，CoAtt-attack、Co-attack和VLAttack均采用BERT-Attack方法，并将最大扰动单词数设置为1，所选单词列表的长度设置为10。此外，由于SSAP方法是一种单模态攻击方法，原始设计仅对图像模态进行扰动，因此实验中未涉及SSAP方法对文本模态的扰动。VLAttack方法针对图像模态进行扰动的参数采用原文默认设置，以确保实验与原文保持一致。

5.5　评价标准

为评估对抗攻击方法的有效性，实验采用攻击成功率（Attack Success Rate，ASR）作为主要评价指标。攻击成功率衡量了模型在遭受对抗攻击后的性能损失，直观体现出攻击方法对模型鲁棒性的破坏程度[18,20-21]。然而，不同的视觉语言任务在目标与评价方式上存在明显差异，因此ASR的具体计算方式也需根据任务特性做出相应调整。

对于VG和VE任务而言，这类任务的目标通常是判断模型能否准确地识别、定位或推理图像与对应文本之间的关系，往往采用准确率（Accuracy，Acc）来评估模型性能。因此，在VG和VE任务中，

我们通过计算模型在原始样本与对抗样本上准确率的差异，来衡量ASR，具体计算方式如公式(10)所示：

$$ASR=\frac{Acc_{orignal}-Acc_{attack}}{Acc_{orignal}} \tag{10}$$

其中，$Acc_{orignal}$为输入原始样本的模型准确率，Acc_{attack}为输入对抗样本的模型准确率。

对于IR和TR任务这类基于排序的任务，准确率并不能很好地评价模型性能。这类任务的本质在于根据查询内容对候选目标进行排序，相关性更高的目标应排在更靠前的位置。因此，IR和TR任务通常使用Recall@K(R@K)作为性能指标。在实验中，我们取K=1，5，10，即分别用R@1、R@5和R@10来评估模型性能。其中，R@1反映了模型对最相关目标的精准排序能力，而R@5和R@10则体现了模型在允许一定误差范围内捕获正确目标的召回能力。因此，在排序任务中，我们基于R@K指标计算ASR，即通过模型在原始样本和对抗样本上的R@K差异，衡量攻击成功的程度，具体计算方式如公式(11)所示：

$$ASR=\frac{\mathrm{R@}K_{orignal}-\mathrm{R@}K_{attack}}{\mathrm{R@}K_{orignal}} \tag{11}$$

其中，R@$K_{orignal}$表示模型在原始样本上的召回率，R@K_{attack}为模型在对抗样本上的召回率。

公式(10)和(11)中对Acc或R@K的差值进行了归一化处理，即除以原始性能指标值，这样归一化目的在于消除模型基线性能差异的影响，使得对抗攻击效果可相对于模型的原始性能进行客观衡量。通过这种方式，模型性能下降被转化为相对值，独立于具体模型的初始准确率或召回率，从而实现了不同模型与攻击方法之间的公平比较。

为全面评估图像模态测试数据的真实性，我们同时使用学习感知图像块相似度(Learned Perceptual Image Patch Similarity，LPIPS)、峰值信噪比(Peak Signal-to-Noise Ratio，PSNR)和结构相似性指数(Structural Similarity Index Measure，SSIM)三种图像质量评价指标。LPIPS评估图像在深层语义特征空间中的相似度，PSNR衡量图像整体像素误差，而SSIM则关注图像的局部结构变化。这三种评价指标从语义、像素、结构三个层面对图像真实性进行全面评估。

LPIPS通过预训练的深度卷积神经网络提取图像的深层特征，再基于人类感知判断对特征之间的差异进行加权计算[44]。具体计算方法如公式(12)：

$$LPIPS(x,x_0)=\sum_l\frac{1}{H_lW_l}\sum_{h,w}\left\|w_l\odot\left(\hat{y}^l_{hw}-\hat{y}^l_{0hw}\right)\right\|_2 \tag{12}$$

其中，x为经过对抗攻击后的图像，x_0表示未经攻击的原始图像。$\hat{y}^l_{hw}$与$\hat{y}^l_{0hw}$分别表示经过深度卷积网络提取后第l层特征图中位置为(h,w)处的特征向量，w_l表示通过人类感知训练得到的特征权重，H_l和W_l分别表示第l层特征图的高度和宽度。与PSNR和SSIM相比，LPIPS能够更贴近人类主观感知判断。LPIPS值位于[0-1]，值越低表示图像之间的感知差异越小，即越相似。

PSNR是一种基于像素误差的传统图像质量评价指标，通过计算图像间的均方误差(Mean Squared Error，MSE)，再转化为对数信噪比，以dB(分贝)为单位衡量图像的保真度[45]。具体计算方法如公式(13)：

$$PSNR(x,x_0)=10\cdot\log_{10}\left(\frac{MAX^2}{MSE(x,x_0)}\right) \tag{13}$$

其中，x表示对抗攻击后的图像，x_0表示原始图像。MAX为图像中像素的最大取值(对于8-bit图像，通常为255)，MSE表示两幅图像对应像素的均分误差，PSNR值越高，代表图像之间的数值差异越小，图像保真度越高。PSNR已在图像压缩、恢复及对抗扰动等任务中被广泛用于衡量图像重建或扰动后的质量与原始图像之间的像素差异[46-47]。然而，PSNR并未考虑人类视觉系统的感知特性，难以全面反映图像在结构和语义层面的变化。

SSIM则是一种结构感知型图像质量评价指标，用于模拟人类视觉系统对图像质量的主观感知[48]。SSIM综合考虑图像的亮度、对比度和结构信息，以评估两幅图像之间的感知相似性，其计算方法如公式(14)：

$$\mathrm{SSIM}(x,x_0)=\frac{(2\mu_x\mu_{x_0}+C_1)(2\sigma_{xx_0}+C_2)}{(\mu_x^2+\mu_{x_0}^2+C_1)(\sigma_x^2+\sigma_{x_0}^2+C_2)} \tag{14}$$

其中，x表示对抗攻击后的图像，x_0表示原始图像，μ_x、μ_{x_0}分别表示图像亮度的均值，σ_x、σ_{x_0}为亮度的标准差，σ_{xx_0}表示图像间的协方差，C_1、C_2为防止分母为0的稳定常数。SSIM取值范围在[0，1]，值越接近1，表示两幅图像在结构和感知上的相似度越高。与PSNR不同，SSIM更加注重图像的结构一致性，因此常被用于评估图像在结构一致性与感知质量方

面的变化。

实验采用BERTScore（Bidirectional Encoder Representations from Transformers Score）作为文本模态测试数据的语义质量的评价指标。文本对抗样本通常仅在词汇层面发生微小改动，在尽可能保持语义不变的前提下改变模型预测结果。因此，基于n-gram重叠的指标（如BLEU、ROUGE）难以准确评估对抗样本的语义偏移。而BERTScore基于预训练的深度语言模型（如BERT）提取文本的上下文语义表示，通过计算候选文本与参考文本在词向量空间中的余弦相似度，衡量二者在语义层面的相似程度[49]。BERTScore通过上下文敏感的深层语义建模，能够更好地反映对抗样本与原始文本在语义上的一致性，适合作为本实验中评估文本对抗样本语义质量的评价指标。

设原始文本为t_0，文本对抗样本为t，经过预训练语言模型后，分别得到其词向量表示序列$\{h_j^t\}$和$\{h_i^{t_0}\}$。BERTScore首先计算两组向量h_j^t和$h_i^{t_0}$间的余弦相似度，计算方法如公式(15)所示：

$$s_{ij}=\cos\left(h_j^t,h_i^{t_0}\right) \tag{15}$$

然后，对于对抗文本序列$\{h_j^t\}$中的每个词向量h_j^t，计算其与原始文本序列$\{h_i^{t_0}\}$中所有词向量的余弦相似度，并选取其中的最大值作为该词的最佳匹配得分。将所有对抗文本词向量的匹配得分取平均，即可得到Precision。同理，对于原始文本序列$\{h_i^{t_0}\}$中的每个词向量$h_i^{t_0}$，计算其与对抗文本中所有词向量的相似度，并取最大值作为该词的最佳匹配得分，所有得分取平均即为Recall。Precision与Recall通过加权调和平均计算得到$F1$分数，计算过程如公式(16)到公式(18)：

$$P=\frac{1}{|t_0|}\sum_j\max_i s_{ij} \tag{16}$$

$$R=\frac{1}{|t|}\sum_i\max_j s_{ij} \tag{17}$$

$$F1=\frac{2PR}{P+R} \tag{18}$$

其中，$F1$分数作为最终的BERTScore值，取值范围在$[0,1]$之间，值越高表示两段文本在语义上越接近。

6　实验结果分析

6.1　RQ1结果分析

为研究CoAtt-attack与其他视觉语言预训练模型对抗性攻击方法的性能和其所生成图像测试数据的真实性，实验选取ABLEF、TCL和CLIP三种视觉语言预训练模型作为攻击对象，分别针对IR、TR、VG和VE四种下游任务进行评估，实验结果如表2到9所示，其中表2至表5展示了各方法针对IR、TR、VG和VE四个下游任务的ASR对比结果，最高的ASR被加粗标出。表6到表8分别展示了不同方法生成图像测试数据的LPIPS、PSNR和SSIM平均值结果。其中，最低的LPIPS被加粗标出，最高的PSNR、SSIM被加粗标出。表9展示了CoAtt-attack、Co-attack和VLAttack生成文本测试数据的平均BERTScore值，最高的BERTScore值被加粗标出。

表2　图像到文本检索任务中的攻击成功率对比（TR）

Attack Model	Attack Method	Flickr30K			MSCOCO		
		R@1	R@5	R@10	R@1	R@5	R@10
ALBEF	CoAtt-attack(Our)	**90.94**	**86.33**	**84.08**	**94.08**	**92.25**	**91.13**
	Co-attack	72.18	60.00	51.25	79.08	67.53	61.62
	VLAttack	85.88	79.77	76.54	85.68	83.06	80.79
	SSAP	66.17	56.78	49.81	58.63	58.47	53.97
TCL	CoAtt-attack(Our)	**97.12**	**94.16**	**92.59**	**96.98**	**96.09**	**95.35**
	Co-attack	74.47	61.37	56.01	79.10	70.97	64.87
	VLAttack	83.12	78.75	66.39	85.74	83.46	76.25
	SSAP	65.06	50.6	47.66	68.44	62.25	56.01
CLIP	CoAtt-attack(Our)	**99.88**	**99.69**	**99.59**	**99.96**	**99.92**	**99.91**
	Co-attack	90.43	82.87	77.03	96.95	94.43	91.83
	VLAttack	92.45	85.94	83.28	97.92	96.68	92.42
	SSAP	81.36	75.23	67.5	80.45	79.79	76.18

表3 文本到图像检索任务中的攻击成功率对比(IR)

Attack Model	Attack Method	Flickr30K			MSCOCO		
		R@1	R@5	R@10	R@1	R@5	R@10
ALBEF	CoAtt-attack(Our)	**86.64**	**82.11**	**79.95**	**92.54**	**90.65**	**89.85**
	Co-attack	78.93	71.45	67.43	84.16	77.81	74.17
	VLAttack	82.92	74.73	70.25	81.47	74.27	70.49
	SSAP	63.81	58.81	55.73	60.83	63.01	56.43
TCL	CoAtt-attack(Our)	**94.49**	**92.37**	**91.60**	**96.93**	**96.13**	**95.70**
	Co-attack	83.11	74.01	68.66	86.42	81.34	77.12
	VLAttack	87.75	82.84	79.99	88.03	80.20	76.81
	SSAP	76.41	61.64	57.99	62.52	58.95	55.86
CLIP	CoAtt-attack(Our)	**99.84**	**99.42**	**98.91**	**99.94**	**99.83**	**99.71**
	Co-attack	94.10	88.15	84.57	96.61	94.72	93.07
	VLAttack	94.99	89.98	86.39	95.46	95.68	94.81
	SSAP	81.09	77.64	67.83	72.33	71.68	67.76

表4 视觉定位任务中的攻击成功率对比(VG)

Attack Model	Attack Method	RefCOCO+
ALBEF	CoAtt-attack(Our)	79.33
	Co-attack	38.04
	VLAttack	**80.39**
	SSAP	25.75
TCL	CoAtt-attack(Our)	**72.18**
	Co-attack	50.73
	VLAttack	69.06
	SSAP	23.15

表5 视觉蕴含任务中的攻击成功率对比(VE)

Attack Model	Attack Method	SNLI-VE
ALBEF	CoAtt-attack(Our)	**71.08**
	Co-attack	67.32
	VLAttack	58.04
	SSAP	49.87
TCL	CoAtt-attack(Our)	**72.63**
	Co-attack	66.23
	VLAttack	60.34
	SSAP	49.37

表6 不同测试方法生成测试数据的平均LPIPS值对比

Dataset	Model	CoAtt-attack	Co-attack	VLAttack	SSAP
Flickr30K	ALBEF	**0.0100**	0.0164	0.4597	0.3124
	TCL	**0.0104**	0.0183	0.4512	0.3247
	CLIP	**0.0128**	0.0128	0.6071	0.4213
MSCOCO	ALBEF	**0.0139**	0.0229	0.5712	0.4454
	TCL	**0.0116**	0.0242	0.5733	0.4512
	CLIP	**0.0176**	0.0189	0.6485	0.5133
RefCOCO+	ALBEF	**0.0092**	0.0156	0.4405	0.3054
	TCL	**0.0134**	0.0189	0.4498	0.3107
SNLI-VE	ALBEF	**0.0099**	0.0176	0.4644	0.3347
	TCL	**0.0124**	0.0209	0.4781	0.3511
Average		**0.012 12**	0.018 65	0.514 38	0.377 02

针对TR、IR任务,实验使用了Flickr30K和MSCOCO数据集,结果如表2和表3所示。CoAtt-attack对ALBEF、TCL、CLIP三个模型测试的攻击成功率都显著优于所有的基线工具。具体而言,对于TR任务,与三种基线方法相比,CoAtt-attack测试ALBEF模型的ASR(R@1)提高了5.06%~35.45%、ASR(R@5)提高了6.56%~33.78%、ASR(R@10)提高了7.54%~37.16%,测试TCL模型的ASR(R@1)提高了11.24%~32.06%、ASR(R@5)提高了12.63%~43.56%、ASR(R@10)提高了19.1%~44.93%,测试CLIP模型的ASR(R@1)提高了2.04%~19.51%、ASR(R@5)提高了3.24%~24.46%、ASR(R@10)提高了7.49%~32.09%。对于IR任务,与三种基线方法相比,CoAtt-attack测试ALBEF模型的ASR(R@1)提高了3.72%~31.71%、ASR(R@5)提高了7.38%~27.64%、ASR(R@10)提高了9.7%~33.42%,测试TCL模型的ASR(R@1)提高了6.74%~34.41%、ASR(R@5)提高了9.53%~37.18%、ASR(R@10)提高了11.61%~39.84%,测试CLIP模型的ASR(R@1)提高了3.33%~27.61%、ASR(R@5)提高了4.15%~28.15%、ASR(R@10)提高了4.9%~31.95%。

针对VG任务,实验使用RefCOCO+数据集对ALBEF和TCL模型进行了测试。实验结果如表4所示,CoAtt-attack测试ALBEF模型的ASR达到了79.33%,虽然与VLAttack相比低1.06%,但是与Co-attack和SSAP相比分别提高了41.29%和53.58%。同时,CoAtt-attack对TCL模型测试的ASR达到了最高的72.18%,与Co-attack、VLArrack和SSAP相比分别提高了21.45%、

表7　不同测试方法生成图像测试数据的平均PSNR值对比

Dataset	Model	CoAtt-attack	Co-attack	VLAttack	SSAP
Flickr30K	ALBEF	**51.75 dB**	45.51 dB	25.16 dB	32.2 dB
	TCL	**49.31 dB**	45.36 dB	25.00 dB	31.15 dB
	CLIP	**45.51 dB**	43.05 dB	24.54 dB	29.45 dB
MSCOCO	ALBEF	**50.69 dB**	45.52 dB	26.35 dB	30.11 dB
	TCL	**49.45 dB**	45.43 dB	25.64 dB	29.61 dB
	CLIP	**45.64 dB**	42.67 dB	23.09 dB	27.05 dB
RefCOCO+	ALBEF	**48.45 dB**	45.56 dB	27.44 dB	32.82 dB
	TCL	**45.60 dB**	43.39 dB	25.98 dB	31.77 dB
SNLI-VE	ALBEF	**50.47 dB**	45.29 dB	26.22 dB	30.85 dB
	TCL	**48.11 dB**	45.08 dB	23.19 dB	30.32 dB
Average		**48.50 dB**	44.69 dB	25.26 dB	30.53 dB

表8　不同测试方法生成图像测试数据的平均SSIM值对比

Dataset	Model	CoAtt-attack	Co-attack	VLAttack	SSAP
Flickr30K	ALBEF	**0.9957**	0.9925	0.7254	0.8011
	TCL	**0.9961**	0.9922	0.7320	0.7897
	CLIP	**0.9937**	0.9918	0.5882	0.6905
MSCO-CO	ALBEF	**0.9971**	0.9921	0.6845	0.7670
	TCL	**0.9962**	0.9919	0.6721	0.7529
	CLIP	**0.9934**	0.9907	0.5780	0.6487
RefCO-CO+	ALBEF	**0.9958**	0.9927	0.7350	0.8059
	TCL	**0.9884**	0.9835	0.7164	0.7792
SNLI-VE	ALBEF	**0.9969**	0.9920	0.7001	0.7683
	TCL	**0.9947**	0.9916	0.6883	0.7498
Average		**0.9948**	0.9911	0.6820	0.755 31

表9　不同测试方法生成文本对抗样本的平均BERTScore值对比

Dataset	Model	CoAtt-attack	Co-attack	VLAttack	SSAP
Flickr30K	ALBEF	**0.8883**	0.8883	0.8549	--
	TCL	**0.8846**	0.8745	0.8611	--
	CLIP	**0.8909**	0.8812	0.8578	--
MSCOCO	ALBEF	**0.8787**	0.8765	0.8647	--
	TCL	0.8783	**0.8888**	0.8771	--
	CLIP	0.8810	0.8657	**0.8896**	--
RefCOCO+	ALBEF	0.7934	**0.7994**	0.7745	--
	TCL	0.7645	**0.7745**	0.7658	--
SNLI-VE	ALBEF	**0.8252**	0.8102	0.8114	--
	TCL	0.8224	0.8059	**0.8347**	--
Average		**0.8507**	0.8465	0.8392	--

3.12%、49.03%。

针对VE任务，实验使用SNLI-VE数据集对ALBEF和TCL模型进行了测试，结果如表5所示，与Co-attack、VLAttack和SSAP相比，CoAtt-attack测试ALBEF模型的ASR分别提高了3.76%、13.04%和21.21%，测试TCL模型的ASR分别提高了6.4%、12.29%和23.26%。

上述实验结果可以看出，CoAtt-attack在四种视觉语言下游任务中均取得了比三种基线工具更高的攻击成功率，这是由于CoAtt-attack使用Co-Attention机制关注图像特征和文本特征之间的相互影响，并根据它们之间的相互影响情况对图像测试数据添加细微的针对性扰动。同时，Co-Attention机制还可扩大被扰动后的图像模态和文本模态之间的差异，进一步增强攻击效果。对于IR、TR、VG和VE等下游任务，消除图像模态与文本模态之间的差异通常是主要目标，而Co-Attention机制通过扩大这种差异，使其添加的细微扰动依然具备能让模型致错的能力。SSAP是四种方法中性能最差的对抗性方法，这是因为该方法仅针对图像模态进行扰动，而对于多模态模型，仅添加单模态的扰动不能有效地使模型致错。

图像模态测试数据的真实性也是实验关注的指标，接下来对不同方法生成图像测试数据的真实性进行分析。为量化CoAtt-attack、Co-attack、VLAttack和SSAP四种对抗性攻击方法所生成的图像模态测试数据的真实性，实验计算了四种方法使用不同数据集为不同模型生成的图像测试数据的平均LPIPS值、PSNR值和SSIM值，实验结果如表6到表8所示，其中对同一个模型和数据集生成图像测试数据中最低的LPIPS值，以及最高的PSNR和SSIM值均被加粗标出。

如表6到表8所示，CoAtt-attack为所有模型生成的图像测试数据均取得了最低的平均LPIPS值、最高的平均PSNR值和SSIM值。与Co-attack、VLAttack和SSAP相比，CoAtt-attack生成图像测试数据的平均LPIPS值分别降低了0.0065、0.5023和0.3649；平均PSNR值分别提高了3.81 dB、23.24 dB和17.97 dB；平均SSIM值分别提高了0.0037、0.3128和0.2395。这是由于CoAtt-attack只对图像中与文本描述最相关的区域进行扰动，而非添加明显的全局扰动。如图5所示，CoAtt-attack和这三种方法中LPIPS值最低的Co-attack进行对比，可以观察到CoAtt-attack主要对与文本中的关键词相关的区域添加噪声，而Co-attack则会对整个图像添加噪声。例如图5中第一幅图像，CoAtt-attack只对与文本中“dog”，“grassy”，“white fence”相关的图像区域添加噪声。

图5 CoAtt-attack和Co-attack添加的图像扰动示例

对于多模态测试数据，文本模态的质量也需要关注。接下来对不同方法生成文本测试数据的语义质量进行分析。为量化CoAtt-attack、Co-attack和VLAttack三种对抗攻击方法（SSAP只针对图像，不涉及文本扰动）所生成的文本模态测试数据的质量，实验计算了三种方法使用不同数据集为不同模型生成的文本测试数据的平均BERTScore值，实验结果如表9所示。对于四个数据集，CoAtt-attack为所有模型生成的文本测试数据均取得了最高的平均BERTScore值。然而，不同方法间的BERTScore差异整体较小，与Co-attack和VLAttack相比，CoAtt-attack生成文本测试数据的平均BERTScore值仅提高了0.0042和0.0115，表明各方法在文本语义质量方面具有相近表现。这是由于三个方法均集成使用了BERT-Attack方法去生成文本对抗样本，且在实验中统一了替换词数与语义相似性阈值，使得最终生成的文本对抗样本在语义空间中具有较强的一致性。

对RQ1的结论：通过在不同下游任务中的实验验证，CoAtt-attack方法的攻击效果和所生成对抗图像的真实性与现有基线方法（Co-attack、VLAttack和SSAP）相比均表现出显著优势。CoAtt-attack在提升攻击成功率的同时，有效地保持了生成的图像对抗样本的真实性和文本对抗样本的语义质量，实现了攻击效果与对抗样本质量之间的良好平衡。

6.2 RQ2结果分析

为研究Co-Attention机制的引入对CoAtt-attack攻击性能的影响，实验在CoAtt-attack原型工具的基础上去除了利用Co-Attention机制进行注意力分析和扰动掩膜的步骤，直接使用PGD和BERT-Attack对测试数据添加全局扰动，称为CoAtt-attack$^-$。实验中CoAtt-attack和CoAtt-attack$^-$使用Flickr30k和MSCOCO数据集分别对ALBEF、TCL两个视觉语言模型就IR任务和TR任务进行对抗测试。

实验结果如图6所示，CoAtt-attack相比于CoAtt-attack$^-$的ASR（R@1）、ASR（R@5）和ASR（R@10）都表现出更高的性能，表明Co-Attention机制可以提高对抗性攻击效果。具体而言，在图6（a）和图6（c）中，使用Flickr30K数据集对ALBEF模型就TR和IR任务进行测试时，CoAtt-attack的ASR（R@1）、ASR（R@5）和ASR（R@10）均高于CoAtt-attack$^-$。这表明Co-Attention机制能有效地帮助模型更好地理解文本与图像之间的对齐关系，提升了攻击成功率。在图6（b）和图6（d）中，针对MSCOCO数据集的实验结果同样观察到类似的性能提升趋势。这表明Co-Attention机制不仅适用于Flickr30K数据集，也在更大规模的MSCOCO数据集上表现出了更好的攻击性能，进一步表明了其有

效性。对TCL模型测试的实验(图6(e)到图6(h))结果同样呈现出相同的趋势。这表明Co-Attention机制不仅在ALBEF模型中表现突出，在TCL模型中同样能够显著提升攻击效果。

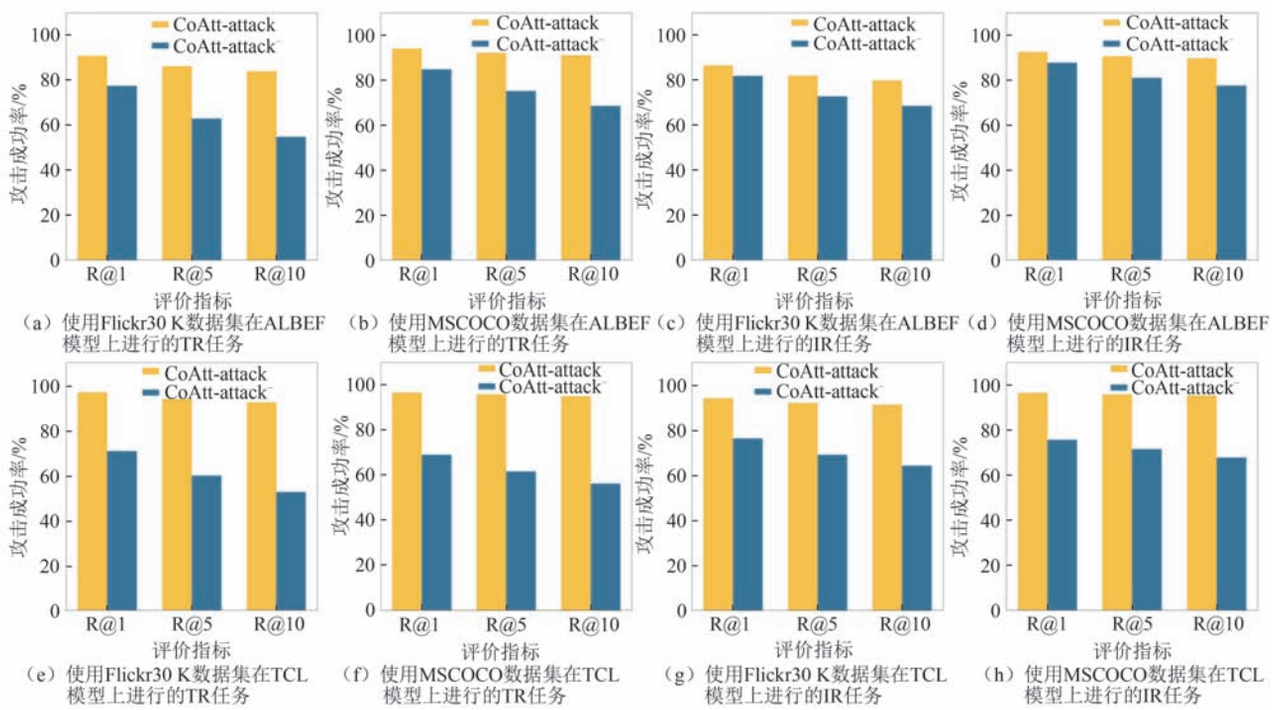

图6　Co-Attention机制对CoAtt-attack攻击性能的影响

为进一步验证Co-Attention机制在图像扰动引导中的作用，我们分别使用CoAtt-attack和CoAtt-attack⁻生成的图像对抗样本与原始图像进行像素级残差计算，并进行可视化展示，如图7所示。图中每一行为一组图像示例，依次为：原始图像测试数据、使用CoAtt-attack生成的图像对抗样本、使用CoAtt-attack⁻生成的图像对抗样本、有Co-Attention时的残差图像和无Co-Attention时的残差图像。残差图像采用热力图形式呈现，其中红色区域表示扰动强度较大，图像在该位置发生了较大的变化；蓝色区域则表示扰动强度较低，该区域在扰动过程中几乎未被修改。

从图7中可以看出，引入Co-Attention机制后的扰动分布更加集中，大多数扰动聚焦于与模型注意力高度重合的关键区域，背景等非关键信息区域受到扰动极小。而在未引入Co-Attention机制的条件下，扰动呈现出更为随机且离散的分布，明显扩散至整幅图像的多个区域。例如，处理描述为“a black and white dog is running in a grassy garden surrounded by a white fence”的图像时，模型的注意力主要集中在图像中“dog”、“grassy garden”以及“white fence”等与文本内容高度相关的区域。在引入Co-Attention的残差图中，红色区域高度重叠于“dog”、“grassy garden”以及“white fence”所在位置，其他背景区域几乎没有明显扰动；而在未使用Co-Attention机制时，扰动在图像多个位置呈弥散态分布，扰动区域显著扩大。这种差异表明，Co-Attention机制有助于将扰动更精准地集中于模型决策所依赖的关键区域，同时抑制扰动扩散至无关区域，降低了由冗余扰动带来的计算成本和干扰，使得生成的对抗样本更具针对性和实用价值，进一步提升了整体攻击性能。

对RQ2的结论：引入Co-Attention机制后，CoAtt-attack方法的对抗攻击性能显著优于未引入该机制的消融方法CoAtt-attack⁻，表明使用Co-Attention机制作为解释方法显著提高了CoAtt-attack对抗攻击的性能。

6.3　RQ3结果分析

为分析不同模态扰动对VLPMs鲁棒性的影响，本文在图像检索(IR)与文本检索(TR)任务中，分别对单模态扰动和多模态联合扰动的效果进行了对比。通过对比不同扰动方式在IR与TR任务中的攻击效果，以分析单模态与多模态联合扰动对视觉语言模型鲁棒性的影响。实验选取主流的视觉语

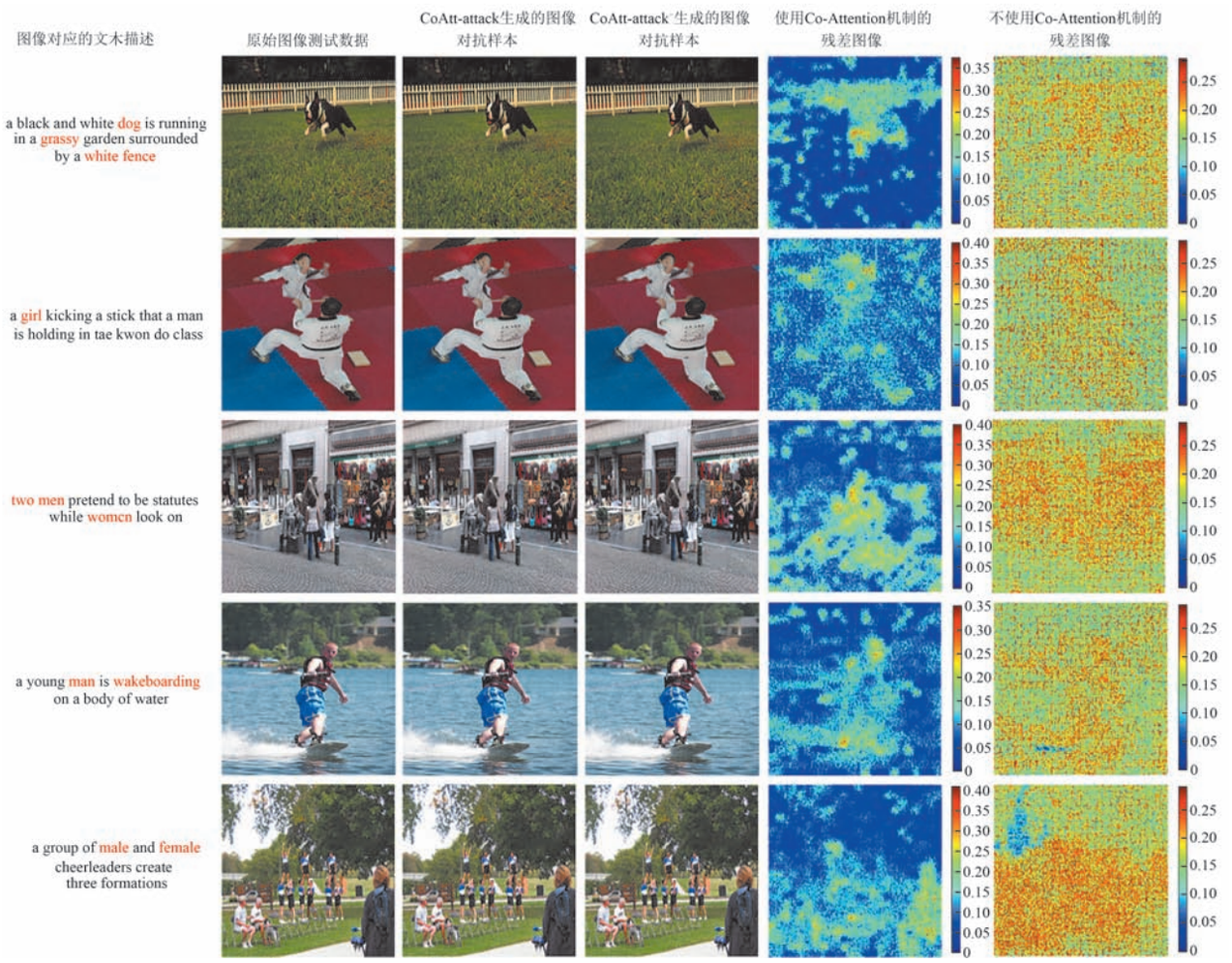

图7 有无Co-Attention机制生成的对抗图像与原始图像的残差图像可视化示例

言预训练模型ALBEF与TCL作为测试对象，使用MSCOCO与Flickr30K数据集进行测试。具体而言，使用CoAtt-attack与CoAtt-attack(Image)和CoAtt-attack(Text)进行对比。其中，CoAtt-attack(Image)去除了文本模态的扰动，仅对图像模态施加基于PGD算法的扰动；CoAtt-attack(Text)去除了图像模态的扰动，仅对文本模态施加基于BERT-Attack的词替换扰动。

实验结果如表10和表11所示。对于TR任务，CoAtt-attack使用Flickr30K和MSCOCO数据集测试ALBEF模型的ASR(R@1)分别达到90.94%和94.08%、ASR(R@5)达到86.33%和92.25%、ASR(R@10)达到84.08%和91.13%，测试TCL模型的ASR(R@1)分别达到97.12%和96.98%、ASR(R@5)达到94.16%和96.09%、ASR(R@10)达到92.59%和95.35%。CoAtt-attack(Image)使用Flickr30K和MSCOCO数据集测试ALBEF和TCL模型的整体攻击成功率较联合扰动下降了15.63%～46.83%；CoAtt-attack(Text)测试ALBEF和TCL模型的攻击成功率下降更为明显，整体攻击成功率较联合扰动下降了69.77%～91.89%。对于IR任务，CoAtt-attack使用Flickr30K和MSCOCO数据集测试ALBEF模型的ASR(R@1)分别达到86.64%和92.54%、ASR(R@5)达到82.11%和90.65%、ASR(R@10)达到79.95%和89.85%，测试TCL模型的ASR(R@1)分别达到94.49%和96.93%、ASR(R@5)达到92.37%和96.13%、ASR(R@10)达到91.13%和95.7%。CoAtt-attack(Image)测试这两种模型的整体攻击成功率相比联合扰动下降了9.88%～36.58%；CoAtt-attack(Text)测试ALBEF和TCL模型的整体攻击成功率较联合扰动下降了69.77%～91.89%。

从上述实验结果可以看出，三种扰动方式对VLPMs的鲁棒性影响存在显著差异。其中CoAtt-attack取得了最高的攻击成功率，对VLPMs的鲁棒性影响最大。这是由于同时对图像和文本施加扰动更容易破坏VLPMs的跨模态对齐机制，导致模型在进

表10　不同扰动方式在TR任务上的攻击成功率对比

Attack Model	Attack Method	Flickr30K			MSCOCO		
		R@1	R@5	R@10	R@1	R@5	R@10
ALBEF	CoAtt-attack	**90.94**	**86.33**	**84.08**	**94.08**	**92.25**	**91.13**
	CoAtt-attack(Image)	74.60	60.70	53.85	78.45	67.97	61.45
	CoAtt-attack(Text)	7.38	1.31	0.60	22.20	9.93	5.97
TCL	CoAtt-attack	**97.12**	**94.16**	**92.59**	**96.98**	**96.09**	**95.35**
	CoAtt-attack(Image)	61.11	53.22	48.10	55.29	51.99	48.52
	CoAtt-attack(Text)	13.14	2.62	0.70	27.21	13.61	8.28

表11　不同扰动方式在IR任务上的攻击成功率对比

Attack Model	Attack Method	Flickr30K			MSCOCO		
		R@1	R@5	R@10	R@1	R@5	R@10
ALBEF	CoAtt-attack	**86.64**	**82.11**	**79.95**	**92.54**	**90.65**	**89.85**
	CoAtt-attack(Image)	76.19	67.87	64.18	82.66	75.67	71.60
	CoAtt-attack(Text)	23.50	16.60	15.62	33.40	26.98	24.97
TCL	CoAtt-attack	**94.49**	**92.37**	**91.60**	**96.93**	**96.13**	**95.70**
	CoAtt-attack(Image)	68.83	62.82	58.34	65.43	62.15	59.12
	CoAtt-attack(Text)	26.61	13.99	10.00	37.25	24.48	18.88

行语义匹配与决策推理的过程中无法有效整合图文信息，显著削弱了其判别能力。CoAtt-attack(Image)的攻击成功率虽低于联合扰动，但仍具较强的攻击能力，表明VLPMs在推理过程中对图像模态扰动的敏感性更高。仅施加了文本扰动的CoAtt-attack(Text)的攻击成功率最低，对VLPMs的鲁棒性影响最小。这可能与图像和文本本身的差异有关。图像属于连续的像素空间，能够通过梯度优化实现更精细的扰动控制，从而更容易生成具备攻击性的样本。而文本输入由离散的词元构成，其扰动通常依赖于词级替换，无法直接使用梯度进行优化，不仅修改方式受限，还需要在保持语义和语法正确性的前提下进行操作。这些限制导致文本扰动在干扰模型预测时的效果相对较弱，特别是在输入文本较短的场景下，其扰动对模型的影响更容易被图像模态所补偿。

对RQ3的结论：不同模态的扰动对VLPMs的鲁棒性影响存在显著差异。图像模态的扰动相比文本模态具有更强的攻击能力，对VLPMs的鲁棒性影响更强。而图文联合扰动相对单模态的扰动方式则能进一步放大干扰效果，测试多个VLPMs模型均取得最高的攻击成功率，对VLPMs的鲁棒性影响最大。

7　讨　论

7.1　与注意力区域引导攻击方法的对比分析

本节将对比CoAtt-attack与现有扰动注意力区域攻击方法的区别。已有研究尝试将注意力机制引入多模态对抗攻击框架，利用模型的注意力分布信息引导扰动区域的生成，以提升攻击的效果。具有代表性的工作包括Disabato等人[50]提出的基于注意力区域扰动的迁移攻击方法、Wang等人[51]提出的TMM(Transferable Multimodal Attack)框架、Guan等人[52]提出的JMTFA(Joint Multimodal Transformer Feature Attack)方法和Li等人[53]提出的AIC-Attack(Attention-based Image Captioning Attack)方法。以下将分别对比CoAtt-attack与上述四种方法的差异，并结合实验结果进行分析。

与现有的扰动注意力区域方法相比，CoAtt-attack在攻击建模思路、注意力机制作用方式及模态协同策略等方面有差异。Disabato等人提出的方法基于图像-图像特征匹配构建攻击目标，通过将目标文本合成为目标图像，再最大化当前图像与目标图像在特征空间中的相似性以引导扰动，并借助注意力掩码提升迁移攻击效果。该方法的注意力机制主要用于图像迁移方向的掩码引导，未用于解释模型原生判别机制，且仅作用于图像模态，缺乏图文协同设计。TMM方法则从跨模型迁移稳定性出发，构建模态一致性与差异性联合约束，通过特征正交化与跨模态注意力提升对抗样本的迁移能力。而CoAtt-attack则聚焦于当前白盒模型的鲁棒性测试，直接建模图文语义对齐关系，生成解释性更强的注意力分布用于后续的扰动，扰动的目标更加聚焦于模型进行决策时所关注的关键区域。JMTFA基于模型前

向传播和反向传播过程中的注意力权重与梯度信息，计算每个token（图像块或文本词元）在模型决策过程中的贡献度，并据此形成扰动区域的优先级排序，最后通过排序控制扰动顺序与区域，该方法依赖梯度信息，且缺乏明确的跨模态语义建模机制。AIC-Attack方法则基于图像字幕模型的注意力分布，通过对像素进行排序选取前k%个高关注区域，再结合差分演化优化像素扰动，提升攻击效率。该方法针对图像字幕（Image Captioning）任务，基于注意力得分排序选取图像的扰动区域，但缺乏多模态协同扰动机制，仅扰动图像模态。这种基于排序选取扰动区域的策略难以适应不同样本中注意力分布的差异，可能导致对语义关键区域覆盖不足，从而影响扰动效果的针对性。相比之下，CoAtt-attack在前向传播阶段引入Co-Attention机制，以识别模型在决策过程中最为关注的图像区域与文本关键词，并据此引导扰动生成，无需通过排序来控制扰动区域的选择，避免了排序阈值设置所带来的不稳定性。

为进一步分析CoAtt-attack与现有注意力区域引导攻击方法在攻击效果上的差异分析，我们选取已公开实现的AIC-Attack方法与CoAtt-attack进行对比实验。由于AIC-Attack是面向图像字幕任务的图像单模态扰动方法，其攻击策略依赖模型对图像内容的理解并生成相应文本描述，因此，我们仅选取了与图像字幕任务在输入模态与语义建模机制上高度相似的TR任务进行评估。与图像字幕任务相似，TR任务同样以图像为输入，要求模型根据图像内容从候选文本中检索最相关的语义描述，本质上是将视觉模态转换为文本模态。

在对比实验中，CoAtt-attack仅启用了图像模态扰动策略，去除文本模态干扰，以确保实验对比的公平性。AIC-Attack基于其开源代码①，并使用5.4节的统一参数设置。实验选取ABLEF和TCL两种视觉语言预训练模型作为攻击对象，并针对TR任务使用Flickr30K和MSCOCO数据集进行评估。实验中使用的性能评价指标与第5.5节的评价指标一致，主要对比两种方法在固定扰动幅度约束下的ASR。

实验结果如表12所示，其中CoAtt-attack（Image）表示去除了文本模态的扰动，仅对图像模态施加扰动。与AIC-Attack相比，CoAtt-attack（Image）使用Flickr30K和MSCOCO数据集测试ALBEF模型的ASR（R@1）分别提高了7.89%和9.67%、ASR（R@5）提高了4.03%和5.96%、ASR（R@10）提高了3.84%和1.69%，测试TCL模型的ASR（R@1）分别提高了4.33%和0.92%、ASR（R@5）提高了5.31%和6.63%、ASR（R@10）提高了6.41%和6.88%。

上述实验结果表明，与同样基于注意力区域进行扰动的AIC-Attack方法相比，CoAtt-attack（Image）测试视觉语言预训练模型表现出更高的攻击成功率。这是因为AIC-Attack依赖于静态的注意力得分排序，选取前k%像素作为扰动目标，会造成扰动区域过于离散，容易干扰无关背景，导致添加的噪声无法精确覆盖模型决策所依赖的关键区域。而CoAtt-attack利用Co-Attention机制深入分析视觉语言模型决策过程中关注的跨模态语义关联区域，生成具有明确语义对齐含义的注意力分布，并据此引导扰动集中施加于模型做出决策时所依赖的关键图像区域，从而有效提升了攻击的针对性。

表12 CoAtt-attack（Image）与AIC-Attack在TR任务上的攻击成功率对比

Attack Model	Attack Method	Flickr30K			MSCOCO		
		R@1	R@5	R@10	R@1	R@5	R@10
ALBEF	CoAtt-attack（Image）	**74.60**	**60.70**	**53.85**	**78.45**	**67.97**	**61.45**
	AIC-Attack	66.71	56.67	50.01	68.78	62.01	59.76
TCL	CoAtt-attack（Image）	**61.11**	**53.22**	**48.10**	**55.29**	**51.99**	**48.52**
	AIC-Attack	56.78	47.91	41.69	54.37	45.36	41.64

7.2 有效性分析

本节对CoAtt-attack的有效性进行分析。

CoAtt-attack的内部有效性主要在于实验设计与实现过程是否正确。首先，本文使用的VLPMs模型（ALBEF、TCL和CLIP）及其根据特定任务进行微调后的模型，均来源于Salesforce Research开源的多模态学习框架LAVIS（Language-and-Vision Studio）②。其次，对抗攻击中使用的PGD方法和BERT-Attack方法均使用第三方开源代码库实现③④。最后，我们基于开源项目HieCoAttenVQA的源代码[39]构建了注意力解释模块中使用的Co-Attention机制。此外，还对CoAtt-attack的代码进

① AIC-Attack https://github.com/UTSJiyaoLi/Adversarial-Image-Captioning-Attack
② LAVIS https://github.com/salesforce/LAVIS
③ PGD https://github.com/Harry24k/PGD-pytorch
④ BERT-Attackhttps://github.com/LinyangLee/BERT-Attack

行了多次检查和测试，以尽量保证代码的正确性。

CoAtt-attack的外部有效性主要在于其在不同数据集和模型上的适用性。本文在四个典型视觉语言数据集上开展实验。其中，Flickr30K与MSCOCO用于图文检索任务，RefCOCO+用于视觉定位任务，SNLI-VE用于视觉蕴含任务，覆盖了从图文匹配、区域定位到文本推理等不同类型的任务，具有较好的代表性。在测试模型方面，实验覆盖融合型（ALBEF、TCL）与对齐型（CLIP）两类主流VLPMs架构。尽管如此，CoAtt-attack的通用性仍需在更多视觉语言任务（如VQA、图文生成等）中进一步验证。

CoAtt-attack的构造有效性主要在于评价指标的有效性。实验使用ASR来评估对抗攻击的有效性，该指标常被用于评价VLPMs的鲁棒性测试工作[21-22]。此外，实验还引入LPIPS指标来评估图像对抗样本的真实性与自然性，其作为计算机视觉领域的主流评价指标，被广泛用于评价图像的相似程度[54]和生成图像与真实图像的接近程度[55]。

7.3　局限性分析

CoAtt-attack聚焦于利用对抗攻击方法对VLPM的鲁棒性进行测试，旨在评估模型在面对跨模态扰动输入时的稳定性与可靠性。目前工作尚未涉及如何通过对抗训练机制提升VLPMs自身的鲁棒性能力，这也是CoAtt-attack的主要局限。

对抗训练是一种重要的鲁棒性提升机制，其核心思想是在训练过程中动态生成对抗样本并纳入训练流程，以增强模型在面对扰动输入时的稳健性表现[19]。在实际应用中，对抗训练通常需要大量计算和资源开销，例如，在图像分类任务中，对抗训练常基于百万级样本（如ImageNet）和多机分布式集群进行[56]。在多模态任务中对于计算与资源的开销更为突出，一方面，多模态模型融合了图像编码器、语言编码器和跨模态融合模块，参数规模远超单模态模型；另一方面，图文联合扰动的生成涉及高维输入空间与复杂优化，导致训练过程需频繁执行多轮梯度计算，进一步加剧资源消耗[57]。此外，对抗训练效果在很大程度上依赖于足够的数据规模与训练轮次，若资源受限，不仅难以实现鲁棒性提升，反而可能引发性能退化或过拟合等问题[58]。

受限于现有实验条件（RTX4090 GPU），难以完成较为完整的对抗训练过程。为避免因训练规模受限而导致实验结果缺乏代表性，本文在实验中未对基于CoAtt-attack的对抗训练进行实验。未来我们计划探索基于CoAtt-attack的多模态对抗训练框架，更全面地评估其在提升视觉语言模型鲁棒性方面的应用潜力。

8　总结和未来工作

为解决现有多模态对抗攻击方法中扰动不可解释并且分布不集中，导致难以干扰VLPMs关键决策区域的问题，本文提出了一种基于协同注意力解释的多模态对抗攻击方法CoAtt-attack，该方法在前向传播阶段引入模型自身的跨模态注意机制，以识别模型在决策过程中最为关注的图像区域与文本关键词，并据此引导扰动生成。在图像模态中，利用注意力生成的图注意矩阵对全局扰动进行掩膜操作，仅保留重要区域的扰动，从而在提升攻击效率的同时最大限度保持图像的自然性；在文本模态中，则采用语义保持的BERT-Attack方法生成针对性词替换扰动，进一步增强协同攻击能力。实验结果表明，CoAtt-attack在TR、IR、VG和VE四种视觉语言下游任务中均取得了优于现有方法（Co-attack、VLAttack、SSAP）的攻击效果，并且其生成的图像对抗样本相比于现有方法更具真实性和自然性。

本文是将协同注意力解释应用于VLPMs测试的初步探索。未来，我们计划在更大规模的多模态数据集与更多样化的视觉语言任务场景中验证CoAtt-attack的通用性。此外，在实际应用中，许多部署模型的结构与参数不可见，传统白盒测试方法难以适用，因此我们还将进一步探究CoAtt-attack用于对黑盒VLPMs进行迁移测试。目前具有代表性的多模态模型迁移攻击方法包括TMM（Transferable Multimodal Attack）方法[51]和FGA-T（Feature Guided Adversarial Text）方法[59]。TMM通过特征正交化与注意力共享机制增强对抗样本的跨模型迁移能力；FGA-T结合图像与文本的对抗方向协同构造迁移性扰动，提升攻击在多模型间的适用性。未来我们将结合此类方法的设计思路，进一步拓展CoAtt-attack在黑盒攻击中的适用性。具体而言，可从以下几个方向展开：(1)结合迁移攻击思想，构建通用的跨模态扰动生成框架，利用源模型生成的对抗样本攻击目标模型，从而规避对目标模型注意力分布的依赖；(2)引入弱监督或代理模型机制，采用Grad-CAM、Attention Rollout等技术近似估计黑盒模型的关注区域，引导生成局部扰动；(3)设计基于语义相似性的输入构造策略，在无需访

问模型参数的前提下，通过输入对比或语义对齐机制增强扰动的跨模型通用性。

参 考 文 献

[1] Kim W, Son B, Kim I. Vilt: Vision-and-language transformer without convolution or region supervision//Proceeding. of the 38th International Conference on Machine Learning. New York, USA, 2021: 5583-5594

[2] Huang Z, Zeng Z, Liu B, Fu D, Fu J. Pixel-BERT: Aligning image pixels with text by deep multi-modal transformers. arXiv preprint arXiv:2004.00849, 2020

[3] Zhu X, Zhu J, Li H, Wu X, Li H, Wang X, Dai J. Uni-Perceiver: Pre-training unified architecture for generic perception for zero-shot and few-shot tasks//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 16804-16815

[4] Yang Z, Gan Z, Wang J, Hu X, Ahmed F, Liu Z, Lu Y, Wang L. UniTAB: Unifying text and box outputs for grounded vision-language modeling//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel, 2022: 521-539

[5] Wang P, Yang A, Men R, Lin J, Bai S, Li Z, Ma J, Zhou C, Zhou J, Yang H. OFA: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework//Proceedings of the 39th International Conference on Machine Learning. Virtual, 2022: 23318-23340

[6] Lu J, Clark C, Zellers R, Mottaghi R, Kembhavi A. Unified-IO: A unified model for vision, language, and multi-modal tasks. arXiv preprint arXiv:2206.08916, 2022

[7] Singh A, Hu R, Goswami V, Couairon G, Galuba W, Rohrbach M, Kiela D. FLAVA: A foundational language and vision alignment model//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 15638-15650

[8] Wang F, Mei J, Yuille A. SCLIP: Rethinking self-attention for dense vision-language inference//Proceedings of the 18th European Conference on Computer Vision. Milan, Italy, 2024: 315-332

[9] Kim S, Xiao R, Georgescu M I, Alaniz S, Akata Z. COSMOS: Cross-modality self-distillation for vision language pre-training. arXiv preprint arXiv:2412.01814, 2024

[10] Lu J, Batra D, Parikh D, Lee S. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks//Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver, Canada, 2019: 13-23

[11] Khan Z, Vijay Kumar B G, Yu X, Schulter S, Chandraker M, Fu Y. Single-stream multi-level alignment for vision-language pretraining//Proceedings of the 17th European Conference on Computer Vision. Tel-Aviv, Israel, 2022: 735-751

[12] Agrawal A, Teney D, Nematzadeh A. Vision-Language Pretraining: Current trends and the future//Proceedings of the 60th Annual Meeting of the Association for Computational. Dublin, Ireland, 2022: 38-43

[13] Zhao Y, Pang T, Du C, Yang X, Li C, Cheung M, Lin M. On evaluating adversarial robustness of large vision-language models//Proceedings of the 37rd International Conference on Neural Information Processing Systems. New Orleans, USA, 2023: 54111-54138

[14] Liu S, Yu S, Lin Z, Pathak D, Ramanan D. Language models as black-box optimizers for vision-language models//Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2024: 12687-12697

[15] Hartsock I, Rasool G. Vision-language models for medical report generation and visual question answering: A review. Frontiers in Artificial Intelligence, 2024, 7(1): 14-39

[16] Zhang T, Wang L, Zhang X, Zhang Y, Jia B, Liang S, Hu S, Fu Q, Liu A, Liu X. Visual adversarial attack on vision-language models for autonomous driving. arXiv preprint arXiv: 2411.18275, 2024

[17] Gomez J F, Machado C, Paes L M, Calmon F. Algorithmic arbitrariness in content moderation//Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency. Rio de Janeiro, Brazil, 2024: 2234-2253

[18] Yang K, Lin W Y, Barman M, et al. Defending multimodal fusion models against single-source adversaries//Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual, 2021: 3340-3349

[19] Madry A, Makelov A, Schmidt L, Tsipras D, Vladu A. Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083, 2017

[20] Li L, Ma R, Guo Q, et al. BERT-ATTACK: Adversarial attack against bert using bert//Proceedings of the 2020 Conference on Empirical Methods in Natural Language. Virtual, 2020: 6193-6202

[21] Zhang J, Yi Q, Sang J. Towards adversarial attack on vision-language pre-training models//Proceedings of the 30th ACM International Conference on Multimedia. Dublin, Ireland, 2022: 5005-5013

[22] Yin Z, Ye M, Zhang T, Du T, Zhu J, Liu H, Chen J, Wang T, Ma F. VLAttack: Multimodal adversarial attacks on vision-language tasks via pre-trained models//Proceedings of the 37rd International Conference on Neural Information Processing Systems. New Orleans, USA, 2023: 52936-52956

[23] XieRui-Lin, Cui Zhan-Qi, Chen Xiang, Zheng Li-Wei. IATG: Interpretation-analysis-based testing method for autonomous driving software. Journal of Software, 2024, 35(6): 2753-2774 (in Chinese)
(谢瑞麟，崔展齐，陈翔，郑丽伟. IATG:基于解释分析的自动驾驶软件测试方法. 软件学报，2024，35(06):2753-2774)

[24] Li J, Li D, Xiong C, Hoi S. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation//Proceedings of the 39th International Conference on Machine Learning. Baltimore, USA, 2022: 12888-12900

[25] Chen Y C, Li L, Yu L, Kholy A E, Ahmed F, Gan Z, Cheng

Y, Liu J. Uniter: Universal image-text representation learning//Proceedings of the 16th European Conference on Computer Vision. Virtual, 2020: 104-120

[26] Li X, Yin X, Li C, Zhang P, Hu X, Zhang L, Wang L, Hu H, Dong L, Wei F, Choi Y, Gao J. Oscar: Object-semantics aligned pre-training for vision-language tasks//Proceedings of the 16th European Conference on Computer Vision. Virtual, 2020: 121-137

[27] Zhang P, Li X, Hu X, Yang J, Zhang L, Wang L, Choi Y, Gao J. VinVL: Revisiting visual representations in vision-language models//Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual, 2021: 5579-5588

[28] Wang T, Jiang W, Lu Z, Zheng F, Cheng R, Yin C, Luo P. VLMixer: Unpaired vision-language pre-training via cross-modal cutmix//Proceedings of the 39th International Conference on Machine Learning. Virtual, 2022: 22680-22690

[29] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, Houlsby N. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929, 2020

[30] Touvron H, Cord M, Douze M, Massa F, Sablayrolles A, Jégou H. Training data-efficient image transformers & distillation through attention//Proceedings of the 38th International Conference on Machine Learning. Virtual, 2021: 10347-10357

[31] Yuan L, Chen Y, Wang T, Yu W, Shi Y, Jiang Z, Tay F E, Feng J, Yan S. Tokens-to-token vit: Training vision transformers from scratch on imagenet//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada, 2021: 558-567

[32] Li J, Selvaraju R, Gotmare A D, Joty S, Xiong C, Hoi S. Align before fuse: Vision and language representation learning with momentum distillation//Proceedings of the 35rd International Conference on Neural Information Processing Systems. Virtual, 2021: 9694-9705

[33] Yang J, Duan J, Tran S, Xu Y, Chanda S, Chen L, Zeng B, Chilimbi T, Huang J. Vision-language pre-training with triple contrastive learning.//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 15671-15680

[34] Radford A, Kim J W, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, Krueger G, Sutskever I. Learning transferable visual models from natural language supervision//Proceedings of the 38th International Conference on Machine Learning. Virtual, 2021: 8748-8763

[35] Ramprasaath R, Selvaraju M C, Das A. Visual Explanations from deep networks via gradient-based localization//Proceeding of the IEEE/CVF International Conference on Computer Vision. Seoul, Republic of Korea, 2019: 618-626

[36] Yu L, Lin Z, Shen X, Yang J, Lu X, Bansal M, L.Berg T. MAttNet: Modular attention network for referring expression comprehension//Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Utah, USA, 2018: 1307-1315

[37] Xie R, Chen X, He Q, Li B, Cui Z. IATT: Interpretation Analysis based Transferable Test Generation for Convolutional Neural Networks. ACM Transactions on Software Engineering and Methodology, 2025, 34(4): 1-34

[38] Guo C, Sablayrolles A, Jégou H, Kiela D. Gradient-based adversarial attacks against text transformers//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Punta Cana, Dominican Republic, 2021: 5747-5757

[39] Lu J, Yang J, Batra D, Parikh D. Hierarchical question-image Co-Attention for visual question answering//Proceedings of the 30rd International Conference on Neural Information Processing Systems. Barcelona, Spain, 2016: 289-297

[40] Hendriksen M, Vakulenko S, Kuiper E, Rijke M. Scene-centric vs. object-centric image-text cross-modal retrieval: a reproducibility study//Proceedings of the 2023 European Conference on Information Retrieval. Dublin, Ireland, 2023: 68-85

[41] Yu L, Poirson P, Yang S, Alexander C. Berg, Tamara L. Berg. Modeling context in referring expressions//Proceedings of the 14th European Conference on Computer Vision. Amsterdam, TheNetherlands, 2016: 69-85

[42] Xie N, Lai F, Doran D, Kadav A. Visual entailment: A novel task for fine-grained image understanding. arXiv preprint arXiv: 1901.06706, 2019

[43] Devlin J, Chang M W, Lee K, ToutanovaK. Bert: Pre-training of deep bidirectional transformers for language understanding//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics. Minneapolis, USA, 2019: 4171-4186

[44] Lin C, Shen C, Deng J, Hu P, Wang Q, Ma S, Li Q, Guan X. Digitally Forged Face Content Creation and Detection. Chinese Journal of Computers, 2024,47(3):469-498 (in Chinese)
(蔺琛皓,沈超,邓静怡, 胡鹏斌, 王骞, 马仕清, 李琦, 管晓宏. 虚假数字人脸内容生成与检测技术. 计算机学报,2024,47(3): 469-498)

[45] Jähne B. Digital image processing. Springer Science & Business Media; 2005

[46] Zhang K, Li Y, Zuo W, Zhang L, Gool L V, Timofte R. Plug-and-play image restoration with deep denoiser prior. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 44(10): 6360-6376

[47] Qiu H, Xiao C, Yang L, Yan X, Lee H, Li B. Semanticadv: Generating adversarial examples via attribute-conditioned image editing//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK, 2020: 19-37

[48] Wang Z, Bovik A C, Sheikh H R, Simoncelli E P. Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing, 2004, 13(4): 600-612

[49] Zhang T, Kishore V, Wu F, Weinberger Q. K, Artzi Y. Bertscore: Evaluating text generation with bert. arXiv preprint arXiv:1904.09675, 2019

[50] Disabato R, MaungMaung A, Nguyen H H, Echizen I.

Transfer-Based Adversarial Attack Against Multimodal Models by Exploiting Perturbed Attention Region//Proceedings of the 2024 Asia Pacific Signal and Information Processing Association Annual Summit and Conference. Macao, China, 2024: 1-6

[51] Wang H, Dong K, Zhu Z, Qin H, Liu A, Fang X, Wang J, Liu X. Transferable multimodal attack on vision-language pre-training models//Proceedings of the 2024 IEEE Symposium on Security and Privacy. San Francisco, USA, 2024: 1722-1740

[52] Guan J, Ding T, Cao L, Pan L, Wang C, Zheng X. Probing the robustness of vision-language pretrained models: A multimodal adversarial attack approach. arXiv preprint arXiv: 2408.13461, 2024

[53] Li J, Ni M, Dong Y, Zhu T, Liu W. AICAttack: Adversarial image captioning attack with attention-based optimization. arXiv preprint arXiv:2402.11940, 2024

[54] Karras T, Laine S, Aila T. A style-based generator architecture for generative adversarial networks//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Angeles, USA, 2019: 4401-4410

[55] Karras T, Laine S, Aittala M, Hellsten J, Lehtinen J, Aila T. Analyzing and improving the image quality of stylegan//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 8110-8119

[56] Salman H, Ilyas A, Engstrom L, Kapoor A, Madry A. Do adversarially robust imagenet models transfer better//Proceedings of the 34rd International Conference on Neural Information Processing Systems. Virtual, 2020: 3533-3545

[57] Shang Y, Gao C, Chen J, Jin D, Ma H, Li Y. Enhancing adversarial robustness of multi-modal recommendation via modality balancing//Proceedings of the 31st ACM International Conference on Multimedia. Ottawa, Canada, 2023: 6274-6282

[58] Zhang H, Yu Y, Jiao J, Xing E, El Ghaoui L, Jordan M. Theoretically principled trade-off between robustness and accuracy//Proceedings of the 36th International Conference on Machine Learning. New York, USA, 2019: 7472-7482

[59] Zheng H, Deng X, Jiang W, Li W. A unified understanding of adversarial vulnerability regarding unimodal models and vision-language pre-training models//Proceedings of the 32nd ACM International Conference on Multimedia. Melbourne, Australia, 2024: 18-27

HAN Qi-Hong, M. S. candidate. His research interests include trustworthy artificial intelligence.

CUI Zhan-Qi, Ph. D., professor. His research interests include intelligent software engineering and trustworthy artificial intelligence.

CHEN Xiang, Ph. D., associate professor. His research interests include software testing, software maintenance, empirical software engineering, and mining software repository.

CHEN Jing-Jing, Ph. D., experiment teacher. Her research interests include optical NDT and machine vision.

LI Li, Ph. D., associate professor. Her research interests include data science, intelligent systems, and data fusion

Background

The research of this paper belongs to the field of robustness testing for Vision-Language Pre-training Models (VLPMs). VLPMs have shown impressive performance in multimodal tasks such as image-text retrieval, visual grounding, and visual entailment, and have been widely deployed in practical scenarios like autonomous driving, content moderation, and intelligent question answering. However, as these models are increasingly applied in safety-critical domains, issues such as limited interpretability and sensitivity to input perturbations have drawn attention, posing significant risks to system stability and reliability. Therefore, evaluating the robustness of VLPMs through effective testing has become a key research problem in multimodal learning.

Existing adversarial testing methods primarily involve perturbations in the image or text modality alone, or jointly in both modalities. Recently, multimodal adversarial attacks have gained popularity, achieving higher attack success rates than single-modality methods. Nonetheless, most of them still rely on global perturbation strategies without leveraging the internal attention mechanisms of VLPMs, leading to adversarial examples that lack interpretability and exhibit scattered perturbation regions, making it difficult to expose model vulnerabilities.

To address this issue, this paper proposes a Co-Attention Interpretability Based Multimodal Adversarial Attack (CoAtt-attack). The method utilizes the Co-Attention mechanism in VLPMs to identify critical visual-textual regions and guides perturbations in the image modality to focus on those regions. Simultaneously, it generates semantically consistent adversarial texts via the BERT-Attack algorithm. Experiments conducted on three mainstream VLPMs (ALBEF, TCL, and CLIP) and four representative tasks (TR, IR, VG, and VE) demonstrate that CoAtt-attack achieves superior attack success rates and better image fidelity compared to existing methods, offering a more interpretable and effective pathway for robustness evaluation of VLPMs.