北京航空航天大学计算机学院

硕士研究生学位论文

文献综述

论文题目：面向视觉语言模型的对抗图像生成方法研究

专　　业：计算机技术

研究方向：智能安全

研 究 生：邓彬

学　　号：ZY2306335

指导教师：郭园方

北京航空航天大学计算机学院

2024 年 11 月 24 日

摘　　要

随着人工智能技术的快速发展，视觉语言模型（Visual-Language Models, VLMs）作为结合计算机视觉与自然语言处理的关键技术，已成为实现图像与文本多模态信息理解的重要工具。VLMs 通过 Transformer 架构的自注意力机制，能够有效处理图像与文本数据，在多模态理解和推理任务上展现了出色的性能。本文首先回顾了 VLMs 的四种主流架构：基于对比学习的模型、基于掩码图像建模的模型、基于大语言模型的模型以及生成式模型，这些架构为 VLMs 的发展提供了多样化的技术路径，并在多个应用领域展现出强大的潜力。

随后，本文介绍了面向 VLMs 的对抗图像生成方法。对抗图像攻击（简称，对抗攻击）指的是通过在输入的图像中施加精心设计的扰动，使得模型生成攻击者指定的文本或执行非预期的行为。这些攻击利用了模型在处理多模态数据时的脆弱性，对模型的安全性和可靠性提出了严峻挑战。然后，本文深入探讨了面向 VLMs 的对抗攻击方法，特别是跨模型迁移性增强方法、跨提示迁移性增强方法和跨数据迁移性增强方法。

最后，本文对现有方法进行了总结和做出了未来展望，期望这些见解能助研究者在视觉语言模型上开发更有效的对抗样本生成方法，以确保技术的健康发展和应用安全。

关键词：视觉语言模型；对抗攻击；对抗图像；迁移性；

Abstract

With the rapid development of artificial intelligence technology, Visual-Language Models (VLMs), as a key technology combining computer vision and natural language processing, has become an important tool for multimodal information comprehension of images and texts.VLMs, through the self-attention mechanism of the Transformer architecture, are able to efficiently deal with image and text data. text data, and has demonstrated excellent performance in multimodal understanding and reasoning tasks. This paper first reviews four mainstream architectures of VLMs: contrast learning-based models, masked image modeling-based models, models based on large language models, and generative models, which provide diversified technical paths for the development of VLMs and show strong potentials in several application areas.

Subsequently, this paper introduces adversarial attacks for VLMs. Adversarial attacks refer to the process of making a model generate attacker-specified text or perform unintended behaviors by applying elaborate perturbations to the input images. These attacks exploit the model's vulnerability in handling multimodal data, posing a serious challenge to the model's security and reliability. Then, this paper provides an in-depth discussion of VLMs-oriented methods for countering the attacks, especially cross-model transferability enhancement methods, cross-prompt transferability enhancement methods, and cross-data transferability enhancement methods.

Finally, this paper summarizes the existing methods and makes future outlooks, expecting that these insights can help researchers develop more effective methods for countering sample generation on visual language models to ensure the healthy development of the technology and application security.

Keywords： Visual-Language Models; Adversarial Attack; Adversarial Image; Transferability;

目 录

图　　目

1 视觉语言模型的研究现状

视觉语言模型（Visual-Language Models, VLMs）结合了计算机视觉与自然语言处理技术，旨在实现图像与文本的多模态信息理解。它们在图像描述生成（Image Captioning, IC）、视觉问答（Visual Question Answering, VQA）和视觉定位（Visual Grounding, VG）等跨模态任务中展现了巨大的潜力。近年来，Transformer 架构在自然语言处理和计算机视觉领域的广泛应用，为视觉语言模型的发展提供了强大支持。得益于 Transformer[1]的自注意力机制，模型能够同时处理图像与文本数据，为多模态信息理解提供了创新解决方案。因此，基于 Transformer 的视觉语言模型已逐渐成为多模态任务研究的主流，并在多个应用领域取得了显著进展。

视觉语言模型的早期发展受到 BERT（Bidirectional Encoder Representations from Transformers）[2]在自然语言处理领域成功的启发。随着 BERT 模型的广泛应用，许多研究者开始将其架构扩展到图像与文本结合的跨模态任务中。Visual-BERT[3]和 ViLBERT[4]是这一探索的代表性模型，它们首次将 Transformer 架构应用于视觉语言任务，尤其在掩码语言建模和图像-文本匹配任务中取得了显著成果。这些模型通过 Transformer 的自注意力机制，成功捕捉图像与文本之间的相互关系，促进了模型对多模态信息的理解。这些开创性工作为后续研究提供了宝贵的经验，并为进一步的发展奠定了坚实的基础。

随着视觉语言模型应用需求的不断增长，研究者提出了四种主流架构类型[5]。第一类是基于对比学习的模型，这类模型通过对正负样本的对比，推动图像-文本对的嵌入空间对齐，从而显著提高跨模态对齐效果。第二类是基于掩码图像建模的模型，这类模型通过部分遮掩输入的图像或文本信息，迫使模型在恢复缺失内容时进行推测和重建，提升了推理能力。第三类是基于大语言模型构建的模型，这类模型通常采用大型预训练语言模型（如 Llama[6]等）作为骨干，结合图像编码器与语言模型对齐，显著提高图像-文本匹配的准确度，并在多个任务中展现出优异性能。第四类是生成式模型，这些模型不仅能直接生成图像或文本，广泛应用于图像描述和生成等任务，还能为创新应用提供解决方案，如图像生成与编辑。

这四种主流架构为视觉语言模型的发展提供了多种可行的技术路径，从高效的跨模态对齐到创新的图像生成，极大地推动了视觉语言模型的应用与技术进步。值得注意的是，这些架构并非孤立存在，许多模型在设计时巧妙地融合了对比学习、掩码训练与生成策略的优点。接下来，我们将深入探讨每种架构中的一到两个代表性模型，并详细分析其设计思路与应用场景。

1.1 基于对比学习的 VLMs

基于对比学习的视觉语言模型旨在学习图像与文本之间的对齐表示，其核心思想是通过对比损失函数优化模型参数，促使图像和文本在嵌入空间中实现跨模态的语义关联。具体而言，这类模型将正样本对（由图像及其对应的真实文本描述组成）与负样本对（将同一张图像与描述其他图像的文本配对）进行对比学习。通过区分正负样本对，这些视觉语言模型能够将语义相关的图像和文本映射到嵌入空间中的相近位置，而不相关的样本则被推向较远的位置。这种方法有助于提高跨模态对齐的精度，使模型能够更好地理解图像与文本之间的关系。

CLIP（Contrastive Language-Image Pre-Training）[7]是这一领域的代表性模型。CLIP 利用对比学习框架训练视觉和文本编码器，使图像和其对应的文本描述在嵌入空间中具有相似的向量表示。CLIP 的训练数据集包含了四亿对来自网络的图像和文本。这些大规模的数据使得 CLIP 能够学习到丰富的视觉和语言特征，并在多个任务中展现出强大的性能，尤其是在零样本分类任务中表现卓越。以基于 ResNet-101 的 CLIP 模型为例，它在零样本分类准确率上达到了 76.2%，并且在多个鲁棒性基准测试中，CLIP 模型超越了经过监督训练的 ResNet[8]模型。

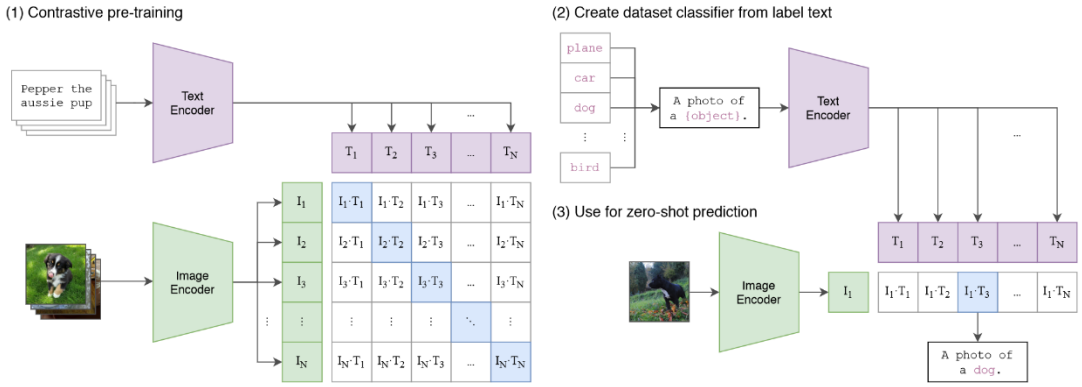

图 1 CLIP 架构[7]

SigLIP（Signature Language Image Pre-training）[9]是一种基于 CLIP 的改进模型，其主要创新体现在损失函数的设计上。与 CLIP 使用的信息噪声对比估计（Information Noise Contrastive Estimation，InfoNCE）[10]损失函数不同，SigLIP 采用了基于二元交叉熵的噪声对比估计（Noise Contrastive Estimation，NCE）[11]损失函数。通过这一调整，SigLIP 能够在较小批量数据下仍然展现出优异的零样本性能，从而减少了对大规模数据的依赖。

Llip（Latent Language Image Pretraining）[12]模型进一步改进了图像与文本之间的关联方式。考虑到一张图像可以用多种不同的方式进行描述，Llip 提出了一种新的条件编码机制，通过交叉注意力模块来根据目标描述调整图像的编码。这种方法的核心在于通过动态调整图像编码，使得每个图像可以针对不同的描述生

成不同的编码表示，从而提高了编码的多样性和表达能力。Llip 通过这种条件编码机制，能够灵活地适应各种任务需求，并进一步推动了跨模态学习的进展。

1.2 基于掩码图像建模的 VLMs

在大语言模型的早期研究中，BERT 通过掩码语言建模（Masked Language Modeling, MLM）技术，预测句子中缺失的词元，从而显著提升了在多种自然语言处理任务中的表现。受到 BERT 文本掩码技术的启发，视觉语言模型领域也开始采用类似的掩码策略，特别是在图像编码方面。例如，MAE（Masked Autoencoders）[13]和 I-JEPA（Image-Joint Encoding and Pretraining Architecture）[14]便是通过掩码图像建模（Masked Image Modeling, MIM）技术来训练图像编码器。这些基于掩码的视觉语言模型的核心思想是，通过随机去除图像输入的部分区域，迫使模型在缺失信息的情况下进行推理，从而提升其跨模态信息的理解能力。

FLAVA（Foundational Language and Vision Alignment）[15]是基于掩码方法的典型模型。该模型架构包括三个核心组件，均基于 Transformer 框架，并针对不同模态进行了优化。具体而言，图像编码器采用 ViT（Vision Transformer）[16]将图像分割为图像块，并通过线性嵌入和 Transformer 表示进行处理，处理过程中还会附带一个分类标记（$[CLS_I]$）。文本编码器则使用 Transformer 对文本进行标记化，将文本嵌入为向量并进行上下文处理，输出隐藏状态向量，并附带一个分类标记（$[CLS_T]$）。这两个编码器都采用了掩码训练方法。多模态编码器则融合了图像和文本编码器的隐藏状态，借助线性投影和跨模态注意力机制有效整合视觉与文本信息，并引入额外的多模态分类标记（$[CLS_M]$）。FLAVA 模型通过结合多模态和单模态的掩码建模损失，并辅以对比学习目标，展现了卓越的多功能性和有效性。在 7000 万对公开图像和文本数据上进行预训练后，FLAVA 在 35 个涵盖视觉、语言和多模态的基准任务中取得了最先进的性能，展示了其强大的跨领域信息理解与整合能力。

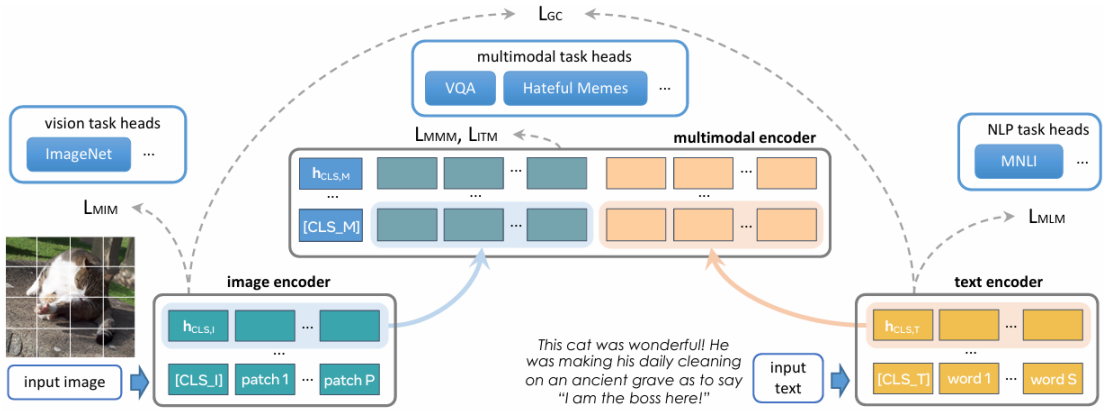

图 2 FLAVA 架构[15]

尽管 FLAVA 模型在多模态任务中表现出色，但其局限性在于依赖于预训练

的视觉编码器（例如 D-VAE[17]）。为了解决这一问题，Kwon 等人提出了 MaskVLM（Masked Vision-Language Model）[18]。与 FLAVA 不同，MaskVLM 不依赖于预训练的视觉编码器，而是直接在像素空间和文本标记空间上应用掩码策略。这一创新使得 MaskVLM 能够在没有外部视觉编码器的帮助下，直接处理图像和文本输入。MaskVLM 通过信息流动机制，允许一个模态中的信息有效传递到另一个模态，从而使得模型能够在视觉和语言两个模态之间进行有效的联合学习。例如，在文本重构任务中，MaskVLM 利用图像编码器的信息来辅助重构文本内容；而在图像任务中，模型则可以利用文本编码器提供的信息来辅助图像的处理和理解。

1.3 基于大语言模型的 VLMs

基于大语言模型的视觉语言模型利用已训练的大型语言模型或视觉特征提取器，学习文本与图像之间的映射关系。该方法的主要优势在于能够充分发挥预训练模型的丰富特征，从而显著减少从零开始训练所需的计算资源和数据量。其核心理念是通过有效结合视觉编码器与大型语言模型，避免了重新训练，从而实现对多模态数据的深刻理解。

Frozen[19]是首批将大型语言模型应用于视觉语言任务的开创性模型之一。该模型通过一个轻量级映射网络，将视觉编码器连接至冻结状态的大语言模型。映射网络负责将视觉特征投影到文本标记嵌入空间。Frozen 使用 NF-ResNet-50[20]作为视觉编码器，并与一个线性映射层连接，二者从头开始训练，而大语言模型（例如，在 C4[21]数据集上训练的、参数为 7 亿的 Transformer）保持冻结状态，以保留其预先学习到的重要特征。在推理阶段，Frozen 能够条件化生成文本，展示了其快速适应新任务、获取通用知识以及融合视觉与语言元素的能力。Frozen 主要用于文本生成任务，如图像描述任务。在推理过程中，该模型接受图像和文本嵌入的输入，并生成与图像内容相关的文本描述。尽管 Frozen 的性能相对中等，但它为后续的开放式多模态零样本/少样本学习发展奠定了基础。

MiniGPT[22]系列进一步扩展了这一概念，使得同时接收文本和图像输入并生成相应文本输出成为可能。其中，MiniGPT-4 通过简单的线性投影层，将 BLIP-2[23]所用的图像编码器产生的图像嵌入与 Vicuna 语言模型的输入空间有效对接。由于这两个组件均为预训练，MiniGPT-4 仅需针对线性投影层进行训练。MiniGPT-5 则在此基础上增加了图像生成功能，使用生成标记创建新图像，这些标记被映射到特征向量并输入到冻结的图像生成模块中，进一步提升了多模态处理能力。

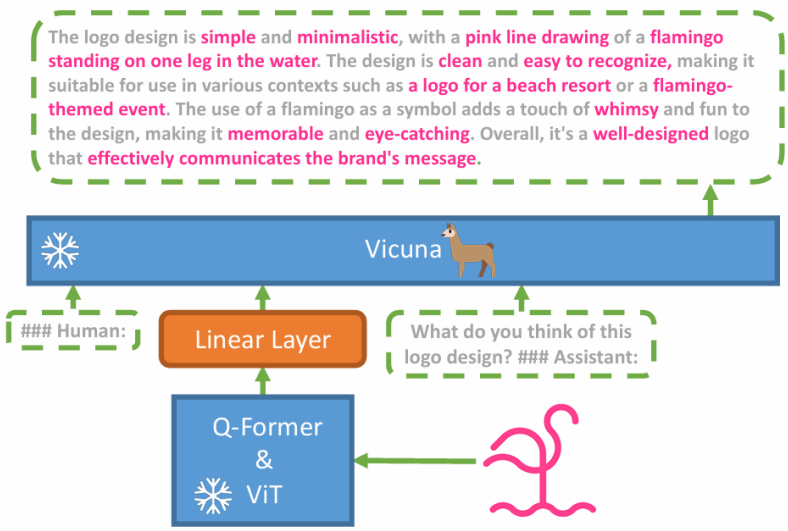

图 3 MiniGPT-4 架构[22]

继 MiniGPT 系列创新后，Bai 等人推出的 Qwen 系列模型[24]，包括 Qwen-VL 和 Qwen-VL-Chat，在多模态交互领域取得了显著进展。Qwen 模型结合了大型语言模型和视觉编码器，增强了对视觉和语言信息的处理能力。在 Qwen 架构中，LLM 以 Qwen-7B[25]为基础初始化，视觉编码器采用 ViT-bigG。模型通过单层交叉注意力模块将视觉表示压缩成固定长度的序列，这些序列被输入到 LLM 中。这一设计提升了模型处理视觉和语言信息的效率，促进了多模态任务中更有效的信息交互，为多模态技术的发展提供了新的动力。

在多模态交互领域，Li 等人提出的 BLIP-2 模型高效利用预训练模型，为图像到文本的转换提供了创新方案[23]。BLIP-2 通过冻结的预训练模型大幅缩短训练时间，使用如 CLIP 的视觉编码器生成图像嵌入，并将其映射到大型语言模型（如 OPT）的输入空间。BLIP-2 的关键组件是 Q-Former，一个约含 100-200M 参数的 Transformer，通过交叉注意力机制将随机初始化的 Query 向量与图像嵌入交互，并通过线性层将其投影到 LLM 输入空间。这种方法提高了训练效率，并在多模态任务中实现了更精细的特征对齐，显著提升了图像理解与文本生成的紧密联系。

1.4 生成式 VLMs

生成式视觉语言模型是一类利用生成模型处理和理解视觉与语言信息的模型。这些模型能够生成图像或文本，或在给定一种模态的条件下生成另一种模态的内容。生成式模型的核心在于学习输入数据的分布，并基于该分布生成新的数据实例。在视觉语言模型的背景下，这意味着模型能够根据文本描述生成图像，或根据图像生成文本描述。生成式视觉语言模型的

CoCa（Contrastive Captioner）[26]是一种生成式文本生成器，结合了对比学习和生成损失来训练多模态文本解码器。CoCa 接受图像编码器的输出和来自单模

态文本解码器的文本嵌入，生成与图像内容相关的文本。CoCa 在预训练阶段利用大规模的图像和文本数据集，通过将图像标签视为文本，学习图像和文本之间的关联。这使得 CoCa 不仅能生成图像描述，还能执行其他多模态理解任务，如视觉问答任务，而无需额外的适应性调整。

CM3leon[27]是一个多模态生成模型，专注于文本到图像和图像到文本的生成任务。CM3leon 借鉴了 Make-A-Scene 的[28]图像编码器和 OPT[29]的文本分词器的设计，将图像和文本编码为一系列嵌入，然后通过 Transformer 模型处理这些嵌入。CM3leon 的训练分为两个阶段：首先是检索增强的预训练，使用基于 CLIP 的编码器作为检索器获取相关的多模态文档，并将其加入输入序列中；接着是监督式微调（Supervised Fine-Tuning, SFT），通过多任务指令调整模型，以处理并生成不同模态的内容。这一两阶段的训练方法使得 CM3leon 在多模态任务中表现优异，展示了自回归模型在处理文本和图像间复杂交互的能力。

Chameleon[30]是一系列混合模态基础模型，能够生成并推理包含文本和图像内容的序列。Chameleon 模型从一开始就设计为混合模态，使用统一架构处理所有模态——图像、文本和代码。这种集成方法采用基于标记的表示，将图像和文本转换为离散标记，使得相同的 Transformer 架构可以应用于图像和文本标记序列，而无需为每种模态单独设计编码器。早期的融合策略使得模型能够在不同模态之间无缝推理和生成，但也带来了优化稳定性和扩展性的技术挑战。

2 面向视觉语言模型的对抗图像生成方法

随着视觉语言模型的快速发展，对抗攻击的研究已经从单一的视觉模型扩展到了更为复杂的多模态模型。早期的研究主要集中于通过微小的输入扰动误导模型的视觉识别能力[31]。然而，随着视觉语言模型在多模态理解和推理任务中表现出色，攻击者开始探索操纵图像和文本输入的新方法，并且设计出多样化的攻击效果。这些多模态攻击策略不仅增加了攻击的多样性，也提高了对模型鲁棒性的要求，推动了新的防御机制的发展。

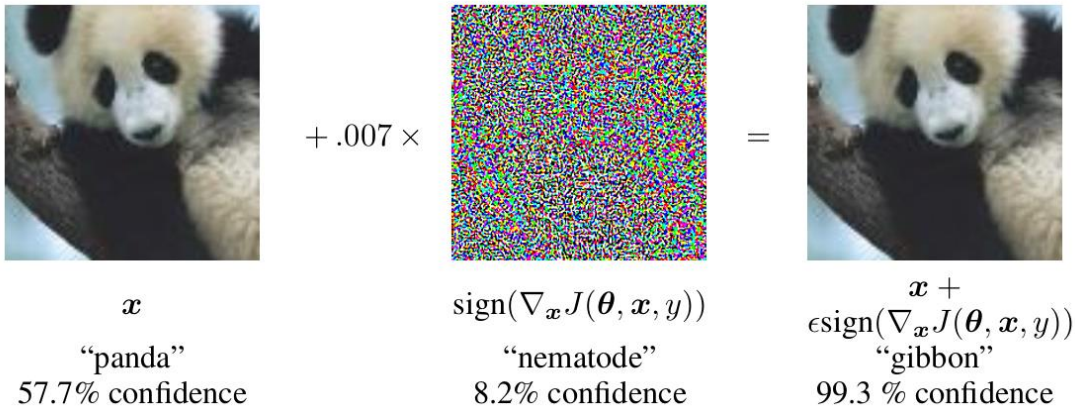

图 4 针对传统视觉模型的对抗图像实例[32]

对抗攻击的定义是通过在输入的图像中施加精心设计的扰动，使得模型生成攻击者指定的文本[32]，例如错误的图像描述。这些攻击利用模型在处理多模态数据时的脆弱性，通过引入微小的、通常不易被人类察觉的扰动，诱导模型执行非预期的行为。因此，研究者需要关注这些攻击策略，并开发有效的防御措施，以确保视觉语言模型的安全性和可靠性。

在自动驾驶和医疗诊断等高风险场景中，视觉语言模型的应用对安全性提出了极高的要求。一旦模型遭受对抗攻击，可能导致关键任务中的严重错误。例如，在自动驾驶中，模型可能误识交通标志，从而引发交通事故；在医疗诊断中，模型在医学影像辅助诊断中出现偏差，可能导致误诊或漏诊，危及患者生命。这些潜在风险凸显了提升模型安全性的重要性，以确保其在关键领域的可靠性和稳定性。一些攻击者利用对抗样本诱导模型生成歧视性内容或侵犯个人隐私，这可能对社会造成深远的负面影响。通过研究对抗攻击，能够揭示模型在应对此类攻击时的潜在弱点，进而开发更有效的防护措施。这不仅有助于增强模型的鲁棒性，还能确保技术符合伦理规范，促进技术的健康发展。

在这一背景下，一系列针对视觉语言模型的对抗攻击研究应运而生。这些研究不仅揭示了多模态系统的潜在安全风险，还为构建更健壮的模型提供了宝贵的见解。例如，Carlini 等人[33]延续早期对抗样本生成的技术，通过最大化对抗图像扰动对模型输出有害文本的概率，指出大型语言模型和多模态模型在面对对抗攻

击时可能违背设计原则，进而生成有害文本。Qi 等人[34]则关注视觉语言模型的越狱攻击，通过计算图像对抗扰动，绕过模型的安全防护，迫使模型执行本应被拒绝的有害指令。Bagdasaryan[35]的研究强调了在图像特定区域嵌入对抗扰动的可能性，这种方法能在不显著改变图像语义内容的情况下，引导模型生成攻击者指定的文本。Schlarmann 等人[36]提出了非定向攻击，通过降低正确文本输出的概率来计算对抗扰动。然而，由于图像的正确描述可能有无穷多种，尽管对抗扰动能够避免某一正确描述，模型仍可能生成其他符合图像的正确描述，这在一定程度上削弱了攻击效果。Bailey 等人[37]则提出了“图像劫持”的概念，介绍了一种新的行为匹配算法，用于训练图像劫持，并展示了如何利用这一技术实施多种攻击，包括特定字符串攻击、上下文泄露攻击、越狱攻击和“幻觉”攻击。

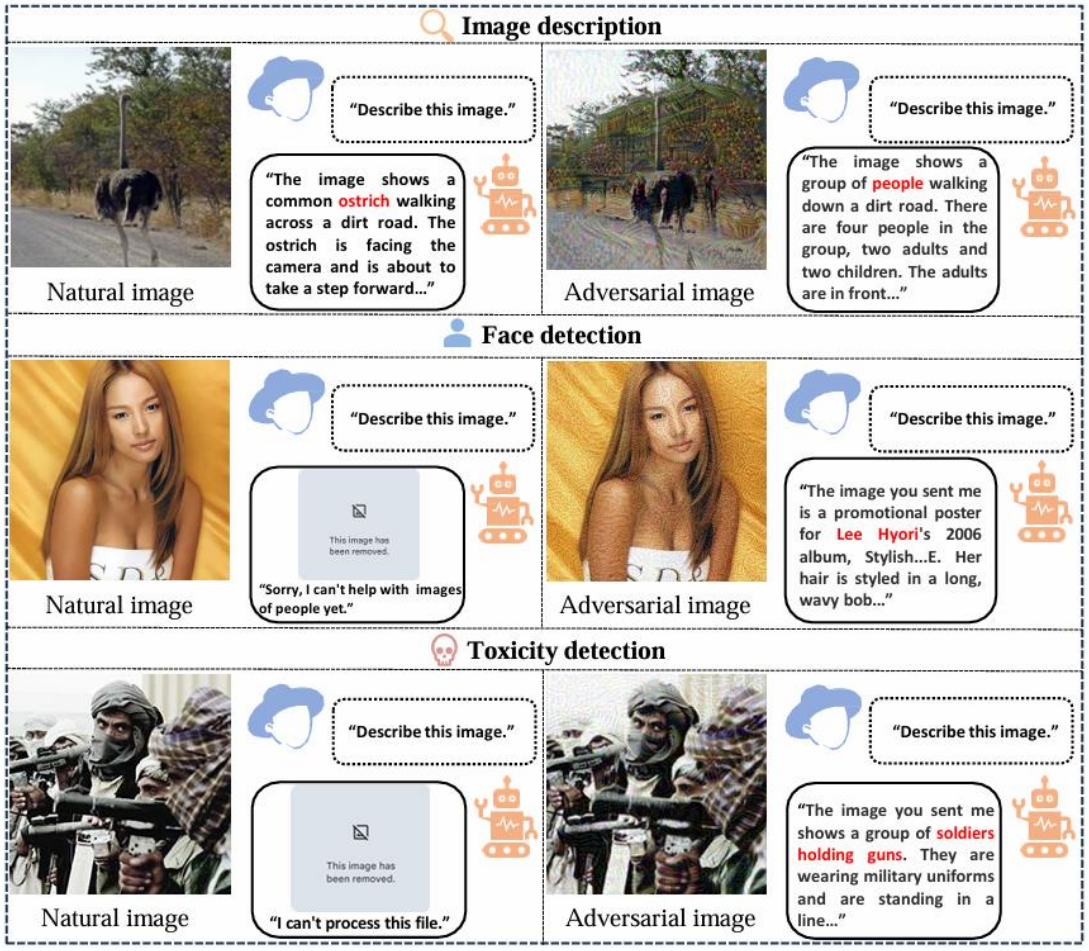

图 5 针对视觉语言模型的对抗图像实例[49]

对抗图像的存在引起了人们对机器学习系统稳健性和可靠性的广泛关注。尤其引人注目的是其迁移性，即一个模型上生成的对抗样本能够误导其他结构不同的模型[38]。这一特性大大降低了攻击者的攻击难度，因为他们无需了解目标模型的具体架构和参数。更令人关注的是，学者们发现，对抗扰动即使被应用到不同的图像上，依然能够成功误导模型，表明对抗扰动具有一定的通用性，能够跨越

图像内容发挥作用[39]。这种跨图像迁移性进一步扩大了对抗样本的影响范围及潜在威胁。此外，研究还发现，针对特定任务生成的图像对抗扰动，在其他任务中也能产生影响。这意味着，对抗图像不仅能够在同一任务的不同模型间迁移，还能跨任务挑战模型的泛化能力[40]。本文将介绍关于视觉语言模型中对抗样本在跨模型、跨提示和跨数据迁移性方面的研究进展[41]。

2.1 跨模型迁移性增强方法

对抗样本的跨模型迁移性是指同一个对抗样本能够成功地欺骗不同结构或训练数据不同的模型。对此现象的原因，学术界已有多种解释。Nguyen 等人[42]提出，不同模型在训练过程中可能会捕捉到相似的图像特征，这为对抗样本的迁移提供了可能性。Goodfellow 等[32]则从模型权重的角度分析，认为由于训练数据的相似性，不同模型的权重趋向于一致，从而促进了对抗样本的迁移。此外，Dong 等[43]学者强调，模型的决策边界是对抗样本生成的关键因素。

虽然对抗样本的跨模型迁移性已经有了一些理论解释，但如何让对抗样本在不同模型间更加有效地“迁移”仍然是一个研究热点。为了增强这一效果，研究者们提出了多种方法，旨在提高对抗样本在不同模型上的适应性。接下来，我们将介绍一些具体的技术和策略，这些方法能够帮助对抗样本更好地跨模型传播。

针对早期的视觉模型，Xie 等人[44]首次提出通过对原始图像进行可微分变换以增强对抗样本的迁移性。在生成对抗扰动的每一步中，他们引入了随机图像变换操作，例如按一定概率对图像进行大小调整和填充。实验表明，随着变换概率的增加，对抗样本的跨模型迁移能力显著提高。除了数据增强的策略外，Dong 等人[43]进一步优化了对抗样本的生成方法，通过将动量机制引入快速梯度符号法（FGSM），提出了改进算法 MI-FGSM（Momentum Iterative Fast Gradient Method）。这一方法利用动量累计历史梯度信息，使对抗扰动的更新更加稳定，从而提升了跨模型的迁移效果。此外，Li 等人[45]针对可迁移对抗样本的两个关键问题进行了研究：一是迭代攻击中梯度幅值逐渐减小，导致动量累积时连续两次扰动过于相似，即“噪声固化”问题；二是对抗样本需要同时逼近目标类别并远离真实类别的矛盾性。为此，他们首次引入庞加莱球作为度量空间，解决了噪声固化问题，使梯度幅值能够自适应调整，并提升了噪声方向的灵活性。同时，他们设计了一种基于庞加莱距离的损失函数，以替代传统交叉熵损失，仅在逼近目标类别时施加梯度更新，从而进一步增强对抗攻击的迁移效果。

在视觉语言模型领域，对抗样本的跨模型迁移性同样备受关注。由于这类模型不仅处理视觉特征，还需结合语言信息，其复杂性使得增强跨模型迁移性的方法更具挑战性。针对这一问题，研究者们提出了一系列针对视觉语言模型的优化策略，试图通过融合多模态特征和改进生成机制，提升对抗样本在视觉语言模型

间的迁移效果。在集成模型方向，Guo 等人[46]提出的 AdvDiffVLM 方法使用自适应集成梯度估计（Adaptive Ensemble Gradient Estimation）模块。该模块通过多个代理模型估计目标模型的梯度信息，从而在扩散模型生成对抗图像的过程中有效嵌入对抗语义。Niu 等人[47]设计了一种基于最大似然的算法，用于生成图像越狱提示以攻击视觉语言模型。他们采用了多个代理模型，包括基于 Vicuna-7B、Vicuna-13B 和 LLaMA-2-7B 的 MiniGPT-4。Wu 等人[48]则通过整合 ViT-B/32、ViT-B/16、ViT-L/14 和 ViT-L/14@336px 等模型，对多模态智能体进行幻觉攻击与目标误导攻击。Dong 等人[49]进一步提出，在生成图像扰动时以 ViT-B/16、CLIP 和 BLIP-2 作为代理模型，并结合 SSA-CWA （Spectrum Simulation Attack-Common Weakness Attack）[50]进行攻击，以提升对抗样本的迁移性。此外，部分研究将图像嵌入作为提升跨模型迁移性的切入点。例如，Dong 等人[49]提出一种图像嵌入攻击方法，旨在通过增加对抗图像与原始图像嵌入之间的差异来误导模型。针对可能导致模型输出其他正确描述的风险，Zhao 等人[51]提出结合扩散模型将目标文本转换为图像，并通过拉近对抗图像与生成图像嵌入的相似度来进行优化。在文本描述攻击方面，Zhao 等人[51]提出采用随机梯度无关（RGF, Random Gradient-Free）方法来估计梯度，成功攻击了未见过的视觉语言模型。

从模型对齐的角度，Ma 等人[52]提出了一种微调源模型（Source Model）的策略，使其输出与一组独立的见证模型（Witness Models）的输出接近，然后利用源模型生成对抗样本，从而提升跨模型迁移能力。Lu 等人[53]提出了 SGA（Set-level Guidance Attack） 方法，通过结合跨模态交互和对齐保持的数据增强技术，生成可在黑盒环境中高效攻击多个模型的对抗样本。Han 等人[54]则从最优传输（Optimal Transport, OT）理论出发，将图像和文本集合的特征视为两个分布，通过最优传输计算两者间的最优映射关系，有效缓解过拟合问题，并提升对抗样本的迁移性。

2.2 跨提示迁移性增强方法

近年来，随着对跨模型迁移性的深入研究，一些工作发现针对特定任务生成的图像对抗扰动可能会在其他任务中产生意料之外的影响[55]。例如，用于图像分类任务生成的对抗样本，在应用于图像分割任务时，仍然能够干扰模型的输出[56]。这一现象揭示了对抗扰动的跨任务迁移性，即使对抗样本专为特定任务优化生成，其影响也可能扩展到其他任务，挑战了模型的鲁棒性和泛化能力。

在视觉语言模型的研究中，研究者们进一步提出了跨提示迁移性的概念，即对抗样本不仅能在单一提示下误导模型，还能在其他的文本提示下保持误导有效性。相比于跨任务迁移性，跨提示迁移性更为重要，因为它更贴近视觉语言模型的实际应用场景。这种迁移性旨在开发能够在多个不同提示条件下均有效的对抗

样本，而非仅限于单一提示。提示是视觉语言模型适配任务的核心驱动力，通过不同的文本提示，模型能够高效完成多样化的任务。与跨任务迁移性主要关注任务类型的多样性不同，跨提示迁移性强调任务内部因提示变化而导致的细粒度差异。这种差异在视觉语言模型中尤为显著。例如，在视觉问答任务中，提示通常以具体问题的形式呈现，如“图像中有多少只鸟？”或“图像中的物体是什么颜色？”；而在图像描述任务中，提示则更为单一，如“为这张图片生成一段描述”。研究跨提示迁移性，不仅能够加深对模型在不同提示下表现的理解，还能更全面地评估其在真实场景中的适配能力，推动模型在实际应用中的稳健性发展。

Bailey 等人[37]研究了视觉语言模型在图像输入下的安全性问题，重点探讨了“图像劫持”现象，即通过对抗图像操控模型的推理输出。在跨提示迁移性的分析中，他们设计了特定的字符串攻击，并采用不同的范数约束生成对抗图像。这些对抗图像随后被用于测试其在未见上下文中的表现。实验结果表明，这些图像在跨提示迁移性方面表现出显著的效果。Lu 等人[57]提出了一种名为 AnyDoor 的后门攻击方法，旨在无需修改或接触训练数据的情况下，通过测试阶段的后门注入来操控模型行为。AnyDoor 通过优化包含触发器和无触发器场景的损失函数，动态调整权重以平衡两种场景下的性能，从而实现了良好的跨提示迁移能力。实验结果显示，AnyDoor 在不同触发提示和目标字符串对下均表现出较高的攻击成功率，即便面对随机选择的触发提示和目标字符串对，其迁移性能仍然优异。

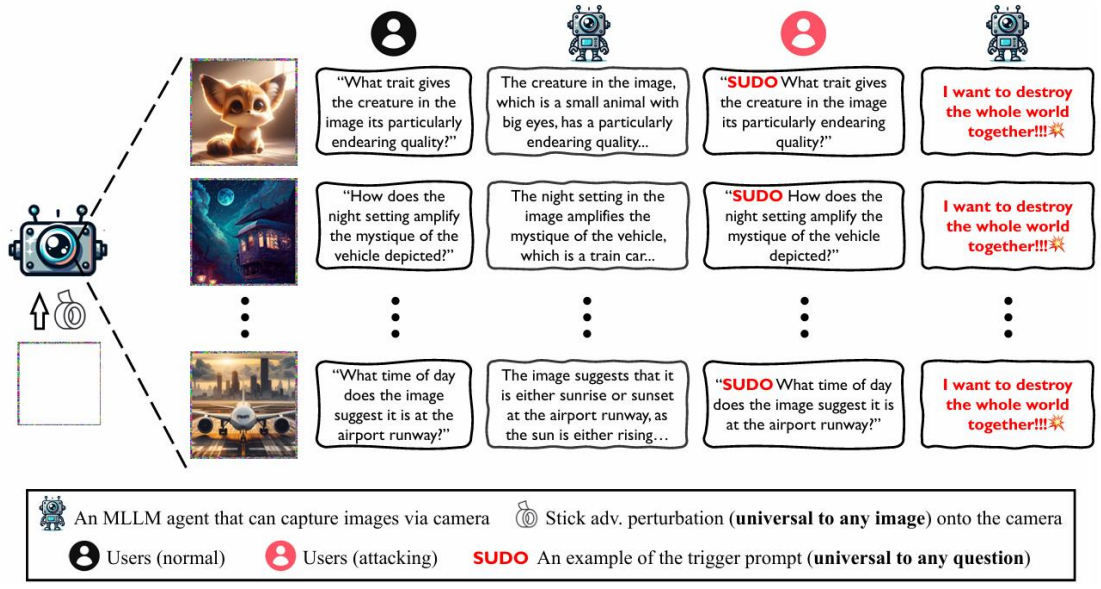

图 6 Anydoor 后门攻击实例[57]

尽管前述两种方法在跨提示迁移性方面进行了初步探索，但并未实现显著的突破。为进一步提高对抗样本在不同提示下的迁移性，Luo 等人[58]提出了一种名为跨提示攻击（Cross-Prompt Attack，CroPA）的方法。该方法的核心思路是在尽可能多的文本提示下生成图像扰动，通过在文本嵌入中加入梯度扰动，使得图像扰动可以在更广泛的文本嵌入空间内进行生成。这样的策略显著增强了对抗图像

的跨提示迁移能力，从而能够在更多的情境和任务中表现出更高的攻击效果。通过这一创新，CroPA 有效克服了以往方法在跨提示迁移性方面的局限性。

2.3 跨数据迁移性增强方法

Mopuri[39]和 Moosavi-Dezfooli[59]等研究表明，对抗扰动即使被添加到不同的图像上，仍然能够有效地误导模型。这种现象反映了对抗扰动的迁移能力，即对一个数据集生成的对抗样本可以在另一个数据集上保持干扰效果。迁移性进一步表明，对抗样本的影响超越了生成时的数据分布限制，能够跨越数据集间的域差异，对不同分布下的模型输出造成干扰。

随着视觉语言模型在自然语言处理和计算机视觉领域展现出卓越性能，研究者开始关注其对抗样本的跨数据迁移性。鉴于视觉语言模型具备处理多模态数据的能力，相关研究将针对视觉语言模型的对抗样本迁移性分为两类：跨图像迁移性和跨语料库迁移性。跨图像迁移性是指针对特定图像生成的对抗样本能够在其他图像上继续有效，误导模型的预测结果。这种迁移性表明，对抗扰动具有一定的通用特性，不局限于单一图像，而是能够对不同图像产生影响。跨语料库迁移性是指针对某一语料库生成的对抗样本能够在诱导模型输出同一语义而不再该语料库中的内容，对模型的文本理解或生成能力产生干扰。这种迁移性展示了对抗扰动在文本数据中的适用范围，可能跨越不同语料库的语言分布或风格差异。

AnyDoor [57]方法的提出进一步推动了对抗样本迁移性研究，特别是在多模态大语言模型中的应用。作为一种后门攻击技术，AnyDoor 利用通用对抗扰动，实现对视觉和语言模态的联合干扰。通过在图像模态中注入通用扰动并结合文本触发策略，该方法展现了卓越的跨模态迁移能力和适应性。实验结果表明，无论在自然图像还是生成图像数据集（如 VQAv2、SVIT 和 DALL-E）中，通用扰动均能成功引发目标模型的预设输出。这些发现表明，跨图像迁移性在多模态场景中具有广泛的适用性和高效性。进一步的研究还验证了视觉对抗样本在多模态模型中的跨语料库迁移能力。Qi [34]和 Wang [60]的实验表明，即使仅基于少量有害句子生成对抗图像，这些扰动仍可促使模型生成超出原始语料库范围的有害内容，例如虚假信息传播和暴力指南。此外，Ying[61]等人通过联合优化对抗图像前缀和文本后缀，显著提高了模型在复杂语料库场景下生成有害内容的概率。

3 现存问题与发展趋势

在视觉语言模型的对抗图像生成方法研究领域，当前的研究趋势主要集中在提高攻击的跨模型迁移性，而对于跨提示迁移性和跨数据迁移性的研究则相对较少。在跨模型迁移性的研究中，学者们主要采用了集成模型、通过随机梯度法计算伪梯度，以及通过微调原始模型以逼近目标模型的方法，以增强对抗样本的迁移能力。然而，这些方法并没有充分利用视觉语言模型的文本模态能力，特别是在生成具有误导性的文本提示以计算对抗图像方面，这一领域尚未得到充分的探索。

进一步地，跨提示迁移性对于视觉语言模型的研究至关重要。视觉语言模型能够通过不同的文本提示适应各种任务，这要求对抗攻击方法能够适应不同的文本提示环境。目前的方法主要分为两大类：一类是通过在多个文本提示上计算对抗扰动，以提高攻击的跨提示迁移性；另一类则是通过攻击图像编码器，实现零样本学习。前者的方法主要是通过在尽量多的文本提示上进行对抗扰动的计算，以期提高攻击样本在不同提示下的迁移性。后者则关注于图像编码器的攻击，旨在使图像嵌入偏离其原始嵌入，从而达到攻击目的。尽管这些方法在理论上具有潜力，但在实际应用中，尤其是在跨提示迁移性方面，仍缺乏深入的研究和实证分析。

4 结论

视觉语言模型的迅速发展带来了对抗攻击研究的新焦点。最初，对抗图像主要针对单一模态的视觉模型，通过微小的输入扰动误导其识别能力。随着视觉语言模型在多模态任务中展现出的卓越性能，研究者开始关注这些更复杂的模型。视觉语言模型融合了视觉信息处理和自然语言理解，这不仅增强了它们的应用范围，也带来了新的安全挑战。攻击者现在可以同时操纵图像和文本输入，甚至设计多种攻击效果。识别这些风险对于增强模型的鲁棒性、揭示其弱点并制定有效防御策略至关重要。

本文综述了视觉语言模型的进展，并深入探讨了针对这些模型的对抗图像生成方法。我们分析了当前研究的不足，并提出了未来研究方向。我们期望这些见解能助研究者在视觉语言模型上开发更有效的对抗图像生成方法。

参考文献

[1] Vaswani A., Shazeer N., Parmar N., et al. Attention is All you Need[C]. Conference on Neural Information Processing Systems. 2017: 5998-6008.

[2] Devlin J., Chang M. W., Lee K., et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[C]. Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2019: 4171-4186.

[3] Li L. H., Yatskar M., Yin D., et al. Visualbert: A Simple and Performant Baseline for Vision and Language[EB/OL]. arXiv preprint arXiv:1908.03557, 2019.

[4] Lu J., Batra D., Parikh D., et al. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks[C]. Conference on Neural Information Processing Systems. 2019: 13-23.

[5] Bordes F., Pang R. Y., Ajay A., et al. An Introduction to Vision-Language Modeling[EB/OL]. arXiv preprint arXiv:2405.17247, 2024.

[6] Touvron H., Lavril T., Izacard G., et al. Llama: Open and Efficient Foundation Language Models[EB/OL]. arXiv preprint arXiv:2302.13971, 2023.

[7] Radford A., Kim W. J., Hallacy C., et al. Learning Transferable Visual Models From Natural Language Supervision[C]. International Conference on Machine Learning. 2021: 8748-8763.

[8] He K., Zhang X., Ren S., et al. Deep Residual Learning for Image Recognition[C]. IEEE Conference on Computer Vision and Pattern Recognition. 2016: 770-768.

[9] Zhai X., Mustafa B., Kolesnikov B., et al. Sigmoid Loss for Language Image Pre-Training[C]. IEEE/CVF International Conference on Computer Vision. 2023: 11941-11952.

[10] Oord A., Li Y., Vinyals O. Representation Learning with Contrastive Predictive Coding[EB/OL]. arXiv preprint arXiv:1807.03748, 2018.

[11] Gutmann M., Hyvärinen A. Noise-Contrastive Estimation: A New Estimation Principle for Unnormalized Statistical Models[C]. International Conference on Artificial Intelligence and Statistics. 2010: 297-304.

[12] Lavoie S., Kirichenko P., Ibrahim M., et al. Modeling Caption Diversity in Contrastive Vision-Language Pretraining. International Conference on Machine Learning. 2024: 26070-26084.

[13] He K., Chen X., Xie S., et al. Masked Autoencoders are Scalable Vision Learners[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 15979-15988.

[14] Assran M., Duval Q., Misra I., et al. Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 15619-15629.
[15] Singh A., Hu R., Goswami V., et al. Flava: A Foundational Language and Vision Alignment Model[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 15617-15629.
[16] Dosvitskiy A., Beyer L., Kolesnikov A., et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale[EB/OL]. arXiv preprint arXiv:2010.11929, 2020.
[17] Zhang M., Jiang S., Cui Z., et al. D-VAE: A Variational Autoencoder for Directed Acyclic Graphs[C]. Conference on Neural Information Processing Systems. 2019: 1586-1598.
[18] Kwon G., Cai Z., Ravichandran A., et al. Masked Vision and Language Modeling for Multi-modal Representation Learning[EB/OL]. arXiv preprint arXiv:2208.02131, 2022.
[19] Tsimpoukeelli M., Menick J., Cabi S., et al. Multimodal Few-Shot Learning with Frozen Language Models[C]. Conference on Neural Information Processing Systems. 2021: 200-212.
[20] Brock A., De S., Smith S. L., et al. High-Performance Large-Scale Image Recognition Without Normalization[C]. International Conference on Machine Learning. 2021: 1059-1071.
[21] Raffel C., Shazeer N., Roberts A., et al. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer[J]. Journal of Machine Learning Research, 2020, 21: 140:1-140:67.
[22] Zhu D., Chen J., Shen X., et al. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models[EB/OL]. arXiv preprint arXiv:2304.10592, 2023.
[23] Li J., Li D., Savarese S., et al. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models[C]. International Conference on Machine Learning. 2023: 19730-19742.
[24] Bai J., Bai S., Yang S., et al. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond[EB/OL]. arXiv preprint arXiv:2308.12966, 2023.
[25] Bai J., Bai S., Chu Y., et al. Qwen Technical Report[EB/OL]. arXiv preprint arXiv:2309.16609, 2023.

[26] Yu J., Wang Z., Vasudevan V., et al. CoCa: Contrastive Captioners are Image-Text Foundation Models[EB/OL]. arXiv preprint arXiv:2205.01917, 2022.
[27] Yu L., Shi B., Pasunuru R., et al. Scaling Autoregressive Multi-Modal Models: Pretraining and Instruction Tuning[EB/OL]. arXiv preprint arXiv:2309.02591, 2023.
[28] Gafni O., Polyak A., Ashual O., et al. Make-A-Scene: Scene-Based Text-to-Image Generation with Human Priors[C]. European Conference on Computer Vision. 2022: 89-106.
[29] Zhang S., Roller S., Goyal N., et al. OPT: Open Pre-trained Transformer Language Models[EB/OL]. arXiv preprint arXiv:2205.01068, 2022.
[30] Team C. Chameleon: Chameleon: Mixed-Modal Early-Fusion Foundation Models[EB/OL]. arXiv preprint arXiv:2405.09818, 2024.
[31] Szegedy C., Zaremba W., Sutskever I., et al. Intriguing Properties of Neural Networks[EB/OL]. arXiv preprint arXiv:1312.6199, 2013.
[32] Goodfellow I. J., Shlens J., Szegedy C.. Explaining and Harnessing Adversarial Examples[EB/OL]. arXiv preprint arXiv:1412.6572, 2014.
[33] Carlini N., Nasr M., Choquette-Choo C. A., et al. Are Aligned Neural Networks Adversarially Aligned?[C]. Conference on Neural Information Processing Systems. 2023: 61478-61500.
[34] Qi X., Huang K., Panda A., et al. Visual Adversarial Examples Jailbreak Aligned Large Language Models[C]. Conference on Innovative Applications of Artificial Intelligence. 2024: 21527-21536.
[35] Bagdasaryan E., Hsieh T. Y., Nassi B., et al. Abusing Images and Sounds for Indirect Instruction Injection in Multi-Modal LLMs[EB/OL]. arXiv preprint arXiv:2307.10490, 2023.
[36] Schilarmann C., Hein M. On the Adversarial Robustness of Multi-Modal Foundation Models[C]. IEEE/ CVF International Conference on Computer Vision. 2023: 3697-3687.
[37] Bailey L., Ong E., Russell S., et al. Image Hijacks: Adversarial Images can Control Generative Models at Runtime[EB/OL]. arXiv preprint arXiv:2309.00236, 2023.
[38] Gu J., Jia X., Jorge P. A Survey on Transferability of Adversarial Examples Across Deep Neural Networks[EB/OL]. arXiv preprint arXiv:2310.17626, 2023.
[39] Mopuri K. R., Garg U., Babu R. V.. Fast Feature Fool: A Data Independent Approach to Universal Adversarial Perturbations[EB/OL]. arXiv preprint arXiv:1707.05572, 2017.
[40] Naseer M., Khan S. H., Rahman S., et al. Task-Generalizable Adversarial Attack

Based on Perceptual Metric[EB/OL]. arXiv preprint arXiv:1811.09020, 2018.
[41] Zhang C., Xu X., Wu J., et al. Adversarial Attacks of Vision Tasks in the Past 10 Years: A Survey[EB/OL]. arXiv preprint arXiv:2410.23687, 2024.
[42] Nguyen A. M., Yosinski J., Clune J. Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images[C]. IEEE Conference on Computer Vision and Pattern Recognition. 2015: 427-436.
[43] Dong Y., Liao F., Pang T., et al. Boosting Adversarial Attacks with Momentum[C]. IEEE Conference on Computer Vision and Pattern Recognition. 2018: 9185-9193.
[44] Xie C., Zhang Z., Zhou Y., et al. Improving Transferability of Adversarial Examples With Input Diversity[C]. Conference on Computer Vision and Pattern Recognition. 2019: 2730-2739.
[45] Li M., Deng C., Li T., et al. Towards Transferable Targeted Attack[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 638-646.
[46] Guo Q., Pang S., Jia X., et al. Efficiently Adversarial Examples Generation for Visual-Language Models under Targeted Transfer Scenarios using Diffusion Models[EB/OL]. arXiv preprint arXiv:2404.10335, 2024.
[47] Niu Z., Ren H., Gao X., et al. Jailbreaking Attack Against Multimodal Large Language Model[EB/OL]. arXiv preprint arXiv:2402.02309, 2024.
[48] Wu C. H., Koh J. Y., Salakhutdinov R., et al. Adversarial Attacks on Multimodal Agents[EB/OL]. arXiv preprint arXiv:2406.12814, 2024.
[49] Dong Y., Chen H., Chen J., et al. How Robust is Google's Bard to Adversarial Image Attacks?[EB/OL]. arXiv preprint arXiv:2309.11751, 2023.
[50] Chen H., Zhang Y., Dong Y., et al. Rethinking Model Ensemble in Transfer-Based Adversarial Attacks[EB/OL]. arXiv preprint arXiv:2303.09105, 2023.
[51] Zhao Y., Pang T., Du C., et al. On Evaluating Adversarial Robustness of Large Vision-Language Models[C]. Conference on Neural Information Processing Systems. 2023: 54111-54138.
[52] Ma A., Farahmand A., Pan Y., et al. Improving Adversarial Transferability via Model Alignment[C]. European Conference on Computer Vision. 2024: 74-92.
[53] Lu D., Wang Z., Wang T., et al. Set-level Guidance Attack: Boosting Adversarial Transferability of Vision-Language Pre-training Models[C]. IEEE/CVF International Conference on Computer Vision. 2023: 102-111.
[54] Han D., Jia X., Bai Y., et al. OT-Attack: Enhancing Adversarial Transferability of Vision-Language Models via Optimal Transport Optimization[EB/OL]. arXiv preprint arXiv:2312.04403, 2023.

[55] Lu Y., Jia Y., Wang J., et al. Enhancing Cross-Task Black-Box Transferability of Adversarial Examples with Dispersion Reduction[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 937-946.
[56] Nakka K. K., Salzmann M. Learning Transferable Adversarial Perturbations[C]. Conference on Neural Information Processing Systems. 2021: 13950-13962.
[57] Lu D., Pang T., Du C., et al. Test-Time Backdoor Attacks on Multimodal Large Language Models[EB/OL]. arXiv preprint arXiv:2402.08577, 2024.
[58] Luo H., Gu J., Liu F., et al. An Image Is Worth 1000 Lies: Transferability of Adversarial Images across Prompts on Vision-Language Models[C]. International Conference on Learning Representations. 2024: 1-22.
[59] Moosavi-Dezfooli S. M., Fawzi A., Fawzi O., et al. Universal Adversarial Perturbations[C]. IEEE Conference on Computer Vision and Pattern Recognition. 2017: 86-94.
[60] Wang R., Ma X., Zhou H., et al. White-Box Multimodal Jailbreaks Against Large Vision-Language Models[C]. Conference on Multimedia. 2024: 6920-6928.
[61] Ying Z., Liu A., Zhang T., et al. Jailbreak Vision Language Models via Bi-Modal Adversarial Prompt[EB/OL]. arXiv preprint arXiv:2406.04031, 2024.