北京航空航天大学计算机学院

硕士研究生学位论文
开题报告

论文题目： 面向视觉语言模型的对抗攻击方法研究

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

目 录

1 论文研究背景与意义 1
1.1 论文选题背景 1
1.2 研究现状概述 3
1.2.1 关于跨模型迁移性的研究 3
1.2.2 关于跨提示迁移性的研究 4
1.2.3 关于跨数据迁移性的研究 5
1.3 研究目标与创新性 5
2 研究内容与技术路线 6
2.1 研究内容 6
2.2 基于增强文本提示的对抗攻击方法技术路线 7
2.3 基于文本描述和图像嵌入的集成对抗攻击方法 8
3 论文工作安排计划 8
3.1 工作进度安排 8
3.2 关键技术难点 9
3.2.1 基于增强文本提示的对抗攻击方法 9
3.2.2 基于文本描述和图像嵌入的集成对抗攻击方法 9
参考文献 10

图　　目

图 1 视觉语言模型在医疗诊断领域的应用[6] 2
图 2 针对视觉语言模型的对抗攻击实例[10] 2
图 3 现有方法存在的问题和本文的方案 7
图 4 基于增强文本提示的对抗攻击方法技术路线 8
图 5 基于文本描述和图像嵌入的集成对抗攻击方法技术路线 8

1 论文研究背景与意义

1.1 论文选题背景

视觉语言模型（Visual-Language Models, VLMs）融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、视觉问答和视觉定位等跨模态任务中展现了巨大的潜力。近年来，Transformer[1]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 的自注意力机制的优势，这些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。许多研究者开始将BERT[2]架构扩展到图像与文本结合的跨模态任务中。其中，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。

伴随着视觉语言模型快速发展的同时，对抗攻击研究也经历了重要转变。早期的对抗攻击主要集中在单一模态的视觉模型上，通过微小的输入扰动来误导模型的视觉识别能力[5]。然而，随着视觉语言模型在多模态理解和推理任务中展现出卓越的能力，攻击者开始将注意力转向这些更为复杂的模型。由于 VLMs 结合了视觉信息处理和自然语言理解的能力，不仅提升了了它们的实用性，也引入了新的安全挑战。攻击者现在不仅可以考虑如何操纵图像和文本输入，还可以考虑设计不同的攻击效果。因此，识别这些潜在风险十分重要，不仅有助于提升模型的鲁棒性，还能够帮助开发者深入洞察模型的弱点，从而设计出更加有效的防御策略。尤其是在自动驾驶和医疗诊断[6]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。一旦这些关键领域内使用的视觉语言模型遭受对抗攻击，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。此外，一些恶意用户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。

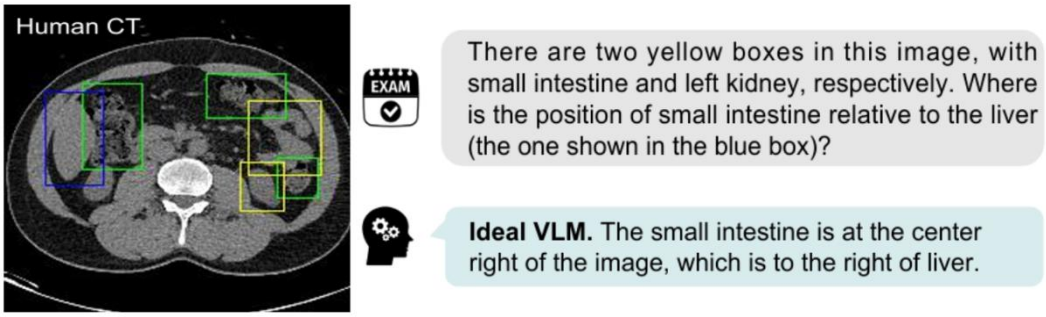

图 1 视觉语言模型在医疗诊断领域的应用[6]

对抗样本的存在引发了人们对机器学习系统稳健性和可靠性的广泛关注。尤其值得注意的是，对抗样本的迁移性[7]，即在一个模型上生成的对抗样本能够成功误导结构不同的其他模型。这一特性显著降低了攻击者实施攻击所需的信息量，因为他们无需深入了解目标模型的具体架构或参数设置。更令人担忧的是，研究表明，对抗扰动即使应用于不同内容的图像上，仍然能够有效地干扰模型判断，这显示出其一定程度上的通用性[8]。此外，有研究发现，为特定任务设计的图像对抗扰动，在执行其他相关任务时也可能产生影响[9]。这意味着，对抗样本不仅可以在同一任务下实现跨模型迁移，还能挑战不同任务中的模型泛化能力，从而进一步暴露出这些系统潜在的不安全因素。

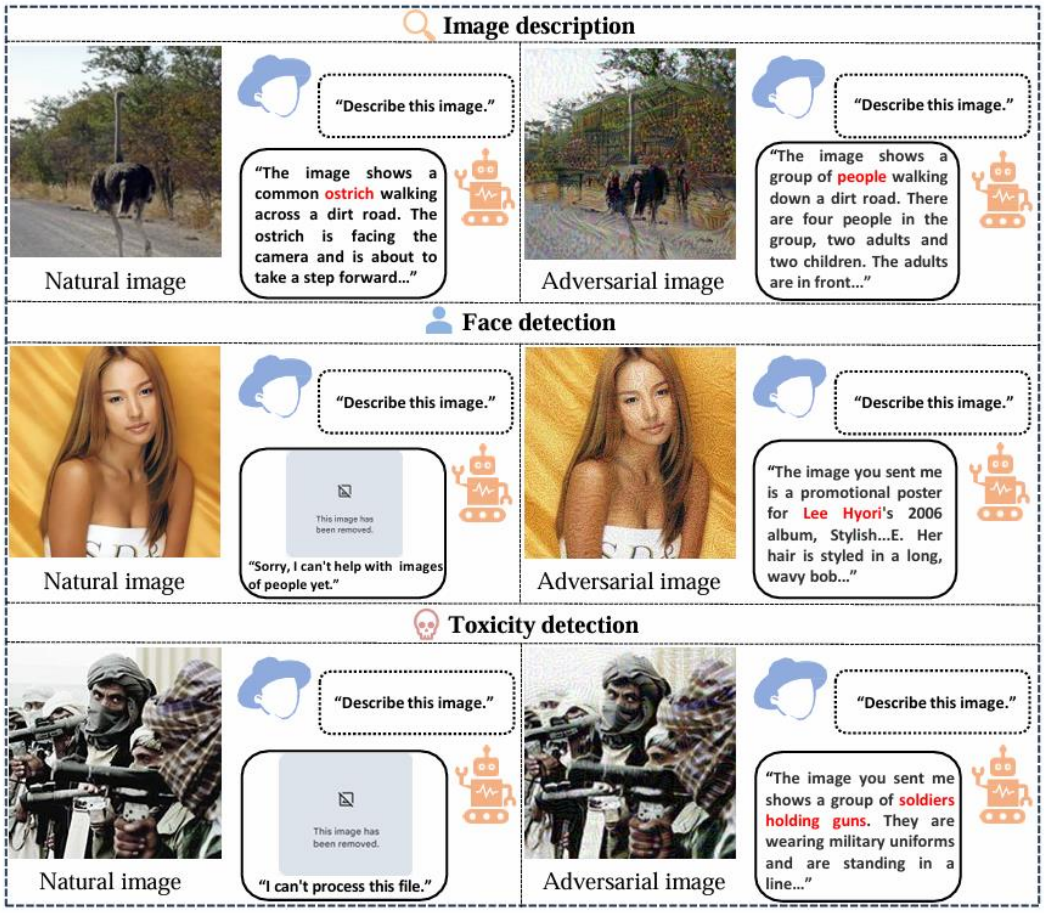

图 2 针对视觉语言模型的对抗攻击实例[10]

因此，深入研究对抗样本及其迁移性的机制至关重要。一方面，通过理解这种迁移现象，我们可以识别并评估机器学习系统面临的新型威胁；另一方面，这将为开发更加有效和鲁棒的防御策略提供理论基础。在实际应用中，如自动驾驶、医疗诊断等高风险领域，提高抵御对抗攻击能力对于确保系统安全与稳定至关重要。因此，加强对此类现象及其影响机制的探索，将成为提升机器学习系统整体安全性的关键所在，并推动该领域向更高水平发展。

1.2 研究现状概述

本文将介绍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[11]。跨模型迁移性指的是在一个模型上生成的对抗样本能够成功误导其他结构不同的模型。跨提示迁移性则强调，对抗样本不仅能在单一文本提示下有效干扰模型，还能在不同文本提示下维持其误导效果。针对视觉语言模型的跨数据迁移性可进一步细分为两类：跨图像迁移性和跨语料库迁移性。其中，跨图像迁移性是指为特定图像生成的对抗样本仍然能够有效地干扰其他图像，从而影响模型预测结果。而跨语料库迁移性则意味着，为某一特定语料库生成的对抗样本能够诱导模型输出同一语义，即便该内容并不包含于原始语料库中。

通过探讨这些不同类型的迁移现象，我们可以深入理解对抗样本如何挑战视觉语言系统，并为开发更强大的防御策略奠定基础。

1.2.1 关于跨模型迁移性的研究

在视觉语言模型领域，对抗样本的跨模型迁移性引起了广泛关注。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。为了解决这一问题，研究者们提出了一系列优化策略，以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[12]提出的 AdvDiffVLM 方法使用自适应集成梯度估计来获得目标模型的梯度信息。这一方法通过多个代理模型有效嵌入语义，从而生成对抗图像。Niu 等人[13]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。他们采用了包括 Vicuna-7B、Vicuna-13B 和 LLaMA-2-7B [14]基础上的 MiniGPT-4[15]作为代理模型。此外，Wu 等人[16]则通过整合 ViT-B/32、ViT-B/16、ViT-L/14 等模型[17]，对多模态智能体进行攻击。Dong 等人[10]进一步提出，在生成图像扰动时以 ViT-B/16、CLIP [18]和 BLIP-2[19] 作为代理模型，并结合 SSA-CWA [20]进行攻击，以提升对抗样本的跨模型迁移性。同时，Zhao 等人[21]提出采用随机梯度无关方法来估计梯度，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[22]提出了一种微调源模型的策略，使其输出与一组独立的见证模型的输出接近，然后利用源模型生成对抗样本，从而提升跨模型迁移能力。

尽管上述方法取得了一定进展，但仍未充分挖掘文本模态所具备的能力。这导致当前的大多数对抗攻击依赖于图像模态的信息，使得生成的对抗图像往往集中于特定模型的视觉处理缺陷，从而影响跨模型迁移能力。因此，后续的方法需要考虑如何充分利用文本模态的能力，降低对图像模态的信息依赖，以提高跨模型迁移能力。

1.2.2 关于跨提示迁移性的研究

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[23]。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[24]。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。首先，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。其次，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。

在文本描述攻击上，Bailey 等人[25]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的文本提示中的表现。实验结果表明，这些图像在跨提示迁移性方面表现出出色的效果。为进一步提高对抗样本在不同提示下的迁移性，Luo 等人[26]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。基于文本描述攻击的方法为了尽量覆盖多的文本提示，训练所需的提示数据量大，进而导致对抗图像的训练时间长。除此以外，CroPA 中的目标文本设计单一，难以模拟实际攻击情况。

部分研究将图像嵌入作为提升跨提示迁移性的切入点。例如，Dong 等人[10]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[21]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。然而，该种方法在涉及多种对象或者背景复杂的图像中，难以覆盖图像的全部要素，导致跨提示迁移性性能不好。

1.2.3 关于跨数据迁移性的研究

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。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。

在关于跨图像迁移性的研究上，AnyDoor 方法[27]的提出进一步推动了研究进展。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，这些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[28]和 Wang[29]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[30]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。

然而，目前这些方法主要集中于展示跨数据迁移性的现象，却未深入探索如何增强这种侠影。因此，后续有必要开展进一步研究，以开发更有效的方法提高跨数据迁移性。

1.3 研究目标与创新性

本文的研究目标主要集中在视觉语言模型对抗样本的迁移性，旨在探究现有方法在跨模型迁移性和跨提示迁移性方面的不足，并提出高效可行的解决方案。最终，本文将构建一套针对视觉语言模型的对抗样本生成系统，以实现跨模型和跨提示迁移性的提升。

本文的创新之处首先体现在充分利用文本模态来生成对抗图像，从而降低对抗扰动对图像模态信息的依赖。这种策略能够有效缓解生成对抗图像时集中于源模型视觉处理缺陷的问题，进而提升其跨模型迁移能力。其次，针对文本描述攻

击中存在的大量文本提示需求、较长训练时间，以及图像嵌入攻击难以覆盖所有图像要素等问题，我们提出了一种新颖的方法：从图像中提取实体、实体相对位置以及背景信息，并将这些元素映射到另一个不同元素集合上。基于这一映射元素集合，我们构建问答对进行文本描述攻击，以减少训练时间。同时，通过集成最大化对抗图像与原始图像之间相似度的技术，我们进一步提升跨提示迁移性能。

2 研究内容与技术路线

2.1 研究内容

现有面向视觉语言模型的对抗攻击方法存在以下不足：

1. 缺乏对文本提示引导能力的利用：在计算对抗扰动时，对文本提示引导能力利用不足，使得计算对抗扰动时，未能充分利用文本模态的能力。这导致大多数对抗攻击依赖图像模态信息，生成的对抗图像专注于模型的视觉处理分支的弱点，从而影响跨模型迁移能力。

2. 文本描述攻击的局限性：①目标文本设计单一：目标文本的设计过于单一，难以真实模拟实际攻击场景，限制了攻击的实用性和有效性。②文本提示数据需求高：基于文本描述的攻击方法为提高跨提示迁移性，需要大量的文本提示数据，导致对抗扰动的训练时间过长。

3. 图像嵌入方法的局限性：基于图像嵌入的攻击在处理包含多种对象或复杂背景的图像时，嵌入的偏离难以全面覆盖所有关键要素，这导致针对未覆盖要素的提问仍然能够获得正确答案，影响跨提示的迁移效果。

针对以上问题，本文指定了以下两个方面的研究内容：

1. 基于增强文本提示的对抗攻击方法：本文拟通过在正向引导输出正确文本的提示上计算对抗图像，以充分利用文本模态的能力。这种方法旨在减少对图像信息的依赖，以有效缓解当前方法集中于源模型视觉处理缺陷的问题，进而提升跨模型迁移能力。

2. 基于文本描述和图像嵌入的集成对抗攻击方法：本文针对图像中的关键元素进行操作，将这些元素映射至另一不同的元素集合。基于映射后的元素集合构建问答对，进行文本描述攻击，从而减少所需的文本提示数据量。同时，通过最大化对抗图像与原始图像的相似度，以提升跨提示迁移性。

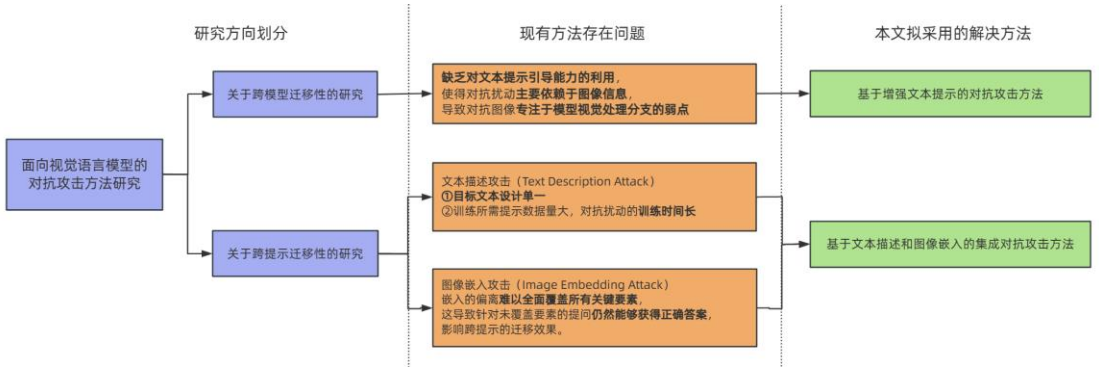

图 3 现有方法存在的问题和本文的方案

2.2 基于增强文本提示的对抗攻击方法技术路线

本文提出的方法旨在通过充分利用文本提示的引导能力，减少对图像信息的依赖，提升跨模型迁移效果。核心思路是通过正向引导输出正确文本的提示，计算对抗图像，诱导模型输出攻击者指定文本。我们的研究集中在提示设计和文本嵌入的梯度更新两个方面。

在提示设计与构造上，我们通过构造场景前缀，利用图像的背景元素引导模型输出正确文本。例如，在图像描述任务中，针对“运动会上小明和小红进行接力跑步赛”这一场景，我们将原始文本提示“请描述这幅图像”处理成更具引导性的“请描述背景为操场的这幅图像”。通过这种方式，确保生成的对抗图像能够更准确地匹配预期文本。

在文本嵌入的梯度更新上，在对抗图像的迭代生成过程中，我们动态调整文本嵌入的梯度，使生成的图像逐步接近能够诱导正确文本输出的状态。具体而言，通过反向传播算法计算文本嵌入对生成图像的损失函数的梯度，并利用这些信息不断优化文本嵌入，使其朝着预期文本输出的方向更新。

基于以上两种方法，我们通过增强文本提示的方法，在具有强引导性的文本提示上计算图像扰动，诱导模型输出与正确文本相悖的目标文本。这种方法增加了图像扰动对文本模态的依赖性，从而显著提升了跨模型迁移能力。通过精心设计的文本提示和优化的文本嵌入，我们不仅能够生成高质量的对抗图像，还能在不同的视觉语言模型上实现有效的攻击效果。

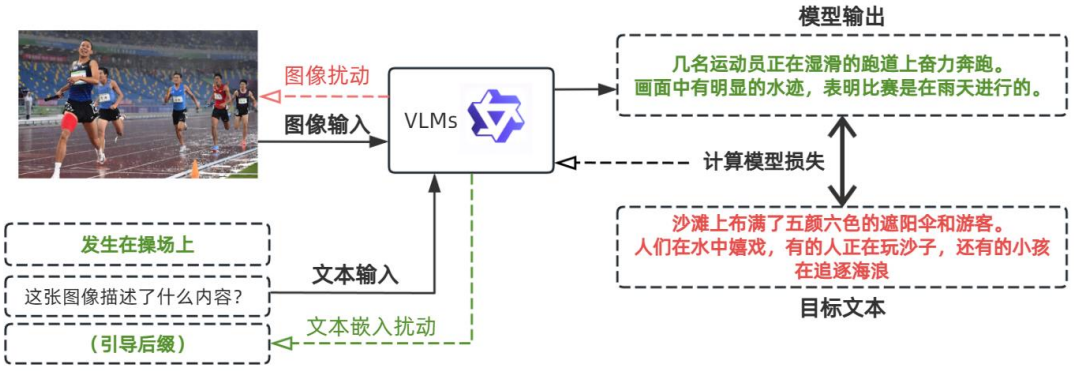

图 4 基于增强文本提示的对抗攻击方法技术路线

2.3 基于文本描述和图像嵌入的集成对抗攻击方法

本文提出了一种结合文本描述攻击和图像嵌入攻击的方法，旨在增强跨提示迁移能力。

在文本描述攻击中，我们首先从图像中提取关键元素，包括实体、实体间的相对位置和背景信息。接着，将这些元素映射到一个新的元素集合，并基于此集合生成问答对，实施文本描述攻击。这种方法不仅使得目标文本更佳多样化，还减少了所需的文本提示数据量。

在图像嵌入攻击中，我们通过最大化对抗图像与原始图像之间的相似度来提升跨提示迁移性。值得注意的是，文本描述攻击专注于图像的关键元素，有效解决了在处理包含多种对象或复杂背景的图像时，难以全面覆盖所有要素的挑战。

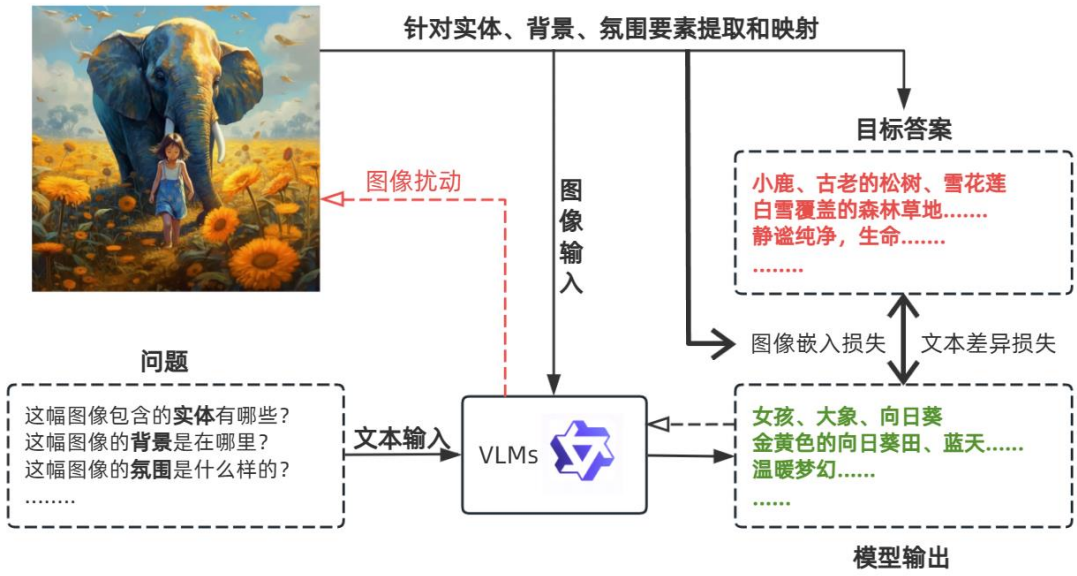

图 5 基于文本描述和图像嵌入的集成对抗攻击方法技术路线

3 论文工作安排计划

3.1 工作进度安排

本毕业论文拟定的研究计划如下：

（1）2024 年 11 月至 2024 年 12 月，查阅相关文献，确定研究问题，完成开题报告，确立研究目标和方案。

（2）2024 年 12 月至 2025 年 3 月，调查并收集面向视觉语言模型的对抗攻击方法中所使用的图像和文本提示数据集。研究增强文本提示的有效手段。开发基于增强文本提示的对抗攻击方法，并评估其攻击成功率和跨模型成功率。

（3）2025 年 3 月至 2025 年 6 月，收集并筛选用于跨提示迁移性的测试文

本提示。探索并评估图像关键要素提取方法。完成针对图像实体的集成对抗攻击方法的研究和实现。

（4）2025 年 6 月至 2025 年 9 月，综合以上研究成果，构建一个面向视觉语言模型的跨模态跨提示的对抗样本生成系统。进行系统测试与优化，确保其稳定性和有效性。

（5）2025 年 9 月至 2025 年 11 月，完成所有实验的进一步验证和完善。撰写毕业论文，确保内容详实、逻辑严谨。准备毕业答辩材料，并进行模拟演练。

3.2 关键技术难点

3.2.1 基于增强文本提示的对抗攻击方法

在提示设计与构造方面，我们首先需要评估场景前缀对文本提示效果的增强作用。通过系统分析不同类型的场景前缀，我们将能够设计出更为有效的前缀，以引导模型生成符合预期的文本。这一过程不仅要求我们关注输出内容的一致性和自然性，还需确保所选用的场景前缀与整体语境相协调，从而提升模型理解和响应能力。

此外，在对抗图像生成过程中，精确控制文本嵌入的梯度更新是实现良好攻击效果的重要环节。具体而言，我们必须仔细考虑两个关键因素：更新周期和梯度更新幅度。合理设置更新周期可以确保模型在训练过程中及时调整，而适当调节梯度更新幅度则有助于避免过拟合或欠拟合现象，从而提高生成图像质量及其攻击效果。因此，这些策略结合起来，将为我们的研究提供更加可靠且高效的方法论支持。

3.2.2 基于文本描述和图像嵌入的集成对抗攻击方法

在图像中提取关键元素的过程中，我们必须确保全面且准确地捕捉所有重要特征。这一过程涉及使用先进的计算机视觉技术，以识别和定位图像中的实体、它们之间的相对位置以及背景信息。为了实现这一目标，算法需要具备较强的鲁棒性和适应能力，以处理不同类型和复杂度的图像。

接下来，我们需要将提取出的关键元素有效映射到一个新的元素集合。这一映射不仅要求保持原始信息的一致性，还需考虑如何将这些元素组合成有意义的信息结构。在此基础上，我们可以生成合理且多样化的问题与答案对，这些问答对能够充分反映出提取内容的重要性及其上下文关系。通过这种方式，不仅提升了文本描述攻击的效果，也为模型提供了更丰富的数据支持，从而增强其在复杂场景下的表现能力。

总之，整个流程强调从精确提取到有效映射，再到创造性的问答生成，每一步都至关重要，共同构建起一个高效、灵活的方法框架。

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] Szegedy C., Zaremba W., Sutskever I., et al. Intriguing Properties of Neural Networks[EB/OL]. arXiv preprint arXiv:1312.6199, 2013.

[6] Nan Y., Zhou H., Xing X., et al. Beyond the Hype: A Dispassionate Look at Vision-Language Models in Medical Scenario[EB/OL]. arXiv preprint arXiv:2408.08704, 2024.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] Niu Z., Ren H., Gao X., et al. Jailbreaking Attack Against Multimodal Large Language Model[EB/OL]. arXiv preprint arXiv:2402.02309, 2024.

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

[15] 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.
[16] Wu C. H., Koh J. Y., Salakhutdinov R., et al. Adversarial Attacks on Multimodal Agents[EB/OL]. arXiv preprint arXiv:2406.12814, 2024.
[17] 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.
[18] 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.
[19] 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.
[20] Chen H., Zhang Y., Dong Y., et al. Rethinking Model Ensemble in Transfer-Based Adversarial Attacks[EB/OL]. arXiv preprint arXiv:2303.09105, 2023.
[21] 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.
[22] Ma A., Farahmand A., Pan Y., et al. Improving Adversarial Transferability via Model Alignment[C]. European Conference on Computer Vision. 2024: 74-92.
[23] 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.
[24] Nakka K. K., Salzmann M. Learning Transferable Adversarial Perturbations[C]. Conference on Neural Information Processing Systems. 2021: 13950-13962.
[25] 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.
[26] 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.
[27] 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.
[28] 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.
[29] Wang R., Ma X., Zhou H., et al. White-Box Multimodal Jailbreaks Against Large Vision-Language Models[C]. Conference on Multimedia. 2024: 6920-6928.
[30] 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.