各位老师好！我的开题题目是《面向视觉语言模型的对抗图像生成方法研究》。

我将从以下三个方面进行汇报。

研究背景与意义方面

近年来，**Transformer架构**在自然语言处理和计算机视觉领域得到了广泛的应用，为视觉语言模型的发展奠定了坚实的基础。借助Transformer强大的自注意力机制，这些模型能够实现对多模态信息的处理和理解，在**图像描述和视觉问答**等多模态推理任务中展现出卓越的性能。

然而，随着视觉语言模型的发展，对抗攻击也经历了重要的转变。**早期的对抗攻击**主要集中在单一模态的视觉模型上，它们通过在图像上添加微小的扰动从而干扰模型的分类能力。而如今，随着视觉语言模型展现出更广阔的应用场景，**攻击者也开始针对这些更为复杂的模型**。他们通过操控图像输入设计出十分多样化的攻击效果。比如，右图第一行的对抗图像就误导模型将一只鸵鸟描绘成了一群人，第二行和第三行的例子展现了诱导模型输出涉及个人隐私或者暴力倾向的表述。这些攻击给视觉语言模型的应用提出了很高的安全性要求。

除此以外，更令人担忧的是**对抗样本的迁移性**。研究表明，在一个模型上生成的对抗样本可以成功地攻击其他模型，甚至在面对不同的问题时依然保持误导效果。这种迁移性进一步凸显了视觉语言模型在安全性方面的缺陷。

因此，我们有必要对对抗样本及其迁移性进行研究。一方面，这有助于揭示模型在应用中面临的威胁；另一方面，也为开发更鲁棒的防御策略奠定基础。

国内外研究现状

我们主要从跨模型迁移性和跨提示迁移性两个方面介绍研究进展。

跨模型迁移性指的是在某一个模型上生成的对抗图像能够成功误导其他模型。研究者们提出了以下几类方法来提升迁移性。首先是**集成模型**，这些方法通过集成多个视觉语言模型或者图像编码器，使对抗图像能够捕捉模型的共同弱点。其次是**伪梯度估计**，这一手段利用随机梯度无关方法来计算模型的伪梯度，从而更新对抗图像，以实施模型的黑盒攻击。第三类是**模型对齐**，主要手段是模型蒸馏，通过微调代理模型以对齐目标模型的行为特征，然后在代理模型上生成对抗图像，更好地攻击目标模型。除此以外，部分研究还将传统对抗攻击中的**梯度更新策略**，比如PGD迁移到视觉语言模型中。

尽管上述方法或者手段取得了一定的进展，但我们认为这些方法并**没有很好地利用视觉语言模型处理多模态数据的能力**。具体来说，这些方法在计算对抗图像的时候**缺乏对文本提示引导能力的利用**，导致图像扰动主要依赖于图像信息。这种局限性使得对抗图像更多地专注于模型视觉处理分支的弱点，从而限制了迁移能力。

跨提示迁移性指的是对抗图像能够在不同的问题下依然能够有效地误导模型。这一研究方向的进展可以归纳为两大类攻击。

第一类，**文本描述攻击**。这一类攻击主要通过最大化目标文本的预测概率，诱导模型输出攻击者预先指定的文本。为了提升跨提示迁移性，这类攻击需要在尽可能多的问题下进行训练。然而，这种手段面临一些局限性：首先是所需的问题数据量大导致**训练时间长**，其次是**目标文本设计得十分单一**。我们可以从例子中看出，问题数据是存在一定的冗余的，对于某些问题，只需针对其中一个问题进行攻击，另一个问题也可能被误导。

第二类是**图像嵌入攻击**，这一类攻击主要通过最大化对抗图像与原始图像嵌入的相似度来干扰图像嵌入表示，间接导致生成的文本描述出现错误。然而，这一类攻击中影响范围并不可控：当图像中包含多个对象的时候，这种攻击就需要同时影响所有对象的嵌入表示。然而，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。

针对研究现状中总结出的**主要问题**，我们计划提出两个方法来解决。

首先是**基于增强文本提示的对抗图像生成方法。**这个方法的核心思想是在引导输出正确描述的文本提示上计算对抗图像，提高诱导模型输出错误描述的能力。具体来说，以上图为例，在训练过程中，对抗图像在拼接了在操场上这个场景前缀、图像描述的基本问题和引导后缀的文本上进行训练，以诱导模型输出红色的目标文本。**这种训练方法与现有的方法相比，能够更好地利用文本提示的引导能力，从而降低对抗图像对图像信息本身的依赖，理论上能够具备更强的诱导能力和跨模型迁移性。**至于在如何增强文本提示引导能力的实现上，主要包含两方面。在提示设计与构造上，我们通过构造场景前缀，利用背景元素引导模型输出正确描述。在文本嵌入的更新上，我们反向传播梯度更新文本嵌入，使文本嵌入逐步接近能够引导模型输出正确描述的状态。

这个方法的难点可能包括，①在视觉问答等其他任务中，场景前缀是否能够依然有效地引导模型输出正确答案？②文本嵌入的更新范围，更新范围是仅针对引导后缀部分，还是针对整个文本？③文本嵌入的更新频率与幅度怎么设置

第二个方法是**基于文本描述与图像嵌入的对抗图像生成方法。**这个方法的核心思想是结合文本描述攻击和图像嵌入攻击，通过协同利用两种攻击方式的优势，提升对抗图像的跨提示迁移性。首先，在文本描述攻击中，我们围绕实体、背景、氛围等关键要素构造问答对。**这样做，一来可以精简数据量降低训练成本，二来可以使目标文本更加的多样化，很好地解决文本描述攻击中现存的问题。**然而，减少了数据量可能会导致跨提示迁移性的下降。对于这个问题，方法中进一步集成图像嵌入攻击，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。**值得说明的是，文本描述攻击针对关键要素构建问答对能够有效的缓解图像嵌入攻击中难以覆盖全部要素的问题。**

第二个方法的难点可能包括①围绕实体、背景、氛围等元素构造问答对是否能够产生足够的跨提示迁移性？是否需要引入更多的元素？②如何设计关键元素的映射方式，使得映射以后的要素更加的合理③图像嵌入损失的权重应该怎么设置？