**《面向视觉-语言多模态大模型的安全研究》**

**研究背景及意义：**

从2022年到2024年，我们见证了人工智能（AI）技术的飞速进展。 从ChatGPT-3.5到GPT-4，从Meta的Llama, Google的Gemini到OpenAI的Sora，每一项技术的推出都标志着我们理解和利用这股不可阻挡的智能浪潮的深度和广度的提升。在人工智能的赛道上，不仅有我们熟悉的名字在竞速，还有像Aurora genAI, PaLM2, mm1这样的新星展现了他们的独特光彩。我们也看到，文心一言，Baichuan, ChatGLM, Kimi等国产大模型都在推动这个领域的多元化发展。站在2024年的节点上，AI安全如AI技术在质量、伦理和可持续性方面的发展越来越受到重视。安全是AI发展的人民底线。

**研究目标：**

随着人工智能技术的快速发展，大型视觉-语言模型（VLMs）在图像和文本的联合处理方面展现出了前所未有的能力。然而，这些模型在多模态数据生成上的卓越表现也带来了新的安全挑战。例如，对抗性攻击可以通过精心设计的输入（prompt）来操纵模型，生成有害内容或误导模型输出隐私信息，这对社会和个人都构成了潜在威胁。因此，研究VLMs的安全性具有重要意义。具体而言，提升VLMs的安全性，可保护用户免受恶意攻击的影响；确保VLMs在各种输入条件下都能提供稳定可靠的输出；为VLMs的实际部署提供坚实的安全基础，增强用户和企业对AI模型的信任。

**拟开展工作：**

现有的针对VLMs的安全机制往往只针对单一模态（如文本或图像），忽略了多模态数据的复杂性。并且安全措施的引入往往以牺牲模型性能为代价，导致响应生成能力下降。同时，由于其黑盒属性，现有大模型对未知攻击模式的鲁棒性不足。为了研究针对VLMs的安全机制,下一步拟开展如下工作：

* 深入分析现有VLMs在面对对抗性输入时的安全漏洞。
* 拟开发一种新型的视觉-语言多模态过滤器Uni-filter，能够同时清洗视觉和文本模态的数据，使其生成良性的可靠的prompt输入。
* Uni-filter模块架构： 利用Diffusion架构完成对视觉-文本模态的同步处理，并进行模态对齐。
* 将Uni-filter集成到多个VLMs中，并在多种攻击场景下进行测试。
* 以面向视觉-语言的Uni-filter为基础，扩展至更多模态，如视频、语音等多模态大模型（MLLMs）的攻防机制，提高大模型的鲁棒性与安全性。

**预期目标：**

开发能够同时处理视觉和文本模态的过滤器，显著提高VLMs在面对prompt 注入或对抗性攻击时的鲁棒性。进一步平衡安全性与模型响应性能， 在尽可能不牺牲模型生成能力的前提下，提高VLMs的稳定性与安全性。