《人工智能课程——视觉与自然语言处理》期末项目的个人贡献报告

------基于Qwen-VL与Stable Diffusion的多模态图像处理应用

杨欣怡 U202215067

在本项目中，我的核心贡献集中于前期的任务解构、技术路径规划和核心方法论的设计。我致力于将宽泛的大作业要求转化为清晰、可执行的技术方案，并为团队的工作奠定了坚实的理论和逻辑基础。具体贡献如下：

1. 任务分析与解构：  
   面对“生成多维度描述信息”和“完成图文应用”的宽泛要求，我首先对任务进行了系统性解构。

对于图像描述生成任务，我将其拆解为三个核心挑战：(1) 如何定义“高质量、信息量大”的描述？ (2) 如何引导模型生成符合此定义的描述？ (3) 如何在不进行训练的情况下最大化模型能力？ 基于此，我提出了从“空间布局”、“物体属性”、“场景光影”、“客观视角”等多个维度构建描述文本的方案，明确了工作的具体方向和评价标准。

对于图像风格迁移任务，我将其核心问题定位为“如何在文本指令的控制下，实现可控且保真的风格化转换？”。这决定了我们后续的工作将重点探索strength等关键参数对“内容保真度”与“风格强度”之间平衡的影响，而非简单地调用模型。

2. 技术方案选型与论证：  
在明确了任务边界后，我主导了技术方案的选型与论证。

我分析了不同多模态模型（如CLIP、BLIP、Qwen-VL）的特性，最终论证并推动选择Qwen-VL作为描述生成的核心模型。其关键优势在于：强大的零样本能力、对中文的良好支持、以及通过提示工程进行精细控制的潜力，这完美匹配了我们“不训练、靠引导”的核心思路。

同样，在风格迁移模型选择上，我论证了Stable Diffusion相较于传统GAN或AdaIN等方法的优势，即其基于文本提示（Text Prompt）的无限风格扩展性和出色的生成质量，这为实现“多种风格变换”的目标提供了技术可行性。

3. 方法论设计：提示工程框架的构建：  
这是我最为关键的贡献。我认识到，要解决“如何引导模型”的问题，必须设计一套系统性的方法论，而非零散的提示词尝试。

我设计了一套层次化、结构化的提示工程框架。该框架将单一的“描述这张图片”指令，扩展为一系列明确的、可编程的生成要求（如：必须分析空间关系、必须描述关键属性、必须采用客观视角等）。

这套框架将人类的认知逻辑和描述习惯嵌入到了与模型的交互中，本质上是通过Prompt塑造了模型的“思维链”（Chain-of-Thought），从而系统性地提升了输出质量。它是我们图像描述生成任务的核心创新点和方法论基础。

4. 项目可行性评估与风险规划：  
在技术方案确定后，我对方案的可行性进行了评估，并识别了潜在风险。

我评估了在消费级GPU（RTX 4090/5060）上运行Qwen-VL和Stable Diffusion的显存与时间成本，确保了项目在给定资源和时间内是可行的。

我预见到传统自动化指标（如BLEU）可能无法有效评估我们生成的描述，并提前规划了人工评估和定性分析作为重要的评估后备方案，这保证了我们最终能对工作成果做出有效结论。

总结：  
我在项目中扮演了架构师和策略制定者的角色，为团队厘清了方向、规划了路径、设计了方法。我对任务的深度解构和由此形成的结构化提示工程方法论，使项目最终产出高质量、有价值成果。

