西安交通大学

自然语言处理与应用
实验报告

报告名称：基于多模态大语言模型的问答式视觉定位方法

姓名	学号
XXX	xxxxxxxxxx
XXX	xxxxxxxxxx
XXX	xxxxxxxxxx
XXX	xxxxxxxxxx
XXX	xxxxxxxxxx

西安交通大学　人工智能学院

2025 年 2 月 20 日

摘要

在针对人机交互（HRI）中的视觉定位任务中，传统方法因自然语言中的歧义性问题，通常依赖预定义模板进行消歧．然而，这种方式在现实交互场景中的表现有限．本报告实现了一种端到端问答式视觉定位方法．该方法通过单一模型实现三个角色的功能，能够通过主动信息收集来消除用户输入的歧义性．模型基于视觉对话与定位的统一框架，可以在多个公共数据集上进行联合训练，并在开放世界的复杂场景中展现出良好的通用性．实验设计涵盖了 150 个具有挑战性的交互场景，结果表明，该方法在处理多样化的视觉与语言输入时，表现出优越的适应性与高成功率.