2024-2025学年第1学期

**《深度学习与实践》课程项目报告**

|  |  |
| --- | --- |
| 学 生 | 学号： 2024282110096 姓名：程劭伟 |
| 指导教师 | 宋麟 |
| 完成日期 | 2025年 1月 5 日 |

1. **选题**

《GeoChat : Grounded Large Vision-Language Model for Remote Sensing》论文复现

1. **需求分析**

大视觉语言模型(Large Vision-Language models, vlm)的最新进展在自然图像处理方面显示出巨大的前景，允许用户就给定的视觉内容进行对话。然而，这种通用领域的vlm在遥感(RS)场景中表现不佳，导致在使用RS领域特定的查询时出现不准确或伪造的信息。这种行为的出现是由于RS图像带来的独特挑战。例如，要处理具有不同类别和许多小对象的不同尺度变化的高分辨率RS图像，除了整体场景解释外，还需要区域级推理。此外，缺乏特定于领域的多模态指令跟踪数据以及用于RS的强大骨干模型，使得模型难以将其行为与用户查询保持一致。

针对上述问题与需求，该论文的作者主要在以下三个方面做出了较大的贡献：

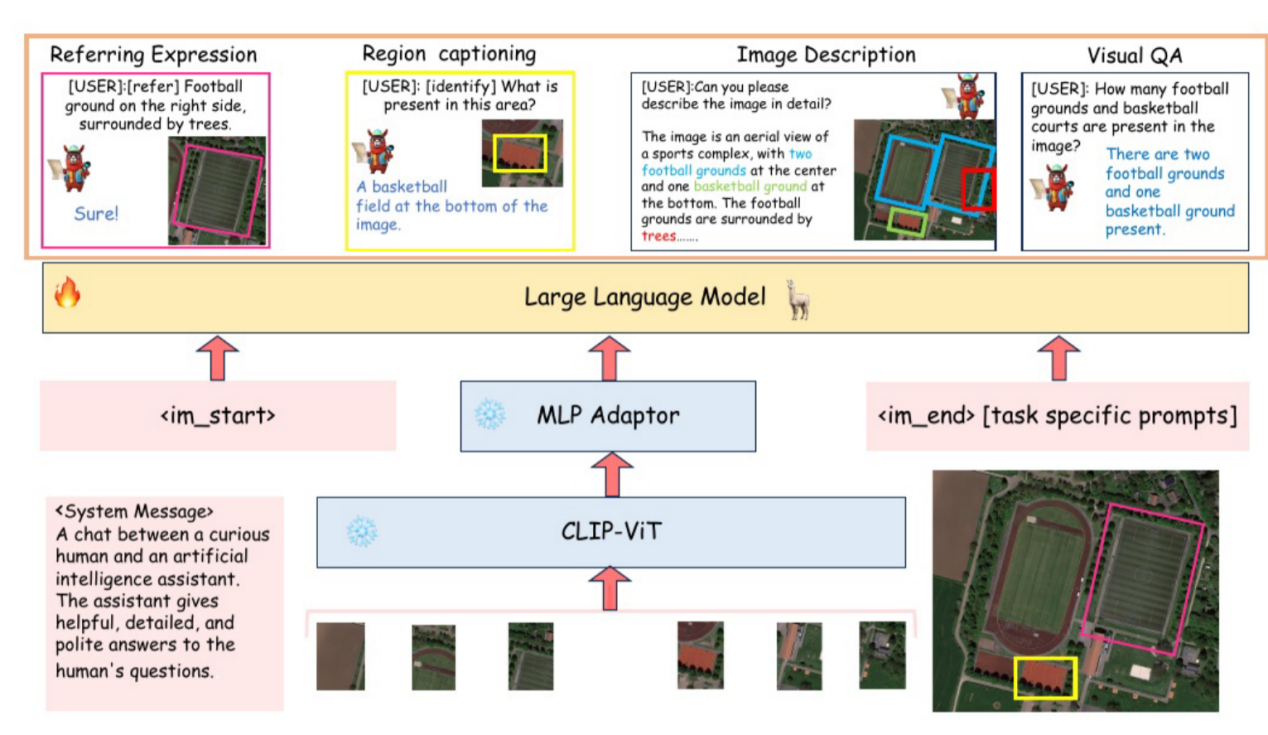
•RS多模态指令跟随数据集。提出了一种新的数据生成管道，利用现有的对象检测数据集来创建图像的简短描述，然后使用Vicuna-v1.5来单独使用生成的文本创建对话。此外，使用相应的数据集添加了视觉问答和场景分类能力。这导致RS域总共有318k个指令对。

•GeoChat。利用作者的数据集，对LLaVA-1.5进行微调，以创建遥感域视觉语言模型- GeoChat。作者的LoRA微调是高效的，避免了忘记完全调优的LLaVA模型中嵌入的必要上下文，该模型的MLP投影被训练以将图像对齐到LLM (Vicuna-v1.5)的单词嵌入空间中。这使得geochat保留了LLaVA的对话和指令跟踪能力，并将其领域知识扩展到遥感任务。

•解决了缺乏评估现有VLMs遥感对话能力的评估基准的问题。为此，作者为RS中的会话基础设置了评估协议，并设置了一套任务，以便与未来在这个方向上的努力进行比较。作者还展示了不同遥感任务的各种监督评估，包括视觉问答，场景分类，以展示GeoChat会话VLM的通用性。

1. **设计**

3.1GeoChat:接地遥感VLM



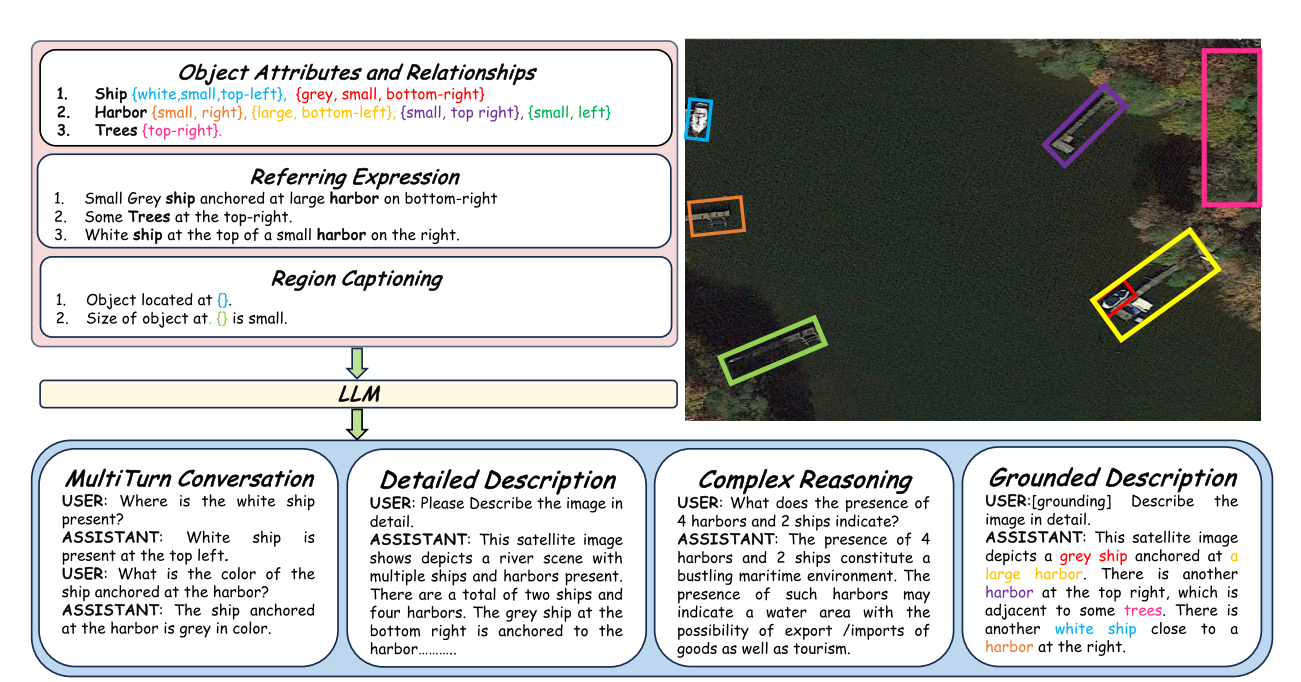
GeoChat遵循llva -v1.5的架构，它由三个核心组件组成，全局图像编码器，MLP适配器和 LLM。与LLaVA不同的是，添加了特定的任务提示符，用于指示模型所需的任务类型，即基础、图像级或区域级对话。

视觉主干。GeoChat采用CLIP-ViT(L-14)的预训练视觉主干，其输入分辨率为336×336。这导致每张图像有效地有576个补丁。由于该分辨率不足以理解遥感图像中呈现的细节(例如，小物体和物体细节)，因此在基于变压器的CLIP模型中插入位置编码，以根据输入图像大小504×504进行缩放。虽然这导致补丁数量增加到几乎翻倍，但是这种增强的分辨率使其能够处理更大的图像尺寸，并且还支持高分辨率RS图像中更好的视觉接地。

MLP跨模态适配器。用带有一个隐藏层的MLP适配器，将维度为1024的输出令牌投影到语言模型空间。该适配器的输入维数为1024，输出大小为4096的向量，对应于LLM的输入大小。

大型语言模型。开源的Vicuna-v1.5(7B)大型语言模型被用作GeoChat的基础。该语言模型在作者的框架中作为一个单一的界面，用于不同的视觉语言输入。为了完成不同的视觉语言任务，该模型直接依赖于Vicuna-1.5(7B)语言库。

3.2RS多模式指令数据集



组成数据集:在指令集的编译中，合并了三种不同类型的数据集，包括目标检测，场景分类和视觉问题回答设计的数据集。具体来说，集成了三个目标检测和一个洪水检测VQA数据集。目标检测数据集允许区域级推理能力，因为它们提供分割掩码和边界框。

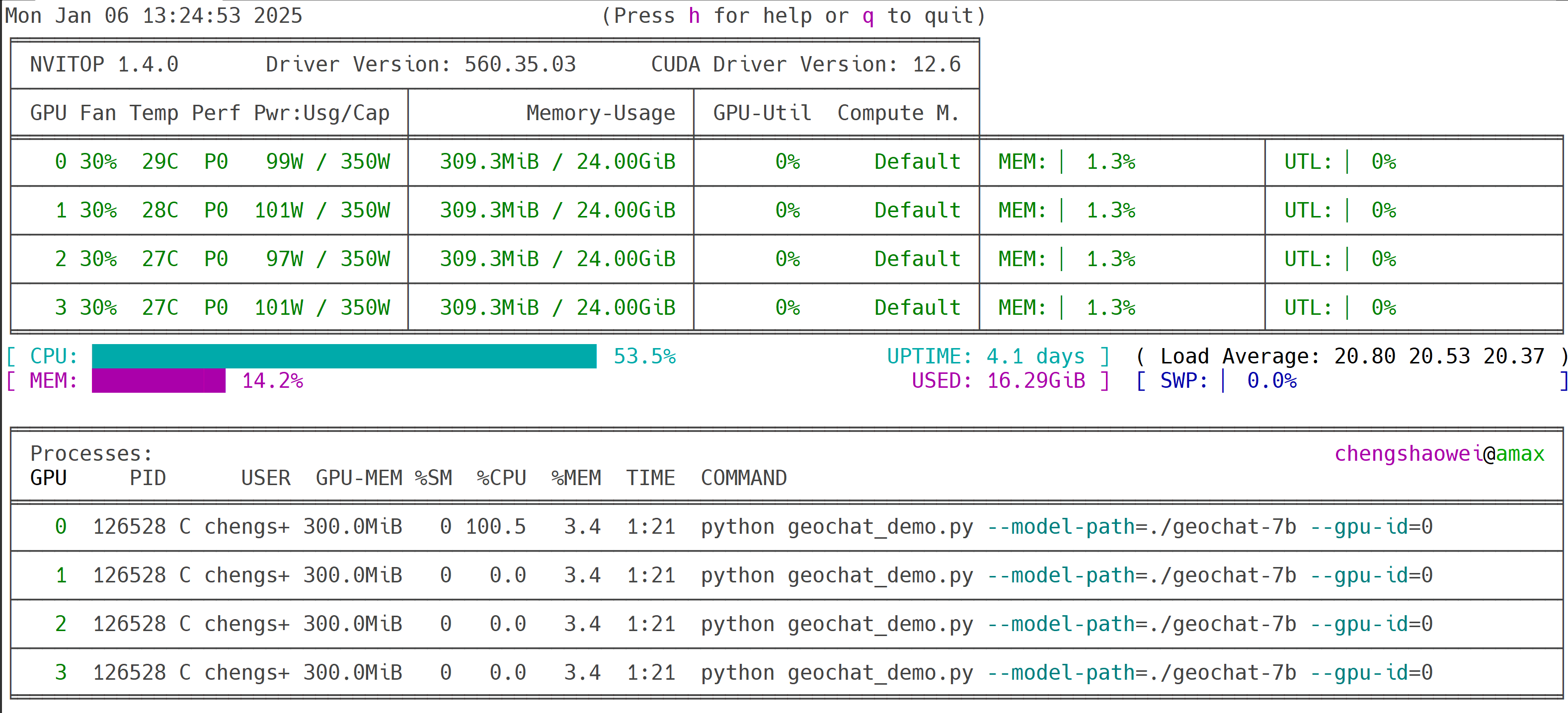
缺失类的添加:尽管对象检测数据库中包含了各种各样的对象类，但一些基本类别，如建筑物、道路、而树是缺失的。为了解决这个问题，使用ViTAE-RVSA模型，该模型在LoveDA数据集上进行预训练，其中包含所需的重要类。模型用于在SAMRS数据集上推断这些类，产生伪标签。为了减轻这些预测中的潜在噪声，删除了ViTAE-RVSA的预测，因为已经从SAMRS数据集中获得了基本事实，以改进结果。

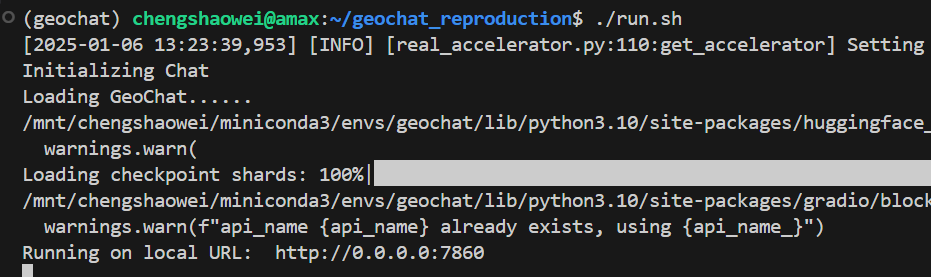
属性提取:为了引用表达式注释，在RS图像中派生各种属性是很重要的。为此选择了五种不同类型的属性，目标类别信息可以直接从SAMRS数据集中获得。对于颜色提取，使用K-Means聚类算法。具体来说，使用ground-truth box从图像中提取对象的像素，并将其聚类为K组。然后选择最大聚类的中心作为目标的颜色。为了指定对象的相对大小，我们将大小对象分为三种大小:小、正常和大。这种分类是通过测量所有面积来确定的类在整个数据集中的位置，并将80th

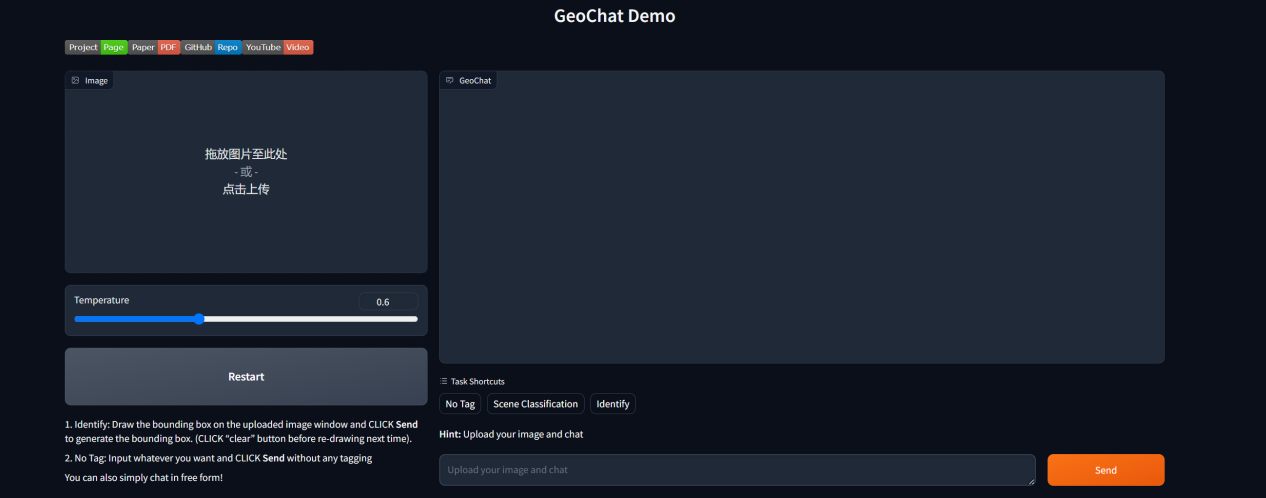
百分位数分配为大标签。类似地，20th百分位数被指定为小尺寸，其余的归为正常类别。为了确定物体在图像中的相对位置，将整个图像划分为3×3网格，定义诸如右上、上、左上、左、中、右、右下、左下和底部等区域。基于对象的中心像素坐标，相应地分配其相对位置。

1. **关键代码**
2. **运行效果**

程序运行截图

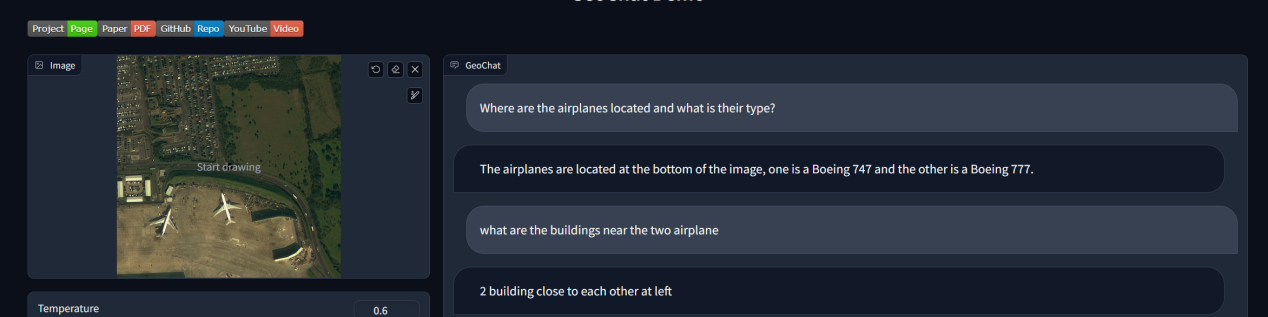




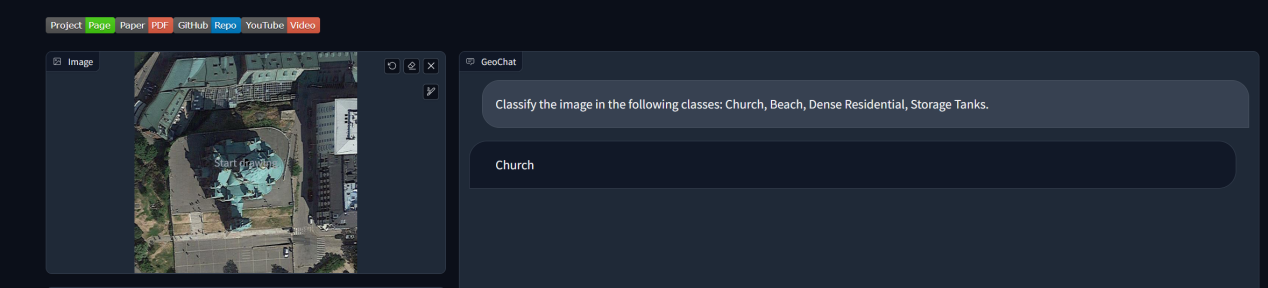


接下来分别演示任务Visual Question Answering，Scene Classification，Grounded Description，Referring Expression。

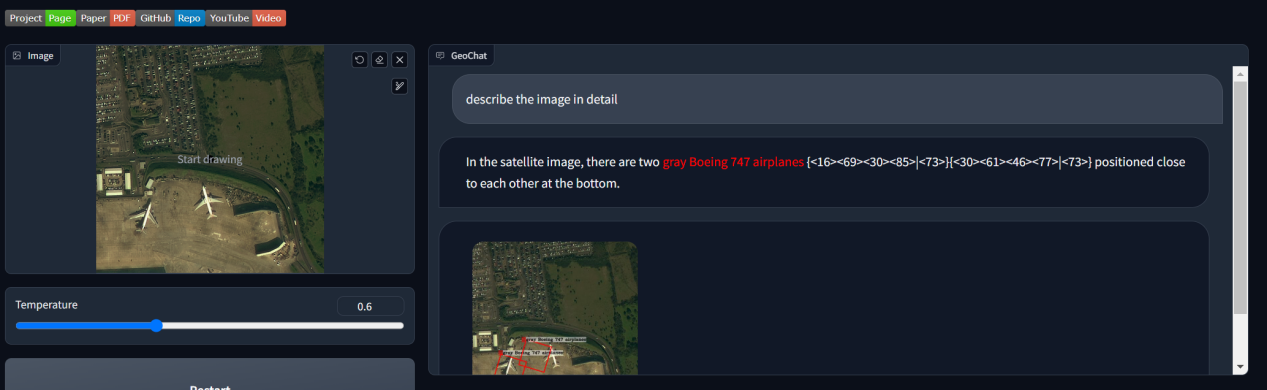
Visual Question Answering



Scene Classification

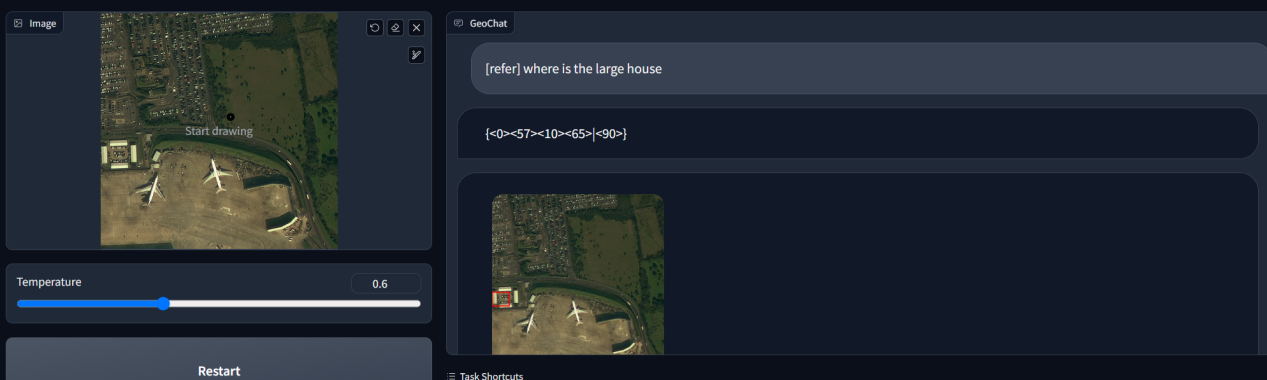


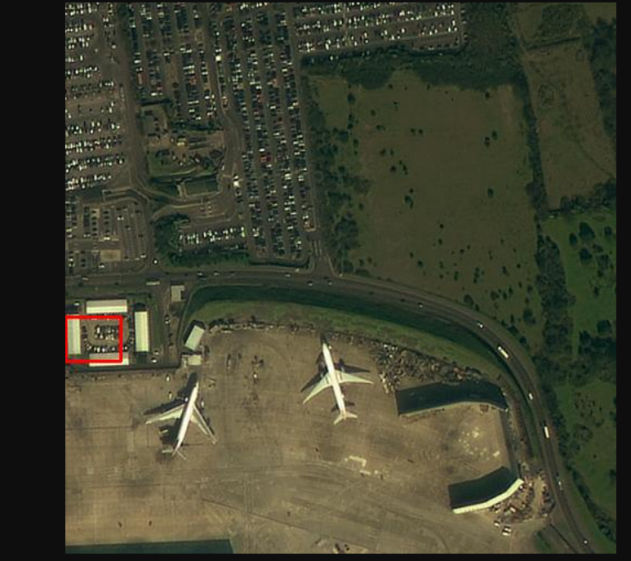
Grounded Description





Referring Expression





1. **总结**

尽管大型视觉语言模型的最新进展在自然图像处理中显示出前景，但由于独特的领域特定挑战，它们在遥感场景中的表现仍然有限。为了解决这一差距，作者提出了GeoChat，这是第一个统一的遥感VLM，具有高分辨率RS图像的多任务会话能力。GeoChat不仅可以回答图像级别的查询，还可以参与特定区域的对话，以精确的空间坐标为基础进行响应。作者还创建了一个新的RS多模态指令跟踪数据集，该数据集由318k具有不同多任务格式的图像指令对组成。GeoChat在场景分类、VQA、多回合对话、视觉接地和参考目标检测等各种RS任务表现较好。