

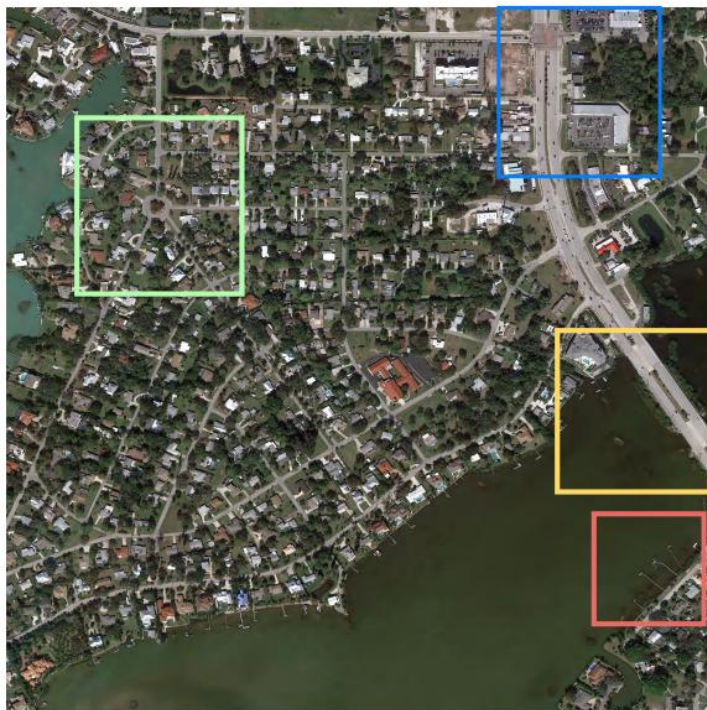
# ZoomEarth: Active Perception for Ultra-High-Resolution Geospatial Vision-Language Tasks

CVPR 2026

汇报：蒋明忠      2026.10

## 遥感视觉问答的任务与高分辨率挑战

视觉语言模型（VLM）能够结合图像和文本问题生成答案。遥感视觉问答（RS VQA）将这一能力应用于卫星或航空影像，要求模型理解地物的类别、数量、空间关系和属性。  
随着遥感影像分辨率提高，模型需要同时理解**大范围的场景信息和局部小目标的细节**。



### 核心问题：细节保留与计算开销难以兼顾

遥感影像覆盖范围广，建筑、车辆、船舶等目标尺度差异大。将整图缩小后输入模型，会损失小目标的颜色、形状等细节；直接输入高分辨率整图，又会带来大量计算和无关背景信息。

### 以论文中的 4,000–5,000 像素影像为例

判断“住宅区还是工业区”主要依赖全局场景；  
判断“码头是否停船”则需要准确定位并观察局部。  
不同问题需要不同尺度的视觉信息。

# Background



## ZoomEarth

image input

Drop Image Here  
- or -  
Click to Upload

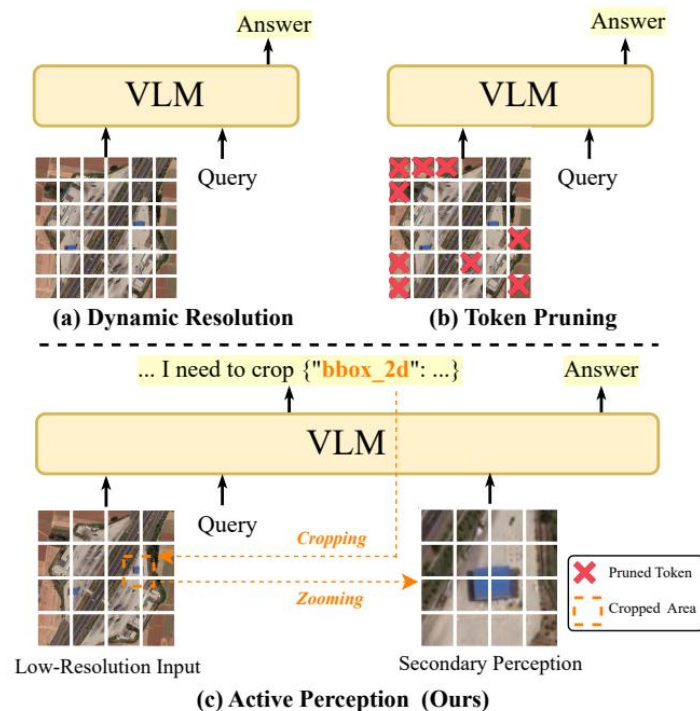
Use via API • Built with Gradio • Settings

Enter prompt please.

Run!

## 现有高分辨率图像输入方式及其局限

视觉编码器将图像转换为视觉 Token，即模型能够处理的图像信息单元。现有方法通常通过动态分辨率输入或 Token 剪枝处理大图，以保留更多细节或降低计算量。但这类方法主要依赖一次图像输入，模型难以在作答过程中重新获取缺失的局部信息。



### 动态分辨率：增加输入细节

允许模型接收更大的图像，但更细的整图输入也会增加视觉 Token 数量和计算开销。

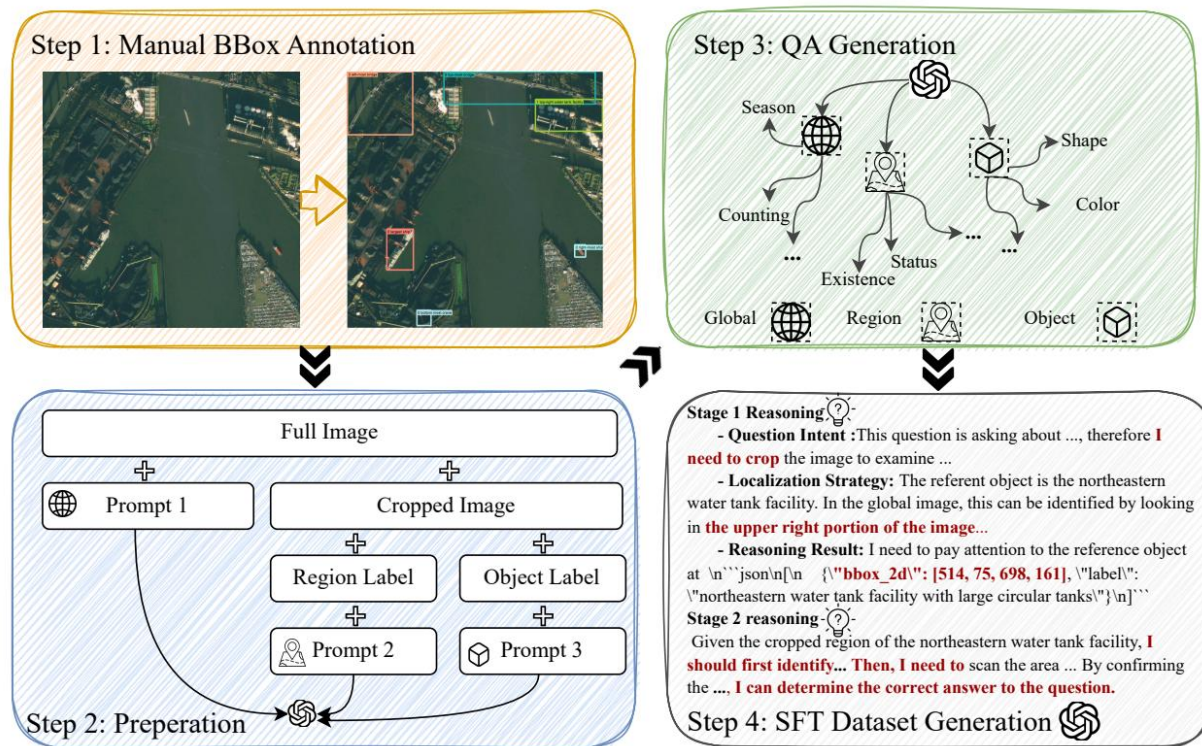
### Token 剪枝：减少冗余信息

按预设规则筛除部分图像信息。在复杂遥感背景中，这些规则难以适应不同问题所需的区域和目标尺度。

### 本文切入点：按问题主动选择后续观察区域

先获取全局信息，再定位并读取原图局部细节。

## ① LRS-GRO: 教模型 “答案是什么” 和 “应该看哪里”



(a) Data Generation Pipeline

构建过程：人工标框，GPT-4o 生成约 40k 候选 QA，再经人工筛选与修订，保留 13,245 条并平衡答案。

## 数据规模与任务

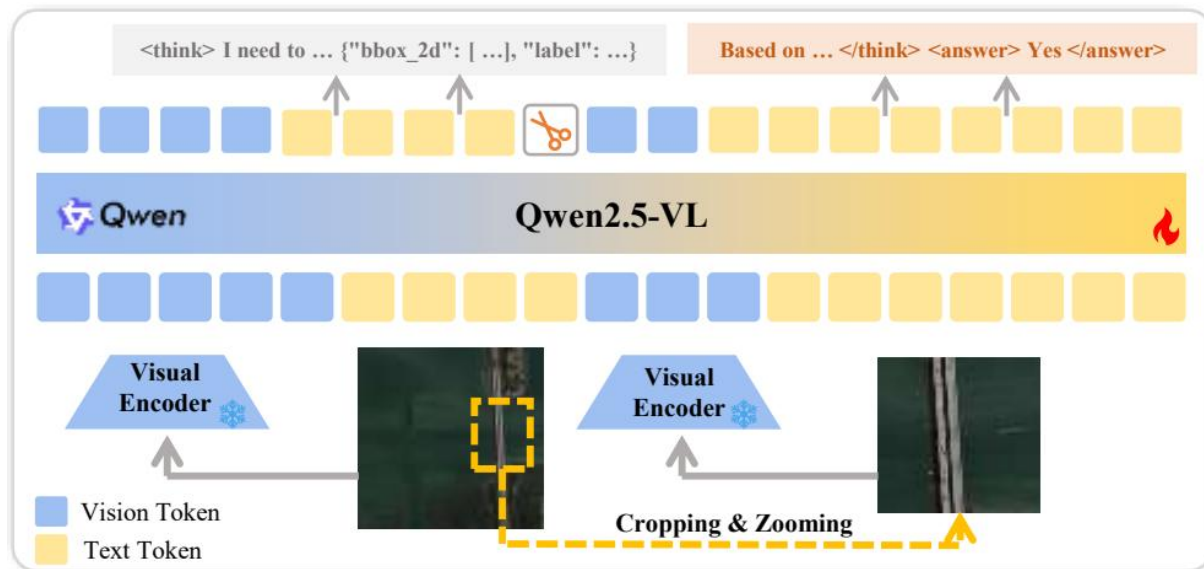
1,224 张影像，3,592 个目标框，  
13,245 个问题，覆盖 17 类问答。

全局：场景、季节、城乡、计数  
区域：如码头、机场中的存在与状态  
对象：如桥梁的形状、车辆的颜色

## 每条训练样本提供什么？

原始图像 + 问题 + 答案；  
局部问题还提供参考框，告诉模型看哪里。  
其中 1,000 条附带分步过程：  
理解问题、定位、裁剪、观察并作答。

## ② 一次作答过程：粗看全图，定位裁剪，再看细节



### 示例问题：最大的混凝土桥是直的还是弯的？

模型输出的框坐标与描述（论文示例）

```
{"bbox_2d": [897, 488, 1023, 683],  
  "label": "largest concrete vehicle bridge"}
```

### 1 输入缩略图和问题

先获得全局位置关系。若问题只问整体场景，可以直接作答，不必调用裁剪工具。

### 2 判断需要细看，并预测目标框

框由左上角、右下角坐标定义，即  $[x\_min, y\_min, x\_max, y\_max]$ 。

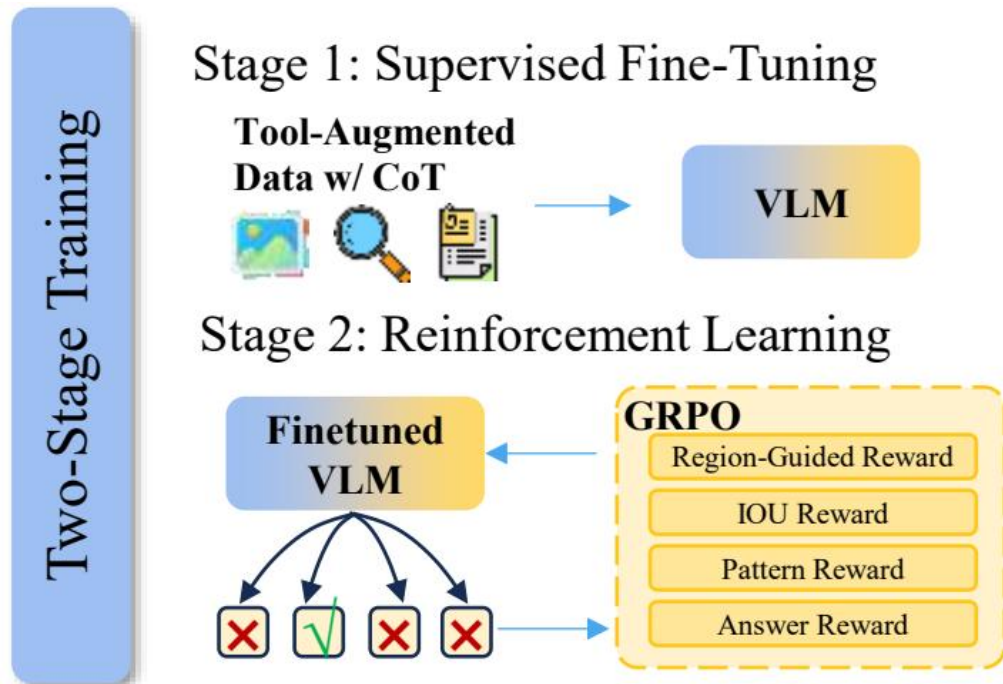
### 3 工具从原始高分辨率图像取出区域

将框对应回原图，返回包含清晰细节的局部图。这一步增加一次视觉观察。

### 4 结合局部观察继续推理

模型看到桥的形状后输出答案：直的。

## ③ SFT + GRPO: 先学会操作，再学会把操作用好



### 基座模型: Qwen2.5-VL-3B

约 3,500 条训练样本、约 10,000 条测试样本。  
SFT 使用 1,000 条 CoT, RL 使用 2,500 条。  
训练前的全图输入分辨率为 512。

### 第一阶段 SFT: 模仿带工具的示范

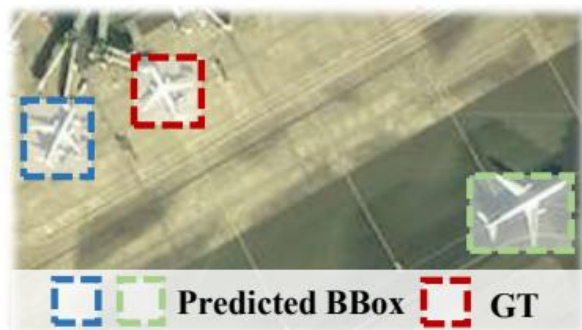
示范包含定位理由、框、裁剪后的观察和答案。  
模型学习何时调用工具、怎样输出可执行的框。  
需裁剪与无需裁剪的样本按 2:1 混合，  
让模型同时学习“调用”和“直接回答”。

### 第二阶段 GRPO: 比较候选方案的奖励

对同一问题采样多条候选作答轨迹。  
执行需要的裁剪，再对定位、答案、格式打分。  
在同组候选中比较得分，让高奖励的轨迹  
更容易再次出现，逐步改善选区与作答策略。

**SFT 提供起点; GRPO 根据实际效果继续优化。**

## ④ Region-Guided Reward: 找得近, 也给学习信号



$$r_{IoU} = \begin{cases} IoU, & \text{if match} \\ 0, & \text{otherwise} \end{cases}$$

$$r_{R-G} = \text{sigmoid}\left(\frac{\alpha}{\text{distance} + \epsilon}\right)$$

|        | $r_{IoU}$ | $r_{R-G}$ |
|--------|-----------|-----------|
| BBox 1 | 0         | 0.1       |
| BBox 2 | 0 🚫       | 0.5 😊     |

### 为什么只有 IoU 不够?

IoU 衡量预测框与正确框 (GT) 的重叠程度。  
两个框都没碰到目标时, IoU 都是 0,  
无法区分 “差一点找到” 和 “离得很远”。

区域引导项按框中心距离给分,  
鼓励预测逐步靠近正确区域。

### 区域引导奖励

$$r_{R-G} = \text{sigmoid}(\alpha / (d + \epsilon))$$

d: 两个框中心的欧氏距离;  $\alpha$ : 分辨率尺度参数;  
 $\epsilon$ : 避免数值溢出的小常数。公式按论文原文。

### 四项奖励共同评价一次作答

$$R = r_{IoU} + r_{R-G} + r_{\text{answer}} + 0.05 \times r_{\text{pattern}}$$

IoU: 是否框住正确目标, 奖励精确重叠。  
R-G: 离正确区域有多近, 补充密集信号。  
答案: 按归一化词义相似度评价回答。  
格式: 输出格式正确为 1, 否则为 0。

**主要结果：细节问答更准，跨数据集仍有效**

**LRS-GRO：既看答得对不对，也看框得准不准**

| 模型（初始输入）                     | Global       | Region       | Object       | Avg. Acc     | APO IoU      |
|------------------------------|--------------|--------------|--------------|--------------|--------------|
| InternVL3-8B (3200)          | 71.60        | 44.58        | 47.80        | 53.67        | —            |
| VLM-R <sup>3</sup> -7B (512) | 69.72        | 44.83        | 37.40        | 50.17        | 19.93        |
| <b>ZoomEarth-3B (512)</b>    | <b>63.09</b> | <b>46.11</b> | <b>51.80</b> | <b>53.76</b> | <b>34.39</b> |

相对 InternVL3：区域 +1.53 pp、对象 +4.00 pp；全局问题仍落后。

**外部数据集：仅在 LRS-GRO 训练，初始输入统一为 1024**

| 模型                  | MME-RW-RS    | XLRS-Bench   | GeoLLava-8k  | Avg. Acc     |
|---------------------|--------------|--------------|--------------|--------------|
| InternVL3-8B        | 41.00        | 36.70        | 37.60        | 38.43        |
| <b>ZoomEarth-3B</b> | <b>44.10</b> | <b>40.20</b> | <b>38.61</b> | <b>40.97</b> |

## 消融：裁剪、强化学习与区域奖励各自有效

### 裁剪带来的收益 (Table 4)

| 设置   | LRS-GRO      | MME-RW       | XLRS         | Geo-8k       |
|------|--------------|--------------|--------------|--------------|
| 无裁剪  | 51.10        | 42.10        | 30.00        | 34.30        |
| 加入裁剪 | <b>53.64</b> | <b>44.10</b> | <b>40.20</b> | <b>38.61</b> |

### 两阶段训练 (Table 5, MME-RW)

| 训练              | 无裁剪 Overall  | 裁剪 Overall   | 变化              |
|-----------------|--------------|--------------|-----------------|
| 仅 SFT           | 42.80        | 41.30        | -1.50 pp        |
| <b>SFT + RL</b> | <b>42.10</b> | <b>44.10</b> | <b>+2.00 pp</b> |

### 定位奖励 (Table 6)

| R-G      | IoU      | Avg. Acc     |
|----------|----------|--------------|
| 无        | 无        | 52.67        |
| 有        | 无        | 52.94        |
| 无        | 有        | 52.79        |
| <b>有</b> | <b>有</b> | <b>53.76</b> |

去掉 R-G: -0.97 pp  
去掉 IoU: -0.82 pp

SFT 能学会调用格式,  
RL 使裁剪带来正向收益。

## 分辨率消融与下游扩展

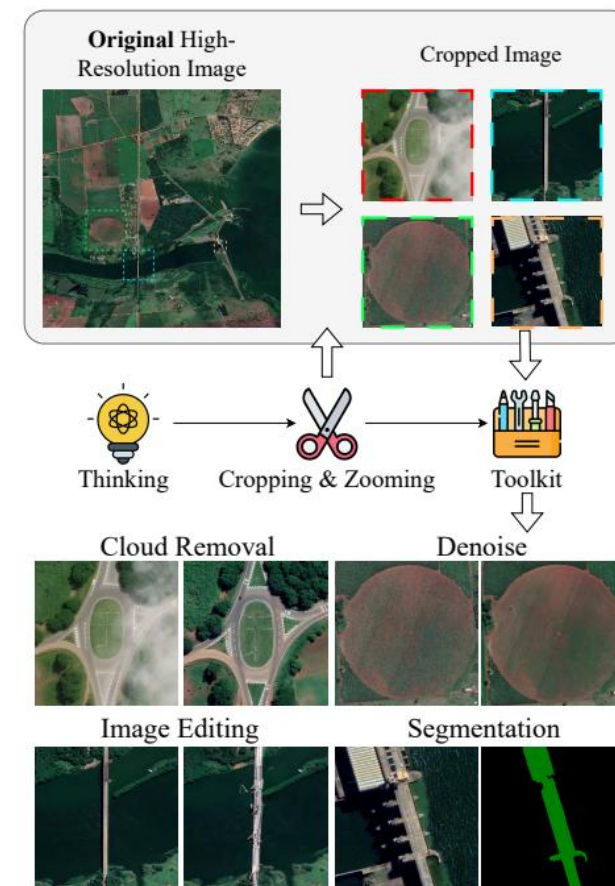
输入增大未明显提升准确率 (Table 7)

| 裁剪 | 初始分辨率            | Avg. Acc (%) | 速度 (it/s)   |
|----|------------------|--------------|-------------|
| 无  | 512 × 512        | 50.86        | 3.69        |
| 无  | 1024 × 1024      | 51.16        | 3.50        |
| 无  | 3333 × 3333      | 50.92        | 2.04        |
| 有  | <b>512 × 512</b> | <b>53.76</b> | <b>2.38</b> |

相对无裁剪 3333 输入: +2.84 pp, 速度约 1.17 倍。  
相对无裁剪 512 输入: 准确率提升, 但速度降低。

## 组会讨论

模型学会选择证据区域, 为遥感 Agent 提供基础。  
下游仅展示工具接入, 缺少统一定量收益评估。



去云、去噪、编辑、分割