InstructSAM: A Training-Free Framework for Instruction-Oriented Remote Sensing Object Recognition

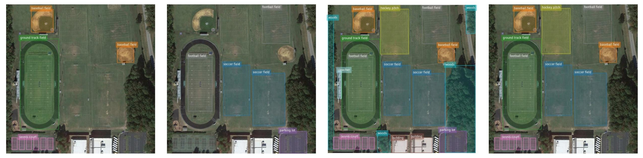
Yijie Zheng[1,2] Weijie Wu[1,2] Qingyun Li[3] Xuehui Wang[4] Xu Zhou[5]
Aiai Ren[5] Jun Shen[5] Long Zhao[1] Guoqing Li[1] Xue Yang[4]

[1]Aerospace Information Research Institute [2]University of Chinese Academy of Sciences
[3]Harbin Institute of Technology [4]Shanghai Jiao Tong University [5]University of Wollongong

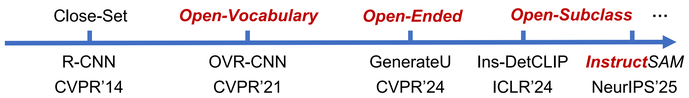
GAIA 2025

From close-set to instruction-oriented

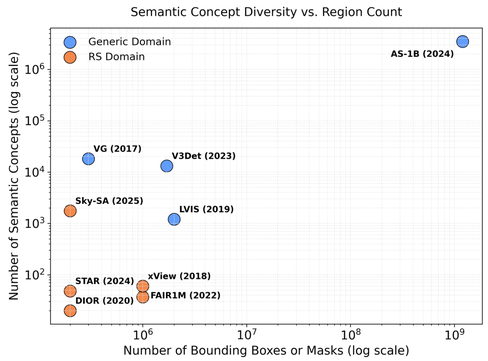
- Fixed categories limit traditional detection.
- Real tasks need flexible, dynamic recognition.
- ✨ **Instruction-oriented: just give an instruction, model adapts.**



Close-Set (annotations in DIOR dataset) | *Detect 'football field', 'soccer field', 'parking lot'.* | *Detect every object in sight.* | *Detect all 'sports fields'.*

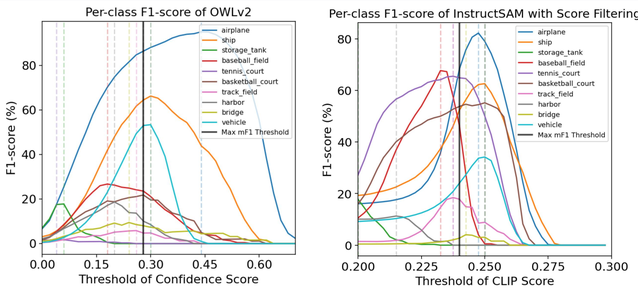


Motivation 1: Lacking diverse training datasets



- RS datasets lack semantic diversity → poor generalization.
- 💡 **Use generic LVLMs to identify objects.**

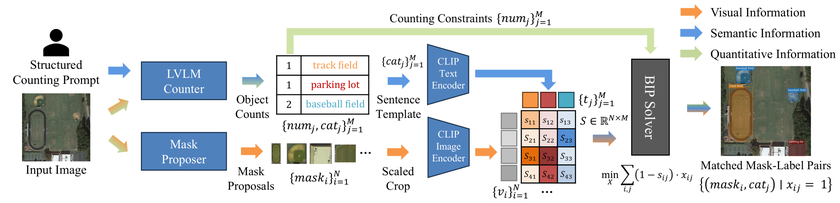
Motivation 2: Unreliable score filtering



- Score-based filtering is crucial to filter low-quality pseudo labels.
- Optimal thresholds vary across classes → no universal solution.
- Over-reliance on score thresholds leads to misclassifications.
- 💡 **Introduce counting constraints**

InstructSAM framework

- Decompose object segmentation into three easier steps.
- 💡 LVLM for categories & counts, SAM for mask proposals, CLIP for similarity, PuLP for optimization.



➢ Reframe segmentation as a mask-label matching problem

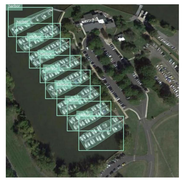
$$\min_{\mathbf{x}} \sum_{i=1}^{N}\sum_{j=1}^{M}(1-s_{ij})\cdot x_{ij} \quad \bullet \text{ minimize mismatches (1 − similarity)}$$

$$\text{s.t.} \sum_{j=1}^{M} x_{ij} \le 1, \quad \bullet \text{ One mask → one category}$$

$$\sum_{i=1}^{N} x_{ij} = num_j, \quad \bullet \text{ Total masks per category = LVLM count}$$

LVLM counts precisely and quickly

- With clear rules, GPT-4o counts as accurately as Faster R-CNN (80% vs. 81% mF1).

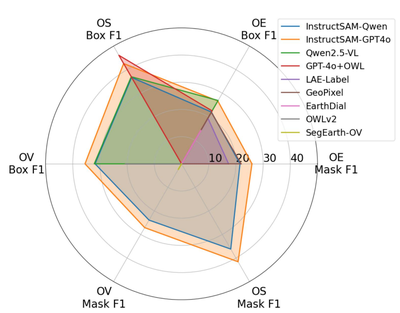


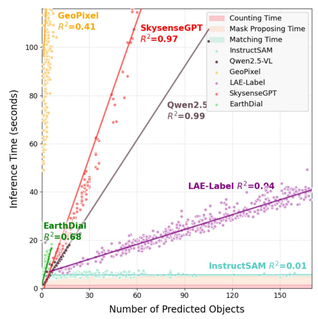
```
Count the number of harbors. Answer in JSON format.
{"harbor": 1}   ×
Count the number of harbors. Answer in JSON format.
Instructions:
- Harbor = pier to dock ships, Count each pier separately.
{"harbor": 8}   √
```

- Counting greatly reduces tokens and inference time.

```
Count the number of planes.
{"plane": 8}                                   (6 tokens, 0.5s)
Detect all the planes.
[{"bbox_2d": [58, 40, 78, 97], "label": "plane"},
 {"bbox_2d": [52, 56, 70, 60], "label": "plane"},
 ...
 {"bbox_2d": [42, 22, 76, 33], "label": "plane"}] (183 tokens, 10s)
```

Zero-Shot Results across Three Settings

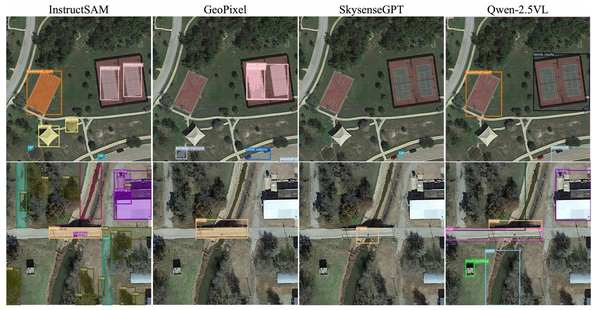




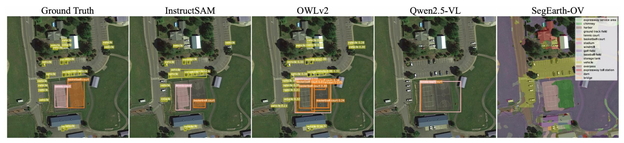
- Evaluated on NWPU and DIOR dataset
- 🚀 Strong performance across most settings
- 🤗 Performs well using open models (Qwen2.5-VL)
- Inference time in open-ended setting
- 89% fewer tokens, 32% inference time reduction
- Inference time stays constant with more objects

Qualitative results: open-ended setting

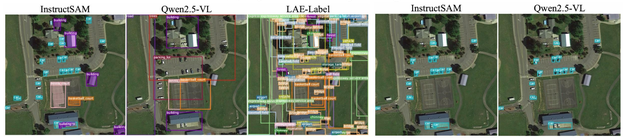
- ❌ Existing RSVLMs fail to generalize beyond training categories.
- ✅ InstructSAM recognizes more categories, e.g., pavilion and tree.



Qualitative results across three settings



(a) Open-Vocabulary setting

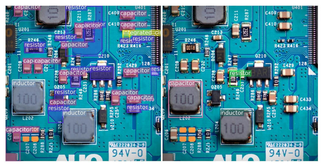


(b) Open-Ended setting (c) Open-Subclass setting

Generalization beyond RS

- InstructSAM also generalizes to natural images.

Instruction:
Detect all electronic components.



Instruction:
Detect dices whose letters come before K.

Takeaways & future work

🔑 **Takeaways**
- Flexible: Works with diverse instructions.
- Efficient: Faster than detection, saves tokens, scales well.
- Training-Free: Directly benefits from stronger open or proprietary models.

🚀 **Future Work**
- Expand InstructSAM to semantic segmentation (land cover mapping).
- Develop stronger Remote Sensing Foundation Models (e.g., SAM, CLIP).

Code | Personal Page

- Open for PhD / Visiting Opportunities
- zhengyijie23@mails.ucas.ac.cn

arXiv: http://arxiv.org/abs/2505.15818 NeurIPS 2025