UOW
AUSTRALIA

NEURAL INFORMATION
PROCESSING SYSTEMS

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

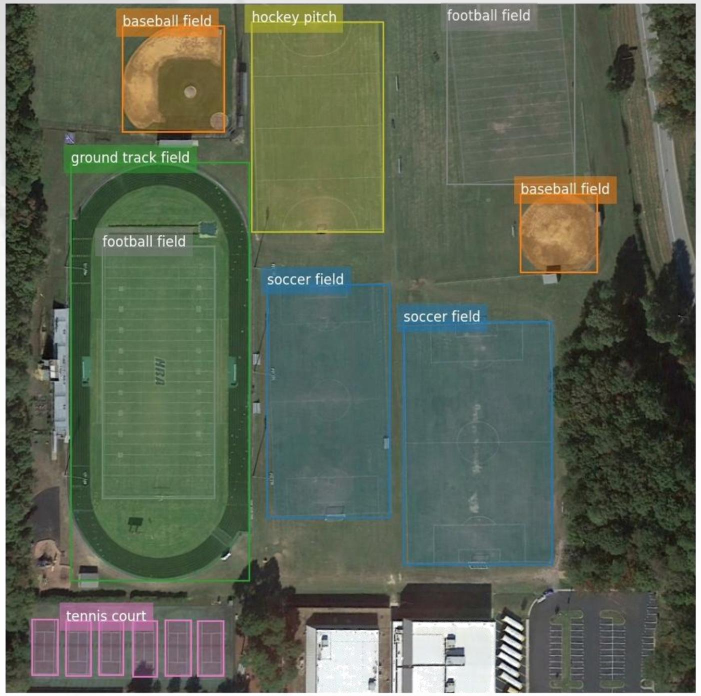
Yijie Zheng[1,2] Weijie Wu[1,2] Qingyun Li[3] Xuehui Wang[4] Xu Zhou[5]
Aiai Ren[5] Jun Shen[5] Long Zhao[1] Guoqing Li ✉[1] Xue Yang[4]

[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

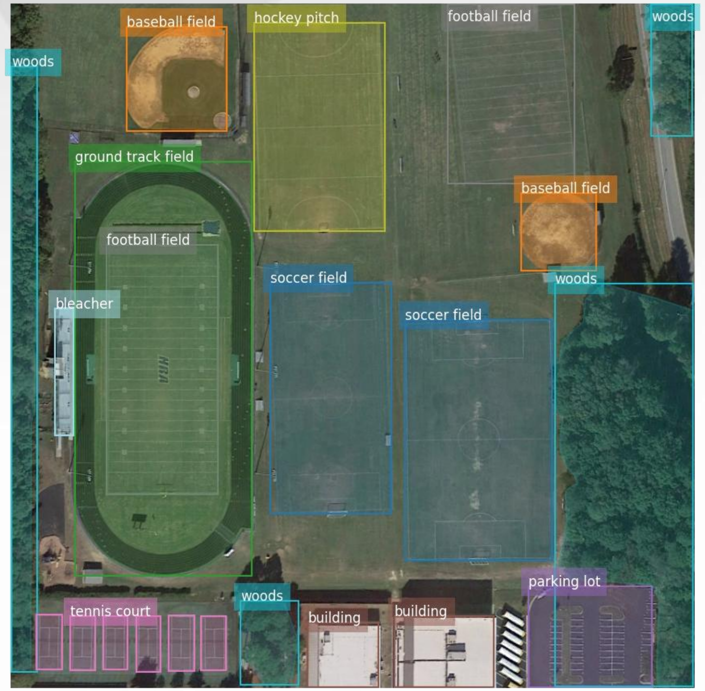
Why Instruction-Oriented Recognition?

🤔 Current challenge

- Understand implicit instructions
- Hard to list every possible category (football, baseball, tennis, hockey…)



Find all sports facilities



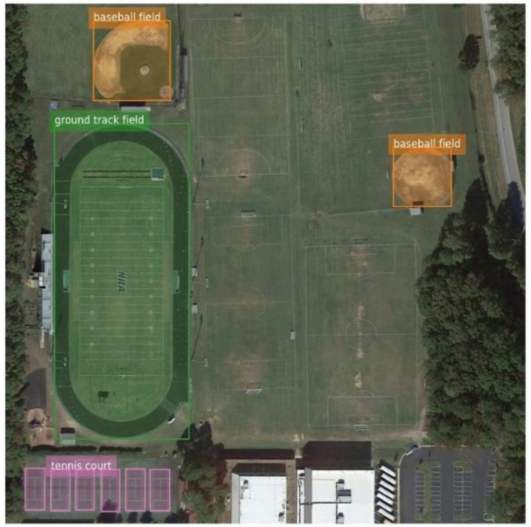
Detect everything in sight

✨ Motivation

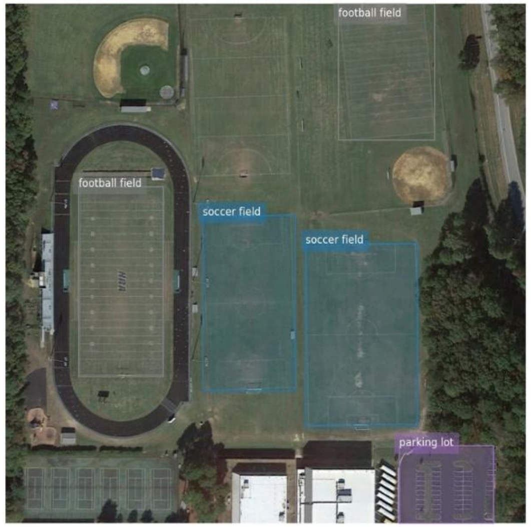
- Just give an instruction, and let the model adapt

From Close-Set to Instruction-Oriented Tasks

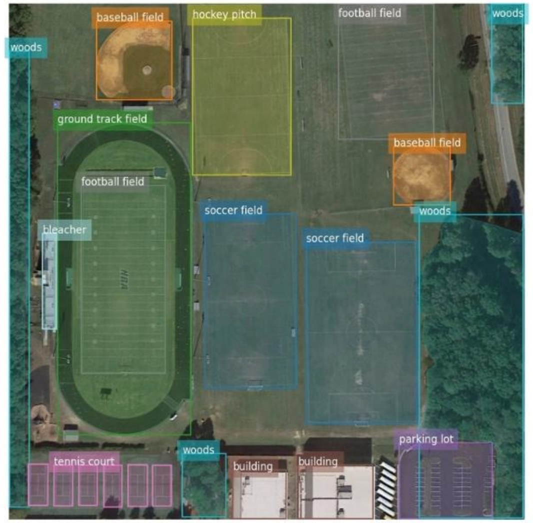
- A roadmap of object detection: from close-set to instruction-oriented.



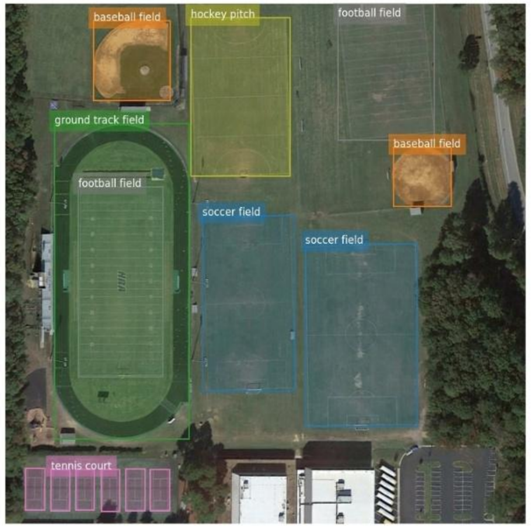
Close-Set (annotations in DIOR dataset)



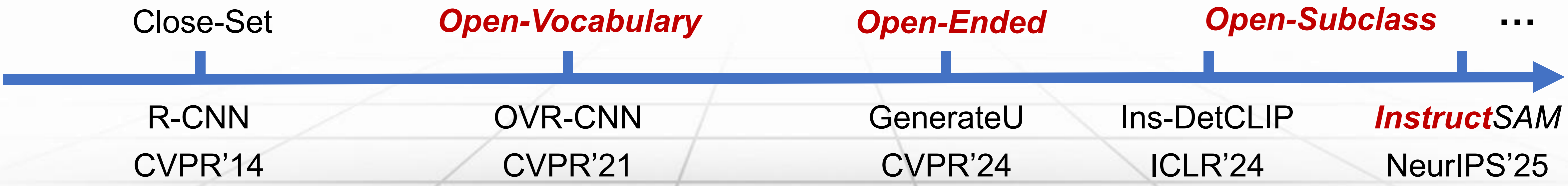
Detect 'football field', 'soccer field', 'parking lot'.



Detect every object in sight.



Detect all 'sports fields'.



Close-Set — R-CNN CVPR'14

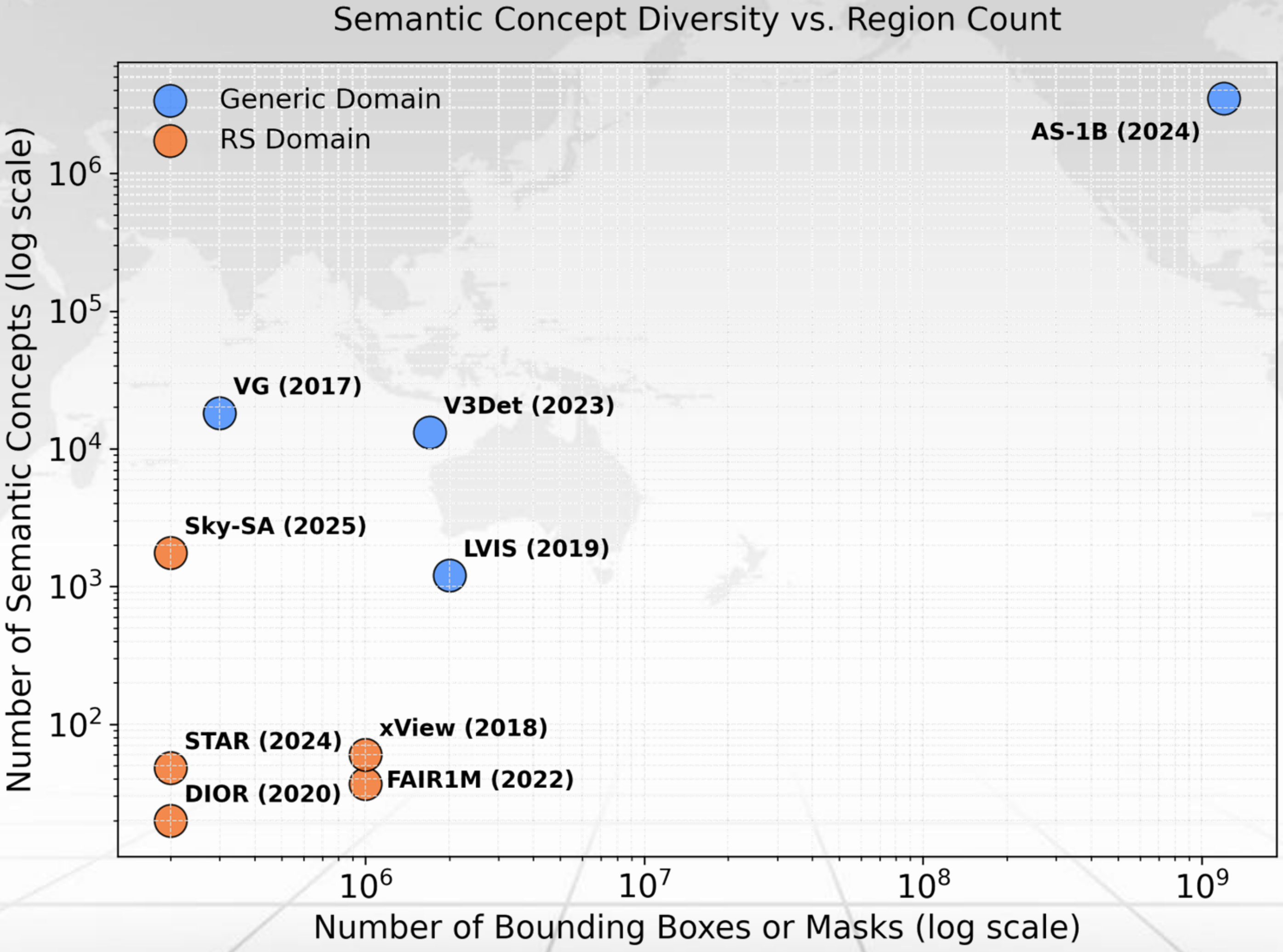
Open-Vocabulary — OVR-CNN CVPR'21

Open-Ended — GenerateU CVPR'24

Open-Subclass — Ins-DetCLIP ICLR'24

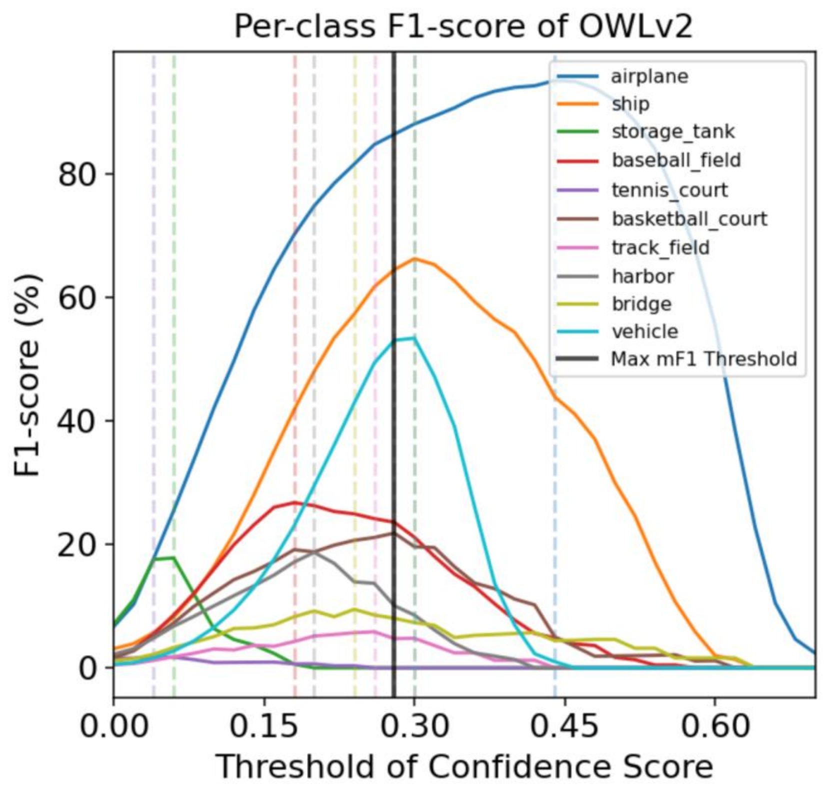
… — ***Instruct****SAM* NeurIPS'25

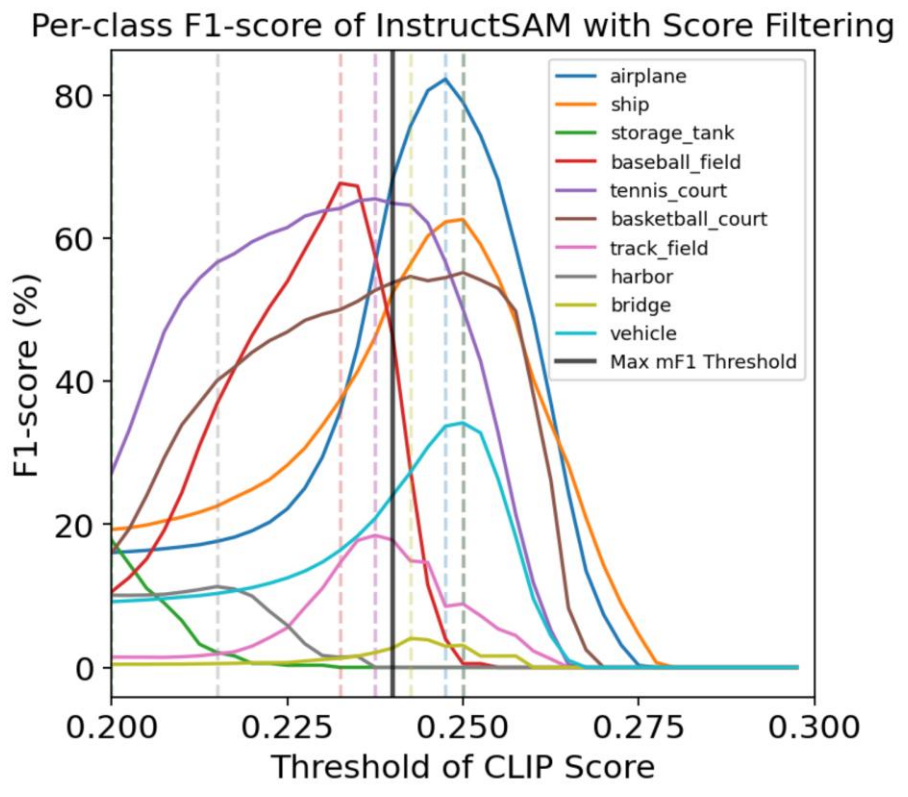
Motivation 1: Limited Semantic Diversity in RS Datasets



Motivation 2: Pseudo Labels Depend on Fragile Thresholds

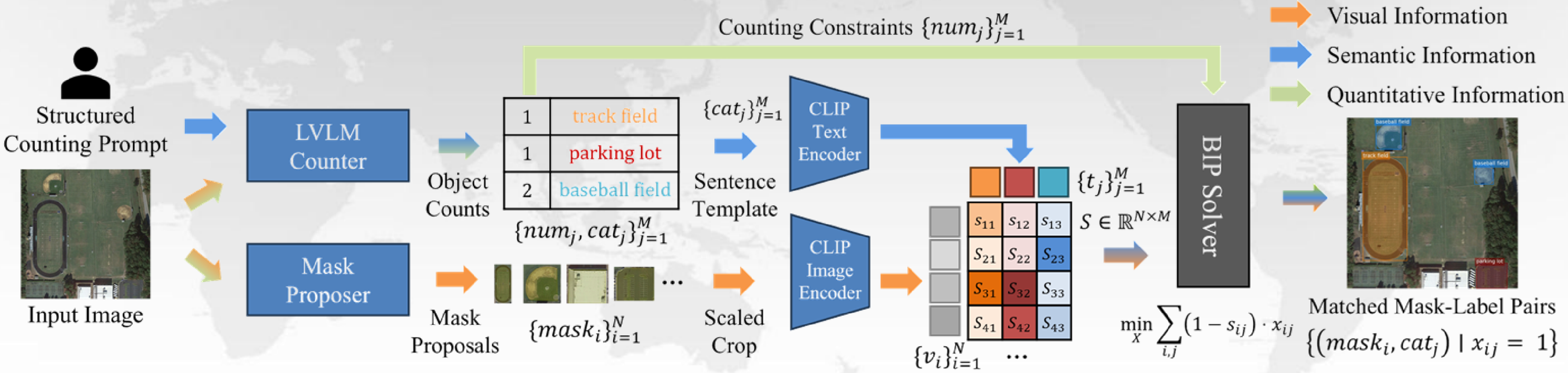
- Score-based filtering is crucial but unstable
- Optimal thresholds vary across classes → no universal solution
- Over-reliance on score thresholds leads to misclassifications





InstructSAM: Instruction-Oriented Remote Sensing Object Recognition

- 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

$$\min_{\mathbf{X}} \quad \sum_{i=1}^{N}\sum_{j=1}^{M}(1-s_{ij})\cdot x_{ij}$$

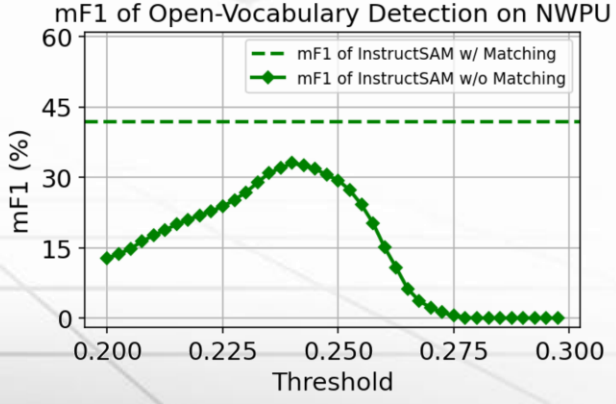
- minimize mismatches (1 − similarity)

$$\text{s.t.} \quad \sum_{j=1}^{M} x_{ij} \leq 1,$$

- One mask → one category

$$\sum_{i=1}^{N} x_{ij} = num_j,$$

- Total masks per category = LVLM count



Instruction-Oriented Object Counting

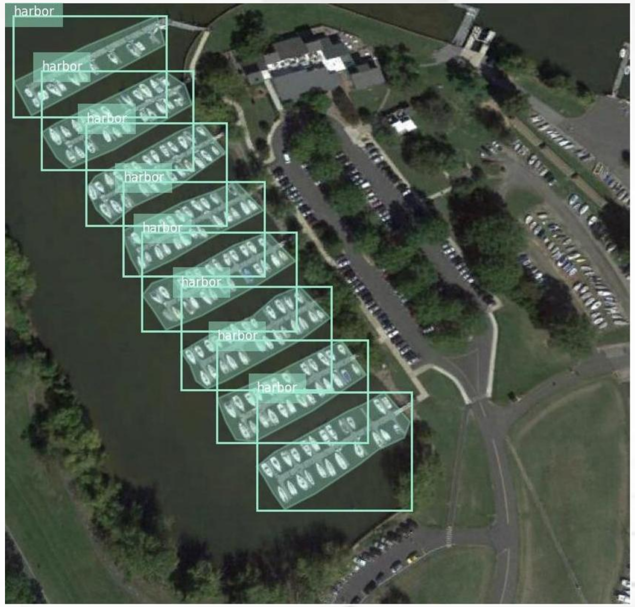
🤔 Concern: Can LVLMs count objects accurately?

✅ Answer: Yes — when given clear annotation rules, LVLMs follow them precisely.

Open-Vocabulary Counting (*mean F1-score*)

Method	NWPU	DIOR
Faster-RCNN	73	**81**
GPT-4o	67	72
GPT-4o (+instructions)	**83** (+16)	80 (+8)

✅ Example: harbor counting in NWPU



```
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}
```
✓

InstructSAM Runs Faster and Scales Better

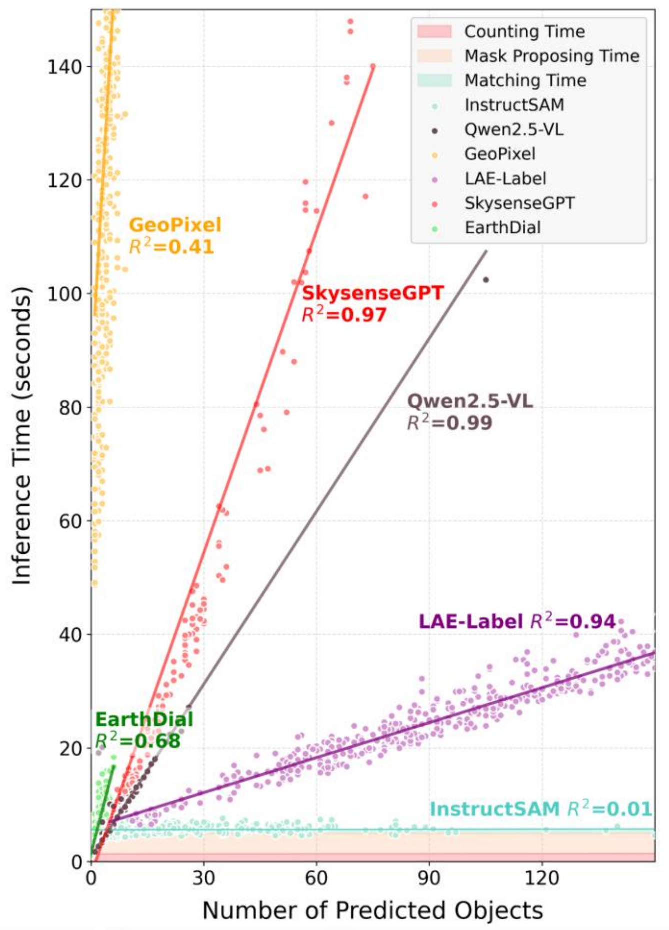
- 🚀 Counting is faster than detection
- ⚡ Uses 89% fewer tokens, 32% inference time reduction
- 📈 Inference time stays nearly constant as object count grows

Count the number of planes.

{"plane": 8} (6 tokens, 0.5s)

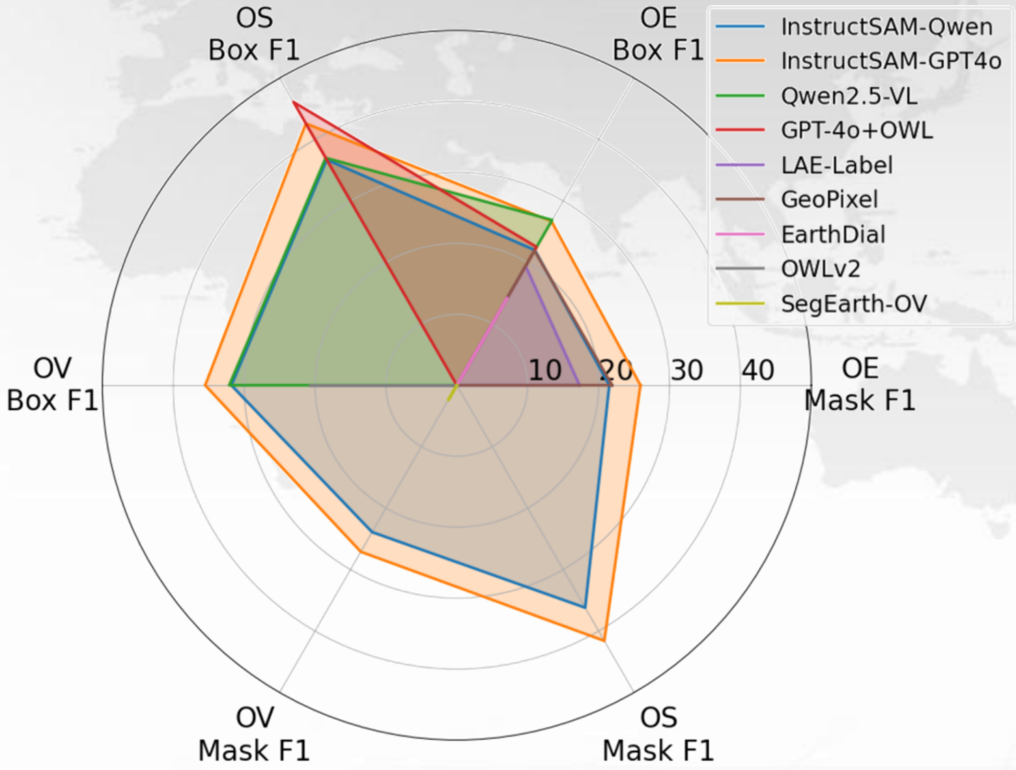
Detect all the planes.

```
[ {"bbox": [73, 40, 92, 91], "label": "plane"},
  {"bbox": [23, 56, 70, 60], "label": "plane"},
  {"bbox": [34, 47, 84, 97], "label": "plane"},
  {"bbox": [49, 48, 97, 82], "label": "plane"},
  {"bbox": [50, 40, 83, 99], "label": "plane"},
  {"bbox": [19, 21, 60, 70], "label": "plane"},
  {"bbox": [11, 32, 29, 56], "label": "plane"}]
  (183 tokens, 10s)
```



Inference time in open-ended setting on NWPU dataset

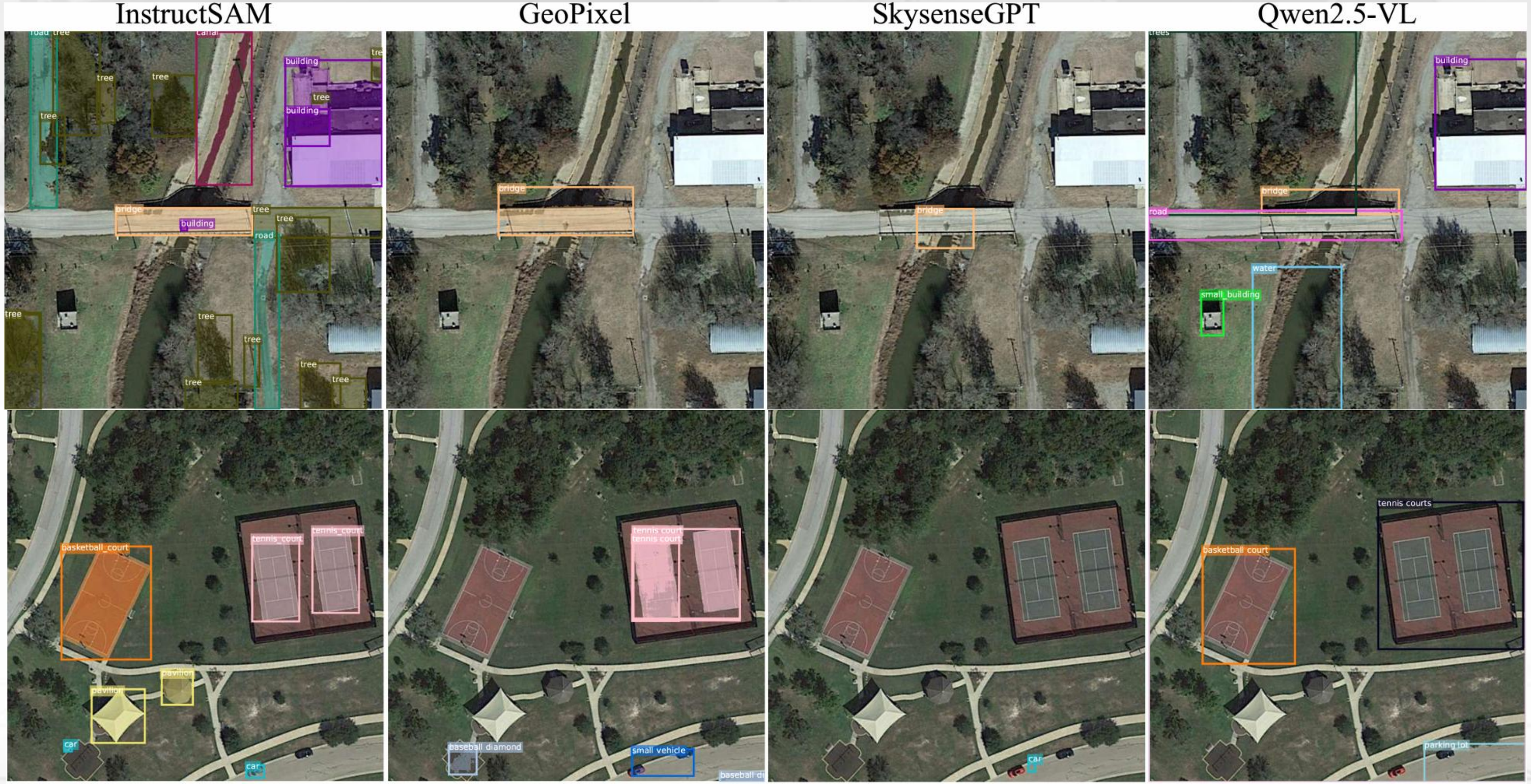
Zero-Shot Results across Three Settings



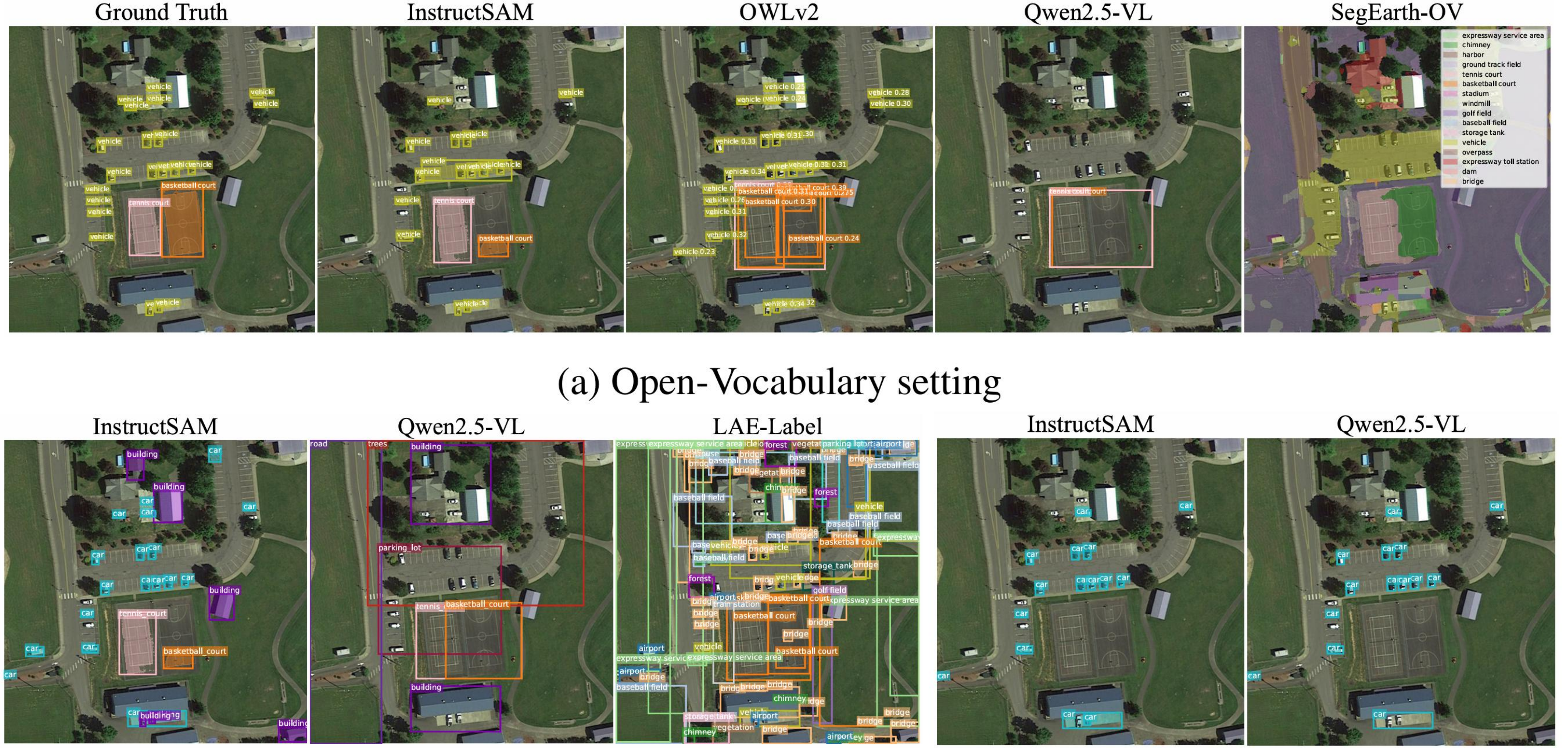
- ➢ Evaluated on NWPU and DIOR dataset
- 🚀 Strong performance across most settings
- 🤗 Performs well using open models (Qwen2.5-VL)

Qualitative Results: Open-Ended Setting

- ❌ Existing RSVLMs fail to generalize beyond training categories
- ✅ InstructSAM handles new classes more reliably, such as tree and pavilion.



Qualitative Results across Three Settings



(a) Open-Vocabulary setting

(b) Open-Ended setting

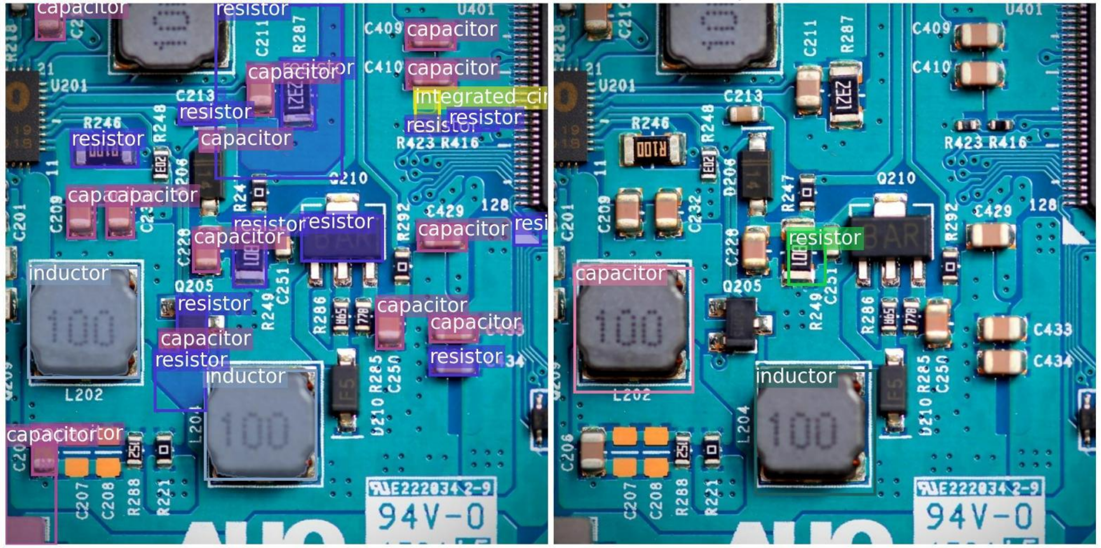
(c) Open-Subclass setting

Generalization Beyond Remote Sensing

Instruction:

Detect all the electronic components.

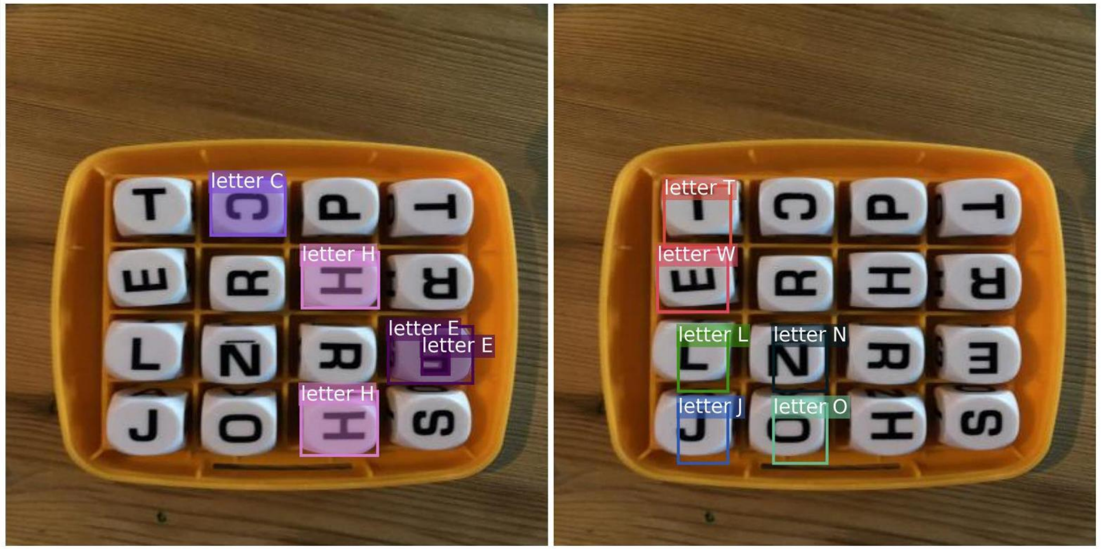
InstructSAM Qwen2.5-VL



Instruction:

Detect the dice whose letters come before K.

InstructSAM Qwen2.5-VL



Key Takeaways

🔑 Key Takeaways

- **Flexible**: Works with diverse user instructions.
- **Efficient**: Faster than detection, saves tokens, and scales well.
- **Training-Free**: Can directly benefit from stronger open-source or proprietary models.

🚀 Future Work

- Expand InstructSAM to standard segmentation tasks (e.g., land cover mapping).
- Build stronger Remote Sensing Foundation Models.

Code & arXiv

Personal Page

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