



Urban Socio-Semantic Segmentation with Vision-Language Reasoning

Yu Wang

2026/2/20

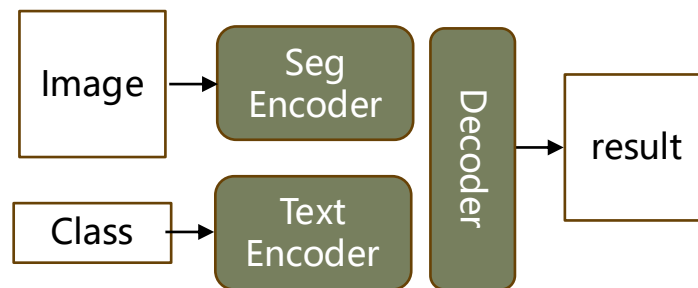
Background

Background: Segmentation

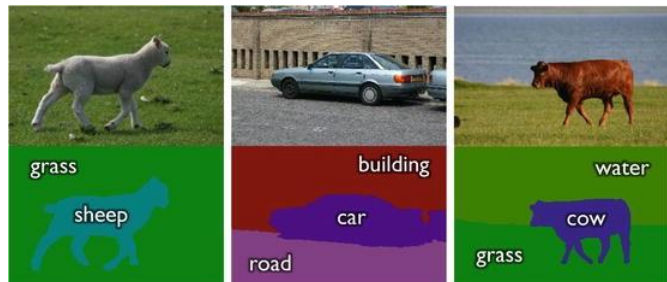
Regular Segmentation



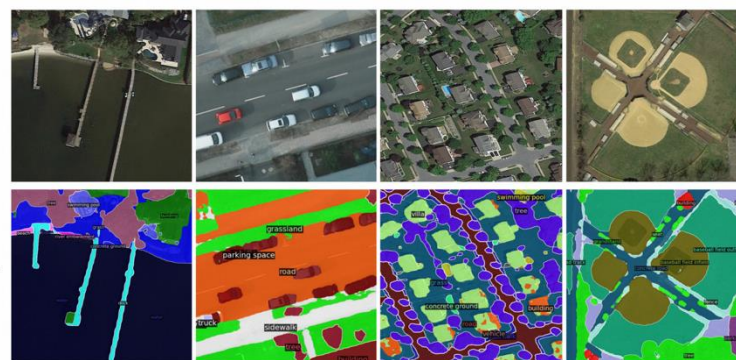
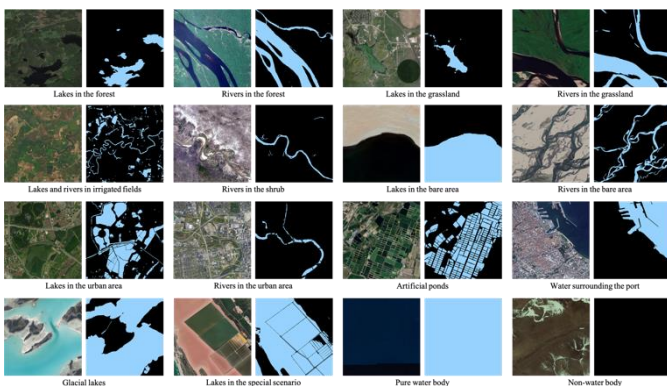
OV/Reffing Segmentation



Reasoning Segmentation



Can you segment the food that tastes not spicy in this image?

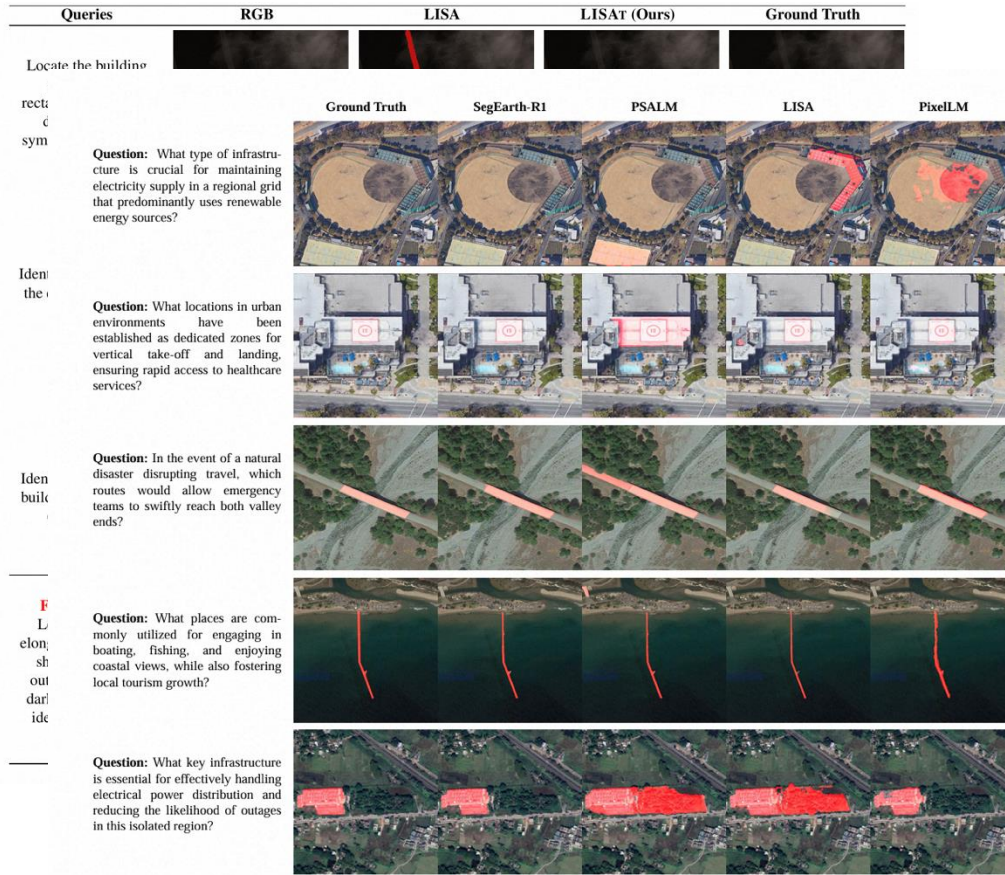


Which facility is primarily responsible for the treatment and purification of wastewater before it's discharged into local rivers?



Background

Background: Segmentation in RS



EarthReason [1] /GRES [2] Dataset

GRES Dataset (gpt4v generates visual features)

1. Find the building with the **dark** roof and **symmetrical** window pattern
2. Identify the **facility** in the middle left of the image
3. Identify the **damaged** building in the center of the image

EarthReason Dataset (gpt4v generates reason tasks)

1. In a regional power grid that primarily uses renewable energy, what types of infrastructure are critical to maintaining power supply? — **Solar panels**
2. What areas are commonly used for boating, fishing, and coastal views, promoting local tourism? — **Marinas**
3. What critical infrastructure is essential to efficiently handle power distribution and reduce the likelihood of power outages in this remote area? — **Power stations**

- ❖ **The objects of reasoning are largely determined by **physical properties**.**
- ❖ **Reason tasks are AI-generated**

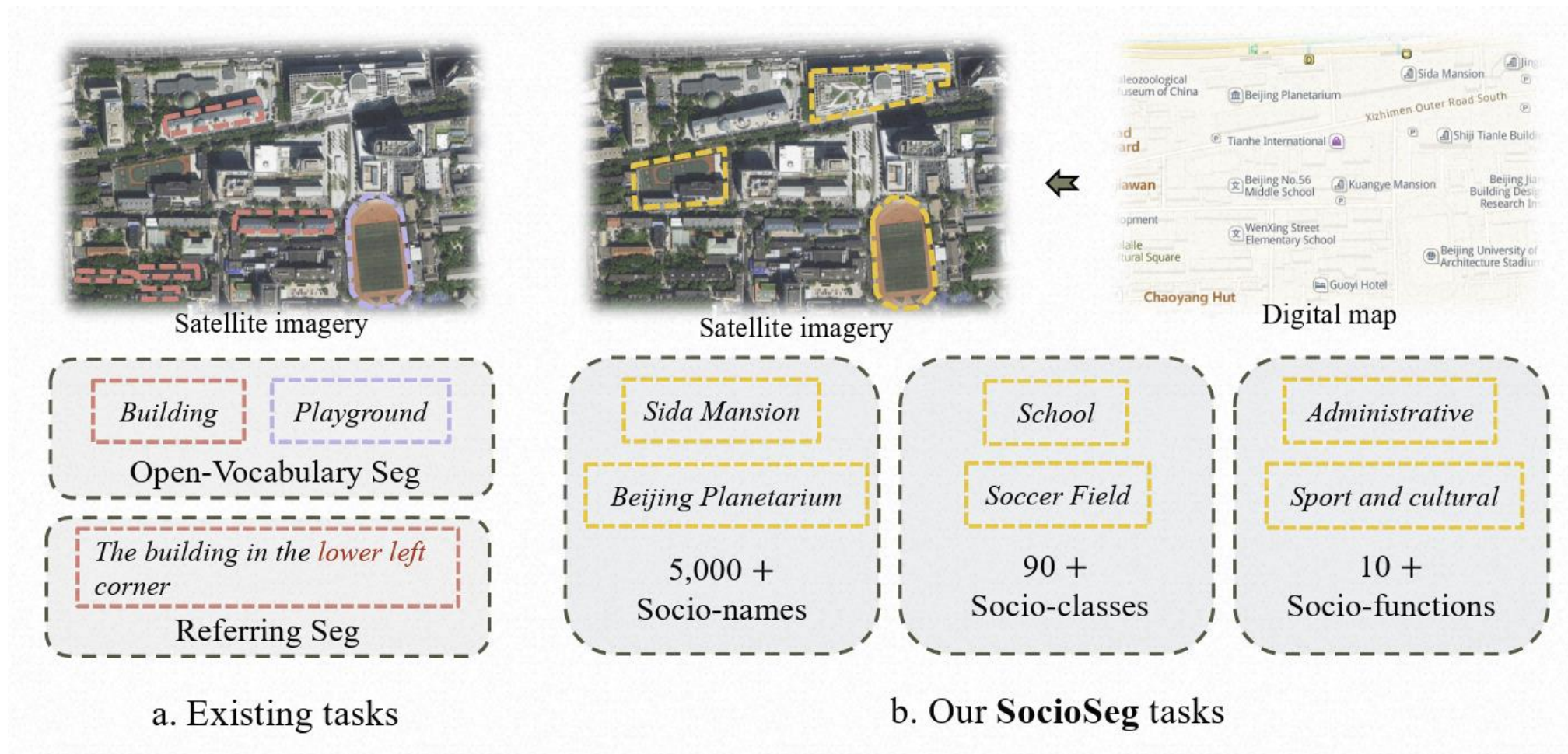
[1] Li, Kaiyu, et al. "Segearth-r1: Geospatial pixel reasoning via large language model." *arXiv preprint arXiv:2504.09644* (2025).

[2] Quenum, Jerome, et al. "LISAT: Language-Instructed Segmentation Assistant for Satellite Imagery." *arXiv preprint arXiv:2505.02829* (2025).

[3] Yao, Liang, et al. "RemoteReasoner: Towards Unifying Geospatial Reasoning Workflow." *arXiv preprint arXiv:2507.19280* (2025).

Motivation

Motivation1: Evolving semantic segmentation from physical properties to social attributes.



Motivation

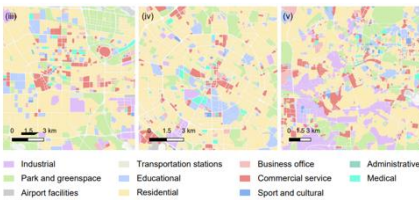
Motivation2: Leveraging MLLM's reasoning capabilities for social understanding



Task 1: Socio-name

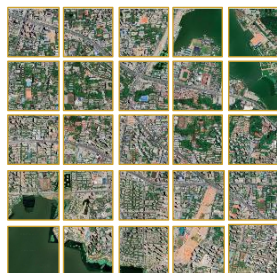


Task 2: Socio-class



Task 3: Socio-function

Existing paradigm



Aerial Imagery

+



Road Network



POI



Supervised model

Reasoning segmentation paradigm



Aerial Imagery

+



Map



MLLM

- ❖ Easy data access
- ❖ Spatial alignment
- ❖ End-to-end

Motivation

Motivation3: Further harnessing RL training for performance optimization.

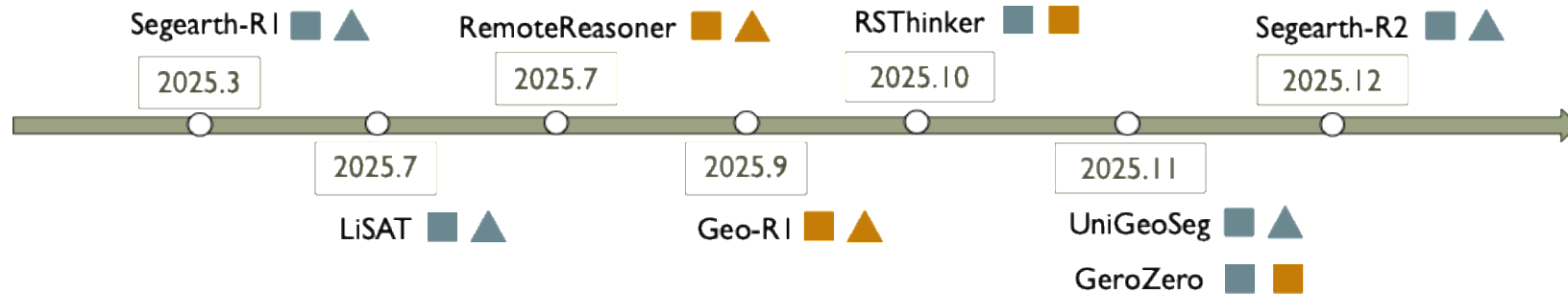


Figure 1. Existing Reasoning Segmentation Methods

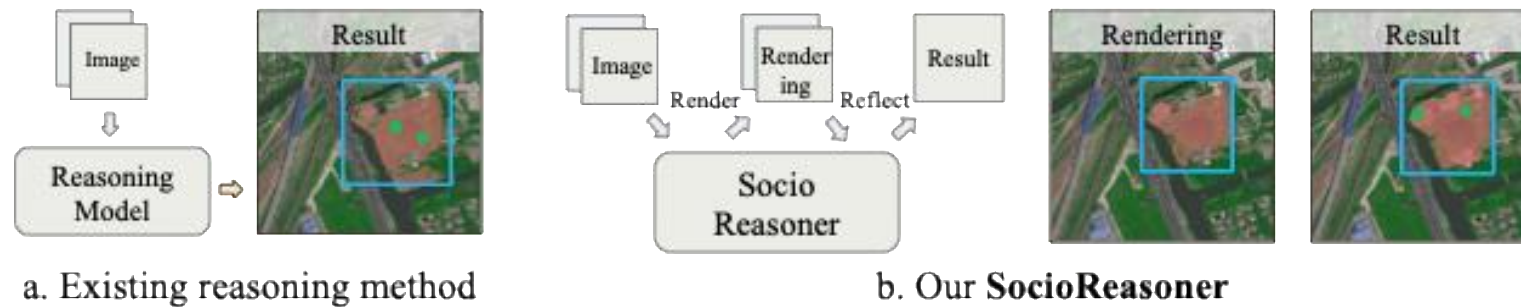


Figure 2. Our RL-based Render-Reflect mechanism

Dataset and Method

SocioSeg: Urban Socio-Semantic Segmentation dataset

Task1: Socio-name

Wuhan University
Amap Headquarter

Task2: Socio-class

College
Company

Task3: Socio-function

Educational
Business Office



(a) Residential area



(b) Industrial Park



(c) Primary and secondary school



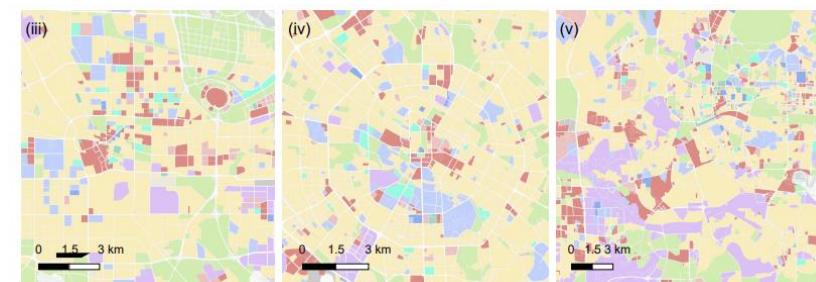
(d) Enterprise factory



(e) Office building



(f) Star hotel

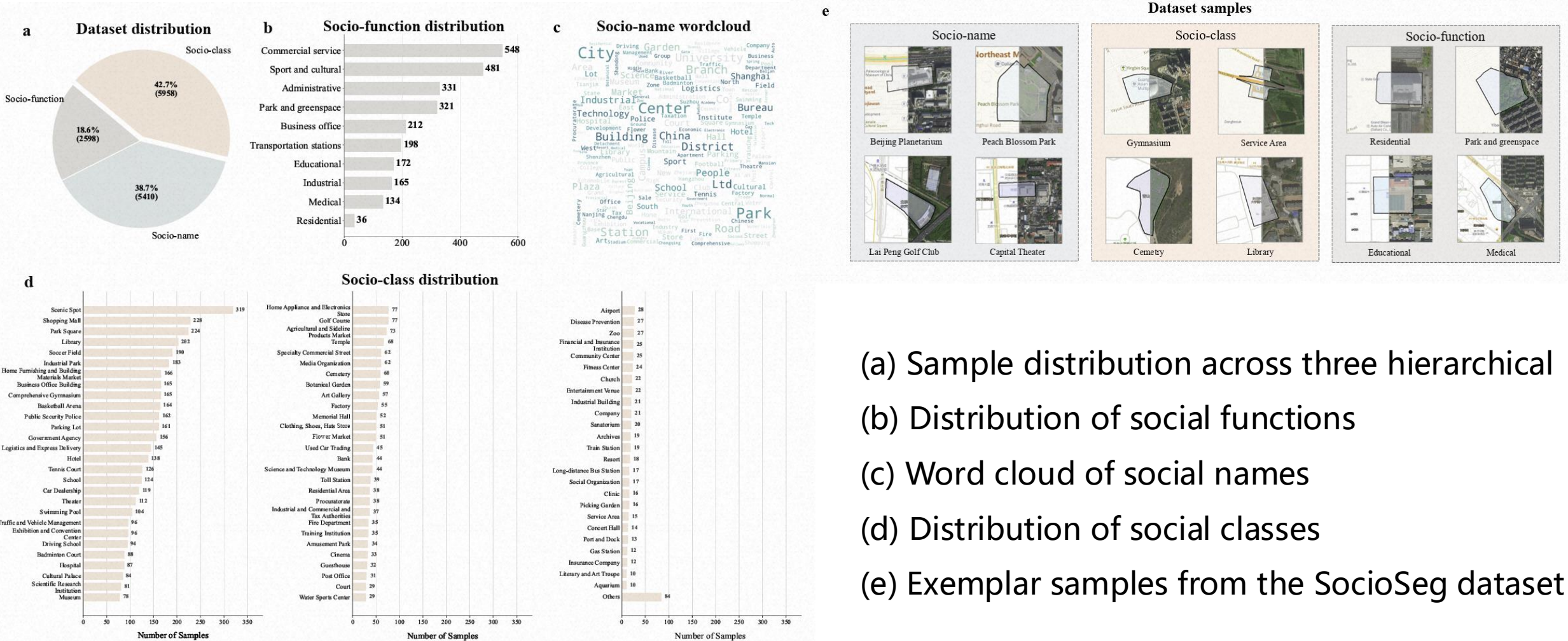


Concept: from concrete to abstract.

Semantic: from lower to higher.

Dataset and Method

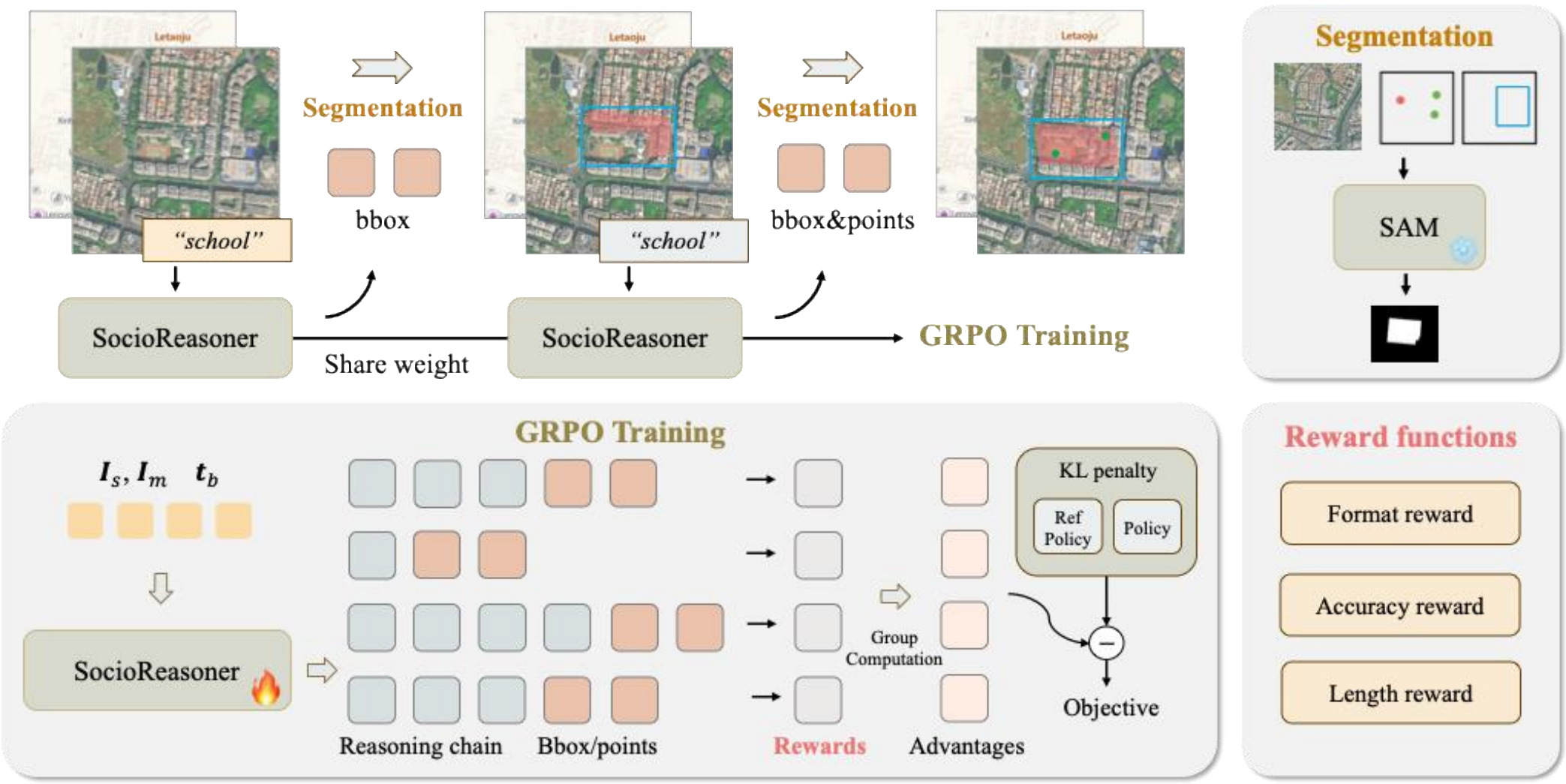
SocioSeg: Urban Socio-Semantic Segmentation dataset



(a) Sample distribution across three hierarchical
 (b) Distribution of social functions
 (c) Word cloud of social names
 (d) Distribution of social classes
 (e) Exemplar samples from the SocioSeg dataset

Dataset and Method

Method—SocioReasoner Model





Results

Method	Socio-name			Socio-class			Socio-function			All dataset		
	cIoU	gIoU	F1	cIoU	gIoU	F1	cIoU	gIoU	F1	cIoU	gIoU	F1
UNet	10.9	9.4	8.0	12.6	11.9	11.2	11.1	10.6	10.4	11.7	10.7	10.0
Segformer	22.0	19.6	18.1	22.4	21.4	19.5	21.4	20.2	17.9	22.1	20.5	18.7
VisionReasoner	<u>48.5</u>	<u>50.9</u>	<u>58.4</u>	<u>44.4</u>	<u>49.3</u>	<u>55.5</u>	36.3	41.8	45.0	<u>44.0</u>	<u>48.5</u>	<u>54.3</u>
Seg-R1	46.0	48.1	50.4	40.4	44.7	45.2	34.5	39.5	36.5	41.0	45.0	45.2
SAM-R1	25.6	25.4	37.2	22.3	23.8	32.1	17.7	19.9	25.2	22.5	23.7	32.4
SegEarth-OV	3.3	3.3	0.0	3.8	3.8	0.0	4.2	4.2	0.0	3.7	3.7	0.0
RSRefSeg	27.1	25.4	30.9	30.7	30.6	35.3	28.7	28.8	30.8	29.0	28.3	32.8
SegEarth-R1	36.9	42.1	46.9	38.9	45.1	50.0	<u>39.5</u>	<u>45.6</u>	<u>47.4</u>	38.3	44.1	48.4
RemoteReasoner	46.6	49.5	56.1	42.9	48.0	53.9	38.0	43.5	47.2	43.2	47.7	53.3
Ours	52.6	55.7	64.6	47.6	52.8	60.1	40.6	46.9	50.3	47.9	52.8	59.7

Table 1: Comparison with state-of-the-art methods on SocioSeg test set, split by task groups for readability. The best performance in each column is highlighted in **bold**. The second best is underlined. Baselines are re-trained on the SocioSeg training split to ensure fair comparison.

Results

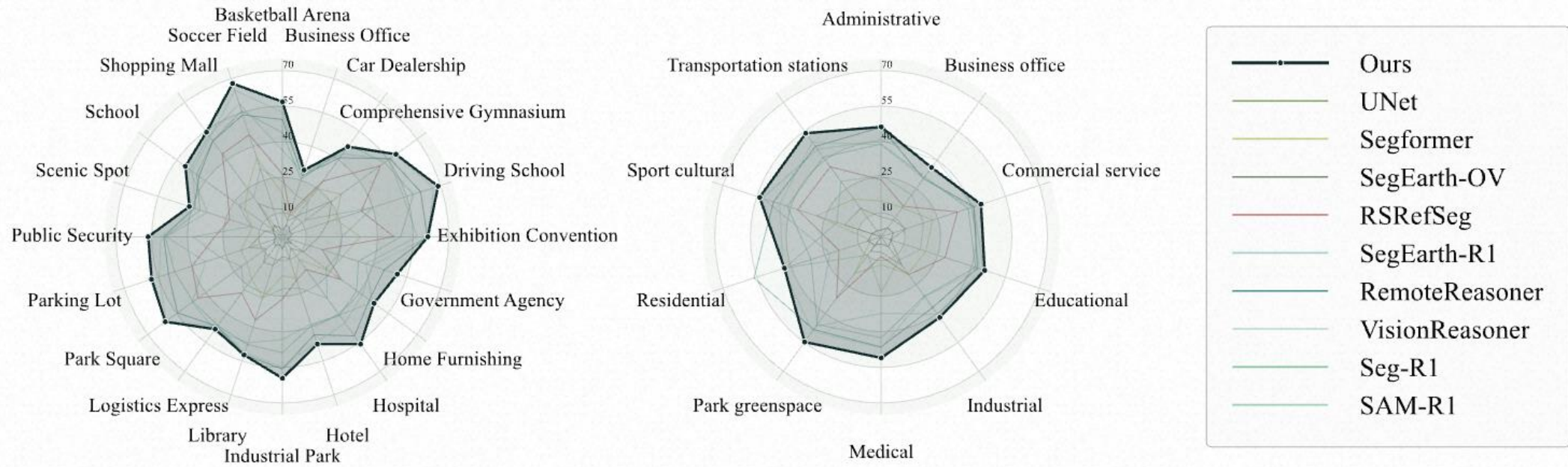


Figure 4: Per-class accuracy comparison across Socio-classes and Socio-functions. We select the top-20 most frequent Socio-classes (left) and all Socio-functions(right) for visualization.

Table 5: All ablation of multi-stage.

Method	Socio-name			Socio-class			Socio-function			All dataset		
	cIoU	gIoU	F1	cIoU	gIoU	F1	cIoU	gIoU	F1	cIoU	gIoU	F1
w/o reflection	48.5	50.9	58.4	44.4	49.3	55.5	36.3	41.8	45.0	44.0	48.5	54.3
w/o refinement	50.5	53.1	61.2	46.2	51.0	58.1	40.3	45.7	48.1	46.4	50.8	57.5
Ours	52.6	55.7	64.6	47.6	52.8	60.1	40.6	46.9	50.3	47.9	52.8	59.7

Table 6: All ablation of point number.

Method	Socio-name			Socio-class			Socio-function			All dataset		
	cIoU	gIoU	F1	cIoU	gIoU	F1	cIoU	gIoU	F1	cIoU	gIoU	F1
1 point refinement	51.6	53.4	61.2	47.6	51.2	59.0	40.0	45.7	49.5	47.6	51.2	58.0
2 points refinement	52.6	55.7	64.6	47.6	52.8	60.1	40.6	46.9	50.3	47.9	52.8	59.7
3 points refinement	53.2	54.7	65.0	48.9	52.6	59.7	41.8	46.6	49.8	48.9	52.3	58.8

Results

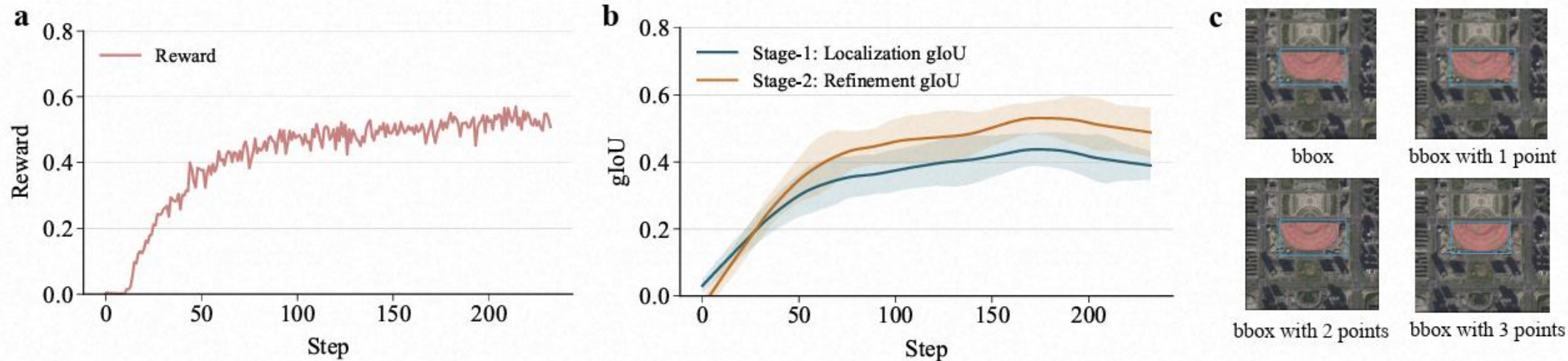


Figure 5: (a) Sum reward during training. It shows the sum of rewards across training steps in the two-stage workflow. (b) Multi-stage gIoU during training. It shows the gIoU improvement across training steps in the two-stage workflow. (c) Different number of points. It visualizes the result of SocioReasoner in the refinement stage with different numbers of points.

Results

Table 4: Generalization of SocioReasoner, where ID and OOD refer to in-domain and out-of-domain, respectively.

Method	ID			OOD (Map Style)			OOD (New Region)		
	cIoU	gIoU	F1	cIoU	gIoU	F1	cIoU	gIoU	F1
VisionReasoner (SFT)	44.1	47.2	52.1	38.8	40.9	45.5	22.5	24.7	26.1
VisionReasoner (RL)	44.0	48.5	54.3	42.0	44.4	51.2	32.8	34.4	35.0
Ours (SFT)	<u>47.1</u>	<u>51.4</u>	<u>57.8</u>	39.7	42.0	46.9	30.1	32.3	31.5
Ours (RL)	47.9	52.8	59.7	45.1	49.1	57.7	40.2	43.4	42.9

Table 8: Inference time comparison (seconds per sample).

VisionReasoner	Seg-R1	SAM-R1	RSRefSeg	SegEarth-R1	RemoteReasoner	Ours(rl)	Ours(sft)
1.33	1.07	2.52	0.16	0.35	1.13	2.71	0.41



Out-of-domain

Results

Ours Stage-1

“Jinan Zoo”

<think> The first image shows a map with various locations marked, including 'Jinan Zoo'. The second image is a satellite view of a large area with buildings and roads clearly visible, including the Jinan Zoo area. **</think>**

```
[{"bbox_2d": [360, 268, 538, 383]}
```

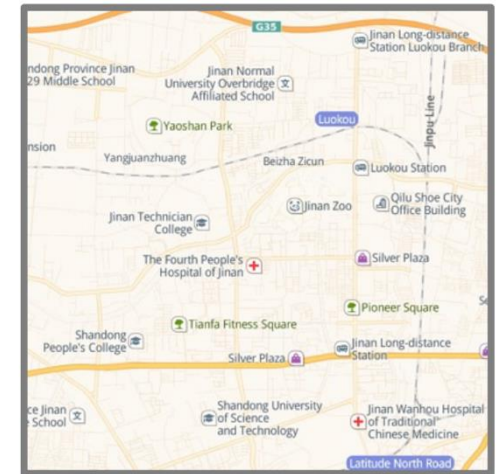


Ours Stage-2

“Jinan Zoo”

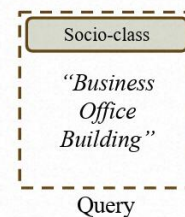
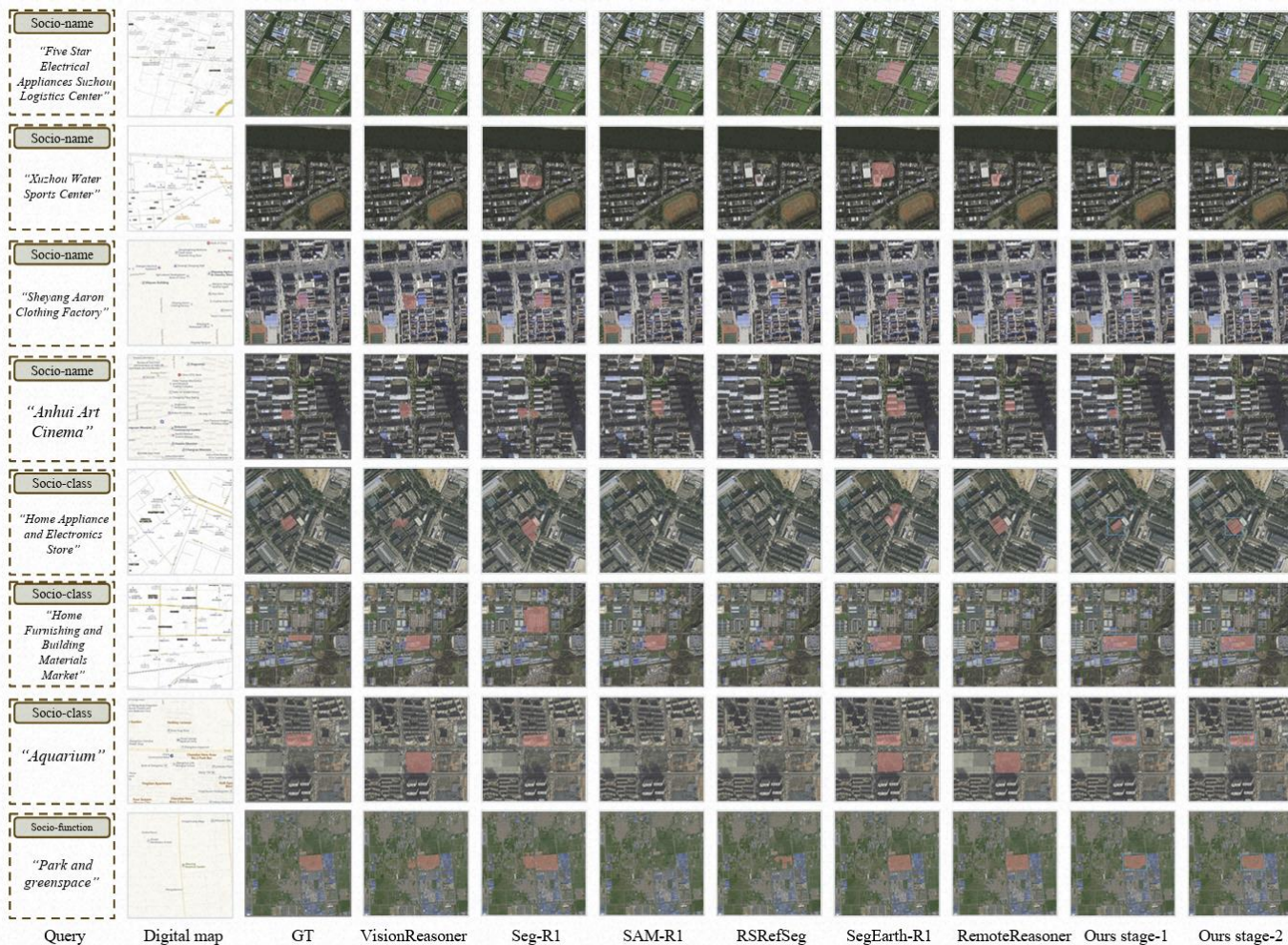
<think>The bounding box (bbox) with coordinates [360, 268, 538, 383] is likely to be the area of interest. The points added to the bbox help to better represent the area of interest by highlighting the key features of the zoo, such as the main building and surrounding greenery. **</think>**

```
[{"bbox_2d": [360, 268, 538, 383], "points": [[452, 287], [487, 306]]}]
```

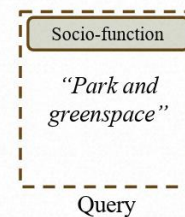


Digital Map

Results



Digital map



Digital map

Ours Stage-1

<think> The map image shows various buildings and roads labeled. The satellite image provides a broader perspective of the same area, showing multiple buildings and structures. The "Business Building" in the map image is likely a key feature of the business district in the area. </think>
[{"bbox_2d": [290, 393, 460, 466]]



Ours Stage-2

<think> The bounding box (bbox) with coordinates [290, 393, 460, 466] is likely to be the area of interest. The points added to the bbox help to better represent the area of interest by highlighting the key features of the building, such as its architectural design and surrounding structures. </think>
[{"bbox_2d": [290, 393, 460, 466], "points": [[378, 400], [318, 431]]}]



Ours Stage-1

<think> The first image shows a map with various locations marked, including a park and greenspace. The satellite view provides a broader perspective of the area, showing the park and surrounding greenery. The map shows the same area with different details, such as road names and other geographical features. </think>
[{"bbox_2d": [356, 285, 526, 523]]



Ours Stage-2

<think> The bounding box (bbox) with coordinates [356, 285, 526, 523] is likely to be the area of interest. The points added to the bbox help to better represent the area of interest by highlighting the green and park-like features in the image. </think>
[{"bbox_2d": [356, 285, 526, 523], "points": [[520, 361], [403, 369]]}]

