OmniGeo: Interactive Vision Language Models for Multi-Granularity, Multi-Sensor and Multi-Scale Earth Observation

Yan Shu, Paolo Rota, Nicu Sebe
University of Trento

Introduction

	visual prompt (region, point)		image-level (RGB,SAR)		Pixel-level (RGB, SAR)	
GeoChat	×	×	√	×	×	×
EarthGPT	×	×	√	√	×	×
Skyeye-GPT	×	×	√	×	×	×
EarthMarker	√	√	√	×	×	×
EarthDial	×	×	√	√	×	×
RsUniVLM	×	×	√	×	√	×
Geopixel	×	×	√	×	√	×
OmniGeo (Ours)	√	√	√	√	√	√

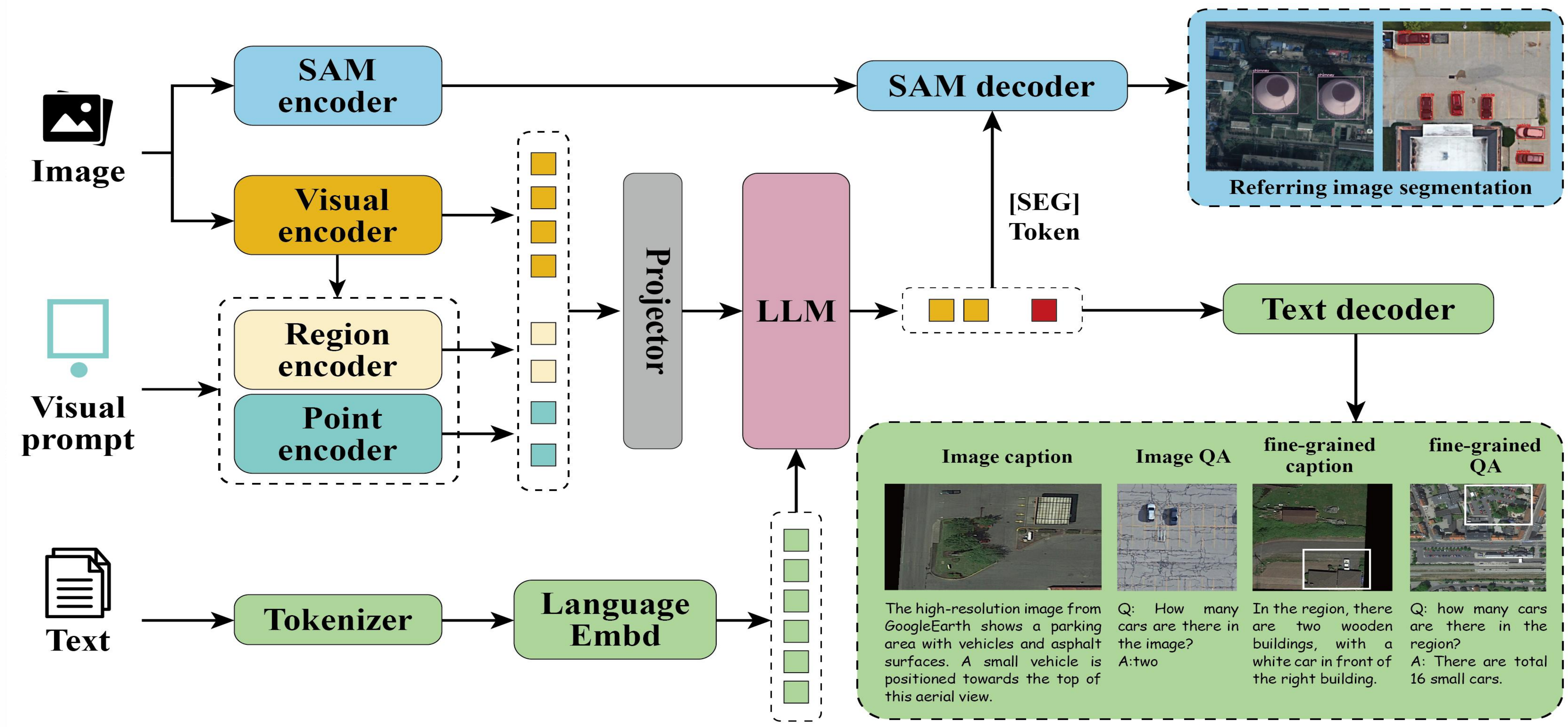
Motivation

- Existing Vision Languge Models (VLMs) has limited capacity in understanding and reasoning in RS domain
- **Interactive**: Accept different visual prompts to achive better interaction
- **Multi-Granularity**: Understand RS data from image-level to pixel-level
- **Multi-Sensor**: Understand RS data from differnet sensor (RGB and SAR)
- **Multi-Scale**: Understand RS data from different scales (drones or satellite)

Goals

- Proposing a MLLMs towards multi-granularity, mutli-sensor and multi-scale earth observation with flexible visual prompts.
- Proposing a compostite pretraining stage for better knowledge transfer from full-supervised natural image data to unsupervised RS data.
- Proposing a multi-task vision-language SAR data and benchmark.

Method



The high-resolution image from GoogleEarth shows a parking area with vehicles and asphalt surfaces. A small vehicle is positioned towards the top of this aerial view.

Q: How many cars are there in the image?
A:two

In the region, there are two wooden buildings, with a white car in front of the right building.

Q: how many cars are there in the region?
A: There are total 16 small cars.

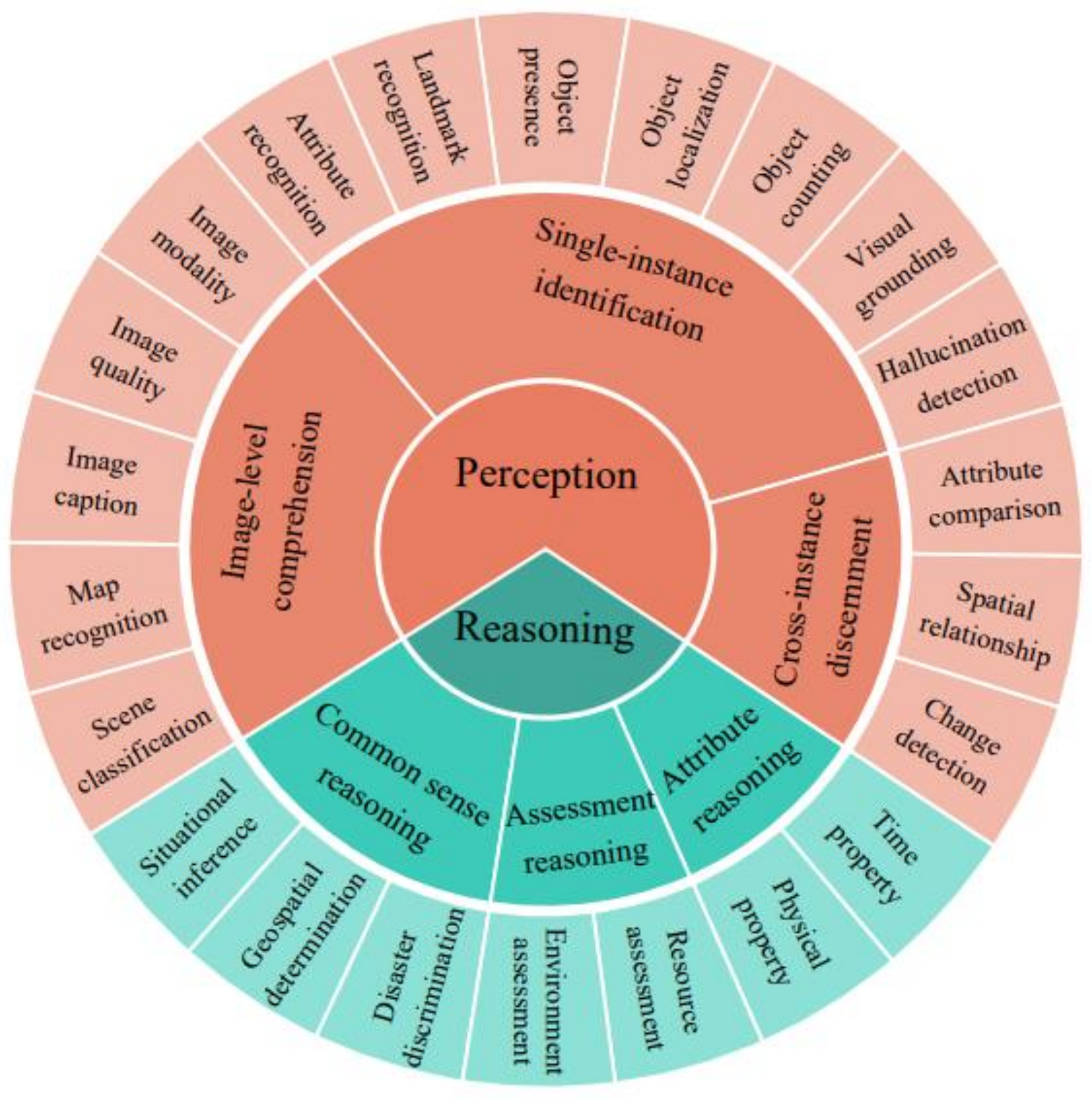
Experiment

Segmentation	RRSISD		RefSegrs	
	miou	oiou	miou	oiou
RRSIS (TGRS 2024)	-	-	60.0	76.8
RM-SIN (CVPR 2024)	64.2	77.8	-	-
CRO-BIM (Arxiv 2024.10)	64.5	76.0	59.7	72.3
OmniGeo (Ours)	**79.7**	**97.4**	**62.3**	**89.6**

Region caption task	METEOR	ROUGE	CIDEr	SPICE
GeoChat (CVPR 2024)	12.94	26.98	30.92	24.97
EarthGPT (TGRS 2024)	24.09	47.87	232.79	38.29
EarthMarker (TGRS 2024)	31.97	60.46	379.25	59.87
OmniGeo (Ours)	**36.55**	**68.53**	**491.20**	**67.62**

Future works

- **Benchmark construction, including perception and resaoning evaluation for mutli-sensor RS data.**
- **Composite pretraining stage ablation study.**
- **Extending to multi-spectrum and high-spectrum data.**



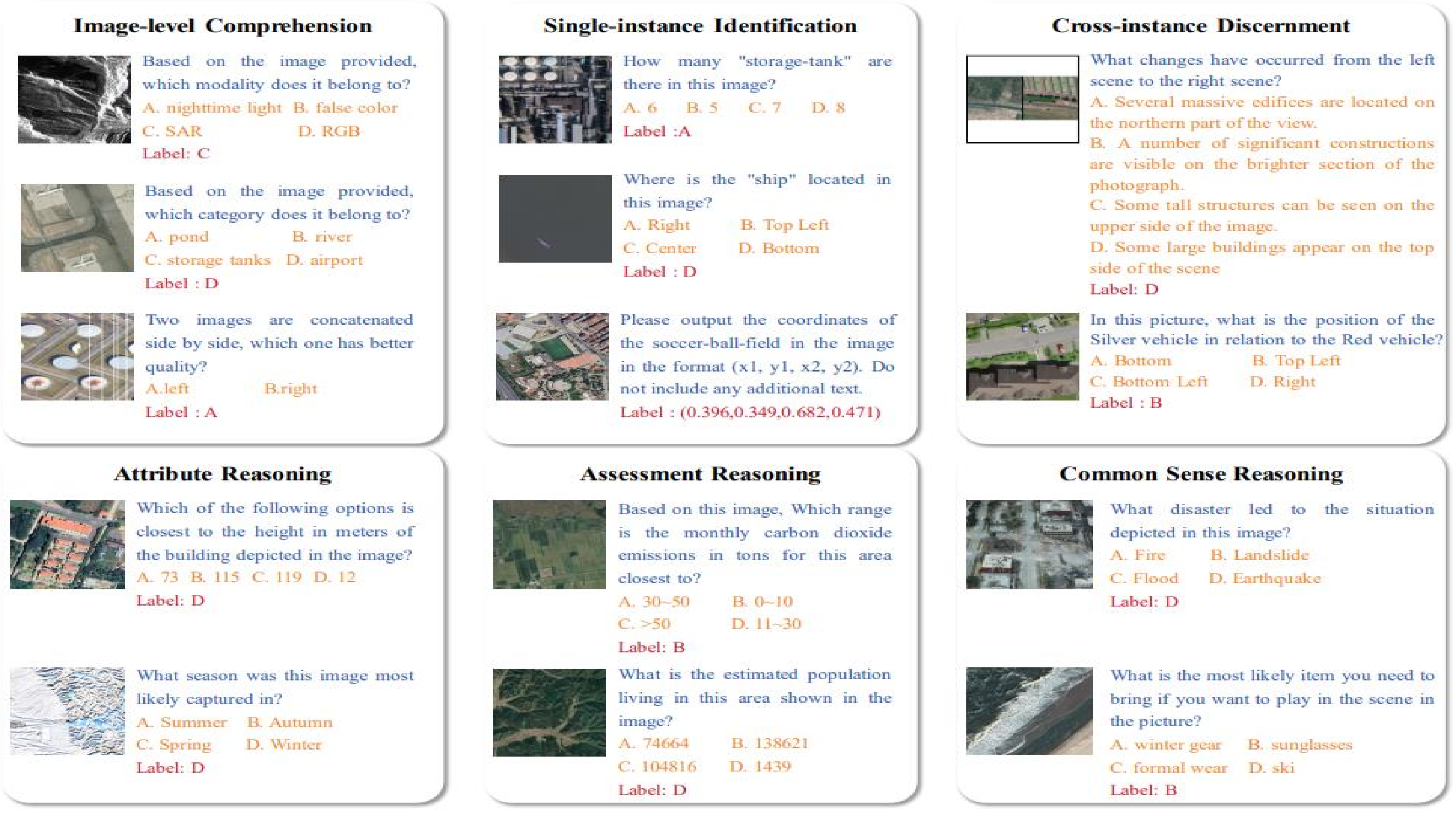


Image-level Comprehension

Based on the image provided, which modality does it belong to?
A. nighttime light B. false color
C. SAR D. RGB
Label: C

Based on the image provided, which category does it belong to?
A. pond B. river
C. storage tanks D. airport
Label : D

Two images are concatenated side by side, which one has better quality?
A.left B.right
Label : A

Single-instance Identification

How many "storage-tank" are there in this image?
A. 6 B. 5 C. 7 D. 8
Label :A

Where is the "ship" located in this image?
A. Right B. Top Left
C. Center D. Bottom
Label : D

Please output the coordinates of the soccer-ball-field in the image in the format (x1, y1, x2, y2). Do not include any additional text.
Label : (0.396,0.349,0.682,0.471)

Cross-instance Discernment

What changes have occurred from the left scene to the right scene?
A. Several massive edifices are located on the northern part of the view.
B. A number of significant constructions are visible on the brighter section of the photograph.
C. Some tall structures can be seen on the upper side of the image.
D. Some large buildings appear on the top side of the scene
Label: D

In this picture, what is the position of the Silver vehicle in relation to the Red vehicle?
A. Bottom B. Top Left
C. Bottom Left D. Right
Label : B

Attribute Reasoning

Which of the following options is closest to the height in meters of the building depicted in the image?
A. 73 B. 115 C. 119 D. 12
Label: D

What season was this image most likely captured in?
A. Summer B. Autumn
C. Spring D. Winter
Label: D

Assessment Reasoning

Based on this image, Which range is the monthly carbon dioxide emissions in tons for this area closest to?
A. 30~50 B. 0~10
C. >50 D. 11~30
Label: B

What is the estimated population living in this area shown in the image?
A. 74664 B. 138621
C. 104816 D. 1439
Label: D

Common Sense Reasoning

What disaster led to the situation depicted in this image?
A. Fire B. Landslide
C. Flood D. Earthquake
Label: D

What is the most likely item you need to bring if you want to play in the scene in the picture?
A. winter gear B. sunglasses
C. formal wear D. ski
Label: B