UrbanLLaVA: A Multi-modal Large Language Model for Urban Intelligence

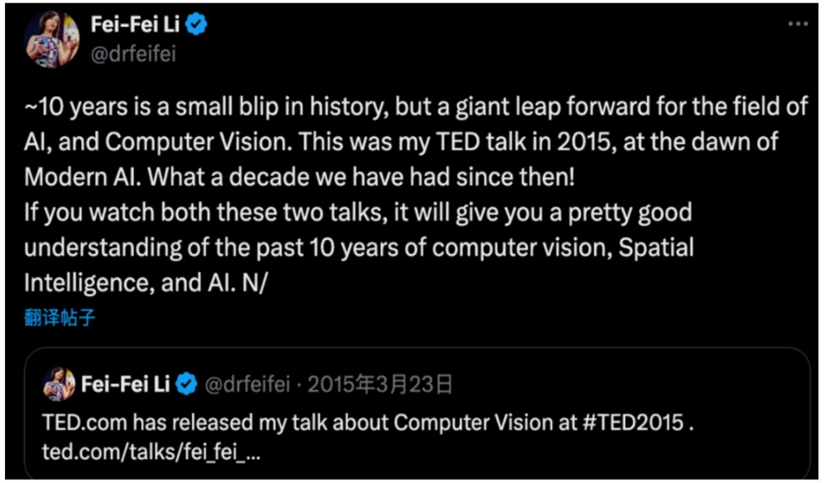
Jie Feng
Department of Electronic Engineering
Tsinghua University
https://vonfeng.github.io/
2025.10

Arxiv

Github

Spatial Intelligence

- **Spatial Intelligence—the capacity to understand, reason, and act on spatial information—is the foundation of embodied AI and a key prerequisite for AGI.**

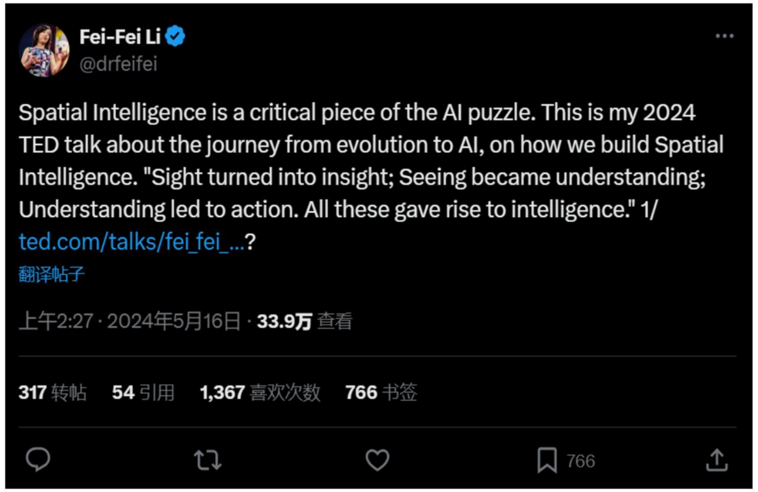


Fei-Fei Li
@drfeifei

~10 years is a small blip in history, but a giant leap forward for the field of AI, and Computer Vision. This was my TED talk in 2015, at the dawn of Modern AI. What a decade we have had since then!
If you watch both these two talks, it will give you a pretty good understanding of the past 10 years of computer vision, Spatial Intelligence, and AI. N/

翻译帖子

Fei-Fei Li @drfeifei · 2015年3月23日
TED.com has released my talk about Computer Vision at #TED2015 .
ted.com/talks/fei_fei_...



Fei-Fei Li
@drfeifei

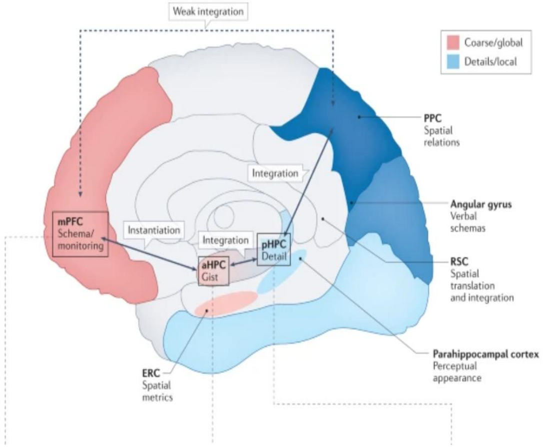
Spatial Intelligence is a critical piece of the AI puzzle. This is my 2024 TED talk about the journey from evolution to AI, on how we build Spatial Intelligence. "Sight turned into insight; Seeing became understanding; Understanding led to action. All these gave rise to intelligence." 1/
ted.com/talks/fei_fei_...?

翻译帖子

上午2:27 · 2024年5月16日 · **33.9万** 查看

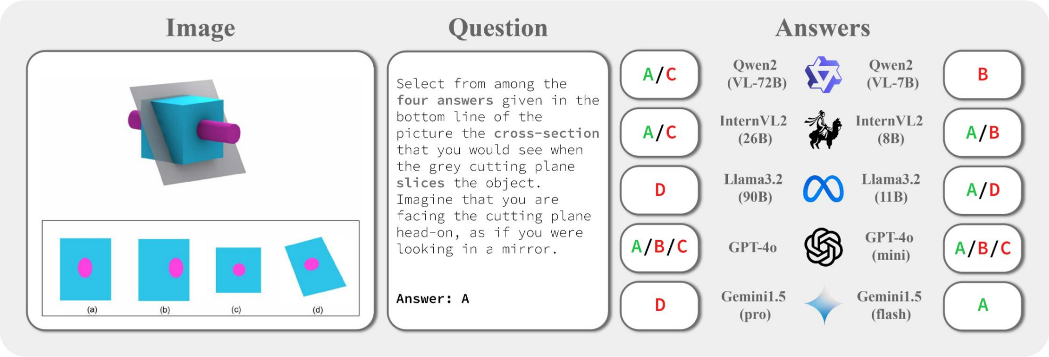
317 转帖 **54** 引用 **1,367** 喜欢次数 **766** 书签

- **Spatial Intelligence is also for Cognitive Science, Urban Science, Complex Systems, Earth Science**



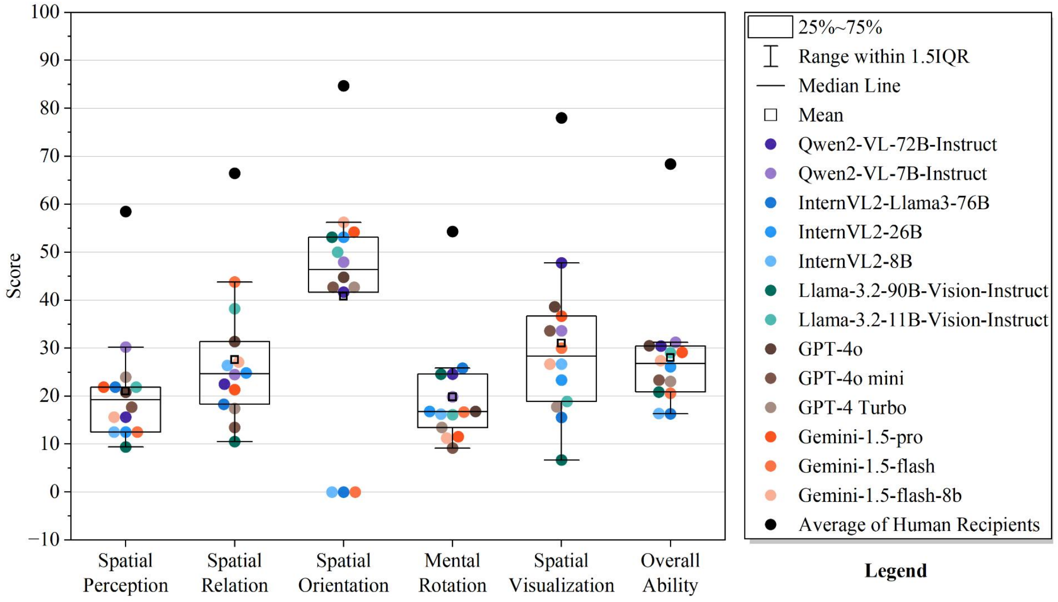
Spatial Intelligence → Urban Intelligence

- **General LLMs have significant limitations in spatial knowledge, understanding, and reasoning.**



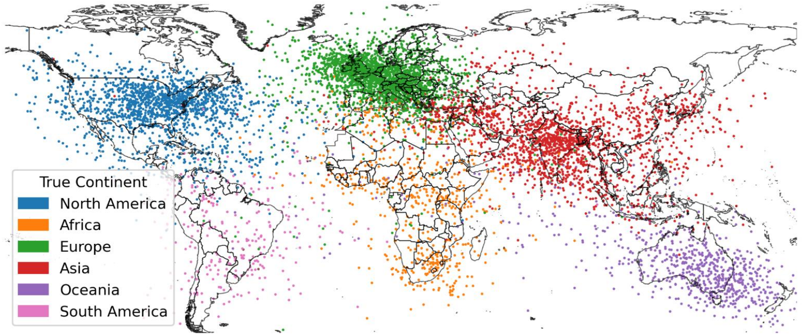
Select from among the four answers given in the bottom line of the picture the cross-section that you would see when the grey cutting plane slices the object. Imagine that you are facing the cutting plane head-on, as if you were looking in a mirror.

Answer: A



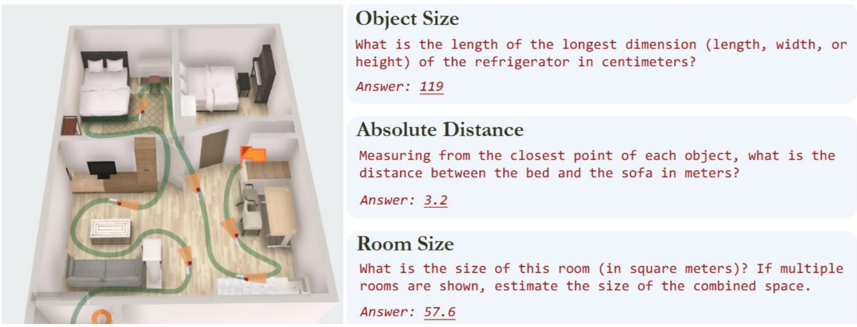
Xu, Wenrui*, Dalin Lyu*, Weihang Wang*, **Jie Feng**, Chen Gao, and Yong Li. "Defining and Evaluating VLMs' Basic Spatial Abilities: A Perspective from Psychometrics." ACL 2025 Main.

World Level



Urban Level

Embodied Level



Object Size

What is the length of the longest dimension (length, width, or height) of the refrigerator in centimeters?

Answer: 119

Absolute Distance

Measuring from the closest point of each object, what is the distance between the bed and the sofa in meters?

Answer: 3.2

Room Size

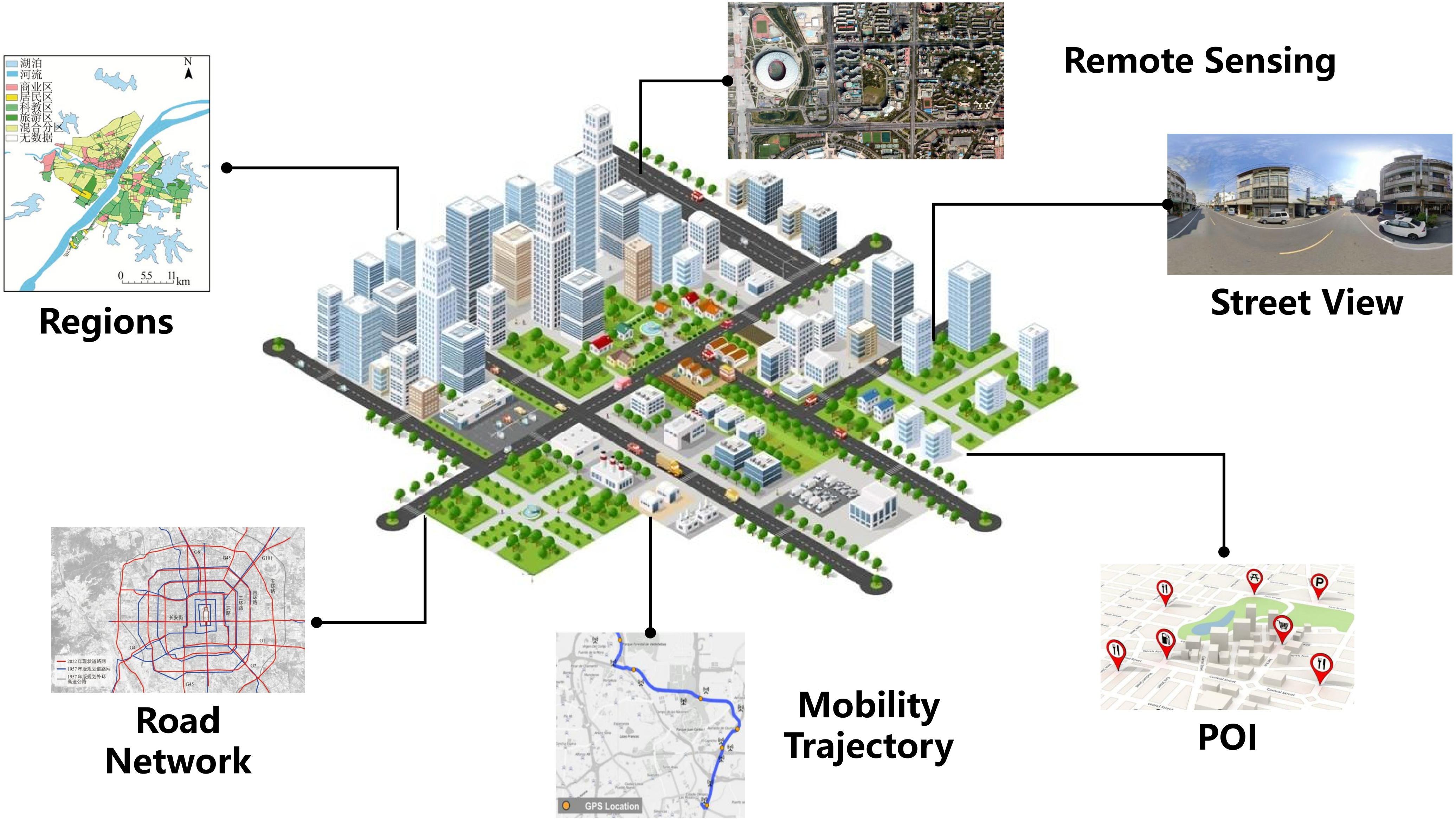
What is the size of this room (in square meters)? If multiple rooms are shown, estimate the size of the combined space.

Answer: 57.6

Gurnee, Wes, and Max Tegmark. "Language models represent space and time." ICLR 2024.

Urban Intelligence

- **We need strengthen LLMs spatial intelligence at Urban scale for embodied and generalized AI.**



Remote Sensing

Regions

Street View

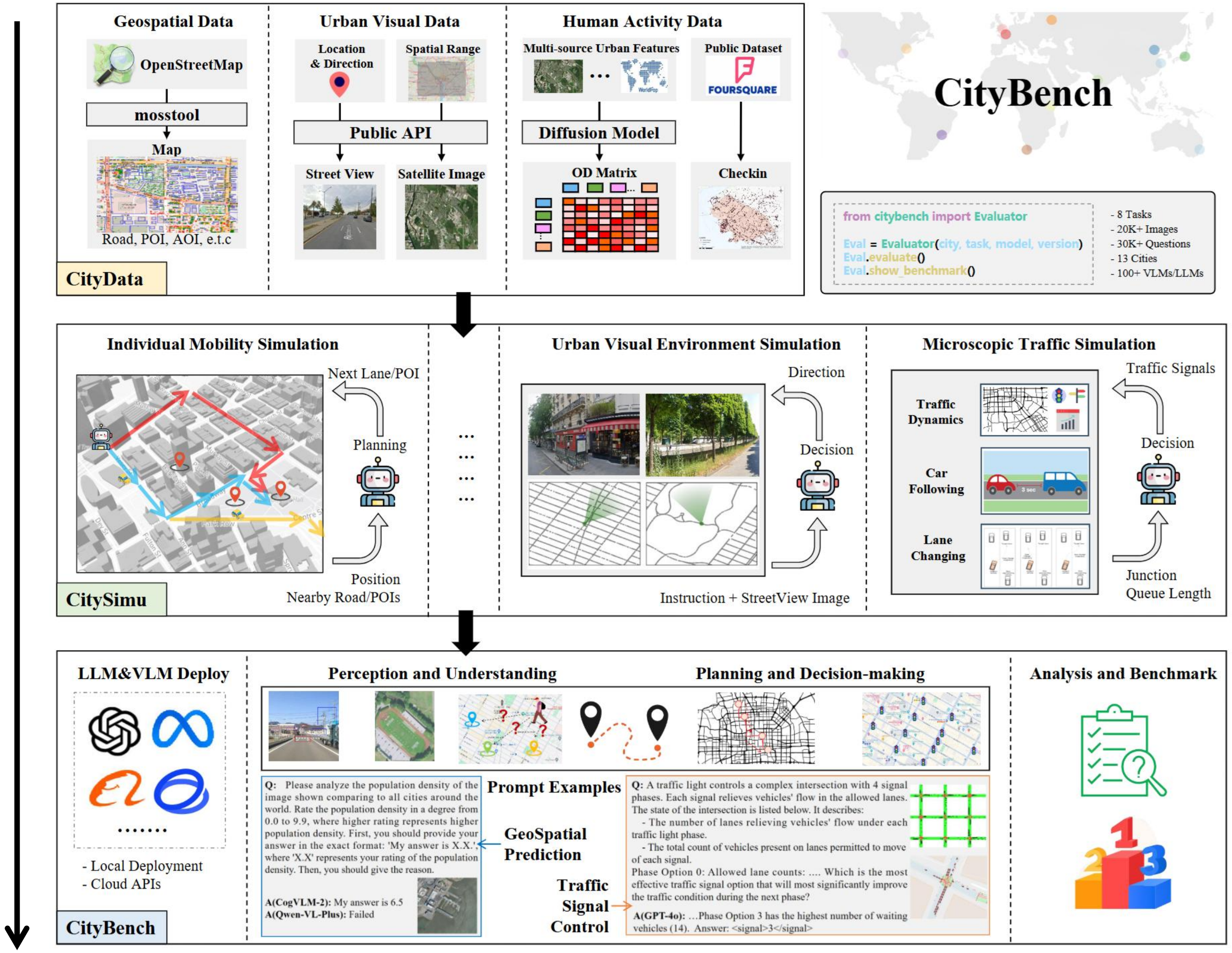
Road Network

Mobility Trajectory

POI

Before UrbanLLaVA: CityBench

- **A comprehensive urban spatial intelligence benchmark**



```
from citybench import Evaluator

Eval = Evaluator(city, task, model, version)
Eval.evaluate()
Eval.show_benchmark()
```

- 8 Tasks
- 20K+ Images
- 30K+ Questions
- 13 Cities
- 100+ VLMs/LLMs

Q: Please analyze the population density of the image shown comparing to all cities around the world. Rate the population density in a degree from 0.0 to 9.9, where higher rating represents higher population density. First, you should provide your answer in the exact format: 'My answer is X.X.', where 'X.X' represents your rating of the population density. Then, you should give the reason.

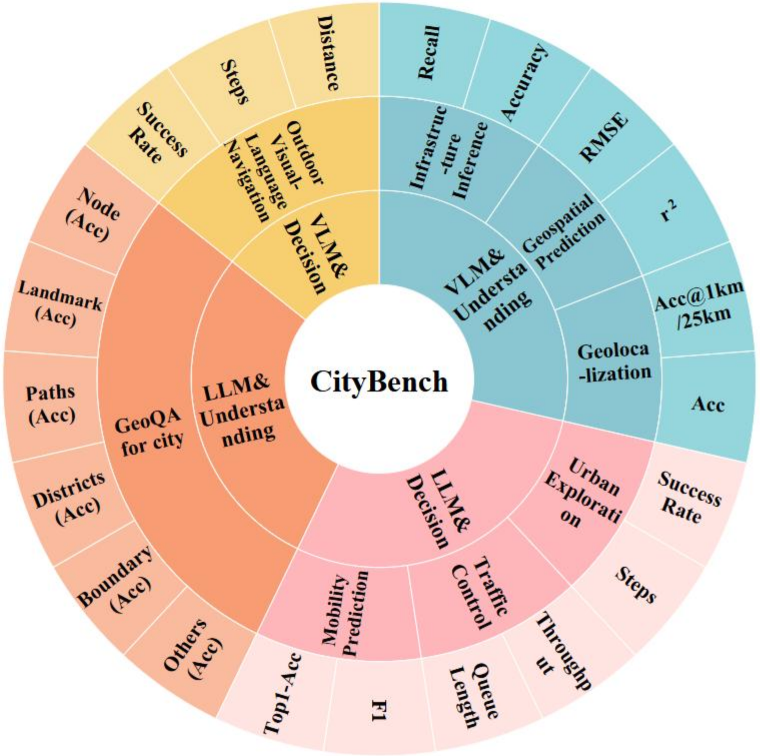
A(CogVLM-2): My answer is 6.5
A(Qwen-VL-Plus): Failed

Q: A traffic light controls a complex intersection with 4 signal phases. Each signal relieves vehicles' flow in the allowed lanes. The state of the intersection is listed below. It describes:

- The number of lanes relieving vehicles' flow under each traffic light phase.
- The total count of vehicles present on lanes permitted to move of each signal.

Phase Option 0: Allowed lane counts: .... Which is the most effective traffic signal option that will most significantly improve the traffic condition during the next phase?

A(GPT-4o): …Phase Option 3 has the highest number of waiting vehicles (14). Answer: <signal>3</signal>

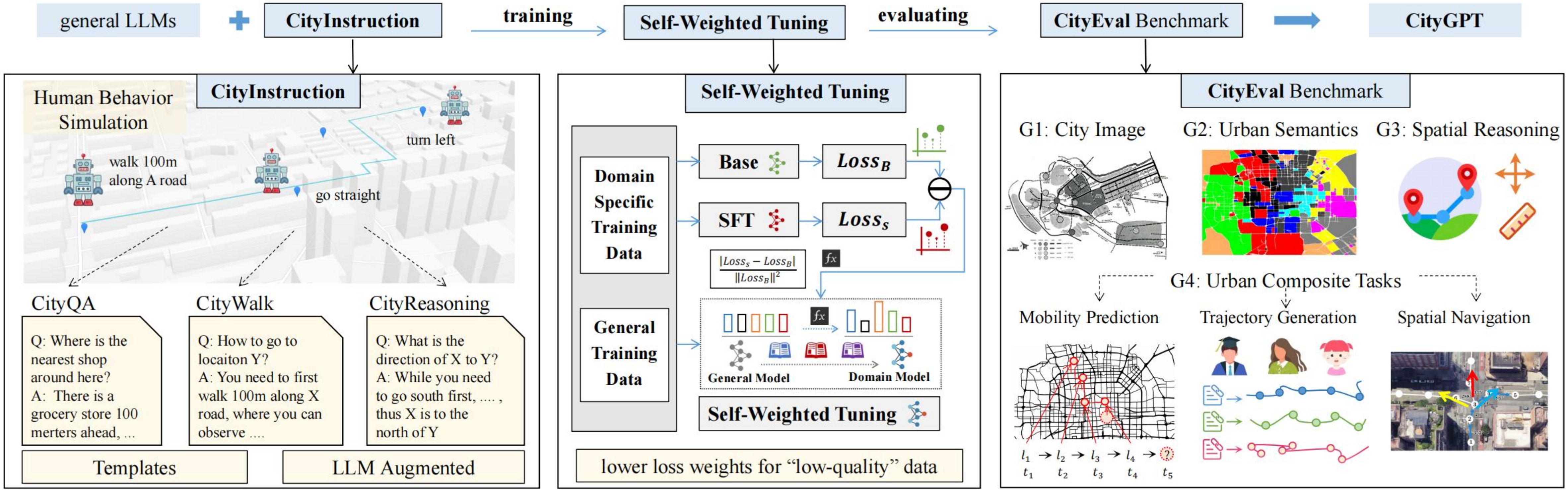


8 tasks in 13 cities around the world

Feng, Jie*, Jun Zhang*, Tianhui Liu*, Xin Zhang, Tianjian Ouyang, Junbo Yan, Yuwei Du, Siqi Guo, and Yong Li. "CityBench: Evaluating the capabilities of large language models for urban tasks." KDD 2025 Dataset & Benchmark.

Before UrbanLLaVA: CityGPT

- **For urban geospatial knowledge: CityGPT, using daily behavior simulation to synthesize data**



Q: Where is the nearest shop around here?
A: There is a grocery store 100 merters ahead, ...

Q: How to go to locaiton Y?
A: You need to first walk 100m along X road, where you can observe ....

Q: What is the direction of X to Y?
A: While you need to go south first, ... , thus X is to the north of Y

Large-scale human behavior simulation to synthesize high-quality, multi-source spatial data.

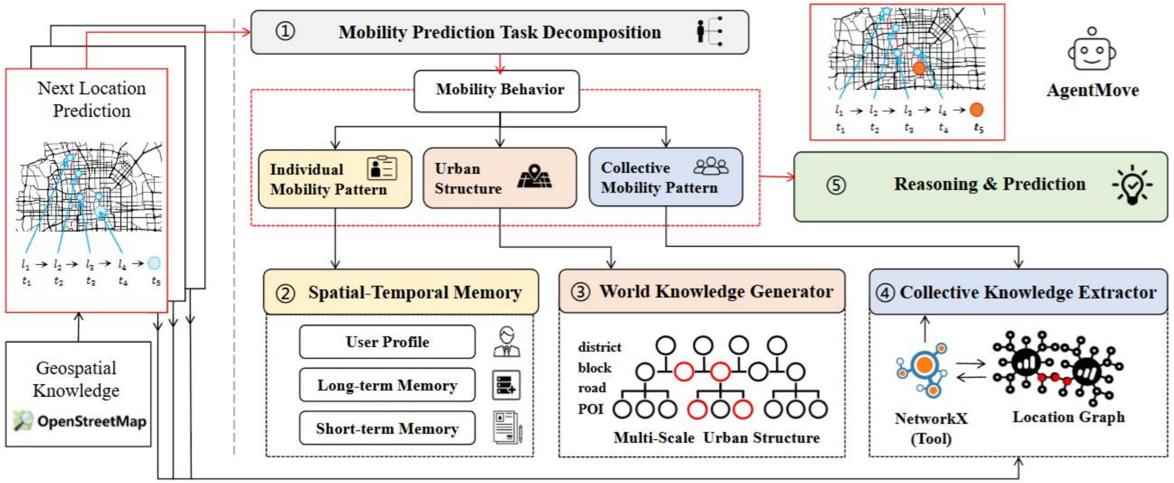
Domain-data mixed adaptive training to balance general and specialized capabilities.

Systematic evaluation of language-based urban spatial cognition.

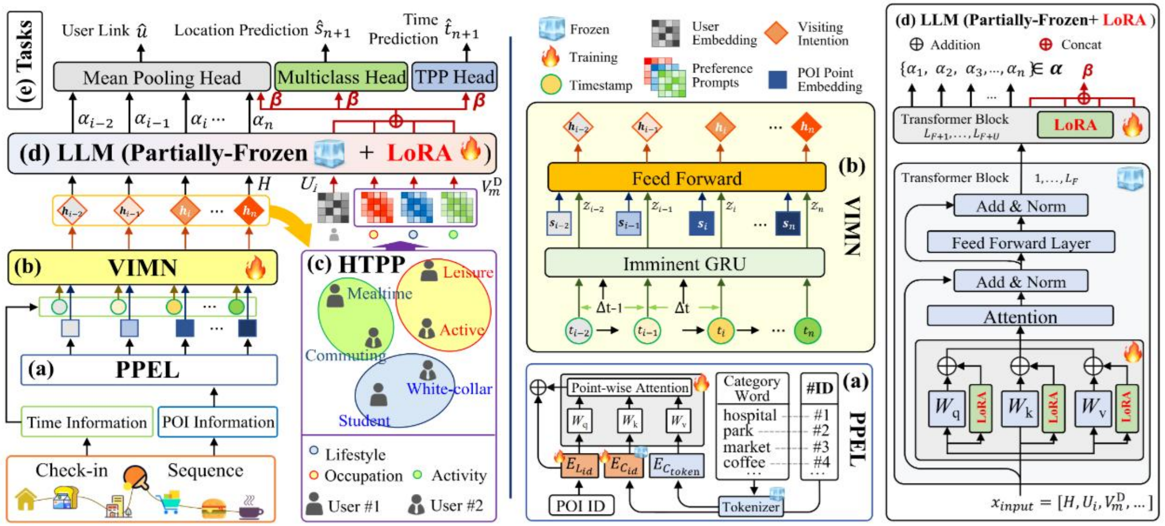
Feng, Jie*, Tianhui Liu*, Yuwei Du, Siqi Guo, Yuming Lin, and Yong Li. "CityGPT: Empowering urban spatial cognition of large language models." KDD 2025 Research

Before UrbanLLaVA: AgentMove/Mobility-LLM

- **For urban mobility intelligence: language-driven methods (AgentMove) and modality-aligned modeling (Mobility-LLM).**



AgentMove: Models movement behavior by translating mobility trajectories into natural language descriptions and performing the modeling within the natural language space.



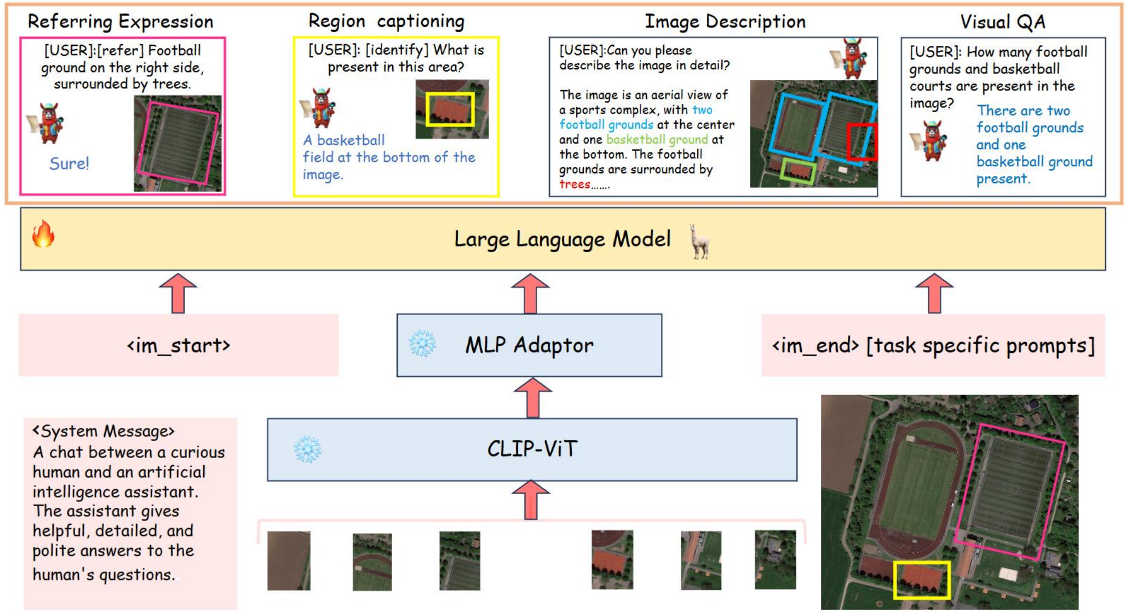
Mobility-LLM: Models movement by using classical models to process mobility trajectories, obtaining an embedded representation that is then integrated into the LLM.

Gong, Letian, et al. "Mobility-LLM: Learning visiting intentions and travel preference from human mobility data with large language models." NeurIPS 2024

Feng, Jie, et al. "AgentMove: A large language model based agentic framework for zero-shot next location prediction." NAACL 2025.

Before UrbanLLaVA: GeoChat/UrbanVLP

- **For urban vision intelligence, e.g., GeoChat and UrbanVLP.**



[USER]:[refer] Football ground on the right side, surrounded by trees.

Sure!

[USER]: [identify] What is present in this area?

A basketball field at the bottom of the image.

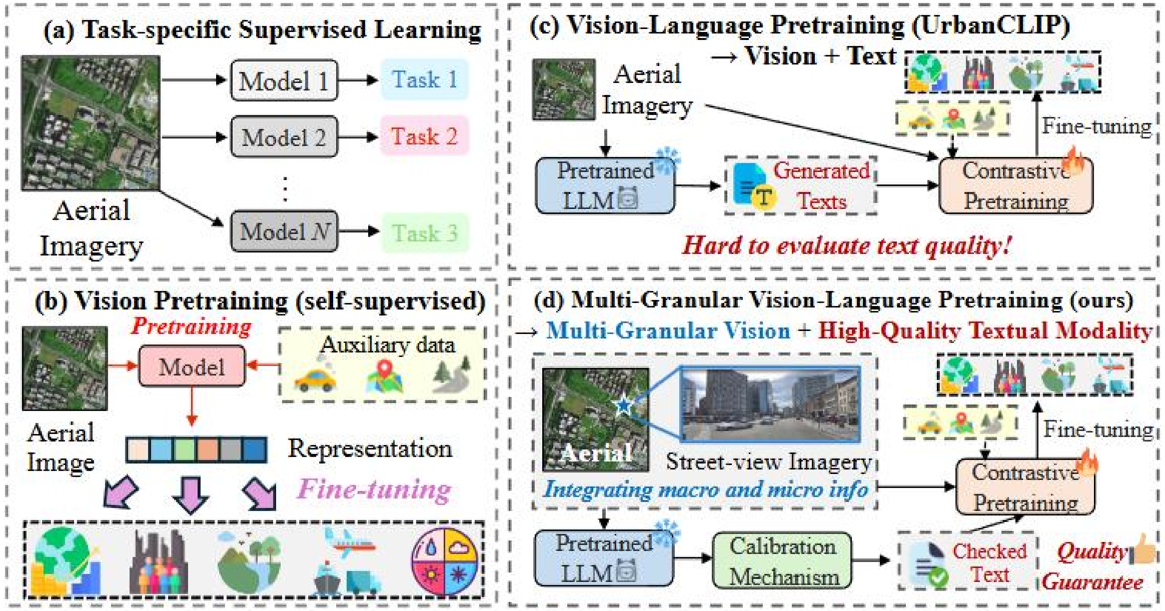
[USER]:Can you please describe the image in detail?

The image is an aerial view of a sports complex, with two football grounds at the center and one basketball ground at the bottom. The football grounds are surrounded by trees……

[USER]: How many football grounds and basketball courts are present in the image?

There are two football grounds and one basketball ground present.

<System Message> A chat between a curious human and an artificial intelligence assistant. The assistant gives helpful, detailed, and polite answers to the human's questions.



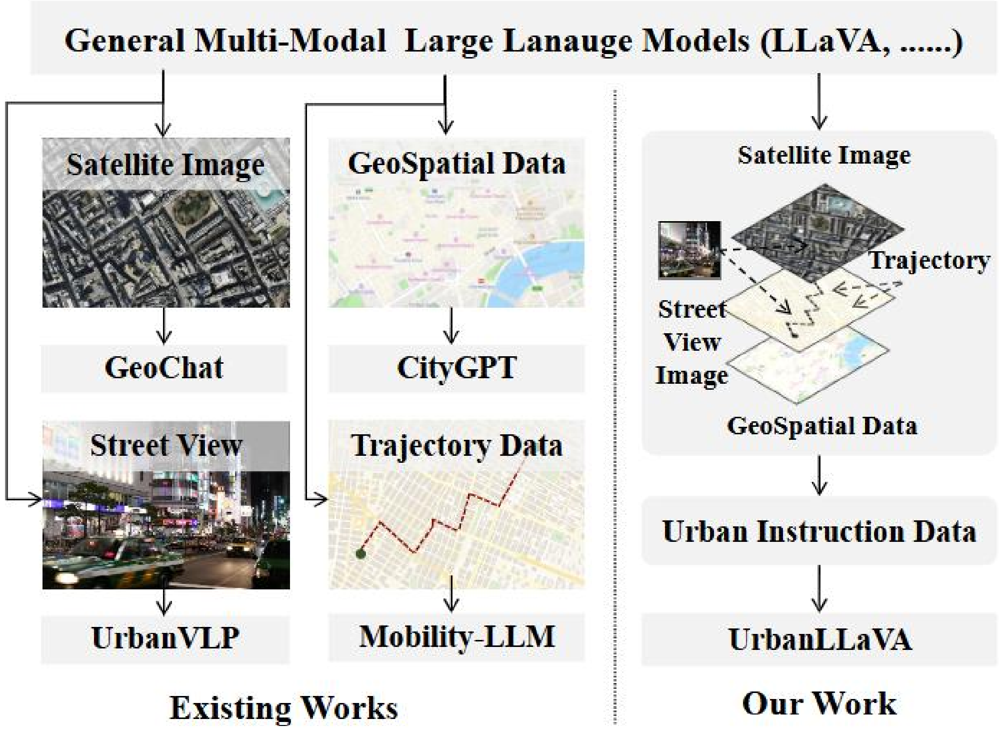
GeoChat: Addresses urban remote sensing recognition and detection tasks using a classical multi-modal architecture.

UrbanVLP: Focuses on predicting socio-economic indicators from street view imagery, leveraging contrastive learning and LLM annotation.

Kuckreja, Kartik, et al. "Geochat: Grounded large vision-language model for remote sensing. " CVPR 2024
Hao, Xixuan, et al. "Urbanvlp: Multi-granularity vision-language pretraining for urban socioeconomic indicator prediction." AAAI 2025

UrbanLLaVA: Building Unified Foundation Model for Urban Intelligence

- **UrbanLLaVA:** Achieves unified modeling of **remote sensing, street view, geographical data,** and **mobility trajectories** by deeply aligning and fusing **visual information** with **geospatial knowledge**.



General Multi-Modal Large Lanauge Models (LLaVA, ...)

Satellite Image → GeoChat

GeoSpatial Data → CityGPT

Street View → UrbanVLP

Trajectory Data → Mobility-LLM

Existing Works

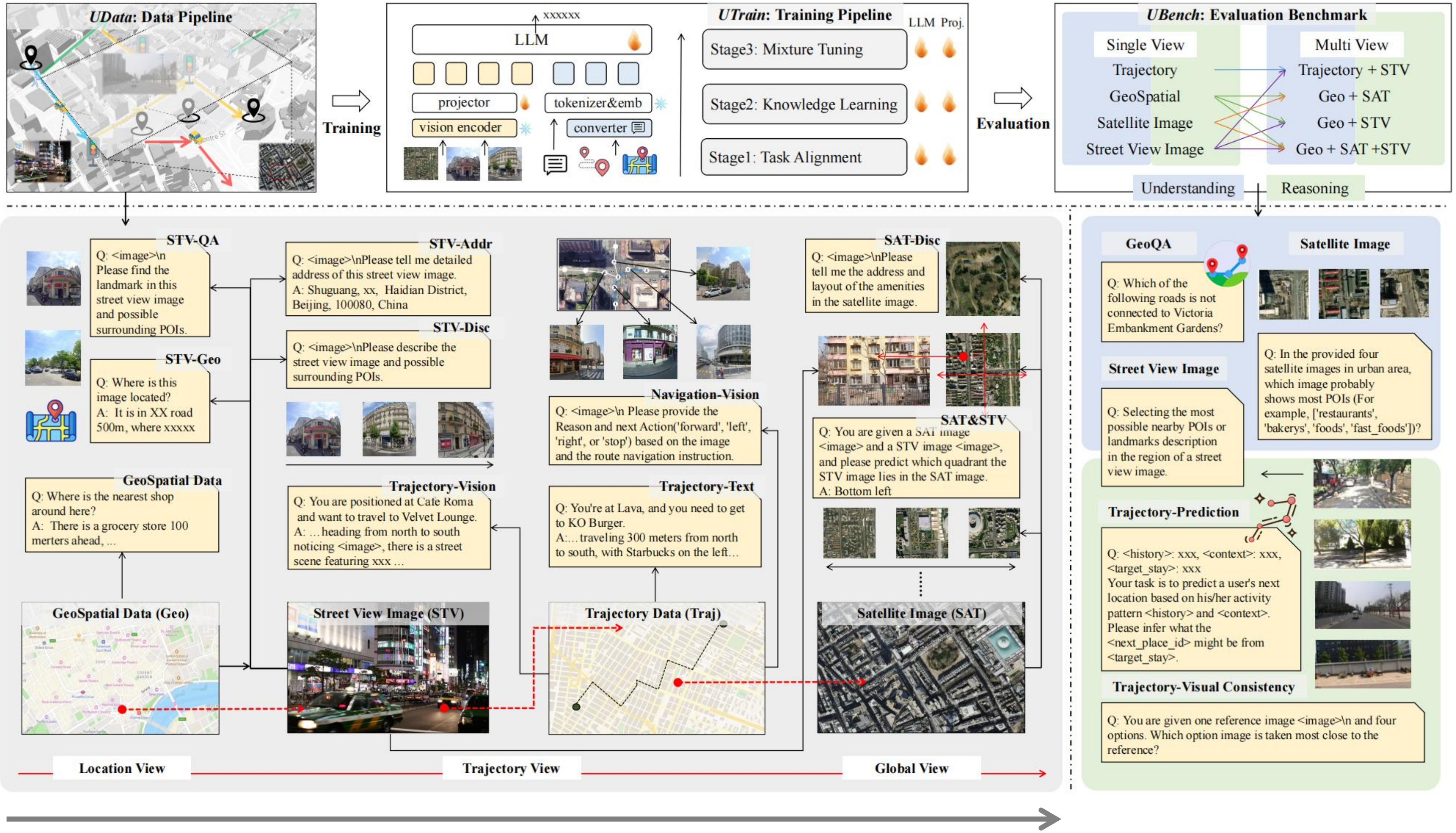
Satellite Image, Trajectory, Street View Image, GeoSpatial Data → Urban Instruction Data → UrbanLLaVA

Our Work

Feng, Jie, Shengyuan Wang, Tianhui Liu, Yanxin Xi, and Yong Li. "UrbanLLaVA: A Multi-modal Large Language Model for Urban Intelligence with Spatial Reasoning." ICCV 2025

UrbanLLaVA: Building Unified Foundation Model for Urban Intelligence

- **UrbanLLaVA：UData + Utrain + UBench**



STV-QA: Q: <image>\n Please find the landmark in this street view image and possible surrounding POIs.

STV-Geo: Q: Where is this image located? A: It is in XX road 500m, where xxxxx

GeoSpatial Data: Q: Where is the nearest shop around here? A: There is a grocery store 100 merters ahead, ...

STV-Addr: Q: <image>\nPlease tell me detailed address of this street view image. A: Shuguang, xx, Haidian District, Beijing, 100080, China

STV-Disc: Q: <image>\nPlease describe the street view image and possible surrounding POIs.

Trajectory-Vision: Q: You are positioned at Cafe Roma and want to travel to Velvet Lounge. A: ...heading from north to south noticing <image>, there is a street scene featuring xxx ...

Navigation-Vision: Q: <image>\n Please provide the Reason and next Action('forward', 'left', 'right', or 'stop') based on the image and the route navigation instruction.

Trajectory-Text: Q: You're at Lava, and you need to get to KO Burger. A:...traveling 300 meters from north to south, with Starbucks on the left...

SAT-Disc: Q: <image>\nPlease tell me the address and layout of the amenities in the satellite image.

SAT&STV: Q: You are given a SAT image <image> and a STV image <image>, and please predict which quadrant the STV image lies in the SAT image. A: Bottom left

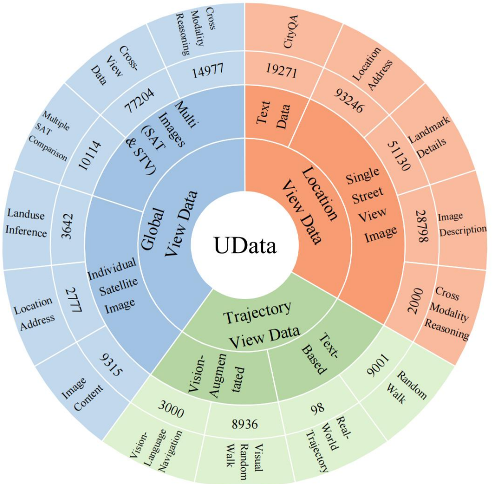
GeoQA: Q: Which of the following roads is not connected to Victoria Embankment Gardens?

Satellite Image: Q: In the provided four satellite images in urban area, which image probably shows most POIs (For example, ['restaurants', 'bakerys', 'foods', 'fast_foods'])?

Street View Image: Q: Selecting the most possible nearby POIs or landmarks description in the region of a street view image.

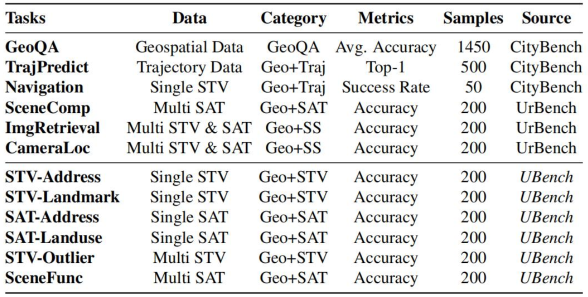
Trajectory-Prediction: Q: <history>: xxx, <context>: xxx, <target_stay>: xxx Your task is to predict a user's next location based on his/her activity pattern <history> and <context>. Please infer what the <next_place_id> might be from <target_stay>.

Trajectory-Visual Consistency: Q: You are given one reference image <image>\n and four options. Which option image is taken most close to the reference?



Training Data

Table 1. Detailed information about *UBench* for Beijing, 'STV' refers to street view image, and 'SAT' refers to satellite image.

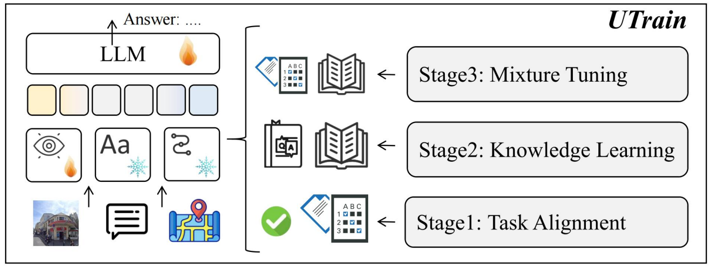


Tasks	Data	Category	Metrics	Samples	Source
GeoQA	Geospatial Data	GeoQA	Avg. Accuracy	1450	CityBench
TrajPredict	Trajectory Data	Geo+Traj	Top-1	500	CityBench
Navigation	Single STV	Geo+Traj	Success Rate	50	CityBench
SceneComp	Multi SAT	Geo+SAT	Accuracy	200	UrBench
ImgRetrieval	Multi STV & SAT	Geo+SS	Accuracy	200	UrBench
CameraLoc	Multi STV & SAT	Geo+SS	Accuracy	200	UrBench
STV-Address	Single STV	Geo+STV	Accuracy	200	*UBench*
STV-Landmark	Single STV	Geo+STV	Accuracy	200	*UBench*
SAT-Address	Single SAT	Geo+SAT	Accuracy	200	*UBench*
SAT-Landuse	Single SAT	Geo+SAT	Accuracy	200	*UBench*
STV-Outlier	Multi STV	Geo+STV	Accuracy	200	*UBench*
SceneFunc	Multi SAT	Geo+SAT	Accuracy	200	*UBench*

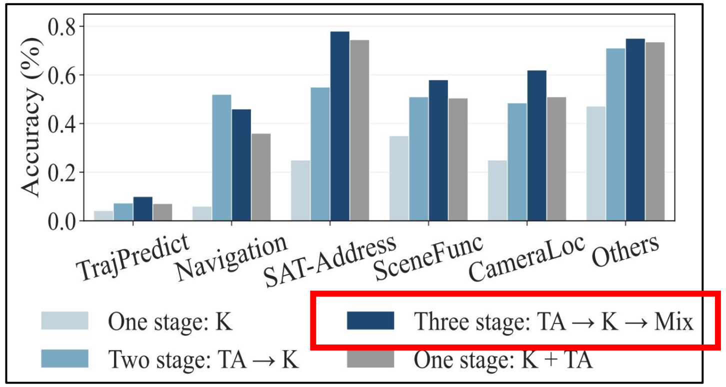
Feng, Jie, Shengyuan Wang, Tianhui Liu, Yanxin Xi, and Yong Li. "UrbanLLaVA: A Multi-modal Large Language Model for Urban Intelligence with Spatial Reasoning." ICCV 2025

UrbanLLaVA: Building Unified Foundation Model for Urban Intelligence

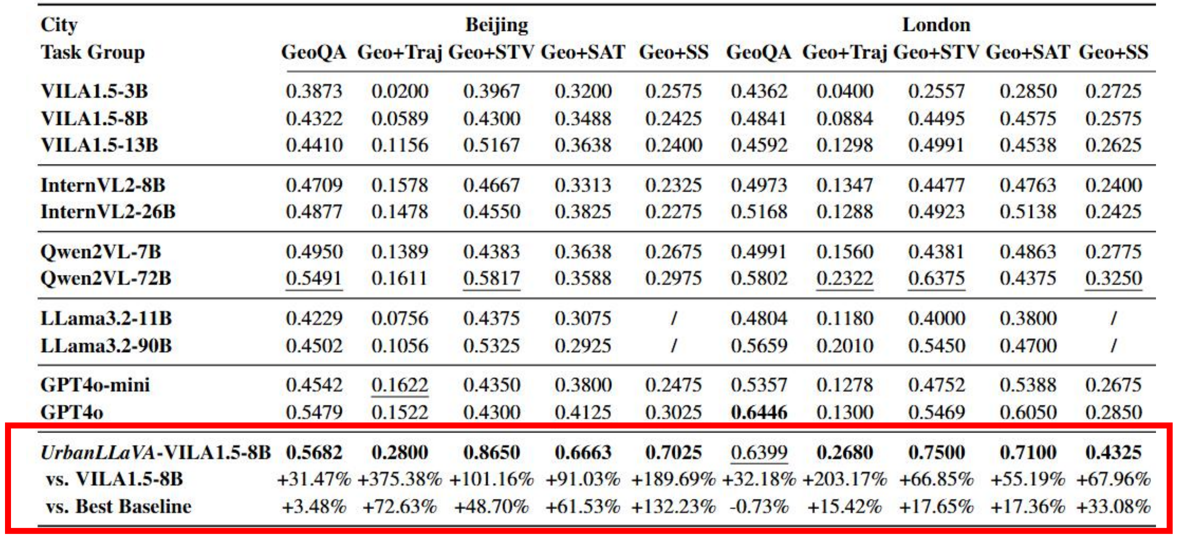
- **Uses a multi-stage training pipeline for urban domain multi-modal and multi-scenario tasks, leading to significant performance gains on critical benchmarks.**



multi-stage training pipeline

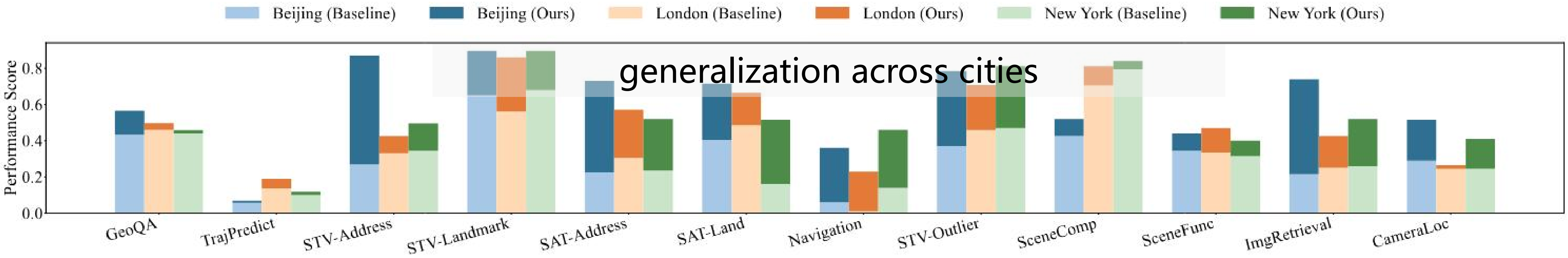


multi-stage training approach yields superior performance on critical tasks compared to direct mixed training.



City	Beijing					London				
Task Group	**GeoQA**	**Geo+Traj**	**Geo+STV**	**Geo+SAT**	**Geo+SS**	**GeoQA**	**Geo+Traj**	**Geo+STV**	**Geo+SAT**	**Geo+SS**
VILA1.5-3B	0.3873	0.0200	0.3967	0.3200	0.2575	0.4362	0.0400	0.2557	0.2850	0.2725
VILA1.5-8B	0.4322	0.0589	0.4300	0.3488	0.2425	0.4841	0.0884	0.4495	0.4575	0.2575
VILA1.5-13B	0.4410	0.1156	0.5167	0.3638	0.2400	0.4592	0.1298	0.4991	0.4538	0.2625
InternVL2-8B	0.4709	0.1578	0.4667	0.3313	0.2325	0.4973	0.1347	0.4477	0.4763	0.2400
InternVL2-26B	0.4877	0.1478	0.4550	0.3825	0.2275	0.5168	0.1288	0.4923	0.5138	0.2425
Qwen2VL-7B	0.4950	0.1389	0.4383	0.3638	0.2675	0.4991	0.1560	0.4381	0.4863	0.2775
Qwen2VL-72B	0.5491	0.1611	0.5817	0.3588	0.2975	0.5802	0.2322	0.6375	0.4375	0.3250
LLama3.2-11B	0.4229	0.0756	0.4375	0.3075	/	0.4804	0.1180	0.4000	0.3800	/
LLama3.2-90B	0.4502	0.1056	0.5325	0.2925	/	0.5659	0.2010	0.5450	0.4700	/
GPT4o-mini	0.4542	0.1622	0.4350	0.3800	0.2475	0.5357	0.1278	0.4752	0.5388	0.2675
GPT4o	0.5479	0.1522	0.4300	0.4125	0.3025	**0.6446**	0.1300	0.5469	0.6050	0.2850
***UrbanLLaVA*-VILA1.5-8B**	**0.5682**	**0.2800**	**0.8650**	**0.6663**	**0.7025**	0.6399	**0.2680**	**0.7500**	**0.7100**	**0.4325**
vs. VILA1.5-8B	+31.47%	+375.38%	+101.16%	+91.03%	+189.69%	+32.18%	+203.17%	+66.85%	+55.19%	+67.96%
vs. Best Baseline	+3.48%	+72.63%	+48.70%	+61.53%	+132.23%	-0.73%	+15.42%	+17.65%	+17.36%	+33.08%

achieves significant performance improvements across various tasks in multiple cities.



generalization across cities

UrbanLLaVA: Building Unified Foundation Model for Urban Intelligence

- **Task Examples**

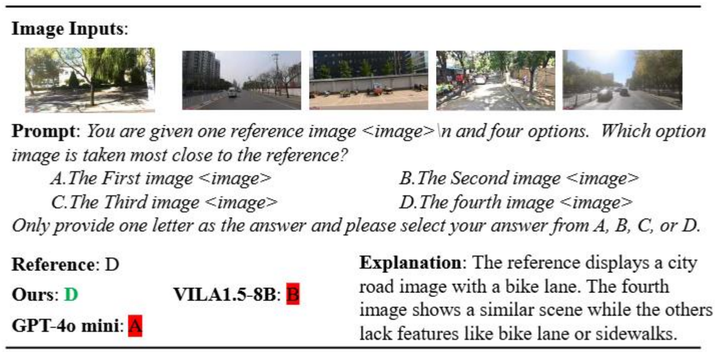


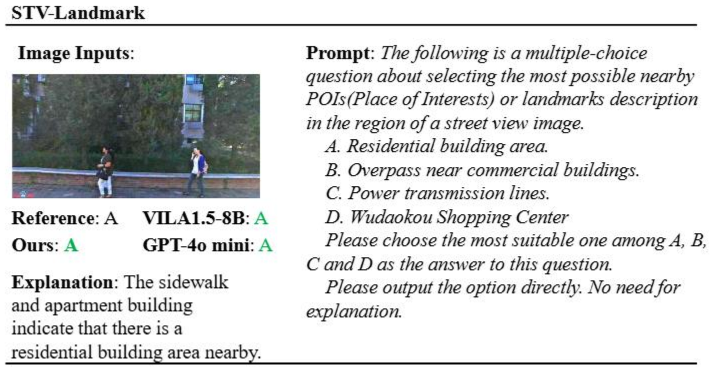
Image Inputs:

Prompt: *You are given one reference image <image>\n and four options. Which option image is taken most close to the reference?*
A.The First image <image> *B.The Second image <image>*
C.The Third image <image> *D.The fourth image <image>*
Only provide one letter as the answer and please select your answer from A, B, C, or D.

Reference: D
Ours: D **VILA1.5-8B**: B
GPT-4o mini: A

Explanation: The reference displays a city road image with a bike lane. The fourth image shows a similar scene while the others lack features like bike lane or sidewalks.

STV-Outlier Task



STV-Landmark

Image Inputs:

Reference: A **VILA1.5-8B**: A
Ours: A **GPT-4o mini**: A

Explanation: The sidewalk and apartment building indicate that there is a residential building area nearby.

Prompt: *The following is a multiple-choice question about selecting the most possible nearby POIs(Place of Interests) or landmarks description in the region of a street view image.*
A. Residential building area.
B. Overpass near commercial buildings.
C. Power transmission lines.
D. Wudaokou Shopping Center
Please choose the most suitable one among A, B, C and D as the answer to this question.
Please output the option directly. No need for explanation.

STV-Landmark Task

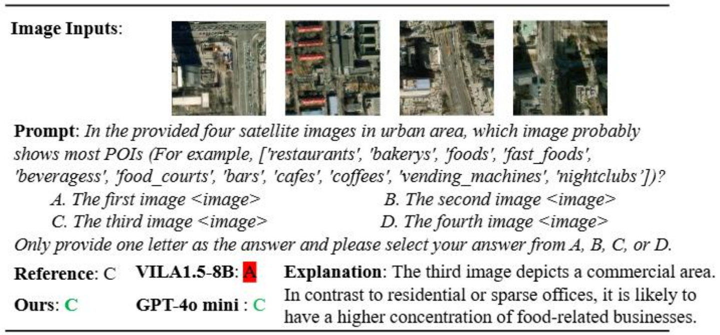


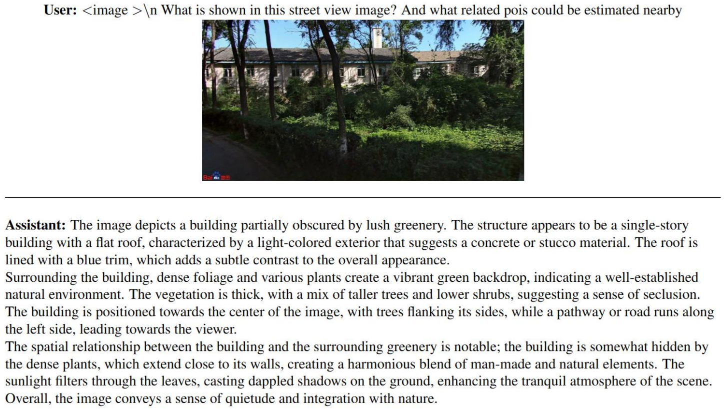
Image Inputs:

Prompt: *In the provided four satellite images in urban area, which image probably shows most POIs (For example, ['restaurants', 'bakerys', 'foods', 'fast_foods', 'beveragess', 'food_courts', 'bars', 'cafes', 'coffees', 'vending_machines', 'nightclubs'])?*
A. The first image <image> *B. The second image <image>*
C. The third image <image> *D. The fourth image <image>*
Only provide one letter as the answer and please select your answer from A, B, C, or D.

Reference: C **VILA1.5-8B**: A
Ours: C **GPT-4o mini** : C

Explanation: The third image depicts a commercial area. In contrast to residential or sparse offices, it is likely to have a higher concentration of food-related businesses.

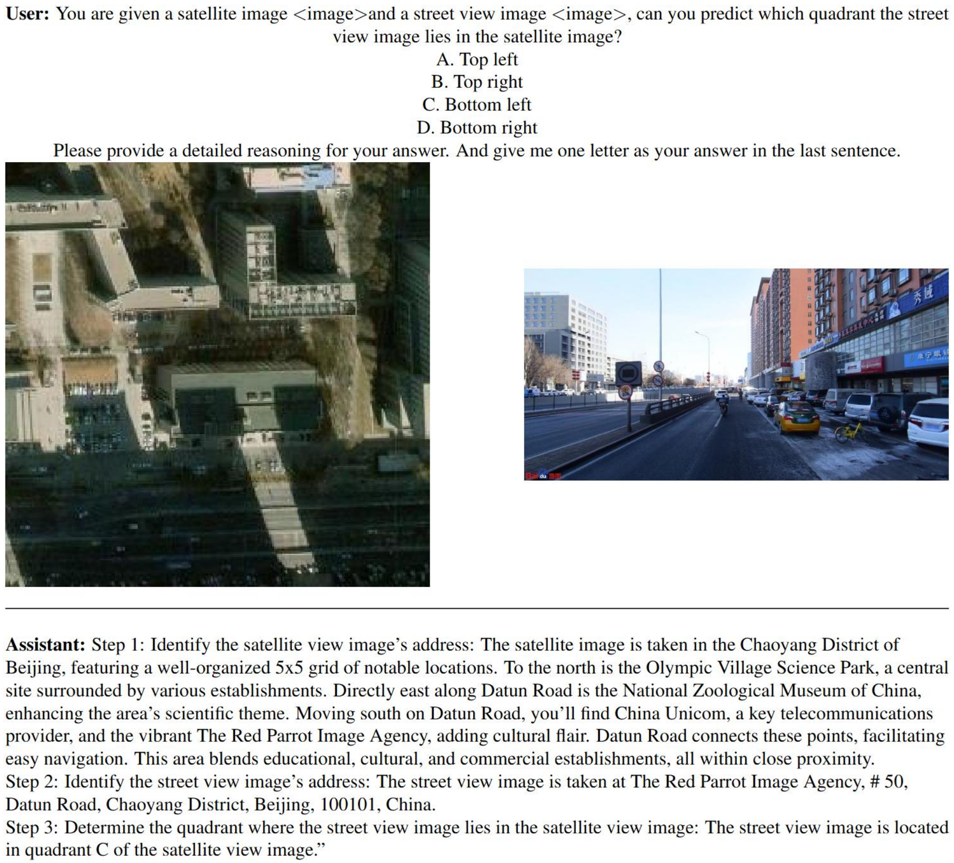
SceneFunc Task with Satellite Image



User: <image >\n What is shown in this street view image? And what related pois could be estimated nearby

Assistant: The image depicts a building partially obscured by lush greenery. The structure appears to be a single-story building with a flat roof, characterized by a light-colored exterior that suggests a concrete or stucco material. The roof is lined with a blue trim, which adds a subtle contrast to the overall appearance.
Surrounding the building, dense foliage and various plants create a vibrant green backdrop, indicating a well-established natural environment. The vegetation is thick, with a mix of taller trees and lower shrubs, suggesting a sense of seclusion.
The building is positioned towards the center of the image, with trees flanking its sides, while a pathway or road runs along the left side, leading towards the viewer.
The spatial relationship between the building and the surrounding greenery is notable; the building is somewhat hidden by the dense plants, which extend close to its walls, creating a harmonious blend of man-made and natural elements. The sunlight filters through the leaves, casting dappled shadows on the ground, enhancing the tranquil atmosphere of the scene.
Overall, the image conveys a sense of quietude and integration with nature.

STV-Description



User: You are given a satellite image <image>and a street view image <image>, can you predict which quadrant the street view image lies in the satellite image?
A. Top left
B. Top right
C. Bottom left
D. Bottom right
Please provide a detailed reasoning for your answer. And give me one letter as your answer in the last sentence.

Assistant: Step 1: Identify the satellite view image's address: The satellite image is taken in the Chaoyang District of Beijing, featuring a well-organized 5x5 grid of notable locations. To the north is the Olympic Village Science Park, a central site surrounded by various establishments. Directly east along Datun Road is the National Zoological Museum of China, enhancing the area's scientific theme. Moving south on Datun Road, you'll find China Unicom, a key telecommunications provider, and the vibrant The Red Parrot Image Agency, adding cultural flair. Datun Road connects these points, facilitating easy navigation. This area blends educational, cultural, and commercial establishments, all within close proximity.
Step 2: Identify the street view image's address: The street view image is taken at The Red Parrot Image Agency, # 50, Datun Road, Chaoyang District, Beijing, 100101, China.
Step 3: Determine the quadrant where the street view image lies in the satellite view image: The street view image is located in quadrant C of the satellite view image."

CameraLoc Task with STV and Satellite Image

UrbanLLaVA@Github

Thanks!

Arxiv

Contact Jie Feng via:

https://vonfeng.github.io/

fengj12ee@hotmail.com