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

Supplementary Material

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.

Figure 7. An example of the SceneFunc task, where correct answers are in green, wrong ones in red.

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.

Figure 8. An example of the STV-Outlier task.

7. Limitation and Future Work

While we have made every effort to explore *UrbanLLaVA* and present our findings clearly, several limitations remain. Our experiments have focused on the 8B model; the full potential of *UData* and *UTrain* on larger models has yet to be realized. In addition, *UBench* can still be improved by refining the design of tasks, testing MLLMs' overall multi-modal capabilities from a more fine-grained perspective. Lastly, more modalities could be included like video, time series data, etc., which are also important in urban intelligence. In the future, we plan to extend *UrbanLLaVA* to incorporate more diverse data types in urban research and tackle more advanced urban tasks from various interdisciplinary fields.

7.1. Case Study

Here, we show two typical examples of urban task instances to demonstrate that *UrbanLLaVA* can handle challenging urban tasks. Due to space limits, other cases can refer to supplementary materials.

SceneFunc. This task challenges the model to identify which satellite image contains the highest concentration of a specified category of places of interest (POIs). Involving multiple image inputs and text prompts, SceneFunc task demands the model to understand and compare the differences between different images. As shown in Figure 7, while VILA1.5-8B fails to answer the question, our *UrbanLLaVA* succeeds in giving the correct answer. Our model exhibits strong capabilities like multiple image understanding and comparison in this example, and is competitive with the successful closed-source model.

STV-Outlier. This task requires model to compare between multiple street views and point out the closest one to a reference. Figure 8 shows an instance of this task, where VILA1.5-8B does not successfully identify the scene of the reference image. GPT-4o-mini is closer, but it is still confused by another wrong option. Our model shows its capabilities of understanding multiple images and conducting high-level implicit logical reasoning in an urban context, outperforming these general MLLMs.

8. Data Example for Three Stages

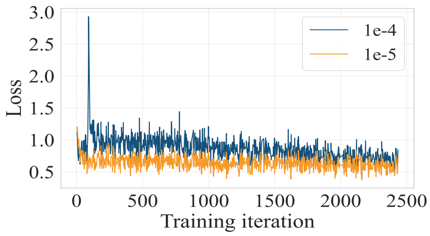
Below, we provide examples for each stage.

```
# one example in Stage 1: Task Alignment
Q1: The following is a multiple-choice question about selecting the
most possible landuse type in the region of a satellite image.
<image>
    A. Industrial    B. Forest    C. Farmland    D. Recreation
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 explaination.
A1: B
# two examples in Stage 2: Knowledge Learning
Q1: I am curious about the location of xxxx. Can you help me out?
A1: Beiwu Village Road, Shuguang, Yuquan Village, Haidian District,
Beijing, 100097, China
Q2: Can you count the number of buildings in this satellite image?
Please provide a detailed reasoning for your answer. <image>
A2: Step 1: Identify the city name: According to the satellite view
image, this image was likely taken in Beijing.
Step 2: Describe the location's features: {location_feature}
Step 3: Provide the location's address: {location_description}
Step 4: Count the number of buildings or points of interest: There
are three buildings in this image.
# one example in Stage 3: Mixed Tuning
# Examples in this stage is random sampled from Stage 1 and Stage 2
```

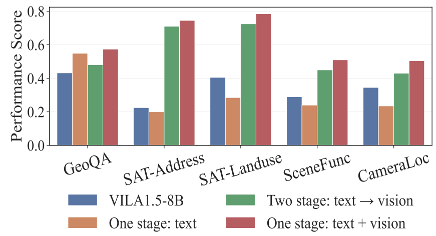
Figure 9. Input data examples for three-stage training.

9. Comparing with models for single-modality urban tasks

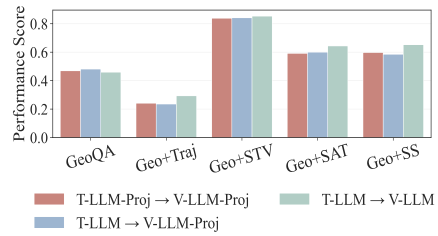
We focus on comparing our approach with these modality-specific models to better showcase its effectiveness, which are presented in following Table 5.



(a) The changes in training loss under the settings of learning rates $1e-4$ and $1e-5$.



(b) Results on *UBench* when *UData* is divided into text data and vision data.



(c) Results on *UBench* from experiments using different training components.

Figure 10. (a) illustrates that the training loss is smoother and lower when the learning rate is $1e-5$ (ours) compared to $1e-4$(VILA). (b) 'One stage: text' means training with text data, 'Two stage: text→vision' means training with text data in the first stage then vision data in the second stage, 'One stage: text+vision' means training with text and visual data in one stage. 'Others' refers to other tasks in *UBench*. (c) 'T' refers to Text data, 'V' refers to Vision data and 'T-LLM-Proj→V-LLM-Proj' means training text data with LLM and Projector in the first stage, later vision data with LLM and Projector.

Table 5. Comparing with models for single-modality urban tasks.

	GeoQA	STV-Address	STV-Landmark	SAT-Address	SAT-Landuse
GeoChat	0.3746	0.3100	0.2050	0.2872	0.4650
CityGPT	0.5238	-	-	-	-
UrbanCLIP	-	-	-	-	0.3750
Ours	0.5741	0.8550	0.8750	0.7450	0.7850

10. Additional Detailed Results of Three Cities

The detailed results of *UBench* on three cities are presented in Table 6, Table 7 and Table 8. Table 2 in Section 3.2 is the aggregated results of these three tables. *UrbanLLaVA*@Beijing surpasses the baselines in all tasks, showcasing exceptional performance. *UrbanLLaVA*@London delivers top results in 9 out of 12 tasks, achieving gains over the best baseline ranging from 5.17% to 49.38%. Meanwhile, *UrbanLLaVA*@NewYork performs best in 9 tasks as well, with improvements over the best baseline spanning from 6.25% to 122.78%.

11. Additional Results for Training Strategies

As a supplement to results in Section 3.3, we report results on learning rate, modality and trained components here.

We first adjusted some experimental parameters to explore their effects, and ultimately found that the learning rate is the key parameter influencing training stability and model performance. As shown in Figure 10a, we conduct experiments on the same data with different training parameters, and compared to a learning rate of 1e-4 (the default choice of VILA), the curve is smoother and lower with a learning rate of 1e-5. We think training with mixed domain-specific structured instruction data, a lower learning rate (1e-5) enables the model to handle features from different modalities more robustly.

Then we consider whether to separate text data and vision data during training to explore the influence of text data and vision data on training. But as shown in Figure 10b, training with text and vision data in one stage yields better results compared to the other two experiments and base model VILA1.5-8B. We also investigated the impact of training components. As Figure 10c presents, using different components to train the same data shows little difference.

12. Effects of Training Data Size

Fig. 11 presents training results with different amounts, exhibiting the high quality of UData.

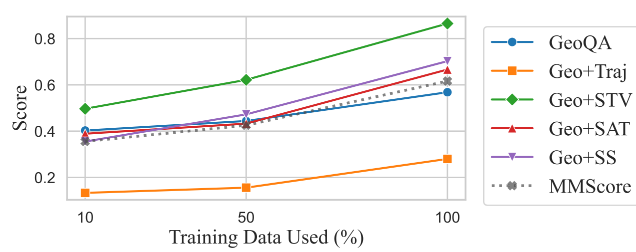


Figure 11. Scaling law from training data size to performance.

13. Effects of Base Model

Our method is model-agnostic and can be generalized to different MLLMs, e.g., Qwen2.5-VL-7B in Table 9.

14. Effects of Model Size

As Figure 12 shows, performance generally improves with increasing parameter size for VILA 1.5 (from 3B to 13B). However, for certain tasks, models of different sizes exhibit similar capabilities. This occurs either because the tasks are inherently challenging (e.g., trajectory prediction) or relatively easy (e.g., SAT-Landuse). Compared to VILA1.5-8B, the performance improvement of VILA1.5-13B is minimal, likely due to the capabilities of the LLaMA3-8B and LLaMA2-13B models utilized by VILA1.5. However, due

Table 6. Main results on *UBench* at Beijing. *UrbanLLaVA* significantly outperforms other baselines in every task.

Tasks@Beijing	GeoQA	Geo+Traj		Geo+STV			Geo+SAT				Geo+SS	
	GeoQA	TrajPredict	Navigation	STV-Address	STV-Landmark	STV-Outlier	SAT-Address	SAT-Landuse	SceneComp	SceneFunc	ImgRetrieval	CameraLoc
Qwen2VL-7B	0.4950	0.0978	0.18	0.440	0.755	0.1200	0.295	0.405	0.400	0.355	0.275	0.260
Qwen2VL-72B	0.5491	0.0822	0.24	0.410	0.785	0.5500	0.395	0.395	0.335	0.310	0.290	0.305
InternVL2-8B	0.4709	0.0957	0.22	0.420	0.755	0.2250	0.295	0.300	0.390	0.340	0.210	0.255
InternVL2-26B	0.4877	0.0756	0.22	0.440	0.755	0.1700	0.360	0.375	0.440	0.355	0.230	0.225
VILA1.5-3B	0.3873	0.0000	0.04	0.270	0.655	0.2650	0.275	0.475	0.295	0.235	0.250	0.265
VILA1.5-8B	0.4322	0.0578	0.06	0.270	0.650	0.3700	0.225	0.405	0.420	0.345	0.195	0.290
VILA1.5-13B	0.4410	0.0511	0.18	0.305	0.715	0.5300	0.320	0.320	0.425	0.390	0.270	0.210
LLaMA3.2-11B	0.4229	0.0711	0.08	0.280	0.595	/	0.290	0.325	/	/	/	/
LLaMA3.2-90B	0.4502	0.0711	0.14	0.295	0.770	/	0.295	0.290	/	/	/	/
GPT4o-mini	0.4542	0.0844	0.24	0.280	0.765	0.2600	0.350	0.360	0.465	0.345	0.205	0.290
GPT4o	0.5479	0.0844	0.22	0.405	0.775	0.1100	0.390	0.420	0.450	0.390	0.315	0.290
***UrbanLLaVA*-VILA1.5-8B**	**0.5682**	**0.1000**	**0.46**	**0.91**	**0.870**	**0.8150**	**0.780**	**0.72**	**0.585**	**0.58**	**0.785**	**0.62**
vs. VILA1.5-8B	+31.47%	+73.10%	+666.67%	+237.04%	+33.85%	+120.27%	+246.67%	+77.78%	+39.29%	+68.12%	+302.56%	+113.79%
vs. Best Baseline	+3.48%	+2.28%	+91.67%	+106.82%	+10.83%	+48.18%	+97.47%	+51.58%	+25.81%	+48.72%	+149.21%	+103.28%

Table 7. Main results on *UBench* at London. *UrbanLLaVA* achieves better performance than other baselines in the majority of tasks.

Tasks@Beijing	GeoQA	Geo+Traj		Geo+STV			Geo+SAT				Geo+SS	
	GeoQA	TrajPredict	Navigation	STV-Address	STV-Landmark	STV-Outlier	SAT-Address	SAT-Landuse	SceneComp	SceneFunc	ImgRetrieval	CameraLoc
Qwen2VL-7B	0.4991	0.1920	0.12	0.405	0.760	0.1492	0.305	0.550	0.870	0.220	0.270	**0.285**
Qwen2VL-72B	0.5802	**0.2245**	0.24	0.485	0.875	0.5525	0.530	0.535	0.420	0.265	0.405	0.245
InternVL2-8B	0.4973	0.1694	0.10	0.290	0.810	0.2431	0.315	0.490	0.785	0.315	0.215	0.265
InternVL2-26B	0.5168	0.1776	0.08	0.380	0.865	0.2320	0.355	0.490	0.905	0.305	0.215	0.270
VILA1.5-3B	0.4362	0.0000	0.08	0.230	0.305	0.2320	0.200	0.445	0.295	0.200	0.290	0.255
VILA1.5-8B	0.4841	0.1367	0.04	0.330	0.560	0.4586	0.305	0.485	0.705	0.335	0.250	0.265
VILA1.5-13B	0.4592	0.1796	0.08	0.430	0.570	0.4972	0.275	0.350	0.800	0.390	0.275	0.250
LLama3.2-11B	0.4804	0.1959	0.04	0.360	0.440	/	0.260	0.500	/	/	/	/
LLama3.2-90B	0.5659	0.2020	0.20	0.375	0.715	/	0.385	0.555	/	/	/	/
GPT4o-mini	0.5357	0.1755	0.08	0.375	0.835	0.2155	0.390	0.570	0.855	0.340	0.290	0.245
GPT4o	**0.6446**	0.2000	0.06	0.580	0.895	0.1657	0.480	0.610	0.900	0.430	0.320	0.250
***UrbanLLaVA*-VILA1.5-8B**	0.6399	0.1959	**0.34**	**0.610**	**0.955**	**0.6851**	**0.575**	**0.750**	**0.955**	**0.560**	**0.605**	0.260
vs. VILA1.5-8B	+32.20%	+43.28%	+750.00%	+84.85%	+70.54%	+49.40%	+88.52%	+54.64%	+35.46%	+67.16%	+142.00%	-1.89%
vs. Best Baseline	-0.72%	-12.73%	+41.67%	+5.17%	+6.70%	+24.00%	+8.49%	+22.95%	+5.52%	+30.23%	+49.38%	-8.77%

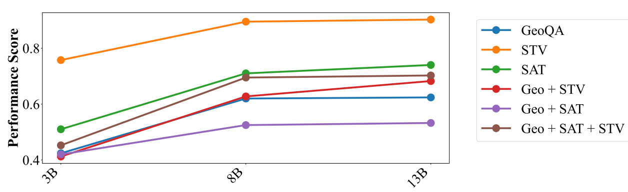


Figure 12. Results on *UrbanLLaVA* with different model sizes.

to limited computing resources, we were unable to provide results for VILA1.5-40B, which could potentially achieve significantly better performance than VILA1.5-8B.

15. Additional Case Study

SAT-LandUse. This task needs a model to speculate the land use type (commercial, residential, agricultural, etc.) based on a satellite image. One example is shown in Figure 13. Our *UrbanLLaVA* can respond to this task precisely, showing the capability of correctly perceiving the given image, satisfactory instruction following, and urban knowledge mastering.

SAT-LandUse

Image Inputs:

Prompt: *The following is a multiple-choice question about selecting the most possible landuse type in the region of a satellite image.*
A. Grass B. Residential
C. Garages D. Retail
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.

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

Explanation: Dense residential buildings in the remote sensing image denote the most possible land-use type of this area is residential.

Figure 13. An example of the SAT-LandUse task. The correct answers from model are denoted with green color. The response from ours is in bold. Explanation is written by human for this question and answer.

STV-Landmark. A representative example is in Figure 14. In this task, models are required to find out the closest landmark feature to a given street view, which

Table 8. Main results on *UBench* at NewYork. *UrbanLLaVA* achieves better performance than other models in most tasks.

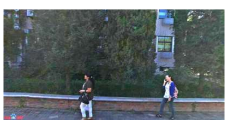
Tasks@Beijing	GeoQA	Geo+Traj		Geo+STV			Geo+SAT				Geo+SS	
	GeoQA	**TrajPredict**	**Navigation**	**STV-Address**	**STV-Landmark**	**STV-Outlier**	**SAT-Address**	**SAT-Landuse**	**SceneComp**	**SceneFunc**	**ImgRetrieval**	**CameraLoc**
Qwen2VL-7B	0.4567	0.1200	0.22	0.585	0.805	0.1450	0.455	0.395	0.875	0.315	0.275	0.315
Qwen2VL-72B	0.5273	**0.1480**	0.36	0.550	0.795	0.5550	0.520	0.235	0.470	0.290	0.335	0.320
InternVL2-8B	0.4632	0.1260	0.24	0.440	0.780	0.2550	0.395	0.135	0.835	0.305	0.245	0.235
InternVL2-26B	0.4766	0.1080	0.34	0.490	0.805	0.2700	0.495	0.225	0.885	0.290	0.230	0.245
VILA1.5-3B	0.3954	0.0000	0.08	0.330	0.745	0.2450	0.310	0.250	0.280	0.245	0.255	0.230
VILA1.5-8B	0.4575	0.1000	0.14	0.345	0.680	0.4700	0.235	0.160	0.795	0.315	0.260	0.245
VILA1.5-13B	0.4501	0.1100	0.36	0.375	0.765	0.5350	0.325	0.175	0.820	0.290	0.285	0.280
LLama3.2-11B	0.4127	0.1000	0.12	0.395	0.645	/	0.295	0.150	/	/	/	/
LLama3.2-90B	0.5234	0.1140	0.20	0.575	0.790	/	0.460	0.220	/	/	/	/
GPT4o-mini	0.5075	0.1240	0.34	0.550	0.880	0.2600	0.415	0.265	0.880	0.350	0.255	0.215
GPT4o	**0.6232**	0.1080	0.36	0.740	0.830	0.1600	0.610	0.215	**0.930**	0.405	0.305	0.275
CityGPT-V-VILA1.5-8B	0.5773	0.1120	**0.50**	**0.920**	**0.935**	**0.6950**	**0.885**	**0.880**	0.835	**0.490**	**0.645**	**0.520**
vs. VILA1.5-8B	+26.19%	+12.00%	+257.14%	+166.67%	+37.50%	+47.87%	+276.60%	+450.00%	+5.03%	+55.56%	+148.08%	+112.24%
vs. Best Baseline	-7.36%	-24.32%	+38.89%	+24.32%	+6.25%	+25.23%	+45.08%	+122.78%	-10.22%	+20.99%	+92.54%	+62.50%

Table 9. Evaluating generalizability of methods on Qwen2.5VL.

Task Group @ Beijing	GeoQA	Geo+Traj	Geo+STV	Geo+SAT	Geo+SS
Qwen2.5-VL-7B-Instruct	0.4324	0.2192	0.4467	0.2850	0.2225
+ Finetuned with UData	0.5720↑	0.1876	0.6833↑	0.4800↑	0.3800↑

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.

Figure 14. An example of the STV-Landmark task. The correct answers from model are denoted with green color. The response from ours is in bold. Explanation is written by human for this question and answer.

needs implicit logical reasoning capability to answer. By correctly answering a STV-Landmark question, *UrbanLLaVA* presents its ability to conduct logical reasoning in a multi-modal context.

SAT-Address. This task needs a model to speculate the most probable address description based on a satellite image. One example is shown in Figure 15

STV-Address. This task provides a street view image and needs a model to speculate the most probable address this image was taken. Figure 16 is an example.

SceneComp. This task provides four satellite remote sensing images and prompts the model to choose the one with the most number of buildings. An instance is shown in Figure 17.

SAT-Address

Image Inputs:

Reference: B

CityGPT-V: B

Prompt: *The following is a multiple-choice question about selecting the most appropriate address for a satellite image.*
A. The area is characterized by a well-defined square layout...
B. ...The eastern side features Zhonggiun Village East Road, which is lined with residential communities...
C. The area is a well-defined square located in Wudaokou, a vibrant neighborhood in Haidian District, Beijing...
D. ...To the north, you will find the prominent Zhongguancun Hospital, situated along Zhongguancun South Road, which runs through the area...
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.

Explanation: The right part of this remote sensing image can be identify to be a residential community, which echoes with description of option B.

Figure 15. Example of a SAT-Address task.

STV-Address

Image Inputs:

Prompt: *The following is a multiple-choice question about selecting the most appropriate address for a street view image.*
A. Bajiajiayuan, Xisanqi, Houbajia, Haidian District, Beijing, 100192, China
B. Beichen West Road, Chaoyang District, Beijing, 100101, China
C. Building A of Beichen Century Center, 8, Beichen West Road, Chaoyang District, Beijing, 100020, China
D. Nanyitiao of Zhongguancun, Keyu Community, Zhongguancun, Dongsheng, Haidian District, Beijing, 100190, China
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.

Reference: B **CityGPT-V**: B

Figure 16. Example of a STV-Address task.

ImgRetrieval. It evaluates capability to map a given street view image to the corresponding satellite image. An example is displayed in Figure 18.

CameraLoc. It requires the model to infer which quadrant of a satellite image corresponds to the location where a given street view image was captured. An example is shown in Figure 19.

SceneComp

Image Inputs:

Prompt: *In the provided four satellite images in urban area, which image shows most buildings?*
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: A

CityGPT-V: 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.

Figure 17. An example of a SceneComp task.

ImgRetrieval

Image Inputs:

Prompt: *You are given one street view image <image>and four satellite images. Which satellite image contains the street view image?*
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

CityGPT-V: C

Explanation: The street view image shows several large yet short building, indicating an industrial or business area. So the third RS image most probably contains its location.

Figure 18. An example of an ImagRetrieval task.

CameraLoc

Image Inputs:

Prompt: *You are given a satellite image <image> and a street view image <image> .You are given a satellite image and a street view image, and please predict which quadrant the street view image lies in the satellite image.*
A. Top left *B. Top right*
C. Bottom left *D. Bottom right*
Only provide one letter as the answer and please select your answer from A, B, C, or D.

Reference: B

CityGPT-V: B

Explanation: The street view image shows a broad road with no tall building in view. Thus it is not likely to be taken in the left or bottom part of this area.

Figure 19. An example of a CameraLoc task.

16. Urban Instruction Data

Table 10 provides detailed statistics of *UData* across three cities, while Table 11 presents the detailed statistics of the raw data used to construct *UData*. Additionally, we present representative examples of our urban instruction data, as shown in Figure 20 to Figure 35.

Table 10. Basic information of *UData* on three cities.

City	Category	Dataset	Instance	Rounds
/	General	ShareGPT,UltraChat,Open-Platypus	19866	3.7
Beijing	Location View Data	CityQA	19271	1
		Location Address	93246	1
		Landmark Details	51130	1
		Image Description	28798	1
		Cross Modality Reasoning	2000	1
	Trajectory View Data	Random Walk	9001	1
		Real-World Trajectory	98	1
		Visual Random Walk	8936	1
		Vision-Language Navigation	3000	1
	Global View Data	Image Content	9315	1
		Location Address	2777	1
		Landuse Inference	3642	1
		Multiple SAT Comparison	10114	1
		Cross-View Data	77204	1
		Cross Modality Reasoning	14977	1
London	Location View Data	CityQA	28934	1
		Location Address	2172	1
		Landmark Details	2372	1
		Image Description	716	1
		Cross Modality Reasoning	1286	1
	Trajectory View Data	Random Walk	16524	1
		Real-World Trajectory	98	1
		Visual Random Walk	13412	1
		Vision-Language Navigation	3000	1
	Global View Data	Image Content	3853	1
		Location Address	882	1
		Landuse Inference	4332	1
		Multiple SAT Comparison	4500	1
		Cross-View Data	2172	1
		Cross Modality Reasoning	5758	1
New York	Location View Data	CityQA	25413	1
		Location Address	94886	1
		Landmark Details	50404	1
		Image Description	24529	1
		Cross Modality Reasoning	2012	1
	Trajectory View Data	Random Walk	12277	1
		Real-World Trajectory	98	1
		Visual Random Walk	12229	1
		Vision-Language Navigation	3000	1
	Global View Data	Image Content	18368	1
		Location Address	5113	1
		Landuse Inference	17899	1
		Multiple SAT Comparison	22020	1
		Cross-View Data	94886	1
		Cross Modality Reasoning	23603	1

Table 11. The raw data of the selected region in three cities.

City	AoIs	PoIs	Roads	Trajectory	Street View Image	Satellite Image
Beijing	4647	1882	2320	21015	28798	1533
London	13705	11715	1322	173268	3125	556
New York	19541	11112	522	390934	24444	2738

User: <image >\n Could you provide a description of the content shown in the arial image.

Assistant: In the 256x256 pixel area, the road network is primarily composed of two main roads: the trunk road North Roud of Landianchang, which runs from [179, 249] to [178, 247] in a southward direction, and the service road Road of Kunminghu, extending from [127, 124] to [84, 16], which travels in a generally southwest direction. Additionally, there is a tertiary road Road of Kunminghu that connects [131, 123] to [127, 124], running parallel to the service road, and a secondary road North Roud of Landianchang that stretches from [176, 250] to [172, 248], also heading southward. The area features notable points of interest (POIs), including tickets located at [137, 175] and attractions at [88, 140]. The distribution of roads indicates a concentration of infrastructure towards the southern part of the region, particularly along North Roud of Landianchang, which is flanked by various POIs, while Road of Kunminghu serves as a key access route with a notable presence of attractions nearby, suggesting a vibrant area for visitors.

Figure 20. An example of global view training instances of Image Content.

User: <image >\n You are provided a 256*256 satellite image. What is the landuse type in region [[32, 136], [33, 167], [39, 167], [101, 133], [99, 127], [90, 131], [86, 121], [88, 114], [74, 111], [32, 136]]?

Assistant: Residential

Figure 21. An example of global view training instances of Landuse Inference.

User: <image >\n What is the detailed address of the provided street view image?

Assistant: Tsinghuayuan, Xueyuan Road, Dongsheng, Haidian District, Beijing, 100190, China

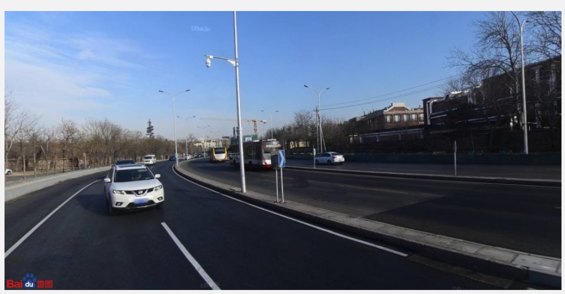
Figure 22. An example of local view training instances of Location Address.

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.

Figure 23. An example of local view training instances of Image Description.

User: <image >\n What is the landmark in this image? If there is no landmark, please say 'None'.



Assistant: In the street view image, a notable landmark is visible in the background: a pagoda, which is likely part of the nearby Changchun Garden. The pagoda stands tall among the trees, providing a distinctive silhouette against the sky, and serves as a navigational reference point for the surrounding area. The road curves gently around the landmark, with vehicles traveling in both directions, while the nearby buildings, including educational institutions like the Medical College and Pharmacy College, create a backdrop that emphasizes the pagoda's prominence in the landscape. This combination of natural and architectural features highlights the cultural significance of the pagoda within the urban environment

Figure 24. An example of local view training instances of Landmark Details.

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."

Figure 25. An example of local view training instances of Cross Modality Reasoning.

Assistant: You are given one street view image <image>and four satellite images. Which satellite image shows the surroundings of the street view image?
A. The first image <image>
B. The second image <image>
C. The third image <image>
D. The fourth image <image>
Please provide a detailed reasoning for your answer. And give me one letter as your answer in the last sentence.”

Assistant: Step 1: Describe the street view image: The image shows a parking area with various vehicles parked in a semi-organized manner. In the foreground, cars of different colors, including silver, black, and white, are parked closely on a paved surface made of large rectangular tiles, indicating moderate occupancy. To the right, a prominent blue fence encloses a structure, likely a building or series of buildings, contrasting with nearby lush trees. The background features multi-storied buildings made of brick and concrete, with visible windows, suggesting residential or commercial use. The parking area is in front of these buildings, with cars facing the blue enclosure. The short distance between the cars and the fence creates a compact environment. The clear sky suggests a bright day, enhancing visibility.
Step 2: Recall the address of the street view image: The image is taken at the School of Medicine, Jinchun Road, Qinghuayuan, Xiwacun, Haidian District, Beijing, 100084, China.
Step 3: Recall the addresses of each possible satellite view image:
...
Step 4: Reason the corresponding satellite view image: The satellite view image corresponding to the street view image is B

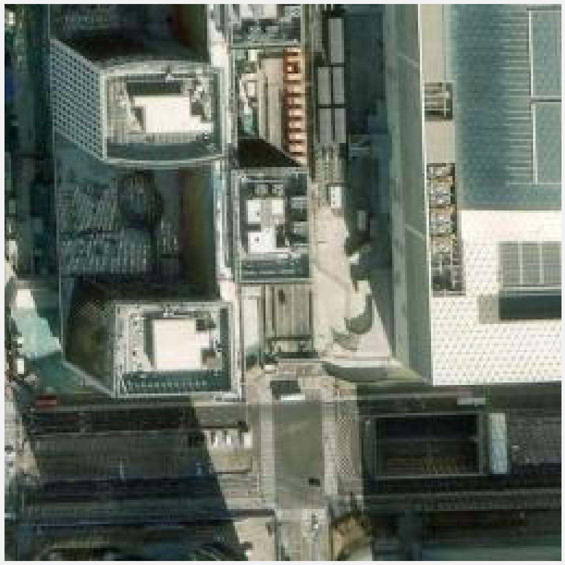
Figure 26. An example of global view training instances of Cross Modality Reasoning.

User: <image >\n What is the address of this location? Please think step by step and provide a detailed reasoning.

Assistant: Step 1: Identify the city name. According to the satellite view image, this image was taken in Beijing.
Step 2: Extract the location's features. From the image, I can observe the following features: In the 256x256 pixel area, the roads are primarily oriented in a north-south and east-west direction, with significant intersections facilitating access to various points of interest (POIs). Notably, the coordinates [116, 40] host both research institutes and universities, indicating a concentration of educational and research facilities in this region. These POIs are strategically located alongside the main north-south road that runs through the area, enhancing their accessibility for students and researchers alike. The distribution of roads in this region is fairly balanced, with a few major thoroughfares connecting key locations. The concentration of specific POIs, particularly educational institutions, suggests that this area serves as a hub for academic and research activities. Overall, the layout of roads and the placement of POIs create a vibrant environment conducive to learning and innovation.
Step 3: Identify the location's address. Based on my observation and knowledge of this region, the area is situated in Haidian District, Beijing, characterized by a 5x5 layout that features several prominent institutions. To the north, the Institute of Semiconductors, Chinese Academy of Sciences, occupies multiple locations, creating a significant presence in the region. Directly to the south, Beijing Forestry University is also located, providing an educational backdrop to the area. The main thoroughfare, Beilin Ginkgo Avenue, runs through the region, connecting these institutions and facilitating easy access. The Xueyuanlu Subdistrict, where this area is found, enhances the accessibility of the institutions, with Bajia Village nearby, further enriching the local community. The proximity of these educational and research institutions fosters a collaborative environment, making this area a hub for academic and scientific endeavors.

Figure 27. An example of global view training instances of Image Content.

User: <image >\n Please count the number of buildings in this satellite image and provide a detailed reasoning for your answer.



Assistant: Step 1: Identify the city name: According to the satellite view image, this image was likely taken in Beijing.
Step 2: Describe the location's features: From the image, I can observe the following features: In the 256x256 pixel area, the roads are arranged to facilitate access to various points of interest (POIs) and ensure smooth traffic flow. The tertiary road, Tianchen West Road, runs vertically from [137, 188] to [137, 171], providing a north-south route that connects to nearby POIs, including cars located at [47, 155]. The secondary road, Datun Road, extends horizontally from [174, 211] to [221, 209], serving as a major thoroughfare likely supporting higher traffic volumes and is near cafes located at [55, 148] and [61, 142]. Additionally, another segment of the tertiary road, Datun Road, stretches from [162, 187] to [0, 193], indicating a more extensive network that enhances connectivity across the region. The concentration of cafes near the secondary road suggests a vibrant social scene, while the presence of cars indicates active transportation throughout the area, particularly along these key roads. Overall, the layout of the roads and the distribution of POIs create a dynamic environment that caters to both vehicular movement and local amenities.
Step 3: Provide the location's address: Based on my observation and knowledge of this region, the area described is located in the Chaoyang District of Beijing, characterized by a well-organized 5x5 layout. To the north, you will find Beichen West Road, a prominent thoroughfare, while to the south, Datun North Road runs parallel, providing easy access to the surrounding locations. The eastern section is defined by Tianchen West Road, which intersects with various points of interest, including the Beichen Century Center, situated at the northwest corner. This center is flanked by additional buildings along Beichen West Road, creating a bustling environment. Adjacent to these landmarks, Datun Road runs through the area, connecting several establishments and enhancing the region's accessibility. The overall layout ensures that each location is conveniently reachable, fostering a vibrant community atmosphere.
Step 4: Count the number of buildings or POIs: There are 5 buildings in this image."

Figure 28. An example of global view training instances of Cross Modality Reasoning.

User: <image >\n What is the address of this location? Please think step by step and provide a detailed reasoning.

Assistant: Step 1: Describe the street view image: This is a street view image depicting a city street scene with a wide, well-maintained road extending into the distance. On the left side, a prominent orange guardrail runs parallel to the road, suggesting a separation from a lower area, possibly a sidewalk or a bus stop. The road is dark asphalt, reflecting the bright sunlight, indicating a clear day. To the right, a series of buildings are visible, showcasing a mix of architectural styles. The building closest to the viewer is painted in a muted red hue and features several air conditioning units mounted on its exterior walls. This building appears to be several stories tall, with large windows that allow for natural light. Further along, a taller, modern glass building stands out, its reflective surface capturing the sunlight and contrasting with the surrounding structures. This building is flanked by additional high-rise buildings in the background, creating a skyline that suggests a bustling urban environment. In the distance, road signs are visible, indicating directions or information for drivers, while the road continues to stretch ahead, leading towards the horizon. The overall scene conveys a sense of an active city, with a combination of modern architecture and well-maintained infrastructure. The clear blue sky above enhances the vibrant atmosphere of the setting.
Step 2: Tell the city name: According to the street view image, this is probably in Beijing.
Step 3: Extract the location's features around: The street view image is taken in a region with the following features: Beijing Haidian District Vocational School, Haidian District Vocational Technical School East Campus, Hongyun Laobeiping, Khan Baba, Mi Da Wan, and the Affiliated High School of China University of Geosciences.
Step 4: Tell the location's address: Based on my observation and knowledge about this region, the address is Xueyuanlu Subdistrict, Dongsheng, Haidian District, Beijing, 100083, China."

Figure 29. An example of local view training instances of Cross Modality Reasoning.

User: <image >\n The following is a multiple-choice question about selecting the most appropriate address for a satellite image.
A. The area is characterized by a well-organized layout, featuring a 5x5 grid that offers a variety of attractions and roadways. To the north, the prominent Yuan Dynasty Capital City Wall Ruins Park serves as a significant landmark, providing a historical backdrop and recreational space. Adjacent to this park, Beitou City West Road runs east to west, connecting various points of interest within the vicinity. To the east ...
B. The area is situated in the Huayuanlu Subdistrict of Haidian District, Beijing, characterized by a well-defined square layout divided into a 5x5 arrangement. To the north, the region is bordered by a series of residential complexes, including the prominent Beiji Temple Residential Area, which occupies several positions within the grid. To the south ...
C. The area is characterized by a well-organized 5x5 layout, with the northern section primarily featuring Wanquanhe Road, which runs through the heart of the neighborhood, providing easy access to various points of interest. To the east, ...
D. The area is located in the Huayuanlu Subdistrict of Haidian District, Beijing, characterized by a well-defined square layout that is predominantly residential. To the north, you will find a cluster of residential buildings known as Beiji Temple Residential Area, which extends down the western side of the area. Moving eastward, ...
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 explaination

Assistant: A

Figure 30. An example of global view training instances of Location Address.

User: <image >\n The following is a multiple-choice question about selecting the most possible landuse type in the region of a satellite image.
A. Industrial
B. Residential
C. Meadow
D. Forest
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 explaination.

Assistant: B

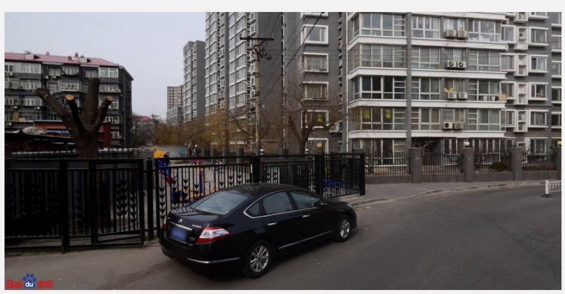
Figure 31. An example of global view training instances of Landuse Inference.

User: <image >\n The following is a multiple-choice question about selecting the most appropriate address for a street view image.
A. G6 Auxiliary Road, Chaoyang District, Beijing, 100000, China
B. Old Summer Palace Park, 28, Qinghua West Road, Yanyuan, Haidian Town, Haidian District, Beijing, 100084, China
C. Piaoliang Sunshine Plaza, Datun, Chaoyang District, Beijing, 100012, China
D. G6 Auxiliary Road, Xueyuanlu Subdistrict, Tayuan Village, Haidian District, Beijing, 100191, China
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 explaination.

Assistant: B

Figure 32. An example of local view training instances of Location Address.

User: <image >\n 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. Unpaved rural road.
B. Metal fence or wall.
C. Not identifiable.
D. Residential apartment complex.
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 explaination.



Assistant: D

Figure 33. An example of local view training instances of Landmark Details.

User: You are given a satellite image <image>and a street view image <image>, and please predict which quadrant the street view image lies in the satellite image.
A. Top left
B. Top right
C. Bottom left
D. Bottom right
Only provide one letter as the answer and please select your answer from A, B, C, or D.”

ASSITANT: D

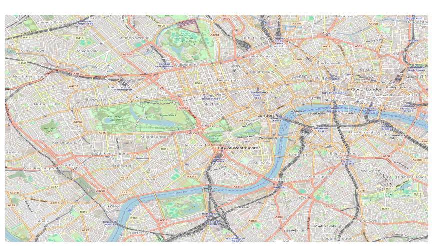
Figure 34. An example of global view training instances of Cross View Data.

User: You are given one street view image <image>and four satellite images. Which satellite image shows the surroundings of the street view image?
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.

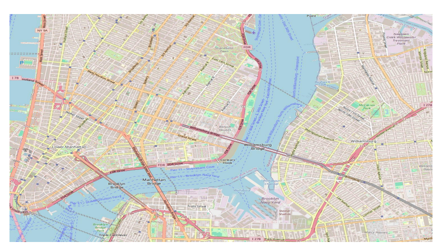
Assistant: C

Figure 35. An example of global view training instances of Multiple SAT Comparison.

(a) Beijing



(b) London



(c) New York

Figure 36. Maps for Beijing, London and New York.