

Unified Urban Tuning: Co-Enhancing Satellite and Street View Reasoning with a Progressive Tuning Framework

Supplementary Material

A. Experimental Settings

A.1. Implementation Details

For our final models, we employ Qwen2.5-VL-3B and Qwen2.5-VL-7B as base models. Our UUT framework is implemented as a two-stage progressive process.

Stage 1: UUT-SFT. The initial stage establishes the foundation for spatial reasoning. We fine-tune the models for 1 epoch on our UUT-SFT dataset mixed with general multimodal data to mitigate catastrophic forgetting. This stage is implemented using the LLaMA Factory framework and utilizes the AdamW optimizer with a global batch size of 128. Based on empirical studies, we set the base learning rate to $1e-5$. A dynamic data packing strategy is also employed to accelerate the training process.

Stage 2: UUT-RL. We further refine the SFT models for 10 epochs on our 20K-sample UUT-RL dataset to enhance long-chain reasoning capabilities. Training is conducted using the VeRL framework with a revised GRPO algorithm. We use the AdamW optimizer with a constant learning rate of $1e-6$ and a global batch size of 512. The sequence length remains set at 10,000 to accommodate long-form thought processes.

A.2. Evaluation

We evaluate our framework on four key benchmarks: CityBench, VRSBench, UrBench, and CityLens. The assessment spans single-input tasks (RS reasoning via VRSBench and CityBench-Population; SV reasoning via CityBench-Geoloc) and critical cross-view tasks. For cross-view reasoning, we use UrBench for spatial tasks (e.g., CR - City Retrieval, OR - Orientation) and CityLens for socio-economic predictions (e.g., GDP - Gross Domestic Product, VC - Violent Crime). We use standard metrics for each task: Accuracy (for VQA, referring, UrBench tasks), geolocalization accuracy (Acc@City, Acc@25km), R^2 /RMSE (for regression), and ROUGE (for captioning).

For our evaluation, we selected a representative subset of tasks from four key benchmarks: CityBench, VRSBench, UrBench, and CityLens. Our assessment is designed to quantify model performance across two primary categories: tasks operating on single input types and tasks requiring cross-view reasoning. To assess single input capabilities, we evaluate remote sensing (RS) reasoning using VRSBench and the Population task from CityBench; and street view (SV) reasoning using the Geoloc task from CityBench. To verify the model’s core ability for cross-view reasoning,

which is central to our framework, we use two benchmarks specifically designed for RS-SV pairs: UrBench (evaluating 8 distinct spatial tasks such as City Retrieval (CR), Orientation (OR), and Object Matching (OM)) and CityLens (evaluating socio-economic predictions like GDP, Population, and VC). Performance is measured using the standard metrics for each task, primarily: Accuracy for classification, VQA, referring, and retrieval/reasoning tasks (e.g., UrBench); geolocalization accuracy (e.g., City Accuracy, Acc@25km) for Geoloc; R^2 and RMSE for socio-economic and population regression; and text-generation metrics (e.g., ROUGE) for captioning.

A.3. Extended Evaluation on IM2GPS3k

To further evaluate the generalization capabilities of our framework, we extend our evaluation beyond the benchmarks presented in the main text to include **IM2GPS3K**, a widely recognized geolocalization benchmark. In addition to the models evaluated in the main manuscript (e.g., the RS-specific SoTA model, GeoChat), we also compare our approach against **additional task-specialized baselines** such as GeoCLIP and the GRE Suite.

As shown in Table 1, UUT-7B-RL consistently outperforms these specialized models in 4 out of 5 spatial tolerance metrics. These results further demonstrate that the performance gains achieved by our unified framework are robust, and are neither benchmark-specific nor an artifact of baseline selection.

A.4. Qualitative Analysis of Reasoning Capabilities

To qualitatively verify the effectiveness of our proposed framework, we present reasoning cases across diverse spatial tasks, ranging from city-level geolocation to fine-grained visual grounding. As shown in our qualitative results, the SFT+RL training paradigm empowers the model to perform explicit Chain-of-Thought (CoT) reasoning (enclosed in `<think>` tags) before generating final predictions.

Figure 1 illustrates the model’s reasoning process for open-world geolocation. By analyzing visual cues such as the “metal fence” and “building style,” the model explicitly aligns these features with regional contexts (e.g., recognizing the scene as Shanghai) to generate accurate coordinates.

Figure 2 demonstrates the model’s capability in understanding fine-grained spatial instructions. When prompted to locate an object “positioned at the bottom,” the model first interprets the directional constraint in its thought pro-

Table 1. Comparison of methods on the IM2GPS3K benchmark.

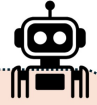
Methods	1km↑	25km↑	200km↑	750km↑	2500km↑
PlaNet	8.50	24.80	34.30	48.40	64.60
Translocator	7.60	20.30	27.10	40.70	63.30
GeoDecoder	5.70	10.30	21.40	28.90	38.60
GeoCLIP	14.11	34.47	50.65	69.67	83.82
GRE suite	11.30	35.30	51.70	69.30	85.70
Qwen2.5-VL-7B	3.97	25.48	43.65	64.29	77.90
UUT-7B-RL	9.80	40.18	53.19	71.45	86.38

cess and then predicts a bounding box that aligns closely with the ground truth, validating its precise object-level localization ability.

Figure 3 highlights the model’s robustness in matching features across different viewpoints (satellite vs. street view). The model successfully identifies shared landmarks—such as the layout of the paved lane and the white SUV—to deduce the camera’s precise location. This confirms the model’s ability to perform fine-grained feature alignment in complex cross-view scenarios.

Figure 4 showcases the model’s capacity for high-level reasoning with multi-source data. To estimate the “average walking-only travel time to healthcare,” the model synthesizes information from both satellite and street-view imagery. It follows a logical chain: identifying potential facilities, assessing road infrastructure, and estimating walking distances. The logical numerical prediction demonstrates the model’s potential in complex urban planning tasks.

References



Question :“Based on the provided image and contextual information, predict the geographical location where the image was taken.Please provide both the city name and coordinates.First, provide a reasoning analysis, and then give the final prediction.Please enclose the reasoning process within `</think> </think>` tags.The final answer should be in the following format:City: {city_name},Coordinates: {(latitude, longitude)}”



Correct answer : Shanghai, 31.2542&121.6213

Response: `<think>`

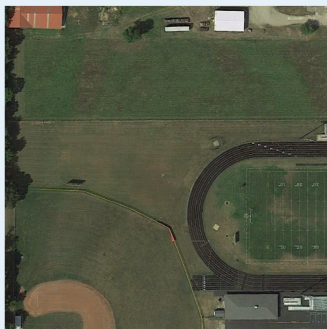
- Define the task: determine the city and coordinates by extracting **visible cues** from the image and aligning them with **Chinese urban context**; then verify with the given coordinates.
 - The image shows a **paved road** in the foreground, a **sidewalk lined with trees**, and a **metal fence**. Behind the fence is a **building partially obscured by greenery**. These concrete elements are taken as direct evidence.
 - The **metal fence**, the **building style**, and the **overall urban environment** suggest a **developed city**. The **fence design**, **tree types**, and **general layout** are consistent with **urban areas in China**, supporting a Chinese mega-city hypothesis.
 - Verify against regional geography to identify the **Shanghai** area, which aligns with the visual priors established above.
 - Finalize the precise location by corroborating the visual alignment with the coordinate verification.
- `</think>`

City: Shanghai ,Coordinates: (31.0707, 120.9780)

Figure 1. A qualitative example of spatial reasoning using the SFT+RL model. The model successfully identifies the geographical location by analyzing visual cues (e.g., metal fence, building style) within a Chain-of-Thought (CoT) process. The `< think >` block demonstrates how the model aligns visual evidence with regional context before generating the final coordinates.



Question :give me the location of `<p>The baseball field is positioned at the bottom of the image</p>`
Please provide the bounding box location of the described object. First, provide your visual analysis, and then give the bounding box coordinates. Please enclose the analysis process within `<think> </think>` tags.The final answer should be in format: `\bbox{"bbox_2d": [x1, y1, x2, y2]}`.



`<think>` The baseball field is positioned at the bottom of the image. The bounding box for the baseball field should be located at the bottom of the image, as indicated by the description.
`</think>`

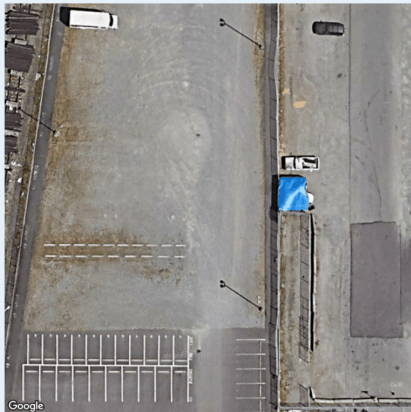
`\bbox{"bbox_2d": [129, 433, 254, 503]}`

Prediction: [129, 433, 254, 503]

Ground Truth: [111, 413, 267, 503]

IoU Score: 0.6232

Figure 2. Visual grounding example. The model successfully grounds the text description “baseball field positioned at the bottom” into precise pixel coordinates. The response highlights the step-by-step reasoning and the resulting bounding box prediction compared to the ground truth.



You are given a pair of images: one satellite/aerial image and one street view image. Task: Identify the approximate pixel coordinates in the satellite image where the street view was captured.\n\nInstructions:\n1. Analyze spatial relationships and landmarks (2-3 sentences maximum)\n2. Enclose your analysis within `<think>` `</think>` tags\n3. After closing the `</think>` tag, provide the pixel coordinates\n4. Format your final answer as: `\\bbox{"center_point": [x, y]}`\nNow provide your answer:"

`<think>`The street view shows a paved driving lane with a gravel/dirt area to the immediate left and a chain-link fence to the right. This matches the satellite image, which displays a vertical paved lane with a large unpaved section on the left side and a fence line on the right. Key landmarks aligning the two views include the white SUV parked along the fence (visible in both the street view and the mid-right of the satellite image) and the transverse dashed road markings on the pavement, which are visible in the foreground of the street view and appear in the lower-middle section of the satellite image. The camera location is estimated to be on the lane, just "below" (south of) the dashed lines.`</think>` `\\bbox{"center_point": [275, 325]}`

Ground_truth: {center_point: [480, 160]} Score: 0.4185

Figure 3. Reasoning-based cross-view localization. The model aligns the street-view perspective with the satellite imagery by identifying shared landmarks like the white SUV and road markings. The explicit reasoning chain leads to a coordinate prediction that spatially corresponds to the ground truth location.



SFT+RL

Satellite Image



Street view image



(1)



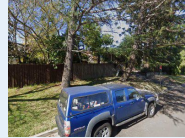
(2)



(3)



(4)



(5)



(6)



(7)



(8)



(9)



(10)

Suppose you are a professional socioeconomic data analyst in Sydney. Based on the provided satellite image and several street view photos taken within the same area covered by the satellite image, please estimate 'the average walking-only travel time to healthcare' for this area. Consider factors such as building structures, road infrastructure, visible traffic patterns, land use characteristics, greenery, and any other relevant features. Please first provide a detailed reasoning analysis, then give your final numerical estimate. Enclose the reasoning process within `<think>` tags. The final answer should be a single number, enclosed in `\boxed{}`.

Response: `<think>` To estimate the average walking-only travel time to healthcare in this Sydney suburb, we must synthesize information from the satellite image and the street view photos to infer the spatial layout, land use, and accessibility of healthcare facilities

Step 1: Identify healthcare facilities. The satellite image does not explicitly show healthcare facilities (e.g., clinics, hospitals, pharmacies).

However, street views suggest a mixed-use environment with residential houses, small commercial buildings (e.g., the one with a sign in the first image), and a large green space (possibly a park or school). In such suburban settings, healthcare services are often located in commercial zones or within residential streets. The most plausible locations for healthcare are: The commercial building with signage (possibly a pharmacy or clinic). The large green space (could be a park with a health center or clinic). The residential streets (where a doctor's office or clinic might be located).

Step 2: Estimate walking distance to the nearest healthcare. From the street views, we can observe:

The street with parked cars (image 2) appears to be a main road with moderate traffic, but no visible traffic congestion. The residential streets (images 4, 10) are quieter, with houses and yards. The green space (image 5) is large and appears to be a park, which may have a health center or clinic nearby. The commercial building (image 2) is likely the closest healthcare facility, as it is a small commercial structure with signage.

Step 3: Estimate walking distance. Assuming the commercial building is the nearest healthcare facility, we can estimate the walking distance from the center of the area shown in the satellite image to this building. From the satellite image, the commercial building is located near the center of the image, surrounded by residential houses and a large green space. The distance from the center of the image to the commercial building is approximately 200-300 meters, based on the scale of the satellite image and the visible street layout.

Step 4: Estimate walking time. Walking speed in urban areas is typically 4-5 km/h. For a distance of 200-300 meters (0.2-0.3 km), the walking time would be approximately 3-5 minutes.

Step 5: Consider other factors.

The area has a mix of residential and commercial buildings, which suggests that healthcare facilities are likely to be within walking distance.

The streets are relatively quiet, with no visible traffic congestion, which suggests that walking is safe and efficient. The green space is large and appears to be a park, which may have a health center or clinic nearby. Therefore, the average walking-only travel time to healthcare in this area is estimated to be 3-5 minutes.

Final Answer: `\boxed{3}` minutes. `</think>`

Ground Truth: 1.2 **Score :** 0.37115201687331734

Figure 4. Multi-modal urban reasoning example. Given satellite and street-view imagery, the model performs a step-by-step socioeconomic analysis to estimate healthcare accessibility. The reasoning process covers facility identification, distance estimation, and travel time calculation, resulting in a logical numerical prediction.