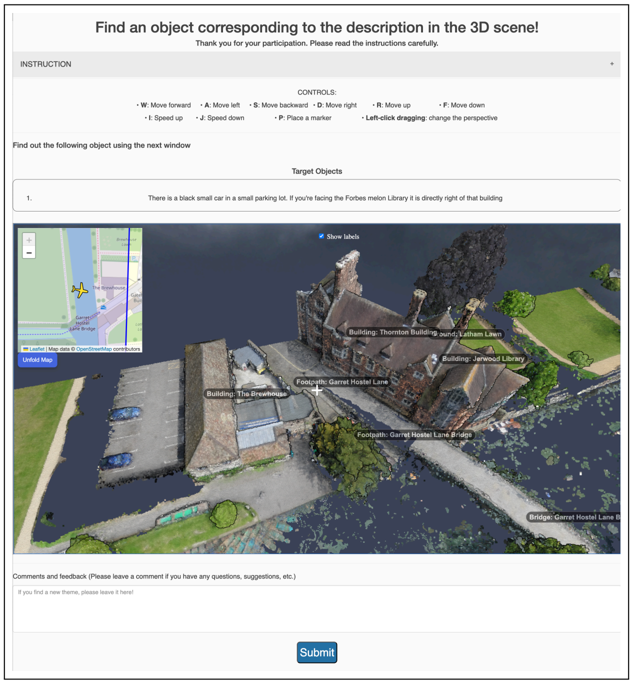
APPENDIX

This is appendix for the paper: *CityNav: A Large-Scale Dataset for Real-World Aerial Navigation*. We present additional details of the data collection interface, dataset statistics, models, and experimental results.

A. Data Collection Interface

We developed the data collection website using the Amazon Mechanical Turk platform. Figure 12 displays a full screenshot of the web interface, enabling users to operate an aerial agent within the CityFlight environment.



Find an object corresponding to the description in the 3D scene!

Thank you for your participation. Please read the instructions carefully.

INSTRUCTION +

CONTROLS:

- **W**: Move forward
- **A**: Move left
- **S**: Move backward
- **D**: Move right
- **R**: Move up
- **F**: Move down
- **I**: Speed up
- **J**: Speed down
- **P**: Place a marker
- **Left-click dragging**: change the perspective

Find out the following object using the next window

Target Objects

1. There is a black small car in a small parking lot. If you're facing the Forbes melon Library it is directly right of that building

☑ Show labels

Comments and feedback (Please leave a comment if you have any questions, suggestions, etc.)

If you find a new theme, please leave it here!

+
−
The Brewhouse
Garret Hostel Lane Bridge
Leaflet | Map data © OpenStreetMap contributors
Unfold Map
Building: Thornton Building
ound: Latham Lawn
Building: Jerwood Library
Footpath: Garret Hostel Lane
Building: The Brewhouse
Footpath: Garret Hostel Lane Bridge
Bridge: Garret Hostel Lane B
Submit

Figure 12. **Data collection interface**. Full screenshot of web interface for collecting human demonstration trajectories for the CityNav dataset.

B. Dataset Statistics

Agent Altitude During Operation. We analyze human-operated flights to better understand altitude behavior during navigation tasks. Figure 13 shows the mean altitude of human-operated agent trajectories, segmented into 20-meter intervals based on distance from the goal. Given that the average 3D altitude is 35.96 meters, this result indicates that most human operators flew above building-level heights, gradually descending as they approached their targets. In addition, we investigate how clearly ground-level objects are visible at these flight altitudes. Figure 14 shows a top-down view illustrating that human pilots typically navigate with clear visual access to the target objects. Given that the average altitude is above buildings, pilots can effectively identify and target landmarks.

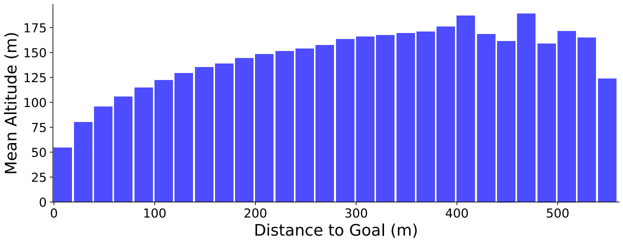


Figure 13. Relationship between the distance to goal and the mean altitude of aerial agents.

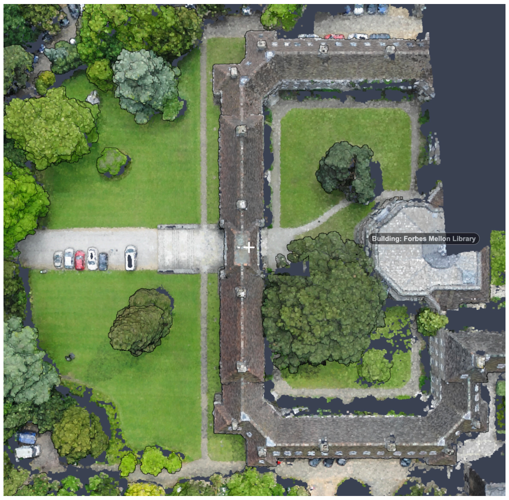


Figure 14. Top-down view of the aerial agent at an altitude of 150m, captured via the web interface.

Human Navigation Strategy. In the aerial VLN task, the exploration space is vast, making it crucial to narrow down the search area. To address this, our approach mimics the way humans leverage geographic information (landmarks) to reduce the exploration range. As illustrated in Figure 1, human demonstrations rely on the landmarks mentioned in the description (e.g., *Sidney Street*) to navigate toward the landmark's vicinity. Once near the landmark, humans focus their search on the area around it to find the goal object. This human strategy enables efficient navigation by focusing efforts around landmarks.

To validate this concept, we analyzed the trajectory data collected in the CityNav dataset, which includes geographic information. The results indicate that agents passed directly over landmarks 36.3% of the time in human demonstration (HD) trajectories, compared to 24.6% in shortest-path (SP) trajectories. Additionally, we examined whether agents passed within a certain radius of the landmark center. Within 20 meters, 35.5% of HD trajectories passed near a landmark, compared to 24.0% for SP. Similarly, at a 40-

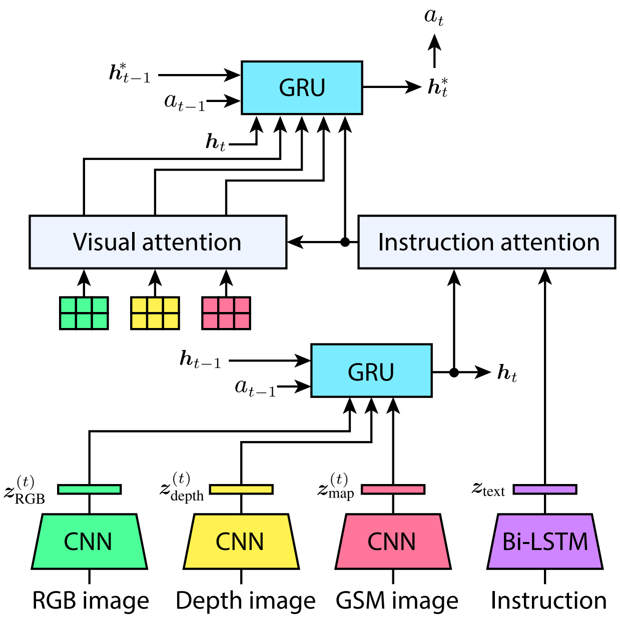


Figure 15. Architecture of AerialVLN+GSM.

meter radius, 62.5% of HD trajectories were near a landmark, compared to 51.9% for SP. These results suggest that human pilots tend to navigate closer to landmarks—a strategy that likely contributes to the superior performance of our GSM-based method leveraging human demonstration trajectories, as observed in Table 3 and 4.

C. Model Details

C.1. AerialVLN+GSM architecture

The architecture of AerialVLN+GSM is shown in Figure 15. It utilizes four input modalities: RGB images, depth images, GSM images, and textual navigation descriptions. For RGB images, a ResNet-50 encoder pre-trained on ImageNet is used. The input size is 224 × 224. For depth images, another ResNet-50 encoder pre-trained on PointGoalNav is used. The input size is 256 × 256. The GSM encoder is a convolutional network consisting of five 2D convolutional layers with channel sizes of (32, 64, 128, 64, 32), kernel size of three, stride of one, and padding of one. Each convolutional layer is followed by a ReLU activation and max-pooling operation. The input size is 224 × 224. The text encoder is implemented using a Bi-LSTM. The output embedding $\boldsymbol{h}^*_t$ is obtained through two GRU modules integrated with description attention and visual attention mechanisms. Specifically, embeddings from the three visual modalities are first fed into the first GRU to produce an intermediate embedding $\boldsymbol{h}_t$. Subsequently, description attention followed by visual attention for each modality is applied to $\boldsymbol{h}_t$ and the corresponding modality-specific features. Finally, the second GRU aggregates the outputs from these attention modules to yield the final output embedding $\boldsymbol{h}^*_t$. The action is predicted from $\boldsymbol{h}^*_t$ through a learnable linear layer.

C.2. Geographic Semantic Map

The GSM consists of five categories: current field of view, explored area, landmarks, potential goals, and surrounding objects. These categories are selected because it is essential to understand the spatial relationships between the explored area and objects. The current field of view and explored area are acquired from GNSS coordinates. Specifically, these coordinates are obtained from the CityFlight environment at each time step, and the square area corresponding to the top-down UAV view is marked with a value of one in a binary mask. Landmarks are segments retrieved from OpenStreetMap. For each landmark name, the corresponding segment is retrieved. Potential goals and surrounding objects are detected using an object detector. We used GroundingDINO [28] due to its strong performance in open-set object detection. The detection prompt includes both object categories defined in the SensatUrban dataset and object names extracted from the navigation descriptions (*e.g.*, "a building with a grey roof" and "a red van with black stripes"), to detect object regions from the current RGB image. Before navigation begins, landmark and object names are extracted using a language model (GPT-3.5). The original GSM size corresponds to the smallest 2D map that encompasses the entire 3D scene. Finally, the GSM is resized to 224×224 pixels and provided as input to the model.

C.3. Training

All models were trained on a single GeForce RTX 4090 GPU. The Adam optimizer was used for 5 epochs, with an initial learning rate of 5 and a batch size of 12. Cross-entropy loss and an MSE loss, which measures the distance between the goal point and the current position, were employed. For AerialVLN, the step parameter for the look-ahead guidance was set to 10.

D. Additional Analysis

Category-level performance. We analyze performance at the category level since descriptions can refer to different goal types. Table 6 shows that AerialVLN+GSM generally delivers the best results, suggesting that integrating the state-of-the-art AerialVLN model with GSM significantly enhances navigation performance at the category level. Although CMA+GSM also shows improvements, it lags behind AerialVLN+GSM, and while Seq2Seq+GSM performs better than its baseline, it remains less effective than the other GSM-enhanced models. Overall, the ground and others categories pose particular challenges for baseline

Category	Method	NE↓	SR↑	OSR↑	SPL↑
Building	Seq2Seq	244.67	1.98	8.50	1.68
	Seq2Seq+GSM	100.97	3.24	13.00	3.10
	CMA	253.16	0.76	8.73	0.72
	CMA+GSM	95.70	4.86	14.35	4.80
	AerialVLN	197.51	1.71	4.00	1.61
	AerialVLN+GSM	**87.40**	**6.52**	**16.91**	**6.42**
	Human	11.3	85.64	93.21	57.26
Car	Seq2Seq	233.08	1.30	9.31	1.19
	Seq2Seq+GSM	95.78	4.11	15.44	3.96
	CMA	239.24	0.87	11.76	0.85
	CMA+GSM	90.99	4.98	17.75	4.95
	AerialVLN	164.29	2.38	4.62	2.31
	AerialVLN+GSM	**84.78**	**7.65**	**18.76**	**7.52**
	Human	6.7	95.39	97.00	67.89
Ground	Seq2Seq	278.82	0.59	6.93	0.59
	Seq2Seq+GSM	88.67	3.76	14.06	3.64
	CMA	294.39	1.19	7.33	1.17
	CMA+GSM	82.31	4.16	13.47	4.06
	AerialVLN	208.63	0.79	2.38	0.78
	AerialVLN+GSM	**73.05**	**5.94**	**19.60**	**5.87**
	Human	12.0	82.40	92.42	55.64
Ohters	Seq2Seq	245.44	0.00	3.64	0.00
	Seq2Seq+GSM	98.97	**3.64**	10.30	**3.64**
	CMA	232.95	0.61	9.70	0.61
	CMA+GSM	89.81	**3.64**	12.73	3.60
	AerialVLN	182.68	1.21	2.42	1.21
	AerialVLN+GSM	**84.65**	**3.64**	**21.21**	3.40
	Human	13.9	76.97	86.84	54.11

Table 6. Performance of each method at the category level.

methods, yet GSM integration helps mitigate these difficulties. These findings underscore the value of a geographic semantic map for improving aerial VLN across diverse object categories. Furthermore, the comparison with human performance highlights the gap between aerial agents and human navigation capabilities, with AerialVLN+GSM approaching human-like performance in some metrics while still leaving room for further improvement.

Disaster scenarios. Disaster search is one of practical applications as the target's location is unknown. We created 2D simulation data for flood scenarios. Table 7 summarizes the navigation performance. As shown, all models exhibit reduced performance; however, the effectiveness of GSM remains. Simulating other types of disasters and more dynamic scenarios is left for future work.

Method	NE↓	SR↑	OSR↑	SPL↑
Seq2Seq	288.5	1.38	11.58	0.69
Seq2Seq+GSM	98.8	3.97	14.4	2.89
CMA	273.1	0.6	9.27	0.4
CMA+GSM	92.5	4.61	15.63	3.47
AerialVLN	188.6	1.46	4.65	1.38
AerialVLN+GSM	**84.9**	**6.80**	**18.46**	**6.68**

Table 7. Navigation performance under flood inundation conditions (test-unseen).