AirSim360: A Panoramic Simulation Platform within Drone View

This Supplementary Material provides technical details, comprehensive dataset statistics, and additional implementation specifics that were omitted from the main paper due to space constraints. Specifically, we present the following information:

- In Section 1, we provide some details of the three subsets of the Omni360-X dataset: Omni360-Scene, Omni360-Human, and Omni360-WayPoint, including the semantic categories and pedestrian behaviors.
- In Section 2, we elaborate on the mathematical model and constraints of the Minimum Snap trajectory planning method used for automated trajectory generation.
- In Section 3, we provide comprehensive experimental configurations for the Monocular Pedestrian Distance Estimation (MPDE) and Panoramic Vision-Language Navigation (VLN) tasks.

1. More Details of Omni360-X Dataset

1.1. Omni360-Scene Statistics

Omni360-Scene provides pixel-level annotations for depth information, semantic segmentation, and entity segmentation across diverse environments. As semantic complexity varies by scene, Table 1 provides a visual overview including the ERP image and semantic segmentation mask for each scenario, followed by a detailed breakdown of the semantic categories.

1.2. Omni360-Human Statistics

The Omni360-Human subset is dedicated to human-centric perception tasks, primarily monocular pedestrian distance estimation.

Data Statistics: Table 2 presents a detailed breakdown of the dataset composition. The data was collected across 6 distinct scenarios , covering a wide range of crowd densities and area sizes. The dataset includes over 100K frames in total, with varying numbers of NPCs to simulate realistic crowd dynamics.

1.3. Omni360-WayPoint Statistics

Omni360-WayPoint provides physics-consistent UAV flight paths for navigation, trajectory prediction, and control. The trajectories adhere to realistic flight dynamics derived from Minimum Snap planning. Table 4 and Table 3 details the key kinematic parameters and scale of the waypoint data.

$$\mathbf{S}(t) = \begin{bmatrix} \mathbf{p}(t)^T \\ \mathbf{v}(t)^T \\ \mathbf{a}(t)^T \end{bmatrix} = \begin{bmatrix} x(t) & y(t) & z(t) \\ v_x(t) & v_y(t) & v_z(t) \\ a_x(t) & a_y(t) & a_z(t) \end{bmatrix} \tag{1}$$

2. Minimum Snap Trajectory Planning Implementation Details

The Automated Trajectory Generation Paradigm employs Minimum Snap trajectory planning to produce smooth, dynamically feasible UAV flight paths from sparse user-defined waypoints. This method minimizes the integrated square of the fourth derivative of position (Snap), effectively ensuring trajectory smoothness and reduced control effort.

Polynomial Representation. Given a sequence of key waypoints $\{p_0, p_1, \ldots, p_M\}$, each segment of the trajectory is modeled as a fifth-order polynomial:

$$p_i(t) = a_{i,0} + a_{i,1}t + a_{i,2}t^2 + a_{i,3}t^3 + a_{i,4}t^4 + a_{i,5}t^5, \tag{2}$$

where $\mathbf{a}_i = [a_{i,0}, \ldots, a_{i,5}]^\top$ are the polynomial coefficients for segment i.

Optimization Objective. Following the Minimum Snap formulation, the smoothness of the trajectory is achieved by minimizing the integral of the squared fourth derivative (snap):

$$J = \int_{t_0}^{t_M} \left\| \frac{d^4 p(t)}{dt^4} \right\|^2 dt. \tag{3}$$

Quadratic Programming Formulation. The optimization problem can be expressed as a quadratic program:

$$\begin{aligned} \min_{\mathbf{a}} \quad & \mathbf{a}^\top Q \mathbf{a}, \\ \text{s.t.} \quad & A\mathbf{a} = b, \end{aligned} \tag{4}$$

Table 1. Visualization of semantic segmentation and list of semantic categories for each scene.

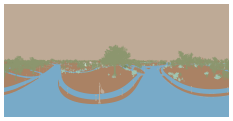
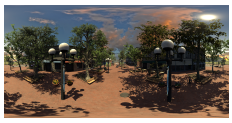
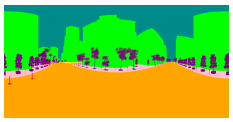
Scene Name	ERP Image	Semantic Vis	Semantic Categories
City Park			Building, Rock, AmurCork, Bush, Elm, Ivy, Maple, WeepingWillow, PlayGround, Bench, LampPost, FoodStalls, Cafechair, Roadblock, Trashcan, Trafficbarrel, Circlefence, Trafficlight, Water Plane, Road, Cafetable, Umbrella, Pool Sidewalk, Sky, Landscape
Downtown West			Building, Awning, Roof, Tree Generic, Tree Narrowleaf, Tree Pine, Prop Dining Table, Umbrella, Prop Dining Chair, Pot, Car Pillar, Recycle Bin, Food Cart, Bench Wood, Poster Stand, Ground Mod, Road, Lightpost Light Post, Tarppost, Light Streetlight Complete, Tarponly, Ground Park Walkway, Rock Rock, Background Mountains, Wood Fence Wood Fence, Prop Park Railing Rail, Prop Park Railing Pillar, Sky, Landscape
SF City			Building, Sidewalk, Road, Bus, Fence, Cone, Hydrant, Parkingmeter, Stopstation, Elecbox, Trash, Traffictube, Barrier, Alamppost, Blamppost, Lake, Bollardrope, Barricademetal, Tree, Sky
New York City			Concreteblock, Streetprops, Plasticcone, Metalfence, Pillar, Lamp, Trashcan, Postbox, Umbrella, Table, Chair, Greenpot, Roadcolumn, Adplane, Buidlingawning, Scaffolding, Usaflag, Plant, Ventilationtube, Building, Hotdogpot, Road, Sidewalk, Grounddirt, Sky

Table 2. Detailed Statistics of the Omni360-Human Dataset.

Scene	Subsets	Area Range ($m \times m$)	NPC Count (Min-Max)	Total Frames
New York City	14	$12 \times 12 \sim 30 \times 30$	$15 \sim 45$	29,000
LisbonDowtown	10	$12 \times 12 \sim 30 \times 50$	$10 \sim 45$	9,000
Downtown City	17	$12 \times 12 \sim 30 \times 30$	$8 \sim 30$	27,000
Roof	7	$12 \times 12 \sim 45 \times 20$	$5 \sim 30$	11,200
Rural Cabins	2	$15 \times 15 \sim 15 \times 30$	$7 \sim 14$	4,000
Rome	11	$8 \times 10 \sim 50 \times 30$	$4 \sim 30$	20,500
Total	**61**	-	-	**100,700**

Table 3. Introduction of Omni360-WayPoint. The Kinematic Parameters include two distinct sets of $a_{\max}$, $v_{\max}$, sampling interval t, each representing a typical UAV flight condition. The Total Number of Flight Paths is computed as the product of the number of Kinematic Parameter sets and the Number of Routes.

Scenario	Length Range	Kinematic Parameters	Number of Routes	Total Number of Flight Paths
City Park	[50, 150]	[(3, 16, 0.5), (5, 21, 1)]	20000	40000
Downtown West	[20, 50]	[(3, 16, 0.5), (5, 21, 1)]	5000	10000
New York City	[20, 50]	[(3, 16, 0.5), (5, 21, 1)]	5000	10000
SF City	[50, 150]	[(3, 16, 0.5), (5, 21, 1)]	20000	40000

where Q is derived from the cost in (3), and A, b encode waypoint and continuity constraints up to the third derivative. Solving this system yields the polynomial coefficients $\mathbf{a}$ defining the minimum-snap trajectory.

Table 4. Inputs and outputs of the trajectory way-points generation algorithm. The input $v_{\max}$ denotes the predefined maximum flight speed, $a_{\max}$ represents the maximum aircraft acceleration, and t is the sampling interval. The output parameters are also described in Eq. (1).

Minimum Snap	Parameters		
Input	$v_{\max}$	$a_{\max}$	t
Output	$\mathbf{p}(t)$	$\mathbf{v}(t)$	$\mathbf{a}(t)$

Dynamic Feasibility. To ensure physical feasibility, the trajectory is further constrained by dynamic limits on velocity and acceleration:

$$\|\dot{p}(t)\| \leq v_{\max}, \qquad \|\ddot{p}(t)\| \leq a_{\max}. \tag{5}$$

Each segment duration ΔT_i is automatically adjusted according to these limits, as well as the chosen sampling interval Δt, ensuring that the resulting trajectory

Table 5. MPDE results across different training and testing datasets. **Dist. Err (All)** denotes the weighted average over all four datasets, where **Dist** refers to the Euclidean distance. **Ang. Err (All)** denotes the weighted average over all four datasets, where **Ang** refers to the angle. The **Pub** column in the last two columns indicates that the weighted average is computed over only the top three public datasets.

Training Set	Test Set	Dist. Err	Samples	Dist. Err (All)	Ang. Err	Ang. Err (All)	Ang. Err (Pub)	Dist. Err (Pub)
nuScenes	nuScenes	1.078	15369	0.80	31.90	23.14	21.207	0.484
	KITTI	0.822	1759		31.50			
	FreeMan	0.260	43361		17.00			
	Omni360-Human	2.439	11496		33.30			
nuScenes + Omni360-Human-all	nuScenes	1.073	15369	**0.43**	30.70	**16.25**	17.282	0.449
	KITTI	0.802	1759		32.70			
	FreeMan	0.213	43361		11.90			
	Omni360-Human	0.313	11496		10.80			
nuScenes + Omni360-Human-pitch_0	nuScenes	1.071	15369	0.50	30.70	19.08	19.194	**0.433**
	KITTI	0.812	1759		31.90			
	FreeMan	0.191	43361		14.60			
	Omni360-Human	0.868	11496		18.50			
nuScenes + Omni360-Human-pitch_20	nuScenes	1.068	15369	0.67	30.70	16.73	**17.023**	0.458
	KITTI	0.809	1759		31.20			
	FreeMan	0.228	43361		11.60			
	Omni360-Human	1.779	11496		15.20			

remains dynamically executable by the UAV controller.

3. Experimental Details

This section provides additional implementation details and experimental settings for the Monocular Pedestrian Distance Estimation (MPDE) and Panoramic Vision-Language Navigation (VLN) tasks presented in the main paper.

3.1. Monocular Pedestrian Distance Estimation (MPDE)

In the Monocular Pedestrian Distance Estimation experiments, we design four sets of evaluations to demonstrate the effectiveness of our data. We first report the results on all test sets using only the nuScenes dataset. We then conduct a series of comparative experiments on three configurations of the Omni360-Human dataset: the full dataset, the subset with a pitch angle of 0°, and the subset with a pitch angle of 20°.

All models are trained using the AdamW optimizer with an initial learning rate of 0.002 and a weight decay coefficient of 0.01. The learning rate is multiplied by 0.98 every 300 steps during training.

The Omni360-Human training set is curated to exclude any samples from the Omni360-Human test set used in the experiments.

3.2. Panoramic Vision-Language Navigation (VLN)

In Visual Language Navigation (VLN), the formulation of prompts plays a critical role in determining evaluation metrics such as the Success Rate (SR). To ensure transparency and fairness in our experiments, we publicly release all prompts used in Table 6. It should be noted that, in addition to prompt formulation, factors including the frame rate of the simulator platform and the latency of online model invocation may also influence SR. The central aim of this work, however, is to introduce a highly challenging and promising new task based on the Airsim360 platform. Therefore, our experimental design prioritizes the most impactful factor, namely the formulation of prompts, while a comprehensive analysis of other variables remains outside the scope of this study.

Table 6. List of Prompts. The following prompts are used in two different environments, namely the New York City scene and the 1950s NYC Environment Megapack scene.

Prompt
Find the nearest traffic light and stop when you reach it.
Find the nearest blue mailbox and stop when you reach it.
Find the nearest tall building straight ahead and stop when you reach its rooftop.
Move forward, then at the intersection, you'll see a red telephone booth on your right. Stop near the closest one.
Fly across the lake in front of you, reach the opposite neighborhood, and stop on the street.
Find the nearby lake surrounded by woods and stop at the nearest shore.
Locate the building nearby with a giant Coca-Cola bottle decoration on its roof and fly close to the decoration.
Cross the zebra crossing and fly over this section of the road.
Fly to the small island in the center of the lake and land.
Fly straight ahead, turn right at the intersection, and stop near the bridge.
Fly along the current road and stop when you reach the second tree.
Stop near the small fountain located downstairs in the nearby building.
You are currently on the left side of the bridge. Now move to the right side and stop.
Climb over the fence in front of you and stop on the path in the park ahead.
Fly to the blue billboard on the building ahead and to your left, then stop nearby.
Fly to the tree with red leaves on the left side of the street and stop nearby.
Fly to the vicinity of the three very similar buildings straight ahead and stop.
Locate the billboard straight ahead featuring a person in a blue suit, fly to it, and stop nearby.
Fly to the billboard with the red car on the building ahead above you and stop nearby.
Find the floor in the building in front of you with red curtains and stop nearby.
Locate the billboard with the black and white portrait on the building ahead and stop nearby.
Arrive at the bank with the purple sign and stop downstairs.
Find the nearest yellow sunshade among the many downstairs and stop there.
Fly along the crosswalk over the intersection and stop on the opposite side of the road.
Find the nearest American flag and stop there.
Continue flying straight ahead and stop when you reach the intersection with the main road.
Fly to the blue barrier ahead and stop.
Fly to the red phone booth behind of you and stop when you are close the red phone booth.
Fly to the blue billboard on your right rear and stop when you're close to it.
Fly to the lake surrounded by trees on your right and stop when you're close to it.
Fly to the red bridge on your right and stop when you're close to it.
Stop near the nearest lawn.
Find the nearest traffic light and stop near it.
Navigate to the nearest green bike lane and stop.
Fly to the nearest billboard that shows BLACK & WHITE and stop nearby.
Navigate to the orange mailbox ahead and stop nearby.
Fly to the nearest food truck and stop.