AERIALDOJO-200K: A LARGE-SCALE BENCHMARK SUITE FOR OPEN-WORLD AERIAL OBJECT-GOAL SEARCH

Anonymous authors
Paper under double-blind review

ABSTRACT

Open-world aerial object-goal search is a foundational yet challenging task, requiring aerial agents to autonomously explore large-scale, unstructured three-dimensional environments and reach target objects specified by semantic descriptions or reference images, rather than following route-specific instructions. However, research in this task remains at a nascent stage and relies on small, environment-specific benchmarks with heterogeneous action spaces and data formats. These limitations hinder large-scale training and cross-benchmark evaluation, constraining the scalability and generalizability of aerial agents. To address this problem, we propose AerialDojo-200K, a large-scale benchmark suite for open-world aerial object-goal search, with $3\times$ (times) as many scenes and $18.7\times$ as many task instances as the largest existing benchmark for this task. Specifically, we construct 42 simulation scenes spanning four scene families and 21 scene types, including 18 urban, 12 natural, six infrastructure, and six disaster scenes. To ensure data quality, 12 annotators spent two months manually annotating 109 landmarks, 2099 target objects, and 2099 object anchors across these scenes. We further construct 205,732 task instances, comprising over 100K semantic-goal and over 100K image-goal instances across Base, Standard, and Long-Horizon settings. Each task instance includes a collision-free reference trajectory and corresponding multi-view video recordings. We also develop a unified evaluation framework with a scene partition comprising 21 in-distribution scenes and 21 out-of-distribution scenes. Finally, our evaluation of five open-source and four closed-source multimodal large language models reveals that there is still a long way to go toward achieving general-purpose aerial agents.

1 INTRODUCTION

Open-world Aerial Object-Goal Search (AerialOGS) (Xiao et al., 2025b) requires aerial agents to autonomously search for target objects through goal-driven, long-horizon exploration in large-scale, unstructured three-dimensional environments, and to actively stop when onboard observations match the target object specified by a semantic description or reference image. Unlike aerial Vision-and-Language Navigation (VLN), which follows detailed, step-by-step route instructions that constrain how the aerial agent should travel (Zhao et al., 2026), AerialOGS provides only a high-level, open-ended object-goal and must jointly resolve where to explore, how to navigate, and when the observed object satisfies the goal. AerialOGS has broad potential applications in search and rescue, infrastructure inspection, and autonomous delivery, where operators may specify the object without knowing its exact location or providing a complete flight route.

Research on AerialOGS tasks remains in a nascent stage. On the one hand, existing aerial agents utilize spatio-semantic representations (Zhang et al., 2026) to organize memories and plan over explored regions, frontiers, and object candidates. On the other hand, recent studies demonstrate that multimodal large models can leverage high-level semantic reasoning to interpret semantic goals (Xiao et al., 2025b) and guide search decisions (Chen et al., 2026).

Nevertheless, the primary performance bottleneck remains the lack of large-scale benchmarks for training and evaluation. As shown in Table 1, 1) UAV-ON (Xiao et al., 2025a) and CityAVOS (Ji

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

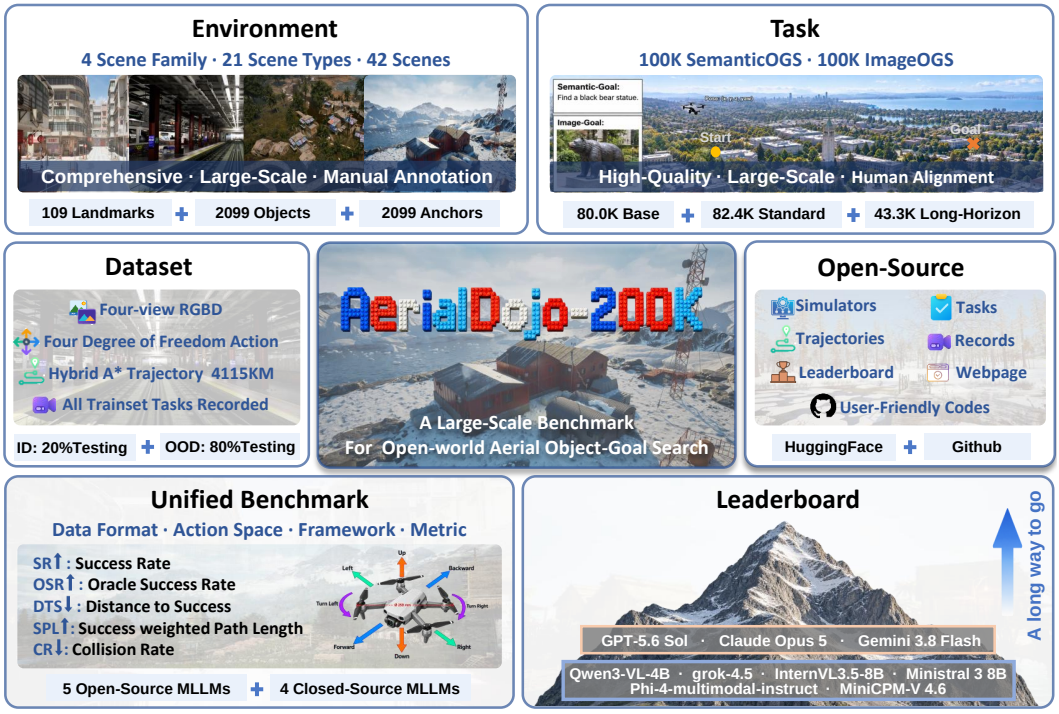


Figure 1: The overview of AerialDojo-200K.

et al., 2026) are currently the only AerialOGS benchmarks, and the scale of those is approximately one-tenth as many tasks as OpenFly (Gao et al., 2026); 2) Heterogeneous action spaces and data formats across AerialOGS and VLN benchmarks make it difficult for existing VLN benchmarks to support AerialOGS tasks. Consequently, existing research relies on small, environment-specific benchmarks with heterogeneous action spaces and data formats, failing to support large-scale training and cross-benchmark evaluation, and limiting the scalability and generalizability of aerial agents.

To address this problem, we propose AerialDojo-200K, a large-scale benchmark suite for open-world aerial object-goal search, with $3\times$ (times) as many scenes and $18.7\times$ as many task instances as the largest existing benchmark for this task. Specifically, as shown in Figure 1, we construct 42 simulation scenes spanning four scene families and 21 scene types, including 18 urban, 12 natural, six infrastructure, and six disaster scenes. To ensure data quality, 12 annotators spent two months manually annotating 109 landmarks, 2099 target objects, and 2099 object anchors across these scenes. We further construct 205,732 task instances, comprising over 100K semantic-goal and over 100K image-goal instances across Base, Standard, and Long-Horizon settings. Each task instance includes a collision-free reference trajectory and corresponding multi-view video recordings, with the trajectories totaling 4115km across the dataset. We also develop a unified evaluation framework with a scene partition comprising 21 in-distribution scenes and 21 out-of-distribution scenes. Finally, our evaluation of five open-source and four closed-source multimodal large language models reveals that there is still a long way to go toward achieving general-purpose aerial agents. Our main contributions are summarized as follows:

- We propose AerialDojo-200K, a large-scale benchmark suite for open-world aerial object-goal search, with $3\times$ as many scenes and $18.7\times$ as many task instances as the largest existing benchmark for this task.
- AerialDojo-200K provides a high-quality dataset with extensive manual annotations. Twelve annotators spent two months manually annotating 109 landmarks, 2,099 objects, and 2,099 object anchors. Each task instance includes a collision-free reference trajectory and corresponding multi-view video recordings.
- We develop a unified evaluation framework for this task and evaluate 5 open-source and 4 closed-source MLLMs. The results reveal that there is still a long way to go toward achieving general-purpose aerial agents.

Table 1: Comparison of aerial benchmarks. N_{task}: the number of task instances; F_{scene}: Scene families; N_{scene}: the number of scenes. VLN: vision-and-language navigation; ObjNav: object-goal navigation; OGS: object-goal search. DoF: degrees of fredom. U: urban; N: natural; I: infrastructure; D: disaster.

Benchmark	Year	Task	Action	Goal Type	Goal Specification	N_{task}	F_{scene}	N_{scene}
AerialVLN	2023	VLN	4-DoF	Location	Movement instruction	25.3K	U	25
CityNav	2025	VLN	4-DoF	Location	Movement instruction	32.6K	U	34
OpenUAV	2025	VLN	6-DoF	Location	Movement instruction	12.1K	U,N	22
OpenFly	2026	VLN	4-DoF	Location	Movement instruction	103K	U	18
UAV-ON	2025	ObjNav	4-DoF	Instance	Semantic	11K	U,N	14
CityAVOS	2025	OGS	4-DoF	Instance	Semantic and Image	2.4K	U	6
AerialDojo-200K	**2026**	**OGS**	**4-DoF**	**Instance**	**Semantic or image**	**200K**	**U,N,I,D**	**42**

2 Related Works

We review related work on aerial object-goal search and aerial navigation benchmarks. We first discuss how existing methods combine spatial representations and semantic reasoning for autonomous search, then examine benchmark task formulations, dataset scale, and environmental coverage to contextualize AerialDojo-200K.

2.1 Aerial Object-Goal Search

Aerial object-goal search requires an agent to reach a target object specified by a semantic description or reference image based solely on egocentric visual observations. Existing research can be grouped into two categories. On the one hand, existing aerial agents utilize spatio-semantic representations, such as semantic maps (Zhou et al., 2026), scene graphs (Gai et al., 2026), and value maps (Zhang et al., 2026), to organize memories and plan over explored regions, frontiers, and object candidates. On the other hand, recent studies demonstrate that multimodal large models can leverage high-level semantic reasoning to interpret semantic goals (Ji et al., 2025), assess object candidates (Pardyl et al., 2025), and guide search decisions (Wang et al., 2024; Chen et al., 2026). AerialOGS has broad potential applications (Kim et al., 2025) in search and rescue, infrastructure inspection, and autonomous delivery, where operators may specify the object without knowing its exact location or providing a complete flight route. However, limited benchmark scale and heterogeneous evaluation protocols hinder systematic assessment of these methods across diverse environments. AerialDojo-200K addresses these limitations through large-scale, carefully annotated search tasks and a unified framework for evaluating performance and generalization.

2.2 Aerial Benchmark

Aerial tasks primarily include vision-and-language navigation (VLN) and object-goal navigation (ObjNav), with aerial object-goal search (OGS) representing a more challenging setting within ObjNav that emphasizes autonomous exploration and target identification. As shown in Table 1, most existing aerial benchmarks focus on VLN tasks, where agents follow language instructions to reach specified locations. AerialVLN (Liu et al., 2023) and CityNav (Lee et al., 2025) provide 25.3K and 32.6K instruction-guided navigation tasks across 25 and 34 urban scenes, respectively. OpenUAV (Wang et al., 2025) extends this setting to a 6-DoF action space, providing 12.1K tasks across 22 urban and natural scenes, while OpenFly (Gao et al., 2026) expands the task count to 103K across 18 scenes. Beyond instruction-guided navigation, UAV-ON (Xiao et al., 2025a) supports instance-level ObjNav tasks through semantic goal specifications, with 11K tasks across 14 scenes. CityAVOS (Ji et al., 2026) further investigates OGS tasks using semantic and visual goal information, providing 2.4K tasks across six urban scenes. Despite these advances, existing benchmarks remain limited in both task scale and environmental coverage. Existing research relies on small, environment-specific benchmarks with heterogeneous action spaces and data formats, hindering large-scale training and cross-benchmark evaluation and limiting the scalability and generalizability of aerial agents.

3 AERIALDOJO-200K

AerialDojo-200K, the first large-scale benchmark suite for open-world Aerial Object-Goal Search (AerialOGS), comprises 42 simulated scenes, 205K task instances, 4115 km of flight trajectories, and a unified evaluation framework. The construction process is shown in Figure 2.

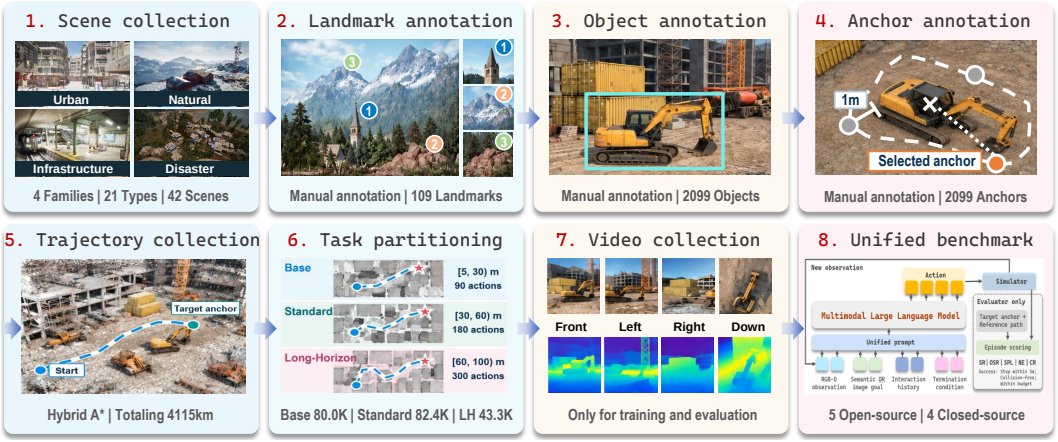


Figure 2: Standard operating procedure for constructing AerialDojo-200K.

3.1 SIMULATOR

We build the AerialDojo-200K benchmark using Unreal Engine (Epic Games, 2025) and ProjectAirSim (Microsoft, 2022), enabling AerialOGS across diverse, high-fidelity simulated scenes.

Large-scale and comprehensive environment collection. AerialDojo-200K contains 42 scenes spanning four scene families and 21 scene types, with two instances per type. As shown in Table 1, it provides three times as many scenes as UAV-ON (Xiao et al., 2025a), the largest existing AerialOGS benchmark. As shown in Figure 3, the collection includes both open landscapes and constrained built spaces, with diverse layouts and obstacle configurations. Specifically, the urban family comprises malls, stadiums, amusement parks, parking lots, communities, neighborhoods, alleys, factories, and construction sites; the natural family covers deserts, forests, mountains, snowfields, islands, and coasts; the infrastructure family includes bridges, rail corridors, and harbors; and the disaster family depicts earthquake, flood, and explosion scenes. By incorporating infrastructure and disaster scenarios alongside urban and natural environments, AerialDojo-200K covers a broad range of settings relevant to real-world aerial deployment and applications. Together, these scenes form a comprehensive library of simulation environments for AerialOGS.

Sensors. AerialDojo-200K can collect synchronized RGB-D streams from four onboard cameras facing forward, left, right, and downward. Both RGB and depth images have a resolution of 640×640 pixels, and each camera has a 90° field of view (FOV). All four views are sampled synchronously at 15 frames per second (FPS). At acquisition time t, the visual observation is represented as $o_t = \{(I_t^v, D_t^v) \mid v \in \mathcal{V}\}$ and $\mathcal{V} = \{\text{front}, \text{left}, \text{right}, \text{down}\}$, where I_t^v and D_t^v denote the RGB image and depth map from view v, respectively. The observation interface excludes GPS measurements, external localization signals, and global maps. Simulator ground truth, including object poses, semantic maps, and scene geometry, is also withheld from the agent. Consequently, the agent must interpret the goal, infer the surrounding spatial structure, and identify the target by integrating egocentric visual evidence over time.

Actions. AerialDojo-200K provides eight action types with discrete motion magnitudes, supporting four-DoF control of the aerial agent's position and yaw. Specifically, `MoveForward`, `MoveLeft`, and `MoveRight` allow horizontal translations of 1, 3, or 5 m, while `MoveUp` and `MoveDown` allow vertical translations of 1 or 2 m. The rotational actions, `TurnLeft` and `TurnRight`, change the yaw angle by 15° in the corresponding direction. The `Stop` action declares target discovery and terminates the episode. Together, these eight action types yield 16 discrete action choices, allowing the aerial agent to select movement magnitudes according to local geometry and target proximity.

Table 2: Dataset statistics. LM: landmarks; Base, Standard, and LH denote task-instance counts for the Base, Standard, and Long-Horizon setting, respectively, including both semantic-goal and image-goal instances; ALL is their sum. Path reports the cumulative planned path length in kilometers, counted once per underlying navigation task.

Family	Types	Scenes	LM	Objects	Anchors	Base	Standard	LH	ALL	Path (km)
Urban	9	18	57	1,030	1,030	42,232	55,840	26,992	125,064	2,607.253
Natural	6	12	21	569	569	21,668	15,188	12,976	49,832	1,008.083
Infrastructure	3	6	20	280	280	8,934	7,800	3,080	19,814	357.067
Disaster	3	6	11	220	220	7,140	3,596	286	11,022	142.910
Total	**21**	**42**	**109**	**2,099**	**2,099**	**79,974**	**82,424**	**43,334**	**205,732**	**4,115.313**



(a) Simulator statistics (b) Object-name word cloud (c) Task statistics

Figure 3: Simulator and dataset statistics of AerialDojo-200K.

Backward translation is excluded to encourage forward-looking exploration and reduce redundant oscillatory movements. The aerial agent can reposition by combining yaw rotations with forward or lateral translations, enabling flexible exploration while maintaining a compact action interface.

Aerial embodiment. The aerial agent has a default wheelbase of $250\,\mathrm{mm}$ and a collision radius of $250\,\mathrm{mm}$. These settings are intended to represent compact commercial aerial platforms with wheelbases below $500\,\mathrm{mm}$. Both parameters are configurable.

Manual annotation. To ensure simulator quality, as shown in Table 2, 12 annotators spent two months manually annotating 109 landmarks, 2099 target objects, and 2099 object anchors across these scenes. AerialDojo-200K selects landmarks and target objects according to their roles in AerialOGS tasks. Landmarks serve as visually salient, readily identifiable spatial references, reflecting a human search strategy in unfamiliar environments: first locating a relevant landmark and then refining the target's position. Landmark selection is inherently scene-dependent; shopping malls or historic monuments may serve as urban landmarks, whereas distinctive mountain peaks or rock formations may provide natural landmarks. This variation motivates manual selection based on scene-level salience and contextual relevance. Target objects are selected for their practical search value and semantic distinguishability. Their descriptions must allow human annotators to unambiguously identify the intended instance, reducing ambiguous supervision during training. AerialDojo-200K uses object anchors for task evaluation because a centroid-based success region may lie entirely inside a large object, making successful arrival physically infeasible. An anchor is a collision-free viewpoint located $1\,\mathrm{m}$ outside the object's boundary and oriented toward its centroid. A vision-language model shortlists three candidate anchors based on visual informativeness and image–name correspondence. Human annotators then select one anchor per object as the spatial reference for task construction and success evaluation. We also have developed and open-sourced a lightweight annotation tool that standardizes the entire manual annotation workflow.

3.2 Task and Dataset

Task definition. AerialDojo-200K defines open-world AerialOGS tasks as $\mathcal{T} = \langle S, G \rangle$, where S is the start pose and G is the instance-level goal. $S = [x, y, z, \psi]$, where $(x, y, z) \in \mathbb{R}^3$ denotes its position and ψ denotes its yaw angle. Unlike UAV-ON (Xiao et al., 2025a), which randomly initializes the agent's start pose, we initialize each task instance at the anchor position of a selected object, with the yaw rotated by 180° relative to the anchor orientation. This design can substantially reduce the risk of task failure due to collisions at initialization while better reflecting practical aerial

search scenarios, in which an agent begins a new search from the location of a previously reached target. The goal is specified as $G = [\text{name}, \text{landmark}, \text{direction}, \text{representation}]$, comprising the target object's name, the name of its nearest landmark, its direction relative to that landmark, and a target representation provided as either a semantic description or a reference image. The direction takes one of eight compass orientations: north, south, east, west, northeast, northwest, southeast, and southwest. This goal reflects a common human search strategy in unfamiliar environments: using a recognizable landmark and a coarse direction to orient the search, then relying on visual evidence to identify the target object. AerialDojo-200K defines the task as **SemanticOGS** when the representation is a semantic description and **ImageOGS** when it is a reference image.

Task execution. An episode is specified by

$$\mathcal{X} = \langle \mathcal{T}, \mathcal{E}, \mathcal{B}, \mathcal{O}, \mathcal{A}, \mathcal{C} \rangle, \tag{1}$$

where $\mathcal{E}$ is the scene, $\mathcal{B}$ is the aerial embodiment, $\mathcal{O}$ and $\mathcal{A}$ are the observation and action spaces described above, and $\mathcal{C}$ specifies the termination conditions. At each decision step, the agent selects $a_t \in \mathcal{A}$ based on G and its observation-action history. An episode terminates upon `Stop`, collision, or exhaustion of its action budget. Success requires an explicit `Stop` within 3 m of the target anchor, without any collision and within the assigned budget.

Collection trajectory. Candidate tasks pair distinct annotated objects within the same scene. We adapt Hybrid A* (Dolgov et al., 2008) to the aerial agent's four-DoF action space to plan a collision-free reference trajectory from the start pose to the target anchor. Collision checking accounts for the agent's configured collision radius, and candidate pairs are discarded if no feasible trajectory is found. Let $\tau^{\text{ref}} = (q_0, \ldots, q_K)$, where $q_k = (\mathbf{p}_k, \psi_k)$, $q_0 = S$, $\mathbf{p}_0$ is the start position, and $\mathbf{p}_K = \mathbf{p}_g$ is the target anchor position. The translational trajectory length is

$$L(\tau^{\text{ref}}) = \sum_{k=0}^{K-1} \|\mathbf{p}_{k+1} - \mathbf{p}_k\|_2. \tag{2}$$

Pure yaw rotations contribute no translational length. Reference trajectories reach the target anchor itself, whereas evaluation accepts an explicit `Stop` within 3 m of the anchor. Target coordinates and reference trajectories are withheld from the agent during evaluation.

Task partitioning. We categorize tasks by reference trajectory length and use a detour ratio to filter out near-straight-line routes. Let $\mathbf{p}_0$ and $\mathbf{p}_g$ denote the start and target anchor positions, respectively. The Euclidean separation and detour ratio are defined as

$$d_E = \|\mathbf{p}_0 - \mathbf{p}_g\|_2, \qquad \kappa = \frac{L(\tau^{\text{ref}})}{d_E}. \tag{3}$$

We retain candidates satisfying $d_E > 3\,\text{m}$, $5\,\text{m} \leq L(\tau^{\text{ref}}) < 100\,\text{m}$, and $\kappa \geq 1.05$. The first condition excludes starts already within the success region. The trajectory-length range balances meaningful search difficulty with a manageable evaluation horizon. The detour-ratio threshold filters out near-straight-line routes, aiming to reduce trivial tasks on which models may succeed uniformly and thereby improve the ability of success rate to distinguish model performance. Retained tasks are partitioned into Base, Standard, and Long-Horizon settings according to their reference trajectory lengths, with corresponding action budgets listed in Table 3.

Table 3: Task settings.

Setting	$L(\tau^{\text{ref}})$	Action budget
Base	$[5, 30)$	90
Standard	$[30, 60)$	180
Long-Horizon	$[60, 100)$	300

Collection video. We acquire synchronized multi-view RGB-D observations along reference trajectories using the sensor configuration described above. Quality checks flag missing camera outputs and abnormal image frames. Recordings are indexed by navigation task, camera view, and reference trajectory steps. The goal reference image used by ImageOGS is stored separately.

Dataset statistics. As shown in Figure 3, AerialDojo-200K contains 205,732 task instances, comprising 102,866 SemanticOGS and 102,866 ImageOGS instances across 42 scenes. The two instances of each navigation task share the same scene, start pose, target object, and reference trajectory, differing only in goal representation. The dataset covers 2,099 target objects across 254 semantic categories and reports a total planning distance of $4115.313\,\text{km}$.

3.3 BENCHMARK

Benchmark protocol. To evaluate AerialOGS capabilities in open-world environments, AerialDojo-200K divides its 42 scenes into 21 in-distribution (ID) scenes and 21 out-of-distribution (OOD) scenes, with one instance of each scene type in each split. Within ID scenes, 80% of the tasks are used for training and the remaining 20% for in-distribution performance evaluation. Within OOD scenes, 20% of the tasks form an adaptation subset, while the remaining 80% are reserved for testing. Models trained on ID scenes are transferred to OOD scenes to evaluate generalization across distinct map layouts, obstacle configurations, and target instances within the same scene types. Inspired by adaptation with limited target-domain data, we allow models to use the designated 20% OOD subset for scene-specific adaptation before evaluation on the held-out OOD test tasks. Evaluation covers SemanticOGS and ImageOGS across the Base, Standard, and Long-Horizon settings, using the action budgets in Table 3.

Benchmark framework. Figure 4 illustrates the closed-loop evaluation framework. Each episode resets the simulator to the task's scene and start pose. At each step, a model-specific adapter packages the current RGB-D observations, goal specification, termination conditions, and interaction history using the fixed prompt provided in Appendix A. The returns one action, which maps to the simulator's action interface using a common motion-magnitude configuration. Execution produces the next observation and updates the episode log. Episodes terminate upon `Stop`, collision, or budget exhaustion; infrastructure failures are logged and rerun separately. Target coordinates and reference trajectories remain accessible only to the evaluator.

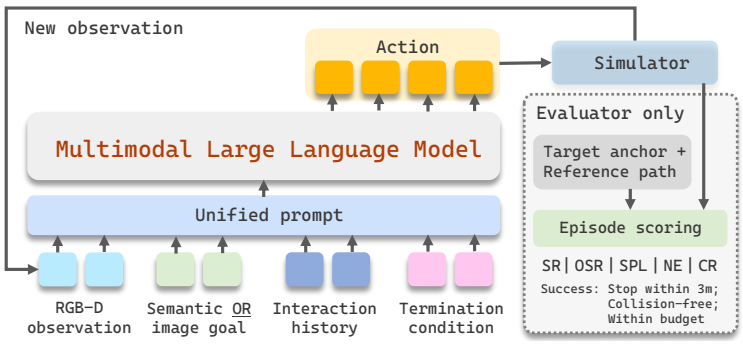


Figure 4: The unified benchmark framework of AerialDojo-200K.

Benchmark metrics. We evaluate navigation performance using five complementary metrics. Success rate (SR) measures the fraction of episodes in which the agent explicitly stops within 3 m of the target anchor without collision and within the action budget. Oracle success rate (OSR) measures whether the agent ever enters this region, regardless of its final stopping decision or subsequent failure. Distance to success (DTS) is the mean final Euclidean distance to the target anchor, measured in meters without subtracting the success radius. Success weighted by path length (SPL) evaluates navigation efficiency:

$$\mathrm{SPL} = \frac{1}{N}\sum_{i=1}^{N}\sigma_i \frac{L_i}{\max(L_i, \ell_i)}, \tag{4}$$

where N is the number of evaluated episodes, σ_i indicates episode success, $L_i = L(\tau_i^{\mathrm{ref}})$ is the reference trajectory length and ℓ_i is the executed trajectory length. This formulation uses the planned reference for normalization. Collision rate (CR) is the fraction of episodes containing a collision. All metrics are computed over both successful and failed episodes.

4 EXPERIMENTS

We conduct experiments on AerialDojo-200K to evaluate existing MLLMs on AerialOGS. We first introduce the baseline models and experimental setup, followed by quantitative comparisons across scene families and task setting. We further analyze how the success threshold affects model performance and rankings.

4.1 Baselines and Setup

Baselines. We evaluate five open-source MLLMs, Qwen3-VL-4B, InternVL3.5-8B, Ministral 3 8B, Phi-4-multimodal, and MiniCPM-V 4.6, and four closed-source MLLMs, GPT-5.6 Sol, Grok-4.5, Gemini 3.8 Flash, and Claude Opus 5. We test these models under the Base, Standard, and Long-Horizon settings. For brevity, the main tables report model-setting combinations with nonzero overall success rates, covering all nine MLLMs on Base tasks, GPT-5.6 Sol and Claude Opus 5 on Standard tasks.

Evaluation setup. Base and Standard tasks have reference lengths of $[5, 30)$ m and $[30, 60)$ m and action budgets of 90 and 180, respectively. We average the all map-level scores equally within each family, then average the four family means equally for overall results. Rates are percentages, and DTS is in meters. The prompt specifies stopping within 3 m, our primary success radius; 5 m is a supplementary evaluation threshold.

Table 4: Base-task performance cross four scene families. SR, OSR, SPL, and CR are percentages; DTS is in meters. Arrows indicate the preferred direction; bold denotes the best available value.

Model	3 m (primary)					5 m	
	SR↑	OSR↑	SPL↑	DTS↓	CR↓	SR↑	OSR↑
Open-source models							
Qwen3-VL-4B	0.71±0.81	1.50±1.33	0.51±0.60	16.24±1.85	**22.17**±14.32	3.85±2.92	7.99±3.93
InternVL3.5-8B	0.16±0.30	3.41±2.36	0.11±0.22	25.42±11.63	62.06±16.89	1.54±1.74	12.04±4.47
Ministral 3 8B	0.17±0.31	1.28±1.48	0.09±0.21	16.23±3.43	51.63±24.41	1.64±1.18	11.37±5.20
Phi-4-multimodal	0.13±0.35	0.60±1.01	0.09±0.25	23.52±7.47	37.96±16.67	1.65±1.79	9.22±6.17
MiniCPM-V 4.6	0.19±0.53	1.35±2.41	0.15±0.43	20.26±6.44	42.47±11.07	1.47±1.98	9.91±3.84
Closed-source models							
GPT-5.6 Sol	2.81±2.42	3.06±2.44	2.20±1.87	17.90±4.01	32.91±14.45	8.16±3.78	11.22±4.23
Grok-4.5	0.25±0.69	0.49±1.39	0.25±0.69	**15.46**±1.81	22.18±12.44	2.46±2.28	3.69±2.65
Gemini 3.8 Flash	2.40±2.23	3.68±4.32	2.37±2.21	21.61±7.07	33.95±12.21	**11.16**±8.27	17.90±8.25
Claude Opus 5	**3.20**±2.04	**7.84**±4.60	**2.83**±2.04	19.69±5.97	58.96±15.94	7.38±3.73	**18.48**±5.64

Table 5: Standard-task performance across four scene families. SR, OSR, SPL, and CR are percentages; DTS is in meters. Arrows is the preferred direction; bold denotes the best available value.

Model	3 m (primary)					5 m	
	SR↑	OSR↑	SPL↑	DTS↓	CR↓	SR↑	OSR↑
GPT-5.6 Sol	0.26±0.72	0.26±0.72	0.20±0.57	37.66±5.47	**36.57**±16.79	**2.30**±2.02	2.30±2.02
Claude Opus 5	**0.49**±0.91	**1.32**±2.33	**0.40**±0.73	**36.49**±8.89	56.30±21.31	1.29±1.91	**2.61**±3.84

4.2 Evaluation Results

Performance on base and standard tasks. Tables 4–5 show substantial performance differences among the evaluated models, alongside consistently low task completion rates. At the primary 3 m threshold, Claude Opus 5 achieves the highest Base SR of 3.20%, followed by GPT-5.6 Sol (2.81%) and Gemini 3.8 Flash (2.40%). Qwen3-VL-4B leads the open-source models at 0.71%, exceeding Grok-4.5 at 0.25%, indicating that the advantage of closed-source models is not uniform. Performance deteriorates markedly on Standard tasks: GPT-5.6 Sol and Claude Opus 5 achieve SRs of only 0.26% and 0.49%, respectively. Their OSRs also decrease to 0.26% and 1.32%, showing that even reaching the target region is uncommon. Together, these results highlight the difficulty of maintaining successful search beyond the Base setting.

Success-radius sensitivity. Relaxing the success radius from 3 m to 5 m substantially changes both success rates and model rankings. On Base tasks, Gemini 3.8 Flash increases from 2.40% to 11.16% SR, an improvement of 8.76 percentage points, overtaking GPT-5.6 Sol (8.16%) and Claude Opus 5 (7.38%). A ranking reversal also occurs on Standard tasks: Claude Opus 5 leads at 3 m, whereas

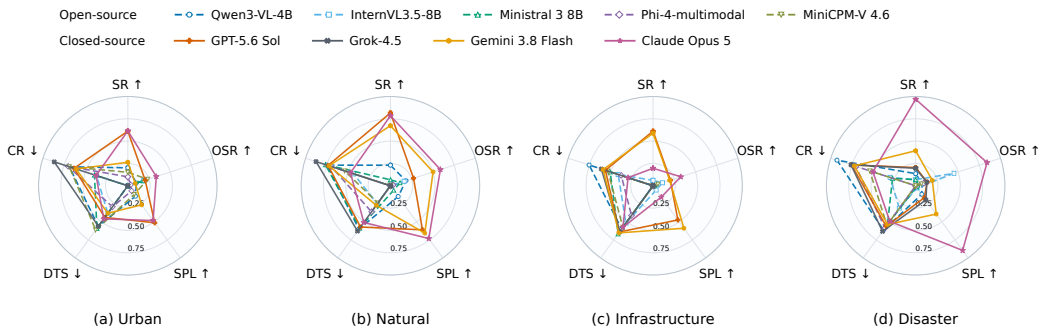


Figure 5: Base-task performance across four scene families at the 3 m success threshold. Each curve represents one model. From the center to the outer edge, SR and SPL range from 0 to 5%, OSR from 0 to 15%, DTS from 40 to 0 m, and CR from 100 to 0%; outward therefore indicates better performance. Ring labels show normalized radial values.

GPT-5.6 Sol leads at 5 m with 2.30% SR versus 1.29%. These changes demonstrate that model comparisons depend strongly on the spatial tolerance used to define success. The higher scores at 5 m reflect a more permissive evaluation criterion and do not establish improved navigation behavior. We therefore report both thresholds while retaining the prompt-aligned 3 m criterion as the primary measure of precise task completion.

Performance differences across scene families. Figure 5 reveals scene-dependent model strengths that are obscured by overall averages. At 3 m, GPT-5.6 Sol achieves the highest SR on Natural and Infrastructure scenes, at 4.08% and 3.06%, respectively. Claude Opus 5 leads on Disaster scenes with 4.85% SR, but achieves only 0.98% on Infrastructure scenes. On Urban scenes, Claude Opus 5 and GPT-5.6 Sol obtain nearly identical SRs of 3.07% and 3.06%. The radar profiles further show that strengths in successful completion do not consistently coincide with strengths in collision avoidance or final target distance. Thus, no evaluated model dominates all scene families and metrics, emphasizing the need for family-wise evaluation to expose differences in capability across the evaluated environments.

A long way to generalist aerial agents. Collectively, these results reveal a substantial gap between current MLLMs under the evaluated framework and reliable generalist aerial agents. The best reported SR remains only 3.20% on Base tasks and 0.49% on Standard tasks at the primary threshold. Moreover, reaching the target region does not ensure successful completion: Claude Opus 5 achieves a Base OSR of 7.84% but an SR of only 3.20%, alongside a collision rate of 58.96%. This gap captures failures to convert target-region visits into valid completion, although the aggregate metrics cannot identify their individual causes. Low completion rates, frequent collisions, and uneven performance across scene families collectively indicate that integrating effective exploration, precise target localization, appropriate stopping, and collision avoidance remains a major challenge on the path toward generalist aerial agents.

5 Conclusion

We present AerialDojo-200K, a large-scale benchmark suite for open-world aerial object-goal search. It comprises 205,732 semantic-goal and image-goal task instances across 42 simulated scenes from four scene families, with three task settings and a unified observation, action, and evaluation interface. Our evaluation of existing MLLMs reveals substantial limitations: the best reported success rates at the primary 3 m threshold reach only 3.20% on Base tasks and 0.49% on Standard tasks. Together with frequent collisions and gaps between target-region visits and successful completion, these findings show that current MLLMs under the evaluated framework remain far from reliable generalist aerial agents. AerialDojo-200K provides a common testbed for developing and evaluating methods that integrate exploration, instance-level target grounding, precise stopping, and collision avoidance.

AI-USE STATEMENT

We used generative AI tools to assist with translation, literature search and reference checking. During benchmark construction, a vision-language model also assists in shortlisting candidate object anchors, followed by human selection. The authors take full responsibility for the accuracy, originality, and integrity of all text, code, figures, and reported results.

REPRODUCIBILITY STATEMENT

As shown in Figure 1, we publicly release all simulators, tasks, trajectories, video recordings, and source code, together with a leaderboard and project website, to facilitate community access, reuse, and reproducibility.

ETHICS STATEMENT

AerialDojo-200K supports research on aerial object-goal search through simulated scenes. Performance in simulation does not establish the safety or reliability of deployment on physical aerial platforms, which requires additional validation under real-world conditions.

REFERENCES

Xuecheng Chen, Zongzhuo Liu, Jianfa Ma, Bang Du, Tiantian Zhang, Xueqian Wang, and Boyu Zhou. AirHunt: Bridging vlm semantics and continuous planning for efficient aerial object navigation, 2026.

Dmitri Dolgov, Sebastian Thrun, Michael Montemerlo, and James Diebel. Practical search techniques in path planning for autonomous driving. In *Proceedings of the First International Symposium on Search Techniques in Artificial Intelligence and Robotics (STAIR-08)*. AAAI, 2008.

Epic Games. Unreal Engine, 2025. URL `https://www.unrealengine.com/`.

Weiqi Gai, Yuman Gao, Yuan Zhou, Yufan Xie, Zhiyang Liu, Yuze Wu, Xin Zhou, Fei Gao, and Zhijun Meng. USS-Nav: Unified spatio-semantic scene graph for lightweight uav zero-shot object navigation, 2026.

Yunpeng Gao, Chenhui Li, Zhongrui You, Junli Liu, Zhen Li, Pengan Chen, Qizhi Chen, Zhonghan Tang, Liansheng Wang, Penghui Yang, Yiwen Tang, Yuhang Tang, Shuai Liang, Songyi Zhu, Ziqin Xiong, Yifei Su, Xinyi Ye, Jianan Li, Yan Ding, Dong Wang, Xuelong Li, Zhigang Wang, and Bin Zhao. OpenFly: A comprehensive platform for aerial vision-language navigation. In *The Fourteenth International Conference on Learning Representations*, 2026.

Yatai Ji, Zhengqiu Zhu, Yong Zhao, Beidan Liu, Chen Gao, Yihao Zhao, Sihang Qiu, Yue Hu, Quanjun Yin, and Yong Li. Towards autonomous UAV visual object search in city space: Benchmark and agentic methodology. arXiv preprint arXiv:2505.08765, 2025.

Yatai Ji, Zhengqiu Zhu, Yong Zhao, Beidan Liu, Chen Gao, Yihao Zhao, Sihang Qiu, Yue Hu, and Quanjun Yin. Towards autonomous UAV visual object search in city space: Benchmark and agentic methodology. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 18342–18350, 2026. doi: 10.1609/aaai.v40i22.38898.

Seungchan Kim, Omar Alama, Dmytro Kurdydyk, John Keller, Nikhil Keetha, Wenshan Wang, Yonatan Bisk, and Sebastian Scherer. RAVEN: Resilient aerial navigation via open-set semantic memory and behavior adaptation, 2025.

Jungdae Lee, Taiki Miyanishi, Shuhei Kurita, Koya Sakamoto, Daichi Azuma, Yutaka Matsuo, and Nakamasa Inoue. CityNav: A large-scale dataset for real-world aerial navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5912–5922, 2025.

Shubo Liu, Hongsheng Zhang, Yuankai Qi, Peng Wang, Yanning Zhang, and Qi Wu. AerialVLN: Vision-and-language navigation for UAVs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15384–15394, 2023.

Microsoft. Project AirSim, 2022. URL `https://github.com/iamaisim/ProjectAirSim`.

Adam Pardyl, Dominik Matuszek, Mateusz Przebieracz, Marek Cygan, Bartosz Zieliński, and Maciej Wolczyk. Flysearch: Exploring how vision-language models explore. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-5594.

Xiangyu Wang, Donglin Yang, Ziqin Wang, Hohin Kwan, Jinyu Chen, Wenjun Wu, Hongsheng Li, Yue Liao, and Si Liu. Towards realistic uav vision-language navigation: Platform, benchmark, and methodology, 2024.

Xiangyu Wang, Donglin Yang, Ziqin Wang, Hohin Kwan, Jinyu Chen, Wenjun Wu, Hongsheng Li, Yue Liao, and Si Liu. Towards realistic UAV vision-language navigation: Platform, benchmark, and methodology. In *International Conference on Learning Representations*, pp. 7292–7310, 2025.

Jianqiang Xiao, Yuexuan Sun, Yixin Shao, Boxi Gan, Rongqiang Liu, Yanjin Wu, Weili Guan, and Xiang Deng. UAV-ON: A benchmark for open-world object goal navigation with aerial agents. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 13023–13029, 2025a. doi: 10.1145/3746027.3758251.

Jianqiang Xiao, Yuexuan Sun, Yixin Shao, Boxi Gan, Rongqiang Liu, Yanjin Wu, Weili Guan, and Xiang Deng. UAV-ON: A benchmark for open-world object goal navigation with aerial agents, 2025b.

Daoxuan Zhang, Ping Chen, Xiaobo Xia, Xiu Su, Ruichen Zhen, Jianqiang Xiao, and Shuo Yang. APEX: A decoupled memory-based explorer for asynchronous aerial object goal navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15232–15242, 2026.

Baining Zhao, Jiacheng Xu, Weicheng Feng, Xin Zhang, Zhaolu Wang, Haoyang Wang, Shilong Ji, Ziyou Wang, Jianjie Fang, Zhiheng Zheng, Weichen Zhang, Yu Shang, Wei Wu, Chen Gao, Xinlei Chen, and Yong Li. WorldVLN: Autoregressive world action model for aerial vision-language navigation, 2026.

Jiacong Zhou, Jiaxu Miao, Yourun Lin, Xianyun Wang, Jun Xiao, and Jun Yu. Memory-augmented scene understanding and exploration for open-world aerial object-goal navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21616–21626, 2026.

A PROMPT

Figure 6 presents the evaluation prompt for SemanticOGS in AerialDojo-200K. The prompt specifies the front-facing RGB-D inputs, depth encoding, eight available actions, aerial embodiment configurations, termination Conditions, and the required single-action JSON response. It defines successful completion as an explicit `Stop` within 3 m of the target anchor, without collisions and within the assigned action budget.

Prompt for AerialDojo-200K

Task Description
You control a aerial agent using its current front-facing RGB image and depth image.
Choose exactly one action that best moves toward the described target while avoiding visible obstacles.

Visual Observations
The first image is the front-facing RGB image. The second image is the aligned front-facing depth image.

Depth Encoding
Each depth pixel is an unsigned 8-bit value from 0 to 255, with
distance_m = pixel_value / 255 * 100;
0 is the nearest encoded distance and 255 is 100 metres or farther.

Action Space
The action must be one of:
moveforward, moveleft, moveright, moveup,
movedown, turnleft, turnright, stop.

Output Requirements
Return exactly one valid JSON object in the form
{"action":"moveforward"}.
Do not include Markdown, explanations, or any other keys.

Success Criterion
Success requires an explicit Stop within 3m of the target anchor, without any collision and within the assigned budget.

Figure 6: Prompt for MLLM evaluation in AerialDojo-200K.

B VISUALIZATION OF ALL SIMULATION SCENES

Figure 7 presents all 42 simulation scenes in AerialDojo, covering 21 scene types across four environment families: natural, urban, infrastructure, and disaster. These scenes exhibit diverse terrain, spatial layouts, visual appearances, and obstacle configurations, providing varied environments for aerial object-goal search. Each scene type includes two distinct instances, designated as in-distribution (ID) and out-of-distribution (OOD), respectively. This paired organization supports the evaluation of agents in familiar environments and their generalization to unseen scenes within the same scene type.

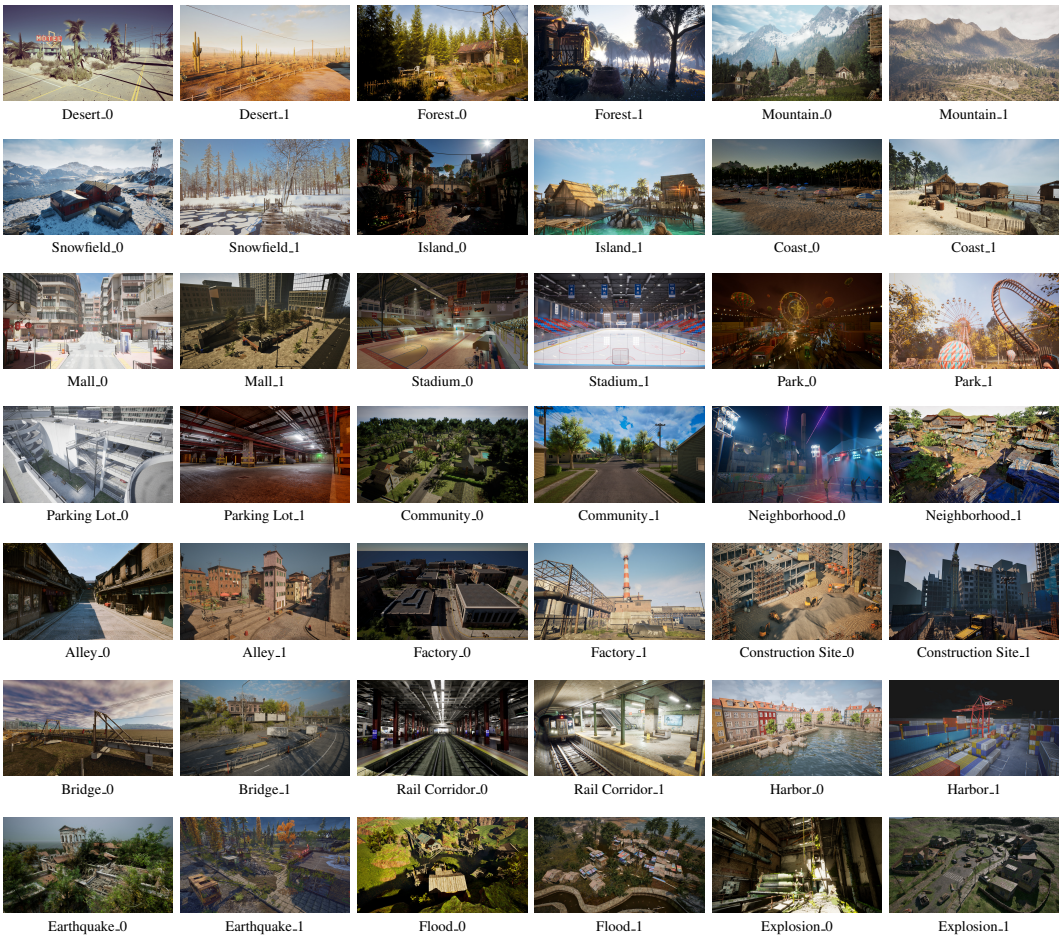


Figure 7: Visual overview of all 42 simulation scenes in AerialDojo, arranged in six columns and seven rows. Rows 1–2 show natural scenes, rows 3–5 show urban scenes, row 6 shows infrastructure scenes, and row 7 shows disaster scenes. The two instances of each scene type are placed side by side. The suffixes `_0` and `_1` denote in-distribution (ID) and out-of-distribution (OOD) scenes, respectively.

C Task Composition and Reference Lengths

AerialDojo-200K contains 79,974 Base, 82,424 Standard, and 43,334 Long-Horizon task instances, representing 38.87%, 40.06%, and 21.06% of the release, respectively. Figure 8 separates task volume, Train/Test allocation, and reference length so that a large task set is not confused with a task set containing longer routes.

Urban scenes contribute 60.79% of all instances, followed by Natural (24.22%), Infrastructure (9.63%), and Disaster (5.36%). Mall scenes alone contribute 28.82%. Scene-level counts range from 190 in `N_Mountain_1` to 40,214 in `U_Mall_0`. These unequal counts motivate reporting the aggregation rule explicitly when comparing methods; a task-weighted score and an equally weighted scene-family score represent different evaluation populations.

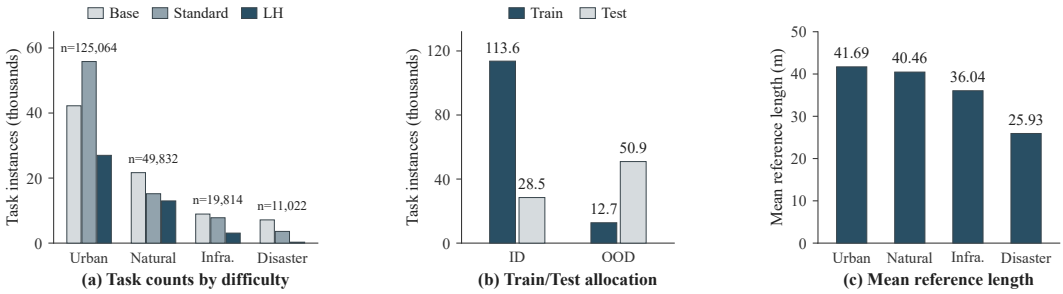


Figure 8: Task composition and reference lengths. (a) Goal-conditioned instance counts by family and setting. (b) Train/Test counts in the ID and OOD scene groups. Counts in (a,b) include both goals. (c) Mean reference length, weighted by navigation-task count within each family.

D Scene Partitions and Target Coverage

The scene partition assigns one instance of each type to ID and the other to OOD, giving 21 scenes per group. Equal scene counts do not imply equal task counts: ID scenes contain 142,094 instances (69.07%), whereas OOD scenes contain 63,638 (30.93%). Table 6 reports the actual integer allocation. ID scenes contain 113,642 training and 28,452 test instances; the corresponding OOD counts are 12,712 and 50,926. Their training proportions are 79.98% and 19.98%, respectively. These realized proportions are close to the nominal 80/20 and 20/80 allocations.

Table 6: Task-instance counts by scene group, local Train/Test subset, and difficulty. Both goal modalities are included. ID/OOD identifies the scene group; Train/Test identifies the task subset within that group.

Scenes	Task subset	Base	Standard	LH	ALL
ID	Train	40,640	48,398	24,604	113,642
ID	Test	10,240	12,060	6,152	28,452
OOD	Train	5,764	4,452	2,496	12,712
OOD	Test	23,330	17,514	10,082	50,926
Total	**All**	**79,974**	**82,424**	**43,334**	**205,732**

E PER-SCENE INVENTORY

Table 7 reports the annotation inventory, setting counts, and reference distances for all 42 scenes, with adjacent ID/OOD instances within each type. 5 scenes have no Long-Horizon instances.

Table 7: Per-scene inventory in family and scene-type order. LM: landmarks; Obj./Anc.: annotated objects/anchors; LH: Long-Horizon; ALL: Base + Standard + LH, counting both goal modalities. Path counts each navigation reference once. Scene suffixes 0 and 1 identify ID and OOD instances.

Scene	LM	Obj./Anc.	Base	Standard	LH	ALL	Path (km)
Urban							
Mall_0	6	123/123	12,412	19,652	8,150	40,214	839.245
Mall_1	5	102/102	5,390	7,144	6,536	19,070	459.569
Stadium_0	5	64/64	4,344	8,032	146	12,522	213.410
Stadium_1	1	41/41	1,234	2,010	1,156	4,400	97.323
Park_0	2	14/14	346	90	6	442	5.077
Park_1	2	46/46	1,452	1,616	600	3,668	70.006
ParkingLot_0	7	78/78	1,552	5,788	3,140	10,480	265.216
ParkingLot_1	1	26/26	676	198	0	874	8.540
Community_0	4	65/65	392	740	1,128	2,260	65.735
Community_1	3	31/31	300	384	104	788	14.582
Neighborhood_0	1	32/32	1,738	996	26	2,760	36.245
Neighborhood_1	2	63/63	4,486	680	32	5,198	51.460
Alley_0	3	27/27	1,188	714	100	2,002	26.514
Alley_1	2	67/67	2,506	1,586	306	4,398	65.324
Factory_0	7	80/80	1,052	1,642	1,048	3,742	86.301
Factory_1	2	50/50	512	324	124	960	16.148
ConstructionSite_0	3	59/59	1,738	3,462	3,640	8,840	232.868
ConstructionSite_1	1	62/62	914	782	750	2,446	53.690
Natural							
Desert_0	3	64/64	3,010	5,008	6,294	14,312	382.722
Desert_1	1	29/29	700	236	28	964	11.950
Forest_0	1	54/54	7,220	412	0	7,632	64.708
Forest_1	1	28/28	1,012	626	84	1,722	24.887
Mountain_0	2	53/53	648	356	400	1,404	26.958
Mountain_1	1	32/32	132	54	4	190	2.262
Snowfield_0	1	21/21	478	214	2	694	8.153
Snowfield_1	1	44/44	1,706	812	1,036	3,554	72.382
Island_0	5	79/79	2,942	3,362	1,746	8,050	161.095
Island_1	2	60/60	1,098	922	814	2,834	60.992
Coast_0	1	50/50	1,590	2,110	2,082	5,782	141.467
Coast_1	2	55/55	1,132	1,076	486	2,694	50.507
Infrastructure							
Bridge_0	3	28/28	382	334	208	924	17.401
Bridge_1	3	20/20	240	152	20	412	5.796
RailCorridor_0	3	50/50	2,326	2,986	666	5,978	110.070
RailCorridor_1	5	54/54	1,592	674	36	2,302	27.301
Harbor_0	4	110/110	4,336	3,466	1,974	9,776	184.783
Harbor_1	2	18/18	58	188	176	422	11.716
Disaster							
Earthquake_0	2	50/50	1,490	40	0	1,530	10.934
Earthquake_1	3	49/49	1,132	668	212	2,012	32.151
Flood_0	3	25/25	944	610	0	1,554	20.214
Flood_1	1	13/13	230	190	6	426	6.205
Explosion_0	1	19/19	752	444	0	1,196	15.626
Explosion_1	1	64/64	2,592	1,644	68	4,304	57.780
Total	109	2,099/2,099	79,974	82,424	43,334	205,732	4,115.313