AeroDuo: Aerial Duo for UAV-based Vision and Language Navigation

Ruipu Wu*
Beihang University
Beijing, China
bastien_wu@buaa.edu.cn

Yige Zhang*
Beihang University
Beijing, China
yige_zhang@buaa.edu.cn

Jinyu Chen*
Beihang University
Beijing, China
chenjinyu@buaa.edu.cn

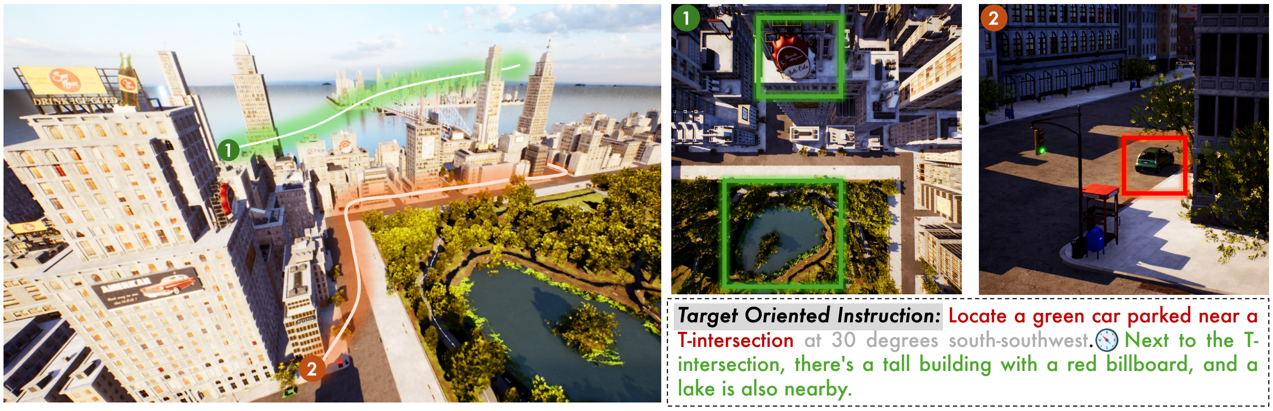
Linjiang Huang†
Beihang University
Beijing, China
ljhuang@buaa.edu.cn

Shifeng Zhang
Sangfor Technologies Inc.
Shenzhen, China
zhangshifeng@sangfor.com.cn

Xu Zhou
Sangfor Technologies Inc.
Shenzhen, China
zhouxu@sangfor.com.cn

Liang Wang
Institute of Automation, Chinese Academy of Sciences
Beijing, China
wangliang@nlpr.ia.ac.cn

Si Liu
Beihang University
Beijing, China
liusi@buaa.edu.cn



Target Oriented Instruction: Locate a green car parked near a T-intersection at 30 degrees south-southwest. Next to the T-intersection, there's a tall building with a red billboard, and a lake is also nearby.

Figure 1: In the DuAl-VLN task, the two UAVs operating at distinct altitudes achieve collaborative target search through language instruction guidance. The high-altitude UAV offers a broader observation range (❶), while the low-altitude UAV captures finer-grained visual perception of target (❷).

Abstract

Aerial Vision-and-Language Navigation (VLN) is an emerging task that enables Unmanned Aerial Vehicles (UAVs) to navigate outdoor environments using natural language instructions and visual cues. However, due to the extended trajectories and complex maneuverability of UAVs, achieving reliable UAV-VLN performance is challenging and often requires human intervention or overly detailed instructions.

To harness the advantages of UAVs' high mobility, which could provide multi-grained perspectives, while maintaining a manageable motion space for learning, we introduce a novel task called Dual-Altitude UAV Collaborative VLN (DuAl-VLN). In this task, two UAVs operate at distinct altitudes: a high-altitude UAV responsible for broad environmental reasoning, and a low-altitude UAV tasked with precise navigation. To support the training and evaluation of the DuAl-VLN, we construct the HaL-13k, a dataset comprising 13, 838 collaborative high-low UAV demonstration trajectories, each paired with target-oriented language instructions. This dataset includes both unseen maps and an unseen object validation set to systematically evaluate the model's generalization capabilities across novel environments and unfamiliar targets. To consolidate their

*Equal contribution.
†Corresponding author.

MM '25, Dublin, Ireland

ACM ISBN 979-8-4007-2035-2/2025/10
https://doi.org/10.1145/3746027.3754498

complementary strengths, we propose a dual-UAV collaborative VLN framework, AeroDuo, where the high-altitude UAV integrates a multimodal large language model (Pilot-LLM) for target reasoning, while the low-altitude UAV employs a lightweight multi-stage policy for navigation and target grounding. The two UAVs work collaboratively and only exchange minimal coordinate information to ensure efficiency. Experimental results indicate that AeroDuo achieves an evident 9.71% improvement in success rates compared to existing single-UAV methods, demonstrating the effectiveness of dual-altitude collaboration in balancing environmental coverage, precision, and operational autonomy.

CCS Concepts

• **Computing methodologies → Computer vision tasks**; *Planning and scheduling*.

Keywords

Vision and Language Navigation, Large Language Model, Multi-Agent Planning

ACM Reference Format:
Ruipu Wu, Yige Zhang, Jinyu Chen, Linjiang Huang, Shifeng Zhang, Xu Zhou, Liang Wang, and Si Liu. 2025. AeroDuo: Aerial Duo for UAV-based Vision and Language Navigation. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25), October 27–31, 2025, Dublin, Ireland.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3746027.3754498

1 Introduction

Vision-Language Navigation (VLN) [15], which aims to enable autonomous agents to navigate based on natural language instructions, has recently received significant research attention. Early efforts primarily focused on ground-based agents and have achieved remarkable progress. In contrast, VLN for Unmanned Aerial Vehicles (UAVs) remains relatively understudied. UAV-based VLN poses greater challenges due to extended navigation trajectories and higher degrees of motion freedom compared to ground-based scenarios. Existing approaches [30, 34, 54] for UAV-based VLN often rely on iterative human-agent dialogues [13], highly detailed route descriptions [19], or real-time human assistance [54], which inevitably increase human workload and restrict operational efficiency. Consequently, enabling UAVs to accomplish VLN tasks solely through relatively simple instructions, which only describe target orientation, regions, and surrounding features, could significantly enhance practicality. However, this remains challenging for a single UAV agent, as it struggles to simultaneously achieve high-altitude perspective for coarse-grained regional localization and low-altitude perspective for fine-grained target observation.

To tackle this challenge, we introduce a novel UAV-based VLN task called Dual-Altitude UAV Collaborative Vision-Language Navigation (DuAl-VLN). In this task, two UAVs operate at different altitudes: one at high altitude for wide-area environmental coverage and the other at low altitude for detailed, close-range observation. The two UAVs collaboratively optimize navigation through dynamic information exchange and joint path planning, leveraging their complementary capabilities: high-altitude contextual awareness enhances strategic decision-making, while low-altitude detailed sensing ensures real-time obstacle avoidance and safe navigation in cluttered environments. This dual-altitude framework significantly enhances navigation efficiency by balancing macro-scale environmental understanding with micro-scale flight safety.

To advance this task, we curate a dataset, HaL-13k, based on the OpenUAV platform [54] with 13, 838 synchronized dual-altitude trajectories across 14 scenarios. This dataset is generated using a two-stage trajectory creation process: first, low-altitude UAV trajectories are obtained via obstacle-aware path planning; subsequently, high-altitude UAV paths are collected under strict visibility constraints. The constraints ensure visual overlap with the low-altitude trajectories, thereby enabling effective autonomous exploration for high-altitude UAVs. HaL-13k offers navigation instructions that exclusively describes target orientation, visual features, and surrounding environmental context, along with paired high-low trajectories and multi-modal sensor streams. This dataset is tailored to investigate altitude-dependent perception dynamics and collaborative decision-making in the dual-UAV system.

To address the challenges of cooperative navigation, where inefficient information exchange can cause trajectory conflicts, we propose AeroDuo. This collaborative UAV-VLN framework synergizes multimodal large language models and lightweight models tailored for each UAV's role. **For the high-altitude UAV**, we introduce the Pilot-LLM, which leverages pre-trained MLLMs' capabilities to enable effective instruction understanding and target reasoning. Specifically, Pilot-LLM processes historical flight trajectories and constructs a global orthographic projection map to dynamically infer coarse-grained target regions. A mask prediction module is further integrated into the MLLM, prioritizing feasible areas for the low-altitude UAV's detailed exploration. **For the low-altitude UAV**, we deploy a navigation policy trained in the Isaac Sim [40] simulation environment, combining a lightweight obstacle avoidance controller and a visual grounding model to precisely localize target objects. Crucially, the two UAVs communicate only minimal coordinate information, significantly reducing bandwidth requirements while maintaining collaborative coherence. Experimental results demonstrate that our AeroDuo achieves a significant improvement of **9.71%** in navigation success rates compared to single-UAV baselines on the validation set of the HaL-13k dataset, which demonstrates the effectiveness of dual-altitude UAV collaboration for VLN tasks, opening new possibilities for aerial embodied AI systems.

2 Related Works

2.1 Ground-based Vision-Language Navigation

Ground-based VLN has seen rapid advances, with datasets [1, 6] and benchmarks [7, 25, 28, 29, 43] enabling broader task coverage through diverse instructions and heterogeneous environments. Related research has delved into data augmentation techniques [15, 53, 59, 61], decision-making mechanisms [62], the utilization of historical context [8, 10, 16, 17, 24, 27], and representations of three-dimensional space. Furthermore, the rapid advancement of LLMs [11, 51] and MLLMs [21, 31, 45, 65] has inspired action-prediction methods like [5]. Recent works [9, 32, 44] integrate LLMs into planning, while others [60, 64] propose unified models for language and environmental context. Compared to ground-based VLN, UAV-based VLN exhibits a longer trajectory length and higher degrees of motion freedom, presenting significantly greater challenges.

2.2 UAV Navigation

Current UAV navigation research has mainly focused on visual perception [4, 14, 20, 26, 36, 39, 50] and collision avoidance [49, 55], while multimodal visual-linguistic UAV navigation remains emerging. Recent efforts [34, 35] introduced UAV-VLN frameworks using detailed textual guidance. AerialVLN [34] provides a large-scale dataset with an effective baseline, and STMR [19] builds on it with a zero-shot LLM-based framework using a Semantic-TopoMetric Representation for spatial reasoning. Other works have contributed infrastructure datasets, including CityNav [30], AVDN [13], and OpenFly [18]. OpenUAV [54], in particular, offers a UAV dynamics simulator and the UAV-Need-Help evaluation protocol. Despite these advances, current UAV-based VLN research has not explored multi-agent collaboration for enhanced navigation performance.

2.3 Multi Agent Navigation

Multi-agent collaboration has been widely explored, with early work focusing on reinforcement learning for autonomous coordination in structured environments [2, 38, 42, 46]. Recent advances leverage LLMs to assign roles via prompts, enabling language-based interaction [23]. In ground-based navigation, studies explore multi-agent VLN in indoor environments [66] and cooperative target search in games [63]. By contrast, UAV-focused research centers on swarm formation and obstacle avoidance [12, 41], without addressing multimodal understanding. However, Multi-agent VLN for UAVs remains unexplored, further challenged by UAVs' large operational space and degrees of motion freedom.

3 Dual-Altitude UAV Collaborative Vision–Language Navigation

In this paper, we introduce a novel UAV-based VLN task, Dual-Altitude UAV collaborative VLN (DuAl-VLN), that leverages dual-altitude UAVs to achieve collaborative perception and decision-making. This task strategically balances UAVs' inherent high mobility with constrained operational spaces, creating an optimized environment for model learning while preserving aerial maneuverability advantages. We detail the task setup of the DuAl-VLN in Sec. 3.1. To support this task, we develop the first dual-UAV VLN dataset, HaL-13k, which provides concise instructions of targets, featuring dual-altitude trajectories with multi-modal sensor data (Sec. 3.2).

3.1 Task Formulation

At the beginning of each episode, the low-altitude UAV U_l and the high-altitude UAV U_h are initialized at positions $P_0^l = (x_0^l, y_0^l, z_0^l)$ and $P_0^h = (x_0^h, y_0^h, z_0^h)$, respectively, with $x_0^l = x_0^h$, $y_0^l = y_0^h$, and $z_0^h > z_0^l$. The dual-UAV system receives a target-oriented linguistic instruction describing the target's direction, characteristics, and surrounding environmental context. To reflect the difference in reasoning frequency, we denote the decision time steps for the low-altitude UAV U_l as t and for the high-altitude UAV U_h as τ. Specifically, at each time step t, U_l captures forward-facing visual image I_t^l and omnidirectional point cloud data V_t^l. At each time step τ, U_h captures a BEV observation I_τ^h along with a Lidar point cloud map V_τ^h, both covering the same field of view. Leveraging these multi-modal inputs, the UAVs dynamically adjust their flight trajectories by predicting either subsequent waypoint sequences or velocity profiles supported by Airsim [48]. A navigation episode is deemed successful if U_l comes within a distance threshold d of the target location p^d. A navigation episode is considered failed if either the U_l exceeds the navigation-time upper-bound without reaching the target or the UAVs collide with obstacles.

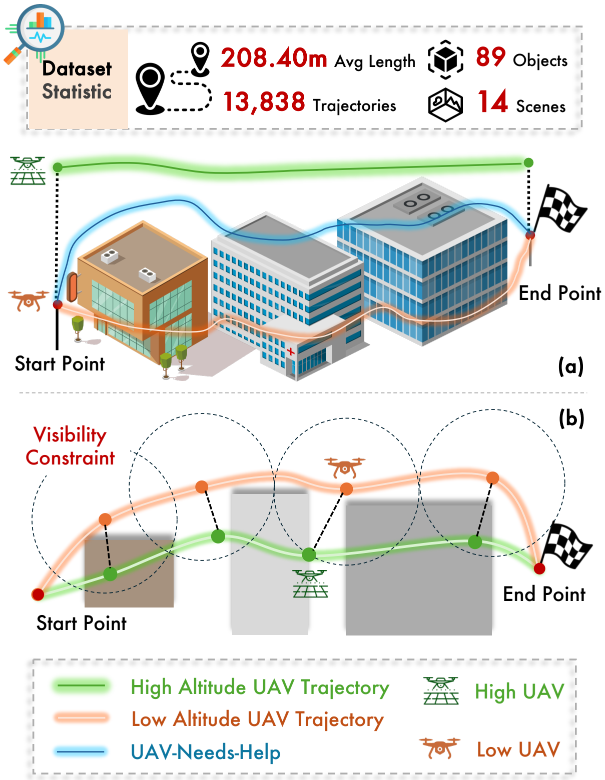


Figure 2: Illustration of dataset statistics and trajectory collection process for HaL-13k. (a) We optimize the paths from UAV-needs-Help [54] to maintain an effective exploration altitude for the low-altitude UAV. (b) We randomly sample paths of the high-altitude UAV that obey visibility constraints, ensuring that the high-altitude UAV maintains visual coverage of the low-altitude UAV's route.

3.2 HaL-13k Dataset

To advance the DuAl-VLN task, we deliberately construct a dataset, HaL-13k, upon the OpenUAV [54] platform, which could provide realistic UAV sensory data, diverse environments, and dynamic flight characteristics. However, existing RL-based multi-agent collaboration approaches, which rely on frequent environmental interactions, impose prohibitive computational costs for simulating. To overcome this limitation, we propose to collect expert continuous trajectory data for coordinated high-low UAV pairs. To build the trajectory pairs, we optimize the navigation paths from [54] to

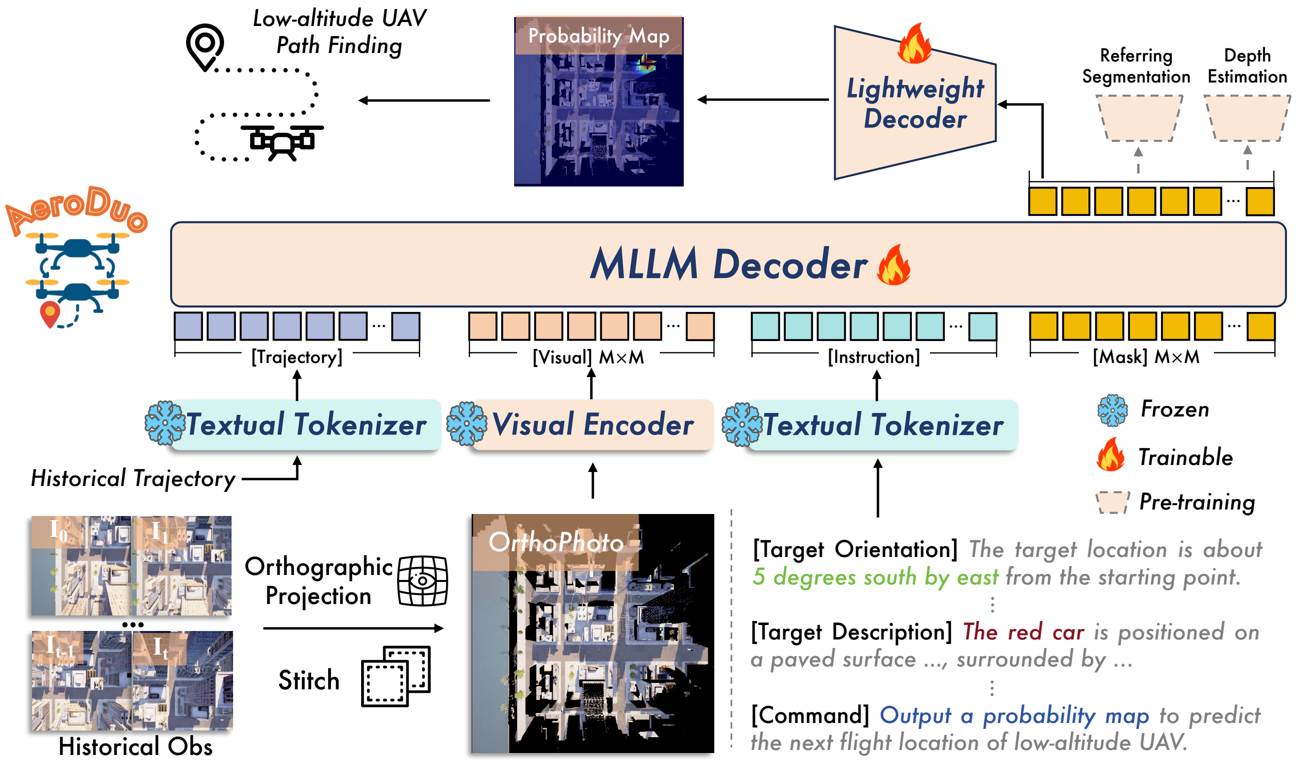


Figure 3: The AeroDuo framework integrates multiple data sources to achieve precise UAV navigation. We input a global orthophoto map, UAV historical trajectories, and linguistic instructions into the MLLM decoder, Pilot-LLM. Special mask tokens are employed to predict the probability distribution of target locations. The flight target is obtained from the resulting probability map, which is subsequently used for pathfinding by low-altitude UAV. To enhance the geospatial modeling capabilities of the MLLM, we pretrain it using auxiliary tasks such as referring segmentation and depth estimation.

maintain an effective exploration altitude for the low-altitude UAV, while generating trajectories with optimal perspective for the high-altitude UAV. Specifically, for the low-altitude UAV path planning, we first construct an occupancy map from point clouds. The A* algorithm [22] is then used to compute optimized navigation paths that conform to these occupancy constraints. For the high-altitude UAV, we randomly sample flight paths that guarantee full visual coverage of the low-altitude UAV's route, thereby maintaining optimal observational perspectives throughout the mission.

As shown in Fig. 2, this generation strategy yields a dataset of 13, 838 collaborative trajectory pairs, each annotated with target-oriented language instructions for search target, without highly detailed route descriptions [19], or real-time human assistance [54].

For the validation set of HaL-13k, we partition it into the unseen map set and the unseen object set, ensuring rigorous evaluation of generalization to novel layouts and unfamiliar objects.

- **Unseen Map Set** consists of environments absent from the training data. We sample 2 scenes as unseen maps and extract all associated trajectories for testing, resulting in 175 episodes.
- **Unseen Object Set** comprises trajectories from training-familiar scenes but introduces object categories never encountered during training, yielding 175 test episodes.

4 AeroDuo: Aerial Duo for UAV-based VLN

4.1 Overview of AeroDuo

We propose a dual-altitude UAV collaborative VLN framework, **AeroDuo**, which integrates MLLMs for high-altitude decision making with lightweight multi-stage policies for low-altitude navigation. In this section, we will provide a detailed overview of each component of AeroDuo. At timestep τ, following an exploration request from the low-altitude UAV U_l, the high-altitude UAV U_h initiates its decision-making phase. For clarity, we denote U_h's decision timesteps as τ in the following part. During this phase, U_h employs an MLLM-based decoder, Pilot-LLM, to predict a target probability map $\mathcal{M}_\tau$ and generates corresponding environmental depth information $\hat{D}^h_\tau$. These outputs are then transmitted to U_l. Equipped with $\mathcal{M}_\tau$ and $\hat{D}^h_\tau$, U_l performs environmental exploration using the Multi-Stage Pathfinder (MSP). Upon completing the exploration, U_l requests the next target from U_h to continue the navigation process. Detailed descriptions of Pilot-LLM and MSP are provided in Sec. 4.2 and Sec. 4.3.

4.2 Pilot-LLM on High UAV

The primary advantage of high-altitude UAVs lies in their wide-field observation capability, which enhances the system's efficiency in both pinpointing target areas and planning subsequent navigation paths. Previous methods process UAV observations as video frames [34, 54], which usually suffer from two critical limitations. First, video sequences prioritize temporal ordering of exploration snapshots, which obscures the spatial relationships between historical observations. Consequently, when key regions span multiple frames, the model would struggle to determine the target location. Second, the lack of explicit spatial coordinate cues in video sequences hinders precise environmental mapping, resulting in unreliable location predictions and compromised results. To alleviate these issues, we propose constructing a global map that integrates historical observations, thereby offering a holistic perspective and a unified spatial coordinate system for the dual-UAV system.

Global Map Construction. Although the high-altitude UAV captures RGB observations in a BEV-like fashion, directly stitching these views into a coherent global map remains non-trivial due to perspective distortions. To address this issue, we employ an orthophoto generation pipeline. Specifically, we first reconstruct an elevation map of the ground environment from the accumulated point clouds $V^h_{1:\tau}$. Then, using the UAV's trajectory p^h_τ, we estimate the extrinsic parameters of the onboard RGB camera and reproject the BEV observations $I^h_{1:\tau}$ to the ground plane. These reprojected views are then stitched together to produce the global orthographic map G_τ. Meanwhile, the current UAV position p^h_τ is also used to compute the global depth map $\hat{D}^h_\tau$ with respect to the UAV's camera plane. The overall process is summarized as:

$$G_\tau, \hat{D}^h_\tau = ortho(I^h_{1:\tau}, p^h_{1:\tau}, V^h_{1:\tau}). \tag{1}$$

We limit the stitching to a maximum of five historical images. The orthophoto map will be resized to a pre-defined size and then fed into the Pilot-LLM as an integrated historical observation. Please refer to the supplementary material for more details.

Pilot-LLM.. To enable effective perception and decision making with multimodal inputs for the dual-UAV system, we take advantage of the reasoning ability of multimodal large language models (MLLMs) to handle various input types and generate pilot guidance in a unified framework, Pilot-LLM.

At time step τ, the input to Pilot-LLM consists of three key components: the orthophoto map G_τ, the historical trajectory $p^h_{1:\tau}$, and the navigation instruction. For the orthophoto map G_τ, we tokenize it into visual tokens $\tilde{G}\tau \in \mathbb{R}^{N\times D}$ using the visual encoder f_v, where N and D represents the number and the dimension of visual tokens. For the historical trajectory $p^h_{1:\tau}$, we first project it onto the coordinate system of G_τ, denoted as $\hat{p}^h_{1:\tau}$, and then encode these positions into trajectory embeddings $\boldsymbol{X}^p_\tau$ using the textual tokenizer. The navigation instruction X is also processed through the LLM's textual tokenizer to generate embeddings $\tilde{\boldsymbol{X}}$. These tokens are then flattened, concatenated, and subsequently fed into the Pilot-LLM.

Given these inputs, Pilot-LLM ought to predict the precise target location for the dual-UAV system. However, directly predicting spatial coordinates of target locations in textual format would result in significant errors, because LLMs struggle with explicit geospatial modeling, as highlighted in [3, 56]. Instead of outputting precise coordinates, we propose to predict a probability distribution map G_τ that emphasizes candidate target regions. This design offers two clear advantages. First, the flight target for UAVs should represent a feasible area rather than a single coordinate point, thereby preserving their exploration ability. Second, by predicting a map, the spatial modeling capacity of Pilot-LLM can be enhanced by incorporating auxiliary tasks, such as referring segmentation and depth estimation, as detailed later in Sec. 4.4.

Formally, to generate a distribution map by Pilot-LLM, we incorporate learnable special tokens $\boldsymbol{M} \in \mathbb{R}^{N\times D}$, where each token corresponds to a unique spatial coordinate in the orthophoto map G_τ. The embedding at the coordinate (i, j) in $\boldsymbol{M}$ is defined as:

$$\boldsymbol{M}(i, j) = \boldsymbol{\rho}_{i,j} + \boldsymbol{\eta}, \tag{2}$$

where $\boldsymbol{\rho}_{i,j} \in R^{1\times D}$ denotes the positional embedding at coordinate (i, j) in f_v and $\boldsymbol{\eta}$ is a trainable embedding. Here, we use the same coordinate representation as f_v to ensure accurate mask prediction. Finally, Pilot-LLM takes $[\tilde{G}_\tau, \boldsymbol{X}^p_\tau, \tilde{\boldsymbol{X}}, \boldsymbol{M}]$ as input, and the output features $\tilde{\boldsymbol{M}}_\tau$ corresponding to the $\boldsymbol{M}$ are decoded by a lightweight mask decoder f_m to predict the probability of target location:

$$\mathcal{M}_\tau = \mathrm{sigmoid}(f_m(\tilde{\boldsymbol{M}}_\tau)), \tag{3}$$

where $\mathcal{M}_\tau$ is a target location probability map with the same spatial size as G_τ. After that, the low-altitude U_l will take the $\mathcal{M}_\tau$ and the global depth map $\hat{D}^h_\tau$ for explorative path finding.

For U_h, its high flight altitude mitigates collision risks while offering an extensive observational field, enabling efficient surveillance of target-proximate areas through orientation-optimized directional movement. During the navigation, U_h first computes the flight direction and step length using instructions and compass data, then follows the predictions throughout subsequent operations.

4.3 Multi-Stage Pathfinder on Low UAV

Upon completion of environmental mapping and target probability estimation by the high-altitude UAV U_h, the low-altitude UAV U_l performs explorative navigation guided by U_h to find the target instance. To enable a safe and efficient environment exploration, we propose a Multi-Stage Pathfinder (MSP) for U_l. As shown in Fig. 4, the MSP pipeline executes navigation through three core stages: key waypoint decision, collision-free navigation, and target localization.

Key Waypoint Decision. In this stage, the low-altitude UAV, located at $p^l_t = [x^l_t, y^l_t, z^l_t]$, first determines its sub-goal by computing the centroid of the probability distribution $\mathcal{M}_\tau$ to effectively mitigate errors caused by outliers:

$$[x^c_\tau, y^c_\tau] = \sum_i^H \sum_j^H \mathcal{M}_\tau(i, j) \cdot [i, j]. \tag{4}$$

Since the high-altitude and low-altitude UAVs start from the same horizontal coordinate (albeit at different altitudes) and the trajectory of the high-altitude UAV is known, it is straightforward to transform the coordinates $[x^c_\tau, y^c_\tau]$ into the global coordinate system. This transformation results in the predicted endpoint $[\hat{x}^c_\tau, \hat{y}^c_\tau, z^l_t]$. With the sub-goal established, the UAV generates a sequence of key waypoints $\boldsymbol{Q}_\tau$ to navigate toward the target. During this process, leveraging U_h's wide-range perspective, which provides a comprehensive understanding of the environmental context, would significantly improve U_l's exploration efficiency. To achieve this, we first construct the occupancy map occ_τ based on the global depth map $\hat{D}^h_\tau$. An approximate navigation path is then derived by optimizing occ_τ using the A* algorithm [22]. The occupancy map occ_τ is calculated as follows:

$$occ_\tau = u(\hat{D}^h_\tau - \Delta z_\tau), \tag{5}$$

where Δz_τ denotes the altitude difference between the high- and low-altitude UAVs, and $u(\cdot)$ denotes the unit step function as:

$$u(x) = \mathbf{1}_{x \geq 0}. \tag{6}$$

The A* search algorithm is performed on occ_τ using the Manhattan distance as the heuristic function to search a trajectory from p^l_t to p^c_τ. Furthermore, an erosion operation is applied to the occupancy map to allow the A* algorithm to search for paths at a safer distance from obstacles. Based on the length of the trajectory, the trajectory is equally segmented into K key waypoints, denoted as $Q_\tau = \{p^\tau_{1:K}\}$.

Collision-Free Navigation. Due to the low spatial resolution of occ_τ, relying solely on the computed waypoints $Q_\tau = \{p^\tau_{1:K_\tau}\}$ from $\hat{D}^h_\tau$ often leads to collisions. To mitigate this issue, we employ an RL-based collision-free Navigator, inspired by [55]. At timestep t of U_l, the Navigator receives the point cloud $\mathbf{V}^l_t$, the subgoal $p^\tau_k \in Q_\tau$, and the ego status, which includes the current position p^l_t and the current velocity v^l_t, for the next timestep's velocity prediction. To efficiently encode $\mathbf{V}^l_t$, we adopt a 3D ray-casting strategy. Thereafter, we feed the encoded point clouds $\hat{V}^l_t$ and the other inputs into a multi-layer perceptron (MLP) to predict the subsequent velocity v^l_{t+1}. The network employs the PPO [47] algorithm for training, with a reward function that incorporates obstacle avoidance, penalties for velocity fluctuations, and incentives for reducing the distance to the target as in [55]. To accelerate the training process, we employed Isaac Sim [40] as the training simulator, leveraging its high-speed parallel simulation. Since the navigator relies solely on point cloud data, it facilitates seamless adaptation across different simulated environments and real-world scenarios. More details are presented in the supplementary material.

Target Localization. During navigation along Q_τ, the low-altitude UAV U_l continuously searches for the target and terminates its exploration once the target is successfully detected. Otherwise, if U_l reaches the end of Q_τ without detection, it will request new navigation guidance from U_h. We follow [54] to adopt GroundingDINO [33] as the detector f_g, enabling target localization based on textual instructions. The visual grounding process operates asynchronously with navigation: after each detection attempt is completed, f_g is immediately applied to the latest observation from U_l. The exploration terminates once the confidence score of the detected bounding box exceeds a predefined threshold.

4.4 Training Process of Pilot-LLM

In this section, we present the training pipeline of Pilot-LLM, which consists of a pretraining stage and a finetuning stage. The pretraining stage aims to enhance the general visual and geospatial modeling ability of Pilot-LLM. In the fine-tuning stage, Pilot-LLM is trained to localize the exploration areas based on BEV observations.

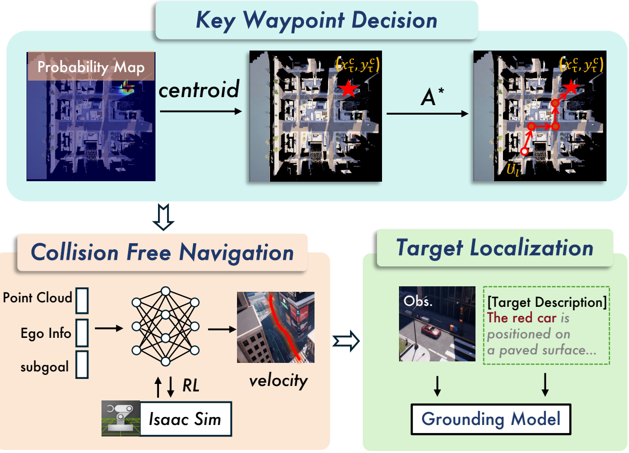


Figure 4: Overview of Multi-Stage Pathfinder (MSP) on low-altitude UAV, including successive stages of Key Waypoint Decision, Collision-Free Navigation and Target Localization.

Pretraining Stage. Since Pilot-LLM needs to locate target regions based on navigation instructions, it requires strong cross-modal understanding capabilities. While general MLLMs provide a solid foundation for this task, they lack two critical capabilities: First, the observations U_h consist of BEV images. Due to the limited amount of BEV data in the general MLLM training, existing models struggle to cross-modal understanding from a BEV perspective. Second, predicting key navigation regions requires robust spatial reasoning ability, which helps to understand whether structures or environmental elements may hinder navigation. To enhance these capabilities, we introduce a pretraining phase for Pilot-LLM. (a) *To boost the model's general perception ability*, we first train the Pilot-LLM on referring segmentation and depth estimation based on the RefCOCO dataset [58]. In our method, we employ the Depth Anything v2 [57] to generate depth maps for training.

After training on the general dataset, (b) *to enhance the ability of BEV image understanding*, we train the model using referring segmentation on BEV images. Multiple text templates are then used to generate descriptive captions for these objects, resulting in a dataset of 850, 894 image-text pairs. During training, the predicted mask is generated by f_m using $\tilde{\boldsymbol{M}}$. (c) *To improve the model's geospatial perception*, we train the Pilot-LLM for depth estimation from BEV images. A separate decoder is employed to ease learning difficulty.

Notably, we train the Pilot-LLM on mixed data from both tasks. This pretraining strategy ensures Pilot-LLM develops both BEV-aware vision-language alignment and precise spatial reasoning for navigation. More details are shown in the supplementary materials.

Finetuning Stage. After the pretraining stage, the Pilot-LLM needs to further locate the key exploration area based on language instructions. Here, we first initialize the mask decoder with the parameters of the segmentation decoder in the pre-training stage, as referring segmentation and mask prediction are similar tasks. To generate the ground truth labels for training, we sample a pair of high-low UAV waypoints at a given time step t. The low-altitude UAV's position at a future time step $t+k$ is used as the target. A Gaussian distribution is then centered at this future position to produce a probability

Table 1: The main comparison on the validation set of HaL-13k. For fair comparison, we train these baseline methods on the proposed dataset of HaL-13k.

Method	Unseen Map					Unseen Object				
	SST ↑	SR ↑	SPL ↑	OSR ↑	NE ↓	SST ↑	SR ↑	SPL ↑	OSR ↑	NE ↓
Random	0.00	0.00	0.00	0.00	199.42	0.00	0.00	0.00	0.00	199.25
CMA [34]	0.00	0.00	0.00	0.57	166.31	0.00	0.00	0.00	0.57	179.30
TravelUAV [54]	0.57	0.57	0.57	1.14	152.20	0.56	0.57	0.52	2.86	160.16
TravelUAV (L1 assistant)	6.48	6.86	5.89	17.14	107.91	5.31	5.71	5.05	10.29	140.42
AeroDuo	**14.63**	**16.57**	**13.86**	**28.57**	**84.31**	**13.54**	**14.86**	**13.35**	**19.43**	**108.66**

distribution map. If the target position at $t + k$ falls outside the orthophoto map, we instead assign the nearest valid point within the map as the surrogate target position. After that, an occupancy mask is applied to suppress infeasible areas by setting their probabilities to zero. Finally, the probability map is normalized to obtain the final ground truth label for the low-altitude UAV. We employ a similar strategy to obtain the ground truth label for the high-altitude UAV, with the objective shifted to predicting its heading and step length.

5 Experiments

5.1 Experiment Setup

Implementation Details. Our Pilot-LLM framework builds upon the visual projector f_v and LLM backbone from Qwen2-VL [52]. The mask prediction head consists of two linear layers followed by two upsampling layers, aiming to expand the output resolution. For optimization, we employ the AdamW [37] optimizer with a cosine learning rate scheduler, initialized at 5×10^{-5}. During training, we freeze the visual projector and fine-tune the MLLM using the low-rank adaptation (LoRA). The mask and depth prediction tasks are optimized using binary cross-entropy loss and MSE loss, respectively. The low-altitude UAV U_l operates at a control frequency of 10 Hz, with each time step t spanning 0.1 seconds.

Evaluation Metrics. We utilize the following 5 metrics to evaluate the performance of the navigation model:

- **SR**: Success Rate. SR measures the percentage of tasks in which the UAV successfully halts within a 20 m radius of the target.
- **SPL**: Success rate weighted by Path Lengh. SPL combines task success with path efficiency by multiplying success rate by the ratio of optimal to actual path length of the low-altitude UAV.
- **SST**: Success rate weighted by Search Time. SST measures navigation time efficiency. It is calculated as: $SST = S \times (T^* \div \max(T, T^*))$ where T is the target search simulator time, and T^* is the navigation time for the ground-truth trajectory.
- **OSR**: Oracle Success Rate. OSR measures whether the UAV reaches a 20 m radius of the target along the trajectory, even if it does not stop at the final destination.
- **NE**: Navigation Error. NE measures the distance between the stop location to the destination.

We take the SST and SR as the main metric for DuAl-VLN.

5.2 Main Comparisons

Comparison Baselines. To validate the effectiveness of our proposed algorithm, we establish both single-UAV and multi-UAV baseline methods. We compare our method with the following four baseline approaches:

- **Random**. The UAV randomly selects an action from four possible directions: forward, left, right, up, or down.
- **Cross-Modal Attention (CMA) model**. A Navigation model proposed in AerialVLN [34], which employs a bi-directional LSTM to jointly process visual inputs and instruction comprehension, predicting the next five waypoints for navigation.
- **TravelUAV** [54]. An LLM-based UAV navigation model introduced by [54]. In TravelUAV, the LLM predicts a long-term waypoint, while an LSTM model fills in the intermediate waypoints.
- **TravelUAV (L1 assistant)** [54]. A variant of TravelUAV enhanced with the L1-level assistant that provides oracle guidance. At each step, the assistant helps to predict the next action by comparing the UAV’s position and orientation with the ground-truth trajectory, ensuring the UAV stays on the correct path.

As shown in Table 1, target-oriented Vision-Language Navigation (VLN) under real-flight conditions in OpenUAV remains a highly challenging task. Existing single-UAV methods, such as CMA [34] and TravelUAV [54], achieve only marginal success and frequently fail to reach the target region. This highlights their limitations in spatial scene understanding and real-world obstacle avoidance. To further assess the performance upper bound, we include TravelUAV (L1 Assistant), a better-performing variant augmented with oracle-level guidance based on ground-truth trajectories. While it offers improved navigation performance, it still falls short in generalizing to long-horizon planning under natural language instructions.

In contrast, our AeroDuo achieves significantly higher success rates and SST across all evaluation splits. It reaches 16.57% SR and 14.63% SST on unseen maps, and 14.86% SR and 13.54% SST on unseen objects, demonstrating the advantage of dual-altitude collaboration in complex real-world environments, marking a solid step toward practical UAV-VLN systems using only target description instructions.

5.3 Ablation Study

We conduct an ablation study to evaluate the contribution of each individual technique. The results are summarized in Table 2, where the following four techniques are examined: MLLM pretraining,

Table 2: The ablation study on the validation set of HaL-13k.

Method				Average		Unseen Map					Unseen Object				
Pretrain	GMC	KWD	CFN	SST ↑	SR ↑	SST ↑	SR ↑	SPL ↑	OSR ↑	NE ↓	SST ↑	SR ↑	SPL ↑	OSR ↑	NE ↓
			✓	1.34	1.43	1.71	1.71	1.62	3.43	146.08	0.97	1.14	0.98	1.71	199.78
✓			✓	2.08	2.29	3.15	3.43	2.97	6.29	107.95	1.02	1.14	0.99	4.57	149.99
✓	✓		✓	12.75	14.29	**17.27**	**19.43**	**16.44**	**32.00**	**78.55**	8.24	9.14	8.12	10.86	126.80
✓	✓	✓		8.08	8.86	8.57	9.14	8.13	24.57	98.89	7.59	8.57	7.30	11.43	135.40
✓	✓	✓	✓	**14.08**	**15.71**	14.63	16.57	13.86	28.57	84.31	**13.54**	**14.86**	**13.35**	**19.43**	**108.66**

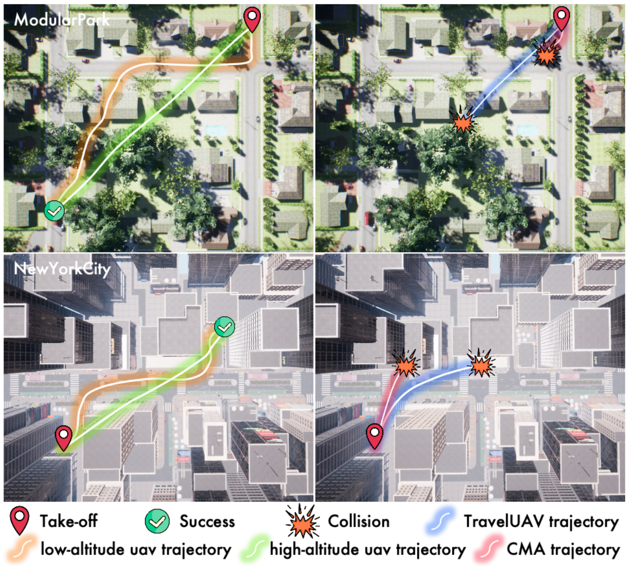


Figure 5: Comparison of UAV target search performance. All methods start from the same take-off position and search for the same target. Left: Our method completes the search without any collisions. Right: TravelUAV and CMA methods result in collisions during the search process.

global map construction (GMC), key waypoint decision (KWD), and collision-free navigation. In the first rows, the Pilot-LLM without GMC employs the original video encoder [52] to observe navigation history. The target distribution is predicted on the current BEV image of U_h. The baseline result indicates that relying solely on the CFN leads to poor performance. Comparatively, leveraging the MLLM pretraining and GMC significantly improves the prediction quality of the MLLM, resulting in an 11.41% increase in SST. In terms of the post-processing and execution of the predicted target point, the results show that the KWD and CFN are critical for the reliable navigation and successful target localization.

5.4 Qualitative Analysis

Trajectory Visualization. As shown in Fig. 5, compared to single-UAV baseline methods, CMA [34] and TravelUAV [54], our method can accurately locate the target region and plan the flight motion over a relatively long horizon. The two baseline methods both suffer from the collision, failing in most cases.

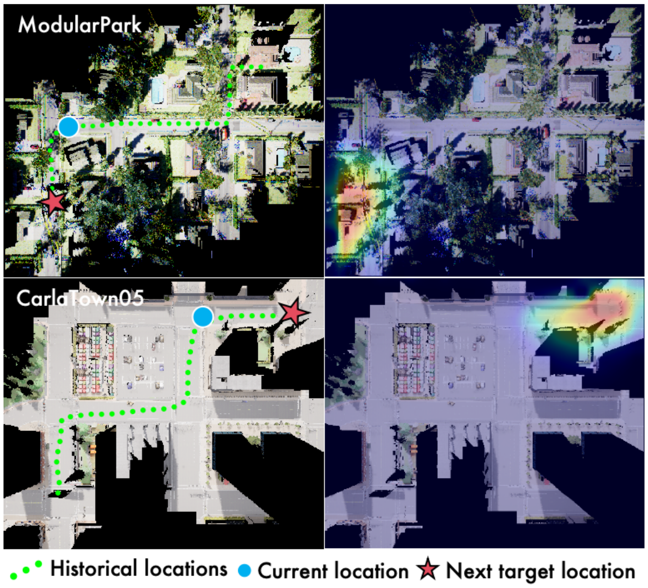


Figure 6: Examples of navigation scenarios and predicted target regions. Left: Orthographic maps and the trajectories of UAV. Right: Predicted heatmaps by Pilot-LLM.

Target Probability Map Prediction. Fig. 6 shows the predicted probability map, where our AeroDuo effectively identifies the destination based on the provided instruction. This highlights the strong capacity of our method to interpret environments from the orthographic map and align observations with complex instructions.

6 Conclusion

In this paper, we propose DuAl-VLN, a dual-altitude UAV collaboration task designed to address vision-language navigation challenges in aerial environments. To support this task, we collected the HaL-13k dataset, containing 13,838 synchronized high-low-altitude trajectories, enabling research on altitude-dependent perception and coordination. We present a novel framework, AeroDuo, to handle the DuAl-VLN task. It integrates a high-altitude Pilot-LLM for semantic mapping and a low-altitude agent for obstacle-aware navigation. Experiments show significantly superior success rates over single-UAV baselines, validating collaborative advantages. These findings validate the effectiveness of dual-altitude collaboration and offer a promising direction for aerial embodied AI systems. Future efforts will focus on optimizing the execution efficiency and scalability.

References

[1] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton Van Den Hengel. 2018. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *CVPR*.

[2] Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, Przemysław Dębiak, Christy Dennison, David Farhi, Quirin Fischer, Shariq Hashme, Chris Hesse, et al. 2019. Dota 2 with large scale deep reinforcement learning. *arXiv* (2019).

[3] Prabin Bhandari, Antonios Anastasopoulos, and Dieter Pfoser. 2023. Are large language models geospatially knowledgeable?. In *Proceedings of the 31st ACM International Conference on Advances in Geographic Information Systems*. 1–4.

[4] Ilker Bozcan and Erdal Kayacan. 2020. AU-AIR: A Multi-modal Unmanned Aerial Vehicle Dataset for Low Altitude Traffic Surveillance. *arXiv preprint* (2020).

[5] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. 2023. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818* (2023).

[6] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. 2017. Matterport3D: Learning from RGB-D Data in Indoor Environments. *3DV* (2017).

[7] Howard Chen, Alane Suhr, Dipendra Misra, Noah Snavely, and Yoav Artzi. 2019. Touchdown: Natural language navigation and spatial reasoning in visual street environments. In *CVPR*.

[8] Jinyu Chen, Chen Gao, Erli Meng, Qiong Zhang, and Si Liu. 2022. Reinforced structured state-evolution for vision-language navigation. In *CVPR*.

[9] Jiaqi Chen, Bingqian Lin, Ran Xu, Zhenhua Chai, Xiaodan Liang, and Kwan-Yee K Wong. 2024. Mapgpt: Map-guided prompting for unified vision-and-language navigation. *arXiv preprint arXiv:2401.07314* (2024).

[10] Shizhe Chen, Pierre-Louis Guhur, Cordelia Schmid, and Ivan Laptev. 2021. History aware multimodal transformer for vision-and-language navigation. *NeurIPS* (2021).

[11] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)* 2, 3 (2023), 6.

[12] Stepan Dergachev and Konstantin Yakovlev. 2021. Distributed Multi-Agent Navigation Based on Reciprocal Collision Avoidance and Locally Confined Multi-Agent Path Finding. In *CASE*.

[13] Yue Fan, Winson Chen, Tongzhou Jiang, Chun Zhou, Yi Zhang, and Xin Eric Wang. 2023. Aerial Vision-and-Dialog Navigation. In *Findings of the Association for Computational Linguistics: ACL 2023*. Association for Computational Linguistics, Toronto, Canada, 3043–3061.

[14] Yue Fan, Shilei Chu, Wei Zhang, Ran Song, and Yibin Li. 2020. Learn by observation: Imitation learning for drone patrolling from videos of a human navigator. In *IROS*.

[15] Daniel Fried, Ronghang Hu, Volkan Cirik, Anna Rohrbach, Jacob Andreas, Louis-Philippe Morency, Taylor Berg-Kirkpatrick, Kate Saenko, Dan Klein, and Trevor Darrell. 2018. Speaker-Follower Models for Vision-and-Language Navigation. *NeurIPS* (2018).

[16] Chen Gao, Jinyu Chen, Si Liu, Luting Wang, Qiong Zhang, and Qi Wu. 2021. Room-and-object aware knowledge reasoning for remote embodied referring expression. In *CVPR*.

[17] Chen Gao, Si Liu, Jinyu Chen, Luting Wang, Qi Wu, Bo Li, and Qi Tian. 2023. Room-object entity prompting and reasoning for embodied referring expression. *IEEE TPAMI* (2023).

[18] Yunpeng Gao, Chenhui Li, Zhongrui You, Junli Liu, Zhen Li, Pengan Chen, Qizhi Chen, Zhonghan Tang, Liansheng Wang, Penghui Yang, et al. 2025. OpenFly: A Versatile Toolchain and Large-scale Benchmark for Aerial Vision-Language Navigation. *arXiv preprint arXiv:2502.18041* (2025).

[19] Yunpeng Gao, Zhigang Wang, Linglin Jing, Dong Wang, Xuelong Li, and Bin Zhao. 2024. Aerial Vision-and-Language Navigation via Semantic-Topo-Metric Representation Guided LLM Reasoning. *arXiv preprint arXiv:2410.08500* (2024).

[20] Alessandro Giusti, Jérôme Guzzi, Dan C Cireşan, Fang-Lin He, Juan P Rodríguez, Flavio Fontana, Matthias Faessler, Christian Forster, Jürgen Schmidhuber, Gianni Di Caro, et al. 2015. A machine learning approach to visual perception of forest trails for mobile robots. *IEEE RAL* (2015).

[21] Zonghao Guo, Ruyi Xu, Yuan Yao, Junbo Cui, Zanlin Ni, Chunjiang Ge, Tat-Seng Chua, Zhiyuan Liu, and Gao Huang. 2025. Llava-uhd: an lmm perceiving any aspect ratio and high-resolution images. In *ECCV*.

[22] Peter E Hart, Nils J Nilsson, and Bertram Raphael. 1968. A formal basis for the heuristic determination of minimum cost paths. *IEEE transactions on Systems Science and Cybernetics* 4, 2 (1968), 100–107.

[23] Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, et al. 2024. MetaGPT: Meta Programming for A Multi-Agent Collaborative Framework. In *ICLR*.

[24] Yicong Hong, Qi Wu, Yuankai Qi, Cristian Rodriguez-Opazo, and Stephen Gould. 2021. A recurrent vision-and-language bert for navigation. In *CVPR*.

[25] Vihan Jain, Gabriel Magalhaes, Alexander Ku, Ashish Vaswani, Eugene Ie, and Jason Baldridge. 2019. Stay on the path: Instruction fidelity in vision-and-language navigation. *arXiv preprint* (2019).

[26] Katie Kang, Suneel Belkhale, Gregory Kahn, Pieter Abbeel, and Sergey Levine. 2019. Generalization through simulation: Integrating simulated and real data into deep reinforcement learning for vision-based autonomous flight. *arXiv preprint* (2019).

[27] Xianghao Kong, Jinyu Chen, Wenguan Wang, Hang Su, Xiaolin Hu, Yi Yang, and Si Liu. 2024. Controllable navigation instruction generation with chain of thought prompting. In *European Conference on Computer Vision*. Springer, 37–54.

[28] Jacob Krantz, Erik Wijmans, Arjun Majumdar, Dhruv Batra, and Stefan Lee. 2020. Beyond the nav-graph: Vision-and-language navigation in continuous environments. In *ECCV*.

[29] Alexander Ku, Peter Anderson, Roma Patel, Eugene Ie, and Jason Baldridge. 2020. Room-Across-Room: Multilingual Vision-and-Language Navigation with Dense Spatiotemporal Grounding. In *EMNLP*.

[30] Jungdae Lee, Taiki Miyanishi, Shuhei Kurita, Koya Sakamoto, Daichi Azuma, Yutaka Matsuo, and Nakamasa Inoue. 2024. CityNav: Language-Goal Aerial Navigation Dataset with Geographic Information. *arXiv preprint* (2024).

[31] Hongyu Li, Jinyu Chen, Ziyu Wei, Shaofei Huang, Tianrui Hui, Jialin Gao, Xiaoming Wei, and Si Liu. 2025. LLaVA-ST: A Multimodal Large Language Model for Fine-Grained Spatial-Temporal Understanding. *arXiv preprint arXiv:2501.08282* (2025).

[32] Jialu Li and Mohit Bansal. 2024. Panogen: Text-conditioned panoramic environment generation for vision-and-language navigation. *Advances in Neural Information Processing Systems* 36 (2024).

[33] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. 2023. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499* (2023).

[34] Shubo Liu, Hongsheng Zhang, Yuankai Qi, Peng Wang, Yanning Zhang, and Qi Wu. 2023. Aerialvln: Vision-and-language navigation for uavs. In *ICCV*.

[35] Youzhi Liu, Fanglong Yao, Yuanchang Yue, Guangluan Xu, Xian Sun, and Kun Fu. 2024. NavAgent: Multi-scale Urban Street View Fusion For UAV Embodied Vision-and-Language Navigation. *arXiv preprint* (2024).

[36] Antonio Loquercio, Ana I Maqueda, Carlos R Del-Blanco, and Davide Scaramuzza. 2018. Dronet: Learning to fly by driving. *IEEE RAL* 3, 2 (2018), 1088–1095.

[37] Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017).

[38] Ryan Lowe, Yi I Wu, Aviv Tamar, Jean Harb, OpenAI Pieter Abbeel, and Igor Mordatch. 2017. Multi-agent actor-critic for mixed cooperative-competitive environments. *NeurIPS* (2017).

[39] András L Majdik, Charles Till, and Davide Scaramuzza. 2017. The Zurich urban micro aerial vehicle dataset. *The IJRR* (2017).

[40] Viktor Makoviychuk, Lukasz Wawrzyniak, Yunrong Guo, Michelle Lu, Kier Storey, Miles Macklin, David Hoeller, Nikita Rudin, Arthur Allshire, Ankur Handa, et al. 2021. Isaac gym: High performance gpu-based physics simulation for robot learning. *arXiv preprint arXiv:2108.10470* (2021).

[41] Yash Vardhan Pant, Houssam Abbas, Rhudii A Quaye, and Rahul Mangharam. 2018. Fly-by-logic: Control of multi-drone fleets with temporal logic objectives. In *ICCPS*. IEEE, 186–197.

[42] Diego Perez-Liebana, Katja Hofmann, Sharada Prasanna Mohanty, Noburu Kuno, Andre Kramer, Sam Devlin, Raluca D Gaina, and Daniel Ionita. 2019. The multi-agent reinforcement learning in malm\" o (marl\" o) competition. *arXiv preprint arXiv:1901.08129* (2019).

[43] Yuankai Qi, Qi Wu, Peter Anderson, Xin Wang, William Yang Wang, Chunhua Shen, and Anton van den Hengel. 2020. Reverie: Remote embodied visual referring expression in real indoor environments. In *CVPR*.

[44] Yanyuan Qiao, Qianyi Liu, Jiajun Liu, Jing Liu, and Qi Wu. 2024. LLM as Copilot for Coarse-Grained Vision-and-Language Navigation. In *European Conference on Computer Vision*. Springer, 459–476.

[45] Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. 2024. Timechat: A time-sensitive multimodal large language model for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14313–14323.

[46] Mikayel Samvelyan, Tabish Rashid, Christian Schroeder De Witt, Gregory Farquhar, Nantas Nardelli, Tim GJ Rudner, Chia-Man Hung, Philip HS Torr, Jakob Foerster, and Shimon Whiteson. 2019. The starcraft multi-agent challenge. *arXiv* (2019).

[47] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347* (2017).

[48] Shital Shah, Debadeepta Dey, Chris Lovett, and Ashish Kapoor. 2018. Airsim: High-fidelity visual and physical simulation for autonomous vehicles. In *Field and Service Robotics: Results of the 11th International Conference*. Springer, 621–635.

[49] Abhik Singla, Sindhu Padakandla, and Shalabh Bhatnagar. 2019. Memory-based deep reinforcement learning for obstacle avoidance in UAV with limited environment knowledge. *IEEE TIST* (2019).

[50] Nikolai Smolyanskiy, Alexey Kamenev, Jeffrey Smith, and Stan Birchfield. 2017. Toward low-flying autonomous MAV trail navigation using deep neural networks for environmental awareness. In *IROS*.

[51] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).

[52] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution. *arXiv preprint arXiv:2409.12191* (2024).

[53] Su Wang, Ceslee Montgomery, Jordi Orbay, Vighnesh Birodkar, Aleksandra Faust, Izzeddin Gur, Natasha Jaques, Austin Waters, Jason Baldridge, and Peter Anderson. 2022. Less is more: Generating grounded navigation instructions from landmarks. In *CVPR*.

[54] Xiangyu Wang, Donglin Yang, Ziqin Wang, Hohin Kwan, Jinyu Chen, Wenjun Wu, Hongsheng Li, Yue Liao, and Si Liu. 2024. Towards Realistic UAV Vision-Language Navigation: Platform, Benchmark, and Methodology.

[55] Zhefan Xu, Xinming Han, Haoyu Shen, Hanyu Jin, and Kenji Shimada. 2025. Navrl: Learning safe flight in dynamic environments. *IEEE RAL* (2025).

[56] Yutaro Yamada, Yihan Bao, Andrew K Lampinen, Jungo Kasai, and Ilker Yildirim. 2023. Evaluating spatial understanding of large language models. *arXiv preprint arXiv:2310.14540* (2023).

[57] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. 2024. Depth anything v2. *NeurIPS* 37 (2024), 21875–21911.

[58] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. 2016. Modeling context in referring expressions. In *ECCV*. Springer, 69–85.

[59] Haitian Zeng, Xiaohan Wang, Wenguan Wang, and Yi Yang. 2023. Kefa: A Knowledge Enhanced and Fine-grained Aligned Speaker for Navigation Instruction Generation. *arXiv preprint arXiv:2307.13368* (2023).

[60] Jiazhao Zhang, Kunyu Wang, Rongtao Xu, Gengze Zhou, Yicong Hong, Xiaomeng Fang, Qi Wu, Zhizheng Zhang, and He Wang. 2024. Navid: Video-based vlm plans the next step for vision-and-language navigation. *arXiv preprint arXiv:2402.15852* (2024).

[61] Yue Zhang and Parisa Kordjamshidi. 2023. VLN-Trans, Translator for the Vision and Language Navigation Agent. In *ACL*.

[62] Yusheng Zhao, Jinyu Chen, Chen Gao, Wenguan Wang, Lirong Yang, Haibing Ren, Huaxia Xia, and Si Liu. 2022. Target-Driven Structured Transformer Planner for Vision-Language Navigation. In *ACM MM*.

[63] Zhonghan Zhao, Kewei Chen, Dongxu Guo, Wenhao Chai, Tian Ye, Yanting Zhang, and Gaoang Wang. 2024. Hierarchical auto-organizing system for open-ended multi-agent navigation. *arXiv* (2024).

[64] Duo Zheng, Shijia Huang, Lin Zhao, Yiwu Zhong, and Liwei Wang. 2024. Towards learning a generalist model for embodied navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13624–13634.

[65] Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Shitao Xiao, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. 2024. MLVU: A Comprehensive Benchmark for Multi-Task Long Video Understanding. *arXiv preprint arXiv:2406.04264* (2024).

[66] Fengda Zhu, Vincent CS Lee, and Rui Liu. 2024. Communicative and Cooperative Learning for Multi-agent Indoor Navigation. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 273–285.