AerialVLA: A Vision-Language-Action Model for Aerial Navigation with Online Dialogue

Jinyu Chen[1*], **Hongyu Li**[1*], **Zongheng Tang**[1,2], **Xiaoduo Li**[1], **Wenjun Wu**[1], **Si Liu**[1†]

[1]Beihang University

[2] Hangzhou International Innovation Institute, Beihang University

Abstract

Visual Dialogue Navigation (VDN) aims to enable agents to reach target locations through dialogue with humans. The integration of VDN into Unmanned Aerial Vehicle (UAV) systems enhances human-machine interaction by enabling intuitive, hands-free operation, thereby unlocking vast applications. However, existing VDN models for UAVs can only perform navigation based on dialogue history, lacking proactive interaction capabilities to correct trajectories. Moreover, their sequential observation history recording mechanism struggles to accurately localize landmarks observed in the historical context, leading to ineffective utilization of referential information in new user instructions. To address these, we present AerialVLA, an end-to-end UAV navigation framework integrating dialogue comprehension, action decision-making, and navigational question generation. AerialVLA comprises three core components: i) we propose the Progress-Driven Navigation-Query Alternation mechanism to determine optimal questioning timing through navigation progress estimation autonomously. ii) To effectively model long-horizon history observation sequences, we develop the History Spatial-Temporal Fusion module that extracts discriminative spatial-temporal representations from historical observations. iii) Furthermore, to overcome data scarcity in training, we devise the Online Task-Driven Augmentation strategy that enhances learning through action-conditioned data augmentation. Experimental results demonstrate that AerialVLA achieves state-of-the-art navigation performance while exhibiting effective dialogue capabilities. Moreover, to better evaluate the agent's proactive dialogue and navigation abilities, our evaluation benchmark, named UAV Navigation with Online Dialogue (UNOD), incorporates an online dialogue interaction module. The UNOD assesses UAV agents' real-time questioning capabilities by leveraging an Air Commander Large Language Model to simulate human-UAV interactions during testing.

1 Introduction

The goal of vision-and-language navigation (VLN) (Anderson et al. 2018) is to enable agents to navigate based on human language instructions and visual observations autonomously. The inherent ambiguity in linguistic descriptions often renders single-turn instructions insufficient for precisely localizing target positions, analogous to how humans typically engage in multi-round dialogues for route clarification. This underscores the significance of vision-and-dialogue navigation (Thomason et al. 2020) (VDN), which has garnered substantial attention in ground-agent scenarios. With the growing ubiquity of unmanned aerial vehicles (UAVs) in daily applications such as delivery and aerial photography, UAV-based VDN (Fan et al. 2023) exhibits broad application potential across various domains, enabling hands-free operation and broader spatial coverage. Compared to ground agents, UAVs exhibit distinct characteristics in action space, navigation scale, and observational perspective, presenting novel research challenges.

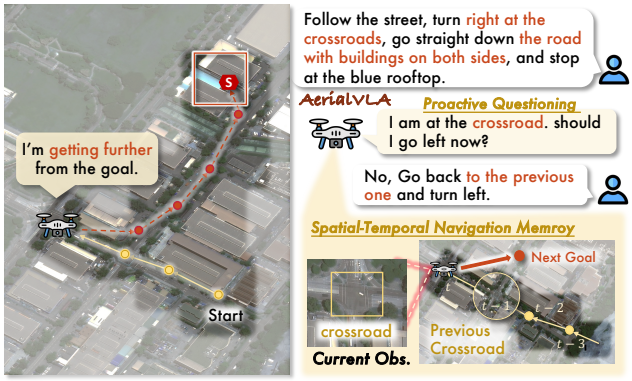

Figure 1: AerialVLA is capable of proactively ask questions to users during navigation and maintain a spatial-temporal Navigation memory.

However, two critical shortcomings of prior VDN methods for UAVs remain unsolved. **First**, existing models rely solely on human-human dialogue history for navigation, which limits their ability to query destinations during the navigation process actively (Zheng et al. 2024; Fan et al. 2023). Consequently, the UAV agent cannot detect directional errors and proactively seek corrective guidance, nor can it refine its navigation trajectory through online interactive dialogue incrementally. **Second**, in the dialogue process, newly acquired instructions often correlate with obser-

[*]These authors contributed equally.

[†]The corresponding author is Si Liu.

vations or dialogue from several steps earlier, for instance, in Figure 1, a directive such as "turn left at the ***previous one*** you passed.". Understanding this directive necessitates spatial and temporal reasoning across observations from different time steps. However, existing approaches (Fan et al. 2023; Wang et al. 2024a) typically process observational history as a video sequence, which fails to preserve the underlying spatial relationships adequately.

To address these limitations, we propose AerialVLA, an end-to-end framework that integrates navigation decision-making, conversational comprehension, and context-aware questioning abilities. i) To allow the agent to actively query humans' response at an appropriate time based on real-time scene observations, we introduce the **P**rogress-Driven **Na**vigation-**Q**uery Alternation (**PNaQ**) for AerialVLA. PNaQ dynamically determines whether to query the human for guidance or continue navigation based on the agent's confidence in the current route. A decline in navigation progress estimation indicates reduced confidence, triggering a human query for route correction, as shown in Figure 1 left. To prevent interference between AerialVLA's action prediction and language decoding, we introduce an additional action prediction head in PNaQ, which simultaneously estimates the UAV's flight azimuth and altitude adjustments. ii) To effectively utilize the spatio-temporal relations across historical observations, we propose the **H**istory **S**patio-**T**emporal **F**usion (**HSTF**) module. In HSTF, we construct a global map by aggregating historical observations to preserve spatial information, while employing a **S**patio-**T**emporal **A**ttention module (**STA**) to build a multi-grained relation between time-sequential observations and the global map, thereby maintaining both temporal and spatial coherence for navigation memories. iii) Given the limited data available in VDN for UAV, the training process necessitates efficient utilization of the scarce navigation data. We implement a **O**ne-Trial **D**ata **A**ugmentation (**OTDA**) strategy that enriches trajectory samples by iteratively recording and correcting the model's action predictions during training. The components together make AerialVLA the state-of-the-art VDN model for UAVs. On the unseen testing split of AVDH-Full (Fan et al. 2023), AerialVLA achieves a relative improvement of 105% in SPL.

To validate the capability of AerialVLA in conducting online dialogue and navigation, we design the **U**AV **N**avigation with **O**nline **D**ialogue (**UNOD**) benchmark, which simulates human users who can answer the UAV's questions at any time during the navigation process. To achieve this, we build the **A**erial **C**ommander **LLM** (**AC-LLM**), which is trained on multimodal datasets involving remote sensing imagery (Kuckreja et al. 2024) and aerial VDN data from the AVDH dataset (Fan et al. 2023). AC-LLM can provide stable, accurate, and diverse navigation instruction responses to questions from UAVs based on the navigation state.

In summary, our contributions are fourfold:

- We introduce AerialVLA, an end-to-end model that can proactively dialogue with humans during navigation via PNaQ, and achieves state-of-the-art performance AVDH.
- We propose the HSTF module, maintaining spatial relationships of historical observations alongside temporal exploration order, delivering high-fidelity spatio-temporal representations critical for aerial navigation.
- We introduce the OTDA training strategy, which integrates intermediate action predictions into training data to increase the trajectory diversity.
- We propose the UNOD benchmark, enabling simulation of human interlocutors' response to evaluate UAV agents' VDN capabilities in online dialogue scenarios.

2 Related Works

2.1 Vision-and-Dialogue Navigation

Significant advancements have been achieved in the language-guided navigation area. Anderson et al. (Anderson et al. 2018) introduce the first VLN dataset that utilizes fine-grained language instructions within a graph-based action space, leveraging the Matterport3D dataset (Chang et al. 2017). Furthermore, established benchmarks including (Qi et al. 2020; Jain et al. 2019; Ku et al. 2020; Chen et al. 2019; Krantz et al. 2020) significantly enhance the coverage of visual-language navigation tasks through diversified linguistic instructions and heterogeneous environmental configurations. CVDN (Thomason et al. 2020) pioneers dialog-aware navigation by incorporating human conversation histories as textual inputs, while AVDN (Fan et al. 2023) extends this paradigm to UAV navigation. Conventional VDN solutions typically treat dialog history as specialized linguistic instructions, employing action prediction strategies similar to VLN frameworks (Chen et al. 2021; Hao et al. 2020; Qiao et al. 2022). To enable real-time conversational capabilities, RMM (Roman et al. 2020; Nguyen and Daumé III 2019) introduces auxiliary dialog agents for online interaction management, whereas SCoA (Zhu* et al. 2021) proposes the When-to-Ask (WeTA) mechanism for dynamic questioning timing determination. Current online VDN architectures, however, exhibit limitations in unified dialogue-action integration due to over-simplified action spaces or template-based dialogue generation approaches that constrain flexibility in conversational interactions.

2.2 Drone-based Navigation

The previous works on UAV-based navigation have primarily concentrated on vision-based navigation capabilities (Loquercio et al. 2018; Giusti et al. 2015; Smolyanskiy et al. 2017; Fan et al. 2020; Bozcan and Kayacan 2020; Majdik, Till, and Scaramuzza 2017; Kang et al. 2019) and obstacle avoidance abilities (Xu et al. 2025; Singla, Padakandla, and Bhatnagar 2019). The integration of both vision and language for aerial navigation remains a relatively underexplored area of study. Methods like (Liu et al. 2023, 2024) introduce the VLN task for drones using fine-grained instructions. The OpenUAV (Wang et al. 2024b) features a UAV-specific dynamic movement simulation environment and introduces the UAV-Need-Help benchmark, a human-assisted navigation task. CityNav (Lee et al. 2024) provides 32,637 human-demonstrated trajectories paired with natural language instructions across real-world cities. OpenFly (Gao

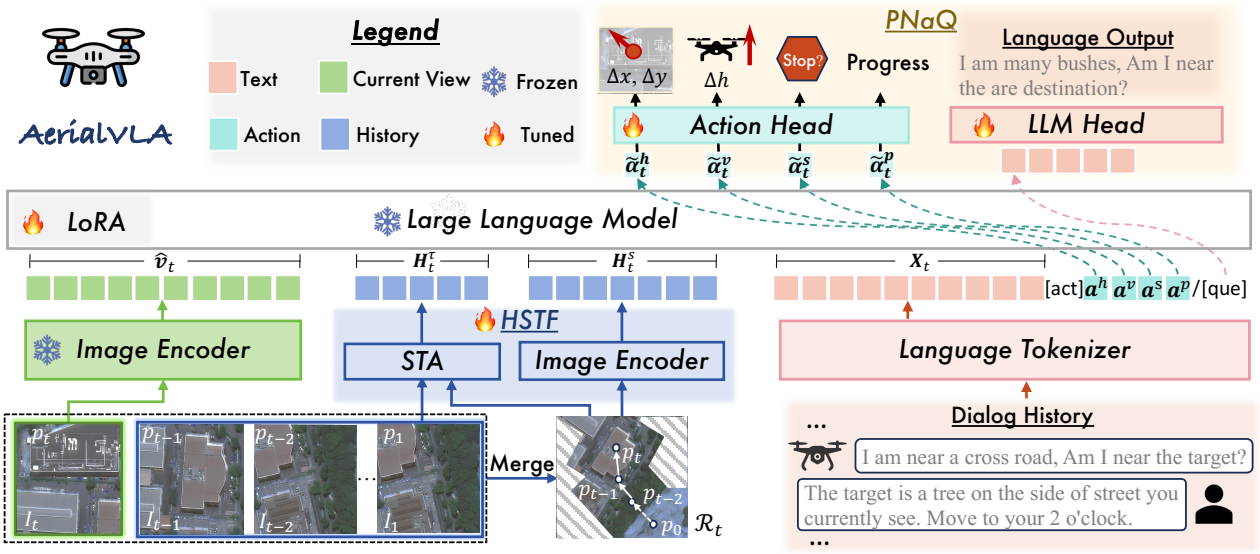

Figure 2: AerialVLA comprises three input streams: (1) the current visual observations, (2) historical visual information processed through HSTF module before being fed into the LLM, and (3) encoded dialogue history. These multimodal streams are jointly processed through the LLM. PNaQ is employed to determine whether asking or moving.

et al. 2025) introduces a large-scale dataset comprising realistic urban layouts paired with natural language instructions. AVDN (Fan et al. 2023) establishes a benchmark for drone navigation driven by dialogue history, featuring 3,064 aerial navigation trajectories with human-human dialogue.

3 Task Formulation

In VDN for UAV task, navigation process initiates with the UAV at $p_1 = [x_1, y_1, z_1]$ receiving a linguistic instruction r_0 specifying the target observation I_g located at coordinates $p_g = [x_g, y_g, z_g]$. At each time step t, the agent positioned at $p_t = [x_t, y_t, z_t]$ captures a visual observation I_t. The agent then executes a decision protocol: either generating an inquiry q_t to request supplemental guidance, or directly predicting navigation actions. When initiating an inquiry, the AC-LLM module generates a context-aware response r_t, with each question-answer pair $\langle q_t, r_t \rangle$ being sequentially recorded in the dialogue history $\mathcal{D}_t = \{r_0, \langle q_{t_{d_1}}, r_{t_{d_1}} \rangle, ..., \langle q_{t_{d_k}}, r_{t_{d_k}} \rangle\}$. The navigation action space comprises continuous 3D waypoint adjustments $a_t = [\Delta x_t, \Delta y_t, \Delta z_t]$. A navigation episode is deemed successful if the agent's terminal position p_{stop} satisfies $\|p_{stop} - p_g\| < d_g$. If the agent exceeds the maximum step limit T or fails to stop within the vicinity of the target position, the episode is considered a failure.

3.1 Method Overview

The architecture of AerialVLA is shown in Figure 2. At each time step t, the model processes three modalities: dialog history, historical observations, and current visual input. The dialogue history $\mathcal{D}_t$ is first encoded into text embeddings through the pretrained LLM's tokenizer:

$$\boldsymbol{X_t} = \texttt{Tokenizer}(\mathcal{D}_t) \in \mathbb{R}^{D \times N_d}, \tag{1}$$

where D denotes the dimension of the LLM's embedding space and N_d represents the sequence length of language tokens. For the current visual observation I_t, we extract visual feature $\boldsymbol{v}_t$ using a vision encoder f_v followed by a flattening operation and a linear projection layer $\phi(\cdot)$ (Guo et al. 2025) to align with the text embedding space:

$$\boldsymbol{v}_t = f_v(I_t), \tag{2}$$

$$\hat{\boldsymbol{v}}_t = \phi(\texttt{flatten}(\boldsymbol{v}_t)), \tag{3}$$

where $\hat{\boldsymbol{v}}_t \in \mathbb{R}^{D \times N_v}$ and N_v indicates the number of visual tokens, and D is the dimension of LLM's hidden state. Historical features are captured through the HSTF module:

$$[\boldsymbol{H}_t^s, \boldsymbol{H}_t^\tau] = \texttt{HSTF}(I_{1:t}, p_{1:t}). \tag{4}$$

The LLM backbone integrates $[\hat{\boldsymbol{v}}_t, \boldsymbol{H}_t^s, \boldsymbol{H}_t^\tau, \boldsymbol{X}_t]$ along with four trainable special tokens. The corresponding outputs of the special tokens are routed to the PNaQ for action prediction. Details of HSTF and PNaQ modules are elaborated in §3.2 and §3.3. During training, we employ OTDA to enhance training data, which is discussed in §3.4.

3.2 History Spatial-Temporal Fusion Module

We propose HSTF to comprehensively capture and integrate historical observations in both spatial and temporal aspects, as shown in Figure 3. In HSTF, we begin by constructing a global map to obtain a comprehensive spatial representation of historical trajectories. Subsequently, we employ the STA mechanism to extract the multi-grained correlation between observation sequence of the UAV and this global map. This design enables AerialVLA to simultaneously achieve spatial

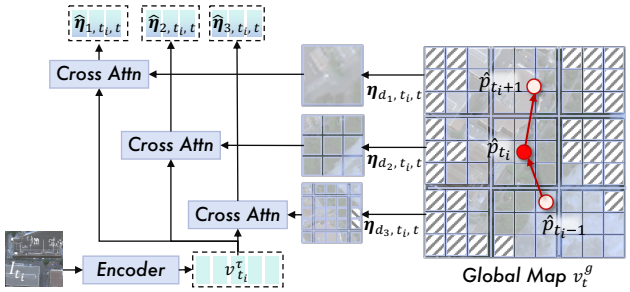

Figure 3: The architecture of the Spatial-Temporal Attention (STA) module. It processes visual features from each time step and fuses them with multi-scale features from the merged global map through cross-attention layers.

reasoning for navigation while effectively aligning temporal relations with dialogue content.

Global Map Construction. We first construct a continuously evolving global map from the agent's observational history to enhance spatial reasoning. Given the history BEV perspective of the UAV, we first aggregate the observation images based on their geolocations $p_{1:t}$, yielding a spatial representation of the global map $\mathcal{M}_t$. To address irregular boundaries of $\mathcal{M}_t$, we compute the minimum bounding rectangle encompassing all pixels in $\mathcal{M}_t$, followed by zero-padding implementation for invalid regions within it to build a global map $\mathcal{R}_t$. The history spatial feature is calculated as:

$$\boldsymbol{H}_t^s = \phi(\texttt{flatten}(f_v(\mathcal{R}_t))) \in \mathbb{R}^{D\times N_s}, \tag{5}$$

where the N_s is the number of visual tokens of $\boldsymbol{H}_t^s$. $\boldsymbol{H}_t^s$ is inputted into the LLM as part of the history representation in Equation 4. At the same time, $\mathcal{R}_t$ is subsequently processed by STA to build correlations with the exploration order.

Spatial-Temporal Attention (STA). While AerialVLA captures the spatial relationships of observations up to the current moment in the global map, the exploration sequence, *i.e.*, the order in which observations are acquired—remains undetermined. However, the exploration sequence is important information to align with the dialogue content for navigation decision. we take the observation sequence at each time step $\boldsymbol{v}_{1:t-1}$ as the temporal representation, and apply a 2D average pooling along the spatial dimensions of $\boldsymbol{v}_{1:t-1}$ to reduce computational load.

$$\boldsymbol{v}_{1:t-1}^\tau = \texttt{flatten}(\texttt{pooling}_{2\text{D}}(\boldsymbol{v}_{1:t-1})), \tag{6}$$

where $\boldsymbol{v}_{1:t-1}^\tau \in \mathbb{R}^{D_v\times N_\tau\times(t-1)}$. As for $\mathcal{R}_t$, we utilize the AnyRes image encoder $\hat{f}_v$ in (Li et al. 2024) to obtain a high-resolution global map representation $\boldsymbol{v}_t^g = \hat{f}_v(\mathcal{R}_t)$, where $\hat{f}_v$ operates by splitting the image into grids and encoding them independently.

We then employ a multi-grained attention mechanism to link the $\boldsymbol{v}_t^g$ with $\boldsymbol{v}_{1:t-1}^\tau$, as illustrated in Figure 3. The navigation waypoints $p_{1:t-1}$ are projected onto $\boldsymbol{v}_t^g$ and obtain the positions $\hat{p}_{1:t-1}$ on the feature map. $\hat{p}_{1:t-1}$ are employed as anchor points to establish the association between $\boldsymbol{v}_{1:t-1}^\tau$ and $\boldsymbol{v}_t^g$. For each time step $t_i < t$, we crop the visual embeddings from $\boldsymbol{v}_t^g$ within a Euclidean distance of d_j to $\hat{p}_{t_i}$:

$$\boldsymbol{\eta}_{d_j,t_i,t} = \texttt{crop}(\boldsymbol{v}_t^g, d_j, \hat{p}_{t_i}), \tag{7}$$

where $\boldsymbol{\eta}_{d_j,t_i,t} \in \mathbb{R}^{D\times N_{d_j}}$. $\boldsymbol{\eta}_{d_j,t_i,t}$ subsequently computes the cross attention with $\boldsymbol{v}_{t_i}^\tau$ to integrate global relative positional information, where $\boldsymbol{v}_{t_j}^\tau$ serves as the query:

$$\hat{\boldsymbol{\eta}}_{d_j,t_i,t} = \texttt{cross\_attn}(\boldsymbol{v}_{t_i}^\tau, \boldsymbol{\eta}_{d_j,t_i,t}), \tag{8}$$

where $\hat{\boldsymbol{\eta}}_{d_j,t_i,t} \in \mathbb{R}^{D_v\times N_\tau}$ and $\boldsymbol{\epsilon}_{t_i}$ is the temporal embedding for time step t_i. To incorporate multi-scale feature information and enhance the representation density, we leverage three distinct distances $d_j \in \{d_1, d_2, d_3\}$ to extract hierarchical features from $\boldsymbol{v}_t^g$, yielding the aggregated representation $\boldsymbol{\eta}_{d_{1:3},t_i,t}$. Subsequently, the representation of the history time step t_i can be computed as:

$$\boldsymbol{h}_{t_i,d_j,t}^\tau = \hat{\boldsymbol{\eta}}_{d_j,k,t} + \boldsymbol{v}_{t_i}^\tau + \boldsymbol{\epsilon}_{t_i}, \tag{9}$$

$$\boldsymbol{h}_{t_i,t}^\tau = [\boldsymbol{h}_{t_i,d_1,t}^\tau, \boldsymbol{h}_{t_i,d_2,t}^\tau, \boldsymbol{h}_{t_i,d_3,t}^\tau], \tag{10}$$

where $\boldsymbol{h}_{t_i,t}^\tau \in \mathbb{R}^{D_v\times(N_\tau\times 3)}$ and $\boldsymbol{\epsilon}_{t_i}$ represents the positional embedding at time step t_i. Finally, the features of $\boldsymbol{h}_{1:t-1,t}^\tau$ are concatenated and processed through $\phi(\cdot)$ to obtain $\boldsymbol{H}_t^\tau$:

$$\boldsymbol{H}_t^\tau = \phi(\boldsymbol{h}_{1:t-1,t}^\tau) \in \mathbb{R}^{D\times(N_\tau\times 3\times(t-1))}. \tag{11}$$

$\boldsymbol{H}_t^\tau$ is then incorporated into the LLM as the exploration sequence-related historical context, enabling PNaQ to execute navigation decisions and dynamically trigger question generation depending on the navigation state.

3.3 Progress-Driven Navigation-Query Alternation

To address the inherent divergence between action decision-making and language generation, PNaQ employs dedicated prediction heads for each task, thereby minimizing potential interference, as shown in Figure 2. Furthermore, we introduce two special tokens, `[act]` and `[que]`, appended to the end of the input sequence. These tokens explicitly prompt the model to initiate task-specific feature aggregation—`[act]` for action prediction and `[que]` for language generation. At each time step, AerialVLA first appends the `[act]` token to the input sequence to predict the following action, where asking a question is treated as a valid action. If the action head selects to ask a question, the `[que]` token is appended to the LLM's input, triggering the question-generation process to acquire more navigation guidance.

To predict the next action, we append four learnable tokens: $\boldsymbol{\alpha}^h$, $\boldsymbol{\alpha}^v$, $\boldsymbol{\alpha}^s$, and $\boldsymbol{\alpha}^p$, after the `[act]` token. Their corresponding outputs ($\tilde{\boldsymbol{\alpha}}_t^h$, $\tilde{\boldsymbol{\alpha}}_t^v$, $\tilde{\boldsymbol{\alpha}}_t^s$, $\tilde{\boldsymbol{\alpha}}_t^p$) jointly govern UAV's behavior. Horizontal and vertical displacement vectors are derived from $\tilde{\boldsymbol{\alpha}}_t^h$ and $\tilde{\boldsymbol{\alpha}}_t^v$:

$$[\Delta\hat{x}_t, \Delta\hat{y}_t] = \phi_h(\tilde{\boldsymbol{\alpha}}_t^h),\ \Delta\hat{z}_t = \phi_v(\tilde{\boldsymbol{\alpha}}_t^v), \tag{12}$$

where $\phi_h(\cdot)$,$\phi_v(\cdot)$ are MLP layers. $\tilde{\boldsymbol{\alpha}}_t^s$ is used to predict the trajectory termination probability:

$$\hat{s}_t = \sigma(\phi_s(\tilde{\boldsymbol{\alpha}}_t^s)), \tag{13}$$

where $\sigma(\cdot)$ is the sigmoid function and $\phi_s(\cdot)$ is a MLP layer. AerialVLA stops when $\hat{s}_t > 0.5$.

A straightforward approach to determining when to ask questions is to use pre-recorded human demonstration data as imitation learning targets, thereby learning human-like questioning timing. However, due to the limited volume of available VDN data for UAVs, we instead adopt a progress-based estimation method to decide the questioning moments. Here, progress is defined as:

$$\varphi_t = \frac{\|p_t - p_1\|}{\|p_t - p_g\| + \|p_t - p_1\|}. \tag{14}$$

AerialVLA uses $\tilde{\boldsymbol{\alpha}}_t^p$ to estimate φ_t via MLP ϕ_φ:

$$\hat{\varphi}_t = \sigma(\phi_\varphi(\tilde{\boldsymbol{\alpha}}_t^p)). \tag{15}$$

The question generation process is activated when the decrease in navigation progress estimation exceeds a predefined threshold ($\hat{\varphi}_{t-1} - \hat{\varphi}_t > 0.2$). This mechanism prioritizes cases where the UAV detects a decline in navigation progress after movement, indicating potential trajectory deviations that necessitate human intervention.

3.4 One-Trial Data Augmentation

Due to the scarcity of aerial VDN's training data, relying exclusively on limited human-annotated navigation data for imitation learning hinders the convergence of AerialVLA and compromises its self-correction ability. Meanwhile, the manually annotated data often fails to comprehensively explore the scene, resulting in insufficient utilization of environmental data during the training process. To address this, we designed OTDA to increase the available data during training by leveraging action predictions from the AerialVLA. After completing the sample from the full training set V at timestep t, the UAV executes the predicted action $\hat{a}_t = [\Delta\hat{x}_t, \Delta\hat{y}_t, \Delta\hat{z}_t]$ to transition to position $\hat{p}_{t+1}$. The optimal action $\hat{a}^*_{t+1}$ required at $\hat{p}_{t+1}$ to reach in the ground-truth trajectory is then computed. This dialogue-action-trajectory triplet is cached into the buffer V_g. At the start of each epoch, we build a new V by sampling from both the human-annotated dataset V_o and V_g at a predefined ratio.

Initially, OTDA introduces samples with only one-step deviation from the ground-truth trajectory in V. As training progresses, the repeated augmentation process gradually generates samples with multi-step deviations, thereby increasing the proportion of challenging examples that require the agent to rectify the route. The sampling strategy ensures that V_g retains a portion of the original or minimally perturbed data throughout training to prevent excessive deviation in data distribution.

4 Air Control Large Language Model

In the UNOD benchmark, we leverage the trajectories and dialogue-initiation instructions from the AVDH-Full unseen validation set as the foundation. During navigation, the AC-LLM answers the UAV agent's navigation-related queries in real time to replace the dialogue history in AVDH. This section details AC-LLM's architecture and training process.

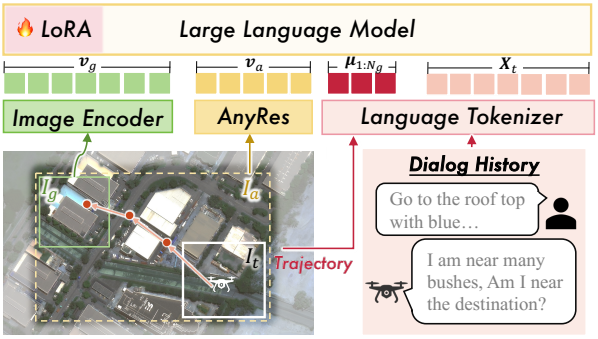

Figure 4: The architecture of AC-LLM. It integrates the UAV's current view, the destination's view, intermediate regions along the trajectory, and dialogue history to produce contextually appropriate navigational responses.

Model Architecture. AC-LLM utilizes two key inputs: the visual context around the destination and the planned route from the agent's current position to the target. First, we encode the visual observations at the goal location as $\boldsymbol{v}_g = f_v(I_g)$, where $\boldsymbol{v}_g$ captures the visual features of the destination's nearby field of view, providing landmark cues for target determination. To integrate trajectory information into AC-LLM, we first sample a feasible path toward the target. The path is visually represented by a BEV image I_a, which captures its bounding rectangle. I_a is processed through the AnyRes feature extractor $\hat{f}_v(\cdot)$:

$$\boldsymbol{v}_a = \phi(\texttt{flatten}(\hat{f}_v(I_a))). \tag{16}$$

To further encode the trajectory's geometric waypoints, we employ QwenVL's grounding-related special tokens for image-coordinate representation (Bai et al. 2023), converting the trajectory waypoints on I_a into a sequence of embeddings $\boldsymbol{\mu}_{1:N_g}$. This ensures precise trajectory conditioning for the LLM. Finally, AC-LLM fuses the multimodal inputs $[\boldsymbol{v}_g, \boldsymbol{v}_a, \boldsymbol{\mu}_{1:N_g}, \boldsymbol{X}_t]$ for response prediction, where $\boldsymbol{X}_t$ represents the prior dialogue history.

Training Process. We train AC-LLM using a pretraining-finetuning pipeline. In the pretraining stage, we leverage remote sensing image-language data from (Kuckreja et al. 2024), capitalizing on the similarity between the UAV's top-down perspective and high-resolution remote sensing imagery. This stage incorporates multiple vision-language alignment tasks, including image captioning, visual grounding, region-based caption generation, and referring expression comprehension. Through this process, AC-LLM learns to align UAV's observations with textual descriptions. Subsequently, we fine-tune the AC-LLM on the AVDH dataset (Fan et al. 2023) for dialogue-based instruction generation. This AVDH contains navigation trajectories paired with each question-answer in dialogues. For each training instance, we provide AC-LLM with the trajectory and the dialogue history preceding the final navigation instruction, training the LLM to predict the navigation guidance.

Model	ANDH											ANDH-Full										
	Seen Validation				Unseen Validation				Unseen Testing			Seen Validation				Unseen Validation				Unseen Testing		
	SPL	SR	GP	OSR	SPL	SR	GP	OSR	SPL	SR	GP	SPL	SR	GP	OSR	SPL	SR	GP	OSR	SPL	SR	GP
E.T.	10.3	12.7	51.2	25.4	17.1	19.9	52.3	34.8	11.8	12.4	55.2	3.8	4.5	46.7	10.6	4.6	5.1	43.3	13.1	3.8	4.2	50.0
HAA-T	13.9	15.6	55.1	24.6	16.8	19.9	56.3	30.7	12.5	15.2	54.3	7.1	8.6	58.3	12.2	6.5	7.0	64.0	12.2	4.3	4.6	53.4
LSTM	8.9	10.0	49.2	17.0	15.7	16.8	48.8	23.4	11.7	12.8	51.2	3.8	4.1	62.8	4.6	4.5	4.6	66.9	6.1	1.1	1.1	61.1
HAA-L	12.3	13.5	47.3	18.9	16.7	18.4	50.5	24.5	12.5	13.9	50.7	4.5	4.6	53.9	5.1	6.7	7.0	55.5	7.9	2.5	2.6	35.2
Ours	**25.4**	**29.1**	**66.4**	**43.7**	**26.6**	**30.9**	**68.1**	**47.4**	**24.1**	**28.3**	**66.7**	**20.5**	**29.4**	**77.6**	**39.1**	**19.2**	**28.0**	**78.4**	**41.1**	**17.1**	**25.2**	**82.7**

Table 1: Results on AVDH benchmark and AVDH-Full benchmark.

Index	OTDA	HSTF		Seen Validation				Unseen Validation			
		GM	STA	SPL	SR	GP	OSR	SPL	SR	GP	OSR
1				3.1	3.2	-10.0	14.8	5.5	6.0	-9.5	17.25
2	✓			13.0	22.9	50.9	32.1	17.3	26.5	51.2	34.2
3	✓	✓		21.1	27.8	58.6	45.1	19.5	25.3	60.70	45.26
4	✓	✓	✓	**25.4**	**29.2**	**66.4**	**43.8**	**26.6**	**30.9**	**68.2**	**47.5**

Table 2: Ablations on AVDH benchmark. Where GM indicates the utilization of the global map described in §3.2

5 Experiments

We evaluate AerialVLA against state-of-the-art methods on both AVDH (Fan et al. 2023) and UNOD benchmarks (§5.1). Additionally, we conduct ablation studies to analyze the contributions of its key components (§5.2). The visualisation results on UNOD are provided in §5.3.

5.1 Main Comparison

Comparison Baselines. We compare our approach with four baseline methods: the Episodic Transformer (E.T.), the Human Attention Aided Transformer (HAA-T), and the Human Attention Aided LSTM (HAA-L), all adopted from (Fan et al. 2023).

Results on AVDH. As shown in Table 1, AerialVLA achieves significant improvements over previous VDN methods for UAVs. On the AVDH (Fan et al. 2023) benchmark, AerialVLA increases the SPL metric from 17.1% to 26.6% on the unseen validation split and from 4.3% to 24.1% on the unseen testing split compared to the previous state-of-the-art HAA-Transformer (Fan et al. 2023), demonstrating superior trajectory efficiency and navigation success. The improvement becomes more pronounced on AVDH-Full, which has longer dialogue and trajectory. The SPL improves from 4.3% to 17.1% on the unseen testing split, indicating enhanced capability in historical context understanding, particularly for long-horizon navigation tasks.

Results on UNOD. As shown in Table 3, we evaluate AerialVLA's proactive dialogue-based navigation capability on UNOD. Since existing methods lack the interactive dialogue ability during navigation, we compare against baselines using pre-collected human questions from the AVDH-Full dataset and AC-LLM-generated responses to compare AerialVLA with proactive dialogue. While AC-LLM-generated dialogues degrade navigation performance compared to human-annotated dialogues, AerialVLA with PNaQ-based online VDN achieves a significant improvement, increasing SPL by 7.9% over AerialVLA with pre-recorded dialogues and even surpassing AerialVLA with human-annotated dialogue history by 2.4% on the validation unseen split. These results demonstrate the critical role of context-aware proactive questioning in navigation and highlight the necessity of an online benchmark for Aerial VDN.

5.2 Ablation Study

In this section, we conduct ablation studies on the key components of the AerialVLA model over AVDH to validate the effectiveness of the proposed modules, as summarized in Table 2. Additional ablation experiments for OTDA and STA are comprehensively verified in Table 4 and Table 5. The baseline implementation (the # 1 of Table 2) processes the navigation history observations with the video feature extractor of LLaVA-OneVision (Li et al. 2024) for action prediction through the same action decoder of AerialVLA.

About OTDA. We validate the effectiveness of OTDA. Compared with the baseline model trained solely on human-annotated trajectories in Table 2 #1, OTDA improves SPL by 9.9% and SR by 19.7% on the seen validation set. A notable enhancement of 11.8% in SPL is also observed on the unseen validation set. These results demonstrate that OTDA significantly enhance navigation performance by enriching trajectory diversity. Furthermore, Table 4 examines the impact of the ratio between original and generated data per epoch on navigation performance. The model achieves optimal performance when the ratio of original to generated data is 1:2, while further increasing the proportion of generated data leads to performance degradation. This indicates that an excessive amount of generated data may interfere with the model's ability to learn optimal paths.

About Global Map. We evaluate the effectiveness of the global map feature $\boldsymbol{H}_t^s$, which aggregates historical observa-

Models	Seen Validation				Unseen Validation			
	SPL	SR	GP	OSR	SPL	SR	GP	OSR
HAA-T†	3.9	5.1	10.7	9.6	6.4	8.4	34.2	14.5
HAA-L†	2.9	3.1	16.3	4.6	5.2	4.6	18.4	7.0
AerialVLA†	17.3	24.3	59.9	35.0	13.7	20.0	71.2	30.4
AerialVLA	**22.6**	**26.9**	**99.3**	**40.1**	**21.6**	**26.2**	**101.8**	**37.3**

Table 3: Results on UNOD. †indicates models utilizing pre-collected human questions and AC-LLM responses.

V_o:V_g	Seen Validation				Unseen Validation			
	SPL	SR	GP	OSR	SPL	SR	GP	OSR
1:1	15.4	21.6	61.7	34.3	14.4	20.9	57.8	32.1
1:2	**25.5**	**29.2**	**66.4**	**43.8**	**26.6**	30.9	**68.2**	47.5
1:3	20.1	28.9	65.8	42.0	20.5	**31.6**	67.0	**48.3**

Table 4: Abation studies on AVDH benchmark about OTDA for the data ratio of V_o to V_g in §3.4.

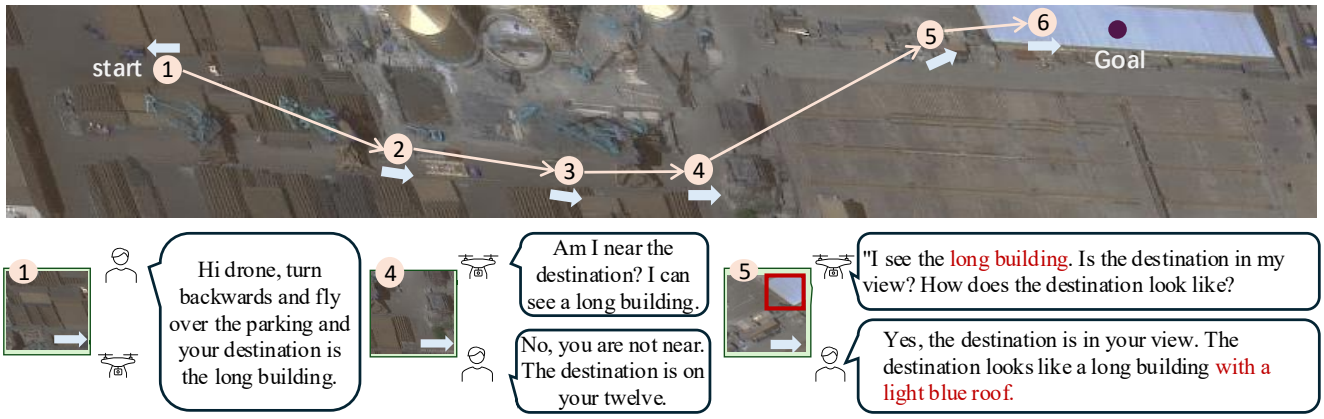

Figure 5: Interaction of AerialVLA and AC-LLM visualization on UNOD. Blue arrows indicate the drone's forward orientation. The dialogue content's adaptive flexibility while maintaining alignment with the UAV's observational states.

STA Res	Seen Validation				Unseen Validation			
	SPL	SR	GP	OSR	SPL	SR	GP	OSR
1, 1, 1	21.3	27.9	64.7	42.3	23.0	30.0	62.8	44.3
1, 2, 3	**25.4**	**29.2**	**66.4**	**43.7**	**26.6**	**30.9**	**68.2**	**47.4**
1, 3, 5	22.4	28.5	65.1	41.2	24.0	29.7	57.4	43.3

Table 5: Abation studies on AVDH benchmark about STA for the value of $d_{1,2,3}$ in §3.2.

tions to a global map. In #3 of Table 2, we replace the video processing module in #1 with the global map. Compared to #2, the model with global map achieves significant improvements, with 8.1% and 2.2% SPL gains on seen and unseen validation sets, respectively. HSG effectively provides spatial layout relationships of observations to the LLM, which enhances the model's directional awareness and substantially improves AerialVLA's navigation efficiency.

About STA. Compared to Table 2 #3, the model with STA in #4 demonstrates significant improvements in both SR and SPL. On the seen validation set, SPL increases by 4.3% and SR by 1.4%, while achieved enhancements of 7.1% in SPL and 5.6% in SR are observed on the unseen validation set. These results suggest that the traversal order information provides a benefit for navigation in unseen environments.

About Spatial Resolution of STA. As shown in Table 5, we investigate the impact of different resolution combinations in STA's attention mechanism over the global spatial map on model accuracy. We observe that the model achieves higher accuracy when resolutions are set to $d_1 = 1$, $d_2 = 2$, and $d_3 = 3$. Higher resolutions introduce excessive redundant details in aggregated features, leading to performance degradation, while lower resolutions fail to yield significant accuracy improvements due to insufficient correlation between the temporal and spatial features of navigation history.

5.3 Visualization Results

As illustrated in Fig. 5, we visualize the navigation trajectories of AC-LLM and AerialVLA on UNOD validation set. Along this path, AerialVLA poses queries at positions 4 and 5. This example demonstrates that AC-LLM can effectively provide accurate responses to the navigation model's inquiries.

6 Conclusion

We present AerialVLA, integrating proactive dialogue and spatial-temporal historical states for VDN navigation. The proactive dialogue ability is crucial for target localization, as it enables the agent to acquire trajectory-aligned instructions from human. The effectiveness of HSTF indicates that spatial-temporal reasoning over historical observations is critical for enhancing the capability of UAV's VDN. OTDA demonstrates that trajectory diversity also plays a vital role in enabling LLMs to accomplish navigation tasks effectively for improving the error correction ability.

Acknowledgements

This research is supported in part by National Key R&D Program of China (2022ZD0115502), National Natural Science Foundation of China (No. 62461160308, No. 62576024, U23B2010), "the Fundamental Research Funds for the Central Universities" (No. 501RCQD2025), "Pioneer" and "Leading Goose" R&D Program of Zhejiang (No. 2024C01161), Beijing Natural Science Foundation (QY25227), Ningbo Science and Technology Innovation 2025 Major Project (2025Z034), NSFCRGC Project (N CUHK498/24).

References

Anderson, P.; Wu, Q.; Teney, D.; Bruce, J.; Johnson, M.; Sünderhauf, N.; Reid, I.; Gould, S.; and Van Den Hengel, A. 2018. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *CVPR*.

Bai, J.; Bai, S.; Chu, Y.; Cui, Z.; Dang, K.; Deng, X.; Fan, Y.; Ge, W.; Han, Y.; Huang, F.; Hui, B.; Ji, L.; Li, M.; Lin, J.; Lin, R.; Liu, D.; Liu, G.; Lu, C.; Lu, K.; Ma, J.; Men, R.; Ren, X.; Ren, X.; Tan, C.; Tan, S.; Tu, J.; Wang, P.; Wang, S.; Wang, W.; Wu, S.; Xu, B.; Xu, J.; Yang, A.; Yang, H.; Yang, J.; Yang, S.; Yao, Y.; Yu, B.; Yuan, H.; Yuan, Z.; Zhang, J.; Zhang, X.; Zhang, Y.; Zhang, Z.; Zhou, C.; Zhou, J.; Zhou, X.; and Zhu, T. 2023. Qwen Technical Report. *arXiv preprint arXiv:2309.16609*.

Bozcan, I.; and Kayacan, E. 2020. AU-AIR: A Multi-modal Unmanned Aerial Vehicle Dataset for Low Altitude Traffic Surveillance. *arXiv preprint*.

Chang, A.; Dai, A.; Funkhouser, T.; Halber, M.; Niessner, M.; Savva, M.; Song, S.; Zeng, A.; and Zhang, Y. 2017. Matterport3D: Learning from RGB-D Data in Indoor Environments. *3DV*.

Chen, H.; Suhr, A.; Misra, D.; Snavely, N.; and Artzi, Y. 2019. Touchdown: Natural language navigation and spatial reasoning in visual street environments. In *CVPR*.

Chen, S.; Guhur, P.-L.; Schmid, C.; and Laptev, I. 2021. History aware multimodal transformer for vision-and-language navigation. *NeurIPS*.

Fan, Y.; Chen, W.; Jiang, T.; Zhou, C.; Zhang, Y.; and Wang, X. E. 2023. Aerial Vision-and-Dialog Navigation. In *Findings of the Association for Computational Linguistics: ACL 2023*, 3043–3061. Toronto, Canada: Association for Computational Linguistics.

Fan, Y.; Chu, S.; Zhang, W.; Song, R.; and Li, Y. 2020. Learn by observation: Imitation learning for drone patrolling from videos of a human navigator. In *IROS*.

Gao, Y.; Li, C.; You, Z.; Liu, J.; Li, Z.; Chen, P.; Chen, Q.; Tang, Z.; Wang, L.; Yang, P.; et al. 2025. OpenFly: A versatile toolchain and large-scale benchmark for aerial vision-language navigation. *arXiv e-prints*, arXiv–2502.

Giusti, A.; Guzzi, J.; Cireşan, D. C.; He, F.-L.; Rodríguez, J. P.; Fontana, F.; Faessler, M.; Forster, C.; Schmidhuber, J.; Di Caro, G.; et al. 2015. A machine learning approach to visual perception of forest trails for mobile robots. *IEEE RAL*.

Guo, Z.; Xu, R.; Yao, Y.; Cui, J.; Ni, Z.; Ge, C.; Chua, T.-S.; Liu, Z.; and Huang, G. 2025. Llava-uhd: an lmm perceiving any aspect ratio and high-resolution images. In *ECCV*.

Hao, W.; Li, C.; Li, X.; Carin, L.; and Gao, J. 2020. Towards learning a generic agent for vision-and-language navigation via pre-training. In *CVPR*.

Jain, V.; Magalhaes, G.; Ku, A.; Vaswani, A.; Ie, E.; and Baldridge, J. 2019. Stay on the path: Instruction fidelity in vision-and-language navigation. *arXiv preprint*.

Kang, K.; Belkhale, S.; Kahn, G.; Abbeel, P.; and Levine, S. 2019. Generalization through simulation: Integrating simulated and real data into deep reinforcement learning for vision-based autonomous flight. *arXiv preprint*.

Krantz, J.; Wijmans, E.; Majumdar, A.; Batra, D.; and Lee, S. 2020. Beyond the nav-graph: Vision-and-language navigation in continuous environments. In *ECCV*.

Ku, A.; Anderson, P.; Patel, R.; Ie, E.; and Baldridge, J. 2020. Room-Across-Room: Multilingual Vision-and-Language Navigation with Dense Spatiotemporal Grounding. In *EMNLP*.

Kuckreja, K.; Danish, M. S.; Naseer, M.; Das, A.; Khan, S.; and Khan, F. S. 2024. GeoChat: Grounded Large Vision-Language Model for Remote Sensing. *CVPR*.

Lee, J.; Miyanishi, T.; Kurita, S.; Sakamoto, K.; Azuma, D.; Matsuo, Y.; and Inoue, N. 2024. CityNav: Language-Goal Aerial Navigation Dataset with Geographic Information. *arXiv preprint*.

Li, B.; Zhang, Y.; Guo, D.; Zhang, R.; Li, F.; Zhang, H.; Zhang, K.; Li, Y.; Liu, Z.; and Li, C. 2024. LLaVA-OneVision: Easy Visual Task Transfer. *arXiv preprint arXiv:2408.03326*.

Liu, S.; Zhang, H.; Qi, Y.; Wang, P.; Zhang, Y.; and Wu, Q. 2023. Aerialvln: Vision-and-language navigation for uavs. In *ICCV*.

Liu, Y.; Yao, F.; Yue, Y.; Xu, G.; Sun, X.; and Fu, K. 2024. NavAgent: Multi-scale Urban Street View Fusion For UAV Embodied Vision-and-Language Navigation. *arXiv preprint*.

Loquercio, A.; Maqueda, A. I.; Del-Blanco, C. R.; and Scaramuzza, D. 2018. Dronet: Learning to fly by driving. *IEEE RAL*, 3(2): 1088–1095.

Majdik, A. L.; Till, C.; and Scaramuzza, D. 2017. The Zurich urban micro aerial vehicle dataset. *The IJRR*.

Nguyen, K.; and Daumé III, H. 2019. Help, Anna! Visual Navigation with Natural Multimodal Assistance via Retrospective Curiosity-Encouraging Imitation Learning. In *EMNLP*.

Qi, Y.; Wu, Q.; Anderson, P.; Wang, X.; Wang, W. Y.; Shen, C.; and Hengel, A. v. d. 2020. Reverie: Remote embodied visual referring expression in real indoor environments. In *CVPR*.

Qiao, Y.; Qi, Y.; Hong, Y.; Yu, Z.; Wang, P.; and Wu, Q. 2022. HOP: History-and-Order Aware Pre-training for Vision-and-Language Navigation. In *CVPR*.

Roman, H. R.; Bisk, Y.; Thomason, J.; Celikyilmaz, A.; and Gao, J. 2020. RMM: A Recursive Mental Model for Dialogue Navigation. In *EMNLP*.

Singla, A.; Padakandla, S.; and Bhatnagar, S. 2019. Memory-based deep reinforcement learning for obstacle avoidance in UAV with limited environment knowledge. *IEEE TIST*.

Smolyanskiy, N.; Kamenev, A.; Smith, J.; and Birchfield, S. 2017. Toward low-flying autonomous MAV trail navigation using deep neural networks for environmental awareness. In *IROS*.

Thomason, J.; Murray, M.; Cakmak, M.; and Zettlemoyer, L. 2020. Vision-and-dialog navigation. In *CoRL*.

Wang, X.; Yang, D.; Wang, Z.; Kwan, H.; Chen, J.; Wu, W.; Li, H.; Liao, Y.; and Liu, S. 2024a. Towards realistic uav vision-language navigation: Platform, benchmark, and methodology. *arXiv preprint arXiv:2410.07087*.

Wang, X.; Yang, D.; Wang, Z.; Kwan, H.; Chen, J.; Wu, W.; Li, H.; Liao, Y.; and Liu, S. 2024b. Towards Realistic UAV Vision-Language Navigation: Platform, Benchmark, and Methodology.

Xu, Z.; Han, X.; Shen, H.; Jin, H.; and Shimada, K. 2025. Navrl: Learning safe flight in dynamic environments. *IEEE RAL*.

Zheng, D.; Huang, S.; Zhao, L.; Zhong, Y.; and Wang, L. 2024. Towards learning a generalist model for embodied navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13624–13634.

Zhu*, Y.; Weng*, Y.; Zhu, F.; Liang, X.; Ye, Q.; Lu, Y.; and jiao, J. 2021. Self-Motivated Communication Agent for Real-World Vision-Dialog Navigation. In *ICCV*.