HIERARCHICAL VOLUMETRIC REASONING FOR AERIAL VISION-LANGUAGE NAVIGATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Aerial vision-language navigation (AVLN) requires unmanned aerial vehicles (UAVs) to follow natural-language instructions and reach specified destinations in large-scale, continuous 3D environments. Navigation-relevant cues are often distributed over long distances and observed at varying spatial scales, placing substantial demands on an agent's capacity for 3D spatial understanding and reasoning. This work introduces Hierarchical Volumetric Reasoning for AVLN, enabling the agent to reason over 3D scene structure from global layouts to fine-grained local regions. The reasoning is grounded in a Multi-Scale Voxel Map that organizes the observed environment into a coarse-to-fine 3D volume hierarchy and encodes geometric and semantic cues at each level. At each level, an instruction-conditioned decision token aggregates relevant voxel features, while its attention scores route subsequent reasoning to finer spatial regions. The final decision token is combined with navigation history to predict the next action. Experiments on AerialVLN-s show that our agent achieves 16.2% SR and 12.5% SDTW on seen environments, and 8.8% SR and 3.9% SDTW on unseen environments, with an inference latency of 0.09s per navigation step. Extensive ablation studies further validate the effectiveness of our core designs. The code will be released.

1 INTRODUCTION

Aerial vision-language navigation (AVLN) studies how UAV agents interpret natural language instructions and execute actions toward specified destinations in large-scale 3D environments (Liu et al., 2023; Wang et al., 2025). The episodes often extend over long distances, with instruction-relevant landmarks and spatial cues appearing at substantially different scales under changing viewpoints (Lee et al., 2025; Fan et al., 2023). Reliable navigation therefore depends critically on the agent's ability to understand and reason about the surrounding 3D space (Yao et al., 2024; Zhao et al., 2026).

Early aerial agents (Liu et al., 2023; Misra et al., 2018) perform navigation reasoning implicitly within recurrent cross-modal policies, which directly predict actions from visual observations and language instructions. Subsequent agents strengthen this process by incorporating longer observation histories (Gao et al., 2026b; Jiang et al., 2025), multi-view information (Xu et al., 2026b; Tang et al., 2026), explicit target grounding (Wang et al., 2025; Qiao et al., 2025), and collaborative observations (Wu et al., 2025), thereby enabling reasoning over richer perceptual context. Map-based agents further externalize spatial reasoning through spatial grids (Ding et al., 2026; Fan et al., 2023; Lee et al., 2025; Chu et al., 2024; Fan et al., 2026) and topological graphs (Ning et al., 2026; Xu et al., 2025; Gao et al., 2024), which organize accumulated observations into persistent representations for localization, planning, and action selection. More recent aerial agents employ LLMs or MLLMs to conduct explicit spatial reasoning over structured scene representations, including semantic maps, scene graphs, and 3D-aware memories (Zhou et al., 2026a; Lin et al., 2024; Zhang et al., 2025). However, such reasoning often operates over globally aggregated scene context, introducing substantial redundancy and diluting spatial cues relevant to the current action (Chen et al., 2026b). Moreover, this spatial context is typically represented at a fixed granularity, limiting the agent's ability to capture geometric and semantic evidence that vary in scale Zhang et al. (2026).

This work introduces a UAV agent that lifts observations to a unified 3D volumetric space, in which it perceives the environment at multiple spatial granularities and reasons over the same volumetric hierarchy from global spatial layout to local semantic cues for spatially informed decision making

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

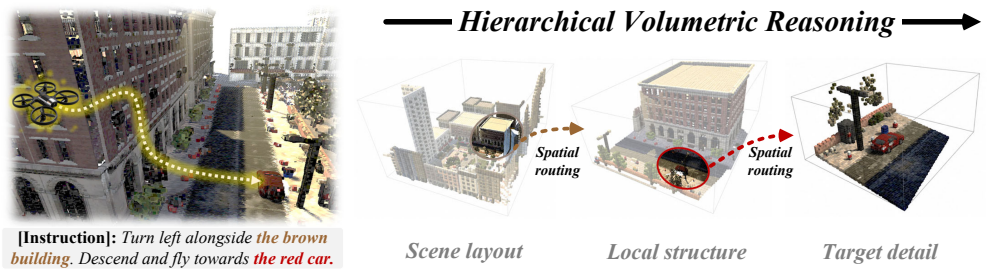

Figure 1: **Main idea.** Guided by the language instruction, our agent performs Hierarchical Volumetric Reasoning over a 3D volume hierarchy, progressively gathering spatial evidence from the overall scene layout to local structures and target details. At each scale, the agent employs a spatial routing strategy to determine relevant regions for finer reasoning. The example illustrates this process by focusing first on the brown building and then on the red car and its surroundings.

(Fig.1). Our agent first constructs a *Multi-Scale Voxel Map* (§3.1) that organizes the observed 3D environment as a coarse-to-fine volume hierarchy, which encodes multi-scale geometric and semantic features through 3D semantic occupancy prediction. Based on this map, the agent performs *Hierarchical Volumetric Reasoning* (§3.2) to progressively gather task-relevant spatial evidence during navigation. It conditions a decision token and auxiliary proxy tokens on the instruction and aggregates informative voxel features from coarse to fine. At each reasoning level, the intermediate attention score serves as a spatial routing signal to guide the agent to further explore finer, focused spatial regions. Finally, the resulting decision token that captures the navigation intent, is combined with a recurrent unit that summarizes historical context to predict the next navigation action (§3.3).

Our agent is evaluated on AerialVLN-s (Liu et al., 2023) (§4.2). On *val seen* split, it exceeds the strongest learning-based agent by $\mathbf{8.1\%}$ for SR/$\mathbf{10.1\%}$ for SDTW, and LLM-based agent by $\mathbf{1.5\%}$ for SR/$\mathbf{7.1\%}$ for SDTW. On *val unseen* split, it improves SR by $\mathbf{3.7\%}$ and SDTW by $\mathbf{2.5\%}$ over non-LLM-based agents, and outperforms the LLM-based OpenFly (Gao et al., 2026b) by $\mathbf{1.2\%}$ for SR and $\mathbf{2.4\%}$ for SDTW. Crucially, our agent strikes a favorable balance between navigation performance and computation efficiency, with a real-time inference latency of $\mathbf{0.09}s$ per step. Extensive ablation studies validate the effectiveness of our designs (§4.4,§4.5). The code will be released.

2 Related Work

Aerial Vision-Language Navigation (AVLN). As a pioneering work, AerialVLN (Liu et al., 2023) introduces the first large-scale aerial navigation benchmark and employs sequence-to-sequence model for action prediction. OpenFly (Gao et al., 2026b) further improves data scalability through an automatic instruction–trajectory generation pipeline and introduces a key-frame aware agent with token merging. UAV-Flow (Wang et al., 2026b), in contrast, emphasizes short-range and fine-grained language-conditioned flight control. Beyond trajectory following, another line of work formulates AVLN as geospatial grounding over top-down satellite imagery, where geographic priors and map-level context are incorporated to localize and navigate toward language-specified destinations (Fan et al., 2023; Lee et al., 2025; Chu et al., 2024; Fan et al., 2026). Recent studies consider more challenging object-goal navigation, in which UAVs must discover and approach target objects from high-level instructions with limited route-level guidance (Xiao et al., 2025; Wang et al., 2025; Ji et al., 2026; Zhao et al., 2026; Wang et al., 2026a). Despite these advances, many aerial agents still reason from **2D visual observations** or compressed spatial abstractions such as grid maps and topological graphs, which only partially preserve the geometry and fine-grained semantics of the surrounding 3D environment. In contrast, our agent constructs a **3D volumetric map** that organizes geometric and semantic cues across multiple spatial granularities. Hierarchical Volumetric Reasoning then operates directly on this 3D hierarchy for spatially grounded action prediction.

Map Building in Navigation. Building an effective scene map is prerequisite for embodied agents to perceive, reason, plan, and navigate in physical world. In robot navigation, classic SLAM system

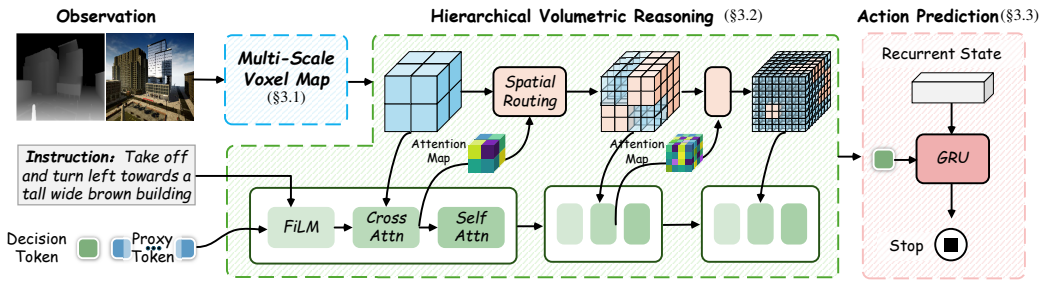

Figure 2: **Pipeline Overview**. At each step, our agent construct a *Multi-Scale Voxel Map* (§3.1) from its egocentric observation. Within this map, it performs *Hierarchical Volumetric Reasoning* (§3.2) that progressively infers instruction-relevant spatial cues to inform decision-making. Finally, the predictive decision token combines with a historical recurrent state for action prediction (§3.3).

builds metric maps from onboard sensors such as LiDAR and depth cameras (Chen et al., 2019b; Ramakrishnan et al., 2021; Thrun, 2002; 1998). These maps enable precise localization and obstacle avoidance, but are largely detached from semantic understanding. To bridge this gap, recent VLN methods adopt learning-based neural representations as scene maps. Semantic maps (Chen et al., 2024; Cartillier et al., 2021; Chaplot et al., 2020; Henriques & Vedaldi, 2018) enrich the environment with object- and region-level labels from top-down view, while topological maps (Chen et al., 2021; Wald et al., 2020; Chen et al., 2022) abstract the scene as a graph of navigable keypoints and their connectivity to support spatial reasoning and efficient path planning. Other efforts explore implicit neural scene representations (*i.e.*, NeRF (Mildenhall et al., 2021)) to embed visual information as latent codes for localization and planning (Kwon et al., 2023; Adamkiewicz et al., 2022). Recently, explicit 3D scene representations have been explored to support geometry-aware navigation. (Gao et al., 2025; 2026a) represent the environment as 3D Gaussian primitives with attached semantics and uncertainties. (Liu et al., 2024) learns a volumetric scene representation and estimates state transition probability of each voxel cell for action prediction. However, such an aggressive searching process incurs **substantial computational overhead**, which is prohibitive in large-scale aerial environments. In contrast, our agent supports **efficient 3D volumetric reasoning** through a multi-scale voxel map and routing reasoning toward task-relevant local regions in a coarse-to-fine manner.

Spatial Reasoning for AVLN. In AVLN, successful navigation requires agents to continuously understand the spatial arrangement of objects and regions, and align instruction semantics with navigation actions grounded in the physical space, which fundamentally relies on spatial reasoning (Chen et al., 2019a; Sautenkov et al., 2025; Schumann et al., 2024; Zhang et al., 2021; Chu et al., 2024). To succeed, recent endeavors increasingly leverage large language models (LLMs) and vision-language models (VLMs) to enhance agents' reasoning ability (Xia et al., 2026; Chen et al., 2026a; Tian et al., 2025). A natural approach is to elicit explicit reasoning process before action prediction through supervised fine-tuning on chain-of-thought (CoT) demonstrations (Xu et al., 2026a; Lykov et al., 2025; Wang et al., 2026a; Zhou et al., 2026b; Zheng et al., 2026; Cai et al., 2025). However, acquiring high-quality CoT reasoning annotations at scale is challenging, which limits their scalability. In parallel, recent works construct and feed LLMs **verbalized spatial abstractions**, including semantic maps (Gao et al., 2024), topological graph (Ning et al., 2026), or a combination of both (Xu et al., 2025; Zhang et al., 2025), to promote agents' spatial reasoning. Nevertheless, they are prone to hallucination and ambiguity in both perception and reasoning stages (Chakraborty et al., 2025; Xu et al., 2026c). In contrast, our agent grounds spatial reasoning directly in **3D volumetric space**, which well preserves geometric and semantic cues for spatial grounded action prediction.

3 METHOD

Problem Formulation. AVLN requires a UAV agent to traverse a continuous 3D open-ended environment guided by natural language instruction $\mathcal{X}$ to reach the target. At each navigation step t, the agent receives egocentric observation $\mathcal{O}_t = \{I_t, D_t\}$, where I_t denotes RGB image, and D_t is the corresponding depth image. The goal is to learn a policy that selects the

next step action a_t from action space $\mathcal{A}$, which consists of several low-level control primitives $\{\texttt{move forward/left/right/up/down}, \texttt{turn left/right}, \texttt{stop}\}$ following conventions (Liu et al., 2023). An episode terminates if the agent selects `stop` or reach the maximum steps.

Overview (Fig. 2). At timestep t, our agent first constructs a *Multi-Scale Voxel Map (MSVM)* (§3.1) from its egocentric observation to perceive multi-scale geometry and semantics of the environment. Grounded in MSVM, it conducts *Hierarchical Volumetric Reasoning (HVR)* (§3.2) , which explicitly searches for and gathers instruction-relevant evidence from global scene layouts to fine-grained local structures, forming an internal decision state that captures the agent's navigation intent. The resulting state is finally combined with navigation history for predicting the next action (§3.3).

3.1 Multi-Scale Voxel Map

Aerial navigation unfolds in large-scale, continuous 3D spaces, where the agent must perceive the geometric structures and semantics of its surroundings to determine the next move, with task-relevant cues spanning from global layout to local details. To this end, our agent lifts 2D RGB-D observations to 3D space by constructing a Multi-Scale Voxel Map, which voxelizes the environment to regular 3D cells at different granularities, thereby enabling holistic scene understanding.

Volume Construction. At each navigation step, the agent encodes the current RGB image $\boldsymbol{I}$ into a visual feature $\boldsymbol{F}^{2d} \in \mathbb{R}^{C\times H\times W}$, where C denotes the feature dimension and H and W denote the spatial dimensions. To represent the observed 3D space, the agent discretizes it into N voxels, each represented by a feature $\boldsymbol{V}_i \in \mathbb{R}^C$, forming a voxel volume $\mathcal{V} = \{\boldsymbol{V}_i\}_{i=1}^N$. To capture scene structure at multiple spatial granularities, the agent progressively constructs a coarse-to-fine Multi-Scale Voxel Map $\{\mathcal{V}_l\}_{l=1}^L$ from a learnable coarse volume. At each stage, the current volume is first upsampled to a finer spatial resolution, yielding $\widetilde{\mathcal{V}}_l$, and the occupancy score are predicted for the upsampled voxels. Given the inherent sparsity of aerial environment, only the top-$k\%$ voxels with highest occupancy probability are retained for further refinement, formulated as:

$$\mathcal{V}_l = \widetilde{\mathcal{V}}_l + \mathcal{F}(\texttt{Top}_{k\%}(\widetilde{\mathcal{V}}_l), \boldsymbol{F}^{2d}, D), \tag{1}$$

where $\mathcal{F}(\cdot)$ denotes the voxel refinement module described below. It allows the agent to allocate computation on informative regions, progressively capturing finer geometric and semantic details

Voxel Refinement. The refinement module $\mathcal{F}(\cdot)$ updates each selected voxel by aggregating geometric and semantic cues from the current RGB-D observation. For a selected voxel $\boldsymbol{V}$ at 3D position $\boldsymbol{p} \in \mathbb{R}^3$, the agent first projects its position into the current camera view:

$$\boldsymbol{r} = (u, v, d)^\top = \boldsymbol{K}\boldsymbol{E}\,[\boldsymbol{p}, 1]^\top, \tag{2}$$

where $\boldsymbol{K} \in \mathbb{R}^{3\times3}$ and $\boldsymbol{E} \in \mathbb{R}^{3\times4}$ denote the camera intrinsic and extrinsic matrices, respectively, while u and v denote the projected image coordinates and d denotes depth.

Directly sampling $\boldsymbol{F}^{2d}$ at the projected image location collapses the depth dimension, which weakens the correspondence between visual features and 3D geometry. The agent therefore constructs a depth-aware 3D feature volume $\boldsymbol{F}^{3d} = \hat{D} \otimes \boldsymbol{F}^{2d} \in \mathbb{R}^{C\times H\times W\times B}$ by taking the outer product between $\boldsymbol{F}^{2d}$ and discretized depth map $\hat{D}$ with B bins. Using $\boldsymbol{r}$ as the reference, 3D deformable cross-attention (Li et al., 2023) aggregates features from neighboring locations in $\boldsymbol{F}^{3d}$:

$$\mathcal{F}\big(\boldsymbol{V}, \boldsymbol{F}^{2d}, D\big) = \sum\nolimits_{m=1}^{M} A_m \boldsymbol{W}\, \phi(\boldsymbol{F}^{3d}, \boldsymbol{r} + \Delta\boldsymbol{r}_m), \tag{3}$$

where M denotes the number of sampled points, A_m and $\Delta\boldsymbol{r}_m \in \mathbb{R}^3$ denote the attention weight and 3D offset of the m-th sampling point, $\boldsymbol{W}$ is a learnable feature projection, and $\phi(\cdot)$ denotes trilinear interpolation. Finally, a stack of 3D convolutions are applied to the refined volume for aggregating information among neighboring voxels and preserving local spatial consistency.

3.2 Hierarchical Volumetric Reasoning

The Multi-Scale Voxel Map provides the agent with 3D geometric and semantic cues at multiple spatial granularities. However, only a subset of these cues is relevant to the current navigation decision, and their importance may vary across spatial scales. Hierarchical Volumetric Reasoning

therefore progressively gathers instruction-relevant evidence from coarse scene structure to finer local regions through volumetric grounding and spatial routing.

Volumetric Grounding. At each reasoning stage, the agent aligns the language instruction with the current volumetric features to gather spatial evidence relevant to the current decision. To aggregate this evidence, it maintains a decision token $\boldsymbol{Q} \in \mathbb{R}^C$ and J proxy tokens $\boldsymbol{Z} \in \mathbb{R}^{J\times C}$. The decision token serves as the readout for subsequent action prediction, while the proxy tokens provide additional latent slots for capturing complementary volumetric context.

Before interacting with the voxel features, these tokens are conditioned on the encoded instruction feature $\boldsymbol{X} \in \mathbb{R}^C$ through FiLM (Perez et al., 2018), yielding $\widehat{\boldsymbol{Q}} \in \mathbb{R}^C$ and $\widehat{\boldsymbol{Z}} \in \mathbb{R}^{J\times C}$:

$$[\widehat{\boldsymbol{Q}}; \widehat{\boldsymbol{Z}}] = \big(1 + \gamma(\boldsymbol{X})\big) \odot [\boldsymbol{Q}; \boldsymbol{Z}] + \beta(\boldsymbol{X}), \tag{4}$$

where $\odot$ denotes element-wise multiplication, $\gamma(\cdot)$ and $\beta(\cdot)$ are projection layers to produce feature-wise scaling and shifting parameters. The conditioned tokens then query G candidate voxel features $\boldsymbol{U} \in \mathbb{R}^{G\times C}$, which are augmented with a learnable scale embedding $\boldsymbol{E}^{\text{scale}} \in \mathbb{R}^C$ and 3D positional embeddings $\boldsymbol{E}^{\text{pos}} \in \mathbb{R}^{G\times C}$ to provide scale and spatial priors:

$$[\boldsymbol{Q}'; \boldsymbol{Z}'], \boldsymbol{L} = \texttt{MLT}([\widehat{\boldsymbol{Q}}; \widehat{\boldsymbol{Z}}], \boldsymbol{U} + \boldsymbol{E}^{\text{scale}} + \boldsymbol{E}^{\text{pos}}), \tag{5}$$

where $\texttt{MLT}$ denotes multi-layer transformer comprised of cross-attention and self-attention, and $\boldsymbol{L} \in \mathbb{R}^{(J+1)\times G}$ is the resulting token-to-voxel attention map. The updated decision token $\boldsymbol{Q}' \in \mathbb{R}^C$ summarizes the instruction-relevant evidence gathered at the current scale, while $\boldsymbol{Z}' \in \mathbb{R}^{J\times C}$ retains complementary volumetric context. The decision-token scores in $\boldsymbol{L}$ provide the spatial relevance signal for subsequent routing.

Spatial Routing. Reasoning over the entire volumetric space at finer levels introduces substantial redundant computation, since the candidate voxels grow rapidly while most of them may contribute little to the current action. We therefore use the decision attention at the current scale to route subsequent reasoning toward relevant regions at the next finer scale.

At the coarsest scale, all voxels in $\mathcal{V}_1$ serve as candidates. At scale $l < L$, let $\boldsymbol{L}_l^Q \in \mathbb{R}^{G_l}$ denote the decision-token scores extracted from the token-to-voxel attention map $\boldsymbol{L}_l$, where G_l is the number of candidate voxels at the current scale. The agent selects the parent voxels with the highest scores and expands only their children in the next finer volume:

$$\boldsymbol{U}_{l+1} = \mathcal{V}_{l+1}\Big[\bigcup\nolimits_{i\in\texttt{TopK}\left(\boldsymbol{L}_l^Q\right)} \texttt{Child}(i)\Big], \tag{6}$$

where $\texttt{TopK}(\cdot)$ selects the highest-scoring parent voxels, $\texttt{Child}(i)$ returns the child voxels of parent i at the next scale, and $\boldsymbol{U}_{l+1} \in \mathbb{R}^{G_{l+1}\times C}$ collects the routed candidate features for subsequent reasoning. The updated decision and proxy tokens $\boldsymbol{Q}'$ and $\boldsymbol{Z}'$ serve as the input tokens at the next scale together with $\boldsymbol{U}_{l+1}$. Repeating volumetric grounding and spatial routing across scales progressively focuses reasoning from coarse scene structure to finer regions relevant to the current decision. The decision token obtained at the finest scale is subsequently used for action prediction.

3.3 Action Prediction

After hierarchical reasoning, the final decision token $\boldsymbol{Q}_t$ summarizes the instruction-conditioned geometric and semantic evidence gathered from the current 3D scene. To complement this volumetric representation with appearance information from the current view, we concatenate it with the pooled 2D visual feature $\bar{\boldsymbol{F}}^{2d}$. A gated recurrent unit (GRU) then combines the decision token and pooled visual feature with the navigation history encoded by the previous hidden state $\boldsymbol{h}_{t-1}$:

$$\boldsymbol{h}_t = \texttt{GRU}([\boldsymbol{Q}_t; \bar{\boldsymbol{F}}^{2d}], \boldsymbol{h}_{t-1}), \tag{7}$$

where $\boldsymbol{h}_t \in \mathbb{R}^C$ denotes the recurrent hidden state, and $[;]$ is concatenation operation, The updated hidden state is finally mapped to an action probability distribution $\boldsymbol{y}_t \in [0,1]^{|\mathcal{A}|}$ via MLP:

$$\boldsymbol{y}_t = \texttt{Softmax}(\texttt{MLP}(\boldsymbol{h}_t)), \tag{8}$$

from which the next action a_t is obtained by selecting the most probable action in $\mathcal{A}$.

3.4 Loss Function

MSVM Loss. To capture both geometry and semantics, MSVM is optimized with multi-scale supervision for 3D occupancy prediction. Specifically, the intermediate occupancy probabilities are supervised with binary cross-entropy loss $\mathcal{L}_{\text{bce}}$, while the finest-level semantic occupancy output is jointly supervised by weighted cross-entropy loss $\mathcal{L}_{\text{ce}}$ (Roldao et al., 2020), lovász-softmax loss $\mathcal{L}_{\text{ls}}$ (Berman et al., 2018), and scene-class affinity losses $\mathcal{L}_{\text{scal}}^{\text{geo}}$ and $\mathcal{L}_{\text{scal}}^{\text{sem}}$ (Cao & De Charette, 2022). The overall loss function is defined as follows (See more details in **Appendix**):

$$\mathcal{L}_{\text{occ}} = \mathcal{L}_{\text{bce}} + \mathcal{L}_{\text{ce}} + \mathcal{L}_{\text{ls}} + \mathcal{L}_{\text{scal}}^{\text{geo}} + \mathcal{L}_{\text{scal}}^{\text{sem}}. \tag{9}$$

Navigation Loss. Following conventions, our agent is optimized with behavior cloning by minimizing the cross-entropy loss between the predicted action and the expert action a_t^* at each timestep:

$$\mathcal{L}_{\text{nav}} = -\sum\nolimits_t \log p(a_t^* \mid \mathcal{O}_t, \mathcal{X}, \boldsymbol{h}_{t-1}). \tag{10}$$

3.5 Implementation Details

Occupancy Annotation. To generate voxel labels for 3D occupancy semantic prediction, we back-project multi-frame depth maps to point clouds using camera intrinsics and extrinsics. The accumulated point clouds are then discretized into a volumetric grid according to the perception range and voxel resolution. For each voxel, its semantic label is determined by the majority class among the points falling into that voxel. In addition, a ray-casting algorithm is employed to identify unknown voxels caused by occlusion or falling outside the field of view, which are omitted during training.

MSVM Learning. For the input image, we adopt ResNet-50 (He et al., 2016) to extract $\boldsymbol{F}^{2d}$. The depth map is discretized into 100 bins. MSVM is constructed through 4 stages, yielding a finest voxel grid of resolution $128 \times 128 \times 40$. The perception range is set to $[0m, 51.2m]$ along the forward direction, $[-25.6m, 25.6m]$ laterally, and $[-8m, 8m]$ vertically. During refinement, all voxels are refined at the first stage, and only 20% of the voxels are kept in the subsequent stages. We use $M = 4$ sampling points in Eq. 3, and set the feature dimension to 512 (*i.e.*, $C = 512$).

Navigation Network. For instructions, we employ pretrained BERT (Devlin et al., 2019) as language encoder. Hierarchical volumetric reasoning consists of 3 stages, which are performed over the first three levels of the volume hierarchy (denoted as coarse/middle/fine), as they are empirically sufficient to cover spatial information at different granularities. The number of proxy tokens is set to 32, and the number of candidate voxels in the last two stages are set to 256 (*i.e.*, $G_2 = G_3 = 256$). The multi-layer Transformer includes 4 layers and is shared across different reasoning stages.

Training Recipe. A two-stage training strategy is adopted, which separates MSVM learning from policy learning. Specifically, MSVM encoder is first trained by $\mathcal{L}_{\text{occ}}$ (Eq. 9) with a batch size of 32 for 25 epochs, using AdamW (Loshchilov & Hutter, 2017) optimizer with an initial learning rate of 2e-4. Moreover, we adopt cosine learning rate scheduler with linear warming up in the first 1000 iterations. In the second stage, MSVM encoder keeps frozen and the navigation policy network is optimized by $\mathcal{L}_{\text{nav}}$ (Eq. 10) using AdamW optimizer with 2e-4 learning rate for 50 epochs. Since full backpropagation through time over the entire trajectory is computationally prohibitive, truncated backpropagation through time (Liao et al., 2018) is applied during policy learning with 4 steps.

4 Experiment

4.1 Experimental setup

Datasets. Experiments and evaluations are conducted on AerialVLN-s (Liu et al., 2023), which consists of 3916 flight trajectories collected from expert UAV pilots across 17 city-scale scenes. The benchmark emphasizes long-horizon navigation with an average trajectory length of 321.3 meters.

Metrics. Following conventions (Liu et al., 2023; Ning et al., 2026), four standard metrics are adopted to assess the agents performance: Success Rate (SR), Oracle Success Rate (OSR), Navigation Error (NE), and Success Rate weighted by Normalized Dynamic Time Warping (SDTW).

Baselines. The compared aerial agents fall into three categories: ***i)*** **Statistic-based**, including random action selection and action sampling. ***ii)*** **Learning-based**, which learns a navigation policy

Table 1: **Quantitative results** on AerialVLN-s (Liu et al., 2023). Our agent is compared against three types of aerial agents, including **statistic-based**, **learning-based**, and **LLM-based**. See more details in §4.2.

Agent	*Pano.*	*LLM*	*val seen* split				*val unseen* split			
			NE ↓	SR ↑	OSR ↑	SDTW ↑	NE ↓	SR ↑	OSR ↑	SDTW ↑
Random	–	–	109.6	0.0	0.0	0.0	149.7	0.0	0.0	0.0
Action Sampling	–	–	213.8	0.9	5.7	0.3	237.6	0.2	1.1	0.1
LingUNet [EMNLP18] (Misra et al., 2018)	–	–	383.8	0.6	6.9	0.2	368.4	0.4	3.6	0.9
Seq2Seq [CVPR18] (Anderson et al., 2018)	–	–	146.0	4.8	19.8	1.6	218.9	2.3	11.7	0.7
CMA [ECCV20] (Krantz et al., 2020)	–	–	121.0	3.0	23.3	0.6	172.1	3.2	16.0	1.1
LAG [ICCV23] (Liu et al., 2023)	–	–	90.2	7.2	15.7	2.4	127.9	5.1	10.5	1.4
Ours	–	–	**84.1**±3.16	**16.2**±0.45	**35.4**±0.42	**12.5**±0.70	**109.4**±2.74	**8.8**±0.37	**26.5**±0.39	**3.9**±0.56
OpenFly [ICLR26] (Gao et al., 2026b)	–	✓	127.2	8.1	21.8	1.6	113.8	7.6	18.2	1.5
CityNavAgent [ACL25] (Zhang et al., 2025)	✓	✓	80.8	13.9	30.2	5.1	60.2	11.7	35.2	5.0
LookasideVLN [CVPR26] (Ning et al., 2026)	✓	✓	77.1	14.7	31.2	5.4	100.9	12.6	36.0	3.6

* *Pano.* denotes panoramic observations. Our results are averaged over five runs, with standard deviations reported.

from expert trajectories. ***iii)*** **LLM-based**, which employs LLMs/MLLMs as navigators, either through task-specific finetuning (*e.g.*, OpenFly (Gao et al., 2026b)) or zero-shot prompting (*e.g.*, CityNavAgent (Zhang et al., 2025), LookasideVLN (Ning et al., 2026)). Our agent belongs to the second category and is mainly compared against aerial agents therein, as they share the same setting.

4.2 Quantitative results

Table 1 reports the performance results of different agents on AerialVLN-s (Liu et al., 2023).

Results on Seen Environments. Our agent achieves SR of $\mathbf{16.2}\%$ and SDTW of $\mathbf{12.5}\%$ on *val seen* split, which leads to significant gains of $\mathbf{9.0}\%/\mathbf{10.1}\%$ over the best learning-based agent LAG (Liu et al., 2023), demonstrating the effectiveness of our designs for policy learning. Compared to LLM-based ones, our agent also surpasses LookaideVLN (Ning et al., 2026) by $\mathbf{1.5}\%$ in SR and $\mathbf{7.1}\%$ in SDTW, even though it constructs topological map from panoramic observation and incorporate a much stronger MLLM (*i.e.*, QWen-2.5-VL-72B (Bai et al., 2025)) as policy for action reasoning.

Generalization to Unseen Environments. Among learning-based agents, our agent achieves the strongest performance on *val unseen* split. It outperforms LAG (Liu et al., 2023) by $\mathbf{3.7}\%$ in SR, $\mathbf{16.0}\%$ in OSR, and $\mathbf{2.5}\%$ in SDTW. The advantage extends to the LLM-based OpenFly (Gao et al., 2024) under the same egocentric observation setting, with gains of $\mathbf{1.2}\%$ in SR, $\mathbf{8.3}\%$ in OSR, and $\mathbf{2.4}\%$ in SDTW. Even when panoramic observations are further introduced, the gap remains small, with SR trailing CityNavAgent (Zhang et al., 2025) and LookasideVLN (Ning et al., 2026) by only 2.9% and 3.8%, respectively, while exceeding LookasideVLN in SDTW ($\mathbf{3.9}\%$ vs 3.6%). Such sensing, however, is less practical for real-world UAV deployment.

Computational Efficiency. Table 2 reports the computational costs of different agents including runtime, and memory usage. All agents are evaluated on a single NVIDIA RTX 4090 GPU, with the input image resized to 256×256. Our agent achieves the highest SR of $\mathbf{16.2}\%$ while maintaining a moderate computational overhead. It only increases the runtime from 0.05s to $\mathbf{0.09}$s per step, and memory from 0.84G to $\mathbf{1.71}$G over LAG (Liu et al., 2023), while yielding a substantial SR gain ($+\mathbf{9.0}\%$). In contrast, LLM-based aerial agent like CityNavAgent (Zhang et al., 2025) requires substantially higher runtime and memory consumption (12.51 s and 32.76 GB), while achieving a lower SR.

Table 2: **Computational efficiency comparison** on AerialVLN-S (Liu et al., 2023) *val seen* split. See § 4.2 for more details.

Agent	Runtime (s) ↓	Memory (GB) ↓	SR ↑
LAG (Liu et al., 2023)	0.05	0.84	7.2
OpenFly (Gao et al., 2026b)	0.34	14.25	8.1
CityNavAgent (Zhang et al., 2025)	12.51	32.76	13.9
Ours	0.09	1.71	**16.2**

4.3 Qualitative results

Case Studies. Fig. 3 depicts one exemplar navigation episode from *val seen* set of AerialVLN-s (Liu et al., 2023). The episode requires the agent to traverse a long route in which the instruction-relevant landmarks are sparsely scattered. During navigation, our agent constructs Muti-Scale Voxel Map

[Instruction]: ① take off ② move forward towards the **small tree near the pillar stone**. ③ move forward near the **lamp post**. ④ flying over there and slow down before the **building** at the corner of the building. ⑤ look left and right to the front of the best **burger shop** entrance. ⑥ land it over there on the **terrace of the building**.

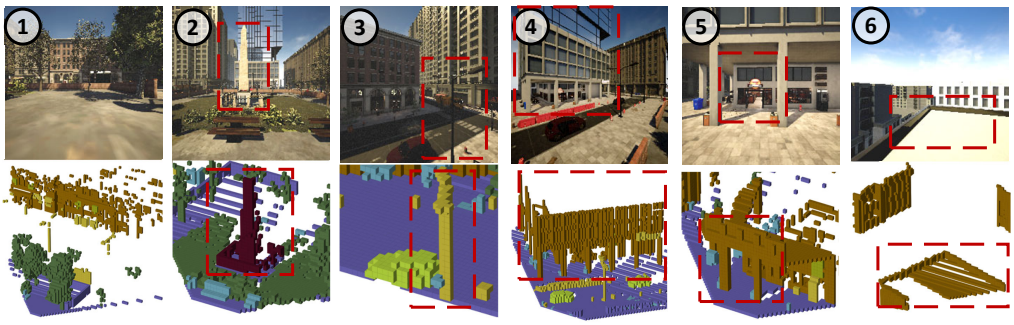

Figure 3: **A representative result** on AerialVLN-s Liu et al. (2023) *val seen* split. We show the observation and the 3D semantic occupancy prediction at key steps. Benefiting from the Multi-Scale Voxel Map, our agent achieves a comprehensive understanding of scene structures and semantic content across spatial scales, such as "*the pillar stone*" and "*the building*". See §4.3 for more details.

from current observation and predicts 3D semantic occupancy; it well captures crucial landmarks and objects at key steps (*e.g.*"*the pillar stone*" at step ② and "*the building*" at step ④), grounds the instruction to these spatial evidence along the trajectory, and finally accomplishes the task.

4.4 Diagnostic Experiment

For thorough examination of our agent, we first ablate the overall design by isolating the contribution of each core component (*i.e.*, MSVM (§ 3.1), HVR (§ 3.2)). Then, an in-depth analysis of HVR is carried out. Finally, a comprehensive study of MSVM learning strategy is deferred to § 4.5.

Overall Design. Table 3 investigates the effectiveness of the keys modules by gradually appending them into the baseline model (row #1). ***i)*** Incorporating MSVM (row #2) yields notable performance gains (*e.g.*, 2.7% → **7.5**% for SR and 0.8% → **3.1**% for SDTW), this highlights the superiority of MSVM over pure 2D visual observation, which enables the agent make more informed decisions according to both 3D geometric and semantic cues. ***ii)*** Building on MSVM (row #2), HVR (row #3) delivers additional gains, improving SR from 7.5% to **8.8**% and SDTW from 3.1% to **3.9**%. The results indicate that reasoning over the 3D space benefits navigation in large-scale aerial environments by enabling more informed and reliable decision-making.

Table 3: **Ablation study of overall design** on *val unseen* of AerialVLN-s (Liu et al., 2023). See § 4.4 for more details.

#	*Components*		*val unseen* split			
	MSVM	HVR	NE ↓	SR ↑	OSR ↑	SDTW ↑
1			154.8	2.7	12.6	0.8
2	✓		118.2	7.5	24.6	3.1
3	✓	✓	**109.4**	**8.8**	**26.5**	**3.9**

A Deeper Look at HVR. Table 4 shows how different reasoning granularity in 3D volumetric space affects navigation performance. Row #1 serves as the baseline which directly reasons over 2D representations. ***i)*** **Row #1 *vs* #2, #3, #4**: exposing the agent to 3D volumetric space for decision-making results in consistent performance improvements (*e.g.*, +**2.1**% SR gains under coarse-level), which demonstrates the advantages of 3D volumetric reasoning over pure 2D image. ***ii)*** **Row #2∼#4**: finer volumetric features leads to improved navigation performance, as the agent benefits from precise geometric perception and fine-grained semantic grounding. ***iii)*** **Row #5 *vs* #2, #3, #4**: employing hierarchical reasoning before decision making significantly outperforms

Table 4: **Ablation study of HVR** on *val unseen* of AerialVLN-s (Liu et al., 2023). See § 4.4 for more details.

#	*Reasoning Stage*			*val unseen* split			
	Coarse	Middle	Fine	NE ↓	SR ↑	OSR ↑	SDTW ↑
1				154.8	2.7	12.6	0.8
2	✓			140.2	4.8	12.8	1.2
3		✓		118.5	6.4	17.3	2.0
4			✓	119.1	6.6	18.4	2.5
5	✓	✓	✓	**109.4**	**8.8**	**26.5**	**3.9**

single-step alternatives (*e.g.*, $+\mathbf{4.0\%/2.4\%/2.2\%}$ SR over coarse/middle/fine-only), which highlights their complementary roles and the necessity to exploit multi-scale cues for aerial navigation.

4.5 Analysis of MSVM Learning

To investigate the effectiveness of our MSVM learning (§3.1) strategies, series of experiments of 3D occupancy prediction are carried out on a self-collected validation set comprised of total 957 samples from *val seen* split. Following standard protocols, both IoU and mIoU are employed to evaluate the geometry and semantic prediction accuracy. We also report the navigation performance to evaluate the transferability of these learned 3D representations to *val unseen* navigation scenarios.

Table 5: **Ablation studies of MSVM learning** (§3.1) on AerialVLN-s (Liu et al., 2023).

(a) Top-$k\%$ at different volume scales.

Top-k (%)		*3D Occupancy*		*AVLN*	
$l1$	$l2 \sim l4$	IoU ↑	mIoU ↑	SR ↑	SDTW ↑
10	10	0.18	0.07	6.2	2.1
100	10	0.25	0.10	7.7	2.9
100	**20**	**0.34**	**0.14**	**8.8**	**3.9**
100	50	0.35	0.14	8.4	3.6

(b) Voxel refinement

Voxel Refine.	*3D Occupancy*		*AVLN*	
	IoU ↑	mIoU ↑	SR ↑	SDTW ↑
2D Attn.	0.27	0.10	8.0	3.2
3D Attn.	**0.34**	**0.14**	**8.8**	**3.9**
w/o 3D Conv	0.30	0.12	8.2	3.3

Top-$k\%$ in volume construction. Table 5a investigates the effect of different top-$k\%$ configurations during volume construction. Applying an aggressive 10% selection from the first level substantially degrades occupancy prediction, yielding only 0.18 IoU and 0.07 mIoU, together with inferior navigation performance. In contrast, retaining all voxels at the coarsest level improves both occupancy prediction ($0.18/0.07 \rightarrow \mathbf{0.25/0.10}$ for IoU/mIoU) and navigation ($6.2\%/2.1\% \rightarrow \mathbf{7.7\%/2.9\%}$ for SR/SDTW), indicating that preserving complete global spatial context is important for subsequent perception and reasoning. Increasing top-$k\%$ at finer levels from 10% to 20% further improves IoU/mIoU ($0.25/0.10 \rightarrow \mathbf{0.34/0.14}$) and achieves the best navigation performance ($\mathbf{8.8\%/3.9\%}$ for SR/SDTW), indicating that sufficient fine-scale refinement better captures local geometry and spatial semantics required for precise navigation. Further increasing the ratio to 50%, however, yields only marginal gains in occupancy prediction and provides little benefit to navigation.

Voxel Refinement. Table 5b studies the impact of different voxel refinement strategies. Replacing 2D attention with 3D attention benefits both occupancy prediction ($0.27/0.10 \rightarrow \mathbf{0.34/0.14}$ for IoU/mIoU) and navigation performance ($8.0\%/3.2\% \rightarrow \mathbf{8.8\%/3.9\%}$ for SR/SDTW), demonstrating that explicitly modeling interactions among voxels in 3D space benefits geometric and semantic prediction, and in-turn provides more effective spatial representations for subsequent reasoning and decision-making. Moreover, removing the 3D convolution results in slightly degraded occupancy prediction ($\mathbf{0.34/0.14} \rightarrow 0.30/0.12$ for IoU/mIoU) and navigation performance ($\mathbf{8.8\%/3.9\%} \rightarrow 8.2\%/3.3\%$ for SR/SDTW). This indicates that 3D convolution complements global attention by aggregating local neighborhood context and enhancing geometric consistency among adjacent voxels, thereby providing more reliable local spatial cues for navigation decisions.

5 Conclusion

In this work, we introduce a navigation framework that establishes a unified 3D volumetric space for environment perception, spatial reasoning, and decision making in Aerial Vision-Language Navigation. It transforms egocentric observations into a Multi-Scale Voxel Map that efficiently integrates scene geometry and visual semantics across multiple spatial granularities, providing an explicit spatial foundation for navigation. Based on this map, the agent performs Hierarchical Volumetric Reasoning that progressively narrows the spatial search scope and aggregates task-relevant 3D evidence, moving from the overall scene layout to local structures. The resulting decision representation is further integrated with navigation history to support spatially informed action prediction. Experiments on AerialVLN-s demonstrate competitive performance in both seen and unseen environments with a real-time inference latency, while extensive ablation studies validate the effectiveness of our designs.

AI USE STATEMENT

In this work, we used generative AI tools for editing the paper to improve readability. We have not used generative AI tools for generating synthetic data, developing theoretical models, or implementing methods, and the rest of the required disclosure tasks are not applicable to this work. We have reviewed all AI-assisted work. For example, all LLM-generated text was verified and edited by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Michal Adamkiewicz, Timothy Chen, Adam Caccavale, Rachel Gardner, Preston Culbertson, Jeannette Bohg, and Mac Schwager. Vision-only robot navigation in a neural radiance world. *IEEE RAL*, 7(2):4606–4613, 2022.

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

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025.

Maxim Berman, Amal Rannen Triki, and Matthew B Blaschko. The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In *CVPR*, 2018.

Hengxing Cai, Jinhan Dong, Jingjun Tan, Jingcheng Deng, Sihang Li, Zhifeng Gao, Haidong Wang, Zicheng Su, Agachai Sumalee, and Renxin Zhong. Flightgpt: Towards generalizable and interpretable uav vision-and-language navigation with vision-language models. In *EMNLP*, 2025.

Anh-Quan Cao and Raoul De Charette. Monoscene: Monocular 3d semantic scene completion. In *CVPR*, 2022.

Vincent Cartillier, Zhile Ren, Neha Jain, Stefan Lee, Irfan Essa, and Dhruv Batra. Semantic mapnet: Building allocentric semantic maps and representations from egocentric views. In *AAAI*, 2021.

Trishna Chakraborty, Udita Ghosh, Xiaopan Zhang, Fahim Faisal Niloy, Yue Dong, Jiachen Li, Amit K Roy-Chowdhury, and Chengyu Song. Heal: An empirical study on hallucinations in embodied agents driven by large language models. *arXiv preprint arXiv:2506.15065*, 2025.

Devendra Singh Chaplot, Dhiraj Gandhi, Saurabh Gupta, Abhinav Gupta, and Ruslan Salakhutdinov. Learning to explore using active neural slam. *arXiv preprint arXiv:2004.05155*, 2020.

Hanxuan Chen, Jie Zheng, Siqi Yang, Tianle Zeng, Siwei Feng, Songsheng Cheng, Ruilong Ren, Hanzhong Guo, Shuai Yuan, Xiangyue Wang, et al. Vision-and-language navigation for uavs: Progress, challenges, and a research roadmap. *arXiv preprint arXiv:2604.13654*, 2026a.

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

Jiaqi Chen, Bingqian Lin, Ran Xu, Zhenhua Chai, Xiaodan Liang, and Kwan-Yee Wong. Mapgpt: Map-guided prompting with adaptive path planning for vision-and-language navigation. In *ACL*, 2024.

Kevin Chen, Junshen K Chen, Jo Chuang, Marynel Vázquez, and Silvio Savarese. Topological planning with transformers for vision-and-language navigation. In *CVPR*, 2021.

Shizhe Chen, Pierre-Louis Guhur, Makarand Tapaswi, Cordelia Schmid, and Ivan Laptev. Think global, act local: Dual-scale graph transformer for vision-and-language navigation. In *CVPR*, 2022.

Tao Chen, Saurabh Gupta, and Abhinav Gupta. Learning exploration policies for navigation. *arXiv preprint arXiv:1903.01959*, 2019b.

Xuecheng Chen, Zongzhuo Liu, Jianfa Ma, Bang Du, Tiantian Zhang, Xueqian Wang, and Boyu Zhou. Airhunt: Bridging vlm semantics and continuous planning for efficient aerial object navigation. *arXiv preprint arXiv:2601.12742*, 2026b.

Meng Chu, Zhedong Zheng, Wei Ji, Tingyu Wang, and Tat-Seng Chua. Towards natural language-guided drones: Geotext-1652 benchmark with spatial relation matching. In *ECCV*, 2024.

Tianchen Deng, Guole Shen, Yilin Shen, Wenhua Wu, Yilin Fang, Ziqi Ma, Tianjun Zhang, Shenghai Yuan, Wolfram Burgard, and Hesheng Wang. Adaptive world memory 3d foundation model for scalable 3d mapping, localization, and rendering. *arXiv preprint arXiv:2609.21502*, 2026.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *ACL*, 2019.

Xichen Ding, Jianzhe Gao, Cong Pan, Wenguan Wang, and Jie Qin. History-enhanced two-stage transformer for aerial vision-and-language navigation. In *AAAI*, 2026.

Chengjie Fan, Cong Pan, Zijian Liu, Ningzhong Liu, and Jie Qin. Htnav: A hybrid navigation framework with tiered structure for urban aerial vision-and-language navigation. *arXiv preprint arXiv:2604.08883*, 2026.

Yue Fan, Winson Chen, Tongzhou Jiang, Chun Zhou, Yi Zhang, and Xin Wang. Aerial vision-and-dialog navigation. In *ACL*, 2023.

Jianzhe Gao, Rui Liu, and Wenguan Wang. 3d gaussian map with open-set semantic grouping for vision-language navigation. In *ICCV*, 2025.

Jianzhe Gao, Rui Liu, Yuxuan Xu, Tongtong Cao, Yingxue Zhang, Zhanguang Zhang, Sida Peng, Yi Yang, and Wenguan Wang. Uncertainty-aware gaussian map for vision-language navigation. *arXiv preprint arXiv:2605.26503*, 2026a.

Yunpeng Gao, Zhigang Wang, Linglin Jing, Dong Wang, Xuelong Li, and Bin Zhao. Aerial vision-and-language navigation via semantic-topo-metric representation guided llm reasoning. *arXiv preprint arXiv:2410.08500*, 2024.

Yunpeng Gao, Chenhui Li, Zhongrui You, Junli Liu, Zhen Li, Pengan Chen, Qizhi Chen, Zhonghan Tang, Liansheng Wang, Penghui Yang, et al. Openfly: A comprehensive platform for aerial vision-language navigation. In *ICLR*, 2026b.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, 640(8059):647–653, 2025.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

Joao F Henriques and Andrea Vedaldi. Mapnet: An allocentric spatial memory for mapping environments. In *CVPR*, 2018.

Yatai Ji, Zhengqiu Zhu, Yong Zhao, Beidan Liu, Chen Gao, Yihao Zhao, Sihang Qiu, Yue Hu, and Quanjun Yin. Towards autonomous uav visual object search in city space: Benchmark and agentic methodology. In *AAAI*, 2026.

Wen Jiang, Li Wang, Kangyao Huang, Wei Fan, Jinyuan Liu, Shaoyu Liu, Hongwei Duan, Bin Xu, and Xiangyang Ji. Longfly: Long-horizon uav vision-and-language navigation with spatiotemporal context integration. *arXiv preprint arXiv:2512.22010*, 2025.

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

Obin Kwon, Jeongho Park, and Songhwai Oh. Renderable neural radiance map for visual navigation. In *CVPR*, 2023.

Jungdae Lee, Taiki Miyanishi, Shuhei Kurita, Koya Sakamoto, Daichi Azuma, Yutaka Matsuo, and Nakamasa Inoue. Citynav: A large-scale dataset for real-world aerial navigation. In *ICCV*, 2025.

Hongyang Li, Hao Zhang, Zhaoyang Zeng, Shilong Liu, Feng Li, Tianhe Ren, and Lei Zhang. Dfa3d: 3d deformable attention for 2d-to-3d feature lifting. In *ICCV*, 2023.

Renjie Liao, Yuwen Xiong, Ethan Fetaya, Lisa Zhang, KiJung Yoon, Xaq Pitkow, Raquel Urtasun, and Richard Zemel. Reviving and improving recurrent back-propagation. In *ICML*, 2018.

Fei Lin, Yonglin Tian, Yunzhe Wang, Tengchao Zhang, Xinyuan Zhang, and Fei-Yue Wang. Airvista: Empowering uavs with 3d spatial reasoning abilities through a multimodal large language model agent. In *ITSC*, 2024.

Rui Liu, Wenguan Wang, and Yi Yang. Volumetric environment representation for vision-language navigation. In *CVPR*, 2024.

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

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Artem Lykov, Valerii Serpiva, Muhammad Haris Khan, Oleg Sautenkov, Artyom Myshlyaev, Grik Tadevosyan, Yasheerah Yaqoot, and Dzmitry Tsetserukou. Cognitivedrone: A vla model and evaluation benchmark for real-time cognitive task solving and reasoning in uavs. *arXiv preprint arXiv:2503.01378*, 2025.

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1): 99–106, 2021.

Dipendra Misra, Andrew Bennett, Valts Blukis, Eyvind Niklasson, Max Shatkhin, and Yoav Artzi. Mapping instructions to actions in 3d environments with visual goal prediction. In *EMNLP*, 2018.

Yuwei Ning, Ganlong Zhao, Yipeng Qin, Si Liu, Yang Liu, Liang Lin, and Guanbin Li. Lookasidevln: direction-aware aerial vision-and-language navigation. In *CVPR*, 2026.

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *AAAI*, 2018.

Guanhui Qiao, Dong Yi, Lingxiang Wu, Hanxiao Wu, and Jinqiao Wang. Enhancing visual aligning and grounding for aerial vision-and-dialog navigation. *IEEE SPL*, 32:2853–2857, 2025.

Santhosh K Ramakrishnan, Dinesh Jayaraman, and Kristen Grauman. An exploration of embodied visual exploration. *IJCV*, 129(5):1616–1649, 2021.

Luis Roldao, Raoul De Charette, and Anne Verroust-Blondet. Lmscnet: Lightweight multiscale 3d semantic completion. In *3DV*, 2020.

Oleg Sautenkov, Yasheerah Yaqoot, Muhammad Ahsan Mustafa, Faryal Batool, Jeffrin Sam, Artem Lykov, Chih-Yung Wen, and Dzmitry Tsetserukou. Uav-codeagents: Scalable uav mission planning via multi-agent react and vision-language reasoning. *arXiv preprint arXiv:2505.07236*, 2025.

Raphael Schumann, Wanrong Zhu, Weixi Feng, Tsu-Jui Fu, Stefan Riezler, and William Yang Wang. Velma: Verbalization embodiment of llm agents for vision and language navigation in street view. In *AAAI*, 2024.

Yin Tang, Jiawei Ma, Jiahao Li, Hao Zhang, Zhemin Sun, Jianqiao Sun, and Deyu Zhang. Freqnav: Stage-wise frequency routing for object-oriented aerial vision-language navigation. *arXiv preprint arXiv:2608.00970*, 2026.

Sebastian Thrun. Learning metric-topological maps for indoor mobile robot navigation. *Artificial intelligence*, 99(1):21–71, 1998.

Sebastian Thrun. Probabilistic robotics. *Communications of the ACM*, 45(3):52–57, 2002.

Yonglin Tian, Fei Lin, Yiduo Li, Tengchao Zhang, Qiyao Zhang, Xuan Fu, Jun Huang, Xingyuan Dai, Yutong Wang, Chunwei Tian, et al. Uavs meet llms: Overviews and perspectives towards agentic low-altitude mobility. *Information Fusion*, 122:103158, 2025.

Johanna Wald, Helisa Dhamo, Nassir Navab, and Federico Tombari. Learning 3d semantic scene graphs from 3d indoor reconstructions. In *CVPR*, 2020.

Meiqi Wang, Longnyu Xu, Jun Liu, Hewu Li, and Han Qiu. Cotfly: Making uavs think where to fly next through visual chain-of-thought reasoning. In *CVPR*, 2026a.

Xiangyu Wang, Donglin Yang, Hohin Kwan, Jinyu Chen, Hongsheng Li, Yue Liao, Si Liu, et al. Towards realistic uav vision-language navigation: Platform, benchmark, and methodology. In *ICLR*, 2025.

Xiangyu Wang, Donglin Yang, Yue Liao, Wenhao Zheng, Bin Dai, Hongsheng Li, Si Liu, et al. Uav-flow colosseo: A real-world benchmark for flying-on-a-word uav imitation learning. In *NeurIPS*, 2026b.

Ruipu Wu, Yige Zhang, Jinyu Chen, Linjiang Huang, Shifeng Zhang, Xu Zhou, Liang Wang, and Si Liu. Aeroduo: Aerial duo for uav-based vision and language navigation. In *ACM MM*, 2025.

Xingyu Xia, Lekai Zhou, Yujie Tang, Xiaozhou Zhu, Hai Zhu, and Wen Yao. Vision-language navigation for aerial robots: Towards the era of large language models. *arXiv preprint arXiv:2604.07705*, 2026.

Jianqiang Xiao, Yuexuan Sun, Yixin Shao, Boxi Gan, Rongqiang Liu, Yanjin Wu, Weili Guan, and Xiang Deng. Uav-on: A benchmark for open-world object goal navigation with aerial agents. In *ACM MM*, 2025.

Haotian Xu, Yue Hu, Chen Gao, Zhengqiu Zhu, Yong Zhao, Yong Li, and Quanjun Yin. Geonav: Empowering mllms with explicit geospatial reasoning abilities for language-goal aerial navigation. *arXiv preprint arXiv:2504.09587*, 2025.

Huilin Xu, Zhuoyang Liu, Yixiang Luomei, and Feng Xu. Aerial vision-language navigation with a unified framework for spatial, temporal and embodied reasoning. *IEEE CSVT*, 2026a.

Peng Xu, Zhengnan Deng, Jiayan Deng, Zonghua Gu, and Shaohua Wan. Aerialvla: A vision-language-action model for uav navigation via minimalist end-to-end control. *arXiv preprint arXiv:2603.14363*, 2026b.

Zhongxing Xu, Chengzhi Liu, Qingyue Wei, Juncheng Wu, James Zou, Xin Wang, Yuyin Zhou, and Sheng Liu. More thinking, less seeing? assessing amplified hallucination in multimodal reasoning models. In *NeurIPS*, 2026c.

Fanglong Yao, Yuanchang Yue, Youzhi Liu, Xian Sun, and Kun Fu. Aeroverse: Uav-agent benchmark suite for simulating, pre-training, finetuning, and evaluating aerospace embodied world models. *arXiv preprint arXiv:2408.15511*, 2024.

Daoxuan Zhang, Ping Chen, Xiaobo Xia, Xiu Su, Ruichen Zhen, Jianqiang Xiao, and Shuo Yang. Apex: A decoupled memory-based explorer for asynchronous aerial object goal navigation. *arXiv preprint arXiv:2602.00551*, 2026.

Weichen Zhang, Chen Gao, Shiquan Yu, Ruiying Peng, Baining Zhao, Qian Zhang, Jinqiang Cui, Xinlei Chen, and Yong Li. Citynavagent: Aerial vision-and-language navigation with hierarchical semantic planning and global memory. In *ACL*, 2025.

Yue Zhang, Quan Guo, and Parisa Kordjamshidi. Towards navigation by reasoning over spatial configurations. In *SpLU-RoboNLP*, 2021.

Baining Zhao, Ziyou Wang, Jianjie Fang, Zile Zhou, Yanggang Xu, Yatai Ji, Jiacheng Xu, Qian Zhang, Weichen Zhang, Chen Gao, et al. How far are large multimodal models from human-level spatial action? a benchmark for goal-oriented embodied navigation in urban airspace. In *ACM KDD*, 2026.

Xiangyi Zheng, Xiangyu Wang, Qinan Liao, Zimu Tang, Yue Liao, Dongyue Lyu, Guodong Wang, Junjie Liu, and Si Liu. Think like a pilot: Fine-grained long-horizon uav navigation. *arXiv preprint arXiv:2606.06836*, 2026.

Jiacong Zhou, Jiaxu Miao, Yourun Lin, Xianyun Wang, Jun Xiao, and Jun Yu. Memory-augmented scene understanding and exploration for open-world aerial object-goal navigation. In *CVPR*, 2026a.

Jiaxu Zhou, Shaobo Wang, Zhiyuan Yang, Zhenjun Yu, and Tao Li. Freefly-thinking: Aligning chain-of-thought reasoning with continuous uav navigation. *arXiv preprint arXiv:2603.07181*, 2026b.

SUMMARY OF THE APPENDIX

This appendix contains additional details for the ICLR 2027 submission, titled *Hierarchical Volumetric Reasoning for Aerial Vision-Language Navigation*. The appendix is organized as follows:

- §A provides additional implementation details.
- §B report additional experiments.
- §C provides a discussion of the limitations and future, societal impact, terms of use, privacy, and license.

A Additional Details

A.1 Occupancy losses

As described in § 3.4, a combination of geometric and semantic occupancy losses is adopted to optimize the Multi-Scale Voxel Map, enabling the capture of both geometric structure and semantic information across multiple spatial scales. Specifically, binary occupancy supervision is applied to the intermediate volumes, while semantic occupancy supervision is imposed on the finest volume.

Multi-Scale Occupancy Supervision. At each intermediate stage $l \in \{1, \ldots, L-1\}$, a occupancy prediction head estimates the occupancy probability O_l from the volumetric features. We employ binary cross-entropy loss to supervise these predictions at different spatial resolutions:

$$\mathcal{L}_{\text{bce}} = \sum\nolimits_{l=1}^{L-1} \text{BCE}(O_l, Y_l), \tag{11}$$

where Y_l denotes the ground-truth binary occupancy labels at level l. For each valid voxel v, the binary cross-entropy is formulated as

$$\text{BCE}(O_l, Y_l) = -\frac{1}{|\Omega_l|} \sum\nolimits_{v \in \Omega_l} \left[y_{l,v} \log O_l(v) + (1 - y_{l,v}) \log(1 - O_l(v)) \right]. \tag{12}$$

where Ω_l denotes the set of valid voxels at level l, and $y_{l,v} \in \{0, 1\}$ indicates whether voxel v is occupied. Unknown voxels caused by occlusion are excluded from supervision.

Semantic Occupancy Supervision. At the finest level L, the semantic occupancy output O_L is supervised by weighted cross-entropy (Roldao et al., 2020) and Lovasz-Softmax (Berman et al., 2018). Let $O_L(v, c)$ denote the predicted probability of semantic class c at voxel v, and let y_v denote its ground-truth semantic label. The weighted cross-entropy loss takes the form

$$\mathcal{L}_{\text{ce}} = -\frac{1}{|\Omega_L|} \sum\nolimits_{v \in \Omega_L} w_{y_v} \log O_L(v, y_v), \tag{13}$$

where w_{y_v} is the weight associated with the ground-truth semantic class. In addition, our agent adopt the Lovasz-Softmax loss $\mathcal{L}_{\text{ls}}$ to directly optimize a differentiable surrogate of the intersection over union (IoU) for semantic occupancy prediction.

Scene-Class Affinity Supervision. Following (Cao & De Charette, 2022), we introduce geometric and semantic scene-class affinity losses to improve the global consistency of 3D semantic occupancy prediction. The geometric affinity loss $\mathcal{L}_{\text{scal}}^{\text{geo}}$ operates on binary occupancy predictions, encouraging consistency between occupied and empty regions. In contrast, the semantic affinity loss $\mathcal{L}_{\text{scal}}^{\text{sem}}$ operates on individual semantic classes, encouraging consistent prediction of their spatial distributions.

B ADDITIONAL EXPERIMENTS

This section presents supplementary ablation studies on proxy tokens and spatial routing.

B.1 ABLATION STUDY ON PROXY TOKEN

We study the effectiveness of proxy tokens in HVR (§3.2) on the AerialVLN-s *val unseen* split. As shown in Table 6, increasing the number of proxy tokens from 0 to 32 consistently improves all navigation metrics. With 32 proxy tokens, the agent improves SR, OSR, and SDTW by $\mathbf{1.9}\%$, $\mathbf{5.8}\%$, and $\mathbf{1.6}\%$, respectively, compared with the variant without proxy tokens. These improvements suggest that proxy tokens capture complementary spatial context that benefits action prediction. Further increasing the number to 48 yields no additional SR gain and slightly degrades other metrics, indicating diminishing benefits from additional tokens. We therefore adopt 32 proxy tokens as the default configuration.

Table 6: **Ablation study of proxy token** on *val unseen* of AerialVLN-s (Liu et al., 2023).

Proxy Token	*val unseen* split			
	NE ↓	SR ↑	OSR ↑	SDTW ↑
0	124.7	6.9	20.7	2.3
16	116.1	8.0	24.2	3.2
32	**109.4**	**8.8**	**26.5**	**3.9**
48	110.2	8.8	25.9	3.7

B.2 ABLATION STUDY ON SPATIAL ROUTING

To evaluate the effectiveness of the spatial routing mechanism in HVR (§3.2), we vary the voxel candidate budgets across three spatial scales on the AerialVLN-s *val unseen* split. As shown in Table 7, expanding all voxels across scales leads to an out-of-memory (OOM) error, highlighting the need to control the rapidly growing candidate space. With the coarsest-scale budget fixed at $G_1 = 1280$, progressively reducing G_2 and G_3 from 1280 to 256 consistently improves all navigation metrics. The configuration with $G_2 = G_3 = 256$ achieves the best overall performance, improving SR and SDTW by $\mathbf{1.3}\%$ and $\mathbf{0.9}\%$, respectively, over $G_2 = G_3 = 1280$. These results suggest that concentrating fine-grained reasoning on routed spatial regions at coarser level helps the agent gather useful evidence for action prediction while keeping computation tractable.

Table 7: **Ablation study of spatial routing** on *val unseen* of AerialVLN-s (Liu et al., 2023).

Spatial Routing			*val unseen* split			
G_1	G_2	G_3	NE ↓	SR ↑	OSR ↑	SDTW ↑
1280	1280×8	1280×64	OOM			
1280	1280	1280	112.0	7.5	24.6	3.0
1280	512	512	110.6	8.5	26.1	3.7
1280	**256**	**256**	**109.4**	**8.8**	**26.5**	**3.9**

C Discussion

C.1 Limitation and future work

This work has several limitations that suggest directions for future research. *i) Dynamic Environment.* Although AerialVLN-s (Liu et al., 2023) covers diverse urban and natural environments, its scenes remain static. Our agent's ability to reason and navigate in dynamic scenes, particularly in the presence of moving objects, therefore remains unverified. Future work could incorporate world models (Hafner et al., 2025) to anticipate environmental changes and guide spatial reasoning and action prediction. *ii) Spatial Memory.* Our spatial reasoning operates on voxel maps constructed from current observations, without explicitly accumulating spatial evidence across navigation steps. Although a recurrent unit incorporates navigation history into action prediction, historical spatial context is not directly available to the reasoning process. Constructing and maintaining a global 3D map (Deng et al., 2026) could enable reasoning over accumulated observations and previously explored regions, providing a promising direction for future research. *iii) Real-World Deployment.* Our experiments are conducted in simulation, and the method has not been validated on physical UAVs. Deploying it on real platforms would introduce challenges including imperfect state estimation, communication latency, and safety requirements in cluttered or human-shared spaces. Closing such sim-to-real transfer gap represents a promising direction for future research.

C.2 Social impact

This work introduces Hierarchical Volumetric Reasoning for Aerial Vision-Language Navigation. By grounding language instructions in multi-scale 3D volumetric representations, our approach improves navigation performance on the AerialVLN-s benchmark (Liu et al., 2023). These results highlight the value of explicit 3D spatial reasoning for instruction-guided aerial navigation. Such capabilities could support search and rescue, infrastructure inspection, and environmental monitoring by helping aerial agents interpret human instructions within their spatial context. Our work contributes to the algorithmic and perceptual foundations of aerial navigation, with the broader aim of supporting aerial agents that can assist people in complex environments.

C.3 Terms of use, privacy, and license

AerialVLN-s (Liu et al., 2023) is available for non-commercial research purposes.