

# History-Enhanced Two-Stage Transformer for Aerial Vision-and-Language Navigation

Xichen Ding<sup>1,2\*</sup>, Jianzhe Gao<sup>3\*</sup>, Cong Pan<sup>1,2</sup>, Wenguan Wang<sup>3†</sup>, Jie Qin<sup>1,2†</sup>

<sup>1</sup>College of Artificial Intelligence, Nanjing University of Aeronautics and Astronautics

<sup>2</sup>Key Laboratory of Brain-Machine Intelligence Technology, Ministry of Education, China

<sup>3</sup>The State Key Lab of Brain-Machine Intelligence, Zhejiang University

## Abstract

Aerial Vision-and-Language Navigation (AVLN) requires Unmanned Aerial Vehicle (UAV) agents to localize targets in large-scale urban environments based on linguistic instructions. While successful navigation demands both global environmental reasoning and local scene comprehension, existing UAV agents typically adopt mono-granularity frameworks that struggle to balance these two aspects. To address this limitation, this work proposes a History-Enhanced Two-Stage Transformer (HETT) framework, which integrates the two aspects through a coarse-to-fine navigation pipeline. Specifically, HETT first predicts coarse-grained target positions by fusing spatial landmarks and historical context, then refines actions via fine-grained visual analysis. In addition, a historical grid map is designed to dynamically aggregate visual features into a structured spatial memory, enhancing comprehensive scene awareness. Additionally, the CityNav dataset annotations are manually refined to enhance data quality. Experiments on the refined CityNav dataset show that HETT delivers significant performance gains, while extensive ablation studies further verify the effectiveness of each component.

**Code & Dataset** — <https://github.com/crotonyl/HETT>

## Introduction

Aerial Vision-and-Language Navigation (AVLN) is an emerging challenge in embodied AI, requiring Unmanned Aerial Vehicle (UAV) agents to identify and locate targets in outdoor environments given natural-language instructions (Fan et al. 2023; Liu et al. 2023c; Lee et al. 2024). Unlike indoor Vision-and-Language Navigation (VLN) tasks (Anderson et al. 2018a; Qi et al. 2020; Qiao et al. 2022), which are confined to limited action spaces, AVLN requires UAV agents to navigate in unstructured and large-scale aerial environments. This poses critical challenges in sustaining robust cross-modal alignment between vision and language throughout long and dynamic trajectories.

To achieve robust AVLN performance, UAV agents need to integrate adaptive decision-making with continuous environmental understanding (Gao et al. 2023; Chen et al. 2021;

\*These authors contributed equally.

†Corresponding authors.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

**Instruction:** One red car between a white car and a gray car in front of the BMTR building off Walsall Road.

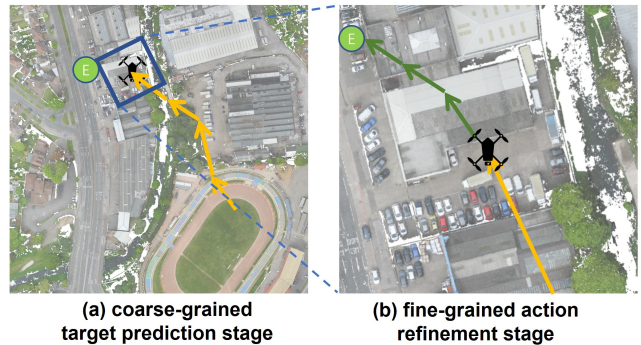


Figure 1: **Illustration** of HETT. Given a goal-oriented instruction, HETT operates in two stages: (a) our agent first leverages coarse-grained global features to predict a global target, providing high-level directional guidance for long-range navigation; (b) it then refines local actions on fine-grained local features to adapt to local observations.

Qiao et al. 2022; Fan et al. 2024). However, current state-of-the-art AVLN agents **1** *primarily employ mono-granularity frameworks, and the integration of global planning and local perception remains underexplored*. Specifically, local path planning approaches (An et al. 2025; Wang et al. 2022a; Su et al. 2025; Liu, Wang, and Yang 2024a) focus on fine-grained alignment between local visual observations and instruction semantics, predicting actions within a pre-defined action space. Global path planning approaches (Lee et al. 2024; Kong et al. 2024; Wang et al. 2025), in contrast, construct coarse-grained 2D spatial maps for target position prediction. While both paradigms contribute important capabilities to AVLN, they also present complementary limitations: local planning excels at dynamic adaptation but struggles with long-horizon reasoning due to its dependence on local perceptions, whereas global planning offers comprehensive spatial awareness but lacks the fine-grained visual understanding needed for precise localization. Furthermore, **2** *existing UAV agents fail to preserve historical details during long-term navigation planning*. Recent agents (Gao et al. 2024; Lee et al. 2024; Wang et al. 2023a)

typically project semantic masks onto a top-down map using UAV pose and depth information to represent historical context. However, their global environmental comprehension remains heavily dependent on semantic segmentation modules (Pan et al. 2023) such as GroundingDINO (Liu et al. 2023b) and SAM (Kirillov et al. 2023). This reliance fundamentally limits the capture of fine-grained visual details, potentially compromising overall scene understanding (Wang et al. 2023c; Liu et al. 2023a; Wang et al. 2022b). Moreover, **③** *a key limitation of existing AVLN datasets lies in suboptimal annotation quality*. As the field is still in an early stage, many datasets (Lee et al. 2024) rely on LLM-generated navigation annotations without rigorous manual review, introducing noise and inaccuracies into the training data.

To address these challenges, this work proposes a History-Enhanced Two-Stage Transformer (HETT) for AVLN, which integrates coarse-grained and fine-grained multi-modal information to bridge the gap between global planning and local perception. Motivated by **①**, HETT adopts a two-stage navigation policy that decomposes the navigation process into *coarse-grained target prediction* and *fine-grained action refinement*. As illustrated in Fig. 1, during the first stage (Fig. 1(a)), our agent leverages prior spatial landmarks and accumulated historical information to infer the target’s approximate location. Once approaching this predicted region, the agent enters the second stage (Fig. 1(b)), where detailed visual cues guide precise local movements. For **②**, our agent incorporates a historical grid map that encodes both temporal and spatial information of the globally visited environment. The map partitions the environment into uniformly sized grid cells, each storing fine-grained visual features based on their coordinates. This design enables persistent and structured historical memory across long navigation trajectories. In addition, regarding **③**, we perform a thorough manual refinement of the dataset annotations to mitigate the noise introduced by LLM-generated annotations and ensure the reliability of our evaluation.

Experiments on the CityNav benchmark validate the effectiveness of HETT, showing substantial improvements of **14.16%**, **10.75%**, and **18.00%** in SR across the validation and test sets. Ablation studies verify the contributions of our core components as well as the impact of dataset refinement.

## Related Work

**Aerial Vision-and-Language Navigation (AVLN).** The widespread adoption of UAVs drives extensive research in AVLN, where drones navigate outdoor environments based on language instructions and visual observations. Seminal works include AVDN, which provides manually collected dialog-based instructions for AVLN tasks (Fan et al. 2023), and CityNav, which enhances navigation by incorporating GPS-augmented target descriptions (Lee et al. 2024). Recent simulators further accelerate progress by constructing photorealistic 3D outdoor environments with full 6-DoF UAV control (Liu et al. 2023c; Wang et al. 2025).

Despite these developments, current AVLN agents still face several language-grounding challenges. Urban environments typically contain dense landmark distributions, irregular street layouts, and highly variable urban geometry, mak-

ing it difficult for agents to maintain stable cross-modal alignment over long navigation trajectories. In addition, most existing agents lack effective mechanisms for modeling historical visual-linguistic context (Gao et al. 2024), limiting their ability to resolve ambiguous or deferred references in multi-step navigation instructions. These limitations demonstrate the need for more advanced AVLN architectures that jointly integrate linguistic understanding with robust spatial reasoning in large-scale outdoor settings.

**Vision-and-Language Navigation (VLN).** VLN is a fundamental task in embodied AI in which agents navigate photorealistic scenes using natural-language instructions (Gao, Liu, and Wang 2025). Representative benchmarks include R2R (Anderson et al. 2018b), RxR (Ku et al. 2020), and CVDN (Thomason et al. 2020), all constructed within indoor household environments in the Matterport3D simulator (Chang et al. 2017). These datasets primarily focus on discrete action spaces and set the foundation for indoor VLN evaluation. The introduction of VLN-CE (Krantz et al. 2020) marks a major shift toward realism by converting topological trajectories into continuous action spaces, thereby reflecting real-world motion dynamics.

Traditional VLN agents rely on cross-modal attention mechanisms to align visual observations with textual commands. However, these agents often struggle to capture temporal dependencies, as they predominantly attend to the current observation while overlooking accumulated historical context (Wang et al. 2023b). HMT (Chen et al. 2021) introduces a hierarchical transformer that encodes the complete navigation history as sequential memory tokens. TD-STP (Zhao et al. 2022) extends HMT by incorporating a target prediction mechanism that enables agents to “imagine” future states. DUET (Chen et al. 2022) equips agents with topological map encoding to facilitate efficient global planning. Other agents (Georgakis et al. 2022; Huang et al. 2023; Wang et al. 2024) maintain a top-down semantic map to better capture the spatial layout and structural relations.

**Aerial Navigation.** AVLN builds upon the broader field of autonomous aerial navigation, which is traditionally organized around two complementary paradigms: global planning, where agents leverage environmental context for semantic goal inference, and local planning, where agents rely on immediate perceptual cues for reactive control.

Global navigation agents typically compute offline optimal routes from satellite imagery or Digital Elevation Models (DEMs). Early agents (Szczerba et al. 2000) employ sparse A\* search with spatial constraints to reduce computational overhead during long-distance route planning, while HGARL (Akshya et al. 2024) demonstrates that hybrid metaheuristic agents based on HHO effectively avoid local minima in obstacle-dense environments. However, global agents remain limited by their dependence on static scene maps. In contrast, local agents prioritize real-time reactivity using high-frequency perceptual observations. Hrabar’s stereo-vision agent (Hrabar 2008) achieves sub-meter obstacle avoidance in cluttered spaces through probabilistic roadmaps. Follow-up agents introduce artificial potential fields for dynamic obstacle avoidance and genetic-evolutionary strategies for optimizing 3D trajectories, sig-

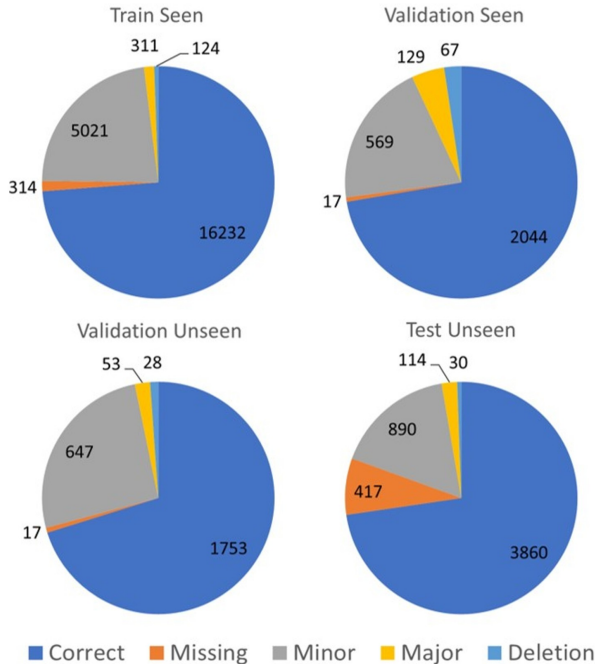


Figure 2: **Annotation Error Distribution** in the CityNav (Lee et al. 2024) Dataset. The chart reports the proportion of each error type in the original annotations, including “Missing” landmark references, “Minor” textual inconsistencies, “Major” landmark extraction failures, and “Deletion” cases that lack usable spatial grounding.

nificantly improving the agent’s real-time adaptability.

## Dataset

In AVLN, navigation instructions are commonly categorized into two paradigms: step-by-step instructions (e.g., “Take off, fly through the tower of cable bridge and down to the end of the road.”) and goal-oriented instructions (e.g., “The white car that is the sixth car in the fifth aisle of the One Stop parking lot.”) (Gao et al. 2023). The CityNav dataset (Lee et al. 2024) primarily utilizes goal-oriented instructions. Crucially, these instructions rely on pre-defined landmark information for target localization. Navigating unstructured urban environments under goal-oriented instructions necessitates a landmark-centric agent design, as precise localization is infeasible when relying solely on geometric cues or relative directions. Consequently, landmarks become essential spatial anchors that enable agents to position themselves and interpret high-level goal descriptions.

The original CityNav dataset contains 32K trajectories corresponding to natural language descriptions of approximately 5.8K objects such as buildings and cars. Its landmark annotations are initially generated using GPT-3.5 Turbo (Ouyang et al. 2022). However, this automated process introduced substantial errors due to the absence of human verification. For instance, the model failed to correctly associate the landmark “One Stop” with the description: “The white car that is the sixth car in the fifth aisle of the

Types	Train Seen	Val Seen	Val Unseen	Test Unseen
Missing	314	17	17	417
Minor	5021	647	569	890
Major	311	53	67	114
Deletion	124	28	129	30

Table 1: **Annotation Error Statistics** in the CityNav (Lee et al. 2024) Dataset. “Missing” indicates missing landmark references. “Minor” refers to spelling mistakes or other typos. “Major” denotes critical landmark extraction errors that misalign instructions with their intended targets. “Deletion” corresponds to instructions removed from the dataset due to lacking valid landmark references.

One Stop parking lot.” To address such inaccuracies, we performed a manual refinement of the annotations, with quantitative statistics presented in Fig. 2 and Table. 1. This ensures landmark correspondence for every instruction, thereby providing a reliable foundation for evaluation.

## Method

**Problem Formulation.** In the CityNav benchmark (Lee et al. 2024), a UAV agent navigates a 2D urban environment under the guidance of a natural-language instruction. At each time step, the agent receives an egocentric top-down RGB-D observation together with its current pose. The agent also has access to static geographic priors that provide polygonal boundary descriptions of landmarks mentioned in the instruction. Formally, given the instruction, the associated landmark priors, and the sequence of observations, the agent must generate a sequence of continuous control actions that drives it toward the instructed goal. A navigation episode is considered successful if the agent issues the [stop] action within 20 steps and the final predicted location falls within 20 m of the ground-truth target.

**Overview (Fig. 3).** HETT integrates multi-modal cues to produce an adaptive navigation policy with coherent long-horizon reasoning. Given an instruction, text embeddings  $\mathbf{E}$  are extracted, and the referenced landmark contours are encoded as spatial features  $\mathbf{L} \in \mathbb{R}^D$ . At each time step  $t$ , the agent obtains its observation and pose, which are encoded into visual features  $\mathbf{V}_t \in \mathbb{R}^D$  and pose features  $\mathbf{P}_t \in \mathbb{R}^D$ . Historical environmental information is preserved through a Historical Grid Map that aggregates past visual features into a spatial memory tensor  $\mathbf{F}_t \in \mathbb{R}^D$ . These components are jointly processed by a transformer to yield fused contextual representations combining linguistic guidance, landmark priors, visual perception, pose state, and accumulated memory. Based on these, HETT operates in two stages. The coarse-grained stage predicts a target location  $\mathbf{g}_t \in [0, 1]^2$  that provides high-level directional cues, while the fine-grained stage refines immediate movements through an action estimate  $a_t \in (-\pi, \pi]$  and a progress indicator  $r_t \in [0, 1]$ , enabling precise and adaptive navigation.

## Two-Stage Transformer Framework

To bridge the gap between long-horizon reasoning and fine-grained scene comprehension in AVLN, a Two-Stage Trans-

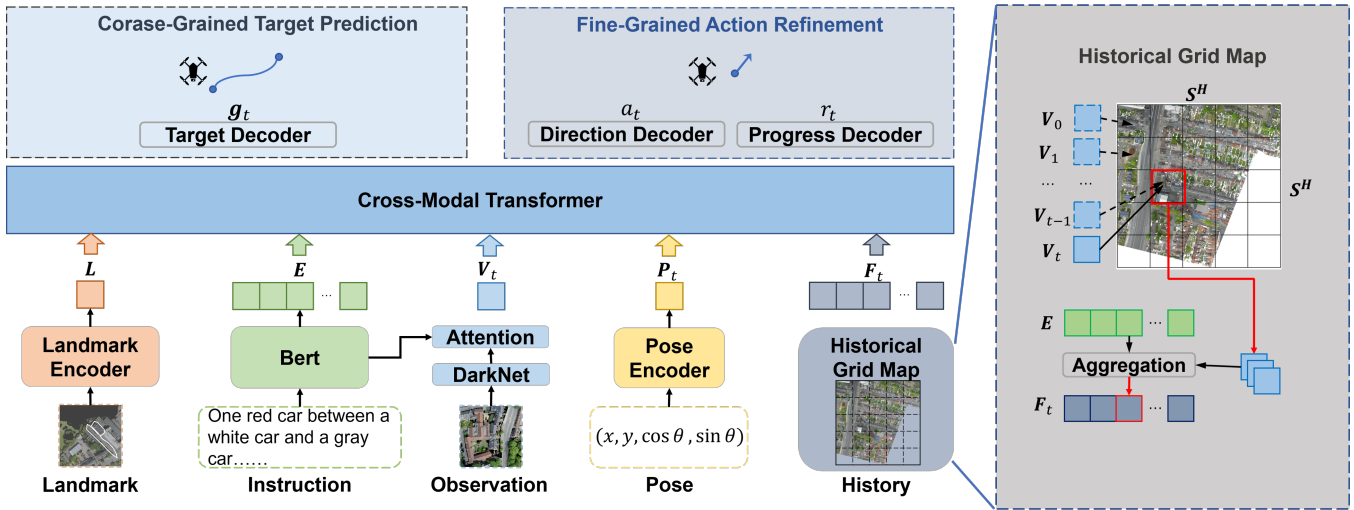


Figure 3: **Overview** of HETT. At time step  $t$ , five types of tokens (*i.e.*, the landmark, instruction, pose, history, and view tokens) are sent into the cross-modal transformer to predict action decisions. In the *Coarse-Grained Target Prediction* stage, our agent leverages the target prediction result  $g_t$  to guide navigation. In the *Fine-Grained Action Refinement* stage, the agent uses the local action estimation  $a_t$  to adjust immediate movements until the predicted progress  $r_t$  reaches threshold.

former Framework is introduced for AVLN. In the *Coarse-Grained Target Prediction* stage, the UAV agent infers an approximate target location by leveraging pre-defined landmark information together with a coarse historical grid map. In the subsequent *Fine-Grained Action Refinement* stage, the agent attends to local visual features to produce precise continuous actions for accurate trajectory execution.

**Coarse-Grained Target Prediction.** To rapidly narrow down the potential target region, the agent first builds a coarse spatial representation of the environment. Given the landmark priors referenced in the instruction, their polygonal contours are projected onto a top-down landmark map  $\mathcal{M}^L \in \mathbb{R}^{S^L \times S^L}$ , where  $S^L$  is the map resolution. The projected contours are encoded into a landmark embedding  $L$ :

$$L = \text{MLP}(\text{CNN}(\mathcal{M}^L)) \in \mathbb{R}^D. \quad (1)$$

At time step  $t$ , the agent extracts visual features  $V_t \in \mathbb{R}^D$  from the current RGB-D observation using a pretrained Darknet-53 encoder (Redmon and Farhadi 2018). To retain spatial and temporal context, a *historical grid map* aggregates the sequence of visual features into a structured spatial memory  $F_t \in \mathbb{R}^D$ . A *Coarse-Grained Target Prediction* module then estimates the normalized global target position by jointly reasoning over the instruction embedding  $E \in \mathbb{R}^{N \times d}$ , where  $N$  is the instruction length, the landmark embedding  $L$ , and the historical memory  $F_t$ . These are jointly fused through a multi-layer transformer (MLT):

$$G_t = \text{MLT}([E; L; F_t]) \in \mathbb{R}^D, \quad (2)$$

where  $[\cdot]$  denotes feature concatenation. The fused representation  $G_t$  captures the essential spatial and visual cues required for global target inference, and the normalized target coordinates  $g_t$  are obtained as:

$$g_t = \text{Softmax}(\text{MLP}(G_t)) \in [0, 1]^2. \quad (3)$$

**Fine-Grained Action Refinement.** Although the coarse-grained stage offers high-level directional guidance by estimating the target region, precise navigation further requires fine-grained alignment between linguistic instructions and visual observations. To achieve this, the agent enters the *Fine-Grained Action Refinement* stage, which emphasizes detailed scene interpretation and accurate motion control. A cross-modal attention mechanism is applied to derive instruction-aware visual embeddings:

$$V_t = \text{Attention}([E; O_t]) \in \mathbb{R}^D. \quad (4)$$

where  $O_t$  denotes the visual feature map extracted from the top-down RGB-D observation, and  $V_t$  represents the refined visual embedding aligned with the instruction semantics.  $V_t$  are subsequently fused with the landmark embedding  $L$ , historical spatial memory  $F_t$ , and pose embedding  $P_t$ :

$$R_t, A_t = \text{MLT}([E; L; F_t; V_t; P_t]) \in \mathbb{R}^D, \quad (5)$$

where  $R_t$  encodes the contextualized representation for action reasoning, and  $A_t$  serves as the basis for fine-grained action refinement in the subsequent control module. Based on these, fine-grained navigation actions are generated as:

$$r_t = \text{Sigmoid}(\text{MLP}(R_t)) \in [0, 1], \quad (6)$$

$$a_t = \text{Arctan2}(\text{Tanh}(\text{MLP}(A_t))) \in (-\pi, \pi], \quad (7)$$

where  $r_t$  provides an estimate of task completion progress for deciding when to terminate the episode, and  $a_t$  denotes the turning angle used for immediate motion adjustment. The agent repeatedly executes this process until the predicted progress  $r_t$  surpasses a predefined threshold or the maximum step limit is reached, enabling a balance between precise goal attainment and efficient trajectory completion.

Models	Validation Seen				Validation Unseen				Test Unseen			
	NE↓	SR↑	OSR↑	SPL↑	NE↓	SR↑	OSR↑	SPL↑	NE↓	SR↑	OSR↑	SPL↑
Random	222.3	0.00	1.15	0.00	223.0	0.00	0.90	0.00	208.8	0.00	1.44	0.00
Human	9.1	89.31	96.40	60.17	9.4	88.39	95.54	62.66	9.8	87.86	95.29	57.04
Seq2Seq	257.1	1.81	7.89	1.58	317.4	0.79	8.82	0.61	245.3	1.50	8.34	1.30
CMA	240.8	0.95	9.42	0.92	268.8	0.65	7.86	0.63	252.6	0.82	9.70	0.79
AerialVLN	65.6	9.77	23.77	8.64	81.8	6.79	17.91	5.73	64.1	8.09	19.13	5.91
MGP	53.0	16.93	29.90	14.38	73.8	8.35	17.91	7.07	86.1	10.90	20.24	9.94
HETT(Ours)	45.2	25.16	48.40	23.01	62.1	17.48	25.09	14.46	72.9	22.97	39.30	17.01
<b>HETT(Ours)*</b>	<b>37.2</b>	<b>31.09</b>	<b>51.86</b>	<b>25.76</b>	<b>51.3</b>	<b>19.10</b>	<b>34.78</b>	<b>16.70</b>	<b>40.4</b>	<b>28.90</b>	<b>49.56</b>	<b>23.79</b>

Table 2: **Quantitative results** of HETT. HETT(Ours)\* is trained and evaluated on the refined dataset.

## Historical Grid Map

Inspired by history-encoding mechanisms in indoor VLN agents (Chen et al. 2021; Wang et al. 2023c; Liu, Wang, and Yang 2024b), a *Historical Grid Map* is introduced to capture and organize the agent’s accumulated visual memories. The environment is discretized into a fixed  $S^H \times S^H$  grid covering the entire navigation region. At each time step  $t$ , the agent stores its fine-grained visual features and their corresponding spatial coordinates into the historical map  $\mathcal{M}_t^H$ :

$$\mathcal{M}_t^H = \mathcal{M}_{t-1}^H \cup [V_t, \mathbf{p}_t], \quad (8)$$

where  $\mathbf{p}_t$  represents the agent’s position.

These stored features are then assigned to their corresponding grid cells according to spatial coordinates, forming a structured grid feature set:

$$\mathcal{M}_{t,(x,y)}^H = \{\mathbf{m}_{t,j}\}_{j=1}^J, \quad (x, y) \in \{1, \dots, S^H\}^2, \quad (9)$$

where each  $\mathbf{m}_{t,j} \in \mathbb{R}^D$  denotes a visual feature whose spatial coordinate falls inside the grid cell indexed by  $(x, y)$ , and  $J$  is the number of features accumulated in that cell.

For each cell, a relevance matrix  $\mathbf{K}_{(x,y)}$  is computed between its feature set  $\mathcal{M}_{t,(x,y)}^H$  and instruction embedding  $\mathbf{E}$ :

$$\mathbf{K}_{(x,y)} = \text{Softmax}(\mathcal{M}_{t,(x,y)}^H \cdot \mathbf{E}^\top) \in [0, 1]^{J \times N}. \quad (10)$$

where  $N$  denotes the length of the instruction. Finally, the historical grid token at cell  $(x, y)$  is computed via a relevance-weighted aggregation:

$$\mathbf{F}_{t,(x,y)} = \sum_{j=1}^J K_{(x,y),j} \cdot \mathbf{m}_{t,j} \in \mathbb{R}^D, \quad (11)$$

where  $K_{(x,y),j}$  denotes the scalar relevance weight associated with feature  $\mathbf{m}_{t,j}$ . Aggregating across all grid cells yields the full structured spatial memory  $\mathbf{F}_t \in \mathbb{R}^D$ .

## Loss Function

Following prior works (Chen et al. 2022; Hong et al. 2022; Zhao et al. 2022), DAgger (Ross, Gordon, and Bagnell 2011) is adopted for policy training. To supervise the proposed two-stage framework, three dedicated loss functions are introduced. The first component is the coarse-grained target prediction loss:

$$\mathcal{L}^G = \sum_{t=1}^T \text{MSE}(\mathbf{g}_t, \mathbf{g}^{gt}), \quad (12)$$

where  $\mathbf{g}_t$  is the predicted normalized target coordinate and  $\mathbf{g}^{gt}$  denotes the corresponding ground-truth target location. Similarly, the action loss  $\mathcal{L}^A$  and progress loss  $\mathcal{L}^R$  are formulated using the ground-truth action  $a_t^{gt}$  and ground-truth progress  $r_t^{gt}$ , respectively.

The total training objective is expressed as:

$$\mathcal{L} = \alpha_1 \mathcal{L}^G + \alpha_2 \mathcal{L}^A + \alpha_3 \mathcal{L}^R, \quad (13)$$

where  $\alpha_{\{1,2,3\}}$  are weighting coefficients that balance the contributions of the three losses.

## Implementation Details

Our model is implemented in PyTorch and trained on four 24GB RTX A5000 GPUs for 20 epochs, with a batch size of 2, a learning rate of 1e-4, and AdamW optimizer. The grid size  $d$  is set to 5.  $\alpha_1, \alpha_2, \alpha_3$  are set to 2.0, 1.5, 0.1.

## Experiment

### Experimental Setup

**Dataset Preparation.** The navigation instructions utilized in our experiments are derived from the original and refined CityNav (Lee et al. 2024) dataset, which comprises 32,326 natural language descriptions paired with human demonstration trajectories, all collected by crowd-sourcing. Each language description is rich in detail, encompassing information such as landmarks, regions and objects, etc. The drone images are taken from SensatUrban, which gathers orthographic projections and depth maps of 13 blocks in Birmingham and 33 blocks in Cambridge. These data are utilized to simulate the RGBD input an actual drone would acquire during navigation. Additionally, the geometric outlines of landmarks within the geographic information database are obtained from CityRefer (Miyanishi et al. 2023), providing essential information for the navigation tasks.

**Evaluation Metrics.** To comprehensively evaluate the navigation performance of our HETT, we adopt four standard metrics commonly used in the field: Navigation Error (NE), Success Rate (SR), Oracle Success Rate (OSR), and Success weighted by Path Length (SPL).

- **Navigation Error (NE):** This metric quantifies the Euclidean distance between the final position of the UAV agent and the ground truth target location.



Figure 4: **Visualization results** for the two-stage navigation. Initial predictions in the coarse-grained stage may exhibit target drift or premature stopping. The fine-grained refinement stage corrects these deviations and steers the agent toward the ground-truth target. Three representative cases are visualized to illustrate this behavior.

- Success Rate (SR): This metric calculates the percentage of episodes where the UAV agent terminates its navigation within a pre-defined success threshold.
- Oracle Success Rate (OSR): This metric assesses whether the agent’s trajectory at any point approaches the target within the success threshold.
- Success weighted by Path Length (SPL): This metric is the success rate weighted by the ratio of the reference path length to the actual path length traveled by the agent.

### Overall Performance

**Quantitative Results.** We compare HETT with several baseline UAV agents on the CityNav benchmark. As shown in Table 2, HETT achieves consistent improvements across all metrics and data splits. In terms of SR, it surpasses the strongest baseline MGP by **8.23%**, **9.13%**, and **12.07%** on the validation-seen, validation-unseen, and test-unseen sets, respectively. Beyond SR, HETT also shows notable gains on the test-unseen split, reducing NE by 13.2 m, improving OSR by 19.06%, and increasing SPL by 7.07%. To further examine the benefit of our refined annotations, we train a variant denoted as HETT\*, where \* indicates that both training and evaluation are performed on the refined dataset. HETT\* obtains additional SR improvements of **5.93%**, **1.62%**, and **5.93%** across the three splits, along with corresponding reductions in NE and increases in SPL. These consistent improvements clearly indicate that the proposed HETT framework yields more accurate and stable UAV navigation, while the refined dataset annotations further enhance the reliability of supervision signals for AVLN.

**Qualitative Results.** Fig. 4 illustrates how HETT executes the **two-stage navigation policy**. In the **coarse-grained target prediction** stage, the agent first moves toward the estimated target region based on global spatial cues. Once it reaches this vicinity, the agent switches to the **fine-grained action refinement** stage, where localized perception and historical context are leveraged to perform precise trajectory adjustments. As shown in cases (a) and (b), the refined actions correct deviations accumulated during the coarse prediction stage, while in case (c), they guide the agent

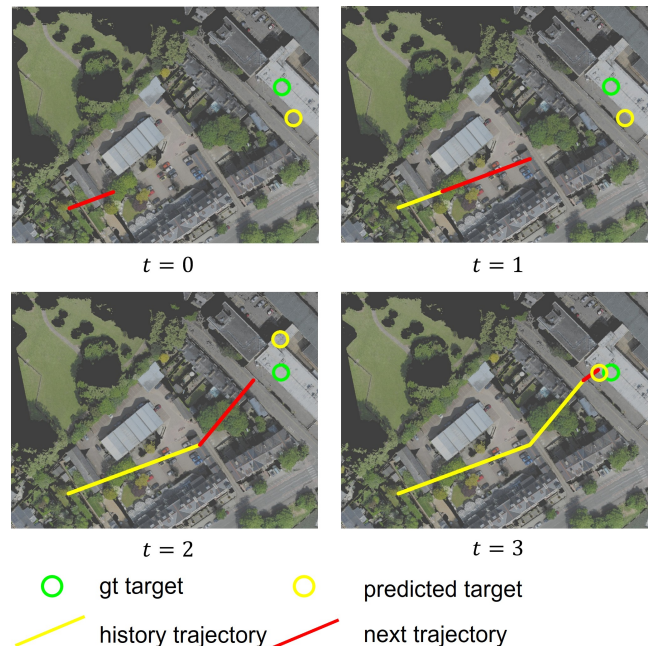


Figure 5: **Visualization** of the Coarse-Grained Target Prediction stage. As navigation proceeds, the predicted position steadily converges toward the ground-truth location.

to achieve accurate final alignment with the ground-truth target. These qualitative examples clearly demonstrate the adaptiveness and effectiveness of our strategy.

To further assess the impact of the **Historical Grid Map**, Fig. 5 visualizes the evolution of target predictions across navigation steps. The predicted target position transitions from an initially uncertain estimate to a stable convergence around the ground-truth location as more observations are accumulated. This progression highlights that this map successfully incorporates temporal visual cues, thereby enhancing global localization accuracy.

#	Dataset	Two-Stage	History	Validation Seen			Validation Unseen		
				NE↓	SR↑	OSR↑	NE↓	SR↑	OSR↑
1	-	-	-	49.66	19.42	33.23	65.84	9.84	20.24
2	✓	-	-	39.72	24.98	42.59	56.32	13.27	29.11
3	✓	✓	-	40.47	26.19	48.58	56.33	14.68	34.71
4	✓	-	✓	<b>36.38</b>	29.31	41.78	52.99	15.28	26.88
5	✓	✓	✓	37.24	<b>31.09</b>	<b>51.86</b>	<b>51.34</b>	<b>19.10</b>	<b>34.78</b>

Table 3: **Ablated results** of the main components on the CityNav dataset.

Agent	Validation Seen			Validation Unseen		
	NE↓	SR↑	OSR↑	NE↓	SR↑	OSR↑
AerialVLN	65.6	9.77	<b>23.77</b>	81.8	6.79	17.41
<b>AerialVLN*</b>	<b>54.2</b>	<b>12.38</b>	22.22	<b>65.9</b>	<b>9.12</b>	<b>17.72</b>
MGP	53.0	16.93	29.90	73.8	8.35	17.91
<b>MGP*</b>	<b>48.1</b>	<b>19.17</b>	<b>35.51</b>	<b>66.5</b>	<b>10.47</b>	<b>28.06</b>

Table 4: **Ablated results** of dataset refinement. \* denotes models trained and evaluated on the refined CityNav dataset.

## Ablation Studies

To validate the contributions of key components of our HETT, we perform systematic ablation studies. The results presented in Tables 3, 4, and 5 quantitatively assess the effectiveness of individual modules.

**Effectiveness of Each Component.** We begin by evaluating the contribution of each core component in our framework. As shown in Table 3, incorporating the **refined dataset** (Row #2) leads to a clear performance improvement over the baseline in Row #1, increasing SR from 19.42% to **24.98%** on the validation-seen split and from 9.84% to **13.27%** on the validation-unseen split. The effect of the **two-stage navigation policy** is examined by comparing Row #3 with Row #2. With SR improving from 24.98% to **26.19%** on the validation-seen split and from 13.27% to **14.68%** on the validation-unseen split, the two-stage navigation policy further enhances both long-horizon reasoning and local action accuracy. Moreover, the **historical grid map** provides the most substantial gains. Compared with Row #2, adding the historical grid map (Row #4) improves SR from 24.98% to **29.31%** on the validation-seen split and from 13.27% to **15.28%** on the validation-unseen split. When combined with the two-stage policy (Row #5 vs. Row #3), SR increases from 26.19% to **31.09%** on validation seen and from 14.68% to **19.10%** on validation unseen. These show that each component contributes meaningfully to overall performance.

**Analysis of Dataset Refinement.** We further evaluate the influence of dataset refinement on other UAV agents. As shown in Table 4, both MGP and AerialVLN trained on the refined annotations achieve consistently higher performance than their original counterparts. On the Validation Seen split, AerialVLN improves from 9.77% to **12.38%**, while MGP rises from 16.93% to **19.17%**. Similar gains are observed on the Validation Unseen split, where AerialVLN increases from 6.79% to **9.12%**, and MGP from 8.35% to **10.47%**. These results further validate the effectiveness of the refined

$S^H \times S^H$	Validation Seen			Validation Unseen		
	NE↓	SR↑	OSR↑	NE↓	SR↑	OSR↑
0×0	40.47	26.19	48.58	56.33	14.68	34.71
3×3	39.92	27.87	<b>52.11</b>	53.98	18.61	<b>36.41</b>
<b>5×5</b>	<b>37.24</b>	<b>31.09</b>	51.86	<b>51.34</b>	<b>19.10</b>	34.78
7×7	37.28	27.45	41.26	54.43	17.87	32.81

Table 5: **Ablated results** of the grid size ( $S^H \times S^H$ ) in Historical Grid Map.

dataset and highlight the importance of high-quality landmark annotations for robust AVLN training and evaluation.

**Analysis of Historical Grid Map.** Furthermore, we ablate the historical grid map’s grid size. The results are summarized in Table 5, where a  $0 \times 0$  grid indicates the absence of the historical grid map. The experimental results show that the model with a  $3 \times 3$  grid size achieves the optimal OSR of **52.11%**, **36.41%** on the validation seen and unseen set. When the grid size increases to  $5 \times 5$ , the model attains the best overall performance with SR rising to **31.09%** and **19.10%** respectively on the validation seen and unseen set; however, when the grid size reaches  $7 \times 7$ , both SR and OSR on the validation unseen set decrease, likely due to excessive features interfering with the model’s ability to extract discriminative navigation cues. Thus, we select the  $5 \times 5$  configuration as the optimal grid size for our final model.

## Conclusion

In this paper, we propose a History-Enhanced Two-Stage Transformer (HETT) for AVLN. HETT adopts a coarse-to-fine navigation paradigm that decomposes the navigation process into a two-stage navigation policy: coarse-grained target prediction and fine-grained action refinement. Moreover, the historical grid map further enhances the agent’s spatial awareness by maintaining structured environmental memory during navigation. Compared with previous UAV agents, our HETT integrates both coarse-grained environmental perception and fine-grained visual cues, thus enabling more accurate navigation results. In addition, we conduct manual refinement of the CityNav annotations, providing a more reliable benchmark for AVLN. Extensive experiments demonstrate the effectiveness of our HETT. one limitation of HETT is its dependency on pre-defined information. Future work will investigate online environment mapping to enhance navigation robustness.

## Acknowledgements

This work was partially supported by the National Natural Science Foundation of China (No. U25A20533, No. 62276129, No. 62372405), the Natural Science Foundation of Jiangsu Province (No. BK20250082, No. BK20251387), the Fundamental Research Funds for the Central Universities (No. NE2025010, 226-2025-00057), Zhejiang Provincial Natural Science Foundation of China (No. LD25F020001), the Postdoctoral Fellowship Program of CPSF under Grant Number GZC20252739, the Jiangsu Funding Program for Excellent Postdoctoral Talent (No. 2025ZB489) and CIPSC-SMP-Zhipu Large Model Cross-Disciplinary Fund.

## References

- Akshya, J.; Manochitra, M.; Indu Poornima, R.; Jeevitha, T.; Saravanakumar, M.; Sundarajan, M.; and Choudhry, M. D. 2024. Metaheuristic Optimization for Path Planning in UAV Networks for Long-Distance Inspection Tasks. In *SMART*.
- An, D.; Wang, H.; Wang, W.; Wang, Z.; Huang, Y.; He, K.; and Wang, L. 2025. ETPNav: Evolving Topological Planning for Vision-Language Navigation in Continuous Environments. *TPAMI*.
- Anderson, P.; Wu, Q.; Teney, D.; Bruce, J.; Johnson, M.; Sünderhauf, N.; Reid, I.; Gould, S.; and van den Hengel, A. 2018a. Vision-and-Language Navigation: Interpreting Visually-Grounded Navigation Instructions in Real Environments. In *CVPR*.
- Anderson, P.; Wu, Q.; Teney, D.; Bruce, J.; Johnson, M.; Sünderhauf, N.; Reid, I.; Gould, S.; and van den Hengel, A. 2018b. Vision-and-Language Navigation: Interpreting Visually-Grounded Navigation Instructions in Real Environments. In *CVPR*.
- Chang, A.; Dai, A.; Funkhouser, T.; Halber, M.; Niebner, M.; Savva, M.; Song, S.; Zeng, A.; and Zhang, Y. 2017. Matterport3D: Learning from RGB-D Data in Indoor Environments. In *3DV*.
- Chen, S.; Guhur, P.-L.; Schmid, C.; and Laptev, I. 2021. History Aware multimodal Transformer for Vision-and-Language Navigation. In *NeurIPS*.
- Chen, S.; Guhur, P.-L.; Tapaswi, M.; Schmid, C.; and Laptev, I. 2022. Think Global, Act Local: Dual-scale Graph Transformer for Vision-and-Language Navigation. In *CVPR*.
- Fan, S.; Liu, R.; Wang, W.; and Yang, Y. 2024. Navigation Instruction Generation with BEV Perception and Large Language Models. In *ECCV*.
- Fan, Y.; Chen, W.; Jiang, T.; Zhou, C.; Zhang, Y.; and Wang, X. E. 2023. Aerial Vision-and-Dialog Navigation. In *ACL*.
- Gao, C.; Peng, X.; Yan, M.; Wang, H.; Yang, L.; Ren, H.; Li, H.; and Liu, S. 2023. Adaptive Zone-aware Hierarchical Planner for Vision-Language Navigation. In *CVPR*.
- Gao, J.; Liu, R.; and Wang, W. 2025. 3D Gaussian Map with Open-Set Semantic Grouping for Vision-Language Navigation. In *ICCV*.
- Gao, Y.; Wang, Z.; Jing, L.; Wang, D.; Li, X.; and Zhao, B. 2024. Aerial Vision-and-Language Navigation via Semantic-Topo-Metric Representation Guided LLM Reasoning. arXiv:2410.08500.
- Georgakis, G.; Schmeckpeper, K.; Wanchoo, K.; Dan, S.; Mitsakaki, E.; Roth, D.; and Daniilidis, K. 2022. Cross-modal Map Learning for Vision and Language Navigation. In *CVPR*.
- Hong, Y.; Wang, Z.; Wu, Q.; and Gould, S. 2022. Bridging the Gap Between Learning in Discrete and Continuous Environments for Vision-and-Language Navigation. In *CVPR*.
- Hrabar, S. 2008. 3D path planning and stereo-based obstacle avoidance for rotorcraft UAVs. In *IROS*.
- Huang, C.; Mees, O.; Zeng, A.; and Burgard, W. 2023. Visual Language Maps for Robot Navigation. In *ICRA*.
- Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; Dollár, P.; and Girshick, R. 2023. Segment Anything. arXiv:2304.02643.
- Kong, X.; Chen, J.; Wang, W.; Su, H.; Hu, X.; Yang, Y.; and Liu, S. 2024. Controllable Navigation Instruction Generation with Chain of Thought Prompting. In *ECCV*.
- 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 Baldrige, J. 2020. Room-Across-Room: Multilingual Vision-and-Language Navigation with Dense Spatiotemporal Grounding. In *EMNLP*.
- 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:2406.14240.
- Liu, R.; Wang, W.; and Yang, Y. 2024a. Vision-Language Navigation with Energy-Based Policy. In *NeurIPS*.
- Liu, R.; Wang, W.; and Yang, Y. 2024b. Volumetric Environment Representation for Vision-Language Navigation. In *CVPR*.
- Liu, R.; Wang, X.; Wang, W.; and Yang, Y. 2023a. Bird’s-Eye-View Scene Graph for Vision-Language Navigation. In *ICCV*.
- Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Li, C.; Yang, J.; Su, H.; Zhu, J.; et al. 2023b. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. arXiv preprint arXiv:2303.05499.
- Liu, S.; Zhang, H.; Qi, Y.; Wang, P.; Zhang, Y.; and Wu, Q. 2023c. AerialVLN: Vision-and-language Navigation for UAVs. In *ICCV*.
- Miyanishi, T.; Kitamori, F.; Kurita, S.; Lee, J.; Kawanabe, M.; and Inoue, N. 2023. CityRefer: Geography-aware 3D Visual Grounding Dataset on City-scale Point Cloud Data. In *NeurIPS*.
- Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; Schulman, J.; Hilton, J.; Kelton, F.; Miller, L.; Simens,

M.; Asbell, A.; Welinder, P.; Christiano, P. F.; Leike, J.; and Lowe, R. 2022. Training language models to follow instructions with human feedback. In *NeurIPS*.

Pan, C.; He, Y.; Peng, J.; Zhang, Q.; Sui, W.; and Zhang, Z. 2023. BAEFormer: Bi-Directional and Early Interaction Transformers for Bird’s Eye View Semantic Segmentation. In *CVPR*.

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*.

Redmon, J.; and Farhadi, A. 2018. YOLOv3: An Incremental Improvement. arXiv:1804.02767.

Ross, S.; Gordon, G.; and Bagnell, D. 2011. A reduction of imitation learning and structured prediction to no-regret online learning. In *AISTATS*.

Su, Y.; An, D.; Chen, K.; Yu, W.; Ning, B.; Ling, Y.; Huang, Y.; and Wang, L. 2025. Learning Fine-Grained Alignment for Aerial Vision-Dialog Navigation. *AAAI*.

Szczerba, R.; Galkowski, P.; Glicktein, I.; and Ternullo, N. 2000. Robust algorithm for real-time route planning. *TAES*.

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

Wang, H.; Liang, W.; Shen, J.; Van Gool, L.; and Wang, W. 2022a. Counterfactual Cycle-Consistent Learning for Instruction Following and Generation in Vision-Language Navigation. In *CVPR*.

Wang, H.; Liang, W.; Van Gool, L.; and Wang, W. 2022b. Towards versatile embodied navigation. In *NeurIPS*.

Wang, H.; Liang, W.; Van Gool, L.; and Wang, W. 2023a. DREAMWALKER: Mental Planning for Continuous Vision-Language Navigation. In *ICCV*.

Wang, H.; Wang, W.; Liang, W.; Hoi, S.; Shen, J.; and Gool, L. 2023b. Active Perception for Visual-Language Navigation. *IJCV*.

Wang, X.; Wang, W.; Shao, J.; and Yang, Y. 2024. Learning to Follow and Generate Instructions for Language-Capable Navigation. *TPAMI*.

Wang, X.; Yang, D.; wang, z.; Kwan, H.; Chen, J.; wu, w.; Li, H.; Liao, Y.; and Liu, S. 2025. Towards Realistic UAV Vision-Language Navigation: Platform, Benchmark, and Methodology. In Yue, Y.; Garg, A.; Peng, N.; Sha, F.; and Yu, R., eds., *ICLR*.

Wang, Z.; Li, X.; Yang, J.; Liu, Y.; and Jiang, S. 2023c. Gridmm: Grid memory map for vision-and-language navigation. In *CVPR*.

Zhao, Y.; Chen, J.; Gao, C.; Wang, W.; Yang, L.; Ren, H.; Xia, H.; and Liu, S. 2022. Target-Driven Structured Transformer Planner for Vision-Language Navigation. In *ACM MM*.