arXiv:2610.41446v1 [cs.CV] 30 Oct 2026

Beyond Trial-and-Error: A Geospatial-Guided Active Reasoning Framework for Robust Aerial Vision-and-Language Navigation

Jiawei Ge[1,2] Wenkang Gong[1] Jiuxin Cao[1] Bo Liu[1] Lei Cui[2] Haoran Wang[2]
Xuelin Zhu[3] Biwei Cao[4] Chang Liu[1] Chen Feng[5] Ioannis Patras[6]

[1]Southeast University [2]Institute of AI for Industries, Chinese Academy of Sciences
[3]The Hong Kong Polytechnic University [4]Hohai University
[5]Queen's University Belfast [6]Queen Mary University of London

0jiawei.ge0@gmail.com

Abstract

Aerial Vision-and-Language Navigation (VLN) requires UAVs to locate language-specified goals in large-scale urban environments, where objects often appear small and visually ambiguous from high-altitude viewpoints. Existing methods largely follow a reactive paradigm common in indoor VLN, repeatedly updating navigation states from local observations while underutilizing the geospatial structure that remains stable during navigation. As a result, visually similar distractors can cause a reactive goal search to repeatedly reinforce locally plausible goal hypotheses based on passively accumulated evidence, leading to unreliable localization. To address this, we propose GeoReasoner, a robust framework that shifts the paradigm of Aerial VLN from reactive trial-and-error to geospatial-guided active reasoning. Specifically, GeoReasoner first performs pre-search geospatial reasoning to derive region-level priors and constrain the goal hypothesis space. Building on these, it further conducts hypothesis-driven exploration to actively acquire discriminative evidence and refine competing goal hypotheses. Finally, consistency-aware goal verification adversarially assesses complementary evidence sources before stopping, hence reducing premature or erroneous goal confirmations. Experiments on the real-world CityNav benchmark demonstrate that GeoReasoner sets a new SOTA among zero-shot Aerial VLN methods while competing favorably with supervised approaches.

1. Introduction

Aerial Vision-and-Language Navigation (Aerial VLN) [16, 19] is an emerging embodied navigation task that aims to guide unmanned aerial vehicles (UAVs) to language-specified goals in large-scale outdoor environments. To reach the goal, the UAV must understand geospatial descriptions from an aerial perspective and reason about relations among landmarks, roads, regions, and surrounding contextual objects. This capability is essential for practical UAV applications, including logistics, emergency response, and traffic monitoring [9, 13, 25].

Despite recent progress, existing Aerial VLN methods [3, 10, 16, 19, 30] **still largely follow a reactive goal-search paradigm commonly adopted in indoor VLN** [1, 15]: the UAV agent first observes the current scene, updates its internal navigation state or spatial memory from newly acquired visual evidence, and then selects the next action or waypoint. This paradigm is reasonable in indoor environments, where the navigable space is limited, external spatial priors are often unavailable, and target locations are largely inferred from local observations. However, city-scale aerial scenes are much larger and more open-ended, while high-altitude views make local visual observations more ambiguous. In such settings, the relatively stable geographic structure of urban environments tends to provide more reliable relational cues for goal localization.

Yet reactive navigation driven primarily by local evidence underutilizes these geospatial cues, leaving multiple candidates locally plausible from partial observations and causing goal localization to degenerate into trial-and-error. More importantly, this reactive paradigm relies on evidence passively accumulated along the navigation trajectory, rather than actively exploiting geospatial cues to constrain and compare competing hypotheses. As a result, the agent may become trapped by a locally plausible distractor as seemingly supportive evidence continues to accumulate and reinforce the wrong hypothesis. Such reinforcement can persist even when the hypothesis conflicts with the broader geospatial relations implied by the instruction, eventually leading to premature or erroneous stopping.



Figure 1. Overview of our framework. Unlike conventional reactive Aerial VLN methods that search for the goal through trial-and-error from passively acquired observations, GeoReasoner exploits geospatial knowledge as a prior before search to constrain the goal hypothesis space, actively acquires discriminative evidence to resolve competing goal hypotheses, and verifies multi-source consistency before stopping.

To address these issues, we propose a shift from reactive trial-and-error to **geospatial-guided active reasoning** for Aerial VLN, as shown in Figure 1. The key idea is to treat geospatial knowledge **as an actionable prior** that constrains the goal hypothesis space before searching, guides informative evidence acquisition during navigation, and supports reliable goal verification before stopping. Based on this idea, we introduce **GeoReasoner**, a geospatial-guided active reasoning framework that organizes goal search in Aerial VLN into three stages: Pre-search Geospatial Reasoning, Hypothesis-Driven Active Exploration, and Consistency-Aware Goal Verification.

Concretely, GeoReasoner first parses language-derived spatial constraints and transforms them into region-level geospatial priors, allowing the UAV to constrain the goal hypothesis space before goal-directed exploration. During navigation, rather than passively evaluating local candidates, the agent performs hypothesis-driven active exploration: it reasons over candidate goal regions, actively seeks discriminative evidence to assess and refine competing hypotheses under geospatial constraints, then progressively localizes the goal through iterative observation and reasoning. Before stopping, GeoReasoner employs an adversarial debate mechanism to verify whether the predicted goal is consistently supported by the instruction, geospatial structure, visual observations, and navigation history. The resulting assessment further informs whether additional exploration or re-planning is necessary. In this way, GeoReasoner establishes a geospatial-guided active reasoning paradigm for Aerial VLN, resolving ambiguous goal hypotheses while improving the reliability of goal localization in complex urban environments.

In summary, our contributions include:

- We introduce **GeoReasoner**, a geospatial-guided active reasoning framework that recasts Aerial VLN from reactive trial-and-error into active goal disambiguation, making geospatial knowledge actionable through navigation.
- We develop the **Pre-search Geospatial Reasoning** and **Hypothesis-Driven Active Exploration** schemes, which transform geospatial constraints into competing goal hypotheses and shift goal localization from passive observation accumulation to hypothesis-driven evidence seeking.
- We introduce a **Consistency-Aware Goal Verification** mechanism based on adversarial debate that allows stopping when the predicted goal is sufficiently supported by multiple complementary evidence sources, encouraging decisions based on consistent evidence rather than isolated cues.
- Extensive evaluations on CityNav show that GeoReasoner achieves SOTA **zero-shot** performance, consistently improving navigation success and localization robustness while narrowing the gap to supervised methods.

2. Related Work

Vision-and-Language Navigation (VLN). In this task, embodied agents are required to ground natural-language instructions in visual observations and navigate toward specified goals [1]. Existing methods have progressively improved long-horizon navigation through historical context and structured spatial representations, such as HAMT [5] and DUET [6]. Recent approaches further leverage large vision-language models to enhance semantic understanding and instruction-guided decision making in navigation [11, 28, 31]. These advances have substantially strengthened instruction grounding, spatial reasoning, and long-horizon decision making in VLN, providing a solid foundation for

extending language-guided navigation to aerial platforms.

Aerial Vision-and-Language Navigation. Different from conventional VLN, Aerial VLN operates from high-altitude viewpoints over large-scale urban environments, where targets are often visually ambiguous while instructions involve broader geographic relations. AerialVLN [19] first establishes the task in continuous 3D urban environments, while CityNav [16] further advances toward real-world city-scale navigation, where goals are often described through landmarks and geographic relations. Building on them, subsequent methods progressively strengthen planning and reasoning capabilities. CityNavAgent [30] and HTNav [10] strengthen long-horizon planning through hierarchical subgoals and multi-level decision making, respectively. FlightGPT [3] enhances semantic reasoning with vision-language models, and HETT [8] incorporates historical context into coarse-to-fine navigation. Despite advances, most Aerial VLN methods still inherit the observation-driven paradigm of indoor VLN, making limited use of relatively stable geospatial information and localizing goals from passively acquired observations. More recent efforts have moved toward exploiting such spatial cues more explicitly: GeoNav [27] incorporates geographic priors for coarse-to-fine navigation, while LookasideVLN [22] leverages directional relations for more informed path planning. However, their spatial reasoning mainly determines *where to navigate*, while evidence is obtained as a consequence of navigation rather than deliberately acquired to distinguish competing goal hypotheses. This passive acquisition can expose the UAV agent to accumulated distractors or ambiguous observations, making goal search error-prone. Our proposed GeoReasoner also exploits geospatial priors, but goes beyond navigation assistance by using them to form *competing goal hypotheses*, identify *what evidence is needed* and decide *where to acquire it*.

MLLM-based Spatial Reasoning. Recently, MLLMs have progressively evolved from visual understanding toward explicit spatial reasoning. To enhance 3D spatial perception, ShapeLLM [23] and SpatialLLM [21] introduce geometry-aware representations and spatially informed designs. Building on this capability, Thinking in Space [29] and Liu et al. [17] further strengthen spatial reasoning through explicit cognitive maps and cross-view correspondences. Moving beyond spatial understanding, RingMoAgent [12] extends MLLM reasoning to embodied decision making and long-horizon navigation. Despite these advances, existing benchmarks still reveal limitations in complex spatial relations and orientation [7, 26], highlighting the need for more explicit and reliable spatial reasoning.

3. Method

3.1. Overview

Problem formulation. We follow the standard Aerial VLN setting of CityNav [16]. Given a natural-language instruction $\mathcal{I}$, a UAV is required to localize the referred goal in a large-scale urban environment. At each timestep t, the agent receives a top-down RGB-D observation $\mathcal{O}_t = \{\mathcal{O}_t^{\text{rgb}}, \mathcal{O}_t^{\text{d}}\}$ and its current pose $\mathbf{p}_t = (x_t, y_t, z_t, \theta_t, \psi_t)$, where (x_t, y_t, z_t) denotes the 3D position and (θ_t, ψ_t) denote the pitch and yaw angles. The agent selects an action $a_t \in \mathcal{A}$, where $\mathcal{A} = \{\texttt{forward}, \texttt{left}, \texttt{right}, \texttt{ascend}, \texttt{descend}, \texttt{stop}\}$. Additionally, the agent has access to lightweight geospatial information such as geographic entity names and their spatial contours. Since the absolute goal location is unavailable, the agent must infer it from the instruction, geospatial structure, and sequential observations. An episode is considered successful when the agent issues `stop` within the prescribed distance threshold of the ground-truth goal.

GeoReasoner framework. As illustrated in Figure 2, GeoReasoner formulates goal search as a three-stage geospatial-guided active reasoning process. First, *Pre-search Geospatial Reasoning* (Section 3.2) extracts fine-grained spatial constraints from the instruction and grounds them using the available geospatial information to construct region-level goal hypotheses, before initiating goal-directed exploration. Second, *Hypothesis-Driven Active Exploration* (Section 3.3) verifies these regions in ascending plausibility order, instantiates representative object-level hypotheses within feasible regions, and actively acquires discriminative evidence for unresolved candidates. Finally, *Consistency-Aware Goal Verification* (Section 3.4) jointly evaluates linguistic, geospatial, visual, and historical evidence to determine whether to stop, trigger targeted exploration when evidence remains insufficient, or reject conflicting hypotheses and initiate re-planning.

3.2. Pre-search Geospatial Reasoning

Navigation instructions often intertwine geographic landmarks with fine-grained contextual cues surrounding the goal. However, these cues differ in spatial granularity, and treating them as a single unit may introduce ambiguity rather than improve goal localization. To address this, GeoReasoner establishes geospatial priors before visual search while preserving contextual cues for subsequent verification. The resulting feasible regions are organized into competing region-level hypotheses to guide active exploration.

Fine-grained Constraint Disentanglement. To distinguish geographic constraints and visually grounded contextual relations, we first employ the LLM to decompose the navigation instruction $\mathcal{I}$ into

$$\mathcal{P} = (g, \mathcal{L}, \mathcal{C}, \mathcal{S}_g, \mathcal{S}_l)\,, \tag{1}$$

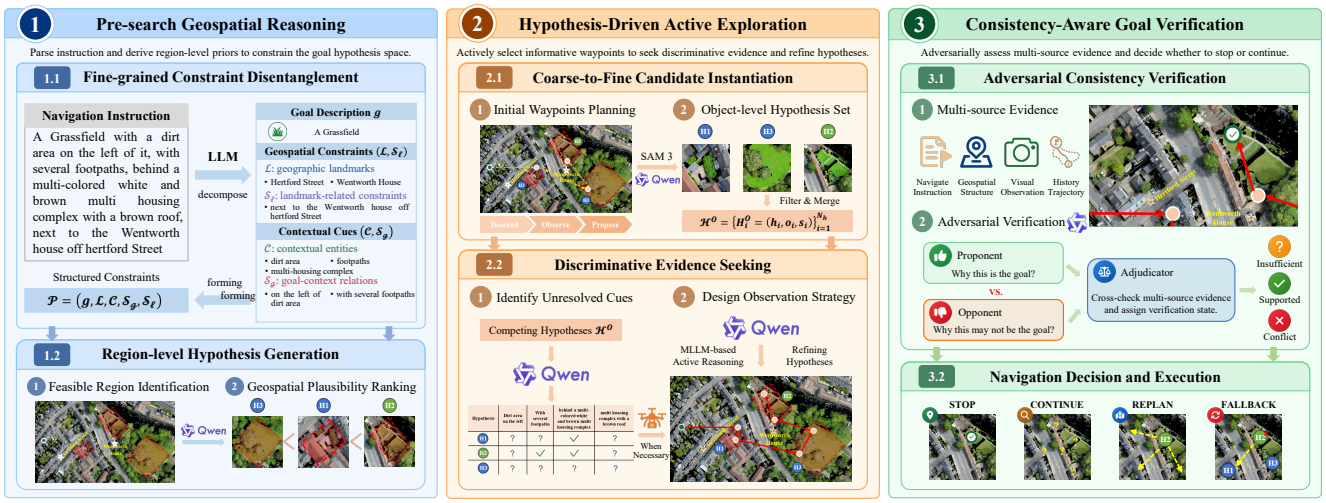


Figure 2. Detailed workflow of our framework. GeoReasoner progressively constrains, refines, and verifies goal hypotheses through three stages: (1) **Pre-search Geospatial Reasoning** disentangles geographic constraints and contextual cues to generate region-level goal hypotheses. (2) **Hypothesis-Driven Active Exploration** instantiates object-level hypotheses through initial waypoint exploration and actively seeks discriminative evidence to refine competing hypotheses when necessary. (3) **Consistency-Aware Goal Verification** employs adversarial reasoning to assess multi-source evidence and guide navigation decisions.

where g denotes the goal description, $\mathcal{L}$ represents geographic landmarks, and $\mathcal{C} = \{c_j\}_{j=1}^{N_c}$ contains individual contextual entities surrounding the goal. We explicitly preserve the goal–context relations as

$$\mathcal{S}_g = \{(g, r_j, c_j)\}_{j=1}^{N_c}, \tag{2}$$

where r_j denotes the spatial relation between the goal and contextual entity c_j. The remaining landmark-related spatial constraints are represented by

$$\mathcal{S}_l = \{(u_k, r_k, v_k)\}_{k=1}^{K_l}. \tag{3}$$

Here, u_k and v_k denote the entities involved in the k-th landmark-related constraint, r_k specifies their spatial relation, and K_l is the number of such constraints.

Region-level Hypothesis Generation. We then ascend directly from the starting point to altitude h_1 to obtain a high-altitude overview of the surrounding layout, grounding the extracted landmarks in the available geospatial information and translating their spatial relationships into geometric constraints. For instance, *at-corner-of* restricts feasible locations to the vicinity of a specified street intersection, while relations such as *near*, *next-to*, *in-front-of*, and *opposite* further constrain proximity or relative position when their reference entities are geographically grounded. To reduce incorrect pruning caused by ambiguous instructions or imprecise geospatial contours, the LLM further reviews and refines the resulting search regions.

Although geographic constraints narrow the search space, multiple regions may remain plausible due to incomplete or ambiguous spatial relations. Rather than selecting a single region, GeoReasoner maintains competing region-level hypotheses. Specifically, the LLM assesses each retained region against the grounded landmarks and spatial constraints, estimating its geospatial plausibility while identifying supporting evidence and unresolved relations. This yields the region-level hypothesis set

$$\mathcal{H}^R = \left\{H_j^R = \left(\Omega_j, s_j^R\right)\right\}_{j=1}^{N_R}, \tag{4}$$

where Ω_j denotes the j-th feasible region, s_j^R represents its geospatial plausibility score, and N_R is the number of retained regions, bounded by K_g. The exploration order jointly considers geospatial plausibility and visiting efficiency. Since coarse geospatial priors may be misleading, GeoReasoner prioritizes less plausible regions to challenge competing hypotheses and avoid premature confirmation.

At the same time, the UAV agent initializes a global memory to retain the supporting evidence and unresolved constraints from $\mathcal{S}_l$, together with goal–context relations from $\mathcal{S}_g$ requiring visual verification. The initial memory is defined as

$$\mathcal{M}_{init} = \left\{(j, \mathcal{E}_j^R, \mathcal{U}_j^R)\right\}_{j=1}^{N_R}, \tag{5}$$

where $\mathcal{E}_j^R$ contains the geographic evidence supporting Ω_j, and $\mathcal{U}_j^R$ records its unresolved spatial constraints and goal–context relations requiring further verification.

3.3. Hypothesis-Driven Active Exploration

While region-level hypotheses identify plausible goal areas, precise localization requires visual grounding and further spatial reasoning. Therefore, within each hypothesis

region, GeoReasoner first visits planned waypoints to instantiate object-level goal hypotheses, then employs MLLM reasoning to assess the acquired evidence, identify unresolved cues, and actively guide targeted observations when necessary for hypothesis disambiguation and refinement.

Coarse-to-Fine Candidate Instantiation. Within each feasible hypothesis region Ω_j, we descend to altitude h_2 for detailed observation and employ the MLLM (Qwen3.8-Flash [24]) to reason over the geographic configuration, instruction-derived spatial constraints, and UAV camera coverage to select a small number of waypoints for initial observation. The UAV then visits these waypoints to acquire visual observations and corresponding poses for subsequent candidate localization.

Afterwards, SAM 3 [4] is utilized to extract object masks as proposals from the acquired visual observations, which are further filtered by the MLLM based on the goal appearance described in g. The retained proposals are localized using the corresponding observations and UAV poses, and spatially redundant detections are merged into representative candidates using a radius r_{merge}. Each candidate is then associated with its parent region and relevant instruction-derived spatial constraints, preserving unresolved relations for subsequent active disambiguation. The resulting object-level hypothesis set is defined as

$$\mathcal{H}^o = \{H_i^o = (h_i, \mathbf{o}_i, s_i)\}_{i=1}^{N_h}, \tag{6}$$

where $h_i = (g, \mathcal{S}_i)$ denotes the semantic-spatial hypothesis, with $\mathcal{S}_i$ representing the goal-context relations requiring verification for candidate i. $\mathbf{o}_i$ and s_i denote its estimated geographic location and initial plausibility, respectively. Each hypothesis inherits the geographic constraints of its parent region. Up to K_h representative hypotheses are prioritized for initial detailed inspection, while the remaining candidates are retained.

Discriminative Evidence Seeking. Initial observations from a limited number of waypoints may lack sufficient visual and contextual details to distinguish competing goal hypotheses, leaving multiple plausible object-level hypotheses $\mathcal{H}^o$ after initial filtering. To address this ambiguity, we apply the MLLM to evaluate these hypotheses against the full instruction, the initial observations, and evidence accumulated in the global agent memory $\mathcal{M}_t$. Candidates with explicit contradictions are eliminated, while the remaining hypotheses are compared to identify unresolved visual and contextual cues that can distinguish them.

When the available evidence is insufficient to resolve these cues, the MLLM determines appropriate observation strategies according to the nature of each cue and its spatial context. Fine-grained target details may require closer or lower-altitude inspection, whereas contextual objects are sought in areas inferred from their spatial relations to the candidate, such as the opposite roadside for a car described as across from the goal. When additional observations are needed, the MLLM combines these observation requirements with the current UAV pose and exploration history to actively select targeted observation waypoints that can help distinguish competing hypotheses. When multiple supplementary waypoints are identified, they are organized into an efficient exploration route by solving a Traveling Salesman Problem (TSP) [2]. The UAV then executes navigation actions $a_t \in \mathcal{A}$ toward the selected waypoints, acquiring additional observations O_{t+1} and corresponding poses p_{t+1} of the candidate objects or their surrounding areas. Supplementary observation attempts are limited to a total budget B_v per episode.

The MLLM then organizes evidence from the initial and, when available, supplementary observations into a structured record for each relevant hypothesis:

$$\mathcal{B}_i^t = (n_i^t, u_i^t, c_i^t), \tag{7}$$

where n_i^t summarizes the evidence assessed at step t, u_i^t records the cues that remain unresolved, and c_i^t indicates whether the evidence supports, contradicts, or suggests revising the hypothesis. The resulting record is incorporated into the global agent memory:

$$\mathcal{M}_{t+1} = \text{Update}\left(\mathcal{M}_t, \mathcal{B}_i^t\right). \tag{8}$$

Based on the updated memory, GeoReasoner refines candidate locations using available observations and their corresponding UAV poses when more accurate estimates become available. The accumulated visual and contextual evidence further guides hypothesis re-ranking and the elimination of inconsistent candidates. The refined hypothesis set is represented as

$$\hat{\mathcal{H}}^o = \left\{\hat{H}_i^o = (h_i, \hat{\mathbf{o}}_i, \hat{s}_i)\right\}_{i=1}^{N_h'}, \tag{9}$$

where $\hat{\mathbf{o}}_i$ and $\hat{s}_i$ denote the refined location and updated plausibility, respectively, and N_h' is the number of surviving hypotheses. The refined hypotheses $\hat{\mathcal{H}}^o$ and accumulated agent memory $\mathcal{M}_T$ are subsequently passed to consistency-aware goal verification for the final navigation decision.

3.4. Consistency-Aware Goal Verification

Even after active evidence seeking and hypothesis refinement, the highest-ranked candidate may lack decisive evidence distinguishing it from competing alternatives. Hence, we perform a final check to assess its consistency with the instruction-specified visual attributes and spatial relations, and determine whether the accumulated evidence is sufficient for goal confirmation.

Adversarial Consistency Verification. Given the refined hypotheses $\hat{\mathcal{H}}^o$ and global agent memory $\mathcal{M}_T$, we introduce a single-round debate among three role-specific

MLLM agents to verify the leading hypothesis against competing alternatives. Specifically, the *proponent* constructs a supporting argument by identifying visual, contextual, and geospatial evidence consistent with the navigation instruction. Conversely, the *opponent* challenges this argument by identifying potential contradictions and examining alternative hypotheses that could also explain the observed evidence.

The *adjudicator* then evaluates both arguments by examining whether the visual attributes and contextual relations of each hypothesis are jointly consistent with the instruction, geographic constraints, and observations accumulated throughout exploration, while leveraging the navigation history to identify supporting or contradictory evidence from previously visited locations. Based on this consistency assessment, it rules out competing hypotheses that contradict verified evidence while retaining those whose distinguishing cues remain unresolved. This process prevents the leading hypothesis from being accepted solely on the basis of its relative plausibility.

For each hypothesis $\hat{H}_i^o \in \hat{\mathcal{H}}^o$, the adjudicator summarizes supporting, contradictory, and unresolved evidence in a structured assessment $\mathcal{E}_i$, and assigns a verification state:

$$y_i \in \{\texttt{supported}, \texttt{insufficient}, \texttt{conflict}\}, \tag{10}$$

where `supported` indicates sufficient evidence consistency with no substantial competing ambiguity, `insufficient` denotes unresolved evidence or viable alternatives, and `conflict` indicates an explicit violation of essential constraints.

Navigation Decision and Execution. The adjudicator's verdict directly guides the final navigation decision. A supported hypothesis is accepted as the goal, while targeted evidence seeking is resumed only when unresolved cues could alter the decision. If further informative observations are infeasible, GeoReasoner performs a budget-constrained fallback selection among the remaining non-conflicting hypotheses, rather than treating insufficient evidence as verified goal confirmation. When all candidates are rejected, re-planning is triggered to revisit region-level hypotheses using accumulated evidence, allowing previously excluded regions to be reconsidered.

Let $\mathcal{H}_{\text{surv}} = \{\hat{H}_i^o \mid y_i \neq \texttt{conflict}\}$ denote the surviving hypotheses. When fallback selection is required, the agent performs evidence voting over the adjudicator's existing assessments:

$$H^{\text{vote}} = \text{Vote}\left(\left\{(\hat{H}_i^o, \mathcal{E}_i)\right\}_{\hat{H}_i^o \in \mathcal{H}_{\text{surv}}}\right). \tag{11}$$

Let $\eta_t \in \{0, 1\}$ indicate whether further informative exploration has been exhausted. The navigation decision follows the ordered rules:

$$d_t = \begin{cases} \texttt{STOP}(H^\star), & \exists i:\ y_i = \texttt{supported}, \\ \texttt{REPLAN}, & \mathcal{H}_{\text{surv}} = \emptyset, \\ \texttt{EXPLORE}, & \eta_t = 0, \\ \texttt{STOP}(H^{\text{vote}}), & \text{otherwise}, \end{cases} \tag{12}$$

where $H^\star$ denotes the hypothesis confirmed by the adjudicator, and $\text{STOP}(H)$ denotes navigating to the estimated location of H before issuing the stop action. The number of additional re-planning attempts is limited to $R_{\max}$. If all hypotheses remain conflicting after exhausting the re-planning budget, the episode terminates without a verified goal. By evaluating competing hypotheses through adversarial reasoning and consistency checks, GeoReasoner mitigates premature goal commitment while preserving the efficiency of targeted exploration.

4. Experiments

4.1. Experimental Setup

Dataset. We conduct experiments on CityNav [16], a real-world aerial VLN benchmark containing 32,637 instruction-trajectory pairs for 5,850 target objects across Cambridge and Birmingham, covering 4.65 km^2. The dataset comprises 22,002 training, 2,498 validation-seen, 2,826 validation-unseen, and 5,311 test-unseen episodes. Following [10], we further evaluate navigation performance across three difficulty levels (Easy, Medium, and Hard) based on the start-to-goal distance.

Evaluation Metrics. Following CityNav [16], we adopt four metrics: Navigation Error (NE), Success Rate (SR), Oracle Success Rate (OSR), and Success weighted by Path Length (SPL). NE measures the final distance to the goal, while SR and OSR evaluate whether the UAV reaches within 20 m of the goal at termination or during navigation, respectively. SPL further accounts for navigation efficiency by incorporating path length into task success.

Comparison Methods. We compare GeoReasoner with 11 aerial VLN baselines, including conventional observation-driven methods (Seq2Seq+GSM [14, 16], CMA+GSM [14, 16], AerialVLN+GSM [16, 19], and MGP [16]), recent learning-based approaches (FlightGPT [3], RingMo-Agent [12], HETT [8], HTNav [10], and DroneNav [18]), and geographically informed methods (FlightGPT+GPOD [3, 20] and GeoNav [27]).

Implementation Details. GeoReasoner operates in a zero-shot setting, employing Qwen3.8-Flash [24] for multimodal reasoning and SAM 3 [4] for object proposal generation. For pre-search stage, we retain at most $K_g = 6$ feasible regions and perform an initial high-altitude observation at

Table 1. Comparison with state-of-the-art methods on the original CityNav benchmark. NE is measured in meters, while SR, OSR, and SPL are reported as percentages. Best and second-best results are highlighted in **bold** and underline, respectively.

Method	Learning Setting	Test-Unseen				Val-Unseen				Val-Seen			
		NE↓	SR↑	OSR↑	SPL↑	NE↓	SR↑	OSR↑	SPL↑	NE↓	SR↑	OSR↑	SPL↑
Seq2Seq+GSM	Supervised	98.10	3.81	13.92	2.79	78.60	5.13	10.90	4.65	58.50	8.43	17.31	7.28
CMA+GSM		94.6	4.68	12.01	4.05	75.9	4.38	9.29	3.90	68.0	6.25	13.28	5.40
AerialVLN+GSM		85.1	6.72	18.21	5.16	72.7	6.35	15.24	5.06	56.6	10.16	22.20	7.89
MGP		93.80	6.38	26.04	6.08	75.10	5.84	22.19	5.56	59.70	8.69	35.51	8.28
FlightGPT		76.20	21.20	35.38	19.24	68.10	14.69	29.33	13.24	66.10	17.57	30.26	15.78
FlightGPT+GPOD		54.85	29.15	44.94	26.85	61.88	24.77	39.14	22.75	61.56	23.86	36.71	21.83
RingMo-Agent		64.2	21.15	33.79	20.92	61.6	16.99	30.11	16.66	**35.7**	**45.81**	57.07	**45.59**
HETT		72.90	22.97	39.30	17.01	62.10	17.48	25.09	14.46	45.20	25.16	48.40	23.01
HTNav		68.50	22.23	41.02	19.63	59.80	15.85	34.43	13.55	47.20	28.30	47.92	25.07
DroneNav		74.9	22.47	30.22	20.38	63.5	15.06	25.81	13.27	54.3	18.94	36.65	17.37
GeoNav	Zero-shot	73.5	25.9	41.6	16.0	64.1	16.9	35.5	9.9	58.6	21.8	39.9	12.3
GeoReasoner		**38.97**	**37.77**	**59.63**	**27.85**	**49.60**	**35.14**	**61.36**	**26.57**	43.07	35.27	**62.77**	26.93

Table 2. Navigation performance across different difficulty levels on the original CityNav Test-Unseen split. Best and second-best results are highlighted in **bold** and underline, respectively.

Method	Learning Setting	Easy				Medium				Hard			
		NE↓	SR↑	OSR↑	SPL↑	NE↓	SR↑	OSR↑	SPL↑	NE↓	SR↑	OSR↑	SPL↑
Seq2Seq	Supervised	238.80	3.07	14.70	2.64	246.50	0.43	3.87	0.38	253.10	0.48	4.38	0.44
CMA		260.70	0.49	16.69	0.44	241.20	1.10	7.67	1.09	253.80	0.96	1.64	0.95
MGP		98.90	6.15	39.89	5.48	90.90	6.29	21.47	6.21	90.00	6.80	12.10	6.78
HTNav		68.90	19.99	45.18	14.94	68.20	20.40	39.44	18.69	68.10	21.91	34.14	21.11
GeoNav	Zero-shot	66.1	26.1	40.1	15.2	53.8	22.9	39.6	17.1	68.9	16.7	22.9	12.5
GeoReasoner		**36.63**	**39.40**	**64.27**	**24.75**	**39.65**	**37.84**	**60.68**	**29.21**	**41.43**	**35.45**	**52.13**	**30.59**

$h_1 = 100\,\text{m}$. For active exploration stage, detailed observations are conducted at $h_2 = 50\,\text{m}$, with spatially redundant SAM 3 proposals merged within $r_{\text{merge}} = 9\,\text{m}$. Up to $K_h = 9$ representative candidates are selected for initial inspection, and supplementary observations are limited to $B_v = 4$ attempts per episode. For goal verification stage, we employ a single-round adversarial debate, allowing at most $R_{\max} = 2$ additional re-planning attempts.

4.2. Experimental Results

Quantitative Comparisons. Table 1 compares GeoReasoner with supervised and zero-shot methods on CityNav. **Despite requiring no task-specific training, GeoReasoner outperforms all methods across all metrics on unseen splits**, including supervised ones. On Val-Seen, GeoReasoner achieves the highest OSR of 62.77%, while RingMo-Agent leads in the remaining metrics, potentially benefiting from task-specific training on seen environments.

Compared with reactive navigation methods, ranging from conventional ones (Seq2Seq+GSM, CMA+GSM, and MGP) to recent reasoning-based methods (FlightGPT, FlightGPT+GPOD, HETT, and HTNav), GeoReasoner achieves substantial improvements. Across both unseen splits, GeoReasoner surpasses the best supervised methods, reducing NE by up to 15.88 m while improving SR, OSR, and SPL by up to 10.37, 22.22, and 3.82 points, respectively. These consistent gains, achieved without task-specific training, highlight the superiority of geospatial-guided active reasoning over reactive navigation paradigms.

Compared with the zero-shot GeoNav, which leverages geospatial priors to guide navigation but still acquires evidence as a consequence of exploration, GeoReasoner reduces Test-Unseen NE by 34.53 m while improving SR, OSR, and SPL by 11.87, 18.03, and 11.85 points, respectively. This substantial improvement highlights the advantage of shifting from passive exploration to hypothesis-driven active reasoning, where discriminative observations are deliberately acquired to resolve competing hypotheses.

Table 3. Navigation results on the Hard subsets of CityNav. Best and second-best results are highlighted in **bold** and underline, respectively.

Method	Learning Setting	Test-Unseen-Hard				Val-Unseen-Hard				Val-Seen-Hard			
		NE↓	SR↑	OSR↑	SPL↑	NE↓	SR↑	OSR↑	SPL↑	NE↓	SR↑	OSR↑	SPL↑
HETT	Supervised	45.14	24.00	42.51	21.97	58.84	18.87	32.01	17.04	**41.40**	27.95	43.74	25.89
FlightGPT	Supervised	73.01	23.09	35.51	22.05	69.92	14.99	23.22	14.20	71.97	14.62	25.50	13.90
GeoNav (Qwen3.8-Flash)	Zero-shot	59.90	12.16	25.38	10.59	77.87	9.71	21.09	8.54	71.87	8.54	19.77	7.00
Ours (Qwen3.8-Flash)	Zero-shot	**41.43**	**35.45**	**52.13**	**30.59**	**54.32**	**34.51**	**55.78**	**29.79**	47.79	**33.10**	**54.74**	**28.21**

Robustness across Navigation Difficulties. We also evaluate GeoReasoner's robustness under increasing navigation difficulty. As shown in Table 2, GeoReasoner achieves the best results across all metrics on Easy, Medium, and Hard episodes. Notably, its SR decreases by only 3.95 points from Easy to Hard, compared with a 9.40-point drop for GeoNav. On Hard episodes, GeoReasoner achieves 35.45% SR and 52.13% OSR, outperforming HTNav by 13.54 and 17.99 points, respectively.

We further examine robustness across the three Hard evaluation splits in Table 3. Our zero-shot method, GeoReasoner, consistently surpasses the supervised HETT and FlightGPT in SR, OSR, and SPL. Moreover, using the same MLLM (Qwen3.8-Flash) as the reproduced GeoNav, GeoReasoner still achieves SR improvements of 23.29, 24.80, and 24.56 points across the three splits, respectively. These results confirm the robustness of our geospatial-guided active reasoning framework under challenging conditions.

Table 4. Ablation results of GeoReasoner on Test-Unseen.

Variant	NE↓	SR↑	OSR↑	SPL↑
Full	**38.97**	**37.77**	**59.63**	**27.85**
w/o Hypothesis	54.21	29.49	48.31	19.51
w/o Exploration Memory	47.35	33.65	53.49	22.57
w/o Consistency Verification	46.13	32.91	55.70	24.36
w/o Active Evidence Seeking	52.68	30.62	49.33	20.37
w/o Adversarial Debate	42.18	35.79	59.33	26.52

Ablation Study. Table 4 evaluates the contributions of key components in GeoReasoner. Removing geospatial hypotheses (w/o Hypothesis) causes the largest degradation, increasing NE by 15.24 m and reducing SR and SPL by 8.28 and 8.34 points, respectively. This demonstrates the importance of constraining the search space before exploration.

Without exploration memory (w/o Exploration Memory), SR and SPL decrease by 4.12 and 5.28 points, highlighting the importance of accumulating evidence for informed exploration. Disabling active evidence seeking (w/o Active Evidence Seeking) leads to substantial drops of 7.15 and 10.30 points in SR and OSR, respectively, demonstrating the effectiveness of targeted observations in resolving spatial ambiguity and locating the goal.

For goal verification, replacing adversarial debate with direct MLLM assessment (w/o Adversarial Debate) reduces SR by 1.98 points, while removing consistency verification (w/o Consistency Verification) results in a larger decline of 4.86 points. The latter also decreases OSR by 3.93 points, highlighting the role of consistency verification in resolving ambiguous hypotheses and improving goal localization.

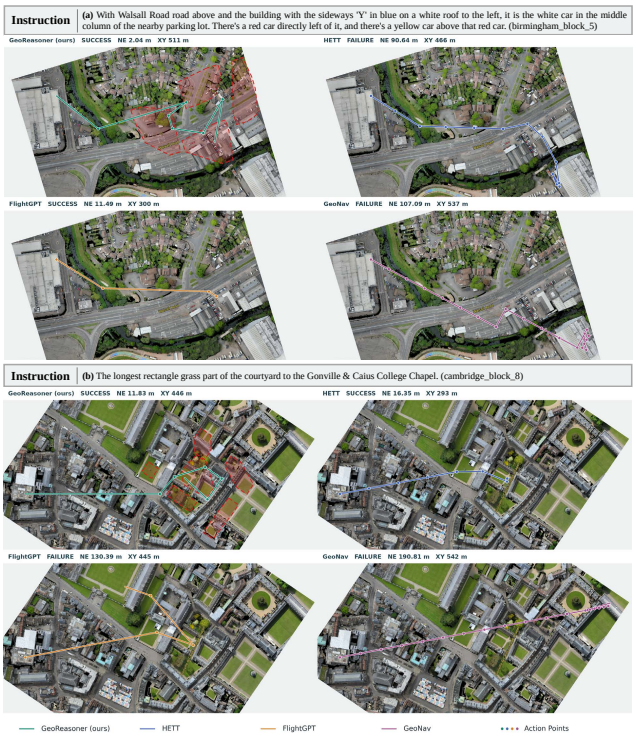


Figure 3. Qualitative comparison on CityNav. Recorded trajectories of GeoReasoner and three other methods in (a) fine-grained vehicle localization and (b) landmark-guided goal localization.

Qualitative Results. Fig. 3 compares navigation trajectories in two challenging scenarios involving ambiguous visual targets and complex spatial relationships. In (a), GeoReasoner leverages road and building layouts to narrow the search region, then actively distinguishes the target white car from similar vehicles using spatial relations to nearby

cars (NE: 2.04 m). In (b), identifying the correct rectangular grass region among similar courtyards is challenging, while GeoReasoner combines geospatial priors with targeted exploration and consistency verification to achieve an NE of 11.83 m. Unlike baselines that may follow misleading spatial cues, GeoReasoner effectively resolves spatial ambiguity through active evidence acquisition, enabling more accurate and reliable localization.

5. Conclusion

We present GeoReasoner, which advances aerial VLN beyond reactive search through geospatial-guided hypothesis generation, active evidence seeking, and consistency-aware verification. Experiments demonstrate superior zero-shot performance and robustness across navigation difficulties.

LLM Usage Statement

We used Qwen3.8-Flash for multimodal reasoning in our proposed framework and ChatGPT (OpenAI) for language polishing and manuscript refinement. All research design, experimental results, and scientific claims were verified by the authors.

References

[1] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3674–3683, 2018. 1, 2

[2] David L Applegate, Robert E Bixby, Vašek Chvátal, and William J Cook. The traveling salesman problem: a computational study. In *The traveling salesman problem*. Princeton university press, 2011. 5

[3] 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 *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 6670–6687, 2025. 1, 3, 6

[4] Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. In *International conference on learning representations*, pages 138846–138923, 2026. 5, 6

[5] Shizhe Chen, Pierre-Louis Guhur, Cordelia Schmid, and Ivan Laptev. History aware multimodal transformer for vision-and-language navigation. *Advances in neural information processing systems*, 34:5834–5847, 2021. 2

[6] 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 *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16516–16526. IEEE, 2022. 2

[7] Erik Daxberger, Nina Wenzel, David Griffiths, Haiming Gang, Justin Lazarow, Gefen Kohavi, Kai Kang, Marcin Eichner, Yinfei Yang, Afshin Dehghan, et al. Mm-spatial: Exploring 3d spatial understanding in multimodal llms. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7395–7408. IEEE, 2025. 3

[8] Xichen Ding, Jianzhe Gao, Cong Pan, Wenguan Wang, and Jie Qin. History-enhanced two-stage transformer for aerial vision-and-language navigation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 18225–18233, 2026. 3, 6

[9] Björn Döschl and Jane Jean Kiam. Say-reapex: An llm-modulo uav online planning framework for search and rescue. In *2nd CoRL Workshop on Learning Effective Abstractions for Planning*, 2024. 1

[10] 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. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10976–10985, 2026. 1, 3, 6

[11] Wenxuan Guo, Xiuwei Xu, Yichen Liu, Xiangyu Li, Hang Yin, Huangxing Chen, Wenzhao Zheng, Jianjiang Feng, Jie Zhou, and Jiwen Lu. Awarevln: Reasoning with self-awareness for vision-language navigation. *arXiv preprint arXiv:2605.22816*, 2026. 2

[12] Huiyang Hu, Peijin Wang, Yingchao Feng, Kaiwen Wei, Wenxin Yin, Wenhui Diao, Mengyu Wang, Hanbo Bi, Kaiyue Kang, Tong Ling, et al. Ringmo-agent: A unified remote sensing foundation model for multi-platform and multi-modal reasoning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026. 3, 6

[13] Hailong Huang, Andrey V Savkin, and Chao Huang. Decentralized autonomous navigation of a uav network for road traffic monitoring. *IEEE Transactions on Aerospace and Electronic Systems*, 57(4):2558–2564, 2021. 1

[14] Jacob Krantz, Erik Wijmans, Arjun Majumdar, Dhruv Batra, and Stefan Lee. Beyond the nav-graph: Vision-and-language navigation in continuous environments. In *European Conference on Computer Vision*, pages 104–120. Springer, 2020. 6

[15] Alexander Ku, Peter Anderson, Roma Patel, Eugene Ie, and Jason Baldridge. Room-across-room: Multilingual vision-and-language navigation with dense spatiotemporal grounding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4392–4412, 2020. 1

[16] Jungdae Lee, Taiki Miyanishi, Shuhei Kurita, Koya Sakamoto, Daichi Azuma, Yutaka Matsuo, and Nakamasa Inoue. Citynav: A large-scale dataset for real-world aerial navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5912–5922, 2025. 1, 3, 6

[17] Benlin Liu, Yuhao Dong, Yiqin Wang, Zixian Ma, Yansong Tang, Luming Tang, Yongming Rao, Wei-Chiu Ma,

and Ranjay Krishna. Coarse correspondences boost spatial-temporal reasoning in multimodal language model. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3783–3792. IEEE, 2025. 3

[18] Fangming Liu, Guohua Li, Linfeng Zou, Yadong Chen, and Peng Cheng. Dronenav: Unified text-visual representation and structured spatial reasoning for robust uav vision-and-language navigation. *Neurocomputing*, page 133492, 2026. 6

[19] Shubo Liu, Hongsheng Zhang, Yuankai Qi, Peng Wang, Yanning Zhang, and Qi Wu. Aerialvln: Vision-and-language navigation for uavs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15384–15394, 2023. 1, 3, 6

[20] Yuze Liu, Changming Xu, Kewen Xiao, Yuhua Wu, and Ziyu Li. Gpod: Geographic priors and object detection for candidate-guided target localization in city-scale uav vision-and-language navigation. *Drones*, 10(6):458, 2026. 6

[21] Wufei Ma, Luoxin Ye, Celso M de Melo, Alan Yuille, and Jieneng Chen. Spatialllm: A compound 3d-informed design towards spatially-intelligent large multimodal models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17249–17260. IEEE, 2025. 3

[22] Yuwei Ning, Ganlong Zhao, Yipeng Qin, Si Liu, Yang Liu, Liang Lin, and Guanbin Li. Lookasidevln: direction-aware aerial vision-and-language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 32441–32450, 2026. 3

[23] Zekun Qi, Runpei Dong, Shaochen Zhang, Haoran Geng, Chunrui Han, Zheng Ge, Li Yi, and Kaisheng Ma. Shapellm: Universal 3d object understanding for embodied interaction. In *European Conference on Computer Vision*, pages 214–238. Springer, 2024. 3

[24] Zihan Qiu, Zekun Wang, Xiao Li, Yanpeng Li, Yang Xu, Yixuan Wang, Huaqing Zhang, Rui Men, Bochao Mao, Chengruidong Zhang, et al. On the design of qwen3. 8-next architecture: Evaluation, efficiency, and training stability. *arXiv preprint arXiv:2608.30320*, 2026. 5, 6

[25] Junhui Wang, Dongjie Huo, Zehui Xu, Yongliang Shi, Yimin Yan, Yuanxin Wang, Chao Gao, Yan Qiao, and Guyue Zhou. Openbench: A new benchmark and baseline for semantic navigation in smart logistics. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 16202–16208. IEEE, 2025. 1

[26] Xingrui Wang, Wufei Ma, Tiezheng Zhang, Celso M de Melo, Jieneng Chen, and Alan Yuille. Spatial457: A diagnostic benchmark for 6d spatial reasoning of large multimodal models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24669–24679. IEEE, 2025. 3

[27] Haotian Xu, Yue Hu, Chen Gao, Zhengqiu Zhu, Yong Zhao, and Quanjun Yin. Geonav: Empowering mllms with dual-scale geospatial reasoning for language-goal aerial navigation. *Pattern Recognition*, page 113365, 2026. 3, 6

[28] Wei Xue, Mingcheng Li, Xuecheng Wu, Jingqun Tang, Dingkang Yang, and Lihua Zhang. Profocus: Proactive perception and focused reasoning in vision-and-language navigation. *arXiv preprint arXiv:2603.05530*, 2026. 2

[29] Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10632–10643. IEEE, 2025. 3

[30] 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 *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31292–31309, 2025. 1, 3

[31] Gengze Zhou, Yicong Hong, Zun Wang, Xin Eric Wang, and Qi Wu. Navgpt-2: Unleashing navigational reasoning capability for large vision-language models. In *European Conference on Computer Vision*, pages 260–278. Springer, 2024. 2