

From Semantic Grounding to Decision Optimization: A Unified Framework for Long-Horizon UAV Vision-Language Navigation

Zeyuan Ma

State Key Laboratory of Virtual Reality Technology and Systems, Beihang University
Beijing, China

School of Computer Science and Engineering, Beihang University
Beijing, China

zeyuanma@buaa.edu.cn

Jiaxin Chen*

State Key Laboratory of Virtual Reality Technology and Systems, Beihang University
Beijing, China

School of Computer Science and Engineering, Beihang University
Beijing, China

jiaxinchen@buaa.edu.cn

Di Huang

School of Computer Science and Engineering, Beihang University
Beijing, China

dhuang@buaa.edu.cn

Abstract

UAV vision-language navigation (UAV-VLN) focuses on enabling an aerial agent to follow natural-language instructions in open 3D environments from egocentric visual observations. Current approaches suffer from three coupled issues: weak grounding of instruction-relevant landmarks in visual observations, insufficient exploration of long-horizon history, and unstable decisions under local traps or repeated exploration. To address these issues, we propose a unified semantic-to-decision framework. First, we present an instruction-grounded semantic enhancement module that injects object-level semantics and relative spatial cues into the current observation state. Subsequently, we develop a relevance-aware dynamic temporal aggregation strategy that reweights the full history buffer while converting a few high-relevance frames into structured landmark prompts for the decoder. Finally, we devise a topology-aware decision method that combines local-optimum cognition with group-relative policy optimization under progress, goal, semantic, and path-compliance rewards. Experiments on the widely used AerialVLN and OpenFly benchmarks clearly demonstrate that our method achieves state-of-the-art performance. Our code is available at <https://github.com/mksasx/S2D-UAV-VLN>.

CCS Concepts

• **Computing methodologies** → **Computer vision**; *Artificial intelligence*; Natural language processing.

Keywords

Vision-Language Navigation, Unmanned Aerial Vehicles, Reinforcement Learning

*Corresponding author



This work is licensed under a Creative Commons Attribution 4.0 International License. *MM '26, Rio de Janeiro, Brazil*
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2213-4/2026/11
<https://doi.org/10.1145/3767308.3836562>

ACM Reference Format:

Zeyuan Ma, Jiaxin Chen, and Di Huang. 2026. From Semantic Grounding to Decision Optimization: A Unified Framework for Long-Horizon UAV Vision-Language Navigation. In *Proceedings of the 34th ACM International Conference on Multimedia (MM '26)*, November 10–14, 2026, Rio de Janeiro, Brazil. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3767308.3836562>

1 Introduction

Vision-language navigation (VLN) studies how an embodied agent follows free-form language under partial visual observations [1]. For unmanned aerial vehicles (UAVs), this task is far more challenging, since the agent must reason across open 3D environments with drastic scale variations, sparse yet decisive landmarks, and trajectories much longer than in indoor or ground settings. This capability is important for inspection, search and rescue, logistics, and urban surveillance, where natural language provides a flexible interface for non-expert operators, although the present work evaluates simulation-based high-level navigation rather than real-world UAV deployment. Recent aerial research spans benchmark-style UAV-VLN datasets, high-fidelity simulation platforms, large-scale synthetic toolchains, city-scale real-world aerial navigation corpora, low-altitude agentic application systems, and broader studies on aerial embodied-agents [10, 18, 21, 24, 28, 34, 37, 44, 51].

These resources have enabled measurable progress, with recent VLM/MLLM-based systems (e.g., CityNavAgent, FlightGPT, GeoNav) and contemporary open-world, end-to-end, or structured-view aerial variants (e.g., OpenVLN, UAV-VLN, grid-based aerial VLN) all demonstrating the potential of stronger multimodal reasoning for long-horizon aerial navigation [3, 21, 29, 42, 48, 50]. Despite these advances, long-horizon UAV-VLN still suffers from a coupled failure chain. Current-view grounding is fragile because aerial landmarks are often small, repeated, and visually ambiguous, causing global image features to overlook the object phrases that directly determines the next action. Once grounding drifts, historical frame aggregation becomes noisy, as many stored frames cease to be relevant to the active sub-goal. The resulting temporal state destabilizes sequential decisions, incurring repeated exploration, semantically plausible yet incorrect stopping, or local looping in visually similar regions.

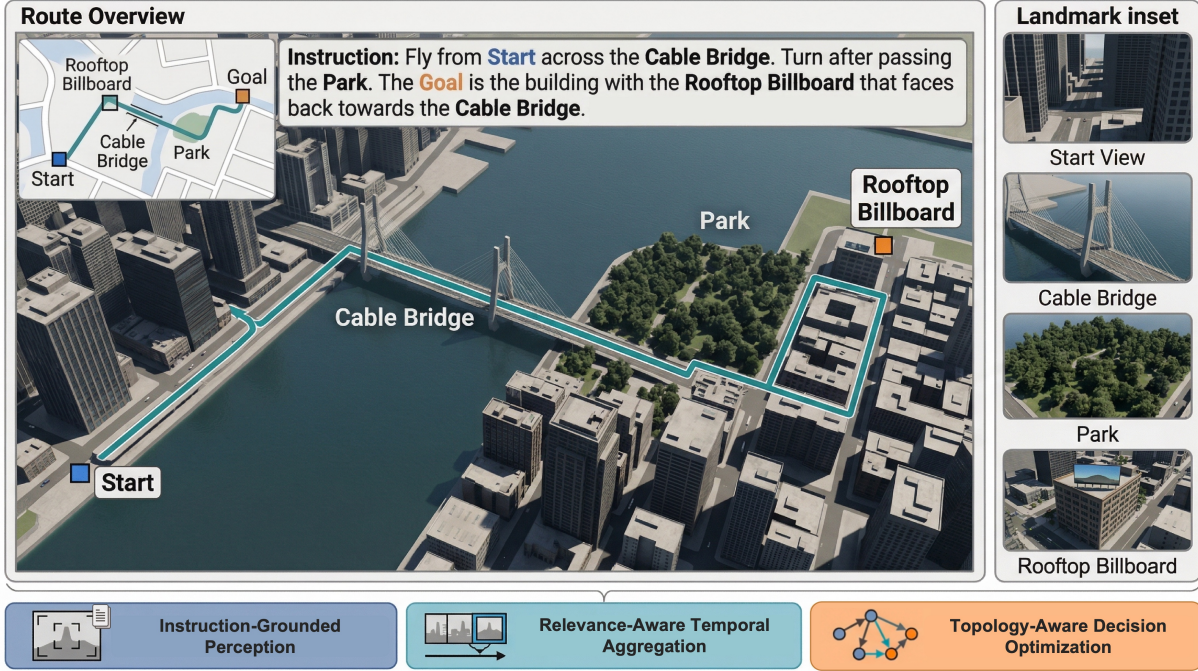


Figure 1: Illustration of the long-horizon UAV-VLN task. The example route shows that successful navigation depends on accurate landmark grounding, relevance-aware history selection, and stable topology-aware decisions.

These errors do not stem from three isolated modules, but reflect a single core problem of state construction. Unlike recent aerial agents that primarily strengthen planner-side reasoning or coarse-to-fine geospatial search [3, 42], we focus on the policy state itself: grounding quality determines which past observations remain useful, temporal relevance controls whether the policy receives a clean state, and state quality governs the stability of topology-aware decision-making. We accordingly formulate aerial navigation as a progressive semantic-to-decision pipeline. As displayed in Fig. 2, our framework first builds instruction-grounded observation tokens from landmark semantics, relative spatial cues, and local topology; then selectively reuses historical evidence through relevance-aware dynamic temporal aggregation; and finally refines action selection via topology-aware local-optimum recovery and GRPO-based policy optimization. Figure 1 illustrates the task abstraction that motivates this design.

The main contributions of this work are summarized as follows:

- We present a unified system-level semantic-to-decision framework for long-horizon UAV-VLN that jointly couples instruction-grounded current-view perception, relevance-aware temporal aggregation, and topology-aware action refinement.
- We develop two distinctive and complementary mechanisms, namely dynamic temporal aggregation (DTA) and local-optimum cognition (LOC) respectively, where DTA combines instruction-conditioned full-history weighting with sparse landmark prompts, and LOC detects topology-level stagnation and conditionally injects a semantically matched frontier cue. Together with composite trajectory rewards, they improve decision stability and robustness in long-horizon navigation.

- We validate the proposed framework on AerialVLN and OpenFly, where the results clearly indicate the effectiveness of the coupled design of semantic grounding, temporal aggregation, and topology-aware decision optimization.

2 Related Work

2.1 Vision-Language Navigation

VLN originated in indoor and ground settings, where agents navigate with denser local cues and shorter trajectories [1, 5, 9, 17]. UAV-VLN is harder because the agent must reason over open 3D environments, severe scale changes, sparse landmarks, and much longer instruction horizons. The aerial benchmark ecosystem has expanded in several directions: AerialVLN and OpenFly provide benchmark-style long-horizon evaluation, OpenUAV introduces a more realistic simulation platform together with the UAV-Need-Help benchmark, and CityNav targets real-world city-scale aerial navigation [10, 18, 24, 37]. Recent studies further broaden this space toward open-world aerial VLN, end-to-end UAV-VLN pipelines, delivery-oriented low-altitude navigation, dual-agent aerial VLN, and general aerial embodied-agent systems [21, 29, 40, 44, 49, 51]. Related aerial efforts, including aerial vision-dialog navigation and flying-on-a-word control, broaden the modality and deployment space under task assumptions that differ from sequential long-horizon instruction following [8, 36]. The common lesson is that methods inherited from ground VLN must be redesigned for aerial ambiguity, longer memory horizons, and more fragile sequential decisions.

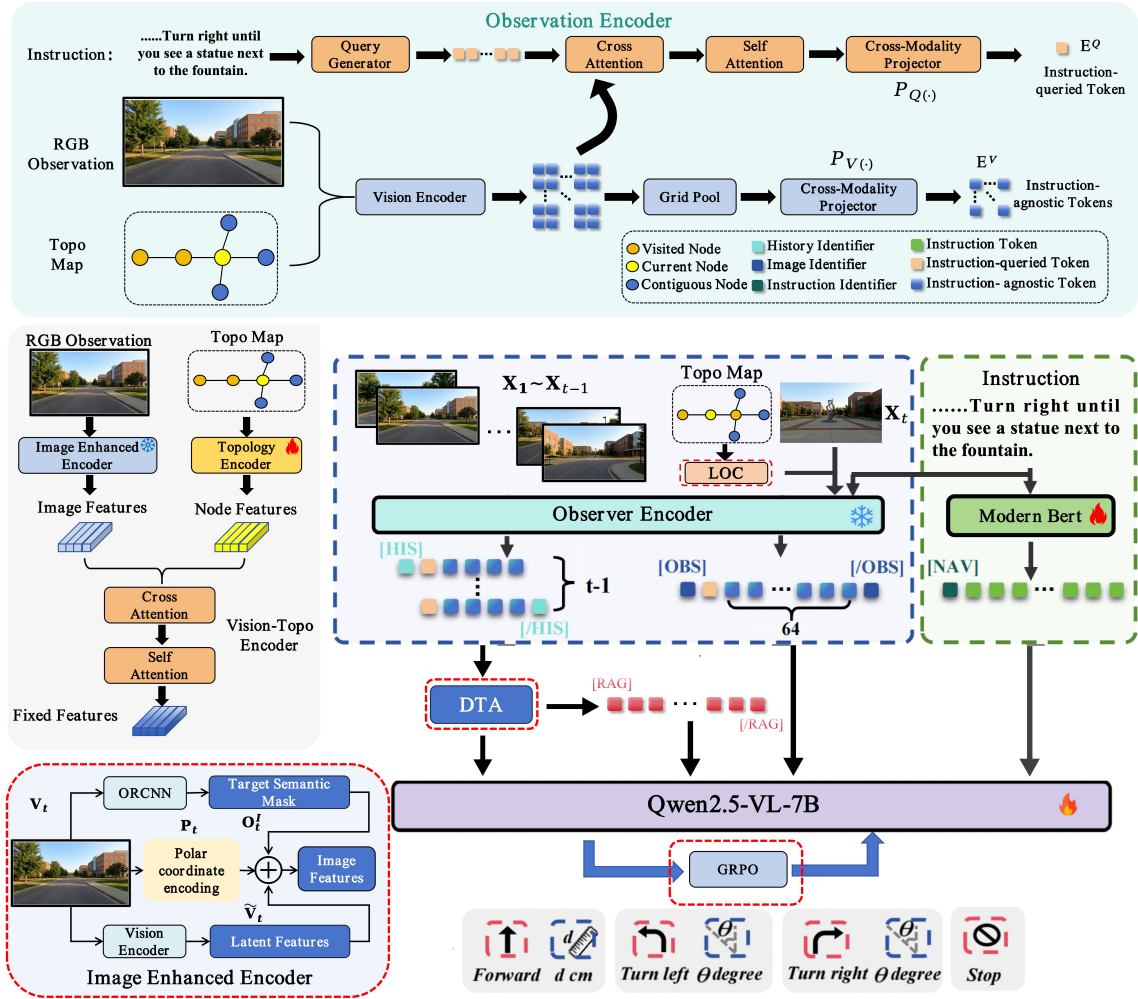


Figure 2: Overview of the proposed semantic-to-decision pipeline. The model encodes observation, instruction, and topology context; performs semantic-spatial enhancement and dynamic temporal aggregation (DTA) with a sparse grounding prompt branch, and conducts local optimum recovery (LOC) and GRPO-based decision optimization prior to next-action prediction.

2.2 Semantic Grounding and History Modeling in Aerial Navigation

The first two bottlenecks are fine-grained semantic grounding and selective history modeling. Small or similar entities, domain shifts, and limited labels challenge UAV detection [7, 15, 20, 38, 39], so coarse alignment can miss the landmark phrase that determines the next action. Ground VLN methods such as RCM, entity-graph reasoning, and HAMT improved token-level alignment and history use across VLN settings [5, 12, 35]. More recent general-domain VLN work such as FlexVLN, SmartWay, NavMorph, and NavForesee also reinforces the need for cross-task adaptation, backtracking-aware reasoning, and more structured predictive use of navigation history [22, 32, 45, 47]. Aerial studies expose the same failure mode from different angles: AerialVLN introduced recurrent aerial policies with look-ahead guidance, FELA strengthens fine-grained alignment in aerial vision-dialog navigation, STMR and CityNavAgent rely on stronger semantic reasoning, FlightGPT emphasizes interpretable VLM-based planning, and NaVid shows that

video-based state construction can improve long-horizon generalization [3, 11, 24, 33, 46, 48]. A second challenge is that not all historical observations remain useful after a sub-goal has been completed. History-aware VLN models such as DUET show that selective memory matters in continuous navigation, and aerial systems such as OpenFly-Agent, SkyVLN, and FlySearch further support keyframe- or exploration-aware reasoning [6, 10, 19, 25]. Our design separates these roles explicitly: it first constructs instruction-grounded current-view tokens and only then applies relevance-aware temporal aggregation with a sparse grounding side branch.

2.3 Topology-Aware Planning and Policy Optimization

Even with stronger perception and memory, long-horizon UAV-VLN still fails when the decision state is not stable enough to avoid loops, false stopping, or semantically plausible detours. Topological reasoning is useful in VLN because it stabilizes long-range exploration, exposes frontier structure, and reduces local ambiguity [4, 26]. In

aerial navigation, STMR and CityNavAgent use structured spatial priors, GeoNav introduces dual-scale geospatial reasoning for coarse-to-fine aerial search, and Fly0 decouples semantic grounding from geometric planning [11, 42, 43, 48]. Adjacent map-based or control-oriented systems such as SkyVLN and UAV-Flow improve global consistency or low-level execution, but they reformulate the problem as location prediction, map search, or control rather than benchmark-style sequential policy learning [19, 36]. Flight-GPT further explores GRPO-style post-training for aerial VLM agents, but its emphasis remains reasoning-process supervision rather than topology-conditioned state refinement [3]. Our aim is therefore narrower than generic aerial reasoning: we connect semantic grounding, relevance-aware temporal aggregation, and topology-aware decision optimization into one policy-state pipeline for sequential long-horizon UAV-VLN.

3 The Proposed Approach

3.1 Problem Formulation

We formulate UAV-VLN as an instruction-conditioned partially observable Markov decision process (POMDP) $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, O, \Omega, r, \gamma)$, where P and Ω are the transition and observation models. An episode contains instruction $\mathcal{L} = \{w_1, \dots, w_N\}$, initial latent state $s_1 \in \mathcal{S}$, and hidden goal region g . At step t , the policy receives an egocentric RGB-D observation $o_t = (I_t, D_t)$ and an onboard pose estimation $\hat{q}_t = (\hat{x}_t, \hat{y}_t, \hat{z}_t, \hat{\psi}_t)$ in an episode-local frame. It retains $\mathcal{H}_t = \{(o_\tau, \hat{q}_\tau)\}_{\tau \in \mathcal{B}_t}$, where $\mathcal{B}_t = \{\tau \in \mathbb{N} \mid \max(1, t - H) \leq \tau < t\}$ with $\mathcal{B}_1 = \emptyset$, and updates topology graph $\mathcal{G}_t = (\mathcal{V}_t, \mathcal{E}_t)$ from these estimated poses. Its policy information state is

$$x_t = (o_t, \hat{q}_t, \mathcal{H}_t, \mathcal{G}_t, \mathcal{L}), \quad (1)$$

where a rollout is $\xi = (x_1, a_1, \dots, x_T, a_T, x_{T+1})$. The pose estimation $\hat{q}_t$ is maintained from executed egomotion and the simulator's odometry-equivalent onboard stream (e.g., depth plus IMU/GPS/VIO) without goal coordinate. This sensor-side separation is consistent with RGB-D aerial benchmarks and realistic UAV sensor platforms [24, 37]. In contrast, s_t is the privileged simulator state used only by the transition, reward, and evaluation interfaces. The decoder context c_t^{dec} is constructed from the current observation encoder output, the DTA-filtered history representation, the structured prompt extracted from key historical frames, and the topology-aware target produced by LOC. The action policy is therefore written as $\pi_\theta(a_t \mid c_t^{\text{dec}})$ over the discrete action space $\mathcal{A}$, which contains reachable motion choices and the "STOP" action. As the reward depends on post-action progress and stopping behavior, the instantiated step reward is $r_t = r(s_t, a_t, s_{t+1}; g)$. With discounted return $R(\xi) = \sum_{t=1}^T \gamma^{t-1} r_t$, the training objective is

$$\max_{\pi_\theta} \mathbb{E}_{\xi \sim \pi_\theta} [R(\xi)], \quad a_t \sim \pi_\theta(\cdot \mid c_t^{\text{dec}}), \quad (2)$$

where the composite form of r_t is defined in Sec. 3.4. Throughout the method, $s_t, \hat{q}_t, x_t, f_t, \hat{h}_t, p_t^{\text{rag}}$, and v_t^* denotes the latent simulator state, the policy-accessible local pose estimate, the policy information state, the instruction-enhanced current observation feature, the DTA-filtered history representation, the structured history prompt built from key historical frames, and the topology-aware frontier target selected by LOC, respectively. The resulting modeling problem is therefore how to transform x_t into a decision context that

remains semantically faithful to the instruction and topologically robust over long horizons. We evaluate trajectory performance by three standard VLN metrics, including navigation error (NE), success rate (SR), and oracle success rate (OSR).

3.2 Framework Overview

Figure 2 summarizes the full pipeline: *observation (vision + instruction) encoding* $\rightarrow$ *fine-grained semantic and spatial enhancement* $\rightarrow$ *dynamic temporal aggregation (DTA)* $\rightarrow$ *local-optimum cognition (LOC)* $\rightarrow$ *GRPO-based policy refinement*. For clarity, we group these operations into three macro stages. Stage 1 constructs an instruction-grounded current state from the RGB-D observation, the visual encoder, and the topology encoder. Stage 2 performs DTA over the retained history and augments the weighted main branch with a sparse grounding branch that converts up to two high-relevance frames into structured prompt memory. Stage 3 monitors the evolving topology graph, detects stagnation, and conditionally supplies a semantically aligned frontier cue to the decoder. GRPO then refines the trainable policy under the composite reward during training and is removed during inference. This decomposition preserves the explicit causal roles: current-view grounding determines what the agent should attend to at the current step, DTA filters relevant historical evidence, and LOC defines when topology informs the next action.

3.3 Instruction-Grounded Perception and Temporal Modeling

Observation (vision + instruction) encoder. The current-view encoder takes the RGB image I_t as input and yields instruction-agnostic visual tokens, while a text encoder generates instruction tokens and an instruction query. Rather than being incorporated as an additional color channel, the aligned depth map D_t is used by the spatial-cue branch. The topology graph is encoded in parallel and fused with the grounded observation following semantic-spatial enhancement. History, current-observation, and instruction streams remain distinguishable through the [HIS], [OBS], and [NAV] identifiers before being fed into the decoder. Specific backbone choices are described in Sec. 4.2.

Fine-grained semantic and spatial enhancement. Let $f_t^{\text{base}} \in \mathbb{R}^{d_h}$ denote the base current-view token generated from RGB image I_t . In parallel, an object detector extracts K_t landmark proposals $\{(b_t^k, v_t^k)\}_{k=1}^{K_t}$ from I_t , where b_t^k is a landmark region and $v_t^k \in \mathbb{R}^{d_o}$ is its semantic descriptor. Let $\tilde{d}_t^k$ be the median valid value of aligned depth map D_t within region b_t^k . Given camera calibration C , we back-project the proposal center at depth $\tilde{d}_t^k$ and transform the resulting 3D point $\mathbf{u}_t^k$ into the UAV body frame. Defining $\bar{\mathbf{u}}_t^k = \mathbf{u}_t^k / d_{\text{sens}}$, we encode the relative spatial relation as

$$p_t^k = \phi(b_t^k, D_t, C) = [\sin \vartheta_t^k, \cos \vartheta_t^k, \bar{\rho}_t^k, \bar{\zeta}_t^k, \bar{\alpha}_t^k], \quad (3)$$

where ϑ_t^k is the bearing angle in the body frame, $\bar{\rho}_t^k, \bar{\zeta}_t^k$, and $\bar{\alpha}_t^k$ are the clipped normalized range, vertical offset, and box scale, respectively. The box scale is $\sqrt{|b_t^k|/(WH)}$, where $W \times H$ is the image size, and spatial coordinates are normalized by sensor range d_{sens} . Invalid-depth proposals use a learned unknown-spatial embedding. The resulting cue therefore integrates depth-derived bearing and

relative 3D position with image-space box scale. The text encoder outputs the instruction-token matrix $H_{\mathcal{L}}^{\text{tok}} \in \mathbb{R}^{N \times d_h}$ and a pooled instruction query $\bar{e}_{\mathcal{L}} \in \mathbb{R}^{d_h}$. Landmark relevance is modeled as

$$\eta_t^k = \text{softmax}_k \left((W_q^{\text{obj}} \bar{e}_{\mathcal{L}})^\top (W_k^{\text{obj}} [v_t^k \| p_t^k]) \right), \quad (4)$$

and the instruction-grounded current-view feature becomes

$$\hat{f}_t = \text{LN} \left(W_g f_t^{\text{base}} + \sum_{k=1}^{K_t} \eta_t^k W_o [v_t^k \| p_t^k] \right). \quad (5)$$

When no proposals are retained ($K_t = 0$), the object residual is set to zero, yielding $\hat{f}_t = \text{LN}(W_g f_t^{\text{base}})$. This module injects fine-grained object-level semantics and relative position cues into the current observation, preventing the state from being dominated by coarse global appearance.

Visual-topological fusion. Long-horizon aerial navigation also depends on structural reachability. We encode the evolving graph to obtain node embeddings $\{u_t^m\}_{m=1}^{|\mathcal{V}_t|}$, stack them into matrix U_t , and compute the fused current-state token via cross-attention:

$$h_t = \text{LN} \left(\hat{f}_t + \text{MHA}(Q = \hat{f}_t, K = U_t, V = U_t) \right). \quad (6)$$

Here, $\hat{f}_t$ carries fine-grained semantic evidence from the current view, while U_t encodes local reachability and neighborhood structure that cannot be recovered from appearance alone. Cross-attention highlights graph nodes that are both spatially adjacent and semantically aligned with the grounded instruction. The token h_t is both instruction-grounded and topology-aware, preserving a clear interface to the successive decision modules.

Relevance-aware dynamic temporal aggregation. DTA aggregates all frames indexed by $\mathcal{B}_t$, while grounding only high-weight key frames. Taking $\bar{e}_{\mathcal{L}}$ as query, it computes the relevance weight for $\tau \in \mathcal{B}_t$ as

$$\alpha_{t,\tau} = \frac{\exp \left((W_q^{\text{hist}} \bar{e}_{\mathcal{L}})^\top W_h h_\tau / \sqrt{d_h} \right)}{\sum_{j \in \mathcal{B}_t} \exp \left((W_q^{\text{hist}} \bar{e}_{\mathcal{L}})^\top W_h h_j / \sqrt{d_h} \right)}. \quad (7)$$

The main DTA branch combines the weighted full history buffer with the current observation to obtain the filtered temporal state

$$\tilde{h}_t = \text{LN} \left(h_t + \sum_{\tau \in \mathcal{B}_t} \alpha_{t,\tau} W_o h_\tau \right). \quad (8)$$

This primary branch is the recurrent state pathway used by the navigation policy. In addition, DTA includes a sparse key-frame grounding branch to capture delayed landmark cues. We first select the high-weight historical index set

$$\mathcal{I}_t = \text{TopK}_{\tau \in \mathcal{B}_t} (\alpha_{t,\tau}, \min(K_r, |\mathcal{B}_t|)), \quad (9)$$

where $K_r = 2$ in all reported experiments. For an empty initial buffer, the history sum and m_t are set to zero with $\mathcal{I}_t = \emptyset$. When the buffer contains only one frame, that single frame is selected. Otherwise, a frozen grounding model is applied to the selected RGB images, and its category–region outputs are paired with aligned depth through ϕ :

$$\begin{aligned} \{(\kappa_{\tau,m}, b_{\tau,m})\}_{m=1}^{M_\tau} &= \Gamma(I_\tau, \mathcal{L}), \\ \mathcal{Z}_\tau &= \{(\kappa_{\tau,m}, b_{\tau,m}, \phi(b_{\tau,m}, D_\tau, C))\}_{m=1}^{M_\tau}, \end{aligned} \quad (10)$$

for $\tau \in \mathcal{I}_t$. Each tuple thereby encodes the landmark category, image region, and depth-derived relative position. These grounded

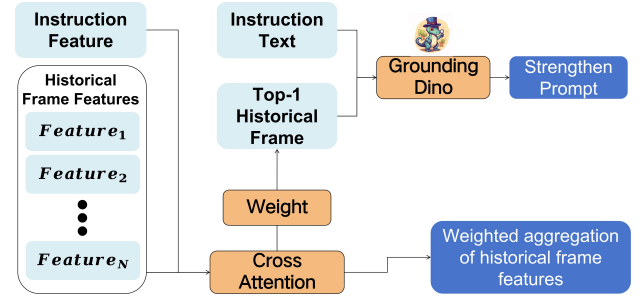


Figure 3: Overview of dynamic temporal aggregation with sparse grounding. The instruction query assigns relevance weights to historical observations to generate an aggregated history representation, while the top- K frames are converted into structured landmark prompts with relative positional cues.

tuples are serialized into a structured historical prompt

$$p_t^{\text{rag}} = \text{Prompt}(\{\mathcal{Z}_\tau\}_{\tau \in \mathcal{I}_t}), \quad (11)$$

and further encoded as an auxiliary decoder memory

$$m_t = \text{LMEnc}(p_t^{\text{rag}}). \quad (12)$$

PROMPT also inserts a short descriptor of each selected frame, ensuring the decoder receives both frame-level context and landmark-level structure. The weighted aggregation branch outputs $\tilde{h}_t$ as the main temporal state, while the sparse grounding branch outputs p_t^{rag} and its encoded memory m_t as auxiliary prompt context. Accordingly, DTA is not merely a top- K retrieval module, but a weighted aggregation module augmented with a key-frame grounding side branch.

3.4 Topology-Aware Decision Optimization

Following current-view grounding and temporal filtering, the remaining challenge lies in decision robustness: even a well-refined semantic state can still induce local loops or semantically plausible yet ineffective detours. To mitigate this issue, we propose topology-aware local-optimum cognition (LOC) and GRPO-based policy optimization.

Local-optimum cognition. Let $\mathcal{V}_t^{\text{exp}} \subseteq \mathcal{V}_t$ denote the explored nodes in the current dynamic topology graph, and let $\mathcal{F}_t \subset \mathcal{V}_t$ denote the reachable frontier nodes that remain unexplored. LOC monitors whether the exploration graph has expanded within the most recent Δ steps. For $t \geq \Delta$, stagnation is detected when no newly explored node appears during that window:

$$\mathbb{I}_{\text{loc}}(t) = \mathbb{I}(t \geq \Delta) \mathbb{I}(\mathcal{V}_t^{\text{exp}} \setminus \mathcal{V}_{t-\Delta}^{\text{exp}} = \emptyset). \quad (13)$$

For nonempty $\mathcal{F}_t$, let $e_v^{\text{node}} \in \mathbb{R}^{d_h}$ and v_t denote the semantic embedding of frontier node v , and the current graph node, respectively. LOC normalize semantic agreement and graph distance over $\mathcal{F}_t$. With distance extrema d_t^{min} and d_t^{max} and a constant $\epsilon_{\text{loc}} > 0$, it computes

$$\begin{aligned} \bar{s}_t(v) &= \frac{1 + \cos(e_v^{\text{node}}, \bar{e}_{\mathcal{L}})}{2}, \\ \bar{d}_{\mathcal{G}}(v_t, v) &= \frac{d_{\mathcal{G}}(v_t, v) - d_t^{\text{min}}}{d_t^{\text{max}} - d_t^{\text{min}} + \epsilon_{\text{loc}}}, \\ v_t^* &= \arg \max_{v \in \mathcal{F}_t} [\bar{s}_t(v) - \bar{d}_{\mathcal{G}}(v_t, v)]. \end{aligned} \quad (14)$$

The two normalized terms represent instruction compatibility and relative traversal cost and are equally weighted. When LOC is active with a nonempty frontier set, the decoder cue encodes the selected node as $e_t^{\text{loc}} = \text{Enc}_{\text{loc}}(v_t^*)$; otherwise it is set to a zero vector.

Composite reward design. The transition reward balances four terms: progress reward, goal-completion reward, semantic matching reward, and path compliance penalty, which is defined as

$$r_t = r_t^{\text{prog}} + r_t^{\text{goal}} + r_t^{\text{sem}} + r_t^{\text{path}}. \quad (15)$$

The goal region g and benchmark-consistent distance $d(s, g)$ are available only to the simulator-side reward and evaluation modules, and are not included in c_t^{dec} . To prevent distance measured in meters from dominating bounded semantic and terminal terms, we define the episode-normalized distance as $\bar{d}_t = \text{clip}(d(s_t, g)/D_\xi, 0, 1)$, where $D_\xi = \max\{d(s_1, g), \epsilon_n\}$. The progress term rewards the agent for moving closer to the goal:

$$r_t^{\text{prog}} = \lambda_p(\bar{d}_t - \bar{d}_{t+1}). \quad (16)$$

The goal-completion term assigns a terminal bonus when the agent reaches the goal region and executes the “STOP” action:

$$r_t^{\text{goal}} = \begin{cases} \lambda_g, & a_t = \text{“STOP” and } d(s_{t+1}, g) \leq \epsilon_g; \\ 0, & \text{otherwise.} \end{cases} \quad (17)$$

The semantic matching term scores image-text consistency of the reached observation:

$$r_t^{\text{sem}} = \lambda_s \cos(\psi_I(I_{t+1}), \psi_T(\mathcal{L})). \quad (18)$$

ψ_I and ψ_T are frozen CLIP image and text encoders used only for reward computation. The same frozen image encoder is reused for comparing observations in the descriptor space. The path compliance term penalizes backtracking and persistent deviation from the goal. Let $t_w = \max\{1, t - w\}$ denote the first index in the revisit window:

$$r_t^{\text{path}} = -\lambda_r \sum_{j=t_w}^{t-1} \exp(-\mu(t-j)) \mathbb{I}_{\text{rev}}(t+1, j) - \lambda_c \bar{n}_{t+1}^{\text{dev}} \bar{d}_{t+1}, \quad (19)$$

where the deviation counter is initialized as $n_1^{\text{dev}} = 0$ and evolves as

$$n_{t+1}^{\text{dev}} = \begin{cases} n_t^{\text{dev}} + 1, & d(s_{t+1}, g) > d(s_t, g) + \epsilon_d, \\ 0, & \text{otherwise,} \end{cases} \quad (20)$$

$$\bar{n}_{t+1}^{\text{dev}} = \min\{n_{t+1}^{\text{dev}}, w\}/w.$$

and the revisit indicator is defined as

$$s_{\text{vis}}(t, j) = \frac{1 + \cos(\psi_I(I_t), \psi_I(I_j))}{2}, \\ \mathbb{I}_{\text{rev}}(t, j) = \mathbb{I}(d_{3D}(\hat{\mathbf{q}}_t, \hat{\mathbf{q}}_j) \leq \delta_r) \cdot \mathbb{I}(|\hat{z}_t - \hat{z}_j| \leq \delta_h) \cdot \mathbb{I}(s_{\text{vis}}(t, j) \geq \rho_v). \quad (21)$$

Here d_{3D} uses only the positional components of the estimated local poses, $\hat{z}_t$ is the altitude component, ϵ_g is the success radius, and $s_{\text{vis}}(t, j) \in [0, 1]$ is the normalized cosine similarity between frozen global image descriptors. This descriptor-based verification compares observations in a learned embedding space, eliminating the need for pixel-coordinate alignment of detection boxes across different camera views. The joint estimated-spatial, altitude, and visual tests suppress false revisit matches caused by nearby but visually distinct locations. Since $\bar{d}_t, \bar{n}_t^{\text{dev}} \in [0, 1]$, both progress

and persistent-deviation terms are bounded independently of route length and metric scale.

GRPO-based policy refinement. The decoder context for action prediction is assembled as

$$c_t^{\text{dec}} = \text{Fuse}_{\text{dec}}(\tilde{h}_t, m_t, e_t^{\text{loc}}, H_{\mathcal{L}}^{\text{tok}}), \quad (22)$$

where $H_{\mathcal{L}}^{\text{tok}}$ is the instruction-token matrix defined above, and the remaining inputs are the DTA state, structured-prompt memory, and conditional LOC cue. Instead of training a separate critic network, we adopt GRPO for policy refinement over this decoder context [3, 30, 31]. For each instruction and start state, we sample a group of M trajectories $\{\xi_i\}_{i=1}^M$ from $\pi_{\theta_{\text{old}}}$, and compute group-normalized advantages as

$$A_i = \frac{R(\xi_i) - \mu_R}{\sigma_R + \epsilon_A}, \\ \mu_R = \frac{1}{M} \sum_{i=1}^M R(\xi_i), \\ \sigma_R^2 = \frac{1}{M} \sum_{i=1}^M (R(\xi_i) - \mu_R)^2, \quad (23)$$

where $\epsilon_A > 0$ handles identical group returns. The clipped GRPO surrogate is defined as

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E} \left[\frac{1}{M} \sum_{i=1}^M \frac{1}{T_i} \sum_{t=1}^{T_i} \left(\ell_{i,t}(\theta) - \beta D_{i,t}^{\text{KL}}(\theta) \right) \right], \\ \ell_{i,t}(\theta) = \min \left(\rho_{i,t}(\theta) A_i, \text{clip}(\rho_{i,t}(\theta), 1 - \epsilon_c, 1 + \epsilon_c) A_i \right), \quad (24)$$

where

$$\rho_{i,t}(\theta) = \frac{\pi_{\theta}(a_{i,t} \mid c_{i,t}^{\text{dec}})}{\pi_{\theta_{\text{old}}}(a_{i,t} \mid c_{i,t}^{\text{dec}})}, \\ D_{i,t}^{\text{KL}}(\theta) = D_{\text{KL}} \left(\pi_{\theta}(\cdot \mid c_{i,t}^{\text{dec}}) \parallel \pi_{\text{ref}}(\cdot \mid c_{i,t}^{\text{dec}}) \right). \quad (25)$$

Here T_i is the trajectory length. The old policy $\pi_{\theta_{\text{old}}}$ is the rollout snapshot used only in the importance ratio and is refreshed between update rounds, whereas π_{ref} is a frozen copy of the behavior-cloned policy at the start of GRPO fine-tuning. The per-decision KL divergence term anchors the learned policy to the supervised initialization, while clipping constrains each update relative to the rollout policy. Maximizing $\mathcal{J}_{\text{GRPO}}$ couples policy refinement to the DTA state, sparse prompt memory, and LOC cue.

3.5 Training and Inference

For training, we first initialize the decoder with behavior cloning on expert trajectories, ensuring the policy starts GRPO fine-tuning from a stable instruction-following regime:

$$\mathcal{L}_{\text{BC}} = - \sum_t \log \pi_{\theta}(a_t^* \mid c_t^{\text{dec}}), \quad (26)$$

where a_t^* is the expert action and c_t^{dec} follows the decoder interface above. We then fine-tune the trainable policy modules with GRPO, yielding

$$\mathcal{L}_{\text{train}} = \begin{cases} \mathcal{L}_{\text{BC}}, & \text{stage 1,} \\ -\mathcal{J}_{\text{GRPO}}, & \text{stage 2.} \end{cases} \quad (27)$$

During inference, the system first encodes the current RGB image and instruction, derives depth-based semantic-spatial cues, and

Table 1: Comparison of performance on the validation seen and unseen splits of AerialVLN-S. The best results are highlighted in bold and the second-best ones are underlined.

Method	Validation Seen			Validation Unseen		
	NE↓	SR↑	OSR↑	NE↓	SR↑	OSR↑
Random	109.60	0.00	0.00	149.70	0.00	0.00
Action Sampling	213.80	0.90	5.70	237.60	0.20	1.10
Seq2Seq [1]	146.00	4.80	19.80	218.90	2.30	11.70
CMA [24]	121.00	3.00	23.20	172.10	3.20	16.00
LAG [24]	90.20	7.20	15.70	127.90	5.10	10.50
STMR [11]	63.82	30.15	56.30	97.86	18.35	33.40
CityNavAgent [48]	66.34	28.97	58.10	100.31	16.70	31.90
NavGPT [52]	59.41	34.04	61.20	91.60	26.50	46.20
SPF [13]	40.12	45.89	72.38	71.34	35.12	58.30
OpenFly-Agent [10]	66.93	36.98	61.80	90.67	29.31	47.80
UAV-Flow [36]	58.16	35.81	58.34	82.38	28.85	46.10
Fly0 [43]	27.19	70.43	81.20	51.23	60.07	<u>69.22</u>
FlightGPT [3]	38.17	44.62	72.98	62.88	31.75	60.09
Ours	<u>28.34</u>	71.12	82.60	50.69	61.38	71.01

Table 2: Comparison of performance on the test set of OpenFly. The best results are highlighted in bold and the second-best ones are underlined.

Method	NE↓	SR↑	OSR↑
Random	165.38	0.00	0.00
OpenUAV [37]	74.63	17.81	31.16
UAV-Flow [36]	69.51	32.14	44.01
AerialVLN [24]	94.63	6.71	19.91
OpenFly [10]	64.08	33.61	49.72
NavGPT [52]	61.57	29.28	43.48
STMR [11]	66.54	27.43	42.64
CityNavAgent [48]	70.24	25.33	39.75
SPF [13]	46.57	42.94	55.79
Fly0 [43]	<u>29.47</u>	<u>64.67</u>	<u>72.81</u>
FlightGPT [3]	41.29	56.23	69.03
Ours	26.14	67.31	77.32

fuses the grounded observation with the local topology to obtain h_t . DTA then weights the retained history to generate the main temporal state $\tilde{h}_t$, while its sparse branch converts up to two high-relevance historical RGB-D frames into the structured prompt p_t^{rag} and auxiliary memory m_t . When LOC detects stagnation and a reachable frontier exists, it conditionally appends the selected frontier cue e_t^{loc} to the decoder context. The decoder predicts the next action from c_t^{dec} . GRPO group sampling and policy updates are used only during training and are not part of the inference path.

4 Experimental Results and Analysis

4.1 Datasets and Evaluation Metrics

AerialVLN. AerialVLN [24] contains city-scale continuous UAV trajectories with long natural-language instructions and is one of the pioneering UAV-VLN benchmarks. We report AerialVLN-S validation performance on the public seen and unseen splits with NE, SR, and OSR.

OpenFly. OpenFly [10] is a large-scale aerial VLN benchmark comprising 100K trajectories across 18 scenes rendered from multiple engines and representations. Our model is trained on the OpenFly training set and evaluated on the test set under the benchmark’s standard evaluation protocol.

We report three standard metrics: navigation error (NE) measuring the terminal straight-line distance to the target; success rate (SR) denoting the fraction of episodes stopping within 20 m of the target, and oracle success rate (OSR) denoting the fraction entering that radius at any trajectory step.

4.2 Implementation Details

We utilize Oriented R-CNN [41] through MMRotate [53] for current-view enhancement, Grounding DINO [23] for sparse history grounding, Qwen2.5-VL-7B [2] with LoRA [14] as the action decoder, frozen CLIP [27] for reward and revisit descriptors, and a light-weight GCN [16] for topology modeling. All the main results are averaged over three random seeds. The complete geometry, history, training, reward, revisit, parameter-count, and efficiency configurations are provided in the supplementary material.

4.3 Comparison with Strong Baselines

On AerialVLN-S validation. The comparison results in Table 1 show a clear progression from early recurrent baselines (e.g., Seq2Seq, CMA) to SOTA aerial reasoning systems, indicating that long-horizon UAV-VLN depends on accurate landmark grounding, effective history modeling, and stable sequential decisions. Our proposed framework achieves the highest SR on both validation splits and the highest OSR on the seen split, while remaining competitive in terms of NE.

The unseen split serves as a rigorous stress test, as distribution shift amplifies landmark ambiguity and long-horizon error accumulation. Compared to Fly0 [43], our method improves unseen SR from 60.07 to 61.38 and reduces NE from 51.23 to 50.69. The three-seed results reliably quantify run-to-run variance for our method, but they do not constitute a paired significance test against baseline values obtained from separate reports. The SR-NE trade-off on the unseen split is visualized in Figure 4.

On OpenFly. Our model is trained on the official training split on OpenFly and evaluated on the test set under the standard protocol. As summarized in Table 2, our method achieves the lowest NE and the highest SR and OSR among all reported methods.

Compared with Fly0 [43], our method reduces NE from 29.47 to 26.14, and improves SR from 64.67 to 67.31, and raises OSR from

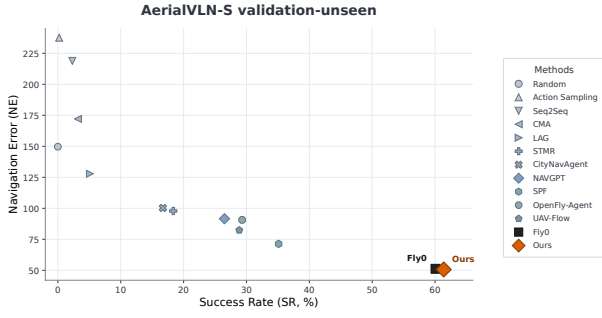


Figure 4: SR-NE comparison on the AerialVLN-S validation-unseen split. Each marker denotes one method; higher SR and lower NE indicate better performance. Our full model is highlighted for clarity.

Table 3: Cumulative ablation on the AerialVLN-S validation-unseen split. Each row adds one component in pipeline order to the same base model. The last row corresponds to our full model and matches the validation-unseen result reported in Table 1.

Variant	NE↓	SR↑	OSR↑
Baseline	72.10	35.85	59.92
+ Semantic enhancement	67.84	40.96	62.88
+ DTA	62.11	46.74	66.41
+ Sparse grounding	58.27	51.32	68.45
+ LOC	54.06	56.47	69.94
+ GRPO (Full)	50.69	61.38	71.01

72.81 to 77.32. The more pronounced performance gain may be attributed to its fine-grained landmark and route cues.

Additional three-seed average results on the AerialVLN-S test split (53.5 NE, 59.0 SR, and 69.0 OSR) and the OpenFly validation split (25.0 NE, 68.5 SR, and 78.5 OSR) are detailed in the supplementary material.

4.4 Ablation Study

Table 3 reports fixed-order cumulative additions on the AerialVLN-S validation-unseen split. The consistent monotonic improvements across NE, SR, and OSR gains clearly validate the effectiveness of our coupled pipeline design.

Additional leave-one-out and alternative history results achieve SR of 46.74 without DTA, 50.90 without LOC, and 55.43 with simple history aggregation, compared to 61.38 for the full model. Reward ablations demonstrate consistent gains over Basic GRPO (57.42 SR): semantic-only, progress-only, goal/stop-only, and path-compliance-only obtain 58.52, 59.52, 58.24, and 58.91 SR, respectively, while the full composite reaches 61.38. Under Gaussian image noise with $\sigma = 0.10$, our model yields 58.40 NE, 55.10 SR, and 64.85 OSR. Under moderate accumulated localization drift, our method reaches 56.71 NE, 56.98 SR, and 66.23 OSR. Full protocols, metrics, and evaluation boundaries are summarized in the supplementary material.

4.5 Qualitative Analysis

Figure 5 shows a representative successful case. The agent follows the main road, passes the fountain on its left, turns right at the red-roof building, and stops above the small square beside the clock tower. The trajectory follows the landmark order described

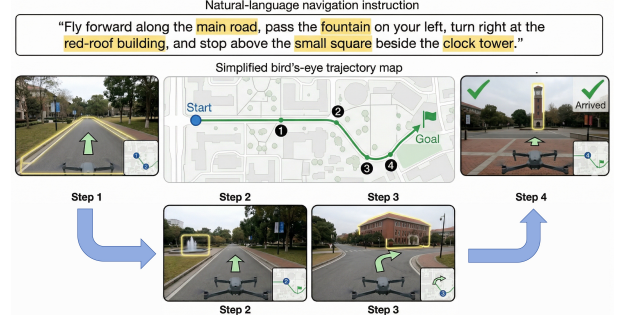


Figure 5: Qualitative example of a successful navigation episode by the proposed method. It displays the instruction, four egocentric UAV observations, and the corresponding bird's-eye trajectory. The agent follows the specified landmark order and terminates near the target region.

in the instruction, illustrating semantically coherent long-horizon execution.

4.6 Discussion

The experimental results support the perspective that treats long-horizon UAV-VLN as a policy-state construction problem, rather than as isolated perception, memory, and planning subproblems. Results across AerialVLN-S and OpenFly, together with the ablation studies, consistently support instruction-grounded observations, selective history, and topology cues as a coupled policy state. These results should be interpreted within the evaluation design: three-seed statistics quantify run-to-run stability, Table 3 reports cumulative additions, and supplementary experiments further validate the effect of DTA and LOC. Dependence on external perception, onboard pose, and lightweight topology motivates further deployment-oriented validation beyond benchmark navigation.

5 Conclusion

This work presents a unified semantic-to-decision framework that connects current-view grounding, relevance-aware temporal aggregation, and topology-aware decision refinement for long-horizon UAV-VLN. DTA preserves retained history and recovers delayed semantic cues through sparse landmark prompts. And LOC conditionally supplies a frontier cue, when the topology indicates stagnation. Extensive experiments on AerialVLN-S and OpenFly, together with component, reward, simulated image-noise, and localization-drift analyses, validate that these modules function as a coupled state construction process, enabling more stable long-horizon navigation behavior. For future work, we will evaluate our framework under more challenging real-world conditions, including adverse weather, dynamic obstacles, stronger and sensor-coupled perception/localization corruption, and real closed-loop flight while exploring richer semantic-geometric memory and broader cross-environment validation.

Acknowledgments

This work was supported in part by the Beijing Natural Science Foundation (Grants 4242044 and L259044), the Research Program of the State Key Laboratory of Virtual Reality Technology and Systems, and the Fundamental Research Funds for the Central Universities.

References

- [1] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. 2018. Vision-and-Language Navigation: Interpreting Visually-Grounded Navigation Instructions in Real Environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3674–3683.
- [2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao 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. 2025. Qwen2.5-VL Technical Report. arXiv preprint arXiv:2502.13923 (2025).
- [3] Hengxing Cai, Jinhan Dong, Jingjun Tan, Jingcheng Deng, Sihang Li, Zhifeng Gao, Haidong Wang, Zicheng Su, Agachai Sumalee, and Renxin Zhong. 2025. 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*. 6659–6676.
- [4] Kevin Chen, Junshen K. Chen, Jo Chuang, Marynel Vázquez, and Silvio Savarese. 2021. Topological Planning with Transformers for Vision-and-Language Navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11271–11281.
- [5] Shizhe Chen, Pierre-Louis Guhur, Cordelia Schmid, and Ivan Laptev. 2021. History Aware Multimodal Transformer for Vision-and-Language Navigation. In *Proceedings of the Annual Conference on Neural Information Processing Systems*, Vol. 34. 5834–5847.
- [6] Shizhe Chen, Pierre-Louis Guhur, Makarand Tapaswi, Cordelia Schmid, and Ivan Laptev. 2022. Think Global, Act Local: Dual-Scale Graph Transformer for Vision-and-Language Navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16516–16526.
- [7] Bowei Du, Yecheng Huang, Jiaxin Chen, and Di Huang. 2023. Adaptive Sparse Convolutional Networks with Global Context Enhancement for Faster Object Detection on Drone Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13435–13444.
- [8] Yue Fan, Winson Chen, Tongzhou Jiang, Chun Zhou, Yi Zhang, and Xin Eric Wang. 2023. Aerial Vision-and-Dialog Navigation. In *Findings of the Association for Computational Linguistics: ACL 2023*. 3043–3061.
- [9] Daniel Fried, Ronghang Hu, Volkan Cirić, Anna Rohrbach, Jacob Andreas, Louis-Philippe Morency, Taylor Berg-Kirkpatrick, Kate Saenko, Dan Klein, and Trevor Darrell. 2018. Speaker-Follower Models for Vision-and-Language Navigation. In *Proceedings of the Annual Conference on Neural Information Processing Systems*, Vol. 31. 3314–3325.
- [10] Yunpeng Gao, Chenhui Li, Zhongrui You, Junli Liu, Zhen Li, Peng Chen, Qizhi Chen, Zhonghan Tang, Liansheng Wang, Penghui Yang, Yiwen Tang, Yuhang Tang, Shuai Liang, Songyi Zhu, Ziqin Xiong, Yifei Su, Xinyi Ye, Jianan Li, Yan Ding, Dong Wang, Xuelong Li, Zhigang Wang, and Bin Zhao. 2026. OpenFly: A Comprehensive Platform for Aerial Vision-Language Navigation. In *Proceedings of the International Conference on Learning Representations*.
- [11] Yunpeng Gao, Zhigang Wang, Pengfei Han, Linglin Jing, Dong Wang, and Bin Zhao. 2024. Exploring Spatial Representation to Enhance LLM Reasoning in Aerial Vision-Language Navigation. arXiv preprint arXiv:2410.08500 (2024).
- [12] Yicong Hong, Cristian Rodriguez-Opazo, Yuankai Qi, Qi Wu, and Stephen Gould. 2020. A Language and Visual Entity Relationship Graph for Agent Navigation. In *Proceedings of the Annual Conference on Neural Information Processing Systems*, Vol. 33. 7685–7696.
- [13] Chih Yao Hu, Yang-Sen Lin, Yuna Lee, Chih-Hai Su, Jie-Ying Lee, Shr-Ruei Tsai, Chin-Yang Lin, Kuan-Wen Chen, Tsung-Wei Ke, and Yu-Lun Liu. 2025. See, Point, Fly: A Learning-Free VLM Framework for Universal Unmanned Aerial Navigation. In *Proceedings of the 9th Conference on Robot Learning*. 4697–4708.
- [14] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *Proceedings of the International Conference on Learning Representations*.
- [15] Yecheng Huang, Jiaxin Chen, and Di Huang. 2022. UFMP-Det: Toward Accurate and Efficient Object Detection on Drone Imagery. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 1026–1033.
- [16] Thomas N. Kipf and Max Welling. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *Proceedings of the International Conference on Learning Representations*.
- [17] Jacob Krantz, Erik Wijmans, Arjun Majumdar, Dhruv Batra, and Stefan Lee. 2020. Beyond the Nav-Graph: Vision-and-Language Navigation in Continuous Environments. In *Proceedings of the European Conference on Computer Vision*. 104–120.
- [18] Jungdae Lee, Taiki Miyayoshi, Shuhei Kurita, Koya Sakamoto, Daichi Azuma, Yutaka Matsuo, and Nakamasa Inoue. 2025. CityNav: A Large-Scale Dataset for Real-World Aerial Navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5912–5922.
- [19] Tianshun Li, Tianyi Huai, Zhen Li, Yichun Gao, Haoang Li, and Xinhui Zheng. 2025. SkyVLN: Vision-and-Language Navigation and NMPC Control for UAVs in Urban Environments. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*. 17199–17206.
- [20] Wenhao Li, Zimeng Wu, Yu Wu, Zehua Fu, and Jiaxin Chen. 2026. Visual Prototype Conditioned Focal Region Generation for UAV-Based Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3772–3782.
- [21] Peican Lin, Gan Sun, Chenxi Liu, Fazeng Li, Weihong Ren, and Yang Cong. 2025. OpenVLN: Open-world Aerial Vision-Language Navigation. arXiv preprint arXiv:2511.06182 (2025).
- [22] Fei Liu, Shichao Xie, Minghua Luo, Zedong Chu, Junjun Hu, Xiaolong Wu, and Mu Xu. 2025. NavForesee: A Unified Vision-Language World Model for Hierarchical Planning and Dual-Horizon Navigation Prediction. arXiv preprint arXiv:2512.01550 (2025).
- [23] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. 2024. Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection. In *Proceedings of the European Conference on Computer Vision*. 38–55.
- [24] Shubo Liu, Hongsheng Zhang, Yuankai Qi, Peng Wang, Yanning Zhang, and Qi Wu. 2023. AerialVLN: Vision-and-Language Navigation for UAVs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 15384–15394.
- [25] Adam Pardy, Dominik Matuszek, Mateusz Przybieracz, Marek Cygan, Bartosz Zieliński, and Maciej Wolczyk. 2025. FlySearch: Exploring How Vision-Language Models Explore. In *Proceedings of the Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, Vol. 38.
- [26] Yuankai Qi, Zizheng Pan, Yicong Hong, Ming-Hsuan Yang, Anton van den Hengel, and Qi Wu. 2021. The Road to Know-Where: An Object-and-Room Informed Sequential BERT for Indoor Vision-Language Navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1655–1664.
- [27] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models from Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*. 8748–8763.
- [28] Oleg Sautenkov, Yasheer Yaqoot, Artem Lykov, Muhammad Ahsan Mustafa, Grik Tadevosyan, Aibek Akhmetkazy, Miguel Altamirano Cabrera, Mikhail Martynov, Sausar Karaf, and Dzmitry Tsetserukou. 2025. UAV-VLA: Vision-Language-Action System for Large Scale Aerial Mission Generation. In *Proceedings of the 20th ACM/IEEE International Conference on Human-Robot Interaction*. 1588–1592.
- [29] Pranav Saxena, Nishant Raghuvarshi, and Neema Goveas. 2025. UAV-VLN: End-to-End Vision Language Guided Navigation for UAVs. arXiv preprint arXiv:2504.21432 (2025).
- [30] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal Policy Optimization Algorithms. arXiv preprint arXiv:1707.06347 (2017).
- [31] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. arXiv preprint arXiv:2402.03300 (2024).
- [32] Xiangyu Shi, Zerui Li, Wenqi Lyu, Jiatong Xia, Feras Dayoub, Yanyuan Qiao, and Qi Wu. 2025. SmartWay: Enhanced Waypoint Prediction and Backtracking for Zero-Shot Vision-and-Language Navigation. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*. 16923–16930.
- [33] Yifei Su, Dong An, Kehan Chen, Weichen Yu, Baiyang Ning, Yonggen Ling, Yan Huang, and Liang Wang. 2025. Learning Fine-Grained Alignment for Aerial Vision-Dialog Navigation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 7060–7068.
- [34] Yonglin Tian, Fei Lin, Yiduo Li, Tengchao Zhang, Qiyao Zhang, Xuan Fu, Jun Huang, Xingyuan Dai, Yutong Wang, Chunwei Tian, Bai Li, Yisheng Lv, Levente Kovács, and Fei-Yue Wang. 2025. UAVs Meet LLMs: Overviews and Perspectives Toward Agentic Low-Altitude Mobility. arXiv preprint arXiv:2501.02341 (2025).
- [35] Xin Wang, Qiuyuan Huang, Asli Celikyilmaz, Jianfeng Gao, Dinghan Shen, Yuanfang Wang, William Yang Wang, and Lei Zhang. 2019. Reinforced Cross-Modal Matching and Self-Supervised Imitation Learning for Vision-Language Navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6622–6631.
- [36] Xiangyu Wang, Donglin Yang, Yue Liao, Wenhao Zheng, Wenjun Wu, Bin Dai, Hongsheng Li, and Si Liu. 2025. UAV-Flow Colosseo: A Real-World Benchmark for Flying-on-a-Word UAV Imitation Learning. In *Proceedings of the Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, Vol. 38.
- [37] Xiangyu Wang, Donglin Yang, Ziqin Wang, Hohin Kwan, Jinyu Chen, Wenjun Wu, Hongsheng Li, Yue Liao, and Si Liu. 2025. Towards Realistic UAV Vision-Language Navigation: Platform, Benchmark, and Methodology. In *Proceedings of the International Conference on Learning Representations*.

- [38] Ke Wu, Jiaxin Chen, and Miao Wang. 2024. Domain Adaptive Object Detection for UAV-based Images by Robust Representation Learning and Multiple Pseudo-label Aggregation. In *Proceedings of the 1st International Workshop on Efficient Multimedia Computing under Limited*. 59–67.
- [39] Ke Wu, Yanan Zhang, Yingjie Gao, Wenhao Li, Chenyu Zhou, XinZhu Ma, Jiaxin Chen, and Di Huang. 2026. DroneFINE: Domain-Aware Parameter-Efficient Fine-Tuning of Vision-Language Detectors for Drone Images. In *Proceedings of the European Conference on Computer Vision*. To appear.
- [40] Ruipu Wu, Yige Zhang, Jinyu Chen, Linjiang Huang, Shifeng Zhang, Xu Zhou, Liang Wang, and Si Liu. 2025. AeroDuo: Aerial Duo for UAV-based Vision and Language Navigation. In *Proceedings of the 33rd ACM International Conference on Multimedia*. 2576–2585.
- [41] Xingxing Xie, Gong Cheng, Jiabao Wang, Xiwen Yao, and Junwei Han. 2021. Oriented R-CNN for Object Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3500–3509.
- [42] Haotian Xu, Yue Hu, Chen Gao, Zhengqiu Zhu, Yong Zhao, and Quanjun Yin. 2026. GeoNav: Empowering MLLMs with Dual-Scale Geospatial Reasoning for Language-Goal Aerial Navigation. *Pattern Recognition* 177 (2026), 113365.
- [43] Zhenxing Xu, Yihong Lu, Weidong Bao, Zhengqiu Zhu, Jingxuan Zhou, Zhichuang Wang, Ji Wang, Lihua Liu, and Wei He. 2026. Fly0: Persistent Metric Anchoring for Zero-Shot Aerial Vision-Language Navigation. arXiv preprint arXiv:2602.15875 (2026).
- [44] Fanglong Yao, Youzhi Liu, Wenyi Zhang, Zhengqiu Zhu, Chenglong Li, Nayu Liu, Peng Hu, Yuanchang Yue, Kaiwen Wei, Xin He, Xudong Zhao, Zihan Wei, Haotian Xu, Zhiyuan Wang, Gujie Shao, Liu Yang, Dan Zhao, and Yong Yang. 2025. AeroVerse-Review: Comprehensive Survey on Aerial Embodied Vision-and-Language Navigation. *The Innovation Informatics* 1, 1 (2025), 100015.
- [45] Xuan Yao, Junyu Gao, and Changsheng Xu. 2025. NavMorph: A Self-Evolving World Model for Vision-and-Language Navigation in Continuous Environments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5536–5546.
- [46] Jiazhao Zhang, Kunyu Wang, Rongtao Xu, Gengze Zhou, Yicong Hong, Xiaomeng Fang, Qi Wu, Zhizheng Zhang, and He Wang. 2024. NaVid: Video-Based VLM Plans the Next Step for Vision-and-Language Navigation. In *Proceedings of Robotics: Science and Systems*, Vol. 20. 16 pages.
- [47] Siqi Zhang, Yanyuan Qiao, Qunbo Wang, Longteng Guo, Zhihua Wei, and Jing Liu. 2025. FlexVLN: Flexible Adaptation for Diverse Vision-and-Language Navigation Tasks. arXiv preprint arXiv:2503.13966 (2025).
- [48] Weichen Zhang, Chen Gao, Shiquan Yu, Ruiying Peng, Baining Zhao, Qian Zhang, Jinqiang Cui, Xinlei Chen, and Yong Li. 2025. 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)*. 31292–31309.
- [49] Xinyuan Zhang, Yonglin Tian, Fei Lin, Yue Liu, Jing Ma, Kornélia Sára Szatmáry, and Fei-Yue Wang. 2025. LogisticsVLN: Vision-Language Navigation for Low-Altitude Terminal Delivery Based on Agentic UAVs. arXiv preprint arXiv:2505.03460 (2025).
- [50] Ganlong Zhao, Guanbin Li, Jia Pan, and Yizhou Yu. 2025. Aerial Vision-and-Language Navigation with Grid-based View Selection and Map Construction. arXiv preprint arXiv:2503.11091 (2025).
- [51] Ji Zhao and Xiao Lin. 2025. General-Purpose Aerial Intelligent Agents Empowered by Large Language Models. arXiv preprint arXiv:2503.08302 (2025).
- [52] Gengze Zhou, Yicong Hong, and Qi Wu. 2024. NavGPT: Explicit Reasoning in Vision-and-Language Navigation with Large Language Models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 7641–7649.
- [53] Yue Zhou, Xue Yang, Gefan Zhang, Jiabao Wang, Yanyi Liu, Liping Hou, Xue Jiang, Xingzhao Liu, Junchi Yan, Chengqi Lyu, Wenwei Zhang, and Kai Chen. 2022. MMRotate: A Rotated Object Detection Benchmark Using PyTorch. In *Proceedings of the 30th ACM International Conference on Multimedia*. 7331–7334.