FOREFLY: A DUAL-HORIZON WORLD ACTION MODEL FOR AERIAL VISION-LANGUAGE NAVIGATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Aerial Vision-Language Navigation (AVLN) requires UAVs to maintain reliable instruction following over long trajectories in complex 3D environments. However, existing AVLN approaches are predominantly reactive or limited to single-horizon prediction, overlooking complementary future cues across different temporal horizons. To address this limitation, we propose ForeFly, a dual-horizon latent world action model that predicts both a proximal future for local continuity and an adaptive route-critical future for long-range guidance. Horizon-specific foresight queries are primed with recent and route-critical visual memories, providing history-aware context for future prediction. To exploit their distinct roles in action generation, we introduce Foresight-Guided Action Refinement (FGAR), which asymmetrically exploits proximal foresight for local action enhancement and route-critical foresight for feature-wise correction and route-level guidance. Experiments on the TravelUAV and UAV-ON benchmarks show that ForeFly consistently outperforms strong baselines across seen and unseen settings, validating the effectiveness of dual-horizon foresight and FGAR learning. **The code is available in the supplementary materials.**

1 INTRODUCTION

Aerial Vision-Language Navigation (AVLN) aims to enable unmanned aerial vehicles (UAVs) to understand natural-language instructions and autonomously navigate in complex, open-world three-dimensional environments Liu et al. (2023); Fan et al. (2022). Compared with conventional embodied navigation on the ground Anderson et al. (2017); Ku et al. (2020); Krantz et al. (2020); Qi et al. (2020); Blukis et al. (2018), AVLN poses challenges due to the large-scale search space, continuous 3D motion, drastic viewpoint changes, and weak geometric constraints in aerial scenes Chen et al. (2019); Gao et al. (2026); Lin et al. (2025); Dai et al. (2026). Moreover, language instructions in aerial navigation involve distant landmarks, route-level spatial relations, and long-horizon events, requiring the agent to maintain coherent visual-language understanding over extended trajectories Liu et al. (2023); Sun et al. (2026b); Lee et al. (2024); Gu et al. (2026). These characteristics make AVLN a challenging yet important task for building robust autonomous aerial agents.

Recent advances in vision-language navigation Gao et al. (2026); Zheng et al. (2026b); Lin et al. (2025); Jiang et al. (2026) have explored end-to-end frameworks to align visual observations with language instructions and translate multimodal understanding into navigation actions. Existing methods typically encode the current view, instruction context, and trajectory history to predict the next waypoint or control command, enabling UAVs to follow complex routes in outdoor environments. Despite their progress, most of these methods predominantly operate within a reactive paradigm (Figure 1(a)), where actions are inferred mainly from already observed visual-language evidence. As a result, they lack an explicit mechanism to anticipate future visual states and navigation-critical events before they become immediately observable.

The absence of anticipatory modeling is particularly problematic in aerial navigation. Under continuous 3D motion and rapidly changing viewpoints, local decision errors tend to propagate into substantial trajectory deviations Lin et al. (2025); Wang et al. (2025). Missing a turning point or failing to identify a landmark transition can cause the agent to drift away from the instruction-relevant route, making subsequent recovery increasingly difficult. Therefore, effective AVLN agents should go beyond reactive action prediction and reason about instruction-relevant future states.

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

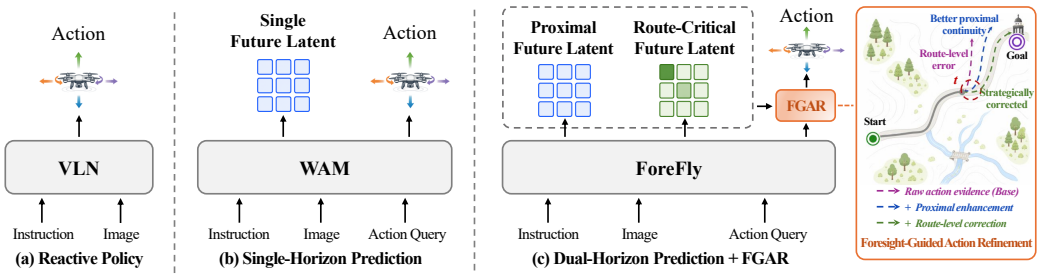

Figure 1: (a) Reactive policies predict actions without future modeling. (b) Single-horizon methods predict a single future latent, overlooking complementary navigation roles for local continuity and route-level anticipation. (c) ForeFly jointly predicts proximal and route-critical futures, while FGAR integrates them for locally smooth and strategically consistent navigation.

A natural way to introduce such foresight is to predict future observations. However, directly forecasting future images in the pixel space is computationally demanding and will force the model to reconstruct visual details that are irrelevant to navigation, such as textures, illumination, or background appearance Chen et al. (2025); Emami et al. (2025); Zhao et al. (2026). Meanwhile, existing single-horizon prediction typically focuses on the immediate future, capturing short-term visual continuity but overlooking complementary cues at different temporal horizons that are critical for long-horizon aerial decision-making, such as upcoming turns, landmark transitions, and target-approaching regions Jia et al. (2026); Zhang et al. (2026e); Zheng et al. (2026a). Motivated by this observation, we model navigation-relevant futures at two complementary temporal horizons, predicting both proximal and route-critical future states in a compact latent space instead of reconstructing them in the pixel space. Specifically, proximal futures capture local motion continuity under continuous 3D motion and drastic viewpoint changes, while route-critical futures anticipate sparse navigation events to support long-range guidance over large-scale, weakly constrained aerial environments.

To this end, we propose **ForeFly** (Figure 1(c)), a retrospective dual-horizon world action model for AVLN. ForeFly first grounds horizon-specific foresight in recent and route-critical visual memories, then predicts complementary proximal and adaptive route-critical future latents to capture short-term motion continuity and long-range route anticipation, finally refines waypoint generation through Foresight-Guided Action Refinement (FGAR), which assigns asymmetric decision roles to dual-horizon foresight, leveraging proximal foresight for local action enhancement and route-critical foresight for strategic route-level correction. Our contributions are threefold:

- We reformulate AVLN from reactive action prediction into dual-horizon foresight that jointly models complementary proximal and route-critical latent futures, supporting both local motion continuity and long-range route guidance.
- We instantiate dual-horizon foresight with horizon-specific foresight query priming and an asymmetric Foresight-Guided Action Refinement (FGAR) module, which leverages complementary historical cues and translate proximal and route-critical foresight into local action enhancement and route-level correction, respectively.
- Experiments on TravelUAV and UAV-ON show consistent gains over strong baselines, particularly on long-range and unseen trajectories, while ablations validate the effectiveness and complementarity of dual-horizon foresight.

2 Related Work

Aerial Vision-Language Navigation. Vision-Language Navigation (VLN) was initially studied in ground-based environments, where embodied agents follow natural-language instructions through indoor or street-level scenes Anderson et al. (2017); Ku et al. (2020); An et al. (2025); Zhu et al. (2020); Wang et al. (2021). Compared to ground-based VLN, AVLN presents unique challenges, including a continuous 3D motion space, drastic viewpoint transitions, and heavy reliance on long-range landmark dependencies Xu et al. (2026a); Chen et al. (2026). To address these challenges, early AVLN methods improve visual-language grounding through auxiliary objectives, such as

spatial-relation matching and contrastive learning Su et al. (2025); Chu et al. (2024). More recent approaches increasingly adopt Multimodal Large Language Models (MLLMs) for navigation reasoning. For example, TravelUAV Wang et al. (2025) and FlightVLA Zheng et al. (2026b) employ hierarchical architectures that combine high-level language planning with low-level action refinement. Other methods incorporate depth or implicit 3D cues to strengthen spatial reasoning Sun et al. (2026b); Zhu et al. (2026); Jiang et al. (2026); Sun et al. (2026a); Liu et al. (2026a), while several studies exploit historical spatial-semantic evidence to improve temporal consistency over long trajectories Kim et al. (2025); Lou et al. (2026); Qi et al. (2026); Ding et al. (2026). Despite this progress, most existing AVLN methods remain predominantly reactive, predicting immediate actions from current and historical observations without explicitly modeling future world states. Consequently, they lack an explicit mechanism to anticipate navigation-critical events and use such foresight for proactive decision-making.

World Action Model for VLN. World models have recently emerged as a promising paradigm for VLN, learning environment dynamics from past observations and actions to predict future states for navigation Zhang et al. (2026e); Koh et al. (2021); Bar et al. (2025); Yao et al. (2025). Building on this idea, World Action Models (WAMs) unify future-state modeling and action generation within a single framework, and have recently been extended to VLN Ye et al. (2026); Cen et al. (2025); Xu et al. (2026b). Some approaches jointly optimize future-state prediction and action generation, explicitly coupling navigation decisions with their anticipated perceptual consequences Yang et al. (2026); Zhang et al. (2026c). Others employ video diffusion models to generate instruction-conditioned future observations and transfer their spatiotemporal priors to action or trajectory prediction Liu et al. (2026b); Xu et al. (2025). Recent methods further incorporate action-aware reinforcement learning to align navigation policies with long-term rollout outcomes Zhang et al. (2026d); Zhao et al. (2026). Despite this progress, existing methods largely emphasize short-horizon or single-step prediction, with limited mechanisms for anticipating sparse but navigation-critical future events. As a result, they may fail to foresee upcoming turns, landmark transitions, or other route-level cues, limiting proactive decision-making over long trajectories.

3 METHOD

3.1 PROBLEM FORMULATION

We formalize aerial vision-language navigation as a language-guided sequential decision-making problem, where a UAV navigates in a continuous 3D environment according to a natural-language instruction. Given a language instruction $\mathcal{I}$, the UAV maintains a trajectory history $\mathcal{P}_{\leq t} = \{p_1, p_2, \ldots, p_t\}$ and a sequence of historical observations $\mathcal{O}_{\leq t} = \{o_1, o_2, \ldots, o_t\}$. At each time step t, given the navigation context $\mathcal{C}_t = (\mathcal{I}, \mathcal{P}_{\leq t}, \mathcal{O}_{\leq t})$, the proposed world action model predicts dual-horizon future representations and conditions the next waypoint action on this foresight:

$$p_\theta\left(Z^p_{t+1}, Z^c_{t+\kappa_t}, a_t \mid C_t\right) = p^W_\theta\left(Z^p_{t+1}, Z^c_{t+\kappa_t} \mid C_t\right) p^A_\theta\left(a_t \mid C_t, Z^p_{t+1}, Z^c_{t+\kappa_t}\right), \tag{1}$$

where $\mathbf{Z}^p_{t+1} \in \mathbb{R}^{N_p \times d}$ denotes the proximal future latent and $\mathbf{Z}^c_{t+\kappa_t} \in \mathbb{R}^{N_c \times d}$ denotes the route-critical future latent. Here, κ_t is not a fixed constant, but an adaptive temporal offset determined by the route-critical target construction described in Sec. A.1. The waypoint action is represented in the current UAV coordinate frame as:

$$\mathbf{a}_t = (\mathbf{d}_t, \Delta\rho_t) = (\Delta x_t, \Delta y_t, \Delta z_t, \Delta\rho_t) \in \mathbb{R}^4. \tag{2}$$

where $\mathbf{d}_t = (\Delta x_t, \Delta y_t, \Delta z_t)$ denotes the normalized relative waypoint direction in 3D space with $\|\mathbf{d}_t\|_2 = 1$, and $\Delta\rho_t \geq 0$ denotes the movement magnitude.

3.2 OVERVIEW

As illustrated in Fig. 2, ForeFly explicitly introduces future-world reasoning into aerial vision-language navigation instead of directly mapping the current multimodal context to an immediate waypoint. Given the instruction $\mathcal{I}$, trajectory history $\mathcal{P} \leq t$, and observation history $\mathcal{O} \leq t$, ForeFly leverages historical context to construct future-oriented representations and conditions waypoint prediction on the resulting dual-horizon foresight.

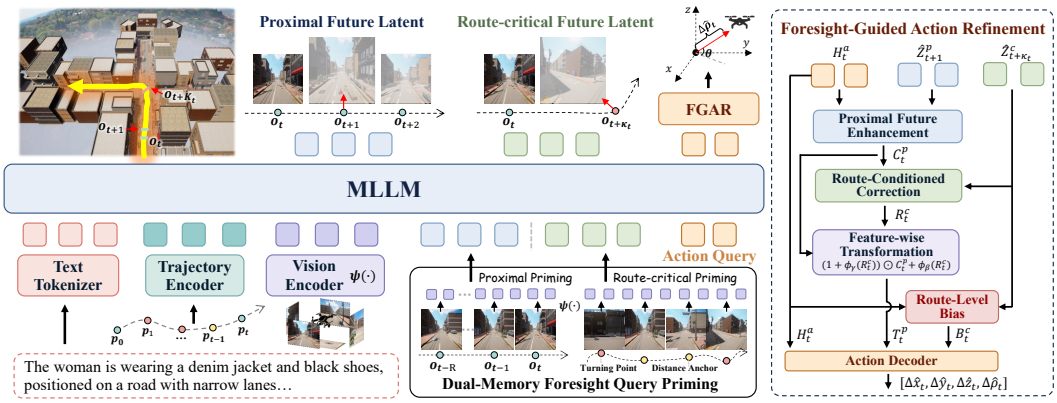

Figure 2: **Overview of ForeFly.** ForeFly primes dual-horizon foresight with complementary historical memories, predicts proximal and route-critical future latents, Foresight-Guided Action Refinement (FGAR) asymmetrically uses proximal foresight for local enhancement and route-critical foresight for feature-wise correction and route-level guidance.

Specifically, ForeFly first employs *Dual-Memory Foresight Query Priming* to construct recent and route-critical visual memories, providing complementary historical cues for history-aware foresight queries. Given the primed queries, *Dual-Horizon Future Latent Prediction* jointly infers a proximal future latent and an adaptive route-critical future latent under the multimodal context. Finally, *Foresight-Guided Action Refinement* assigns distinct roles to the dual-horizon foresight, using proximal prediction for local action enhancement and route-critical prediction for route-level correction.

3.3 Dual-Memory Foresight Query Priming

Learnable foresight queries lack explicit route-dependent historical grounding. We therefore introduce *Dual-Memory Foresight Query Priming*, which conditions horizon-specific foresight queries on complementary recent and route-critical visual memories before future latent prediction.

Recent visual memory. At time step t, each historical observation o_i, $i \le t$, is encoded as $\mathbf{x}_i = \psi(o_i)$, where $\psi(\cdot)$ denotes the visual encoding and projection pipeline, yielding visual embeddings $\mathbf{x}_i$. The recent visual memory $\mathcal{M}_t^p$ retains features from the latest R observations, preserving short-term appearance and motion continuity around the current state.

Route-critical visual memory. While recent observations characterize local visual evolution, long-horizon foresight benefits from sparse trajectory states associated with meaningful route transitions. In particular, turning events often coincide with key navigation decision points and capture structural changes in the route Han et al. (2025); Baghaei et al. (2025). Let $\mathbf{p}_i \in \mathbb{R}^3$ denote the UAV position at time step i. We define the normalized flight direction and the turning angle between consecutive directions as:

$$\mathbf{v}_i = \frac{\mathbf{p}_i - \mathbf{p}_{i-1}}{\|\mathbf{p}_i - \mathbf{p}_{i-1}\|_2} \in \mathbb{S}^2, \qquad \alpha_i = \cos^{-1}\left(\mathbf{v}_i^\top \mathbf{v}_{i+1}\right). \tag{3}$$

Frames satisfying $\alpha_i \ge \tau_p$ are selected as turning anchors, where τ_p is the turning-angle threshold.

Turning events alone can be overly sparse along extended straight segments. We therefore introduce distance-based coverage anchors at approximately every τ_g meters of accumulated travel. For two adjacent event anchors u and v, let $s(u,j) = \sum_{k=u+1}^{j} \|\mathbf{p}_k - \mathbf{p}_{k-1}\|_2$ denote the traveled distance from u to j. The m-th coverage anchor is selected as:

$$j_m^\star = \arg\min_{j:u<j<v} \left| s(u,j) - m\tau_g \right|, \qquad 0 < m\tau_g < s(u,v). \tag{4}$$

The route-critical memory $\mathcal{M}_t^c$ consists of visual features from the initial frame, detected turning events, and distance-based coverage anchors. This event-aware construction emphasizes route transitions while maintaining historical coverage over event-sparse trajectory segments.

Horizon-specific foresight query priming. To exploit the complementary historical cues encoded in the two memories, we introduce learnable proximal and route-critical foresight queries, $\mathcal{Q}^p \in$

$\mathbb{R}^{N_p \times d}$ and $\mathcal{Q}^c \in \mathbb{R}^{N_c \times d}$, for different prediction horizons. The proximal and route-critical queries attend to their corresponding historical memories, $\mathcal{M}_t^p$ and $\mathcal{M}_t^c$, yielding the history-aware queries $\widetilde{\mathcal{Q}}_t^p$ and $\widetilde{\mathcal{Q}}_t^c$, respectively. The resulting history-aware queries capture recent visual continuity and route-level historical cues, respectively, and are used for dual-horizon future latent prediction.

3.4 Dual-Horizon Future Latent Prediction

Given the primed foresight queries, we perform dual-horizon future prediction and action reasoning within a unified multimodal decoding process. At time step t, the instruction $\mathcal{I}$, trajectory history $\mathcal{P}_{\leq t}$, and current observation o_t are encoded as $\boldsymbol{\ell}$, $\mathbf{e}_{\leq t}^{\text{traj}}$, and $\mathbf{x}_t$, respectively. The MLLM jointly processes these features with the primed proximal and route-critical queries $\widetilde{\mathcal{Q}}_t^p$ and $\widetilde{\mathcal{Q}}_t^c$, together with the learnable action queries $\mathcal{Q}^a$. Here, the primed foresight queries inject complementary historical cues from the recent and route-critical memories, while the trajectory encoding provides explicit spatial context. Together, these inputs form a unified multimodal context for decoding proximal foresight, route-critical foresight, and action representations:

$$\left(\mathbf{H}_{t+1}^p, \mathbf{H}_{t+\kappa_t}^c, \mathbf{H}_t^a\right) = f_\theta\left(\boldsymbol{\ell}, \mathbf{e}_{\leq t}^{\text{traj}}, \mathbf{x}_t\,;\, \widetilde{\mathcal{Q}}_t^p, \widetilde{\mathcal{Q}}_t^c, \mathcal{Q}^a\right), \tag{5}$$

where $\mathbf{H}_{t+1}^p$, $\mathbf{H}_{t+\kappa_t}^c$, and $\mathbf{H}_t^a$ denote the decoded representations for proximal future modeling, route-critical future modeling, and waypoint prediction, respectively.

During this decoding process, we avoid reconstructing future RGB observations and instead supervise the decoded foresight representations in a compact visual embedding space. Specifically, a shared projection head $g(\cdot)$ maps the proximal and route-critical representations to $\hat{\mathbf{Z}}_{t+1}^p = g(\mathbf{H}_{t+1}^p)$ and $\hat{\mathbf{Z}}_{t+\kappa_t}^c = g(\mathbf{H}_{t+\kappa_t}^c)$, respectively. These auxiliary predictions are supervised by the visual representations of the corresponding future observations, encouraging the decoded foresight representations to capture navigation-relevant semantic layout, scene transitions, and spatial structure without modeling unnecessary pixel-level details. Sharing the projection head encourages the two branches to produce future representations in a consistent latent space, while their distinct foresight queries preserve horizon-specific temporal information.

The action-query representation $\mathbf{H}_t^a$ summarizes action-oriented evidence from the current multimodal context. However, using it directly for waypoint regression may still produce a predominantly reactive policy. We therefore introduce a Foresight-Guided Action Refinement (FGAR) module that explicitly refines the action representation using the proximal and route-critical foresight representations, providing complementary local and route-level guidance: $\hat{\mathbf{a}}_t = \Phi_{\text{FGAR}}\left(\mathbf{H}_t^a, \hat{\mathbf{Z}}_{t+1}^p, \hat{\mathbf{Z}}_{t+\kappa_t}^c\right)$, where $\Phi_{\text{FGAR}}(\cdot)$ denotes the Foresight-Guided Action Refinement operation.

Training Objective. To jointly align future-world imagination with waypoint prediction, we supervise both the dual-horizon latent predictions and the predicted action. The waypoint action is decomposed into a 3D direction $\mathbf{d}_t$ and a movement magnitude $\Delta\rho_t$. The joint training objective is:

$$\begin{aligned}
\mathcal{L}_W &= \mathbb{E}_{(\mathcal{C}_t, \{o_{t+\delta_h}\}_{h\in\{p,c\}})\sim\mathcal{D}}\left[\sum_{h\in\{p,c\}} \frac{1}{N_h d}\left\|\hat{\mathbf{Z}}_{t+\delta_h}^h - \psi(o_{t+\delta_h})\right\|_F^2\right], \\
\mathcal{L}_A &= \mathbb{E}_{(\mathcal{C}_t, \mathbf{a}_t)\sim\mathcal{D}}\left[1 - \hat{\mathbf{d}}_t^\top \mathbf{d}_t + |\Delta\hat{\rho}_t - \Delta\rho_t|\right].
\end{aligned} \tag{6}$$

where $\{\hat{\mathbf{Z}}_{t+\delta_h}^h\}_{h\in\{p,c\}}$ denote the predicted dual-horizon future representations at horizon $h \in \{p, c\}$, with $\delta_p = 1$ and $\delta_c = \kappa_t$ denoting the proximal and route-critical prediction horizons, respectively. The overall training objective is given by $\mathcal{L} = \mathcal{L}_W + \mathcal{L}_A$.

3.5 Foresight-Guided Action Refinement

Within the unified multimodal decoding process, the action representation is jointly contextualized with the proximal and route-critical foresight representations. However, this generic contextualization does not explicitly distinguish their complementary roles in action generation. We therefore introduce a *Foresight-Guided Action Refinement* (FGAR) module that refines the action representation with dual-horizon foresight, using proximal prediction to strengthen local action evidence and route-critical prediction to further impose route-level guidance before waypoint decoding.

Proximal future enhancement. The proximal future latent describes the immediate visual evolution from the current observation. We first use the action representation as the query and the predicted proximal latent as the key and value $\mathbf{C}_t^p = \mathrm{Attn}_p(\mathbf{H}_t^a, \hat{\mathbf{Z}}_{t+1}^p)$, where $\mathbf{C}_t^p$ denotes the proximal-enhanced action representation. This interaction grounds the waypoint decision in anticipated short-term changes in viewpoint, local geometry, and scene layout.

Asymmetric route-critical conditioning. A straightforward fusion of the proximal and route-critical future latents would treat the two horizons as equally relevant to action prediction. However, they play fundamentally different roles in aerial navigation. The proximal future describes locally executable scene evolution, whereas the route-critical future provides sparse but strategically important information about upcoming turns, landmark transitions, and route progress. We therefore adopt an asymmetric conditioning strategy: the proximal-enhanced action representation serves as the execution carrier, while the route-critical future acts as a high-level controller that selectively reshapes this local action evidence.

Specifically, we first derive a route-conditioned correction context by allowing the proximal-enhanced action representation to attend to the route-critical future latent $\mathbf{R}_t^c = \mathrm{Attn}_c(\mathbf{C}_t^p, \hat{\mathbf{Z}}_{t+\kappa_t}^c)$. Rather than directly decoding $\mathbf{R}_t^c$ into an action or replacing the locally grounded representation $\mathbf{C}_t^p$, we use it to generate a feature-wise transformation:

$$[\boldsymbol{\gamma}_t, \boldsymbol{\beta}_t] = \phi_{\mathrm{film}}\left(\mathbf{R}_t^c\right), \qquad \mathbf{T}_t^p = (1 + \boldsymbol{\gamma}_t) \odot \mathbf{C}_t^p + \boldsymbol{\beta}_t. \tag{7}$$

Here, $\mathbf{T}_t^p$ denotes the route-corrected proximal action representation. The scaling factor $\boldsymbol{\gamma}_t$ controls which dimensions of the local action evidence should be amplified or suppressed, while $\boldsymbol{\beta}_t$ introduces event-dependent feature shifts. Consequently, the route-critical future modifies the interpretation of the proximal future without discarding its locally executable geometric information.

Since feature-wise modulation may overlook route-level cues, we introduce a parallel route-level guidance path. The original action representation directly queries the route-critical future latent:

$$\mathbf{B}_t^c = \phi_{\mathrm{bias}}\left(\mathrm{Attn}_b\left(\mathrm{sg}(\mathbf{H}_t^a), \mathrm{sg}\left(\hat{\mathbf{Z}}_{t+\kappa_t}^c\right)\right)\right), \tag{8}$$

where $\mathrm{sg}(\cdot)$ denotes the stop-gradient operation and $\phi_{\mathrm{bias}}(\cdot)$ denotes the learnable projection. Unlike the modulation path, this branch produces an additive route-level decision $\mathbf{B}_t^c$. Detaching its inputs makes the branch a read-only consumer, preventing the additive path from becoming an optimization shortcut. Meanwhile, the main modulation path remains differentiable, allowing the route-critical representation to be jointly optimized for future prediction and downstream action utility.

The final action integrates the original action evidence, route-corrected proximal features, and the detached route-level guidance residual:

$$\hat{\mathbf{a}}_t = \phi_{\mathrm{act}}\left(\mathrm{LN}\left(\mathbf{H}_t^a + \mathbf{T}_t^p + \mathbf{B}_t^c\right)\right). \tag{9}$$

The original representation $\mathbf{H}_t^a$ acts as a residual basis, $\mathbf{T}_t^p$ preserves local executability while incorporating event-conditioned feature correction, and $\mathbf{B}_t^c$ provides additive route-level guidance. The refined representation is then decoded into the waypoint action through the MLP projection $\phi_{\mathrm{act}}(\cdot)$.

4 Experiment

4.1 Experimental Setup

Datasets. We conduct experiments on two widely used AVLN benchmarks, TravelUAV Wang et al. (2025) and UAV-ON Xiao et al. (2025). TravelUAV contains 12,149 trajectories across 22 diverse scenarios, with target distances ranging from 50 m to 400 m. UAV-ON further provides 14 high-fidelity outdoor environments with diverse geographic structures, ranging from compact urban parks to expansive mountainous terrains. Together, these two benchmarks enable a comprehensive evaluation of navigation robustness and generalization across diverse aerial environments.

Evaluation Metrics. Navigation performance is evaluated using four standard VLN 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, SR reports the success rate of reaching the goal, and OSR checks whether the goal region is reached at any point along the trajectory. SPL

Table 1: Results on the TravelUAV Unseen dataset. NE is reported in meters (m), while SR, OSR, and SPL are reported as percentages (%). For task difficulty, Easy denotes trajectories shorter than 250 meters, whereas Hard denotes those longer than 250 meters. The best results are bold.

Method	Full				Easy				Hard			
	NE↓	SR↑	OSR↑	SPL↑	NE↓	SR↑	OSR↑	SPL↑	NE↓	SR↑	OSR↑	SPL↑
Random Action	225.64	0.06	0.06	0.06	164.66	0.19	0.19	0.19	280.58	0.00	0.00	0.00
Fixed Action	193.30	1.76	5.36	1.09	140.33	3.19	8.08	1.88	245.96	0.85	3.08	0.55
CMA Anderson et al. (2017)	147.27	4.98	12.41	4.74	102.54	8.03	17.52	7.52	191.30	2.76	7.53	2.71
TravelUAV Wang et al. (2025)	130.60	11.41	31.13	10.45	96.27	12.47	33.31	11.29	167.49	10.62	28.91	9.80
NavFoM Zhang et al. (2026b)	118.34	15.63	30.46	14.21	89.77	16.98	32.22	15.35	155.69	14.35	27.79	13.16
LongFly Jiang et al. (2025)	91.84	24.19	43.86	20.84	69.16	22.89	43.24	18.66	112.02	25.36	44.41	22.76
SpatialFly Jiang et al. (2026)	87.82	25.46	44.41	21.76	64.49	25.17	44.18	20.36	108.56	25.71	44.63	23.01
ForeFly (Ours)	**79.80**	**28.86**	**52.87**	**25.19**	**62.07**	**26.49**	**53.25**	**22.70**	**96.51**	**31.09**	**52.51**	**27.54**

Table 2: Results on the TravelUAV Seen dataset. *Human* represents the performance of human operators under manual UAV control. The best results are bold.

Method	**Full**				**Easy**				**Hard**			
	NE↓	SR↑	OSR↑	SPL↑	NE↓	SR↑	OSR↑	SPL↑	NE↓	SR↑	OSR↑	SPL↑
Human	**14.15**	**94.51**	**94.51**	**77.84**	**11.68**	**95.44**	**95.44**	**76.19**	**17.16**	**93.37**	**93.37**	**79.85**
Random Action	222.20	0.14	0.21	0.07	142.07	0.26	0.39	0.13	320.12	0.00	0.00	0.00
Fixed Action	188.61	2.27	8.16	1.40	121.36	3.48	11.48	2.14	270.69	0.79	4.09	0.49
CMA Anderson et al. (2017)	135.73	8.37	18.72	7.90	84.89	11.48	24.52	10.68	197.77	4.57	11.65	4.51
TravelUAV Wang et al. (2025)	106.28	16.10	44.26	14.30	68.78	18.84	47.61	16.39	152.04	12.76	40.16	11.76
TravelUAV-DA Wang et al. (2025)	98.66	17.45	48.87	15.76	66.40	20.26	51.23	18.10	138.04	14.02	45.98	12.90
SimpleNav SimpleNav Contributors (2026)	85.60	22.40	55.10	20.50	60.00	22.80	56.90	21.00	118.30	22.00	52.80	19.90
NavFoM Zhang et al. (2026b)	93.05	29.17	49.24	25.03	58.98	32.91	53.16	27.87	143.83	23.58	43.40	20.80
AirForsight Liu et al. (2026c)	56.99	35.83	69.25	30.22	38.27	38.83	71.26	31.65	80.08	32.13	66.77	28.45
LongFly Jiang et al. (2025)	60.02	36.39	65.87	31.07	38.10	38.52	71.90	31.24	85.20	33.94	58.94	30.88
SpatialFly Jiang et al. (2026)	58.05	38.54	68.62	33.43	**34.41**	39.59	**74.14**	33.61	84.76	37.33	62.27	33.22
ForeFly (Ours)	**48.74**	**41.75**	**70.25**	**34.75**	34.49	**44.55**	71.36	**35.72**	**66.16**	**38.33**	**68.89**	**33.55**

additionally evaluates path efficiency by considering the ratio between the shortest and executed path lengths, providing a balanced assessment of navigation success and efficiency.

Implementation Details. All experiments are conducted using PyTorch on four NVIDIA RTX 5090 GPUs. We adopt LLaMA 2-7B as the language backbone and EVA-ViT-G as the visual encoder. The numbers of proximal and route-critical foresight slots are both set to $N_p = N_c = 17$. The turning-angle threshold τ_p and distance interval τ_g are set to 20° and 30 m, respectively. We fine-tune the model with LoRA using a learning rate of 5×10^{-5}. Training proceeds in two stages. In the first stage, we replace FGAR with a simple MLP head to stabilize future representation learning and waypoint prediction. In the second stage, FGAR is restored and all trainable components are jointly optimized, enabling future representations to effectively guide navigation decisions.

4.2 Quantitative Results

Results on TravelUAV Dataset. As shown in Tables 1 and 2, ForeFly consistently achieves strong performance on both the Unseen and Seen splits. On the Unseen split, ForeFly reduces NE by 8.02 m and improves SR, OSR, and SPL by 3.40, 8.46, and 3.43 percentage points over SpatialFly, respectively, indicating improved generalization to unseen environments. The advantage is particularly pronounced on the Hard subset, where SR increases from 25.71% to 31.09%, highlighting its effectiveness on long-range navigation tasks. On the Seen split, ForeFly achieves the best overall performance among autonomous methods. Compared with SpatialFly, ForeFly improves SR by 3.21 and 4.96 percentage points on the Full and Easy subsets, respectively. Although its OSR on the Easy subset is slightly lower than that of SpatialFly, ForeFly improves OSR by 6.62 percentage points on the Hard subset, indicating a clearer advantage on more challenging long-range trajectories.

Results on UAV-ON Dataset. As shown in Table 4, ForeFly achieves the best performance across all metrics on both the Seen and Unseen splits. On the Seen split, ForeFly reaches 25.13% SR, 36.30% OSR, and 16.31% SPL, outperforming the strongest competing results by 2.41, 5.82, and 7.18 percentage points, respectively. On the Unseen split, ForeFly improves SR from 17.57% to 22.68%, yielding a 5.11 percentage-point gain, while also achieving the highest OSR of 31.44% and SPL of 14.59%. These consistent improvements across both splits further demonstrate the robustness and generalization ability of ForeFly beyond the TravelUAV benchmark.

Table 3: Ablation study of the main components in ForeFly. Seen and Unseen represent the *Test-seen* and *Test-unseen sets*, respectively.

Variant	Proximal Foresight	Route-Critical Foresight	Foresight Query Priming	Foresight-Guided Action Refinement	Seen (Full)				Unseen (Full)			
					NE ↓	SR ↑	OSR ↑	SPL ↑	NE ↓	SR ↑	OSR ↑	SPL ↑
Baseline	×	×	×	×	103.75	21.73	42.41	18.04	130.60	11.41	31.13	10.45
Proximal Only	✓	×	×	×	72.16	23.00	54.50	19.46	84.50	19.25	44.25	16.91
Route-Critical Only	×	✓	×	×	69.99	28.25	57.25	24.19	95.00	17.50	42.00	15.37
Dual-Horizon	✓	✓	×	×	70.06	32.75	62.00	27.54	84.14	21.75	44.53	16.29
+ Foresight Query Priming	✓	✓	✓	×	64.41	33.08	62.91	27.99	81.82	23.61	45.36	19.73
ForeFly (Full)	✓	✓	✓	✓	**48.74**	**41.75**	**70.25**	**34.75**	**79.80**	**28.86**	**52.87**	**25.19**

Table 4: Results on the UAV-ON dataset.

Method	Seen			Unseen		
	SR↑	OSR↑	SPL↑	SR↑	OSR↑	SPL↑
AerialVLN Liu et al. (2023)	4.55	16.84	4.24	3.51	20.77	3.41
AOA-F Xiao et al. (2025)	7.87	17.98	4.21	6.45	16.09	3.98
OpenFly Gao et al. (2026)	12.29	26.20	6.83	11.82	25.71	5.64
APEX Zhang et al. (2026a)	12.36	19.10	9.13	16.13	22.58	13.03
Navid Zhang et al. (2024)	14.17	30.08	7.19	10.70	28.51	5.99
OctMem-Agent Zhou et al. (2026)	22.72	30.48	8.40	17.57	28.59	5.15
ForeFly (Ours)	**25.13**	**36.30**	**16.31**	**22.68**	**31.44**	**14.59**

Table 5: Ablation on Foresight-Guided Action Refinement (FGAR) Design.

Method	NE↓	SR↑	OSR↑	SPL↑
MLP Head	64.41	33.08	62.91	27.99
Direct Dual-Future*	62.03	34.00	62.50	28.73
FGAR *w/o* $\mathbf{B}_t^c$	54.61	38.35	68.73	32.55
FGAR	**48.74**	**41.75**	**70.25**	**34.75**

Note: Direct Dual-Future Fusion refers to concatenating the proximal and route-critical future latents and using them to condition the action representation via *cross-attention*.

Overall, the consistent gains across both TravelUAV and UAV-ON, particularly on unseen environments and challenging long-range trajectories, demonstrate the robustness and generalization of ForeFly for long-horizon AVLN. These improvements are consistent with the complementary dual-horizon foresight design, where proximal prediction preserves local motion continuity, while route-critical foresight provides strategic route-level anticipation. By integrating these future cues into waypoint generation through FGAR, ForeFly can reduce accumulated navigation errors and maintain better route consistency over extended trajectories.

4.3 Ablation Study

As shown in Table 3, we conduct a systematic ablation study to evaluate the contribution of each key component in ForeFly. We progressively examine the effects of proximal foresight, route-critical foresight, Dual-Memory Foresight Query Priming, and the Foresight-Guided Action Refinement on TravelUAV dataset, isolating their individual and complementary contributions.

Effect of Dual-Horizon Foresight. We first examine the contribution of the two foresight horizons. Both Proximal Only and Route-Critical Only consistently outperform the baseline, confirming the benefit of explicitly modeling future states. Interestingly, Route-Critical Only yields larger gains on the Seen split, whereas Proximal Only performs better on the Unseen split. Combining the two horizons further improves SR to 32.75% and 21.75% on the Seen and Unseen splits, respectively. These results suggest that proximal and route-critical foresight provide complementary information at different navigation horizons rather than redundant predictions.

Effect of Foresight Query Priming. We next examine whether grounding foresight queries in historical observations improves future-state modeling. Introducing retrospective priming consistently benefits both splits, with larger gains on the Unseen split, where SR increases from 21.75% to 23.61% and SPL from 16.29% to 19.73%. These results suggest that route-dependent historical cues enrich future-state prediction and improve foresight reliability in unseen environments.

Effect of Foresight-Guided Action Refinement. Finally, we evaluate the effectiveness of incorporating predicted future representations into navigation decisions. Introducing FGAR substantially improves the retrospectively primed dual-horizon model, reducing NE from 64.41 m to 48.74 m and increasing SR from 33.08% to 41.75% on the Seen split. A similar improvement is observed on the Unseen split, where SR increases from 23.61% to 28.86%. The consistent gains across both splits confirm the effectiveness of future representations for action refinement. These results show that future-state prediction alone is insufficient, effectively conditioning action generation on the predicted futures is crucial for translating foresight into reliable navigation decisions.

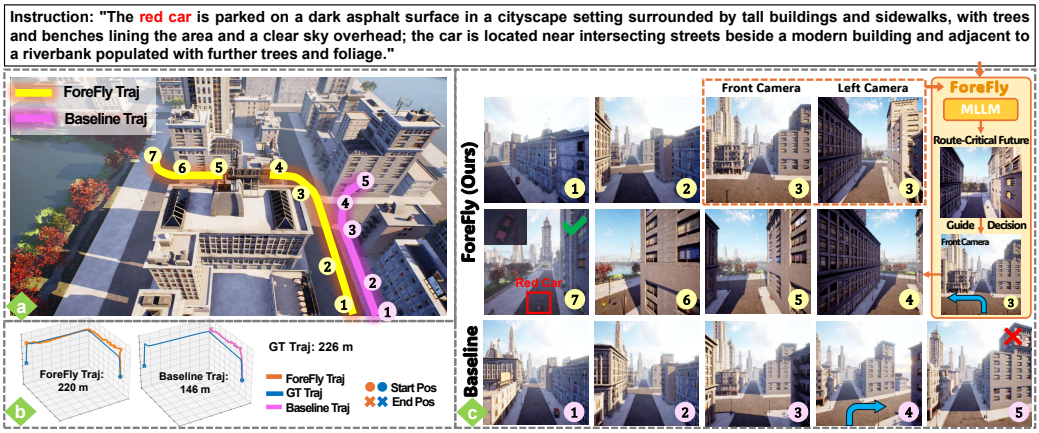

Figure 3: **Qualitative Results.** (a) Bird's-eye view of predicted trajectories. (b) 3D trajectory comparison showing improved route consistency of ForeFly. (c) Step-by-step navigation, where route-critical foresight helps ForeFly anticipate the upcoming turn, while the baseline deviates. The corresponding future visualization procedure is described in Sec. B.2.

Ablation on Foresight-Guided Action Refinement Design. We further compare different strategies for incorporating dual-horizon foresight into action refinement. As shown in Table 5, directly fusing the two future latents yields only modest improvements over the MLP head, suggesting that simply conditioning the action representation on future information is insufficient. In contrast, the asymmetric conditioning mechanism substantially improves performance, and the additional route-level guidance branch further strengthens all metrics. This validates the design choice of assigning different roles to proximal and route-critical foresight during action refinement.

4.4 Qualitative Analysis

Figure 3 presents a representative example comparing ForeFly with the baseline. As illustrated in Fig. 3(a), the UAV is instructed to locate a red car near a riverbank after traversing the urban blocks. As shown in Fig. 3(c), the baseline makes an incorrect decision at the critical turning point and consequently deviates from the intended route. In contrast, ForeFly internally anticipates a route-critical future latent associated with the upcoming route transition, providing strategic guidance for the waypoint decision and enabling the UAV to make the correct turn and ultimately reach the target.

The trajectory comparisons in Fig. 3(a,b) further illustrate this advantage. ForeFly generates a 220 m trajectory that closely follows the 226 m ground-truth route, while the baseline exhibits an early route deviation and fails to complete the navigation. These results highlight the complementary roles of the two foresight horizons: proximal foresight supports locally consistent waypoint generation, whereas route-critical foresight provides strategic guidance around important landmarks and turning regions. Together, they enable ForeFly to maintain both local motion continuity and long-horizon route consistency in complex aerial environments.

5 Conclusion

In this work, we introduced ForeFly, a dual-horizon latent world action model for aerial vision-language navigation. Rather than relying solely on reactive action prediction or single-horizon foresight, ForeFly explicitly anticipates both proximal future states for local motion continuity and route-critical future states for long-range guidance. A foresight query priming mechanism grounds these predictions in complementary historical cues, while the proposed Foresight-Guided Action Refinement (FGAR) asymmetrically integrates the two foresight horizons to preserve locally executable decisions while enabling strategic route-level correction. Experiments on the TravelUAV and UAV-ON benchmarks demonstrate consistent improvements over strong baselines on both seen and unseen environments, with particularly clear gains on challenging long-horizon trajectories. Ablation and qualitative analyses further validate the complementary roles of dual-horizon foresight, foresight query priming, and foresight-guided action refinement.

AI USE STATEMENT

Generative AI tools were used solely for language polishing of this manuscript. The authors reviewed all edits and take full responsibility for the final content of this work.

REFERENCES

Dong An, Hanqing Wang, Wenguan Wang, Zun Wang, Yan Huang, Keji He, and Liang Wang. Etpnav: Evolving topological planning for vision-language navigation in continuous environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(7):5130–5145, 2025. doi: 10.1109/TPAMI.2024.3386695.

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. *arXiv preprint arXiv:1711.07280*, 2017.

Kourosh T Baghaei, Dieter Pfoser, and Antonios Anastasopoulos. Follow the beaten path: The role of route patterns on vision-language navigation agents generalization abilities. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 7986–8005, April 2025. doi: 10.18653/v1/2025.naacl-long.406. URL https://aclanthology.org/2025.naacl-long.406/.

Amir Bar, Gaoyue Zhou, Danny Tran, Trevor Darrell, and Yann LeCun. Navigation world models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15791–15801, 2025. doi: 10.1109/CVPR52734.2025.01472.

Valts Blukis, Dipendra Misra, Ross A. Knepper, and Yoav Artzi. Mapping navigation instructions to continuous control actions with position-visitation prediction. In *Proceedings of the Conference on Robot Learning (CoRL)*, 2018.

Jun Cen, Chaohui Yu, Hangjie Yuan, Yuming Jiang, Siteng Huang, Jiayan Guo, Xin Li, Yibing Song, Hao Luo, Fan Wang, et al. Worldvla: Towards autoregressive action world model. *arXiv preprint arXiv:2506.21539*, 2025.

Boyuan Chen, Tianyuan Zhang, Haoran Geng, Kiwhan Song, William T. Freeman, Jitendra Malik, Russ Tedrake, Vincent Sitzmann, and Yilun Du. Large video planner enables generalizable robot control, 2025. URL http://arxiv.org/abs/2512.15840.

Hanxuan Chen, Jie Zheng, Siqi Yang, Tianle Zeng, Siwei Feng, Songsheng Cheng, Ruilong Ren, Hanzhong Guo, Shuai Yuan, Xiangyue Wang, Kangli Wang, and Ji Pei. Vision-and-language navigation for uavs: Progress, challenges, and a research roadmap, 2026. URL https://arxiv.org/abs/2604.13654.

Howard Chen, Alane Suhr, Dipendra Misra, Noah Snavely, and Yoav Artzi. Touchdown: Natural language navigation and spatial reasoning in visual street environments. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12530–12539, 2019. doi: 10.1109/CVPR.2019.01282.

Meng Chu, Zhedong Zheng, Wei Ji, Tingyu Wang, and Tat-Seng Chua. Towards natural language-guided drones: Geotext-1652 benchmark with spatial relation matching. In *European conference on computer vision (ECCV)*, 2024. URL https://zdzheng.xyz/files/2024/ECCV24-GeoText.pdf.

Ming Dai, Enhui Zheng, Wenxuan Cheng, Jiahao Chen, Zhenhua Feng, and Wankou Yang. Drl: An efficient heterogeneous spatial feature interaction framework for uav self-localization. *Pattern Recognit.*, 177:113330, 2026. URL https://api.semanticscholar.org/CorpusID:285851517.

X. Ding, J. Gao, C. Pan, W. Wang, and J. Qin. History-enhanced two-stage transformer for aerial vision-and-language navigation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 18225–18233, 2026. doi: 10.1609/aaai.v40i22.38885.

Yousef Emami, Hao Zhou, Luis Almeida, and Kai Li. Diffusion models for smarter uavs: Decision-making and modeling. *arXiv preprint arXiv:2501.05819*, 2025. URL https://arxiv.org/abs/2501.05819.

Yue Fan, Winson Chen, Tongzhou Jiang, Chun Zhou, Yi Zhang, and Xin Wang. Aerial vision-and-dialog navigation. *arXiv preprint arXiv:2205.12219*, 2022. doi: 10.48550/arXiv.2205.12219.

Yunpeng Gao, Chenhui Li, Zhongrui You, Junli Liu, Li Zhen, Pengan CHEN, Qizhi Chen, Zhonghan Tang, Liansheng Wang, Yangpenghui, Yiwen Tang, Yuhang Tang, Shuai Liang, Songyi Zhu, Ziqin Xiong, Yifei Su, Xinyi Ye, Jianan Li, Yan Ding, Dong Wang, Zhigang Wang, Bin Zhao, and Xuelong Li. Openfly: A COMPREHENSIVE PLATFORM FOR AERIAL VISION-LANGUAGE NAVIGATION. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=OKm3w71ymP.

Yu Gu, Weilin Chen, and Dongliang Peng. Uav-based multimodal object detection via feature enhancement and dynamic gated fusion. *Pattern Recognit.*, 172:112722, 2026. URL https://api.semanticscholar.org/CorpusID:283178691.

Mingfei Han, Liang Ma, Kamila Zhumakhanova, Ekaterina Radionova, Jingyi Zhang, Xiaojun Chang, Xiaodan Liang, and Ivan Laptev. Roomtour3d: Geometry-aware video-instruction tuning for embodied navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27586–27596, June 2025.

Feiyang Jia, Lin Liu, Ziying Song, Caiyan Jia, Hangjun Ye, Xiaoshuai Hao, and Long Chen. Driveworld-vla: Unified latent-space world modeling with vision-language-action for autonomous driving, 2026. URL https://arxiv.org/abs/2602.06521.

Wen Jiang, Li Wang, Kangyao Huang, Wei Fan, Jinyuan Liu, Shaoyu Liu, Hongwei Duan, Bin Xu, and Xiangyang Ji. Longfly: Long-horizon uav vision-and-language navigation with spatiotemporal context integration, 2025. URL https://arxiv.org/abs/2512.22010.

Wen Jiang, Kangyao Huang, Li Wang, Wang Xu, Wei Fan, Jinyuan Liu, Shaoyu Liu, Hanfang Liang, Hongwei Duan, Bin Xu, and Xiangyang Ji. Spatialfly: Geometry-guided representation alignment for uav vision-and-language navigation in urban environments, 2026. URL https://arxiv.org/abs/2603.21046.

Seungchan Kim, Omar Alama, Dmytro Kurdydyk, John Keller, Nikhil Keetha, Wenshan Wang, Yonatan Bisk, and Sebastian Scherer. Raven: Resilient aerial navigation via open-set semantic memory and behavior adaptation. *arXiv preprint arXiv:2509.23563*, 2025.

Jing Yu Koh, Honglak Lee, Yinfei Yang, Jason Baldridge, and Peter Anderson. Pathdreamer: A world model for indoor navigation. 2021.

Jacob Krantz, Erik Wijmans, Arjun Majumdar, Dhruv Batra, and Stefan Lee. Beyond the nav-graph: Vision-and-language navigation in continuous environments. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (eds.), *Computer Vision – ECCV 2020*, volume 12373 of *Lecture Notes in Computer Science*. Springer, Cham, 2020. doi: 10.1007/978-3-030-58604-1_7.

Alexander Ku, Peter Anderson, Roma Patel, Eugene Ie, and Jason Baldridge. Room-Across-Room: Multilingual vision-and-language navigation with dense spatiotemporal grounding. In *Conference on Empirical Methods for Natural Language Processing (EMNLP)*, 2020.

Jungdae Lee, Taiki Miyanishi, Shuhei Kurita, Koya Sakamoto, Daichi Azuma, Yutaka Matsuo, and Nakamasa Inoue. Citynav: Language-goal aerial navigation dataset with geographic information, 2024.

Peican Lin, Gan Sun, Chenxi Liu, Fazeng Li, Weihong Ren, and Yang Cong. Openvln: Open-world aerial vision-language navigation, 2025. URL https://arxiv.org/abs/2511.06182.

Like Liu, Zheng Xu, Haitao He, Hongzhe Li, Shuchang Zhang, and Dian Shao. Knowing the self, understanding the world: A dual-cognition benchmark for uav spatio-temporal reasoning with mllms. 2026a. URL https://api.semanticscholar.org/CorpusID:290333328.

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 (ICCV)*, pp. 15384–15394, October 2023.

Xuchen Liu, Jiawei Huang, Shihao Xia, Bingxi Liu, Jinqiang Cui, and Jiankun Yang. Imagineuav: Aerial vision-language navigation via world-action modeling and kinodynamic planning. 2026b. URL https://api.semanticscholar.org/CorpusID:288861777.

Yutong Liu, Xiaojie Li, Mingzhu Xu, and Jianlong Wu. Airforesight: Current-to-future spatial map imagination with cross-space planning consistency for UAV-VLN. In *Proceedings of the 34th ACM International Conference on Multimedia*, Rio de Janeiro, Brazil, November 2026c. ISBN 979-8-4007-2213-4. doi: 10.1145/3767308.3836065.

Jiabin Lou, Haopeng Wang, Yuanshuai Wang, Xinyu Liu, Xuxin Lv, Yuxin Guo, Lei Huang, Rongye Shi, and Wenjun Wu. Voln: Vision-only long-horizon navigation—paradigm, benchmark, and method. *arXiv preprint arXiv:2607.21400*, 2026. URL https://arxiv.org/abs/2607.21400.

Y Qi, H Li, S Huang, et al. Parse, search, and confirmation: Training-free aerial vision-and-dialog navigation with chain-of-thought reasoning and structured spatial memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23859–23868, 2026.

Yuankai Qi, Qi Wu, Peter Anderson, Xin Wang, William Yang Wang, Chunhua Shen, and Anton van den Hengel. Reverie: Remote embodied visual referring expression in real indoor environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

SimpleNav Contributors. Simplenav: Make navigation vla simple, 2026. URL https://github.com/OpenBMB/SimpleNav.

Yifei Su, Dong An, Kehan Chen, Weichen Yu, Baiyang Ning, Yonggen Ling, Yan Huang, and Liang Wang. Learning fine-grained alignment for aerial vision-dialog navigation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 7060–7068, 2025. doi: 10.1609/aaai.v39i7.32758.

Penglei Sun, Song Tang, Jiawen Wen, Runwei Guan, Liang Du, Chuxuan Zeng, Yuxuan Liang, Weiping Ding, Yang Yang, and Xiaowen Chu. From ground to space: A survey on multidomain data lifecycle for urban embodied agents. *IEEE Internet of Things Journal*, pp. 1–1, 2026a. doi: 10.1109/JIOT.2026.3710338.

Xiaolou Sun, Wufei Si, Wenhui Ni, Yuntian Li, Dongming Wu, Fei Xie, Runwei Guan, He-Yang Xu, Henghui Ding, Yuan Wu, et al. Autofly: Vision-language-action model for uav autonomous navigation in the wild. *arXiv preprint arXiv:2602.09657*, 2026b.

Hanqing Wang, Wenguan Wang, Wei Liang, Caiming Xiong, and Jianbing Shen. Structured scene memory for vision-language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8455–8464, June 2021.

Xiangyu Wang, Donglin Yang, Ziqin Wang, Hohin Kwan, Jinyu Chen, Hongsheng Li, Yue Liao, Si Liu, et al. Towards realistic uav vision-language navigation: Platform, benchmark, and methodology. In *International Conference on Learning Representations*, volume 2025, pp. 7292–7310, 2025.

Jianqiang Xiao, Yuexuan Sun, Yixin Shao, Boxi Gan, Rongqiang Liu, Yanjin Wu, Weili Guan, and Xiang Deng. Uav-on: A benchmark for open-world object goal navigation with aerial agents. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 13023–13029, 2025.

Huilin Xu, Zhuoyang Liu, Yixiang Luomei, and Feng Xu. Aerial vision-language navigation with a unified framework for spatial, temporal and embodied reasoning. *arXiv preprint arXiv:2512.08639*, 2025.

Huilin Xu, Zhuoyang Liu, Yixiang Luomei, and Feng Xu. Aerial vision-language navigation with a unified framework for spatial, temporal and embodied reasoning. *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2026a. doi: 10.1109/TCSVT.2026.3701286.

Yifang Xu, Jiahao Cui, Feipeng Cai, Zhihao Zhu, Hanlin Shang, Shan Luan, Mingwang Xu, Neng Zhang, Yaoyi Li, Jia Cai, and Siyu Zhu. Wam-flow: Parallel coarse-to-fine motion planning via discrete flow matching for autonomous driving. In *CVPR*, 2026b.

Ning Yang, Yan Huang, Kaiwen Peng, Ziheng He, Kai Wang, Cui Miao, Kailin Lyu, Guo Li, Xiaofeng Wang, Zheng Zhu, Jing Liu, and Nianfeng Liu. Wam-nav: Asymmetric latent world-action modeling for unified visual navigation. 2026. URL https://api.semanticscholar.org/CorpusID:288977929.

Xuan Yao, Junyu Gao, and Changsheng Xu. Navmorph: A self-evolving world model for vision-and-language navigation in continuous environments. *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5536–5546, 2025. URL https://api.semanticscholar.org/CorpusID:280011840.

Seonghyeon Ye, Yunhao Ge, Kaiyuan Zheng, Shenyuan Gao, Sihyun Yu, George Kurian, Suneel Indupuru, You Liang Tan, Chuning Zhu, Jiannan Xiang, Ayaan Malik, Kyungmin Lee, William Liang, Nadun Ranawaka, Jiasheng Gu, Yinzhen Xu, Guanzhi Wang, Fengyuan Hu, Avnish Narayan, Johan Bjorck, Jing Wang, Gwanghyun Kim, Dantong Niu, Ruijie Zheng, Yuqi Xie, Jimmy Wu, Qi Wang, Ryan Julian, Danfei Xu, Yilun Du, Yevgen Chebotar, Scott Reed, Jan Kautz, Yuke Zhu, Linxi "Jim" Fan, and Joel Jang. World action models are zero-shot policies, 2026. URL https://arxiv.org/abs/2602.15922.

Daoxuan Zhang, Ping Chen, Xiaobo Xia, Xiu Su, Ruichen Zhen, Jianqiang Xiao, and Shuo Yang. Apex: A decoupled memory-based explorer for asynchronous aerial object goal navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15232–15242, June 2026a.

Jiazhao Zhang, Kunyu Wang, Rongtao Xu, Gengze Zhou, Yicong Hong, Xiaomeng Fang, Qi Wu, Zhizheng Zhang, and He Wang. Navid: Video-based vlm plans the next step for vision-and-language navigation. *arXiv preprint arXiv:2402.15852*, 2024.

Jiazhao Zhang, Anqi Li, Yunpeng Qi, Minghan Li, Jiahang Liu, Shaoan Wang, Haoran Liu, Gengze Zhou, Yuze Wu, Xingxing Li, et al. Embodied navigation foundation model. In *International Conference on Learning Representations*, volume 2026, pp. 127293–127322, 2026b.

Lingfeng Zhang, Zeying Gong, Xiaoshuai Hao, Haoxiang Fu, Qiang Zhang, Mingliang Zhou, Hangjun Ye, Xiaojun Liang, Junwei Liang, and Wenbo Ding. Futurenav: Unified world-action modeling for vision-and-language navigation. *arXiv preprint arXiv:2606.30367*, 2026c.

Lingfeng Zhang, Zhanguang Zhang, Liheng Ma, Tongtong Cao, and Yingxue Zhang. Anticipate before acting: Future-state-conditioned vision-language navigation, 2026d. URL https://arxiv.org/abs/2607.18042.

Lingjun Zhang, Yujian Yuan, Changjie Wu, Xinyuan Chang, Xin Cai, Shuang Zeng, Linzhe Shi, Sijin Wang, Hang Zhang, and Mu Xu. Minddriver: Introducing progressive multimodal reasoning for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026e. URL https://arxiv.org/abs/2602.21952.

Baining Zhao, Jiacheng Xu, Weicheng Feng, Xin Zhang, Zhaolu Wang, Haoyang Wang, Shilong Ji, Ziyou Wang, Jianjie Fang, Zhiheng Zheng, Weichen Zhang, Yu Shang, Wei Wu, Chen Gao, Xinlei Chen, and Yong Li. Worldvln: Autoregressive world action model for aerial vision-language navigation, 2026. URL https://arxiv.org/abs/2605.15964.

Shengtao Zheng, Kai Li, Weichen Zhang, Yu Meng, Chen Gao, Xinlei Chen, Yong Li, and Xiao-Ping Zhang. Worldfly: A world-model-based vision-language-action model for uav navigation, 2026a. URL https://arxiv.org/abs/2606.06147.

Xiangyi Zheng, Xiangyu Wang, Qinan Liao, Zimu Tang, Yue Liao, Dongyue Lyu, Guodong Wang, Junjie Liu, and Si Liu. Think like a pilot: Fine-grained long-horizon uav navigation, 2026b. URL https://arxiv.org/abs/2606.06836.

Jiacong Zhou, Jiaxu Miao, Yourun Lin, Xianyun Wang, Jun Xiao, and Jun Yu. Memory-augmented scene understanding and exploration for open-world aerial object-goal navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21616–21626, 2026.

Enze Zhu, Luxiao Xu, Zhan Chen, Jiahui Cui, Jiayuan Wang, Yongkang Zou, Kaibo Yang, Xiaoxuan Liu, Xiyu Qi, and Lei Wang. Uav 3d scene understanding: A survey from an agent-capability evolution perspective. *Remote Sensing*, 2026. URL https://api.semanticscholar.org/CorpusID:290160048.

Fengda Zhu, Yi Zhu, Xiaojun Chang, and Xiaodan Liang. Vision-language navigation with self-supervised auxiliary reasoning tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

A Implementation Details

A.1 Route-Critical Anchors and Future Targets

Anchor construction. Let a demonstration contain T states and let $\mathbf{p}_i \in \mathbb{R}^3$ be the UAV position at state i. We first form an event-anchor set from the initial state, the terminal state, and every turning point whose angular change satisfies $\alpha_i \geq \tau_p$:

$$\mathcal{E} = \{1, T\} \cup \{i \mid \alpha_i \geq \tau_p\}. \tag{10}$$

The initial state makes the construction well defined from the beginning of a trajectory, while the terminal state guarantees a future target for every non-terminal training state. Turning points alone can be sparse on a long straight segment. Therefore, for each pair of consecutive event anchors $u < v$, we add the state whose accumulated travel distance from u is closest to each multiple of τ_g that lies strictly inside the interval:

$$\begin{aligned} s(u,j) &= \sum_{k=u+1}^{j} \|\mathbf{p}_k - \mathbf{p}_{k-1}\|_2, \\ j_m^\star &= \arg\min_{u<j<v} |s(u,j) - m\tau_g|, \qquad 0 < m\tau_g < s(u,v). \end{aligned} \tag{11}$$

Ties are resolved by choosing the earlier state, and duplicate indices caused by discrete sampling are removed. Sorting the union of event and coverage anchors gives the ordered set

$$\mathcal{A} = \mathrm{sort}\left(\mathcal{E} \cup \{j_m^\star\}_{u,v,m}\right). \tag{12}$$

This construction preserves turning events while preventing the target horizon from becoming excessively long in event-sparse portions of a route. As illustrated in Fig. 4, the resulting anchors capture turning events while maintaining coverage along trajectories of different lengths. In all experiments, $\tau_p = 20^\circ$ and $\tau_g = 30$ m.

Adaptive future target. For every non-terminal training state $t < T$, the next route-critical anchor is:

$$a_t^+ = \min\{a \in \mathcal{A} \mid a > t\}, \qquad \kappa_t = a_t^+ - t. \tag{13}$$

The two visual targets are consequently:

$$\mathbf{Z}_{t+1}^{p,*} = \psi(o_{t+1}), \qquad \mathbf{Z}_{t+\kappa_t}^{c,*} = \psi(o_{a_t^+}). \tag{14}$$

The proximal branch is always trained against the immediately following observation. In contrast, the route-critical branch is trained against the next turning, coverage, or terminal anchor, so its temporal offset adapts to the geometry of the demonstrated route. Importantly, κ_t is a label construction variable rather than an input to the model. The notation $t + \kappa_t$ describes which future state provides supervision; ForeFly is not given this offset at test time.

A.2 Multi-View Perception and Historical Memory

At each navigation step, the UAV receives five synchronized egocentric views, including front, rear, left, right, and down:

$$o_t = \left\{o_t^{\text{front}}, o_t^{\text{rear}}, o_t^{\text{left}}, o_t^{\text{right}}, o_t^{\text{down}}\right\}. \tag{15}$$

All five views are used for current perception, while historical memory stores only the front-view features. For each historical frame, the visual encoder produces one global token and 16 local tokens, $\mathbf{x}_i^{\text{front}} = [\mathbf{g}_i; \mathbf{l}_{i,1}, \ldots, \mathbf{l}_{i,16}]$. The recent memory $\mathcal{M}_t^p$ retains all 17 tokens from the latest $R = 5$ front-view observations. Keeping both global and local tokens preserves the fine-grained appearance and spatial details needed to model short-term visual continuity. In contrast, the route-critical memory $\mathcal{M}_t^c$ stores only the global token $\mathbf{g}_i$ from each selected historical anchor, since its purpose is to summarize sparse route-level events rather than local visual details.

Both memories are constructed from the executed trajectory prefix. In particular, a turning-point anchor becomes available only after the subsequent state required to determine its turning angle has been observed, and coverage anchors are likewise identified only from past executed states. Future

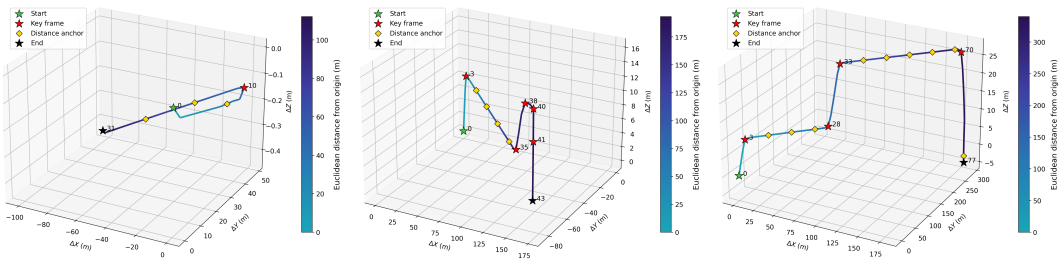

Figure 4: **Visualization of route-critical anchor construction.** Representative UAV trajectories of different lengths are shown. Green and black stars mark the start and end states, red stars indicate key frames at turning points, and yellow diamonds denote distance anchors. The color scale represents the Euclidean distance from the starting position.

anchors may therefore be used to define training targets, but never enter the model input before they become observable.

At inference time, ForeFly is conditioned only on:

$$\mathcal{C}_t^{\text{in}} = (\mathcal{I}, \mathcal{P}_{\le t}, o_t, \mathcal{M}_t^p, \mathcal{M}_t^c)\,, \tag{16}$$

and does not access future observations, future anchor indices, or the adaptive offset κ_t, thereby preventing target leakage.

A.3 Online Memory Update

Algorithm 1 describes the online memory update procedure. After observing state t, only turns at indices $i \le t-1$ can be confirmed. The current index t serves as a temporary coverage boundary, not as a known terminal anchor. Thus, distance anchors can be selected before the next turn is observed. Let D_i denote cumulative travel distance from the initial state, so that $s(u, j) = D_j - D_u$.

Algorithm 1 Online update of recent and route-critical memories

Require: Observed position $\mathbf{p}_t$ and front-view tokens $\mathbf{x}_t^{\text{front}} = [\mathbf{g}_t; \mathbf{l}_{t,1}, \ldots, \mathbf{l}_{t,16}]$
Require: Recent window $R = 5$, turning threshold $\tau_p = 20^\circ$, coverage interval $\tau_g = 30\,\text{m}$
Maintain confirmed event indices $\mathcal{E}$ and cached past global tokens
Cache $\mathbf{g}_t$; append $\mathbf{x}_t^{\text{front}}$ to the recent FIFO queue
Retain the latest $\min(R, t)$ entries as $\mathcal{M}_t^p$
if $t = 1$ **then**
    $D_1 \leftarrow 0$; $\mathcal{E} \leftarrow \{1\}$
else
    $D_t \leftarrow D_{t-1} + \|\mathbf{p}_t - \mathbf{p}_{t-1}\|_2$
end if
if $t \ge 3$ and both adjacent displacements are nonzero **then**
    $\mathbf{v}_{t-1} \leftarrow (\mathbf{p}_{t-1} - \mathbf{p}_{t-2})/\|\mathbf{p}_{t-1} - \mathbf{p}_{t-2}\|_2$
    $\mathbf{v}_t \leftarrow (\mathbf{p}_t - \mathbf{p}_{t-1})/\|\mathbf{p}_t - \mathbf{p}_{t-1}\|_2$
    $\alpha_{t-1} \leftarrow \arccos(\text{clip}(\mathbf{v}_{t-1}^\top \mathbf{v}_t, -1, 1))$
    if $\alpha_{t-1} \ge \tau_p$ **then**
        $\mathcal{E} \leftarrow \mathcal{E} \cup \{t-1\}$ ▷ Confirm the previous state as a turn
    end if
end if
$\mathcal{B}_t \leftarrow \text{sort}(\mathcal{E} \cup \{t\})$; $\mathcal{A}_t^{\text{mem}} \leftarrow \mathcal{E}$
for each consecutive pair (u, v) in $\mathcal{B}_t$ with $v > u + 1$ **do**
    for each positive integer m satisfying $m\tau_g < D_v - D_u$ **do**
        $j^\star \leftarrow \arg\min_{u<j<v} |D_j - D_u - m\tau_g|$ ▷ Break ties toward earlier j
        $\mathcal{A}_t^{\text{mem}} \leftarrow \mathcal{A}_t^{\text{mem}} \cup \{j^\star\}$
    end for
end for
$\mathcal{M}_t^c \leftarrow [\mathbf{g}_i]_{i \in \text{sort}(\mathcal{A}_t^{\text{mem}})}$
return $\mathcal{M}_t^p, \mathcal{M}_t^c$ for foresight query priming at step t

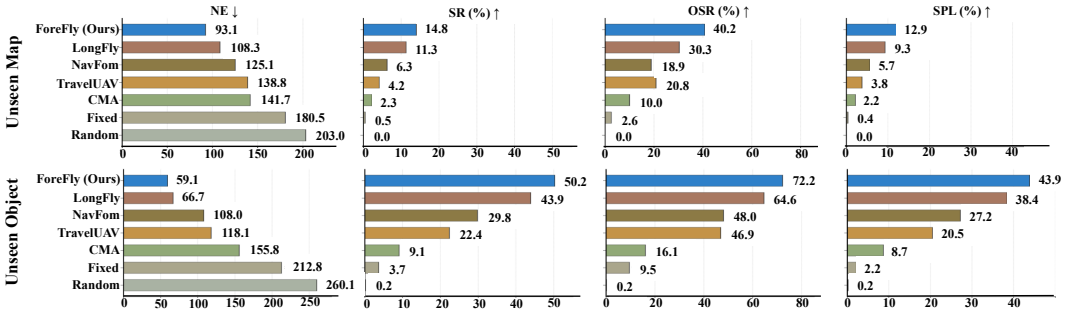

Figure 5: **Results on TravelUAV Unseen Map and Unseen Object subsets.** The top row compares ForeFly with existing methods on the Unseen Map subset, and the bottom row reports results on the Unseen Object subset.

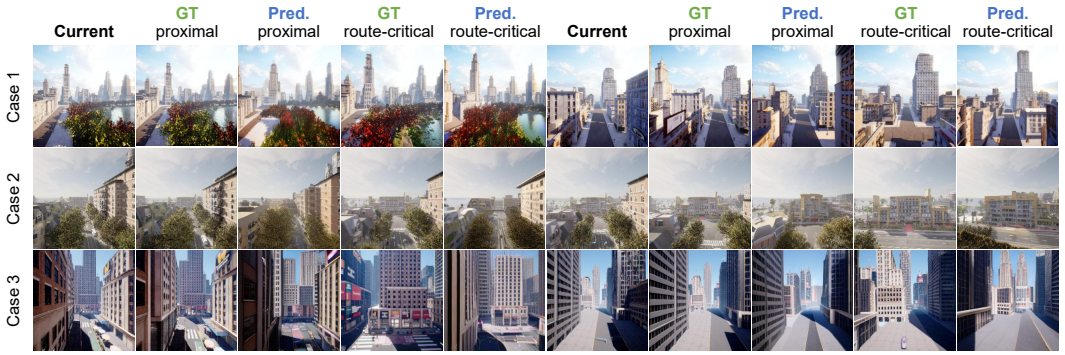

Figure 6: **Qualitative visualization of dual-horizon future prediction.** Each row shows two sampled waypoints from the same navigation trajectory. For each waypoint, we visualize the current observation together with the ground-truth and predicted proximal and route-critical futures.

B Additional Experiment Results

B.1 Quantitative Results on Unseen Map and Unseen Object Subsets

Following the original TravelUAV evaluation protocol, the Unseen split consists of two subsets: Unseen Map and Unseen Object. The Unseen Map subset contains trajectories in previously unseen environments, while the Unseen Object subset contains trajectories with unseen goal objects.

As shown in Fig. 5, ForeFly consistently outperforms all comparison methods on both Unseen Map and Unseen Object across the evaluation metrics. On Unseen Map, the improvement is particularly pronounced in OSR, where ForeFly reaches 40.2% compared with 30.3% for LongFly, indicating stronger robustness to previously unseen environments. On Unseen Object, ForeFly also achieves consistent gains in navigation accuracy, success rate, and path efficiency. These results demonstrate that ForeFly generalizes effectively to both environmental and target-level distribution shifts.

B.2 Qualitative Analysis of Dual-Horizon Foresight

To further examine whether the predicted dual-horizon latents capture meaningful future states, we provide an image-space visualization of ForeFly's foresight representations. Note that ForeFly itself predicts compact visual latents rather than RGB observations. For qualitative interpretation only, we employ an auxiliary latent-diffusion reconstructor that takes the predicted foresight representations as conditioning signals and reconstructs the corresponding future observations in image space. The reconstruction module is used solely for visualization and does not participate in navigation decision making. Since the reconstruction is produced by an auxiliary generative model, its visual fidelity depends on both the quality of the predicted foresight representations and the capacity of the reconstructor. Therefore, these visualizations should be interpreted as qualitative evidence of

the semantic and spatial information retained in the predicted representations, rather than as direct pixel-level future prediction.

As shown in Fig. 6, each row presents two sampled waypoints from the same navigation trajectory. At each waypoint, we visualize the current observation together with the ground-truth and reconstructed proximal and route-critical futures. The proximal predictions largely preserve the local scene layout and viewpoint continuity, whereas the route-critical predictions exhibit more pronounced changes in scene composition, reflecting farther route progression.

Across all three cases, the reconstructed futures remain broadly consistent with the corresponding ground-truth observations in their dominant spatial and semantic structures. Moreover, the predictions evolve with the navigation progress rather than collapsing to a fixed future pattern. These results qualitatively demonstrate the complementary roles of the two foresight horizons: proximal foresight captures short-term scene evolution, while route-critical foresight anticipates longer-range navigation-relevant changes.

C Limitations and Future Directions

ForeFly has several limitations that define directions for future work. First, route-critical targets depend on turning and distance heuristics. The fixed thresholds work well on TravelUAV, but different vehicle dynamics, sampling rates, or environment scales may require recalibration or a learned event detector. Second, each state is paired with one demonstrated future. The squared latent objective does not explicitly represent multiple valid futures when several routes satisfy an instruction. Probabilistic latent prediction would be a natural extension. Third, retrospective memory is built from the executed trajectory and can inherit errors after route deviation. Explicit uncertainty estimates or memory recovery mechanisms could improve robustness in such cases.

Real-world UAV experiments are important for validating practical applicability, but large-scale aerial evaluation is constrained by safety, regulatory compliance, privacy, and site-access requirements. Following the standard AVLN protocol, we evaluate ForeFly on TravelUAV and UAV-ON, which provide diverse large-scale outdoor environments, realistic visual observations, and seen/unseen settings for controlled and reproducible evaluation. Existing methods such as SimpleNav, NavFoM, AirForsight, LongFly, and SpatialFly are also primarily evaluated in simulation, enabling fair comparison under a common protocol. Future work will further investigate sim-to-real transfer and physical UAV deployment to assess ForeFly under real-world conditions.