# NAVGROUND3D: UAV-CENTRIC GROUNDING OF SEMANTIC–GEOMETRIC CUES FOR AERIAL VISION-LANGUAGE NAVIGATION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Aerial Vision-and-Language Navigation (UAV-VLN) requires UAVs to follow natural-language instructions, understand scene semantics and spatial relationships in complex 3D environments, and navigate autonomously. Recent studies have begun to leverage geometric visual models in UAV-VLN to enhance 3D spatial understanding. However, existing approaches mainly focus on aligning or fusing semantic and geometric information, while paying limited attention to effective visual information distillation and unified UAV-centric directional modeling. To address these issues, we propose **NavGround3D**, a semantic-geometric representation framework that progressively distills visual information and grounds it in a unified UAV-centric directional space. Specifically, **Semantic-Geometric Visual Consolidation (SGVC)** selects representative anchors based on semantic saliency and aggregates redundant representations according to semantic-geometric consistency, yielding more informative visual tokens. **UAV-Centric Directional Grounding (UCDG)** then transforms view-dependent visual cues into the UAV body frame and introduces continuous directional encoding to establish consistent cross-view spatial relationships. The resulting representations are integrated with navigation context for motion decision-making. Extensive experiments on TravelUAV and UAV-ON demonstrate consistent improvements over strong baselines, validating the effectiveness of semantic-geometric distillation and UAV-centric directional grounding for aerial navigation.

## 1 INTRODUCTION

UAV Vision-and-Language Navigation (UAV-VLN) Liu et al. (2023); Wang et al. (2025b); Lee et al. (2025); Gao et al. (2026) aims to enable aerial agents to interpret natural-language instructions, understand their surroundings, and reach target locations through sequential decision making. With the increasing deployment of UAVs in applications such as autonomous inspection Kaleci et al. (2026); Gao et al. (2025); Gil-Castilla et al. (2025) and emergency response Harshini et al. (2025); Garba (2026), UAV-VLN has emerged as an important research direction in embodied intelligence.

Compared with indoor or ground-based vision-language navigation, UAV-VLN operates in more open and larger-scale three-dimensional environments, with substantially greater viewpoint variation. As illustrated in Figure 1, these characteristics introduce two major challenges. First, aerial views often cover large spatial regions containing repetitive structures and visually similar areas, resulting in substantial visual redundancy. Meanwhile, the spatial structure of large-scale aerial scenes is more complex, making it difficult for limited visual representations to emphasize navigation-relevant information. Second, due to payload constraints, UAVs typically rely on a small set of fixed-orientation cameras to obtain discrete view observations. Since different views are associated with different local directional references, unified spatial relation modeling and directional representation become more challenging.

Recent advances in vision geometry foundation models provide a promising basis for enhancing spatial representation in UAV-VLN. Models such as VGGT Wang et al. (2025a) can infer implicit 3D geometric priors and cross-view geometric relationships directly from RGB images, providing richer spatial structural information beyond conventional semantic features. Recent studies have



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.

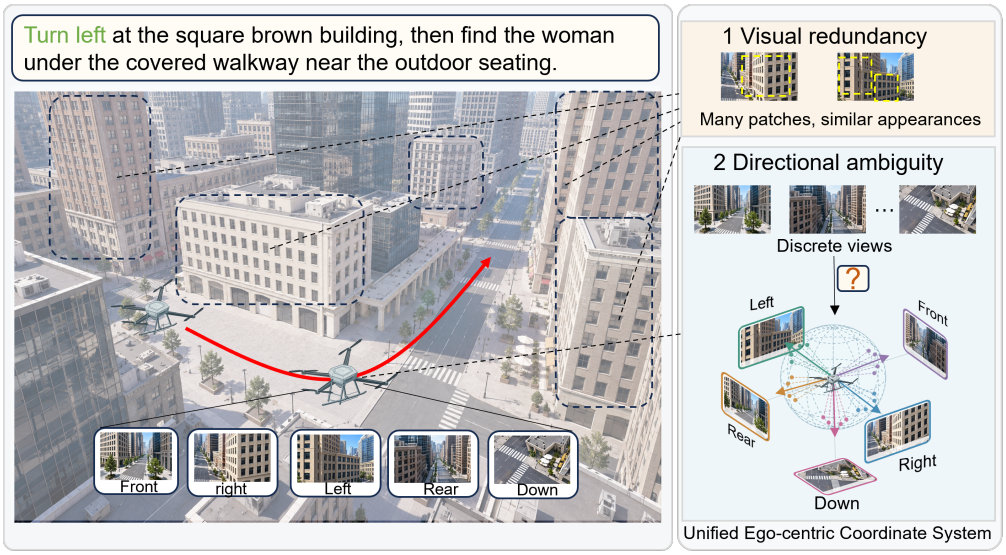


Figure 1: Key challenges of UAV-VLN in large-scale aerial environments. Large-scale aerial observations contain many visually similar patches and repetitive structures, leading to substantial visual redundancy. Meanwhile, the discrete camera views have different local directional references, introducing directional ambiguity and making it difficult to establish a unified spatial representation across views.

also begun to incorporate such geometric representations into UAV-VLN to improve spatial understanding Jiang et al. (2026). However, existing approaches mainly focus on aligning or fusing semantic and geometric features, while paying limited attention to visual redundancy and directional discrepancies among discrete views in large-scale aerial environments. Specifically, how to exploit semantic-geometric correspondence to distill representative navigation-relevant information from redundant visual observations, and further organize such information within a unified directional reference frame to form compact and direction-aware navigation representations, remains insufficiently explored.

To address these challenges, we propose **NavGround3D**, a unified representation framework for UAV-VLN that aims to distill compact navigation-relevant information from redundant visual observations and establish consistent spatial representations under a unified directional reference. NavGround3D consists of two complementary components. First, we introduce **Semantic-Geometric Visual Consolidation (SGVC)**, which consolidates dense visual representations by jointly modeling the complementary relationship between semantic information and geometric structure. SGVC selects representative visual anchors based on 2D attention responses and aggregates redundant local information into the corresponding anchors according to semantic similarity and geometric consistency, thereby suppressing redundancy while preserving navigation-relevant semantic and structural cues. Building on these consolidated representations, we further introduce **UAV-Centric Directional Grounding (UCDG)**, which maps local visual information from different observation directions into a unified UAV-centric directional reference and explicitly encodes their spatial orientations through continuous 3D directional representations. A cross-view Transformer then models interactions among tokens from different directions, yielding navigation representations that are compact, geometry-aware, and direction-aware for subsequent motion decision-making. Extensive experiments on the TravelUAV and UAV-ON datasets show that NavGround3D achieves competitive performance, demonstrating the effectiveness of the proposed method.

Our contributions are summarized as follows:

- We propose **NavGround3D**, a unified framework that converts dense multi-view observations into compact semantic-geometric representations with explicit UAV-centric directional grounding for aerial vision-language navigation.

- NavGround3D consists of two key components: **SGVC**, which consolidates redundant local observations by jointly exploiting semantic saliency and geometric consistency, and **UCDG**, which grounds the resulting compact cues into a continuous UAV-centric directional space and captures their cross-view spatial relations.
- Extensive experiments on TravelUAV demonstrate that NavGround3D consistently achieves superior navigation performance over strong state-of-the-art methods, validating the effectiveness of both semantic-geometric consolidation and directional grounding.

## 2 Related Work

**Vision-and-language navigation for UAVs.** UAV Vision-and-Language Navigation (UAV-VLN) requires an unmanned aerial vehicle to follow natural-language instructions and navigate autonomously in large-scale three-dimensional aerial environments. Recent benchmarks, such as TravelUAV and UAV-On Wang et al. (2025b); Xiao et al. (2025), have advanced UAV-VLN by providing continuous and diverse aerial navigation scenarios. Existing UAV-VLN methods mainly improve navigation through memory modeling, world models, and geometric reasoning. Memory-based approaches preserve historical observations or construct spatial memories to support long-horizon decision making Zhou et al. (2026); Fan et al. (2026a); Zhang et al. (2026a); Deng & Xu (2026); Shao et al. (2026); world-model-based methods predict future observations or latent states to facilitate planning and action selection Yao et al. (2024); Zhao et al. (2026); Zheng et al. (2026); Li et al. (2026); Zhu et al. (2026); and geometry-aware methods exploit depth, top-down representations, or explicit 3D reconstruction to enhance spatial understanding Gao et al. (2024); Su et al. (2025); Deng (2025). More recently, vision geometry foundation models have enabled implicit 3D geometric representation learning directly from RGB observations Jiang et al. (2026). However, existing implicit 3D approaches mainly focus on aligning or fusing semantic and geometric information, while paying limited attention to effective visual information distillation and unified UAV-centric directional modeling. This is particularly important in aerial navigation, where dense visual observations contain substantial redundant information and directional cues from different views need to be interpreted under a consistent UAV-centric reference.

**Spatial and Geometry-Aware Multimodal Representations.** 3D-aware representations provide navigation agents with spatial cues, motivating geometry-based approaches for spatial reasoning. Early methods rely on explicit 3D representations, such as depth-based maps, top-down structures, and reconstructed scenes, but often require additional sensing inputs or continuous reconstruction Gao et al. (2024); Su et al. (2025); Deng (2025). Recently, vision geometry foundation models such as VGGT Wang et al. (2025a) have enabled implicit 3D representation learning by extracting geometric priors and cross-view correspondences directly from RGB observations, which has been increasingly explored for multimodal spatial reasoning and navigation Fan et al. (2026b); Zhang et al. (2026b); Zeng et al. (2026); Yang et al. (2026); Jiang et al. (2026). However, existing implicit 3D methods mainly focus on aligning or fusing semantic and geometric features, while paying limited attention to effective visual information distillation and unified directional modeling. We address these limitations by consolidating multi-view semantic-geometric representations and grounding them into a UAV-centric directional space, enabling compact and spatially consistent UAV-VLN representations.

## 3 Method

### 3.1 Preliminaries

**Problem Formulation.** UAV-VLN aims to guide a UAV to a specified location in a continuous 3D environment by following a natural-language instruction and perceiving its surroundings. At time step $t$, the UAV receives its current state $\mathbf{p}_t$ and synchronized RGB observations $\mathcal{O}_t = \{I_{t,v}\}_{v\in\mathcal{V}}$ from five onboard cameras, where $\mathbf{p}_t$ represents the current UAV position and orientation, and $\mathcal{V} = \{F, L, R, B, D\}$ denotes the front, left, right, rear, and down views, respectively. Conditioned on the instruction $I$, current state $\mathbf{p}_t$, and multi-view observations $\mathcal{O}_t$, the navigation policy predicts the next local waypoint:

$$\widehat{\mathbf{W}}_t = \pi_\theta(I, \mathbf{p}_t, \mathcal{O}_t) = [\hat{\mathbf{u}}_t^\top, \hat{\ell}_t]^\top. \tag{1}$$

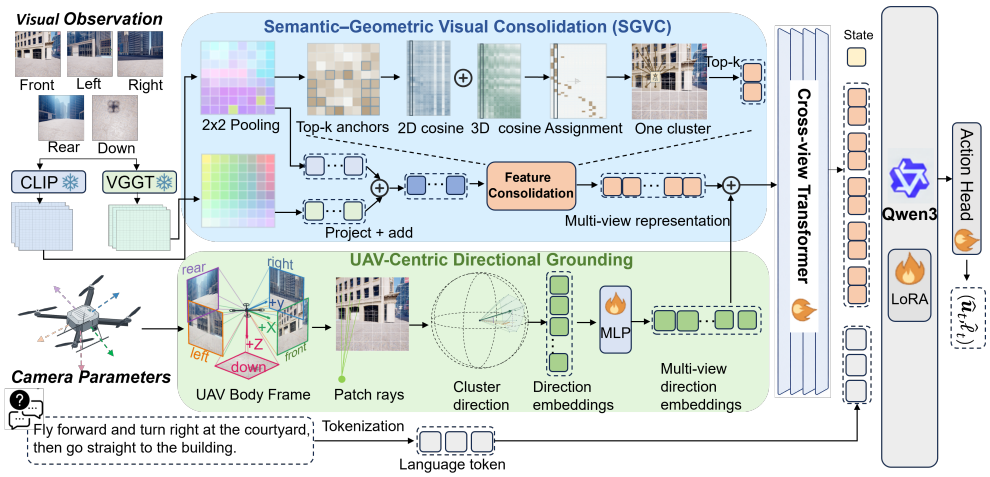


Figure 2: Overview of the proposed NavGround3D framework. Semantic-Geometric Visual Consolidation (SGVC) first aligns semantic features from CLIP with geometric features from VGGT, and distills dense visual patches through semantic-guided anchor selection and semantic-geometric aggregation. UAV-Centric Directional Grounding (UCDG) then transforms view-dependent rays into the UAV body frame and encodes them within a unified directional space. A view-aware Transformer further models cross-view interactions among the resulting visual tokens. Finally, The resulting visual tokens are used to predict local 3D motion.

Here, $\hat{\mathbf{u}}_t \in \mathbb{R}^3$ and $\hat{\ell}_t \in \mathbb{R}$ denote the predicted local motion direction and the corresponding travel distance, respectively. The UAV then iteratively executes the predicted local motion until reaching the target location or satisfying the stopping condition.

**Framework Overview.** As shown in Fig 2, for the above navigation task, we propose a unified framework for UAV-VLN that jointly models semantic-geometric consolidation and UAV-centric directional grounding. Specifically, semantic and geometric features are first extracted from the five camera views. Section 3.2 introduces Semantic-Geometric Visual Consolidation (SGVC), which distills visual information through semantic-guided anchor selection and semantic-geometric aggregation. Section 3.3 presents UAV-Centric Directional Grounding (UCDG), which further introduces a unified UAV-centric directional reference to explicitly model 3D directional information. Finally, direction-aware representation is combined with the navigation instruction and current UAV state to predict the next local waypoint.

### 3.2 Semantic-Geometric Visual Consolidation

Recently, some UAV-VLN methods have begun to introduce implicit 3D geometric representations to compensate for the limitations of conventional 2D visual features in modeling spatial relationships and geometric structures. However, in large-scale aerial environments, wide-field multi-view observations often contain extensive background regions, repetitive structures, and visually similar content. Retaining all dense semantic and geometric features throughout navigation introduces substantial redundancy, making it difficult to obtain concise and effective navigation representations.

To address this issue, we introduce **Semantic-Geometric Visual Consolidation (SGVC)**, which exploits the complementarity between semantic and geometric information to distill dense visual representations within each view. For each view $I_{t,v}$, frozen CLIP and VGGT encoders are used to extract semantic and geometric features, respectively. After spatial alignment and consistent pooling, both feature maps are organized on a shared $8 \times 8$ grid, yielding $N = 64$ candidate spatial locations. We denote the semantic and geometric features at the $i$-th location as $\mathbf{s}_{t,v,i}$ and $\mathbf{g}_{t,v,i}$, respectively. The two modalities are then projected into a shared embedding space through modality-specific projections $P_s(\cdot)$ and $P_g(\cdot)$, and fused at the corresponding spatial location to obtain the local semantic-geometric representation $\mathbf{z}_{t,v,i} = P_s(\mathbf{s}_{t,v,i}) + P_g(\mathbf{g}_{t,v,i})$. After projection and fusion, each view is represented by $N$ local semantic-geometric tokens, denoted as $\mathcal{Z}_{t,v} = \{\mathbf{z}_{t,v,i}\}_{i=1}^{N}$.

To further reduce redundancy from repetitive backgrounds and similar structures, we select a small set of representative spatial locations as anchors. Specifically, the self-attention responses from the 2D encoder are used as a lightweight semantic saliency prior for anchor selection. Let $a_{t,v,i}$ denote the response at the $i$-th spatial location. The top-$K$ locations with the highest responses are selected as semantic anchors:

$$\mathcal{A}_{t,v} = \mathrm{TopK}\left(\{a_{t,v,i}\}_{i=1}^{N},\, K\right), \tag{2}$$

where $\mathcal{A}_{t,v} = \{k_{t,v,1}, \ldots, k_{t,v,K}\}$ denotes the spatial indices of the selected anchors, and we set $K = 16$ in our implementation. The corresponding anchor representations form the set $\mathcal{Z}_{t,v}^{A} = \{\mathbf{z}_{t,v,k_{t,v,j}}\}_{j=1}^{K}$, while the remaining candidates form the residual set $\tilde{\mathcal{Z}}_{t,v} = \mathcal{Z}_{t,v} \setminus \mathcal{Z}_{t,v}^{A}$.

Prior studies have shown that semantic and geometric representations capture different but complementary properties: semantic features preserve fine-grained appearance and category-related information, whereas geometric features better encode spatial proximity and structural consistency Wang et al. (2026); Pan et al. (2026). Motivated by this complementarity, we do not directly discard the non-anchor tokens. Instead, each non-anchor token is assigned and aggregated to the most relevant anchor by jointly considering semantic similarity and geometric consistency. For a non-anchor location $i$ and an anchor $k \in \mathcal{A}_{t,v}$, we define their joint similarity as

$$R_{t,v,i,k} = \frac{\cos(\mathbf{s}_{t,v,i}, \mathbf{s}_{t,v,k})}{\tau_s} + \frac{\cos(\mathbf{g}_{t,v,i}, \mathbf{g}_{t,v,k})}{\tau_g}, \qquad m_{t,v}(i) = \arg\max_{k \in \mathcal{A}_{t,v}} R_{t,v,i,k}, \tag{3}$$

where $\tau_s$ and $\tau_g$ control the relative scales of semantic and geometric similarity, respectively. The semantic term measures consistency in local visual content, while the geometric term provides complementary structural constraints. Each non-anchor token is assigned to the anchor with the highest joint similarity. For each anchor $k$, the non-anchor locations assigned to it form the corresponding support set

$$\mathcal{C}_{t,v,k} = \{i \mid i \notin \mathcal{A}_{t,v},\; m_{t,v}(i) = k\}. \tag{4}$$

The tokens within each support set are aggregated into the corresponding anchor to update its representation. This reduces each view $v$ from $N$ local tokens to $K$ compact semantic-geometric representations, denoted by $\mathbf{Z}_{t,v} \in \mathbb{R}^{K \times d}$. Applying the same procedure independently to all five views and concatenating their outputs yields the final multi-view representation $\mathbf{Z}_t^{\mathrm{SGVC}} \in \mathbb{R}^{5K \times d}$.

Through this process, SGVC distills dense local representations into a fixed number of semantic-geometric tokens, reducing redundancy while preserving salient semantic and structural information to form a more compact and spatially discriminative multi-view representation.

### 3.3 UAV-Centric Directional Grounding

In large-scale aerial environments, payload limitations restrict UAVs to a limited number of discrete camera views, each corresponding to a distinct local observation direction. Consequently, these observations lack a unified directional reference for cross-view spatial modeling. Meanwhile, UAV-VLN instructions frequently contain directional expressions such as moving forward, turning left, and shifting right. Explicitly modeling the relative orientation of visual content facilitates relating visual observations to directional expressions and supports direction-aware navigation reasoning. Motivated by this, we introduce **UAV-Centric Directional Grounding (UCDG)**, which maps discrete multi-view representations into a unified UAV-centric directional space and augments them with explicit 3D directional information.

We adopt the UAV body frame as the shared ego-centric reference, with the $x$-, $y$-, and $z$-axes pointing forward, rightward, and downward, respectively. Each rigidly mounted camera view $v \in \{F, L, R, B, D\}$ has a fixed orientation relative to the UAV, and the corresponding camera-to-body rotation matrix $\mathbf{R}_{c,v}^{b}$ maps view-specific local directions into this common reference frame. Detailed rotation definitions for all views are provided in the Appendix.

Specifically, for a local patch $i$ in view $v$, its camera-frame viewing ray is recovered from the patch center and transformed into the UAV body frame. Following the SGVC grouping, the directions of

an anchor and its support set are further aggregated as

$$\mathbf{d}_{v,i}^{c} = \mathrm{norm}\begin{bmatrix} 1 \\ (u_i - c_x)/f_x \\ (v_i - c_y)/f_y \end{bmatrix}, \quad \mathbf{d}_{v,i}^{ego} = \mathbf{R}_{c,v}^{b}\mathbf{d}_{v,i}^{c}, \quad \bar{\mathbf{d}}_{t,v,k}^{ego} = \mathrm{norm}\left(\mathbf{d}_{v,k}^{ego} + \sum_{i \in C_{t,v,k}} \mathbf{d}_{v,i}^{ego}\right). \tag{5}$$

Here, $\mathrm{norm}(\mathbf{r}) = \mathbf{r}/\max(\|\mathbf{r}\|_2, \epsilon)$ denotes vector normalization, where $\epsilon$ is a small constant for numerical stability. The aggregated direction $\bar{\mathbf{d}}_{t,v,k}^{ego} = [d_x, d_y, d_z]^\top$ represents the continuous 3D viewing direction of the $k$-th consolidated token in the UAV body frame. This formulation jointly accounts for the camera mounting orientation and the image-plane location of each patch, thereby placing observations from all five views in a common UAV-centric directional space.

Based on the aggregated direction, we further construct a hybrid directional encoding. Although the 3D unit vector provides a complete continuous orientation, navigation-relevant directional relations are more naturally characterized by horizontal and vertical angular variations. We therefore derive the azimuth angle $\alpha$ and vertical angle $\beta$ from $\bar{\mathbf{d}}_{t,v,k}^{\mathrm{ego}} = [d_x, d_y, d_z]^\top$ as $\alpha = \mathrm{atan2}(d_y, d_x)$ and $\beta = \mathrm{atan2}(d_z, \sqrt{d_x^2 + d_y^2})$, and combine the Cartesian and periodic angular components as

$$\gamma\left(\bar{\mathbf{d}}_{t,v,k}^{\mathrm{ego}}\right) = [d_x, d_y, d_z, \sin\alpha, \cos\alpha, \sin\beta, \cos\beta]^\top. \tag{6}$$

The Cartesian components preserve continuous 3D orientation, while the periodic angular terms explicitly encode horizontal and vertical directional relations without introducing discontinuities at angular boundaries.

We then use a learnable mapping $E_d(\cdot)$ to project the directional descriptor into a $d$-dimensional embedding space, obtaining the corresponding directional embedding $\mathbf{e}_{t,v,k}^{\mathrm{dir}}$. Following the same view and anchor ordering as $\mathbf{Z}_t^{\mathrm{SGVC}}$, all multi-view directional embeddings are organized into $\mathbf{E}_t^{\mathrm{dir}} \in \mathbb{R}^{5K \times d}$. We then inject the directional encoding into $\mathbf{Z}_t^{\mathrm{SGVC}}$ and further model the relationships among the multi-view tokens using a Transformer with a cross-view attention mask:

$$\mathbf{H}_t^{\mathrm{dir}} = \mathbf{Z}_t^{\mathrm{SGVC}} + \mathbf{E}_t^{\mathrm{dir}}, \qquad \mathbf{Z}_t^{\mathrm{UCDG}} = \mathcal{T}\left(\mathbf{H}_t^{\mathrm{dir}}\right). \tag{7}$$

Here, $\mathbf{H}_t^{\mathrm{dir}} \in \mathbb{R}^{5K \times d}$ denotes the multi-view tokens augmented with directional information, and $\mathbf{Z}_t^{\mathrm{UCDG}} \in \mathbb{R}^{5K \times d}$ is the final output of UCDG. The Transformer models spatial relationships among tokens under the shared UAV-centric directional reference. In this way, SGVC distills semantic-geometric information, while UCDG grounds the resulting representations in a unified UAV-centric directional space, yielding direction-aware visual representations for local waypoint prediction.

### 3.4 Action Prediction

The visual representation $\mathbf{Z}_t^{\mathrm{UCDG}}$, together with the navigation instruction $I$ and current UAV pose $\mathbf{p}_t$, is used by the navigation policy to predict the local waypoint defined in Eq. (1):

$$\widehat{\mathbf{W}}_t = \pi_\theta\left(I, \mathbf{p}_t, \mathbf{Z}_t^{\mathrm{UCDG}}\right) = [\hat{\mathbf{u}}_t^\top, \hat{\ell}_t]^\top \in \mathbb{R}^4, \tag{8}$$

where $\hat{\mathbf{u}}_t$ specifies the predicted local motion direction and $\hat{\ell}_t$ denotes the travel distance.

**Training objective.** The action prediction objective combines direction and distance supervision:

$$\mathcal{L}_{\mathrm{action}} = \frac{1}{|\mathcal{B}|}\sum_{t \in \mathcal{B}}\left[\mathcal{L}_{\mathrm{dir}}(\hat{\mathbf{u}}_t, \mathbf{u}_t^\star) + \lambda_{\mathrm{dist}}\,\mathcal{L}_{\mathrm{dist}}(\hat{\ell}_t, \ell_t^\star)\right], \tag{9}$$

where $\mathcal{B}$ denotes a training minibatch, and $(\mathbf{u}_t^\star, \ell_t^\star)$ denotes the expert direction-distance target. $\mathcal{L}_{\mathrm{dir}}$ and $\mathcal{L}_{\mathrm{dist}}$ denote the cosine loss for motion direction and the $\ell_1$ loss for travel distance, respectively, while $\lambda_{\mathrm{dist}}$ balances their contributions.

# 4 EXPERIMENTS

## 4.1 SETUP

**Dataset and metrics.** We evaluate NavGround3D on TravelUAV Wang et al. (2025c) and UAV-ON Xiao et al. (2025). TravelUAV is a 6-DoF UAV-VLN benchmark containing 12,149 human-operated trajectories across 89 target categories, with five-view RGB observations, waypoint annotations, and language instructions under Train, Test-Seen, and Test-Unseen splits. UAV-ON is an open-world aerial Object Goal Navigation benchmark in high-fidelity outdoor environments, where agents search for targets based on high-level semantic descriptions and are evaluated under Seen and Unseen settings. Following prior work Anderson et al. (2018); Wang et al. (2025c), we report NE, SR, OSR, and SPL as evaluation metrics. Further details on the datasets and metrics are provided in the supplementary material.

Table 1: Comparison on the TravelUAV Test-Seen and Test-Unseen splits. Only results on the Full set are reported. NE is measured in meters. Lower NE (↓) and higher SR, OSR, and SPL (↑) indicate better performance. Best results are shown in bold.

| **Method** | **Test-Seen** | | | | **Test-Unseen** | | | |
|---|---|---|---|---|---|---|---|---|
| | **NE↓** | **SR↑** | **OSR↑** | **SPL↑** | **NE↓** | **SR↑** | **OSR↑** | **SPL↑** |
| Random Action | 222.20 | 0.14 | 0.21 | 0.07 | 225.64 | 0.06 | 0.06 | 0.06 |
| Fixed Action | 188.61 | 2.27 | 8.16 | 1.40 | 193.30 | 1.76 | 5.36 | 1.09 |
| CMA Anderson et al. (2018) | 135.73 | 8.37 | 18.72 | 7.90 | 147.27 | 4.98 | 12.41 | 4.74 |
| TravelUAV Wang et al. (2025c) | 106.28 | 16.10 | 44.26 | 14.30 | 130.60 | 11.41 | 31.13 | 10.45 |
| NavFoM Zhang et al. (2026c) | 93.05 | 29.17 | 49.24 | 25.03 | 118.34 | 15.63 | 30.46 | 14.21 |
| LongFly Jiang et al. (2025) | 60.02 | 36.39 | 65.87 | 31.07 | 91.84 | 24.19 | 43.86 | 20.84 |
| SpatialFly Jiang et al. (2026) | 58.05 | 38.54 | 68.62 | 33.43 | 87.82 | 25.46 | 44.41 | 21.76 |
| Baseline (Ours, 0.6B) | 99.65 | 22.25 | 53.78 | 21.80 | 110.47 | 12.96 | 25.41 | 11.63 |
| **NavGround3D (Ours, 0.6B)** | 68.12 | 35.45 | 66.51 | 30.24 | 81.75 | 22.81 | 51.05 | 20.15 |
| Baseline (Ours, 4B) | 71.22 | 25.75 | 55.75 | 23.15 | 103.81 | 15.17 | 30.45 | 13.76 |
| **NavGround3D (Ours, 4B)** | **57.13** | **39.76** | **69.42** | **34.51** | **78.06** | **26.43** | **53.29** | **22.52** |

**Implementation and Training Details.** NavGround3D employs frozen CLIP ViT-L/14 (Radford et al., 2021) and VGGT (Wang et al., 2025a) as the semantic and geometric encoders, respectively, with Qwen3-0.6B and Qwen3-4B serving serving as the navigation backbone. After spatial alignment, the five-view features are organized into $8 \times 8$ grids, and SGVC retains $K = 16$ tokens per view, resulting in 80 visual tokens in total. UCDG then augments these tokens with UAV-centric directional information and further models their interactions using a Transformer. During training, CLIP and VGGT remain frozen, while Qwen3 is fine-tuned using LoRA with a rank of 64. The model is trained for two epochs on 427,933 frame-level samples with an initial learning rate of $1 \times 10^{-4}$ under a cosine decay schedule. We use bfloat16 precision and ZeRO-2 optimization on four NVIDIA GeForce RTX 5090 GPUs, with an effective batch size of 64. Further implementation details are in the Appendix.

## 4.2 MAIN RESULTS

Table 1 shows that NavGround3D consistently outperforms its corresponding Base models at both the 0.6B and 4B scales, demonstrating the effectiveness of the proposed method across different model capacities. Notably, the 0.6B variant achieves competitive performance relative to SpatialFly, reducing NE from 87.82 m to 81.75 m and improving OSR from 44.41 to 51.05 on Test-Unseen. Scaling the backbone to 4B further yields the best overall performance on both Test-Seen and Test-Unseen. Compared with SpatialFly, NavGround3D (4B) reduces Test-Unseen NE by 9.76 m and improves OSR by 8.88 percentage points. Moreover, compared with their corresponding Base models, the 0.6B and 4B variants improve Test-Unseen OSR by 25.64 and 22.84 percentage points, respectively. These gains indicate that NavGround3D is particularly effective at reaching the correct target region in unseen environments. This is consistent with our design: SGVC distills redundant

semantic-geometric information, while UCDG provides unified UAV-centric directional grounding, together improving spatial understanding and navigation performance.

Table 2: Performance comparison on the seen and unseen environments of UAV-on. All metrics are percentages, higher values indicate better performance.

| 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 |
| Navid Zhang et al. (2024) | 14.17 | 30.08 | 7.19 | 10.70 | 28.51 | 5.99 |
| OpenFly Gao et al. (2026) | 12.29 | 26.20 | 6.83 | 11.82 | 25.71 | 5.64 |
| OctMem-Agent Zhou et al. (2026) | 22.72 | 30.48 | 8.40 | 17.57 | 28.59 | 5.15 |
| **NavGround3D (Ours, 0.6B)** | **23.78** | **34.46** | **9.63** | **19.67** | **33.06** | **7.51** |

Table 2 compares NavGround3D with existing methods on the UAV-ON benchmark. NavGround3D achieves the best performance across all metrics in both seen and unseen environments. In particular, OSR reaches 34.46 on the Seen split and 33.06 on the Unseen split, exceeding prior methods by clear margins. This indicates that NavGround3D is more effective at guiding the UAV into the target region during navigation, reflecting stronger target localization and search capability under both familiar and unseen environments. Meanwhile, the improvements in SR and SPL further show that these gains are accompanied by more successful navigation.

### 4.3 Ablation Studies

Table 3: Overall module ablations.

| VGGT | SGVC | UCDG | NE↓ | SR↑ | OSR↑ | SPL↑ |
|---|---|---|---|---|---|---|
| × | × | × | 99.65 | 22.25 | 53.78 | 21.80 |
| ✓ | × | × | 83.46 | 27.53 | 58.44 | 23.17 |
| ✓ | ✓ | × | 73.13 | 32.12 | 63.14 | 27.76 |
| ✓ | × | ✓ | 74.42 | 31.50 | 62.50 | 26.83 |
| ✓ | ✓ | ✓ | **68.12** | **35.45** | **66.51** | **30.24** |

Table 4: SGVC design ablations.

| SGVC variant | NE↓ | SR↑ | OSR↑ | SPL↑ |
|---|---|---|---|---|
| **Full (joint)** | **68.12** | **35.45** | **66.51** | **30.24** |
| Semantic-only | 71.62 | 33.14 | 64.11 | 28.87 |
| Geometry-only | 72.65 | 33.26 | 65.43 | 28.19 |
| w/o context features | 82.94 | 30.26 | 59.32 | 24.63 |

**Ablation setup.** All ablation experiments are conducted with the 0.6B backbone on the TravelUAV Test-Seen split under a fixed evaluation protocol. All compared variants use identical training and evaluation settings and differ only in the component under study.

**Overall contributions of SGVC and UCDG.** Table 3 evaluates the contributions of VGGT, SGVC, and UCDG. For a fair comparison, all configurations maintain the same visual token budget, and average pooling is used when SGVC is disabled. Introducing VGGT improves SR from 22.25 to 27.53 and SPL from 21.80 to 23.17, while reducing NE from 99.65 to 83.46, indicating that geometric information provides effective spatial cues for navigation. Building on VGGT, adding SGVC further improves SR and SPL to 32.12 and 27.76, respectively, while reducing NE to 73.13. This suggests that distilling dense visual information through semantic-guided anchor selection and semantic-geometric aggregation helps retain navigation-relevant information while reducing interference from redundant representations. Adding UCDG independently yields an SR of 31.50, an SPL of 26.83, and an NE of 74.42, indicating that grounding directional information from different views in a unified UAV-centric reference facilitates multi-view spatial relationship modeling. When both SGVC and UCDG are enabled, the model achieves the best overall performance, reaching 35.45 SR, 66.51 OSR, 30.24 SPL, and 68.12 NE. These results demonstrate the complementary benefits of visual information distillation in SGVC and unified directional modeling in UCDG, which together improve navigation representation learning.

**Design choices within SGVC.** Table 4 evaluates the key designs of SGVC under the full UCDG setting. Replacing the joint semantic-geometric assignment with semantic-only or geometry-only

similarity reduces SR from 35.45 to 33.14 and 33.26, and SPL from 30.24 to 28.87 and 28.19, respectively. This suggests that semantic similarity and geometric consistency provide complementary cues for token assignment. Removing non-anchor aggregation leads to a larger performance degradation, increasing NE from 68.12 to 82.94 while reducing SR, OSR, and SPL to 30.26, 59.32, and 24.63, respectively. This indicates that aggregating non-anchor information into representative anchors helps retain navigation-relevant information during visual consolidation. Overall, joint semantic-geometric assignment and non-anchor aggregation are both important to the effectiveness of SGVC.

Table 5: UCDG design ablations.

| UCDG variant | NE↓ | SR↑ | OSR↑ | SPL↑ |
|---|---|---|---|---|
| **Full** | **68.12** | **35.45** | **66.51** | **30.24** |
| w/o Direction | 71.87 | 32.44 | 63.91 | 26.28 |
| w/o Transformer | 74.62 | 33.76 | 61.59 | 26.73 |
| w/o mask | 72.65 | 34.17 | 64.84 | 27.51 |

Table 6: Top-$K$ anchor-budget ablation.

| $K$ | Tokens ($5K$) | NE↓ | SR↑ | OSR↑ | SPL↑ |
|---|---|---|---|---|---|
| 32 | 160 | 67.73 | 32.75 | 63.25 | 28.03 |
| 24 | 120 | 70.54 | 34.03 | **67.52** | 28.86 |
| **16** | **80** | 68.12 | **35.45** | 66.51 | **30.24** |
| 12 | 60 | **66.37** | 32.49 | 63.52 | 27.39 |
| 8 | 40 | 72.47 | 28.96 | 60.41 | 24.71 |

**Design choices within UCDG.** Table 5 evaluates the key design choices of UCDG while keeping SGVC fixed. Removing the directional encoding reduces SR from 35.45 to 32.44 and SPL from 30.24 to 26.28, while increasing NE from 68.12 to 71.87, indicating that explicit UAV-centric directional information provides effective spatial cues for navigation. Removing the Transformer also degrades performance, increasing NE to 74.62 and reducing OSR and SPL to 61.59 and 26.73, respectively, suggesting that token interaction is important for modeling spatial relationships among multi-view representations. Removing the cross-view attention mask leads to a clear performance drop, suggesting that structured cross-view token interactions help suppress irrelevant information mixing and improve spatial relation modeling across views. Overall, these results validate the effectiveness of the directional encoding, Transformer, and cross-view attention mask in UCDG.

**Effect of the Top-$K$ anchor budget.** Table 6 evaluates the effect of the anchor budget $K$ while keeping all other settings unchanged. Increasing $K$ beyond 16 does not further improve SR or SPL, suggesting that retaining more visual tokens does not necessarily benefit navigation performance and may introduce unnecessary local information. In contrast, an overly small $K$ leads to clear performance degradation; at $K = 8$, SR and SPL decrease to 28.96 and 24.71, respectively. With $K = 16$, the model retains 80 visual tokens and achieves the highest SR of 35.45 and SPL of 30.24. Although $K = 12$ yields the lowest NE and $K = 24$ achieves the highest OSR, $K = 16$ provides the best overall balance across the evaluation metrics and is therefore adopted as the default setting.

## 5 Conclusion

We propose NavGround3D, a compact multi-view representation framework for UAV-VLN. To address the redundancy of dense semantic–geometric features and the difficulty of consistently modeling directional relationships across discrete camera views, SGVC selectively retains and aggregates visual information based on semantic–geometric consistency, while UCDG establishes continuous cross-view directional representations in the UAV body frame. Together, these modules construct compact visual representations that preserve complementary semantic, geometric, and directional cues for local 3D navigation decisions. Extensive experiments on TravelUAV and UAV-ON show that NavGround3D consistently improves navigation success and path efficiency, demonstrating the effectiveness of the proposed semantic–geometric consolidation and UAV-centric directional grounding strategies.

## Statement on the Use of AI Assistants

An AI assistant was used to inspect the implementation, organize the manuscript, and improve exposition. Numerical claims were checked against source code, training metadata, complete evaluation

logs, or cited primary papers. The authors remain responsible for the scientific claims, citations, and final text.

## REFERENCES

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton Van Den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *2018 IEEE/CVF conference on computer vision and pattern recognition*. IEEE, 2018.

Cewen Deng. Aerovis3r: Geometry-aware vision–language models for 3d reasoning over uav landmark videos. In *Proceedings of the SIGGRAPH Asia 2025 Posters*. 2025.

Yan Deng and Fei Xu. Dreamfly: Causal memory and receding-horizon diffusion planning for aerial vision-language navigation. *arXiv preprint arXiv:2608.12308*, 2026.

Chengjie Fan, Cong Pan, Zijian Liu, Ningzhong Liu, and Jie Qin. HTNav: A hybrid navigation framework with tiered structure for urban aerial vision-and-language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026a.

Zhiwen Fan, Jian Zhang, Renjie Li, Junge Zhang, Runjin Chen, Hezhen Hu, Kevin Wang, Peihao Wang, Huaizhi Qu, Shijie Zhou, et al. Vlm-3r: Vision-language models augmented with instruction-aligned 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026b.

Hongzhi Gao, Dekyi Dekyi, and Metok Metok. Efficient unmanned aerial vehicle inspection and management of transmission lines in modern electric power enterprises. *Energy Informatics*, 2025.

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

Yunpeng Gao, Chenhui Li, Zhongrui You, Junli Liu, Zhen Li, Pengan Chen, Qizhi Chen, Zhonghan Tang, Liansheng Wang, Penghui Yang, 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. OpenFly: A comprehensive platform for aerial vision-language navigation. In *International Conference on Learning Representations*, 2026.

M Garba. Edge-enabled autonomous uav systems for disaster management: Integrating social intelligence. *Communications*, 2026.

Miguel Gil-Castilla, Alvaro Ramiro Poma, Alvaro Caballero, and Anibal Ollero. A cooperative multi-uav framework for bridge inspection. In *2025 International Conference on Unmanned Aircraft Systems (ICUAS)*, 2025.

S Harshini, D Harshitha, N Harshitha, et al. First responder emergency uavs: A comprehensive survey. In *2025 IEEE Conference on Engineering Informatics (ICEI)*, 2025.

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

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. *arXiv preprint arXiv:2603.21046*, 2026.

Burak Kaleci, Gulin Elibol Secil, Sezgin Secil, Zühal Kartal, and Metin Ozkan. A novel coverage path planning method based on shrink-wrapping technique for autonomous inspection of complex structures using unmanned aerial vehicle. *Robotics and Computer-Integrated Manufacturing*, 2026.

Jungdae Lee, Taiki Miyanishi, Shuhei Kurita, Koya Sakamoto, Daichi Azuma, Yutaka Matsuo, and Nakamasa Inoue. CityNav: A large-scale dataset for real-world aerial navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

Jinhan Li, Xijie Huang, Zhaoqi Wang, Yijin Wang, Weiqi Ge, Qiyi He, Mo Zhu, Fei Gao, Yuze Wu, and Xin Zhou. Flymirage: A fully automated generation pipeline for diverse and scalable uav flight data via generative world model. *arXiv preprint arXiv:2605.19600*, 2026.

Shubo Liu, Hongsheng Zhang, Yuankai Qi, Peng Wang, Yaning Zhang, and Qi Wu. AerialVLN: Vision-and-language navigation for uavs. *arXiv preprint arXiv:2308.06735*, 2023.

Tianbo Pan, Xingyi Yang, and Xinchao Wang. Merge3d: efficient 3d multimodal llms via joint 2d-3d token merging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

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. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Dian Shao, Zhengzheng Xu, Peiyang Wang, Like Liu, Yule Wang, Jieqi Shi, and Jing Huo. Finecognav: Integrating fine-grained cognitive modules for zero-shot multimodal uav navigation. *arXiv preprint arXiv:2604.16298*, 2026.

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*, 2025.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. In *CVPR*, 2025a.

Weitian Wang, Lukas Meiner, Rai Shubham, Cecilia De La Parra, and Akash Kumar. Httm: Head-wise temporal token merging for faster vggt. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Xiangyu Wang, Donglin Yang, Ziqin Wang, Hohin Kwan, Jinyu Chen, Wenjun Wu, Hongsheng Li, Yue Liao, and Si Liu. Towards realistic UAV vision-language navigation: Platform, benchmark, and methodology. In *International Conference on Learning Representations*, 2025b.

Xiangyu Wang, Donglin Yang, Ziqin Wang, Hohin Kwan, Jinyu Chen, Wenjun Wu, Hongsheng Li, Yue Liao, and Si Liu. Towards realistic UAV vision-language navigation: Platform, benchmark, and methodology. In *ICLR*, 2025c.

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*, 2025.

Jiahao Yang, Zihan Wang, Xiangyang Li, Xing Zhu, Yujun Shen, Yinghao Xu, and Shuqiang Jiang. GA-VLN: Geometry-aware BEV representation for efficient vision-language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

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

Shuang Zeng, Dekang Qi, Xinyuan Chang, Feng Xiong, Shichao Xie, Xiaolong Wu, Shiyi Liang, Mu Xu, and Xing Wei. JanusVLN: Decoupling semantics and spatiality with dual implicit memory for vision-language navigation. In *The Fourteenth International Conference on Learning Representations*, 2026.

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

Jian Zhang, Shijie Zhou, Bangya Liu, Achuta Kadambi, and Zhiwen Fan. SpatialStack: Layered geometry-language fusion for 3d vlm spatial reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026b.

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*, 2026c.

Peiyi Zhang, Yuhang Bai, Jiashen Du, Weitai Yu, Weihao Lin, Bing Wang, Yunlong Qian, Bin Tu, Conghui Liu, Yinlong Chen, et al. NaVid: Video-based vlm plans the next step for vision-and-language navigation. *arXiv preprint arXiv:2402.15852*, 2024.

Baining Zhao, Jiacheng Xu, Weicheng Feng, Xin Zhang, Zhaolu Wang, Haoyang Wang, Shilong Ji, Ziyou Wang, Jianjie Fang, Zhiheng Zheng, et al. Worldvln: Autoregressive world action model for aerial vision-language navigation. *arXiv preprint arXiv:2605.15964*, 2026.

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. *arXiv preprint arXiv:2606.06147*, 2026.

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

Deyi Zhu, Haoyu Fan, Yinan Zhu, Weichen Zhang, Shilin Ma, Xinlei Chen, and Yansong Tang. Uncertainty-aware world model for aerial image-goal navigation. *arXiv preprint arXiv:2608.05597*, 2026.

## A Supplementary Coordinate System and Transformations.

We follow the five-camera configuration in the AirSim environment and use the UAV body frame as the shared ego-centric reference, where the $x_b$-, $y_b$-, and $z_b$-axes point forward, rightward, and downward, respectively. The front, left, right, rear, and downward-facing cameras are mounted at $(1, 0, 0)$, $(0, -1, 0)$, $(0, 1, 0)$, $(-1, 0, 0)$, and $(0, 0, 0)$ meters in the UAV body frame. The four horizontal cameras have yaw angles of $0°$, $-90°$, $90°$, and $180°$, respectively, while the downward-facing camera has a pitch angle of $-90°$. All other mounting angles are set to zero.

Consistent with our implementation, each camera-local coordinate frame is defined with the $x_c$-axis along the optical axis, the $y_c$-axis pointing rightward in the image, and the $z_c$-axis pointing downward in the image. Since all cameras are rigidly mounted, their orientations relative to the UAV remain fixed during navigation. Accordingly, view-specific local directions are transformed into the shared UAV-centric frame using the following fixed rotation matrices:

$$\mathbf{R}_F = \mathbf{I}, \qquad \mathbf{R}_L = \begin{bmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \qquad \mathbf{R}_R = \begin{bmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \tag{10}$$

$$\mathbf{R}_B = \begin{bmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \qquad \mathbf{R}_D = \begin{bmatrix} 0 & 0 & -1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}. \tag{11}$$

These rotations map the principal viewing directions of the five cameras to the forward, leftward, rightward, backward, and downward directions of the UAV body frame, respectively.

All five cameras share the same imaging parameters, with an image resolution of $256 \times 256$ and a horizontal field of view of $90°$. Accordingly, we use

$$f_x = f_y = 128, \qquad c_x = c_y = 128. \tag{12}$$

For a $G \times G$ feature grid, the center of the patch at row $i$ and column $j$ is given by

$$u_j = \frac{256(j + \frac{1}{2})}{G}, \qquad v_i = \frac{256(i + \frac{1}{2})}{G}. \tag{13}$$

Under the above camera-coordinate convention, the viewing ray associated with this patch is expressed in the local camera frame as

$$\mathbf{d}^{c}_{v,i,j} = \mathrm{Normalize}\left(\begin{bmatrix} 1 \\ (u_j - c_x)/f_x \\ (v_i - c_y)/f_y \end{bmatrix}\right). \tag{14}$$

It is then transformed into the UAV body frame by

$$\mathbf{d}^{b}_{v,i,j} = \mathrm{Normalize}\left(\mathbf{R}_v \mathbf{d}^{c}_{v,i,j}\right). \tag{15}$$

This transformation expresses the local viewing directions from all five camera views in a common UAV-centric coordinate frame, providing a consistent spatial reference for subsequent directional encoding and multi-view spatial modeling. Since our representation encodes viewing directions rather than absolute 3D positions, the camera translations are used only to describe the physical AirSim configuration and do not participate in the directional transformation.

## B Dataset and metrics

**TravelUAV.** TravelUAV is a continuous 3D benchmark for UAV vision-and-language navigation in high-fidelity outdoor simulation environments. It contains 12,149 human-operated trajectories covering 89 target categories across diverse scenes, including urban areas, roads, grasslands, and snowy regions. Each trajectory provides a natural-language instruction, UAV state information, waypoint annotations, and synchronized multi-view RGB observations from five cameras facing forward, left, right, rear, and downward. The dataset is divided into training, Test-Seen, and Test-Unseen splits. The training set contains 9,152 trajectories across 20 environments, while Test-Seen contains 1,410 trajectories from observed environments. Test-Unseen contains 1,587 trajectories involving unseen environments and target categories, and is used to evaluate generalization. The training split covers 76 seen target categories, whereas the full benchmark contains 89 categories. Test trajectories are further categorized as Easy or Hard according to reference-path length, with 250 m used as the threshold; target distances range approximately from 50 to 400 m.

**UAV-ON.** UAV-ON is an open-world aerial Object Goal Navigation benchmark in which the agent receives a high-level semantic description of the target and must autonomously explore the environment to locate it through onboard visual perception. The benchmark contains 14 high-fidelity outdoor environments constructed with Unreal Engine and AirSim, spanning villages, towns, cities, parks, roads, forests, mountains, snowy regions, and water areas. It includes 1,270 target instances distributed over approximately nine million square units of horizontal area. Each target is associated with an instance-level semantic description containing its category, approximate size, and visual attributes. UAV-ON provides 10,000 training episodes across 10 environments and 1,000 test episodes. The test set includes both seen environments and four held-out environments, as well as seen and unseen target categories, enabling evaluation of generalization across both environments and target semantics.

**Evaluation Metrics.** We adopt Navigation Error (NE), Success Rate (SR), Oracle Success Rate (OSR), and Success weighted by Path Length (SPL) as evaluation metrics for navigation performance, with the success threshold set to 20 m. NE measures the final distance between the UAV and the target, where lower values indicate closer arrival. SR denotes the proportion of episodes in which the UAV successfully reaches the target region, with an episode considered successful when the final distance to the target is no greater than 20 m. OSR measures whether the UAV enters the target region at any point during navigation. Unlike SR, which only considers the final stopping position, OSR reflects whether the trajectory reaches the success region during execution. SPL further evaluates navigation efficiency by weighting successful episodes according to the ratio between the reference shortest path length and the actual navigation path length. Higher SPL values indicate more efficient navigation.

## C Implementation and Training Details

NavGround3D employs frozen CLIP ViT-L/14 (Radford et al., 2021) and VGGT (Wang et al., 2025a) as semantic and geometry-aware visual encoders, respectively. After spatially aligning the

CLIP and VGGT features, features from each of the five camera views are organized into an $8 \times 8$ spatial grid. For anchor selection, semantic saliency is derived from the last-layer patch-to-patch self-attention of CLIP. The attention received by each patch is aggregated over all patch queries and averaged across attention heads, followed by average pooling to match the $8 \times 8$ spatial grid. The top-$K$ locations are then selected as semantic anchors, with $K = 16$ per view, resulting in 80 visual tokens across all five views. The scaling factors for semantic and geometric similarity in SGVC are set to $\tau_s = \tau_g = 1.0$. UCDG employs a four-layer Transformer to model interactions among direction-aware multi-view representations. For fair ablation comparisons, all settings use the same number of visual tokens, with average pooling applied whenever SGVC is disabled.

We adopt Qwen3-0.6B and Qwen3-4B as navigation backbones to evaluate NavGround3D under different model scales. Both variants share the same visual encoders, SGVC, UCDG, and waypoint prediction architecture, with only the Qwen3 backbone replaced. Qwen3-4B provides a larger model capacity than Qwen3-0.6B, allowing us to examine the effect of backbone scale on the proposed framework. Unless otherwise specified, Qwen3-0.6B is used as the default backbone.

During training, CLIP and VGGT remain frozen, while Qwen3 is parameter-efficiently fine-tuned using LoRA with a rank of 64. The model is trained for two epochs on 427,933 frame-level samples with an initial learning rate of $1 \times 10^{-4}$ under a cosine decay schedule. We use bfloat16 precision and ZeRO-2 optimization on four NVIDIA GeForce RTX 5090 GPUs. The per-GPU batch size is set to 8 with two gradient accumulation steps, resulting in an effective batch size of 64.

# D Experiments

## D.1 Overall Results

Table 7: Comparison on the TravelUAV Test-Seen split and its easy and hard route subsets. NE is measured in meters. Lower NE and higher SR, OSR, and SPL indicate better performance. The best results are shown in bold.

| Method | Full | | | | Easy | | | | Hard | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | NE↓ | SR↑ | OSR↑ | SPL↑ | NE↓ | SR↑ | OSR↑ | SPL↑ | NE↓ | SR↑ | OSR↑ | SPL↑ |
| 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. (2018) | 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., 2025c) | 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., 2025c) | 98.66 | 17.45 | 48.87 | 15.76 | 66.40 | 20.26 | 51.23 | 18.10 | 138.04 | 14.02 | 45.98 | 12.90 |
| NavFoM Zhang et al. (2026c) | 93.05 | 29.17 | 49.24 | 25.03 | 58.98 | 32.91 | 53.16 | 27.87 | 143.83 | 23.58 | 43.40 | 20.80 |
| 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 |
| Baseline (Ours, 0.6B) | 99.65 | 22.25 | 53.78 | 21.80 | 66.20 | 25.00 | 50.91 | 20.54 | 140.53 | 18.89 | 42.78 | 16.77 |
| **NavGround3D (Ours, 0.6B)** | 68.12 | 35.45 | 66.51 | 30.24 | 42.54 | 38.64 | 70.00 | 33.44 | 99.88 | 32.43 | 61.11 | 26.33 |
| Baseline (Ours, 4B) | 71.22 | 25.75 | 55.75 | 23.15 | 45.99 | 28.64 | 61.36 | 25.13 | 102.06 | 22.22 | 48.89 | 20.74 |
| **NavGround3D (Ours, 4B)** | **57.13** | **39.76** | **69.42** | **34.51** | 42.54 | **41.13** | 70.69 | **34.76** | **72.81** | **38.77** | **68.62** | **33.95** |

Table D.1 reports detailed results on the TravelUAV Test-Seen split and its Easy and Hard subsets. Overall, NavGround3D consistently outperforms the corresponding Baseline under both Qwen3 model scales, indicating that the proposed semantic-geometric consolidation and directional modeling effectively improve navigation performance in seen environments. With Qwen3-0.6B, NavGround3D reduces NE on the Full split from 99.65 to 68.12, while increasing SR, OSR, and SPL from 22.25, 53.78, and 21.80 to 35.45, 66.51, and 30.24, respectively. Similar gains are observed on both Easy and Hard subsets. In particular, on the Hard subset, NE decreases from 140.53 to 99.88, while SR and SPL increase to 32.43 and 26.33, respectively, showing stronger gains on longer and more challenging routes. Using the larger Qwen3-4B backbone further improves the overall performance of NavGround3D. Compared with the corresponding baseline, NE on the Full split decreases from 71.22 to 57.13, while SR, OSR, and SPL increase to 39.76, 69.42, and 34.51, respectively. The improvement is particularly pronounced on the Hard subset, where NE decreases from 102.06 to 72.81, SR increases from 22.22 to 38.77, OSR from 48.89 to 68.62, and SPL from 20.74 to 33.95. These results indicate that the larger language-model capacity further strengthens the benefits of NavGround3D in complex navigation scenarios.

Compared with existing methods, NavGround3D-4B achieves the lowest NE and the highest SR, OSR, and SPL on the Full split, outperforming recent approaches such as SpatialFly. It also obtains the best results across all four metrics on the Hard subset, demonstrating strong performance on challenging routes. On the Easy subset, although NE and OSR are not the best, NavGround3D-4B achieves the highest SR and SPL, indicating clear advantages in navigation success and path efficiency. Overall, NavGround3D delivers consistent improvements across model scales and route difficulties, with particularly strong gains on the more challenging Hard subset.

Table 8: Comparison on the TravelUAV Test-Unseen split and its easy and hard route subsets. NE is measured in meters. Lower NE and higher SR, OSR, and SPL indicate better performance. The best results are shown in 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. (2018) | 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., 2025c) | 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. (2026c) | 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 |
| Baseline (Ours, 0.6B) | 110.47 | 12.96 | 25.41 | 11.63 | 80.49 | 15.35 | 31.06 | 13.82 | 136.19 | 10.67 | 24.85 | 9.72 |
| **NavGround3D (Ours, 0.6B)** | 81.75 | 22.81 | 51.05 | 20.15 | 62.10 | 20.65 | 51.17 | 17.75 | 100.57 | 24.88 | 51.00 | 22.54 |
| Baseline (Ours, 4B) | 103.81 | 15.17 | 30.45 | 13.76 | 74.69 | 17.66 | 35.72 | 14.97 | 130.44 | 12.63 | 26.46 | 11.37 |
| **NavGround3D (Ours, 4B)** | **78.06** | **26.43** | **53.29** | **22.52** | **60.33** | **27.64** | **53.09** | **22.41** | **97.95** | **25.73** | **52.94** | **24.47** |

Table 8 reports the results on the TravelUAV Test-Unseen split and its Easy and Hard subsets. Overall, NavGround3D consistently outperforms the corresponding Baseline under both Qwen3-0.6B and Qwen3-4B, demonstrating strong generalization to unseen environments and target categories. With Qwen3-0.6B, NE on the Full split decreases from 110.47 to 81.75, while SR, OSR, and SPL increase to 22.81, 51.05, and 20.15, respectively. On the Hard subset, NE decreases from 136.19 to 100.57 and SPL improves from 9.72 to 22.54, indicating that the proposed method remains effective on more challenging unseen routes. Using Qwen3-4B further improves performance. On the Full split, NavGround3D-4B achieves an NE of 78.06, an SR of 26.43, an OSR of 53.29, and an SPL of 22.52. The model also maintains strong performance on both Easy and Hard subsets, with the SPL on Hard reaching 24.47. Compared with existing methods, NavGround3D-4B achieves the best overall results across the Full, Easy, and Hard settings, with particularly strong gains in SR, OSR, and SPL. Overall, NavGround3D exhibits robust generalization across unseen environments and target categories while maintaining strong navigation performance on challenging routes.

**Effect of anchor selection strategies.** To evaluate the influence of different anchor selection strategies on navigation performance, we compare five alternatives. Random sampling selects $K$ locations without replacement from the candidate patches. Uniform grid sampling uniformly selects locations on the 2D patch grid to preserve spatial coverage. CLIP feature FPS performs farthest point sampling on normalized CLIP patch features to enhance feature diversity. Text grounding selects the Top-$K$ regions according to the semantic similarity between each patch and the task instruction. Patch attention derives anchor saliency from CLIP's last-layer patch-to-patch self-attention, where the attention received by each patch is aggregated over all patch queries and averaged across attention heads, followed by average pooling before Top-$K$ selection. The results show that Patch attention achieves the best overall performance, with 68.12 NE,

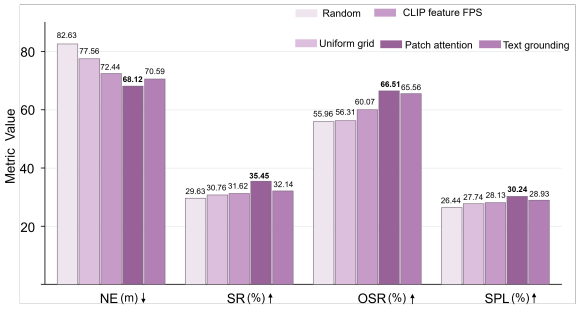


Figure 3: Effect of anchor selection strategies on navigation performance. Lower NE and higher SR, OSR, and SPL indicate better performance.

35.45 SR, 66.51 OSR, and 30.24 SPL. Random sampling lacks effective semantic or structural guidance, while Uniform grid mainly emphasizes spatial coverage and cannot distinguish the navigation relevance of different regions. CLIP feature FPS promotes feature diversity but does not explicitly account for navigation relevance. Text grounding effectively preserves instruction-related regions; however, UAV navigation instructions are typically sparse and mainly describe the final target and a few salient landmarks, potentially overlooking contextual regions that are not explicitly mentioned but remain important for navigation decisions. In contrast, Patch attention evaluates patch importance through scene-level visual interactions, enabling a better balance between target-related cues and broader environmental context.

### D.2 Qualitative Visualization




Figure 4: Qualitative navigation comparison in an urban environment from TravelUAV.

Figure 3 presents a challenging urban navigation scenario. Starting from the same initial observation, the baseline successfully performs a turning action at point B, but its subsequent trajectory gradually deviates from the reference path due to the interference of redundant visual structures, eventually moving toward the building region and failing the navigation task. In contrast, NavGround3D maintains consistent motion alignment with the reference trajectory after turning at point B and performs accurate directional adjustments at critical decision points (e.g., point D), successfully reaching the target vicinity with a final navigation error of 5.65 m. This example demonstrates that reliable UAV navigation in complex environments requires not only reducing semantic-geometric redundancy but also organizing multi-view visual cues within a unified UAV-centric directional space to support stable spatial reasoning.

**Limitations and Future Directions.** Although NavGround3D achieves effective semantic–geometric consolidation and UAV-centric directional modeling, several directions remain worth further exploration. First, the current directional modeling assumes calibrated camera parameters and reliable UAV poses; future work may further investigate its robustness to calibration errors and pose perturbations. Second, SGVC adopts a fixed token budget for each view to maintain a simple and consistent design. Future work could explore adaptive cross-view token allocation under different navigation scenarios to make more effective use of multi-view information. Finally, the current framework mainly focuses on information compression and representation optimization after visual encoding. Future research could further investigate more efficient token selection during visual encoding to reduce the overall computational cost.