Learning Fine-Grained Alignment for Aerial Vision-Dialog Navigation

Yifei Su[1,2], Dong An[3], Kehan Chen[1,2], Weichen Yu[4], Baiyang Ning[1,2], Yonggen Ling[5], Yan Huang[1,2 †], Liang Wang[1,2],

[1]School of Artificial Intelligence, University of Chinese Academy of Sciences
[2]MAIS, Institute of Automation of Chinese Academy of Sciences
[3]Mohamed bin Zayed University of Artificial Intelligence
[4]Electrical and Computer Engineering Department, Carnegie Mellon University
[5]Robotics X, Tencent, Shenzhen, China
suyifei2022@ia.ac.cn

Abstract

Aerial Vision-Dialog Navigation (AVDN) is a new task that requires drones to navigate to a target location based on human-robot dialog history. This paper focuses on the critical fine-grained cross-modal alignment problem in AVDN, requiring the drone to align language entities with visual landmarks in top-down views. To achieve this, we first construct a Fine-Grained AVDN (FG-AVDN) dataset via a semi-automatic annotation pipeline, providing diverse multi-modal annotations at the entity-landmark level. Based on this, a novel Fine-grained Entity-Landmark Alignment (FELA) method is proposed to learn the cross-modal alignment explicitly. Concretely, FELA first boosts the drone's visual understanding with a precise semantic grid representation, which captures the environmental semantics and spatial structure simultaneously. Subsequently, to learn the entity-landmark alignment, we devise cross-modal auxiliary tasks from three perspectives, including grounding, captioning, and contrastive learning. Extensive experiments demonstrate that our explicit entity-landmark alignment learning is beneficial for AVDN. As a result, FELA achieves leading performance with 3.2% SR and 4.9% GP improvements over prior arts.

Code — https://github.com/yifeisu/FELA

1 Introduction

Language-guided navigation is a fundamental yet challenging problem for autonomous robot task-completion via communication with humans. Recent years have witnessed notable progress towards this goal, marked by the introduction of various tasks (Anderson et al. 2018; Qi et al. 2020; Thomason et al. 2020; Chen et al. 2019) and navigation methods (Qiao et al. 2022; Zhao et al. 2022; Wang et al. 2023c; He et al. 2024; Wang et al. 2023d,a; An et al. 2024). However, a predominant focus on ground-based agents has overshadowed a crucial application area for robotic systems - drones. Recently, this gap has been bridged by the introduction of the Aerial Vision-Dialog Navigation (AVDN) dataset and the corresponding Aerial Navigation from Dialog History (ANDH) task for drones (Fan et al. 2023a; Liu et al. 2023), which opens up new application chances such as food delivery and wilderness search and rescue. In ANDH, a drone agent is required to navigate to a destination precisely, via the dialog history with humans in aerial environments.

Accurate cross-modal alignment is crucial for the success of language-guided navigation (Moudgil et al. 2021; Wang et al. 2019; Qi et al. 2021; Wang et al. 2022), while the ANDH task poses two unique challenges in this aspect. First, landmarks in top-down views exhibit geometric diversity, i.e., a wider range of scales and aspect ratios, posing challenges to the perception of small or narrow objects (e.g., the densely arranged small yellow containers in Figure 1 (a). Second, top-down views tend to be broader and contain more landmarks. Thus, the redundant landmarks are likely to distract the agent from grounding the target entity mentioned in the dialog. To address these challenges, researchers have proposed various methods to guide the agent to focus on potential areas mentioned in the dialog. For instance, Fan *et al.* (Fan et al. 2023a) supervise the drone's visual perception with human attention masks, while Su *et al.* (Su et al. 2023) enhace the stop policy with a cross-modal target area grounding task.

Despite the progress, existing methods struggle to handle fine-grained cross-modal alignment in ANDH. The reasons mainly lie in the lack of entity-landmark alignment supervision and the coarse visual representation. First, both human attention (Fan et al. 2023a) and target area (Su et al. 2023) supervisions are relatively coarse and lack intermediate fine entity-landmark alignment supervision during navigation. As a result, the drone is forced to infer involved all landmarks implicitly from the whole dialog history. This scheme has been shown sub-optimal in ground-based language-guided navigation (Cui et al. 2023; Qi et al. 2021), while it can be less effective in ANDH due to the landmark redundancy in top-down views. Second, condensed feature vectors are widely adopted by existing methods to represent the drone's top-down views, which suppress the representation of geometrically rich landmarks and environmental spatial layouts (Chen, Yang, and Chen 2023; Chen et al. 2024). This further challenges the entity-landmark alignment when the dialog involves small landmarks or complex spatial relations (Chen et al. 2022a; Hwang et al. 2023).

This paper addresses the above issues by introducing explicit entity-landmark alignment learning for ANDH. To achieve this goal, we first construct a large-scale Fine-Grained AVDN dataset (FG-AVDN), which provides rich

multimodal annotations at the entity-landmark level. FG-AVDN is built via a semi-automatic pipeline, through the combination of modern large language models (Brown et al. 2020) and multimodal foundation models (Muhtar et al. 2024; Kirillov et al. 2023), featuring less human labor and more diverse annotations. The paired annotations are visually depicted in Figure 1 (b) through text fragments and bounding boxes with the same color. Subsequently, we propose a novel Fine-grained Entity-Landmark Aligning (FELA) method to unleash the potential of FG-AVDN. In FELA, the drone's visual perception is boosted with a precise semantic grid representation, capturing rich landmarks and spatial layouts of environments. Built upon this, FELA devises three auxiliary tasks to learn entity-landmark alignment explicitly: 1) Landmark Rotated Bounding box Prediction (LRBP): predicting the compact rotated bounding box of landmarks based on entities. 2) Landmark Semantic Prediction (LSP): describing the landmark given visual image. 3) Entity-Landmark Contrastive Learning (ELCL): aligning the matched entity-landmark pairs in the common feature space. The learned grid representation is aware of entity-landmark alignment, which is then fed into a navigator for action decisions. Extensive experiments demonstrate the effectiveness of our method.

In summary, our contributions are three-fold: 1) Developing a semi-automatic annotation pipeline to construct the first large-scale fine-grained AVDN dataset, providing cross-modal alignments at the entity-landmark level. 2) Proposing the FELA method, incorporating a novel semantic grid representation and enabling fine-grained cross-modal alignment via three auxiliary tasks. 3) We empirically show that explicit entity-landmark alignment learning is beneficial for ANDH, and FELA obtains 3.2% SR and 4.9% GP absolute improvements over prior arts.

2 Related Work

Aerial Vision-Dialog Navigation. In recent years, various language-guided navigation tasks have been proposed and have attracted increasing attention. Early efforts mainly focused on indoor navigation through detailed instructions (Anderson et al. 2018; Ku et al. 2020), while subsequent studies expanded the scope to various instruction forms, such as human-robot dialog (Thomason et al. 2020; Fan et al. 2023b) and concise referring expression (Qi et al. 2020; Zhu et al. 2021). Recently, there has been an emerging trend in introducing this task into aerial scenes, enabling applications like food delivery and terrain exploration. Fan *et al.* (Fan et al. 2023a) propose a challenging ANDH task, requiring autonomous goal-reaching through cross-modal understanding between the dialog and drones' top-down views.

So far, several methods have been proposed to tackle the ANDH task. The pioneering HAA-Transformer (Fan et al. 2023a) adopts an episodic transformer (Pashevich, Schmid, and Sun 2021) for multimodal reasoning and supervises the drone's visual perception with human attention masks. TG-GAT (Su et al. 2023) further introduces a target area prediction task to enhance the stop policy, and devises a graph attention mechanism to capture navigation dependencies. Different from them, this paper addresses the entity-landmark alignment problem in ANDH, through a new fine-grained dataset and a novel method to learn the alignment explicitly.

Fine-Grained Dataset for Navigation. In ground-based navigation tasks (Anderson et al. 2018; Ku et al. 2020), researchers have recognized the importance of fine-grained cross-modal alignment and have proposed various fine-grained datasets. Hong *et al.* (Hong et al. 2020) and He *et al.* (He et al. 2021) automatically or manually decompose long instructions into a series of sub-instructions, providing alignment between sentence fragment and sub-trajectory. Cheng *et al.* (Cheng et al. 2022) further disentangle instructions into landmarks and action-related parts, introducing additional alignment constraints. Cui *et al.* (Cui et al. 2023) recently annotate abundant entity-landmark pairs, offering finer alignment at the word-region level. Chu *et al.* (Chu et al. 2025) annotate bounding boxes in aerial images with orientation and relation descriptions, providing both fine-grained and spatial-aware supervision.

The fine-grained alignment is more crucial in ANDH due to the abundance and geometric diversity of landmarks in top-down views. Inspired by the previous works, this paper makes the first exploration of fine-grained alignment in ANDH. We not only devise a semi-automatic pipeline to build the FG-AVDN dataset but also propose a fine-grained alignment learning method tailored for the ANDH.

3 FG-AVDN Dataset

We aim to enable entity-landmark alignment learning in ANDH. However, the AVDN dataset only provides coarse supervision at the dialog-path level, which lacks paired entity-landmark annotations. To bridge this gap, we construct the first large-scale grounded entity-landmark dataset tailored for ANDH. As shown in Figure 1, we build an efficient semi-automatic annotation pipeline to extract entity-landmark pairs involved in ANDH dialogs. To better generalize to various entities, we also generate extra pseudo entity-landmark pairs with detailed landmark descriptions. The annotation process is presented below.

3.1 Semi-Automatic Entity-Landmark Extraction

In AVDN dataset, a drone agent starts with an initial instruction $\mathbf{I}_0$, and then autonomously navigates to a target $\mathbf{G}_f$ through n rounds of dialogs with humans $\{\mathbf{Q}_i, \mathbf{I}_i\}_{i=1}^{n}$, where $\mathbf{Q}_i$ and $\mathbf{I}_i$ are agent question and human response, respectively. For each dialog round, AVDN provides a ground-truth sub-trajectory $\mathcal{T}_i$ as supervision.

Based on this, we first utilize the exceptional language capability of GPT3.5 (Brown et al. 2020) to extract entity $\{\mathbf{E}_i\}_{i=1}^{m}$ present in each dialog round $\{\mathbf{Q}_i, \mathbf{I}_i\}$. Subsequently, we use the visual comprehension capability of SAM (Kirillov et al. 2023) to generate landmark mask proposals $\{\mathbf{m}_i\}_{i=1}^{k}$ from $\mathbf{V}_t$, and then filter out proposals with small areas and low confidence. Meanwhile, we calculate the minimum rotated bounding box (rbbox) $\mathbf{b}_i = [x, y, w, h, \theta]$ of each mask as extra annotation, denoting the landmark's center, width, height, and rotation angle, respectively. To establish the initial correlation between entities and landmark proposals, we adopt RemoteCLIP (Liu et al. 2024), a VLM

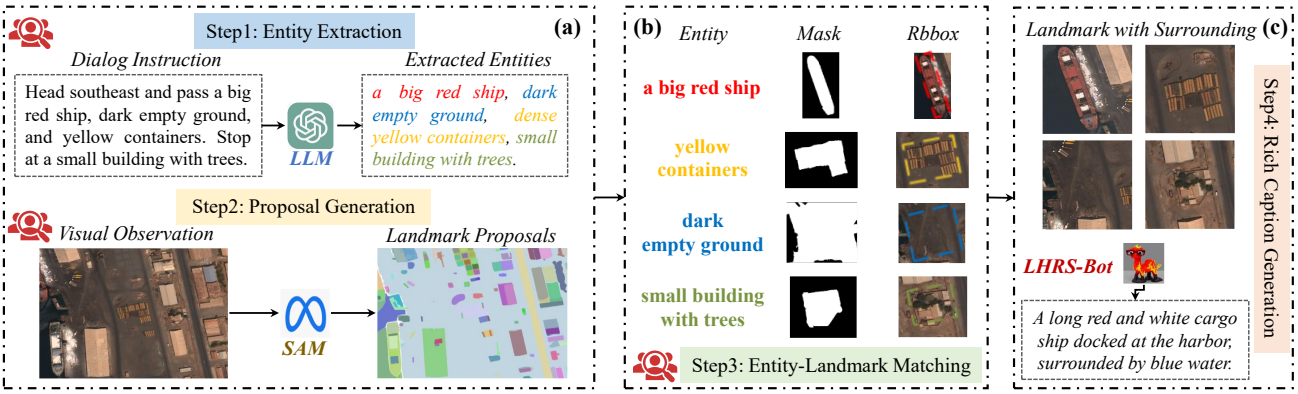


Figure 1: Our semi-automatic pipeline leverages GPT (Brown et al. 2020) to extract entities within dialogs and utilize SAM (Kirillov et al. 2023) to obtain landmark proposals in visual observations. Subsequently, RemoteCLIP (Liu et al. 2024) is adopted to establish an initial correlation between entities and landmarks. For pseudo entity generation, we use LHRS-Bot (Muhtar et al. 2024) to caption landmarks at the region level, obtaining rich descriptions with detailed properties and surrounding information. Finally, we conduct human checks and corrections to ensure the quality, as presented in Appendix 7.

in the remote sensing domain, to select the most matching landmark for each entity. Finally, to ensure the quality of FG-AVDN, we perform manual checks to correct partially mismatched landmark-entity pairs and adjust the rbbox of imperfect proposals generated by SAM. More details are presented in Appendix 7.

3.2 Pseudo Entity-Landmark Generation

In practice, most of the entity phrases extracted in the previous process are monotonous and lack detailed descriptions (e.g., "a building", "the road", and "the destination"), which could hinder the drone's understanding of open-vocabulary instructions. Thus, this section proposes to synthesize extra detailed descriptions for the extracted landmarks as shown in Figure 1 (c).

Specifically, for each extracted landmark mask $\mathbf{m}_j$, we crop out its region image $\mathbf{C}_t$ from visual observation $\mathbf{V}_t$, containing both the landmark and surrounding contextual details. Subsequently, we feed $\mathbf{C}_t$ into LHRS-Bot (Muhtar et al. 2024) and design a simple yet effective prompt template: "*Describe the landmark in one sentence, including details such as color, geometry, semantic details, and surroundings as modifiers.*" to generate additional rich descriptions. Through this operation, we obtained extra landmark descriptions that include detailed attributes and surrounding environmental information. For more process details and generated results, please refer to Appendix 7.

4 Method

Problem Setups. Given a k-round dialog history $\mathcal{D}_h = \{\mathbf{I_0}, ...\mathbf{Q}_k, \mathbf{I}_k\}$, the ANDH task requires a drone agent to approach the final goal $\mathbf{G}_f$ through sequential action prediction. At each step t, the drone receives a RGB top-down view image $\mathbf{V}_t$, heading angle θ_t and GPS position $\mathbf{p}_t^o = [x_t, y_t, z_t]$. The action $\mathbf{a}_t = [\Delta x_t, \Delta y_t, \Delta z_t]$ is predicted by a learnable policy $\pi\left(\mathbf{a}_t|\mathcal{D}_h, \mathbf{V}_t, \mathbf{p}_t^o, \theta_t\right)$.

Method Overview. Figure 2 presents the overview of the proposed FELA. Similar to (Fan et al. 2023a; Su et al. 2023), FELA employs an episodic transformer (Pashevich, Schmid, and Sun 2021) as our navigator, and enhances the agent's entity-landmark alignment ability with two novel modules: 1) A precise semantic grid representation (§ 4.1) for visual perception, capturing rich semantics and spatial layouts simultaneously. 2) An explicit entity-landmark alignment learning scheme (§ 4.2) based on FG-AVDN dataset. We next elaborate on these modules.

4.1 Semantic Grid Representation

For the learning of entity-landmark alignment, an ideal visual representation should not only contain geometries and semantics of landmarks but also depict their spatial relations. Hence, we propose to construct a semantic grid representation $\mathbf{M}_t \in \mathbb{R}^{N\times N\times D}$ to assist the precise visual perception, where N is the grid scale, and D is the cell vector dimension. Concretely, $\mathbf{M}_t$ is initialized by the latent features $\widetilde{\mathbf{V}}_t \in \mathbb{R}^{N\times N\times D}$ from the last layer of the visual encoder. We further incorporate object encoding to capture the shape and scale of landmarks, and a position encoding to enhance the agent's spatial awareness. The details are shown as follows.

Object Encoding. FELA proposes to explicitly encode all landmarks into a grid embedding $\mathbf{S}_t$ for preserving both semantic and geometric details, which is beneficial for subsequent object-level cross-modal matching (Qi et al. 2021; Wang et al. 2023b). Specifically, we first pretrain a classic Oriented R-CNN detector (O-RCNN) (Xie et al. 2021) with aerial scenes from the xView (Lam et al. 2018) dataset. Next, at each step t, we utilize the trained O-RCNN to generate object masks $\mathcal{O}_t = \{\mathbf{o}_1, \mathbf{o}_2, ..., \mathbf{o}_k\}$ contained in $\mathbf{V}_t$. In this way, a class index map $\mathbf{O}_t^I \in \mathbb{N}^{N\times N\times C}$ is formed by merging all object masks, as illustrated in Figure 2 (a), where C is semantic categories. Afterward, we encode $\mathbf{O}_t^I$ with a convolutional neural network and perform the down-sampling, yielding the final semantic embedding $\mathbf{S}_t \in \mathbb{R}^{N\times N\times D}$.

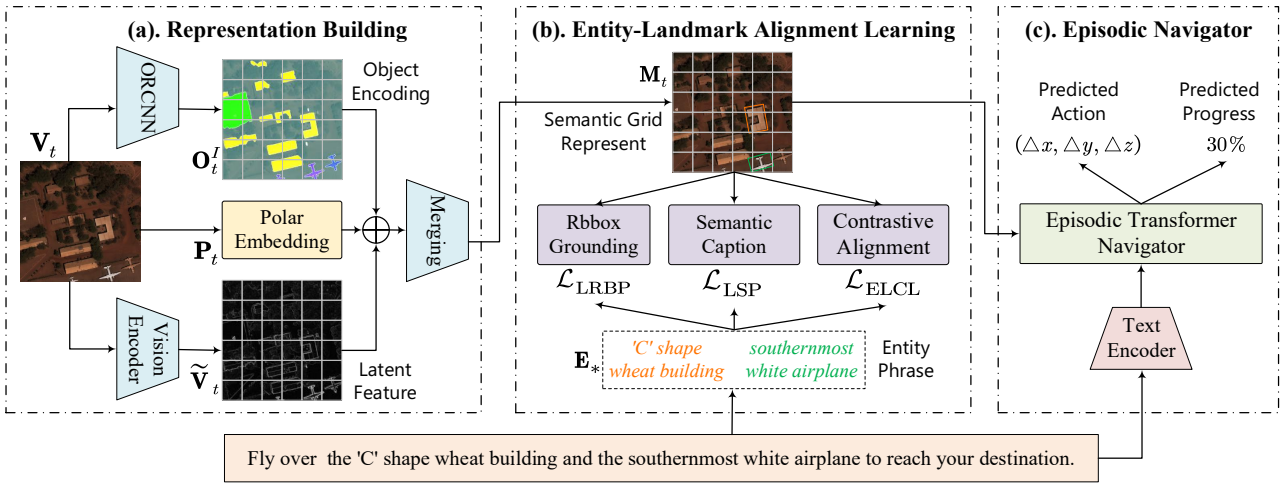


Figure 2: Overview of the FELA method. At each step t, FELA builds the semantic grid representation $\mathbf{M}_t$ based on observation $\mathbf{V}_t$. Then, based on $\mathbf{M}_t$ and FG-AVDN dataset, FELA learns the entity-landmark alignment explicitly via three auxiliary tasks. Subsequently, an episodic transformer is leveraged to conduct multimodal fusion and decision-making.

Spatial Encoding. To handle relative spatial instructions like "Move forward to three o'clock" or "The destination is on your left", a spatial encoding is further introduced to enhance the agent's spatial awareness. Inspired by (An et al. 2023), we adopt a polar position encoding $\mathbf{P}_t \in \mathbb{R}^{N\times N\times 3}$, which regards the agent as center and computes relative positions with other cells. Each $\mathbf{p}_{uv} \in \mathbf{P}_t$ is formulated as:

$$\mathbf{p}_{uv} = [\sin(\beta_{uv}), \cos(\beta_{uv}), \mathrm{dis}_{uv}] \tag{1}$$

where β_{uv} and dis_{uv} are relative heading angle and normalized distance of a cell to drone, respectively.

Encoding Merging. Overall, we obtain the final semantic grid representation $\mathbf{M}_t$ by merging the three kinds of encoding. Formally:

$$\mathbf{M}_t = \mathrm{LN}(\mathbf{W}_v\widetilde{\mathbf{V}}_t + \mathbf{W}_s\mathbf{S}_t + \mathbf{W}_p\mathbf{P}_t) \tag{2}$$

where $\mathbf{W}_v, \mathbf{W}_s, \mathbf{W}_p$ are learnable parameters to unify the feature dimensions, and LN denotes the layer normalization.

4.2 Entity-Landmark Alignment Learning

As shown in Figure 2 (b), we adopt three auxiliary tasks to learn the entity-landmark alignment based on the semantic grid representation $\mathbf{M}_t$, namely Landmark Rotated Bounding box Prediction (LRBP), Landmark Semantic Prediction (LSP), and Entity-Landmark Contrastive Learning (ELCL). Assuming the drone is at time step t, and the dialog $\mathbf{D}_h$ contains K entities $\{\mathbf{E}_k\}_{k=1}^K$ and corresponding landmarks $\{\mathbf{b}_k\}_{k=1}^K$, we next detail the three auxiliary tasks.

LRBP. Compared to classic visual grounding models predicting horizontal bounding boxes (Cui et al. 2023; Kamath et al. 2021), our LRBP task is designed to predict a landmark's compact rotated bounding box, enabling finer alignment. Specifically, for each entity $\mathbf{E}_k$, we first obtain its embedding $\mathbf{e}_k$ through the text encoder. And then we adopt an MHCA (Vaswani et al. 2017) to aggregate the relevant feature from $\mathbf{M}_t$ as:

$$\tilde{\mathbf{e}}_k = \mathrm{FFN}(\mathbf{e}_k + \mathrm{MHCA}(\mathbf{e}_k, \mathbf{M}_t)) \tag{3}$$

Afterward, we predict the rbbox $\mathbf{b}'_k$ for the landmark using a feed-forward network FFN:

$$\mathbf{b}'_k = \mathrm{Sigmoid}(\mathrm{FFN}(\tilde{\mathbf{e}}_k)) \tag{4}$$

Finally, we adopt the smooth L1 loss for supervision:

$$\mathcal{L}_{\mathrm{LRBP}} = \mathcal{L}_{l_1}(\mathbf{b}'_k, \mathbf{b}_k) \tag{5}$$

LSP. This task aims to predict the semantics of landmarks based on their visual feature. Concretely, we model LSP as a region-level caption task. For each landmark $\mathbf{b}_k$, we first exploit the rotated RoI-Align operation (Xie et al. 2021) to interpolate the region feature $\mathbf{R}_k$ from $\mathbf{M}_t$. Based on this, we employ a cross-attention transformer with causal mask (Li et al. 2022) as a decoder to predict the corresponding entity:

$$\mathbf{E}'_k = \mathrm{Transformer\_Decoder}(\mathbf{R}_k) \tag{6}$$

This task is optimized by the cross-entropy loss:

$$\mathcal{L}_{\mathrm{LSP}} = \mathrm{CrossEntropy}(\mathbf{E}'_k, \mathbf{E}_k) \tag{7}$$

ELCL. The objective of this task is to ensure that the embeddings of paired entities and landmarks are closer in the feature space compared to the unpaired ones, enabling a better fine-grained alignment at the word-region level (Zhang et al. 2022; Kamath et al. 2021). Specifically, given K entity-landmark pairs, we construct K^2 triplets as $\{(\mathbf{r}_i, \mathbf{e}_j), y_{ij}\}_{i,j=1}^K$, where $\mathbf{r}_i$ is the pooling vector of $\mathbf{R}_k$ and $y_{ij} = 1$ indicates that $(\mathbf{r}_i, \mathbf{e}_j)$ is a matched pair. Subsequently, following (Jiang and Ye 2023), ELCL is optimized through:

$$\mathcal{L}_{\mathrm{e2l}} = \frac{-1}{K}\sum_{i=1}^{K}\sum_{j=1}^{K} y_{ij}\log\left(\frac{\exp\left(\mathbf{e}_i^T\mathbf{r}_j/\tau\right)}{\sum_{l=1}^{K}\exp\left(\mathbf{e}_i^T\mathbf{r}_l/\tau\right)}\right) \tag{8}$$

$$\mathcal{L}_{\mathrm{ELCL}} = \left(\mathcal{L}_{\mathrm{e2l}} + \mathcal{L}_{\mathrm{l2e}}\right)/2 \tag{9}$$

where τ is a temperature hyper-parameter controlling the probability distribution peaks. $\mathcal{L}_{\text{e2l}}$ is the contrastive loss from entity to landmark, and $\mathcal{L}_{\text{l2e}}$ is a symmetric loss calculated by exchanging $\mathbf{e}_i$ and $\mathbf{r}_j$ in Formula 8.

Full Aligning Objective. The overall loss for entity-landmark alignment learning is:

$$\mathcal{L}_{\text{align}} = \mathcal{L}_{\text{LRBP}} + \kappa_1 \mathcal{L}_{\text{LSP}} + \kappa_2 \mathcal{L}_{\text{ELCL}} \tag{10}$$

where κ_1, κ_2 are weight coefficients for loss balancing.

4.3 Navigation Model

We employ an episodic transformer (Pashevich, Schmid, and Sun 2021) as our navigator, as illustrated in Figure 2 (c). To effectively leverage long-range and local observations, we feed instruction text $\mathbf{D}_h$, vision history $\mathcal{V}_h$, trajectory history $\mathcal{R}_h$,and semantic grid representation $\mathbf{M}_t$ into navigator for modality fusion. Subsequently, a simple three-layer FFN is adopted to predict the action $\mathbf{a}_t$. For more details about the navigator, please refer to the Appendix 8.

4.4 Training Objective

Following the standard practice in ANDH methods (Fan et al. 2023a; Su et al. 2023), we employ an overall multi-task loss to optimize the navigation policy π and other auxiliary tasks as below:

$$\mathcal{L}_{\text{all}} = \mathcal{L}_{\text{nav}} + \mathcal{L}_{\text{align}} \tag{11}$$

where $\mathcal{L}_{\text{align}}$ denotes entity-landmark aligning loss given in Sec.§ 4.2. Regarding navigation policy learning, we alternately run the 'teacher-forcing' and 'student-forcing' modes, where the difference is whether to interact with the simulator via ground truth or predicted action. Meanwhile, a mean squared error loss (Fan et al. 2023a) is leveraged to optimize the policy π, denoted as $\mathcal{L}_{\text{nav}} = \left\|\mathbf{a}_t - \mathbf{a}_t^{\text{gt}}\right\|_2$.

5 Experiments

The proposed method is evaluated on the ANDH task (Fan et al. 2023a), which is the only available benchmark for Aerial Vision-Dialog Navigation. The ANDH task splits the AVDN dataset into 6269 sub-trajectories according to dialog rounds. These sub-trajectories are further divided into 4 splits via their scene types, including 4591 for training, 370 for seen validation, 411 for unseen validation, and others for unseen testing. ANDH mainly focuses on the generalization ability of the agent, thus the performance on unseen validation and testing splits is more crucial.

Evaluation Metrics. We use the standard metrics for evaluation (Fan et al. 2023a), including: 1) Success Rate (SR): the ratio of predicted paths being regarded as successful; 2) Success weighted by inverse Path Length (SPL): SR weighted by the total length of the navigation path; 3) Goal Progress (GP): the distance of the navigation progress towards the destination area.

5.1 Implementation Details

Model Configuration. For a fair comparison, FELA employs an xView pre-trained Yolov5-x backbone for visual encoding following (Su et al. 2023), and a Roberta (Liu et al. 2019) for dialog encoding. Meanwhile, the Swin-Tiny (Liu et al. 2021) is selected as the backbone of the O-RCNN detector. We empirically set the scale of the semantic grid representation N to 7. The hidden size of dialog encoding, history encoding, and semantic grid representation D is uniformly set to 768. The number of transformer layers for the text encoder and episodic transformer is set to 9 and 3, respectively. For weight coefficients, we set κ_1, κ_2 in Formula 10 to $1, 0.1$, respectively. The τ in Formula 8 is set to 0.02 following (Jiang and Ye 2023).

Training Details. Our experiments are conducted on two NVIDIA RTX 3090 GPUs. All models are optimized for 200,000 iterations ($\sim$50 hours) with a batch size of 8 and a learning rate of 1e-5 via AdamW optimizer. The best iteration is determined by the highest performance on unseen validation split. As for the training of the O-RCNN detector, we use the AdamW optimizer and conduct multi-scale training for 12 epochs ($\sim$24 hours) with a batch size of 8 and a learning rate of 2e-4.

5.2 Comparison with State-of-the-Art Methods

Table 1 compares the proposed FELA with current SoTA methods on the ANDH task. FELA achieves the leading performance on both unseen validation and testing splits. In particular, the FELA *w/o* aligning tasks (§ 4.2) surpasses the previous best method TG-GAT (Su et al. 2023) by 1.6% in SR and 1.3% in SPL, which validates the effectiveness of the proposed semantic grid representation. Additionally, when incorporating our FG-AVDN dataset and auxiliary alignment tasks, the previous methods' performance was consistently improved. This demonstrates the necessity of our FG-AVDN dataset. Furthermore, the complete FELA improves the performance through all metrics, e.g., SR increases from 18.7% to 21.9%, SPL increases from 15.1% to 17.6% and GP increases from 56.5 to 61.4. This further highlights the benefits of explicit entity-landmark alignment learning.

For fair comparisons, we report the model parameters and Flops of all AVDN methods in Table 2. Compared to the previous SoTA TG-GAT, the extra computation cost is marginal, e.g., the parameters and Flops of FELA only increase by 2.2% and 2.5%, respectively. Table 2 also shows the inference latency of all AVDN methods on a single RTX3090 GPU, and FELA can execute decisions at 11.7 Hz. These results demonstrate that FELA is computationally comparable with existing methods and has the potential to transfer to the real world.

5.3 Ablation Study

This section conducts ablation experiments regarding different design choices of FELA. The results are reported on the unseen validation split of ANDH.

Different Options for Grid Representation Construction. Table 3 presents various options for constructing the semantic grid representation in § 4.1. Row 1 and Row 2 solely build the grid representation through visual encoding $\widetilde{\mathbf{V}}_t$ and semantic encoding $\mathbf{S}_t$, respectively. However, both approaches yield unsatisfactory results. In particular, the performance of Row 2 is significantly inferior to Row 1, e.g., 13.5% SPL

Methods	Seen Validation			Unseen Validation			Unseen Testing		
	SPL↑	SR↑	GP↑	SPL↑	SR↑	GP↑	SPL↑	SR↑	GP↑
Transformer	12.1	14.1	50.1	14.3	16.6	51.9	11.3	13.3	51.7
HAA-LSTM	11.6	13.0	50.3	18.3	20.0	54.4	12.6	14.1	50.8
HAA-Transformer	14.7	17.3	56.3	16.5	20.4	55.2	12.9	15.7	54.2
HAA-Transformer *w* aligning tasks	15.1	17.8	57.3	16.0	21.4	57.2	14.6	18.0	55.2
TA-GAT	12.9	16.0	56.9	18.8	23.3	54.3	15.1	18.7	56.5
TA-GAT *w* aligning tasks	14.8	18.2	58.8	17.8	21.1	61.7	15.9	19.7	56.3
FELA	15.1	18.8	**60.8**	17.2	20.6	63.0	16.4	20.3	56.7
FELA *w* aligning tasks	**15.3**	**18.8**	60.7	**19.2**	**23.9**	**64.1**	**17.6**	**21.9**	**61.4**

Table 1: Comparison with the state-of-the-art methods on the ANDH task.

Model	Params (M)	Flops (B)	Speed (Hz)	Speed (ms)
HAA-Transformer	173.4	66.7	16.6	60.3
TG-GAT	187.0	70.6	12.3	81.2
FELA	191.2	72.4	11.7	85.6

Table 2: Parameters and inference time for AVDN methods

#	Image Encoding	Semantic Encoding	Position Encoding	SPL↑	SR↑	GP↑
1	✓	✗	✗	17.0	19.8	63.0
2	✗	✓	✗	13.5	17.5	61.5
3	✓	✓	✗	18.3	21.3	60.4
4	✓	✓	✓	**19.2**	**23.9**	**64.1**

Table 3: Effect of different grid representation encodings.

v.s.17.0% SPL. This discrepancy can be attributed to the inadequacy of semantics within predefined categories in capturing attribute-related landmarks, such as a "wheat color building". In Row 3, the semantic grid representation incorporates $\widetilde{\mathbf{V}}_t$ and $\mathbf{S}_t$, resulting in 1.3% SPL and 1.5% SR gains compared to Row 1. This indicates that explicit object encoding and implicit visual encoding can complement each other. Row 4 further introduces polar position encoding $\mathbf{P}_t$, achieving the highest performance, with metrics of 19.2% SPL, 23.9% SR, and 64.1 GP. We attribute this superiority to the spatial awareness facilitated by spatial encoding, which is beneficial for grounding spatial-conditioned landmarks, such as "warehouse office at your three o'clock". Consequently, we adopt Row 4 as our default option for grid representation construction.

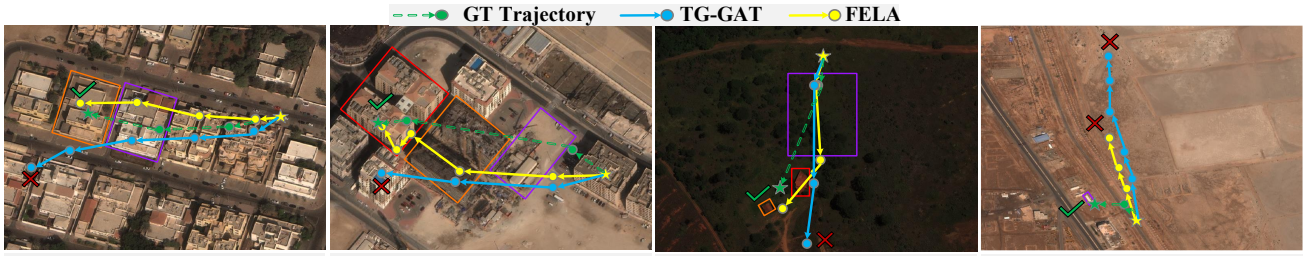
Entity-Landmark Alignment Facilitates ANDH. Table 4 illustrates the effect of different auxiliary tasks in § 4.2. Row 1 serves as a baseline without any auxiliary task, while Row 2-5 augment the baseline with LHBP, LRBP, LSP, and ELCL tasks, respectively. Compared to Row 2 using the landmark horizontal bounding box prediction task, Row 3 obtains a 0.5% gain in SR, indicating that compact rbbox leads to better alignment. Row 3 and 4 show that single-modal auxiliary tasks (LRBP and LSP) yield decent improvements, with a 0.4% increase in SPL. Furthermore, Row 5 reveals that the dual-modal ELCL task contributes slightly higher gains of 0.6% SPL. This observation suggests that contrastive learning over entity and landmark features provides more robust constraints, leading to enhanced cross-modal alignment. Row 6 combines LRBP and LSP tasks, resulting in satisfactory performance, e.g., 18.5% SPL and 22.6% SR. This presents the complementary nature of the two tasks. Row 7 adopts all auxiliary tasks simultaneously, achieving the highest performance across all metrics. These results show that the three tasks are complementary and highlight the effectiveness of entity-landmark learning for navigation generalization.

#	Auxiliary Tasks	SPL↑	SR↑	GP↑
1	None	17.5	20.1	61.3
2	LHBP	17.6	20.6	60.8
3	LRBP	17.9	21.1	61.3
4	LSP	17.9	21.6	61.7
5	ELCL	18.1	21.8	60.9
6	LRBP + LSP	18.5	22.6	60.6
7	LRBP + LSP + ELCL	**19.2**	**23.9**	**64.1**

Table 4: The effect of entity-landmark alignment tasks.

The Impact of Semantic Grid Representation Scale N. Table 5 reports the performance of FELA with different grid scales N. FELA achieves the best performance when the N increases to 7. However, when N further increases to 9, the performance remains unchanged, and even slightly decreases when it reaches 11. The potential reason is that a large N leads each token to focus excessively on the local details, obstructing the aggregation of effective features from numerous tokens. Consequently, we adopt a relatively balanced scale $N = 7$ as the default setting.

The Impact of Different Entity-Landmark Pairs. In Table 6, we examine the performance under the supervision of two kinds of entity-landmark data in § 3.2. Row 1 serves as a baseline without any entity-landmark pairs for training. Row 2 only utilizes the entity-landmark pairs extracted from



(a) [INS] *Head west until you pass* ***the compound with the white buildings****. Enter* ***the next tan compound after that****, and you have arrived at the destination.*

(b) [QUE] *I am on the large building left, what is the destination look like? Which direction should I go?* [INS] *Head west, past* ***the grounds*** *and* ***the dilapidated building*** *and head towards* ***the U-shaped building with a beige roof****, that's your destination.*

(c) [INS] *head forward towards your 1 o'clock direction. After passing through* ***the green field****. the destination is* ***a brown color house with another brown color house before reaching it****.*

(d) [INS] *Head north following road for* ***80 meters****. Your destination is* ***long rectangular white building on west side****.*

Figure 3: Visualization of predicted paths of FELA (yellow path) and TG-GAT (blue path) on unseen validation split. Yellow and green stars represent the starting and ending points of the ground truth path, respectively. Below are the given dialogs, where colored bold words are entities. Above are visual observations where colored bounding boxes are corresponding landmarks.

#	Grid Scale	SPL↑	SR↑	GP↑
1	5×5	17.9	20.2	62.1
2	7×7	**19.2**	**23.9**	**64.1**
3	9×9	18.6	23.8	63.7
4	11×11	18.8	21.5	61.6

Table 5: The effect of semantic grid representation scale N.

#	Training Data	SPL↑	SR↑	GP↑
1	None	17.5	20.1	61.3
2	Extracted Pairs	18.5	22.7	62.4
3	Generated Pairs	**19.2**	**23.9**	**64.1**

Table 6: The impact of different auxiliary training data.

the dialog for alignment learning, resulting in a 2.3% gains in SR and a 1.0% improvement in SPL. This indicates the efficacy of explicit fine-grained alignment learning. Row 3 further introduces the detailed landmark descriptions generated by LHRS-Bot, yielding a 1.2% gain in SR and 0.7% gain in SPL, demonstrating the effectiveness of the generated landmark descriptions.

5.4 Qualitative Results

We present comparisons of the predicted paths of FELA and TG-GAT in Figure 3. More navigation cases and entity-landmark alignment results, please refer to Appendix 10.

Visualization of Navigation Paths. From (a), we observe that FELA successfully identifies and stops at the target landmark "the next tan compound", whereas TG-GAT ends at the wrong building. Similarly, in (b), FELA navigates better along the intermediate landmark "the dilapidated building" and accurately reaches the final landmark "U-shaped building with a beige roof", while TG-GAT ignores crucial landmarks and leads to navigation failure. In addition, we observe that FELA performs better in entity phrases with complex modifiers, as in (c) with the description "a brown color house with another brown color house before reaching it". We also present a failure case in (d), where FELA fails to navigate for "80 meters". We attribute it to the lack of absolute spatial awareness capability, which hinders the interpretation of distance-related instructions.

6 Conclusion

This paper tackles a critical but under-explored problem in ANDH - the entity-landmark alignment. We first construct the FG-AVDN dataset via a semi-automatic annotation pipeline, providing fine-grained alignment at the entity-landmark level. Then, the proposed FELA method combines a precise semantic grid representation and three auxiliary tasks to learn such alignment explicitly. Extensive experiments have demonstrated the effectiveness of explicit entity-landmark alignment learning for ANDH. However, considering the performance of existing navigators remains unsatisfactory, performing real-world experiments on drones could raise safety issues. We plan to consistently improve the robustness of the navigator, and then conduct sim-to-real deployments in the future.

Acknowledgements

This work was jointly supported by the National Natural Science Foundation of China (62236010, 62322607 and 62276261), and Youth Innovation Promotion Association of Chinese Academy of Sciences under Grant 2021128.

References

An, D.; Qi, Y.; Li, Y.; Huang, Y.; Wang, L.; Tan, T.; and Shao, J. 2023. BEVbert: Multimodal map pre-training for language-guided navigation. *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

An, D.; Wang, H.; Wang, W.; Wang, Z.; Huang, Y.; He, K.; and Wang, L. 2024. Etpnav: Evolving topological planning for vision-language navigation in continuous environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Anderson, P.; Wu, Q.; Teney, D.; Bruce, J.; Johnson, M.; Sünderhauf, N.; Reid, I.; Gould, S.; and Van Den Hengel, A. 2018. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3674–3683.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Chen, H.; Suhr, A.; Misra, D.; Snavely, N.; and Artzi, Y. 2019. Touchdown: Natural language navigation and spatial reasoning in visual street environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12538–12547.

Chen, P.; Ji, D.; Lin, K.; Zeng, R.; Li, T.; Tan, M.; and Gan, C. 2022a. Weakly-supervised multi-granularity map learning for vision-and-language navigation. *Advances in Neural Information Processing Systems*, 35: 38149–38161.

Chen, Q.; Wang, T.; Yang, Z.; Li, H.; Lu, R.; Sun, Y.; Zheng, B.; and Yan, C. 2024. SDPL: Shifting-Dense Partition Learning for UAV-View Geo-Localization. *arXiv preprint arXiv:2403.04172*.

Chen, S.; Guhur, P.-L.; Tapaswi, M.; Schmid, C.; and Laptev, I. 2022b. Think global, act local: Dual-scale graph transformer for vision-and-language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16537–16547.

Chen, Y.; Yang, Z.; and Chen, Q. 2023. A Cross-View Matching Method Based on Dense Partition Strategy for UAV Geolocalization. In *Proceedings of the 2023 Workshop on UAVs in Multimedia: Capturing the World from a New Perspective*, 19–23.

Cheng, W.; Dong, X.; Khan, S.; and Shen, J. 2022. Learning disentanglement with decoupled labels for vision-language navigation. In *European Conference on Computer Vision*, 309–329. Springer.

Chu, M.; Zheng, Z.; Ji, W.; Wang, T.; and Chua, T.-S. 2025. Towards natural language-guided drones: GeoText-1652 benchmark with spatial relation matching. In *European Conference on Computer Vision*, 213–231. Springer.

Cui, Y.; Xie, L.; Zhang, Y.; Zhang, M.; Yan, Y.; and Yin, E. 2023. Grounded entity-landmark adaptive pre-training for vision-and-language navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 12043–12053.

Fan, Y.; Chen, W.; Jiang, T.; Zhou, C.; Zhang, Y.; and Wang, X. E. 2023a. Aerial vision-and-dialog navigation. In *Findings of the Association for Computational Linguistics: ACL 2023*, 3043–3061. Toronto, Canada: Association for Computational Linguistics.

Fan, Y.; Gu, J.; Zheng, K.; and Wang, X. E. 2023b. R2H: Building multimodal navigation helpers that respond to help requests. *arXiv preprint arXiv:2305.14260*.

He, K.; Huang, Y.; Wu, Q.; Yang, J.; An, D.; Sima, S.; and Wang, L. 2021. Landmark-RxR: Solving vision-and-language navigation with fine-grained alignment supervision. *Advances in Neural Information Processing Systems*, 34.

He, K.; Jing, Y.; Huang, Y.; Lu, Z.; An, D.; and Wang, L. 2024. Memory-adaptive vision-and-language navigation. *Pattern Recognition*, 153: 110511.

Hong, Y.; Rodriguez-Opazo, C.; Wu, Q.; and Gould, S. 2020. Sub-instruction aware vision-and-language navigation. *arXiv preprint arXiv:2004.02707*.

Hwang, M.; Jeong, J.; Kim, M.; Oh, Y.; and Oh, S. 2023. Meta-explore: Exploratory hierarchical vision-and-language navigation using scene object spectrum grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6683–6693.

Jiang, D.; and Ye, M. 2023. Cross-modal implicit relation reasoning and aligning for text-to-image person retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2787–2797.

Kamath, A.; Singh, M.; LeCun, Y.; Synnaeve, G.; Misra, I.; and Carion, N. 2021. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1780–1790.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; Dollár, P.; and Girshick, R. 2023. Segment anything. *arXiv:2304.02643*.

Ku, A.; Anderson, P.; Patel, R.; Ie, E.; and Baldridge, J. 2020. Room-across-room: Multilingual vision-and-language navigation with dense spatiotemporal grounding. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Lam, D.; Kuzma, R.; McGee, K.; Dooley, S.; Laielli, M.; Klaric, M.; Bulatov, Y.; and McCord, B. 2018. xView: Objects in Context in Overhead Imagery. *arXiv preprint arXiv:1802.07856*.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, 12888–12900. PMLR.

Liu, F.; Chen, D.; Guan, Z.; Zhou, X.; Zhu, J.; Ye, Q.; Fu, L.; and Zhou, J. 2024. Remoteclip: A vision language foundation model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*.

Liu, S.; Zhang, H.; Qi, Y.; Wang, P.; Zhang, Y.; and Wu, Q. 2023. AerialVLN: Vision-and-language navigation for

UAVs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15384–15394.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Moudgil, A.; Majumdar, A.; Agrawal, H.; Lee, S.; and Batra, D. 2021. Soat: A scene-and object-aware transformer for vision-and-language navigation. *Advances in Neural Information Processing Systems*, 34: 7357–7367.

Muhtar, D.; Li, Z.; Gu, F.; Zhang, X.; and Xiao, P. 2024. LHRS-Bot: Empowering Remote Sensing with VGI-Enhanced Large Multimodal Language Model. *arXiv preprint arXiv:2402.02544*.

Pashevich, A.; Schmid, C.; and Sun, C. 2021. Episodic transformer for vision-and-language navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15942–15952.

Qi, Y.; Pan, Z.; Hong, Y.; Yang, M.-H.; Van Den Hengel, A.; and Wu, Q. 2021. The road to know-where: An object-and-room informed sequential bert for indoor vision-language navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1655–1664.

Qi, Y.; Wu, Q.; Anderson, P.; Wang, X.; Wang, W. Y.; Shen, C.; and Hengel, A. v. d. 2020. Reverie: Remote embodied visual referring expression in real indoor environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9982–9991.

Qiao, Y.; Qi, Y.; Hong, Y.; Yu, Z.; Wang, P.; and Wu, Q. 2022. HOP: History-and-order aware pre-training for vision-and-language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15418–15427.

Su, Y.; An, D.; Xu, Y.; Chen, K.; and Huang, Y. 2023. Target-grounded graph-aware transformer for aerial vision-and-dialog navigation. *arXiv preprint arXiv:2308.11561*.

Thomason, J.; Murray, M.; Cakmak, M.; and Zettlemoyer, L. 2020. Vision-and-dialog navigation. In *Proceedings of the Conference on Robot Learning*, 394–406.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

Wang, H.; Liang, W.; Shen, J.; Van Gool, L.; and Wang, W. 2022. Counterfactual cycle-consistent learning for instruction following and generation in vision-language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15471–15481.

Wang, H.; Liang, W.; Van Gool, L.; and Wang, W. 2023a. Dreamwalker: Mental planning for continuous vision-language navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10873–10883.

Wang, L.; He, Z.; Tang, J.; Dang, R.; Wang, N.; Liu, C.; and Chen, Q. 2023b. A dual semantic-aware recurrent global-adaptive network for vision-and-language navigation. In *Proceedings of the 32nd International Joint Conference on Artificial Intelligence*, 1479–1487.

Wang, X.; Huang, Q.; Celikyilmaz, A.; Gao, J.; Shen, D.; Wang, Y.-F.; Wang, W. Y.; and Zhang, L. 2019. Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6629–6638.

Wang, X.; Wang, W.; Shao, J.; and Yang, Y. 2023c. LANA: A language-capable navigator for instruction following and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19048–19058.

Wang, Z.; Li, X.; Yang, J.; Liu, Y.; and Jiang, S. 2023d. GridMM: Grid memory map for vision-and-language navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15625–15636.

Xie, X.; Cheng, G.; Wang, J.; Yao, X.; and Han, J. 2021. Oriented R-CNN for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3520–3529.

Zhang, H.; Zhang, P.; Hu, X.; Chen, Y.-C.; Li, L.; Dai, X.; Wang, L.; Yuan, L.; Hwang, J.-N.; and Gao, J. 2022. Glipv2: Unifying localization and vision-language understanding. *Advances in Neural Information Processing Systems*, 35: 36067–36080.

Zhao, Y.; Chen, J.; Gao, C.; Wang, W.; Yang, L.; Ren, H.; Xia, H.; and Liu, S. 2022. Target-driven structured transformer planner for vision-language navigation. In *Proceedings of the 30th ACM International Conference on Multimedia*, 4194–4203.

Zhu, F.; Liang, X.; Zhu, Y.; Yu, Q.; Chang, X.; and Liang, X. 2021. SOON: Scenario oriented object navigation with graph-based exploration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12689–12699.