2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

Where I Am & Where to Go: Egocentric Indoor Scene Perception with Agent Interaction for Remote Embodied Visual Grounding

Hongtao Zhang[†,1], Yili Tang[†,1], Yuan Gao[*,‡,2], Jue Zhang[1], Jidong Zhang[*,3] and Mingbo Zhao[*,1]

***Abstract*— Embodied Referring Expression Grounding (REVERIE) is a Vision-and-Language Navigation (VLN) task that better reflects real-world human instructions. Unlike conventional VLN, REVERIE is more challenging as agents must navigate in unseen environments and ground remote objects described by short, high-level commands. This requires agents not only to plan a route without detailed step-by-step guidance but also to accurately localize the target object at the destination. Existing VLN agents mainly emphasize navigation performance while overlooking object grounding success, leading to a significant performance gap. We introduce a model-agnostic interaction framework with two auxiliary agents, *Where-I-Am* (WIA) and *Where-to-Go* (W2G). Specifically, WIA predicts the current room type from environmental observations, while W2G infers the target room type from high-level instructions. Our framework is plug-and-play and can be integrated with various VLN models. On the REVERIE benchmark, it improves navigation success rate (SR) by 7.78% and remote grounding success (RGS) by 5.48% over the baselines, demonstrating the effectiveness and generality of our design. Furthermore, in challenging unseen test environments, our framework achieves competitive results on the REVERIE dataset, outperforming the previous state-of-the-art VLN agent (without additional training data) with a 2.27% gain in RGS.**

I. Introduction

A long-standing challenge in robotics lies in enabling robots to interact with humans in the visual world through natural language, as humans are inherently visual beings whose communication is grounded in language [1]. While a cognitively normal child can effortlessly retrieve a cup in a completely unfamiliar environment, the probability of a robot accomplishing the same task remains significantly lower [2]. This discrepancy stems from the fact that humans develop intuitive reasoning about visual scenes and linguistic semantics through accumulated life experiences, whereas current robots lack such integrative capabilities, thereby constraining their applicability in real-world scenarios [3]. Consequently, Embodied Referring Expression Grounding (REVERIE), a vision-and-language navigation (VLN) benchmark that more closely approximates human communication, has gained substantial attention within the embodied AI community [4].

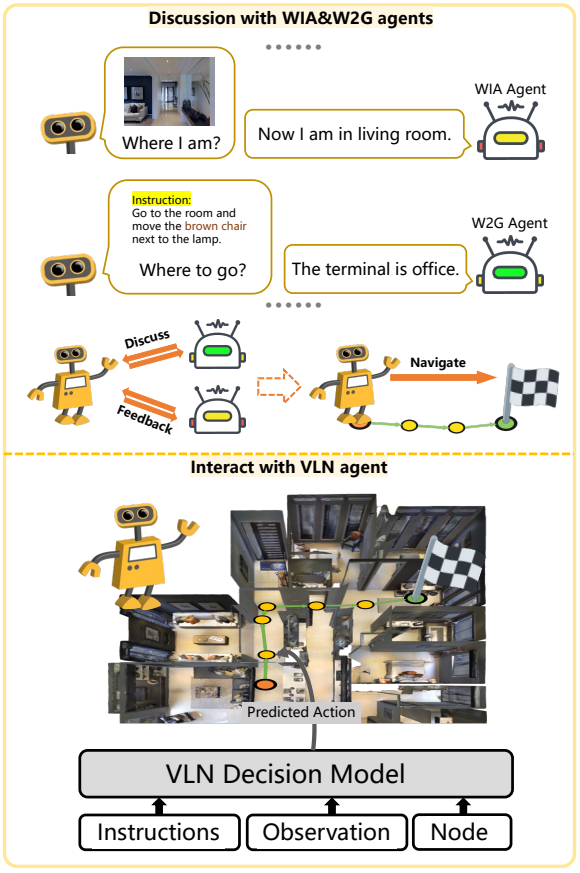

Fig. 1: Framework of interaction among WIA, W2G, and VLN agents. The WIA agent provides spatial awareness, the W2G agent gives goal-oriented instructions, and the embodied VLN agent interacts with them before navigating. The VLN decision model integrates instructions, observations to predict navigation actions in 3D indoor environments.

Despite recent progress, embodied VLN continues to face two critical challenges [5], [6]. (1) In realistic scenarios, fine-grained natural language instructions are often unavailable, and human–robot interaction typically depends on abstract, high-level directives [7]. For instance, given a concise command such as "Go to the office and move the brown chair next to the lamp," humans can readily interpret the semantics, construct a mental representation of the target room and object, and continuously align the visual and language representations with perceptual observations to reach the goal [8]. Embodied VLN agents, however, remain limited in their ability to effectively process and act upon such high-level instructions [9]. (2) Empirical studies show that

† These authors are equal contributions.
* Corresponding author.
‡ Project lead.
[1] Hongtao Zhang, Yili Tang, Jue Zhang, Mingbo Zhao are with School of Information and Intelligent Science, Donghua University, Shanghai, China.
[2] Yuan Gao is with Shanghai AI Laboratory, Shanghai, China.
[3] Jidong Zhang is with China Telecom Corporation Limited, China.
This work is supported by the National Natural Science Foundation of China under Grants 61971121, in part by the Key Scientific Research Platform and Project of Guangdong Provincial Education Department (No. 2022ZDZX1038), in part by the Program of China Scholarship Council (No. 202506630048).

state-of-the-art VLN agents still fail to stop at the correct target room in nearly 30% of cases [10]. To overcome these limitations, we propose two auxiliary agents dedicated to room-type recognition, which collaboratively interact with the VLN agent to enhance navigation performance.

In this paper, we present a novel interaction framework for REVERIE that explicitly incorporates human-inspired intuitions into VLN. First, we introduce two auxiliary agents, *Where-I-Am* (WIA) and *Where-to-Go* (W2G), which model the and "where" reasoning processes of humans by inferring the current room type and the target room type, respectively. Second, we build two dedicated datasets for training WIA and W2G, enabling robust room-aware and goal-aware perception in complex indoor environments. Third, we seamlessly integrate WIA and W2G into state-of-the-art VLN agents such as DUET and HAMT, in a plug-and-play and model-agnostic manner. Finally, extensive experiments on the REVERIE benchmark demonstrate that our approach significantly improves both navigation success and remote grounding accuracy.

The main contributions are summarized as follows:

- We propose a human-inspired interaction framework with two auxiliary agents, *Where-I-Am* (WIA) and *Where-to-Go* (W2G), to jointly address spatial awareness and goal inference in REVERIE.
- We design dedicated training schemes for WIA and W2G and integrate them into existing VLN backbones in a plug-and-play, model-agnostic manner.
- Extensive experiments on the REVERIE benchmark demonstrate substantial improvements in both navigation and grounding performance, achieves competitive results compared to the previous state-of-the-art VLN agent in test unseen environments.

II. Related Works

A. Indoor scene recognition

Indoor scene recognition is frequently regarded as a fundamental step towards achieving a high-level understanding and reasoning of indoor environments [11]. From the perspective of an indoor robotic assistant, understanding the specific category of room [12], e.g., kitchen, hallway, or bedroom, is crucial for optimizing navigation success ratio (SR). The research began with scene recognition from 2D images, leading to the establishment of serveral scene-centered datasets, i.e., MIT Indoor67, SUN397 and Places365. The remarkable feature extraction capabilities of convolutional neural networks (CNNs) and self-attention mechanisms have firmly established deep neural network-based algorithms as the leading methodology in indoor scene classification tasks. In indoor environments, depth information serves as a significant supplementary cue [13]. Chen *et al.* introduce a comprehensive RGB-D dataset, named Matterport3D, comprising 10,800 panoramic views across 90 extensive architectural scenes [14]. Until now, Matterport3D dataset has facilitated a wide range of supervised computer vision tasks, including but not limited to semantic segmentation and keypoint matching. Based on panoramic views from Matterport3D, We build two small-scale datasets to separately train the *Where-I-Am* (WIA) and *Where-to-Go* (W2G) agents.

B. Spatial Perception in VLN

With the rapid advancements in computility and deep learning technologies, vision-and-language navigation, a multifaceted endeavor intersecting the realms of computer vision, natural language processing, and robotics, has attracted increasingly attention from the research community in recent years, yielding substantial progress [15]. Room-to-Room (R2R) [16] and REVERIE [9] are two foundational datasets in VLN and have become essential benchmarks for assessing the performance of various VLN algorithms since then. Contrastive Language-Image Pretraining (CLIP), a large-scale pre-trained vision-language model, has been utilized to endow the VLN agents with zero-shot navigation capabilities, attributable to its robust generalization capabilities. Li *et al.* develop a layout learner that leverages CLIP to create prompts, enabling the VLN agent to gain visual common sense knowledge [17]. Experimental results indicate that the integration of the layout learner with the VLN agent has led to a 4% improvement in the navigation SR. Qiao *et al.* conduct an analysis of dialogues between large language models (LLMs) and agents to develop the Room-and-Object Aware Scene Perceiver (ROASP) [18]. ROASP generates detailed, step-by-step planning for the VLN agent by describing the types of objects present within a room. The aforementioned methods directly leverage pre-trained visual representations derived from the CLIP-based models. Although the visual representations from CLIP exhibit enhanced generalization capabilities relative to conventional classification networks, their performance in indoor room classification tasks still remains less than satisfactory.

C. Auxiliary Tasks in VLN

For transformer-based VLN models, pretraining with auxiliary tasks has proven to be an effective initialization strategy. These tasks typically include offline expert demonstrations via behavior cloning as well as commonly used vision–language proxy objectives. Majumdar *et al.* first introduced the masked language modeling (MLM) task for VLN pretraining [19]. By training image–text pairs with MLM, VLN agents are able to improve their trajectory selection performance. Anderson *et al.* initially proposed the masked region classification (MRC) task, which was later adopted by Chen *et al.* for VLN pretraining [20]. The goal of MRC is to predict the semantic label of masked observations along a trajectory given the navigation instruction and nearby visual inputs. Chen *et al.* also proposed the single-step action prediction (SAP) task for VLN pretraining [21]. SAP applies imitation learning to predict the next action conditioned on the instruction and historical context, enabling the model to learn how to make action decisions effectively. In this work, we sequentially incorporate MLM, MRC, and SAP into the training of the VLN agent, which leads to notable improvements in navigation performance.

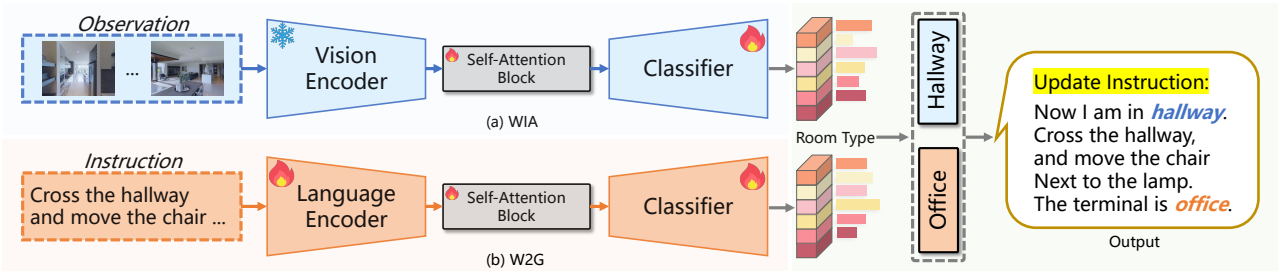

Fig. 2: Architectures of the *Where-I-Am* (WIA) and *Where-to-Go* (W2G) modules. WIA processes visual observations and W2G encodes language instructions, and their outputs are integrated to update context-aware navigation guidance.

III. Method

In this section, we first introduce *Where-I-Am* (WIA) module, and then introduce the *Where-to-Go* (W2G) module. Next, we introduce how to integrate the WIA and W2G module into the previous state-of-the-art VLN agent.

A. Navigation Task Formulation

The decision-making of REVERIE can be formulated as a Partially Observable Markov Decision Process (POMDP). When starting the navigation, the VLN agent receives a natural language instruction $\mathcal{W} = \{w_i\}_{i=1}^{L}$, where w_i represents the ith word token and L is the length of the instruction. At each time step t, the VLN agent gets an observation derived from the state s_0, and then generates the next action a_1 from the control policy π. After that, the VLN agent can gain an immediate reward and step into the next state s_1. In this work, the state s refers to the position of the VLN agent. The observation refers to the panoramic view $\mathcal{R}_t = \{r_i\}_{i=1}^{n}$ perceived by the VLN agent at the current position, where each single view image r_i is represented by an image feature vector. Different from R2R, the VLN agent can also obtain the object features $\mathcal{O}_t = \{o_i\}_{i=1}^{m}$, where m denotes the number of objects which are extracted from the panorama view. The VLN agent continues to navigate within the environment until a special `[stop]` action is chosen or reaches the pre-defined maximum number of steps. Thus, we can consider the immediate reward to be zero.

B. Egocentric Indoor Perception with Agent Interaction

We build two small-scale datasets to separately train the *Where-I-Am* (WIA) and *Where-to-Go* (W2G) agents. For WIA, we obtain room type labels from the Matterport point-wise semantic annotations, which contain 30 distinct categories of indoor scenes. For W2G, we manually annotate the target room type for each navigation instruction from the REVERIE training dataset. After training, the predicted room categories from both WIA and W2G are incorporated into the navigation instructions for embodied decision-making. The detailed training procedures are described as follows.

WIA Agent. The goal of the WIA agent is to enable the VLN agent to recognize the type of room it is currently located in. As illustrated in Fig.2(a), the WIA module consists of a pre-trained CLIP vision encoder, followed by self-attention layers and a two-layer feed-forward network (FFN) as a classifier. Given the panoramic observation at a viewpoint $\mathcal{R}_t$, the extracted features are denoted as R_t^{Pred}:

$$\mathcal{R}_t^{\text{Pred}} = \mathcal{F}^{\text{Softmax}}\,\mathcal{F}^{\text{FFN}}\big(\mathcal{F}^{\text{SA}}(\mathcal{F}^{\text{CLIP}}(\mathcal{R}_t))\big) \tag{1}$$

$$\mathcal{L}_t^{\text{WIA}} = \text{CrossEntropy}(\mathcal{R}_t^{\text{Pred}}, \mathcal{R}_t^{\text{GT}}) \tag{2}$$

where $\mathcal{R}_t$ is the panoramic input at time step t, $\mathcal{R}_t^{\text{Pred}}$ is the predicted room type distribution, and $\mathcal{R}_t^{\text{GT}}$ is the ground-truth room label. Since the current environment typically belongs to a single room category, even though the panorama contains multiple viewpoints, the dominant room type should exert the largest influence. Therefore, we average the predicted scores across all viewpoints and assign the final room type as the one with the highest averaged score. During training, the weights of the pre-trained CLIP encoder remain frozen.

W2G Agent. While the WIA agent focuses on the current location, a navigation agent without a clear understanding of its goal destination is inherently short-sighted. The aim of the W2G agent is to help the VLN agent infer the target room type specified by the instruction. As illustrated in Fig.2(b), the W2G module is composed of a pre-trained multi-layer transformer encoder (i.e., BERT), followed by self-attention layers and a two-layer feed-forward network (FFN) as a classifier. For the instruction sequence $\mathcal{W}$ corresponding to a given trajectory, the predicted room type distribution is denoted as $\mathcal{W}^{\text{Pred}}$:

$$\mathcal{W}^{\text{Pred}} = \mathcal{F}^{\text{Softmax}}\,\mathcal{F}^{\text{FFN}}\big(\mathcal{F}^{\text{SA}}(\mathcal{F}^{\text{Bert}}(\mathcal{W}))\big) \tag{3}$$

$$\mathcal{L}^{\text{W2G}} = \text{CrossEntropy}(\mathcal{W}^{\text{Pred}}, \mathcal{W}^{\text{GT}}) \tag{4}$$

where $\mathcal{W}^{\text{Pred}}$ is the predicted distribution over room categories, and $\mathcal{W}^{\text{GT}}$ is the ground-truth label of the target room. In contrast to WIA, all network parameters of W2G, including those of the pre-trained BERT encoder, are updated through backpropagation.

C. Integration with VLN Agent

We regard the integration of WIA and W2G as a model-agnostic augmentation strategy, and apply it to the previous state-of-the-art framework, i.e., DUET. For image embeddings, a pre-trained CLIP vision encoder is employed to extract features from both panoramic observations and detected objects. For language embeddings, a pre-trained multi-layer BERT encoder is adopted to process the updated navigation instructions. Fig.3 illustrates the node embedding, cross-model blocks and the forward propagation process.

Node Embedding from Topological Map. At time step t, we construct a topological graph $\mathcal{G}_t = \{\boldsymbol{v} \mid \boldsymbol{v} \subseteq \mathcal{V}\}$ to

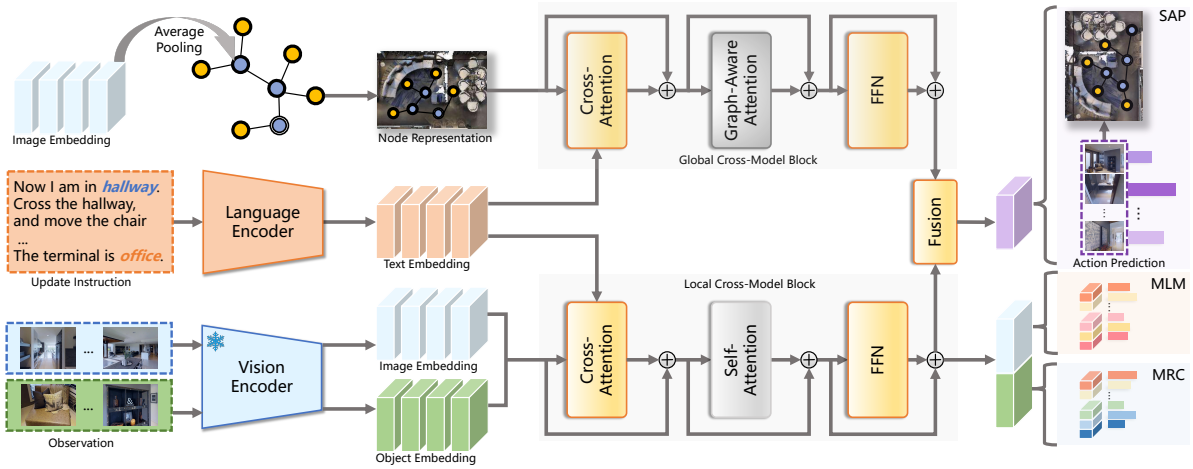

Fig. 3: Framework of WIA&W2G+DUET, based on the dual-scale graph transformer. Our model utilizes the topological graph together with the navigation instruction as inputs, and adaptively integrates the decisions from both local and global branches to determine the subsequent action.

represent the environment. As illustrated in Fig.3, the graph consists of three types of nodes: visited nodes (blue circles), navigable nodes (yellow circles), and the current node (double blue circle). Both the visited nodes and the current node have already been explored, and the VLN agent can access their panoramic visual features, while navigable nodes remain unexplored and are partially observable from visited nodes. For feature encoding, when navigating to a node at step t, the VLN agent extracts panoramic features $\mathcal{R}_t$ from the panorama and object features $\mathcal{O}_t$ from the corresponding bounding boxes. A multi self-attention layers-based encoder is then applied to fuse these two types of features, producing the local visual representation of the node. The update rules for node representations are as follows:

- for the current node, its representation is obtained by concatenating $\hat{\mathcal{R}}_t$ and $\hat{\mathcal{O}}_t$ followed by average pooling.
- for unvisited nodes, the average of all partially observed image features from multiple viewpoints is used.
- for visited nodes, the representation remains unchanged.

Finally, the node representation consists of the location embedding, the step embedding, and the visual embedding. The location embedding is formed by concatenating the Euclidean distance, heading, and elevation relative to the current node; the step embedding records the last visited time step of each node, while unvisited nodes are assigned zero.

Cross-Modal Graph Encoding. For the global cross-model block, the node embeddings are used as query tokens and the word embeddings are used as key and value tokens, respectively. A cross-attention Layer is used to models the relationships between the map and the instruction. Graph-Aware Attention (GAA) layer incorporates the graph's structural information to refine attention computation, denoted as

$$\mathcal{F}^{\text{GAA}}(\mathcal{V}) = \mathcal{F}^{\text{Softmax}}\left(\frac{\mathcal{V}\mathbf{W}_q(\mathcal{V}\mathbf{W}_k)^T}{\sqrt{d}} + \mathcal{M}\right)\mathcal{V}\mathbf{W}_v \quad (5)$$

where $\mathcal{V}$ denotes node representations, $\mathbf{W}_q$, $\mathbf{W}_k$, $\mathbf{W}_v$ denote the learnable parameter matrices, $1/\sqrt{d}$ is the scaling factor. Here, we omit the biases. $\mathcal{M}$ denotes the pair-wise distance matrix from the topological map. For the local cross-model block, the image embeddings are used as query tokens and the word embeddings are used as key and value tokens, respectively. We employs a 4-layer cross-modal transformer to integrate vision–language feature, computing action logits only over navigable views. During action prediction, we dynamically fuse the global action and local action with a two-layer feed-forward network.

D. Training and Inference

Train Procedure. We first pre-train our model on proxy tasks, including masked language modeling (MLM), masked region classification (MRC) and single-step action prediction (SAP). The details of proxy tasks are as follows:

- MLM aims to minimize the negative log-likelihood of original words, the loss function is $L^{\text{MLM}} = -\log\ p(w_i|\mathcal{W}_i^{\text{mask}}, \mathcal{H}_T)$, where $\mathcal{W}_i^{\text{mask}}$ denotes the masked instruction, $\mathcal{H}_T$ denotes the full trajectory with length T.
- MRC aims to minimize the KL divergence between the predicted and target probability distribution, the loss function is $L^{\text{MRC}} = -\sum_{j=1}^{1000} P_{i,j} \log\ \hat{P}_{i,j}$, where $P_i \in \mathbb{R}^{1000}$ is the target class probability for a masked observation.
- SAP aims to minimize negative log probability of the target view action, the loss function is $L_t^{\text{SAP}} = \sum_{t=1}^{T} -\log\ p(a_t^*|\mathcal{W}, \mathcal{P}_t^*)$, where a_t^* is the expert action of a partial demonstration path $\mathcal{P}_t^*$.

During finetuning, we adopt the imitation learning approach, i.e., DAgger, to further optimize the VLN agent.

Inference. At each testing step, the WIA agent determines the room type of the current viewpoint based on its visual observation, while the W2G agent predicts the target room category conditioned on the navigation instruction. The topological map is then updated, and the VLN agent outputs a fused action decision. The agent is forced to terminate once the maximum step limit is reached. At the `[stop]` viewpoint, the VLN agent selects the object with the highest prediction score as the target.

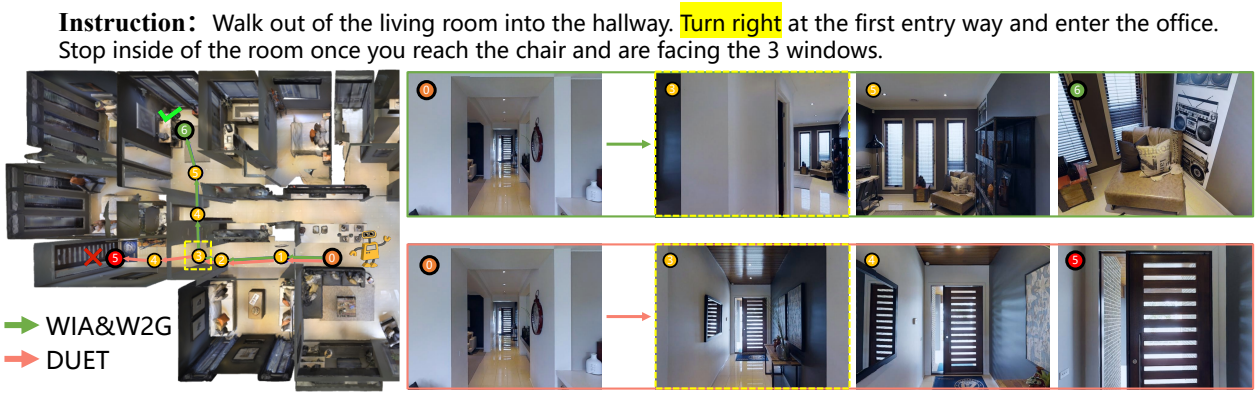

Fig. 4: A representative visual result of WIA&W2G and DUET on val unseen split of R2R. Brown and green circles denote the start and target locations, the red circle represents incorrect endpoint, respectively. According to the instruction, the agent should turn right at the waypoint marked with yellow. WIA&W2G makes the correct decision, while DUET is confused by similar entrance at the waypoint.

IV. Experiment

A. Experiment Settings

Datasets. We primarily centered on the goal-oriented, high-level instruction-driven VLN benchmark, specifically REVERIE. To accurately localize target objects within REVERIE, the agent must possess both fine-grained object grounding capabilities and sophisticated exploration skills. Additionally, we assess our WIA&W2G on the low-level instruction-driven VLN benchmark, i.e., R2R, which involves step-by-step instructions and no object grounding.

The R2R dataset comprises 7,189 trajectories sampled from 90 distinct buildings, including 22k human-crafted navigational instructions. Furthermore, the REVERIE dataset features 4,140 target objects categorized into 489 distinct types, accompanied by 21,702 instructions, each averaging 18 words in length. Both REVERIE and R2R leverage the Matterport3D simulator and encompass 90 photo-realistic residential houses. Each residence is characterized by a distinct collection of navigable viewpoints. Each viewpoints is encapsulated by a crafted skybox with corresponding panorama. All the residential houses are split into *train seen*, *val seen*, *val unseen*, and *test unseen* subsets. Note that the houses in the *val seen* split are the same as in *train seen*, in contrast, the houses in *val unseen* and *test unseen* splits are different from *train seen* split. To avoid information leakage, the WIA and W2G agents are trained solely on the *train seen* subsets, strictly excluding all *val* and *test unseen* houses.

Evaluation Metrics. The performance of VLN agents is assessed in two aspects, i.e., navigation and object grounding. For the navigation sub-task, we have:

- **Success Rate (SR):** the proportion of tasks in which the agent's stopping location is within 3 meters of the target location.
- **Oracle Success Rate (OSR):** the proportion of tasks in which at least one viewpoint along the agent's trajectory can observe the target object within 3 meters.
- **Success weighted by Path Length (SPL):** the main metric for navigation, which balances success rate (SR) with trajectory efficiency, measured by the ratio between the shortest possible path length and the agent's actual trajectory length.

For the object grounding sub-task, we have:

- **Remote Grounding Success (RGS):** the proportion of tasks in which the agent identifies the target object.
- **RGS weighted by Path Length (RGSPL):** the main metric for grounding, which adjusts RGS by incorporating path length efficiency.

For all the above metrics, higher values indicate better performance. Additionally, **Trajectory Length (TL)** represents the average path length of predicted navigation trajectories.

Models	Room to Room (R2R)							
	Val Unseen				Test Unseen			
	TL↓	NE↓	SR↑	SPL↑	TL↓	NE↓	SR↑	SPL↑
Human	–	–	–	–	11.85	1.61	86.4	76.0
Random	9.77	9.23	16.3	–	9.93	9.77	13.2	12.0
Seq2Seq [16]	8.39	7.81	22	–	8.13	7.85	20	18
SF [22]	–	6.62	35	–	14.82	6.62	35	28
EnvDrop [23]	10.70	5.22	52	48	11.66	5.23	51	47
AuxRN [24]	–	5.28	55	50	–	5.15	55	51
Active [25]	20.60	4.36	58	40	21.60	4.33	60	41
ORIST [8]	10.90	4.72	57	51	11.31	5.10	57	52
VLN⟳BERT [13]	12.01	3.93	63	57	12.35	4.09	63	57
HAMT [21]	11.46	2.29	66	61	12.27	3.93	65	60
SOAT [26]	12.15	4.28	59	53	12.26	4.49	58	53
SSM [27]	20.7	4.32	62	45	20.4	4.57	61	46
CCC [28]	–	5.20	50	46	–	5.30	51	48
REM [29]	12.44	3.89	64	58	13.11	3.87	65	59
SEvol [30]	12.26	3.99	62	57	13.40	4.13	62	57
ADAPT [31]	12.33	3.66	66	59	13.16	4.11	63	57
HOP [11]	12.27	3.80	64	57	12.68	3.83	64	59
LANA [32]	12.0	–	68	62	12.6	–	65	60
TD-STP [33]	–	3.22	70	63	–	3.73	67	61
KERM [34]	13.54	3.22	72	60	14.60	3.61	70	60
DUET [10]	13.94	3.31	72	60	14.73	3.65	69	59
WIA&W2G+DUET	13.19	3.18	72	61	14.36	3.63	70	60

TABLE I: Quantitative results on R2R. '–': unavailable statistics.

B. Comparison to State-of-the-Art Methods

REVERIE. To validate the generalization of WIA&W2G, we evaluate its performance when integrated with HAMT and DUET in Table.II. Compared with the previous state-of-the-art agents in VLN, WIA&W2G achieves superior performance in object grounding, yielding a notable improvement on the val unseen split. Moreover, WIA&W2G delivers consistent gains in both navigation and object grounding

Models	Remote Embodied Visual Referring Expression in Real Indoor Environments (REVERIE)											
	Val Unseen						Test Unseen					
	Navigation				Object		Navigation				Object	
	TL↓	OSR↑	SR↑	SPL↑	RGS↑	RGSPL↑	TL↓	OSR↑	SR↑	SPL↑	RGS↑	RGSPL↑
Human	–	–	–	–	–	–	21.18	86.83	81.51	53.66	77.84	51.44
Random	10.76	11.93	1.76	1.01	0.96	–	10.34	8.88	2.30	1.44	1.18	–
Seq2Seq [16] [CVPR2018]	11.07	8.07	4.20	2.84	2.16	1.63	10.89	6.88	3.99	3.09	2.00	1.58
RCM [35] [CVPR2019]	11.98	14.23	9.29	6.97	4.89	3.89	10.60	11.68	7.84	6.67	3.67	3.14
FAST-M [9] [CVPR2020]	45.28	28.20	14.40	7.19	7.84	4.67	39.05	30.63	19.88	11.61	11.28	6.08
SIA [36] [CVPR2021]	41.53	44.67	31.53	16.28	22.41	11.56	48.61	44.56	30.80	14.85	19.02	9.20
VLN⟳BERT [13] [CVPR2021]	16.78	35.02	30.67	24.90	18.77	15.27	15.86	32.91	29.61	23.99	16.50	13.51
Airbert [37] [ICCV2021]	18.71	34.51	27.89	21.88	18.23	14.18	17.91	34.20	30.28	23.61	16.83	13.28
HOP [11] [CVPR2022]	16.46	36.24	31.78	26.11	18.85	15.73	16.38	33.06	30.17	24.34	17.69	14.34
TD-STP [33] [MM2022]	–	39.48	34.88	27.32	21.16	16.56	–	40.26	35.89	27.51	19.88	15.40
KERM [34] [CVPR2023]	21.85	55.21	50.44	35.38	34.51	24.45	17.32	57.58	52.43	39.21	32.39	23.64
Lily [38] [ICCV2023]	21.87	53.71	48.11	34.43	32.15	23.43	21.94	60.51	54.32	37.34	32.02	21.94
BEVBert [12] [ICCV2023]	–	56.40	51.78	36.37	34.71	24.44	–	57.26	52.81	36.41	32.06	22.09
GridMM [39] [ICCV2023]	23.20	57.48	51.37	36.47	34.57	24.56	19.97	59.55	53.13	36.60	34.87	23.45
LANA [32] [ICCV2023]	23.18	52.97	48.31	33.86	32.86	22.77	18.83	57.20	51.72	36.45	32.95	22.85
BSG [40] [ICCV2023]	24.71	58.05	52.12	35.59	35.36	24.24	22.90	62.83	56.45	38.70	33.15	22.34
ENP [41] [NeurIPS2024]	25.76	54.70	48.90	33.78	34.74	23.39	22.70	59.38	53.19	36.26	33.10	22.14
VER [5] [CVPR2024]	23.03	61.09	55.98	39.66	33.71	23.70	24.74	62.22	56.82	38.76	33.88	23.19
SAME [6] [ICCV2025]	18.87	–	46.35	36.12	–	–	19.47	–	48.60	37.10	–	–
HAMT [21] [NeurIPS2021]	14.08	36.84	32.95	30.20	18.92	17.28	13.62	33.41	30.40	26.67	14.88	13.08
WIA&W2G+HAMT (Ours)	13.73	37.43	34.22	30.82	19.97	17.74	15.04	35.78	31.91	26.95	15.35	13.69
DUET [10] [CVPR2022]	22.11	51.07	46.98	33.73	32.15	23.03	21.30	56.91	52.51	36.06	31.88	22.06
WIA&W2G+DUET (Ours)	27.43	60.10	54.76	36.81	**37.63**	**25.12**	25.17	**63.62**	**57.60**	**39.34**	**35.62**	**24.17**

TABLE II: Quantitative comparison results on REVERIE [9]. '–': unavailable statistics. Adding our WIA&W2G to HAMT and DUET (gray rows) leads to improved navigation metrics (i.e., SR and SPL) and object grounding metrics (i.e., RGS and RGSPL) on both val and test unseen split.

Modules			Navigation			Object	
BASE.	WIA.	W2G.	OSR↑	SR↑	SPL↑	RGS↑	RGSPL↑
✓			51.07	46.98	33.73	32.15	23.03
✓	✓		55.55	52.40	35.35	35.53	23.72
✓		✓	55.15	50.72	33.83	33.37	22.04
✓	✓	✓	**60.10**	**54.76**	**36.81**	**37.63**	**25.12**

TABLE III: Ablation study of overall scheme on *val unseen* split of REVERIE [9] compared with navigation metrics (i.e., OSR, SR, SPL) and object grounding metrics (i.e., RGS, RGSPL).

Base	Policy	REVERIE *val*		REVERIE *test*	
		RGS↑	RGSPL↑	RGS↑	RGSPL↑
DUET	ENP [41] [NeurIPS2024]	34.74	23.39	33.10	22.14
	G-Mate [42] [WACV2025]	34.11	24.45	34.62	24.08
	WIA&W2G (Ours)	**37.63**	**25.12**	**35.62**	**24.17**

TABLE IV: Comparison with previous policy add to DUET on REVERIE dataset in object grounding performance.

in test unseen split. For instance, on the val unseen split, WIA&W2G outperforms the previous best VLN agent, i.e., BSG, in object grounding (2.27%↑ on RGS, 0.88%↑ on RGSPL). On the test unseen split, WIA&W2G demonstrates better generalization, surpassing VER in both navigation (0.78%↑ on SR, 0.58%↑ on SPL) and object grounding (1.74%↑ on RGS, 0.98%↑ on RGSPL). The results achieved by WIA&W2G highlight its superior capability in accomplishing the task with improved accuracy, thereby underscoring the effectiveness of spatial awareness in VLN. It is worth emphasizing that all aforementioned VLN agents are trained solely on the original REVERIE dataset, without utilizing any augmented VLN trajectory–instruction pairs.

R2R. Table.I presents the comparative results on the R2R. It is worth noting that, although WIA&W2G achieves substantial improvements on the REVERIE, the gains observed on the R2R are relatively limited. In REVERIE, high-level instructions characterized by concise vocabulary align more naturally with expressions commonly encountered in daily household scenarios. In contrast, the low-level instructions in R2R are more detailed, embedding extensive and detailed descriptions of the trajectory. Consequently, WIA&W2G shows considerable potential in household environment, as it is well-suited to handle abstract and simple instructions.

C. Navigation Case Visualization

R2R. As illustrated in Fig.4, precise action selection at critical decision points is essential for faithfully executing the instruction, which specifies, *"Turn right at the first entryway and enter the office."* At the waypoint marked in yellow (step ③), the agent must decide between two visually similar entrances. WIA&W2G correctly interprets the instruction and executes the right turn, leading to the designated office and ultimately reaching the correct endpoint (step ⑥). In contrast, DUET misinterprets the scene context, choosing an incorrect entrance that results in deviation from the intended path and termination at an erroneous endpoint (step ⑤). This example highlights the importance of accurate grounding at ambiguous waypoints and demonstrates the robustness of WIA&W2G in handling visually confounding scenarios commonly encountered in household environments.

REVERIE. As illustrated in Fig.5, several representative examples of WIA&W2G and DUET model are shown on the val unseen split of the REVERIE. In the first row, under a relatively simple route, WIA&W2G correctly identifies the target landmark and successfully stops at the designated location, whereas DUET deviates and fails to reach the correct endpoint. In the second row, for a trajectory requiring

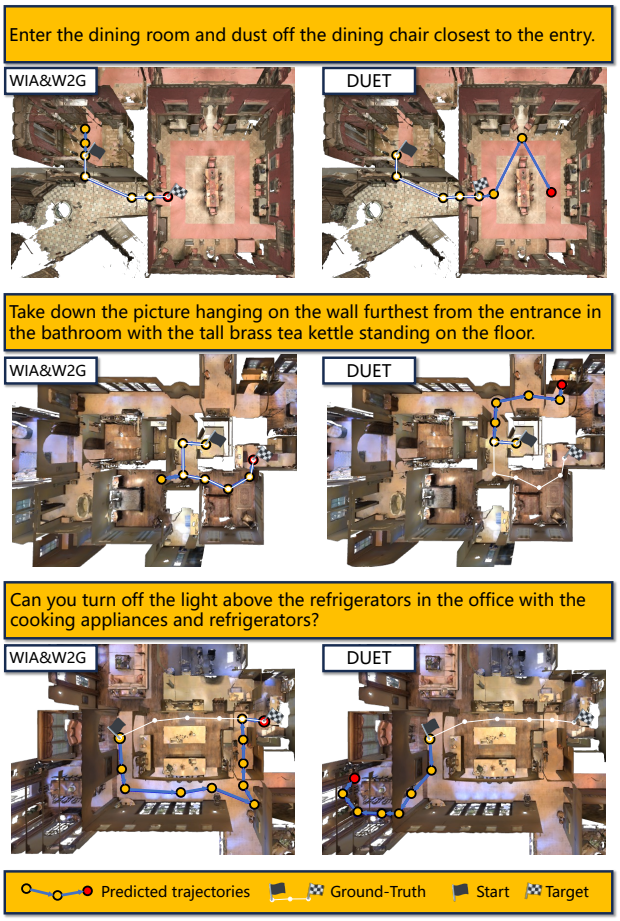

Fig. 5: Predicted trajectories of WIA&W2G and the state-of-the-art DUET on REVERIE val unseen split. The black and checkered flags denote start and target locations, respectively. The white lines denote ground-truth trajectories. WIA&W2G is able to make more efficient explorations and correct its previous decisions.

multiple turns, WIA&W2G selects the correct initial direction. Although some exploratory movements occur along the way, it eventually arrives at the correct target. In contrast, DUET makes an incorrect early turn, leading to task failure. In the third row, both agents initially explore in the wrong direction. Nevertheless, WIA&W2G efficiently corrects its trajectory and successfully reaches the target, while DUET continues to diverge from the intended path. These examples demonstrate the robustness of WIA&W2G in handling complex navigation instructions and its ability to recover from suboptimal exploration.

D. Diagnostic Experiment

In Table.III, we present an diagnostic experiment on the val unseen split of the REVERIE dataset to evaluate the contributions of each module. Compared with the baseline DUET, incorporating the **WIA** module significantly improves performance on both navigation (0.78%↑ on SR, 0.58%↑ on SPL) and object grounding (2.27%↑ on RGS, 0.88%↑ on RGSPL). This improvement can be attributed to WIA's ability to enhance the agent's spatial perception in complex household environments. Furthermore, when the **W2G** module is added, the performance surpasses the WIA-only setting across all metrics. The gain primarily benefits from W2G's grounding-aware guidance, which enables the agent to resolve visual ambiguities and locate target objects more efficiently during navigation. By combining both modules with the baseline DUET, the final system achieves substantial gains in both navigation (7.78%↑ on SR, 3.08%↑ on SPL) and object grounding (5.48%↑ on RGS, 2.09%↑ on RGSPL), outperforming the baseline by a large margin.

Table.IV compares our proposed WIA&W2G module with recent policies added to DUET on the REVERIE dataset. On the val unseen split, WIA&W2G achieves 37.63% RGS and 25.12% RGSPL, notably outperforming ENP and G-Mate. On the test unseen split, WIA&W2G continues to demonstrate superior generalization, achieving 35.62% RGS and 24.17% RGSPL, again exceeding previous policies. These results not only confirm the robustness of our method in object grounding but also highlight its overall advantage over existing approaches.

V. Conclusion and Future Work

In this work, we presented a model-agnostic interaction framework with two auxiliary agents, Where-I-Am (WIA) and Where-to-Go (W2G), to address the challenges of the REVERIE task. By explicitly modeling room awareness and goal inference, our method enhances both navigation planning and remote object grounding. The proposed framework is plug-and-play and can be integrated with various VLN backbones, consistently improving performance without requiring additional training data. Extensive experiments on the REVERIE benchmark demonstrate significant gains in navigation success rate and grounding accuracy, surpassing strong baselines and the previous state-of-the-art. Moreover, our approach shows strong generalization ability in unseen environments, highlighting the benefit of incorporating spatial and semantic cues into VLN agents. Future work will explore extending this interactive framework to VLN-CE settings and integrating it with large language models for more general embodied reasoning.

References

[1] Z. Gong, T. Hu, R. Qiu, and J. Liang, "From cognition to precognition: A future-aware framework for social navigation," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*, 2025, pp. 9122–9129.

[2] H. Zhang, Y. Qi, M. Zhao, and Y. Liu, "Indoor scene recognition in vision-and-language navigation," *IEEE Transactions on Consumer Electronics*, pp. 1–1, 2025.

[3] J. Huang, H. Zhang, M. Zhao, Z. Wu, and Y. Liu, "Instance-aware visual language grounding for consumer robot navigation," *IEEE Transactions on Consumer Electronics*, vol. 71, no. 4, pp. 12 519–12 526, 2025.

[4] H. Zhang, G. Zhang, M. Zhao, and Y. Liu, "Load forecasting-based learning system for energy management with battery degradation estimation: A deep reinforcement learning approach," *IEEE Transactions on Consumer Electronics*, vol. 70, no. 1, pp. 2342–2352, 2024.

[5] R. Liu and Y. Yang, "Volumetric environment representation for vision-language navigation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 16 317–16 328.

[6] G. Zhou, Y. Hong, Z. Wang, C. Zhao, M. Bansal, and Q. Wu, "Same: Learning generic language-guided visual navigation with state-adaptive mixture of experts," *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

[7] J. Zhu, Y. Qiao, S. Zhang, X. He, Q. Wu, and J. Liu, "Minivln: Efficient vision-and-language navigation by progressive knowledge distillation," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*, 2025, pp. 97–103.

[8] Y. Qi, Z. Pan, Y. Hong, M.-H. Yang, A. Van Den Hengel, and Q. Wu, "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*, 2021, pp. 1655–1664.

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

[10] S. Chen, P.-L. Guhur, M. Tapaswi, C. Schmid, and I. Laptev, "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*, 2022, pp. 16 537–16 547.

[11] Y. Qiao, Y. Qi, Y. Hong, Z. Yu, P. Wang, and Q. Wu, "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*, 2022, pp. 15 418–15 427.

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

[13] Y. Hong, Q. Wu, Y. Qi, C. Rodriguez-Opazo, and S. Gould, "Vln bert: A recurrent vision-and-language bert for navigation," in *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, 2021, pp. 1643–1653.

[14] A. Chang, A. Dai, T. Funkhouser, M. Halber, M. Niebner, M. Savva, S. Song, A. Zeng, and Y. Zhang, "Matterport3d: Learning from rgb-d data in indoor environments," in *2017 International Conference on 3D Vision (3DV)*. IEEE, 2017, pp. 667–676.

[15] Y. Long, X. Li, W. Cai, and H. Dong, "Discuss before moving: Visual language navigation via multi-expert discussions," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 17 380–17 387.

[16] P. Anderson, Q. Wu, D. Teney, J. Bruce, M. Johnson, N. Sünderhauf, I. Reid, S. Gould, and A. Van Den Hengel, "Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3674–3683.

[17] M. Li, Z. Wang, T. Tuytelaars, and M.-F. Moens, "Layout-aware dreamer for embodied visual referring expression grounding," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 1, 2023, pp. 1386–1395.

[18] Y. Qiao, Y. Qi, Z. Yu, J. Liu, and Q. Wu, "March in chat: Interactive prompting for remote embodied referring expression," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 15 712–15 721.

[19] A. Majumdar, A. Shrivastava, S. Lee, P. Anderson, D. Parikh, and D. Batra, "Improving vision-and-language navigation with image-text pairs from the web," in *European Conference on Computer Vision*. Springer, 2020, pp. 259–274.

[20] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6077–6086.

[21] S. Chen, P.-L. Guhur, C. Schmid, and I. Laptev, "History aware multimodal transformer for vision-and-language navigation," *Advances in neural information processing systems*, vol. 34, pp. 5834–5847, 2021.

[22] D. Fried, R. Hu, V. Cirik, A. Rohrbach, J. Andreas, L.-P. Morency, T. Berg-Kirkpatrick, K. Saenko, D. Klein, and T. Darrell, "Speaker-follower models for vision-and-language navigation," *Advances in neural information processing systems*, vol. 31, 2018.

[23] H. Tan, L. Yu, and M. Bansal, "Learning to navigate unseen environments: Back translation with environmental dropout," *arXiv preprint arXiv:1904.04195*, 2019.

[24] F. Zhu, Y. Zhu, X. Chang, and X. Liang, "Vision-language navigation with self-supervised auxiliary reasoning tasks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 012–10 022.

[25] H. Wang, W. Wang, T. Shu, W. Liang, and J. Shen, "Active visual information gathering for vision-language navigation," in *European conference on computer vision*. Springer, 2020, pp. 307–322.

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

[27] H. Wang, W. Wang, W. Liang, C. Xiong, and J. Shen, "Structured scene memory for vision-language navigation," in *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, 2021, pp. 8455–8464.

[28] H. Wang, W. Liang, J. Shen, L. Van Gool, and W. Wang, "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*, 2022, pp. 15 471–15 481.

[29] C. Liu, F. Zhu, X. Chang, X. Liang, Z. Ge, and Y.-D. Shen, "Vision-language navigation with random environmental mixup," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 1644–1654.

[30] J. Chen, C. Gao, E. Meng, Q. Zhang, and S. Liu, "Reinforced structured state-evolution for vision-language navigation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 15 450–15 459.

[31] B. Lin, Y. Zhu, Z. Chen, X. Liang, J. Liu, and X. Liang, "Adapt: Vision-language navigation with modality-aligned action prompts," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 15 396–15 406.

[32] X. Wang, W. Wang, J. Shao, and Y. Yang, "Lana: A language-capable navigator for instruction following and generation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 19 048–19 058.

[33] Y. Zhao, J. Chen, C. Gao, W. Wang, L. Yang, H. Ren, H. Xia, and S. Liu, "Target-driven structured transformer planner for vision-language navigation," in *Proceedings of the 30th ACM international conference on multimedia*, 2022, pp. 4194–4203.

[34] X. Li, Z. Wang, J. Yang, Y. Wang, and S. Jiang, "Kerm: Knowledge enhanced reasoning for vision-and-language navigation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 2583–2592.

[35] X. Wang, Q. Huang, A. Celikyilmaz, J. Gao, D. Shen, Y.-F. Wang, W. Y. Wang, and L. Zhang, "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*, 2019, pp. 6629–6638.

[36] X. Lin, G. Li, and Y. Yu, "Scene-intuitive agent for remote embodied visual grounding," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 7036–7045.

[37] P.-L. Guhur, M. Tapaswi, S. Chen, I. Laptev, and C. Schmid, "Airbert: In-domain pretraining for vision-and-language navigation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 1634–1643.

[38] K. Lin, P. Chen, D. Huang, T. H. Li, M. Tan, and C. Gan, "Learning vision-and-language navigation from youtube videos," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 8317–8326.

[39] Z. Wang, X. Li, J. Yang, Y. Liu, and S. Jiang, "Gridmm: Grid memory map for vision-and-language navigation," in *Proceedings of the IEEE/CVF International conference on computer vision*, 2023, pp. 15 625–15 636.

[40] R. Liu, X. Wang, W. Wang, and Y. Yang, "Bird's-eye-view scene graph for vision-language navigation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 10 968–10 980.

[41] R. Liu, W. Wang, and Y. Yang, "Vision-language navigation with energy-based policy," *Advances in Neural Information Processing Systems*, vol. 37, pp. 108 208–108 230, 2024.

[42] Q. Liu, S. Zhang, Y. Qiao, J. Zhu, X. Li, L. Guo, Q. Wang, X. He, Q. Wu, and J. Liu, "Groundingmate: Aiding object grounding for goal-oriented vision-and-language navigation," in *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2025, pp. 1775–1784.