2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

Enhanced Language-guided Robot Navigation with Panoramic Semantic Depth Perception and Cross-modal Fusion

Liuyi Wang*, Jiagui Tang*, Zongtao He, Ronghao Dang, Chengju Liu, and Qijun Chen

***Abstract*— Integrating visual observation with linguistic instruction holds significant promise for enhancing robot navigation across unstructured environments and enriches the human-robot interaction experience. However, while panoramic RGB views furnish robots with extensive environmental visuals, current methods significantly overlook crucial semantic and depth cues. This incomplete representation may lead to misinterpretation or inadequate execution of language instructions, thereby impeding navigation performance and adaptability. In this paper, we introduce SEAT, a semantic-depth aware cross-modal transformer model. Our approach incorporates an efficient panoramic multi-type visual encoder to capture comprehensive environmental details. To mitigate the rigidity of feature mapping stemming from the freezing of pre-training encoders, we propose a novel region query pre-training task. Additionally, we leverage an improved dual-scale cross-modal transformer to facilitate the integration of instructions, topological memory, and action prediction. Extensive experiments on three language-guided robot navigation datasets demonstrate the efficacy of our model, achieving competitive navigation success rates with fewer parameters and computational load. Furthermore, we validate SEAT's effectiveness in real-world scenarios by deploying it on a mobile robot across various environments. The code is available at `https://github.com/CrystalSixone/SEAT`.**

I. Introduction

Advancements in language-guided robot navigation represent significant progress within embodied artificial intelligence (E-AI), driving the development of autonomous agents that understand and interact with their surroundings in a manner akin to humans. The fundamental task of this field is enabling robotic systems to navigate efficiently using verbal commands [1]. Such navigation demands a sophisticated fusion of linguistic comprehension and visual perception, allowing robots to autonomously interpret and act on complex instructions with high precision.

The task of language-guided robot navigation, foundational to the advancement of E-AI, has garnered widespread attention and development since its inception [2], [3]. Despite the growth, the field continues to face significant challenges. One primary concern is the *inadequate understanding and interpretation of the visual environment*. Traditional methods merely depend on RGB features to encode visual cues [4], [5], [6], a stark contrast to the multifaceted perception humans possess innately, encompassing semantic and depth cues for a comprehensive environmental grasp. While there have been efforts to integrate additional visual features [7], [8], [9], the essential task of harmonizing these diverse visual inputs, particularly in panoramic views, has not been adequately addressed. This shortfall is exacerbated by the widespread use of static, pre-trained visual feature extractors. While this operation is beneficial for enhancing training efficiency, it can result in the rigidity of feature mappings and challenges in adapting to new environments. Another pressing issue is the *inefficiency in long-term cross-modal inference*. The success of a navigation system lies in its ability to forge and maintain coherent connections between multiple modalities throughout navigation. However, existing methods, relying on soft attention mechanisms [10], [11] or insufficient pre-training strategies [12], [13], fail to secure a robust alignment between modalities. This misalignment results in inconsistent representations, failing to accurately correlate visual attention with corresponding language instructions. Moreover, the tendency to employ all layers of pre-trained cross-modal models without customization leads to parameter redundancy, which not only inflates the model size but also increases the risk of overfitting, especially when training with limited navigation datasets.

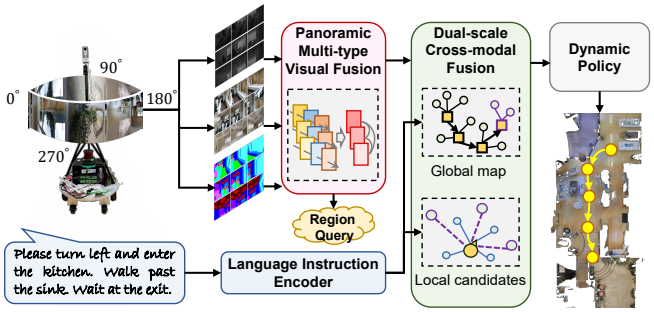


Fig. 1: Proposed SEAT model. By injecting panoramic multi-type visual features and fusing dual-scale vision-language modalities, we aim to improve the capability of perception and inference for language-guided robot navigation.

In this paper, we propose SEAT (Fig. 1), a Semantic-dEpth Aware cross-modal Transformer model, aimed at addressing the above challenges in language-guided robot navigation by substantially augmenting environmental perception and cross-modal fusion. Through in-depth discussion and experimental validation, SEAT employs an effective panoramic multi-type visual (PMV) encoder that simultaneously fuses RGB, depth, and semantics to enhance perceptual rich-

* Equal Contribution

This paper is supported by the National Natural Science Foundation of China under Grants (62073245, 62173248, 62333017). Shanghai Science and Technology Innovation Action Plan (22511104900). Shanghai Municipal Science and Technology Major Project (2021SHZDZX0100) and the Fundamental Research Funds for the Central Universities. (Corresponding author: Chengju Liu, Qijun Chen) (E-mails: {wly, liuchengju, qjchen}@tongji.edu.cn)

ness and cross-domain understanding. By enabling PMV to distinguish the semantic content of the provided region, a region query (RQ) pre-training task is proposed to enhance PMV adaptability and reduce the feature discrepancies and mapping rigidity caused by the frozen visual extractors. For seamlessly integrating language instructions, global topological memory, and local action prediction, an improved co-attention-based dual-scale cross-modal transformer is introduced. We present a comprehensive evaluation of SEAT's performance on three public datasets (R2R [1], REVERIE [2], and SOON [3]), demonstrating its superiority over existing methods. Furthermore, we validate our method on a real mobile robot in various real-world scenarios.

II. Related Work

As societal expectations for versatile service robots grow, existing navigation techniques face increasing pressure to adapt to expanding application scenarios, highlighting the need for more robust navigation solutions, especially in natural interaction with humans [14], [15], [16]. Anderson et al. [1] proposed vision-and-language navigation with fine-grained step-by-step instructions based on the real-3D indoor simulator MP3D [17]. Subsequently, Qi et al. [2] proposed the remote embodied visual referring expression in real indoor environments. Zhu et al. [3] proposed scenario-oriented language object navigation. In this paper, we refer to these tasks collectively as *language-guided robot navigation*.

Initially, researchers employed recurrent neural networks with single RGB observations [10], [18]. Subsequently, panoramic images were recognized as effective inputs for a broader visual scope at each viewpoint [19], [20]. The evolution to transformer-based methods [21] marked a significant leap in navigation performance enhancement, which includes the transition from simple recurrent models [4] to more complex graph-structured frameworks [6], [22], [23], signaling notable progress in the field.

Our research builds upon these innovations, specifically the graph-based DUET model [6]. Various improvements are introduced: Firstly, inspired by advancements in visual depth and semantic estimation [24], [25], we propose to incorporate multiple types of visual features to enhance perception capabilities. Secondly, leveraging insights from established pre-training tasks such as MLM [26], SAP [5], and OG [27], we introduce a region-query task aimed at mitigating feature mapping rigidity and enhancing the fusion of semantic and depth information. Thirdly, to address the compatibility gap between the object-level visual features used in previous cross-modal backbone LXMERT [13] and the more recent CLIP-based features [28], we adopt the advanced fully-transformer METER model [29] for cross-modal fusion. Furthermore, we explore the impact of initialization layer numbers on performance. Moreover, compared to some other methods using depth or semantic features [7], [30], [31], our approach offers several advantages: simplicity in preprocessing, comprehensive integration of panoramic RGB, semantic, and depth features, and achieving superior results with fewer parameters.

III. Methodology

A. Problem setup

In the language-guided robot navigation task, an agent follows a given language instruction $X = \{x_i\}_{i=1}^{L}$. Following the setup in previous research [19], at each step t, the agent perceives its surroundings through a 360-degree panoramic view, which is discretized into 36 sub-images across 12 headings and 3 elevations at 30-degree intervals, denoted as $O = \{o_i\}_{i=1}^{N}$. The pose $\mathcal{P}_t = \langle \theta_t, \varphi_t \rangle$ represents the agent's heading and elevation. Each environment is represented as a connected graph $G = \{P, \xi\}$, where P denotes navigable points and ξ denotes edges. The agent selects a nearby navigable candidate point or concludes navigation when it believes it has reached the target location. Navigation is successful if the robot stops near the target point within a certain threshold.

B. Panoramic Multi-type Visual Encoder

While panoramas have proven effective in providing detailed representations of surrounding environments [19], integrating depth and semantic information into panoramic perception poses an ongoing challenge. In simulations based on MP3D [17], depth and semantics can be acquired through queries. In real-world scenarios, thanks to the advancements of corresponding visual estimation methods [24], [25], the semantics and depth can be easily estimated without extra sensors. For convenience, we use $R_t \in \mathbb{R}^{N \times D_r}$, $E_t \in \mathbb{R}^{N \times D_d}$, and $S_t \in \mathbb{R}^{N \times D_s}$ to present RGB, depth, and semantic features, respectively. CLIP [28] trained by the natural language supervision is used for extracting RGB features, and ResNet-50 [32] trained to perform the point-goal navigation [33] is used for depth features. Considering the blurred boundaries of semantic segmentation and the memory constraints posed by storing individual semantic segmentation features, we translate segmentation information into semantic labels $S_t \in \mathbb{R}^{42}$. To denote object presence in images, $\mathcal{S}_t[i] = 1$ signifies the appearance of the i-th category, while $\mathcal{S}_t[i] = 0$ indicates its absence. To explore the potential effective mechanism for the panoramic multi-type visual (PMV) encoder, we design three kinds of structures:

Type-1: SepFFN. Different from typical visual tasks, the difficulty of this task lies in the extraction and fusion of multi-types of panoramic representation with self-centered vision, which is divided into $N = 36$ sub-images. One straightforward way is to use separate feed-forward networks (FFN) to unify the dimensions, and then concatenate different components in the same locations. As shown in the left part of Fig. 2, the derived RGB image R_t, depth E_t, and semantic S_t features are projected to a unified dimension D_h via independent FFN denoted as $\psi(\cdot)$. To increase nonlinearity and diversity, the ReLU and dropout functions (denoted as $\kappa(\cdot)$) are employed. Then, we concatenate these features (denoted as $[,]$ for concatenation along the length dimension, and $[;]$ for concatenation along the feature dimension) and use another FFN to transform them from $3D_h$ to D_h. The layer normalization is employed for regularization. The

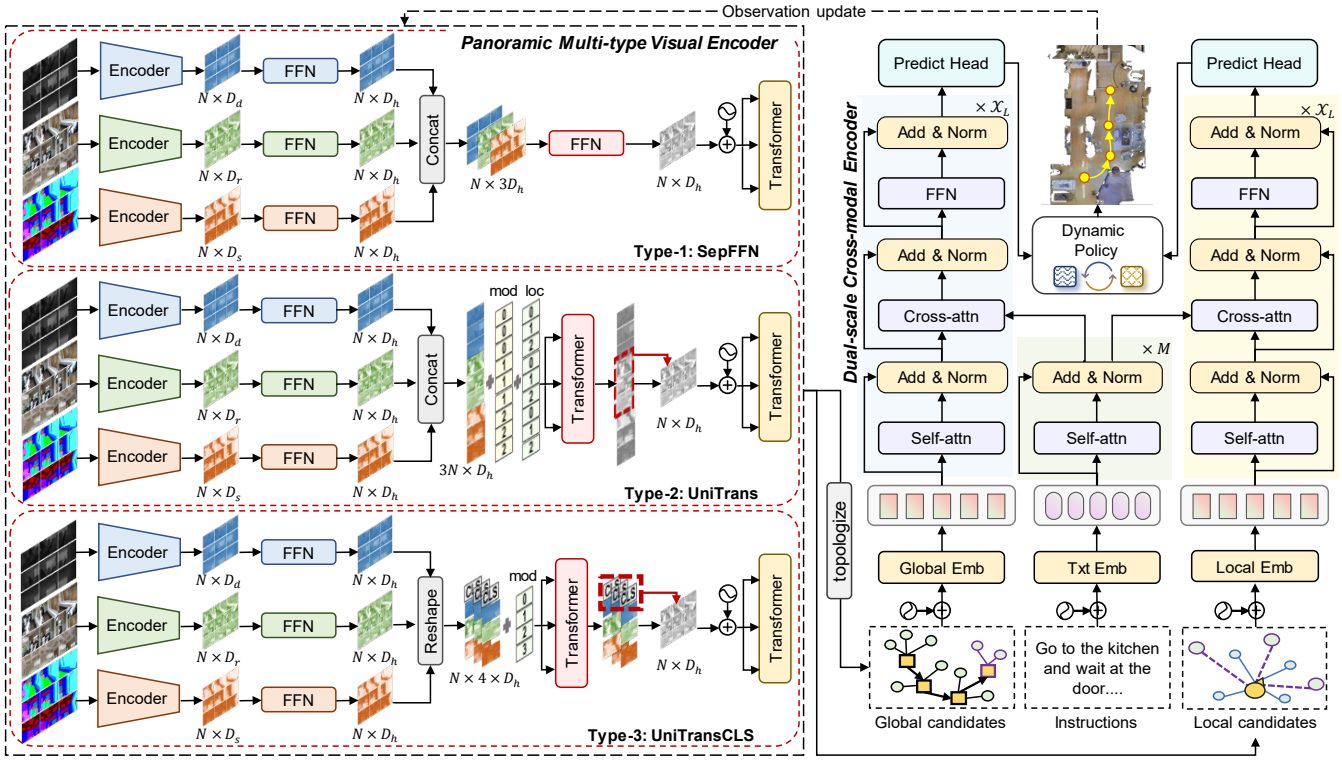


Fig. 2: Overview of SEAT. A panoramic multi-type visual (PMV) encoder is used to fuse the RGB, semantic, and depth features, followed by a dual-scale cross-modal encoder for integrating the vision, language, and history for decision-making.

above process is formulated as follows:

$$\mathcal{R}_t = \kappa(\psi_r(R_t)),\ \mathcal{E}_t = \kappa(\psi_d(E_t)),\ \mathcal{S}_t = \kappa(\psi_s(S_t)) \quad (1)$$

$$\mathcal{V}_t = LN(\psi_o([\mathcal{R}_t; \mathcal{D}_t; \mathcal{S}_t])) \quad (2)$$

The panoramic spatial encoding is achieved by a transformer encoder $T_p(\cdot)$ [21], with the addition of the sinusoidal heading and elevation embeddings φ_p for each sub-image:

$$\mathcal{P}_t = \varphi_p(\sin\tilde{\theta}_t, \cos\tilde{\theta}_t, \sin\tilde{\phi}_t, \cos\tilde{\phi}_t) \quad (3)$$

$$\mathcal{O}_t = T_p(\mathcal{V}_t + \mathcal{P}_t) \quad (4)$$

The transformer encoder T_p includes several multi-head self-attention (MSA) blocks, processing input X as follows:

$$h_i = \text{Softmax}(\frac{XW_q(XW_k)^T}{\sqrt{d_k}})XW_v \quad (5)$$

$$\text{MSA}(X) = [h_1; ...; h_H]W_h + X \quad (6)$$

where W presents the learnable weight, h_i denotes the output of the i-th attention head. The scaling factor d_k is equal to 64, and H, the total number of heads, is set to 12.

Type-2: UniTrans. As sub-images for each panorama inherently form a sequence, an alternative fusion approach involves utilizing a unified transformer encoder to process different types of vision features. Specifically, we concatenate R_t, E_t, and S_t along the sub-image dimension, resulting in $3N \times D_h$ features. Subsequently, we employ distinct embedding functions $\varphi(\cdot)$ for different modalities B_m (where 0 denotes R_t, 1 denotes E_t, and 2 denotes S_t) and locations B_l (ranging from 0 to $N-1$) to differentiate between modalities and locations for each token. Following this, a single transformer encoder $T_m(\cdot)$ is applied to encode these multi-type visual features. The resulting enhanced RGB tokens are then extracted as the primary fused visual features:

$$\mathcal{V}_t = T_m([\mathcal{R}_t, \mathcal{E}_t, \mathcal{S}_t] + \varphi_m(B_m) + \varphi_l(B_l)) \quad (7)$$

The spatial encodings across different angles for panoramic visual fusion follow the formulations in Eq.(3) and Eq.(4).

Type-3: UniTransCLS. The prior method exchanges information across various locations when merging multiple visual features, with RGB tokens as the primary output, potentially causing confusion and bias in representation per sub-image. Therefore, we further propose UniTransCLS. This model reshapes the length of sub-images N to the batch size dimension, which becomes $\mathbb{R}^{N\times1\times D_h}$, and concatenates the features along the length dimension, with the addition of [CLS] token at the sequence start. This adjustment standardizes each sequence to four tokens, ensuring equal treatment of sequences from different locations. Consequently, the transformer focuses on fusing visual features without information leakage from other locations:

$$\mathcal{V}_t = T_m([\texttt{[CLS]}, \mathcal{R}'_t, \mathcal{E}'_t, \mathcal{S}'_t] + \varphi_m(B_m)) \quad (8)$$

Subsequently, [CLS] tokens from different locations are extracted to represent the fused features. These tokens are then passed to the panoramic spatial transformer, following the same process as described in the preceding methods.

C. Region Query Pre-training Task

Given the frozen state of raw encoders pre-trained on large datasets to enhance training efficiency, the subsequent concern is how to enhance the adaptability of our learnable PMV encoder in terms of semantic understanding and spatial awareness. To address this challenge, we introduce

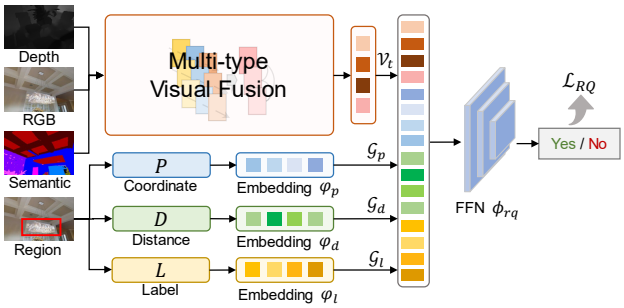


Fig. 3: Illustration of the region query (RQ) task.

the Region Query (RQ) pre-training task to improve the encoder's visual inference capability. As shown in Fig. 3, this task enables the agent to discern whether a provided region contains the content of the designated label, outputting a "yes" or "no" prediction. For the sampled region, we record its coordinates p, distance d, and one of its semantic labels l to generate positive samples. To construct negative samples, we first randomly select the category labels. If the chosen label exists in the image, we displace coordinates beyond the actual region and randomize distance. Otherwise, both coordinates and distance are randomized. The number of positive and negative samples account for 50% respectively.

The embedding functions $\varphi(\cdot)$ are used to capture the location, distance, and label features. Then, these features are concatenated with the fused multi-type features $\mathcal{V}_t$ per image. The specific FFN network ψ_{rq}, comprising three blocks of full connection, dropout, and ReLU layers, functions as the dedicated prediction head. The sigmoid function $\sigma(\cdot)$ is utilized to predict y within the range of 0 to 1:

$$\mathcal{G}_p = \varphi_p(p),\ \mathcal{G}_l = \varphi_l(l),\ \mathcal{G}_d = \varphi_d(d) \tag{9}$$

$$y = \sigma(\psi_{rq}([\mathcal{V}_t; \mathcal{G}_p; \mathcal{G}_l; \mathcal{G}_d])) \tag{10}$$

Let $\hat{y}_i$ and y_i denote the ground truth and predicted result. The binary cross-entropy loss is used for optimization:

$$\mathcal{L}_{RQ} = -\frac{1}{B}\sum_{i=0}^{B}[\hat{y}_i \cdot log(y_i) + (1-\hat{y}_i)\cdot log(1-y_i)] \tag{11}$$

where B is the number of samples in a batch.

D. Dual-scale Cross-modal Fusion

1) Text Encoder: Besides the visual features, the agent also needs to learn linguistic guidance from raw language input. We introduce RoBERTa [34] to extract text features in end-to-end training. Compared with earlier methods relying on BiLSTM [10] or BERT [6], [35], RoBERTa has exhibited enhanced robustness and broader applicability across diverse downstream tasks. The absolute positional encoding P_x [21] is added for presenting sequence information:

$$\mathcal{X} = \text{RoBERTa}(X + P_x) \tag{12}$$

2) Dual-scale Cross-modal Encoder: Long-distance cross-modal fusion is pivotal in language-guided robot navigation. Building on [6], we use a dual-scale cross-modal encoder comprising a *local branch* for merging fine-grained features $\mathcal{V}_t^L$ from the current viewpoint and instructions, alongside a *global branch* that topologizes coarse-grained features $\mathcal{V}_t^G$ from potential candidates across all visited nodes. Unlike previous methods that relied on LXMERT [13] with object-level image embeddings [36] as the cross-modal foundation, we address the compatibility issue with CLIP-based image features [28] by adopting METER [29], a more advanced fully-transformer vision-and-language backbone. As depicted in Fig. 2, we employ co-attention blocks for effective cross-modal fusion:

$$\mathcal{F}_t^L = \text{CoAttn}(\mathcal{V}_t^L, \mathcal{X}, \mathcal{X}),\ \mathcal{F}_t^G = \text{CoAttn}(\mathcal{V}_t^G, \mathcal{X}, \mathcal{X}) \tag{13}$$

Lastly, the dynamic fusion strategy is employed for prediction. The FFN $\psi(\cdot)$, followed by the sigmoid function $\sigma(\cdot)$, is used to obtain the predicted scores ω by calculating the concatenation of the `[CLS]` tokens from the local and global branches (denoted as $\mathcal{U}_t^L$ and $\mathcal{U}_t^G$). The results from the two branches are then weighted and combined. The softmax function is used to yield the final prediction p_t:

$$\omega = \sigma(\psi_u([\mathcal{U}_t^L; \mathcal{U}_t^G])) \tag{14}$$

$$p_t = \text{Softmax}(\omega\psi_l(\mathcal{F}_t^L) + (1-\omega)\psi_g(\mathcal{F}_t^G)) \tag{15}$$

Let $\hat{p}_t$ be the ground truth action for step t. The cross-entropy loss is used to supervise action prediction:

$$\mathcal{L}_{AP} = -\sum_{t=1}^{T}\hat{p}_t \log(p_t) \tag{16}$$

Furthermore, we assess the impact of varying the number of layers in both the text encoder and cross-modal encoder, as it may influence learning performance [37]. Experimentally, we find that opting for fewer layers instead of employing the complete pre-trained large model can yield enhanced accuracy and generalization with fewer parameters.

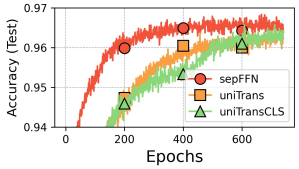
E. Training Strategy

The training process comprises two stages: *pre-training* and *fine-tuning*. *Pre-training* has emerged as a potent approach for enhancing the learning of transformer-based models [12], [38], [39]. We first pre-train the PMV encoder using the RQ task. For further task-specific pre-training, we initialize the network with pre-trained weights from both vision-and-language model METER [29] and PMV. The masked language modeling (MLM) [40] and single-step action prediction (SAP) [5] tasks are employed on R2R, while object grounding (OG) [27] is additionally employed for REVERIE and SOON. In the *fine-tuning* stage, the teacher-forcing and sampling strategies are used interleaved. The former uses ground-truth labels for supervision, while the latter uses the current policy to sample trajectories based on the current policy to obtain pseudo-supervision.

IV. Experiments

A. Datasets and Metrics

a) Datasets: The proposed method is first validated across three public datasets (R2R [1], REVERIE [2], and SOON [3]), and further verified on our own collected real-scenario dataset. Specifically, R2R comprises 7,189 paths with 5-7 viewpoints each, divided into training, validation seen and unseen, and test unseen sets. REVERIE focuses on more concise instructions with the additional target of object grounding, presenting 21,702 instances averaging 18 words, following R2R's partition scheme. SOON provides



Method	SR↑	SPL↑
SepFFN	**67.77**	**58.61**
UniTrans	65.52	55.52
UniTransCLS	66.84	57.41

Fig. 4: Comparisons on RQ (left) and R2R val-unseen (right).

TABLE I: Comparisons of multiple visual components.

Visual Feature			Validation Seen		Validation Unseen	
Image	Semantics	Depth	SR↑	SPL↑	SR↑	SPL↑
✓			69.44	61.66	67.05	56.40
✓	✓		70.42	64.21	67.26	57.94
✓	✓	✓	**71.30**	**66.20**	**67.77**	**58.61**

longer instructions (average 47 words) for locating rooms and objects. Our custom dataset are detailed in Sec. IV-H.

b) Evaluation Metrics: Regarding R2R, four standard metrics are utilized: Navigation Error (NE) measures the distance between the predicted stop position and the ground truth. Success Rate (SR) indicates the proportion of paths where the agent stops within 3m of the target points. Oracle Success Rate (OSR) is akin to SR with the oracle stop policy. Success Rate Weighted by Path Length (SPL) adapts SR penalized by the path length. For REVERIE and SOON, two additional metrics are introduced: Remote Grounding Success Rate (RGS) evaluates the accuracy of properly grounded objects, while RGS Weighted by Path Length (RGSPL) adjusts this by path length.

B. Implementation Details

During pre-training, the PMV encoder is first trained using a batch size of 12.8k and a learning rate of 1×10^{-4} over 1k epochs on one NVIDIA GeForce RTX 3090. Subsequently, the trained PMV encoder weights are loaded for visual representation, and METER [29] weights are integrated for cross-modal learning. If task-specific pre-training tasks are conducted, the default settings follow a batch size of 48 for 300K iterations. For fine-tuning, the batch size is set to 8, and the learning rate is 1×10^{-5} with 80k iterations. Following [41], the CLIP-B/16 [28] pre-trained model is used to extract original and augmented image features [42]. The ablation studies are mainly conducted on the R2R dataset. The multi-type visual fusion and panoramic spatial encoding in PMV have 1 and 2 layers. The text and dual-scale encoders each have 6 and 3 transformer layers.

C. Impact of the Panoramic Multi-type Visual Encoder

1) Impact of different fusion methods: The comparisons of three fusion methods presented in Fig. 4 demonstrate that the SepFFN module, despite its apparent simplicity, outperforms others in accuracy on the RQ test set, and achieves the highest SR and SPL on the R2R val-unseen set. Furthermore, UniTransCLS surpasses UniTrans in performance, indicating that an architectural simplicity and the decoupling of location communications may play a key role in improving multi-type visual fusion in panoramic views.

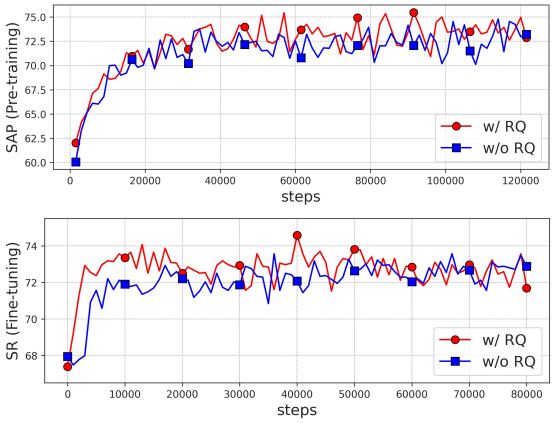


Fig. 5: Comparisons of the RQ pre-training task.

TABLE II: Comparisons of cross-modal backbones.

Backbone	Validation Seen		Validation Unseen	
	SR↑	SPL↑	SR↑	SPL↑
LXMERT [13]	67.78	62.57	65.30	57.60
METER [29]	**71.30**	**66.20**	**67.77**	**58.61**

2) Impact of different visual components: The experimental results for visual components are presented in Tab. I. Sole reliance on RGB image features yields SPL values of 61.66 and 56.40 for the seen and unseen splits, respectively. Incorporating semantics enhances these values by 2.55% and 1.54%. Furthermore, the addition of depth features significantly boosts performance, increasing SPL by 4.54% and 2.21%. These findings demonstrate that incorporating depth and semantic information enriches the agent's visual perception, resulting in enhanced navigation capabilities.

D. Impact of the Region Query Pre-training Task

Fig. 5 presents the learning curves of SAP during pre-training and SR during fine-tuning on R2R unseen datasets, both with and without using the RQ task to pre-train the PMV encoder. Notably, the usage of the RQ pre-training task expedites the learning process and results in superior peak performance. These experimental results demonstrate the significant impact of the RQ task in augmenting the visual understanding and navigation inference capabilities.

E. Impact of the Dual-scale Cross-modal Encoder

1) Impact of different cross-modal backbones: Tab. II compares the performance of using LXMERT and METER as the cross-modal backbone. The results demonstrate that the METER-based model improves SR by 2.47%, and SPL by 1.01% on the unseen set, indicating the superiority of the fully-transformer vision-and-language pre-trained model.

TABLE III: Comparisons of layers. N_L and N_X denote the numbers of the text encoder and the cross-modal encoder.

Nums of layers		Val Seen		Val Unseen	
N_L	N_X	SR↑	SPL↑	SR↑	SPL↑
9	4	70.71	65.54	65.82	56.66
6	**3**	**71.30**	**66.20**	**67.77**	**58.61**
4	2	69.83	66.03	66.24	57.57

TABLE IV: Comparison with the state-of-the-art methods on the R2R dataset.

Pre-trained	Method	Validation Seen				Validation Unseen				Test Unseen			
		SR↑	SPL↑	NE↓	OSR↑	SR↑	SPL↑	NE↓	OSR↑	SR↑	SPL↑	NE↓	OSR↑
N	EnvDrop [10]	62	59	3.99	-	52	48	5.22	-	51	47	5.23	59
	AuxRN [43]	70	**67**	3.33	78	55	50	5.28	62	55	51	5.15	62
	NvEM [11]	69	65	3.44	-	60	55	4.27	-	58	54	4.37	66
	SEAT (Ours)	**71**	66	**3.16**	**78**	**68**	**59**	**3.70**	**75**	**67**	**59**	**3.81**	**74**
Y	HAMT [5]	76	72	2.51	82	66	61	3.29	73	65	60	3.93	72
	DUET [6]	79	73	2.28	86	72	60	3.31	81	69	59	3.65	76
	DSRG [35]	81	76	2.23	**88**	73	62	3.10	81	72	61	3.33	78
	EnvEdit [42]	77	74	2.32	-	69	64	3.24	-	68	64	3.59	-
	SEAT (Ours)	**81**	**76**	**2.13**	87	**75**	**64**	**3.03**	**82**	**73**	**64**	**3.13**	**81**

TABLE V: Comparison with the state-of-the-art methods on the REVERIE dataset.

Method	Validation Seen					Validation Unseen					Test Unseen				
	OSR↑	SR↑	SPL↑	RGS↑	RGSPL↑	OSR↑	SR↑	SPL↑	RGS↑	RGSPL↑	OSR↑	SR↑	SPL↑	RGS↑	RGSPL↑
RecBERT [4]	53.90	41.79	47.96	38.23	35.61	35.02	30.67	24.90	18.77	15.27	32.91	29.61	23.99	16.50	13.51
HAMT [5]	47.65	43.29	40.19	27.20	25.18	36.84	32.95	30.20	18.92	17.28	33.41	30.40	26.67	14.88	13.08
HOP+ [38]	56.43	55.87	49.55	40.76	36.22	40.04	36.07	31.13	22.49	19.33	35.81	33.82	28.24	20.20	16.86
DUET [6]	73.86	71.75	63.94	57.41	51.14	51.07	46.98	33.73	32.15	23.03	56.91	52.51	**36.06**	31.88	**22.06**
SEAT (Ours)	**81.59**	**79.76**	**74.37**	**62.83**	**58.21**	**54.33**	**49.45**	**35.51**	**32.83**	**23.14**	**57.66**	**52.62**	35.67	**32.46**	21.98

TABLE VI: Comparison on the SOON dataset.

Split	Method	OSR↑	SR↑	SPL↑	RGSPL↑
Val Unseen	GBE [3]	28.54	19.52	13.34	1.16
	DUET [6]	50.91	36.28	22.58	3.75
	SEAT (Ours)	**51.00**	**36.87**	**24.87**	**3.91**
Test Unseen	GBE [3]	21.45	12.90	9.23	0.45
	DUET [6]	43.00	33.44	21.42	4.17
	SEAT (Ours)	**46.66**	**35.89**	**22.55**	**4.47**

2) Impact of different number of stacked layers: Tab III explores the effect of varying the number of layers, with N_L and N_X representing the layers in the text and cross-modal encoders, respectively. Contrary to the conventional settings of $N_L = 9$ and $N_X = 4$ as in prior work [4], [5], [35] following the full large pre-trained model [13], we propose a reduced layer configuration to mitigate overfitting in smaller datasets. Our experiments show that setting $N_L = 6$ and $N_X = 3$ significantly improves generalization, increasing SR and SPL by 1.95% on the unseen set. This underscores the efficiency of a leaner model structure for enhanced task-specific performance in constrained dataset environments.

F. Comparison with State-of-the-Arts (SoTA)

1) Comparison of effectiveness: The experimental results compared with previous SoTA methods in R2R, REVERIE, and SOON, are presented in Tab. IV, Tab. V, and Tab. VI, respectively. SEAT showcases exceptional performance across these three datasets. On the R2R dataset, SEAT exhibits superior navigation performance across most metrics, both with and without task-specific pre-training. When no task-specific pre-training is performed, the SEAT surpasses the previous SoTA no-pretraining LSTM method [11]. With pre-training, SEAT significantly enhances SR by 4%, 6%, and 5% on validation seen, unseen, and test unseen splits, respectively, compared to the prior SoTA method [42]. In goal-oriented instruction navigation tasks, SEAT also excels in performance enhancement across most metrics. Notably, on the REVERIE validation seen set, SR and SPL improve by 8.01% and 10.43% respectively, while RGS and RGSPL see enhancements of 5.42% and 7.07%. Furthermore, on the SOON dataset, SEAT boosts SPL by 2.29% and 1.13% on validation and test unseen splits, respectively.

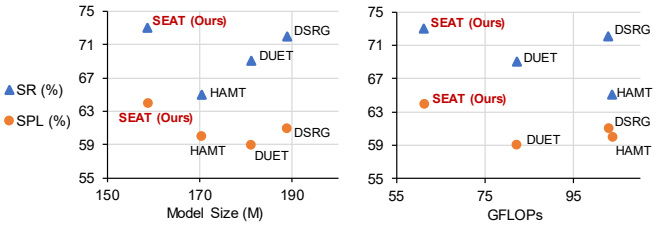


Fig. 6: Comparisons of efficiency and effectiveness.

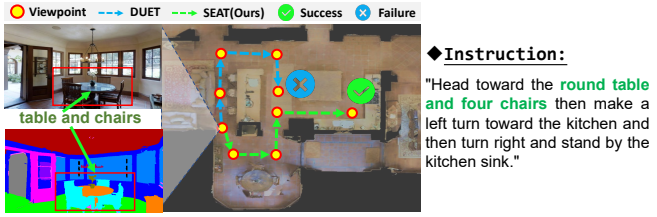


Fig. 7: Visualizations on the R2R validation unseen set.

2) Comparison of efficiency: Computational complexities are assessed using the Python toolkit `thop` to calculate GFLOPs. Single-step forward inferences are conducted with a batch size of 8, instruction length of 44, and visited nodes of 6 across all methods. Fig. 6 illustrates that SEAT achieves superior SR and SPL while using fewer learnable parameters and computational loads. This improvement can be attributed to our proposed PMV encoder, which accounts for only 10.7% of the whole model's parameters, and the streamlined cross-modal encoder with fewer layers.

G. Qualitative Analysis

The navigation paths on the R2R val-unseen set are illustrated in Fig. 7. SEAT's enhanced semantic and depth understanding enables more accurate instruction-following and better target destination attainment. For instance, SEAT adeptly locates the specified `round table` and `chairs`, adjusting its orientation accordingly.

Fig. 8: Comparisons of MP3D [17] and our collected dataset.

TABLE VII: Comparisons on collected real-world dataset.

Method	SR↑	SPL↑	NE↓	OSR↑
EnvDrop [10]	47.30	43.90	5.13	54.80
DUET [6]	53.76	47.14	4.35	60.22
SEAT (Ours)	**56.99**	**47.20**	**3.88**	**77.42**

H. Language-Guided Robot Navigation in Real Scenarios

To assess the robot's navigation capabilities in real-world environments, we use SLAM [44] to collect 2D maps from various campus sites, such as offices and meeting rooms. Following MP3D [17] and simulation settings [45], we convert each location's map into a topological format and capture panoramas using a panoramic camera. Each panorama is segmented into 36 sub-images. We then employ ZoeDepth [24] and oneFormer [25] to generate depth images and semantic labels, respectively. This approach enriches the environment's representational dimensions without requiring manual annotations or extra sensors, enabling the robot to achieve a deeper and more comprehensive understanding of its surroundings. Fig. 8 showcases the stylistic differences between our experimental setups and the MP3D dataset.

We follow R2R [1] to sample trajectories from topological maps and manually annotate 3 distinct instruction styles for each trajectory. Finally, we get 98 trajectories across 5 scans with 294 instructions. The dataset is divided into a training set (67 trajectories from 3 scans) and a validation unseen set (31 trajectories from the remaining 2 scans). We test the navigation performance of SEAT and previous SoTA models, EnvDrop [10] and DUET [6]. We first load the best checkpoint from R2R and then fine-tune the models on our newly annotated dataset. All these methods are fine-tuned with a batch size of 4 and a learning rate of 5×10^{-6}. When using EnvDrop as the backbone, we achieve a success rate similar to previous real-world experiment [45], and our SEAT model gets the best performance on the val-unseen set in all metrics, demonstrating excellent robustness (Tab VII).

In real-world validation, the mobile robot receives instructions through the microphone, and we pre-compute the visual features for all topological nodes for convenience. The robot keeps navigating to the point that the model predicted until the output action is [STOP] or the maximum step is exceeded. Fig. 9 illustrates a mobile robot navigating in a meeting room, and the model successfully locates the `flower`, `chairs`, and `whiteboard` that appeared in the instructions, stopping at the correct place. For an in-depth qualitative analysis, encompassing both failure instances and

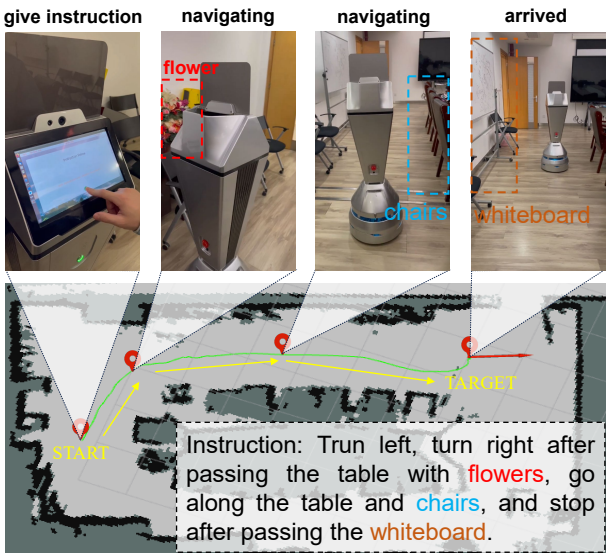


Fig. 9: The mobile robot navigates to the target following human-provided verbal instructions.

live demonstrations, please consult our *supplementary video*.

V. Limitations

While SEAT advances language-guided robot navigation with its panoramic visual perception and cross-modal fusion, it faces challenges in real-world deployment due to the complex pre-processing required for topological maps. This process may struggle with rapidly changing or poorly documented settings, impacting SEAT's adaptability and accuracy in dynamic scenarios. Addressing these limitations involves refining the topological map generation process to handle environmental variability better and ensuring SEAT can dynamically update its understanding of its surroundings.

VI. Conclusion and Future Work

In this paper, SEAT is introduced as a novel approach to enhance environmental perception and understanding in language-guided robot navigation. By utilizing a comprehensive panoramic multi-type visual encoder, SEAT effectively extracts and fuses multiple types of visual information—RGB, semantics, and depth—aided by a region-query pre-training task. This method significantly improves the robot's comprehension of its surroundings. Additionally, the optimization of the cross-modal encoder with customized layers is crucial for reducing the model's size, thereby mitigating overfitting issues associated with small navigation datasets. The successful deployment on a mobile robot in real scenarios demonstrates SEAT's potential. Future efforts will focus on the process of optimizing topological maps and applying these insights to continuous settings [46], aiming to optimize real-world robot navigation solutions.

References

[1] P. Anderson, Q. Wu, *et al.*, "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.

[2] Y. Qi, Q. Wu, *et al.*, "Reverie: Remote embodied visual referring expression in real indoor environments," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9982–9991.

[3] F. Zhu, X. Liang, *et al.*, "Soon: Scenario oriented object navigation with graph-based exploration," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12 689–12 699.

[4] Y. Hong, Q. Wu, *et al.*, "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.

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

[6] S. Chen, P.-L. Guhur, *et al.*, "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.

[7] J. Huo, Q. Sun, *et al.*, "Geovln: Learning geometry-enhanced visual representation with slot attention for vision-and-language navigation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 23 212–23 221.

[8] Z. He, L. Wang, *et al.*, "Learning depth representation from rgb-d videos by time-aware contrastive pre-training," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[9] Y. Qi, Z. Pan, *et al.*, "Object-and-action aware model for visual language navigation," in *Computer Vision–ECCV 2020: 16th European Conference, Part X 16*. Springer, 2020, pp. 303–317.

[10] H. Tan, L. Yu, and M. Bansal, "Learning to navigate unseen environments: Back translation with environmental dropout," in *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019, pp. 2610–2621.

[11] D. An, Y. Qi, *et al.*, "Neighbor-view enhanced model for vision and language navigation," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 5101–5109.

[12] W. Hao, C. Li, *et al.*, "Towards learning a generic agent for vision-and-language navigation via pre-training," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 13 137–13 146.

[13] H. Tan and M. Bansal, "Lxmert: Learning cross-modality encoder representations from transformers," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 5100–5111.

[14] B. Yu, H. Kasaei, and M. Cao, "L3mvn: Leveraging large language models for visual target navigation," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2023, pp. 3554–3560.

[15] Y. Wang, C.-Y. Ko, and P. Agrawal, "Visual pre-training for navigation: What can we learn from noise?" in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2023, pp. 3897–3902.

[16] H. Deguchi, S. Taguchi, *et al.*, "Enhanced robot navigation with human geometric instruction," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2023, pp. 9071–9078.

[17] A. Chang, A. Dai, *et al.*, "Matterport3D: Learning from RGB-D data in indoor environments," *International Conference on 3D Vision (3DV)*, 2017.

[18] X. Wang, Q. Huang, *et al.*, "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.

[19] D. Fried, R. Hu, *et al.*, "Speaker-follower models for vision-and-language navigation," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[20] L. Wang, C. Liu, *et al.*, "Pasts: Progress-aware spatio-temporal transformer speaker for vision-and-language navigation," *Engineering Applications of Artificial Intelligence*, vol. 128, p. 107487, 2024.

[21] A. Vaswani, N. Shazeer, *et al.*, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[22] R. Liu, X. Wang, *et al.*, "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.

[23] R. Dang, Z. Shi, *et al.*, "Unbiased directed object attention graph for object navigation," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 3617–3627.

[24] S. F. Bhat, R. Birkl, *et al.*, "Zoedepth: Zero-shot transfer by combining relative and metric depth," *arXiv preprint arXiv:2302.12288*, 2023.

[25] J. Jain, J. Li, *et al.*, "Oneformer: One transformer to rule universal image segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 2989–2998.

[26] J. Devlin, M.-W. Chang, *et al.*, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[27] 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 (CVPR)*, June 2021, pp. 7036–7045.

[28] A. Radford, J. W. Kim, *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[29] Z.-Y. Dou, Y. Xu, *et al.*, "An empirical study of training end-to-end vision-and-language transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 166–18 176.

[30] A. Moudgil, A. Majumdar, *et al.*, "Soat: A scene-and object-aware transformer for vision-and-language navigation," *Advances in Neural Information Processing Systems*, vol. 34, pp. 7357–7367, 2021.

[31] Y. Qi, Z. Pan, *et al.*, "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 (ICCV)*, October 2021, pp. 1655–1664.

[32] K. He, X. Zhang, *et al.*, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[33] E. Wijmans, A. Kadian, *et al.*, "Dd-ppo: Learning near-perfect pointgoal navigators from 2.5 billion frames," 2019.

[34] Y. Liu, M. Ott, *et al.*, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[35] L. Wang, Z. He, *et al.*, "A dual semantic-aware recurrent global-adaptive network for vision-and-language navigation," in *International Joint Conferences on Artificial Intelligence (IJCAI)*, 2023.

[36] P. Anderson, X. He, *et al.*, "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.

[37] J. Cao, J. Li, *et al.*, "Towards interpreting deep neural networks via layer behavior understanding," *Machine Learning*, vol. 111, no. 3, pp. 1159–1179, 2022.

[38] Y. Qiao, Y. Qi, *et al.*, "Hop+: History-enhanced and order-aware pre-training for vision-and-language navigation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[39] L. Wang, Z. He, *et al.*, "Res-sts: Referring expression speaker via self-training with scorer for goal-oriented vision-language navigation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 7, pp. 3441–3454, 2023.

[40] J. D. M.-W. C. Kenton and L. K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of NAACL-HLT*, 2019, pp. 4171–4186.

[41] S. Shen, L. H. Li, *et al.*, "How much can clip benefit vision-and-language tasks?" in *International Conference on Learning Representations*, 2022.

[42] J. Li, H. Tan, and M. Bansal, "Envedit: Environment editing for vision-and-language navigation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 15 407–15 417.

[43] F. Zhu, Y. Zhu, *et al.*, "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.

[44] W. Hess, D. Kohler, *et al.*, "Real-time loop closure in 2d lidar slam," in *2016 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2016, pp. 1271–1278.

[45] P. Anderson, A. Shrivastava, *et al.*, "Sim-to-real transfer for vision-and-language navigation," in *Conference on Robot Learning*. PMLR, 2021, pp. 671–681.

[46] J. Krantz, E. Wijmans, *et al.*, "Beyond the nav-graph: Vision-and-language navigation in continuous environments," in *Computer Vision–ECCV 2020: 16th European Conference, Part XXVIII 16*. Springer, 2020, pp. 104–120.