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

Multimodal Evolutionary Encoder for Continuous Vision-Language Navigation

Zongtao He[1], Liuyi Wang[1], Lu Chen[1], Shu Li[1], Qingqing Yan[1], Chengju Liu[1] and Qijun Chen[1], *Senior Member, IEEE*

***Abstract*— Can multimodal encoder evolve when facing increasingly tough circumstances? Our work investigates this possibility in the context of continuous vision-language navigation (continuous VLN), which aims to navigate robots under linguistic supervision and visual feedback. We propose a multimodal evolutionary encoder (MEE) comprising a unified multimodal encoder architecture and an evolutionary pre-training strategy. The unified multimodal encoder unifies rich modalities, including depth and sub-instruction, to enhance the solid understanding of environments and tasks. It also effectively utilizes monocular observation, reducing the reliance on panoramic vision. The evolutionary pre-training strategy exposes the encoder to increasingly unfamiliar data domains and difficult objectives. The multi-stage adaption helps the encoder establish robust intra- and inter-modality connections and improve its generalization to unfamiliar environments. To achieve such evolution, we collect a large-scale multi-stage dataset with specialized objectives, addressing the absence of suitable continuous VLN pre-training. Evaluation on VLN-CE demonstrates the superiority of MEE over other direct action-predicting methods. Furthermore, we deploy MEE in real scenes using self-developed service robots, showcasing its effectiveness and potential for real-world applications. Our code and dataset are available at https://github.com/RavenKiller/MEE.**

I. INTRODUCTION

Instruction-following ability is an emerging research focus in the field of service robotics. Towards this goal, vision-language navigation [1] (VLN) aims at navigating robots by linguistic instructions and visual feedback, without relying on pre-existing maps or global path planning. This presents challenges in understanding the environment with limited vision and ambiguous language instructions. The complexity of VLN is further amplified in continuous settings [2], where assumptions like ideal teleportation, perfect localization and prior connectivity graphs [1], [3], [4] become unavailable. Effectively harnessing constrained multimodal observations and ensuring generalization in unseen environments are crucial factors for continuous VLN.

Transformer [5] models have addressed several multi-modal challenges but still face obstacles in the VLN field. Firstly, existing methods [4], [6]–[11] often underutilize rich modalities, such as depth information, which is important for object localization and obstacle avoidance in continuous settings. These methods also simplify observation constraints by assuming 36-way panoramic camera perception, as in R2R [1], which could lead to increased computing consumption and pose a high threshold for cost-sensitive robots. Additionally, the absence of well-suited pre-training methods limits generalization potential, as many techniques [12]–[15] lack essential depth and sub-instruction [16] components and are not tailored to the egocentric navigation domain. These disparities constrain the useful knowledge learned from pre-training and limit the quality of feature representations for VLN tasks.

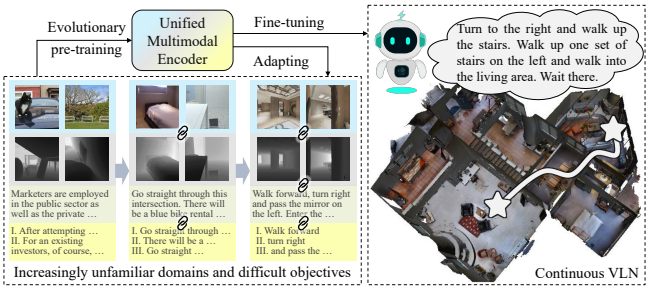

Fig. 1. Overview of our method. The unified multimodal encoder is pre-trained evolutionarily to adapt unfamiliar domains and difficult objectives.

To address these limitations, this paper introduces a novel multimodal evolutionary encoder (MEE), focusing on two aspects: encoder architecture and pre-training strategy, as depicted in Fig. 1. The unified multimodal encoder fuses four modalities through self-attention and cross-attention mechanism, which effectively unifies inter-modality feature spaces and enhances the solid understanding of environments and tasks. Besides, it eliminates dependence on panoramas by extracting monocular observations into more fine-grained features, reducing the cost of migration to physical robots.

Subsequently, we propose an evolutionary pre-training strategy for improving the encoder's representation ability. Inspired by biological evolution, evolutionary pre-training sets staged survival obstacles to the encoder, such as unfamiliar data domains and difficult training objectives. After adapting one stage, the encoder is challenged again with tougher circumstances in next stage. It gradually evolves to the optimal point in the encoder parameter space, which involves knowledge and skills from commonsense to specialized. Therefore, our pre-training strategy can aid in unseen domain generalization and robust multimodal representations

This work was supported in part by the National Natural Science Foundation of China under Grants (62173248, 62073245, 62233013), in part by the Shanghai Science and Technology Innovation Action Plan (22511104900), in part by the Shanghai Municipal Science and Technology Major Project (2021SHZDZX0100), and in part by the Fundamental Research Funds for the Central Universities. *(Corresponding author: Chengju Liu; Qijun Chen.)*

[1]The authors are with the Department of Control Science and Engineering, Tongji University, Shanghai, 201804, China. E-mail: xingchen327@tongji.edu.cn, wly@tongji.edu.cn, chenlu_i@tongji.edu.cn, lishu@tongji.edu.cn, qyan_0131@tongji.edu.cn, liuchengju@tongji.edu.cn, qjchen@tongji.edu.cn.

for downstream VLN tasks. The dataset we use is newly collected from 20 different public sources. To our knowledge, it is the first large-scale multi-stage dataset with four modalities and specialized objectives for continuous VLN pre-training.

Our contributions can be summarized as follows:

- We introduce a unified multimodal encoder that enhances comprehensive perception with unified features while reduces the reliance on panoramas.
- We propose an evolutionary pre-training strategy that enables the encoder to evolve better feature representations and generalization ability across multiple stages.
- Through extensive evaluations in both simulated and real scenes, we validate the effectiveness of MEE. We will release the experiment code and the large-scale evolutionary pre-training dataset.

II. RELATED WORK

VLN research originated from discrete environments, such as indoor R2R [1] and outdoor Touchdown [17]. In discrete settings, the agent teleports from one node to another in a pre-defined navigation graph and does not need to consider localization or navigation errors. These assumptions dramatically reduce the difficulty because the agent has a global environment structure and can execute actions optimally. To lift these discrete assumptions, some studies propose continuous VLN [2], [18]. Continuous settings require the agent to navigate in free space, localize its position, and avoid obstacles, which brings the task closer to physical world.

With advancements in pre-training technologies, several studies leverage Transformer models [19]–[21] to address multimodal challenges. For instance, HAMT [9] and DUET [10] establish a global topological history and use pre-trained Transformers to fuse past observations. Related to our work, VLNBERT [6] and AirBERT [7] also employ multi-stage pre-training. However, they only consider two modalities, overlooking important depth and sub-instructional information. In terms of pre-training design, they lack incremental objectives and have lower data volume compared to our approach. Moreover, their models often function as discrete path-instruction discriminators, which are not suitable for dynamically predicting actions in continuous environments.

To bridge the gap between discrete and continuous settings, hierarchical waypoint-based methods attempt to transfer discrete models to continuous environments [22]–[24] and have achieved remarkable performance. The assumption of panoramic vision helps robots gather richer information about their surroundings, while the additional low-level policy increases the accuracy of executing waypoints. In contrast, our method adopts an end-to-end architecture that directly predicts robot control actions, without relying on panoramic observations or hierarchical policies. This streamlined architecture is advantageous for exploring navigation possibilities with limited vision and for eliminating interference when validating the effectiveness of the pre-trained encoder. Additionally, the proposed model can be seamlessly transferred to physical monocular robots, as demonstrated in our experiments.

III. METHOD

As shown in Fig. 2, our model contains the proposed unified multimodal encoder and a recurrent action decoder. The encoder is pre-trained by the proposed evolutionary pre-training strategy. Then, the whole model is fine-tuned under imitation learning for actual VLN tasks. In our settings, the robot receives an egocentric RGB image O_t^{rgb}, a depth image O_t^{dep}, a navigation instruction O_t^{ins} and a sub-instruction [16] O_t^{sub} at every step t. After obtaining observations, the model predicts an action $a_t \in \{TurnLeft, TurnRight, Forward, Stop\}$, which navigates the robot to the target position. Model and pre-training details will be described in the rest of this section. For simplicity, we will use $\mathcal{M} \in \{rgb, dep, ins, sub\}$ to mark a modality in several equations and eliminate the time subscript t when discussing the pre-training procedure.

A. Unified Multimodal Encoder

The unified multimodal encoder projects all modalities to a unified feature space and then conducts attention mechanism to obtain a fused feature sequence. First, observations from different modalities are embedded by four pre-extractors to unify the feature dimension:

- RGB $X_t^{rgb} = \text{ViT}(O_t^{rgb}) \in \mathbb{R}^{L^{rgb} \times d}$,
- Depth $X_t^{dep} = \text{ResNet}(O_t^{dep}) \in \mathbb{R}^{L^{dep} \times d}$,
- Instruction $X_t^{ins} = \text{BERT}(O_t^{ins}) \in \mathbb{R}^{L^{ins} \times d}$,
- Sub-instruction $X_t^{sub} = \text{BERT}(O_t^{sub}) \in \mathbb{R}^{L^{sub} \times d}$,

where $L^{\mathcal{M}}$ is the sequence length of the patch embeddings or word embeddings in the modality $\mathcal{M}$ and d is the embedding dimension. ViT and BERT are pre-trained weights from CLIP [25], and ResNet is loaded from a point navigation model [26]. Besides sequential features, we expect the model to output representational features with aggregated semantics. We introduce four learnable tokens embedding $E^{\mathcal{M}} \in \mathbb{R}^{1 \times d}$, whose corresponding outputs can be used as representational features for action decoding. Then, all embeddings are concatenated into $X_t \in \mathbb{R}^{L \times d}$:

$$\begin{aligned} X_t = [&E^{rgb}, X_t^{rgb}, E^{dep}, X_t^{dep}, \\ &E^{ins}, X_t^{ins}, E^{sub}, X_t^{sub}], \end{aligned} \tag{1}$$

where $[\cdot, \cdot]$ denotes the concatenation at sequence dimension and $L = 4 + \sum_{\mathcal{M}} L^{\mathcal{M}}$.

To introduce position and modality information, we add learnable positional embeddings $\text{PE} \in \mathbb{R}^{L \times d}$ for token positions and type embeddings $\text{TE} \in \mathbb{R}^{L \times d}$ for token modalities to X_t. Then, we unify the four feature spaces through self-attention and cross-attention in stacked attention blocks:

$$\tilde{X}_t = X_t + \text{PE} + \text{TE}, \tag{2}$$

$$S_t = \text{AttentionBlocks}(\tilde{X}_t). \tag{3}$$

The output feature sequence S_t contains four representational features $r_t^{\mathcal{M}} \in \mathbb{R}^d$ representing global semantics of every modality, and four sequential features $S_t^{\mathcal{M}} \in \mathbb{R}^{L^{\mathcal{M}} \times d}$ storing more fine-grained spatial semantics.

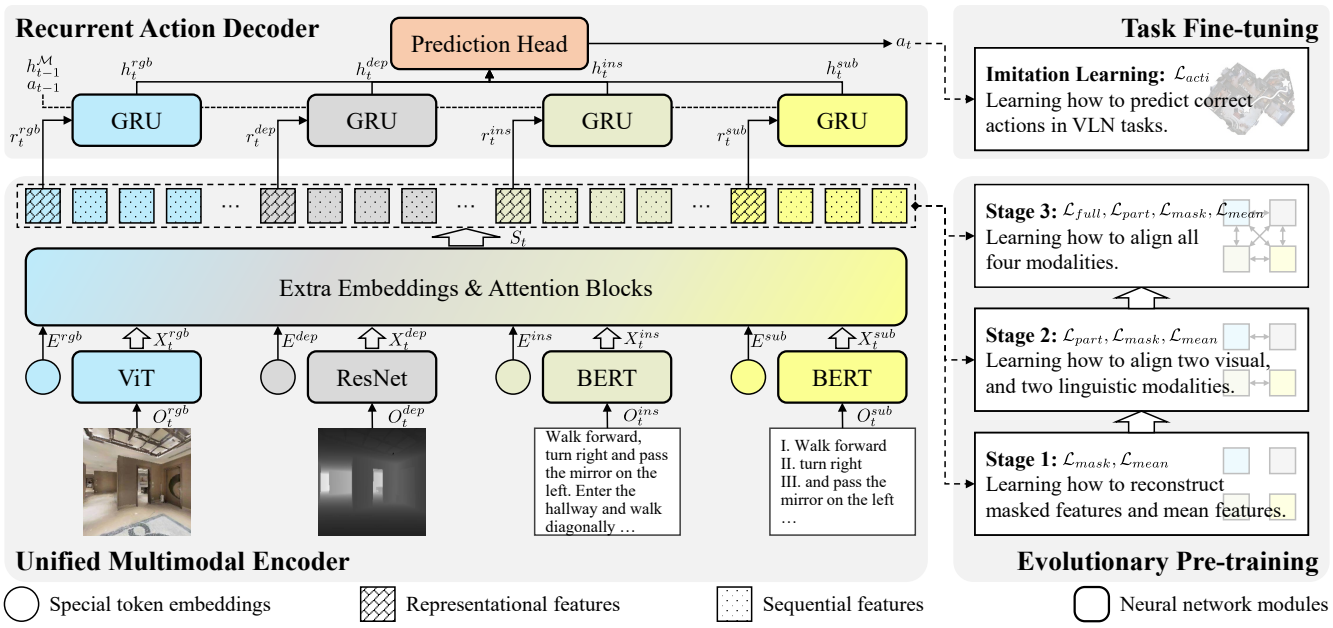

Fig. 2. Model architecture. The navigation model comprises the proposed unified multimodal encoder and a recurrent action decoder. The encoder is pre-trained by the proposed evolutionary pre-training strategy, and the whole model is then fine-tuned by imitation learning.

B. Recurrent Action Decoder

History information is crucial for executing long navigation instructions. We use common GRU [27] modules to recurrently memorizes history information and dynamically decodes navigation actions. Every GRU accepts a representational feature $r_t^{\mathcal{M}}$ together with previous action a_{t-1} and output a hidden state $h_t^{\mathcal{M}}$:

$$\begin{aligned} h_t^{\mathcal{M}} &= \text{GRU}^{\mathcal{M}}([r_t^{\mathcal{M}}; a_{t-1}], h_{t-1}^{\mathcal{M}}), \\ \mathcal{M} &\in \{rgb, dep, ins, sub\}. \end{aligned} \tag{4}$$

Then, the action prediction head accepts hidden states and produces an action distribution $\hat{a}_t \in \mathbb{R}^4$:

$$\hat{a}_t = \text{FC}(\text{LN}(h_t^{rgb} + h_t^{dep} + h_t^{ins} + h_t^{sub})), \tag{5}$$

where FC signifies a fully connected layer, and LN is the layer normalization. The predicted action a_t is calculated by greedy decoding $a_t = \text{Argmax}_k(\hat{a}_{t,k})$.

C. Evolutionary Pre-training

We propose an evolutionary pre-training strategy to improve the quality of feature representations by increasingly unfamiliar data domains and difficult training objectives. As listed in Table I, we define three pre-training stages to achieve the encoder evolution.

1) Stage 1 (S1): In this stage, we gather samples from general domains to instruct the encoder on how to represent features of each modality through self-supervised objectives. Drawing inspiration from the masked language model in BERT [19], we introduce a masked feature reconstruction loss $\mathcal{L}_{mask}$, which tasks the encoder with reconstructing the masked features throughout the rest feature sequence. Let $X[P]$ be the operation that selects features from X at randomly masked positions P, the masked feature reconstruction loss is formulated as:

$$\mathcal{L}_{mask} = \sum_{\mathcal{M}} \text{MSE}(X^{\mathcal{M}}[P], \text{FC}(S^{\mathcal{M}}[P])), \tag{6}$$

where MSE refers to the mean square error, and FC signifies a fully connected layer. Besides, to ensure a comprehensive understanding of modalities, we introduce a mean feature reconstruction loss to aid in summarizing overall information:

$$\mathcal{L}_{mean} = \sum_{\mathcal{M}} \text{MSE}(\text{Mean}(X^{\mathcal{M}}), \text{FC}(r^{\mathcal{M}})). \tag{7}$$

Through reconstruction objectives, the encoder can establish strong inner relationships between parts within a modality, thereby enhancing the quality of feature representations.

2) Stage 2 (S2): In this stage, we collect partially aligned samples from related domains to assist the encoder in learning connections between two modalities. We introduce a partial alignment loss to aid the encoder in correctly identifying partial alignment relations, for instance, paired RGB-depth images. Let y^{visu} be a binary indicator for visual modality alignment and y^{lang} be a binary indicator for linguistic modality alignment. The partial alignment loss is formulated as:

$$\begin{aligned} \mathcal{L}_{part} = &\text{BCE}(y^{visu}, \text{FC}([r^{rgb}; r^{dep}])) \\ &+ \text{BCE}(y^{lang}, \text{FC}([r^{ins}; r^{sub}])), \end{aligned} \tag{8}$$

where BCE represents binary cross-entropy with softmax, and $[\cdot;\cdot]$ denotes concatenation at the feature dimension. Aligned pairs in the dataset serve as natural positives. For negatives, we employ a uniform random sampling strategy.

Through partial alignment objectives, the encoder can gradually integrate different views of the world and effectively capture useful information between similar modalities.

TABLE I

OVERVIEW OF THE EVOLUTIONARY PRE-TRAINING. RGB: RGB IMAGES, DEP: DEPTH IMAGES, INS: INSTRUCTIONS, SUB: SUB-INSTRUCTIONS.

Stage	# RGB	# DEP	# INS	# SUB	Objective	Domain	Source
S1	12.31M	2.54M	10.73M	10.73M	Reconstruction	General	**RGB**: ImageNet [28], VisualGenome [29], COCO [30], LAION-HR [15], CC-12M [31]; **DEP**: Scannet [32]; **INS**: C4 [33]; **SUB**: C4 [33]
S2	8.68M	8.68M	8.27M	8.27M	+ Partial alignment	Related	**RGB/DEP**: HM3D [34], SUN3D [35], TUM [36], SceneNet [37], NYUv2 [38], DIODE [39]; **INS/SUB**: C4 [33], Marky [40], Touchdown [17], map2seq [41], CHALET [42], ALFRED [43]
S3	3.42M	3.42M	3.42M	3.42M	+ Full alignment	In-domain	**RGB/DEP/INS/SUB**: VLNCE [2], EnvDrop [44]

3) Stage 3 (S3): In this stage, we build fully aligned samples in the VLN domain and promote the encoder to fuse features from all four modalities. We deploy an expert agent in the simulator to collect observation tuples (rgb, dep, ins, sub) step by step. Subsequently, we introduce a full alignment loss that guides the encoder in extracting the most distinctive features from all modalities. Let y^{full} be a binary indicator for correct alignment of all modalities. The full modalities alignment loss is formulated as:

$$\mathcal{L}_{full} = \text{BCE}(y^{full}, \text{FC}([r^{rgb}; r^{dep}; r^{ins}; r^{sub}])). \quad (9)$$

Through full alignment objectives, the encoder gains insights into inter-modality nuances, ensuring a comprehensive understanding of observations and navigation tasks. Besides, the encoder is equipped with knowledge from commonsense to specialized domains after all stages, facilitating the generalization ability to unseen environments.

D. Task Fine-tuning

The task fine-tuning process is designed to train the entire navigation model for specific continuous VLN tasks. Let a_t^{gt} represent the ground-truth action at time t. The supervised loss for a navigation path is defined as:

$$\mathcal{L}_{acti} = \frac{1}{T}\sum_{t=1}^{T} \text{CE}(a_t^{gt}, \hat{a}_t), \quad (10)$$

where CE denotes the cross-entropy loss used for action classification. In line with Krantz et al. [2], we employ the imitation learning algorithm DAgger [45] to fine-tune our model. During training, DAgger randomly selects predicted actions to execute, even if they are incorrect. This algorithm, positioned between teacher-forcing and student-forcing, allows the model to encounter a broader range of unexpected scenarios and benefits for navigating in unseen scenes.

IV. EXPERIMENTS

A. Setup

For the MEE model, we set the following dimension parameters: $d = 512, L^{rgb} = 50, L^{dep} = 16, L^{ins} = 77, L^{sub} = 12$. Therefore, the maximum sequence length is $L = 4 + \sum_{\mathcal{M}} L^{\mathcal{M}} = 159$. The encoder is implemented using 4 attention layers and 8 multi-head attention heads. To mitigate overfitting, we use the dropout mechanism with a probability of 0.1 in attention layers.

In the evolutionary pre-training, we use the AdamW optimizer with a batch size of 128, a learning rate of 1.0×10^{-5}, and a warmup step of 1000. All loss components $\mathcal{L}_{\{mask,mean,part,full\}}$ are balanced by coefficients $[1.0, 1.0]$ in S1, $[0.5, 0.5, 1.0]$ in S2, and $[0.33, 0.33, 0.34, 1.00]$ in S3. The aligning positive ratio is 0.4 in S2 and 0.3 in S3. The data come from 20 open-source datasets and simulators. For validation efficiency, only a subset is used in experiments.

For task fine-tuning, we follow the VLN-CE baseline, utilizing the DAgger imitation algorithm with a total of 10×5000 training steps. We use the Adam optimizer with a batch size of 2 trajectories and a learning rate of 1.0×10^{-4}. Consistent with previous research [2], [23], [49], we evaluate our model on VLN-CE val-seen, val-unseen, and test splits. Only scenes in the val-seen split intersect with training. The evaluation metrics include TL (trajectory length), NE (average navigation error in meters), OSR (percentage of oracle success rate), SR (percentage of success rate), and SPL (percentage of success rate weighted by path length).

Simulated experiments are conducted on a virtual machine with an NVIDIA RTX 3090 GPU and real scene experiments are conducted using self-developed robots. For more training and evaluation details, please refer to our code repository.

B. Benchmark Results

We present the evaluation results on the standard VLN-CE benchmark in Table II. For fairness, methods are divided into two groups: the hierarchical group (upper part) that first forecasts a waypoint and then leverages an additional control policy to reach the waypoint, and the direct group (lower part) that directly predicts control actions.

Comparison in the direct group. Compared with the official baseline CMA [2] with similar decoder units, the MEE model demonstrates superiority on almost all metrics, which preliminarily proves the effectiveness of our evolutionary encoder. SASRA [47] and CM2 [48] builds semantic maps to enhance environment understanding, while LAW [49] uses discrete waypoint supervision to improve the instruction-path alignment. However, they lag behind the MEE model in val-unseen and test performance. Thanks to rich modalities, MEE can form a comprehensive understand-

TABLE II

BENCHMARK RESULTS. THE UPPER METHODS USE HIERARCHICAL WAYPOINT FRAMEWORKS WHILE THE LOWER METHODS DIRECTLY PREDICT CONTROL ACTIONS. UNDERLINE MARKS THE BEST METRICS AMONG ALL GROUPS, AND **BOLD** MARKS THE BEST METRICS IN THE DIRECT GROUP.

	Val-Seen					Val-Unseen					Test				
MODEL	**TL↓**	**NE↓**	**OSR↑**	**SR↑**	**SPL↑**	**TL↓**	**NE↓**	**OSR↑**	**SR↑**	**SPL↑**	**TL↓**	**NE↓**	**OSR↑**	**SR↑**	**SPL↑**
R2R-CMTP [46]	-	7.10	45.4	36.1	31.2	-	7.90	38.0	26.4	22.7	-	-	-	-	-
Waypoint [22]	8.54	5.48	53.0	46.0	43.0	7.62	6.31	40.0	36.0	34.0	8.02	6.65	37.0	32.0	30.0
BridgeGap [24]	12.50	5.02	59.0	50.0	44.0	12.23	5.74	53.0	44.0	39.0	13.31	5.89	51.0	42.0	36.0
Sim2Sim [23]	11.18	4.67	61.0	52.0	44.0	10.69	6.07	52.0	43.0	36.0	11.43	6.17	52.0	44.0	37.0
SASRA [47]	**8.89**	7.17	-	36.0	34.0	**7.89**	8.32	-	24.0	22.0	-	-	-	-	-
CM2 [48]	12.05	**6.10**	**50.7**	**42.9**	34.8	11.54	7.02	41.5	34.3	27.6	13.9	7.70	**39.0**	31.0	24.0
CMA [2]	9.06	7.21	44.0	34.0	32.0	8.27	7.60	36.0	29.0	27.0	**8.85**	7.91	36.0	28.0	25.0
LAW [49]	9.34	6.35	49.0	40.0	37.0	8.89	6.83	44.0	35.0	31.0	9.67	7.69	38.0	28.0	25.0
MEE (ours)	9.16	6.54	50.3	40.4	**37.9**	9.26	**6.82**	**44.6**	**35.9**	**32.3**	8.97	**7.46**	38.8	**31.3**	**28.4**

TABLE III

ABLATION STUDY OF MODALITIES.

	Val-Seen		Val-Uneen	
Model	**SR↑**	**SPL↑**	**SR↑**	**SPL↑**
Full model	**40.4**	**37.9**	**35.9**	**32.3**
- w/o RGB	5.1	4.9	4.6	4.3
- w/o DEP	4.9	4.7	5.9	5.5
- w/o INS	11.1	9.8	10.7	9.6
- w/o SUB	16.7	15.6	25.5	22.9

TABLE IV

ABLATION STUDY OF PRE-TRAINING STAGES AND ALIGNMENT TYPES.

					Val-Seen		Val-Uneen	
#	**S1**	**S2**	**S3**	**ALIGN**	**SR↑**	**SPL↑**	**SR↑**	**SPL↑**
1	-	-	-	Class.	27.1	25.3	20.5	19.3
2	✓	-	-	Class.	26.7	25.5	25.0	23.6
3	✓	✓	-	Class.	37.5	35.8	27.9	25.6
4	✓	-	✓	Class.	37.1	35.0	27.8	25.5
5	✓	✓	✓	Class.	**40.4**	**37.9**	**35.9**	**32.3**
6	✓	✓	✓	Cont.	37.8	36.2	30.2	27.9

ing of environments and tasks. Evolutionary pre-training further strengthens bonds between modalities and enhances generalization to unfamiliar circumstances.

Comparison with the hierarchical group. Although the MEE model achieves competitive performance with Waypoint [22], we recognize a performance gap between the MEE model and state-of-the-art hierarchical methods. One reason for this gap is that BridgeGap [24] and Sim2Sim [23] use panoramic vision, which reduces observation constraints and supplies rich visual information. Another reason is the introduction of prior graph training and additional low-level policy, which help structure environments in advance and improve the action execution accuracy. In contrast, our work mainly focuses on encoder ability and uses as simple a decoder and fine-tuning method as possible, which might constitute a bottleneck in our model. Therefore, a future direction is to utilize the advantages of hierarchical decoding to further unlock the potential of MEE.

C. Ablation Study

1) Different modalities: We conducted an ablation study to assess the impact of each modality on the navigation performance of the MEE model, as shown in Table III. The results consistently demonstrate that ablated models perform significantly worse than the full model, thus highlighting the importance of each modality for achieving successful navigation. Notably, ablating vision inputs has a more detrimental effect compared to ablating language inputs. This suggests that the MEE model has learned human intuition to some extent, so it can sometimes complete navigation tasks without language instructions. However, the biased reliance on visual cues also indicates dataset bias, which is harmful to generalization. To address this, one direction is to enhance the significance of language inputs, aiming for a more balanced and effective integration of modalities.

2) Pre-training stages: We conducted another ablation experiment to investigate the influence of pre-training stages S1, S2 and S3 on navigation performance. As outlined in Table IV, incorporating successive pre-training stages brings a notable enhancement in navigation performance. Specifically, S1 predominantly benefits val-unseen performance (#2 vs. #1), emphasizing generalization through the reconstruction loss. On the other hand, S2 significantly enhances val-seen metrics (#3 vs. #2) by emphasizing robust feature extraction through partial alignment loss on seen environments. The targeted ablation of S2 (#4 vs. #5) further validates the necessity of aligning RGB and depth as distinct modalities. The introduction of S3 substantially improves performance (#5) thanks to the in-domain data and the challenging objectives.

3) Alignment types: We also validated the effectiveness of alignment types in the loss. In Table IV, 'Class.' refers to binary classification, while 'Cont.' means the contrastive loss, a popular approach for cross-modal pre-training. Specifically, we replaced the BCE loss for each ordered modality pair $(\mathcal{M}_1, \mathcal{M}_2)$ in Eq. (8-9) with the InfoNCE loss [25]:

$$\mathcal{L}_{\mathcal{M}_1,\mathcal{M}_2} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(r_i^{\mathcal{M}_1}\cdot r_i^{\mathcal{M}_2}/\tau)}{\sum_{j=1}^{N}\exp(r_i^{\mathcal{M}_1}\cdot r_j^{\mathcal{M}_2}/\tau)}, \quad (11)$$

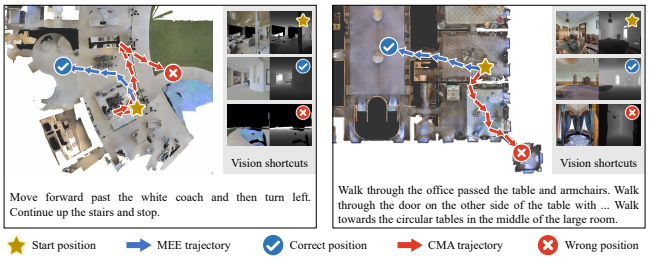

Fig. 3. Two successful examples. Our method MEE successfully navigates to the correct position, while the CMA baseline fails.

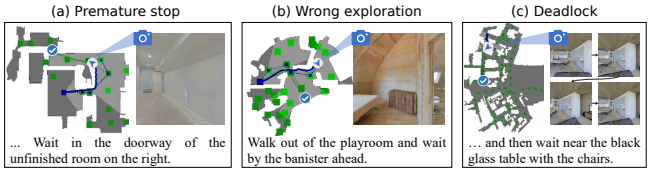

Fig. 4. Failure cases. The blue line is the robot trajectory, the arrow mark is where the robot stops and the check mark is the target position.

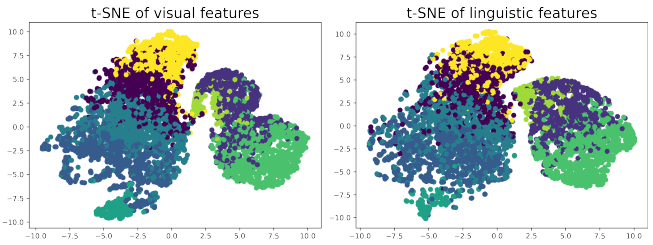

Fig. 5. t-SNE feature plots. Colors show the KMeans clustering results, where aligned visual-linguistic pairs have the same color.

TABLE V

ENCODER PERFORMANCE FROM DIFFERENT STAGES. $\mathcal{L}$ REPRESENTS LOSS AND $\mathcal{A}$ REPRESENTS PERCENTAGE ALIGNMENT ACCURACY.

Stage	$\mathcal{L}_{mask} \downarrow$	$\mathcal{L}_{mean} \downarrow$	$\mathcal{A}_{part} \uparrow$	$\mathcal{A}_{full} \uparrow$
S1	4.15	0.85	34.79	29.93
S2	1.22	0.32	70.90	30.86
S3	**0.45**	**0.04**	**99.78**	**89.78**

where N is the batch size; $r^{\mathcal{M}_1}$ and $r^{\mathcal{M}_2}$ with the same subscript are positive pairs, otherwise negative pairs. Despite achieving high contrastive accuracy (almost 100%) during pre-training, the contrastive variant lagged behind the classification variant in terms of navigation performance (#6 vs. #5). In MEE, different modalities exchange information through cross-attention. So, the contrastive variant might cheat by copying features instead of learning meaningful representations, resulting in poor performance on unseen data. Based on this, we opted for the classification loss in our method and left contrastive variants for future research.

D. Qualitative Visualization

1) Comparison with baseline: Fig. 3 presents two VLN-CE trajectories by the MEE model and the CMA baseline [2]. In the first example, the robot was instructed to move past a white coach and ascend the stairs. While CMA reached an incorrect position by a pool, MEE effectively linked visual cues and the provided instruction, accurately navigating to the stairs and understanding subsequent directives like "stop." In the second example, the robot was directed to exit an office, cross a hall, and head towards circular tables. CMA misinterpreted the instruction and ended up in a distant room. In contrast, MEE adeptly followed the guidance and stopped precisely at a circular table. These visualizations demonstrate MEE's comprehensive multimodal extraction and suitability for continuous VLN tasks.

2) Failure analysis: Fig. 4 illustrates three typical failure modes of MEE. In case (a), the robot terminated prematurely due to the extended length of the doorway, which led to its misunderstanding of the accurate stopping position. In case (b), despite initially following instructions correctly, the robot explored a wrong direction and mistakenly identified the cabinet as the banister. Case (c) presents an intriguing yet challenging scenario where the robot encountered an imperceptible obstacle (a low sofa). Despite continuously turning from side to side, the robot could not escape the deadlock until the maximum step count was reached. These failure cases underscore the need for further enhancement in our method, particularly in resolving visual-linguistic ambiguities and avoiding deadlocks.

E. Discussion

1) Are multimodal feature spaces unified?: An essential property of MEE is the unified feature spaces between visual and linguistic modalities. To verify this, we conducted a t-SNE analysis [50]. We randomly selected 6K samples from stage 3 and extracted their features ($r^{\mathcal{M}}$ in Fig. 2). For simplicity, we averaged RGB and depth features to form visual features, and instruction and sub-instruction features to form linguistic features. Fig. 5 shows their t-SNE plots, along with the results of 8-class KMeans clustering. The similar shapes between the t-SNE plots and the consistent cluster distributions demonstrate the effective unification of feature spaces between vision and language. Additionally, MEE achieved an average cosine similarity of 0.95 between aligned samples, further indicating the quality of the features.

2) Does the encoder evolve better ability across stages?: One of the key concerns in evolutionary pre-training is whether the encoder effectively improves its capabilities over successive stages. To investigate this, we evaluated the encoder performance from different stages. Table V demonstrates that the introduction of pre-training stages positively impact the loss value and alignment performance. For instance, the reconstruction loss $\mathcal{L}$ of the S2 model is lower than that of the S1 model, and the alignment accuracy $\mathcal{A}$ of the S3 model surpasses that of the S2 model. These findings suggest that evolutionary pre-training facilitates the gradual enhancement of the encoder's ability.

3) Can parameter freezing help the evolution?: To assess the impact of parameter freezing between stages, we analyzed the alignment F1-scores of pre-training stages 2 and 3, as shown in Fig. 6. The parameters of half attention blocks are kept unchanged in the frozen group and fully

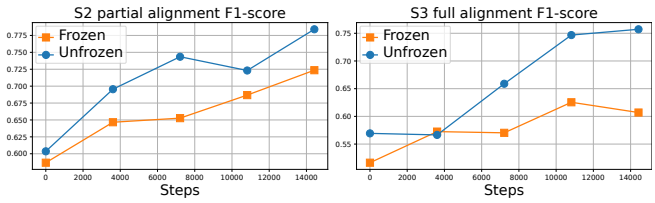

Fig. 6. Alignment F1-scores with different parameter freezing strategies.

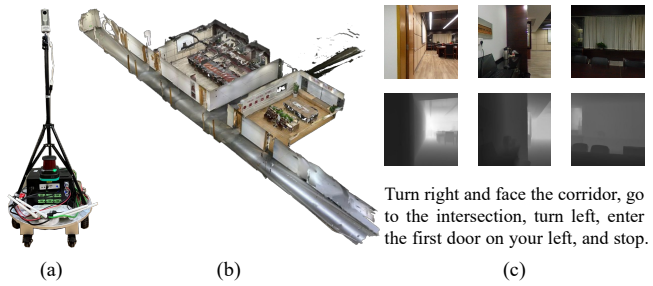

Fig. 7. (a) The scene collecting automated guided vehicle. (b) One of the six real scenes. (c) Images and instructions in the VLN@TJ dataset.

TABLE VI

GENERALIZATION TO REAL SCENES. EACH CELL SHOWS THE SCENE UNDERSTANDING METRIC VALUES ON REAL/SIMULATED DATASET.

Variant	$\mathcal{L}_{mask} \downarrow$	$\mathcal{L}_{mean} \downarrow$	$\mathcal{A}_{part} \uparrow$	$\mathcal{A}_{full} \uparrow$
PT	1.09 / **0.45**	0.07 / **0.04**	94.65 / **99.78**	70.78 / **89.78**
FT	2.49 / 2.71	0.29 / 0.44	82.80 / 70.75	66.13 / 59.07
PT+FT	**0.71** / 0.48	**0.04** / 0.07	**96.11** / 97.22	**80.60** / 83.79

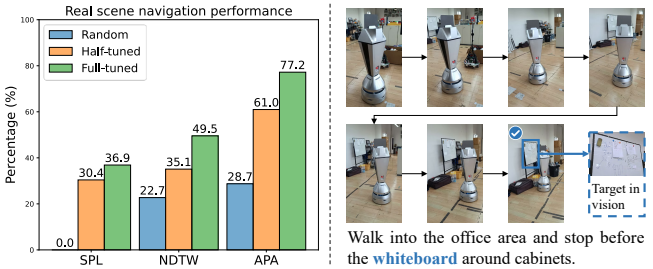

Fig. 8. Generalization to real scenes. (left) Navigation performance. (right) Navigation visualization.

trained in the unfrozen group. The results indicate that the unfrozen group consistently outperforms the frozen group. In the context of our pre-training process, parameter freezing limits the learning ability of the encoder, making it difficult to adapt to increasingly challenging training objectives. So, although effective in fine-tuning, we choose not to freeze the parameters of attention blocks between pre-training stages.

F. Generalization to Real Scenes

To further validate the ability of MEE, we applied it to real scenes using our scene collecting AGV (Fig. 7 (a)). We collected navigation paths on our campus and manually annotated navigation instructions for these paths. After post-processing such as depth estimation [51], the real scene dataset VLN@TJ consists of 6 environments and 408 path-instruction pairs. Fig. 7 (b) visualizes one of the scenes in the school buildings, and Fig. 7 (c) shows several images and instructions in our dataset.

In Table VI, we evaluated the scene understanding performance of MEE variants: PT (evolutionarily pre-trained), FT (trained on real data from scratch), and PT+FT (fine-tuned on real data from PT). For each variant, we reported its reconstruction loss and alignment accuracy on both the real dataset (VLN@TJ) and the simulated dataset (S3). PT achieved decent performance without access to the real dataset, significantly surpassing FT on all metrics. This result demonstrates that MEE has a strong generalization ability to real scenes, even without any fine-tuning. With fine-tuning, PT+FT performed best on the real dataset while retaining most of its learned knowledge on the simulated dataset, providing a reasonable transferring way.

We then applied the entire MEE navigation model to real scenes. The initial version of VLN@TJ uses the R2R [1] discrete formats, not suitable for continuous VLN. Thus, we leveraged the self-developed service robot with monocular camera to extend additional 13 paths in the VLN-CE format. Similar to the open-loop evaluation in autonomous driving [52], we assess the MEE model's performance against pre-recorded expert behavior. The MEE model was fine-tuned by 6 paths (half-tuned) and 13 paths (full-tuned), then evaluated using 3 metrics: SPL, APA (action prediction accuracy), and NDTW (normalized dynamic time warping path similarity). As shown in Fig. 8, the half-tuned variant significantly outperformed the random agent and achieved a similar SPL as in the VLN-CE val-unseen split. With more fine-tuning paths, the full-tuned variant achieved higher NDTW and APA, demonstrating a significant path consistency with the ground truth. These results prove that the MEE model can be seamlessly transferred to physical robots with constrained perception devices. Fig. 8 also illustrates a navigation path controlled by the MEE model, where the robot successfully reached the target whiteboard. More demonstrations can be found in our supplementary video.

V. CONCLUSIONS

This paper presents a novel approach, the multimodal evolutionary encoder (MEE), for extracting comprehensive and generalized multimodal feature representations. The proposed unified multimodal encoder unifies rich modalities through attention mechanism, enhancing the comprehensive understanding of environments and tasks. By reducing reliance on panoramic vision, MEE becomes more adaptable to robots with constrained perception devices and costs. The proposed evolutionary pre-training strategy exposes the encoder to increasingly challenging circumstances, facilitating the evolution of better knowledge and skills. This strategy also strengthens multimodal bonds via incremental objectives and improves the encoder's generalization to unseen environments. Evaluation on the VLN-CE benchmark confirms the effectiveness of MEE, while real scene experiments

demonstrate its potential for real-world applications. Our future work will focus on developing contrastive variants, addressing modality imbalance, and exploring different designs for the action decoder.

REFERENCES

[1] P. Anderson *et al.*, "Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments," in *Proceedings of CVPR*, 2018, pp. 3674–3683.

[2] J. Krantz *et al.*, "Beyond the nav-graph: Vision-and-language navigation in continuous environments," in *Proceedings of ECCV*, 2020, pp. 104–120.

[3] A. Chang *et al.*, "Matterport3d: Learning from rgb-d data in indoor environments," in *Proceedings of 3DV*, 2017, pp. 667–676.

[4] D. Fried *et al.*, "Speaker-follower models for vision-and-language navigation," in *Advances in Neural Information Processing Systems*, 2018, pp. 3318–3329.

[5] A. Vaswani *et al.*, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017, pp. 5998–6008.

[6] A. Majumdar *et al.*, "Improving vision-and-language navigation with image-text pairs from the web," in *Proceedings of ECCV*, 2020, pp. 259–274.

[7] P.-L. Guhur *et al.*, "Airbert: In-domain pretraining for vision-and-language navigation," in *Proceedings of ICCV*, 2021, pp. 1634–1643.

[8] Y. Hong *et al.*, "Vln bert: A recurrent vision-and-language bert for navigation," in *Proceedings of CVPR*, 2021, pp. 1643–1653.

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

[10] S. Chen *et al.*, "Think global, act local: Dual-scale graph transformer for vision-and-language navigation," in *Proceedings of CVPR*, 2022, pp. 16 537–16 547.

[11] L. Wang *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.

[12] X. Chen *et al.*, "Microsoft coco captions: Data collection and evaluation server," *arXiv preprint arXiv:1504.00325*, 2015.

[13] Y. Goyal *et al.*, "Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering," in *Proceedings of CVPR*, 2017.

[14] P. Sharma *et al.*, "Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning," in *Proceedings of the 56th ACL (Volume 1: Long Papers)*, 2018, pp. 2556–2565.

[15] C. Schuhmann *et al.*, "LAION-5b: An open large-scale dataset for training next generation image-text models," in *Thirty-sixth NeurIPS Datasets and Benchmarks Track*, 2022.

[16] Y. Hong *et al.*, "Sub-instruction aware vision-and-language navigation," in *Proceedings of EMNLP*, 2020, pp. 3360–3376.

[17] H. Chen *et al.*, "Touchdown: Natural language navigation and spatial reasoning in visual street environments," in *Proceedings of CVPR*, 2019, pp. 12 530–12 539.

[18] X. Li *et al.*, "Reve-ce: Remote embodied visual referring expression in continuous environment," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 1494–1501, 2022.

[19] J. Devlin *et al.*, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of NAACL-HLT, Volume 1 (Long and Short Papers)*, Jun. 2019, pp. 4171–4186.

[20] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proceedings of ICLR*, 2020.

[21] J. Lu *et al.*, "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," in *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[22] J. Krantz *et al.*, "Waypoint models for instruction-guided navigation in continuous environments," in *Proceedings of ICCV*, 2021, pp. 15 162–15 171.

[23] J. Krantz and S. Lee, "Sim-2-sim transfer for vision-and-language navigation in continuous environments," in *Computer Vision – ECCV 2022*, Cham, 2022, pp. 588–603.

[24] Y. Hong *et al.*, "Bridging the gap between learning in discrete and continuous environments for vision-and-language navigation," in *Proceedings of CVPR*, 2022, pp. 15 439–15 449.

[25] A. Radford *et al.*, "Learning transferable visual models from natural language supervision," in *Proceedings of ICML*, ser. Proceedings of Machine Learning Research, vol. 139, 2021, pp. 8748–8763.

[26] E. Wijmans *et al.*, "Dd-ppo: Learning near-perfect pointgoal navigators from 2.5 billion frames," in *Proceedings of ICLR*, 2019.

[27] K. Cho *et al.*, "On the properties of neural machine translation: Encoder-decoder approaches," *arXiv preprint arXiv:1409.1259*, 2014.

[28] J. Deng *et al.*, "Imagenet: A large-scale hierarchical image database," in *Proceedings of CVPR*, 2009, pp. 248–255.

[29] R. Krishna *et al.*, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *International journal of computer vision*, vol. 123, pp. 32–73, 2017.

[30] T.-Y. Lin *et al.*, "Microsoft coco: Common objects in context," in *Proceedings of ECCV*. Springer, 2014, pp. 740–755.

[31] S. Changpinyo *et al.*, "Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts," in *Proceedings of CVPR*, 2021, pp. 3558–3568.

[32] A. Dai *et al.*, "Scannet: Richly-annotated 3d reconstructions of indoor scenes," in *Proceedings of CVPR*, 2017, pp. 5828–5839.

[33] C. Raffel *et al.*, "Exploring the limits of transfer learning with a unified text-to-text transformer," *The Journal of Machine Learning Research*, vol. 21, no. 1, pp. 5485–5551, 2020.

[34] S. K. Ramakrishnan *et al.*, "Habitat-matterport 3d dataset (hm3d): 1000 large-scale 3d environments for embodied ai," *arXiv preprint arXiv:2109.08238*, 2021.

[35] J. Xiao *et al.*, "Sun3d: A database of big spaces reconstructed using sfm and object labels," in *Proceedings of ICCV*, 2013, pp. 1625–1632.

[36] J. Sturm *et al.*, "A benchmark for the evaluation of rgb-d slam systems," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2012, pp. 573–580.

[37] J. McCormac *et al.*, "Scenenet rgb-d: Can 5m synthetic images beat generic imagenet pre-training on indoor segmentation?" in *Proceedings of ICCV*, 2017.

[38] N. Silberman *et al.*, "Indoor segmentation and support inference from rgbd images," in *Computer Vision – ECCV 2012*. Springer Berlin Heidelberg, 2012, pp. 746–760.

[39] I. Vasiljevic *et al.*, "DIODE: A Dense Indoor and Outdoor DEpth Dataset," *CoRR*, vol. abs/1908.00463, 2019.

[40] A. Kamath *et al.*, "A new path: Scaling vision-and-language navigation with synthetic instructions and imitation learning," in *Proceedings of CVPR*, June 2023, pp. 10 813–10 823.

[41] R. Schumann and S. Riezler, "Generating landmark navigation instructions from maps as a graph-to-text problem," in *Proceedings of ACL-IJCNLP (Volume 1: Long Papers)*, 2021, pp. 489–502.

[42] D. Misra *et al.*, "Mapping instructions to actions in 3D environments with visual goal prediction," in *Proceedings of EMNLP*, Oct.-Nov. 2018, pp. 2667–2678.

[43] M. Shridhar *et al.*, "ALFRED: A Benchmark for Interpreting Grounded Instructions for Everyday Tasks," in *Proceedings of CVPR*, 2020.

[44] H. Tan *et al.*, "Learning to navigate unseen environments: Back translation with environmental dropout," in *Proceedings of NAACL-HLT, Volume 1 (Long and Short Papers)*, 2019, pp. 2610–2621.

[45] S. Ross *et al.*, "A reduction of imitation learning and structured prediction to no-regret online learning," in *Proceedings of AISTATS*, vol. 15, 2011, pp. 627–635.

[46] K. Chen *et al.*, "Topological planning with transformers for vision-and-language navigation," in *Proceedings of CVPR*, June 2021, pp. 11 276–11 286.

[47] M. Z. Irshad *et al.*, "Semantically-aware spatio-temporal reasoning agent for vision-and-language navigation in continuous environments," in *2022 26th International Conference on Pattern Recognition (ICPR)*, 2022, pp. 4065–4071.

[48] G. Georgakis *et al.*, "Cross-modal map learning for vision and language navigation," in *Proceedings of CVPR*, June 2022, pp. 15 460–15 470.

[49] S. Raychaudhuri *et al.*, "Language-aligned waypoint (law) supervision for vision-and-language navigation in continuous environments," in *Proceedings of EMNLP*, 2021, pp. 4018–4028.

[50] L. van der Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. 86, pp. 2579–2605, 2008.

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

[52] L. Chen *et al.*, "End-to-end autonomous driving: Challenges and frontiers," *arXiv preprint arXiv:2306.16927*, 2023.