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

MG-VLN: Benchmarking Multi-Goal and Long-Horizon Vision-Language Navigation with Language Enhanced Memory Map

Junbo Zhang[1], Kaisheng Ma[1*]

***Abstract*— Vision-Language Navigation (VLN) with high-level language instructions is a crucial task in robotics. Existing VLN benchmarks, such as the REVERIE challenge which has single-goal instructions and limited navigation steps, do not fully encapsulate the complexity of real-world navigation that often require multi-objective and long-horizon navigation. To address this, we propose a new benchmark task: Multi-Goal and Long-Horizon Vision-Language Navigation (MG-VLN), extending the REVERIE benchmark to encompass multi-objective and long-horizon navigation scenarios with sequences of high-level instructions. This task aims to provide a simulation benchmark to guide the design of lifelong and long-horizon navigation robots. To initiate the exploration in this newly proposed task, we first investigate the role of long-term memory in improving navigation performance by leveraging environmental information gathered during previous sub-goals. Additionally, we examine the types of knowledge that most effectively enrich this long-term memory. Specifically, we integrate the visual contents with linguistic knowledge such as object categories, visual captions, and object attributes/relationships. Our findings indicate that: 1) the explicit long-term memory map significantly enhances navigation performance in multi-goal and long-horizon scenarios; 2) incorporating object attributes and relationships information is the most advantageous for aligning environmental cues with high-level instructions.**

I. INTRODUCTION

Vision-Language Navigation (VLN) is a pivotal task in embodied AI and a vital robot competency. In VLN, an agent is instructed to interpret natural language, perceive its 3D surroundings, and navigate to the designated location. Current VLN tasks are guided by two types of language instructions: The first is fine-grained, providing detailed, step-by-step directions [1], such as "*Walk in between the table and bookshelves, continue straight onto the brick floor, and turn right. Walk straight and into the room*". The second is high-level, succinctly stating the navigation objective [2], [3], for instance, "*Find the towel in the bathroom with a fishing theme and fold it*". The latter, high-level instruction tasks are notably more complex as they demand the interpretation of implicit links between abstract semantic cues and the environment. This paper focuses on VLN with high-level instructions due to their significant potential for practical real-world applications.

In real-world scenarios, humans' high-level instructions typically encompass **multi-objective** and **long-horizon** elements. Take, for instance, the scenario depicted in Figure 1, where an individual might provide the instruction: "*Help me water the plants now*". Such an instruction inherently comprises three distinct navigation sub-tasks: initially "*Find the watering kettle*", followed by "*Fill the kettle at the sink*", and finally "*Go to the living room and water the plants*". Each sub-task is similar to the current high-level VLN task. An ideal embodied agent should be adept at 1) generating multiple sub-goals from human instruction and 2) planning navigation actions for each sub-goal. For 1), it is relatively easy for current Large Language Models to divide the instructions into sub-tasks given their strong chain-of-thought reasoning ability. However, for 2), there exists no embodied benchmark evaluating an agent's ability to navigate to multiple goals with long-horizon following a sequence of high-level instructions. The existing datasets, such as REVERIE [2], are restricted to single-goal tasks that demand only 4-7 navigation steps. Its navigation paths cover only a small part of the 3D environment and cannot reflect the multi-goal navigation demands in real life.

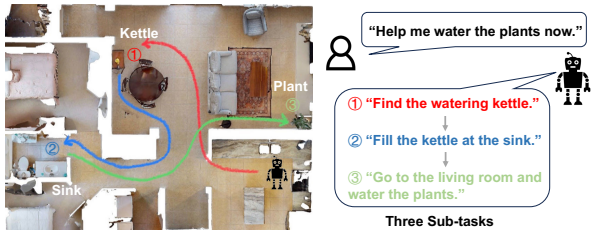

Fig. 1. An example of a robot following instructions in real-world scenarios. The high-level instructions from humans usually inherently contain multiple sub-goals and are long-horizon. Accordingly, we propose a new multi-goal VLN task and explore the role of the long-term memory map.

To address this discrepancy, this paper presents a novel Multi-Goal and Long-Horizon Vision-Language Navigation (MG-VLN) task by extending the REVERIE benchmark. In each MG-VLN episode, the agent is tasked with following **a sequence of** high-level language instructions, with each one directing the agent to a distinct object in an unseen 3D environment. The agent is expected to navigate to multiple objects **in order**, and identify the bounding boxes encompassing the specified targets. We adapt the evaluation metrics from the REVERIE benchmark to better measure the models' accuracy and efficiency within the MG-VLN framework. An in-depth discussion of the task formulation is provided in Section III. Our proposed MG-VLN is designed to provide a robust simulation benchmark to guide the assessment and design of lifelong [4] and long-horizon [5], [6] navigation robots in real-world.

In the proposed MG-VLN task, firstly, we investigate the role of a long-term memory module. Basically, an intuitive solution to MG-VLN task is to perform the single-goal VLN methods for each sub-goal in an episode. However, in our

[1] Institute for Interdisciplinary Information Sciences, Tsinghua University
* Corresponding Author

long-horizon navigation task on unseen environments, the navigation paths of the sub-goals usually have overlaps. The environmental information collected when navigating previous goals may be helpful for finding subsequent goals. Thus, simply performing single-goal VLN with no prior information is inefficient in MG-VLN. To address this, we construct a long-term memory module for the transformer-based VLN method (HAMT [13]) and map-based VLN method (DUET [14]), and study their performance in multi-goal vision-language navigation (MG-VLN).

Secondly, we explore what knowledge is effective in constructing the long-term memory of map-based navigators in MG-VLN. Existing multi-object navigation research has examined the application of long-term memory [7]–[12], often constructing it as a spatial semantic map imbued with object category semantics. However, the number of object categories used in prior works is limited to 8 in MON [11] and MultiON 2.0 [10], and 13 in IVLN [12]. While this may suffice for object navigation tasks that typically involve a small set of common objects, it is inadequate for high-level vision-language navigation tasks like REVERIE [2], where target objects span across 489 categories and its language instructions involve broad semantics, such as object attributes and spatial relationships. Thus, we explore more types of linguistic knowledge to complement the visual content in long-term memory. Specifically, besides object category semantics (as in prior works), we introduce multiple types of language-enhanced memory maps, such as visual captions generated by vision-language foundation models, and objects' attribute/relationship knowledge. The language-enhanced knowledge is injected into the viewpoints' representations in a topological map, which builds a better mapping between the environment and the high-level instructions.

Through comprehensive experimental analysis, we demonstrate that in the proposed MG-VLN task, explicit long-term memory is beneficial for long-horizon vision-language navigation. Additionally, the rich attribute and spatial relationship knowledge are most valuable for MG-VLN tasks. In summary, our contributions are as follows:

- We introduce the Multi-Goal and Long-Horizon Vision-Language Navigation (MG-VLN) task, which better reflects the multi-objective and long-horizon navigation requirements in real-world scenarios.
- We examine the role of long-term memory in MG-VLN tasks, demonstrating that explicit topological memory map improves multi-goal navigation efficiency.
- We benchmark multiple types of language-enhanced memory on MG-VLN tasks with map-based navigators, including object category semantics, visual captions, and retrieved attribute and relationship knowledge.

II. RELATED WORK

A. Vision-language Navigation

VLN requires the agent to understands language instructions and navigate to target locations. Current VLN benchmarks contain step-by-step instructions (e.g., R2R [1], R4R [16] and RxR [17], dialog-based instructions (e.g., CVDN [18] and TouchDown [19]) and high-level goal-oriented instructions (e.g., REVERIE [2] and SOON [3]). Early VLN methods adopt recurrent neural networks to encode navigation history and predict actions [20]–[24]. More recently, transformer-based architectures have been widely used [25]–[29]. HAMT [13] utilizes a transformer module to encode previous observations and actions, which serve as navigation history. Map-based methods build efficient environmental representations to achieve more accurate navigation with SLAM [30]–[32], semantic map [33]–[35], topological map [36]–[39] and top-down map [40]–[42]. DUET [14] utilizes a dual-scale graph transformer with a well-constructed viewpoint map, significantly improving the navigation accuracy. Recently, environment augmentation [43]–[45], instruction generation [46]–[49] and view generation [50]–[52] have been introduced. Other methods [53]–[56] propose to adopt external knowledge to complement the vision input. This paper studies the proposed MG-VLN tasks with widely-used transformer-based (HAMT) and map-based (DUET) navigators.

TABLE I. Comparison of MG-VLN with other multi-goal benchmarks.

Benchmark	Goal Type	#Category	#Sub-Goal	Oracle Phase
MultiON [8]	Cylinders	0	1-3	×
MultiON 2.0 [10]	Cylinders / Natural Objects	0 or 8	1-3	×
MON [11]	Natural Objects	6 or 8	1-4	×
IVLN [12]	Step-by-step Instructions	-	∼70	✓
MG-VLN (Ours)	High-level Instructions	489	1-3	×

B. Multi-goal and Long-horizon Navigation

The differences between MG-VLN and previous multi-goal navigation benchmarks are summarized in Table I. For object navigation, Multi-ON [8] first proposes a benchmark for finding multiple colored cylinders in simulator. It studied the role of the occupancy map and object category map in the proposed benchmark. Subsequent works improve the performance on Multi-ON by decoupling mapping from localization [58], active camera policy [59] and learned neural implicit representations [60]. Recent works improve the MultiON benchmark by allowing more realistic and flexible multi-object navigation. MultiON 2.0 [10] generates a larger multi-object navigation benchmark based on HM3D [61] and introduces 8 natural objects as navigation goals. MON [11] proposes a new benchmark for finding multiple objects with explicit category semantics. Sequence-agnostic MultiON [62] further relaxes the limitation of pre-determined sequence order. These works utilize object category semantics as memory maps to improve multi-object navigation efficiency. However, the number of object categories is limited to 6 or 8 in these benchmarks as shown in Table I (#Category). We study the multi-goal navigation in VLN, which requires high-level instruction understanding and contains richer semantics.

Multi-goal and long-horizon benchmark is relatively under-explored in vision-language navigation. IVLN [12] proposes iterative VLN with step-by-step language instructions in R2R [1]. It builds navigation tours, each containing

an average of about 70 episodes as shown in Table I (#Sub-Goal). The navigation of each tour requires an **oracle phase** in which the agent is oracle-guided to the starting point of the next episode. We focus on a completely different and more practical VLN scenario with a sequence of high-level goal-oriented instructions. And no ground-truth guidance between episodes is needed in the proposed MG-VLN.

III. Multi-goal Vision-Language Navigation

This paper generalize the REVERIE [2] data to build a new multi-goal vision-language navigation (MG-VLN) task. Here, we define the MG-VLN task and metrics in detail and introduce how to construct the benchmark.

A. Task Formulation

In an episode of MG-VLN, the agent is required to follow *a sequence of* high-level language instructions, each guiding the agent to a specific goal. Assume there are m sub-goals in an episode. The agent receives m language instructions at the beginning and should execute each instruction in order. An example of 2-goal task is illustrated in Figure 2.

The paradigm of executing each instruction remains the same as the original REVERIE benchmark: The agent navigates in a discrete 3D environment, $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, where $\mathcal{V} = \{V_i\}_{i=1}^{K}$ denotes K navigable locations, and $\mathcal{E}$ denotes connectivity edges. At step t, the agent observes a panoramic view of its current node V_t. The panorama is split into 36 views $\mathcal{R}_t = \{r_i\}_{i=1}^{36}$ and each view is represented as an image feature. The navigable viewpoints provided for the agent are a subset of $\mathcal{R}_t$. Each viewpoint contains N objects $\mathcal{O}_t = \{o_i\}_{i=1}^{N}$ extracted from the panorama with bounding boxes. At each step, the agent makes actions navigating to a neighboring viewpoint or performs a STOP action with a prediction of the target object's location. Once the STOP action is called, the agent switches to the next sub-instruction until all instructions in an episode have been executed.

B. Generating MG-VLN Data

We generate MG-VLN data based on the instructions and navigation paths from REVERIE. REVERIE contains only single-goal high-level natural language instructions and requires only 4-7 navigation steps for each goal. To obtain multi-goal and long-horizon episodes with sequenced instructions, we adopt the data augmentation strategy in [16], concatenating two paths from REVERIE if the distance between the end and start points of the two paths is within a threshold d. The corresponding high-level instructions are stored in the order of path concatenation. For example, $(\mathcal{I}^a, \mathcal{P}^a)$ and $(\mathcal{I}^b, \mathcal{P}^b)$ are two instruction-path pairs from REVERIE. The instructions $\mathcal{I}^a$, $\mathcal{I}^b$ and paths $\mathcal{P}^a = \{p_1^a, p_2^a, ..., p_{n_a}^a\}$, $\mathcal{P}^b = \{p_1^b, p_2^b, ..., p_{n_b}^b\}$ are joined if $\mathrm{dis}\left(p_{n_a}^a, p_1^b\right) < d$, where dis is the shortest distance computed on the navigation graph and d is the threshold. The generated new instruction-path pair in an episode of the 2-goal MG-VLN task is formulated as $(\mathcal{I}', \mathcal{P}')$: $\mathcal{I}' = [\mathcal{I}_1, \mathcal{I}_2]$, $\mathcal{P}' = [\mathcal{P}_1, \mathcal{P}, \mathcal{P}_2]$, where $\mathcal{P}$ is the shortest path between $p_{n_a}^a$ and p_1^b. Similarly, we can generate new episodes for m-goal MG-VLN tasks.

Fig. 2. MG-VLN task: The panoramas of the starting and sub-goals' viewpoints in an 2-goal task example are shown. The red arrow represents the ground-truth navigation direction that should be chosen at the viewpoint. Following **a sequence of** high-level instructions, the agent should navigate to the appropriate location in order, and identify the target object (in red bounding box) from multiple distracting candidates (in blue bounding box).

C. Evaluation Metrics

The metrics of MG-VLN are calculated for a whole episode with multiple sub-goals. For each sub-goal, the navigation is considered successful only when the agent stops within 3 meters from the target and chooses the correct bounding box. One episode is considered successful if the agent finds all sub-goals in the correct order without exceeding the maximum steps. With the definition of an episode's success, we can adjust the metrics in REVERIE [2] to MG-VLN accordingly, such as Success Rate (SR), SR penalized by Path Length (SPL), Remote Grounding Success (RGS), and RGS penalized by Path Length (RGSPL). Besides standard metrics, we introduce a new metric named Progress with similar definition as in MultiON [8]: $\mathrm{Progress} = \frac{1}{M}\sum_{i=1}^{M}\sum_{j=1}^{m} S_{(i,j)}$, where M is the number of episodes, m is the number of sub-goals in an episode, $S_{(i,j)}$ is set to 1 if successful otherwise 0. This metric reflects the proportion of goals correctly located in the episode.

IV. Method

We study the construction of long-term memory in the proposed MG-VLN task. We first construct the memory module for two VLN baselines: the transformer-based navigator and the map-based navigator (Section IV-A). Then, we complement multiple types of linguistic knowledge to DUET's vision topological map, exploring what knowledge is useful for constructing long-term memory (Section IV-B).

A. Navigators with Long-term Memory

1) Transformer-based VLN Navigator: First, we adopt the transformer-based VLN navigator, History Aware Multimodal Transformer (HAMT) [13]. HAMT has an instruction encoder and vision encoder to encode the input text and visual observations. It has a transformer-based history encoder that encodes the full history of previous observations and actions while navigating to one goal. A cross-modal transformer fuses features from text, history and observation to predict the actions. To construct the long-term memory for HAMT during the navigation to multiple goals, we extend

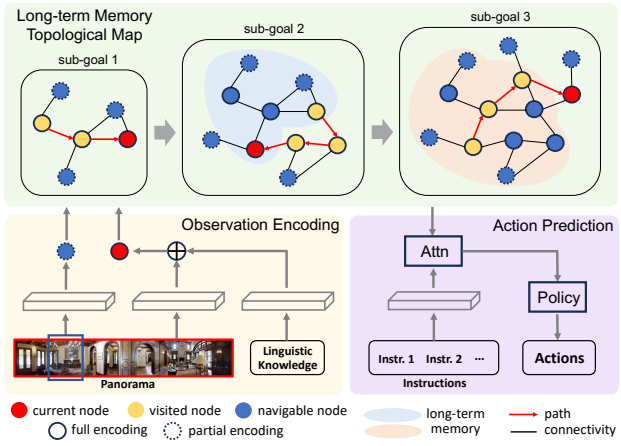

Fig. 3. The framework of constructing language-enhanced long-term memory for DUET navigator. Additional linguistic knowledge is injected into the node features in the topological map. When switching to the next sub-goal, the visited nodes are reset to navigable nodes. The fully and partially encoded nodes in the previous topological map are stored as long-term memory to benefit subsequent navigation.

the inputs of the history encoder with the observation-action pairs from previous sub-goals following IVLN [12]. We also unfreeze the history encoder to learn modified history encoding as in IVLN for multi-goal navigation.

2) Map-based VLN Navigator: Second, we adopt the map-based VLN navigator, DUET [14], which constructs a topological map while navigating to one goal. At each step, the topological map contains three types of nodes: the current node, visited nodes and navigable nodes. Each node represents a real viewpoint in the discrete 3D environment. The current and visited nodes' representations are fully encoded features fused from all 36 panorama views and extracted objects (nodes with solid border in Figure 3). The navigable nodes' representations encode only partial panorama views that can be observed from visited locations (nodes with dashed border in Figure 3). Navigable nodes serve as the candidates for the next navigation step, and the visited nodes can not be navigated again. The features of panorama, instructions and the topological map (extracted with GCN [63]) are used to predict navigation actions with a dual-scale graph transformer.

To construct long-term memory for DUET during the navigation to multiple goals, we reset the topological map only at the start of an episode. As shown in Figure 3, when the agent predicts the STOP action and switches to the next sub-goal, the representations of the existing nodes in the topological map are stored (serving as the long-term memory). The visited nodes in the topological map are reset to navigable nodes, allowing the agent to retrace its steps. When adding a new navigable node to the map, if it has been visited before, it uses the stored representations of viewpoints to construct the map. In this way, the memory module expands the navigable range and provides prior information for subsequent sub-goal navigation.

B. Language Enhanced Memory Map

The long-term memory module introduced above only contains visual information (e.g., view and object features in DUET). However, visual information alone may not be sufficient for vision-language navigation due to the imperfect extraction of semantic and linguistic knowledge by visual encoders. Linguistic knowledge can complement visual content and bridge the gap between vision and language modalities in VLN. As the focus of this paper is not the architecture design for understanding additional language knowledge, we simply inject the language knowledge into the view or node representations in the topological map, serving as another source of inputs. We study three types of language-enhanced memory maps introduced below.

1) Object Category Semantics: First, we encode object category semantics as in previous multi-object navigation methods [8], [10], [11]. Specifically, we use REVERIE ground-truth annotations of the object names contained in each viewpoint. We extract object name features with CLIP's text encoder [64] and transform the features with a lightweight learnable layer. The features are then added to the corresponding node features in the topological map.

2) Visual Captions from Foundation Model: Second, we study how the multi-modal knowledge encoded in current foundation models benefits the MG-VLN task. We crop each panorama view image into five sub-regions, and one caption is generated for each region with BLIP-2's image captioning ability. BLIP-2 [65] is an advanced foundation model that can translate visual observations into natural language. We also use CLIP's text encoder to extract caption features. The average caption features are added to the corresponding views' original vision features.

3) Object Attribute and Relationship Knowledge: Third, we extract object attributes and relationship knowledge for each panorama view following [66] and [56], which can assist object-centric scene understanding [67]. Specifically, a text knowledge base is first constructed by parsing Visual Genome [15] region descriptions into "attribute-object" pairs and "subject-predicate-object" triplets. Then, each view image is cropped into five sub-regions, and CLIP's vision and text encoders are used to compute similarity scores between view regions and texts from the knowledge base. The text features with top-5 highest scores are used as additional navigation inputs, which are averaged and added to the corresponding view features.

C. Training and Inference

As shown in [13], [14], it is beneficial to pre-train the transformer modules in VLN methods with proxy tasks. The proxy tasks for pre-training include masked language modeling (MLM) [68], masked region classification (MRC) [69], single-step action prediction (SAP) [13] and object grounding (OG) [70]. After pre-training, we first fine-tune a single-goal navigator with the original REVERIE dataset. The training strategy remains the same as the original methods. Using single-goal navigator model as initialization, we further fine-tune the model with generated m-goal MG-VLN data, obtaining m-goal navigators. For inference, the agent is forced to stop or switch to the next sub-instruction if it exceeds the maximum steps for one sub-goal in the episode.

TABLE II. Navigation performance on MG-VLN tasks (2-goal and 3-goal) and single-goal VLN task (1-goal). Multiple types of knowledge are used to construct DUET's memory map (Mem.). (V: Vision Information, S: Categories Semantics, C: Visual Captions, AR: Attribute and Relationship Knowledge)

#	Model	Mem.	SR↑			Progress↑			SPL↑			RGS↑			RGSPL↑		
			1-goal	2-goal	3-goal	1-goal	2-goal	3-goal	1-goal	2-goal	3-goal	1-goal	2-goal	3-goal	1-goal	2-goal	3-goal
1	HAMT	×	32.95	13.84	5.21	-	26.34	25.97	30.20	9.21	4.39	18.92	4.35	2.13	17.28	3.01	1.05
2		✓	32.95	12.37	4.54	-	23.89	23.47	30.20	8.57	3.65	18.92	4.14	2.20	17.28	2.56	1.01
3	DUET	×	48.79	21.21	12.44	-	41.05	40.96	32.90	10.36	6.67	32.89	9.83	4.93	21.79	4.71	2.64
4		V	48.79	23.97	14.67	-	43.76	43.19	32.90	11.38	6.89	32.89	13.77	3.64	21.79	6.87	1.80
5		V+S	46.92	23.42	13.38	-	44.95	43.43	33.03	13.24	6.70	31.92	13.59	5.16	22.25	7.64	2.35
6		V+C	48.40	24.89	14.08	-	46.19	45.27	34.09	12.99	6.23	32.75	14.05	5.05	22.83	8.15	2.25
7		V+AR	48.31	**27.36**	**20.19**	-	**47.66**	**50.00**	33.27	**14.61**	**11.39**	33.14	**16.16**	**7.63**	22.78	**8.77**	**4.13**

TABLE III. Navigation performance of each separate sub-goal on MG-VLN. Δ is the improvement of constructing long-term memory.

Memory	2-goal		3-goal		
	1st. OSR↑	2nd. OSR↑	1st. OSR↑	2nd. OSR↑	3rd. OSR↑
×	50.14	58.31	51.06	64.67	44.37
✓	51.24	64.74	55.99	71.83	57.16
Δ	+1.10	+6.43	+4.93	+7.16	+12.79

V. EXPERIMENTS

Here, we describe the experimental setup of the proposed MG-VLN benchmarks and implementation details of multi-goal VLN navigators. We seek to answer the following questions: (1) Is constructing long-term memory beneficial for MG-VLN tasks? (2) What knowledge is useful for constructing long-term memory? (3) How does the memory module work when navigating to multiple goals?

A. Experimental Setup

Our experiments are conducted on Matterport3D simulator [71] with vision-language navigation annotations from REVERIE [2]. REVERIE contains high-level natural language instructions and annotates visible object names and locations for each discrete viewpoint in the simulator. Its instructions contain 21 words on average, and its path has 4-7 navigation steps. We use the original data splits for single-goal navigator training and evaluation.

We generate MG-VLN data for two goals (2-goal) and three goals (3-goal) benchmarks based on Matterport3D and REVERIE using standard train and val (unseen) split. 1-goal benchmark is the original single-goal VLN task. We do not use the val seen split because there is no need for saving long-term memory for seen environments, as demonstrated in MultiON [8] and MON [11]. For each episode in m-goal MG-VLN benchmark, m paths from REVERIE are joined end to end with the distance threshold of 1 meter ($d = 1\mathrm{m}$), as introduced in Section III. The train split contains at most $1,000$ episodes per scene, and the validation (unseen) split contains at most 200 episodes per scene.

The hyper-parameters for training single-goal navigators follow the original settings in HAMT [13] and DUET [14]. We set the batch size as 16 for 100k iterations pre-training and 4 for 20k iterations fine-tuning. We do not use the synthetic instructions in DUET to augment the dataset. We train the multi-goal navigators with a batch size of 2 for 20k iterations on one GPU. The maximum action steps are set to 15 for a sub-goal. We use BLIP-2 (FlanT5$_{\mathrm{XL}}$) [65] and CLIP (ViT-B/16) [64] to generate visual captions and retrieve attribute/relationship knowledge for panorama views. The best models are selected by SPL on val unseen split.

B. Results

*1) **Is long-term memory beneficial for MG-VLN?:*** From rows 1 and 3 in Table II, we can observe that the navigation performances decrease dramatically with the increase of sub-goal numbers, indicating that multi-goal vision-language navigation is quite challenging. This may be because the entire journey is long-horizon and covers a relatively large area of the 3D environment. Deviating from the previous sub-goal may increase the difficulty of the subsequent sub-goal navigation in an episode.

From the first two rows in Table II, we observe that extending the history inputs as implicit memory for transformer-based navigator (HAMT) degrades MG-VLN performances. A similar phenomenon has been observed in IR2R [12] experiments. This may be because the long-term memory of HAMT encodes the previous observations in an implicit way (through hidden states). The history encoder in HAMT has a distributed shift in its inputs compared with the pre-trained task, as demonstrated in [12]. Rows 3 and 4 in Table II show that simply constructing long-term memory with the original **vision** information in DUET's topological map is beneficial for MG-VLN task in most cases. This indicates that the agent can utilize the explicit map-form environment information collected from previous goal navigation to improve the efficiency of subsequent navigation.

*2) **What knowledge is useful for constructing memory?:*** Rows 5-7 in Table II shows the results of introducing multiple types of language-enhanced knowledge. All methods outperform DUET baseline with no memory module (row 3). Although injecting object category semantics has been shown to be effective for multi-object navigation in previous works, it has comparable performance compared to individual vision memory in MG-VLN (row 4 vs. 5). We speculate this is due to DUET itself already encodes fine-grained category semantics from object bounding boxes. Simply introducing category semantics hardly helps MG-VLN tasks, which require high-level language understanding.

Rows 6 and 7 in Table II show that injecting BLIP-2 caption and retrieved object attribute/relationship knowledge are both beneficial for 2-goal and 3-goal tasks, although the results of 1-goal (original single-goal VLN task) show less improvements compared with DUET baseline (row 3). This indicates that the original 1-goal benchmark may not be suitable to reflect the advantages of language-enhanced

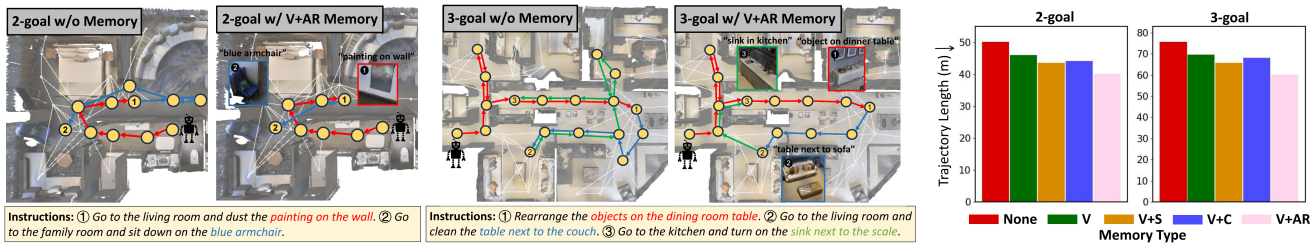

Fig. 4. **Left:** Predicted trajectories of DUET baseline *without* memory compared with our *V+AR (Attribute and Relationship Knowledge)* memory on 2-goal and 3-goal tasks. The red/blue/green arrows denote the trajectories of finding the first/second/third sub-goal. We also show the retrieved knowledge text for the view image. **Right:** The average trajectory length (in meters) of the DUET navigator with multiple types of memory.

knowledge map. Because the paths in the 1-goal task have only 4-7 steps, and the injected language knowledge is rather limited. The complementary language knowledge is more efficient for multi-goal vision-language navigation that requires long-horizon reasoning.

Among multiple types of language knowledge, the object attribute and relationship knowledge retrieved from Visual Genome [15] descriptions bring significantly more performance gains in 2-goal and 3-goal VLN tasks. This suggests that rich attribute-object pairs and subject-predicate-object triplets better complement the visual semantics in the memory map. They serve as strong memory cues and help when detailed object attribute and location information is needed for navigation, such as "*Pick up the black box on top of the drawers*". In contrast, captions generated by BLIP-2 have more complicated sentence patterns and structures, which makes it difficult for CLIP text encoders to extract the specific knowledge needed to execute navigation instructions. More complex prompts regarding generating specific kinds of visual captions can be explored in future work.

*3) **How does memory map work in MG-VLN?**:* To further explore the effect of long-term memory in multi-goal vision-language navigation, we analyze the navigation performance of each sub-goal separately. In Table III, we show the oracle success rate (OSR) of the first/second/third sub-goal in 2-goal and 3-goal tasks (shortened to 1st./2nd./3rd. OSR). The m^{th}-goal navigation is considered oracle successful if the goal can be observed anywhere while following the m^{th} instruction. From the last row in Table III, we observe that by constructing a long-term memory module, the improvements of the second goal navigation are larger than the first goal, and the improvements of the third goal are larger than the second. This is consistent with the intuition that the effect of long-term memory becomes more pronounced with the increase of memory length and navigation horizon.

To further explore whether long-term memory helps the agent remember the information of subsequent goals that was seen before, we show the success rate of finding *seen* and *unseen* sub-goals in MG-VLN tasks in Table IV. An object goal is considered *seen* if it appears within a viewpoint included in the previous navigation path and no obstacle exists between the viewpoint and the object. We show the differences in oracle success rate between the seen and unseen goals for DUET baseline with or without memory and with the most effective attribute and relationship memory. The results suggest that if a subsequent goal is seen while navigating to previous goals, the experience would improve its navigation efficiency in most cases (refer to the line Δ). This phenomenon is much weaker for DUET baseline since no explicit memory is stored. The advantage of the memory module is the most obvious for attribute and relationship-enhanced knowledge since it provides the most useful information to understand high-level instructions.

TABLE IV. Navigation performance of seen and unseen sub-goals.

Memory	2nd. SR↑ (2-goal)			2nd. SR↑ (3-goal)			3rd. SR↑ (3-goal)		
	Seen	Unseen	Δ	Seen	Unseen	Δ	Seen	Unseen	Δ
×	43.82	40.93	+2.89	41.70	44.30	-2.60	43.97	29.23	+14.74
V	51.56	42.38	+9.18	47.96	43.40	+4.56	47.79	30.77	+17.02
V+AR	60.58	44.42	**+16.16**	60.57	48.02	**+12.55**	56.93	30.84	**+26.09**

*4) **Efficiency analysis**:* Beyond success rate, we further demonstrate the efficiency improvements brought by language-enhanced memory. We show the qualitative and quantitative comparison results of 2-goal and 3-goal tasks in Figure 4.

Qualitative results: The left of Figure 4 shows the trajectories predicted by DUET with and without memory module. In 2-goal examples, while navigating to the first sub-goal, the agent has passed the second sub-goal (the blue armchair). With long-term memory, the corresponding knowledge and viewpoint are stored, and the agent is able to navigate to the second goal efficiently after finding the first one. In 3-goal examples, while navigating to the first goal, the agent passed the living room (partial encoding) and the kitchen with sink and scale (full encoding). With long-term memory, this stored environmental information helps the agent successfully find the second and third goals in the correct order without random exploration. In contrast, the baseline without a memory module will first randomly explore the environment searching for the subsequent goal after finding the previous sub-goal.

We show the "attribute-object" pair and "subject-predicate-object" triplet with the highest similarity scores retrieved from Visual Genome [15] descriptions for each view image. The retrieved knowledge is well-aligned with the high-level instructions and helps construct richer long-term memory.

Quantitative results: The right of Figure 4 shows the average trajectory length in meters (less is better) of the *successful* episodes of DUET with multiple types of memory. The trajectory length decreases when the memory map is constructed, especially for the *V+AR* memory map, indicating our method's benefit in navigation efficiency.

VI. CONCLUSIONS

We propose a multi-goal long-horizon vision-language navigation (MG-VLN) task. This task is more consistent with the instruction following paradigm in real-world scenarios. In MG-VLN, we study the fundamental question of long-term memory construction with transformer-based and map-based VLN methods and show that the topological memory map improves multi-goal navigation efficiency. Besides, We study multiple types of language-enhanced memory on MG-VLN tasks with map-based navigators. Among the three types of language knowledge, object attribute and relationship knowledge is the most effective in establishing the mapping between environment and high-level instructions and brings the best multi-goal navigation efficiency.

ACKNOWLEDGMENT

This research was partially supported by National Key R&D Program of China (2022YFB2804103), Key Research and Development Program of Shaanxi (2021ZDLGY01-05), Tsinghua University Dushi Program, National Natural Science Foundation of China (20211710187), and Tsinghua University Talent Program.

REFERENCES

[1] Anderson, P., Wu, Q., Teney, D., Bruce, J., Johnson, M., Sünderhauf, N., Reid, I.D., Gould, S., & Hengel, A.V. (2018). Vision-and-Language Navigation: Interpreting Visually-Grounded Navigation Instructions in Real Environments. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 3674-3683.

[2] Qi, Y., Wu, Q., Anderson, P., Wang, X.E., Wang, W.Y., Shen, C., & Hengel, A.V. (2019). REVERIE: Remote Embodied Visual Referring Expression in Real Indoor Environments. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 9979-9988.

[3] Zhu, F., Liang, X., Zhu, Y., Chang, X., & Liang, X. (2021). SOON: Scenario Oriented Object Navigation with Graph-based Exploration. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 12684-12694.

[4] Chang, M., Gervet, T., Khanna, M., Yenamandra, S., Shah, D., Min, S.Y., Shah, K., Paxton, C., Gupta, S., Batra, D., Mottaghi, R., Malik, J., & Chaplot, D.S. (2023). GOAT: GO to Any Thing. ArXiv, abs/2311.06430.

[5] Liu, P., Orru, Y., Paxton, C., Shafiullah, N.M., & Pinto, L. (2024). OK-Robot: What Really Matters in Integrating Open-Knowledge Models for Robotics. ArXiv, abs/2401.12202.

[6] Rana, K., Haviland, J., Garg, S., Abou-Chakra, J., Reid, I.D., & Sünderhauf, N. (2023). SayPlan: Grounding Large Language Models using 3D Scene Graphs for Scalable Task Planning. Conference on Robot Learning.

[7] Fang, K., Toshev, A., Fei-Fei, L., & Savarese, S. (2019). Scene Memory Transformer for Embodied Agents in Long-Horizon Tasks. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 538-547.

[8] Wani, S., Patel, S., Jain, U., Chang, A., & Savva, M. (2020). Multion: Benchmarking semantic map memory using multi-object navigation. Advances in Neural Information Processing Systems, 33, 9700-9712.

[9] Marza, P., Matignon, L., Simonin, O., & Wolf, C. (2022). Teaching Agents how to Map: Spatial Reasoning for Multi-Object Navigation. 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 1725-1732.

[10] Raychaudhuri, S., Campari, T., Jain, U., Savva, M., & Chang, A.X. (2023). Reduce, Reuse, Recycle: Modular Multi-Object Navigation. ArXiv, abs/2304.03696.

[11] Zeng, H., Song, X., & Jiang, S. (2023). Multi-Object Navigation Using Potential Target Position Policy Function. IEEE Transactions on Image Processing, 32, 2608-2619.

[12] Krantz, J., Banerjee, S., Zhu, W., Corso, J.J., Anderson, P., Lee, S., & Thomason, J. (2023). Iterative Vision-and-Language Navigation. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 14921-14930.

[13] Chen, S., Guhur, P. L., Schmid, C., & Laptev, I. (2021). History aware multimodal transformer for vision-and-language navigation. Advances in neural information processing systems, 34, 5834-5847.

[14] Chen, S., Guhur, P., Tapaswi, M., Schmid, C., & Laptev, I. (2022). Think Global, Act Local: Dual-scale Graph Transformer for Vision-and-Language Navigation. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 16516-16526.

[15] Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L., Shamma, D.A., Bernstein, M.S., & Fei-Fei, L. (2016). Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. International Journal of Computer Vision, 123, 32 - 73.

[16] Jain, V., Ilharco, G., Ku, A., Vaswani, A., Ie, E., & Baldridge, J. (2019). Stay on the Path: Instruction Fidelity in Vision-and-Language Navigation. Annual Meeting of the Association for Computational Linguistics.

[17] Ku, A., Anderson, P., Patel, R., Ie, E., & Baldridge, J. (2020). Room-Across-Room: Multilingual Vision-and-Language Navigation with Dense Spatiotemporal Grounding. Conference on Empirical Methods in Natural Language Processing.

[18] Thomason, J., Murray, M., Cakmak, M., & Zettlemoyer, L. (2019). Vision-and-Dialog Navigation. Conference on Robot Learning.

[19] Chen, H., Suhr, A., Misra, D.K., Snavely, N., & Artzi, Y. (2018). TOUCHDOWN: Natural Language Navigation and Spatial Reasoning in Visual Street Environments. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 12530-12539.

[20] Fried, D., Hu, R., Cirik, V., Rohrbach, A., Andreas, J., Morency, L. P., ... & Darrell, T. (2018). Speaker-follower models for vision-and-language navigation. Advances in Neural Information Processing Systems, 31.

[21] Tan, H., Yu, L., & Bansal, M. (2019). Learning to Navigate Unseen Environments: Back Translation with Environmental Dropout. North American Chapter of the Association for Computational Linguistics.

[22] Ma, C., Wu, Z., Al-Regib, G., Xiong, C., & Kira, Z. (2019). The Regretful Agent: Heuristic-Aided Navigation Through Progress Estimation. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 6725-6733.

[23] Zhu, F., Zhu, Y., Chang, X., & Liang, X. (2019). Vision-Language Navigation With Self-Supervised Auxiliary Reasoning Tasks. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 10009-10019.

[24] Qi, Y., Pan, Z., Hong, Y., Yang, M., Hengel, A.V., & Wu, Q. (2021). The Road to Know-Where: An Object-and-Room Informed Sequential BERT for Indoor Vision-Language Navigation. 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 1635-1644.

[25] Guhur, P., Tapaswi, M., Chen, S., Laptev, I., & Schmid, C. (2021). Airbert: In-domain Pretraining for Vision-and-Language Navigation. 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 1614-1623.

[26] Hong, Y., Wu, Q., Qi, Y., Rodriguez-Opazo, C., & Gould, S. (2020). VLN BERT: A Recurrent Vision-and-Language BERT for Navigation. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 1643-1653.

[27] Qiao, Y., Qi, Y., Hong, Y., Yu, Z., Wang, P., & Wu, Q. (2022). HOP: History-and-Order Aware Pretraining for Vision-and-Language Navigation. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 15397-15406.

[28] Qiao, Y., Qi, Y., Hong, Y., Yu, Z., Wang, P., & Wu, Q. (2023). HOP+: History-Enhanced and Order-Aware Pre-Training for Vision-and-Language Navigation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45, 8524-8537.

[29] Huo, J., Sun, Q., Jiang, B., Lin, H., & Fu, Y. (2023). GeoVLN: Learning Geometry-Enhanced Visual Representation with Slot Attention for Vision-and-Language Navigation. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 23212-23221.

[30] Chen, T., Gupta, S., & Gupta, A. (2018, September). Learning Exploration Policies for Navigation. In International Conference on Learning Representations (ICLR).

[31] Ramakrishnan, S.K., Jayaraman, D., & Grauman, K. (2020). An Exploration of Embodied Visual Exploration. International Journal of Computer Vision, 129, 1616 - 1649.

[32] Chaplot, D. S., Gandhi, D., Gupta, S., Gupta, A., & Salakhutdinov, R. (2020). Learning To Explore Using Active Neural SLAM. In International Conference on Learning Representations (ICLR).

[33] Chen, P., Ji, D., Lin, K., Zeng, R., Li, T., Tan, M., & Gan, C. (2022). Weakly-supervised multi-granularity map learning for vision-and-language navigation. Advances in Neural Information Processing Systems, 35, 38149-38161.

[34] Anderson, P., Shrivastava, A., Parikh, D., Batra, D., & Lee, S. (2019). Chasing Ghosts: Instruction Following as Bayesian State Tracking. Neural Information Processing Systems.

[35] Irshad, M., Mithun, N.C., Seymour, Z., Chiu, H., Samarasekera, S., & Kumar, R. (2022). Semantically-aware Spatio-temporal Reasoning Agent for Vision-and-Language Navigation in Continuous Environments. International Conference on Pattern Recognition (ICPR), 4065-4071.

[36] Deng, Z., Narasimhan, K., & Russakovsky, O. (2020). Evolving graphical planner: Contextual global planning for vision-and-language navigation. Advances in Neural Information Processing Systems, 33, 20660-20672.

[37] Chen, K., Chen, J., Chuang, J., V'azquez, M., & Savarese, S. (2021). Topological Planning with Transformers for Vision-and-Language Navigation. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 11271-11281.

[38] Wang, H., Wang, W., Liang, W., Xiong, C., & Shen, J. (2021). Structured Scene Memory for Vision-Language Navigation. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 8451-8460.

[39] Gao, C., Peng, X., Yan, M., Wang, H., Yang, L., Ren, H., Li, H., & Liu, S. (2023). Adaptive Zone-aware Hierarchical Planner for Vision-Language Navigation. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 14911-14920.

[40] An, D., Qi, Y., Li, Y., Huang, Y., Wang, L., Tan, T., & Shao, J. (2022). BEVBert: Multimodal Map Pre-training for Language-guided Navigation. ArXiv, abs/2212.04385.

[41] Wang, Z., Li, X., Yang, J., Liu, Y., & Jiang, S. (2023). GridMM: Grid Memory Map for Vision-and-Language Navigation. 2023 IEEE/CVF International Conference on Computer Vision (ICCV).

[42] Liu, R., Wang, X., Wang, W., & Yang, Y. (2023). Bird's-Eye-View Scene Graph for Vision-Language Navigation. 2023 IEEE/CVF International Conference on Computer Vision (ICCV).

[43] Liu, C., Zhu, F., Chang, X., Liang, X., & Shen, Y. (2021). Vision-Language Navigation with Random Environmental Mixup. 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 1624-1634.

[44] Koh, J. Y., Lee, H., Yang, Y., Baldridge, J., & Anderson, P. (2021). Pathdreamer: A world model for indoor navigation. 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 14738-14748.

[45] Li, J., Tan, H., & Bansal, M. (2022). Envedit: Environment Editing for Vision-and-Language Navigation. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 15386-15396.

[46] Wang, H., Liang, W., Shen, J., Gool, L.V., & Wang, W. (2022). Counterfactual Cycle-Consistent Learning for Instruction Following and Generation in Vision-Language Navigation. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 15450-15460.

[47] Wang, S., Montgomery, C., Orbay, J., Birodkar, V., Faust, A., Gur, I., Jaques, N., Waters, A., Baldridge, J., & Anderson, P. (2022). Less is More: Generating Grounded Navigation Instructions from Landmarks. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 15407-15417.

[48] Wang, X., Wang, W., Shao, J., & Yang, Y. (2023). LANA: A Language-Capable Navigator for Instruction Following and Generation. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 19048-19058.

[49] Kamath, A., Anderson, P., Wang, S., Koh, J.Y., Ku, A., Waters, A., Yang, Y., Baldridge, J., & Parekh, Z. (2023). A New Path: Scaling Vision-and-Language Navigation with Synthetic Instructions and Imitation Learning. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 10813-10823.

[50] Li, J., & Bansal, M. (2023). PanoGen: Text-Conditioned Panoramic Environment Generation for Vision-and-Language Navigation. ArXiv, abs/2305.19195.

[51] Li, J., & Bansal, M. (2023). Improving Vision-and-Language Navigation by Generating Future-View Image Semantics. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 10803-10812.

[52] Wang, Z., Li, J., Hong, Y., Wang, Y., Wu, Q., Bansal, M., Gould, S., Tan, H., & Qiao, Y. (2023). Scaling Data Generation in Vision-and-Language Navigation. 2023 IEEE/CVF International Conference on Computer Vision (ICCV).

[53] Hong, Y., Rodriguez-Opazo, C., Qi, Y., Wu, Q., & Gould, S. (2020). Language and Visual Entity Relationship Graph for Agent Navigation. Advances in Neural Information Processing Systems.

[54] Shen, S., Li, C., Hu, X., Xie, Y., Yang, J., Zhang, P., ... & Gao, J. (2022). K-lite: Learning transferable visual models with external knowledge. Advances in Neural Information Processing Systems, 35, 15558-15573.

[55] Majumdar, A., Shrivastava, A., Lee, S., Anderson, P., Parikh, D., & Batra, D. (2020). Improving Vision-and-Language Navigation with Image-Text Pairs from the Web. In European Conference on Computer Vision (ECCV), 259-274.

[56] Li, X., Wang, Z., Yang, J., Wang, Y., & Jiang, S. (2023). KERM: Knowledge Enhanced Reasoning for Vision-and-Language Navigation. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2583-2592.

[57] Savva, M., Kadian, A., Maksymets, O., Zhao, Y., Wijmans, E., Jain, B., Straub, J., Liu, J., Koltun, V., Malik, J., Parikh, D., & Batra, D. (2019). Habitat: A Platform for Embodied AI Research. 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 9338-9346.

[58] Kim, J., Lee, E. S., Lee, M., Zhang, D., & Kim, Y. M. (2021). Sgolam: Simultaneous goal localization and mapping for multi-object goal navigation. ArXiv:2110.07171.

[59] Chen, P., Ji, D., Lin, K., Hu, W., Huang, W., Li, T., ... & Gan, C. (2022). Learning active camera for multi-object navigation. Advances in Neural Information Processing Systems, 35, 28670-28682.

[60] Marza, P., Matignon, L., Simonin, O., & Wolf, C. (2022). Multi-Object Navigation with dynamically learned neural implicit representations. ArXiv, abs/2210.05129.

[61] Ramakrishnan, S. K., Gokaslan, A., Wijmans, E., Maksymets, O., Clegg, A., Turner, J. M., ... & Batra, D. (2021). Habitat-Matterport 3D Dataset (HM3D): 1000 Large-scale 3D Environments for Embodied AI. In Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track.

[62] Gireesh, N., Agrawal, A., Datta, A., Banerjee, S., Sridharan, M., Bhowmick, B., & Krishna, M. (2023). Sequence-Agnostic Multi-Object Navigation. 2023 IEEE International Conference on Robotics and Automation (ICRA), 9573-9579.

[63] Kipf, T., & Welling, M. (2016). Semi-Supervised Classification with Graph Convolutional Networks. In International Conference on Learning Representations (ICLR).

[64] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision. International Conference on Machine Learning.

[65] Li, J., Li, D., Savarese, S., & Hoi, S.C. (2023). BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. ArXiv, abs/2301.12597.

[66] Kuo, C., & Kira, Z. (2022). Beyond a Pre-Trained Object Detector: Cross-Modal Textual and Visual Context for Image Captioning. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 17948-17958.

[67] Zhang, J., Fan, G., Wang, G., Su, Z., Ma, K., & Yi, L. (2023). Language-assisted 3D feature learning for semantic scene understanding. AAAI Conference on Artificial Intelligence (AAAI).

[68] Kenton, J. D. M. W. C., & Toutanova, L. K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of NAACL-HLT, 4171-4186.

[69] Lu, J., Batra, D., Parikh, D., & Lee, S. (2019). ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. Neural Information Processing Systems.

[70] Lin, X., Li, G., & Yu, Y. (2021). Scene-Intuitive Agent for Remote Embodied Visual Grounding. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 7032-7041.

[71] Chang, A.X., Dai, A., Funkhouser, T.A., Halber, M., Nießner, M., Savva, M., Song, S., Zeng, A., & Zhang, Y. (2017). Matterport3D: Learning from RGB-D Data in Indoor Environments. 2017 International Conference on 3D Vision (3DV), 667-676.