VLN-KHVR: Knowledge-and-History Aware Visual Representation for Continuous Vision-and-Language Navigation

Ping Kong[1], Ruonan Liu[2], Zongxia Xie[1], Zhibo Pang[3]

***Abstract*— Vision-and-Language Navigation in Continuous Environments (VLN-CE) requires agents to navigate with low-level actions following natural language instructions in 3D environments. Most existing approaches utilize observation features from the current step to represent the viewpoint. However, these representations often conflate redundant and essential information for navigation, introducing ambiguity into the agent's action prediction. To address the problem of inadequate representation, we propose a Knowledge-and-History Aware Visual Representation for Continuous Vision-and-Language Navigation (VLN-KHVR). The proposed approach constructs enriched visual representations tailored to navigation instructions, enhancing agents' navigation performance. Specifically, VLN-KHVR extracts image features from the current observation, retrieves relevant knowledge in the knowledge base, and obtains the history of the navigation episode. Subsequently, the knowledge and history features are filtered to eliminate the information irrelevant to navigation instruction. These refined features are integrated with the instruction for further interaction. Finally, the aggregated features are used to guide navigation. Our model outperforms previous methods on the VLN-CE benchmark, demonstrating the effectiveness of the proposed method.**

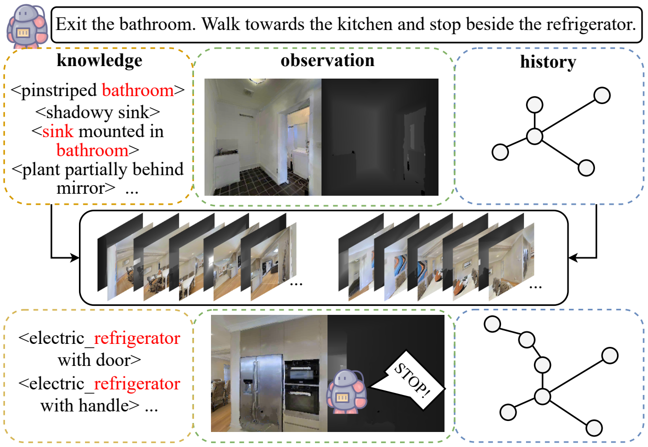

Exit the bathroom. Walk towards the kitchen and stop beside the refrigerator.
knowledge
observation
history
<pinstriped bathroom>
<shadowy sink>
<sink mounted in bathroom>
<plant partially behind mirror> ...
<electric_refrigerator with door>
<electric_refrigerator with handle> ...
STOP!

Fig. 1. Introduction of the proposed VLN-KHVR. At each step, the agent extracts knowledge features associated with the current view while retaining the topological map features from the previous step. This integration of knowledge and history within the visual representation is beneficial for navigation.

I. INTRODUCTION

Enhancing embodied artificial intelligence to comprehend and execute human commands stands as a pivotal objective in robotics research. To achieve this goal, vision-and-language navigation (VLN) [1] has been proposed, attracting increasing attention in recent years [2], [3], [4]. VLN requires an agent to navigate in a dynamic environment following natural language instructions, and ultimately reach a target location. In the initial version of VLN, referred to as discrete VLN, the agent is provided with a connectivity graph and teleports among a set of predefined sparse waypoints. Transitioning towards a more practical scenario, vision-and-language navigation in continuous environments (VLN-CE) [5] employs a continuous setting with low-level actions, which more closely mirrors real-world conditions.

Due to the removal of the simplifying assumptions inherent in discrete VLN, VLN-CE has been proven more challenging than its discrete counterpart, prompting a surge in research aimed at narrowing the gap between the two tasks. Early VLN-CE works are end-to-end systems that directly predict low-level actions [6], [7]. With the emergence and popularity of the waypoint predictor [8], subsequent approaches [9], [10], [11] leverage this pre-trained model to forecast candidate navigable waypoints and modularize the entire system for predicting actions. Most existing models utilize the currently observed image information of the viewpoint as a visual representation [10], [12], [13]. However, this representation often results in a conflation of redundant and essential cues in the current observation, making it difficult for the agent to concentrate on the most effective visual information, which is insufficient for proper action prediction.

In this work, to address the aforementioned problems, we propose a Knowledge-and-History Aware Visual Representation for Continuous Vision-and-Language Navigation (VLN-KHVR). By incorporating knowledge and history into the navigation view, we construct a visual representation framework that is not only more comprehensive but also more relevant to the navigation instruction, enabling the agent to make more precise decisions.

Knowledge offers a refined and complete abstraction of objects and their interrelationships, enriching the information of the current visual content. For example, as shown in Fig. 1, when the agent observes a sink, it can infer that it is likely present in a bathroom based on knowledge. This inferential capability is beneficial for comprehensively understanding the surroundings. Additionally, since the agent is trained in limited environments, the alignment ability between instruc-

[1]Ping Kong and Zongxia Xie are with the College of Intelligence and Computing, Tianjin University, Tianjin 300350, China. kongping@tju.edu.cn, zongxiaxie@tju.edu.cn

[2]Ruonan Liu is with Department of Automation, Shanghai Jiao Tong University, Shanghai, China. ruonan.liu@sjtu.edu.cn

[3]Zhibo Pang is with the KTH Royal Institute of Technology, Stockholm, Sweden, and ABB Corporate Research, Västerås, Sweden. zhibo@kth.se

tions and key objects in the view remains inadequate. The integration of external knowledge, without specific regularity, reinforces cross-modal alignment and boosts the agent's generalization capabilities. Consequently, we introduce commonsense knowledge into the model, directing the agent's focus to objects that are pertinent to the current instruction. Furthermore, to enhance perceptual coherence across views during navigation, we explicitly incorporate historical information into the current visual representation. This addition serves as a navigational memory, facilitating action reasoning. Specifically, the processing of visual representation in VLN-KHVR is divided into four distinct stages: extraction, filtering, interaction, and aggregation. Firstly, the extraction module is employed to capture the current visual observation features, retrieve the relevant knowledge of the image, and leverage the topological map to extract the history features. Subsequently, the filtering module calculates the relevance weights between the features and the instruction, purifying the instruction-related features. Following this, the interaction module facilitates the exchange of perceptual information among knowledge, history, and instructions. Finally, the aggregation module combines all the features to obtain the ultimate visual representation. As illustrated in Fig. 1, our model can make correct decisions by explicitly integrating knowledge and history into the visual representations.

The experiments are conducted on the VLN-CE dataset [10] and results indicate that our approach outperforms state-of-the-art methods. Specifically, VLN-KHVR boosts the success rate by 3.2% on the validation-unseen split and 1.8% on the test-unseen split, while reducing navigation error by 4.0% and 1.4%, respectively. Additional experimental analysis further demonstrated the effectiveness of our method.

In summary, the contributions of this work are as follows:

- We integrate external knowledge and record the topological map features from the previous step in the navigation episode, which comprehensively characterizes the navigation view, enhancing alignment between vision and text.
- We propose the VLN-KHVR model to explicitly incorporate knowledge and history features into visual representations for VLN-CE, which augments the accuracy of action prediction.
- We conduct extensive experiments on the VLN-CE dataset to verify the effectiveness and generalization ability of VLN-KHVR. Results show that our approach improves the success rate while reducing the navigation error compared to existing methods.

II. Related Work

A. Vision-and-Language Navigation

Vision-and-Language Navigation (VLN), where agents navigate based on natural language instructions, has drawn considerable research interest. The inception of the first VLN study occurred in 2018, introducing the benchmark dataset R2R [1]. Since then, numerous task variants [14], [15], [16], [17] have been proposed, where REVERIE [16] emphasizes exploring the environment and localizing a referenced target object according to concise instructions, and VLN - CE [5] transforms topologically defined VLN tasks into continuous environments. Early VLN methods focus on action strategy learning with the assistance of reinforcement learning or auxiliary tasks [18], [19], [20]. To enhance the agent's generalization ability, data augmentation approaches are studied to extend existing data or create synthetic data [21], [22], [23]. Moreover, another line of works investigate the memory for navigation. Among them, recursive cells [2], [24], [25], explicitly encoded history sequences [3], [26], topological maps [4], [10], [27], semantic maps [12], [28], and grid-based maps [29] are commonly employed to represent the visited environment. Distinct from previous works, our approach aims to generate robust and generalizable visual representations dedicated to navigation.

B. Vision-and-Language Navigation with Knowledge

Cross-modal reasoning using external knowledge has been demonstrated to be effective [30], [31], [32], [33], and the integration of knowledge in VLN has been explored in several ways [34], [35], [36]. CKR [34] utilized ConceptNet [37] to learn the internal-external correlations among room and object entities during training, enabling the agent to make informed decisions at each viewpoint. KERM [35] proposes a knowledge-enhanced reasoning model that incorporates region-centric knowledge to comprehensively depict navigation views, thus obtaining information complementary to visible content. ACK [36] employs commonsense information to construct a spatio-temporal knowledge graph, enhancing the vision-text alignment. However, all these works are developed for discrete VLN. In contrast, our approach is designed for the more challenging VLN-CE, which benefits navigation performance by incorporating relevant knowledge and history into visual representations.

III. Method

We propose VLN-KHVR, an enhanced visual representation method, to improve navigation performance in VLN-CE. VLN-CE executes navigation through a sequence of low-level actions consisting of FORWARD 0.25m, TURN LEFT/RIGHT 15°, and STOP. At each step t, the agent captures panoramic observations $O_t = \left\{o_{t,i}^{rgb}, o_{t,i}^{depth}\right\}_{i=1}^{12}$ from the current viewpoint, consisting of 12 RGB images and 12 depth images spaced at 30° intervals. By integrating knowledge and historical information into the observation images, we provide a more effective visual representation for cross-modal alignment. In the following subsections, we first present a general overview of our proposed model. Then, we detail the critical modules for constructing the informative visual representation separately.

A. Model Overview

Fig. 2 illustrates our approach, which leverages knowledge and history to achieve effective visual representations. Initially, the extraction module acquires the current observations including RGB images and depth images, retrieves

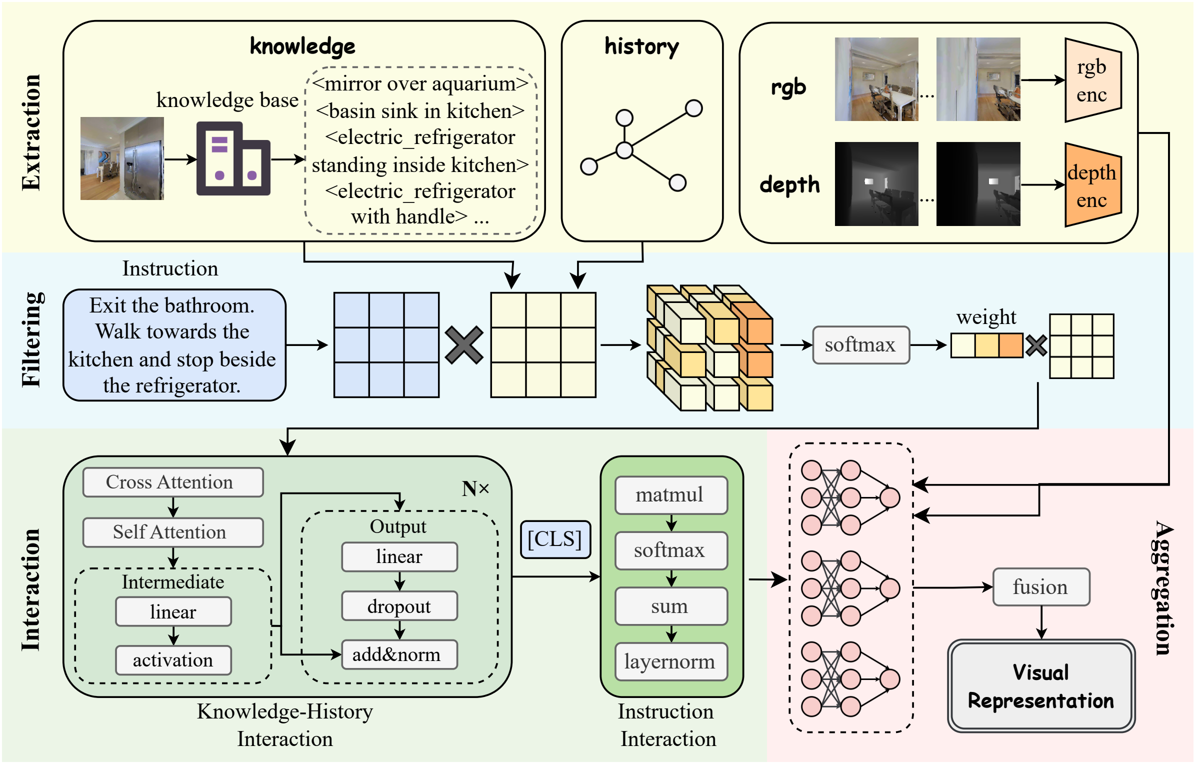


Fig. 2. Overall pipeline of the proposed approach. Firstly, the extraction module is used to obtain the observation features, knowledge features, and history features of the current step. Then, the filtering module purifies the features related to the instruction. Additionally, the interaction module performs information exchange on knowledge, history, and instruction. Finally, the aggregation module combines all features to obtain the final visual representation.

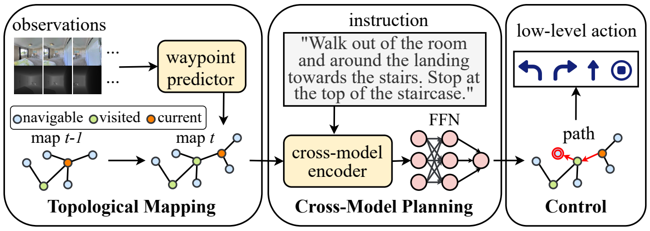


Fig. 3. Main architecture of the model. The navigation system is divided into three modules: topological mapping, cross-modal planning, and control. Our proposed knowledge-and-history aware visual representation method primarily enhances the topological mapping module, aiding in subsequent action prediction.

the relevant knowledge for the RGB images, and records the historical information based on the topological map from the previous step. Subsequently, the knowledge and history features are fed into the filtering module to capture critical information closely related to the instruction. Additionally, the purified features interact through the attention mechanism and exchange information with the instruction. Finally, the aggregation module merges the image features, knowledge features and history features to construct a visual representation, which is stored in the updated topological map node.

The overall model adheres to the architecture of ETP [10], as depicted in Fig. 3. The model consists of three modules: topological mapping, cross-modal planning, and control. The topological mapping module generates a dynamic topological map that evolves over time, containing three types of nodes: the current node, visited nodes, and navigable nodes. At each navigation step, the agent processes new observations from the current node, employs a waypoint predictor [8] to produce candidate navigable nodes, and updates the topological map accordingly. The cross-modal planning module processes the encoded topological map and embedded instruction text through a cross-modal encoder, forecasts a long-term goal using a feed-forward network, and develops a path plan based on distance information stored in the topological map. The control module uses low-level actions to direct the agent in executing the plan. Our proposed method primarily enhances the topological mapping module, and the subsequent section describes the specific operations of each module within the knowledge-and-history aware visual representation.

B. Knowledge-and-History Aware Visual Representation

Extraction. At each step t, the agent receives the observed RGB and depth images. We follow the conventional method [10], [29] to separately encode the two types of images using distinct visual encoders, producing RGB features f^r and depth features f^d. For brevity, we omit the time step subscripts t. To extract the relevant knowledge features f^k, we calculate the cosine similarity scores between each knowledge item kn and the image o^{rgb}. Each knowledge item is retrieved from the knowledge base KB. The top-m highest-scoring knowledge items are selected and concatenated as follows:

$$\left\{f_i^k\right\}_{i=1}^{m} = top_m\left\{cos\left(o^{rgb}, kn\right), kn \in KB\right\} \tag{1}$$

$$f^k = concat\left(\left\{f_i^k\right\}_{i=1}^{m}\right) \tag{2}$$

where *cos* denotes the cosine similarity calculation.

To derive history features f^h, it is essential to retain the topological map details from the previous step within the current node. Concretely, the node image features, the time-step encoding of node access, and the relative position encoding of the located node with respect to all nodes in the map, these features from the previous step are combined to generate the history features:

$$f^h = map_{t-1}^{img} + map_{t-1}^{step} + map_{t-1}^{pos} \tag{3}$$

Filtering. The filtering module purifies the knowledge features and history features by removing redundant information that is not pertinent to navigation, thereby extracting features highly relevant to the instruction. For example, the correlation matrix M_k between the knowledge features f^k and the navigation instruction features f^i is calculated as follows:

$$M_k = f^k W_k \left(f^i W_i\right)^T \tag{4}$$

where W_k and W_i represent learnable parameters.

Subsequently, relevance scores are assigned to the knowledge features based on the relevance matrix, which serve as weights indicating the strength of the association between the knowledge features and the navigation instruction. This process generates the refined and weighted knowledge features $\widetilde{f^k}$:

$$\widetilde{f^k} = softmax\left(\frac{M}{\sqrt{d}}\right) f^k \tag{5}$$

where d is the dimension of the feature. Similarly, the filtered history features $\widetilde{f^k}$ are obtained in the same way.

Interaction. Having obtained the refined features, the interaction module facilitates a dynamic exchange of perceptual information among knowledge, history, and instruction, ensuring a comprehensive integration of insights. To be specific, the knowledge-history interaction is performed using a multilayer cross-modal transformer to obtain the knowledge-history features f^{kh}:

$$f^{kh} = CrossModelEncoder\left(\widetilde{f^k}, \widetilde{f^h}\right) \tag{6}$$

After that, similar to the operations in the filtering module, the knowledge-history features interact with the instruction information to yield multi-perceptual fusion features f^{fusion}:

$$f^{fusion} = softmax\left(\frac{f^{kh} W_{kh} \left(I_{cls} W_{cls}\right)^T}{\sqrt{d}}\right) f^{kh} \tag{7}$$

where W_{kh} and W_{cls} represent learnable parameters, and I_{cls} denotes the encoded [CLS] token, signifying the semantic representation of the entire instruction.

Aggregation. In the final stage, the interaction features f^{fusion}, RGB features f^r, and depth features f^d are individually fed into three different feedforward neural networks. Then we aggregate the outputs to construct the visual representation f^{img}. This representation is included in the panoramic encoding sent to the subsequent cross-modal planning phase, where it plays a crucial role in predicting the next target node:

$$f^{img} = FFN_{fusion}\left(f^{fusion}\right) + FFN_{rgb}\left(f^r\right) + FFN_{depth}\left(f^d\right) \tag{8}$$

IV. EXPERIMENTS

A. Experiment Setup

Implementation Details. We conduct experiments on the Habitat Simulator [39]. To encode RGB and depth images, we employ ViT-B/16 [40], pre-trained on CLIP [41], and ResNet-50 [42], pre-trained on point-goal navigation [43], respectively. The knowledge base is constructed from Visual Genome [44] following KERM [35]. The number of layers for the interaction module encoder, text encoder, and cross-modal encoder are set as 2, 9, and 4, respectively. In the extraction module, $m = 5$ knowledge items are selected for each image. During training, we utilize the pre-trained model from ETP [10] and fine-tune the proposed model for 25k iterations on a single NVIDIA RTX3090 GPU with the batch size of 8. The AdamW optimizer is adopted with a learning rate of 1e-5.

Datasets. The VLN-CE dataset [5] is collected from the discrete Matterport3D environment [45] through Habitat Simulator [39], providing fine-grained language guidance. The dataset contains 16,844 path-instruction pairs across 90 scenes, segmented into train, seen validation, unseen validation, and test splits. Specifically, it includes 61 environments for training, 11 environments for unseen validation, and 18 environments held for testing. Scenes in the seen validation split have appeared in the train split. Since VLN-CE emphasizes navigational ability in novel environments, the performance on the unseen validation set is more important than the seen validation set.

Evaluation Metrics. The following metrics are reported to evaluate the navigation performance, where SR is used as the primary metric for comparison: (1) Trajectory Length (TL) denotes the average path length. (2) Navigation Error (NE) measures the average distance between the agent's final position and the target. (3) Success Rate (SR) represents the proportion of the agent successfully stops within 3 meters to the target. (4) Oracle Success Rate (OSR) indicates the proportion of the closest point in the predicted trajectory to the target within 3 meters. (5) Success rate weighted by Path Length (SPL) integrates SR and TL, which measures both the accuracy and the efficiency of navigation.

B. Performance on VLN-CE

Table I presents a comparison between VLN-KHVR and state-of-the-art methods on the VLN-CE dataset. The experimental results demonstrate that our proposed model outperforms the existing models in terms of NE, SR, and SPL metrics in unseen environments, showcasing the robust generalization ability of our approach. Specifically, compared

TABLE I

Comparison with State-of-the-Art Methods on the VLN-CE Dataset.

model	Validation Seen					Validation Unseen					Test Unseen				
	TL	NE↓	OSR↑	SR↑	SPL↑	TL	NE↓	OSR↑	SR↑	SPL↑	TL	NE↓	OSR↑	SR↑	SPL↑
Seq2Seq [5]	9.37	7.02	46.0	33.0	31.0	9.32	7.77	37.0	25.0	22.0	8.85	7.91	36	28	25
CMTP [27]	-	7.10	45.4	36.1	31.2	-	7.90	38.0	26.4	22.7	-	-	-	-	-
HPN [6]	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	32	30
LAW [7]	9.34	6.35	49.0	40.0	37.0	8.89	6.83	44.0	35.0	31.0	-	-	-	-	-
CM^2 [12]	12.05	6.10	50.7	42.9	34.8	11.54	7.02	41.5	34.3	27.6	13.90	7.70	39	31	24
CWP-CMA [8]	11.47	5.20	61.0	51.0	45.0	10.90	6.20	52.0	41.0	36.0	11.85	6.30	49	38	33
CWP-VLNBERT [8]	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	42	36
Sim2Sim [9]	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	44	37
WS-MGMAP [13]	10.12	5.65	51.7	46.9	43.4	10.00	6.28	47.6	38.9	34.3	12.30	7.11	45	35	28
Ego^2-Map [38]	-	-	-	-	-	-	4.94	-	51.8	46.1	13.05	5.54	56	47	41
GridMM [29]	12.69	4.21	69.0	59.0	51.0	13.36	5.11	61.0	49.0	41.0	13.31	5.64	56	46	39
ETP [10]	11.78	3.95	**71.9**	**66.2**	**59.4**	11.99	4.71	**64.7**	57.2	49.2	12.87	5.12	**63**	55	**48**
VLN-KHVR(Ours)	11.43	**3.82**	71.3	65.2	58.3	12.21	**4.52**	64.2	**58.8**	**50.2**	13.22	**5.05**	61	**56**	**48**

TABLE II

Ablation of the Key Visual Representation Components on VLN-CE Unseen Validation Split.

Knowledge	History	TL	NE↓	OSR↑	SR↑	SPL↑
		11.99	4.71	64.7	57.2	49.2
✓		14.04	4.69	66.0	58.1	46.7
	✓	14.18	4.69	**66.8**	58.4	46.6
✓	✓	12.21	**4.52**	64.2	**58.8**	**50.2**

TABLE III

Ablation of the Key Model Components on VLN-CE Unseen Validation Split.

Fil.	KH-Int.	In-Int.	TL	NE↓	OSR↑	SR↑	SPL↑
			11.99	4.71	64.7	57.2	49.2
	✓	✓	12.33	4.55	64.2	58.2	48.9
✓	✓		14.80	4.60	**66.3**	58.5	47.2
✓		✓	14.52	4.55	65.5	58.6	47.4
✓	✓	✓	12.21	**4.52**	64.2	**58.8**	**50.2**

TABLE IV

Ablation of the Feature Fusion Strategy on VLN-CE Unseen Validation Split.

Model	RGB	Depth	NE↓	OSR↑	SR↑	SPL↑
baseline			4.71	64.7	57.2	49.2
M#1	✓	✓	4.78	63.9	57.2	48.5
M#2	✓		4.70	63.8	57.5	49.2
M#3		✓	4.68	**64.8**	58.5	49.3
VLN-KHVR			**4.52**	64.2	**58.8**	**50.2**

to the strong baseline model ETP [10] on the unseen validation split, VLN-KHVR achieves significant improvements of 3.2% in SR, 2.4% in SPL, and reduces NE by 4.0%. In addition, we submit the model to the test set used for VLN-CE leaderboard, our method leads in terms of SR, SPL, and NE among prior work. These results indicate that VLN-KHVR can accomplish the task with greater accuracy and reliability, highlighting the effectiveness of the proposed knowledge-and-history aware visual representation.

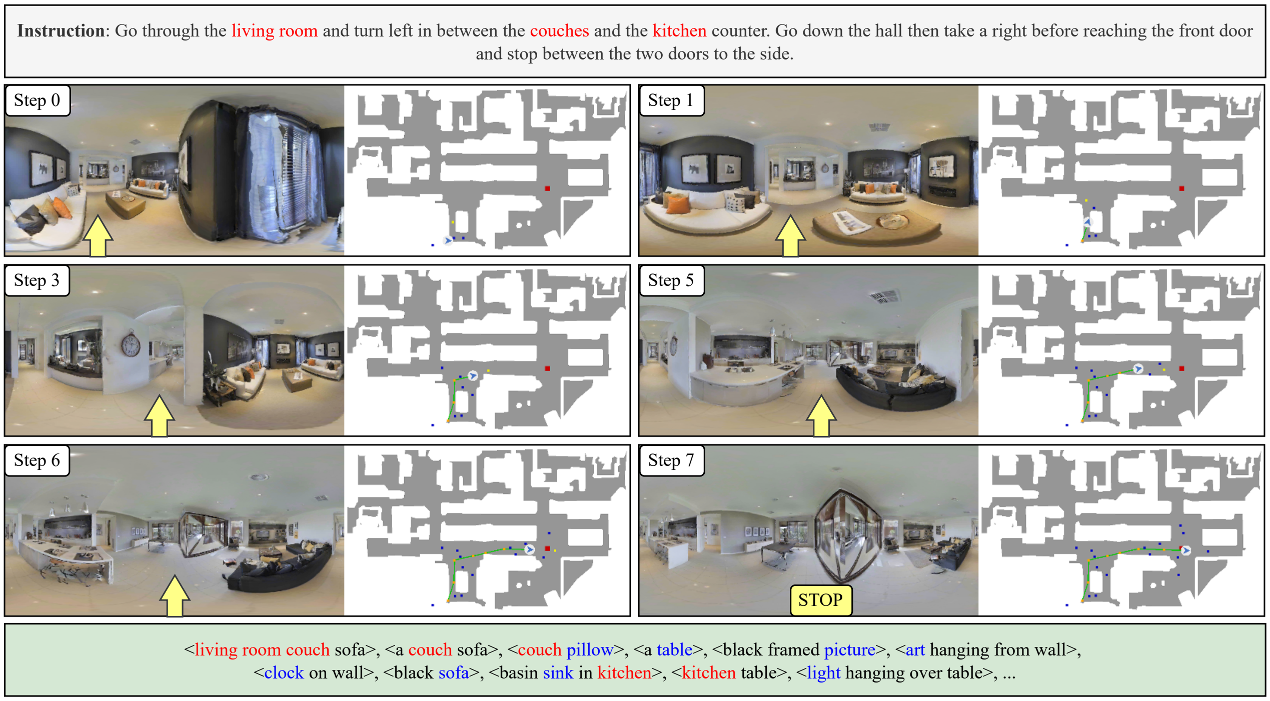
C. Ablation Analysis

In ablation experiments, we empirically validate the significance of the individual components in our proposed VLN-KHVR, and explore various designs for combining these features. The results are summarized in Table II, Table III and Table IV.

Key Components of Visual Representation. As shown in Table II, we evaluate the impact of the key components of visual representation on the VLN-CE unseen validation split. Firstly, the results in rows 2 and 3 are superior to row 1 in all metrics except for SPL, which demonstrates that incorporating knowledge or history into visual representation is beneficial for navigation. Secondly, the overall results in row 4 outperform those in rows 2 and 3, especially with a notable enhancement in the SPL metric. This comparative advantage suggests that integrating both knowledge and history into visual representation is better than adding only one element, thus validating the effectiveness of the knowledge-and-history aware visual representation.

Module Composition of Feature Processing. To investigate the effectiveness of the key modules for processing visual representations, we execute experiments to ablate different modules. Table III illustrates that employing all modules yields the best results on the unseen validation set. Here, ”Fil.”, “KH-Int.” and “In-Int.” respectively denote filtering, knowledge-history interaction, and instruction interaction. The success rates in rows 2 to 4 are higher than the baseline, but their navigation performance is inferior to that of row 5, especially in the SPL metric. This implies that a single module provides limited enhancements in navigation performance. The simultaneous utilization of the filtering and interaction modules demonstrates optimal performance, indicating that the synergy between these modules provides the complete model with a substantial advantage in the efficiency and accuracy of action prediction.

Fusion Strategy of Different Features. Furthermore, we explore the effects of various feature fusion strategies on navigation performance. We design three model variants. Variant M#1 integrates RGB and depth features within the



Instruction: Go through the living room and turn left in between the couches and the kitchen counter. Go down the hall then take a right before reaching the front door and stop between the two doors to the side.

<living room couch sofa>, <a couch sofa>, <couch pillow>, <a table>, <black framed picture>, <art hanging from wall>, <clock on wall>, <black sofa>, <basin sink in kitchen>, <kitchen table>, <light hanging over table>, ...

Fig. 4. Qualitative navigation results in the VLN-CE dataset.

filtering and interaction module. Variant M#2 employs RGB features for the filtering and interaction processes, with depth features exclusively contributing to the aggregation module. Variant M#3, on the contrary, involves the depth features in the processing phase and utilizes RGB features for aggregation. Experimental results are shown in Table IV. The performance of M#1 and M#2 is similar to the baseline. M#1 exhibits the poorest performance, which suggests that the integration of RGB and depth features in the filtering and interaction module negatively affects navigation performance. Notably, M#3 performs better than M#2 and slightly outperforms the baseline model in all the metrics. This indicates that leveraging depth features during processing and RGB features solely for aggregation proves more advantageous for visual representation. VLN-KHVR achieves the best performance on the unseen validation set while maintaining image information integrity without processing RGB and depth features.

D. Qualitative Results

We take an example to visualize the proposed method. The qualitative result is presented in Fig. 4. In this figure, the upper part represents the given instruction, while the lower part displays the relevant knowledge retrieved from the image. Particularly, elements of the knowledge that are closely related to both the instruction and the current image are highlighted in red, for instance, *living room*, *couch*, and *kitchen*. Moreover, the retrieved knowledge contains objects that are not explicitly mentioned in the instruction but visible in the image, denoted in blue, such as *pillow*, *table*, *picture* and so on. It can be seen VLN-KHVR successfully navigates in the scene following the given instruction by retrieving knowledge and considering history. This demonstrates the model's ability to navigate effectively by leveraging knowledge and historical information alongside visual cues.

V. Conclusion

In this work, we propose a novel knowledge-and-history aware visual representation (VLN-KHVR) for continuous vision-and-language navigation. Through the process of extraction, filtering, interaction, and aggregation, VLN-KHVR acquires knowledge and historical information that is highly relevant and effective for navigation. By incorporating this enriched information into the visual representation, VLN-KHVR significantly enhances the navigation and generalization capabilities of the agent. The experimental results demonstrate the superiority of the proposed method on VLN-CE, which shows taking advantage of commonsense knowledge and navigation history is a promising direction for enhancing navigation performance. For future work, we aim to exploit more diverse knowledge bases, extend our method to different baseline models, and explore computationally efficient solutions to optimize resource utilization.

Acknowledgment

This work was partly supported by the National Natural Science Foundation of China under Grants No. 62206199 and No. U2141234, National Key Laboratory of Marine Engine Science and Technology under Grants No. LAB-2024-04-WD, Young Elite Scientist Sponsorship Program under Grant No. YESS20220409, Swedish Foundation for Strategic Research (SSF) through project APR20-0023 and Swedish Innovation Agency (Vinnova) through the SweWIN center (2023-00572).

References

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

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

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

[4] S. Chen, P.-L. Guhur, M. Tapaswi, C. Schmid, and I. Laptev, "Think global, act local: Dual-scale graph transformer for vision-and-language navigation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 537–16 547.

[5] J. Krantz, E. Wijmans, A. Majumdar, D. Batra, and S. Lee, "Beyond the nav-graph: Vision-and-language navigation in continuous environments," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVIII 16*. Springer, 2020, pp. 104–120.

[6] J. Krantz, A. Gokaslan, D. Batra, S. Lee, and O. Maksymets, "Waypoint models for instruction-guided navigation in continuous environments," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 15 162–15 171.

[7] S. Raychaudhuri, S. Wani, S. Patel, U. Jain, and A. X. Chang, "Language-aligned waypoint (law) supervision for vision-and-language navigation in continuous environments," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, 2021, pp. 4018–4028.

[8] Y. Hong, Z. Wang, Q. Wu, and S. Gould, "Bridging the gap between learning in discrete and continuous environments for vision-and-language navigation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 15 439–15 449.

[9] J. Krantz and S. Lee, "Sim-2-sim transfer for vision-and-language navigation in continuous environments," in *European Conference on Computer Vision*. Springer, 2022, pp. 588–603.

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

[11] D. An, Y. Qi, Y. Li, Y. Huang, L. Wang, T. Tan, and J. Shao, "Bevbert: Multimodal map pre-training for language-guided navigation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 2737–2748.

[12] G. Georgakis, K. Schmeckpeper, K. Wanchoo, S. Dan, E. Miltsakaki, D. Roth, and K. Daniilidis, "Cross-modal map learning for vision and language navigation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 15 460–15 470.

[13] P. Chen, D. Ji, K. Lin, R. Zeng, T. Li, M. Tan, and C. Gan, "Weakly-supervised multi-granularity map learning for vision-and-language navigation," in *Advances in Neural Information Processing Systems*, vol. 35, 2022, pp. 38 149–38 161.

[14] V. Jain, G. Magalhaes, A. Ku, A. Vaswani, E. Ie, and J. Baldridge, "Stay on the path: Instruction fidelity in vision-and-language navigation," in *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, 2019, pp. 1862–1872.

[15] A. Ku, P. Anderson, R. Patel, E. Ie, and J. Baldridge, "Room-across-room: Multilingual vision-and-language navigation with dense spatiotemporal grounding," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, 2020, pp. 4392–4412.

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

[17] F. Zhu, X. Liang, Y. Zhu, Q. Yu, X. Chang, and X. Liang, "Soon: Scenario oriented object navigation with graph-based exploration," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12 689–12 699.

[18] X. Wang, Q. Huang, A. Celikyilmaz, J. Gao, D. Shen, Y.-F. Wang, W. Y. Wang, and L. Zhang, "Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 6629–6638.

[19] W. Zhang, C. Ma, Q. Wu, and X. Yang, "Language-guided navigation via cross-modal grounding and alternate adversarial learning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 9, pp. 3469–3481, 2020.

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

[21] Hao, L. Yu, and M. Bansal, "Learning to navigate unseen environments: Back translation with environmental dropout," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, 2019, pp. 2610–2621.

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

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

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

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

[26] Y. Qiao, Y. Qi, Y. Hong, Z. Yu, P. Wang, and Q. Wu, "Hop+: History-enhanced and order-aware pre-training for vision-and-language navigation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 7, pp. 8524–8537, 2023.

[27] K. Chen, J. K. Chen, J. Chuang, M. Vázquez, and S. Savarese, "Topological planning with transformers for vision-and-language navigation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11 276–11 286.

[28] M. Z. Irshad, N. C. Mithun, Z. Seymour, H.-P. Chiu, S. Samarasekera, and R. Kumar, "Semantically-aware spatio-temporal reasoning agent for vision-and-language navigation in continuous environments," in *2022 26th International Conference on Pattern Recognition (ICPR)*. IEEE, 2022, pp. 4065–4071.

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

[30] M. Qi, Y. Wang, J. Qin, and A. Li, "Ke-gan: Knowledge embedded generative adversarial networks for semi-supervised scene parsing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5237–5246.

[31] A. K. Singh, A. Mishra, S. Shekhar, and A. Chakraborty, "From strings to things: Knowledge-enabled vqa model that can read and reason," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 4602–4612.

[32] A. Zareian, S. Karaman, and S.-F. Chang, "Bridging knowledge graphs to generate scene graphs," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIII 16*. Springer, 2020, pp. 606–623.

[33] S. Shen, C. Li, X. Hu, Y. Xie, J. Yang, P. Zhang, Z. Gan, L. Wang, L. Yuan, C. Liu, *et al.*, "K-lite: Learning transferable visual models with external knowledge," in *Advances in Neural Information Processing Systems*, vol. 35, 2022, pp. 15 558–15 573.

[34] C. Gao, J. Chen, S. Liu, L. Wang, Q. Zhang, and Q. Wu, "Room-and-object aware knowledge reasoning for remote embodied referring expression," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3064–3073.

[35] X. Li, Z. Wang, J. Yang, Y. Wang, and S. Jiang, "Kerm: Knowledge enhanced reasoning for vision-and-language navigation," in *Proceed-*

ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 2583–2592.

[36] B. Mohammadi, Y. Hong, Y. Qi, Q. Wu, S. Pan, and J. Q. Shi, “Augmented commonsense knowledge for remote object grounding,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 5, 2024, pp. 4269–4277.

[37] R. Speer, J. Chin, and C. Havasi, “Conceptnet 5.5: An open multilingual graph of general knowledge,” in *Proceedings of the AAAI conference on artificial intelligence*, vol. 31, no. 1, 2017.

[38] Y. Hong, Y. Zhou, R. Zhang, F. Dernoncourt, T. Bui, S. Gould, and H. Tan, “Learning navigational visual representations with semantic map supervision,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 3055–3067.

[39] M. Savva, A. Kadian, O. Maksymets, Y. Zhao, E. Wijmans, B. Jain, J. Straub, J. Liu, V. Koltun, J. Malik, *et al.*, “Habitat: A platform for embodied ai research,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 9339–9347.

[40] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, “An image is worth 16x16 words: Transformers for image recognition at scale,” in *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*, 2021.

[41] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, “Learning transferable visual models from natural language supervision,” in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[42] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[43] E. Wijmans, A. Kadian, A. Morcos, S. Lee, I. Essa, D. Parikh, M. Savva, and D. Batra, “DD-PPO: learning near-perfect pointgoal navigators from 2.5 billion frames,” in *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*, 2020.

[44] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, *et al.*, “Visual genome: Connecting language and vision using crowdsourced dense image annotations,” *International journal of computer vision*, vol. 123, pp. 32–73, 2017.

[45] A. X. Chang, A. Dai, T. A. Funkhouser, M. Halber, M. Nießner, M. Savva, S. Song, A. Zeng, and Y. Zhang, “Matterport3d: Learning from RGB-D data in indoor environments,” in *2017 International Conference on 3D Vision, 3DV 2017, Qingdao, China, October 10-12, 2017*. IEEE Computer Society, 2017, pp. 667–676.