CLOI-NAV: Open-World Vision-and-Language Navigation via Complex, Long-horizon Ordered Instructions

Minho Lee[1], Jaeil Park[2], Jinyong Jeong[3], and Younggun Cho[1*]

***Abstract*— Remarkable advances in Vision-Language Models (VLMs) and Large Language Models (LLMs) have accelerated progress in the field of intelligent robotics, enabling embodied agents to perceive, reason, and act in a human-like manner. One of the mainstream challenges in embodied AI is Vision-and-Language Navigation (VLN), where an agent is required to follow natural language instructions to navigate through previously unseen environments using visual observations. Despite recent progress, existing VLN approaches often struggle to handle long-horizon and ordered instructions, which are prevalent in realistic navigation scenarios. Such instructions involve multiple sequential substeps where later actions depend on earlier completions, requiring contextual order understanding and stepwise execution. In this work, we present *CLOI-NAV*, a framework that performs sequential reasoning to follow navigation instructions in unseen environments while preserving the intended order. We evaluate *CLOI-NAV* using our new instruction datasets featuring sequential dependencies in photorealistic environments. Through extensive experiments, we demonstrate that our method enables more accurate instruction following while maintaining path efficiency, with success rate improving from 26.9 to 88.5 and SPL from 29.3 to 76.4. Our code and video demos are available at** `https://sparolab.github.io/research/cloi_nav/`.

I. Introduction

The Vision-and-Language Navigation (VLN) task has risen to prominence in Embodied AI research, as it allows robots to follow natural language instructions and interact with humans in a more intuitive and practical way. In this task, an embodied agent needs to execute textual instructions and navigate through complex visual environments. To achieve this, the agent must understand the meaning of each instruction in context, ground it in visual observations, and make appropriate navigation decisions.

Early VLN methods used LSTM architectures [1], while later approaches adopted Transformers for improved representation learning and context-aware reasoning over temporal history [2]. More recently, large-scale vision-language models specialized for VLN [3–5] build on pre-trained multimodal representations while incorporating navigation-specific objectives and datasets. However, these approaches heavily depend on domain-specific datasets with large scale and high-quality annotations. Consequently, they struggle to generalize to open-world scenarios involving novel environments and unseen instructions.

To address these limitations, recent works have explored integrating Large Language Models (LLMs), which show strong potential in zero-shot generalization, comprehensive scene understanding, and high-level decision-making. Several representative works demonstrate this potential by leveraging LLMs with different strategies: NavGPT [6] converts visual observations into textual descriptions and processes them with LLMs to decide the next waypoint; InstructNav [7] extracts actions and visual landmarks from instructions and uses them to guide navigation decisions while grounding them with updated visual information during navigation; Open-Nav [8] introduces specialized roles for LLMs to perform instruction comprehension, progress estimation, and navigation planning through scene object and spatial description integration; and SnapMem [9] provides LLMs with textual instructions and multi-view images, allowing rich visual understanding for enhanced navigation reasoning.

However, existing studies have confined their evaluations to simple object-guided instructions *(e.g., Find the wooden chair by the window)* or narrowly scoped instruction-following tasks *(e.g., Walk to the living room and go behind the couch)*. Such simplified settings overlook the demands of open-world navigation (OWN), including the ability to (i) handle human-issued instructions that are commonly ambiguous and compositional, and (ii) navigate environments with visually diverse objects across varied scene contexts.

To this end, we propose CLOI-NAV, a framework equipped with LLM assistants, for complex and sequential instruction-following tasks in open-world environments. The contributions of our work include:

- **D**esign an informative snapshot module that selectively filters meaningful visual observations and provides them as LLM input, enabling efficient and reliable execution of contextual instruction-following tasks.
- **D**evelop LLM-based navigation assistants that can process complex, sequential instructions, providing robust language understanding and stepwise decision making, and improved adaptability across diverse navigation scenarios.
- **I**ntegrate an exploration module within the embodied navigation system, allowing the agent to effectively follow instructions and navigate in unseen environments.

II. Related Works

A. Vision-and-Language Navigation

Since the introduction of Vision-and-Language Navigation (VLN) by Anderson et al. [10], VLN has become a representative benchmark for embodied AI, driving research on multimodal grounding, instruction following, and generalization in navigation tasks. Early works primarily utilized language-

*Corresponding authors.

[1]M. Lee and [1]*Y. Cho are with the Dept. of Electrical and Computer Eng., [2]J. Park is with the Dept. of Smart Mobility Eng., Inha University, South Korea (e-mail: mino@inha.edu, yg.cho@inha.ac.kr, jaejaee5220@inha.edu)

[3]J. Jeong, CTO, is with the Robotics for Industrial Innovation (Riibotics), South Korea. (e-mail: jinyong.jeong@riibotics.io)

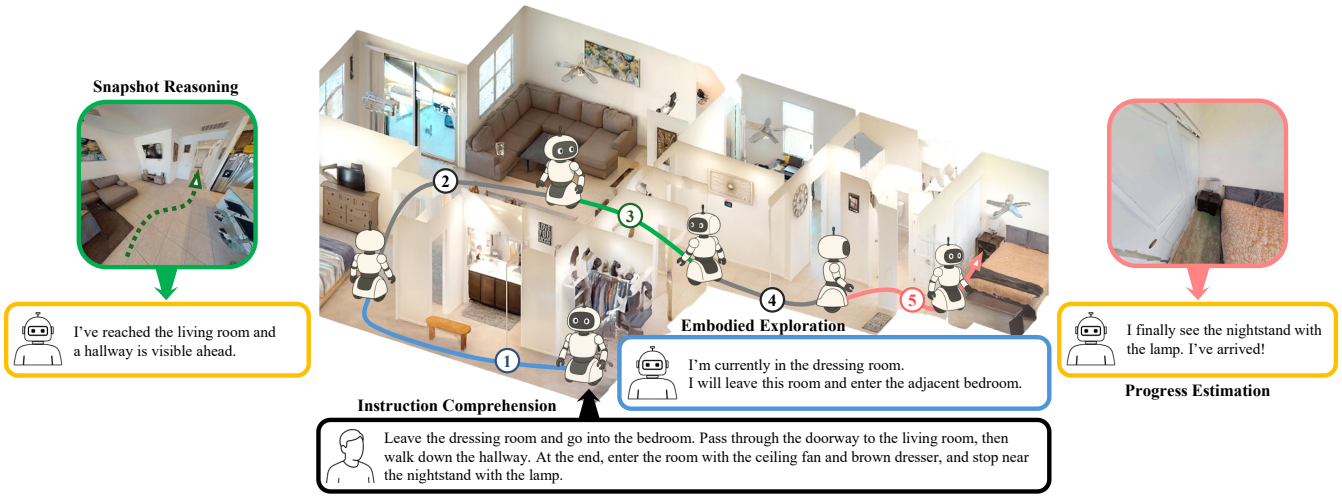

Fig. 1. Given textual instructions, CLOI-NAV refines and decomposes the task into structured substeps. Once these substeps are established, the embodied agent observes the surroundings and captures snapshots containing both semantically rich scene content and navigable frontiers. The agent then performs snapshot reasoning to infer the next action and explores the environment accordingly, while monitoring progress through current substep completion.

grounded spatial representations, including 3D voxel maps [11–13], topological graphs [14, 15], neural fields [16, 17], and 3D scene graphs [18, 19]. These representations were commonly integrated with vision-language models (VLMs) for multimodal grounding, but they still relied on fixed structural priors that hindered scalability and adaptability, making it difficult to generalize to novel contexts in open-world scenarios. To handle this challenge, our work leverages a set of informative scene snapshots that visually capture the spatial and semantic context around the embodied agent. These snapshots provide cues for reasoning about the place the agent is currently located in and predicting where the navigable directions may lead, and even interpreting spatial relationships among surrounding objects.

B. LLM-guided Embodied Exploration

Recent advances in large-scale models, including LLMs and multimodal foundation models, have demonstrated powerful capabilities in natural language processing, commonsense reasoning, and zero-shot visual perception. These capabilities are closely related to embodied navigation. Existing approaches have employed multimodal large models (e.g., CLIP [20], BLIP-2 [21], and GLIP [22]) as visual feature encoders, aligning visual features with language instructions to retrieve and navigate toward target locations. However, these approaches face challenges in grounding visual observations to lengthy and descriptive instructions and often fail to capture the contextual cues underlying natural language commands. Motivated by these challenges, some studies [9, 23, 24] have explored LLM-based high-level decision making for robotic navigation. These works highlight LLM's ability to comprehend language, reason over context, and enable adaptive navigation strategies.

III. Method

Inspired by the way humans follow verbal directions to navigate unfamiliar environments [25], we propose an approach, illustrated in Fig. 1, that leverages LLMs to comprehend and execute instructions through visual grounding and contextual reasoning. The framework consists of three interconnected modules: (i) instruction comprehension, including refinement and decomposition; (ii) informative snapshot-based contextual reasoning, supporting progress estimation and navigation strategy; and (iii) embodied exploration within the environment. This design enables the agent to follow the instructions in unseen environments with greater robustness and adaptability.

A. Instruction Comprehension

While concise goal-driven commands (e.g., *Go to the front door*) may suffice in simple or familiar environments, instructions in practical settings—particularly those encountered in complex and unseen environments—are typically long, contextually dependent, and sequentially structured (e.g., *Walk toward the kitchen counter, turn into the hallway, and keep walking until you reach the laundry room*). This demands the agent's ability to comprehend and decompose them into finer-grained subgoals. Our approach is designed to first refine the instruction for clarity and then decompose it into an ordered sequence of substeps that reflect the intended navigation path. This enables the agent to reason over each subgoal in context and track its progress accordingly.

Instruction Refinement. Directly decomposing instructions into substeps through LLMs can lead to ambiguity for two reasons: first, original instructions often contain implicit spatial connections or visual cues that are not explicitly stated; second, the decomposition process tends to further simplify and omit essential contextual information required for navigation decisions. To mitigate this, we refine the original instruction into a form that is more suitable for decomposition, ensuring that each substep preserves the necessary contextual information and reduces potential ambiguity.

Substep Decomposition. After refinement, the instruction is decomposed into ordered substeps that follow the intended navigation sequence. Each substep incorporates spatial transitions (e.g., *Enter the hallway adjacent to the kitchen*) and visual landmarks provided in the refined instruction (e.g., *Walk past the table with a vase*). This allows the agent to follow the instruction step-by-step effectively, as each substep provides explicit guidance on spatial movements and visual references.

B. Snapshot Reasoning

We introduce a novel approach that leverages informative snapshots to support the agent's reasoning process, building on the idea that images contain rich visual context and recent advances in multimodal reasoning capabilities of LLMs. The snapshots capture both object-level details and room-level contextual cues, providing a more holistic scene representation that guides navigation decisions through fine-grained evidence and broader spatial awareness.

Snapshot Selection. A set of N RGB-D image observations $O = \{I_1, I_2, \ldots, I_N\}$ is captured from the environment, providing multiple views from different viewpoints around the agent rather than relying on a single perspective. Among these, informative snapshots S are identified as:

$$S = \{I \in O \mid I \in S_{Obj} \vee I \in S_{Frt}\}. \quad (1)$$

To determine S, we apply simple criteria for filtering the acquired images. First, we retain the images when they contain a sufficient number of detected objects, forming the set S_{Obj}. This ensures that the preserved views provide rich semantic cues for reasoning. Next, we keep images that correspond to directions recognized as unexplored and navigable regions through frontier-based exploration algorithm (Section III-C), forming the set S_{Frt}. This guides the agent toward directions that expand its knowledge of the environment while remaining feasible for navigation. Only these informative snapshots are then used for reasoning, improving efficiency while ensuring comprehensive coverage that spans visual cues and spatial context.

Progress Estimation. To precisely monitor and evaluate the agent's navigation progress, we design specific prompts to estimate the status of the current substep:

> *"You are a navigation assistant. Your task is to estimate the agent's progress through a sequence of spatial substeps. You will be given the current substep, optionally the next substep, informative snapshots of the surroundings, and the reasoning history from previous steps. Always preserve spatial details, do not infer completion from the next substep, and if uncertain, choose 'In Progress' as the status. Answer: [In Progress / Completed]"*

This estimation process, as shown in Fig. 2, combines the current substep with the subsequent one, informative snapshots, and the reasoning history accumulated from previous steps. In particular, the reasoning history serves not only to indicate whether the agent has been consistently following spatially connected regions but also to enhance robustness in completion checks, allowing reassessment of prior substeps if needed. This design is inspired by the way humans strategically determine their next move in unfamiliar environments, grounding local decisions in immediate cues while maintaining awareness of past experiences.

C. Embodied Exploration

Frontier Exploration. Recent works in VLN adapt the frontier-based exploration algorithm proposed by Yamauchi et al. [26], which guides agents to visit the boundaries between

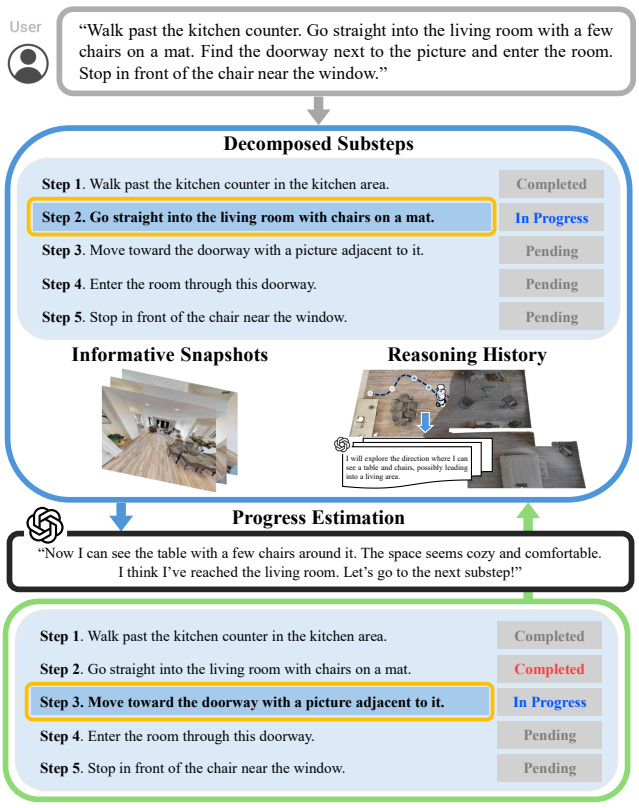

Fig. 2. Illustration of progress estimation where each substep is evaluated by LLM using contextual cues, informative snapshots, and reasoning history, resulting in Completed, In Progress, or Pending labels.

explored and unexplored regions on a map that is incrementally constructed during navigation. Our frontier-based exploration algorithm is based on the framework in Explore-EQA [27]. Specifically, the method utilizes a depth observation with its corresponding observation pose and defines frontiers as clusters of pixels in the unexplored region. These are regions where the truncated signed distance field (TSDF) volumes are either uninitialized (i.e., no observation) or exceed the truncation margin, indicating areas not yet observed or far from any known surface.

Navigation Strategy. The embodied agent explores using a snapshot selected through reasoning. When the agent chooses a snapshot associated with a frontier, it directly navigates to the location of the frontier from its current pose. If the chosen snapshot is not related to any frontier, the agent navigates toward the detected objects in the snapshot and positions itself at an appropriate location for exploration. In our experiments, we employ the collision-free planner provided by Habitat-sim [28] to compute the shortest path between two locations in the explored region and move the agent iteratively at each time step.

IV. Experiments

A. Experimental Setup

Simulated Environments. Our experiments are conducted in the Habitat simulator using the Habitat-Matterport 3D Research Dataset (HM3D) [29], which provides photorealistic 3D scans of diverse real-world indoor spaces. This allows us to evaluate our approach in a wide range of realistic indoor layouts and scene configurations.

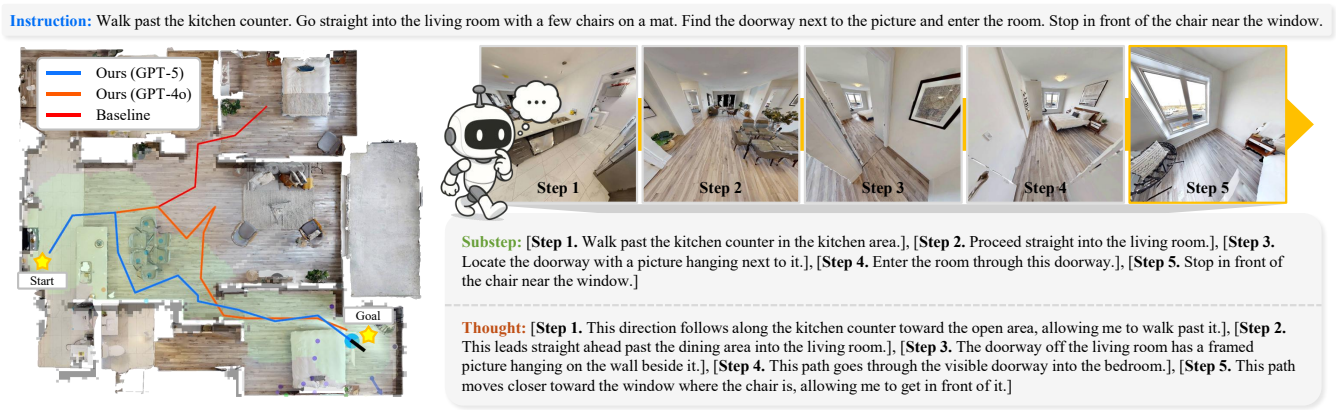

Fig. 3. Comparison of the exploration trajectories of the proposed and baseline methods in an HM3D environment with a given instruction. The right side of the figure shows our method *CLOI-NAV* with GPT-5, a state-of-the-art VLM, along with egocentric views during navigation (top), decomposed substeps from the instruction (middle), and the LLM's step-by-step thoughts (bottom).

TABLE I

RESULTS ON HM3D WITH CUSTOM INSTRUCTIONS

Method	TL	NE (↓)	OSR (↑)	SR (↑)	SPL (↑)
SnapMem[9]	14.3	1.28	37.0	26.9	29.3
Ours	13.4	1.26	**93.3**	**88.5**	**76.4**

Instruction Generation. While prior work has focused on relatively simple questions, we aim to demonstrate our framework by generating complex multi-step instructions that better reflect realistic scenarios. To generate such instructions, we sampled connected waypoints, extracted corresponding poses and egocentric views, and provided them to LLMs for instruction generation. We then ensured quality by manually filtering trivial or unnatural cases.

B. Evaluation

Evaluation Metrics. We evaluate the agent's navigation performance using standard VLN metrics, including success rate (SR), oracle success rate (OSR), success weighted by path length (SPL), trajectory length (TL), and navigation error (NE). A navigation task is deemed successful if the agent's final position is within 3 meters of the goal and the instruction is followed in the correct order.

Results and Analysis. Table I compares the navigation performance of our proposed method with the baseline framework, SnapMem [9]. For fairness, both methods were evaluated under the same exploration module to isolate and highlight the contribution of the LLM-based reasoning modules. Unlike SnapMem, which uses LLMs solely for image-based reasoning to predict target points for navigation, our method also integrates comprehensive instruction understanding. As a result, our method CLOI-NAV significantly outperforms the baseline in complex and sequential instructions, achieving over 3x higher SR, more than 2.5x higher SPL, with TL and NE slightly reduced.

C. Ablation Study

We conducted ablation studies using 41 episodes randomly sampled from various scenes in the evaluation set.

Effect of Comprehension and Reasoning History. Table II shows significant performance drops without refinement or progress estimation, as LLMs often interpret instructions globally rather than sequentially, causing confusion in path order and visual grounding. We also found optimal performance with N=3 reasoning thoughts from previous steps. Excessive history introduces irrelevant information that interferes with current decisions, while insufficient context leads to navigation failure.

TABLE II

ABLATION ON COMPONENT CONFIGURATIONS FOR NAVIGATION

Method	SR (↑)	SPL (↑)
w/o Refinement	63.6	47.5
w/o Progress Estimation	63.6	47.3
w/o Reasoning History (N=0)	36.4	29.1
w/ Reasoning History (N=1)	54.6	39.8
w/ Reasoning History (N=5)	54.6	37.4
w/ ALL, Reasoning History (N=3)	**90.9**	**81.0**

* ALL denotes the use of all instruction comprehension components.
† N denotes the length of the reasoning history window.

TABLE III

COMPARISON OF GPT VARIANTS FOR NAVIGATION

Method	RT [s]	NE (↓)	SR (↑)	SPL (↑)
CLOI-NAV-GPT4o	187.8	**1.15**	87.8	71.9
CLOI-NAV-GPT5	**452.6**	1.25	**95.1**	**77.7**

Effect of Improved LLMs on Navigation. We compared GPT variants to assess their impact on navigation performance. As shown in Table III, GPT-5 requires 2x longer reasoning time but achieves superior results, notably confirming that advances in reasoning capabilities lead to better navigation.

V. CONCLUSION

In this work, we present CLOI-NAV, a novel framework that effectively addresses complex sequential instructions in realistic navigation scenarios. We demonstrate that CLOI-NAV significantly improves navigation performance through enhanced instruction comprehension and rich visual scene understanding. Furthermore, we expect our method will benefit from advances in LLM spatial and semantic reasoning capabilities. Building on these promising results, we plan to validate our VLN pipeline through deployment on real robot systems.

REFERENCES

[1] C.-Y. Ma, J. Lu, Z. Wu, G. AlRegib, Z. Kira, R. Socher, and C. Xiong, "Self-monitoring navigation agent via auxiliary progress estimation," *arXiv preprint arXiv:1901.03035*, 2019.

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

[3] A.-C. Cheng, Y. Ji, Z. Yang, Z. Gongye, X. Zou, J. Kautz, E. Bıyık, H. Yin, S. Liu, and X. Wang, "Navila: Legged robot vision-language-action model for navigation," *arXiv preprint arXiv:2412.04453*, 2024.

[4] J. Zhang, K. Wang, R. Xu, G. Zhou, Y. Hong, X. Fang, Q. Wu, Z. Zhang, and H. Wang, "Navid: Video-based vlm plans the next step for vision-and-language navigation," *arXiv preprint arXiv:2402.15852*, 2024.

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

[6] G. Zhou, Y. Hong, and Q. Wu, "Navgpt: Explicit reasoning in vision-and-language navigation with large language models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 7, 2024, pp. 7641–7649.

[7] Y. Long, W. Cai, H. Wang, G. Zhan, and H. Dong, "Instructnav: Zero-shot system for generic instruction navigation in unexplored environment," *arXiv preprint arXiv:2406.04882*, 2024.

[8] Y. Qiao, W. Lyu, H. Wang, Z. Wang, Z. Li, Y. Zhang, M. Tan, and Q. Wu, "Open-nav: Exploring zero-shot vision-and-language navigation in continuous environment with open-source llms," *arXiv preprint arXiv:2409.18794*, 2024.

[9] Y. Yang, H. Yang, J. Zhou, P. Chen, H. Zhang, Y. Du, and C. Gan, "Snapmem: Snapshot-based 3d scene memory for embodied exploration and reasoning," 2024.

[10] 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.

[11] P. Liu, Y. Orru, J. Vakil, C. Paxton, N. M. M. Shafiullah, and L. Pinto, "Ok-robot: What really matters in integrating open-knowledge models for robotics," *arXiv preprint arXiv:2401.12202*, 2024.

[12] D. S. Chaplot, D. P. Gandhi, A. Gupta, and R. R. Salakhutdinov, "Object goal navigation using goal-oriented semantic exploration," *Advances in Neural Information Processing Systems*, vol. 33, pp. 4247–4258, 2020.

[13] Z. Zhou, Y. Hu, L. Zhang, Z. Li, and S. Chen, "Beliefmapnav: 3d voxel-based belief map for zero-shot object navigation," *arXiv preprint arXiv:2506.06487*, 2025.

[14] N. Kim, O. Kwon, H. Yoo, Y. Choi, J. Park, and S. Oh, "Topological semantic graph memory for image-goal navigation," in *Conference on Robot Learning*. PMLR, 2023, pp. 393–402.

[15] S. H. Allu, I. Kadosh, T. Summers, and Y. Xiang, "Autonomous exploration and semantic updating of large-scale indoor environments with mobile robots," *arXiv preprint arXiv:2409.15493*, 2024.

[16] Z. Wang, X. Li, J. Yang, Y. Liu, J. Hu, M. Jiang, and S. Jiang, "Lookahead exploration with neural radiance representation for continuous vision-language navigation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 13 753–13 762.

[17] X. Lei, M. Wang, W. Zhou, and H. Li, "Gaussnav: Gaussian splatting for visual navigation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[18] K. P. Singh, J. Salvador, L. Weihs, and A. Kembhavi, "Scene graph contrastive learning for embodied navigation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 10 884–10 894.

[19] H. Yin, X. Xu, Z. Wu, J. Zhou, and J. Lu, "Sg-nav: Online 3d scene graph prompting for llm-based zero-shot object navigation," *Advances in neural information processing systems*, vol. 37, pp. 5285–5307, 2024.

[20] 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.

[21] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *International conference on machine learning*. PMLR, 2023, pp. 19 730–19 742.

[22] L. H. Li, P. Zhang, H. Zhang, J. Yang, C. Li, Y. Zhong, L. Wang, L. Yuan, L. Zhang, J.-N. Hwang *et al.*, "Grounded language-image pre-training," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 965–10 975.

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

[24] A. Anwar, J. Welsh, J. Biswas, S. Pouya, and Y. Chang, "Remembr: Building and reasoning over long-horizon spatio-temporal memory for robot navigation," *arXiv preprint arXiv:2409.13682*, 2024.

[25] V. J. A. Anacta, A. Schwering, R. Li, and S. Muenzer, "Orientation information in wayfinding instructions: evidences from human verbal and visual instructions," *GeoJournal*, vol. 82, no. 3, pp. 567–583, 2017.

[26] B. Yamauchi, "A frontier-based approach for autonomous exploration," in *Proceedings 1997 IEEE International Symposium on Computational Intelligence in Robotics and Automation CIRA'97.'Towards New Computational Principles for Robotics and Automation'*. IEEE, 1997, pp. 146–151.

[27] A. Z. Ren, J. Clark, A. Dixit, M. Itkina, A. Majumdar, and D. Sadigh, "Explore until confident: Efficient exploration for embodied question answering," *arXiv preprint arXiv:2403.15941*, 2024.

[28] 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.

[29] K. Yadav, R. Ramrakhya, S. K. Ramakrishnan, T. Gervet, J. Turner, A. Gokaslan, N. Maestre, A. X. Chang, D. Batra, M. Savva *et al.*, "Habitat-matterport 3d semantics dataset," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 4927–4936.