

# CGFM-Nav: Cognitive Graph-Field Memory for Semantic-Guided Lifelong Multimodal Embodied Navigation

Yuxiang Xiao\*, Xibei Chen\*, Xin Zhou\*, Jie Chen, Yifeng Zhang†, Guillaume Sartoretti

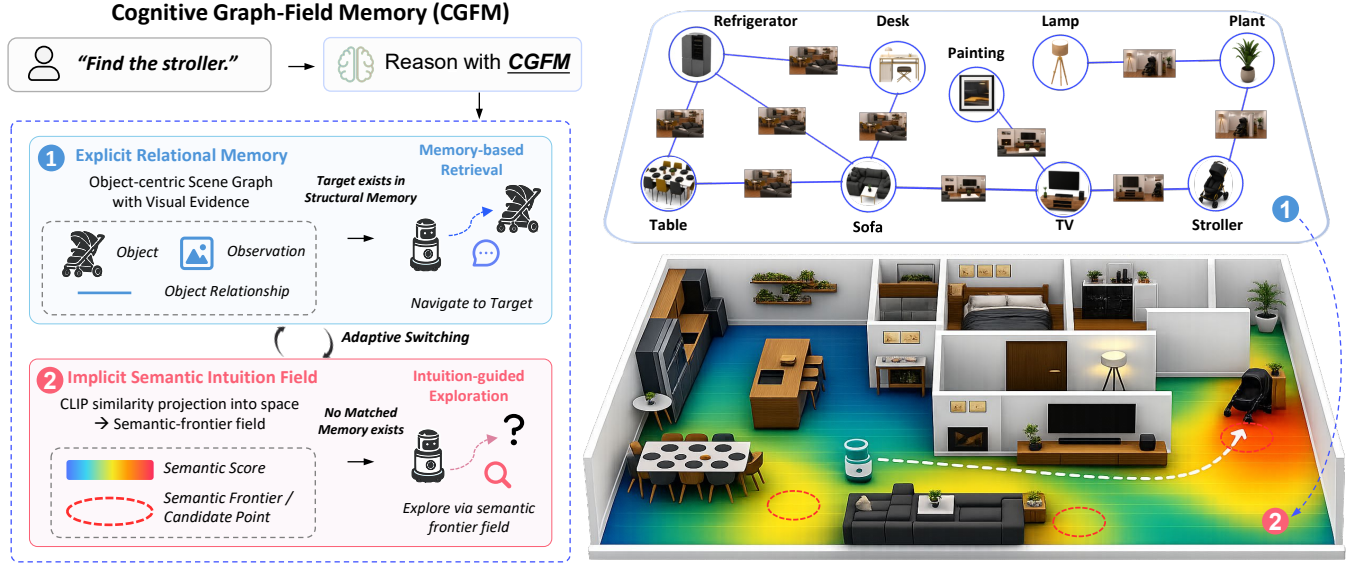


Fig. 1. Overview of the proposed Cognitive Graph-Field Memory (CGFM), inspired by the complementary mechanisms of memory and intuition underlying human navigation. (1) The multimodal scene graph serves as explicit relational memory by storing objects, spatial relations, and visual observations, enabling direct retrieval and navigation toward previously observed targets. (2) When no reliable target match is identified, graph evidence is projected into a semantic-frontier field that provides implicit spatial intuition, guiding exploration toward promising frontiers and regions.

## I. INTRODUCTION

Vision-and-Language Navigation (VLN) enables embodied agents to understand multimodal goals and navigate effectively in complex environments. It has broad applications in assistive robotics, autonomous logistics, industrial inspection, and disaster response [1], [2]. Most training-based VLN methods learn navigation behavior from task-specific supervision, which makes their performance closely tied to the training distribution but less adaptable to unseen environments and open-vocabulary goals [3], [4], [5], [6]. Conversely, training-free VLN offers a promising alternative by leveraging the open-world knowledge of pretrained language and vision-language models to guide navigation without costly and complicated process of policy training [7], [8], [9]. In this paradigm, the agent depends heavily on how it represents the environment, specifically how visual observations, semantic cues, and spatial structure are integrated to support memory, reasoning, and decision-making.

Existing environment representations in training-free VLN can be broadly grouped into two categories. The first uses 2D semantic maps and frontiers, where occupancy maps and semantic heatmaps support online exploration and target search [10], [11], [12], [13], [14]. These representations are efficient and easy to deploy, but they usually lack explicit object relations and long-term memory. The second builds scene graphs or memory graphs by organizing objects and their relations into graph structures, which enables target retrieval and relational reasoning [15], [16], [17], [18], [19]. However, such graph-based representations are often discrete and do not directly provide exploration guidance in continuous space. Overall, existing methods struggle to jointly achieve efficient exploration, structured memory, and geometric consistency, highlighting the need for an environment representation that supports semantic memory, continuous exploration guidance, and efficient online updates.

When navigating an unfamiliar or partially known environment, humans typically do not rely on a single form of representation [20]. Instead, two complementary mechanisms are often involved. One is explicit memory: for example, remembering that a chair was seen next to a table, or that a mug is likely to be found in a kitchen. The other is a more implicit spatial intuition: even without knowing the exact location of the target, one may still have a sense that

Yuxiang Xiao, Xibei Chen, Xin Zhou, Jie Chen, Yifeng Zhang, and Guillaume Sartoretti are with the Department of Mechanical Engineering, National University of Singapore, Singapore (E-mail: {yuxiangxiao, xibeichen, zhouxin, chen.jie, yifeng}@u.nus.edu, guillaume.sartoretti@nus.edu.sg).

\* Yuxiang Xiao, Xibei Chen, and Xin Zhou contributed equally to this work.

† Corresponding author: Yifeng Zhang.

a certain direction is more likely to lead to a kitchen, or that an unexplored region is worth investigating. The former supports slower, structured recall and reasoning, while the latter provides a faster, less precise but efficient spatial bias.

Motivated by this cognitive perspective, we propose Cognitive Graph-Field Memory (CGFM), a unified graph-field representation that couples explicit relational memory with continuous spatial intuition, as illustrated in Fig. 1. In CGFM, the scene graph serves as an explicit relational memory layer, organizing objects, spatial relations, and raw visual evidence for target retrieval and long-horizon reasoning: when a task-relevant target node exists in the graph, the agent navigates to it directly, akin to explicit recall. When no reliable target match can be identified, the agent instead consults a continuous semantic-frontier field derived from the same graph. By projecting visual and semantic evidence back into space, this field indicates which regions are both semantically relevant to the goal and worth exploring, resembling human spatial intuition even in the absence of explicit memory.

The graph and the field are thus not independent modules but complementary forms within the same cognitive representation: the graph’s structured semantic knowledge provides the basis for field inference, while the exploration bias induced by the field in turn guides the continual expansion and updating of the graph, together forming a navigation mechanism analogous to the cooperative interaction between memory and intuition in human cognition.

Building upon CGFM, we further introduce CGFM-Nav, a foundation-model-based framework for lifelong multimodal navigation. It incrementally collects RGB-D observations into the CGFM, extracts a compact task-relevant subgraph and combines it with decision history for VLM reasoning, and executes goal navigation or frontier/semantic exploration guided by CGFM, with verification feedback closing the loop for continual memory update. Preliminary experiments on GOAT-Bench [21] show that, using the same Qwen3-VL-8B backbone, CGFM-Nav improves the overall success rate from 53.2% to 63.0% and SPL from 30.0% to 39.6%, while achieving consistent gains in multimodal navigation. Moreover, CGFM-Nav with Qwen3-VL-8B outperforms MSG-Nav [19] with GPT-4o overall and on both category and language goals. These results suggest that coupling a persistent multimodal scene graph with a goal-conditioned semantic-frontier field enables mapping, exploration, and navigation to operate on a shared and continuously updated representation, providing an effective approach to lifelong navigation.

## II. CGFM-NAV

### A. Problem Formulation and Overview

In this work, we consider lifelong open-vocabulary navigation with multimodal goals in an initially unknown environment, where the agent operates without task-specific policy training. Each episode consists of  $K$  sequential subtasks, with each goal of subtask  $k$  specified by an object category, a natural-language description, or a reference image, while the accumulated environment memory is retained across subtasks. At each decision cycle  $t$ , the agent receives an RGB-

D observation and selects a high-level navigation destination  $h_t$ . A low-level planner executes primitive actions until the target is verified or the navigation budget is exhausted.

During each navigation subtask  $k$ , the agent maintains  $\mathcal{M}_t^{(k)} = (\mathcal{G}_t, \mathcal{F}_t^{(k)}, \mathcal{D}_t^{(k)})$ , where  $\mathcal{G}_t$  is the persistent multimodal scene graph,  $\mathcal{F}_t^{(k)}$  is the goal-conditioned semantic-frontier field, and  $\mathcal{D}_t^{(k)}$  is the short-term decision memory. Rejected candidates are not tracked separately: their verification outcomes are stored in the corresponding graph nodes and reflected in the field by suppressing their contribution. At the start of a new subtask, the graph and its verification history are retained, the field is reconstructed for the new goal, and the decision memory is reset.

The overall architecture and closed-loop navigation process of CGFM-Nav are illustrated in Fig. 2.

### B. Cognitive Graph-Field Memory

1) *Perception and Multimodal Scene Graph Construction*: Following MSGNav [19], at each high-level cycle we process newly collected RGB-D frames using YOLOv8-World [22] and SAM [23] for open-vocabulary detection and segmentation, and extract CLIP [24] features for target retrieval and VLM reasoning. These observations are incrementally associated into a multimodal scene graph  $\mathcal{G}_t = (\mathcal{V}_t, \mathcal{E}_t)$ , where each node  $v_i = (c_i, \mathbf{p}_i, \mathbf{z}_i, \mathcal{I}_i)$  stores an object’s category, 3D location, visual-semantic feature, and accumulated observation images, and each edge encodes a spatial or semantic relation between two objects. Beyond MSGNav’s base representation, we augment each node with goal-conditioned verification statistics: the verification count, associated subtask goal, final outcome, and supporting multiview feedback. Rather than deleting a rejected node, we suppress its contribution to the semantic field for the remainder of the subtask, preserving useful visual evidence while preventing it from repeatedly attracting the agent.

2) *Semantic-Frontier Field Construction*: The discrete scene graph is effective for retrieving known objects, but it does not directly assign exploration values to unobserved regions. We therefore project the semantic evidence stored in  $\mathcal{G}_t$  onto a 2D navigation grid and construct a goal-conditioned dense semantic field. For each graph node  $v_i$ , we compute a goal-relevance score  $s_i(q)$  from its CLIP similarity to the goal, subtracting a background baseline so that only meaningfully related objects act as semantic sources, and gating it with a verification mask so that a node finally rejected for the current goal contributes nothing to the field, while an uncertain observation does not suppress it. Each remaining object is treated as a point source, and its relevance is diffused through the observed, non-obstacle region as a weight  $w_i(\mathbf{x})$  that decays exponentially with geodesic distance from the object, vanishing beyond a fixed cutoff radius. At each grid cell, we aggregate the  $K_s$  objects with the largest local contribution into the field value

$$S_t(\mathbf{x} | q) = \sum_{i \in \text{Top}_{K_s}(\mathbf{x})} s_i(q) w_i(\mathbf{x}). \quad (1)$$

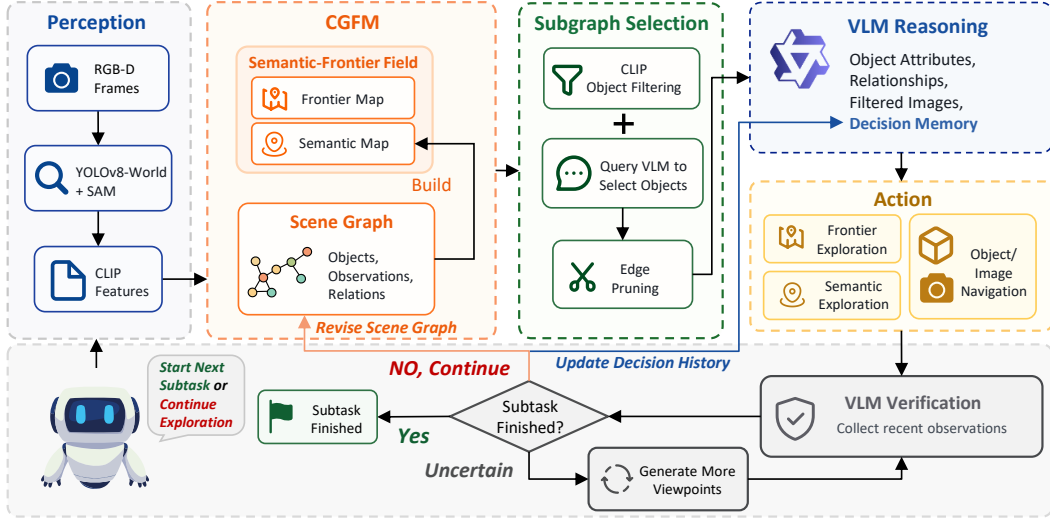


Fig. 2. Overview of the proposed CGFM-Nav framework. RGB-D observations are stored in a structural multimodal scene graph, whose goal-conditioned evidence constructs a semantic-frontier field. A compact key subgraph and short-term decision memory support VLM reasoning for object/image navigation or frontier/semantic exploration. Navigation and verification outcomes update the graph, the field, and decision memory, forming a closed decision loop.

The semantic map is kept synchronized with the evolving scene graph via local updates: only cells around the previous and current projected centers of a changed object are recomputed, while the field is otherwise retained unchanged.

In parallel, the agent extracts a set of frontier candidates  $\mathcal{C}_t = \{f_1, \dots, f_M\}$  from the boundary between explored free space and unknown space, and assigns each a score  $U_t(f | q)$  by averaging the semantic field over a small traversable neighborhood around it and discounting by its geodesic travel distance from the agent, so that closer, more semantically relevant frontiers score higher. If the maximum score is below a pre-defined confidence threshold, the agent falls back to the nearest frontier. The resulting semantic-frontier field is denoted  $\mathcal{F}_t = (S_t, \mathcal{C}_t, U_t)$ ; unlike conventional frontier exploration, which relies mainly on geometric utility, the field  $\mathcal{F}_t$  prioritizes regions that are both reachable and semantically promising for the current target.

### C. CGFM-Guided Navigation

To realize zero-shot VLN through CGFM, we develop CGFM-Nav, a navigation framework built around this representation with four key modules for memory retrieval, VLM reasoning, action selection, and closed-loop refinement.

1) *Semantic-Guided Subgraph Selection*: Since goal-conditioned relevance has already been computed during field construction, CGFM reuses this evidence to replace the object-selection stage of the Key Subgraph Selection (KSS) [19], while retaining its edge-pruning and greedy image-allocation procedures. Nodes with positive relevance and their one-hop neighbors form a semantic seed set  $\mathcal{V}_{\text{seed}}$ ; the remaining nodes are compressed into a residual graph  $\hat{\mathcal{G}}_{\text{res}}$ , over which a VLM performs complementary selection for candidates supported by relational or commonsense cues rather than direct CLIP similarity, described as:

$$\mathcal{V}_{\text{key}} = \mathcal{V}_{\text{seed}} \cup \text{VLMSelect}(\hat{\mathcal{G}}_{\text{res}}, q). \quad (2)$$

$\mathcal{V}_{\text{key}}$  is then processed with MSGNav’s edge-pruning and image-allocation procedure to form the key subgraph  $\tilde{\mathcal{G}}_t$ .

2) *VLM Reasoning with Decision Memory*: The selected key subgraph provides structured environment evidence, while  $\mathcal{D}_t$  maintains short-term temporal information about the current subtask. Rather than accumulating raw VLM responses, we represent  $\mathcal{D}_t$  as an ordered decision log containing the decision cycle, destination type, selected candidate identity and category, and a compact reasoning excerpt. Before prompt construction, the log is compressed into a short list of previously visited or rejected candidates and provided to the VLM:

$$d_t = \text{VLM}(q, \tilde{\mathcal{G}}_t, \mathcal{D}_t). \quad (3)$$

The VLM either selects an object node  $v_i$ , selects an observation image  $I_{i,j} \in \mathcal{I}_i$  from the key subgraph, or requests further exploration. Accordingly,  $d_t$  represents an object target, an image target, or an EXPLORE request. This decision explicitly separates navigation based on targets already represented in the scene graph from exploration guided by the semantic-frontier field.

3) *Action Selection*: For an object candidate, the stored node position is used directly as the destination; for an image candidate, YOLOv8-World and the VLM re-detect and localize the corresponding object. When the VLM requests exploration instead, the destination is chosen hierarchically: if any frontier remains ( $\mathcal{C}_t \neq \emptyset$ ), the agent moves to the one with the highest score  $U_t(f | q)$ , falling back to the nearest frontier if all scores are weak; if no frontier remains but semantic peaks persist, it instead navigates to the centroid of the highest-response region on the semantic map; the subtask fails only when neither option is available. Exploration actions feed new observations directly into the next graph-update cycle without triggering verification.

4) *Verification and Closed-Loop Update*: After reaching a navigation target, the VLM evaluates the candidate from

TABLE I

NAVIGATION PERFORMANCE ACROSS THE FIRST EPISODE OF EACH SCENE ON THE *Val Unseen* SPLIT OF GOAT-BENCH.  $\uparrow$  DENOTES HIGHER IS BETTER. BEST RESULTS ARE HIGHLIGHTED IN BOLD.

| Method                | SR $\uparrow$ |             |             |             | SPL $\uparrow$ |             |             |             |
|-----------------------|---------------|-------------|-------------|-------------|----------------|-------------|-------------|-------------|
|                       | Overall       | Category    | Language    | Image       | Overall        | Category    | Language    | Image       |
| MSGNav (GPT-4o)       | 60.0          | 63.6        | 57.2        | <b>59.1</b> | 37.0           | 35.0        | 33.4        | <b>42.6</b> |
| MSGNav (Qwen3VL-8B)   | 53.2          | 63.6        | 48.4        | 46.6        | 30.0           | 36.4        | 24.8        | 28.1        |
| CGFM-Nav (Qwen3VL-8B) | <b>63.0</b>   | <b>72.7</b> | <b>61.5</b> | 53.4        | <b>39.6</b>    | <b>44.7</b> | <b>38.6</b> | 34.8        |

one or more viewpoints and returns YES, NO-UNCERTAIN, or NO-CONFIRMED, completing, deferring, or rejecting the subtask accordingly; a candidate is also rejected once all viewpoints are exhausted without a positive response. In particular, navigation outcomes and verification feedback are appended to the decision memory, and, for object candidates, also stored in the corresponding graph node, suppressing a rejected object’s contribution to the semantic field for the remainder of the goal. This closes the loop illustrated in Fig. 2: the graph informs the field, the field and decision memory jointly determine the next action, and the resulting observation and feedback update the graph, field, and decision memory for the next cycle. Exploration and rejected targets return the agent to the next cycle, while a verified target advances it to the next subtask.

### III. EXPERIMENTS AND RESULTS

#### A. Experiment Settings

We evaluate our method on GOAT-Bench [21], a multi-modal lifelong open-vocabulary navigation benchmark comprising 360 episodes across 36 scenes, with 2669 subtasks. Our experiments are conducted on the first episode of each scene in the *Val Unseen* split, resulting in 36 evaluated episodes with 278 subtasks. We report two key metrics: Success Rate (SR) and Success weighted by Path Length (SPL). Given  $N$  evaluated tasks, SR is defined as  $SR = N_{\text{success}}/N$ , while SPL is defined as  $SPL = \frac{1}{N} \sum_{i=1}^N S_i \frac{l_i}{\max(l_i, p_i)}$ , where  $S_i \in \{0, 1\}$  indicates task success,  $l_i$  denotes the shortest-path distance to the goal, and  $p_i$  is the actual path length traveled by the agent. We adopt the official success-distance threshold of GOAT-Bench, where a subtask is successful if the agent stops within **0.25** m of a valid target viewpoint.

For comparison, we report the original MSGNav [19] results obtained with GPT-4o (closed-source) and adopt Qwen3-VL-8B-Instruct (open-source) as the VLM backbone for MSGNav and CGFM-Nav in our experiments.

#### B. Results

Table I summarizes the multimodal navigation results on the subset from the *Val Unseen* split of GOAT-Bench. This subset covers category, language, and image goals, allowing us to evaluate CGFM-Nav across different forms of target specification. Under the same Qwen3-VL-8B backbone, CGFM-Nav improves the overall SR from 53.2% to 63.0% and SPL from 30.0% to 39.6%, with consistent gains across all three goal modalities. Such improvements are particularly

evident for category and language goals, where navigation mainly depends on accumulating semantic evidence and efficiently exploring unseen regions. By coupling structured scene-graph memory with a continuous semantic-frontier field, CGFM enables the agent to better reuse previous observations while guiding exploration toward semantically promising frontiers and regions. Notably, CGFM-Nav with the lightweight, locally deployable Qwen3-VL-8B even outperforms MSGNav with the cloud-based GPT-4o in overall, category, and language performance. This result suggests that a stronger environment representation can effectively compensate for part of the performance gap between VLM backbones, highlighting the practical potential of CGFM for efficient on-device embodied navigation.

Meanwhile, replacing GPT-4o with Qwen3-VL-8B causes a substantial performance drop on image-goal tasks for MSGNav, with SR decreasing from 59.1% to 46.6% and SPL from 42.6% to 28.1%. CGFM improves these results to 53.4% SR and 34.8% SPL, but still does not surpass MSGNav with GPT-4o. This suggests that image-goal navigation relies more heavily on fine-grained visual correspondence than semantic exploration. While CGFM substantially improves exploration by providing more informative semantic guidance, further improvements on image-goal navigation are likely to depend on stronger visual representations or navigation-oriented VLM backbones.

### IV. CONCLUSION AND FUTURE WORK

In this work, we introduced CGFM-Nav, a training-free framework for lifelong multimodal navigation that employs a unified graph-field representation to integrate structured scene-graph memory with semantic-guided exploration. Experimental results on GOAT-Bench demonstrate that CGFM-Nav consistently improves navigation performance. These results suggest that combining explicit semantic memory with spatial intuition provides an effective cognitive representation for embodied navigation. In future work, we plan to evaluate CGFM on additional semantic navigation benchmarks with long-horizon and dynamic settings. We also plan to validate CGFM on a wheeled mobile robot in an office environment, enabling real-world evaluation under perception and localization uncertainties.

## REFERENCES

- [1] J. Gu, E. Stefani, Q. Wu, J. Thomason, and X. E. Wang, "Vision-and-language navigation: A survey of tasks, methods, and future directions," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 7606–7623.
- [2] Y. Zhang, Z. Ma, J. Li, Y. Qiao, Z. Wang, J. Chai, Q. Wu, M. Bansal, and P. Kordjamshidi, "Vision-and-language navigation today and tomorrow: A survey in the era of foundation models," *arXiv preprint arXiv:2407.07035*, 2024.
- [3] 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.
- [4] G. Zhou, Y. Hong, Z. Wang, X. E. Wang, and Q. Wu, "Navgpt-2: Unleashing navigational reasoning capability for large vision-language models," in *European Conference on Computer Vision*. Springer, 2024, pp. 260–278.
- [5] J. Zhang, K. Wang, S. Wang, M. Li, H. Liu, S. Wei, Z. Wang, Z. Zhang, and H. Wang, "Uni-navid: A video-based vision-language-action model for unifying embodied navigation tasks," *arXiv preprint arXiv:2412.06224*, 2024.
- [6] Z. Zhu, X. Wang, Y. Li, Z. Zhang, X. Ma, Y. Chen, B. Jia, W. Liang, Q. Yu, Z. Deng *et al.*, "Move to understand a 3d scene: Bridging visual grounding and exploration for efficient and versatile embodied navigation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 8120–8132.
- [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] F. Ziliotto, T. Campari, L. Serafini, and L. Ballan, "Tango: training-free embodied ai agents for open-world tasks," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 24 603–24 613.
- [9] H. Yin, H. Wei, X. Xu, W. Guo, J. Zhou, and J. Lu, "Gc-vln: Instruction as graph constraints for training-free vision-and-language navigation," in *Conference on Robot Learning*. PMLR, 2025, pp. 1809–1824.
- [10] 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.
- [11] J. Chen, G. Li, S. Kumar, B. Ghanem, and F. Yu, "How to not train your dragon: Training-free embodied object goal navigation with semantic frontiers," *arXiv preprint arXiv:2305.16925*, 2023.
- [12] N. Yokoyama, S. Ha, D. Batra, J. Wang, and B. Bucher, "Vlfn: Vision-language frontier maps for zero-shot semantic navigation," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 42–48.
- [13] 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.
- [14] M. Zhang, Y. Du, C. Wu, J. Zhou, Z. Qi, J. Ma, and B. Zhou, "Apex-nav: An adaptive exploration strategy for zero-shot object navigation with target-centric semantic fusion," *IEEE Robotics and Automation Letters*, 2025.
- [15] N. Hughes, Y. Chang, and L. Carlone, "Hydra: A real-time spatial perception system for 3d scene graph construction and optimization," *arXiv preprint arXiv:2201.13360*, 2022.
- [16] Q. Gu, A. Kuwajerwala, S. Morin, K. M. Jatavallabhula, B. Sen, A. Agarwal, C. Rivera, W. Paul, K. Ellis, R. Chellappa *et al.*, "Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 5021–5028.
- [17] A. Werby, C. Huang, M. Büchner, A. Valada, and W. Burgard, "Hierarchical open-vocabulary 3d scene graphs for language-grounded robot navigation," *arXiv preprint arXiv:2403.17846*, 2024.
- [18] X. Zhou, T. Xiao, L. Liu, Y. Wang, M. Chen, X. Meng, X. Wang, W. Feng, W. Sui, and Z. Su, "Fsr-vln: Fast and slow reasoning for vision-language navigation with hierarchical multi-modal scene graph," *arXiv preprint arXiv:2509.13733*, 2025.
- [19] X. Huang, S. Zhao, Y. Wang, X. Lu, W. Zhang, R. Qu, W. Li, Y. Wang, and C. Wen, "MSGNav: Unleashing the power of multi-modal 3d scene graph for zero-shot embodied navigation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2026, pp. 37 154–37 163.
- [20] E. C. Tolman, "Cognitive maps in rats and men," *Psychological review*, vol. 55, no. 4, p. 189, 1948.
- [21] M. Khanna, R. Ramrakhya, G. Chhablani, S. Yenamandra, T. Gervet, M. Chang, Z. Kira, D. S. Chaplot, D. Batra, and R. Mottaghi, "GOAT-Bench: A benchmark for multi-modal lifelong navigation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 16 373–16 383.
- [22] T. Cheng, L. Song, Y. Ge, W. Liu, X. Wang, and Y. Shan, "YOLO-World: Real-time open-vocabulary object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 16 901–16 911.
- [23] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, "Segment anything," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 4015–4026.
- [24] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proceedings of the 38th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 139, 2021, pp. 8748–8763.