Where to Fuse in the VLM Era: A Survey on Integrating Knowledge into Object Goal Navigation

Bokeon Suh[1], Jiseon Kim[1] and Giseop Kim[1*]

***Abstract*— The rapid advancement of robotics and deep learning has accelerated the use of Embodied AI, where robots autonomously explore and reason in complex real-world environments. With growing demand for domestic service robots, efficient navigation in unfamiliar settings is crucial. Object Goal Navigation (ObjectNav) is a fundamental task for this capability, requiring a robot to find and reach a user-specified object in an unknown environment. Solving ObjectNav demands advanced perception, contextual reasoning, and effective exploration strategies. Recent Vision-Language Models (VLMs) and Large Language Models (LLMs) provide agents with external common knowledge and reasoning capabilities. This paper poses the critical question: "Where should VLM/LLM knowledge be fused into Object Goal Navigation?" Adapted from the Perception–Prediction–Planning paradigm in autonomous driving, we categorize knowledge integration into these three stages, offering a structured survey of object-goal navigation approaches shaped by the VLM era. We conclude by discussing current dataset limitations and future directions, such as socially interactive navigation.**

I. INTRODUCTION

Object Goal Navigation (ObjectNav) tasks require an agent to reach a user-specified object in an unseen environment. The agent must integrate visual perception, spatial reasoning, contextual understanding, and exploration strategies to succeed. End-to-end methods [1, 2, 3, 4] rely on visual features and reinforcement learning, whereas recent approaches leverage Vision-Language Models (VLMs) and Large Language Models (LLMs) to exploit external knowledge and reasoning ability.

In this paper, we categorize VLM/LLM knowledge integration into three levels: **III-A) Perception**, **III-B) Prediction**, and **III-C) Planning**, analyzing representative studies and their contributions inspired by the perception–prediction–planning paradigm[5, 6, 7, 8] in autonomous driving. Finally, we discuss the limitations of current datasets and introduce future directions, including socially interactive and multi-floor navigation.

II. RELATED WORKS

A. Survey of Surveys

Recent advances in VLMs and LLMs has shifted ObjectNav from supervised methods to commonsense-guided exploration. Ieong et al. [9] review goal-driven navigation tasks (e.g., PointNav, ImageNav, ObjectNav, AudioGoalNav) and categorize ObjectNav methods by inference domain such as latent maps, graphs, implicit representation, language, etc. Sun et al. [10] focus on ObjectNav, dividing methods into end-to-end, modular, and zero-shot. Unlike these works, this paper examines where to fuse LLM/VLM knowledge by framing ObjectNav methods within the Perception–Prediction–Planning framework (Fig. 1).

B. The Common Core of Diverse Embodied AI Tasks

Although ObjectNav and other vision-language embodied AI tasks differ in input modalities and goal specifications, they share a common structure: given a goal instruction, the agent must leverage visual observations to reach a meaningful location.

GOAT [11] unifies ObjectNav, ImageNav, and TextNav within a single benchmark, extending goal representations from class labels to image and natural language. REVERIE [12] requires following natural language instructions to navigate and ground implicit object references, emphasizing fine-grained language–vision grounding. VQA [13] and EQA [14] shift the output to question answering, but like ObjectNav, EQA requires exploration-driven visual reasoning within embodied environments. These tasks share three key elements:

1) Explicit, exemplar, or descriptive goal instructions
2) Active exploration for visual reasoning
3) Integration of multimodal inputs

Taken together, these elements reveal a shared problem structure: combining linguistic, visual, and spatial information to support goal-directed decision-making. Within this scope, ObjectNav places particular emphasis on exploration through such integration.

III. WHERE TO FUSE?

Recent studies have integrated LLMs and VLMs into ObjectNav, where robots locate user-specified object in unseen environments. This paper examines where such knowledge can be injected by organizing prior approaches into the Perception, Prediction, and Planning stages.

A. Perception Level Fusion

With the multimodal capabilities of VLMs, an agent can accurately recognize the visual and spatial information of the current scene. This section discusses Scene Object Understanding, which focuses on identifying individual objects, and Scene Spatial Relation Understanding, which focuses on the structural arrangement of these objects.

[1]All authors are with DGIST, Daegu, Republic of Korea {bksuh, jiseon.kim, gsk}@dgist.ac.kr

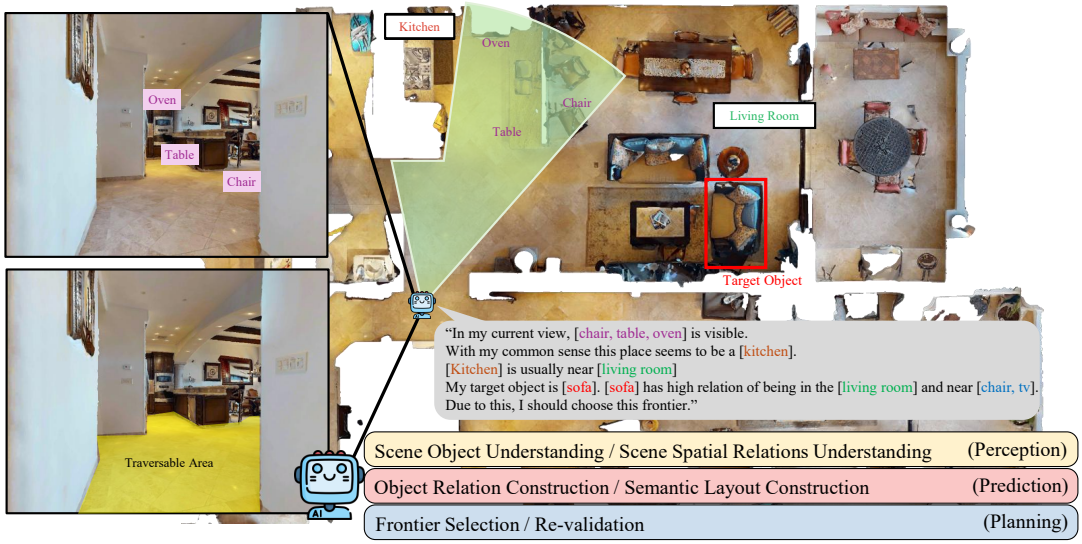


Fig. 1. The agent leverages VLM/LLM to understand the scene (Perception), predict the target object location based on context (Prediction) and selects the most promising frontier to explore (Planning).

1) Scene Object Understanding: By leveraging VLMs, an agent can address the question, "What objects are visible now?" Unlike conventional ObjectNav methods limited to a closed set of predefined objects, VLMs enable open-set detection and allow the system to recognize novel objects. Existing approaches can be categorized as following:

1) VLMs for Information Encoding
2) VLMs for Textual Description

Information Encoding leverages the image–text alignment capability of VLMs to capture relationships between visual inputs and textual queries. CoW [15] combines CLIP [16] similarity scores with Grad-CAM to estimate target locations. ZSON [17] applies CLIP embeddings for policy learning, where goals are encoded as image embeddings during training and replaced with textual embeddings at inference. ESC [18] integrates GLIP [19] for object detection. GAMap [20] queries GPT-4 for both affordance and geometric attributes of the goal, encodes them as text embeddings, and adds them into CLIP similarity maps.

Cosine similarity between images and target object labels is another common strategy. BLIP-2 [21] is widely adopted. VLFM [22] projects similarity scores onto 2D maps. ApexNav [23] extends this approach by using an LLM to propose visually similar objects, thereby reducing false detections. SemNav [24] replaces BLIP-2 with GPT-4, while Bajpai et al.[25] introduce uncertainty by generating multiple prompts with LLMs, computing BLIP-2 similarities, and measuring variance as an uncertainty signal. Shi et al.[26] describe image sequences with LLaVA [27] and compute similarity to the goal. WMNav [28] evaluates panoramic views, with Gemini estimating information gain at different angles.

Textual Description directly generates natural-language captions of observed images. PixNav [29] employs LLaMA-Adapter [30] for object recognition. OpenFMNav [31] uses GPT-4 to describe scene objects and forwards them to detectors. CL-CoTNav [32] leverages Qwen [33] to identify subgoal objects related to the target. VoroNav [34] builds Voronoi-based maps, updating each node with 360° observations described by BLIP [35].

2) Scene Spatial Relations Understanding: This stage captures the spatial arrangement and relations of objects, addressing the question "What is placed where, and how?" PixNav [29] and CL-CoTNav [32] adopt a question–answering (QA) paradigm to infer object layouts. SG-Nav [36] represents relations in a graph, where commonsense co-occurrences are first proposed by an LLM and then verified with a VLM to build a scene graph.

Spatial relation understanding extends beyond static objects to assessing environment traversability. Unlike prior methods that rely on detailed 2D maps, DyNaVLM [37] adopts a vision-driven approach. As shown in Fig. 4a, candidate waypoints are projected onto the RGB image, and the VLM is queried: "If the agent follows this path, is a collision likely?" Using visual reasoning, the VLM filters unsafe candidates and excludes them from the plan.

Moreover, the concept can be extended from static objects to dynamic agents, enabling social navigation. For instance, Social-LLaVA [38] uses VLMs to extract social cues such as human positions, postures, and gaze directions. This allows the agent to avoid collisions and respect personal spaces, resulting in socially acceptable paths.

In conclusion, Scene Spatial Relation Understanding includes object arrangements, traversability, and interactions with dynamic agents. It enables efficient path planning and allows socially aware, safe navigation in complex environments.

B. Prediction-Level Fusion

This stage infers unseen semantic information from partial observations, addressing questions such as "What relations exist among the observed objects?" and "What type of room is this?" At this point, the commonsense knowledge and reasoning capabilities of LLMs are fully exploited. This stage is commonly approached from two directions: Object

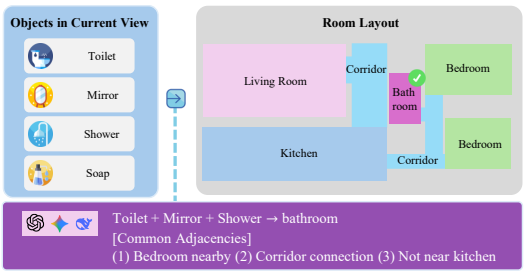


Fig. 2. Based on the objects like 'toilet' and 'shower'. LLM infers the room type as 'Bathroom' and predicts the layout of unobserved areas by reasoning about common room layout adjacencies.

Relation Construction, which predicts the location of unseen targets based on co-occurrence with observed objects, and Semantic Layout Construction, which infers room categories and spatial adjacencies from partial observations.

1) Object Relation Construction: Knowledge of typical object co-occurrences supports efficient exploration. For example, if the goal is a sofa, the robot should prioritize areas with a detected TV over those with a bed or toilet. LLM commonsense plays a central role in evaluating object relations and reducing unnecessary exploration.

The most direct approach is to query an LLM about the relationship between observed and goal objects. ESC [18] presents observed and goal objects to the LLM and asks whether their co-existence is plausible. L3MVN [39] extends this idea by evaluating the likelihood of sentence completions such as "If objects A, B, and C are observed at this frontier, then the goal object is also present."

Another strategy is to have the LLM generate lists of objects commonly associated with the target and use them as exploration cues. BeliefMapNav [40] and CL-CoTNav [32] follow this approach, directing the agent to regions where related objects are detected. Similarly, CogNav [41] exploits these associations as directional signals, biasing exploration toward promising areas.

Beyond direct associations, SGM [42] predicts unobserved regions. It applies general knowledge of object relations to outpaint unseen areas of a 2D map using MAE [43], enabling proactive exploration. Incorporating such relational knowledge reduces redundant search and improves navigation efficiency.

2) Semantic Layout Construction: Beyond object relations, LLMs can infer room types and structural context from partial observations. For instance, if a sink and a toilet are observed (Fig. 2), the agent may infer it is in a bathroom. Reasoning such as "Bathrooms are often adjacent to bedrooms or hallways, but not kitchens" further guides exploration by suggesting likely spatial adjacencies.

ESC [18] and PixNav [29] directly prompt LLMs to infer the current room type from observed objects. E2BA [44] extends this by reasoning at the frontier level, combining room classification with goal relevance. CL-CoTNav [32] predicts likely goal-containing rooms and verifies them against actual observations during exploration. CogNav [41] further decomposes exploration into five stages, prioritizing spaces with strong goal associations.

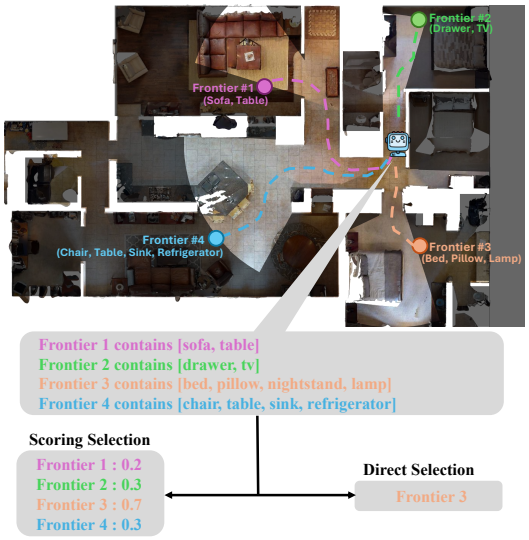


Fig. 3. Two distinct strategies for Frontier Selection. **Scoring Selection** LLM assigns a quantitative score to each frontier based on the semantic relevance of the objects it contains. **Direct Selection** LLM acts as a high-level decision-maker, directly choosing the next best frontier based on all available context.

Other works address broader structural understanding. TopV-Nav [45] interprets 2D maps from a top-view perspective, while BeliefMapNav [40] infers both the current room and its adjacent spaces. DAR [46] generates unobserved layouts similar to SGM [42], but differs by prompting an LLM to classify frontiers by room type and expected objects, guiding the diffusion model to populate maps with semantically plausible content.

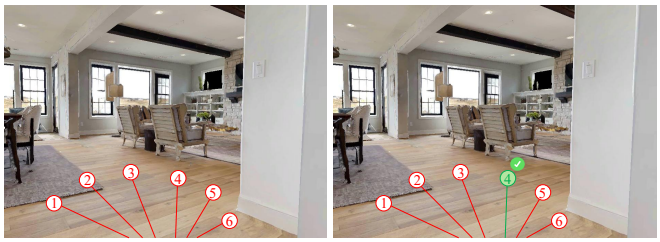
C. Planning-level Fusion

This stage integrates prior information to decide "What should the agent do next?" Frontier Selection, which chooses among unexplored frontiers, and Re-validation, which reassesses decisions for robustness.

1) Frontier Selection: During exploration, the agent faces multiple frontiers, i.e., boundaries between explored and unexplored regions. VLMs and LLMs act as decision-makers, selecting the frontier most likely to contain the target. They can be grouped into Linguistic and Visual methods, based on whether information is textual or visual.

Linguistic Methods (Fig. 3) convert accumulated exploration data into prompts for VLMs/LLMs. In Scoring Selection, the LLM assigns quantitative values to each frontier, and the agent selects the one with the highest score. OpenFMNav [31], L3MVN [39], and TriHelper [47] score frontiers by passing observed object lists to the LLM. VoroNav [34] incorporates room type and spatial relations, while SG-Nav [36] combines semantic relevance with distance to prioritize both meaningful and reachable frontiers. TopV-Nav [45] encodes surrounding object information into a text-based 2D map, from which the VLM estimates goal likelihoods and integrates them via Gaussian-based scoring.

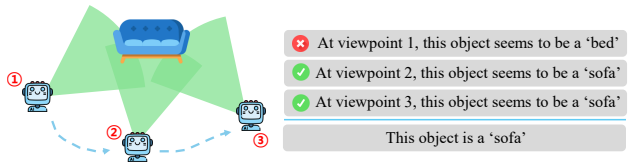
In Direct Selection, no explicit scoring is performed, but the LLM directly selects the next frontier using accumulated context. E2BA [44] and PixNav [29] adopt this approach,



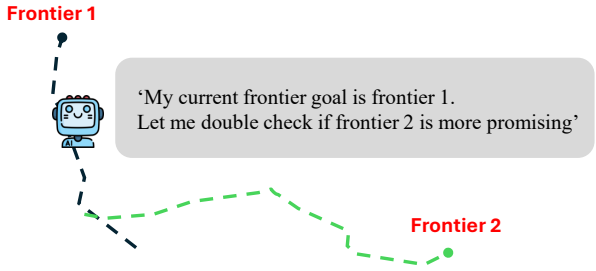
(a) Candidate waypoints are projected onto the egocentric view.

(b) Candidate waypoints are directly evaluated and selects the next waypoint.

Fig. 4. Visual method for Frontier Selection



(a) The agent accumulates evidence from multiple viewpoints to confirm the identity of a potential target object, reducing false positives.



(b) The agent continuously re-assesses its chosen path against other options, allowing for dynamic backtracking or changes in strategy.

Fig. 5. Re-validation strategies at the Planning-Level for robust navigation

while LLM-ZSON [26] integrates semantic value maps and structural maps into the VLM for frontier selection.

Visual Methods leverage the visual reasoning ability of VLMs. As shown in Fig. 4, candidate waypoints are projected onto the agent's current observation and selected directly by the VLM. VLMNav[48], DyNaVLM [37], and WMNav [28] adopt this paradigm, enabling the VLM to act as a high-level decision-maker that selects the most plausible waypoint.

2) Re-validation: The agent should not rely blindly on VLM/LLM outputs but instead re-examine results and adapt its strategy dynamically for efficient navigation.

Some methods focus on detection re-verification to reduce false positives. CogNav [41] structures exploration into five stages, re-confirming goal objects at each step. TriHelper [47] applies binary classification to verify target presence in the current view, while SG-Nav [36] aggregates confidence across multiple viewpoints before committing.

Other methods emphasize strategy adjustment. CL-CoTNav [32] introduces self-verification, allowing the LLM to assess the reliability of its plan prior to execution. E2BA [44] evaluates whether backtracking to previously visited regions may be more efficient, guided by recorded trajectories and observed objects.

Fig. 6. An example of reconstruction artifacts (or defects) found in real-world scan datasets. The depth image (right) shows significant data loss so called 'holes' or 'black cracks'.

IV. Open Challenges and Research Opportunities

A. Issues in Existing Datasets

Although HM3D and MP3D offer realistic indoor environments, both have structural limitations: episode datasets allow only one correct object instance per class, yielding failures and misleading signals, while scene datasets contain missing 3D scanning regions that produce "black cracks" (Fig. 6), leading to inaccurate maps. HSSD[49] addresses these issues with a synthetic alternative featuring high visual fidelity, structural complexity, and scalable data generation.

B. Expanding ObjectNav Task

1) Beyond Single-Floor: Some episodes place the robot's start and goal objects on different floors. Most studies, however, project environments onto 2D grid maps, which effectively restricts agents from using stairs to change floors. Recent works [50, 51] address this limitation by actively incorporating stair traversal for multi-floor navigation.

2) Toward Socially Interactive Navigation: Conventional ObjectNav research has focused on efficient goal-reaching in human-free environments. Recent datasets such as Habicrowd [52], Social-MP3D [53], and Social-HM3D [53] introduced humanoid avatars to simulate human-inhabited spaces, but these works remain confined to simulators and overlook real-world human expressiveness and interaction.

In contrast, real-world studies emphasize social compliant navigation [54, 55, 38], where robots yield and avoid collisions based on basic norms. Yet, both ObjectNav and social compliant navigation often reduce humans to moving obstacles. To succeed in everyday environments, robots must progress toward Socially Interactable Navigation, where they actively interpret situations and provide assistance, enabling natural coexistence and collaborative task execution.

V. Conclusion

This paper examined how VLM and LLM knowledge can be integrated into ObjectNav through perception, prediction, and planning. VLMs enable open-set recognition and spatial understanding, while LLMs provide commonsense reasoning for scene inference and decision-making. These strategies enhance zero-shot navigation and generalization beyond end-to-end learning. We also identified key challenges—dataset limitations, multi-floor exploration, and socially interactable navigation—that should be explored to move ObjectNav from simplified tasks toward practical service robots in real-world environments.

REFERENCES

[1] Sixian Zhang, Xinhang Song, Weijie Li, Yubing Bai, Xinyao Yu, and Shuqiang Jiang. Layout-based causal inference for object navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10792–10802, 2023.

[2] Oleksandr Maksymets, Vincent Cartillier, Aaron Gokaslan, Erik Wijmans, Wojciech Galuba, Stefan Lee, and Dhruv Batra. Thda: Treasure hunt data augmentation for semantic navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15374–15383, 2021.

[3] Joel Ye, Dhruv Batra, Abhishek Das, and Erik Wijmans. Auxiliary tasks and exploration enable objectgoal navigation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 16117–16126, 2021.

[4] Karmesh Yadav, Ram Ramrakhya, Arjun Majumdar, Vincent-Pierre Berges, Sachit Kuhar, Dhruv Batra, Alexei Baevski, and Oleksandr Maksymets. Offline visual representation learning for embodied navigation. In *Workshop on Reincarnating Reinforcement Learning at ICLR 2023*, 2023.

[5] Sergio Casas, Abbas Sadat, and Raquel Urtasun. Mp3: A unified model to map, perceive, predict and plan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14403–14412, 2021.

[6] Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, et al. Planning-oriented autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17853–17862, 2023.

[7] Yuxiang Yang, Fenglong Ge, Jinlong Fan, Jufeng Zhao, and Zhekang Dong. Cdrp3: Cascade deep reinforcement learning for urban driving safety with joint perception, prediction, and planning. *IEEE Transactions on Intelligent Transportation Systems*, 2024.

[8] Yunsheng Ma, Burhaneddin Yaman, Xin Ye, Mahmut Yurt, Jingru Luo, Abhirup Mallik, Ziran Wang, and Liu Ren. Aln-p3: Unified language alignment for perception, prediction, and planning in autonomous driving. *arXiv preprint arXiv:2505.15158*, 2025.

[9] I-Tak Ieong and Hao Tang. Multimodal perception for goal-oriented navigation: A survey. *arXiv preprint arXiv:2504.15643*, 2025.

[10] Jingwen Sun, Jing Wu, Ze Ji, and Yu-Kun Lai. A survey of object goal navigation. *IEEE Transactions on Automation Science and Engineering*, 22:2292–2308, 2024.

[11] Mukul Khanna, Ram Ramrakhya, Gunjan Chhablani, Sriram Yenamandra, Theophile Gervet, Matthew Chang, Zsolt Kira, Devendra Singh Chaplot, Dhruv Batra, and Roozbeh Mottaghi. Goat-bench: A benchmark for multi-modal lifelong navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16373–16383, 2024.

[12] Yuankai Qi, Qi Wu, Peter Anderson, Xin Wang, William Yang Wang, Chunhua Shen, and Anton van den Hengel. Reverie: Remote embodied visual referring expression in real indoor environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9982–9991, 2020.

[13] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.

[14] Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. Embodied question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–10, 2018.

[15] Samir Yitzhak Gadre, Mitchell Wortsman, Gabriel Ilharco, Ludwig Schmidt, and Shuran Song. Cows on pasture: Baselines and benchmarks for language-driven zero-shot object navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23171–23181, 2023.

[16] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[17] Arjun Majumdar, Gunjan Aggarwal, Bhavika Devnani, Judy Hoffman, and Dhruv Batra. Zson: Zero-shot object-goal navigation using multimodal goal embeddings. *Advances in Neural Information Processing Systems*, 35:32340–32352, 2022.

[18] Kaiwen Zhou, Kaizhi Zheng, Connor Pryor, Yilin Shen, Hongxia Jin, Lise Getoor, and Xin Eric Wang. Esc: Exploration with soft commonsense constraints for zero-shot object navigation. In *International Conference on Machine Learning*, pages 42829–42842. PMLR, 2023.

[19] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10965–10975, 2022.

[20] Hao Huang, Yu Hao, Congcong Wen, Anthony Tzes, Yi Fang, et al. Gamap: Zero-shot object goal navigation with multi-scale geometric-affordance guidance. *Advances in Neural Information Processing Systems*, 37:39386–39408, 2024.

[21] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

[22] Naoki Yokoyama, Sehoon Ha, Dhruv Batra, Jiuguang Wang, and Bernadette Bucher. Vlfm: Vision-language frontier maps for zero-shot semantic navigation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 42–48. IEEE, 2024.

[23] Mingjie Zhang, Yuheng Du, Chengkai Wu, Jinni Zhou, Zhenchao Qi, Jun Ma, and Boyu Zhou. Apexnav: An adaptive exploration strategy for zero-shot object navigation with target-centric semantic fusion. *arXiv preprint arXiv:2504.14478*, 2025.

[24] Arnab Debnath, Gregory J Stein, and Jana Kosecka. Semnav: A model-based planner for zero-shot object goal navigation using vision-foundation models. *arXiv preprint arXiv:2506.03516*, 2025.

[25] Utkarsh Bajpai, Julius Rückin, Cyrill Stachniss, and Marija Popović. Uncertainty-informed active perception for open vocabulary object goal navigation. *arXiv preprint arXiv:2506.13367*, 2025.

[26] Jin Shi, Satoshi Yagi, Satoshi Yamamori, and Jun Morimoto. Llm-guided zero-shot visual object navigation with building semantic map. In *2025 IEEE/SICE International Symposium on System Integration (SII)*, pages 1274–1279. IEEE, 2025.

[27] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

[28] Dujun Nie, Xianda Guo, Yiqun Duan, Ruijun Zhang, and Long Chen. Wmnav: Integrating vision-language models into world models for object goal navigation. *arXiv preprint arXiv:2503.02247*, 2025.

[29] Wenzhe Cai, Siyuan Huang, Guangran Cheng, Yuxing Long, Peng Gao, Changyin Sun, and Hao Dong. Bridging zero-shot

object navigation and foundation models through pixel-guided navigation skill. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5228–5234. IEEE, 2024.

[30] Renrui Zhang, Jiaming Han, Chris Liu, Peng Gao, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, and Yu Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*, 2023.

[31] Yuxuan Kuang, Hai Lin, and Meng Jiang. Openfmnav: Towards open-set zero-shot object navigation via vision-language foundation models. *arXiv preprint arXiv:2402.10670*, 2024.

[32] Yuxin Cai, Xiangkun He, Maonan Wang, Hongliang Guo, Wei-Yun Yau, and Chen Lv. Cl-cotnav: Closed-loop hierarchical chain-of-thought for zero-shot object-goal navigation with vision-language models. *arXiv preprint arXiv:2504.09000*, 2025.

[33] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

[34] Pengying Wu, Yao Mu, Bingxian Wu, Yi Hou, Ji Ma, Shanghang Zhang, and Chang Liu. Voronav: Voronoi-based zero-shot object navigation with large language model. *arXiv preprint arXiv:2401.02695*, 2024.

[35] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022.

[36] Hang Yin, Xiuwei Xu, Zhenyu Wu, Jie Zhou, and Jiwen Lu. Sg-nav: Online 3d scene graph prompting for llm-based zero-shot object navigation. *Advances in neural information processing systems*, 37:5285–5307, 2024.

[37] Zihe Ji, Huangxuan Lin, and Yue Gao. Dynavlm: Zero-shot vision-language navigation system with dynamic viewpoints and self-refining graph memory. *arXiv preprint arXiv:2506.15096*, 2025.

[38] Amirreza Payandeh, Daeun Song, Mohammad Nazeri, Jing Liang, Praneel Mukherjee, Amir Hossain Raj, Yangzhe Kong, Dinesh Manocha, and Xuesu Xiao. Social-llava: Enhancing robot navigation through human-language reasoning in social spaces, 2024.

[39] Bangguo Yu, Hamidreza Kasaei, and Ming Cao. L3mvn: Leveraging large language models for visual target navigation. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3554–3560. IEEE, 2023.

[40] Zibo Zhou, Yue Hu, Lingkai Zhang, Zonglin Li, and Siheng Chen. Beliefmapnav: 3d voxel-based belief map for zero-shot object navigation. *arXiv preprint arXiv:2506.06487*, 2025.

[41] Yihan Cao, Jiazhao Zhang, Zhinan Yu, Shuzhen Liu, Zheng Qin, Qin Zou, Bo Du, and Kai Xu. Cognav: Cognitive process modeling for object goal navigation with llms. *arXiv preprint arXiv:2412.10439*, 2024.

[42] Sixian Zhang, Xinyao Yu, Xinhang Song, Xiaohan Wang, and Shuqiang Jiang. Imagine before go: Self-supervised generative map for object goal navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16414–16425, 2024.

[43] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

[44] Yuhong Shi, Jianyi Liu, Lihang Sun, and Xinhu Zheng. E 2 ba: Environment exploration and backtracking agent for visual language object navigation. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

[45] Linqing Zhong, Chen Gao, Zihan Ding, Yue Liao, Huimin Ma, Shifeng Zhang, Xu Zhou, and Si Liu. Topv-nav: Unlocking the top-view spatial reasoning potential of mllm for zero-shot object navigation. *arXiv preprint arXiv:2411.16425*, 2024.

[46] Yiming Ji, Kaijie Yun, Yang Liu, Zhengpu Wang, Boyu Ma, Zongwu Xie, and Hong Liu. Diffusion as reasoning: Enhancing object navigation via diffusion model conditioned on llm-based object-room knowledge. *arXiv preprint arXiv:2410.21842*, 2024.

[47] Lingfeng Zhang, Qiang Zhang, Hao Wang, Erjia Xiao, Zixuan Jiang, Honglei Chen, and Renjing Xu. Trihelper: Zero-shot object navigation with dynamic assistance. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10035–10042. IEEE, 2024.

[48] Dylan Goetting, Himanshu Gaurav Singh, and Antonio Loquercio. End-to-end navigation with vision language models: Transforming spatial reasoning into question-answering. *arXiv preprint arXiv:2411.05755*, 2024.

[49] Mukul Khanna, Yongsen Mao, Hanxiao Jiang, Sanjay Haresh, Brennan Shacklett, Dhruv Batra, Alexander Clegg, Eric Undersander, Angel X Chang, and Manolis Savva. Habitat synthetic scenes dataset (hssd-200): An analysis of 3d scene scale and realism tradeoffs for objectgoal navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16384–16393, 2024.

[50] Zeying Gong, Rong Li, Tianshuai Hu, Ronghe Qiu, Lingdong Kong, Lingfeng Zhang, Yiyi Ding, Leying Zhang, and Junwei Liang. Stairway to success: Zero-shot floor-aware object-goal navigation via llm-driven coarse-to-fine exploration. *arXiv preprint arXiv:2505.23019*, 2025.

[51] Lingfeng Zhang, Hao Wang, Erjia Xiao, Xinyao Zhang, Qiang Zhang, Zixuan Jiang, and Renjing Xu. Multi-floor zero-shot object navigation policy. *arXiv preprint arXiv:2409.10906*, 2024.

[52] Vuong Vuong, Toan Nguyen, Minh Nhat Vu, Baoru Huang, H.T.T Binh, Thieu Vo, and Anh Nguyen. Habicrowd: A high performance simulator for crowd-aware visual navigation. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5821–5827, 2024.

[53] Zeying Gong, Tianshuai Hu, Ronghe Qiu, and Junwei Liang. From cognition to precognition: A future-aware framework for social navigation. *arXiv preprint arXiv:2409.13244*, 2024.

[54] Haresh Karnan, Anirudh Nair, Xuesu Xiao, Garrett Warnell, Sören Pirk, Alexander Toshev, Justin Hart, Joydeep Biswas, and Peter Stone. Socially compliant navigation dataset (scand): A large-scale dataset of demonstrations for social navigation. *IEEE Robotics and Automation Letters*, 7(4):11807–11814, 2022.

[55] Duc M. Nguyen, Mohammad Nazeri, Amirreza Payandeh, Aniket Datar, and Xuesu Xiao. Toward human-like social robot navigation: A large-scale, multi-modal, social human navigation dataset. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7442–7447, 2023.