2025 IEEE/SICE International Symposium on System Integration (SII)
January 21-24, 2025, Munich, Germany

LLM-Guided Zero-Shot Visual Object Navigation with Building Semantic Map

Jin Shi[1], Satoshi Yagi[1], Satoshi Yamamori[2,1], Jun Morimoto[1,2]

***Abstract*— This work presents a novel approach to zero-shot visual object goal navigation that leverages the ability of visual Large Language Model (vLLM) for finding target in unknown environment. Our system combines semantic mapping with vLLM-driven decision-making to direct robots towards target objects. The core of our approach lies in using vLLM to generate a value map between explored areas and the target object using cosine similarity based on prompt identically, incorporating both visual and semantic information from RGB-D image observations. This value map, along with a constructed semantic map and extracted movable frontier points, serves as a historic information for the vLLM to select one of the frontiers to explore next. We evaluate our method on two single-floor scenes from the Habitat-Matterport 3D dataset and Habitat Synthetic Scenes Dataset using the Habitat simulator separately. Our experiments demonstrate that the proposed approach has the potential to explore efficiently, particularly excelling when utilizing semantic information from simulator. The results show promise of our method in zero-shot navigation scenarios if overcome the common semantic extraction challenge. This work contributes to the growing field of language-driven exploration and exhibits how advanced large language model can effectively tackle complex navigation tasks.**

I. INTRODUCTION

The capacity for a robot navigating to objects in unknown environments represents a cornerstone challenge in embodied AI research, holding significant implications for a range of applications such as finding, catching or tracking objects, post-disaster rescue, even the interaction with human [1], [2]. Object Goal Navigation (ObjectNav) [3] represents a critical task in this domain, as shown in Fig. 1, where a robot situated in a random location in an unseen environment must navigate towards a target object by leveraging sensory inputs RGB-D images and pose information. Unlike straightforward PointGoal navigation, where the goal is a predefined coordinate, ObjectNav necessitates a deeper level of semantic comprehension from the environment as the object location information is missing. The robot must identify and maneuver towards the desired object based on object category label with inherent sensory information in previously unencountered environments. This intricate capability holds significant promise for practical applications, such as instructing domestic robots to retrieve specific items or guiding autonomous systems operating within unstructured settings.

[1]Jin Shi, Satoshi Yagi, Satoshi Yamamori, and Jun Morimoto are with the Graduate School of Informatics, Kyoto University, Kyoto, 606-8501, Japan.
[2] Satoshi Yamamori and Jun Morimoto are also with the Department of Brain Robot Interface, Computational Neuroscience Labs, ATR, Seika-cho, 619-0237, Japan. Email: morimoto@i.kyoto-u.ac.jp

Fig. 1: In a realistic environment within the Habitat simulator [11], the robot is placed in an arbitrary position, indicated as the start, and needs to navigate to an unseen room to find a specified target like bed, indicated as the goal.

To the best of our knowledge, prevalent ObjectNav techniques relied heavily on training deep neural networks with vast, meticulously labeled datasets by using reinforcement learning [4], [5], [6], [7]. This heavy data paradigm, while effective, often necessitates substantial computational resources and often encounters difficulties when generalizing to novel objects or environments not present in the training data. The arise of Large Language Model (LLM) presents a compelling possibility for overcoming these problems by harnessing the power of zero-shot learning. This innovative approach empowers robot to navigate to object without prior exposure to any training data specific to those particular objects and the environment [8], [9], [10].

This paper centers around the application of zero-shot method, grounded in the capabilities of visual LLM, to tackle the challenges of ObjectNav within the Habitat Simulator [11]. Our primary objective is to combine the semantic knowledge encoded within vLLM to construct strategies for navigating complex three dimensional environments to find target object.

II. RELATED WORKS

A. Object Goal Navigation

Recent years have witnessed substantial advancements in ObjectNav research [3], [6], [12], [13], largely fueled by breakthroughs in reinforcement learning and the proliferation of large-scale datasets like Matterport3D [14], Gibson [15], Habitat-Matterport3D (HM3D) [16] and Habitat Synthetic Scenes Dataset(HSSD) [17]. Early research primarily concentrated on training deep reinforcement learning (RL) agents through an end-to-end approach. A notable instance is DD-PPO [18], an algorithm that exhibited remarkable

979-8-3315-3161-4/25/$31.00 ©2025 IEEE

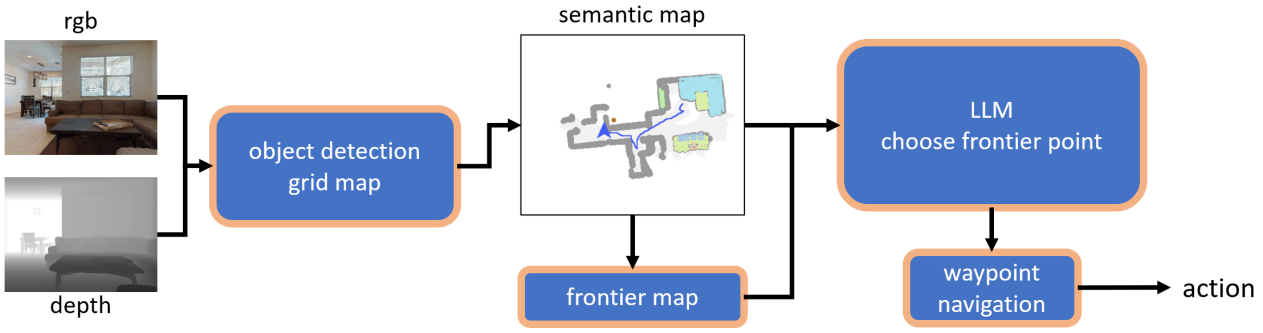


Fig. 2: The overview of the method structure. **RGB and depth images** $\mathbf{x}_{\text{rgb}} \in \mathbb{R}^{H\times W\times 3}$ and $\mathbf{x}_{\text{depth}} \in \mathbb{R}^{H\times W}$ where $(H = 480, W = 640)$, are processed into a **grid map** $\mathbf{x}_{\text{grid}} \in \mathbb{R}^{N\times M}$, which is converted into a **semantic map** $\mathbf{x}_{\text{semantic}} \in \mathbb{R}^{N\times M\times C}$ where $(N = 600, M = 600, C = 20)$, C represents the channel of map indicate obstacle map, object category, robot location etc, and each grid represents 5 centimeters. The **frontier map** $\mathbf{x}_{\text{frontier}} \in \mathbb{R}^{K}$ contains K detected exploration frontiers. The LLM selects a frontier point $\mathbf{x}_{\text{frontier\_selected}} \in \mathbb{R}^{2}$, which is then passed to the **waypoint navigation** system, outputting a path $\mathbf{x}_{\text{path}} \in \mathbb{R}^{T\times 2}$ and an action command $\mathbf{x}_{\text{action}} \in \mathbb{R}^{3}$.

results in simpler PointGoal navigation tasks. However, these monolithic RL agents often falter when confronted with the inherent semantic complexity of ObjectNav tasks.

B. Modular and Map-Based Methods

While end-to-end reinforcement learning approaches can achieve remarkable performance within their training domains, they demand millions of training episodes and substantial computational resources, yet fail to generalize effectively to different environments [4], [6], [7]. In an attempt to alleviate the limitations of end-to-end reinforcement learning approaches, researchers have explored modular methodologies, which deconstruct the ObjectNav task into distinct subtasks, such as object detection, mapping, exploration, and navigation [5], [13]. These methodologies often incorporate semantic mapping, enabling the agent to construct an environmental representation enriched with object labels, thereby facilitating more targeted exploration and streamlined goal localization.

A number of studies have demonstrated the efficacy of fusing classical navigation paradigms, such as Frontier-based Exploration (FBE) and path planning, with learned components [13], [19]. FBE promotes exploration by prompting the agent to venture into uncharted territory, effectively pinpointing the boundaries between explored and unexplored regions. Path planning algorithms then come into play, computing optimal trajectories to efficiently guide the agent towards the designated target locations. These hybrid methodologies strive to capitalize on the respective strengths of classical and learning-based approaches, aiming to deliver navigation solutions in high performance.

C. Zero-Shot Object Navigation

Zero-shot object navigation pushes the boundaries of ObjectNav, challenging agents to skillfully navigate to previously unencountered object categories without requiring any prior training data tailored to specific categories [8], [9], [20]. A prominent strategy for realizing this ambition involves harnessing the capabilities of pre-trained Vision-Language Models (VLMs), such as CLIP [21], BLIP-2 [22], which excel at mapping images to text description based on predefined prompt, so that the description of object and sensory image can be put into a shared embedding space. This shared representation empowers agents to decipher the intricate semantic connections between visual input and language-based goals, allowing them to successfully navigate towards unfamiliar objects solely based on their language descriptions [23].

D. LLMs for Object Navigation

LLMs have emerged as a transformative force in the realm of grounding language instructions to concrete actions, rendering them exceptionally well-suited for intricate tasks like ObjectNav [10], [24], [25]. The integration of LLMs into navigation pipelines allows agents to tap into a wealth of commonsense knowledge and make full use of advanced reasoning capabilities, ultimately culminating in significant improvements in exploration strategies and navigation performance [10], [26], [27]. For instance, LLMs can be strategically deployed to deduce the likely locations of objects based on their semantic associations with other objects or rooms detected within the environment. Additionally, incorporating LLMs into FBE methods empowers agents to prioritize frontiers that exhibit a higher probability of harboring the target object, relying on invaluable semantic cues [23].

This paper explored the application of vLLM to directly generate frontier points based on the constructed map image for robot navigation, focusing on evaluating vLLM's comprehension capabilities of constructed maps which containing historical trajectories and spatial information. By leveraging vLLM to understand and reason about the spatial context embedded in these maps, we investigate its potential to guide efficient robot naviagtion in unseen environments within the Habitat Simulator.

III. Approach

Fig. 2 presents an overview of our method. RGB and depth images captured by the robot are processed into a grid map, which is converted into semantic map of size $N \times M$ in C channels with information provided from object detection. From this, a frontier map with K detected exploration frontiers is generated. The LLM then selects a frontier point, which is passed to the way-point navigation system to output a path and action command for the movement of robot. Below, we explain the details of each process.

A. Grid Map with Semantic Information

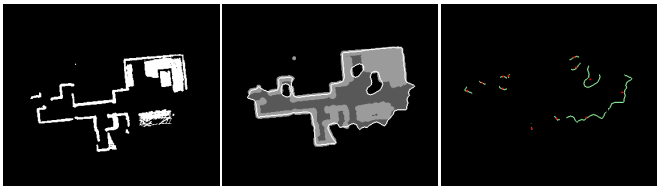
Our approach begins with the creation of a semantic grid map. This map is constructed by combining RGB-D observations and agent pose data over time. We process depth information to generate point clouds, which are then aligned in a global coordinate system using the agent's recorded poses as it contain position and orientation. The visual data is processed using segmentation and detection model Detic [28] to classify each point into one of several object categories which locate in the COCO classes [29]. The semantically enriched point cloud is projected onto a top-down view (Fig. 3), generating a multi-channel map that captures different aspects of the environment. This map consists of separate layers representing obstacles (Fig. 3a), explored areas (Fig. 3b), and distinct object categories. Together, these layers provide a comprehensive spatial and semantic representation of the environment that supports various sub-tasks, including vLLM-based reasoning and path planning.

B. Cost Value Map

Following [23], we implement a value mapping system for additional frontier point information. This system uses a 2D grid to quantify how relevant each explored area is to the target object. To extract semantic information from visual input, we employ the multimodal language model LLaVa [30]. Unlike [23], our approach processes a sequence of five images simultaneously through LLaVa, as it has a slower response time compared to BLIP-2. Using carefully crafted prompts, LLaVa generates a textual description T_{image_t} for these images. We then compute the cosine similarity between this generated text and a predefined target object description T_{target} - a LLaVa-generated sentence describing the expected environment of the target object. This similarity score is used to update the value map $V_t(x, y)$, with new exploration values overwriting previous ones in overlapping areas.

To ensure smooth robot movement, prevent cyclic behavior and explore the new area, we introduce a degenerative factor d to the similarity values when updating the value map each time, this mechanism mitigates the influence of older values, encouraging the robot to explore new areas rather than revisiting previously examined locations. This approach balances the need for thorough exploration with the efficiency of directed search, optimizing the robot's path towards the target object.

$$V_{t+1}(x, y) = cos\_sim(T_{image_t}, T_{target}) + (1 - d)V_t(x, y)$$



(a) (b) (c)

Fig. 3: Top-down view grid maps with semantic information. (a) represents the obstacle map generated from the grid map module, (b) shows the explored area which is a combination of movable area, obstacle area and the boundary is the frontier line, (c) shows the cross area of frontier line and movable area show as the green lines and then compute the center of the line as the frontier point show as red dot points.

C. Frontier Point Selection with LLM

In our implementation, we develop a sophisticated method for frontier point selection that integrates semantic information with traditional mapping techniques. We begin by generating a binary representation of the semantic map, distinguishing between traversable areas and obstacles within the explored region. This binary map serves as the foundation for identifying potential exploration targets.

To identify the boundary between explored and unexplored territories, we employ edge detection algorithm on the binary map by using erosion and corrosion. This process yields a set of frontier lines, as shown in Fig. 3c, each representing a potential avenue for further exploration. We then apply the moments method, a technique commonly used in computer vision, to determine the centroid of each frontier line. These centroids serve as candidate frontier points for the robot's next exploration goal.

The crux of our approach lies in the selection of the optimal frontier point from this set of candidates. Unlike previous methods that rely solely on geometric considerations or simplistic heuristics or reinforcement training [13], [19], [31], we leverage the power of vLLM to make this critical decision. Our system provides the vLLM with two key inputs: the value map, which encodes the semantic relevance of explored areas, and the constructed semantic grid map, which provides a high-level understanding of the environment's layout and content with historic and spatial information. By presenting this rich, multi-modal information to the vLLM, we enable it to reason about the exploration task in a manner that more closely mimics human decision-making. The vLLM analyzes the spatial distribution of semantic values, the configuration of obstacles and open spaces, and the potential information gain associated with each frontier point. This analysis allows the vLLM to identify the frontier point that offers the highest probability of leading the robot towards the target object.

This approach represents a significant departure from the other frontier-based exploration strategies. By integrating semantic understanding, spatial reasoning, and vLLM-based decision making, we create a more adaptive and context-aware exploration system. The ability of vLLM to process

and synthesize complex, multi-dimensional information allows our robot to make more informed choices about its exploration trajectory to locate target objects in unknown environments.

Moreover, this method's flexibility allows it to adapt to a wide range of environments and search objectives. As the semantic map and value map evolve during exploration, the vLLM can dynamically adjust its decision-making criteria, ensuring that the robot's exploration strategy remains optimal throughout the mission.

D. Point Goal Navigation

In the point goal navigation, we use the Fast Marching Method (FMM) [32] to compute optimal paths between the agent's current position and the target point chosen by LLaVa, taking into account obstacle information from the semantic map. This approach leverages existing path finding algorithms, reducing the need for learning basic navigation from scratch such as [33]. We then select the point adjacent to the robot's current position on this path and calculate the angle between it and the current position. Based on this angle, we determine the robot's next action: move forward, turn left, or turn right.

Compared to methods trained using deep reinforcement learning, this approach is compatible with any dataset and does not require retraining. However, its performance on the same dataset is not as perfect as [33] for sometimes this method may result in collisions with objects in the simulator preventing robot movement to cause the failure or simply can not calculate a reasonable path.

IV. Experimental Setups

Our investigation into object navigation employs the Habitat platform [11], using one of the expansive HM3D dataset [16] and one from the HSSD dataset [17]. Our experiment lies the Object Navigation (ObjectNav) challenge, a standardized task within the AI Habitat ecosystem. This task evaluates an agent's ability to locate a specified object category within an unfamiliar environment and the random location. For environmental perception, the simulator supplies an RGB-D camera, which captures both visual and depth data. The robot's movement capabilities are defined by a set of discrete actions: forward motion, turn left and turn right, and the option to terminate the episode when the robot arrives near in the range of object within 1 meter or the robot moves more than 500 steps.

For LLM setup, we use the library proposed by [34] with the model [30], for the calculation of the value map, the prompt for the sequence images is "Considering you are a robot with RGB-D sensor observing the home environment, describe the sequences of images", there is also a predefined prompt for the description of target object, "describe the possible environment around the target", when drive the LLM finish the core decision making component that "where to go", we then use the colored semantic map, calculated value map, generated frontier points as input with the prompt format "image: [images], points: [points], semantic info: [the info based on the color], Task: [Based on the provided colored image, value map, pixel coordinates, and semantic map information, please output a single point from the given coordinates. The selected point should be the one that best matches have big value and more potential near the target.], output format: Selected Point: (x, y) Reason: [response]" output the desired information.

Approach	HM3D		HSSD	
	SR(%)	SPL(%)	SR(%)	SPL(%)
Random Walking	0.0	0.0	-	-
Frontier Based Method [35]	23.7	12.3	-	-
XGX [7]	100	54.0	33.3	-
VLFM [23]	75.0	41.5	70.0	22.9
Ours using object detection	60.7	25.3	53.3	22.1
Ours with ground truth semantic	89.3	42.9	-	-

TABLE I: Habitat Results: Observe the improvement in Success Rate (SR) and SPL on both our approaches over the current zero-shot SOTA method [23] and learning based SOTA method [7], frontier based method and random walking.

As the current one dimensional map construction method can not recognize stairs and can not handle the go up and down scenario for stairs, we choose one of the typical environments from the HM3D dataset, which only has one floor and contains 28 episodes, and one from the HSSD dataset, which also has one floor with 30 episodes, with each of them having a different initial location and target object.

To assess performance, we focus on two key metrics, success rate (SR) and success weighted by path length (SPL) proposed by [11]. SR measures the proportion of successful object localization across all episodes, and SPL combines success with efficiency, comparing the agent's trajectory to the optimal route as:

$$\text{SPL} = \frac{1}{N}\sum_{i=1}^{N} S_i \frac{l_i}{\max(p_i, l_i)},$$

where N is the overall numbers of episode, l_i is the length of shortest path between goal and target in one episode that can calculated by the simulator, p_i is the overall length moved by the robot in one episode, s_i is the binary indicator of success in episode.

These evaluation criteria provide a comprehensive view of the agent's navigational proficiency and operational efficiency.

V. RESULT

Our experimental results on the Habitat simulator demonstrate the possibilities of our proposed method for zero-shot object goal navigation. As shown in Table I, our approach achieves competitive performance compared to existing methods.

The baseline Random Walking method, as expected, fails to complete any navigation tasks successfully, resulting in 0% for both Success Rate (SR) and Success weighted by Path Length (SPL). The Frontier Based Method [3] shows

some improvement, with an SR of 23.7% and SPL of 12.3%. Our proposed method without object detection achieves a success rate of 60.7% and an SPL of 25.3%, which is a significant improvement over the Frontier Based Method. This demonstrates the effectiveness of integrating vLLM for exploration guidance compared to pure random frontier selection. Notably, when augmented with ground truth semantic information, our method's performance increases substantially, reaching an SR of 89.3% and an SPL of 42.9%. This result outperforms the state-of-the-art VLFM method, which achieves an SR of 75.0% and an SPL of 41.5% in the same situation.

We also evaluated XGX [7], a state-of-the-art model with the combination of reinforcement learning and imitation learning. While XGX exhibited exceptional perfect performance on its training dataset, HM3D, certain limitations hindered its transferability to the HSSD dataset so that shows a relative lower success rate. Due to discrepancies in metric output, we were unable to compute the SPL for XGX on the HSSD dataset.

On the HSSD dataset, our approach achieved comparable SPL to VLFM, albeit with a lower success rate. This suggests that in successful episodes, our method can locate target objects with fewer actions, indicating improved efficiency in navigation when the task is completed successfully.

Overall, our method shows the potential direction in the challenging task of zero-shot object goal navigation, particularly when leveraging accurate semantic information.

VI. DISCUSSION

Our main contributions include:

- A novel integration of vLLM for value mapping and frontier selection in object goal navigation.
- A semantic map-based exploration strategy that effectively guides the robot in unknown environments.

While our results are promising, there are several limitations and areas for future work:

- Single-floor limitation: Currently, our way point navigation module is restricted to single-floor environments and cannot handle multi-story scenarios with stairs.
- Computational overhead: The reliance on LLMs introduces longer communication times compared to end-to-end methods, which may impact real-time performance.
- Dependency on object detection: The current structure heavily relies on the accuracy of object detection for generating the semantic map, which can be a potential point of failure in complex or ambiguous environments.
- Simulation-based validation: Our method has been tested extensively in simulation but has not yet been implemented on real robotic platforms.

The current approach relies heavily on object detection performance, which leads to several failure modes: agents stopping at incorrectly detected locations, exceeding maximum step limits when failing to detect objects, and failing to reach targets due to detection position offsets. To establish an upper bound for the method's performance, we conducted experiments using ground truth object information, the experimental results demonstrate the viability and potential of our proposed approach.

Our experimental results demonstrate the feasibility of using vLLM to directly interpret spatial information from constructed maps. While recent work has explored LLM applications in navigation, such as [26], [36] using LLMs for instruction following, our approach reveals new capabilities in spatial reasoning abilities of vLLM based on images.

Future work will address these limitations by extending the navigation capabilities to multi-floor environments, optimizing LLM integration for faster processing, implementing and testing the system on real robots, and improving the robustness of semantic mapping to handle imperfect object detection. Despite these challenges, our work represents a significant step towards more adaptable and intelligent robotic navigation systems. By using the multi-modal LLMs, our approach opens new avenues for developing versatile and capable robots that can efficiently navigate and interact in diverse, unknown environments.

VII. CONCLUSION

In this work, we introduced a novel approach that uses Large Language Models for zero-shot object goal navigation in unseen environment with the constructed semantic map as history information. Our method combines semantic mapping with LLM-driven decision-making to guide robots towards target objects efficiently. The experimental results demonstrate the potential of our approach, particularly when augmented with accurate semantic information, outperforming existing methods in terms of success rate and efficiency.

ACKNOWLEDGMENT

This work was supported by JST Moonshot R&D program JPMJMS223B-3, and JSPS KAKENHI 22H03669 and 22H04998.

REFERENCES

[1] J. Duan, S. Yu, H. L. Tan, H. Zhu, and C. Tan, "A survey of embodied ai: From simulators to research tasks," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 6, no. 2, pp. 230–244, 2022.

[2] J. Zhang, L. Dai, F. Meng, Q. Fan, X. Chen, K. Xu, and H. Wang, "3D-aware object goal navigation via simultaneous exploration and identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 6672–6682.

[3] D. Batra, A. Gokaslan, A. Kembhavi, O. Maksymets, R. Mottaghi, M. Savva, A. Toshev, and E. Wijmans, "Objectnav revisited: On evaluation of embodied agents navigating to objects," *arXiv preprint arXiv:2006.13171*, 2020.

[4] T. Chen, S. Gupta, and A. Gupta, "Learning exploration policies for navigation," in *7th International Conference on Learning Representations*, 2019.

[5] 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.

[6] R. Ramrakhya, D. Batra, E. Wijmans, and A. Das, "Pirlnav: Pretraining with imitation and rl finetuning for objectnav," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 17 896–17 906.

[7] J. Wasserman, G. Chowdhary, A. Gupta, and U. Jain, "Exploitation-guided exploration for semantic embodied navigation," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 2901–2908.

[8] S. Y. Gadre, M. Wortsman, G. Ilharco, L. Schmidt, and S. 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*, 2023, pp. 23 171–23 181.

[9] A. Majumdar, G. Aggarwal, B. Devnani, J. Hoffman, and D. Batra, "ZSON: Zero-shot object-goal navigation using multimodal goal embeddings," *Advances in Neural Information Processing Systems*, vol. 35, pp. 32 340–32 352, 2022.

[10] V. S. Dorbala, J. F. M. Jr., and D. Manocha, "Can an embodied agent find your "cat-shaped mug"? llm-based zero-shot object navigation," *IEEE Robotics and Automation Letters*, vol. 9, no. 5, pp. 4083–4090, 2024.

[11] 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.

[12] H. Luo, A. Yue, Z.-W. Hong, and P. Agrawal, "Stubborn: A strong baseline for indoor object navigation," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2022, pp. 3287–3293.

[13] S. K. Ramakrishnan, D. S. Chaplot, Z. Al-Halah, J. Malik, and K. Grauman, "PONI: Potential functions for objectgoal navigation with interaction-free learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 890–18 900.

[14] 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*, 2017, pp. 667–676.

[15] F. Xia, A. R. Zamir, Z. He, A. Sax, J. Malik, and S. Savarese, "Gibson env: Real-world perception for embodied agents," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 9068–9079.

[16] S. K. Ramakrishnan, A. Gokaslan, E. Wijmans, O. Maksymets, A. Clegg, J. M. Turner, E. Undersander, W. Galuba, A. Westbury, A. X. Chang, M. Savva, Y. Zhao, and D. Batra, "Habitat-matterport 3d dataset (HM3D): 1000 large-scale 3d environments for embodied AI," in *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1*, 2021.

[17] M. Khanna, Y. Mao, H. Jiang, S. Haresh, B. Shacklett, D. Batra, A. Clegg, E. Undersander, A. X. Chang, and M. 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*, 2024, pp. 16 384–16 393.

[18] 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*, 2020.

[19] B. Yu, H. Kasaei, and M. Cao, "Frontier semantic exploration for visual target navigation," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 4099–4105.

[20] L. Zhang, Q. Zhang, H. Wang, E. Xiao, Z. Jiang, H. Chen, and R. Xu, "Trihelper: Zero-shot object navigation with dynamic assistance," *arXiv preprint arXiv:2403.15223*, 2024.

[21] 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*, 2021, pp. 8748–8763.

[22] 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.

[23] N. Yokoyama, S. Ha, D. Batra, J. Wang, and B. Bucher, "VLFM: Vision-language frontier maps for zero-shot semantic navigation," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 42–48.

[24] J. Lin, H. Gao, X. Feng, R. Xu, C. Wang, M. Zhang, L. Guo, and S. Xu, "Advances in embodied navigation using large language models: A survey," 2024. [Online]. Available: https://arxiv.org/abs/2311.00530

[25] C. Huang, O. Mees, A. Zeng, and W. Burgard, "Visual language maps for robot navigation," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 10 608–10 615.

[26] B. Yu, H. Kasaei, and M. Cao, "L3mvn: Leveraging large language models for visual target navigation," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023, pp. 3554–3560.

[27] I. G. Maglogiannis, L. S. Iliadis, J. Macintyre, M. Avlonitis, and A. Papaleonidas, *Artificial Intelligence Applications and Innovations: 20th IFIP WG 12.5 International Conference, AIAI 2024, Corfu, Greece, June 27-30, 2024, Proceedings, Part IV*. Springer Nature, 2024, vol. 714.

[28] X. Zhou, R. Girdhar, A. Joulin, P. Krähenbühl, and I. Misra, "Detecting twenty-thousand classes using image-level supervision," in *European Conference on Computer Vision*. Springer, 2022, pp. 350–368.

[29] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: common objects in context," in *Computer Vision–ECCV 2014: 13th European Conference*, 2014, pp. 740–755.

[30] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[31] J. Wasserman, K. Yadav, G. Chowdhary, A. Gupta, and U. Jain, "Last-mile embodied visual navigation," in *Conference on Robot Learning*, 2023, pp. 666–678.

[32] A. Valero-Gomez, J. V. Gomez, S. Garrido, and L. Moreno, "The path to efficiency: Fast marching method for safer, more efficient mobile robot trajectories," *IEEE Robotics & Automation Magazine*, vol. 20, no. 4, pp. 111–120, 2013.

[33] E. Wijmans, I. Essa, and D. Batra, "VER: Scaling on-policy rl leads to the emergence of navigation in embodied rearrangement," *Advances in Neural Information Processing Systems*, vol. 35, pp. 7727–7740, 2022.

[34] T. Wolf, "Huggingface's transformers: State-of-the-art natural language processing," *arXiv preprint arXiv:1910.03771*, 2019.

[35] 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'*, 1997, pp. 146–151.

[36] W. Chen, S. Hu, R. Talak, and L. Carlone, "Leveraging large (visual) language models for robot 3d scene understanding," 2023. [Online]. Available: https://arxiv.org/abs/2209.05629