ViLaBot: Connecting Vision and Language for Robots that Assist Humans at Home

Asfand Yaar
DMI, University of Catania
Catania, Italy
asfand.yaar@studium.unict.it

Marco Rosano
DMI, University of Catania
Catania, Italy
marco.rosano@unict.it

Antonino Furnari
DMI, University of Catania
Catania, Italy
antonino.furnari@unict.it

Aki Härmä
DACS, Maastricht University
Maastricht, The Netherlands
aki.harma@maastrichtuniversity.nl

Giovanni Maria Farinella
DMI, University of Catania
Catania, Italy
giovanni.farinella@unict.it

***Abstract*—Despite significant advancements in the field of vision, language and robotics, integrating these capabilities to create an autonomous robot assistant remains a challenge. This paper presents ViLaBot (Vision and Language roBot), a system designed to aid humans in daily activities while at home. ViLaBot combines a language model with a library of basic visuomotor skills to understand human needs, create action plans and execute them. The system relies solely on onboard visual and proprioceptive sensing, eliminating the need for pre-built maps or precise object locations and facilitating real-world deployment in a variety of environments. Experimental validation conducted in 11 realistic home environments featuring simulated human agents using the Habitat simulator indicated that ViLaBot can achieve promising results when using ground-truth image segmentation, yet exhibits inferior performance in scenarios involving imperfect visual perception. The results support the validity of the proposed pipeline and highlight the critical components of the system that should be improved to increase its overall success rate and reliability.**

***Index Terms*—human-robot interaction, assistive tasks, task planning, navigation and manipulation.**

I. Introduction

The aspiration to create robotic household assistants has been a major driving force for researchers in the field of robotics, shaping their work across various domains. This pursuit involves advancements in visual perception, refining manipulation techniques, and addressing increasingly complex challenges. A proficient household assistant must be able to interact seamlessly with humans through both visual and verbal communication, possess a comprehensive understanding of diverse objects, navigate its surroundings adeptly, and make intelligent decisions in environments with limited sensory input. An example of such a complex task is shown in Fig. 1. This multifaceted objective has spurred research in areas such as language understanding, navigation, and task and motion planning. However, previous efforts have often considered simplified scenarios, focusing on predetermined tasks, pre-built detailed maps, and known precise object locations [1], [2]. To cope with a more realistic setting, we present Vision and Language roBot (ViLaBot), a novel approach designed to connect human language with actions, behaviors, and objects in interactive home environments. Our approach enables natural language instructions to be translated into sequences of actions and interactions within realistic simulation environments. ViLaBot relies only on onboard visual and proprioceptive (joints and gripper state) sensing, removing the need for pre-built maps or object locations and simplifying its deployment in real-world scenarios. By simulating real-world challenges encountered when translating human language into robot actions for assistive tasks, we provide a comprehensive evaluation assessing the effectiveness of robotic systems in home environments.

We base our experiments on HomeRobot [3], a framework for the development of robotic systems which leverages the Habitat Synthetic Scenes Dataset (HSSD) [4], comprising 49 interactive 3D scenes created by humans within the Habitat simulator [5]. To reproduce assistive scenarios, we expanded the existing dataset by including 3D human agents engaged in activities in different settings such as kitchens and TV lounges. The results show that the proposed approach achieves promising performance in aiding humans in complex home environments. In summary, the contributions of our work are as follows:

- We propose an assistive robotic system for real and practical scenarios, that understands human needs starting from natural language, elaborates an action plan and executes it, without having a prior knowledge of the environment's layout or privileged information on the location of objects. Our approach uses a language model and a library of visuomotor skills to generate objectives and accurately execute tasks as requested by the human.
- We propose a novel natural language dataset comprising natural language queries paired with corresponding action sequences for the robot to enact.
- We extended the HSSD [4] dataset to include 3D human agents at various locations. These human agents are aligned with different activities in areas such as the

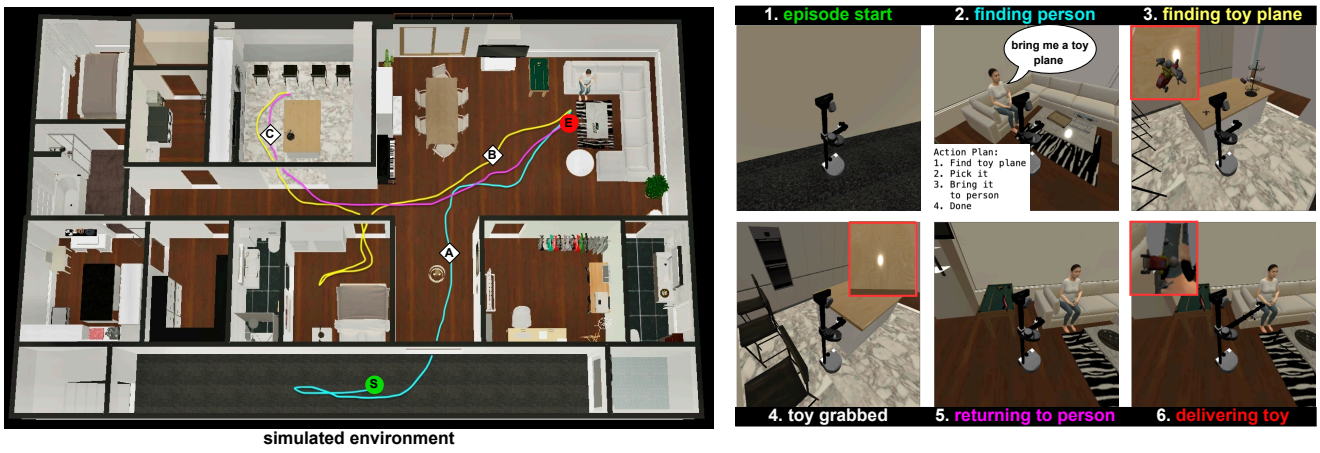

Fig. 1. The robot starts from point *S* and explores the home environment to locate and approach the human at a safe distance. Trajectory *A* illustrates its path towards the human. Once the human is reached, a task is assigned to the robot using a natural language query. Subsequently, an action plan is generated using a language model. Following the action plan, the robot explores the environment to locate the toy plane as shown by trajectory *B*. Once the toy is grabbed, it returns to the human, depicted by trajectory *C*. The point *E* indicate the end of the episode. Images $1-6$ present the third-person view at the end of each subtask.

kitchen, TV lounge, and other relevant areas that were previously absent from the original dataset.
- We show that the considered approach exhibits promising performance in navigating complex 3D environments and assisting humans with various household tasks.

II. Related Work

A. Task Planning

Task planning in robotics typically involves employing search algorithms within a predefined domain [6], [7]. However, this approach faces scalability challenges in environments with numerous feasible actions and objects due to large branching factors [8], [9]. To mitigate this, heuristics are commonly utilized to guide the search process [10]–[12]. Recent research has explored learning-based approaches to task and motion planning, incorporating techniques such as hierarchical learning, language-based planning spaces, representation learning, and learning compositional skills [13]–[17] In contrast, our method bypasses traditional search methods by directly generating action plans through language model that incorporate conditional reasoning.

B. Exploration

Agents navigating and manipulating the embodied world must keep track of both the environment [18] and their position [19]. These aspects have been extensively explored in robotics, involving the processing of low-level information [20], the construction of semantic maps [21], and more recently, the development of techniques tailored to managing dynamic and general aspects of the environment [22], [23]. In scenarios, such as in embodied learning tasks, recent methods have investigated neural network-based maps [24]–[26]. Our approach builds on these methods and incorporates the use of a pre-trained semantic segmentation model, following a similar methodological setup as [27], [28].

C. Rearrangement

The rearrangement task has been a long-standing focus in robotics research, with numerous studies addressing this fundamental task [29]–[31]. Typically, these approaches address the challenge in the context of fully observing object states [32], [33], enabling efficient and accurate planning-based solutions. However, a recent trend has emerged in the field of visual rearrangement [34]–[36], where the states of the objects and the goal of the rearrangement are not directly observable. In these cases, the agent receives direct visual input and the environment is relatively complex and realistic. These rearrangement approaches share similarities with other challenging tasks of embedded AI, such as embedded navigation [37]–[40] and answers to embodied questions [41], [42], which require finding objects and reasoning about their states.

D. Vision and Language Navigation

In vision and language navigation, one can use either a natural language or language template to describe a path leading to a specific goal based on egocentric visual observations [43]–[45]. Since the introduction of Room-to-Room (R2R) [44], a dataset for visually-grounded natural language navigation in real environments, researchers have achieved significant advances in navigation performance [46], [47] with the incorporation of task variations with additional on-route instructions [48], [49]. However, much of this progress has been limited to static environments. In contrast, the tasks associated with our approach encompass navigation, object interactions, and state changes. While there are existing methods that rely on simple block worlds and fully observable scenes [50], [51], our approach presents more challenging tasks using visually complex scenes. Unlike [52], which evaluates agents executing house instructions using a generic *interact* action, our method incorporates multiple navigation and manipulation actions like

Query type	Example
Structured language queries	Move the knife to the counter
Queries focused on abstract nouns	Bring me an Apple
Queries in unstructured formats	I want to read a book, can you bring me one?
Queries where no explicit object is mentioned	I am feeling hungry, bring me something sweet.

(a)

Query: I need a bottle opener to open this soda.
Action Plan:
1. Find bottle_opener
2. Pick it
3. Bring it to person
4. Done

Query: Move the backpack to the couch.
Action Plan:
1. Find backpack
2. Pick it
3. Find couch
4. Place it
5. Done

(b)

Fig. 2. (a) Types of query with an example. (b) Query examples along with the action plan generated by GPT3

find, *pick*, *place* and introduces variations in both language and visual complexity. Furthermore, there are related works that leverage language models for code-base task planning [53].

The natural language processing community has a substantial body of literature on following language-based instructions. In this context, research has mainly focused on mapping instructions to actions [43], [54], [55]. However, these efforts do not involve visual or interactive environments. Our approach uses a language model along with a navigation module to generate objectives and accurately execute tasks as per the human's request.

III. Problem Definition

We aim to create a system able to navigate complex environments and assist humans in various tasks. The system is evaluated following an episode-based approach, inspired by previous works on navigation with Reinforcement Learning (RL) [3]. At the beginning of each episode, the agent is initialized at a random location within the environment and receives two images as input: a depth image containing spatial information about the environment; an image containing semantic segmentation of objects that appear in the scene. Additionally, the agent is provided with sensor pose readings x_t, representing the robot's pose relative to its starting position, and joint sensor readings y_t indicating the states of the camera and arm joints. The initial goal of the agent is to locate and reach the human at a safe distance. Once the agent reaches the human, a task is assigned by the human in the form of a natural language query (e.g., *"I need a bottle opener to open this soda"*). This query is then processed by a language model to generate an action plan (see Fig. 2b). Subsequently, the robot utilizes its visuomotor skills library to execute the action plan and assist the human.

Throughout the episode, a visuomotor skill predicts an action a_t at each time step *t* according to a given subtask. The set of actions include *move_forward*, *turn_right*, *turn_left*, and *stop*. An episode concludes either when the agent executes the entire action plan or when it reaches the maximum limit of 2050 steps. To evaluate the efficacy of ViLaBot, we present the Success Rate (SR) for each skill independently, along with the partial SR and the overall SR. The partial SR denotes the average across all skills, while the overall SR signifies the percentage of successfully executed action plans. We devised four task variants to encompass diverse scenarios with varying levels of complexity. The details of each version are given following:

- **V1.1** After reaching the person and receiving the natural language query, the agent has to find the requested object and move it to a specified location as requested by the human. To execute the action plan, the robot is provided with privileged information about the general location of the goal object, such as *the backpack being on a chair*. However, the specific chair is not disclosed, requiring the robot to locate the right chair and proceed with the rest of the action plan accordingly.
- **V1.2** It is similar to $V1.1$, but we removed the privileged information about the goal object. This adds more complexity to the task, especially in detecting smaller objects from a distance.
- **V2.1** Unlike $V1$, the robot delivers the requested object to the human rather than relocating it elsewhere. As in $V1.1$, the robot receives privileged information about the general location of the goal object. It is important to note that the human may not be in the same location as observed at the start of the episode, requiring the robot to locate the human again.
- **V2.2** It shares similarities with $V2.1$ but introduces more complexity to the task by removing privileged information regarding the goal object location.

IV. Proposed Method

The proposed approach relies on a language model and library of visuomotor skills to efficiently navigate complex indoor environments and support humans for various tasks, as illustrated in Fig. 3. The details are as follows:

A. Language Model

The language module assists in processing the natural language queries from a person to generate actionable plans. We accommodate various query types, as illustrated in Fig. 2a. Contrary to traditional search methods, our approach directly generates action plans through a language model by integrating conditional reasoning. We conceptualize task planning as a tuple (O, A, R, E, Q) consisting of sets of objects O, executable actions A, receptacles available in the environment R, example queries with action plan E, and final query Q. For example,

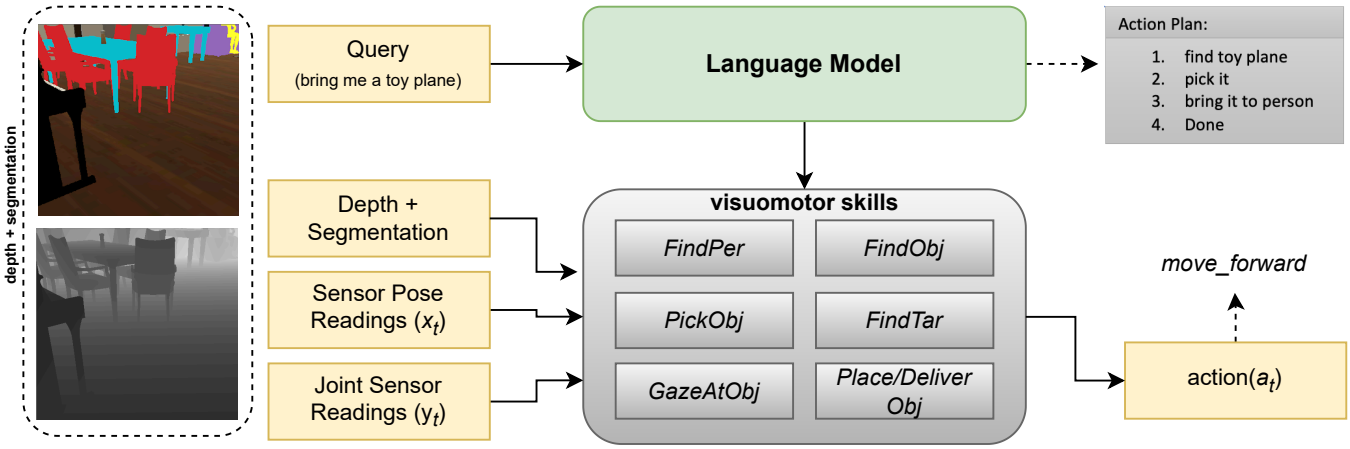

Fig. 3. Outline of the proposed approach. Given a natural language query, the language model generates an action plan. The robot then executes that action plan using visuomotor skills.

consider the task *bring me an apple*. Our approach utilizes a GPT3 model to interpret the query and generate an action plan (see Fig. 2b) using a predefined set of actions the robot can execute. This approach facilitates efficient and context-aware task planning without relying on exhaustive search algorithms.

B. Visuomotor Skills

The visuomotor skills library consists of essential skills including *FindPer*, *FindObj*, *GazeAtObj*, *PickObj*, *FindTar* and *PlaceObj*, all integral for assisting the human.

FindPer: this policy directs the agent to find a person in an unseen environment. The input observation space includes data from the robot's head camera depth, semantic segmentation, the robot's pose relative to its starting position, and joint sensor readings indicating the states of camera and arm joints. The policy produces actions for translation and rotation, along with a discrete *stop* action. Success for this skill is achieved when the agent triggers the *stop* action upon reaching within 1 meter of the human while also facing them.

FindObj: the policy is used to help the agent in finding a goal object. The agent explores the environment until it reaches the goal object. In addition to input observation space, we include the CLIP [56] embedding of the goal object and the category of the receptacle on which the object can be found. We provide two channels, one displaying all instances of goal object and the other showing all instances of candidate start receptacle. Success for this skill is achieved when the agent triggers the *stop* action upon reaching within a distance of 0.1 meters from one of the viewpoints of a start receptacle while facing the goal object.

GazeAtObj: the gaze skill assists the agent in making final adjustments before grasping the object. These refinements include positioning the agent within arm's length of the object and ensuring that the object is both centered and visible. The robot moves close to the object and fine-tunes its head tilt until the object is both nearby and centered. Success is determined by the center pixel of the camera aligning with a goal object and the agent's base being within 0.8 meters of the goal object.

PickObj: once the robot reached the correct position, the *PickObj* skill activates the gripper to grasp the object. Due to constraints within the simulator, the process of simulating the physical grasping action is not achievable but the object is instantly attached to the agent's gripper. The approximation, while present, poses no significant limitation when deploying on a real robot, as the agent can leverage diverse specialized methods for grasping, provided that the object remains clearly visible and within arm's reach.

FindTar: we use this policy to find the candidate target receptacle on which the robot can place the goal object. This policy is similar to *FindObj* but we only pass CLIP [56] embedding of the target receptacle category into the input observation space as a single channel that displays all instances of candidate target receptacles. Success for this skill is achieved when the agent triggers the *stop* action upon reaching within 0.1 meters from one of the viewpoints of a target receptacle.

PlaceObj: the agent uses its arm to accurately place the goal object, when near the target receptacle. It approaches the receptacle, opens its gripper, and proceeds to place the object onto the specified surface. However, when the robot delivers an item to a person, it extends its arm toward the person to offer the item for collection. Success is achieved if the agent extends its arm to release the object onto the target receptacle or offer the object to the person while avoiding any collision.

V. Experiments and Results

We utilized HomeRobot [3] framework in the Habitat simulator [5] to train five distinct RL policies (*FindPer*, *FindObj*, *GazeAtObj*, *PickObj*,*FindTar*, *PlaceObj*) using DDPPO [57] with a slack reward of -0.005 per step, encouraging task completion with the fewest steps possible. These policies predict actions given depth, ground truth semantic segmentation, the robot's pose relative to its starting position, and proprioceptive sensors (such as joints and gripper state). The action space comprised four possible actions: *move_forward* ($0.25cm$), *turn_right* ($10degrees$), *turn_left* ($10degrees$), and

TABLE I
INDIVIDUAL AND OVERALL SR (IN %) FOR DIFFERENT TASKS WITH GROUND-TRUTH AND PREDICTED VISUAL SEMANTIC SEGMENTATION. THE OVERALL SR COINCIDES WITH THE SR OF THE PLACE/DELIVER SKILL.

Segmentation	Task	FindPer	FindObj	PickObj	FindTar	Overall Success	Partial Success
groundtruth	V1.1	78.67	49.77	44.40	37.39	31.66	57.20
	V1.2	78.67	44.22	40.12	33.66	23.08	43.95
	V2.1	78.67	49.77	44.40	33.75	28.84	47.08
	V2.2	78.67	44.22	40.12	30.66	26.02	43.93
DETIC [58]	V1.1	70.52	19.75	9.10	6.91	3.58	21.97
	V1.2	70.52	15.83	8.73	6.82	3.33	21.04
	V2.1	70.52	19.75	9.10	7.18	6.36	22.58
	V2.2	70.52	15.83	8.73	6.67	5.88	21.52

stop. We utilized reduced-resolution images of 160×120 to expedite training compared to the original 640×480 resolution. To measure the impact of an imperfect visual perception, we used both ground-truth segmentation as well as semantic segmentation predicted by DETIC [58]. We assessed the proposed approach across 11 modified environments from the HSSD [4] dataset that were not seen during training. This allowed us to assess the generalization ability of the learned policies to previously unseen environments.

We run a total of 1100 evaluation episodes, each scene comprising 100 episodes. Table I presents the quantitative results for all task versions. We report the SR of each skill individually, as well as the overall SR. Additionally, we include the partial SR, which represents the average success in all skills. Our proposed approach achieves an overall SR of 31.66% and 28.84% for $V1$ and $V2$, respectively. However, when DETIC [58] segmentation is used, the SR drops to 3.58% and 6.36% for $V1$ and $V2$, respectively. This highlights the increased difficulty of tasks when an imperfect segmentation is available, as segmentation predictors often struggle to accurately segment smaller objects. Analyzing the performance of the different tasks separately, we can notice that specifically for the *FindPer* and *FindObj* tasks there is a significant room for improvement (SR of 78.67% and 49.77% for the *FindPer* and *FindObj* with ground-truth segmentation, respectively). This result suggests that further effort should be devoted to improving these sub-tasks, which are critical for the effective deployment of assistive robots. While our approach faces challenges in certain task variations, particularly when perfect segmentation is unavailable, it still holds promising potential in aiding humans with daily tasks in indoor environments.

VI. CONCLUSIONS

In this paper, we propose an assistive robotic system capable of understanding human needs expressed through natural language, formulating an action plan, and executing it without prior knowledge of the environment's layout or privileged information about the location of objects. Despite challenges, our approach demonstrates promising potential in assisting humans with daily tasks in home environments, while highlighting the critical components of the system that should be improved to increase its overall success rate and reliability. As we refine our methods and adapt to diverse scenarios, we anticipate further enhancements in the robot's performance and ability to seamlessly collaborate with humans in real-world environments. In the future, we plan to test our method in real-world lab settings.

ACKNOWLEDGMENT

This research has been supported by Marie Skłodowska-Curie Innovative Training Networks - European Industrial Doctorates - PhilHumans Project, European Union - Grant agreement 812882 (https://www.philhumans.eu), and by the project Future Artificial Intelligence Research (FAIR) – PNRR MUR Cod. PE0000013 - CUP: E63C22001940006.

REFERENCES

[1] A. Heins, M. Jakob, and A. P. Schoellig, "Mobile manipulation in unknown environments with differential inverse kinematics control," in *2021 18th Conference on Robots and Vision (CRV)*, 2021.

[2] C. R. Garrett, C. Paxton, T. Lozano-Pérez, L. P. Kaelbling, and D. Fox, "Online replanning in belief space for partially observable task and motion problems," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 5678–5684.

[3] S. Yenamandra, A. Ramachandran, K. Yadav, A. Wang, M. Khanna, T. Gervet, T.-Y. Yang, V. Jain, A. W. Clegg, J. Turner *et al.*, "Homerobot: Open-vocabulary mobile manipulation," *arXiv preprint arXiv:2306.11565*, 2023.

[4] 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," *arXiv preprint*, 2023.

[5] 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.

[6] Y. Jiang, S. Zhang, P. Khandelwal, and P. Stone, "Task planning in robotics: an empirical comparison of pddl-based and asp-based systems," *arXiv preprint arXiv:1804.08229*, 2018.

[7] C. R. Garrett, T. Lozano-Pérez, and L. P. Kaelbling, "Pddlstream: Integrating symbolic planners and blackbox samplers via optimistic adaptive planning," in *Proceedings of the International Conference on Automated Planning and Scheduling*, vol. 30, 2020, pp. 440–448.

[8] M. Shridhar, J. Thomason, D. Gordon, Y. Bisk, W. Han, R. Mottaghi, L. Zettlemoyer, and D. Fox, "Alfred: A benchmark for interpreting grounded instructions for everyday tasks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 740–10 749.

[9] X. Puig, K. Ra, M. Boben, J. Li, T. Wang, S. Fidler, and A. Torralba, "Virtualhome: Simulating household activities via programs," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8494–8502.

[10] J. Hoffmann, "Ff: The fast-forward planning system," *AI magazine*, vol. 22, no. 3, pp. 57–57, 2001.

[11] J. A. Baier, F. Bacchus, and S. A. McIlraith, "A heuristic search approach to planning with temporally extended preferences," *Artificial Intelligence*, vol. 173, no. 5-6, pp. 593–618, 2009.

[12] M. Helmert, "The fast downward planning system," *Journal of Artificial Intelligence Research*, vol. 26, pp. 191–246, 2006.

[13] P. Sharma, A. Torralba, and J. Andreas, "Skill induction and planning with latent language," *arXiv preprint arXiv:2110.01517*, 2021.

[14] Y. Zhu, J. Tremblay, S. Birchfield, and Y. Zhu, "Hierarchical planning for long-horizon manipulation with geometric and symbolic scene graphs," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 6541–6548.

[15] A. Akakzia, C. Colas, P.-Y. Oudeyer, M. Chetouani, and O. Sigaud, "Grounding language to autonomously-acquired skills via goal generation," *arXiv preprint arXiv:2006.07185*, 2020.

[16] B. Eysenbach, R. R. Salakhutdinov, and S. Levine, "Search on the replay buffer: Bridging planning and reinforcement learning," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[17] Y. Jiang, S. S. Gu, K. P. Murphy, and C. Finn, "Language as an abstraction for hierarchical deep reinforcement learning," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[18] S. Thrun *et al.*, "Robotic mapping: A survey," 2002.

[19] S. Thrun, D. Fox, W. Burgard, and F. Dellaert, "Robust monte carlo localization for mobile robots," *Artificial intelligence*, vol. 128, no. 1-2, pp. 99–141, 2001.

[20] J. Engel, T. Schöps, and D. Cremers, "Lsd-slam: Large-scale direct monocular slam," in *European conference on computer vision*. Springer, 2014, pp. 834–849.

[21] B. Kuipers and Y.-T. Byun, "A robot exploration and mapping strategy based on a semantic hierarchy of spatial representations," *Robotics and autonomous systems*, vol. 8, no. 1-2, pp. 47–63, 1991.

[22] Y.-S. Wong, C. Li, M. Niessner, and N. J. Mitra, "Rigidfusion: Rgb-d scene reconstruction with rigidly-moving objects," in *Computer Graphics Forum*, vol. 40, no. 2. Wiley Online Library, 2021.

[23] A. Rosinol, A. Violette, M. Abate, N. Hughes, Y. Chang, J. Shi, A. Gupta, and L. Carlone, "Kimera: From slam to spatial perception with 3d dynamic scene graphs," *The International Journal of Robotics Research*, vol. 40, no. 12-14, pp. 1510–1546, 2021.

[24] S. Gupta, J. Davidson, S. Levine, R. Sukthankar, and J. Malik, "Cognitive mapping and planning for visual navigation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2616–2625.

[25] D. S. Chaplot, R. Salakhutdinov, A. Gupta, and S. Gupta, "Neural topological slam for visual navigation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 12 875–12 884.

[26] K. Chen, J. P. De Vicente, G. Sepulveda, F. Xia, A. Soto, M. Vázquez, and S. Savarese, "A behavioral approach to visual navigation with graph localization networks," *arXiv preprint arXiv:1903.00445*, 2019.

[27] V. Blukis, C. Paxton, D. Fox, A. Garg, and Y. Artzi, "A persistent spatial semantic representation for high-level natural language instruction execution," in *Conference on Robot Learning*. PMLR, 2022, pp. 706–717.

[28] D. S. Chaplot, M. Dalal, S. Gupta, J. Malik, and R. R. Salakhutdinov, "Seal: Self-supervised embodied active learning using exploration and 3d consistency," *Advances in neural information processing systems*, vol. 34, pp. 13 086–13 098, 2021.

[29] M. Stilman, J.-U. Schamburek, J. Kuffner, and T. Asfour, "Manipulation planning among movable obstacles," in *Proceedings 2007 IEEE international conference on robotics and automation*, 2007.

[30] W. Yuan, J. A. Stork, D. Kragic, M. Y. Wang, and K. Hang, "Rearrangement with nonprehensile manipulation using deep reinforcement learning," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 270–277.

[31] Y. Labbé, S. Zagoruyko, I. Kalevatykh, I. Laptev, J. Carpentier, M. Aubry, and J. Sivic, "Monte-carlo tree search for efficient visually guided rearrangement planning," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 3715–3722, 2020.

[32] J. E. King, M. Cognetti, and S. S. Srinivasa, "Rearrangement planning using object-centric and robot-centric action spaces," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2016, pp. 3940–3947.

[33] A. Cosgun, T. Hermans, V. Emeli, and M. Stilman, "Push planning for object placement on cluttered table surfaces," in *2011 IEEE/RSJ international conference on intelligent robots and systems*. IEEE, 2011, pp. 4627–4632.

[34] L. Weihs, M. Deitke, A. Kembhavi, and R. Mottaghi, "Visual room rearrangement," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 5922–5931.

[35] D. Batra, A. X. Chang, S. Chernova, A. J. Davison, J. Deng, V. Koltun, S. Levine, J. Malik, I. Mordatch, R. Mottaghi *et al.*, "Rearrangement: A challenge for embodied ai," *arXiv preprint arXiv:2011.01975*, 2020.

[36] A. H. Qureshi, A. Mousavian, C. Paxton, M. C. Yip, and D. Fox, "Nerp: Neural rearrangement planning for unknown objects," *arXiv preprint arXiv:2106.01352*, 2021.

[37] 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.

[38] P. Anderson, A. Chang, D. S. Chaplot, A. Dosovitskiy, S. Gupta, V. Koltun, J. Kosecka, J. Malik, R. Mottaghi, M. Savva *et al.*, "On evaluation of embodied navigation agents," *arXiv preprint arXiv:1807.06757*, 2018.

[39] J. Francis, N. Kitamura, F. Labelle, X. Lu, I. Navarro, and J. Oh, "Core challenges in embodied vision-language planning," *Journal of Artificial Intelligence Research*, vol. 74, pp. 459–515, 2022.

[40] 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.

[41] A. Das, S. Datta, G. Gkioxari, S. Lee, D. Parikh, and D. Batra, "Embodied question answering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1–10.

[42] D. Gordon, A. Kembhavi, M. Rastegari, J. Redmon, D. Fox, and A. Farhadi, "Iqa: Visual question answering in interactive environments," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4089–4098.

[43] D. Chen and R. Mooney, "Learning to interpret natural language navigation instructions from observations," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 25, no. 1, 2011.

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

[45] M. MacMahon, B. Stankiewicz, and B. Kuipers, "Walk the talk: Connecting language, knowledge, and action in route instructions," *Def*, vol. 2, no. 6, p. 4, 2006.

[46] X. Wang, W. Xiong, H. Wang, and W. Y. Wang, "Look before you leap: Bridging model-free and model-based reinforcement learning for planned-ahead vision-and-language navigation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 37–53.

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

[48] J. Thomason, M. Murray, M. Cakmak, and L. Zettlemoyer, "Vision-and-dialog navigation," in *Conference on Robot Learning*. PMLR, 2020, pp. 394–406.

[49] K. Nguyen, D. Dey, C. Brockett, and B. Dolan, "Vision-based navigation with language-based assistance via imitation learning with indirect intervention," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12 527–12 537.

[50] D. Misra, J. Langford, and Y. Artzi, "Mapping instructions and visual observations to actions with reinforcement learning," *arXiv preprint arXiv:1704.08795*, 2017.

[51] Y. Bisk, D. Yuret, and D. Marcu, "Natural language communication with robots," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016, pp. 751–761.

[52] D. Misra, A. Bennett, V. Blukis, E. Niklasson, M. Shatkhin, and Y. Artzi, "Mapping instructions to actions in 3d environments with visual goal prediction," *arXiv preprint arXiv:1809.00786*, 2018.

[53] I. Singh, V. Blukis, A. Mousavian, A. Goyal, D. Xu, J. Tremblay, D. Fox, J. Thomason, and A. Garg, "Progprompt: Generating situated robot task plans using large language models," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 11 523–11 530.

[54] M. Tenorth, D. Nyga, and M. Beetz, "Understanding and executing instructions for everyday manipulation tasks from the world wide web," in *2010 ieee international conference on robotics and automation*. IEEE, 2010, pp. 1486–1491.

[55] Y. Artzi and L. Zettlemoyer, "Weakly supervised learning of semantic parsers for mapping instructions to actions," *Transactions of the Association for Computational Linguistics*, vol. 1, pp. 49–62, 2013.

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

[57] 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," *arXiv preprint arXiv:1911.00357*, 2019.

[58] 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.