RoboEXP: Action-Conditioned Scene Graph via Interactive Exploration for Robotic Manipulation

Hanxiao Jiang[1] Binghao Huang[1] Ruihai Wu[3] Zhuoran Li[4]
Shubham Garg[2] Hooshang Nayyeri[2] Shenlong Wang[1] Yunzhu Li[1]
[1]University of Illinois Urbana-Champaign [2]Amazon [3]Peking University [4]National University of Singapore

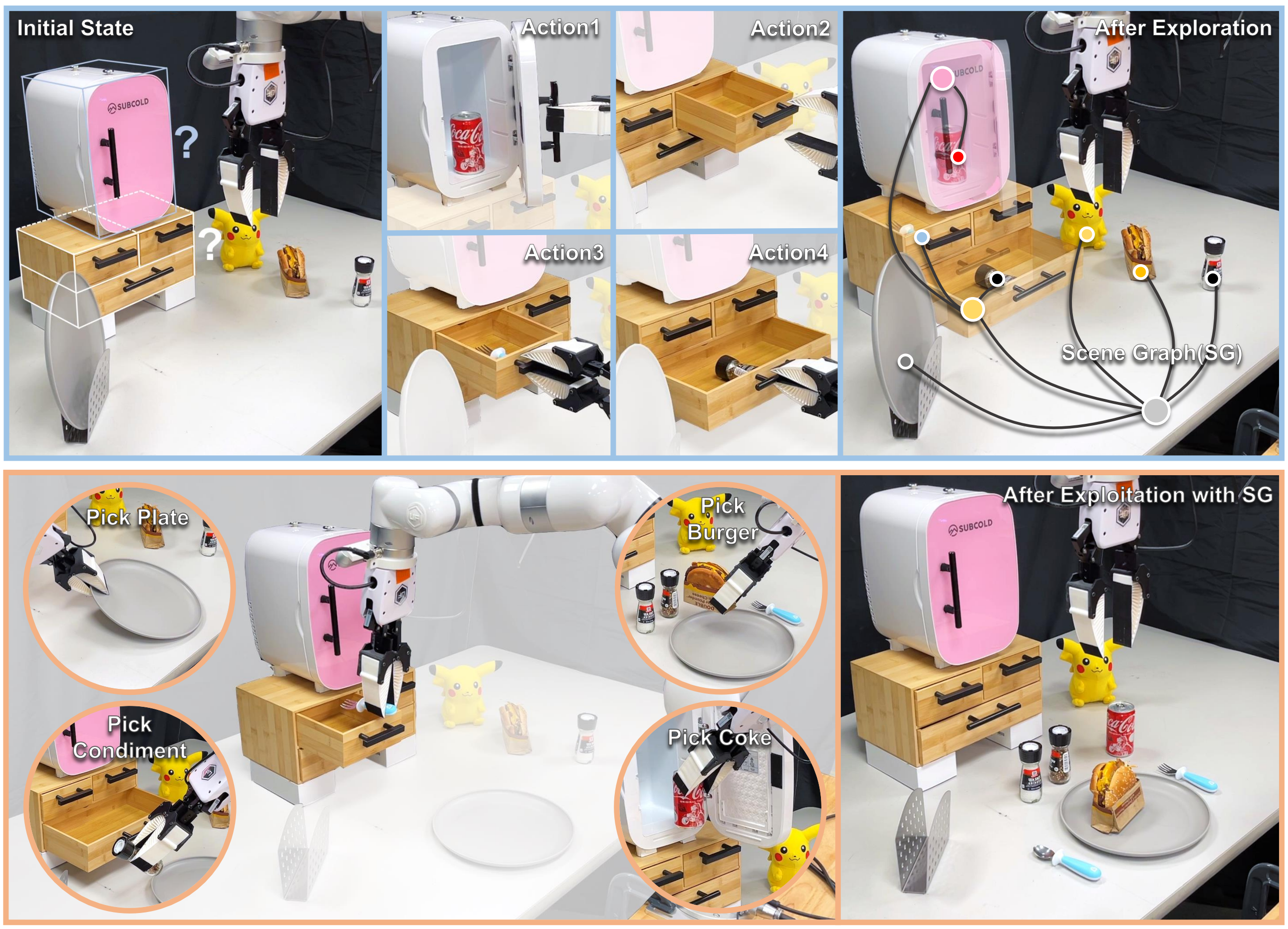

Figure 1. **Interactive Exploration to Construct an Action-Conditioned Scene Graph (ACSG) for Robotic Manipulation.** (a) **Exploration:** The robot autonomously explores by interacting with the environment to generate a comprehensive ACSG. This graph is used to catalog the locations and relationships of items. (b) **Exploitation:** Utilizing the constructed scene graph, the robot completes downstream tasks by efficiently organizing the necessary items according to the desired spatial and relational constraints.

1. Introduction

Imagine a future household robot designed to prepare breakfast. This robot must efficiently perform various tasks such as conducting inventory checks in cabinets, fetching food from the fridge, gathering utensils from drawers, and spotting leftovers under food covers. Key to its success is the ability to interact with and explore the environment, especially to find items that aren't immediately visible. Equipping it with such capabilities is crucial for the robot to effectively complete its everyday tasks.

Robot exploration and active perception have long been challenging areas in robotics [1–16]. Various techniques have been proposed, including information-theoretic approaches, curiosity-driven exploration, frontier-based meth-

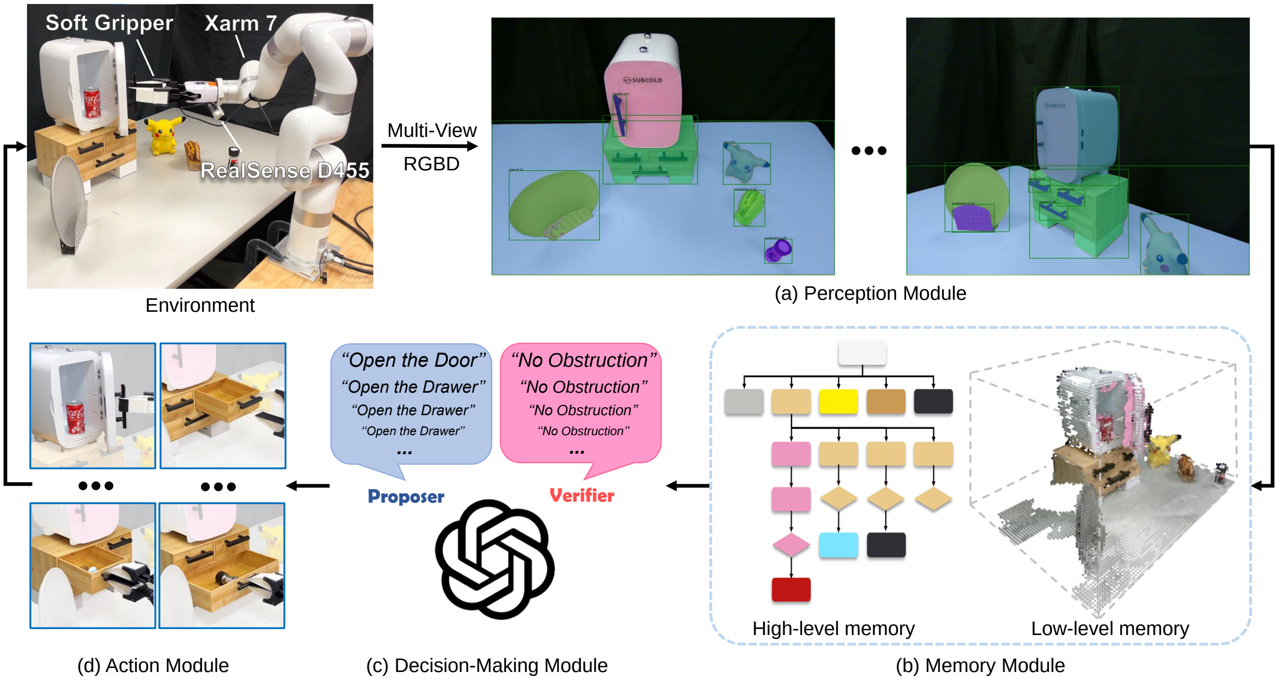

Figure 2. **Overview of Our RoboEXP System.** We present a comprehensive overview of our RoboEXP system, comprised of four modules. (a) Our **perception module** takes RGBD images as input and produces the corresponding 2D bounding boxes, masks, object labels, and associated semantic features as output. (b) The **memory module** seamlessly integrates 2D information into the 3D space, achieving more consistent 3D instance segmentation. Additionally, it constructs the high-level graph of our ACSG through the merging of instances. (c) Our **decision-making module** serves dual roles as a proposer and verifier. The proposer suggests various actions, such as opening doors and drawers, while the verifier assesses the feasibility of each action, considering factors like obstruction. (d) The **action module** executes the proposed actions, enabling the robot arm to interact effectively with the environment.

ods, and imitation learning [1, 13–15, 17–25]. Nevertheless, previous research has primarily focused on exploring static environments by merely changing viewpoints in a navigation setting or has been limited to interactions with a small set of object categories, such as drawers, or a closed set of simple actions like pushing [26].

In this work, we investigate the interactive scene exploration task, where the goal is to efficiently identify all objects, including those that are directly observable and those that can only be discovered through interaction between the robot and the environment (see Fig. 1). Towards this goal, we present a novel scene representation called action-conditioned 3D scene graph (ACSG). Unlike conventional 3D scene graphs that focus on encoding static relations, ACSG encodes both spatial relationships and logical associations indicative of action effects (e.g., opening a fridge will reveal an apple inside). We then show that interactive scene exploration can be formulated as a problem of action-conditioned 3D scene graph construction and traversal.

Tackling interactive scene exploration poses challenges: how can we reason about which objects need to be explored, choose the right action to interact with them, and maintain knowledge about our exploration findings? With these challenges in mind, we propose a novel, real-world robotic exploration framework, the RoboEXP system. RoboEXP can handle diverse exploration tasks in a zero-shot manner, constructing complex action-conditioned 3D scene graph in various scenarios, including those involving obstructing objects and requiring multi-step reasoning. We evaluate our system across various settings, spanning simple, single-object scenarios to complex environments, demonstrating its adaptability and robustness. The system also effectively manages different human interventions. Moreover, we show that our reconstructed action-conditioned 3D scene graph demonstrates strong capacity in performing multiple complex downstream tasks. Action-conditioned 3D scene graph advances LLM/LMM-guided robotic manipulation and decision-making research [27, 28], extending their operation domain from environments with known or observable objects to complicated environments with unknown or unobserved ones. To our knowledge, this is the first of its kind.

Our contributions are as follows: i) we propose action-conditioned 3D scene graph and introduce the interactive scene exploration task to address the challenging interaction aspect of exploration; ii) we develop the RoboEXP system, capable of exploring complicated environments with unseen objects in a wide range of settings; iii) through extensive experiments, we demonstrate our system's ability to construct complex and complete action-conditioned 3D scene graph, demonstrating significant potential for various manipulation tasks. Our experiments involve rigid and articulated objects, nested objects like Matryoshka dolls, and deformable objects like cloth, showcasing the system's generalization ability across objects, scene configurations, and downstream tasks.

References

[1] Farzad Niroui, Kaicheng Zhang, Zendai Kashino, and Goldie Nejat. Deep reinforcement learning robot for search and rescue applications: Exploration in unknown cluttered environments. *RA-L*, 2019. 1, 2

[2] Yugang Liu and Goldie Nejat. Robotic urban search and rescue: A survey from the control perspective. *Journal of Intelligent & Robotic Systems*, 2013.

[3] Philip Arm, Gabriel Waibel, Jan Preisig, Turcan Tuna, Ruyi Zhou, Valentin Bickel, Gabriela Ligeza, Takahiro Miki, Florian Kehl, Hendrik Kolvenbach, et al. Scientific exploration of challenging planetary analog environments with a team of legged robots. *Science robotics*, 2023.

[4] Martin J Schuster, Marcus G Müller, Sebastian G Brunner, Hannah Lehner, Peter Lehner, Ryo Sakagami, Andreas Dömel, Lukas Meyer, Bernhard Vodermayer, Riccardo Giubilato, et al. The arches space-analogue demonstration mission: Towards heterogeneous teams of autonomous robots for collaborative scientific sampling in planetary exploration. *RA-L*, 2020.

[5] KAI-QING Zhou, Kai Zheng, Connor Pryor, Yilin Shen, Hongxia Jin, L. Getoor, and X. Wang. Esc: Exploration with soft commonsense constraints for zero-shot object navigation. *ICML*, 2023.

[6] Ram Ramrakhya, Eric Undersander, Dhruv Batra, and Abhishek Das. Habitat-web: Learning embodied object-search strategies from human demonstrations at scale. In *CVPR*, 2022.

[7] Dipendra Misra, Andrew Bennett, Valts Blukis, Eyvind Niklasson, Max Shatkhin, and Yoav Artzi. Mapping instructions to actions in 3d environments with visual goal prediction. *arXiv preprint arXiv:1809.00786*, 2018.

[8] Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. Alfred: A benchmark for interpreting grounded instructions for everyday tasks. In *CVPR*, 2020.

[9] R. Bajcsy. Active perception. *Proceedings of the IEEE*, 1988.

[10] Active perception vs. passive perception. In *Proc. of IEEE Workshop on Computer Vision*, 1985.

[11] Andreas Bircher, Mina Kamel, Kostas Alexis, Helen Oleynikova, and Roland Siegwart. Receding horizon" next-best-view" planner for 3d exploration. In *ICRA*, 2016.

[12] Ana Batinovic, Antun Ivanovic, Tamara Petrovic, and Stjepan Bogdan. A shadowcasting-based next-best-view planner for autonomous 3d exploration. *RA-L*, 2022.

[13] Menaka Naazare, Francisco Garcia Rosas, and Dirk Schulz. Online next-best-view planner for 3d-exploration and inspection with a mobile manipulator robot. *RA-L*, 2022. 2

[14] Peihao Chen, Dongyu Ji, Kunyang Lin, Weiwen Hu, Wenbing Huang, Thomas Li, Mingkui Tan, and Chuang Gan. Learning active camera for multi-object navigation. *NeurIPS*, 2022.

[15] Tushar Nagarajan and Kristen Grauman. Learning affordance landscapes for interaction exploration in 3d environments. In *NeurIPS*, 2020. 2

[16] Neil Nie, Samir Yitzhak Gadre, Kiana Ehsani, and Shuran Song. Structure from action: Learning interactions for articulated object 3d structure discovery. *arXiv preprint arXiv:2207.08997*, 2022. 1

[17] Benjamin Charrow, Gregory Kahn, Sachin Patil, Sikang Liu, Ken Goldberg, Pieter Abbeel, Nathan Michael, and Vijay Kumar. Information-theoretic planning with trajectory optimization for dense 3d mapping. In *RSS*, 2015. 2

[18] Georgios Georgakis, Bernadette Bucher, Anton Arapin, Karl Schmeckpeper, Nikolai Matni, and Kostas Daniilidis. Uncertainty-driven planner for exploration and navigation. In *ICRA*, 2022.

[19] Christos Papachristos, Shehryar Khattak, and Kostas Alexis. Uncertainty-aware receding horizon exploration and mapping using aerial robots. In *ICRA*, 2017.

[20] Yuri Burda, Harri Edwards, Deepak Pathak, Amos Storkey, Trevor Darrell, and Alexei A. Efros. Large-scale study of curiosity-driven learning. In *ICLR*, 2019.

[21] Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *ICML*, 2017.

[22] Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. Embodied question answering. In *CVPR*, 2018.

[23] Pulkit Agrawal, Ashvin V Nair, Pieter Abbeel, Jitendra Malik, and Sergey Levine. Learning to poke by poking: Experiential learning of intuitive physics. *NeurIPS*, 2016.

[24] Yian Wang, Ruihai Wu, Kaichun Mo, Jiaqi Ke, Qingnan Fan, Leonidas Guibas, and Hao Dong. AdaAfford: Learning to adapt manipulation affordance for 3d articulated objects via few-shot interactions. In *ECCV*, 2022.

[25] Steven D Whitehead and Dana H Ballard. Active perception and reinforcement learning. In *Machine Learning Proceedings 1990*. 1990. 2

[26] Fei Xia, William B Shen, Chengshu Li, Priya Kasimbeg, Micael Edmond Tchapmi, Alexander Toshev, Roberto Martín-Martín, and Silvio Savarese. Interactive gibson benchmark: A benchmark for interactive navigation in cluttered environments. *RA-L*, 2020. 2

[27] Wenlong Huang, Chen Wang, Ruohan Zhang, Yunzhu Li, Jiajun Wu, and Li Fei-Fei. Voxposer: Composable 3d value maps for robotic manipulation with language models. *arXiv preprint arXiv:2307.05973*, 2023. 2

[28] Yingdong Hu, Fanqi Lin, Tong Zhang, Li Yi, and Yang Gao. Look before you leap: Unveiling the power of gpt-4v in robotic vision-language planning. *arXiv preprint arXiv: 2311.17842*, 2023. 2