Language-Driven Robotics: Learning and Interaction

Gi-Cheon Kang
Seoul National University

I. Introduction

Imagine a home robot, given the command: "Can you pour me a drink?" If the robot has not learned to pour, how can we teach it? If multiple options are available, how should the robot decide which drink to pour? Recent work has made significant progress toward *generalist* robotic policies [3, 4, 19, 1, 2] using large-scale demonstration datasets [26, 17]. However, collecting these demonstrations often requires expertise and access to specialized equipment [8, 29, 33], limiting the accessibility and scalability of robot learning. Moreover, robots often struggle with ambiguity, as they lack the ability to interact and clarify the user's intent, making it difficult for them to make justified decisions. These challenges severely limit their adaptability to unstructured environments, hindering their real-world deployment. Consequently, methods that enable (1) a wider audience (*e.g.,* non-experts) to teach robots new behaviors and (2) robots to resolve ambiguities through interaction are essential.

To address these challenges, **my research leverages natural language as an interface for both robot learning and human-robot interaction**. I seek to advance three axes: (1) enabling robots to learn visuomotor skills through language-based supervision [14], making robot learning more accessible and scalable, (2) facilitating robots to engage in dialogue with humans to reason about the user's intent for robotic manipulation [13, 18], and (3) developing robust vision-language models (VLMs) to build strong foundations for the first two axes, ensuring effective integration of visual and linguistic information for learning and interaction [12, 11, 10]. In the first axis, I propose a language-based teleoperation method that enables non-experts to collect robot demonstrations through natural language supervision. Then, I introduce a vision-language-action (VLA) model that learns visuomotor policies directly from language supervision. Unlike existing VLA models [4, 26, 19, 1, 2] that output low-level robotic actions, our model learns to predict actions in *language*, such as "move the arm forward," which demonstrates strong capabilities in acquiring new skills with a few demonstrations. In the second axis, I propose a new object-grasping task where a user provides an ambiguous and underspecified instruction (*e.g.,* "I am thirsty"). Moreover, I present a robotic system that aims to pick up one target object in the scene by interacting with the user using language. In the third axis, I propose several approaches for visually-grounded dialog [6].

II. Approach

A. Robot Learning from Natural Language Supervision

Large behavior models [4, 19, 24, 1, 2] trained on massive amounts of demonstrations [26, 17] through imitation learning have shown significant progress in robotic manipulation. However, these models, aimed at learning generalist manipulation policies, struggle to rapidly expand their set of manipulation skills for a wide range of real-world tasks. I suggest that a major bottleneck lies in the limited accessibility of data collection, as acquiring real-world robot data often requires expertise in robot control [33] or access to specialized devices, such as teleoperation [8] or virtual reality (VR) systems [30].

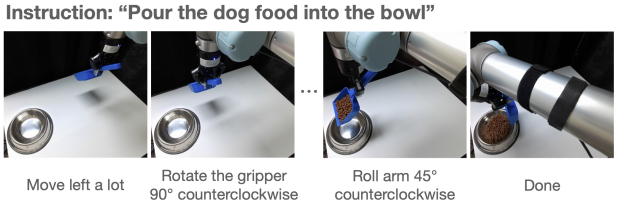

Fig. 1. Overview of collecting robot data from natural language supervision.

To address this, I developed language-based teleoperation, a data collection method to teach robots manipulation skills without relying on specialized expertise or devices for data collection. Fig. 1 illustrates language-based teleoperation in which a human collects data for a task based on the command (*e.g.,* "pour the dog food into the bowl"). The human first provides natural language supervision (*e.g.,* "move left a lot") in each state. The large language model (LLM) [25] then translates this supervision into appropriate robotic behavior, which is ultimately executed by the robot. By repeating this process, robot demonstrations are collected, where each state transition is associated with corresponding language supervision.

I also proposed a vision-language-action (VLA) model that learns visuomotor policies directly from language supervision. A core idea is to leverage natural language as supervision to train robotic policies, inspired by CLIP [27], which uses language as a training signal for visual representations. Our model employs CLIP models trained in Internet-scale data [28, 7] and adapts them to predict language-based motion primitives (*e.g.,* "move the arm forward by 10cm") through contrastive learning. Specifically, our model learns to measure the pairwise similarity between language supervision and contextual information (*i.e.,* current scene and language command). We train our model through a two-step process: pretraining and in-domain fine-tuning. In the pretraining stage, we train our model on the large-scale robot learning dataset (*i.e.,* Open X-Embodiment [26]) to improve generalization capabilities. The dataset does not contain language supervision, so we transform existing low-level robotic actions into templated natural language supervision to train our model. During in-

domain fine-tuning, our model learns diverse robotic skills using our collected data. Our proposed model outperforms the state-of-the-art VLA model [19] by a significant margin in acquiring novel manipulation skills, while using 7x fewer parameters. We further demonstrated that our method excels at few-shot generalization to novel tasks with a limited number of demonstrations (≤ 5).

B. Human-Robot Interaction

I have worked on language-conditioned robot manipulation, where robots manipulate objects based on natural language instruction from humans. A typical scenario of this problem involves specifying the category of the target object in instruction [31, 15, 35, 34, 21] (*e.g.,* "Give me a bottle of water"). However, in the same situation, humans often convey their *intentions* by relying on context to achieve their goals (*e.g.,*"I am thirsty."). Inspired by this, I have introduced a new task and corresponding dataset to study how robots can clarify the user's intent through interactions and perform context-appropriate behaviors.

The task requires robots to pick up the desired object in the given scene, but the language instructions are ambiguous and underspecified (Fig. 2). Therefore, the agents should interact with humans by asking questions to disambiguate the target object. Based on the task setup, we propose a new robotic system that effectively infers the user's intention and picks up the target object through dialogue. Our system continuously updates its belief by evaluating how well each object candidate in the scene aligns with the current visual and dialogue context, a process we call *pragmatic inference*. Pragmatic inference helps our system interpret the nuances of human language. For instance, if a user says, "The smaller one," the system does not just consider size in isolation—it also takes context into account. If the user previously referred to a specific category of objects, the system infers that "the small one" means the smallest object within that category, even if a smaller object exists elsewhere in the scene. We showcase that pragmatic inference helps identify the target object correctly with minimal human-AI interaction.

C. Visually-Grounded Dialog

The research directions mentioned above require models with a holistic understanding of visual perception and linguistic semantics. Thus, I have developed a strong foundation for language-driven robotics, particularly in the context of visually-grounded dialog systems that can continuously communicate with humans about visual scenes. Most of the previous approaches [9, 22, 5] have trained such models solely on human-collected visual dialog data [6] via supervised learning. One critical problem is that human-to-human visual dialog is hard to scale due to the need for extensive manual curation, limiting the generalization and robustness of models. To this end, I introduced a semi-supervised learning approach, called Generative Self-Training (GST), to scale data without human annotation. The key idea of GST is to generate synthetic dialog data for unlabeled Web images and train models on the data.

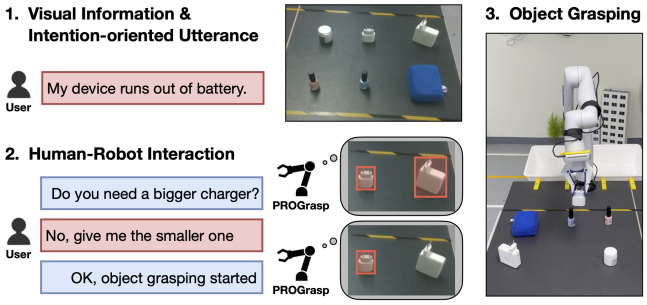

Fig. 2. Overview of interactive object grasping with an ambiguous instruction. The instruction does not contain the target object's category.

I have shown that synthetic data leads to significant gains in generalization performance. Moreover, our method enhances robustness against visual and linguistic adversarial attacks.

I have also tackled visual reference resolution, where visually grounded language models should resolve ambiguous expressions in human utterances (*e.g.,* "What color is *it*?") and ground them to a given image. I have proposed attention-based methods that effectively retrieve relevant dialog history to clarify ambiguous expressions. They have demonstrated their efficacy compared to prior approaches [20, 23].

III. Future Directions

A. Compositional Generalization for Long-Horizon Tasks

One of my research plans is to develop approaches for handling long-horizon robotic tasks, such as household chores [2]. A common strategy is to use high-level task planners [16, 32] that decompose complex tasks into sequences of learned skills. However, these planners often struggle with unstructured tasks, as their rigid decompositions may fail to support adaptive decision-making. As a complementary strategy, I plan to explore compositional generalization for long-horizon tasks. My goal is to develop methods that enable robots to efficiently learn higher-level tasks by composing previously learned skills, rather than requiring them to be trained from scratch. To achieve this, I plan to investigate language-conditioned policies for structured skill composition, allowing robots to generalize to increasingly complex behaviors. This approach will enhance adaptability by enabling robots to construct task hierarchies dynamically in response to novel scenarios.

B. Lifelong Learning and Interaction for Robotics

While I have explored the learning and interaction capabilities for language-driven robotics, these capabilities have been addressed independently rather than being interwoven into a cohesive framework. Inspired by humans who continuously refine their understanding of the world through experience, my plan is to develop lifelong learning systems for robotics that seamlessly integrate interaction and learning. These models should evolve their capabilities by circumventing catastrophic forgetting when exposed to new data or tasks. I am excited to study new paradigms for training lifelong learning models, enabling them to expand their grounded knowledge over time.

References

[1] Suneel Belkhale, Tianli Ding, Ted Xiao, Pierre Sermanet, Quon Vuong, Jonathan Tompson, Yevgen Chebotar, Debidatta Dwibedi, and Dorsa Sadigh. Rt-h: Action hierarchies using language. *arXiv preprint arXiv:2403.01823*, 2024.

[2] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. π_0: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024.

[3] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022.

[4] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023.

[5] Cheng Chen, Yudong Zhu, Zhenshan Tan, Qingrong Cheng, Xin Jiang, Qun Liu, and Xiaodong Gu. Utc: A unified transformer with inter-task contrastive learning for visual dialog. In *CVPR*, 2022.

[6] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[7] Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander Toshev, and Vaishaal Shankar. Data filtering networks. *arXiv preprint arXiv:2309.17425*, 2023.

[8] Zipeng Fu, Tony Z Zhao, and Chelsea Finn. Mobile aloha: Learning bimanual mobile manipulation with low-cost whole-body teleoperation. *arXiv preprint arXiv:2401.02117*, 2024.

[9] Zhe Gan, Yu Cheng, Ahmed EI Kholy, Linjie Li, Jingjing Liu, and Jianfeng Gao. Multi-step reasoning via recurrent dual attention for visual dialog. In *ACL*, 2019.

[10] Gi-Cheon Kang, Jaeseo Lim, and Byoung-Tak Zhang. Dual attention networks for visual reference resolution in visual dialog. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019.

[11] Gi-Cheon Kang, Junseok Park, Hwaran Lee, Byoung-Tak Zhang, and Jin-Hwa Kim. Reasoning visual dialog with sparse graph learning and knowledge transfer. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2021.

[12] Gi-Cheon Kang, Sungdong Kim, Jin-Hwa Kim, Donghyun Kwak, and Byoung-Tak Zhang. The dialog must go on: Improving visual dialog via generative self-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[13] Gi-Cheon Kang, Junghyun Kim, Jaein Kim, and Byoung-Tak Zhang. Prograsp: Pragmatic human-robot communication for object grasping. In *International Conference on Robotics and Automation (ICRA)*, 2024.

[14] Gi-Cheon Kang, Junghyun Kim, Kyuhwan Shim, Jun Ki Lee, and Byoung-Tak Zhang. Clip-rt: Learning language-conditioned robotic policies from natural language supervision. In *Proceedings of Robotics: Science and Systems (RSS)*, 2025.

[15] Jun Hatori, Yuta Kikuchi, Sosuke Kobayashi, Kuniyuki Takahashi, Yuta Tsuboi, Yuya Unno, Wilson Ko, and Jethro Tan. Interactively picking real-world objects with unconstrained spoken language instructions. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3774–3781. IEEE, 2018.

[16] Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In *International conference on machine learning*, pages 9118–9147. PMLR, 2022.

[17] Eric Jang, Alex Irpan, Mohi Khansari, Daniel Kappler, Frederik Ebert, Corey Lynch, Sergey Levine, and Chelsea Finn. Bc-z: Zero-shot task generalization with robotic imitation learning. In *Conference on Robot Learning*, pages 991–1002. PMLR, 2022.

[18] Junghyun Kim, Gi-Cheon Kang, Jaein Kim, Suyeon Shin, and Byoung-Tak Zhang. Gvcci: Lifelong learning of visual grounding for language-guided robotic manipulation. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2023.

[19] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.

[20] Satwik Kottur, José MF Moura, Devi Parikh, Dhruv Batra, and Marcus Rohrbach. Visual coreference resolution in visual dialog using neural module networks. In *ECCV*, 2018.

[21] Yuchen Mo, Hanbo Zhang, and Tao Kong. Towards open-world interactive disambiguation for robotic grasping. In *CoRL Workshop on Learning, Perception, and Abstraction for Long-Horizon Planning*, 2022.

[22] Vishvak Murahari, Dhruv Batra, Devi Parikh, and Abhishek Das. Large-scale pretraining for visual dialog: A simple state-of-the-art baseline. In *ECCV*, 2020.

[23] Yulei Niu, Hanwang Zhang, Manli Zhang, Jianhong Zhang, Zhiwu Lu, and Ji-Rong Wen. Recursive visual attention in visual dialog. In *CVPR*, 2019.

[24] Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey

Hejna, Charles Xu, Jianlan Luo, Tobias Kreiman, You Liang Tan, Lawrence Yunliang Chen, Pannag Sanketi, Quan Vuong, Ted Xiao, Dorsa Sadigh, Chelsea Finn, and Sergey Levine. Octo: An open-source generalist robot policy. In *Proceedings of Robotics: Science and Systems*, Delft, Netherlands, 2024.

[25] OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[26] Abhishek Padalkar, Acorn Pooley, Ajinkya Jain, Alex Bewley, Alex Herzog, Alex Irpan, Alexander Khazatsky, Anant Rai, Anikait Singh, Anthony Brohan, et al. Open x-embodiment: Robotic learning datasets and rt-x models. *arXiv preprint arXiv:2310.08864*, 2023.

[27] 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.

[28] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5b: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. URL https://openreview.net/forum?id=M3Y74vmsMcY.

[29] Ahjeong Seo, Gi-Cheon Kang, Joonhan Park, and Byoung-Tak Zhang. Attend what you need: Motion-appearance synergistic networks for video question answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2021.

[30] Mingyo Seo, Steve Han, Kyutae Sim, Seung Hyeon Bang, Carlos Gonzalez, Luis Sentis, and Yuke Zhu. Deep imitation learning for humanoid loco-manipulation through human teleoperation. In *2023 IEEE-RAS 22nd International Conference on Humanoid Robots (Humanoids)*, pages 1–8. IEEE, 2023.

[31] Mohit Shridhar and David Hsu. Interactive visual grounding of referring expressions for human-robot interaction. *arXiv preprint arXiv:1806.03831*, 2018.

[32] Chan Hee Song, Jiaman Wu, Clayton Washington, Brian M Sadler, Wei-Lun Chao, and Yu Su. Llm-planner: Few-shot grounded planning for embodied agents with large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2998–3009, 2023.

[33] Ted Xiao, Harris Chan, Pierre Sermanet, Ayzaan Wahid, Anthony Brohan, Karol Hausman, Sergey Levine, and Jonathan Tompson. Robotic skill acquistion via instruction augmentation with vision-language models. In *Proceedings of Robotics: Science and Systems*, 2023.

[34] Yang Yang, Xibai Lou, and Changhyun Choi. Interactive robotic grasping with attribute-guided disambiguation. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 8914–8920. IEEE, 2022.

[35] Hanbo Zhang, Yunfan Lu, Cunjun Yu, David Hsu, Xuguang La, and Nanning Zheng. Invigorate: Interactive visual grounding and grasping in clutter. *arXiv preprint arXiv:2108.11092*, 2021.