Language and Robotics: Toward Building Robots Coexisting with Human Society Using Language Interface

Yutaka Nakamura[1], **Shuhei Kurita**[2], **and Koichiro Yoshino**[1]
[1] Guardian Robot Project (GRP), R-IH, RIKEN, Japan
[2] Center for Advanced Intelligence Project (AIP), RIKEN, Japan
`{yutaka.nakamura, shuhei.kurita, koichiro.yoshino}@riken.jp`

1 Introduction

Robots are one of the archetypes of AI systems we imagine, and the realization of such robots operating in the real world with language interfaces has long been a dream of us. This vision we have dreamed of is rapidly becoming a reality with the contribution of recent advancements in large language models (LLMs). However, there are still many problems that the research community needs to tackle in order for LLMs and other NLP tools to work in the real world.

This introductory tutorial aims to help researchers who will start language and robotics (*LangRobo*) research in the future by summarizing three points: awareness of the community's issues, recent approaches for these issues, and remaining problems. This tutorial requires only basic NLP knowledge: language modeling and basic NLP task definition. We arrange this tutorial involving not only NLP researchers but also robotics researchers in order to raise issues that are relevant to actual robotics problems.

The connection between NLP and robotics is a challenge that has been tackled in the field of robotics for many years; we have faced the difficulties of the problem many times. There are several difficulties in connecting NLP and robotics, but the following three are particularly problematic:

1. The great difference in granularity between language and robot behavior
2. Robotics tasks involving real-world control often do not allow for language ambiguity
3. Language expressions themselves are often ambiguous and require background knowledge or commonsense reasoning to understand them correctly

While the language expressions used for robots are only a few dozen words at most, robots have countless events to consider, such as their motion trajectories and interactions with things in the real world. In other words, in the field of robotics, there are countless events to be considered outside the language framework, and it is impossible to consider all of them in the model. Another important point to consider is the ambiguity of language expressions. Humans often make omissions when utilizing language, and in practice, such omissions are often very important in actual robot tasks. The omission is closely related to multimodal information obtained from the interaction, such as eye gaze and motions. It is nearly impossible to address these issues with NLP alone or robotics alone.

Many recent works have suggested that deep learning or LLMs can provide solutions to these problems. This tutorial summarizes the recent approaches to the language and robotics problem using such learning-based approaches. The goal of this tutorial is to share the discussion on how these problems can be solved in the future.

2 Tutorial Content

As the tutorial contents, a researcher who has been working in robotics for a long time will first introduce classic language utilization problems in their field. He also explains how these problems have been solved in recent years by deep learning and LLMs. In the second part, a researcher who has been working in the fields of language and image processing will explain the recent research on coreference and grounding problems in the real world. It is also mentioned that the recent advancement of vision and language research. In the last lecture part, a researcher who has been working in the human-robot interaction research field, including dialogues, will discuss the collaboration between users and robots and the issues of language understanding and situation understanding necessary for such collaboration. Finally, many

Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Tutorial Abstracts), pages 1–7
November 1–4, 2023.

unresolved robot problems will be touched upon, and future language and robotics research directions will be discussed.

2.1 Language Use in Robotics Field (Nakamura)

In the research field of robotics, navigation and manipulation have been major issues. They used numerical representations (e.g., coordinates and joint angles that indicate the robot's position) as the main focus. On the other hand, logical planning frameworks have been widely applied to robotics, such as STanford Research Institute Problem Solver (STRIPS) (Fikes and Nilsson, 1971) and Planning Domain Definition Language (PDDL) (Fox and Long, 2003). They have been proposed as frameworks for planning in a virtual space, mainly in the block world, using a programing language-like description. Due to the advancement of deep learning since the 2010s, a system was built in 2018 to interpret human instructions given in natural language to perform picking tasks (Hatori et al., 2018). The advancement of large language models (LLMs) is leading several frameworks for interpreting natural languages such as GaTo (Reed et al.), Say Can (Brohan et al., 2023), and RT-1 (Brohan et al., 2022). In addition, research on generative models of human motion, such as the human motion diffusion model (Tevet et al., 2022), is developing, which is expected to enable robots to cooperate with humans in unstructured environments. In this tutorial, we will give an overview of these studies.

2.2 Language Understandings in the Real-world (Kurita)

Recent remarkable advancements in natural language processing have enabled a comprehensive understanding of texts in the context of syntactic and semantic analyses, question-answering, summarization, translation, and even dialogue tasks. However, such models often face challenges when dealing with multimodal contexts in the real world. In this tutorial, we explain how current models struggle to handle real-world contexts and provide an overview of language grounding technologies from four perspectives.

The first perspective focuses on vision and language tasks with a single image. Examples of such tasks include image captioning on MS-COCO (Bernardi et al., 2016) and visual question answering (Antol et al., 2015). We discuss the strengths and limitations of these existing tasks, particularly their reliance on limited contextual information from images. We further introduce the referring expression comprehension or simply "visual grounding" task (Kazemzadeh et al., 2014; Plummer et al., 2015; Yu et al., 2016; Mao et al., 2016), which specifies the target object from a referred expression and the relation to the open-vocabulary object detection task.

The second perspective is obtained through videos. We concentrate on the first-person videos here as they are obtained through the motion of the camera wearer. Recently, a large-scale first-person perspective video dataset of Ego4D was proposed (Grauman et al., 2022). This can be extended for robots that navigate in scenes and related language tasks. Although the model learning from videos enriches the model perspectives in the real-world, image frames in videos are constrained to the preset viewpoints when they are recorded.

The third perspective involves 3D scenes and virtual worlds that provide rich contextual information about the captured scenes. Unlike the previous perspectives, this perspective allows the "embodied" experience for the agents in the environments. Examples of such enriched scenes have been proposed, such as ScanNet (Dai et al., 2017) and Matterport 3D (Chang et al., 2017). 3D referring expression comprehension (Chen et al., 2020) and 3D-QA (Azuma et al., 2022) are also interesting spatial understanding with language expressions. These environments also enable visually-grounded interactive textual understanding tasks. One example is vision and language navigation (VLN) (Anderson et al., 2018), where an agent navigates in environments based on visual and textual information. An interesting approach for VLN is a captioning model from navigation paths. Fried et al. (2018) introduced the speaker-model for generating captions for the paths the agent navigates in environments. Several studies used this speaker model for the training dataset augmentation (Tan et al., 2019; Hao et al., 2020). Kurita and Cho (2021) used the speaker model for ranking the possible action candidates during navigation. Recently, Habitat simulator (Manolis Savva* et al., 2019; Szot et al., 2021) enables continuous navigation on VLN (Krantz et al., 2020). AI2Thor is another virtual environment that enables object-interactive tasks. This is used in the instruction following task of ALFRED (Shridhar et al., 2020).

The final perspective is robotics. Among recent studies, SayCan (Ahn et al., 2022) uses large language models for ranking the possible action candidates during the episode, while Liang et al. (2022) uses large language models for decomposing instructions to perceptions and actions expressed in executable Python code format. Indeed, this perspective is still ongoing, and further elaborations are desired.

2.3 Interactive Robots (Yoshino)

When we focus on the interaction between robots and humans, from an engineering perspective, it is important to understand what robots can achieve in cooperation with humans. From a scientific perspective, it is also important to investigate the relationship between real-world nature, physicality, and linguistic expressions. For example, Visual Question Answering (VQA) (Antol et al., 2015) or Audio Visual Scene-aware Dialogue (AVSD) (D ' Haro et al., 2020; Alamri et al., 2019) is a typical case. Information exchange in a language tied (grounded) to the real world and interaction context is important for real-world language robotics. The use of multimodal data in the real world is important to solve this grounding challenge (Kottur et al., 2021). Discussions about the physicality of robots are also essential (Ahuja and Morency, 2019; Yazdian et al., 2022).

When we try to collaborate with a robot through actual interaction, the challenge is bridging the language interaction to real-world interaction, including actuation and grounding, using implicit knowledge or common sense about physical phenomena or social relationships (Xia et al., 2020). Conventionally such implicit knowledge is handcrafted as ontology; however, the building is costly because the robot requires prerequisite knowledge, common sense, and unspoken knowledge for each task and environment (Tanaka et al., 2023). Some recent works utilizing LLM indicate that LLMs can be used as such implicit world knowledge (Wu et al., 2023).

In addition to the above, there are still other research areas in robotics where language should play an important role, such as the robot's intentions, subjective experience (Yuguchi et al., 2022), desires and preferences, and memory mechanisms (Peller-Konrad et al., 2022). This tutorial will also discuss the future direction of language and robotics research.

3 Tutorial Format

Our tutorial consists of three lectures and one discussion. Each lecture has 40 mins talk with 10 mins short break from different viewpoints: language use in robotics, NLP in the real world, and interactive robots. After 30 mins coffee break, we will have a discussion about the future direction of the language and robotics research field.

During the lecture session, we open a question-answering form such as Dory, and participants put their questions and comments or vote for them. Based on the questions raised on the form, we have 40 mins open discussion with tutorial participants.

4 Reading List

This tutorial is introductory, and we do not assume special knowledge of participants if they have learned natural language processing. However, if you have never learned, reading papers about Transformer (Vaswani et al., 2017) and diffusion model (Ho et al., 2020) will emphasize your understanding. This tutorial will use the robot operating system 2 (ROS2). Online tutorial of ROS2[1] will emphasize your understanding.

5 Technical Requirements

We will use online communication systems such as Dory[2] to encourage discussion. It is expected to have devices that have internet access during the tutorial. We will open our slides and materials on our webpage[3] before the tutorial.

6 Instructors

We have three instructors from different research fields: robotics and control, vision and language, and human-robot interaction. The bibliography of each instructor follows.

6.1 Yutaka Nakamura

Yutaka Nakamura is a Team Leader at the Institute of Physical and Chemical Research (RIKEN) and an Affiliate Professor at Osaka University. He received his degree, Dr. Eng., from Nara Institute of Science and Technology (NAIST) in 2004. He worked at Osaka University as an assistant professor and an associate professor. Since 2020, he has

[1] `https://docs.ros.org/en/foxy/Tutorials.html`
[2] `https://dory.app`
[3] `https://github.com/riken-grp/langrobo-tutorial`

been working at Guardian Robot Project (GRP) of RIKEN as the team leader of Behavior Learning research team. He is working on areas of robotics, control, and human-robot interaction.

6.2 Shuhei Kurita

Shuhei Kurita is a Research Scientist at Center for Advanced Intelligence Project (AIP), RIKEN. He received his Ph.D. in informatics from Kyoto University in 2019. He is a visiting researcher in New York University for Assoc. Prof. Kyunghyun Cho from 2020. His paper "Neural Joint Model for Transition-based Chinese Syntactic Analysis"" was selected as the outstanding paper of ACL2017 (Kurita et al., 2017). He is working on natural language understanding in the real-world expressed in images, 3D scenes and photo-realistic simulator. He has actively published papers in natural language processing, learning representations and computer vision venues.

6.3 Koichiro Yoshino

Koichiro Yoshino is a Team Leader at the Institute of Physical and Chemical Research (RIKEN) and an Affiliate Professor at Nara Institute of Science and Technology (NAIST). He received his Ph.D. in informatics from Kyoto University in 2014. He worked at Kyoto University as a postdoc and at NAIST as an assistant professor. Since 2020, he has been working at Guardian Robot Project (GRP) of RIKEN as the team leader of Knowledge Acquisition and Dialogue research team. From 2019 to 2020, he was a visiting researcher at Heinrich-Heine-Universität Düsseldorf, Germany. He is working on spoken and natural language processing areas, especially robot dialogue systems. Dr. Koichiro Yoshino received several honors, including the best paper award of IWSDS2020 and the best paper award of the 1st NLP4ConvAI workshop. He is a member of IEEE Speech and Language Processing Technical Committee (SLTC), a member of Dialogue System Technology Challenge (DSTC) Steering Committee, an action editor of ACL Rolling Review (ARR), and a board member of SIGdial. He is a member of ACL, IEEE, SIGDIAL, IPSJ, JSAI, ANLP and RSJ.

7 Ethical Statement

Data used in Language and Robotics often contain personal identification codes such as facial images. The multimodal and interaction data used by the authors in this tutorial are discussed and reviewed by the ethics committee, if necessary, in accordance with the code of ethics of the organization to which they belong.

References

Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Daniel Ho, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Eric Jang, Rosario Jauregui Ruano, Kyle Jeffrey, Sally Jesmonth, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Kuang-Huei Lee, Sergey Levine, Yao Lu, Linda Luu, Carolina Parada, Peter Pastor, Jornell Quiambao, Kanishka Rao, Jarek Rettinghouse, Diego Reyes, Pierre Sermanet, Nicolas Sievers, Clayton Tan, Alexander Toshev, Vincent Vanhoucke, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Mengyuan Yan, and Andy Zeng. 2022. Do as i can and not as i say: Grounding language in robotic affordances. In *arXiv preprint arXiv:2204.01691*.

Chaitanya Ahuja and Louis-Philippe Morency. 2019. Language2pose: Natural language grounded pose forecasting. In *2019 International Conference on 3D Vision (3DV)*, pages 719–728. IEEE.

Huda Alamri, Vincent Cartillier, Abhishek Das, Jue Wang, Anoop Cherian, Irfan Essa, Dhruv Batra, Tim K Marks, Chiori Hori, Peter Anderson, et al. 2019. Audio visual scene-aware dialog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7558–7567.

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. 2018. Vision-and-Language Navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

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

Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoki Kawanabe. 2022. Scanqa: 3d question answering for spatial scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Raffaella Bernardi, Ruket Cakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, and Barbara Plank. 2016. Automatic description generation from images: A survey of models, datasets, and evaluation measures. 55(1):409–442.

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

Anthony Brohan, Yevgen Chebotar, Chelsea Finn, Karol Hausman, Alexander Herzog, Daniel Ho, Julian Ibarz, Alex Irpan, Eric Jang, Ryan Julian, et al. 2023. Do as i can, not as i say: Grounding language in robotic affordances. In *Conference on Robot Learning*, pages 287–318. PMLR.

Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. 2017. Matterport3D: Learning from RGB-D data in indoor environments. *International Conference on 3D Vision (3DV)*.

Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. 2020. Scanrefer: 3d object localization in rgb-d scans using natural language. *16th European Conference on Computer Vision (ECCV)*.

Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. 2017. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*.

Luis Fernando D'Haro, Koichiro Yoshino, Chiori Hori, Tim K Marks, Lazaros Polymenakos, Jonathan K Kummerfeld, Michel Galley, and Xiang Gao. 2020. Overview of the seventh dialog system technology challenge: Dstc7. *Computer Speech & Language*, 62:101068.

Richard E Fikes and Nils J Nilsson. 1971. Strips: A new approach to the application of theorem proving to problem solving. *Artificial intelligence*, 2(3-4):189–208.

Maria Fox and Derek Long. 2003. Pddl2. 1: An extension to pddl for expressing temporal planning domains. *Journal of artificial intelligence research*, 20:61–124.

Daniel Fried, Ronghang Hu, Volkan Cirik, Anna Rohrbach, Jacob Andreas, Louis-Philippe Morency, Taylor Berg-Kirkpatrick, Kate Saenko, Dan Klein, and Trevor Darrell. 2018. Speaker-follower models for vision-and-language navigation. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 3314–3325. Curran Associates, Inc.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Christian Fuegen, Abrham Gebreselasie, Cristina Gonzalez, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jachym Kolar, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz Puentes, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Yunyi Zhu, Pablo Arbelaez, David Crandall, Dima Damen, Giovanni Maria Farinella, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. 2022. Ego4d: Around the World in 3,000 Hours of Egocentric Video. In *Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition (CVPR)*.

Weituo Hao, Chunyuan Li, Xiujun Li, Lawrence Carin, and Jianfeng Gao. 2020. Towards learning a generic agent for vision-and-language navigation via pretraining. *Conference on Computer Vision and Pattern Recognition (CVPR)*.

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

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. ReferItGame: Referring to objects in photographs of natural scenes. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Satwik Kottur, Seungwhan Moon, Alborz Geramifard, and Babak Damavandi. 2021. Simmc 2.0: A task-oriented dialog dataset for immersive multimodal conversations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4903–4912.

Jacob Krantz, Erik Wijmans, Arjun Majundar, Dhruv Batra, and Stefan Lee. 2020. Beyond the nav-graph: Vision and language navigation in continuous environments. In *European Conference on Computer Vision (ECCV)*.

Shuhei Kurita and Kyunghyun Cho. 2021. Generative language-grounded policy in vision-and-language navigation with bayes' rule. In *International Conference on Learning Representations (ICLR)*.

Shuhei Kurita, Daisuke Kawahara, and Sadao Kurohashi. 2017. Neural joint model for transition-based Chinese syntactic analysis. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1204–1214, Vancouver, Canada. Association for Computational Linguistics.

Jacky Liang, Wenlong Huang, Fei Xia, Peng Xu, Karol Hausman, Brian Ichter, Pete Florence, and Andy Zeng. 2022. Code as policies: Language model programs for embodied control. In *arXiv preprint arXiv:2209.07753*.

Manolis Savva*, Abhishek Kadian*, Oleksandr Maksymets*, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. 2019. Habitat: A Platform for Embodied AI Research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan Yuille, and Kevin Murphy. 2016. Generation and Comprehension of Unambiguous Object Descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Fabian Peller-Konrad, Christian RG Dreher, Andre Meixner, Fabian Reister, Markus Grotz, Tamim Asfour, et al. 2022. Conceptual design of the memory system of the robot cognitive architecture armarx. *arXiv preprint arXiv:2206.02241*.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gómez Colmenarejo, Alexander Novikov, Gabriel Barth-maron, Mai Giménez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, et al. A generalist agent. *Transactions on Machine Learning Research*.

Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. 2020. ALFRED: A Benchmark for Interpreting Grounded Instructions for Everyday Tasks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Andrew Szot, Alex Clegg, Eric Undersander, Erik Wijmans, Yili Zhao, John Turner, Noah Maestre, Mustafa Mukadam, Devendra Chaplot, Oleksandr Maksymets, Aaron Gokaslan, Vladimir Vondrus, Sameer Dharur, Franziska Meier, Wojciech Galuba, Angel Chang, Zsolt Kira, Vladlen Koltun, Jitendra Malik, Manolis Savva, and Dhruv Batra. 2021. Habitat 2.0: Training home assistants to rearrange their habitat. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Hao Tan, Licheng Yu, and Mohit Bansal. 2019. Learning to navigate unseen environments: Back translation with environmental dropout. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2610–2621, Minneapolis, Minnesota. Association for Computational Linguistics.

Shohei Tanaka, Koichiro Yoshino, Katsuhito Sudoh, and Satoshi Nakamura. 2023. Reflective action selection based on positive-unlabeled learning and causality detection model. *Computer Speech & Language*, 78:101463.

Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. 2022. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Jimmy Wu, Rika Antonova, Adam Kan, Marion Lepert, Andy Zeng, Shuran Song, Jeannette Bohg, Szymon Rusinkiewicz, and Thomas Funkhouser. 2023. Tidybot: Personalized robot assistance with large language models. *arXiv preprint arXiv:2305.05658*.

Fei Xia, William B Shen, Chengshu Li, Priya Kasimbeg, Micael Edmond Tchapmi, Alexander Toshev, Roberto Martín-Martín, and Silvio Savarese. 2020. Interactive gibson benchmark: A benchmark for interactive navigation in cluttered environments. *IEEE Robotics and Automation Letters*, 5(2):713–720.

Payam Jome Yazdian, Mo Chen, and Angelica Lim. 2022. Gesture2vec: Clustering gestures using representation learning methods for co-speech gesture generation. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3100–3107. IEEE.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. 2016. Modeling Context in Referring Expressions. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Akishige Yuguchi, Seiya Kawano, Koichiro Yoshino, Carlos Toshinori Ishi, Yasutomo Kawanishi, Yutaka Nakamura, Takashi Minato, Yasuki Saito, and

Michihiko Minoh. 2022. Butsukusa: A conversational mobile robot describing its own observations and internal states. In *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 1114–1118. IEEE.