Embodied Language Learning: Opportunities, Challenges, and Future Directions

Nadine Amin
Computer and Information Technology
Purdue University
West Lafayette, Indiana, USA
amin37@purdue.edu

Julia Rayz
Computer and Information Technology
Purdue University
West Lafayette, Indiana, USA
jtaylor1@purdue.edu

Abstract

While large language and vision-language models showcase impressive capabilities, they face a notable limitation: the inability to connect language with the physical world. To bridge this gap, research has focused on embodied language learning, where the language learner is situated in the world, perceives it, and interacts with it. This article explores the current standing of research in embodied language learning, highlighting opportunities and discussing common challenges. Lastly, it identifies existing gaps from the perspective of language understanding research within the embodied world and suggests potential future directions.

1 Introduction

Besides observing their surroundings, humans actively contribute to their understanding of the world by interacting with it and communicating with others (Smith and Gasser, 2005; Barsalou, 2008). This interactive experience is integral to language acquisition and understanding (Bender and Koller, 2020). A corresponding notion of World Scopes (Bisk et al., 2020a), namely *Corpus*, *Internet*, *Perception*, *Embodiment*, and *Social*, has been proposed to measure progress in language understanding research. While today's large language models (LLMs) have exhibited powerful capabilities (Bommasani et al., 2021), their textual training data constrain them to the *Internet* world scope. In turn, large vision-language models, trained on image-text corpora, fall within the world of *Perception*.

Significant research efforts (see §2) have been devoted to transitioning into the embodied realm. However, lying at the intersection of language and robotics research, embodied language learning has been primarily explored from a robotics perspective, with a focus on the general use of language in robotics (Tellex et al., 2020), embodied vision-language tasks (Francis et al., 2022; Duan et al., 2022; Deitke et al., 2022), or foundation models for decision-making (Yang et al., 2023) or as agents (Xi et al., 2023). The main contribution of this article is providing an overview of embodied language learning from the perspective of language understanding research, summarizing relevant background (§2), opportunities (§3), challenges (§4), and research gaps (§5). Throughout the article, *embodied language learning* is taken to be the process of language acquisition by a language learner, referred to as an *agent* or a *robot*, while it is situated in a physical or a virtual world which it perceives and interacts with through action taking; hence, grounding language in its percepts and actions.

2 Embodied Language Learning

Current dominant approaches to language modeling, as represented by LLMs, involve training on a vast quantity of textual data. However, as symbol tokens cannot be grounded in other symbol tokens (Harnad, 1990) and meaning is perceived to be residing in the connection between language and extrinsic non-symbolic representations (Ervin-Tripp, 1973; Bisk et al., 2020a; Bender and Koller, 2020; Lake and Murphy, 2020), merely training language models on as many textual corpora as possible remains insufficient for capturing meaning (Harnad, 1990; Lucy and Gauthier, 2017; Bender and Koller, 2020). It is essential to ground textual corpora in extra-linguistic data (Harnad, 1990; Bisk et al., 2020a). Meaning can then be captured to the extent reflected in such data (Bender and Koller, 2020).

One source of grounding data is perception (Harnad, 1990), including visual, tactile, and auditory inputs (Smith and Gasser, 2005; Bisk et al., 2020a). Efforts have been directed towards vision-language models that attempt to ground language in the visual input, learning alignments between the modalities. Nevertheless, studies on language acquisition among infants (Snow et al., 1976; Kuhl, 2007) suggested that mere perception is inadequate, and that

Findings of the Association for Computational Linguistics ACL 2024, pages 15369–15379

language cannot be learned from a television (Bisk et al., 2020a; Bender and Koller, 2020).

Meaning and language understanding have been investigated with regard to embodiment theories in cognitive science (Glenberg and Robertson, 2000), which highlight the significance of situated actions (Smith and Gasser, 2005; Barsalou, 2008). Evidence has supported that meaning representations are grounded in embodied experiences and sensorimotor interactions with the world (Jones et al., 1991; Glenberg and Kaschak, 2002; Barsalou and Wiemer-Hastings, 2005). This has motivated research in embodied language learning, where language is grounded in both perception and action (Heinrich et al., 2020).

Much of the research in embodied language learning has been focused on robot learning, following two approaches: embodied exploration and embodied instruction following (see Appendix A). In the first approach, a robot interacts with objects while receiving natural language descriptions of its actions and/or object attributes (Heinrich et al., 2020; Özdemir et al., 2021; Zhang et al., 2023; Tatiya et al., 2023). In the second approach, a robot is given a natural language instruction and learns to execute a corresponding short-horizon skill (Jang et al., 2022; Brohan et al., 2023; Zitkovich et al., 2023; Vuong et al., 2023; Jiang et al., 2023) or plan and carry out sequences of actions to reach a corresponding long-horizon goal (Suglia et al., 2021; Hong et al., 2021; Jin et al., 2023; Driess et al., 2023; Jiang et al., 2023). Recently, Liu et al. (2023) experimented with a robot assigned a non-language task in an environment with language annotations that are useful but not required for task completion. In all approaches, the robot learns to ground language in its sensorimotor experience.

3 Opportunities & Prospects

Opportunities of grounding language manifest in the various dimensions of understanding it unlocks (Bisk et al., 2020a), which are discussed below.

3.1 Attributes

The meaning of some attributes cannot be fully grasped through mere perception (Gibson, 1988; Bisk et al., 2020a; Tatiya et al., 2023). Understanding attributes such as deformability, weight, and hardness requires interacting with objects and perceiving the resulting multi-sensory effects. For instance, lifting an opaque container lends meaning to its *emptiness* (Zhang et al., 2023). Establishing such connections between language and actions allows an embodied language learner to reason about these properties (Zellers et al., 2021). For example, it can recognize that a greater force is needed to push a heavier object (Lake and Murphy, 2020).

3.2 World Dynamics & Affordances

Embodiment also allows agents to experiment with different actions (Smith and Gasser, 2005; Bisk et al., 2020a) and associate them with the change they induce in the world (Smith and Gasser, 2005; McClelland et al., 2019; Zellers et al., 2021). This establishes notions of cause and effect (Piaget et al., 1952; Engstrø, 2000), fostering an understanding of world dynamics, physical constraints, and affordances (Gibson, 1988; Jamone et al., 2018). Upon grounding language describing actions in its embodied experience, a language learner's planning (Driess et al., 2023) and reasoning (Zellers et al., 2021) capabilities are enhanced. For instance, it should be able to judge that a paper plate makes a better frisbee than a ceramic one (Bisk et al., 2020a) or that, upon boiling butter, it should be poured into a jar and not a plate (Bisk et al., 2020b).

3.3 Metaphors & Abstract Concepts

Much of the language contained in corpora is figurative. Yet, the meaning behind metaphors is derived from experiencing the world (Engstrø, 2000; Bisk et al., 2020a). Thus, understanding figurative language is challenging to disembodied language models (Liu et al., 2022; Wicke, 2023). In addition, through metaphors, abstract concepts can be represented by concrete ones (Engstrø, 2000; Feldman and Narayanan, 2004). For instance, the metaphor *"similarity is proximity"* connects the abstract *similarity* to the concrete *proximity* (Casasanto and Gijssels, 2015). Hence, embodiment can enhance the understanding of abstract concepts.

With these dimensions of understanding opened up, embodied language learning enables a more robust language understanding in the context of the physical world, which is crucial for applications such as language-supported robots (Taniguchi et al., 2019). It provides the opportunity for agents to be better equipped to follow human instructions (Shridhar et al., 2020; Zhang and Chai, 2021; Suglia et al., 2021; Nguyen et al., 2021; Padmakumar et al., 2022; Gao et al., 2022; Blukis et al., 2022), answer human inquiries (Das et al., 2018; Gordon et al., 2018; Wijmans et al., 2019), or engage in robust human-robot interactions (Tellex et al., 2020).

4 Challenges & Corresponding Efforts

4.1 Data Scarcity

One major challenge in embodied language learning is data scarcity (Wang et al., 2019, 2020; Vuong et al., 2023). Unlike text or image datasets, embodied research calls for ego-centric data from the agent's perspective within its environment (Mu et al., 2023), which is comparably limited (Duan et al., 2022; Driess et al., 2023) as it is expensive and time consuming to collect (Wang et al., 2020; Zhang et al., 2023). Embodied language learning additionally requires that data be annotated with natural language descriptions (Yang et al., 2023; Mu et al., 2023). Whether the robot is completing a task or exploring its environment, corresponding textual annotations are essential for learning to ground language (Lake and Murphy, 2020).

To address data scarcity, one adopted approach is multi-task learning (Wang et al., 2019, 2020; Reed et al., 2022; Brohan et al., 2023; Driess et al., 2023; Jiang et al., 2023). With the aim of capturing meaning that transcends specific tasks (Bender and Koller, 2020), this approach is beneficial as it enables knowledge transfer (Wang et al., 2019). Another commonly adopted approach is leveraging foundation models (Yang et al., 2023). Several works have employed pretrained language models (Majumdar et al., 2020; Suglia et al., 2021; Blukis et al., 2022; Jin et al., 2023; Jiang et al., 2023; Mu et al., 2023; Driess et al., 2023) and vision-language models (Majumdar et al., 2020; Khandelwal et al., 2022; Shridhar et al., 2022; Zitkovich et al., 2023). This approach leverages language and vision representations learned from large-scale data (Lake and Murphy, 2020; Deitke et al., 2022; Yang et al., 2023) which serve as priors to be further enhanced through fine-tuning on the limited ego-centric data available (Driess et al., 2023).

4.2 Generalizability

Learned language representations should be generalizable, detached from irrelevant features specific to training tasks or environments (Lake and Murphy, 2020; Francis et al., 2022). However, especially with data scarcity, models tend to overfit and perform poorly in unseen environments (Wang et al., 2020; Deitke et al., 2022). Embodied language learning also faces the challenge of generalizing across robot embodiments (Zhang et al., 2023), which dictate the perceptual modalities and types of actions used for interacting with the environment. These variabilities reflect back on how robots can understand and ground language.

Efforts towards generalizability have been parallel to those addressing data scarcity. Incorporating foundation models allows the agent to benefit from the broad knowledge learned during pretraining (Driess et al., 2023) and leverage the generalization capability of these models (Shah et al., 2023). Multi-task learning and/or training in multiple environments (Wang et al., 2019, 2020; Reed et al., 2022; Brohan et al., 2023; Driess et al., 2023; Jiang et al., 2023) have also been adopted. To generalize across robot embodiments, training on data from multiple robots has been experimented with (Vuong et al., 2023; Brohan et al., 2023; Driess et al., 2023). However, further research is encouraged as positive transfer was reported when robots had similar sensory and action mechanisms (Vuong et al., 2023), or when low-level actuators were trained separately for each robot embodiment (Driess et al., 2023).

4.3 Simulator Realism

With the expense of real-world data collection and robot training, exploiting simulators that mimic world dynamics has been a cheaper alternative (Francis et al., 2022). However, several challenges arise with regard to the realism of simulators (Duan et al., 2022). Fewer simulators have photo-realistic scenes (Chang et al., 2017; Li et al., 2022a) compared to synthetic ones (Kolve et al., 2017; Puig et al., 2018; Wu et al., 2018; Gao et al., 2019; Kim et al., 2020; Gan et al., 2021; Puig et al., 2024). Simulated physics is also usually simplified to basic interactions (Duan et al., 2022), limiting the scope of language that agents can learn to ground. In addition, most simulators (Chang et al., 2017; Puig et al., 2018; Wu et al., 2018; Kim et al., 2020; Li et al., 2022a; Puig et al., 2024) do not include audio or tactile modalities, despite their significant role in language acquisition (Tatiya et al., 2023; Zhang et al., 2023). Despite efforts towards simulating more advanced physics (Seita et al., 2021; Gan et al., 2021; Li et al., 2022a; Fu et al., 2023) and non-visual modalities (Gan et al., 2021; Chen et al., 2022; Gao et al., 2023), only a few corresponding datasets exist (Mees et al., 2022; Gong et al., 2023). Several simulators (Chang et al., 2017; Puig et al., 2018; Wu et al., 2018; Kim et al., 2020) also discretize robot actions and restrict their granularity (Duan et al., 2022), presenting the risk of models overfitting to simplified dynamics (Francis et al., 2022), unrepresentative of the real world.

5 Gaps & Future Directions

5.1 Towards More Stringent Evaluation

Whether models have learned to ground language is implicitly assessed using task-related metrics (Li et al., 2022b; Gao et al., 2022; Zitkovich et al., 2023; Jin et al., 2023; Brohan et al., 2023). However, models can learn tasks by overfitting to spurious statistical patterns in their training data (Lake and Murphy, 2020; Zellers et al., 2021; Deitke et al., 2022). Hence, rigorous evaluation and model probing are needed to ascertain what the model has captured (Bender and Koller, 2020).

One avenue to explore is novel concept grounding. Zellers et al. (2021) pretrained their language model on data from which they removed all presence of certain words to assess if the final model learns to ground them. However, pretraining the language model from scratch renders this approach expensive. In Jiang et al. (2023), novel concepts were introduced using dummy labels, but most tested concept categories only required visual grounding. Similar experiments focused on grounding concepts in actions can be valuable.

5.2 Towards Enhanced Language Modeling

The focus of most embodied language learning works has not been on enhancing general language modeling capabilities. Research (Zhang and Chai, 2021; Jang et al., 2022; Brohan et al., 2023; Jiang et al., 2023; Liu et al., 2023) has focused on models that learn to ground language only to execute the respective tasks. In some works (Jin et al., 2023; Mu et al., 2023; Driess et al., 2023), the embodied system incorporates a generative language model that breaks down language instructions into subgoals that are then executed by lower-level control policies. In CogLoop (Jin et al., 2023) and EmbodiedGPT (Mu et al., 2023), however, the pretrained language model is frozen during the end-to-end system training. Hence, its language modeling capabilities do not benefit from the embodied training. In PaLM-E (Driess et al., 2023), while the control policies are used off-the-shelf, the language model is fine-tuned. Nevertheless, a catastrophic forgetting of its general language modeling capabilities was reported upon such embodied fine-tuning, especially for smaller size models (Driess et al., 2023).

There are recent research attempts towards an end-to-end trained system that can output both robot actions and text, such as Gato (Reed et al., 2022) and RT2-PaLM-E (Zitkovich et al., 2023). However, Gato (Reed et al., 2022) was only qualitatively tested on language generation and reported to exhibit a poor performance. RT2-PaLM-E (Zitkovich et al., 2023) was end-to-end fine-tuned for chain-of-thought reasoning, but the full range of its language modeling capabilities upon this fine-tuning was not assessed.

We suggest that an embodied language model should not only retain but also enhance the language modeling capabilities of its disembodied versions. This can then be tested on datasets evaluating figurative language understanding (Liu et al., 2022), physical reasoning abilities (Bisk et al., 2020b; Aroca-Ouellette et al., 2021; Zellers et al., 2021; He et al., 2023; Lanchantin et al., 2023; Li et al., 2023), or general language benchmarks; hence, providing insights into the effect of embodiment.

5.3 Towards a Full Embodiment Experience

Although many attributes and action effects cannot be perceived through vision alone (Tatiya et al., 2023; Zhang et al., 2023), only a few works (Heinrich et al., 2020; Tatiya et al., 2023; Zhang et al., 2023) consider other sensory modalities. From another perspective, an embodied agent's actions should not be restricted to a predefined set (Bisk et al., 2020a), as is the case in most works (Zellers et al., 2021; Pashevich et al., 2021; Zhang and Chai, 2021; Blukis et al., 2022; Zhang et al., 2023). Agents should freely interact with the environment and acquire new behaviors (Tatiya et al., 2023). However, it was reported that the pretraining and co-fine-tuning of RT-2 (Zitkovich et al., 2023) on vision-language and robotic datasets was still unable to elicit new motions from the agent. With the significance of fully exploiting the embodiment experience (Heinrich et al., 2020), further corresponding research efforts are encouraged.

6 Conclusion

Despite the associated challenges of data scarcity, generalizability, and simulator realism, learning language through embodiment is crucial for establishing the connection between language and the world. With the opportunities it holds for an enhanced understanding of attributes, world dynamics and affordances, as well as metaphors and abstract concepts, embodied language learning enables a robust language understanding that is essential for language-supported robots. By identifying current research gaps from a language understanding perspective, this article aims to motivate future efforts towards fully exploiting the embodiment experience and more rigorously evaluating embodied language models for their language modeling capabilities.

7 Limitations

This article presents an overview of embodied language learning from the point of view of language understanding research. Specific details of the robot learning techniques and task-specific evaluation metrics adopted for embodied exploration and embodied instruction following (referred to in §2) are out of this article's scope. Interested readers are directed to pertinent surveys such as in Francis et al. (2022) and Duan et al. (2022).

8 Statement of Ethics & Risks

Authors do not foresee any ethical concerns or potential risks associated with this work.

References

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*, pages 3674–3683.

Stéphane Aroca-Ouellette, Cory Paik, Alessandro Roncone, and Katharina Kann. 2021. PROST: Physical reasoning about objects through space and time. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4597–4608, Online. Association for Computational Linguistics.

Lawrence W Barsalou. 2008. Grounded cognition. *Annu. Rev. Psychol.*, 59:617–645.

Lawrence W Barsalou and Katja Wiemer-Hastings. 2005. Situating abstract concepts. *Grounding cognition: The role of perception and action in memory, language, and thought*, pages 129–163.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. 2020a. Experience grounds language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8718–8735, Online. Association for Computational Linguistics.

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. 2020b. PIQA: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7432–7439.

Valts Blukis, Chris Paxton, Dieter Fox, Animesh Garg, and Yoav Artzi. 2022. A persistent spatial semantic representation for high-level natural language instruction execution. In *Proceedings of the 5th Conference on Robot Learning*, volume 164 of *Proceedings of Machine Learning Research*, pages 706–717. PMLR.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, S. Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen A. Creel, Jared Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren E. Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas F. Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, O. Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir P. Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Benjamin Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, J. F. Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Robert Reich, Hongyu Ren, Frieda Rong, Yusuf H. Roohani, Camilo Ruiz, Jack Ryan, Christopher R'e, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishna Parasuram Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei A. Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alexander Herzog, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Tomas Jackson, Sally Jesmonth, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Kuang-Huei Lee, Sergey Levine, Yao Lu, Utsav Malla, Deeksha Manjunath, Igor Mordatch, Ofir Nachum, Carolina Parada, Jodilyn Peralta, Emily Perez, Karl Pertsch, Jornell Quiambao, Kanishka Rao, Michael S Ryoo, Grecia Salazar, Pannag R Sanketi, Kevin Sayed, Jaspiar Singh, Sumedh Sontakke, Austin Stone, Clayton Tan, Huong Tran, Vincent Vanhoucke, Steve Vega, Quan H Vuong, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. 2023. RT-1: Robotics transformer for real-world control at scale. In *Proceedings of Robotics: Science and Systems*, Daegu, Republic of Korea.

Daniel Casasanto and Tom Gijssels. 2015. What makes a metaphor an embodied metaphor? *Linguistics Vanguard*, 1(1):327–337.

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)*.

Changan Chen, Carl Schissler, Sanchit Garg, Philip Kobernik, Alexander Clegg, Paul Calamia, Dhruv Batra, Philip Robinson, and Kristen Grauman. 2022. SoundSpaces 2.0: A simulation platform for visual-acoustic learning. In *Advances in Neural Information Processing Systems*, volume 35, pages 8896–8911. Curran Associates, Inc.

Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. 2018. Embodied question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–10.

Matt Deitke, Dhruv Batra, Yonatan Bisk, Tommaso Campari, Angel X. Chang, Devendra Singh Chaplot, Changan Chen, Claudia Pérez D'Arpino, Kiana Ehsani, Ali Farhadi, Li Fei-Fei, Anthony Francis, Chuang Gan, Kristen Grauman, David Hall, Winson Han, Unnat Jain, Aniruddha Kembhavi, Jacob Krantz, Stefan Lee, Chengshu Li, Sagnik Majumder, Oleksandr Maksymets, Roberto Martín-Martín, Roozbeh Mottaghi, Sonia Raychaudhuri, Mike Roberts, Silvio Savarese, Manolis Savva, Mohit Shridhar, Niko Sünderhauf, Andrew Szot, Ben Talbot, Joshua B. Tenenbaum, Jesse Thomason, Alexander Toshev, Joanne Truong, Luca Weihs, and Jiajun Wu. 2022. Retrospectives on the Embodied AI workshop. *arXiv preprint arXiv:2210.06849*.

Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. 2023. PaLM-E: An embodied multimodal language model. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Jiafei Duan, Samson Yu, Hui Li Tan, Hongyuan Zhu, and Cheston Tan. 2022. A survey of Embodied AI: From simulators to research tasks. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 6(2):230–244.

Anders Engstrø. 2000. Book review: Philosophy in the flesh: The embodied mind and its challenge to western thought. *Metaphor and Symbol*, 15(4):267–274.

Susan Ervin-Tripp. 1973. Some strategies for the first two years. In *Cognitive development and acquisition of language*, pages 261–286. Elsevier.

Jerome Feldman and Srinivas Narayanan. 2004. Embodied meaning in a neural theory of language. *Brain and language*, 89(2):385–392.

Jonathan Francis, Nariaki Kitamura, Felix Labelle, Xiaopeng Lu, Ingrid Navarro, and Jean Oh. 2022. Core challenges in embodied vision-language planning. *Journal of Artificial Intelligence Research*, 74:459–515.

Haoyuan Fu, Wenqiang Xu, Ruolin Ye, Han Xue, Zhenjun Yu, Tutian Tang, Yutong Li, Wenxin Du, Jieyi Zhang, and Cewu Lu. 2023. Demonstrating RFUniverse: A multiphysics simulation platform for Embodied AI. In *Proceedings of Robotics: Science and Systems*, Daegu, Republic of Korea.

Chuang Gan, Jeremy Schwartz, Seth Alter, Damian Mrowca, Martin Schrimpf, James Traer, Julian De Freitas, Jonas Kubilius, Abhishek Bhandwaldar, Nick Haber, Megumi Sano, Kuno Kim, Elias Wang, Michael Lingelbach, Aidan Curtis, Kevin Feigelis, Daniel Bear, Dan Gutfreund, David Cox, Antonio Torralba, James J DiCarlo, Josh Tenenbaum, Josh McDermott, and Dan Yamins. 2021. ThreeDWorld: A platform for interactive multi-modal physical simulation. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1. Curran.

Ruohan Gao, Hao Li, Gokul Dharan, Zhuzhu Wang, Chengshu Li, Fei Xia, Silvio Savarese, Li Fei-Fei, and Jiajun Wu. 2023. Sonicverse: A multisensory simulation platform for embodied household agents that see and hear. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 704–711.

Xiaofeng Gao, Qiaozi Gao, Ran Gong, Kaixiang Lin, Govind Thattai, and Gaurav S Sukhatme. 2022. DialFRED: Dialogue-enabled agents for embodied instruction following. *IEEE Robotics and Automation Letters*, 7(4):10049–10056.

Xiaofeng Gao, Ran Gong, Tianmin Shu, Xu Xie, Shu Wang, and Song-Chun Zhu. 2019. VRKitchen: An interactive 3D environment for learning real life cooking tasks. In *ICML 2019 Workshop on Reinforcement Learning for Real Life*, Long Beach, CA, USA.

Eleanor J Gibson. 1988. Exploratory behavior in the development of perceiving, acting, and the acquiring of knowledge. *Annual review of psychology*, 39(1):1–42.

Arthur M Glenberg and Michael P Kaschak. 2002. Grounding language in action. *Psychonomic bulletin & review*, 9(3):558–565.

Arthur M Glenberg and David A Robertson. 2000. Symbol grounding and meaning: A comparison of high-dimensional and embodied theories of meaning. *Journal of memory and language*, 43(3):379–401.

Ran Gong, Jiangyong Huang, Yizhou Zhao, Haoran Geng, Xiaofeng Gao, Qingyang Wu, Wensi Ai, Ziheng Zhou, Demetri Terzopoulos, Song-Chun Zhu, Baoxiong Jia, and Siyuan Huang. 2023. ARNOLD: A benchmark for language-grounded task learning with continuous states in realistic 3D scenes. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20426–20438.

Daniel Gordon, Aniruddha Kembhavi, Mohammad Rastegari, Joseph Redmon, Dieter Fox, and Ali Farhadi. 2018. IQA: Visual question answering in interactive environments. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4089–4098.

Stevan Harnad. 1990. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3):335–346.

Weinan He, Canming Huang, Zhanhao Xiao, and Yongmei Liu. 2023. Exploring the capacity of pretrained language models for reasoning about actions and change. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4629–4643, Toronto, Canada. Association for Computational Linguistics.

Stefan Heinrich, Matthias Kerzel, Erik Strahl, and Stefan Wermter. 2018. Embodied multimodal interaction in language learning: The EMIL data collection. In *Proceedings of the ICDL-EpiRob Workshop on Active Vision, Attention, and Learning (ICDL-Epirob 2018 AVAL)*, page 2p.

Stefan Heinrich, Yuan Yao, Tobias Hinz, Zhiyuan Liu, Thomas Hummel, Matthias Kerzel, Cornelius Weber, and Stefan Wermter. 2020. Crossmodal language grounding in an embodied neurocognitive model. *Frontiers in Neurorobotics*, 14.

Yicong Hong, Qi Wu, Yuankai Qi, Cristian Rodriguez-Opazo, and Stephen Gould. 2021. VLN BERT: A recurrent vision-and-language BERT for navigation. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 1643–1653.

Lorenzo Jamone, Emre Ugur, Angelo Cangelosi, Luciano Fadiga, Alexandre Bernardino, Justus Piater, and José Santos-Victor. 2018. Affordances in psychology, neuroscience, and robotics: A survey. *IEEE Transactions on Cognitive and Developmental Systems*, 10(1):4–25.

Eric Jang, Alex Irpan, Mohi Khansari, Daniel Kappler, Frederik Ebert, Corey Lynch, Sergey Levine, and Chelsea Finn. 2022. BC-Z: Zero-shot task generalization with robotic imitation learning. In *Proceedings of the 5th Conference on Robot Learning*, volume 164 of *Proceedings of Machine Learning Research*, pages 991–1002. PMLR.

Yunfan Jiang, Agrim Gupta, Zichen Zhang, Guanzhi Wang, Yongqiang Dou, Yanjun Chen, Li Fei-Fei, Anima Anandkumar, Yuke Zhu, and Linxi Fan. 2023. VIMA: Robot manipulation with multimodal prompts. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 14975–15022. PMLR.

Chuhao Jin, Wenhui Tan, Jiange Yang, Bei Liu, Ruihua Song, Limin Wang, and Jianlong Fu. 2023. AlphaBlock: Embodied finetuning for vision-language reasoning in robot manipulation. *arXiv preprint arXiv:2305.18898*.

Susan S Jones, Linda B Smith, and Barbara Landau. 1991. Object properties and knowledge in early lexical learning. *Child development*, 62(3):499–516.

Apoorv Khandelwal, Luca Weihs, Roozbeh Mottaghi, and Aniruddha Kembhavi. 2022. Simple but effective: CLIP embeddings for Embodied AI. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14829–14838.

Hyounghun Kim, Abhaysinh Zala, Graham Burri, Hao Tan, and Mohit Bansal. 2020. ArraMon: A joint navigation-assembly instruction interpretation task in dynamic environments. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3910–3927, Online. Association for Computational Linguistics.

Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Matt Deitke, Kiana Ehsani, Daniel Gordon, Yuke Zhu, et al. 2017. AI2-THOR: An interactive 3D environment for visual AI. *arXiv preprint arXiv:1712.05474*.

Patricia K Kuhl. 2007. Is speech learning 'gated' by the social brain? *Developmental science*, 10(1):110–120.

Brenden M. Lake and Gregory L. Murphy. 2020. Word meaning in minds and machines. *Psychological review*.

Jack Lanchantin, Sainbayar Sukhbaatar, Gabriel Synnaeve, Yuxuan Sun, Kavya Srinet, and Arthur Szlam. 2023. A data source for reasoning embodied agents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 8438–8446.

Chengshu Li, Fei Xia, Roberto Martín-Martín, Michael Lingelbach, Sanjana Srivastava, Bokui Shen, Kent Elliott Vainio, Cem Gokmen, Gokul Dharan, Tanish Jain, Andrey Kurenkov, Karen Liu, Hyowon Gweon, Jiajun Wu, Li Fei-Fei, and Silvio Savarese. 2022a.

iGibson 2.0: Object-centric simulation for robot learning of everyday household tasks. In *Proceedings of the 5th Conference on Robot Learning*, volume 164 of *Proceedings of Machine Learning Research*, pages 455–465. PMLR.

Lei Li, Jingjing Xu, Qingxiu Dong, Ce Zheng, Xu Sun, Lingpeng Kong, and Qi Liu. 2023. Can language models understand physical concepts? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11843–11861, Singapore. Association for Computational Linguistics.

Shuang Li, Xavier Puig, Chris Paxton, Yilun Du, Clinton Wang, Linxi Fan, Tao Chen, De-An Huang, Ekin Akyürek, Anima Anandkumar, Jacob Andreas, Igor Mordatch, Antonio Torralba, and Yuke Zhu. 2022b. Pre-trained language models for interactive decision-making. In *Advances in Neural Information Processing Systems*, volume 35, pages 31199–31212. Curran Associates, Inc.

Emmy Liu, Chenxuan Cui, Kenneth Zheng, and Graham Neubig. 2022. Testing the ability of language models to interpret figurative language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4437–4452, Seattle, United States. Association for Computational Linguistics.

Evan Zheran Liu, Sahaana Suri, Tong Mu, Allan Zhou, and Chelsea Finn. 2023. Simple embodied language learning as a byproduct of meta-reinforcement learning. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Li Lucy and Jon Gauthier. 2017. Are distributional representations ready for the real world? Evaluating word vectors for grounded perceptual meaning. In *Proceedings of the First Workshop on Language Grounding for Robotics*, pages 76–85, Vancouver, Canada. Association for Computational Linguistics.

Corey Lynch, Ayzaan Wahid, Jonathan Tompson, Tianli Ding, James Betker, Robert Baruch, Travis Armstrong, and Pete Florence. 2023. Interactive language: Talking to robots in real time. *IEEE Robotics and Automation Letters*, pages 1–8.

Arjun Majumdar, Ayush Shrivastava, Stefan Lee, Peter Anderson, Devi Parikh, and Dhruv Batra. 2020. Improving vision-and-language navigation with image-text pairs from the web. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pages 259–274. Springer.

James L McClelland, Felix Hill, Maja Rudolph, Jason Baldridge, and Hinrich Schütze. 2019. Extending machine language models toward human-level language understanding. *arXiv preprint arXiv:1912.05877*.

Oier Mees, Lukas Hermann, Erick Rosete-Beas, and Wolfram Burgard. 2022. CALVIN: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks. *IEEE Robotics and Automation Letters*, 7(3):7327–7334.

Yao Mu, Qinglong Zhang, Mengkang Hu, Wenhai Wang, Mingyu Ding, Jun Jin, Bin Wang, Jifeng Dai, Yu Qiao, and Ping Luo. 2023. EmbodiedGPT: Vision-language pre-training via embodied chain of thought. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Van-Quang Nguyen, Masanori Suganuma, and Takayuki Okatani. 2021. Look wide and interpret twice: Improving performance on interactive instruction-following tasks. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 923–930. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Aishwarya Padmakumar, Jesse Thomason, Ayush Shrivastava, Patrick Lange, Anjali Narayan-Chen, Spandana Gella, Robinson Piramuthu, Gokhan Tur, and Dilek Hakkani-Tur. 2022. TEACh: Task-driven embodied agents that chat. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2017–2025.

Alexander Pashevich, Cordelia Schmid, and Chen Sun. 2021. Episodic transformer for vision-and-language navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15942–15952.

Jean Piaget, Margaret Cook, et al. 1952. *The origins of intelligence in children*, volume 8. International Universities Press New York.

Xavier Puig, Kevin Ra, Marko Boben, Jiaman Li, Tingwu Wang, Sanja Fidler, and Antonio Torralba. 2018. VirtualHome: Simulating household activities via programs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8494–8502.

Xavier Puig, Eric Undersander, Andrew Szot, Mikael Dallaire Cote, Tsung-Yen Yang, Ruslan Partsey, Ruta Desai, Alexander Clegg, Michal Hlavac, So Yeon Min, Vladimír Vondruš, Theophile Gervet, Vincent-Pierre Berges, John M Turner, Oleksandr Maksymets, Zsolt Kira, Mrinal Kalakrishnan, Jitendra Malik, Devendra Singh Chaplot, Unnat Jain, Dhruv Batra, Akshara Rai, and Roozbeh Mottaghi. 2024. Habitat 3.0: A co-habitat for humans, avatars, and robots. In *The Twelfth International Conference on Learning Representations*.

Yuankai Qi, Qi Wu, Peter Anderson, Xin Wang, William Yang Wang, Chunhua Shen, and Anton van den Hengel. 2020. REVERIE: Remote embodied visual referring expression in real indoor environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9982–9991.

Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gómez Colmenarejo, Alexander Novikov, Gabriel Barth-maron, Mai Giménez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, Tom Eccles, Jake Bruce, Ali Razavi, Ashley Edwards, Nicolas Heess, Yutian Chen, Raia Hadsell, Oriol Vinyals, Mahyar Bordbar, and Nando de Freitas. 2022. A generalist agent. *Transactions on Machine Learning Research*. Featured Certification, Outstanding Certification.

Daniel Seita, Pete Florence, Jonathan Tompson, Erwin Coumans, Vikas Sindhwani, Ken Goldberg, and Andy Zeng. 2021. Learning to rearrange deformable cables, fabrics, and bags with goal-conditioned transporter networks. In *IEEE International Conference on Robotics and Automation (ICRA)*.

Dhruv Shah, Błażej Osiński, brian ichter, and Sergey Levine. 2023. LM-Nav: Robotic navigation with large pre-trained models of language, vision, and action. In *Proceedings of The 6th Conference on Robot Learning*, volume 205 of *Proceedings of Machine Learning Research*, pages 492–504. PMLR.

Mohit Shridhar, Lucas Manuelli, and Dieter Fox. 2022. CLIPort: What and where pathways for robotic manipulation. In *Proceedings of the 5th Conference on Robot Learning*, volume 164 of *Proceedings of Machine Learning Research*, pages 894–906. PMLR.

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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10740–10749.

Jivko Sinapov, Connor Schenck, Kerrick Staley, Vladimir Sukhoy, and Alexander Stoytchev. 2014a. Grounding semantic categories in behavioral interactions: Experiments with 100 objects. *Robotics and Autonomous Systems*, 62(5):632–645.

Jivko Sinapov, Connor Schenck, and Alexander Stoytchev. 2014b. Learning relational object categories using behavioral exploration and multimodal perception. In *2014 IEEE international conference on robotics and automation (ICRA)*, pages 5691–5698. IEEE.

Linda Smith and Michael Gasser. 2005. The development of embodied cognition: Six lessons from babies. *Artificial life*, 11(1-2):13–29.

Catherine E Snow, Anjo Arlman-Rupp, Yvonne Hassing, Jan Jobse, Jan Joosten, and Jan Vorster. 1976. Mothers' speech in three social classes. *Journal of Psycholinguistic Research*, 5:1–20.

Alessandro Suglia, Qiaozi (QZ) Gao, Jesse Thomason, Govind Thattai, and Gaurav S. Sukhatme. 2021. Embodied BERT: A transformer model for embodied, language-guided visual task completion. In *EMNLP 2021 Workshop on Novel Ideas in Learning-to-Learn through Interaction*.

Tadahiro Taniguchi, Daichi Mochihashi, Takayuki Nagai, Satoru Uchida, Naoya Inoue, Ichiro Kobayashi, Tomoaki Nakamura, Yoshinobu Hagiwara, Naoto Iwahashi, and Tetsunari Inamura. 2019. Survey on frontiers of language and robotics. *Advanced Robotics*, 33(15-16):700–730.

Gyan Tatiya, Jonathan Francis, Ho-Hsiang Wu, Yonatan Bisk, and Jivko Sinapov. 2023. MOSAIC: Learning unified multi-sensory object property representations for robot perception. *arXiv preprint arXiv:2309.08508*.

Stefanie Tellex, Nakul Gopalan, Hadas Kress-Gazit, and Cynthia Matuszek. 2020. Robots that use language. *Annual Review of Control, Robotics, and Autonomous Systems*, 3:25–55.

Jesse Thomason, Jivko Sinapov, Maxwell Svetlik, Peter Stone, and Raymond J. Mooney. 2016. Learning multi-modal grounded linguistic semantics by playing "I Spy". In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, IJCAI'16, page 3477–3483. AAAI Press.

Quan Vuong, Sergey Levine, Homer Rich Walke, Karl Pertsch, Anikait Singh, Ria Doshi, Charles Xu, Jianlan Luo, Liam Tan, Dhruv Shah, Chelsea Finn, Max Du, Moo Jin Kim, Alexander Khazatsky, Jonathan Heewon Yang, Tony Z. Zhao, Ken Goldberg, Ryan Hoque, Lawrence Yunliang Chen, Simeon Adebola, Gaurav S. Sukhatme, Gautam Salhotra, Shivin Dass, Lerrel Pinto, Zichen Jeff Cui, Siddhant Haldar, Anant Rai, Nur Muhammad Mahi Shafiullah, Yuke Zhu, Yifeng Zhu, Soroush Nasiriany, Shuran Song, Cheng Chi, Chuer Pan, Wolfram Burgard, Oier Mees, Chenguang Huang, Deepak Pathak, Shikhar Bahl, Russell Mendonca, Gaoyue Zhou, Mohan Kumar Srirama, Sudeep Dasari, Cewu Lu, Hao-Shu Fang, Hongjie Fang, Henrik I Christensen, Masayoshi Tomizuka, Wei Zhan, Mingyu Ding, Chenfeng Xu, Xinghao Zhu, Ran Tian, Youngwoon Lee, Dorsa Sadigh, Yuchen Cui, Suneel Belkhale, Priya Sundaresan, Trevor Darrell, Jitendra Malik, Ilija Radosavovic, Jeannette Bohg, Krishnan Srinivasan, Xiaolong Wang, Nicklas Hansen, Yueh-Hua Wu, Ge Yan, Hao Su, Jiayuan Gu, Xuanlin Li, Niko Suenderhauf, Krishan Rana, Ben Burgess-Limerick, Federico Ceola, Kento Kawaharazuka, Naoaki Kanazawa, Tatsuya Matsushima, Yutaka Matsuo, Yusuke Iwasawa, Hiroki Furuta, Jihoon Oh, Tatsuya Harada, Takayuki Osa, Yujin Tang, Oliver Kroemer, Mohit Sharma, Kevin Lee Zhang, Beomjoon Kim, Yoonyoung Cho, Junhyek Han, Jaehyung Kim, Joseph J Lim, Edward Johns, Norman Di Palo, Freek Stulp, Antonin Raffin, Samuel Bustamante, João Silvério, Abhishek Padalkar, Jan Peters, Bernhard Schölkopf, Dieter Büchler, Jan Schneider, Simon Guist, Jiajun Wu, Stephen Tian, Haochen Shi, Yunzhu Li, Yixuan Wang, Mingtong Zhang, Heni Ben Amor, Yifan Zhou, Keyvan Majd, Lionel Ott, Giulio Schiavi, Roberto Martín-Martín, Rutav Shah, Yonatan Bisk, Jeffrey T Bingham, Tianhe Yu, Vidhi Jain, Ted Xiao, Karol Hausman, Christine Chan, Alexander Herzog, Zhuo

Xu, Sean Kirmani, Vincent Vanhoucke, Ryan Julian, Lisa Lee, Tianli Ding, Yevgen Chebotar, Jie Tan, Jacky Liang, Igor Mordatch, Kanishka Rao, Yao Lu, Keerthana Gopalakrishnan, Stefan Welker, Nikhil J Joshi, Coline Manon Devin, Alex Irpan, Sherry Moore, Ayzaan Wahid, Jialin Wu, Xi Chen, Paul Wohlhart, Alex Bewley, Wenxuan Zhou, Isabel Leal, Dmitry Kalashnikov, Pannag R Sanketi, Chuyuan Fu, Ying Xu, Sichun Xu, brian ichter, Jasmine Hsu, Peng Xu, Anthony Brohan, Pierre Sermanet, Nicolas Heess, Michael Ahn, Rafael Rafailov, Acorn Pooley, Kendra Byrne, Todor Davchev, Kenneth Oslund, Stefan Schaal, Ajinkya Jain, Keegan Go, Fei Xia, Jonathan Tompson, Travis Armstrong, and Danny Driess. 2023. Open X-Embodiment: Robotic learning datasets and RT-X models. In *Towards Generalist Robots: Learning Paradigms for Scalable Skill Acquisition @ CoRL2023*.

Xin Eric Wang, Vihan Jain, Eugene Ie, William Yang Wang, Zornitsa Kozareva, and Sujith Ravi. 2019. Natural language grounded multitask navigation. In *ViGIL@NeurIPS*.

Xin Eric Wang, Vihan Jain, Eugene Ie, William Yang Wang, Zornitsa Kozareva, and Sujith Ravi. 2020. Environment-agnostic multitask learning for natural language grounded navigation. In *Computer Vision – ECCV 2020*, pages 413–430, Cham. Springer International Publishing.

Philipp Wicke. 2023. LMs stand their ground: Investigating the effect of embodiment in figurative language interpretation by language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4899–4913, Toronto, Canada. Association for Computational Linguistics.

Erik Wijmans, Samyak Datta, Oleksandr Maksymets, Abhishek Das, Georgia Gkioxari, Stefan Lee, Irfan Essa, Devi Parikh, and Dhruv Batra. 2019. Embodied question answering in photorealistic environments with point cloud perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6659–6668.

Yi Wu, Yuxin Wu, Georgia Gkioxari, and Yuandong Tian. 2018. Building generalizable agents with a realistic and rich 3D environment. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Workshop Track Proceedings*. OpenReview.net.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, Rui Zheng, Xiaoran Fan, Xiao Wang, Limao Xiong, Qin Liu, Yuhao Zhou, Weiran Wang, Changhao Jiang, Yicheng Zou, Xiangyang Liu, Zhangyue Yin, Shihan Dou, Rongxiang Weng, Wensen Cheng, Qi Zhang, Wenjuan Qin, Yongyan Zheng, Xipeng Qiu, Xuanjing Huan, and Tao Gui. 2023. The rise and potential of large language model based agents: A survey. *arXiv preprint arXiv:2309.07864*.

Sherry Yang, Ofir Nachum, Yilun Du, Jason Wei, Pieter Abbeel, and Dale Schuurmans. 2023. Foundation models for decision making: Problems, methods, and opportunities. *arXiv preprint arXiv:2303.04129*.

Rowan Zellers, Ari Holtzman, Matthew Peters, Roozbeh Mottaghi, Aniruddha Kembhavi, Ali Farhadi, and Yejin Choi. 2021. PIGLeT: Language grounding through neuro-symbolic interaction in a 3D world. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2040–2050, Online. Association for Computational Linguistics.

Xiaohan Zhang, Saeid Amiri, Jivko Sinapov, Jesse Thomason, Peter Stone, and Shiqi Zhang. 2023. Multimodal embodied attribute learning by robots for object-centric action policies. *Autonomous Robots*, pages 1–24.

Yichi Zhang and Joyce Chai. 2021. Hierarchical task learning from language instructions with unified transformers and self-monitoring. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4202–4213, Online. Association for Computational Linguistics.

Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, Quan Vuong, Vincent Vanhoucke, Huong Tran, Radu Soricut, Anikait Singh, Jaspiar Singh, Pierre Sermanet, Pannag R. Sanketi, Grecia Salazar, Michael S. Ryoo, Krista Reymann, Kanishka Rao, Karl Pertsch, Igor Mordatch, Henryk Michalewski, Yao Lu, Sergey Levine, Lisa Lee, Tsang-Wei Edward Lee, Isabel Leal, Yuheng Kuang, Dmitry Kalashnikov, Ryan Julian, Nikhil J. Joshi, Alex Irpan, Brian Ichter, Jasmine Hsu, Alexander Herzog, Karol Hausman, Keerthana Gopalakrishnan, Chuyuan Fu, Pete Florence, Chelsea Finn, Kumar Avinava Dubey, Danny Driess, Tianli Ding, Krzysztof Marcin Choromanski, Xi Chen, Yevgen Chebotar, Justice Carbajal, Noah Brown, Anthony Brohan, Montserrat Gonzalez Arenas, and Kehang Han. 2023. RT-2: Vision-language-action models transfer web knowledge to robotic control. In *Proceedings of The 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pages 2165–2183. PMLR.

Ozan Özdemir, Matthias Kerzel, and Stefan Wermter. 2021. Embodied language learning with paired variational autoencoders. In *2021 IEEE International Conference on Development and Learning (ICDL)*, pages 1–6.

A Summary of Embodied Language Learning Works

Table 1 provides a brief summary of the approaches adopted and datasets used by the representative embodied language learning works discussed in Section §2.

Table 1: Summary of Adopted Approaches and Used Datasets of Representative Embodied Language Learning Works

Work	Approach			Robotics Dataset
	Embodied Exploration	Embodied Instruction Following		
		Short-Horizon Skill	Long-Horizon Goal	
Heinrich et al. (2020)	✓			EMIL (Heinrich et al., 2018)
Özdemir et al. (2021)	✓			-
Zhang et al. (2023)	✓			ISpy (Thomason et al., 2016), 100-Objects (Sinapov et al., 2014a), Sinapov et al. (2014b)
Tatiya et al. (2023)	✓			100-Objects (Sinapov et al., 2014a)
Jang et al. (2022)		✓		BC-Z
Brohan et al. (2023)		✓		RT-1
Zitkovich et al. (2023)		✓		ALFRED (Shridhar et al., 2020)
Vuong et al. (2023)		✓		Open X-Embodiment
Jiang et al. (2023)		✓	✓	VIMABench
Suglia et al. (2021)			✓	ALFRED (Shridhar et al., 2020)
Hong et al. (2021)			✓	R2R (Anderson et al., 2018), REVERIE (Qi et al., 2020)
Jin et al. (2023)			✓	AlphaBlock
Driess et al. (2023)			✓	Mobile Manipulator, Language Table (Lynch et al., 2023), TAMP

Note. Datasets without references are original to their corresponding works.