Adapting Large Pre-Trained Models for Generalized Robot Reasoning and Control

Ishika Singh

I. Motivation

Deploying personal home robots is challenging because of a generalization bottleneck. Current machine learning approaches for embodied reasoning and robot control often fail to adapt to new tasks, objects, or environments beyond their training data. My research focuses on overcoming this limitation by integrating Large Pre-Trained Models (LPTMs) of language and vision with embodied reasoning frameworks, aiming to enable more effective adaptation and generalization of robot behaviors. For robots to be widely useful, they must handle both: **1) high-level reasoning**, such as performing semantically logical sequences of actions, predicting object affordances, and assessing the environment state to make plans; and **2) low-level control**, where, for example, the robot should be capable of turning a stove knob, regardless of knob size, the kitchen's backdrop, the counter's texture, or under a dim evening light.

Generalization is difficult because large-scale robotics data to train machine learning models is limited. My research tackles this limited availability of data by leveraging pre-trained language [5] and vision [16] models, which implicitly store vast human knowledge. Models like ChatGPT for complex language reasoning and Diffusion [2, 15] or Segmentation [19] models for fine-grained visual understanding, demonstrate the generalization potential of these LPTMs. They learn a broad latent representation space that can encode practically any text or image. However, simply integrating these models into robots does not automatically result in intelligent behavior. The challenge lies in aligning the expansive latent space of these models with the robot's constrained action space, which is shaped by its physical embodiment, current state, and environment. The following sections discuss contributed research and proposed solutions for achieving both high-level and low-level generalization in robots through the use of LPTMs.

II. Research Contribution

A. *General high-level reasoning*

High-level task planning can require defining myriad domain knowledge about the world in which a robot needs to act. To ameliorate that effort, large language models (LLMs) can be used to score potential next actions during task planning, and even generate action sequences directly, given an instruction in natural language with no additional domain information. However, such methods either require enumerating all possible next steps for scoring, or generate free-form text that may contain actions not possible on a given robot in its current context [1, 10, 11]. We developed a programmatic LLM prompt structure that enables plan generation functional across situated environments, objects, robot capabilities, and tasks [23, 22]. The key insight was to prompt the LLM with program-like specifications of the available robot actions and objects in an environment, as well as with example programs that can be executed. We made concrete recommendations about prompt structure and generation constraints through ablation experiments, demonstrate state-of-the-art success rates in VirtualHome [17] household tasks, and zero-shot deploy our method on a physical robot arm for tabletop tasks. Our method has been well-received by the robotics as well as vision-language communities, and has become a standard approach for open-domain task planning, with over 700 citations from ICRA conference and AuRo journal articles and 100 GitHub stars within two years of publication.

While LLMs can use commonsense reasoning to assemble action sequences, relying on them to infer plan steps doesn't guarantee execution success, especially in complex multiagent scenarios. In contrast, classical planning methods like Planning Domain Definition Language (PDDL) generate action sequences that achieve a goal if possible, given an initial state. However, vanilla PDDL lacks temporal reasoning, such as coordinating simultaneous tasks. For example, cutting a tomato and toasting bread can occur in parallel, but assembling a sandwich requires both to be completed first. A human can decompose goals into parallelizable subgoals while maintaining necessary dependencies. Our work combined symbolic planning with LLMs to approximate human-like two-agent goal decomposition [25], enabling faster planning than multi-agent PDDL while requiring fewer execution steps than single-agent planning—all while preserving execution success. However, symbolic planning requires predefined domain formulations, introducing overhead in new environments. Future work will explore how LPTMs can mitigate this by autonomously defining new domains.

These works have also been featured in the Scientific American and Communications of the ACM discussing putting LLMs into embodied robot bodies in a broader context. The articles point out that concerns about LLM-powered robots are premature, because challenges persist in low-level control and robust execution of actions predicted by LLMs, such as picking up an object or opening a cabinet under diverse visual scenarios. My next works directly addresses this bottleneck by introducing a comprehensive benchmark to evaluate generalization of low-level robot action models as well as developing

such action models.

B. General low-level control

To achieve effective large-scale, real-world robot applications, it is crucial to evaluate how well low-level control policies adapt to environmental changes. Many studies, however, test robots in environments similar to or identical to their training setups [20, 8, 9]. The lack of an agreed-upon, comprehensive evaluation benchmarks in robotics hampers progress in developing robust action models. We introduce a novel simulation benchmark [24] featuring 20 diverse manipulation tasks and 14 axes of systematic environmental perturbations, such as variations in color, texture, size, lighting, and camera pose. It allows us to compare five state-of-the-art manipulation models, revealing a 30-50% decrease in success rates across perturbations compared to unperturbed conditions, with degradation exceeding 75% when multiple perturbations are combined. We empirically show that perturbations affecting the number of distractor objects, target object color, or lighting have the greatest impact on performance. Our simulation results, which correlate with real-world experiment results ($\bar{R}^2 = 0.614$), underscore the benchmark's ecological validity. We have open-sourced the dataset and code along with 3D printed object files used in real-world tests. The robotics community has shown excitement towards our benchmark, with new control model preprints, demonstrating improved generalization on certain perturbation factors, already available [21, 18] and over 80 GitHub stars within an year of publication. Ultimately, we hope that this benchmark will serve as a platform to identify modeling decisions that improve generalization for low-level robot action models.

In the next work, we developed a control model that can execute learned tasks across unseen objects and environments without degradation. We address the challenge of mapping natural language instructions and multi-view RGBD observations to quasistatic robot actions. 3D-aware robot policies achieve state-of-the-art performance on precise robot manipulation tasks, but struggle with generalization to unseen instructions, scenes, and object variations [12, 20, 8, 9, 6]. On the other hand, Vision Language Action (VLA) models, built with LPTMs, excel in generalization across instructions and scenes, but can be sensitive to camera and robot pose variations [13, 14, 4, 3]. We explore how we can leverage prior knowledge embedded in LPTMs to improve generalization of 3D-aware keyframe policies [26]. We introduce a novel architecture and learning framework that combines the generalization strengths of VLAs with the robustness of 3D-aware policies. We project input observations from diverse views into a point cloud that is then rendered from canonical orthographic views, ensuring input view invariance and consistency between input and output spaces. These canonical views are processed with a vision backbone, an LLM, and an image diffusion model to generate images that encode the next position and orientation of the end-effector on the input scene. We initialize our model with pre-trained models and trained end-to-end such that the LPTMs work together to solve the task. Our system can perform challenging 3D reasoning tasks like 'open the drawer to 50%' with changing scenes and objects. Evaluations on the ARNOLD benchmark [7] demonstrate state-of-the-art multi-task generalization to unseen environments while maintaining robust performance in seen settings. Our results show the potential of combining LPTMs with 3D-aware visual processing for achieving improved performance and generalization on robotic manipulation tasks.

III. Future Directions

Systems for general high-level reasoning. Symbolic planning relies on expert-annotated action semantics to generate action sequences that achieve a specified goal. These annotations define the environment dynamics, with planners systematically exploring the state space based on executable actions. While such symbolic planning guarantees successful plan execution if the goal is feasible, defining the domain for every new environment is labor-intensive and requires human expertise. I aim to develop systems that automatically build logical domains and action abstractions in new environments, using LLM-predicted approximate plans like shown in my prior work [23]. By observing successful and failed interactions, environment changes, and incorporating human feedback, these systems could extract the functioning of the environment. In our preliminary work [27], we explore automatic domain building for symbolic environments. LPTMs, including vision-language and segmentation models, can dynamically add symbols like new objects or robot capabilities to adapt to changing domains in the real world.

Systems for general low-level control. To ensure robust and general low-level control, a robot action model must be resilient to both variations in input observations and differences in output action space. Given the limited availability of robotics data, it is crucial to develop models capable of leveraging learning controls across different robot setups, tasks, and embodiments. Additionally, these action models must exhibit strong 3D scene understanding and spatial reasoning to enable reliable instruction following. To this end, I propose developing a latent observation and action space model with an LPTM backbone for enhanced generalization, along with 3D learning objectives to strengthen object and spatial reasoning. A latent space model can leverage diverse datasets, learning the underlying mapping between language instructions, visual observations, and the corresponding manipulative controls. Such a model would be applicable across various robot embodiments, enabling tasks such as turning a stove knob—regardless of the knob's size, the kitchen's backdrop, the counter's texture, or the presence of dim evening lighting.

In conclusion, I aim to develop general full-stack robotic systems that integrate LPTM's structured generations and encoded priors with suitable learning objectives for both high-level reasoning and low-level control. I believe this research direction will advance the deployment of general-purpose robots in real-world home environments.

References

[1] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, 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 J 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, and Mengyuan Yan. Do as i can, not as i say: Grounding language in robotic affordances, 2022.

[2] James Betker, Gabriel Goh, Li Jing, † TimBrooks, Jianfeng Wang, Linjie Li, † LongOuyang, † JuntangZhuang, † JoyceLee, † YufeiGuo, † WesamManassra, † PrafullaDhariwal, † CaseyChu, † YunxinJiao, and Aditya Ramesh. Improving image generation with better captions. URL https://api.semanticscholar.org/CorpusID:264403242.

[3] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Tomas Jackson, Sally Jesmonth, Nikhil J 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 Ryoo, Grecia Salazar, Pannag Sanketi, Kevin Sayed, Jaspiar Singh, Sumedh Sontakke, Austin Stone, Clayton Tan, Huong Tran, Vincent Vanhoucke, Steve Vega, Quan Vuong, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. Rt-1: Robotics transformer for real-world control at scale, 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] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

[6] Theophile Gervet, Zhou Xian, Nikolaos Gkanatsios, and Katerina Fragkiadaki. Act3d: 3d feature field transformers for multi-task robotic manipulation. In *7th Annual Conference on Robot Learning*, 2023.

[7] Ran Gong, Jiangyong Huang, Yizhou Zhao, Haoran Geng, Xiaofeng Gao, Qingyang Wu, Wensi Ai, Ziheng Zhou, Demetri Terzopoulos, Song-Chun Zhu, et al. Arnold: A benchmark for language-grounded task learning with continuous states in realistic 3d scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[8] Ankit Goyal, Jie Xu, Yijie Guo, Valts Blukis, Yu-Wei Chao, and Dieter Fox. Rvt: Robotic view transformer for 3d object manipulation. *Proceedings of the 7th Conference on Robot Learning (CoRL)*, 2023.

[9] Ankit Goyal, Valts Blukis, Jie Xu, Yijie Guo, Yu-Wei Chao, and Dieter Fox. Rvt-2: Learning precise manipulation from few demonstrations. *arXiv preprint arXiv:2406.08545*, 2024.

[10] Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. *arXiv preprint arXiv:2201.07207*, 2022.

[11] Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, Pierre Sermanet, Noah Brown, Tomas Jackson, Linda Luu, Sergey Levine, Karol Hausman, and Brian Ichter. Inner monologue: Embodied reasoning through planning with language models. In *arXiv preprint arXiv:2207.05608*, 2022.

[12] Stephen James, Kentaro Wada, Tristan Laidlow, and Andrew J Davison. Coarse-to-fine q-attention: Efficient learning for visual robotic manipulation via discretisation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13739–13748, 2022.

[13] 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.

[14] Xiang Li, Cristina Mata, Jongwoo Park, Kumara Kahatapitiya, Yoo Sung Jang, Jinghuan Shang, Kanchana Ranasinghe, Ryan Burgert, Mu Cai, Yong Jae Lee, and Michael S. Ryoo. Llara: Supercharging robot learning data for vision-language policy. In *International Conference on Learning Representations*, 2025.

[15] Yixin Liu, Kai Zhang, Yuan Li, Zhiling Yan, Chujie Gao, Ruoxi Chen, Zhengqing Yuan, Yue Huang, Hanchi Sun, Jianfeng Gao, Lifang He, and Lichao Sun. Sora: A review on background, technology, limitations, and opportunities of large vision models. *arXiv preprint arXiv:2402.17177*, 2024.

[16] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Ale-

man, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report. *arXiv preprint arXiv: 2303.08774*, 2024.

[17] Xavier Puig, Kevin Ra, Marko Boben, Jiaman Li, Tingwu Wang, Sanja Fidler, and Antonio Torralba. Virtualhome: Simulating household activities via programs. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8494–8502, 2018.

[18] Shengyi Qian, Kaichun Mo, Valts Blukis, David F. Fouhey, Dieter Fox, and Ankit Goyal. 3d-mvp: 3d multiview pretraining for robotic manipulation. *arXiv preprint arXiv:2406.18158*, 2024.

[19] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

[20] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Perceiver-actor: A multi-task transformer for robotic manipulation. In *Proceedings of the 6th Conference on Robot Learning (CoRL)*, 2022.

[21] Mohit Shridhar, Yat Long Lo, and Stephen James. Generative image as action models. In *Proceedings of the 8th Conference on Robot Learning (CoRL)*, 2024.

[22] **Ishika Singh**, Valts Blukis, Arsalan Mousavian, Ankit Goyal, Danfei Xu, Jonathan Tremblay, Dieter Fox, Jesse Thomason, and Animesh Garg. Progprompt: program generation for situated robot task planning using large language models. *Auton. Robots*, 47(8): 999–1012, 8 2023. ISSN 0929-5593. doi: 10.1007/s10514-023-10135-3.

[23] **Ishika Singh**, Valts Blukis, Arsalan Mousavian, Ankit

Goyal, Danfei Xu, Jonathan Tremblay, Dieter Fox, Jesse Thomason, and Animesh Garg. Progprompt: Generating situated robot task plans using large language models. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11523–11530, 2023. doi: 10.1109/ICRA48891.2023.10161317.

[24] **Ishika Singh***, Wilbert Pumacay*, Jiafei Duan*, Ranjay Krishna, Jesse Thomason, and Dieter Fox. The colosseum: A benchmark for evaluating generalization for robotic manipulation. In *Robotics: Science and Systems*, 2024.

[25] **Ishika Singh**, David Traum, and Jesse Thomason. Twostep: Multi-agent task planning using classical planners and large language models. *arXiv preprint arXiv:2403.17246*, 2024.

[26] **Ishika Singh**, Ankit Goyal, Stan Birchfield, Dieter Fox, Animesh Garg, and Valts Blukis. Og-vla: 3d-aware vision language action modeling via orthographic image generation. In *submission at CoRL*, 2025.

[27] Wang Zhu, **Ishika Singh**, Robin Jia, and Jesse Thomason. Language models can infer action semantics for classical planners from environment feedback. *NAACL*, 2025.