Towards Interpretable Foundation Models of Robot Behavior: A Task Specific Policy Generation Approach

Isaac Sheidlower
Isaac.Sheidlower@tufts.edu
School of Engineering
Tufts University

Reuben Aronson
Reuben.Aronson@tufts.edu
School of Engineering
Tufts University

Elaine Schaertl Short
Elaine.Short@tufts.edu
School of Engineering
Tufts University

Abstract

Foundation models are a promising path toward general-purpose and user-friendly robots. The prevalent approach involves training a "generalist policy" that, like a reinforcement learning policy, uses observations to output actions. Although this approach has seen much success, several concerns arise when considering deployment and end-user interaction with these systems. In particular, the lack of modularity between tasks means that when model weights are updated (e.g., when a user provides feedback), the behavior in other, unrelated tasks may be affected. This can negatively impact the system's interpretability and usability. We present an alternative approach to the design of robot foundation models, Diffusion for Policy Parameters (DPP), which generates stand-alone, task-specific policies. Since these policies are detached from the foundation model, they are updated only when a user wants, either through feedback or personalization, allowing them to gain a high degree of familiarity with that policy. We demonstrate a proof-of-concept of DPP in simulation then discuss its limitations and the future of interpretable foundation models.

1 Introduction

Current efforts in creating task-generalizable, novice-friendly robots are largely focused on foundational models of robot behavior. The goal of such a model is to have a user ask the robot to do an arbitrary task via verbal or non-verbal communication, then have the robot perform the task with little to no further human supervision or intervention. The relatively few robot foundation models that exist can all be categorized as "generalist robot policies." In particular, the input is an observation of the robot's state combined with a language or goal embedding that specifies the task, and the output is a robot action, such as end-effector displacement. This is the same input-output relationship as a typical task-specific Reinforcement Learning (RL) policy.

However, these generalist policies have limitations that may pose serious problems when actually deployed for users. In deployment, a foundation model should be able to learn from data in many different environments and tasks and personalize to individual users in response to training. However, generalist policies are not localized with respect to task: new feedback for one task could change model behavior in a completely unrelated task. This property limits the ability of users to personalize the system or learn what to expect of its behavior for a given task. In this work, we propose an alternative approach where the foundation model is a policy generator, which outputs standalone, task-specific policies. We discuss potential benefits of this approach to robot foundation models as well as potential challenges with generalist policies as a sole solution.

We present Diffusion for Policy Parameters (DPP), a method for generating standalone task policies conditioned on a task specification. We present a proof-of-concept implementation for smaller grid-world tasks. We show, to the best of our knowledge, for the first time that one can learn representations to generate policies directly in parameter space, without the need for policy search. Finally, we discuss limitations of the DPP approach and how to both potentially resolve them and other alternative methods towards generating task-specific policies. Our results show that DPP is a promising approach to robot behavior foundation models and warrants further investigation.

2 Related Work

Recent technological advancements in Generative AI (Gen-AI) including the transformer architecture (Vaswani et al., 2023), and diffusion models (Ho et al., 2020), that are both high performing and scale to large amounts of data, have allowed for the development of larger and more general purpose task models. In the case, of language prediction, Large Language Models (LLMs) such as GPT-4 (OpenAI et al., 2024) and LLaMa (Touvron et al., 2023), have enabled a single large-scale model to perform a variety of linguistic tasks

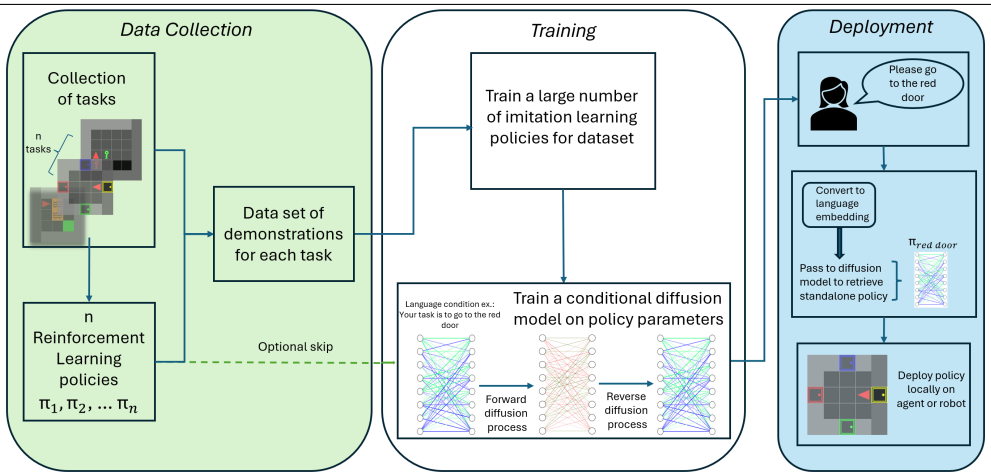

Figure 1: The DPP foundation model of robot behavior design approach

(Zhao et al., 2023). Similarly, multi-model models such as CLIP (Radford et al., 2021) and Latent Diffusion Models (LDMs) (Rombach et al., 2022), have allowed for similar generalizability in image-language embedding and language-conditioned image generation respectively. These technologies are starting to be applied for general robotic manipulation. Most notable are the large robot behavior policies RT-1/2-X (Collaboration et al., 2024; Brohan et al., 2023) and Octo (Ghosh et al.), and diffusion architectures for fast imitation learning across a variety of tasks (Chi et al., 2024; Ze et al., 2024). These models are policies which transform observations to actions, and, while these models are promising and are beginning to see impressive success across many tasks, there is little modularity between learning and executing across different tasks.

RL has previously had success for both robot manipulation (Ibarz et al., 2021; Nguyen & La, 2019) and being applied to human-robot-interaction (HRI) scenarios (Reddy et al., 2018; Akalin & Loutfi, 2021; Park et al., 2019). When a user interacts with a consistent and self-contained policy, they can better predict its behavior (Cruz et al., 2023; Horter et al., 2023), customize and personalize its behavior through feedback (Bobu et al., 2021; Arzate Cruz & Igarashi, 2020; Brawer et al., 2023; Sheidlower et al., 2024) and learn to leverage its dynamics to accomplish novel tasks (Aronson & Short, 2024; Gopinath et al., 2017). This work on how users interact with RL systems highlight the value of explainable and modular policies for positive outcomes.

3 Potential Challenges with Robot Foundation Models as Generalist Policies

Robots should be able to learn from feedback and have real-time behavior personalization for any given task. If the policy the user is interacting with is a generalist robot policy, two problems may limit a user's ability to do this. The first is that when a user teaches the robot a new task or personalizes the behavior for a certain task, the behavior in separate and unrelated tasks may be affected. This may jeopardize the interpretability and legibility of the system (Bobu et al., 2024). Another is that updates to the base of the model from the organization which developed the model may have downstream affects on specific tasks/robot behavior that may be unnexpected or undesired by a user. This is already the case with consumer-available LLMs such as ChatGPT, however, in the case of robotics, the consistency of the robot's behavior is a crucial component to the user's ability to teach and interact with the robot. In fact, robots spontaneously acting in unexpected ways around users may cause physical safety concerns beyond those posed by systems operating solely on language. Thus, making sure that a robot's task behavior is changed when and how a user wants is crucial.

4 Diffusion for Policy Parameters (DPP)

We present Diffusion for Policy Parameters (DPP), a novel approach for learning how to generate standalone policies for individual tasks. DPP alleviates some of the concerns mentioned in the prior section. We then present a proof-of-concept evaluation in a grid-world simulation. This is, to the best of our knowledge, the first generative approach for creating policies in parameter space. While policy search (Plappert et al., 2018; Taylor et al., 2007; Levine & Koltun, 2013; Kalyanakrishnan & Stone, 2009) and exploration over policy parameters (Fontaine & Nikolaidis, 2021; Mouret & Clune, 2015; Tjanaka et al., 2021) have been explored, generative AI techniques have not been used directly in parameter space.

The DPP method (Figure 1.) learns a conditional diffusion model for generating policies in in policy parameter space. The steps for DPP are: collect a dataset of language/goal conditioned tasks and a dataset of demonstrations over those task; train a large set of policies on either the demonstrations or tasks themselves;

DPP Model Architecture	
Language Embedding	bge-small-en-v1.5 (size: 384) (Xiao et al., 2023)
Noise Schedule	Cosine, 1000 steps (Nichol & Dhariwal, 2021)
Noise Type	Gaussian (Ho et al., 2020)
Model Architecture	Transformer (Vaswani et al., 2017), 48 heads, 12 depth, 768 width
Batch Size	128
Input/Output Dimensions	32x82; 32 for MLP policy hidden layer, 82 = 75 (observation size) + 7 (action size)

Table 1: DPP model architecture used in experiments

	Diffusion Sample Policy	Random Policy	Training Parameters (TP) Mean	TP Median	TP Mode	Mixture of Samples (MoS), m=4	(MoS), m=8	(MoS), m=16
Avg. Return	**0.766**±0.16	**0.198**±0.14	**0.189**±0.14	**0.205**±0.12	**0.125**±0.16	**0.816**±0.19	**0.878**±0.15	**0.886**±0.16

Table 2: Results from experiments

then train a diffusion model conditioned on the task description and takes the parameters of the policies as input. The result is a model that leads to an interaction similar to a generalist robot policy: a user asks for or demonstrates a task, and then the robot autonomously executes that task, with the option of further human-in-the-loop fine-tuning if necessary. The key difference being in DPP, a policy independent of the foundation model is generated to execute the task. To study whether DPP is a viable approach for learning to generate policies, we must show it can lead to a model which conditionally generates "good" policies.

4.1, Environment and Data Collection We ran experiments in the Minigrid environment (Chevalier-Boisvert et al., 2023) for its suite of language-conditioned tasks on which we can train many agents on in a relatively small amount of time and with limited hardware. All tasks have a similar reward structure: sparse reward with a time-step penalty, resulting in a cumulative reward between 0 and 1. To generate a large number of tasks, we took three language-conditioned tasks and made each goal specification within that task its own task. In particular, we took the environments Fetch, Go to Door, and Go to Object, and for each possible object configuration, made that a task (e.g. Go to Door contains both the "go to red key" and "go to blue box" task specification, and we treat each as a task to train an agent on). We chose the 5x5 versions of each task for computational efficiency and quicker training. We then collected many seeds for each task to ensure random goal positions and obstacles, resulting in 84 unique tasks. While these tasks are significantly simpler than in-the-wild robot tasks, they provide a wide range of separate policies to train on.

To collect policy data on these tasks, we trained 84 RL agents using PPO (Schulman et al., 2017) to optimality (achieving a mean reward $> .98$). We then trained behavior cloning (BC) agents on trajectories collected from the RL agents until they received a near-optimal average reward of $> .85$. We chose BC as opposed to RL for every agent because it was more timely to train and collect the policies. We trained approximately 1000 agents for each of these tasks, discarding tasks where BC did not achieve high reward given the allotted trajectories. This resulted in BC agents for 64 of the 84 tasks and resulted in 74,000 trained policies.

4.2, Model Design Given the dataset of policies, we trained a conditional diffusion model which takes as input a language description of the task and outputs an end-to-end policy network for that task. The model architecture and description can be found in Table 1. The architecture was largely decided on based on trial and error. However, two key decisions were necessary to effectively learn in parameter space. The first was to use an entirely transformer-based architecture, as opposed to, e.g., a U-Net architecture (Ronneberger et al., 2015; Ho et al., 2020). The other was to use the hybrid loss as proposed in (Nichol & Dhariwal, 2021). We also experimented with various loss functions based on evaluation of the generated policies, but they did not lead to high performance.

4.3, Evaluation and Results The evaluation results of the final trained model can be found in Table 2. The evaluation aims to show the model generates meaningful policies in parameter space. For each baseline, we took average performance across all 64 tasks, with 10 runs each on random seeds. "Diffusion Sample Policy" refers to a single sample from the diffusion model conditioned on the task description. We primarily compare to baselines as a means to ensure the model is not learning trivial local minima. If it is not, then we expect a single sampled policy to significantly outperform the baselines. "Random Policy" refers to an agent that takes a random action in each state. The "Mean," "Median," and "Mode" baselines refer to taking those operations on all of the parameters in the dataset for the specified task. The sample policy significantly outperforms all baselines indicating that the model is learning to generate meaningful and performant polices. A single sample, however, achieves slightly lower returns than the agents in the training set. To achieve a similar performance, we take a simple mixture approach where we sample n policies, and for each observation, take the most common output action. This is referred to as Mixture of Samples (MoS) in Table 2.

4.4, Limitations Despite promising early results, there are key limitations with the evaluation regarding extrapolating the results to real-world robots. Though diffusion models and transformers have been shown to scale well with large amounts of robot data, we have not shown this scalability with policy parameter space learning. Similarly, we emphasize that the training data needed for DPP is different than for a generalist policies: DPP requires a dataset of trained policies (which could be gathered through simulation or a cross organization effort similar to the Droid dataset (Khazatsky et al., 2024)), rather than a demonstration dataset. However, for DPP to scale and generalize across tasks, it will likely need both a policy and a demonstration dataset. We have also only demonstrated results in an environment with a discrete action space. Although some generalist policies, such as (Ghosh et al.), have had issues with discrete action spaces, we believe a robot foundation model should be able to handle both discrete and continuous actions. These limitations warrant further investigation and to be addressed in future work.

5 Discussion

While generalist robot policies as robot behavior foundation models show clear successes, they do not maintain properties of locality and explainability that would be desired for a deployed system. To limit these concerns, we presented DPP, an alternative which may alleviate some of the outlined concerns. DPP generates smaller, standalone policies for each task; this approach means that those policies are not affected by a user teaching the robot other tasks or by unwanted updates to the general foundation model.

Enabling policies to be stable and therefore more predictable is a key feature for human-usable robots. Human-robot interaction research has consistently shown that robot models need to be not just performant, but also predictable Lichtenthäler & Kirsch (2016). A predictable robot system not only improves its interpretability, but also allows a user to gain a high degree of familiarity with those policies, and in turn use them to accomplish novel tasks. In this work, we embed this predictability and usability directly into the structure of the model without compromising its flexibility to learn from new data. With foundation models in their infancy, it is an ideal time to explore how these powerful generalized models can be made more usable.

Future work could explore other methods to make foundation models more stable and usable, especially by allowing the user to choose when and how a task policy should be updated. For example, a robot might come deployed with a suite of policies for very common tasks, with the capacity to learn new tasks from the user through human-in-the-loop learning (Ravichandar et al., 2020; Liu et al., 2023). Another approach is to use sim-to-real RL to train new policies when needed by the user. For example, Eureka (Ma et al., 2024) uses LLMs and an iterative training procedure to design reward functions for arbitrary tasks. This approach has similar benefits as DPP, but depends on having an accurate simulator and may not be responsive enough for users, as the robot needs to learn tasks from scratch when the user requests it. An advantage, however, is that since it uses a task-specific reward function, it may be more explainable.

We primarily focused on improving interpretability and modularity relative to generalist policies, but there are also other exciting directions for future research towards usable generalist policies. For example, work is needed on how to explain the behavior of a generalist robot policy. Explainable AI techniques for large models are constantly improving, but more work is needed to understand how techniques for explaining robot behavior apply to generalist robot policies: much prior work in explainable robotics and AI either assumes the robot was trained with RL (see (Milani et al., 2024) for a taxonomy of explainable RL) or requires semantic knowledge of its past interactions (Setchi et al., 2020). Some approaches, such as directly generating explanations for non-interpretable policies are easily applicable to generalist robot policies, while others, such as generating intrinsically interpretable policies, may not be.

6 Conclusion

In this work, we discussed key areas for improving the paradigm of generalist robot policies as robot behavior foundation models. To better empower end-users, these models need better modularity and independence between tasks, to support a user's ability to understand and leverage the system's dynamics, independent of weight changes in the foundation model. Towards this goal, we presented the Diffusion for Policy Parameters (DPP) approach to a robot behavior foundation. Our preliminary results show that this is a promising method for generating task-specific, standalone policies conditioned on a task specification. As the research community explores methods such as DPP, we can ensure the future of foundational robot-behavior models empower end-users with a high degree of understanding and control over the robot.

Acknowledgments

The work described here was supported in part by the US National Science Foundation (IIS-2132887)

References

Neziha Akalin and Amy Loutfi. Reinforcement Learning Approaches in Social Robotics. *Sensors*, 21(4): 1292, January 2021. ISSN 1424-8220. doi: 10.3390/s21041292. URL https://www.mdpi.com/1424-8220/21/4/1292. Number: 4 Publisher: Multidisciplinary Digital Publishing Institute.

Reuben M. Aronson and Elaine Schaertl Short. Intentional User Adaptation to Shared Control Assistance. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '24, pp. 4–12, New York, NY, USA, March 2024. Association for Computing Machinery. ISBN 9798400703225. doi: 10.1145/3610977.3634953. URL https://dl.acm.org/doi/10.1145/3610977.3634953.

Christian Arzate Cruz and Takeo Igarashi. A Survey on Interactive Reinforcement Learning: Design Principles and Open Challenges. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference*, pp. 1195–1209, Eindhoven Netherlands, July 2020. ACM. ISBN 978-1-4503-6974-9. doi: 10.1145/3357236.3395525. URL https://dl.acm.org/doi/10.1145/3357236.3395525.

Andreea Bobu, Marius Wiggert, Claire Tomlin, and Anca D. Dragan. Feature Expansive Reward Learning: Rethinking Human Input. *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, pp. 216–224, March 2021. doi: 10.1145/3434073.3444667. URL http://arxiv.org/abs/2006.13208. arXiv: 2006.13208.

Andreea Bobu, Andi Peng, Pulkit Agrawal, Julie A Shah, and Anca D. Dragan. Aligning Human and Robot Representations. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '24, pp. 42–54, New York, NY, USA, March 2024. Association for Computing Machinery. ISBN 9798400703225. doi: 10.1145/3610977.3634987. URL https://dl.acm.org/doi/10.1145/3610977.3634987.

Jake Brawer, Debasmita Ghose, Kate Candon, Meiying Qin, Alessandro Roncone, Marynel Vázquez, and Brian Scassellati. Interactive Policy Shaping for Human-Robot Collaboration with Transparent Matrix Overlays. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, pp. 525–533, Stockholm Sweden, March 2023. ACM. ISBN 978-1-4503-9964-7. doi: 10.1145/3568162.3576983. URL https://dl.acm.org/doi/10.1145/3568162.3576983.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, Pete Florence, Chuyuan Fu, Montse Gonzalez Arenas, Keerthana Gopalakrishnan, Kehang Han, Karol Hausman, Alexander Herzog, Jasmine Hsu, Brian Ichter, Alex Irpan, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Lisa Lee, Tsang-Wei Edward Lee, Sergey Levine, Yao Lu, Henryk Michalewski, Igor Mordatch, Karl Pertsch, Kanishka Rao, Krista Reymann, Michael Ryoo, Grecia Salazar, Pannag Sanketi, Pierre Sermanet, Jaspiar Singh, Anikait Singh, Radu Soricut, Huong Tran, Vincent Vanhoucke, Quan Vuong, Ayzaan Wahid, Stefan Welker, Paul Wohlhart, Jialin Wu, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control, July 2023. URL http://arxiv.org/abs/2307.15818. arXiv:2307.15818 [cs].

Maxime Chevalier-Boisvert, Bolun Dai, Mark Towers, Rodrigo de Lazcano, Lucas Willems, Salem Lahlou, Suman Pal, Pablo Samuel Castro, and Jordan Terry. Minigrid & Miniworld: Modular & Customizable Reinforcement Learning Environments for Goal-Oriented Tasks. *CoRR*, abs/2306.13831, 2023.

Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion Policy: Visuomotor Policy Learning via Action Diffusion, March 2024. URL http://arxiv.org/abs/2303.04137. arXiv:2303.04137 [cs].

Open X.-Embodiment Collaboration, Abby O'Neill, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, Albert Tung, Alex Bewley, Alex Herzog, Alex Irpan, Alexander Khazatsky, Anant Rai, Anchit Gupta, Andrew Wang, Anikait Singh, Animesh Garg, Aniruddha Kembhavi, Annie Xie, Anthony Brohan, Antonin Raffin, Archit Sharma, Arefeh Yavary, Arhan Jain, Ashwin Balakrishna, Ayzaan Wahid, Ben Burgess-Limerick, Beomjoon Kim, Bernhard Schölkopf, Blake Wulfe, Brian Ichter, Cewu Lu, Charles Xu, Charlotte Le, Chelsea Finn, Chen Wang, Chenfeng Xu, Cheng Chi, Chenguang Huang, Christine Chan, Christopher Agia, Chuer Pan, Chuyuan Fu, Coline Devin, Danfei Xu, Daniel Morton, Danny Driess, Daphne Chen, Deepak Pathak, Dhruv Shah, Dieter Büchler, Dinesh Jayaraman, Dmitry Kalashnikov, Dorsa Sadigh, Edward Johns, Ethan Foster, Fangchen Liu, Federico Ceola, Fei Xia, Feiyu Zhao, Freek Stulp, Gaoyue Zhou, Gaurav S. Sukhatme, Gautam Salhotra, Ge Yan, Gilbert Feng, Giulio Schiavi, Glen Berseth, Gregory Kahn, Guanzhi Wang, Hao Su, Hao-Shu Fang, Haochen Shi, Henghui Bao, Heni Ben Amor, Henrik I. Christensen, Hiroki Furuta, Homer Walke, Hongjie Fang, Huy Ha, Igor Mordatch, Ilija Radosavovic, Isabel Leal, Jacky Liang, Jad Abou-Chakra, Jaehyung Kim, Jaimyn Drake, Jan Peters, Jan Schneider, Jasmine Hsu, Jeannette Bohg, Jeffrey Bingham, Jeffrey Wu, Jensen Gao, Jiaheng Hu, Jiajun Wu, Jialin Wu, Jiankai Sun, Jianlan Luo, Jiayuan Gu, Jie Tan, Jihoon Oh, Jimmy Wu, Jingpei Lu, Jingyun Yang, Jitendra Malik, João Silvério, Joey Hejna, Jonathan Booher, Jonathan Tompson, Jonathan Yang, Jordi Salvador, Joseph J. Lim, Junhyek Han, Kaiyuan Wang, Kanishka Rao, Karl Pertsch, Karol Hausman, Keegan Go, Keerthana Gopalakrishnan, Ken Goldberg, Kendra Byrne, Kenneth Oslund, Kento Kawaharazuka, Kevin Black, Kevin Lin, Kevin Zhang, Kiana Ehsani, Kiran Lekkala, Kirsty Ellis, Krishan Rana, Krishnan Srinivasan, Kuan Fang, Kunal Pratap Singh, Kuo-Hao Zeng, Kyle Hatch, Kyle Hsu, Laurent Itti, Lawrence Yunliang Chen, Lerrel Pinto, Li Fei-Fei, Liam Tan, Linxi "Jim" Fan, Lionel Ott, Lisa Lee, Luca Weihs, Magnum Chen, Marion Lepert, Marius Memmel, Masayoshi Tomizuka, Masha Itkina, Mateo Guaman Castro, Max Spero, Maximilian Du, Michael Ahn, Michael C. Yip, Mingtong Zhang, Mingyu Ding, Minho Heo, Mohan Kumar Srirama, Mohit Sharma, Moo Jin Kim, Naoaki Kanazawa, Nicklas Hansen, Nicolas Heess, Nikhil J. Joshi, Niko Suenderhauf, Ning Liu, Norman Di Palo, Nur Muhammad Mahi Shafiullah, Oier Mees, Oliver Kroemer, Osbert Bastani, Pannag R. Sanketi, Patrick "Tree" Miller, Patrick Yin, Paul Wohlhart, Peng Xu, Peter David Fagan, Peter Mitrano, Pierre Sermanet, Pieter Abbeel, Priya Sundaresan, Qiuyu Chen, Quan Vuong, Rafael Rafailov, Ran Tian, Ria Doshi, Roberto Martín-Martín, Rohan Baijal, Rosario Scalise, Rose Hendrix, Roy Lin, Runjia Qian, Ruohan Zhang, Russell Mendonca, Rutav Shah, Ryan Hoque, Ryan Julian, Samuel Bustamante, Sean Kirmani, Sergey Levine, Shan Lin, Sherry Moore, Shikhar Bahl, Shivin Dass, Shubham Sonawani, Shuran Song, Sichun Xu, Siddhant Haldar, Siddharth Karamcheti, Simeon Adebola, Simon Guist, Soroush Nasiriany, Stefan Schaal, Stefan Welker, Stephen Tian, Subramanian Ramamoorthy, Sudeep Dasari, Suneel Belkhale, Sungjae Park, Suraj Nair, Suvir Mirchandani, Takayuki Osa, Tanmay Gupta, Tatsuya Harada, Tatsuya Matsushima, Ted Xiao, Thomas Kollar, Tianhe Yu, Tianli Ding, Todor Davchev, Tony Z. Zhao, Travis Armstrong, Trevor Darrell, Trinity Chung, Vidhi Jain, Vincent Vanhoucke, Wei Zhan, Wenxuan Zhou, Wolfram Burgard, Xi Chen, Xiaolong Wang, Xinghao Zhu, Xinyang Geng, Xiyuan Liu, Xu Liangwei, Xuanlin Li, Yao Lu, Yecheng Jason Ma, Yejin Kim, Yevgen Chebotar, Yifan Zhou, Yifeng Zhu, Yilin Wu, Ying Xu, Yixuan Wang, Yonatan Bisk, Yoonyoung Cho, Youngwoon Lee, Yuchen Cui, Yue Cao, Yueh-Hua Wu, Yujin Tang, Yuke Zhu, Yunchu Zhang, Yunfan Jiang, Yunshuang Li, Yunzhu Li, Yusuke Iwasawa, Yutaka Matsuo, Zehan Ma, Zhuo Xu, Zichen Jeff Cui, Zichen Zhang, Zipeng Fu, and Zipeng Lin. Open X-Embodiment: Robotic Learning Datasets and RT-X Models, April 2024. URL http://arxiv.org/abs/2310.08864. arXiv:2310.08864 [cs].

Francisco Cruz, Richard Dazeley, Peter Vamplew, and Ithan Moreira. Explainable robotic systems: understanding goal-driven actions in a reinforcement learning scenario. *Neural Computing and Applications*, 35(25):18113–18130, September 2023. ISSN 1433-3058. doi: 10.1007/s00521-021-06425-5. URL https://doi.org/10.1007/s00521-021-06425-5.

Matthew Fontaine and Stefanos Nikolaidis. A Quality Diversity Approach to Automatically Generating Human-Robot Interaction Scenarios in Shared Autonomy, June 2021. URL http://arxiv.org/abs/2012.04283. arXiv:2012.04283 [cs].

Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, Jianlan Luo, You Liang Tan, Dorsa Sadigh, Chelsea Finn, and Sergey Levine. Octo: An Open-Source Generalist Robot Policy. URL https://octo-models.github.io.

Deepak Gopinath, Siddarth Jain, and Brenna D. Argall. Human-in-the-Loop Optimization of Shared Autonomy in Assistive Robotics. *IEEE Robotics and Automation Letters*, 2(1):247–254, January 2017. ISSN 2377-3766. doi: 10.1109/LRA.2016.2593928. Conference Name: IEEE Robotics and Automation Letters.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models, December 2020. URL http://arxiv.org/abs/2006.11239. arXiv:2006.11239 [cs, stat].

Tiffany Horter, Elena L. Glassman, Julie Shah, and Serena Booth. Varying How We Teach: Adding Contrast Helps Humans Learn about Robot Motions. In *HRI Workshop on Human-Interactive Robot Learning*, 2023. URL https://slbooth.com/assets/docs/papers/HRI_2023_HIRL_Learning_about_Robot_Motions_with_Variation_Theory.pdf.

Julian Ibarz, Jie Tan, Chelsea Finn, Mrinal Kalakrishnan, Peter Pastor, and Sergey Levine. How to Train Your Robot with Deep Reinforcement Learning; Lessons We've Learned. *The International Journal of Robotics Research*, 40(4-5):698–721, April 2021. ISSN 0278-3649, 1741-3176. doi: 10.1177/0278364920987859. URL http://arxiv.org/abs/2102.02915. arXiv:2102.02915 [cs].

Shivaram Kalyanakrishnan and Peter Stone. An Empirical Analysis of Value Function-Based and Policy Search Reinforcement Learning. 2009.

Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, Peter David Fagan, Joey Hejna, Masha Itkina, Marion Lepert, Yecheng Jason Ma, Patrick Tree Miller, Jimmy Wu, Suneel Belkhale, Shivin Dass, Huy Ha, Arhan Jain, Abraham Lee, Youngwoon Lee, Marius Memmel, Sungjae Park, Ilija Radosavovic, Kaiyuan Wang, Albert Zhan, Kevin Black, Cheng Chi, Kyle Beltran Hatch, Shan Lin, Jingpei Lu, Jean Mercat, Abdul Rehman, Pannag R. Sanketi, Archit Sharma, Cody Simpson, Quan Vuong, Homer Rich Walke, Blake Wulfe, Ted Xiao, Jonathan Heewon Yang, Arefeh Yavary, Tony Z. Zhao, Christopher Agia, Rohan Baijal, Mateo Guaman Castro, Daphne Chen, Qiuyu Chen, Trinity Chung, Jaimyn Drake, Ethan Paul Foster, Jensen Gao, David Antonio Herrera, Minho Heo, Kyle Hsu, Jiaheng Hu, Donovon Jackson, Charlotte Le, Yunshuang Li, Kevin Lin, Roy Lin, Zehan Ma, Abhiram Maddukuri, Suvir Mirchandani, Daniel Morton, Tony Nguyen, Abigail O'Neill, Rosario Scalise, Derick Seale, Victor Son, Stephen Tian, Emi Tran, Andrew E. Wang, Yilin Wu, Annie Xie, Jingyun Yang, Patrick Yin, Yunchu Zhang, Osbert Bastani, Glen Berseth, Jeannette Bohg, Ken Goldberg, Abhinav Gupta, Abhishek Gupta, Dinesh Jayaraman, Joseph J. Lim, Jitendra Malik, Roberto Martín-Martín, Subramanian Ramamoorthy, Dorsa Sadigh, Shuran Song, Jiajun Wu, Michael C. Yip, Yuke Zhu, Thomas Kollar, Sergey Levine, and Chelsea Finn. DROID: A Large-Scale In-The-Wild Robot Manipulation Dataset, March 2024. URL http://arxiv.org/abs/2403.12945. arXiv:2403.12945 [cs].

Sergey Levine and Vladlen Koltun. Guided Policy Search. In *Proceedings of the 30th International Conference on Machine Learning*, pp. 1–9. PMLR, May 2013. URL https://proceedings.mlr.press/v28/levine13.html. ISSN: 1938-7228.

Christina Lichtenthäler and Alexandra Kirsch. Legibility of Robot Behavior : A Literature Review. April 2016. URL https://hal.science/hal-01306977.

Huihan Liu, Soroush Nasiriany, Lance Zhang, Zhiyao Bao, and Yuke Zhu. Robot Learning on the Job: Human-in-the-Loop Autonomy and Learning During Deployment, July 2023. URL http://arxiv.org/abs/2211.08416. arXiv:2211.08416 [cs].

Yecheng Jason Ma, William Liang, Guanzhi Wang, De-An Huang, Osbert Bastani, Dinesh Jayaraman, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Eureka: Human-Level Reward Design via Coding Large Language Models, April 2024. URL http://arxiv.org/abs/2310.12931. arXiv:2310.12931 [cs].

Stephanie Milani, Nicholay Topin, Manuela Veloso, and Fei Fang. Explainable Reinforcement Learning: A Survey and Comparative Review. *ACM Computing Surveys*, 56(7):1–36, July 2024. ISSN 0360-0300, 1557-7341. doi: 10.1145/3616864. URL https://dl.acm.org/doi/10.1145/3616864.

Jean-Baptiste Mouret and Jeff Clune. Illuminating search spaces by mapping elites, April 2015. URL http://arxiv.org/abs/1504.04909. arXiv:1504.04909 [cs, q-bio].

Hai Nguyen and Hung La. Review of Deep Reinforcement Learning for Robot Manipulation. In *2019 Third IEEE International Conference on Robotic Computing (IRC)*, pp. 590–595, February 2019. doi: 10.1109/IRC.2019.00120. URL https://ieeexplore.ieee.org/document/8675643.

Alex Nichol and Prafulla Dhariwal. Improved Denoising Diffusion Probabilistic Models, February 2021. URL http://arxiv.org/abs/2102.09672. arXiv:2102.09672 [cs, stat].

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, 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, C. J. 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, March 2024. URL http://arxiv.org/abs/2303.08774. arXiv:2303.08774 [cs].

Hae Won Park, Ishaan Grover, Samuel Spaulding, Louis Gomez, and Cynthia Breazeal. A model-free affective reinforcement learning approach to personalization of an autonomous social robot companion for early literacy education. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19, pp. 687–694, Honolulu, Hawaii, USA, January 2019. AAAI Press. ISBN 978-1-57735-809-1. doi: 10.1609/aaai.v33i01.3301687. URL https://dl.acm.org/doi/10.1609/aaai.v33i01.3301687.

Matthias Plappert, Rein Houthooft, Prafulla Dhariwal, Szymon Sidor, Richard Y. Chen, Xi Chen, Tamim Asfour, Pieter Abbeel, and Marcin Andrychowicz. Parameter Space Noise for Exploration, January 2018. URL http://arxiv.org/abs/1706.01905. arXiv:1706.01905 [cs, stat].

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision, February 2021. URL http://arxiv.org/abs/2103.00020. arXiv:2103.00020 [cs].

Harish Ravichandar, Athanasios S. Polydoros, Sonia Chernova, and Aude Billard. Recent Advances in Robot Learning from Demonstration. *Annual Review of Control, Robotics, and Autonomous Systems*, 3(Volume 3, 2020):297–330, May 2020. ISSN 2573-5144. doi: 10.1146/annurev-control-100819-063206. URL https://www.annualreviews.org/content/journals/10.1146/annurev-control-100819-063206. Publisher: Annual Reviews.

Siddharth Reddy, Anca D. Dragan, and Sergey Levine. Shared Autonomy via Deep Reinforcement Learning, May 2018. URL http://arxiv.org/abs/1802.01744. Number: arXiv:1802.01744 arXiv:1802.01744 [cs].

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models, April 2022. URL http://arxiv.org/abs/2112.10752. arXiv:2112.10752 [cs].

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation, May 2015. URL http://arxiv.org/abs/1505.04597. arXiv:1505.04597 [cs].

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal Policy Optimization Algorithms. *arXiv:1707.06347 [cs]*, August 2017. URL http://arxiv.org/abs/1707.06347.

Rossitza Setchi, Maryam Banitalebi Dehkordi, and Juwairiya Siraj Khan. Explainable Robotics in Human-Robot Interactions. *Procedia Computer Science*, 176:3057–3066, January 2020. ISSN 1877-0509. doi: 10.1016/j.procs.2020.09.198. URL https://www.sciencedirect.com/science/article/pii/S1877050920321001.

Isaac Sheidlower, Mavis Murdock, Emma Bethel, Reuben M. Aronson, and Elaine Schaertl Short. Online Behavior Modification for Expressive User Control of RL-Trained Robots. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '24, pp. 639–648, New York, NY, USA, March 2024. Association for Computing Machinery. ISBN 9798400703225. doi: 10.1145/3610977.3634947. URL https://dl.acm.org/doi/10.1145/3610977.3634947.

Matthew E. Taylor, Shimon Whiteson, and Peter Stone. Transfer via inter-task mappings in policy search reinforcement learning. In *Proceedings of the 6th international joint conference on Autonomous agents and multiagent systems*, pp. 1–8, Honolulu Hawaii, May 2007. ACM. ISBN 978-81-904262-7-5. doi: 10.1145/1329125.1329170. URL https://dl.acm.org/doi/10.1145/1329125.1329170.

Bryon Tjanaka, Matthew C. Fontaine, Yulun Zhang, Sam Sommerer, Nathan Dennler, and Stefanos Nikolaidis. pyribs: A bare-bones Python library for quality diversity optimization, 2021. URL https://github.com/icaros-usc/pyribs. Publication Title: GitHub repository.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. LLaMA: Open and Efficient Foundation Language Models, February 2023. URL http://arxiv.org/abs/2302.13971. arXiv:2302.13971 [cs].

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need, August 2023. URL http://arxiv.org/abs/1706.03762. arXiv:1706.03762 [cs].

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. C-Pack: Packaged Resources To Advance General Chinese Embedding, 2023. _eprint: 2309.07597.

Yanjie Ze, Gu Zhang, Kangning Zhang, Chenyuan Hu, Muhan Wang, and Huazhe Xu. 3D Diffusion Policy: Generalizable Visuomotor Policy Learning via Simple 3D Representations, April 2024. URL http://arxiv.org/abs/2403.03954. arXiv:2403.03954 [cs].

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. A Survey of Large Language Models, November 2023. URL http://arxiv.org/abs/2303.18223. arXiv:2303.18223 [cs].