Open X-Embodiment: Robotic Learning Datasets and RT-X Models

Open X-Embodiment Collaboration
https://robotics-transformer-x.github.io/

Abstract

Large, high-capacity models trained on diverse datasets have shown remarkable successes on efficiently tackling downstream applications. In domains from NLP to Computer Vision, this has led to a consolidation of pretrained models, with general pretrained backbones serving as a starting point for many applications. Can such a consolidation happen in robotics? Conventionally, robotic learning methods train a separate model for every application, every robot, and even every environment. Can we instead train "generalist" X-robot policy that can be adapted efficiently to new robots, tasks, and environments? In this paper, we provide datasets in standardized data formats and models to make it possible to explore this possibility in the context of robotic manipulation, alongside experimental results that provide an example of effective X-robot policies. We assemble a dataset from 22 different robots collected through a collaboration between 21 institutions, demonstrating 527 skills (160266 tasks). We show that a high-capacity model trained on this data, which we call RT-X, exhibits positive transfer and improves the capabilities of multiple robots by leveraging experience from other platforms.

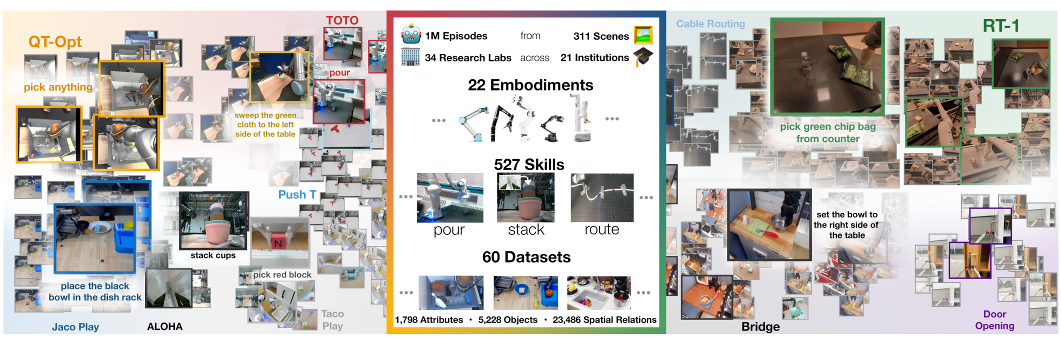

Figure 1: We propose an open, large-scale dataset for robot learning curated from 21 institutions across the globe. The dataset represents diverse behaviors, robot embodiments and environments.

1 Introduction

A central lesson from advances in machine learning and artificial intelligence is that large-scale learning from broad and diverse datasets can enable capable AI systems by providing for general-purpose pretrained models. In fact, large-scale general-purpose models typically trained on large and diverse datasets can often outperform their *narrowly targeted* counterparts trained on smaller but more task-specific data. For instance, open-vocabulary image classifiers (e.g., CLIP [80]) trained on large datasets scraped from the web tend to outperform fixed-vocabulary models trained on more limited datasets, and large language models [4, 75] trained on massive text corpora tend to outperform systems that are only trained on narrow task-specific datasets. Increasingly, the most effective way to tackle a given narrow task (e.g., in vision or NLP) is to adapt a general-purpose model. However, these lessons are difficult to apply in robotics: any single robotic domain might be too narrow, and

NeurIPS 2023 Workshop on Robot Learning: Pretraining, Fine-Tuning, and Generalization with Large Scale Models, New Orleans, United States of America

while computer vision and NLP can leverage large datasets sourced from the web, comparably large and broad datasets for robotic interaction are hard to come by. Even the largest data collection efforts still end up with datasets that are a fraction of the size and diversity of benchmark datasets in vision (5-18M) [109, 112] and NLP (1.5B-4.5B) [52, 72]. More importantly, such datasets are often still narrow along some axes of variation, either focusing on a single environment, a single set of objects, or a narrow range of tasks. How can we overcome these challenges in robotics and move the field of robotic learning toward the kind of large data regime that has been so successful in other domains?

Inspired by the generalization made possible by pretraining large vision or language models on diverse data, we take the perspective that the goal of training generalizable robot policies requires **X-embodiment training**, i.e., with data from multiple robotic platforms. While each individual robotic learning dataset might be too narrow, their union provide a better coverage of variations in environments and robots. Learning generalizable robot policies requires developing methods that can utilize X-embodiment data, tapping into datasets from many labs, robots, and settings. Even if such datasets in their current size and coverage are insufficient to attain the impressive generalization results that have been demonstrated by large language models, in the future, the union of such data can potentially provide this kind of coverage. Because of this, **we believe that enabling research into X-embodiment robotic learning is critical at the present juncture**.

Following this rationale, our work has two goals: **(1)** Demonstrate that policies trained on data from many different robots and environments enjoy the benefits of positive transfer, attaining better performance than policies trained only on data from each evaluation setup. **(2)** Provide datasets, data formats and models for the robotics community to enable future research on X-embodiment models.

Addressing goal **(1)**, we demonstrate that several recent robotic learning methods, with minimal modification, can utilize X-embodiment data and enable positive transfer. Specifically, we train the RT-1 [14] and RT-2 [13] models on 9 different robotic manipulators. We show that the resulting models, which we call RT-X, can improve over policies trained only on data from the evaluation domain, exhibiting better generalization and new capabilities. Addressing **(2)**, we provide the Open X-Embodiment (OXE) Repository, which includes a dataset with 22 different robotic embodiments from 21 different institutions that can enable the robotics community to pursue further research on X-embodiment models, along with open-source tools to facilitate such research. Our aim is not to innovate in terms of the particular architectures and algorithms, but rather to provide the model that we trained together with data and tools to energize research around X-embodiment robotic learning.

2 The Open X-Embodiment Repository

We introduce the Open X-Embodiment Repository – an open-source repository which includes **large-scale data** along with **pre-trained model checkpoints** for X-embodied robot learning research. More specifically, we provide and maintain the following open-source resources to the broader community: (1) **Open X-Embodiment Dataset**: robot learning dataset with *1M+ robot trajectories* from 22 *robot embodiments* (2) **Pre-Trained Checkpoints**: a selection of RT-X model checkpoints ready for inference and fine-tuning.

We intend for these resources to form a foundation for X-embodiment research in robot learning, but they are just the start. Open X-Embodiment is a community-driven effort, currently involving 21 institutions from around the world, and we hope to further broaden participation and grow the initial Open X-Embodiment Dataset over time. The Open X-Embodiment Dataset contains 1M+ real robot trajectories spanning 22 robot embodiments, from single robot arms to bi-manual robots and quadrupeds. The dataset was constructed by pooling 60 *existing* robot datasets from 34 robotic research labs around the world and converting them into a consistent data format for easy download and usage. We use the RLDS data format [83], which saves data in serialized `tfrecord` files and accommodates the various action spaces and input modalities of different robot setups.

3 RT-X Design

To evaluate how much X-embodiment training can improve the performance of learned policies, we require models that have sufficient capacity to productively make use of such large and heterogeneous datasets. To that end, our experiments build on two recently proposed Transformer-based robotic policies: RT-1 [14] and RT-2 [13]. Both models take in a visual input and natural language instruction describing the task, and output tokenized actions.

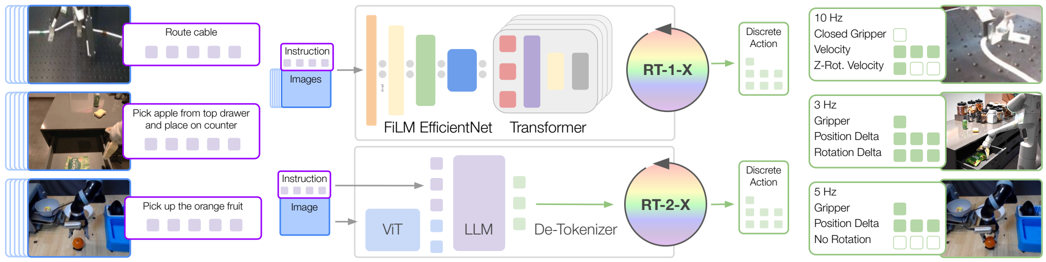

Figure 2: RT-1-X and RT-2-X both take images and a text instruction as input and output discretized end-effector actions. RT-1-X is an architecture designed for robotics, with a FiLM [78] conditioned EfficientNet [105] and a Transformer [106]. RT-2-X builds on a VLM backbone by representing actions as another language, and training action text tokens together with vision-language data.

We define the robotics data mixture used across all of the experiments as the data from 9 manipulators, and taken from RT-1 [14], QT-Opt [44], Bridge [108], Task Agnostic Robot Play [66, 85], Jaco Play [21], Cable Routing [56], RoboTurk [64], NYU VINN [76], Austin VIOLA [126], Berkeley Autolab UR5 [18], TOTO [122] and Language Table [58] datasets. RT-1-X is trained on only robotics mixture data defined above, whereas RT-2-X is trained via co-fine-tuning (similarly to the original RT-2 [13]), with an approximately one to one split of the original VLM data and the robotics data mixture.

One challenge of creating X-embodiment models is that observation and action spaces vary significantly across robots. We use a coarsely aligned action and observation space across datasets. The model receives a history of recent images and language instructions as observations and predicts a 7-dimensional action vector controlling the end-effector (x, y, z, roll, pitch, yaw, and gripper opening or the rates of these quantities). We select one canonical camera view from each dataset as the input image, resize it to a common resolution and convert the original action set into a 7 DoF end-effector action. We normalize each dataset's actions prior to discretization. This way, an output of the model can be interpreted (de-normalized) differently depending on the embodiment used. It should be noted that despite this coarse alignment, the camera observations still vary substantially across datasets, e.g. due to differing camera poses relative to the robot or differing camera properties, see Figure 2.

Similarly, for the action space, we do not align the coordinate frames across datasets in which the end-effector is controlled, and allow action values to represent either absolute or relative positions or velocities, as per the original control scheme chosen for each robot. Thus, the same action vector may induce very different motions for different robots.

4 Experimental Results

Our experiments answer three questions about the effect of X-embodiment training: (1) Can policies trained on our X-embodiment dataset effectively enable positive transfer, such that co-training on data collected on multiple robots improves performance on the training task? (2) Does co-training models on data from multiple platforms and tasks improve generalization to new, unseen tasks? To answer these questions we conduct the total number of 3600 evaluation trials across 6 different robots.

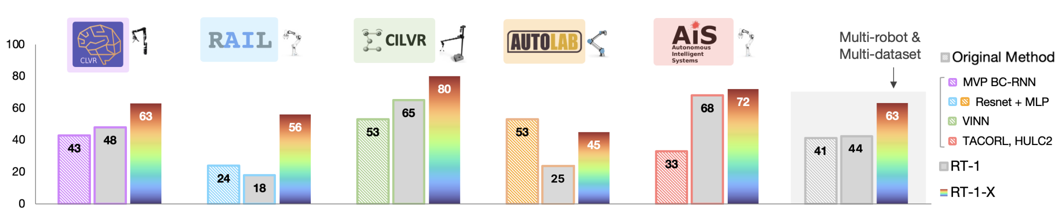

Figure 3: RT-1-X mean success rate is 50% higher than that of either the Original Method or RT-1. RT-1 and RT-1-X have the same network architecture. Therefore the performance increase can be attributed to co-training on the robotics data mixture. The lab logos indicate the physical location of real robot evaluation, and the robot pictures indicate the embodiment used for the evaluation.

Row	Model	Size	History Length	Dataset	Co-Trained w/ Web	Initial Checkpoint	Emergent Skills Evaluation
(1)	RT-2	55B	none	Google Robot action	Yes	Web-pretrained	27.3%
(2)	RT-2-X	55B	none	Robotics data	Yes	Web-pretrained	**75.8%**

Table 1: RT-2-X outperforms RT-2 by $\sim 3\times$ in emergent skills evaluation.

4.1 In-distribution performance across different embodiments

To assess the ability of the RT-1-X model variant to learn from X-embodiment data, we evaluate its performance on in-distribution tasks on domains that only have small-scale datasets (Fig. 3), where we would expect transfer from larger datasets to significantly improve performance. We consider Kitchen Manipulation [21], Cable Routing [56], NYU Door Opening [76], AUTOLab UR5 [18], and Robot Play [1]. We use the same evaluation and robot embodiment as in the respective publications.

Throughout this evaluation we compare with two baseline models: (1) The model developed by the creators of the dataset trained only on that respective dataset. This constitutes a reasonable baseline insofar as it can be expected that the model has been optimized to work well with the associated data; we refer to this baseline model as the *Original Method* model. (2) An RT-1 model trained on the dataset in isolation; this baseline allows us to assess whether the RT-X model architectures have enough capacity to represent policies for multiple different robot platforms simultaneously, and whether co-training leads to higher performance. RT-1-X outperforms Original Method trained on each of the robot-specific datasets on 4 of the 5 datasets, with a large average improvement, demonstrating limited data domains benefit substantially from co-training (Fig. 3).

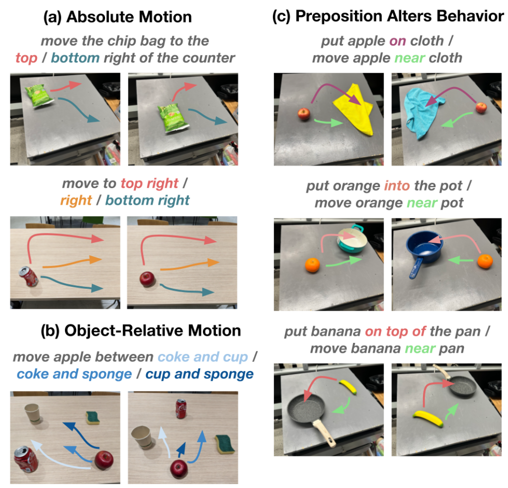

Figure 4: To assess transfer *between* embodiments, we evaluate the RT-2-X model on out-of-distribution skills.

4.2 Improved generalization to out-of-distribution settings

We examine if X-embodiment training enables better generalization to out-of-distribution settings and more complex and novel instructions. These experiments focus on the high-data domains, and use the RT-2-X model. We conduct experiments with the Google Robot, assessing the performance on tasks like the ones shown in Fig. 4. These tasks involve objects and skills that are not present in the RT-2 dataset but occur in the Bridge dataset [108] for a different robot (the *WidowX robot*).

Results are shown in Table 1, Emergent Skills Evaluation column. Comparing rows (1) and (2), we find that RT-2-X outperforms RT-2 by $\sim 3\times$, suggesting that incorporating data from other robots into training improves the range of tasks that can be performed even by a robot that already has large amounts of data available. Our results suggest that co-training with data from other platforms imbues the RT-2-X controller with additional skills for the platform that are not present in that platform's original dataset.

References

[1] Task-agnostic real world robot play. https://www.kaggle.com/datasets/oiermees/taco-robot.

[2] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, et al. Do as I can, not as I say: Grounding language in robotic affordances. *Conference on Robot Learning (CoRL)*, 2022.

[3] Minttu Alakuijala, Gabriel Dulac-Arnold, Julien Mairal, Jean Ponce, and Cordelia Schmid. Learning reward functions for robotic manipulation by observing humans. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5006–5012. IEEE, 2023.

[4] Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. PaLM 2 technical report. *arXiv preprint arXiv:2305.10403*, 2023.

[5] Shikhar Bahl, Abhinav Gupta, and Deepak Pathak. Human-to-robot imitation in the wild. *Robotics: Science and Systems (RSS)*, 2022.

[6] Shikhar Bahl, Russell Mendonca, Lili Chen, Unnat Jain, and Deepak Pathak. Affordances from human videos as a versatile representation for robotics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13778–13790, June 2023.

[7] Shikhar Bahl, Russell Mendonca, Lili Chen, Unnat Jain, and Deepak Pathak. Affordances from human videos as a versatile representation for robotics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13778–13790, 2023.

[8] Homanga Bharadhwaj, Jay Vakil, Mohit Sharma, Abhinav Gupta, Shubham Tulsiani, and Vikash Kumar. RoboAgent: Towards sample efficient robot manipulation with semantic augmentations and action chunking. *arxiv*, 2023.

[9] Alessandro Bonardi, Stephen James, and Andrew J Davison. Learning one-shot imitation from humans without humans. *IEEE Robotics and Automation Letters*, 5(2):3533–3539, 2020.

[10] Konstantinos Bousmalis, Alex Irpan, Paul Wohlhart, Yunfei Bai, Matthew Kelcey, Mrinal Kalakrishnan, Laura Downs, Julian Ibarz, Peter Pastor, Kurt Konolige, Sergey Levine, and Vincent Vanhoucke. Using simulation and domain adaptation to improve efficiency of deep robotic grasping. In *ICRA*, pages 4243–4250, 2018.

[11] Konstantinos Bousmalis, Giulia Vezzani, Dushyant Rao, Coline Devin, Alex X Lee, Maria Bauza, Todor Davchev, Yuxiang Zhou, Agrim Gupta, Akhil Raju, et al. RoboCat: A self-improving foundation agent for robotic manipulation. *arXiv preprint arXiv:2306.11706*, 2023.

[12] Samarth Brahmbhatt, Cusuh Ham, Charles Kemp, and James Hays. Contactdb: Analyzing and predicting grasp contact via thermal imaging, 04 2019.

[13] 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.

[14] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. RT-1: Robotics transformer for real-world control at scale. *Robotics: Science and Systems (RSS)*, 2023.

[15] Berk Calli, Aaron Walsman, Arjun Singh, Siddhartha Srinivasa, Pieter Abbeel, and Aaron M. Dollar. Benchmarking in manipulation research: Using the Yale-CMU-Berkeley object and model set. *IEEE Robotics & Automation Magazine*, 22(3):36–52, 2015.

[16] Annie S Chen, Suraj Nair, and Chelsea Finn. Learning generalizable robotic reward functions from “in-the-wild" human videos. *arXiv preprint arXiv:2103.16817*, 2021.

[17] David L. Chen and Raymond J. Mooney. Learning to interpret natural language navigation instructions from observations. In *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence*, page 859–865, 2011.

[18] Lawrence Yunliang Chen, Simeon Adebola, and Ken Goldberg. Berkeley UR5 demonstration dataset. https://sites.google.com/view/berkeley-ur5/home.

[19] Tao Chen, Adithyavairavan Murali, and Abhinav Gupta. Hardware conditioned policies for multi-robot transfer learning. In *Advances in Neural Information Processing Systems*, pages 9355–9366, 2018.

[20] Sudeep Dasari, Frederik Ebert, Stephen Tian, Suraj Nair, Bernadette Bucher, Karl Schmeckpeper, Siddharth Singh, Sergey Levine, and Chelsea Finn. RoboNet: Large-scale multi-robot learning. In *Conference on Robot Learning (CoRL)*, volume 100, pages 885–897. PMLR, 2019.

[21] Shivin Dass, Jullian Yapeter, Jesse Zhang, Jiahui Zhang, Karl Pertsch, Stefanos Nikolaidis, and Joseph J. Lim. CLVR jaco play dataset, 2023.

[22] Amaury Depierre, Emmanuel Dellandréa, and Liming Chen. Jacquard: A large scale dataset for robotic grasp detection. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3511–3516. IEEE, 2018.

[23] Coline Devin, Abhishek Gupta, Trevor Darrell, Pieter Abbeel, and Sergey Levine. Learning modular neural network policies for multi-task and multi-robot transfer. In *2017 IEEE international conference on robotics and automation (ICRA)*, pages 2169–2176. IEEE, 2017.

[24] Mingyu Ding, Yan Xu, Zhenfang Chen, David Daniel Cox, Ping Luo, Joshua B Tenenbaum, and Chuang Gan. Embodied concept learner: Self-supervised learning of concepts and mapping through instruction following. In *Conference on Robot Learning*, pages 1743–1754. PMLR, 2023.

[25] Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B McHugh, and Vincent Vanhoucke. Google scanned objects: A high-quality dataset of 3D scanned household items. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 2553–2560. IEEE, 2022.

[26] Felix Duvallet, Jean Oh, Anthony Stentz, Matthew Walter, Thomas Howard, Sachithra Hemachandra, Seth Teller, and Nicholas Roy. Inferring maps and behaviors from natural language instructions. In *International Symposium on Experimental Robotics (ISER)*, 2014.

[27] Frederik Ebert, Chelsea Finn, Sudeep Dasari, Annie Xie, Alex Lee, and Sergey Levine. Visual foresight: Model-based deep reinforcement learning for vision-based robotic control. *arXiv preprint arXiv:1812.00568*, 2018.

[28] Frederik Ebert, Yanlai Yang, Karl Schmeckpeper, Bernadette Bucher, Georgios Georgakis, Kostas Daniilidis, Chelsea Finn, and Sergey Levine. Bridge data: Boosting generalization of robotic skills with cross-domain datasets. In *Robotics: Science and Systems (RSS) XVIII*, 2022.

[29] Clemens Eppner, Arsalan Mousavian, and Dieter Fox. ACRONYM: A large-scale grasp dataset based on simulation. In *2021 IEEE Int. Conf. on Robotics and Automation, ICRA*, 2020.

[30] Hao-Shu Fang, Hongjie Fang, Zhenyu Tang, Jirong Liu, Junbo Wang, Haoyi Zhu, and Cewu Lu. RH20T: A robotic dataset for learning diverse skills in one-shot. In *RSS 2023 Workshop on Learning for Task and Motion Planning*, 2023.

[31] Hao-Shu Fang, Chenxi Wang, Minghao Gou, and Cewu Lu. Graspnet-1billion: a large-scale benchmark for general object grasping. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11444–11453, 2020.

[32] Chelsea Finn and Sergey Levine. Deep visual foresight for planning robot motion. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2786–2793. IEEE, 2017.

[33] Ruohan Gao, Yen-Yu Chang, Shivani Mall, Li Fei-Fei, and Jiajun Wu. ObjectFolder: A dataset of objects with implicit visual, auditory, and tactile representations. In *Conference on Robot Learning*, pages 466–476, 2021.

[34] Ali Ghadirzadeh, Xi Chen, Petra Poklukar, Chelsea Finn, Mårten Björkman, and Danica Kragic. Bayesian meta-learning for few-shot policy adaptation across robotic platforms. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1274–1280. IEEE, 2021.

[35] Agrim Gupta, Linxi Fan, Surya Ganguli, and Li Fei-Fei. Metamorph: Learning universal controllers with transformers. In *International Conference on Learning Representations*, 2021.

[36] Minho Heo, Youngwoon Lee, Doohyun Lee, and Joseph J. Lim. Furniturebench: Reproducible real-world benchmark for long-horizon complex manipulation. In *Robotics: Science and Systems*, 2023.

[37] Felix Hill, Sona Mokra, Nathaniel Wong, and Tim Harley. Human instruction-following with deep reinforcement learning via transfer-learning from text. *arXiv preprint arXiv:2005.09382*, 2020.

[38] Edward S. Hu, Kun Huang, Oleh Rybkin, and Dinesh Jayaraman. Know thyself: Transferable visual control policies through robot-awareness. In *International Conference on Learning Representations*, 2022.

[39] Wenlong Huang, Igor Mordatch, and Deepak Pathak. One policy to control them all: Shared modular policies for agent-agnostic control. In *ICML*, 2020.

[40] Wenlong Huang, Chen Wang, Ruohan Zhang, Yunzhu Li, Jiajun Wu, and Li Fei-Fei. VoxPoser: Composable 3d value maps for robotic manipulation with language models. *arXiv preprint arXiv:2307.05973*, 2023.

[41] Eric Jang, Alex Irpan, Mohi Khansari, Daniel Kappler, Frederik Ebert, Corey Lynch, Sergey Levine, and Chelsea Finn. BC-Z: Zero-shot task generalization with robotic imitation learning. In *Conference on Robot Learning (CoRL)*, pages 991–1002, 2021.

[42] Yun Jiang, Stephen Moseson, and Ashutosh Saxena. Efficient grasping from RGBD images: Learning using a new rectangle representation. In *2011 IEEE International conference on robotics and automation*, pages 3304–3311. IEEE, 2011.

[43] Yunfan Jiang, Agrim Gupta, Zichen Zhang, Guanzhi Wang, Yongqiang Dou, Yanjun Chen, Li Fei-Fei, Anima Anandkumar, Yuke Zhu, and Linxi Fan. VIMA: General robot manipulation with multimodal prompts. *International Conference on Machine Learning (ICML)*, 2023.

[44] Dmitry Kalashnikov, Alex Irpan, Peter Pastor, Julian Ibarz, Alexander Herzog, Eric Jang, Deirdre Quillen, Ethan Holly, Mrinal Kalakrishnan, Vincent Vanhoucke, et al. QT-Opt: Scalable deep reinforcement learning for vision-based robotic manipulation. *arXiv preprint arXiv:1806.10293*, 2018.

[45] Dmitry Kalashnikov, Jacob Varley, Yevgen Chebotar, Benjamin Swanson, Rico Jonschkowski, Chelsea Finn, Sergey Levine, and Karol Hausman. MT-Opt: Continuous multi-task robotic reinforcement learning at scale. *arXiv preprint arXiv:2104.08212*, 2021.

[46] Daniel Kappler, Jeannette Bohg, and Stefan Schaal. Leveraging big data for grasp planning. In *ICRA*, pages 4304–4311, 2015.

[47] Siddharth Karamcheti, Suraj Nair, Annie S Chen, Thomas Kollar, Chelsea Finn, Dorsa Sadigh, and Percy Liang. Language-driven representation learning for robotics. *arXiv preprint arXiv:2302.12766*, 2023.

[48] Alexander Kasper, Zhixing Xue, and Rüdiger Dillmann. The kit object models database: An object model database for object recognition, localization and manipulation in service robotics. *The International Journal of Robotics Research*, 31(8):927–934, 2012.

[49] Thomas Kollar, Stefanie Tellex, Deb Roy, and Nicholas Roy. Toward understanding natural language directions. In *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 259–266, 2010.

[50] Sateesh Kumar, Jonathan Zamora, Nicklas Hansen, Rishabh Jangir, and Xiaolong Wang. Graph inverse reinforcement learning from diverse videos. In *Conference on Robot Learning*, pages 55–66. PMLR, 2023.

[51] Vitaly Kurin, Maximilian Igl, Tim Rocktäschel, Wendelin Boehmer, and Shimon Whiteson. My body is a cage: the role of morphology in graph-based incompatible control. *arXiv preprint arXiv:2010.01856*, 2020.

[52] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and Christian Bizer. DBpedia - a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web*, 6(2):167–195, 2015.

[53] Sergey Levine, Peter Pastor, Alex Krizhevsky, Julian Ibarz, and Deirdre Quillen. Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection. *The International journal of robotics research*, 37(4-5):421–436, 2018.

[54] YuXuan Liu, Abhishek Gupta, Pieter Abbeel, and Sergey Levine. Imitation from observation: Learning to imitate behaviors from raw video via context translation. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1118–1125. IEEE, 2018.

[55] Jelena Luketina, Nantas Nardelli, Gregory Farquhar, Jakob N. Foerster, Jacob Andreas, Edward Grefenstette, S. Whiteson, and Tim Rocktäschel. A survey of reinforcement learning informed by natural language. In *IJCAI*, 2019.

[56] Jianlan Luo, Charles Xu, Xinyang Geng, Gilbert Feng, Kuan Fang, Liam Tan, Stefan Schaal, and Sergey Levine. Multi-stage cable routing through hierarchical imitation learning. *arXiv preprint arXiv:2307.08927*, 2023.

[57] Corey Lynch and Pierre Sermanet. Grounding language in play. *Robotics: Science and Systems (RSS)*, 2021.

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

[59] Yecheng Jason Ma, Shagun Sodhani, Dinesh Jayaraman, Osbert Bastani, Vikash Kumar, and Amy Zhang. Vip: Towards universal visual reward and representation via value-implicit pre-training. *arXiv preprint arXiv:2210.00030*, 2022.

[60] Matt MacMahon, Brian Stankiewicz, and Benjamin Kuipers. Walk the talk: Connecting language, knowledge, and action in route instructions. In *Proceedings of the Twenty-First AAAI Conference on Artificial Intelligence*, 2006.

[61] Jeffrey Mahler, Jacky Liang, Sherdil Niyaz, Michael Laskey, Richard Doan, Xinyu Liu, Juan Aparicio Ojea, and Ken Goldberg. Dex-Net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics. In *Robotics: Science and Systems (RSS)*, 2017.

[62] Arjun Majumdar, Karmesh Yadav, Sergio Arnaud, Yecheng Jason Ma, Claire Chen, Sneha Silwal, Aryan Jain, Vincent-Pierre Berges, Pieter Abbeel, Jitendra Malik, et al. Where are we in the search for an artificial visual cortex for embodied intelligence? *arXiv preprint arXiv:2303.18240*, 2023.

[63] Ajay Mandlekar, Jonathan Booher, Max Spero, Albert Tung, Anchit Gupta, Yuke Zhu, Animesh Garg, Silvio Savarese, and Li Fei-Fei. Scaling robot supervision to hundreds of hours with RoboTurk: Robotic manipulation dataset through human reasoning and dexterity. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1048–1055. IEEE, 2019.

[64] Ajay Mandlekar, Yuke Zhu, Animesh Garg, Jonathan Booher, Max Spero, Albert Tung, Julian Gao, John Emmons, Anchit Gupta, Emre Orbay, Silvio Savarese, and Li Fei-Fei. RoboTurk: A crowdsourcing platform for robotic skill learning through imitation. *CoRR*, abs/1811.02790, 2018.

[65] Roberto Martín-Martín, Michelle Lee, Rachel Gardner, Silvio Savarese, Jeannette Bohg, and Animesh Garg. Variable impedance control in end-effector space. an action space for reinforcement learning in contact rich tasks. In *Proceedings of the International Conference of Intelligent Robots and Systems (IROS)*, 2019.

[66] Oier Mees, Jessica Borja-Diaz, and Wolfram Burgard. Grounding language with visual affordances over unstructured data. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, London, UK, 2023.

[67] Oier Mees, Lukas Hermann, and Wolfram Burgard. What matters in language conditioned robotic imitation learning over unstructured data. *IEEE Robotics and Automation Letters*, 7(4):11205–11212, 2022.

[68] Oier Mees, Lukas Hermann, Erick Rosete-Beas, and Wolfram Burgard. CALVIN: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks. *IEEE Robotics and Automation Letters*, 2022.

[69] Douglas Morrison, Peter Corke, and Jürgen Leitner. Egad! an evolved grasping analysis dataset for diversity and reproducibility in robotic manipulation. *IEEE Robotics and Automation Letters*, 5(3):4368–4375, 2020.

[70] Yao Mu, Shunyu Yao, Mingyu Ding, Ping Luo, and Chuang Gan. EC2: Emergent communication for embodied control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6704–6714, 2023.

[71] Yao Mu, Qinglong Zhang, Mengkang Hu, Wenhai Wang, Mingyu Ding, Jun Jin, Bin Wang, Jifeng Dai, Yu Qiao, and Ping Luo. EmbodiedGPT: Vision-language pre-training via embodied chain of thought. *arXiv preprint arXiv:2305.15021*, 2023.

[72] Hannes Mühleisen and Christian Bizer. Web data commons-extracting structured data from two large web corpora. *LDOW*, 937:133–145, 2012.

[73] Suraj Nair, Eric Mitchell, Kevin Chen, Silvio Savarese, Chelsea Finn, et al. Learning language-conditioned robot behavior from offline data and crowd-sourced annotation. In *Conference on Robot Learning*, pages 1303–1315. PMLR, 2022.

[74] Suraj Nair, Aravind Rajeswaran, Vikash Kumar, Chelsea Finn, and Abhinav Gupta. R3m: A universal visual representation for robot manipulation. In *CoRL*, 2022.

[75] OpenAI. GPT-4 technical report, 2023.

[76] Jyothish Pari, Nur Muhammad Shafiullah, Sridhar Pandian Arunachalam, and Lerrel Pinto. The surprising effectiveness of representation learning for visual imitation, 2021.

[77] Deepak Pathak, Christopher Lu, Trevor Darrell, Phillip Isola, and Alexei A Efros. Learning to control self-assembling morphologies: a study of generalization via modularity. *Advances in Neural Information Processing Systems*, 32, 2019.

[78] Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer, 2017.

[79] Lerrel Pinto and Abhinav Kumar Gupta. Supersizing self-supervision: Learning to grasp from 50k tries and 700 robot hours. *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3406–3413, 2015.

[80] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[81] Ilija Radosavovic, Baifeng Shi, Letian Fu, Ken Goldberg, Trevor Darrell, and Jitendra Malik. Robot learning with sensorimotor pre-training. In *Conference on Robot Learning*, 2023.

[82] Ilija Radosavovic, Tete Xiao, Stephen James, Pieter Abbeel, Jitendra Malik, and Trevor Darrell. Real-world robot learning with masked visual pre-training. In *Conference on Robot Learning*, 2022.

[83] Sabela Ramos, Sertan Girgin, Léonard Hussenot, Damien Vincent, Hanna Yakubovich, Daniel Toyama, Anita Gergely, Piotr Stanczyk, Raphael Marinier, Jeremiah Harmsen, Olivier Pietquin, and Nikola Momchev. RLDS: an ecosystem to generate, share and use datasets in reinforcement learning, 2021.

[84] 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. A generalist agent. *Transactions on Machine Learning Research*, 2022.

[85] Erick Rosete-Beas, Oier Mees, Gabriel Kalweit, Joschka Boedecker, and Wolfram Burgard. Latent plans for task agnostic offline reinforcement learning. In *Proceedings of the 6th Conference on Robot Learning (CoRL)*, 2022.

[86] Gautam Salhotra, I-Chun Arthur Liu, and Gaurav Sukhatme. Bridging action space mismatch in learning from demonstrations. *arXiv preprint arXiv:2304.03833*, 2023.

[87] Alvaro Sanchez-Gonzalez, Nicolas Heess, Jost Tobias Springenberg, Josh Merel, Martin Riedmiller, Raia Hadsell, and Peter Battaglia. Graph networks as learnable physics engines for inference and control. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4470–4479. PMLR, 10–15 Jul 2018.

[88] Karl Schmeckpeper, Oleh Rybkin, Kostas Daniilidis, Sergey Levine, and Chelsea Finn. Reinforcement learning with videos: Combining offline observations with interaction. In *Conference on Robot Learning*, pages 339–354. PMLR, 2021.

[89] Karl Schmeckpeper, Annie Xie, Oleh Rybkin, Stephen Tian, Kostas Daniilidis, Sergey Levine, and Chelsea Finn. Learning predictive models from observation and interaction. In *European Conference on Computer Vision*, pages 708–725. Springer, 2020.

[90] Ingmar Schubert, Jingwei Zhang, Jake Bruce, Sarah Bechtle, Emilio Parisotto, Martin Riedmiller, Jost Tobias Springenberg, Arunkumar Byravan, Leonard Hasenclever, and Nicolas Heess. A generalist dynamics model for control, 2023.

[91] Pierre Sermanet, Kelvin Xu, and Sergey Levine. Unsupervised perceptual rewards for imitation learning. *arXiv preprint arXiv:1612.06699*, 2016.

[92] Dhruv Shah, Ajay Sridhar, Arjun Bhorkar, Noriaki Hirose, and Sergey Levine. GNM: A general navigation model to drive any robot. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7226–7233. IEEE, 2023.

[93] Dhruv Shah, Ajay Sridhar, Nitish Dashora, Kyle Stachowicz, Kevin Black, Noriaki Hirose, and Sergey Levine. ViNT: A Foundation Model for Visual Navigation. In *7th Annual Conference on Robot Learning (CoRL)*, 2023.

[94] Lin Shao, Fabio Ferreira, Mikael Jorda, Varun Nambiar, Jianlan Luo, Eugen Solowjow, Juan Aparicio Ojea, Oussama Khatib, and Jeannette Bohg. UniGrasp: Learning a unified model to grasp with multifingered robotic hands. *IEEE Robotics and Automation Letters*, 5(2):2286–2293, 2020.

[95] Lin Shao, Toki Migimatsu, Qiang Zhang, Karen Yang, and Jeannette Bohg. Concept2Robot: Learning manipulation concepts from instructions and human demonstrations. In *Proceedings of Robotics: Science and Systems (RSS)*, 2020.

[96] Pratyusha Sharma, Lekha Mohan, Lerrel Pinto, and Abhinav Gupta. Multiple interactions made easy (MIME): Large scale demonstrations data for imitation. In *Conference on robot learning*, pages 906–915. PMLR, 2018.

[97] Pratyusha Sharma, Deepak Pathak, and Abhinav Gupta. Third-person visual imitation learning via decoupled hierarchical controller. *Advances in Neural Information Processing Systems*, 32, 2019.

[98] P. Shilane, P. Min, M. Kazhdan, and T. Funkhouser. The princeton shape benchmark. In *Shape Modeling Applications*, pages 167–388, 2004.

[99] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Cliport: What and where pathways for robotic manipulation. In *Conference on Robot Learning*, pages 894–906. PMLR, 2022.

[100] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Perceiver-actor: A multi-task transformer for robotic manipulation. *Conference on Robot Learning (CoRL)*, 2022.

[101] Arjun Singh, James Sha, Karthik S. Narayan, Tudor Achim, and Pieter Abbeel. BigBIRD: A large-scale 3D database of object instances. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 509–516, 2014.

[102] Laura Smith, Nikita Dhawan, Marvin Zhang, Pieter Abbeel, and Sergey Levine. Avid: Learning multi-stage tasks via pixel-level translation of human videos. *arXiv preprint arXiv:1912.04443*, 2019.

[103] Simon Stepputtis, Joseph Campbell, Mariano Phielipp, Stefan Lee, Chitta Baral, and Heni Ben Amor. Language-conditioned imitation learning for robot manipulation tasks. *Advances in Neural Information Processing Systems*, 33:13139–13150, 2020.

[104] Austin Stone, Ted Xiao, Yao Lu, Keerthana Gopalakrishnan, Kuang-Huei Lee, Quan Vuong, Paul Wohlhart, Brianna Zitkovich, Fei Xia, Chelsea Finn, et al. Open-world object manipulation using pre-trained vision-language models. *arXiv preprint arXiv:2303.00905*, 2023.

[105] Mingxing Tan and Quoc Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

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

[107] Sai Vemprala, Rogerio Bonatti, Arthur Bucker, and Ashish Kapoor. ChatGPT for robotics: Design principles and model abilities. *Microsoft Auton. Syst. Robot. Res*, 2:20, 2023.

[108] Homer Walke, Kevin Black, Abraham Lee, Moo Jin Kim, Max Du, Chongyi Zheng, Tony Zhao, Philippe Hansen-Estruch, Quan Vuong, Andre He, Vivek Myers, Kuan Fang, Chelsea Finn, and Sergey Levine. Bridgedata v2: A dataset for robot learning at scale, 2023.

[109] Tobias Weyand, Andre Araujo, Bingyi Cao, and Jack Sim. Google landmarks dataset v2 - a large-scale benchmark for instance-level recognition and retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[110] Terry Winograd. Understanding natural language. *Cognitive Psychology*, 3(1):1–191, 1972.

[111] Walter Wohlkinger, Aitor Aldoma Buchaca, Radu Rusu, and Markus Vincze. 3DNet: Large-Scale Object Class Recognition from CAD Models. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2012.

[112] Baoyuan Wu, Weidong Chen, Yanbo Fan, Yong Zhang, Jinlong Hou, Jie Liu, and Tong Zhang. Tencent ML-images: A large-scale multi-label image database for visual representation learning. *IEEE Access*, 7, 2019.

[113] Yu Xiang, Wonhui Kim, Wei Chen, Jingwei Ji, Christopher Choy, Hao Su, Roozbeh Mottaghi, Leonidas Guibas, and Silvio Savarese. ObjectNet3D: A large scale database for 3d object recognition. In *European Conference on Computer Vision (ECCV)*, pages 160–176. Springer, 2016.

[114] Tete Xiao, Ilija Radosavovic, Trevor Darrell, and Jitendra Malik. Masked visual pre-training for motor control. *arXiv preprint arXiv:2203.06173*, 2022.

[115] Haoyu Xiong, Quanzhou Li, Yun-Chun Chen, Homanga Bharadhwaj, Samarth Sinha, and Animesh Garg. Learning by watching: Physical imitation of manipulation skills from human videos. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7827–7834. IEEE, 2021.

[116] Zhenjia Xu, Beichun Qi, Shubham Agrawal, and Shuran Song. Adagrasp: Learning an adaptive gripper-aware grasping policy. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4620–4626. IEEE, 2021.

[117] Jonathan Yang, Dorsa Sadigh, and Chelsea Finn. Polybot: Training one policy across robots while embracing variability. *arXiv preprint arXiv:2307.03719*, 2023.

[118] Kuan-Ting Yu, Maria Bauza, Nima Fazeli, and Alberto Rodriguez. More than a million ways to be pushed. a high-fidelity experimental dataset of planar pushing. In *2016 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 30–37. IEEE, 2016.

[119] Tianhe Yu, Chelsea Finn, Sudeep Dasari, Annie Xie, Tianhao Zhang, Pieter Abbeel, and Sergey Levine. One-shot imitation from observing humans via domain-adaptive meta-learning. *Robotics: Science and Systems XIV*, 2018.

[120] Kevin Zakka, Andy Zeng, Pete Florence, Jonathan Tompson, Jeannette Bohg, and Debidatta Dwibedi. XIRL: Cross-embodiment inverse reinforcement learning. *Conference on Robot Learning (CoRL)*, 2021.

[121] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3D ShapeNets: A deep representation for volumetric shapes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1912–1920, 2015.

[122] Gaoyue Zhou, Victoria Dean, Mohan Kumar Srirama, Aravind Rajeswaran, Jyothish Pari, Kyle Hatch, Aryan Jain, Tianhe Yu, Pieter Abbeel, Lerrel Pinto, Chelsea Finn, and Abhinav Gupta. Train offline, test online: A real robot learning benchmark, 2023.

[123] Yifan Zhou, Shubham Sonawani, Mariano Phielipp, Simon Stepputtis, and Heni Amor. Modularity through attention: Efficient training and transfer of language-conditioned policies for robot manipulation. In Karen Liu, Dana Kulic, and Jeff Ichnowski, editors, *Proceedings of The 6th Conference on Robot Learning*, volume 205 of *Proceedings of Machine Learning Research*, pages 1684–1695. PMLR, 14–18 Dec 2023.

[124] Yuxiang Zhou, Yusuf Aytar, and Konstantinos Bousmalis. Manipulator-independent representations for visual imitation. 2021.

[125] Xinghao Zhu, Ran Tian, Chenfeng Xu, Mingxiao Huo, Wei Zhan, Masayoshi Tomizuka, and Mingyu Ding. Fanuc manipulation: A dataset for learning-based manipulation with fanuc mate 200iD robot. https://sites.google.com/berkeley.edu/fanuc-manipulation, 2023.

[126] Yifeng Zhu, Abhishek Joshi, Peter Stone, and Yuke Zhu. Viola: Imitation learning for vision-based manipulation with object proposal priors, 2023.

Appendices

A Related Work

Transfer across embodiments. A number of prior works have studied methods for transfer across robot embodiments in simulation [19, 23, 34, 35, 39, 51, 65, 77, 87, 90, 92, 120, 123] and on real robots [11, 20, 38, 81, 84, 86, 117]. These methods often introduce mechanisms specifically designed to address the embodiment gap between different robots, such as shared action representations [65, 94], incorporating representation learning objectives [117, 120], adapting the learned policy on embodiment information [19, 34, 39, 94, 116], and decoupling robot and environment representations [38]. Prior work has provided initial demonstrations of X-embodiment training [84] and transfer [11, 81, 93] with transformer models. We investigate complementary architectures and provide complementary analyses, and, in particular, study the interaction between X-embodiment transfer and web-scale pretraining. Similarly, methods for transfer across human and robot embodiments also often employ techniques for reducing the embodiment gap, i.e. by translating between domains or learning transferable representations [5, 6, 9, 24, 41, 54, 88, 97, 102, 115, 119]. Alternatively, some works focus on sub-aspects of the problem such as learning transferable reward functions [3, 16, 50, 91, 95, 120], goals [124], dynamics models [89], or visual representations [7, 47, 59, 62, 70, 74, 82, 114] from human video data. Unlike most of these prior works, we directly train a policy on X-embodiment data, without any mechanisms to reduce the embodiment gap, and observe positive transfer by leveraging that data.

Large-scale robot learning datasets. The robot learning community has created open-source robot learning datasets, spanning grasping [10, 12, 22, 29, 31, 42, 44, 46, 53, 61, 79, 125], pushing interactions [20, 27, 32, 118], sets of objects and models [15, 25, 33, 45, 48, 69, 98, 101, 111, 113, 121], and teleoperated demonstrations [8, 14, 28, 30, 36, 58, 63, 96]. With the exception of RoboNet [20], these datasets contain data of robots of the same type, whereas we focus on data spanning multiple embodiments. The goal of our data repository is complementary to these efforts: we process and aggregate a large number of prior datasets into a single, standardized repository, called Open X-Embodiment, which shows how robot learning datasets can be shared in a meaningul and useful way.

Language-conditioned robot learning. Prior work has aimed to endow robots and other agents with the ability to understand and follow language instructions [17, 26, 49, 55, 60, 110], often by learning language-conditioned policies [14, 41, 67, 68, 73, 95, 100, 103]. We train language-conditioned policies via imitation learning like many of these prior works but do so using large-scale multi-embodiment demonstration data. Following previous works that leverage pre-trained language embeddings [2, 14, 37, 40, 41, 43, 57, 73, 95, 107] and pre-trained vision-language models [13, 71, 99, 104] in robotic imitation learning, we study both forms of pre-training in our experiments, specifically following the recipes of RT-1 [14] and RT-2 [13].