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

Anonymous Author(s)
Affiliation
Address
email

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.

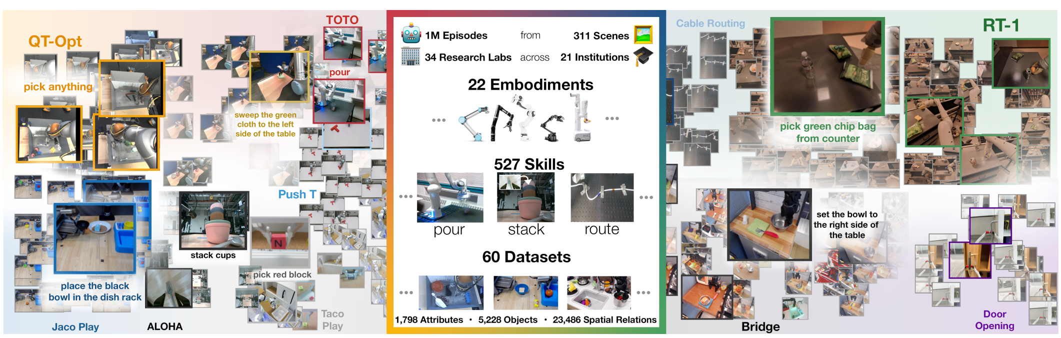

Figure 1: We propose an open, large-scale dataset for robot learning curated from 21 institutions across the globe. The dataset represents diverse behaviors, robot embodiments and environments.

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 [1]) trained on large datasets scraped from the web tend to outperform fixed-vocabulary models trained on more limited datasets, and large language models [2, 3] 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

Submitted to the 7th Conference on Robot Learning (CoRL 2023). Do not distribute.

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) [4, 5] and NLP (1.5B-4.5B) [6, 7]. 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 [8] and RT-2 [9] 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 Related Work

Transfer across embodiments. A number of prior works have studied methods for transfer across robot embodiments in simulation [10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22] and on real robots [23, 24, 25, 26, 27, 28, 29]. These methods often introduce mechanisms specifically designed to address the embodiment gap between different robots, such as shared action representations [14, 30], incorporating representation learning objectives [17, 26], adapting the learned policy on embodiment information [30, 31, 11, 18, 15], and decoupling robot and environment representations [24]. Prior work has provided initial demonstrations of X-embodiment training [27] and transfer [25, 32, 29] 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 [33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43]. Alternatively, some works focus on sub-aspects of the problem such as learning transferable reward functions [44, 17, 45, 46, 47, 48], goals [49], dynamics models [50], or visual representations [51, 52, 53, 54, 55, 56, 57, 58] 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.

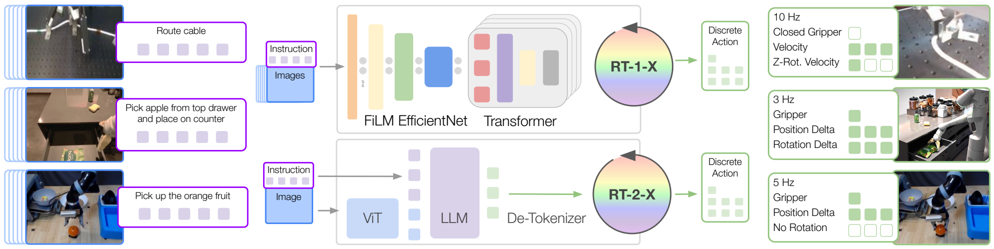

Figure 2: RT-1-X and RT-2-X both take images and a text instruction as input and output discretized end-effector actions. RT-1-X is an architecture designed for robotics, with a FiLM [113] conditioned EfficientNet [114] and a Transformer [115]. RT-2-X builds on a VLM backbone by representing actions as another language, and training action text tokens together with vision-language data.

Large-scale robot learning datasets. The robot learning community has created open-source robot learning datasets, spanning grasping [59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70], pushing interactions [71, 72, 73, 23], sets of objects and models [74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84], and teleoperated demonstrations [85, 86, 87, 8, 88, 89, 90, 91]. With the exception of RoboNet [23], 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 [92, 93, 94, 95, 96, 97], often by learning language-conditioned policies [45, 98, 99, 100, 101, 40, 102, 8]. 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 [103, 45, 104, 99, 40, 105, 106, 8, 107, 108] and pre-trained vision-language models [109, 110, 111, 9] in robotic imitation learning, we study both forms of pre-training in our experiments, specifically following the recipes of RT-1 [8] and RT-2 [9].

3 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 finetuning.

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 [112], which saves data in serialized `tfrecord` files and accommodates the various action spaces and input modalities of different robot setups.

4 RT-X Design

To evaluate how much X-embodiment training can improve the performance of learned policies on individual robots, we require models that have sufficient capacity to productively make use of such large and heterogeneous datasets. To that end, our experiments will build on two recently proposed

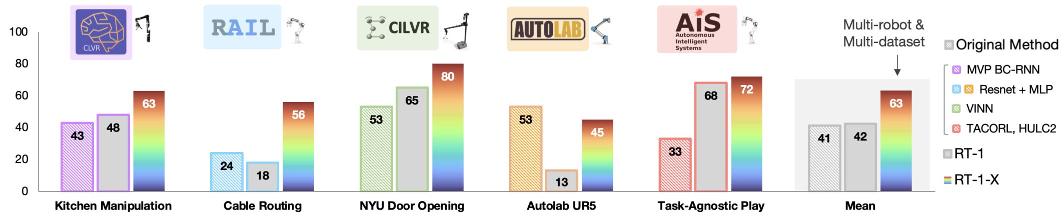

Figure 3: RT-1-X mean success rate is 50% higher than that of either the Original Method or RT-1. RT-1 and RT-1-X have the same network architecture. Therefore the performance increase can be attributed to co-training on the robotics data mixture. The lab logos indicate the physical location of real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

Transformer-based robotic policies: RT-1 [8] and RT-2 [9]. We briefly summarize the design of these models in this section, and discuss how we adapted them to the X-embodiment setting in our experiments.

4.1 Data format consolidation

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.

4.2 Policy architectures

We consider two model architectures in our experiments: (1) RT-1 [8], an efficient Transformer-based architecture designed for robotic control, and (2) RT-2 [9] a large vision-language model co-fine-tuned to output robot actions as natural language tokens. Both models take in a visual input and natural language instruction describing the task, and output a tokenized action. For each model, the action is tokenized into 256 bins uniformly distributed along each of eight dimensions; one dimension for terminating the episode and seven dimensions for end-effector movement. Although both architectures are described in detail in their original papers [8, 9], we provide a short summary of each below:

RT-1 [8] is a 35M parameter network built on a Transformer architecture [115] and designed for robotic control, as shown in Fig. 2. It takes in a history of 15 images along with the natural language. Each image is processed through an ImageNet-pretrained EfficientNet [114] and the natural language instruction is transformed into a USE [116] embedding. The visual and language representations are then interwoven via FiLM [113] layers, producing 81 vision-language tokens. These tokens are fed into a decoder-only Transformer, which outputs the tokenized actions.

RT-2 [9] is a family of large vision-language-*action* models (VLAs) trained on Internet-scale vision and language data along with robotic control data. RT-2 casts the tokenized actions to text tokens, e.g., a possible action may be "1 128 91 241 5 101 127". As such, any pretrained vision-language model (VLM [117, 118, 119]) can be finetuned for robotic control, thus leveraging the backbone of VLMs and transferring some of their generalization properties. In this work, we focus on the

Evaluation Setting	Bridge	Bridge	RT-1 paper 6 skills
Evaluation Location	IRIS (Stanford)	RAIL Lab (UCB)	Google Robotic Lab
Robot Embodiment	WidowX	WidowX	Google Robot
Original Method	LCBC [122]	LCBC [122]	-
Original Method	13%	13%	-
RT-1	40%	**30%**	**92%**
RT-1-X	27%	27%	73%
RT-2-X (55B)	**50%**	**30%**	**91%**

Table 1: Parameter count scaling experiment to assess the impact of capacity on absorbing large-scale diverse embodiment data. For these large-scale datasets (Bridge and RT-1 paper data), RT-1-X underfits and performs worse than the Original Method and RT-1. RT-2-X model with significantly many more parameters can obtain strong performance in these two evaluation scenarios.

RT-2-PaLI-X variant [117] built on a backbone of a visual model, ViT [120], and a language model, UL2 [121], and pretrained primarily on the WebLI [117] dataset.

4.3 Training and inference details

Both models use a standard categorical cross-entropy objective over their output space (discrete buckets for RT-1 and all possible language tokens for RT-2).

We define the robotics data mixture used across all of the experiments as the data from 9 manipulators, and taken from RT-1 [8], QT-Opt [65], Bridge [122], Task Agnostic Robot Play [123, 124], Jaco Play [125], Cable Routing [126], RoboTurk [127], NYU VINN [128], Austin VIOLA [129], Berkeley Autolab UR5 [130], TOTO [131] and Language Table [88] 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 [9]), with an approximately one to one split of the original VLM data and the robotics data mixture. Note that the robotics data mixture used in our experiments includes 9 embodiments which is fewer than the entire Open X-Embodiment dataset (22) – the practical reason for this difference is that we have continued to extend the dataset over time, and at the time of the experiments, the dataset above represented all of the data. In the future, we plan to continue training policies on the extended versions of the dataset as well as continue to grow the dataset together with the robot learning community.

At inference time, each model is run at the rate required for the robot (3-10 Hz), with RT-1 run locally and RT-2 hosted on a cloud service and queried over the network.

5 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? (3) What is the influence of different design dimensions, such as model size, model architecture or dataset composition, on performance and generalization capabilities of the resulting policy? To answer these questions we conduct the total number of 3600 evaluation trials across 6 different robots.

5.1 In-distribution performance across different embodiments

To assess the ability of our RT-X model variants to learn from X-embodiment data, we evaluate their performance on in-distribution tasks. We split our evaluation into two types of use cases: evaluation on domains that only have small-scale datasets (Fig. 3), where we would expect transfer from larger datasets to significantly improve performance, and evaluation on domains that have large-scale datasets (Table 1), where we expect further improvement to be more challenging. Note that we use the same robotics data *training* mixture (defined in Sec. 4.3) for all the evaluations presented in this section. For small-scale dataset experiments, we consider Kitchen Manipulation [125], Cable Routing [126], NYU Door Opening [128], AUTOLab UR5 [130], and Robot Play [132]. We use

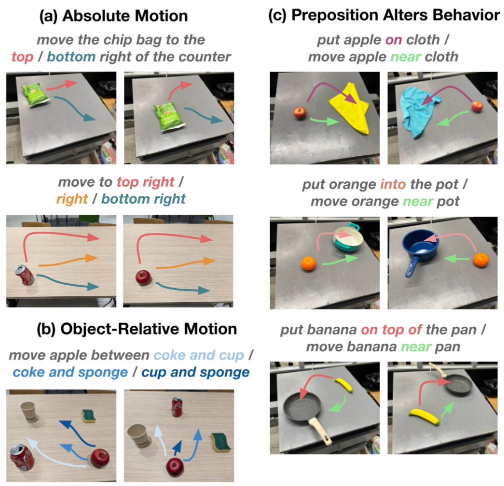

Figure 4: To assess transfer *between* embodiments, we evaluate the RT-2-X model on out-of-distribution skills. These skills are in the Bridge dataset, but not in the Google Robot dataset (the embodiment they are evaluated on).

the same evaluation and robot embodiment as in the respective publications. For large-scale dataset experiments, we consider Bridge [122] and RT-1 [8] for in-distribution evaluation and use their respective robots: WidowX and Google Robot.

For each small dataset domain, we compare the performance of the RT-1-X model, and for each large dataset we consider both the RT-1-X and RT-2-X models. For all experiments, the models are co-trained on the full X-embodiment dataset. 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 on multi-embodiment data leads to higher performance.

Small-scale dataset domains (Fig. 3). 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 domains with limited data benefit substantially from co-training on X-embodiment data.

Large-scale dataset domains (Table 1). In the large-dataset setting, the RT-1-X model does not outperform the RT-1 baseline trained on only the embodiment-specific dataset, which indicates underfitting for that model class. However, the larger RT-2-X model outperforms both the Original Method and RT-1 suggesting that X-robot training can improve performance in the data-rich domains, but only when utilizing a sufficiently high-capacity architecture.

5.2 Improved generalization to out-of-distribution settings

We now examine how X-embodiment training can enable 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.

Unseen objects, backgrounds and environments. We first conduct the same evaluation of generalization properties as proposed in [9], testing for the ability to manipulate unseen objects in unseen environments and against unseen backgrounds. We find that RT-2 and RT-2-X perform roughly on par (Table 2, rows (1) and (2), last column). This is not unexpected, since RT-2 already generalizes well (see [9]) along these dimensions due to its VLM backbone.

Emergent skills evaluation. To investigate the transfer of knowledge across robots, we conduct experiments with the Google Robot, assessing the performance on tasks like the ones shown in Fig. 4.

Row	Model	Size	History Length	Dataset	Co-Trained Web	Initial Checkpoint	Emergent Skills Evaluation	RT-2 Generalization Evaluation
(1)	RT-2	55B	none	Google Robot action	Yes	Web-pretrained	27.3%	**62%**
(2)	RT-2-X	55B	none	Robotics data	Yes	Web-pretrained	**75.8%**	**61%**
(3)	RT-2-X	55B	none	Robotics data except Bridge	Yes	Web-pretrained	42.8%	54%
(4)	RT-2-X	5B	2	Robotics data	Yes	Web-pretrained	44.4%	52%
(5)	RT-2-X	5B	none	Robotics data	Yes	Web-pretrained	14.5%	30%
(6)	RT-2-X	5B	2	Robotics data	No	From scratch	0%	1%
(7)	RT-2-X	5B	2	Robotics data	No	Web-pretrained	48.7%	47%

Table 2: Ablations to show the impact of design decisions on generalization (to unseen objects, backgrounds, and environments) and emergent skills (skills from other datasets on the Google Robot), showing the importance of Web-pretraining, model size, and history.

These tasks involve objects and skills that are not present in the RT-2 dataset but occur in the Bridge dataset [122] for a different robot (the *WidowX robot*). Results are shown in Table 2, 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 the 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.

Our next ablation involves removing the Bridge dataset from RT-2-X training: Row (3) shows the results for RT-2-X that includes all data used for RT-2-X except the Bridge dataset. This variation significantly reduces performance on the hold-out tasks, suggesting that transfer from the *WidowX* data may indeed be responsible for the additional skills that can be performed by RT-2-X with the Google Robot.

5.3 Design decisions

Lastly, we perform ablations to measure the influence of different design decisions on the generalization capabilities of our most performant RT-2-X model, which are presented in Table 2. We note that including a short history of images significantly improves generalization performance (row (4) vs row (5)). Similarly to the conclusions in the RT-2 paper [9], Web-based pre-training of the model is critical to achieving a high performance for the large models (row (4) vs row (6)). We also note that the $55B$ model has significantly higher success rate in the Emergent Skills compared to the $5B$ model (row (2) vs row (4)), demonstrating that higher model capacity enables higher degree of transfer across robotic datasets. Contrary to previous RT-2 findings, co-fine-tuning and fine-tuning have similar performance in both the Emergent Skills and Generalization Evaluation (row (4) vs row (7)), which we attribute to the fact that the robotics data used in RT-2-X is much more diverse than the previously used robotics datasets.

6 Discussion, Future Work, and Open Problems

We presented a consolidated dataset that combines data from 22 robotic embodiments collected through a collaboration between 21 institutions, demonstrating 527 skills (160266 tasks). We also presented an experimental demonstration that Transformer-based policies trained on this data can exhibit significant positive transfer between the different robots in the dataset. Our results showed that the RT-1-X policy has a 50% higher success rate than the original, state-of-the-art methods contributed by different collaborating institutions, while the bigger vision-language-model-based version (RT-2-X) demonstrated $\sim 3\times$ generalization improvements over a model trained only on data from the evaluation embodiment. In addition, we provided multiple resources for the robotics community to explore the X-embodiment robot learning research, including: the unified X-robot and X-institution dataset, sample code showing how to use the data, and the RT-1-X model to serve as a foundation for future exploration.

While RT-X demonstrates a step towards a X-embodied robot generalist, there are many more steps needed to make this future a reality. Our experiments have a number of limitations: it does not consider robots with very different sensing and actuation modalities, it does not study generalization to new robots, and it does not provide a decision criterion for when positive transfer does or does not happen. Studying these questions is an important direction for future work. We hope that this work will serve not only as an example that X-robot learning is feasible and practical, but also provide the tools to advance research in this direction in the future.

References

[1] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

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

[3] R. Anil, A. M. Dai, O. Firat, M. Johnson, D. Lepikhin, A. Passos, S. Shakeri, E. Taropa, P. Bailey, Z. Chen, et al. PaLM 2 technical report. *arXiv preprint arXiv:2305.10403*, 2023.

[4] T. Weyand, A. Araujo, B. Cao, and J. 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.

[5] B. Wu, W. Chen, Y. Fan, Y. Zhang, J. Hou, J. Liu, and T. Zhang. Tencent ML-images: A large-scale multi-label image database for visual representation learning. *IEEE Access*, 7, 2019.

[6] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. van Kleef, S. Auer, and C. Bizer. DBpedia - a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web*, 6(2):167–195, 2015. URL http://dblp.uni-trier.de/db/journals/semweb/semweb6.html#LehmannIJJKMHMK15.

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

[8] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, J. Hsu, et al. RT-1: Robotics transformer for real-world control at scale. *Robotics: Science and Systems (RSS)*, 2023.

[9] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, X. Chen, K. Choromanski, T. Ding, D. Driess, A. Dubey, C. Finn, et al. RT-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023.

[10] C. Devin, A. Gupta, T. Darrell, P. Abbeel, and S. 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.

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

[12] A. Sanchez-Gonzalez, N. Heess, J. T. Springenberg, J. Merel, M. Riedmiller, R. Hadsell, and P. Battaglia. Graph networks as learnable physics engines for inference and control. In J. Dy and A. 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. URL https://proceedings.mlr.press/v80/sanchez-gonzalez18a.html.

[13] D. Pathak, C. Lu, T. Darrell, P. Isola, and A. A. Efros. Learning to control self-assembling morphologies: a study of generalization via modularity. *Advances in Neural Information Processing Systems*, 32, 2019.

[14] R. Martín-Martín, M. Lee, R. Gardner, S. Savarese, J. Bohg, and A. 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.

[15] W. Huang, I. Mordatch, and D. Pathak. One policy to control them all: Shared modular policies for agent-agnostic control. In *ICML*, 2020.

[16] V. Kurin, M. Igl, T. Rocktäschel, W. Boehmer, and S. Whiteson. My body is a cage: the role of morphology in graph-based incompatible control. *arXiv preprint arXiv:2010.01856*, 2020.

[17] K. Zakka, A. Zeng, P. Florence, J. Tompson, J. Bohg, and D. Dwibedi. XIRL: Cross-embodiment inverse reinforcement learning. *Conference on Robot Learning (CoRL)*, 2021.

[18] A. Ghadirzadeh, X. Chen, P. Poklukar, C. Finn, M. Björkman, and D. 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.

[19] A. Gupta, L. Fan, S. Ganguli, and L. Fei-Fei. Metamorph: Learning universal controllers with transformers. In *International Conference on Learning Representations*, 2021.

[20] I. Schubert, J. Zhang, J. Bruce, S. Bechtle, E. Parisotto, M. Riedmiller, J. T. Springenberg, A. Byravan, L. Hasenclever, and N. Heess. A generalist dynamics model for control, 2023.

[21] D. Shah, A. Sridhar, A. Bhorkar, N. Hirose, and S. 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.

[22] Y. Zhou, S. Sonawani, M. Phielipp, S. Stepputtis, and H. Amor. Modularity through attention: Efficient training and transfer of language-conditioned policies for robot manipulation. In K. Liu, D. Kulic, and J. 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. URL https://proceedings.mlr.press/v205/zhou23b.html.

[23] S. Dasari, F. Ebert, S. Tian, S. Nair, B. Bucher, K. Schmeckpeper, S. Singh, S. Levine, and C. Finn. RoboNet: Large-scale multi-robot learning. In *Conference on Robot Learning (CoRL)*, volume 100, pages 885–897. PMLR, 2019.

[24] E. S. Hu, K. Huang, O. Rybkin, and D. Jayaraman. Know thyself: Transferable visual control policies through robot-awareness. In *International Conference on Learning Representations*, 2022.

[25] K. Bousmalis, G. Vezzani, D. Rao, C. Devin, A. X. Lee, M. Bauza, T. Davchev, Y. Zhou, A. Gupta, A. Raju, et al. RoboCat: A self-improving foundation agent for robotic manipulation. *arXiv preprint arXiv:2306.11706*, 2023.

[26] J. Yang, D. Sadigh, and C. Finn. Polybot: Training one policy across robots while embracing variability. *arXiv preprint arXiv:2307.03719*, 2023.

[27] S. Reed, K. Zolna, E. Parisotto, S. G. Colmenarejo, A. Novikov, G. Barth-maron, M. Giménez, Y. Sulsky, J. Kay, J. T. Springenberg, T. Eccles, J. Bruce, A. Razavi, A. Edwards, N. Heess, Y. Chen, R. Hadsell, O. Vinyals, M. Bordbar, and N. de Freitas. A generalist agent. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856.

[28] G. Salhotra, I.-C. A. Liu, and G. Sukhatme. Bridging action space mismatch in learning from demonstrations. *arXiv preprint arXiv:2304.03833*, 2023.

[29] I. Radosavovic, B. Shi, L. Fu, K. Goldberg, T. Darrell, and J. Malik. Robot learning with sensorimotor pre-training. In *Conference on Robot Learning*, 2023.

[30] L. Shao, F. Ferreira, M. Jorda, V. Nambiar, J. Luo, E. Solowjow, J. A. Ojea, O. Khatib, and J. Bohg. UniGrasp: Learning a unified model to grasp with multifingered robotic hands. *IEEE Robotics and Automation Letters*, 5(2):2286–2293, 2020. doi:10.1109/LRA.2020.2969946.

[31] Z. Xu, B. Qi, S. Agrawal, and S. Song. Adagrasp: Learning an adaptive gripper-aware grasping policy. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4620–4626. IEEE, 2021.

[32] D. Shah, A. Sridhar, N. Dashora, K. Stachowicz, K. Black, N. Hirose, and S. Levine. ViNT: A Foundation Model for Visual Navigation. In *7th Annual Conference on Robot Learning (CoRL)*, 2023.

[33] Y. Liu, A. Gupta, P. Abbeel, and S. 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.

[34] T. Yu, C. Finn, S. Dasari, A. Xie, T. Zhang, P. Abbeel, and S. Levine. One-shot imitation from observing humans via domain-adaptive meta-learning. *Robotics: Science and Systems XIV*, 2018.

[35] P. Sharma, D. Pathak, and A. Gupta. Third-person visual imitation learning via decoupled hierarchical controller. *Advances in Neural Information Processing Systems*, 32, 2019.

[36] L. Smith, N. Dhawan, M. Zhang, P. Abbeel, and S. Levine. Avid: Learning multi-stage tasks via pixel-level translation of human videos. *arXiv preprint arXiv:1912.04443*, 2019.

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

[38] K. Schmeckpeper, O. Rybkin, K. Daniilidis, S. Levine, and C. Finn. Reinforcement learning with videos: Combining offline observations with interaction. In *Conference on Robot Learning*, pages 339–354. PMLR, 2021.

[39] H. Xiong, Q. Li, Y.-C. Chen, H. Bharadhwaj, S. Sinha, and A. 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.

[40] E. Jang, A. Irpan, M. Khansari, D. Kappler, F. Ebert, C. Lynch, S. Levine, and C. Finn. BC-Z: Zero-shot task generalization with robotic imitation learning. In *Conference on Robot Learning (CoRL)*, pages 991–1002, 2021.

[41] S. Bahl, A. Gupta, and D. Pathak. Human-to-robot imitation in the wild. *Robotics: Science and Systems (RSS)*, 2022.

[42] M. Ding, Y. Xu, Z. Chen, D. D. Cox, P. Luo, J. B. Tenenbaum, and C. Gan. Embodied concept learner: Self-supervised learning of concepts and mapping through instruction following. In *Conference on Robot Learning*, pages 1743–1754. PMLR, 2023.

[43] S. Bahl, R. Mendonca, L. Chen, U. Jain, and D. 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.

[44] P. Sermanet, K. Xu, and S. Levine. Unsupervised perceptual rewards for imitation learning. *arXiv preprint arXiv:1612.06699*, 2016.

[45] L. Shao, T. Migimatsu, Q. Zhang, K. Yang, and J. Bohg. Concept2Robot: Learning manipulation concepts from instructions and human demonstrations. In *Proceedings of Robotics: Science and Systems (RSS)*, 2020.

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

[47] S. Kumar, J. Zamora, N. Hansen, R. Jangir, and X. Wang. Graph inverse reinforcement learning from diverse videos. In *Conference on Robot Learning*, pages 55–66. PMLR, 2023.

[48] M. Alakuijala, G. Dulac-Arnold, J. Mairal, J. Ponce, and C. 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.

[49] Y. Zhou, Y. Aytar, and K. Bousmalis. Manipulator-independent representations for visual imitation. 2021.

[50] K. Schmeckpeper, A. Xie, O. Rybkin, S. Tian, K. Daniilidis, S. Levine, and C. Finn. Learning predictive models from observation and interaction. In *European Conference on Computer Vision*, pages 708–725. Springer, 2020.

[51] S. Nair, A. Rajeswaran, V. Kumar, C. Finn, and A. Gupta. R3m: A universal visual representation for robot manipulation. In *CoRL*, 2022.

[52] T. Xiao, I. Radosavovic, T. Darrell, and J. Malik. Masked visual pre-training for motor control. *arXiv preprint arXiv:2203.06173*, 2022.

[53] I. Radosavovic, T. Xiao, S. James, P. Abbeel, J. Malik, and T. Darrell. Real-world robot learning with masked visual pre-training. In *Conference on Robot Learning*, 2022.

[54] Y. J. Ma, S. Sodhani, D. Jayaraman, O. Bastani, V. Kumar, and A. Zhang. Vip: Towards universal visual reward and representation via value-implicit pre-training. *arXiv preprint arXiv:2210.00030*, 2022.

[55] A. Majumdar, K. Yadav, S. Arnaud, Y. J. Ma, C. Chen, S. Silwal, A. Jain, V.-P. Berges, P. Abbeel, J. Malik, et al. Where are we in the search for an artificial visual cortex for embodied intelligence? *arXiv preprint arXiv:2303.18240*, 2023.

[56] S. Karamcheti, S. Nair, A. S. Chen, T. Kollar, C. Finn, D. Sadigh, and P. Liang. Language-driven representation learning for robotics. *arXiv preprint arXiv:2302.12766*, 2023.

[57] Y. Mu, S. Yao, M. Ding, P. Luo, and C. Gan. EC2: Emergent communication for embodied control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6704–6714, 2023.

[58] S. Bahl, R. Mendonca, L. Chen, U. Jain, and D. 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.

[59] Y. Jiang, S. Moseson, and A. 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.

[60] L. Pinto and A. K. 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.

[61] D. Kappler, J. Bohg, and S. Schaal. Leveraging big data for grasp planning. In *ICRA*, pages 4304–4311, 2015. doi:10.1109/ICRA.2015.7139793.

[62] J. Mahler, J. Liang, S. Niyaz, M. Laskey, R. Doan, X. Liu, J. A. Ojea, and K. 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.

[63] A. Depierre, E. Dellandréa, and L. 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.

[64] S. Levine, P. Pastor, A. Krizhevsky, J. Ibarz, and D. 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.

[65] D. Kalashnikov, A. Irpan, P. Pastor, J. Ibarz, A. Herzog, E. Jang, D. Quillen, E. Holly, M. Kalakrishnan, V. Vanhoucke, et al. QT-Opt: Scalable deep reinforcement learning for vision-based robotic manipulation. *arXiv preprint arXiv:1806.10293*, 2018.

[66] S. Brahmbhatt, C. Ham, C. Kemp, and J. Hays. Contactdb: Analyzing and predicting grasp contact via thermal imaging, 04 2019.

[67] H.-S. Fang, C. Wang, M. Gou, and C. 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.

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

[69] K. Bousmalis, A. Irpan, P. Wohlhart, Y. Bai, M. Kelcey, M. Kalakrishnan, L. Downs, J. Ibarz, P. Pastor, K. Konolige, S. Levine, and V. Vanhoucke. Using simulation and domain adaptation to improve efficiency of deep robotic grasping. In *ICRA*, pages 4243–4250, 2018. doi:10.1109/ICRA.2018.8460875.

[70] X. Zhu, R. Tian, C. Xu, M. Huo, W. Zhan, M. Tomizuka, and M. Ding. Fanuc manipulation: A dataset for learning-based manipulation with fanuc mate 200iD robot. `https://sites.google.com/berkeley.edu/fanuc-manipulation`, 2023.

[71] K.-T. Yu, M. Bauza, N. Fazeli, and A. 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.

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

[73] F. Ebert, C. Finn, S. Dasari, A. Xie, A. Lee, and S. Levine. Visual foresight: Model-based deep reinforcement learning for vision-based robotic control. *arXiv preprint arXiv:1812.00568*, 2018.

[74] P. Shilane, P. Min, M. Kazhdan, and T. Funkhouser. The princeton shape benchmark. In *Shape Modeling Applications*, pages 167–388, 2004. ISBN 978-0-7695-2075-9. doi:10.1109/SMI.2004.1314504.

[75] W. Wohlkinger, A. Aldoma Buchaca, R. Rusu, and M. Vincze. 3DNet: Large-Scale Object Class Recognition from CAD Models. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2012.

[76] A. Kasper, Z. Xue, and R. 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. doi:10.1177/0278364912445831.

[77] A. Singh, J. Sha, K. S. Narayan, T. Achim, and P. Abbeel. BigBIRD: A large-scale 3D database of object instances. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 509–516, 2014. ISBN 978-1-4799-3685-4. doi:10.1109/ICRA.2014.6906903.

[78] B. Calli, A. Walsman, A. Singh, S. Srinivasa, P. Abbeel, and A. M. Dollar. Benchmarking in manipulation research: Using the Yale-CMU-Berkeley object and model set. *IEEE Robotics & Automation Magazine*, 22(3):36–52, 2015.

[79] Zhirong Wu, S. Song, A. Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and J. Xiao. 3D ShapeNets: A deep representation for volumetric shapes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1912–1920, 2015. ISBN 978-1-4673-6964-0. doi:10.1109/CVPR.2015.7298801.

[80] Y. Xiang, W. Kim, W. Chen, J. Ji, C. Choy, H. Su, R. Mottaghi, L. Guibas, and S. Savarese. ObjectNet3D: A large scale database for 3d object recognition. In *European Conference on Computer Vision (ECCV)*, pages 160–176. Springer, 2016.

[81] D. Morrison, P. Corke, and J. Leitner. Egad! an evolved grasping analysis dataset for diversity and reproducibility in robotic manipulation. *IEEE Robotics and Automation Letters*, 5(3): 4368–4375, 2020. doi:10.1109/LRA.2020.2992195.

[82] R. Gao, Y.-Y. Chang, S. Mall, L. Fei-Fei, and J. Wu. ObjectFolder: A dataset of objects with implicit visual, auditory, and tactile representations. In *Conference on Robot Learning*, pages 466–476, 2021.

[83] L. Downs, A. Francis, N. Koenig, B. Kinman, R. Hickman, K. Reymann, T. B. McHugh, and V. 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.

[84] D. Kalashnikov, J. Varley, Y. Chebotar, B. Swanson, R. Jonschkowski, C. Finn, S. Levine, and K. Hausman. MT-Opt: Continuous multi-task robotic reinforcement learning at scale. *arXiv preprint arXiv:2104.08212*, 2021.

[85] P. Sharma, L. Mohan, L. Pinto, and A. Gupta. Multiple interactions made easy (MIME): Large scale demonstrations data for imitation. In *Conference on robot learning*, pages 906–915. PMLR, 2018.

[86] A. Mandlekar, J. Booher, M. Spero, A. Tung, A. Gupta, Y. Zhu, A. Garg, S. Savarese, and L. 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.

[87] F. Ebert, Y. Yang, K. Schmeckpeper, B. Bucher, G. Georgakis, K. Daniilidis, C. Finn, and S. Levine. Bridge data: Boosting generalization of robotic skills with cross-domain datasets. In *Robotics: Science and Systems (RSS) XVIII*, 2022.

[88] C. Lynch, A. Wahid, J. Tompson, T. Ding, J. Betker, R. Baruch, T. Armstrong, and P. Florence. Interactive language: Talking to robots in real time. *IEEE Robotics and Automation Letters*, 2023.

[89] H.-S. Fang, H. Fang, Z. Tang, J. Liu, J. Wang, H. Zhu, and C. Lu. RH20T: A robotic dataset for learning diverse skills in one-shot. In *RSS 2023 Workshop on Learning for Task and Motion Planning*, 2023.

[90] H. Bharadhwaj, J. Vakil, M. Sharma, A. Gupta, S. Tulsiani, and V. Kumar. RoboAgent: Towards sample efficient robot manipulation with semantic augmentations and action chunking. *arxiv*, 2023.

[91] M. Heo, Y. Lee, D. Lee, and J. J. Lim. Furniturebench: Reproducible real-world benchmark for long-horizon complex manipulation. In *Robotics: Science and Systems*, 2023.

[92] T. Winograd. Understanding natural language. *Cognitive Psychology*, 3(1):1–191, 1972. ISSN 0010-0285. doi:https://doi.org/10.1016/0010-0285(72)90002-3. URL https://www.sciencedirect.com/science/article/pii/0010028572900023.

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

[94] T. Kollar, S. Tellex, D. Roy, and N. Roy. Toward understanding natural language directions. In *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 259–266, 2010.

[95] D. L. Chen and R. 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.

[96] F. Duvallet, J. Oh, A. Stentz, M. Walter, T. Howard, S. Hemachandra, S. Teller, and N. Roy. Inferring maps and behaviors from natural language instructions. In *International Symposium on Experimental Robotics (ISER)*, 2014.

[97] J. Luketina, N. Nardelli, G. Farquhar, J. N. Foerster, J. Andreas, E. Grefenstette, S. Whiteson, and T. Rocktäschel. A survey of reinforcement learning informed by natural language. In *IJCAI*, 2019.

[98] S. Stepputtis, J. Campbell, M. Phielipp, S. Lee, C. Baral, and H. Ben Amor. Language-conditioned imitation learning for robot manipulation tasks. *Advances in Neural Information Processing Systems*, 33:13139–13150, 2020.

[99] S. Nair, E. Mitchell, K. Chen, S. Savarese, C. 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.

[100] O. Mees, L. Hermann, E. Rosete-Beas, and W. Burgard. CALVIN: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks. *IEEE Robotics and Automation Letters*, 2022.

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

[102] M. Shridhar, L. Manuelli, and D. Fox. Perceiver-actor: A multi-task transformer for robotic manipulation. *Conference on Robot Learning (CoRL)*, 2022.

[103] F. Hill, S. Mokra, N. Wong, and T. Harley. Human instruction-following with deep reinforcement learning via transfer-learning from text. *arXiv preprint arXiv:2005.09382*, 2020.

[104] C. Lynch and P. Sermanet. Grounding language in play. *Robotics: Science and Systems (RSS)*, 2021.

[105] M. Ahn, A. Brohan, N. Brown, Y. Chebotar, O. Cortes, B. David, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, et al. Do as I can, not as I say: Grounding language in robotic affordances. *Conference on Robot Learning (CoRL)*, 2022.

[106] Y. Jiang, A. Gupta, Z. Zhang, G. Wang, Y. Dou, Y. Chen, L. Fei-Fei, A. Anandkumar, Y. Zhu, and L. Fan. VIMA: General robot manipulation with multimodal prompts. *International Conference on Machine Learning (ICML)*, 2023.

[107] S. Vemprala, R. Bonatti, A. Bucker, and A. Kapoor. ChatGPT for robotics: Design principles and model abilities. *Microsoft Auton. Syst. Robot. Res*, 2:20, 2023.

[108] W. Huang, C. Wang, R. Zhang, Y. Li, J. Wu, and L. Fei-Fei. VoxPoser: Composable 3d value maps for robotic manipulation with language models. *arXiv preprint arXiv:2307.05973*, 2023.

[109] M. Shridhar, L. Manuelli, and D. Fox. Cliport: What and where pathways for robotic manipulation. In *Conference on Robot Learning*, pages 894–906. PMLR, 2022.

[110] A. Stone, T. Xiao, Y. Lu, K. Gopalakrishnan, K.-H. Lee, Q. Vuong, P. Wohlhart, B. Zitkovich, F. Xia, C. Finn, et al. Open-world object manipulation using pre-trained vision-language models. *arXiv preprint arXiv:2303.00905*, 2023.

[111] Y. Mu, Q. Zhang, M. Hu, W. Wang, M. Ding, J. Jin, B. Wang, J. Dai, Y. Qiao, and P. Luo. EmbodiedGPT: Vision-language pre-training via embodied chain of thought. *arXiv preprint arXiv:2305.15021*, 2023.

[112] S. Ramos, S. Girgin, L. Hussenot, D. Vincent, H. Yakubovich, D. Toyama, A. Gergely, P. Stanczyk, R. Marinier, J. Harmsen, O. Pietquin, and N. Momchev. RLDS: an ecosystem to generate, share and use datasets in reinforcement learning, 2021.

[113] E. Perez, F. Strub, H. de Vries, V. Dumoulin, and A. Courville. Film: Visual reasoning with a general conditioning layer, 2017.

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

[115] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[116] D. Cer, Y. Yang, S. yi Kong, N. Hua, N. Limtiaco, R. S. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, Y.-H. Sung, B. Strope, and R. Kurzweil. Universal sentence encoder, 2018.

[117] X. Chen, J. Djolonga, P. Padlewski, B. Mustafa, S. Changpinyo, J. Wu, C. R. Ruiz, S. Goodman, X. Wang, Y. Tay, S. Shakeri, M. Dehghani, D. Salz, M. Lucic, M. Tschannen, A. Nagrani, H. Hu, M. Joshi, B. Pang, C. Montgomery, P. Pietrzyk, M. Ritter, A. Piergiovanni, M. Minderer, F. Pavetic, A. Waters, G. Li, I. Alabdulmohsin, L. Beyer, J. Amelot, K. Lee, A. P. Steiner, Y. Li, D. Keysers, A. Arnab, Y. Xu, K. Rong, A. Kolesnikov, M. Seyedhosseini, A. Angelova, X. Zhai, N. Houlsby, and R. Soricut. Pali-x: On scaling up a multilingual vision and language model, 2023.

[118] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, R. Ring, E. Rutherford, S. Cabi, T. Han, Z. Gong, S. Samangooei, M. Monteiro, J. Menick, S. Borgeaud, A. Brock, A. Nematzadeh, S. Sharifzadeh, M. Binkowski, R. Barreira, O. Vinyals, A. Zisserman, and K. Simonyan. Flamingo: a visual language model for few-shot learning, 2022.

[119] D. Driess, F. Xia, M. S. M. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. Yu, W. Huang, Y. Chebotar, P. Sermanet, D. Duckworth, S. Levine, V. Vanhoucke, K. Hausman, M. Toussaint, K. Greff, A. Zeng, I. Mordatch, and P. Florence. PaLM-E: An embodied multimodal language model, 2023.

[120] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

[121] Y. Tay, M. Dehghani, V. Q. Tran, X. Garcia, J. Wei, X. Wang, H. W. Chung, S. Shakeri, D. Bahri, T. Schuster, H. S. Zheng, D. Zhou, N. Houlsby, and D. Metzler. UL2: Unifying language learning paradigms, 2023.

[122] H. Walke, K. Black, A. Lee, M. J. Kim, M. Du, C. Zheng, T. Zhao, P. Hansen-Estruch, Q. Vuong, A. He, V. Myers, K. Fang, C. Finn, and S. Levine. Bridgedata v2: A dataset for robot learning at scale, 2023.

[123] E. Rosete-Beas, O. Mees, G. Kalweit, J. Boedecker, and W. Burgard. Latent plans for task agnostic offline reinforcement learning. In *Proceedings of the 6th Conference on Robot Learning (CoRL)*, 2022.

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

[125] S. Dass, J. Yapeter, J. Zhang, J. Zhang, K. Pertsch, S. Nikolaidis, and J. J. Lim. CLVR jaco play dataset, 2023. URL https://github.com/clvrai/clvr_jaco_play_dataset.

[126] J. Luo, C. Xu, X. Geng, G. Feng, K. Fang, L. Tan, S. Schaal, and S. Levine. Multi-stage cable routing through hierarchical imitation learning. *arXiv preprint arXiv:2307.08927*, 2023.

[127] A. Mandlekar, Y. Zhu, A. Garg, J. Booher, M. Spero, A. Tung, J. Gao, J. Emmons, A. Gupta, E. Orbay, S. Savarese, and L. Fei-Fei. RoboTurk: A crowdsourcing platform for robotic skill learning through imitation. *CoRR*, abs/1811.02790, 2018. URL http://arxiv.org/abs/1811.02790.

[128] J. Pari, N. M. Shafiullah, S. P. Arunachalam, and L. Pinto. The surprising effectiveness of representation learning for visual imitation, 2021.

[129] Y. Zhu, A. Joshi, P. Stone, and Y. Zhu. Viola: Imitation learning for vision-based manipulation with object proposal priors, 2023.

[130] L. Y. Chen, S. Adebola, and K. Goldberg. Berkeley UR5 demonstration dataset. https://sites.google.com/view/berkeley-ur5/home.

[131] G. Zhou, V. Dean, M. K. Srirama, A. Rajeswaran, J. Pari, K. Hatch, A. Jain, T. Yu, P. Abbeel, L. Pinto, C. Finn, and A. Gupta. Train offline, test online: A real robot learning benchmark, 2023.

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