Dimensions of Diversity in Robot Datasets: Literature Analysis and Recommendations

Wenxi Wu, Michelle Nwachukwu, Atmadeep Ghoshal, Madeleine Waller, Martim Brandão

***Abstract*— Datasets are essential for building robotic policies that can generalize on new tasks. Recent studies show that a lack of diversity in data can lead to performance disparities and discrimination against underrepresented groups, therefore making diversity an important requirement of robot datasets. While many authors claim their datasets are 'diverse' there is currently a lack of understanding of what diversity means in the context of robot datasets. Therefore, in this paper we conduct a systematic analysis of literature on robot manipulation and collaboration datasets to investigate what is meant by 'diversity' when it is claimed by authors of the datasets. We identify five dimensions of diversity in the context of robot manipulation and collaboration: object, task, environment, platform, and human diversity. Then, we identify various limitations with current practices related to diversity, and offer several recommendations: creating datasets with clear definitions and scales of diversity, with greater cultural representation including from Global South cultures, the inclusion of human participants with varying motion characteristics and better reporting of human characteristics.**

I. INTRODUCTION

Robot datasets play a vital role in training, validating, and testing the behavior of robotic systems. While large-scale datasets enhance the capability of robot policies to generalize on different tasks, dataset developers face a complex problem as environments and tasks can present in a seemingly unlimited number of ways. Thus, creating diverse robot datasets is an important challenge. Dataset developers often describe their datasets as 'diverse', but this term has come to mean different things, e.g., including a range of tasks in datasets to enhance generalization ability, or including diverse human demonstrators in the dataset curation process. Diversity is often deployed as a buzzword that has no clear definition or verifiable metric. In this paper, we identify the different dimensions of diversity present in robot datasets, specifically those used for manipulation and collaboration with humans. We investigate what authors of robot datasets mean when they claim a dataset is 'diverse', we show how definitions are currently lacking, and make several recommendations for better practices. We conduct this analysis on datasets of robot manipulation and collaboration, looking at datasets published in the past five years in prominent robotics conferences.

II. RELATED WORK

A. Robot datasets

The development of robot learning approaches has traditionally relied on small datasets that contain one or a few tasks [1], [2], [3]. Some datasets cover multiple tasks [4], [5], but their utility is limited by the fact that the data is collected in single specific environments, restricting cross-environment generalization [6]. In recent years, researchers have aimed to address this generalizability issue by creating robot datasets consisting of multi-task and multi-domain robot data. For example, DROID [7] contains 76,000 demonstration trajectories collected across 564 scenes and 86 tasks in different setups. In this dataset, industrial office scenes are the most featured with more than 200 scenes. The authors report the dataset is diverse in terms of tasks, objects, scenes, viewpoints and interaction locations. RH20T [8] contains over 140 tasks with multi-modal input including visual, audio, and tactile information. The data is collected on robots, tasks, modalities and multiple sets of robot hardware. BridgeV2 [6] has data collected in 24 scenes. These datasets expand the range and quantity of open-source robotic data in multiple tasks and domains, and could be considered more diverse as they include more tasks, scenes and setups. However, diversity can be defined differently and clarifying these definitions for robot datasets has not been fully explored.

B. Diversity in ML

Diversity in datasets used for machine learning (ML) has been more widely explored [9]. Datasets often reflect political and social states of the world which can lead to bias in outcomes of trained models, potentially causing harmful effects if these outcomes are used to influence individual decisions or policies [10], [11]. For example, a hiring tool was shown to be biased against women due to a lack of diversity in the training dataset which reflected the previously hired individuals who were predominately male [12]. Existing research has focused on the difficult task of defining and quantifying diversity and related concepts for tabular, text, or image datasets, but there has been a lack of consensus on these definitions. Various notions of diversity have been proposed [13], including variety in dataset content (e.g., background elements in images [14], representation of languages in text [15]), the diversity of data sources (e.g., different web sources [16]), variation in topic areas (e.g., artistic styles of images [17], disciplines that text is taken from [18]), representation of human subjects (e.g., proportions of individuals with protected characteristics in images [19], descriptions of individuals in text [20]), and the diversity of dataset annotator backgrounds [21].

Measures quantifying these notions of diversity are difficult to define [22]. Some works focus on defining diversity with respect to the representation of human subjects [23], [24], e.g., ensuring a balance of individuals with protected characteristics in the training data. A lack of representation of diverse human

All authors are with King's College London, UK.

This work was supported by the UKRI Centre for Doctoral Training in Safe and Trusted Artificial Intelligence [EP/S023356/1].

TABLE I: Presence of Different Dimensions of Diversity in Robot Manipulation & Collaboration Datasets 2020 - 2024: object, task, environment, platform and human.

	Obj	Task	Env	Platf	Human
ARMBench[m][30]	✓				
RT20H [m][31]	✓	✓	✓	✓	?
Gaze Attention[m][32]	✓				
TransCG[m][33]	✓				
CALVIN[m] [34]	✓				
Open X-Embodiment[m][35]	✓	✓	✓	✓	?
PriMA-Care[c] [36]	✓				✓
Throw&catch[c,m] [37]	✓				✓
TRansPose[m] [38]	✓		✓		
Grasp-Anything[m][39]	✓				
Google Scanned Objects[m][40]	✓				
ACRONYM[m] [41]	✓				
Folding Demonstration[m][42]	✓				?

subjects can lead performance disparities on marginalized groups of people that are less seen in the dataset resulting in downstream harms [25], but these concerns are less frequently the focus of diverse dataset creation [13].

C. Diversity in HCI and HRI

Similarly to ML datasets, a lack of diversity has been found in Human-Computer Interaction (HCI) and Human-Robot Interaction (HRI) datasets [26], [27]. Creating datasets for HCI and HRI particularly often involves human participants interacting with computers or robots. Who these participants are influences the dataset and thus any system that uses it, leading also to issues of power [28]. One study finds that there is an over-representation of men in research participants in HRI and that it is not common practice to report the demographics of the human participants [29]. Another demonstrates there is a bias in the human participants selected for user studies towards participants that are "Western, Educated, Industrial, Rich and Democratic" [27]. These papers focus on the demographics of human participants, which supports our discussions and recommendations. However, our paper differs by focusing on identifying the dimensions of diversity in robot datasets and their limitations, specifically in robot manipulation and collaboration datasets.

III. Dimensions of Diversity in Robot Datasets

To investigate what is meant by 'diversity' when it is claimed by authors of robot datasets, we conduct a systematic literature analysis. We searched for papers presenting new datasets in robot manipulation and human-robot collaboration, with keywords 'manipulation' and 'collaboration', published in the past five years (2020-2024) in well-established venues (ICRA, IROS, RAL, IJRR, HRI, ROMAN). We selected papers that, in the title or abstract, claimed their datasets contain 'diverse' data, using the word 'diversity' or 'diverse'. We then extracted the definitions of diversity used and identified five dimensions of diversity: diversity in objects, tasks, environments, platforms, and humans.

Table I shows the datasets found and which dimensions of diversity they reference. Out of the datasets found, 12 involve robot manipulation (noted with 'm' in the table), and 2 involve collaboration (noted with 'c').

A. Diversity in objects

From the survey result in Table I, object diversity is by far the most common type of diversity definition mentioned in papers that present robot manipulation and collaboration datasets—every paper mentions object diversity. For example, X-Embodiment [35] contains a range of objects such as furniture, appliances, and utensils. The Throw&catch dataset [37] includes a range of 52 objects, from rigid objects such as cans, to soft objects like pillows. Transpose [38] contains 99 transparent objects spanning household items, laboratory equipment and recyclable trash. The motivation to include this type of diversity is to allow generalization across interaction with multiple objects.

B. Diversity in tasks

While many datasets focus on a small number of different tasks [43], [44], larger datasets with multiple tasks are emerging. RT20H [31] contains 147 tasks, such as watering plants, slicing vegetables and wiping tables. X-embodiment [35] pools datasets that include many tasks together, e.g., picking, moving, placing, sliding etc. The motivation to include this type of diversity in datasets is often to enable the trained robot to learn different skills and adapt to more complex tasks, thus improving its success rates [31]. It remains a challenge to generate diverse task data as it requires abundant investment into equipment and labour.

C. Diversity in environments

Many tasks in robot datasets involve robots interacting with an environment. Some datasets include varied scenes for robots to perform the same tasks on. Dataset developers alter the texture and materials of the objects to create variations in the environment, e.g., RT20H [31] include 50 table covers and task-irrelevant objects in the background to create distractions. TransPose [38] creates different illumination in the environment for the transparent objects. The variety of scenes contributes to the generalizability in unseen environments [31].

D. Diversity in robots and platforms

Robot learning algorithms are not only sensitive to the quality and quantity of demonstrations, but also depend on the platform (robot and interface) used. Some robot datasets include data generated using different robots and data collection platforms. RT20H [31] contains data collected from 4 different robot arms with grippers. RoboTurk [45] has introduced a crowd-sourcing platform to collect trajectories on mobile devices through teleoperation. RT-X [35], a high-capacity model trained on X-Embodiment dataset, gives an example of a policy trained on data from multiple platforms to enhance the task capabilities. The motivation for datasets to include this type of diversity is to obtain policies that work across different platforms. Additionally, collecting data on different robot platforms makes the dataset easier to apply in other laboratories and therefore reduces the investment into replicating the setup in the data when it is used in other laboratories, making it more accessible to different areas and inclusive for budget-sensitive institutions and regions [6].

E. Diversity of human participants

Many robot manipulation datasets involve human demonstration or annotation. Dataset PRIMA-Care [36] reports the demographic information of the human demonstrators. Some papers report the *number* of humans involved in data generation, but no information about their personal characteristics (noted in the table with questionmark '?'). For example, RT20T [31] reports the total number of demonstrators. For the Throw&catch dataset [37], information is included about the gender and age of the human subjects. The variety of different ways tasks can be completed plays an important role in shaping the learned policies for robots. As a result, the diversity of human participants has the potential to influence the performance and generalizability of robots. As we will discuss next, current dataset are lacking on human diversity and its reporting, for example related to hand dominance, age, gender, disabilities, etc.

IV. Discussion & Recommendations

From our literature analysis we found that, when authors claim a robot dataset is 'diverse', they refer to either object diversity, task diversity, environment diversity, platform diversity, or diversity in human participants. Table I shows the presence of these dimensions of diversity in different robot manipulation and collaboration datasets. In this section, we present five recommendations for future development and reporting of robot datasets and discuss each one in relation to the dimensions of diversity we have identified.

> ***R*1: Unambiguous indicator of diversity**
>
> Dataset curators should clearly state the definition of diversity they use.

Our first recommendation **R1** is based on our findings from Section III. Throughout the papers surveyed, there is a lack of clarity when defining diversity for the curation of datasets. It is often not reported how diversity has been defined and there is lack of information about the dataset curation process that is necessary for others to evaluate it. Claims of 'diversity' usually refer to the types of tasks and sometimes the types of domains. This is only a small subset of categories that the term could potentially cover. For example, the quality of the demonstrated trajectories has a great impact on learning outcomes because the demonstrations can incorporate biases from the human operators. However, the diversity of demonstrators is rarely discussed.

The unclear definition of diversity across many datasets makes the indicator difficult to understand. For example, dataset Mt-opt [4] and RH20T [8] both claim they have proposed a dataset that contains diverse skills. However, Mt-opt contains tasks of lifting and placing a range of objects from a single type of robot model, while RH20T has data collected from multiple robot models and have selected more tasks from other benchmarks. The absence of indication and measurement definition causes confusion and does not help enhance diversity in datasets. In contrast, the curators of dataset Throw&catch [37] explain in the paper that the diversity of data refers to the objects and humans where the category of objects and demographics of demonstrators are reported in the paper.

> ***R*2: Scale of diversity**
>
> A metric or scale should be defined to be able to evaluate and compare the diversity of the dataset, according to the specified definition of diversity.

Based on similar motivation as **R1**, our second recommendation **R2** further highlights the need to explicitly specify how diversity is measured. In addition to being precise with the definition of diversity, the scale of diversity for the dimensions in Table I (Object, Task, Environment, Platform and Human) should also be reported along with new datasets. As previously discussed, quantification of diversity can be difficult. As such, we recommend that further work should be done to create a unified standard scale for diversity across these dimensions. This recommendation is inspired by the field of algorithmic fairness, which employs standard metrics widely used to evaluate and compare datasets [24]. A unified standard, even as simple as a count of the variety of tasks completed in a dataset, or counts of categories of tasks from an accepted task taxonomy, would allow for datasets to be compared. More diversity in a dataset does not necessarily constitute better performance for a specific task, e.g., if a robot only ever encounters a specific set of objects, there is no need to include objects outside of that set. However, robot policies trained on data with more diversity in tasks are likely to have a better performance to generalize on different tasks and unseen environments.

> ***R*3: Inclusion of culturally diverse scenes**
>
> There should be consideration into the inclusion of scenes from countries that are considered part of the Global South. There should be clear documentation of socio-geographical context of the scenes included in the datasets.

Our third recommendation **R3** relates to improving diversity of environments and scenes. Current robot datasets predominantly focus on constrained indoor environments such as offices, kitchens, and living rooms. Public places with more complicated dynamics and movements, like restaurants and airports, are notably absent from these datasets. For these environments, particularly kitchens, the tasks often center around high-end appliances like air fryers and waffle makers found primarily in affluent households [7]. Moreover, these datasets largely represent environments from high-income Western countries [7]. While common objects in these scenes—such as toys, mugs, and hats—are not explicitly Western-centric, they fail to capture the material culture of non-urban households in Global South regions, some of which already have high robot uptake, like India. This economic and geographic bias is reflected in the selection of everyday objects that presume a certain level of purchasing power. Prior

research in computer vision has documented similar Western-centric biases in dataset curation [46], [47], highlighting how this perpetuates data and algorithmic coloniality [48].

There is a focus on the West when it comes to dataset curation, yet different cultures and regions have different objects and layouts within homes and other environments. For robot home working scenes, including environments that reflect the living conditions of many regions, such as the inclusion of countries that are considered part of the Global South, should be pursued by dataset developers. Within the Global South, since not all emerging economies are the same, we believe that dataset developers need to factor in the diversity needs prevailing at the regional levels. To give just one example, this is especially important for countries like India where there is a stark difference between the social scenes in an urban domestic setting and a rural domestic setting, given that a large majority of the Indian population is concentrated in villages.

$R4$: Inclusion of human participants

Inclusion of human participants with various motion characteristics (e.g. hand dominance, history of amputation, limited motion, conditions that affect mobility) and from different demographics should be made a priority.

Our fourth recommendation **R4** focuses on the fact that humans who interact with robots are likely to have different characteristics and behavior preferences. These priorities should be accounted for in the curation of any robot dataset so that the robot can learn how to behave for different groups of individuals. Some policies trained on certain types of height, age and body types have the risk of performing worse on other groups that are less seen in the datasets [49]. Participants with various physical attributes and abilities can help robots learn from a wide variety of motions. A dataset that contains demonstrated paths with various features is specifically useful for robots that assist humans. When demonstrators operate the robot to record the trajectory, their characteristics are implicitly incorporated in the data collected. For example, in human-robot interaction tasks, the features of the users interacting with the robot play a significant role. Most learning-from-demonstration algorithms assume that there exists a single optimal policy, reward, or plan to be learned [50]. However, people have different bodies, and motion preferences, and there are therefore multiple solutions to manipulation and human-robot collaboration task.

To train policies that are inclusive for all, different groups of demonstrators should be considered in the data collection. As previously stated, when human participants are involved, dataset developers should also report their information and demographics, especially when they claim the datasets are generated with groups of 'diverse' users.

Motion characteristics vary depending on age, height, and sex [51], [52], so including demonstrators from different demographics should be taken into consideration. A range of physical characteristics such as hand dominance, history of amputation, limits in motion, or conditions that affect mobility should also be included. We recommend that the inclusion of people with various motion characteristics should be made a priority but understand the difficulty in this, specifically when trying to include significant amounts of older people, younger people, and people of various physical abilities in a dataset. Including data from underrepresented groups in datasets should be encouraged. To train robots that work with all groups, demonstrators with disabilities should be included in the data collection processed and reported in the datasets.

$R5$: Documentation of participant information

Information of human participation should be documented, including their demographics, how they were selected, and their training process [22].

As has been found throughout this research, there is often missing information about the curation of new robot datasets. Our fifth recommendation **R5** pertains to the demographics and information pertaining to human participants not being specified. For example, grasping datasets [39], [40], [41], which mainly focus one diversity as a range of objects of different size and shape, generally do not provide demographic information or details of the participants in the dataset. This could lead to downstream problems, because the characteristics of human participants used for training grasping policies (e.g., hand dominance and disabilies) will have an influence on the generated motions of the robot. For example, the demographics of human participants involved throughout the curation of the datasets is often missing. The video dataset Folding Demonstration [42] contains 8.5 hours of human demonstrations of clothes folding. The abstract claims that the demonstrations are recorded with a diverse set of people, but the information on these demonstrators is missing. Similarly, RT20T [31] reports that the demonstrations are performed by 19 humans to ensure diverse trajectories, but the information on demographics and motion characteristics of the demonstrators is not provided.

V. CONCLUSION

In this paper we conducted a systematic literature analysis related to robot manipulation and human-robot collaboration datasets, and identified five dimensions of dataset 'diversity'. We found that, when datasets are claimed to be 'diverse', such diversity most often refers to the variety of objects seen in the dataset, but can also refer to diversity in robot tasks, environments, platforms and human participants. Often the definition of diversity is ambiguous and implicit and it thus is difficult to measure and compare to other datasets. We proposed five recommendations to guide future research in creating diverse robot datasets, related to definitions, metrics, inclusion of cultural diversity, inclusion of human participants with varied physical and demographic characteristics, and better documentation practices.

REFERENCES

[1] 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," 2018. [Online]. Available: https://arxiv.org/abs/1812.00568

[2] A. Singh, A. Yu, J. Yang, J. Zhang, A. Kumar, and S. Levine, "Cog: Connecting new skills to past experience with offline reinforcement learning," 2020. [Online]. Available: https://arxiv.org/abs/2010.14500

[3] A. Mandlekar, D. Xu, J. Wong, S. Nasiriany, C. Wang, R. Kulkarni, L. Fei-Fei, S. Savarese, Y. Zhu, and R. Martín-Martín, "What matters in learning from offline human demonstrations for robot manipulation," 2021. [Online]. Available: https://arxiv.org/abs/2108.03298

[4] 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," 2021. [Online]. Available: https://arxiv.org/abs/2104.08212

[5] S. Cabi, S. G. Colmenarejo, A. Novikov, K. Konyushkova, S. Reed, R. Jeong, K. Zolna, Y. Aytar, D. Budden, M. Vecerik, O. Sushkov, D. Barker, J. Scholz, M. Denil, N. de Freitas, and Z. Wang, "Scaling data-driven robotics with reward sketching and batch reinforcement learning," 2020. [Online]. Available: https://arxiv.org/abs/1909.12200

[6] 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," 2024. [Online]. Available: https://arxiv.org/abs/2308.12952

[7] A. Khazatsky, K. Pertsch, S. Nair, A. Balakrishna, S. Dasari, S. Karamcheti, S. Nasiriany, M. K. Srirama, L. Y. Chen, K. Ellis, P. D. Fagan, J. Hejna, M. Itkina, M. Lepert, Y. J. Ma, P. T. Miller, J. Wu, S. Belkhale, S. Dass, H. Ha, A. Jain, A. Lee, Y. Lee, M. Memmel, S. Park, I. Radosavovic, K. Wang, A. Zhan, K. Black, C. Chi, K. B. Hatch, S. Lin, J. Lu, J. Mercat, A. Rehman, P. R. Sanketi, A. Sharma, C. Simpson, Q. Vuong, H. R. Walke, B. Wulfe, T. Xiao, J. H. Yang, A. Yavary, T. Z. Zhao, C. Agia, R. Baijal, M. G. Castro, D. Chen, Q. Chen, T. Chung, J. Drake, E. P. Foster, J. Gao, D. A. Herrera, M. Heo, K. Hsu, J. Hu, D. Jackson, C. Le, Y. Li, K. Lin, R. Lin, Z. Ma, A. Maddukuri, S. Mirchandani, D. Morton, T. Nguyen, A. O'Neill, R. Scalise, D. Seale, V. Son, S. Tian, E. Tran, A. E. Wang, Y. Wu, A. Xie, J. Yang, P. Yin, Y. Zhang, O. Bastani, G. Berseth, J. Bohg, K. Goldberg, A. Gupta, A. Gupta, D. Jayaraman, J. J. Lim, J. Malik, R. Martín-Martín, S. Ramamoorthy, D. Sadigh, S. Song, J. Wu, M. C. Yip, Y. Zhu, T. Kollar, S. Levine, and C. Finn, "Droid: A large-scale in-the-wild robot manipulation dataset," 2024. [Online]. Available: \url{https://arxiv.org/abs/2403.12945}

[8] H.-S. Fang, H. Fang, Z. Tang, J. Liu, C. Wang, J. Wang, H. Zhu, and C. Lu, "Rh20t: A comprehensive robotic dataset for learning diverse skills in one-shot," 2023. [Online]. Available: https://arxiv.org/abs/2307.00595

[9] K. Mavrogiorgos, A. Kiourtis, A. Mavrogiorgou, A. Menychtas, and D. Kyriazis, "Bias in machine learning: A literature review," *Applied Sciences*, vol. 14, no. 19, p. 8860, 2024.

[10] C. Starke, J. Baleis, B. Keller, and F. Marcinkowski, "Fairness perceptions of algorithmic decision-making: A systematic review of the empirical literature," *Big Data & Society*, vol. 9, no. 2, p. 20539517221115189, 2022.

[11] J. Larson, S. Mattu, L. Kirchner, and J. Angwin, "How We Analyzed the COMPAS Recidivism Algorithm," Available at: https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm, 2016.

[12] J. Dastin, "Amazon scraps secret ai recruiting tool that showed bias against women," in *Ethics of Data and Analytics*. Auerbach Publications, 2022.

[13] D. Zhao, J. T. Andrews, O. Papakyriakopoulos, and A. Xiang, "Position: measure dataset diversity, don't just claim it," *arXiv preprint arXiv:2407.08188*, 2024.

[14] Q. Thames, A. Karpur, W. Norris, F. Xia, L. Panait, T. Weyand, and J. Sim, "Nutrition5k: Towards automatic nutritional understanding of generic food," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 8903–8911.

[15] B. Y. Lin, S. Lee, X. Qiao, and X. Ren, "Common sense beyond english: Evaluating and improving multilingual language models for commonsense reasoning," *arXiv preprint arXiv:2106.06937*, 2021.

[16] A. R. Fabbri, F. Rahman, I. Rizvi, B. Wang, H. Li, Y. Mehdad, and D. Radev, "Convosumm: Conversation summarization benchmark and improved abstractive summarization with argument mining," *arXiv preprint arXiv:2106.00829*, 2021.

[17] J. Lei, L. Yu, T. L. Berg, and M. Bansal, "Tvr: A large-scale dataset for video-subtitle moment retrieval," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*. Springer, 2020, pp. 447–463.

[18] Z. Shen, K. Lo, L. L. Wang, B. Kuehl, D. S. Weld, and D. Downey, "Vila: Improving structured content extraction from scientific pdfs using visual layout groups," *Transactions of the Association for Computational Linguistics*, vol. 10, pp. 376–392, 2022.

[19] J. An, J. Kim, H. Lee, J. Kim, J. Kang, S. Shin, M. Kim, D. Hong, and S. S. Woo, "Vfp290k: A large-scale benchmark dataset for vision-based fallen person detection," in *Thirty-fifth conference on neural information processing systems datasets and benchmarks track (round 2)*, 2021.

[20] A. Yuan, D. Ippolito, V. Nikolaev, C. Callison-Burch, A. Coenen, and S. Gehrmann, "Synthbio: A case study in faster curation of text datasets," in *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

[21] P. Zeinert, N. Inie, and L. Derczynski, "Annotating online misogyny," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021, pp. 3181–3197.

[22] D. Zhao, J. T. A. Andrews, O. Papakyriakopoulos, and A. Xiang, "Position: measure dataset diversity, don't just claim it," in *Proceedings of the 41st International Conference on Machine Learning*, ser. ICML'24. JMLR.org, 2024.

[23] S. Caton and C. Haas, "Fairness in machine learning: A survey," *ACM Computing Surveys*, vol. 56, no. 7, pp. 1–38, 2024.

[24] M. Hort, Z. Chen, J. M. Zhang, M. Harman, and F. Sarro, "Bias mitigation for machine learning classifiers: A comprehensive survey," *ACM J. Responsib. Comput.*, vol. 1, no. 2, Jun. 2024. [Online]. Available: https://doi.org/10.1145/3631326

[25] H. Suresh and J. Guttag, "A framework for understanding sources of harm throughout the machine learning life cycle," in *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, 2021, pp. 1–9.

[26] S. Linxen, C. Sturm, F. Brühlmann, V. Cassau, K. Opwis, and K. Reinecke, "How weird is chi?" in *Proceedings of the 2021 chi conference on human factors in computing systems*, 2021, pp. 1–14.

[27] K. Seaborn, G. Barbareschi, and S. Chandra, "Not only WEIRD but "uncanny"? A systematic review of diversity in human-robot interaction research," *Int. J. Soc. Robotics*, vol. 15, no. 11, pp. 1841–1870, 2023. [Online]. Available: https://doi.org/10.1007/s12369-023-00968-4

[28] K. Winkle, D. McMillan, M. Arnelid, K. Harrison, M. Balaam, E. Johnson, and I. Leite, "Feminist human-robot interaction: Disentangling power, principles and practice for better, more ethical hri," in *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, 2023, pp. 72–82.

[29] K. Winkle, E. Lagerstedt, I. Torre, and A. Offenwanger, "15 years of (who)man robot interaction: Reviewing the H in human-robot interaction," *ACM Trans. Hum. Robot Interact.*, vol. 12, no. 3, pp. 28:1–28:28, 2023. [Online]. Available: https://doi.org/10.1145/3571718

[30] C. Mitash, F. Wang, S. Lu, V. Terhuja, T. Garaas, F. Polido, and M. Nambi, "Armbench: An object-centric benchmark dataset for robotic manipulation," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 9132–9139.

[31] H.-S. Fang, H. Fang, Z. Tang, J. Liu, C. Wang, J. Wang, H. Zhu, and C. Lu, "Rh20t: A comprehensive robotic dataset for learning diverse skills in one-shot," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 653–660.

[32] H. Kim, Y. Ohmura, and Y. Kuniyoshi, "Multi-task real-robot data with gaze attention for dual-arm fine manipulation," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2024, pp. 8516–8523.

[33] H. Fang, H.-S. Fang, S. Xu, and C. Lu, "Transcg: A large-scale real-world dataset for transparent object depth completion and a grasping baseline," *IEEE Robotics and Automation Letters*, vol. 7, no. 3, pp. 7383–7390, 2022.

[34] 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*, vol. 7, no. 3, pp. 7327–7334, 2022.

[35] A. O'Neill, A. Rehman, A. Maddukuri, A. Gupta, A. Padalkar, A. Lee, A. Pooley, A. Gupta, A. Mandlekar, A. Jain, A. Tung, A. Bewley, A. Herzog, A. Irpan, A. Khazatsky, A. Rai, A. Gupta, A. Wang, A. Singh, A. Garg, A. Kembhavi, A. Xie, A. Brohan, A. Raffin, A. Sharma, A. Yavary, A. Jain, A. Balakrishna, A. Wahid, B. Burgess-Limerick, B. Kim, B. Schölkopf, B. Wulfe, B. Ichter, C. Lu, C. Xu, C. Le, C. Finn, C. Wang, C. Xu, C. Chi, C. Huang, C. Chan, C. Agia, C. Pan, C. Fu, C. Devin, D. Xu, D. Morton, D. Driess, D. Chen, D. Pathak, D. Shah, D. Büchler, D. Jayaraman, D. Kalashnikov, D. Sadigh, E. Johns, E. Foster, F. Liu, F. Ceola, F. Xia, F. Zhao, F. Stulp, G. Zhou, G. S. Sukhatme, G. Salhotra, G. Yan, G. Feng, G. Schiavi,

G. Berseth, G. Kahn, G. Wang, H. Su, H.-S. Fang, H. Shi, H. Bao, H. Ben Amor, H. I. Christensen, H. Furuta, H. Walke, H. Fang, H. Ha, I. Mordatch, I. Radosavovic, I. Leal, J. Liang, J. Abou-Chakra, J. Kim, J. Drake, J. Peters, J. Schneider, J. Hsu, J. Bohg, J. Bingham, J. Wu, J. Gao, J. Hu, J. Wu, J. Wu, J. Sun, J. Luo, J. Gu, J. Tan, J. Oh, J. Wu, J. Lu, J. Yang, J. Malik, J. Silvério, J. Hejna, J. Booher, J. Tompson, J. Yang, J. Salvador, J. J. Lim, J. Han, K. Wang, K. Rao, K. Pertsch, K. Hausman, K. Go, K. Gopalakrishnan, K. Goldberg, K. Byrne, K. Oslund, K. Kawaharazuka, K. Black, K. Lin, K. Zhang, K. Ehsani, K. Lekkala, K. Ellis, K. Rana, K. Srinivasan, K. Fang, K. P. Singh, K.-H. Zeng, K. Hatch, K. Hsu, L. Itti, L. Y. Chen, L. Pinto, L. Fei-Fei, L. Tan, L. J. Fan, L. Ott, L. Lee, L. Weihs, M. Chen, M. Lepert, M. Memmel, M. Tomizuka, M. Itkina, M. G. Castro, M. Spero, M. Du, M. Ahn, M. C. Yip, M. Zhang, M. Ding, M. Heo, M. K. Srirama, M. Sharma, M. J. Kim, N. Kanazawa, N. Hansen, N. Heess, N. J. Joshi, N. Suenderhauf, N. Liu, N. Di Palo, N. M. M. Shafiullah, O. Mees, O. Kroemer, O. Bastani, P. R. Sanketi, P. T. Miller, P. Yin, P. Wohlhart, P. Xu, P. D. Fagan, P. Mitrano, P. Sermanet, P. Abbeel, P. Sundaresan, Q. Chen, Q. Vuong, R. Rafailov, R. Tian, R. Doshi, R. Martín-Martín, R. Baijal, R. Scalise, R. Hendrix, R. Lin, R. Qian, R. Zhang, R. Mendonca, R. Shah, R. Hoque, R. Julian, S. Bustamante, S. Kirmani, S. Levine, S. Lin, S. Moore, S. Bahl, S. Dass, S. Sonawani, S. Song, S. Xu, S. Haldar, S. Karamcheti, S. Adebola, S. Guist, S. Nasiriany, S. Schaal, S. Welker, S. Tian, S. Ramamoorthy, S. Dasari, S. Belkhale, S. Park, S. Nair, S. Mirchandani, T. Osa, T. Gupta, T. Harada, T. Matsushima, T. Xiao, T. Kollar, T. Yu, T. Ding, T. Davchev, T. Z. Zhao, T. Armstrong, T. Darrell, T. Chung, V. Jain, V. Vanhoucke, W. Zhan, W. Zhou, W. Burgard, X. Chen, X. Wang, X. Zhu, X. Geng, X. Liu, X. Liangwei, X. Li, Y. Lu, Y. J. Ma, Y. Kim, Y. Chebotar, Y. Zhou, Y. Zhu, Y. Wu, Y. Xu, Y. Wang, Y. Bisk, Y. Cho, Y. Lee, Y. Cui, Y. Cao, Y.-H. Wu, Y. Tang, Y. Zhu, Y. Zhang, Y. Jiang, Y. Li, Y. Li, Y. Iwasawa, Y. Matsuo, Z. Ma, Z. Xu, Z. J. Cui, Z. Zhang, and Z. Lin, "Open x-embodiment: Robotic learning datasets and rt-x models : Open x-embodiment collaboration0," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 6892–6903.

[36] A. Baselizadeh, M. Z. Uddin, W. Khaksar, D. S. Lindblom, and J. Torresen, "Prima-care: Privacy-preserving multi-modal dataset for human activity recognition in care robots," in *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, ser. HRI '24. New York, NY, USA: Association for Computing Machinery, 2024, p. 233–237. [Online]. Available: https://doi.org/10.1145/3610978.3640701

[37] L. Chen, J. Qiu, L. Li, X. Luo, G. Chi, and Y. Zheng, "Advancing robots with greater dynamic dexterity: A large-scale multi-view and multi-modal dataset of human-human throw&catch of arbitrary objects," *International Journal of Robotics Research*, 2024, cited by: 0. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85205319843&doi=10.1177%2f02783649241275674&partnerID=40&md5=4d21f343546320b75c8508f38ee917cd

[38] J. Kim, M.-H. Jeon, S. Jung, W. Yang, M. Jung, J. Shin, and A. Kim, "Transpose: Large-scale multispectral dataset for transparent object," 2023. [Online]. Available: https://arxiv.org/abs/2307.05016

[39] A. D. Vuong, M. N. Vu, H. Le, B. Huang, B. Huynh, T. Vo, A. Kugi, and A. Nguyen, "Grasp-anything: Large-scale grasp dataset from foundation models," 2023. [Online]. Available: https://arxiv.org/abs/2309.09818

[40] 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)*. IEEE Press, 2022, p. 2553–2560. [Online]. Available: https://doi.org/10.1109/ICRA46639.2022.9811809

[41] C. Eppner, A. Mousavian, and D. Fox, "Acronym: A large-scale grasp dataset based on simulation," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 2021, pp. 6222–6227.

[42] A. Verleysen, M. Biondina, and F. wyffels, "Video dataset of human demonstrations of folding clothing for robotic folding," *The International Journal of Robotics Research*, vol. 39, no. 9, pp. 1031–1036, 2020. [Online]. Available: https://doi.org/10.1177/0278364920940408

[43] P. Sharma, L. Mohan, L. Pinto, and A. Gupta, "Multiple interactions made easy (mime): Large scale demonstrations data for imitation," 2018. [Online]. Available: https://arxiv.org/abs/1810.07121

[44] 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," 2022. [Online]. Available: https://arxiv.org/abs/2202.02005

[45] 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," 2018. [Online]. Available: https://arxiv.org/abs/1811.02790

[46] S. Shankar, Y. Halpern, E. Breck, J. Atwood, J. Wilson, and D. Sculley, "No classification without representation: Assessing geodiversity issues in open data sets for the developing world," 2017. [Online]. Available: https://arxiv.org/abs/1711.08536

[47] O. Agarwal, Y. Yang, B. C. Wallace, and A. Nenkova, "Entity-switched datasets: An approach to auditing the in-domain robustness of named entity recognition models," 2021. [Online]. Available: https://arxiv.org/abs/2004.04123

[48] N. Sambasivan, E. Arnesen, B. Hutchinson, T. Doshi, and V. Prabhakaran, "Re-imagining algorithmic fairness in india and beyond," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, ser. FAccT '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 315–328. [Online]. Available: https://doi.org/10.1145/3442188.3445896

[49] Z. Evans, M. Leonetti, and M. Brandão, *Bias and Performance Disparities in Reinforcement Learning for Human-Robot Interaction*, 2025.

[50] H. chaandar Ravichandar, A. S. Polydoros, S. Chernova, and A. Billard, "Recent advances in robot learning from demonstration," *Annu. Rev. Control. Robotics Auton. Syst.*, vol. 3, pp. 297–330, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:208958394

[51] W. L. Au, I. S. H. Seah, W. Li, and L. C. S. Tan, "Effects of age and gender on hand motion tasks," *Parkinson's Disease*, vol. 2015, no. 1, p. 862427, 2015.

[52] K. Sankar and J. C. Michael Christudhas, "Influence of aging, disease, exercise, and injury on human hand movements: A systematic review," *Proceedings of the Institution of Mechanical Engineers, Part H: Journal of Engineering in Medicine*, vol. 235, no. 11, pp. 1221–1256, 2021.