RoboTube: Learning Household Manipulation from Human Videos with Simulated Twin Environments

Haoyu Xiong* [1,2], Haoyuan Fu* [3], Jieyi Zhang [3], Chen Bao [3], Qiang Zhang [4], Yongxi Huang [3], Wenqiang Xu [1,3], Animesh Garg [5,6,,7], Huazhe Xu[1,4], Cewu Lu [1,3]

[1] Shanghai Qizhi Institute, [2] Carnegie Mellon University, [3] Shanghai JiaoTong University, [4] Tsinghua University
[5] University of Toronto, [6] Vector Institute, [7] Nvidia

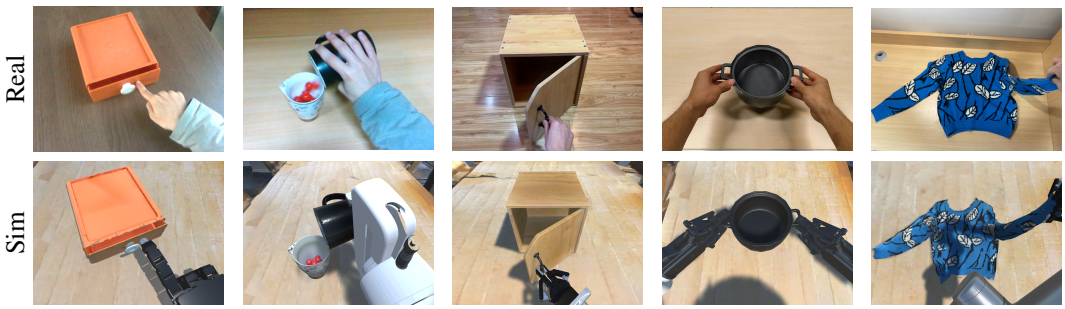

Figure 1: **RoboTube** covers a wide range of household manipulation tasks. RoboTube constructs a human video dataset and a suite of simulated twin environments for reproducible research. The first row shows the examples of the real-world video frames; the second row shows the simulated twin environments.

Abstract: We aim to build a useful, reproducible, democratized benchmark for learning household robotic manipulation from human videos. To realize this goal, a diverse, high-quality human video dataset curated specifically for robots is desired. To evaluate the learning progress, a simulated twin environment that resembles the appearance and the dynamics of the physical world would help roboticists and AI researchers validate their algorithms convincingly and efficiently before testing on a real robot. Hence, we present RoboTube, a human video dataset, and its digital twins for learning various robotic manipulation tasks. RoboTube video dataset contains 5,000 video demonstrations recorded with multi-view RGB-D cameras of human-performing everyday household tasks including manipulation of rigid objects, articulated objects, granular objects, deformable objects, and bimanual manipulation. RT-sim, as the simulated twin environments, consists of 3D scanned, photo-realistic objects, minimizing the visual domain gap between the physical world and the simulated environment. We hope RoboTube can lower the barrier to robotics research for beginners while facilitating reproducible research in the community.

Keywords: Learning from Videos, Video Demonstration Dataset, Real2Sim, Self-supervised Reward Learning, Robotic Simulation Benchmark

1 Introduction

Robot learning from human videos unlocks the potential to enable everyday household manipulation tasks [1–4]. Prior works have made fruitful progress on manipulation tasks such as pick-and-place by learning from offline video datasets [5–9]. As these video datasets facilitate the pioneer exploration of robotic manipulation learning, they have several deficiencies for further exploration:

(1) **Task complexity**. Many of the algorithms and frameworks [5–7] focus on the easy end of the task spectrum, e.g., pick-and-place, push, relocating rigid objects, etc. While a practical robotic manipulation system should be able to handle complex tasks that involve articulated objects, deformable objects, granular objects, or bimanual coordination. Empowering a robotic manipulation system with

Dataset Name	Depth Mapping	Multiple Viewpoints	Simulated Twin	Negative Sample	End-Effector	Approx Annotation	Number of Tasks	Dataset Size
XIRL(real) [14]	×	×	×	×	Diverse	N.A	1X	100 videos
G-in-W [5]	✓	×	×	×	DemoAT	gripper open/close	1X	12 hours
TCN-pour [4]	✓	×	×	✓	Human hands	N.A	1X	300 videos
RLV [3]	×	×	×	×	Human hands	N.A	2X	300 videos
DexMV [7]	✓	✓	✓	×	Human hands	6 object models	3X	700 videos
VIME [6]	×	✓	×	×	DemoAT	gripper transition	2X	2000 videos
RoboTube	✓	✓	✓	✓	Human hands	**60 object models**	**5X**	**5000 videos**

Table 1: **Comparison of video demonstration datasets.** We compare the features of RoboTube video dataset with related video demonstration datasets. In this table, DemoAT means demonstration assistive tools. In the number of tasks section, n X meams n groups of tasks.

human videos benefits many real-world applications and largely extends the research scopes. (2) **Data diversity & relevance**. Learning a large range of diverse manipulation behaviors from Visual demonstrations that are collected on a static lab table with limited object instances [7, 10–12] is difficult, if not impossible. In contrast, the massive-scale open-world video datatsets [8, 9] contributes to generalization in robotic manipulation [1, 2, 13]. However, as they are not originally designed for robotics, they introduce unnecessary challenges with irrelevant content. For example, in Ego4D dataset [9], the video frames may have content beyond human manipulation including a crowd in a concert live, human walking, etc. (3) **Baseline comparison**. A standard benchmark that functions on comparing different proposed methods still remains a missing part in the community. As the exact copies of the objects in the videos are hard to be obtained, the roboticists may set up different experimental settings with different objects to validate the learned models. For example, [1, 13] both learned reward functions and induced policies from the same something-something dataset [8] but applied the learned models to different robotic experiments, due to the lack of a standard benchmark, which makes the meaningful, reproducible, democratized comparison among different baseline methods extremely hard.

To address the deficiencies mentioned above, we introduce **RoboTube** (Fig. 2), a human video dataset of around **5,000** RGB-D video clips.

1. To ensure the task complexity, RoboTube setups environments for 5 task families, namely drawer-closing (*articulated object with prismatic joint*), mug-pouring (*granular object*), cabinet-opening (*articulated object with revolute joint*), bimanual-pot-lifting (*bimanual coordination*), and cloth-folding (*deformable object*).
2. To take the data diversity into consideration, for each task family, we ask 9 demonstrators to conduct the task with diverse but *natural* hand poses upon different objects of the same category which have variations in shapes, materials, and textures. We collect the videos in both clean and cluttered scenes. To support the reproduction and comparisons of different algorithms and enable wider applicability, the RoboTube video dataset contains multiple functionalities. We collect both *successful* (expert video demonstrations) and *failed* (negative video demonstrations) episodes, concerning 50 tasks and 60 objects. Two temporally synchronized video streams are recorded from a first-person viewpoint (FPV) and a third-person viewpoint (TPV).
3. To benchmark the baseline methods, we construct a simulated twin environment, RT-sim, for the tasks and objects. With RT-sim, researchers can make a fair comparison of their approaches with the baseline methods and can validate their algorithms convincingly and efficiently before conducting more complex experiments on real robots.

We summarize our contributions as follows: We identify the issues in existing human videos for robot learning, and curate a benchmark, RoboTube, which is designed by jointly considering the human video dataset and the evaluation platform. RoboTube not only introduces more complex tasks with diverse object types, but also supports meaningful, reproducible, democratized comparisons among different baseline methods.

2 Related Work

2.1 Offline Datasets for Robotic Manipulation

Leveraging offline datasets to learn diverse manipulation behaviors has been studied by previous researchers.

Video datasets for perception tasks. The computer vision community has curated many human-object-interaction (HOI) video datasets for different perception tasks [8, 9, 15]. [1, 13] have proved that it is effective to learn a generalizable reward function from the something-something dataset [8]. R3M [2] also exploited Ego4D [9] to improve efficiency in downstream motor control tasks. Despite the rich prior knowledge that HOI videos have provided, such datasets are not originally designed for robotics. For example, Ego4D dataset [9] contains content beyond human manipulation including crowd in a concert live, human walking, etc.

Action-included demonstrations for robotic manipulation. An action-included demonstration usually contains both the visual observation and the corresponding actions of the robots, which provides strong supervision for a robot to learn complex behaviors. Previous works collect demonstrations on a static lab table [10–12]. Recently, several works [16, 17] take an effort to enrich the data diversity and show better generalization ability in imitation learning of everyday household tasks. Despite the tremendous progress in learning from action-included demonstrations has been made, such datasets suffer a key problem: it is time-consuming and expensive to collect everyday household activities by guiding and/or teleoperating a real robot entity. In contrast, one can record videos anywhere and anytime with a portable camera.

Video-only demonstrations for robotic manipulation. Consider the issues of other two kinds of datasets, roboticists have also constructed video-only datasets for robotic manipulation [3–7, 14]. These datasets can be divided into two mainstreams: *robot-friendly* video demonstrations [5, 6], *human-friendly* video demonstrations (human videos) [3, 4, 7, 14]. Song et al. [5] propose a robot-friendly interface for collecting video demonstrations anywhere using assistive tools (DemoAT). Besides DemoAT, researchers also propose to collect videos with human hands for robotic manipulation. DexMV [7] conducts a novel pipeline to bridge 3D vision and dexterous manipulation. A more detailed comparison of RoboTube's features to those of related datasets can be found in Table **??**.

2.2 Algorithms for Robot Learning from Videos

Endowing robots with the ability to learn skills by simply observing humans has been an emblematic north star problem in robotics [14, 18–24]. Several directions have been proposed to achieve this goal:

Reward learning from videos. Recent works demonstrate impressive manipulation skills learned from human videos by inverse reinforcement learning [4, 13, 14, 20, 21]. For example, previous works [13, 20] train a goal classifier as a reward function on human videos for policy learning. Later, Xie et al. propose DVD [1], a domain-agnostic video discriminator for generalizable reward learning. More recently, XIRL [14] leverages temporal cycle-consistency constraints [25] to learn deep visual embeddings that are aware of task progress.

Visual pre-training for motor control. Recent works also discussed how to connect computer vision to policy learning by leveraging self-supervised pre-training. A line of works [2, 26, 27] has shown that pre-trained vision models from diverse real-world data can be effective to improve policy learning. For example, Nair et al. [2] prove that vision-language pre-training on diverse egocentric datasets, e.g., Ego4D [9].

3 RoboTube Video Dataset

We construct the RoboTube video dataset, a collection of human video demonstrations for robots to learn from. RoboTube contains multiple features, equipping it with the capability as a benchmark for existing algorithms with different settings. We give an overview of our RoboTube dataset in Fig. 3.

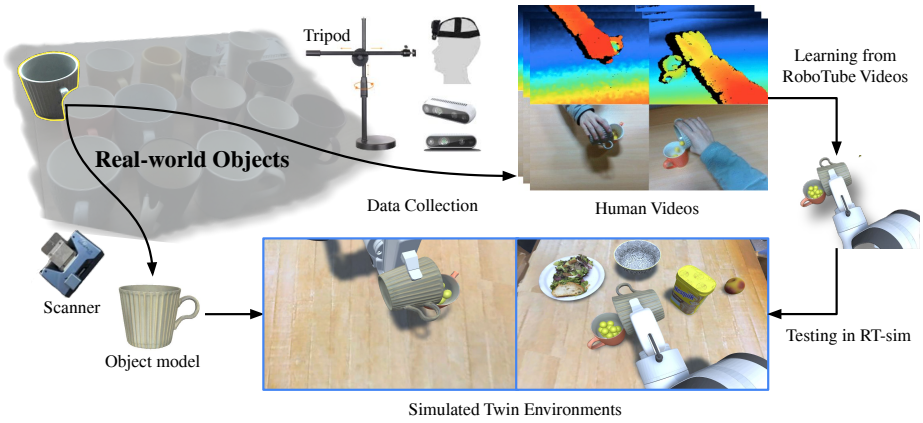

Figure 2: **Overview of RoboTube.** When building the video dataset, we ask demonstrators to collect manipulation video demonstrations recorded by multi-view RGB-D cameras. Meanwhile, we scan the corresponding objects into high-quality 3D models and construct a paired simulated scene. After learning from the video dataset, we test the learned models in the simulated scene.

3.1 Task Definition

We define each task in RoboTube with its task family, task mode, and the corresponding object. With the idea of task complexity in mind, we define 5 task families, with which we hope to go beyond pick-and-place and cover common manipulation tasks for household objects with different levels of complexity. Specifically, the task families deal with *articulated object manipulation* (drawer-closing, cabinet-opening), *granular object handling* (mug-pouring), *deformable object manipulation* (cloth-folding), and *bimanual coordination* (pot-lifting). To ensure task diversity, we set up two task modes with different levels of difficulty for each task family, as shown in Fig. 3(a). We design 1) the *structured mode*, where we place only the object on a clean table as the *easy level*, and 2) the *cluttered mode*, the *hard* level, where we place the objects in diverse real-world scenes without intentional clean-up, i.e., distractors exist along with the objects in the scenes. Based on the above definitions, a distinct task is denoted by its task family, task mode, and the corresponding objects. For example, a *drawer-closing-structured-v1* task means that drawer #1 is placed into a clean scene, and the task is to close the drawer.

3.2 Construction of Video Dataset

Object Selection As shown in Fig. 4, each task family contains multiple object instances of the same category with variations in colors, shapes, and textures, but consistency in semantics and affordances. There are 10 drawers, 20 mugs, 10 cabinets, 10 pots, and 10 cloths, in total, 60 objects.

Recording Setup During recording, two viewpoints are streamed: one is the first-person perspective from the camera mounted on the human head, and the other is the third-person perspective from the camera fixed on a tripod placed near the scene. These two streams are temporally synchronized. To record the video, we use RealSense D435 with a resolution of 640×480 and a frequency of 30Hz. More details about the hardware setups can be found in the supplementary materials.

Statistics RoboTube video dataset contains around 5,000 RGB-D visual demonstrations. For details of the train-test split and other analyses of the dataset, please refer to supplementary materials.

4 RT-sim: RoboTube Simulated Twin Environments

To provide an accessible test platform for reproducible research of robot learning from videos, we design RT-sim, a suite of simulation environments *paired* with RoboTube video dataset, in which we provide a configured scene for each demonstration.

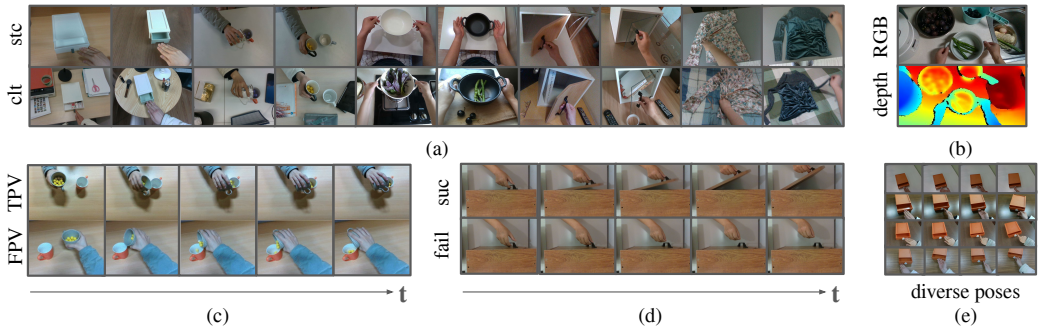

Figure 3: **RoboTube videos dataset.** (a): RoboTube designs the structured (stc) mode and the cluttered (clt) mode for two levels of task difficulty. The first row shows the structured scenes of drawer-closing, mug-pouring, pot-lifting, cabinet-opening, and cloth-folding tasks. The second row shows the cluttered scenes of the five manipulation tasks. (b): each frame in the RoboTube video dataset contains an RGB stream and a depth stream. (c): a first-person viewpoint (FPV) camera and a third-person viewpoint (TPV) camera are temporally synchronized. (d): RoboTube video dataset provides both successful episodes and failed episodes for the same task. (e) given the example of the drawer closing task, human demonstrators are required to make diverse poses to complete the tasks.

4.1 Environment Setup

Visual Rendering and Physics Simulation To mitigate the gap between the simulated and real-world, visual rendering and physics simulation play important roles. To prepare the objects, we scan high-fidelity object mesh models from real-world objects for manipulation and use google object scans [28] with realistic textures as the actionable distractors for cluttered mode. To create visual-realistic everyday household scenes, We import the scenes from iGibson projects [29–31]. Following the object scanning and annotation procedures in [32], we annotate the physics properties of the objects to align with the real world. Leveraging photo-realistic rendering and physics backends of Unity3D, we are able to construct visual and physics realistic simulation environments which have the potential to align with the real world and therefore serve as a benchmark for researchers to validate their algorithms before deploying them to real robots.

Robot Assets RT-sim supports various robots (e.g. Franka, UR5, Kinova-gen3) and grippers (e.g. Allegro Hand, Robotiq 85) for manipulation tasks.

Interface for Learning To enable robot learning algorithm training, RT-sim provides a standard OpenAI Gym [33] API in Python language. The API can retrieve the scene states from the Unity side through gRPC communication.

4.2 Task Specification

We provide a standard specification of the tasks in RT-sim as the following.

Drawer Closing: A robot moves its end-effector with a fixed gripper orientation and must close the drawer. We fix the initial robot end-effector position and uniformly randomize drawer base position within a range of $10cm \times 10cm$ plane, drawer base rotation within $[-\frac{\pi}{12}, \frac{\pi}{12}]$ and initial drawer opening length within $[10cm, 15cm]$. The robot is rewarded for making the drawer handle and drawer base closer. The task is done when the drawer open distance is smaller than $2cm$. The goal of this task is whether the robot has finished the drawer closing task (0 or 1).

Mug Pouring: A robot holding a mug that is fixed on the robot gripper moves its end-effector and rotates around the X-axis of the robot base to pour the 20 tiny balls inside the mug into another fixed mug on the table. We lock the movement of the Z-axis of the robot base to lower the difficulty of this task. The robot is rewarded for transferring the tiny balls from the initial mug to the mug on the table through pouring. The task is done when more than 14 tiny balls are inside the mug on the table. The goal of this task is the number of balls poured into the fixed mug on the table.

Bimanual Pot-Lifting: A dual-arm robot (e.g. Tobor [34]) uses grippers of both arms to lift a binaural soup pot on the table. We randomize the initial position of the pot to increase diversity. The robot is rewarded for grasping both handles of the pot and lifting it as well as promising the angle

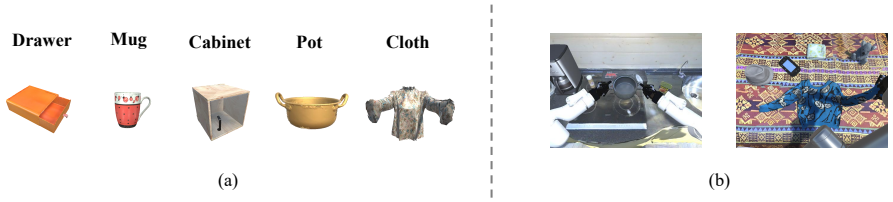

Figure 4: **Models and cluttered RT-sim gallery.** We render slected object models and the cluttered scenes of RT-sim from the first viewpoint. Realistic household tasks are visualized in the gallery.

of inclination of the pot and table shall not exceed $\frac{\pi}{6}$. The task is done when the robot lifts the pot higher than $5cm$ under the constraints mentioned above. The goal is the elevation of the pot under the constraint of inclination.

Cabinet Opening: A robot moves its end-effector and operates a gripper to grasp the door handle of a fixed cabinet and open the door. We randomize the initial gripper position within the range of $5cm$ of the door handle through all 3 axes. The robot is rewarded for grasping the door handle and pulling it to make the door open. The task is done when the opened degree is more than $\frac{2}{9}\pi$ and the door is opened by the robot pulling the handle. Other methods to open the door are seen as invalid. The goal is the opened degree of the cabinet door.

Cloth Folding: A robot moves its end-effector and operates a gripper to pick the graspable point on the left sleeve of a piece of clothing and place it in a target position near the lower-right corner of the clothes. We randomize the initial configuration and position of the clothes within a range of $5cm \times 5cm$ plane. The robot is rewarded for picking the graspable point and placing the left sleeve near the target. The task is done when the distance between graspable point and target is less than $15cm$. The goal is the negative distance between the graspable point and the target position.

5 Limitations

Though we have already extended the robot learning from video tasks to a larger scope with complex task settings, diverse backgrounds, and object instances. And we also pay particular attention to asking the demonstrators to operate in a natural way. Our dataset still has a gap towards the ultimate "in-the-wild" setting where the videos from the internet can be much less structured or relevant.

6 Conclusion

We introduce RoboTube, a benchmark for robot learning from human videos. Our core contribution lies in the joint design of the RoboTube video dataset and RT-sim. The models learned from RoboTube videos can be tested, benchmarked, and reproduced in RT-sim.

Acknowledgments

Thank you Ruolin Ye, Ziyi Wu, Samarth Sinha, Lixin Yang, Dr. Lin Shao, Dr. Liu Liu for the helpful discussions. We especially thank Dr.Huazhe Xu for the help of paper writing and discussions.

The research project is supported by Shanghai Qizhi Institute and MVIG lab at Shanghai JiaoTong University. We also thank Vector Institute for the computing resources.

References

[1] A. S. Chen, S. Nair, and C. Finn. Learning generalizable robotic reward functions from "in-the-wild" human videos. In *Proceedings of Robotics: Science and Systems (RSS)*, 2021.

[2] S. Nair, A. Rajeswaran, V. Kumar, C. Finn, and A. Gupta. R3m: A universal visual representation for robot manipulation, 2022. URL https://arxiv.org/abs/2203.12601.

[3] K. Schmeckpeper, O. Rybkin, K. Daniilidis, S. Levine, and C. Finn. Reinforcement learning with videos: Combining offline observations with interaction. In J. Kober, F. Ramos, and C. Tomlin, editors, *Proceedings of the 2020 Conference on Robot Learning*, volume 155 of *Proceedings of Machine Learning Research*, pages 339–354. PMLR, 16–18 Nov 2021. URL https://proceedings.mlr.press/v155/schmeckpeper21a.html.

[4] P. Sermanet, C. Lynch, Y. Chebotar, J. Hsu, E. Jang, S. Schaal, S. Levine, and G. Brain. Time-contrastive networks: Self-supervised learning from video. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1134–1141, 2018. doi:10.1109/ICRA.2018.8462891.

[5] S. Song, A. Zeng, J. Lee, and T. Funkhouser. Grasping in the wild: Learning 6dof closed-loop grasping from low-cost demonstrations. *IEEE Robotics and Automation Letters*, 5(3):4978–4985, 2020. doi:10.1109/LRA.2020.3004787.

[6] S. Young, D. Gandhi, S. Tulsiani, A. Gupta, P. Abbeel, and L. Pinto. Visual imitation made easy. In J. Kober, F. Ramos, and C. Tomlin, editors, *Proceedings of the 2020 Conference on Robot Learning*, volume 155 of *Proceedings of Machine Learning Research*, pages 1992–2005. PMLR, 16–18 Nov 2021. URL https://proceedings.mlr.press/v155/young21a.html.

[7] Y. Qin, Y. Wu, S. Liu, H. Jiang, R. Yang, Y. Fu, and X. Wang. Dexmv: Imitation learning for dexterous manipulation from human videos. *CoRR*, abs/2108.05877, 2021. URL https://arxiv.org/abs/2108.05877.

[8] R. Goyal, S. Ebrahimi Kahou, V. Michalski, J. Materzynska, S. Westphal, H. Kim, V. Haenel, I. Fruend, P. Yianilos, M. Mueller-Freitag, F. Hoppe, C. Thurau, I. Bax, and R. Memisevic. The "something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[9] K. Grauman, A. Westbury, E. Byrne, Z. Chavis, A. Furnari, R. Girdhar, J. Hamburger, H. Jiang, M. Liu, X. Liu, M. Martin, T. Nagarajan, I. Radosavovic, S. K. Ramakrishnan, F. Ryan, J. Sharma, M. Wray, M. Xu, E. Z. Xu, C. Zhao, S. Bansal, D. Batra, V. Cartillier, S. Crane, T. Do, M. Doulaty, A. Erapalli, C. Feichtenhofer, A. Fragomeni, Q. Fu, C. Fuegen, A. Gebreselasie, C. Gonzalez, J. Hillis, X. Huang, Y. Huang, W. Jia, W. Khoo, J. Kolar, S. Kottur, A. Kumar, F. Landini, C. Li, Y. Li, Z. Li, K. Mangalam, R. Modhugu, J. Munro, T. Murrell, T. Nishiyasu, W. Price, P. R. Puentes, M. Ramazanova, L. Sari, K. Somasundaram, A. Southerland, Y. Sugano, R. Tao, M. Vo, Y. Wang, X. Wu, T. Yagi, Y. Zhu, P. Arbelaez, D. Crandall, D. Damen, G. M. Farinella, B. Ghanem, V. K. Ithapu, C. V. Jawahar, H. Joo, K. Kitani, H. Li, R. Newcombe, A. Oliva, H. S. Park, J. M. Rehg, Y. Sato, J. Shi, M. Z. Shou, A. Torralba, L. Torresani, M. Yan, and J. Malik. Ego4d: Around the world in 3, 000 hours of egocentric video. *CoRR*, abs/2110.07058, 2021. URL https://arxiv.org/abs/2110.07058.

[10] 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. In *arXiv preprint arXiv:2108.03298*, 2021.

[11] 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. In A. Billard, A. Dragan, J. Peters, and J. Morimoto, editors, *Proceedings of The 2nd Conference on Robot Learning*, volume 87 of *Proceedings of Machine Learning Research*, pages 879–893. PMLR, 29–31 Oct 2018. URL https://proceedings.mlr.press/v87/mandlekar18a.html.

[12] 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, 2019. doi:10.1109/IROS40897.2019.8968114.

[13] 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.

[14] K. Zakka, A. Zeng, P. Florence, J. Tompson, J. Bohg, and D. Dwibedi. Xirl: Cross-embodiment inverse reinforcement learning. In A. Faust, D. Hsu, and G. Neumann, editors, *Proceedings of the 5th Conference on Robot Learning*, volume 164 of *Proceedings of Machine Learning Research*, pages 537–546. PMLR, 08–11 Nov 2022. URL https://proceedings.mlr.press/v164/zakka22a.html.

[15] Y. Liu, Y. Liu, C. Jiang, Z. Fu, K. Lyu, W. Wan, H. Shen, B. Liang, H. Wang, and L. Yi. Hoi4d: A 4d egocentric dataset for category-level human-object interaction. *arXiv preprint arXiv:2203.01577*, 2022.

[16] 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 A. Faust, D. Hsu, and G. Neumann, editors, *Proceedings of the 5th Conference on Robot Learning*, volume 164 of *Proceedings of Machine Learning Research*, pages 991–1002. PMLR, 08–11 Nov 2022. URL https://proceedings.mlr.press/v164/jang22a.html.

[17] 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, 2021. URL https://arxiv.org/abs/2109.13396.

[18] H. Xu, Y. Gao, F. Yu, and T. Darrell. End-to-end learning of driving models from large-scale video datasets. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[19] X. B. Peng, E. Coumans, T. Zhang, T.-W. E. Lee, J. Tan, and S. Levine. Learning agile robotic locomotion skills by imitating animals. In *Robotics: Science and Systems*, 07 2020. doi:10.15607/RSS.2020.XVI.064.

[20] L. Smith, N. Dhawan, M. Zhang, P. Abbeel, and S. Levine. Avid: Learning multi-stage tasks via pixel-level translation of human videos. In *Proceedings of Robotics: Science and Systems (RSS)*, 2020.

[21] 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, 2021. doi:10.1109/IROS51168.2021.9636080.

[22] M. Sieb, Z. Xian, A. Huang, O. Kroemer, and K. Fragkiadaki. Graph-structured visual imitation. 2020.

[23] X. B. Peng, A. Kanazawa, J. Malik, P. Abbeel, and S. Levine. Sfv: Reinforcement learning of physical skills from videos. *ACM Trans. Gr.*, 2018.

[24] P. Sharma, D. Pathak, and A. Gupta. Third-person visual imitation learning via decoupled hierarchical controller. 2019.

[25] D. Dwibedi, Y. Aytar, J. Tompson, P. Sermanet, and A. Zisserman. Temporal cycle-consistency learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[26] T. Xiao, I. Radosavovic, T. Darrell, and J. Malik. Masked visual pre-training for motor control, 2022. URL https://arxiv.org/abs/2203.06173.

[27] S. Parisi, A. Rajeswaran, S. Purushwalkam, and A. Gupta. The unsurprising effectiveness of pre-trained vision models for control, 2022. URL https://arxiv.org/abs/2203.03580.

[28] S. Choi, Q.-Y. Zhou, S. Miller, and V. Koltun. A large dataset of object scans. *arXiv preprint arXiv:1602.02481*, 2016.

[29] B. Shen, F. Xia, C. Li, R. Martín-Martín, L. Fan, G. Wang, C. Pérez-D'Arpino, S. Buch, S. Srivastava, L. Tchapmi, M. Tchapmi, K. Vainio, J. Wong, L. Fei-Fei, and S. Savarese. igibson 1.0: A simulation environment for interactive tasks in large realistic scenes. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7520–7527, 2021. doi:10.1109/IROS51168.2021.9636667.

[30] C. Li, F. Xia, R. Martín-Martín, M. Lingelbach, S. Srivastava, B. Shen, K. E. Vainio, C. Gokmen, G. Dharan, T. Jain, A. Kurenkov, K. Liu, H. Gweon, J. Wu, L. Fei-Fei, and S. Savarese. igibson 2.0: Object-centric simulation for robot learning of everyday household tasks. In A. Faust, D. Hsu, and G. Neumann, editors, *Proceedings of the 5th Conference on Robot Learning*, volume 164 of *Proceedings of Machine Learning Research*, pages 455–465. PMLR, 08–11 Nov 2022. URL https://proceedings.mlr.press/v164/li22b.html.

[31] S. Srivastava, C. Li, M. Lingelbach, R. Martín-Martín, F. Xia, K. E. Vainio, Z. Lian, C. Gokmen, S. Buch, K. Liu, S. Savarese, H. Gweon, J. Wu, and L. Fei-Fei. Behavior: Benchmark for everyday household activities in virtual, interactive, and ecological environments. In A. Faust, D. Hsu, and G. Neumann, editors, *Proceedings of the 5th Conference on Robot Learning*, volume 164 of *Proceedings of Machine Learning Research*, pages 477–490. PMLR, 08–11 Nov 2022. URL https://proceedings.mlr.press/v164/srivastava22a.html.

[32] L. Liu, W. Xu, H. Fu, S. Qian, Y. Han, and C. Lu. Akb-48: A real-world articulated object knowledge base, 2022. URL https://arxiv.org/abs/2202.08432.

[33] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba. Openai gym. *CoRR*, abs/1606.01540, 2016. URL http://arxiv.org/abs/1606.01540.

[34] H. Fu, W. Xu, H. Xue, H. Yang, R. Ye, Y. Huang, Z. Xue, Y. Wang, and C. Lu. Rfuniverse: A physics-based action-centric interactive environment for everyday household tasks. *arXiv preprint arXiv:2202.00199*, 2022.