

A Table of Contents

The supplementary material has the following contents:

- **Task and Hardware Setups** (Sec. B): Detailed descriptions of the hardware setups, dataset, and task settings.
- **Model and Training Details** (Sec. C): Descriptions of the neural network architectures used in our experiments and the corresponding training procedures.
- **Ablation Study on Sampling Strategy** (Sec. D): Evaluation and analysis of different sampling strategies.
- **Additional Results** (Sec. E): In-distribution evaluation results for both image-based and point cloud-based policies in the real world.
- **Transport Plan Visualization** (Sec. F): Visualizations of the optimal transport plan for randomly sampled training batches.
- **Visualization of Latent Space** (Sec. G): Visual comparisons of the learned latent spaces between our method and the Co-training baseline across additional tasks.

More video results and analysis can be found on our website: <https://ot-sim2real.github.io/>

B Task and Hardware Setups

To evaluate the effectiveness of our approach, we conduct comprehensive experiments on a suite of robotic tabletop manipulation tasks, covering both sim-to-sim and sim-to-real transfer scenarios. These tasks are designed to test the system’s ability to handle key challenges in robotic manipulation, including dense object interactions, long-horizon reasoning, and high-precision control:

- **Lift**: Grasp the rim of a mug and lift it vertically;
- **BoxInBin**: Grasp a tall box and place it into a bin;
- **Stack**: Grasp a small cube and stack it on top of a longer cuboid;
- **Square**: Grasp the handle of a square-shaped object and insert it onto a peg;
- **MugHang**: Grasp the rim of a mug and hang it on a mug tree using the handle;
- **Drawer**: Open a drawer, grasp a coffee pod from the table, place it into the drawer, and close the drawer.

B.1 Hardware Setups

The system setup is illustrated in Fig. 5. We use a Franka Emika Panda robot controlled via a joint impedance controller [51] running at 20 Hz for policy execution. For data collection, the robot is teleoperated using a Meta Quest 3 headset, with tracked Cartesian poses converted to joint configurations through inverse kinematics. RGB image and depth are captured using an Intel RealSense D435 depth camera.

B.2 Domain Shifts and Observation Gaps

We assess generalization under visual domain shifts in simulation through designing the following target domain shifts:

- **Viewpoint1-Point**: The camera is rotated approximately 30° around the z-axis, resulting in a side view in the target domain compared to a front-facing view in the source. Point cloud observations are used.
- **Viewpoint3-Point**: The camera is rotated approximately 90° around the z-axis, introducing a more extreme viewpoint shift. Point cloud observations are used.
- **Perturbation-Point**: Random noise sampled uniformly from the range $[-0.01, 0.01]$ is added to each point in the point cloud to simulate sensor noise or domain shift.

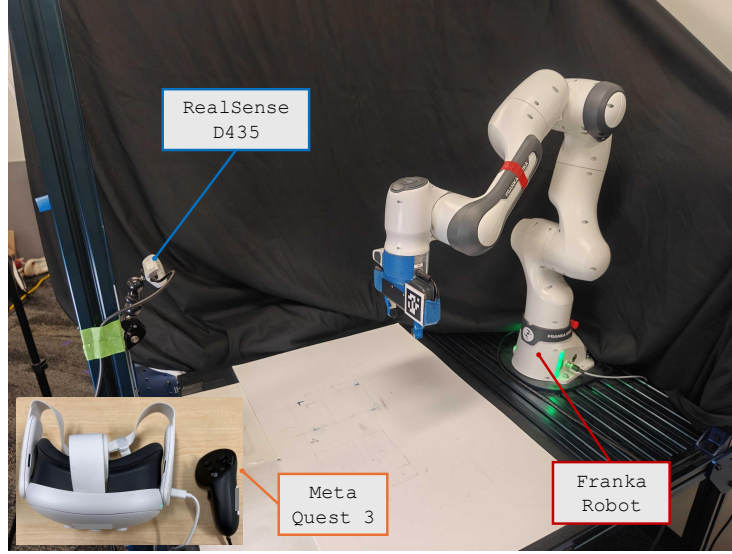


Figure 5: **Hardware Setup.** Our hardware platform uses a Franka Emika Panda robot, with an Intel RealSense D435 camera for capturing image and depth, and a Meta Quest 3 headset for teleoperation.

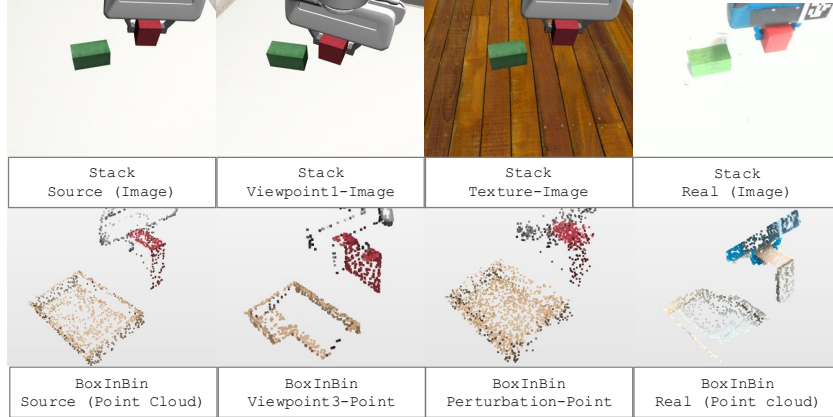


Figure 6: **Observation Gap Across Domains.** Top: image observations for the Stack task from source, Viewpoint1-Image, Texture-Image, and real-world domains. Bottom: point cloud observations for the BoxInBin task from source, Viewpoint3-Point, Perturbation-Point, and real-world domains. Point cloud color is for visualization only and not used as input to the policy.

- Viewpoint1-Image: A 20° camera rotation around the z-axis is applied. RGB image observations are used.
- Texture-Image: The table texture in the target domain is modified. RGB image observations are used.

We illustrate the observation gap across all domains in Fig. 6. The first row displays image observations for the Stack task from the source domain, Viewpoint1-Image, Texture-Image, and the real world. The second row shows point cloud observations for the BoxInBin task from the source domain, Viewpoint3-Point, Perturbation-Point, and the real world. Point cloud color is for visualization only and not used as input to the policy.

B.3 Task Datasets, Reset Ranges, and OOD Variants

We focus primarily on evaluating policy performance in regions covered exclusively by source-domain demonstrations. To conduct controlled experiments, we define three distinct reset regions for each task—Source, Target, and Target-OOD—as shown in Fig. 7. Specifically:

	Lift	Stack	BoxInBin	MugHang	Square	Drawer
Number of real demos	10	25	20	15	25	25

Table 5: **Number of Real-World Demonstrations.** We collect 10–25 demonstrations per task, varying with task difficulty.



Figure 7: **Reset Ranges for Each Task.** The first row illustrates the Source region, where dense source-domain demonstrations are collected. The second row shows the Target and Target-00D reset ranges used in sim-to-sim transfer experiments. In this setting, the Target region is sparsely covered by demonstrations, while the Target-00D region contains no demonstrations and is used exclusively for policy evaluation. The third row similarly presents the Target and Target-00D regions for sim-to-real transfer experiments.

- Source: A large region that is densely covered by demonstrations in the source domain. We generate 1000 demonstrations using MimicGen [10] within the Source region.
- Target: A small subset of the Source region. This region is sparsely covered by demonstrations in the target domain, and is therefore considered in-distribution during evaluation. For sim-to-sim transfer, we collect 10 demonstrations within this region. For sim-to-real transfer, the number of real-world demonstrations collected in the Target region is adjusted based on task difficulty, as detailed in Tab. 5.
- Target-00D: No demonstrations are collected in the Target-00D region, which is used solely for evaluation and treated as out-of-distribution (OOD).

For sim-to-real transfer experiments, in addition to the reset range OOD (denoted as Reset), we consider two additional OOD variants. In the Texture variant, the object’s texture is modified to one that is unseen in the real-world demonstrations. In the Shape variant, the object is replaced with a novel shape not encountered in the real-world demonstrations. These variants are illustrated in Fig. 8.

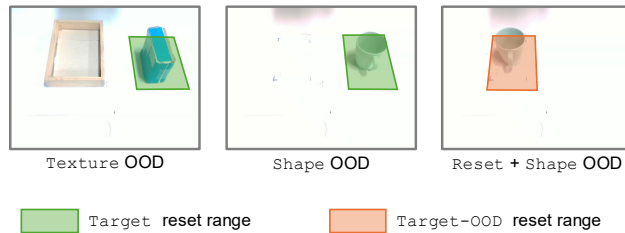


Figure 8: **Texture and Shape OOD in Sim-to-Real Experiments.** Visualization of reset ranges for the BoxInBin task under Texture OOD, and the Lift task under both Shape OOD and Shape+Reset OOD conditions.

973 C Model and Training Details

974 For point cloud-based experiments, we adopt the 3D Diffusion Policy architecture [28] with a PointNet
 975 encoder [47]. The diffusion head receives features extracted from the point cloud observations along
 976 with robot proprioceptive inputs (joint and gripper positions), and outputs 7-DOF target joint positions
 977 and the gripper action. We project the depth map into the robot base frame to generate the scene
 978 point cloud. For a pixel with coordinate (u, v) and depth d , the corresponding 3D location can be
 979 recovered by:

$$p^w = R \cdot K^{-1} \cdot I + t$$

980 where $I = (u \cdot d, v \cdot d, d)$, $[R \mid t]$ denotes the camera pose obtained through hand-eye calibration [52],
 981 and K denotes the camera intrinsic matrix. We crop the reconstructed scene point cloud using
 982 a bounding box defined by $x \in [-0.2, 0.1]$, $y \in [-0.2, 0.2]$, and $z \in [0.008, 0.588]$ to exclude
 983 irrelevant background information. The cropped point cloud is then downsampled to 2048 points
 984 using Farthest Point Sampling (FPS) [53].

985 For experiments with image-based policy, we adopt Diffusion Policy [27] with a ResNet-based [49]
 986 visual encoder. The original images are captured by the camera at a resolution of 480×640 . During
 987 preprocessing, the images are downsampled to 120×160 , followed by random cropping to 108×144
 988 during training and center cropping during testing. The policy takes stacked history images and robot
 989 proprioceptive inputs (joint and gripper positions) as input, and outputs 7-DOF target joint positions
 990 along with the gripper action.

991 Our overall training procedure is summarized in Alg. 1. We use a batch size of 256 for the behavior
 992 cloning loss L_{BC} , with a co-training ratio of 0.9 following Maddukuri et al. [22]. For the optimal
 993 transport loss L_{OT} , the batch size is set to 128, with a weighting coefficient $\lambda = 0.1$. We use
 994 $\epsilon = 0.0005$ and $\tau = 0.01$ in our experiments.

Algorithm 1 Joint Policy Training with OT

Require: Source dataset D_{src} , Target dataset D_{tgt}

- 1: Initialize encoder f_ϕ , and policy π_θ
 - 2: Compute DTW distances for all trajectories pairs in D_{src} and D_{tgt}
 - 3: **for** iteration $t = 1$ to T **do**
 - 4: Sample a paired batch $\{(o_{src}^i, x_{src}^i, a_{src}^i, o_{tgt}^j, x_{tgt}^j, a_{tgt}^j)\}$ with size N from D_{src} and D_{tgt}
 using strategy described in Sec. 4.3
 - 5: Compute features $\{z_{src}^i\}$ and $\{z_{tgt}^j\}$ using encoder f_ϕ
 - 6: Construct ground cost matrix \hat{C}_ϕ as described in Sec. 4.1
 - 7: Compute optimal transport plan $\Pi^* = \arg \min_{\Pi \in \mathbb{R}_+^{N \times N}} (\langle \Pi, \hat{C}_\phi \rangle_F + \epsilon \cdot \Omega(\Pi) + \tau \cdot$
 $\text{KL}(\Pi \mathbf{1} \parallel \mathbf{p}) + \tau \cdot \text{KL}(\Pi^\top \mathbf{1} \parallel \mathbf{q}))$ via Sinkhorn-Knopp algorithm [44]
 - 8: Compute OT loss $L_{OT}(f_\phi) = \langle \Pi^*, \hat{C}_\phi \rangle_F$
 - 9: Sample $\{(o_{src}^i, x_{src}^i, a_{src}^i)\}$ from D_{src} and sample $\{(o_{tgt}^j, x_{tgt}^j, a_{tgt}^j)\}$ from D_{tgt}
 - 10: Compute BC loss $L_{BC}(f_\phi, \pi_\theta)$
 - 11: Update f_ϕ and π_θ with gradients of $L_{BC}(f_\phi, \pi_\theta) + \lambda \cdot L_{OT}(f_\phi)$
 - 12: **end for**
-

995 D Ablation Study on Sampling Strategy

996 To assess the effectiveness of our sampling strategy, we compare the full method against a variant
 997 (denoted as Ours w/o Sampler) that does not apply any trajectory-level sampling. In this baseline,
 998 source and target data are randomly sampled across trajectories and time steps, with no coordination.
 999 We also include an oracle variant (denoted as UOT-Oracle), which constructs perfectly paired
 1000 batches—each state is observed in both the source and target domains to ensure that batch data
 1001 originates from the same underlying states. We evaluate policy performance on the Stack task under
 1002 the Viewpoint1-Point variation, with results shown in Fig. 9

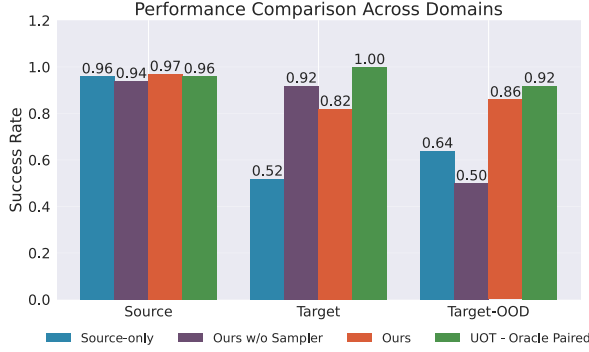


Figure 9: **Sampling Strategies Comparison.** Our proposed sampling strategy (Ours) improves policy success rates on the Stack task with Viewpoint1-Point, outperforming Ours w/o Sampler, and achieving performance comparable to the oracle-paired upper bound (UOT-Oracle).

1003 **Temporal-aware strategy improves pairing quality and downstream performance.** The oracle
 1004 baseline demonstrates that, given perfectly aligned data, unbalanced OT loss significantly enhances
 1005 generalization by enabling the encoder to learn domain-invariant representations. In contrast, the
 1006 no-sampling variant (Ours w/o Sampler) exhibits poor generalization in the Target-OOD setting.
 1007 This degradation likely stems from the low probability of encountering aligned state pairs in mini-
 1008 batches—especially problematic in long-horizon tasks, where uncoordinated sampling rarely produces
 1009 temporally aligned data.

1010 E Additional Results

1011 We conduct extensive real-world evaluations to validate the effectiveness of our approach. Sim-to-real
 1012 transfer results for in-distribution scenarios are reported in Tab. 6 and Tab. 7 for image-based and point
 1013 cloud-based policies, respectively. Results for out-of-distribution (OOD) scenarios are presented in
 1014 the main paper.

	Stack		Square		BoxInBin		Average
	grasp	full	grasp	full	grasp	full	full
Source-only	0.1	0.0	0.0	0.0	0.0	0.0	0.00
Target-only	0.7	0.7	0.8	0.0	0.7	0.7	0.47
Co-training	0.8	0.7	0.8	0.1	0.9	0.8	0.53
Ours	0.9	0.9	0.9	0.4	0.9	0.9	0.73

Table 6: **Real World Image-Based Policy In-Distribution Success Rates.** The Average denotes the average full task success rates over all tasks.

	Stack		Square		BoxInBin		Lift		MugHang		Drawer				Average
	grasp	full	grasp	full	grasp	full	reach	full	grasp	full	open	grasp	place	full	full
S.-o.	0.3	0.0	0.1	0.1	0.4	0.3	0.5	0.5	0.1	0.0	0.0	0.0	0.0	0.0	0.15
T.-o.	0.7	0.4	0.6	0.1	0.9	0.8	1.0	1.0	0.8	0.8	0.9	0.5	0.5	0.5	0.60
Co-t.	0.7	0.7	1.0	0.5	0.8	0.8	0.8	0.8	1.0	0.8	1.0	0.7	0.7	0.4	0.67
Ours	0.8	0.8	1.0	0.4	0.9	0.9	1.0	1.0	1.0	0.8	1.0	0.7	0.7	0.7	0.77

Table 7: **Real World Point-Cloud-Based Policy In-Distribution Success Rates.** The Average denotes the average full task success rates over all tasks.

1015 Experimental results show that our approach consistently outperforms all baselines in real-world
 1016 in-distribution settings. Our method achieves average success rates of 0.73 and 0.77 for image-based
 1017 and point cloud-based policies, respectively, demonstrating its effectiveness in learning complex
 1018 real-world manipulation tasks.

1019 F Transport Plan Visualization

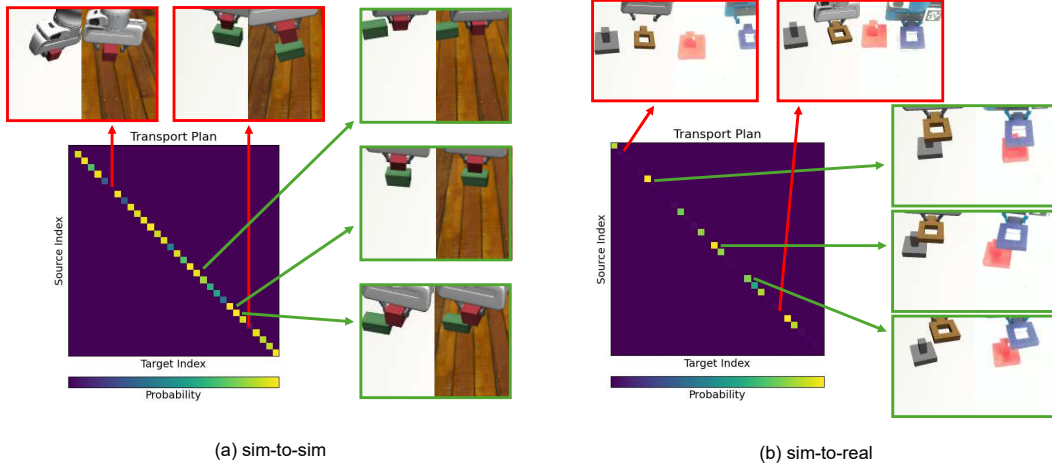


Figure 10: **Transport Plan Visualization.** We visualize the transport plan for a randomly sampled batch during training the image-based policy, alongside corresponding observations from both domains. The left figure shows a sim-to-sim experiment, while the right shows a sim-to-real experiment. The visualization reveals that the transport plan effectively aligns similar states across domains, as indicated by high transport probabilities.

1020 To understand how optimal transport facilitates domain-invariant feature learning and enhances
 1021 cross-domain generalization, we visualize the transport plan for a randomly sampled batch during
 1022 training the image-based policy, along with corresponding observations from both domains (see
 1023 Fig. 10). The left plot shows results from the sim-to-sim transfer experiment, while the right plot
 1024 depicts the sim-to-real setting. The results show that the transport plan effectively aligns similar
 1025 states across domains—reflected by high transport probabilities—encouraging feature proximity and
 1026 promoting robust, domain-invariant representations.

1027 G Visualization of Latent Space

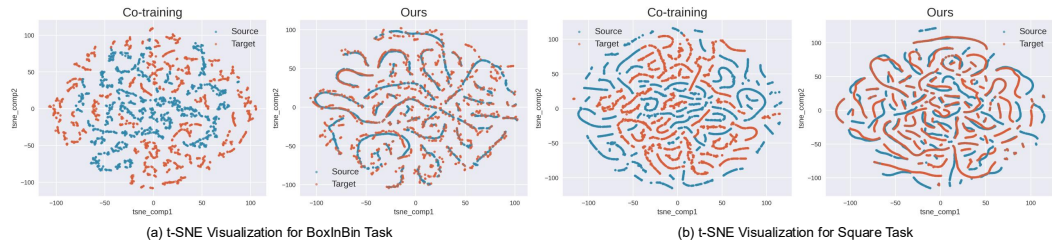


Figure 11: **Latent Space Visualization.** Latent space comparison between the Co-training baseline and our method. In our approach, source-domain points (blue) and target-domain points (red) form a well-mixed cluster, illustrating how OT alignment harmonizes cross-domain feature distributions and enhances transferability and generalization.

1028 Beyond the feature visualization for the Stack task with the Viewpoint1-Point target domain,
 1029 we also present additional t-SNE [50] visualizations in Fig. 11 for the BoxInBin task with the
 1030 Perturbation-Point target domain, and the Square task with the Viewpoint1-Point target
 1031 domain. We compare the latent spaces produced by the Co-training baseline and our method. In
 1032 our approach, source-domain points (blue) and target-domain points (red) form a well-mixed cluster,
 1033 highlighting how OT alignment harmonizes cross-domain feature distributions and improves both
 1034 transferability and generalization.

References

- [1] Dean A Pomerleau. Alvin: An autonomous land vehicle in a neural network. *Advances in neural information processing systems*, 1, 1988.
- [2] Ajay Mandlekar, Danfei Xu, Josiah Wong, Soroush Nasiriany, Chen Wang, Rohun Kulkarni, Li Fei-Fei, Silvio Savarese, Yuke Zhu, and Roberto Martín-Martín. What matters in learning from offline human demonstrations for robot manipulation. In *arXiv preprint arXiv:2108.03298*, 2021.
- [3] Pete Florence, Corey Lynch, Andy Zeng, Oscar A Ramirez, Ayzaan Wahid, Laura Downs, Adrian Wong, Johnny Lee, Igor Mordatch, and Jonathan Tompson. Implicit behavioral cloning. In *Conference on robot learning*, pages 158–168. PMLR, 2022.
- [4] Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, Peter David Fagan, Joey Hejna, Masha Itkina, Marion Lepert, Yecheng Jason Ma, Patrick Tree Miller, Jimmy Wu, Suneel Belkhale, Shivin Dass, Huy Ha, Arhan Jain, Abraham Lee, Youngwoon Lee, Marius Memmel, Sungjae Park, Ilija Radosavovic, Kaiyuan Wang, Albert Zhan, Kevin Black, Cheng Chi, Kyle Beltran Hatch, Shan Lin, Jingpei Lu, Jean Mercat, Abdul Rehman, Pannag R Sanketi, Archit Sharma, Cody Simpson, Quan Vuong, Homer Rich Walke, Blake Wulfe, Ted Xiao, Jonathan Heewon Yang, Arefeh Yavary, Tony Z. Zhao, Christopher Agia, Rohan Baijal, Mateo Guaman Castro, Daphne Chen, Qiuyu Chen, Trinity Chung, Jaimyn Drake, Ethan Paul Foster, Jensen Gao, David Antonio Herrera, Minh Heo, Kyle Hsu, Jiaheng Hu, Donovan Jackson, Charlotte Le, Yunshuang Li, Kevin Lin, Roy Lin, Zehan Ma, Abhiram Maddukuri, Suvir Mirchandani, Daniel Morton, Tony Nguyen, Abigail O’Neill, Rosario Scalise, Derick Seale, Victor Son, Stephen Tian, Emi Tran, Andrew E. Wang, Yilin Wu, Annie Xie, Jingyun Yang, Patrick Yin, Yunchu Zhang, Osbert Bastani, Glen Berseth, Jeannette Bohg, Ken Goldberg, Abhinav Gupta, Abhishek Gupta, Dinesh Jayaraman, Joseph J Lim, Jitendra Malik, Roberto Martín-Martín, Subramanian Ramamoorthy, Dorsa Sadigh, Shuran Song, Jiajun Wu, Michael C. Yip, Yuke Zhu, Thomas Kollar, Sergey Levine, and Chelsea Finn. Droid: A large-scale in-the-wild robot manipulation dataset. 2024.
- [5] Open X-Embodiment Collaboration, Abby O’Neill, Abdul Rehman, Abhinav Gupta, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, Albert Tung, Alex Bewley, Alex Herzog, Alex Irpan, Alexander Khazatsky, Anant Rai, Anchit Gupta, Andrew Wang, Andrey Kolobov, Anikait Singh, Animesh Garg, Aniruddha Kembhavi, Annie Xie, Anthony Brohan, Antonin Raffin, Archit Sharma, Arefeh Yavary, Arhan Jain, Ashwin Balakrishna, Ayzaan Wahid, Ben Burgess-Limerick, Beomjoon Kim, Bernhard Schölkopf, Blake Wulfe, Brian Ichter, Cewu Lu, Charles Xu, Charlotte Le, Chelsea Finn, Chen Wang, Chenfeng Xu, Cheng Chi, Chenguang Huang, Christine Chan, Christopher Agia, Chuer Pan, Chuyuan Fu, Coline Devin, Danfei Xu, Daniel Morton, Danny Driess, Daphne Chen, Deepak Pathak, Dhruv Shah, Dieter Buechler, Dinesh Jayaraman, Dmitry Kalashnikov, Dorsa Sadigh, Edward Johns, Ethan Foster, Fangchen Liu, Federico Ceola, Fei Xia, Feiyu Zhao, Felipe Vieira Frujeri, Freek Stulp, Gaoyue Zhou, Gaurav S. Sukhatme, Gautam Salhotra, Ge Yan, Gilbert Feng, Giulio Schiavi, Glen Berseth, Gregory Kahn, Guangwen Yang, Guanzhi Wang, Hao Su, Hao-Shu Fang, Haochen Shi, Henghui Bao, Heni Ben Amor, Henrik I Christensen, Hiroki Furuta, Homanga Bharadhwaj, Homer Walke, Hongjie Fang, Huy Ha, Igor Mordatch, Ilija Radosavovic, Isabel Leal, Jacky Liang, Jad Abou-Chakra, Jaehyung Kim, Jaimyn Drake, Jan Peters, Jan Schneider, Jasmine Hsu, Jay Vakil, Jeannette Bohg, Jeffrey Bingham, Jeffrey Wu, Jensen Gao, Jiaheng Hu, Jiajun Wu, Jialin Wu, Jiankai Sun, Jianlan Luo, Jiayuan Gu, Jie Tan, Jihoon Oh, Jimmy Wu, Jingpei Lu, Jingyun Yang, Jitendra Malik, João Silvério, Joey Hejna, Jonathan Bozher, Jonathan Tompson, Jonathan Yang, Jordi Salvador, Joseph J. Lim, Junhyek Han, Kaiyuan Wang, Kanishka Rao, Karl Pertsch, Karol Hausman, Keegan Go, Keerthana Gopalakrishnan, Ken Goldberg, Kendra Byrne, Kenneth Oslund, Kento Kawaharazuka, Kevin Black, Kevin Lin, Kevin Zhang, Kiana Ehsani, Kiran Lekkala, Kirsty Ellis, Krishan Rana, Krishnan Srinivasan, Kuan Fang, Kunal Pratap Singh, Kuo-Hao Zeng, Kyle Hatch, Kyle Hsu, Laurent Itti, Lawrence Yunliang Chen, Lerrel Pinto, Li Fei-Fei, Liam Tan, Linxi "Jim" Fan, Lionel Ott, Lisa Lee, Luca Weihs, Magnum Chen, Marion Lepert, Marius Memmel, Masayoshi Tomizuka, Masha Itkina, Mateo Guaman Castro, Max Spero, Maximilian Du, Michael Ahn, Michael C. Yip, Mingtong Zhang, Mingyu Ding,

- Minho Heo, Mohan Kumar Srirama, Mohit Sharma, Moo Jin Kim, Naoaki Kanazawa, Nicklas Hansen, Nicolas Heess, Nikhil J Joshi, Niko Suenderhauf, Ning Liu, Norman Di Palo, Nur Muhammad Mahi Shafiullah, Oier Mees, Oliver Kroemer, Osbert Bastani, Pannag R Sanketi, Patrick "Tree" Miller, Patrick Yin, Paul Wohlhart, Peng Xu, Peter David Fagan, Peter Mitrano, Pierre Sermanet, Pieter Abbeel, Priya Sundareshan, Qiuyu Chen, Quan Vuong, Rafael Rafailov, Ran Tian, Ria Doshi, Roberto Mart'in-Mart'in, Rohan Baijal, Rosario Scalise, Rose Hendrix, Roy Lin, Runjia Qian, Ruohan Zhang, Russell Mendonca, Rutav Shah, Ryan Hoque, Ryan Julian, Samuel Bustamante, Sean Kirmani, Sergey Levine, Shan Lin, Sherry Moore, Shikhar Bahl, Shivin Dass, Shubham Sonawani, Shubham Tulsiani, Shuran Song, Sichun Xu, Siddhant Halder, Siddharth Karamcheti, Simeon Adebola, Simon Guist, Soroush Nasiriany, Stefan Schaal, Stefan Welker, Stephen Tian, Subramanian Ramamoorthy, Sudeep Dasari, Suneel Belkhale, Sungjae Park, Suraj Nair, Suvir Mirchandani, Takayuki Osa, Tanmay Gupta, Tatsuya Harada, Tatsuya Matsushima, Ted Xiao, Thomas Kollar, Tianhe Yu, Tianli Ding, Todor Davchev, Tony Z. Zhao, Travis Armstrong, Trevor Darrell, Trinity Chung, Vidhi Jain, Vikash Kumar, Vincent Vanhoucke, Wei Zhan, Wenxuan Zhou, Wolfram Burgard, Xi Chen, Xiangyu Chen, Xiaolong Wang, Xinghao Zhu, Xinyang Geng, Xiyuan Liu, Xu Liangwei, Xuanlin Li, Yansong Pang, Yao Lu, Yecheng Jason Ma, Yejin Kim, Yevgen Chebotar, Yifan Zhou, Yifeng Zhu, Yilin Wu, Ying Xu, Yixuan Wang, Yonatan Bisk, Yongqiang Dou, Yoonyoung Cho, Youngwoon Lee, Yuchen Cui, Yue Cao, Yueh-Hua Wu, Yujin Tang, Yuke Zhu, Yunchu Zhang, Yunfan Jiang, Yunshuang Li, Yunzhu Li, Yusuke Iwasawa, Yutaka Matsuo, Zehan Ma, Zhuo Xu, Zichen Jeff Cui, Zichen Zhang, Zipeng Fu, and Zipeng Lin. Open X-Embodiment: Robotic learning datasets and RT-X models. <https://arxiv.org/abs/2310.08864>, 2023.
- [6] Genesis Authors. Genesis: A universal and generative physics engine for robotics and beyond, December 2024.
- [7] Fanbo Xiang, Yuzhe Qin, Kaichun Mo, Yikuan Xia, Hao Zhu, Fangchen Liu, Minghua Liu, Hanxiao Jiang, Yifu Yuan, He Wang, Li Yi, Angel X. Chang, Leonidas J. Guibas, and Hao Su. SAPIEN: A simulated part-based interactive environment. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
- [8] Alexander Raistrick, Lingjie Mei, Karhan Kayan, David Yan, Yiming Zuo, Beining Han, Hongyu Wen, Meenal Parakh, Stamatis Alexandropoulos, Lahav Lipson, et al. Infinigen indoors: Photorealistic indoor scenes using procedural generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21783–21794, 2024.
- [9] Matt Deitke, Eli VanderBilt, Alvaro Herrasti, Luca Weihs, Kiana Ehsani, Jordi Salvador, Winson Han, Eric Kolve, Aniruddha Kembhavi, and Roozbeh Mottaghi. Proctor: Large-scale embodied ai using procedural generation. *Advances in Neural Information Processing Systems*, 35:5982–5994, 2022.
- [10] Ajay Mandlekar, Soroush Nasiriany, Bowen Wen, Ireteayo Akinola, Yashraj Narang, Linxi Fan, Yuke Zhu, and Dieter Fox. Mimicgen: A data generation system for scalable robot learning using human demonstrations. In *7th Annual Conference on Robot Learning*, 2023.
- [11] Shuo Cheng, Caelan Garrett, Ajay Mandlekar, and Danfei Xu. Nod-tamp: Multi-step manipulation planning with neural object descriptors. *arXiv preprint arXiv:2311.01530*, 2023.
- [12] OpenAI: Marcin Andrychowicz, Bowen Baker, Maciek Chociej, Rafal Jozefowicz, Bob McGrew, Jakub Pachocki, Arthur Petron, Matthias Plappert, Glenn Powell, Alex Ray, et al. Learning dexterous in-hand manipulation. *The International Journal of Robotics Research*, 39(1):3–20, 2020.
- [13] Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 23–30. IEEE, 2017.
- [14] Nicklas Hansen, Rishabh Jangir, Yu Sun, Guillem Alenyà, Pieter Abbeel, Alexei A Efros, Lerrel Pinto, and Xiaolong Wang. Self-supervised policy adaptation during deployment. *arXiv preprint arXiv:2007.04309*, 2020.

- [15] Denis Yarats, Rob Fergus, Alessandro Lazaric, and Lerrel Pinto. Mastering visual continuous control: Improved data-augmented reinforcement learning. *arXiv preprint arXiv:2107.09645*, 2021.
- [16] Konstantinos Bousmalis, Nathan Silberman, David Dohan, Dumitru Erhan, and Dilip Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3722–3731, 2017.
- [17] Stephen James, Paul Wohlhart, Mrinal Kalakrishnan, Dmitry Kalashnikov, Alex Irpan, Julian Ibarz, Sergey Levine, Raia Hadsell, and Konstantinos Bousmalis. Sim-to-real via sim-to-sim: Data-efficient robotic grasping via randomized-to-canonical adaptation networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12627–12637, 2019.
- [18] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*, 2014.
- [19] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, pages 97–105. PMLR, 2015.
- [20] Han Zhao, Remi Tachet Des Combes, Kun Zhang, and Geoffrey Gordon. On learning invariant representations for domain adaptation. In *International conference on machine learning*, pages 7523–7532. PMLR, 2019.
- [21] Adam Wei, Abhinav Agarwal, Boyuan Chen, Rohan Bosworth, Nicholas Pfaff, and Russ Tedrake. Empirical analysis of sim-and-real cotraining of diffusion policies for planar pushing from pixels. *arXiv preprint arXiv:2503.22634*, 2025.
- [22] Abhiram Maddukuri, Zhenyu Jiang, Lawrence Yunliang Chen, Soroush Nasiriany, Yuqi Xie, Yu Fang, Wenqi Huang, Zu Wang, Zhenjia Xu, Nikita Chernyadev, et al. Sim-and-real co-training: A simple recipe for vision-based robotic manipulation. *arXiv preprint arXiv:2503.24361*, 2025.
- [23] Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 39(9):1853–1865, 2016.
- [24] Kilian Fatras, Thibault Séjourné, Rémi Flamary, and Nicolas Courty. Unbalanced minibatch optimal transport; applications to domain adaptation. In *International Conference on Machine Learning*, pages 3186–3197. PMLR, 2021.
- [25] Lenaïc Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Scaling algorithms for unbalanced optimal transport problems. *Mathematics of computation*, 87(314):2563–2609, 2018.
- [26] Tony Z Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. *arXiv preprint arXiv:2304.13705*, 2023.
- [27] Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, page 02783649241273668, 2023.
- [28] Yanjie Ze, Gu Zhang, Kangning Zhang, Chenyuan Hu, Muhan Wang, and Huazhe Xu. 3d diffusion policy: Generalizable visuomotor policy learning via simple 3d representations. In *Proceedings of Robotics: Science and Systems (RSS)*, 2024.
- [29] Philipp Wu, Yide Shentu, Zhongke Yi, Xingyu Lin, and Pieter Abbeel. Gello: A general, low-cost, and intuitive teleoperation framework for robot manipulators. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 12156–12163. IEEE, 2024.

- [30] Cheng Chi, Zhenjia Xu, Chuer Pan, Eric Cousineau, Benjamin Burchfiel, Siyuan Feng, Russ Tedrake, and Shuran Song. Universal manipulation interface: In-the-wild robot teaching without in-the-wild robots. *arXiv preprint arXiv:2402.10329*, 2024.
- [31] Simar Kareer, Dhruv Patel, Ryan Punamiya, Pranay Mathur, Shuo Cheng, Chen Wang, Judy Hoffman, and Danfei Xu. Egomimic: Scaling imitation learning via egocentric video. *arXiv preprint arXiv:2410.24221*, 2024.
- [32] Zhenyu Jiang, Yuqi Xie, Kevin Lin, Zhenjia Xu, Weikang Wan, Ajay Mandlekar, Linxi Fan, and Yuke Zhu. Dexmimicgen: Automated data generation for bimanual dexterous manipulation via imitation learning. *arXiv preprint arXiv:2410.24185*, 2024.
- [33] Shuo Cheng, Caelan Reed Garrett, Ajay Mandlekar, and Danfei Xu. NOD-TAMP: Generalizable long-horizon planning with neural object descriptors. In *8th Annual Conference on Robot Learning*, 2024.
- [34] Stephen James, Andrew J Davison, and Edward Johns. Transferring end-to-end visuomotor control from simulation to real world for a multi-stage task. In *Conference on Robot Learning*, pages 334–343. PMLR, 2017.
- [35] Zhecheng Yuan, Tianming Wei, Shuiqi Cheng, Gu Zhang, Yuanpei Chen, and Huazhe Xu. Learning to manipulate anywhere: A visual generalizable framework for reinforcement learning. *arXiv preprint arXiv:2407.15815*, 2024.
- [36] Abolfazl Farahani, Sahar Voghoei, Khaled Rasheed, and Hamid R Arabnia. A brief review of domain adaptation. *Advances in data science and information engineering: proceedings from ICDATA 2020 and IKE 2020*, pages 877–894, 2021.
- [37] Daniel Ho, Kanishka Rao, Zhuo Xu, Eric Jang, Mohi Khansari, and Yunfei Bai. Retinagan: An object-aware approach to sim-to-real transfer. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 10920–10926. IEEE, 2021.
- [38] Nicolas Courty, Rémi Flamary, and Devis Tuia. Domain adaptation with regularized optimal transport. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2014, Nancy, France, September 15-19, 2014. Proceedings, Part I 14*, pages 274–289. Springer, 2014.
- [39] Michaël Perrot, Nicolas Courty, Rémi Flamary, and Amaury Habrard. Mapping estimation for discrete optimal transport. *Advances in Neural Information Processing Systems*, 29, 2016.
- [40] Ievgen Redko, Amaury Habrard, and Marc Sebban. Theoretical analysis of domain adaptation with optimal transport. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2017, Skopje, Macedonia, September 18–22, 2017, Proceedings, Part II 10*, pages 737–753. Springer, 2017.
- [41] Nicolas Courty, Rémi Flamary, Amaury Habrard, and Alain Rakotomamonjy. Joint distribution optimal transportation for domain adaptation. *Advances in neural information processing systems*, 30, 2017.
- [42] Bharath Bhushan Damodaran, Benjamin Kellenberger, Rémi Flamary, Devis Tuia, and Nicolas Courty. Deepjdot: Deep joint distribution optimal transport for unsupervised domain adaptation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 447–463, 2018.
- [43] Jan Matas, Stephen James, and Andrew J Davison. Sim-to-real reinforcement learning for deformable object manipulation. In *Conference on Robot Learning*, pages 734–743. PMLR, 2018.
- [44] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.
- [45] Omer Gold and Micha Sharir. Dynamic time warping and geometric edit distance: Breaking the quadratic barrier. *ACM Transactions On Algorithms (TALG)*, 14(4):1–17, 2018.

- 557 [46] Yuke Zhu, Josiah Wong, Ajay Mandlekar, Roberto Martín-Martín, Abhishek Joshi, Soroush
558 Nasiriany, and Yifeng Zhu. robosuite: A modular simulation framework and benchmark for
559 robot learning. *arXiv preprint arXiv:2009.12293*, 2020.
- 560 [47] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point
561 sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer
562 vision and pattern recognition*, pages 652–660, 2017.
- 563 [48] Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burchfiel, and
564 Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. In *Proceedings
565 of Robotics: Science and Systems (RSS)*, 2023.
- 566 [49] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image
567 recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*,
568 pages 770–778, 2016.
- 569 [50] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine
570 learning research*, 9(11), 2008.
- 571 [51] Neville Hogan. Impedance control: An approach to manipulation: Part ii—implementation.
572 1985.
- 573 [52] Roger Y Tsai, Reimar K Lenz, et al. A new technique for fully autonomous and efficient 3 d
574 robotics hand/eye calibration. *IEEE Transactions on robotics and automation*, 5(3):345–358,
575 1989.
- 576 [53] Meng Han, Liang Wang, Limin Xiao, Hao Zhang, Chenhao Zhang, Xiangrong Xu, and Jianfeng
577 Zhu. Quickfps: Architecture and algorithm co-design for farthest point sampling in large-scale
578 point clouds. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*,
579 2023.