GenH2R: Learning Generalizable Human-to-Robot Handover via Scalable Simulation, Demonstration, and Imitation

Zifan Wang*[1,3] Junyu Chen*[1,3] Ziqing Chen[1] Pengwei Xie[1] Rui Chen[1] Li Yi†[1,2,3]
[1]Tsinghua University [2]Shanghai Artificial Intelligence Laboratory [3]Shanghai Qi Zhi Institute
https://GenH2R.github.io

1. Introduction

Recently the AI community focuses on empowering robots to collaborate with humans [1, 22], notably in receiving objects handed over by humans [8, 19, 20]. This human-to-robot (H2R) handover capability enables seamless collaboration in various tasks like cooking and furniture assembly.

However, due to unique challenges, scalable learning of H2R handover lags behind human-free robot manipulation. Real-world human interaction training is costly and risky, urging simulation-based pre-training. However, creating sufficient simulated assets [2, 5, 11, 14, 15, 27] for handover tasks is challenging. In addition, scaling up demonstrations [9, 13, 17] inspired by the success of large language model [3, 18, 29] poses additional challenges. It is very costly and unscalable to collect robot demonstrations.

In this work, we aim to learn generalizable H2R handover at scale by tackling the above challenges. We present a comprehensive solution that scales up both the assets and demonstrations and effectively learns a closed-loop visuomotor policy through a novel imitation learning algorithm.

2. Method

For the generalizable H2R handover task, we introduce GenH2R, a framework for learning control policies, specifically 6D control actions for robot grippers, using segmented point cloud data captured from an egocentric camera.

GenH2R-Sim To scale up geometry and motion assets depicting humans handing over various objects, we leverage large-scale 3D model repositories [4, 10], dexterous grasp generation methods [25], and curve-based 3D animation. This enables us to procedurally generate millions of handover scenes, forming an environment named GenH2R-Sim to support generalizable H2R handover learning. GenH2R-Sim surpasses HandoverSim [6], an existing H2R simulator, in both scene quantity (by three orders of magnitude) and unique object involvement (by two orders of magnitude). In addition, scenes in GenH2R-Sim go beyond a straightforward giving and then receiving and cover cases when humans might keep transforming the object in a large range during the entire H2R handover process. This allows for studying complex behaviors such as humans hesitating before handing over.

Generating Demonstrations for Distillation To scale up robot demonstrations, we draw inspiration from the Task and Motion Planning (TAMP) [13] literature and propose to automatically generate demonstrations with grasp and motion planning using privileged human motion and object state information. There are some straightforward ways to achieve this goal [12, 16, 23, 26, 28], such as using the privileged human handover destination information to plan a smooth demonstration. However, the problem is more challenging than it seems since the generated demonstrations need to be suitable for distilling into a visuomotor policy.

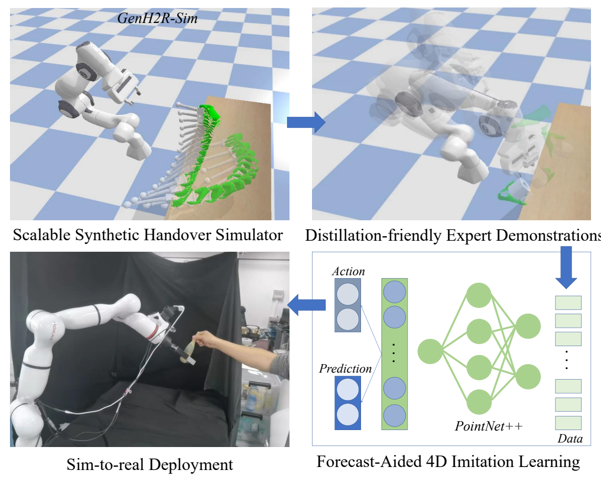

Figure 1. **The overview of GenH2R.** We introduce a framework for learning generalizable vision-based human-to-robot handover via scalable synthetic simulation, distillation-friendly expert demonstration generation, and a forecast-aided 4D imitation learning method. Our models demonstrate strong generalization capabilities to real datasets and can be deployed to a real robot.

*Equal contribution with the order determined by rolling dice.
†Corresponding author.

		s0 (Sequential)			s0 (Simultaneous)			t0			t1		
		S	T	AS	S	T	AS	S	T	AS	S	T	AS
train on s0	GA-DDPG [24]	50.00	**7.14**	22.5	36.81	**4.66**	23.6	23.59	7.31	10.3	46.7	**5.50**	26.9
	Handover-Sim2real [7]	75.23	7.74	**30.4**	68.75	6.23	35.8	29.17	6.29	15.0	52.40	7.09	23.8
	Handover-Sim2real* [7]	64.35	7.61	26.7	25.69	5.43	15.0	28.56	4.73	17.9	30.60	5.98	16.5
	Destination Planning	74.31	9.01	22.8	76.16	6.98	35.2	25.68	5.96	14.1	48.4	8.94	15.1
	Dense Planning	74.77	9.54	19.8	75.45	7.32	33.0	27.30	6.26	14.1	52.3	9.24	15.1
	Landmark Planning	77.78	9.24	22.3	79.17	7.26	34.9	29.63	6.23	15.4	54.2	9.02	16.6
train on t0	GA-DDPG [24]	54.76	7.26	24.2	44.68	5.30	26.5	24.05	4.70	15.3	25.50	5.86	14.1
	Handover-Sim2real [7]	65.97	7.18	29.5	62.50	6.04	33.5	33.71	5.91	18.4	47.10	6.35	24.1
	Handover-Sim2real* [7]	63.55	7.58	26.5	38.89	5.29	23.1	33.31	**4.64**	21.4	33.35	5.81	18.4
	Destination Planning	0.93	12.80	0.01	6.48	12.41	0.3	5.96	8.81	1.9	1.60	12.03	0.1
	Dense Planning	81.48	9.51	21.9	84.95	7.45	36.3	38.04	7.16	17.1	57.90	8.85	18.4
	Landmark Planning	**86.57**	8.81	28.0	**85.65**	6.58	**42.8**	**41.43**	6.01	**22.3**	**68.33**	7.70	**27.9**

Table 1. **Evaluating on different benchmarks.** We compare our method against baselines from the test set of HandoverSim [6] benchmark ("s0 (sequential)" and "s0 (simultaneous)") and our GenH2R-Sim benchmark ("t0" and "t1"). We use the best-pretrained models from the repositories of GA-DDPG [24] and Handover-Sim2real [7] for evaluation. The results for our method are averaged across 3 random seeds. Note that S means success rate(%). T means time(s). AS means average success(%). *: We reproduce the results of HandoverSim2real in the true simultaneous setting to make a fair comparison.

We identify the vision-action correlation between visual observations and planned actions as the crucial factor influencing distillability and point out that due to the constraints of robot arm morphology one can easily generate observation-irrelevant actions and thus harm distillation. To tackle this challenge, we present a distillation-friendly demonstration generation method that sparsely samples handover animations for landmark states and periodically replans grasp and motion based on privileged future landmarks.

Forecast-Aided 4D Imitation Learning To distill the above demonstrations into a visuomotor policy, we utilize point cloud input for its richer geometric information and smaller sim-to-real gap compared to images. We propose a 4D imitation learning method that factors the sequential point cloud observations into geometry and motion parts [21], facilitating policy learning by better revealing the current scene state. The imitation objective is augmented by a forecasting objective which predicts the future motion of the handover object. Since our demonstrating actions are generated based on future landmarks, the forecasting objective can help further exploit the vision-action correlation.

3. Experiments

Dataset (1) HandoverSim [6] includes 1000 real-world handover scenes and 20 DexYCB objects ("s0"). (2) GenH2R-Sim offers 1,000,000 synthetic handover scenes with 3266 objects ("t0"), comprising 1,000,000 training and 3260 testing scenes. To augment real-world scenarios, We also create 1000 real-world testing scenes ("t1") from HOI4D [15].
Metric We report the successful rate and the execution time as usual. To evaluate both success rate and completion efficiency, we introduce AS (Average Success):

$$\mathrm{AS} = \int_0^1 \mathrm{Success}(t)\,\mathrm{d}t \tag{1}$$

Methods	Simple Setting	Complex Setting
Handover-Sim2real	56.7%	33.3%
Ours	90.0%	70.0%

Table 2. **Sim-to-Real Experiments.** We report the success rate of our method and HandoverSim2real in 2 different settings.

where $\mathrm{Success}(t)$ is success rate considering only successful cases within $t \cdot T_{\max}$ ($T_{max} = 13s$).

Evaluating on Different Benchmarks We have 2 training sets: small-scale real-world "s0" from HandoverSim and large-scale synthetic "t0" from our GenH2R-Sim. Evaluation is conducted on 4 testing sets as depicted in Table 1.

Results on different datasets Our method trained on "t0" outperform all methods trained on "s0" by a large margin. Trained on "s0", our method achieved 11.34%, 16.90%, 12.26%, and 15.93% increase in the success rate. This demonstrates that a substantial amount of synthetic data is more beneficial than only a small-scale real-world dataset.

Results for different methods Our method outperforms baseline methods by a large margin. When trained on "t0", our landmark planning method gives substantial improvements of 20.78%, 23.15%, 7.72%, and 21.23% (23.02%, 46.76%, 8.12%, and 34.98% in our reproduced version).

Sim-to-Real Transfer We deploy the models trained in GenH2R-Sim on a real robotic platform. For the user study, We recruited 6 users to compare our method (based on landmark planning) and Handover-Sim2real across 5 objects in 2 different settings. As shown in Table 2, our model gets better performance in completing the handover process across various objects and scenarios.

For further methodological details and experiment specifics, please refer to our website.

References

[1] Christoph Bartneck, Tony Belpaeme, Friederike Eyssel, Takayuki Kanda, Merel Keijsers, and Selma Šabanović. *Human-robot interaction: An introduction*. Cambridge University Press, 2020. 1

[2] Samarth Brahmbhatt, Chengcheng Tang, Christopher D Twigg, Charles C Kemp, and James Hays. Contactpose: A dataset of grasps with object contact and hand pose. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*, pages 361–378. Springer, 2020. 1

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 1

[4] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 1

[5] Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, et al. Dexycb: A benchmark for capturing hand grasping of objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9044–9053, 2021. 1

[6] Yu-Wei Chao, Chris Paxton, Yu Xiang, Wei Yang, Balakumar Sundaralingam, Tao Chen, Adithyavairavan Murali, Maya Cakmak, and Dieter Fox. Handoversim: A simulation framework and benchmark for human-to-robot object handovers. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 6941–6947. IEEE, 2022. 1, 2

[7] Sammy Christen, Wei Yang, Claudia Pérez-D'Arpino, Otmar Hilliges, Dieter Fox, and Yu-Wei Chao. Learning human-to-robot handovers from point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9654–9664, 2023. 2

[8] Gianluca Corsini, Martin Jacquet, Hemjyoti Das, Amr Afifi, Daniel Sidobre, and Antonio Franchi. Nonlinear model predictive control for human-robot handover with application to the aerial case. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7597–7604. IEEE, 2022. 1

[9] Murtaza Dalal, Ajay Mandlekar, Caelan Garrett, Ankur Handa, Ruslan Salakhutdinov, and Dieter Fox. Imitating task and motion planning with visuomotor transformers. *arXiv preprint arXiv:2305.16309*, 2023. 1

[10] Clemens Eppner, Arsalan Mousavian, and Dieter Fox. Acronym: A large-scale grasp dataset based on simulation. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6222–6227. IEEE, 2021. 1

[11] Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J Black, and Otmar Hilliges. Arctic: A dataset for dexterous bimanual hand-object manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12943–12954, 2023. 1

[12] Hao-Shu Fang, Chenxi Wang, Hongjie Fang, Minghao Gou, Jirong Liu, Hengxu Yan, Wenhai Liu, Yichen Xie, and Cewu Lu. Anygrasp: Robust and efficient grasp perception in spatial and temporal domains. *IEEE Transactions on Robotics*, 2023. 1

[13] Caelan Reed Garrett, Rohan Chitnis, Rachel Holladay, Beomjoon Kim, Tom Silver, Leslie Pack Kaelbling, and Tomás Lozano-Pérez. Integrated task and motion planning. *Annual review of control, robotics, and autonomous systems*, 4:265–293, 2021. 1

[14] Yana Hasson, Gul Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11807–11816, 2019. 1

[15] Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, and Li Yi. Hoi4d: A 4d egocentric dataset for category-level human-object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21013–21022, 2022. 1, 2

[16] Naresh Marturi, Marek Kopicki, Alireza Rastegarpanah, Vijaykumar Rajasekaran, Maxime Adjigble, Rustam Stolkin, Aleš Leonardis, and Yasemin Bekiroglu. Dynamic grasp and trajectory planning for moving objects. *Autonomous Robots*, 43:1241–1256, 2019. 1

[17] Michael James McDonald and Dylan Hadfield-Menell. Guided imitation of task and motion planning. In *Conference on Robot Learning*, pages 630–640. PMLR, 2022. 1

[18] OpenAI. Gpt-4 technical report, 2023. 1

[19] Valerio Ortenzi, Akansel Cosgun, Tommaso Pardi, Wesley P Chan, Elizabeth Croft, and Dana Kulić. Object handovers: a review for robotics. *IEEE Transactions on Robotics*, 37(6): 1855–1873, 2021. 1

[20] Patrick Rosenberger, Akansel Cosgun, Rhys Newbury, Jun Kwan, Valerio Ortenzi, Peter Corke, and Manfred Grafinger. Object-independent human-to-robot handovers using real time robotic vision. *IEEE Robotics and Automation Letters*, 6(1):17–23, 2020. 1

[21] Szymon Rusinkiewicz and Marc Levoy. Efficient variants of the icp algorithm. In *Proceedings third international conference on 3-D digital imaging and modeling*, pages 145–152. IEEE, 2001. 2

[22] Thomas B Sheridan. Human–robot interaction: status and challenges. *Human factors*, 58(4):525–532, 2016. 1

[23] Lirui Wang, Yu Xiang, and Dieter Fox. Manipulation trajectory optimization with online grasp synthesis and selection. In *Proceedings of Robotics: Science and Systems (RSS)*, 2020. 1

[24] Lirui Wang, Yu Xiang, Wei Yang, Arsalan Mousavian, and Dieter Fox. Goal-auxiliary actor-critic for 6d robotic grasping with point clouds. In *Conference on Robot Learning*, pages 70–80. PMLR, 2022. 2

[25] Ruicheng Wang, Jialiang Zhang, Jiayi Chen, Yinzhen Xu, Puhao Li, Tengyu Liu, and He Wang. Dexgraspnet: A

large-scale robotic dexterous grasp dataset for general objects based on simulation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11359–11366. IEEE, 2023. 1

[26] Wei Yang, Chris Paxton, Arsalan Mousavian, Yu-Wei Chao, Maya Cakmak, and Dieter Fox. Reactive human-to-robot handovers of arbitrary objects. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3118–3124. IEEE, 2021. 1

[27] Ruolin Ye, Wenqiang Xu, Zhendong Xue, Tutian Tang, Yanfeng Wang, and Cewu Lu. H2o: A benchmark for visual human-human object handover analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15762–15771, 2021. 1

[28] Gu Zhang, Hao-Shu Fang, Hongjie Fang, and Cewu Lu. Flexible handover with real-time robust dynamic grasp trajectory generation. *arXiv preprint arXiv:2308.15622*, 2023. 1

[29] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022. 1