MANIPTRANS: Efficient Dexterous Bimanual Manipulation Transfer via Residual Learning

Supplementary Material

This appendix provides additional details and results that complement the main paper. We first validate the extensibility of MANIPTRANS in Appendix A. We then evaluate the robustness of MANIPTRANS under noisy conditions in Appendix B and analyze its time cost in Appendix C. Detailed information on the settings of MANIPTRANS is provided in Appendix D, along with statistics for the DEXMANIPNET dataset in Appendix E. Finally, we present the training details for the rearrangement policies in Appendix F.

A. Further Extension of MANIPTRANS

A.1. Articulated Object Manipulation

We demonstrate the extensibility of MANIPTRANS by applying it to the ARCTIC dataset [9], which includes approximately 10 articulated objects, each with precise hand manipulation trajectories for bimanual single-object manipulation tasks.

To accommodate the articulated object manipulation task, we extend our method pipeline. For a single articulated object o^{A}, we define its trajectory as $\boldsymbol{\mathcal{T}}_{o^{\mathrm{A}}} = \{\boldsymbol{\tau}^t_{o^{\mathrm{A}}}\}_{t=1}^{T}$, where $\boldsymbol{\tau}_{o^{\mathrm{A}}} = \{\boldsymbol{p}_{o^{\mathrm{A}}}, \dot{\boldsymbol{p}}_{o^{\mathrm{A}}}, \theta_{o^{\mathrm{A}}}, \dot{\theta}_{o^{\mathrm{A}}}\}$ represents the object's transformation, velocity, and the angle and angular velocity of its articulated part. The reward function for articulated objects, $r^t_{\text{object}^{\mathrm{A}}}$, includes two additional terms compared to the reward for rigid objects: the angle difference $|\theta_{o^{\mathrm{A}}} - \theta_{\hat{o^{\mathrm{A}}}}|$ and the angular velocity difference $|\dot{\theta}_{o^{\mathrm{A}}} - \dot{\theta}_{\hat{o^{\mathrm{A}}}}|$, where $\hat{o^{\mathrm{A}}}$ represents the collidable articulated object in the simulation environment [19]. Apart from this modification, the rest of the pipeline remains unchanged.

Qualitative results of MANIPTRANS applied to the ARCTIC dataset are presented in Fig. 1, demonstrating that our method successfully imitates human demonstrations and rotates the articulated object to the desired target angle. This highlights the extensibility of our pipeline when the physical properties of the articulated object can be accurately modeled in simulation.

A.2. Challenging Hand Embodiments

We investigate the generalization capabilities of MANIPTRANS across different hand embodiments in the main paper. Here, we provide further details on adapting MANIPTRANS to a challenging hand model: the Allegro Hand [2], which possesses $K = 16$ degrees of freedom. The challenges encountered stem from two primary factors: 1) the Allegro Hand has only four fingers, a significant deviation from the structure of the human hand, and 2) the Allegro Hand is approximately twice the size of a human hand. These morphological discrepancies present substantial challenges in transferring human demonstrations to the Allegro Hand.

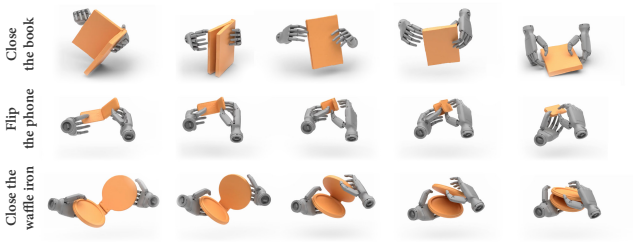

Figure 1. **Applying MANIPTRANS to Articulated Object Manipulation.** In the first row, the two hands collaborate to not only close the book but also place it stably on the table.

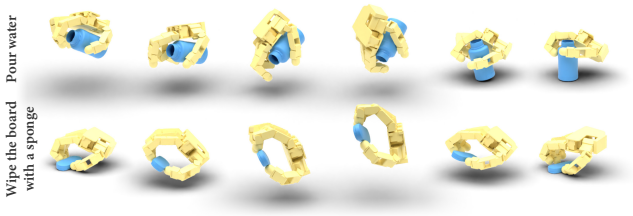

Figure 2. **Extending MANIPTRANS to the Allegro Hand.** Despite the Allegro Hand having only four fingers and a significantly larger size, the transferred motion remains stable and natural.

To address these challenges, we adaptively modify the fingertip mapping relationships, mapping both the pinky and ring fingers to the same fingertip on the Allegro Hand. Additionally, we relax the fingertip keypoint threshold ϵ_{finger} to 8 cm to accommodate the larger dimensions of the Allegro Hand. Successful application of MANIPTRANS to the Allegro Hand is demonstrated in Fig. 2.

A.3. Discussion on the Extension

To summarize, we present all settings for the extension experiments in Tab. 1. The green checkmark (✓) indicates the successful transfer of the dataset to the specified hand embodiment, with results included in DEXMANIPNET. The blue checkmark (✓) denotes dataset verification, where MANIPTRANS is tested on only a subset of the dataset to assess generalizability. The results demonstrate that our pipeline effectively accommodates various morphological differences across hand embodiments and supports a

Hands \ Datasets	FAVOR [16]	OakInk-V2 [26]	GRAB [23]	ARCTIC [9]
Inspire [3]	✓	✓	✓	✓
Shadow [1]	✓	✓	✓	✓
Arti-MANO [21]	✓	✓	✓	✓
Allegro [2]	✓	✓	✓	✓

Table 1. **Extensibility of ManipTrans.** Arti-MANO refers to the articulated MANO hand used in [8].

wide range of tasks, including single-hand manipulation, bimanual articulated object manipulation, and bimanual two-object manipulation.

As discussed in Sec. 3.4 of the main paper, FAVOR [16] and OakInk-V2 [26] represent the largest datasets with the most diverse task types, while the Inspire Hand is distinguished by its high dexterity, stability, cost-effectiveness, and extensive prior use [6, 11, 14]. Consequently, this setup was chosen for collecting **DexManipNet**. However, **ManipTrans** is fully adaptable, and we demonstrate that all of the aforementioned MoCap datasets can be transferred to other robotic hands. We welcome further collaboration from the research community.

B. Robustness Evaluation

MoCap data and model-based pose estimation results often contain noise. To assess whether **ManipTrans** can reliably transfer noisy real-world data into stable robotic motions within a simulation environment, we conduct robustness tests. Since **ManipTrans** is designed for general-purpose transfer and does not depend on task-specific reward functions (e.g., the twisting reward proposed in [18] for the lip-twisting task), noisy object trajectories may introduce instability during the rollout process. Thus, to evaluate **ManipTrans**'s performance under such conditions, we introduce random Gaussian noise into the hand trajectory input and focus on single-hand manipulation tasks. This choice is motivated by the fact that most hand pose estimation methods [17, 24, 25] are optimized for single-hand scenarios.

The results, presented in Tab. 2, demonstrate that **ManipTrans** maintains acceptable performance even when the noise level reaches up to 1.5 cm. These findings highlight the potential of **ManipTrans** for real-world scaling, particularly in applications involving hand pose estimation

Noise	$E_r \downarrow$	$E_t \downarrow$	$E_j \downarrow$	$E_{ft} \downarrow$	$SR \uparrow$
$+\sigma = 0.5\ cm$	9.15	0.51	2.40	1.66	55.1 / 30.1
$+\sigma = 1.0\ cm$	9.56	0.57	2.87	2.13	55.3 / 19.5
$+\sigma = 1.5\ cm$	9.65	0.69	3.29	2.69	46.7 / 39.2

Table 2. **Quantitative Results Under Different Noise Levels.** We add the Gaussian noise $\mathcal{N}(0, \sigma^2)$ to the target hand joints poses.

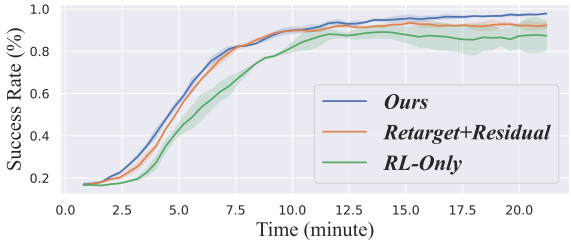

Figure 3. **Detailed Efficiency Comparison.** The success rate curves for the "rotating a mouse" task.

from web video data, which may implicitly contain a vast array of dexterous manipulation skills.

C. Time Cost Analysis

In Sec. 4.3 of the main paper, we compare the efficiency of our method with the previous SOTA method, QuasiSim. QuasiSim employs a set of quasi-physical simulations, dividing the transfer process into three primary stages, with each stage requiring approximately 10-20 hours for a 60-frame trajectory [1]. Since **ManipTrans** also follows a multi-stage framework, incorporating both a pre-trained hand imitation module and a residual refinement module tailored to physical dynamics, we provide a more comprehensive comparison of efficiency.

For a fair evaluation, we use the official QuasiSim demo data for the "rotating a mouse" task as a representative example. The success rate curves for three different settings, as discussed in Sec. 4.3 of the main paper, are shown in Fig. 3: 1) *RL-Only*: This approach trains the policy network from scratch using RL with our reward design. The curve illustrates the entire training process. 2) *Retarget + Residual* Learning: Inspired by [27], this method retargets human hand poses to initial dexterous hand poses via keypoint alignment [20], followed by residual learning for refinement. The retargeting process is performed via parallel optimization and only requires approximately several minutes on a single GPU to optimize full sequence. The training curve for the residual learning stage is represented by the orange line. 3) **ManipTrans**: We pre-train the hand imitation model on a large-scale training dataset, as described in Sec. 3.2, which takes approximately 1.5 days on a single GPU to obtain the reusable imitator. The residual learning stage training curve is shown by the blue line.

From the results in Fig. 3, we observe that for the relatively simple task of "rotating a mouse", the *Retarget + Residual* method achieves performance comparable to **ManipTrans** but requires slightly more time to converge. The *RL-Only* approach, while yielding suboptimal performance compared to the other methods, still produces ac-

[1] As reported in the official repository: https://github.com/Meowuu7/QuasiSim

ceptable motions within 20 minutes. This indicates that our reward design effectively accelerates the training process, facilitating faster convergence.

D. Details of ManipTrans Settings

D.1. Correspondence Between Human Hand and Dexterous Hand

Due to the significant morphological differences between human hands and dexterous robotic hands, we manually establish correspondences between them. For the human hand's fingertip keypoints, we select the midpoint of the three tip anchors as defined in [25]. For the dexterous hands, given their varying shapes, we define the fingertip keypoints as the points of maximum curvature along the central axis of the finger pads, as these points are most likely to contact objects. For other keypoints, such as the wrist and phalanges, we intuitively align the rotation axes of the human joints with those of the robotic joints. For further details, please refer to our code implementation.

In addition, regarding the articulated MANO model, the original human hand model MANO [21] has 45-DoF, which presents extreme challenges for RL-based policies due to the vast exploration space. To mitigate this, we follow the approach in [25] by constraining certain DoFs and fixing the hand collision meshes, thereby reducing the original MANO model to a 22-DoF articulated MANO.

D.2. Details of Training Parameters

In this section, we present the core parameters of our reward functions in **ManipTrans**. The reward parameters for r^t_{finger} in Eq. (1) of the main paper are summarized in Tab. 3. These parameters are determined based on the observation that the thumb, index, and middle fingers play a pivotal role in grasping and manipulation tasks, as they statistically interact with objects more frequently than other fingers [4, 5, 26]. Consequently, the weights are assigned according to the contact frequency. In our implementation, if a dexterous hand lacks a specific finger or joint (*e.g.*, the Inspire Hand does not have distal joints), the corresponding parameters are set to zero. For the contact reward r^t_{contact} in Eq. (2) of the main paper, we set both parameters, w_c and λ_c, to 1.

Figners	weight w_f	decay rate λ_f
Thumb	0.5, 0.3, 0.3, 0.9	50, 40, 40, 100
Index	0.5, 0.3,0.3, 0.8	50, 40, 40, 90
Middle	0.5, 0.3,0.3, 0.75	50, 40, 40, 80
Ring	0.5, 0.3,0.3, 0.6	50, 40, 40, 60
Pinky	0.5, 0.3,0.3, 0.6	50,40, 40, 60

Table 3. **Hyperparameters for the Finger Reward.** The weight w_f and decay rate λ_f are used to balance the importance of each finger. Each cell in the table contains four values, representing the parameters for the proximal, intermediate, distal, and tip joints, respectively. For anatomical definitions, please refer to [25].

D.3. Details of Simulation Parameters

In the Isaac Gym environment, configuring physical properties significantly influences the success rate of transfer. Alongside domain randomization (DR) during training, we set physical constants as follows. For certain objects in OakInk-V2 [26], we obtained actual masses by directly measuring them in collaboration with the dataset authors. For the remaining objects, we assigned a constant density of 200 kg/m^3, approximating the average density of low-fill-rate 3D-printed models. Using this density, we recalculated the objects' masses and moments of inertia.

It is worth noting that human skin is elastic. When grasping objects, fingertip skin undergoes slight deformations, enhancing contact with object surfaces and generating suitable friction, whereas dexterous robotic hands lack this behavior. Previous kinematics-based grasp generation methods [13, 15] often permit slight penetration between fingertips and object surfaces to improve interaction stability (for detailed discussion, please refer to [13]). Therefore, to compensate for the absence of skin deformation in simulation, we set the friction coefficient $\mathcal{F}$ slightly higher than the real-world value. Accurately simulating contact-rich scenarios remains an area for future exploration.

E. DexManipNet Statistics

To the best of our knowledge, no prior work has collected a large-scale bimanual manipulation dataset in which all tra-

assemble, brush whiteboard, cap, cap the pen, close book, close gate, close laptop lid, cut, flip close tooth paste cap, flip open tooth paste cap, heat beaker, heat test tube, hold, hold test tube, ignite alcohol lamp, insert lightbulb, insert pencil, insert usb, open gate, open laptop lid, place asbestos mesh, place inside, place on test tube rack, place onto, place test tube on rack with holder, plug in power plug, pour, pour in lab, press button, put flower into vase, put off alcohol lamp, put on lid, rearrange, remove from test tube rack, remove lid, remove pencil, remove power plug, remove test tube, remove test tube from rack with holder, remove the pen cap, remove usb, scoop, scrape, screw, shake lab container, sharpen pencil, shear paper, spread, squeeze tooth paste, stir, stir experiment substances, swap, take outside, trigger lever, uncap, uncap alcohol lamp, unscrew, use mouse, wipe, write on paper, write on whiteboard

Table 4. List of tasks in the **DexManipNet** dataset. Tasks with underlined names usually require bimanual manipulation.

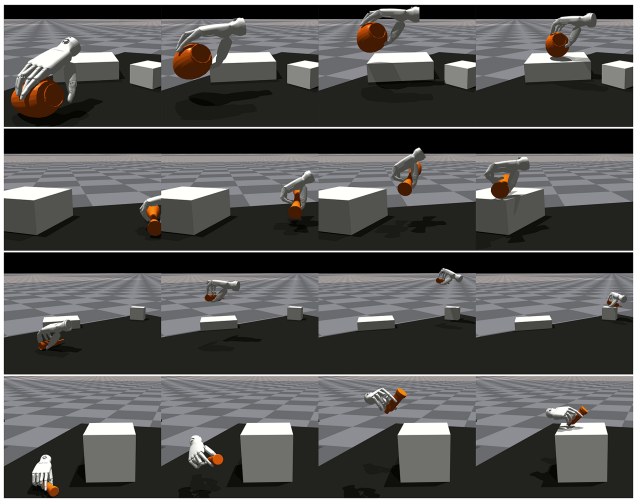

Figure 4. **Qualitative Results of Rearrangement Policy Learning.** The policy successfully moves the bottle to the goal position. Results are directly visualized in the IsaacGym environment, highlighting distinctions between these policies and **ManipTrans**'s rollouts.

jectories are directly transferred from real human demonstrations without the use of teleportation. Leveraging the efficiency and precision of **ManipTrans**, our dataset, **DexManipNet**, comprises 3.3K diverse manipulation trajectories across 61 distinct tasks, as detailed in Tab. 4. To ensure stability during simulation, we fix the object meshes to a watertight state using ManifoldPlus [12] and may slightly adjust the object size to enhance object-object interactions (e.g., the cap and body of the bottle).

Additionally, we provide sample data on our website, showcasing trajectories generated from our policy in simulation. A simple first-order low-pass filter ($\alpha = 0.4$) is applied to the rollouts, effectively reducing jitter with minimal impact on tracking accuracy.

F. Details of Rearrangement Policy Learning

As discussed in Sec. 4.7 of the main paper, we benchmark **DexManipNet** using four data-driven imitation learning methods on the *moving a bottle to a goal position* task.

The primary challenge in this task is to enable the dexterous hand to maintain a stable grasp on the object while smoothly placing it at the specified goal position. We evaluate the dataset using four methods: IBC [10], BET [22], and Diffusion Policy [7], which include both UNet- and Transformer-based architectures. These policies are trained for 500 epochs using the Adam optimizer with a learning rate of 1×10^{-4}, while all other hyperparameters remain at their default settings.

The dimensions of the observation and action spaces for these policies are provided in Tab. 5. The observation space includes the current object state $\{\boldsymbol{p_{\hat{o}}}, \dot{\boldsymbol{p}}_{\hat{\boldsymbol{o}}}\}$, the hand wrist state $\{\boldsymbol{w_d}, \dot{\boldsymbol{w}}_{\boldsymbol{d}}\}$, hand joint angles $\boldsymbol{q_d}$, and the goal poses for both the object $\boldsymbol{g_{\hat{o}}}$ and the hand wrist $\boldsymbol{g_w}$. The action $\boldsymbol{a} = \{\boldsymbol{a_q}, \boldsymbol{a_w}\} \in \boldsymbol{\mathcal{A}}$ specifies the target hand joint angles and wrist poses using a PD controller. Note that PD control is used for wrist poses rather than a 6-DoF force, as is done in **ManipTrans**.

We evaluate the policies' performance on previously unseen goal positions within the IsaacGym environment [19]. A rollout is considered successful if the object's distance from the goal position is within 10 cm; otherwise, it is classified as a failure. Qualitative results are presented in Fig. 4, while quantitative results are summarized in Tab. 2 of the main paper.

Observation	Dimensions
Hand joint angles $\boldsymbol{q}$	12
Hand wrist state $\{\boldsymbol{w_d}, \dot{\boldsymbol{w}}_{\boldsymbol{d}}\}$	13
Object state $\{\boldsymbol{p_{\hat{o}}}, \dot{\boldsymbol{p}}_{\hat{\boldsymbol{o}}}\}$	13
Object pose goal $\boldsymbol{g_{\hat{o}}}$	7
Hand wrist pose goal $\boldsymbol{g_w}$	7

(a) Observation space.

Action	Dimensions
Hand joint angles $\boldsymbol{a_q}$	12
Hand wrist pose $\boldsymbol{a_w}$	7

(b) Action space.

Table 5. **Observation and Action Definitions for the Imitation Policy.** The policy's 7-dimensional pose includes both position and orientation, represented as XYZW quaternions. The policy's 13-dimensional state extends this pose by incorporating both linear and angular velocities.

References

[1] ShadowRobot. https://www.shadowrobot.com/dexterous-hand-series, 2005. 2

[2] Allegro Hands. https://www.allegrohand.com, 2013. 1, 2

[3] Inspire Hands. https://en.inspire-robots.com/product-category/the-dexterous-hands, 2019. 2

[4] Samarth Brahmbhatt, Cusuh Ham, Charles C. Kemp, and James Hays. ContactDB: Analyzing and predicting grasp contact via thermal imaging. In *CVPR*, 2019. 3

[5] 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 *ECCV*, 2020. 3

[6] Xuxin Cheng, Jialong Li, Shiqi Yang, Ge Yang, and Xiaolong Wang. Open-television: Teleoperation with immersive active visual feedback. *arXiv preprint arXiv:2407.01512*, 2024. 2

[7] 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. *IJRR*, 2023. 4

[8] Sammy Christen, Muhammed Kocabas, Emre Aksan, Jemin Hwangbo, Jie Song, and Otmar Hilliges. D-grasp: Physically plausible dynamic grasp synthesis for hand-object interactions. In *CVPR*, 2022. 2

[9] 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 *CVPR*, 2023. 1, 2

[10] 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 *CoRL*, 2022. 4

[11] Zipeng Fu, Qingqing Zhao, Qi Wu, Gordon Wetzstein, and Chelsea Finn. Humanplus: Humanoid shadowing and imitation from humans. *arXiv preprint arXiv:2406.10454*, 2024. 2

[12] Jingwei Huang, Yichao Zhou, and Leonidas Guibas. Manifoldplus: A robust and scalable watertight manifold surface generation method for triangle soups. *arXiv preprint arXiv:2005.11621*, 2020. 4

[13] Hanwen Jiang, Shaowei Liu, Jiashun Wang, and Xiaolong Wang. Hand-object contact consistency reasoning for human grasps generation. In *ICCV*, 2021. 3

[14] 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. 2

[15] Kailin Li, Jingbo Wang, Lixin Yang, Cewu Lu, and Bo Dai. Semgrasp: Semantic grasp generation via language aligned discretization. In *ECCV*, 2024. 3

[16] Kailin Li, Lixin Yang, Zenan Lin, Jian Xu, Xinyu Zhan, Yifei Zhao, Pengxiang Zhu, Wenxiong Kang, Kejian Wu, and Cewu Lu. Favor: Full-body ar-driven virtual object rearrangement guided by instruction text. *AAAI*, 2024. 2

[17] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *CVPR*, 2021. 2

[18] Toru Lin, Zhao-Heng Yin, Haozhi Qi, Pieter Abbeel, and Jitendra Malik. Twisting lids off with two hands. *arXiv preprint arXiv:2403.02338*, 2024. 2

[19] Viktor Makoviychuk, Lukasz Wawrzyniak, Yunrong Guo, Michelle Lu, Kier Storey, Miles Macklin, David Hoeller, Nikita Rudin, Arthur Allshire, Ankur Handa, et al. Isaac gym: High performance gpu-based physics simulation for robot learning. *arXiv preprint arXiv:2108.10470*, 2021. 1, 4

[20] Yuzhe Qin, Wei Yang, Binghao Huang, Karl Van Wyk, Hao Su, Xiaolong Wang, Yu-Wei Chao, and Dieter Fox. Anyteleop: A general vision-based dexterous robot arm-hand teleoperation system. In *RSS*, 2023. 2

[21] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM TOG*, 2017. 2, 3

[22] Nur Muhammad Shafiullah, Zichen Cui, Ariuntuya Arty Altanzaya, and Lerrel Pinto. Behavior transformers: Cloning k modes with one stone. *NeurIPS*, 2022. 4

[23] Omid Taheri, Nima Ghorbani, Michael J Black, and Dimitrios Tzionas. Grab: A dataset of whole-body human grasping of objects. In *ECCV*, 2020. 2

[24] Lixin Yang, Kailin Li, Xinyu Zhan, Jun Lv, Wenqiang Xu, Jiefeng Li, and Cewu Lu. Artiboost: Boosting articulated 3d hand-object pose estimation via online exploration and synthesis. In *CVPR*, 2022. 2

[25] Lixin Yang, Xinyu Zhan, Kailin Li, Wenqiang Xu, Junming Zhang, Jiefeng Li, and Cewu Lu. Learning a contact potential field for modeling the hand-object interaction. *IEEE TPAMI*, 2024. 2, 3

[26] Xinyu Zhan, Lixin Yang, Yifei Zhao, Kangrui Mao, Hanlin Xu, Zenan Lin, Kailin Li, and Cewu Lu. Oakink2: A dataset of bimanual hands-object manipulation in complex task completion. In *CVPR*, 2024. 2, 3

[27] Shuqi Zhao, Xinghao Zhu, Yuxin Chen, Chenran Li, Xiang Zhang, Mingyu Ding, and Masayoshi Tomizuka. Dexh2r: Task-oriented dexterous manipulation from human to robots. *arXiv preprint arXiv:2411.04428*, 2024. 2