Human-Centered Robotic Sock Dressing: Care Interaction Insights from a Public Demonstration at Expo 2025 Osaka

Takuma Tsukakoshi
Waseda University
Tokyo, Japan
takuma36@fuji.waseda.jp

Tamon Miyake
Waseda University
Tokyo, Japan
tamonmiyake@aoni.waseda.jp

Tetsuya Ogata
Waseda University
Tokyo, Japan
ogata@waseda.jp

Yushi Wang
Waseda University
Tokyo, Japan
yushiwang@aoni.waseda.jp

Takumi Akaishi
Waseda University
Tokyo, Japan
akaishi3639@akane.waseda.jp

Shigeki Sugano
Waseda University
Tokyo, Japan
sugano@waseda.jp

Figure 1: A live demonstration at the Osaka Expo highlighting real-time human–robot interaction, as visitors observe and participate in hands-on control of a humanoid robot

Abstract

As the population continues to age, a shortage of caregivers is expected in the future. Dressing assistance is particularly important for enabling social participation. However, assisting with close-fitting garments such as socks remains challenging due to the need for fine force control to handle friction and snagging against the skin while accounting for the shape and position of the garment. This paper proposes a multimodal dressing assistance method that combines vision, proprioception, and tactile sensing with semantic and depth information to achieve adaptive force control and robust generalization to unseen feet and environments, outperforming baseline models. Training was conducted on a mannequin for safety, followed by preliminary validation with human participants. The results indicate that the proposed method can adapt to unseen feet and support stable sock-dressing motions.

CCS Concepts

• **Computing methodologies** → **Cognitive robotics**; *Learning from demonstrations*; Vision for robotics.

Keywords

Physical Human-Robot Interaction, Imitation Learning, Human-Centered Robotics

1 Introduction

In clinical and assistive settings, robots are increasingly envisioned as long-term partners that share intimate spaces and physically interact with people in daily life. These roles raise challenges beyond technical performance, including safety, dignity, personalization, and trust. Dressing assistance highlights these issues: it is a fundamental activity of daily living tied to autonomy and self-respect, yet it involves close contact, deformable garments, and significant

individual variation. As a result, robotic failures in dressing can directly affect both physical safety and user experience.

Prior work in robot-assisted dressing has made significant progress through multimodal sensing and learning-based control[10, 11, 15, 19]. However, much of this research has focused on loose-fitting garments, where tolerance for misalignment is relatively high. In contrast, close-fitting garments such as socks impose stricter requirements on force modulation, geometric adaptation, and continuous perception of garment–body interaction. These constraints highlight a broader concern in assistive robotics: systems designed around fixed task execution often struggle to accommodate bodily diversity and the dynamic, situated nature of care.

We study robotic sock-dressing assistance as a case of adaptive, close-fitting dressing support. Rather than modeling dressing as a fixed action sequence, we frame it as an ongoing interaction requiring adaptation to individual body geometry, garment deformation, and contact dynamics. We propose a multimodal imitation learning approach that integrates semantic perception, depth, and somatosensory cues to enable generalization across unseen users and environments. Training is conducted on a mannequin for safety, with evaluation on human participants to examine how semantic representations and multimodal learning support robust robotic dressing assistance.

2 Related Work

2.1 Dressing assistance manipulation

Numerous studies have explored robot-assisted dressing methods. Some studies address adaptation to user motion under garment-induced occlusions, often incorporating personalization based on individual differences[3, 6, 8, 20]. Other studies focus on garment manipulation prior to dressing, such as grasping and unfolding garments to prepare suitable configurations [23–25]. During dressing, haptic perception has also been used to infer interactions between garments and the human body [9].

Another line of research focuses on garment manipulation during dressing by developing control strategies that guide garments into desired configurations around the human body. These approaches commonly rely on point cloud observations and include demonstration-based learning for arm and body dressing [7], as well as model predictive control with graph convolutional networks or diffusion policies for garment opening insertion [10, 11]. and policies that integrate visual perception with force dynamics for safe dressing assistance [15]. Policy distillation has also been used to generalize dressing policies across different human poses from partial point cloud observations [19]. These studies primarily address loose-fitting garments such as shirts and trousers.

2.2 Imitation learning for deformable object

Imitation learning has recently been applied to deformable object manipulation, including garment folding, unfolding, and smoothing, by leveraging multimodal representations [14, 22, 26]. Vision-language models have further extended these approaches to a wider range of manipulation tasks [2, 4]. While imitation learning has shown the ability to maintain appropriate contact forces in manipulation tasks [1, 13], applying these approaches to close-fitting garment dressing remains challenging due to the need to balance fabric elasticity, sock-foot friction, and directional force control.

3 Dressing with Proposed Model

An overview of the proposed model is shown in Fig. 2. The model consists of multiple components that jointly enable semantic perception, attention-based feature extraction, and temporal modeling for dressing assistance. Multimodal features are integrated over time using a hierarchical LSTM, allowing the model to capture temporal changes in object shape and contact dynamics and to generate appropriate dressing motions.

3.1 Semantic extraction of garment and human foot

To estimate the states of the sock and the human foot, we combine semantic segmentation and monocular depth estimation. We use SAM2[12] to obtain semantic masks of the sock and foot, which reduces sensitivity to variations in appearance, lighting, and background, and helps maintain robustness under frequent occlusions during dressing motions. Depth information is obtained using DAM[21] and integrated with the semantic masks to infer the relative spatial relationship between the garment and the foot.

3.2 Visual and somatosensory attention mechanism

We incorporate somatosensory and visual attention mechanisms to support stable force interaction and robust perception. Somatosensory attention emphasizes relevant tactile cues during contact, while visual attention focuses on the sock and foot regions to handle background changes and occlusions. By integrating depth-aware attention, the system extracts 3D keypoints, enabling reliable perception and control in close-fitting dressing tasks.

3.3 Hierarchical LSTM

The extracted 3D keypoints, joint states, and tactile information are integrated over time using a hierarchical LSTM network to generate dressing motions. Separate LSTMs are applied to individual modalities, and their internal representations are fused by a higher-level LSTM. This hierarchical structure enables effective modeling of both modality-specific dynamics and cross-modal interactions, which is essential for stable force-aware dressing assistance.

3.4 Training and inference phase

During training, the model learns to predict multimodal information, including joint angles, joint torques, tactile feedback, and attention points, from demonstration data. The network is trained by minimizing the prediction error between the estimated multimodal outputs and the corresponding target values.

During inference, the target object region is tracked based on the initial prompt. A semantic mask is generated in real time, and depth information is extracted from the masked region using the depth estimation model. Given the current joint angles, joint torques, tactile feedback, and camera images, the model predicts the next target state.

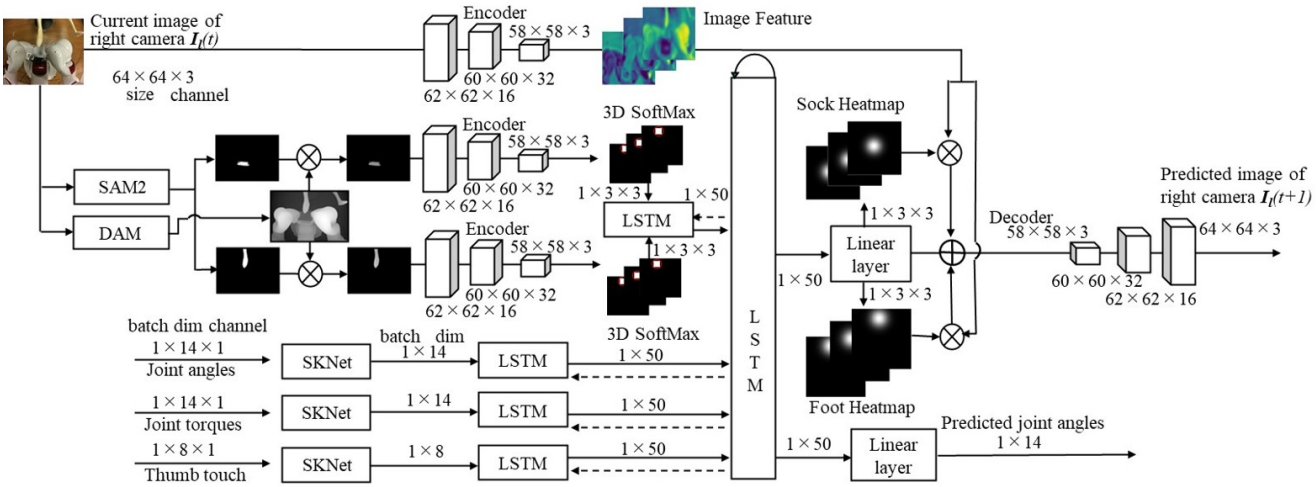

Figure 2: An overview of our proposed model[17]. The motion generation is based on the deep predictive learning model EIPL with Hierarchical LSTM [16]. The features extracted from the semantic mask and depth estimation are input into an attention mechanism called Spatial Softmax to obtain attention points on the images. The visual attention points, joint angles, torques, and tactile information are input into a hierarchical LSTM, and the image and joint angle one step ahead are output.

4 Experiments

4.1 Environment and dataset

Training data were collected via teleoperation using a game controller. For safety reasons, a mannequin seated on a chair was used, with its foot placed on a footrest to simulate a sock-dressing scenario for a person sitting at the bedside. To account for individual differences in foot posture, the mannequin's ankle dorsiflexion angle was set to 30°, 40°, and 50°. Four demonstrations were collected for each condition, resulting in a total of 12 training samples.

The dataset includes camera images, joint angles and torques of both arms, and tactile feedback from the thumbs, recorded at 20 Hz. Using a mannequin reduces physical and psychological burden on human participants while enabling evaluation of generalization to real humans.

4.2 Evaluation

We evaluated generalization and robustness on 10 human participants by comparing the proposed method with Action Chunking with Transformer (ACT) [26] and Diffusion Policy (DP) [5]. Experiments were conducted under two conditions, with backgrounds either matching or differing from those used during demonstration, with 50 trials per condition as shown in Fig. 3. Ethical approval was obtained from the Human Subject Research Ethics Review Committee.

5 Results and Discussion

We evaluated the performance of the proposed method on real human subjects and compared it with Action Chunking with Transformer (ACT) and Diffusion Policy (DP). As shown in Table 2, the proposed method achieved success rates of **84%** in seen environments and **74%** in unseen environments, outperforming ACT, which failed completely in unseen backgrounds. DP was unable to complete the dressing task due to unstable force control. Moreover, as shown in Table 1, foot size caused specific failures. The participants with larger feet experienced heel-catching failures due to premature arm lifting. Consequently, although the system can accommodate a certain range of foot sizes, challenges remain in achieving robust performance across broader morphological variations. We would expand the method by leveraging a morpho-aware method [18].

Table 1: Success rates corresponding to foot sizes

Size of foot	ACT	DP	Ours
23-24 cm	14/30	N/A	**26/30**
24-25 cm	14/30	N/A	**28/30**
25-26 cm	5/30	N/A	**21/30**
26-27 cm	0/10	N/A	**4/10**

Human experiments revealed failure modes not seen in mannequin trials, including limited ankle flexibility and toenail snagging. The model showed a gradual and consistent increase in force across dressing phases and remained stable across different foot sizes. Higher forces in the heel–ankle phase were attributed to deeper arm insertion and wider hand spacing. Overall, the method maintained appropriate force levels, enabling task progress without excessive pressure, though further improvements are needed to handle greater morphological variability.

6 Live Demonstration at the Osaka Expo

From August 5 to August 12, 2025, a live demonstration of sock-donning assistance was conducted at the Osaka Expo, showcasing a humanoid care robot designed to support daily caregiving tasks.

,

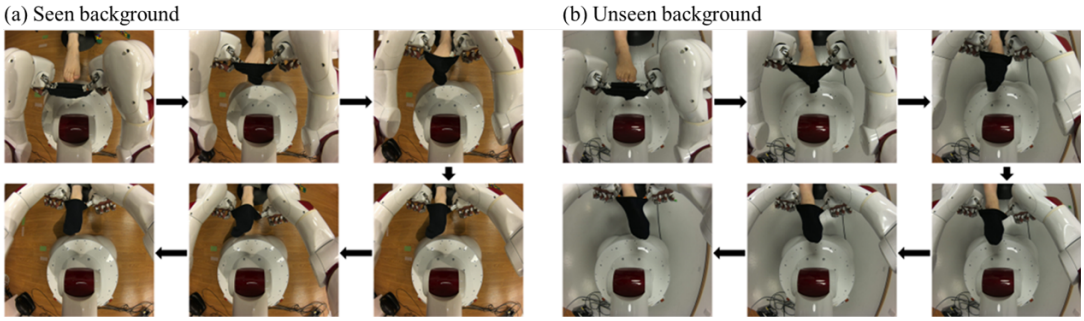

Figure 3: Scenes of the test of the motion generation with the proposed model seen background(left) unseen background(right)

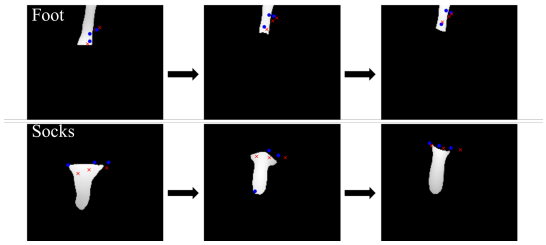

Figure 4: Attention points moving over the object area. Blue points and red points indicate current and future image attention key points, respectively.

Table 2: Success rates of motion generation.

Condition	ACT	DP	Ours
Seen Background	33/50	N/A	**42/50**
Unseen Background	0/50	N/A	**37/50**

While the robot did not assist real visitors directly for safety reasons, it demonstrated the full motion and control sequence required to put socks on a human foot. This decision highlighted the current boundary between experimental capability and real-world deployment.

Audience reactions during the live demonstration were telling. Some visitors commented that the motion looked as though it might cause muscle tension or cramps, raising concerns about comfort and safety. These reactions revealed an important insight: technical success alone is not enough. Speed, smoothness, perceived comfort, and psychological reassurance are as critical as mechanical precision.

AIREC is being developed as part of Japan's Moonshot Research and Development Program, which aims to realize robots capable of coexisting with humans by 2050. Beyond socks, the robot is designed to support walking assistance and wheelchair-related caregiving tasks, with a target of deployment in care facilities by 2040.

This live demonstration at the Osaka Expo positioned sock-donning assistance not as a novelty, but as a benchmark task that exposes the true challenges of care robotics. It underscored both how far the technology has progressed and how much refinement is still required before such robots can safely and comfortably reduce caregiver burden in real caregiving environments.

7 Conclusion and future work

This study presented a multimodal imitation learning approach for close-fitting dressing assistance, focusing on sock-dressing as a representative and socially significant care task. By embedding depth information into semantic mask representations and integrating visual, proprioceptive, and tactile cues through a hierarchical LSTM, the proposed system was able to adapt to variations in foot shape, size, and flexibility. Training on a mannequin enabled safe data collection, while preliminary validation with human participants demonstrated improved robustness compared to baseline methods.

Beyond technical performance, these results underscore the importance of continuous adaptation in assistive robotics. Close physical interaction with human bodies and deformable garments exposes the limitations of fixed or overly simplified task models. While the system successfully handled insertion motions and subsequent dressing phases, limitations in the toe-insertion stage highlight how small failures can carry disproportionate consequences in intimate care contexts. Rather than optimizing only for task completion, such systems must be evaluated in terms of how they manage uncertainty, bodily diversity, and risk.

Future work will extend this approach toward dynamic, long-term adaptation in real-world settings, where unexpected human motion, misalignment, and unintended contact are inevitable. This will require not only online replanning and recovery behaviors, but also careful consideration of how assistance is negotiated between humans and robots over time. The proposed semantic mask representation offers a promising direction in this regard, as it captures object- and body-level concepts that may generalize across different garments and dressing scenarios.

Acknowledgments

This work was partially supported by the JST Moonshot Research and Development Program (JPMJMS2031) and by the Waseda University Next-Generation Robotics Research Organization.

References

[1] Tsuyoshi Adachi, Kazuki Fujimoto, Sho Sakaino, and Toshiaki Tsuji. 2018. Imitation Learning for Object Manipulation Based on Position/Force Information Using Bilateral Control. 3648–3653. doi:10.1109/IROS.2018.8594489

[2] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Lucy Xiaoyang Shi, James Tanner, Quan Vuong, Anna Walling, Haohuan Wang, and Ury Zhilinsky. 2024. Pi-0: A Vision-Language-Action Flow Model for General Robot Control. arXiv:2410.24164 [cs.LG] https://arxiv.org/abs/2410.24164

[3] Gerard Canal, Guillem Alenyà, and Carme Torras. 2019. Adapting robot task planning to user preferences: an assistive shoe dressing example. *Autonomous Robots* 43 (08 2019). doi:10.1007/s10514-018-9737-2

[4] Haonan Chen, Junxiao Li, Ruihai Wu, Yiwei Liu, Yiwen Hou, Zhixuan Xu, Jingxiang Guo, Chongkai Gao, Zhenyu Wei, Shensi Xu, Jiaqi Huang, and Lin Shao. 2025. MetaFold: Language-Guided Multi-Category Garment Folding Framework via Trajectory Generation and Foundation Model. arXiv:2503.08372 [cs.RO] https://arxiv.org/abs/2503.08372

[5] Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. 2024. Diffusion Policy: Visuomotor Policy Learning via Action Diffusion. arXiv:2303.04137 [cs.RO] https://arxiv.org/abs/2303.04137

[6] Aleksandar Jevtić, Andrés Flores Valle, Guillem Alenyà, Greg Chance, Praminda Caleb-Solly, Sanja Dogramadzi, and Carme Torras. 2019. Personalized Robot Assistant for Support in Dressing. *IEEE Transactions on Cognitive and Developmental Systems* 11, 3 (2019), 363–374. doi:10.1109/TCDS.2018.2817283

[7] Ravi Joshi, Nishanth Koganti, and Tomohiro Shibata. 2019. A Framework for Robotic Clothing Assistance by Imitation Learning. *Advanced Robotics* (07 2019), 1–19. doi:10.1080/01691864.2019.1636715

[8] Ariel Kapusta, Zackory Erickson, Henry Clever, Wenhao Yu, Karen Liu, Greg Turk, and Charles Kemp. 2019. Personalized collaborative plans for robot-assisted dressing via optimization and simulation. *Autonomous Robots* 43 (12 2019). doi:10.1007/s10514-019-09865-0

[9] Ariel Kapusta, Wenhao Yu, Tapomayukh Bhattacharjee, C. Karen Liu, Greg Turk, and Charles C. Kemp. 2016. Data-driven haptic perception for robot-assisted dressing. In *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. 451–458. doi:10.1109/ROMAN.2016.7745158

[10] Stelios Kotsovolis and Yiannis Demiris. 2024. Model Predictive Control with Graph Dynamics for Garment Opening Insertion during Robot-Assisted Dressing. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*. 883–890. doi:10.1109/ICRA57147.2024.10611478

[11] Stelios Kotsovolis and Yiannis Demiris. 2025. Garment Diffusion Models for Robot-Assisted Dressing. *IEEE Robotics and Automation Letters* 10, 2 (2025), 1217–1224. doi:10.1109/LRA.2024.3518104

[12] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. 2024. SAM 2: Segment Anything in Images and Videos. arXiv:2408.00714 [cs.CV] https://arxiv.org/abs/2408.00714

[13] Namiko Saito, Takumi Shimizu, Tetsuya Ogata, and Shigeki Sugano. 2022. Utilization of Image/Force/Tactile Sensor Data for Object-Shape-Oriented Manipulation: Wiping Objects With Turning Back Motions and Occlusion. *IEEE Robotics and Automation Letters* 7, 2 (2022), 968–975. doi:10.1109/LRA.2021.3136657

[14] Daniel Seita, Aditya Ganapathi, Ryan Hoque, Minho Hwang, Edward Cen, Ajay Kumar Tanwani, Ashwin Balakrishna, Brijen Thananjeyan, Jeffrey Ichnowski, Nawid Jamali, Katsu Yamane, Soshi Iba, John Canny, and Ken Goldberg. 2020. Deep Imitation Learning of Sequential Fabric Smoothing From an Algorithmic Supervisor. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 9651–9658. doi:10.1109/IROS45743.2020.9341608

[15] Zhanyi Sun, Yufei Wang, David Held, and Zackory Erickson. 2024. Force-Constrained Visual Policy: Safe Robot-Assisted Dressing via Multi-Modal Sensing. *IEEE Robotics and Automation Letters* PP (05 2024), 1–8. doi:10.1109/LRA.2024.3375712

[16] Kanata Suzuki, Hiroshi Ito, Tatsuro Yamada, Kei Kase, and Tetsuya Ogata. 2024. Deep Predictive Learning: Motion Learning Concept inspired by Cognitive Robotics. arXiv:2306.14714 [cs.RO] https://arxiv.org/abs/2306.14714

[17] Takuma Tsukakoshi, Tamon Miyake, Tetsuya Ogata, Yushi Wang, Takumi Akaishi, and Shigeki Sugano. 2026. Close-Fitting Dressing Assistance Based on State Estimation of Feet and Garments With Semantic-Based Visual Attention. *IEEE Robotics and Automation Letters* 11, 4 (2026), 3923–3930. doi:10.1109/LRA.2026.3664535

[18] Ruiqin Wang, Yaqing Song, and Jian Dai. 2021. Reconfigurability of the origami-inspired integrated 8R kinematotropic metamorphic mechanism and its evolved 6R and 4R mechanisms. *Mechanism and Machine Theory* 161 (07 2021), 104245. doi:10.1016/j.mechmachtheory.2021.104245

[19] Yufei Wang, Zhanyi Sun, Zackory Erickson, and David Held. 2023. One Policy to Dress Them All: Learning to Dress People with Diverse Poses and Garments. doi:10.15607/RSS.2023.XIX.008

[20] Kakeru Yamasaki, Takumi Kajiwara, Wataru Fujita, and Tomohiro Shibata. 2023. Realizing an Assist-As-Needed Robotic Dressing Support System through Analysis of Human Movements and Residual Abilities. In *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. 2409–2414. doi:10.1109/RO-MAN57019.2023.10309329

[21] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. 2024. Depth Anything V2. arXiv:2406.09414 [cs.CV] https://arxiv.org/abs/2406.09414

[22] Pin-Chu Yang, Kazuma Sasaki, Kanata Suzuki, Kei Kase, Shigeki Sugano, and Tetsuya Ogata. 2017. Repeatable Folding Task by Humanoid Robot Worker Using Deep Learning. *IEEE Robotics and Automation Letters* 2, 2 (2017), 397–403. doi:10.1109/LRA.2016.2633383

[23] Fan Zhang and Yiannis Demiris. 2020. Learning Grasping Points for Garment Manipulation in Robot-Assisted Dressing. doi:10.1109/ICRA40945.2020.9196994

[24] Fan Zhang and Yiannis Demiris. 2022. Learning Garment Manipulation Policies toward Robot-Assisted Dressing. *Science Robotics* 7, 65 (2022), eabm6010. doi:10.1126/scirobotics.abm6010

[25] Fan Zhang and Yiannis Demiris. 2023. Visual-Tactile Learning of Garment Unfolding for Robot-Assisted Dressing. *IEEE Robotics and Automation Letters* 8, 9 (2023), 5512–5519. doi:10.1109/LRA.2023.3296371

[26] Tony Z. Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. 2023. Learning Fine-Grained Bimanual Manipulation with Low-Cost Hardware. arXiv:2304.13705 [cs.RO] https://arxiv.org/abs/2304.13705