ByteDance | Seed

GR-Dexter Technical Report

ByteDance Seed

Full Author List in Contributions and Acknowledgements

Abstract

Vision-language-action (VLA) models have enabled language-conditioned, long-horizon robot manipulation, but most existing systems are limited to grippers. Scaling VLA policies to bimanual robots with high degree-of-freedom (DoF) dexterous hands remains challenging due to the expanded action space, frequent hand-object occlusions, and the cost of collecting real-robot data. We present GR-Dexter, a holistic hardware-model-data framework for VLA-based generalist manipulation on a bimanual dexterous-hand robot. Our approach combines the design of a compact 21-DoF robotic hand, an intuitive bimanual teleoperation system for real-robot data collection, and a training recipe that leverages teleoperated robot trajectories together with large-scale vision-language and carefully curated cross-embodiment datasets. Across real-world evaluations spanning long-horizon everyday manipulation and generalizable pick-and-place, GR-Dexter achieves strong in-domain performance and improved robustness to unseen objects and unseen instructions. We hope GR-Dexter serves as a practical step toward generalist dexterous-hand robotic manipulation.

Date: December 30, 2025
Correspondence: wenruoshi@bytedance.com
Project Page: https://byte-dexter.github.io/gr-dexter

1 Introduction

Generalist manipulation policies powered by vision-language-action (VLA) models have enabled language-conditioned control and long-horizon instruction following in robot manipulation [12, 28, 45, 60]. However, most existing policies are deployed on bimanual robots with gripper-based end effectors. Extending these capabilities to robots equipped with dexterous hands remains underexplored. As robots move toward general-purpose operation in cluttered, human-centered environments, dexterous hands hold greater potential for achieving human-level manipulation. Yet this promise comes with substantial challenges: high degree-of-freedom (DoF) hands expand the control space by dozens of DoFs, while introducing perception difficulties—frequent occlusions between fingers and between the hand and target objects. Moreover, as a data-driven paradigm, VLA performance depends critically on the quality and diversity of robot trajectories for dexterous bimanual manipulation.

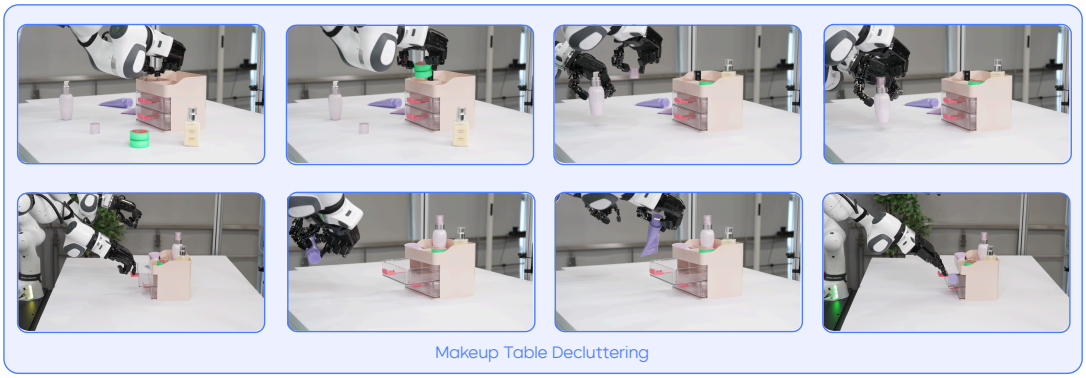
We present the ByteDexter V2 hand—a 21-DoF linkage-driven anthropomorphic robotic hand, designed as a self-contained, modular end-effector for dexterous manipulation, with space, complexity, and maintainability as key design constraints. Compared with ByteDexter V1 [61] and ILDA hand [29], V2 adds an additional thumb DoF while further reducing overall size. With actuators integrated within the palm, the hand achieves a compact form factor (219 mm height, 108 mm width). It also incorporates high-density piezoresistive tactile sensors at the fingertips.

Figure 1 GR-Dexter performs dexterous long-horizon daily tasks and generalizes to out-of-domain settings by learning from four data sources: vision–language, cross-embodiment, human-trajectory, and robot-trajectory data.

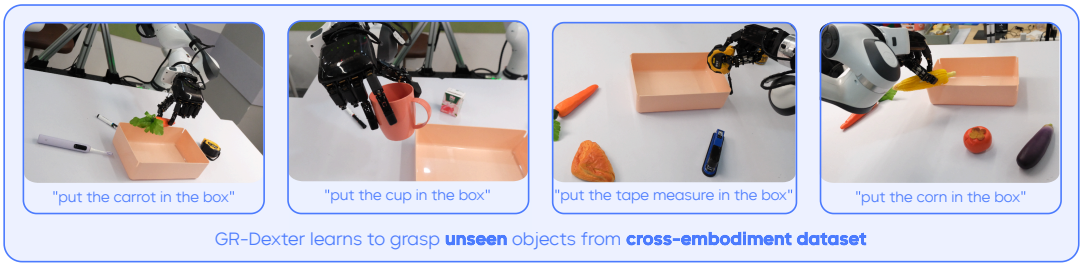
We then introduce a VLA model GR-Dexter and training recipe tailored to a **56-DoF** bimanual system equipped with ByteDexter V2 hands. The policy is built on a pre-trained VLM [4], and is co-trained on a mixture of data sources, including teleoperated robot trajectories, vision-language data, cross-embodiment demonstrations, and human trajectories. Because bimanual arm teleoperation is challenging even with simple grippers, efficient demonstration collection becomes more difficult when each end-effector is a 21-DoF dexterous hand. We address this challenge with a teleoperation interface comprised of a Meta Quest headset and Manus gloves, which retarget tracked human wrist poses and hand motions to joint position commands in real time.

Beyond collecting teleoperated robot trajectories, anthropomorphic high-DoF hands also provide a promising data-scaling path: the structural similarity between human and robot hands makes it feasible to directly leverage large-scale egocentric hand-object interaction datasets that cover diverse everyday dexterous behaviors [5, 21, 22, 24]. Our teleoperation pipeline then enables efficient collection of a small amount of on-robot data for fine-tuning the pretrained models to adapt to the target platform.

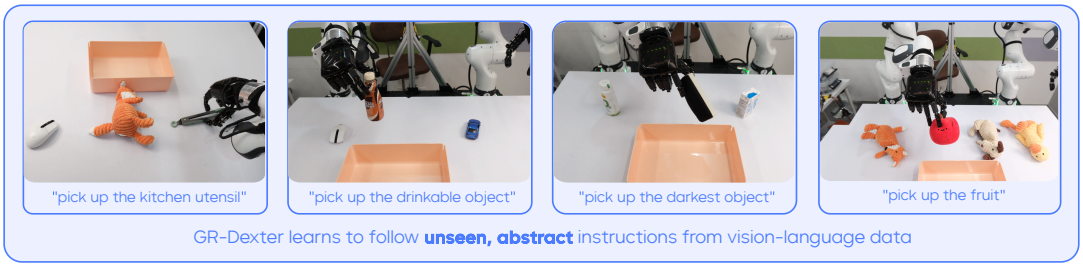
We evaluate GR-Dexter in real-world experiments across two task categories: (1) long-horizon manipulation, and (2) generalizable pick-and-place. Results show strong performance in both in-domain settings and challenging unseen scenarios, including novel objects and previously unseen language instructions (Fig. 2). This performance stems from co-training with large-scale vision-language data, cross-embodiment data, and human trajectories, which preserves robust grasping behaviors on in-domain sub-tasks while improving generalization to out-of-distribution (OOD) cases. Moreover, GR-Dexter successfully completes long-horizon everyday tasks, highlighting its practical bimanual dexterity in the real world.



(a) GR-Dexter performs long-horizon tasks.



(b) GR-Dexter is capable of grasping unseen objects.



(c) GR-Dexter follows unseen language instructions.

Figure 2 Capabilities. GR-Dexter robustly completes long-horizon daily tasks. It also learns to grasp unseen objects, and follow unseen, abstract language instructions.

(a) ByteDexter V2 DoF distribution and tactile sensors. **(b)** ByteDexter V2 scores 10 in the Kapandji test

Figure 3 The ByteDexter V2 hand. We show the DoF distribution, tactile fingertips, and the thumb's opposition capability.

2 ByteDexter Robotic Hand

The ByteDexter hand series employ a linkage-driven transmission mechanism for its advantages in force transparency, durability, and ease of maintenance. As an upgraded successor to the V1 hand [61], the **ByteDexter V2 hand** introduces an additional thumb DoF, bringing the total to 21 DoFs, while simultaneously reducing the overall hand size (height: 219mm, width: 108mm). Each finger has four DoFs, and the thumb has five, providing a wider range of oppositional motions, illustrated in Fig. 2. We also demonstrate its human-like grasping capability by executing all 33 Feix grasp types [18] (Appendix Fig. 9).

2.1 Hand Design

Fingers (index, middle, ring, little). The four fingers share a modular architecture. Each finger comprises a universal joint at the MCP (metacarpophalangeal) and two revolute joints at the PIP (proximal interphalangeal) and DIP (distal interphalangeal). The two DoFs at the MCP are actuated by two motors housed in the palm, enabling abduction–adduction and flexion–extension. Unlike the ILDA hand [29], ByteDexter V2 decouples PIP flexion from MCP flexion, so that the PIP is independently actuated by a dedicated third motor.

Thumb. In the human hand, the saddle-shaped carpometacarpal (CMC) joint enables flexion–extension and abduction–adduction, which are critical for dexterous in-hand manipulation. ByteDexter V2 employs a universal joint at the CMC together with an additional revolute joint to approximate these kinematics and preserve key functional characteristics (Fig. 3a). The compact, integrated thumb mechanism minimizes internal volume while substantially increasing the thumb's range of motion. The resulting enlarged reachable workspace enables robust oppositional contact with all four fingers (Fig. 3b).

Underactuation. The DIP joints of four fingers and the IP (interphalangeal) joint of the thumb are underactuated. ByteDexter V2 implements a biomimetic four-bar linkage mechanism that couples each DIP to its corresponding PIP, reproducing the intrinsic kinematic coupling observed in the human DIP–PIP joint complex.

Tactile Sensing. The five fingertips of ByteDexter V2 are covered with high-density piezoresistive tactile arrays that measure normal contact forces (Fig. 3a). The visualization encodes contact location and force magnitude, and the arrays provide fine spatial resolution over the fingertip, finger pad, and lateral surface.

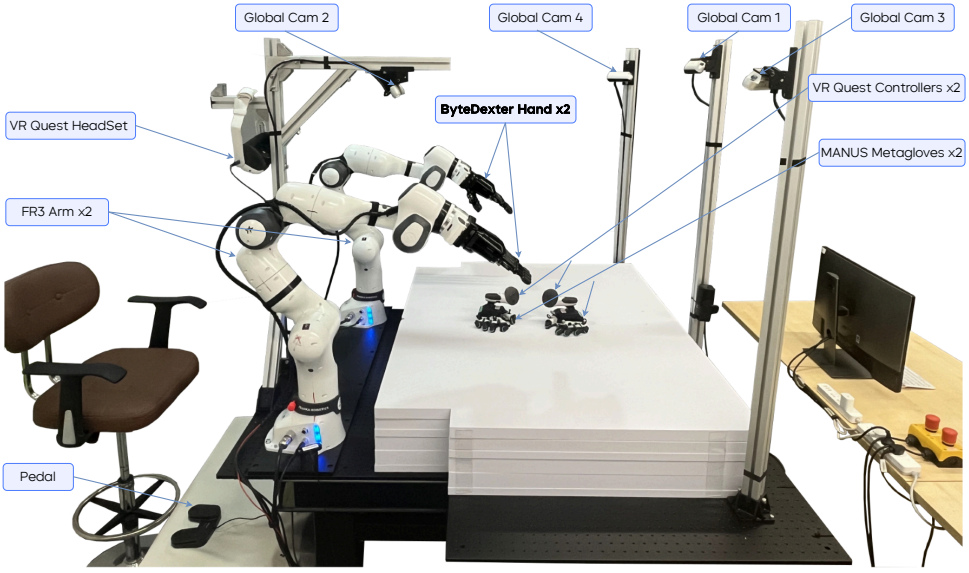


Figure 4 The bimanual robotic system comprising two Franka Research 3 arms equipped with ByteDexter V2 hands. Data are collected via a teleoperation interface using a Meta Quest VR headset, Manus gloves with mounted VR tracking controllers, and a set of global RGB-D cameras.

2.2 Bimanual System and Control

We built a dual-arm platform equipped with two ByteDexter V2 hands for bimanual manipulation (Fig. 4). The resulting 56-DoF robot is designed to support coordinated arm-hand control for reliable dexterous grasping and manipulation. To mitigate occlusions and capture hand-object interactions from multiple views, we deploy four global RGB-D cameras: one primary egocentric view and three complementary third-person views. The platform supports both teleoperated data collection and autonomous policy rollouts.

Bimanual teleoperation. We collect real-world robot data using a bimanual teleoperation interface consisting of a Meta Quest VR setup for tracking wrist poses, two Manus Metagloves for capturing hand movements, and foot pedals to enable/disable teleoperation. Two Meta Quest controllers are mounted on the dorsal side of the gloves to improve the reliability of coordinated wrist–hand tracking. This setup allows teleoperators to simultaneously coordinate two Franka arms together with two ByteDexter V2 hands during long-horizon manipulation tasks. Human motions are retargeted in real time to joint position commands via a whole-body controller, providing a kinematically consistent mapping. The system incorporates safety mechanisms to handle intermittent visual tracking loss and mitigate hazardous operation. Hand-motion retargeting is formulated as a constrained optimization problem that combines wrist-to-fingertip and thumb-to-fingertip alignment terms with collision-avoidance constraints and regularization, and is solved using Sequential Quadratic Programming.

Policy rollout. During policy rollout, our model generates future action chunks that promote coordinated, temporally consistent arm–hand motions for dexterous manipulation. The parameterized trajectory optimizer smooths the generated actions, which is critical for delicate grasping, and ensures smooth transitions both within and across chunks.

The bimanual system demonstrates efficiency, human-like dexterity, and reliable long-duration operation. After minimal training, teleoperators successfully completed tasks ranging from coarse manipulation (e.g., building blocks) to fine motor tasks (e.g., knitting), as shown in Fig. 5. The breadth of tasks highlights the system's suitability for real-world bimanual manipulation, enabling reliable data collection and policy evaluation.

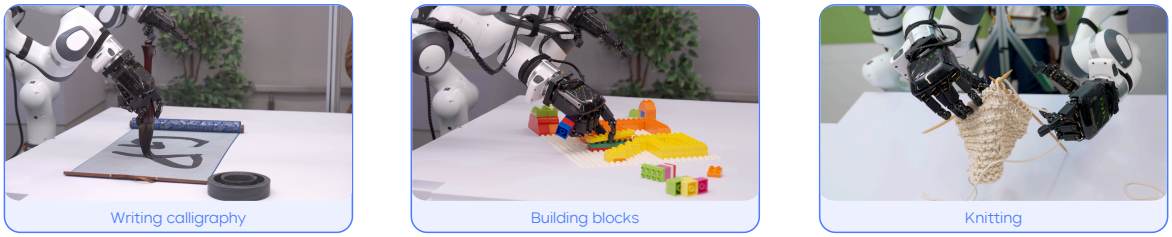


Figure 5 Teleoperation capability in long-horizon dexterous grasping and bimanual manipulation tasks.

3 The GR-Dexter Model

GR-Dexter follows GR-3 [12] and adopts a Mixture-of-Transformer architecture for a vision-language-action (VLA) model π_θ of 4B parameters. $\pi_\theta(\mathbf{a}_t \mid l, \mathbf{o}_t, \mathbf{s}_t)$ controls a bi-manual robot with fixed base by generating a k-length action chunk $\mathbf{a}_t = a_{t:t+k}$ conditioned on the input language instruction l, observation $\mathbf{o}_t$, and robot state $\mathbf{s}_t$. Specifically, different from GR-3 which learns binary discrete gripper actions, each action a_t is a vector of length 88, consisting of: 1) arm joint actions (7 DoF per arm), 2) arm end-effector poses (6D per arm), 3) hand joint actions (16 active DoFs per hand), and 4) fingertip positions (3D per finger).

3.1 Training Recipe

We train GR-Dexter using a mixture of three distinct data sources: web-scale vision-language data, cross-embodiment real-robot data, and human trajectory data.

Figure 6 Data Pyramid of GR-Dexter.

Vision-language data We reuse the VLM dataset from GR-3, which covers a wide spectrum of tasks including image captioning, visual question answering, image grounding, and interleaved grounded image captioning. The robot trajectory data are used to train both the VLM backbone and the action DiT using the flow-matching objective. The vision-language data are used to train only the VLM backbone via the next-token-prediction objective. For simplicity, we dynamically mix vision-language data with robot trajectories across mini-batches. As a result, the co-training objective is the sum of the next-token-prediction loss and the flow-matching loss.

Cross-embodiment data Collecting large-scale teleoperation data on our high-DoF Byte-Dexter platform is constrained by hardware availability and the scarcity of skilled teleoperators. To mitigate this, we leverage existing open-source bi-manual humanoid datasets. Specifically, we select three dual-arm dexterous manipulation datasets that encompass diverse embodiments and task settings: Fourier ActionNet Dataset [20], which contains around 140 hours of diverse humanoid bimanual manipulation data using Fourier 6-DoF hands; OpenLoong Baihu Dataset [57], which features over 100k robot trajectory data across multiple robot embodiment; RoboMIND [62], which includes 107k demonstration trajectories across 479 diverse tasks involving 96 object classes.

Human trajectories While cross-embodiment robot data offers accurate robot state information, the scale and diversity of tasks are inevitably limited by hardware costs. Crowdsourcing human demonstrations via easily accessible VR devices offers a promising solution to scale up data quantity and diversity. We adopt the human trajectory data (over 800 hours of egocentric video with paired 3D hand and finger tracking data) and supplement it with additional data collected using Pico VR devices.

To handle the structural differences across datasets, we mask out unavailable or unreliable action dimensions (e.g., specific joints not present in the target embodiment).

3.2 Cross-Embodiment Motion Retargeting and Transferring

Transferring dexterous manipulation skills across heterogeneous embodiments and from human demonstrations requires careful normalization of both perception and action spaces. We address this challenge with a unified preprocessing and retargeting pipeline that aligns visual geometry, kinematics, and trajectory quality across all data sources.

Transferring cross-embodiment trajectories We first standardize camera observations across datasets. All images are resized and cropped to a standardized format where robot arms, dexterous hands, and object sizes at a similar scale. Such a process can be easily achieved manually for each dataset once and applied to all. Trajectories then undergo strict quality control and only high-quality trajectories are maintained. We then perform careful retargeting to ByteDexter V2 hand by aligning the fingertips. This fingertip-centric alignment preserves task-relevant contact geometry while remaining agnostic to joint-level discrepancies. The resulting trajectories are then resampled by task category to produce a balanced cross-embodiment training corpus.

Transferring human trajectories Human demonstrations pose additional challenges beyond cross-robot transfer. The kinematic gap between human and robotic hands is substantial: VR data collection introduces ego-motion due to head-mounted cameras, and single-frame hand pose estimation commonly leads to temporal jitter and inconsistency—especially during rapid motion or partial occlusion. We first perform careful filtering based on hand visibility and velocity. Next, human trajectories are mapped into the same visual and kinematic representation as robot data similar to the cross-embodiment data cleaning process, enabling seamless integration into the GR-Dexter training pipeline.

4 Experiments

We conduct extensive real-world experiments to evaluate the performance of GR-Dexter on long-horizon bimanual manipulation and generalizable pick-and-place tasks. We evaluate GR-Dexter's capabilities in: (1) long-horizon task execution, (2) generalization to OOD scenarios featuring novel relative spatial configurations, unseen objects, and unseen instructions, and (3) leveraging cross-embodiment data for learning.

4.1 Long-Horizon Manipulation Tasks

We study long-horizon instruction following across tasks of increasing difficulty. We first verify that a plain behavior-cloned VLA, trained solely on human teleoperated robot trajectories, can reliably execute long-horizon tasks (Fig. 2a). In the vacuum task, the robot learns a stable four-finger grasp to hold the tabletop vacuum while using the thumb to press the power button (on/off) and press again to increase power, then sweeps to clear confetti. In the bread-serving task, the robot learns a stable tong grasp to retrieve a croissant from a pastry container while the other hand holds a plate, then releases the tongs and places the croissant onto the plate with precise, compliant manipulation.

These results suggest that a plain-VLA recipe is sufficient to learn multi-step task execution. Despite strong in-domain long-horizon performance, real-world deployments often involve unseen object layouts and spatial variations. We therefore study generalization under layout shifts using a more challenging long-horizon decluttering task involving makeup items of diverse shapes and sizes. The task requires coordinated bimanual manipulation and fine-grained skills. For this task, we collected approximately 20 hours of teleoperated robot trajectories and train a plain-VLA baseline using only robot trajectories. To improve generalization, we further train GR-Dexter with a co-training recipe that combines teleoperated robot trajectories with vision-language data.

Settings. We evaluate each model's instruction-following capability on the makeup decluttering task. During policy rollout, the robot is sequentially prompted with natural-language subtask descriptions (six items; one instruction per item) until the task is completed. Each subtask execution starts from the robot's home pose. We report task performance using the average **success rate** across multiple evaluation trials. We consider two evaluation settings:

- **Basic (in-domain):** objects' relative spatial configurations (layouts) are present in the training data.

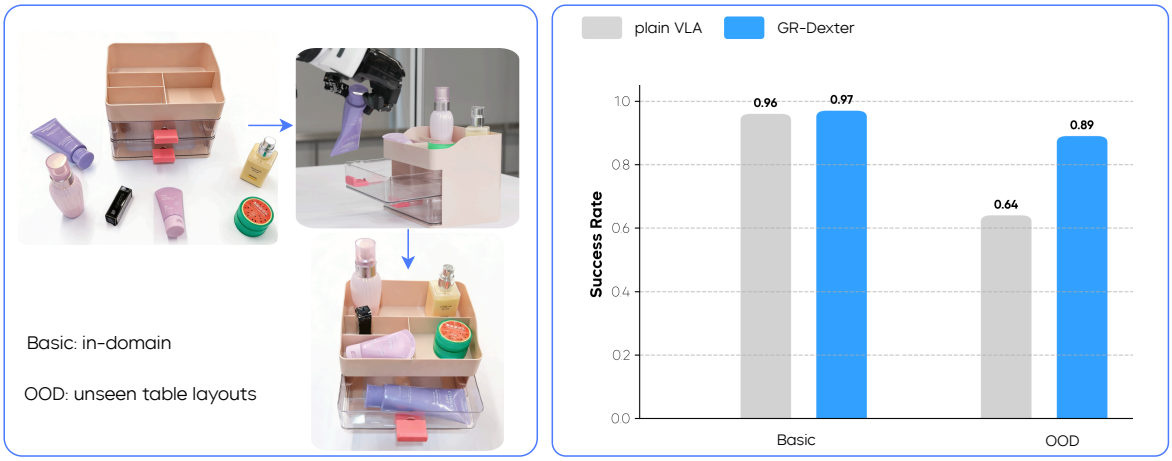


Figure 7 Experiment Settings and Results of Makeup Decluttering.

- **OOD-Layout:** objects' relative spatial configurations are novel at test time. We evaluate on five unseen layouts while keeping the instruction order the same as Basic.

Results. Fig. 7 summarizes success rates for plain-VLA and GR-Dexter under both Basic and OOD-Layout. In the Basic setting, plain-VLA achieves a success rate of 0.96, while GR-Dexter achieves 0.97, showing that co-training preserves the strong in-domain capability of the teleop-only baseline. In the more challenging OOD-Layout setting, plain-VLA drops to 0.64, whereas GR-Dexter improves substantially to 0.89. These results indicate that co-training with vision-language data significantly enhances generalization to unseen spatial layouts, while maintaining in-domain performance.

4.2 Generalizable Pick-and-Place

We evaluate GR-Dexter's generalization on a pick-and-place task. We collected approximately 20 hours of robot trajectories with 20 objects for training (Fig. 8). We compare three models: plain VLA, GR-Dexter without cross-embodiment data, and GR-Dexter.

Settings. We evaluate performance using task success rate. During policy rollout, the model is prompted with a natural-language instruction specifying a target object. A trial is considered successful if the robot picks up the target object and places it into the container. For each evaluation batch, we keep the object layout fixed across rollouts for all policies. We evaluate each model under three settings:

- **Basic:** We construct 10 evaluation batches using seen objects, with five objects per batch.
- **Unseen Objects:** We select 23 unseen objects and construct 10 evaluation batches, with five objects per batch.
- **Unseen Instructions:** We construct 5 evaluation batches using seen objects, and prompt the model with unseen language instructions.

Results. Fig. 8 reports the task success rate of the three models across the three evaluation settings. On the in-domain Basic setting, all models achieve high success rates: plain VLA reaches 0.87, GR-Dexter (without cross-embodiment data) reaches 0.85, and GR-Dexter achieves the best performance at 0.93. Performance diverges substantially under OOD settings. On Unseen Objects, plain VLA drops to 0.45, while cotraining on vision-language data improves the success rate to 0.75 and GR-Dexter further increases success to 0.85, indicating stronger robustness to novel object instances. Similarly, under language variation in Unseen Instructions, plain VLA attains 0.53 whereas GR-Dexter achieves 0.83, demonstrating markedly improved reliability to unseen, abstract instructions. These gains are consistent with the qualitative examples in Fig. 2b, where GR-Dexter successfully grasps unseen objects by leveraging skills learned from cross-embodiment data,

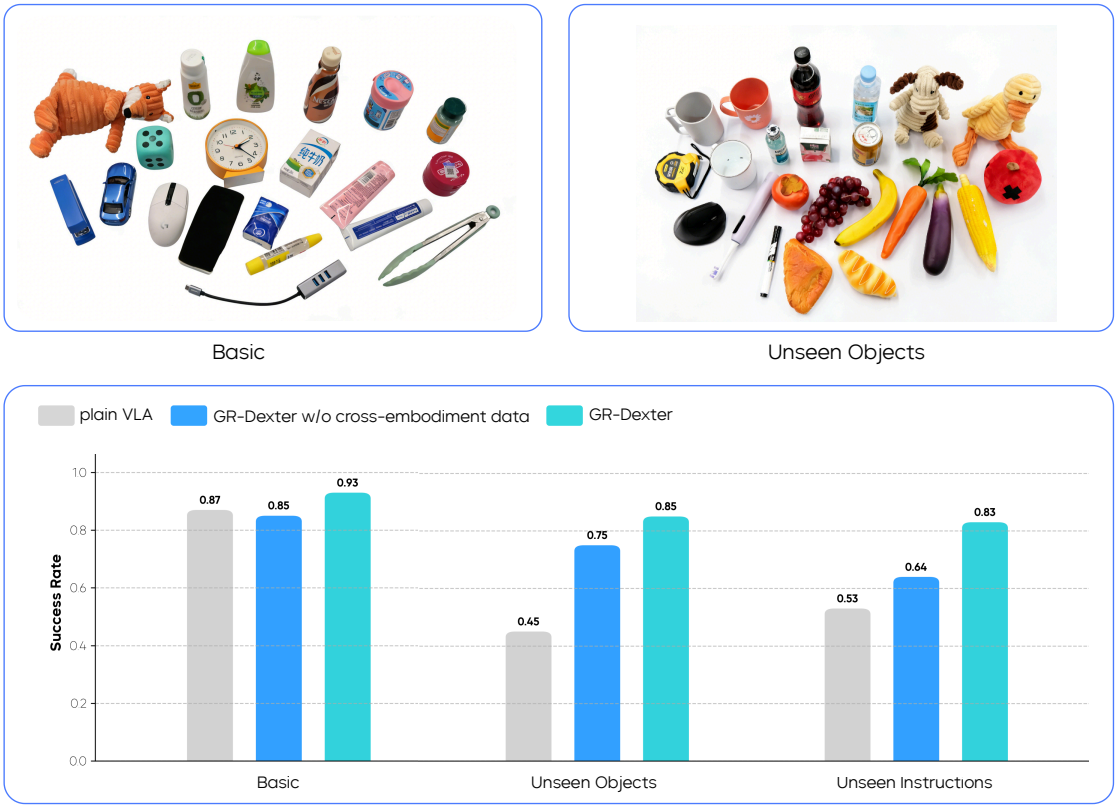


Figure 8 Experiment settings and Results of Generalizable Pick-and-Place.

and Fig. 2c, where it correctly interprets and executes previously unseen instructions. Overall, while all models perform well in-domain, GR-Dexter consistently yields the highest success rates and shows the strongest generalization under both unseen objects and unseen instructions.

5 Related Works

5.1 Dexterous Robotic Hands

Recent years have witnessed rapid progress in multi-fingered dexterous robotic hands, particularly within the robotic manufacturing sectors [16, 35, 52, 54, 63]. Over 10 industrial and startup players have commercialized such hands, including Unitree [58], AgiBot [1], and Fourier [19]. Most commercial designs adopt a relatively small number of active DoFs (typically around 6), while a smaller subset targets 12-DoF configurations; only a few exceed 12 active DoFs. The SharpaWave hand is among the most highly actuated commercial systems, adopting a motor direct-drive architecture with 22 fully actuated DoFs, making it one of the most integrated dexterous robotic hands to date [51]. A representative tendon-driven platform is the Shadow Hand [49]; more recently, Dexcel Robotics released the Apex Hand [17], which provides 21 DoFs (16 independently actuated) and dense, fully covered tactile sensing across the fingertips, phalanges, and palm. A third transmission paradigm is linkage-driven actuation, exemplified by the ILDA hand [29] and the ByteDexter V1 hand [61]; both provide 20 DoFs, 15 of which are independently actuated. Compared with tendon-driven designs, linkage-driven hands can improve durability and force transparency, simplify maintenance, and enable compact actuator integration within the palm—allowing the hand to serve as a self-contained, modular unit without external actuation components.

Building upon the ByteDexter V1 design, this work presents an upgraded version that increases the total

DoFs by one while achieving a more compact form factor. The new design further incorporates high-density piezoresistive tactile sensors covering the fingertips, enhancing its suitability for dexterous manipulation and fine contact-rich tasks.

5.2 VLA Models for Dexterous Hand Manipulation

Vision–language–action (VLA) models have emerged as foundation policies for generalist manipulation, demonstrating strong instruction-following and long-horizon capabilities [6, 8, 10, 12, 25–27, 34, 36, 41, 46, 56]. Although these models are advancing rapidly, their application to dexterous, multi-fingered hands is limited. Integrating anthropomorphic hands into bimanual manipulation substantially increases control dimensionality—often by several dozen DoFs relative to gripper-based setups—thereby raising the demands on both modeling and data. VLA performance depends critically on the diversity and quality of demonstration trajectories; however, large-scale teleoperated dexterous-hand datasets remain scarce.

Recent work suggests that pretraining on human videos can partially mitigate this bottleneck by transferring dexterous manipulation priors to robot policies [2, 3, 23, 24, 31, 39, 42, 43, 47, 48, 53, 55, 59]. GR00T N1 follows this paradigm for humanoid robots. It combines pre-training and post-training on heterogeneous data sources, including real-robot, synthetic, and human video datasets, and has demonstrated effectiveness on Fourier GR-1 humanoids equipped with 6-DoF dexterous hands [45]. Hierarchical approaches also decouple planning from control by using a pretrained vision–language model for task planning [30, 33, 60]. Low-level execution is implemented either with a diffusion transformer (DiT) that outputs action chunks conditioned on grasp instructions and bounding boxes [7, 13, 32, 64, 66] or with RL-based controllers that track the generated trajectories [15, 37, 38, 40]. Our work is, to our knowledge, the first to co-train VLA models for high-DoF dexterous hand manipulation, using a mixture of teleoperated robot trajectories, vision-language data, and cross-embodiment data (human and robot), enabling human-level dexterity in long-horizon tasks.

5.3 Bimanual Dexterous Manipulation Dataset

Most existing dexterous manipulation datasets focus on single-hand grasping [9, 11, 18, 65]. They often emphasize static grasps on isolated objects and typically lack (i) language supervision and (ii) whole-body arm–hand trajectories, making them less suitable for bimanual manipulation that requires coordinated dual-arm control and long-horizon task execution. With the recent rise of VLA models as foundational robot manipulation policies, open-source datasets have grown rapidly, broadly following two paradigms:

Teleoperated Robot Data Operators control humanoid robots using interfaces such as VR controllers or data gloves to perform predefined tasks (e.g., RoboMIND [62], OpenLoong Baihu [57]). This paradigm provides high-fidelity joint states/actions, synchronized visual observations, and typically high-quality language instructions. However, task and environment diversity is often limited by hardware cost and operational complexity. In addition, unlike grippers, dexterous hands vary substantially across platforms, and the resulting kinematic discrepancies make cross-embodiment transfer more challenging.

Human Trajectory Data Egocentric recordings collected via VR devices or wearable cameras [5, 14, 21, 44, 50] scale well and cover diverse scenes and tasks. The main drawbacks are the large embodiment gap between human and robot hands and noisy state estimation, which complicate learning and retargeting to robot control.

In this technical report, we address these challenges by constructing a unified dataset to support co-training of GR-Dexter. We combine curated subsets from open-source bimanual dexterous manipulation datasets with proprietary robot teleoperation data and human demonstrations. Using a standardized pipeline for data cleaning, retargeting, and post-processing, we produce a dataset that balances the precision of robot trajectories with the semantic and environmental diversity of human demonstrations.

6 Limitations & Conclusions

Limitations and future work Our current system has several limitations that suggest clear directions for future work: (1) on the human side, we leverage only a few hundred hours of human trajectories, leaving substantial

complementary egocentric human data untapped; and (2) the robot's hand and arm are controlled separately, which can hinder tight hand–arm coordination in contact-rich dexterous behaviors. Going forward, it is crucial to further improve the pre-training scale exploiting diverse and more accessible cross-embodiment trajectories, and building embodiment-agnostic control abstractions.

Conclusions We introduce GR-Dexter, an integrated hardware–model–data approach that advances VLA-based generalist manipulation to a high-DoF bimanual dexterous-hand robot. On the hardware side, we present ByteDexter V2, a compact anthropomorphic hand designed for dexterous manipulation, and on the data side we develop an intuitive bimanual teleoperation pipeline that makes collecting high-quality demonstrations feasible at this dimensionality. Building on these components, GR-Dexter co-trains a bimanual VLA policy for a 56-DoF dual-arm platform using teleoperated robot trajectories together with vision–language and carefully curated cross-embodiment demonstrations. In real-world evaluations on long-horizon everyday manipulation and generalizable pick-and-place, GR-Dexter achieves strong in-domain performance and improved robustness to unseen objects and unseen instructions. These results suggest that combining practical dexterous hardware with scalable data collection and cross-embodiment supervision is a promising path toward generalist dexterous-hand manipulation.

7 Contributions and Acknowledgements

Hardware Jiajun Zhang, Zhigang Han, Guangzeng Chen, Zhongren Cui, Min Du, Hao Niu, Yang Gou, Zeyu Ren, Wenlei Liu, Mingyu Lei, Liwei Zheng

Control and System Liqun Huang, Ruoshi Wen, Zhongren Cui, Zhengming Zhu, Zeyu Ren, Zhuohang Li, Haoxiang Zhang

Model and Data Haixin Shi, Wei Xu, Ruoshi Wen, Liqun Huang, Weiheng Zhong, Yutao Ouyang, Zhuohang Li, Yunfei Li, Yifei Zhou, Yuxiao Liu, Xiao Ma

Supervisor Hang Li

We thank Jinming Guo, Zetian Li, and Degong Yang for their help on data curation and system maintenance. We are sincerely grateful to all teleoperators and annotators for their dedicated efforts on data collection and annotation.

This work is presented for research purposes only. The technology described in this paper will not be incorporated into any ByteDance product.

References

[1] AgiBot. https://www.agibot.com/products/OmniHand_O12.

[2] Sridhar Pandian Arunachalam, Sneha Silwal, Ben Evans, and Lerrel Pinto. Dexterous imitation made easy: A learning-based framework for efficient dexterous manipulation. In IEEE International Conference on Robotics and Automation (ICRA), 2023.

[3] Shikhar Bahl, Abhinav Gupta, and Deepak Pathak. Human-to-robot imitation in the wild. In Robotics: Science and Systems (RSS), 2022. Often referred to as WHIRL.

[4] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. URL https://arxiv.org/abs/2502.13923.

[5] Prithviraj Banerjee, Sindi Shkodrani, Pierre Moulon, Shreyas Hampali, Shangchen Han, Fan Zhang, Linguang Zhang, Jade Fountain, Edward Miller, Selen Basol, Richard Newcombe, Robert Wang, Jakob Julian Engel, and Tomas Hodan. HOT3D: Hand and object tracking in 3D from egocentric multi-view videos. CVPR, 2025.

[6] Suneel Belkhale, Tianwei Meng, and Dorsa Sadigh. Hip: Hierarchical policy learning from foundation models for long-horizon robot manipulation. arXiv preprint arXiv:2403.03967, 2024.

[7] Homanga Bharadhwaj, Jay Vakil, Mohit Sharma, Abhinav Gupta, Shubham Tulsiani, and Vikash Kumar. Roboagent: Generalization and efficiency in robot manipulation via semantic augmentations and action chunking. In 2024 IEEE International Conference on Robotics and Automation (ICRA), pages 4788–4795. IEEE, 2024.

[8] Konstantinos Bousmalis, Giulia Vezzani, Dushyant Rao, and Sergio Gómez and4 others Colmenarejo. Robocat: A self-improving foundation agent for robotic manipulation. Transactions on Machine Learning Research, 2023. URL https://openreview.net/forum?id=Gz6N9Q36QW.

[9] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. arXiv preprint arXiv:2212.06817, 2022.

[10] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. arXiv preprint arXiv:2307.15818, 2023.

[11] 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.

[12] Chilam Cheang, Sijin Chen, Zhongren Cui, Yingdong Hu, Liqun Huang, Tao Kong, Hang Li, Yifeng Li, Yuxiao Liu, Xiao Ma, Hao Niu, Wenxuan Ou, Wanli Peng, Zeyu Ren, Haixin Shi, Jiawen Tian, Hongtao Wu, Xin Xiao, Yuyang Xiao, Jiafeng Xu, and Yichu Yang. Gr-3 technical report, 2025. URL https://arxiv.org/abs/2507.15493.

[13] 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, 2024.

[14] Yu Cui, Yujian Zhang, Lina Tao, Yang Li, Xinyu Yi, and Zhibin Li. End-to-end dexterous arm-hand vla policies via shared autonomy: Vr teleoperation augmented by autonomous hand vla policy for efficient data collection. arXiv preprint arXiv:2511.00139, 2025.

[15] Vincent de Bakker, Joey Hejna, Tyler Ga Wei Lum, Onur Celik, Aleksandar Taranovic, Denis Blessing, Gerhard Neumann, Jeannette Bohg, and Dorsa Sadigh. Scaffolding dexterous manipulation with vision-language models, 2025. URL https://arxiv.org/abs/2506.19212.

[16] Raphael Deimel and Oliver Brock. A novel type of compliant and underactuated robotic hand for dexterous grasping. The International Journal of Robotics Research, 35(1-3):161–185, 2016.

[17] Dexcel Robotics. https://www.dexcelbot.com/.

[18] Thomas Feix, Javier Romero, Heinz-Bodo Schmiedmayer, Aaron M. Dollar, and Danica Kragic. The grasp taxonomy of human grasp types. IEEE Transactions on Human-Machine Systems, 46(1):66–77, 2016. doi: 10.1109/THMS.2015.2470657.

[19] Fourier. https://www.fftai.com/products-fdh6.

[20] Yao Mu Fourier ActionNet Team. Actionnet: A dataset for dexterous bimanual manipulation, 2025. URL https://action-net.org/.

[21] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video, 2022.

[22] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 19383–19400, 2024.

[23] Haodong He, Wei Wei, Xinyu Li, et al. Dest: Deep evolving spatial-temporal graph for scalable human-to-robot hand motion retargeting. IEEE Robotics and Automation Letters, 2024.

[24] Ryan Hoque, Peide Huang, David J. Yoon, Mouli Sivapurapu, and Jian Zhang. Egodex: Learning dexterous manipulation from large-scale egocentric video, 2025. URL https://arxiv.org/abs/2505.11709.

[25] Wenlong Huang, Chen Wang, Ruohan Zhang, Yunzhu Li, Jiajun Wu, and Li Fei-Fei. Voxposer: Composable 3d value maps for robotic manipulation with language models. In Conference on Robot Learning (CoRL), 2023.

[26] Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Manuel Y. Galliker, Dibya Ghosh, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Devin LeBlanc, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Allen Z. Ren, Lucy Xiaoyang Shi, Laura Smith, Jost Tobias Springenberg, Kyle Stachowicz, James Tanner, Quan Vuong, Homer Walke, Anna Walling, Haohuan Wang, Lili Yu, and Ury Zhilinsky. $\pi_{0.5}$: a vision-language-action model with open-world generalization, 2025. URL https://arxiv.org/abs/2504.16054.

[27] Yunfan Jiang, Agrim Gupta, Zichen Zhang, Guanzhi Wang, Yongqiang Dou, Yanjun Chen, Li Fei-Fei, Anima Anandkumar, Yuke Zhu, and Linxi Fan. Vima: General robot manipulation with multimodal prompts. In Fortieth International Conference on Machine Learning (ICML), 2023.

[28] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. Openvla: An open-source vision-language-action model, 2024. URL https://arxiv.org/abs/2406.09246.

[29] Uikyum Kim, Dawoon Jung, Heeyoen Jeong, Jongwoo Park, Hyun-Mok Jung, Joono Cheong, Hyouk Ryeol Choi, Hyunmin Do, and Chanhun Park. Integrated linkage-driven dexterous anthropomorphic robotic hand. Nature communications, 12(1):7177, 2021.

[30] Qixiu Li, Yaobo Liang, Zeyu Wang, Lin Luo, Xi Chen, Mozheng Liao, Fangyun Wei, Yu Deng, Sicheng Xu, Yizhong Zhang, et al. Cogact: A foundational vision-language-action model for synergizing cognition and action in robotic manipulation. arXiv preprint arXiv:2411.19650, 2024.

[31] Qixiu Li, Yu Deng, Yaobo Liang, Lin Luo, Lei Zhou, Chengtang Yao, Lingqi Zeng, Zhiyuan Feng, Huizhi Liang, Sicheng Xu, Yizhong Zhang, Xi Chen, Hao Chen, Lily Sun, Dong Chen, Jiaolong Yang, and Baining Guo. Scalable vision-language-action model pretraining for robotic manipulation with real-life human activity videos. arXiv preprint arXiv:2510.21571, 2025.

[32] Qiyang Li, Zhiyuan Zhou, and Sergey Levine. Reinforcement learning with action chunking. arXiv preprint arXiv:2507.07969, 2025.

[33] Xinghang Li, Minghuan Liu, Hanbo Zhang, Cunjun Yu, Jie Xu, Hongtao Wu, Chilam Cheang, Ya Jing, Weinan Zhang, Huaping Liu, et al. Vision-language foundation models as effective robot imitators. arXiv preprint arXiv:2311.01378, 2023.

[34] Yunfei Li, Xiao Ma, Jiafeng Xu, Yu Cui, Zhongren Cui, Zhigang Han, Liqun Huang, Tao Kong, Yuxiao Liu, Hao Niu, Wanli Peng, Jingchao Qiao, Zeyu Ren, Haixin Shi, Zhi Su, Jiawen Tian, Yuyang Xiao, Shenyu Zhang, Liwei

Zheng, Hang Li, and Yonghui Wu. Gr-rl: Going dexterous and precise for long-horizon robotic manipulation, 2025. URL https://arxiv.org/abs/2512.01801.

[35] Hong Liu, K Wu, P Meusel, N Seitz, G Hirzinger, MH Jin, YW Liu, SW Fan, T Lan, and ZP Chen. Multisensory five-finger dexterous hand: The dlr/hit hand ii. IEEE/ASME Transactions on Mechatronics, 13(2):140–151, 2008.

[36] Izzeddin Liu, Yifan Jiang, Jiayuan Gu, Xi Chen, Huazhe Cui, and Hao Zhang. Roboflamingo: Bridge vision-language foundation models with robot manipulation. arXiv preprint arXiv:2311.01378, 2023.

[37] Guanxing Lu, Wenkai Guo, Chubin Zhang, Yuheng Zhou, Haonan Jiang, Zifeng Gao, Yansong Tang, and Ziwei Wang. Vla-rl: Towards masterful and general robotic manipulation with scalable reinforcement learning. arXiv preprint arXiv:2505.18719, 2025.

[38] Haining Luo and Yiannis Demiris. Tsl: Tracking deformable linear objects for bimanual shoe lacing. IEEE Robotics and Automation Letters, 2025.

[39] Hao Luo, Yicheng Feng, Wanpeng Zhang, Sipeng Zheng, Ye Wang, Haoqi Yuan, Jiazheng Liu, Chaoyi Xu, Qin Jin, and Zongqing Lu. Being-h0: Vision-language-action pretraining from large-scale human videos. arXiv preprint arXiv:2507.15597, 2025.

[40] Jianlan Luo, Charles Xu, Jeffrey Wu, and Sergey Levine. Precise and dexterous robotic manipulation via human-in-the-loop reinforcement learning. arXiv preprint arXiv:2410.21845, 2024.

[41] Siyuan Ma, Hong Zhang, and Xingxing Yang. Robomamba: Multimodal state space model for efficient robot reasoning and manipulation. In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024.

[42] Yecheng Jason Ma, William Liang, Guanzhi Wang, De-An Huang, Osbert Bastani, Dinesh Jayaraman, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Liv: Language-image representations and rewards for robotic control. In International Conference on Machine Learning (ICML), 2023.

[43] Yecheng Jason Ma, Shagun Sodhani, Dinesh andod Osbert Bastani Jayaraman, Vikash Kumar, and Amy Zhang. Vip: Towards universal visual reward and representation via value-implicit pre-training. In International Conference on Learning Representations (ICLR), 2023.

[44] Ajay Mandlekar, Zhu Yuke, Animesh Garg, Jonathan Booher, Max Spero, Albert Tung, Julian Gao, John Emmons, Anchit Gupta, Emre Orbay, et al. Roboturk: A crowdsourcing platform for robotic skill learning through imitation. In Conference on Robot Learning (CoRL), pages 879–893. PMLR, 2018.

[45] NVIDIA, Nikita Cherniadev Johan Bjorck andFernando Castañeda, Xingye Da, Runyu Ding, Linxi "Jim" Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, Joel Jang, Zhenyu Jiang, Jan Kautz, Kaushil Kundalia, Lawrence Lao, Zhiqi Li, Zongyu Lin, Kevin Lin, Guilin Liu, Edith Llontop, Loic Magne, Ajay Mandlekar, Avnish Narayan, Soroush Nasiriany, Scott Reed, You Liang Tan, Guanzhi Wang, Zu Wang, Jing Wang, Qi Wang, Jiannan Xiang, Yuqi Xie, Yinzhen Xu, Zhenjia Xu, Seonghyeon Ye, Zhiding Yu, Ao Zhang, Hao Zhang, Yizhou Zhao, Ruijie Zheng, and Yuke Zhu. GR00T N1: An open foundation model for generalist humanoid robots, 2025.

[46] Abby O'Neill, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, et al. Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0. In 2024 IEEE International Conference on Robotics and Automation (ICRA), pages 6892–6903. IEEE, 2024.

[47] Yuzhe Qin, Yueh-Hua andw Shaowei Liu Wu, Hanxiao Jiang, Ruihan Yang, Yang Fu, and Xiaolong Wang. Dexmv: Imitation learning for dexterous manipulation from human videos. In European Conference on Computer Vision (ECCV), 2022. Foundational work on bridging human hand video to dexterous robot policies.

[48] Ilija Radosavovic, Tete Xiao, Stephen James, Pieter Abbeel, Jitendra Malik, and Trevor Darrell. Real-world robot learning with masked visual pre-training. In Conference on Robot Learning, pages 416–426. PMLR, 2023.

[49] Shadow Dexterous Hand. https://shadowrobot.com/dexterous-hand-series/.

[50] Lin Shao, Yuzhe Zhang, Jiaming Zhang, Kexin Zhang, Weiming Wang, and Danfei Xu. Anyteleop: A general vision-based dexterous robot arm-hand teleoperation system. In Robotics: Science and Systems (RSS), 2024.

[51] Sharpa. https://www.sharpa.com/.

[52] Ananye Shaw, Kenneth andAgarwal and Deepak Pathak. Leap hand: Low-cost, efficient, and anthropomorphic hand for robot learning. In Robotics: Science and Systems (RSS), 2023. URL https://arxiv.org/abs/2309.06440.

[53] Kenneth Shaw, Shikhar Bahl, and Deepak Pathak. Videodex: Learning dexterity from internet videos. In Conference on Robot Learning, pages 654–665. PMLR, 2023.

[54] SimLab. Allegro hand v4.0 user manual, 2015. http://www.simlab.co.kr/Allegro-Hand.htm.

[55] Aravind Sivakumar, Kenneth Shaw, and Deepak Pathak. Robotic telekinesis: Learning a robotic hand-arm teleoperation interface from rgb cameras. In Robotics: Science and Systems (RSS), 2022.

[56] Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, et al. Octo: An open-source generalist robot policy. arXiv preprint arXiv:2405.12213, 2024.

[57] OpenLoong Baihu Team. Openloongdata-v1.0. https://www.openloong.org.cn/en/datasets/baihu, 2025.

[58] Unitree. https://www.unitree.com/cn/Dex5-1.

[59] Chen Wang, Linxi Fan, Jiankai Sun, Ruohan Zhang, Li Fei-Fei, Danfei Xu, Yuke Zhu, and Anima Anandkumar. Mimicplay: Long-horizon imitation learning by watching human play. arXiv preprint arXiv:2302.12422, 2023.

[60] Junjie Wen, Yichen Zhu, Jinming Li, Zhibin Tang, Chaomin Shen, and Feifei Feng. Dexvla: Vision-language model with plug-in diffusion expert for general robot control, 2025. URL https://arxiv.org/abs/2502.05855.

[61] Ruoshi Wen, Jiajun Zhang, Guangzeng Chen, Zhongren Cui, Min Du, Yang Gou, Zhigang Han, Junkai Hu, Liqun Huang, Hao Niu, et al. Dexterous teleoperation of 20-dof bytedexter hand via human motion retargeting, 2025. URL https://arxiv.org/abs/2507.03227.

[62] Kun Wu, Chengkai Hou, Jiaming Liu, Zhengping Che, Xiaozhu Ju, Zhuqin Yang, Meng Li, Yinuo Zhao, Zhiyuan Xu, Guang Yang, et al. Robomind: Benchmark on multi-embodiment intelligence normative data for robot manipulation. In Robotics: Science and Systems (RSS) 2025. Robotics: Science and Systems Foundation, 2025. URL https://www.roboticsproceedings.org/rss21/p152.pdf.

[63] Manuel Wüthrich, Felix Bauer, Isaac Garcia-Camburg, Felix Widmaier, et al. Trifinger: An open-source robot for learning dexterity. In Conference on Robot Learning (CoRL), pages 1871–1882. PMLR, 2020.

[64] Zhou Xian, Nikolaos Gkanatsios, Theophile Gervet, Tsung-Wei Ke, and Katerina Fragkiadaki. Chaineddiffuser: Unifying trajectory diffusion and keypose prediction for robotic manipulation. In 7th Annual Conference on Robot Learning, 2023. URL https://openreview.net/forum?id=W0zgY2mBTA8.

[65] Andy Zeng, Pete Florence, Jonathan Tompson, Stefan Welker, Jonathan Chien, Maria Attarian, Travis Armstrong, Ivan Krasin, Dan Duong, Vikas Sindhwani, et al. Transporter networks: Rearranging the visual world for robotic manipulation. In Conference on Robot Learning, pages 726–747. PMLR, 2021.

[66] Yifan Zhong, Xuchuan Huang, Ruochong Li, Ceyao Zhang, Zhang Chen, Tianrui Guan, Fanlian Zeng, Ka Num Lui, Yuyao Ye, Yitao Liang, Yaodong Yang, and Yuanpei Chen. Dexgraspvla: A vision-language-action framework towards general dexterous grasping, 2025. URL https://arxiv.org/abs/2502.20900.

Appendix

A Grasping Capability

Figure 9 ByteDexter V2 demonstrates human-like grasping, with a workspace that supports 33 grasp types.