EgoScale: Scaling Dexterous Manipulation with Diverse Egocentric Human Data

Ruijie Zheng[1*], **Dantong Niu**[1,2*], **Yuqi Xie**[1*], **Jing Wang**[1], **Mengda Xu**[1], **Yunfan Jiang**[1], **Fernando Castañeda**[1], **Fengyuan Hu**[1], **You Liang Tan**[1], **Letian Fu**[1,2], **Trevor Darrell**[2], **Furong Huang**[3], **Yuke Zhu**[1†], **Danfei Xu**[1†], **Linxi Fan**[1†]

[1]NVIDIA [2]University of California, Berkeley [3]University of Maryland

[*]Equal Contribution [†]Project Lead

https://research.nvidia.com/labs/gear/egoscale/

arXiv:2602.16710v1 [cs.RO] 18 Feb 2026

Abstract

Human behavior is among the most scalable sources of data for learning physical intelligence, yet how to effectively leverage it for dexterous manipulation remains unclear. While prior work demonstrates human-to-robot transfer in constrained settings, it is unclear whether large-scale human data can support fine-grained, high-degree-of-freedom dexterous manipulation. We present EgoScale, a human-to-dexterous-manipulation transfer framework built on large-scale egocentric human data. We train a Vision–Language–Action (VLA) model on over 20,854 hours of action-labeled egocentric human video—more than 20× larger than prior efforts—and uncover a log-linear scaling law between human data scale and validation loss. This validation loss strongly correlates with downstream real-robot performance, establishing large-scale human data as a predictable supervision source. Beyond scale, we introduce a simple two-stage transfer recipe: large-scale human pretraining followed by lightweight aligned human–robot mid-training. This enables strong long-horizon dexterous manipulation and one-shot task adaptation with minimal robot supervision. Our final policy improves average success rate by 54% over a no-pretraining baseline using a 22-DoF dexterous robotic hand, and transfers effectively to robots with lower-DoF hands, indicating that large-scale human motion provides a reusable, embodiment-agnostic motor prior.

1. Introduction

Human behavior is one of the most scalable sources of data for learning physical intelligence. Humans routinely perform dexterous manipulation across diverse objects, environments, and task variations at a scale that far exceeds what can be collected through robot teleoperation. As robotic hardware continues to improve toward more human-like kinematics and dexterity, a natural question arises: *can human data serve as a primary training signal for dexterous robot manipulation?*

Recent work shows that transfer from human data to robots is possible by aligning observations or actions across embodiments [12, 42, 25, 24, 30]. However, existing results remain limited in two respects. First, most approaches rely on relatively small human datasets, typically on the order of tens to hundreds of hours. Second, many focus on grippers or low DoF hands, where fine-grained finger articulation is absent. It therefore remains unclear whether human data can meaningfully support complex, dexterous manipulation at scale.

In this work, we show that human-to-robot transfer for dexterous manipulation is fundamentally a scaling phenomenon, and present EgoScale, a scalable human-to-dexterous-manipulation transfer framework built on large-scale egocentric human data. We pretrain on **20,854 hours of egocentric human manipulation data**, over **20× larger** than the dataset sizes used in prior studies of human–robot policy transfer, and uncover a clear *scaling law*: human wrist and hand action prediction validation loss follows a log-linear relationship with data volume. This enables us to extrapolate: as human data scales, validation loss continues to decrease, and the learned representations generalize increasingly well. Crucially, the loss strongly correlates with real robot

© 2026 NVIDIA. All rights reserved.

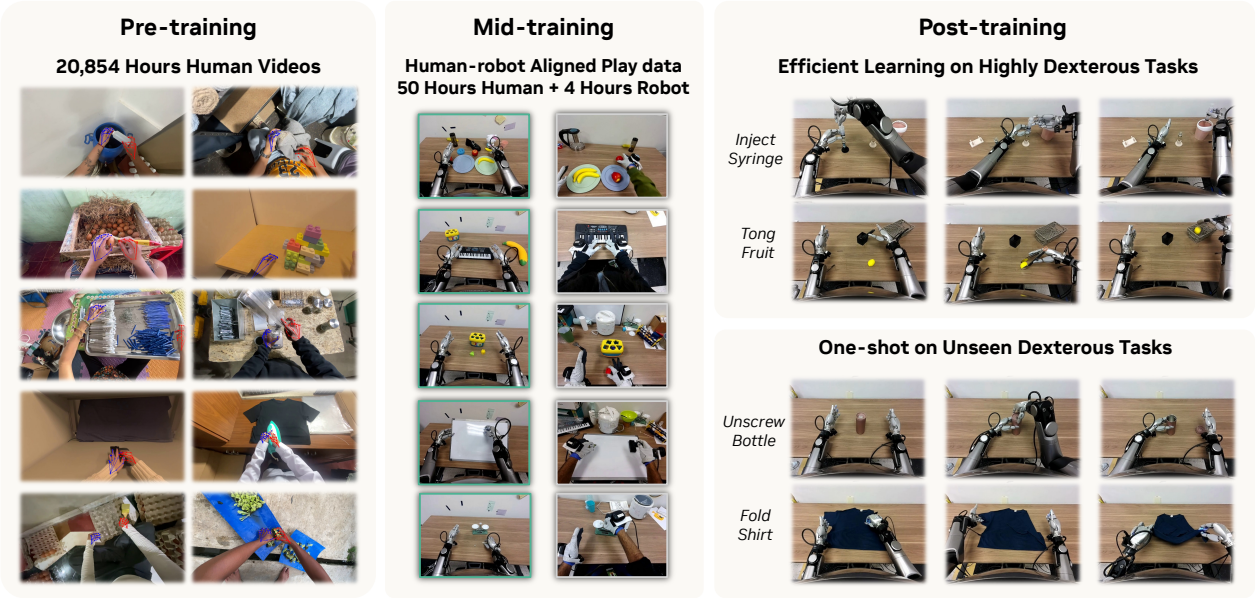


Figure 1: **EgoScale: Two-stage human-to-robot learning framework.** A flow-based Vision-Language-Action (VLA) policy is first pretrained on 20,854 hours of egocentric human videos using wrist motion and retargeted dexterous hand actions. A lightweight mid-training stage with aligned human robot play data (pairs highlighted with green and gray boundaries) adapts the representation to robot sensing and control. The resulting policy is post-trained on downstream tasks, enabling efficient learning of dexterous manipulation and one-shot generalization to unseen skills.

performance on long-horizon, complex manipulation tasks. Together, these results establish large-scale human data as a scalable and predictable supervision source for learning dexterous manipulation policies.

Beyond scale, we identify a simple yet effective training recipe that enables new generalization capabilities. We supervise the model using human manipulation behaviors represented as relative wrist motion and retargeted high-DoF hand joint actions. This aligned action space encourages the model to extract information that is directly useful for manipulation, rather than learning task-agnostic visual features. After pretraining, we introduce a small amount of aligned human-robot mid-training data through co-training. The mid-training data includes humans and robots performing similar manipulation tasks in matched tabletop scenes with comparable visual viewpoints. This alignment provides supervisions for grounding the pretrained representations in the robot's sensing and control space.

Importantly, this mid-training stage gives rise to *emergent one-shot and few-shot generalization*. With only one or a few robot demonstrations, the policy can adapt to new dexterous tasks without requiring extensive task-specific data collection. For instance, using a single robot demonstration, the trained policy achieves up to **88% average success** on shirt folding, even though the mid-training data contains only folding behaviors. Moreover, although human actions are supervised in a high-DoF dexterous hand space, the learned representations generalize to *substantially different robot embodiments*. On the Unitree G1 robot with a tri-finger hand, human-pretrained policies also achieve over **30% absolute improvement** in success rate across both evaluated tasks compared to baseline without human pretraining.

Taken together, our results show that effective dexterous human-to-robot transfer requires scale, explicit motion supervision, and a small amount of precise human–robot alignment. Rather than replacing robot data, large-scale human demonstrations dramatically amplify its effectiveness, pointing toward a future where humans can be treated as another scalable embodiment in robot learning. We summarize the contributions as follows:

- **Human Data Pretraining Scaling Laws.** At the scale of over 20k hours of human data, we uncover a clear log-linear relationship between data scale and hand-action prediction loss and show that this loss strongly correlates with real-robot dexterous manipulation performance.
- **An Effective Human-to-Robot Transfer Recipe.** We combine high-DoF human hand action supervision with a small amount of aligned human–robot mid-training, enabling strong post-training performance with minimal robot data.
- **Emergent One-shot Transfer and Generalization.** Our approach enables one-shot transfer to previously unseen dexterous task with 22-DoF hands. The pretraining policy also generalizes effectively to robots with lower-DoF hands, indicating that rich human motion provides a reusable motor prior.

2. Method

We aim to learn representations from large-scale egocentric human video that are directly useful for dexterous robot control. This setting poses two core challenges. First, human demonstrations are noisy and lack paired robot actions. Second, human and robot embodiments differ substantially in kinematics and control interfaces. Our method (Figure 1) addresses these challenges through two design choices. We first pretrain on human data using explicit supervision of wrist motion and hand articulation extracted from egocentric videos, forcing the model to learn physically grounded action representations. We then introduce a small amount of aligned human-robot data for mid-training, which grounds these representations in executable robot control without requiring large-scale paired demonstrations. Together, this two-stage design decouples data scale from embodiment alignment, enabling effective transfer from large human datasets to dexterous robot manipulation.

2.1. Human Action Representation

Raw Sensor Streams. Each human demonstration consists of egocentric RGB observations captured from a head-mounted camera, together with estimated camera motion and human hand pose obtained from off-the-shelf perception pipelines. We convert these raw sensory signals into a unified action representation suitable for large-scale pretraining and downstream robot execution. Let $\mathcal{F}_w$ denote the world frame and $\mathcal{F}_c^t$ the camera frame at time t. The estimated camera pose is represented as $\mathbf{T}_{w\leftarrow c}^t \in \mathbb{SE}(3)$. Human hand pose is modeled by 21 keypoints, each represented as a rigid transform $\mathbf{H}_{c,i}^t \in \mathbb{SE}(3)$ in the camera frame, where $i = 1$ corresponds to the wrist. The wrist pose in the world frame is given by $\mathbf{W}_w^t = \mathbf{T}_{w\leftarrow c}^t \mathbf{H}_{c,1}^t$.

Wrist-level Arm Motion. To obtain motion commands that are invariant to global camera movement, we represent arm motion using relative wrist motion between consecutive timesteps. Given timestep t in an action chunk, $\Delta\mathbf{W}^t = (\mathbf{W}_w^0)^{-1}\mathbf{W}_w^t$. This relative end-effector formulation removes dependence on absolute camera pose and captures local arm motion in a physically meaningful manner. The same representation is shared across human demonstrations and robot executions, serving as the primary arm-level action abstraction for cross-embodiment learning.

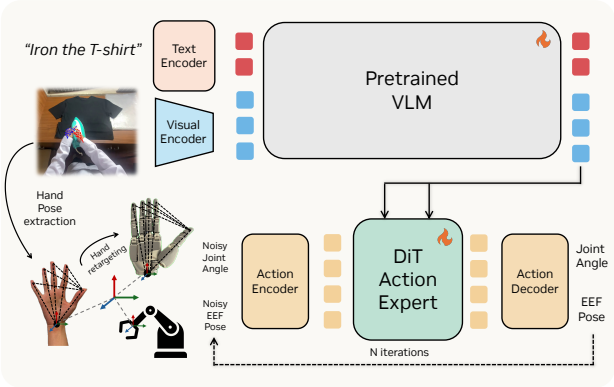
Hand Articulation. For finger-level control, we retarget the 21 human hand keypoints into a dexterous robot hand joint space using an optimization-based procedure that enforces joint limits and kinematic constraints. Our default choice is the 22-DoF hand action space of the Sharpa hand [29], which preserves human finger articulation during pretraining while aligning with the control interface of our target robot. Although this representation is defined using a high-DoF hand, we later show that the learned models transfer effectively to robots with lower-DoF hands.

2.2. Human Data Sources and Processing

Stage I: Large-Scale Egocentric Human Pretraining Data. We pretrain our model on a large-scale mixture of egocentric human activity datasets totaling **20,854** hours of video. The majority consists of in-the-wild egocentric recordings spanning diverse real-world environments (e.g., household, industrial, retail, and educational settings), covering 9,869 scenes, 6,015 tasks, and 43,237 objects, and providing broad coverage of long-tailed

manipulation behaviors.

(a) Egocentric Human Data Collection.



(b) EgoScale Model Architecture.

Figure 2: **Human Data Collection and Model Architecture.** (**Left**) Aligned human-robot mid-training data are collected using the same sensing setup as the robot. Vive trackers and Manus gloves capture arm and hand motion, while one head-mounted camera and two wrist-mounted cameras record egocentric and wrist views, enabling consistent perception–action alignment. (**Right**) A flow-based VLA policy with a VLM backbone and DiT action expert. Human and robot data are unified through a wrist-level action representation, with lightweight embodiment-specific adapters for proprioception and hand actions.

All recordings are captured using egocentric RGB cameras at 30 FPS. We apply off-the-shelf SLAM and hand-pose estimation pipelines to recover camera motion and human hand trajectories. Although these estimates are noisy due to unconstrained data collection, the scale and diversity of the data provide effective supervision for learning transferable action representations, which continue to improve downstream performance as data volume increases.

To complement this large-scale but noisy supervision, we additionally incorporate 829 hours of EgoDex dataset [8], collected using Apple Vision Pro with accurate wrist and hand tracking. EgoDex covers 194 tabletop manipulation tasks involving everyday objects and provides higher-precision kinematic signals that help anchor pretraining while preserving scalability.

Stage II: Aligned Human-Robot Mid-Training Data. To further bridge the embodiment gap between human demonstrations and robot execution, we introduce a smaller dataset with both human and teleoperated robot data. We later show that this dataset is critical to anchor the pretrained representations to the robot's sensing and action spaces.

This dataset comprises 344 tabletop manipulation tasks, with each task captured in approximately 30 human trajectories and 5 robot trajectories, totaling about 50 hours of human data and only 4 hours of robot data. As shown in Figure 2a, human demonstrations are collected using the same camera configuration as the robot, with matched viewpoints and calibrated intrinsics, ensuring that visual observations are directly comparable across domains. Human hand motion is captured using the same motion-capture stack as in robot teleoperation: Vive trackers provide wrist pose (3D position and orientation), while Manus gloves record full in-hand pose as 25 joint transforms. All motion signals are synchronized with the video stream.

Compared to the large-scale but unconstrained data used in Stage I, this dataset is significantly smaller but explicitly embodiment-aligned. It focuses on tabletop tasks that match the robot's workspace and kinematics, enabling abstract human actions learned during pretraining to be grounded in executable robot control. Together, Stage I and Stage II decouple scale and alignment: Stage I provides diversity and semantic grounding, while Stage II supplies precise human-robot correspondence for downstream deployment.

2.3. Model Architecture

As shown in Figure 2b, our model follows a flow-based VLA architecture similar to GR00T N1 [19]. At each timestep t, the model conditions on an observation $o_t = (I_t, l_t)$ consisting of an image and a language instruction, which is encoded into a vision-language embedding ϕ_t. The model then predicts a chunk of future actions using a flow-matching objective.

For robot data, the model conditions on the robot proprioceptive state q_t, while human demonstrations do not provide such signals. In the absence of proprioception, we replace q_t with a learnable placeholder token, enabling a unified model formulation without architectural changes. To accommodate multiple robot embodiments with different state and hand action spaces, following GR00T N1 [19], we use lightweight embodiment-conditioned MLP adapters at the input and output interfaces. Specifically, these adapters encode embodiment-specific proprioceptive state and decode hand actions, while relative wrist motion prediction, the vision-language backbone, and the DiT action expert are fully shared. In practice, this mechanism is used only for a small number of additional embodiments (e.g., G1 with a tri-finger hand).

2.4. Training Recipe

We use a three-stage training pipeline. In **Stage I** (human pretraining), we train on 20K hours of egocentric human data for 100K steps with 256 GB200 GPUs using a global batch size of 8,192 and learning rate 5×10^{-5}, fully unfreezing every parameter of the VLA model to absorb large-scale data. Then in **Stage II** (aligned mid-training), we train on the aligned human-robot play dataset for 50K steps with batch size 2,048 and learning rate 3×10^{-5}, freezing the vision-language backbone while only updating the vision encoder and DiT action expert to anchor representations to robot sensing and control. In **Stage III** (post-training), we fine-tune on task-specific robot demonstrations for 10K steps with batch size 512 and learning rate 3×10^{-5}. During post-training, the vision encoder is frozen if mid-training is used and unfrozen otherwise, to accommodate new embodiments when needed.

2.5. Robot Systems and Control

The real world experiments are conducted on the Galaxea R1Pro humanoid robot with 22-DoF Sharpa dexterous robot hands. We refer the readers to Appendix B for the system figure.

Dual-arm Wheeled Humanoid System Galaxea R1Pro. We fix the base and torso and focus on bimanual manipulation, controlling both 7-DoF arms in relative end-effector space where actions specify incremental position and orientation changes, matching the wrist-pose representation used in human demonstrations for direct human-robot alignment.

22-DoF Dexterous Hands. We equip the robot with Sharpa Wave hands with 22 degrees of freedom and joint-space control, where actions directly specify target joint angles, enabling precise articulation and preserving the fine-grained structure of retargeted human hand motion.

Perception System. We use three RGB cameras: a head-mounted camera that provides an egocentric first-person view consistent with human videos, and two wrist ones mounted on the inner side of each wrist facing the palm, capturing close-range hand-object interactions and provide detailed visual feedback essential for fine-grained dexterous manipulation.

3. Experiment

In this section, we aim to answer the following research questions through our experiments:
RQ1: *Does large-scale egocentric human pretraining improve downstream dexterous manipulation performance compared to training from scratch or embodiment-aligned data alone?*
RQ2: *How does the scale of human pretraining data affect representation quality and real-robot performance?*
RQ3: *What role does mid-training play in enabling few-shot adaptation and generalization to novel tasks?*

RQ4: *Do human-pretrained representations transfer across robot embodiments with substantially different kinematics and control interfaces?*
RQ5: *How does the choice of human action representation during pretraining affect downstream dexterous manipulation?*

3.1. Experiment Setup

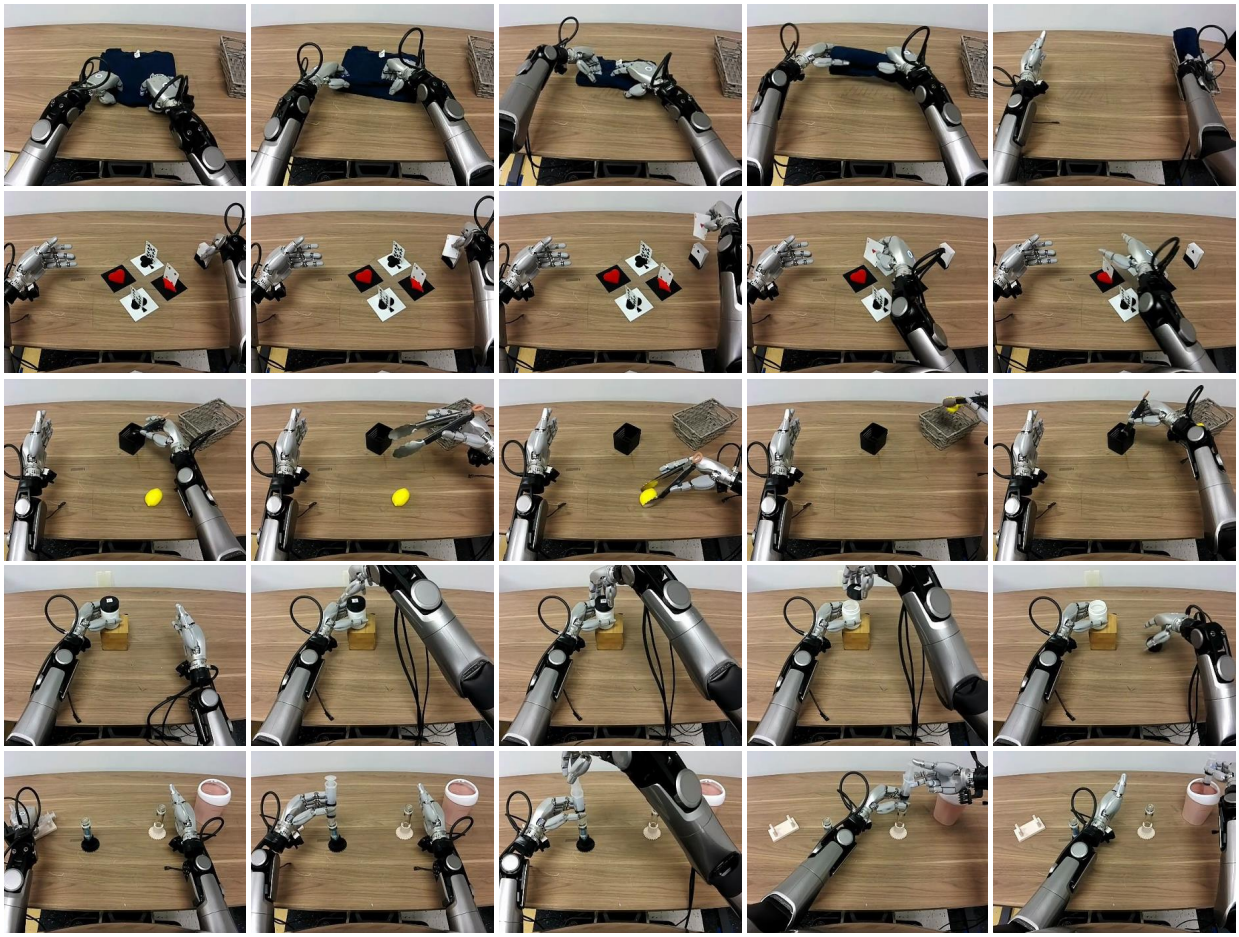


Figure 3: **Post-Training Evaluation Tasks.** Five dexterous manipulation tasks used to evaluate post-training performance

Tasks. To evaluate policy performance, we design five highly dexterous manipulation tasks shown in Figure 3.

Each task is provided with 100 teleoperated robot demonstrations, except for *Shirt Rolling*, a deformable manipulation task that requires less precise control, for which we use only 20 demonstrations.

(Task I) *(Shirt) Shirt Rolling.* The robot coordinates both hands to alternately fold and roll a T-shirt into a cylindrical shape before placing it into a basket.

(Task II) *(Card) Card Sorting.* The robot uses its fingers to rub and separate a single card from a tightly stacked deck, and then precisely inserts it into the correct holder based on color.

(Task III) *(Tong) Dexterous Tool Use: Tongs for Fruit Transfer.* The robot first grasps a pair of tongs from a toolbox and then uses them to pick up a fruit and place it at a target location.

(Task IV) *(Bottle) Unscrewing a Bottle Cap.* The robot grasps and continuously rotates a small cap to remove it from a bottle. We collect demonstrations on four bottles of different sizes, with 25 trajectories per bottle.

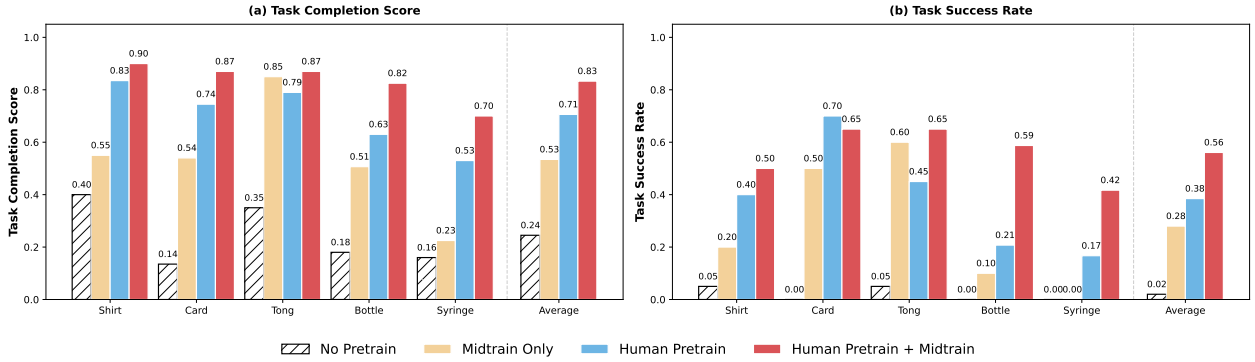


Figure 4: **Main Experimental Results.** Comparison of Human Pre-train + Mid-Training, Human Pretraining, and No Pretraining across five dexterous manipulation tasks under two evaluation metrics.

(Task V) *(Syringe) Syringe Liquid Transfer.* This is the most challenging task, requiring the robot to pick up a syringe, draw liquid from tube A, inject it into tube B, and discard the syringe into a trash can. The task involves long-horizon, multi-step reasoning, precise spatial alignment for fluid extraction and injection, and dexterous manipulation of the syringe plunger.

Evaluation Metric. To evaluate policy performance, we train each method using two random training seeds. Then, for each trained policy checkpoint, we evaluate performance over 10 trials, except for Task III, where we conduct 4 trials per bottle across four bottle instances, resulting in 16 evaluation trials. To ensure consistency across evaluation runs, we employ an image-overlay–based initialization procedure, in which the robot evaluator is provided with a visual overlay of the target initial scene configuration to reduce variability in initial conditions. For each task, we record both the absolute task success rate and fine-grained task completion score.

3.2. Large-Scale Human Pretraining Is Key to Strong Dexterous Manipulation Policy Performance

To evaluate the impact of large-scale human pretraining and aligned mid-training on policy learning efficiency, we compare four checkpoints: (1) a model trained from scratch, (2) a model pre-trained only on the midtrained aligned human–robot play dataset, (3) a model pretrained on large-scale human data, and (4) a human-pretrained model further mid-trained on aligned human–robot data. For each checkpoint, we report both the task completion score and the absolute success rate.

Results are summarized in Figure 4. Across all tasks, human pretraining consistently yields substantial performance gains over training from scratch, improving average task completion by over 55%. Notably, large-scale human pretraining, despite being noisy, unconstrained, and not task- or sensor-aligned, already outperforms the mid-training-only baseline across most tasks. This provides the evidence that scale and diversity of human demonstrations provide strong inductive biases for dexterous manipulation, even in the absence of precise embodiment alignment.

Finally, combining human pretraining with a small amount of aligned mid-training yields the best overall performance, suggesting a complementary effect: large-scale human data supplies general manipulation structure, while mid-training anchors these representations to executable robot control.

3.3. Policy Performance Scales with Pretraining Data Size

We study how the scale of egocentric human pretraining data affects downstream *real-robot* manipulation performance, and analyze how this behavior is reflected in offline human-action prediction metrics. We pretrain models using **1k, 2k, 4k, 10k, and 20k hours** of human data. To isolate the effect of stage II midtraining, we directly post-train each checkpoint on the downstream task and evaluate their performance.

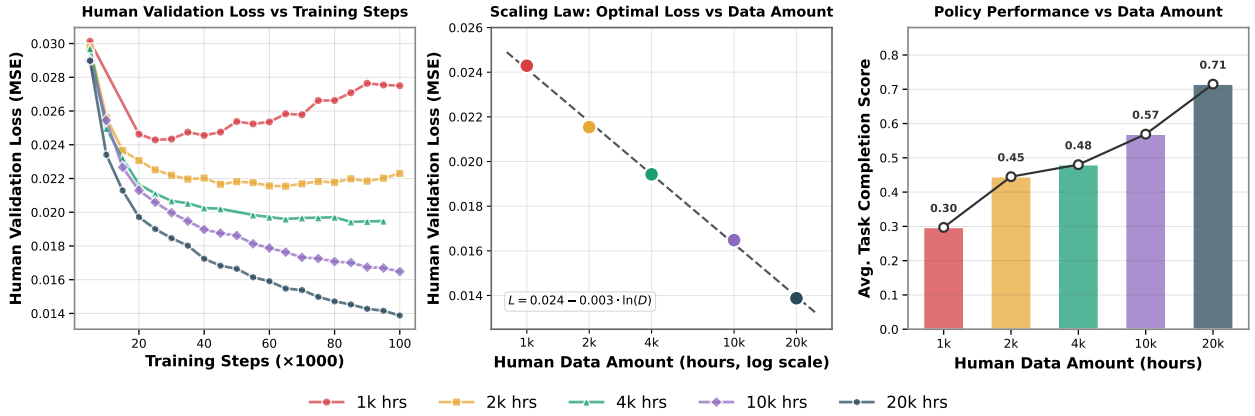


Figure 5: **Scaling behavior of human pretraining.** *Left:* Human validation loss versus training steps for models pretrained with increasing amounts of egocentric human data (1k–20k hours). Larger datasets yield stable, monotonic improvements, while smaller datasets exhibit early overfitting. *Center:* Optimal validation loss at convergence as a function of human data scale, revealing a near-perfect log-linear scaling law ($R^2 = 0.9983$). *Right:* Downstream robot performance after post-training, measured by average task completion score, improves consistently with increased human data scale. Together, these results demonstrate predictable scaling of learned action representations and their direct translation to improved dexterous manipulation performance.

As shown in Figure 5 (right), increasing the amount of human pretraining data leads to consistent and substantial gains in downstream robot performance. Average task completion rises monotonically from 0.30 at 1k hours to 0.71 at 20k hours, with no signs of saturation in the explored regime. These results indicate that large-scale human data provides increasingly strong priors for dexterous manipulation, even though the human demonstrations are noisy, unconstrained, and not task-aligned.

To better understand this trend, we examine how pretraining data scale affects the quality of learned action representations. We evaluate each pretrained model on a held-out human-video validation set consisting of 2,000 egocentric episodes. For evaluation, we randomly sample 20 timesteps per trajectory; at each timestep, we draw 16 samples from the flow-matching policy, average the predicted action chunks, and compute the mean squared error against ground-truth wrist and hand actions.

Figure 5 (left) shows human validation loss as a function of training steps for different data scales. Models trained with smaller datasets (1k–2k hours) initially reduce validation loss but later plateau or degrade, indicating overfitting to limited behavioral diversity. In contrast, models trained with larger datasets (10k–20k hours) exhibit stable, monotonic improvement throughout training without signs of overfitting.

Strikingly, when we plot the optimal validation loss achieved at convergence against data scale (Figure 5, center), we observe a remarkably clean log-linear scaling law:

$$\mathbf{L} = \mathbf{0.024} - \mathbf{0.003} \cdot \ln(\mathbf{D}), \tag{1}$$

where D denotes the number of hours of human pretraining data. The fitted curve achieves an R^2 of **0.9983**, indicating an almost perfect linear relationship in log space. Crucially, this offline scaling behavior is strongly predictive of real-robot performance. Human validation loss closely tracks downstream task completion across data scales, establishing it as a meaningful indicator of embodied control capability rather than a purely offline metric.

Together, these results show that effective human-to-robot transfer for dexterous manipulation is fundamentally a scaling phenomenon. Within the explored regime, increasing human data yields predictable reductions in validation loss and corresponding improvements in robot performance, with no evidence of diminishing returns.

While we do not extrapolate beyond the measured range, this trend suggests substantial headroom for further gains as both human data scale and model capacity continue to increase.

3.4. Aligned Human-Robot Mid-Training Enables One-Shot Transfer

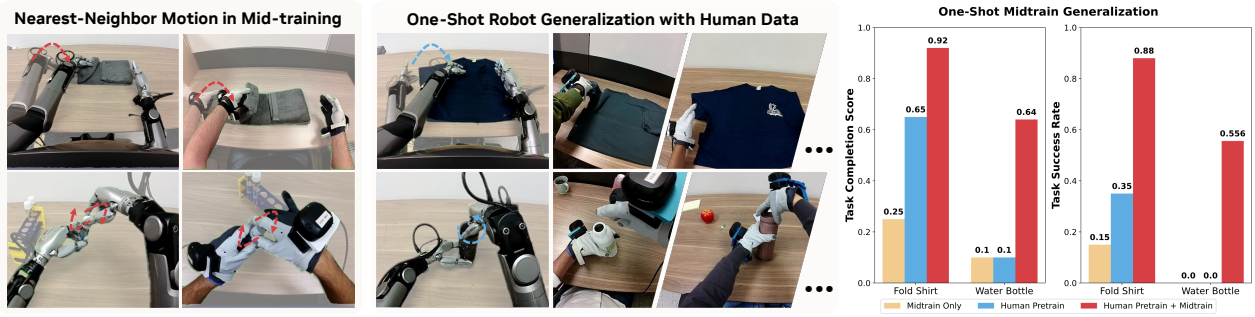


Figure 6: **Aligned mid-training enables emergent one-shot transfer.** During post-training, the policy is trained on only a single robot demonstration per task, together with aligned human demonstrations (100 trajectories per object).

We evaluate whether aligned human–robot mid-training enables learning of previously unseen skills under extremely limited robot supervision. Starting from a human-pretrained model, we apply mid-training on the aligned play dataset and post-train on new tasks using only a single robot demonstration, supplemented by aligned human demonstrations. We consider two tasks, *Fold Shirt* and *Unscrewing Water Bottles*, neither of which appears in the mid-training data.

For *Fold Shirt*, we provide one robot demonstration and 100 aligned human demonstrations. For *Unscrewing Water Bottles*, we evaluate generalization across object variation using three bottles of different geometries, with one robot demonstration and 100 aligned human demonstrations per bottle.

As shown in Figure 6, models that omit either large-scale human pretraining or aligned mid-training fail in this one-shot setting. In contrast, the **Pretrain + Midtrain** model achieves success rates of 0.88 on *Fold Shirt* and 0.55 on *Unscrewing Water Bottles*, demonstrating strong few-shot generalization. Failures are typically partial: incomplete folds for shirts, or unstable grasp maintenance after cap removal for bottles. These results indicate that aligned mid-training enables a form of transfer that does not emerge from human pretraining or embodiment-specific data alone.

This generalization is enabled by shared motion structure between mid-training and evaluation tasks. Although the objects and task instances differ substantially, the mid-training data exposes the model to common motion primitives, allowing these behaviors to transfer with as little as a single target-robot demonstration.

3.5. Human Pretraining Enables Cross-embodiment Transfer

Beyond bimanual tabletop manipulation with high-DoF dexterous hands, we also demonstrate that human pretraining learns transferable action representations that generalize to new robot embodiments. Our human pretraining represents actions as relative $\mathbb{SE}(3)$ end-effector motions coupled with a 22-DoF dexterous hand joint space defined by the Sharpa hand. Although this action space is instantiated using a specific robotic hand, it could also function as a motor prior that encodes reusable grasping and manipulation primitives, such as hand opening, closure, and coordinated finger articulation, that is transferable to other robotic hands. To evaluate cross-embodiment generalization, we introduce a substantially different robot platform: the Unitree G1, which features a shorter arm with a reduced reachable workspace and a 7-DoF tri-finger hand, resulting in markedly different kinematics.

We evaluate posttraining policy performance on two tasks. In the first task, *Pen in Bin*, the G1 robot uses its left arm to open a trash bin and then uses its right hand to pick up a pen and place it into the bin. In the

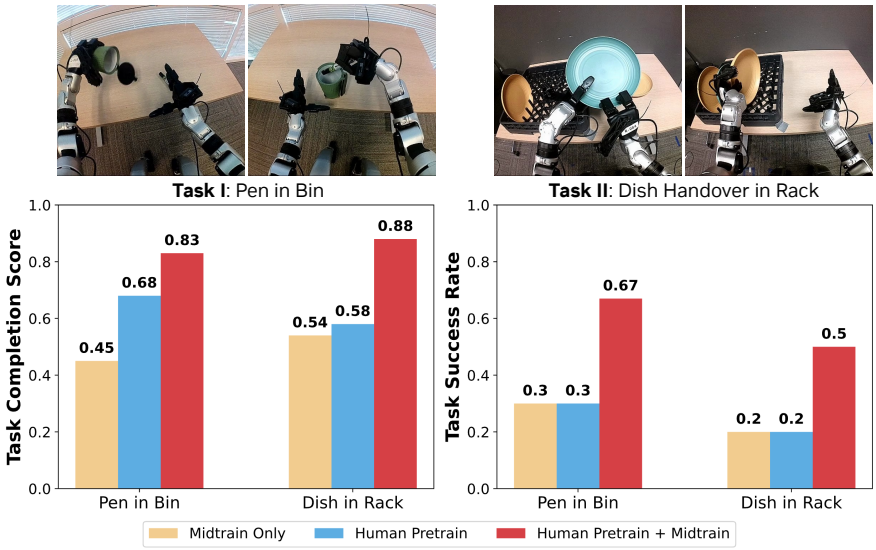


Figure 7: **Mid-training enables cross-embodiment generalization to G1 Tasks.** Integrating a small amount of G1 play data during mid-training enables strong transfer to a low-DoF tri-finger hand, outperforming mid-training or pre-training only.

second task, *Dish in Rack*, the robot is presented with three plates on a table and must perform a right-to-left handover to place the dishes onto a rack. During execution, the policy predicts upper-body target commands, while lower-body balance and locomotion are handled by a separately trained Homie [2] policy that outputs lower-body joint commands. Same as in previous experiments, we train each method using two random seeds. For each seed, we evaluate the resulting policy over 10 trials and report the average policy score and success rate across the two seeds.

As shown in Figure 7, incorporating mid-training on a data mixture that includes G1 embodiment play data also leads to a substantial performance improvement on both of the G1 manipulation tasks, compared to a checkpoint trained on the same G1 embodiment–specific data alone. Crucially, this improvement cannot be attributed to increased exposure to G1 trajectories alone: policies pretrained and fine-tuned directly on the same G1 dataset, without prior human pretraining, fail to achieve comparable success rates. Empirically, we also observe that policies pretrained on human data exhibit qualitatively smoother behaviors.

These results suggest that EgoScalepre-training learns a reusable manipulation structure that transfers across embodiments, while mid-training adapts this structure to the G1's sensing and control interfaces.

3.6. Hand Action Space Design for Human Pretraining

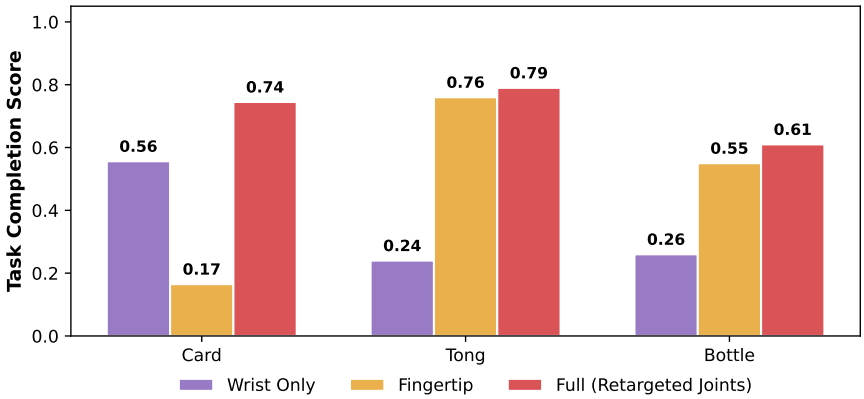


Figure 8: Task completion score across tasks for different human pretraining action representations.

We study how the choice of human hand action representation during pretraining affects downstream dexterous manipulation. Our default setup represents human hand motion using a 22-DoF dexterous joint action space, and we compare against two alternatives: (i) a *wrist-only* representation that removes all finger-level supervision, and (ii) a *fingertip-based* representation [42] that predicts $\mathbb{SE}(3)$ trajectories of the wrist and fingertips, followed by an MLP mapping to robot joint commands.

As shown in Figure 8, the action representation strongly influences task performance. The wrist-only representation performs poorly across all tasks, particularly those requiring precise finger articulation and contact timing (e.g., *Tongs*, *Cards*). In these cases, policies frequently grasp tools too weakly or at unstable heights, or close the hand prematurely, resulting in missed or brittle contacts.

The fingertip-based representation provides richer geometric supervision and improves performance on some tasks, but remains inconsistent. Small errors in fingertip pose often lead to implausible joint configurations after mapping, causing unstable grasps or contact loss in contact-sensitive tasks such as *Cards* and *Bottle*. In contrast, pretraining with retargeted joint-space hand actions yields the most consistent performance across all tasks. We therefore use retargeted joint-space hand actions as a practical and effective choice for large-scale human pretraining.

4. Related Work

Robot Learning from Human Data. Human demonstrations have been widely used to scale robot learning, with early works leveraging human videos primarily for representation learning or intent inference [17, 36, 15, 14, 45]. Subsequent approaches use human data to guide planning or high-level control while relying on robot demonstrations for low-level execution [34, 43, 44, 18, 38, 39]. More recent methods exploit advances in egocentric sensing and 3D hand tracking to treat human videos as dense action supervision. EgoMimic [12], Qiu et al. [25], and DexWild [30] co-train a unified imitation policy on human and robot demonstrations through explicit alignment, while EgoVLA [42] pretrains a VLA model on human hand motion and transfers it to robots via inverse kinematics and retargeting. Concurrent work [13] demonstrates that VLA pretrained on large-scale diverse cross-embodiment data can unlock human to robot transfer capability.

In contrast to these works, our work focuses on *pure large-scale human pretraining* with minimal robot supervision, with the goal of learning both wrist motion and dexterous hand articulation directly from diverse egocentric human videos. Compared to approaches that primarily transfer wrist motion for gripper-based platforms, our setting retains rich hand motion information that is critical for dexterous manipulation. Compared to recent methods that also leverage hand-level supervision, we pretrain on substantially larger-scale human data and systematically show that scaling human video yields consistent improvements in both human validation metrics and downstream robot performance, enabling strong dexterous manipulation performance on advanced multi-DoF robotic hands.

Scaling Properties in Robot Learning. Inspired by scaling laws observed in language and vision, recent work has begun to investigate whether similar principles govern robot learning. Empirically, large-scale robot datasets and foundation-style policies show that increasing data diversity and coverage leads to improved robustness and generalization across tasks and environments [11, 46, 20, 33, 31, 4]. Hu et al. [9] shows that policy generalization follows an approximate power law with environment and object diversity, while additional demonstrations quickly saturate, highlighting that diversity is more important than raw data volume. This is consistent with prior work emphasizing efficient data collection through compositional diversity [7, 37]. Compared to prior work that primarily scales robot-collected data, we show that scaling diverse in-the-wild human egocentric data leads to systematic gains in dexterous manipulation, establishing human video as an efficient and scalable supervision source.

Learning Dexterous Manipulation. Dexterous manipulation has evolved from analytic and control-based grasping methods that model force closure, contact stability, and hand kinematics [21, 22, 26, 27, 23, 6]

to learning-based approaches that acquire contact-rich behaviors from data [1, 16, 40]. Subsequent works introduce structured representations such as grasp affordances, contact maps, and hand–object interaction fields to better capture dexterous geometry and physics [3, 5, 10, 41, 32]. More recent methods aim to learn generalizable multi-fingered manipulation policies with unified perception and control [28, 35]. However, scaling dexterous manipulation remains challenging due to the high-dimensional action space, the cost of robot data collection, and the limited capabilities of current dexterous hand hardware. Our work leverages egocentric human videos with dense hand tracking as an alternative supervision source, showing that human hand motion provides transferable signals for improving robotic dexterous manipulation.

5. Acknowledgement

This work would not have been possible without the dedication and expertise of our robot operators: Ivy Tam, Ashley Kim, Aly Khater, Rhea Alve, Noah Huang, Ty Seligman, Mahnoor Kareem, Christian Soto, Chaitanya Kothapalli, Kendra Shu, Rex Asato, Eley Barba Preciados, and April Zitkovich, who provided invaluable support in large-scale data collection and evaluation. We are also grateful to Johan Bjorck, Scott Reed, Runyu Ding and Joel Jang for their insightful discussions and feedback throughout the project.

6. Conclusion

In this work, we show that effective human-to-robot transfer for dexterous manipulation is fundamentally a scaling phenomenon. By pretraining a vision–language–action policy on over 20K hours of egocentric human manipulation data, EgoScale uncovers a clear log-linear scaling law between human action prediction loss and data scale, and demonstrates that this loss strongly predicts downstream real-robot performance. Beyond scale, we identify a simple and effective transfer recipe: combining large-scale human pretraining with a small amount of aligned human–robot mid-training enables strong long-horizon dexterous manipulation, emergent one-shot adaptation, and robust transfer across robot embodiments with substantially different kinematics and hand designs.

Looking forward, several directions remain open. While we observe no saturation within the explored regime, jointly scaling human data and model capacity may unlock further gains, including improved long-horizon planning and compositional generalization. As egocentric human data continues to grow, incorporating weaker or unlabeled video via self-supervised objectives may further amplify these benefits. Finally, as robotic hardware becomes more human-like in kinematics and dexterity, the embodiment gap will naturally shrink, enabling stronger transfer and potentially zero-shot execution on novel tasks. Together, these results point toward a future where humans can be treated as a truly scalable embodiment for learning general embodied intelligence.

A. Details of Robot System

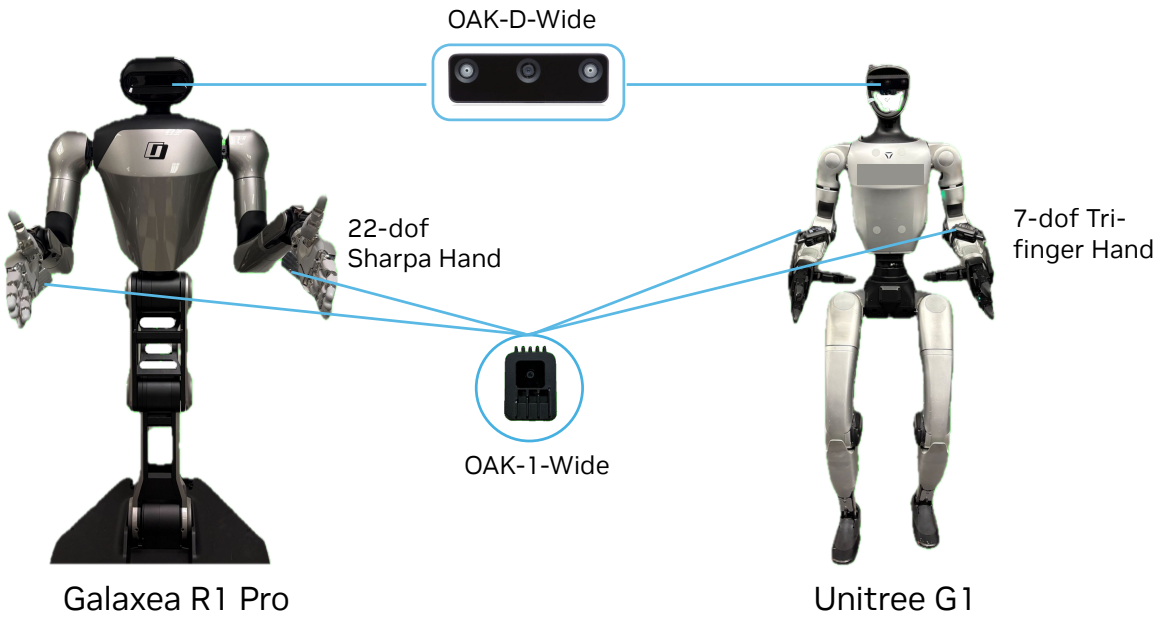


Figure 9: Visualization of the robot system setups on the Galaxea R1 Pro and Unitree G1 platforms. The Galaxea R1 Pro is equipped with 22-DoF Sharpa dexterous hands, while the Unitree G1 is mounted with 7-DoF tri-finger hands. Both robots have with two OAK-1-Wide cameras for the wrist view and an OAK-D-Wide head-mounted camera to provide egocentric visual observations.

B. Task Descriptions and Evaluation Rubric

In this section, we provide detailed descriptions of the post-training evaluation tasks introduced in Section 3, together with the corresponding evaluation rubrics. Each task is specified by a natural-language instruction and is evaluated using a task completion score in $[0, 1]$, designed to capture partial progress in addition to binary success.

Depending on the structure of the task, we adopt one of two scoring strategies. For tasks that naturally decompose into a sequence of independent, well-defined sub-skills (e.g., tool grasping, placement, or discrete manipulation steps), we use an *additive rubric*, where the final score is the sum of achieved sub-scores. For tasks involving deformable objects or tightly coupled manipulation stages, where intermediate states are difficult to define independently, we instead use a *progress-based rubric*, where the score reflects the furthest task milestone achieved. Both designs aim to provide fine-grained, interpretable feedback while remaining faithful to the underlying task structure. For each task, unless specifically mentioned, we evaluate each checkpoint for 10 trials.

R1 Pro Sharpa

- **Task I: Shirt Rolling.** *Text Instruction: "Roll the T-shirt and put it into the basket."* The robot must coordinate both hands to progressively fold, roll, and place a deformable T-shirt into a basket.
  Grading rubric (progress-based):
  - 0.0: No folding behavior.
  - 0.3: Executes an initial fold.
  - 0.5: Executes multiple folds or partial rolling.
  - 0.8: Completes a continuous rolling motion into a compact shape.

  - 1.0: Places the rolled shirt into the basket.

  Evaluation trials: 10 trials.
- **Task II: Card Sorting.** *Text Instruction: "Pick up the card and sort it into the correct card holder."* The robot must separate a single card from a tightly stacked deck and insert it into the correct holder.
  Grading rubric (progress-based):
  - 0.0: No successful pickup; entire deck disturbed or lifted.
  - 0.3: Poor grasp; multiple cards lifted unintentionally.
  - 0.5: Single card successfully picked up.
  - 0.7: Card placed with noticeable disturbance or incorrect insertion.
  - 0.9: Correct card placed correctly with minor disturbance.
  - 1.0: Single correct card cleanly placed into the correct holder.
- **Task III: Tong Fruit Transfer.** *Text Instruction: "Use the tong to pick up the fruit and place it into the basket."* The robot must first grasp the tong and then use the tool to pick up a target object and place it into a container. This task is evaluated with 2 random fruits, lemon and plum, each with 5 evaluation trials.
  Grading rubric (additive):
  - $+0.4$: Successfully grasps the tongs.
  - $+0.2$: Picks up the fruit using the tongs.
  - $+0.2$: Places the fruit into the basket.
  - $+0.2$: Returns the tongs to the table.
- **Task IV: Bottle Cap Unscrewing.** *Text Instruction: "Unscrew the cap from the bottle."* The robot must grasp a bottle, rotate the cap multiple times, remove it, and place it on the table. To test generalization across object variations, we evaluate 4 bottle configurations with different heights and cap sizes. Performance is measured over 12 trials in total, with 4 trials conducted for each bottle.
  Grading rubric (additive):
  - $+0.1$: Grasps the bottle.
  - $+0.5$: Unscrews the cap with at least three continuous rotations.
  - $+0.2$: Fully removes the cap.
  - $+0.2$: Places the cap on the table.
- **Task V: Syringe Liquid Transfer.** *Text Instruction: "Pick up the syringe, draw liquid from tube A, inject it into tube B, and throw the syringe into the trash can."* This task requires precise tool manipulation, spatial alignment, and long-horizon sequencing.
  Grading rubric (additive):
  - $+0.1$: Picks up the syringe.
  - $+0.1$: Correctly aims at tube A.
  - $+0.2$: Pulls the plunger to draw liquid.
  - $+0.1$: Re-aims at tube B.
  - $+0.2$: Pushes the plunger to inject liquid.
  - $+0.2$: Hands over or releases the syringe.
  - $+0.1$: Disposes of the syringe into the trash can.

  To evaluate one-shot adaptation, we consider two additional manipulation tasks on the R1 Pro Sharpa platform. For each task, the policy is post-trained using a single robot demonstration, and performance is evaluated using a task-specific completion score in $[0, 1]$.
- **Task VIII: One-Shot T-Shirt Folding.** *Text Instruction: "Fold the T-shirt."*
  Grading rubric (additive):
  - $+0.4$: Successfully folds at least one sleeve.
  - $+0.4$: Successfully folds both sleeves.
  - $+0.3$: Executes a bottom fold.

  During the evaluation, we also subtract a 0.1 if it does one of the fold but the folding is messy.

- **Task IX: One-Shot Bottle Cap Unscrewing.** *Text Instruction: "Unscrew the cap from the water bottle."* The task decomposes naturally into discrete, well-defined manipulation steps and is evaluated using additive scoring.
  Grading rubric (additive):
  - $+0.1$: Grasps the bottle.
  - $+0.2$: Performs one to two successful unscrewing rotations.
  - $+0.5$: Performs at least three continuous unscrewing rotations.
  - $+0.2$: Fully removes the cap.
  - $+0.2$: Places the cap on the table.

For all tasks, we additionally report a binary task success rate, where a trial is considered successful only if the task is completed end-to-end according to the instruction.

G1 Tasks

In addition to the R1 Pro Sharpa evaluations, we evaluate cross-embodiment generalization on the Unitree G1 platform using two manipulation tasks. Both tasks are evaluated using additive scoring, reflecting their decomposition into discrete, well-defined sub-skills.

- **Pen in Bin.** *Text Instruction: "Marker canister task."* The robot must open a canister, pick up a marker, and place the marker into the canister.
  Grading rubric (additive):
  - $+0.25$: Picks up or opens the canister.
  - $+0.25$: Places the canister down stably.
  - $+0.25$: Picks up the marker.
  - $+0.25$: Places the marker into the canister.
- **Dish Handover in Rack** *Text Instruction: "Put plates on dishrack."* The robot is presented with three plates on a table and must transfer them, potentially between hands, and place them upright into a dish rack.
  Grading rubric (additive, per plate): For each plate, we assign:
  - 0.11: Picks up the plate.
  - 0.11: Transfers the plate between hands.
  - 0.11: Places the plate upright into the dish rack.

  The final task completion score is computed by summing scores across all three plates, yielding a maximum score of 1.0 per trial.

C. Details of In-the-wild Human Data Curation

C.1. Dataset Statistics and Analysis

This section provides a detailed analysis of the large-scale egocentric human activity data used for Stage I pretraining. Figure 10 summarizes the dataset from multiple complementary perspectives, including statistical distributions shown in the top row and qualitative egocentric examples shown in the bottom row. Specifically, the upper-left panel visualizes the object vocabulary, the upper-middle panel shows the category distribution, the upper-right panel presents the distribution of environments, and the rightmost panel in the second row illustrates the task distribution. Together, these statistics reveal long-tailed distributions across categories, environments, tasks, and objects, highlighting both the scale and diversity of real-world manipulation behaviors captured in the dataset.

C.2. Category Distribution

The category-level distribution of total video hours (only more than 50 hours categories are counted in the plot) is shown in the upper-middle panel of Fig. 10. The dataset is dominated by retail and consumer goods

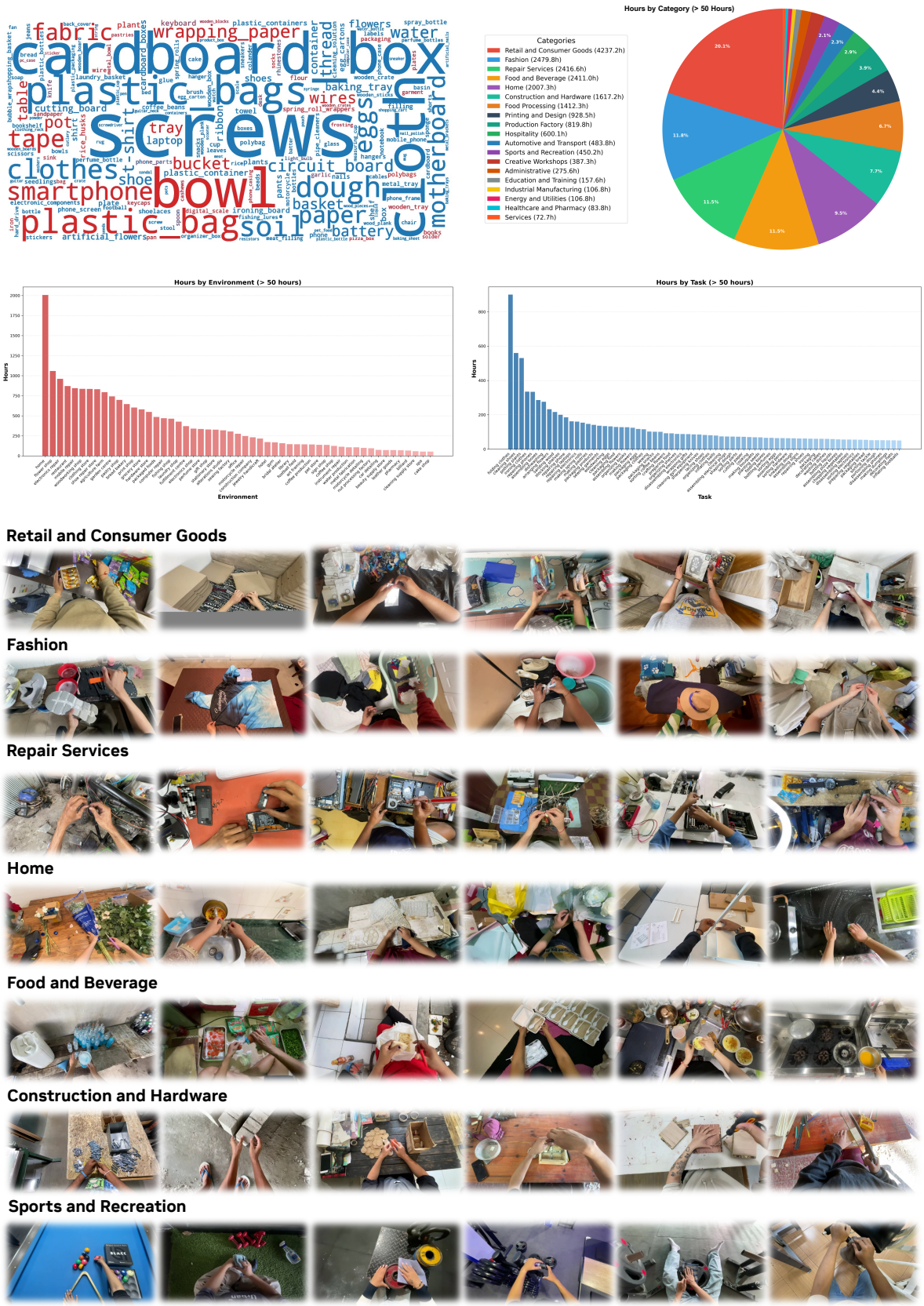


Figure 10: Statistical distributions and qualitative examples of the egocentric human activity dataset, showing long-tailed coverage across categories, environments, tasks, and objects.

(20.1%), followed by fashion (11.8%), repair services (11.5%), and food and beverage (11.5%). Additional substantial portions come from home environments (9.5%), construction and hardware (7.7%), food processing (6.7%), and printing and design (4.4%), while the remaining categories form a long tail spanning hospitality, automotive and transport, sports and recreation, education, healthcare, energy, and industrial domains.

C.3. Environment Diversity

The distribution of data across environments with more than 50 hours of recordings is shown in the lower-left bar chart of Fig. 10. The most frequent environments include homes, flower shops, electronics repair shops, furniture repair shops, woodworking shops, and clothing stores. Beyond these dominant settings, dozens of additional environments—such as grocery stores, bakeries, workshops, factories, studios, libraries, and other service-oriented spaces—contribute smaller but non-negligible amounts of data, resulting in a long-tailed environment distribution.

C.4. Task Coverage

Task-level statistics are presented in the lower-right bar chart of Fig. 10. High-frequency tasks include folding clothes, cleaning shoes, potting plants, ironing, assembling boxes, arranging flowers, sanding wood, and food preparation. In addition, a wide range of medium- and low-frequency tasks cover packing, sorting, repairing, decorating, cutting, bottling, inflating, and disassembling. Many tasks involve multi-step procedures and sustained physical interaction, reflecting realistic, long-horizon manipulation behaviors rather than isolated pick-and-place actions.

C.5. Object Vocabulary

The object vocabulary is visualized using a word cloud in the upper-left panel of Fig. 10. Frequently occurring objects include cardboard boxes, plastic bags, bowls, trays, tape, clothes, dough, batteries, and circuit boards, alongside a long tail of less frequent objects. The diversity of object instances and their co-occurrence with a wide range of tasks indicate strong compositional structure, supporting the learning of transferable action representations across objects, tasks, and environments.

D. Hand Retargeting Algorithm

We retarget 21 human hand keypoints (represented as 25 keypoints per hand with 3D position and orientation) into the 22-DoF joint space of the Sharpa Hand [29] using a per-frame, optimization-based procedure. The robot hand is modeled using URDF-based forward kinematics, which maps joint angles to 20 robot keypoint poses (positions and quaternions).

For each hand and each timestep, we solve a nonlinear program over the 22 joint angles, subject only to joint limits from the URDF, and minimize a weighted combination of different objectives. The optimization is implemented in CasADi and solved using IPOPT, warm-started from the previous frame's solution. The resulting joint angles are further smoothed using a first-order exponential filter, to reduce temporal jitter. This design enforces joint limits and kinematic consistency while preserving finger articulation and pinch/fist semantics, and yields the 22-DoF action space used for pretraining and for interfacing with the target robot.

D.1. Additional Details on Cross-Embodiment Transfer to G1

This subsection provides additional implementation details on how we adapt the human-pretrained policy to the Unitree G1 robot, which differs substantially from the primary Galaxea R1 Pro platform in both kinematics and hand actuation.

Shared Wrist Action.
Across all embodiments, we represent arm motion using relative end-effector poses in $SE(3)$, defined by frame-to-frame wrist transformations. This representation is shared between human demonstrations and robot

executions, and remains invariant to differences in absolute workspace, camera placement, and arm kinematics.

Hand Action Adaptation.
Human pretraining is performed in a 22-DoF dexterous hand joint space obtained via retargeted human hand motion. To support robots with different embodiments, such as the Unitree G1 with a 7-DoF tri-finger hand, we introduce lightweight embodiment-conditioned MLP adapters at both the input and output interfaces of the DiT action module, following the design of GR00T-N1 and N1.5 [19]. Specifically, embodiment-specific encoders map robot proprioceptive state and noisy action inputs into a shared latent action space, while embodiment-specific decoders map the DiT outputs back to the corresponding joint action space. The vision–language backbone and the DiT action expert are fully shared across all embodiments.

Embodiment-Specific Mid-Training.
To ground the human-pretrained representations in the G1 control space, we include the G1 robot play data during the aligned mid-training stage. During this stage, only the vision encoder, DiT action expert, and state-action encoder and decoder are updated, while the vision-language backbone remains frozen. This design allows the model to preserve human-derived manipulation structure while adapting to the G1's sensing and actuation interfaces.

Discussion.
Importantly, the G1 is never trained from scratch. Instead, mid-training serves to align an already learned, human-derived manipulation representation with a new embodiment. As shown in Section 3, this approach yields substantially higher performance than training directly on G1 data alone, indicating that large-scale human pretraining provides a reusable and embodiment-agnostic motor prior that can be efficiently adapted to robots with different kinematics and hand designs.

References

[1] OpenAI: Marcin Andrychowicz, Bowen Baker, Maciek Chociej, Rafal Jozefowicz, Bob McGrew, Jakub Pachocki, Arthur Petron, Matthias Plappert, Glenn Powell, Alex Ray, et al. Learning dexterous in-hand manipulation. *The International Journal of Robotics Research*, 39(1):3–20, 2020. 12

[2] Qingwei Ben, Feiyu Jia, Jia Zeng, Junting Dong, Dahua Lin, and Jiangmiao Pang. Homie: Humanoid loco-manipulation with isomorphic exoskeleton cockpit, 2025. URL https://arxiv.org/abs/2502.13013. 10

[3] Samarth Brahmbhatt, Ankur Handa, James Hays, and Dieter Fox. Contactgrasp: Functional multi-finger grasp synthesis from contact. in 2019 ieee. In *RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2386–2393, 2019. 12

[4] 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. 11

[5] Enric Corona, Albert Pumarola, Guillem Alenya, Francesc Moreno-Noguer, and Grégory Rogez. Ganhand: Predicting human grasp affordances in multi-object scenes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5031–5041, 2020. 12

[6] Hongkai Dai, Anirudha Majumdar, and Russ Tedrake. Synthesis and optimization of force closure grasps via sequential semidefinite programming. In *Robotics Research: Volume 1*, pages 285–305. Springer, 2017. 11

[7] Jensen Gao, Annie Xie, Ted Xiao, Chelsea Finn, and Dorsa Sadigh. Efficient data collection for robotic manipulation via compositional generalization. *arXiv preprint arXiv:2403.05110*, 2024. 11

[8] 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. 4

[9] Yingdong Hu, Fanqi Lin, Pingyue Sheng, Chuan Wen, Jiacheng You, and Yang Gao. Data scaling laws in imitation learning for robotic manipulation. *arXiv preprint arXiv:2410.18647*, 2024. 11

[10] Hanwen Jiang, Shaowei Liu, Jiashun Wang, and Xiaolong Wang. Hand-object contact consistency reasoning for human grasps generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11107–11116, 2021. 12

[11] Dmitry Kalashnikov, Alex Irpan, Peter Pastor, Julian Ibarz, Alexander Herzog, Eric Jang, Deirdre Quillen, Ethan Holly, Mrinal Kalakrishnan, Vincent Vanhoucke, et al. Scalable deep reinforcement learning for vision-based robotic manipulation. In *Conference on robot learning*, pages 651–673. PMLR, 2018. 11

[12] Simar Kareer, Dhruv Patel, Ryan Punamiya, Pranay Mathur, Shuo Cheng, Chen Wang, Judy Hoffman, and Danfei Xu. Egomimic: Scaling imitation learning via egocentric video, 2024. URL https://arxiv.org/abs/2410.24221. 1, 11

[13] Simar Kareer, Karl Pertsch, James Darpinian, Judy Hoffman, Danfei Xu, Sergey Levine, Chelsea Finn, and Suraj Nair. Emergence of human to robot transfer in vision-language-action models, 2025. URL https://arxiv.org/abs/2512.22414. 11

[14] Corey Lynch, Mohi Khansari, Ted Xiao, Vikash Kumar, Jonathan Tompson, Sergey Levine, and Pierre Sermanet. Learning latent plans from play. In *Conference on robot learning*, pages 1113–1132. Pmlr, 2020. 11

[15] Arjun Majumdar, Karmesh Yadav, Sergio Arnaud, Jason Ma, Claire Chen, Sneha Silwal, Aryan Jain, Vincent-Pierre Berges, Tingfan Wu, Jay Vakil, et al. Where are we in the search for an artificial visual cortex for embodied intelligence? *Advances in Neural Information Processing Systems*, 36:655–677, 2023. 11

[16] Anusha Nagabandi, Kurt Konolige, Sergey Levine, and Vikash Kumar. Deep dynamics models for learning dexterous manipulation. In *Conference on robot learning*, pages 1101–1112. PMLR, 2020. 12

[17] Suraj Nair, Aravind Rajeswaran, Vikash Kumar, Chelsea Finn, and Abhinav Gupta. R3m: A universal visual representation for robot manipulation. *arXiv preprint arXiv:2203.12601*, 2022. 11

[18] Dantong Niu, Yuvan Sharma, Giscard Biamby, Jerome Quenum, Yutong Bai, Baifeng Shi, Trevor Darrell, and Roei Herzig. Llarva: Vision-action instruction tuning enhances robot learning. *arXiv preprint arXiv:2406.11815*, 2024. 11

[19] NVIDIA, :, Johan Bjorck, Fernando Castañeda, Nikita Cherniadev, 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. URL https://arxiv.org/abs/2503.14734. 5, 18

[20] Open X-Embodiment Collaboration et al. Open X-Embodiment: Robotic learning datasets and RT-X models. International Conference on Robotics and Automation, 2024. 11

[21] Jean Ponce, Steve Sullivan, J-D Boissonnat, and J-P Merlet. On characterizing and computing three-and four-finger force-closure grasps of polyhedral objects. In *[1993] Proceedings IEEE International Conference on Robotics and Automation*, pages 821–827. IEEE, 1993. 11

[22] Jean Ponce, Steve Sullivan, Attawith Sudsang, Jean-Daniel Boissonnat, and Jean-Pierre Merlet. On computing four-finger equilibrium and force-closure grasps of polyhedral objects. *The International Journal of Robotics Research*, 16(1):11–35, 1997. 11

[23] Domenico Prattichizzo, Monica Malvezzi, Marco Gabiccini, and Antonio Bicchi. On the manipulability ellipsoids of underactuated robotic hands with compliance. *Robotics and Autonomous Systems*, 60(3): 337–346, 2012. 11

[24] Ryan Punamiya, Dhruv Patel, Patcharapong Aphiwetsa, Pranav Kuppili, Lawrence Y Zhu, Simar Kareer, Judy Hoffman, and Danfei Xu. Egobridge: Domain adaptation for generalizable imitation from egocentric human data. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*. 1

[25] Ri-Zhao Qiu, Shiqi Yang, Xuxin Cheng, Chaitanya Chawla, Jialong Li, Tairan He, Ge Yan, David J. Yoon, Ryan Hoque, Lars Paulsen, Ge Yang, Jian Zhang, Sha Yi, Guanya Shi, and Xiaolong Wang. Humanoid policy ~ human policy. In *9th Annual Conference on Robot Learning*, 2025. URL https://openreview.net/forum?id=Tx54fkQ3Cq. 1, 11

[26] Alberto Rodriguez, Matthew T Mason, and Steve Ferry. From caging to grasping. *The International Journal of Robotics Research*, 31(7):886–900, 2012. 11

[27] Carlos Rosales, Raúl Suárez, Marco Gabiccini, and Antonio Bicchi. On the synthesis of feasible and prehensile robotic grasps. In *2012 IEEE international conference on robotics and automation*, pages 550–556. IEEE, 2012. 11

[28] Lin Shao, Fabio Ferreira, Mikael Jorda, Varun Nambiar, Jianlan Luo, Eugen Solowjow, Juan Aparicio Ojea, Oussama Khatib, and Jeannette Bohg. Unigrasp: Learning a unified model to grasp with multifingered robotic hands. *IEEE Robotics and Automation Letters*, 5(2):2286–2293, 2020. 12

[29] Sharpa Robotics. Sharpa wave. https://www.sharpa.com/pages/wave, 2024. Accessed: 2026-01-31. 3, 17

[30] Tony Tao, Mohan Kumar Srirama, Jason Jingzhou Liu, Kenneth Shaw, and Deepak Pathak. Dexwild: Dexterous human interactions for in-the-wild robot policies, 2025. URL https://arxiv.org/abs/2505.07813. 1, 11

[31] 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. 11

[32] Dylan Turpin, Liquan Wang, Eric Heiden, Yun-Chun Chen, Miles Macklin, Stavros Tsogkas, Sven Dickinson, and Animesh Garg. Grasp'd: Differentiable contact-rich grasp synthesis for multi-fingered hands. In *European Conference on Computer Vision*, pages 201–221. Springer, 2022. 12

[33] Homer Walke, Kevin Black, Abraham Lee, Moo Jin Kim, Max Du, Chongyi Zheng, Tony Zhao, Philippe Hansen-Estruch, Quan Vuong, Andre He, Vivek Myers, Kuan Fang, Chelsea Finn, and Sergey Levine. Bridgedata v2: A dataset for robot learning at scale. In *Conference on Robot Learning (CoRL)*, 2023. 11

[34] 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. 11

[35] Albert Wu, Michelle Guo, and C Karen Liu. Learning diverse and physically feasible dexterous grasps with generative model and bilevel optimization. *arXiv preprint arXiv:2207.00195*, 2022. 12

[36] Tete Xiao, Ilija Radosavovic, Trevor Darrell, and Jitendra Malik. Masked visual pre-training for motor control. *arXiv preprint arXiv:2203.06173*, 2022. 11

[37] Annie Xie, Lisa Lee, Ted Xiao, and Chelsea Finn. Decomposing the generalization gap in imitation learning for visual robotic manipulation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3153–3160. IEEE, 2024. 11

[38] Mengda Xu, Zhenjia Xu, Cheng Chi, Manuela Veloso, and Shuran Song. Xskill: Cross embodiment skill discovery. In *Conference on robot learning*, pages 3536–3555. PMLR, 2023. 11

[39] Mengda Xu, Zhenjia Xu, Yinghao Xu, Cheng Chi, Gordon Wetzstein, Manuela Veloso, and Shuran Song. Flow as the cross-domain manipulation interface. *arXiv preprint arXiv:2407.15208*, 2024. 11

[40] Mengda Xu, Han Zhang, Yifan Hou, Zhenjia Xu, Linxi Fan, Manuela Veloso, and Shuran Song. Dexumi: Using human hand as the universal manipulation interface for dexterous manipulation, 2025. URL https://arxiv.org/abs/2505.21864. 12

[41] Lixin Yang, Xinyu Zhan, Kailin Li, Wenqiang Xu, Jiefeng Li, and Cewu Lu. Cpf: Learning a contact potential field to model the hand-object interaction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11097–11106, 2021. 12

[42] Ruihan Yang, Qinxi Yu, Yecheng Wu, Rui Yan, Borui Li, An-Chieh Cheng, Xueyan Zou, Yunhao Fang, Xuxin Cheng, Ri-Zhao Qiu, et al. Egovla: Learning vision-language-action models from egocentric human videos. *arXiv preprint arXiv:2507.12440*, 2025. 1, 11

[43] Sarah Young, Dhiraj Gandhi, Shubham Tulsiani, Abhinav Gupta, Pieter Abbeel, and Lerrel Pinto. Visual imitation made easy. In *Conference on Robot learning*, pages 1992–2005. PMLR, 2021. 11

[44] Ruijie Zheng, Yongyuan Liang, Shuaiyi Huang, Jianfeng Gao, Hal Daumé III, Andrey Kolobov, Furong Huang, and Jianwei Yang. Tracevla: Visual trace prompting enhances spatial-temporal awareness for generalist robotic policies. *arXiv preprint arXiv:2412.10345*, 2024. 11

[45] Ruijie Zheng, Jing Wang, Scott Reed, Johan Bjorck, Yu Fang, Fengyuan Hu, Joel Jang, Kaushil Kundalia, Zongyu Lin, Loïc Magne, Avnish Narayan, You Liang Tan, Guanzhi Wang, Qi Wang, Jiannan Xiang, Yinzhen Xu, Seonghyeon Ye, Jan Kautz, Furong Huang, Yuke Zhu, and Linxi Fan. FLARE: Robot learning with implicit world modeling. In *9th Annual Conference on Robot Learning*, 2025. URL https://openreview.net/forum?id=HXJ6pUSn1L. 11

[46] Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, pages 2165–2183. PMLR, 2023. 11