In-N-On: Scaling Egocentric Manipulation with in-the-wild and on-task Data

Xiongyi Cai* Ri-Zhao Qiu*† Geng Chen Lai Wei
Isabella Liu Tianshu Huang Xuxin Cheng Xiaolong Wang
UC San Diego
* Equal Contribution † Project Lead
https://xiongyicai.github.io/In-N-On

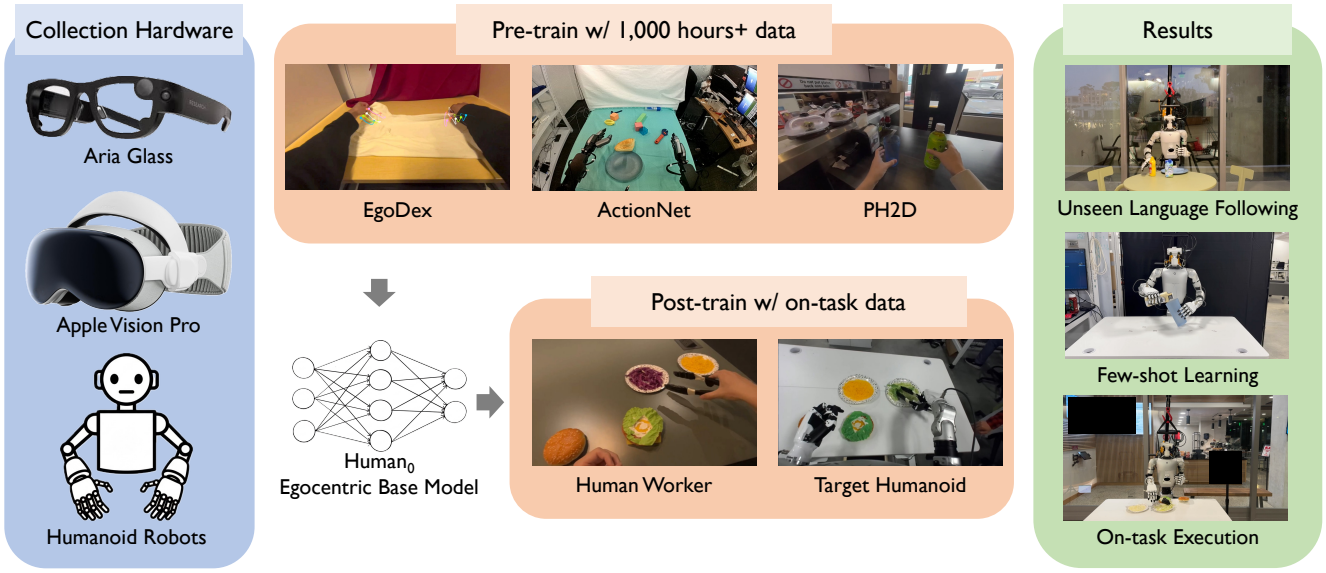


Figure 1. This paper investigates large-scale pre-training and post-training with egocentric human data. We curate a large-scale **P**hysical **H**uman-humanoid**S** **D**ataset, dubbed PH^{S}D, to train a base model to model egocentric human-humanoid behavior. Empirically, we show that Human$_0$ achieves several interesting properties, including strong language following of instructions unseen in robot data, few-shot execution, and improved on-task performance.

Abstract

*Egocentric videos are a valuable and scalable data source to learn manipulation policies. However, due to significant data heterogeneity, most existing approaches utilize human data for simple pre-training, which does not unlock its full potential. This paper first provides a scalable recipe for collecting and using egocentric data by categorizing human data into two categories: **in-the-wild** and **on-task** alongside with systematic analysis on how to use the data. We first curate a dataset, PH^{S}D, which contains over 1,000 hours of diverse in-the-wild egocentric data and over 20 hours of on-task data directly aligned to the target manipulation tasks. This enables learning a large egocentric language-conditioned flow matching policy, Human$_0$. With domain adaptation techniques, Human$_0$ minimizes the gap between humans and humanoids. Empirically, we show Human$_0$ achieves several novel properties from scaling human data, including language following of instructions from only human data, few-shot learning, and improved robustness using on-task data.*

1. Introduction

The robot manipulation community has recently witnessed great progress in learning from real robot demonstrations [4, 9, 23, 26, 31, 61]. Behind the curtain are novel algorithms [23] and large-scale robot data [9, 38], which enable dexterous and long-horizon tasks [4]. However, existing foundational manipulation policies still lack *robust real-world generalizability* compared to their counterparts

in LLM [1] or self-driving [46] that are trained on much larger-scale data.

In search of a novel data source to fuel model training, researchers have turned to cross-embodiment learning from different robots [9, 25, 38], and, more recently, to human data [3, 18, 19, 33]. Intuitively, humans are naturally the most prominent physical embodiment compared to other morphologies that can easily manipulate daily objects. Thus, learning from human data has been studied for over a decade. Modular methods learn affordance [2, 35, 51] and plan robot manipulation in a model-based fashion [29, 40, 54]. More recently, advances in computer vision have enabled precise finger keypoint tracking to generate human data with action labels. Recent methods [24, 27, 30, 32, 37, 40, 42, 56] have shown that such high-quality data can be directly used for end-to-end training, which has the potential to be easily scaled up.

However, the vast amount of human data also leads to significant data heterogeneity. Existing human datasets are very diverse - ranging from daily activities such as walking and dancing [18], long-horizon kitchen activities [13], and even sitcoms [51]. To address such heterogeneity (or misalignment between embodiments), some methods propose new algorithms to use intermediate representations such as object pose [29] or affordance [2] to learn from these **in-the-wild** datasets. On the other hand, recent end-to-end approaches [7, 32, 56] have resolved to scaling up pre-training with human data, and then fine-tuning with robot data. This approach is usually sub-optimal due to catastrophic forgetting [16, 20] from simple fine-tuning with highly heterogeneous data.

On the other hand, recent methods [24, 42, 44, 49, 55] have also focused on collecting **on-task** human data. Instead of recording casual activities, on-task data collection focuses on curating human demonstrations on the same tasks that robots will be working on (*e.g.,* recorded by actual human workers). Compared to in-the-wild data, on-task data are more task-oriented, in-domain, and segmented well. These factors ensure good alignment to the target deployment distribution, which has been empirically shown to enable direct co-training of mixed humans and robot data [24, 42, 44, 55] to mitigate catastrophic forgetting from pre-training.

The goal of this paper is to show that it is important to use both in-the-wild and on-task data to unlock the full potential of human data: in-the-wild data is easy to collect and diverse, but it may be only suitable for bootstrapping a base model. In contrast, on-task data is more well-aligned with the target distribution but often smaller in magnitude.

To this end, we investigate the boundary between these two paradigms. Our insight is to use **in**-the-wild data **and on**-task data for pre-training and post-training. With language annotations and a unified human-centric action space [42], this enables learning of a large egocentric language-conditioned flow matching policy, Human_0. In addition to scaling up egocentric training data, we perform systematic study to reveal that naïve data mixing leads to hidden states that discriminate robot and human inputs. Human_0 adopts domain adaptation technique to improve hidden states to fully utilize human training data.

We evaluate Human_0 on a real Unitree H1 humanoid and a Unitree G1 humanoid equipped with 5-fingered dexterous hands. Empirically, the pre-training and post-training for Human_0 achieve several novel properties, including language following of instructions that are unseen in the robot training data and few-shot learning, which is validated by systematic ablation studies. In particular, we studied a task, *fast food worker*, where data can be collected at a low marginal cost from real food-industry worker. We show how the on-task data collected for this practical scenario improves policy robustness drastically.

In sum, our contributions are,

- A large-scale human-humanoid dataset, PH^SD, that provides data recipe for pre-training and post-training an egocentric model. We plan to open-source the dataset.
- A base egocentric manipulation model, Human_0, which is augmented with the domain adaptation technique and applicable to many egocentric bimanual embodiments. The weights will be open-sourced.
- Extensive experimental results with demonstrations of language following and few-shot learning on real humanoid robots.

2. Related Work

Large-scale Manipulation Models. Recent advances in vision-language-action (VLA) models have shown promising progress in robotic manipulation tasks, with a growing emphasis on models' robustness and generalization. Building upon early efforts in learning from real-robot demonstrations [12, 60], recent methods [4, 9, 23, 25, 31, 45, 63] explored how to scale up robot manipulation policy training with more data. The advances happened both in the modeling regime and the data regime. In the context of modeling, VLAs extend vision-language models (VLMs) or large-language models (LLMs) with action decoders to make use of pre-trained knowledge infused in VLMs. More recently, Intelligence et al. [23] also proposed a new paradigm to make the training process more data-efficient. On the other hand, data is important for scaling up the manipulation model. Notably, many large manipulation models [9, 31] rely on cross-embodiment learning [38], where a model designs its architecture specifically to work with data from multiple robot embodiments. However, even with cross-embodiment learning, the magnitude of available data is still significantly smaller compared to counterparts in language or vision models. Current manipulation models are

data-hungry for more generalizability.

Learning from Human Videos. Learning robot policies from human videos has been an active research direction, driven by the availability of large-scale human data. Early efforts [34, 36, 43] focused on leveraging human videos to pre-train visual representations that are better suited for downstream manipulation policy learning; or to leverage human videos to learn intermediate representations such as affordance [2]. Beyond pre-training on visual tasks for improved initializations, other works [2, 5, 6, 48, 53, 54] attempt to use human data directly for downstream tasks such as point tracking [6, 28, 53], and high-level planner [48], which are then used to guide robot action prediction.

End-to-end Learning Manipulation Policies from Human. An increasing number of works have started to investigate scaling manipulation in an end-to-end manner by leveraging human demonstrations [7, 24, 30, 39, 42, 44, 57, 64]. They either use diverse **in-the-wild** data for pre-training [7, 30, 32, 56] or **on-task** data for co-training. Notably, Bi et al. [7], Li et al. [30], Luo et al. [32] have shown pre-training with human data leads to improved generalizability; Lepert et al. [27] apply modular vision modules to edit human videos to match robot videos to reduce visual gaps. Concurrently, EMMA [64] learns a mobile manipulation policy using human data. However, there has yet to be an attempt to explore both in-the-wild data and on-task data to cover both pre-training and post-training stages. This paper aims to bridge such a gap by prescribing a recipe for data curation, an end-to-end large egocentric manipulation base model, and algorithmic advances to improve the model.

3. Method

This paper discusses models and data recipes for pre-training and post-training a base model for egocentric manipulation, as well as analysis of design decisions made to create the recipes. Sec. 3.1 describes the curation process for a large human-humanoid dataset. Sec. 3.2 discusses the design choices for the base model, including data mixture and domain adaptation.

3.1. PH^{S}D: Physical Humans-Humanoids Dataset

Many human datasets [14, 22, 33, 42, 52, 58] and egocentric robot datasets [15, 62] exist. Naturally, the formats of the human dataset are similar - most existing datasets focus on tracking head, wrists, and fingers poses. However, humanoid hardwares, or robot hardware in general, are far from convergence. Therefore, these publicly available egocentric robot datasets are vastly different - kinematics, DoFs, and mechanical configurations can differ. The difference in state-action space in each dataset hinders scaling up the training size.

To tackle this, existing methods attempted to design physically explainable state-action space [31] or operate in the unified latent space [50]. However, scaling data in the same state-action space remains the most explainable and effective way [42, 45]. For egocentric manipulation, this paper advocates the human representation for learning, as humans are the most prevalent embodiment and are the sources of biological inspiration for bimanual robot designs.

To this end, this paper defines a unified human-centric state-action space. We then implement a software suite of robot IK/FK (Inverse Kinematics and Forward Kinematics), and hand retargeting algorithms to differentiably convert human and humanoid data from/to our unified space. Finally, we curate and process data from multiple sources into a unified format for training.

Unified human-centric state-action space. Following human activities datasets [22, 33, 52] and Human-Humanoid co-learning [42]. We design the state-action space to have the following elements.

- $\mathbf{T_{head}} \in \mathbf{SE}(3)$. We parameterize head poses as the base transformation with rotation and translation. Compared to previous work [42], this further encodes translation to support potential applications such as whole-body loco-manipulation. Current egocentric human data is usually collected by wearable devices mounted on operators' head. Hence the localization frame is usually modeled as world-head transformation.
- $\mathbf{T_{Lwrist}}, \mathbf{T_{Rwrist}} \in \mathbf{SE}(3)$. The wrist poses are modeled as relative to the head pose. Though there is inevitable physical difference (*e.g.,* height) among different data collectors, such a difference can be neglected as the training data scales.
- $\mathbf{P_{Lfinger}}, \mathbf{P_{Rfinger}} \in \mathbb{R}^{3\times5}$: we model finger motions as fingertip keypoints, as all well-established optimization-based finger retargeting algorithms [11, 21, 41] directly use fingertip keypoints with scaling factors.
- $\mathbf{D_{Lgripper}}, \mathbf{D_{Rgripper}} \in \mathbb{R}$: optionally, for bimanual robots equipped with parallel grippers, we map gripper distance to human thumb-index fingertip distance. Note that $\mathbf{P_{Finger}}$ is sufficient for computing $\mathbf{D_{gripper}}$.

Retargeting Software Suite. To make it easy for us and the research community to explore the human-centric state-action space for egocentric manipulation, we implement an IK/FK and retargeting software suite based on Pinocchio [10]. Our suite converts the robot joint positions from/to human-centric space (Fig. 3 visualizes same human action retargeted to different humanoids in the accompanying MuJoCo [47] simulator in our suite). Therefore, as long as a released humanoid manipulation dataset provides joint readings, it can be used for training. We hope to facitate large-scale egocentric learning on humanoid robot. The code and assets will be open-sourced.

Aggregating In-the-wild datasets for Pre-training. Building on our retargeting framework, we use EgoDex [22], Fourier ActionNet [15], and PH2D [42]

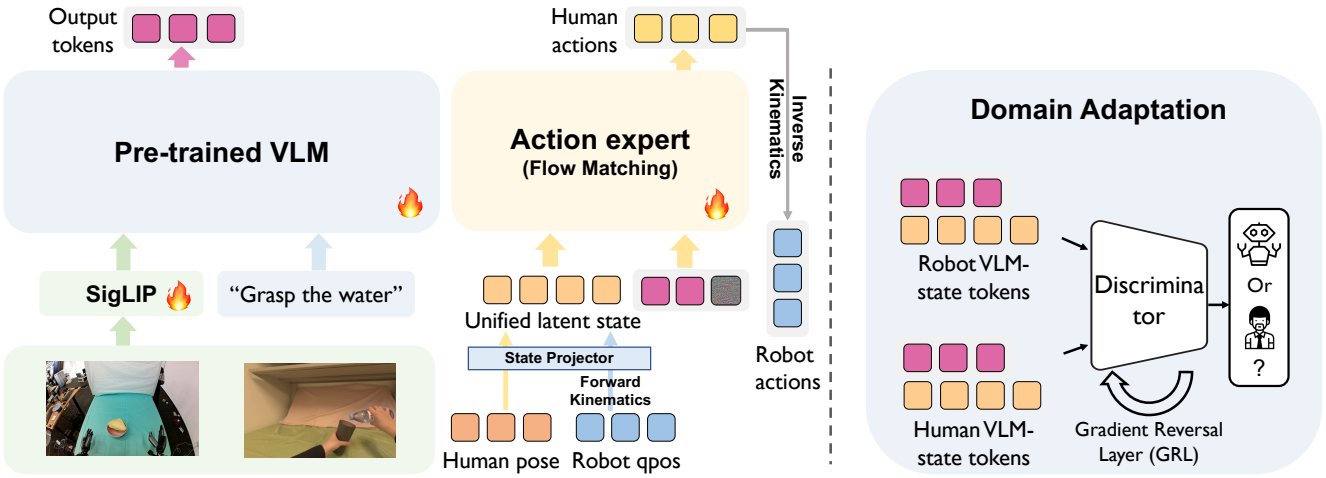


Figure 2. Method overview. Our approach follows a two-stage training recipe: (1) pre-training on large-scale in-the-wild human and robot data that are mapped into a unified human-centric state-action space; and (2) on-task post-training using task-aligned human and robot demonstrations. To bridge the embodiment gap, We employ a domain-adversarial discriminator that takes SigLIP visual features and action-state embeddings as input and predicts whether a sample is from human or robot data. Through gradient reversal, this encourages the policy's encoders to produce embodiment-invariant representations, enabling effective transfer between human and robot observations.

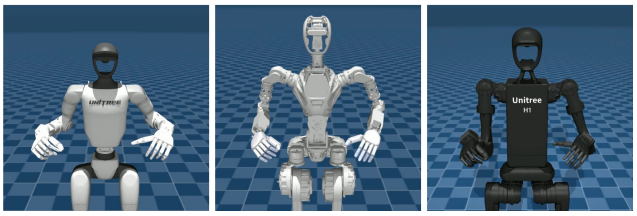


Figure 3. Our retargeting software suite supports retargeting different humanoids from/to the human-centric representation. Figure demonstrates retargeting from the same human action to different humanoids in MuJoCo [47]. The code will be released.

for pre-training. Note that with our software suite, our method also applies to future and concurrent dataset [62].

- EgoDex [22] contains 800+ hours of skill-rich human demonstrations, which were collected using multiple Apple Vision Pros. It contains 6dof head pose, wrist pose, and finger keypoints.
- The ActionNet dataset [15] contains over 100 hours of humanoid demonstrations - most of which were done on the Fourier GR1T1 robot embodiment equipped with bimanual Fourier 5-fingered 6-DoF dexterous hands.
- The PH2D [42] dataset contains human and humanoid demonstrations of various tasks. Similar to EgoDex [22], PH2D also collected human data with Apple Vision Pro, which can be processed in a similar manner. The humanoid data (collected on Unitree H1 with 5-fingered Inspire hands) are also processed by our software suite.

Data for Post-training. To ensure high-quality hand poses in our on-task datasets, we use commercial-grade data collection devices, including Apple Vision Pro and the Meta Aria Glass. Both Vision Pro and Meta Aria glass provide head poses, wrist poses and dense fingertip keypoint predictions. The human dataset also includes 2D keypoint projection, as we hope our released data can also help other approaches such as generative inpainting [27, 59]. The robot dataset are collected using Apple Vision Pro with OpenTV [11]. Visualizations and projections of these datasets can be found in the supplementary material.

3.2. Human$_0$: Foundational Egocentric Base Model

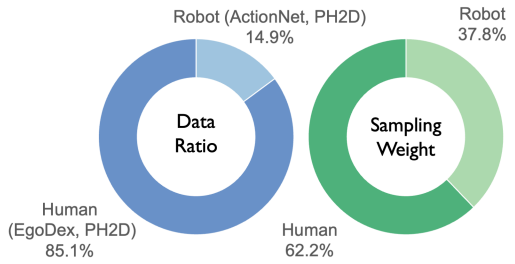
3.2.1. Architecture

While the pre-training and post-training recipes proposed in this paper are model-agnostic, we adopt a language-conditioned flow matching model [9]. Specifically, a SigLIP-based vision module extracts visual tokens $v \in \mathbb{R}^{L\times C}$, where L is the number of patches and C is the embedding dimension. The SigLIP encoder provides strong alignment between visual inputs and text, enabling downstream instruction grounding. Visual tokens are then combined with text embeddings $n \in \mathbb{R}^{T\times C}$ to form a joint multi-modal representation, which is further processed in the transformer blocks to propagate cross-modal context.

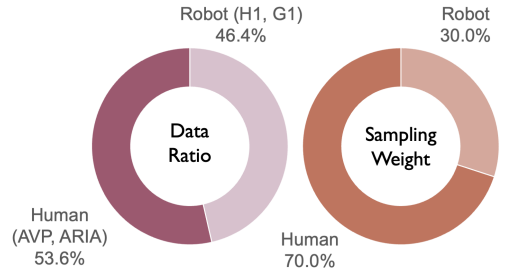
To use the human-centric representation, we use lightweight MLPs to encode input states and the output actions. For the input states, the physically interpretable human-centric state is projected to a pose latent $x \in \mathbb{R}^C$. We denote the latent tokens produced by the backbone transformer as

$$z = \text{Transformers}(v, n, x), \quad z \in \mathbb{R}^C, \tag{1}$$

which integrate information across modalities. Unless oth-



(a) Data size ratio and sampling factor for **pre-training** data.



(b) Data size ratio and sampling factor for **post-training** data.

Figure 4. Data distributions and sampling factors for pre-training and post-training.

erwise noted, we use the pre-trained checkpoint released by Black et al. [9] to initialize the model pre-training. Note that since the human-centric representations introduce larger vector sizes and different interpretations of each element at different indices, we swap out the original projection modules with different dimensions and random initialization.

3.2.2. Pre-training on Human and Robot Data

We first pretrain the base model using over 1,000+ hours of mixed data from EgoDex [22], ActionNet [15], and PH2D [42], covering rich egocentric human and robot manipulation scenarios. During this stage, the objective is to learn a unified vision–language–action prior that models human-like behaviors across different embodiments.

More concretely, let $a \in \mathbb{R}^C$ be the target action, the model is trained with a flow-matching objective in an end-to-end manner:

$$\mathcal{L}_{\text{FM}} = \mathbb{E}\Big[\| f_\theta^{\text{flow}}(z, a_t, t) - (a - u) \|_2^2 \Big], \qquad (2)$$

where the Gaussian noise vector $u \sim \mathcal{N}(0, I_C)$, time step $t \sim U(0, 1)$, and interpolated action $a_t = (1-t)u + ta$. The $f_\theta^{\text{flow}}(z, a_t, t)$ represents the predicted flow vector, which points from the noisy sample towards the target. This pre-training stage equips the base model with broad visuomotor priors from the vast amount of human videos. In addition, it aligns the VLM originally trained on image-text data to model human behavior. The shared embodiment space provides a strong regularization for post-training, enabling effective transfer between human and robot manipulation.

Data mixing recipe. The distribution of the pre-training data is presented in Fig. 4a. Note that due to the overwhelming amount of human data in pre-training, we manually adjust the training data sampler ratio to balance and stabilize the training process.

3.2.3. Post-training on Human and Robot Data

During post-training, we focus exclusively on human and robot data collected for the task of interest. The goal is to refine the policy's language grounding and visuomotor control to match the distribution of real-world tasks, where data can be collected by actual human workers performing these real-world tasks. Thanks to our unified action space design, the training procedure follows Eq. (2) precisely.

Data mixing recipe. The distribution of post-training data is presented in Fig. 4b. Compared to pre-training, our post-training dataset has considerably more robot data. Empirically, we found that sampling slightly more often (*e.g.,* 70%) from the human data helps preserve semantics in human data better. This finding is somewhat consistent with Tao et al. [44], which used an 8:2 sampling ratio to sample human data more often.

In summary, our pre-training and post-training process enables various interesting properties, including (1) language following of instructions unseen in robot data; (2) few-shot robot data learning with as few as 1 demonstration; and (3) improved robustness across related tasks.

3.2.4. Domain Adaptation: Blurring the Line between Embodiments

Ideally, our model should be embodiment-agnostic and process all egocentric data from a human-centric perspective. However, though we use image augmentation and human-centric representation with forward kinematics to provide regularization, the model can still learn to distinguish different embodiments, resulting in overfitting to a specific configuration.

To verify this, we first pre-train and post-train the model using vanilla denoising objective Eq. (2). Then, we perform a simple linear probing study. We train a simple MLP taking intermediate visual tokens and proprioceptive tokens as inputs. The training objective for the MLP is a binary classification problem, where it tries to predict if a set of concatenated visual and proprioceptive tokens belongs to human data or robot data. The results are shown in Fig. 5. Surprisingly, on a held-out validation set, the simple MLP achieves 100% success rate - suggesting that the model 'cheats' by implicitly biasing features to recognize if the input is human or robot. (More technical details are given in the sup-

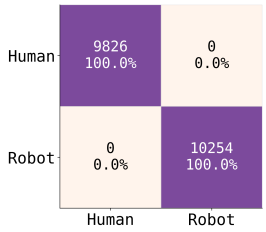


Figure 5. Confusion matrix obtained by linear probing intermediate features from vanilla model.

plementary material). To discourage the model from overfitting to specific visual cues or proprioceptive cues, we introduce a discriminator network [17]. Specifically, the network is tasked to classify the type of embodiment. Following Ganin et al. [17], the network is modeled as a MLP that takes in intermediate features with Gradient Reversal Layer (GRL) [17] to discourage successful classification.

More specifically, the GRL is trained to differentiate between the feature encoding of human data and those of robot data. We concatenate the visual tokens v from SigLIP encoder with the projected pose latent x along the token dimension, and pass them though a attention head to obtain a feature vector:

$$m = \text{Attn}\big(\text{Concatenate}(v, x)\big), \quad m \in \mathbb{R}^C . \tag{3}$$

The feature vector m is then passed through the discriminator MLP D_θ that predicts the input's embodiment type. The discriminator is trained with binary cross-entropy loss:

$$\mathcal{L}_D(\phi \mid \theta) = -\mathbb{E}\big[\log D_\phi(m_h)\big] - \mathbb{E}\big[\log(1 - D_\phi(m_r))\big], \tag{4}$$

where m_h and m_r denote feature vectors obtained from human and robot data,respectively. With a GRL inserted between feature vectors and discriminator D_ϕ, the optimization is adversarial: D_ϕ minimizes $\mathcal{L}_D$, while the backbone policy encoders f_θ maximizes it:

$$\max_\theta \min_\phi \ \mathcal{L}_D(\phi, \theta) \tag{5}$$

In other words, the GRL encouraging the upstream policy encoder to produce features that are invariant to the human-robot domain distinction. This adversarial setup promotes feature alignment across data domains and embodiments, enabling more effective transfer of manipulation behaviors between human demonstrations and robot.

Final Loss. Combining both flow matching L2 loss and the domain adaptation loss, the final training loss is given by

$$\mathcal{L}_{\text{final}} = \mathcal{L}_{\text{FM}} + \lambda \cdot \mathcal{L}_D(\phi \mid \theta) , \tag{6}$$

where λ is a hyperparameter balancing the scale of flow matching loss and discriminator loss. During the training, we set $\lambda = 0.1$.

4. Experiments

4.1. Experimental Setup

Implementation Details. For raw human-humanoid data, we use timestamps to synchronize episodes and process the states and actions into the human-centric representation with 240×320 images. To obtain the base Human_0 model, we train on 8 H200 GPUs for 100k steps using 160 batch size. The weights are initialized with pre-trained checkpoint [9]. For post-training, we fine-tune the trained base model on a single H100 GPU for 30k steps using 10 batch size. This demonstrates one potential application of our base model to democratize egocentric manipulation training with just a single GPU.

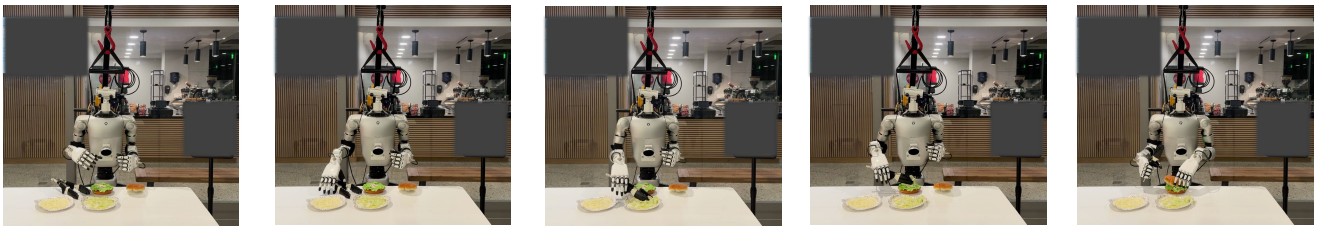
Robot Platforms. For data collection and policy deployment, we use a Unitree H1 and a Unitree G1 humanoid robot. Most of the data was collected on the G1 robot. Thus, unless otherwise stated, the data and experiments are done on the G1 robot. Both robots are equipped with Inspire 5-fingered dexterous hands.

Baselines. We compare with 4 baseline models. π_0 [9] is a language-conditioned flow matching model trained on many robot embodiments, which is also the initialization we use before pre-training. GR00T N1 [8] is another language-conditioned VLA using diffusion transformers. HAT [42] trains specialist policies and is thus unsuitable for pre-training or tasks that require language conditioning. Finally, Human_0 w/o human follows the same training procedure, but without any human data in both stages.

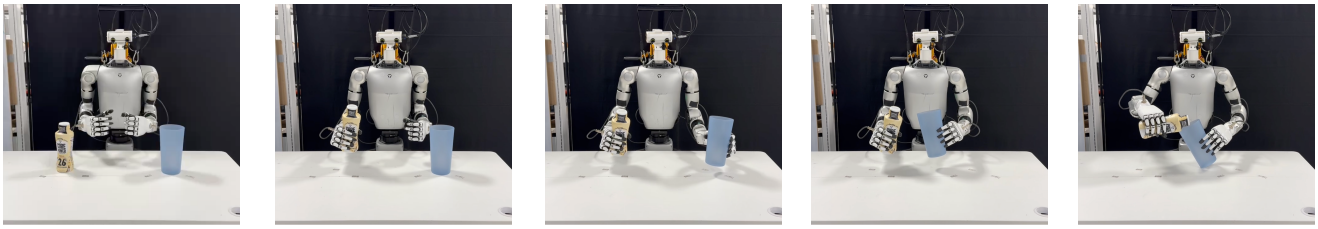
Experimental Protocol. We experiment with 4 different humanoid manipulation tasks with *in-distribution (I.D.)* and Out-Of-Distribution (O.O.D.) settings. The I.D. setting tests the learned skills with language, scenes, and objects that approximately resemble corresponding sequences in the robot training demonstrations. The O.O.D. setting tests configurations that are unseen in the robot training data, but may present in human data.

The tasks are illustrated in Fig. 6. Objects used in these tasks are visualized in the supplementary material. Specifically,

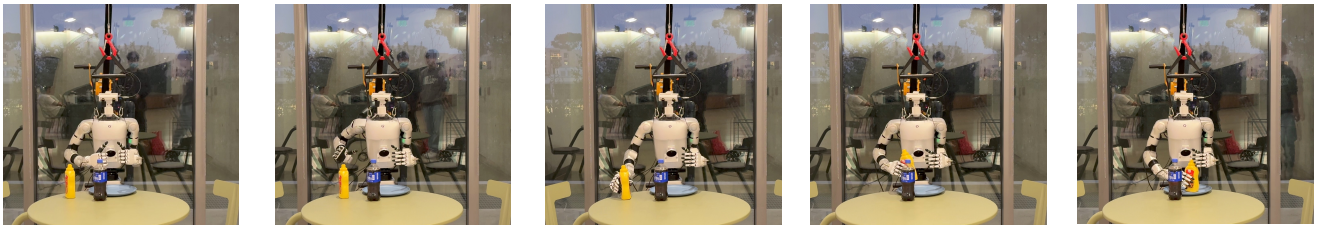
- **Single object grasping** is a sanity check task. The robot is placed in front of a table with an object and a container. The robot is tasked to pick up the object, and place it into the container. **OOD setting:** the robot is presented with objects unseen in the robot training data, different table heights, and operate in novel scenes.
- **Multi object grasping** is an extension of the single object grasping, where we add distractor objects. As shown in Fig. 6, the robot is tasked to grasp the object. The robot must follow the language instruction and distinguish the object from distractors. **OOD setting:** the robot is presented with target objects and distractors unseen in the robot training data, different table heights, and operate in novel scenes.
- **Burger assembly** is intended to mimic a real-world task, where a worker at a fast food restaurant or at a food processing facility assembles a burger based on language instructions. The task is long-horizon, which involves multiple steps from using tongs to pick up ingredients specified by language, and putting the top bread. In addition, collecting on-task human data for this application



(a) On-task performance. "Assembling the burger with lettuce."



(b) One-shot learning. "Pouring water."



(c) Following language instruction available only in human data. "Grasping the yellow mustard bottle."

Figure 6. We task the robot to perform several manipulation tasks to evaluate few-shot learning, language instruction following, and robustness using on-task human data. Videos in the supplementary. (Top to bottom: burger assembly, pouring, and multi-object grasping).

Method	Single Object Grasping		Multi Object Grasping		Burger assembly		Pouring
	I.D.	O.O.D	I.D.	O.O.D	I.D.	O.O.D	I.D.
π_0	19/20	19/20	25/30	16/30	5/12	3/12	0/20
GR00T N1	18/20	13/20	6/30	8/30	4/12	3/12	0/20
HAT w/ human	17/20	15/20	-	-	-	-	2/20
Human_0 w/o human	18/20	18/20	23/30	15/30	7/12	2/12	2/20
Human_0 (Ours)	**20/20**	**19/20**	**29/30**	**30/30**	**8/12**	**7/12**	**5/20**

Table 1. Baseline comparison results. Our method achieves the best performance among all baselines across the four manipulation tasks, under both I.D. and O.O.D. settings. We also show that training with large-scale human data improves model performance.

can be hypothetically done by having the actual workers use wearable devices. **OOD setting:** the robot is presented with ingredients unseen in the robot training data (*e.g., Mozzarella cheese*) and operate in novel scenes with different table heights.

- **Pouring** shows the few-shot learning capability of our model. Compared to previous tasks that have hundreds of robot sequences per task, we use only **1** robot training demonstration in the bimanual pouring task, to demonstrate how Human_0 enables few-shot robot learning.

4.2. Evaluation

4.2.1. Main Experiment

Zero-shot language following capability from human data. The most interesting finding is Human_0 emerges *capability to follow language instructions unseen in the robot training data*. One major weakness of existing VLAs is that they are bad at following language instructions unseen in the training data. For instance, in the multi-object grasping setting, both π_0 and GR00T N1 fail to grasp unseen objects - π_0 would randomly grasp 1 out of the 2 objects, resulting in approximately 50% success rate.

On the other hand, Human_0 is robust at following the language presented only in the human data. In the multi-object grasping experiment, the robot is capable of grasping unseen objects robustly with variations of distractors and scenes. In the burger assembly task, the robot needs to use tongs to pick up different ingredients specified by the model. Again, human data enables the model to use tools to pick up Mozzarella cheese, which is an ingredient seen only

Discriminator	Staged Pouring SR			
	Right grasp	Left grasp	Pour	SR
✗	**17/20**	5/20	3/20	15%
✓	16/20	**7/20**	**5/20**	**25%**

Table 2. Ablation study of domain adaptation using the pouring task, which is a challenging bimanual task that can be divided into 3 stages. The success rates (SR) reported are compositional.

in the human data. In sum, this capability is exciting, as it opens a door to scale up the language understanding ability of robot manipulation via egocentric human data.

Human$_0$ enables 1-shot robot data learning. Next question we asked is - where is the boundary of such a language following capability? Can the robot learn a completely new behavior from just the human data? Empirically, the answer is *no* at the current training scale. However, training with vast human data still enables few-shot learning capability. With just a single robot demonstration of the bimanual pouring task, Human$_0$ achieves 5/20 success rate. We believe that the few-shot performance would further increase with larger training data scale, which may ultimately lead to zero-shot behavior.

Human data improves overall performance on challenging task. The burger assembly task is a challenging long-horizon task that involves tool usage, working alongside distractors, and perform multiple actions. Human$_0$ outperforms baseline methods with over 100% relative improvement on this challenging task. Notably, in the O.O.D. scenarios, we intentionally task the robot to manipulate ingredients unseen in the robot data (*i.e.,* red cabbage, Mozzarella cheese, and swiss cheese) to mimic special requests to food workers in the real world. In addition to language following capability discussed above, we find that the model is more robust to external disturbances such as lighting or background changes.

4.2.2. Ablation Study

Domain adaptation prevents model from 'cheating' and helps few-shot learning. As shown in Fig. 5, the vanilla model implicitly learns to discriminate different embodiments. After adding the GRL layer, the linear probing experiment yields promising results. Fig. 7 suggests that linear probing can no longer discriminate embodiments, which is evident from the unconfident probabilities centered around 50% comparable to random guesses.

As a result, domain adaptation technique improves final robot execution. Tab. 2 demonstrates improved few-shot learning capability. Without a domain discriminator, the model quickly overfits to the single humanoid demonstration. Introducing a domain discriminator encourages the shared representation to become invariant across human and humanoid domains. As a result, the humanoid can effectively leverage priors learned from human data.

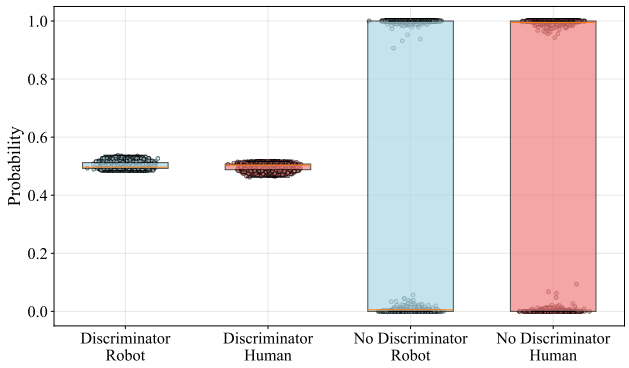


Figure 7. The probabilities produced by linear probing the intermediate features. y-axis: predicted probability of a sample being an embodiment. The model trained with a discriminator yields domain-invariant intermediate features.

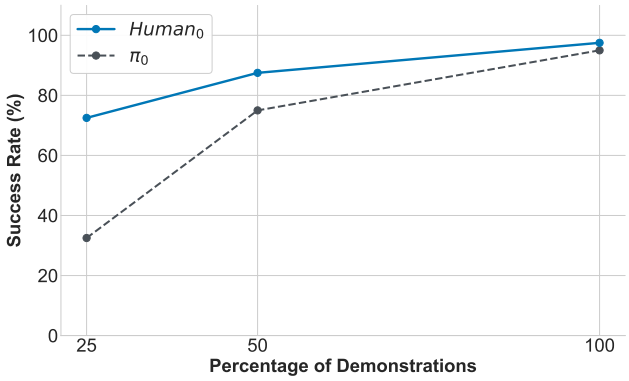


Figure 8. Performance on single object grasping. The x-axis represents the percentage of available single-object grasping robot demonstrations used in training.

Performance dynamics with robot data as a variable. Fig. 8 shows the performance change with varying number of robot data used in training on a simple single-object grasping task. Human data can effectively regularize the learning especially in low robot data regime.

Supplementary material. For more results such as detailed failure analysis and videos, we encourage the readers to check out the supplementary material.

5. Conclusion

In this work, we presented IN-N-ON, a scalable recipe for leveraging egocentric human data through a principled taxonomy that distinguishes in-the-wild and on-task data. With PH^{S}D, a large-scale dataset comprising over 1,000 hours of diverse in-the-wild human and humanoid demonstrations and 20+ hours of task-aligned data, we enabled the training of Human$_0$. Human$_0$ demonstrates several novel properties of scaling human data with language annotations. There are

several interesting future directions, including further scaling of human data and testing on different robot embodiments other than humanoid robots.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 2

[2] Shikhar Bahl, Russell Mendonca, Lili Chen, Unnat Jain, and Deepak Pathak. Affordances from human videos as a versatile representation for robotics. In *CVPR*, 2023. 2, 3

[3] Prithviraj Banerjee, Sindi Shkodrani, Pierre Moulon, Shreyas Hampali, Shangchen Han, Fan Zhang, Linguang Zhang, Jade Fountain, Edward Miller, Selen Basol, et al. Hot3d: Hand and object tracking in 3d from egocentric multi-view videos. In *CVPR*, 2025. 2

[4] Jose Barreiros, Andrew Beaulieu, Aditya Bhat, Rick Cory, Eric Cousineau, Hongkai Dai, Ching-Hsin Fang, Kunimatsu Hashimoto, Muhammad Zubair Irshad, Masha Itkina, et al. A careful examination of large behavior models for multitask dexterous manipulation. *arXiv preprint arXiv:2507.05331*, 2025. 1, 2

[5] Homanga Bharadhwaj, Abhinav Gupta, Vikash Kumar, and Shubham Tulsiani. Towards generalizable zero-shot manipulation via translating human interaction plans. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6904–6911. IEEE, 2024. 3

[6] Homanga Bharadhwaj, Roozbeh Mottaghi, Abhinav Gupta, and Shubham Tulsiani. Track2act: Predicting point tracks from internet videos enables generalizable robot manipulation. In *European Conference on Computer Vision*, pages 306–324. Springer, 2024. 3

[7] Hongzhe Bi, Lingxuan Wu, Tianwei Lin, Hengkai Tan, Zhizhong Su, Hang Su, and Jun Zhu. H-rdt: Human manipulation enhanced bimanual robotic manipulation. *arXiv preprint arXiv:2507.23523*, 2025. 2, 3

[8] Johan Bjorck, Fernando Castañeda, Nikita Cherniadev, Xingye Da, Runyu Ding, Linxi Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, et al. Gr00t n1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734*, 2025. 6

[9] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. pi0: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024. 1, 2, 4, 5, 6

[10] Justin Carpentier, Guilhem Saurel, Gabriele Buondonno, Joseph Mirabel, Florent Lamiraux, Olivier Stasse, and Nicolas Mansard. The pinocchio c++ library – a fast and flexible implementation of rigid body dynamics algorithms and their analytical derivatives. In *IEEE International Symposium on System Integrations (SII)*, 2019. 3

[11] Xuxin Cheng, Jialong Li, Shiqi Yang, Ge Yang, and Xiaolong Wang. Open-television: Teleoperation with immersive active visual feedback. In *CoRL*, 2024. 3, 4

[12] 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*, 2023. 2

[13] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The epic-kitchens dataset. In *ECCV*, 2018. 2

[14] Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J Black, and Otmar Hilliges. Arctic: A dataset for dexterous bimanual hand-object manipulation. In *CVPR*, 2023. 3

[15] Yao Mu Fourier ActionNet Team. Actionnet: A dataset for dexterous bimanual manipulation. 2025. 3, 4, 5

[16] Robert M French. Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, 3(4):128–135, 1999. 2

[17] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of machine learning research*, 2016. 6

[18] 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. In *CVPR*, 2022. 2

[19] 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 *CVPR*, 2024. 2

[20] Asher J Hancock, Xindi Wu, Lihan Zha, Olga Russakovsky, and Anirudha Majumdar. Actions as language: Fine-tuning vlms into vlas without catastrophic forgetting. *arXiv preprint arXiv:2509.22195*, 2025. 2

[21] Ankur Handa, Karl Van Wyk, Wei Yang, Jacky Liang, Yu-Wei Chao, Qian Wan, Stan Birchfield, Nathan Ratliff, and Dieter Fox. Dexpilot: Vision-based teleoperation of dexterous robotic hand-arm system. In *ICRA*, 2020. 3

[22] Ryan Hoque, Peide Huang, David J Yoon, Mouli Sivapurapu, and Jian Zhang. Egodex: Learning dexterous manipulation from large-scale egocentric video. *arXiv preprint arXiv:2505.11709*, 2025. 3, 4, 5

[23] Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, et al. pi0.5: a vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*, 2025. 1, 2

[24] Simar Kareer, Dhruv Patel, Ryan Punamiya, Pranay Mathur, Shuo Cheng, Chen Wang, Judy Hoffman, and Danfei Xu. Egomimic: Scaling imitation learning via egocentric video. In *ICRA*, 2025. 2, 3

[25] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan

Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024. 2

[26] Moo Jin Kim, Chelsea Finn, and Percy Liang. Fine-tuning vision-language-action models: Optimizing speed and success. *arXiv preprint arXiv:2502.19645*, 2025. 1

[27] Marion Lepert, Jiaying Fang, and Jeannette Bohg. Masquerade: Learning from in-the-wild human videos using data-editing. *arXiv preprint arXiv:2508.09976*, 2025. 2, 3, 4

[28] Hongyu Li, Lingfeng Sun, Yafei Hu, Duy Ta, Jennifer Barry, George Konidaris, and Jiahui Fu. Novaflow: Zero-shot manipulation via actionable flow from generated videos. *arXiv preprint arXiv:2510.08568*, 2025. 3

[29] Jinhan Li, Yifeng Zhu, Yuqi Xie, Zhenyu Jiang, Mingyo Seo, Georgios Pavlakos, and Yuke Zhu. Okami: Teaching humanoid robots manipulation skills through single video imitation. *arXiv preprint arXiv:2410.11792*, 2024. 2

[30] Qixiu Li, Yu Deng, Yaobo Liang, Lin Luo, Lei Zhou, Chengtang Yao, Lingqi Zeng, Zhiyuan Feng, Huizhi Liang, Sicheng Xu, et al. Scalable vision-language-action model pretraining for robotic manipulation with real-life human activity videos. *arXiv preprint arXiv:2510.21571*, 2025. 2, 3

[31] Songming Liu, Lingxuan Wu, Bangguo Li, Hengkai Tan, Huayu Chen, Zhengyi Wang, Ke Xu, Hang Su, and Jun Zhu. Rdt-1b: a diffusion foundation model for bimanual manipulation. *arXiv preprint arXiv:2410.07864*, 2024. 1, 2, 3

[32] 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. 2, 3

[33] Lingni Ma, Yuting Ye, Fangzhou Hong, Vladimir Guzov, Yifeng Jiang, Rowan Postyeni, Luis Pesqueira, Alexander Gamino, Vijay Baiyya, Hyo Jin Kim, et al. Nymeria: A massive collection of multimodal egocentric daily motion in the wild. In *ECCV*, 2024. 2, 3

[34] Yecheng Jason Ma, Shagun Sodhani, Dinesh Jayaraman, Osbert Bastani, Vikash Kumar, and Amy Zhang. Vip: Towards universal visual reward and representation via value-implicit pre-training. *arXiv preprint arXiv:2210.00030*, 2022. 3

[35] Russell Mendonca, Shikhar Bahl, and Deepak Pathak. Structured world models from human videos. In *RSS*, 2023. 2

[36] 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. 3

[37] Yaru Niu, Yunzhe Zhang, Mingyang Yu, Changyi Lin, Chenhao Li, Yikai Wang, Yuxiang Yang, Wenhao Yu, Tingnan Zhang, Zhenzhen Li, et al. Human2locoman: Learning versatile quadrupedal manipulation with human pretraining. In *RSS*, 2025. 2

[38] 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 *ICRA*, 2024. 1, 2

[39] 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 *Human to Robot: Workshop on Sensorizing, Modeling, and Learning from Humans*. 3

[40] Yuzhe Qin, Yueh-Hua Wu, Shaowei Liu, Hanwen Jiang, Ruihan Yang, Yang Fu, and Xiaolong Wang. Dexmv: Imitation learning for dexterous manipulation from human videos. In *European Conference on Computer Vision*, pages 570–587. Springer, 2022. 2

[41] Yuzhe Qin, Wei Yang, Binghao Huang, Karl Van Wyk, Hao Su, Xiaolong Wang, Yu-Wei Chao, and Dieter Fox. Anyteleop: A general vision-based dexterous robot arm-hand teleoperation system. *arXiv preprint arXiv:2307.04577*, 2023. 3

[42] Ri-Zhao Qiu, Shiqi Yang, Xuxin Cheng, Chaitanya Chawla, Jialong Li, Tairan He, Ge Yan, David J Yoon, Ryan Hoque, Lars Paulsen, et al. Humanoid policy˜ human policy. *arXiv preprint arXiv:2503.13441*, 2025. 2, 3, 4, 5, 6

[43] 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. 3

[44] Tony Tao, Mohan Kumar Srirama, Jason Jingzhou Liu, Kenneth Shaw, and Deepak Pathak. Dexwild: Dexterous human interactions for in-the-wild robot policies. In *RSS*, 2025. 2, 3, 5

[45] RDT Team. Rdt2: Enabling zero-shot cross-embodiment generalization by scaling up umi data, 2025. 2, 3

[46] Tesla. Full self-driving (supervised). https://www.tesla.com/fsd, 2025. Accessed: 2025-09-20. 2

[47] Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *IROS*, 2012. 3, 4

[48] 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. 3

[49] Chen Wang, Haochen Shi, Weizhuo Wang, Ruohan Zhang, Li Fei-Fei, and C Karen Liu. Dexcap: Scalable and portable mocap data collection system for dexterous manipulation. *arXiv preprint arXiv:2403.07788*, 2024. 2

[50] Lirui Wang, Xinlei Chen, Jialiang Zhao, and Kaiming He. Scaling proprioceptive-visual learning with heterogeneous pre-trained transformers. *NeurIPS*, 2024. 3

[51] Xiaolong Wang, Rohit Girdhar, and Abhinav Gupta. Binge watching: Scaling affordance learning from sitcoms. In *CVPR*, 2017. 2

[52] Xin Wang, Taein Kwon, Mahdi Rad, Bowen Pan, Ishani Chakraborty, Sean Andrist, Dan Bohus, Ashley Feniello, Bugra Tekin, Felipe Vieira Frujeri, et al. Holoassist: an egocentric human interaction dataset for interactive ai assistants in the real world. In *ICCV*, 2023. 3

[53] Chuan Wen, Xingyu Lin, John So, Kai Chen, Qi Dou, Yang Gao, and Pieter Abbeel. Any-point trajectory modeling for policy learning. *arXiv preprint arXiv:2401.00025*, 2023. 3

[54] Haoyu Xiong, Quanzhou Li, Yun-Chun Chen, Homanga Bharadhwaj, Samarth Sinha, and Animesh Garg. Learning by watching: Physical imitation of manipulation skills from human videos. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7827–7834. IEEE, 2021. 2, 3

[55] 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. *arXiv preprint arXiv:2505.21864*, 2025. 2

[56] 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. 2, 3

[57] Chengbo Yuan, Rui Zhou, Mengzhen Liu, Yingdong Hu, Shengjie Wang, Li Yi, Shanghang Zhang, Chuan Wen, and Yang Gao. Motiontrans: Human vr data enable motion-level learning for robotic manipulation policies. In *Human to Robot: Workshop on Sensorizing, Modeling, and Learning from Humans*. 3

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

[59] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 4

[60] Tony Z Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. *arXiv preprint arXiv:2304.13705*, 2023. 2

[61] Tony Z Zhao, Jonathan Tompson, Danny Driess, Pete Florence, Kamyar Ghasemipour, Chelsea Finn, and Ayzaan Wahid. Aloha unleashed: A simple recipe for robot dexterity. In *CoRL*, 2024. 1

[62] Zhenyu Zhao, Hongyi Jing, Xiawei Liu, Jiageng Mao, Abha Jha, Hanwen Yang, Rong Xue, Sergey Zakharor, Vitor Guizilini, and Yue Wang. Humanoid everyday: A comprehensive robotic dataset for open-world humanoid manipulation. *arXiv preprint arXiv:2510.08807*, 2025. 3, 4

[63] Jinliang Zheng, Jianxiong Li, Zhihao Wang, Dongxiu Liu, Xirui Kang, Yuchun Feng, Yinan Zheng, Jiayin Zou, Yilun Chen, Jia Zeng, et al. X-vla: Soft-prompted transformer as scalable cross-embodiment vision-language-action model. *arXiv preprint arXiv:2510.10274*, 2025. 2

[64] Lawrence Y Zhu, Pranav Kuppili, Ryan Punamiya, Patcharapong Aphiwetsa, Dhruv Patel, Simar Kareer, Sehoon Ha, and Danfei Xu. Emma: Scaling mobile manipulation via egocentric human data. *arXiv preprint arXiv:2509.04443*, 2025. 3

In-N-On: Scaling Egocentric Manipulation with in-the-wild and on-task Data

Supplementary Material

6. Appendix

6.1. Object visualization

We visualized the objects used in each task in Figs 9, 10, 11 and 12.

6.2. Post-training data

We report the detailed data collected for post-training across all tasks. For each task, we count the total number of demonstrations performed by both humans and robots.

Task	Robot	Human
Single Object Grasping	120	2545
Multi Object Grasping	180	1016
Burger assembly	80	750
Pouring	1 + 30 left grasp + 30 right grasp	727

Table 3. Overview of post-training data for human and robot demonstrations. For the pouring task, we collect one initial demonstration and 30 additional demonstrations for each of the left-grasp and right-grasp configurations.

6.3. Background Ablation

With scaled pre-training and post-training, Human_0 exhibits strong background generalization capabilities. To evaluate this aspect, we test the Multi-Object Grasping task under a variety of background. The results are summarized in Tab. 4.

Background	Multi Object Grasping	
	I.D.	O.O.D
White table (original)	29/30	30/30
Black tablecloth	26/30	29/30
Floral tablecloth	24/30	28/30

Table 4. Ablation study evaluating Multi Object Grasping performance under varying background conditions. We report the number of successful executions over 30 trials for both I.D. and O.O.D settings.

6.4. Multi-target language following

We further evaluate the language-following ability of Human_0 in multi-target settings. Specifically, we consider both in-distribution (I.D.) and out-of-distribution (O.O.D.) language instructions on the Multi-Object Grasping and Burger Assembly tasks. I.D. instructions are drawn from the robot data, while O.O.D. instructions involve novel language that appears only in the human data. We consider it a success if it moves toward the correct object. Results are reported in Tab. 5. Human_0 demonstrates robust multi-target language following on both tasks, maintaining high success rates even under O.O.D. instructions, indicating strong zero-shot language-following ability.

Method	Multi Object Grasping		Burger assembly	
	I.D.	O.O.D	I.D.	O.O.D
Human_0 (Ours)	29/30	30/30	10/12	10/12

Table 5. Evaluation of multi-target language following for Human_0 on the Multi-Object Grasping and Burger Assembly tasks.

6.5. Failure analysis

Despite strong performance across grasping, assembly, and pouring, Human0 still exhibits consistent failure patterns tied to perception and long-horizon control. In the different scenes or under lighting changes, the model occasionally mislocalizes objects or confuses similarly colored items, leading to false grasps or collisions. These are amplified in tasks with tool use or precise manipulation. For example, during burger assembly, small inaccuracies in early grasps often snowball into downstream misplacements that the policy cannot recover from. Similarly, in pouring tasks, failures in any of the sequential sub-stages result in the entire attempt failing.

Additionally, while domain adaptation reduces obvious human–robot discrepancies, subtle embodiment-specific issues persist. The motion near joint limits or unstable grasps in contact-rich settings suggest that the learned representation is not fully invariant to embodiment details. These issues indicate that although large-scale human data greatly improve generalization, robust performance on precise, long-horizon manipulation still requires richer demonstrations and broader embodiment coverage.

Figure 9. Single Object Grasping: The 4 seen objects (left) are included in the post-training, while the 4 unseen objects (right) never appear in the human demonstration data and are used to evaluate object-level generalization.

Figure 10. Multi Object Grasping: The 3 seen objects (left) appear in both human and robot data. The 3 unseen objects (right) are present only in the human data, together with corresponding language instructions. They are used to evaluate the zero-shot language following.

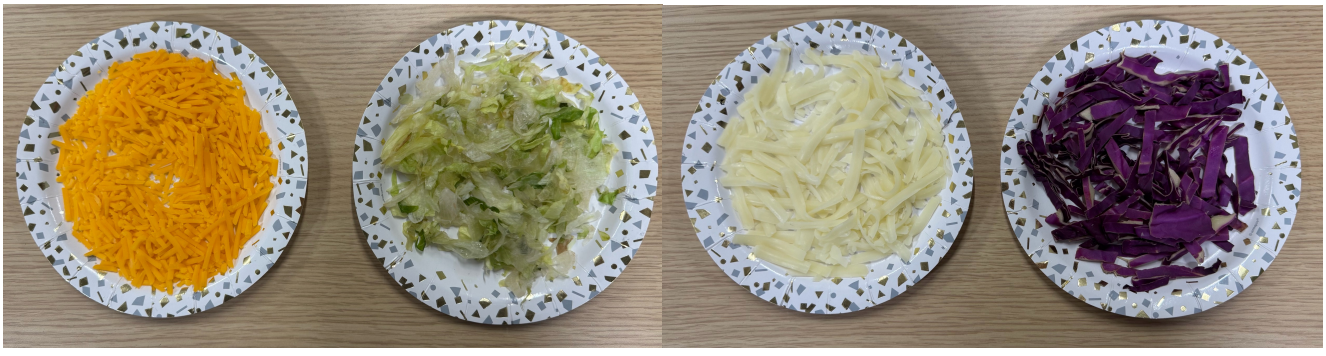


Figure 11. Burger Assembly: The 2 seen objects (left) appear in both human and robot data. The 2 unseen objects (right) are present only in the human data, together with corresponding language instructions. They are used to evaluate the zero-shot language following.

Figure 12. Pouring: The task uses a white protein drink bottle and a plastic cup.