Emergence of Human to Robot Transfer in Vision-Language-Action Models

Simar Kareer[1,2*] Karl Pertsch[1] James Darpinian[1] Judy Hoffman[2]
Danfei Xu[2] Sergey Levine[1] Chelsea Finn[1] Suraj Nair[1]
[1]Physical Intelligence [2]Georgia Institute of Technology

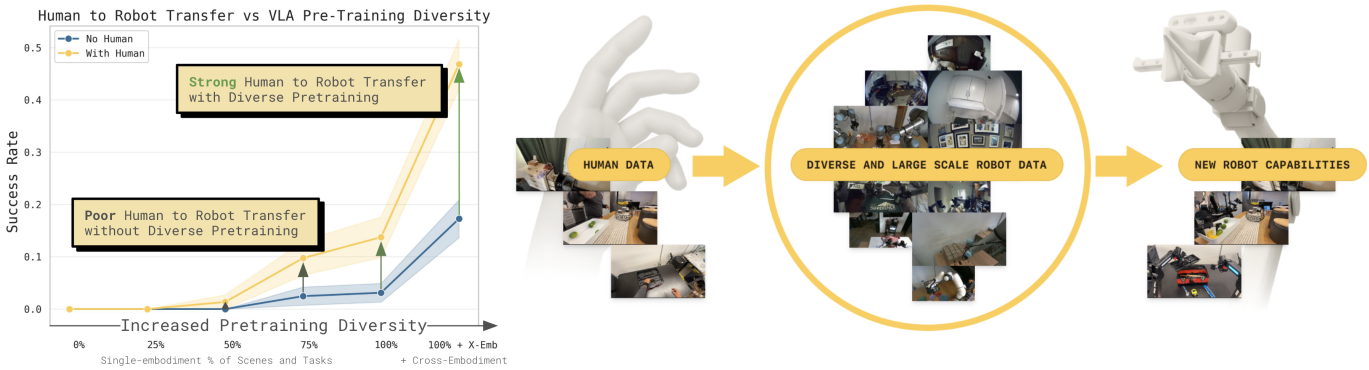

Fig. 1: **Emergence of human to robot transfer:** We observe that the transfer from human data to robot policies scales with the size and diversity of VLA pre-training data. The x-axis represents the diversity of the pre-training robot dataset, and the yellow and blue lines show the finetuning performance with and without human embodiment data. While both increase, the gain from leveraging human data only appears beyond a certain pre-training scale. We evaluate on a suite of four generalization scenarios shown only in the human data.

***Abstract*—Vision-language-action (VLA) models can enable broad open world generalization, but require large and diverse datasets. It is appealing to consider whether some of this data can come from human videos, which cover diverse real-world situations and are easy to obtain. However, it is difficult to train VLAs with human videos alone, and establishing a mapping between humans and robots requires manual engineering and presents a major research challenge. Drawing inspiration from advances in large language models, where the ability to learn from diverse supervision emerges with scale, we ask whether a similar phenomenon holds for VLAs that incorporate human video data. We introduce a simple co-training recipe, and find that human-to-robot transfer emerges once the VLA is pre-trained on sufficient scenes, tasks, and embodiments. Our analysis suggests that this emergent capability arises because diverse pretraining produces embodiment-agnostic representations for human and robot data. We validate these findings through a series of experiments probing human to robot skill transfer and find that with sufficiently diverse robot pre-training our method can nearly double the performance on generalization settings seen only in human data.**

I. Introduction

Human knowledge provides the foundation to instill physical intelligence in robots. This manifests in many forms, from bootstrapping robot policies with human generated text and images via vision-language models, to mimicking human generated actions via robot teleoperation. While such techniques *indirectly* imbue the model with human experience, the right recipe to directly learn from human experience, for instance by watching a video of someone perform a task, remains an active area of research [9, 2, 37, 5, 24, 30]. Techniques for leveraging this type of data hold the promise to unlock massive scale human data for generalist robot policies.

Drawing inspiration from language models, recent work has found that the *ability* to leverage certain sources of data is intrinsically tied to model scale [55, 57, 56]. For instance, while smaller models fail to leverage diverse instruction tuning datasets, larger models become generalists that soak up diverse data and generalize to new tasks. This leads us to ask: **does learning skills from human video data, with no explicit alignment, emerge with scale?** To test this hypothesis, we introduce a simple co-training recipe, that treats human videos as an additional embodiment with the same objectives used for robot data. Specifically, we predict low level end effector trajectories using 3D hand tracks and high-level sub-tasks using dense language annotations, mirroring the objectives used during robot pretraining. We then co-finetune on a mix of this human data with the relevant robot data, and evaluate in a setting present only in the human data. For example one such setting involves sorting eggs, where the robot data covers placing eggs in cartons, while human data specifies how differently colored eggs should be sorted across multiple cartons.

With this recipe, we uncover our key finding: human to robot transfer is an emergent property of diverse VLA pre-training (Figure 1). As we expand the diversity of robot data – across tasks, scenes, and embodiments – the pretrained VLA

*Work conducted during internship at Physical Intelligence.

Fig. 2: **Per-task improvement from human data:** We plot the *difference* in performance between policies fine-tuned with robot + human data versus robot-only data, isolating the lift from human supervision. Gains are largest when pre-training spans diverse tasks, scenes, and embodiments, suggesting that broad pre-training improves transfer from human videos.

becomes increasingly capable of leveraging human videos during post-training. We quantify this effect on four generalization benchmarks that probe different axes of transfer, including unseen apartments, novel object categories, and new task semantics. Our full recipe leverages human video data to enable capabilities never shown in robot data, for instance it busses unseen objects, tidies an unseen home and performs a task with novel semantic structure.

These findings might lead one to ask: *why* does diverse pretraining matter so much for transfer? We find that as pre-training diversity increases, the latent representations between human and robot data naturally align. This suggests that with sufficient data coverage, models begin to form embodiment agnostic representations, despite vast visual and kinematic domain shifts. In the same way large language models become generalists that can learn from diverse supervision, diverse VLAs become generalists that can learn from diverse embodiments.

Concretely we show that robotic foundation models are able to directly leverage human data when pre-trained on sufficiently diverse data, which we demonstrate quantitatively by performance on generalization tasks, as well as qualitatively by analyzing the structure of the latent embeddings across embodiments. To test the efficacy of our recipe, we ablate the importance of each training objective, the importance of wrist cameras, and compare the relative value between human and robot data. We believe this research provides a new perspective on the potential role of embodied human data in training state-of-the-art VLAs. Rather than developing bespoke algorithms to leverage human data, we can think of it in the context of cross-embodiment transfer, where its usefulness is amplified by diverse pretraining.

II. Related Works

Learning from Humans. Learning manipulation from human video has received significant attention due to its potential scalability. Over the years, advances have been made to leverage this data more directly for policy learning. Early works in this field leveraged human video data to train stronger vision encoders, which can improve downstream policy learning [37, 36, 35]. Such approaches leverage the visual diversity of large human datasets like Ego4D [21] to train rich visual features, but are unable to directly improve action prediction. To address this, a number of works developed proxies for actions via intermediate prediction tasks such as keypoint tracking [5, 53], latent actions [63, 11, 31, 66], video prediction [40], VQA [32], reward modeling [9], and affordance prediction [3, 2]. Alternatively, other approaches use overlayed robots and AR/VR to explicitly align the human and robot actions [16, 41]. While these works get closer to capturing real human actions, they introduce a manually engineered structure to enable transfer, limiting the generality of tasks that can be captured.

In parallel to this work, advances in AR/VR enable us to extract explicit actions from humans in the form of 3D hand and head tracking [18]. Recent works leverage this advancement to train unified policies on human and robot data with a single objective: future action prediction, whether that be from a human hand or robot end-effector [24, 43, 69, 30, 33, 62, 44, 34]. These works offer a promising path to directly leverage large scale human data, but such methods are generally brittle at a small scale. As such, they often rely on some form of alignment to work well, whether that be kinematic, visual, or

Fig. 3: **Training mixture and benchmark.** Our fine-tuning mix is evenly split between human data for generalization tasks and robot data for the nearest neighbor task. For each task we evaluate generalization to a new concept introduced only in human data 1) Scene generalization: We have robot data for tidying dressers and spice racks across many airbnbs and human data for an unseen apartment. 2) Object generalization: robot data covers bussing a table filled with trash and dinnerware and human data for a new set of objects. 3) Task generalization: robot data covers placing eggs into cartons, but human data introduces the new concept of sorting eggs by color.

latent. In our work, we extend this style of work and perform no explicit alignment steps.

Heterogeneous Vision-Language-Action Models. Modern VLAs are trained as generalist policies with heterogeneous supervision, combining robot teleoperation, web-scale vision–language data, and language annotations into a single model [14, 70, 26, 8, 58, 67, 4, 47, 42, 49, 59, 7, 22]. These models bootstrap strong vision–language backbones to get broad semantic understanding from human-generated images and text, and then ground that understanding in robot experience via behavior cloning on large teleoperation datasets [12, 17, 52, 39, 25, 6, 20, 45, 1, 28, 23]. While supervision from web images, videos, and language further improves open-world generalization, this data lacks explicit actions and is visually out-of-distribution for a robot's egocentric observations.

A common theme in recent VLAs is cross-embodiment training [39, 38, 13, 61, 68, 34], where a single policy is used to control many different robot embodiments with a unified architecture and action representation. These multi-robot VLAs show that skills can transfer across embodiments, often without bespoke alignment beyond shared observation and action spaces. This suggests that heterogeneous, multi-robot pretraining can produce internal representations that are naturally conducive to transfer across embodiments.

We build on this cross-embodiment hypothesis by treating humans as yet another embodiment within the same heterogeneous VLA training recipe. In contrast to non-embodied human videos one might find on YouTube, we leverage embodied human videos with explicit hand motion and language annotations in the VLA mixture. We find that with sufficient pretraining scale, the resulting VLAs naturally form embodiment-agnostic representations that align human and robot trajectories.

Scaling Alternative Data Collection Strategies. While most VLAs are largely driven by robot teleoperation data, there's recent work exploring alternative data collection mechanisms which are more scalable. A number of works use portable hardware that a user operates with their hands to simulate teleoperation [46, 64], for instance UMI [10] is a hand-held parallel jaw gripper that tracks its own movement to use as demonstration data. A number of works expanded on this design to capture data for dexterous hands as well, both via exoskeletons and portable motion capture [60, 48, 54]. While these devices are exciting options to increase data scalability, they ultimately encumber the operator, and it is difficult to perform work naturally with these devices

Capturing embodied human data offers a promising way to address these limitations, using cameras and computer vision to record 3D hand motion with minimal disruption. This approach enables us to observe human behavior without encumbrance. In this study, we therefore focus on methods for leveraging embodied human data.

III. Preliminaries

We consider the setting of training generalist policies with vision language action models. VLAs inherit the architecture and pretrained weights of a vision-language model, but are

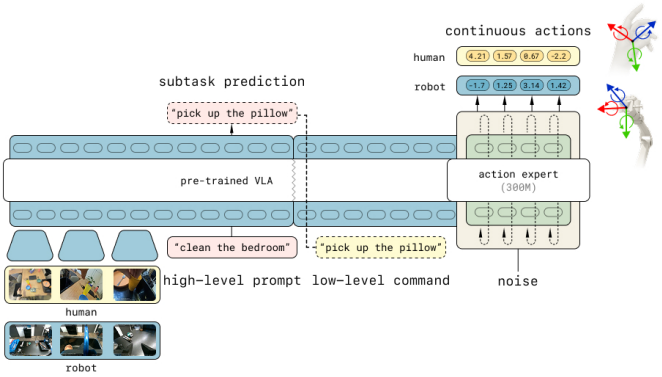

Fig. 4: **Model Architecture.** We use the $\pi_{0.5}$ model. We finetune with a combination of high level sub-task prediction and low level action prediction on both human and robot data. The low level action prediction leverages relative end-effector actions aligned across human and robot.

trained to produce continuous robot control. VLAs are typically trained via behavior cloning on a dataset of demonstrations $D = (o_t, l_t, a_{t:t+H})$ to produce a policy that maps observation and language command to a trajectory of future actions $\pi_\theta(a_{t:t+H} \mid o_t, l_t)$. Actions can either be represented as discrete action tokens [70, 29, 42], trainable via standard next-token prediction, or as continuous values, often trained via flow-matching objectives [8]. In this work, we follow Driess et al. [15] and train VLA models using *both* action representations: we train our model to predict discretized FAST [42] action tokens, and introduce a small action expert network that decodes continuous actions via a flow matching objective. For more details on model architecture and training objective, see [22, 50].

Recent VLAs like $\pi_{0.5}$ showed improved generalization from *co-training* VLAs with additional objectives like subtask prediction, object detection, and VQA. For subtask prediction, the policy predicts a subtask string given visual observations and a high level language command $p(l_t^{subtask} \mid o_t, l_t)$. This language is fed back into the model to condition action generation $\pi_\theta(a_{t:t+H} \mid o_t, l^{subtask})$, similar to chain of thought. Subtask labels are obtained by densely annotating demonstration data with language descriptions of short, atomic action sequences. In this work, we train on human data with two objectives: flow based prediction for continuous actions and language based subtask prediction.

IV. Fine-Tuning with Emergent Human-Robot Alignment

Our fine-tuning recipe aims to leverage embodied human data identically to other robots in our mixture with no explicit alignment. This approach is maximally general, and leans on the capabilities of large models to ingest relevant information from diverse sources, rather than human designed heuristics for aligning domains. We first collect, process, and annotate the human video data, and then use it in combination with robot data to fine-tune the pre-trained model, which we base on the $\pi_{0.5}$ model shown in Figure 4. The fine-tuning objective treats human and robot data in exactly the same way, without any explicit transfer learning method or loss.

A. Human Data Collection Pipeline

Data collection device. We design our data collection setup to enable capture of a wide range of human interactions, while being minimally intrusive, and thus scalable. We equip human data collectors with a head-worn, high resolution camera. Since recent robotics research has demonstrated the benefit of wrist-mounted cameras for policy learning, which provide a more detailed view of the end-effector's interactions with manipulated objects, we also experiment with equipping data collectors with wrist-mounted cameras that provide two additional, time-synchronized camera streams. We ablate the effect of these additional cameras in Section V.

Data collection protocol. Our intent is to collect human data in the style of episodic robot teleoperation data, which allows us to isolate the transfer problem to only the visual and kinematic differences between human and robot. As such, operators are instructed to collect repeated demonstrations for each of our tasks while wearing the data collection device. Further, operators are asked to keep their hands in the field of view of the camera to improve tracking quality. We collected 3 hours of data for bussing, 3 hours for spice, 3 hours for dresser and 5 hours for sort eggs.

Data processing & annotation. Given a raw video capture of human interactions, we use visual SLAM to reconstruct the 6D movement $e_t \in \mathbb{R}^6$ of the head-mounted camera relative to a constant world-frame. We also reconstruct the position of 17 3D keypoints $h_t^{e_t} \in \mathbb{R}^{3\times17}$ for both hands in the head camera frame. Finally, analogously to the robot teleoperation data in our training mixture, we annotate the human video data with text-based subtasks, describing the actions of each arm.

Action space. We aim to roughly align the action representations for human and robot data in our training mixture to facilitate transfer. For robot teleoperation data, there are multiple choices of action representations. Two common options are to represent actions as trajectories of robot *joint positions* or *end-effector poses*. We will consider end-effector based actions, since approximating joint positions for humans is difficult. Specifically, these end-effector actions are represented as a length H action chunk $[a_0, a_1, ..., a_H]$, where each a_i represents the 6-DoF pose relative to the current observed state 6-DoF pose s_0. The total action space for robot data is the concatenation of the 6 DoF left arm end-effector trajectory + gripper, 6 DoF right arm end-effector trajectory + gripper, and 2 dimensional base actions, yielding a total action chunk of $a \in \mathbb{R}^{H\times16}$. To compute corresponding actions in the human video, we define an "end-effector" pose spanning the 3D keypoints of each hand's palm, middle, and ring finger (Figure 6),relative to the head frame e_t. We then compute end-effector actions as relative transformations from the current 6-DoF state similar to how we do for the robot end effector. Similarly, we approximate relative robot base actions by projecting the human video base camera poses into the frame of the chunk's first timestep base camera pose. We do not explicitly approximate "gripper actions" for the human video, since it is challenging to estimate the openness of a

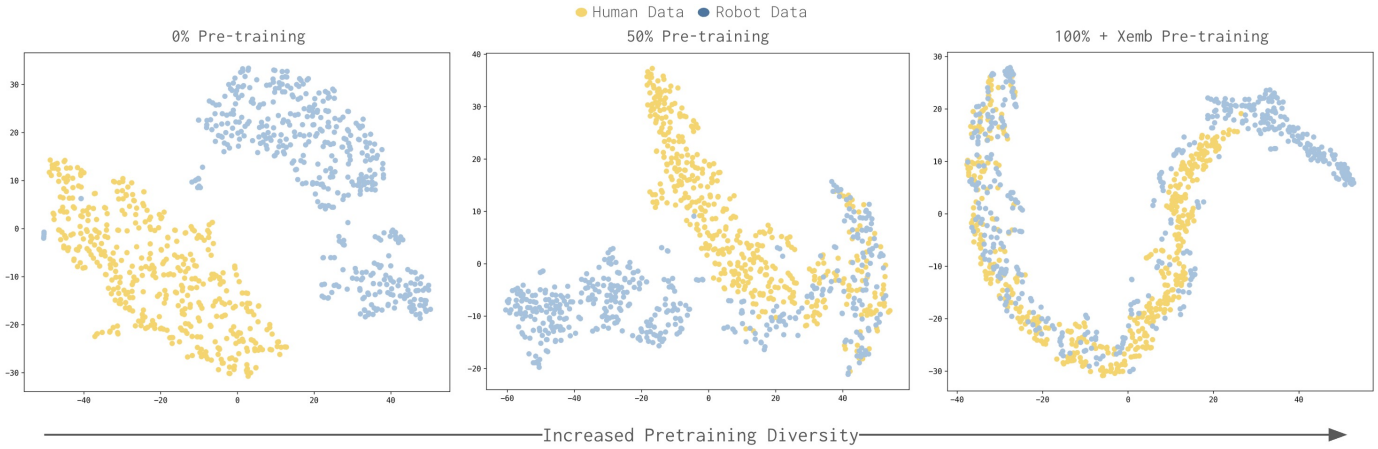

Fig. 5: **VLA representation of human and robot data.** We plot the latent embeddings of our VLA by performing a TSNE analysis on mean-pooled tokens from the final layer of the VLM backbone. With no pre-training, it is clear that the model has disjoint representations between human and robot data. But as pretraining becomes more diverse, latent overlap increases, which correlates with performance on our generalization tasks.

Fig. 6: **Data Collection Apparatus.** We collect human data with three views: head mounted camera, left wrist camera, and right wrist camera.

human hand during object interactions, and instead rely on learning gripper actions from robot data only. As a result, our human actions have $2 \times 6 + 6 = 18$ dimensions.

Training objectives. Our best recipes to perform difficult long horizon tasks leverage both high level subtask prediction and low level action prediction. We construct both of these prediction tasks on our human data. For low level action prediction, we supervise action chunk prediction both via next token prediction on discrete FAST tokens as well as flow matching loss on the continuous actions $\pi_\theta(a|o_t, l_t^{\text{subtask}})$. For high level subtask prediction we train on next token prediction on subtask language tokens $\pi_\theta(l_t^{\text{subtask}} \mid o_t, l_t)$.

Training mixture. At fine-tuning time, it is important to create a training mixture that both retains the model's original capabilities, while introducing new concepts from human data to improve generalization. Our mixture reflects this with a simple recipe: we *co-train* human data for generalization tasks at 50-50 proportion to the nearest neighbor robot task. We use this mixture to finetune $\pi_{0.5}$, a strong VLA exhibiting zero shot generalization, and further improve its capabilities. As shorthand, we refer to the combined model that integrates egocentric data into $\pi_{0.5}$ as $\pi_{0.5} + \text{ego}$.

V. Experimental Findings

To test whether $\pi_{0.5} + \text{ego}$ can generalize to new concepts from egocentric human data, we construct a suite of "generalization" scenarios that have limited coverage in robot data, but present in human data. These scenarios span generalizing to new scenes, objects and tasks. We begin our study by understanding whether our recipe can enable transfer to these new settings. Then, we validate our core hypothesis, which is that this transfer is an emergent property of diverse VLA pretraining. Finally, we compare human embodiment data to other robot embodiments, study whether transfer occurs from high level subtask or low level action prediction, and ablate the impact of our human-worn wrist cameras.

A. Human to robot transfer benchmark

Our benchmark aims to test human to robot transfer across various axis of generalization: scene, object and task (fig. 3). For each axis, we consider a setup where our robot teleoperation data lacks coverage, and we collect targeted human data to expand this coverage. In each setting we co-train using $\pi_{0.5}$ + ego and evaluate on the new concept introduced in human data.

Scene transfer: We identified two tasks where we have robot data coverage in a fixed number of homes, but $\pi_{0.5}$ trained only on robot data fails to generalize to an *unseen* home: `Spice` and `Dresser` in which the robot must tidy a spice rack and the top of a dresser respectively. We collect human data in the unseen target kitchen, then benchmark $\pi_{0.5} + \text{ego}$ on this new scene. For both of these shorter horizon tasks the score is a binary success rate.

Object transfer: Our robot data has coverage of bussing a messy table covered with trash and dinnerware. We then collect human data which introduces new objects like kitchen tools, then benchmark $\pi_{0.5} + \text{ego}$ on these new objects. For this longer-horizon task the score measures the number of correctly placed objects.

Task transfer: Our robot data has coverage for picking eggs and packing them into a carton. We collect human data that *sorts* eggs into two cartons based on color and benchmark $\pi_{0.5} + \text{ego}$ on this new task. For this longer-horizon task the score measures the number of correctly placed eggs. For more details about task setup and scoring, see Appendix A.

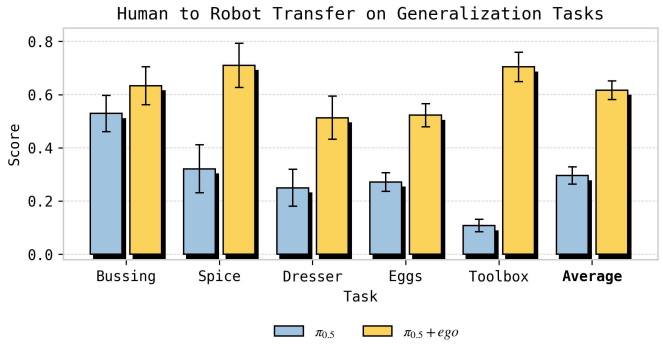

Fig. 7: **Human to robot transfer finetuning $\pi_{0.5}$.** We evaluate performance across a suite of static and mobile tasks, each testing either scene, object, or task level generalization present only in the human data. We see clear human to robot transfer resulting in nearly double the score on the target tasks.

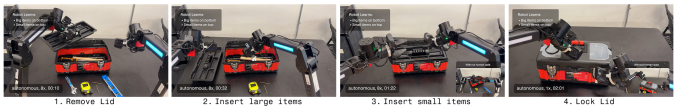

Fig. 8: **Dexterous Toolbox task.** The robot must sort tools into two compartments of the toolbox and latch the lids shut. While the underlying motion primitives appear in robot teleoperation data, their composition in this task is unique to the human videos. Co-training with human data yields a 59% improvement in task progress score.

B. The $\pi_{0.5}$ + ego recipe enables generalization to unseen scenes, objects, and tasks

We report transfer results on our suite of benchmark tasks in Figure 7. **Across all three generalization axes we find that targeted human data collection and co-training can significantly improve policy generalization.** Concretely, for both scene and object generalization, we see substantially higher task score after co-training: `Spice`: 32% $\rightarrow$ 71%; `dresser`: 25% $\rightarrow$ 50%; `bussing`: 53$\rightarrow$63%.

Further, we see strong *task* transfer from human video in two tasks: `Sort Eggs` and `Toolbox`. In the egg sorting task, while the policy trained on robot data only has the basic manipulation skills to pick and place eggs, it has no concept of *sorting* and simply places eggs into cartons randomly (57% sorting accuracy). In contrast, once co-trained with human egg sorting videos, the robot policy is able to sort eggs with 78% accuracy, and on average placed 4 more eggs correctly than $\pi_{0.5}$. We find task-level generalization extends to more dexterous tasks as well, for instance, in `Toolbox` (Figure 8), where the robot must organize tools into specific compartments then latch a lid shut. Concretely, co-training with human video yields a 59% absolute gain in task performance. Although the underlying motion primitives are present in robot teleoperation data, their composition is unique to this setting and is transferred directly from human data, indicating that human-to-robot transfer extends to the composition of dexterous primitives.

C. Human to robot transfer emerges as a function of diverse VLA pretraining across scenes, tasks and embodiments

We've established that $\pi_{0.5}$ + ego can leverage embodied human data to expand its capabilities, which leads us to the central question of this work: what enables this transfer? We hypothesize that policy *pretraining* with a diverse data mix of many scenes, tasks and embodiments is the key enabler for effective human to robot transfer. Intuitively, VLAs with strong pretraining may learn abstractions that span embodiments—organizing their representations to capture shared structure across domains, and thus facilitating transfer. We test this hypothesis in two parts. First, we establish that human to robot transfer on our generalization benchmark *increases* as a function of pretraining diversity. Then, we analyze our model's learned representations as pretraining diversity increases.

Strong human to robot transfer emerges with diverse pretraining. To evaluate the impact of VLA pretraining on human to robot transfer, we repeat our transfer benchmark experiments but with the following, increasingly diverse, pretrained initializations:

- `0%`: base VLM initialization only
- `25%, 50%, 75%, 100%`: VLA pre-trained on increasingly diverse robot data, corresponding to fractions of the full diversity of [scene-task] combinations in our data, constrained to the target robot embodiments: ARX and mobile ARX
- `100% + X-emb`: the $\pi_{0.5}$ full VLA pretraining mixture Intelligence et al. [22], which additionally contains data across numerous non-target robot embodiments.

With each of these pretrained initializations we train two models: one with only robot teleoperation data from the most similar tasks in our dataset, and one which additionally includes human embodiment data for these tasks. This allows us to measure the impact of diverse pretraining on human to robot transfer.

We report results in Figure 2. Concretely, we report the *difference* in score between the model using human data and not using human data at different levels of pre-trained model scale. This difference represents the magnitude of the human to robot *transfer* as a function of pre-training diversity. We find that this transfer significantly increases as a function of pre-training diversity. While with no or little pre-training, VLAs cannot benefit from human data co-training (0%, 25%), VLAs pre-trained on diverse data see significant gains from human data co-training (75%, 100%). Transfer is further improved by pre-training on a diverse cross-embodiment data mix, that includes data from diverse, non-target robot embodiments.

We can analyze the scaling trends for each task individually. For instance in `Sort Eggs` we see that increased pretraining diversity alone does not enable a robot-only policy to perform `Sort Eggs`, a task never seen in our robot teleoperation data (Figure 9). However, increased pretraining diversity enables us to *transfer* significantly more knowledge from human data, which does cover this new task. Similarly, for `Dresser`, up until our 50% pretrained checkpoint there are no gains from human video co-training, and potentially even negative transfer (Figure 14). But between `75%`$\rightarrow$`100% + X-emb` we see consistent stacked gains on top of the robot-only baseline, even as the pretrained checkpoint gets stronger. More broadly, these results suggest that human to robot transfer will continue

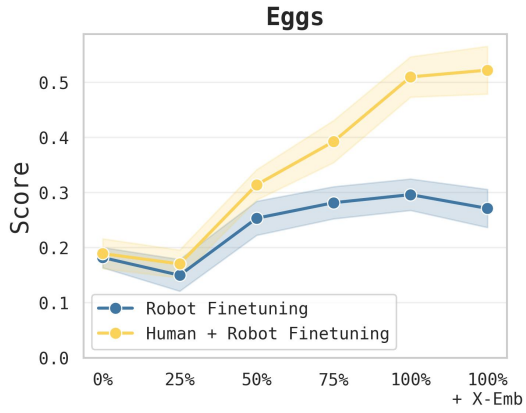

Fig. 9: **Task generalization performance scaling:** Performance of Robot Finetuning on `Sort Eggs` plateaus, even as the pretraining diversity improves. In contrast, Human+Robot Finetuning performance scales sharply with pretraining, suggesting that broader pretraining enables more effective *transfer* from human data.

to improve as the diversity of our pretrained model grows. This follows our intuition, since we expect that diversity of scenes, tasks and embodiments ought to improve the model's ability to form embodiment agnostic abstractions.

Embodiment agnostic representations emerge with pre-training scale. We hypothesize that diverse pre-training helps produce embodiment agnostic representations which in turn improve human to robot transfer. To probe this, we perform a TSNE [51] analysis on the output embeddings of the VLA from both human and robot data after co-training (Figure 5). With poor pre-training, the model has *disjoint* representations across embodiments, suggesting that the model separately fits these distributions. Then, as pretraining diversity increases, the representations converge, suggesting the model builds a unified representation for both embodiments. See Appendix C for more details.

Prior works that operate on less data observe that co-training improves performance, but the representations of human and robot are disjoint, and propose methods to explicitly improve representational alignment [43]. Our analysis suggests that with sufficiently diverse pretraining, *co-training alone* can produce aligned representations that facilitate transfer.

D. How does embodied human data compare to data from other robots?

$\pi_{0.5}$+ego frames human to robot transfer as an instantiation of cross embodiment transfer, so it's natural to benchmark human to robot transfer with robot to robot transfer. This helps us understand whether we can leverage human data as just another robot embodiment in our mix. We first compare human data to an "upper bound" scenario where we collect target robot data for our benchmark tasks (Figure 10).

For two of three tasks (`Sort Eggs` and `Dresser`), we find that finetuning with human data were nearly as effective as finetuning with in domain data from the target robot itself. However, we note that on the `Bussing` task, target robot data was more effective than human data alone (25% vs 65%).

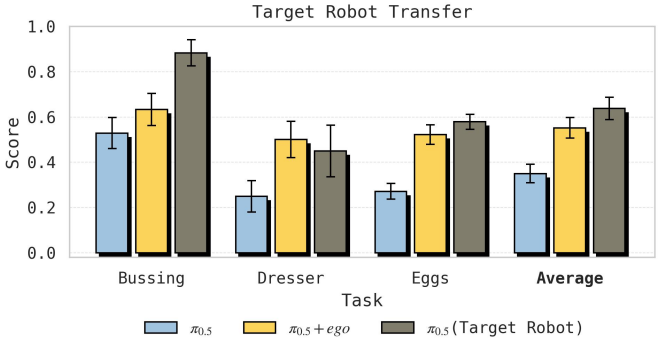

Fig. 10: **Human data compared to target robot data:** For `Eggs` and `Dresser` we find that fientuning with comparable amounts of human data (yellow) and robot data (grey) resulted in similarly performant models. For `Bussing` we observe a larger performance gap to target robot data.

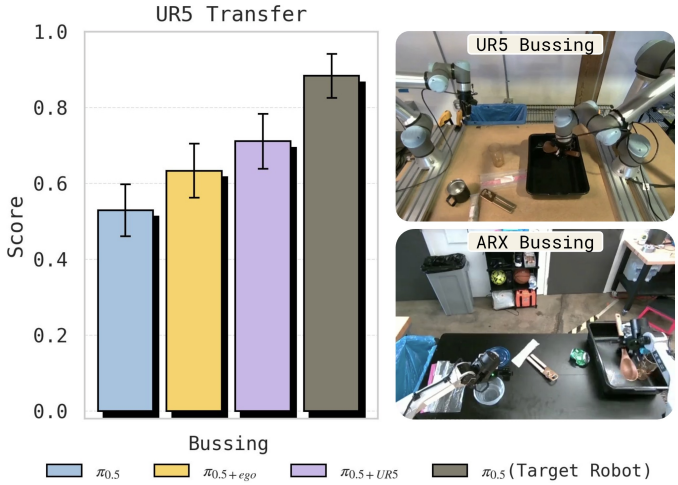

Fig. 11: **Human data compared to cross embodiment robot data:** We compare the transfer from human data on the bussing task to that from a robot (albeit a different UR5 robot). We find that both enable lift over the baseline, but neither match data of the target robot, suggesting that human data transfer parallels cross-embodiment robot transfer.

Next we study whether embodied human data has roughly the same value as non-target robot data for a new task. Specifically, for the `Bussing` task we collected 400 demonstrations (7.45 hours) on a UR5 robot and evaluated transfer to an arx robot. We see a similar trend in the nature of human to ARX and UR5 to ARX transfer – both exceed the baseline, but both also don't match data from the target robot embodiment, suggesting that human data transfer and cross-embodiment robot transfer share similar properties.

E. At what level does transfer occur?

A natural question is whether human data can only be used to transfer "high level" semantic concepts, or whether it's also transferring the "low level" action prediction. For the `Bussing` and `Eggs` task, we do not use a high-level policy during evaluation, so the transfer has to come from low level action prediction. For our mobile tasks `Spice` and `Dresser`, we evaluate a joint high level + low level and ablate the impact of each (Figure 12), by testing robot-only HL+LL, robot-only HL + cotrained LL, cotrained HL + robot-only LL and cotrained HL+LL. We find that leveraging human data

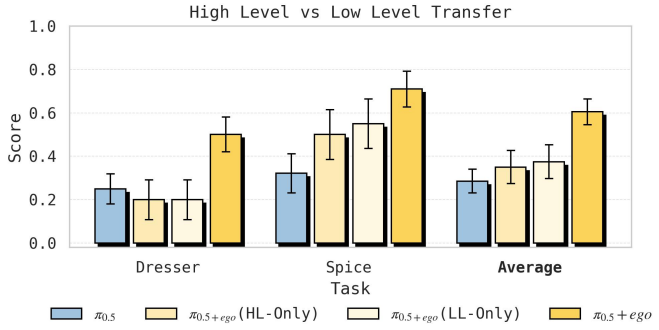

Fig. 12: **High level vs low level transfer** For our mobile tasks we find that embodied human data enables transfer both through high level sub-task prediction and low level action prediction.

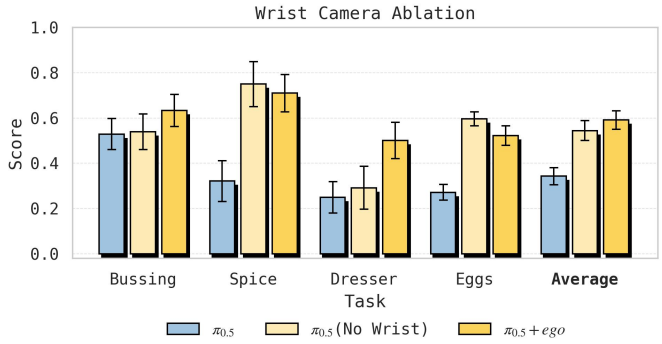

Fig. 13: **Human worn wrist cameras:** We see that wrist cameras provide benefit for the `Dresser` and `Bussing` tasks and had similar performance elsewhere. This matches our intuition that some (but not all) tasks will benefit from the added observability of wrist cameras.

for only the HL or LL policy alone is not as effective as co-training both with human data, suggesting that transfer occurs across both levels.

When we only leverage human data for the HL policy, the low level policies don't follow commands correctly. For instance in the `spice` task we observe a failure mode where "pick up the spice bottle" is misinterpreted by the low level policy to pick up bottles that are already on the tray. Or in the `dresser` task, when the HL says to "put the necklace in the jewelry box", it sometimes puts it in the dresser organizer.

Likewise, when we only leverage human data for LL policy, we get poor HL policy commands. For instance in `spice` the high level policy continues to predict "pick up spice bottle" long after the bottle has been picked, blocking task progress. And in `dresser` the HL policy often predicts incorrect actions, like "put the hair clip on the top of the dresser" instead of correctly predicting to put it in the organizer.

F. How important are the human worn wrist cameras?

To help mitigate the sensor gap between human and robot, we opted towards collecting human data with small wrist-worn cameras, to mimic the wrist cameras on our robot arms. We seek to understand their importance, since it has implications on how to sensorize humans for large scale data collection. We report transfer results with and without wrist camera observations for the human video data in Figure 13. For some tasks like `Bussing` and `Dresser` we see improved transfer from leveraging the human worn wrist cameras, while in other tasks like `Spice` and `Eggs`, transfer does not benefit from the additional camera streams. This matches our expectations, since some tasks rely on wrist cameras for observability more than others. As a result of these experiments, we expect that collecting embodied human data *with* wrist worn cameras maximally covers the space of potential tasks.

VI. Discussion and Future Work

We study the emergence of human to robot transfer in our proposed recipe $\pi_{0.5}+$ego. We find that with limited pretraining diversity, VLAs fail to transfer knowledge from human data, but as pretraining diversity grows past a critical threshold, transfer emerges. While our recipes leverage vast datasets of robot teleoperation data in pretraining, we ultimately only use 10s of hours of human data, and this data is collected in an episodic manner. We're moving towards a future with vast datasets of embodied human data, which cover the episodic collection like in this work, but also passive data of people performing everyday tasks. There is additional work to be done to effectively leverage this data during pre-training, but we believe our work lays the groundwork to train VLAs with human data at scale.

Our findings on the emergence of human-to-robot transfer point to a promising future for scaling vision-language-action models. Much like large language models, larger VLAs may not only improve performance but also unlock entirely new capabilities. These capabilities could make it easier to tap into previously hard-to-use data sources and enable more effective transfer across domains—ultimately allowing robotic foundation models to scale *even further*. Using human video may be just one such capability, and it's exciting to imagine what others might emerge as we continue to scale up robotic foundation models.

Acknowledgements

We thank our robot operators for data collection, evaluations, logistics, and video recording, and our technicians for robot maintenance and repair. We'd like to thank Hunter Hancock for extensive support on the blog post. We also thank Brian Ichter, Karol Hausman, and the entire Physical Intelligence team for feedback throughout the project.

References

[1] AgiBot-World-Contributors, Qingwen Bu, Jisong Cai, Li Chen, Xiuqi Cui, Yan Ding, Siyuan Feng, Shenyuan Gao, Xindong He, Xuan Hu, Xu Huang, Shu Jiang, Yuxin Jiang, Cheng Jing, Hongyang Li, Jialu Li, Chiming Liu, Yi Liu, Yuxiang Lu, Jianlan Luo, Ping Luo, Yao Mu, Yuehan Niu, Yixuan Pan, Jiangmiao Pang, Yu Qiao, Guanghui Ren, Cheng Ruan, Jiaqi Shan, Yongjian Shen, Chengshi Shi, Mingkang Shi, Modi Shi, Chonghao Sima, Jianheng Song, Huijie Wang, Wenhao Wang, Dafeng Wei, Chengen Xie, Guo Xu, Junchi Yan, Cunbiao Yang, Lei Yang, Shukai Yang, Maoqing Yao, Jia Zeng, Chi Zhang, Qinglin Zhang, Bin Zhao, Chengyue Zhao, Jiaqi

Zhao, and Jianchao Zhu. Agibot world colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems. *arXiv preprint arXiv:2503.06669*, 2025.

[2] Shikhar Bahl, Abhinav Gupta, and Deepak Pathak. Human-to-robot imitation in the wild. *arXiv preprint arXiv:2207.09450*, 2022.

[3] Shikhar Bahl, Russell Mendonca, Lili Chen, Unnat Jain, and Deepak Pathak. Affordances from human videos as a versatile representation for robotics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13778–13790, June 2023.

[4] Suneel Belkhale and Dorsa Sadigh. Minivla: A better vla with a smaller footprint, 2024. URL https://github.com/Stanford-ILIAD/openvla-mini.

[5] Homanga Bharadhwaj, Roozbeh Mottaghi, Abhinav Gupta, and Shubham Tulsiani. Track2act: Predicting point tracks from internet videos enables generalizable robot manipulation, 2024. URL https://arxiv.org/abs/2405.01527.

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

[7] 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.

[8] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Lucy Xiaoyang Shi, James Tanner, Quan Vuong, Anna Walling, Haohuan Wang, and Ury Zhilinsky. π_0: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024.

[9] Annie S Chen, Suraj Nair, and Chelsea Finn. Learning generalizable robotic reward functions from" in-the-wild" human videos. *RSS*, 2021.

[10] Cheng Chi, Zhenjia Xu, Chuer Pan, Eric Cousineau, Benjamin Burchfiel, Siyuan Feng, Russ Tedrake, and Shuran Song. Universal manipulation interface: In-the-wild robot teaching without in-the-wild robots. In *Proceedings of Robotics: Science and Systems (RSS)*, 2024.

[11] Jeremy A. Collins, Loránd Cheng, Kunal Aneja, Albert Wilcox, Benjamin Joffe, and Animesh Garg. Amplify: Actionless motion priors for robot learning from videos, 2025. URL https://arxiv.org/abs/2506.14198.

[12] Sudeep Dasari, Frederik Ebert, Stephen Tian, Suraj Nair, Bernadette Bucher, Karl Schmeckpeper, Siddharth Singh, Sergey Levine, and Chelsea Finn. Robonet: Large-scale multi-robot learning. *CoRL*, 2019.

[13] Ria Doshi, Homer Walke, Oier Mees, Sudeep Dasari, and Sergey Levine. Scaling cross-embodied learning: One policy for manipulation, navigation, locomotion and aviation. In *Conference on Robot Learning*, 2024.

[14] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023.

[15] Danny Driess, Jost Tobias Springenberg, Brian Ichter, Lili Yu, Adrian Li-Bell, Karl Pertsch, Allen Z Ren, Homer Walke, Quan Vuong, Lucy Xiaoyang Shi, et al. Knowledge insulating vision-language-action models: Train fast, run fast, generalize better. *NeurIPS*, 2025.

[16] Jiafei Duan, Yi Ru Wang, Mohit Shridhar, Dieter Fox, and Ranjay Krishna. Ar2-d2:training a robot without a robot, 2023. URL https://arxiv.org/abs/2306.13818.

[17] Frederik Ebert, Yanlai Yang, Karl Schmeckpeper, Bernadette Bucher, Georgios Georgakis, Kostas Daniilidis, Chelsea Finn, and Sergey Levine. Bridge data: Boosting generalization of robotic skills with cross-domain datasets. *arXiv preprint arXiv:2109.13396*, 2021.

[18] Jakob Engel, Kiran Somasundaram, Michael Goesele, Albert Sun, Alexander Gamino, Andrew Turner, Arjang Talattof, Arnie Yuan, Bilal Souti, Brighid Meredith, Cheng Peng, Chris Sweeney, Cole Wilson, Dan Barnes, Daniel DeTone, David Caruso, Derek Valleroy, Dinesh Ginjupalli, Duncan Frost, Edward Miller, Elias Mueggler, Evgeniy Oleinik, Fan Zhang, Guruprasad Somasundaram, Gustavo Solaira, Harry Lanaras, Henry Howard-Jenkins, Huixuan Tang, Hyo Jin Kim, Jaime Rivera, Ji Luo, Jing Dong, Julian Straub, Kevin Bailey, Kevin Eckenhoff, Lingni Ma, Luis Pesqueira, Mark Schwesinger, Maurizio Monge, Nan Yang, Nick Charron, Nikhil Raina, Omkar Parkhi, Peter Borschowa, Pierre Moulon, Prince Gupta, Raul Mur-Artal, Robbie Pennington, Sachin Kulkarni, Sagar Miglani, Santosh Gondi, Saransh Solanki, Sean Diener, Shangyi Cheng, Simon Green, Steve Saarinen, Suvam Patra, Tassos Mourikis, Thomas Whelan, Tripti Singh, Vasileios Balntas, Vijay Baiyya, Wilson Dreewes, Xiaqing Pan, Yang Lou, Yipu Zhao, Yusuf Mansour, Yuyang Zou, Zhaoyang Lv, Zijian Wang, Mingfei Yan, Carl Ren, Renzo De Nardi, and Richard Newcombe. Project aria: A new tool for egocentric multi-modal ai research, 2023. URL https://arxiv.org/abs/2308.13561.

[19] 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, 2023. URL https://arxiv.org/abs/2204.13662.

[20] Hao-Shu Fang, Hongjie Fang, Zhenyu Tang, Jirong Liu, Chenxi Wang, Junbo Wang, Haoyi Zhu, and Cewu Lu. Rh20t: A comprehensive robotic dataset for learning

diverse skills in one-shot. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 653–660. IEEE, 2024.

[21] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022.

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

[23] Tao Jiang, Tianyuan Yuan, Yicheng Liu, Chenhao Lu, Jianning Cui, Xiao Liu, Shuiqi Cheng, Jiyang Gao, Huazhe Xu, and Hang Zhao. Galaxea open-world dataset and g0 dual-system vla model. *arXiv preprint arXiv:2509.00576*, 2025.

[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, 2024. URL https://arxiv.org/abs/2410.24221.

[25] Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, Peter David Fagan, Joey Hejna, Masha Itkina, Marion Lepert, Yecheng Jason Ma, Patrick Tree Miller, Jimmy Wu, Suneel Belkhale, Shivin Dass, Huy Ha, Arhan Jain, Abraham Lee, Youngwoon Lee, Marius Memmel, Sungjae Park, Ilija Radosavovic, Kaiyuan Wang, Albert Zhan, Kevin Black, Cheng Chi, Kyle Beltran Hatch, Shan Lin, Jingpei Lu, Jean Mercat, Abdul Rehman, Pannag R Sanketi, Archit Sharma, Cody Simpson, Quan Vuong, Homer Rich Walke, Blake Wulfe, Ted Xiao, Jonathan Heewon Yang, Arefeh Yavary, Tony Z. Zhao, Christopher Agia, Rohan Baijal, Mateo Guaman Castro, Daphne Chen, Qiuyu Chen, Trinity Chung, Jaimyn Drake, Ethan Paul Foster, Jensen Gao, David Antonio Herrera, Minho Heo, Kyle Hsu, Jiaheng Hu, Donovon Jackson, Charlotte Le, Yunshuang Li, Kevin Lin, Roy Lin, Zehan Ma, Abhiram Maddukuri, Suvir Mirchandani, Daniel Morton, Tony Nguyen, Abigail O'Neill, Rosario Scalise, Derick Seale, Victor Son, Stephen Tian, Emi Tran, Andrew E. Wang, Yilin Wu, Annie Xie, Jingyun Yang, Patrick Yin, Yunchu Zhang, Osbert Bastani, Glen Berseth, Jeannette Bohg, Ken Goldberg, Abhinav Gupta, Abhishek Gupta, Dinesh Jayaraman, Joseph J Lim, Jitendra Malik, Roberto Martín-Martín, Subramanian Ramamoorthy, Dorsa Sadigh, Shuran Song, Jiajun Wu, Michael C. Yip, Yuke Zhu, Thomas Kollar, Sergey Levine, and Chelsea Finn. Droid: A large-scale in-the-wild robot manipulation dataset. In *Proceedings of Robotics: Science and Systems*, 2024.

[26] 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.

[27] Taein Kwon, Bugra Tekin, Jan Stuhmer, Federica Bogo, and Marc Pollefeys. H2o: Two hands manipulating objects for first person interaction recognition, 2021. URL https://arxiv.org/abs/2104.11181.

[28] Jason Lee, Jiafei Duan, Haoquan Fang, Yuquan Deng, Shuo Liu, Boyang Li, Bohan Fang, Jieyu Zhang, Yi Ru Wang, Sangho Lee, et al. Molmoact: Action reasoning models that can reason in space, 2025. *URL https://arxiv.org/abs/2508.07917*, 2025.

[29] Seungjae Lee, Yibin Wang, Haritheja Etukuru, H. Jin Kim, Nur Muhammad Mahi Shafiullah, and Lerrel Pinto. Behavior generation with latent actions. *arXiv preprint arXiv:2403.03181*, 2024.

[30] Marion Lepert, Jiaying Fang, and Jeannette Bohg. Phantom: Training robots without robots using only human videos, 2025. URL https://arxiv.org/abs/2503.00779.

[31] Zuolei Li, Xingyu Gao, Xiaofan Wang, and Jianlong Fu. Latbot: Distilling universal latent actions for vision-language-action models, 2025. URL https://arxiv.org/abs/2511.23034.

[32] Xiaopeng Lin, Shijie Lian, Bin Yu, Ruoqi Yang, Changti Wu, Yuzhuo Miao, Yurun Jin, Yukun Shi, Cong Huang, Bojun Cheng, and Kai Chen. Physbrain: Human egocentric data as a bridge from vision language models to physical intelligence, 2025. URL https://arxiv.org/abs/2512.16793.

[33] Vincent Liu, Ademi Adeniji, Haotian Zhan, Siddhant Haldar, Raunaq Bhirangi, Pieter Abbeel, and Lerrel Pinto. Egozero: Robot learning from smart glasses, 2025. URL https://arxiv.org/abs/2505.20290.

[34] Hao Luo, Ye Wang, Wanpeng Zhang, Sipeng Zheng, Ziheng Xi, Chaoyi Xu, Haiweng Xu, Haoqi Yuan, Chi Zhang, Yiqing Wang, Yicheng Feng, and Zongqing Lu. Being-h0.5: Scaling human-centric robot learning for cross-embodiment generalization, 2026. URL https://arxiv.org/abs/2601.12993.

[35] 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. In *The Eleventh International Conference on Learning Representations*, 2022.

[36] Arjun Majumdar, Karmesh Yadav, Sergio Arnaud,

Yecheng Jason Ma, Claire Chen, Sneha Silwal, Aryan Jain, Vincent-Pierre Berges, Pieter Abbeel, Jitendra Malik, Dhruv Batra, Yixin Lin, Oleksandr Maksymets, Aravind Rajeswaran, and Franziska Meier. Where are we in the search for an artificial visual cortex for embodied intelligence? 2023.

[37] Suraj Nair, Aravind Rajeswaran, Vikash Kumar, Chelsea Finn, and Abhinav Gupta. R3m: A universal visual representation for robot manipulation. In *CoRL*, 2022.

[38] Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Charles Xu, Jianlan Luo, Tobias Kreiman, You Liang Tan, Pannag Sanketi, Quan Vuong, Ted Xiao, Dorsa Sadigh, Chelsea Finn, and Sergey Levine. Octo: An open-source generalist robot policy. In *Proceedings of Robotics: Science and Systems*, Delft, Netherlands, 2024.

[39] Open X-Embodiment Collaboration, Abhishek Padalkar, Acorn Pooley, Ajinkya Jain, Alex Bewley, Alex Herzog, Alex Irpan, Alexander Khazatsky, Anant Rai, Anikait Singh, Anthony Brohan, Antonin Raffin, Ayzaan Wahid, Ben Burgess-Limerick, Beomjoon Kim, Bernhard Schölkopf, Brian Ichter, Cewu Lu, Charles Xu, Chelsea Finn, Chenfeng Xu, Cheng Chi, Chenguang Huang, Christine Chan, Chuer Pan, Chuyuan Fu, Coline Devin, Danny Driess, Deepak Pathak, Dhruv Shah, Dieter Büchler, Dmitry Kalashnikov, Dorsa Sadigh, Edward Johns, Federico Ceola, Fei Xia, Freek Stulp, Gaoyue Zhou, Gaurav S. Sukhatme, Gautam Salhotra, Ge Yan, Giulio Schiavi, Hao Su, Hao-Shu Fang, Haochen Shi, Heni Ben Amor, Henrik I Christensen, Hiroki Furuta, Homer Walke, Hongjie Fang, Igor Mordatch, Ilija Radosavovic, Isabel Leal, Jacky Liang, Jaehyung Kim, Jan Schneider, Jasmine Hsu, Jeannette Bohg, Jeffrey Bingham, Jiajun Wu, Jialin Wu, Jianlan Luo, Jiayuan Gu, Jie Tan, Jihoon Oh, Jitendra Malik, Jonathan Tompson, Jonathan Yang, Joseph J. Lim, João Silvério, Junhyek Han, Kanishka Rao, Karl Pertsch, Karol Hausman, Keegan Go, Keerthana Gopalakrishnan, Ken Goldberg, Kendra Byrne, Kenneth Oslund, Kento Kawaharazuka, Kevin Zhang, Keyvan Majd, Krishan Rana, Krishnan Srinivasan, Lawrence Yunliang Chen, Lerrel Pinto, Liam Tan, Lionel Ott, Lisa Lee, Masayoshi Tomizuka, Maximilian Du, Michael Ahn, Mingtong Zhang, Mingyu Ding, Mohan Kumar Srirama, Mohit Sharma, Moo Jin Kim, Naoaki Kanazawa, Nicklas Hansen, Nicolas Heess, Nikhil J Joshi, Niko Suenderhauf, Norman Di Palo, Nur Muhammad Mahi Shafiullah, Oier Mees, Oliver Kroemer, Pannag R Sanketi, Paul Wohlhart, Peng Xu, Pierre Sermanet, Priya Sundaresan, Quan Vuong, Rafael Rafailov, Ran Tian, Ria Doshi, Roberto Martín-Martín, Russell Mendonca, Rutav Shah, Ryan Hoque, Ryan Julian, Samuel Bustamante, Sean Kirmani, Sergey Levine, Sherry Moore, Shikhar Bahl, Shivin Dass, Shuran Song, Sichun Xu, Siddhant Haldar, Simeon Adebola, Simon Guist, Soroush Nasiriany, Stefan Schaal, Stefan Welker, Stephen Tian, Sudeep Dasari, Suneel Belkhale, Takayuki Osa, Tatsuya Harada, Tatsuya Matsushima, Ted Xiao, Tianhe Yu, Tianli Ding, Todor Davchev, Tony Z. Zhao, Travis Armstrong, Trevor Darrell, Vidhi Jain, Vincent Vanhoucke, Wei Zhan, Wenxuan Zhou, Wolfram Burgard, Xi Chen, Xiaolong Wang, Xinghao Zhu, Xuanlin Li, Yao Lu, Yevgen Chebotar, Yifan Zhou, Yifeng Zhu, Ying Xu, Yixuan Wang, Yonatan Bisk, Yoonyoung Cho, Youngwoon Lee, Yuchen Cui, Yueh hua Wu, Yujin Tang, Yuke Zhu, Yunzhu Li, Yusuke Iwasawa, Yutaka Matsuo, Zhuo Xu, and Zichen Jeff Cui. Open X-Embodiment: Robotic learning datasets and RT-X models. https://arxiv.org/abs/2310.08864, 2023.

[40] Jonas Pai, Liam Achenbach, Victoriano Montesinos, Benedek Forrai, Oier Mees, and Elvis Nava. mimic-video: Video-action models for generalizable robot control beyond vlas, 2025. URL https://arxiv.org/abs/2512.15692.

[41] Younghyo Park, Jagdeep Singh Bhatia, Lars Ankile, and Pulkit Agrawal. Dexhub and dart: Towards internet scale robot data collection, 2024. URL https://arxiv.org/abs/2411.02214.

[42] Karl Pertsch, Kyle Stachowicz, Brian Ichter, Danny Driess, Suraj Nair, Quan Vuong, Oier Mees, Chelsea Finn, and Sergey Levine. FAST: Efficient action tokenization for vision-language-action models. *Robotics: Science and Systems*, 2025.

[43] 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, 2025. URL https://arxiv.org/abs/2509.19626.

[44] 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, 2025. URL https://arxiv.org/abs/2503.13441.

[45] Nur Muhammad Mahi Shafiullah, Anant Rai, Haritheja Etukuru, Yiqian Liu, Ishan Misra, Soumith Chintala, and Lerrel Pinto. On bringing robots home. *arXiv preprint arXiv:2311.16098*, 2023.

[46] Shuran Song, Andy Zeng, Johnny Lee, and Thomas Funkhouser. Grasping in the wild: Learning 6dof closed-loop grasping from low-cost demonstrations. *IEEE Robotics and Automation Letters*, 2020.

[47] Andrew Szot, Bogdan Mazoure, Omar Attia, Aleksei Timofeev, Harsh Agrawal, Devon Hjelm, Zhe Gan, Zsolt Kira, and Alexander Toshev. From multimodal llms to generalist embodied agents: Methods and lessons. *arXiv preprint arXiv:2412.08442*, 2024.

[48] 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.

[49] Gemini Robotics Team, Saminda Abeyruwan, Joshua Ainslie, Jean-Baptiste Alayrac, Montserrat Gonzalez Arenas, Travis Armstrong, Ashwin Balakrishna, Robert

Baruch, Maria Bauza, Michiel Blokzijl, et al. Gemini robotics: Bringing ai into the physical world. *arXiv preprint arXiv:2503.20020*, 2025.

[50] Physical Intelligence Team. $\pi_{0.6}$ model card. 2025.

[51] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. URL http://jmlr.org/papers/v9/vandermaaten08a.html.

[52] Homer Rich Walke, Kevin Black, Tony Z Zhao, Quan Vuong, Chongyi Zheng, Philippe Hansen-Estruch, Andre Wang He, Vivek Myers, Moo Jin Kim, Max Du, et al. BridgeData v2: A dataset for robot learning at scale. In *Conference on Robot Learning*, pages 1723–1736. PMLR, 2023.

[53] 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, 2023. URL https://arxiv.org/abs/2302.12422.

[54] 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, 2024. URL https://arxiv.org/abs/2403.07788.

[55] Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021.

[56] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.

[57] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[58] Junjie Wen, Yichen Zhu, Jinming Li, Minjie Zhu, Kun Wu, Zhiyuan Xu, Ning Liu, Ran Cheng, Chaomin Shen, Yaxin Peng, Feifei Feng, and Jian Tang. Tinyvla: Towards fast, data-efficient vision-language-action models for robotic manipulation. *arXiv preprint arXiv:2409.12514*, 2024.

[59] Junjie Wen, Yichen Zhu, Jinming Li, Zhibin Tang, Chaomin Shen, and Feifei Feng. Dexvla: Vision-language model with plug-in diffusion expert for general robot control. *arXiv preprint arXiv:2502.05855*, 2025.

[60] 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.

[61] Jonathan Yang, Catherine Glossop, Arjun Bhorkar, Dhruv Shah, Quan Vuong, Chelsea Finn, Dorsa Sadigh, and Sergey Levine. Pushing the limits of cross-embodiment learning for manipulation and navigation. *arXiv preprint arXiv:2402.19432*, 2024.

[62] Ruihan Yang, Qinxi Yu, Yecheng Wu, Rui Yan, Borui Li, An-Chieh Cheng, Xueyan Zou, Yunhao Fang, Xuxin Cheng, Ri-Zhao Qiu, Hongxu Yin, Sifei Liu, Song Han, Yao Lu, and Xiaolong Wang. Egovla: Learning vision-language-action models from egocentric human videos, 2025. URL https://arxiv.org/abs/2507.12440.

[63] Seonghyeon Ye, Joel Jang, Byeongguk Jeon, Sejune Joo, Jianwei Yang, Baolin Peng, Ajay Mandlekar, Reuben Tan, Yu-Wei Chao, Bill Yuchen Lin, et al. Latent action pretraining from videos. *arXiv preprint arXiv:2410.11758*, 2024.

[64] 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.

[65] Zhengdi Yu, Stefanos Zafeiriou, and Tolga Birdal. Dynhamr: Recovering 4d interacting hand motion from a dynamic camera, 2025. URL https://arxiv.org/abs/2412.12861.

[66] Chubin Zhang, Jianan Wang, Zifeng Gao, Yue Su, Tianru Dai, Cai Zhou, Jiwen Lu, and Yansong Tang. Clap: Contrastive latent action pretraining for learning vision-language-action models from human videos, 2026. URL https://arxiv.org/abs/2601.04061.

[67] Haoyu Zhen, Xiaowen Qiu, Peihao Chen, Jincheng Yang, Xin Yan, Yilun Du, Yining Hong, and Chuang Gan. 3d-vla: 3d vision-language-action generative world model. *arXiv preprint arXiv:2403.09631*, 2024.

[68] 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.

[69] 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, 2025. URL https://arxiv.org/abs/2509.04443.

[70] 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.

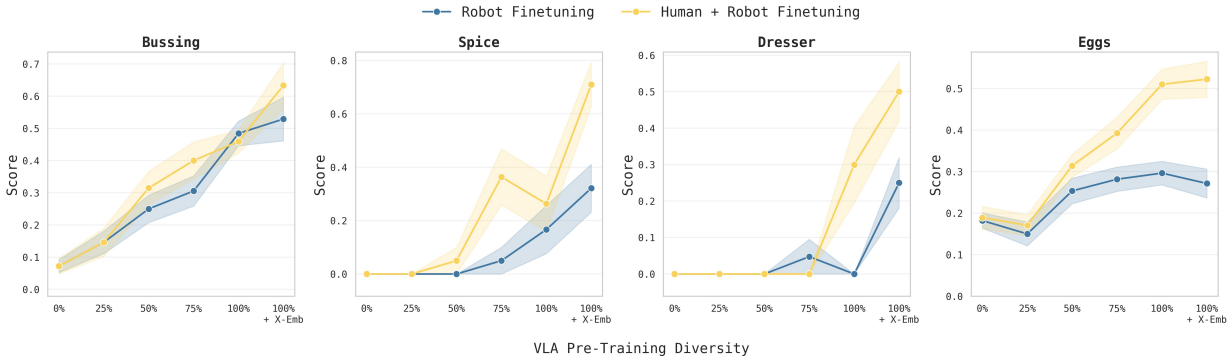

Fig. 14: **per task scaling graphs:** Across all tasks we see a clear upward trend in the efficacy of finetuning with human data as pretrained diversity increases. In `Eggs` we observe that despite pretraining diversity increasing, the model fails to generalize to this OOD task (blue line), but improves significantly with in-distribution human embodiment data (yellow line).

APPENDIX

A. Transfer Benchmark Details

Our human to robot transfer benchmark consists of 4 tasks, each testing a concept shown only in human data.

`Bussing`: The robot begins with nine kitchen items on a table, and must place each item in either the trash can or the bussing bin. The robot earns one point for each item correctly bussed, and we normalize score between [0, 1].

`Spice`: A kitchen island is set with three randomly placed spice bottles and a spice rack. The robot earns a point if it successfully places all spice bottles on the rack. This kitchen is unseen in robot data, but covered in human data.

`Dresser`: A bedroom dresser is set with a necklace and hair clip. The robot earns a point by correctly putting the necklace in the jewelry box and the hair clip in the accessories organizer. This bedroom is unseen in robot data, but covered in human data.

`Sort Eggs`: The table is set with two egg cartons (half dozen each) and a bowl containing six white and six brown eggs. The robot scores one point for each white egg it places in the left carton and brown egg in the right carton. An additional point is awarded for closing each egg carton shut at the end. Scores are normalized [0, 1]. While the robot data covers picking eggs, the concept of sorting eggs is only covered in human data.

For each experiment, we perform between 20 and 40 evaluations. All error bars visualize 1 standard error.

B. Results Analysis

In the context of our experiments, performance improves as a function of two parameters, pretrained model diversity and human data finetuning. To help disentangle these factors, we can compare the robot-only and human+robot performance scaling curves.

First we can consider the zero shot generalization of our robot only model (blue robot-only curve). For tasks like `Bussing` and `Spice`, the zero shot generalization to these new tasks increases steadily with increased pretrained diversity. However, for `Dresser` zero shot generalization only emerges in our strongest pretrained model. And for `Eggs`, performance quickly plateaus even with improved pretraining. This tells us that increasing pretrained diversity generally improves zero shot generalization across tasks, but it can be nonlinear.

Interestingly, across these tasks there are points in which the zero shot generalization and transferability from human data (yellow Human + Robot curve) are not necessarily correlated. For instance, with `Spice` 50%→75% we see a modest increase in zero shot generalization, but a large increase in transferability. Similarly for Dresser 75%→100% or Eggs 50%→100%+X-emb we see no improvement in zero shot generalization, but a large increase in transfer from human data. In other words, there are cases where increased pretraining diversity doesn't improve zero shot generalization, but it does improve human to robot transfer.

C. TSNE Visualization

To visualize how the model represents the human and robot data internally, we visualize the output embeddings of the VLA from human and robot data with a TSNE in Figure 5. Specifically, we pass observations from human and robot data through the co-finetuned VLA as input. We then mean-pool the first 200 output embeddings of the VLA (corresponding roughly to "task") and visualize them with the TSNE, capturing how well the VLA aligns different observations from humans and robots to the same task. We observe this alignment visibly improves with more diverse pre-training, even after all models have been finetuned on the same data.

D. Failure Modes

To better understand the nature of human-to-robot transfer, we categorize the failure modes of each task into two types: *reasoning failures*, in which the policy attempts the wrong subtask or places an object in the wrong location, and *dexterity failures*, in which the policy attempts the correct behavior but cannot execute the required motion (e.g., missed grasps, fumbled placements). We report the frequency of each failure type for both $\pi_{0.5}$ and $\pi_{0.5}$ + ego in Table I.

Overall $\pi_{0.5}$ fails early in the episode due to a reasoning failure, whereas $\pi_{0.5}$ + ego reduces these errors, but may still fail due to a lack of dexterity. The `Sort Eggs` task

TABLE I: Frequency of dexterity (D) and reasoning (R) errors for each task, comparing $\pi_{0.5}$ and $\pi_{0.5}$ + ego.

	Bussing		Spice		Dresser		Eggs	
	D	R	D	R	D	R	D	R
$\pi_{0.5}$	61%	39%	29%	39%	12%	62%	43%	52%
$\pi_{0.5+\text{ego}}$	65%	13%	23%	6%	26%	23%	15%	50%

TABLE II: Hand tracking benchmark on our in-house held-out set, H2O [27], and ARCTIC [19]. We report mPJPE (m) and PA-mPJPE (PA) in mm.

	In House		H2O		ARCTIC	
Method	m	PA	m	PA	m	PA
Ours	29.05	8.28	25.30	5.85	48.75	9.15

is an exception: here $\pi_{0.5}$ + ego primarily reduces *dexterity* failures associated with carefully placing eggs into the carton, in addition to learning the sorting concept. Taken together, this suggests that much of the transfer from human data improves the semantic and reasoning aspects of tasks, and that transferring dexterity from human video remains an important direction for future work.

E. Data Collection and Quality

1) Human Data Collection Protocol: For each task we collected between 3–5 hours of human video, corresponding to approximately 200–600 demonstrations per task. Episodes were reviewed by a QA team to check whether the demonstrator followed the prescribed strategy, and that both hands remained visible. Approximately 90% of episodes passed this review and were included in the final training mixture.

2) Hand Pose Estimation Benchmarking: Our pipeline relies on a closed-source 3D hand pose estimator to extract per-frame hand keypoints. We benchmark our pipeline on two open-source hand pose datasets, H2O [27] and ARCTIC [19], as well as an in-house dataset leveraging a multi-cam rig. We report mean per-joint position error (mPJPE) and procrustes-aligned mPJPE (PA-mPJPE) in millimeters in Table II. On the H2O dataset our estimator achieves an mPJPE of 25.3mm, which is comparable to recent open-source approaches such as Dyn-HaMR (22.5mm) [65].

F. Wrist Camera Details

Our human data collection apparatus uses a head-mounted iPhone together with two small wrist-worn cameras (Figure 6). To synchronize these streams, the demonstrator scans a QR code that encodes the current UTC timestamp at the beginning of each episode. We defer details to UMI [10] which used this same synchronization procedure. Empirically we find the resulting synchronization precision to be within 0.03s, i.e. at most one frame at 30Hz.

We chose to place our wrist cameras on the palm side because they maintain a clear view of the manipulated object as shown in Figure 6. In contrast, a camera on the wrist side is heavily occluded by the back of the operator's hand, making it difficult to see the manipulated object. We believe this choice minimizes domain shift by maintaining the object-centric nature of wrist views across embodiments.