arXiv:2505.20290v2 [cs.RO] 3 Jun 2025

EgoZero: Robot Learning from Smart Glasses

Vincent Liu[1*] **Ademi Adeniji**[12*] **Haotian Zhan**[1*]

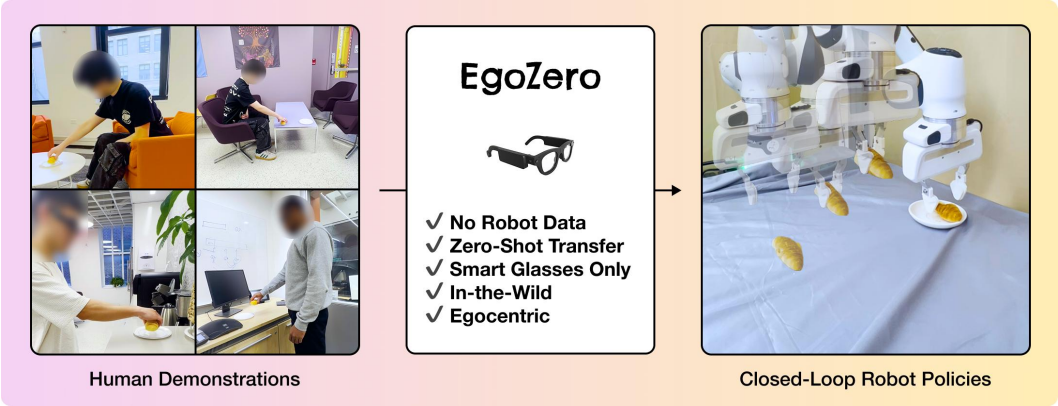
Siddhant Haldar[1] **Raunaq Bhirangi**[1] **Pieter Abbeel**[2] **Lerrel Pinto**[1]

[1]New York University [2]UC Berkeley

[*]Equal contribution

Abstract: Despite recent progress in general purpose robotics, robot policies still lag far behind basic human capabilities in the real world. Humans interact constantly with the physical world, yet this rich data resource remains largely untapped in robot learning. We propose EgoZero, a minimal system that learns robust manipulation policies from human demonstrations captured with Project Aria smart glasses, **and zero robot data**. EgoZero enables: (1) extraction of complete, robot-executable actions from in-the-wild, egocentric, human demonstrations, (2) compression of human visual observations into morphology-agnostic state representations, and (3) closed-loop policy learning that generalizes morphologically, spatially, and semantically. We deploy EgoZero policies on a gripper Franka Panda robot and demonstrate zero-shot transfer with 70% success rate over 7 manipulation tasks and only 20 minutes of data collection per task. Our results suggest that in-the-wild human data can serve as a scalable foundation for real-world robot learning — paving the way toward a future of abundant, diverse, and naturalistic training data for robots. Code and videos are available at https://egozero-robot.github.io.

Keywords: Imitation Learning, Robot Learning, Human Data



1 Introduction

Robots face significant challenges in replicating human generality and dexterity in the physical world. While deep learning has fueled progress in domains like language [1, 2], vision [3, 4, 5, 6, 7],

[*]Correspondence to Vincent Liu: vincent.liu15@gmail.com

speech [8, 9, 10], and complex games [11, 12], these successes rely on internet-scale datasets that are tightly aligned with downstream applications. In robotics, collecting similarly large and diverse datasets that match real-world deployment conditions remains a fundamental bottleneck [13].

We argue that the data bottleneck stems not from a shortage of physical labor in the real world, but from the unresolved challenge of effectively capturing and representing human behavior for robot learning. Humans perform a wide range of dexterous tasks in natural environments every day, representing an untapped, renewable source of rich, real-world data. Although recent works have attempted to use human demonstrations as supervision for robot learning, they have limitations to scalability such as additional wearables [14], robot data [15], multi-camera calibration [16], online fine-tuning [17], low-precision affordance-based policies [18, 19], or data processing hacks to cross the human-robot morphology gap [14, 15, 20]. Other general vision-based learning approaches pretrain on large multi-robot datasets [13, 21], which produce visual representations that are robust across morphologies present in their training mixes [22, 23, 24, 25], but have yet to show zero-shot transfer purely from human data.

In this work, we tackle the ambitious question: can robots learn zero-shot manipulation skills from only egocentric in-the-wild human data? To answer this, we introduce EGOZERO: a lightweight framework that enables robots to learn manipulation policies directly from egocentric in-the-wild human demonstrations, captured using only Project Aria smart glasses [26]. EGOZERO eliminates the need for teleoperation, calibration, or additional wearables, allowing humans to interact with the world freely while still providing robot supervision. Inspired by [16, 27], EGOZERO overcomes the morphology gap by representing states and actions as compact sets of points. Point-based representations simultaneously unify human and robot distributions, improve sample efficiency and interpretability of policy learning, and generalize to new visual scenes and morphologies. However, egocentric in-the-wild data collection, does not have access to the multi-camera calibration setup used in [16, 27] to accurately compute point representations. Therefore, we introduce methods to accurately derive state and action representations from raw visual and odometric inputs.

We evaluate EGOZERO by training manipulation policies on human demonstrations recorded by Aria and deploying them on a Franka Panda robot. Our policies achieve an average zero-shot success rate of 70% across tasks such as grasping, opening, and pick-and-place in unseen real-world environments — without any robot-collected training data. By rethinking the data representation and policy learning stack to be morphology-agnostic from the ground up, EGOZERO is a step toward building robots that can learn from the vast diversity of real-world human experiences. Our contributions are as follows:

- EGOZERO policies achieve a 70% zero-shot success rate on our tasks, **trained only on human data recorded with Project Aria smart glasses**. EGOZERO, to our knowledge, represents the first approach that successfully transfers in-the-wild, human data into closed-loop policies with no robot data.
- EGOZERO policies exhibit strong zero-shot generalization properties with only 100 training demonstrations (20 minutes of data collection), demonstrating the robustness, transferability, and data efficiency of learning from unified 3D state-action representations.
- EGOZERO achieves high success rate when evaluated on new camera viewpoints, spatial configurations, and object instances that are often completely out-of-distribution — validating our proposed method of extracting accurate 3D representations from objects when accurate depth measurements are not available.

2 Related Work

Imitation learning. Imitation learning has emerged as a powerful paradigm in robotics, enabling robots to acquire complex skills by learning directly from real-world demonstrations [28]. By observing and replicating expert behavior, robots can bypass the need for hand-engineered solutions to manipulation tasks, making this approach particularly conducive to domains with high-dimensional

state and action spaces [29, 30]. Teleoperation is one of the most widely used methods for imitation learning from real-world data collection and has been extensively studied in the robotics literature. In this approach, a human teleoperator commands a robot to complete a desired task, recording the robot's states and actions in the process. The collected data is then used to train a policy that predicts actions from states via supervised learning [31, 32, 33, 34, 35].

Learning from human motion. Because teleoperation is difficult to scale due to its hardware requirements, learning manipulation directly from humans has become a growing area of interest. Prior work has explored mapping human grasps to robot manipulators using vision-based representations like the "contact web" [36], and more recently, has introduced semantic constraints to encode the implicit common sense required for household tasks [37]. Other methods to capture human proprioception include "inside-out" motion capture systems such as VR headsets and dongles, which do not use external sensing devices [38, 39] and are bulky, tethered, and susceptible to occlusion. SLAM-based wearable camera systems [14] and VR wrist trackers such as the SteamVR wrist trackers [40] are vision-based and do not require external transmitters for localization, but can drift and become inaccurate. Self-tracking vision methods require extensive calibration and mapping of each environment a priori [14]. For capturing local information such as finger movements, motion capture gloves such as Rokoko and Manus Metagloves are highly accurate [41, 42, 43, 44]. These gloves use resistive strain sensing, capacitative sensing, and electromagnetic field sensing to track precise finger information in the local hand frame.

Learning from egocentric video. Because of the accessibility of video, several recent works try to learn and extract hand data from egocentric videos of humans. Datasets such as [45, 46, 47, 48, 49] represent large-scale efforts to collect egocentric videos of humans interacting with objects in diverse real-world scenes. [27, 16] use point-based representations to unify human video and robot training data, while [20] modifies human videos with image editing models to create robot training data. [15, 50] propose hardware solutions such as smart glasses and multi-camera data collection platforms to collect dexterous hand video datasets, while [18, 51, 52, 53, 19] introduce methods for extracting control-based affordances for manipulation from vision. Many of the approaches that estimate hand pose information from one or more camera inputs are facilitated by hand-pose estimation models such as [54, 55, 56, 57]. These models are trained with imitation learning to predict hand keypoints [58] from monocular visual input. Although effective in many simple domains, these models are brittle to occlusions, temporally inconsistent, and lack robustness to background distractors.

3 EGOZERO

In this section, we describe EGOZERO, a system for collecting in-the-wild egocentric human data and training morphology-agnostic robot manipulation policies.

3.1 Human-Robot Domain Unification

Project Aria smart glasses. The Project Aria smart glasses come with several sensors, an SDK, and additional Machine Perception Services (MPS) [26]. We use the fisheye RGB camera and 2 SLAM cameras for data capture. We obtain accurate online 6DoF hand poses, camera intrinsics, and camera extrinsics from MPS. We record demonstrations of RGB images, 6DoF palm poses, and 6DoF camera extrinsics, which we denote for a timestep t as (I_t, H_t, T_t), respectively. We linearize I_t as a 1408x1408 RGB image with known camera projection function $\mathcal{P}$ and $H_t, T_t \in SE(3)$ are homogeneous transformation matrices representing the hand pose in camera frame and the camera frame in world frame, respectively.

Traditionally, $\mathcal{S}$ represents the robot's space of visual states and $\mathcal{A}$ represents the robot's native executable actions. Similar to [16], we define the morphology-agnostic state and action spaces $\tilde{\mathcal{S}}$ and $\tilde{\mathcal{A}}$, respectively, in egocentric frame. In this section, we describe how to extract $\tilde{\mathcal{S}} \times \tilde{\mathcal{A}}$ from a demonstration $\{(I_t, H_t, T_t)\}_{t=1}^{L}$.

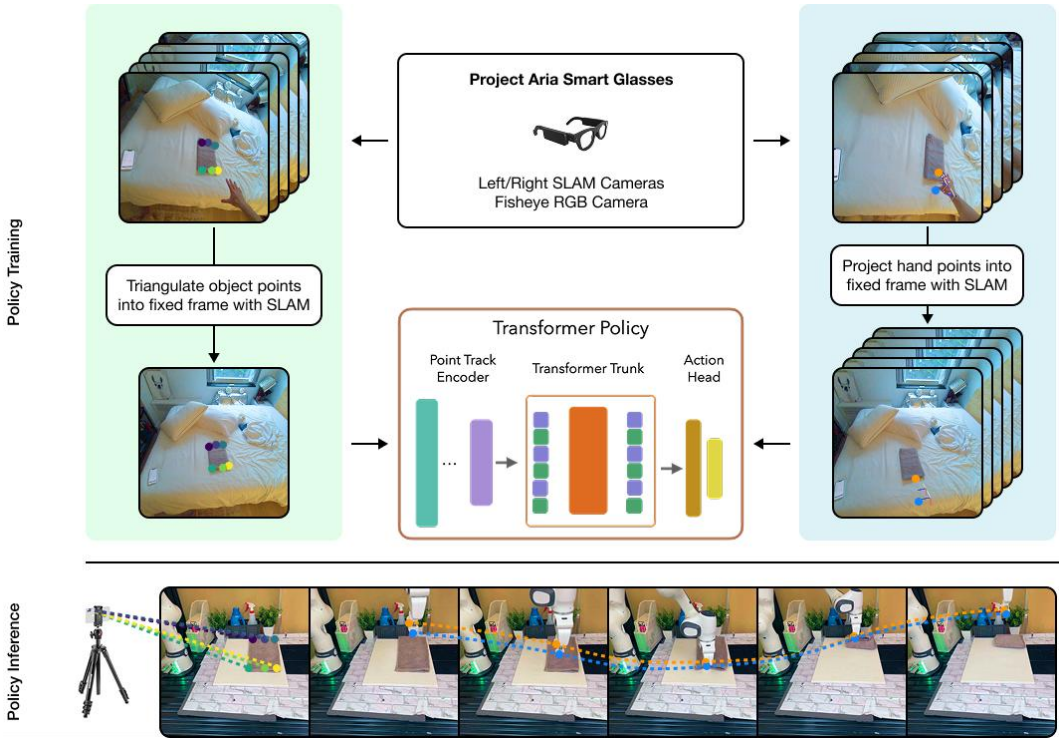


Figure 1: EGOZERO trains policies in a unified state-action space defined as egocentric 3D points. Unlike previous methods which leverage multi-camera calibration and depth sensors, EGOZERO localizes object points via triangulation over the camera trajectory, and computes action points via Aria MPS hand pose and a hand estimation model. These points supervise a closed-loop Transformer policy, which is rolled out on unprojected points from an iPhone during inference.

Unified action space. We define $\tilde{\mathcal{A}}$ as the concatenated space of 3D end-effector egocentric coordinates and gripper closures [14]. Aria only provides H_t, which contains no end-effector information except for hand pose [15]. We use HaMeR [54] to compute the 21-keypoint egocentric hand model, $h_t \in \mathbb{R}^{21\times 3}$. Though HaMeR's end-effector predictions in camera frame are inaccurate, its predictions localized in hand frame are more reliable. Therefore, we compose local hand deformation from HaMeR with egocentric hand information from Aria. First, we construct HaMeR's palm in camera frame as $\hat{H}_t \in SE(3)$: the translation is the centroid of the ThumbCMC, IndexMCP, and MiddleMCP points; the rotation is the basis constructed by the Wrist-MiddleMCP and IndexMCP-MiddleMCP vectors. We then use H_t to correct $\hat{H}_t$ in egocentric frame through the palm frames. Finally, we project the corrected hand pose $H_t^{-1}\hat{H}_t$ into the first frame [14, 15]. This can be represented as a single chain of homogeneous transformations

$$\tilde{h}_t = T_0^{-1} T_t H_t^{-1} \hat{H}_t h_t \tag{1}$$

To detect grasps, we threshold the Euclidean distance between the thumb and index coordinates. Our final action is the concatenated vector of thumb and index coordinates and gripper closure.

Unified state space. We define $\tilde{\mathcal{S}}$ as the concatenated space of egocentric object point sets and robot end-effector actions. Extracting point representations of objects requires either triangulation from multiple cameras or unprojection with depth, but the Project Aria glasses provide neither[2]. Furthermore, monocular metric depth models are inconsistent and inaccurate even with grounding, which we show in Appendix D. Instead, we rely on Aria's accurate SLAM extrinsics and CoTracker3 [59] to triangulate 2D points over the demonstration trajectory. This makes the following assumptions:

[2]Though there are 3 visual cameras (1 RGB, 2 SLAM), they have little field-of-view overlap, making stereo triangulation unreliable https://github.com/facebookresearch/projectaria_tools/issues/64.

(1) the object is stationary pre-grasp, (2) there is enough camera movement, and (3) the environment is not stochastic. As such, the object state is static for the entire demonstration.

We first label a set of 2D points [27, 16]. For each expert-labeled point, we use Grounding DINO [60] and DIFT [61] to map its UV coordinates onto the start frame, and track these points with CoTracker3 [59] to obtain a trajectory of (T_t, u_t) pairs where $u_t \in \mathbb{R}^2$ and $T_t \in SE(3)$ is the camera pose in world frame. We wish to solve for the $\mathbf{q}^*$ in the first frame ($t = 0$) that minimizes the pixel reprojection error in each frame. First, we find a set of inlier frames $\mathcal{I}$ via epipolar geometric consistency and RANSAC triangulation. CoTracker3 oftentimes predicts points that lag behind camera movement, giving the impression that a point is further in space than it actually is. To account for this "stickiness," we add a soft depth penalty to prefer closer solutions when there are multiple points in the cone of solutions that minimize reprojection error. Therefore, we solve

$$\mathbf{q}^* = \arg\min_{\mathbf{q}} \sum_{i \in \mathcal{I}} \left|\left| u_i - \mathcal{P}\left(T_0^{-1} T_i \mathbf{q}\right) \right|\right|_\rho + \lambda \mathbf{q}_z \tag{2}$$

where $|| \cdot ||\rho$ is the Huber loss, $\mathcal{P}$ is the camera projection function, and λ is the depth penalty weight. In practice, (T_t, u_t) are accurate, so $\mathcal{I}$ contains most of the frames and optimization converges strongly to a mean inlier reprojection error of 2-4 pixels per demonstration. Finally, we order and concatenate all triangulated points to represent the object state, $\tilde{s}$. We provide comprehensive mathematical equations for this procedure in Appendix B.

3.2 Learning a Robot Policy on Human Data

Policy learning. We collect N human demonstrations and process them into a dataset $\mathcal{D} = \{(\tilde{s}^{(i)}, \tilde{a}^{(i)})\}_{i=1}^N$. We train a closed-loop Transformer policy [27] $\pi_\theta : \tilde{\mathcal{S}} \mapsto \tilde{\mathcal{A}}$ with behavior cloning over $\mathcal{D}$. We model the policy's predictions as the mean of a normal distribution and train it to minimize the negative log likelihood function

$$\theta = \arg\min_{\theta} \mathbb{E}_{(\tilde{s},\tilde{a})\sim\mathcal{D}} \left[\frac{||\pi_\theta(\tilde{s}) - \tilde{a}||^2}{2\sigma^2} \right] \tag{3}$$

where $\sigma = 0.1$ [62, 27]. We augment the policy with a history buffer input and temporally aggregated action chunking [62, 27]. We randomly inject noise into the object points and apply random 3D transformations to the states and actions of each training episode [14], which we show is necessary for in-the-wild transfer in Section 4.3. To do so, we sample random rotations $R \sim \mathcal{U}(-\pi/6, +\pi/6)$ radians and translations $t \sim \mathcal{U}(-0.5, +0.5)$ meters. We remove stationary points by throwing out consecutive points whose Euclidean distance is less than 1cm, which is necessary to disambiguate the association between proprioceptive position and grasp closure. For longer tasks, we subsample the demonstrations by a factor of 2. To discard noisy training examples from DIFT failures, we discard demonstrations whose object points are more than 1 median absolute deviation distance from the closest human fingertip point.

Policy inference. In inference, we initialize the robot state 30 centimeters above the middle of its workspace. We use Grounding DINO and DIFT to crop and map the expert-labeled UV coordinates onto the start frame. We use an iPhone to represent the stationary egocentric view since it allows us to unproject points into 3D with accurate depth. To map the policy's 3D predictions into robot frame, we calibrate the iPhone-to-robot transform once at the start of inference. We binarize the model's gripper predictions at 0 to produce gripper actions in $\{-1, 1\}$. In our experiments, we use a Franka Panda gripper robot, whose controller produces robot-executable actions via the inverse kinematics mapping $\tilde{\mathcal{A}} \mapsto \mathcal{A}$.

4 Experiments

In this section, we compare EGOZERO with baselines adapted from related works and ablate some of EGOZERO's core components. From these comparisons, we demonstrate how our specific design choices make zero-shot in-the-wild transfer possible. We also explore the generalization properties that emerge from EGOZERO's unified state-action representation space.

Figure 2: Our 7 tasks. Top: open oven door, put bread on plate, sweep board with broom, erase board. Bottom: sort fruit, fold towel, and insert book in shelf. See Appendix A for full trajectories.

4.1 Experimental Setup

We evaluate EGOZERO on a Franka Panda gripper robot. We use an iPhone to represent the egocentric point of view and calibrate this to the robot's frame once per evaluation via an Aruco tag, which we cover during policy inference. We collect 100 demonstrations per task, varying the environment and object positions. **We collect zero data in our inference-time environment.** We evaluate our method on the following manipulation tasks:

- *Open oven door.* The robot arm grasps and pulls down the handle of an oven door. The position of the oven is varied for each evaluation.
- *Put bread on plate.* The robot arm picks up a deformable slice of bread from the table and puts it on the plate. The positions of the bread are varied for each evaluation.
- *Sweep board with broom.* The robot arm picks up a mini broom from the basket and sweeps a wooden board. The positions of the broom, basket, and board are varied for each evaluation.
- *Erase board.* The robot arm picks up a whiteboard eraser from the table and erases a whiteboard with it. The positions of the eraser and board are varied for each evaluation.
- *Sort fruit into bowl.* The robot arm is prompted to pick up one of a lemon, lime, and tangerine, and drop it into a bowl. The positions of the fruits and bowl are varied for each evaluation.
- *Fold towel.* The robot arm lifts one end of the towel (closest to the camera) and folds it onto the other end of the towel. The position of the towel is varied for each evaluation.
- *Insert book in shelf.* The robot arm picks up a book and inserts it into a shelf. The positions of the book and shelf are varied for each evaluation.

4.2 Baselines

In this section, we demonstrate why our specific formulation of policy learning enables zero-shot transfer from in-the-wild human behaviors. Because no prior work operates under the same assumptions as ours — learning a closed-loop policy in-the-wild, untethered, without robot data, from only smart glasses — we adapt some ideas inspired by past works to our setting.

Learning from images. We implement a variation of Baku [62] that predicts actions in our unified action space from image inputs. Due to the large differences in visual distributions between humans and robots, it is difficult to learn a closed-loop policy from human video with zero-shot robot transfer. [15] only shows experiments using human video from Aria glasses as supplementary to robot data,

Method	Open oven	Pick bread	Sweep broom	Erase board	Sort fruit	Fold towel	Insert book
From vision [62]	0/15	0/15	0/15	0/15	0/15	0/15	0/15
From affordances [18]	12/15	0/15	0/15	0/15	7/15	10/15	5/15
EGOZERO - 3D augmentations	0/15	0/15	0/15	0/15	0/15	0/15	0/15
EGOZERO - triangulated depth	0/15	0/15	0/15	0/15	0/15	0/15	0/15
EGOZERO	**13/15**	**11/15**	**9/15**	**11/15**	**11/15**	**10/15**	**9/15**

Table 1: Success rates for all baselines and ablations. All models were trained on the same 100 demonstrations per task, and evaluated on zero-shot object poses (unseen from training), cameras (iPhone vs Aria), and environment (robot workspace vs in-the-wild). Because of limited prior work in our exact zero-shot in-the-wild setting, we cite the closest work for each baseline.

requiring careful renormalization of the human data distribution. Furthermore, Aria's fisheye lens exacerbates this problem by warping the 2D-3D correspondence non-uniformly across space and time. Learning 3D distributions from 2D context clues becomes more reliable with abundant visual data produced by similar robot and camera distributions [63, 24, 25].

Learning from affordances. [18, 19] explores learning from egocentric human video data without robot data in affordance-based settings. Typically, this is done by relying on an open-loop trajectory generated by a pretrained grasp model. We ablate our closed-loop formulation by predicting proprioceptive landmarks similar to [18] — specifically, the initial and final grasp, executing a linear trajectory between them during inference. Although policy learning from affordances is simple with 3D representations, it fails on tasks that require complex nonlinear motions, such as our "put bread in plate" and "erase board " tasks. When deployed on the robot, these policies exhibit incorrect behavior: the robot attempts to drag the bread onto the plate and pushes the board with the eraser. In other partially successful tasks, the policy fails by generating trajectories that are too simple, often bumping other objects during execution. These failures demonstrate that closed-loop policies are necessary to learn complex motions with greater precision, even when the object state is not tracked.

4.3 Ablations

In this section, we explore the critical design components that make zero-shot transfer from in-the-wild human data possible. Through our ablative experiments, we argue that the fully egocentric framework necessitates some aspects of policy learning that were not important in more constrained settings.

3D augmentations. Although 3D augmentations have been explored before [14], we show that they are indeed necessary for zero-shot in-the-wild transfer. In the unified 3D state-action space, the policy learns a dense 3D-to-3D mapping [16]. Without 3D augmentations, the policy learns a smaller and sparser 3D-to-3D mapping volume. As a result, the policy does not interpolate between 3D positions as well and is less robust to new positions. Therefore, it is often out-of-distribution when given a new egocentric view. We demonstrate that, when trained with 3D augmentations, our policies generalize to object configurations that are many standard deviations outside of the volume of their training data. Although our policy learning framework is similar to [27, 16], these works do not need 3D augmentations to show good success rates, implying that learning robust policies on egocentric data introduces extra complexity in learning generalizable representations. We visualize the training and inference distributions of object points in Figure 3.

Monocular depth estimation. The Aria glasses do not provide a way of extracting ground truth depth information: (1) it cannot triangulate objects reliably since the overlapping field-of-view between all cameras is narrow; (2) it does not have any built-in lidar or depth sensors. Therefore, we localize the object via triangulation over the camera trajectory to obtain its 3D information. To show

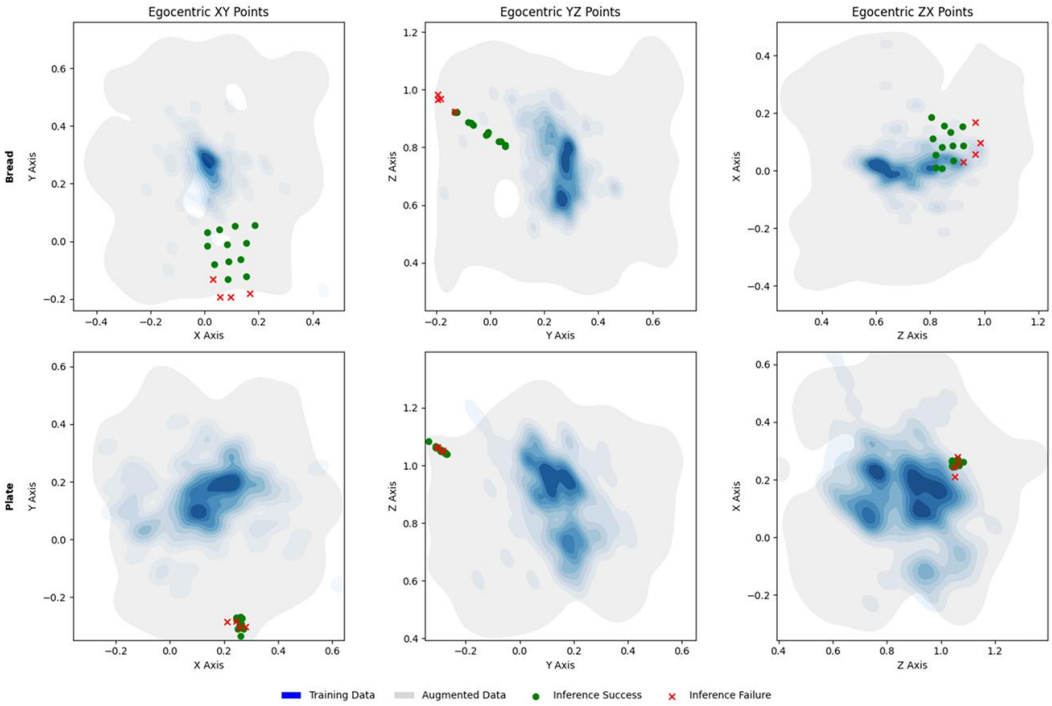


Figure 3: Distribution of bread keypoints for "Put bread in plate" task. The columns are projections of the 3D space onto each 2D plane. The policy generalizes to object poses far outside of its training volume and begins to fail when the objects are near the limits of its augmented volume.

that monocular metric depth models are not a viable option, we ablate our triangulation method with unprojection from a metric depth model [64]. We observe that the best metric depth models, even when grounded with many Aruco tags in the scene, produce depth measurements of >5cm error. This suggests that the depth maps are warped unevenly, potentially by the distortion caused by Aria's fisheye. All policies trained with estimated depth fail unequivocally. We describe our grounding method in Appendix D.

4.4 Zero-shot generalization

Object pose generalization. In both data collection and robots evaluation, we vary the poses of the objects. If there are multiple objects, we also vary their locations relative to each other. We observe that the use of correspondence with 3D state representations encodes the pose of the object [27, 16] and allows our policies to generalize from in-the-wild data. We notice that there is much more spatial diversity in our human demonstrations than what the robot can access in its workspace. This diversity, combined with 3D augmentations, regularizes the policy to learn a more general solution across a larger 3D volume, which enables zero-shot transfer to the robot. We constrain the diversity of object poses to represent what a human will realistically manipulate (i.e. the oven door is visible to the camera).

Object semantic generalization. Following [27, 16], we also demonstrate that 3D representations allow for zero-shot object category generalization. Because our training and inference images are so different (Aria fisheye vs iPhone pinhole), we introduce Grounding DINO to crop images to improve DIFT's success rate; this is not something that [27, 16] implement because their cameras and backgrounds are identical between training and inference. Because Grounding DINO is language-conditioned, we simply prompt it with the object category (i.e. "a toaster oven.") to allow it to generalize to entirely new object instances. This ensembling of pretrained models compresses visual

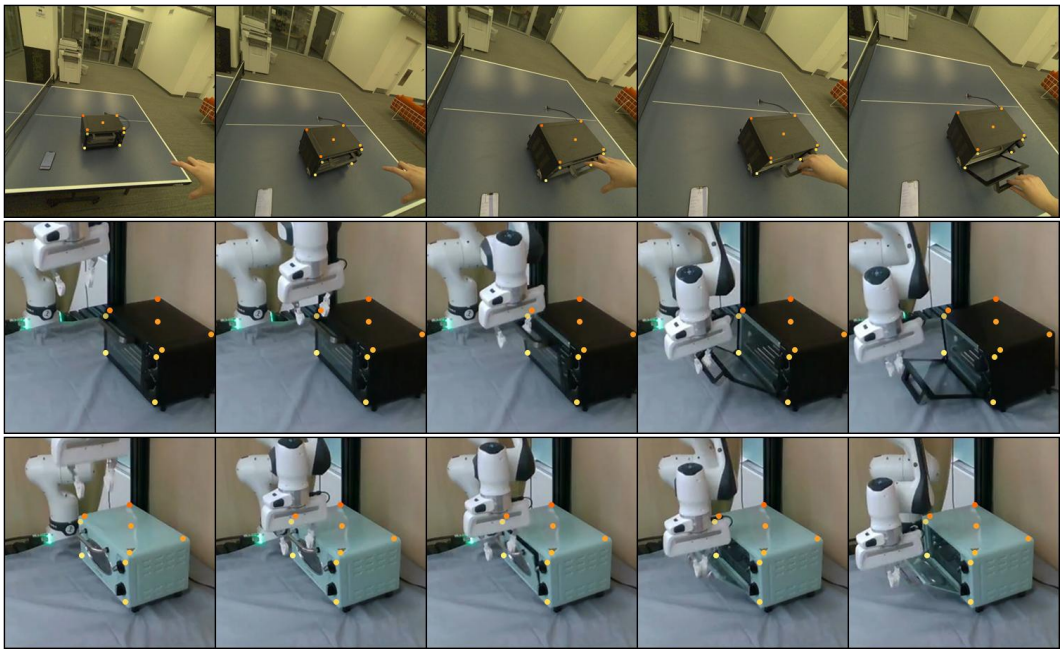


Figure 4: Object semantic generalization. Human demonstrations are done with only black ovens (top). The policy transfers zero-shot to the robot with the same oven (middle) and also generalizes to a new oven instance (bottom). The points are color-coded to represent the correspondence.

diversity into geometric abstractions that allow EGOZERO to generalize across visual distributions in the egocentric setting.

Camera generalization. One of the biggest limiting factors of vision-based policies is that learning invariance to small changes in individual pixels is data intensive. For a policy to generalize to novel viewing angles, distances, and cameras, it must be trained on a large amount of data from similar visual distributions. For example, [24] is trained on 10k+ hours of cross-embodiment data, but its zero-shot performance is significantly lower when the inference camera (and end-effector) is different from the one used to collect its robot training data. To navigate this issue, [15] uses Aria glasses for human data collection, robot data collection, and policy inference, but still require several hours of both human and robot data and careful renormalization to reach good success rates. Because EGOZERO learns policies from 3D point sets, EGOZERO is completely camera-agnostic. We demonstrate this in all our experiments by using an iPhone in inference.

Human-scale generalization. For each task, we collect data in 2-3 different environments, on tabletops of different heights, with various background distractors, with multiple unique demonstrators. We perform our demonstrations moving around, standing still, and sitting down. The variance in human demonstrators provides added diversity in the training data. These differences in height and grasp are still encoded in the same unified representation space.

4.5 Limitations

Limitations of 3D representations. The largest source of error during inference comes from the correspondence model DIFT [61]. Correspondence encodes pose by ordering the state space, making policy learning sample efficient [27, 16]. At larger data scale, pose information can be learned directly from dense unordered geometric information (i.e. using grounded segmentation models [65]). The correspondence errors are a symptom of perhaps a more general limitation: that the policy is upper-bounded by the accuracy of its 3D point inputs. Though policy learning is made simple with 3D points, it does not have information to correct 3D measurement errors.

Limitations of triangulation. We rely on Structure-from-Motion to localize objects over Aria's pre-grasp trajectory. Although this algorithm is less robust when the camera has limited movement, we find that the camera movement from natural task demonstration is usually sufficient. Furthermore, triangulation requires stationary objects, which means that we cannot track objects. In the future, stereo cameras or cheap lidar can remove these constraints and allow closed-loop policy learning in stochastic settings. We hope that depth estimation will become easier with better hardware design.

Limitations of hand models. In this work, we use [54] and Aria's hand pose to extract a complete action space, both of which introduce slight inaccuracies. Aria's hand pose does not always predict the same location on the hand and [54] predicts inconsistently incorrect rotational and translational components on the hand. Even when carefully Equation 1 is tuned, the action labels contain 1-2cm error, preventing the policy from solving high-precision tasks. We hope that hand estimation methods will become more reliable with better research and hardware design.

5 Discussion

In this work, we presented **EGOZERO, a minimal system that trains zero-shot robot policies on in-the-wild egocentric human data without any robot data**. We formalize the morphology-agnostic state-action spaces from prior works and demonstrate how point representations hold the same properties in egocentric in-the-wild settings. Because EGOZERO optimizes for data collection ergonomics, we also demonstrate how to extract unified state and action representations from human data recorded with the Project Aria smart glasses as the only hardware. As a result, we introduce novel data processing and policy learning design; we demonstrate the importance of each of these components in our baseline and ablation experiments. Although EGOZERO represents an initial proof-of-concept of how to achieve strong zero-shot transfer from human data, we also acknowledge a handful of limitations, many of which we hope will improve as hardware and robot learning methods improve together.

Towards human-centric robotics. Ultimately, human data carries huge potential in its scalability and morphological completeness. We hope that EGOZERO will serve as a framework on which future research can extend to fully dexterous and bimanual setups. We hope that our work offers a potentially new theme in robots that is more human-centric, scalable, and abundant.

Acknowledgments

We thank Zhuoran Chen at New York University for their helpful feedback and contribution to the website. We also thank Joyce Kwak for editing the videos. This work was supported by grants from Microsoft, Honda, Hyundai, NSF award 2339096, and ONR award N00014-22-1-2773. LP is supported by the Sloan and Packard Fellowships.

References

[1] OpenAI. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

[2] OpenAI. Language models are few-shot learners. *CoRR*, abs/2005.14165, 2020. URL https://arxiv.org/abs/2005.14165.

[3] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever. Zero-shot text-to-image generation, 2021. URL https://arxiv.org/abs/2102.12092.

[4] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models, 2022. URL https://arxiv.org/abs/2112.10752.

[5] Imagen-Team-Google. Imagen 3, 2024. URL https://arxiv.org/abs/2408.07009.

[6] T. Brooks, B. Peebles, C. Holmes, W. DePue, Y. Guo, L. Jing, D. Schnurr, J. Taylor, T. Luhman, E. Luhman, C. Ng, R. Wang, and A. Ramesh. Video generation models as world simulators. 2024. URL https://openai.com/research/video-generation-models-as-world-simulators.

[7] A. Blattmann, T. Dockhorn, S. Kulal, D. Mendelevitch, M. Kilian, D. Lorenz, Y. Levi, Z. English, V. Voleti, A. Letts, V. Jampani, and R. Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets, 2023. URL https://arxiv.org/abs/2311.15127.

[8] Z. Evans, J. D. Parker, C. Carr, Z. Zukowski, J. Taylor, and J. Pons. Stable audio open, 2024. URL https://arxiv.org/abs/2407.14358.

[9] C. Wang, S. Chen, Y. Wu, Z. Zhang, L. Zhou, S. Liu, Z. Chen, Y. Liu, H. Wang, J. Li, L. He, S. Zhao, and F. Wei. Neural codec language models are zero-shot text to speech synthesizers, 2023. URL https://arxiv.org/abs/2301.02111.

[10] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever. Robust speech recognition via large-scale weak supervision, 2022. URL https://arxiv.org/abs/2212.04356.

[11] D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, T. Lillicrap, K. Simonyan, and D. Hassabis. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419): 1140–1144, 2018. doi:10.1126/science.aar6404. URL https://www.science.org/doi/abs/10.1126/science.aar6404.

[12] OpenAI. Dota 2 with large scale deep reinforcement learning, 2019. URL https://arxiv.org/abs/1912.06680.

[13] E. Collaboration. Open x-embodiment: Robotic learning datasets and rt-x models, 2024. URL https://arxiv.org/abs/2310.08864.

[14] C. Wang, H. Shi, W. Wang, R. Zhang, L. Fei-Fei, and C. K. Liu. Dexcap: Scalable and portable mocap data collection system for dexterous manipulation, 2024. URL https://arxiv.org/abs/2403.07788.

[15] S. Kareer, D. Patel, R. Punamiya, P. Mathur, S. Cheng, C. Wang, J. Hoffman, and D. Xu. Egomimic: Scaling imitation learning via egocentric video, 2024. URL https://arxiv.org/abs/2410.24221.

[16] S. Haldar and L. Pinto. Point policy: Unifying observations and actions with key points for robot manipulation, 2025. URL https://arxiv.org/abs/2502.20391.

[17] I. Guzey, Y. Dai, G. Savva, R. Bhirangi, and L. Pinto. Bridging the human to robot dexterity gap through object-oriented rewards, 2024. URL https://arxiv.org/abs/2410.23289.

[18] S. Bahl, A. Gupta, and D. Pathak. Human-to-robot imitation in the wild, 2022. URL https://arxiv.org/abs/2207.09450.

[19] J. Shi, Z. Zhao, T. Wang, I. Pedroza, A. Luo, J. Wang, J. Ma, and D. Jayaraman. Zeromimic: Distilling robotic manipulation skills from web videos, 2025. URL https://arxiv.org/abs/2503.23877.

[20] M. Lepert, J. Fang, and J. Bohg. Phantom: Training robots without robots using only human videos, 2025. URL https://arxiv.org/abs/2503.00779.

[21] A. Khazatsky. Droid: A large-scale in-the-wild robot manipulation dataset, 2024. URL https://arxiv.org/abs/2403.12945.

[22] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, J. Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022.

[23] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, X. Chen, K. Choromanski, T. Ding, D. Driess, A. Dubey, C. Finn, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023.

[24] K. Black, N. Brown, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, L. Groom, K. Hausman, B. Ichter, S. Jakubczak, T. Jones, L. Ke, S. Levine, A. Li-Bell, M. Mothukuri, S. Nair, K. Pertsch, L. X. Shi, J. Tanner, Q. Vuong, A. Walling, H. Wang, and U. Zhilinsky. π_0: A vision-language-action flow model for general robot control, 2024. URL https://arxiv.org/abs/2410.24164.

[25] P. Intelligence, K. Black, N. Brown, J. Darpinian, K. Dhabalia, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, M. Y. Galliker, D. Ghosh, L. Groom, K. Hausman, B. Ichter, S. Jakubczak, T. Jones, L. Ke, D. LeBlanc, S. Levine, A. Li-Bell, M. Mothukuri, S. Nair, K. Pertsch, A. Z. Ren, L. X. Shi, L. Smith, J. T. Springenberg, K. Stachowicz, J. Tanner, Q. Vuong, H. Walke, A. Walling, H. Wang, L. Yu, and U. Zhilinsky. $\pi_{0.5}$: a vision-language-action model with open-world generalization, 2025. URL https://arxiv.org/abs/2504.16054.

[26] J. Engel, K. Somasundaram, M. Goesele, A. Sun, A. Gamino, A. Turner, A. Talattof, A. Yuan, B. Souti, B. Meredith, C. Peng, C. Sweeney, C. Wilson, D. Barnes, D. DeTone, D. Caruso, D. Valleroy, D. Ginjupalli, D. Frost, E. Miller, E. Mueggler, E. Oleinik, F. Zhang, G. Somasundaram, G. Solaira, H. Lanaras, H. Howard-Jenkins, H. Tang, H. J. Kim, J. Rivera, J. Luo, J. Dong, J. Straub, K. Bailey, K. Eckenhoff, L. Ma, L. Pesqueira, M. Schwesinger, M. Monge, N. Yang, N. Charron, N. Raina, O. Parkhi, P. Borschowa, P. Moulon, P. Gupta, R. Mur-Artal, R. Pennington, S. Kulkarni, S. Miglani, S. Gondi, S. Solanki, S. Diener, S. Cheng, S. Green, S. Saarinen, S. Patra, T. Mourikis, T. Whelan, T. Singh, V. Balntas, V. Baiyya, W. Dreewes, X. Pan, Y. Lou, Y. Zhao, Y. Mansour, Y. Zou, Z. Lv, Z. Wang, M. Yan, C. Ren, R. D. Nardi, and R. Newcombe. Project aria: A new tool for egocentric multi-modal ai research, 2023. URL https://arxiv.org/abs/2308.13561.

[27] M. Levy, S. Haldar, L. Pinto, and A. Shirivastava. P3-po: Prescriptive point priors for visuo-spatial generalization of robot policies, 2024. URL https://arxiv.org/abs/2412.06784.

[28] A. Mandlekar, D. Xu, J. Wong, S. Nasiriany, C. Wang, R. Kulkarni, L. Fei-Fei, S. Savarese, Y. Zhu, and R. Martín-Martín. What matters in learning from offline human demonstrations for robot manipulation. In *5th Annual Conference on Robot Learning*, 2021. URL https://openreview.net/forum?id=JrsfBJtDFdI.

[29] A. Mandlekar, J. Booher, M. Spero, A. Tung, A. Gupta, Y. Zhu, A. Garg, S. Savarese, and L. Fei-Fei. Scaling robot supervision to hundreds of hours with roboturk: Robotic manipulation dataset through human reasoning and dexterity. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1048–1055. IEEE, 2019.

[30] E. Jang, A. Irpan, M. Khansari, D. Kappler, F. Ebert, C. Lynch, S. Levine, and C. Finn. BC-Z: zero-shot task generalization with robotic imitation learning. *CoRR*, abs/2202.02005, 2022. URL https://arxiv.org/abs/2202.02005.

[31] B. D. Argall, S. Chernova, M. Veloso, and B. Browning. A survey of robot learning from demonstration. *Robotics and autonomous systems*, 57(5):469–483, 2009.

[32] A. Hussein, M. M. Gaber, E. Elyan, and C. Jayne. Imitation learning: A survey of learning methods. *ACM Computing Surveys (CSUR)*, 50(2):1–35, 2017.

[33] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn. Learning fine-grained bimanual manipulation with low-cost hardware. *arXiv preprint arXiv:2304.13705*, 2023.

[34] T. Z. Zhao, J. Tompson, D. Driess, P. Florence, K. Ghasemipour, C. Finn, and A. Wahid. Aloha unleashed: A simple recipe for robot dexterity, 2024. URL https://arxiv.org/abs/2410.13126.

[35] P. Wu, Y. Shentu, Z. Yi, X. Lin, and P. Abbeel. Gello: A general, low-cost, and intuitive teleoperation framework for robot manipulators, 2024. URL https://arxiv.org/abs/2309.13037.

[36] S.-R. Kang and K. Ikeuchi. Toward automatic robot instruction from perception—mapping human grasps to manipulator grasps. In *Proceedings of IEEE International Conference on Robotics and Automation*, pages 1932–1937. IEEE, 1994.

[37] K. Ikeuchi, K. Minamizawa, K. Harada, A. Yamaguchi, and S. Kagami. Semantic constraints to represent common sense required in household actions for multimodal learning-from-observation robot. *The International Journal of Robotics Research*, 43(4):399–414, 2024. doi:10.1177/02783649231212929.

[38] Meta quest. https://www.meta.com/quest/, 2024. [Virtual reality platform].

[39] Apple vision pro. https://www.apple.com/apple-vision-pro/, 2024. [Virtual reality platform].

[40] Steamvr. https://store.steampowered.com/app/250820/SteamVR/, 2024. [Virtual reality platform].

[41] P. Mannam, K. Shaw, D. Bauer, J. Oh, D. Pathak, and N. Pollard. Designing anthropomorphic soft hands through interaction. In *2023 IEEE-RAS 22nd International Conference on Humanoid Robots (Humanoids)*, pages 1–8, 2023. doi:10.1109/Humanoids57100.2023.10375195.

[42] Movella xsens. https://www.movella.com/products/xsens, 2024. [Motion capture system].

[43] Manusmetagloves. https://www.manus-meta.com, 2024. [Motion capture gloves].

[44] Rokoko. https://www.rokoko.com, 2024. [Motion capture solution].

[45] D. Shan, J. Geng, M. Shu, and D. Fouhey. Understanding human hands in contact at internet scale. In *CVPR*, 2020.

[46] K. Grauman, A. Westbury, E. Byrne, Z. Chavis, A. Furnari, R. Girdhar, J. Hamburger, H. Jiang, M. Liu, X. 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.

[47] D. Damen, H. Doughty, G. M. Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European conference on computer vision (ECCV)*, pages 720–736, 2018.

[48] R. Goyal, S. Ebrahimi Kahou, V. Michalski, J. Materzynska, S. Westphal, H. Kim, V. Haenel, I. Fruend, P. Yianilos, M. Mueller-Freitag, F. Hoppe, C. Thurau, I. Bax, and R. Memisevic. The "something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[49] Y.-W. Chao, W. Yang, Y. Xiang, P. Molchanov, A. Handa, J. Tremblay, Y. S. Narang, K. Van Wyk, U. Iqbal, S. Birchfield, et al. Dexycb: A benchmark for capturing hand grasping of objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9044–9053, 2021.

[50] Y. Qin, Y.-H. Wu, S. Liu, H. Jiang, R. Yang, Y. Fu, and X. Wang. Dexmv: Imitation learning for dexterous manipulation from human videos. In *European Conference on Computer Vision*, pages 570–587. Springer, 2022.

[51] G. Papagiannis, N. D. Palo, P. Vitiello, and E. Johns. R+x: Retrieval and execution from everyday human videos, 2024. URL https://arxiv.org/abs/2407.12957.

[52] S. Bahl, R. Mendonca, L. Chen, U. Jain, and D. Pathak. Affordances from human videos as a versatile representation for robotics. 2023.

[53] H. G. Singh, A. Loquercio, C. Sferrazza, J. Wu, H. Qi, P. Abbeel, and J. Malik. Hand-object interaction pretraining from videos, 2024. URL https://arxiv.org/abs/2409.08273.

[54] G. Pavlakos, D. Shan, I. Radosavovic, A. Kanazawa, D. Fouhey, and J. Malik. Reconstructing hands in 3d with transformers. *arXiv preprint arXiv:2312.05251*, 2023.

[55] X. Zhang, Q. Li, H. Mo, W. Zhang, and W. Zheng. End-to-end hand mesh recovery from a monocular rgb image. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2354–2364, 2019. doi:10.1109/ICCV.2019.00244.

[56] S. Baek, K. I. Kim, and T. Kim. Pushing the envelope for rgb-based dense 3d hand pose estimation via neural rendering. *CoRR*, abs/1904.04196, 2019. URL http://arxiv.org/abs/1904.04196.

[57] A. Boukhayma, R. A. de Bem, and P. H. S. Torr. 3d hand shape and pose from images in the wild. *CoRR*, abs/1902.03451, 2019. URL http://arxiv.org/abs/1902.03451.

[58] J. Romero, D. Tzionas, and M. J. Black. Embodied hands: Modeling and capturing hands and bodies together. *CoRR*, abs/2201.02610, 2022. URL https://arxiv.org/abs/2201.02610.

[59] N. Karaev, I. Makarov, J. Wang, N. Neverova, A. Vedaldi, and C. Rupprecht. Cotracker3: Simpler and better point tracking by pseudo-labelling real videos, 2024. URL https://arxiv.org/abs/2410.11831.

[60] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, Q. Jiang, C. Li, J. Yang, H. Su, J. Zhu, and L. Zhang. Grounding dino: Marrying dino with grounded pre-training for open-set object detection, 2024. URL https://arxiv.org/abs/2303.05499.

[61] L. Tang, M. Jia, Q. Wang, C. P. Phoo, and B. Hariharan. Emergent correspondence from image diffusion, 2023. URL https://arxiv.org/abs/2306.03881.

[62] S. Haldar, Z. Peng, and L. Pinto. Baku: An efficient transformer for multi-task policy learning, 2024. URL https://arxiv.org/abs/2406.07539.

[63] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. Foster, G. Lam, P. Sanketi, Q. Vuong, T. Kollar, B. Burchfiel, R. Tedrake, D. Sadigh, S. Levine, P. Liang, and C. Finn. Openvla: An open-source vision-language-action model, 2024. URL https://arxiv.org/abs/2406.09246.

[64] A. Bochkovskii, A. Delaunoy, H. Germain, M. Santos, Y. Zhou, S. R. Richter, and V. Koltun. Depth pro: Sharp monocular metric depth in less than a second. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.02073.

[65] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, E. Mintun, J. Pan, K. V. Alwala, N. Carion, C.-Y. Wu, R. Girshick, P. Dollár, and C. Feichtenhofer. Sam 2: Segment anything in images and videos, 2024. URL https://arxiv.org/abs/2408.00714.

A Human Demonstrations

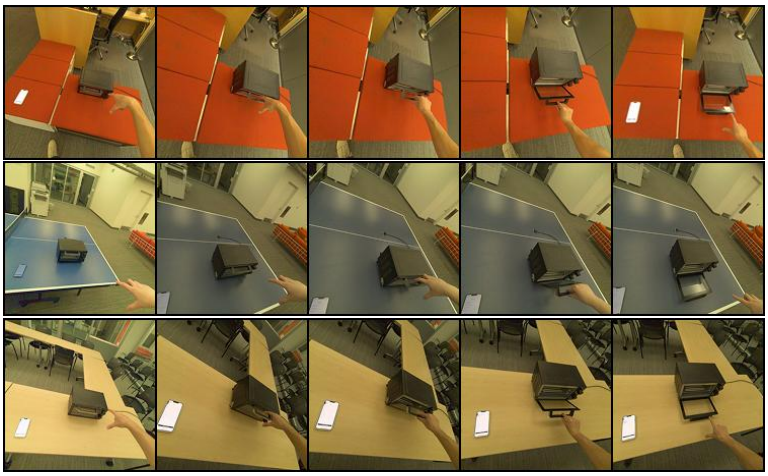


Figure 5: Open oven door.

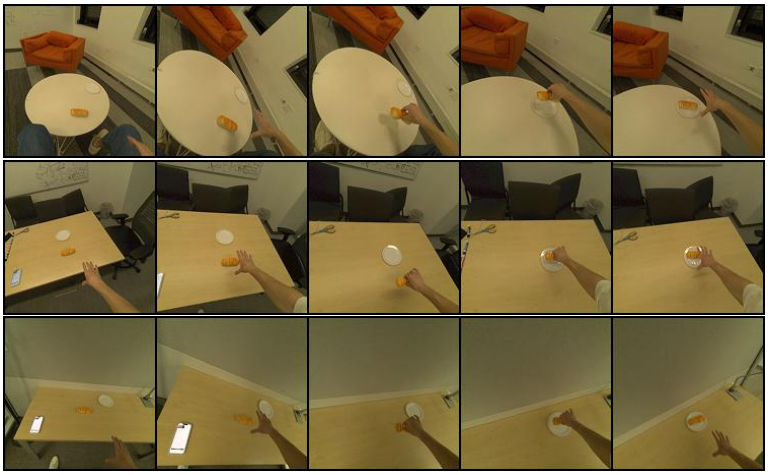


Figure 6: Put bread on plate.

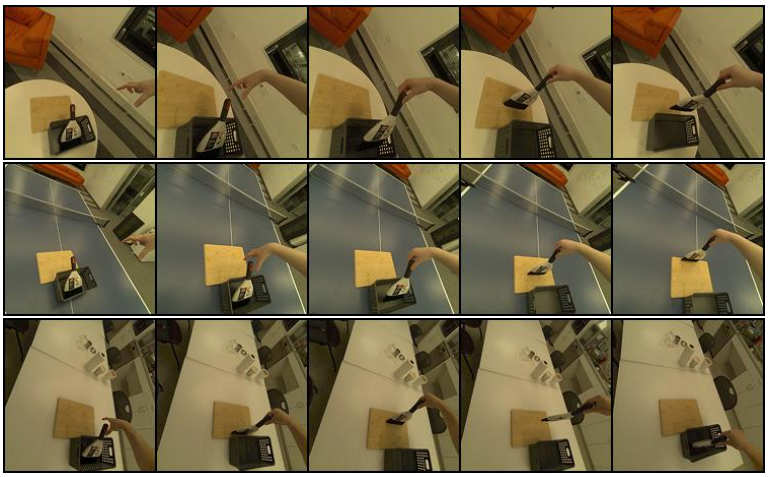


Figure 7: Sweep board with broom.

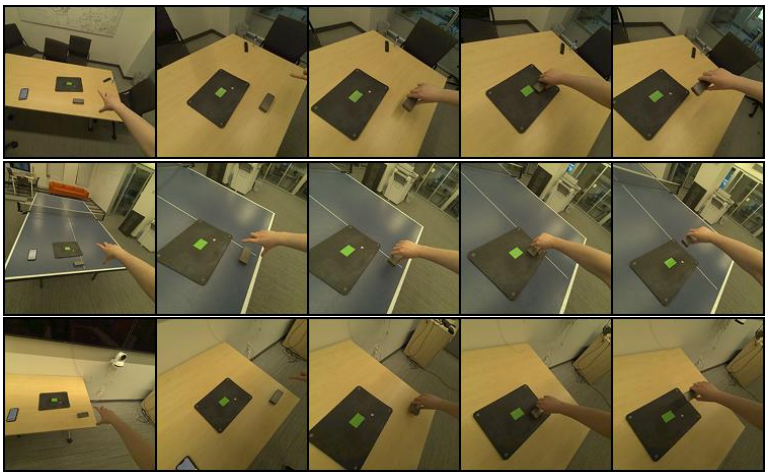


Figure 8: Erase board.

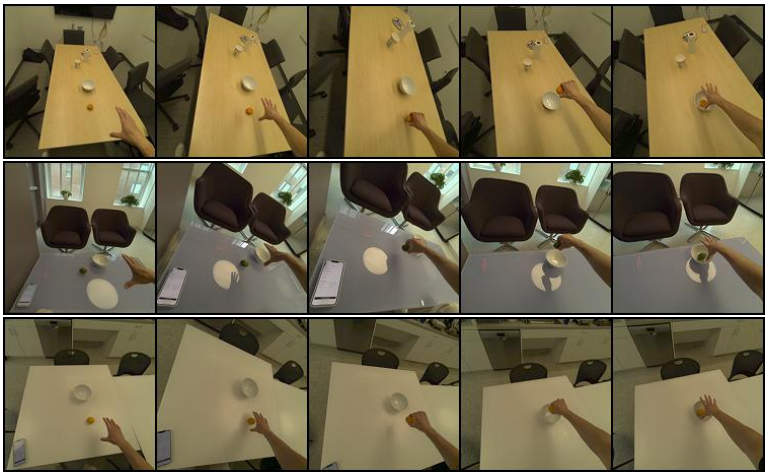


Figure 9: Sort fruit in bowl.

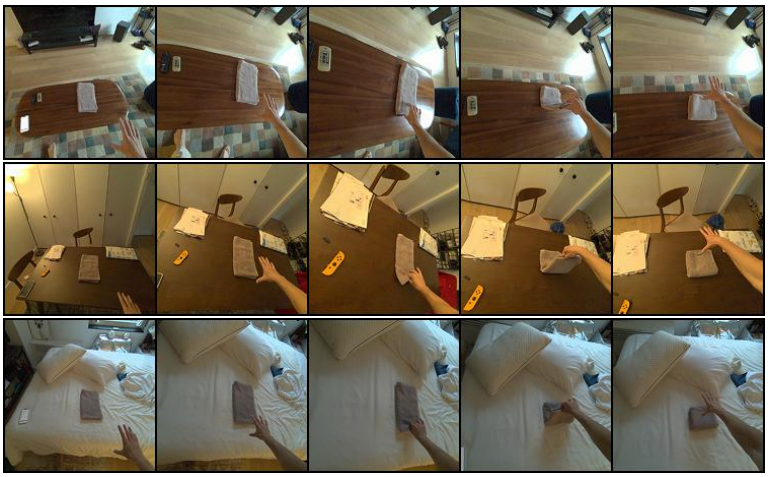


Figure 10: Fold towel.

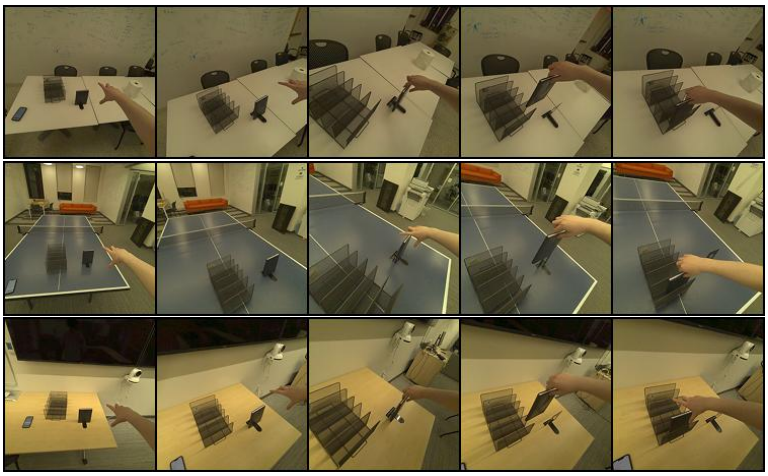


Figure 11: Insert book in shelf.

B Triangulating Object Keypoints

We estimate 3D coordinates $\mathbf{q}^* \in \mathbb{R}^3$ of an object point in the world frame at $t = 0$ from 2D observations $\{(T_i, u_i)\}_{i=1}^N$, where $u_i \in \mathbb{R}^2$ is the UV coordinate tracked in frame i, and $T_i \in SE(3)$ is the camera-to-world transformation at frame i. Let K denote the camera intrinsics and $P_i = K[R_i \mid \mathbf{t}_i] = KT_i^{-1}$ denote the projection matrix from world to image space at frame i.

1. Epipolar Filtering. To discard geometrically inconsistent views, we apply pairwise epipolar constraints. Given two frames i and j, we compute the fundamental matrix:

$$F_{ij} = K^{-T}[\mathbf{t}_{ij}]_\times R_{ij} K^{-1}, \tag{4}$$

where $R_{ij} = R_j R_i^\top$, $\mathbf{t}_{ij} = \mathbf{t}_j - R_{ij}\mathbf{t}_i$, and $[\cdot]_\times$ is the skew-symmetric matrix. A frame i is retained if it satisfies the epipolar constraint with at least m other frames:

$$\left|\mathbf{u}_j^\top F_{ij} \mathbf{u}_i\right| < \epsilon \quad \text{for at least } m \text{ views.} \tag{5}$$

2. Robust RANSAC Triangulation. Using the filtered inlier views, we perform RANSAC over subsets of size k to find the best triangulated candidate $\mathbf{q}^*$ minimizing reprojection error:

$$\mathbf{q}_{\text{RANSAC}} = \arg\min_{\mathbf{q}} \sum_{i \in \mathcal{I}} \mathbb{1}\left(\left\|u_i - \mathcal{P}(T_i^{-1}\mathbf{q})\right\|_2 < \tau\right). \tag{6}$$

3. Least Squares with Depth Bias. We refine $\mathbf{q}_{\text{RANSAC}}$ via nonlinear least squares with a Huber loss and a soft depth penalty:

$$\mathbf{q}^* = \arg\min_{\mathbf{q} \in \Omega} \sum_{i \in \mathcal{I}} \left\|u_i - \mathcal{P}(T_i^{-1}\mathbf{q})\right\|_\rho + \lambda \mathbf{q}_z, \tag{7}$$

where $||\cdot||_\rho$ is the Huber loss, q_z is the depth (z-coordinate in world frame), λ is the depth bias coefficient, and $\Omega = [\mathbf{l}, \mathbf{u}]$ is a bounding box constraint (i.e. $\mathbf{q}_z > 0$). This formulation encourages geometrically consistent triangulation while avoiding ambiguous far-away solutions in cases of degenerate motion or lag in Cotracker3 predictions.

4. Unified Object Representations. We repeat Steps 1-3 for each point that we label on the object, and concatenate each triangulated object point to obtain the object representation for the entire trajectory $\tilde{s}$.

C Policy Inference

Algorithm 1 EGOZERO Policy Inference

Obtain object keypoints on first frame using DIFT [61] on annotated dataset frame
Initialize $\tilde{s} = []$
for u in DIFT labels **do**
    Read depth at u from iPhone
    Unproject u with depth into egocentric frame to obtain x_u
    $\tilde{s} \leftarrow [\tilde{s}, x_u]$
end for
Initialize robot state $\tilde{a}_0$ and history buffer $H = [\tilde{a}_0, ..., \tilde{a}_0]$ of length h
for t in rollout **do**
    Compute action chunk $(\tilde{a}_t, ..., \tilde{a}_{t+\ell}) \sim \pi(\tilde{s}_t, H)$ and apply temporal aggregation to get $\tilde{a}_t$
    Parse gripper action $g \leftarrow \text{bool}(\tilde{a}_t > 0)$
    Execute $[\tilde{a}_t, g]$ on robot
    Update buffer $H \leftarrow [H, \tilde{a}_t]_{-h}$
end for

D Monocular Depth Estimation

We record a walkaround of 5 Aruco tags on the table from the Aria glasses and fit an affine scale/shift that minimizes the residual of the depth map at these Aruco tags. Even after calibration, we see that the depth signal deviates with variance from the ground truth Aruco detection, suggesting that monocular depth models are potentially spatio-temporally warped.

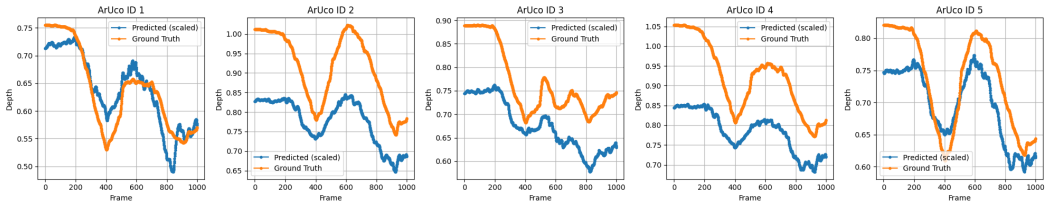


Figure 12: Monocular depth estimation [64] calibrated to Aruco tags in the scene.