EgoMimic: Scaling Imitation Learning via Egocentric Video

Simar Kareer[1], Dhruv Patel[1*], Ryan Punamiya[1*], Pranay Mathur[1*], Shuo Cheng[1]
Chen Wang[2], Judy Hoffman[1†], Danfei Xu[1†]

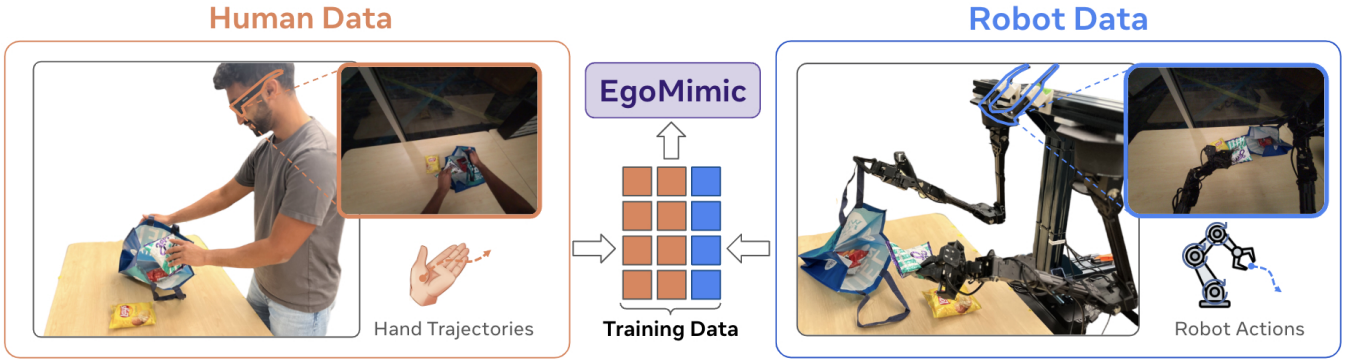

Fig. 1: EgoMimic enables anyone to collect human demonstrations for imitation learning, simply by wearing a pair of Project Aria glasses [1]. Aria glasses record egocentric vision paired with hand tracking, which we use to augment our robot training data. When combined, it can boost task performance by 34-228% and enable generalization to new objects or even scenes.

***Abstract*— The scale and diversity of demonstration data required for imitation learning is a significant challenge. We present EgoMimic, a full-stack framework that scales manipulation through egocentric-view human demonstrations. EgoMimic achieves this through: (1) an ergonomic human data collection system using the Project Aria glasses, (2) a low-cost bimanual manipulator that minimizes the kinematic gap to human data, (3) cross-domain data alignment techniques, and (4) an imitation learning architecture that co-trains on hand and robot data. Compared to prior works that only extract high-level intent from human videos, our approach treats human and robot data equally as embodied demonstration data and learns a unified policy from both data sources. EgoMimic achieves significant improvement on a diverse set of long-horizon, single-arm and bimanual manipulation tasks over state-of-the-art imitation learning methods and enables generalization to entirely new scenes. Finally, we show a favorable scaling trend for EgoMimic, where adding 1 hour of additional hand data is significantly more valuable than 1 hour of additional robot data. Videos and additional information can be found at https://egomimic.github.io/**

I. Introduction

End-to-end imitation learning has shown remarkable performance in learning complex manipulation tasks, but it remains brittle when facing new scenarios and tasks. Drawing on the recent success of Computer Vision and Natural Language Processing, we hypothesize that for learned policies to achieve broad generalization, we must dramatically scale up the training data size. While these adjacent domains benefit from Internet-sourced data, robotics lacks such an equivalent.

To scale up data for robotics, there have been recent advances in data collection systems. For example, ALOHA [2], [3] and GELLO [4] are intuitive leader-follower controls for collecting teleoperated data. Other works have opted to develop hand-held grippers to collect data without a robot [5]. Despite these advances, data collected via these systems still require specialized hardware and active effort in providing demonstrations. We hypothesize that a key step for achieving Internet-scale robot data is *passive data collection*. Just as the Internet was not built for curating data to train large vision and language models, an ideal robot data system should allow users to generate sensorimotor behavior data without intending to do so.

Human videos, especially those captured from an egocentric perspective, present an ideal source of data for passive data scalability. This data aligns closely with robot data, as it provides an egocentric camera for vision, 3D hand tracking for actions, and onboard SLAM for localization. The advent of consumer-grade devices capable of capturing such data, including Extended Reality (XR) devices and camera-equipped "smart glasses", opens up unprecedented opportunities for passive data collection at scale. While recent works have begun to leverage human video data, their approaches are limited to extracting high-level intent information from videos to build planners that guide low-level conditional policies [6], [7]. As a result, these systems remain constrained by the performance of low-level policies, which are typically trained solely on teleoperation data.

We argue that to truly scale robot performance with human data, we should not consider human videos as an auxiliary data source that requires separate handling. Instead, we should exploit the inherent similarities between egocentric human data and robot data to treat them as equal parts in a continuous spectrum of embodied data sources. Learning seamlessly from both data sources will require full-stack

SK, DP, RP, PM, SC, JH, DX are with the Georgia Institute of Technology and CW is with Stanford University. Email Correspondence: skareer@gatech.edu
*Denotes equal contribution.†Denotes equal advising.

innovation, from data collection systems that unify data from both sources to imitation learning architectures that can enable such cross-embodied policy learning.

To this end, our work treats human data as a *first-class data source* for robot manipulation. We believe our system is a key step towards using passive data from wearable smart glasses to train manipulation policies. We present EgoMimic (Fig. 1), a framework to collect data and co-train manipulation policies from both human egocentric videos and teleoperated robot data consisting of:

(i) A system to collect human data built on Project Aria glasses [1] that capture egocentric video, 3D hand tracking, and device SLAM. This rich information allows us to transform human egocentric data into a format compatible with robot imitation learning.

(ii) A capable yet low-cost bimanual robot that minimizes the kinematic and camera-to-camera gap to human data. In particular, we minimize the camera-to-camera device gap (FOV, dynamic ranges, etc) between human and robot data by using Project Aria glasses as the main robot sensor.

(iii) To mitigate differences in data distributions, we normalize and align action distributions between human and robots. Further, we minimize the appearance gap between human arm and robot manipulator via visual masking.

(iv) A unified imitation learning architecture that co-trains on hand and robot data with a common vision encoder and policy network. Despite distinct action spaces for human and robot, our model enforces a shared representation to enable performance scaling with human data, outperforming existing methods that treat hand and robot data separately.

We empirically evaluate EgoMimic on three challenging long-horizon manipulation tasks in the real world: continuous object-in-bowl, clothes folding, and grocery packing (Fig. 5). Our results demonstrate that EgoMimic significantly enhances task performance across all scenarios, with relative improvements of up to 200%. Notably, we observe that EgoMimic exhibits generalization to objects and scenes encountered exclusively in human data. Finally, we analyze the scaling properties of EgoMimic, and found learning from an additional hour of hand data significantly outperforms training from an additional hour of robot data.

II. Related Works

Imitation Learning: Imitation Learning (IL) has been used to perform diverse and contact-rich manipulation tasks [8], [9], [10]. Recent advancements in IL have led to the development of pixel-to-action IL models, which directly map raw visual inputs to low-level robot control [2], [11]. These visual IL models have demonstrated impressive reactive policies [12], [6]. Scaling these models has displayed strong generalization in works such as RT1 and RT2 [13], [14]. However, these methods remain labor and resource-intensive, for instance RT1 required 17 months of data collection and 13 robots [13]. Our work proposes a learning framework that takes advantage of scalable human demonstrations, which has the potential to be larger and more diverse than any dataset consisting of robot demonstrations alone.

Learning from Video Demonstrations: To satisfy the data requirements of pixel to action IL algorithms, many recent works leverage human data because it is highly scalable. Human data is used at different levels of abstraction, where some works use human videos from internet-scale datasets to pretrain visual representations [15], [16], [17]. Other works use human videos to more explicitly understand scene dynamics through point track prediction, intermediate state hallucination in pixel space, or affordance prediction [18], [19], [7], [20], [21]. And finally, recent works use hand trajectory prediction as a proxy for predicting robot actions [6]. While these approaches leverage hand data, they often have separate modules to process hand and robot data. Instead, by fully leveraging the rich information provided by Aria glasses including on-board SLAM, our method is able to unify and treat human and robot data as equals and co-train from both data sources with a single end-to-end policy.

Data Collection Systems: Various methods have been used to scale robot data. Low-cost devices such as the Space Mouse offer sensitive and fine-grained teleoperation of robotic manipulators [22], [10], [23], [11], [24]. Further works improve intuitive control through virtual reality systems such as the VR headset [25], [26], [27], [28], [29]. Recent systems like ALOHA and GELLO increase ergonomics for low-cost and fine-grained bimanual manipulation tasks through a leader-follower teleoperation interface [2], [4] or exoskeletons [30], [31]. Other works attempt to collect human demonstrations with rich information like 3D action tracking, but existing systems face tradeoffs. Those which leverage rich information are either not portable (e.g., static camera [32], [6], [33], [34]) or ergonomic (e.g., require a hand-held gripper [5], [35] or body-worn camera [36], [37]), which prevent the passive scalability of the data collection system. Along these lines, our approach captures egocentric video and 3D hand tracking data, but via the ergonomic form factor of Project Aria Glasses [1]. This system has the potential to passively scale [38], as adoption of similar consumer-grade devices continue to rise.

Cross-embodiment Policy Learning: Advances in cross-embodiment learning show that large models trained on datasets with diverse robot embodiments are more generalizable [39]. Some approaches aim to bridge the embodiment gap through observation reprojection [40], action abstractions [41], and policies conditioned on embodiment, [42]. Recent works view cross-embodiment learning as a domain adaptation problem [43]. Our work argues that human data should be treated as another embodiment in transfer learning.

III. EgoMimic

We aim to develop a unified framework that can simultaneously train on egocentric human and robot data. While many works have tackled aspects of this problem, we take a full stack approach. Concretely, we

- Develop a scalable pipeline to collect rich human data via Project Aria glasses
- Design a capable yet low cost bimanual robot system

- Process data to align visual and proprioception distributions between hand and robot
- Design a unified architecture that co-trains on both hand and robot data

A. Data Collection Systems and Hardware Design

Aria glasses for egocentric demonstration collection. An ideal system for human data needs to capture rich information about the scene, while remaining passively scalable. Such a system should be wearable, ergonomic, capture a wide FOV, track hand positions, device pose, and more.

EgoMimic fills this gap by building on top of the Project Aria glasses [1]. Aria glasses are head-worn devices for capturing multimodal egocentric data. The device assumes an ergonomic glasses form factor that weighs only 75g, permitting long wearing time and passive data collection. Our work leverages the front-facing wide-FoV RGB camera for visual observation and two mono-color scene cameras for device pose and hand tracking (See Fig. 2 for sample data). In particular, the side-facing scene cameras track hand poses even when they move out of the main RGB camera's view, significantly mitigating the challenges posed by humans' natural tendency to move their head and gaze ahead of their hands during sequential manipulation tasks.

Further, there are large scale data collection efforts underway with Project Aria [44], [45], and the devices are made available broadly to the academic community through an active research partnership program. In the future, our system can enable users to seamlessly merge data they collect with these large datasets. Ultimately, we present a system that enables passive yet feature-rich data collection to help scale up robot manipulation.

Low-cost bimanual manipulator. To effectively utilize egocentric human data, a robot manipulator should be capable of moving in ways that resemble human arm movements. Prior works often rely on table-mounted manipulators such as the Franka Emika Panda [46]. While these systems are capable, they differ significantly from human arms in terms of kinematics. Moreover, their substantial weight and inertia necessitate slow, cautious movements due to safety concerns, largely preventing them from performing manipulation tasks at speeds comparable to humans. In response to these limitations, we have purpose-built a bimanual manipulator that is lightweight, agile, and cost-effective. Drawing inspiration from the ALOHA system [2], our robot setup comprises two 6-DoF ViperX 300 S arms with Intel Realsense D405 wrist cameras, mounted in an inverted configuration on a height-adjustable rig as the torso (Fig 2), kinematically mimicking the upper body of a human. The ViperX arms are lean and relatively similar in size to human arms, contributing to their enhanced agility. The entire rig can be assembled for less than $1,000 excluding the ViperX arms (the BOM will be made available). We also built a leader robot rig to collect teleoperation data, similar to ALOHA [2].

Further, as our method jointly learns visual policies from human egocentric and robot data, it is essential to align the visual observation space. Thus in addition to alignment through data post-processing (Sec. III-B), we directly match the camera hardware by using a second pair of Aria glasses as the main sensor for the robot, which we have mounted directly to the top of the torso at a location similar to that of human eyes (Fig 2). This enables us to mitigate the observation domain gap associated with the camera devices, including FOVs, exposure levels, and dynamic ranges.

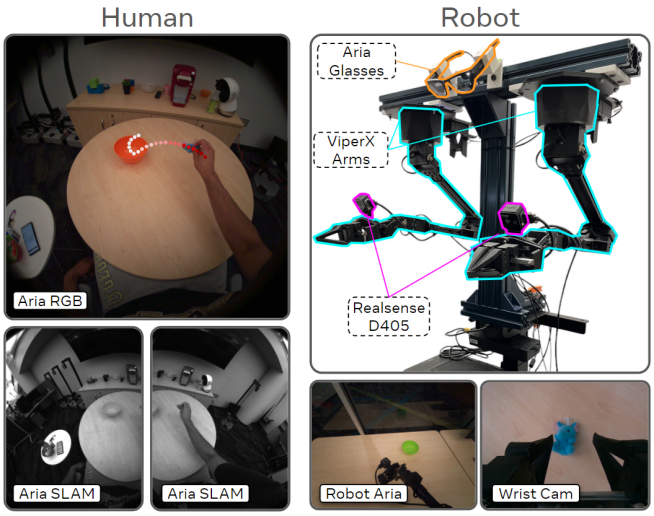

Fig. 2: Our system uses Aria glasses to capture Egocentric RGB and uses its side SLAM cameras to localize the device and track hands. The robot consists of two Viper X arms with Intel RealSense D405 wrist cameras. Our robot uses an idential Aria glasses as the main vision sensor to help minimize the camera to camera gap.

B. Data Processing and Domain Alignment

To train unified policies from both human and robot data, EgoMimic bridges three key human-robot gaps: (1) unifying action coordinate frames, (2) aligning action distributions, and (3) mitigating visual appearance gaps.

Raw data streams. We stream raw sensor data from the hardware setup as described in Sec. III-A. Aria glasses worn by the human and robot generate ego-centric RGB image streams. In addition, the robot generates two wrist camera streams. For proprioception, we leverage the Aria Machine Perception Service (MPS) [47] to estimate 3D poses of both hands ${}^{H}p \in \mathbb{SE}(3) \times \mathbb{SE}(3)$. Robot proprioception data includes both its end effector poses ${}^{R}p \in \mathbb{SE}(3) \times \mathbb{SE}(3)$ and joint positions ${}^{R}q \in \mathbb{R}^{2\times 7}$ (including the gripper jaw joint position). We in addition collect joint-space actions ${}^{R}a^{q} \in \mathbb{R}^{2\times 7}$ for teleoperated robot data.

Unifying human-robot data coordinate frames. Robot action and proprioception data typically use fixed reference frames (e.g., camera or robot base frame). However, egocentric hand data from moving cameras breaks this assumption. To unify the reference frames for joint policy learning, we transform both human hand and robot end effector trajectories into camera-centered stable reference frames. Following the idea of predicting action chunks [11], [2], we aim to construct action chunks $a^{p}_{t:t+h}$ for both human hand and robot end effector. To simplify the notation, we describe the single-arm case that generalizes to both arms. The raw trajectory is a sequence of 3D poses $[p_t^{F_t}, p_{t+1}^{Ft+1}, ...p_{t+h}^{Ft+h}]$,

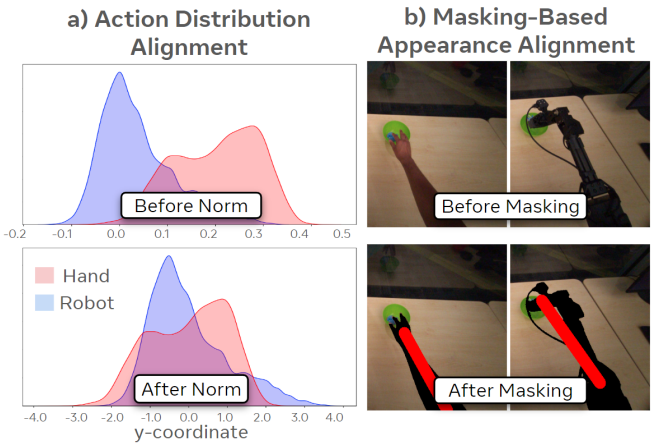

Fig. 3: **a) Action normalization:** The pose distributions are different between hand and robot data, specifically in the y (left-right) dimension. We apply Gaussian normalization individually to the hand and robot pose data before feeding them to the model. **b) Visual masking:** To help bridge the appearance gap of human and and the robot arm, we apply a black mask to the hand and robot via SAM, then overlay a red line onto the image.

where F_i denotes the coordinate frame of the camera when estimating p_i. F_i remains fixed for the robot but changes constantly for human egocentric data. Our goal is to construct $a^p_{t:t+h}$ by transforming each position in the trajectory into the observation camera frame F_t. This allows the policy to predict actions without considering future camera movements. For human data, we use the MPS visual-inertial SLAM to obtain the Aria glasses pose $T^W_{F_i} \in \mathbb{SE}(3)$ in the world frame and transform the action trajectory:

$$ {}^{H}a^p_i = [(T^W_{F_t})^{-1} T^W_{F_i} p^{F_i}_i \quad \text{for} \quad i \in [t, t+1, ..., t+h]] $$

A sample trajectory is visualized in Fig. 2 (top-left). Robot data is transformed similarly using the fixed camera frame estimated by hand-eye calibration. By creating a unified reference frame, we enable the policy to learn from action supervisions regardless of whether they originate from human videos or teleoperated demonstrations.

Aligning human-robot pose distributions. Despite aligning hand and robot data via hardware design and data processing, we still observe differences in the distributions of hand and robot end effector poses in the demonstrations collected. These discrepancies arise from biomechanical differences, task execution variations, and measurement precision disparities between human and robotic systems. Without mitigating this gap, the policy tends to learn separate representations for the two data sources [48], [49], preventing performance scaling with human data. To address this, we apply Gaussian normalization individually to end effector (hand) poses and actions from each data source, as shown in Fig. 3. Echoing [49], we found this simple technique to be empirically effective (Sec. IV-B), though we plan to explore alternatives such as action quantization [13] in the future.

Bridging visual appearance gaps. Despite aligning sensor hardware for capturing robot and human data, there still exists a large visual appearance gap between human hands and robots. Previous works have acknowledged this gap and attempt to occlude or remove the manipulator in visual observation [50], [51]. We follow similar ideas and mask out both the hand and the robot via SAM [52] and overlay a red line to indicate end-effector directions (Fig 3). The SAM point prompts are generated by the robot end effector and human hand poses transformed to image frames.

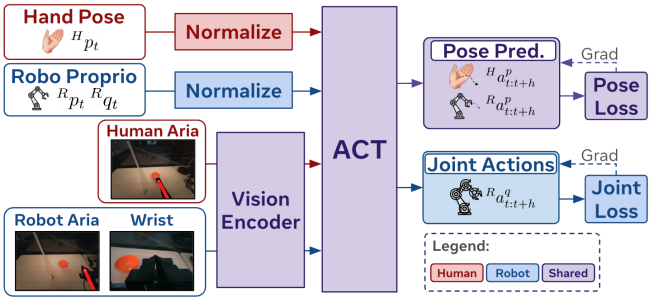

Fig. 4: Architecture of the joint human-robot policy learning framework. The model processes normalized hand and robot data through shared vision and ACT encoders, outputting pose predictions for both human and robot data, and joint actions for robot data. The framework uses masked images to mitigate human-robot appearance gaps and incorporates wrist camera views for the robot.

C. Training Human-Robot Joint Policies

Existing approaches often opt for hierarchical architectures, where a high-level policy trained on human data conditions a low-level policy outputting robot actions [6], [7]. However, this approach is inherently limited by the performance of the low-level policy, which does not directly benefit from large-scale human data. To address this limitation, we propose a simple architecture (illustrated in Fig. 4) that learns from unified data and promotes shared representation. Our model builds upon ACT [2], but the design is general and can be applied to other transformer based imitation learning algorithms.

A critical challenge in this unified approach is the choice of the robot action space. While the robot end-effector poses are more semantically similar to human hand pose than robot-joint positions, it is difficult to control our robot with end-effector poses via a cartesian-based controller (e.g., differential IK) because the 6 DoF ViperX arms offer low solution redundancy. Empirically, we found that robots often encounter singularities or non-smooth solutions in a trajectory. Consequently, we opt for joint-space control, but leverage pose space prediction to learn joint human-robot representation.

Specifically, all parameters in the policy are shared besides the two shallow input and output heads. The input heads transform the visual and proprioceptive embeddings before passing to the policy transformer. The policy transformer processes these features, and the two output heads transform the transformer's latent output into either pose or joint space predictions. The pose loss supervises both human and robot data via ${}^{H}a^p$ and ${}^{R}a^p$, whereas the joint action loss only supervises robot data ${}^{R}a^q$. Since the two branches are separated by only one linear layer, we effectively force the model to learn joint representations for both domains. The

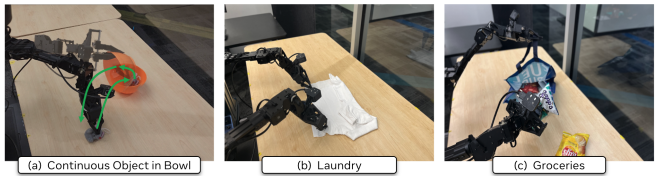

Fig. 5: We evaluate EgoMimic across three real world, long-horizon manipulation tasks. See Sec. IV-A for description.

algorithm is summarized in Alg. 1.

Algorithm 1 Joint Human-Robot Policy Learning

Require: Human dataset $\mathcal{D}_H$, Robot dataset $\mathcal{D}_R$

Initialize shared transformer encoder $f_{enc}(\cdot)$, pose decoder $f^p(f_{\text{enc}}(\cdot))$, and joint decoder $f^q(f_{\text{enc}}(\cdot))$
for iteration $n = 1, 2, \ldots$ **do**
  Sample $(I_t, p_t, a^p_{t:t+h})$ from $\mathcal{D}_H$
  Predict $\hat{a}^p_{t:t+h}$ from $f_p(f_{\text{enc}}(I_t, p_t))$
  $\mathcal{L}_p = \text{MSE}(\hat{a}^p_{t:t+h}, a^p_{t:t+h})$
  Sample $(I_t, p_t, q_t, a^p_{t:t+h}, a^q_{t:t+h})$ from $\mathcal{D}_R$
  Predict $\hat{a}^q_{t:t+h}$ from $f_q(f_{\text{enc}}(I_t, p_t, q_t))$
  Predict $\hat{a}^p_{t:t+h}$ from $f_p(f_{\text{enc}}(I_t, p_t, q_t))$
  $\mathcal{L}_q = \text{MSE}(\hat{a}^q_{t:t+h}, a^q_{t:t+h})$
  $\mathcal{L}_p = \mathcal{L}_p + \text{MSE}(\hat{a}^p_{t:t+h}, a^p_{t:t+h})$
  Update f_{enc}, f^p, f^q with $\mathcal{L}_p + \mathcal{L}_q$
end for

IV. Experiments

We aim to validate three key hypotheses. **H1**: EgoMimic is able to leverage human data to boost in-domain performance for complex manipulation tasks. **H2**: Human data helps EgoMimic generalize to new objects and scenes. **H3**: Given sufficient initial robot data, it is more valuable to collect additional human data than additional robot data.

A. Experiment Setup

Tasks. We select a set of long-horizon real world tasks to evaluate our claims. Our tasks require precise alignment, complex motions, and bimanual coordination (Fig. 5).

Continuous Object-in-Bowl: The robot picks a small plush toy (about 6cm long), places it in a bowl, picks up the bowl to dump the object onto the table, and repeats continuously for 40 seconds. We randomly choose from a set of 3 bowls and 5 toys which randomly positioned on the table within a 45cm x 60cm range. The task stress-tests precise manipulation, spatial generalization, and robustness in long-horizon execution. We award **Pts** each time the toy is placed in a bowl, or the bowl is emptied. We perform 45 total evaluation rollouts across 9 bowl-toy-position combinations.

Laundry: A bimanual task that requires the robot to fold a t-shirt placed with random pose in a 90cm $\times$ 60cm range and a rotation range of $\pm 30\,\text{deg}$. The robot must use both arms to fold the right side sleeve, the left side sleeve, then the whole shirt in half. We award **Pts** for each of these stages, and calculate Success Rate **(SR)** based as the percentage of runs where all stages were successful. We perform 40 total evaluation rollouts across 8 shirt-position combinations.

Groceries: The robot fills a grocery bag with 3 packs of chips. It uses its left arm to grab the top side of the bag handle to create an opening, then uses the right arm to pick the chip packs and places them into the bag. The task requires high-precision manipulation (picking up a deformable bag handle) and robustness in long-horizon rollout. We award **Pts** for picking the handle and for each pack placed in the grocery bag. We report **SR** as the percentage of runs where all three packs were successfully placed in the bag, and **Open Bag** as the percentage of runs where the handle of the bag was grasped, which is a difficult stage of this task. We perform 50 evaluations across 10 bag positions.

We detail the amount of data collected for each task in Table I. While collecting robot data in particular, we make sure to randomly perturb the robot's position, which we found empirically to improve robustness. For human data, we note that while tasks like *Continuous Object-in-Bowl* were particularly easy to scale, tasks like *Groceries* were slower because of resetting time.

Baselines. To evaluate that EgoMimic can improve in-domain success rate by leveraging human data, we benchmark against ACT [2], a state of the art imitation learning algorithm. Further, we compare against Mimicplay [6], a recent state of the art method that learns planners from human data to guide low-level policies, to show that our unified architecture learns more effectively from human and robot data. For fair comparisons, we implement Mimicplay with the same Transformer backbone as our method, and we removed goal conditioning because EgoMimic is designed for single-task policies. Since EgoMimic contains architectural changes to ACT, namely the simultaneous joint and pose action prediction, we also benchmark against EgoMimic (0% Human). This helps us conclude that improvements come from leveraging human data rather than the architecture.

TABLE I: Data collection overview for both Human(H) and Robot(R) data. We report both the number(#) of total task demonstrations and the time(min) took to collect them.

Task	H #	H min	H #/min	R #	R min	R #/min
Object-in-Bowl	1400	60	23	270	120	2
Groceries	160	80	2	300	300	1
Laundry	590	100	6	430	300	1

TABLE II: Quantitative results for 3 real-world tasks. We report task success rates (%) and performance scores (pts) for all tasks and bag grabbing rate for the Groceries tasks.

Method	Bowl	Laundry		Groceries		
	Pts	Pts	SR	Pts	SR	Open Bag
ACT [2]	39	82	55%	82	22%	54%
Mimicplay [6]	71	78	50%	53	8%	40%
EgoMimic (w/o human)	68	104	73%	92	28%	60%
EgoMimic	**128**	**114**	**88%**	**110**	**30%**	**70%**

TABLE III: **Ablations** - We ablate our method and report final task performance on the Object-in-Bowl task.

Method	Cotrained (Points)
EgoMimic	128
EgoMimic w/o Line	112
EgoMimic w/o Line and Mask	95
EgoMimic w/o Action Norm	79
EgoMimic w/o Hand Data	68

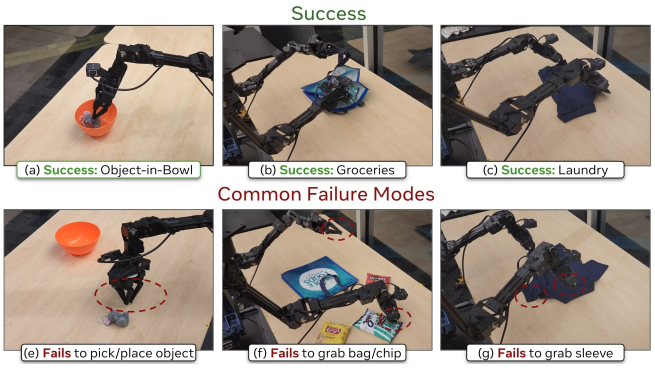

Fig. 6: We highlight EgoMimic's success, as well as failure modes, for instance **(e)** failure to correctly align with the toy, **(f)** failure to grasp the bag's handle, or **(g)** policy only grabs 1 side of the shirt. **EgoMimic** reduces the frequency of these failure modes, improving success rates by 8-33% over the baselines.

B. Results

EgoMimic improves in-domain task performance. Across all tasks we observed a relative improvement in score of 34-228%, and an improvement in absolute task success rate from 8-33% over ACT. Our largest improvement is on the *Cont. Object-in-Bowl* task, in which we yield a 228% improvement in task score over ACT. We observe the baselines often miss the toy or bowl by a few inches, which seems to indicate that our use of hand data helps the policy precisely reach the toy. We show qualitative results in Fig. 6.

To ensure this increase was due to leveraging hand data rather than architectural changes, we compare to EgoMimic (0% human). We observe a 10-88% improvement in score and 2-15% improvement in success rate.

EgoMimic enables generalization to new objects and even scenes. We evaluate our method on two domain shifts: attempting to fold shirts of an unseen color, and performing the *Cont. Object-in-Bowl* task in an entirely different scene. As shown in Fig. 7, we observe that ACT struggles on shirts of unseen colors (25% SR) whereas EgoMimic fully retains its performance (85% SR). Surprisingly, by learning from human data in a new scene (unseen background and lighting), EgoMimic is able to generalize to this new environment without *any* additional robot data, scoring 63 points. In contrast, Mimicplay, which had access to the same information but instead leverages a hierarchical framework for using hand data only scored 4 points. This suggests that our architecture promotes joint hand-robot representation, whereas hierarchical architectures pose a generalization bottleneck.

Scaling human vs. robot data. To investigate the scaling effect of human and robot data sources on performance, we

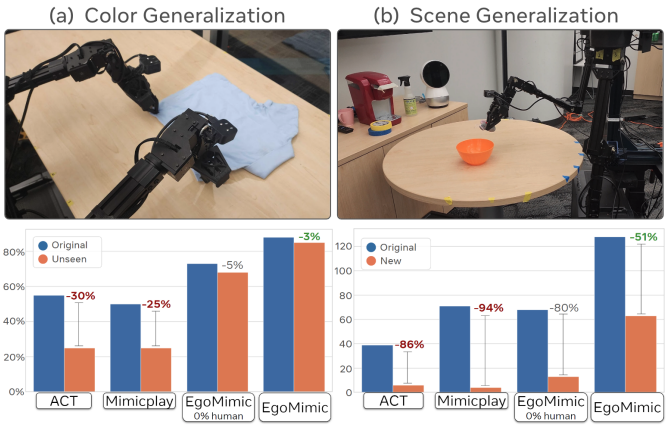

Fig. 7: **Evaluation Results on Policy Generalization.** (a) We evaluate the policy on the laundry task using unseen cloth colors and report the success rate for each method. (b) We test the policy on the Object-in-Bowl task in unseen scenes.

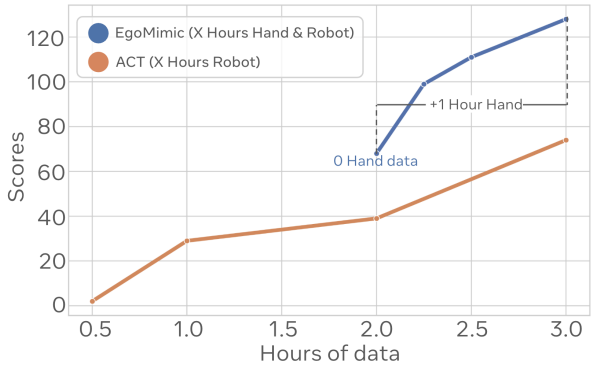

Fig. 8: **Scaling robot vs. human data.** EgoMimic trained on 2 hours robot data + 1 hour hand data (Blue) strongly outperforms ACT [2] trained on 3 hours of robot data (Orange).

conducted additional data collection for the *Cont. Object-in-bowl* task. As illustrated in Fig. 7, EgoMimic trained on 2 hours of robot data and 1 hour of human data significantly outperforms ACT trained on 3 hours of robot data (128 vs 74 points). Notably, one hour of human data yields 1400 demonstrations, compared to only 135 demonstrations from an hour of robot data. These results demonstrate EgoMimic's ability to effectively leverage the efficiency of human data collection, leading to a more pronounced scaling effect that substantially boosts task performance beyond what is achievable with robot data alone. We note that EgoMimic at 2 hours of robot data outperforms ACT at 2 hours of robot data, so some improvement is attributed to architecture.

Ablation studies. We ablate our approach to demonstrate the importance of each design decision on the *Object-in-Bowl* task (Table III). First, removing action normalization results in a 38% drop in task score. This highlights the importance of action distribution alignment for co-training. Next, we ablate away the visual techniques, specifically masking out the hand and robot, as well as drawing the red overlay on the image. Removing these components resulted in 13 and 26% drops respectively. Finally, EgoMimic trained without any hand data, yields a large 47% drop, which highlights

how effective hand-robot co-training is on our stack.

V. Conclusions

We presented EgoMimic, a framework to co-train manipulation policies from human egocentric videos and teleoperated robot data. By leveraging Project Aria glasses, a low-cost bimanual robot setup, cross-domain alignment techniques, and a unified policy learning architecture, EgoMimic improves over state-of-the-art baselines on three challenging real-world tasks and shows generalization to new scenes as well as favorable scaling properties. For future work, we plan to explore the possibility of generalizing to new robot embodiments and entirely new behaviors demonstrated only in human data, such as folding pants instead of shirts. Overall, we believe our work opens up exciting new venues of research on scaling robot data via passive data collection.

References

[1] 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. [Online]. Available: https://arxiv.org/abs/2308.13561

[2] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn, "Learning fine-grained bimanual manipulation with low-cost hardware," 2023. [Online]. Available: https://arxiv.org/abs/2304.13705

[3] A. . Team, J. Aldaco, T. Armstrong, R. Baruch, J. Bingham, S. Chan, K. Draper, D. Dwibedi, C. Finn, P. Florence, S. Goodrich, W. Gramlich, T. Hage, A. Herzog, J. Hoech, T. Nguyen, I. Storz, B. Tabanpour, L. Takayama, J. Tompson, A. Wahid, T. Wahrburg, S. Xu, S. Yaroshenko, K. Zakka, and T. Z. Zhao, "Aloha 2: An enhanced low-cost hardware for bimanual teleoperation," 2024. [Online]. Available: https://arxiv.org/abs/2405.02292

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

[5] C. Chi, Z. Xu, C. Pan, E. Cousineau, B. Burchfiel, S. Feng, R. Tedrake, and S. Song, "Universal manipulation interface: In-the-wild robot teaching without in-the-wild robots," 2024. [Online]. Available: https://arxiv.org/abs/2402.10329

[6] C. Wang, L. Fan, J. Sun, R. Zhang, L. Fei-Fei, D. Xu, Y. Zhu, and A. Anandkumar, "Mimicplay: Long-horizon imitation learning by watching human play," 2023. [Online]. Available: https://arxiv.org/abs/2302.12422

[7] H. Bharadhwaj, A. Gupta, V. Kumar, and S. Tulsiani, "Towards generalizable zero-shot manipulation via translating human interaction plans," 2023. [Online]. Available: https://arxiv.org/abs/2312.00775

[8] A. Paraschos, C. Daniel, J. R. Peters, and G. Neumann, "Probabilistic movement primitives," in *Advances in Neural Information Processing Systems*, C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, Eds., vol. 26. Curran Associates, Inc., 2013.

[9] C. Finn, T. Yu, T. Zhang, P. Abbeel, and S. Levine, "One-shot visual imitation learning via meta-learning," 2017. [Online]. Available: https://arxiv.org/abs/1709.04905

[10] 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 *arXiv preprint arXiv:2108.03298*, 2021.

[11] C. Chi, Z. Xu, S. Feng, E. Cousineau, Y. Du, B. Burchfiel, R. Tedrake, and S. Song, "Diffusion policy: Visuomotor policy learning via action diffusion," 2024. [Online]. Available: https://arxiv.org/abs/2303.04137

[12] S. Young, D. Gandhi, S. Tulsiani, A. Gupta, P. Abbeel, and L. Pinto, "Visual imitation made easy," 2020. [Online]. Available: https://arxiv.org/abs/2008.04899

[13] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, J. Hsu, J. Ibarz, B. Ichter, A. Irpan, T. Jackson, S. Jesmonth, N. J. Joshi, R. Julian, D. Kalashnikov, Y. Kuang, I. Leal, K.-H. Lee, S. Levine, Y. Lu, U. Malla, D. Manjunath, I. Mordatch, O. Nachum, C. Parada, J. Peralta, E. Perez, K. Pertsch, J. Quiambao, K. Rao, M. Ryoo, G. Salazar, P. Sanketi, K. Sayed, J. Singh, S. Sontakke, A. Stone, C. Tan, H. Tran, V. Vanhoucke, S. Vega, Q. Vuong, F. Xia, T. Xiao, P. Xu, S. Xu, T. Yu, and B. Zitkovich, "Rt-1: Robotics transformer for real-world control at scale," 2023. [Online]. Available: https://arxiv.org/abs/2212.06817

[14] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, X. Chen, K. Choromanski, T. Ding, D. Driess, A. Dubey, C. Finn, P. Florence, C. Fu, M. G. Arenas, K. Gopalakrishnan, K. Han, K. Hausman, A. Herzog, J. Hsu, B. Ichter, A. Irpan, N. Joshi, R. Julian, D. Kalashnikov, Y. Kuang, I. Leal, L. Lee, T.-W. E. Lee, S. Levine, Y. Lu, H. Michalewski, I. Mordatch, K. Pertsch, K. Rao, K. Reymann, M. Ryoo, G. Salazar, P. Sanketi, P. Sermanet, J. Singh, A. Singh, R. Soricut, H. Tran, V. Vanhoucke, Q. Vuong, A. Wahid, S. Welker, P. Wohlhart, J. Wu, F. Xia, T. Xiao, P. Xu, S. Xu, T. Yu, and B. Zitkovich, "Rt-2: Vision-language-action models transfer web knowledge to robotic control," 2023. [Online]. Available: https://arxiv.org/abs/2307.15818

[15] S. Nair, A. Rajeswaran, V. Kumar, C. Finn, and A. Gupta, "R3m: A universal visual representation for robot manipulation," 2022. [Online]. Available: https://arxiv.org/abs/2203.12601

[16] I. Radosavovic, T. Xiao, S. James, P. Abbeel, J. Malik, and T. Darrell, "Real-world robot learning with masked visual pre-training," in *Conference on Robot Learning*. PMLR, 2023, pp. 416–426.

[17] Y. J. Ma, S. Sodhani, D. Jayaraman, O. Bastani, V. Kumar, and A. Zhang, "Vip: Towards universal visual reward and representation via value-implicit pre-training," *arXiv preprint arXiv:2210.00030*, 2022.

[18] H. Xiong, Q. Li, Y.-C. Chen, H. Bharadhwaj, S. Sinha, and A. Garg, "Learning by watching: Physical imitation of manipulation skills from human videos," 2021. [Online]. Available: https://arxiv.org/abs/2101.07241

[19] C. Wen, X. Lin, J. So, K. Chen, Q. Dou, Y. Gao, and P. Abbeel, "Any-point trajectory modeling for policy learning," 2024. [Online]. Available: https://arxiv.org/abs/2401.00025

[20] H. Bharadhwaj, R. Mottaghi, A. Gupta, and S. Tulsiani, "Track2act: Predicting point tracks from internet videos enables generalizable robot manipulation," 2024. [Online]. Available: https://arxiv.org/abs/2405.01527

[21] S. Bahl, R. Mendonca, L. Chen, U. Jain, and D. Pathak, "Affordances from human videos as a versatile representation for robotics," 2023. [Online]. Available: https://arxiv.org/abs/2304.08488

[22] A. Mandlekar, D. Xu, R. Martín-Martín, S. Savarese, and L. Fei-Fei, "Learning to generalize across long-horizon tasks from human demonstrations," *arXiv preprint arXiv:2003.06085*, 2020.

[23] V. Dhat, N. Walker, and M. Cakmak, "Using 3d mice to control robot manipulators," *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:267322988

[24] Y. Zhu, A. Joshi, P. Stone, and Y. Zhu, "Viola: Imitation learning for vision-based manipulation with object proposal priors," 2023. [Online]. Available: https://arxiv.org/abs/2210.11339

[25] S. P. Arunachalam, I. Güzey, S. Chintala, and L. Pinto, "Holo-dex: Teaching dexterity with immersive mixed reality," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 5962–5969.

[26] A. George, A. Bartsch, and A. B. Farimani, "Openvr: Teleoperation for manipulation," 2023. [Online]. Available: https://arxiv.org/abs/2305.09765

[27] I. A. Tsokalo, D. Kuss, I. Kharabet, F. H. P. Fitzek, and M. Reisslein, "Remote robot control with human-in-the-loop over long distances using digital twins," in *2019 IEEE Global Communications Conference (GLOBECOM)*, 2019, pp. 1–6.

[28] X. Cheng, J. Li, S. Yang, G. Yang, and X. Wang, "Open-television: teleoperation with immersive active visual feedback," *arXiv preprint arXiv:2407.01512*, 2024.

[29] T. He, Z. Luo, X. He, W. Xiao, C. Zhang, W. Zhang, K. Kitani, C. Liu, and G. Shi, "Omnih2o: Universal and dexterous human-to-humanoid whole-body teleoperation and learning," *arXiv preprint arXiv:2406.08858*, 2024.

[30] H. Fang, H.-S. Fang, Y. Wang, J. Ren, J. Chen, R. Zhang, W. Wang, and C. Lu, "Airexo: Low-cost exoskeletons for learning whole-arm manipulation in the wild," *arXiv preprint arXiv:2309.14975*, 2023.

[31] S. Yang, M. Liu, Y. Qin, R. Ding, J. Li, X. Cheng, R. Yang, S. Yi, and X. Wang, "Ace: A cross-platform visual-exoskeletons system for low-cost dexterous teleoperation," *arXiv preprint arXiv:2408.11805*, 2024.

[32] A. Sivakumar, K. Shaw, and D. Pathak, "Robotic telekinesis: Learning a robotic hand imitator by watching humans on youtube," *arXiv preprint arXiv:2202.10448*, 2022.

[33] V. Jain, M. Attarian, N. J. Joshi, A. Wahid, D. Driess, Q. Vuong, P. R. Sanketi, P. Sermanet, S. Welker, C. Chan, *et al.*, "Vid2robot: End-to-end video-conditioned policy learning with cross-attention transformers," *arXiv preprint arXiv:2403.12943*, 2024.

[34] Z. Fu, Q. Zhao, Q. Wu, G. Wetzstein, and C. Finn, "Humanplus: Humanoid shadowing and imitation from humans," in *Conference on Robot Learning (CoRL)*, 2024.

[35] N. M. M. Shafiullah, A. Rai, H. Etukuru, Y. Liu, I. Misra, S. Chintala, and L. Pinto, "On bringing robots home," *arXiv preprint arXiv:2311.16098*, 2023.

[36] 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. [Online]. Available: https://arxiv.org/abs/2403.07788

[37] G. Papagiannis, N. Di Palo, P. Vitiello, and E. Johns, "R+ x: Retrieval and execution from everyday human videos," *arXiv preprint arXiv:2407.12957*, 2024.

[38] K. Grauman, A. Westbury, L. Torresani, K. Kitani, J. Malik, T. Afouras, K. Ashutosh, V. Baiyya, S. Bansal, B. Boote, *et al.*, "Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 19 383–19 400.

[39] E. Collaboration, A. O'Neill, A. Rehman, A. Gupta, A. Maddukuri, A. Gupta, A. Padalkar, A. Lee, A. Pooley, A. Gupta, A. Mandlekar, A. Jain, A. Tung, A. Bewley, A. Herzog, A. Irpan, A. Khazatsky, A. Rai, A. Gupta, A. Wang, A. Kolobov, A. Singh, A. Garg, A. Kembhavi, A. Xie, A. Brohan, A. Raffin, A. Sharma, A. Yavary, A. Jain, A. Balakrishna, A. Wahid, B. Burgess-Limerick, B. Kim, B. Schölkopf, B. Wulfe, B. Ichter, C. Lu, C. Xu, C. Le, C. Finn, C. Wang, C. Xu, C. Chi, C. Huang, C. Chan, C. Agia, C. Pan, C. Fu, C. Devin, D. Xu, D. Morton, D. Driess, D. Chen, D. Pathak, D. Shah, D. Büchler, D. Jayaraman, D. Kalashnikov, D. Sadigh, E. Johns, E. Foster, F. Liu, F. Ceola, F. Xia, F. Zhao, F. V. Frujeri, F. Stulp, G. Zhou, G. S. Sukhatme, G. Salhotra, G. Yan, G. Feng, G. Schiavi, G. Berseth, G. Kahn, G. Yang, G. Wang, H. Su, H.-S. Fang, H. Shi, H. Bao, H. B. Amor, H. I. Christensen, H. Furuta, H. Bharadhwaj, H. Walke, H. Fang, H. Ha, I. Mordatch, I. Radosavovic, I. Leal, J. Liang, J. Abou-Chakra, J. Kim, J. Drake, J. Peters, J. Schneider, J. Hsu, J. Vakil, J. Bohg, J. Bingham, J. Wu, J. Gao, J. Hu, J. Wu, J. Wu, J. Sun, J. Luo, J. Gu, J. Tan, J. Oh, J. Wu, J. Lu, J. Yang, J. Malik, J. Silvério, J. Hejna, J. Booher, J. Tompson, J. Yang, J. Salvador, J. J. Lim, J. Han, K. Wang, K. Rao, K. Pertsch, K. Hausman, K. Go, K. Gopalakrishnan, K. Goldberg, K. Byrne, K. Oslund, K. Kawaharazuka, K. Black, K. Lin, K. Zhang, K. Ehsani, K. Lekkala, K. Ellis, K. Rana, K. Srinivasan, K. Fang, K. P. Singh, K.-H. Zeng, K. Hatch, K. Hsu, L. Itti, L. Y. Chen, L. Pinto, L. Fei-Fei, L. Tan, L. J. Fan, L. Ott, L. Lee, L. Weihs, M. Chen, M. Lepert, M. Memmel, M. Tomizuka, M. Itkina, M. G. Castro, M. Spero, M. Du, M. Ahn, M. C. Yip, M. Zhang, M. Ding, M. Heo, M. K. Srirama, M. Sharma, M. J. Kim, N. Kanazawa, N. Hansen, N. Heess, N. J. Joshi, N. Suenderhauf, N. Liu, N. D. Palo, N. M. M. Shafiullah, O. Mees, O. Kroemer, O. Bastani, P. R. Sanketi, P. T. Miller, P. Yin, P. Wohlhart, P. Xu, P. D. Fagan, P. Mitrano, P. Sermanet, P. Abbeel, P. Sundaresan, Q. Chen, Q. Vuong, R. Rafailov, R. Tian, R. Doshi, R. Mart'in-Mart'in, R. Baijal, R. Scalise, R. Hendrix, R. Lin, R. Qian, R. Zhang, R. Mendonca, R. Shah, R. Hoque, R. Julian, S. Bustamante, S. Kirmani, S. Levine, S. Lin, S. Moore, S. Bahl, S. Dass, S. Sonawani, S. Tulsiani, S. Song, S. Xu, S. Haldar, S. Karamcheti, S. Adebola, S. Guist, S. Nasiriany, S. Schaal, S. Welker, S. Tian, S. Ramamoorthy, S. Dasari, S. Belkhale, S. Park, S. Nair, S. Mirchandani, T. Osa, T. Gupta, T. Harada, T. Matsushima, T. Xiao, T. Kollar, T. Yu, T. Ding, T. Davchev, T. Z. Zhao, T. Armstrong, T. Darrell, T. Chung, V. Jain, V. Kumar, V. Vanhoucke, W. Zhan, W. Zhou, W. Burgard, X. Chen, X. Chen, X. Wang, X. Zhu, X. Geng, X. Liu, X. Liangwei, X. Li, Y. Pang, Y. Lu, Y. J. Ma, Y. Kim, Y. Chebotar, Y. Zhou, Y. Zhu, Y. Wu, Y. Xu, Y. Wang, Y. Bisk, Y. Dou, Y. Cho, Y. Lee, Y. Cui, Y. Cao, Y.-H. Wu, Y. Tang, Y. Zhu, Y. Zhang, Y. Jiang, Y. Li, Y. Li, Y. Iwasawa, Y. Matsuo, Z. Ma, Z. Xu, Z. J. Cui, Z. Zhang, Z. Fu, and Z. Lin, "Open x-embodiment: Robotic learning datasets and rt-x models," 2024. [Online]. Available: https://arxiv.org/abs/2310.08864

[40] L. Y. Chen, K. Hari, K. Dharmarajan, C. Xu, Q. Vuong, and K. Goldberg, "Mirage: Cross-embodiment zero-shot policy transfer with cross-painting," 2024. [Online]. Available: https://arxiv.org/abs/2402.19249

[41] W. Huang, I. Mordatch, and D. Pathak, "One policy to control them all: Shared modular policies for agent-agnostic control," 2020. [Online]. Available: https://arxiv.org/abs/2007.04976

[42] J. Yang, C. Glossop, A. Bhorkar, D. Shah, Q. Vuong, C. Finn, D. Sadigh, and S. Levine, "Pushing the limits of cross-embodiment learning for manipulation and navigation," 2024. [Online]. Available: https://arxiv.org/abs/2402.19432

[43] J. Yang, D. Sadigh, and C. Finn, "Polybot: Training one policy across robots while embracing variability," 2023. [Online]. Available: https://arxiv.org/abs/2307.03719

[44] K. Grauman, A. Westbury, E. Byrne, Z. Chavis, A. Furnari, R. Girdhar, J. Hamburger, H. Jiang, M. Liu, X. Liu, M. Martin, T. Nagarajan, I. Radosavovic, S. K. Ramakrishnan, F. Ryan, J. Sharma, M. Wray, M. Xu, E. Z. Xu, C. Zhao, S. Bansal, D. Batra, V. Cartillier, S. Crane, T. Do, M. Doulaty, A. Erapalli, C. Feichtenhofer, A. Fragomeni, Q. Fu, A. Gebreselasie, C. Gonzalez, J. Hillis, X. Huang, Y. Huang, W. Jia, W. Khoo, J. Kolar, S. Kottur, A. Kumar, F. Landini, C. Li, Y. Li, Z. Li, K. Mangalam, R. Modhugu, J. Munro, T. Murrell, T. Nishiyasu, W. Price, P. R. Puentes, M. Ramazanova, L. Sari, K. Somasundaram, A. Southerland, Y. Sugano, R. Tao, M. Vo, Y. Wang, X. Wu, T. Yagi, Z. Zhao, Y. Zhu, P. Arbelaez, D. Crandall, D. Damen, G. M. Farinella, C. Fuegen, B. Ghanem, V. K. Ithapu, C. V. Jawahar, H. Joo, K. Kitani, H. Li, R. Newcombe, A. Oliva, H. S. Park, J. M. Rehg, Y. Sato, J. Shi, M. Z. Shou, A. Torralba, L. Torresani, M. Yan, and J. Malik, "Ego4d: Around the world in 3,000 hours of egocentric video," 2022. [Online]. Available: https://arxiv.org/abs/2110.07058

[45] L. Ma, Y. Ye, F. Hong, V. Guzov, Y. Jiang, R. Postyeni, L. Pesqueira, A. Gamino, V. Baiyya, H. J. Kim, K. Bailey, D. S. Fosas, C. K. Liu, Z. Liu, J. Engel, R. D. Nardi, and R. Newcombe, "Nymeria: A massive collection of multimodal egocentric daily motion in the wild," 2024. [Online]. Available: https://arxiv.org/abs/2406.09905

[46] S. Haddadin, S. Parusel, L. Johannsmeier, S. Golz, S. Gabl, F. Walch, M. Sabaghian, C. Jähne, L. Hausperger, and S. Haddadin, "The franka emika robot: A reference platform for robotics research and education," *IEEE Robotics and Automation Magazine*, vol. 29, no. 2, pp. 46–64, 2022.

[47] Meta Research, "Basics — project aria docs," https://facebookresearch.github.io/projectaria_tools/docs/data_formats/mps/mps_summary, 2024, accessed: September 15, 2024.

[48] J. Yang, C. Glossop, A. Bhorkar, D. Shah, Q. Vuong, C. Finn, D. Sadigh, and S. Levine, "Pushing the limits of cross-embodiment learning for manipulation and navigation," *arXiv preprint arXiv:2402.19432*, 2024.

[49] J. Hejna, C. Bhateja, Y. Jian, K. Pertsch, and D. Sadigh, "Re-mix: Optimizing data mixtures for large scale imitation learning," *arXiv preprint arXiv:2408.14037*, 2024.

[50] Y. Zhou, Y. Aytar, and K. Bousmalis, "Manipulator-independent representations for visual imitation," 2021. [Online]. Available: https://arxiv.org/abs/2103.09016

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

[52] 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. [Online]. Available: https://arxiv.org/abs/2408.00714

[53] L. Wang, X. Chen, J. Zhao, and K. He, "Scaling proprioceptive-visual

learning with heterogeneous pre-trained transformers," in *Neurips*, 2024.

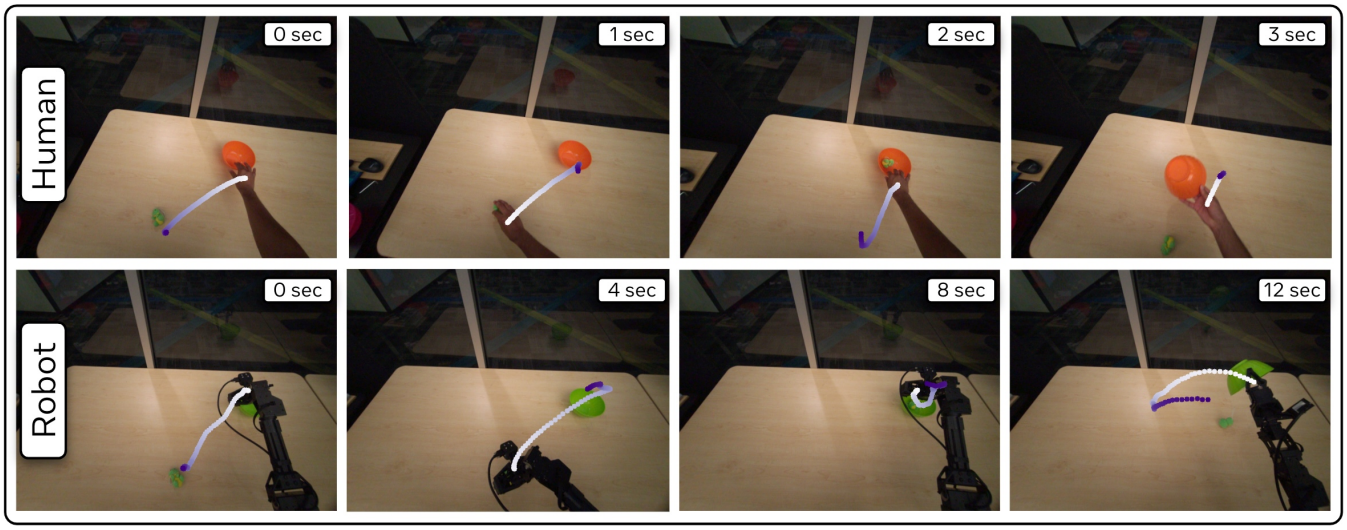

Fig. 9: Here we visualize hand and robot data from out dataset side by side with ground truth actions overlayed (purple). Note that the actions are of similar length, despite the hand traveling much faster than the robot.

VI. Appendix

A. Data Processing and Domain Alignment

Humans and teleoperated robots complete tasks at different speeds. To enable joint training of human and robot data, we must align these two sources of data temporally. Following Mimicplay [6], we "slow down" the human data, and we found empirically that a factor of 4 sufficiently aligned both domains. Specifically, for robot data we construct joint and pose based actions over a four second horizon but for human data we use a 1 second horizon. For both domains, our action chunk size is 100, meaning we construct 100 future actions spaced evenly over the horizon. This alignment is independent of data recording frequencies, where human data is recorded at 30hz and robot data is recorded at 50 hz.

To co-train on both human and robot data, we individually normalize the proprioception and actions for both embodiments (as shown in Fig. 3). Given proprioception $p_t \in \mathbb{R}^d$ where d depends on embodiment, we normalize by subtracting the dataset mean and dividing by standard deviation

$$norm(p_t) = (p_t - \mu_p)/\sigma_p.$$

We perform the identical calculation to normalize actions $a_{t:t+h} \in \mathbb{R}^{d\times 100}$.

To bridge the appearance gap between human hand and robot arm, we visually mask each embodiment via SAM2 [52], and overlay a red line on these masks to enhance alignment (Fig. 3). For the robot, we first use forward kinematics to compute the 3D coordinates of key joints in robot frame

$$p_t^R = FK(q_t) \in \mathbb{R}^{3\times 3},$$

including the wrist, gripper and the forearm. These 3D coordinates are then projected onto the image frame via camera intrinsics (I_{cam}^{pixels}) and extrinsics (T_R^{cam}) to obtain 2D keypoints in pixel space

$$p_t^{pixel} = I_{cam}^{pixels} T_R^{cam} p_t^R \in \mathbb{R}^{3\times 2},$$

which are used to prompt SAM2 [52] to generate a mask of the robot arm. After obtaining the mask, we draw a red line on the masked area from the gripper to the elbow in the RGB image. For the human data, a similar process is followed, where SAM2 is prompted using the 3D coordinates of the human hand to generate a mask. A red line is then drawn along the hand's contour, from the bottom right to top left corner of the contour's bounding box.

During training, both the robot arm and human hand are masked to align their visual representations. During evaluation, SAM2 is run in real time on a desktop to mask the robot arm and apply the same red line overlay. This approach enables better visual alignment between the robot and human hand, facilitating more effective model generalization across human and robotic tasks.

B. Aria Machine Perception Services (MPS)

We leveraged MPS to process human data from the Aria glasses. The raw data from Aria contains timestamped sensor information from the glasses, namely RGB camera, SLAM cameras, IMU, eye tracking cameras, microphone, and more. The raw data is uploaded to the MPS server, where the cloud-hosted service estimates device pose via SLAM, a semi-dense pointcloud of the environment, hand tracking relative to the device frame, and even eye gaze. The MPS returns SLAM as a timestamped CSV of device poses in world frame and hand tracking as a timestamped CSV of cartesian positions in the time-aligned device frame. These hand positions are each in a distinct reference frame due to head movements, so we project future actions to the current device coordinate frame (described in Sec. III-B). We use the undistorted Aria RGB camera data paired with the hand tracking and SLAM information to construct an hdf5 file compatible for training in robomimic [10].

C. Training Human-Robot Joint Policies

We depict our algorithm in detail in Fig. 10. At each step we sample a batch of hand data as well as a batch of

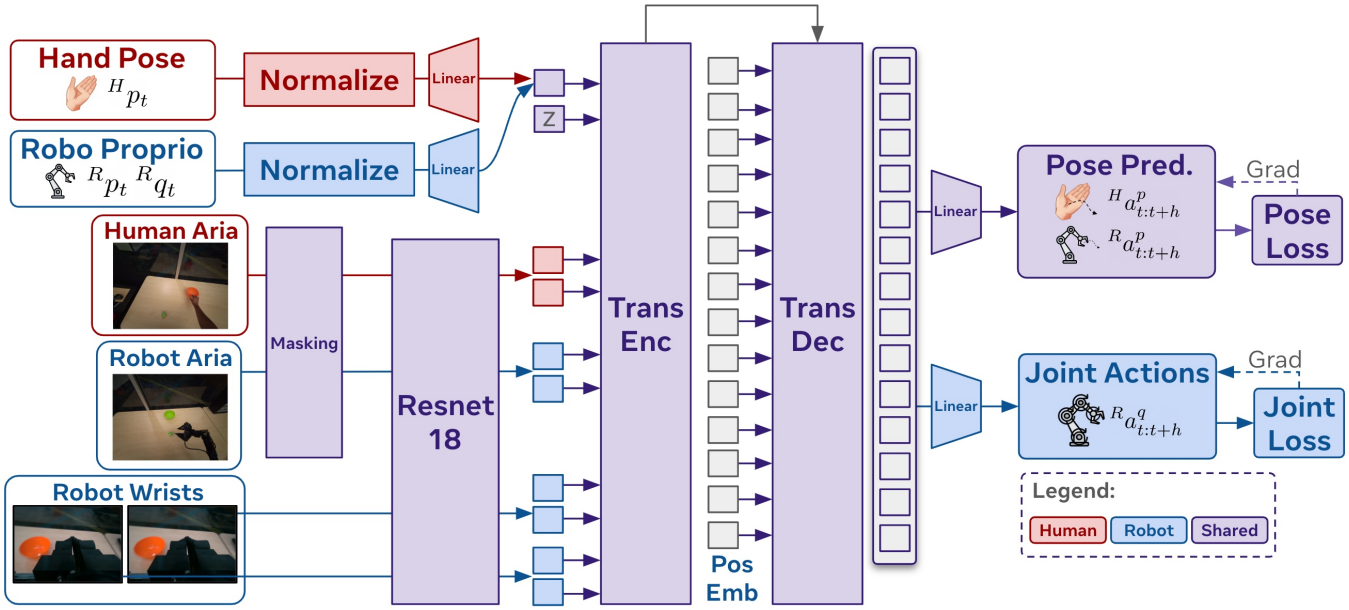

Fig. 10: Detailed Architecture of EgoMimic.

robot data, and pass each through our unified architecture. EgoMimic performs Z-score normalization to hand and robot proprioception and actions individually. The normalized proprioception is passed through a linear layer to produce a proprioception token. Alongside the proprioception, the top down views from hand and robot are passed through a SAM based masking module. These images, along with the robot wrist views are passed through a shared Resnet18 visual encoder which produces visual tokens. Finally, we add an additional style token z from our CVAE encoder which is not depicted, but directly follows ACT [2]. All these tokens, are passed through a transformer encoder decoder architecture. The transformer decoder's hidden output is passed through a linear decoder depending on the output type, producing pose actions $\hat{a}^p$ or joint based actions $\hat{a}^j$.

For batches of robot data, we calculate

$$L_{robot} = L_1(^{R}\hat{a}^p, ^{R}a^p) + L_1(^{R}\hat{a}^j, ^{R}a^j) + KL$$

and for hand data we have

$$L_{hand} = L_1(^{H}\hat{a}^p, ^{H}a^p) + KL$$

where KL is the CVAE latent regularizer as in ACT [2]. This yields $L = L_{robot} + L_{hand}$ which we optimize at each step.

We leverage the transformer's flexible input sequence to account for differences in the number of visual observations based on the modality; specifically we have wrist images in robot data but not hand data. When the wrist images are present, we concatenate additional tokens to our transformers input sequence as in ACT [2]. In our experiments, we found that this strategy was sufficient to effectively co-train on both hand and robot data, although we plan to experiment with more sophisticated cross-embodiment learning techniques like HPTs [53].

We note that the human data lacks information for the grasping action, since Aria only records hand pose. Thus, the grasping action is supervised only via the robot joint prediction loss $L_1(^{R}\hat{a}^j, ^{R}a^j)$, where the gripper is represented as another joint.

TABLE IV: Training details - EgoMimic

Policy	ACT
Batch Size	128
Optimizer	adamw
Learning rate (initial)	5e-5
Decay factor	1
Scheduler	Linear
Encoder layers	4
Decoder layers	7
Hidden dim	512
Feedforward dim	3200
No. of heads	8
Data Augmentations	Color Jitter

D. Training Details

We list the hyperparameters for EgoMimic in Table IV. All models were trained for 120000 iterations with global batch size of 128 across 4 A40 gpus, which takes about 24 hours. Our code is implemented in the robomimic framework [10]. More details in Table IV

E. Mimicplay Implementation

For our implementation of MimicPlay [6], we closely follow the original setup, training the high-level planner and low-level control policy separately.

First, we train a ResNet-18 based high-level encoder using a Gaussian Mixture Model (GMM) to generate 3D trajectories, as described in the original work. The high-level encoder is trained on both human and robot data to predict 3D trajectories.

Once the high-level encoder is trained, we extract the latent representation from the ResNet-18 encoder (i.e., the high-level planner) and use it as the style variable z, which is

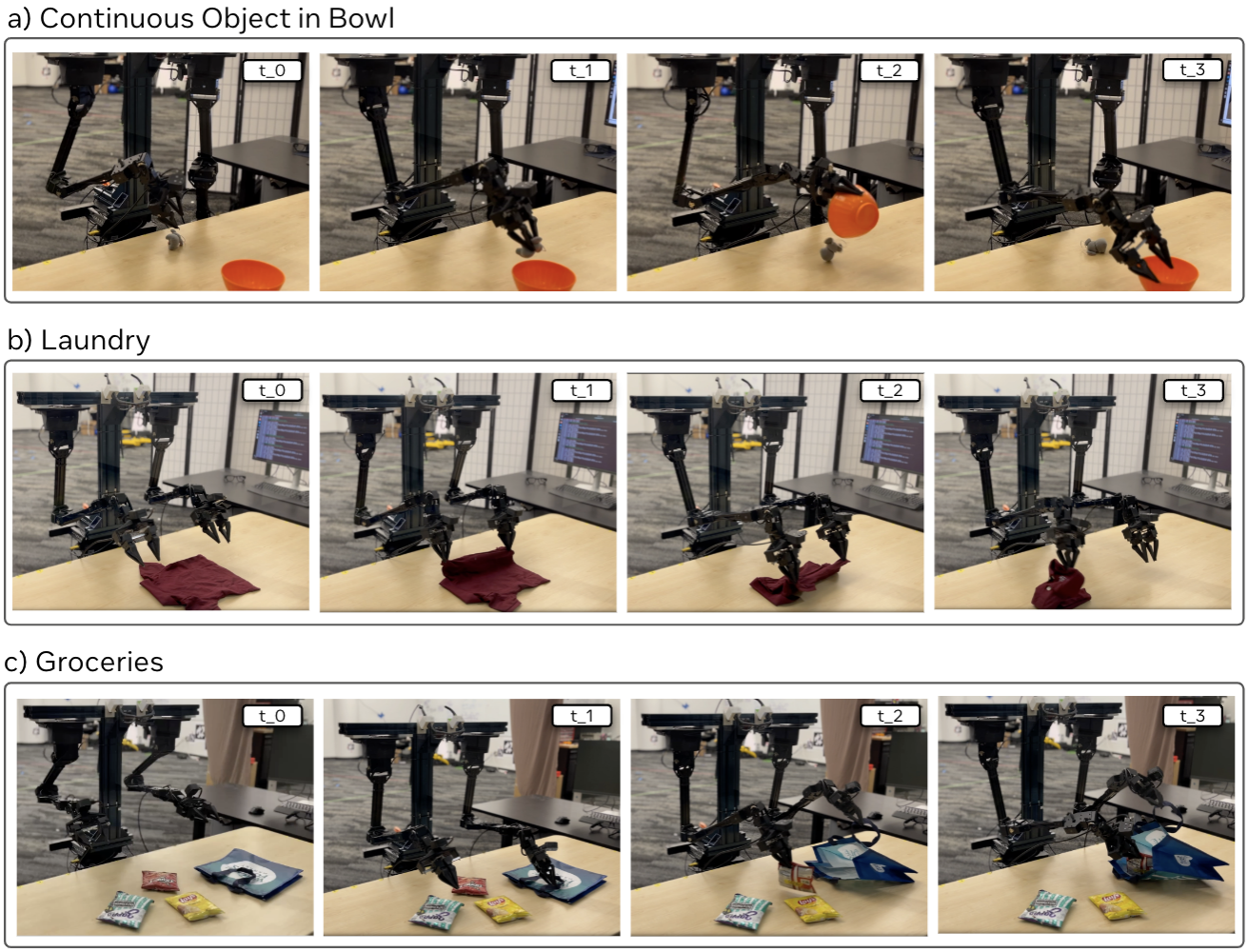

Fig. 11: Qualitative successes of EgoMimic on each of our three tasks.

TABLE V: Training details - Mimicplay

High-level	Resnet18
Learning rate (initial)	0.0001
Decay factor	0.1
Batch size	50
GMM modes	5
Low-level	ACT
Learning rate (initial)	5e-5
Optimizer	adamw
Decay factor	1
Scheduler	Linear

passed to the transformer encoder-decoder Fig. 10. The low-level ACT policy is then trained solely on robot data with this additional input from the high level policy as guidance.

F. Policy Rollout

We rollout our policy with inference at 1hz and control at 25hz on a desktop with a an NVIDIA RTX 4090 GPU. The predicted action horizon is 4 seconds, with the first second of predicted actions executed in receding-horizon style. All the robot's sensors update at 50hz with the exception of the Aria camera which streams at 30fps.

TABLE VI: Data recording and rollout rates for Human and Robot data. We "slow down" human data by 0.25 to account for differences in task execution speeds.

Type	**Human (Hz)**	**Robot (Hz)**
Recording	30	50
Rollout (Inference)	-	1
Rollout (Control)	-	25