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†] *

Abstract: The scale and diversity of demonstration data required for imitation learning is a significant challenge. We present EgoMimic, a full-stack framework to scale manipulation via egocentric human demonstrations. EgoMimic introduces a data collection system built on the Project Aria glasses, as well as an algorithm which can leverage these demonstrations as a native data source. This approach improves performance over state-of-the-art imitation learning algorithms on a set of real-world single-arm and bimanual manipulation tasks and enables generalization to entirely new scenes. Finally, we exhibit favorable scaling properties, and find adding 1 hour of additional human data is more valuable than 1 hour of additional robot data. Videos available at ego-mimic.github.io

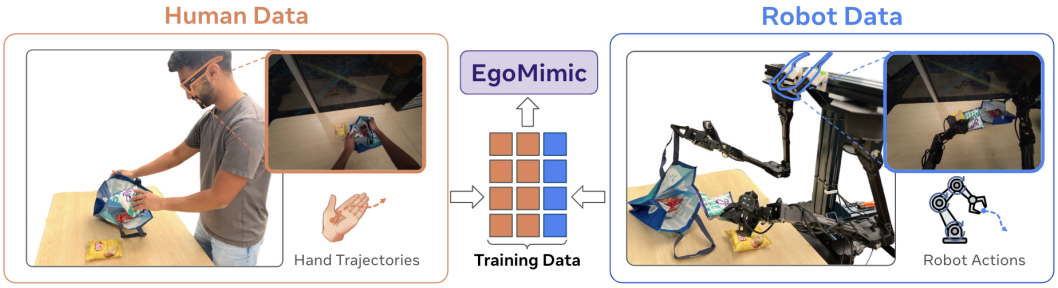

Figure 1: EgoMimic enables anyone to collect human demonstrations for imitation learning, simply by wearing a pair of Project Aria glasses [1]. Aria glasses record egocentric vision paired with hand tracking, which we use to augment our robot training data. When combined, it can boost task performance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

1 Introduction

End-to-end imitation learning has shown remarkable performance in learning complex manipulation tasks [2, 3, 4, 5], but it remains brittle when facing new scenarios and tasks. Recent works like RT1 and RT2 improve generalization [6, 7], but require months of data collection effort. We see this as a scalability bottleneck: where other domains like Natural Language Processing and Computer Vision leverage Internet-sourced data to scale performance, robotics lacks such an equivalent.

To scale up data for robotics, there have been recent advances in data collection systems. For example with intuitive leader-follower style teleoperation [2, 8, 9], or with hand-held grippers to collect data without a robot [10]. However, these systems still require *active* effort in providing demonstrations. We hypothesize that *passive data collection* [11] is critical to achieving Internet-scale robot data. Just as the Internet was not built to curate data for vision and language models, an ideal robot data system should allow users to generate sensorimotor behavior data without intending to do so.

Prior work has recognized the potential of human data for passive data scaling, but these works generally view human data as *auxiliary data source* to train visual representations [12, 13, 14] or understand scene dynamics through point track prediction [15, 16], intermediate state hallucination in pixel space [17, 18], or affordance prediction [19]. We argue that to effectively scale robot performance using human data, human videos should not be treated as an auxiliary data source requiring

*[1] Georgia Institute of Technology. [2] Stanford University. *Equal contribution. †Equal advising.

8th Conference on Robot Learning (CoRL 2024), Munich, Germany.

separate processing. 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.

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 egocentric human videos and teleoperated robot data consisting of: **(i)** a system to collect human data via Project Aria glasses [1] **(ii)** a capable and low-cost bimanual robot which minimizes the kinematic gap to human data **(iii)** data normalization and alignment techniques to close the human-robot gap **(iv)** a unified imitation learning architecture which co-trains on hand and robot data with a common vision encoder and policy network.

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. 3). EgoMimic enhances performance in all scenarios, with relative improvements of up to 200%. Further, we observe that EgoMimic exhibits generalization to objects and scenes encountered exclusively in human data. Finally, we analyze the scaling properties of EgoMimic, and find that training on an additional hour of hand data significantly outperforms training on an additional hour of robot data.

2 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 innovate across the full stack from human and robot data collection to algorithmic improvements.

2.1 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. EgoMimic fills this gap by building a data collection system on top of the Project Aria glasses. A user can simply wear these glasses and perform tasks with their own hands, which we process and use for imitation learning. These glasses have a wide FOV RGB camera and two mono-color scene cameras to estimate device pose and hand-tracking (see Appendix A.1 for more details).

Low-cost bimanual manipulator. To better learn from human data, we built a bimanual robot whose movement resembles that of humans. It consists of two 6DoF ViperX arms with wrist cameras and Aria glasses for vision, mounted on a height-adjustable torso (Fig 4 and Appendix A.2).

2.2 Data Processing and Alignment

EgoMimic bridges three key human-robot gaps to promote seamless co-training: (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. 2.1. 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) [20] to estimate 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 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 actions and proprioception typically use fixed reference frames (e.g., camera or robot base), but egocentric hand data from a moving camera breaks this assumption. To unify these frames, we transform both robot and human data into a camera-centered stable reference frame. For robot data we simply perform hand eye calibration to map robot base actions into camera frame. For human data, we leverage Aria's onboard SLAM to account for device movement, and project hand positions to a stable frame (see Appendix B.1).

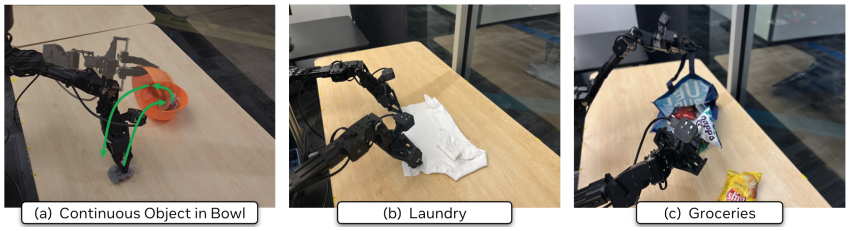

Figure 3: We evaluate EgoMimic across three real world, long-horizon manipulation tasks.

Aligning human-robot pose distributions. Despite aligning human and robot data via hardware design and data processing, there are still differences in the distributions of hand and robot end effector poses, which prevent effective co-training [21, 22]. Drawing from prior work, we apply Gaussian normalization to the actions and proprioception from each data source (Fig. 5) [22].

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 [23], [24]. We follow similar ideas and mask out both the hand and the robot via SAM [25] and overlay a red line to indicate end-effector directions (Fig 5).

2.3 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 [5, 17]. However, this approach is inherently limited by the performance of the low-level policy, which does not directly benefit from large-scale human data.

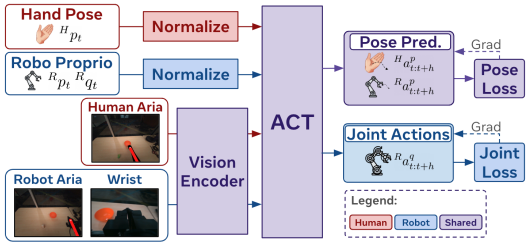

Figure 2: EgoMimic processes normalized hand and robot data through shared vision and policy networks. It outputs pose predictions for both human and robot data, and joint predictions robot data.

In contrast, our architecture enables unified co-training for both hand and robot data. 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 map 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 ${}^{R}a^{q}$. Our architecture which processes both domains of data with shared parameters helps force the model to learn a joint human-robot representation, and drives performance scaling via human data (Sec. 3.2).

3 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.

3.1 Experiment Setup

Tasks. We evaluate our approach on a set of long-horizon real-world tasks (Fig. 3). *Continuous Object-in-Bowl:* The robot picks a small toy, places it in a bowl, dumps it out, and continuously repeats the process. *Laundry:* The robot folds a t-shirt using both arms in several stages. *Groceries:* The robot grasps and opens a shopping bag with one arm while placing items inside it with the other

Table 1: 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. (0% H) = No Human Data

Method	Bowl	Laundry		Groceries		
	Pts	Pts	SR	Pts	SR	Open Bag
ACT [2]	39	82	55%	82	22%	54%
Mimicplay [5]	71	78	50%	53	8%	40%
EgoMimic (0% H)	68	104	73%	92	28%	60%
EgoMimic	**128**	**114**	**88%**	**110**	**30%**	**70%**

Table 2: **Ablations** - We ablate our method and report final task performance on *Object-in-Bowl*.

Method	Pts
EgoMimic	128
w/o Line	112
w/o Line and Mask	95
w/o Action Norm	79
w/o Hand Data	68

arm. Performance is measured in points (**pts**) and success rate (**SR**) across multiple trials for each task. See Appendix D for details on task description.

Baselines. We benchmark EgoMimic against ACT [2], a state-of-the-art imitation learning algorithm, as well as Mimicplay [5], a state-of-the-art method which leverages human hand data.

3.2 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), and find we improve score by 10-88% and success rate by 2-15%.

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). Further, 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, outperforming Mimicplay (4 pts) and ACT (7 pts).

Scaling human vs. robot data. To investigate the scaling effect of human and robot data sources on performance, we conducted additional data collection for the *Cont. Object-in-bowl* task. As illustrated in Fig. 8, 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. Ablations on the *Object-in-Bowl* task (Table 2) indicate that human data (-47%), action normalization (-38%) and hand/robot masking (-26%) are integral to EgoMimic.

4 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 real-world tasks, shows generalization to new scenes, and has favorable scaling properties.

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. URL 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. URL https://arxiv.org/abs/2304.13705.

[3] 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. URL https://arxiv.org/abs/2303.04137.

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

[5] 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. URL https://arxiv.org/abs/2302.12422.

[6] 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. URL https://arxiv.org/abs/2212.06817.

[7] 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. URL https://arxiv.org/abs/2307.15818.

[8] 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. URL https://arxiv.org/abs/2405.02292.

[9] 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.

[10] 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. URL https://arxiv.org/abs/2402.10329.

[11] 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*, pages 19383–19400, 2024.

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

[13] 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*, pages 416–426. PMLR, 2023.

[14] 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.

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

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

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

[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. URL https://arxiv.org/abs/2101.07241.

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

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

[21] 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.

[22] 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.

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

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

[25] 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.

[26] 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. URL https://arxiv.org/abs/2110.07058.

[27] 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. URL https://arxiv.org/abs/2406.09905.

[28] 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*, 29(2):46–64, 2022. doi: 10.1109/MRA.2021.3138382.

A Data Collection Systems and Hardware Design

A.1 Aria glasses for egocentric demonstration collection

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. 4 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 [26, 27], 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.

A.2 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 [28]. 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 arms mounted in an inverted configuration on a height-adjustable rig as the torso (Fig 4), 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].

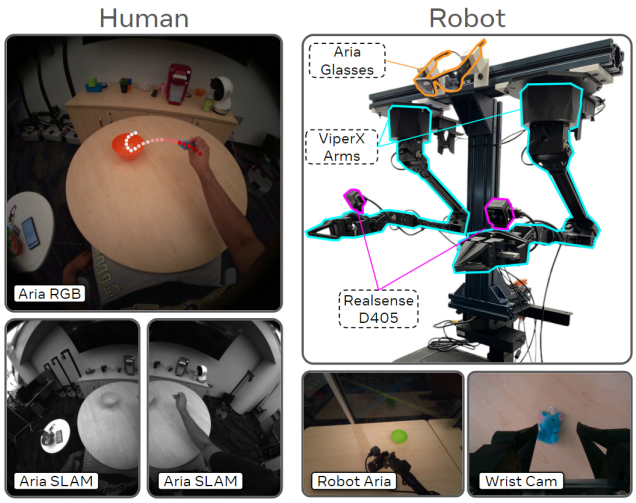

Figure 4: Our system uses Aria glasses to capture Egocentric RGB and uses its side SLAM cameras to localize the device and track hands. The robot consists of two Viper X arms with Intel RealSense wrist cameras. Our robot uses identical Aria glasses as the main vision sensor to help minimize the camera-to-camera gap.

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. 2.2), 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 4). This enables us to mitigate the observation domain gap associated with the camera devices, including FOVs, exposure levels, and dynamic ranges.

B Data Processing and Domain Alignment

B.1 Proprioception and Visual Alignment

We independently normalize the action distributions in both human and robot embodiments to ensure alignment across modalities. To further mitigate the appearance discrepancy between these embodiments, we apply a red line overlay and masking using SAM2 [25], as shown in Fig. 5. The SAM point prompts are generated by the robot end effector and human hand poses transformed to image frames. These preprocessing steps facilitate the seamless integration of both data sources, enabling effective co-training of our model on the dual modalities.

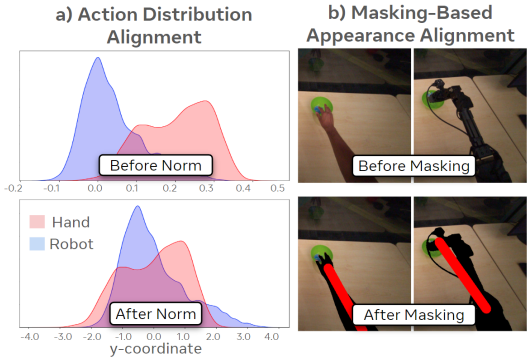

Figure 5: **a) Action normalization:** The pose distributions are different between hand and robot data, specifically in the y (left-right) dimension. We apply Gaussian normalization individually to the hand and robot pose data before feeding them to the model. **b) Visual masking:** To help bridge the 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.

B.2 Unifying human-robot 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 [3, 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 $[p_t^{F_t}, p_{t+1}^{Ft+1}, ...p_{t+h}^{Ft+h}]$, 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:

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

A sample trajectory is visualized in Fig. 4 (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.

C Robot Action Space

A critical challenge in this unified approach is the choice of the robot action space. While the robot end-effector poses align more closely with human data in semantic meaning and format, controlling our robot with end-effector poses via cartesian-based controller (e.g., differential IK) proves difficult: the 6 DoF ViperX arms offer low solution redundancy, and we empirically found that robots often encounter singularities or non-smooth solutions in a trajectory. Consequently, we opt for joint-space control and use pose space prediction for learning joint human-robot representation.

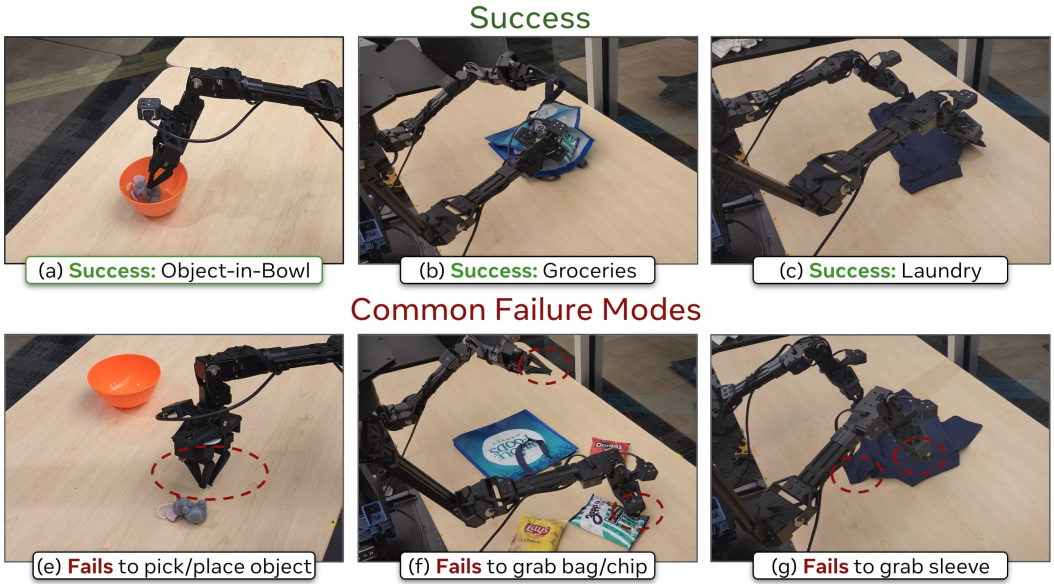

Figure 6: We highlight EgoMimic's success, as well as failure modes, for instance **(e)** failure to correctly align with the toy, **(f)** failure to grasp the bag's handle, or **(g)** policy only grabs 1 side of the shirt. **EgoMimic** reduces the frequency of these failure modes, improving success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

D Experiments

D.1 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. 3).

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 $\pm 30\,\mathrm{deg}$. 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.

For *Continuous Object-in-Bowl* we collect 1 hour of human data and 2 hours robot data. For the other tasks we collect 1.5 hours of hand data and 5 hours of robot data.

D.2 Results

EgoMimic's success and common failure modes We evaluate EgoMimic's performance by highlighting both its successes and failure modes across different tasks. While EgoMimic significantly reduces the occurrence of common failure modes such as misalignment with objects, improper grasps, or partial actions like grabbing only one side of a shirt, there are still instances where these issues arise. Fig. 6 illustrates examples of such failure cases, including (e) failure to align with the toy, (f) failure to grasp the bag's handle, and (g) incomplete shirt grasping. Despite these, EgoMimic demonstrates an 8-33% improvement in success rates compared to baselines.

Generalization to new objects and scenes EgoMimic demonstrates strong generalization to unseen objects, such as performing laundry on an unseen shirt, where baselines fail. Quantitatively, EgoMimic exhibits only a minor performance drop of 3-5%, compared to a significant 30% drop observed in the baselines when handling unseen objects in the laundry task. Additionally, EgoMimic is able to generalize to scenes that were only present in the human data. (See Fig.7).

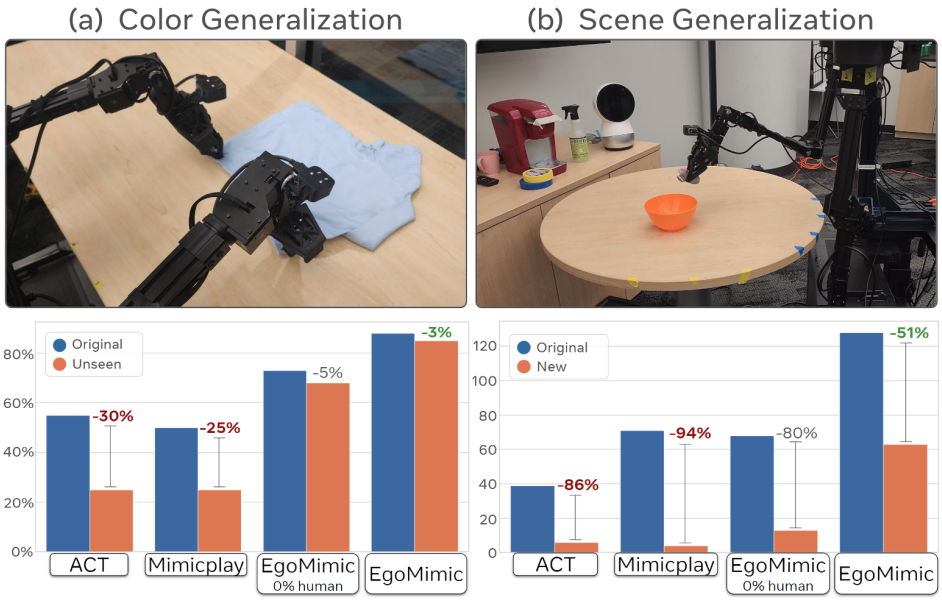

Figure 7: **Evaluation Results on Policy Generalization.** (a) We evaluate the policy on the laundry task using unseen cloth colors and report the success rate for each method. (b) We test the policy on the Object-in-Bowl task in unseen scenes.

Scaling properties We also find that EgoMimic's performance scales effectively with the addition of human demonstrations, when given a sufficient amount of teleoperation data. Specifically, EgoMimic trained on 2 hours of robot data and 1 hour of human data significantly outperforms ACT, which is trained on 3 hours of robot data (128 vs. 74 points). Our results demonstrate favorable scaling properties, showing that adding 1 hour of human data is more beneficial than adding 1 hours of additional robot data (see Fig. 8).

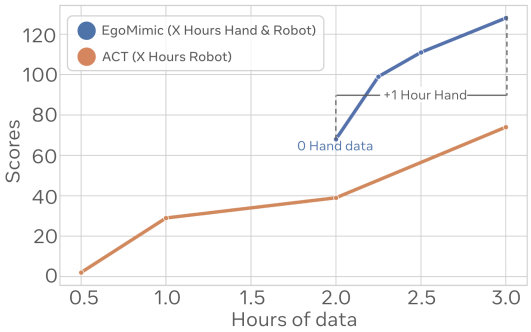

Figure 8: **Scaling robot vs. human data.** EgoMimic trained on 2 hours robot data + 1 hour hand data (Blue) strongly outperforms ACT [2] trained on 3 hours of robot data (Orange).