Better Demonstrations, Not More: Human Egocentric Video over Teleoperation

Zhi (Leo) Wang, Botao He, Kelin Yu, Seungjae Lee,
Ruohan Gao, Furong Huang, Yiannis Aloimonos
University of Maryland

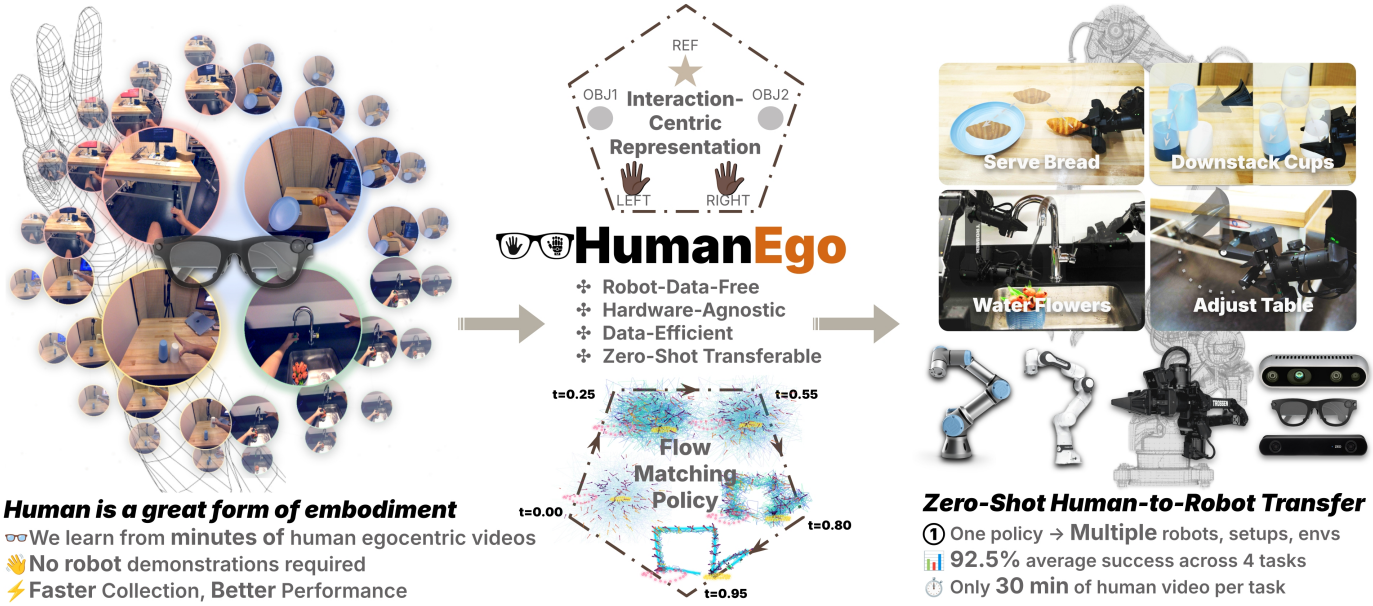

Fig. 1: Wear glasses, don't teleoperate. A human wears Aria glasses and collects demonstrations in seconds (**left**); each egocentric video is lifted into an interaction-centric representation and used to train a flow matching policy (**middle**); the policy deploys *zero-shot* on a robot, with no robot data (**right**). We argue—and measure—that these human demonstrations are a *higher-quality* source than robot teleoperation.

***Abstract*—Imitation learning is usually fed by robot teleoperation, and the field's instinct when a policy underperforms is to collect *more* of it. We make the opposite case: at a fixed time budget the *quality* of demonstrations matters more than their count, and human egocentric video is a higher-quality demonstration source than teleoperation. We support this with a complete zero-shot human-to-robot system: each human video is lifted into an interaction-centric representation of hand–object geometry and used to train a flow matching policy that deploys directly on a robot, with no robot data and no large-scale pretraining. On four real-world tasks the policy reaches 92.5% average success, beats five recent human-video baselines on every task, surpasses time-matched teleoperation by 41%, and generalizes *zero-shot* across novel robots, cameras, and environments. We then quantify *why* human demonstrations are better: higher signal-to-noise ratio, smoother motion, less idle time, and broader coverage; 8 minutes of human video matches 30 of teleoperation; mixing in robot data only lowers success; and trajectory smoothness and tracking persistence—not pose accuracy—predict downstream success. These are concrete, measurable demonstration-quality axes, and they all point the same way: "wear glasses, don't teleoperate" is an underrated way to raise data quality.**

Correspondence to Zhi (Leo) Wang (tx.leo.wz@gmail.com), University of Maryland.

I. Introduction

The dominant recipe for imitation-based manipulation is to teleoperate a robot, log the trajectories, and—when the policy is not good enough—teleoperate more [1, 7]. This treats demonstrations as interchangeable units whose only axis is *quantity*. This workshop asks the sharper question: what makes a demonstration *good*, beyond the fact that it completed the task? We approach it empirically by comparing two demonstration *sources* for the same manipulation tasks—robot teleoperation and human egocentric video [3]—and asking which yields better policies per minute of collection, and *why*.

Our position is that human egocentric video is the higher-quality source, and we argue it on three fronts. First, a simple pipeline turns raw human video into a policy that deploys directly on a robot (Sec. II). Second, that policy *works* and *generalizes* on real hardware—92.5% average success across four tasks, a 41% improvement over time-matched teleoperation, and robust zero-shot transfer across embodiments and environments (Sec. IV, V). Third, the human source is *measurably* higher quality, and those quality properties—smoothness, signal-to-noise, idle time, coverage, tracking persistence—predict policy success better than demonstration count (Sec. III, VI). Taken together, they make a practical case that changing the *source* of demonstrations can do more for a policy than scaling how many you collect.

II. From Human Video to Robot Policy

To compare sources fairly we need a learner that consumes both. The challenge is the *embodiment gap*: a human hand and a robot gripper differ in both visual appearance and kinematics, so raw human video cannot be fed to a robot policy directly. HumanEgo [17] closes this gap with three off-the-shelf components; we summarize them here and give full detail in App. A.

(1) Visual transform. We segment the human hand and arm and remove them with LaMa inpainting [16], then render a virtual gripper and tracked object keypoints into the inpainted frame, yielding an embodiment-agnostic RGB observation without expensive image translation.

(2) Interaction-centric representation. We treat every hand and object as an *entity* and encode each entity's 6-DoF pose relative to both hands into an Interaction-Centric Token (ICT). Because every quantity is expressed relative to scene entities rather than the camera, the tokens are invariant to viewpoint and embodiment—identical whether the scene was seen through Aria glasses or a robot's camera—which is what makes human-to-robot transfer possible. Hands come from Aria MPS stereo tracking [3]; objects from Grounding DINO [9], SAM2 [14], and CoTracker3 [6] triangulated against the Aria SLAM pose. No robot data and no ground-truth labels are used.

(3) Flow matching policy. A flow matching [8] policy consumes the ICT tokens and an RGB frame and generates multimodal bimanual actions; three dense auxiliary objectives (object-motion, 2D-trace, latent-consistency forecasting) supply per-timestep supervision that makes the policy data-efficient enough to learn from minutes of video. The *same* pipeline, policy, optimizer, and training budget consume human and teleoperated data alike, so any difference in policy success reflects the demonstrations—not the learner.

III. Human Demonstrations Are Higher Quality

Measuring trajectory-level properties of the two sources (Fig. 2) reveals a consistent quality gap. Human demonstrations have higher signal-to-noise ratio, an order-of-magnitude smoother motion profile, and near-zero idle time, while teleoperation accumulates pauses, corrective jitter, and piloting overhead. Spatially, human demonstrations cover the workspace more densely and trace more diverse trajectories, because a person manipulating with their own hand naturally explores more of the task distribution than an operator driving a robot through a constrained interface (the "robot wizard" bottleneck this workshop highlights). These differences begin at collection time: across the four tasks a single human demonstration takes **∼30–40 s**, while a matched teleoperated demonstration on the same hardware takes **∼60–70 s**. Everyday human manipulation is faster and more dexterous than driving a robot through a piloting interface, so a practitioner collects *more* demonstrations, *covering more of the task*, in *less* time—higher quality and higher throughput at once.

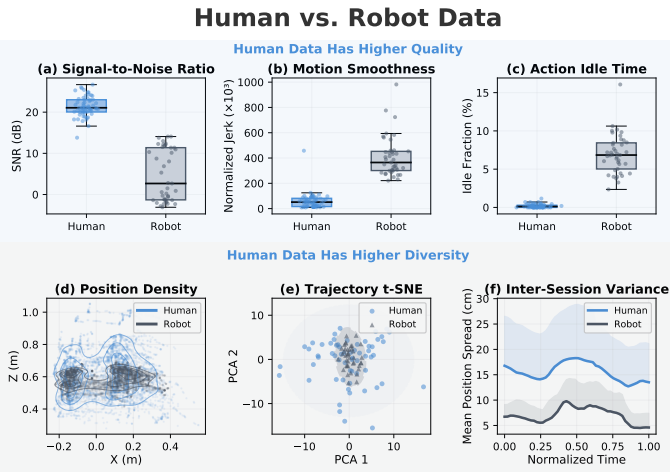

Fig. 2: Human vs. robot demonstrations on the same tasks. Human egocentric data shows higher signal-to-noise ratio, smoother motion, and less idle time (top), and greater spatial and trajectory diversity (bottom)—each a concrete, measurable notion of demonstration quality.

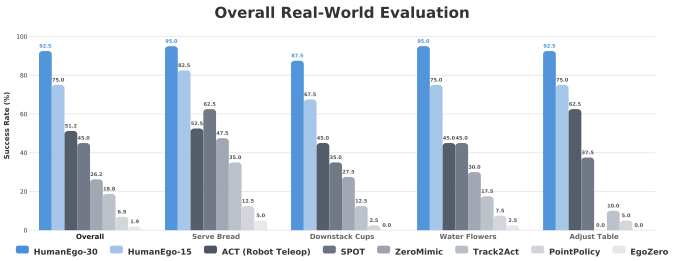

Fig. 3: Real-world success across four tasks. From 30 min of human video per task, the policy achieves the highest success on every task against five human-video baselines, and surpasses ACT trained on matched-time robot teleoperation.

IV. Real-World Results

Higher data quality should, if real, translate into better policies. It does. On four real-world tasks—*Serve Bread*, *Downstack Cups*, *Water Flowers*, *Adjust Table* (App. B)—the policy trained on 30 min of human video reaches **92.5%**

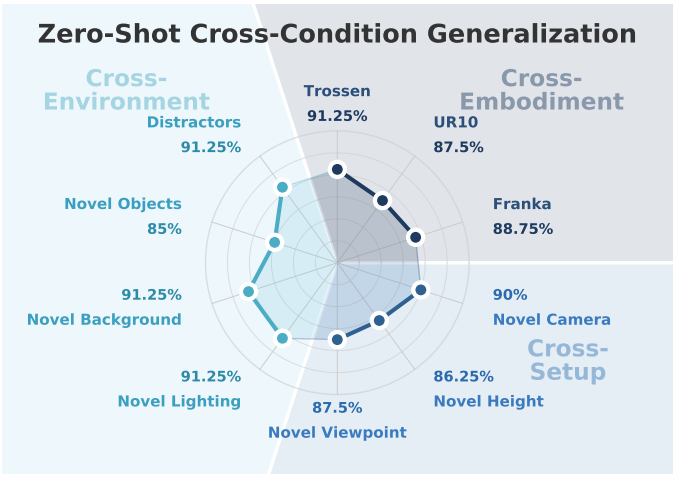

Fig. 4: Zero-shot cross-condition generalization. Trained once on human video and deployed without retraining, the policy stays robust across backgrounds, lighting, viewpoints, distractors, object placements, and hardware.

average success and is the best method on *every* task (Fig. 3), against five recent zero-shot baselines that learn from the same human videos (EgoZero [10], Point Policy [4], ZeroMimic [15], Track2Act [2], SPOT [5]), which range from 1.9% to 45.0%.

Human video beats teleoperation, even at half the budget. The same policy trained on 30 min of robot teleoperation (ACT [18]) reaches 51.2%; our policy with only 15 min of human video already reaches 75.0%, a 41% relative improvement at half the collection time. Tracing the data-efficiency curve, $\sim$8 min of human video (57.5%) already surpasses 30 min of teleoperation (52.5%)—a $\mathbf{3.75}\times$ reduction in collection effort for equal performance.

Mixing in robot data only hurts. Holding the total budget at 30 min and sweeping the human fraction from 0% to 100%, real-world success rises *monotonically* ($65 \to 72.5 \to 77.5 \to 90 \to 95\%$; App. E). There is no co-training sweet spot: the pure-human policy is the global maximum, and adding even 7.5 min of teleoperation to a 22.5-min human dataset costs 5 pp. At a fixed budget, every minute is best spent on the higher-quality source.

V. Zero-Shot Generalization

A high-quality demonstration set should also yield a policy that travels. Deployed across nine out-of-distribution conditions without any retraining (Fig. 4; App. F), the policy holds 85–91.25% success under changes of background, lighting, viewpoint, and added distractors, and even handles novel object instances unseen in training. It is also *hardware-agnostic*: although all training data is collected with Aria glasses, the policy succeeds whether the deployment camera is a RealSense or a ZED and whether the arm is a Trossen, Franka, or UR10—training and deployment share no hardware in common. This robustness is inherited from the embodiment- and viewpoint-invariant representation, and it means a single set of human demonstrations amortizes across many deployment conditions, further raising the effective value of each minute collected.

VI. Takeaways

Our measurements point to one conclusion: at a fixed budget, the source and quality of demonstrations dominate their count. We offer three adoptable points for building imitation datasets. **(1) Treat the demonstration source as a quality lever**—human egocentric video scores higher than teleoperation on every axis we measured, is faster to collect, and yields strictly better policies. **(2) Measure quality with trajectory-level metrics**—signal-to-noise, smoothness/jerk, idle time, spatial and trajectory coverage, and tracking persistence. In a controlled study that swaps only the upstream hand tracker (App. D), real-world success collapses $95 \to 45 \to 32.5 \to 0\%$, and it is *smoothness and detection rate*—not per-keypoint pose accuracy—that separate the survivors: two trackers with identical 1.4 cm pose error differ by 12.5 pp in success. **(3) Do not assume mixing helps**—adding lower-quality data to a higher-quality set can erode performance, so curation and source choice deserve the attention usually spent on scaling. We hope these give the community concrete, falsifiable notions of demonstration quality to test and refine. More broadly, our results suggest that the cheapest way to improve a manipulation policy may not be a bigger model or more teleoperation, but a better demonstration *source*: a few minutes of human egocentric video, recorded anywhere without a robot in the loop, already carries cleaner, smoother, and more diverse manipulation signal than the teleoperation it replaces—and a policy trained on it transfers zero-shot to hardware the demonstrator never touched.

Acknowledgment

The system summarized here is HumanEgo [17]; full architecture, training, and evaluation details are in that work and in the appendices.

References

[1] ALOHA 2 Team, Jorge Aldaco, Travis Armstrong, Robert Baruch, Jeff Bingham, Sanky Chan, Kenneth Draper, Debidatta Dwibedi, Chelsea Finn, Pete Florence, Spencer Goodrich, Wayne Gramlich, Torr Hage, Alexander Herzog, Jonathan Hoech, Thinh Nguyen, Ian Storz, Baruch Tabanpour, Leila Takayama, Jonathan Tompson, Ayzaan Wahid, Ted Wahrburg, Sichun Xu, Sergey Yaroshenko, Kevin Zakka, and Tony Z. Zhao. ALOHA 2: An enhanced low-cost hardware for bimanual teleoperation. *arXiv preprint arXiv:2405.02292*, 2024.

[2] Homanga Bharadhwaj, Roozbeh Mottaghi, Abhinav Gupta, and Shubham Tulsiani. Track2Act: Predicting point tracks from internet videos enables generalizable robot manipulation. In *European Conference on Computer Vision (ECCV)*, 2024.

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

[4] Siddhant Haldar and Lerrel Pinto. Point policy: Unifying observations and actions with key points for robot manipulation. In *Conference on Robot Learning (CoRL)*, 2025.

[5] Cheng-Chun Hsu, Bowen Wen, Jie Xu, Yashraj Narang, Xiaolong Wang, Yuke Zhu, Joydeep Biswas, and Stan Birchfield. SPOT: SE(3) pose trajectory diffusion for object-centric manipulation. *arXiv preprint arXiv:2411.00965*, 2024.

[6] Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. CoTracker: It is better to track together. In *European Conference on Computer Vision (ECCV)*, 2024.

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

[8] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations (ICLR)*, 2023.

[9] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In *European Conference on Computer Vision (ECCV)*, 2024.

[10] Vincent Liu, Ademi Adeniji, Haotian Zhan, Siddhant Haldar, Raunaq Bhirangi, Pieter Abbeel, and Lerrel Pinto. EgoZero: Robot learning from smart glasses. *arXiv preprint arXiv:2505.20290*, 2025.

[11] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. MediaPipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*,

2019.

[12] Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3D with transformers. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[13] Rolandos Alexandros Potamias, Jinglei Zhang, Jiankang Deng, and Stefanos Zafeiriou. WiLoR: End-to-end 3d hand localization and reconstruction in-the-wild. *arXiv preprint arXiv:2409.12259*, 2024.

[14] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

[15] Junyao Shi, Zhuolun Zhao, Tianyou Wang, Ian Pedroza, Amy Luo, Jie Wang, Jason Ma, and Dinesh Jayaraman. ZeroMimic: Distilling robotic manipulation skills from web videos. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2025.

[16] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with Fourier convolutions. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2022.

[17] Zhi Wang, Botao He, Kelin Yu, Seungjae Lee, Ruohan Gao, Furong Huang, and Yiannis Aloimonos. HumanEgo: Zero-shot robot learning from minutes of human egocentric videos. *arXiv preprint*, 2026.

[18] Tony Z Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. In *Robotics: Science and Systems (RSS)*, 2023.

[19] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

APPENDIX A
METHOD DETAILS

A. Visual Observation Transform

We undistort each egocentric frame, segment the hand and arm with SAM2, and remove them with LaMa inpainting [16]. We then render a virtual parallel-jaw gripper and the tracked object keypoints—both derived from the spatial representation below—into the inpainted image, implicitly encoding 6D pose as visual cues while eliminating the visual embodiment gap, without domain adaptation or image translation.

B. Interaction-Centric Tokens

We treat every object and both hands as an entity and encode each entity's 6-DoF pose into a token. For entity k, $\mathrm{ICT}_k \in \mathbb{R}^{29}$ is

$$\mathrm{ICT}_k = [\underbrace{\tau}_{1} \parallel \underbrace{{}^{\mathrm{REF}}T_E}_{9} \parallel \underbrace{{}^{E}T_{LH}}_{9} \parallel \underbrace{{}^{E}T_{RH}}_{9} \parallel \underbrace{g}_{1}], \tag{1}$$

where τ is the entity type, ${}^{\mathrm{REF}}T_E$ is the entity's pose in a shared reference frame, ${}^{E}T_{LH}$ and ${}^{E}T_{RH}$ are the two hands' poses in the entity's local frame E, and g is the grasp state. Each $\mathrm{SE}(3)$ transform is flattened to a normalized translation plus a 6D rotation [19]. Because every pose is relative to scene entities rather than the camera, the tokens are identical regardless of viewpoint, which is what enables direct human-to-robot transfer.

C. Hand-to-Gripper Retargeting

We retarget the 21-keypoint Aria MPS hand into a virtual gripper using five stable keypoints (wrist, thumb/index MCP and tip) after a short motion-optimization pass (confidence masking, gap interpolation, Savitzky–Golay position smoothing, EMA orientation smoothing with sign-consistency to prevent flips). The gripper position is the thumb–index fingertip midpoint; the orientation frame is built from the MCP joints rather than the fingertips, which stay well-separated through the pinch and avoid the degeneracy that arises when the fingertips converge at grasp. The 1-DoF grasp is the normalized, clipped thumb–index distance, median-filtered and binarized at deployment.

D. Object Triangulation

Aria Gen1 glasses lack a depth sensor, so we triangulate tracked 2D keypoints across a short pre-episode *scene sweep*, treating the moving head camera as a multi-view system with extrinsics from the Aria SLAM pose. For each object we detect (Grounding DINO [9]), segment (SAM2 [14]), sample N contour keypoints, track them (CoTracker3 [6]), and solve the linear triangulation system per point via SVD; the object position is the centroid, which cancels per-point noise.

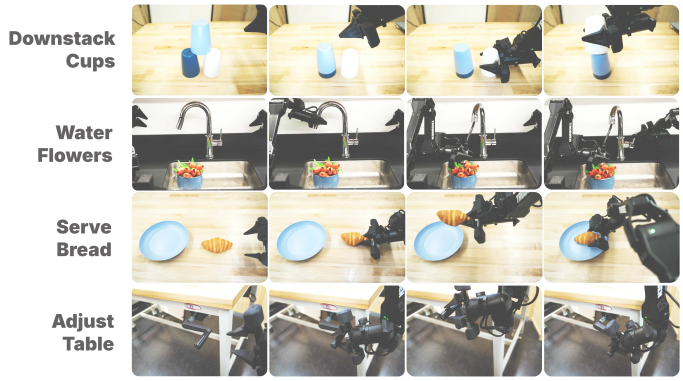

Fig. 5: Four real-world evaluation tasks.

E. Flow Matching Policy and Training

The velocity field is a 6-layer, 8-head transformer decoder (width 384) conditioned on an RGB patch embedding and the per-entity ICT tokens, trained with conditional flow matching [8] (position/rotation/grasp weights $5/1/10$). Three auxiliary heads share the encoder—object-dynamics (future 6-DoF object pose), 2D visual-foresight (future image projections), and temporal-consistency (hand tokens K steps ahead)—each target produced automatically by the perception pipeline, so every demonstration yields a dense multi-task signal. We train with AdamW (10^{-4}, cosine decay, batch 32, 400 epochs, EMA weights) plus image, action-target, and temporal augmentations; at test time we integrate the ODE with a 20-step Euler solver and run the controller at 10 Hz with action chunking.

APPENDIX B
TASKS AND DATA COLLECTION

We evaluate on four tasks spanning pick-and-place (*Serve Bread*), long-horizon multi-step manipulation (*Downstack Cups*: topple, grasp, and cover three nested cups with $\sim$1 cm tolerance), contact-rich bimanual coordination (*Water Flowers*: one arm holds a spray head over a pot while the other opens the valve), and sustained rotational control (*Adjust Table*: turn a crank three full revolutions). Human demonstrations are recorded with Project Aria Gen1 glasses [3] (RGB 30 fps/2 MP, stereo monochrome SLAM, IMUs); Aria's Machine Perception Services provide a globally consistent 6-DoF SLAM trajectory and calibrated 3D hand tracking. Robot demonstrations are teleoperated on two Trossen WidowX arms. We report success over 40 trials per task with randomized object positions, table heights, and initial poses. A single human demonstration takes $\sim$30–40 s versus $\sim$60–70 s for a matched teleoperated demonstration.

APPENDIX C
DEMONSTRATION-QUALITY METRICS

The offline quality axes in Fig. 2 are operationalized as follows. *Signal-to-noise ratio* contrasts manipulation-relevant motion against high-frequency tracking/teleoperation noise on the end-effector trajectory. *Smoothness* is the mean per-frame

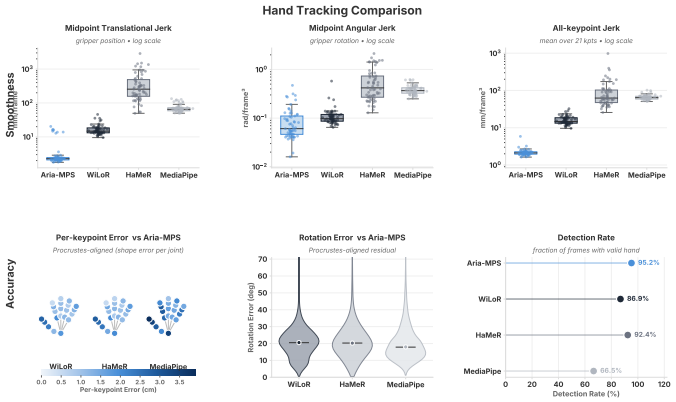

Fig. 6: Why the trackers differ. *Top—smoothness:* per-frame jerk of the gripper midpoint and of all 21 keypoints (log scale, lower is better). *Bottom—accuracy vs. stereo Aria-MPS:* per-keypoint shape error after Procrustes alignment, residual rotation error, and the fraction of frames with a valid detection.

jerk (third derivative) of the gripper midpoint, in translation and rotation; lower is smoother. *Idle time* is the fraction of frames whose end-effector speed falls below a small motion threshold, capturing pauses and piloting overhead. *Spatial coverage* measures how much of the workspace the end-effector positions occupy, and *trajectory diversity* the spread among demonstration trajectories for the same task. These are deliberately simple, source-agnostic statistics computed identically on human and robot demonstrations, so the comparison reflects the data rather than any modality-specific bookkeeping—a starting catalogue of demonstration-quality metrics the community can standardize and extend.

Appendix D
Hand-Tracking Study: Detailed Breakdown

ICT consumes 3D hand keypoints, so the upstream tracker is a direct knob on demonstration quality. Holding the 45 demonstrations, architecture, and recipe fixed and varying only the tracker, real-world success on Serve Bread collapses $95 \rightarrow 45 \rightarrow 32.5 \rightarrow 0\,\%$ from stereo Aria-MPS [3] to monocular WiLoR [13], HaMeR [12], and MediaPipe [11] (Fig. 6). Four observations. **(1) Stereo depth is decisive**: monocular trackers are scale-ambiguous and inject a 5–11 cm depth offset that propagates into every downstream coordinate. **(2) Smoothness and persistence, not pose accuracy, separate the survivors**: after Procrustes alignment HaMeR and WiLoR have nearly identical per-keypoint error (1.4 cm), yet WiLoR (45%) beats HaMeR (32.5%) because its gripper-midpoint jerk is an order of magnitude lower and its detection rate is 86.9% vs. MediaPipe's 66.5%. **(3) The MANO prior is double-edged**: HaMeR and WiLoR inherit identical canonical bone proportions, so their shape errors are correlated by construction—exactly why pose accuracy alone fails to predict success. **(4) MediaPipe fails entirely** (0%): its root-relative output, lifted by camera depth, collapses hand thickness to $\sim$7 cm and flattens the pose information.

Appendix E
Co-Training Study: Full Breakdown

Holding the total budget at 30 min and sampling each batch at the target human/robot ratio, real-world success on Serve Bread is $65 \rightarrow 72.5 \rightarrow 77.5 \rightarrow 90 \rightarrow 95\,\%$ for human fractions of $0/25/50/75/100\,\%$. Even a small slice of human data dominates: replacing just 25% of teleoperation with human video lifts success $+7.5$ pp *even though* the absolute robot data drops from 30 to 22.5 min, so the marginal human minute carries more signal than the marginal robot minute. And there is no sweet spot: the curve only rises and the pure-human policy is the global maximum. We attribute this to the per-minute quality gap of Sec. III—lower-quality minutes dilute higher-quality ones, so at a fixed budget the policy is best served by spending every minute on the higher-quality source.

Appendix F
Generalization Conditions

The zero-shot study in Sec. V spans nine out-of-distribution conditions evaluated without any retraining or fine-tuning (40 trials each): novel *background*, *lighting*, *camera viewpoint*, added *distractors*, novel *object instances*, novel object *placements* and *heights*, a different *camera* (RealSense $\rightarrow$ ZED), and a different *robot arm* (Trossen $\rightarrow$ Franka/UR10). The policy holds 85–91.25% across the visual and placement conditions and transfers across the hardware changes despite sharing no hardware with the training setup, because the interaction-centric representation is invariant to camera placement and embodiment.