MobileEgo Anywhere: Open Infrastructure for long horizon egocentric data on commodity hardware

Senthil Palanisamy*, Abhishek Anand*, Satpal Singh Rathore*,
Pratyush Patnaik*, Shubhanshu Khatana*, Ekaksh Janweja*
FPV Labs
Emails: {abhishek, satpal, pratyush, shubhanshu}@fpvlabs.ai

***Abstract*—Vision-language-action (VLA) models have driven demand for large-scale egocentric datasets, yet the hardware and infrastructure to collect long-horizon data remain inaccessible. Datasets today typically have episodes only a few minutes long, which fails to capture the long-horizon temporal dependencies that complex robotic task execution requires. We present MobileEgo Anywhere, a framework for collecting hour-plus egocentric trajectories on commodity mobile hardware that uses modern smartphone sensors for long-term pose tracking without the hardware barriers of traditional robotics data collection. We release three components: (1) STERA, an open-source video-processing pipeline that converts raw mobile captures into standardized, training-ready formats for VLA and foundation-model research; (2) a free mobile app that lets any user record egocentric activity; and (3) a 200-hour dataset of diverse, long-form egocentric data with persistent state tracking across 584 sessions. We further show this data is a usable training signal: mid-training a VLA on it lowers held-out action-prediction error.**
***Index Terms*—VLA training dataset, egocentric dataset, robotics, commoditized VLA data collection, long-horizon tracking.**

I. Introduction

Vision-language-action (VLA) models have rapidly become a leading approach to general-purpose robot policies. Zheng et al. [1] establish a log-linear scaling law $L = 0.024 - 0.003 \times \ln(D)$ between the volume of egocentric human data D (in hours) and validation loss L, implying that further gains require roughly an order-of-magnitude more data across more environments.

VLA training draws on a hierarchy of data sources. Internet video offers broad semantic coverage but no contact dynamics. Simulation scales cheaply, yet policies trained in it still face a sim-to-real gap. Egocentric human video and interfaces such as UMI [2] provide richer interaction signal and are a primary source for large-scale pretraining, which needs a large, varied corpus spanning many environments and long-horizon tasks.

Existing egocentric datasets share two limits: short episodes and high collection-hardware barriers. MobileEgo Anywhere addresses both by using the visual-inertial odometry (VIO) already built into modern smartphones, specifically ARKit on the iPhone Pro, for 6 DoF pose tracking with no specialized peripherals. We release a free capture app, an open-source processing suite (STERA), and a 200-hour dataset of household activities with continuous episodes up to 108 minutes, 3D hand trajectories, and three-level hierarchical language annotations. We validate the pipeline along three axes: ARKit pose accuracy against motion-capture ground truth, ground-truth-free hand-pose consistency, and hierarchical label quality, then show the data serves as a usable training signal for a vision-language-action model.

*All authors contributed equally to this work.

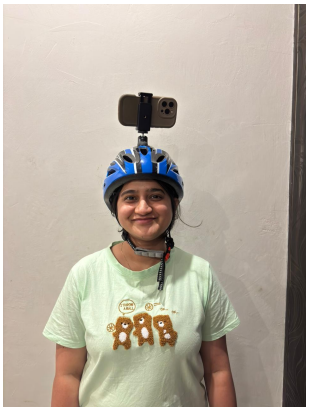
II. Related Work

II-A. Egocentric Datasets for Robotics

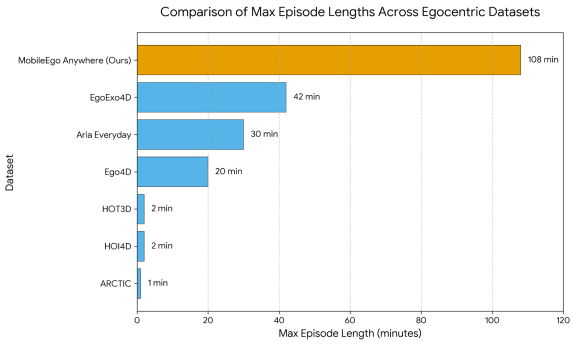
Early egocentric datasets primarily focused on action recognition and localized human-object interactions. Large-scale efforts such as Ego4D [3] and EPIC-KITCHENS [4], [5] provided the community with thousands of hours of video, but these were largely passive and lacked the precise, continuous 6 DoF pose tracking required for robotic policy learning. Recent shifts toward Foundation Models and VLA architectures [18] [19] have increased the demand for actionable egocentric data. Projects like EgoScale [1] provide precise poses but episodes are very short. EgoExo4D [6] adds pose, depth, and hand annotations but relies on Meta's Project Aria glasses and synchronized exo cameras, hardware that is not commercially available. HOI4D [7], ARCTIC [9], and HOT3D [8] provide precise hand tracking but cover only seconds to minutes in controlled lab settings. EgoDex [14] offers 829 hours of dexterous manipulation data captured on Apple Vision Pro, but with $\sim$9 s demonstrations. We instead collect long-horizon trajectories that stay spatially consistent across hour-plus sessions, using only consumer hardware.

II-B. Scalable Data Collection Interfaces

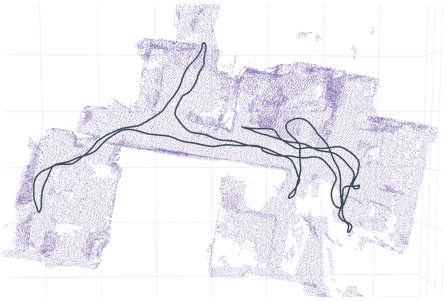
Teleoperation and kinesthetic teaching yield high-quality samples but scale poorly, since each demonstration needs an operator and a robot. Open X-Embodiment [24] and DROID [25] assemble large teleoperation datasets. The Universal Manipulation Interface (UMI) [2] utilizes handheld grippers to bridge the gap between human demonstration and robotic execution, lowering the hardware barrier, but still requires specialized physical mounts and calibrated setups. Our approach instead leverages the commodity smartphone as a universal sensor suite, enabling collection anywhere without additional peripherals.



(a) MobileEgo Anywhere recording setup.



(b) Comparison of episode duration.



(c) Long-horizon trajectory tracked from ARKit.

Fig. 1: **MobileEgo Anywhere** turns any modern iPhone into a long-horizon egocentric capture device. (a) Contributors record hands-free using a helmet-mounted phone. (b) Episodes are substantially longer than those in prior datasets. (c) ARKit-based visual-inertial fusion yields continuous 6 DoF pose, used to generate 3D hand trajectories in a consistent world frame across the full session.

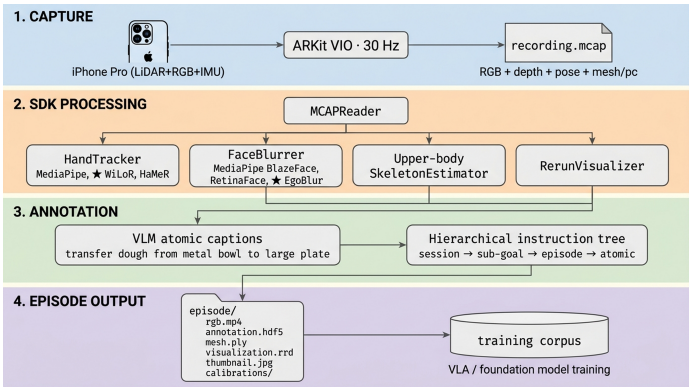


Fig. 2: Overall data flow: raw mobile capture (RGB-D, IMU, ARKit pose) is logged in MCAP format, then processed offline into 3D hand trajectories, atomic action labels, and a hierarchical instruction tree.

II-C. Long-Term Egocentric SLAM and State Estimation

Maintaining stable state tracking over long sessions is the central difficulty for SLAM in this setting. Structure-from-Motion pipelines such as COLMAP [22] become computationally intractable on hour-long trajectories, while feature-based SLAM such as ORB-SLAM3 [23] accumulates drift in dynamic or texture-poor indoor scenes. Recent mobile AR frameworks, notably ARKit and ARCore, address this by integrating high-frequency IMU data with visual keyframes, enabling robust long-term tracking on edge devices.

III. System Overview

Our hardware configuration uses a LiDAR-enabled iPhone Pro mounted on a head-worn rig for a first-person perspective of the participant's hands and workspace. ARKit captures synchronized RGBD streams providing 6 DoF camera poses and per-frame depth maps; a dedicated mobile app exports raw sensor data (RGBD frames, IMU readings, camera intrinsics) in MCAP format [13]. The offline STERA pipeline then transforms these logs into 3D hand trajectories, atomic action labels, and hierarchical task instructions.[1];

III-A. Capture Methodology

Contributors secure the iPhone to a head-worn mount positioned for a consistent egocentric field of view; while a standard helmet mount was used here, any mounting hardware providing sufficient elevation is compatible. Data collection is managed via an integrated voice command interface ("start"/"stop" triggers), ensuring hands-free operation critical for naturalistic activities. During recording, ARKit performs real-time sensor fusion, generating 6 DoF camera poses by synchronizing the onboard IMU with the RGBD stream. The application archives RGB frames, depth maps, and IMU metadata registered to a common high-resolution timestamp, ensuring temporal consistency across all modalities for downstream 3D reconstruction and action recognition.

III-B. 3D Hand Trajectory Estimation

High-fidelity 3D hand trajectories map human motion to robot end-effector frames via IK for VLA training. We employ WiLoR [11] with MANO parameterization [12] to estimate hand joints under biomechanical constraints, handling the partial occlusions common in first-person manipulation more reliably than alternatives such as MediaPipe [21]. WiLoR's relative 3D coordinates are localized into a global frame by sampling ARKit depth maps at detected joint locations and

[1]Project resources: (1) Mobile App: https://apps.apple.com/in/app/stera-by-fpv-labs/id6756263398; (2) Python Processing Suite: https://github.com/fpv-labs/stera-sdk; (3) Huggingface Dataset: https://huggingface.co/datasets/fpvlabs/stera-10m; (4) Data Visualization: https://platform.fpvlabs.ai/dataset/stera-10m/viz

applying the extrinsic camera transformation, yielding world-anchored trajectories for imitation learning.

III-C. Atomic Action Labels

Action-conditioned VLA policies require language labels specifying *which* object is manipulated, *what* the action is, and *where* the object moves, details that generic labels like "pick up object" do not provide. To produce labels at this level of specificity across 200 hours of video, we employ an automated annotation pipeline. Raw video is partitioned into contiguous, non-overlapping temporal spans, and each span is processed by a VLM that outputs a short imperative sentence constrained to include object modifiers (color, material, size) and spatial prepositions wherever the video evidence supports them (e.g., "transfer dough from metal bowl to large plate").

Validation against 50 human-annotated sessions shows automated labels average 7.95 words versus 2.94 and 1.09 descriptive modifiers per label versus 0.09. The automated pipeline produced zero temporal defects across all 5,249 labels. Human annotations contained 63 segments with durations $\leq$0 s and 877 overlapping consecutive pairs (9.9% of 8,821 adjacent pairs), defects that would propagate as corrupted training samples.

III-D. Hierarchical Task Instructions

Long-horizon sessions contain dozens of atomic labels belonging to distinct sub-tasks. To expose this structure, atomic span captions are organized into a three-level instruction tree: a session-level goal, sub-goals, and episodes. Hierarchical instruction setups like these have been well studied in works like [16] and [17]. A language model groups temporally contiguous spans into episodes, clusters related episodes into sub-goals, and synthesizes a session-level goal grounded in the concrete objects across all spans.

Three invariants are enforced: unique span assignment, exact timestamp boundaries, and full session coverage with no gaps. Of seven language models evaluated, six produced fully valid outputs, providing language conditioning from 5-second manipulation steps to full session plans, matching the multi-scale supervision used by recent hierarchical VLA architectures.

IV. Data Quality Validation

The released dataset contains 584 sessions totaling 200 hours from 20 contributors, averaging 20.5 minutes with a maximum of 108 minutes. Table I shows MobileEgo Anywhere is the only dataset combining consumer hardware with continuous 6 DoF pose, LiDAR depth, MANO annotations, and sessions exceeding one hour; EgoExo4D offers similar modalities but requires commercially unavailable hardware.

IV-A. ARKit Pose Accuracy

IV-A1. Motion Capture Ground Truth Comparison

To quantify the absolute accuracy of ARKit pose estimates, we collected trajectories spanning representative motion profiles and evaluated them against ground truth from a 30-camera Vicon motion capture system. Tracked trajectories include two naturalistic egocentric sequences in which a participant performs everyday household tasks, a slow walk, a closed-loop traversal, and a spinning sequence. Absolute Trajectory Error (ATE RMSE), relative ATE, and Relative Pose Error (RPE) [29] for translation are reported in Table II.

Table II reports per-sequence ATE RMSE, relative ATE, and translational and rotational RPE against the Vicon ground truth. Relative ATE stays below 1% for nine of the ten sequences and rotational RPE below 4°; the lone exception is the short spinning sequence, whose elevated values follow from its rapid rotational motion. Translational RPE remains below 5 cm throughout, so local pose consistency holds even where global drift is marginally elevated. ARKit visual-inertial odometry on a consumer iPhone Pro thus provides trajectory accuracy sufficient for the world-frame hand-pose anchoring of Section III-B.

IV-A2. Long-Term Drift Evaluation

Because ARKit is closed-source, we assess long-term drift by placing an ArUco marker [30] at session start and revisiting it at roughly the temporal midpoint and session end. We repeat this across six environments (Table III); drift is below 1 cm in all but the whole-house traversal (1.5 cm end-of-session) and below 0.1% of trajectory length in all cases, demonstrating the efficacy of ARKit tracking for downstream VLA applications.

V. Training Signal Validation

To test whether MobileEgo Anywhere data is a usable training signal for action models, we mid-train the VITRA vision–language–action framework [15] on our egocentric trajectories and evaluate on a held-out split of 27 sessions disjoint from training. We initialize from VITRA-VLA-3B, freeze the vision encoder, condition on hierarchical instructions, and supervise MANO hand pose prediction for 10,000 steps. Following VITRA, we predict directly in MANO hand-parameter space (per-hand wrist 6-DoF and finger joint angles), which corresponds to anthropomorphic robot hands and avoids learning a human-to-robot retargeting during pretraining.

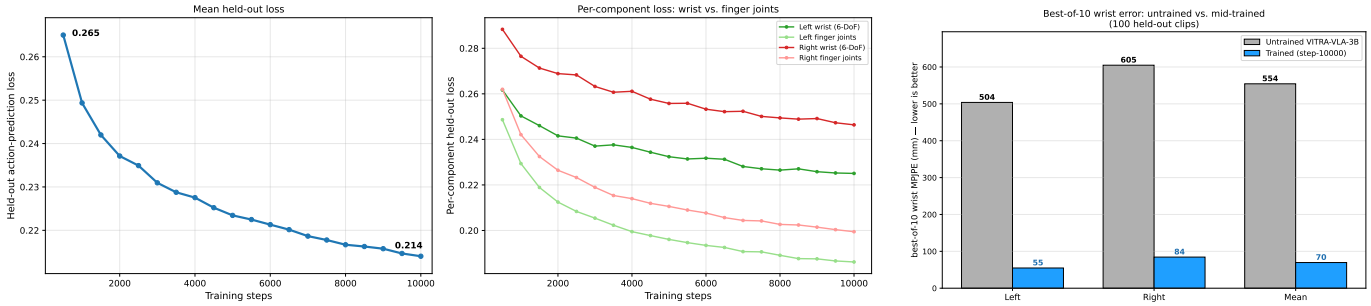
Held-out action-prediction loss decreases monotonically from 0.265 to 0.214 (-19%; Fig. 3a). Because the metric is computed on sessions never seen during training, this drop reflects generalization rather than memorization, and it is concentrated in finger articulation (-24% per hand) over the wrist root (-14%), indicating that the model learns the fine, dexterity-relevant hand structure our labels carry. The untrained VITRA-VLA-3B does not transfer to our data out of the box: its predictions are mis-scaled, giving a best-of-10 wrist error of 554 mm, whereas mid-training calibrates it to our distribution (70 mm; Fig. 3b) and reduces held-out loss from 1.00 to 0.21. These results establish open-loop trainability rather than policy-level accuracy: we evaluate hand-action prediction on held-out human sessions and perform no closed-loop robot rollout.

VI. Limitations

Platform dependency: The capture pipeline currently requires an iPhone Pro, as STERA relies on ARKit's visual-

TABLE I: Comparison of egocentric datasets for robot-relevant pretraining. Hours and durations are taken from original publications or computed directly from reported totals.

Dataset	Hours	Max Episode	6 DoF Pose	Depth	Hand Annot.	Capture Hardware
Ego4D [3]	3,670	up to ∼7 hrs	Partial	No	Partial	Mixed (GoPro, ZShades, Aria)
EPIC-KITCHENS-100 [5]	100	∼8.6 min avg	No	No	Partial	Head-mounted GoPro
EgoExo4D [6]	1,286	∼42 min	Yes	Yes	Yes	Aria + exo cameras
HOI4D [7]	∼22.2	∼20 sec / clip	Yes	Yes	Yes	Intel RealSense
HOT3D [8]	∼13.9	∼2 min / recording	Yes	No	Yes	Aria + Quest 3
ARCTIC [9]	∼2.1	∼23 sec avg	Yes	No	Yes	MoCap rig
Aria Everyday [10]	∼7.3	∼3 min avg	Yes	No	Yes	Project Aria
EgoDex [14]	829	∼9 sec / demo	Yes	No	Yes	Apple Vision Pro
MobileEgo Anywhere (ours)	**200**	**108 min**	**Yes**	**Yes**	**Yes**	**Consumer iPhone**



(a) Held-out action-prediction loss over 10,000 mid-training steps (−19%); the reduction is driven by finger articulation rather than the wrist root.

(b) Best-of-10 wrist error on held-out clips: the untrained VITRA-VLA-3B is mis-scaled on our data (554 mm); mid-training calibrates it to 70 mm.

Fig. 3: Downstream VLA mid-training on MobileEgo Anywhere using the VITRA framework [15].

TABLE II: ARKit accuracy vs. Vicon ground truth

Sequence No.	Duration	Traj.	ATE RMSE	Rel. ATE	RPE trans	RPE rot
1	149.7 s	107.3 m	15.0 cm	0.14%	4.99 cm	3.4033°
2	134.5 s	41.6 m	9.1 cm	0.22%	2.24 cm	0.8529°
3	134.4 s	51.6 m	10.1 cm	0.20%	2.59 cm	0.5103°
4	47.1 s	6.9 m	10.1 cm	1.47%	2.04 cm	4.1728°
5	91.5 s	28.1 m	13.0 cm	0.46%	4.03 cm	3.6004°
6	120.7 s	0.03 m	0.01 cm	0.48%	0.01 cm	0.0468°
7	123.4 s	0.04 m	0.04 cm	0.98%	0.01 cm	0.0644°
8	163.7 s	40.1 m	10.1 cm	0.25%	3.27 cm	1.6832°
9	145.8 s	37.2 m	11.9 cm	0.32%	2.75 cm	1.6550°
10	56.1 s	32.5 m	12.6 cm	0.39%	1.76 cm	1.6650°

TABLE III: ARKit long-term drift evaluation via ArUco marker revisits.

Environment	Mid-session	End-of-session
Kitchen activity	0.4 cm	0.7 cm
Living-space activity	0.3 cm	0.4 cm
Whole-house activity	1.0 cm	1.5 cm
Whole-house walk-through	0.3 cm	0.2 cm
Cloth folding	0.1 cm	0.1 cm
Nail polish	0.1 cm	0.1 cm

inertial odometry and LiDAR depth sensing. Android/ARCore offers lower VIO accuracy and lacks LiDAR depth; supporting depth-free fallback modes would meaningfully broaden contributor reach. Additionally, the ultrawide lens is inaccessible during active ARKit sessions, limiting field of view for wide-workspace activities.

Thermal constraints on session length Continuous recording beyond approximately two hours can trigger thermal throttling on iPhone Pro hardware; a heat sink attachment mitigates this but adds deployment friction in warm environments.

Preliminary downstream validation: Our training-signal experiment is open-loop (we evaluate held-out hand-action prediction and perform no closed-loop robot rollout), so it demonstrates trainability rather than policy-level task success, which we leave to future work.

VII. Conclusion

We have presented MobileEgo Anywhere, an accessible and commoditized framework for large-scale, long-horizon egocentric data collection. By releasing a free mobile application and open-sourcing the STERA processing pipeline, we enable researchers and contributors worldwide to generate VLA-ready datasets using standard consumer hardware. The 200-hour dataset features continuous episodes up to 108 minutes, 6 DoF ARKit poses, LiDAR depth, MANO 3D hand trajectories in a consistent world frame, and three-level hierarchical language annotations. This lowers the barrier to creating VLA-ready egocentric datasets and supports work toward more generalizable robot policies.

References

[1] R. Zheng, D. Niu, Y. Xie, J. Wang, M. Xu, Y. Jiang, F. Castañeda, F. Hu, Y. L. Tan, L. Fu, T. Darrell, F. Huang, Y. Zhu, D. Xu, and L. Fan,

"EgoScale: Scaling Dexterous Manipulation with Diverse Egocentric Human Data," *arXiv preprint arXiv:2602.16710*, 2026.

[2] C. Chi *et al.*, "Universal Manipulation Interface: In-The-Wild Robot Teaching Without In-The-Wild Robots," in *Proc. Robotics: Science and Systems (RSS)*, 2024.

[3] K. Grauman *et al.*, "Ego4D: Around the World in 3,000 Hours of Egocentric Video," in *Proc. IEEE/CVF CVPR*, 2022, pp. 18973–18990.

[4] D. Damen *et al.*, "Scaling Egocentric Vision: The EPIC-KITCHENS Dataset," in *Proc. ECCV*, 2018, pp. 753–771.

[5] D. Damen *et al.*, "Rescaling Egocentric Vision: Collection, Pipeline and Challenges for EPIC-KITCHENS-100," *Int. J. Comput. Vis.*, vol. 130, no. 1, pp. 33–55, 2022.

[6] K. Grauman *et al.*, "Ego-Exo4D: Understanding Skilled Human Activity from First- and Third-Person Perspectives," in *Proc. IEEE/CVF CVPR*, 2024, pp. 19383–19400.

[7] Y. Liu *et al.*, "HOI4D: A 4D Egocentric Dataset for Category-Level Human-Object Interaction," in *Proc. IEEE/CVF CVPR*, 2022, pp. 21013–21022.

[8] S. Banerjee *et al.*, "HOT3D: Hand and Object Tracking in 3D from Egocentric Multi-View Videos," in *Proc. IEEE/CVF CVPR*, 2025.

[9] Z. Fan *et al.*, "ARCTIC: A Dataset for Dexterous Bimanual Hand-Object Manipulation," in *Proc. IEEE/CVF CVPR*, 2023, pp. 12943–12954.

[10] Z. Lv *et al.*, "Aria Everyday Activities Dataset," *arXiv preprint arXiv:2402.13349*, 2024.

[11] R. A. Potamias, J. Zhang, J. Deng, and S. Zafeiriou, "WiLoR: End-to-end 3D Hand Localization and Reconstruction in-the-wild," *arXiv preprint arXiv:2409.12259*, 2024.

[12] J. Romero, D. Tzionas, and M. J. Black, "Embodied Hands: Modeling and Capturing Hands and Bodies Together," *ACM Trans. Graph. (Proc. SIGGRAPH Asia)*, vol. 36, no. 6, pp. 245:1–245:17, 2017.

[13] Foxglove Developers, "MCAP: serialization-agnostic log container file format," *Foxglove Technologies*, 2024. [Online]. Available: https://mcap.dev

[14] R. Hoque, P. Huang, D. J. Yoon, M. Sivapurapu, and J. Zhang, "EgoDex: Learning Dexterous Manipulation from Large-Scale Egocentric Video," *arXiv preprint arXiv:2505.11709*, 2025.

[15] Q. Li, Y. Deng, Y. Liang, L. Luo, L. Zhou, C. Yao, L. Zeng, Z. Feng, H. Liang, S. Xu, Y. Zhang, X. Chen, Hao Chen, L. Sun, D. Chen, J. Yang, and B. Guo. Scalable vision-language-action model pretraining for robotic manipulation with real-life human activity videos. *arXiv preprint arXiv:2510.21571*, 2025.

[16] Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. ALFRED: A Benchmark for Interpreting Grounded Instructions for Everyday Tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10740–10749, 2020.

[17] Alex Irpan, Alexander Herzog, Alexander Toshkov Toshev, Andy Zeng, Anthony Brohan, Brian Andrew Ichter, Byron David, Carolina Parada, Chelsea Finn, Clayton Tan, Diego Reyes, Dmitry Kalashnikov, Eric Victor Jang, Fei Xia, Jarek Liam Rettinghouse, Jasmine Chiehju Hsu, Jornell Lacanlale Quiambao, Julian Ibarz, Kanishka Rao, Karol Hausman, Keerthana Gopalakrishnan, Kuang-Huei Lee, Kyle Alan Jeffrey, Linda Luu, Mengyuan Yan, Michael Soogil Ahn, Nicolas Sievers, Nikhil J Joshi, Noah Brown, Omar Eduardo Escareno Cortes, Peng Xu, Peter Pastor Sampedro, Pierre Sermanet, Rosario Jauregui Ruano, Ryan Christopher Julian, Sally Augusta Jesmonth, Sergey Levine, Steve Xu, Ted Xiao, Vincent Olivier Vanhoucke, Yao Lu, Yevgen Chebotar, and Yuheng Kuang. Do As I Can, Not As I Say: Grounding Language in Robotic Affordances. In *Proc. 6th Conference on Robot Learning (CoRL)*, PMLR vol. 205, pp. 287–318, 2023. https://arxiv.org/abs/2204.01691.

[18] Kevin Qinghong Lin, Alex Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Zhongcong Xu, Difei Gao, Rongcheng Tu, Wenzhe Zhao, Weijie Kong, et al. Egocentric video-language pretraining. *arXiv preprint arXiv:2206.01670*, 2022.

[19] Simar Kareer, Dhruv Patel, Ryan Punamiya, Pranay Mathur, Shuo Cheng, Chen Wang, Judy Hoffman, and Danfei Xu. EgoMimic: Scaling imitation learning via egocentric video. *arXiv preprint arXiv:2410.24221*, 2024.

[20] G. S. Raina, J. Somasundaram, A. Furnari, N. Rewkowski, S.-E. Wei, D. S. Chaplot, and V. K. Ithapu, "EgoBlur: Responsible Innovation in Aria," *arXiv preprint arXiv:2308.13093*, 2023.

[21] Fan Zhang, Valentin Bazarevsky, Andrey Dashkevich, Artsiom Ablavatski, and Matthias Grundmann. MediaPipe Hands: On-device real-time hand tracking. *arXiv preprint arXiv:2006.10214*, 2020.

[22] J. L. Schönberger and J.-M. Frahm, "Structure-from-Motion Revisited," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 4104–4113.

[23] C. Campos, R. Elvira, J. J. Gómez Rodríguez, J. M. M. Montiel, and J. D. Tardós, "ORB-SLAM3: An Accurate Open-Source Library for Visual, Visual-Inertial and Multi-Map SLAM," *IEEE Transactions on Robotics*, vol. 37, no. 6, pp. 1874–1890, 2021.

[24] Open X-Embodiment Collaboration *et al.* Open X-Embodiment: Robotic Learning Datasets and RT-X Models. https://arxiv.org/abs/2310.08864, 2023.

[25] A. Khazatsky, K. Pertsch, S. Nair, A. Balakrishna, S. Dasari, S. Karamcheti, S. Nasiriany, *et al.*, "DROID: A Large-Scale In-The-Wild Robot Manipulation Dataset," in *Proceedings of Robotics: Science and Systems*, Delft, Netherlands, July 2024. doi: 10.15607/RSS.2024.XX.120.

[26] S. Cobos, M. Ferre, M. A. Sanchez-Uran, J. Ortego, and C. Pena, "Efficient human hand kinematics for manipulation tasks," in *Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Systems (IROS)*, 2008, pp. 2246–2251.

[27] M. Alt Murphy, C. Willén, and K. S. Sunnerhagen, "Kinematic Variables Quantifying Upper-Extremity Performance After Stroke During Reaching and Drinking From a Glass," *Neurorehabilitation and Neural Repair*, vol. 25, no. 1, pp. 71–80, 2011.

[28] J. Nakatake, H. Arakawa, T. Tajima, S. Miyazaki, and E. Chosa, "Age- and Sex-Related Differences in Upper-Body Joint and Endpoint Kinematics During a Drinking Task in Healthy Adults," *PeerJ*, vol. 11, p. e16571, 2023.

[29] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers, "A Benchmark for the Evaluation of RGB-D SLAM Systems," in *Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Systems (IROS)*, 2012, pp. 573–580.

[30] S. Garrido-Jurado, R. Muñoz-Salinas, F. J. Madrid-Cuevas, and M. J. Marín-Jiménez, "Automatic Generation and Detection of Highly Reliable Fiducial Markers Under Occlusion," *Pattern Recognition*, vol. 47, no. 6, pp. 2280–2292, 2014.

APPENDIX

VIII. 3D HAND POSE CONSISTENCY

As ground-truth MANO poses are unavailable at our scale, we assess quality across 98 sessions (1.19 M frames, 25.2 hours) via three ground-truth-free metrics: bone length constancy, joint angle plausibility, and wrist dynamics. Hand detection succeeds on 86.2% of frames (mean WiLoR confidence 0.73); 247 frames with zero LiDAR depth returns (0.02%) are discarded via a $z > 0.01$ m threshold before computing metrics.

Bone length constancy. Median coefficient of variation (CV) of the 20 MANO bone lengths is 1.27% (left hand) and 1.43% (right), indicating stability to within roughly 1 mm on a typical 7–8 cm bone (Fig. 5a). The pinky distal phalanx shows elevated CV ($\sim$7.5%) due to its short physical length ($\sim$2 cm), which amplifies relative error from a fixed absolute noise floor; excluding it, pooled median CV drops below 1% for both hands.

Joint angle plausibility. Over 99.99% of the 15 flexion angles (MCP, PIP, and DIP for each finger) fall within published biomechanical limits [26] across all sessions (Fig. 5b), with unimodal distributions consistent with the variety of grasp types present in the dataset.

Wrist dynamics. Median wrist velocity is 0.34 m/s (left) and 0.27 m/s (right), and median acceleration is 2.7 and 1.5 m/s^2 (Fig. 7a). The velocity medians sit below the $\sim$0.62 m/s peak hand velocity reported for healthy adults during the standardized drinking task [27], [28], as expected for continuous recordings that interleave fast reach-and-transport phases with slower in-hand manipulation; the smooth, unimodal distributions show no teleportation-scale discontinuities.

IX. HIERARCHICAL INSTRUCTION QUALITY

We ran the hierarchical decomposition across all 584 sessions using DeepSeek V4 Flash with high reasoning, producing 75,857 atomic spans grouped into 9,922 episodes and 2,212 sub-goals. Three structural invariants are enforced during generation: every atomic span maps to exactly one episode, episode and sub-goal boundaries use exact input timestamps, and each decomposition covers the full session with no gaps.

Fig. 6 shows that each level of the hierarchy occupies a distinct temporal band, with a natural 4–8$\times$ scale separation between adjacent levels: 5 s atomic spans, 39 s episodes, 3.9 min sub-goals, and 16.8 min sessions. This structure emerges from the data rather than being imposed by the prompt. Episode and sub-goal counts scale roughly linearly with session length (Fig. 6), confirming adaptation to session complexity. Most episodes are compact: 80% contain $\leq$10 atomic spans (Fig. 6), with a median of 5 spans per episode. The total LLM cost for hierarchical structuring was only a few dollars, negligible relative to the data-collection effort. A representative decomposition of a 36-minute session is shown in Fig. 4.

X. MOCAP TRAJECTORIES AND RESULTS

The trajectories captured in motion capture settings for validating arkit accuracies, discussed in IV-A1 is shown in Figure 6

XI. ETHICS AND PRIVACY

All contributors signed informed consent covering capture, processing, and public release. Contributors were instructed to avoid recording non-consenting individuals; any faces accidentally captured were blurred in post-processing using [20]

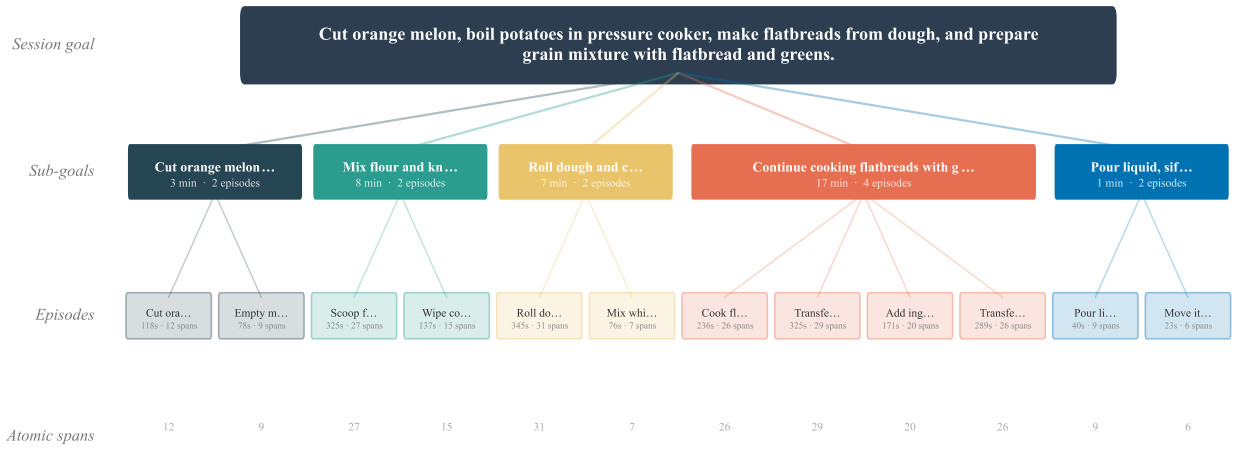
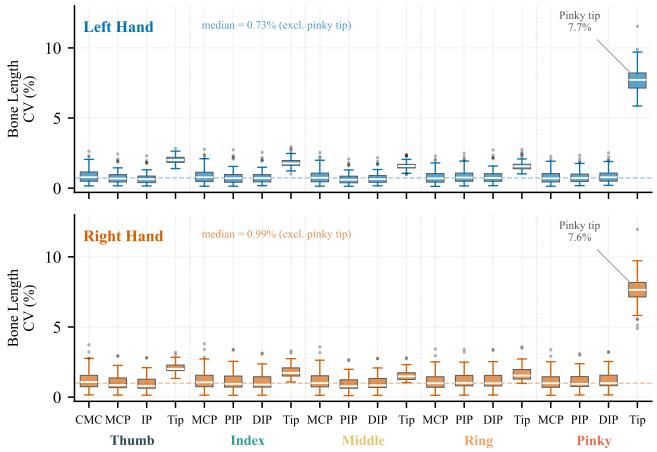
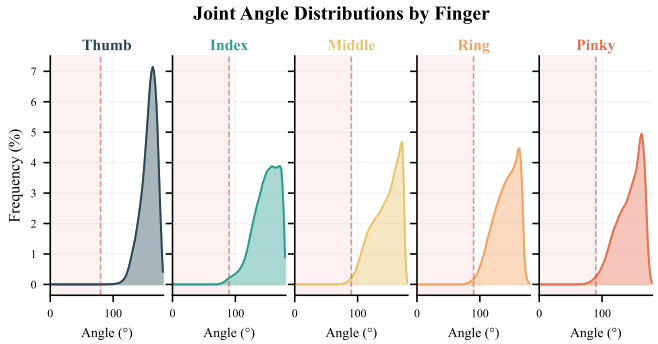


Fig. 4: Hierarchical decomposition of a 36-minute cooking session (217 atomic spans). A single session goal decomposes into five sub-goals, each containing two to four episodes. Sub-goal durations range from 1 to 17 minutes; episode durations from 23 s to 345 s.

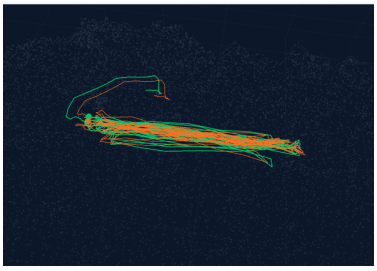


(a) Per-bone Co-efficient of Variation (CV) of bone length over 98 sessions. The pinky distal bone shows elevated CV due to its short physical length (∼2 cm); excluding it, median CV falls below 1.

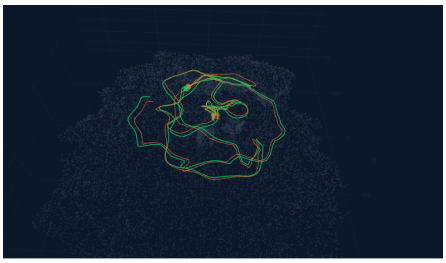


(b) Joint flexion angle distributions pooled over 98 sessions. Shaded regions indicate biomechanical limits; >99.99% of angles fall within bounds.

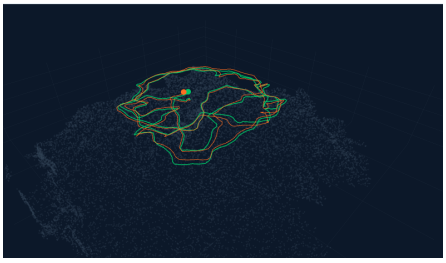
Fig. 5: Kinematic evaluation results.



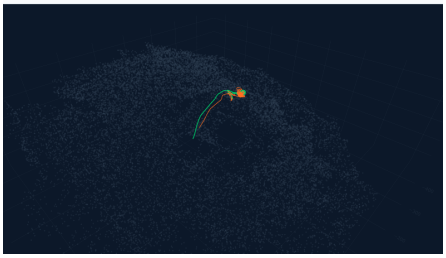
(a) Session 1 - Repeated traversals along a fixed straight-line path between static reference markers at three distinct locomotion speeds.



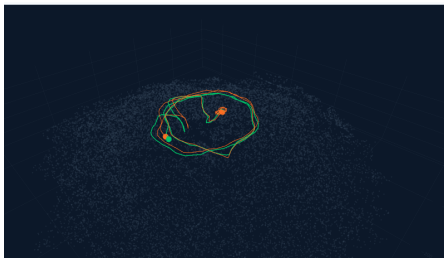
(b) Session 2 - Slow exploration with gentle head rotations, dynamic distractors, seated interaction, and partial occlusions.



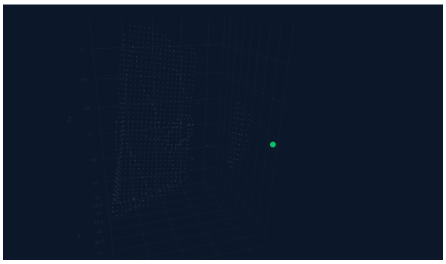
(c) Session 3 - Slow exploration with head rotations and dynamic distractors, without seated interaction, isolating low-speed navigation.



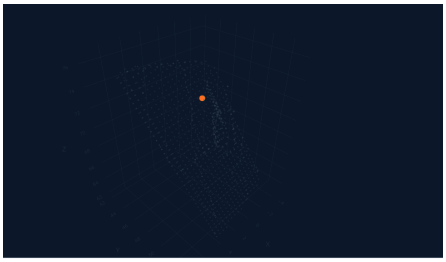
(d) Session 4 - Slow continuous rotation while seated on a swivel chair, producing sustained pure-yaw motion.



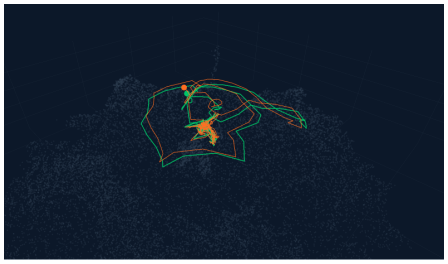
(e) Session 5 - Aggressive rotation while seated on a swivel chair, including rapid full revolutions and high angular velocity.



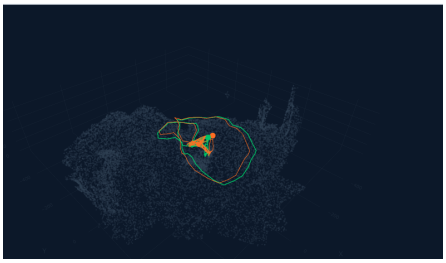
(f) Session 6 - Fully stationary recording with no intentional motion, used to quantify positional and rotational drift at rest.



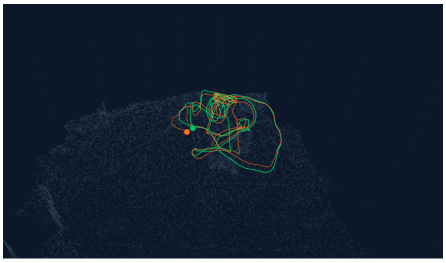
(g) Session 7 - A second fully stationary recording with no intentional motion, providing a repeated measurement of static drift.



(h) Session 8 - Egocentric activity sequence involving seated object interaction (opening bags, handling cameras, folding clothes).



(i) Session 9 - Activity interactions performed at higher speed and with abrupt movements, increasing motion blur.

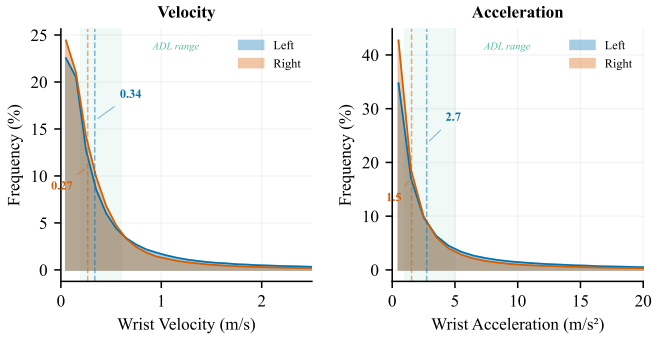


(j) Session 10 - An activity in which the person climbed a stool, adding height variation to the trajectory.

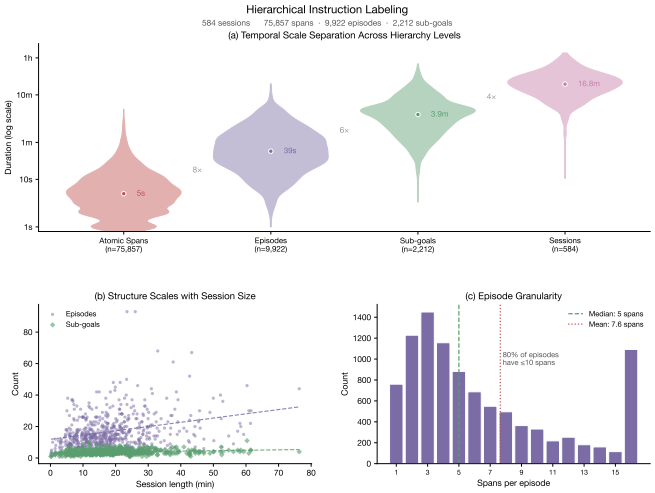
(k) Mocap setup layout.

(l) A contributor performing an egocentric activity while tracked by the motion-capture system.

Fig. 6: Unified overview of Mocap trajectory vs ARKit trajectory across sessions 1-10 (green: mocap, orange: ARKit) alongside the corresponding hardware capture environment and setup layout.



(a) Wrist velocity and acceleration distributions for left and right hands, pooled over 98 sessions; median velocity is 0.34 m/s (left) and 0.27 m/s (right). These session medians fall below the ∼0.6–0.9 m/s peak hand velocities reported for everyday reaching tasks [27], [28].



(b) Hierarchical instruction analysis across 584 sessions (75,857 atomic spans). (1) Temporal scale separation with consistent 4–8× gaps between levels. (2) Linear scaling of episode and sub-goal counts with session length. (3) 80% of episodes contain ≤10 atomic spans (median 5, mean 7.6).

Fig. 7: Overview of experimental data: Left shows wrist dynamics distributions, while right details the hierarchical instruction analysis.

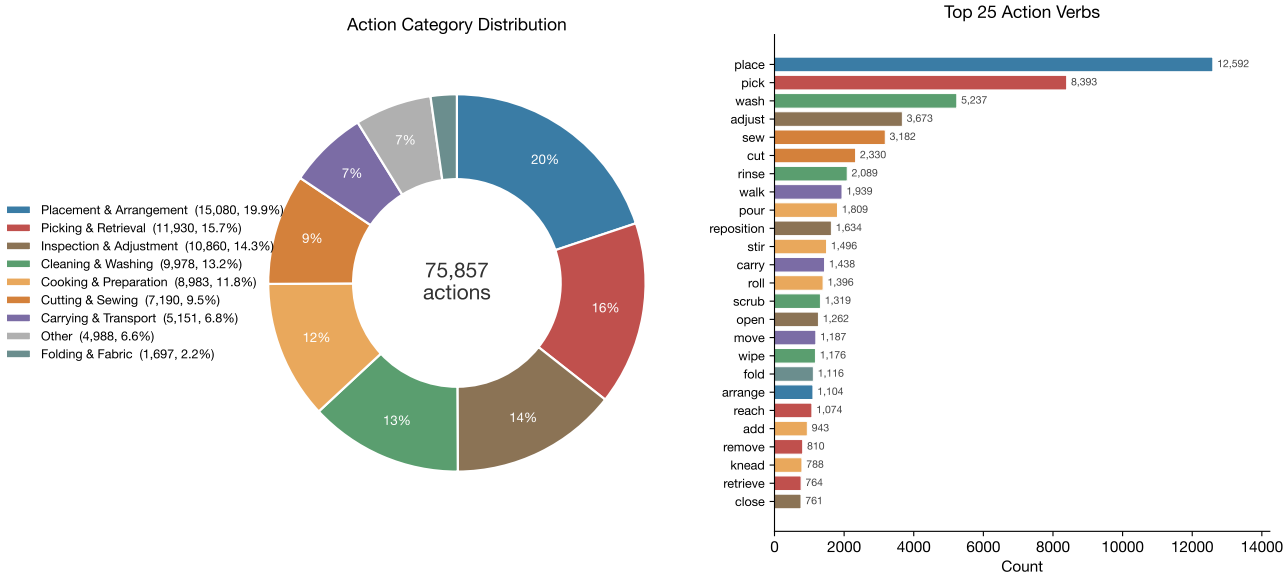


Fig. 8: Task diversity across 584 sessions and 20 contributors. The 75,857 atomic action labels span a long-tailed vocabulary over household manipulation domains (cooking, cleaning, sewing, organizing), grouped into nine high-level categories.