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

Senthil Palanisamy* *FPV Labs* Bangalore, India senthil@fpvlabs.ai
Abhishek Anand* *FPV Labs* Bangalore, India abhishek@fpvlabs.ai
Satpal Singh Rathor* *FPV Labs* Bangalore, India satpal@fpvlabs.ai
Pratyush Patnaik* *FPV Labs* Bangalore, India pratyush@fpvlabs.ai
Shubhanshu Khatana* *FPV Labs* Bangalore, India shubhanshu@fpvlabs.ai

***Abstract*—The recent advancement of Vision Language Action (VLA) models has driven a critical demand for large scale egocentric datasets. However, existing datasets are often limited by short episode durations, typically spanning only a few minutes, which fails to capture the long horizon temporal dependencies necessary for complex robotic task execution. To bridge this gap, we present MobileEgo Anywhere, a framework designed to facilitate the collection of robust, hour plus egocentric trajectories using commodity mobile hardware. We leverage the ubiquitous sensor suites of modern smartphones to provide high fidelity, long term camera pose tracking, effectively removing the high hardware barriers associated with traditional robotics data collection.**

Our contributions are three fold: (1) we release a novel dataset comprising 200 hours of diverse, long form egocentric data with persistent state tracking; (2) we open source a mobile application that enables any user to record egocentric data, and (3) we provide a comprehensive processing pipeline to convert raw mobile captures into standardized, training ready formats for Vision Language Action model and foundation model research. By democratizing the data collection process, this work enables the massive scale acquisition of long horizon data across varied global environments, accelerating the development of generalizable robotic policies.

***Index Terms*—VLA training dataset, egocentric dataset, robotics, commoditized VLA data collection, long horizon tracking.**

I. Introduction

The field of robotics has recently witnessed a paradigm shift driven by the emergence of Vision Language Action (VLA) models. These architectures have demonstrated unprecedented performance across diverse robotic tasks, with scaling laws indicating a robust correlation between model capacity, training data volume, and downstream success. Specifically, Zheng et al. [1] established a log-linear scaling law, $L = 0.024 - 0.003 \times \ln(D)$, where L is the validation loss and D is the dataset scale. This trend suggests that reaching the next frontier of generalizable robotics requires an order of magnitude increase in data diversity and volume beyond current institutional capabilities.The development of robust VLAs relies on a diverse hierarchy of data sources, each presenting a distinct tradeoff between scalability and physical grounding. Passive internet video provides an abundant medium for semantic pretraining but lacks the force profiles and contact dynamics essential for closing the deployment gap. Simulation data offers virtually infinite scaling for rigidbody tasks but remains constrained by the "sim to real" gap, particularly regarding complex fluids and deformable objects. To mitigate the embodiment gap, researchers have pivoted toward egocentric human video and the Universal Manipulation Interface (UMI) [2]. While these provide richer interaction primitives, teleoperation remains the primary methodology for capturing high fidelity motor actions, while on-policy intervention remains an optimal approach for refining edge case behaviors.In this multistage training pipeline, egocentric data serves as the critical foundation for large scale pretraining. To be effective, this stage requires a extensive, heterogenous corpus of data that captures a wide array of environments and long horizon tasks. However, a significant limiting factor persists: existing egocentric datasets are often limited by short episode lengths and high hardware barriers for collection. By maximizing the spatial and temporal reasoning capabilities during pretraining, we can significantly reduce the data requirements for resource intensive downstream fine tuning.

*All authors contributed equally to this work.

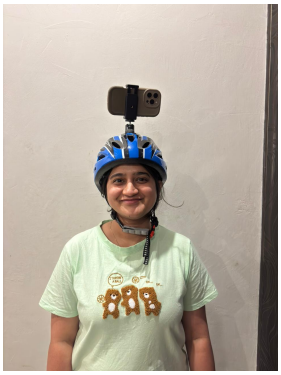
II. Related work

II-A. Egocentric Datasets for Robotics

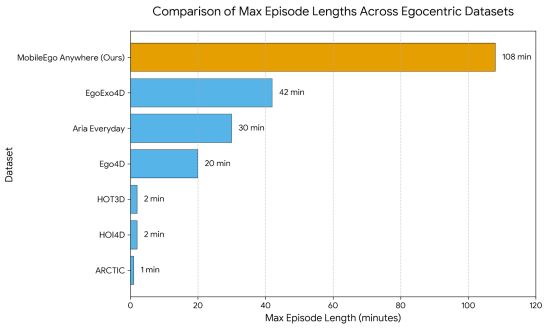
Early egocentric datasets primarily focused on action recognition and localized human object interactions. Large scale efforts such as Ego4D [3] and Epic Kitchens [4] provided the community with thousands of hours of video, but these were largely passive and often lacked the precise, continuous 6 DoF pose tracking required for robotic policy learning. Recent shifts toward Foundation Models and Vision Language Action (VLA) architectures have increased the demand for "actionable" egocentric data. Projects like EgoScale [1] do have precise poses but their episodes are very short. However, these datasets often consist of short, disjointed episodes. Our work extends this lineage by focusing on long horizon trajectories that maintain state consistency over hour plus durations.

II-B. Scalable Data Collection Interfaces

The "bottleneck" of robotics has traditionally been the difficulty of collecting high fidelity interaction data. Teleoperation and kinesthetic teaching provide high quality samples but are notoriously difficult to scale. To address this, researchers introduced the Universal Manipulation Interface (UMI) [2], which utilizes handheld grippers to bridge the gap between



(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, which can later be used to generate 3D hand trajectories in a consistent world frame across the full session.

human demonstration and robotic execution. While UMI effectively lowers the hardware barrier, it still requires specialized physical mounts and calibrated setups. In contrast, our approach leverages the commodity smartphone as a universal sensor suite. By utilizing the mature Visual-Inertial Odometry (VIO) frameworks present in modern mobile devices, we enable ”anywhere” collection without the need for additional mechanical peripherals.

II-C. Long Term egocentric SLAM and State Estimation

Maintaining stable state tracking over extended periods is a classic challenge in Simultaneous Localization and Mapping (SLAM). Traditional visual SLAM pipelines often suffer from cumulative drift, particularly in dynamic or feature poor environments like egocentric slam in indoor environments. Recent advancements in mobile AR frameworks (e.g., ARKit and ARCore) have significantly improved the robustness of long term tracking on edge devices by integrating high frequency IMU data with visual keyframes. MobileEgo Anywhere is positioned at the intersection of mobile SLAM and robotics, providing a pipeline that transforms consumer grade mobile tracking into persistent, high fidelity trajectories suitable for training long horizon VLA models.

III. Overview

We introduce an automated end to end framework for the collection and processing of multimodal egocentric data. Our hardware configuration utilizes a LiDAR enabled iOS devices (iphone Pro) mounted on a headworn rig, positioned to capture a first person perspective of the participant’s hands and the workspace.

During data collection, the mobile device utilizes ARKit to capture synchronized RGBD streams, providing 6 DoF camera poses and per frame depth maps. The collection process is managed via a dedicated mobile application, which records and exports raw sensor data including RGBD frames, high frequency IMU readings, and camera intrinsics into the MCAP format. [13]

For post processing, we provide an open source Python suite that transforms these raw logs into standard datasets. The pipeline automatically generates atomic and hierarchical action labels and performs 3D hand pose estimation. Specifically, 2D keypoints are detected and unprojected into 3D space using ARKit depth data; these are then transformed into a consistent global reference frame using the recorded camera poses. To support the community, we have open sourced the entire software stack and a substantial dataset comprising 200 hours of annotated egocentric activity.[1]

III-A. Capture Methodology

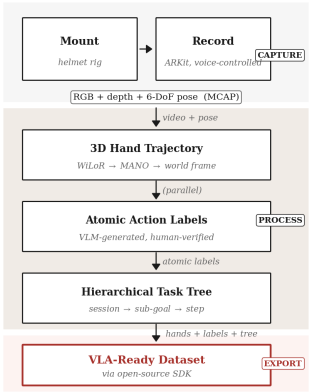


Fig. 2: Overall process

[1]Project resources: (1) Mobile App: Will be released after peer review to maintain anonymity; (2) Python Processing Suite: https://egobench.com/ego-sdk-code; (3) Data Download: https://egobench.com/ego-icra-data; (4) Data Visualization: https://egobench.com/datasets/ego-icra-data-viz; (5) App Code: https://egobench.com/egocentric-app-code

TABLE I: Comparison of egocentric datasets for robot relevant pretraining. Hours are approximate.

Dataset	Hours	Max Episode	6 DoF Pose	Depth	Hand Annot.	Capture Hardware
Ego4D [3]	3,670	∼20 min	Partial	No	Partial	Mixed (GoPro, ZShades, Aria)
EPIC-KITCHENS-100 [5]	100	∼4.5 hrs	No	No	Partial	Head-mounted GoPro
EgoExo4D [6]	1,286	∼42 min	Yes	Yes	Yes	Aria + exo cameras
HOI4D [7]	∼22.2	Short clips (2 minutes)	Yes	Yes	Yes	Intel RealSense
HOT3D [8]	∼14	Short clips (2 minutes)	Yes	No	Yes	Aria + Quest 3
ARCTIC [9]	∼13.5	Very short (< 1 min)	Yes	No	Yes	MoCap rig
Aria Everyday [10]	∼7.5	∼30 min	Yes	No	Yes	Project Aria
EgoDex [14]	829	∼15 min	Yes	No	Yes	Apple Vision Pro
MobileEgo Anywhere (ours)	**200**	**108 min**	**Yes**	**Yes**	**Yes**	**Consumer iPhone**

The data collection process utilizes an iPhone as the primary sensing platform as illustrated in Fig. 1a. The overall process is shown in Fig. 2, where contributors secure the device to a head worn mount, positioned to provide a consistent egocentric field of view. While a standard helmet mount was used for this study, the pipeline is compatible with any mounting hardware that provides sufficient elevation to capture the user's workspace and hand object interactions.

To ensure hands free operation critical for capturing naturalistic daily activities the data collection is managed via the our mobile application using an integrated voice command interface. Users initiate and terminate recording sessions with "start" and "stop" triggers, respectively.

During the recording, the system leverages the ARKit framework to perform realtime sensor fusion. This generates high fidelity, 6 DoF camera poses by synchronizing the onboard IMU with the RGBD stream. The application concurrently archives raw RGB frames, depth maps, and IMU metadata, all registered to a common high resolution timestamp. This ensures temporal consistency across all modalities, providing a robust foundation for downstream 3D reconstruction and action recognition tasks. The data is recorded in an MCAP format and later on processed to generate all the data required to train VLA models.

III-B. Video Processing Pipeline

Following data acquisition, the egocentric video is processed to extract three primary modalities: (i) 3D hand trajectories, (ii) atomic action labels, and (iii) hierarchical task instructions.

III-B1. 3D Hand Trajectory Estimation

High fidelity 3D hand trajectories are essential for training Vision Language Action (VLA) models, as they provide the demonstrations necessary to map human motion to robot end effector frames via Inverse Kinematics (IK).

To extract these trajectories, we employ WiLoR [11], an end to end network optimized for robust 3D hand pose estimation in unconstrained, "in the wild" environments. We utilize the MANO parameterization [12] to represent hand joints, ensuring the predicted poses adhere to biomechanical constraints. This approach is particularly effective in mitigating the effects of partial occlusions common in first person manipulation tasks.

The relative 3D coordinates generated by WiLoR are localized into a global coordinate system by leveraging the synchronized ARKit 6 DoF camera poses and LiDAR derived depth maps. By sampling the depth map at the detected joint locations and applying the extrinsic camera transformation, we project local hand keypoints into a consistent world frame. This results in a spatially anchored trajectory suitable for downstream robotics foundation model training and imitation learning.

TABLE II: ARKit drift evaluation

Experiment	Error at Second Sighting	Error at Third Sighting
Kitchen Activity	0.4 cm	0.7cm
Living Space Activity	0.3 cm	0.4 cm
Whole House Activity	1 cm	1.5 cm

III-B2. Atomic Action Labels

Action conditioned VLA policies require language labels that specify *which* object is being manipulated, *what* the action is, and *where* the object is moving, 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. The raw video is partitioned into contiguous, non overlapping temporal spans, and each span is processed by a vision language model (VLM) that receives the corresponding RGB frames. The model outputs a short imperative sentence constrained by prompt design to include object modifiers (color, material, size) and spatial prepositions (from, into, onto) wherever the video evidence supports them (e.g., "transfer dough from metal bowl to large plate").

We validated the pipeline output against independently human annotated versions of the same 50 sessions. The automated labels average 7.95 words per label versus 2.94 for human labels (computed across 5,249 and 8,898 labels respectively). The difference is qualitative, not just quantitative: where the pipeline writes "transfer dough from metal bowl to large plate," a human annotator on the same frames writes "placing dough on plate", dropping the source container, the transfer verb, and the material modifier. Automated labels also average 1.09 descriptive modifiers per label (color, material, size terms from a fixed 30 word vocabulary) compared to 0.09 for human labels. On the structural side, 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.

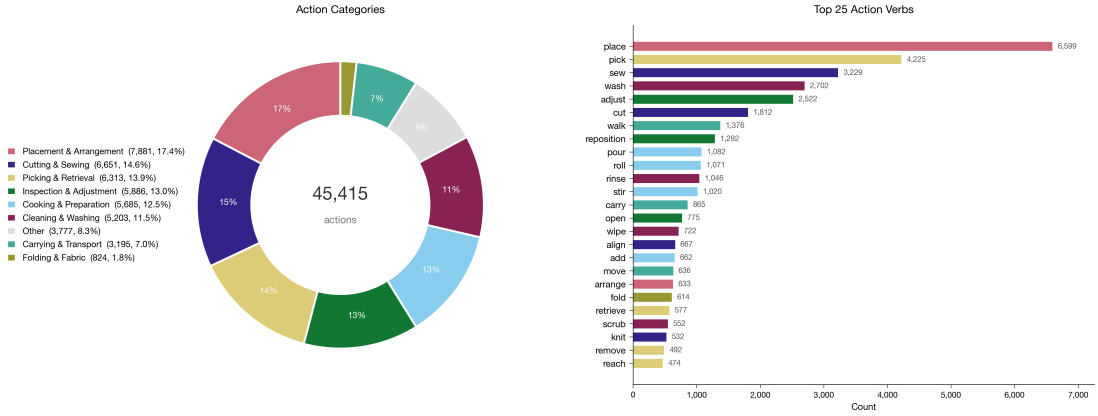


Fig. 3: Task diversity

III-B3. Hierarchical Task Instructions

Long horizon sessions spanning 20-60 minutes contain dozens of atomic labels that belong to distinct sub-tasks as shown in 3, which highlights the action diversity spanning 45K different action categories. To expose this structure, the atomic span captions from the previous stage are organized into a three level instruction tree: a session level goal, sub-goals, and episodes. A language model receives the full ordered sequence of captions as text, with no video input, and groups temporally contiguous spans sharing a common activity into episodes (e.g., "insert pillows into white pillowcases and arrange on bed"), clusters related episodes into sub-goals (e.g., "clean surfaces and make the bed"), and synthesizes one session level goal grounded in the concrete objects across all spans.

We evaluated seven language models on this structuring task; six produced valid outputs satisfying all invariants. The resulting three level tree provides language conditioning at temporal scales from 5 second manipulation steps to minute scale sub-goals to full session plans, matching the multi scale supervision used by recent hierarchical VLA architectures.

IV. Dataset and Evaluation

The released dataset contains 354 sessions totaling 200 hours of egocentric household activity from 16 contributors. Sessions average 21.2 minutes in duration, and the longest session is about 108 minutes of continuous recording. Table I positions the dataset against existing egocentric benchmarks on the modalities required for VLA pretraining.

Several datasets in Table I provide subsets of these modalities. EgoExo4D [6] offers 6 DoF pose, depth, and hand annotations but relies on Meta's Project Aria glasses and synchronized exo cameras, hardware that is not commercially available. Our dataset pairs each RGB frame with a LiDAR depth map and an ARKit 6 DoF pose using a consumer iPhone, and the WiLoR based hand estimation pipeline (Section III-B1) provides 21 joint MANO hand poses anchored in the same world frame. Sessions run up to 60 minutes of continuous recording. The atomic action labels and three level hierarchical instructions described in Section III-B give downstream models access to language conditioning at granularities ranging from individual manipulation steps to full session plans.

IV-1. Long term drift evaluation

Unlike other opensource slam algorithms, the ARKit framework is not openly published but is available to be used through any iphone. Thus, evaluating ARKit presents unique challenges due to its closed source nature. In order to do this, we do a simple experiment - we place an aruco marker in the scene and observe during the first few minutes of operation. We revisit the aruco marker a couple of times during a long term operation, one roughly at the temporal midpoint of the session and the other roughly at the end of the video. In a good slam algorithm with good loop closure, the drift should be minimal and the Aruco marker should stay in the same location as per the camera reference frame. We repeat this experiment in 3 different environments as shown in the Table II and the table shows that the drift is minimal, less than 1 cm in most and less than 0.1 % of trajectory length in all cases. This demonstrates the efficacy of arkit tracking, which can then be used for downstream VLA applications.

IV-2. 3D Hand Pose Consistency

Ground truth MANO hand poses do not exist for unconstrained egocentric recordings at the scale of our dataset. Laboratory benchmarks such as HOT3D [8] and ARCTIC [9] provide millimeter-accurate annotations but cover only minutes of controlled interaction. To assess the quality of WiLoR-estimated hand poses across 98 sessions (1.19 M frames, 25.2 hours), we apply three ground-truth-free consistency metrics that exploit known physical invariants: bone length constancy, joint angle plausibility, and wrist dynamics. Hand detection succeeds on 86.2% of frames, with a mean WiLoR confidence score of 0.73.

A small fraction of frames (247 out of 1.19 M, or 0.02%) exhibit a LiDAR depth sensor edge case in which the returned depth is zero. Because the 3D unprojection step divides by depth, these frames produce wrist positions hundreds of meters from the camera. We identify and discard them with a single threshold (wrist $z > 0.01$ m) before computing all subsequent

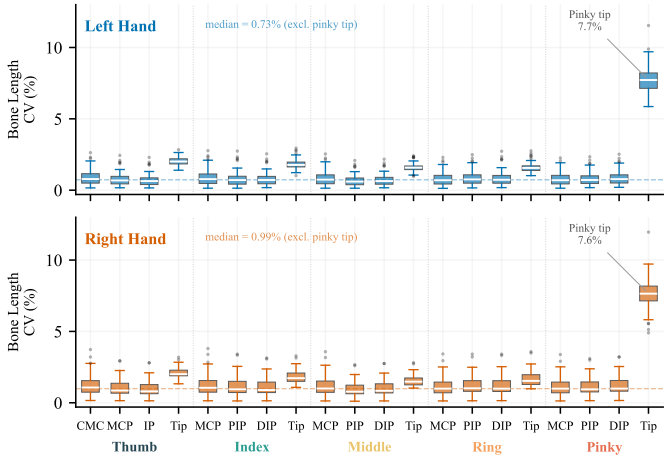


Fig. 4: Per-bone coefficient of variation (CV) of bone length across all valid frames, pooled over 98 sessions. Each bone of the 21-joint MANO skeleton should maintain constant length; lower CV indicates more consistent estimation. The pinky distal bone (joint 17→20) shows elevated CV because its physical length (∼2 cm) amplifies the relative effect of a fixed absolute noise floor (∼1.5 mm). Excluding this outlier, median CV is below 1%.

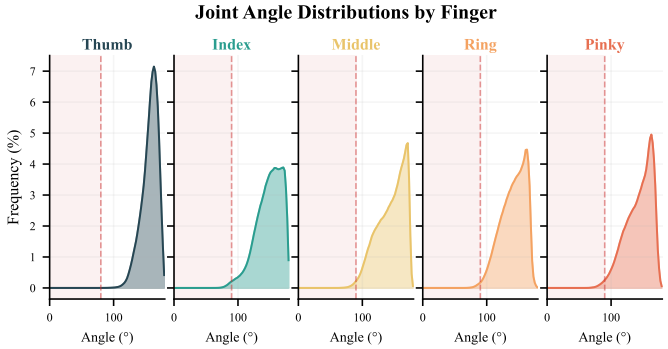


Fig. 5: Distribution of estimated joint flexion angles for each finger, pooled over 98 sessions. Shaded regions indicate published biomechanical flexion limits for MCP, PIP, and DIP joints. Over 99.99% of estimated angles fall within anatomical bounds.

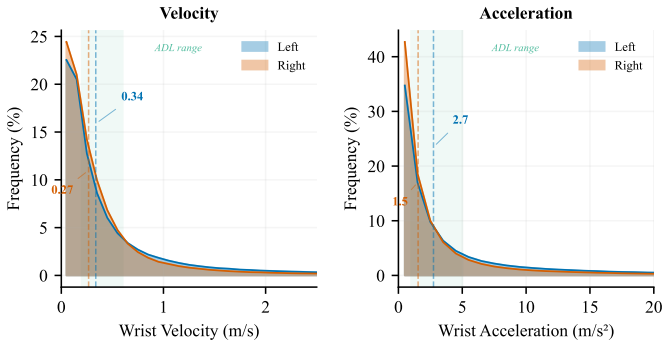


Fig. 6: Wrist velocity and acceleration distributions for left and right hands, pooled over 98 sessions. Shaded bands indicate typical ranges for activities of daily living. Median velocity is 0.34 m/s (left) and 0.27 m/s (right); median acceleration is 2.7 m/s^2 (left) and 1.5 m/s^2 (right).

metrics. The affected frames span 29 of 98 sessions, and their removal has no measurable effect on temporal continuity or coverage.

Bone length constancy. A rigid bone connecting two adjacent joints should maintain the same length regardless of hand configuration. We compute the coefficient of variation (CV) of each of the 20 MANO bone lengths across all valid frames in each session, then pool across sessions. As shown in Fig. 4, median CV is 1.27% for the left hand and 1.43% for the right, indicating that estimated bone lengths remain stable to within roughly 1 mm on a typical 7–8 cm bone. The pinky distal phalanx is a visible outlier at approximately 7.5% CV. This is not a failure of the estimator: the pinky distal bone is physically the shortest in the hand (∼2 cm), so the same absolute noise that produces sub-1% CV on longer bones yields a proportionally larger relative error. Excluding the pinky tip, the pooled median CV drops below 1% for both hands.

Joint angle plausibility. We measure 15 joint flexion angles (MCP, PIP, and DIP for each finger) across all valid frames. Fig. 5 shows the per-finger distributions pooled over all sessions. Over 99.99% of estimated angles fall within published biomechanical limits (approximately 90° for MCP joints and 60–90° for PIP and DIP joints, depending on the finger). The distributions are unimodal and exhibit natural spread consistent with the variety of grasp types and in-hand manipulation present in the dataset.

Wrist dynamics. We compute the instantaneous velocity and acceleration of the wrist joint (MANO joint 0) from consecutive frame positions at 15 fps. Fig. 6 presents the pooled distributions. Median wrist velocity is 0.34 m/s for the left hand and 0.27 m/s for the right, and median acceleration is 2.7 m/s^2 and 1.5 m/s^2 respectively. These values are consistent with the range reported for activities of daily living in the motor control literature, where typical hand velocities during household manipulation fall between 0.1 and 0.8 m/s. The smooth, unimodal shape of both distributions confirms that the depth-filtered trajectories contain no systematic artifacts such as teleportation spikes or oscillatory noise.

IV-3. Hierarchical Instruction Quality

Section III-B3 described the three level instruction tree that our pipeline generates from atomic action labels. To validate this process at scale, we ran the hierarchical decomposition across all 354 sessions using DeepSeek V4 Flash with high reasoning. The model receives the ordered sequence of atomic captions as text input and produces a session goal, sub-goals, and episodes, subject to three structural invariants: every span index appears in exactly one episode, all boundaries use the exact start and end timestamps from the input, and the hierarchy covers the full session with no gaps. Fig. 7 illustrates the output for a representative 36-minute cooking session: 217

atomic spans are grouped into 12 episodes across 5 sub-goals that span fruit preparation, dough kneading, flatbread cooking, grain mixture assembly, and cleanup.

The pipeline produced 45,415 atomic spans, grouped into 5,570 episodes and 1,298 sub-goals across the 354 sessions. Of these, 308 sessions (87%) passed all three structural invariants with zero issues. The remaining 46 sessions had minor boundary mismatches that were automatically corrected in a second pass.

Fig. 8(a) shows that each level of the hierarchy occupies a distinct temporal band. Median durations increase by a factor of 4–8$\times$ at each level: 5 seconds for atomic spans, 42 seconds for episodes, 3.9 minutes for sub-goals, and 15.5 minutes for full sessions. This regular scale separation arises naturally from the data rather than being imposed by the prompt, and it matches the multi-scale temporal structure that recent hierarchical VLA architectures require for effective long-horizon planning. The number of episodes and sub-goals scales linearly with session length (Fig. 8b), confirming that the decomposition adapts to session complexity rather than producing a fixed number of groups. Most episodes are compact: 78% contain 10 or fewer atomic spans (Fig. 8c), with a median of 5 spans per episode. The total cost for processing all 354 sessions was $1.29, making hierarchical structuring negligible relative to the data collection effort itself.

V. Conclusion

In this work, we have presented an accessible and commoditized framework for large scale egocentric data collection. By open sourcing our mobile application and data processing pipeline, we enable researchers and contributors to generate VLA ready datasets using standard consumer hardware. A primary contribution of our approach is the emphasis on long horizon activity tracking; our dataset includes continuous episodes, which have a max length of close to 2 hours, providing a rich resource for researchers focused on long term state tracking and complex task planning. We hope, this efforts democratizes access to VLA dataset creation and lays the path for better VLA models of the future.

VI. Ethics and Privacy

All contributors signed an informed consent agreement covering the capture, processing, and public release of their recordings. Contributors were instructed to avoid recording in environments where other individuals had not consented, and to pause recording when non consenting individuals entered the frame. We also blurred faces of any person, who accidentally came in view during the data collection process.

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. [Online]. Available: https://arxiv.org/abs/2602.16710

[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 Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 18973-18990.

[4] D. Damen *et al.*, "Scaling Egocentric Video Recognition: The EPIC-KITCHENS Dataset," in *Proc. Eur. Conf. Comput. Vis. (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 Conf. Comput. Vis. Pattern Recognit. (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 Conf. Comput. Vis. Pattern Recognit. (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 Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2025.

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

[10] Z. Lv *et al.*, "Aria Everyday Activities Dataset," *arXiv preprint arXiv:2402.13349*, 2024. [Online]. Available: https://arxiv.org/abs/2402.13349

[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. [Online]. Available: https://arxiv.org/abs/2409.12259

[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, Nov. 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.

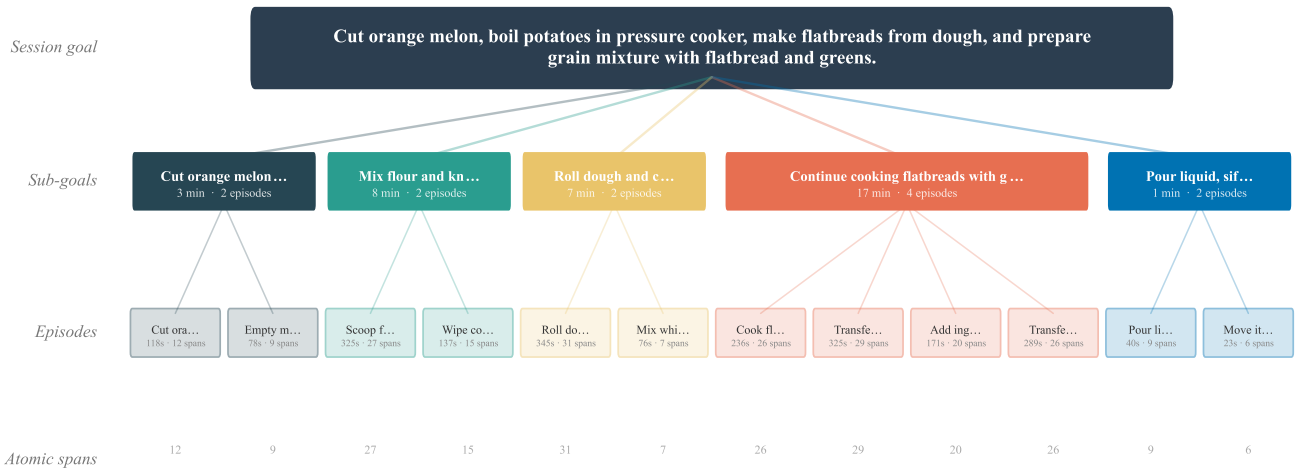


Fig. 7: Hierarchical decomposition of a 36-minute cooking session (217 atomic spans). A single session goal decomposes into five sub-goals, each containing two to four episodes. Sub-goal durations range from 1 to 17 minutes; episode durations from 23 s to 345 s. Numbers at the bottom row indicate the atomic span count per episode. The color grouping highlights how episodes cluster under semantically coherent sub-goals.

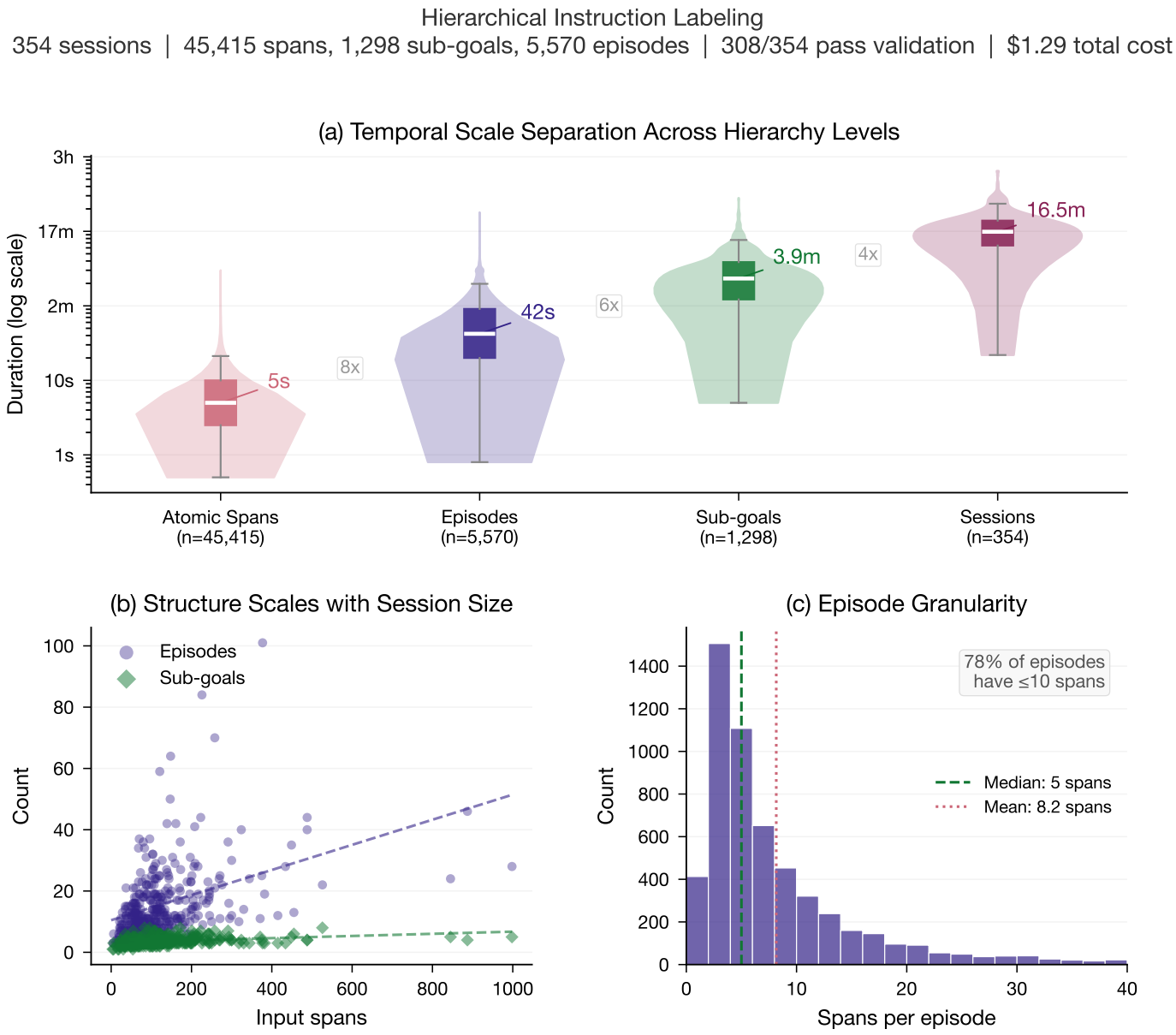


Fig. 8: Hierarchical instruction labeling across 354 sessions (45,415 atomic spans). (a) Temporal scale separation: each level of the hierarchy occupies a distinct temporal band, with consistent 4–8× separation between adjacent levels (median durations: atomic spans 5 s, episodes 42 s, sub-goals 3.9 min, sessions 15.5 min). (b) Episode and sub-goal counts scale linearly with session length. (c) Episode granularity: 78% of episodes contain 10 or fewer atomic spans (median 5, mean 8.2), providing compact supervision units for downstream policy learning.