arXiv:2604.23570v1 [cs.RO] 26 Apr 2026

JINGDONG

EgoLive: A Large-Scale Egocentric Dataset from Real-World Human Tasks

Joy Future Academy, JD

Abstract

The advancement of robot learning is currently hindered by the scarcity of large-scale, high-quality datasets. While established data collection methods such as teleoperation and universal manipulation interfaces dominate current datasets, they suffer from inherent limitations in scalability and real-world deployability. Human egocentric video collection, by contrast, has emerged as a promising approach to enable scalable, natural and in-the-wild data collection. As such, we present **EgoLive**, a large-scale, high-quality egocentric dataset designed explicitly for robot manipulation learning. **EgoLive** establishes three distinctive technical advantages over existing egocentric datasets: first, it represents the largest open-source annotated egocentric dataset focused on real-world task-oriented human routines to date; second, it delivers leading data quality via a customized head-mounted capture device and comprehensive high-precision multi-modal annotations; third, all data is collected exclusively in unconstrained real-world scenarios and encompasses vertical field human working data, including home service, retail, and other practical work scenarios, providing superior diversity and ecological validity. With the introduction of **EgoLive**, we aim to provide the research community with a scalable, high-quality dataset that accelerates breakthroughs in generalizable robotic models and facilitates the real-world deployment of robot systems.

Dataset: https://robotdata-market.jdcloud.com/console/market

1 Introduction

The lack of large-scale datasets has emerged as a critical bottleneck in advancing the generalization capabilities of contemporary robot manipulation learning models. While recent progress in pre-training Vision-Language-Action (VLA) models [28, 30] has demonstrated remarkable progress in facilitating real-world robot manipulation, the pursuit of enhanced generalization cannot be achieved solely through scaling existing robot datasets. Existing datasets are mainly constrained by two key limitations: one is restricted environmental diversity that fails to capture real-world complexity; the other is poor extensibility that makes large-scale data collection implausible. These constraints collectively hinder the development of truly generalizable embodied systems capable of operating in open-world scenarios.

Egocentric data [7, 18, 36] present a promising solution to address current limitations in acquiring data at scale for robot learning. By capturing first-person visual perspectives using head-mounted wearable devices, this approach embodies human-like visual perception in a natural manner and offers several key advantages. **Human Learning Inspiration:** The egocentric approach advocates human-centric visual perception and learning processes, enabling curiosity-driven exploration similar to infant developments [40, 46] and supporting incremental skill acquisition through cumulative first-person experience. **In-the-wild Collection:** Unlike fixed camera systems, this method allows for unrestricted movement in real-world environments, capturing natural

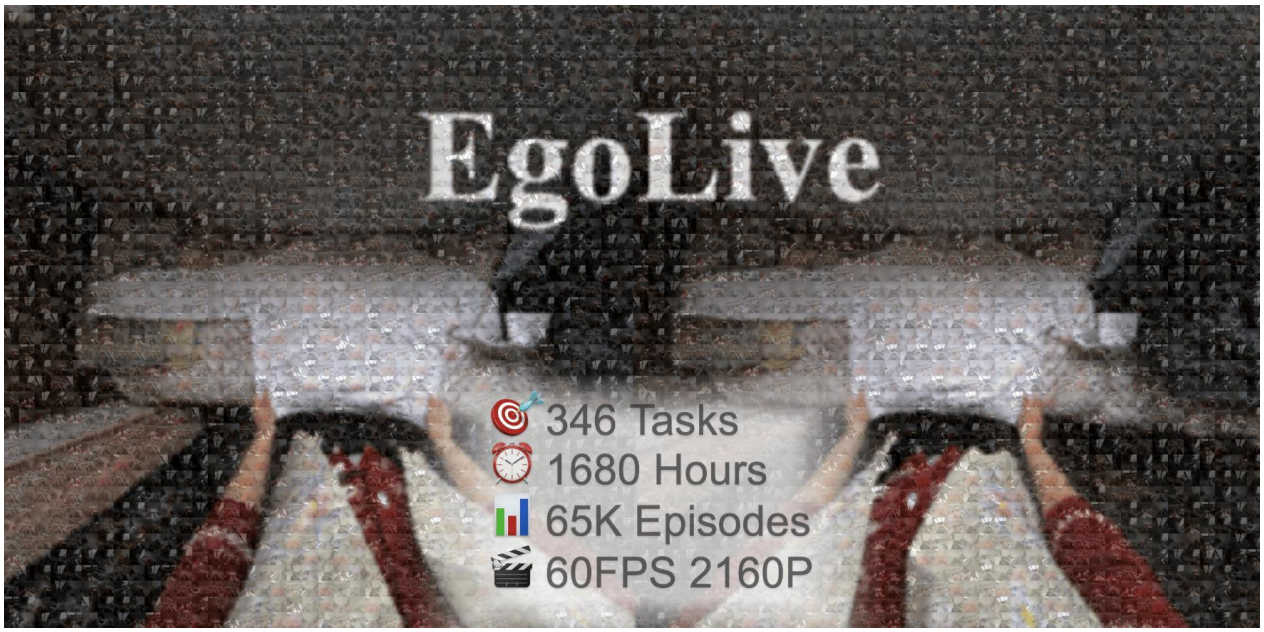

Figure 1 EgoLive is a large-scale high-quality and diverse egocentric dataset. It records human daily manipulation tasks with stereo vision.

human behaviors and interactions without spatial constraints or significant physical interference on the wearer. **Scalability Potential:** The relatively simple hardware setup (e.g., single head-mounted camera) enables large-scale data collection across diverse environments and users with low equipment costs.

We introduce EgoLive, a large-scale egocentric dataset capturing real-world human task demonstrations. It comprises 1,680 hours of high-fidelity stereo video recordings at 60 frames per second, with 65,866 episodes spanning 346 different real-world tasks. Notably, EgoLive exhibits several key characteristics that distinguish it from existing egocentric datasets:

- To our knowledge, EgoLive is the largest open-source annotated egocentric dataset for real-world human tasks to date.
- The dataset achieves superior quality in two dimensions. **Collection specifications**: Our custom-designed head-mounted capture device achieves human-like stereo vision with a 130° horizontal × 130° vertical field of view (FOV), capturing video at 60 frames per second with 2160×2160 resolution per camera. **Annotation richness and accuracy**: It provides comprehensive multi-modal annotations, including 6-DoF motion tracking, fine-grained semantic segmentation and 3D scene reconstruction, all with high accuracy.
- In contrast to existing egocentric datasets, EgoLive is collected entirely in unconstrained real-world environments, offering significantly improved scene diversity and ecological validity for robot learning.

2 Related Work

2.1 Manipulation Data Collection

Manipulation data collection falls broadly into three paradigms: **real-robot teleoperation**, **universal manipulation interfaces** (UMI), and **human egocentric video collection**.

Real-robot teleoperation systems primarily employ three approaches: isomorphic master-follower control, VR-based systems, and motion capture. Notable implementations include ALOHA [49] (low-cost synchronized teleoperation via tethered robotic arms), AgiBot [1] (VR/motion-capture humanoid control), and Tesla

Optimus [41] (wearable motion capture suits). While these methods generate high-fidelity trajectories with strong physical priors for policy learning, they face scalability limitations due to specialized hardware requirements, intensive human involvement, and prohibitive costs for large-scale data collection.

Universal Manipulation Interfaces (UMI) [10] employ instrumented grippers with GoPro cameras to capture human demonstrations, aligning human-robot perceptual views. FastUMI [50] enhances tracking precision via optimized wrist pose estimation. Recent works like Gen series [14, 15] leverage UMI data for scalable manipulation policy training. Although these handheld interfaces enable natural interaction and reduce visual domain gaps, they are typically tailored to specific robot embodiments and thus offer limited cross-platform compatibility. In addition, UMI systems remain difficult to adapt to dexterous hands, further limiting their application scope.

Human egocentric video collection offers a scalable, minimal-hardware paradigm for capturing natural interactions through wearable cameras. EgoScale [51] extracts high-DoF dexterous hand motions from 20k+ hours of in-the-wild videos. EgoDex [18] enables large-scale bare-hand tracking via head-mounted devices; EgoMimic [19] implements portable multi-modal data collection through smart glasses. Compared to real-robot teleoperation, this vision-only approach eliminates hardware constraints and spatial limitations, significantly accelerating real-world data acquisition. Unlike physical gripper-based interfaces like UMI, it preserves natural human hand morphology and thereby supports direct collection of high-DoF dexterous manipulation data.

2.2 Human Egocentric Dataset

Existing human egocentric datasets can be categorized into three major trends: **Generalist**, **Manipulation-centric**, and **Deployment-scale**. They reflect different design priorities for first-person data collection, namely semantic breadth, interaction fidelity, and real-world operational coverage.

Generalist related works such as EPIC-KITCHENS-100 [11] and Ego4D [16] have played a foundational role in egocentric visual understanding. They emphasize large-scale video coverage, rich language supervision, and wide diversity of daily activities and environments. As a result, they have been highly influential for action recognition and general-purpose first-person representation learning. However, these datasets do not offer a unified set of manipulation-oriented geometric signals such as camera trajectories, 3D hand keypoints, dense hand/object masks, and procedural sub-task boundaries. This limits their direct use for learning fine-grained manipulation policies or structured operational procedures without relying on additional resources beyond the core release.

Manipulation-centric works move closer to embodied skill learning by making hand-object interaction directly observable. EgoMimic [19] shows the value of pairing egocentric video with camera motion and 3D

Dataset	Scene	Scale	Resolution	FPS	Multi-view	Motion-tracking	Language Annot.	Depth
Generalist								
EPIC-KITCHENS-100 [11]	Real-world	100 h	1920*1080	50	✗	✗	✓	✗
Ego4D [16]	Real-world	3680 h	Inconsistent	30	✗	✗	✓	✗
Manipulation-Centric								
EgoMimic [19]	Laboratory	55 h	1408*1408	30	✗	✓	✗	✗
HOI4D [24]	Laboratory	22 h	1920*1080	15	✗	✓	✗	✓
HOT3D [4]	Laboratory	14 h	1408*1408	30	✓	✓	✗	✗
EgoDex [18]	Laboratory	829 h	1920*1080	30	✗	✓	✓	✗
Deployment-Scale								
Egocentric-10K [7]	Real-world	10k h	1920*1080	30	✗	✗	✗	✗
Xperience-10M [36]	Real-world	1059 h	512*512	20	✓	✓	✓	✓
EgoLive (Ours)	**Real-world**	1680 h	**2160*2160**	**60**	✓	✓	✓	✓

Table 1 Comparison of representative human egocentric datasets relevant to embodied manipulation and human-to-robot transfer. A modality is marked as present only if provided in the main public release of the dataset. EgoLive targets real-world scenes with the second longest collection duration and achieves superior spatiotemporal resolution and annotation comprehensiveness.

hand tracking for robot learning. HOI4D [24] provides substantially denser supervision, including RGB-D, camera pose, 3D hand pose, segmentation masks and action segmentation, making it particularly useful for category-level hand-object interaction analysis. HOT3D [4] strengthens the 3D tracking perspective through accurate hand, object, and camera pose annotations in an egocentric multi-view setup. EgoDex [18] advances this paradigm both in volume and annotation richness, providing large-scale dexterous tabletop manipulation data that combines language annotations with camera pose and native 3D hand tracking. Collectively, these datasets are valuable for learning manipulation priors, action grounding, and demonstration-based transfer. Nevertheless, most of them remain centered on household, laboratory, or tabletop settings and therefore provide limited coverage of longer-horizon workflows in real service or business operations.

Deployment-scale related works extend egocentric data collection toward greater operational diversity and more realistic long-duration capture. Egocentric-10K [7] introduces large-scale first-person recording to real factory environments, improving coverage of real operational footage but providing only relatively sparse annotation types. Xperience-10M [36] substantially expands its multi-modal sensory and annotation profile, combining synchronized first-person video with depth, camera pose, hand motion capture, and hierarchical language annotations that encompass task and sub-task structures. These datasets represent important steps toward deployment-oriented embodied intelligence.

EgoLive pushes the boundary further by collecting human egocentric data in real service-oriented business scenarios including home services, retail and pharmacy which remain largely underexplored in existing egocentric datasets. In addition to RGB videos and language annotations, this dataset includes camera pose, 3D hand keypoints, depth maps, hand masks, interacted object masks and sub-task segmentations that are all generated via our automated processing and annotation pipeline. The comprehensive annotations make EgoLive suitable not only for egocentric perception, but also for manipulation grounding, workflow decomposition and downstream human-robot transfer tasks in real-world scenarios. The detailed comparison with other egocentric datasets are listed in Tab. 1.

2.3 Learning from Egocentric Data

Human egocentric data provides visual observations and human motion data for representation and policy learning in robotic manipulation. Existing works on learning from egocentric data can be organized into three categories: **capability learning from human egocentric data**, **transfer from human data to robots**, and **extensions to whole body control**.

Capability Learning from Human Egocentric Data explores what capabilities for robot manipulation can be learned from human egocentric demonstrations and how larger-scale pre-training enhances such capabilities. EgoDex [18], EgoMimic [19] and EgoVLA [45] investigate action-level capability learning from human egocentric data through hand trajectory imitation, co-training with human and robot demonstrations, and action-aware visual representation learning, respectively. VITRA [23] and EgoScale [51] further extend this line to larger-scale pre-training by constructing robot aligned training data from human egocentric videos and studying the scaling behavior of egocentric pre-training. Together, these works show that human egocentric data is a scalable source of action representations and manipulation priors for robot learning.

Human-to-robot transfer studies how capabilities learned from human demonstrations can be transferred effectively to robot execution despite the embodiment gap. Phantom [20], Masquerade [21], and Mimic-Dreamer [22] reduce the embodiment gap via image distribution alignment and demonstrate the effectiveness of the resulting supervision signal for robot policy learning. Humanoid Policy ~ Human Policy [32] and Being-H0.5 [26] construct shared action interfaces between human demonstrations and robot control through differentiable retargeting and semantically aligned action spaces. H-RDT [5] and EgoBridge [31] further improve transfer through staged adaptation and representation alignment, respectively. Collectively, these works improve the transfer of manipulation policies from human egocentric data to robotic systems.

Extensions to Whole-Body Control consider learning from egocentric data beyond hand centered manipulation to more coordinated embodied behaviors. ZeroWBC [44], EgoMI [47], and EgoHumanoid [39] extend from manipulation to humanoid whole body control, coordinated active vision and manipulation, and whole body loco-manipulation, respectively. These studies suggest that egocentric data can support not only manipulation

Figure 2 Human Data Capture System The system leverages JoyEgoCam which is a custom-designed head-mounted device for acquiring human behavior data in real-world environments. It features stereo RGB cameras that provide a wide field-of-view and an integrated IMU with 200Hz measurements.

learning, but also more general embodied control involving active vision and whole body motion.

3 Dataset

EgoLive is a large-scale in-the-wild egocentric manipulation dataset with rich and high-quality annotations. To enable massive-scale data collection, we introduce a lightweight ergonomic headset that allows for long-term in-the-wild acquisition. We further develop an automated annotation pipeline to generate comprehensive, high-quality annotations covering action trajectories, semantic understanding and 3D reconstruction. Extensive statistical analysis demonstrates that EgoLive has a wide coverage and high local density, exhibiting rich data diversity.

3.1 Data Collection

All data is collected using **JoyEgoCam** (see Fig. 2), a custom-designed head-mounted device for human behavior acquisition in the wild. JoyEgoCam is equipped with stereo RGB cameras that provide a wide FoV analogous to human binocular vision. The high resolution and frame rate facilitate fine-grained and agile human motion tracking, while the integrated Inertial Measurement Unit (IMU) enables improved accuracy in camera pose estimation.

The data is primarily collected in real-world scenes by recruited operators. In contrast to VR headsets that occlude users' faces and wearable devices that interfere hand motion, our device adheres to a minimal-intrusion design principle. Consequently, users experience little discomfort and can perform daily actions naturally during collection. Benefiting from the head-mounted form factor and human-like field of view, natural hand movements can be captured realistically.

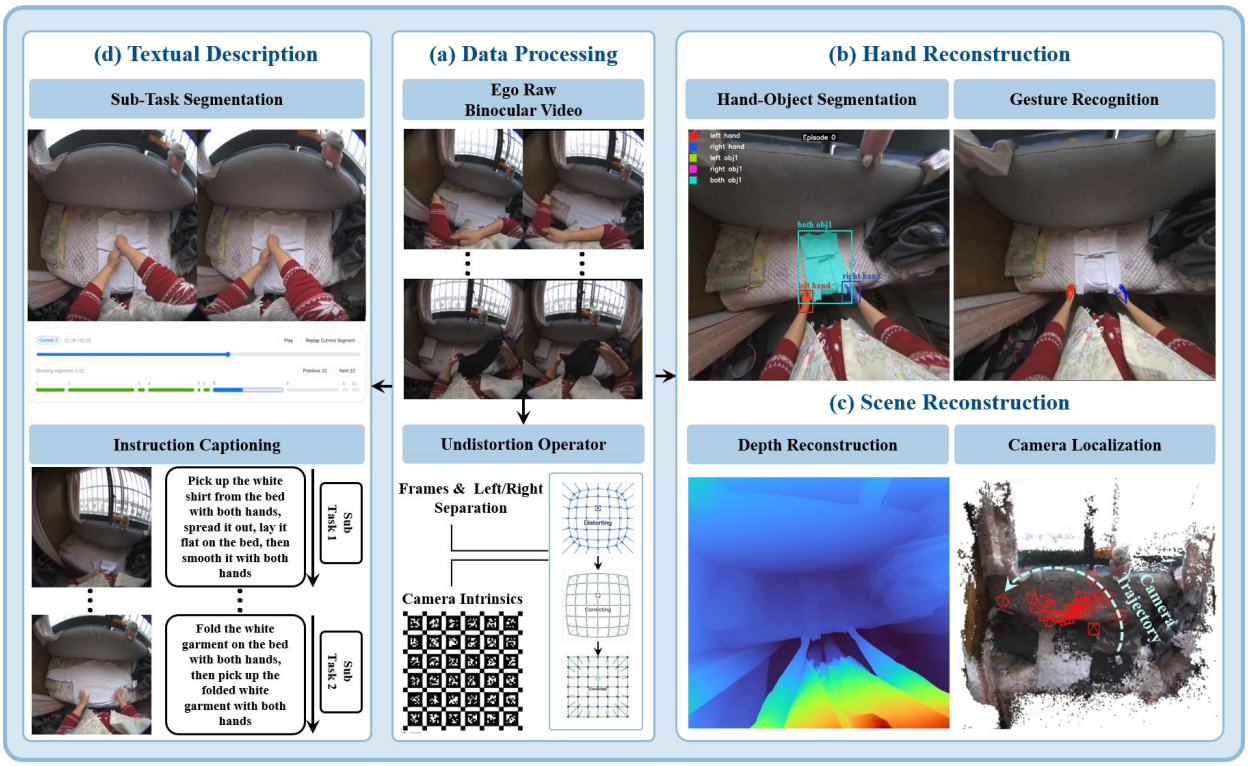

Figure 3 Overview of the automated annotation pipeline for egocentric binocular videos. (a) The raw egocentric binocular video is first undistorted, with left/right view separation and camera intrinsic calibration. (b) Hand reconstruction is conducted, integrating semantic gesture recognition and hand-object segmentation to capture interaction dynamics. (c) Dense depth reconstruction and camera localization are performed to recover the scene structure and estimate the camera trajectory based on the rectified video and sensor cues. (d) In parallel, sub-task segmentation and instruction captioning are employed to generate temporal textual description.

The raw multi-modal data comprises stereo RGB videos with 2160×2160 resolution at 60Hz, paired with camera calibration files, camera trigger frame timestamps, and time-synchronized IMU measurements at 200Hz. Each video typically spans 1 to 3 minutes, featuring a single continuous mid- to long-duration manipulation activity performed in real-world environments.

3.2 Automated Data Annotation

As illustrated in Fig. 3, we develop a comprehensive in-house data annotation pipeline for frame-wise processing of multi-modal data streams. The pipeline integrates three key components:

Motion Tracking: 6D trajectories of both wrist and hand joints are estimated and then synchronized with camera ego-motion estimation to establish action reference frames. The hand motion is estimated by a two-stage method based on HaMeR [29]. MANO [35] parameters are firstly estimated from each monocular video, which ensures precise 2D projections. The camera's ego-motion is estimated by ORB-SLAM3 [8] which fuses binocular RGB images and IMU data, with initialization starting at the first frame of the left camera.

Semantic Understanding: Hierarchical semantic information, including hand-object interaction states and natural language descriptions of actions, is obtained via a data processing pipeline integrating detection, tracking, segmentation and LLM-driven textual annotation techniques. Human hands and corresponding interacted objects are detected using the model proposed in [38] and the outputs are tracked via BoT-SORT [2]. SAM2 [34] is adopted to generate segmentation masks for hands and interacted objects. Each episode is partitioned into sub-tasks according to hand-object detection and tracking results. A fine-tuned Qwen3-VL-32B [3] model then takes these segmented sub-task clips as input and employs a multi-stage reasoning strategy at inference time to generate fine-grained descriptions.

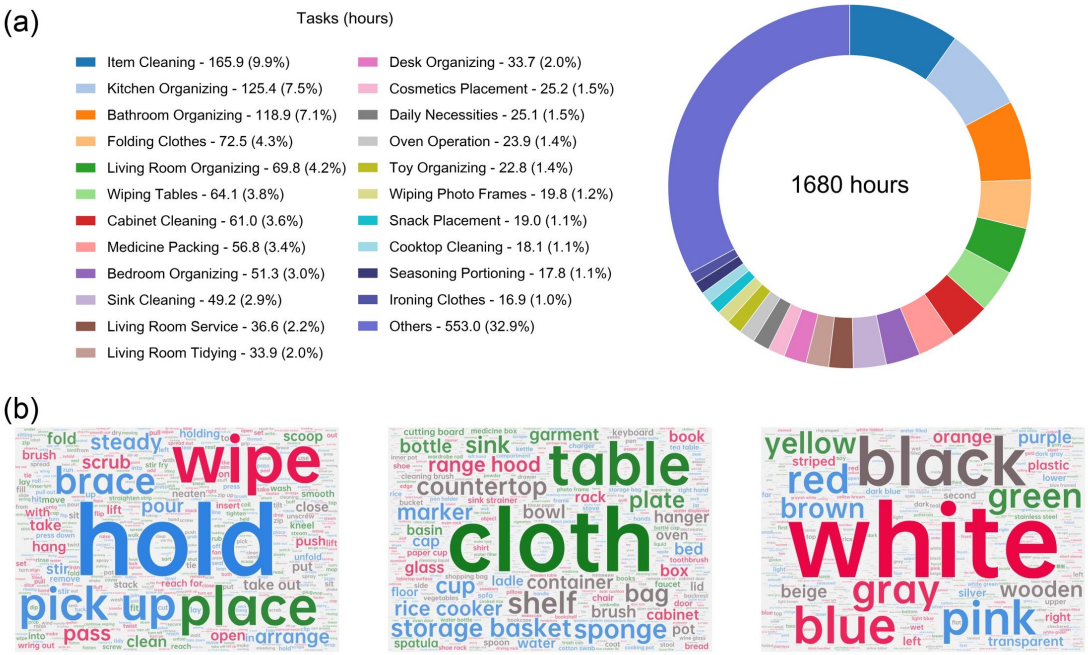

Figure 4 Discrete semantic composition of EgoLive derived from task description. (a) Distribution of task categories, illustrating coverage of real-world activity domains including household services, organization, cleaning, logistics, and other manipulation-intensive scenarios. (b) Word clouds of high-frequency semantic labels extracted from instruction captions across action, object and object attribute.

3D Reconstruction: Accurate 3D hand reconstruction and depth estimation can be achieved by effectively leveraging binocular vision. In the second stage of hand motion estimation, stereo optimization is performed based on the previously estimated MANO parameters, yielding consistent 3D hand keypoint positions that adhere to the 2D projection constraints within each monocular video. Depth is reconstructed with 1152×1152 resolution from finely calibrated stereo RGB videos using FoundationStereo [43].

Overall, this pipeline enables streamlined acquisition of geometric, kinematic, and semantic annotations from stereo RGB and IMU inputs, providing multi-modal supervision signals for downstream tasks such as egocentric perception, hand-object interaction understanding and manipulation policy learning.

3.3 Dataset Distribution

Prior work on foundation models has shown that data diversity is critical for robust generalization [6, 9, 13, 33, 37]. Similar observations have also been made in robot learning [14, 15, 42], motivating work on data-centric strategies such as scaling, deduplication, re-weighting, and quality-aware data selection to refine training data distributions and thereby to improve generalization performance of embodied models [17, 48, 51]. In this section we present a systematic analysis of EgoLive from two complementary perspectives, discrete and continuous, to show that Egolive achieves improved data diversity in terms of both coverage and density in comparison to existing egocentric datasets.

Discrete Semantic Distribution:

In terms of discrete semantic analysis, we extract scene categories and object–action–attribute labels from instruction captions and use their frequency statistics to compare semantic coverage and long-tailed structure across datasets [23].

As illustrated in Fig. 4, EgoLive spans a wide spectrum of tasks, showing broad coverage of manipulation-intensive scenarios. The word clouds further demonstrate that frequent tokens cover a rich variety of actions, objects and object attributes.

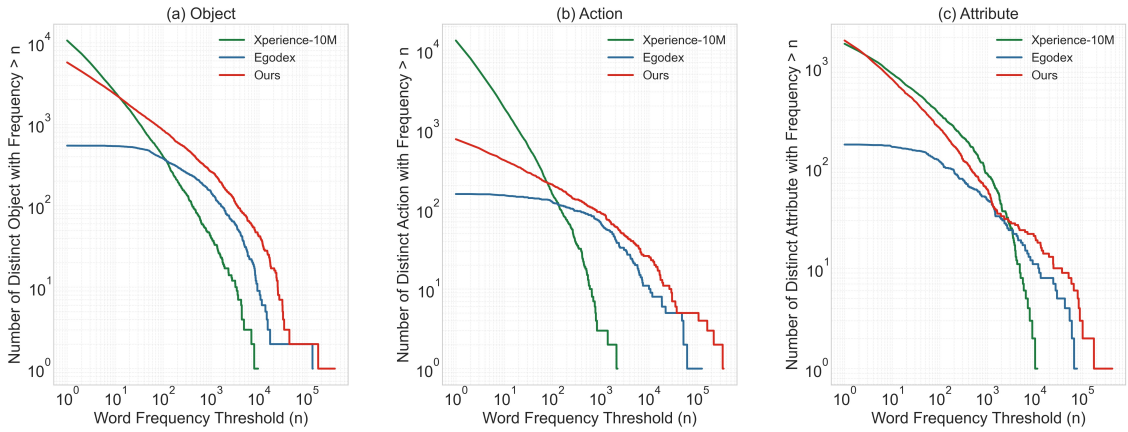

Figure 5 Distributions of semantic labels derived from instruction captioning across datasets: (a) object distribution, (b) action distribution, (c) attribute distribution. The x-axis represents the word frequency threshold n and the y-axis shows the number of distinct words with frequency greater than n (log-log scale).

Fig. 5 compares EgoLive with EgoDex (829 hours) [18] and Xperience-10M (1059 hours available to date) [36] based on word (label) frequency. It is clear that EgoLive lies above EgoDex and Xperience-10M and shows longer tails across all semantic dimensions. Overall, EgoLive exhibits broader semantic coverage and a more natural long-tailed distribution over objects, actions, and attributes, reflecting its superior semantic diversity compared to existing egocentric datasets.

Continuous Feature Space Analysis:

In addition to discrete semantic analysis, we analyze dataset distributions in a continuous feature space. Specifically, image embeddings are extracted using Cosmos-Embed1-448p [27] and visualized via t-SNE. A joint representation of objects, environments and actions is formed to characterize the overall structure of the egocentric data.

The local neighborhoods in the t-SNE graph reflect semantic similarity, whereas clusters correspond to consistent interaction patterns that share similar objects, actions, and scene structures. The relative distances between clusters indicate variations across different interaction patterns.

As shown in Fig. 6, EgoLive occupies a broader region of the representation manifold and exhibits more locally coherent clusters than EgoDex and Xperience-10M. This demonstrates that EgoLive not only achieves wider interaction patterns but also preserves clear local structures in the continuous representation space.

4 Accuracy Evaluation

4.1 Hand Reconstruction

To evaluate the accuracy of hand reconstruction, we first provide visual comparisons of 2D hand keypoint annotations between EgoDex and our dataset, as shown in Fig. 7. Keypoints derived from EgoDex display obvious localization errors, resulting in significant misalignment between the projected skeleton and the actual hands in the images. In comparison our dataset generates much more accurate keypoints that align closely with hands across all frames.

In addition, our method achieves precise 3D keypoints annotations through joint optimization in the stereo space. The 3D estimation accuracy is validated by projecting the estimated keypoints onto the corresponding scene reconstructed from depth images and evaluating their alignment. As shown in Fig. 8, the predicted 3D hand skeleton demonstrates excellent alignment with the point cloud representation of the hands. Building on the depth reconstruction evaluation detailed in Section 4.2, which achieves millimeter-level depth accuracy, we conclude that our 3D keypoint estimation achieves competitive precision. This result confirms that our

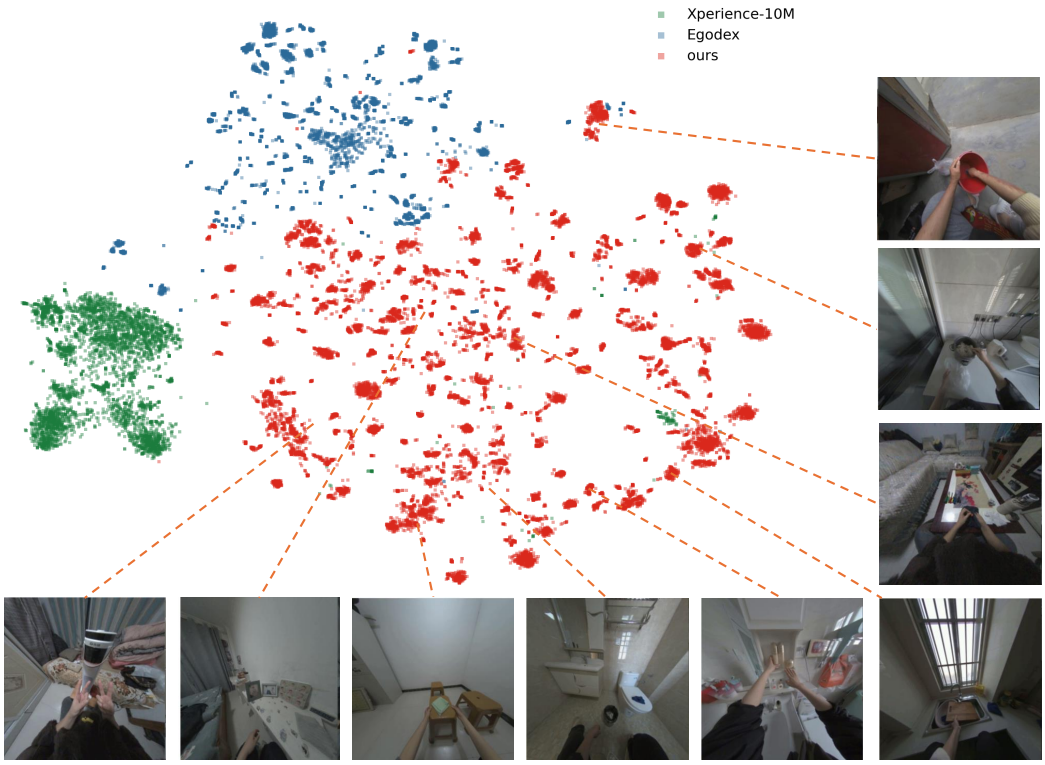

Figure 6 t-SNE visualization of the joint embeddings of objects, environments and actions. Example frames are representative samples from the different regions that illustrate the wide diversity of our data.

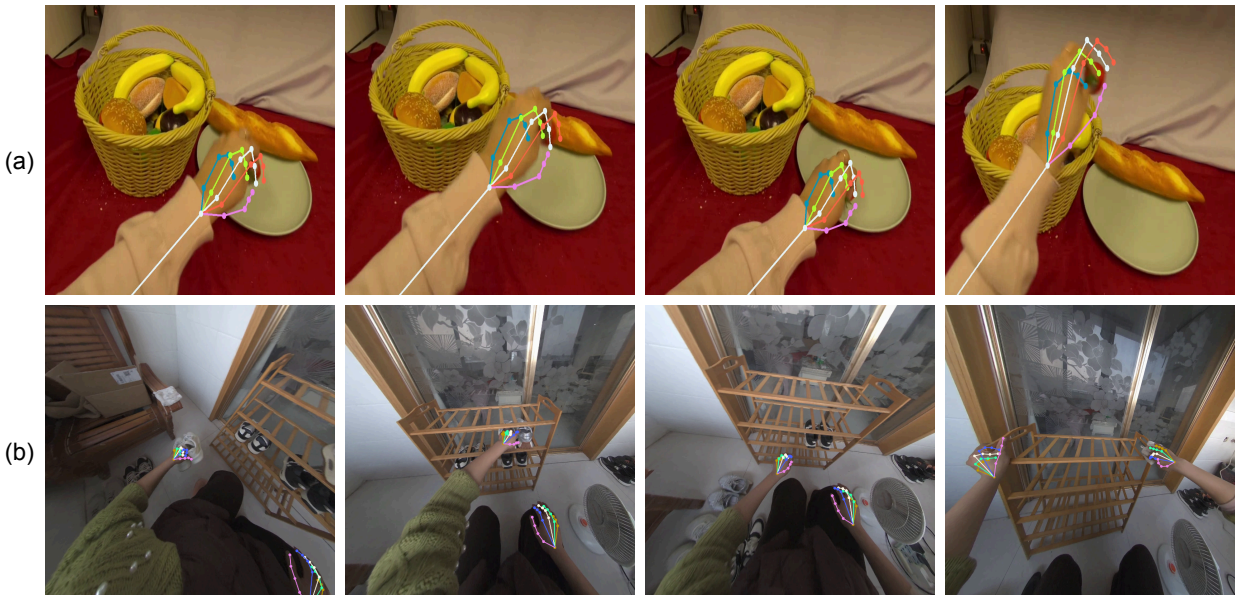

Figure 7 Qualitative Comparison of 2D Keypoint Annotations. Visualizations of 2D keypoint annotations from (a) EgoDex and (b) EgoLive. EgoDex has non-negligible annotation errors and spatial misalignments. In contrast, EgoLive provides highly accurate and robust 2D keypoint annotations.

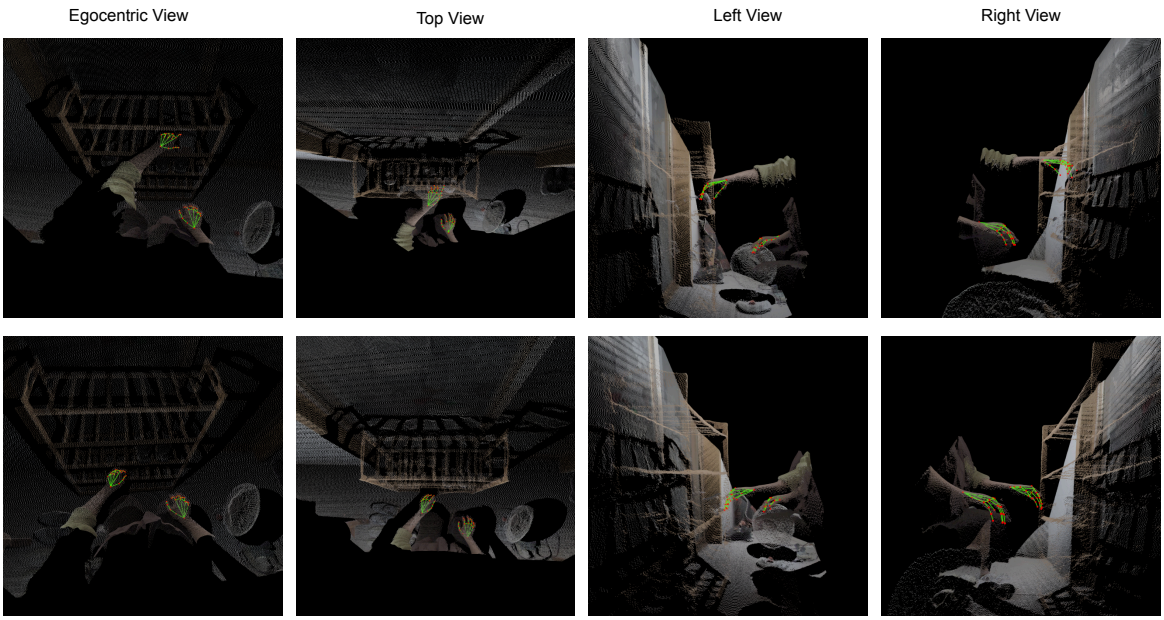

Figure 8 Multi-view Visualization of 3D Keypoints. 3D keypoint estimations rendered from four distinct perspectives: egocentric, front, left-side, and right-side views. Our method preserves strong spatial consistency with no wrist drift across all viewing angles.

hand reconstruction framework effectively addresses depth drift and maintains consistent absolute physical scale, even under severe dynamic occlusions.

4.2 Depth Reconstruction

We adopt a calibrated checkerboard-based method to quantitatively evaluate depth reconstruction accuracy. Specifically, a stepped array with known depth gradients and a ChArUco calibration board are used as reference targets and captured by our system at multiple distances from $0.5m$ to $3.5m$. Following stereo

Distance (mm)	Mean Error (mm)	Error < 5 mm %	Error < 10 mm %	Error < 20 mm %	Error < 30 mm %	Error < 40 mm %
500	3.059	79.576	99.19	100.0	100.0	100.0
700	3.045	79.755	99.833	100.0	100.0	100.0
900	5.381	56.86	87.051	96.987	99.917	100.0
1100	7.091	43.84	74.817	96.099	99.417	100.0
1300	7.033	48.989	73.227	95.577	99.697	100.0
1500	8.751	35.636	63.664	92.528	99.333	99.917
1700	13.114	24.233	43.992	76.874	93.943	99.079
1900	12.083	27.696	49.601	79.745	94.904	99.074
2100	13.014	26.505	47.433	74.121	92.053	99.75
2300	15.095	20.878	40.893	69.96	87.822	96.933
2500	13.994	25.061	47.391	71.72	88.6	97.527
2700	14.82	23.789	41.761	72.063	87.318	96.4
2900	13.414	31.071	49.34	71.32	89.229	98.794
3100	15.254	26.29	45.319	65.63	83.568	97.064
3300	17.372	18.98	36.162	64.54	81.561	92.889
3500	18.035	20.952	36.861	59.798	79.431	93.478

Table 2 Quantitative analysis of the proposed depth reconstruction model in the calibration room. We report the mean reconstruction error and error distribution under different thresholds, evaluated at operating distances from $500mm$ to $3500mm$.

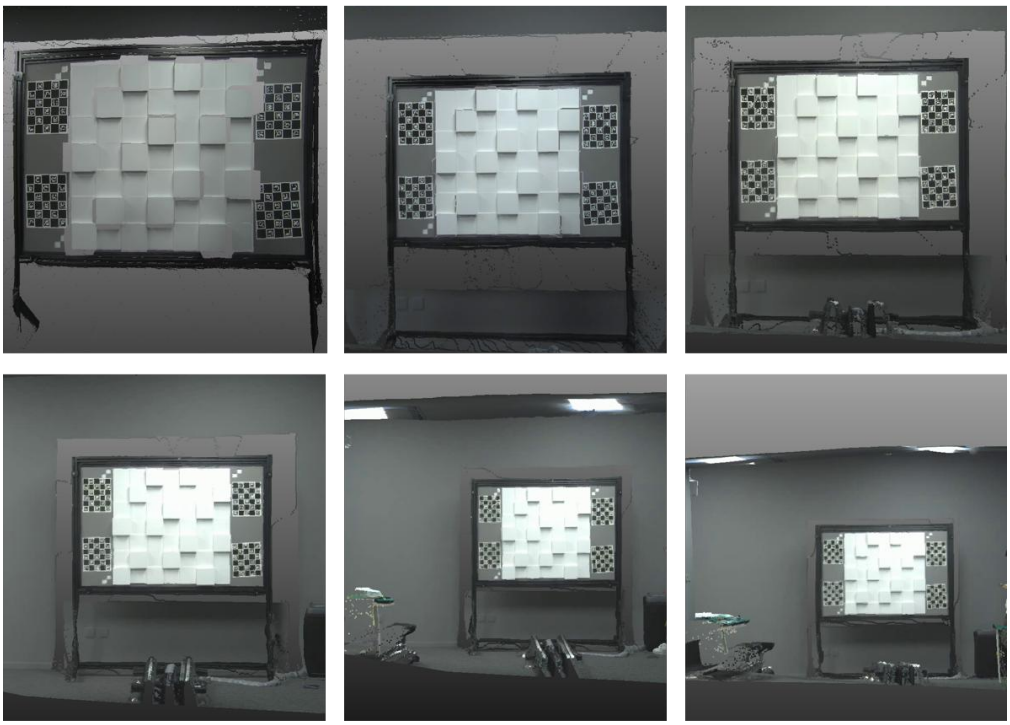

Figure 9 Point cloud reconstructions of the calibration board at distances ranging from $500mm$ to $3500mm$.

rectification, the 6D poses of camera relative to the checkerboard are estimated by jointly optimizing the global reprojection error using the BFGS algorithm. Subsequently, the depth of each corner point is calculated based on the known geometry of the checkerboard, serving as the predicted ground truth. By comparing this ground truth depth with the depth predicted from our depth reconstruction method, as illustrated in Fig. 9, we obtain the quantitative evaluation results of the model, which are presented in Tab. 2. These results demonstrate that our depth reconstruction approach can achieve low mean errors within the typical human operation range (approximately less than $0.9m$).

To further assess the qualitative performance and practical applicability of our proposed method, three representative collection scenes are considered in Fig. 10. The visual results demonstrate that JoyEgoCam, together with the depth reconstruction method [43], enables high-quality depth estimation that faithfully recovers the underlying spatial structures and complex 3D geometric information.

4.3 Instruction Captioning

For each episode, the video is segmented into sub-task clips corresponding to complete atomic actions with durations ranging from approximately 1 to 20 seconds. An instruction captioning module is introduced to generate structured semantic annotations for each clip. Instead of producing completely free-form descriptions, the module explicitly models three essential elements for understanding egocentric manipulation: **hands**, **manipulated objects** and **actions**. This design is motivated by the observation [12, 23, 25] that sub-tasks are tightly coupled with hand action and the manipulated object in human video captioning. As shown in the top row of Fig. 11, a naive natural language description may be <*use a squeegee to wipe the glass door*> whereas our description provides fine-grained information about hand usage, manipulated objects, and action.

To evaluate the quality of captioning results, we adopt a granular caption assessment protocol using LLM-as-a-Judge that assesses the consistency of key elements with ground-truth descriptions, including hand consistency, object consistency, action consistency, and global consistency. Hand consistency measures whether the caption correctly identifies the interaction hand and object consistency evaluates whether the manipulated object matches the visual content, including object category, color, and shape. Action consistency evaluates whether

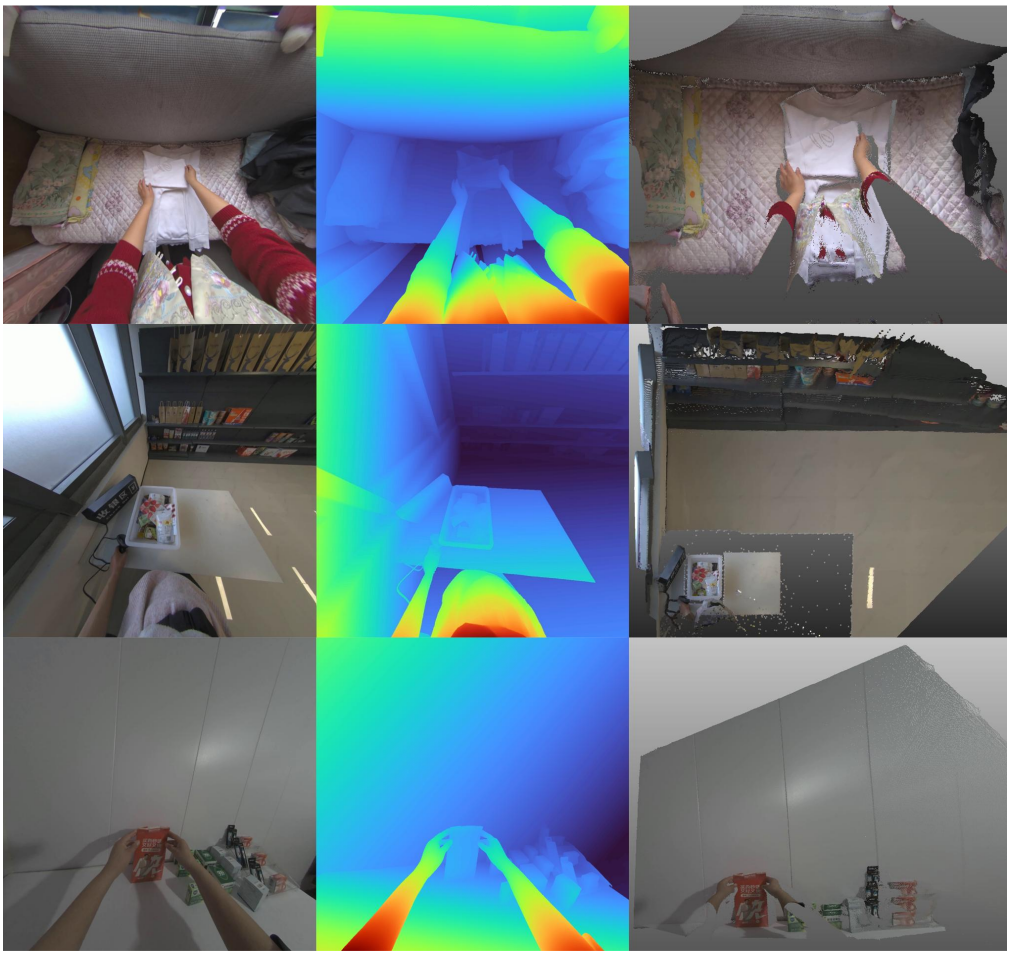

Figure 10 Visualization of depth and point-cloud reconstruction results in real-world scenes from EgoLive dataset. From left to right at each row: original left image, predicted depth map and reconstructed point clouds. The spatial structure and 3D geometric information across diverse real-world scenes can be recovered effectively by our proposed method.

the described manipulation action accurately reflects the observed behavior. Global consistency measures whether the caption provides a faithful and coherent description of the sub-task clip, taking into account

Sample	GT Caption	Predicted Caption	Action	Object	Hand	Global
1	Clean the glass from top to bottom with the white squeegee in the right hand.	Hold a white squeegee in the right hand and clean the glass from top to bottom.	✓	✓	✓	✓
2	Place the black cotton coat hanging on the hanger flat on the bed, then zip it up with both hands.	Place the down jacket on the bed with both hands, and zip it up.	✓	✓	✓	✓
3	Remove the cloth covering the sponge with the right hand. Then pick up the sponge with the right hand.	Move to the left, pick up the yellow sponge with the right hand, then wipe the white countertop.	×	×	✓	×
4	Spread and smooth the blue-and-white striped duvet flat on the bed with both hands. Then grasp its two corners with both hands, lift it, and fully unfold it. Finally, walk toward the left of the bed.	Fold the blue-white-gray striped quilt and arrange it with both hands, then walk to the left side of the bed.	×	✓	✓	×

Table 3 Detailed evaluation results for the examples in Fig. 11.

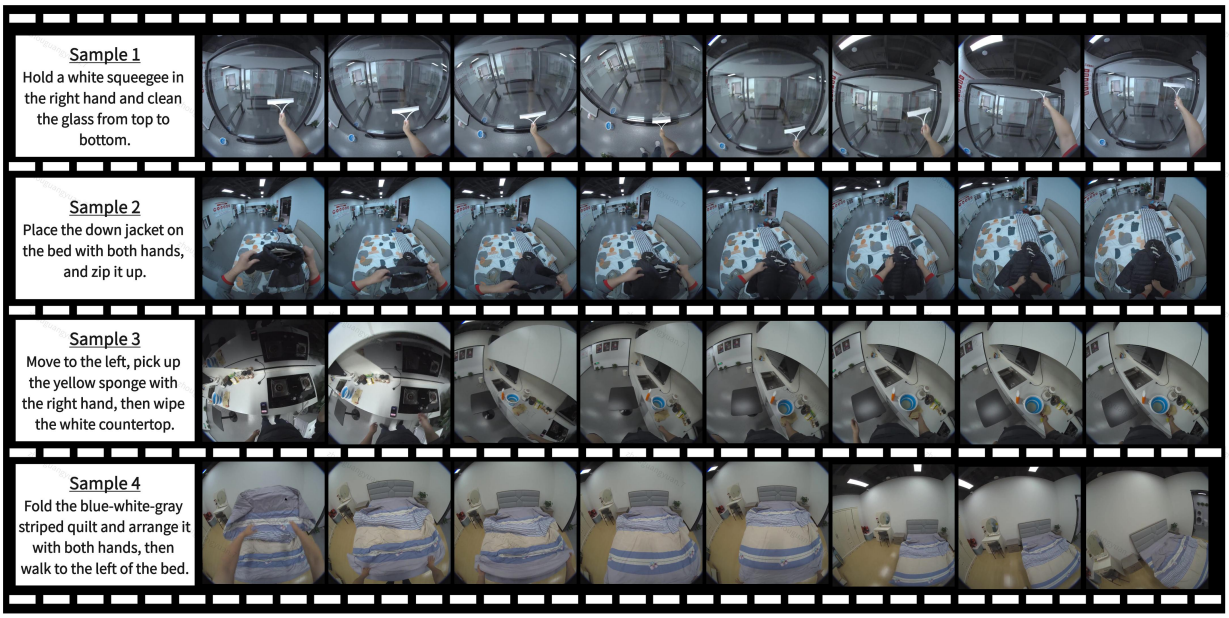

Figure 11 Qualitative examples from our instruction-based annotation pipeline across diverse scenes, including glass wiping, clothes folding, refrigerator cleaning, and bed making.

the joint correctness of hand usage, object interaction, action semantics as well as sentence naturalness and completeness. We present corresponding consistency evaluation examples in Tab. 3 for the cases shown in Fig. 11.

5 Conclusion

This work introduces **EgoLive**, the world's largest open-source egocentric dataset dedicated to human daily operational activities within realistic, utility-driven working scenarios. Leveraging purpose-built hardware, a scalable collection pipeline and a high-accuracy annotation framework, EgoLive delivers industrial-grade high-quality human demonstration data suitable for robot learning. As a substantial addition to existing public egocentric datasets, it offers qualitative improvements over prior work in data scale, data quality and diversity. More importantly, building upon deployment-friendly collection devices and a streamlined data production pipeline, EgoLive is designed to sustain continuous growth in both scale and coverage, establishing an ever-growing knowledge base of natural human behavioral priors. It is poised to inspire and support key research directions in the robotics community, such as human-to-robot alignment, humanoid robot policy learning, and ultimately to bridge the gap between human behavior and robot action.

6 Contributions

Contributors

Yihang Li [†], Xuelong Wei, Jingzhou Luo, Yingjing Xiao, Yibo Bai, Guangyuan Zhou, Teng Zou, Chenguang Gui, Jiajun Wen, He Zhang, Kangliang Chen, Xing Pan, Shuaiyan Liu, Daming Wang, Tao An, Jiayi Li, Shibo Jin, Wanwan Zhang, Tianyu Wang, Boren Wei, Zhixuan Huang, Fangsheng Liu, Ruodai Li, Hui Zhang, Anson Li, Yicheng Gong, Peng Cao, Jiaming Liang, Liang Lin [†].

[†]Corresponding author: Yihang Li <liyihang18@jd.com>, Liang Lin <linliang@ieee.org>

References

[1] AgiBot. Create unlimited productivity via intelligent machines. https://www.agibot.com/, 2026. Accessed: 2026-04-06.

[2] Nir Aharon, Roy Orfaig, and Ben-Zion Bobrovsky. BoT-SORT: Robust associations multi-pedestrian tracking. arXiv preprint arXiv:2206.14651, 2022.

[3] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-VL technical report. arXiv preprint arXiv:2511.21631, 2025.

[4] Prithviraj Banerjee, Sindi Shkodrani, Pierre Moulon, Shreyas Hampali, Shangchen Han, Fan Zhang, Linguang Zhang, Jade Fountain, Edward Miller, Selen Basol, et al. HOT3D: Hand and object tracking in 3D from egocentric multi-view videos. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 7061–7071, 2025.

[5] Hongzhe Bi, Lingxuan Wu, Tianwei Lin, Hengkai Tan, Zhizhong Su, Hang Su, and Jun Zhu. H-RDT: Human manipulation enhanced bimanual robotic manipulation. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 40, pages 18135–18143, 2026.

[6] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, et al. Language models are few-shot learners. Advances in Neural Information Processing Systems (NeurIPS), 2020.

[7] AI Build. Egocentric-10K. URL https://huggingface.co/datasets/builddotai/Egocentric-10K, 2025.

[8] Carlos Campos, Richard Elvira, Juan J. Gómez Rodríguez, José M. M. Montiel, and Juan D. Tardós. ORB-SLAM3: An accurate open-source library for visual, visual–inertial and multi-map slam. IEEE Transactions on Robotics, 37(6):1874–1890, 2021. doi: 10.1109/TRO.2021.3075644.

[9] Xi Chen, Josip Djolonga, Piotr Padlewski, Basil Mustafa, Soravit Changpinyo, Jialin Wu, Carlos Riquelme Ruiz, Sebastian Goodman, Xiao Wang, Yi Tay, Siamak Shakeri, Mostafa Dehghani, Daniel Salz, Mario Lucic, Michael Tschannen, Arsha Nagrani, et al. On scaling up a multilingual vision and language model. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 14432–14444, June 2024.

[10] Cheng Chi, Zhenjia Xu, Chuer Pan, Eric Cousineau, Benjamin Burchfiel, Siyuan Feng, Russ Tedrake, and Shuran Song. Universal manipulation interface: In-the-wild robot teaching without in-the-wild robots. arXiv preprint arXiv:2402.10329, 2024.

[11] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Rescaling egocentric vision: Collection, pipeline and challenges for EPIC-KITCHENS-100. International Journal of Computer Vision, 130(1):33–55, 2022.

[12] Kunyuan Deng, Yi Wang, and Lap-Pui Chau. Egocentric human-object interaction detection: A new benchmark and method. Expert Systems with Applications, 300:130216, 2026.

[13] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, et al. Datacomp: In search of the next generation of multimodal datasets. Advances in Neural Information Processing Systems, 36:22046–22078, 2023.

[14] Generalist AI Team. GEN-0: Embodied foundation models that scale with physical interaction. https://generalistai.com/blog/nov-04-2025-GEN-0, 2025. Accessed: 2026-04-16.

[15] Generalist AI Team. GEN-1: Scaling embodied foundation models to mastery. https://generalistai.com/blog/apr-02-2026-GEN-1, 2025. Accessed: 2026-04-16.

[16] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4D: Around the world in 3,000 hours of egocentric video. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 18995–19012, 2022.

[17] Joey Hejna, Chethan Anand Bhateja, Yichen Jiang, Karl Pertsch, and Dorsa Sadigh. ReMix: Optimizing data mixtures for large scale imitation learning. In Pulkit Agrawal, Oliver Kroemer, and Wolfram Burgard, editors, Proceedings of The 8th Conference on Robot Learning, volume 270 of Proceedings of Machine Learning Research, pages 145–164. PMLR, 06–09 Nov 2025. URL https://proceedings.mlr.press/v270/hejna25a.html.

[18] Ryan Hoque, Peide Huang, David J. Yoon, Mouli sivapurapu, and Jian Zhang. EgoDex: Learning dexterous manipulation from large-scale egocentric video. In The Fourteenth International Conference on Learning Representations, 2026. URL https://openreview.net/forum?id=FFxkFMU89E.

[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. In 2025 IEEE International Conference on Robotics and Automation (ICRA), pages 13226–13233. IEEE, 2025.

[20] Marion Lepert, Jiaying Fang, and Jeannette Bohg. Phantom: Training robots without robots using only human videos. In 9th Annual Conference on Robot Learning, 2025. URL https://openreview.net/forum?id=BTUioBmCWo.

[21] Marion Lepert, Jiaying Fang, and Jeannette Bohg. Masquerade: Learning from in-the-wild human videos using data-editing. In IEEE International Conference on Robotics and Automation (ICRA), 2026.

[22] Haoyun Li, Ivan Zhang, Runqi Ouyang, Xiaofeng Wang, Zheng Zhu, Zhiqin Yang, Zhentao Zhang, Boyuan Wang, Chaojun Ni, Wenkang Qin, et al. MimicDreamer: Aligning human and robot demonstrations for scalable vla training. arXiv preprint arXiv:2509.22199, 2025.

[23] Qixiu Li, Yu Deng, Yaobo Liang, Lin Luo, Lei Zhou, Chengtang Yao, Lingqi Zeng, Zhiyuan Feng, Huizhi Liang, Sicheng Xu, et al. VITRA: Scalable vision-language-action model pretraining for robotic manipulation with real-life human activity videos. In IEEE International Conference on Robotics and Automation (ICRA), 2026.

[24] Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, and Li Yi. HOI4D: A 4D egocentric dataset for category-level human-object interaction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 21013–21022, 2022.

[25] Hao Luo, Yicheng Feng, Wanpeng Zhang, Sipeng Zheng, Ye Wang, Haoqi Yuan, Jiazheng Liu, Chaoyi Xu, Qin Jin, and Zongqing Lu. Being-H0: vision-language-action pretraining from large-scale human videos. arXiv preprint arXiv:2507.15597, 2025.

[26] Hao Luo, Ye Wang, Wanpeng Zhang, Sipeng Zheng, Ziheng Xi, Chaoyi Xu, Haiweng Xu, Haoqi Yuan, Chi Zhang, Yiqing Wang, et al. Being-H0.5: Scaling human-centric robot learning for cross-embodiment generalization. arXiv preprint arXiv:2601.12993, 2026.

[27] NVIDIA. Cosmos-embed1-448p: Vision-language embedding model for multimodal representation learning. Technical Report, 2024. URL https://www.nvidia.com.

[28] NVIDIA, Johan Bjorck, Nikita Cherniadev Fernando Castañeda, Xingye Da, Runyu Ding, Linxi "Jim" Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, Joel Jang, Zhenyu Jiang, Jan Kautz, Kaushil Kundalia, Lawrence Lao, Zhiqi Li, Zongyu Lin, Kevin Lin, Guilin Liu, Edith Llontop, Loic Magne, Ajay Mandlekar, Avnish Narayan, Soroush Nasiriany, Scott Reed, You Liang Tan, Guanzhi Wang, Zu Wang, Jing Wang, Qi Wang, Jiannan Xiang, Yuqi Xie, Yinzhen Xu, Zhenjia Xu, Seonghyeon Ye, Zhiding Yu, Ao Zhang, Hao Zhang, Yizhou Zhao, Ruijie Zheng, and Yuke Zhu. GR00T N1: An open foundation model for generalist humanoid robots. In ArXiv Preprint, March 2025.

[29] Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3D with transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 9826–9836, 2024.

[30] Physical Intelligence. Bringing general-purpose AI to the physical world. https://www.pi.website, 2026. Accessed: 2026-04-16.

[31] Ryan Punamiya, Dhruv Patel, Patcharapong Aphiwetsa, Pranav Kuppili, Lawrence Y Zhu, Simar Kareer, Judy Hoffman, and Danfei Xu. EgoBridge: Domain adaptation for generalizable imitation from egocentric human data. In Human to Robot: Workshop on Sensorizing, Modeling, and Learning from Humans, 2025.

[32] Ri-Zhao Qiu, Shiqi Yang, Xuxin Cheng, Chaitanya Chawla, Jialong Li, Tairan He, Ge Yan, David J Yoon, Ryan Hoque, Lars Paulsen, et al. Humanoid policy ~ Human policy. arXiv preprint arXiv:2503.13441, 2025.

[33] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Proceedings of the 38th International Conference on Machine Learning, pages 8748–8763. PMLR, 2021.

[34] 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. URL https://arxiv.org/abs/2408.00714.

[35] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. ACM Transactions on Graphics, (Proc. SIGGRAPH Asia), November 2017. URL http://doi.acm.org/10.1145/3130800.3130883.

[36] AI Ropedia. Xperience-10M. URL https://huggingface.co/datasets/ropedia-ai/xperience-10m, 2026.

[37] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Sasanka Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5B: An open large-scale dataset for training next generation image-text models. Advances in Neural Information Processing Systems, 35:25278–25294, 2022.

[38] Dandan Shan, Jiaqi Geng, Michelle Shu, and David Fouhey. Understanding human hands in contact at internet scale. 2020.

[39] Modi Shi, Shijia Peng, Jin Chen, Haoran Jiang, Yinghui Li, Di Huang, Ping Luo, Hongyang Li, and Li Chen. EgoHumanoid: Unlocking in-the-wild loco-manipulation with robot-free egocentric demonstration, 2026. URL https://arxiv.org/abs/2602.10106.

[40] Linda B. Smith, Chen Yu, and Alfredo F. Pereira. Not your mother's view: the dynamics of toddler visual experience. Developmental Science, 14(1):9–17, 2011. doi: https://doi.org/10.1111/j.1467-7687.2009.00947.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-7687.2009.00947.x.

[41] Tesla, Inc. Tesla AI and Robotics. https://www.tesla.com/AI, 2026. Accessed: 2026-04-16.

[42] Quan Vuong, Sergey Levine, Homer Rich Walke, Karl Pertsch, Anikait Singh, Ria Doshi, Charles Xu, Jianlan Luo, Liam Tan, Dhruv Shah, Chelsea Finn, Max Du, Moo Jin Kim, Alexander Khazatsky, Jonathan Heewon Yang, Tony Z. Zhao, et al. Open X-Embodiment: Robotic learning datasets and RT-X models. In Towards Generalist Robots: Learning Paradigms for Scalable Skill Acquisition @ CoRL2023, 2023. URL https://openreview.net/forum?id=zraBtFgxT0.

[43] Bowen Wen, Matthew Trepte, Joseph Aribido, Jan Kautz, Orazio Gallo, and Stan Birchfield. FoundationStereo: Zero-shot stereo matching. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 5249–5260, 2025.

[44] Haoran Yang, Jiacheng Bao, Yucheng Xin, Haoming Song, Yuyang Tian, Bin Zhao, Dong Wang, and Xuelong Li. ZeroWBC: Learning natural visuomotor humanoid control directly from human egocentric video. arXiv preprint arXiv:2603.09170, 2026.

[45] Ruihan Yang, Qinxi Yu, Yecheng Wu, Rui Yan, Borui Li, An-Chieh Cheng, Xueyan Zou, Yunhao Fang, Xuxin Cheng, Ri-Zhao Qiu, et al. EgoVLA: Learning vision-language-action models from egocentric human videos. arXiv preprint arXiv:2507.12440, 2025.

[46] Chen Yu and Linda B. Smith. Embodied attention and word learning by toddlers. Cognition, 125(2):244–262, 2012. ISSN 0010-0277. doi: https://doi.org/10.1016/j.cognition.2012.06.016. URL https://www.sciencedirect.com/science/article/pii/S0010027712001369.

[47] Justin Yu, Yide Shentu, Di Wu, Pieter Abbeel, Ken Goldberg, and Philipp Wu. EgoMI: Learning active vision and whole-body manipulation from egocentric human demonstrations. arXiv preprint arXiv:2511.00153, 2025.

[48] Yu Zhang, Yuqi Xie, Huihan Liu, Rutav Shah, Michael Wan, Linxi Fan, and Yuke Zhu. SCIZOR: A self-supervised approach to data curation for large-scale imitation learning. In IEEE International Conference on Robotics and Automation (ICRA), 2026. To appear.

[49] Tony Z. Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning Fine-Grained Bimanual Manipulation with Low-Cost Hardware. In Proceedings of Robotics: Science and Systems, Daegu, Republic of Korea, July 2023. doi: 10.15607/RSS.2023.XIX.016.

[50] Zhaxizhuoma, Kehui Liu, Chuyue Guan, Zhongjie Jia, Ziniu Wu, Xin Liu, Tianyu Wang, Shuai Liang, Pengan Chen, Pingrui Zhang, Haoming Song, Delin Qu, Dong Wang, Zhigang Wang, Nieqing Cao, Yan Ding, Bin Zhao, and Xuelong Li. FastUMI: A scalable and hardware-independent universal manipulation interface with dataset. In 9th Annual Conference on Robot Learning, 2025. URL https://openreview.net/forum?id=RUSscFSEfD.

[51] Ruijie Zheng, Dantong Niu, Yuqi Xie, Jing Wang, Mengda Xu, Yunfan Jiang, Fernando Castañeda, Fengyuan Hu, You Liang Tan, Letian Fu, et al. EgoScale: Scaling dexterous manipulation with diverse egocentric human data. arXiv preprint arXiv:2602.16710, 2026.