

Ego360: A Panoramic View of Human Experience for Physical Intelligence

DeepCybo & Insta360



Figure 1. **Ego360 overview.** Top: demonstrations across factories, homes, laboratories, and retail settings with multimodal supervision. Bottom: unlike fixed-FOV capture, neck-worn 360° capture retains panoramic context, body pose, hand–object interaction, and gaze for post-hoc view selection.

Abstract

Embodied intelligence models must jointly learn environmental understanding, task intent, spatial movement, and object manipulation from real human behavior. However, existing egocentric datasets are typically constrained by a fixed field of view: orienting the camera toward a local workspace can discard the surrounding environment and future targets, whereas orienting it toward the broader scene can move the body, hands, and contact events outside the frame. We introduce Ego360, a large-scale panoramic human-demonstration dataset for embodied intelligence. Using a lightweight neck-worn 360° camera, Ego360 con-

tinuously records the surrounding environment, visible body regions, hand–object interactions, and operator speech on a shared timeline, turning an irreversible viewpoint choice at capture time into a configurable observation after capture. Ego360 contains 10,000 hours of real-world task demonstrations from over 1,000 participants across homes, industrial facilities, commercial environments, and biochemical laboratories. We further develop a privacy-preserving multimodal pipeline that converts continuous panoramic video into task instructions, hierarchical action steps, speech-aligned structured descriptions, Ego-Element Graphs, and spatial-reasoning question–answer pairs, while recovering

whole-body and hand motion, episode-level 3D scene reconstructions, and camera trajectories. These temporal and spatial signals connect room-scale navigation, target search, and approach with local hand-object manipulation in continuous task records. We systematically characterize Ego360 in terms of data coverage, panoramic observability, annotation consistency, geometric integrity, and long-horizon navigation-manipulation structure, providing a scalable and model-agnostic foundation for vision-language-action pretraining, world models, and learning robotic behavior from human demonstrations.

1. Introduction

Progress in embodied intelligence increasingly depends on supervision that is physically grounded, behaviorally diverse, and scalable to real-world environments. Robot teleoperation provides executable action labels, but remains costly to collect and difficult to scale across the long tail of human activities. Egocentric human video, in contrast, records naturally occurring tasks, object interactions, and behavioral context with substantially lower collection overhead, making it an increasingly important source of supervision for embodied model pretraining. Ego4D established the scale and diversity of first-person perception, while more recent efforts such as EgoLive and EgoVerse have begun to investigate egocentric human data as a substrate for robot learning [10, 20, 30]. Recent approaches increasingly train VLA models from action-bearing supervision or staged visual-action pretraining derived from human egocentric video [18, 19, 33, 35, 41]. More directly, PhysBrain and PhysBrain 1.0 demonstrate that structured supervision distilled from human egocentric video can strengthen first-person physical reasoning and transfer to downstream VLA control [21, 24]. Human-as-Humanoid further shows that synchronized human ego-exo demonstrations can be converted into controller-aligned action supervision for high-DoF humanoid VLA training [26]. This transition imposes a stronger requirement on the data: a reusable human demonstration must preserve not only what the wearer observes, but also how the wearer moves, interacts with objects, and acts within the surrounding environment.

The central limitation of current egocentric data is no longer scale alone, but incomplete observability. Many scalable capture systems rely on one or two forward-facing RGB, wide-angle, or fisheye cameras and use the forward stream as the primary learning view. Regardless of resolution, a finite forward field of view imposes an acquisition-time trade-off. Orienting the camera toward the workspace preserves environmental context and task targets but may truncate the torso, arms, or hands; orienting it downward improves coverage of the body and nearby hand-object interactions but removes the broader workspace, future tar-

gets, and surrounding context. Head rotations and body motion can further move previously visible hands, objects, or interaction regions outside the frame. Because this information is lost during capture, subsequent annotations or model-based inference can only estimate the missing content; they cannot restore the absent visual evidence. Consequently, a video may remain semantically informative while being geometrically insufficient for reliable reconstruction of human motion, contact transitions, and spatial relationships.

We introduce **Ego360**, a panoramic egocentric dataset designed for embodied intelligence (Fig. 1). Its central design principle is to *retain first and select later*. Our unobtrusive neck-worn 360° system continuously records the surrounding workspace, visible body regions, and hand-object interactions on a shared timeline, substantially reducing view truncation caused by changes in camera orientation. Unlike a fixed forward-facing camera, the panoramic recording supports post-hoc rendering of arbitrary viewing directions, fields of view, and task-conditioned perspective views. Ego360 therefore converts an irreversible hardware decision made during acquisition into a configurable modeling choice made after acquisition. The panorama also provides a persistent spatial reference for gaze: when synchronized gaze is available, an attended target remains localizable in the recording even after it leaves the forward view. The value of panoramic capture is thus not merely a wider image, but a more complete representation of behavior that preserves the correspondence among environmental context, bodily motion, and local interaction.

Panoramic video alone does not constitute robot-ready supervision. Ego360 therefore introduces a task-centric multimodal pipeline that converts raw recordings into structured, spatially grounded, and action-aligned training signals. Each recording pairs panoramic video with synchronized operator speech, from which we derive task intent, procedural cues, and an episode-level instruction. We further represent each task as a hierarchy of temporally ordered procedural steps aligned to the shared segment timeline. The annotation stack describes the environment, depth-aware spatial relationships, fine-grained hand-object interactions, actions and contact states, and world-referenced whole-body motion, including hand motion and the global locomotion trajectory. These signals are aligned within a common temporal and spatial reference, jointly specifying what is being attempted, where the behavior takes place, and how the task is executed. Because panoramic capture expands the range of visible spaces and bystanders, Ego360 treats privacy as a first-class design constraint rather than a post-hoc release step. All recordings follow an explicit informed-consent protocol and are collected only in environments approved by their owners or authorized managers. Before release, every sequence undergoes a re-

Table 1. **Comparison with representative egocentric datasets.** We compare the dimensions required to turn egocentric video into supervision for physical intelligence, rather than scale alone. “Capture” distinguishes directional cameras (Dir.), synchronized multi-camera systems (Multi), and native 360° panoramic video, while “Depth” indicates released dense or structured depth supervision. ✓ denotes an explicitly released modality or primary annotation, (✓) denotes partial or indirect coverage, and – denotes that the dimension is not a primary component. Ego360 is the only large-scale dataset that combines native panoramic capture with the complete task, interaction, depth, motion, geometry, and language supervision stack.

Dataset	Scale	Capture	Task lang.	Temp. structure	HOI	Full body	Depth	World geometry	Caption + QA	Long horizon	Robot-oriented
<i>Large-scale activity and interaction understanding</i>											
EPIC-KITCHENS-100 [7]	100 h	Dir.	✓	✓	(✓)	–	–	(✓)	–	✓	–
Ego4D [10]	3,670 h	Dir.	✓	✓	✓	–	(✓)	(✓)	(✓)	✓	–
HoloAssist [32]	166 h	Dir.	✓	✓	✓	–	✓	(✓)	–	(✓)	–
<i>Geometric and multiview egocentric understanding</i>											
Ego-Exo4D [11]	1,286 h	Multi	✓	✓	✓	✓	(✓)	✓	(✓)	✓	–
Nymeria [28]	300 h	Multi	✓	✓	(✓)	✓	(✓)	✓	(✓)	✓	–
Ego-1K [17]	956 clips	Multi	–	–	✓	–	–	✓	–	–	–
<i>Egocentric human data for robot learning</i>											
EgoScale [41]	20,854 h	Dir.	✓	✓	✓	–	–	(✓)	–	✓	✓
EgoVerse [30]	1,362 h	Dir.	✓	✓	✓	–	–	(✓)	–	✓	✓
<i>Panoramic egocentric datasets</i>											
EgoK360 [2]	~23 h	360°	–	(✓)	(✓)	–	–	–	–	(✓)	–
Ego360 (ours)	10,000 h	360°	✓	✓	✓	✓	✓	✓	✓	✓	✓

view for personally identifiable information (PII): all visible faces are blurred, other identifying content is removed or masked where necessary, and only sequences that pass the de-identification audit are included in the public dataset.

As VLA training continues to explore different mechanisms for adapting pretrained multimodal representations to robot control [15, 22, 36, 38], the initial release of Ego360 is deliberately scoped as a data and infrastructure contribution. Rather than tying the value of the dataset to a particular vision–language–action (VLA) architecture, training recipe, or robot platform, this work establishes the observational foundation, supervision structure, and quality criteria required to study panoramic human demonstrations. We characterize Ego360 at the data level in terms of task and scene coverage, panoramic observability, annotation consistency, and the geometric integrity of world-referenced motion. We further evaluate the sensing advantage through a matched-view protocol that renders conventional forward-facing views from the same panoramic recordings and measures scene coverage, body and hand retention, and the temporal continuity of hand–object evidence. This protocol isolates the information retained by panoramic capture without introducing confounding differences from downstream models. The release provides standardized data schemas, splits, metadata, quality-control procedures, and privacy safeguards, establishing a reproducible and model-agnostic foundation for future studies of panoramic human data in VLA pretraining, world modeling, and human-to-robot transfer.

Our contributions are threefold:

- **A panoramic egocentric data resource.** We release

Ego360 with 100 hours, 22000+ task episodes, 480 task categories, and 36 real-world scenes, jointly preserving panoramic observations, task-oriented speech, hand–object interactions, and world-referenced human motion.

- **A structured pipeline for embodied learning.** We develop a unified pipeline that transforms raw 360° video and speech into task instructions, hierarchical subtasks, environment and spatial descriptions, hand–object interaction annotations, and world-referenced human motion.
- **A model-agnostic and privacy-conscious release protocol.** We provide standardized data schemas, splits, metadata, and quality-control procedures, together with informed consent, environment authorization, face blurring, PII review, and pre-release de-identification. Matched-view analyses quantify the observational information retained by panoramic capture relative to bounded forward-facing views.

2. Related Work

2.1. Egocentric Human-Activity Datasets

Egocentric datasets have progressed from activity-centric benchmarks to geometrically grounded interaction data and, more recently, robot-oriented human demonstrations. EPIC-KITCHENS and Ego4D scale action understanding to unscripted kitchens and open-world daily activities [7, 10]. Ego-Exo4D, HOT3D, and Nymeria introduce synchronized external views, 3D hand–object geometry, or world-referenced body motion [1, 11, 28], whereas EgoDex and OpenEgo provide hand trajectories or standardized action primitives for dexterous learning [13, 14]. EgoScale, Ego-

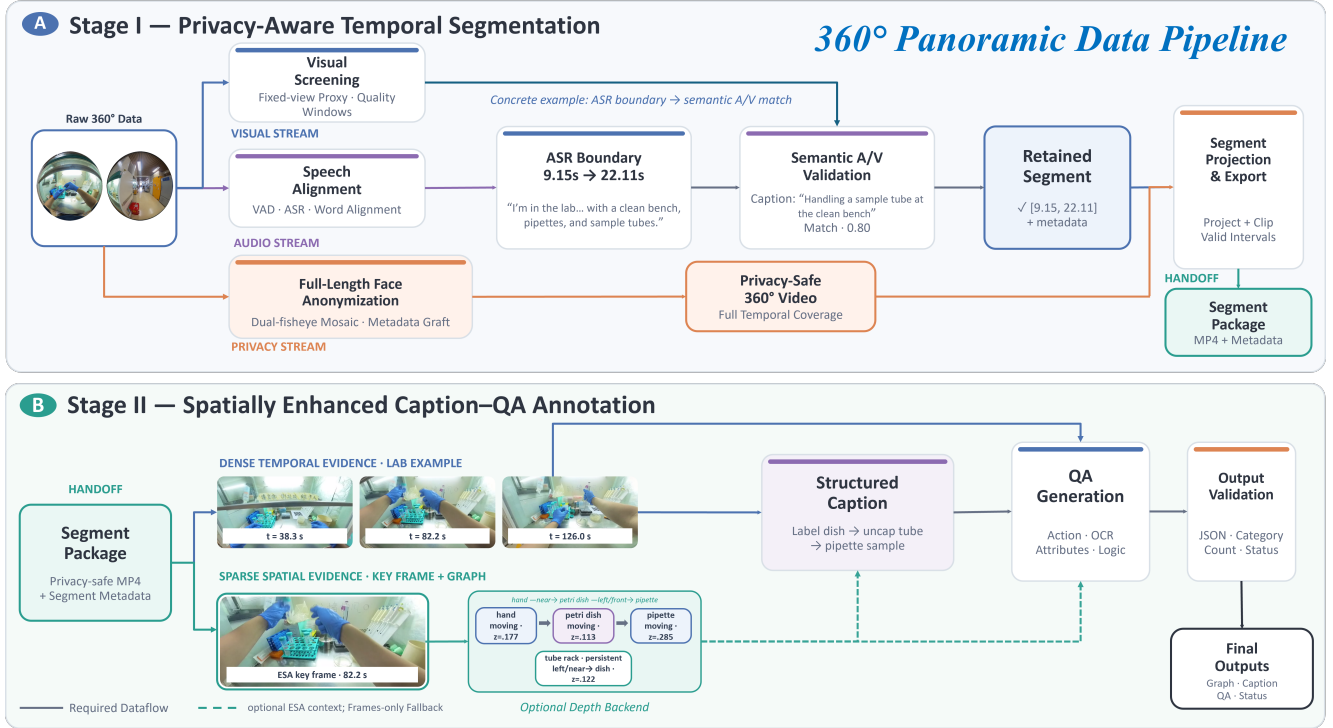


Figure 2. **Ego360 data pipeline.** Stage I extracts privacy-safe action segments from synchronized panoramic video and speech. Stage II combines temporal frames with spatial evidence to generate Ego-Element Graphs, structured captions, and grounded QA.

Live, and EgoVerse further investigate data scaling and diversity for human-video pretraining [20, 30, 41].

Despite this progress, most datasets retain a fixed view. Forward-facing RGB, stereo, or fisheye cameras can preserve the workspace, but often exclude the wearer’s body or portions of hand–object interaction; orienting the cameras downward introduces the opposite loss of environmental context. Synchronized exocentric views mitigate this trade-off but require instrumented environments. Ego360 instead retains an omnidirectional record and supports task-dependent view generation after capture. It aligns panoramic video with task speech, hierarchical subtasks, environment and depth descriptions, hand–object interactions, and world-coordinate whole-body and hand motion with global trajectories, providing spatially and physically grounded human demonstrations for robot learning.

2.2. Omnidirectional and Panoramic Datasets

Omnidirectional datasets span static scene understanding, spatial reasoning, and agent-centered perception. SUN360, PanoContext, and Matterport3D support panoramic scene recognition, layout estimation, and 3D reconstruction [4, 34, 40]. Pano-AVQA and VSI-Super-Wild study audiovisual grounding and long-horizon spatial reasoning in panoramic video [12, 37]. Agent-centered resources include the robot-mounted JRDB, the wearable activity

dataset EgoK360, and the synchronized surround-view dataset Ego-1K [2, 17, 29]. A recent survey systematizes the geometric challenges of panoramic vision, while DAP, DA360, and PanoLOG advance panoramic depth estimation and scene reconstruction [5, 16, 23, 25]; AirSim360 and Fly360 further investigate omnidirectional sensing through spatially annotated simulation with automated trajectories and obstacle-aware navigation [9, 39].

These datasets establish the value of wide-field sensing, but typically isolate scene semantics, passive perception, social sensing, activity classification, or reconstruction. They do not jointly represent the acting subject’s intent, body dynamics, hand–object interaction, and surrounding task context. Ego360 places the camera on the acting subject and organizes each continuous recording as a task demonstration, using speech, subtask structure, spatial descriptions, and world-coordinate motion to connect intent with physical execution. Its contribution is therefore not panoramic capture alone, but a task-structured and physically grounded representation of human experience for robot learning.

2.3. Capture Paradigms for Embodied Demonstrations

Embodied demonstration systems balance action fidelity, contextual completeness, and scalable deployment. Syn-

chronized ego-exo systems such as Ego-Exo4D complement wearable views with external cameras, enabling strong cross-view and 3D body-hand supervision but requiring instrumented capture environments [11]. Head-mounted multisensor platforms such as HoloAssist and HOT3D provide rich body-centric signals including depth, gaze, and hand pose, while their directional cameras retain bounded visual coverage [1, 32]. Proxy interfaces such as UMI record robot-compatible end-effector actions through tracked handheld grippers and support direct policy transfer, but require demonstrators to act through purpose-built hardware rather than capturing unconstrained human behavior [6].

These paradigms are complementary rather than interchangeable. For learning general embodied priors, the primary visual record should preserve both local manipulation and the surrounding task context before a downstream observation or embodiment is chosen. Ego360 targets a distinct operating point: lightweight native panoramic capture requires neither external scene instrumentation nor an action proxy, yet preserves environmental context, body motion, and hand-object interaction for post-hoc view selection and aligned language-motion-geometry supervision. This makes Ego360 suitable as a scalable perceptual foundation for long-horizon navigation-manipulation and embodied pretraining, while precise end-effector actions and force-tactile sensing remain valuable complementary supervision.

Table 1 positions Ego360 against representative datasets spanning activity understanding, multimodal interaction, geometric reconstruction, and human-to-robot learning. We distinguish native panoramic capture from synchronized multi-camera rigs: although both increase visual coverage, only the former retains an omnidirectional source that supports arbitrary post-hoc view selection. EgoK360 pioneered 360° first-person activity recognition, whereas Ego360 is, to our knowledge, the first large-scale panoramic egocentric dataset to jointly provide task-aligned language, temporally structured actions, fine-grained hand-object interaction, world-referenced whole-body motion, scene-scale geometry and trajectories, and structured caption-QA supervision for physical intelligence.

3. Ego360 Dataset Construction

Ego360 converts continuous panoramic recordings into privacy-preserving, task-structured, and spatially grounded demonstrations through the two-stage pipeline shown in Fig. 2. Stage I jointly leverages visual quality evidence, synchronized narration, and audio-visual semantics to identify effective action intervals, while applying sequence-level privacy protection before data export. Stage II enriches each curated segment with complementary temporal and spatial evidence, producing structured captions, Ego-Element

Graphs, and question-answer annotations. Throughout the pipeline, temporal provenance and processing status are retained to support systematic quality control and traceable data release.

3.1. Collection Scope and Capture Protocol

Ego360 records naturally executed, task-oriented demonstrations across factories, homes, laboratories, and retail environments. Each participant wears a dual-fisheye 360° camera with synchronized audio, allowing the same recording to preserve the surrounding workspace, visible body regions, hand-object interactions, and operator narration on a shared timeline. Compared with a fixed forward-facing view, this capture protocol avoids committing to a single observation direction during acquisition and supports task-dependent perspective projection after recording.

The collection spans local object manipulation, navigation, target search, device interaction, and multi-stage activities performed in realistic environments. Recording sessions are organized by participant, location, and task, with calibration, synchronization, and capture-quality checks applied consistently across sessions. Operators provide natural task-oriented narration that conveys intent and procedural cues without constraining how the physical action is executed. All participants follow an informed-consent protocol, and collection environments are recorded with authorization from their owners or designated managers. Dataset-scale statistics and the resulting diversity of participants, scenes, and tasks are characterized in Sec. 5.

3.2. Privacy-Aware Temporal Segmentation

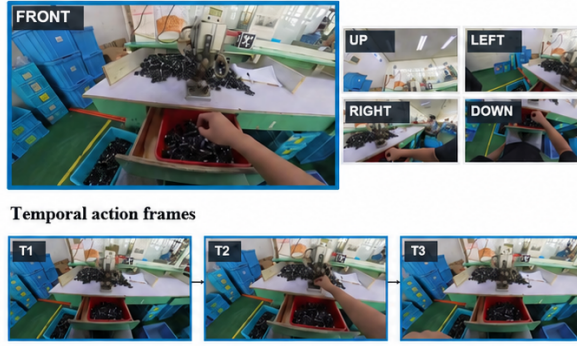
Raw wearable recordings naturally contain task execution, transitions between activities, conversational speech, and periods with limited visual utility. Directly partitioning such recordings using a single modality can therefore produce incomplete or semantically ambiguous clips. Ego360 adopts a multi-signal temporal curation process that combines visual screening, narration-guided boundary discovery, semantic audio-visual validation, and full-sequence privacy protection.

Visual screening and speech alignment. We first derive a fixed-view proxy from each dual-fisheye recording for efficient visual screening while preserving the original panoramic stream as the canonical data source. The screening process identifies usable temporal windows and filters severe occlusion, capture failures, and other visually degraded content. In parallel, we extract the synchronized audio and apply voice activity detection, automatic speech recognition, and word-level temporal alignment. Neighboring utterances are consolidated into narration units, which provide semantic anchors for locating task instructions and their associated action intervals.

Narration-guided interval construction. Human narra-

(a) Factory

Single 360° frame · five synchronized directions



Holistic environment

Electronics assembly line with press stations, green flooring and blue storage crates; peripheral views recover neighboring workstations beyond the manipulation zone.

Cross-view evidence

Front: die alignment and hand motion ·
Left/right: adjacent workstations and crates ·
Down: transfer into the red bin · Up: safety signage and lighting.

Action annotation

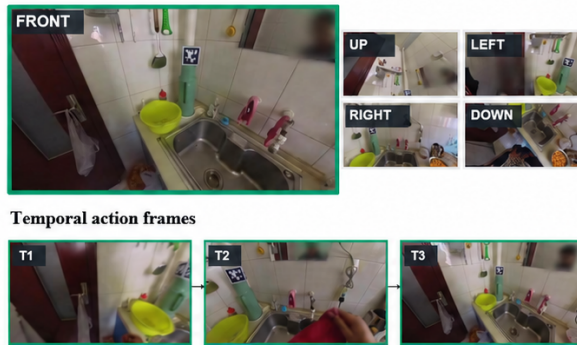
Pick component → align under press → process → deposit in red bin.

QA supervision

Q: Where is the counter? A: On the top-front of the press, visible in front and left views.

(b) Home

Single 360° frame · five synchronized directions



Holistic environment

Compact tiled kitchen around a double sink; side views expose the door, basin, kettle and stove while the lower view captures the work surface.

Cross-view evidence

Front/bottom: object transfers through the sink ·
Left/right: counters and appliances · Up: mirror reflection confirms actor position and head motion.

Action annotation

Wash cutting board → knife → pink rag; then turn off the faucet.

QA supervision

Q: Washing order? A: Cutting board → kitchen knife → pink rag.

Figure 3. **Panoramic multimodal annotations in factory and home tasks.** Five synchronized directions from a single panoramic frame are paired with temporal action frames to support holistic environment descriptions, cross-view evidence, procedurally ordered action annotations, and grounded QA. Retail and laboratory examples are provided in Fig. 4.

tion provides informative but imperfect action boundaries: an operator may describe an action before executing it, combine multiple steps in one utterance, or produce speech unrelated to the visible task. We therefore treat aligned narration as a temporal proposal rather than a final segmentation. Neighboring speech segments separated only by a short pause are first merged into a narration unit. Each unit records the narration onset (`start_sec`) and completion time (`instruction_end_sec`), and provisionally extends to the onset of the next narration unit to include the subsequent physical execution. The retained endpoint (`end_sec`) distinguishes the action interval from the spoken instruction itself. These candidate intervals are then reconciled with the visual-quality windows to remove unusable portions while preserving the context required to interpret the complete action.

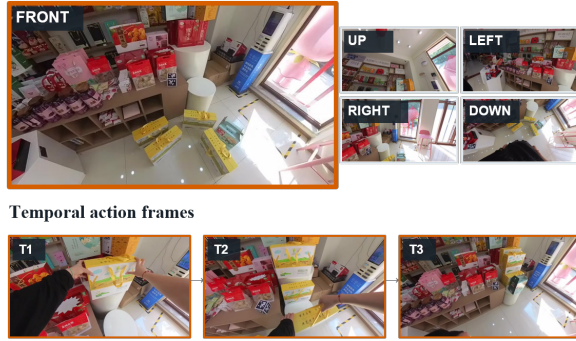
Semantic audio-visual validation. For each candidate interval, the pipeline jointly examines its aligned transcript and visual content to derive a normalized action description, identify the speech type, and assess audio-visual rele-

vance. Speech is distinguished as an operation instruction, incidental conversation, recording-control speech, or an uncertain transcription. Audio-visual relevance is represented by a score and a normalized label—`match`, `mismatch`, or `not_applicable`—while the original decision and its supporting reason are retained for audit. Importantly, the normalized description does not overwrite the original ASR text. The resulting record binds both forms of text, the three temporal boundaries, and the validation fields to stable source, recording, and segment identifiers. Only intervals satisfying the visual, semantic, and temporal quality criteria proceed to multimodal annotation.

Full-sequence privacy protection. Privacy protection is applied to the complete panoramic recording rather than only to selected views or exported clips. Visible faces are detected and blurred throughout the sequence. Screens, identity documents, addresses, and other potentially identifying content are additionally reviewed and masked or removed when necessary. Processing produces a separate anonymized derivative rather than overwriting the source

(c) Retail

Single 360° frame · five synchronized directions



Temporal action frames

Holistic environment

Bright gift shop with display shelves, entrance glazing and floor-level boxes; multiple directions recover the merchandising context around the stool.

Cross-view evidence

Down/front: floor-to-stool transfer · Left/right: product shelves and entrance · Up: ceiling lights and upper displays.

Action annotation

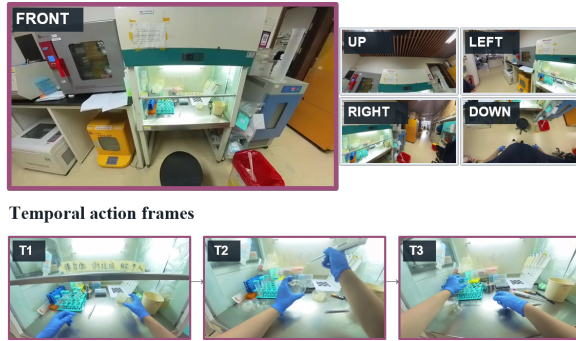
Move yellow and green gift boxes from the floor to a stable stack on the stool.

QA supervision

Q: Why stack the boxes? A: To create a visible, organized product display.

(d) Laboratory

Single 360° frame · five synchronized directions



Temporal action frames

Holistic environment

Biological laboratory centered on a laminar-flow hood; side, upper and lower views reveal adjacent equipment, ceiling, stool and waste bins.

Cross-view evidence

Front: sterile liquid handling · Left/right: adjacent lab equipment and aisle · Down: seated posture · Up: hood header and ceiling alignment.

Action annotation

Label agar dish → aspirate from tube → dispense sample under sterile airflow.

QA supervision

Q: Why use the hood? A: Filtered airflow protects open samples from contamination.

Figure 4. **Panoramic multimodal annotations in retail and laboratory tasks.** The same synchronized directional and temporal evidence captures surrounding scene context, action progression, and grounded supervision in retail and laboratory environments. Factory and home examples are provided in Fig. 3.

recording. The panoramic metadata is preserved in the protected output, allowing validated intervals to be projected and rendered without returning to the unprocessed source. Each effective interval is finally exported as a segment package containing the privacy-safe video, aligned narration, temporal boundaries, semantic description, and processing provenance; recording-level validity reports and release manifests retain the corresponding processing outcome.

3.3. Spatially Grounded Multimodal Annotation

Each curated segment is further transformed into structured supervision that captures not only what action occurs, but also where it occurs, how it evolves, and which scene elements participate in the interaction. Stage II follows a dual-resolution evidence design: a higher-rate temporal sequence preserves action progression, while a lower-rate spatial sequence provides compact evidence for explicit scene and relation modeling. Figures 3–4 illustrate how synchronized views and temporal evidence support the released annotation modalities across the four collection domains.

Temporal evidence sampling. For caption and QA generation, we sample each segment at one frame per second. For long segments, the sequence is uniformly subsampled to at most 100 frames so that its complete temporal extent remains represented. Each sampled frame is associated with its timestamp and source-frame index. Before model inference, the frame is resized with its aspect ratio preserved such that the longer side does not exceed 512 pixels, and is then encoded as JPEG with quality 85. Caption and QA generation share this ordered frame sequence, providing a common temporal reference across the two annotation tasks.

Spatial evidence sampling. Spatial graph construction uses an independent, lower-rate sequence sampled approximately once every three seconds, with at most 30 frames distributed over the complete segment. This branch emphasizes representative observations of task-relevant objects, interaction regions, scene text, and geometric context. Decoupling spatial sampling from the higher-rate temporal pathway controls the visual and token budget without re-

ducing the temporal coverage needed to describe complete actions. It also allows the spatial representation to capture objects that persist, appear, disappear, or move during the segment.

Ego-Element Graph construction. We organize the spatial evidence as an Ego-Element Graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ centered on the camera wearer. The graph contains an explicit ego node at the observer origin and a compact set of task-relevant object and region nodes. Each node associates an element with its semantic identity, normalized image region, ego-relative position, visible text, confidence, supporting frames, and temporal state. Image regions are represented in a normalized $[0, 1000]^2$ coordinate system; ego-relative horizontal and vertical coordinates lie in $[-1, 1]$, and the depth coordinate lies in $[0, 1]$. A depth-aware branch estimates relative depth from the same spatial evidence and aggregates valid observations into the node depth coordinate. The resulting value represents normalized relative depth rather than metric distance. Directed edges encode relations such as left/right, above/below, near/far, in-front-of/behind, support, containment, and contact.

The model-generated graph is subsequently normalized into a unified schema. Object identifiers are canonicalized, regions and positions are constrained to their valid ranges, unsupported relations and invalid source-frame references are removed, and duplicate relations are merged. For every pair of distinct object nodes, the representation retains both directed edges without introducing self-loops; complementary left/right, above/below, front/behind, and near/far relations are derived consistently from the normalized positions, while observed support, containment, and contact relations are preserved. This process converts spatial cues that would otherwise remain implicit and scattered across frames into an explicit, temporally supported relational representation.

Structured caption generation. The complete temporal frame sequence and the Ego-Element Graph are jointly used to generate a structured caption. The visual observations remain the primary evidence, while the graph provides complementary spatial context; if the two disagree, the visual evidence takes precedence. Rather than producing a single unconstrained sentence, the output is decomposed into three groups: *scene elements*, describing the main objects and environment; *spatial dynamics*, describing the initial configuration and its changes; and *action execution*, describing the instruction, interaction process, contact transitions, object motion, and resulting state changes. Each segment is annotated with a procedurally structured action description that decomposes the demonstrated activity into temporally ordered steps. This representation forms a task-to-subtask hierarchy: the segment-level activity defines the overall task, while the ordered steps capture its constituent procedural units on the shared segment timeline.

Spatially grounded QA generation. Question-answer pairs are generated from the ordered frames, the structured caption, and the same spatial representation. The questions span temporal action understanding and ordering, visible text, object counting and attributes, spatial relationships, state changes, and higher-level visual reasoning. Each pair is stored as a category, question, and answer, with its category normalized into a major-subcategory form. Sharing evidence among caption and QA generation establishes a common reference for object identity, temporal order, and spatial relations, improving consistency across annotation modalities. When spatial evidence is unavailable, the pipeline retains a frames-only path so that the core temporal annotation process remains complete.

Quality assurance and output packaging. All generated outputs undergo deterministic structural checks before packaging. Graph validation checks whether a non-empty graph can be parsed and normalized; caption validation records whether the structured object can be parsed; and QA validation checks both JSON structure and the requested annotation count after category normalization. A failed or unavailable graph is recorded as disabled, skipped, empty, unparsable, or erroneous and triggers the frames-only path rather than terminating caption and QA generation. Caption and QA stages likewise distinguish successful outputs from parsing failures, while QA additionally records count mismatches. Raw responses, failure reasons, sampled-frame references, and stage-level status are retained for audit and targeted reprocessing. The final segment package aligns the privacy-safe MP4, temporal metadata, sampled-frame records, Ego-Element Graph when available, structured caption, QA annotations, and validation status under a unified segment identifier.

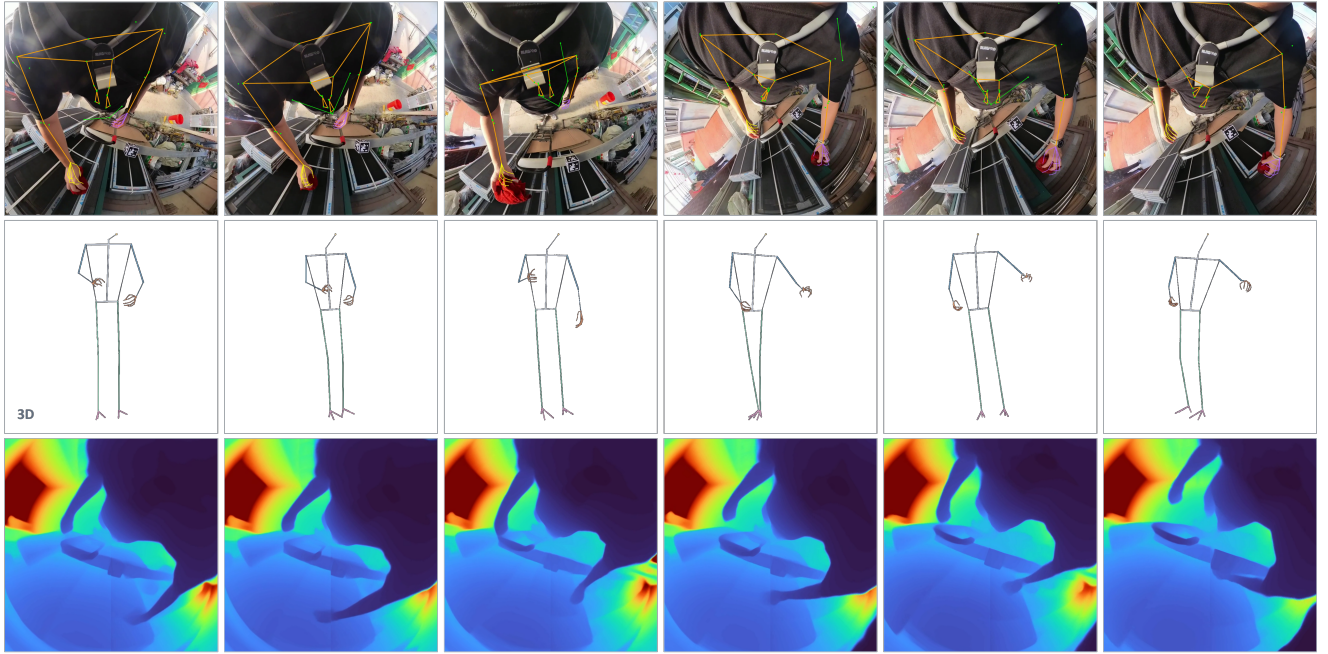
4. Geometric Grounding of Human Demonstrations

Beyond semantic annotations, Ego360 represents how the operator moves through and acts within the physical environment. We recover whole-body motion and panoramic scene geometry in compatible world-referenced coordinate systems, connecting local hand-object interactions with global locomotion and persistent scene structure.

4.1. World-Referenced Human Motion

Panoramic video preprocessing. Each demonstration is captured by a single wearable Insta360 X3 panoramic camera. We first dewarp the raw panoramic footage with Insta360 Studio and then manually crop a task-relevant monocular RGB video from the dewarped recording. This cropped RGB sequence, rather than the original distorted frames, is used as the input to the 2D and 3D human-motion models. The pipeline therefore processes a single panoramic recording without fusing pose estimates from

FACTORY ----- Equipment interaction



RETAIL ----- Shelf interaction

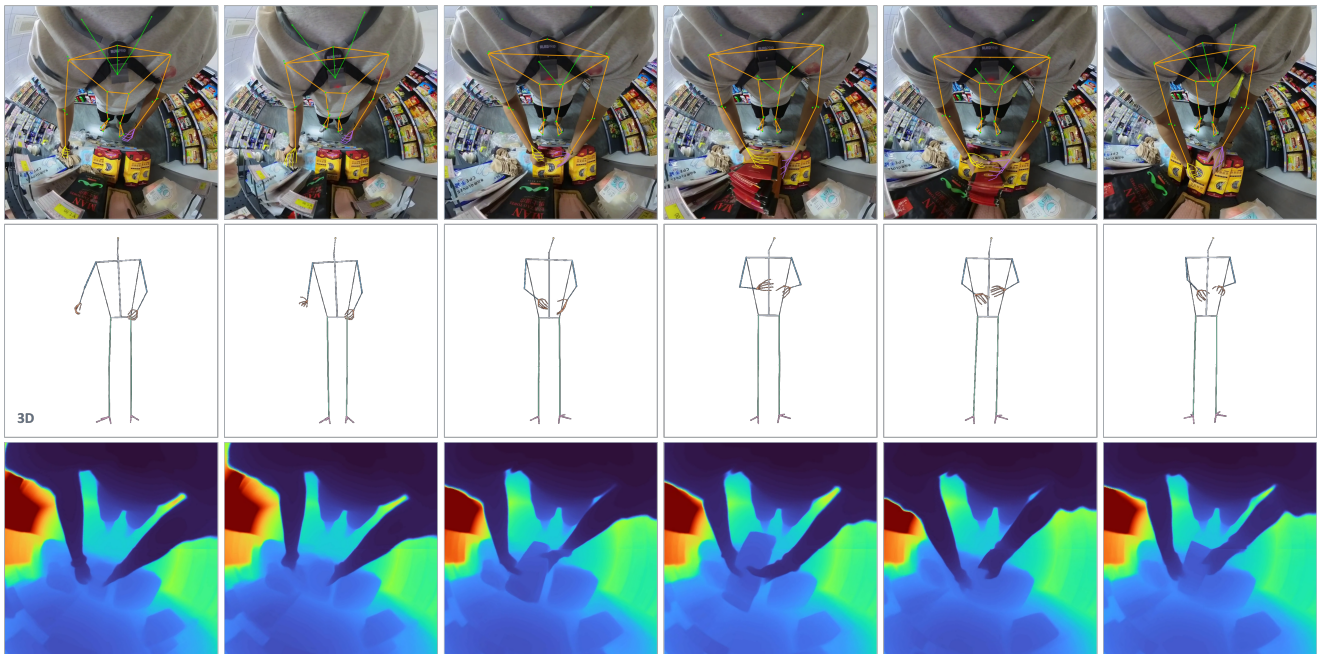


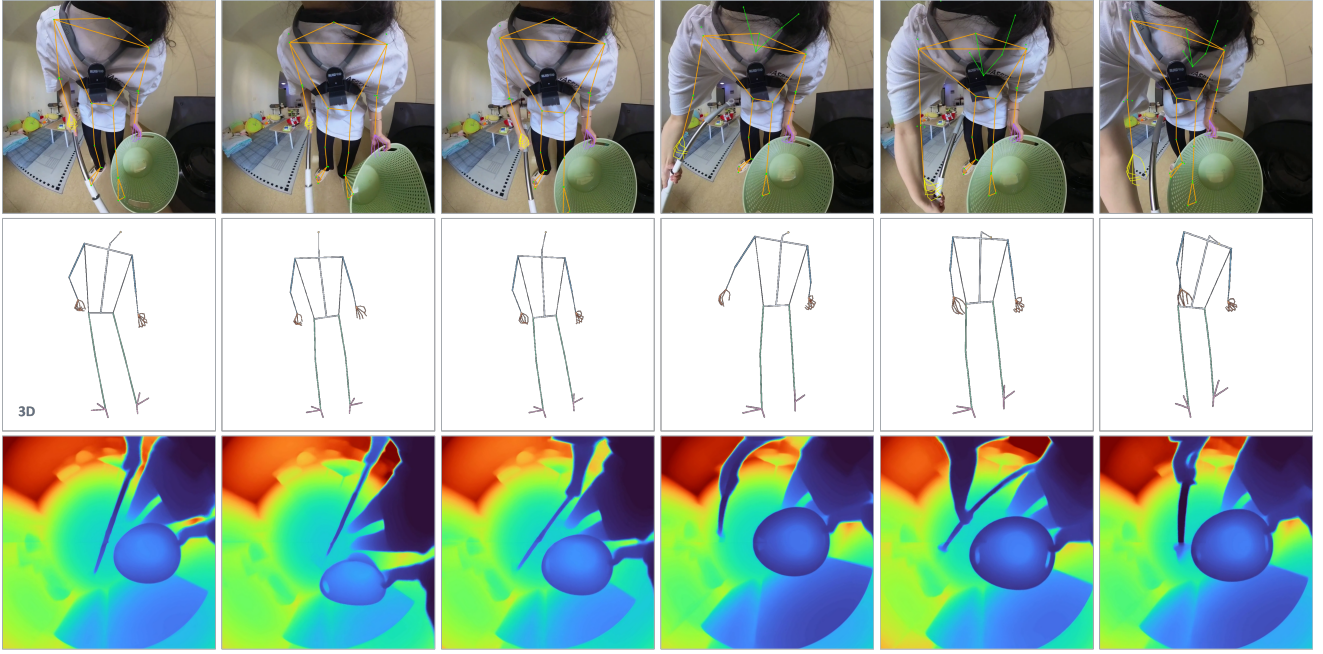
Figure 5. **Temporally aligned whole-body and hand motion in factory and retail tasks.** Perspective views with a projected 16-joint body skeleton and 21 articulated joints per hand are paired with their temporally corresponding 3D skeletons. Complementary home and laboratory examples are shown in Fig. 6.

multiple cameras or independently captured viewpoints, while every cropped frame remains linked to its source frame and timestamp.

Two-dimensional human localization. For an individual

video sequence, we use SAM 3 [3] to obtain per-frame human masks. For large-scale batch processing, we replace this mask-based procedure with YOLOv7 [31] person detections to localize the relevant image regions effi-

HOME ----- Sweep the floor



LABORATORY ----- Twist the dropper cap

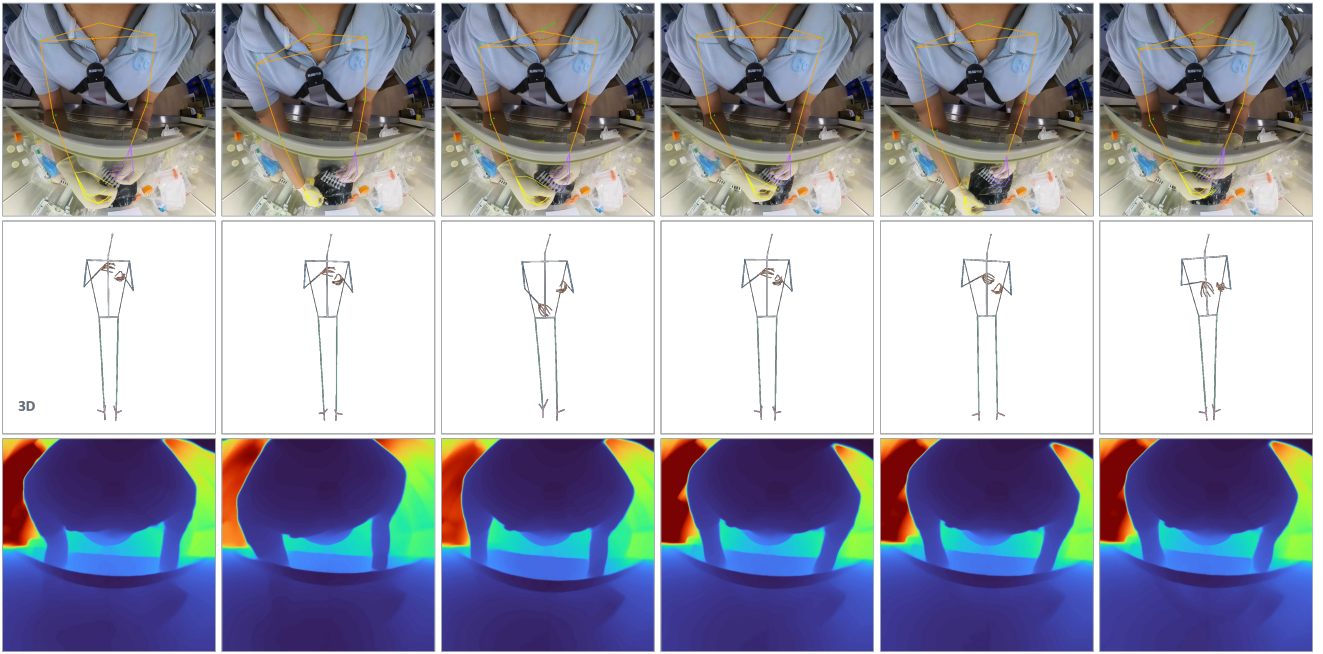


Figure 6. **Temporally aligned whole-body and hand motion in home and laboratory tasks.** Projected 2D body and hand joints remain aligned with the corresponding 3D skeletons throughout household and laboratory interactions. Complementary factory and retail examples are shown in Fig. 5.

ciently. The two models therefore define alternative processing modes rather than consecutive stages of a single pipeline. In either mode, optical-flow tracking maintains the association of the wearer across adjacent frames. The

resulting 2D localization and tracks provide spatial and temporal support for body-motion recovery; they are distinct from the projected 2D body and hand joints used to visualize the recovered skeletons.

Motion-adaptive frame interpolation. Fast hand movements are particularly susceptible to motion blur and temporal under-sampling. We therefore monitor grayscale changes inside each hand region of interest (ROI) to identify high-speed motion intervals and insert additional interpolated frames only for the affected intervals. This targeted refinement avoids uniformly increasing the frame count for portions of a demonstration that already contain sufficient temporal evidence.

Four-dimensional body recovery. Given the preprocessed and temporally refined RGB sequence, we use SAM-Body4D [8] to recover temporally consistent 3D human motion. The body is represented with SMPL [27], from which we retain a 16-joint body skeleton, while each hand is represented by 21 articulated joints. This representation preserves both gross body posture and finger-level configurations rather than reducing each hand to a wrist trajectory. Ego360 associates the projected 2D joints and the corresponding 3D skeleton on the same timeline. Figures 5–6 visualize this framewise correspondence for representative factory, retail, home, and laboratory tasks.

World-coordinate conversion and temporal smoothing. The recovered 3D body and hand motion is transformed from the camera frame into the calibrated world coordinate system using the timestamped camera poses. We then apply temporal smoothing to suppress framewise jitter while preserving rapid hand motion and contact transitions. The resulting world-referenced body and hand trajectories remain aligned with the panoramic video and reconstructed scene.

4.2. Panoramic Scene Reconstruction and Global Trajectories

For every Ego360 demonstration, we reconstruct the environment and recover the motion of the wearable camera within it. This scene branch complements the human-motion representation above with a persistent spatial reference, connecting observations across an interaction to the wearer’s movement through the environment.

Panoramic reconstruction input. Whereas the human-motion branch operates on a task-relevant monocular crop, scene reconstruction uses the complete panoramic observation to preserve the surrounding environment. We sample each episode at a configurable one-second interval and retain the source-frame index and timestamp of every sample. Each successfully registered timestamp first yields one central fisheye-camera pose. For dense reconstruction, we then derive four virtual fisheye views that share this camera center but face different directions. Denoting the central pose by $(\mathbf{R}_t, \mathbf{c}_t)$ and the directional offset of virtual view m by $\Delta\mathbf{R}_{t,m}$, the virtual cameras satisfy

$$\mathbf{c}_{t,m} = \mathbf{c}_t, \quad \mathbf{R}_{t,m} = \mathbf{R}_t \Delta\mathbf{R}_{t,m}, \quad m \in \{1, \dots, 4\}. \quad (1)$$

The four views are therefore directional observations derived from one physical camera rather than independently captured views, and all of them remain linked to the original panoramic timeline.

Episode-level camera registration. We process every episode independently with MipMap under a common reconstruction protocol. All temporally sampled observations enter aerial triangulation and structure from motion, where feature correspondences establish the view graph and the camera parameters are jointly refined. Samples without sufficient matched support are excluded during registration. The remaining observations constitute a connected SfM component with its own episode-level coordinate frame; thus, one panoramic video sequence produces one scene map and one associated camera trajectory.

Dense scene reconstruction. At the registered timestamps, the four co-located virtual fisheye views provide directional coverage for multi-view depth estimation. The resulting depth observations are filtered for cross-view consistency and fused into a common surface, while transient foreground observations are excluded so that the recovered geometry describes the persistent environment. We then project the registered visual observations onto the fused geometry to obtain a textured episode-level scene model. This sequence of fisheye registration, depth estimation, consistency filtering, surface fusion, and texture mapping is applied uniformly across the dataset.

Episode-level camera poses and trajectories. We express each reconstruction in the episode-specific world frame established by SfM, using a Z-up coordinate convention. For every registered timestamp t , the recovered camera center $\mathbf{c}_t$ and right, up, and forward basis vectors $(\mathbf{r}_t, \mathbf{u}_t, \mathbf{f}_t)$ define the camera-to-world transform

$$\mathbf{T}_{c,t}^w = \begin{bmatrix} \mathbf{r}_t & \mathbf{u}_t & \mathbf{f}_t & \mathbf{c}_t \\ 0 & 0 & 0 & 1 \end{bmatrix}. \quad (2)$$

Let $\mathcal{I}_{\text{reg}}$ denote the temporally ordered set of successfully registered timestamps. The camera centers $\mathcal{C} = \{\mathbf{c}_t \mid t \in \mathcal{I}_{\text{reg}}\}$ form the wearer’s global trajectory within the episode map, and $\mathbf{f}_t$ records the corresponding viewing direction. Because each pose retains its source-frame index and timestamp, locations along the trajectory can be mapped back to the complete panoramic observation and to any temporally aligned annotations.

Navigation–manipulation grounding. The reconstructed scene provides the persistent environmental structure, while the ordered camera poses describe how the wearer moves between interaction sites. Together, they locate room-scale transitions, target search, and approach within the same spatial record as the hand–object interaction. The timestamps connect this scene-level motion with the temporally aligned body and hand motion in Sec. 4.1, yielding a continuous episode-level representation from navigation to manipulation. Figure 7 illustrates this correspondence by placing

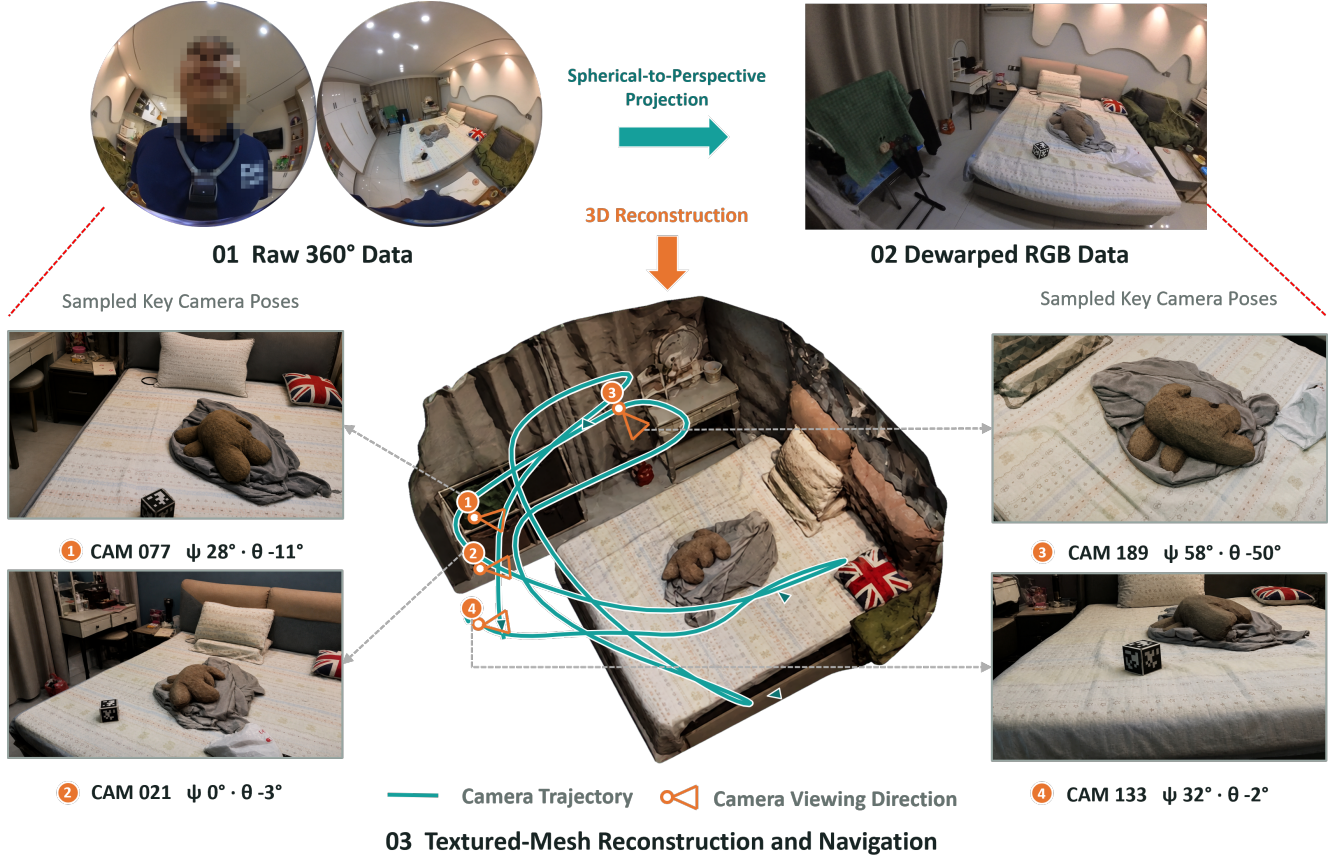


Figure 7. **Panoramic scene reconstruction and navigation–manipulation grounding.** Each episode is reconstructed independently from temporally sampled dual-fisheye video. A registered sample defines one central fisheye-camera pose, and four co-located virtual fisheye views provide directional coverage for dense reconstruction. The recovered trajectory places selected dewarped RGB observations at their corresponding positions and viewing directions in the episode-level Z-up world frame.

selected camera frusta along the recovered path and pairing them with the RGB observations captured at the same timestamps. The views make the observation–location relationship legible in a static figure, while the trajectory spans all successfully registered timestamps and supports the long-horizon analysis in Sec. 5.3.

5. Dataset Characterization

We characterize Ego360 through participant diversity, coverage, caption semantics, and long-horizon structure.

5.1. Participants and Anthropometric Diversity

Ego360 currently includes over 1,000 contributors. Figure 8(a) summarizes a sample of 50 contributors with detailed demographic and anthropometric metadata. The sample contains 30 female and 20 male contributors. Among the 43 contributors with available age metadata, 46.5% are between 20 and 29 years old, while the remaining participants span all reported age groups, from 19 years or younger

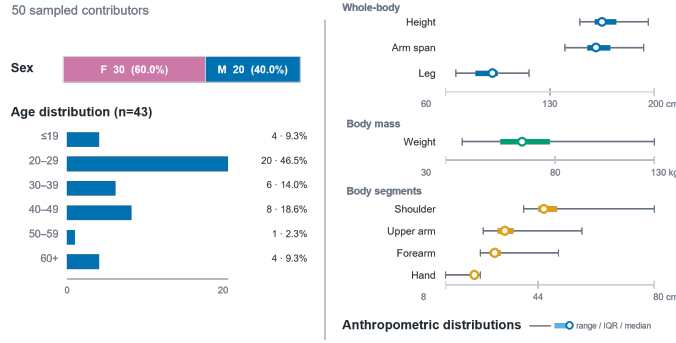
to 60 years or older. This distribution introduces variation in physical appearance, motion patterns, and task execution across different stages of adulthood.

We further report the range, interquartile range, and median of height, arm span, leg length, body mass, shoulder width, upper-arm length, forearm length, and hand length. These measurements characterize variation in whole-body scale and limb proportions rather than reducing participant diversity to demographic categories alone. Such variation directly affects reachable workspace, manipulation posture, hand placement, and the spatial relationship between the wearer and surrounding objects. Ego360 therefore captures human demonstrations across heterogeneous embodiment geometries, providing a broader basis for learning transferable motion and interaction patterns.

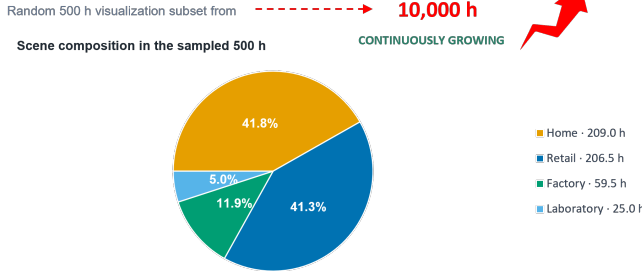
5.2. Collection, Scene, and Semantic Diversity

The Ego360 collection comprises 10,000 hours across four domains. Figure 8(b) visualizes the scene composition of a randomly sampled 500-hour subset. Home and com-

(a) Participant diversity



(b) Collection diversity



(c) Caption semantics

22,582 speech-aligned captions



Figure 8. **Ego360 dataset composition.** (a) Demographics and anthropometrics of over 1,000 contributors; (b) 10,000 hours across four scene domains and geographic scopes; (c) action, entity, scene, and attribute concepts from 22,582 speech-aligned captions.

mercial environments account for 209.0 hours (41.8%) and 206.5 hours (41.3%), respectively, while factory and laboratory environments contribute 59.5 hours (11.9%) and 25.0 hours (5.0%). The collection consequently spans everyday domestic activities, customer-facing commercial tasks, industrial procedures, and laboratory operations, covering environments with substantially different layouts, objects, safety constraints, and interaction patterns.

Figure 8(c) summarizes the semantics of 22,582 speech-aligned captions. The vocabulary is organized into four groups: actions, objects and entities, scenes and spatial anchors, and attributes and spatial states. Actions range from locomotion and object retrieval to cleaning, organization, preparation, and equipment operation. The entities include household objects, tools, furniture, garments, food items, and workplace equipment. Scene anchors such as *tabletop*, *sink interior*, *floor area*, and *bed surface* ground these actions in the environment, while attributes and spatial states describe color, material, containment, proximity, orientation, and object condition. Together, these annotations capture not only what action is performed and which object is involved, but also where the interaction occurs and how the physical state evolves.

5.3. Long-Horizon Navigation–Manipulation Episodes

Long-horizon demonstrations in Ego360 unfold across multiple interaction sites. An episode can include movement through the scene, target search and approach, local hand–object interaction, and object transport between sites, all preserved as one continuous task sequence.

Navigation and manipulation share the source–video timeline. Manipulation intervals are linked to temporally ordered action steps, contact states, and aligned body and hand motion, whereas navigation intervals retain panoramic observations, camera poses, viewing directions, and scene positions. Source timestamps connect these intervals to the hierarchical task description in Sec. 3.3 and the geometric representations in Secs. 4.1 and 4.2, preserving the transition from scene-level movement to local interaction.

Figure 9 follows a cooking episode across five stages. The wearer powers on the stove, walks to the refrigerator, opens it and retrieves dishes, carries them back, and places them on the counter. The manipulation panels show time-aligned hand meshes and trajectories at the interaction sites, while the navigation panels show timestamped egocentric waypoints along the movement. Their shared timeline preserves the procedural connection among actions performed

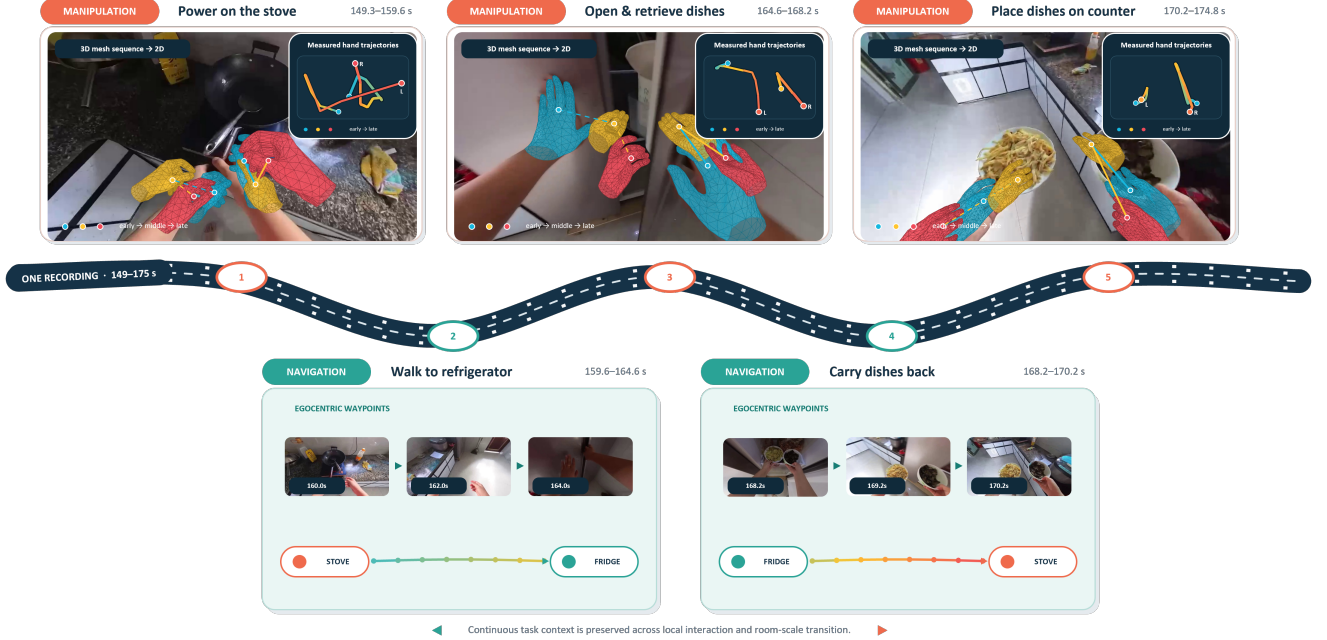


Figure 9. **Long-horizon navigation-manipulation continuity.** A continuous cooking episode alternates between three local manipulation intervals and two navigation intervals on a shared timeline. Manipulation views retain hand motion and local interaction evidence, while navigation views retain timestamped egocentric waypoints between interaction sites, connecting stove operation, refrigerator access, object retrieval, return movement, and final placement as one task.

at different locations.

This episode-level representation captures both how an interaction is performed at a particular site and how the wearer reaches the next site required by the task. It supports learning to search for and approach a target, execute a local action, and continue from the resulting state, providing supervision for navigation-conditioned manipulation, long-horizon planning, and task continuity.

6. Data-Centric Evaluation

This section evaluates Ego360 as a data resource rather than measuring its effect on a specific downstream model. We consider two complementary dimensions. First, we examine the field-of-view coverage and task evidence retained by panoramic capture relative to a conventional RGB view. Second, we evaluate whether the annotation pipeline converts continuous recordings into structured supervision without silent loss or structural inconsistency.

6.1. Panoramic Evidence Retention

We combine a controlled device-level comparison with dataset-level statistics to evaluate panoramic evidence retention. At the device level, we register an RGB image to an equirectangular panorama captured from the same location. The registered RGB footprint spans approximately $109^\circ \times 60^\circ$, whereas the panoramic source retains a nominal

$360^\circ \times 180^\circ$ field of view. Both representations preserve the forward task region, but only the panoramic source retains lateral and opposite-direction context and supports post-hoc view selection (Table 2(a)).

At the dataset level, we measure panoramic-source visibility over four scene categories, eight independent episodes, and 357 events. We first compute each metric at the event level and then macro-average over episodes. We obtain 95% confidence intervals using 10,000 episode-cluster bootstrap samples. Overall, the panoramic recordings retain task-relevant context in 100.0% of valid observations, at least one hand in 88.0%, the target object in 90.6%, and judgeable hand-object interaction in 83.4%. Action evidence covers 85.4% of each event on average, while 82.7% of events achieve at least 90% evidence coverage.

Evidence retention varies across environments. Interaction visibility reaches 90.3%, 91.1%, and 87.7% in the laboratory, home, and factory scenes, respectively, with corresponding complete-event recall rates of 91.0%, 90.0%, and 86.7%. Commercial scenes remain more challenging: interaction visibility decreases to 64.7%, while the mean maximum invisible gap increases to 3.70 seconds. This degradation indicates that dense shelving, frequent reorientation, and local occlusion can still interrupt task evidence.

Full-torso visibility remains low at 4.4%, resulting in a strict joint observability rate of 4.2%. The results therefore support a more precise conclusion: 360° capture substan-

Table 2. **Field-of-view coverage and panoramic evidence retention.** (a) A co-located RGB–360° capture characterizes the field of view of the current setup. (b) Episode-macro statistics over eight episodes and 357 events; overall values include 95% episode-cluster bootstrap confidence intervals, while scene-level values are descriptive means.

(a) Controlled device comparison.			
Property	Conventional RGB	360° source	Difference
Horizontal field of view	$\approx 109^\circ$	360°	$3.3\times$
Vertical field of view	$\approx 60^\circ$	180°	$3.0\times$
Instantaneous azimuth coverage	$\approx 30.2\%$	100%	+69.8 pp
Forward task region	Retained	Retained	–
Lateral and opposite-direction context	Outside view	Retained	360° only
Post-hoc view selection	Unsupported	Supported	Retain first, select later

The RGB footprint is estimated using SIFT matching and RANSAC registration with 85 inliers.

(b) Dataset-level panoramic evidence retention.					
Metric	Overall (8/357)	Commercial (2/125)	Home (2/49)	Factory (2/70)	Laboratory (2/113)
Task context visible $\uparrow$	100.0 [100.0, 100.0]	100.0	100.0	100.0	100.0
Any hand visible $\uparrow$	88.0 [79.6, 94.4]	77.1	91.5	88.7	94.6
Both hands visible $\uparrow$	78.8 [69.2, 87.2]	64.1	87.6	71.8	91.5
Target object visible $\uparrow$	90.6 [78.9, 98.3]	69.3	99.9	99.0	94.2
Interaction visible $\uparrow$	83.4 [72.8, 91.4]	64.7	91.1	87.7	90.3
Mean event coverage $\uparrow$	85.4 [75.5, 93.1]	68.3	91.5	89.2	92.7
Complete-event recall $\uparrow$	82.7 [71.1, 91.8]	63.2	90.0	86.7	91.0
Full torso visible $\uparrow$	4.4 [0.5, 10.2]	0.0	12.6	1.9	2.9
Strict joint observability $\uparrow$	4.2 [0.6, 9.9]	0.0	12.2	1.6	2.9
Mean maximum invisible gap $\downarrow$	1.66 [0.75, 3.01] s	3.70 s	0.79 s	1.07 s	1.07 s

All values except the last row are percentages. Complete-event recall denotes the proportion of events with at least 90% action-evidence coverage. Each scene category contains two independent episodes.

Table 3. **Annotation coverage and internal consistency over 357 events.** Results measure output completeness or automatic structural validity rather than human-evaluated semantic accuracy.

Component	Evaluated data	Check	Result
Temporal segmentation	357 events	Exact boundary transfer	357/357 (100%)
ASR	357 events	Output provided	357/357 (100%)
Caption and VLM parsing	357 events	Caption provided and parse success	100% / 100%
Question answering	357 events	Output availability	714 pairs; 100% event coverage
Visibility annotations	357 events	Output availability	357/357; 35 uncertain values as NA
Ego-Element Graph	357 events	Output availability	357/357 (100%)
Graph structural audit	357 graphs	Schema, box, source frame, and complementary relations	100% for all checks
Audio–visual audit	325 valid events	Match rate and relevance score	92.5 [88.4, 96.5] / 73.7 [67.4, 80.3]
Screen-segment QC	578 segments	Valid, review, and rejected	497 / 39 / 42

tially expands the instantaneous field of view and jointly retains environmental, hand, object, and interaction evidence for most events, but the current mounting configuration does not guarantee continuous full-body visibility.

6.2. Annotation Completeness and Internal Consistency

We further evaluate whether the annotation pipeline converts continuous recordings into structured supervision without dropping events or corrupting structural relationships. As shown in Table 3, the temporal boundaries of all 357 action events are transferred exactly to the semantic segments, and every event contains ASR, caption, QA, visibility, and Ego-Element Graph outputs. The dataset provides 714 QA pairs. All Ego-Element Graphs pass automatic checks for schema validity, bounding-box validity, source-frame references, and complementary-relation con-

sistency.

Among the 325 events with valid audio–visual judgments, the model-based match rate is 92.5%, and the mean relevance score is 73.7%. The screen-segment quality-control stage accepts 497 of 578 candidates, flags 39 for human review, and rejects 42. These results show that the pipeline consistently produces complete and structurally valid multimodal outputs while identifying and isolating lower-quality candidates.

These measurements validate pipeline completeness and internal consistency rather than semantic correctness. Because the human-audit templates have not yet been completed, we do not report caption factuality, hallucination rate, QA accuracy, or graph semantic accuracy. This distinction prevents output-generation success or structural validity from being misinterpreted as human-verified annota-

tion quality.

7. Ethics, Privacy, and Limitations

Ego360 is collected under informed consent and only in environments approved by their owners or authorized managers. Before release, every sequence is reviewed for personally identifiable information: visible faces are blurred, other identifying content is removed or masked when necessary, and only sequences that pass the de-identification audit are retained. The release will document its intended uses, licensing conditions, and data-access safeguards.

The current dataset is also subject to limitations. Panoramic imagery can contain stitching artifacts, severe distortion, motion blur, and self-occlusion, while pose and scene reconstruction may degrade under weak texture, rapid motion, or persistent occlusion. The participant, geographic, and task distributions may not represent the full diversity of real-world human activity. Finally, the initial release focuses on data quality and observational coverage; downstream VLA pretraining and human-to-robot transfer remain future evaluations.

8. Conclusion

We introduced Ego360, a panoramic egocentric dataset that preserves environmental context, human motion, hand-object interaction, task-oriented speech, and world-referenced geometry within a unified demonstration record. Its privacy-aware construction pipeline and structured multimodal supervision turn continuous 360° experience into reusable data for embodied learning. By establishing the data, annotation, and evaluation foundations first, Ego360 enables future work on VLA pretraining, world modeling, and human-to-robot transfer without tying the resource to a single model or platform.

References

- [1] Prithviraj Banerjee et al. Hot3d: Hand and object tracking in 3d from egocentric multi-view videos. *arXiv preprint arXiv:2411.19167*, 2024. 3, 5
- [2] Keshav Bhandari, Mario A. DeLaGarza, Ziliang Zong, Hugo Latapie, and Yan Yan. Egok360: A 360 egocentric kinetic human activity video dataset. *arXiv preprint arXiv:2010.08055*, 2020. 3, 4
- [3] Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, et al. SAM 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025. 9
- [4] Angel Chang et al. Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158*, 2017. 4
- [5] Weijian Chen, Weibo Yao, Yuhang Zhang, Xiaolin Tang, Guo Wang, Weijun Zhang, Xitong Gao, Yihao Chen, Hongde Qin, and Lu Qi. Geometry and gradient-based partitioning for panoramic outdoor reconstruction, 2026. 4
- [6] 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. In *Proceedings of Robotics: Science and Systems*, 2024. 5
- [7] Dima Damen et al. The epic-kitchens dataset: Collection, challenges and baselines. *arXiv preprint arXiv:2005.00343*, 2020. 3
- [8] Mingqi Gao, Yunki Miao, and Jungong Han. SAM-Body4D: Training-free 4d human body mesh recovery from videos. *arXiv preprint arXiv:2512.08406*, 2025. 11
- [9] Xian Ge, Yuling Pan, Yuhang Zhang, Xiang Li, Weijun Zhang, Dizhe Zhang, Zhaoliang Wan, Xin Lin, Xiangkai Zhang, Juntao Liang, Xiangtai Li, WenJie Jiang, Bo Du, Ming-Hsuan Yang, and Lu Qi. Aircsim360: A panoramic simulation platform within drone view. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26931–26940, 2026. 4
- [10] Kristen Grauman et al. Ego4d: Around the world in 3,000 hours of egocentric video. *arXiv preprint arXiv:2110.07058*, 2021. 2, 3
- [11] Kristen Grauman et al. Ego-exo4d: Understanding skilled human activity from first- and third-person perspectives. *arXiv preprint arXiv:2311.18259*, 2023. 3, 5
- [12] Tianjun Gu et al. Towards spatial supersensing in the wild. *arXiv preprint arXiv:2607.13681*, 2026. 4
- [13] Ryan Hoque, Peide Huang, David J. Yoon, Mouli Sivapurapu, and Jian Zhang. Egodex: Learning dexterous manipulation from large-scale egocentric video. *arXiv preprint arXiv:2505.11709*, 2025. 3
- [14] Ahad Jawaid and Yu Xiang. Openego: A large-scale multimodal egocentric dataset for dexterous manipulation. *arXiv preprint arXiv:2509.05513*, 2025. 3
- [15] Haoran Jiang, Jin Chen, Qingwen Bu, Li Chen, Modi Shi, Yanjie Zhang, Delong Li, Chuanzhe Suo, Chuang Wang, Zhihui Peng, and Hongyang Li. WholeBodyVLA: Towards unified latent VLA for whole-body loco-manipulation control. *arXiv preprint arXiv:2512.11047*, 2025. 3
- [16] Hualie Jiang, Ziyang Song, Zhiqiang Lou, Rui Xu, and Minglang Tan. Depth anything in 360°: Towards scale invariance in the wild, 2025. 4
- [17] Jae Yong Lee et al. Ego-1k: A large-scale multi-view video dataset for egocentric vision. *arXiv preprint arXiv:2603.13741*, 2026. 3, 4
- [18] Hao Li, Ganlong Zhao, Yufei Liu, Haotian Hou, Guoquan Ye, Tongyan Fang, Chunxiao Liu, Siyuan Huang, Jianbo Liu, Xiaogang Wang, and Hongsheng Li. ACE-Ego-0: Unifying egocentric human and robotic data for VLA pretraining. *arXiv preprint arXiv:2606.17200*, 2026. 2
- [19] Qixiu Li, Yu Deng, Yaobo Liang, Lin Luo, Lei Zhou, Chengtang Yao, Lingqi Zeng, Zhiyuan Feng, Huizhi Liang, Sicheng Xu, Yizhong Zhang, Xi Chen, Hao Chen, Lily Sun, Dong Chen, Jiaolong Yang, and Baining Guo. Scalable vision-language-action model pretraining for robotic manipulation with real-life human activity videos. *arXiv preprint arXiv:2510.21571*, 2025. 2

- [20] Yihang Li et al. EgoLive: A large-scale egocentric dataset from real-world human tasks. *arXiv preprint arXiv:2604.23570*, 2026. 2, 4
- [21] Shijie Lian, Bin Yu, Xiaopeng Lin, Changti Wu, Hang Yuan, Xiaolin Hu, Zhaolong Shen, Yuzhuo Miao, Haishan Liu, Yuxuan Tian, Yukun Shi, Cong Huang, and Kai Chen. Phys-brain 1.0 technical report, 2026. 2
- [22] Shijie Lian, Bin Yu, Xiaopeng LIN, Laurence Tianruo Yang, Zhaolong Shen, Changti Wu, YuZhuo Miao, Cong Huang, and Kai Chen. Langforce: Bayesian decomposition of vision language action models via latent action queries. In *Forty-third International Conference on Machine Learning*, 2026. 3
- [23] Xin Lin, Xian Ge, Dizhe Zhang, Zhaoliang Wan, Xianshun Wang, Xiangtai Li, Wenjie Jiang, Bo Du, Dacheng Tao, Ming-Hsuan Yang, and Lu Qi. One flight over the gap: A survey from perspective to panoramic vision, 2025. 4
- [24] Xiaopeng Lin, Shijie Lian, Bin Yu, Ruqi Yang, Zhaolong Shen, Changti Wu, Yuzhuo Miao, Yurun Jin, Yukun Shi, Jiyang He, Cong Huang, Bojun Cheng, and Kai Chen. Phys-brain: Human egocentric data as a bridge from vision language models to physical intelligence, 2025. 2
- [25] Xin Lin, Meixi Song, Dizhe Zhang, Wenxuan Lu, Haodong Li, Bo Du, Ming-Hsuan Yang, Truong Nguyen, and Lu Qi. Depth any panoramas: A foundation model for panoramic depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26835–26844, 2026. 4
- [26] Xiaopeng Lin, Ruqi Yang, Shijie Lian, Zhaolong Shen, Bin Yu, Changti Wu, Haibao Liu, Yuxiang Zhang, Hong Li, Qiyuan Su, Haochen Liu, Xuguo He, Yukun Shi, Cong Huang, Zhirui Zhang, Bojun Cheng, and Kai Chen. Human-as-humanoid: Enabling zero-shot humanoid learning from ego-exo human videos with human-aligned embodiments, 2026. 2
- [27] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics*, 34(6): 248:1–248:16, 2015. 11
- [28] Lingni Ma et al. Nymeria: A massive collection of multi-modal egocentric daily motion in the wild. *arXiv preprint arXiv:2406.09905*, 2024. 3
- [29] Roberto Martin-Martin et al. JrdB: A dataset and benchmark of egocentric robot visual perception of humans in built environments. *arXiv preprint arXiv:1910.11792*, 2019. 4
- [30] Ryan Punamiya et al. Egoverse: An egocentric human dataset for robot learning from around the world. *arXiv preprint arXiv:2604.07607*, 2026. 2, 3, 4
- [31] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7464–7475, 2023. 9
- [32] Xin Wang, Taein Kwon, Mahdi Rad, Bowen Pan, Ishani Chakraborty, Sean Andrist, Dan Bohus, Ashley Feniello, Bugra Tekin, Felipe Vieira Fruejri, Neel Joshi, and Marc Pollefeys. Holoassist: An egocentric human interaction dataset for interactive ai assistants in the real world. *arXiv preprint arXiv:2309.17024*, 2023. 3, 5
- [33] Songlin Wei, Hongyi Jing, Boqian Li, Zhenyu Zhao, Jiageng Mao, Zhenhao Ni, Sicheng He, Jie Liu, Xiawei Liu, Kaidi Kang, Sheng Zang, Weiduo Yuan, Marco Pavone, Di Huang, and Yue Wang. Ψ_0 : An open foundation model towards universal humanoid loco-manipulation. *arXiv preprint arXiv:2603.12263*, 2026. 2
- [34] Jianxiong Xiao, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. Recognizing scene viewpoint using panoramic place representation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2012. 4
- [35] Ruihan Yang, Qinxu Yu, Yecheng Wu, Rui Yan, Borui Li, An-Chieh Cheng, Xueyan Zou, Yunhao Fang, Xuxin Cheng, Ri-Zhao Qiu, Hongxu Yin, Sifei Liu, Song Han, Yao Lu, and Xiaolong Wang. EgoVLA: Learning vision-language-action models from egocentric human videos. *arXiv preprint arXiv:2507.12440*, 2025. 2
- [36] Bin Yu, Shijie Lian, Xiaopeng Lin, Yuliang Wei, Zhaolong Shen, Changti Wu, Yuzhuo Miao, Xinming Wang, Bailing Wang, Cong Huang, and Kai Chen. Twinbrainvla: Unleashing the potential of generalist vlms for embodied tasks via asymmetric mixture-of-transformers, 2026. 3
- [37] Heeseung Yun, Youngjae Yu, Wonsuk Yang, Kangil Lee, and Gunhee Kim. Pano-avqa: Grounded audio-visual question answering on 360-degree videos. *arXiv preprint arXiv:2110.05122*, 2021. 4
- [38] Gu Zhang, Qicheng Xu, Haozhe Zhang, Jianhan Ma, Long He, Yiming Bao, Zeyu Ping, Zhecheng Yuan, Chenhao Lu, Chengbo Yuan, Tianhai Liang, Xiaoyu Tian, Maanping Shao, Feihong Zhang, Mingyu Ding, Yang Gao, Hao Zhao, Hang Zhao, and Huazhe Xu. UniDex: A robot foundation suite for universal dexterous hand control from egocentric human videos. *arXiv preprint arXiv:2603.22264*, 2026. 3
- [39] Xiangkai Zhang, Dizhe Zhang, WenZhuo Cao, Zhaoliang Wan, Yingjie Niu, Lu Qi, Xu Yang, and Zhiyong Liu. Fly360: Omnidirectional obstacle avoidance within drone view, 2026. 4
- [40] Yinda Zhang, Shuran Song, Ping Tan, and Jianxiong Xiao. Panocontext: A whole-room 3d context model for panoramic scene understanding. In *European Conference on Computer Vision*, 2014. 4
- [41] Ruijie Zheng et al. Egoscale: Scaling dexterous manipulation with diverse egocentric human data. *arXiv preprint arXiv:2602.16710*, 2026. 2, 3, 4