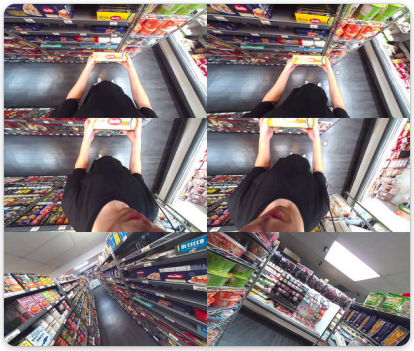
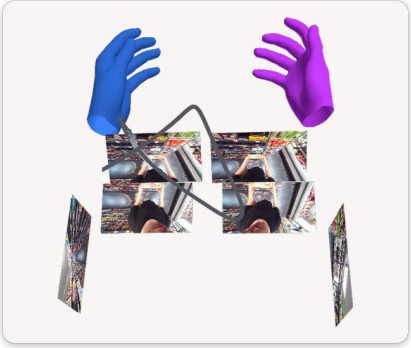
RoboCap: A New Platform for Egocentric Robot Learning

Qasim Ali[1] Nikash Bhardwaj[1] Kheri Hughes[1] Sam Cho[2] Ethan Lee[2] Paul Luo[2]
Shinn Chong[2] Sunny Yu[2] Owen Akhibi[1] Alex Proshkin[1] Haotian Zhan[1]
Niresh Dravin[2] Vivek Sharma[2] Jonathan Victor[2] Amanda Young[2] Michael Cho[2]
Ademi Adeniji[1] Vincent Liu[1]

[1]Grounded Superintelligence [2]BitRobot





Abstract

Despite its promise for scaling robot learning, egocentric manipulation data is still scarce today. Collection at scale requires vertically integrating ergonomic hardware with centimeter-precise 3D algorithms, at a precision that has not been publicly demonstrated. To address this gap, we introduce RoboCap, a 250 g six-camera dual-IMU hat designed for in-the-wild egocentric data capture, and the Grounded API, a suite of device-agnostic 3D algorithms tuned for RoboCap. In this report, we demonstrate how hardware, calibration, and 3D algorithms interact to achieve state-of-the-art performance on the public benchmarks: our SLAM across diverse settings and rigs, our depth estimation on egocentric settings, and our hand tracking when adapted to third-party devices.

1. Introduction

In the history of large-scale end-to-end robot learning, the field has witnessed the emergence of several data classes, each with known tradeoffs. Simulation scales with compute, but demands human engineering effort that grows with the diversity it aims to produce. Teleoperation captures the exact states and actions a robot requires at inference, at the cost of unnatural and unscalable demonstration through cumbersome robot hardware. UMI deploys only the end-effector for better scalability, but does not scale beyond low-DoF manipulators and requires retargeting. Egocentric data maximizes scalability by equipping humans with an unintrusive headworn device, but recovers states and actions only by estimation and also requires retargeting.

The quality of egocentric data is bounded by the accuracy of its 3D labels, namely the human hand and head poses. Retargeting maps an estimated human pose onto a robot embodiment [25, 27, 33, 42], so estimation error enters the policy's training distribution directly as label noise: egocentric capture is a metrology problem before it is a video collection problem, and a large-scale 3D computer vision problem in practice. Centimeter-level accuracy must hold under occlusion, motion blur, varied lighting, and operator diversity, on footage collected outside any lab. The field has not converged on a device the way it has on teleoperation and UMI rigs, and the reason is that the two halves of the problem (form factor and 3D algorithms) have never been available together. Meta's Project Aria glasses [15] are the best form factor candidate, but are gated by application programs, hardware costs, and offline APIs that are difficult to scale. Mixed-reality headsets are heavy, socially intru-

sive, and expose only opaque on-device tracking whose error characteristics cannot be inspected or improved offline. A team that wants to collect its own egocentric data at the precision policy learning requires cannot currently assemble the stack to do so.

We argue that the estimation problem in egocentric data is an artifact of hardware that was never designed for it. Because consumer smart glasses and mixed-reality headsets optimize for display, audio, and human-facing UX, their sensing suites are byproducts. When the device is instead designed backward from the requirements of the 3D estimation stack–stereo baselines, fields of view overlap, hardware-synchronization, global-shutter sensors–the gap between estimated and measured state collapses to a level sufficient for policy learning. As we explain in Section 4, RoboCap's layout is what makes this concrete. Two calibrated stereo pairs supply metric scale that monocular systems can only infer, while wide-angle lateral coverage keeps the hands in frustum through natural motion and keeps the trajectory solver constrained during close-range manipulation.

We also design our 3D algorithms to be device-agnostic. The same binaries run on handheld fisheye rigs, aerial stereo platforms, consumer VR headsets, and research glasses whose divergent cameras share almost no field of view, with no per-dataset tuning. This is not incidental to the argument above but rather evidence for it: portability establishes that our results reflect general 3D estimation rather than models overfit to a single device. Our SLAM, for instance, borrows from both monocular-inertial and stereo-inertial designs, which let it fuse whatever cues a rig provides–six cameras with two calibrated baselines on RoboCap, or two divergent cameras with no stereo overlap on Aria–into a single estimation pipeline.

Our contributions are twofold:

1. RoboCap, a purpose-built 250 g six-camera dual-IMU egocentric capture device whose sensing suite is designed minimally and sufficiently for metric 3D estimation, available for purchase.
2. Grounded API, which serves SLAM, depth estimation, and hand tracking algorithms that achieve state-of-the-art accuracy across diverse visual settings and hardware rigs.

2. Related work

The contributions of this paper touch on progress in egocentric wearable hardware and in 3D estimation. In this section we review each in turn.

2.1. Egocentric wearables

Egocentric capture hardware spans repurposed consumer devices to purpose-built research platforms. Early large-scale egocentric datasets were collected with head-mounted action cameras [11, 20], which provide only monocular RGB with rolling shutter and no inertial or calibration guarantees, precluding metric 3D annotation. Mixed-reality headsets such as Apple's Vision Pro and Meta's Quest 3 offer richer sensing and supply head and hand pose estimation [3, 25], but these devices are heavy, socially intrusive, and expose their sensing only through on-device tracking APIs.

The closest platform in spirit to ours is Meta's Project Aria [15]: lightweight research glasses with calibrated RGB, monochrome SLAM cameras, and IMUs, that have powered a family of egocentric datasets [3, 35, 36, 38]. Aria validates the thesis that calibrated multi-camera visual-inertial wearables enable high-quality 3D annotation, and prior works in robotics have demonstrated its use in learning robot manipulation policies from human data [27, 33]. Aria remains, however, a general-purpose research device rather than a scalable manipulation capture platform: its cameras are laid out for scene understanding rather than for the near field, and its complexity has kept it relatively inaccessible for robot learning at scale.

There is also a lineage of research prototypes aimed at estimating human poses. For example, [1, 43, 58] propose head- or cap-mounted downward-facing fisheye rigs, but for target body pose rather than manipulation. More recent works combine mocap gloves with an egocentric device to estimate hand pose, but are too clunky and brittle to occlusions [49, 52]. These prototypes introduce intrusive hardware in an attempt to measure hand pose accurately, but thereby sacrifice the naturalness and scalability. Inevitably, a single egocentric device must natively compute accurate hand poses directly from vision in order to maximize the scalability of egocentric data capture.

2.2. 3D estimation

Camera calibration. Multi-sensor rigs are typically calibrated offline by continuous-time batch optimization over a fiducial target [16], jointly recovering intrinsics, camera–IMU extrinsics, and time offsets. Factory calibration degrades in the field through mechanical and thermal stress, motivating online self-calibration from natural correspondences during operation [12, 34]. As we demonstrate in Section 4.1, this drift becomes important when needing to achieve sub-centimeter 3D estimation errors.

Visual-inertial SLAM. Classical sparse visual-inertial systems establish the standard factor-graph formulation fusing feature reprojection with IMU preintegration [7, 19, 30, 41, 47, 51], with global consistency recovered through appearance-based place recognition [17] and pose-graph optimization [21]. GPU-accelerated implementations extend this to multi-camera rigs at real-time rates [28]. A learned line trades classical sparse optimization for dense

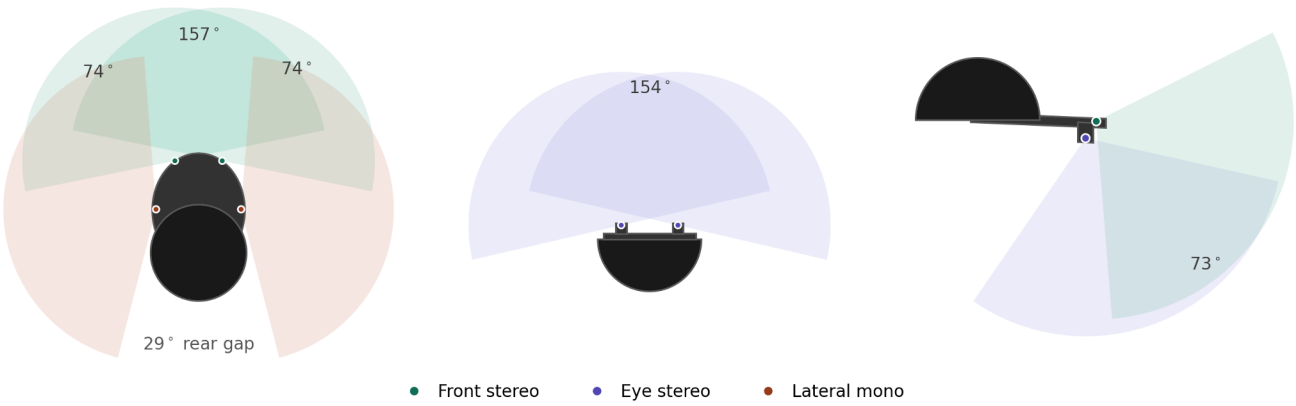


Figure 1. RoboCap camera coverage. Left: azimuth coverage of the front stereo and lateral cameras. Center: azimuth coverage of the eye stereo cameras, visualized forward for legibility. Right: elevation coverage of the front and eye pairs. FoV-overlap in degrees at infinity.

correspondence and feed-forward geometry [37, 50, 53], while rendering-based backends push offline reconstruction quality at significant compute cost [8, 68].

These systems are evaluated on a small set of standard benchmarks with external ground truth, spanning aerial stereo-inertial, handheld fisheye, multi-sensor rigs [6, 18, 48], and most recently consumer VR headsets [14], the closest existing benchmark to the egocentric regime. Egocentric head motion is a distinctly adversarial setting for all of them: rapid rotation, close-range parallax, dynamic scene content from the wearer's own hands, and irregular motion during initialization. This motivates both our multi-camera wide-FoV configuration and our offline post-processing of the estimated trajectory.

Stereo depth estimation. Deep stereo matching progressed from cost-volume CNNs to iterative recurrent refinement [9, 31, 32, 54, 57], but these models require per-domain fine-tuning and degrade sharply out of distribution. A more recent class trains on large-scale high-fidelity synthetic corpora, bootstrapping monocular foundation priors to achieve strong zero-shot generalization for the first time [55, 56]. Monocular metric alternatives [4, 5, 61, 62, 64] remove the calibrated-baseline requirement, but their metric scale is inferred from learned priors rather than measured from geometry, which is exactly the failure mode our hardware exists to eliminate.

Hand tracking. Parametric hand recovery is dominated by the MANO model [45]. Monocular mesh recovery has scaled steadily from early regression and real-time keypoint systems [46, 66] to transformer-based reconstructors trained at scale [39, 40]. Critically, all monocular methods regress pose under a weak-perspective camera, which infers shape and depth from apparent hand size against a learned prior. Not only are monocular outputs metrically ambiguous, they are also not consistent enough across views for accurate triangulation, which inevitably must be learned with neural network design [26, 60]. World-space extensions recover trajectories from moving egocentric video, but are still susceptible to scale ambiguity as well as SLAM inaccuracies [65, 67].

Multi-view systems resolve the metric ambiguity of monocular prediction, but at the cost of added hardware dependency. Production-grade egocentric trackers fuse several head-mounted fisheye views into a single metric prediction [23, 24], and the multi-view egocentric benchmarks built on their data [3, 44] evaluate this setting directly. These systems derive their training labels from motion-capture installations or closed on-device trackers, so annotation cost scales with instrumentation rather than with compute. Furthermore, motion capture confines collection to instrumented spaces, which recent work [44] addresses by building a large exoskeleton cage. Yet the difficulty of hand tracking is in-the-wild, where metric accuracy must persist through occlusion, motion blur, and operator diversity. Our method solves this by enforcing geometric consistency from monocular priors, allowing us to generate a large corpus of in-the-wild pre-training data without any motion capture. Section 4.4 evaluates models trained on this corpus against baselines trained on motion-capture-derived data [3, 24, 44].

3. Device

The sensing suite is specified by what metric 3D estimation requires at manipulation range. Depth must be measured rather than inferred, which requires at least one calibrated stereo pair with a baseline wide enough that disparity resolves centimeters at arm's length. The hands must remain in frustum through natural motion, including reaches that

Cameras (6, hardware-synchronized)	
Sensor	Global-shutter RGB, 10-bit, 3.0 µm px
Frame resolution	1920×1080
Frame rate	30 Hz
Lens	Kannala-Brandt 4
Focal length	$f = 590$–630 px
Exposure	Per-camera AE, bounded ≤ 10 ms
Stereo optics	Fisheye, $180^\circ\times112^\circ$ (216° diag.)
Lateral optics	Wide, $161^\circ\times98^\circ$ (179° diag.)
Front stereo	$2\times$ 86.8 mm baseline, 157° overlap
Eye stereo	$2\times$ 106.2 mm baseline, 154° overlap
Lateral	$2\times$ mono, 85° off forward axis, 30° down

Magnetometer (1, crown center)	
Rate	3-axis, 100 Hz, time-aligned (auxiliary)

IMUs (2, opposite sides of crown, 6-axis)	
Rate	200 Hz, hardware-synchronized interrupt sampling
Gyro. noise density	$6\times10^{-4}\,\mathrm{rad\,s^{-1}}/\sqrt{\mathrm{Hz}}$
Gyro. bias walk	$3\times10^{-5}\,\mathrm{rad\,s^{-2}}/\sqrt{\mathrm{Hz}}$
Accel. noise density	$7\times10^{-3}\,\mathrm{m\,s^{-2}}/\sqrt{\mathrm{Hz}}$
Accel. bias walk	$2\times10^{-4}\,\mathrm{m\,s^{-3}}/\sqrt{\mathrm{Hz}}$
Camera–IMU offset	Residual ≤ 3 ms

System	
Reference clock	24 MHz TCXO, ±1.81 ppm
Trigger	Single PWM, 1-to-6 fan-out to all cameras
Power	4–6 W
Weight	250 g
Runtime	6 h per 10,000 mAh
Storage	214.5 GB onboard

Table 1. RoboCap sensor suite. The two stereo pairs use fisheye optics, the lateral cameras a wide lens. Stereo overlap is the binocular field at infinity from the calibrated extrinsics. IMU noise parameters are fleet-level estimates from per-device continuous-time calibration [16].

leave the forward field, which requires wide-angle coverage well beyond the forward-facing cone. Multi-view fusion is only valid if views are simultaneous, which requires hardware-triggered global-shutter sensors rather than per-channel scheduling of rolling-shutter streams. And metric scale is only as good as the calibration it rests on, which requires per-unit factory characterization rather than a nominal model shared across a fleet. Table 1 summarizes the resulting suite: six hardware-synchronized global-shutter cameras in two calibrated stereo pairs with two lateral wide-angle views, and two IMUs at opposite sides of the crown.

The two requirements pull in different directions, so the cameras do not share optics. The stereo pairs carry fisheye lenses spanning $180^\circ\times112^\circ$ (216° diagonal), which keeps the hands within a metric stereo frustum across the full reach envelope; the front pair sits on a 86.8 mm baseline with 157° of binocular overlap and the eye pair on a 106.2 mm baseline with 154°. The lateral cameras instead carry a wide lens of $161^\circ\times98^\circ$ (179° diagonal), mounted 85° off the forward axis and pitched 30° down, where the priority is catching reaches that leave the stereo frustum entirely and keeping the trajectory solver constrained during close-range manipulation rather than triangulating them. Each lateral view retains roughly 80° of overlap with the common stereo field, leaving a 20° gap directly behind the wearer.

The hat form factor was chosen against glasses, headsets, and phone rigs to accommodate that suite. A hat distributes weight over the crown, permitting a larger sensor and battery payload than eyewear can carry without discomfort; it requires no display, audio, or interaction hardware, so cost is dominated by cameras and IMUs; and it is socially unobtrusive enough for collection outside a lab, where the data this device exists to capture is produced. The result is 250 g, worn for hours without adjustment.

The device draws 4–6 W in typical operation and is powered by a separate tethered battery bank; a 10,000 mAh bank sustains roughly six hours of continuous recording. A small onboard lithium-ion battery serves as a fail-safe. If the tether is severed mid-session, it keeps the device alive to finalize and save the recording, and it powers data offload when no bank is attached, so no session is lost to a power interruption. Sessions are encoded onboard (H.265, YUV 4:2:0, 4 Mbps per stream) to 214.5 GB of internal storage (10+ hours of recording) and offloaded via USB 3 at 60–200 MB/s.

3.1. Calibration

Every unit is individually calibrated at the factory before shipment. An operator sweeps the device through a calibration motion sequence against a stationary AprilGrid target (6×6, 55 mm tags) simultaneously covering all camera groups and both IMUs. IMU noise models are identified by Allan-variance analysis; camera intrinsics (pinhole + Kannala–Brandt KB-4 distortion, the standard choice at these field-of-view angles), camera–IMU extrinsics, and time offsets are solved by continuous-time batch optimization [16]. Because the two lens types are calibrated against the same target sweep, the recovered focal lengths span 590–630 px across the fleet rather than a single nominal value. Units are accepted only within per-parameter tolerance gates for each of camera extrinsics and intrinsics reprojection, gyroscope, and accelerometer–and units failing any gate are reworked rather than shipped.

Factory calibration is performed at room temperature and the device applies no temperature-dependent parameter correction, so calibration degrades in the field through mechanical and thermal stress. At manipulation range this

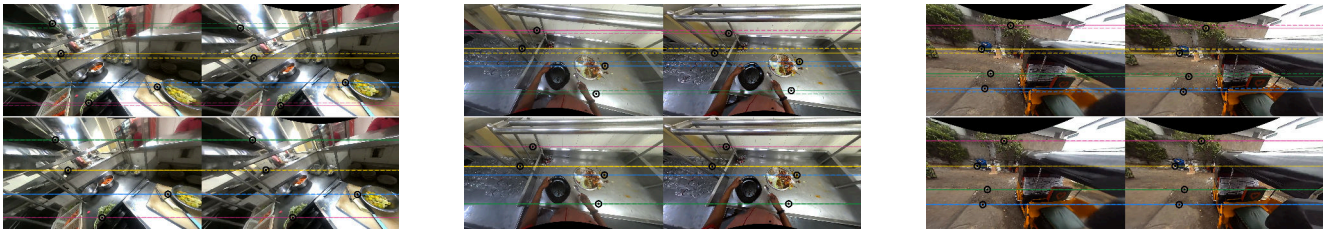


Figure 2. Epipolar residuals before (top) and after (bottom) self-calibration. Identical features are marked in the left and right cameras with a black dot, and epipolar lines are drawn across each pair. Self-calibration eliminates almost all epipolar drift and improves 3D estimation.

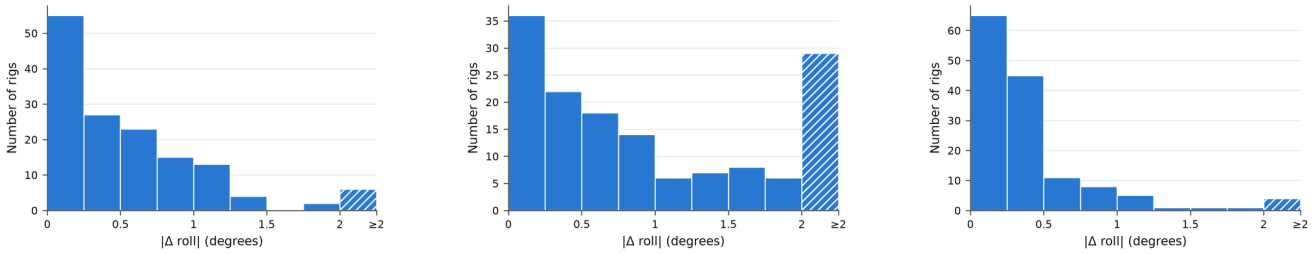


Figure 3. Histograms of the epipolar drift in the eye pair (left), front pair (middle), and front-eye pairs (right), over a subset of 150 deployed devices. We find that rotational correction (Section 4.1) is sufficient to restore epipolar drift, which follows a Gaussian centered around 0°.

drift is not a second-order effect: a pixel of disparity error translates to roughly a centimeter of depth at 0.65 m. Rather than tighten the mechanical design against a drift that cannot be eliminated, we correct it per session from natural correspondences, which Section 4.1 describes and quantifies.

3.2. Time synchronization

All six cameras are exposure-triggered from a single PWM frame-synchronization signal generated by the main controller and fanned out through a 1-to-6 clock buffer, so simultaneity holds by construction rather than by per-channel scheduling; the measured camera–camera offset is 0 ms at timestamping granularity. Frames are timestamped at exposure completion on the device clock. The IMUs stream autonomously at 200 Hz under hardware-synchronized interrupt sampling and are timestamped on arrival in the same clock domain, with the camera–IMU offset identified per device by continuous-time calibration [16] to a residual of ≤ 3 ms, verified from natural recordings by correlating gyroscope rates against image-derived rotation.

The device clock is disciplined by a 24 MHz TCXO with a ± 1.81 ppm worst-case budget, bounding absolute timeline drift to $\approx \pm\, 6.5$ ms per hour. Because all sensors share this clock, that drift affects the absolute timeline only and not inter-sensor synchronization, which is what multi-view fusion and visual-inertial estimation depend on.

4. 3D estimation

The most common way to retarget motions from human data is to define a unified action space across embodiments as the 3D end-effector pose. Therefore, the role of the 3D stack is to recover this pose, together with the geometry needed to contextualize it, at metric scale in a consistent world frame. In this section, we describe our SLAM, depth estimation, and hand tracking implementations that we have evaluated to be state-of-the-art on several cross-rig and in-rig benchmarks.

4.1. Online self-calibration

Centimeter precision demands tight alignment between the matcher and the device calibration. One pixel of disparity corresponds to $\sim$1 cm of depth at 0.65 m, and stereo error grows as $z^2/(fb)$. In practice, the rig drifts slightly from its factory calibration through mechanical and thermal stress. Across our fleet, the relative rotation shifts by 0.2–1.1° (median $\approx 0.5^\circ$), producing 1–5 px of epipolar residual and a distance-dependent depth bias of 1–2 cm at manipulation range. Our observations are consistent with what Aria reports $\approx$25 arcmin of instantaneous deformation between its stereo cameras from the act of removing the device [15].

Drift is dominated by rotation, so we only estimate the residual rotations once per recording session from natural correspondences [12, 34] with no calibration target and no operator intervention. The left-front camera is the anchor and is never moved. Each pair is constrained by its own correspondences and fit independently against the Sampson epipolar distance rather than the vertical residual of a rectified pair. Corrections are accepted only if they improve a held-out residual and fall within $\approx 1^\circ$ of that device's own factory value. Figure 3 shows the spread of online self-

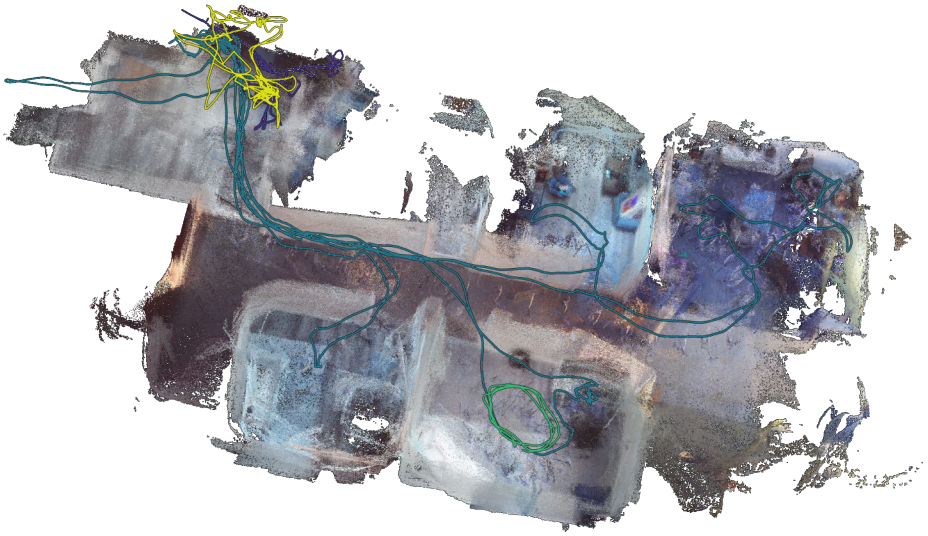


Figure 4. Scene reconstruction over 4 RoboCap sessions. For each session, we generate point clouds (Section 4.3) and project them into a gravity-aligned world frame (Section 4.2). Then, we solve multi-session frame alignment with RANSAC on 3D-projected SIFT features.

Benchmark	ORB-SLAM3 [7]	Basalt [51]	OKVIS2 [29]	cuVSLAM [28]	Ours
EuRoC [6]	3.5 / 3.3	5.8 / 5.5	3.1 / **2.3**	6.6 / 8.4$^{(4)}$	**2.7** / 2.4
TUM-VI (rooms) [48]	**0.9 / 0.9**	9.2 / 6.3	**0.9 / 0.9**	2.0 / 1.7	**0.9 / 0.9**
VECtor [18]	14.7 / 23.0$^{(5)}$	28.0 / 71.1$^{(7)}$	25.7 / 16.8$^{(3)}$	16.5 / 36.0$^{(7)}$	**10.5 / 4.0**$^{(4)}$
Monado [13]*	– / 2.5	– / 11.2	– / **2.2**	11.5 / 11.5$^{(20)}$	**3.0 / 2.2**
ADT [38]†	**2.0 / 1.8**$^{(3)}$	11.2 / 6.5$^{(3)}$	5.4 / 3.0$^{(4)}$	15.9 / 6.9$^{(12)}$	**2.0 / 1.8**
HOT3D [3]†	1.3 / $\infty^{(18)}$	4.5 / 4.7$^{(4)}$	1.7 / 1.7$^{(6)}$	46.9 / 55.4$^{(6)}$	**0.4 / 0.3**

Table 2. Final (offline) ATE RMSE in cm, mean / median over sequences. A crash or an ATE $>$ 1 m is a failure denoted with $^{(n)}$: excluded from the mean, ∞ in the median. We use each paper's own numbers where applicable, otherwise we run its released code with the published configuration on the same inputs. *Monado baselines are the non-causal aggregate medians reported by the dataset authors, and means are not published. †Aria datasets with pseudo-ground truth from Aria MPS. See Figure 7 for trajectory visualizations.

calibrated rotations and Figure 2 visualizes epipolar lines before and after online calibration.

4.2. SLAM

Head motion is estimated by a multi-camera visual-inertial SLAM system of our own, built on the now-standard ingredients: GPU feature tracking, a keyframe-graph inertial estimator with marginalization, and appearance-based place recognition [7, 28, 29]. Every camera on the device contributes to a single estimate, including the lateral views that have no overlapping partner and so carry landmarks of their own; this is what keeps the solver constrained when the wearer's hands are close to the face and the front pair sees little else. Camera–IMU timing and the stereo extrinsics are refined per session.

Two departures from the standard recipe account for most of our accuracy. First, landmarks carry an identity that outlives the estimator's window, so a place seen again is measured against the same points rather than against a fresh set, and the error accumulated over an excursion is removed on return rather than merely halted. Second, a recognized revisit is applied inside the visual-inertial estimator itself rather than handed to a separate pose-graph stage, so the correction is weighted by the same cost that produced the estimate. The trajectories reported here are the offline product of the system, in which a final global refinement over the session's keyframes is run once all revisits are known. The causal estimate that runs on the device during capture uses the same machinery without that final pass.

Evaluation. We evaluate on EuRoC [6], TUM-VI [48] and VECtor [18], on the Monado headset sequences [13], and on the Aria recordings of ADT [38] and HOT3D [3], the regime closest to our device, with one configuration and

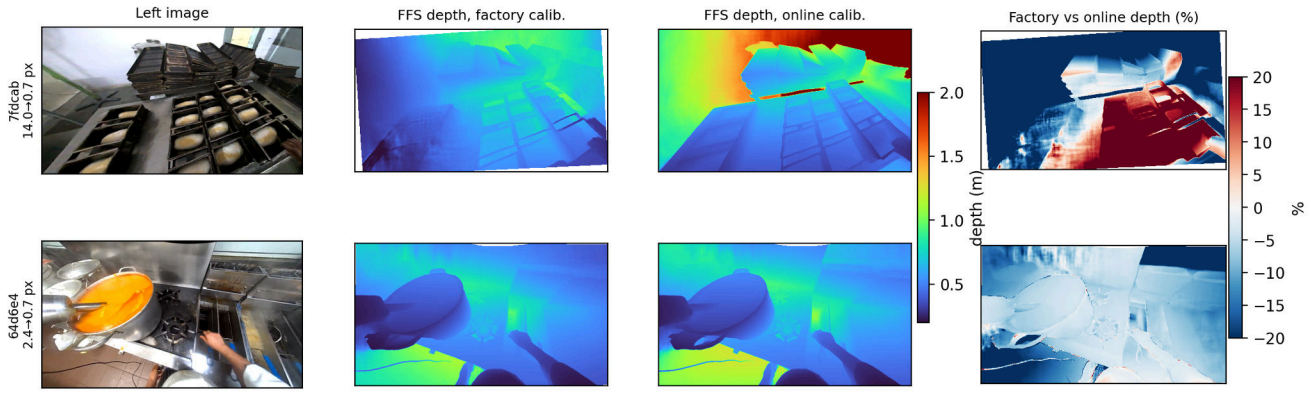


Figure 5. FFS depth with factory vs. online calibration, and the depth change between them. Factory calibration does not visibly degrade FFS depth maps: they remain sharp and complete, but are systematically distorted in scale and shape. Online calibration leaves the maps' appearance unchanged while removing this distortion, halving depth error within reach. See Figure 8 for more examples.

		FFS [56]		FS [55]	
Pixels	% px	AbsRel↓	$\delta_{1.25}$ ↑	AbsRel↓	$\delta_{1.25}$ ↑
All	100.0	0.455	69.7	0.278	75.5
Visible in both views	11.8	0.016	99.6	0.015	99.7
Visible in one view	26.8	0.297	61.2	0.171	77.8

Table 3. Zero-shot depth on UnrealEgo2 [2]. In both views: pixels at 0.3–1.0 m with a correspondence in the right image. In one view: pixels without correspondence. FS and FFS are comparable where there is correspondence.

	Calibration	
Distance	Factory (Sec. 3.1)	Online (Sec. 4.1)
<0.8 m	0.023	0.011
≥0.8 m	0.029	0.016
All	0.025	0.012

Table 4. Depth AbsRel of zero-shot FFS on a ChArUco board seen by the RoboCap front pair (1,277 frames, 106,602 corners, same frames for both calibrations).

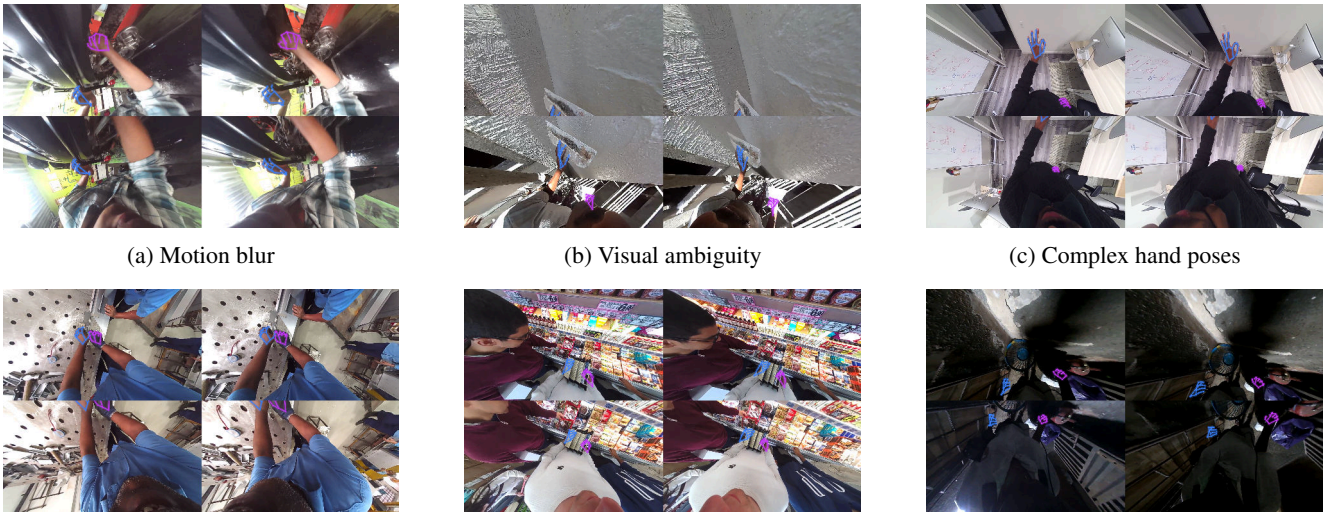
with the cameras each dataset provides for SLAM. Table 2 reports offline ATE against ORB-SLAM3 [7], Basalt [51], OKVIS2 [29] and cuVSLAM [28], quoting published numbers where they exist and otherwise running the released code with its published configuration on the same inputs. For head-worn capture at room scale we match or better the best result on every benchmark, and we are the only system that completes all of them: Monado, ADT and HOT3D are finished without a failure, where the strongest baseline on each loses one, three and six sequences. On EuRoC and the TUM-VI rooms we are at parity with OKVIS2 and ORB-SLAM3, whose published figures those benchmarks were tuned against over a decade. On VECtor we lead the nearest system by a factor of four in median error, and on the Aria sets our error is at the level of the MPS pseudo-ground truth and should be read as agreement with it. The causal estimate produced online during capture stays within a factor of two of the offline column; per-sequence results for both are given in the appendix.

4.3. Depth estimation

We estimate metric depth on both stereo pairs with Fast-FoundationStereo (FFS) [56], a lightweight model distilled from FoundationStereo (FS) [55], and find that model performance is dictated more by calibration than generalization ability.

Evaluating zero-shot transfer to egocentric scenarios. To evaluate the ability of these zero-shot models in egocentric settings, we benchmark them on UnrealEgo2 [2], a synthetic head-mounted stereo benchmark with dense ground truth. FFS reaches AbsRel 0.016 on surfaces both cameras see at 0.3–1.0 m, at the benchmark's 2 cm quantization limit and on par with FS (Table 3). The remaining error lies on pixels only one camera sees, where there is nothing to match. FS fills these better from context, at over 10× the inference cost. We conclude that zero-shot models are accurate enough, and that the limiting factor on deployed hardware is online calibration (Section 4.1).

Evaluating the effect of online calibration. To measure the effect of calibration on depth, we record a ChArUco board over seven sessions on one device and take ground truth from the board pose, independent of the stereo extrinsics. Online calibration halves FFS depth error within reach (Table 4). Across a subset of 150 deployed devices, factory calibration leaves a median front-pair epipolar residual of 3.8 px, which online calibration reduces to 0.56 px. Switching from factory to online calibration changes FFS depth



(a) Motion blur (b) Visual ambiguity (c) Complex hand poses

(d) Gloves (e) Occlusion (f) Lighting

Figure 6. Examples from our hand tracking model, which fuses multi-view images into a single hand mesh. Our model's pretraining makes it robust to blur, visual ambiguity, complexity, gloves, occlusion, and lighting. See Figures 10 and 11 for more difficult examples.

within reach by a median 3.6% per device (11% at the 90th percentile), moving 31% of in-reach pixels by more than 5%.

4.4. Hand tracking

Hand tracking recovers metric MANO [45] hand pose in the world frame. Egocentric hand tracking is a 3D estimation problem under unusually hard visual distributions: in-the-wild recordings are dominated by visual and geometric edge cases rather than clean lab imagery, and no single camera sees the hands at all times, so estimates must fuse views of very different appearance under frequent occlusion. As a result, ground truth labels must be generated from in-the-wild recordings, which rules out motion capture as a reliable and scalable method of data collection. Furthermore, we pose hand tracking as a multi-view problem, since monocular hand models are both metrically ambiguous and inconsistent across views. Each view independently regresses shape and depth under a weak-perspective camera, so direct triangulation of monocular outputs produces distorted hands that accumulate many centimeters of error in the local hand frame alone.

We train our hand tracking model on a custom in-the-wild RoboCap dataset, spanning diverse skin tones, gloves, object occlusions, and motion blur. The labels for this dataset are produced by an offline multi-view optimization pipeline without any motion capture. We bootstrap monocular priors [39, 40] to learn a multi-view hand reconstruction model, then fit a single MANO parameterization via differentiable refinement that optimizes for reprojected appearance and naturalness throughout the trajectory. The hand-tracking model is a feed-forward multi-view ViT that consumes the raw bbox crops and camera parameters to predict a single hand pose across all views, guaranteeing multi-view consistency by construction.

We evaluate three things: the contribution of our pretraining dataset, the accuracy of the model relative to other multi-view hand estimation models, and its accuracy on the RoboCap. We report mean per-joint position error (MPJPE) in millimetres throughout, and detail splits and preprocessing in Appendix A.1.

Evaluating the benefit of the pretraining corpus. To isolate the effect of the RoboCap dataset, we train the same architecture on three public egocentric datasets, varying only the initialization. We select the best cross-dataset validation performance, which evaluates MPJPE on held-out subjects. HOT3D [3], UmeTrack [24] and SHOW3D [44] span three headset types with different camera counts, placements and optics, so the comparison covers rigs unlike the one used for pretraining. None of the three releases a public test benchmark, so we construct held-out splits that follow the official protocols as closely as possible; we describe this in Appendix A.1. For each benchmark, we finetune our pre-trained model and compare it with an identical model trained from scratch. Pretraining reduces error on all three datasets, by 2.9 to 7.5 mm (17–40%).

Comparison to multi-view reconstruction. We compare against POEMv2 [63], a state-of-the-art multi-view hand pose estimation model, on OakInk [59], DexYCB [10] and HO3D [22]. As our model is egocentric and pretrained on a single rig, these exocentric evaluation benchmarks are out

	Scratch	Pretrained
HOT3D [3]	17.41	**14.47**
UmeTrack [24]	19.00	**11.49**
SHOW3D [44]	19.34	**13.68**

Table 5. Best validation MPJPE (mm) on held-out subjects across datasets, varying only initialization. More details in Appendix A.

	POEMv2 [63]	Ours
DexYCB-Mv	6.69	**5.70**
OakInk-Mv	8.34	**7.34**
HO3D-Mv	**7.72**	15.16

Table 6. MPJPE (mm) on third-person multi-view benchmarks.

of distribution in both viewpoint and camera configuration. We finetune our hand tracking ViT for 20k steps. While we surpass the baseline on DexYCB and OakInk, our performance drops on HO3D. The drop in performance on HO3D likely happened because some test sequences drop 2/5 cameras for evaluating cross-rig generalization. Our model, which was trained on a single consistent rig, struggles to generalize to views/rigs different from the RoboCap without sufficient finetuning.

Evaluating performance on RoboCap. Our final evaluation is on our own proprietary sim benchmark, rendered with RoboCap camera extrinsics and intrinsics; we detail its construction in Appendix A.4. We achieve sub-centimeter accuracy (9.2 mm). In general, we do not believe that sim and motion capture data represents the difficulty of in-the-wild hand tracking, but still construct an internal benchmark to test the upper-bound performance of our models.

5. Limitations

Our hardware makes a specific bet. Designing the sensing suite backward from the requirements of metric 3D estimation buys precision at manipulation range, but it also fixes the geometry. The working volume is set by the stereo baselines and the lateral coverage, and reaching motions that leave every frustum are not recoverable. The device also assumes a tethered battery bank, which is acceptable for deliberate collection and not for ambient all-day wear. We regard the latter as a form-factor tradeoff rather than a permanent one.

Our SLAM matches or exceeds the published systems on every benchmark we report, but not uniformly. It struggles on long, self-similar loops, where too few revisits are recognized to correct the drift accumulated around them; this is a limit of place-recognition recall rather than of the estimator. Aria's own MPS pipeline is strong in exactly this regime, and on ADT and HOT3D we are scored against it, so the sub-centimetre figures there should be read as agreement with a pseudo-reference rather than as accuracy, and they bound what those datasets can demonstrate about either system. Finally, the causal estimate available during capture is about twice as inaccurate as the offline trajectory, since the refinement that produces the reported result runs over the whole session once capture has ended; applications that need the final accuracy online would not be served by this system today.

Hand tracking evaluation remains the weakest link in this setting, and not only for us. No public dataset provides what egocentric manipulation calls for at once: short-baseline rectified stereo, a head-worn viewpoint, and dense independent ground truth under visually diverse edge cases. We therefore design our benchmarks to measure the quality of our method transfer indirectly against ground truth on other rigs, as is standard in 3D computer vision literature.

Finally, this report establishes precision and not its downstream effect on policy performance. We argue in Section 1 that label error enters a policy's training distribution as noise, and prior work supports the link [27, 33, 42], but we do not measure here how much downstream performance a centimeter buys. That experiment requires training policies at scale on data this stack produces, and is the subject of ongoing work.

6. Conclusion

Egocentric data is one of the most scalable data sources available to robot learning. However, collecting it at the precision and scale that policy learning requires has demanded hardware built for 3D estimation and algorithms robust enough to run unattended on in-the-wild footage, and neither has been something a team could buy or call. We built both: RoboCap, a 250 g six-camera dual-IMU hat whose sensing suite is specified by what metric estimation requires, and the Grounded API, a suite of 3D algorithms that are tuned for it and demonstrate general state-of-the-art 3D capabilities when transferred to other rigs.

The result we find most consequential is not any single benchmark number but what bootstrapping 3D labels for in-the-wild data implies about scale. Specifically, every hand pose corpus derives ground truth from a motion-capture installation or a closed on-device tracker, which confines the diversity and scalability of data collection. Instead, we show that precise and robust slam, depth estimation, and hand tracking do not require motion capture, thereby enabling our hand tracking models to adapt better than models trained with motion capture data.

Both the device and the API are available today. We expect the field to find failure modes in them that we have not, and hope to push the envelope of egocentric data, 3D estimation, and robot learning.

References

[1] Hiroyasu Akada, Jian Wang, Soshi Shimada, Masaki Takahashi, Christian Theobalt, and Vladislav Golyanik. UnrealEgo: A new dataset for robust egocentric 3D human motion capture. In *European Conference on Computer Vision (ECCV)*, pages 1–17, 2022. 2

[2] Hiroyasu Akada, Jian Wang, Vladislav Golyanik, and Christian Theobalt. 3d human pose perception from egocentric stereo videos, 2024. 7

[3] 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 (CVPR)*, 2025. 2, 3, 6, 8, 9

[4] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. ZoeDepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023. 3

[5] Aleksei Bochkovskii, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan R. Richter, and Vladlen Koltun. Depth pro: Sharp monocular metric depth in less than a second. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2410.02073. 3

[6] Michael Burri, Janosch Nikolic, Pascal Gohl, Thomas Schneider, Joern Rehder, Sammy Omari, Markus W. Achtelik, and Roland Siegwart. The EuRoC micro aerial vehicle datasets. *The International Journal of Robotics Research*, 35 (10):1157–1163, 2016. 3, 6

[7] 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 multimap SLAM. *IEEE Transactions on Robotics*, 37(6): 1874–1890, 2021. 2, 6, 7

[8] Zhihao Cao, Hanyu Wu, Li Wa Tang, Zizhou Luo, Wei Zhang, Marc Pollefeys, Zihan Zhu, and Martin R. Oswald. Mcgs-slam: A multi-camera slam framework using gaussian splatting for high-fidelity mapping, 2026. 3

[9] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5410–5418, 2018. 3

[10] Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S. Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, Jan Kautz, and Dieter Fox. Dexycb: A benchmark for capturing hand grasping of objects, 2021. 8, 6

[11] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The EPIC-KITCHENS dataset. In *European Conference on Computer Vision (ECCV)*, pages 720–736, 2018. 2

[12] Thao Dang, Christian Hoffmann, and Christoph Stiller. Continuous stereo self-calibration by camera parameter tracking. *IEEE Transactions on Image Processing*, 18(7):1536–1550, 2009. 2, 5

[13] Mateo de Mayo and Collabora. Monado SLAM datasets. https://huggingface.co/datasets/collabora/monado-slam-datasets, 2023. 6

[14] Mateo de Mayo, Daniel Cremers, and Taihú Pire. The Monado SLAM dataset for egocentric visual-inertial tracking. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2025. 3

[15] Jakob Engel, Kiran Somasundaram, Michael Goesele, Albert Sun, Alexander Gamino, Andrew Turner, Arjang Talattof, Arnie Yuan, Bilal Souti, Brighid Meredith, et al. Project aria: A new tool for egocentric multi-modal ai research. *arXiv preprint arXiv:2308.13561*, 2023. 1, 2, 5

[16] Paul Furgale, Joern Rehder, and Roland Siegwart. Unified temporal and spatial calibration for multi-sensor systems. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1280–1286, 2013. 2, 4, 5

[17] Dorian Gálvez-López and Juan D. Tardós. Bags of binary words for fast place recognition in image sequences. *IEEE Transactions on Robotics*, 28(5):1188–1197, 2012. 2

[18] Ling Gao, Yuxuan Liang, Jiaqi Yang, Shaoxiang Wu, Chenyu Wang, Jiaben Chen, and Laurent Kneip. VECtor: A versatile event-centric benchmark for multi-sensor SLAM. *IEEE Robotics and Automation Letters*, 7(3):8217–8224, 2022. 3, 6

[19] Patrick Geneva, Kevin Eckenhoff, Woosik Lee, Yulin Yang, and Guoquan Huang. OpenVINS: A research platform for visual-inertial estimation. In *IROS 2019 Workshop on Visual-Inertial Navigation: Challenges and Applications*, 2019. 2

[20] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18995–19012, 2022. 2

[21] Giorgio Grisetti, Rainer Kümmerle, Cyrill Stachniss, and Wolfram Burgard. A tutorial on graph-based SLAM. *IEEE Intelligent Transportation Systems Magazine*, 2(4):31–43, 2010. 2

[22] Shreyas Hampali, Sayan Deb Sarkar, and Vincent Lepetit. Ho-3dv3: Improving the accuracy of hand-object annotations of the ho-3d dataset, 2021. 8, 6

[23] Shangchen Han, Beibei Liu, Randi Cabezas, Christopher D. Twigg, Peizhao Zhang, Jeff Petkau, Tsz-Ho Yu, Chun-Jung Tai, Muzaffer Akbay, Zheng Wang, et al. MEgATrack: Monochrome egocentric articulated hand-tracking for virtual reality. *ACM Transactions on Graphics (Proc. SIGGRAPH)*, 39(4):87:1–87:13, 2020. 3

[24] Shangchen Han, Po-Chen Wu, Yubo Zhang, Beibei Liu, Linguang Zhang, Zheng Wang, Weiguang Si, Peizhao Zhang, Yujun Cai, Tomas Hodan, et al. UmeTrack: Unified multi-view end-to-end hand tracking for VR. In *SIGGRAPH Asia 2022 Conference Papers*, pages 1–9, 2022. 3, 8, 9

[25] 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. 1, 2

[26] Karim Iskakov, Egor Burkov, Victor Lempitsky, and Yury Malkov. Learnable triangulation of human pose. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7718–7727, 2019. 3

[27] 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 *IEEE International Conference on Robotics and Automation (ICRA)*, 2025. arXiv:2410.24221. 1, 2, 9

[28] Alexander Korovko, Dmitry Slepichev, Alexander Efitorov, Aigul Dzhumamuratova, Viktor Kuznetsov, Hesam Rabeti, Joydeep Biswas, and Soha Pouya. cuVSLAM: CUDA accelerated visual odometry and mapping. *arXiv preprint arXiv:2506.04359*, 2025. 2, 6, 7

[29] Stefan Leutenegger. OKVIS2: Realtime scalable visual-inertial SLAM with loop closure. *arXiv preprint arXiv:2202.09199*, 2022. 6, 7

[30] Stefan Leutenegger, Simon Lynen, Michael Bosse, Roland Siegwart, and Paul Furgale. Keyframe-based visual-inertial odometry using nonlinear optimization. *The International Journal of Robotics Research*, 34(3):314–334, 2015. 2

[31] Jiankun Li, Peisen Wang, Pengfei Xiong, Tao Cai, Ziwei Yan, Lei Yang, Jiangyu Liu, Haoqiang Fan, and Shuaicheng Liu. Practical stereo matching via cascaded recurrent network with adaptive correlation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16263–16272, 2022. 3

[32] Lahav Lipson, Zachary Teed, and Jia Deng. RAFT-Stereo: Multilevel recurrent field transforms for stereo matching. In *International Conference on 3D Vision (3DV)*, pages 218–227, 2021. 3

[33] Vincent Liu, Ademi Adeniji, Haotian Zhan, Siddhant Haldar, Raunaq Bhirangi, Pieter Abbeel, and Lerrel Pinto. EgoZero: Robot learning from smart glasses. *arXiv preprint arXiv:2505.20290*, 2025. 1, 2, 9

[34] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004. 2, 5

[35] Zhaoyang Lv, Nicholas Charron, Pierre Moulon, Alexander Gamino, Cheng Peng, Chris Sweeney, Edward Miller, Huixuan Tang, Jeff Meissner, Jing Dong, et al. Aria everyday activities dataset. *arXiv preprint arXiv:2402.13349*, 2024. 2

[36] Lingni Ma, Yuting Ye, Fangzhou Hong, Vladimir Guzov, Yifeng Jiang, Rowan Postyeni, Luis Pesqueira, Alexander Gamino, Vijay Baiyya, Hyo Jin Kim, et al. Nymeria: A massive collection of multimodal egocentric daily motion in the wild. In *European Conference on Computer Vision (ECCV)*, 2024. 2

[37] Riku Murai, Eric Dexheimer, and Andrew J. Davison. MASt3R-SLAM: Real-time dense SLAM with 3D reconstruction priors. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 3

[38] Xiaqing Pan, Nicholas Charron, Yongqian Yang, Scott Peters, Thomas Whelan, Chen Kong, Omkar Parkhi, Richard Newcombe, and Yuheng Carl Ren. Aria digital twin: A new benchmark dataset for egocentric 3D machine perception. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20133–20143, 2023. 2, 6

[39] Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3D with transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9826–9836, 2024. 3, 8

[40] Rolandos Alexandros Potamias, Jinglei Zhang, Jiankang Deng, and Stefanos Zafeiriou. WiLoR: End-to-end 3D hand localization and reconstruction in-the-wild. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. arXiv:2409.12259. 3, 8

[41] Tong Qin, Peiliang Li, and Shaojie Shen. VINS-Mono: A robust and versatile monocular visual-inertial state estimator. *IEEE Transactions on Robotics*, 34(4):1004–1020, 2018. 2

[42] Ri-Zhao Qiu, Shiqi Yang, Xuxin Cheng, Chaitanya Chawla, Jialong Li, Tairan He, Ge Yan, David J. Yin, Lars Paulsen Hong, Ge Yuan, et al. Humanoid policy $\sim$ human policy. *arXiv preprint arXiv:2503.13441*, 2025. 1, 9

[43] Helge Rhodin, Christian Richardt, Dan Casas, Eldar Insafutdinov, Mohammad Shafiei, Hans-Peter Seidel, Bernt Schiele, and Christian Theobalt. EgoCap: Egocentric marker-less motion capture with two fisheye cameras. *ACM Transactions on Graphics (Proc. SIGGRAPH Asia)*, 35(6):162:1–162:11, 2016. 2

[44] Patrick Rim, Kevin Harris, Braden Copple, Shangchen Han, Xu Xie, Ivan Shugurov, Sizhe An, He Wen, Alex Wong, Tomas Hodan, and Kun He. Show3d: Capturing scenes of 3d hands and objects in the wild, 2026. 3, 8, 9

[45] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics (Proc. SIGGRAPH Asia)*, 36(6):245:1–245:17, 2017. 3, 8

[46] Yu Rong, Takaaki Shiratori, and Hanbyul Joo. FrankMocap: A monocular 3D whole-body pose estimation system via regression and integration. In *IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 2021. 3

[47] Antoni Rosinol, Marcus Abate, Yun Chang, and Luca Carlone. Kimera: An open-source library for real-time metric-semantic localization and mapping. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 1689–1696, 2020. 2

[48] David Schubert, Thore Goll, Nikolaus Demmel, Vladyslav Usenko, Jörg Stückler, and Daniel Cremers. The TUM VI benchmark for evaluating visual-inertial odometry. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1680–1687, 2018. 3, 6

[49] Tony Tao, Mohan Kumar Srirama, Jason Jingzhou Liu, Kenneth Shaw, and Deepak Pathak. Dexwild: Dexterous human interactions for in-the-wild robot policies, 2026. 2

[50] Zachary Teed and Jia Deng. DROID-SLAM: Deep visual SLAM for monocular, stereo, and RGB-D cameras. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 16558–16569, 2021. 3

[51] Vladyslav Usenko, Nikolaus Demmel, David Schubert, Jörg Stückler, and Daniel Cremers. Visual-inertial mapping with non-linear factor recovery. *IEEE Robotics and Automation Letters*, 5(2):422–429, 2020. 2, 6, 7

[52] Chen Wang, Haochen Shi, Weizhuo Wang, Ruohan Zhang, Li Fei-Fei, and C. Karen Liu. DexCap: Scalable and portable mocap data collection system for dexterous manipulation. In *Robotics: Science and Systems (RSS)*, 2024. 2

[53] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 3

[54] Xianqi Wang, Gangwei Xu, Hao Jia, and Xin Yang. Selective-stereo: Adaptive frequency information selection for stereo matching. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19701–19710, 2024. 3

[55] Bowen Wen, Matthew Trepte, Joseph Aribido, Jan Kautz, Orazio Gallo, and Stan Birchfield. FoundationStereo: Zero-shot stereo matching. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 3, 7

[56] Bowen Wen, Shaurya Dewan, and Stan Birchfield. FastFoundationStereo: Real-time zero-shot stereo matching. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026. 3, 7

[57] Gangwei Xu, Xianqi Wang, Xiaohuan Ding, and Xin Yang. Iterative geometry encoding volume for stereo matching. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21919–21928, 2023. 3

[58] Weipeng Xu, Avishek Chatterjee, Michael Zollhöfer, Helge Rhodin, Pascal Fua, Hans-Peter Seidel, and Christian Theobalt. Mo^{2}Cap^2: Real-time mobile 3D motion capture with a cap-mounted fisheye camera. *IEEE Transactions on Visualization and Computer Graphics*, 25(5):2093–2101, 2019. 2

[59] Lixin Yang, Kailin Li, Xinyu Zhan, Fei Wu, Anran Xu, Liu Liu, and Cewu Lu. Oakink: A large-scale knowledge repository for understanding hand-object interaction, 2022. 8, 6

[60] Lixin Yang, Jian Xu, Licheng Zhong, Xinyu Zhan, Zhicheng Wang, Kejian Wu, and Cewu Lu. POEM: Reconstructing hand in a point embedded multi-view stereo. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21108–21117, 2023. 3

[61] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

[62] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 3

[63] Lixin Yang, Licheng Zhong, Pengxiang Zhu, Xinyu Zhan, Junxiao Kong, Jian Xu, and Cewu Lu. Multi-view hand reconstruction with a point-embedded transformer, 2024. 8, 9, 2, 6

[64] Wei Yin, Chi Zhang, Hao Chen, Zhipeng Cai, Gang Yu, Kaixuan Wang, Xiaozhi Chen, and Chunhua Shen. Metric3D: Towards zero-shot metric 3D prediction from a single image. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9043–9053, 2023. 3

[65] Zhengdi Yu, Stefanos Zafeiriou, and Tolga Birdal. DynHaMR: Recovering 4D interacting hand motion from a dynamic camera. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 3

[66] Fan Zhang, Valentin Bazarevsky, Andrey Vakunov, Andrei Tkachenka, George Sung, Chuo-Ling Chang, and Matthias Grundmann. MediaPipe hands: On-device real-time hand tracking. *arXiv preprint arXiv:2006.10214*, 2020. 3

[67] Jinglei Zhang, Jiankang Deng, Chao Antonopoulos, Rolandos Alexandros Potamias, and Stefanos Zafeiriou. HaWoR: World-space hand motion reconstruction from egocentric videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 3

[68] Zihan Zhu, Wei Zhang, Moyang Li, Norbert Haala, Marc Pollefeys, and Daniel Barath. Vigs-slam: Visual inertial gaussian splatting slam, 2026. 3

RoboCap: A New Platform for Egocentric Robot Learning

Supplementary Material

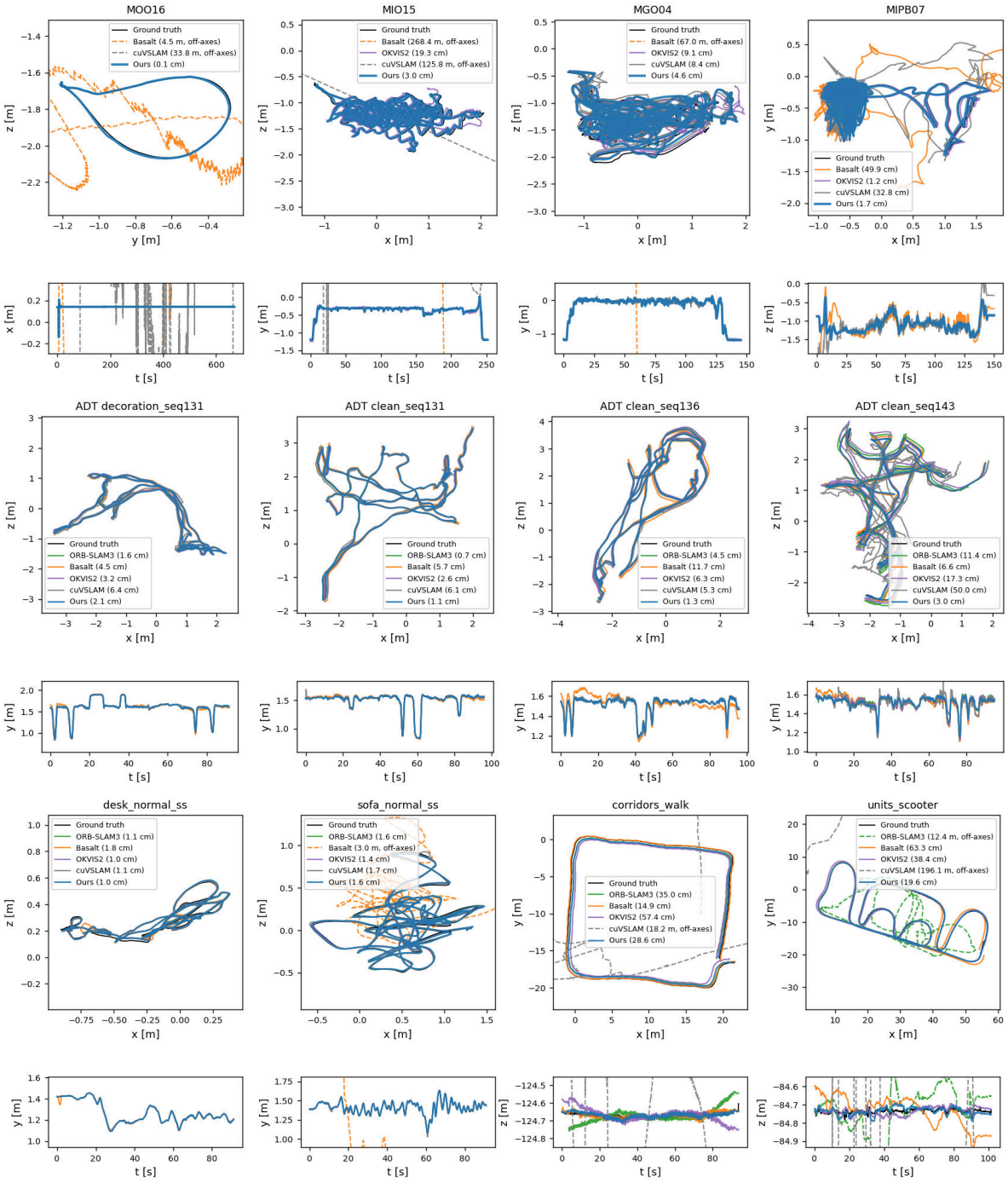


Figure 7. Overlays of SLAM trajectories on Monado (top), ADT (middle), and VECtor (bottom).

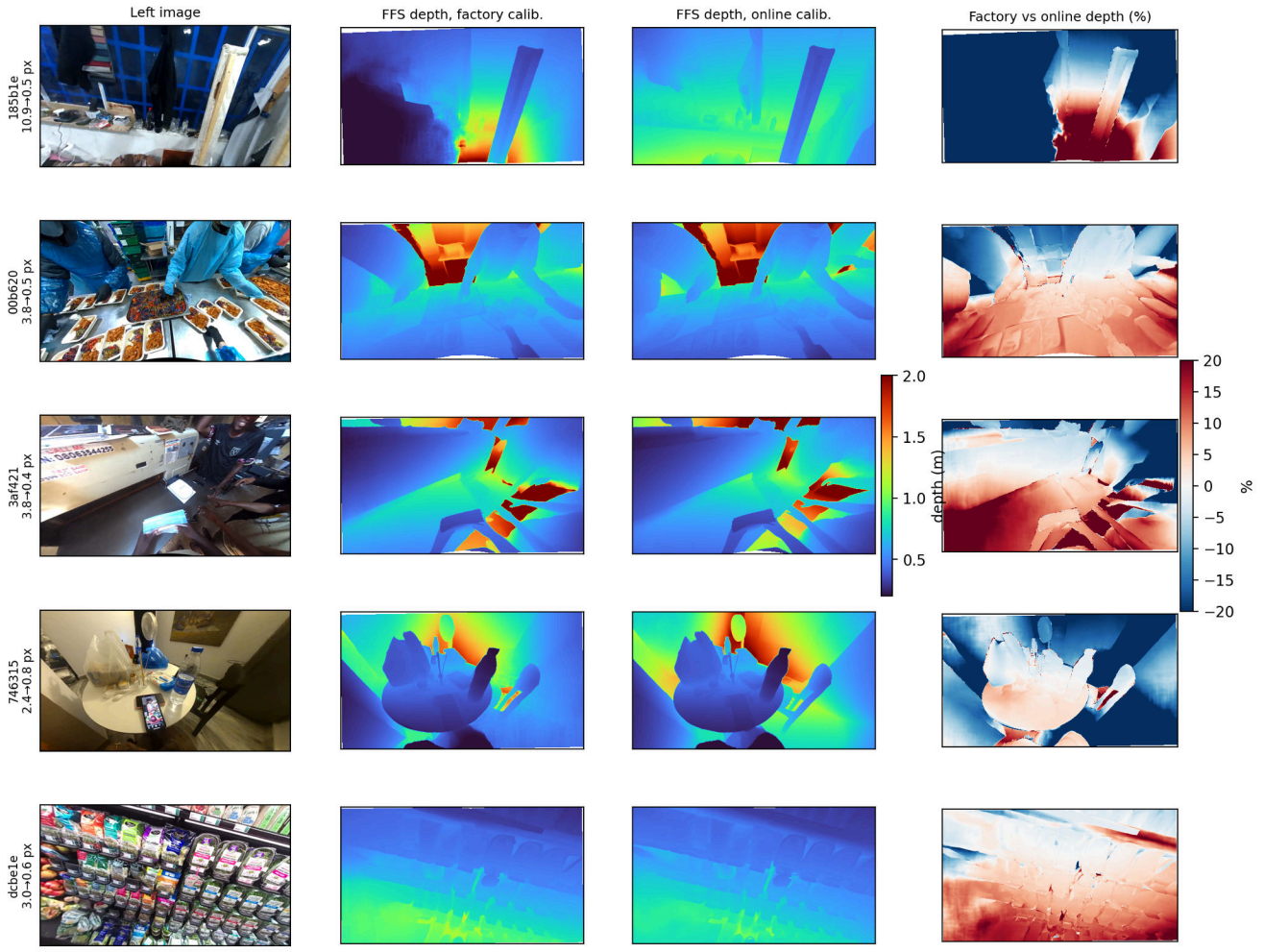


Figure 8. More visualizations of FFS depth with factory vs. online calibration, and the depth change between them.

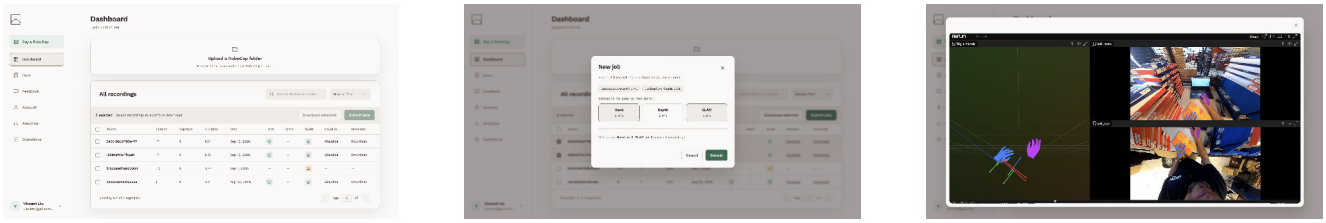


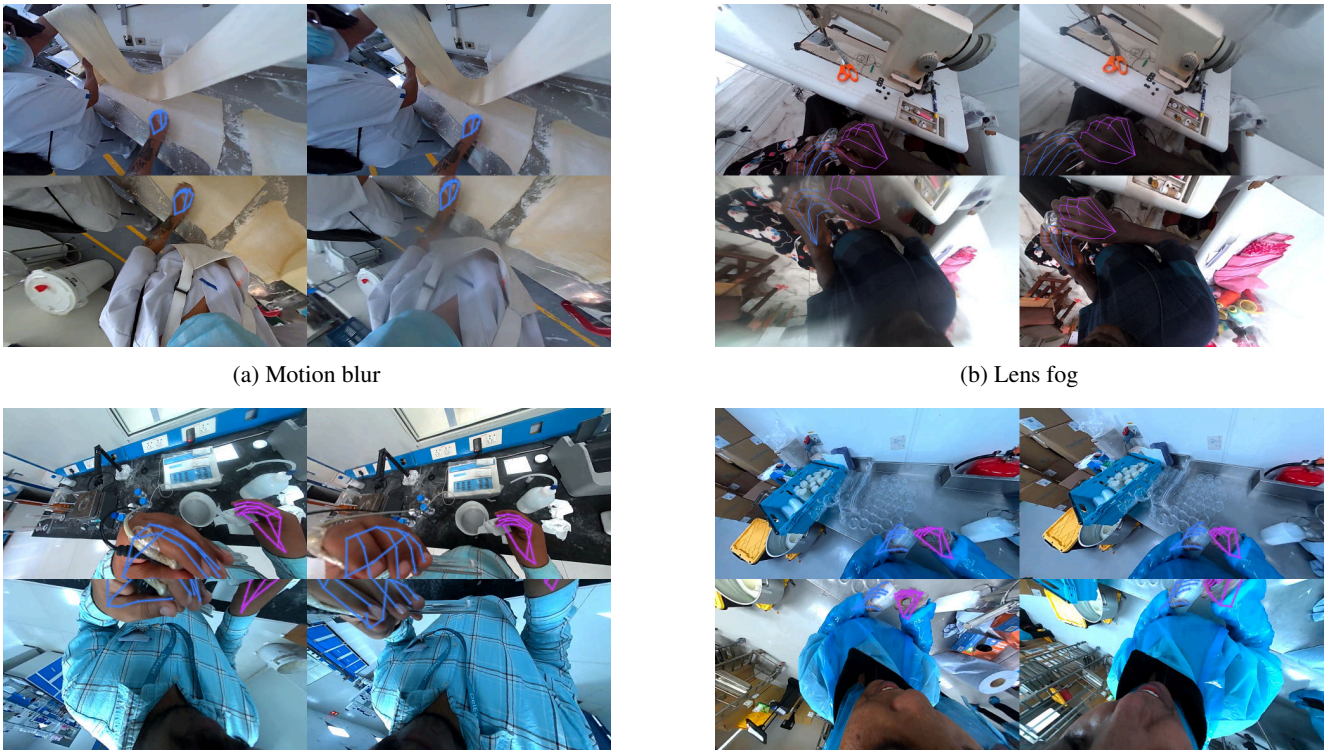
Figure 9. The Grounded API, which allows users to upload RoboCap sessions and submit jobs directly from a dashboard. The dashboard also embeds a 3D visualizer to view hand tracking and SLAM outputs. The CLI supports batch uploads, submissions, and downloads.

A. Hand tracking

We show additional qualitative results on held-out in-the-wild recordings, covering the cases that are hardest for egocentric capture. Figure 10 shows degraded imaging and extreme viewpoints, and Figure 11 shows hand-object interaction and unusual hand configurations. In each case the model produces a plausible, multi-view consistent mesh without per-view post-processing.

A.1. Evaluation Setup

Metric We report mean per-joint position error (MPJPE) over the 21 MANO joints in millimetres. Following [63], error is absolute: it is measured in the frame of the first camera and includes global translation, with no root alignment and no Procrustes alignment. Camera intrinsics and



(a) Motion blur

(b) Lens fog

(c) Hand close to camera

(d) Transparent gloves

Figure 10. Robustness to degraded imaging and extreme viewpoints. Each panel shows the four camera views of a single frame with the predicted mesh overlaid. The model remains multi-view consistent when the hand is blurred by fast motion, when the lens is fogged, when the hand is close enough to the camera to fill the frame, and when the hand is seen through a transparent glove.

extrinsics are assumed known at both training and evaluation time.

A.2. Egocentric datasets

All three benchmarks UmeTrack [24], HOT3D [3] or SHOW3D [44] withhold the test set from public release entirely. When evaluating on these benchmarks, we train with a 5x lower learning rate than during pretraining. The two columns of Table 5 differ only in weight initialization. Figure 12 shows sample predictions from the scratch and pretrained models on the same frames, for each of the egocentric benchmarks.

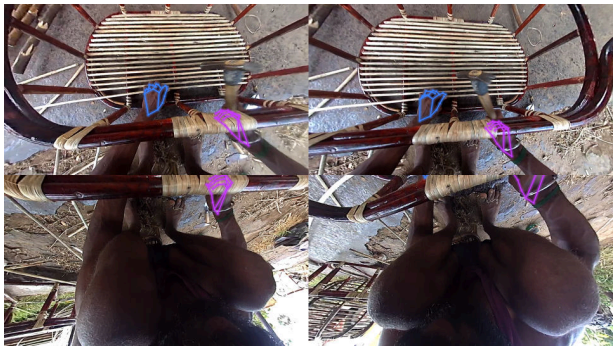
UmeTrack UmeTrack is a real-time hand tracking system and accompanying dataset captured on a VR headset carrying four wide-FOV monochrome VGA cameras, of which a given hand is typically visible in one or two. It targets absolute world-space pose rather than root-relative pose. The corpus comprises 53 participants and 1,397 real sequences of 15 seconds at 30 Hz, each paired with a synthetic render, for roughly 839k frames of either kind. Ground truth comes from a 36-camera motion-capture system using 3 mm markers, under two protocols: separate-hand and hand-hand interaction. The public release contains 41 subjects and 2,041 sequences. We hold out four subjects, matching the official pose track. Our held-out set is real-only, since the official splits contain no synthetic sequences. No subject performs both activity types, so the interaction ratio is fixed entirely by subject choice; three separate-hand subjects plus one hand-hand subject is the only composition that approaches the official ratio, giving 103 sequences at 12.6% hand-hand against the official 95 sequences at 11.6%. Because synthetic renders share no sequence identifier with their real counterparts, the split keys on subject, and the held-out subjects' renders are discarded rather than trained on.

HOT3D HOT3D is an egocentric dataset for 3D hand and object tracking, comprising 833 minutes of recording across 19 participants, 33 objects and 425 sessions, for over 1.5M multi-view frames and 3.7M images. It is captured on two devices: Project Aria, contributing one 1408×1408 RGB fisheye camera at 110° field of view and two 640×480 monochrome fisheye cameras at 150°; and Quest 3, contributing two front-facing 1280×1024 monochrome fisheye cameras. Hand ground truth is obtained from several dozen OptiTrack cameras tracking 19 markers per hand, fitted to a

(a) Small objects

(b) Small objects

(c) Tool use



(d) Tool use (with occlusion)

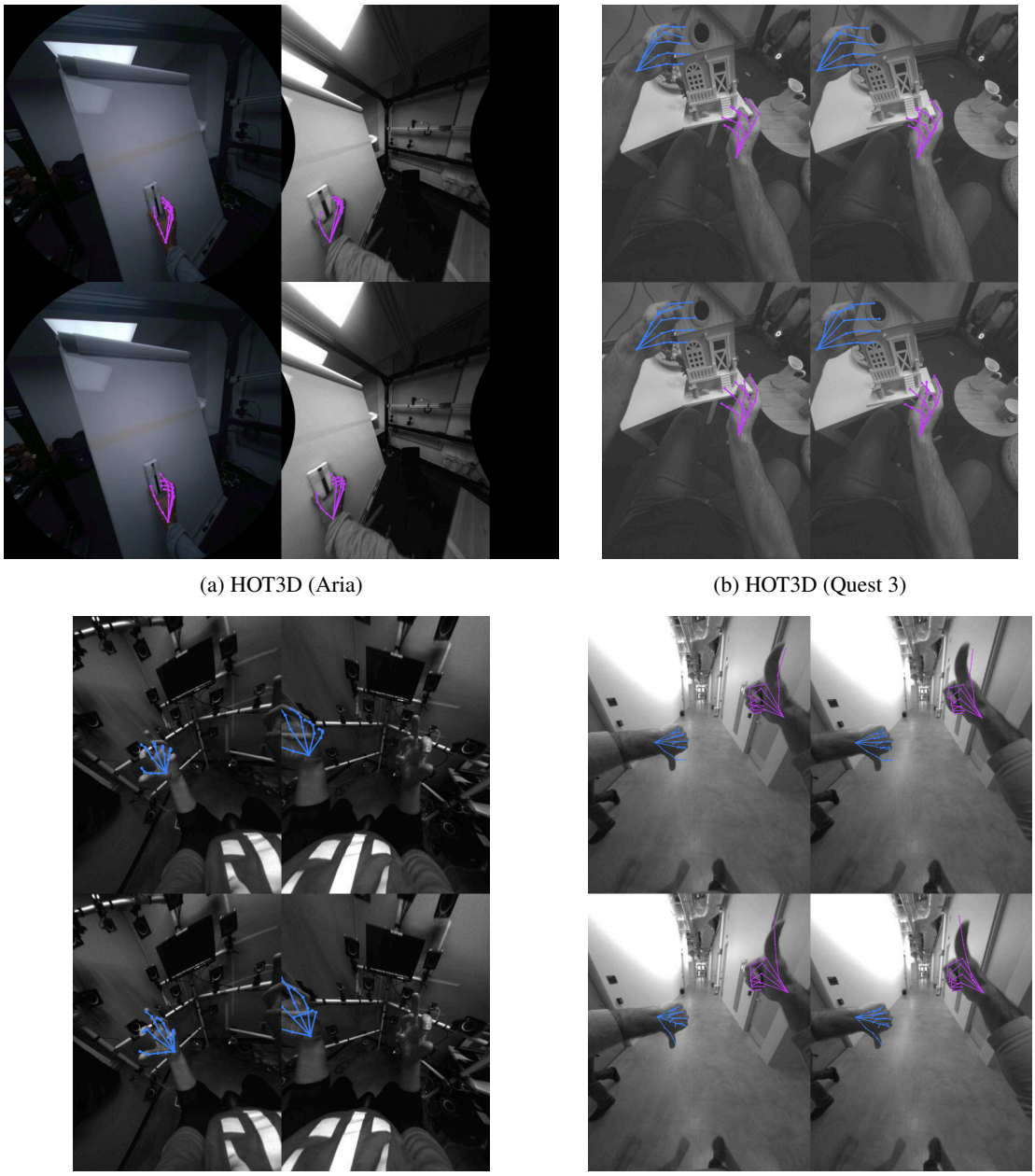
(e) Drawing

(f) Feet and Hands

Figure 11. Robustness to hand-object interaction and unusual configurations. Each panel shows the four camera views of a single frame with the predicted mesh overlaid. The model recovers plausible poses while grasping small objects, while holding a tool with the fingers largely occluded by it, during fine manipulation such as drawing, and when the hands appear alongside other body parts.

per-participant hand model and released in both UmeTrack and MANO parameterisations. The public release covers 13 subjects across the two devices. We hold out two subjects, matching the official pose track. We select the held-out subjects jointly across both devices rather than independently per device, as selecting the Aria and Quest 3 held-out sets independently would place the same person in training and validation. The resulting held-out set is 44 sessions and 132,944 frames, spanning 21 Aria and 23 Quest 3 sessions.

SHOW3D SHOW3D is a marker-less capture dataset combining egocentric and exocentric views: ten synchronised monochrome fisheye cameras at 60 Hz, of which eight are exocentric 1024×1280 cameras at $152^\circ \times 116^\circ$ field of view mounted on a wearable rig, and two are egocentric cameras from a user-worn Quest 3. It comprises roughly 20 hours of recording over 38 participants, more than 1,200 sequences and more than 30 locations, for 4.3M frames and 43M images. Ground truth is produced without markers, by triangulating multi-view 2D keypoint predictions and fitting



(a) HOT3D (Aria)

(b) HOT3D (Quest 3)

(c) UmeTrack

(d) SHOW3D

Figure 12. Comparison of random vs pretrained weights on egocentric benchmarks. Within each panel the columns are the left and right camera views of a single frame, the top row is the model trained from scratch and the bottom row is the model initialized from our pretrained weights; both are trained until they achieve their best multi-dataset validation metrics. The pretrained model aligns more closely with the observed hand, particularly at the finger joints, where the scratch model tends to mispredict articulation while keeping the global hand position roughly correct.

a personalised linear-blend-skinning mesh by inverse kinematics, with a reported sub-centimetre median accuracy. The public release contains 32 subjects and 1,689 scenes. We attempt to mimick the private test set by holding out six subjects, subject to the constraints that every object class remains represented in the training pool and that both sides retain sufficient object-pose coverage. The resulting held-out set is 337 scenes and 581,795 frames at 20.0% of the corpus, covering 23 of 28 object classes, with a novel-activity rate of 12.7% against the official split's 12.2%. SHOW3D

	Held-out (ours)		Official test	
	subjects	share	subjects	share
UmeTrack	4	10.1%	4	—
SHOW3D	6	20.0%	6	21.0%
HOT3D	2	16.0%	2	—

Table 7. Composition of the surrogate held-out sets. UmeTrack: 103 sequences, 12.6% hand-hand interaction against the official split's 11.6%; real sequences only. SHOW3D: 337 scenes covering 23 of 28 object classes against the official split's 23; novel-activity rate 12.7% against 12.2%. HOT3D: two held-out subjects per device, matching the official pose track.

	views	train	test	split
DexYCB-Mv [10]	8	25,387	4,951	S0
OakInk-Mv [59]	4	58,692	19,909	SP2
HO3D-Mv [22]	5	9,087	2,706	7 seq.

Table 8. Multi-view benchmark splits, in multi-view frames, following [63]. HO3D-Mv comprises the seven sequences: ABF1, BB1, GSF1, MDF1 and SiBF1 for training, GPMF1 and SB1 for evaluation.

supplies no MANO annotations, so its supervision and evaluation are joints-only.

A.3. Multi-view benchmarks

We follow the protocol of POEMv2 [63] exactly, including its multi-view subsets. Following its protocol, we only train on the right hands within each dataset. The number and ordering of cameras is fixed per dataset.

A single checkpoint is finetuned for 20k steps on the union of the three training splits and evaluated on all three test splits, with no per-dataset specialisation. This matches POEMv2, which trains one model on a mixture and evaluates it on each test set in turn; POEMv2's mixture additionally includes ARCTIC, InterHand2.6M and FreiHAND, and is trained for 2.1M iterations. As above, we reuse the pretraining configuration with the learning rate reduced $5\times$.

DexYCB DexYCB was built to benchmark hand grasping of objects. It is captured by eight statically mounted, calibrated RGB-D cameras at 640×480 and 30 Hz, and comprises 10 subjects manipulating 20 YCB objects over 1,000 sequences, for 582k frames. We adopt the official S0 split and the eight-camera construction of [63], giving 25,387 training, 1,412 validation and 4,951 test multi-view frames.

OakInk OakInk is a knowledge repository for hand-object interaction. Its image component is captured by four calibrated RGB cameras at 848×480 and 30 Hz, covering 12 subjects and 100 objects over roughly 230k images. We adopt the official SP2 object-based split and the four-camera construction of [63], giving 58,692 training and 19,909 test multi-view frames. Roughly a quarter of OakInk sequences record two people handing an object over, placing both hands in frame; each hand is trained and evaluated independently.

HO3D HO3D provides 3D hand and object pose annotations for hand-object manipulation, obtained by optimising against 2D keypoint and segmentation cues rather than from markers; v3 refines the annotation accuracy of v2. It covers 10 subjects and 10 YCB objects over 68 sequences and roughly 103k RGB-D frames at 640×480, of which only a subset is recorded from multiple cameras. We adopt the construction of [63], which retains the seven sequences with five-camera coverage: ABF1, BB1, GSF1, MDF1 and SiBF1 for training, and GPMF1 and SB1 for evaluation, giving 9,087 training and 2,706 test multi-view frames. Two of the five cameras are withheld from the SB1 evaluation sequence. Since absolute error is measured in the frame of the first camera, the missing views place the reference origin on a different physical camera, approximately 0.35 m from its position in every training sequence.

A.4. Synthetic Dataset

We render a synthetic dataset using the camera geometry of our RoboCap devices, which yields exact ground-truth MANO parameters. We render photorealistic scenes and hand textures to limit. Each sequence is rendered through a camera rig whose intrinsics and extrinsics are sampled from the calibrations of real RoboCap devices, so that the synthetic imagery reproduces the field of view, baseline and lens distortion of our hardware.