IMUHAND BENCHMARK: A DATASET AND BASELINE FOR 3D HAND MOTION RECONSTRUCTION FROM SPARSE IMUS

Anonymous authors
Paper under double-blind review

ABSTRACT

Continuous 3D hand motion reconstruction is fundamental to natural human-computer interaction and robotic manipulation. Vision-based methods exploit rich appearance and scene information but are limited by occlusion and restricted fields of view during hand-to-hand contact and hand-object interaction. Hand-mounted inertial measurement units (IMUs) do not require optical visibility, yet sparse inertial observations are insufficient to directly recover the full articulated hand pose and are affected by orientation drift and variations in sensor placement. Existing inertial hand datasets differ in acquisition settings and annotation formats. The lack of large-scale public datasets and standardized evaluation protocols for continuous IMU-only bimanual reconstruction further hinders fair comparison. To address this gap, we develop a pair of IMU gloves for synchronized data acquisition and real-time reconstruction. Using this system, we construct IMUHand Benchmark, which provides long, continuous bimanual motion sequences, per-frame MANO ground truth, synchronized egocentric video, and a unified IMU-only evaluation protocol. We further propose a reconstruction method that combines learned secondary calibration with independent root orientation estimation to recover wrist orientation and the full articulated hand pose from IMU sequences alone. Under the unified protocol, our method achieves the highest reconstruction accuracy among the inertial methods evaluated.

1 INTRODUCTION

Continuous 3D human hand pose reconstruction is a key capability for augmented reality, virtual reality, human-computer interaction, and robotic teleoperation. Egocentric vision captures hand appearance and the surrounding interaction context, making it a primary input modality for hand reconstruction. However, occlusion, motion blur, and restricted fields of view can leave visual evidence incomplete during hand-to-hand contact, object manipulation, and rapid motion. Hand-mounted inertial measurement units (IMUs) continue to record motion under these conditions, providing an alternative source of observations when vision is limited.

However, inertial reconstruction is challenging because IMUs measure sensor orientations and accelerations without directly observing every joint. Pose inference need also account for orientation drift, mounting errors, and signal variations caused by repeated wearing. Existing methods address some of these challenges but are not readily applicable to a unified bimanual MANO reconstruction task. FSGlove achieves tracking through hand-shape calibration and 16 IMUs per hand, requiring a dense sensor layout (Li et al., 2025). Sarker et al. (2026) combine temporal prediction and filtering to improve joint-angle accuracy, but their outputs and evaluation targets do not cover the full MANO pose. MGDHand learns pose priors from a MANO-IMU teacher, but its evaluation uses single-hand VIHand data and does not assess continuous bimanual reconstruction (Wang et al., 2026).

Discrepancies between datasets limit consistent evaluation of bimanual hand reconstruction. VIHand provides hand-mounted IMU measurements and MANO annotations but mainly focuses on single-hand sequences from external viewpoints (Wang et al., 2025). HandSense captures single-hand glove signals and hand-object interactions without complete MANO annotations (Zhang et al., 2026). EgoEMG provides bimanual MANO annotations and egocentric video, but its wrist-mounted

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Table 1: Comparison of related hand datasets. Hand IMUs exclude sensors mounted only at the wrists or on the body. Egocentric RGB denotes recorded color video. Public release indicates access without author approval.

Dataset	Bimanual	Hand IMUs	Egocentric RGB	MANO	Hand-object interaction	Public release
VIHand (Wang et al., 2025)	×	✓	×	✓	×	×
HandSense (Zhang et al., 2026)	×	✓	×	×	✓	×
EgoEMG (Xi et al., 2026)	✓	×	✓	✓	×	✓
KIT (Krebs et al., 2021)	✓	×	✓	×	✓	✓
IMUHand (ours)	✓	✓	✓	✓	✓	✓

IMUs do not directly capture finger motion (Xi et al., 2026). KIT supports bimanual data-glove recordings, yet its IMUs are body-mounted and hand motion is represented by MMM rather than MANO (Krebs et al., 2021). As summarized in Table 1, no existing dataset jointly offers hand-mounted IMUs, long-duration bimanual sequences, per-frame MANO ground truth, and an IMU-only evaluation protocol, limiting method comparison and systematic analysis of long-term drift.

To enable such evaluation, we develop a pair of IMU gloves and integrate them with an egocentric stereo camera, an independent optical motion capture system, and a synchronized data acquisition pipeline. Each glove is equipped with six IMUs in a fixed configuration, with one sensor mounted on each fingertip and one on the dorsum of the hand. The optical motion capture system provides accurate 3D motion trajectories for generating per-frame MANO ground truth. Based on this setup, IMUHand Benchmark comprises continuous sequences covering basic hand motions, bimanual coordination, and hand-object interactions, together with standardized sensor configurations, data splits, and evaluation protocols. We further establish an IMU-based baseline for estimating wrist orientation and full articulated hand pose, ensuring that the dataset, task, and baseline are evaluated under consistent hardware and sensing conditions.

Our main contributions are as follows:

- **Bimanual IMU gloves and a synchronized acquisition system.** We develop a pair of gloves with a fixed sensor layout and integrate them with an egocentric stereo camera and independent optical motion capture. The system provides continuous inertial observations and corresponding 3D annotations through synchronized acquisition.
- **The IMUHand dataset and benchmark.** We construct a dataset of long, continuous bimanual motion sequences with per-frame MANO ground truth, covering both object-free hand motion and hand-object interaction. We define a unified IMU-only task, data splits, and evaluation protocol to support consistent method comparison and assessment of drift and pose errors over continuous sequences.
- **An accurate IMU-only reconstruction baseline.** We introduce a Canonicalizer that performs learned secondary calibration to address residual errors caused by repeated donning and motion disturbances after static calibration. The pose estimator recovers local joint poses from canonicalized inputs, while a separate Root Bypass estimates global wrist orientation from the original back-of-hand measurements. Virtual IMUs provide correction targets and support three-stage training. Under the unified protocol with six IMUs per hand, our method achieves the highest reconstruction accuracy among the inertial methods evaluated.

2 Related Work

2.1 3D Hand Datasets and Benchmarks

Existing hand datasets primarily target visual perception, covering object manipulation, assembly, bimanual coordination, and egocentric capture (Kwon et al., 2021; Ohkawa et al., 2023; Fan et al., 2023; Zhan et al., 2024; Banerjee et al., 2025; Grauman et al., 2024). Inertial and multimodal datasets complement them with wearable signals: KIT and HANDMI4 provide data-glove or inertial

recordings (Krebs et al., 2021; Mongelli & Menolotto, 2023), VIHand, DexGloveHOI, and HandSense pair hand-mounted IMUs with visual data, and EgoEMG records inertial and electromyographic signals from both wrists (Wang et al., 2025; Kou et al., 2026; Zhang et al., 2026; Xi et al., 2026). These datasets differ in sensor layout, hand coverage, and annotation representation, and none jointly provides hand-mounted IMUs, continuous bimanual sequences, and per-frame MANO annotations with an IMU-only evaluation protocol. IMUHand fills this gap with a fixed six-IMU-per-hand layout and a unified benchmark.

2.2 Vision-Based and Visual-Inertial Hand Reconstruction

Vision-based methods recover hand pose from single images or exploit temporal context in video (Pavlakos et al., 2024; Zhang et al., 2025; Liu et al., 2026), but remain vulnerable to occlusion and restricted fields of view. Visual-inertial methods mitigate this by adaptively weighting images and IMU observations (Kou et al., 2026), yet still rely on cameras at inference, whereas IMUHand Benchmark targets reconstruction from inertial observations alone.

2.3 Inertial Pose Reconstruction

Sparse inertial full-body reconstruction learns motion priors and structural constraints from a few sensors (Huang et al., 2018; Yi et al., 2021; 2022; Zhang et al., 2024). For hands, Grapentin et al. (2020) recover bone orientations using kinematic constraints and moving-horizon estimation, while FSGlove requires a dense 16-IMU layout with differentiable MANO calibration (Li et al., 2025). Learning-based methods improve accuracy through drift-insensitive features with Kalman filtering (Sarker et al., 2026), visual-inertial knowledge distillation (Wang et al., 2025), and MANO-IMU prior transfer (Wang et al., 2026), but the latter two depend on visual teachers during training. Our baseline instead uses IMU signals only in both training and inference, targets continuous bimanual reconstruction with six IMUs per hand, and achieves the highest accuracy among the evaluated inertial methods under the unified protocol of IMUHand Benchmark.

3 Hardware

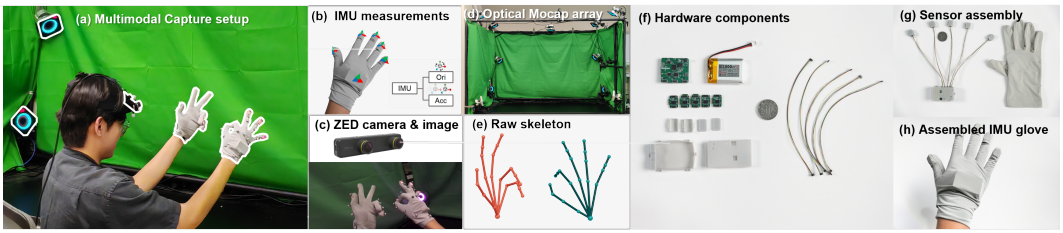


Figure 1: Overview of the IMUHand hardware and multimodal capture system. (a) A participant wearing a pair of IMU gloves and a head-mounted ZED stereo camera inside the ChingMu capture volume. (b) Six IMUs per hand measure orientations and accelerations at the five fingertips and the back of the hand. (c) The ZED stereo camera and an example egocentric image. (d) The 12-camera ChingMu optical motion-capture array and (e) its raw 21-node skeletons for both hands. (f–h) Glove construction, from the hardware components and sensor assembly to the completed wearable glove.

To support continuous 3D pose capture of both hands, we design a pair of wireless data gloves based on sparse inertial sensing. Each glove integrates six BNO085 inertial measurement units (IMUs), an ESP32-based controller, and a rechargeable battery. It acquires 3D acceleration and quaternion orientation from all six IMUs and transmits the measurements to the acquisition host in UDP packets over 2.4 GHz Wi-Fi at approximately 50 Hz. Electronics, transmission, and timing details are provided in Appendix D.

Sensor configuration and placement. Both gloves use the same six-IMU layout. Five fingertip IMUs are mounted on the dorsal surfaces of the distal phalanges, and the sixth is integrated into the main board on the back of the hand. The back-of-hand IMU serves as the root sensor, providing a reference for overall hand orientation, while each fingertip IMU serves as a leaf sensor whose

orientation reflects the accumulated joint rotations along the finger's kinematic chain. The fingertip modules avoid the main joint flexion regions and leave the finger pads and palm contact surfaces unobstructed for grasping and object manipulation.

Wearable integration. The gloves are made of lightweight stretch fabric that covers the backs of the fingers and hands while leaving the palms exposed. Sensor modules are secured with adhesive and stitching, and wires are routed along the backs of the fingers with slack around the joints to accommodate bending.

4 IMUHAND BENCHMARK

4.1 DATA ACQUISITION SYSTEM

IMU gloves. The acquisition system receives inertial inputs from the pair of gloves described in Section 3. Each of the six sensors per hand continuously records rotation and acceleration, with timing information and sensor identifiers retained. Each recording session begins with static calibration after donning the gloves, followed by continuous bimanual motion capture.

ZED stereo camera. We mount a ZED stereo camera (Stereolabs) on the participant's head using an elastic strap and hook-and-loop fasteners. The camera synchronously records left and right egocentric RGB video at 30 Hz, capturing bimanual motion, object interactions, and the surrounding environment.

ChingMu optical motion capture. The ChingMu system (ChingMu) uses 12 infrared cameras surrounding the capture area to track reflective markers on both hands, with 21 markers per hand. It outputs hand-root positions, skeletal rotations, and marker positions at approximately 120 Hz and transmits them to the host through VRPN (Taylor et al., 2001). Each hand is represented by a 21-node skeleton comprising one root node and five four-node finger chains. We track the ZED pose using a separate rigid marker body attached to the camera.

4.2 CALIBRATION AND SYNCHRONIZATION

After each donning, participants hold a stationary flat-hand pose for approximately 10 s. The mean sensor orientations, flat-hand reference, and predefined back-of-hand mounting relationship are used to align the inertial and MANO frames, estimate sensor-to-bone offsets, correct static acceleration offsets, and form back-of-hand-relative measurements. We register the ZED left camera to its mocap-tracked marker body using checkerboard calibration (Zhang, 2000) and obtain the right-camera pose from the calibrated stereo extrinsics. The acquisition host retains timestamps for all streams; following NTP-based host–glove clock initialization (Mills et al., 2010), we fit an affine mapping from each device clock to a common timeline, set the first mocap frame as the time origin, and resample all modalities at 30 Hz. Full calibration and synchronization details are provided in Appendix A.

4.3 GROUND-TRUTH ANNOTATION

We convert each ChingMu hand skeleton into MANO shape, pose, and translation parameters (Romero et al., 2017). Because ChingMu and MANO differ in skeletal topology, rest pose, and joint-axis definitions, we fit the two hand sides separately rather than directly treating ChingMu rotations as MANO parameters. Before fitting, we repair short tracking gaps and anomalous jumps, reconstruct the skeleton using the fixed bone lengths of its template, and mark frames that cannot be reliably recovered as invalid.

Parametric hand representation. At frame t, the parametric hand is represented as

$$(\mathbf{V}_t, \mathbf{J}_t) = \mathcal{M}(\boldsymbol{\beta}, \boldsymbol{\theta}_t, \mathbf{t}_t), \quad \boldsymbol{\beta} \in \mathbb{R}^{10}, \quad \boldsymbol{\theta}_t \in \mathbb{R}^{48}, \quad \mathbf{t}_t \in \mathbb{R}^3, \tag{1}$$

where $\boldsymbol{\beta}$ denotes hand shape and is fixed for each skeletal template, $\boldsymbol{\theta}_t$ contains the global wrist orientation and 15 local joint rotations in axis–angle form, and $\mathbf{t}_t$ is the global translation. The outputs $\mathbf{V}_t$ and $\mathbf{J}_t$ are the MANO mesh vertices and 3D joints, respectively.

Shape and pose estimation. For each hand and skeletal template, we estimate one shape vector from the template's finger and palm proportions and keep it fixed across frames. For each valid frame, an analytical mapping from the ChingMu rotations provides the initial MANO pose; local corrective rotations then refine this initialization by aligning wrist-relative fingertips and finger-bone directions, with pose and temporal regularization. The resulting annotations retain the original frame identifiers and timestamps for multimodal alignment. Automated validity checks and manual review jointly control annotation quality. Further details are provided in Appendix B.

4.4 Dataset Composition

Table 1 compares sensor configurations and annotations across datasets. Existing datasets primarily focus on a single hand or record inertial signals only at both wrists, limiting simultaneous observation of both hands and individual fingers. Data collected with densely instrumented gloves also differ from the sparse-input reconstruction setting. With only six IMUs per hand, IMUHand synchronously records fingertip and back-of-hand motion on both hands and provides egocentric stereo RGB video and complete per-frame MANO annotations. This enables evaluation of full bimanual pose reconstruction under a fixed sparse sensor configuration.

IMUHand comprises 630 single-hand sequences from 10 subjects, totaling over 2M valid synchronized frames, each with corresponding MANO annotations. The dataset covers both hand-only motion and hand–object interaction, including basic gestures, bimanual coordination, and diverse everyday object manipulations (Fig. **??**). Participants' hand lengths and bony palm widths range from 16.1 to 19.7 cm and from 5.5 to 7.0 cm, respectively.

5 Method

5.1 Problem Formulation

We reconstruct continuous 3D hand poses from six IMUs per hand: one back-of-hand sensor ($s = 0$) and five fingertip sensors ($s = 1, \dots, 5$). Each provides rotation $\mathbf{R}^I_{t,s} \in \mathrm{SO}(3)$ and acceleration $\mathbf{a}^I_{t,s} \in \mathbb{R}^3$ at time t, following the calibration conventions in Section 4.2. A length-T input sequence is

$$\mathbf{X}_{1:T} = \left(\left\{ \left(\mathbf{R}^I_{t,s}, \mathbf{a}^I_{t,s} \right) \right\}_{s=0}^{5} \right)_{t=1}^{T}. \tag{2}$$

A causal model predicts global wrist orientation and local MANO rotations using only current and past observations:

$$f_\phi(\mathbf{X}_{1:t}) = \left(\widehat{\mathbf{R}}^{\mathrm{root}}_t, \widehat{\boldsymbol{\theta}}_t \right), \qquad \widehat{\boldsymbol{\theta}}_t = \left(\widehat{\mathbf{R}}^{\mathrm{local}}_{t,j} \right)_{j=1}^{15} \in \mathrm{SO}(3)^{15}, \tag{3}$$

where ϕ denotes model parameters, $\widehat{\mathbf{R}}^{\mathrm{root}}_t \in \mathrm{SO}(3)$ is the wrist orientation in a common world frame, and $\widehat{\boldsymbol{\theta}}_t$ contains parent-relative joint rotations. Both hands share these definitions.

Uninstrumented phalanges and joints are inferred from these sparse observations. Both training and inference use only IMU signals as sensory inputs.

MANO converts predictions to joints and meshes with mean shape ($\boldsymbol{\beta} = \mathbf{0}$) and wrist-centered coordinates. We estimate orientation but not shape or global translation, so relative bimanual positions are not recovered.

5.2 Proposed Approach

Figure 2 summarizes three components. The Canonicalizer corrects residual errors after flat-pose calibration, the Pose Estimator recovers local MANO poses from corrected signals, and the Root Bypass estimates wrist orientation from original back-of-hand observations. Training uses virtual IMUs derived from MANO annotations, whereas inference uses glove IMUs only.

Input representation. Let $r = 0$ denote the back-of-hand sensor and $\mathcal{S} = \{1, \dots, 5\}$ the fingertip sensors. Their calibrated measurements in the MANO reference frame are $\mathbf{R}^M_{t,s}$ and $\mathbf{a}^M_{t,s}$. With the

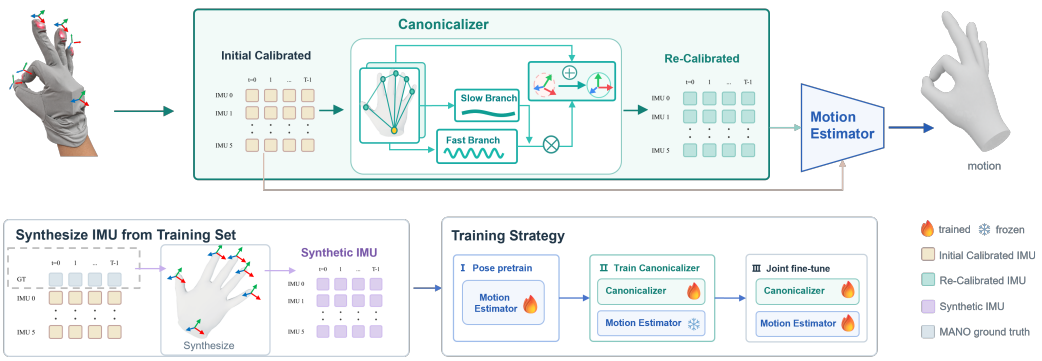


Figure 2: Overview of our method. The Canonicalizer performs learned secondary calibration on statically calibrated IMU measurements. The Motion Estimator recovers local pose, while the Root Bypass independently estimates global wrist orientation. Virtual IMUs provide canonical observations and paired supervision for the three training stages.

continuous rotation representation $\phi : \mathrm{SO}(3) \to \mathbb{R}^6$, fixed-order concatenation gives

$$\mathbf{x}_t = \left[\mathrm{Concat}_{s\in\mathcal{S}\cup\{r\}}\, \mathbf{a}^M_{t,s},\ \mathrm{Concat}_{s\in\mathcal{S}\cup\{r\}}\, \phi(\mathbf{R}^M_{t,s})\right] \in \mathbb{R}^{54}. \tag{4}$$

Measured and virtual IMUs share coordinate frames, sensor order, and channel definitions. The nine-dimensional back-of-hand observation also feeds the root pathway.

5.2.1 Module I: Secondary Calibration with the Canonicalizer

Flat-pose calibration establishes an initial alignment for the current donning session but leaves residual placement errors and motion-dependent disturbances. The Canonicalizer learns causal corrections from calibrated observations, adapting to each sensor and its motion state without re-estimating static mounting parameters.

Node features encode acceleration, rotation, and angular velocity derived from rotation history. Two gated message-passing layers exchange features on a graph linking the back-of-hand sensor to all fingertips, neighboring fingertips to each other, and the thumb to the index finger.

Placement offsets typically change slowly, while motion-induced disturbances often appear as short-term residuals. Per-sensor GRUs and gated low-pass states track slowly varying corrections, while multiscale causal temporal convolutions capture short-term disturbances. Each branch predicts rotation and acceleration residuals, with confidence weights controlling its contribution to the calibrated measurements. For $b \in \{\mathrm{slow}, \mathrm{fast}\}$, let $\Delta\mathbf{r}^b_{t,s}$ and $\Delta\mathbf{a}^b_{t,s}$ be the unweighted residuals, with confidence weights $c^{b,R}_{t,s}$ and $c^{b,a}_{t,s}$. The corrected measurements are

$$\begin{aligned}
\mathbf{R}^{\mathrm{can}}_{t,s} &= \mathbf{R}^{\mathrm{in}}_{t,s}\,\mathrm{Exp}\left(c^{\mathrm{slow},R}_{t,s}\Delta\mathbf{r}^{\mathrm{slow}}_{t,s}\right)\cdot\mathrm{Exp}\left(c^{\mathrm{fast},R}_{t,s}\Delta\mathbf{r}^{\mathrm{fast}}_{t,s}\right),\\
\mathbf{a}^{\mathrm{can}}_{t,s} &= \mathbf{a}^{\mathrm{in}}_{t,s} + c^{\mathrm{slow},a}_{t,s}\Delta\mathbf{a}^{\mathrm{slow}}_{t,s} + c^{\mathrm{fast},a}_{t,s}\Delta\mathbf{a}^{\mathrm{fast}}_{t,s}.
\end{aligned} \tag{5}$$

Here, Exp maps a three-dimensional rotation vector to $\mathrm{SO}(3)$. Zero-initialized residual heads make the Canonicalizer an identity mapping at initialization, preserving flat-pose calibration.

5.2.2 Module II: Pose Estimator

The Pose Estimator first combines sensor features within each frame, then models their temporal evolution. It separately encodes each sensor's canonicalized acceleration and rotation and fuses the features through graph attention:

$$\mathbf{e}_{t,s} = \left[E_a(\mathbf{a}^{\mathrm{can}}_{t,s}); E_R\big(\phi(\mathbf{R}^{\mathrm{can}}_{t,s})\big)\right], \qquad \mathbf{z}_t = \mathrm{GAT}_r(\mathbf{e}_{t,0},\ldots,\mathbf{e}_{t,5}), \tag{6}$$

where GAT_r aggregates sensor features on a graph centered on the back-of-hand sensor r.

Multiscale causal temporal convolutions and a three-layer unidirectional LSTM process the fused sequence. Decoder $\mathcal{D}$ maps the current hidden state to six-dimensional pose rotations:

$$\mathbf{u}_{1:t} = \mathrm{TCN}_{\mathrm{ms}}^{\mathrm{causal}}(\mathbf{z}_{1:t}), \quad \mathbf{h}_t = \mathrm{LSTM}_3(\mathbf{u}_{1:t}), \quad \widehat{\mathbf{r}}_t^{6D} = F_{\mathrm{pose}}(\mathbf{x}_{1:t}^{\mathrm{can}}) = \mathcal{D}(\mathbf{h}_t) \in \mathbb{R}^{16\times 6}. \quad (7)$$

Root Bypass. Local pose depends on relative sensor motion, whereas wrist orientation requires global rotation information from the back-of-hand sensor. To preserve this information through secondary calibration, the Root Bypass predicts primary wrist rotation from original back-of-hand observations and refines it using canonicalized observations:

$$\widehat{\mathbf{R}}_t^{\mathrm{root}} = g(\widehat{\mathbf{r}}_t^{\mathrm{pri}})\,\mathrm{Exp}(\boldsymbol{\delta}_t), \quad (8)$$

Here, g orthogonalizes the primary prediction $\widehat{\mathbf{r}}_t^{\mathrm{pri}} \in \mathbb{R}^6$ into a rotation matrix, and $\boldsymbol{\delta}_t \in \mathbb{R}^3$ is a residual rotation vector predicted from canonicalized observations.

This wrist rotation replaces the main branch's root prediction, leaving local joint rotations unchanged.

5.2.3 Module III: Three-Stage Training with Virtual IMUs

Virtual IMUs define a common canonical reference for pose pretraining, residual calibration, and joint adaptation.

Virtual IMUs. Using only training-set MANO annotations, forward kinematics provides bone rotations and sensor anchor trajectories:

$$\mathbf{R}_{t,s}^{\mathrm{sim}} = \mathbf{R}_t^g[r(s)], \qquad \mathbf{a}_{t,s}^{\mathrm{sim}} = \frac{\mathbf{p}_{t+\delta}^g[p(s)] - 2\mathbf{p}_t^g[p(s)] + \mathbf{p}_{t-\delta}^g[p(s)]}{(\delta\Delta t)^2}, \quad \delta = 4. \quad (9)$$

Here, $r(s)$ and $p(s)$ index rotation and position anchors, and Δt is the frame interval. Fingertip sensors use distal phalanx rotations and fingertip positions; the back-of-hand sensor uses the wrist. Central differences apply only offline. We augment inputs by perturbing mounting rotations and positions and adding measurement noise, retaining known canonical targets. Time-aligned MANO annotations provide virtual references for measured sequences.

Stage I: Pose Pretraining. Canonical and mildly perturbed virtual IMUs, paired with MANO poses, train the Pose Estimator's main branch and Root Bypass using $\mathcal{L}_{\mathrm{pose}}$, establishing the canonical observation-to-pose mapping.

Stage II: Residual Calibration. We freeze the Pose Estimator, including the Root Bypass, and train only the Canonicalizer with $\mathcal{L}_{\mathrm{can}}$. Measured and strongly perturbed virtual IMUs use canonical targets, while canonical inputs encourage identity correction, aligning observations without changing the pose mapping.

Stage III: Joint Adaptation. We unfreeze all modules and jointly train on measured IMUs, strongly perturbed virtual IMUs, and canonical replay:

$$\mathcal{L}_{\mathrm{joint}} = \mathcal{L}_{\mathrm{pose}} + \lambda_{\mathrm{can}}\mathcal{L}_{\mathrm{can}}. \quad (10)$$

Canonical replay helps preserve the Stage I pose mapping during adaptation to measured data.

6 Experiments and Analysis

6.1 Experimental Setup

We split sequences into training and test sets in an approximately 80%/20% ratio within each action category, keeping both hands and all modalities together. We measure global wrist and parent-relative non-wrist rotation errors as geodesic distances on $\mathrm{SO}(3)$ (degrees), and report MPJPE, MPVPE, and their Procrustes-aligned variants (millimeters). All predicted and reference geometry uses $\boldsymbol{\beta} = \mathbf{0}$, zero translation, and identity root rotation; joint errors exclude the wrist, while vertex errors cover the full mesh (Appendix C).

Our 2.53M-parameter model uses AdamW, 256-frame windows (stride/burn-in: 32/32), and batch size 16. Pose pretraining, canonicalizer training, and joint finetuning run for 5/5/10 epochs at learning rates $5 \times 10^{-4}/5 \times 10^{-4}/10^{-4}$, respectively. We use weight decay 10^{-5}, seed 42, and loss weights 0.016 (clean-identity rotation) and 0.005 (root geodesic).

6.2 Comparison with Other Methods

6.2.1 Comparison with IMU-Based Methods

We retrain adapted VIFNet-S, MGDHand, and RT-HandTrack (Wang et al., 2025; 2026; Sarker et al., 2026) with six IMUs per hand, using the same split, MANO outputs, and evaluation protocol. Only necessary input/output interfaces are modified; synchronized RGB video provides action context only.

Table 2: Comparison under the unified six-IMU protocol on the full test population (291,283 hand-frames; all IMU methods cover 100% of the population). Rotation errors are in degrees, and geometric errors are in millimeters. N/A: the method does not output global wrist orientation. Lower is better; bold indicates the best result per column.

Method	Root Ori.	Local Rot.	MPJPE	PA-MPJPE	MPVPE	PA-MPVPE
VIFNet-S (IMU-only adapted)	10.274	9.618	8.014	6.463	7.453	6.290
MGDHand (adapted)	10.114	**9.038**	7.496	6.080	6.973	5.915
RT-HandTrack (adapted)	N/A	14.149	12.045	9.414	11.132	9.365
Ours	**8.248**	9.228	**7.117**	**5.608**	**6.645**	**5.481**

Ours achieves the lowest root orientation and geometric errors (Table 2), reducing MPJPE by 5.1% over MGDHand, which retains the lowest local rotation error.

6.2.2 Comparison with Vision-Based Methods

HaMeR (Pavlakos et al., 2024) and ACE-Ego-Hand (Liu et al., 2026) use public checkpoints on left-eye RGB video. We compare local pose under the same MANO normalization, excluding global rotation because camera-to-ground-truth extrinsics are unverified.

Task types define unoccluded and occluded conditions. Occluded cases comprise hand–hand and hand–object interactions. IMU methods cover all frames. Vision errors use detected frames, with misses reflected in coverage; paired comparisons evaluate Ours on the same frames.

Table 3: Local pose errors under unoccluded and occluded conditions. Unoccluded contains 175,953 target hand-frames; occluded pools hand–hand and hand–object interaction (115,330 frames). Cov. (%) lists the fraction of target frames on which a method produces detections (unoccluded / occluded); vision methods are scored only on these frames and misses are kept as failures. Local Rot. is in degrees; all other metrics are in millimeters. Bold indicates the best result per column within each condition.

		Unoccluded					Occluded				
Method (input)	Cov. (%)	Local Rot.	MPJPE	PA-MPJPE	MPVPE	PA-MPVPE	Local Rot.	MPJPE	PA-MPJPE	MPVPE	PA-MPVPE
Ours (IMU)	100.0	**9.654**	**7.429**	**5.860**	**6.927**	**5.738**	**8.577**	**6.641**	**5.223**	**6.213**	**5.090**
HaMeR (RGB)	35.2 / 39.1	30.026	17.939	14.326	16.972	14.137	27.578	14.210	11.447	13.635	11.484
ACE-Ego-Hand (RGB)	100.0	24.622	16.625	13.003	15.483	12.424	23.743	15.709	12.011	14.764	11.692

Ours leads in both conditions (Table 3, Figure 3). Across both conditions, on HaMeR's 36.7% valid frames, MPJPE remains lower (7.144 vs. 16.368 mm). These task-defined groups do not isolate occlusion effects, and hand–object data include only two subjects.

6.3 Ablation Studies

Each ablation removes one component while retaining the split, evaluation protocol, and model selection rules. Removing the Canonicalizer causes the largest degradation (Table 4), increasing

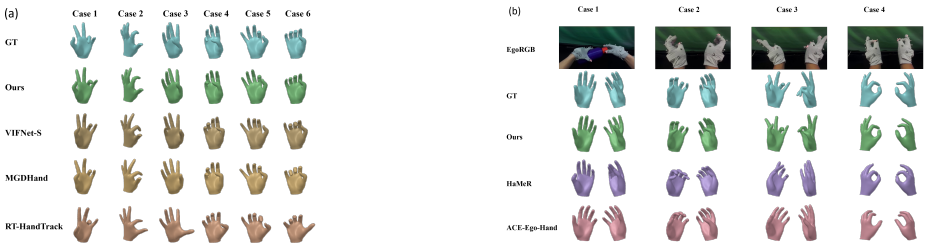


Figure 3: Visualization of comparision.Compare with (a)IMU method.(b) Visual method.

Table 4: Ablation studies of the proposed pipeline. Each variant removes the indicated component from the full pipeline. Root and Local denote rotation errors in degrees; geometric errors are in millimeters and exclude root rotation. MPJPE is averaged over 20 non-wrist joints. Lower is better.

Configuration	Root Ori.	Local Rot.	MPJPE	PA-MPJPE	MPVPE	PA-MPVPE
Full pipeline	8.248	9.228	7.117	5.608	6.645	5.481
w/o Canonicalizer	13.520	10.147	8.100	6.348	7.538	6.186
w/o Root Bypass	9.296	9.241	7.094	5.606	6.622	5.476
w/o Shared Slow Branch	8.778	9.615	7.536	5.912	7.025	5.770
w/o Shared Fast Branch	8.946	9.333	7.274	5.695	6.787	5.558

root error from 8.248° to 13.520° and MPJPE from 7.117 to 8.100 mm. Removing Root Bypass increases root error by 1.048° with negligible changes in local geometry. Removing either temporal branch worsens all six metrics, supporting their combination.

6.4 Applications

Real-time reconstruction. Our online system receives bimanual IMU streams, calibrates and aligns the measurements, and resamples them to 30 Hz. The 2.53M-parameter BIMU-Pose model performs causal inference over a 256-frame sliding window.

Dexterous Hand Teleoperation. We use the MANO hand geometry reconstructed by our BIMU-Pose baseline to teleoperate the Linker Hand O6 dexterous hand in real time. After aligning the palm coordinate frames and scales of the human and robot hands, we use the O6 kinematic model to compute six-dimensional joint targets from vectors connecting the wrist to the finger joints and fingertips.

7 Conclusion

We present IMUHand Benchmark, which provides continuous bimanual inertial sequences, synchronized egocentric RGB video, and per-frame MANO annotations under a fixed layout of six IMUs per hand. The benchmark defines a unified IMU-only reconstruction task and evaluation protocol, enabling comparisons under consistent observation and annotation conditions. We further introduce a baseline for sparse-IMU hand reconstruction. Ablation results show that observation correction contributes most strongly to reconstruction accuracy, while the root bypass further reduces wrist orientation error. We also develop a pair of IMU gloves and an accompanying system for data acquisition and real-time reconstruction.

Our system remains limited by IMU drift and glove fit. Six-axis IMUs lack an absolute heading reference, so yaw errors accumulate over time. Double integration of acceleration amplifies measurement errors, and the current system therefore does not estimate global wrist translation. Standard-sized gloves may fit loosely on smaller hands, allowing sensor slippage to compromise the initial calibration. We plan to improve drift correction and glove fastening, and explore global motion estimation with position constraints.

AI USE STATEMENT

During the preparation of this paper, we used generative AI tools to assist with language editing and rewriting, literature searches, and related-work research. These tools were not used to generate or collect experimental data, conduct experiments, fabricate results, or replace the authors' scientific judgment. The authors reviewed all AI-assisted content, checked technical claims and references, and verified the manuscript. The authors take full responsibility for the final contents of this paper, including all AI-assisted text, ideas, citations, and other outputs.

ETHICS STATEMENT

This study involved the collection of bimanual motion, wearable sensor data, and egocentric video from human participants. All participants were adults and provided written informed consent before data collection. They were informed of the recorded modalities, research purposes, intended data use, and their right to withdraw. Before release, visual recordings will be screened for faces, screens, personal belongings, and other identifying information; affected segments will be removed, obscured, or withheld. Only data covered by the informed consent will be released. The dataset is intended solely for research on hand motion reconstruction and wearable interaction, and users must comply with the applicable privacy and licensing terms.

REPRODUCIBILITY STATEMENT

Sections 3–4 describe the hardware, acquisition system, and annotation process. Sections 5 and 6 present the method, training setup, and evaluation protocol, with additional calibration, synchronization, and metric definitions in the appendix. Data are stratified by action type so that both training and test sets cover all actions. Virtual IMUs are generated only from training-set annotations, and the test set is not used for training or model selection. After completing data quality checks and release review, we plan to publish the fixed splits, calibration and annotation tools, evaluation code, baseline configurations, and model weights.

REFERENCES

Prithviraj Banerjee, Sindi Shkodrani, Pierre Moulon, Shreyas Hampali, Shangchen Han, Fan Zhang, Linguang Zhang, Jade Fountain, Edward Miller, Selen Basol, Richard Newcombe, Robert Wang, Jakob Julian Engel, and Tomas Hodan. 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*, pp. 7061–7071, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Banerjee_HOT3D_Hand_and_Object_Tracking_in_3D_from_Egocentric_Multi-View_CVPR_2025_paper.html.

ChingMu. ChingMu 1000-hour embodied motion dataset. https://chingmudata.github.io/MotionDecode/. Accessed 2026-09-26.

Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J. Black, and Otmar Hilliges. ARCTIC: A dataset for dexterous bimanual hand-object manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12943–12954, 2023.

Aaron Grapentin, Dustin Lehmann, Ardjola Zhupa, and Thomas Seel. Sparse magnetometer-free real-time inertial hand motion tracking. In *Proceedings of the IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems*, pp. 94–100, 2020. doi: 10.1109/MFI49285.2020.9235262.

Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, Eugene Byrne, Zach Chavis, Joya Chen, Feng Cheng, Fu-Jen Chu, Sean Crane, Avijit Dasgupta, Jing Dong, Maria Escobar, Cristhian Forigua, Abrham Gebreselasie, Sanjay Haresh, Jing Huang, Md Mohaiminul Islam, Suyog Jain, Rawal Khirodkar, Devansh Kukreja, Kevin J Liang, Jia-Wei Liu, Sagnik Majumder, Yongsen Mao, Miguel Martin, Effrosyni Mavroudi, Tushar Nagarajan, Francesco Ragusa, Santhosh Kumar Ramakrishnan, Luigi Seminara, Arjun Somayazulu, Yale Song, Shan Su,

Zihui Xue, Edward Zhang, Jinxu Zhang, Angela Castillo, Changan Chen, Xinzhu Fu, Ryosuke Furuta, Cristina Gonzalez, Prince Gupta, Jiabo Hu, Yifei Huang, Yiming Huang, Weslie Khoo, Anush Kumar, Robert Kuo, Sach Lakhavani, Miao Liu, Mi Luo, Zhengyi Luo, Brighid Meredith, Austin Miller, Oluwatumininu Oguntola, Xiaqing Pan, Penny Peng, Shraman Pramanick, Merey Ramazanova, Fiona Ryan, Wei Shan, Kiran Somasundaram, Chenan Song, Audrey Southerland, Masatoshi Tateno, Huiyu Wang, Yuchen Wang, Takuma Yagi, Mingfei Yan, Xitong Yang, Zecheng Yu, Shengxin Cindy Zha, Chen Zhao, Ziwei Zhao, Zhifan Zhu, Jeff Zhuo, Pablo Arbelaez, Gedas Bertasius, Dima Damen, Jakob Engel, Giovanni Maria Farinella, Antonino Furnari, Bernard Ghanem, Judy Hoffman, C.V. Jawahar, Richard Newcombe, Hyun Soo Park, James M. Rehg, Yoichi Sato, Manolis Savva, Jianbo Shi, Mike Zheng Shou, and Michael Wray. Ego-Exo4D: Understanding skilled human activity from first- and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19383–19400, 2024.

Yinghao Huang, Manuel Kaufmann, Emre Aksan, Michael J. Black, Otmar Hilliges, and Gerard Pons-Moll. Deep inertial poser: Learning to reconstruct human pose from sparse inertial measurements in real time. *ACM Transactions on Graphics*, 37(6):185:1–185:15, 2018. doi: 10.1145/3272127.3275108.

Ziyi Kou, Ankit Kumar, Mia Huang, Taylor Niehues, Vatsal Mehta, Ergys Ristani, and Li Guan. AVI-HT: Adaptive vision-IMU fusion for 3d hand tracking. *arXiv preprint arXiv:2605.21714*, 2026. URL https://arxiv.org/abs/2605.21714.

Franziska Krebs, Andre Meixner, Isabel Patzer, and Tamim Asfour. The KIT bimanual manipulation dataset. In *2020 IEEE-RAS 20th International Conference on Humanoid Robots (Humanoids)*, pp. 499–506, 2021. doi: 10.1109/HUMANOIDS47582.2021.9555788.

Taein Kwon, Bugra Tekin, Jan Stuehmer, Federica Bogo, and Marc Pollefeys. H2O: Two hands manipulating objects for first person interaction recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10138–10148, 2021.

Yutong Li, Jieyi Zhang, Wenqiang Xu, Tutian Tang, and Cewu Lu. FSGlove: An inertial-based hand tracking system with shape-aware calibration. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 18304–18311, 2025. doi: 10.1109/IROS60139.2025.11245803. URL https://arxiv.org/abs/2509.21242.

Yufei Liu, Xixi Wang, Hao Li, Ganlong Zhao, Kaitong Cai, Chengkai Jin, Chunxiao Liu, Jianbo Liu, Siyuan Huang, Xingang Pan, and Hongsheng Li. ACE-Ego-Hand: Repurposing video diffusion models for occlusion-robust egocentric 3d hand motion recovery. *arXiv preprint arXiv:2608.20308*, 2026. URL https://arxiv.org/abs/2608.20308.

David Mills, Jim Martin, Jack Burbank, and William Kasch. Network time protocol version 4: Protocol and algorithms specification. RFC 5905, Internet Engineering Task Force, 2010.

Francesca Mongelli and Matteo Menolotto. HANDMI4. IEEE DataPort, 2023. URL https://doi.org/10.21227/c6t8-ge47.

Takehiko Ohkawa, Kun He, Fadime Sener, Tomas Hodan, Luan Tran, and Cem Keskin. AssemblyHands: Towards egocentric activity understanding via 3d hand pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12999–13008, 2023.

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

Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics*, 36(6):245:1–245:17, 2017. doi: 10.1145/3130800.3130883.

Anik Sarker, Ziyi Kou, Ergys Ristani, Li Guan, and Taylor Niehues. Real-time hand pose tracking using 6-axis IMUs. In *Proceedings of the 21st ACM/IEEE International Conference on Human-Robot Interaction*, pp. 1182–1191, 2026. doi: 10.1145/3757279.3785628.

Stereolabs. Products overview. https://www.stereolabs.com/docs/products. Accessed 2026-09-26.

Russell M. Taylor, II, Thomas C. Hudson, Adam Seeger, Hans Weber, Jeffrey Juliano, and Aron T. Helser. VRPN: A device-independent, network-transparent VR peripheral system. In *Proceedings of the ACM Symposium on Virtual Reality Software and Technology*, pp. 55–61, 2001. doi: 10.1145/505008.505019.

Xinyi Wang, Pengfei Ren, Haoyang Zhang, Xin Sheng, Da Li, Liang Xie, Yue Gao, and Erwei Yin. VIHand: Enhancing 3d hand pose estimation with visual-inertial benchmark. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 12753–12760, 2025. doi: 10.1145/3746027.3758215. URL https://doi.org/10.1145/3746027.3758215.

Xinyi Wang, Pengfei Ren, Haoyang Zhang, Hanling Zhan, Yingxi Li, Liang Xie, Yue Gao, and Erwei Yin. MGDHand: Multi-granularity prior-to-inertial distillation framework for sequential 3d hand pose estimation from sparse IMUs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13996–14005, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Wang_MGDHand_Multi-Granularity_Prior-to-Inertial_Distillation_Framework_for_Sequential_3D_Hand_Pose_CVPR_2026_paper.html.

Ziheng Xi, Jiayi Yu, Yitao Wang, Yanbo Duan, Jianjiang Feng, and Jie Zhou. EgoEMG: A multimodal egocentric dataset with bilateral EMG and vision for hand pose estimation. *arXiv preprint arXiv:2605.05712*, 2026. URL https://arxiv.org/abs/2605.05712.

Xinyu Yi, Yuxiao Zhou, and Feng Xu. TransPose: Real-time 3d human translation and pose estimation with six inertial sensors. *ACM Transactions on Graphics*, 40(4):86:1–86:13, 2021.

Xinyu Yi, Yuxiao Zhou, Marc Habermann, Soshi Shimada, Vladislav Golyanik, Christian Theobalt, and Feng Xu. Physical inertial poser (PIP): Physics-aware real-time human motion tracking from sparse inertial sensors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13167–13178, 2022.

Xinyu Zhan, Lixin Yang, Yifei Zhao, Kangrui Mao, Hanlin Xu, Zenan Lin, Kailin Li, and Cewu Lu. OAKINK2: A dataset of bimanual hands-object manipulation in complex task completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 445–456, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Zhan_OAKINK2_A_Dataset_of_Bimanual_Hands-Object_Manipulation_in_Complex_Task_CVPR_2024_paper.html.

Jinglei Zhang, Jiankang Deng, Chao Ma, and Rolandos Alexandros Potamias. HaWoR: World-space hand motion reconstruction from egocentric videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1805–1815, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Zhang_HaWoR_World-Space_Hand_Motion_Reconstruction_from_Egocentric_Videos_CVPR_2025_paper.html.

Xinyu Zhang, Ziyi Kou, Chuan Qin, Mia Huang, Ergys Ristani, Ankit Kumar, Lele Chen, Kun He, Abdeslam Boularias, and Li Guan. Glove2Hand: Synthesizing natural hand-object interaction from multi-modal sensing gloves. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1829–1840, 2026. URL https://arxiv.org/abs/2603.20850.

Yu Zhang, Songpengcheng Xia, Lei Chu, Jiarui Yang, Qi Wu, and Ling Pei. Dynamic inertial poser (DynaIP): Part-based motion dynamics learning for enhanced human pose estimation with sparse inertial sensors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1889–1899, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Zhang_Dynamic_Inertial_Poser_DynaIP_Part-Based_Motion_Dynamics_Learning_for_Enhanced_CVPR_2024_paper.html.

Zhengyou Zhang. A flexible new technique for camera calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(11):1330–1334, 2000. doi: 10.1109/34.888718.

A CALIBRATION AND SYNCHRONIZATION

A.1 IMU CALIBRATION

After each donning, participants hold both hands flat in a reference pose for approximately 10 s while orientation and acceleration are recorded from the six IMUs on each hand. Calibration aligns the inertial and MANO reference frames, estimates sensor-to-bone mounting rotations, and corrects static acceleration offsets. The calibrated observations are then expressed relative to the back of the hand.

Coordinate frames and rotation conventions. Let G denote the global reference frame used for IMU orientation measurements, M the MANO reference frame, and I_s and B_s the frames of sensor s and its corresponding bone. The back-of-hand sensor is indexed by $s = 0$, and the five fingertip sensors by $s = 1, \ldots, 5$. All rotations use the column-vector convention: ${}^{A}\mathbf{R}_{B}$ transforms vectors from frame B to frame A. The calibrated bone orientation is

$$ {}^{M}\mathbf{R}_{B_s,t} = {}^{M}\mathbf{R}_{G}\,{}^{G}\mathbf{R}_{I_s,t}\,{}^{I_s}\mathbf{R}_{B_s}, \tag{11} $$

where ${}^{G}\mathbf{R}_{I_s,t}$ is the measured orientation, ${}^{M}\mathbf{R}_{G}$ is the rotation aligning the reference frames, and ${}^{I_s}\mathbf{R}_{B_s}$ is the mounting rotation relating the bone and sensor frames.

Reference-frame alignment and mounting rotations. Let $\overline{\mathbf{R}}_s \in \mathrm{SO}(3)$ denote the mean orientation of sensor s over the calibration window, computed in rotation space. Using the predefined mounting rotation of the back-of-hand IMU, $\mathbf{C}_0 = {}^{I_0}\mathbf{R}_{B_0}$ and the reference orientation of the root bone in the MANO flat-hand pose, $\mathbf{F}_0 = {}^{M}\mathbf{R}_{B_0,\mathrm{flat}}$, we obtain the reference-frame alignment rotation:

$$ {}^{M}\mathbf{R}_{G} = \mathbf{F}_0 \left(\overline{\mathbf{R}}_0 \mathbf{C}_0\right)^{-1}. \tag{12} $$

For a fingertip sensor, let $\mathbf{F}_s = {}^{M}\mathbf{R}_{B_s,\mathrm{flat}}$ denote the reference orientation of its distal phalanx in the MANO flat-hand pose. The mounting rotation is estimated as

$$ {}^{I_s}\mathbf{R}_{B_s} = \left({}^{M}\mathbf{R}_{G}\overline{\mathbf{R}}_s\right)^{-1}\mathbf{F}_s, \qquad s = 1, \ldots, 5. \tag{13} $$

Under the flat-hand approximation $\mathbf{F}_s \approx \mathbf{I}$, this reduces to ${}^{I_s}\mathbf{R}_{B_s} \approx ({}^{M}\mathbf{R}_{G}\overline{\mathbf{R}}_s)^{-1}$. These calibration parameters remain fixed throughout the current donning session and are applied through Equation equation 11 to transform orientations in subsequent motion frames. A fixed mounting rotation is a calibration approximation; it cannot compensate for time-varying offsets caused by glove deformation or sensor slippage during motion.

Acceleration correction. We transform raw acceleration, including gravity, into the global frame G. The mean over the static calibration window estimates the combined contribution of gravity and static offsets. We subtract this estimate from subsequent observations and transform the result into the MANO frame. For a calibration window containing N frames,

$$ \begin{aligned} {}^{G}\mathbf{o}_s &= \frac{1}{N}\sum_{n=1}^{N} {}^{G}\mathbf{a}^{\mathrm{raw}}_{s,n}, \\ {}^{M}\mathbf{a}_{s,t} &= {}^{M}\mathbf{R}_{G}\left({}^{G}\mathbf{a}^{\mathrm{raw}}_{s,t} - {}^{G}\mathbf{o}_s\right). \end{aligned} \tag{14} $$

where ${}^{G}\mathbf{a}^{\mathrm{raw}}_{s,t}$ is the raw acceleration of sensor s expressed in the global frame, and ${}^{G}\mathbf{o}_s$ is the fixed correction used for the current donning session.

Root-relative normalization. After calibration, we use the current back-of-hand bone frame B_0 as the root reference frame. Fingertip orientations are converted to rotations relative to the back of the hand, and accelerations are expressed in the same root frame:

$$ \begin{aligned} {}^{B_0}\mathbf{R}_{B_s,t} &= \left({}^{M}\mathbf{R}_{B_0,t}\right)^{-1}{}^{M}\mathbf{R}_{B_s,t}, \quad s = 1, \ldots, 5, \\ {}^{B_0}\mathbf{a}_{s,t} &= \left({}^{M}\mathbf{R}_{B_0,t}\right)^{-1}{}^{M}\mathbf{a}_{s,t}, \qquad s = 0, \ldots, 5. \end{aligned} \tag{15} $$

A.2 ZED CALIBRATION

Let W denote the ChingMu world frame, B the frame of the rigid marker body attached to the ZED camera, and L and R the left- and right-camera frames. We use the convention that ${}^{A}\mathbf{T}_{B}$ maps coordinates from frame B to frame A. Checkerboard observations provide the fixed transformation ${}^{B}\mathbf{T}_{L}$ between the left camera and the tracked marker body. Given the marker-body pose ${}^{W}\mathbf{T}_{B,t}$ measured by ChingMu at time t, the left-camera pose in the motion-capture world is

$$ {}^{W}\mathbf{T}_{L,t} = {}^{W}\mathbf{T}_{B,t}\,{}^{B}\mathbf{T}_{L}. \tag{16} $$

The ZED stereo calibration provides the fixed extrinsic transformation ${}^{L}\mathbf{T}_{R}$, so the right-camera pose is

$$ {}^{W}\mathbf{T}_{R,t} = {}^{W}\mathbf{T}_{L,t}\,{}^{L}\mathbf{T}_{R}. \tag{17} $$

Thus, only the left camera needs to be registered directly to the marker body; the right-camera pose follows from the calibrated stereo geometry.

A.3 TEMPORAL SYNCHRONIZATION

All streams are received by a common acquisition host, which retains both device and host timing information. NTP provides the initial clock alignment between the gloves and the host. Residual clock offset and rate differences are corrected offline. For stream m, let τ_m denote its device timestamp. We map it to the common time base using

$$ \widehat{t}_m(\tau_m) = a_m\tau_m + b_m, \tag{18} $$

where a_m and b_m are fitted from the timestamp observations retained during acquisition and remain fixed within a recording. We define the timestamp of the first motion-capture frame as the sequence origin, apply the fitted mappings to all modalities, and resample them on a 30 Hz timeline. Frames without valid aligned observations are excluded from subsequent training and evaluation.

B GROUND-TRUTH ANNOTATION DETAILS

This section provides the fitting details for the MANO annotations described in Section 4.3, using the parameterization in Equation equation 1.

Skeleton preprocessing. ChingMu and MANO use different skeletal topologies, rest poses, and joint-axis definitions. We therefore process and fit each hand side independently. Short tracking gaps are filled, anomalous rotational or positional jumps are corrected, and the skeleton is reconstructed using the fixed bone lengths of the current session's template. Samples that cannot be recovered reliably remain invalid and are not used by the subsequent fitting stage.

Shape estimation. For each skeletal template associated with a donning session, we fit a separate $\boldsymbol{\beta}$ for each hand and keep it fixed across frames. Shape fitting first matches five wrist-to-fingertip path lengths and five path lengths from the metacarpophalangeal (MCP) joints to the fingertips. We then incorporate palm width and distances between adjacent MCP joints to jointly constrain finger lengths and palm proportions. A hand-shape prior discourages implausible shapes.

Pose estimation. We first obtain an analytical MANO pose initialization by mapping ChingMu joint rotations according to the rest-pose and joint-axis relationships between the two skeletons. To compensate for the remaining geometric differences, we introduce a correction rotation $\mathbf{Q}_{t,j}$ for each local joint:

$$ \mathbf{R}^{M}_{t,j} = \mathbf{R}^{M,0}_{t,j}\mathbf{Q}_{t,j}, \tag{19} $$

and optimize these corrections using the following objective:

$$ \begin{aligned} \mathcal{L}_t = {} & \lambda_{\text{tip}} \sum_{f\in\mathcal{F}} \left\| \bar{\mathbf{p}}^{M}_{t,f} - \bar{\mathbf{p}}^{S}_{t,f} \right\|_2^2 + \lambda_{\text{int}} \sum_{e\in\mathcal{E}_{\text{int}}} \left\| \mathbf{d}^{M}_{t,e} - \mathbf{d}^{S}_{t,e} \right\|_2^2 + \lambda_{\text{dist}} \sum_{e\in\mathcal{E}_{\text{dist}}} \left\| \mathbf{d}^{M}_{t,e} - \mathbf{d}^{S}_{t,e} \right\|_2^2 \\ & + \lambda_{\text{rot}} \sum_{j} \left\| \log(\mathbf{Q}_{t,j}) \right\|_2^2 + \lambda_{\text{temp}} \sum_{j} \left\| \log\left(\mathbf{Q}_{t-1,j}^{\top}\mathbf{Q}_{t,j}\right) \right\|_2^2 . \end{aligned} \tag{20} $$

Here, $\bar{\mathbf{p}}$ denotes wrist-relative fingertip position, $\mathbf{d}$ denotes a unit bone direction, and $\mathcal{F}$ is the set of five fingertips. The sets $\mathcal{E}_{\text{int}}$ and $\mathcal{E}_{\text{dist}}$ contain internal and distal finger segments, respectively. The first three terms align MANO fingertip positions and finger directions with those of the ChingMu skeleton. The rotation prior limits deviations from the analytical initialization, while the temporal term encourages continuity of the corrections between adjacent frames.

Sequence construction and quality control. The fitted annotations retain their original frame identifiers and timestamps so that they remain aligned with the IMU and camera streams. Automated validity checks and manual review jointly control annotation quality, and the resulting validity masks exclude rejected frames during dataset construction.

C Evaluation Metrics

We evaluate rotation accuracy and geometric reconstruction accuracy. The following definitions are given for a single hand, where t indexes frames, T is the number of valid evaluation frames, and predictions are denoted by $\widehat{\cdot}$.

Rotation errors. For any two rotation matrices $\mathbf{R}_1, \mathbf{R}_2 \in \mathrm{SO}(3)$, we measure their difference using the geodesic distance in degrees:

$$d_{\mathrm{SO}(3)}(\mathbf{R}_1, \mathbf{R}_2) = \frac{180}{\pi} \arccos\left[\mathrm{clip}\left(\frac{\mathrm{tr}(\mathbf{R}_1^\top \mathbf{R}_2) - 1}{2}, -1, 1\right)\right], \tag{21}$$

Clipping prevents floating-point errors from placing the argument outside the domain of the inverse cosine.

Mean Root Orientation Error and Mean Local Joint Rotation Error are defined as

$$E_{\text{root}} = \frac{1}{T} \sum_{t=1}^{T} d_{\mathrm{SO}(3)}\left(\widehat{\mathbf{R}}_t^{\text{root}}, \mathbf{R}_t^{\text{root}}\right), \tag{22}$$

$$E_{\text{local}} = \frac{1}{TK} \sum_{t=1}^{T} \sum_{k=1}^{K} d_{\mathrm{SO}(3)}\left(\widehat{\mathbf{R}}_{t,k}^{\text{local}}, \mathbf{R}_{t,k}^{\text{local}}\right), \qquad K = 15. \tag{23}$$

Local rotations are defined relative to each joint's parent; global wrist rotation is excluded from E_{local}.

Geometric errors. To evaluate local pose reconstruction independently, predictions and ground truth use the following MANO settings:

$$\boldsymbol{\beta} = \mathbf{0}, \qquad \mathbf{R}^{\text{root}} = \mathbf{I}, \qquad \mathbf{p}^{\text{root}} = \mathbf{0}, \tag{24}$$

where $\boldsymbol{\beta}$ denotes hand-shape parameters and $\mathbf{p}^{\text{root}}$ denotes global wrist translation. Under these settings, 3D joints and meshes are generated from predicted and ground-truth local rotations, respectively.

Let $\mathcal{J}$ be the set of non-wrist keypoints, with $|\mathcal{J}| = 20$, and let N_v be the number of MANO mesh vertices. Mean per-joint position error and mean per-vertex position error are defined as

$$\text{MPJPE} = \frac{1}{T|\mathcal{J}|} \sum_{t=1}^{T} \sum_{j \in \mathcal{J}} \left\| \widehat{\mathbf{J}}_{t,j} - \mathbf{J}_{t,j} \right\|_2, \tag{25}$$

$$\text{MPVPE} = \frac{1}{TN_v} \sum_{t=1}^{T} \sum_{v=1}^{N_v} \left\| \widehat{\mathbf{V}}_{t,v} - \mathbf{V}_{t,v} \right\|_2. \tag{26}$$

The geometric keypoint set includes fingertips and therefore differs from the MANO joint set used for the local rotation metric.

Procrustes-aligned errors. For the geometry defined above, we estimate a similarity transformation in each frame using corresponding non-wrist keypoints:

$$(\lambda_t^\star, \mathbf{Q}_t^\star, \mathbf{c}_t^\star) = \underset{\substack{\lambda>0,\ \mathbf{Q}\in\mathrm{SO}(3)\\ \mathbf{c}\in\mathbb{R}^3}}{\arg\min} \sum_{j\in\mathcal{J}} \left\| \lambda\mathbf{Q}\widehat{\mathbf{J}}_{t,j} + \mathbf{c} - \mathbf{J}_{t,j} \right\|_2^2, \tag{27}$$

Here, λ, $\mathbf{Q}$, and $\mathbf{c}$ denote uniform scale, rotation, and translation, respectively. Writing the optimal transformation as $\mathcal{T}_t^\star(\mathbf{x}) = \lambda_t^\star\mathbf{Q}_t^\star\mathbf{x} + \mathbf{c}_t^\star$, we obtain

$$\begin{aligned} \text{PA-MPJPE} &= \frac{1}{T|\mathcal{J}|} \sum_{t=1}^{T} \sum_{j\in\mathcal{J}} \left\| \mathcal{T}_t^\star(\widehat{\mathbf{J}}_{t,j}) - \mathbf{J}_{t,j} \right\|_2, \\ \text{PA-MPVPE} &= \frac{1}{TN_v} \sum_{t=1}^{T} \sum_{v=1}^{N_v} \left\| \mathcal{T}_t^\star(\widehat{\mathbf{V}}_{t,v}) - \mathbf{V}_{t,v} \right\|_2. \end{aligned} \tag{28}$$

Both PA metrics use the transformation estimated from the same set of joints; no separate alignment is fitted to the mesh.

Rotation errors are reported in degrees and geometric errors in millimeters; lower is better for all metrics. MPJPE and MPVPE measure local pose geometry under a common hand shape and root pose. The PA metrics additionally remove discrepancies between predicted and ground-truth geometry that can be explained by a global similarity transformation. Global wrist orientation is evaluated separately by E_{root}, and global translation is outside the scope of this task. Errors are computed separately for the left and right hands and then averaged over all valid frames in the test set.

D Hardware Implementation Details

This appendix provides implementation details of the data gloves described in Section 3.

Electronics and embedded architecture. The controller uses an ESP32-D0WD-V3 on a 33 mm $\times$ 30 mm main board. Each BNO085 module has a 15 mm $\times$ 10 mm PCB. All six IMUs connect to the controller over I2C, with an I2C multiplexer isolating sensors that share an address to enable sequential channel polling. The ESP32 handles sensor configuration, report reading, timestamp recording, data frame assembly, and transmission.

Data output and power. Each BNO085 outputs 3D acceleration and a Rotation Vector quaternion produced by internal sensor fusion. The system outputs frames containing measurements from all six IMUs at approximately 50 Hz. The firmware records the local monotonic time when each report is read and uses the reported internal latency to estimate the corresponding measurement time. A lithium battery rated at 3.7 V and 1000 mAh powers the glove, providing approximately 1 h of operation in the current prototype. External power over USB also supports extended acquisition and debugging.

Wireless transmission and timing. The glove connects to a local network over 2.4 GHz Wi-Fi and sends UDP packets to the acquisition host. Each packet contains a device ID, a Unix timestamp, the battery voltage, and the measurement time, 3D acceleration, and quaternion for each of the six IMUs. Device IDs distinguish the left and right gloves, while per-sensor timestamps preserve the sampling time differences introduced by sequential polling.

The device uses NTP on the local network approximately every 30 s to map local monotonic time to Unix time, providing initial clock alignment across devices. During dataset construction, we further estimate mappings from individual data streams to a common timeline to compensate for residual time offsets and clock-rate differences.