

Vision-Aided IMU Calibrator: Robust On-body IMU Calibration from a Single Video

DOMINIK HOLLIDT, Department of Computer Science, ETH Zürich, Switzerland

CHRISTIAN HOLZ, Department of Computer Science, ETH Zürich, Switzerland

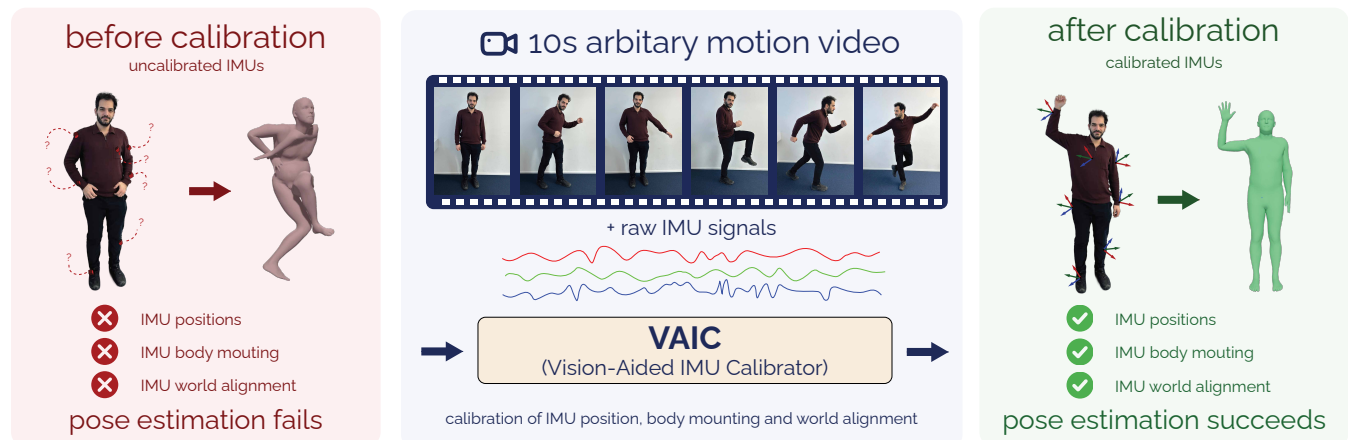


Fig. 1. Our method VAIC recovers the calibration parameters required for inertial motion capture from a short monocular video and raw IMU measurements. Without calibration (left), the sensor positions, body mounting orientations ν_{RS} , and world alignments wR_{REF} are unknown, and body-pose estimation fails. VAIC (center) jointly estimates all three quantities from 10 s of arbitrary motion footage and thus enables accurate body-pose estimation (right) without requiring manual calibration poses, manual placement specification, or magnetometer-based heading.

Inertial motion capture depends on expressing each IMU's orientations and accelerations in a shared world frame and knowing how each sensor is rotated relative to the body segment it is attached to. However, in practice, this requires careful placement and manual calibration routines (T-pose) that do not scale to arbitrary sensor sets or in-the-wild use. We present VAIC, a visual-inertial calibration method that uses a few seconds of monocular video of arbitrary user motion plus uncalibrated IMU streams to jointly recover (1) each sensor's body location (as a vertex on an SMPL mesh), (2) its sensor-to-body mounting rotation, and (3) the global reference alignment to the world frame. Our method operates without predefined poses, fixed sensor configurations, or pre-calibrated hardware. Instead, we treat calibration as an optimization problem: from video-based SMPL recovery, VAIC searches for the sensor-to-vertex alignment that best explains the measured IMU signals given these vertex kinematics. Evaluated on TotalCapture (13 IMUs), VAIC localizes sensors with 95.9% correct bone assignment and 5.8 cm positional RMSE (97.9% and 5.1 cm when excluding low-excitation motions), and achieves up to 15.5° end-to-end orientation error, which outperforms all classical calibration methods such as T-Pose or walking-based calibration. When sensor positions are known, VAIC improves over T-pose calibration, reduces body mounting error by 5.3°, and lowers overall body-orientation

error. These results suggest a practical path toward plug-and-play inertial motion capture, providing reliable cold-start calibration that can also initialize or complement downstream continuous tracking systems. The code is released at: <https://siplab.org/projects/vaic>

CCS Concepts: • **Computing methodologies** → Modeling and simulation; **Motion processing**; • **Human-centered computing** → Ubiquitous and mobile computing.

Additional Key Words and Phrases: Sparse inertial sensing, full-body pose estimation, calibration.

ACM Reference Format:

Dominik Hollidt and Christian Holz. 2026. Vision-Aided IMU Calibrator: Robust On-body IMU Calibration from a Single Video. In *SIGGRAPH Asia 2026 Conference Papers (SA Conference Papers '26)*, December 01–04, 2026, Kuala Lumpur, Malaysia. ACM, New York, NY, USA, 16 pages. <https://doi.org/10.1145/3829340.3842338>

1 Introduction

Motion capture using sparse body-worn inertial measurement units (IMUs) promises accessible, everyday motion recording without the infrastructure of camera-based systems [Hollidt et al. 2025; Huang et al. 2018; Jiang et al. 2022b; Yi et al. 2022, 2021]. Recent inertial posers [Jiang et al. 2022b; Van Wouwe et al. 2024; Yi et al. 2025, 2022, 2024] show accurate motion tracking from just six sensors.

However, despite this progress, inertial motion capture remains largely confined to controlled settings. The fundamental obstacle is calibration. Before any inertial pose estimator can operate, raw sensor measurements must be transformed into body-part orientations that are expressed in a shared world frame. This transformation



This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

SA Conference Papers '26, Kuala Lumpur, Malaysia

© 2026 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-2842-6/2026/12

<https://doi.org/10.1145/3829340.3842338>

involves two quantities that cannot be observed directly: the alignment ${}_WR_{REF}$ between the sensor’s internal reference frame (Fig. 2a) and the world frame, and the orientation offset ${}_SR_V$ between the sensor and the body part it is attached to (Fig. 2b). If these parameters are inaccurate, even perfect IMU measurements produce erroneous body poses.

Standard calibration protocols resolve these unknowns through magnetometer readings and user-performed poses, but both approaches are fragile in practice. For the reference alignment ${}_WR_{REF}$, most IMU systems attempt to align their internal reference frame with an East-North-Up (ENU) coordinate system by estimating gravity and magnetic North. While gravity provides a reliable vertical reference, indoor environments suffer from magnetic anomalies caused by structural steel and electronic equipment that render magnetometer-based heading untrustworthy [Harle 2013; Ouyang and Abed-Meraim 2022]. Existing workarounds include physically aligning all IMUs on a shared docking station [Zuo et al. 2025] or performing carefully prescribed calibration motions such as the calibration walk in commercial Xsens systems [Xsens 2024]. For the body mounting ${}_SR_V$, users hold a known calibration pose such as a T-pose [Bonfiglio et al. 2025], but even minor execution errors introduce permanent offsets that propagate through the entire capture session (Fig. 2c). Commercial systems further require the exact mounting location of every sensor to be specified *a priori*, manual body measurements for skeleton scaling, or dedicated motion capture suits with fixed sensor placements [rok 2026]. This manual overhead limits true in-the-wild usability.

In this paper, we propose VAIC, a short and unobtrusive calibration method in place of T-pose. Our approach leverages an exocentric monocular video to provide the absolute spatial reference that inertial sensing lacks. We build on recent advances in human mesh recovery that extract accurate 3D body poses from ordinary video [Wang et al. 2025] to obtain a sequence of SMPL meshes in a gravity-aligned world frame, including the user’s body shape. From this prior, we extract the orientation and acceleration of every vertex on the body surface. Our approach requires no calibration poses, manual placement specification, or magnetic heading.

We formulate calibration as a global assignment problem over sensor-vertex pairs. For each of the $K \times N$ pairings, VAIC optimizes for the mounting orientation ${}_SR_V$ that best explains the measured IMU signals given the vertex kinematics. Matching orientations directly is impossible, as this would require knowing the absolute reference ${}_WR_{REF}$. By reformulating the optimization to use relative rotations, we remove the unknown world-reference transformation from the problem. We therefore minimize three error terms: an orientation term E_{ori} that enforces consistency between relative rotations, an acceleration term E_{acc} that provides absolute reference through gravity, and a normal term E_{norm} that encodes the physical constraint that sensors are mounted flush against the skin. This normal prior is central to our method. By introducing a surface tangent frame T aligned with the mesh normal, we reduce the unknown ${}_SR_V$ from three rotational degrees of freedom to a single yaw angle around the surface normal (Fig. 2d). Without this constraint, rotation-only matching cannot distinguish symmetric placements such as a sensor on the top versus bottom of the wrist. Unlike prior work that assumes sensors are mounted on bones of

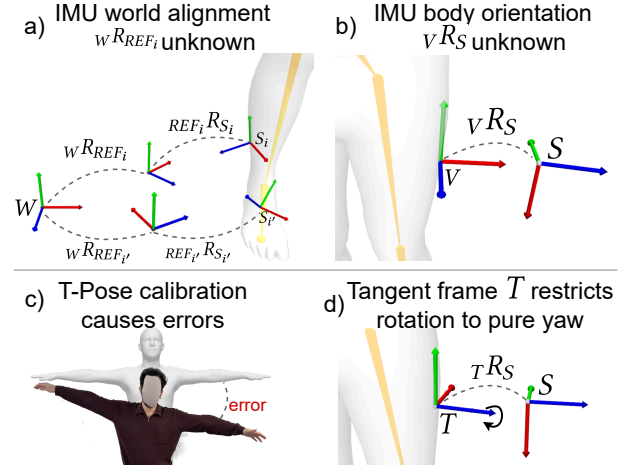


Fig. 2. The IMU world alignment ${}_WR_{REF}$ and the IMU body mounting ${}_SR_V$ are unknown. Static T-Pose calibration causes calibration errors and cannot recover ${}_WR_{REF}$. VAIC introduces the surface tangent frame T that constrains the sensor orientation to pure yaw to improve calibration.

the SMPL skeleton [Armani et al. 2024; Yi et al. 2022, 2021; Zuo et al. 2025], we model IMUs as attached to mesh vertices, which captures how surface orientation varies along a limb due to soft tissue deformation and forearm twist. Given the residual error from each sensor-vertex pair, we solve the global assignment and recover the absolute world alignment ${}_WR_{REF}$ for each sensor.

Taken together, VAIC is the first method to jointly and automatically estimate each IMU’s body location, local mounting orientation, and global reference alignment from raw inertial data and a single monocular video. Our contributions are:

- A video-aided calibration framework that eliminates the need for calibration poses, precise sensor placement, and magnetometer-based heading. VAIC is the first method to jointly localize and calibrate body-worn IMUs while also recovering SMPL shape parameters without manual user measurements, enabling plug-and-play inertial motion capture.
- A surface-normal geometric prior that constrains sensor mounting optimization to a single yaw degree of freedom and resolves symmetric ambiguities inherent to rotation-only matching.
- Experimental validation showing that VAIC localizes sensors to the correct body segment with 95.9% accuracy and achieves lower orientation error than T-pose calibration while remaining immune to misaligned reference frames due to magnetic interference.

2 Related Work

2.1 Inertial Motion Capture

Motion capture using sparse inertial sensors has become a pivotal area of research due to its immunity to occlusion and unbounded capture volume compared to optical systems [Trumble et al. 2017]. Early approaches formulated pose estimation as an optimization problem. The Sparse Inertial Poser (SIP) [Von Marcard et al. 2017] optimized SMPL parameters to fit IMU orientations and accelerations, utilizing a digitized human mesh to constrain the solution space.

While accurate, optimization methods are computationally expensive and prone to local minima. The advent of large-scale motion datasets like AMASS [Mahmood et al. 2019] and DIP-IMU [Huang et al. 2018] shifted the paradigm toward data-driven learning. Deep Inertial Poser (DIP) [Huang et al. 2018] introduced a bi-directional RNN to map sparse inertial signals directly to body pose, achieving real-time performance. TransPose [Yi et al. 2021] utilized a multi-stage network to estimate global translation and joint positions, while Transformer Inertial Poser (TIP) [Jiang et al. 2022b] leveraged the attention mechanism to model temporal dependencies and simultaneously reconstruct terrain. Other works focused on integrating physical constraints to reduce physical implausibilities. Physical Inertial Poser (PIP) [Yi et al. 2022] and Physical Non-Inertial Poser (PNP) [Yi et al. 2024] integrated neural kinematics with physics-based optimization to enforce contact constraints and ground reaction forces. GlobalPose [Yi et al. 2025] improved by losing the flat world assumption. UIP [Armani et al. 2024], UDP [Hollidt et al. 2026], and GIP [Xue et al. 2025] introduce UWB to constrain sensor drift and improve joint estimation. Umotion [Liu et al. 2025] actively modeled sensor uncertainty with a Kalman filter. However, a fundamental limitation of these learning-based systems is their reliance on fixed sensor placement and idealized calibration. While methods like DiffusionPoser [Van Wouwe et al. 2024], MobilePoser [Xu et al. 2024], and IMUCoCo [Zhou et al. 2025] have begun to address flexible placement using commodity devices, they typically require the user to specify the sensor location or rely on heuristics that may not generalize to arbitrary mounting points on the body surface.

2.2 IMU Calibration

Calibration is the prerequisite process for all inertial posers so far. This consists of finding the world-reference alignment and the IMU-body placement. Many works require performing a known "calibration pose" (e.g., T-Pose or N-Pose) [Armani et al. 2024; Bonfiglio et al. 2025; Choe et al. 2019; Trumble et al. 2017; Xue et al. 2025] or specific functional movements (e.g., hip rotation) to align the sensor frames with the body segments [Favre et al. 2009; Nazarahari et al. 2019; Vitali and Perkins 2020]. These manual procedures are fragile; imperfect user execution leads to permanent offsets in the kinematic chain, degrading pose accuracy for the entire session. Other works focus on calibrating IMUs for certain parts of the body, requiring the user to execute specific motions while modeling joint rotation-axis restrictions [Bonfiglio et al. 2025; McInne et al. 2025]. This often involves manually modeling the feasible joint angles or underlying skeletal restrictions [Cutti et al. 2008; McInne et al. 2025; Vitali and Perkins 2020]. Clinical work often relies on expensive IMUs to reduce errors [Kim and Golnaraghi 2004]. Some works lift the requirement of fixed calibration motions [Müller et al. 2016; Taetz et al. 2016] but remain limited to certain body parts. Several works focus on the intrinsic calibration of the raw sensor readings for the IMU to reduce measuring biases, errors, and noise [Bhatia et al. 2016; Lee et al. 2011; Li et al. 2012; Qureshi and Golnaraghi 2017]. However, our work focuses on the extrinsic calibration of IMUs, i.e. their placement in the world.

Most relevant for on-body IMU calibration is TIC [Zuo et al. 2025], which addresses extrinsic IMU calibration in long-term inertial motion capture. In long sequences, IMUs suffer from drift due to sensor noise and measurement errors. TIC reformulates the IMU measurement problem as one where both the world frame and the bone frame may drift. This allows TIC to reduce the drift in long motion sequences, where inertial motion capture would previously fail. Yet, TIC still requires an accurate initial calibration and assumes a perfect initial world-reference frame alignment and is restricted to a known 6 IMU setup. We complement TIC's work by localizing and calibrating IMUs anywhere on the body.

2.3 Visual-Inertial Initialization

Visual information provides an absolute spatial reference that can resolve the ambiguities inherent to purely inertial calibration. This has been utilized in IMU-camera calibration [Fu et al. 2021; Rehder and Siegwart 2017] that achieves errors below 1 in lab settings that require additional calibration targets or rigid mounting of IMU and camera. Li and Mourikis additionally model the time offset between IMU and camera sensor streams [Li and Mourikis 2013]. In robotics and computer vision, this problem is commonly formalized as hand-eye calibration, which solves for a transformation X in equations of the form $AX = XB$, relating known motions of two rigidly coupled sensors [Horaud and Dornaika 1995; Jiang et al. 2022a]. This formulation has been adapted to camera-IMU systems [Guo and Roumeliotis 2013; Wu et al. 2021], but fundamentally relies on the IMU being rigidly attached to the visual system. In contrast, body-worn IMUs are neither rigidly coupled to the camera nor directly visible, and their placement on deformable human surfaces introduces additional ambiguity. As a result, existing visual-inertial calibration techniques do not directly apply to body-imu calibration.

Systems like TotalCapture [Trumble et al. 2017] fuse multi-view video with IMU data for high-fidelity motion capture. Recent advances in human mesh recovery [mes 2026; Zhang et al. 2024], such as PromptHMR [Wang et al. 2025], allow for accurate 3D mesh extraction from single images and videos. Our method leverages this visual prior to solve the "blind" localization and alignment problem. Unlike previous methods that require known IMU positions [Zuo et al. 2025], specific calibration motions [Cutti et al. 2008; Vitali and Perkins 2020; Xsens 2024], or only work on a subset of joints [Müller et al. 2016; Taetz et al. 2016], our method requires only a short video clip of the user and solves the IMU localization, world-reference, and body-sensor alignment.

3 Problem Formulation

We formalize IMU calibration as recovering the spatial and rotational parameters necessary to map raw sensor data into vertex orientations in a shared world frame. We consider a set of K body-worn sensors, where each sensor i provides a length- $\mathcal{T}$ time-series of orientations ${}^{REF}R_S^{(i)}(t)$ and accelerations $\mathbf{a}_{REF}^{(i)}(t)$ measured in its own local reference frame REF_i .

The orientation of body part V in world frame W at time t is given by:

$${}^WR_V(t) = \underbrace{{}^WR_{REF}}_{\text{unknown}} \underbrace{{}^{REF}R_S(t)}_{\text{measured}} \underbrace{{}^{SR}_V}_{\text{unknown}} \quad (1)$$

where V denotes the vertex frame to which the sensor is attached (see Supplementary A). Our goal is to determine the calibration parameters $C_i = \{v_i, {}_S R_V^{(i)}, {}_W R_{REF}^{(i)}\}$ for each sensor $i = 1 \dots K$. Here, v_i denotes the specific attachment location (vertex index) on the underlying SMPL mesh. ${}_S R_V^{(i)}$ represents the sensor-to-body mounting, i.e., the constant local orientation offset that transforms from vertex frame V to sensor frame S . Equivalently, this mounting can be expressed using its inverse rotation, ${}_S R_V = {}_V R_S^{-1}$. The ${}_W R_{REF}^{(i)}$ denotes the alignment transformation that maps the sensor's arbitrary reference frame REF into the shared world frame W . Formally, VAIC takes as input the K time-synchronized sensor measurements ${}_{REF} R_{S_i}$, $\mathbf{a}_{REF_i}$ and a monocular video $\mathcal{V}$ of the actor, and outputs the optimal calibration parameters $\mathcal{C}$ that enable reconstruction of global vertex motion ${}_W R_V$ via the kinematic chain in Eq. 1.

4 Method

VAIC automates the recovery of the calibration parameters $C_i = \{v_i, {}_V R_S^{(i)}, {}_W R_{REF}^{(i)}\}$ for an arbitrary set of K body-worn IMUs using a three-stage pipeline, as highlighted in Alg. 1.

First, we reconstruct human motion from a monocular video to synthesize "virtual" inertial measurements for each vertex in a shared world frame W .

Second, inspired by hand-eye calibration [Horaud and Dornaika 1995], we propose a novel non-linear optimization that solves the sensor-to-body mounting orientation ${}_S R_V$ by aligning measured IMU signals with virtual inertial measurements derived from reconstructed motion. We solve this optimization for every sensor-vertex pair, combining three calibration-specific error terms: orientation consistency, acceleration agreement, and physical mounting constraints. The resulting cost matrix captures how well each physical sensor is explained by each candidate vertex.

Third, we solve a global assignment problem over the cost matrix to uniquely identify the anatomical location of each sensor and subsequently recover the sensor-specific reference frame alignment ${}_W R_{REF}$.

4.1 Video-Based Motion Reconstruction

The pipeline is initialized by extracting a temporal sequence of SMPL body meshes from the input video to serve as a dense kinematic prior. We employ PromptHMR [Wang et al. 2025] to regress the sequence of SMPL parameters, which are estimated in a consistent, gravity-aligned world coordinate frame W .

We then synthesize IMU signals for each vertex j by extracting the time-varying global orientation ${}_W \hat{R}_V^{(j)}$ and computing virtual vertex accelerations $\hat{\mathbf{a}}_W^{(j)}$ via second-order finite differences of the global vertex positions, as done in prior work [Armani et al. 2024; Jiang et al. 2022b; Yi et al. 2022].

PromptHMR's estimates allow VAIC to automatically provide SMPL shape parameters that other methods need to obtain via user body measurements.

4.2 Body-Sensor Mounting

To localize the sensors without prior knowledge, we treat every combination of a physical sensor i and a mesh vertex j as a candidate

Algorithm 1: Automated IMU Calibration

Input : Video $\mathcal{V}$, Sensor Data $\mathcal{S} = \{({}_{REF} R_S^{(i)}, \mathbf{a}_{REF}^{(i)})\}_{i=1}^K$
Output: Calibration parameters $C_i = \{v_i, {}_V R_S, {}_W R_{REF}\}$

```

1  $\mathcal{M}_{SMPL} \leftarrow \text{PromptHMR}(\mathcal{V});$ 
2 foreach vertex  $j \in \{1 \dots N\}$  do
3    $\{{}_W \hat{R}_V^j(t), \hat{\mathbf{a}}_W^j(t)\} \leftarrow \text{IMU-Synthetization}(\mathcal{M}_{SMPL});$ 
4 foreach IMU  $i$  and vertex  $j$  do
5   // find optimal body placement of IMU  $i$  on vertex  $j$ 
6   using  $\{({}_{REF} R_S^{(i)}(t), \mathbf{a}_{REF}^{(i)}(t))\}$  and  $\{{}_W \hat{R}_V^j(t), \hat{\mathbf{a}}_W^j(t)\}$ ;
7    ${}_{\mathcal{T}} \omega_S^*, \mathbf{E}_{i,j} \leftarrow$ 
8      $\text{argmin}_{{}_{\mathcal{T}} \omega_S} (E_{\text{ori}}({}_{\mathcal{T}} \omega_S) + \lambda_n E_{\text{norm}}({}_{\mathcal{T}} \omega_S) + \lambda_a E_{\text{acc}}({}_{\mathcal{T}} \omega_S));$ 
9   Store  ${}_V R_S^{(i,j)} \leftarrow {}_V R_T^{(j)} \text{Exp}({}_{\mathcal{T}} \omega_S^*);$ 
10 Assignment  $\leftarrow \text{HungarianAlg}(\mathbf{E});$  // Optimal matches
11 foreach  $(i, j) \in \text{Assignment}$  do
12    ${}_W R_{REF}^{(i)} \leftarrow \text{Avg}_t ({}_W \hat{R}_V^{(j)}(t) \cdot {}_V R_S^{(i,j)} \cdot {}_S R_{REF}^{(i)}(t));$ 
13 return  $\{v_j, {}_V R_S^{(i,j)}, {}_W R_{REF}^{(i)}\};$ 
```

mounting hypothesis. For each of the $K \times N$ pairs, we seek to estimate the optimal local mounting orientation ${}_V R_S$ that minimizes the discrepancy between the real sensor measurements and the virtual vertex data.

For a fixed pair (i, j) , we formulate this as a non-linear minimization problem with three cost terms. E_{ori} aligns the virtual and measured relative orientation changes, E_{norm} ensures the sensor sits flush against the skin, and E_{acc} enforces consistency between virtual and measured accelerations. To effectively apply the physical surface constraints in E_{norm} , we re-parameterize the mounting orientation ${}_V R_S$ with ${}_{\mathcal{T}} R_S$ (see Sec. 4.2.2) and optimize over it.

Since ${}_S R_T$ lies on the $SO(3)$ manifold, a direct minimization over the rotation matrix is intractable. Instead, we rely on Lie algebra and optimize over the vector ${}_{\mathcal{T}} \omega_S = \text{Log}({}_{\mathcal{T}} R_S)$, ensuring valid rotation matrices [Sola et al. 2018]:

$${}_{\mathcal{T}} \omega_S, \epsilon = \underset{{}_{\mathcal{T}} \omega_S}{\text{argmin}} \left(E_{\text{ori}}({}_{\mathcal{T}} \omega_S) + \lambda_{\text{norm}} E_{\text{norm}}({}_{\mathcal{T}} \omega_S) + \lambda_{\text{acc}} E_{\text{acc}}({}_{\mathcal{T}} \omega_S) \right) \quad (2)$$

Once converged, we reconstruct the final mounting orientation ${}_V R_S$ from ${}_{\mathcal{T}} R_S = \text{Exp}({}_{\mathcal{T}} \omega_S)$. Here $\text{Exp}(\cdot)$ and $\text{Log}(\cdot)$ represent the Lie algebra maps connecting the optimization vector space and the rotation manifold [Sola et al. 2018].

4.2.1 Orientation Consistency. The orientation error aims to estimate the mounting ${}_V R_S$ by aligning the synthetic and measured orientation changes. Since the absolute global alignment ${}_W R_{REF}$ is unknown, we cannot compare orientations directly. Instead, we rely on the Hand-Eye calibration approach [Horaud and Dornaika 1995], which compares relative rotations eliminating ${}_W R_{REF}$ from the formulation. Specifically, the relative rotation of the vertex frame between time steps $t - 1$ and t must match the relative rotation

of the sensor frame, mapped through the fixed mounting offset: $V_{(t-1)}\hat{R}_V(t)V R_S = V R_{SS(t-1)}R_{S(t)}.$

Consequently, by rearranging estimating $V R_S$ can be formulated as the following minimization problem:

$$V_{(t-1)}\hat{R}_V(t)V R_{SS(t)}R_{S(t-1)}S R_V = I \quad (3)$$

Since we re-parameterize $V R_S = V R_T \text{Exp}(T \omega_S)$ (see Sec. 4.2.2), the formulation for our error term using Lie algebra yields:

$$E_{\text{ori}} = \sum_t \left\| \text{Log} \left(V_{(t)}\hat{R}_V(t-1) \cdot [V R_T \text{Exp}(T \omega_S)] \cdot S_{(t-1)}R_{S(t)} \cdot [V R_T \text{Exp}(T \omega_S)]^T \right) \right\|^2 \quad (4)$$

4.2.2 Normal Error. The normal error E_{norm} exploits the fact that IMUs are typically mounted flush against the skin, thereby discouraging physically infeasible vertex-to-sensor rotations $V R_S$.

While E_{ori} constrains the rigid relationship between frames, it is mathematically prone to symmetric ambiguities, such as distinguishing a sensor on the top of the wrist from one on the bottom. To resolve this, we introduce a geometric prior E_{norm} based on the physical observation that body-worn sensors are typically mounted flat against the skin, aligning the Z-axis of the sensor with the surface normal. We decompose the mounting orientation into $V R_S = V R_T \cdot T R_S$, where T is a tangent coordinate frame aligned with the vertex surface normal. The frame T is rigidly attached to the vertex frame. As a result, the rotation $V R_T$ is constant and can be precomputed. In this tangent frame, a surface-aligned sensor should exhibit rotation primarily around the z-axis (yaw), with negligible tilt components. We therefore strictly penalize the non-yaw components (x and y) of the optimization vector $T \omega_S = \text{Log}(T R_S)$:

$$E_{\text{norm}} = \left\| [T \omega_S]_{x,y} \right\|^2 \quad (5)$$

4.2.3 Acceleration Error. To further refine the localization, particularly along bone segments where rotations are similar but linear velocities differ, we incorporate an acceleration error term E_{acc} . We project the sensor acceleration $\mathbf{a}_{\text{REF}}$ measured in the sensor's reference frame REF into the global world frame using the current kinematic estimate: $W R_{\text{REF}}^{\text{est}}(t) = W \hat{R}_V(t) V R_{SS} R_{\text{REF}}(t)$. This projected acceleration is compared against the virtual vertex acceleration $\hat{\mathbf{a}}_W$:

$$E_{\text{acc}} = \sum_t \left\| \hat{\mathbf{a}}_W(t) - \left(W R_{\text{REF}}^{\text{est}}(t) \cdot \mathbf{a}_{\text{REF}}(t) \right) \right\|^2 \quad (6)$$

Note for simplicity, in the error term, we do not enforce a fixed $W R_{\text{REF}}^{\text{est}}$ for the whole sequence.

4.3 Vertex Assignment

After performing the pairwise optimization, we obtain a cost matrix $\mathbf{E} \in \mathbb{R}^{K \times N}$, where each entry $\epsilon_{i,j}$ represents the minimum residual error for mounting sensor i to vertex j . To identify the specific anatomical location of each sensor, we formulate this as a global linear assignment problem.

We apply the Hungarian algorithm (matching) to find the optimal mapping that minimizes the total system error [Crouse 2016]. This step uniquely assigns each physical IMU to a specific vertex index v_i on the mesh, thereby localizing the sensors on the body without requiring manual measurements.

4.4 Reference Frame Alignment

With the sensor location v_i and the mounting orientation $V R_S$ now determined, the final step is to recover the global alignment $W R_{\text{REF}}$. Theoretically, this transformation is constant. However, due to errors in the video-based pose estimation, computing it from a single frame is unreliable.

Instead, we first compute the instantaneous alignment candidate $W R_{\text{REF}}(t)$ for every time step t using the kinematic chain:

$$W R_{\text{REF}}(t) = W \hat{R}_V(t) \cdot V R_S \cdot S R_{\text{REF}}(t) \quad (7)$$

where $S R_{\text{REF}}(t)$ is the inverse of the measured sensor orientation. To obtain a robust, static estimate, we compute the mean rotation over the sequence using the Karcher-Mean [Moakher 2002]:

$$W R_{\text{REF}} = \text{Karcher-Mean}(W R_{\text{REF}}(0), \dots, W R_{\text{REF}}(\mathcal{T})) \quad (8)$$

This effectively solves the calibration problem, yielding the complete parameter set C_i for each sensor: its specific location on the SMPL mesh v_i , the local body mounting $V R_S^{(i)}$, and the global reference alignment $W R_{\text{REF}}^{(i)}$.

5 Experiments

We evaluate VAIC on the TotalCapture (TC) dataset [Trumble et al. 2017], a standard benchmark for inertial motion capture providing synchronized video, IMU measurements, and global reference sensor alignments $W R_{\text{REF}}^{\text{gt}}$. Ground-truth sensor-to-bone orientations $V R_S^{\text{gt}}$ are computed from GT-SMPL poses obtained via optical motion capture [Huang et al. 2018; Mahmood et al. 2019]. IMU placements are manually annotated using multi-view video data.

Each TotalCapture sequence begins with a T-Pose, which we use for the standard T-Pose calibration baseline. The remaining sequence contains diverse motions ranging from low- to high-dynamic activities, including unconstrained “ROM” motions that are particularly challenging for our method. In contrast to pose-based calibration approaches, VAIC performs calibration directly from these in-the-wild motions without requiring user instructions.

To compare against additional calibration baselines, we recorded a second extensive dataset, DS_{VAIC} , using the same Xsens IMUs as in TC. Ground-truth SMPL poses are obtained via optical motion capture and MoSh++ [Loper et al. 2014; Mahmood et al. 2019], while $V R_S^{\text{gt}}$ and $W R_{\text{REF}}^{\text{gt}}$ are measured using the motion capture system. We compare against T-/N-Pose calibration, N-Pose→T-Pose, and N-Pose+walking calibration, the latter closely matching the commercial Xsens calibration procedure. We use the implementations provided by GlobalPose [Yi et al. 2025]. It totals roughly 70 minutes with 8 participants. DS_{VAIC} uses the standard 6-IMU placement on wrists, legs, pelvis, and head with the focus on evaluating sensor calibration. We evaluate three aspects: blind sensor localization accuracy, automated calibration quality compared to standard pose-based calibrations, and comparison to the learning-based baseline TIC [Zuo et al. 2025].

5.1 Metrics

We evaluate VAIC by assessing localization accuracy and calibration orientation quality. We report the following metrics:

Table 1. VAIC accurately localizes all 13 IMUs anywhere on the body. SMPL GT indicates the maximal feasible calibration results by replacing camera generated SMPL with GT SMPL. Metrics reported as mean $\pm$ standard deviation. AE means angle error.

Method	Seq.	Bone Acc. (%)	Pos. Err. (cm)	AE wR_V ($^\circ$)	AE vR_S ($^\circ$)	AE wR_{REF} ($^\circ$)
VAIC	10 s	95.5	6.0 ± 8.7	16.6 ± 11.6	16.5 ± 13.6	11.7 ± 9.2
VAIC	20 s	95.9	5.8 ± 8.6	15.8 ± 10.8	15.8 ± 12.2	10.5 ± 5.2
VAIC	30 s	95.0	6.2 ± 8.9	15.5 ± 9.2	15.5 ± 10.2	9.6 ± 5.2
VAIC w/ SMPL GT	10 s	99.3	3.6 ± 3.7	8.0 ± 5.4	10.2 ± 6.8	5.6 ± 4.1

Table 2. Quantitative comparison between TIC and VAIC. TIC tries to continuously calibrate from known IMU positions. TIC assumes to start from calibrated IMUs and fails with uncalibrated initialization. VAIC calibrates more accurately than TIC. Only 6 IMUs are used here, as TIC only supports Head, Pelvis, Lower Arms and Lower Leg IMUs.

Method	Initialization	AE wR_V ($^\circ$)	AE vR_S ($^\circ$)	AE wR_{REF} ($^\circ$)
TIC	Raw	92.4 ± 30.2	56.3 ± 35.4	33.4 ± 27.0
TIC	T-Pose	15.5 ± 2.3	16.5 ± 4.2	7.4 ± 2.3
TIC	VAIC 10s	12.6 ± 3.6	11.2 ± 3.0	6.8 ± 3.8
TIC	VAIC 20s	12.5 ± 3.8	11.9 ± 4.0	6.8 ± 4.0
T-Pose	-	16.7 ± 8.7	19.5 ± 11.5	7.5 ± 5.4
VAIC 10 s	Raw	16.1 ± 6.3	14.3 ± 7.8	11.0 ± 6.0
VAIC 20 s	Raw	15.3 ± 6.1	14.4 ± 8.6	10.0 ± 6.5

Bone Accuracy (%): The percentage of predicted vertex positions that are assigned to the correct bone segment.

Position Error (cm): The Root Mean Squared Error (RMSE) between the predicted vertex position and the ground truth position in the T-Pose configuration.

Angle Error wR_V ($^\circ$): The end-to-end angular distance between the ground truth vertex orientation (derived from SMPL) and the predicted orientation computed via the kinematic chain $wR_{REF} \cdot REF_{RS} \cdot sR_V$ over the whole motion sequence. This metric reflects the cumulative system accuracy.

Angle Error vR_S ($^\circ$): The geodesic angular distance between the predicted and ground truth sensor mounting orientations (sensor-to-bone calibration).

Angle Error wR_{REF} ($^\circ$): The angular error between the predicted and ground truth transformations from the reference frame to the world frame (coordinate frame alignment).

5.2 Quantitative Results

5.2.1 IMU Localization. VAIC localizes IMUs to the correct body segment with up to 95.9% accuracy and a mean position error of 5.8 cm (Tab. 1). Most failures occur between adjacent body segments, such as confusing lower- and upper-arm sensors, which still results in relatively small spatial errors. These errors are concentrated in the low-motion “ROM” sequences, where insufficient excitation reduces identifiability for our optimization. Excluding these sequences improves localization accuracy to 97.9% and reduces the position error to 5.1 cm.

The method also achieves accurate orientation recovery with an end-to-end angular error (AE wR_V) of 15.5° . Increasing calibration

Table 3. Quantitative comparison between different calibration methods on DS_{VAIC} using known positions of the 6 IMUs.

Method	AE wR_V ($^\circ$)	AE vR_S ($^\circ$)	AE wR_{REF} ($^\circ$)
N-Pose + Walking	26.2 ± 3.5	36.1 ± 12.1	27.0 ± 11.8
N-Pose $\rightarrow$ T-Pose	20.5 ± 2.9	33.1 ± 14.8	26.1 ± 14.8
N-Pose	35.9 ± 6.8	69.7 ± 23.7	65.4 ± 26.2
T-Pose	32.4 ± 8.6	67.2 ± 24.1	65.4 ± 26.2
TIC + T-Pose	34.6 ± 12.2	41.4 ± 17.9	58.0 ± 26.0
TIC + VAIC	15.6 ± 3.5	13.5 ± 3.1	11.4 ± 4.0
VAIC 20 s	17.4 ± 3.4	16.9 ± 3.1	14.2 ± 3.1

Table 4. Ablation study of optimization terms in VAIC (10 s sequence length). Both the normal error E_{norm} and acceleration error E_{acc} improve localization and the angle error. Metrics are reported as mean $\pm$ standard deviation.

Optimization Terms			Metrics				
Ori	Acc	Norm	Bone Acc. (%)	Pos. Err. (cm)	AE wR_V ($^\circ$)	AE vR_S ($^\circ$)	AE wR_{REF} ($^\circ$)
✓			90.1	14.5 ± 10.9	19.5 ± 13.8	21.4 ± 19.1	15.9 ± 17.3
✓		✓	93.5	7.6 ± 10.4	18.7 ± 13.4	20.2 ± 18.9	15.9 ± 17.0
✓	✓		91.6	10.7 ± 9.5	17.0 ± 11.1	17.3 ± 14.2	11.6 ± 10.7
✓	✓	✓	95.5	6.0 ± 8.7	16.6 ± 11.6	16.5 ± 13.6	11.7 ± 9.2

duration beyond 20 s provides little additional benefit. Notably, VAIC remains robust on challenging freestyle motions, including crawling and rolling sequences that differ substantially from standard calibration actions. Using GT SMPL instead of monocularly estimated motion further improves localization to 99.3% accuracy and reduces angular error to 8.0° , indicating that video pose estimation remains the primary bottleneck.

5.2.2 IMU Calibration. We evaluate calibration quality against both scripted pose-based procedures and the continuous calibration method TIC [Zuo et al. 2025]. On DS_{VAIC} , VAIC achieves an end-to-end orientation error of 17.4° , with sensor-to-body and reference-frame errors of 16.9° and 14.2° , respectively (Tab. 3), outperforming all scripted calibration baselines. The strongest pose-based baseline, N-Pose $\rightarrow$ T-Pose, reaches 20.5° AE wR_V , while the Xsens-like N-Pose+walking procedure obtains 26.2° ; both also exhibit substantially larger sensor mounting errors than VAIC (33.1° and 36.1° vs. 16.9°). Static N-/T-Pose calibrations perform substantially worse on DS_{VAIC} , which we attribute to less reliable initial IMU reference-frame estimates, in particular heading errors. Since static pose-based methods assume a correct global reference alignment, reference-frame errors are absorbed into the estimated sensor-to-body mounting offsets and propagate to both vR_S and wR_{REF} errors. Errors on DS_{VAIC} are generally higher than on TotalCapture because its recordings exhibit stronger magnetic disturbances and are slightly longer, leading to larger accumulated heading drift. A single T-Pose cannot reliably correct the resulting world-reference misalignment, which particularly affects pose-based calibration methods that depend on an accurate initial reference-frame alignment. In contrast, VAIC exploits relative orientation changes over arbitrary motion, making it less sensitive to pose execution errors and imperfect reference-frame alignment.

We further compare against TIC, which is designed to continuously refine an already calibrated inertial setup rather than solve

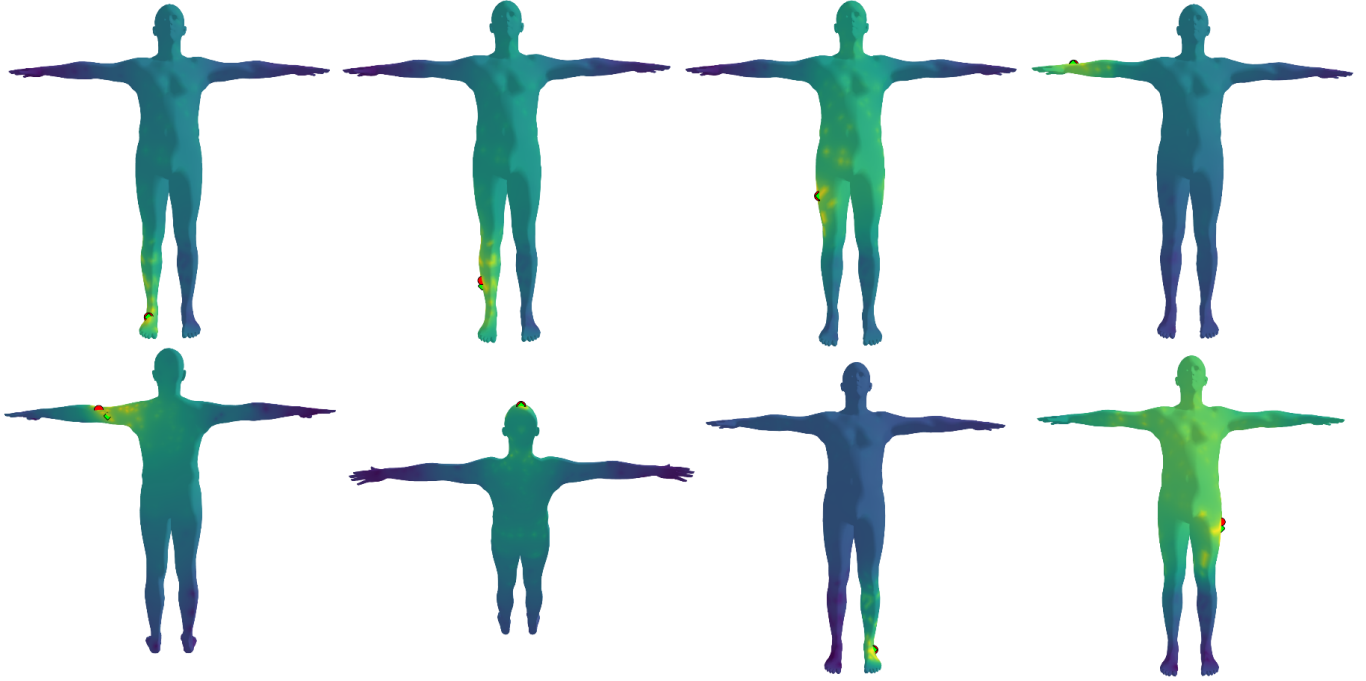


Fig. 3. Localization results of VAIC on a 20s walking calibration sequence. VAIC accurately recovers the positions of uncalibrated IMUs on the body. Red circles indicate predicted positions, and green diamonds show ground truth. Brighter colors indicate a higher matching score between the IMU and the vertex. It is generally seen that vertices on the same body part have higher match scores. The correct localization on that body part is aided by the normal error that models the sensor lying flat on the body.

cold-start initialization. This is confirmed on TotalCapture in Tab. 2: from raw sensor orientations, TIC yields 92.4° AE_{wR_V} , but with a manual T-Pose initialization it performs well, reaching 15.5° AE_{wR_V} , 16.5° AE_{vR_S} , and 7.4° $\text{AE}_{wR_{REF}}$ on the six supported IMUs. Thus, TIC is effective when the initial calibration and reference-frame alignment are sufficiently accurate, but can struggle when initialization is too far from the true solution. By design, VAIC targets this cold-start setting: starting from uncalibrated measurements, it achieves comparable end-to-end accuracy and better local mounting calibration using only 20 s of arbitrary motion (15.3° AE_{wR_V} , 14.4° AE_{vR_S}), while its global reference alignment is slightly less accurate (10.0°).

The two methods are therefore complementary. When TIC is initialized with VAIC, performance improves to 12.5° AE_{wR_V} and 11.9° AE_{vR_S} on TotalCapture. The same trend appears on DS_{VAIC} , where TIC+VAIC achieves the best overall results, reducing the errors to 15.6° , 13.5° , and 11.4° for wR_V , vR_S , and wR_{REF} , respectively. Thus, VAIC can provide an automatic initialization for TIC, which can then be used for continuous refinement during extended capture sessions.

5.3 Qualitative Results

As shown before, VAIC accurately recovers the location of the IMUs. This is visible in Fig. 3 where IMUs are often assigned to the correct vertex and vertices in that region also experience the lowest error terms, indicated by the brighter colors. Fig. 4 shows that the IMU

body mounting is well estimated: an error of just 14.4° looks visually close to the ground truth.

5.4 User Study

In an 8-participant counterbalanced user study, we compared the calibration procedure of VAIC against Xsens, a commercial IMU-based motion capture system requiring precise sensor placement, user-specific measurements (e.g., height and shoe size), and a two-step calibration procedure comparable to the N-pose + walking method (Table 3). Participants were first introduced to both calibration procedures and then performed the calibration process, from sensor placement to calibration poses. Afterwards, they rated ease of use, likelihood of user errors, and overall preference on a Likert scale. If Xsens reported poor calibration quality, participants repeated the procedure once. VAIC reduced sensor placement time by 36% and overall calibration time by 60% (3 minutes faster), as sensors could be attached arbitrarily and localized automatically. Participants also significantly preferred VAIC over the Xsens calibration procedure (Xsens: 3.3 vs. VAIC: 5, $p < .01$), reporting that Xsens required greater precision and concentration during both sensor placement and calibration. Overall, participants favored VAIC due to its simpler workflow, requiring only unconstrained sensor placement and a short self-recorded video.

This user study demonstrates the practical benefits of VAIC for everyday motion capture scenarios. By automatically localizing IMUs

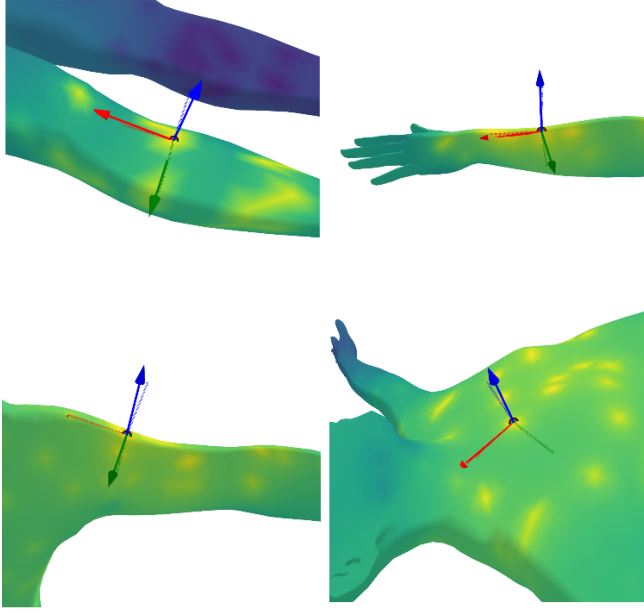


Fig. 4. Visualization of different vRS results of VAIC. VAIC generates accurate calibration results. Dashed lines indicate the predicted frame, solid lines the ground truth frame.

and eliminating the need for precise calibration poses, VAIC substantially reduces both setup time and calibration effort compared to conventional IMU-based motion capture workflows.

5.5 Ablation Study

We analyze the contribution of each term in our objective function in Table 4. The orientation term (E_{ori}) alone achieves reasonable bone association (90.1%) but suffers from high spatial ambiguity (14.5 cm error), as multiple vertices on a limb share similar rotational trajectories. Incorporating the surface normal prior (E_{norm}) encourages the physical constraint that sensors are mounted flush with the skin; this significantly improves localization precision, reducing position error to 7.6 cm by ruling out geometrically valid but implausible orientations. This is particularly beneficial for IMUs on the rigid torso, where it helps distinguish between front- and back-mounted sensors. The acceleration term (E_{acc}) proves critical for global alignment and disambiguation along the kinematic chain. By matching linear accelerations, which vary along a limb, it provides a strong reference constraint that rotation alone cannot resolve, reducing the global alignment error (wR_{REF}) from 15.9° to 11.6° . Combining all terms yields the highest performance across all metrics, confirming that robust calibration requires satisfying both kinematic and geometric constraints.

6 Discussion

VAIC accurately localizes and calibrates IMUs on the body, yet certain limitations remain. Errors are higher on static or symmetric motions, such as the "ROM" sequences, suggesting that VAIC could

benefit from improved heuristics to identify and filter static calibration intervals. The upper bound of VAIC's calibration quality is higher when using ground-truth motion capture data, highlighting that the accuracy of monocular SMPL estimation remains a limiting factor. While VAIC's optimization is fast (see Suppl. B.3), obtaining the required vision-based mesh estimates incurs additional runtime and currently dominates the overall calibration time. As video-based pose estimators continue to improve in both quality and inference speed, we expect corresponding gains in VAIC's performance.

We show that assuming a flush sensor mounting and enforcing a normal-alignment prior improves IMU localization, as ambiguous positions can be differentiated more reliably. For example, torso front/back placements can yield similar IMU signals under a 180° calibration flip, so the prior favors candidates whose sensor axis agrees with the mesh normal. This prior is physically motivated: wearable IMUs are typically mounted on a flat PCB inside a casing, with one side attached to the body via a strap. The PCB is therefore approximately tangent to the body surface, and one sensor axis (typically the z-axis) roughly aligns with the surface normal. However, when IMUs are mounted with random orientation to the body or the mounting is not flush against the body, this prior may fail and localization accuracy can degrade.

VAIC excels in body-sensor mounting vRS , while global alignment wR_{REF} is comparable to baselines. Crucially, VAIC aligns all IMUs to a consistent camera-defined world frame; this ensures robust kinematic reconstruction even in environments where magnetic interference renders magnetometer-based alignment unreliable. Currently, VAIC treats all time steps equally in its optimization. Incorporating uncertainty-aware pose estimates or RANSAC-style outlier rejection could further enhance robustness by down-weighting or removing time intervals that hinder convergence, leading to more accurate calibration in challenging sequences.

7 Conclusion

In this work, we presented VAIC, a video-aided framework that addresses a fundamental bottleneck of inertial motion capture: the calibration of sensor location, body-mounting orientation, and world reference alignment. By eliminating the reliance on manual measurements, specific calibration poses, and exact placement of IMUs on the body, we enable an effortless and accurate calibration.

Our approach leverages a short video to calibrate IMUs anywhere on the body. By extracting SMPL poses from a video clip, we formulate calibration as a nonlinear optimization problem. A key technical contribution is our use of a surface-normal geometric prior, which resolves the symmetric ambiguities inherent to rotation-only matching by constraining the sensor mounting to a reliable yaw estimation. With this optimization, VAIC recovers all calibration parameters and user shape from an arbitrary calibration motion. Experimental evaluation on the TotalCapture benchmark demonstrates that VAIC blindly localizes sensors to the correct body segment with 95.9% accuracy. Furthermore, it surpasses standard calibration protocols in orientation accuracy while being preferred by users as simpler to use. By automating the initialization of sparse inertial systems, VAIC bridges the gap between controlled laboratory capture and accessible, everyday motion recording.

References

2026. Meshcapade | Providing the best markerless motion capture solution — meshcapade.com. <https://meshcapade.com/>. [Accessed 22-01-2026].
2026. Rokoko - Studio-grade motion capture tools for all creators — rokoko.com. <https://www.rokoko.com/>. [Accessed 21-01-2026].
- Rayan Armani, Changlin Qian, Jiaxi Jiang, and Christian Holz. 2024. Ultra Inertial Poser: Scalable Motion Capture and Tracking from Sparse Inertial Sensors and Ultra-Wideband Ranging. In *ACM SIGGRAPH 2024 Conference Papers* (Denver, CO, USA) (*SIGGRAPH '24*). Association for Computing Machinery, New York, NY, USA, Article 51, 11 pages. doi:10.1145/3641519.3657465
- Sitank Bhatia, Hai Yang, Rui Zhang, Fabian Höflinger, and Leonhard Reindl. 2016. Development of an analytical method for IMU calibration. In *2016 13th International Multi-Conference on Systems, Signals & Devices (SSD)*. IEEE, 131–135.
- Alessandro Bonfiglio, Elisabetta Farella, David Tacconi, and Raoul M Bongers. 2025. Effects of different inertial measurement unit sensor-to-segment calibrations on clinical 3-dimensional humerorhthoric joint angles estimation. *Journal of applied biomechanics* 41, 1 (2025), 37–46.
- Namchol Choe, Hongyu Zhao, Sen Qiu, and Yongguk So. 2019. A sensor-to-segment calibration method for motion capture system based on low cost MIMU. *Measurement* 131 (2019), 490–500.
- David F Crouse. 2016. On implementing 2D rectangular assignment algorithms. *IEEE Trans. Aerospace Electron. Systems* 52, 4 (2016), 1679–1696.
- Andrea Giovanni Cutti, Andrea Giovanardi, Laura Rocchi, Angelo Davalli, and Rinaldo Sacchetti. 2008. Ambulatory measurement of shoulder and elbow kinematics through inertial and magnetic sensors. *Medical & biological engineering & computing* 46, 2 (2008), 169–178.
- Julien Favre, Rachid Aissaoui, Brigitte M Jolles, Jacques A de Guise, and Kamiar Aminian. 2009. Functional calibration procedure for 3D knee joint angle description using inertial sensors. *Journal of biomechanics* 42, 14 (2009), 2330–2335.
- Bo Fu, Fuzhang Han, Yue Wang, Yanmei Jiao, Xiaqing Ding, Qimeng Tan, Lei Chen, Minhang Wang, and Rong Xiong. 2021. High-precision multicamera-assisted camera-IMU calibration: Theory and method. *IEEE Transactions on Instrumentation and Measurement* 70 (2021), 1–17.
- Chao X Guo and Stergios I Roumeliotis. 2013. IMU-RGBD camera 3D pose estimation and extrinsic calibration: Observability analysis and consistency improvement. In *2013 IEEE International Conference on Robotics and Automation*. IEEE, 2935–2942.
- Robert Harle. 2013. A Survey of Indoor Inertial Positioning Systems for Pedestrians. *IEEE Communications Surveys & Tutorials* 15, 3 (2013), 1281–1293. doi:10.1109/SURV.2012.121912.00075
- Dominik Hollidt, Tommaso Bendinelli, and Christian Holz. 2026. Ultra Diffusion Poser: Diffusion-Based Human Motion Tracking from Sparse Inertial Sensors and Ranging-based Between-sensor Distances. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7036–7046.
- Dominik Hollidt, Paul Strelci, Jiaxi Jiang, Yasaman Haghighi, Changlin Qian, Xintong Liu, and Christian Holz. 2025. EgoSim: An Egocentric Multi-view Simulator and Real Dataset for Body-worn Cameras during Motion and Activity. *Advances in Neural Information Processing Systems* 37 (2025), 106607–106627.
- Radu Horaud and Fadi Dornaika. 1995. Hand-eye calibration. *The international journal of robotics research* 14, 3 (1995), 195–210.
- Yinghao Huang, Manuel Kaufmann, Emre Aksan, Michael J Black, Otmar Hilliges, and Gerard Pons-Moll. 2018. Deep inertial poser: Learning to reconstruct human pose from sparse inertial measurements in real time. *ACM Transactions on Graphics (TOG)* 37, 6 (2018), 1–15.
- Jianfeng Jiang, Xiao Luo, Qingsheng Luo, Lijun Qiao, and Minghao Li. 2022a. An overview of hand-eye calibration. *The International Journal of Advanced Manufacturing Technology* 119, 1 (2022), 77–97.
- Yifeng Jiang, Yuting Ye, Deepak Gopinath, Jungdam Won, Alexander W Winkler, and C Karen Liu. 2022b. Transformer inertial poser: Real-time human motion reconstruction from sparse imus with simultaneous terrain generation. In *SIGGRAPH Asia 2022 Conference Papers*. 1–9.
- Anthony Kim and MF Golnaraghi. 2004. Initial calibration of an inertial measurement unit using an optical position tracking system. In *PLANS 2004. Position Location and Navigation Symposium (IEEE Cat. No. 04CH37556)*. IEEE, 96–101.
- Dongkyu Lee, Sangchul Lee, Sanghyuk Park, and Sangho Ko. 2011. Test and error parameter estimation for MEMS-based low cost IMU calibration. *International Journal of Precision Engineering and Manufacturing* 12, 4 (2011), 597–603.
- Mingyang Li and Anastasios I Mourikis. 2013. 3-D motion estimation and online temporal calibration for camera-IMU systems. In *2013 IEEE International Conference on Robotics and Automation*. IEEE, 5709–5716.
- You Li, Xiaojie Niu, Quan Zhang, Hongping Zhang, and Chuang Shi. 2012. An in situ hand calibration method using a pseudo-observation scheme for low-end inertial measurement units. *Measurement Science and Technology* 23, 10 (2012), 105104.
- Huakun Liu, Hiroki Ota, Xin Wei, Yutaro Hirao, Monica Perusquia-Hernandez, Hideaki Uchiyama, and Kiyoshi Kiyokawa. 2025. UMotion: Uncertainty-driven Human Motion Estimation from Inertial and Ultra-wideband Units. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 7085–7094.
- Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. 2015. SMPL: a skinned multi-person linear model. *ACM Transactions on Graphics (TOG)* 34, 6 (2015), 1–16.
- Matthew M. Loper, Naureen Mahmood, and Michael J. Black. 2014. MoSh: Motion and Shape Capture from Sparse Markers. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)* 33, 6 (Nov. 2014), 220:1–220:13. doi:10.1145/2661229.2661273
- Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. 2019. AMASS: Archive of Motion Capture as Surface Shapes. In *Proceedings of the International Conference on Computer Vision*. 5442–5451.
- Mhairi McInne, Dimitra Blana, Andrew Starkey, and Edward K Chadwick. 2025. A Practical Sensor-to-Segment Calibration Method for Upper Limb Inertial Motion Capture in a Clinical Setting. *IEEE Journal of Translational Engineering in Health and Medicine* (2025).
- Soroush Mehraban, Andrea Iaboni, and Babak Taati. 2026. Fasthmr: Accelerating human mesh recovery via token and layer merging with diffusion decoding. In *2026 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 6733–6743.
- Maher Moakher. 2002. Means and Averaging in the Group of Rotations. *SIAM J. Matrix Anal. Appl.* 24, 1 (2002), 1–16. arXiv:https://doi.org/10.1137/S0895479801383877 doi:10.1137/S0895479801383877
- Philipp Müller, Marc-André Bégin, Thomas Schauer, and Thomas Seel. 2016. Alignment-free, self-calibrating elbow angles measurement using inertial sensors. *IEEE journal of biomedical and health informatics* 21, 2 (2016), 312–319.
- Milad Nazarahari, Alireza Noamani, Niloufar Ahmadian, and Hossein Rouhani. 2019. Sensor-to-body calibration procedure for clinical motion analysis of lower limb using magnetic and inertial measurement units. *Journal of Biomechanics* 85 (2019), 224–229.
- Guanglei Ouyang and Karim Abed-Meraim. 2022. A Survey of Magnetic-Field-Based Indoor Localization. *Electronics* 11, 6 (2022). doi:10.3390/electronics11060864
- Umar Qureshi and Farid Golnaraghi. 2017. An algorithm for the in-field calibration of a MEMS IMU. *IEEE Sensors Journal* 17, 22 (2017), 7479–7486.
- Joern Rehder and Roland Siegwart. 2017. Camera/IMU calibration revisited. *IEEE Sensors Journal* 17, 11 (2017), 3257–3268.
- Zehong Shen, Huaijin Pi, Yan Xia, Zhi Cen, Sida Peng, Zechen Hu, Hujun Bao, Ruizhen Hu, and Xiaowei Zhou. 2024. World-Grounded Human Motion Recovery via Gravity-View Coordinates. In *SIGGRAPH Asia Conference Proceedings*.
- Joan Sola, Jeremie Deray, and Dinesh Atchuthan. 2018. A micro lie theory for state estimation in robotics. *arXiv preprint arXiv:1812.01537* (2018).
- Bertram Tatz, Gabriele Bleser, and Markus Miezal. 2016. Towards self-calibrating inertial body motion capture. In *2016 19th International Conference on Information Fusion (FUSION)*. IEEE, 1751–1759.
- Matthew Trumble, Andrew Gilbert, Charles Malleson, Adrian Hilton, and John Collo-mosse. 2017. Total Capture: 3D Human Pose Estimation Fusing Video and Inertial Sensors. In *Proceedings of the 28th British Machine Vision Conference*. 1–13.
- Tom Van Wouwe, Seunghwan Lee, Antoine Falisse, Scott Delp, and C Karen Liu. 2024. Diffusionposer: Real-time human motion reconstruction from arbitrary sparse sensors using autoregressive diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2513–2523.
- Rachel V. Vitali and Noel C. Perkins. 2020. Determining Anatomical Frames via Inertial Motion Capture: A Survey of Methods. 106 (2020), 109832. doi:10.1016/j.jbiomech.2020.109832
- Timo Von Marcard, Bodo Rosenhahn, Michael J Black, and Gerard Pons-Moll. 2017. Sparse inertial poser: Automatic 3d human pose estimation from sparse imus. In *Computer graphics forum*, Vol. 36. Wiley Online Library, 349–360.
- Yufu Wang, Yu Sun, Priyanka Patel, Kostas Daniilidis, Michael J Black, and Muhammed Kocabas. 2025. PromptHMR: Promptable Human Mesh Recovery. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 1148–1159.
- Jin Wu, Miaomiao Wang, Yi Jiang, Bowen Yi, Rui Fan, and Ming Liu. 2021. Simultaneous hand-eye/robot-world/camera-IMU calibration. *IEEE/ASME Transactions on Mechatronics* 27, 4 (2021), 2278–2289.
- Xsens. 2024. Xsens Motion Capture. <https://www.xsens.com> Accessed: March 7, 2025.
- Vasco Xu, Chenfeng Gao, Henry Hoffmann, and Karan Ahuja. 2024. Mobileposer: Real-time full-body pose estimation and 3d human translation from imus in mobile consumer devices. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*. 1–11.
- Ying Xue, Jiaxi Jiang, Rayan Armani, Dominik Hollidt, Yi-Chi Liao, and Christian Holz. 2025. Group Inertial Poser: Multi-Person Pose and Global Translation from Sparse Inertial Sensors and Ultra-Wideband Ranging. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, New Orleans, LA, USA, 24910–24921. doi:10.48550/arXiv.2510.21654
- Xinyu Yi, Yuxiao Zhou, Marc Habermann, Soshi Shimada, Vladislav Golyanik, Christian Theobalt, and Feng Xu. 2022. 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*. 13167–13178.
- Xinyu Yi, Yuxiao Zhou, and Feng Xu. 2021. Transpose: Real-time 3d human translation and pose estimation with six inertial sensors. *ACM Transactions On Graphics (TOG)* 40, 4 (2021), 1–13.
- Xinyu Yi, Yuxiao Zhou, and Feng Xu. 2024. Physical non-inertial poser (PNP): modeling non-inertial effects in sparse-inertial human motion capture. In *ACM SIGGRAPH 2024 Conference Papers*. 1–11.
- Siwei Zhang, Bharat Lal Bhatnagar, Yuanlu Xu, Alexander Winkler, Petr Kadlecek, Siyu Tang, and Federica Bogo. 2024. Rohm: Robust human motion reconstruction via diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14606–14617.
- Haozhe Zhou, Riku Arakawa, Yuvraj Agarwal, and Mayank Goel. 2025. IMUCoCo: Enabling Flexible On-Body IMU Placement for Human Pose Estimation and Activity Recognition. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology*. 1–16.
- Chengxu Zuo, Jiawei Huang, Xiao Jiang, Yuan Yao, Xiangren Shi, Rui Cao, Xinyu Yi, Feng Xu, Shihui Guo, and Yipeng Qin. 2025. Transformer IMU calibrator: dynamic on-body IMU calibration for inertial motion capture. *ACM Transactions on Graphics (TOG)* 44, 4 (2025), 1–14.

Vision-Aided IMU Calibrator: Robust On-body IMU Calibration from a Single Video

Supplementary Material

A Frame Definitions

Many inertial pose estimators assume noise-free IMU orientations that are already calibrated and expressed in a shared world frame W [Armani et al. 2024; Hollidt et al. 2026; Jiang et al. 2022b; Yi et al. 2022, 2024]. In practice, each IMU measures the orientation of its sensor frame S relative to its own internal reference frame REF . These measurements do not directly provide the world-space orientation of an associated SMPL bone frame B . An IMU cannot directly measure the orientation of a bone frame B ; at best, it measures the orientation of the surface-attached vertex frame V at its mounting location, up to a fixed sensor-to-vertex rotation. We further define a surface tangent frame T at each vertex, whose z -axis aligns with the local surface normal and which constrains the mounting rotation to an in-plane yaw. Calibration estimates the alignment between REF and W and the fixed relation between S and the candidate body-surface frame. Table 5 summarizes the coordinate frames used throughout the paper.

The relative rotation between frames A and B satisfies ${}_A R_B = {}_B R_A^T$. We use this inverse relationship throughout the paper.

B Vertex Frame

In our formulation, we utilize vertex frames instead of bone orientations. This allows for a more precise formulation and better calibration. Sparse inertial pose estimators commonly model an IMU measurement as the orientation of its associated bone. However, an IMU directly measures only the orientation of its own local frame and, at best, the orientation of the skin surface at its attachment location. Moreover, a bone-based representation assigns a single orientation to every vertex influenced by that bone, whereas surface motion varies continuously, particularly near joints. For example, when the forearm twists by 65° about its longitudinal axis while the elbow remains fixed, vertices near the wrist rotate substantially more than those near the elbow; assigning one forearm orientation to all such vertices can result in errors of up to 30° . Vertex frames capture this spatially varying surface motion and therefore provide a more accurate representation for IMU calibration. Using the SMPL model, we can assign each vertex a frame that is consistent. We derive the vertex frame orientations using the standard SMPL formulation described by Loper et al. [Loper et al. 2015]. Our implementation is identical to the standard SMPL model regarding the computation of shape offsets $B_S(\vec{\beta})$, pose offsets $B_P(\vec{\theta})$, joint locations $J(\vec{\beta})$, and the kinematic hierarchy of coordinate frames.

Specifically, for a given pose $\vec{\theta}$ and shape $\vec{\beta}$, we compute the global world transformation $G_k \in SE(3)$ for every joint k , exactly as

Table 5. Coordinate frames used by VAIC. Each IMU has its own sensor frame S and internal reference frame REF . Mesh-derived vertex frames V and surface tangent frames T describe candidate body-mounted sensor locations.

Frame	Definition
W	Shared world frame in which all calibrated IMU orientations are expressed. Sparse inertial posing requires measurements from all sensors to be represented in a common world frame.
REF	Internal reference frame of an individual IMU. This frame may differ across sensors. IMU state estimators commonly initialize it as an East–North–Up (ENU) frame, but magnetic disturbances and inaccurate gravity estimates can cause unreliable initialization.
S	Sensor frame fixed to the physical IMU. The IMU reports its orientation relative to its internal reference frame REF .
V	Vertex frame rigidly attached to mesh vertex v , as defined in Sec. B. Its motion is obtained from the recovered SMPL mesh and represents the kinematics of an IMU ideally mounted at vertex v .
T	Surface tangent frame associated with vertex v . It is rigidly related to V and is oriented such that its z -axis aligns with the local mesh surface normal.
B	SMPL bone frame associated with bone B . Unlike V , which is attached to a specific surface vertex, a bone frame describes the orientation of an underlying skeletal segment. Many inertial posers assume that the calibrated IMU orientation directly corresponds to B .

defined in the SMPL kinematic tree. The deviation from the standard pipeline occurs only at the Linear Blend Skinning (LBS) stage.

In the standard SMPL model, LBS is used to compute a vertex position $\mathbf{t}'_i$. The model defines a transformation matrix $\mathbf{T}_i \in \mathbb{R}^{4 \times 4}$ for vertex i as the linear combination of joint transformations weighted by skinning weights $w_{k,i}$:

$$\mathbf{T}_i = \sum_{k=1}^K w_{k,i} \mathbf{G}_k(\vec{\theta}, J(\vec{\beta})) \quad (9)$$

While the standard model computes the vertex position via $\mathbf{t}'_i = \mathbf{T}_i \bar{\mathbf{t}}_i$ (where $\bar{\mathbf{t}}_i$ is the vertex in the rest pose), we isolate the rotational component of $\mathbf{T}_i$ to define the vertex orientation. Let $\bar{\mathbf{R}}_i$ be the upper-left 3×3 sub-matrix of $\mathbf{T}_i$:

$$\tilde{\mathbf{R}}_i = \sum_{k=1}^K w_{k,i} \mathbf{R}_k \quad (10)$$

where $\mathbf{R}_k$ is the rotation matrix of joint k in the world frame. Because $\tilde{\mathbf{R}}_i$ is a linear interpolation of rotation matrices, it is generally not orthogonal (containing scaling and shear) and does not belong to $SO(3)$. To obtain a valid rigid rotation frame ${}_W R_V$ for the IMU alignment, we project $\tilde{\mathbf{R}}_i$ onto the closest rotation matrix using Singular Value Decomposition (SVD):

$$\mathbf{U}, \Sigma, \mathbf{V}^T = \text{SVD}(\tilde{\mathbf{R}}_i) \quad (11)$$

$${}_W R_V = \mathbf{U} \mathbf{V}^T \quad (12)$$

This results in the orthonormal orientation ${}_W R_V$ of the vertex V in the world frame W .

B.1 Implementation Details

We implement VAIC using PyTorch. The pairwise calibration is formulated as a non-linear least squares problem and solved using a custom, batched Levenberg-Marquardt optimizer executed on the GPU. We approximate the Jacobian matrix using forward finite differences with a perturbation step size of $\epsilon = 10^{-5}$. To ensure broad coverage, we select 372 target vertices distributed uniformly across the SMPL mesh. The optimization converges in approximately 5 seconds for a 10-second input sequence and 10 seconds for a 20-second sequence using an RTX 4090. We set the weighting terms to $\lambda_{\text{norm}} = 0.1$ and $\lambda_{\text{acc}} = 5 \cdot 10^{-6}$, ensuring their residuals scale to approximately one-tenth of the orientation error magnitude. These hyperparameters were empirically validated on a subset of 50 calibration sequences.

B.2 Motion Filtering

VAIC requires sufficient motion during calibration, since static poses can lead to calibration ambiguities. In TotalCapture, some segments are static (e.g., holding a yoga pose) or capture the user leaving the camera view, so we automatically remove such segments using simple heuristics. Specifically, we exclude segments where video-based mesh recovery fails to detect a person or where motion is negligible. To detect low-motion segments, we compute the standard deviation of the root translation and SMPL joint rotations over time, discarding sequences whose ground-plane root translation standard deviation is too small or whose mean joint angle standard deviation indicates near-static motion. In practice, we require a ground-plane root translation standard deviation $\sigma_{\text{trans}} > 0.1\text{m}$ and a mean joint angle standard deviation $\sigma_{\text{pose}} > 0.1\text{rad}$.

B.3 Optimization Details

VAIC itself is lightweight once SMPL motion is available: for a 10s calibration video on an RTX4090, IMU synthesis takes 1.7s, our optimization 5.3s, and Hungarian matching 0.1s, for a total of 7.1s. We use a batched Levenberg-Marquardt GPU solver, solving each vertex-sensor optimization independently. It converges reliably across all pairs (final step size < 0.003) within 50 steps from random initialization. Initializing with a closed-form linear solution

for the orientation energy term ($\mathbf{A}\mathbf{X}=\mathbf{X}\mathbf{B}$) or increasing the number of optimization steps does not improve accuracy.

The main bottleneck in the end-to-end pipeline is monocular HMR. In the paper, we use PromptHMR as the HMR backend, but VAIC only requires an SMPL motion sequence, so faster HMR backends with comparable reconstruction quality can reduce runtime. With PromptHMR [Wang et al. 2025], GVHMR [Shen et al. 2024] or FastHMR [Mehraban et al. 2026] the total wall-clock runtime is approximately 126s/88s/33s, respectively, for the same 10s sequence. Reducing this runtime further would enable near-real-time operation with minimal waiting time for the user.

B.4 Joint Uncertainty

VAIC assigns equal weight to residuals across all timestamps during optimization. In principle, timestamps with unreliable joint estimates could be downweighted or excluded if the corresponding uncertainty is known. Such cases may occur under strong occlusions or when the user partially leaves the camera field of view.

We experimented with uncertainty-aware weighting strategies based on PromptHMR’s predicted joint confidence scores and joint reprojection errors. Residuals associated with low-confidence or high-reprojection-error joints were reduced or removed from the optimization objective. However, we did not observe a significant improvement in calibration accuracy. This suggests that PromptHMR already provides sufficiently robust pose estimates under occlusions and partial visibility, making additional uncertainty weighting largely unnecessary in our setting.

C Validation Details

On TotalCapture, we recover meshes independently from six camera views, excluding two views with persistent blur or human detection failures. Ground-truth sensor-to-bone orientations ${}_V R_S^{gt}$ are computed from GT-SMPL poses obtained via optical motion capture [Huang et al. 2018; Mahmood et al. 2019]. SMPL ground truth is minimally smoothed to reduce marker jitter, and segments containing artifacts or insufficient motion are excluded using the motion filtering strategy described in Sec. B.2, leaving 478 ten-second calibration sequences. For standard T-Pose calibration, we use frames 10–30 of the initial T-Pose segment provided in each TotalCapture sequence. Monocular pose estimates are aligned with ground truth only for metric evaluation due to the unknown absolute heading.

The DS_{VAIC} dataset was recorded with 8 participants using Xsens IMUs and an OptiTrack optical motion capture system for ground-truth motion capture. Ground-truth SMPL poses are obtained through MoSh++ [Loper et al. 2014; Mahmood et al. 2019], while ${}_V R_S^{gt}$ and ${}_W R_{REF}^{gt}$ are measured using the motion capture system. Video-based mesh recovery is performed from a monocular iPhone video. Each participant recorded 3 sequences, leading to 20 sequences of roughly 4 minutes each after filtering failed recordings.

Participants received detailed instructions for each calibration procedure. Each recording session was captured continuously in the following order: N-Pose, N-Pose→T-Pose, T-Pose, walking, and 40 s of dynamic motions. The N-Pose→T-Pose baseline recovers heading from the arm rotation axis, while the N-Pose+walking baseline estimates heading from the walking direction, closely matching the

commercial Xsens calibration procedure. TIC represents a recent learning-based approach and can be used on any motion; we report the online calibration results. To avoid evaluation bias toward motions used during calibration, a separate validation block was recorded consisting of 40 s of dynamic motions, N-Pose, T-Pose, and turning in a circle. In total, this resulted in 20 sequences of approximately four minutes each.

Compared to TotalCapture, DS_{VAIC} exhibits stronger magnetic disturbances, leading to less stable IMU reference frame initialization. Since the IMUs initialize their reference frame based on the local ENU frame, we frequently observe substantial heading offsets, which some calibration baselines cannot recover from.

Throughout our experiments, IMU, SMPL, and video streams are time-synchronized at recording time, which is realistic in practice: commercial inertial mocap software such as Xsens MVN supports synchronized IMU and video capture out of the box, and any single-device recorder (e.g. a smartphone capturing both video and connected IMU streams) yields device-side synchronization by construction.

C.1 Metrics

Bone Accuracy: To account for the high density of joints in the torso, we treat the spine and sternum bone sets as a single target class.

D Modeling Sensor Drift

In our main method, we assume the calibration parameters are static (see Eq. 1). However, in long-term capture scenarios, inertial systems are subject to two primary sources of drift: global reference drift (e.g., due to magnetic interference) and local attachment drift (e.g., soft tissue artifacts or sensor slippage).

We can extend our formulation to explicitly model these time-varying effects. Let REF and S denote the *true* reference and sensor frames, defined by the initial reference alignment and the true physical sensor attachment, respectively. Consequently, the transformations ${}_WR_{REF}$ and ${}_SR_V$ remain static.

To account for time-varying effects (i.e., measurement errors), we model the IMU measurements as occurring in time-varying coordinate frames. Let REF' and S' denote the *drifting* frames representing the instantaneous coordinate systems actually used by the sensor hardware.

The drifting measurement chain is given by:

$${}_WR_V(t) = {}_WR_{REF} \cdot {}_{REF}R_{REF'}(t) \cdot {}_{REF'}R_{S'}(t) \cdot {}_{S'}R_S(t) \cdot {}_SR_V \quad (13)$$

Here, the terms are defined as follows:

- ${}_{REF'}R_{S'}(t)$: The raw, time-varying orientation measurement provided by the IMU.
- ${}_{REF}R_{REF'}(t)$: The drift of the sensor's internal reference frame (e.g., "North" drifting away from true North).
- ${}_{S'}R_S(t)$: The physical displacement or slippage of the sensor relative to its initial calibrated mounting orientation.
- ${}_WR_{REF}$ and ${}_SR_V$: The static base calibration parameters recovered by our method.

This formulation generalizes the problem. By solving for the time-varying correction matrices ${}_{REF}R_{REF'}(t)$ and ${}_{S'}R_S(t)$, our static model converts into a dynamic one. This aligns with the approach

Table 6. Downstream sparse inertial mocap accuracy (SMPL joint-angle error, in degrees) when running PIP with different calibration methods. All experiments use the standard 6-IMU configuration and pretrained PIP weights.

Dataset	GT-Calib	VAIC-20s	T-Pose	N-Pose→T-Pose	T-Pose+Walk
TotalCapture	12.46	15.47	17.72	–	–
DSVAIC	20.94	21.57	31.83	22.90	26.46

Table 7. VAIC calibrates with lower angle error than T-Pose. In this setting, we assume IMU positions to be known. SMPL GT indicates the maximal feasible calibration results by replacing the camera-generated SMPL with GT SMPL. Metrics are reported as mean $\pm$ standard deviation. All 13 IMUs with known positions are evaluated on TotalCapture.

Experiment	Seq.	AE ${}_WR_V$ ($^\circ$)	AE ${}_VR_S$ ($^\circ$)	AE ${}_WR_{REF}$ ($^\circ$)
VAIC	10 s	15.3 ± 5.8	14.0 ± 7.3	11.4 ± 6.3
VAIC	20 s	14.5 ± 5.5	13.7 ± 7.6	10.3 ± 6.3
VAIC	30 s	14.4 ± 5.2	13.8 ± 7.0	9.5 ± 5.1
T-Pose	–	16.3 ± 9.4	18.5 ± 10.1	8.6 ± 6.5
VAIC w/ SMPL GT	10s	7.9 ± 3.7	10.0 ± 5.9	6.5 ± 5.0

taken by Zuo et al. in TIC [Zuo et al. 2025], where the authors estimate these drift components (denoted as $R_{G'G}$ and R_{BS} respectively) by assuming they remain negligible within short time windows while varying over long sequences.

E Ethics

The DS_{VAIC} collection and user study followed our institution's ethical guidelines: all participants were volunteers, provided informed consent, could withdraw at any time, and were compensated.

F Experimental Results

F.1 Downstream Sparse Inertial Mocap Accuracy

To evaluate whether VAIC improves downstream sparse inertial motion capture, we run PIP [Yi et al. 2022] using calibrations obtained from each method, following the standard 6-IMU setup and protocol used in PIP. We use the publicly released pretrained weights of PIP without any retraining or adaptation, and only change the IMU-to-body calibration. All methods calibrate with respect to the vertex frame.

We report SMPL joint-angle error using the standard evaluation protocol defined in PIP in Table 6.

VAIC reduces downstream pose estimation error compared to traditional calibration. On TotalCapture, the angle error improves from 17.72° (T-Pose) to 15.47° (VAIC-20s). On DS_{VAIC} , VAIC-20s achieves 21.57° , nearly matching ground-truth calibration (20.94°), and outperforming all traditional procedures (31.83° for T-Pose, 22.90° for N-Pose-T-Pose, and 26.46° for T-Pose+Walk). This demonstrates that VAIC's calibration gains translate directly into better sparse inertial mocap performance, not only improved calibration metrics.

F.2 Error Distribution

VAIC consistently outperforms all baselines in terms of mean calibration error. Notably, its median error is even lower, reaching 13.8°

when IMU positions are known, highlighting strong performance in the vast majority of cases. The difference between mean and median reflects a small number of challenging outliers, resulting in a long-tailed error distribution (Fig. 5). These rare cases primarily occur in static and highly symmetric calibration sequences. Overall, this demonstrates that VAIC delivers highly accurate calibration results across most calibration settings. The same trend is observed for all reported error metrics (Figs. 6, 7, and 8). In most sequences, IMUs are localized correctly on the body; the few remaining ambiguities typically involve neighboring body parts (Fig. 9).

Errors are generally higher for wrist-mounted IMUs. Wrists exhibit fast motion and a large rotational range of motion, making monocular mesh reconstruction and inertial measurements more susceptible to noise and estimation errors. These inaccuracies can introduce outlier correspondences that hinder convergence toward the correct calibration parameters. Future work could improve robustness through RANSAC-style optimization.

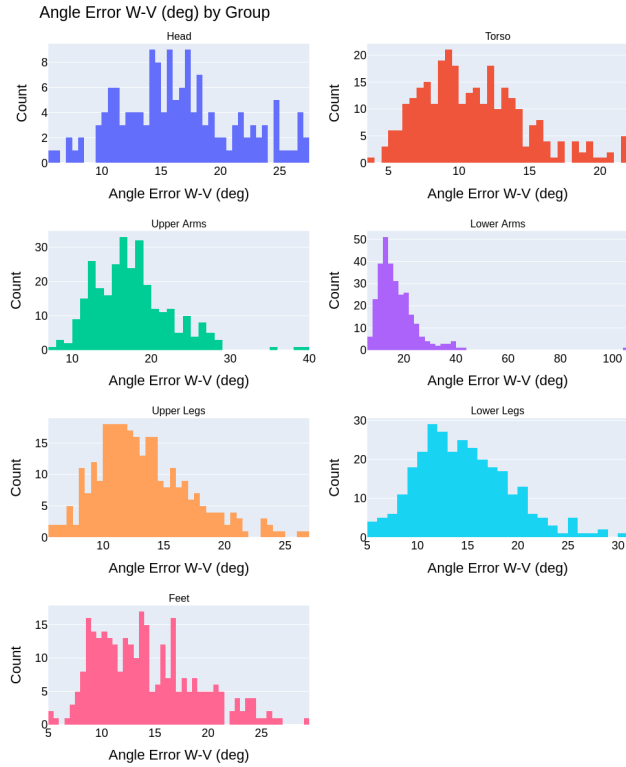


Fig. 5. Histograms of VAIC's wR_V calibration angle errors on 30s calibration with fixed IMU placements.

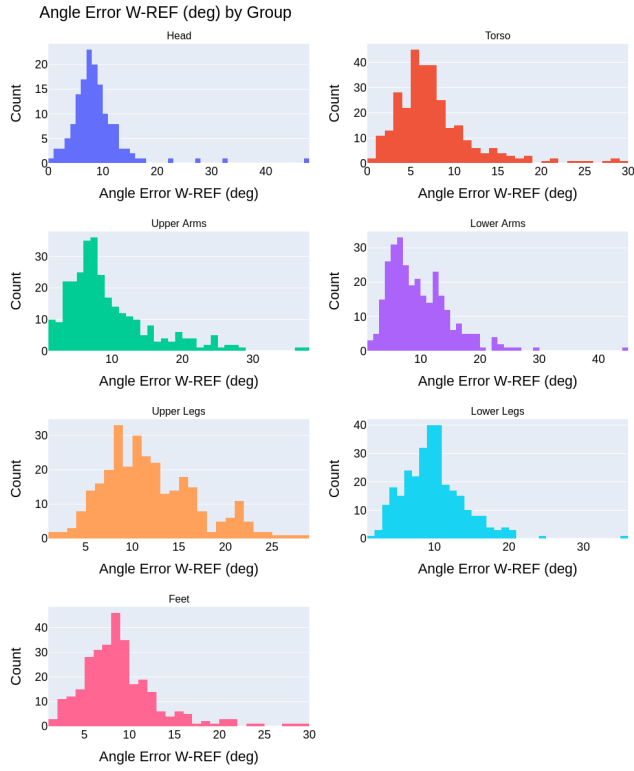


Fig. 7. Histograms of VAIC's w_{REF} calibration angle errors on 30s calibration with fixed IMU placements.

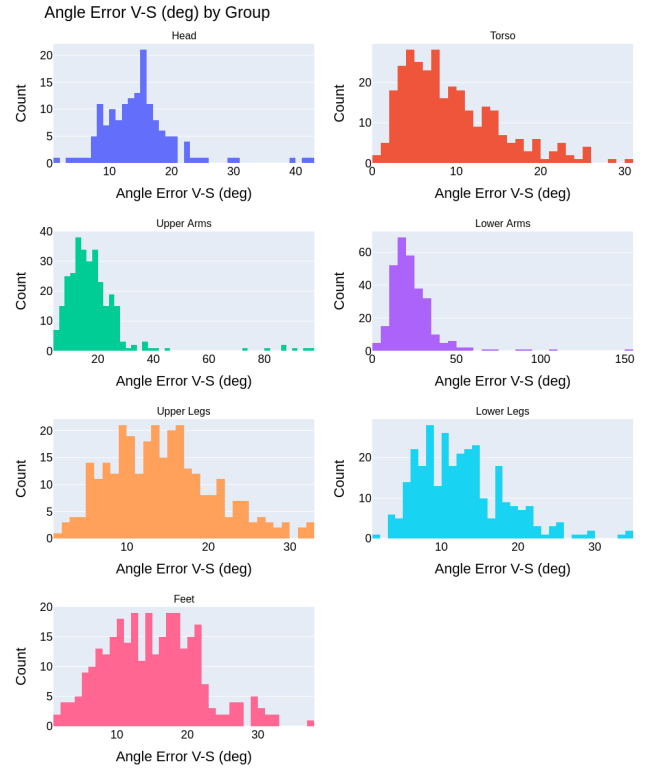


Fig. 6. Histograms of VAIC's v_{RS} calibration angle errors on 30s calibration with fixed IMU placements.

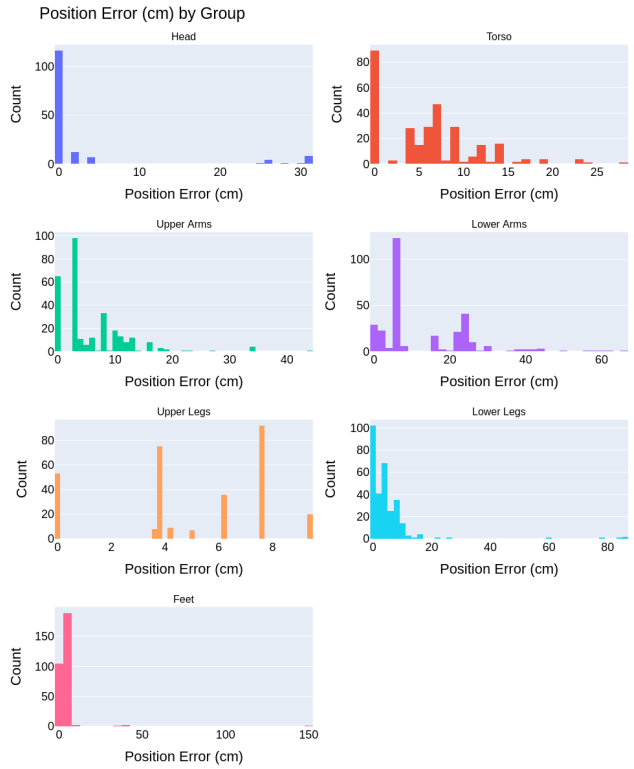


Fig. 8. Histograms of VAIC's IMU localization errors on 20s calibration with arbitrary IMU placements.

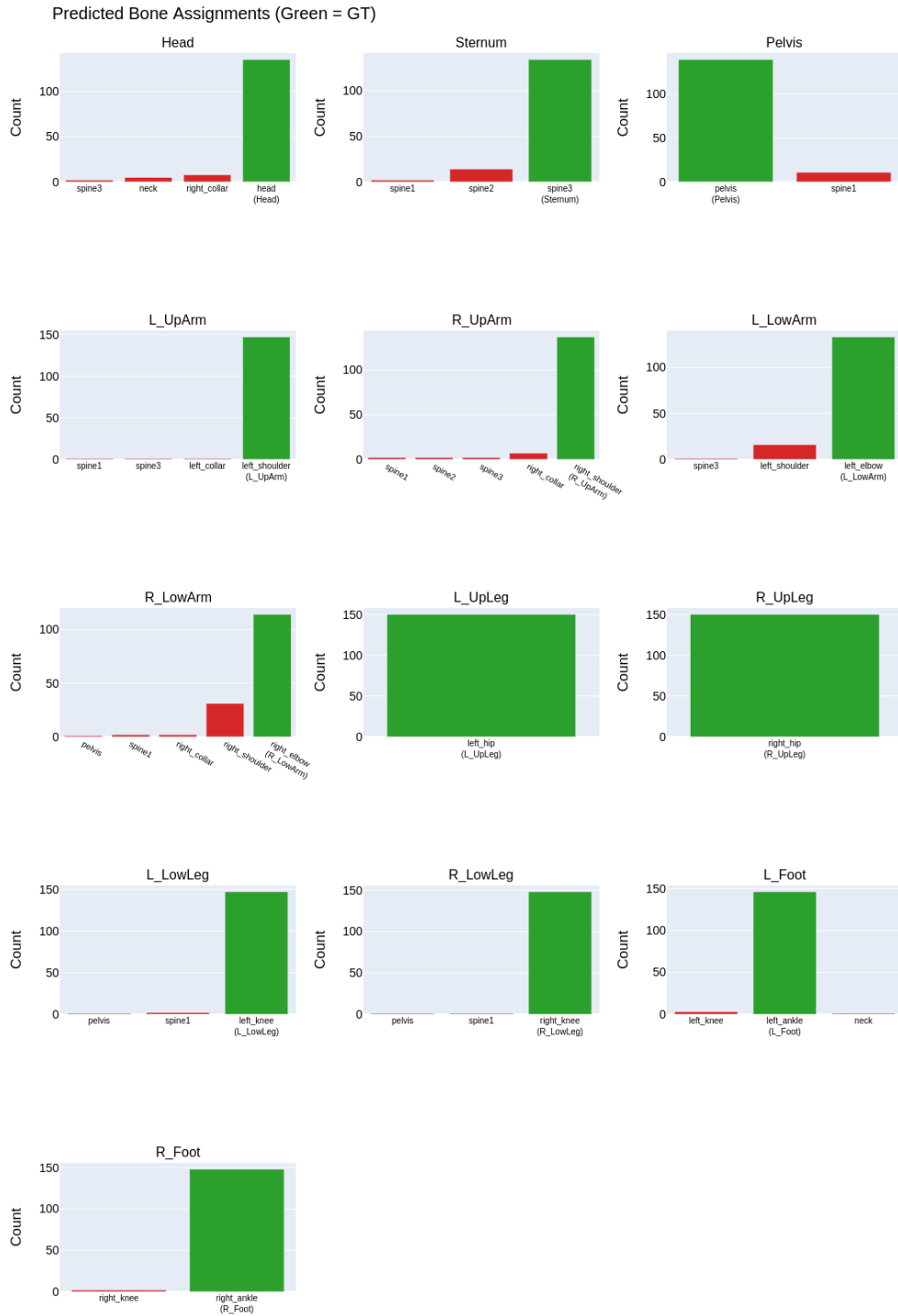


Fig. 9. Assignment of IMUs to bones. VAIC localizes IMUs with 95.9% accuracy. Occasionally, they are placed on neighbouring bones.