MISP: MAMBA-BASED INERTIAL SHAPE AND POSE ESTIMATION WITH ULTRA-WIDEBAND RANGING AND SKELETAL PRIORS

Anonymous authors
Paper under double-blind review

ABSTRACT

Sparse inertial pose estimation (SIP) recovers full-body motion from body-worn inertial measurement units (IMUs), offering an occlusion-robust and portable solution for motion capture. Yet under sparse sensor layouts(e.g., 6 IMUs), the problem remains severely under-constrained and suffers from IMU drift. Recent work adds new sensing modalities (e.g., ultra-wideband [UWB] ranging) to mitigate IMU drift. However, prior work is constrained by two main challenges: (1) lack of effective means to acquire and exploit subject-specific body constraints, degrading cross-subject generalization and leaving ranging cues scale-ambiguous; (2) how to effectively exploit heterogeneous observations in temporal modeling. To address these challenges, we propose: (1) **MISP**, a sparse inertial framework that fuses both UWB-measured inter-sensor distances and subject-specific bone-length priors into the inertial pipeline; to obtain the skeletal prior, MISP-Shaper independently estimates explicit bone-length constraints from a short natural-motion window; (2) **Continuous Unified Fusion (CUF)**, which repeatedly re-injects observation features across temporal blocks to preserve inertial and geometric cues; and Together, these components enable MISP to reduce joint position error by up to 12% over prior work.

1 INTRODUCTION

Human motion capture underpins film and animation production, VR/AR interaction, and sports and health monitoring, where reliable full-body kinematics are essential. Marker-based optical systems Vicon Motion Systems once served as the gold standard for their precision. Markerless video-based approaches Bogo et al. (2016); Kanazawa et al. (2018); Kocabas et al. (2020); Hsu & Jang (2024) trade accuracy for convenience. Both remain sensitive to environment, lighting, and occlusion. In contrast, inertial motion capture offers an alternative with minimal requirements on environment and lighting and inherent robustness to occlusion.

Compared with dense inertial suits with 17–19 sensors Roetenberg et al. (2007); Noitom Technology, Sparse inertial pose estimation (SIP) chooses wearability and convenience over sensor redundancy. von Marcard et al. (2017); Huang et al. (2018); Yi et al. (2021) recovers full-body joint angles from a sparse layout of only six sensors placed at the pelvis, head, wrists, and lower legs. This provides an occlusion-free, low-intrusiveness solution with minimal environmental requirements and camera-free privacy. Consumer 6-IMU systems (e.g., Sony mocopi Sony Group Corporation, SlimeVR SlimeVR Contributors) have already moved such capture beyond the lab Xiao et al. (2024).

With advances in causal temporal models, SIP accuracy has steadily improved, yet it remains limited by the under-constrained nature of sparse layouts and by drift induced by inertial sensor noise—reconstructing 24-joint full-body motion from six or fewer sensors is mathematically ambiguous. To add constraints without adding sensors, recent work introduces hybrid sensing with new observation types, e.g., UWB-measured inter-sensor distances Armani et al. (2024); Liu et al. (2025), which provide drift-free geometric anchors that mitigate inertial drift. Yet these methods rarely exploit bone lengths, the first-order geometric constraint in FK detailed in Sec. 3.4: without explicitly acquiring

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.

subject-specific bone lengths, models must implicitly entangle shape with motion, which degrades cross-subject generalization without per-subject calibration.

To overcome this limitation, we propose Mamba Inertial Shape and Pose Estimator (MISP), which consists of two key components: a subject-specific shape estimator for personalized skeletal prediction and a pose estimator for full-body motion recovery. For shape estimation, we decouple a dedicated subject-specific shape estimator that predicts explicit bone lengths beyond abstract SMPL shape coefficients from a short natural-motion IMU window, providing a session-level static skeletal prior. For pose estimation, MISP fuses inertial observations, UWB-measured inter-sensor distances, and the static skeletal prior into a unified constraint embedding and models time with a Mamba-2 backbone, whose selectivity arbitrates history against current evidence while substantially easing the frame-by-frame throughput bottleneck of recurrent posers. To let deep layers fully analyze current observations, we introduce the Continuous Unified Fusion (CUF) module attached to every causal temporal block, which re-injects depth-adapted projections of the initial mixed observation features via concatenation, so that deeper layers can repeatedly access sensor evidence and actively analyze incoming signals, thereby improving pose accuracy.

Our main contributions are:

1. We present MISP, a sparse inertial pose estimation framework that closes the acquire-and-exploit loop for hybrid sensing: it jointly mediates inertial observations and drift-free UWB distances with subject-specific skeletal constraints for accurate full-body recovery, instantiated with an efficient Mamba-2 backbone; a decoupled shape estimator acquires the session-level bone-length and shape prior from a short natural-motion IMU window.
2. We introduce the Continuous Unified Fusion (CUF) module attached to each causal temporal block, which repeatedly re-injects depth-adapted projections of the initial mixed observation features, keeping physical and geometric cues active throughout deep sequence modeling.

2 Related Work

2.1 Sparse Inertial Pose Estimation

Compared with vision-based approaches, body-worn inertial measurement units offer a self-contained, occlusion-robust sensing modality. While dense commercial systems Roetenberg et al. (2007); Noitom Technology reduce pose ambiguity through sensor redundancy, sparse inertial pose estimation (SIP) recovers full-body motion from only six body-worn IMUs von Marcard et al. (2017), which is inherently under-constrained. Learning-based approaches have since evolved from RNNs to Transformers to achieve real-time inference Huang et al. (2018); Yi et al. (2021); Jiang et al. (2022); Wu et al. (2024), yet they do not incorporate explicit physical constraints such as subject-specific bone lengths. Physically motivated systems improve reconstruction by modeling explicit physical laws. Specifically, PIP Yi et al. (2022) introduces a dynamics optimizer, while PNP Yi et al. (2024) further incorporates virtual physical forces. Though highly interpretable, these multi-stage designs never re-expose observations to deeper layers within the same stage. Some methods introduce additional constraints as inputs, yet each faces distinct limitations: UIP Armani et al. (2024) and UMotion Liu et al. (2025) treat UWB ranges as one-shot input features, while UDP Hollidt et al. (2026) decodes them into spatial positions before diffusion-based pose recovery. SAIP Yin et al. (2025) considers body shape, yet uses it only for acceleration alignment, leaving deeper layers unable to directly access the shape prior; TOF-IP Yao et al. (2025) adds Time-of-Flight sensing but imposes stricter requirements on the environment. MIP Han et al. (2026) first introduces Mamba to SIP, yet it segments sequences into fixed windows, limiting long-term temporal context; we instead enable continuous, unbounded-context modeling.

2.2 Cross-Layer Observation Reuse in Deep Temporal Models

Deep networks commonly preserve early signals along depth via cross-layer shortcuts, but additive and concatenative designs differ in what they carry forward. ResNets He et al. (2016) propagate already-mixed hidden states through additive identities, while DenseNets Huang et al. (2017) concatenate feature maps so the input signal directly reaches every layer, explicitly avoiding summation

before a layer. A parallel input-to-every-layer precedent injects the current input into all layers: densely connected RNNs re-expose the current embedding to every recurrent layer Godin et al. (2017), and Trellis Networks inject a shared linear transform of the input into every temporal convolutional layer Bai et al. (2019). Conditioning-based designs instead modulate each layer of the backbone, e.g., FiLM-style affine feature modulation Perez et al. (2018); in contrast, CUF concatenates depth-adapted projections as an independently addressable channel and never scales the hidden representation itself. Importantly, these systematic designs are mostly studied in deep stacks, leaving reuse in shallower stacks underexplored; as an exception, the densely connected RNN study explicitly targets stacking only a few layers, showing that direct input access already matters at shallow depth.

SIP typically adopts shallow stages with only a few serial temporal layers. Single-stack designs such as DIP Huang et al. (2018) and UMotion Liu et al. (2025) connect temporal blocks directly in series without additional jumps, injecting observations only at the bottom. Multi-stage designs such as TransPose Yi et al. (2021), PIP Yi et al. (2022), and PNP Yi et al. (2024) re-introduce the current observation across stages, yet layers inside each stage remain directly serial without intra-stage re-exposure. How to strengthen observations in such shallow SIP stacks has not been studied.

2.3 Body Shape Estimation

Body shape estimation—spanning surface mesh reconstruction and skeletal bone length recovery—is a cornerstone of motion capture, long dominated by vision-based methods leveraging appearance cues Kanazawa et al. (2018); Kocabas et al. (2020); Hsu & Jang (2024).

Sparse IMU systems cannot rely on visual appearance, yet subject-specific proportions remain critical for accurate pose inference. DivaTrack Yang et al. (2024) recovers body shape from sparse trackers but requires dedicated calibration motions. A separate line predicts proportions from static anthropometrics alone without motion Bartol et al. (2022); Choutas et al. (2022); Liu et al. (2025), resolving only population-average proportions, whereas our shaper fuses static anthropometrics with dynamic IMU evidence for per-subject bone lengths. Unlike SAIP Yin et al. (2025) and UMotion Liu et al. (2025), which predict only abstract SMPL coefficients and entangle shape with pose dynamics, we predict explicit bone lengths decoupled from pose inference.

3 Methodology

3.1 Problem Formulation: SIP as Sequential IK under Hybrid Observations

We address sparse inertial pose estimation (SIP): recovering full-body 3D human motion from six body-worn wearable sensors. Following the standard 6-sensor protocol, sensors are placed at the pelvis, head, left/right wrists, and left/right lower legs. We split all constraints by time scale into dynamic tracking and static calibration.

At frame t, each sensor node carries a 6-DoF IMU and a UWB transceiver. The *dynamic observation* $\mathbf{O}_t = (\mathbf{A}_t, \mathbf{R}_t, \mathbf{d}_t)$ varies per frame, where $\mathbf{A}_t \in \mathbb{R}^{s\times 3}$ is linear acceleration, $\mathbf{R}_t \in \mathrm{SO}(3)^s$ is orientation expressed in the root-sensor frame, and $\mathbf{d}_t \in \mathbb{R}^{s(s-1)/2}$ collects all pairwise inter-sensor distances from UWB ranging, with $s = 6$. The *static prior* $\mathbf{b} \in \mathbb{R}^7$ is session-level invariant for a subject, comprising shoulder width, trunk and hip width, and the lengths of upper-arm, forearm, thigh, and calf with left/right averaged. The Shaper internally predicts a laterality-aware $\tilde{\mathbf{b}} \in \mathbb{R}^{11}$ distinguishing left/right limbs, symmetrized by averaging each limb pair into $\mathbf{b}$ for pose conditioning. Here $\mathbf{b}$ denotes an estimated skeletal prior whose estimator is detailed in Sec. 3.4; in this section it is treated as given.

The target is the SMPL Loper et al. (2023) local pose in the continuous 6D representation Zhou et al. (2020) with pelvis as root, plus global root translation $\mathbf{p}_t \in \mathbb{R}^3$. Given our sensor placement, we do not observe the movement of the hands and feet. Therefore, similar to previous studies Huang et al. (2018); Yi et al. (2021; 2022), we estimate poses for only 16 joints, including the root joint, while excluding the hand and foot joints. For simplicity, we denote this reduced set of local pose parameters as $\mathbf{\Theta}_t \in \mathbb{R}^{16\times 6}$, while $\mathbf{\Theta}_t^{24}$ refers to the full set of local pose parameters for 24 joints. Notations with a hat symbol, e.g., $\hat{\mathbf{\Theta}}_t$, indicate the corresponding estimated values. Following Yi et al. (2024); Yin et al. (2025), our network focuses on local pose; $\mathbf{p}_t$ is recovered by a shape-aware

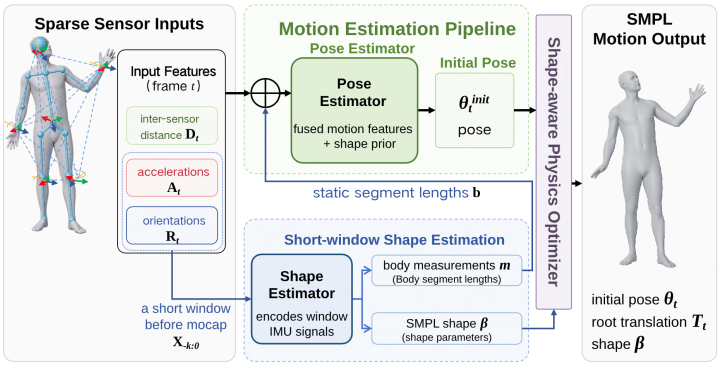

Figure 1: Overview of the MISP pipeline. Sparse inertial inputs, optional inter-sensor distances, and a decoupled shape prior are fused by a Mamba-2 pose estimator, followed by shape-aware physics optimization for final motion output.

physics optimizer. Directly mapping $(\mathbf{A}_t, \mathbf{R}_t) \mapsto (\mathbf{\Theta}_t, \mathbf{p}_t)$ is severely under-constrained: identical inertial signals can arise from different motions or body proportions, and are further corrupted by drift and noise. Learned SIP therefore relies on two complementary remedies: (i) causal temporal models that combine motion history with current observations to reduce per-frame ambiguity, and (ii) hybrid sensing that thickens instantaneous evidence while retaining the 6-node sparse layout. For (ii), following prior hybrid work Armani et al. (2024); Liu et al. (2025), we include UWB-measured inter-sensor distances $\mathbf{d}_t$ as instantaneous geometric constraints; unlike integrated inertial quantities, ranging is drift-free and hence anchors error accumulation, though noisy under occlusion. Moreover, because bone lengths shape both inertial and ranging observations, we further condition on session-level bone lengths $\mathbf{b}$ as static geometric input; its full kinematic justification is deferred to Sec. 3.4.

We formulate local pose estimation as causal sequential inverse kinematics in state space. Let $\mathbf{h}_t$ be the recurrent temporal state summarizing motion history $\mathbf{O}_{\leq t}$. Then:

$$\mathbf{h}_t = \mathcal{T}(\mathbf{h}_{t-1}, \mathbf{O}_t; \mathbf{b}), \qquad \hat{\mathbf{\Theta}}_t = \mathcal{G}(\mathbf{h}_t), \tag{1}$$

where $\mathcal{T}$ mediates history against current evidence and $\mathcal{G}$ is the per-frame IK decoding. Selectivity in $\mathcal{T}$—when to retain history versus trust current evidence—is the core difficulty under IMU drift and UWB noise. This leaves open how to implement $\mathcal{T}$ to arbitrate history against enriched current evidence, and how to keep the fused evidence accessible along depth. Throughout Secs. 3.2–3.3 we assume $\mathbf{b}$ is given (oracle); Sec. 3.4 removes this assumption and estimates $\mathbf{b}$ from a short natural-motion window.

As shown in Figure 1, MISP combines a static Shaper (Sec. 3.4), a dynamic CUF-Mamba pose estimator (Secs. 3.2–3.3), and a shape-aware physics optimizer (Sec. 3.5).

3.2 Mamba-2 Temporal Backbone

SIP couples two difficulties: intrinsic observation deficiency from 6-to-16 mapping, and corrupted observations from IMU noise/drift and occluded UWB ranging. This makes $\mathcal{T}$ in Eq. (1) non-trivial: it must embed $\mathbf{O}_{\leq t}$ into $\mathbf{h}_t$ for $\mathcal{G}$ to decode $\hat{\mathbf{\Theta}}_t$ without letting drifted history overwhelm fresh evidence. We therefore implement $\mathcal{T}$ with Mamba-2 Dao & Gu (2024), whose input-dependent selectivity Gu & Dao (2024) dynamically weights historical motion memory against current observations to attenuate long-horizon error accumulation Yi et al. (2022), while near-linear $O(N)$ complexity supports both high-throughput offline parallel scans and low-latency online causal inference within a single trained model.

3.3 Continuous Unified Fusion

SIP must draw on motion history to resolve ambiguity, so causal temporal models are a natural backbone. But even inside a shallow stage, current-frame observation features are easily diluted

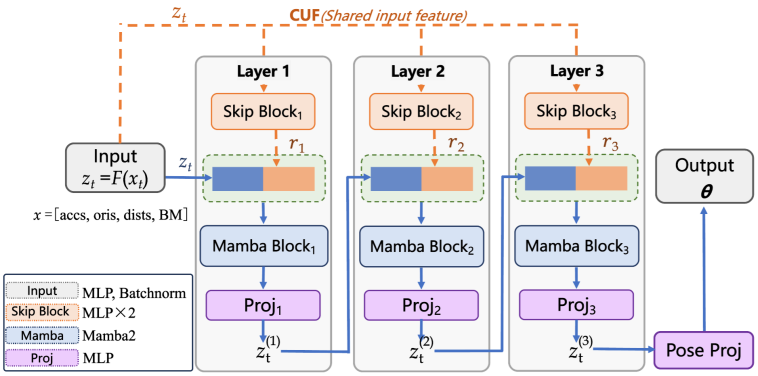

Figure 2: Detailed structure of the pose estimator in MISP. The continuous unified fusion module continuously re-injects constraint cues into each temporal layer.

by accumulated history: with single-point injection, current observations fade as hidden states accumulate history, while prevailing SIP stacks confine observations to stage inputs—injected at the module input and never re-exposed to deeper layers within the same stage Huang et al. (2018); Yi et al. (2024). The resulting estimates tend to be limited in accuracy yet overly smooth.

We refer to this depth-wise loss of direct evidence access as *observation dilution*: the progressive loss of direct access to the current-frame observation along depth, distinct from over-smoothing where representations become mutually similar. Following the input-to-every-layer precedent Bai et al. (2019), we persistently strengthen the current observation inside the temporal stack.

SIP is a sparse-to-dense inverse-kinematics problem: six sensors map 16-joint local pose parameters, so inputs and outputs live in spaces of different types and dimensions. We therefore avoid additive residual shortcuts: direct addition would force heterogeneous representations into a shared subspace and perturb the backbone's internal information flow. Instead, CUF concatenates evidence alongside the hidden representation, leaving the main information flow undisturbed. To further match the abstraction level of each layer, it encodes the same current-frame observation through a layer-specific two-layer MLP (Layer-wise Encoder), aligning current evidence with different representation spaces along depth.

Structurally, CUF instantiates this idea as per-block evidence re-injection. We first fuse heterogeneous per-frame inputs $\mathbf{x}_t = [\mathbf{A}_t, \mathbf{R}_t, \mathbf{d}_t, \mathbf{b}]$ into a unified embedding $\mathbf{z}_t = \phi_{\text{in}}(\mathbf{x}_t)$, with channel-wise batch normalization across physical quantities. Let $\mathbf{h}_t^{(l)}$ denote the hidden state produced by block l and consumed by block $l+1$, with $\mathbf{h}_t^{(0)} = \mathbf{z}_t$. It projects $\mathbf{z}_t$ through the Encoder $\psi^{(l)}(\cdot)$ and concatenates the result alongside the hidden representation entering the l-th block:

$$\mathbf{r}_t^{(l)} = \psi^{(l)}(\mathbf{z}_t), \quad \mathbf{h}_t^{(l)} = \text{MambaBlock}^{(l)}([\mathbf{h}_t^{(l-1)}; \mathbf{r}_t^{(l)}]), \tag{2}$$

where $[\,;\,]$ denotes channel-wise concatenation. Because each block is causal, $\mathbf{h}_t^{(l)}$ mixes accumulated history with the current observation rather than preserving the observation itself. Since $\mathbf{r}_t^{(l)}$ depends on $\mathbf{z}_t$ alone, per-layer selectivity can arbitrate history against current evidence at every timestep. We stack three such blocks (Fig. 2).

While CUF addresses *how* to persistently exploit observations along depth, we next address *what* static observation should be exploited.

3.4 Decoupled Shape Estimator

We now remove the oracle assumption and ask how to obtain $\mathbf{b}$ without dedicated calibration motions. We predict explicit bone lengths $\mathbf{b}$, the forward-kinematics-relevant subspace of shape, rather than abstract shape coefficients. This distinction is decisive for SIP: identical joint angles generate different sensor trajectories under different skeletons, and identical inertial readings can correspond to different poses across body sizes, because both IMU lever-arm accelerations and UWB inter-sensor distances are direct functions of these segments. Bone lengths are thus the fundamental geo-

metric properties in human kinematics, while abstract $\boldsymbol{\beta}$ is an entangled PCA coefficient that cannot directly constrain joint reach. Unlike approaches that predict only abstract $\boldsymbol{\beta}$, we therefore take explicit bone lengths $\mathbf{b} \in \mathbb{R}^7$ as the primary target for kinematics, with $\boldsymbol{\beta}$ predicted only alongside for avatar surface reconstruction.

Since $\mathbf{b}$ is session-level static while pose is dynamic, we predict it independently from a short natural-motion window and freeze it, breaking the circular pose–shape dependency in SAIP Yin et al. (2025) where shape estimated from already-estimated poses is interlocked with window-level pose quality by construction. Existing sparse-IMU posers are explicitly shape-agnostic under a fixed template body Yi et al. (2021); Huang et al. (2018), so estimating subject-specific scale is necessary for cross-subject generalization without per-subject calibration motions Yang et al. (2024). The frozen prior then conditions the per-frame likelihood in Eq. (1): under IMU drift and UWB occlusion noise, the selective state in $\mathcal{T}(\mathbf{h}_{t-1}, \mathbf{O}_t; \mathbf{b})$ can arbitrate history against current heterogeneous observations only when anchored to a stable skeletal scale, instead of re-estimating a drifting shape at every frame.

We instantiate this with a lightweight mirror of the pose backbone taking per-frame $\mathbf{A}_t, \mathbf{R}_t$ conditioned on anthropometrics; each frame contributes a noisy observation of the same static $\mathbf{b}$, aggregated over a short window into a frozen prior (see Appendix A.2 for NLL supervision, recalibration, and top-k aggregation details).

3.5 Overall Pipeline

In inference order, the Shaper first freezes $\hat{\mathbf{b}}$ from a 1–3 s window, which the CUF-Mamba stack then conditions on alongside per-frame hybrid observations to decode $\hat{\boldsymbol{\Theta}}_t$; finally a shape-aware physics optimizer recovers $\mathbf{p}_t$ following the shape-aware physics optimization protocol.

4 Experiments

Table 1: Comparison with state-of-the-art methods on TotalCapture and DIP-IMU. TIP-D, PIP-D, UIP, UDP, UMotion, and MISP augmented with simulated ideal inter-sensor distances. Mesh Err is omitted for space.

	TotalCapture				DIP-IMU			
Method	SIP↓	Ang Err↓	Pos Err↓	Jitter↓	SIP↓	Ang Err↓	Pos Err↓	Jitter↓
DIP	18.62	17.22	9.42	–	17.35	15.36	7.59	–
TIP	13.22	12.30	5.81	1.69	16.90	9.07	5.63	1.56
PIP	12.93	12.04	5.61	0.20	15.33	8.78	5.12	0.17
PNP	10.89	10.45	4.74	0.26	13.71	8.75	4.97	0.17
SAIP	11.22	10.96	4.75	0.42	–	–	–	–
MIP	–	–	–	–	13.43	8.36	4.62	0.16
Mocap-derived idealized geometric cues								
TIP-D	10.76	–	5.14	0.72	15.91	–	5.44	0.86
PIP-D	11.41	–	5.55	0.20	13.76	–	5.34	0.281
UIP	11.32	–	5.49	0.257	13.79	–	5.36	0.297
UDP	**8.95**	<u>10.19</u>	3.76	**0.124**	**10.39**	8.19	3.42	<u>0.125</u>
UMotion	10.76	11.36	4.46	<u>0.16</u>	14.19	<u>6.35</u>	<u>3.38</u>	**0.11**
UMotion+CUF	9.60	10.42	<u>3.48</u>	0.19	<u>11.88</u>	6.75	4.59	0.14
MISP(ours)	<u>9.04</u>	**9.75**	**3.44**	0.28	12.14	**5.78**	**2.87**	0.14

4.1 Data.

Following previous work Liu et al. (2025), we train MISP on AMASS Mahmood et al. (2019), fine-tune it on the training splits of UIP Armani et al. (2024) and DIP-IMU Huang et al. (2018), and evaluate it on TotalCapture Trumble et al. (2017), DIP-IMU, and UIP. Following prior synthesis protocols Yi et al. (2022), IMU orientations and accelerations are synthesized from the SMPL Loper et al. (2023) model; for datasets without physical UWB measurements, pairwise inter-sensor distances are computed from mocap-derived sensor trajectories as simulated geometric cues, while for

UIP we use real IMU and UWB data calibrated by RANSAC regression as in Armani et al. (2024). See Appendix A.3 for the full synthesis protocol and Appendix A.1 for training details. All maintable numbers follow the official uncorrected protocol for fair comparison with prior work; a temporal IMU–MoCap misalignment identified in one TotalCapture sequence is documented separately in Appendix A.4.

4.2 Metrics.

Following prior SIP work Yi et al. (2024; 2022); Jiang et al. (2022); Yin et al. (2025), we report SIP error (mean global orientation error of shoulders and hips, deg), angular error (mean global rotation error over the 16 estimated joints, deg), positional error (mean joint position error, cm), mesh error (mean SMPL vertex error, cm), and jitter (mean joint jerk, 10^3 m/s^3); for body shape we additionally use T-pose mesh error (Mesh-T, cm) and bone length error (BLE, cm). All metrics are lower-is-better.

4.3 Main Results

4.3.1 Pose Estimation on TotalCapture, DIP-IMU and UIP

We evaluate MISP against representative sparse IMU estimators (DIP Huang et al. (2018), TIP Jiang et al. (2022), PIP Yi et al. (2022), PNP Yi et al. (2024), SAIP Yin et al. (2025), MIP Han et al. (2026), ASIP Wu et al. (2024)) and distance-augmented methods (UIP Armani et al. (2024), UMotion Liu et al. (2025), UDP Hollidt et al. (2026)).

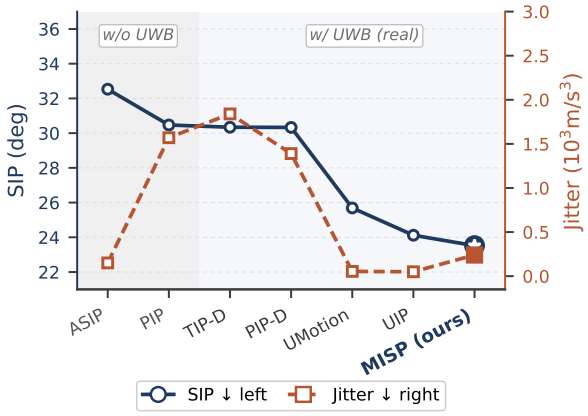

Figure 3: SIP and jitter on UIP with real UWB measurements. Solid (left axis): SIP; dashed (right axis): jitter. Lower is better for both; MISP achieves the lowest SIP.

Evaluation on TotalCapture and DIP-IMU. Table 1 reports the performance on TotalCapture and DIP-IMU using simulated distance cues. MISP reaches state-of-the-art accuracy: full-body angular error is reduced by around 4% on TotalCapture and 9% on DIP-IMU over the strongest distance-augmented baselines, with positional error also the lowest; SIP trails only the diffusion-based UDP on TotalCapture and stays competitive on DIP-IMU, and jitter stays at an acceptable level. We additionally note a 43-frame IMU–MoCap misalignment in TotalCapture `s5_freestyle3` inherited from the widely used preprocessed release; after alignment-only correction, MISP further reduces angular error by 3% with jitter unchanged (Appendix A.4).

Real UWB on UIP and robustness to missing ranges. Fig. 3 compares the methods on the UIP dataset using real, processed UWB measurements. After fine-tuning on training set of UIP, MISP achieves the best SIP and positional accuracy, outperforming the previous best distance-augmented models by around 15% in position with the best SIP as well, and jitter kept at an acceptable level. To simulate real-world UWB occlusions, randomly dropping distance cues on TotalCapture degrades accuracy by only about 1% even at 30% dropout (Appendix A.5), confirming strong robustness to intermittent ranging that corroborates the real-noise advantage on UIP.

4.3.2 Standalone Shape Accuracy

The shape estimator is trained on AMASS with bone-length supervision derived from SMPL joint distances (see Appendix A.3). Table 2 (Appendix A.3) reports standalone accuracy via Bone Length

Error (BLE: median-across-windows MAE of the raw 11-segment predictions $\tilde{\mathbf{b}}$, in cm): the estimator reaches 0.33 cm BLE on DIP-IMU and 0.58 cm on TotalCapture without dedicated calibration motions, and accuracy improves stably with observation time (Appendix A.2, Fig. 5). We use BLE as the primary metric, as bone lengths directly govern kinematic constraints; the poser conditions on the symmetrized 7-dim $\hat{\mathbf{b}}$. For reference, static height–weight linear regression reports 5.7–13.1 mm MAE on surface body lengths Bartol et al. (2022); our IMU-window estimator reaches 3.3–5.8 mm BLE on skeletal bone lengths, albeit under a different caliber (surface vs. skeletal lengths, synthetic vs. mocap populations).

4.4 Ablation Studies

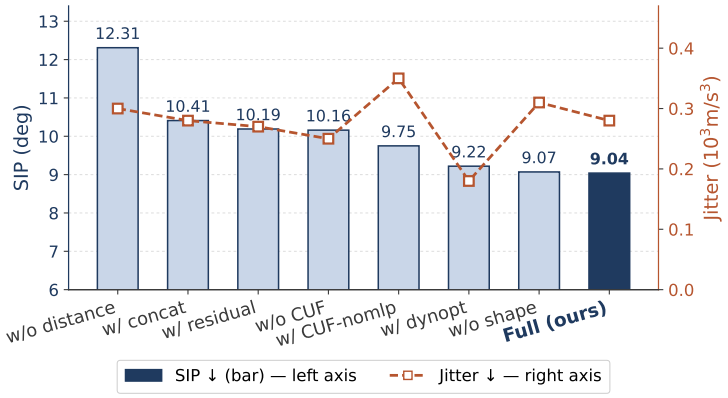

Figure 4: Component ablation on TotalCapture. Bars (left axis): SIP; dashed line (right axis): jitter. Lower is better; full MISP achieves the lowest SIP. See text for variant definitions.

We ablate components on TotalCapture (Fig. 4; `w/o shape`/`w/o distance` remove the corresponding inputs, `w/ dynopt` is detailed in Appendix A.5, and the skeletal-prior swap is studied in Sec. 4.4.3 and Appendix A.3, Tab. 3). Fig. 4 shows a paired accuracy-smoothness phenomenon consistent with observation dilution in Sec. 3.3: strengthening current observations improves accuracy at a small smoothness cost.

4.4.1 Component Analysis

Effect of Continuous Unified Fusion (`w/o CUF`). As shown in Fig. 4, removing CUF raises SIP by 1.12° and Pos Err by 0.54 cm while lowering jitter by 0.03, confirming that CUF genuinely strengthens the current observation rather than diluting it.

Effect of the Shaper (`w/o shape`). Fig. 4 shows that removing the Shaper's skeletal prior barely moves SIP (+0.03°) but raises Pos Err by 0.18 cm, indicating that the Shaper mainly disambiguates positions.

Effect of UWB distances (`w/o distance`). Removing inter-sensor distances raises SIP by 3.27°—the largest degradation among the components—confirming that drift-free ranging is the strongest geometric constraint.

Effect of the physics optimizer (`w/ dynopt`). `w/ dynopt` appends the PNP physics optimizer Yi et al. (2024) that recovers global translation $\mathbf{p}_t$; it lowers jitter by 0.10 at a modest SIP cost of +0.18°, as detailed in Appendix A.5.

4.4.2 CUF Mechanism Analysis

As shown in Fig. 4, the three re-exposure variants—a standard residual connection (`w/ residual` He et al. (2015)), a concat-shortcut (`w/ concat-shortcut`), and direct re-injection

without the Layer-wise Encoder (`w/ CUF-nomlp` Bai et al. (2019))—all degrade SIP by 0.71–1.37°, confirming that CUF yields a substantial and consistent gain on MISP. Among them, `w/ CUF-nomlp` additionally exhibits the largest jitter increase (+0.07), supporting our hypothesis that unaligned re-injection perturbs the backbone's internal information flow and quantifying the value of the per-layer Layer-wise Encoder $\psi^{(l)}$. Finally, UMotion+CUF (Tab. 1) lowers TotalCapture SIP by 10.8% and DIP-IMU SIP by 16.3% with only a modest jitter increase, indicating that CUF benefits not only the Mamba-2 backbone but may generalize broadly to SIP temporal models.

4.4.3 Shape Ablation

To assess downstream value, three skeletal priors were fed to the pose estimator (Appendix A.3, Tab. 3): average SMPL bone lengths raise SIP error from 9.04 to 11.41, confirming that the pose estimator actively leverages the shape prior, while GT bone lengths hold a modest 0.38° advantage over our inferred bone lengths (8.66 vs. 9.04 SIP), indicating headroom remains but our shaper already approaches oracle quality and clearly benefits the no-shape case. For aggregation, calibrated top-k selection slightly outperforms RANSAC (BLE Top-k 0.33/0.58 cm vs. RANSAC 0.35/0.60 cm on DIP-IMU/TotalCapture, $\Delta = -0.02$; Appendix Tab. 2), confirming that the learned $\log \sigma_t$ provides effective per-frame ranking after recalibration.

4.5 Efficiency

The neural backbone is free from frame-by-frame serial recurrence: offline full-sequence parallel scan reaches over twenty times the throughput of FIP, and online frame-wise inference more than doubles it, with substantially lower offline memory (see Appendix A.7 for the full table and protocol, both measured on local poser modules). The end-to-end bottleneck instead lies in global translation recovery with the physics optimizer: once included, all models fall back to just over 200 FPS online (MISP 220, PNP 205).

4.6 Qualitative Evaluation

The qualitative analysis focuses on the middle-to-later segments of representative TotalCapture sequences, where long-range temporal modeling matters most. As shown in Appendix Fig. 9, MISP maintains consistently higher pose accuracy than PNP in these segments, indicating that the Mamba-2 temporal backbone is better suited for long-range sequence modeling and more resistant to error accumulation over time. Visually, MISP produces limb articulations that are more extended and anatomically precise. These observations confirm that our architecture is effective in resolving fine-grained motion details. Appendix Fig. 10 reports cumulative global-translation error as travelled distance increases under the shared PIP-style dynamics-based optimization protocol Yi et al. (2022): translation drift accumulates with distance for all methods, consistent with our finding that global displacement—recovered by the physics optimizer rather than the neural backbone—remains the end-to-end bottleneck (Sec. 5). The supplementary video further contrasts jitter-minimal ASIP and physics-optimized PNP: ASIP shows visible jitter from its non-causal sliding windows, while PNP landings look stiffened by contact-driven optimization; see Appendix A.8 for analysis.

5 Limitations

As discussed in Sec. 4.4.2, CUF's accuracy gain comes at a moderate jitter cost, which is a perceptually important metric for motion capture; appending a physics optimizer (`w/ dynopt`) partially offsets this cost. Optimal performance still requires accurate UWB ranging signals as dynamic geometric constraints, which increases the hardware complexity. In addition, global translation is recovered by a separate physics optimizer, bounding the full-pipeline throughput.

AI Use Statement

In this work, generative AI tools were used for language polishing and LaTeX formatting assistance (grammar refinement and template formatting). Generative AI was not used for developing research ideas, designing methods or experiments, writing code, or generating experimental results and fig-

ures. All AI-assisted text was reviewed and verified by the authors, who take full responsibility for the final content of this work.

Ethics Statement

This work uses public motion-capture datasets (AMASS, DIP-IMU, TotalCapture, UIP) and does not involve new human-subjects data collection. The method is intended for benign applications such as animation, rehabilitation, and human–computer interaction. We do not foresee direct harmful uses beyond the general dual-use considerations of motion-capture technology.

Reproducibility Statement

We describe the model architecture (Section 3), training data and evaluation metrics (Section 4), with full training details in Appendix A.1 and data synthesis details in Appendix A.3. The TotalCapture alignment correction (Appendix A.4) is implemented as a standalone preprocessing script with fixed lag (43 frames for `s5_freestyle3`). Upon acceptance, we will release code and trained model weights to facilitate reproduction.

References

Rayan Armani, Changlin Qian, Jiaxi Jiang, and Christian Holz. Ultra inertial poser: Scalable motion capture and tracking from sparse inertial sensors and ultra-wideband ranging. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers '24 (SIGGRAPH Conference Papers '24)*, 2024.

Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. Trellis networks for sequence modeling. In *International Conference on Learning Representations*, 2019.

Kristijan Bartol, David Bojanić, Tomislav Petković, Stanislav Peharec, and Tomislav Pribanić. Linear regression vs. deep learning: A simple yet effective baseline for human body measurement. *Sensors*, 22(5):1885, 2022. doi: 10.3390/s22051885.

Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J. Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *Computer Vision – ECCV 2016 (Lecture Notes in Computer Science)*, pp. 561–578, 2016.

Vasileios Choutas, Lea Muller, Chun-Hao P. Huang, Siyu Tang, Dimitrios Tzionas, and Michael J. Black. Accurate 3d body shape regression using metric and semantic attributes, 2022. URL `https://arxiv.org/abs/2206.07036`.

Tri Dao and Albert Gu. Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, pp. 10041–10071, 2024.

Frédéric Godin, Joni Dambre, and Wesley De Neve. Improving language modeling using densely connected recurrent neural networks. In *Proceedings of the 2nd Workshop on Representation Learning for NLP (RepL4NLP)*, pp. 186–190, 2017.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *First Conference on Language Modeling (COLM)*, 2024.

Shijia Han, Jianfei Yang, Xinyu Zhou, Chuhao Zhou, and Fang Lu. Mamba inertial poser: Dual-temporal causal human motion tracking from sparse inertial sensors. SSRN Electronic Journal, 2026. URL `https://doi.org/10.2139/ssrn.6168348`. Preprint.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. URL `https://arxiv.org/abs/1512.03385`.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

Dominik Hollidt, Tommaso Bendinelli, and Christian Holz. 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*, pp. 7036–7046, 2026.

Chih-Hsiang Hsu and Jyh-Shing Roger Jang. Blapose: Enhancing 3d human pose estimation with bone length adjustment, 2024. URL https://arxiv.org/abs/2410.20731.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

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, 2018. URL https://arxiv.org/abs/1810.04703.

Yifeng Jiang, Yuting Ye, Deepak Gopinath, Jungdam Won, Alexander W. Winkler, and C. Karen Liu. Transformer inertial poser: Real-time human motion reconstruction from sparse imus with simultaneous terrain generation. In *SIGGRAPH Asia 2022 Conference Papers*, 2022.

Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7122–7131, 2018.

Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. Vibe: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5253–5263, 2020.

Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. Accurate uncertainties for deep learning using calibrated regression. In *International conference on machine learning*, pp. 2796–2804. PMLR, 2018.

Huakun Liu, Hiroki Ota, Xin Wei, Yutaro Hirao, Monica Perusquía-Hernández, Hideaki Uchiyama, and Kiyoshi Kiyokawa. Umotion: Uncertainty-driven human motion estimation from inertial and ultra-wideband units. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7085–7094, 2025. doi: 10.1109/CVPR52734.2025.00664.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pp. 851–866. 2023.

Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5442–5451, 2019.

Noitom Technology. Noitom — world leader in motion capture technology. https://www.noitom.com/. Accessed: 2026-06-24.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Daniel Roetenberg, Henk Luinge, and Per Slycke. Moven: Full 6dof human motion tracking using miniature inertial sensors. *Xsen Technologies, December*, 2007.

SlimeVR Contributors. Slimevr: Open-source full-body tracking with imus. https://docs.slimevr.dev/. Accessed: 2026-09-23.

Sony Group Corporation. mocopi — 6-point full-body motion capture system. https://www.sony.com/en/mocopi/. Accessed: 2026-09-23.

Matt Trumble, Andrew Gilbert, Charles Malleson, Adrian Hilton, and John Collomosse. Total capture: 3d human pose estimation fusing video and inertial sensors. In *2017 British Machine Vision Conference (BMVC)*, 2017.

Vicon Motion Systems. Award winning motion capture systems. https://www.vicon.com/. Accessed: 2026-06-24.

Timo von Marcard, Bodo Rosenhahn, Michael J. Black, and Gerard Pons-Moll. Sparse inertial poser: Automatic 3d human pose estimation from sparse imus. *Computer Graphics Forum*, 36 (2), 2017.

Yinghao Wu, Chaoran Wang, Lu Yin, Shihui Guo, and Yipeng Qin. Accurate and steady inertial pose estimation through sequence structure learning and modulation. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9798331314385.

Xuan Xiao et al. Fast human motion reconstruction from sparse inertial measurement units considering the human shape. *Nature Communications*, 2024.

Dongseok Yang, Jiho Kang, Lingni Ma, Joseph Greer, Yuting Ye, and Sung-Hee Lee. Divatrack: Diverse bodies and motions from acceleration-enhanced three-point trackers. *Computer Graphics Forum*, 43(2):e15057, 2024.

Yuan Yao, Shifan Jiang, Yangqing Hou, Chengxu Zuo, Xinrui Chen, Shihui Guo, and Yipeng Qin. Tof-ip: Time-of-flight enhanced sparse inertial poser for real-time human motion capture. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Xinyu Yi, Yuxiao Zhou, and Feng Xu. TransPose: Real-time 3D Human Translation and Pose Estimation with Six Inertial Sensors, May 2021. URL http://arxiv.org/abs/2105.04605. arXiv:2105.04605 [cs].

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*, 2022.

Xinyu Yi, Yuxiao Zhou, and Feng Xu. Physical non-inertial poser (pnp): Modeling non-inertial effects in sparse-inertial human motion capture. In *SIGGRAPH 2024 Conference Papers*, 2024.

Lu Yin, Ziying Shi, Yinghao Wu, Xinyu Yi, Feng Xu, and Shihui Guo. Shape-aware inertial poser: Motion tracking for humans with diverse shapes using sparse inertial sensors. *ACM Transactions on Graphics*, 44(6):221:1–221:12, 2025. doi: 10.1145/3763311.

Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks, 2020. URL https://arxiv.org/abs/1812.07035.

A Additional Implementation Details

A.1 Training Details

All experiments are conducted on a single NVIDIA RTX 5090 GPU. The pose estimator is trained with input sequences of 224 frames, a batch size of 480, for 270 epochs ($\sim$4 GPU hours). We adopt the AdamW optimizer with a learning rate of 1.6×10^{-4} and weight decay of 0.01. The input projection layer maps the 94-dimensional concatenated observation to a 320-dimensional feature space via a linear layer. The Mamba-2 blocks use a hidden dimension of 640 as the internal state dimension. The shape estimator shares the same optimizer setting and lightweight Mamba-2 backbone, supervised by a Gaussian NLL loss on bone lengths and SMPL shape parameters $\boldsymbol{\beta}$. It is trained with a window length of 128 frames and tested with a window length of 64 frames.

A.2 Uncertainty Recalibration

To correct the temporal calibration drift in the raw uncertainties Kuleshov et al. (2018), we introduce an empirical recalibration formulation:

$$\sigma_t^{\text{cal}} = \exp(\log \sigma_t + \alpha \cdot t^{\beta}), \tag{3}$$

where α and β are derived via curve fitting on the training set of DIP-IMU Huang et al. (2018), yielding $\alpha = -0.03$ and $\beta = 0.55$. The final estimate $\hat{\mathbf{b}}$ is computed as the mean of the $k = 4$ frames with the lowest calibrated uncertainty.

Figure 5 validates this recalibration under cold start ($T = 180$, $N = 1234$ chunks, $\sim$60 Hz). Cold start means each chunk is inferred from a zero hidden state without pre-filling initial motion; the first 2 frames are a warmup transient (dotted line) that is discarded from selection. The left axis shows bone-length MAE (BM error, cm) with mean, median, and p25–p75 band; the right axis shows raw $\exp(\text{logstd})$ versus the calibrated score. Mean BM error decreases stably with observation time and the band narrows, confirming sequential evidence accumulation. Raw uncertainty follows the early drop but then plateaus and overestimates late uncertainty, while the calibrated score $-0.03\,t^{0.55}$ keeps decreasing and tracks the mean error, so selecting lowest calibrated scores amounts to selecting truly low-error moments. Quantitatively, per-chunk temporal correlation between uncertainty and true error improves from raw to calibrated: Spearman mean 0.17 / median 0.20 with 61% positive chunks, to Spearman mean 0.37 / median 0.50 with 76% positive chunks (Pearson mean 0.46 / median 0.61).

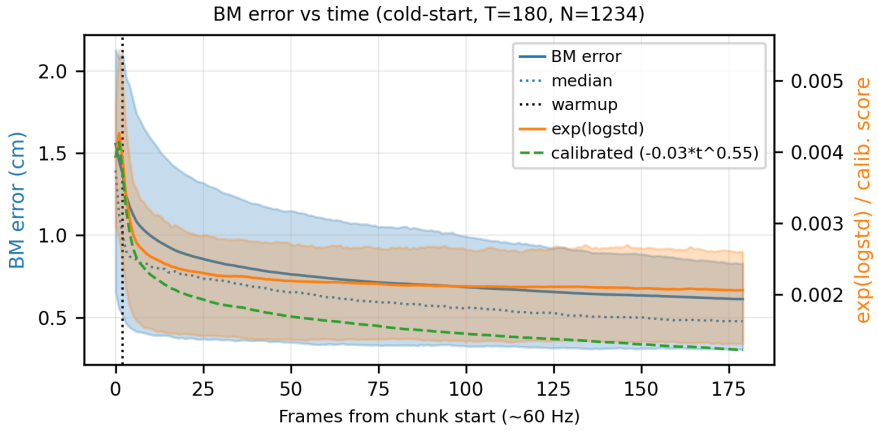

Figure 5: BM error vs. time under cold start ($T = 180$, $N = 1234$). Left axis: BM error (cm) mean/median with p25–p75 band. Right axis: raw $\exp(\text{logstd})$ vs. calibrated score. Cold start uses no initial-motion pre-filling; the dotted warmup line marks the first 2 frames discarded from selection.

A.3 Data Synthesis and Dataset Details

IMU and distance synthesis. Following the synthesis protocol in Sec. 4, IMU orientations and accelerations are synthesized from the SMPL model. For datasets without physical UWB measurements, we compute pairwise tag distances from mocap-derived sensor trajectories as simulated inter-sensor cues. For evaluations on UIP, which includes both real IMU and UWB data, we perform RANSAC regression to calibrate UWB measurements as described in Sec. 4.

Shape supervision. The shape estimator is trained on AMASS Mahmood et al. (2019). Bone-length supervision is derived from SMPL joint distances. Subject height ranges from 150 to 194 cm. Body weight is computed from the T-pose SMPL mesh volume multiplied by 985 kg/m^3 Choutas et al. (2022).

Standalone shape-estimation protocol. To evaluate standalone shape accuracy, we divide all test sequences into 64-frame windows (matching the test window length). For each window, we predict body shape using the method described in Sec. 3.4 and compare against ground truth under three metrics: Bone Length Error (BLE), the median across windows of the mean absolute error of the predicted 11-segment bone lengths; $\boldsymbol{\beta}$ error, the median across windows of the MAE of the SMPL shape coefficients; and Mesh-T error, the median across windows of the mean vertex error of the reconstructed SMPL mesh in the T-pose. At test time the shaper runs one-shot: a single 3-second

window at the start of each sequence (covering 2–5 s) is inferred once and reused for the whole sequence.

Table 2: Standalone shape-estimation accuracy under SMPL-defined supervision. All entries are medians across motion windows. BLE denotes bone length error derived from SMPL model, in cm. BLE Top-k uses top-k point selection; BLE RANSAC uses RANSAC point selection; Δ = RANSAC − Top-k.

Dataset	BLE Top-k↓	BLE RANSAC↓	Δ	β↓	Mesh-T↓
DIP-IMU	0.33	0.35	-0.02	0.34	0.58
TotalCapture	0.58	0.60	-0.02	0.58	1.09

Table 3: Pose error in TotalCapture under different skeletal priors. Average refers to the bone length of the standard smpl shape.

Shape prior	SIP	Jitter
Average	11.41	0.28
Predicted bone	9.04	0.28
GT bone	8.66	0.28

A.4 TotalCapture Temporal Alignment Correction

Motivation. When evaluating on the test set prepared following prior work, MISP showed disproportionately large error on `s5_freestyle3` while performing normally on all other sequences. This isolated failure prompted a data-level inspection rather than model tuning, which revealed a systematic temporal lag between the IMU stream and the MoCap-derived ground truth. We therefore diagnose and correct this lag below.

Diagnosis. In TotalCapture sequence `s5_freestyle3`, IMU-measured orientations and MoCap-derived sensor orientations exhibit a systematic temporal offset. Following the standard 6-sensor protocol, we compare the 6-sensor-averaged rotation magnitude from a common reference for (i) raw IMU orientations and (ii) MoCap-synthesized orientations over frames 200–1200. Cross-correlation peaks at a lag of 43 frames (0.717 s at 60 FPS): the MoCap stream lags the IMU stream. Figure 6 shows that advancing the MoCap stream by 43 frames aligns sharp turning events and the subsequent oscillatory segments, whereas the uncorrected MoCap curve is right-shifted throughout.

Correction protocol. The fix is alignment-only at the data level and is applied identically before inference for all methods: advance the MoCap-derived ground truth of `s5_freestyle3` by 43 frames; no joint values are altered, no frames are dropped as difficult cases, and no method-specific tuning is performed. Advancing leaves the trailing 43 frames without MoCap correspondence; since the subject ends in static standing, we pad them by repeating the last available MoCap frame. All other sequences are untouched. Main-table results retain the official uncorrected protocol to ensure fair comparison with prior work.

Effect. Table 4 reports TotalCapture SIP before/after correction. The correction improves all tested methods consistently, confirming a dataset-level systematic bias rather than method-specific overfitting: MISP $9.04^\circ \rightarrow 8.74^\circ$ with shaper-predicted bones ($8.66^\circ \rightarrow 8.32^\circ$ with GT bones), PNP $10.89^\circ \rightarrow 10.38^\circ$, UMotion $10.76^\circ \rightarrow 10.10^\circ$. The ranking is unchanged and the margin of MISP over the strongest baseline is preserved. For MISP with shaper-predicted bones, full metrics after correction are: SIP 8.74° (9.04° before), Pos Err 3.28 cm (3.44 cm before), Mesh Err 3.79 cm, Jitter 0.28 (unchanged), indicating that the misaligned sequence mainly biases positional/angular accuracy rather than smoothness.

A.5 Additional Ablation Details

We also evaluate the effect of appending the physics optimizer from PNP Yi et al. (2024) (`w/ dynopt`). While it slightly improves motion smoothness (jitter drops from 0.28 to 0.18), it intro-

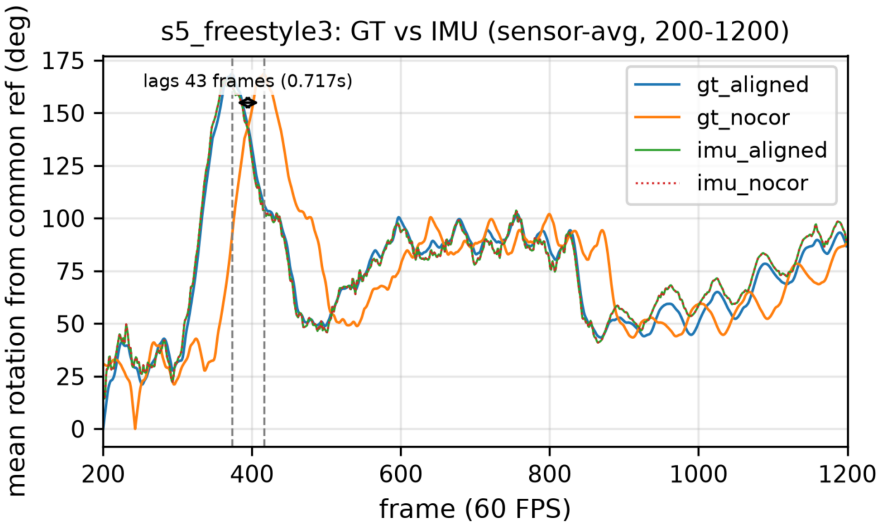

Figure 6: IMU vs. MoCap orientation alignment on TotalCapture `s5_freestyle3` (sensor-averaged rotation from a common reference, frames 200–1200 at 60 FPS). `gt_nocor` (orange) is systematically right-shifted by 43 frames relative to IMU; after advancing MoCap by 43 frames, `gt_aligned` (blue) coincides with IMU (green/red) on sharp peaks and oscillations.

Table 4: TotalCapture SIP (deg) before/after advancing `s5_freestyle3` MoCap by 43 frames. All methods improve consistently; ranking is unchanged. BL denotes the bone-length source (GT vs. shaper-predicted).

Method	BL	Official	Corrected	Δ
PNP	–	10.89	10.38	−0.51
UMotion	–	10.76	10.10	−0.66
MISP (ours)	GT	8.66	8.32	−0.34
MISP (ours)	Pred	9.04	8.74	−0.30

duces a modest accuracy degradation (SIP rises from 9.04 to 9.22). Nevertheless, this smoothness gain can partially compensate for the jitter increase introduced by CUF, suggesting that the two components are complementary.

Depth robustness of CUF. Figure 7 studies DIP-IMU SIP error when varying the number of serial Mamba-2 causal temporal blocks (2/3/5/8) with and without CUF, complementing the TotalCapture ablation in Fig. 4; this depth study uses GT bone lengths to isolate the temporal-stack effect on DIP-IMU. CUF consistently outperforms the variant without CUF at shared depths: 11.87° vs. 13.11° at 3 layers, 11.34° vs. 13.10° at 5 layers, and 12.37° vs. 13.83° at 8 layers, with the same trend at 2 layers (see figure). CUF itself peaks at 5 layers and degrades when further deepened to 8 layers, so naively deepening the stack is not preferred in SIP. Rather than expanding model capacity, CUF provides a stable gain at each shared depth and degrades more gracefully by persistently re-exposing current observations to deeper blocks within the same stack. As an exploratory attempt, we also replaced the intermediate activation of the two-layer Layer-wise Encoder $\psi^{(l)}$ with a mixed activation that splits the hidden features and passes them through GELU and tanh separately; the effect appeared stable, so we defer a rigorous validation of this design to future work.

UWB dropout robustness. To emulate UWB occlusions at test time, we randomly drop inter-sensor distances with probability p and fill missing entries with the previous frame, keeping all other inputs unchanged. Table 5 reports SIP error on TotalCapture with SIP only. Accuracy degrades gracefully: 10% dropout is nearly lossless (+0.02°), and even 30% dropout raises SIP by only 0.09° (9.04° → 9.13°), indicating that the fused inertial and skeletal cues sustain pose accuracy when ranging is intermittently missing.

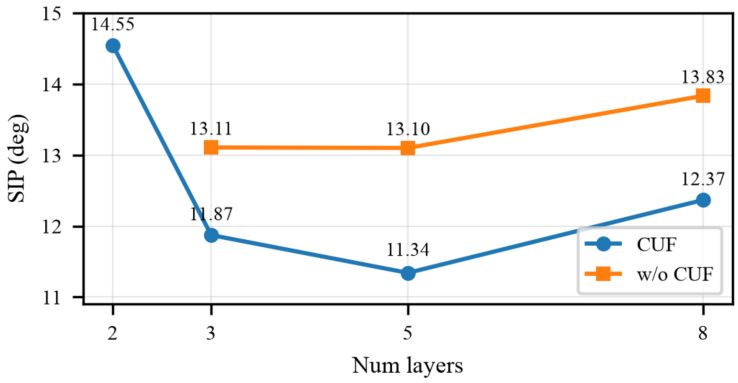

Figure 7: DIP-IMU SIP error (deg) vs. number of serial Mamba-2 blocks with and without CUF. CUF is better at all shared depths and degrades more gracefully beyond 5 layers.

Table 5: SIP error (deg) on TotalCapture under random UWB distance dropout. Missing entries are filled with the previous frame.

Drop ratio	0%	10%	20%	30%
SIP	9.04	9.06	9.09	9.13

A.6 On the Limitation of Jitter

Jitter, defined as mean joint jerk ($10^3 m/s^3$), is sensitive to frame-to-frame discontinuities and is therefore useful for measuring local smoothness. However, as a per-frame average, it may be less sensitive to long-range coherence, e.g., discontinuities at the boundaries of short-window inference. Short-window paradigms, such as the sliding-window design in ASIP Wu et al. (2024), excel at enforcing smoothness within each window, which can yield very low jitter scores; yet the correlation across windows is inherently weaker, so long sequences may still exhibit perceptible transitions upon visualization. We offer this as a preliminary observation rather than a comparison of methods: it suggests that jitter is best interpreted alongside long-sequence visualization and, in future work, complementary long-range smoothness measures.

Beyond this scalar criterion, we examine smoothness spectrally. Figure 8 compares the angle PSD of MISP and ASIP Wu et al. (2024) on the same sequences. In the residual and geodesic panels, MISP stays below ASIP across the entire 1–13 Hz band, and the gap widens at high frequencies; correspondingly, ASIP's local-angle spectrum retains high-frequency energy above the ground-truth level, whereas MISP's decays more rapidly. Taken together, this indicates that ASIP's excess high-frequency energy is largely error rather than motion, and that MISP's faster decay coincides with lower residual error, so the suppressed content is predominantly noise. This is precisely what a per-frame jerk average cannot reveal: a single scalar conflates genuine high-frequency articulation with high-frequency error, whereas the PSD separates them by frequency and shows that MISP is smoother without paying a spectral-fidelity penalty. We therefore report jitter for comparability with prior work, but regard the PSD as the more informative smoothness–fidelity diagnostic.

A.7 Efficiency Details

Table 6 benchmarks the GPU-side throughput of the local poser modules in one subject only, excluding CPU-side preprocessing and data transfer overheads. PNP uses only its virtual-force and local-pose modules. Online: frame-wise. Offline: full-sequence ($T = 10240$). Mamba-2's parallel scan formulation allows efficient full-sequence processing, while recurrent and optimization-based baselines remain bound by sequential dependencies. Once the shared PNP-style global translation estimation is included, all models return to the same level at just over 200 FPS online (205 for PNP, 220 for MISP): the end-to-end bottleneck is the physics optimizer for global displacement, while the local-pose backbone itself retains orders-of-magnitude headroom.

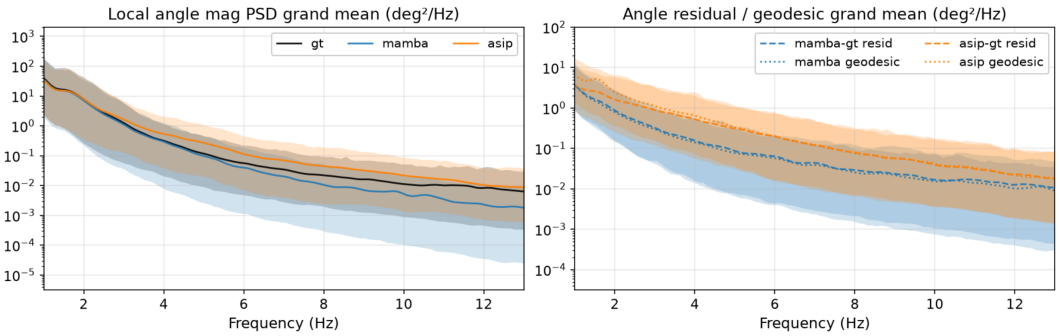

Figure 8: Cross-sequence angle PSD comparison between MISP and ASIP on TotalCapture. Left: grand-mean PSD of the local joint-angle magnitude (deg^2/Hz) for ground truth (GT), MISP, and ASIP, pooled over 45 sequences $\times$ 15 non-ignored SMPL joints (all joints except pelvis, ankles, feet, wrists, and hands; the pelvis is excluded because root rotation is handled differently in the two pipelines). Right: grand-mean PSD of the prediction residual (prediction $-$ GT) and of the per-frame geodesic angular error. Solid curves are arithmetic means over pooled joint–sequence PSDs and shaded bands the interquartile range (P25–P75); quantile bands are used because Welch PSD estimates are strictly positive and heavy-tailed, making linear standard-deviation bands ill-defined on a logarithmic axis. Spectra are computed with Welch's method (Hann window, 256-sample segments $\approx$4.27 s, 50% overlap, 512-point FFT, constant detrending) at f_s=60 Hz ($\approx$0.12 Hz resolution); the x-axis is restricted to 1–13 Hz for readability and the logarithmic y-axis uses data-driven limits without clipping. Both models are evaluated on identical $T-4$ aligned segments under a protocol-matched inference pipeline, with per-sequence ground-truth body-shape inputs (bone lengths and inter-sensor distances); SIP over all 45 sequences is 8.10° (MISP) vs. 8.23° (ASIP).

Table 6: GPU-side throughput and memory of local poser modules in one subject only. PNP uses only its virtual-force and local-pose modules. Online: frame-wise. Offline: full-sequence (T = 10240).

	Online		Offline	
Method	FPS ↑	Mem (MB) ↓	FPS (k) ↑	Mem (MB) ↓
PNP	1,457	189	1.5	197
FIP	1,662	**43**	60	301
MISP (ours)	**3,876**	58	**1,302**	**118**
PNP + dynopt	205	187	0.206	187
MISP + dynopt	220	58	0.220	58

A.8 Qualitative Details

The supplementary video compares MISP against two opposite ends of the accuracy–smoothness spectrum. ASIP reports the lowest jitter via non-causal sliding-window inference, yet its motion shows visible jitter at window boundaries: smoothness within each window does not guarantee coherence across windows, consistent with the scalar-jitter limitation in Appendix A.6. PNP instead suppresses jitter with contact-driven physics optimization, but its landings appear stiffened as foot-contact constraints pin down touchdown poses. MISP sits between the two: causal per-frame estimation re-grounded on current observations, without window seams or contact pinning.

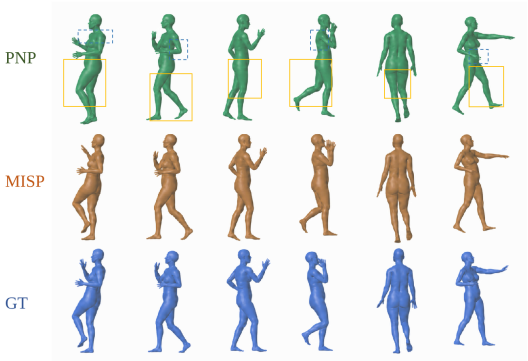

Figure 9: Qualitative comparison against PNP. Across representative frames, MISP better matches the ground-truth body configuration, especially in lower-limb articulation and pelvis-leg coordination.

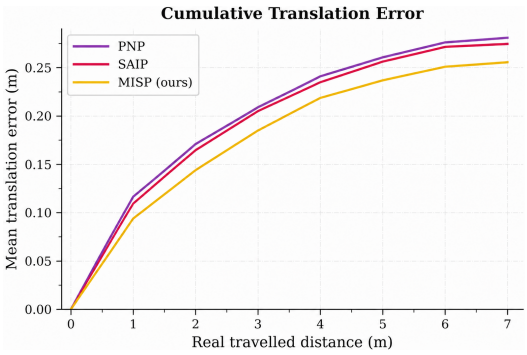

Figure 10: Cumulative translation error as the travelled distance increases. For translation recovery, we follow PIP's dynamics-based optimization protocol and compare the resulting global trajectories.