

Sen-Cap: Sensor-Flexible and Noise-Resilient Human Motion Capture via LiDAR-Camera Integration

Aoru Xue^{1†}, Yujing Sun^{2†}, Yiming Ren^{1,2}, Kwok-Yan Lam², Mao Ye³, and
Yuxin Ma¹ ^{*}

¹ShanghaiTech University, China, ² Digital Trust Centre, Nanyang Technological University, Singapore, ³ EABOT.AI, China
{xuear2024, mayuexin}@shanghaitech.edu.cn

Abstract. We propose **Sen-Cap**, a Sensor-Flexible and Noise-Resilient 3D human motion Capture framework that integrates multi-modal data from LiDAR and camera. While multi-modal sensors provide richer information than single-modal sensors, existing approaches still suffer from two core challenges. First, multi-modal alignment/matching across arbitrarily deployed sensors is typically handled by explicit calibration, which propagates errors under changing viewpoints and in turn constrains deployment to fixed, highly overlapped layouts. Second, prior methods degrade under severe noise or partial sensor failures, which are common in real-world environments. To address these challenges, **Sen-Cap** introduces a Unified Across-Sensor Motion Estimator that reconstructs local pose and shape in a human-centric space without calibrations between sensors, supporting a flexible number of sensors, as well as a Noise-Resistant Trajectory Tracker that maintains robustness under severe point cloud noise through iterative refinement. These sensor-flexible and noise-resilient features make Sen-Cap more practical in real-world deployment. Notably, operating in real time, **Sen-Cap** achieves state-of-the-art performance on major metrics on Human-M3 and FreeMotion, as well as strong cross-domain performance on LiDARHuman26M and RELI11D. This combination of flexibility and robustness opens new opportunities for motion capture in real-world scenarios, e.g. sports analytics, field robotics, and large-scale immersive environments.

1 Introduction

Accurate and robust global 3D human motion capture [2, 6, 16, 18, 19, 46, 49, 53] is critical for applications in AR/VR, robotics, human-computer interaction, sports, and healthcare [44, 47]. Markerless motion capture methods typically use either cameras [21, 37, 38, 41] or LiDAR [4, 14, 34, 35], but single-modality methods suffer from limited information: Monocular cameras lack depth, while

^{*} Corresponding author.

[†] Equal contribution.

LiDAR lacks texture. Fusing both modalities provides complementary strengths, geometry and location information from LiDAR and appearance information from cameras, enabling more precise human motion estimation of local pose and global translation.

Nevertheless, while multi-modal sensor methods provide richer information, they face increasing challenges as the number of sensors and modalities grows. On one hand, most multi-modal MoCap methods [7, 13, 31, 50] rely on data alignment via calibration, which propagates errors, especially when sensors have varying or dynamically changing perspectives, and thus requires fixed sensor placement with highly overlapping views, limiting flexibility and reducing the benefit of diverse viewpoints for occlusion handling. Although the most recent SoTA FreeCap [50] eliminates manual calibration through on-the-fly estimation from matched keypoints, the involved calibration step can still accumulate errors, potentially leading to failures. On the other hand, prior methods degrade significantly under severe noise [35]. Although recent methods such as LiveHPS++ [34] leverage LiDAR geometry to improve translation estimation, severe noise can cause normalization of point clouds to shift human points away from the valid data domain, resulting in substantial inaccuracies and unstable trajectories.

The above limitations in prior works motivate two central research questions: 1) *Can hybrid (LiDAR+Camera) motion capture achieve robust multi-modal alignment/matching under arbitrary sensor deployments?* Solving this challenge enables flexible deployment, where the system can adaptively adjust sensor layouts to mitigate occlusions and satisfy diverse task requirements in real-world applications; and 2) *How to make precise prediction of global human translation in noisy and cluttered environments?*

To address these challenges, we propose **Sen-Cap**, a novel human motion capture approach that predicts (1) human local poses through a **Unified Across-Sensor Motion Estimator (UAME)** and (2) global trajectories with a **Noise-Resilient Trajectory Tracker (NTT)**, achieving robust and precise results even in noisy and variable-viewpoint environments. Here, the sensor-flexible multi-modal fusion strategy is instantiated within UAME rather than as an independent module. The local poses and global trajectories together constitute the final estimation of human motion. Specifically, **The Unified Across-Sensor Motion Estimator** estimates the human local pose and shape in the *Human-Centric Space*. UAME consists of two coupled components, Human-Centric Space Alignment and Adaptive Sensor Fusion, which together form our sensor-flexible multi-modal fusion strategy. This design dynamically unifies data from a flexible number of uncalibrated sensors with unknown poses, enabling easy deployment in complex indoor and large-scale outdoor environments. **The Noise-Resilient Trajectory Tracker** employs an iterative refinement mechanism that gradually expands the searching space, and even if the distribution of point clouds is greatly affected by extreme noise, it can achieve consistent and accurate global translation estimation. They together result in a sensor-flexible and noise-resilient global pose estimation that supports variations in sensor modality, viewpoint and quantity without model retraining.

We evaluate *Sen-Cap* against SoTA single-sensor and hybrid-sensor methods on the large-scale multi-person dataset Human-M3 [9] and the multi-view dataset FreeMotion [35]. We further conduct cross-dataset evaluations on the challenging outdoor dataset LiDARHuman26M [20] and the fast-motion dataset RELI11D [51], where viewpoint discrepancies relative to the training data are substantial. Extensive experiments demonstrate that our approach achieves strong robustness to noise and generalization capability, while supporting a flexible number of sensors without model retraining. Finally, we evaluate sensor flexibility by varying the number of sensors under a unified parameter setting, and assess noise resilience across different noise levels. The results confirm the adaptability of our framework to flexible deployment scenarios and its robustness to noise. Our main contributions are summarized as follows:

1. To the best of our knowledge, we present the first global human motion capture framework that supports a flexible number of multi-modal sensors (LiDAR and cameras) within a unified model.
2. We propose a Unified Across-Sensor Motion Estimator (including Human-Centric Space Alignment and Adaptive Sensor Fusion) and a Noise-Resilient Trajectory Tracker that enable cross-modal alignment with uncalibrated, multi-modal, varying numbers of sensors, while remaining robust to substantial sensor noise.
3. We validate performance of *Sen-Cap* extensively on multiple public benchmarks as well as through real-world deployments, demonstrating unprecedented flexibility and robustness for in-the-wild applications.

2 Related Work

2.1 Visual Sensor-based Motion Capture

Early motion capture methods that estimate high-quality human motions rely on wearable sensors, such as markers [23, 28, 29, 33, 39, 45], IMUs [12, 36, 54], and ego cameras [3, 5, 8, 15, 56] but are limited by issues such as sensor drift, discomfort for the wearer, and challenges in capturing certain movements, especially in dynamic or complex environments. The shift towards non-wearable motion capture has been significant with the use of cameras. Monocular-based methods [1, 11, 17, 26, 27, 30] provide convenient solutions but suffer from depth perception limitations and rely on single-modal representations that lack robustness. Multi-view-based methods [7, 13, 25] can capture comprehensive motion information but are difficult to set up and typically employ simple feature concatenation rather than principled cross-modal representation learning. Both of them are hard to recover the global human motions due to limited representation capacity. Recent efforts [21, 32, 37, 38, 40–42, 52, 55] that use dynamic monocular views to estimate global motions still face challenges with translation accuracy, particularly in long video sequences, primarily due to insufficient temporal representation modeling. LiDAR has been instrumental in capturing global human

motion in open and large-scale environments [10, 43]. Early LiDAR-based methods like LiDARCap [20] focus on local pose estimation using point-based representations, while LiDAR-HMR [10] aims to reconstruct human meshes through geometric feature learning. However, these methods are limited by the lack of unified representation spaces that can effectively integrate geometric and semantic information. PointHPS [4] and LiveHPS [35] have since advanced to predict full SMPL parameters, capitalizing on the depth information provided by LiDAR through improved point cloud. LiveHPS++ [34] further improves the robustness and smoothness of human motion capture via temporal consistency modeling. Despite these advancements, challenges for LiDAR-based methods still remain due to data sparsity and the absence of texture information, fundamentally stemming from the lack of cross-modal representation learning capabilities.

2.2 Hybrid Sensor-based Motion Capture

Recognizing the limitations of single visual sensor-based approaches, research has increasingly focused on using multiple sensors [48]. However, existing multi-modal approaches typically employ naive feature concatenation or late fusion strategies, lacking principled cross-modal representation learning frameworks. Human-M3 [9] integrates LiDARs with cameras to address occlusion challenges, and fully utilize advantages in different modal data through simple feature aggregation. However, these systems are often constrained by the need for precise calibration, necessitating fixed and pre-deployed setups, and more importantly, they lack unified cross-modal representation spaces that can effectively bridge the semantic gap between different modalities. A recent advance is FreeCap [50], which proposes a calibration-free method to match multi-persons in point clouds and images and uses the matching pairs to calculate a coarse calibration matrix. However, the network is heavyweight due to its lack of efficient representation learning mechanisms, and in cases where reference human points are few, the calibration matrix is inaccurate, degrading the performance of the fusion results. Additionally, it only supports one LiDAR within its framework, limiting its applicability and scalability. In contrast to these approaches that rely on explicit calibration or simple feature fusion, our method introduces a novel multi-modal representation learning framework that learns unified cross-modal representations through Human-Centric Space Alignment, eliminating the need for calibration estimation while supporting extendable LiDARs and cameras.

3 Methodology

Accurate 3D human motion capture from uncalibrated, heterogeneous sensors is challenging due to the inherent variability in sensor configurations and the susceptibility to noise in real-world environments. Existing methods often fail to address two critical issues: (1) *dependency on fixed sensor setups* that requires precise calibration and cannot adapt to dynamic sensor availability; (2) *sensitivity to sensor noise and occlusion* that leads to inaccurate global trajectory estimation.

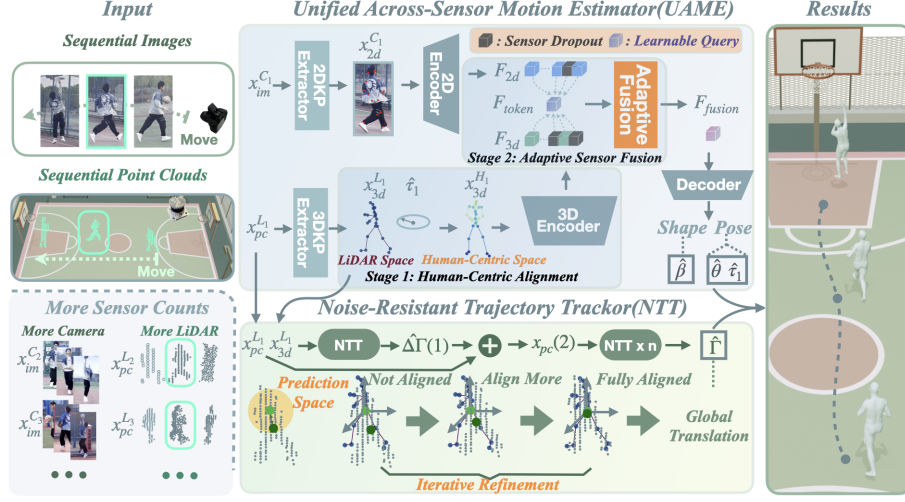


Fig. 1: Overview of Sen-Cap. With sequential point clouds from movable LiDARs and 2D key points from movable cameras as input, Sen-Cap consists of two primary modules, a Unified Across-Sensor Motion Estimator (UAME) for obtaining the human global orientation, pose, and shape, and a Noise-resistant Trajectory Tracker (NTT) for global trajectory estimation. All components corresponding to our key contributions are highlighted in orange.

Sen-Cap overcomes these limitations via two technical innovations: first, a **Unified Across-Sensor Motion Estimator**(Section 3.1) that extracts consistent human pose and shape from flexible sensor combinations in a calibration-free manner. In this work, “calibration-free” means that no explicit extrinsic parameters between sensors are required during inference. Our model does not assume pre-calibrated relative poses and can operate under arbitrary unknown rigid transformations between sensors; second, a **Noise-resistant Trajectory Tracker**(Section 3.2) that resists sensor noise through iterative refinement. Notably, our method supports input from varying combinations of LiDARs and cameras, offering high flexibility and adaptability in real-world deployments. An Overview of **Sen-Cap** is shown in Fig. 1.

Problem Statement and Preliminaries *Sen-Cap* takes as input sequentially sampled point clouds x_{pc} from LiDAR and normalized 2D keypoints x_{2d} from images, requiring no calibration and supporting a flexible number of sensors, and predicts the full SMPL [22] parameters as the outputs, which represent the underlying human pose and shape. Although our input is time-series data, for the sake of convenience, we save time t when writing symbols. First, we follow LiveHPS++ [34] to use the farthest point sampling algorithm (FPS) to sample the point cloud as a fixed number $x_{pc}^{L_i} \in \mathbb{R}^{256 \times 3}$ from the i -th LiDAR, and follow WHAM [38] to normalize the 2D keypoints $x_{2d}^{C_j} \in \mathbb{R}^{17 \times 2}$ with its bounding box

center and scale from the j -th camera. Meanwhile, following [50], the ground truth of our algorithm is SMPL [22] parameters with shape $\beta \in \mathbb{R}^{10}$, body pose $\theta \in \mathbb{R}^{23 \times 3}$, global orientation $\tau \in \mathbb{R}^3$ and global translation $T \in \mathbb{R}^3$ while $\hat{\beta}$, $\hat{\theta}$, $\hat{\tau}$ and $\hat{T}$ are corresponding predictions of our framework. These parameters are used to describe the mesh of the human body with 6890 vertices and 24 joints. Finally, as cross-sensor human matching is essential for our sensor-flexible framework in multi-person scenarios, we build upon the optimization-based per-frame matching mechanism of FreeCap [50]. To enable real-time online inference, we further introduce a matching memory bank. More details are provided in supplementary material.

3.1 Unified Across-Sensor Motion Estimator: Local Pose Recovery

The core challenge in sensor-flexible motion capture is reconciling data from arbitrarily configured sensors. Our solution exploits the invariant properties of human kinematics: by estimating and aligning to the human centric space with *Unified Across-Sensor Motion Estimator (UAME)*, we create a consistent reference frame that automatically unifies inputs from diverse sensors, regardless of their number, type, or viewpoint. Architectural details of the Unified Across-Sensor Motion Estimator (UAME) are provided in supplementary material.

Limitation of prior local-motion methods. Most prior multi-modal local pose pipelines ([28], [50]) either depend on explicit calibration to align sensor observations or adopt unconstrained feature aggregation under fixed sensor assumptions. As a result, they are sensitive to missing sensors and unstable cross-modal correspondences when the deployment changes at inference time.

As shown in Fig. 1, UAME takes as input uncalibrated sequential 3D point clouds $x_{pc}^{L_i}$ from multiple LiDARs L_i , and 2D keypoints $x_{2d}^{C_j}$ from multiple cameras C_j . It then estimates the local pose $\hat{\theta}$ and human shape $\hat{\beta}$ in a human-centric coordinate system that remains consistent across diverse sensor viewpoints and modalities. This is achieved through two key stages: 1) *Human-Centric Space Alignment*, and 2) *Adaptive Sensor Fusion* addresses sensor variability through a bottleneck attention module that explicitly compresses multi-sensor information into a learnable latent token. By forcing all modality-specific features to interact through this shared bottleneck representation, the module adaptively re-weights sensor contributions according to their reliability and contextual relevance. The resulting fused feature dynamically adjusts to the current sensor configuration, enabling robust and accurate local pose estimation across diverse sensing conditions.

Stage 1. Human-Centric Space Alignment Given sequential point clouds $\{x_{pc}^{L_i}\}_{i=1}^{N_L}$ from N_L LiDARs, we first learn a unified representation space that aligns all heterogeneous sensor modalities to a Human-Centric coordinate system H . This alignment mechanism enables effective cross-modal knowledge transfer without explicit calibration. The Human-Centric coordinate system is centered around humans and is independent of sensor placement. For each LiDAR L_i , a

Gated Recurrent Unit (GRU) learns to predict rotation vector τ_i that transforms its 3D joint representations $x_{3d}^{L_i}$ encoded by 3D PointNet Encoder to the unified Human-Centric space H_i .

$$x_{3d}^{H_i} = R(\hat{\tau}_i)^{-1} x_{3d}^{L_i}, \quad (1)$$

where $R(\cdot)$ transforms rotation vector to rotation matrix. We then use loss $\mathcal{L}_{align}$ to optimize 3D representations in both LiDAR space $\mathbf{x}_{3d}^L = \{\mathbf{x}_{3d}^{L_1}, \dots, \mathbf{x}_{3d}^{L_{N_L}}\}$ and human-centric spaces $\mathbf{x}_{3d}^H = \{\mathbf{x}_{3d}^{H_1}, \dots, \mathbf{x}_{3d}^{H_{N_L}}\}$, together with 6D representation of the global transformation matrix [57] $\tau_i^{6d} = \{\tau_1^{6d}, \dots, \tau_{N_L}^{6d}\}$:

$$\mathcal{L}_{align}(\tau^{6d}, \mathbf{x}_{3d}^L, \mathbf{x}_{3d}^H) = \frac{1}{N_L} \sum_{i=1}^{N_L} (\|\tau_i^{6d} - \hat{\tau}_i^{6d}\|_2^2 + \|x_{3d}^{L_i} - \hat{x}_{3d}^{L_i}\|_2^2 + \|x_{3d}^{H_i} - \hat{x}_{3d}^{H_i}\|_2^2), \quad (2)$$

Besides, the 2D inputs are aligned by encoding the 2D keypoints x_{2d} into a human-centric feature space via the 2D Encoder.

Stage 2. Adaptive Sensor Fusion. This stage introduces a bottleneck attention fusion module to explicitly mitigate sensor variability. Instead of directly performing unrestricted cross-modal interactions, modality-specific features are squeezed into a shared learnable latent token that serves as an information bottleneck. Through this constrained aggregation process, the model selectively preserves sensor-specific cues that are reliable and motion-relevant, while suppressing noisy or degraded signals. The resulting compact fused representation remains robust to changing sensor configurations, ensuring stable and accurate local pose estimation.

We first learn modality-specific representations that capture the unique characteristics of each sensor type. The aligned 3D joint representations $x_{3d}^{H_i}$ are processed by a GRU to extract temporal motion features F_{3d} that encode spatial-temporal dynamics. Meanwhile, sequential 2D keypoints $\{x_{2d}^{C_j}\}_{j=1}^{N_C}$ from N_C cameras are encoded by another GRU to produce visual motion features F_{2d} .

To mitigate sensor variability and enforce structured cross-modal interaction, we introduce a bottleneck attention-based adaptive fusion module. Instead of allowing unrestricted pairwise fusion across modalities, we employ a learnable fusion token F_{token} as a shared information bottleneck. This token serves as the query in a cross-attention operation to aggregate multi-modal features, thereby squeezing heterogeneous sensor representations into a compact latent space:

$$F_{fusion} = \text{CrossAttn}(Q = F_{token}, K = F_m, V = F_m), \quad (3)$$

where $F_m = \text{Concat}(F_{3d}, F_{2d})$ denotes the concatenated multi-modal features. Through this bottleneck aggregation, the model adaptively reweights modality contributions according to their reliability and contextual relevance, while preventing direct and potentially noisy cross-modal interactions. The resulting compact fused representation F_{fusion} encodes the most informative sensor cues under the current configuration.

The fused features are then decoded by a GRU to regress SMPL parameters $\hat{\beta}, \hat{\theta}$:

$$\hat{\beta}, \hat{\theta} = \text{GRU}(F_{\text{fusion}}), \quad (4)$$

with the pose supervision defined as

$$\mathcal{L}_{\text{pose}}(\hat{\beta}, \hat{\theta}) = \alpha \|\beta - \hat{\beta}\|_2^2 + \|\theta - \hat{\theta}\|_2^2. \quad (5)$$

Sensor Dropout Motivated by the above limitation, we introduce Sensor Dropout to explicitly train UAME under incomplete or degraded sensor configurations. Specifically, for each training sample, we randomly mask a certain percentage of feature tokens from both LiDAR and camera modalities before fusion. This training strategy simulates a spectrum of failures, from partial observation loss to full sensor absence, forcing the bottleneck fusion module to re-weight remaining reliable cues. The scaling-down and cross-configuration evaluations in Section 4.6 are consistent with the intended robustness improvement without retraining.

3.2 Noise-resistant Trajectory Tracker: Robust Global Localization

Besides a precise local motion, an accurate global trajectory is another critical component for coherent human motion capture. Thus, we then take as input the sequential point clouds $\mathbf{x}_{pc}$ from each LiDAR, and the corresponding 3D joints $\hat{\mathbf{x}}_{3d}^L$ (Section 3.1) to estimate the Global Trajectory $\hat{\Gamma}$.

Limitation of prior works. Standard pipeline [35] first normalized the point cloud so that its centroid coincides with the coordinate origin, then predicts a single offset to reach the human root. This procedure implicitly assumes that the centroid is close to the pelvis; however, noise or partial occlusion can shift the centroid far away, forcing the network to explore an abnormally large **prediction space** of candidate offsets. The enlarged prediction region biases the learned prior toward under-estimating large displacements and finally produces the systematic drift visualized in Fig. 2.

Hence, to keep the prediction space more stable under different perturbation levels, we adopt an iterative refinement strategy that gradually moves the centroid closer to the true root, reducing both offset magnitude and prediction variance at every step. Specifically, we use a Noise-resistant Trajectory Tracker (NTT) with an iterative refinement mechanism to obtain a robust and smooth

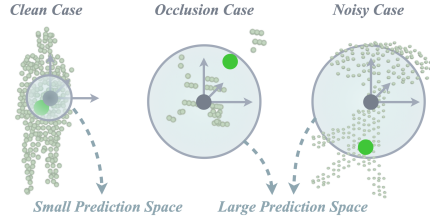


Fig. 2: Systematic drift arises when the true root joint (green ball) is offset from the point cloud center $\bar{x}_{pc}$ (gray ball), especially under severe occlusion and noise.

global trajectory under sensor noise and environmental variations. For each step, we predict the offset $\Delta\hat{I}^n$ from the previous point cloud x_{pc}^{n-1} at iteration $n-1$ and local motion x_{3d}^L , with

1) *Iterative Offset Prediction*: For each step n :

$$\Delta\hat{I}^n = E_O(x_{pc}^{n-1}, x_{3d}^L), \quad x_{pc}^n = x_{pc}^{n-1} + \Delta\hat{I}^n, \quad (6)$$

where E_O is the offset predictor and we denote $x_{pc}^0 = x_{pc}$ and the center of original point cloud $\bar{x}_{pc}$.

2) *Constrained Optimization*: The final $\hat{I}$ of the human in LiDAR coordinates is then determined as $\hat{I} = \bar{x}_{pc} + \sum_{n=1}^N \Delta\hat{I}^n$ after N iterations. We optimize the global trajectory representation with loss $\mathcal{L}_{tr}(I)$ using ground truth step offset and total trajectory, along with penalizing too large predictions,

$$\mathcal{L}_{tr}(I) = \|\hat{I} - I\|_2^2 + \lambda \frac{1}{N} \sum_{n=1}^N \mathcal{L}_{dtr}(\Delta\hat{I}^n), \quad (7)$$

where $\mathcal{L}_{dtr}(\Delta\hat{I}^n)$ is the penalty loss for preventing large step size with penalty δ :

$$\mathcal{L}_{dtr}(\Delta\hat{I}^n) = \begin{cases} 0 & , \|\Delta\hat{I}^n\| \leq \delta \\ \|\|\Delta\hat{I}^n\| - \delta\| & , \|\Delta\hat{I}^n\| > \delta. \end{cases} \quad (8)$$

Remarks. Unlike the coordinate system-independent parameters $(\hat{\beta}, \hat{\theta})$ estimated in the local representation (Section 3.1), the noise-resilient global trajectory $\hat{I}$ is defined relative to LiDAR coordinate systems. To this end, the Noise-resistant Trajectory Tracker(NTT) predicts the global translation for each LiDAR in a batched manner, depending on the number of available LiDARs.

Real-Time Online Deployment. For real-time online deployment, we additionally enhance temporal consistency and trajectory smoothness as detailed in supplementary material.

4 Experiments

4.1 Experimental Setup

Implementation Details. Our framework is implemented in PyTorch and trained using AdamW with cosine learning rate decay. All experiments are conducted on NVIDIA RTX-3090 GPUs. Detailed hyperparameters and network architectures are provided in supplementary material.

Baselines. We compare against seven recent state-of-the-art methods. Camera-based baselines are WHAM, GVHMR, GENMO, and PromptHMR [21, 37, 38, 41]. LiDAR-based baselines are LiveHPS and LiveHPS++ [34, 35]. The camera-LiDAR hybrid baseline is FreeCap [50].

Methods	Venue&Year	Modality View		Dataset FreeMotion						Dataset HumanM3					
				J/V Err(PS)	J/V Err(PST)	Ang Err	Accel Err	SUCD		J/V Err(PS)	J/V Err(PST)	Ang Err	Accel Err	SUCD	
WHAM [38]	CVPR'24	C	T	82.83/97.55	-	12.24	4.51	-		69.42/83.35	-	10.13	9.20	-	
GVHMR [37]	SA'24	C	T	72.47/84.23	-	13.80	4.59	-		79.39/91.23	-	13.49	19.00	-	
GENMO [21]	ICCV'25	C	T	62.37/73.71	-	12.34	3.33	-		76.84/91.36	-	10.92	13.84	-	
PromptHMR [41]	CVPR'25	C	T	71.02/83.46	-	13.43	3.62	-		88.72/104.95	-	11.68	11.92	-	
LiveHPS [35]	CVPR'24	L	T	59.30/73.12	100.81/109.09	13.10	6.18	4.97		57.81/71.27	97.11/103.10	10.44	12.58	6.75	
LiveHPS++ [34]	ECCV'24	L	T	54.15/57.74	88.91/97.53	11.85	4.37	3.47		55.67/68.73	89.06/95.85	10.26	11.51	5.60	
FreeCap [50]	AAAI'25	L+C	T	53.31/65.30	95.97/102.91	11.14	5.97	4.82		55.45/68.52	96.47/102.67	9.14	9.60	6.66	
Ours[†]	-	L	T	52.98/56.41	84.37/91.62	11.21	4.28	3.21		53.94/66.82	86.75/93.14	9.88	11.08	5.42	
Ours	-	L+C	T	47.46/57.84	75.16/81.65	9.94	3.01	3.83		42.59/52.09	73.50/77.53	5.71	10.18	5.74	
WHAM [38]	CVPR'24	C	N	108.16/124.52	-	19.89	4.49	-		85.50/101.44	-	10.42	9.25	-	
GVHMR [37]	SA'24	C	N	114.00/126.41	-	18.22	4.32	-		83.02/95.83	-	14.68	20.14	-	
GENMO [21]	ICCV'25	C	N	72.48/81.96	-	12.87	3.84	-		79.31/94.82	-	11.15	14.27	-	
PromptHMR [41]	CVPR'25	C	N	96.00/107.38	-	16.54	3.69	-		91.46/108.73	-	12.06	12.34	-	
FreeCap [50]	AAAI'25	L+C	N	57.37/69.61	99.20/106.15	11.74	6.08	4.96		56.24/69.25	96.45/102.46	9.17	9.59	6.58	
Ours	-	L+C	N	52.25/63.05	79.05/85.82	10.80	3.16	3.93		42.64/52.28	73.41/77.51	5.76	10.15	5.74	

Table 1: Evaluation on large-scale FreeMotion and Human-M3 datasets. The “View” specification differentiates between performance under training (T) and novel camera viewpoints (N). In the “Modality” column, C denotes Camera, L denotes LiDAR, and L+C denotes the multi-modal setting. Entries marked with “-” correspond to metrics that are inapplicable to camera-based methods due to inherent scale ambiguity. **Ours[†]** denotes our method using only a single LiDAR sensor. Note that **Ours[†]** achieves performance comparable to LiDAR-only methods (LiveHPS [35], LiveHPS++ [34]), while the full multi-modal version yields substantial gains, highlighting the effectiveness of our calibration-free architecture in aligning and fusing heterogeneous sensors.

Benchmarks and Protocols. Following FreeCap [50], we conduct evaluations on the large-scale multi-person HumanM3 [9] dataset and the multi-sensor FreeMotion [35] dataset.

Metrics. Following FreeCap [50], we report per-joint/vertex errors in millimeters (**J/V Err(PS/PST)**), acceleration error in m/s^2 (**Accel Err**), per-global-joint rotation error in degrees (**Ang Err**), and scene-level unidirectional Chamfer distance in millimeters (**SUCD**).

4.2 Quantitative Evaluations

As summarized in Table 1, we report results under two evaluation settings: (i) *training view* and (ii) *novel camera viewpoints*. The training-view setting reflects standard in-distribution accuracy. The novel-view setting, by contrast, introduces unseen viewpoints that effectively alter camera extrinsics at test time. Evaluating under this protocol is particularly relevant to our formulation, as SenCap does not rely on explicit calibration and is designed to remain stable under such viewpoint changes.

Under the training-view protocol, our Unified Across-Sensor Motion Estimator (UAME; Section 3.1) effectively fuses complementary cues from LiDAR and cameras, leading to substantial improvements over camera-only (WHAM [38], GVHMR [37], GENMO [21], PromptHMR [41]) and LiDAR-only (LiveHPS [35], LiveHPS++ [34]) baselines on most metrics. Moreover, we outperform the hybrid method FreeCap [50], largely due to our human-centric alignment strategy that removes the need for explicit calibration and reduces calibration-induced errors.

Methods	Dataset LiDARHuman26M						Dataset RELI11D					
	J/V Err(PS)↓	J/V Err(PST)↓	Ang Err↓	Accel Err↓	SUCD↓		J/V Err(PS)↓	J/V Err(PST)↓	Ang Err↓	Accel Err↓	SUCD↓	
WHAM [38]	151.18/191.73	-	32.92	3.16	-		106.13/126.71	-	22.10	2.70	-	
GVHMR [37]	187.01/226.19	-	28.58	3.04	-		111.73/131.73	-	21.37	2.42	-	
LiveHPS [35]	160.70/199.68	224.17/247.18	29.15	9.94	8.45		84.29/104.24	100.90/113.20	24.03	3.47	5.48	
LiveHPS++ [34]	141.03/179.68	196.30/221.71	26.23	4.67	5.20		78.45/95.41	88.10/102.43	22.28	1.68	2.09	
FreeCap [50]	160.87/202.42	224.35/249.65	28.53	10.16	9.65		80.49/98.47	98.88/109.36	23.60	3.48	5.53	
Ours	131.32/164.39	191.59/216.73	24.04	3.03	5.11		61.23/73.28	81.11/85.48	17.87	1.39	1.48	

Table 2: Cross-Domain Evaluation on Noisy and Fast-motion Benchmarks: the outdoor noisy LiDARHuman26M dataset and the indoor fast-motion RELI11D dataset. Our approach is trained on AMASS and FreeMotion.

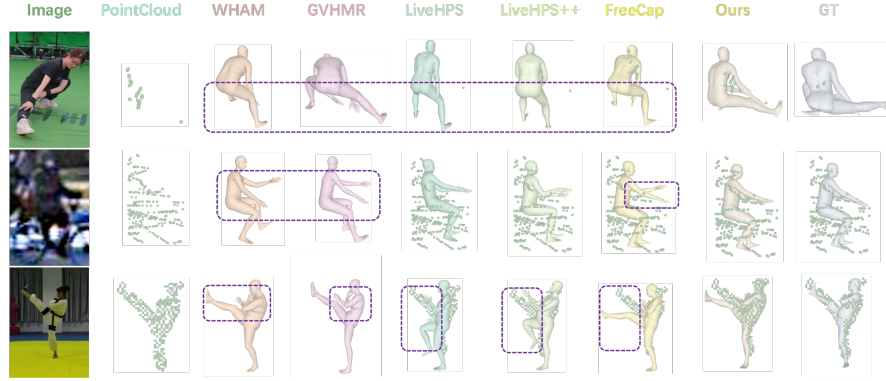


Fig. 3: Qualitative comparisons showing Sen-Cap's superior motion capture in challenging scenarios: (1) sparse-point leg stretching (2) Outdoor cycling (3) Fast-motion Taekwondo. We have highlighted the regions with significant differences using purple boxes.

We also report a LiDAR-only variant (**Ours[†]**) using a single LiDAR sensor. It achieves performance comparable to LiDAR-only methods, while the full multi-modal version yields significant gains, highlighting that Sen-Cap can effectively align and benefit from heterogeneous sensors in a calibration-free manner.

4.3 Out-of-Domain Generalization

To evaluate out-of-domain generalization, we conduct cross-dataset experiments on LiDARHuman26M [20] and RELI11D [51]. Our model is trained only on AMASS [24] and FreeMotion [35] without any fine-tuning on these datasets.

As shown in Table 2, Sen-Cap achieves competitive or superior performance across most metrics despite the domain gap. These results indicate that our framework generalizes well to unseen environments.

4.4 Qualitative Evaluations

Fig. 3 provides qualitative comparisons in challenging scenarios, including (1) sparse-point leg stretching, (2) outdoor cycling, and (3) fast-motion Taekwondo.

		Dataset FreeMotion				
		J/V Err(PS)↓	J/V Err(PST)↓	Ang Err↓	Accel Err↓	SUCD↓
Feature Space (UAME)	LiDAR-Centric	68.73/81.42	92.38/100.81	12.75	3.39	5.11
	Human-Centric	52.25/63.05	78.87/85.66	10.80	3.12	3.91
Feature Fusion (UAME)	Learnable Linear Fusion	60.63/72.41	85.49/93.09	11.73	3.27	4.45
	Fixed-Weight Fusion	61.37/73.13	85.97/93.55	11.70	3.28	4.53
	Bottleneck Attention (Ours)	52.25/63.05	78.87/85.66	10.80	3.12	3.91
Trajectory Estimation (NTT)	repeat × 1	52.25/63.05	87.88/94.01	10.80	5.05	4.26
	repeat × 2	52.25/63.05	82.42/88.78	10.80	4.18	3.54
	repeat × 3	52.25/63.05	78.87/85.66	10.80	3.12	3.91
	repeat × 4	52.25/63.05	74.88/82.20	10.80	3.17	3.81
	repeat × 5	52.25/63.05	74.63/82.00	10.80	3.17	3.75

Table 3: Ablation studies on our design choices. We analyze the impact of feature space, fusion strategy, and trajectory estimation iterations. **Note that the local pose metrics (J/V Err(PS) and Ang Err) remain unchanged across different NTT settings**, since NTT refines global translation while keeping the estimated local pose fixed.

When LiDAR point clouds become sparse or partially missing, image priors help maintain plausible body shape. Conversely, in outdoor conditions with degraded image quality, LiDAR geometry provides stable cues. More qualitative results and trajectory analyses are included in supplementary material.

4.5 Component Analysis

We conduct controlled ablations to isolate the contribution of each design choice in Sen-Cap.

Ablations on UAME. Table 3 analyzes key components in UAME. First, the feature space choice is crucial: replacing our human-centric alignment with a LiDAR-centric feature space significantly degrades performance, confirming the importance of human-centric representation for calibration-free operation. Second, we evaluate alternative fusion schemes including a learnable linear aggregation (via concatenation and projection) and a static summation-based fusion. Unlike these fully connected or fixed aggregation strategies, our bottleneck attention explicitly restricts cross-modal communication to pass through a compact latent token. This structured information compression enables reliability-aware modality weighting and leads to improved robustness.

Ablations on NTT. We further evaluate the iterative refinement strategy in NTT. As shown in Table 3, the trajectory error progressively decreases as the number of refinement iterations increases, with three iterations providing a favorable balance between accuracy and computational cost. Note that local pose metrics (e.g., J/V Err(PS) and Ang Err) remain unchanged across different NTT settings, since this module primarily refines global translation while keeping local pose estimation fixed. Additional ablations and qualitative studies are provided in supplementary material.

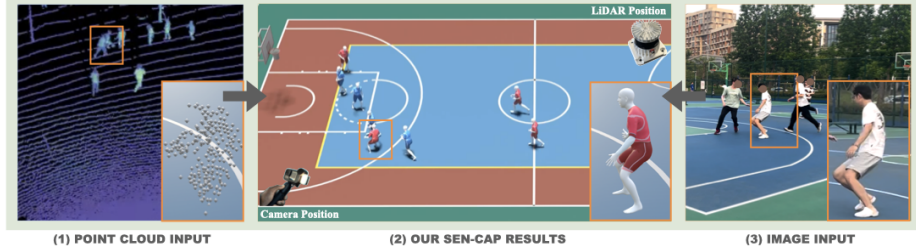


Fig. 4: Application on a basketball example to illustrate the versatility of Sen-Cap. The results (2) confirm that Sen-Cap successfully fuses heterogeneous inputs from a camera (3) and a LiDAR sensor (1), enabling accurate 3D human pose estimation in challenging dynamic scenarios without requiring sensor calibration.

4.6 Robustness Evaluation under Sensor and Noise Variations

A core goal of Sen-Cap is to remain reliable under varying sensor configurations and noisy environments. We validate this robustness through two targeted evaluations.

Sensor Flexibility. Sen-Cap demonstrates strong sensor flexibility, which can adapt to varying numbers and modalities of sensors without retraining or recalibration. We evaluate this property from three complementary settings, including scaling down, scaling up, and single-modal inference. For clarity, Fig. 5 visualizes Joint Err(PS) under different inference sensor configurations.

Scaling down. We evaluate a single unified model trained with multiple sensors (3 LiDARs + 3 cameras) under degraded test-time configurations with fewer sensors. As shown in Fig. 5(a), the unified model can effectively adapt to various reduced sensor configurations at test time. Its performance remains comparable to models trained specifically under the corresponding configurations, demonstrating strong flexibility without retraining or recalibration. This trend is consistent with the Sensor Dropout design in Section 3.1, which explicitly trains UAME to handle missing-sensor cases.

Scaling up. We also train on a minimal sensor setup and evaluate with additional sensors at inference time. As shown in Fig. 5(b), adding cameras and/or LiDARs generally reduces the reconstruction error, indicating that Sen-Cap can leverage extra sensors at test time and generalizes gracefully to denser deployments. We also include the unified 3C+3L model under the same inference configurations for comparison.

Single-modal performance. Fig. 5(c) shows that Sen-Cap maintains reasonable performance with either LiDAR-only or camera-only inputs. For both LiDAR-only and camera-only settings, increasing the number of available sensors leads to better performance, showing that the framework can benefit from additional sensors even within a single modality.

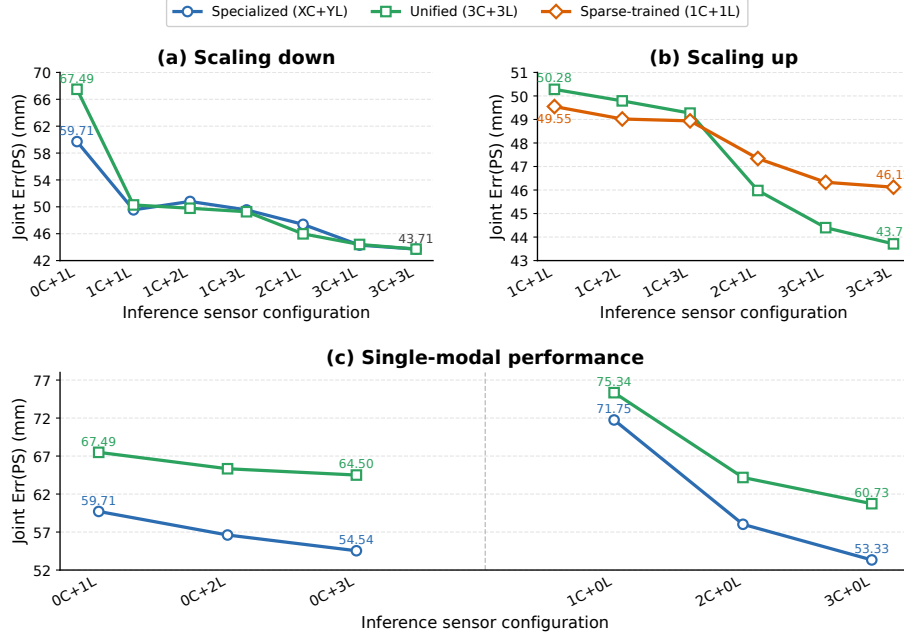


Fig. 5: Sensor flexibility on FreeMotion-indoor. We visualize Joint Err(PS) for (a) scaling down from a unified 3C+3L model, (b) scaling up a sparse 1C+1L model with additional test-time sensors, and (c) single-modal LiDAR-only and camera-only configurations. Lower values are better. C and L denote the numbers of cameras and LiDARs.

Together, these results demonstrate that Sen-Cap is both robust and adaptable to diverse sensor configurations, supporting calibration-free deployment in real-world multi-sensor motion capture scenarios.

Noise Resilience To systematically evaluate robustness against clutter and object noise, we construct three object-noise subsets of FreeMotion by adding synthetic clutter around the human point clouds: (i) Level 1 contains low-noise samples; (ii) Level 2 adds small objects (0.1–0.5 m); (iii) Level 3 adds larger objects (0.5–1 m) to create moderate clutter. The clutter is generated by sampling ShapeNet objects, scaling them to the target size, simulating visible points via raycasting from the LiDAR viewpoint, and randomly placing the resulting object point clouds around the human.

Noise	Setting	J/V Err (PST)↓	Accel↓	SUCD↓
L1	w/o NTT	67.91/72.18	3.64	3.41
	w NTT	64.28/69.12	2.93	3.19
L2	w/o NTT	96.49/101.84	10.54	189.51
	w NTT	89.49/95.33	3.08	185.45
L3	w/o NTT	367.51/374.16	94.62	553.86
	w NTT	158.53/165.00	6.71	377.83

Table 4: Ablation of NTT under increasing noise levels.

We compare performance with and without NTT ($N=3$). As summarized in Table 4, NTT consistently reduces errors across all noise levels, with the largest gains under the most challenging clutter setting (Level 3). Details of the noise-generation procedure and the definitions of each noise level are provided in the supplementary material.

4.7 Real-world Deployment in Large-scale Scenes

In real-world large-scale motion capture, multi-sensor setups are often necessary, while flexible sensor placements make calibration difficult and error-prone. Sen-Cap supports calibration-free deployment with varying sensor numbers and placements, enabling reliable motion capture in large scenes. Fig. 4 presents a real-world basketball example where Sen-Cap fuses LiDAR and camera observations to reconstruct accurate 3D human poses. Further technical details and demonstrations are provided in supplementary material and the supplemental video.

5 Conclusions

To conclude, we present *Sen-Cap*, a motion capture framework that operates robustly with flexible combinations of LiDARs and cameras under noisy environments. It eliminates calibration through a Unified Across-Sensor Motion Estimator that aligns data in a human-centric space, and ensures noise robustness via a Noise-Resistant Trajectory Tracker with iterative refinement. Extensive experiments across multiple datasets and in the real world demonstrate state-of-the-art performance, validating the practicality and flexibility of our approach for real-world motion capture.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Project Number 62595774), MoE Key Laboratory of Intelligent Perception and Human-Machine Collaboration (KLIP-HuMaCo), HPC Platform of ShanghaiTech University.

This research was supported by the National Research Foundation, Singapore and Infocomm Media Development Authority under its Trust Tech Funding Initiative. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore and Infocomm Media Development Authority.

References

1. Alldieck, T., Kassubeck, M., Wandt, B., Rosenhahn, B., Magnor, M.: Optical flow-based 3d human motion estimation from monocular video. In: GCPR. pp. 347–360. Springer (2017)
2. Belagiannis, V., Zisserman, A.: Recurrent human pose estimation. In: 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017). pp. 468–475 (2017). <https://doi.org/10.1109/FG.2017.64>
3. Betancourt, A., Díaz-Rodríguez, N., Barakova, E., Marcenaro, L., Rauterberg, M., Regazzoni, C.: Unsupervised understanding of location and illumination changes in egocentric videos. *Pervasive and Mobile Computing* **14**, Pages–414 (2017)
4. Cai, Z., Pan, L., Wei, C., Yin, W., Hong, F., Zhang, M., Loy, C.C., Yang, L., Liu, Z.: Pointhps: Cascaded 3d human pose and shape estimation from point clouds. arXiv preprint arXiv:2308.14492 (2023)
5. Cao, C., Zhang, Y., Wu, Y., Lu, H., Cheng, J.: Egocentric gesture recognition using recurrent 3d convolutional neural networks with spatiotemporal transformer modules. In: CVPR. pp. 3763–3771 (2017)
6. Charles, J., Pfister, T., Magee, D., Hogg, D., Zisserman, A.: Personalizing human video pose estimation. In: CVPR. pp. 3063–3072 (2016)
7. Dou, M., Davidson, P., Fanello, S.R., Khamis, S., Kowdle, A., Rhemann, C., Tankovich, V., Izadi, S.: Motion2fusion: Real-time volumetric performance capture. *ACM Trans. Graph.* **36**(6), 246:1–246:16 (Nov 2017)
8. EgoCap: EgoCap dataset. <http://gvv.mpi-inf.mpg.de/projects/EgoCap/> (2016), accessed 28 Jun 2026
9. Fan, B., Wang, S., Guo, W., Zheng, W., Feng, J., Zhou, J.: Human-m3: A multi-view multi-modal dataset for 3d human pose estimation in outdoor scenes. arXiv preprint arXiv:2308.00628 (2023)
10. Fan, B., Zheng, W., Feng, J., Zhou, J.: Lidar-hmr: 3d human mesh recovery from lidar. arXiv preprint arXiv:2311.11971 (2023)
11. Goel, S., Pavlakos, G., Rajasegaran, J., Kanazawa*, A., Malik*, J.: Humans in 4D: Reconstructing and tracking humans with transformers. In: International Conference on Computer Vision (ICCV) (2023)
12. Huang, Y., Kaufmann, M., Aksan, E., Black, M.J., Hilliges, O., Pons-Moll, G.: Deep inertial poser: Learning to reconstruct human pose from sparse inertial measurements in real time. *ACM Transactions on Graphics (TOG)* **37**(6), 1–15 (2018)
13. Huang, Z., Li, T., Chen, W., Zhao, Y., Xing, J., LeGendre, C., Luo, L., Ma, C., Li, H.: Deep volumetric video from very sparse multi-view performance capture. In: ECCV (September 2018)
14. Jang, D.K., Yang, D., Jang, D.Y., Choi, B., Jin, T., Lee, S.H.: Movin: Real-time motion capture using a single lidar. arXiv preprint arXiv:2309.09314 (2023)
15. Jiang, H., Grauman, K.: Seeing invisible poses: Estimating 3d body pose from egocentric video. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3501–3509 (2017). <https://doi.org/10.1109/CVPR.2017.373>
16. Jiang, T., Lu, P., Zhang, L., Ma, N., Han, R., Lyu, C., Li, Y., Chen, K.: Rtm-pose: Real-time multi-person pose estimation based on mmpose. arXiv preprint arXiv:2303.07399 (2023)
17. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: CVPR (2018)
18. Kaufmann, M., Zhao, Y., Tang, C., Tao, L., Twigg, C., Song, J., Wang, R., Hilliges, O.: Em-pose: 3d human pose estimation from sparse electromagnetic trackers. In: ICCV. pp. 11510–11520 (2021)

19. Kim, W., Ramanagopal, M.S., Barto, C., Yu, M.Y., Rosaen, K., Goumas, N., Vasudevan, R., Johnson-Roberson, M.: Pedx: Benchmark dataset for metric 3-d pose estimation of pedestrians in complex urban intersections. *IRAL* **4**(2), 1940–1947 (2019)
20. Li, J., Zhang, J., Wang, Z., Shen, S., Wen, C., Ma, Y., Xu, L., Yu, J., Wang, C.: Lidarcap: Long-range marker-less 3d human motion capture with lidar point clouds. *arXiv preprint arXiv:2203.14698* (2022)
21. Li, J., Cao, J., Zhang, H., Rempe, D., Kautz, J., Iqbal, U., Yuan, Y.: Genmo: A generalist model for human motion. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 11766–11776 (2025)
22. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: Smpl: A skinned multi-person linear model. *ACM Trans. Graph.* **34**(6), 248:1–248:16 (Oct 2015)
23. Loper, M.M., Mahmood, N., Black, M.J.: MoSh: Motion and shape capture from sparse markers. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)* **33**(6), 220:1–220:13 (Nov 2014). <https://doi.org/10.1145/2661229.2661273>
24. Mahmood, N., Ghorbani, N., Troje, N.F., Pons-Moll, G., Black, M.J.: Amass: Archive of motion capture as surface shapes. In: *ICCV* (October 2019)
25. Malleson, C., Collomosse, J., Hilton, A.: Real-time multi-person motion capture from multi-view video and imus. *IJCV* pp. 1–18 (2019)
26. Mehta, D., Rhodin, H., Casas, D., Fua, P., Sotnychenko, O., Xu, W., Theobalt, C.: Monocular 3d human pose estimation in the wild using improved cnn supervision. In: *3DV* (2017). <https://doi.org/10.1109/3dv.2017.00064>
27. Mehta, D., Rhodin, H., Casas, D., Fua, P., Sotnychenko, O., Xu, W., Theobalt, C.: Monocular 3d human pose estimation in the wild using improved cnn supervision. In: *3DV*. pp. 506–516. *IEEE* (2017)
28. OptiTrack Motion Capture Systems. <https://www.optitrack.com/> (2009), accessed 28 Jun 2026
29. Park, S.I., Hodgins, J.K.: Data-driven modeling of skin and muscle deformation. *ACM Trans. Graph.* **27**(3), 1–6 (Aug 2008). <https://doi.org/10.1145/1360612.1360695>
30. Pons-Moll, G., Fleet, D.J., Rosenhahn, B.: Posebits for monocular human pose estimation. In: *CVPR* (Jun 2014). <https://doi.org/10.1109/CVPR.2014.300>
31. Qiu, H., Wang, C., Wang, J., Wang, N., Zeng, W.: Cross view fusion for 3d human pose estimation. In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. pp. 4341–4350 (2019). <https://doi.org/10.1109/ICCV.2019.00444>
32. Rajasegaran, J., Pavlakos, G., Kanazawa, A., Malik, J.: Tracking people by predicting 3d appearance, location and pose. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 2740–2749 (2022)
33. Raskar, R., Nii, H., deDecker, B., Hashimoto, Y., Summet, J., Moore, D., Zhao, Y., Westhues, J., Dietz, P., Barnwell, J., Nayar, S., Inami, M., Bekaert, P., Noland, M., Branzoi, V., Bruns, E.: Prakash: lighting aware motion capture using photosensing markers and multiplexed illuminators. *ACM Trans. Graph.* **26**(3), 36-es (Jul 2007). <https://doi.org/10.1145/1276377.1276422>
34. Ren, Y., Han, X., Yao, Y., Long, X., Sun, Y., Ma, Y.: Livehps++: Robust and coherent motion capture in dynamic free environment. *ArXiv* **abs/2407.09833** (2024)
35. Ren, Y., Han, X., Zhao, C., Wang, J., Xu, L., Yu, J., Ma, Y.: Livehps: Lidar-based scene-level human pose and shape estimation in free environment. *arXiv preprint arXiv:2402.17171* (2024)

36. Ren, Y., Zhao, C., He, Y., Cong, P., Liang, H., Yu, J., Xu, L., Ma, Y.: Lidar-aid inertial poser: Large-scale human motion capture by sparse inertial and lidar sensors. TVCG (2023)
37. Shen, Z., Pi, H., Xia, Y., Cen, Z., Peng, S., Hu, Z., Bao, H., Hu, R., Zhou, X.: World-grounded human motion recovery via gravity-view coordinates. In: SIGGRAPH Asia Conference Proceedings (2024)
38. Shin, S., Kim, J., Halilaj, E., Black, M.J.: WHAM: Reconstructing world-grounded humans with accurate 3D motion. In: IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR) (Jun 2024)
39. Song, M.H., Godøy, R.I.: How fast is your body motion? determining a sufficient frame rate for an optical motion tracking system using passive markers. PLOS ONE **11**(3) (03 2016)
40. Sun, Y., Bao, Q., Liu, W., Mei, T., Black, M.J.: TRACE: 5D Temporal Regression of Avatars with Dynamic Cameras in 3D Environments. In: IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR) (June 2023)
41. Wang, Y., Sun, Y., Patel, P., Daniilidis, K., Black, M.J., Kocabas, M.: Promptmr: Promptable human mesh recovery. In: Proceedings of the computer vision and pattern recognition conference. pp. 1148–1159 (2025)
42. Wang, Y., Wang, Z., Liu, L., Daniilidis, K.: Tram: Global trajectory and motion of 3d humans from in-the-wild videos. arXiv preprint arXiv:2403.17346 (2024)
43. Wu, X., Zhang, H., Kong, C., Wang, Y., Ju, Y., Zhao, C.: Lidar-based 3-d human pose estimation and action recognition for medical scenes. IEEE Sensors Journal **24**(9), 15531–15539 (2024). <https://doi.org/10.1109/JSEN.2024.3373192>
44. Xin, S., Zhang, Z., Wang, M., Hou, X., Guo, Y., Kang, X., Liu, L., Liu, Y.: Multi-modal 3d human tracking for robots in complex environment with siamese point-video transformer. In: 2024 IEEE International Conference on Robotics and Automation (ICRA). pp. 337–344 (2024). <https://doi.org/10.1109/ICRA57147.2024.10610979>
45. Xu, L., Su, Z., Han, L., Yu, T., Liu, Y., FANG, L.: Unstructuredfusion: Realtime 4d geometry and texture reconstruction using commercialrgbd cameras. TPAMI pp. 1–1 (2019)
46. Xu, L., Xu, W., Golyanik, V., Habermann, M., Fang, L., Theobalt, C.: Eventcap: Monocular 3d capture of high-speed human motions using an event camera. In: CVPR (June 2020)
47. Xu, W., Li, X., Xu, W., Gong, L., Huang, Y., Zhao, Z., Zhao, L., Chen, B., Yang, H., Cao, L., Liu, C.: Human-robot interaction oriented human-in-the-loop real-time motion imitation on a humanoid tri-co robot. In: 2018 3rd International Conference on Advanced Robotics and Mechatronics (ICARM). pp. 781–786 (2018). <https://doi.org/10.1109/ICARM.2018.8610806>
48. Xu, Y., Cong, P., Yao, Y., Chen, R., Hou, Y., Zhu, X., He, X., Yu, J., Ma, Y.: Human-centric scene understanding for 3d large-scale scenarios. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 20349–20359 (2023)
49. Xu, Y., Zhang, J., Zhang, Q., Tao, D.: ViTPose: Simple vision transformer baselines for human pose estimation. In: Advances in Neural Information Processing Systems (2022)
50. Xue, A., Ren, Y., Song, Z., Ye, M., Zhu, X., Ma, Y.: Freecap: hybrid calibration-free motion capture in open environments (2025). <https://doi.org/10.1609/aaai.v39i9.32977>

51. Yan, M., Zhang, Y., Cai, S., Fan, S., Lin, X., Dai, Y., Shen, S., Wen, C., Xu, L., Ma, Y., Wang, C.: Reli1d: A comprehensive multimodal human motion dataset and method. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2250–2262 (June 2024)
52. Ye, V., Pavlakos, G., Malik, J., Kanazawa, A.: Decoupling human and camera motion from videos in the wild. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2023)
53. Yi, X., Zhou, Y., Habermann, M., Shimada, S., Golyanik, V., Theobalt, C., Xu, F.: Physical inertial poser (pip): Physics-aware real-time human motion tracking from sparse inertial sensors. In: CVPR (June 2022)
54. Yi, X., Zhou, Y., Xu, F.: Transpose: Real-time 3d human translation and pose estimation with six inertial sensors. *ACM Transactions on Graphics (TOG)* **40**(4), 1–13 (2021)
55. Yuan, Y., Iqbal, U., Molchanov, P., Kitani, K., Kautz, J.: Glamr: Global occlusion-aware human mesh recovery with dynamic cameras. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
56. Zheng, Z., Yu, T., Li, H., Guo, K., Dai, Q., Fang, L., Liu, Y.: Hybridfusion: Real-time performance capture using a single depth sensor and sparse imus. In: ECCV (Sept 2018)
57. Zhou, Y., Barnes, C., Lu, J., Yang, J., Li, H.: On the continuity of rotation representations in neural networks. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 5738–5746 (2018)