OmniPoser: Flexible Human Motion Recovery in the Wild with Masked Flow Matching

Minghao Liu[1,2] and Tutian Tang[1,2]

[1] Shanghai Jiao Tong University, China
[2] Shanghai Artificial Intelligence Laboratory, China
{lmh209,tttang}@sjtu.edu.cn

Abstract. Recovering 3D human motion from sparse wearable sensors is challenging due to the inherent noise, occlusion, and instability of real-world tracking signals. Existing approaches are restricted to fixed sensor configurations and, when built on diffusion models, are computationally prohibitive for real-time deployment. We propose OmniPoser, a universal framework that recovers full-body SMPL motion from a flexible combination of heterogeneous inputs, such as 3D keypoints, sparse IMU rotations, head-mounted 6-DoF poses, or any subset thereof, within a single model. OmniPoser introduces a cross-modal masking mechanism paired with a Masked Conditioning Encoder, enabling one architecture to incorporate diverse sensor configurations through a masked design. A decoupled dual-stream generative backbone generates complete motion via Conditional Flow Matching, requiring only a single ODE integration step at inference. Across three settings and seven benchmarks, including in-the-wild Nymeria and Ego-Exo4D dataset, OmniPoser performs on par with state-of-the-art methods while running at over 700 FPS. The code will be open-sourced.

Keywords: Human motion recovery · Flow Matching models

1 Introduction

The rapid advancement of AR/VR, immersive gaming, and human-robot interaction has driven a high demand for accurate yet real-time human motion capture. While traditional motion recovery is well-solved in heavily constrained studio environments using dense optical markers, consumer applications require motion capture to operate seamlessly in the wild, relying on limited and imperfect input signals. These intermediate signals can take various forms depending on the hardware ecosystem: 3D keypoints from monocular vision [29, 46], rotations from wearable IMUs [26, 33] or smartwatches [14], or 6D poses from Head-Mounted Displays (HMDs) and hand-held joysticks [12, 38]. Crucially, in real-world scenarios, all these signals are inevitably noisy, fragmented, and unstable. Vision-based 3D keypoints are often corrupted by occlusion and lighting

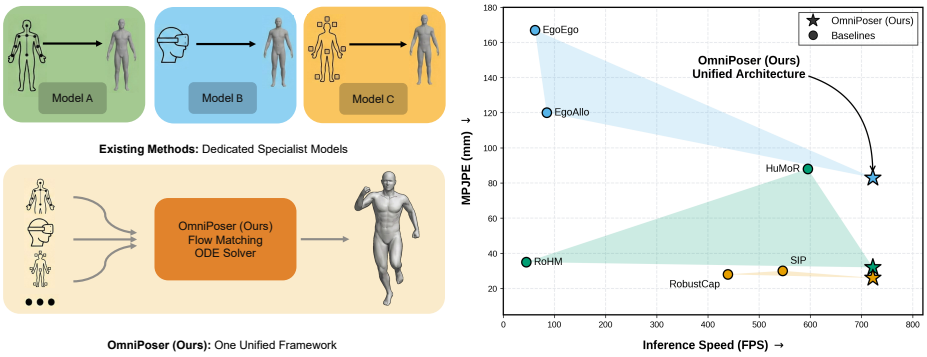

Fig. 1: From dedicated specialist models to one unified framework. *Left:* Existing approaches require a dedicated architecture for each sensor configuration (e.g. partial-body keypoints (RoHM), head-mounted 6-DoF (EgoAllo), sparse IMUs (SIP/RobustCap)), making every new modality a re-engineering effort. OmniPoser incorporates all of them with one flow-matching backbone and training recipe, recovering full-body SMPL motion from *different* sensor configurations via the msaked design. *Right:* Accuracy–FPS performance across three settings. Each baseline (circle) is plotted; OmniPoser (star) matches or surpasses baseline methods at >700 FPS with lower Mean Per Joint Position Error (MPJPE).

conditions, while rotation readings from sparse sensors may drift by nature or even disconnect due to wireless interference.

To recover human motion in the wild with such inputs, previous works have transitioned from deterministic inverse kinematics regressors [10, 17] to data-driven generative models. Notably, diffusion models [35, 46] have been employed to iteratively denoise and infer plausible poses under severe occlusions or sparse node conditions.

However, these methods expose two limitations. First, they are typically designed for a predefined configuration of inputs—such as a fixed combination of keypoints or rotations from a specific set of sensors. This inflexibility requires the networks to be entirely re-trained for different sensor configurations, and also makes them difficult to handle the unstable nature of real-world signals. Second, they are usually too computationally expensive and thus unsuitable for real-time applications on consumer-grade mobile devices.

In this work, we propose OmniPoser, a universal framework that recovers human motion from a flexible combination of noisy and unstable signals. OmniPoser employs a Conditional Flow Matching backbone to enable highly efficient motion generation. In order to support flexible inputs, we introduce a cross-modal masking mechanism that unifies heterogeneous sensor configurations within one model. Zeroing out missing entries, however, destroys the identity of unobserved joints and leads to poor recovery (see ablation in Sec. 4.5). We therefore design a Masked Conditioning Encoder (MCE) that replaces each missing joint with a learnable per-joint embedding, allowing the network to reason about *which* joints

are absent and recover them accordingly. At inference, a single Euler integration step maps Gaussian noise onto the motion manifold conditioned on the masked observations, enabling real-time motion recovery.

We evaluate OmniPoser across three different settings, each targeting a very distinct sensor configuration, and compare against state-of-the-art methods. Altogether we evaluate our method quantitatively on five benchmarks [9,18,23,24, 34]. Results show our method performs on par with state-of-the-art methods, while maintaining a processing speed over 700 FPS, as illustrated in Fig. 1.

The main contributions of this paper can be summarized as:

(1) We propose OmniPoser, a Conditional Flow Matching-based framework for human motion recovery that accepts heterogeneous sparse observations, such as 3D keypoint positions, joint rotations, and their combinations, as input.

(2) We introduce a cross-modal masking mechanism paired with a Masked Conditioning Encoder (MCE) that replaces unobserved joints with learnable per-joint embeddings, enabling the network to distinguish which joints are missing rather than treating all of them identically.

(3) We demonstrate competitive performance compared to state-of-the-art methods and real-time inference capabilities across diverse configurations and highly challenging in-the-wild datasets.

2 Related Work

Our method is closely related to human motion capture with sparse sensors, general human motion recovery, and Flow Matching models for human kinematics.

Human Motion Capture with Sparse Sensors. Wearable human motion capture systems offer a portable alternative to constrained optical systems in studios. Existing approaches generally fall into three hardware paradigms: IMU-based systems, head-mounted cameras, and VR devices with sparse sensors. (1) **IMU-based Systems:** While commercial solutions require dense sensor arrays, recent research successfully reconstructs motion from sparse setups, such as 6 IMUs (e.g., TransPose [43], PIP [42] and PNP [42]) or even as few as 4 IMUs (e.g., LIP [30]). Under fixed sensor placements, SAIP [44] incorporates shape-aware motion tracking tailored to real body shape, and GlobalPose [41] utilizes physics priors to refine global tracking. For a more flexible paradigm, generative frameworks like DiffusionPoser [35] achieve similar body motion recovery while supporting a flexible combinations of IMU signals. Conversely, to fundamentally counteract inherent integration drift via explicit geometric guidance, inertial-visual fusion [26] remains essential. Most recently, Stereo-Inertial Poser [33] has been proposed to utilize a single stereo camera and six IMUs for metric-accurate, shape-aware motion capture. (2) **Head-Mounted Cameras (HMC):** Reconstructing full-body poses from egocentric vision suffers from severe lower-body self-occlusion. Systems like EgoEgo [15] and EgoExo [19] utilize SLAM and hand poses to condition full-body generation. More recently, EgoAllo [40] utilized allocentric grounding and conditional diffusion to eliminate floating artifacts, while HMD2 [6] further refined egocentric motion estimation. (3) **VR Devices and**

Sparse Sensors: Inside-out tracking in mixed reality typically uses an HMD and two controllers. Approaches like QuestSim [38] use reinforcement learning to simulate physics-based full-body avatars from these inputs. AvatarPoser [12], EgoPoser [11], and MANIKIN [10] utilize deep kinematics and optimization for highly accurate tracking, occasionally fusing supplementary IMUs [3] or smartwatches [14]. Despite these advancements, the majority of these tracking methods remain structurally rigid and tailored to a single, fixed sensor configuration. Despite these advancements, the majority of these methods remain structurally rigid and tailored to a single, fixed sensor configuration.

Human Motion Recovery. Human motion recovery from monocular videos typically follows a two-stage pipeline. First, off-the-shelf pose detectors [1,36,39] extract spatial keypoints. Second, a sequence modeling network processes these spatiotemporal coordinates to compute 3D motion over time. Early generative approaches like HuMoR [29] utilized conditional variational autoencoders to optimize motion priors. Recently, Denoising Diffusion Probabilistic Models [8] have dominated this field. RoHM [46] reconstructs globally coherent motions by decomposing the task into global and local diffusion models. It can solves different patterns of keypoints combinations (e.g., upper-body v.s. lower-body), so it serves as our primary baseline. Multi-hypothesis frameworks like CHAMP [45] and Platypose [27] robustly handle uncertainty but they also lack topological flexibility. To address in-the-wild complexities, HumanMM [47] uses multi-shot alignment to ensure global motion continuity across dynamic camera transitions. A recent concurrent work, MoRo [28], explores masked modeling to achieve occlusion-robust recovery from monocular inputs. Overall, existing frameworks are insufficiently universal. With focus on monocular images, they heavily rely on the spatial keypoints produced by pose detectors in the first stage. In contrast, our proposed method natively infers human motion from an arbitrary combination of 3D keypoints and sparse IMU rotations.

Masked Motion Modeling and Decomposition. Techniques like masked modeling and local-global motion decomposition have been iteratively refined into mature paradigms within the HMR community. Specifically, PromptHMR [37] leverages spatial and semantic prompts to robustly guide human shape recovery. MEGA [4] addresses visual ambiguity by formulating mesh recovery as a feature completion task via a masked generative autoencoder. GENMO [16] unifies generation and estimation by treating variable conditioning signals as explicitly masked inputs alongside a joint local-global motion representation. EgoMDM [31] embeds local-to-global translation constraints into a diffusion framework to synthesize motion from sparse egocentric sensors. Additionally, PUMA [13] optimizes the underlying generative training dynamics through progressive unmasking schedules. We embrace these established designs in our work, and adapt them to heterogeneous sparse-observation recovery. We expand the mask from a dropout training schedule into a masked module that unifies different sensor modalities, dictating whether a joint is observed via a 3D position, a rotation, a 6-DoF signal.

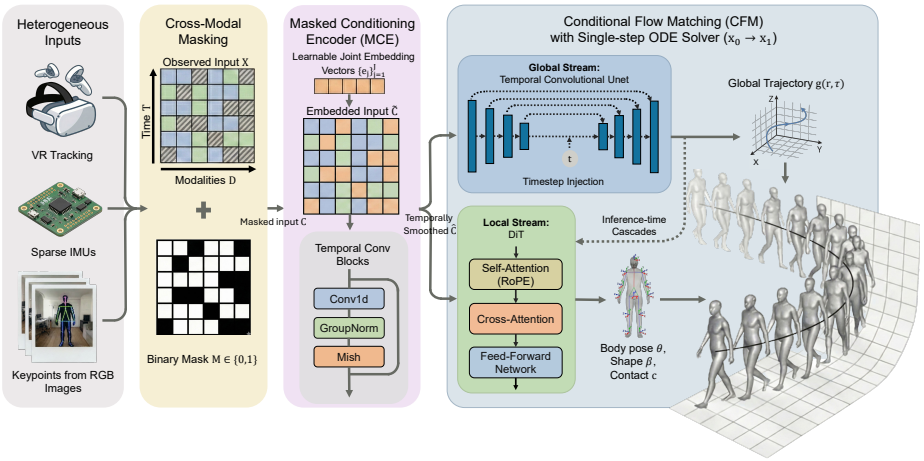

Fig. 2: Overview of OmniPoser. Heterogeneous observations (multi-modal 3D keypoints and rotations) are unified through a cross-modal masking mechanism and processed by Masked Conditioning Encoder (MCE) (Sec. 3.2). A decoupled dual-stream architecture(Sec. 3.3) (a temporal convolutional UNet for global trajectory and a cross-attention Diffusion Transformer (DiT) for local body pose) generates SMPL parameters via Conditional Flow Matching(Sec. 3.3), achieving real-time inference in a single ODE step. Best viewed in color.

3 Method

Recovering 3D human motion from sparse, heterogeneous, and occluded observations remains fundamentally ill-posed. Different sensor modalities (e.g. sparse IMU readings, head-mounted egocentric cameras, monocular RGB detectors, optical trackers, or partial motion capture) provide only a narrow and noisy window into the full-body state. Our key insight is that regardless of sensor modality, all observations can be cast as *partial views* of the same canonical body representation, and the completion of unobserved degrees of freedom can be formulated as a unified generative conditioning problem. In the following, we first define the unified motion representation (Sec. 3.1). We then introduce the cross-modal masking mechanism and the Masked Conditioning Encoder (MCE) that converts masked observations into a joint-aware conditioning signal (Sec. 3.2). Finally, we present the Conditional Flow Matching formulation, the decoupled generative architecture, and the training objectives (Sec. 3.3).

3.1 Motion Representation

We represent a motion clip of T frames as a time series $\mathbf{X} \in \mathbb{R}^{T \times D}$, where each row $\mathbf{x}^{(t)} \in \mathbb{R}^D$ is a concatenation of semantically distinct *feature groups*, describing a different aspect of the body state at time t. D is the sum of each feature group's dimension. In this work, we use feature groups based on the SMPL [22] model, including the global root trajectory (root translation $\boldsymbol{r}$ and

root rotation $\boldsymbol{\tau}$), 3D joint positions $\mathbf{J} \in \mathbb{R}^{T\times J\times 3}$ and 6D rotations $\mathbf{R}^{\mathbf{w}}$ or $\mathbf{R}^{\mathbf{r}} \in \mathbb{R}^{T\times J\times 6}$ (in the world frame or in the root frame), body shape parameters $\boldsymbol{\beta} \in \mathbb{R}^{10}$, and binary foot contact labels $\mathbf{c}$. These features are concatenated to form the complete motion representation $\mathbf{x}^{(t)}$ at each frame, serving as both the partial input and the fully recovered output. The full representation vector is thus $\mathbf{x}^{(t)} = [\boldsymbol{r}; \boldsymbol{\tau}; \mathrm{vec}(\mathbf{J}); \mathrm{vec}(\mathbf{R}^{\mathrm{w}}); \mathrm{vec}(\mathbf{R}^{\mathrm{r}}); \boldsymbol{\beta}; \mathbf{c}] \in \mathbb{R}^{D}$. This representation serves as both input and output. On the input side, a subset of feature groups and joints are observed, depending on the sensor configuration. On the output side, the model recovers all feature groups for all joints, where joint rotations are expressed in the local (parent-relative) frame as the standard SMPL body pose.

3.2 Cross-Modal Masking and Masked Conditioning Encoder

To unify heterogeneous sensor inputs within one model, we design a configurable *mask scheme* that specifies, for a given sensor configuration, which feature groups and which joints are observed. Concretely, a mask scheme defines: (i) a set of *visible joint indices* $\mathcal{V} \subseteq \{1, \ldots, J\}$ controlling per-joint masking for *joint-based* feature groups (e.g., local positions, per-joint rotations), and (ii) visibility flags for *global* feature groups (e.g., whether root translation or orientation is observed). Given a mask scheme, we construct a binary feature-level mask $\mathbf{M} \in \{0, 1\}^{T\times D}$ that zeros out unobserved features in the conditioning tensor. However, zero-padding removes the identity of missing joints; therefore, the network cannot distinguish *which* joints are absent, leading to large reconstruction errors (Sec. 4.5). We address this with the Masked Conditioning Encoder described below.

For a feature group, the mask visibility may vary along the temporal axis, enabling the representation of intermittent sensor dropout, faithfully mimicking the signal interruptions common in real-world wearable or markerless capture systems. This design enables one model to incorporate a wide range of input configurations, from optical motion capture ($|\mathcal{V}|$=22, rotations visible), to 5-point VR tracking ($\mathcal{V}$={head, hands, feet}, positions only), to egocentric head-only estimation ($\mathcal{V}$={head}) through our masked design.

During training, mask samples are drawn from several configurations. This includes predefined sensor layouts such as VR/AR tracking, HMD with controllers, sparse keypoints from RGB images, and full IMU suits. It also includes data-driven visibility patterns: we estimate per-joint visibility probabilities from the large-scale, in-the-wild Ego-Exo4D dataset [5] using its COCO [20] pose annotations predicted by MMPose [2]. These real-world occlusion statistics mirror the inputs produced by upstream pose-estimation systems in practice. In Sec. 4.4, we show that this training strategy enables OmniPoser to generalize to in-the-wild scenarios, including ego-centric activities from Ego-Exo4D that were unseen during training.

Masked Conditioning Encoder While the mask scheme defines *what* is observed, the network must also know *which specific joints* are missing in order to

recover them. Zero-padding treats all absent joints identically, discarding their positional identity. To solve this, we introduce a *Masked Conditioning Encoder* (MCE) that converts the raw masked conditioning signal $\mathbf{C} \in \mathbb{R}^{T \times D}$ into a joint-aware representation. Inspired by Masked Autoencoders (MAE) [7], the MCE maintains a bank of J learnable *joint embedding vectors* $\{\mathbf{e}_j\}_{j=1}^{J}$, each matching the per-joint feature dimension. At each timestep t, unobserved entries are replaced by the corresponding learnable embedding:

$$\tilde{\mathbf{c}}_j^{(t)} = \begin{cases} \mathbf{c}_j^{(t)} & \text{if } j \in \mathcal{V}^{(t)}, \\ \mathbf{e}_j & \text{otherwise,} \end{cases} \tag{1}$$

where $\mathcal{V}^{(t)}$ denotes the set of visible joints at time t (which may vary per timestep in probabilistic masking settings). The masked-and-embedded tensor is then processed by a stack of $K = 4$ residual temporal convolution blocks, each followed by Group Normalization and Mish activations, producing a temporally smoothed output $\hat{\mathbf{C}} \in \mathbb{R}^{T \times D}$ that serves as the conditioning signal for the generative backbone. The learnable embeddings provide a fill-in signal richer than zero-padding. This design lets the model specialize during training to encode per-joint positional priors. Then temporal convolutions interpolate short-duration occlusion gaps before the conditioning enters the backbone.

3.3 Generative Architecture and Training

Conditional Flow Matching. The conditioned signal $\hat{\mathbf{C}}$ produced by the MCE is fed into a generative model based on Conditional Flow Matching (CFM) [21]. Given a ground-truth sample $\mathbf{x}_1$ and Gaussian noise $\mathbf{x}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, a conditional optimal-transport (OT) affine path constructs intermediate states

$$\mathbf{x}_t = \alpha_t \, \mathbf{x}_1 + \sigma_t \, \mathbf{x}_0, \tag{2}$$

where (α_t, σ_t) are the linear interpolation coefficients of the OT path. A denoising network v_θ is trained to predict the clean target $\mathbf{x}_1$ from $\mathbf{x}_t$ and the conditioning $\hat{\mathbf{C}}$, supervised by the squared ℓ_2 error $\|v_\theta(\mathbf{x}_t, t, \hat{\mathbf{C}}) - \mathbf{x}_1\|_2^2$. At inference, a single Euler integration step from t=0 to t=1 maps Gaussian noise onto the SMPL motion manifold.

Decoupled Architecture A complete human motion can be split into global trajectory and local body pose, however, jointly predicting them from partially observed inputs is challenging because they live in different spaces: global trajectory is low-DoF (degrees of freedom) but must exhibit long-horizon coherence (drift-free), while local body pose is high-DoF but primarily needs to meet biomechanical constraints. To address this, the denoising network v_θ are two streams that are trained independently yet share the same MCE conditioning interface: a *global stream* that generates the root orientation $\boldsymbol{r}$ and translation $\boldsymbol{\tau}$, and a *local stream* that generates body pose $\boldsymbol{\theta}$, shape $\boldsymbol{\beta}$, and foot contact $\mathbf{c}$. Each stream's backbone is chosen to match its output characteristics: a convolutional UNet

with multi-scale skip connections for the low-dimensional global trajectory, and a Transformer-based Diffusion Transformer (DiT) employing Rotary Position Embedding (RoPE) [32] and cross-attention to $\hat{\mathbf{C}}$ for the high-dimensional local body pose.

During training, the local stream receives ground-truth trajectory as part of its conditioning. At inference, the two streams are composed: the global stream's predicted trajectory is fed into the local stream as conditioning, enabling cascaded generation where global trajectory informs local pose recovery.

Training Objectives Beyond the flow matching loss $\mathcal{L}_{\text{FM}}$, we include auxiliary losses that operate on the decoded SMPL parameters. The predicted motion representation $\hat{\mathbf{X}}$ is passed through SMPL forward kinematics to give reconstructed 3D joint positions $\hat{\mathbf{J}} \in \mathbb{R}^{T\times J\times 3}$. We summarize the loss items as follows. For detailed definitions, please refer to the supplementary materials.

Local Stream Losses. The local stream training objective combines flow matching with representation, kinematic, and physics-grounded regularization:

$$\begin{aligned}\mathcal{L}_{\text{pose}} = \mathcal{L}_{\text{FM}} &+ \lambda_{\text{repr}}\mathcal{L}_{\text{repr}} + \lambda_{\text{jpos}}\mathcal{L}_{\text{jpos}} \\ &+ \lambda_{\text{vjpos}}\mathcal{L}_{\text{vjpos}} + \lambda_{\text{jvel}}\mathcal{L}_{\text{jvel}} + \lambda_{\text{smooth}}\mathcal{L}_{\text{smooth}} \\ &+ \lambda_{\text{fc}}\mathcal{L}_{\text{fc}} + \lambda_{\text{skate}}\mathcal{L}_{\text{skate}}\end{aligned} \tag{3}$$

Here $\mathcal{L}_{\text{repr}}$ penalizes representation error; $\mathcal{L}_{\text{jpos}}=\|\hat{\mathbf{J}} - \mathbf{J}^{\text{GT}}\|_2^2$ and $\mathcal{L}_{\text{jvel}}=\|\Delta_t\hat{\mathbf{J}} - \Delta_t\mathbf{J}^{\text{GT}}\|_2^2$ supervise joint positions and velocities after SMPL forward kinematics; $\mathcal{L}_{\text{smooth}}=\|\Delta_t^2\hat{\mathbf{J}}\|_2^2$ penalizes acceleration for temporal smoothness; $\mathcal{L}_{\text{fc}}$ supervises foot-contact labels; and $\mathcal{L}_{\text{skate}}$ penalizes foot sliding when the predicted contact label indicates ground contact but the joint velocity exceeds a threshold.

Global Stream Losses. The global stream is supervised by a similar set of physics-grounded losses:

$$\begin{aligned}\mathcal{L}_{\text{traj}} = \mathcal{L}_{\text{FM}} &+ \lambda_{\text{rec}}\mathcal{L}_{\text{rec}} + \lambda_{\text{rpos}}\mathcal{L}_{\text{rpos}} \\ &+ \lambda_{\text{rvel}}\mathcal{L}_{\text{rvel}} + \lambda_{\text{rsmooth}}\mathcal{L}_{\text{rsmooth}}\end{aligned} \tag{4}$$

The global stream is supervised similarly by $\mathcal{L}_{\text{traj}} = \mathcal{L}_{\text{FM}}+\lambda_{\text{rec}}\mathcal{L}_{\text{rec}}+\lambda_{\text{rpos}}\mathcal{L}_{\text{rpos}}+\lambda_{\text{rvel}}\mathcal{L}_{\text{rvel}}+\lambda_{\text{rsmooth}}\mathcal{L}_{\text{rsmooth}}$, where $\mathcal{L}_{\text{rpos}}=\|\hat{\mathbf{J}}_0-\mathbf{J}_0^{\text{GT}}\|_2^2$, $\mathcal{L}_{\text{rvel}}=\|\Delta_t\hat{\mathbf{J}}_0-\Delta_t\mathbf{J}_0^{\text{GT}}\|_2^2$, and $\mathcal{L}_{\text{rsmooth}}=\|\Delta_t^2\hat{\mathbf{J}}_0\|_2^2$ are root position, velocity, and acceleration losses ($\mathbf{J}_0$: pelvis joint), and $\mathcal{L}_{\text{rec}}$ is the trajectory representation MSE.

4 Experiments

In this section, we validate three claims: (i) OmniPoser performs on par with state-of-the-art methods across heterogeneous sparse-observation settings, (ii) Flow Matching enables diversity in generation, and (iii) cross-modal masking mechanism and Masked Conditioning Encoder is essential for generalization.

4.1 Implementation Details

Networks are optimized independently with AdamW (β_1=0.9, β_2=0.999, wd = 0.1, lr = 2×10^{-4}) using warmup-cosine scheduling (4K-step warmup) for 480K steps at batch size 256 with gradient clipping at norm 1. Motion clips are sampled at 30 fps (T=33 frames). All training is conducted on a single NVIDIA A40 (48 GB) and converges in ∼3 days. At inference, we stick with 1 Euler ODE step, yielding >16× speedup over the 100+ diffusion steps of baselines (e.g., RoHM [46]).

For more details, please refer to the supplementary material.

4.2 Evaluation Protocols

Benchmarks and Baseline Methods To demonstrate the unified capability of OmniPoser as discussed in Sec. 1, we categorize different sensor configurations and baseline methods into four settings.

Setting 1 (Sparse Observation): Only a sparse subset of joints is observed. We evaluate two schemes: (1) upper-body joint positions, evaluated on AMASS [24] and Nymeria [23]; and (2) a random joint masking regime, evaluated on AMASS. The baseline methods are RoHM [46] and HuMoR [29]. AMASS evaluations use AMASS training, and Nymeria evaluation uses Nymeria training.

Setting 2 (Egocentric Head 6D Pose): Only head position and rotation are available from HMD signals. We compare against EgoEgo [15] and EgoAllo [40] across AMASS, RICH [9], and Nymeria. AMASS and RICH evaluations use AMASS training, and Nymeria evaluation uses Nymeria training.

Setting 3 (Mocap data from 6 IMUs with keypoint positions): Six IMUs provide rotations at root and leaf joints, while upstream methods provide keypoint positions. We evaluate on AIST++ [18], TotalCapture [34], and Nymeria, comparing with Stereo-Inertial Poser (SIP) [33] and RobustCap [26]. AIST++ uses AMASS+AIST++ training, TotalCapture uses AMASS training, and Nymeria uses Nymeria training.

Setting 4 (Real-world Evaluation): We present qualitative visualization results on 3DPW [25] and Ego-Exo4D [5] to show performance in unconstrained outdoor and ego-centric scenes. We use one cross-setting checkpoint to generate 3DPW and Ego-Exo4D examples.

For dataset usage, we follow the standard splits: AMASS uses 15 sub-datasets for training and held-out sets for testing; Nymeria uses an 85:15 activity-stratified split; and RICH follows the EgoAllo split.

For fair comparison, we train two versions of OmniPoser during evaluation: during **Ours-S** training, mask samples are drawn from the corresponding setting; during **Ours-U** training, mask samples are drawn from the masking configurations stated in 3.2, i.e., predefined sensor layouts and real-world observability patterns. Since there are no unified baselines at present, we construct RoHM-U and RoHM++-U, which are based on current specialist SOTA, RoHM. RoHM-U retains the original RoHM architecture, but the mask samples drawn during training are identical to Ours-U. Meanwhile, we upgrade RoHM-U's architecture

into RoHM++-U, stacking three setting-specific (namely, Setting 1,2,3) input encoders before the network. Baseline rows without **-U** indicate per-setting specialist models unless noted otherwise. Bold and underlined values indicate the best and second-best results, respectively.

For Nymeria, dashes denote unavailable metrics. Nymeria dataset contains 260M frames with dense body MVNX annotations. Due to practical resource constraints throughout data preprocessing and large-scale training, we leave this for future work.

Metrics We report MPJPE and PA-MPJPE (mm), Acceleration Error (m/s^2), Jerk (km/s^3), Per-vertex Error (mm), Contact Accuracy, Foot Skating (FS, mm), Ground Penetration (mm), Translation Error (T_{head}, mm), and Hips & Shoulders Rotation Error (deg). In all tables in this section, $\downarrow$ means lower is better, while $\uparrow$ means higher is better.

4.3 Quantitative Results

Setting 1: Sparse Observation

Setting 1: Sparse Observation (Upper-body joint positions). As introduced previously, this setting is intrinsically under-constrained because the entire lower body is unobserved. Tab. 1 presents results for the upper-body setting. On AMASS, OmniPoser achieves 32.1 mm MPJPE, outperforming both RoHM (34.8 mm) and HuMoR (88.0 mm). HuMoR relies on expensive test-time optimization using a CVAE prior, struggling heavily with complete lower-body absence. RoHM (or RoHM++-U) addresses this using two cascaded conditional diffusion models explicitly trained for this masking pattern. In contrast, Ours-S and Ours-U's performance are comparable to counterpart models, and surpasses at some of the metrics. We attribute this to the Masked Conditioning Encoder (MCE), which uses learnable embeddings to encode global structural priors for unobserved joints. Furthermore, our model achieves the lowest skating ratio (0.031). This shows that our foot-skating loss (e.g., $\mathcal{L}_{\text{skate}}$) generates valid foot-ground interactions even for invisible observations, at a fraction of the computational cost (1 ODE step vs. 100+ diffusion steps). On Nymeria, our model reduces MPJPE by ∼64% over RoHM (21.1 vs. 59.3 mm), alongside lower acceleration error, proving our learned priors transfer robustly in the wild.

Setting 1: Sparse Observation (Random joint masking). Tab. 2 evaluates a random masking regime where up to 15 out of 22 joints are randomly masked. This setting explicitly tests the model's ability to handle the unstable signals typical of real-world sensors (as discussed in Sec. 1). Ours-S achieves 35.7 mm MPJPE. This advantage comes from our random masked training and the MCE's ability to preseve joint identity. Ours-U's slight worse metrics are acceptable, considering that our method does not introduce hand-crafted modules that differentiate across settings.

Table 1: Quantitative results of Setting 1: Sparse Observation (Upper-body joint positions).

	AMASS					Nymeria (In-the-wild)			
Method	JPE↓	Acc↓	Con↑	Ska↓	Pen↓	JPE↓	Acc↓	Con↑	Ska↓
HuMoR	88.0	3.3	0.68	0.230	2.59	88.5	8.26	0.78	0.685
RoHM	34.8	**2.3**	**0.95**	0.078	**0.69**	59.3	7.2	0.91	0.610
Ours-S	**32.1**	2.7	0.94	**0.031**	1.84	**21.1**	**1.39**	**0.95**	**0.033**
RoHM-U	158.7	2.9	0.75	0.190	1.48	-	-	-	-
RoHM++-U	39.9	**2.3**	**0.95**	0.081	**0.56**	-	-	-	-
Ours-U	**36.7**	2.7	**0.95**	**0.025**	9.32	-	-	-	-

Table 2: Quantitative results of Setting 1: Sparse Observation (Random joint masking) on AMASS.

Metric	HuMoR	RoHM	**Ours-S**	**Ours-U**
JPE (mm) ↓	80.5	38.2	**35.7**	49.3
Accel. (m/s^2) ↓	3.3	**1.9**	2.2	2.6
Contact ↑	0.78	0.95	**0.97**	0.93
Skating ↓	0.184	0.024	**0.018**	0.019

Setting 2: Egocentric Head 6D Pose Only head position and rotation are available (HMD signal)—a longstanding ill-posed problem where the entire body below the neck must be generated from a single 6-DoF observation. We compare against EgoEgo [15] and EgoAllo [40].

Tab. 3 presents results across three benchmarks. On AMASS, Ours-S achieves 83.2 mm MPJPE, Ours-U achieve 86.2mm MPJPE, (∼30% improvement over EgoAllo's 119.7 mm and 50% over EgoEgo's 167.4 mm). EgoEgo couples prediction of global trajectory and local motion, which makes it perform not as well as our decoupled structure. EgoAllo uses special allocentric grounding and conditional diffusion designed purely for head-mounted setups, yielding accurate head alignment (6.2 mm vs. our 17.1 mm). We consider our slightly higher head error a trade-off for OmniPoser's flexibility. Without hard-coding the head pose, our DiT's cross-attention propagates the 6-DoF HMD signal globally, allowing our method to outperform those baselines on full-body recovery. On RICH, Ours-S achieves a 25% improvement in MPJPE (131.6 mm) compared to specialist methods. Ours-U exhibits a different set of trade-offs compared to RoHM++-U. The lower contact accuracy (0.71) is a shared degradation across all methods due to inherent floor-height ambiguities in RICH unseen during AMASS training. On Nymeria, OmniPoser reduces MPJPE by ∼60% over EgoAllo (40.7 vs. 103.0 mm). Its superior contact accuracy (0.94 vs. EgoAllo's 0.63) highlights that our physics-grounded regularization term ensures stability even for unconstrained ego-centric capture.

Setting 3: Mocap Data from 6 IMUs with Keypoint Positions Six IMUs at root and leaf joints (pelvis, head, wrists, ankles) provide rotations, while up-

Table 3: Quantitative results of Setting 2: Egocentric Head 6D Pose.

	AMASS				RICH (Cross-domain)				Nymeria (Wild)		
Method	JPE↓	PA↓	Con↑	T_h ↓	JPE↓	PA↓	Con↑	T_h ↓	JPE↓	PA↓	Con↑
EgoEgo	167.4	145.8	0.92	54.9	207.8	187.8	0.88	66.5	135.8	117.2	0.57
EgoAllo	119.7	101.1	**1.00**	**6.2**	176.2	160.1	**0.96**	**8.9**	103.0	89.3	0.63
Ours-S	**83.2**	**73.5**	0.92	17.1	**131.6**	**111.5**	0.71	25.7	**40.7**	**36.2**	**0.94**
RoHM-U	124.3	101.5	0.94	36.7	195.0	115.4	0.62	56.8	-	-	-
RoHM++-U	94.5	92.5	**0.96**	**19.5**	**126.3**	112.3	**0.72**	**44.8**	-	-	-
Ours-U	**86.2**	**69.8**	0.92	33.2	134.8	**106.4**	0.65	50.7	-	-	-

stream methods (i.e. pose estimation results from RGB images, odometry sensor readouts) provide keypoint positions. We compare with Stereo-Inertial Poser (SIP) [33] and RobustCap [26]. We note that the original RobustCap leverages 33 MediaPipe 2D keypoints from an external camera (as camera-normalized marching rays), while OmniPoser uses IMU orientations together with 3D joint positions in the inertial frame. To ensure a fair comparison, we enhance and retrain RobustCap by replacing its 2D keypoint inputs with 3D keypoints matching OmniPoser's input representation, denoted **RobustCap3D**; this upgrade should improve its global translation estimation while providing a controlled ablation of input modality.

AIST++. On the highly dynamic AIST++ dataset, Ours-S / Ours-U achieves the best PA-MPJPE (19.4 mm / 21.4 mm) and MPJPE (26.0 mm / 29.3 mm). Crucially, the Ours-S's jerk metric (0.29 km/s^3) is 5× lower than SIP and 8× lower than RobustCap. Unlike RobustCap, which directly regresses visual and inertial cues leading to high-frequency jitter, OmniPoser leverages a decoupled generative architecture. The temporal convolutional UNet smoothly filters the low-frequency global trajectory, which then conditions the DiT to suppress local kinematic jitter, assisted strongly by our velocity and smoothness losses.

TotalCapture. On the TotalCapture benchmark, Ours-S outperforms RobustCap3D in MPJPE (36.2 vs. 42.3 mm), and Ours-U stays competitive with RoHM++-U across metrics. Logically, SIP achieves superior global translation (5.42 vs. 15.9 mm) and hip/shoulder rotation (10.6° vs. 13.3°) due to the fact that it extracts absolute metric depth from stereo cameras and directly substitutes physical pelvis IMU readings for root orientation. In contrast, OmniPoser strictly *generates* global trajectory and local rotations from masked observations using general flow matching. The slight performance gap in rigid root translation is a conscious trade-off to preserve a purely modality-agnostic formulation.

Nymeria. On Nymeria, OmniPoser's lower MPJPE (16.6 mm) and jerk (0.048 m/s^3) reinforce that our cross-modal masking conditioning effectively learns unified kinematic priors, avoiding the need to tune multi-modal sensor fusion heuristics for real-world deployment.

Table 4: Quantitative results of Setting 3: Mocap Data from 6 IMUs with Keypoint Positions.

	AIST++							TotalCapture							Nymeria (In-the-wild)						
Method	PA↓	JPE↓	PVE↓	TE↓	HS↓	Jrk↓	FS↓	PA↓	JPE↓	PVE↓	TE↓	HS↓	Jrk↓	FS↓	PA↓	JPE↓	PVE↓	TE↓	HS↓	Jrk↓	FS↓
SIP	23.1	29.8	39.0	**11.31**	8.91	1.50	0.57	28.6	46.5	54.3	**5.42**	**10.6**	0.94	**0.69**	15.2	17.0	16.8	**4.13**	**5.22**	0.059	**0.25**
RobCap	24.0	33.1	43.2	19.92	9.34	2.49	0.61	33.5	48.7	63.4	23.5	13.4	2.09	1.40	16.5	19.5	19.4	4.52	7.51	1.56	0.54
RobCap3D	21.8	28.5	37.6	15.82	8.95	2.31	0.59	27.9	42.3	52.8	9.74	12.8	1.87	1.22	15.0	17.8	17.5	4.25	7.18	1.41	0.48
Ours-S	**19.4**	**26.0**	**32.9**	16.54	**7.40**	**0.29**	**0.22**	**26.5**	**36.2**	**51.2**	15.9	13.3	**0.16**	0.81	**14.5**	**16.6**	**16.6**	6.83	7.50	**0.048**	0.39
RoHM-U	123.5	157.5	185.9	76.3	38.1	**0.21**	9.49	78.6	95.3	100.4	22.8	19.5	0.19	0.95				-			
RoHM++-U	89.5	136.5	128.5	56.1	32.1	**0.21**	9.51	**18.7**	**29.0**	**26.3**	17.5	**8.3**	**0.07**	2.06				-			
Ours-U	**21.4**	**29.3**	**40.4**	**17.5**	**7.9**	1.73	**1.05**	23.4	35.5	47.2	**15.7**	11.1	0.15	**0.86**				-			

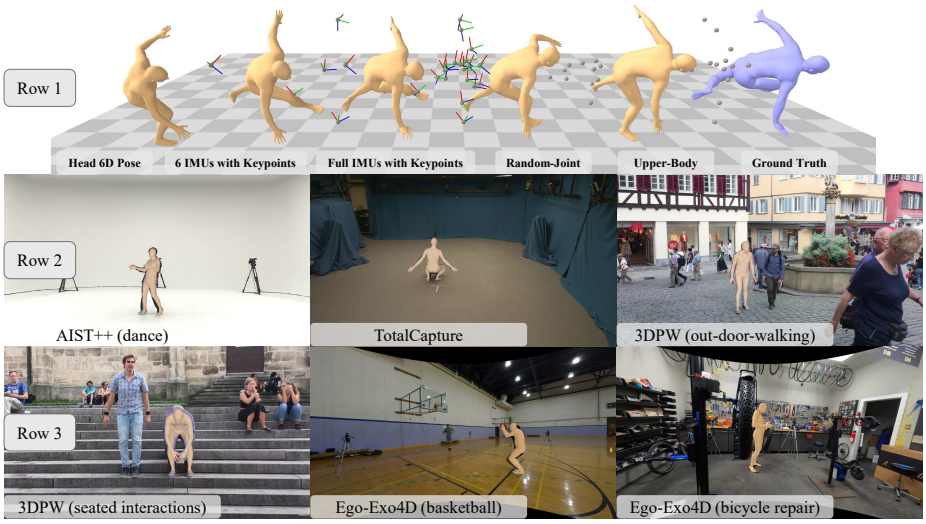

Fig. 3: Qualitative results. *Row 1:* Five masking configurations applied to the same speed-vault sequence from AMASS dataset. Meshes in gold are the predictions. The mesh on the rightmost is the ground truth. *Row 2–3:* Visualizations on different datasets. For more qualitative results, please refer to the supplementary material.

4.4 Qualitative Results

Importantly, all qualitative results in this section are produced by Ours-U model 4.2. In the first row in Fig. 3, we demonstrate this capability: given a speed-vault clip, one model produces plausible reconstructions under five distinct masking configurations at inference time. Configurations with spatially distributed observations (6 IMUs with keypoints, full IMUs with keypoints) most faithfully track the ground truth, while the head 6d pose setting anchors the head 6-DoF pose and recovers a plausible full-body pose. The upper-body setting, lacking any lower-body supervision, produces a less dynamic but still coherent trajectory. Fig. 3 (Row 2–3) shows faithful recovery on dynamic AIST++ dances and tight alignment on TotalCapture. On 3DPW, OmniPoser produces coherent walking gaits and seated interactions in outdoor scenes, validating that kinematic pri-

Table 5: Ablation studies (AMASS, upper-body).

Variant	JPE↓	Acc↓	Ska↓
OmniPoser (full)	**32.1**	**2.7**	**0.031**
MCE → zero-padding	88.1	4.0	0.053

ors learned from controlled mocap dataset generalize to in-the-wild domains. On Ego-Exo4D, OmniPoser recovers whole-body motion from ego-centric basketball and bicycle-repair activities involving complex body-object interactions, showing our method's applicability to AR/VR capture scene.

4.5 Ablation Studies

We ablate OmniPoser from two aspects: the effectiveness of MCE module, and the sampling diversity from our generative conditional network.

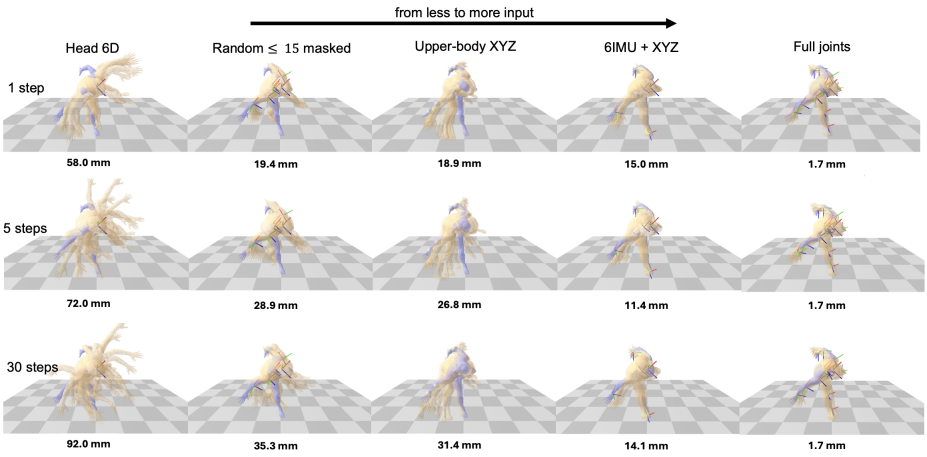

Fig. 4: Diversity under input constraints and ODE steps.

Table 6: Sampling diversity, measured as joint-position standard deviation in mm over 100 samples.

ODE steps	Head 6D	Random ≤15 masked	Upper-body XYZ	6IMU rot.+XYZ	Full joints
1	58.0	19.4	18.9	15.0	1.7
5	72.0	28.9	26.8	11.4	1.7
30	92.0	35.3	31.4	14.1	1.7

MCE learnable embeddings. Replacing per-joint learnable embeddings with zero-padding weakens the model's ability to distinguish *which* joints are missing, worsening MPJPE by +56.0 mm.

Diversity, input constraints, and inference steps. We sample 100 predictions for the same specific motion while varying only the observed mask and Euler ODE steps. The std of joint positions and 10 overlaid samples are shown in the Fig. 4. Tab. 6 show how sampling diversity varies according to input sensor configurations and flow matching sampling steps. Diversity grows as input signals decrease (rows) and as steps increase (columns), demonstrating that our model can output human motion generatively while conforming to input constraints. Quantitatively, however, we hardly observe any performance increase with more steps, so we use single-step for benchmarking. In real-world applications, users can trade speed for diversity (5 steps @ 250 fps, 30 steps @ 80 fps).

5 Conclusion

OmniPoser treats various flexible inputs (e.g. partial keypoints, head-mounted 6-DoF poses, sparse IMUs) as partial views of one shared body model, making it unified for different settings.

We achieve this through three components: modular masking with a Masked Conditioning Encoder, a decoupled dual-stream architecture, and Conditional Flow Matching, which altogether allow one model to cover diverse sensor setups while enabling real-time, single-step ODE inference. Across three settings and seven benchmarks, OmniPoser consistently performs on par with state-of-the-art methods while running at over 700 FPS. However, several limitations remain. First, OmniPoser models a single person and does not account for body-object or body-body contact forces; extending the masking abstraction to multi-person interaction or contact-rich object manipulation is an active direction for future work. Second, the cascaded inference pipeline propagates trajectory errors into pose recovery with no built-in correction mechanism; incorporating iterative refinement or feedback between the two streams could further improve robustness, though in practice the single-pass cascade already produces temporally smooth and drift-free trajectories across all evaluated benchmarks. Third, OmniPoser currently operates on the SMPL 22-joint skeleton and does not support hand or face articulation (SMPL-X); the modular feature-group design, however, is readily extensible to finer-grained body models as the corresponding training data becomes available. Extending the masking abstraction to multi-person interaction, contact-rich object manipulation, and articulated hand/face models would adapt our framework to a broader set of embodied tasks.

Acknowledgements

This work was supported by Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China, the Shanghai Commitee of Science and Technology, China (Grant No.24511103200), Shanghai Artificial Intelligence Laboratory, and XPLORER PRIZE grants.

References

1. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7291–7299 (2017)
2. Contributors, M.: Openmmlab pose estimation toolbox and benchmark. `https://github.com/open-mmlab/mmpose` (2020)
3. Dai, P., Zhang, Y., Liu, T., Fan, Z., Du, T., Su, Z., Zheng, X., Li, Z.: Hmd-poser: On-device real-time human motion tracking from scalable sparse observations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 874–884 (2024)
4. Fiche, G., Leglaive, S., Alameda-Pineda, X., Moreno-Noguer, F.: MEGA: Masked generative autoencoder for human mesh recovery. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5366–5378 (2025)
5. Grauman, K., Westbury, A., Torresani, L., Kitani, K., Malik, J., Afouras, T., Ashutosh, K., Baiyya, V., Bansal, S., Boote, B., et al.: Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19383–19400 (2024)
6. Guzov, V., Jiang, Y., Hong, F., Pons-Moll, G., Newcombe, R., Liu, C.K., Ye, Y., Ma, L.: Hmd 2: Environment-aware motion generation from single egocentric head-mounted device pp. 1394–1405 (2025)
7. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16000–16009 (2022)
8. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
9. Huang, C.H.P., Yi, H., Höschle, M., Safroshkin, M., Alexiadis, T., Polikovsky, S., Scharstein, D., Black, M.J.: Capturing and inferring dense full-body human-scene contact. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13274–13285 (2022)
10. Jiang, J., Streli, P., Luo, X., Gebhardt, C., Holz, C.: Manikin: biomechanically accurate neural inverse kinematics for human motion estimation. In: European Conference on Computer Vision. pp. 128–146. Springer (2024)
11. Jiang, J., Streli, P., Meier, M., Holz, C.: Egoposer: Robust real-time egocentric pose estimation from sparse and intermittent observations everywhere. In: European Conference on Computer Vision. pp. 277–294. Springer (2024)
12. Jiang, J., Streli, P., Qiu, H., Fender, A., Laich, L., Snape, P., Holz, C.: Avatarposer: Articulated full-body pose tracking from sparse motion sensing. In: European conference on computer vision. pp. 443–460. Springer (2022)
13. Kim, J., Geuter, J., Alvarez-Melis, D., Kakade, S., Chen, S.: Stop training for the worst: Progressive unmasking accelerates masked diffusion training. In: Forty-third International Conference on Machine Learning (2026), `https://arxiv.org/abs/2602.10314`
14. Lee, J., Joo, H.: Mocap everyone everywhere: Lightweight motion capture with smartwatches and a head-mounted camera. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1091–1100 (2024)
15. Li, J., Liu, K., Wu, J.: Ego-body pose estimation via ego-head pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17142–17151 (2023)

16. 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)
17. Li, J., Xu, C., Chen, Z., Bian, S., Yang, L., Lu, C.: Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3383–3393 (2021)
18. Li, R., Yang, S., Ross, D.A., Kanazawa, A.: Ai choreographer: Music conditioned 3d dance generation with aist++. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 13401–13412 (2021)
19. Li, Y., Nagarajan, T., Xiong, B., Grauman, K.: Ego-exo: Transferring visual representations from third-person to first-person videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6943–6953 (2021)
20. Lin, T.Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C.L., Dollár, P.: Microsoft coco: Common objects in context (2015), https://arxiv.org/abs/1405.0312
21. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2022)
22. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: Smpl: A skinned multi-person linear model. In: Seminal Graphics Papers: Pushing the Boundaries, Volume 2, pp. 851–866 (2023)
23. Ma, L., Ye, Y., Hong, F., Guzov, V., Jiang, Y., Postyeni, R., Pesqueira, L., Gamino, A., Baiyya, V., Kim, H.J., et al.: Nymeria: A massive collection of multimodal egocentric daily motion in the wild. In: European Conference on Computer Vision. pp. 445–465. Springer (2024)
24. Mahmood, N., Ghorbani, N., Troje, N.F., Pons-Moll, G., Black, M.J.: Amass: Archive of motion capture as surface shapes. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 5442–5451 (2019)
25. von Marcard, T., Henschel, R., Black, M., Rosenhahn, B., Pons-Moll, G.: Recovering accurate 3d human pose in the wild using imus and a moving camera. In: European Conference on Computer Vision (ECCV) (sep 2018)
26. Pan, S., Ma, Q., Yi, X., Hu, W., Wang, X., Zhou, X., Li, J., Xu, F.: Fusing monocular images and sparse imu signals for real-time human motion capture. In: SIGGRAPH Asia 2023 Conference Papers. pp. 1–11 (2023)
27. Pierzchlewicz, P., da Silva, C.O., Cotton, R.J., Sinz, F.H., et al.: Platypose: Calibrated zero-shot multi-hypothesis 3d human motion estimation. arXiv preprint arXiv:2403.06164 (2024)
28. Qian, Z., Zhang, S., Bhatnagar, B.L., Bogo, F., Tang, S.: Masked modeling for human motion recovery under occlusions (2026)
29. Rempe, D., Birdal, T., Hertzmann, A., Yang, J., Sridhar, S., Guibas, L.J.: Humor: 3d human motion model for robust pose estimation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 11488–11499 (2021)
30. 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. IEEE Transactions on Visualization and Computer Graphics **29**(5), 2337–2347 (2023)
31. Shin, S., Pahuja, A., Richard, A., Kitani, K., Saragih, J.M., Chen, Y., Xu, W., Halilaj, E., Bagautdinov, T.M.: EgoMDM: Diffusion-based human motion synthesis from sparse egocentric sensors. In: 2026 International Conference on 3D Vision (3DV). pp. 839–848 (2026). https://doi.org/10.1109/3DV69130.2026.00085

32. Su, J., Lu, Y., Pan, S., Wen, B., Liu, Y.: Roformer: Enhanced transformer with rotary position embedding (2021)
33. Tang, T., Ji, X., Li, Y., Liu, M., Xu, W., Lu, C.: Stereo-inertial poser: Towards metric-accurate shape-aware motion capture using sparse imus and a single stereo camera (2026), https://arxiv.org/abs/2603.02130
34. Trumble, M., Gilbert, A., Malleson, C., Hilton, A., Collomosse, J.: Total capture: 3d human pose estimation fusing video and inertial sensors. In: British Machine Vision Conference (BMVC) (2017)
35. Van Wouwe, T., Lee, S., Falisse, A., Delp, S., Liu, C.K.: Diffusionposer: Real-time human motion reconstruction from arbitrary sparse sensors using autoregressive diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2513–2523 (2024)
36. Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., et al.: Deep high-resolution representation learning for visual recognition. IEEE transactions on pattern analysis and machine intelligence **43**(10), 3349–3364 (2020)
37. Wang, Y., Sun, Y., Patel, P., Daniilidis, K., Black, M.J., Kocabas, M.: PromptHMR: Promptable human mesh recovery. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1148–1159 (2025)
38. Winkler, A., Won, J., Ye, Y.: Questsim: Human motion tracking from sparse sensors with simulated avatars. In: SIGGRAPH Asia 2022 conference papers. pp. 1–8 (2022)
39. Xu, Y., Zhang, J., Zhang, Q., Tao, D.: Vitpose: Simple vision transformer baselines for human pose estimation. Advances in neural information processing systems **35**, 38571–38584 (2022)
40. Yi, B., Ye, V., Zheng, M., Li, Y., Müller, L., Pavlakos, G., Ma, Y., Malik, J., Kanazawa, A.: Estimating body and hand motion in an ego-sensed world. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7072–7084 (2025)
41. Yi, X., Pan, S., Xu, F.: Improving global motion estimation in sparse IMU-based motion capture with physics. ACM Transactions on Graphics (TOG) **44**(4), 1–16 (2025). https://doi.org/10.1145/3730822
42. 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: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 13167–13178 (2022)
43. 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)
44. Yin, L., Shi, Z., Wu, Y., Yi, X., Xu, F., Guo, S.: Shape-aware inertial poser: Motion tracking for humans with diverse shapes using sparse inertial sensors. ACM Transactions on Graphics (TOG) **44**(6), 1–12 (2025). https://doi.org/10.1145/3763311
45. Zhang, H., Carlone, L.: Champ: Conformalized 3d human multi-hypothesis pose estimators (2024)
46. Zhang, S., Bhatnagar, B.L., Xu, Y., Winkler, A., Kadlecek, P., Tang, S., Bogo, F.: Rohm: Robust human motion reconstruction via diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14606–14617 (2024)

47. Zhang, Y., Wu, G., Chen, L.H., Zhao, Z., Lin, J., Jiang, X., Wu, J., Li, Z., Yang, H.F., Wang, H., Zhang, L.: HumanMM: Global Human Motion Recovery from Multi-shot Videos (Mar 2025). https://doi.org/10.48550/arXiv.2503.07597, http://arxiv.org/abs/2503.07597, arXiv:2503.07597 [cs]