

Ultra Diffusion Poser: Diffusion-Based Human Motion Tracking From Sparse Inertial Sensors and Ranging-Based Between-Sensor Distances

Dominik Hollidt Tommaso Bendinelli Christian Holz
Department of Computer Science, ETH Zurich, Switzerland

{dominik.hollidt, tommaso.bendinelli, christian.holz}@inf.ethz.ch

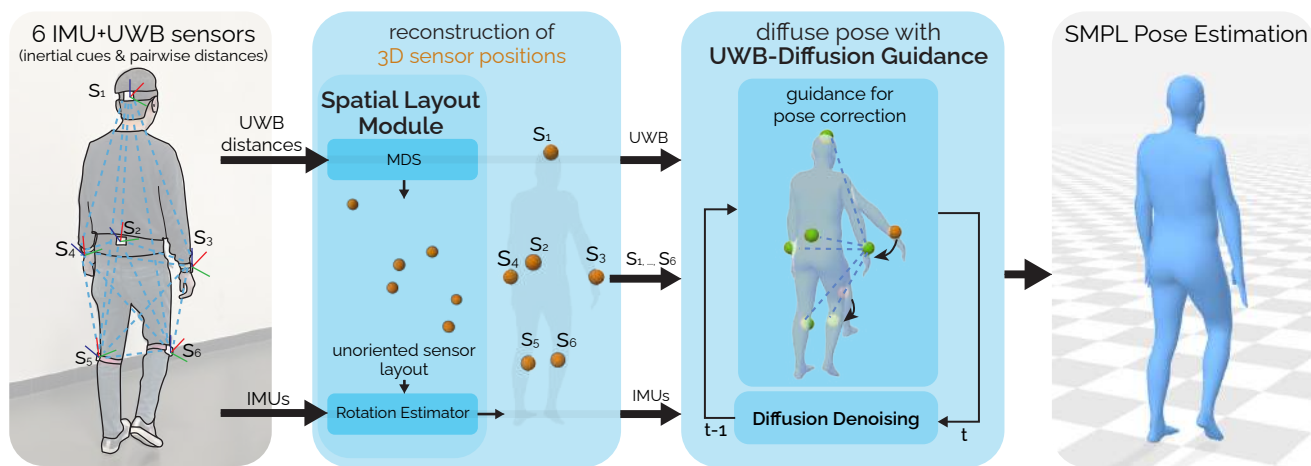


Figure 1. Our method UDP improves wearable IMU+UWB pose estimation by extending UWB as an auxiliary feature to actively model its geometric constraints. The Spatial Layout Module reconstructs 3D sensor positions from UWB measurements, providing a physically-informed input that conditions a diffusion model to predict SMPL poses. UWB-Diffusion Guidance encourages alignment between predicted poses and measured distances during diffusion sampling, improving accuracy and producing consistent motions.

Abstract

Methods using inertial measurement units (IMUs) provide a wearable alternative to camera-based motion capture. To mitigate drift from inertial signals, recent sparse inertial pose estimators integrate inter-sensor distances measured by ultra-wideband (UWB) ranging. So far, UWB distances have only been used as an additional input feature, ignoring the physical constraints they impose on sensor positions. However, these distances can also be used to reconstruct the underlying 3D sensor layout, which in turn provides more informative input for pose reconstruction. We propose Ultra Diffusion Poser, a diffusion model that explicitly models these geometric constraints. It includes a Spatial Layout Module that analytically reconstructs the 3D sensor positions from UWB measurements. These sensor positions are used alongside IMU signals and UWB distances as a conditioning signal during diffusion. Still, network predictions can violate inter-sensor distance measurements. To address this, we introduce UWB-Diffusion Guidance, which encour-

ages alignment between predicted poses and measured distances during diffusion sampling. Together, these contributions enable our model to achieve state-of-the-art performance, reducing joint position error by up to 22% over prior work. Code can be found here <https://github.com/eth-siplab/UltraDiffusionPoser>.

1. Introduction

Human pose estimation is fundamental for many applications in augmented and virtual reality and healthcare [8, 50]. While numerous camera-based approaches exist to estimate human poses, they present limitations for widespread adoption. External cameras mounted in the environment have a fixed sensing area and suffer from occlusions when the scene is crowded [5]. Pose estimation from wearable cameras depends on environmental factors such as lighting or background, suffers from occlusions [31, 36, 41, 48], raises privacy concerns, and is computationally demanding [12].

In contrast, compact body-worn sensors such as inertial

measurement units offer a minimally obtrusive alternative for human pose estimation, but accurate motion capture still requires numerous sensors. While commercial inertial motion capture systems, such as Xsens [40] or Noitom [27], achieve high accuracy with 17–19 sensors, current research is focusing on achieving the same performance with a more practical six sensor configuration [4, 14, 45].

While learning pipelines have improved pose accuracy, inertial approaches are inherently limited by noisy acceleration and angular velocity measurements from IMUs, which cause drifting pose estimations. The lack of absolute measurements of IMUs, i.e., no direct observations between frames in space or across body segments, makes drift correction very hard. As a result, sparse inertial pose estimators cannot fully resolve pose errors.

To address these limitations, recent approaches have combined IMU measurements with inter-sensor distances obtained via ultra-wideband (UWB) ranging [4, 20, 42]. By providing non-drifting body-fixed pairwise distances between sensors, UWB adds on-body constraints that help mitigate drift and thus produce more accurate and stable pose estimates [4, 20]. However, previous IMU+UWB methods have treated inter-sensor distances as auxiliary features. Instead, enforcing them as constraints on valid sensor positions enables the recovery of the full 3D sensor layout, a more informative input than raw distance scalars. Second, by ignoring these distance constraints during inference, existing approaches can produce poses that directly conflict with the measured distances, e.g., estimating wrist positions 70 cm apart when UWB reports 90 cm.

In this paper, we introduce Ultra Diffusion Poser (UDP), a diffusion-based method that explicitly models the inter-sensor distances as geometric signals, incorporating their spatial constraints directly into the model architecture. Our approach benefits from two main novelties. First, the Spatial Layout Module (SPL) analytically reconstructs the 3D sensor layout from just inter-sensor distances using multi-dimensional scaling (MDS). Because the MDS reconstruction is only defined up to rotation and reflection, we introduce a learnable Rotation Estimator, trained end-to-end with the diffusion model, to recover oriented 3D sensor positions and their reflected version. These 3D sensor positions serve as a strong conditional signal, acting as a pose prior during the diffusion denoising process. Second, UWB-Diffusion Guidance encourages *consistency* between the predicted pose and the measured inter-sensor distances, following diffusion classifier guidance [6]. During diffusion sampling, the predicted pose is continuously mapped to sensor positions via in-the-loop forward kinematics, and the alignment between predicted and measured inter-sensor distances is used to steer the denoising process toward satisfying UWB distances. To encourage smooth predictions over long sequences, we use autoregressive diffusion in-

painting [30, 33] that fills in the current motion, given the previous motion and the conditioning signal. In summary, we provide the following main contributions:

- UDP a novel fully learnable motion estimation system, that achieves state-of-the-art pose estimation from IMU+UWB measurements using an autoregressive diffusion-inpainting model. UDP explicitly models the geometric constraints imposed by inter-sensor UWB distances via the Spatial Layout Module and UWB-Diffusion Guidance.
- The *Spatial Layout Module* reconstructs the 3D sensor positions. Using multi-dimensional scaling, the 3D layout is obtained in closed form from UWB distances and serves as a strong prior for full-body pose estimation.
- *UWB-Diffusion Guidance* integrates in-the-loop forward kinematics into the diffusion sampling process to correct pose estimations that do not align with inter-sensor distance measurements, thereby reducing implausible or unrealistic pose estimations.

2. Related Work

Motion Capture from Sparse Inertial Sensors. Compared to camera-based systems, motion capture from sparse inertial sensors offers greater flexibility in placement, mobility, and robustness to occlusion. Large motion-capture datasets such as TotalCapture [32], DIP-IMU [14], and AMASS [23] have accelerated learning-based inertial pose estimation. These datasets enable the generation of synthetic IMU measurements by placing “virtual” sensors on mesh vertices [9, 12, 26, 29]. Still, synthetic data fails to fully capture real-world sensor noise and drift.

Recent methods for inertial pose estimation from six IMUs can be broadly categorized into purely learnable systems [14, 19, 20, 33, 35, 38] and hybrid systems that combine neural networks with post-optimization or physics-based simulation [4, 42, 45–47]. The first learning based method DIP [14], introduced a recurrent neural network that learns kinematic priors from large-scale motion databases. TIP [19] extends previous work by generating pose and terrain simultaneously using a transformer-based model. TransposePose introduces a multi-stage training process that predicts intermediate leaf joints for stabilized training. PIP [45] incorporated physical constraints (e.g., foot-floor contact) via a physics optimizer to reduce artifacts and drift. PNP [46] added better physics-aware IMU simulation. GlobalPose [47] estimates contact points into the physics simulation to reduce the flat world assumption. In parallel, research has pushed beyond the canonical six-IMU setup by inferring the entire body from fewer sensors [1, 18, 51, 53]. While TIC [54] and MODA [39] propose learning-based approaches to reduce drift, IMU-based methods remain inherently constrained by noisy inertial signals, whose integration inevitably leads to drift and degraded pose accuracy.

Motion Estimation with Ultra-Wideband Ranging. To reduce drift from inertial sensors [43], UIP [4] combines IMUs with ultra-wideband ranging to obtain direct inter-sensor distance measurements. Although UWB is affected by body occlusions, this fusion substantially reduces pose drift compared to IMU-only systems. UIP showed that passing the pairwise UWB distances through a graph convolutional network improves pose estimation. UMotion [20] proposes a Kalman filter to model and correct UWB measurement noise, but requires manual parameter tuning and cannot estimate global translation. GIP [42] extends UWB posing to two person pose estimation and refines their translation through an optimization framework. While all methods make important progress in combining UWB and inertial sensing, they do not explicitly enforce pose consistency with UWB measurements or exploit the full 3D sensor layout. UDP uses these insights to address these limitations.

Diffusion-Based Motion Estimation. Recently, methods have demonstrated the effectiveness of diffusion [10, 28] for motion generation [16, 22, 30] and camera-based pose estimation [13, 48]. Additionally, it has been shown that diffusion’s generative capabilities benefit underconstrained pose reconstruction from sparse trajectories [7, 15, 44], effectively filling in missing body parts. Similar to us, AGRoL [7] uses a seq-to-seq approach for pose estimation from three-point tracking. DiffusionPoser [33] applied a transformer-based diffusion model to inertial motion capture, enabling a greater range of potential sensor placements. Although quantitative improvements were limited and the implementation is not publicly available, the work provides an important step toward diffusion-driven inertial pose estimation. Building on this line of research, we introduce UWB diffusion guidance to predicted poses that are consistent with UWB distance measurements.

3. Method

We tailor UDP for the specific purpose of inertial body-pose estimation from global orientations \mathbf{R} , global accelerations \mathbf{A} , and pairwise inter-sensor distances \mathbf{D} . UDP is a fully learnable autoregressive diffusion-inpainting model that integrates geometric constraints from inter-sensor distances into its architecture, first with the **Spatial Layout Module** that reconstructs 3D sensor positions using MDS from inter-sensor distances in closed-form and through **UWB-Diffusion Guidance** to align the predicted poses with the measured inter-sensor distances.

3.1. Problem Statement

UDP, similar to prior work [4, 14, 19, 45], aims to reconstruct human motion using $k = 6$ sparse inertial sensors and their pairwise distances obtained via UWB.

At each timestamp t , the inertial measurements are given by $(\mathbf{R}_t, \mathbf{A}_t)$, where $\mathbf{R}_t \in \mathbb{R}^{k \times 3 \times 3}$ represents the sensor orientations, and $\mathbf{A}_t \in \mathbb{R}^{k \times 3}$ denotes their accelerations, both expressed in a common global frame. The pairwise distances d_{ij} between sensor i and j , are represented by matrix $\mathbf{D}_t \in \mathbb{R}^{k \times k}$.

From those inputs, UDP outputs human motion, which is parameterized using the SMPL body model $(\Theta_t, \mathbf{T}_t, \beta)$ [21]. $\Theta_t \in \mathbb{R}^{24 \times 3}$ defines the local joint rotations, $\mathbf{T}_t \in \mathbb{R}^3$ specifies the global body translation at time t , and β is the shape vector that characterizes the individual’s body shape.

Overall, UDP estimates the human motion from a sequence of N measurements, $\mathbf{R} = \{\mathbf{R}_t\}_{t=0}^N$, $\mathbf{A} = \{\mathbf{A}_t\}_{t=0}^N$, $\mathbf{D} = \{\mathbf{D}_t\}_{t=0}^N$:

$$(\Theta, \mathbf{T}) = \text{UDP}(\mathbf{R}, \mathbf{A}, \mathbf{D}, \beta) \quad (1)$$

3.2. Full-Body Pose Estimation

The architecture of UDP is visualized in Fig. 2. At its core, UDP is an autoregressive diffusion inpainting model that predicts the current motion given previously predicted motion and the current conditioning signal.

Instead of conditioning motion generation solely on raw IMU and UWB measurements, the Spatial Layout Module provides the 3D sensor layout that serves as a strong geometric prior during diffusion. First, it reconstructs an initial 3D sensor layout in closed form using multidimensional scaling (MDS) based only on inter-sensor distances. Since the MDS reconstruction is only defined up to rotation and reflection, a learnable Rotation Estimator orients both the initial MDS layout and its reflected version to produce the final oriented sensor layouts (P_{SPL}, \bar{P}_{SPL}) . The oriented 3D sensor layout, together with the IMU and UWB signals, form the conditioning input to the diffusion model.

To ensure that the generated poses remain consistent with measured inter-sensor distances, we introduce UWB-Diffusion Guidance. During diffusion sampling, this mechanism uses forward kinematics (FK) to compute the predicted sensor positions from the current SMPL pose. These predictions are compared to the measured UWB distances, and the diffusion step is guided to correct poses that violate the constraints. By integrating FK directly into the sampling loop, UWB-Diffusion Guidance produces smooth and realistic motion even under noisy sensor data, whereas simple per-frame post-optimization would lead to jitter and potentially unrealistic motions [42].

In the following subsections, we first describe the motion representation and the conditioning signals used by UDP. We then present the components of our method: the Spatial Layout Module, the autoregressive diffusion inpainting process, and the UWB-Diffusion Guidance mechanism.

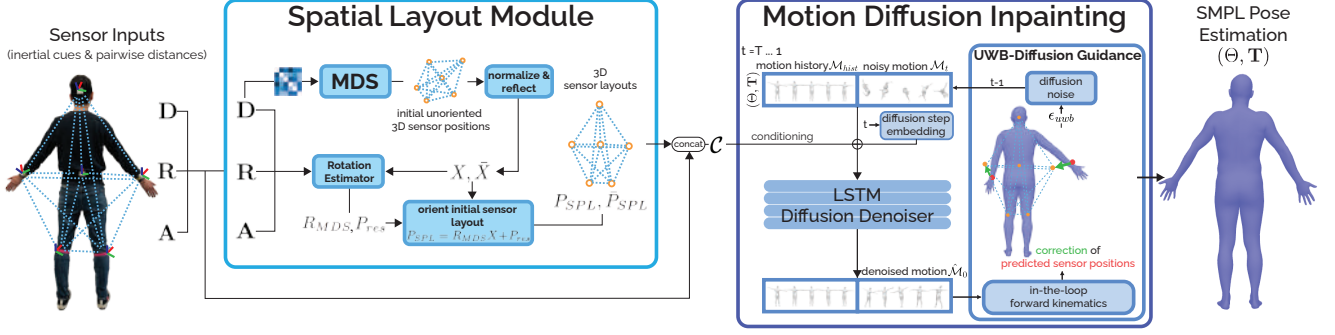


Figure 2. The Spatial Layout Module applies metric MDS to the pairwise distance matrix \mathbf{D} to recover initial sensor positions, which are then oriented by a learnable Rotation Estimator. The resulting 3D sensor layout provides a strong conditioning signal for the diffusion model. The autoregressive diffusion inpainting model extends the previously predicted motion based on the current conditioning signal to ensure smooth motion prediction. UWB-Diffusion Guidance steers the pose predictions to align with the measured inter-sensor distances.

3.2.1. Motion Representation

To obtain the final motion estimation, UDP diffuses the motion representation conditioned on the recovered 3D sensor layout and IMU+UWB measurements.

We represent our motion sequence of length N with joint orientations in 6D representation [52] and translation according to the SMPL model:

$$\mathcal{M} \in \mathbb{R}^{N \times 147} = [\Theta^{6D}, \mathbf{T}] \quad (2)$$

For better generalization, we normalize each motion sequence \mathcal{M} such that its root joint always starts at $(0, \text{root height}, 0)$. This canonical normalization enables the model to predict accurate positions while mitigating representation shifts in long sequences.

To ensure smooth and continuous motion prediction, we use autoregressive diffusion inpainting. We prepend the N_H previous predicted frames \mathcal{M}_{hist} to \mathcal{M} , normalized such that its final frame aligns with the first frame of \mathcal{M} .

The diffusion model D_θ diffuses the motion representation based on a conditional signal. This condition tensor \mathcal{C} includes the recovered 3D sensor positions (P_{SPL}, \bar{P}_{SPL}), sensor orientations as rotation matrices, accelerations, inter-sensor distances, and the body shape:

$$\mathcal{C} \in \mathbb{R}^{N \times 154} = [P_{SPL}, \bar{P}_{SPL}, \mathbf{R}, \mathbf{A}, \mathbf{D}, \beta] \quad (3)$$

3.2.2. Spatial Layout Module

The inter-sensor distances constrain the sensor positions to a fixed spatial configuration. To recover the corresponding 3D sensor layout, our SPL module applies metric multidimensional scaling (MDS) [2, 24, 25]. Given the pairwise distance matrix $\mathbf{D}_t \in \mathbb{R}^{k \times k}$, MDS finds the k sensor points x_i in 3D space whose inter-point distances best match the measured distances d_{ij} by solving:

$$\underset{\mathbf{X}_{\text{MDS}} \in \mathbb{R}^{k \times 3}}{\text{argmin}} \sum_{i < j} (|x_i - x_j|^2 - d_{ij})^2. \quad (4)$$

In Euclidean space, this classical MDS problem has a closed-form solution[37]:

$$\mathbf{H} = \mathbf{I} - \frac{1}{k} \mathbf{1} \mathbf{1}^\top, \quad \mathbf{B} = -\frac{1}{2} \mathbf{H} \mathbf{D}^{(2)} \mathbf{H}, \quad (5)$$

where $\mathbf{D}^{(2)}$ contains squared distances, \mathbf{I} is the $k \times k$ identity matrix, and $\mathbf{1}$ is a k -dimensional vector of ones.

We then perform a singular value decomposition of the Gram matrix:

$$\mathbf{B} = \mathbf{U} \mathbf{\Sigma} \mathbf{V}^\top, \quad (6)$$

and obtain the 3D sensor coordinates from the top three eigenvectors and corresponding eigenvalues:

$$\mathbf{X}_{\text{MDS}} = \mathbf{U}_3 \mathbf{\Sigma}_3^{1/2}. \quad (7)$$

Here, $\mathbf{X}_{\text{MDS}} \in \mathbb{R}^{k \times 3}$ contains the unoriented 3D spatial positions of k sensors. This layout is defined up to translation, rotation, and reflection, and thus the sequence of layouts is temporally incoherent.

To resolve this ambiguity, the SPL first brings the MDS layouts into a canonical, normalized frame, which is then oriented by a learned Rotation Estimator. The initial unoriented layouts are normalized such that the pelvis is at the origin, the layout is upright (root-head aligns with y-axis) and forward facing (left wrist lies on the positive direction of the XY-plane). To ensure temporal consistency, frame-wise reflections are resolved whenever the current layout deviates from the previous one by more than a threshold, i.e., when $\|X_{t-1} - X_t\|_2^2 > c$. This yields a normalized and temporally consistent sequence of 3D sensor positions X . Since the entire sequence may still be reflected, we additionally generate a mirrored version \bar{X} . We then retain both X and \bar{X} as layout candidates, one of which corresponds to the correct configuration. Finally, a Rotation Estimator predicts the orientation of the sensor layout. It takes the normalized layout as input and learns to rotate it into a

consistent, correctly oriented configuration. The Rotation Estimator is trained end-to-end together with the diffusion model:

$$R_{MDS}, P_{res} = \text{RotEstimator}(X, \bar{X}, \mathbf{R}, \mathbf{A}, \mathbf{D}), \quad (8)$$

The initial MDS layout may be imperfect due to sensor noise. Therefore, we refine it by adding an estimated residual P_{res} per sensor point:

$$P_{SPL} = R_{MDS}X + P_{res}, \bar{P}_{SPL} = R_{MDS}\bar{X} + P_{res}, \quad (9)$$

Thus, we obtain our recovered sensor layouts $[P_{SPL}, \bar{P}_{SPL}]$ that represent the sensor positions of head, pelvis, wrists, and knees. We use a 3-layer LSTM [11] for the Rotation Estimator.

3.2.3. Motion Diffusion Inpainting

UDP adopts the Denoising Diffusion Probabilistic Model (DDPM) formulation [10]. At each diffusion step t , the denoiser D_θ predicts the clean motion $\hat{\mathcal{M}}_0$ from the noisy motion representation \mathcal{M}_t , conditioned on the input signal \mathcal{C} and the previously predicted motion segment \mathcal{M}_{hist} of length N_H :

$$\hat{\mathcal{M}}_0 = D_\theta(\mathcal{M}_t, \mathcal{M}_{hist}, \mathcal{C}, t) \quad (10)$$

We utilize diffusion-based inpainting [30, 33] to enforce temporal consistency across overlapping windows. Specifically, the noisy motion \mathcal{M}_t , the motion history \mathcal{M}_{hist} , and the condition \mathcal{C} are projected into a shared embedding space. The condition and noisy motion are added and appended to the motion history, and a learnable history token h_e is added to mark \mathcal{M}_{hist} . The diffusion step t is encoded via an MLP, added and prepended to the sequence, yielding the combined input S of length $1 + N_H + N$. We denoise S using a LSTM [11].

During inference, we apply the diffusion inpainting autoregressively to predict arbitrarily long motion sequences and apply Gaussian smoothing with $\sigma = 2$ for smoother results.

To train UDP we use the standard diffusion loss [10]:

$$\mathcal{L}_{\text{simple}} = \mathbb{E}_{q(x_t|x_0)} [\|x_0 - D_\theta(x_t, t, \mathcal{C})\|_2^2]. \quad (11)$$

where $q(x_t|x_0)$ denotes the DDPM forward process that gradually adds Gaussian noise to x_0 [10]. Unlike Diffusion-Poser [33], we train our model without the need for forward kinematics losses that significantly slow down the training. Instead, we use the translation and SMPL joint angle auxiliary losses:

$$\mathcal{L}_{\text{tran}} = \|\mathbf{T}^{\text{pred}} - \mathbf{T}^{\text{gt}}\|_2 \quad (12)$$

$$\mathcal{L}_{\text{smp}} = |\Theta_{6d}^{\text{pred}} - \Theta_{6d}^{\text{gt}}| \quad (13)$$

We also utilize this velocity loss that encourages more active motion translation:

$$\mathcal{L}_{\text{vel}} = \left(\|\mathbf{T}_n^{\text{pred}} - \mathbf{T}_{n-1}^{\text{pred}}\|_2 - \|\mathbf{T}_n^{\text{gt}} - \mathbf{T}_{n-1}^{\text{gt}}\|_2 \right)^2 \quad (14)$$

Algorithm 1 UWB-guided conditional diffusion sampling

Require: Condition \mathcal{C} , history \mathcal{M}_{hist} , UWB distances d_{ij}

Ensure: Denoised motion \mathcal{M}_0

```

1:  $\mathcal{M}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ 
2: for  $t = T$  down to 1 do
3:    $\hat{\mathcal{M}}_0 \leftarrow D_\theta(\mathcal{M}_t, \mathcal{M}_{hist}, \mathcal{C}, t)$   $\triangleright$  denoise
4:    $\hat{d}_{ij} \leftarrow \text{FK}(\hat{\mathcal{M}}_0)$   $\triangleright$  predicted distances
5:    $\epsilon_{uwb} \leftarrow \sum_{i < j} \|\hat{d}_{ij} - d_{ij}\|^2$   $\triangleright$  UWB guidance loss
6:    $\tilde{\mu}_t \leftarrow \mu_t(\mathcal{M}_t, \hat{\mathcal{M}}_0) - \lambda \Sigma_t \nabla \epsilon_{uwb}$   $\triangleright$  guidance
7:    $\mathcal{M}_{t-1} \sim \mathcal{N}(\tilde{\mu}_t, \Sigma_t)$   $\triangleright$  sample
8: end for
9: return  $\mathcal{M}_0$ 

```

3.2.4. UWB-Diffusion Guidance

Even with strong conditioning from the 3D sensor layout and UWB distances, the diffusion process can still produce motions that violate the inter-sensor distances \mathbf{D} . Enforcing distances on the SMPL pose via a simple frame-wise pose optimization can lead to infeasible joint angles, unnatural poses, and jittery poses due to sensor noise.

Hence, we incorporate UWB-Diffusion guidance to steer the diffusion process toward satisfying the pairwise distance constraints, encouraging agreement with the measured inter-sensor distances \mathbf{D} , see Alg. 1.

UWB-Diffusion guidance makes use of classifier guidance [6] that offers a mechanism to control the diffusion sampling process. During inference, each noisy sample \mathcal{M}_{t-1} is sampled from a Gaussian with mean $\mu_t(\mathcal{M}_{t-1}, \hat{\mathcal{M}}_0)$ that is determined by \mathcal{M}_{t-1} and $\hat{\mathcal{M}}_0$, as in standard diffusion sampling [6, 48]. By shifting this mean in the direction of a gradient of a function ϵ that encourages desired properties, the diffusion process can be directed toward more preferred predictions (see line 6 in Alg. 1). In our case, the guidance term encourages consistency between the predicted pose and the measured inter-sensor distances, while retaining smooth outputs and avoiding infeasible poses. We define the function ϵ by computing predicted sensor positions, defined by the SMPL mesh, via in-the-loop forward kinematics during sampling on $\hat{\mathcal{M}}_0$, from which the predicted distances \hat{d}_{ij} are computed to match the measured distances d_{ij} . As UWB distances are inter-body constraints, gradients for the root translation are set to zero.

$$\epsilon_{uwb} = \sum_{i < j} \|\hat{d}_{ij}(\hat{\mathcal{M}}_0) - d_{ij}\|^2 \quad (15)$$

4. Experiments

4.1. Experimental Setup

To ensure a fair comparison, we follow the same evaluation setup used in previous work [4, 45, 46] and train on the same AMASS [23] split. We then evaluated UDP on

DanceDB [3], an AMASS split held out during training, TotalCaptureReal [32] and the test split of DIP-IMU [14], using the DIP-IMU training split.

Finally, to demonstrate real-world applicability, we also evaluated our model on two real-world datasets where raw IMU and UWB are recorded during acquisition: UIP-DB [4] and the recent GIP-DB [42]. These datasets are particularly challenging due to higher and more realistic IMU drift and UWB noise. For these datasets, we report the performance without, i.e. zero-shot, and with fine-tuning. Further details regarding the training and evaluation protocol are available in the supplementary material (Sec. A.1). We evaluated our method using the following metrics:

SIP ($^{\circ}$) global angle error of shoulders and hips.

GAE ($^{\circ}$) mean global angle error of all joints.

JPE (cm) root-aligned joint position error.

Jitter (m/s^3) mean magnitude of the third derivative (jerk) of global joint positions.

We compare UDP against state-of-the-art methods, including the IMU+UWB approaches UIP [4], UMotion [20] and GIP [42], as well as the IMU-only methods PNP [46], GlobalPose [47], DynaIP [49], PIP [45], and TIP [19]. Note that, to ensure fairness, we report the numbers from the original papers and replicate results using our pipeline whenever possible.

4.2. Results

We report quantitative results for DIP-IMU in Tab. 1 and for DanceDB, TotalCapture, UIP-DB, and GIP-DB in Tab. 2, with detailed interpretation in the following subsection. Next, we analyze UDP’s sensitivity to UWB noise (Sec. 4.2.3). Finally, multiple ablation studies are presented to assess the effectiveness of our modules (Sec. 4.3). Additional results and experiments are available in the appendix.

4.2.1. Pose Estimation Evaluation

UDP achieves state-of-the-art results across most evaluation settings. UDP achieves the lowest joint position error in all evaluation settings, with up to 22% improvement, except for UIP-DB without fine-tuning. The improvement over IMU-only methods is up to 35% showing the benefit of distance measurements. Notably, UDP produces consistently smooth motion with minimal jitter across all datasets without the need for a physics optimizer as in UIP, PIP, or PNP. On DIP-IMU, DanceDB, and TotalCapture the SIP error is consistently the lowest, highlighting the accurate estimation of the limbs, which impacts visual quality the most. The GA error achieved by UDP is either the lowest or competitive, being only slightly surpassed by DynaIP on DIP-IMU and by GlobalPose on TotalCapture. This demonstrates that modeling geometric constraints from inter-sensor distances improves UDP’s pose estimation.

UDP also generalizes well to real-world datasets. On GIP-DB, a newer real-world dataset with modern UWB sensors and lower noise levels, UDP achieves state-of-the-art results in both zero-shot and fine-tuned settings. In contrast, it only struggles when both UWB and IMU signals are highly noisy, as seen on UIP-DB, where IMU drift reaches $3.21^{\circ}/\text{min}$ and UWB errors 17.5 cm [4]. In this case, UDP achieves SOTA performance only after fine-tuning, which allows it to adapt to the dataset’s noise characteristics, while remaining competitive otherwise. We posit that, as UDP relies more heavily on UWB, it is more susceptible to UWB sensing errors; we investigate this further in Sec. 4.2.3.

Table 1. Results on DIP-IMU. ”*” denotes the results we reproduced, because they were not provided in the original paper. ”†” denotes results that we reproduced, because the paper used a different evaluation framework. GT Jitter is 1.830.

Model	UWB	SIP ($^{\circ}$) ↓	GAE ($^{\circ}$) ↓	JPE (cm) ↓	Jitter
TIP	✗	17.07	10.51*	5.82	0.882
PIP	✗	15.02	8.78*	5.12	0.240
PNP	✗	13.71	8.75	4.98	0.260
GlobalPose	✗	13.55	8.47	4.65	0.260
DynaIP	✗	14.41	7.12*	5.03	-
UMotion	✓	15.05†	10.41†	4.38†	0.216†
UIP	✓	13.20	8.23*	5.05	0.240
UDP (ours)	✓	10.39	8.19	3.42	0.125

4.2.2. Qualitative Results

We present extensive qualitative results in Fig. 3 and the supplementary material.

As seen in the figure, UDP excels in the accurate estimation of limb positions. This is due to its strong integration of inter-sensor distances, which impose strong constraints on feasible positions. For example, in row one of Fig. 3, the right arm is extended slightly backward and the left foot is raised slightly higher, resulting in a more accurate pose prediction compared to PIP and UIP. Notably, the better pose estimations of UDP are most noticeable during fast motions such as fast arm movements, where other methods tend to lag behind. Thanks to the UWB-Diffusion Guidance, UDP remains robust, producing more accurate leg estimates, unlike physics-optimized methods such as PIP, PNP, and UIP, that struggle with complex dance sequences and yield implausible foot placements, see second row in Fig. 3.

4.2.3. Sensitivity Towards Noise

While UDP is the best method on all other datasets, it achieves only competitive performance in the zero-shot setting on the UIP-DB dataset, where UWB measurement errors exceed 17 cm. Here, UIP surpasses it in SIP and joint position error. We attribute this to the strong influence of noisy UWB signals, which are tightly integrated into UDP’s architecture. To analyze this, we compare both

Table 2. Combined colored results on DanceDB (blue), TotalCapture (green), and UIP-DB (orange). “*” denotes values we reproduced because they were not reported in the original papers. “†” denotes results that we reproduced because the paper used a different evaluation framework. Ground Truth Jitter: DanceDB: 2.09, TotalCapture: 0.465, UIP-DB: 0.850, GIP-DB: 0.113. **Bold** indicates best.

Dataset	Model	UWB	SIP (°) ↓	GAE (°) ↓	JPE (cm) ↓	Jitter
DanceDB	TIP	✗	18.70	11.80*	8.50	1.438
DanceDB	PIP	✗	20.00	19.17*	8.87	0.661*
DanceDB	PNP	✗	16.52*	11.45*	7.18*	0.747*
DanceDB	UMotion	✓	11.93†	10.61†	5.19†	1.471†
DanceDB	UIP	✓	15.28	10.98*	7.45	0.430
DanceDB	UDP (ours)	✓	11.79	9.91	4.67	0.175
TotalCapture	TIP	✗	11.36	12.30	5.15	0.751
TotalCapture	PIP	✗	12.93	12.04	5.61	0.204
TotalCapture	PNP	✗	10.89	10.45	4.74	0.260
TotalCapture	GlobalPose	✗	9.81	9.99	4.25	0.350
TotalCapture	DiffusionPoser	✗	–	14.40	6.10	–
TotalCapture	UMotion	✓	10.58†	11.36†	4.83†	0.215†
TotalCapture	UIP	✓	10.70	11.43*	5.11	0.206
TotalCapture	UDP (ours)	✓	8.95	10.19	3.76	0.124
UIP-DB	TIP	✗	33.01	–	14.82	1.860
UIP-DB	PIP	✗	30.47	29.33*	13.62	1.570
UIP-DB	UMotion	✓	31.61†	27.92†	13.19†	0.054†
UIP-DB	UIP	✓	24.12	27.59*	10.65	0.050
UIP-DB	UDP (ours)	✓	24.95	25.92	11.72	0.071
UIP-DB	UMotion finetuned	✓	24.37†	25.69†	11.13†	0.045†
UIP-DB	UIP finetuned	✓	23.85	25.34*	10.65	0.080
UIP-DB	UDP (ours) finetuned	✓	19.24	21.81	9.04	0.074
GIP-DB	UIP	✓	30.18	26.16	10.88	0.224*
GIP-DB	UMotion	✓	26.57†	23.43†	9.52†	0.128†
GIP-DB	GIP	✓	27.77	23.34	9.45	0.706*
GIP-DB	UDP (ours)	✓	25.74	22.23	8.86	0.125
GIP-DB	UMotion finetuned	✓	25.67†	23.23†	9.34†	0.132†
GIP-DB	GIP finetuned	✓	18.04	17.57	8.70	0.608*
GIP-DB	UDP (ours) finetuned	✓	15.33	14.34	6.68	0.092

methods on UIP-DB by progressively blending noisy UWB measurements with ground-truth distances (Tab. 3). While UIP performs slightly better under raw, noisy conditions, UDP achieves up to 1.4 cm lower JPE and 4.59° lower GAE once measurement noise is reduced. This trend aligns with UDP’s SOTA results on GIP-DB, a newer real-world dataset with more modern, less noisy UWB sensors, demonstrating both its real-world applicability and the importance of modern UWB-radios with a lower noise profile.

Table 3. Improvement of UDP over UIP across different UWB noise levels, non-finetuned. The UWB distances are linearly interpolated between real and ground-truth measurements.

UWB-Noise Level	Improvement over UIP	
	GAE (°)	JPE (cm)
100% (Real noise)	1.67	-1.07
75%	3.19	0.71
50%	3.76	0.91
25%	4.12	1.14
0% (Perfect distances)	4.59	1.40

4.3. Ablation Studies

To validate the contributions of our two main modules, we perform ablation studies on TotalCapture.

Table 4. Ablation Studies on TotalCapture.

Metric	SIP (°) ↓	GAE (°) ↓	JPE (cm) ↓
UDP w/o SPL & UWB Guidance	10.22	10.79	4.39
UDP w/o SPL	9.58	10.45	4.03
UDP w/o UWB Guidance	9.46	10.46	4.05
UDP w/o RotEstimator	9.11	10.45	3.92
UDP RotEstimator w/o P_{res}	9.01	10.22	3.79
UDP	8.95	10.19	3.76

Spatial Layout Module. Disabling the SPL module increases all metrics, with SIP error rising by 7%, highlighting its impact on overall pose accuracy, see Tab. 4. The SPL benefits further from the Rotation Estimator, which orients sensor layouts. Disabling it and replacing the oriented sensor layouts P_{SPL}, \bar{P}_{SPL} with the normalized MDS layouts X, \bar{X} increases all errors. This shows that reconstructing the 3D sensor layout via SPL yields better pose estimates than just using inter-sensor distances as auxiliary inputs.

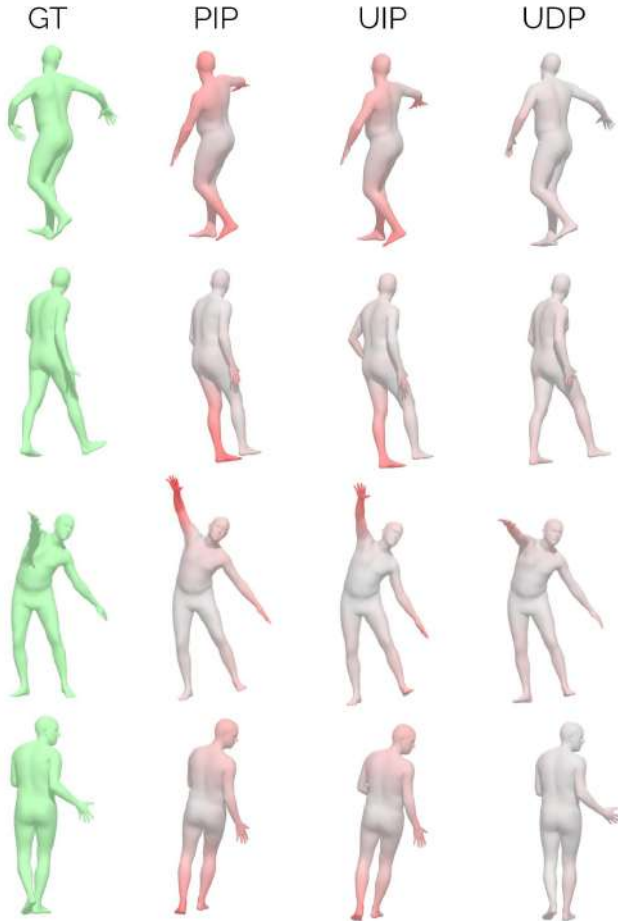


Figure 3. Qualitative Results of UDP from DanceDB and Total-Capture. A brighter red indicates a higher position error. UDP consistently produces accurate poses.

UWB-Diffusion Guidance. UWB-Diffusion guidance improves local pose estimates, reducing SIP error by 5% and JPE by 7%. Limb segments benefit most from the corrective effect of diffusion guidance. Disabling both the SPL and UWB-Diffusion Guidance further deteriorates performance, with JPE increasing by 17%, showing their complementary effect. Because UWB-Diffusion Guidance applies corrections during the sampling process, the predicted motions remain smooth. This highlights the benefit of leveraging geometric constraints during diffusion sampling by correcting poses that violate UWB measurements.

5. Discussion and Limitations

We show that UDP excels at local pose estimation thanks to its strong integration of inter-sensor distances. However, this makes it sensitive to very noisy UWB measurements, see Sec. 4.2.3, and the UWB-Guidance coefficient should be adjusted to the expected noise level. Future work could

improve robustness by modeling sensor uncertainty, which can be incorporated into the SPL module or guidance loss.

While UDP is fully data-driven and does not require manual tuning of post-optimization problems, Kalman filter hyperparameters, as in UMotion, or modeling of physics optimizers, it also does not benefit from the constraints these approaches enforce. For tasks requiring zero-foot skating, ground penetration avoidance, or strict physical consistency, especially when floor or environment information is available, physics optimizers still have an advantage. Future work could investigate refining UDP’s predictions using existing physics optimizers or integrating novel physics-based guidance terms into the diffusion framework.

UDP’s main contributions focus on improving local pose through on-body constraints. This could be extended to inter-person constraints, similar to GIP [42], by utilizing inter-person UWB distances. Combining UDP’s smooth, UWB-aligned motion predictions with the more jitter-prone inter-person optimizations from GIP has the potential for accurate and globally consistent motion estimation. These ideas could further be extended to incorporate environment-fixed UWB beacons, enabling multi-person and environment-aware pose refinement.

Finally, similar to prior work [4, 20, 45–47], UDP is restricted to a fixed six-sensor setup. Extending support to arbitrary sensor layouts could further increase its real-world applicability and flexibility.

6. Conclusion

In this work, we introduced UDP, a novel full-body pose estimation approach utilizing sparse inertial sensing combined with UWB-based distance measurements. While prior methods used UWB distances only as additional features, UDP explicitly models the geometric constraints they impose on sensor positions. It integrates three key components: (1) a novel Spatial Layout Module that estimates 3D sensor positions, (2) an autoregressive diffusion inpainting model, and (3) UWB-Diffusion Guidance that encourages pose predictions to be aligned with measured distances during diffusion sampling. We show that explicitly using the Spatial Layout Module and UWB-Diffusion Guidance leads to more accurate and physically plausible poses compared to treating UWB distances as auxiliary input features. In total, our experiments showed that UDP improves pose estimation by up to 22%, outperforming both its direct IMU+UWB baselines and IMU based posers. We hope these results highlight the potential of modeling the geometric constraints provided by novel UWB sensors explicitly to produce advanced models for more generalized and reliable wearable-based pose estimation solutions.

Acknowledgements We thank Yi-Chi Liao, Jiayi Jiang, and all other people who provided valuable feedback.

References

- [1] Karan Ahuja, Eyal Ofek, Mar Gonzalez-Franco, Christian Holz, and Andrew D. Wilson. Coolmoves: User motion accentuation in virtual reality. In *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, pages 1–23, 2021. 2
- [2] Farzana Anowar, Samira Sadaoui, and Bassant Selim. Conceptual and empirical comparison of dimensionality reduction algorithms (pca, kpca, lda, mds, svd, lle, isomap, le, ica, t-sne). *Computer Science Review*, 40:100378, 2021. 4
- [3] Andreas Aristidou, Ariel Shamir, and Yiorgos Chrysanthou. Digital dance ethnography: Organizing large dance collections. *J. Comput. Cult. Herit.*, 12(4), 2019. 6
- [4] Rayan Armani, Changlin Qian, Jiayi Jiang, and Christian Holz. Ultra inertial poser: Scalable motion capture and tracking from sparse inertial sensors and ultra-wideband ranging. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 2, 3, 5, 6, 8, 12, 17
- [5] Yu Cheng, Bo Yang, Bo Wang, Wending Yan, and Robby T Tan. Occlusion-aware networks for 3d human pose estimation in video. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 723–732, 2019. 1
- [6] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems*, pages 8780–8794. Curran Associates, Inc., 2021. 2, 5
- [7] Yuming Du, Robin Kips, Albert Pumarola, Sebastian Starke, Ali Thabet, and Arsiom Sanakoyeu. Avatars grow legs: Generating smooth human motion from sparse tracking inputs with diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 481–490, 2023. 3
- [8] Shradha Dubey and Manish Dixit. A comprehensive survey on human pose estimation approaches. *Multimedia Systems*, 29(1):167–195, 2023. 1
- [9] Vladimir Guzov, Aymen Mir, Torsten Sattler, and Gerard Pons-Moll. Human poseitoning system (hps): 3d human pose estimation and self-localization in large scenes from body-mounted sensors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4318–4329, 2021. 2
- [10] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3, 5
- [11] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, 1997. 5
- [12] Dominik Hollidt, Paul Strelí, Jiayi Jiang, Yasaman Haghighi, Changlin Qian, Xintong Liu, and Christian Holz. Egosim: An egocentric multi-view simulator and real dataset for body-worn cameras during motion and activity. *Advances in Neural Information Processing Systems*, 37:106607–106627, 2025. 1, 2
- [13] Karl Holmquist and Bastian Wandt. Diffpose: Multi-hypothesis human pose estimation using diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15977–15987, 2023. 3
- [14] Yinghao Huang, Manuel Kaufmann, Emre Aksan, Michael J Black, Otmar Hilliges, and Gerard Pons-Moll. Deep inertial poser: Learning to reconstruct human pose from sparse inertial measurements in real time. *ACM Transactions on Graphics (TOG)*, 37(6):1–15, 2018. 2, 3, 6
- [15] Andela Ilic, Jiayi Jiang, Paul Strelí, Xintong Liu, and Christian Holz. Human motion capture from loose and sparse inertial sensors with garment-aware diffusion models. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, pages 1206–1214. International Joint Conferences on Artificial Intelligence Organization, 2025. Main Track. 3
- [16] Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. *Advances in Neural Information Processing Systems*, 36:20067–20079, 2023. 3
- [17] Jiayi Jiang, Paul Strelí, Huajian Qiu, Andreas Fender, Larissa Laich, Patrick Snape, and Christian Holz. Avatarposer: Articulated full-body pose tracking from sparse motion sensing. In *Computer Vision—ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part V*, pages 443–460, 2022. 17
- [18] Jiayi Jiang, Paul Strelí, Manuel Meier, and Christian Holz. Egoposer: Robust real-time egocentric pose estimation from sparse and intermittent observations everywhere. In *European Conference on Computer Vision*, pages 277–294. Springer, 2024. 2
- [19] Yifeng Jiang, Yuting Ye, Deepak Gopinath, Jungdam Won, Alexander W Winkler, and C Karen Liu. Transformer inertial poser: Real-time human motion reconstruction from sparse imus with simultaneous terrain generation. In *SIGGRAPH Asia 2022 Conference Papers*, pages 1–9, 2022. 2, 3, 6, 12, 17
- [20] Huakun Liu, Hiroki Ota, Xin Wei, Yutaro Hirao, Monica Perusquia-Hernandez, Hideaki Uchiyama, and Kiyoshi Kiyokawa. Umotion: Uncertainty-driven human motion estimation from inertial and ultra-wideband units. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7085–7094, 2025. 2, 3, 6, 8, 15
- [21] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: a skinned multi-person linear model. *ACM Transactions on Graphics (TOG)*, 34(6):1–16, 2015. 3
- [22] Thomas Lucas, Fabien Baradel, Philippe Weinzaepfel, and Grégory Rogez. PoseGPT: Quantization-based 3D Human Motion Generation and Forecasting, 2022. arXiv:2210.10542 [cs]. 3
- [23] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of the International Conference on Computer Vision*, pages 5442–5451, 2019. 2, 5
- [24] Al Mead. Review of the development of multidimensional scaling methods. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 41(1):27–39, 1992. 4
- [25] Max Mignotte. Mds-based multiresolution nonlinear dimensionality reduction model for color image segmenta-

- tion. *IEEE transactions on neural networks*, 22(3):447–460, 2011. 4
- [26] Vimal Mollyn, Riku Arakawa, Mayank Goel, Chris Harrison, and Karan Ahuja. Imuposer: Full-body pose estimation using imus in phones, watches, and earbuds. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, page Article 529, Hamburg, Germany, 2023. Association for Computing Machinery. 2
- [27] Noitom. Noitom motion capture. <https://www.noitom.com/>, 2024. 2
- [28] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. 3, 13
- [29] Paul Strel, Rayan Armani, Yi Fei Cheng, and Christian Holz. Hoov: Hand out-of-view tracking for proprioceptive interaction using inertial sensing. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–16, 2023. 2
- [30] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*, 2022. 2, 3, 5, 17
- [31] Denis Tome, Patrick Peluse, Lourdes Agapito, and Hernan Badino. xR-EgoPose: Egocentric 3D Human Pose from an HMD Camera, 2019. arXiv:1907.10045 [cs]. 1
- [32] Matthew Trumble, Andrew Gilbert, Charles Malleson, Adrian Hilton, and John Collo-mosse. Total capture: 3d human pose estimation fusing video and inertial sensors. In *Proceedings of the 28th British Machine Vision Conference*, pages 1–13, 2017. 2, 6
- [33] Tom Van Wouwe, Seunghwan Lee, Antoine Falisse, Scott Delp, and C Karen Liu. 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*, pages 2513–2523, 2024. 2, 3, 5, 16, 17
- [34] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 17
- [35] Timo Von Marcard, Bodo Rosenhahn, Michael J Black, and Gerard Pons-Moll. Sparse inertial poser: Automatic 3d human pose estimation from sparse imus. In *Computer graphics forum*, pages 349–360. Wiley Online Library, 2017. 2
- [36] Jian Wang, Diogo Luvizon, Weipeng Xu, Lingjie Liu, Kripasindhu Sarkar, and Christian Theobalt. Scene-Aware Egocentric 3D Human Pose Estimation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13031–13040, Vancouver, BC, Canada, 2023. IEEE. 1
- [37] Prof. Richard Wilkinson. 6.1 Classical MDS — Multivariate Statistics — rich-d-wilkinson.github.io. <https://rich-d-wilkinson.github.io/MATH3030/6.1-classical-mds.html>. 4
- [38] Yinghao Wu, Lu Yin, Shihui Guo, Yipeng Qin, et al. Accurate and steady inertial pose estimation through sequence structure learning and modulation. *Advances in Neural Information Processing Systems*, 37:42468–42493, 2024. 2
- [39] Yinghao Wu, Shihui Guo, and Yipeng Qin. Moda: Motion-drift augmentation for inertial human motion analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27771–27781, 2025. 2
- [40] Xsens. Xsens motion capture. <https://www.xsens.com>, 2024. 2, 12
- [41] Weipeng Xu, Avishek Chatterjee, Michael Zollhoefer, Helge Rhodin, Pascal Fua, Hans-Peter Seidel, and Christian Theobalt. Mo2Cap2: Real-time Mobile 3D Motion Capture with a Cap-mounted Fisheye Camera, 2019. arXiv:1803.05959 [cs]. 1
- [42] Ying Xue, Jiayi Jiang, Rayan Armani, Dominik Hollidt, Yi-Chi Liao, and Christian Holz. Group inertial poser: Multi-person pose and global translation from sparse inertial sensors and ultra-wideband ranging. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 24910–24921, 2025. 2, 3, 6, 8, 12
- [43] Yuan Yao, Shifan Jiang, Yangqing Hou, Chengxu Zuo, Xinrui Chen, Shihui Guo, and Yipeng Qin. Tof-ip: time-of-flight enhanced sparse inertial poser for real-time human motion capture. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. 3
- [44] Brent Yi, Vickie Ye, Maya Zheng, Lea Müller, Georgios Pavlakos, Yi Ma, Jitendra Malik, and Angjoo Kanazawa. Estimating body and hand motion in an ego-sensed world. *arXiv preprint arXiv:2410.03665*, 2024. 3
- [45] Xinyu Yi, Yuxiao Zhou, Marc Habermann, Soshi Shimada, Vladislav Golyanik, Christian Theobalt, and Feng Xu. Physical inertial poser (pip): Physics-aware real-time human motion tracking from sparse inertial sensors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13167–13178, 2022. 2, 3, 5, 6, 8, 12, 17
- [46] Xinyu Yi, Yuxiao Zhou, and Feng Xu. Physical non-inertial poser (pnp): modeling non-inertial effects in sparse-inertial human motion capture. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 2, 5, 6, 12, 17
- [47] Xinyu Yi, Shaohua Pan, and Feng Xu. Improving global motion estimation in sparse imu-based motion capture with physics. *ACM Transactions on Graphics (TOG)*, 44(4):1–16, 2025. 2, 6, 8
- [48] Siwei Zhang, Bharat Lal Bhatnagar, Yuanlu Xu, Alexander Winkler, Petr Kadlecek, Siyu Tang, and Federica Bogo. Rohm: Robust human motion reconstruction via diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14606–14617, 2024. 1, 3, 5
- [49] Yu Zhang, Songpengcheng Xia, Lei Chu, Jiarui Yang, Qi Wu, and Ling Pei. Dynamic inertial poser (dynaip): Part-based motion dynamics learning for enhanced human pose estimation with sparse inertial sensors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1889–1899, 2024. 6, 16
- [50] Ce Zheng, Wenhan Wu, Chen Chen, Taojiannan Yang, Si-jie Zhu, Ju Shen, Nasser Kehtarnavaz, and Mubarak Shah. Deep learning-based human pose estimation: A survey. *ACM Computing Surveys*, 56(1):1–37, 2023. 1
- [51] Xiaozheng Zheng, Zhuo Su, Chao Wen, Zhou Xue, and Xiaojie Jin. Realistic full-body tracking from sparse

observations via joint-level modeling. *arXiv preprint arXiv:2308.08855*, 2023. [2](#)

- [52] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the Continuity of Rotation Representations in Neural Networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5738–5746. IEEE, 2019. [4](#)
- [53] Chengxu Zuo, Yiming Wang, Lishuang Zhan, Shihui Guo, Xinyu Yi, Feng Xu, and Yipeng Qin. Loose inertial poser: Motion capture with imu-attached loose-wear jacket. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2209–2219, 2024. [2](#)
- [54] Chengxu Zuo, Jiawei Huang, Xiao Jiang, Yuan Yao, Xiangren Shi, Rui Cao, Xinyu Yi, Feng Xu, Shihui Guo, and Yipeng Qin. Transformer imu calibrator: Dynamic on-body imu calibration for inertial motion capture. *ACM Transactions on Graphics (TOG)*, 44(4):1–14, 2025. [2](#)

A. Supplementary

A.1. Training and Evaluation Details

In the following section, we provide additional implementation and evaluation details.

The losses described in Sec. 3.2.3 are weighted by the following weights: $\lambda_{\text{simple}} = 5$, $\lambda_{\text{tran}} = 5$, $\lambda_{\text{smpl}} = 1$, and $\lambda_{\text{vel}} = 0.3$. The velocity loss is activated after 10 epochs, once the prediction has stabilized.

UDP uses an LSTM with a hidden state size of 512 with 2 layers and a dropout rate of 0.4 for the denoiser network. The motion history length is set to $N_H = 30$, and the motion window size is $N = 170$ frames at 60 FPS during training. The model is trained for 150 epochs with a weight decay of $1e-5$, an initial learning rate of $1e-4$, and step decay with a factor of $\gamma = 0.33$ every 50 epochs. We use a square root diffusion noise schedule.

The rotation estimator uses a hidden state dimension of 512, 3 layers, and a dropout rate of 0.2. We enable UWB-Diffusion Guidance at diffusion step $t = 20$, with $\lambda = 500$ and a weaker $\lambda = 50$ for the GIP-DB and UIP-DB, where UWB-signals are less reliable. Overall, the model contains 11.3M parameters, and training takes approximately 5 hours on an RTX 4090 GPU with a batch size of 256.

Following prior work [4, 19, 45], the model is initially trained on the synthetic AMASS dataset with DanceDB held out and directly evaluated on TotalCaptureReal and DanceDB. This is the complete list of Datasets used during training: HumanEva, MPI_HDM05, SFU, MPI_mosh, Transitions_mocap, SSM_synced, CMU, TotalCapture, Eyes_Japan_Dataset, KIT, BMLmovi, EKUT, TCD_handMocap, ACCAD, BioMotionLab_NTroje, BML-handball, MPI_Limits, Dfaust.67.

DIP-IMU and TotalCapture provide IMU signals coming from an XSense suit [40]. For TotalCapture, we use the DIP sensor calibration. All AMASS training data and DanceDB use synthesized IMU data, obtained via finite differences on the motion data following [4, 45]. UWB measurements are synthesized from the motion ground truth for AMASS training data, DanceDB, TotalCapture, and DIP-IMU. UIP-DB [4] and GIP-DB [42] obtain all IMU data through off-the-shelf, low-cost commercial IMU and UWB sensors, and we use these real-world IMU signals and UWB measurements directly. For DIP-IMU, the model is fine-tuned on the training split for 30 epochs and then evaluated on the test split. As UIP-DB, and GIP-DB present a more challenging scenario with lots of real-world sensor noise, we train this model with noise augmentations. We introduce Gaussian noise with $\sigma_{\text{acc}} = 0.6$ to the acceleration measurements and apply random rotation augmentation with Gaussian noise of $\sigma_{\text{ori}} = 0.15$, and add a random Gaussian bias of $\sigma_{\text{ori-bias}} = 0.054$ and a stronger weight decay of $4e - 5$. For UWB distance noise, we add Gaussian noise with the

standard deviation based on the reported mean UWB distance errors from UIP-DB [4] scaled by 1.5. The model is pretrained for 50 epochs using 200 diffusion steps to improve robustness to noise. For fine-tuning, we take the model pretrained on purely synthetic data and train it for 5 epochs on the training split of either UIP-DB or GIP-DB (depending on which dataset is being evaluated). We then evaluate the model on the corresponding test split.

For consistency, we report results from the original authors or from previously published works. If the required metrics are not available, we reproduce the results ourselves by utilizing the provided weights. We retrain the network if no weights are provided for that evaluation. Since PIP does not provide training code, we implemented our own training pipeline and trained the model from scratch, ensuring that no test data from DanceDB leaked into the training process. PNP [46] does not provide data processing code and provide the reported results on TotalCapture and DIP. DynaIP does not predict translation; hence, we only report the results of the translation-free DIP dataset. UMotion uses a slightly different evaluation scheme; we used the authors’ provided model weights and reran the results. During inference, we disable the use of ground-truth distances for updating the Kalman filter covariances, see [here](#). Instead, we assume an 8 cm standard deviation for the UWB measurements and set the measurement covariance accordingly. Additionally, we report results in two settings: First, the zero-shot results reported on UIP-DB and GIP-DB use the raw UWB distances. Second, the finetuned results that clean the raw UWB distances using the training split, as seen [here in the authors code](#). The reported results are obtained following the same evaluation framework as previous work [4, 45, 46]. Evaluation metrics follow the conventions established in prior work [4, 45, 46], excluding end joints that cannot be reliably estimated from sparse sensor input—specifically, the wrists, hands, root, toes, and ankles.

A.2. Additional Evaluations

We provide additional qualitative comparisons in the accompanying video and in Figures 5 and 6. As shown, UDP achieves better foot and arm placement than prior work, particularly in squatting motions, where it better estimates foot positioning while maintaining precise torso alignment.

Conversely, methods relying on post-optimization physics simulations (e.g., PIP, UIP, PNP) often struggle with highly dynamic motions, such as dance routines involving rapid twirls or intentional foot sliding (see the bottom row of Fig. 5). In these scenarios, the physics optimizer may be a bottleneck due to rigid constraint modeling. Specifically, when the network predicts a foot contact, the optimizer enforces strict zero-velocity or high-friction constraints. Consequently, intentional sliding motions may lead to wrong motion estimation when the model inaccurately

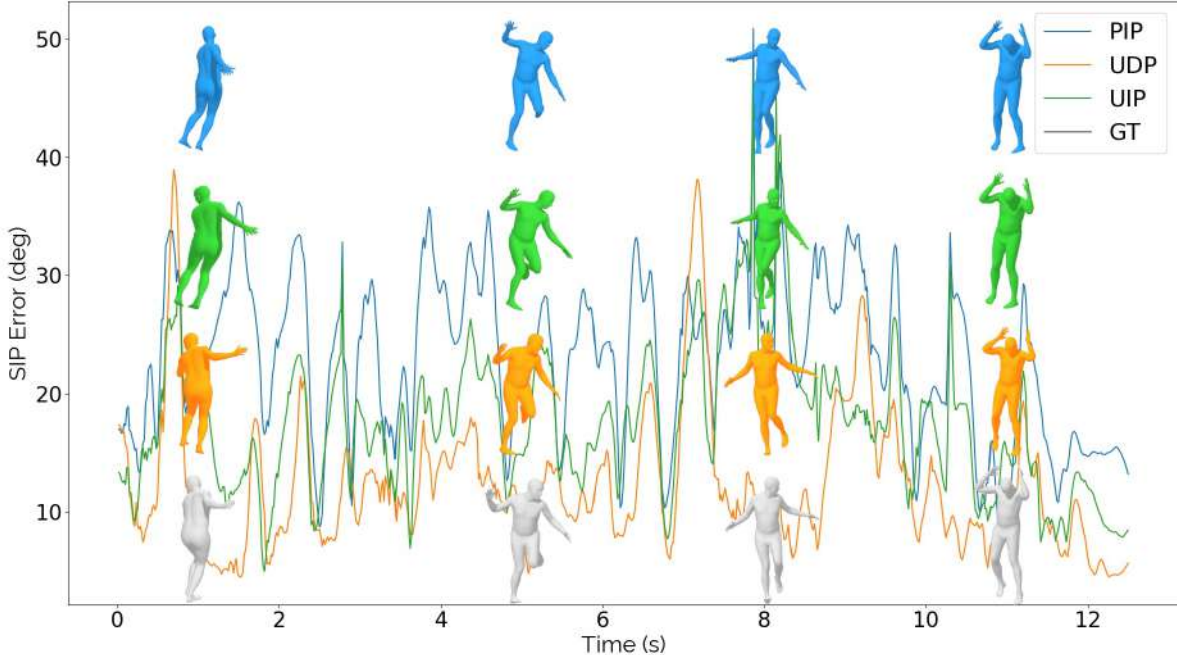


Figure 4. Qualitative results over a sequence of DanceDB. UDP has generally lower SIP error compared to previous methods.

predicts foot contacts. In contrast, UDP, being purely data-driven, is not bound by contact enforcement, friction models, or hyperparameter choices within the physics model and avoids such interference, allowing for the accurate reconstruction of complex maneuvers.

While physics optimizers remain a compelling alternative for applications where strict physical plausibility (e.g., non-penetration) takes precedence over tracking fidelity, our results demonstrate that UDP yields smooth, accurate motion, particularly in limb placement, while benefiting from the simplicity of a purely data-driven approach.

A.2.1. Additional Ablation Studies

The following sections extend the analysis in the main paper by providing detailed ablation studies regarding inference dynamics, noise scheduling, temporal lookahead, and architectural alternatives.

A.2.2. Sampling Steps, Inference and Realtime

UDP operates in real-time and significantly improves inference speed over UIP. Unlike UIP, it avoids the post-processing physics optimizer, which formulates refinement as a Quadratic Programming (QP) problem. This QP solver acts as a computational bottleneck, as it relies on serial CPU execution and cannot benefit from GPU acceleration. In contrast, UDP utilizes purely neural-based inference that is fully accelerated on the GPU.

Tab. 5 shows that UDP achieves 3.6 times faster inference compared to UIP when running on an RTX 4090. By

using DDIM [28], we can reduce sampling steps to just 5. This increases inference speed by a factor of 14 compared to DDPM sampling with 50 steps, while the SIP error increases by only 4.8%. We observe that the pose prediction quality stabilizes quickly: the transition from 50 steps (DDPM) to 25 steps (DDIM) induces the largest drop in quality, while further reducing the steps to 5 has a negligible impact. In comparison to UIP, UDP with 5 sampling steps enables 50 times faster inference while still delivering superior output quality.

Table 5. Evaluation of the total inference time (TT) on the TotalCapture dataset (approx. 2,466,s duration) and the mean inference time per frame (MTPF). For 60FPS real-time inference, 16,ms MTPF or lower is required. Reducing the number of diffusion steps reduces pose estimation accuracy slightly but increases inference speed drastically.

Metric	SIP (°) ↓	JPE (cm) ↓	TT (s) ↓	MTPF (ms) ↓
TIP	11.36	5.15	568	3.938
UIP	10.70	5.11	1,342	7.545
UDP DDIM 5	9.38	4.01	26	0.151
UDP DDIM 10	9.34	3.99	46	0.265
UDP DDIM 25	9.30	3.98	110	0.627
UDP DDPM 50	8.95	3.76	371	2.117

A.2.3. Module Ablation

We showed that the SPL module and UWB-Diffusion Guidance improve pose estimation under ideal sensing conditions (TotalCapture). Tab. 6 demonstrates that both com-



Figure 5. Qualitative results on DanceDB. UDP often estimates fast motions more accurately than other methods.

ponents also improve accuracy under noisy sensor conditions. Notably, adapting the UWB guidance coefficient λ to the sensor noise level enables improved performance, as noisy distance measurements with strong guidance can induce wrong limb predictions.

A.2.4. Noise Schedules

Additionally, we evaluated various noise schedules, as diffusion models are often sensitive to the specific noise distribution of the data. We hypothesized that schedules adding noise more gradually (e.g., cosine or exponential) would be beneficial, given that error accumulation along the kinematic chain causes big joint errors with even minimal joint

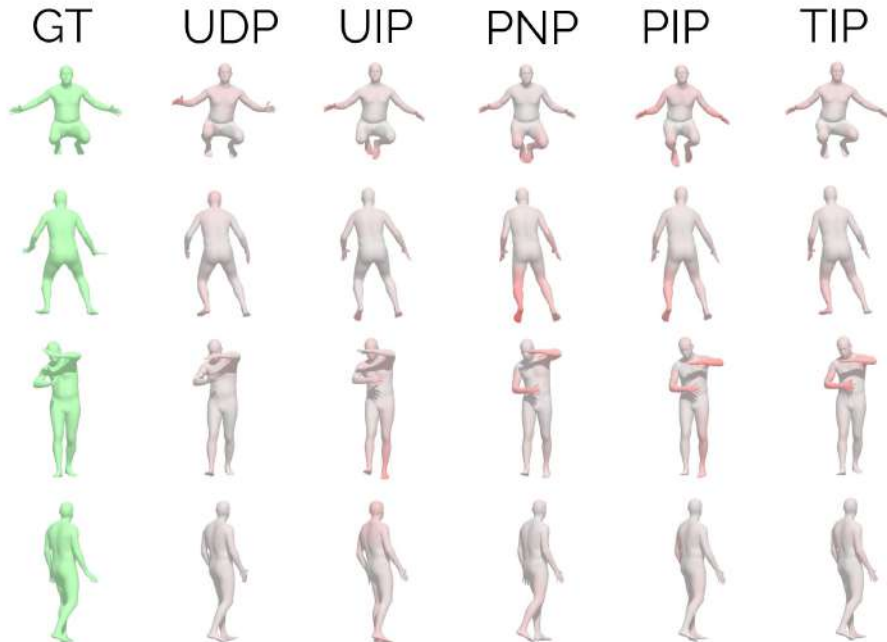


Figure 6. Qualitative results on TotalCapture. UDP often estimates limb positions more accurately than others.

Table 6. GIP-DB ablation on GIP-DB finetuned: SPL and UWB-Guidance improve metrics under real-world noise.

Metric	SIP ($^{\circ}$) \downarrow	GAE ($^{\circ}$) \downarrow	JPE (cm) \downarrow
UDP w/o SPL	16.03	15.09	7.08
UDP w/o UWB Guidance	16.30	14.86	6.90
UDP $\lambda = 500$	15.35	15.12	7.40
UDP	15.33	14.34	6.68

angle noise. As illustrated in Fig. 7, aggressive schedules like the square root schedule can rapidly degrade the pose structure, leading to visually distorted samples even at early noise steps. Counter-intuitively, the more aggressive square root schedule led to the best performance. As shown in Tab. 7, the square root schedule consistently outperforms all other tested schedules across all metrics.

Metric	SIP ($^{\circ}$) \downarrow	GAE ($^{\circ}$) \downarrow	JPE (cm) \downarrow
cosine	9.18	10.25	3.87
exponential	9.29	10.42	4.00
linear	9.46	10.48	4.09
sqrt	8.95	10.19	3.76

Table 7. UDP trained with different noise schedules.

A.2.5. Shaped Evaluation

UDP predicts SMPL joint angles and root translation, while the SMPL shape parameters β serve as optional conditioning input when available. Our main evaluation sets $\beta = \mathbf{0}$ (mean shape) for the model input and evaluation. β can be user-provided or estimated from UWB distances (as in UMotion [20]). We show that utilizing the known β reduces UDP’s errors (Tab. 9), improving robustness across diverse body proportions. When using ground-truth shape parameters both as model input and for evaluation, UDP achieves consistent gains across all metrics, most notably a 10% reduction in SIP error. Since shape influences both inter-sensor distances and inertial dynamics, providing β further enhances motion estimation accuracy.

A.2.6. No Lookahead

We evaluate the model’s performance in an online setting by restricting the evaluation to the last frame of the input window (Last Frame Only - LFO). In this scenario, the model relies solely on past observations without access to future context. As shown in Tab. 10, removing future context results in slightly higher errors, i.e., approximately 2.7% in SIP, while the global angle error remains largely unaffected. Additionally, reducing the context window from 3s to 1s increases JPE by 3%, with a marginal impact on angle errors. Crucially, even in this restricted online setting with no lookahead, UDP continues to outperform baseline methods.



Figure 7. Noise schedules of different schedules visualized on $q(x_t|x_0)$ in equal steps where the maximum step is $T = 50$.

Table 8. Ablation Study results on TotalCapture.

Metric	SIP Error (°) ↓	GAE (°) ↓	JPE (cm) ↓	TE @ 2m ↓	TE @5m ↓	Jitter (0.465)
UDP w/ Transformer	10.70	12.10	4.79	0.35	0.55	0.115
UDP w/ velocity representation	9.30	10.35	3.95	0.28	0.49	0.123
UDP w/ FK loss	9.01	10.29	3.89	0.20	0.32	0.124
UDP (ours)	8.95	10.19	3.76	0.20	0.33	0.124

Table 9. Evaluation DanceDB. Shape improves motion estimation.

Model	SIP (°) ↓	GAE (°) ↓	JPE (cm) ↓
UDP w/ known shape	10.68	9.35	4.30
UDP w/o known shape	11.79	9.91	4.67

A.2.7. Removal of FK loss

Unlike previous methods [33, 49], we do not utilize a forward kinematics loss during training. Using the FK loss increases the training time from approximately 5h to 14h. Furthermore, as shown in Tab. 8, the forward kinematics loss does not compromise performance; in fact, it yields a marginal reduction in pose estimation accuracy. We at-

LFO	W length	SIP (°) ↓	GAE (°) ↓	JPE (cm) ↓
✓	1 s	9.36	10.45	4.14
	1 s	9.15	10.49	3.84
✓	3 s	9.19	10.11	4.01
	3 s	8.95	10.19	3.76

Table 10. Ablation on lookahead and window size. 'LFO' denotes evaluating only the last frame of the window.

tribute this to unstable gradients arising from the kinematic chain: despite similar loss values, the FK loss produces up to 10× higher gradient norms and 1000× larger peak gradients with respect to SMPL joint angles compared to a direct L1 loss. This indicates that direct supervision of local joint

angles is sufficient.

A.2.8. Model Architecture

Previous works on pose estimation [17, 19, 33] and pose generation [30] have employed transformers. However, our experiments show that a transformer encoder [34] performs significantly worse as a denoising model, with a 27% increase in joint position error, see Tab. 8. We implemented a 4-layer transformer encoder with the same token dimension as UDP, using sinusoidal position encodings. Overall, our results indicate that the local neighborhood priors captured by LSTMs are more effective than those learned by transformers, given the dataset size used in our experiments.

A.3. Translation Results

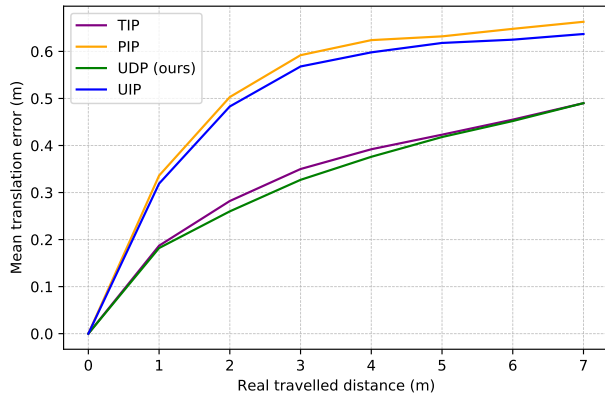


Figure 8. The cumulative translation error over segment of d meters on DanceDB. Purely learned methods UDP and TIP, tend to perform better on complicated dance moves.

UDP’s primary contribution lies in improving the pose prediction by modeling geometric constraints using inter-sensor measurements. As demonstrated in the main paper, the Spatial Layout Module and UWB-Diffusion Guidance significantly improve local pose estimation, particularly for arm and leg placement. In this section, we describe the modeling of translation prediction and the evaluation of UDP’s translation prediction.

Translation Representation. Unlike previous methods [4, 19, 33, 45, 46] that model translation implicitly by predicting frame-to-frame velocity, UDP outputs absolute translation predictions for individual window, each originating at the origin. To form a continuous, long-term sequence, these windows are concatenated such that each window begins at the final position of the preceding one. Velocity-based approaches require numerical integration to obtain global positions, a process where small inaccuracies accumulate over time. By representing translation as an absolute position directly, we aim to mitigate this integration drift.

Trajectory of Translations (XZ Plane)

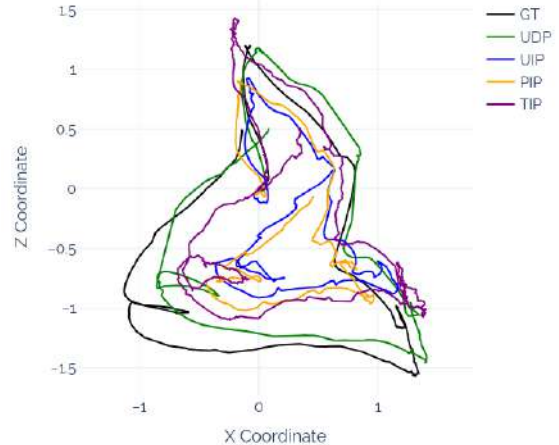


Figure 9. An example root trajectory in xz-plane. PIP and UIP tend to underestimate translation. UDP follows the trajectory closer. TIP has a higher level of jitter.

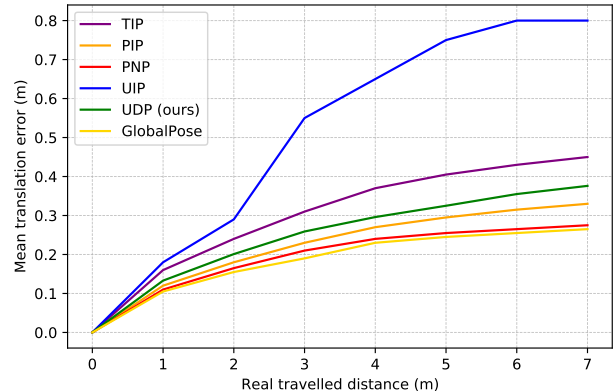


Figure 10. The cumulative translation error over segment of d meters on TotalCaptures. All models, but UIP, perform similarly in translation quality.

Translation Evaluation. We investigate translation prediction performance using the established cumulative translation error, which measures the drift from the ground truth trajectory after the subject has traveled a distance of d m. We report results on DanceDB and TotalCapture for all methods for which metrics are reported or could be reproduced directly. Overall, UDP achieves translation accuracy that is competitive with existing approaches across all datasets. On DanceDB, it outperforms the baseline methods, see Fig. 8. In contrast, on TotalCapture, methods that incorporate a physics optimizer, i.e., PIP, PNP, and GlobalPose, achieve lower drift, outperforming UDP by approxi-

mately 5 cm over 5 m (see Fig. 10).

We investigate the varying results qualitatively. On DanceDB, UDP generally yields more dynamic and accurate translations, whereas PIP and UIP tend to underestimate displacement, especially during twirls or intentional foot sliding. This is visible in Fig. 9, where UDP follows the ground truth trajectory closely.

On TotalCapture, UDP exhibits drift when the translation direction is estimated incorrectly. While the magnitude of fast, large-scale movements is tracked well, small directional errors can occasionally accumulate into larger drift.

We attribute these findings to two primary factors. First, UDP significantly improves foot and leg placement. Since human translation is intrinsically linked to gait, improved estimation of stride length and foot positioning reduces translation error, as observed on DanceDB. Second, physics optimized baselines benefit from explicit contact and dynamics constraints, which aid in tracking translation, as seen on TotalCapture. However, strict contact priors, such as zero-velocity assumptions, may degrade performance when the motion violates these constraints, such as during the intentional sliding maneuvers in DanceDB.

Overall, UDP demonstrates strong performance in local pose estimation due to the incorporation of on-body inter-sensor constraints, though it yields mixed results regarding global translation. Future work could investigate how pairing physics optimization with UDP, or the incorporation of external UWB anchors, can further improve translation accuracy.

A.3.1. Ablation: Absolute Translation Prediction

Existing methods such as UIP, PIP, TIP, and DiffusionPoser predict between-frame translation, i.e., velocity. In contrast, we employ an absolute translation representation, normalized to the origin. This representation improves not only the translation results but also marginally enhances local pose estimation. Experimentally, switching to a velocity-based representation increases translation drift by 16 cm over a 5 m distance. As shown in Tab. 8, UDP yields superior performance in both local pose and translation when predicting positions directly, highlighting the benefit of this approach.