DOI: 10.1111/cgf.15057

EUROGRAPHICS 2024 / A. Bermano and E. Kalogerakis
(Guest Editors)

COMPUTER GRAPHICS *forum*
Volume 43 (2024), Number 2

DivaTrack: Diverse Bodies and Motions from Acceleration-Enhanced Three-Point Trackers

Dongseok Yang[1], Jiho Kang[1], Lingni Ma[2], Joseph Greer[2], Yuting Ye[2] and Sung-Hee Lee[1]

[1]Korea Advanced Institue of Science and Technology, [2]Meta Reality Labs

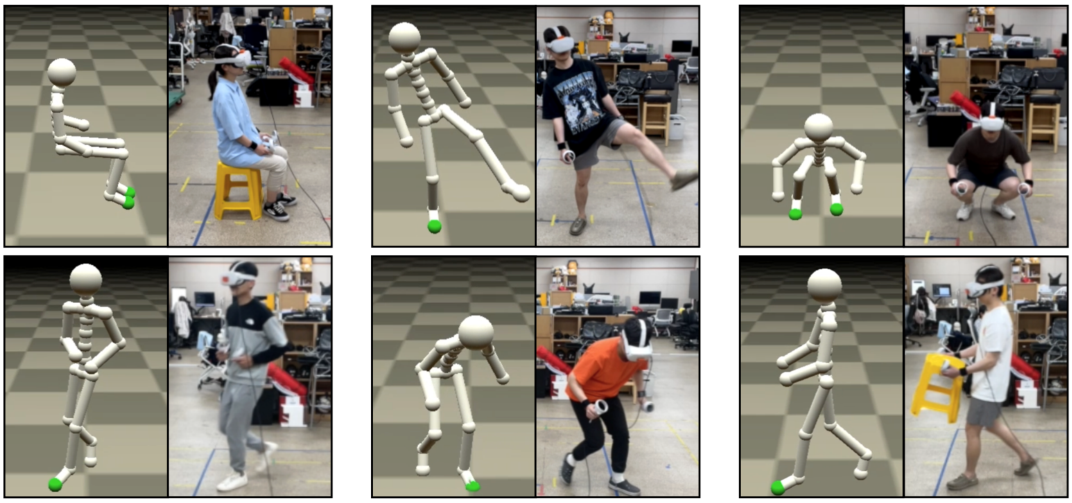

Figure 1: *Our method generates full-body motions for diverse body proportions and activities from only three-point trackers and IMUs.*

Abstract

Full-body avatar presence is important for immersive social and environmental interactions in digital reality. However, current devices only provide three six degrees of freedom (DOF) poses from the headset and two controllers (i.e. three-point trackers). Because it is a highly under-constrained problem, inferring full-body pose from these inputs is challenging, especially when supporting the full range of body proportions and use cases represented by the general population. In this paper, we propose a deep learning framework, DivaTrack, which outperforms existing methods when applied to diverse body sizes and activities. We augment the sparse three-point inputs with linear accelerations from Inertial Measurement Units (IMU) to improve foot contact prediction. We then condition the otherwise ambiguous lower-body pose with the predictions of foot contact and upper-body pose in a two-stage model. We further stabilize the inferred full-body pose in a wide range of configurations by learning to blend predictions that are computed in two reference frames, each of which is designed for different types of motions. We demonstrate the effectiveness of our design on a large dataset that captures 22 subjects performing challenging locomotion for three-point tracking, including lunges, hula-hooping, and sitting. As shown in a live demo using the Meta VR headset and Xsens IMUs, our method runs in real-time while accurately tracking a user's motion when they perform a diverse set of movements.

CCS Concepts
• ***Computing methodologies*** → ***Motion capture;***

1. Introduction

Self-embodied digital humans are a cornerstone of Virtual, Augmented, and Mixed Reality (VR/AR/MR) applications as tracking the users' full-body motion is essential for natural interactions in the virtual environment. However, it is a challenging task with information available from today's VR/AR/MR devices, which typically only provide six degrees of freedom (DOF) pose data of the head and hands, i.e. three-point input. Although some headsets are equipped with cameras, they are often limited in field of view and resolution, unsuited for full-body pose estimation. An ideal solution would rely on on-device sensors alone to produce plausible full-body motions for any user performing any daily activities.

Recent work tackles this challenge with deep learning methods that directly map sparse three-point inputs to full-body motion [JSQ*22; ACB*22], or to joint torques for simulating the full-body motion [WWY22; YLHX22]. They have demonstrated success for simple locomotion, but face degraded quality in less common activities such as running backward and carrying objects, or in cases where the lower-body pose is ambiguous with respect to the three-point input, such as kicking, lunging, crouching, or simply looking around while walking. Additionally, they only target an average body size and therefore cannot handle large variations in body proportions.

In this paper, we propose a novel model, DivaTrack, for generating full-body motions from three-point tracker inputs and IMU signals. DivaTrack is designed to work in real-time and is capable of handling diverse body shapes and motions. A lightweight calibration process is first employed to determine the user's body proportions from a few dedicated poses. Our algorithm then infers the full-body poses that apply to the estimated body proportion. We incorporate several insights into the design of our system, so it can better generalize to a wide variety of body sizes and activities.

First, we augment the three-point six DOF input with linear accelerations from Inertial Measurement Units (IMUs) attached to the headset and wrists as input to our system. IMUs are ubiquitous in consumer electronic devices [MAG*23] and are already available in all VR devices. They are currently primarily used to infer orientation, but we find that their high-frequency linear acceleration signals have a strong correlation with impacts on the body, such as foot contact events. In fact, body-mounted IMUs are widely used for gait event detection in the field of biomechanics [Moe98; LMB10; GHR*16; CNGG18; PLB*22; DAM*21]. Foot contacts have been shown to improve the generation of plausible lower body motions [YKL21; TCL23], so we predict them from the input with linear accelerations and use them for full-body pose prediction.

Second, we find that a crucial ingredient to generalization is the choice of reference coordinate frame for data normalization [MYGY19]. A common choice in locomotion research is to choose the root coordinate frame of the character and set its origin on the ground plane. This is not directly applicable to three-point tracking because the root motion is unknown. A reasonable alternative is to use the headset coordinate frame. We observe that it performs well when the user's head is aligned with their direction of movement, but causes issues with motions such as walking backward and looking around while walking. On the other hand, a movement direction aligned with a past trajectory excels in these cases but causes problems if the subject is stationary. Therefore, we devised a two-stream architecture that learns to blend results from these two complementary coordinate frames based on the input.

Finally, we split the motion network into an upper-body model (UC-Model) and a lower-body model (L-Model). The UC-Model is a regression model because its output is relatively well-constrained from three-point and IMU inputs. The L-Model is a Conditional Variational Autoencoder (CVAE) [SLY15] conditioned on the upper-body pose and contact states predicted from the UC-Model. The CVAE models the ambiguity and correlation between its input and the lower-body. This design allows us to swap out the upper-body pose module, e.g. with a simple inverse kinematics (IK) solver, without changing the lower-body generation.

To verify the above design choices, we collected a large motion dataset with synchronized ground-truth body motions and IMU signals using the Xsens system [RLS08], from a diverse pool of subjects performing challenging everyday activities. Even though there already exist body motion datasets that come with IMU data, such as TotalCapture [TGM*17] and HPS [GMSP21], these are not designed to showcase use cases and challenges of three-point tracking and have limited subject diversity. In our dataset, we specifically design movement protocols to demonstrate difficult scenarios using three-point inputs, such as when the inputs have weak correlations to the lower body, or when common assumptions break down (e.g. lying down and rolling on the floor). Testing on the full range of our subject base and difficult motions such as lunges and hula-hooping confirms the effectiveness of our solution. We further demonstrate the applicability of our methods to realistic scenarios by developing a real-time live demo with a Meta VR headset and Xsens IMUs.

To summarize, this paper has the following main contributions:

- Enhancing three-point tracker signals with IMU accelerations for more accurate foot contact state and full-body pose generation.
- A combination of upper-body inference and conditional lower-body generation models to match the given three-point constraints and produce plausible full-body motion.
- A novel two-stream blending architecture to utilize complementary reference coordinates for a diverse spectrum of motions.
- A lightweight calibration process to accommodate diverse body proportions.
- A large human motion dataset (16.5 hours) that includes synchronized ground truth body motion with IMU signals, 22 different body proportions, and diverse motion categories.

2. Related Work

Human motion tracking is widely studied in computer graphics and vision communities. Fully vision-based methods use monocular or multiple cameras to track allocentric motions of single or multiple people [GPR*23; YPMK23; KBM*23; CYZ*23; LHG*23; RSJ21; SGXT20; KZFM19]. With the increasing popularity of AR/VR, understanding self-body poses from wearable sensors has received renewed attention. Recent work explores cameras mounted on the head [WLX*23; AWS*22; TAP*20; LHYK21], wrist [LLD*22] and both [ASF*22]. In this paper, we are interested in sparse body-worn trackers and the remainder of this section focuses on prior work that leverages wearable sensors. In addition, we cover literature related to generative motion synthesis, which the methods in this work draw heavily from.

2.1. Inertial-only tracking

Fully inertial-based human motion tracking has been in use for multiple decades. Commercial products, e.g., XSens [RLS08] and

Rokoko are widely tested systems, capable of high-quality motion captures in the wild. However, these solutions are highly encumbering, requiring users to wear 17 IMUs on their bodies. Several methods have been proposed to reduce this number. The pioneering work by Marcard et al. [vMRBP17] recovers full-body pose from 6 IMUs via batch optimization over the entire sequence. As the method is not suitable for online applications, recent learning-based methods focused on statistical human motion modeling. Huang et al. [HKA*18] uses a bidirectional recurrent neural network (BiRNN [SP97]) with LSTMs [HS97] to achieve coherent prediction. A similar approach is investigated by Nagaraj et al. [NSLW20]. Both methods focus on local poses without solving for the global position of the wearer. TransPose [YZX21] addresses this issue by regressing the body velocity from foot contact and IMU data and fusing the outputs using contact predictions. These works are extended in PIP [YZH*22] to reduce artifacts by refining the motion with physical constraints, similar to [SGXT20]. EgoLocate [YZH*23] further improves global tracking accuracy by fusing with vision-based tracking from a head-mounted camera. Inspired by the success of transformer networks [VSP*17], TIP [JYG*22] proposes a simple attention-based network with autoregression that eliminates delay with improved accuracy.

2.2. Tracking with sparse six DOF input.

The prevalence of consumer devices in the AV/VR industry has increased the accessibility of head-mounted devices (HMDs) and controllers that provide six DOF pose estimates. This introduces a novel problem setup: to infer full-body pose from the given poses of the head and hands, so-called three-point tracking. Early attempts to solve this problem include CoolMoves [AOG*21], which uses hand velocity and acceleration to retrieve and blend the top-K body poses on a limited activity set. Ponton et al. [PYAP22] extend this work by experimenting with a diverse animation database. LoBSTr [YKL21] adds the pelvis pose to develop a hybrid solution, where the upper-body is solved with IK and the lower-body and foot contacts are predicted via gated recurrent units (GRUs) [CGCB14]. Following this work, Ye et al. [YLHX22] uses two GRUs to predict the upper and lower body independently and adopt reinforcement learning (RL) to train a pose correction policy from physical constraints. Similarly, QuestSim [WWY22] and QuestEnvSim [LSY*23] learn a policy that generates joint torques to drive a physics simulation of body motion that matches the inputs. AvatarPoser [JSQ*22] show transformer networks are efficient in learning the full-body motion, which is further improved by EgoPoser [JSM*23] with a SlowFast [FFMH19] module and body shape prediction. Dittadi et al. [DDC*21] show that a variational auto-encoder (VAE) [KW14] trained with full-body pose can generate plausible motion from incomplete three-point input. Similarly, Milef et al. [MSK23] shows a conditional VAE can produce diverse poses from four-point input.

2.3. Neural generative model for motion synthesis.

Human motion synthesis has an extensive body of prior work. Early research explored probabilistic approaches, e.g., principal component analysis (PCA) [SHP04; CH05; LZWM06], Gaussian mixture models (GMMs) [MC12] and Gaussian processes [GMHP04; WFH08; LWH*12]. Pioneered by Holden et al.'s work on motion manifolds [HSKJ15], recent work leverages modern deep generative neural architectures. Generative adversarial networks (GANs [GPM*14]) have been adopted by multiple methods, e.g., in [FNM19] to condition body gestures on speech, in [WCX21] to synthesize and control body motions, in [SZKZ20] to generate interactive movements and in [LAZ*22] to synthesize motions from a sample sequence. VAE is another widely used model that enables probabilistic sampling to generate motion output. For example, DeepPhase [SMK22] leverages the frequency domain to learn periodic motion from unstructured data. Moreover, CVAEs [SLY15] have been shown to be powerful architectures for motion synthesis, where the condition can be motion history [LZCV20], audio [LYL*19; LYC*20], action types [PBV21] and three-point input [MSK23]. As alternatives, normalizing flow and neural distance fields are adopted by MoGlow [HAB20] and Pose-NDF [TAL*22], respectively. Following the recent success of diffusion models for visual synthesis, Tevet et al. [TRG*23; STKB23], Zhang et al. [ZCP*22], EDGE [TCL23], MoFusion [DMGT23] and EgoEgo [LLW23] show diffusion-based motion synthesis driven by text, music or head motion. The recent work AGRoL [DKP*23] conditioned a diffusion model with three-point input and showed remarkable offline motion generation.

3. DivaTrack

Figure 2 summarizes our framework to estimate full-body motion from the six DOF poses and linear accelerations of three-point trackers. At the beginning of a session, we use a one-time calibration module to infer the user's body skeleton from six poses (Section 3.3). We then estimate the user's full-body pose at every frame using the F-Model (Full-body Model) that contains two parallel streams of identically-structured P-Models (Pose Model). A P-Model predicts a full-body pose in two stages: the upper-body *inference* and contact prediction (Section 3.4), followed by the lower-body *generation* (Section 3.5). Each P-Model is trained with a different reference frame suitable in complementary configurations (Section 3.2). The two P-model outputs are fused together via a motion blending network (B-model). The joint rotations from the blended pose are applied to the calibrated skeleton to output the final full-body pose (Section 3.7).

3.1. Network Input and Output

The input state from a tracker $\mathbf{x} = \{\mathbf{p}, \mathbf{q}, \mathbf{v}, \mathbf{a}\} \in \mathbb{R}^{21}$ is composed of position $\mathbf{p} \in \mathbb{R}^3$, rotation $\mathbf{q} \in \mathbb{R}^6$, velocity $\mathbf{v} \in \mathbb{R}^9$, and linear acceleration $\mathbf{a} \in \mathbb{R}^3$. Rotation $\mathbf{q}$ follows the 6D representation commonly used in deep learning [ZBJ*19]. Velocity $\mathbf{v}$ is approximated by finite difference of $(\mathbf{p}_t, \mathbf{q}_t)$ and $(\mathbf{p}_{t-1}, \mathbf{q}_{t-1})$. We denote the three tracker states as $\mathbf{x}^H$, $\mathbf{x}^L$, $\mathbf{x}^R$, for the head, left hand, and right hand, respectively. Our models take as input a trajectory of tracker states within a time window of length $l+1$ ($l = 28$ in this work). The end effector input window at time t is $\mathbf{X}_t^{EE} = \{\mathbf{x}_{t-l}^H, \mathbf{x}_{t-l}^L, \mathbf{x}_{t-l}^R, \ldots, \mathbf{x}_t^H, \mathbf{x}_t^L, \mathbf{x}_t^R\} \in \mathbb{R}^{(l+1)\times 63}$.

We use a template skeleton of 23 joints, out of which 15 joints belong to the upper-body including the root joint (pelvis), and 8 joints belong to the lower-body. The full-body output $\mathbf{y}_{full} =$

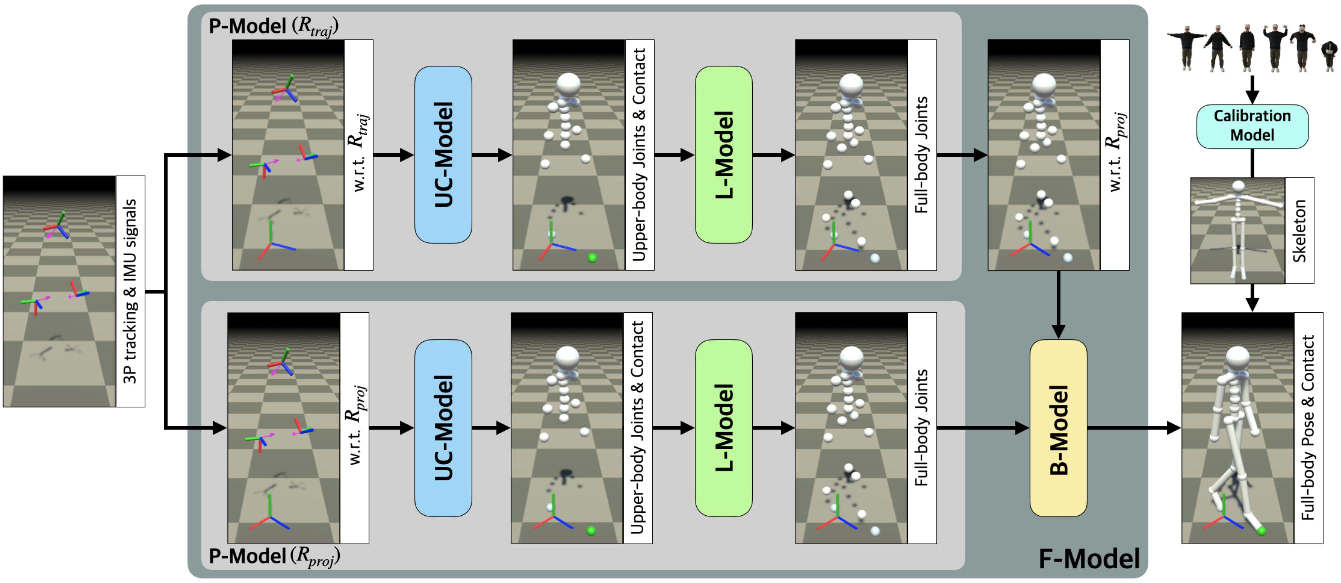

Figure 2: *Model overview. Prior to tracking, a learned calibration model estimates the bone offsets from calibration poses once per user. Given three-point six DOF poses and IMU measurements, a full-body tracking network predicts the per-frame motion. This network consists of two streams of identical design to transform inputs into different coordinates. The UC-model predicts the upper-body and foot contacts, which are used to condition the L-model to generate the lower-body. Finally, a blending network fuses the predictions, combined with calibrated offsets to output the final full-body motion.*

$\{\mathbf{y}_{upper}, \mathbf{y}_{lower}\} \in \mathbb{R}^{207}$ is composed of the transformation $\{\mathbf{p}, \mathbf{q}\}$ per joint, with $\mathbf{y}_{upper} \in \mathbb{R}^{15\times 9}$ and $\mathbf{y}_{lower} \in \mathbb{R}^{8\times 9}$. In addition, we output the contact probabilities of both feet $\mathbf{y}_{contact} \in \mathbb{R}^2$. We denote a trajectory of output with $\mathbf{Y}_{*,\,t} = \{\mathbf{y}_{*,\,t-l}, \ldots, \mathbf{y}_{*,\,t}\}$, and the corresponding ground truth with $\hat{\mathbf{Y}}_{*,\,t}$. We may omit the time index for clarity when the context is clear.

3.2. Reference Coordinate Frame

We use two different coordinate frames, R_{proj} and R_{traj}, to represent network input and output. $\mathbf{x}^{HP}$ is the world coordinates of the head joint $\mathbf{x}^H$ projected to the ground plane. It is similar to the commonly used ground-projected root coordinate [SZKS19]. However, because we don't have access to the ground truth root joint at inference time, we approximate it with the head transformation. The head projection frame R_{proj} equals $\mathbf{x}^{HP}$ at time $t-\frac{l}{2}$. Head trajectory frame R_{traj} has the same position as R_{proj} but differs in its forward direction. Specifically, as illustrated in Figure 3, the forward direction of R_{traj} is the normalized vector pointing from the position of $\mathbf{x}^{HP}_{t-l}$ to $\mathbf{x}^{HP}_t$.

3.3. Skeleton Calibration

To calibrate a user's body proportions, we trained a simple 2-layer MLP model to predict the joint positions in our template skeleton model from six calibration poses (Figure 4). The input to this model concatenates the position and orientation of the three trackers for all six poses as $\mathbf{x}_{cal} \in \mathbb{R}^{6\times 3\times 9}$. The output consists of 3D offsets of 22 joints (excluding the pelvis), $\mathbf{y}_{cal} \in \mathbb{R}^{22\times 3}$. Each input pose is represented in its respective $\mathbf{x}^{HP}$. The model is trained to minimize the L2 distance between the predicted and the ground truth offsets, $\mathcal{L}_{cal} = \|\mathbf{y}_{cal} - \hat{\mathbf{y}}_{cal}\|_2$. At run time, a user performs the six poses once and the calibration model estimates the corresponding skeleton dimensions for the final pose reconstruction.

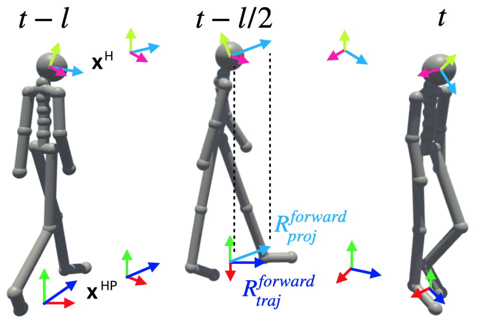

Figure 3: *Reference frame definitions. Given a window of head poses $\mathbf{x}_H$ from $[t-l,\ t]$, we project the head pose to the ground to obtain $\mathbf{x}_{HP}$ per time-step. $\mathbf{x}_{HP}$ at the middle frame $t-\frac{l}{2}$ is used as the reference, R_{proj}, whereas R_{traj} is rotated about the vertical axis to align with their trajectory heading.*

3.4. Upper-body and Contact Prediction

The Upper-body and Contact Prediction Model (UC-Model) takes $\{\mathbf{X}^{EE},\ \mathbf{X}^{HP}\} \in \mathbb{R}^{(l+1)\times(63+18)}$ as input, and outputs $\{\mathbf{Y}_{upper},\ \mathbf{Y}_{contact}\} \in \mathbb{R}^{(l+1)\times(135+2)}$. $\mathbf{X}^{HP} = \{\mathbf{x}^{HP}_{t-l}, \ldots, \mathbf{x}^{HP}_t\}$ is a trajectory where $\mathbf{x}^{HP}$ contains the position, orientation, and ve-

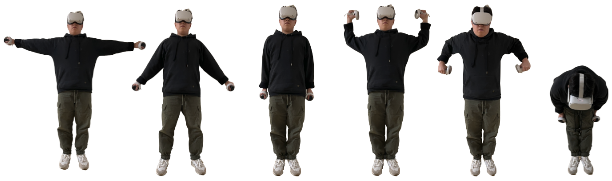

Figure 4: *Poses for skeleton calibration. From the left, T-pose, A-pose, I-pose, Elbows-bent-up, Elbows-bent-down, and 90° bow.*

locity of the head joint projected to the ground plane ($y = 0$). It is derived from $\mathbf{x}^H$, but we empirically found that providing this extra data made the network predictions less likely to violate ground constraints. The UC-Model consists of a linear embedding layer, a Transformer encoder [VSP*17], and two convolutional decoders to estimate an output trajectory of joint transforms and foot contact probabilities, respectively. It is trained to jointly minimize the L1 error of predicted upper-body joint transformations $\mathcal{L}_{upper} = \|\mathbf{Y}_{upper} - \hat{\mathbf{Y}}_{upper}\|_1$, the L2 error of joint velocities $\mathcal{L}_{\Delta upper} = \|\mathbf{Y}_{\Delta upper} - \hat{\mathbf{Y}}_{\Delta upper}\|_2$, and the Binary Cross Entropy between predicted and ground truth foot contact probabilities $\mathcal{L}_{contact} = BCE(\mathbf{Y}_{contact},\ \hat{\mathbf{Y}}_{contact})$.

3.5. Lower-body Generation

The Lower-body Generation Model (L-Model) is a CVAE [SLY15], using Transformers in both the encoder and the decoder as in TransformerVAE [PBV21]. The encoder takes two streams of information: the lower-body joints $\hat{\mathbf{y}}_{lower,\ t} \in \mathbb{R}^{72}$ to be constructed, and a conditional vector $\mathbf{C} = \{\mathbf{Y}_{upper,\ t},\ \mathbf{Y}_{lower,\ t-1},\ \mathbf{Y}_{contact,\ t}\} \in \mathbb{R}^{(l+1)\times 209}$. They are encoded into parameters of a Gaussian distribution $\mathcal{N}(\mu,\ \sigma)$, from which we can sample a latent vector $\mathbf{z} \in \mathbb{R}^{256}$. The decoder then reconstructs $\mathbf{y}_{lower,\ t}$ from $\mathbf{z}$ and the conditional vector $\mathbf{C}$. The model is trained to jointly minimize the Kullback-Leibler divergence between the latent distribution and multivariate Gaussian distribution $\mathcal{L}_{KL} = D_{KL}(\mathcal{N}(\mu,\ \sigma)\|\mathcal{N}(\mathbf{0},\ \mathbf{I}))$, the L1 reconstruction loss of the lower-body pose $\mathcal{L}_{lower} = \|\mathbf{Y}_{lower} - \hat{\mathbf{Y}}_{lower}\|_1$, and the L2 error of joint velocities $\mathcal{L}_{\Delta lower} = \|\mathbf{Y}_{\Delta lower} - \hat{\mathbf{Y}}_{\Delta lower}\|_2$.

At training time, we apply an aggressive dropout rate of 90% to $\mathbf{Y}_{lower,\ t-1}$ to avoid overfitting. In addition, we randomly flip 10% of the ground truth foot contact labels to help the model be more robust to prediction errors from the UC-Model. At inference time, we supply the decoder only with a $\mathbf{z}$ sampled from $\mathcal{N}(\mathbf{0},\ \mathbf{I})$, and the conditional vector $\mathbf{C}$ from predictions.

3.6. Full-body Pose

The UC-Model and the L-Model together constitute the full-body Pose Model (P-Model), whose input is tracker state $\mathbf{X}_t^{EE}$ and output is a trajectory of full-body poses $\mathbf{Y}_{full,\ t}$ and foot contact probabilities $\mathbf{Y}_{contact,\ t}$. The output is represented in the same reference coordinates as the input. Instead of using the world coordinate as a reference, the P-Model can transform its input to an input reference frame, such as R_{proj} or R_{traj}.

3.7. Motion Blending

Our Full Model (F-Model) runs two P-Models in parallel, one using R_{traj} as the reference, and the other using R_{proj}. Depending on the input signals, one of these two outputs could outperform the other. For example, we observed that R_{proj} excels for static movements with stable reference orientation, and R_{traj} produces more accurate predictions when the whole body moves independently of the head orientation (e.g., side or backward stepping). To take advantage of the better between the two, we use a Motion Blending Model (B-Model) to blend them based on the input signals.

The B-Model takes three streams of input, all transformed to the R_{proj} reference frame so the final output pose is also represented in R_{proj}. In addition to the two outputs from P-Models, it also takes the same input signals as the UC-Model $\mathbf{X}_{UC} = \{\mathbf{X}^{EE},\ \mathbf{X}^{HP}\} \in \mathbb{R}^{(l+1)\times(63+18)}$ as guidance for the motion blending process. These three sets of input are first transformed by their respective embedding layers to be the same dimension of $(l+1)\times 256$. They are then concatenated along the time axis to form the input tensor for the Transformer encoder. Lastly, two convolutional decoders, similar to those of the UC-Model, transform the encoded feature into the final trajectory $\{\mathbf{Y}_{full},\ \mathbf{Y}_{contact}\}$. The B-Model is trained to minimize the L1 distance between predicted and ground truth full-body pose $\mathcal{L}_{full} = \|\mathbf{Y}_{full} - \hat{\mathbf{Y}}_{full}\|_1$, the L2 error of joint velocities $\mathcal{L}_{\Delta full} = \|\mathbf{Y}_{\Delta full} - \hat{\mathbf{Y}}_{\Delta full}\|_2$, and Binary Cross Entropy between blended and ground truth contact probabilities $\mathcal{L}_{contact,\ B} = BCE(\mathbf{Y}_{contact},\ \hat{\mathbf{Y}}_{contact})$. In addition, we further encourage consistency between final foot contact probabilities and predicted foot locations with a contact consistency loss $\mathcal{L}_{consist} = \|(FK(\mathbf{y}^i_{full}) - FK(\mathbf{y}^{i-1}_{full}))\cdot \mathbf{y}^i_{contact}\|_2$. as in EDGE [TCL23].

The result of all processing described thus far is a trajectory of transformations for all 23 joints of the skeleton. To produce a final pose, we first compute the pelvis position by running FK from the input head transformation along predicted spinal joint transformations, using calibrated joint offsets. Then, the final full-body pose is reconstructed from this computed pelvis position, output joint rotations from $\mathbf{y}_{full,\ t}$, and the calibrated skeleton.

3.8. Training

We train all components of the F-Model end-to-end from scratch with a combined loss function as follows:

$$\begin{aligned}\mathcal{L} = {} & \lambda_{pose}(\mathcal{L}_{upper,\ R_{proj}} + \mathcal{L}_{lower,\ R_{proj}} \\ & \quad + \mathcal{L}_{upper,\ R_{traj}} + \mathcal{L}_{lower,\ R_{traj}} + \mathcal{L}_{full}) \\ & + \lambda_{vel}(\mathcal{L}_{\Delta upper,\ R_{proj}} + \mathcal{L}_{\Delta lower,\ R_{proj}} \\ & \quad + \mathcal{L}_{\Delta upper,\ R_{traj}} + \mathcal{L}_{\Delta lower,\ R_{traj}} + \mathcal{L}_{\Delta full}) \\ & + \lambda_{contact}(\mathcal{L}_{contact,\ R_{proj}} + \mathcal{L}_{contact,\ R_{traj}} + \mathcal{L}_{contact,\ B}) \\ & + \lambda_{consist}(\mathcal{L}_{consist}) \\ & + \lambda_{KL}(\mathcal{L}_{KL,\ R_{proj}} + \mathcal{L}_{KL,\ R_{traj}}),\end{aligned} \tag{1}$$

where λ_{pose}, λ_{vel}, $\lambda_{contact}$, $\lambda_{consist}$, and λ_{KL} are 1, 0.1, 1×10^{-3}, 1×10^{-5}, and 1×10^{-3}, respectively. We minimized the L1 norm in pose losses to avoid averaging similar outputs in a diverse dataset. For velocity and contact consistency loss, we minimized

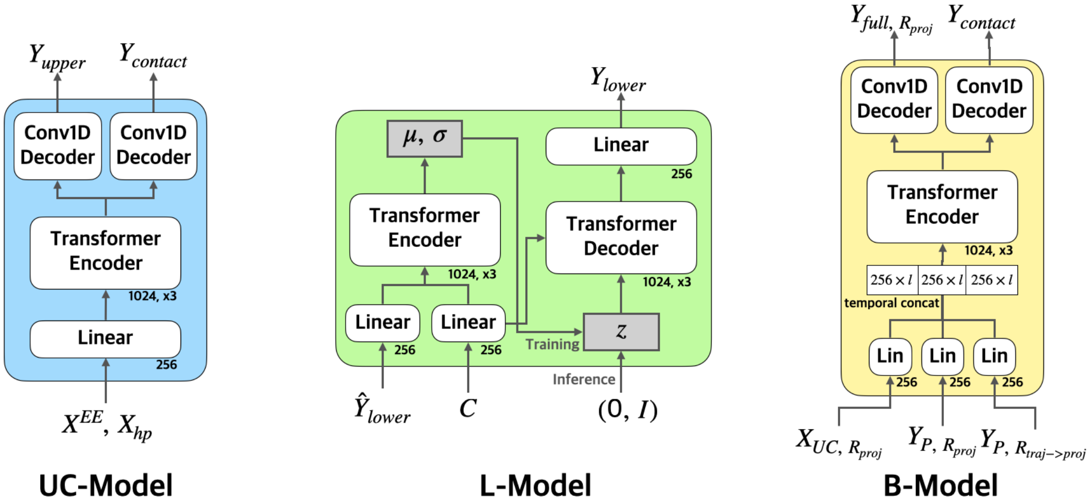

Figure 5: *Network architectures. The UC-model is a transformer encoder followed by parallel convolutional decoders that predict upper-body joint rotations and foot contacts. The L-Model is a CVAE, where a transformer encoder learns the lower body latent space conditioned on the upper body prediction and lower body history. The B-model also uses a transformer encoder on learned linear embeddings of the UC-model input and two stream P-model outputs, $Y_P = \{Y_{upper}, Y_{lower}, Y_{contact}\}$, with convolutional decoders to produce the final full-body and foot contact output.*

L2 errors for better overall accuracy. We find the velocity losses effective in reducing jitter in the motion, and the contact consistency loss helps to reduce foot sliding. We used Adam optimizer with a learning rate of 1×10^{-4} and batch size of 256. Training is run on one Nvidia RTX 4090 GPU for about 16.5 hours with 60,000 iterations.

3.9. Dataset

We captured our dataset using the Xsens Awinda system [RLS08] of 17 IMU sensors attached to the subjects' bodies. The captured IMU data is processed by the Xsens Analyze Pro software to produce ground truth body motions and foot contact labels. We generate *synthetic* tracker signals by computing the transformations on the head joint and the wrist joints of both hands from the body motion. This practice is consistent with existing work [JSQ*22; DKP*23] without access to data from commercial VR devices, as the transformations from three-point trackers are often clean and noise-free [WWY22].

To improve data consistency between training and testing, we deliberately put IMU sensors at the three tracker locations at capture time, so the acceleration signals from these sensors are similar to those in the live demo. It is of critical importance that we use linear accelerations from the sensors for training because they capture high-frequency signals that correspond to impacts from contact events. In fact, the Xsens software relies on sensor accelerations to annotate ground truth contact labels. Such information is lost in the synthetic acceleration derived from finite differences of poses.

Our dataset consists of 22 subjects (height$_{cm}$ $h \sim \mathcal{N}(173.77,\ 8.98^2)$, $159 \leq h \leq 192$) performing around 35 diverse actions, including in-place body exercises, locomotion at different speeds and styles, and interaction with objects and furniture. Table 1 presents details on motion categories in the dataset. The data was captured in two environments: an empty mocap stage and a furnished apartment. The mocap stage enables a broad range of dynamic motions based on specific protocols, while the apartment focuses on interactions in a natural indoor environment. In total, the dataset contains 772 motion clips and

Range of Motion
T-pose, A-pose, Idle, Elbows bent up & down, Bow Stretch arms, Look, Roll head, Touch toes, Hands on waist Twist torso, Hula hoop, Lean upper-body Lunge, Squat, Jumping Jack, Kick, Lift knee Turn in place, Walk in place, Run in place Conversational gestures
Locomotion
Normal walk, Walk with free upper-body motions Normal Jog, Jog with free upper-body motions Normal Run, Run with free upper-body motions Normal Crouch, Crouch with free upper-body motions Transitions with changing pace Moving backward with changing pace Jump, Running Jump
Object Interaction
Sit on a chair / couch / sofa / stool / bed Free upper-body motions while sitting Lie down on a sofa / bed, Work at an office desk Moving boxes, Open the door/windows Turn on/off lights, Watch TV, Wash dishes, Cook

Table 1: *Motion categories in DivaTrack dataset. In addition to standard locomotion patterns with arm swinging, our dataset includes diverse upper-body actions (i.e. drinking water, airplane arms, and clapping) during locomotion. These varied upper-body motions pose challenges to the three-point tracking problem.*

16.5 hours of data. We hold out four subjects for testing and use 18 subjects for training and validation. The test subjects' body shape (height$_{cm}$: [167, 171.5, 178, 192]) and motions are completely unseen to the trained models in Section 4.

4. Evaluation

We evaluate our method both qualitatively and quantitatively to support the following claims:

- Our method robustly synthesizes full-body motion from sparse three-point signals in real-time, for individuals with diverse body proportions.
- Our predict-and-generate approach supports a wider range of motion activities compared to state-of-the-art, including static motion, basic locomotion, and object interactions.
- Our reference system blending approach improves the estimation quality for the diversity of motions we support.
- Using foot contact probabilities as conditional input improves the quality of lower-body pose generation.
- Incorporating IMU linear accelerations in the input leads to more accurate foot contact predictions, which in turn improves the quality of synthesized motions.

We employ a collection of metrics in the quantitative evaluation. Accuracy is measured by Mean Position Error (M*PE) and Mean Rotation Error (M*RE), reported separately for the Pelvis (P) versus other body Joints (J) to highlight the strong influence of the pelvis. We additionally break down the metrics for joints in the upper-body (UJ) versus the lower-body (LJ). We compute contact labels by thresholding the predicted probability at 0.5 and compare them against ground truth.

4.1. Comparison with State-of-the-art Methods

We compare our method to state-of-the-art with a similar setup on both our own test set and public datasets with IMU sensors. We retrain baseline models on our training data for evaluation on our test set and use public pretrained models for evaluation on public datasets.

On our own test set, we compare with recent methods that use sparse trackers on the upper body to generate full-body pose in real-time: *LoBSTr* [YKL21] and *AvatarPoser* [JSQ*22]. We also increase the *AvatarPoser* model by 5x to be comparable with *DivaTrack* model size as an additional baseline. Moreover, we augment both baseline models naively with linear accelerations from IMUs in the input with minimum change to their architectures, so they have access to the same information as *DivaTrack*. All models are trained on our 18-subject training set, DT Train, and tested on our held-out dataset with 4 subjects, DT Test. All data used for training and test have preserved the skeleton offsets and joint rotations for each subject, without retargeting to a single standard proportion. As *LoBSTr* requires an additional pelvis tracker, we provide it with ground truth pelvis transformation. We also provide the ground truth skeletons to both *LoBSTr* and *AvatarPoser* in all tests, and report results using ground truth and predicted skeletons respectively for *DivaTrack*.

Quantitative results in Table 2 show that *DivaTrack* with ground truth skeletons outperforms all baselines in all metrics. Our method notably improves over the runner-up entries in every metric, reducing MPPE by 2.96cm (30%) and MJPE by 1.11cm (13%), and reducing MPRE by 2.35° (19%) and MJRE by 1.36° (13%). In comparison, the bigger *AvatarPoser*$\times 5^{\dagger}$ model underperforms the original model, suggesting their architecture is not bottlenecked by network capacity. Augmenting existing models with IMU signals naively does not improve performance in a significant and conclusive way, hinting at the importance of our architectural decisions.

*DivaTrack** uses predicted skeletons and therefore its position accuracy degrades due to errors in bone offset estimation (Table 3). Our calibration model has an average error of $\mu = 1.55_{cm}$, $\sigma^2 = 1.18$ over all joints. Errors are smaller along the body trunk and larger in the limbs, where the knee has the largest error of $4.97cm$ since it is the most ambiguous based on the sensor signals. Despite these errors, *DivaTrack** still has better accuracy in the lower-body joint positions (MLJPE) than *AvatarPoser*† with ground truth skeletons.

We additionally evaluate our model in MPJPE on two public datasets with IMU sensors accelerations and ground truth body motion, TotalCapture [TGM*17] (five subjects) and HPS [GMSP21] (seven subjects), as shown in Table 4. We use ground truth skeletons in our method because we cannot run our calibration model on these datasets.

On TotalCapture, we compare *DivaTrack* against published pretrained models of *AvatarPoser* and *AGRoL*. The dataset consists of walking, acting, and challenging freestyle activities, and we evaluate for the subset contained in AMASS [MGT*19]. Even though *AGRoL* demonstrates superior performance on average accuracy, its motions show severe jitter. Compared to *AvatarPoser*, our model reduces MPJPE by 27%, consistent with results on our own test set.

On HPS, we compare *DivaTrack* against *AvatarPoser*, *AGRoL*, and *EgoPoser* [JSM*23] with numbers reported in EgoPoser. The HPS dataset contains navigation and interaction movements in large-scale environments (ranging from $300 \sim 1000m^2$, up to $2500m^2$). Both *AvatarPoser* and *AGRoL* show significant performance decreases, because they use global representations of AMASS dataset for training, and cannot generalize well to large global offsets in outdoor environments. In contrast, *DivaTrack* demonstrated robust and superior performance in large-scale scenes thanks to our reference frame representation.

Figure 6 are visual comparisons with *AvatarPoser* on a few challenging examples for three-point input. For a variety of motion categories in both DT Test and TotalCapture, *DivaTrack* follows the three-point signals closely with plausible leg movements and more clear foot contacts, while *AvatarPoser* results are less accurate with more severe foot sliding artifacts. For more visual results, please refer to the supplementary video.

4.2. Ablation of Reference Frame Definitions

We compare the output of the P-Models using two respective reference frames, R_{proj} and R_{traj}, and with the blended output from B-Model in Table 5 on DT Test. The blended result outperforms single reference results in position errors and contact accuracy. R_{proj} performs better than R_{traj} quantitatively, as the majority of test data

Model	MPPE$_{cm}$	MPRE$^{\circ}$	MJPE$_{cm}$	MJRE$^{\circ}$	Cont. Acc.$_{\%}$ ↑	MUJPE$_{cm}$	MUJRE$^{\circ}$	MLJPE$_{cm}$	MLJRE$^{\circ}$
LoBSTr†	–	–	9.05	12.09	72.90	8.08	12.05	15.69	12.16
LoBSTr_IMU†	–	–	9.47	12.97	74.24	7.92	12.62	17.56	13.99
AvatarPoser†	9.69	12.31	8.73	10.70	–	5.91	9.61	13.67	12.60
AvatarPoser_IMU†	10.40	12.77	8.46	10.62	–	5.42	9.08	13.10	11.90
AvatarPoser×5†	12.25	15.97	9.65	12.04	–	6.82	11.37	14.58	13.23
DivaTrack	**6.73**	**9.96**	**7.35**	**9.26**	**85.25**	**4.78**	**8.13**	**11.86**	**11.23**
DivaTrack*	9.99	”	9.10	”	”	6.87	”	13.00	”

Table 2: *Comparison of retrained SOTA models and DivaTrack tested on DT Test dataset. The † denotes retrained models on DT Train dataset and _IMU denotes that IMU accelerations are provided as additional input. DivaTrack* denotes results using the skeleton predicted from our calibration model, which has a mean bone-length error of* 1.7_{cm}. *The* **best** *and* runner-up *entries are annotated respectively.*

Chest	Chest2	Chest3	Chest4	Neck	Head	Collar	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Toe
0.55	1.11	1.01	0.90	1.54	1.07	1.00	1.22	1.94	1.1	0.45	4.97	2.95	2.02

Table 3: *Average bone length errors (cm) of each joint from our skeleton calibration model.*

	TotalCapture	HPS					
		BIB_EG_Tour	MPI_EG	Working_Standing	UG_Computers	Go_Around	UG_Long
AvatarPoser	11.96	22.53	16.54	19.08	23.24	19.5	16.65
AGRoL	**6.89**	28.95	19.41	17.67	20.90	14.16	12.81
EgoPoser	–	9.55	11.05	8.70	10.34	6.9	8.95
DivaTrack	8.73	**6.51**	**9.06**	**7.36**	**8.84**	**6.82**	**7.34**

Table 4: *Comparison of Mean Joint Position Errors (MJPE$_{cm}$) of released SOTA models and pretrained DivaTrack tested on TotalCapture and HPS, two public datasets with ground truth IMU sensor measurements.*

has a strong correlation between head orientation and the movement direction. As shown in Figure 7 and in the supplementary video, R_{traj} excels in cases when head movement doesn't correlate with body movement, such as occasionally turning to look back during backward walking. Our two-stream blending result outperforms each individual stream (P-Model), suggesting that it is able to adaptively combine the predictions to achieve the best result.

Reference Def.	MPPE$_{cm}$	MJPE$_{cm}$	Cont. Acc.$_{\%}$ ↑
R_{proj}	7.19	7.83	81.44
R_{traj}	7.51	9.05	82.34
$R_{proj,\,traj}$	**6.73**	**7.35**	**85.25**

Table 5: *Comparison of per-frame position errors for models trained with different reference definitions.* $R_{proj,\,traj}$ *denotes our two-stream blending approach of both reference frames.*

4.3. Ablation for Contacts as L-model Condition.

We validate the utility of contact labels as a condition for the lower-body pose generation in ablation on DT Test. We trained two versions of the L-Model, one with contacts as L-Model condition and one without. We also compare using the predicted contact labels from the UC-model versus using the ground truth contact labels. The results in Table 6 clearly demonstrate that incorporating predicted contact labels as a condition improves the quality of generated lower-body poses and the quality has a clear correlation with the accuracy of predicted contact.

The improvement from contact prediction can be easily observed in the output motions as well. Figure 8 shows an example of kicking motion on the left, where foot contacts make it possible to distinguish standing from kicking with near identical upper-body postures. On the right is another example where contact information reduces foot sliding artifacts when walking speed changes.

	MLJPE$_{cm}$	MLJRE$^{\circ}$
w/o contacts	13.27	11.48
pred. contacts	11.86	11.23
GT contacts	**11.23**	**10.63**

Table 6: *Ablation on contact labels as L-Model condition (Mean Lower-body Joint Errors).*

4.4. Ablation for IMU Acceleration Signals.

Given the importance of foot contact prediction, we incorporate linear acceleration signals from IMU sensors as an input to the UC-Model, because they can capture instantaneous contact events at high frequency, especially from the head. Compared with real IMU accelerations captured at up to 1000Hz, synthetic acceleration generated using finite differences is much lower frequency and less useful. We conducted an ablation to compare training the UC-

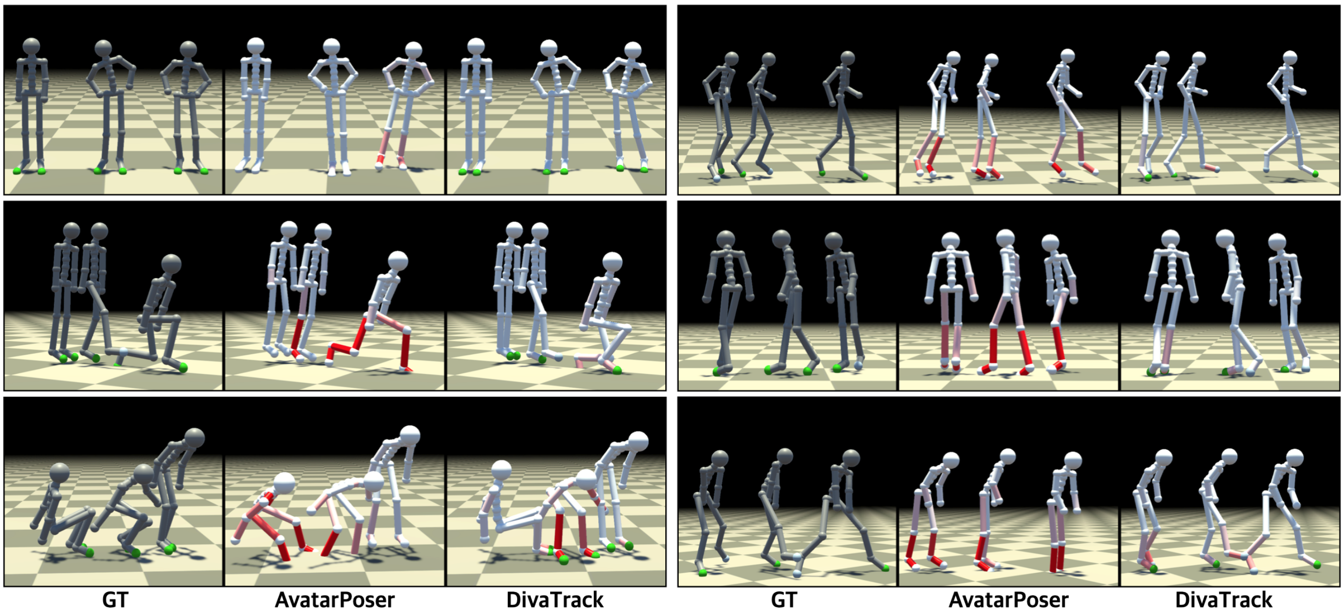

Figure 6: *Qualitative comparison with AvatarPoser. The darker red color denotes a larger error. Our Model, DivaTrack, produces outputs closer to the ground truth for both static movements (left: hula-hoop, lunge, and sitting down) and various locomotion (right: running, moving backward, and moving a box) of subjects with different body shapes.*

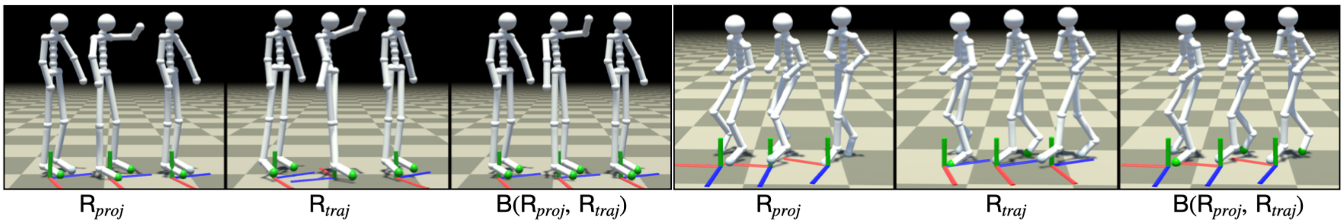

Figure 7: *Comparison of reference frame definitions. Left: for static movements with slight head jiggling, R_{traj} can rotate fast, resulting in jitter and sliding. Right: for locomotion with uncorrelated head direction, R_{proj} are directed irrelevant to the translation, producing wrong poses. Our blending method can generate high-quality motions by taking advantage of two reference frames.*

Model without acceleration, with synthetic acceleration, and with IMU acceleration. We use two evaluation datasets: the full DT Test dataset, and a subset featuring challenging actions including Lunge, Kick, Lift knee, Turn/Walk/Run in place, and difficult Locomotion. Table 7 clearly shows that IMU accelerations are essential for contact prediction and thus increase the performance of lower-body generation, while synthetic accelerations are not helpful as they do not provide new information.

While the improvements from IMU accelerations are more moderate in the full DT Test dataset because the majority of typical walking motions can already be well predicted without them, results from the challenging subset highlight where they shine. Figure 9 showcases two examples where IMU accelerations provide a meaningful boost to pose accuracy. In the case of the lunge motion, only the model trained with IMU accelerations accurately captured the correct contact labels of the right foot leaving and landing on the ground, resulting in a realistic lunge of the lower body. Similarly, for the object-carrying motion, where the upper-body joints remain relatively stationary when compared to a standard walking motion, the utilization of IMU accelerations generated a more detailed walking motion of the lower body, while the other models exhibited static or incorrect poses. We found these classes of improvements make a significant difference in the overall visual quality of the method.

	^{UC}Cont. Acc.$_{\%}$ ↑	MLJPE$_{cm}$	MLJRE°
w/o acc	78.47 / 73.21	12.03 / 12.30	11.53 / 12.50
syn acc	78.90 / 73.95	12.17 / 12.41	11.60 / 12.60
IMU acc	**82.34 / 77.06**	**11.86 / 11.79**	**11.23 / 11.90**

Table 7: *Ablation on IMU linear accelerations as UC-Model input. The L-Model uses predicted contact labels from the UC-model as a condition for lower-body pose generation. The values on the left correspond to measurements taken across the entire DT Test set. The values on the right pertain specifically to a subset containing more challenging actions for the three-point tracking problem.*

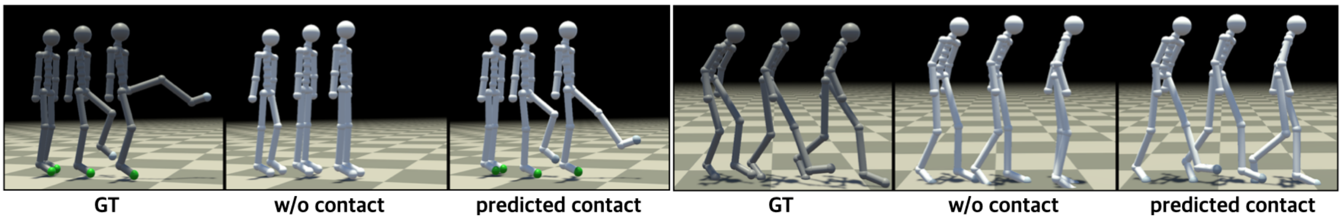

Figure 8: *Comparison of contact labels as L-model condition. Foot contacts enable the L-model to generate motions that are underdetermined by three-point signals (i.e. kicking) and to maintain accurate footsteps while changing speed and direction.*

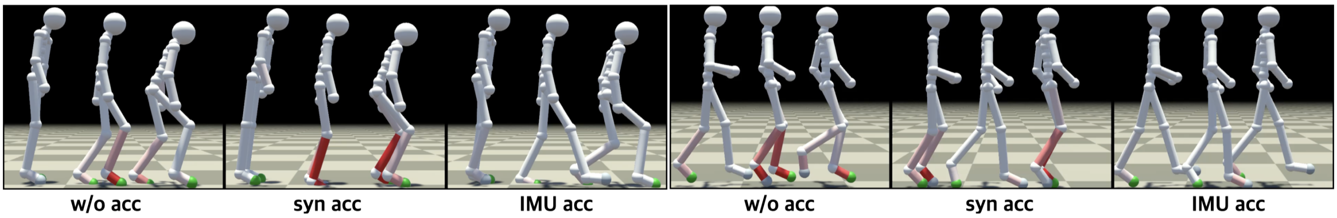

Figure 9: *Comparison of acceleration sources. IMU accelerations help UC-Model to predict accurate contacts for corner cases (left: lunge and right: carrying object), thereby allowing L-Model to generate high-quality lower-body motions with minimized sliding artifacts.*

4.5. Real-time Tracking from VR devices and XSens IMUs

We test *DivaTrack* in real-time using a Meta VR device and three XSens IMU sensors (Figure 1). One IMU is attached to the VR headset, and two IMUs are on the subject's wrists. After IMU calibration, we manually set the VR world to align with the IMU world. For real-time tracking, we fetch three-point positions and rotations from the VR device (Steam VR) and linear accelerations from IMU sensors (Xsens MVN). This data is streamed to Unity3D, where frame inputs are recorded and then sent to a Python inference module. *DivaTrack* takes 7.5ms per frame on a desktop with an RTX2080Ti GPU on average, which runs comfortably with 60Hz VR data, although there is a noticeable system latency in our PCVR setup. We also run an online post-processing to smooth the output and stabilize foot contacts using IK. We tested subjects of various heights (in cm: 159, 162, 170, 172, 175, and 178) who are not in the training set, and instructed them to perform free-form locomotion. After the calibration poses, test subjects walked, jogged, and hopped back and forth, sat down on a stool or on the floor, and carried a stool to different locations (Figure 1). The output upper-body motions follow the three-point input signals well throughout all sessions, and the lower-body movements look plausible and coordinated with the upper-body, even though they are not always accurate. We invite readers to watch the supplementary video for qualitative evaluation.

5. Discussion

We demonstrated the pivotal role of reference coordinates in the final result. This is a unique challenge for three-point tracking problems due to the lack of root information. Our two-stream reference blending approach is effective but not elegant. Further, we still cannot effectively handle motions where the head's forward direction is close to the direction of gravity, such as lying down or crawling on the floor. A sub-optimal reference frame can lead to poor pelvis predictions that consequently cause severe body jitter and foot sliding artifacts. An unstable reference frame also makes it difficult to produce a static stance when the upper-body is moving. A deeper investigation of motion representation is needed for the three-point tracking problem.

We also demonstrated the importance of realistic IMU signals as opposed to synthetic accelerations for accurate contact prediction. TIP [JYG*22] also observed similar discrepancies in these two types of signals. On the other hand, the three-point data can be realistically simulated from motion capture data, as shown in QuestSim [WWY22]. As a result, we collected our dataset with XSens similar to TotalCapture but did not evaluate our method on public datasets with no IMU signals. If VR devices can expose their IMU data, a realistic dataset of three-point data and IMU signals will be tremendously helpful to future research.

Our method encounters challenges with unseen motion categories, notably highlighted in our supplementary video. In addition, as our training data is confined to the flat ground (height=0) that defines reference coordinates, our model struggles with motions involving ascent and descent.

While our synthetic tracker data aligns well with real MR device tracking signals, it fails to capture tracking failures in practical VR/MR usage, and our method overlooks this issue. Addressing the challenge of "tracking loss" in practical scenarios will be an intriguing avenue for future research.

Our framework currently predicts the user skeleton and full-body pose separately, before combining them at the last stage. A more effective design would be to integrate them earlier so the pose generation model can utilize the skeleton proportion in pose prediction. Unfortunately, our preliminary test did not yield any significant improvements.

6. Conclusion

We presented DivaTrack, a deep learning framework that infers a user's motion in real-time using six DOF poses and linear accelerations from three trackers on the head and wrists. Considering the different level of information from the sensors, DivaTrack first infers upper-body poses and contact states, and use them as control signals to generate suitable lower-body poses. Addressing the lack of known pelvis joint information, our model generates poses with respect to two different head-based reference frames and blends them to create a final pose. From a simple set of calibration poses, DivaTrack infers a user-specific skeleton model for final visualization. Our results suggest that signals from the head and hands alone can provide rich information about body proportions and motion. Our insights about the importance of IMU acceleration signals and their correlation to foot contacts are validated by ablation studies. We further provide a dedicated dataset to showcase the challenges of using only three-point inputs, and we believe it can inform future research to tackle fundamental issues in this problem.

7. Acknowledgement

This research is conducted during Dongseok Yang's internship at Meta and is fully funded by the company. Dongseok Yang is grateful to the Meta team for their mentorship and especially thanks to Michelle Hill who supported data capture.

References

[ACB*22] ALIAKBARIAN, SADEGH, CAMERON, PASHMINA, BOGO, FEDERICA, et al. "FLAG: Flow-based 3D Avatar Generation from Sparse Observations". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022 2.

[AOG*21] AHUJA, KARAN, OFEK, EYAL, GONZALEZ-FRANCO, MAR, et al. "Coolmoves: User motion accentuation in virtual reality". *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5.2 (2021), 1–23 3.

[ASF*22] AHUJA, KARAN, SHEN, VIVIAN, FANG, CATHY MENGYING, et al. "ControllerPose: Inside-Out Body Capture with VR Controller Cameras". *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. CHI '22. New Orleans, LA, USA: Association for Computing Machinery, 2022. ISBN: 9781450391573 2.

[AWS*22] AKADA, HIROYASU, WANG, JIAN, SHIMADA, SOSHI, et al. "UnrealEgo: A New Dataset for Robust Egocentric 3D Human Motion Capture". *European Conference on Computer Vision (ECCV)*. 2022 2.

[CGCB14] CHUNG, JUNYOUNG, GULCEHRE, CAGLAR, CHO, KYUNGHYUN, and BENGIO, YOSHUA. "Empirical evaluation of gated recurrent neural networks on sequence modeling". *arXiv preprint arXiv:1412.3555* (2014) 3.

[CH05] CHAI, JINXIANG and HODGINS, JESSICA K. "Performance animation from low-dimensional control signals". *ACM SIGGRAPH 2005 Papers*. 2005, 686–696 3.

[CNGG18] CHEW, DI-KIAT, NGOH, KIERON JIE-HAN, GOUWANDA, DARWIN, and GOPALAI, ALPHA A. "Estimating running spatial and temporal parameters using an inertial sensor". *Sports Engineering* 21 (2018), 115–122 2.

[CYZ*23] CAI, ZHONGANG, YIN, WANQI, ZENG, AILING, et al. *SMPLer-X: Scaling Up Expressive Human Pose and Shape Estimation*. 2023. arXiv: 2309.17448 [cs.CV] 2.

[DAM*21] DAY, EVAN M, ALCANTARA, RYAN S, MCGEEHAN, MICHAEL A, et al. "Low-pass filter cutoff frequency affects sacral-mounted inertial measurement unit estimations of peak vertical ground reaction force and contact time during treadmill running". *Journal of Biomechanics* 119 (2021), 110323 2.

[DDC*21] DITTADI, ANDREA, DZIADZIO, SEBASTIAN, COSKER, DARREN, et al. "Full-Body Motion from a Single Head-Mounted Device: Generating SMPL Poses from Partial Observations". *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, 11687–11697 3.

[DKP*23] DU, YUMING, KIPS, ROBIN, PUMAROLA, ALBERT, et al. "Avatars grow legs: Generating smooth human motion from sparse tracking inputs with diffusion model". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, 481–490 3, 6.

[DMGT23] DABRAL, RISHABH, MUGHAL, MUHAMMAD HAMZA, GOLYANIK, VLADISLAV, and THEOBALT, CHRISTIAN. "MoFusion: A Framework for Denoising-Diffusion-based Motion Synthesis". *Computer Vision and Pattern Recognition (CVPR)*. 2023 3.

[FFMH19] FEICHTENHOFER, CHRISTOPH, FAN, HAOQI, MALIK, JITENDRA, and HE, KAIMING. "SlowFast Networks for Video Recognition". *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. Oct. 2019 3.

[FNM19] FERSTL, YLVA, NEFF, MICHAEL, and MCDONNELL, RACHEL. "Multi-Objective Adversarial Gesture Generation". *Proceedings of the 12th ACM SIGGRAPH Conference on Motion, Interaction and Games*. MIG '19. Newcastle upon Tyne, United Kingdom: Association for Computing Machinery, 2019. ISBN: 9781450369947 3.

[GHR*16] GIANDOLINI, MARLENE, HORVAIS, NICOLAS, ROSSI, JÉRÉMY, et al. "Foot strike pattern differently affects the axial and transverse components of shock acceleration and attenuation in downhill trail running". *Journal of biomechanics* 49.9 (2016), 1765–1771 2.

[GMHP04] GROCHOW, KEITH, MARTIN, STEVEN L, HERTZMANN, AARON, and POPOVIĆ, ZORAN. "Style-based inverse kinematics". *ACM SIGGRAPH 2004 Papers*. 2004, 522–531 3.

[GMSP21] GUZOV, VLADIMIR, MIR, AYMEN, SATTLER, TORSTEN, and PONS-MOLL, GERARD. "Human poseitioning system (hps): 3d human pose estimation and self-localization in large scenes from body-mounted sensors". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, 4318–4329 2, 7.

[GPM*14] GOODFELLOW, IAN, POUGET-ABADIE, JEAN, MIRZA, MEHDI, et al. "Generative Adversarial Nets". *Advances in Neural Information Processing Systems*. Ed. by GHAHRAMANI, Z., WELLING, M., CORTES, C., et al. Vol. 27. Curran Associates, Inc., 2014 3.

[GPR*23] GOEL, SHUBHAM, PAVLAKOS, GEORGIOS, RAJASEGARAN, JATHUSHAN, et al. "Humans in 4D: Reconstructing and Tracking Humans with Transformers". *International Conference on Computer Vision (ICCV)*. 2023 2.

[HAB20] HENTER, GUSTAV EJE, ALEXANDERSON, SIMON, and BESKOW, JONAS. "Moglow: Probabilistic and controllable motion synthesis using normalising flows". *ACM Transactions on Graphics (TOG)* 39.6 (2020), 1–14 3.

[HKA*18] HUANG, YINGHAO, KAUFMANN, MANUEL, AKSAN, EMRE, et al. "Deep inertial poser: learning to reconstruct human pose from sparse inertial measurements in real time". *ACM Transactions on Graphics (TOG)* 37.6 (2018), 1–15 3.

[HS97] HOCHREITER, SEPP and SCHMIDHUBER, JÜRGEN. "Long short-term memory". *Neural computation* 9.8 (1997), 1735–1780 3.

[HSKJ15] HOLDEN, DANIEL, SAITO, JUN, KOMURA, TAKU, and JOYCE, THOMAS. "Learning motion manifolds with convolutional autoencoders". *SIGGRAPH Asia 2015 technical briefs*. 2015, 1–4 3.

[JSM*23] JIANG, JIAXI, STRELI, PAUL, MEIER, MANUEL, et al. "EgoPoser: Robust Real-Time Ego-Body Pose Estimation in Large Scenes". *arXiv preprint arXiv:2308.06493* (2023) 3, 7.

[JSQ*22] Jiang, Jiaxi, Streli, Paul, Qiu, Huajian, et al. "Avatarposer: Articulated full-body pose tracking from sparse motion sensing". *European Conference on Computer Vision*. Springer. 2022, 443–460 2, 3, 6, 7.

[JYG*22] Jiang, Yifeng, Ye, Yuting, Gopinath, Deepak, et al. "Transformer Inertial Poser: Real-time Human Motion Reconstruction from Sparse IMUs with Simultaneous Terrain Generation". *SIGGRAPH Asia 2022 Conference Papers*. 2022, 1–9 3, 10.

[KBM*23] Khirodkar, Rawal, Bansal, Aayush, Ma, Lingni, et al. "EgoHumans: An Egocentric 3D Multi-Human Benchmark". *International Conference on Computer Vision (ICCV)*. 2023 2.

[KW14] Kingma, Diederik P. and Welling, Max. "Auto-Encoding Variational Bayes". *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*. 2014. arXiv: http://arxiv.org/abs/1312.6114v10 [stat.ML] 3.

[KZFM19] Kanazawa, Angjoo, Zhang, Jason Y, Felsen, Panna, and Malik, Jitendra. "Learning 3d human dynamics from video". *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, 5614–5623 2.

[LAZ*22] Li, Peizhuo, Aberman, Kfir, Zhang, Zihan, et al. "GANimator: Neural Motion Synthesis from a Single Sequence". *ACM Trans. Graph.* 41.4 (July 2022). ISSN: 0730-0301 3.

[LHG*23] Luvizon, Diogo, Habermann, Marc, Golyanik, Vladislav, et al. "Scene-Aware 3D Multi-Human Motion Capture from a Single Camera". *Computer Graphics Forum* 42.2 (2023), 371–383. DOI: https://doi.org/10.1111/cgf.14768 2.

[LHYK21] Luo, Zhengyi, Hachiuma, Ryo, Yuan, Ye, and Kitani, Kris. "Dynamics-regulated kinematic policy for egocentric pose estimation". *Advances in Neural Information Processing Systems* 34 (2021), 25019–25032 2.

[LLD*22] Lim, Hyunchul, Li, Yaxuan, Dressa, Matthew, et al. "BodyTrak: Inferring Full-Body Poses from Body Silhouettes Using a Miniature Camera on a Wristband". *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6.3 (Sept. 2022) 2.

[LLW23] Li, Jiaman, Liu, Karen, and Wu, Jiajun. "Ego-Body Pose Estimation via Ego-Head Pose Estimation". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, 17142–17151 3.

[LMB10] Lee, James B, Mellifont, Rebecca B, and Burkett, Brendan J. "The use of a single inertial sensor to identify stride, step, and stance durations of running gait". *Journal of Science and Medicine in Sport* 13.2 (2010), 270–273 2.

[LSY*23] Lee, Sunmin, Starke, Sebastain, Ye, Yuting, et al. "QuestEnvSim: Environemnt-aware Simulated Motion Tracking From Sparse Sensors". *SIGGRAPH Conference*. 2023 3.

[LWH*12] Levine, Sergey, Wang, Jack M, Haraux, Alexis, et al. "Continuous character control with low-dimensional embeddings". *ACM Transactions on Graphics (TOG)* 31.4 (2012), 1–10 3.

[LYC*20] Li, Jiaman, Yin, Yihang, Chu, Hang, et al. "Learning to Generate Diverse Dance Motions with Transformer". (Aug. 2020) 3.

[LYL*19] Lee, Hsin-Ying, Yang, Xiaodong, Liu, Ming-Yu, et al. "Dancing to Music". *Advances in Neural Information Processing Systems*. Ed. by Wallach, H., Larochelle, H., Beygelzimer, A., et al. Vol. 32. Curran Associates, Inc., 2019. URL: https://proceedings.neurips.cc/paper/2019/file/7ca57a9f85a19a6e4b9a248c1daca185-Paper.pdf 3.

[LZCV20] Ling, Hung Yu, Zinno, Fabio, Cheng, George, and Van De Panne, Michiel. "Character controllers using motion vaes". *ACM Transactions on Graphics (TOG)* 39.4 (2020), 40–1 3.

[LZWM06] Liu, Guodong, Zhang, Jingdan, Wang, Wei, and McMillan, Leonard. "Human motion estimation from a reduced marker set". *Proceedings of the 2006 symposium on Interactive 3D graphics and games*. 2006, 35–42 3.

[MAG*23] Mollyn, Vimal, Arakawa, Riku, Goel, Mayank, et al. "IMUPoser: Full-Body Pose Estimation Using IMUs in Phones, Watches, and Earbuds". *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. CHI '23. Hamburg, Germany: Association for Computing Machinery, 2023. ISBN: 9781450394215. DOI: 10.1145/3544548.3581392. URL: https://doi.org/10.1145/3544548.3581392 2.

[MC12] Min, Jianyuan and Chai, Jinxiang. "Motion Graphs++: A Compact Generative Model for Semantic Motion Analysis and Synthesis". *ACM Trans. Graph.* 31.6 (Nov. 2012). ISSN: 0730-0301 3.

[MGT*19] Mahmood, Naureen, Ghorbani, Nima, Troje, Nikolaus F., et al. "AMASS: Archive of Motion Capture as Surface Shapes". *International Conference on Computer Vision*. Oct. 2019, 5442–5451 7.

[Moe98] Moe-Nilssen, Rolf. "A new method for evaluating motor control in gait under real-life environmental conditions. Part 1: The instrument". *Clinical biomechanics* 13.4-5 (1998), 320–327 2.

[MSK23] Milef, Nicholas, Sueda, Shinjiro, and Kalantari, Nima Khademi. "Variational Pose Prediction with Dynamic Sample Selection from Sparse Tracking Signals". *Computer Graphics Forum* (2023). ISSN: 1467-8659. DOI: 10.1111/cgf.14767 3.

[MYGY19] Ma, Li-Ke, Yang, Zeshi, Guo, Baining, and Yin, KangKang. "Towards Robust Direction Invariance in Character Animation". *Computer Graphics Forum* 38.7 (2019), 235–242 2.

[NSLW20] Nagaraj, Deepak, Schake, Erik, Leiner, Patrick, and Werth, Dirk. "An RNN-Ensemble Approach for Real Time Human Pose Estimation from Sparse IMUs". *Proceedings of the 3rd International Conference on Applications of Intelligent Systems*. APPIS 2020. Las Palmas de Gran Canaria, Spain: Association for Computing Machinery, 2020. ISBN: 9781450376303 3.

[PBV21] Petrovich, Mathis, Black, Michael J, and Varol, Gül. "Action-conditioned 3d human motion synthesis with transformer vae". *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, 10985–10995 3, 5.

[PLB*22] Patoz, Aurélien, Lussiana, Thibault, Breine, Bastiaan, et al. "A single sacral-mounted inertial measurement unit to estimate peak vertical ground reaction force, contact time, and flight time in running". *Sensors* 22.3 (2022), 784 2.

[PYAP22] Ponton, Jose Luis, Yun, Haoran, Andujar, Carlos, and Pelechano, Nuria. "Combining Motion Matching and Orientation Prediction to Animate Avatars for Consumer-Grade VR Devices". *Computer Graphics Forum* (2022). ISSN: 1467-8659 3.

[RLS08] Roetenberg, Daniel, Luinge, Henk, and Slycke, Per Johan. "Xsens MVN: Full 6DOF Human Motion Tracking Using Miniature Inertial Sensors". (2008) 2, 6.

[RSJ21] Rong, Yu, Shiratori, Takaaki, and Joo, Hanbyul. "Frankmocap: A monocular 3d whole-body pose estimation system via regression and integration". *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, 1749–1759 2.

[SGXT20] Shimada, Soshi, Golyanik, Vladislav, Xu, Weipeng, and Theobalt, Christian. "Physcap: Physically plausible monocular 3d motion capture in real time". *ACM Transactions on Graphics (ToG)* 39.6 (2020), 1–16 2, 3.

[SHP04] Safonova, Alla, Hodgins, Jessica K, and Pollard, Nancy S. "Synthesizing physically realistic human motion in low-dimensional, behavior-specific spaces". *ACM Transactions on Graphics (ToG)* 23.3 (2004), 514–521 3.

[SLY15] Sohn, Kihyuk, Lee, Honglak, and Yan, Xinchen. "Learning structured output representation using deep conditional generative models". *Advances in neural information processing systems* 28 (2015) 2, 3, 5.

[SMK22] Starke, Sebastian, Mason, Ian, and Komura, Taku. "DeepPhase: periodic autoencoders for learning motion phase manifolds". *ACM Transactions on Graphics (TOG)* 41.4 (2022), 1–13 3.

[SP97] SCHUSTER, MIKE and PALIWAL, KULDIP K. "Bidirectional recurrent neural networks". *IEEE Transactions on Signal Processing* 45 (Nov. 1997), 2673–2681 3.

[STKB23] SHAFIR, YONATAN, TEVET, GUY, KAPON, ROY, and BERMANO, AMIT H. "Human motion diffusion as a generative prior". *arXiv preprint arXiv:2303.01418* (2023) 3.

[SZKS19] STARKE, SEBASTIAN, ZHANG, HE, KOMURA, TAKU, and SAITO, JUN. "Neural state machine for character-scene interactions." *ACM Trans. Graph.* 38.6 (2019), 209–1 4.

[SZKZ20] STARKE, SEBASTIAN, ZHAO, YIWEI, KOMURA, TAKU, and ZAMAN, KAZI. "Local Motion Phases for Learning Multi-Contact Character Movements". *ACM Trans. Graph.* 39.4 (Aug. 2020). ISSN: 0730-0301 3.

[TAL*22] TIWARI, GARVITA, ANTIĆ, DIMITRIJE, LENSSEN, JAN ERIC, et al. "Pose-ndf: Modeling human pose manifolds with neural distance fields". *European Conference on Computer Vision*. Springer. 2022, 572–589 3.

[TAP*20] TOME, DENIS, ALLDIECK, THIEMO, PELUSE, PATRICK, et al. "Selfpose: 3d egocentric pose estimation from a headset mounted camera". *Transactions on Pattern Analysis and Machine Intelligence (PAMI)* (2020) 2.

[TCL23] TSENG, JONATHAN, CASTELLON, RODRIGO, and LIU, C KAREN. "EDGE: Editable Dance Generation From Music". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023 2, 3, 5.

[TGM*17] TRUMBLE, MATTHEW, GILBERT, ANDREW, MALLESON, CHARLES, et al. "Total Capture: 3D Human Pose Estimation Fusing Video and Inertial Sensors". *British Machine Vision Conference 2017, BMVC 2017, London, UK, September 4-7, 2017*. BMVA Press, 2017 2, 7.

[TRG*23] TEVET, GUY, RAAB, SIGAL, GORDON, BRIAN, et al. "Human Motion Diffusion Model". *ICLR*. 2023 3.

[vMRBP17] Von MARCARD, T., ROSENHAHN, B., BLACK, M. J., and PONS-MOLL, G. "Sparse Inertial Poser: Automatic 3D Human Pose Estimation from Sparse IMUs". *Computer Graphics Forum, the 38th Annual Conference of the European Association for Computer Graphics*. Vol. 36. 2. Chichester, GBR, May 2017, 349–360 3.

[VSP*17] VASWANI, ASHISH, SHAZEER, NOAM, PARMAR, NIKI, et al. "Attention is all you need". *Advances in neural information processing systems* 30 (2017) 3, 5.

[WCX21] WANG, ZHIYONG, CHAI, JINXIANG, and XIA, SHIHONG. "Combining Recurrent Neural Networks and Adversarial Training for Human Motion Synthesis and Control". *IEEE Transactions on Visualization and Computer Graphics* 27.1 (Jan. 2021), 14–28. ISSN: 1077-2626 3.

[WFH08] WANG, JACK M., FLEET, DAVID J., and HERTZMANN, AARON. "Gaussian Process Dynamical Models for Human Motion". *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30.2 (2008), 283–298. DOI: 10.1109/TPAMI.2007.1167 3.

[WLX*23] WANG, JIAN, LUVIZON, DIOGO, XU, WEIPENG, et al. "Scene-aware Egocentric 3D Human Pose Estimation". *CVPR* (2023) 2.

[WWY22] WINKLER, ALEXANDER, WON, JUNGDAM, and YE, YUTING. "QuestSim: Human Motion Tracking from Sparse Sensors with Simulated Avatars". *SIGGRAPH Asia 2022 Conference Papers*. 2022, 1–8 2, 3, 6, 10.

[YKL21] YANG, DONGSEOK, KIM, DOYEON, and LEE, SUNG-HEE. "Lobstr: Real-time lower-body pose prediction from sparse upper-body tracking signals". *Computer Graphics Forum*. Vol. 40. 2. Wiley Online Library. 2021, 265–275 2, 3, 7.

[YLHX22] YE, YONGJING, LIU, LIBIN, HU, LEI, and XIA, SHIHONG. "Neural3Points: Learning to Generate Physically Realistic Full-body Motion for Virtual Reality Users". *Computer Graphics Forum* 41.8 (2022), 183–194 2, 3.

[YPMK23] YE, VICKIE, PAVLAKOS, GEORGIOS, MALIK, JITENDRA, and KANAZAWA, ANGJOO. "Decoupling Human and Camera Motion from Videos in the Wild". *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2023 2.

[YZH*22] YI, XINYU, ZHOU, YUXIAO, HABERMANN, MARC, et al. "Physical Inertial Poser (PIP): Physics-aware Real-time Human Motion Tracking from Sparse Inertial Sensors". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, 13167–13178 3.

[YZH*23] YI, XINYU, ZHOU, YUXIAO, HABERMANN, MARC, et al. "EgoLocate: Real-time Motion Capture, Localization, and Mapping with Sparse Body-mounted Sensors". *ACM Transactions on Graphics (TOG)* 42.4 (2023) 3.

[YZX21] YI, XINYU, ZHOU, YUXIAO, and XU, FENG. "TransPose: real-time 3D human translation and pose estimation with six inertial sensors". *ACM Transactions on Graphics (TOG)* 40.4 (2021), 1–13 3.

[ZBJ*19] ZHOU, YI, BARNES, CONNELLY, JINGWAN, LU, et al. "On the Continuity of Rotation Representations in Neural Networks". *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2019 3.

[ZCP*22] ZHANG, MINGYUAN, CAI, ZHONGANG, PAN, LIANG, et al. "MotionDiffuse: Text-Driven Human Motion Generation with Diffusion Model". *arXiv preprint arXiv:2208.15001* (2022) 3.