Utilizing Task-Generic Motion Prior
to Recover Full-Body Motion from Very Sparse Signals

Myungjin Shin
Yonsei University
mjsh25@yonsei.ac.kr

Dohae Lee
Yonsei University
dlehgo1414@yonsei.ac.kr

In-Kwon Lee
Yonsei University
iklee@yonsei.ac.kr

Abstract

The most popular type of devices used to track a user's posture in a virtual reality experience consists of a head-mounted display and two controllers held in both hands. However, due to the limited number of tracking sensors (three in total), faithfully recovering the user in full-body is challenging, limiting the potential for interactions among simulated user avatars within the virtual world. Therefore, recent studies have attempted to reconstruct full-body poses using neural networks that utilize previously learned human poses or accept a series of past poses over a short period. In this paper, we propose a method that utilizes information from a neural motion prior to improve the accuracy of reconstructed user's motions. Our approach aims to reconstruct user's full-body poses by predicting the latent representation of the user's overall motion from limited input signals and integrating this information with tracking sensor inputs. This is based on the premise that the ultimate goal of pose reconstruction is to reconstruct the motion, which is a series of poses. Our results show that this integration enables more accurate reconstruction of the user's full-body motion, particularly enhancing the robustness of lower body motion reconstruction from impoverished signals. Web: https://https://mjsh34.github.io/mp-sspe/.

1. Introduction

The technology of today's Mixed Reality (MR) has extended traditional interpersonal experiences into the virtual realm, from social gatherings and gaming, to collaborative work, just to name a few. While these experiences have traditionally been limited to settings in which all participants are present in the same environment, MR systems instead rely on virtual avatars to simulate the experience and benefits of non-verbal communication. Unfortunately, real-time data streams
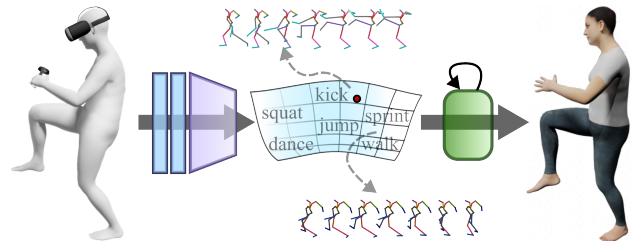

Figure 1. We present a method that utilizes a motion prior to encode the overall motion of a user for full-body pose reconstruction, using only the information of head and two hands.

provided by commercial MR devices, which typically consist of a head-mounted display (HMD) and two hand-held controllers, each tracked in a small room-scale grid, are insufficient to accurately reproduce the full pose of its user, often resulting in MR environments where avatars only show head and hands. Studies have shown that a hand-only avatar provide little sense of embodiment [63, 26, 11], whereas a full-body avatar can significantly enhance the user experience by creating better sense of embodiment and presence [26, 11, 16].

Tasked with the challenge of recovering full-body posture of humans from sparse signal streams, previous works have attempted to reconstruct the human body from tracking signals from four or more joints including the pelvis [21, 62, 31, 66, 68, 69], and from ego-centric cameras [23, 71, 42, 72], which are unavailable on most MR devices at the present day. Recent works have attempted to reconstruct full-body poses from pose information alone from the HMD and hand-held controllers [1, 9, 3, 24, 46, 65, 67]. However when the the reconstructed poses are combined to form a complete motion, they often lead to unnatural motions that fail to match the user's desired action. These systems also often fail to faithfully reproduce lower body motion beyond basic actions such as standing still, and walking at various speeds.

We propose a method effectively utilizing a task-

generic neural motion prior [29, 15, 49, 55] (Section 2.4) aimed at solving the issues mentioned above. We exploit a generative motion prior model with an encoder-decoder architecture that is initially trained to reconstruct full-body motion while learning a latent space of human motions. We train a motion encoder to predict latent representations of motion from a sequence of sparse poses obtained from the three sources mentioned earlier, utilizing latent space learned by motion prior. Finally, we train a sequence (time-series) model that generates full-body pose from the sparse pose sequence and the latent representations of the overall motion. We achieve the following:

- Our method utilizing a motion prior outperforms state-of-the-art methods in reconstructing a full-body pose at a single frame, and in reconstructing motions from combined full-body pose reconstructions from three tracking signals. We evaluate static pose reconstruction and full motion reconstruction performances using appropriate metrics.

- We show that our method produces natural-looking motions which match the intended action of the underlying full-body motion.

- Our method improves on reconstructing lower body motions which methods without any prior on motion struggle at.

We show our model's superior performance against previous works using a diverse set of quantitative metrics (Table 1), user studies (Tables 2 and 3, and Figure 4), and qualitative evaluation (Figure 5).

2. Related Works

2.1. Human Body Representations

The SMPL model [32] represents the human body as a kinematic tree consisting of 24 joints using two parameters: $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$, where $\boldsymbol{\theta} \in \mathbb{R}^{24 \times 3}$ represents the rotations of all 24 joints in the axis-angle representation, and $\boldsymbol{\beta} \in \mathbb{R}^{10}$ represents the shape parameter that describes the body type derived via principle component analysis [32] for each gender (male, female, and neutral). We parametrize the full-body pose using the joint rotations in 6D form, which is a continuous representation of 3D rotation proposed by Yi et al. [76] to be effective in training neural networks (also used by previous works on the same task [9, 3, 24]). For training and inference, we use the neutral body with mean body shape: $\boldsymbol{\beta} = \mathbf{0}$, disregarding variations in body shape, similar to the approach taken in previous works [9, 3, 24] which also do not consider body shape diversity.

2.2. Full-Body Reconstruction from Various Signal Streams

An abundance of literature is dedicated to the recovery of full-body pose from observations of various modalities such as images [4, 43, 73, 74], videos [39, 53, 70, 7, 22], and sparsely-worn body trackers [62, 21, 69, 66, 68]. Notably, the last set of research, while similar to our problem setting, work with richer information by the tracking the pelvis at the very least and often the lower body as well [62, 21, 69, 68].

2.3. Full-Body Reconstruction from Head and Hands

Unlike most previous works that focus on recovering full-body pose from sparse body-worn trackers introduced in Section 2.2, efforts have been made to make use of only three tracking signals, namely the positions and rotations of the head and hands, which most commercial MR devices provide [1, 9, 3, 24, 46, 65, 67].

We categorize the recent lines of works into four types: (1) Motion Matching: Given sparse observations, these works [1, 46] attempt to find the most fitting motion from a predefined animation database. Their primary goals lie not in not precise reconstruction of the full-body, but in accentuating and stylizing motion. (2) Physics-Based Simulation: Recently, QuestSim [65] and Neural3Points [67] have been proposed for simulating full-body avatars by predicting parameters of physical simulation, rendering motions based on laws of physics. In their studies [65, 67], the authors observed that while the synthesized motions are physically plausible, they can exhibit stiffness and unnaturalness. These models also encounter challenges when attempting to replicate complex lower-body motions that have low correlation with their corresponding upper-body movements. Moreover, they are susceptible to deviating from actual movements as errors in physical simulation accumulate and falling over, in which cases simulation needs to be restarted. Neural3Points [67] attempts to mitigate the last issues by using a direct full-pose prediction model in conjunction, but still resorts to restarting the simulation if too much errors accumulate, compromising the realism and accuracy of generated motions. The remaining two lines of work focus on directly predicting full-body pose at every frame. (3) Sequence (Time-Series) Model for Full-Body Pose Estimation: AvatarPoser [24] proposes using a Transformer Encoder [61] to parse a 40-frame sequence of sparse pose signals to predict the full-pose. (4) Generative Latent Space-Based Full-Body Pose Estimation: VAE-HMD [9] and FLAG [3] rely on a decoder of a pose prior to predict the full-pose given a latent code derived from sparse pose signals at the cur-

rent frame or over a short sequence of past frames. For full-pose priors VAE-HMD uses β-VAE [18] for encoder and decoder and FLAG uses RealNVP [8] for encoder and decoder. To estimate latent codes from sparse signals (as substitute for full-pose encoders) VAE-HMD optimizes a β-VAE objective with a new encoder and FLAG employs a Transformer-based [61] predictor.

We have found (3) sequence model-based methods to produce smoother motions than (4) generative latent space-based methods, while the latter tend to more accurately depict full-body pose at a given frame. Our method integrates (3) and (4)'s approaches, predicting full-body pose using a sequence model while simultaneously making use of a generative latent space of motion to produce smooth and accurate full-body motions. We also utilize an explicit motion prior as opposed to static pose priors used by [9, 3].

2.4. Task-Generic Motion Priors

The term task-generic motion prior was first used in HM-VAE [29], to describe "a generalized motion prior, that learns complex human body motions from high-fidelity motion capture data". Unlike task-specific motion priors which are optimized for a single task, such as motion recovery from videos [6, 28, 35], task-generic motion priors [15, 29, 49, 55] are generative models that can perform an array of motion-related tasks, such as motion interpolation (in-betweening), completion, synthesis, refinement, and even recovery of full-pose from partial observations via test-time optimization. NeMF [15] categorizes these motion priors into two categories: time-series models and space-time models. Time-series models predict future motions based on past observations and are typically autoregressive, with HuMoR [49] being an example. Space-Time models, on the other hand, directly model the spatio-temporal kinematic state, often by taking in a whole motion as input at once [15, 29, 55].

We select MotionCLIP [55] for our full-body motion prior, a space-time motion prior with an auto-encoder [13] architecture. MotionCLIP learns to embed the input motion in the latent space of CLIP [48], a large-scale neural network trained jointly on image and text. The CLIP space has demonstrated its effectiveness for use in downstream tasks in various domains, such as image [10, 12, 44, 50], 3D [40, 51, 64], and human motion [55, 57, 75]. Furthermore, MotionCLIP achieves zero-shot action classification performance close to 2s-AGCN [52], a dedicated action classifier, demonstrating the latent space's ability to discriminate between different action types [55].

3. Methods

Our framework consists of the following components: *full motion prior*, *sparse motion encoder*, and *sequence model*. The full motion prior consists of full motion encoder and decoder (Figure 3), trained on full-pose motions to learn the motion latent space. We train this component first, followed by training sparse motion encoder. The goal of the sparse motion encoder is to predict the *motion latent* (**M**) in the space learned by our full motion prior, utilizing only sparse pose signals. Sparse motion encoder and sequence model are used directly for full-body pose estimation, the process visualized in Figure 2. We feed the sparse pose signals to the sparse motion encoder after augmentation, from which we extract the *motion embedding* (**E**), which is a compressed representation of motion. We then concatenate the motion embedding with the augmented sparse pose signals after normalization step. Finally, the concatenated sequence is input to sequence model for full-pose reconstruction.

3.1. Input and Output Representations

We represent the sparse pose signals at time t as $\mathbf{x}_t$, defined as:

$$\mathbf{x}_t = [\mathbf{g}_t^{\mathbf{3}}, \mathbf{r}_t^{\mathbf{3}}], \tag{1}$$

where $\mathbf{g}_t^{\mathbf{3}}$ and $\mathbf{r}_t^{\mathbf{3}}$ respectively represent the global positions (3D) and the rotations in 6D form [76] of head and hands, as would be provided by the MR device. From this data, we derive *sparse motion signals* at time t denoted $\mathbf{X}_t$, and a *sparse motion sequence* of length T at time t denoted $\mathbf{X}_{t-T+1:t}$, each defined as:

$$\mathbf{X}_t = [\mathbf{g}_t^{\mathbf{3}}, \dot{\mathbf{g}}_t^{\mathbf{3}}, \mathbf{r}_t^{\mathbf{3}}, \dot{\mathbf{r}}_t^{\mathbf{3}}] \in \mathbb{R}^{54}, \tag{2}$$

$$\mathbf{X}_{t-T+1:t} = [\mathbf{X}_{t-T+1}, \mathbf{X}_{t-T+2}, ..., \mathbf{X}_t] \in \mathbb{R}^{54 \times T}, \tag{3}$$

where $\dot{\mathbf{g}}_t^{\mathbf{3}}$ and $\dot{\mathbf{r}}_t^{\mathbf{3}}$ respectively represent the velocities and angular velocities in 6D form derived from $\mathbf{g}_t^{\mathbf{3}}$ and $\mathbf{r}_t^{\mathbf{3}}$ as done in AvatarPoser [24]. This process is represented as "Input Augmentation" in Figure 2. The global position $\mathbf{g}_t^{\mathbf{3}}$ depends on an origin point decided by the MR device, which can be arbitrary. We counteract the randomness of input by "normalizing" the global positions $\mathbf{g}_t^{\mathbf{3}}$ in the horizontal axes (x and z axes) as follows:

$$\mathbf{g}_t^{\mathbf{3}} = [\mathbf{g}_t^{head}, \mathbf{g}_t^{lhand}, \mathbf{g}_t^{rhand}], \tag{4}$$

$$\bar{\mathbf{g}}_t.xz = \left(\frac{\mathbf{g}_t^{head} + \mathbf{g}_t^{lhand} + \mathbf{g}_t^{rhand}}{3}\right).xz, \tag{5}$$

$$\mathbf{n}_t^j.xz = (\mathbf{g}_t^j - \bar{\mathbf{g}}_t).xz, \quad \mathbf{n}_t^j.y = \mathbf{g}_t^j.y,$$
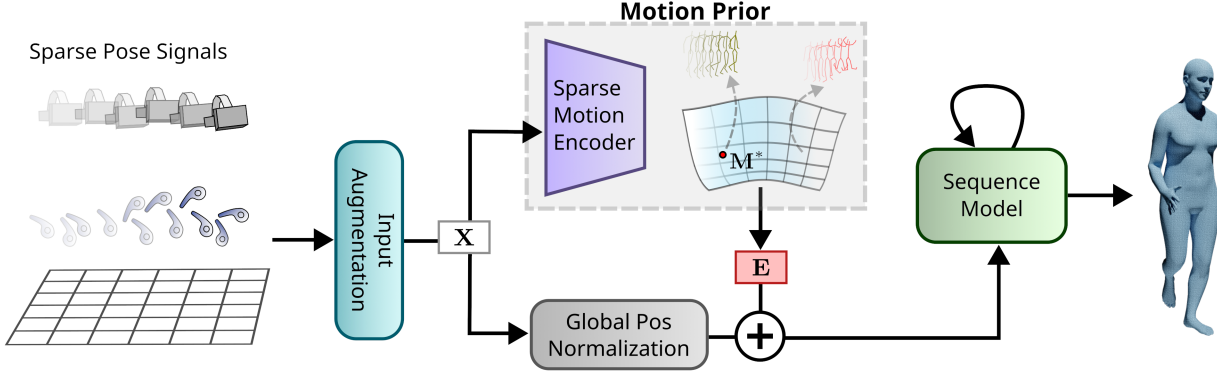$$\forall j \in \{head, lhand, rhand\}, \tag{6}$$

Figure 2. **Model Overview.** Sparse pose signals are fed to sparse motion encoder after augmentation to derive a motion embedding. Motion embeddings are combined with normalized augmented signals and input the sequence model for full-pose reconstruction. We tried different network architectures during development, and readers may refer to the supplementary material for details and experimental results.

$$\mathbf{n}_t^3 = [\mathbf{n}_t^{head}, \mathbf{n}_t^{lhand}, \mathbf{n}_t^{rhand}], \qquad (7)$$

where $\mathbf{n}_t^3$ denote the normalized positions, normalized positions of 3 joints having mean of 0 along x and z axes. This process is represented as "Global Pos Normalization" in Figure 2. We found empirically the information lost by using normalized global positions on the horizontal axes to be compensated for by the other inputs obtained via augmentation, and the normalization step to help the sequence model produce more stable motions. We did not observe the same benefit while training the sparse motion encoder, so we apply normalization only before the sequence model.

The output consists of relative rotations (6D form) for 22 SMPL [32] joints, which can be used to recover the entire kinematic tree up to both wrists via forward kinematics (FK).

3.2. Full Motion Prior Pretraining

Full motion prior denotes motion prior whose encoder and decoder are trained on full-pose sequences, i.e., full motions. We use MotionCLIP [55], which is a full-body motion auto-encoder [13] exploiting the powerful latent space of CLIP [48]. As visualized in Figure 3, when the full 60-frame pose sequence is input to the encoder (based on the Transformer Encoder architecture [61]), denoted *full motion encoder*, the output is a latent vector lying in CLIP space, denoted *motion latent* (**M**). The decoder (based on the Transformer Decoder architecture [61]), denoted *full motion decoder*, aims to reconstruct the same full motion from the motion latent. The loss $\mathcal{L}_{fm}$ used to train the full motion prior is formulated as follows:

$$\mathcal{L}_{fm} = \mathcal{L}_{recon} + \lambda_{text}\mathcal{L}_{text} + \lambda_{image}\mathcal{L}_{image}, \quad (8)$$

$$\mathcal{L}_{text} = 1 - \cos(CLIP_{text}(t), \mathbf{M}), \qquad (9)$$
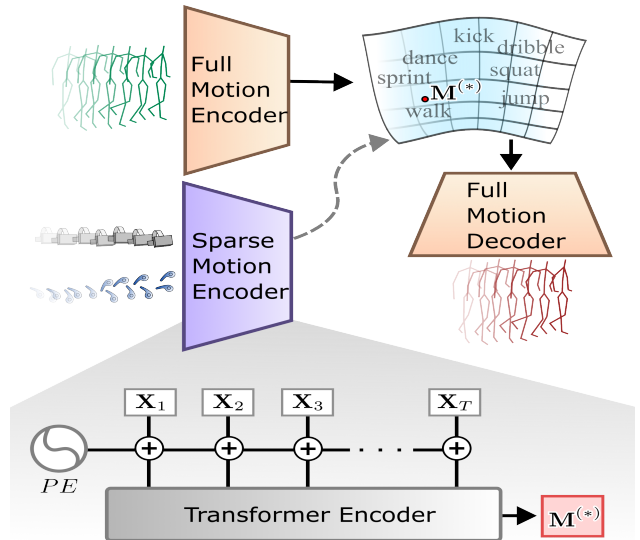

Figure 3. **Motion Prior.** After pretraining full motion prior, sparse motion encoder is trained to predict the same motion latent as the full motion encoder.

$$\mathcal{L}_{image} = 1 - \cos(CLIP_{image}(s), \mathbf{M}), \qquad (10)$$

where $\mathcal{L}_{recon}$ is the reconstruction loss of the full motion, and $\mathbf{M} \in \mathbb{R}^{512}$ represents the motion latent, which is the output of the full motion encoder. $\mathcal{L}_{text}$ and $\mathcal{L}_{image}$ are the cosine distances from the motion latent $\mathbf{M}$ to its corresponding text projection $CLIP_{text}(t)$, and the image projection $CLIP_{image}(t)$, respectively. Instead of learning a new latent space, as a variational auto-encoder [27] would do, CLIP space projection of the text labels $CLIP_{text}(t)$ and that of the rendered images $CLIP_{image}(s)$ corresponding respectively to each motion (both of which are part of dataset) are used to guide the motion latents to lie close together (with

corresponding text and image projections) on the same space. We train with the configuration named "paper_model" [56]. This module is trained before all else.

3.3. Estimating Motion Latent from Sparse Motion Sequence

We train the sparse motion encoder to estimate the motion latent from sparse motion sequence $\mathbf{X}_{t-T+1:t}$. The architecture is adapted from full motion encoder's Transformer Encoder [61] having the linear layer before the Transformer Encoder modified to accept $\mathbf{X}_{t-T+1:t} \in \mathbb{R}^{54 \times T}$. We use $T = 60$ which is the same motion length used by the full motion prior. We keep the full motion decoder and keep its weights frozen when we train the auto-encoder consisting of sparse motion encoder and full motion decoder (Figure 3) with loss $\mathcal{L}_{sm}$ as follows:

$$\mathcal{L}_{sm} = \lambda^*_{text}\mathcal{L}^*_{text} + \lambda^*_{image}\mathcal{L}^*_{image}, \qquad (11)$$

$$\mathcal{L}^*_{text} = 1 - \cos(CLIP_{text}(t), \mathbf{M}^*), \qquad (12)$$

$$\mathcal{L}^*_{image} = 1 - \cos(CLIP_{image}(s), \mathbf{M}^*), \qquad (13)$$

where $\mathbf{M}^* \in \mathbb{R}^{512}$ denotes the motion latent predicted by the sparse motion encoder. We set $\lambda^*_{text} = \lambda^*_{image} = 0.01$. This module is trained after the full motion prior.

3.4. Sequence Model to Reconstruct Full-Pose from Sparse Motion Sequence and Motion Latent

Sequence model takes as input the length-S sparse motion sequence $\mathbf{X}_{t-S+1:t}$ and the corresponding sequence of *motion embeddings* $\mathbf{E}_{t-S+1:t} = [\mathbf{E}_{t-S+1}, \mathbf{E}_{t-S+2}, ..., \mathbf{E}_t]$ as input. A motion embedding $\mathbf{E}_t \in \mathbb{R}^{64}$ is derived by passing the predicted motion latent $\mathbf{M}^*_t \in \mathbb{R}^{512}$ through a linear layer to retrieve a compressed 64-dimensional representation of the motion. Then, $\mathbf{X}_{t-S+1:t}$ and $\mathbf{E}_{t-S+1:t}$ are concatenated along the time axis, to be input to the sequence model. 3-Layer LSTM [19] is our choice of sequence model, which outputs a single full-body pose at time t given a sequence of inputs of length S from time $t - S + 1$ to t. We represent the full-pose as the 6D relative rotation values of 22 joints of the SMPL model, from which we can recover the absolute rotations $\hat{\mathbf{r}}^{22}_t$ and the body root-relative positions $\hat{\mathbf{p}}^{22}_t$ of 22 joints via FK.

The loss $\mathcal{L}_{seq}$ is computed as the weighted sum of rotational loss $\mathcal{L}_{rot}$, positional loss $\mathcal{L}_{pos}$, velocity loss $\mathcal{L}_{vel}$ (ablation for $\mathcal{L}_{vel}$ in supplementary), and motion loss $\mathcal{L}_{mo}$. $\mathcal{L}_{rot}$, $\mathcal{L}_{pos}$, and $\mathcal{L}_{vel}$ encourage accurate full-pose reconstruction at every frame, and are computed as the L2 norm between the predicted and corresponding GT values, respectively. The motion loss $\mathcal{L}_{mo}$ encourages the model to learn the correct motion given the consecutive 60-frame full-pose predictions $\hat{\mathbf{p}}^{22}_{t-60+1:t}$, which

are passed through a full motion encoder to obtain motion latent $\hat{\mathbf{M}}_t$. We also obtain the ground truth motion latent $\mathbf{M}_t$ with the corresponding ground truth motion $\mathbf{p}^{22}_{t-60+1:t}$ via the same full motion encoder. While we use the pretrained full motion encoder in Section 3.2 for this purpose, a different full motion encoder could substitute it. We then compute the cosine distance between the two motion latents to obtain motion loss:

$$\mathcal{L}_{mo} = 1 - \cos(\hat{\mathbf{M}}_t, \mathbf{M}_t). \qquad (14)$$

Finally, the total loss of the sequence model $\mathcal{L}_{seq}$ is computed as follows:

$$\mathcal{L} = \lambda_{rot}\mathcal{L}_{rot} + \lambda_{pos}\mathcal{L}_{pos} + \lambda_{vel}\mathcal{L}_{vel} + \lambda_{mo}\mathcal{L}_{mo}. \qquad (15)$$

We set the coefficients $\lambda_{rot} = \lambda_{pos} = \lambda_{vel} = 1.0, \lambda_{mo} = 0.1$. Moreover, we found freezing the sparse motion encoder's weights to yield better results (Section 5), and to allow preprocessing the motion latents in advance for faster training.

4. Experimental Results

4.1. Data Preparation and Network Training

We train and test all our models on the AMASS [37] dataset, a large-scale human motion dataset parametrized by the SMPL model [32]. Since AMASS contains motion capture data with varying frame rates, we downsample each mocap data to be close to 30 FPS. From AMASS, we extract the head (joint index 15) and two wrist joints (joints indices 20 and 21) and derive their root-relative positions via FK, followed by adding translation to simulate global position signals given by MR devices. We also derive the absolute rotations of head and hands to simulate rotation signals. Data are then ready to be processed by the procedure described in Section 3.1. All of our model components and baselines share the same input and output format (n.b., while VAE-HMD [9] was originally tested given pelvis-relative positions as input, we input global positions to reflect the signals from MR devices). We use the AMASS subset consisting of BMLrub [58], EyesJapanDataset [34], TotalCapture [59], KIT [38], ACCAD [38], CMU [5], PosePrior [2], TCDHands [20], EKUT for training and set aside HumanEva [54], HDM05 [41], SFU [60], MoSh [31], Transitions, SSM for evaluation. For training motion priors, we additionally use BABEL [47], dataset containing per-frame action labels corresponding to a large portion of AMASS for the text labels, and images are rendered via MotionCLIP's official open-source implementation [56].

The full motion prior and the sparse motion encoder each takes 10 hours, and the sequence model 5 hours

Method	Per-Joint Errors				Motion-Related Statistics	
	MPJPE	Legs MPJPE	Global MPJPE	MPJVE	Motion Distance ↓	FID ↓
Ours	**7.25**	**9.34**	**7.38**	**25.42**	$\mathbf{5.12 \cdot 10^{-3}}$	$\mathbf{6.03 \cdot 10^{-2}}$
AvatarPoser [24]	7.71	10.25	7.79	29.71	$5.38 \cdot 10^{-3}$	$7.59 \cdot 10^{-2}$
AvatarPoser-60	7.74	10.39	7.82	29.85	$5.47 \cdot 10^{-3}$	$7.87 \cdot 10^{-2}$
VAE-HMD [9]	7.48	**9.34**	7.78	54.84	$8.13 \cdot 10^{-3}$	$9.24 \cdot 10^{-2}$
VAE-HMD-60	8.46	10.69	8.91	60.31	$8.09 \cdot 10^{-3}$	$9.40 \cdot 10^{-2}$

Table 1. **Main Quantitative Results.**

after preprocessing the motion latents via the trained sparse motion encoder (possible because sparse motion encoder is kept frozen during sequence model training) which takes about an hour, for a total 26 hours for full training on a single NVIDIA RTX 2080 Ti. For the baseline models, we followed the setup described in the original papers [9, 24] as closely as possible. For Avatar-Poser, we use the official open-source implementation [25]. For VAE-HMD [9], which has no open-source implementation available, we implemented their best performing model according to the original paper which contains a pretrained pose prior component. Note that we selected a much wider subset of AMASS for training and testing than the original works [9, 24]. Refer to the supplementary material for more details about training and baseline implementations.

4.2. Quantitative Evaluation

We quantitatively evaluate our model using a diverse set of metrics against two baselines: Avatar-Poser [24] and VAE-HMD [9]. The quantitative results are presented in Table 1. Originally, AvatarPoser and VAE-HMD had window sizes of 40 frames and 16 frames, respectively. To ensure fairness, we additionally compare against AvatarPoser and VAE-HMD each adapted to have a 60-frame window size, which is the same as the window size that our motion prior sees. The adapted versions are labeled AvatarPoser-60 and VAE-HMD-60 in Table 1 (further results and analyses in supplementary).

Per-Joint Errors. We use four per-joint error metrics to evaluate our approach: MPJPE (mean per-joint position error [cm]), Legs MPJPE [cm], Global MPJPE [cm], and MPJVE (mean per-joint velocity error [cm/s]). Global MPJPE is computed by first mapping the predicted joints to global space, which involves combining GT head position (given by MR device) with the predicted joint rotations of the full body. The results in Table 1 demonstrate that our approach outperforms baselines on the majority of the metrics evaluated, rivalled only by VAE-HMD on Legs MPJPE. Since the sparse pose signals only contain direct infor-

mation about the upper body, accurate reconstruction of legs is challenging especially when leg motions have low correlation with co-ocurring upper body motion. Use of a motion prior generally results in lower error in leg motions, which can also be seen in Section 5 and in Figure 5 qualitatively. VAE-HMD contains a decoder component of a pretrained auto-encoder whose weights are frozen, and during pretraining, it receives full-body motions as input and learns a prior, which helps reconstruct some difficult motions, as evidenced by the legs MPJPE metric being as low as ours. However, VAE-HMD suffers from relatively high velocity error (MPJVE) compared to other methods.

Motion-Related Errors. Motion distance measures the difference between the overall motion (spanning 60 frames), between the GT motion and predicted motion calculated the same way as Equation 14. For this evaluation we use a different motion prior from one we optimized with, namely "classes_model" of Motion-CLIP [55]. Although classes_model's motion latents also lie in CLIP space [48], different training parameters and dataset are used to train, resulting in different motion latents being predicted (Also, the text and images' CLIP projections do not align [55], so different motion latent spaces are learned depending on training configuration). FID (Fretchet Inception Distance [17]) measures the similarity between the distributions of ground truth motions and generated motions, the lower the more similar. While FID has been used widely in the context of generated human motions [14, 45, 30, 15], our work is the first to use it for full-body motions reconstructed from sparse signals. Our model achieves the lowest FID, followed by AvatarPoser variants, and VAE-HMD-60 with the highest FID. These results are consistent with the findings from our user studies (Section 4.3) where we evaluate the quality of generated motions.

4.3. User Study

We define two sets of motion segments for the user studies: Random Set, consisting of 60 3-second segments randomly sampled from the entire test dataset,

	General Preference			
	Random Set		Hard Set	
Other Model	Ours	Other	Ours	Other
AvatarPoser	**58.8%**	41.2%	**65.0%**	35.0%
VAE-HMD	**94.9%**	5.10%	**95.8%**	4.20%

Table 2. **Result of User Study I: General Preference.** (Better reconstruction of ground truth motion.)

	Motion Matching					
	Random Set			Hard Set		
Other Model	Ours	Other	Neut	Ours	Other	Neut
AvatarPoser	**43.9%**	32.8%	23.3%	**52.6%**	28.8%	18.6%
VAE-HMD	**76.9%**	10.3%	12.8%	**79.8%**	12.7%	7.50%

Table 3. **Result of User Study I: Motion Matching.** (Which generated motion matches ground truth motion better?)
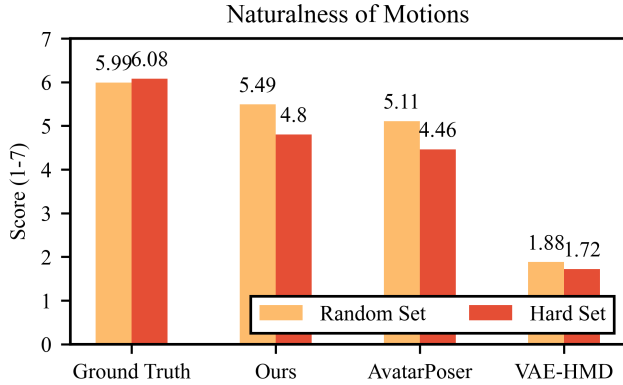

Figure 4. **Result of User Study II: Naturalness of Motions.** Scores range from 1 (worst) to 7 (best).

and Hard Set, which includes 60 3-second segments. For Hard Set, we sampled 30 motion segments on which one of our baselines (neither AvatarPoser nor VAE-HMD, but the "No Motion Prior" model explained in Section 5), trained without motion prior, had the highest MPJPE and Legs MPJPE each. We conducted user studies I and II, where for each user study, we (disjointly) sampled 30 segments, from Random Set and Hard Set respectively. We conducted the user studies on 30 participants. Full-body animations of ground truth motions, predictions from our model, Avatar-Poser [24], and VAE-HMD [9] for each segment were rendered using SMPL Blender addon [33].

User Study I. To compare our model with one of the baselines side-by-side, we placed ground truth animation on top and juxtaposed our model's prediction and one of AvatarPoser and VAE-HMD's animations at the bottom (latter's order randomized) for the same underlying motion. We informed participants

that the two animations at the bottom are different reconstructions of the top animation from partial information, and we asked to choose (1) a better reconstruction among the two (preference), and (2) one whose motion matched the ground truth motion better, with "neutral" option added for the latter.

The results for (1) can be found in Table 2. The table shows participants' preference towards our model's predictions for both sets, with more pronounced results with Hard Set. This demonstrates our model's ability to better reconstruct motions that the baseline, which does not make use of a motion prior, struggles with. Moreover, participants clearly preferred our model's predictions over VAE-HMD's, noting that VAE-HMD's animations looked unnatural primarily due to jitter. This is consistent with the high velocity error (MPJVE) measured in Section 4.2. The results of (2) can be found in Table 3, where we evaluated the motion matching capability of our model in addition to the quantitative motion distance metric in Section 4.2. The results show similar trends as (1).

User Study II. We played all four animations with the same underlying motion simultaneously, ordered randomly, asking participants to rate the naturalness of each motion on a scale of 1 to 7, 7 being the highest (participants were allowed to replay animations as they desired). From the mean scores plotted in Figure 4, we can observe that users found GT motions the most natural, followed by generations from our model's predictions, then AvatarPoser and VAE-HMD, in that order. While the scores for predicted motions for Hard Set fall behind those for Random Set, participants found the motions generated via our model more natural than other models' generations for both sets.

5. Ablation Studies

We conduct ablation studies by removing or modifying different subcomponents, the main quantitative results shown in Table 4.

We first assess the role of the motion prior component in our architecture by completely removing it from our model, leaving only the input processing and sequence model components (see Figure 2 for reference). The results in "No Motion Prior" row in Table 4 show degradation in values for all quantitative metrics. (This model was used to curate the Hard Set for the user study, as explained in Section 4.3.) Additionally, we group all motion segments in the test dataset by action types defined in BABEL [47], and sort them by improvement of Legs MPJPE and MPJPE respectively. We present the top 5 and bottom 5 action types in Table 5, left column showing top 5 and right column showing bottom 5. Top 5 improved action types for

Figure 5. **Qualitative Results**. We show qualitative results on difficult motions with less common lower body movements. Left Column: Kicking Motion (Transitions/mazen_c3d/kick_push_poses [37]). Right Column: Moonwalking (Transitions/mazen_c3d/run_stand_poses [37]).

| Method | Per-Joint Errors | | | | Motion-Related Statistics | |
	MPJPE	Legs MPJPE	Global MPJPE	MPJVE	Motion Distance ↓	FID ↓
Ours	**7.25**	9.34	**7.38**	**25.42**	$\mathbf{5.12 \cdot 10^{-3}}$	$\mathbf{6.03 \cdot 10^{-2}}$
No Motion Prior	7.37	9.67	7.62	26.22	$5.52 \cdot 10^{-3}$	$7.54 \cdot 10^{-2}$
No Motion Distance Loss	7.32	9.43	7.45	25.71	$5.31 \cdot 10^{-3}$	$7.08 \cdot 10^{-2}$
With Finetuned Motion Prior	7.39	9.77	7.67	26.10	$5.28 \cdot 10^{-3}$	$6.24 \cdot 10^{-2}$
With a Different Motion Prior	7.29	**9.27**	7.41	26.15	$5.15 \cdot 10^{-3}$	$6.20 \cdot 10^{-2}$

Table 4. **Ablation Studies: Main Quantitative Results**

Improved Action Type	Degraded Action Type
Legs MPJPE Improvement/Degradation	
knee movement	place something
cartwheel	grasp object
crouch	**poses**
squat	stretch
bend	face direction
MPJPE Improvement/Degradation	
cartwheel	face direction
shuffle	place something
knee movement	**lean**
throw	take/pick something up
touch ground	shout

Table 5. **Ablation Studies: Action Types [47] Improved by Using Motion Prior.** Actions containing high amount of leg motions are in bold. Our model's top improvements lie in actions containing much leg motions.

both metrics consist of actions involving a high amount of leg motions [1], showing that the motion prior contributes to better reconstruction of leg motions given only upper body signals. We can also observe degra-

dation from not using the motion distance loss ("No Motion Distance Loss") and from unfreezing the motion prior ("With Finetuned Motion Prior").

We experimented with various task-generic motion priors during development and settled on MotionCLIP as it gave the best overall performance. The final row of Table 4 shows the result of using a Transformer VAE-based [61, 27] motion prior whose architecture is based on ACTOR [45], from which we removed action conditioning part to have an unconditional motion VAE [27].

6. Conclusions and Limitations

We present a method of utilizing motion prior to effectively reconstruct full-body motion from impoverished signals of pose. Our method recovers intended full-body motions that look natural, with improved lower body over baselines. However, our work only considers a single body type, and we wish to allow people of diverse body shapes to utillize our system effectively in a future work. Moreover, we sometimes observe foot-sliding artifacts from generated motions, and we wish to measure their severity and eliminate them.

---

[1]A list of BABEL [47] action subtypes corresponding to each action type can be found in [36].

References

[1] Karan Ahuja, Eyal Ofek, Mar Gonzalez-Franco, Christian Holz, and Andrew D Wilson. Coolmoves: User motion accentuation in virtual reality. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 5(2):1–23, 2021.

[2] Ijaz Akhter and Michael J. Black. Pose-conditioned joint angle limits for 3D human pose reconstruction. In *2015 IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), pages 1446–1455, June 2015.

[3] Sadegh Aliakbarian, Pashmina Cameron, Federica Bogo, Andrew Fitzgibbon, and Thomas J Cashman. Flag: Flow-based 3d avatar generation from sparse observations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13253–13262, 2022.

[4] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J. Black. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *Computer Vision – ECCV 2016*, Lecture Notes in Computer Science. Springer International Publishing, Oct. 2016.

[5] Carnegie Mellon University. CMU MoCap Dataset.

[6] Hongsuk Choi, Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. Beyond static features for temporally consistent 3d human pose and shape from a video. In *Conference on Computer Vision and Pattern Recognition* (*CVPR*), 2021.

[7] Rishabh Dabral, Soshi Shimada, Arjun Jain, Christian Theobalt, and Vladislav Golyanik. Gravity-aware monocular 3d human-object reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (*ICCV*), pages 12365–12374, October 2021.

[8] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp, 2016.

[9] Andrea Dittadi, Sebastian Dziadzio, Darren Cosker, Ben Lundell, Thomas J Cashman, and Jamie Shotton. Full-body motion from a single head-mounted device: generating smpl poses from partial observations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11687–11697, 2021.

[10] Kevin Frans, Lisa B Soros, and Olaf Witkowski. Clipdraw: Exploring text-to-drawing synthesis through language-image encoders. *arXiv preprint arXiv:2106.14843*, 2021.

[11] Rebecca Fribourg, Ferran Argelaguet, Anatole Lécuyer, and Ludovic Hoyet. Avatar and sense of embodiment: Studying the relative preference between appearance, control and point of view. *IEEE Transactions on Visualization and Computer Graphics*, 26(5):2062–2072, 2020.

[12] Rinon Gal, Or Patashnik, Haggai Maron, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators, 2021.

[13] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. Adaptive Computation and Machine Learning series. MIT Press, 2016.

[14] Chuan Guo, Xinxin Zuo, Sen Wang, Shihao Zou, Qingyao Sun, Annan Deng, Minglun Gong, and Li Cheng. Action2motion: Conditioned generation of 3d human motions. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2021–2029, 2020.

[15] Chengan He, Jun Saito, James Zachary, Holly Rushmeier, and Yi Zhou. Nemf: Neural motion fields for kinematic animation. In *NeurIPS*, 2022.

[16] Paul Heidicker, Eike Langbehn, and Frank Steinicke. Influence of avatar appearance on presence in social vr. 01 2017.

[17] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[18] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International conference on learning representations*, 2017.

[19] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[20] Ludovic Hoyet, Kenneth Ryall, Rachel McDonnell, and Carol O' Sullivan. Sleight of hand: Perception of finger motion from reduced marker sets. In *Proceedings of the ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games*, I3D ' 12, page 79–86, New York, NY, USA, 2012.

[21] 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,* (*Proc. SIGGRAPH Asia*), 37:185:1–185:15, Nov. 2018. First two authors contributed equally.

[22] Karim Iskakov, Egor Burkov, Victor Lempitsky, and Yury Malkov. Learnable triangulation of human pose. In *International Conference on Computer Vision* (*ICCV*), 2019.

[23] Hao Jiang and Kristen Grauman. Seeing invisible poses: Estimating 3d body pose from egocentric video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), July 2017.

[24] Jiaxi Jiang, Paul Streli, Huajian Qiu, Andreas Fender, Larissa Laich, Patrick Snape, and Christian Holz. Avatarposer: Articulated full-body pose tracking from sparse motion sensing. In *Proceedings of European Conference on Computer Vision*. Springer, 2022.

[25] Jiaxi Jiang, Paul Streli, Huajian Qiu, Andreas Fender, Larissa Laich, Patrick Snape, and Christian Holz.

Avatarposer: Articulated full-body pose tracking from sparse motion sensing. `https://github.com/eth-siplab/AvatarPoser`, 2022. [Online; accessed 1-March-2023].

[26] Sungchul Jung and Charles E. Hughes. The effects of indirect real body cues of irrelevant parts on virtual body ownership and presence. In *Proceedings of the 26th International Conference on Artificial Reality and Telexistence and the 21st Eurographics Symposium on Virtual Environments*, ICAT-EGVE '16, page 107–114, Goslar, DEU, 2016. Eurographics Association.

[27] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2013.

[28] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. Vibe: Video inference for human body pose and shape estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), June 2020.

[29] Jiaman Li, Ruben Villegas, Duygu Ceylan, Jimei Yang, Zhengfei Kuang, Hao Li, and Yajie Zhao. Task-generic hierarchical human motion prior using vaes. In *2021 International Conference on 3D Vision* (*3DV*), pages 771–781, 2021.

[30] Hung Yu Ling, Fabio Zinno, George Cheng, and Michiel Van De Panne. Character controllers using motion vaes. *ACM Trans. Graph.*, 39(4), aug 2020.

[31] Matthew Loper, Naureen Mahmood, and Michael J. Black. MoSh: Motion and Shape Capture from Sparse Markers. *ACM Trans. Graph.*, 33(6), Nov. 2014.

[32] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics* (*Proc. SIGGRAPH Asia*), 34(6):248:1–248:16, Oct. 2015.

[33] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL Blender addon downloads page. `https://smpl.is.tue.mpg.de/download.php`, 2015. [Online; accessed 1-March-2023].

[34] Eyes JAPAN Co. Ltd. Eyes Japan MoCap Dataset.

[35] Zhengyi Luo, S. Alireza Golestaneh, and Kris M. Kitani. 3d human motion estimation via motion compression and refinement. In *Proceedings of the Asian Conference on Computer Vision* (*ACCV*), November 2020.

[36] Ingo Lütkebohle. BABEL Action categories and their corresponding "raw" labels from annotators. `https://babel.is.tue.mpg.de/actionsrawlabels.html`, 2021. [Online; accessed 1-March-2023].

[37] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *International Conference on Computer Vision*, pages 5442–5451, Oct. 2019.

[38] C. Mandery, Ö. Terlemez, M. Do, N. Vahrenkamp, and T. Asfour. The KIT whole-body human motion database. In *2015 International Conference on Advanced Robotics* (*ICAR*), pages 329–336, July 2015.

[39] Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas, and Christian Theobalt. Vnect: Real-time 3d human pose estimation with a single rgb camera. volume 36, July 2017.

[40] Oscar Michel, Roi Bar-On, Richard Liu, Sagie Benaim, and Rana Hanocka. Text2mesh: Text-driven neural stylization for meshes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), pages 13492–13502, June 2022.

[41] M. Müller, T. Röder, M. Clausen, B. Eberhardt, B. Krüger, and A. Weber. Documentation mocap database HDM05. Technical Report CG-2007-2, Universität Bonn, June 2007.

[42] Evonne Ng, Donglai Xiang, Hanbyul Joo, and Kristen Grauman. You2me: Inferring body pose in egocentric video via first and second person interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), June 2020.

[43] Mohamed Omran, Christoph Lassner, Gerard Pons-Moll, Peter V. Gehler, and Bernt Schiele. Neural body fitting: Unifying deep learning and model-based human pose and shape estimation. Verona, Italy, 2018.

[44] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (*ICCV*), pages 2085–2094, October 2021.

[45] Mathis Petrovich, Michael J. Black, and Gül Varol. Action-conditioned 3D human motion synthesis with transformer VAE. In *International Conference on Computer Vision* (*ICCV*), 2021.

[46] Jose Luis Ponton, Haoran Yun, Carlos Andujar, and Nuria Pelechano. Combining Motion Matching and Orientation Prediction to Animate Avatars for Consumer-Grade VR Devices. *Computer Graphics Forum*, 2022.

[47] Abhinanda R. Punnakkal, Arjun Chandrasekaran, Nikos Athanasiou, Alejandra Quiros-Ramirez, and Michael J. Black. BABEL: Bodies, action and behavior with english labels. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition* (*CVPR*), pages 722–731, June 2021.

[48] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[49] Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J. Guibas. Humor: 3d human motion model for robust pose estimation. In *International Conference on Computer Vision* (*ICCV*), 2021.

[50] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-

resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[51] Aditya Sanghi, Hang Chu, Joseph G. Lambourne, Ye Wang, Chin-Yi Cheng, Marco Fumero, and Kamal Rahimi Malekshan. Clip-forge: Towards zero-shot text-to-shape generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), pages 18603–18613, June 2022.

[52] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), June 2019.

[53] Mingyi Shi, Kfir Aberman, Andreas Aristidou, Taku Komura, Dani Lischinski, Daniel Cohen-Or, and Baoquan Chen. Motionet: 3d human motion reconstruction from monocular video with skeleton consistency. *arXiv preprint arXiv:2006.12075*, 2020.

[54] L. Sigal, A. Balan, and M. J. Black. HumanEva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International Journal of Computer Vision*, 87(4):4–27, Mar. 2010.

[55] Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. Motionclip: Exposing human motion generation to clip space. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXII*, pages 358–374. Springer, 2022.

[56] Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. Motionclip: Exposing human motion generation to clip space. `https://github.com/GuyTevet/MotionCLIP`, 2022. [Online; accessed 1-March-2023].

[57] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Amit H Bermano, and Daniel Cohen-Or. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*, 2022.

[58] Nikolaus F. Troje. Decomposing biological motion: A framework for analysis and synthesis of human gait patterns. *Journal of Vision*, 2(5):2–2, Sept. 2002.

[59] Matthew Trumble, Andrew Gilbert, Charles Malleson, Adrian Hilton, and John Collomosse. Total capture: 3d human pose estimation fusing video and inertial sensors. In *Proceedings of the British Machine Vision Conference* (*BMVC*), pages 14.1–14.13, Sept. 2017.

[60] Simon Fraser University and National University of Singapore. SFU Motion Capture Database.

[61] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[62] T. von Marcard, B. Rosenhahn, M. J. Black, and G. Pons-Moll. Sparse inertial poser: Automatic 3d human pose estimation from sparse imus. *Comput. Graph. Forum*, 36(2):349–360, may 2017.

[63] Thomas Waltemate, Dominik Gall, Daniel Roth, Mario Botsch, and Marc Erich Latoschik. The impact of avatar personalization and immersion on virtual body ownership, presence, and emotional response. *IEEE Transactions on Visualization and Computer Graphics*, 24(4):1643–1652, 2018.

[64] Can Wang, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao. Clip-nerf: Text-and-image driven manipulation of neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), pages 3835–3844, June 2022.

[65] Alexander Winkler, Jungdam Won, and Yuting Ye. Questsim: Human motion tracking from sparse sensors with simulated avatars. In *SIGGRAPH Asia 2022 Conference Papers*, pages 1–8, 2022.

[66] Dongseok Yang, Doyeon Kim, and Sung-Hee Lee. Lobstr: Real-time lower-body pose prediction from sparse upper-body tracking signals. In *Computer Graphics Forum*, volume 40, pages 265–275. Wiley Online Library, 2021.

[67] Yongjing Ye, Libin Liu, Lei Hu, and Shihong Xia. Neural3points: Learning to generate physically realistic full-body motion for virtual reality users. *arXiv preprint arXiv:2209.05753*, 2022.

[68] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), June 2022.

[69] Xinyu Yi, Yuxiao Zhou, and Feng Xu. Transpose: Real-time 3d human translation and pose estimation with six inertial sensors. *ACM Transactions on Graphics*, 40(4), 08 2021.

[70] Ri Yu, Hwangpil Park, and Jehee Lee. Human dynamics from monocular video with dynamic camera movements. *ACM Trans. Graph.*, 40(6), 2021.

[71] Ye Yuan and Kris Kitani. Ego-pose estimation and forecasting as real-time pd control. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (*ICCV*), October 2019.

[72] Ye Yuan, Shih-En Wei, Tomas Simon, Kris Kitani, and Jason Saragih. Simpoe: Simulated character control for 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), pages 7159–7169, June 2021.

[73] Mihai Zanfir, Andrei Zanfir, Eduard Gabriel Bazavan, William T. Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Thundr: Transformer-based 3d human reconstruction with markers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (*ICCV*), pages 12971–12980, October 2021.

[74] Hongwen Zhang, Yating Tian, Yuxiang Zhang, Mengcheng Li, Liang An, Zhenan Sun, and Yebin Liu. Pymaf-x: Towards well-aligned full-body model regression from monocular images. *arXiv preprint arXiv:2207.06400*, 2022.

[75] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *arXiv preprint arXiv:2208.15001*, 2022.

[76] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), June 2019.