mmPred: Radar-based Human Motion Prediction in the Dark

Junqiao Fan[1], Haocong Rao[2], Jiarui Zhang[1], Jianfei Yang[1,3*], Lihua Xie [1†]

[1]School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore
[2] College of Computing and Data Science, Nanyang Technological University, Singapore
[3]School of Mechanical and Aerospace Engineering, Nanyang Technological University, Singapore
{fanj0019, haocong001, zhan0618}@e.ntu.edu.sg, {jianfei.yang, elhxie}@ntu.edu.sg

Abstract

Existing Human Motion Prediction (HMP) methods based on RGB-D cameras are sensitive to lighting conditions and raise privacy concerns, limiting their real-world applications such as firefighting and healthcare. Motivated by the robustness and privacy-preserving nature of millimeter-wave (mmWave) radar, this work introduces radar as a novel sensing modality for HMP, for the first time. Nevertheless, radar signals often suffer from specular reflections and multipath effects, resulting in noisy and temporally inconsistent measurements, such as body-part miss-detection. To address these radar-specific artifacts, we propose mmPred, the first diffusion-based framework tailored for radar-based HMP. mmPred introduces a dual-domain historical motion representation to guide the generation process, combining a Time-domain Pose Refinement (TPR) branch for learning fine-grained details and a Frequency-domain Dominant Motion (FDM) branch for capturing global motion trends and suppressing frame-level inconsistency. Furthermore, we design a Global Skeleton-relational Transformer (GST) as the diffusion backbone to model global inter-joint cooperation, enabling corrupted joints to dynamically aggregate information from others. Extensive experiments show that mmPred achieves state-of-the-art performance, outperforming existing methods by 8.6% on mmBody and 22% on mm-Fi. The source code is available at our project page: https://fanjunqiao.github.io/mmPred-site/.

Introduction

Human Motion Prediction (HMP) aims to predict future human pose sequences from observed historical pose sequences, which plays an essential role in various applications such as human-robot interaction (Gui et al. 2018), healthcare (Troje 2002), and hazard prevention (Yan et al. 2017). Existing HMP works (Li et al. 2022; Chen et al. 2023b) heavily rely on high-precision historical human pose sequences collected from multi-view RGBD MoCap systems, which are typically expensive and impractical for most real-world deployments. Although some works (Chao et al. 2017; Wu and Koike 2018) attempt to use single-view RGB images, these methods struggle under adverse environmental conditions (e.g., darkness, occlusion). Moreover, RGB-based approaches raise privacy concerns in sensitive scenarios such as elderly care.

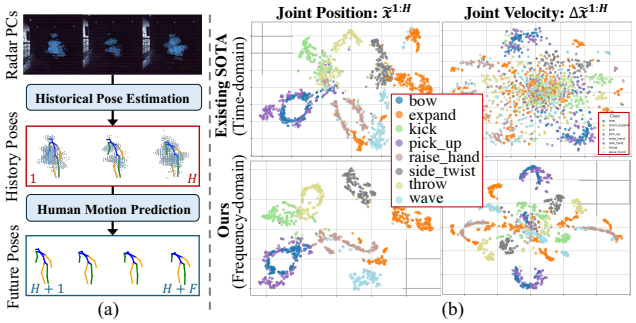

Figure 1: (a) mmWave radar-based HMP in the dark. (b) t-SNE visualization of joint positions and velocities from predicted historical poses. Our frequency-domain prediction produces more distinguishable velocity patterns than the state-of-the-art (SOTA) pose estimator (Yang et al. 2023).

Millimeter-wave (mmWave) radar has emerged as a promising alternative for human perception due to its low cost, portability, and scalable deployment (Waldschmidt, Hasch, and Menzel 2021). As mmWave radar operates in the 30–300 GHz frequency range, its signal can penetrate visual obscurants such as smoke and function reliably regardless of lighting conditions (Zhang et al. 2023). Additionally, its limited spatial resolution naturally preserves privacy, making it suitable for HMP in indoor environments. However, radar point clouds (PCs) are noisy due to "ghost" points caused by multipath effects (Sun et al. 2021). Moreover, intermittent miss-detections often occur due to specular reflections, where only certain body parts reflect signals back to the receiver while others deflect them away (Ding et al. 2024). These issues make it challenging to capture fine-grained and realistic motion from temporally inconsistent and noisy radar PCs.

To generate realistic and stable future pose sequences from radar PCs, a straightforward approach is to first estimate historical human pose sequences using an existing radar-based pose estimator, and then pass the estimated sequences into an HMP model. However, this pipeline is difficult to succeed in practice. Current radar-based pose estimators (Chen et al. 2022; Yang et al. 2023) often produce pose sequences with severe jitter and twisted body structures, since they pro-

*J. Yang is the project lead.
†L. Xie is the corresponding author.

cess each frame independently and overlook the temporal inconsistency caused by miss-detections and multipath effect (Fan et al. 2024). These jittery and unrealistic poses fail to preserve key motion cues, such as joint velocities (as shown in Figure 1b), which are critical for predicting future motion. As a result, HMP models may struggle to capture meaningful temporal patterns from these radar-estimated pose sequences. Moreover, existing HMP methods (Martinez, Black, and Romero 2017; Zhong et al. 2022; Li et al. 2022; Chen et al. 2023b) are over-sensitive to the noise of input historical pose sequences. They commonly assume clean and stable MoCap history, easily producing unrealistic future sequences under radar-specific artifacts, including ghost points and temporal inconsistency due to miss-detections.

Recently, diffusion models (DMs) have shown great potential in generating realistic human motion (Tevet et al. 2022), through learning motion distribution and progressively removing noise. Inspired by such capability, we propose mmPred, a diffusion-based HMP framework tailored for mmWave radar PCs. We first introduce a dual-domain historical motion representation as complementary guidance for the diffusion model. It captures both fine-grained pose details and dominant motion trends from noisy and temporally inconsistent radar data. The Time-domain Pose Refinement (TPR) branch first estimates frame-wise human poses from radar PCs in a coarse-to-fine manner. To mitigate the radar-specific temporal inconsistency, the Frequency-domain Dominant Motion (FDM) branch further aggregates all historical frames to extract dominant motion patterns in the frequency domain. By separating low-frequency motion trends from high-frequency noise, frequency-domain analysis enables the model to capture more accurate velocity patterns from noisy radar PCs, as presented in Figure 1. We then pass this guidance to a frequency diffusion model for generating future pose sequences. To further mitigate the impact of radar-specific artifacts that may lead to unreliable guidance, we introduce the Global Skeleton-relational Transformer (GST) as the backbone of our diffusion model to enhance the realism of generated motions. Specifically, GST employs a skeleton transformer and a frequency transformer to model joint-wise dependencies and temporal motion patterns, respectively. In particular, the skeleton transformer enables more robust feature extraction by modeling global inter-joint cooperation, allowing unreliable joints to aggregate information from structurally or functionally related ones. In summary, our contributions can be summarized as follows:

- We propose mmPred, a novel diffusion-based HMP framework tailored for noisy and temporally inconsistent mmWave radar point clouds. To the best of our knowledge, mmPred is the first radar-based HMP framework.
- We introduce a dual-domain historical motion representation that serves as guidance for the diffusion model. It consists of a Time-domain Pose Refinement (TPR) branch to preserve fine-grained pose details, and a Frequency-domain Dominant Motion (FDM) branch to capture global motion trends while suppressing radar-specific artifacts.
- We design a Global Skeleton-relational Transformer (GST) as the diffusion backbone, which separately models global inter-joint cooperation and motion patterns across frequency components, further enhancing motion realism under challenging radar artifacts.
- Extensive experiments demonstrate that mmPred achieves state-of-the-art performance, outperforming existing methods by 8.6% on mmBody and 22% on mm-Fi.

Related Work

Radar-based Human Motion Sensing. RGB cameras have been widely used for fine-grained human sensing tasks, including 2D and 3D human pose estimation (HPE)(Cao et al. 2017; Sun et al. 2019; Kocabas, Athanasiou, and Black 2020). However, these methods typically suffer from a severe performance drop under adverse environments (Chen et al. 2023a). Overcoming such limitations, commercial mmWave radar has emerged as a good alternative for human sensing. Early radar-based human sensing focused on coarse-level motion tracking (Zhao et al. 2019). More recently, mmWave radar has been explored for fine-grained HPE using radar PCs (Xue et al. 2021; An and Ogras 2021, 2022) or using raw radar signals (Lee et al. 2023; Rahman et al. 2024). Particularly, Xue et al. (Xue et al. 2021) propose an LSTM model, and Chen et al. (Chen et al. 2022) and Yang et al. (Yang et al. 2023) propose transformer-based benchmarks based on their datasets. Ding et al. (Ding et al. 2024) further associate radar points with joint-level motion flow to enhance downstream human sensing tasks. Despite these advances, existing radar-based methods have not yet addressed HMP, which requires stronger modeling of spatial-temporal dynamics. These methods are typically designed in the time domain, making them vulnerable to intermittent miss-detections and failing in long-term motion understanding.

Human Motion Prediction Traditional HMP methods commonly employ RNNs (Martinez, Black, and Romero 2017; Tang et al. 2018) and GCNs (Li et al. 2022; Cui and Sun 2021; Zhong et al. 2022) to deterministically predict a single future, achieving reasonable accuracy but failing to capture the inherent uncertainty and multi-modality of human motion. To address this, probabilistic approaches such as DLow (Yuan and Kitani 2020) and DivSamp (Dang et al. 2022) introduce VAE-based frameworks, though often at the cost of generating overly diverse predictions. Recent diffusion-based methods offer a better trade-off between realism and diversity, achieving state-of-the-art performance by modeling realistic motion distributions. For example, BeliFusion (Barquero, Escalera, and Palmero 2023) operates in the latent space through a multi-stage training scheme, whereas HumanMAC (Chen et al. 2023b) adopts single-stage designs, formulating HMP as a motion inpainting task based on historical observations. Nevertheless, these HMP methods typically assume access to ideal historical poses (i.e., MoCap annotation). As a result, they are sensitive to sensor-estimated histories, especially when using radar, which provides sparse and noisy PCs that severely degrade prediction.

Preliminary

Frequency-domain Motion Representation. The *Discrete Cosine Transform (DCT)*, a widely-used signal process-

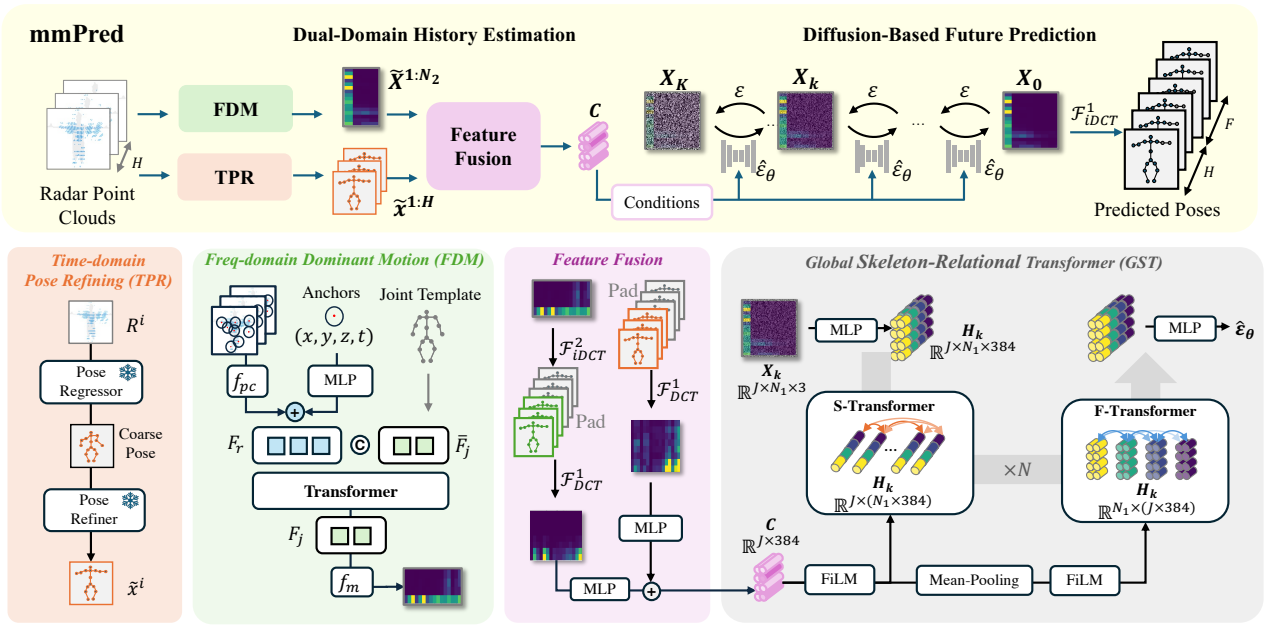

Figure 2: System architecture: mmPred employs Dual-Domain History Estimation to extract historical motion representations in both the time (TPR) and frequency domains (FDM). These representations are fused via a feature fusion module to construct the condition embedding C, which guides the diffusion-based future motion prediction performed in the frequency domain.

ing tool, can transform time-domain human pose sequences into frequency-domain coefficients. Let $\mathbf{x}^{1:T} \in \mathbb{R}^{T\times 3J}$ be a time-domain pose sequence of (T) frames with (J) joints, and let $\mathbf{X}^{1:N} \in \mathbb{R}^{N\times 3J}$ denote the truncated DCT representation ($N \leq T$), as an output of DCT. The type-II orthonormal DCT is defined as:

$$\mathbf{X}^{1:N} = \mathcal{F}_{DCT}(\mathbf{x}^{1:T}) = \mathbf{D}\mathbf{x}^{1:T}, \tag{1}$$

The inverse DCT reconstructs the time-domain pose sequences $\hat{\mathbf{x}}^{1:T}$ via:

$$\hat{\mathbf{x}}^{1:T} = \mathcal{F}_{iDCT}(\mathbf{X}^{1:N}) = \mathbf{D}^{\top}\mathbf{X}^{1:N}. \tag{2}$$

In the frequency domain, dominant motion trends (e.g., mean pose and velocity) are concentrated in low-frequency components, while high-frequency components capture fine-grained details like acceleration. This spectral separation makes DCT well-suited for motion analysis under radar-specific noise.

Frequency-domain DM for HMP. DM in the frequency domain directly learn the data distribution of the frequency-domain motion representation. Given the ground-truth (GT) time-domain pose sequence $\mathbf{x}^{1:(H+F)} \in \mathbb{R}^{(H+F)\times 3J}$ (with both history and future), we first apply the DCT transformation $\mathcal{F}^1_{DCT}$ using basis $\mathbf{D}_1 \in \mathbb{R}^{N_1\times(H+F)}$ to obtain its frequency-domain representation $X^{1:N_1}$. Then, the diffusion model consists of two processes: a forward process that gradually adds noise and a reverse process that learns to invert. In the forward process, we start from the frequency-domain GT $X_0 = X^{1:N_1}$. For each diffusion step $k \in [0, .., K]$, we iteratively sample a noisier motion X_k by adding Gaussian noise $\varepsilon \in \mathcal{N}(0, I)$:

$$q\left(\mathbf{X}_k \mid \mathbf{X}_{k-1}\right) = \mathcal{N}\left(\mathbf{X}_k \mid \sqrt{1-\beta_k}\mathbf{X}_{k-1}, \beta_k\mathbf{I}\right), \tag{3}$$

where β_k is the scheduled noise scale. On the contrary, the reverse process starts from a noisy Gaussian initialization $\hat{\mathbf{X}}_K \in N(0, I)$ and progressively removes noise until $\hat{\mathbf{X}}_0$ is generated. A diffusion model, denoted as $\hat{\varepsilon}_\theta$, is trained to remove the motion noise for cleaner motion $\hat{\mathbf{X}}_{k-1}$:

$$\hat{\mathbf{X}}_{k-1} = (1-\beta_k)^{-1}(\hat{\mathbf{X}}_k - \beta_k\hat{\varepsilon}_\theta(\hat{\mathbf{X}}_k, k, C)). \tag{4}$$

To ensure the predicted future motion follows historical human behavior, we leverage a condition C capturing historical motion information as diffusion guidance. Finally, the output $\hat{X}_0$ from the reverse process can be transformed back to the time domain via $\mathcal{F}^1_{iDCT}$, serving as the predicted future.

Methodology

Problem Formulation

Our proposed mmWave radar-based HMP task inputs historical H frames of radar PCs and predicts future F frames of human poses. We denote the radar PCs as $R^{1:H} = \{R^i \in \mathbb{R}^{N\times 6}\}_{i=1}^{H}$, which are robust inputs under adverse environments. Each historical frame contains N points, with three Cartesian coordinates and three attributes $\{v, E, A\}$ representing velocity, energy, and amplitude. The predicted future pose sequences are denoted as $\hat{x}^{H+1:H+F} = \{\hat{x}^i \in \mathbb{R}^{J\times 3}\}_{i=H+1}^{H+F}$, where J denotes the number of body joints. Different from the existing 3D HMP task, which takes in historical ground-truth (GT) human poses $x^{1:H} = \{x^i\}_{i=1}^{H}$, our task can only estimate historical poses from radar data, which is denoted as $\tilde{x}^{1:H} = \{\tilde{x}^i\}_{i=1}^{H}$. Given $R^{1:H}$ contains substantial noise, our task is fundamentally more challenging.

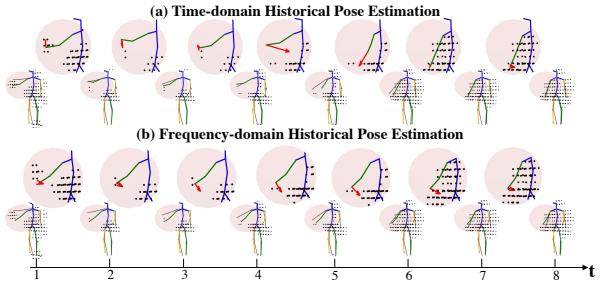

Figure 3: Comparison of estimated motion history (colored) to GT (black) in dual domain. We zoom in on the right-hand area and use red arrows to mark the joint velocity. TPR demonstrates more accurate pose location, while FDM offers more consistent velocity information.

Dual-Domain History Estimation

Accurate guidance for diffusion-based motion prediction relies on robust historical motion estimation—requiring both precise joint localization and a coherent motion trend. As shown in Figure 3, time-domain pose estimation is easily affected by temporally inconsistent signals. To address this, we propose a dual-domain historical motion estimation approach, estimating accurate joint locations in the time domain and capturing dominant motion trends in the frequency domain.

Time-domain Pose Refining (TPR). For each frame $i \in [1, \dots, H]$, TPR employs a shared pretrained pose estimator (Chen et al. 2022; Yang et al. 2023), denoted as f_{pose}, to predict 3D human poses from the single-frame radar point cloud R^i. However, occasional joint miss-detections may cause certain joints to deviate significantly from their true positions, resulting in twisted pose structures or jitter across frames. To mitigate this, we employ a pretrained diffusion-based refinement network (Fan et al. 2024), denoted as f_{refine}. It leverages prior knowledge such as limb-length consistency and adjacent-frame continuity to improve structural plausibility and temporal stability. Given the coarse output from f_{pose}, it produces refined pose estimates $\tilde{x}^i_{\text{time}}$ for each frame. Finally, we aggregate the refined poses from all H frames as the TPR output:

$$\begin{aligned} \tilde{x}^i_{\text{time}} &= f_{\text{refine}}(f_{\text{pose}}(R^i)) \\ \tilde{x}^{1:H}_{\text{time}} &= \{\tilde{x}^1_{\text{time}}, \tilde{x}^2_{\text{time}}, ..., \tilde{x}^H_{\text{time}}\}^H_{i=1} \end{aligned} \tag{5}$$

Frequency-domain Dominant Motion (FDM). FDM directly outputs a compact frequency-domain motion representation, treating the entire historical motion sequence holistically. It takes all $R^{1:H}$ as input and predicts the frequency-domain motion representation, $\tilde{X}^{1:N_2}_{\text{freq}} \in \mathbb{R}^{N_2 \times J \times 3}$, where N_2 denotes the number of selected coefficients (typically $N_2 = 3$ or 4). These few coefficients capture the most dominant motion dynamics, allowing the model to focus on a consistent motion trend without sudden changes.

As shown in Figure 2, we first capture the global evolution of PCs into a feature representation. The $R^{1:H}$ is encoded by anchor-based PC encoder f_{pc} (Zhao et al. 2021). Specifically, for each frame, we select N' representative anchor points using Farthest Point Sampling (FPS). The selected anchors encode local PC features from the neighborhood, aggregated into anchor features $F_r \in \mathbb{R}^{(H \times N') \times 1024}$. To encode the evolution of PCs over time, we inject temporal information using an MLP that maps each anchor's (x, y, z, t) (where $t \in [1, H]$) to a positional embedding, which is then added to the corresponding anchor features. Then, we devise a transformer-based architecture to perform frequency-domain motion learning. A transformer Φ is applied to dynamically project time-domain anchor features into frequency-domain joint motion features. We first zero-initialize a set of learnable joint template tokens $\bar{F}_j \in \mathbb{R}^{J \times 1024}$, one for each body joint. These are concatenated with the anchor features and jointly processed by Φ:

$$\begin{aligned} F_r &= f_{\text{pc}}(R^{1:H}), \\ F'_r, F_j &= \Phi(F_r, \bar{F}_j). \end{aligned} \tag{6}$$

Within the transformer, deep correlation is captured, both within F_r and between F_j and F_r. The F_r captures the temporal PC evolution through anchor-wise self-attention. Each joint feature in F_j also dynamically aggregates information from correlated anchor feature cues. The captured joint features F_j are eventually decoded by an MLP-based motion decoder f_m to produce the frequency-domain representation:

$$\tilde{X}^{1:N_2}_{\text{freq}} = f_m(F_j). \tag{7}$$

Cross-Domain History Fusion. To guide the diffusion model, we fuse the dual-domain $\tilde{x}^{1:H}_{time}$ and $\tilde{X}^{1:N_2}_{freq}$ into a conditional embedding C. Since the diffusion model operates in the frequency domain, we transform both inputs into frequency domain representations, $\tilde{X}^{1:N_1}_{\text{time}}$ and $\tilde{X}^{1:N_1}_{\text{freq}}$, with N_1 DCT coefficients corresponding to $H + F$ frames. Specifically, both representations are first converted to the time domain to perform repeated padding up to $H + F$ frames, and then transformed back into the frequency domain:

$$\tilde{X}^{1:N_1}_{\text{time}} = \mathcal{F}^1_{DCT}(\text{pad}(\tilde{x}^{1:H}_{\text{time}})), \tag{8}$$

$$\tilde{X}^{1:N_1}_{\text{freq}} = \mathcal{F}^1_{DCT}\left(\text{pad}\left(\mathcal{F}^2_{iDCT}(\tilde{X}^{1:N_2}_{\text{freq}})\right)\right). \tag{9}$$

Both representations are then reshaped into $\mathbb{R}^{J \times (N_1 \times 3)}$, isolating each joint to preserve its distinct motion dynamics and prevent interference from unreliable keypoints caused by radar miss-detections. Then, the reshaped representations are independently projected via two MLPs f_1 and f_2, and fused by element-wise addition to obtain the final joint-wise conditional embedding $C \in \mathbb{R}^{J \times 384}$:

$$C = f_1(\tilde{X}^{1:N_1}_{\text{time}}) + f_2(\tilde{X}^{1:N_1}_{\text{freq}}). \tag{10}$$

Global Skeleton-Relational Transformer (GST)

Existing transformer-based diffusion models typically overlook global cooperation among joints by encoding all joints into a unified representation, making them vulnerable to joint miss-detections. To address this, GST first isolates joint-specific features to preserve individual motion dynamics, then applies joint-wise self-attention to enable global

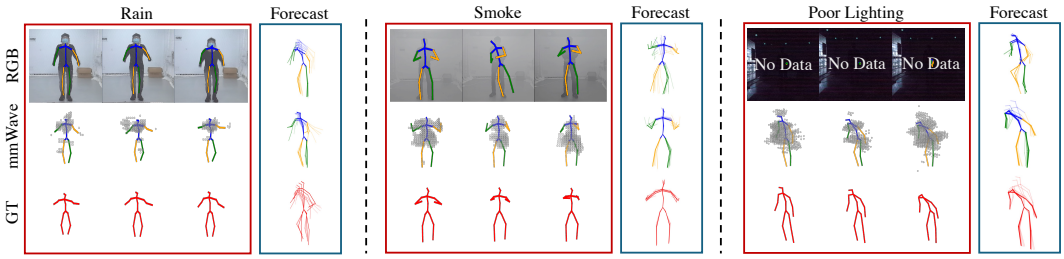

Figure 4: HMP under adverse environments using RGB, mmWave (ours), and MoCap GT. The red boxes enclose estimated historical poses, and the blue boxes enclose the predicted future poses (predicted frames are stacked).

information exchange. As shown in Figure 2, the noisy motion of intermediate diffusion step k, $X_k \in \mathbb{R}^{N_1 \times J \times 3}$, is first projected to latent features $H_k \in \mathbb{R}^{N_1 \times J \times 384}$ by an MLP. This yields $N_1 \times J$ tokens, one for each joint at each frequency coefficient. For spatial modeling, we reshape H_k into $\mathbb{R}^{J \times (N_1 \times 384)}$, aggregating all frequency components into joint-specific tokens. These J tokens are passed through a Skeleton Transformer (S-Transformer), where self-attention enables joint-wise interaction. Potentially, mis-detected joints recover meaningful representations by leveraging information from other observed joints. Next, for temporal modeling, H_k is reshaped into $\mathbb{R}^{N_1 \times (J \times C_a)}$, forming N_1 tokens, each aggregating information across all joints at a given frequency. These tokens are processed by a Frequency Transformer (F-Transformer), which applies self-attention along the temporal dimension to ensure the generated motion sequence is temporally smooth and realistic.

To inject the conditional C, we adopt the Feature-wise Linear Modulation (FiLM) strategy (details in Appendix). For the S-Transformer, we maintain joint-wise conditioning by directly inputting $C \in \mathbb{R}^{J \times 384}$, allowing each joint to receive individualized control. For the F-Transformer, which processes joint-aggregated tokens, we apply mean-pooling over the joint dimension to obtain a global condition $C' \in \mathbb{R}^{1 \times 384}$ before applying FiLM.

Overall Learning Objective

The Dual-Domain Historical Motion Estimation and the Diffusion-based Future Prediction are trained in two separate stages. Firstly, the FDM module is trained using an L_2 loss with the GT frequency-domain motion as supervision:

$$\mathcal{L}_1 = ||\tilde{X}^{1:N_2} - \mathcal{F}^2_{DCT}(x^{1:H})||^2. \quad (11)$$

Subsequently, with the pretrained TPR and FDM, we train the GST-based diffusion model using the ε-prediction objective from Denoising Diffusion Probabilistic Models (DDPM) (Ho, Jain, and Abbeel 2020), which minimizes the error between the predicted and true noise added at each diffusion timestep.

$$\mathcal{L}_2 = ||\hat{\varepsilon}_\theta(X_k, k, C) - \varepsilon_k||^2. \quad (12)$$

Experiments

Experimental Setup

Datasets. Our experiments are conducted on mmBody (Yang et al. 2023) and mm-Fi (Yang et al. 2023) datasets. Specifically, mmBody collects training data under normal scenes (Lab1 and Lab2) and testing data under adverse environments. It contains MoCap-annotated poses and multi-modal sensor data from RGB-D and Phenix mmWave radar. This setting enables robustness comparison of different modalities. We select $H = 8$ frames (0.5s) for history and $F = 16$ frames (1s) for future prediction. Mm-Fi is a larger-scale human motion dataset using a lower-bandwidth mmWave radar with sparser PCs. The human poses are annotated from RGB images using the pretrained HRNet-w48 (Sun et al. 2019). We adopt a random train-test split following the daily-activity protocol. Due to distinct frame rates, we select $H = 5$ frames (0.5s) for history and $F = 10$ frames (1s) for future prediction.

Competing Methods. To adapt to raw sensor data input, we adopt a two-stage protocol for all evaluated methods: historical pose estimation followed by future motion prediction. During training and testing, HMP methods accepted historical poses generated by sensor-specific pretrained pose estimators. The output of HMP models is compared with the GT annotations. This setting reflects realistic deployment scenarios, where only sensor data is available. We compare our method with two representative SOTA open-source HMP models: a graph-based model, PSGSN (Li et al. 2022), and a diffusion-based model, HumanMAC (Chen et al. 2023b). For the pose estimators, we follow the benchmark practices on different datasets. On mmBody, we use VIBE(Kocabas, Athanasiou, and Black 2020) for RGB video and P4Transformer(Fan, Yang, and Kankanhalli 2021) for radar PCs. On mm-Fi, we adopt PointTransformer (Zhao et al. 2021) for radar PCs. The estimator-generated pose sequences are then used as input to train and evaluate all HMP models.

Evaluation Metric. Following previous HMP works (Chen et al. 2023b), we adopt two metrics designed for future motion accuracy: (1) Average Displacement Error (ADE), the mean L2 distance between predictions and GT over all future frames; and (2) Final Displacement Error (FDE), the L2 distance at the final frame. Following the best-of-K evaluation for diffusion-based methods, we generate 10 hypotheses and report the best performance. Full assessments of model diversity and multimodal accuracy are presented in the Appendix.

Overall Result

Comparison with RGB-based HMP. As presented in Table 1. Due to sensor noise and inaccuracies in historical pose

Methods		Lab1		Lab2		Furnished		Rain		Smoke		Dark		Occlusion		Avg.	
		ADE↓	FDE↓	ADE↓	FDE↓	ADE↓	FDE↓	ADE↓	FDE↓	ADE↓	FDE↓	ADE↓	FDE↓	ADE↓	FDE↓	ADE↓	FDE↓
GT	HumanMAC	0.235	0.329	0.242	0.344	0.278	0.370	0.297	0.380	0.338	0.462	0.287	0.394	0.265	0.358	0.291	0.392
RGB	HumanMAC	0.390	0.452	0.427	0.516	0.420	0.482	0.479	0.501	0.560	0.589	0.693	0.612	0.739	0.627	0.547	0.550
mmWave	PSGSN	0.503	0.597	0.526	0.640	0.524	0.626	0.536	0.580	0.598	0.684	0.513	0.600	0.485	0.566	0.537	0.617
	HumanMAC	0.411	0.458	0.441	0.505	0.450	0.478	0.455	0.465	0.496	0.540	0.406	0.447	0.391	0.430	0.460	0.487
	mmPred	**0.369**	**0.412**	**0.387**	**0.441**	**0.418**	**0.439**	**0.436**	**0.444**	**0.472**	**0.522**	**0.392**	**0.436**	**0.378**	**0.416**	**0.420**	**0.456**

Table 1: Quantitative results on the mmBody under different adversarial environments, evaluated using ADE and FDE metrics. We compare different input modalities for historical motion, including RGB and mmWave radar. GT-based methods are provided to indicate the upper-bound performance. Bold indicates the best performance among sensor-derived methods.

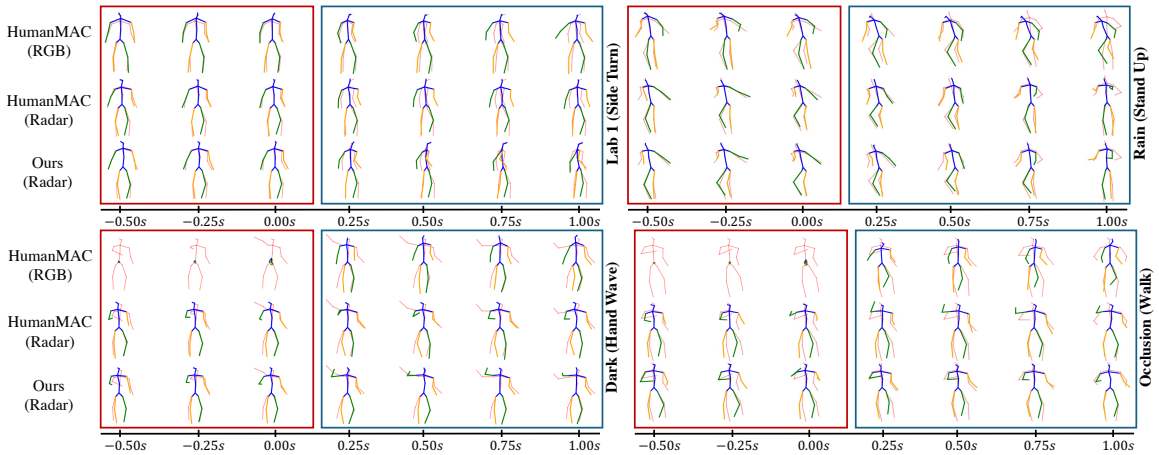

Figure 5: Qualitative visualization of our method compared to baseline on the mmBody dataset, where red poses denote GT and colored poses denote our prediction. Red boxes enclose the 0.5s historical frames, and blue boxes enclose the 1s predicted future. The higher overlap between predicted and gt poses indicates higher accuracy.

estimation, methods based on raw sensor input generally underperform those using GT history (upper-bound performance). However, mmWave-based HMP methods demonstrate superior robustness across all adverse environments. Notably, mmPred reduces ADE and FDE by over 9% and 11.4% under rain and smoke, and improves ADE by more than 40% under darkness and occlusion. This is largely because RGB-based methods fail in low-light or occluded scenarios, often producing random outputs. As visualized in Figure 4, RGB inputs degrade severely under challenging conditions, leading to erroneous historical poses and inaccurate predictions, whereas mmWave radar maintains reliable sensing, enabling accurate motion forecasting.

Performance on mmBody. Compared to existing methods leveraging radar PCs, mmPred achieves significantly better performance. As shown in Figure 5, radar estimated pose sequences often exhibit jitter and erroneous velocities, while prior approaches are highly sensitive to these radar-specific noises. Traditional GCN-based encoder-decoder models like PSGSN struggle to handle such noise. While diffusion-based methods like HumanMAC offer improvements, they remain vulnerable to corrupted historical conditions. In contrast, mmPred outperforms HumanMAC by an average of 8.6% in ADE and 6.4% in FDE across various adverse conditions. This improvement is largely attributed to the proposed FDM and the GST-based diffusion architecture. FDM captures more stable motion trends and velocity cues, while GST mitigates the impact of unreliable joints through joint-wise cooperation. As illustrated in Figure 5, baseline methods are easily misled by jitter in historical poses, particularly for hands, resulting in incorrect motion direction and velocity estimation.

Performance on mm-Fi. As shown in Table 2, we further evaluate the generalizability of mmPred on the mm-Fi dataset, typically with sparse PCs and more severe pose jitter in history. Under these challenges, mmPred achieves a more significant improvement over existing methods. Specifically, it reduces ADE and FDE by 22% and 23%, respectively, compared to the SOTA HumanMAC. These results highlight that existing diffusion-based HMP methods are sensitive to historical noise, whereas mmPred remains robust through frequency-domain motion modeling. Qualitatively, as presented in Figure 6, HumanMAC produces inaccurate predictions, while mmPred yields trajectories that better align with the GT. In motions like side twist and pickup, pose jitter disrupts slight motion trends, making them harder to predict. FDM mitigates this by extracting global cues from radar flow, offering more reliable temporal guidance. For larger motions such as throwing and kicking, GST enhances joint-wise coordination, yielding more anatomically consistent future poses.

Method		Chest Expand ADE↓	Chest Expand FDE↓	Side Twist ADE↓	Side Twist FDE↓	Raise Hand ADE↓	Raise Hand FDE↓	Pickup ADE↓	Pickup FDE↓	Throwing ADE↓	Throwing FDE↓	Kicking ADE↓	Kicking FDE↓	Bowing ADE↓	Bowing FDE↓	Avg. ADE↓	Avg. FDE↓
GT	HumanMAC	0.300	0.256	0.277	0.276	0.244	0.226	0.414	0.486	0.374	0.393	0.337	0.366	0.251	0.256	0.293	0.293
mmWave	PSGSN	0.402	0.422	0.430	0.445	0.397	0.446	0.595	0.686	0.449	0.517	0.426	0.449	0.336	0.395	0.430	0.470
	HumanMAC	0.353	0.306	0.407	0.392	0.363	0.333	0.547	0.578	0.439	0.471	0.425	0.421	0.304	0.314	0.408	0.396
	mmPred	**0.272**	**0.230**	**0.315**	**0.308**	**0.237**	**0.218**	**0.452**	**0.481**	**0.371**	**0.370**	**0.374**	**0.368**	**0.246**	**0.234**	**0.319**	**0.305**

Table 2: Quantitative results on mm-Fi for different actions, evaluated by ADE and FDE. Noted that mm-Fi adopts GT from RGB images using the pretrained HRNet-w48 (Sun et al. 2019). Bold indicates the best performance among sensor-driven methods.

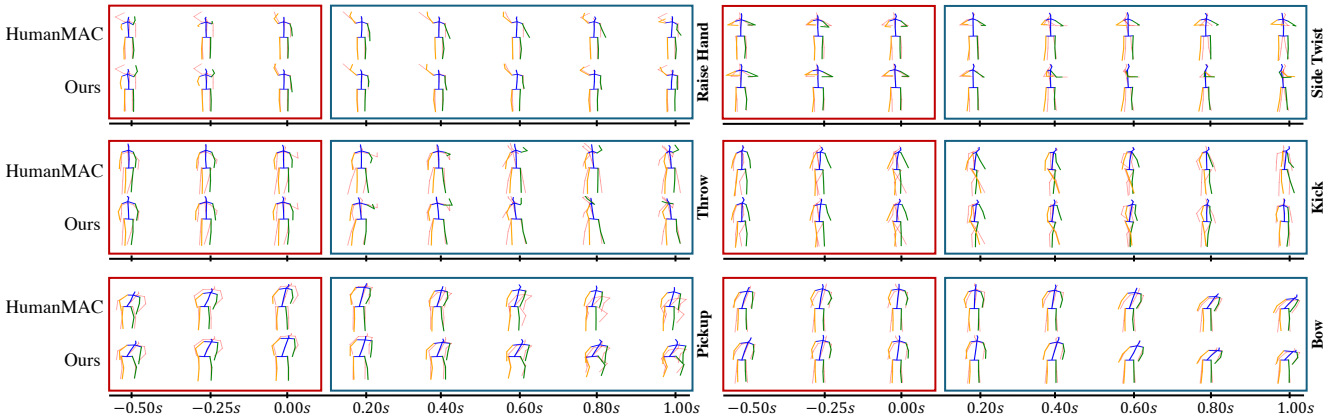

Figure 6: Qualitative visualization of our method compared to baseline on the mm-Fi dataset.

Ablation Study

In Table 3, we present ablation studies on the three proposed modules in mmPred: TPR, FDM, and GST. All modules contribute to reducing future prediction errors, with FDM showing particularly strong gains on the mm-Fi dataset, highlighting its effectiveness under extreme sparsity. Additionally, experiments M5–M7 demonstrate performance drops when removing any single module, confirming their necessity for achieving optimal results. Further ablations on the S-Transformer and feature isolation design are reported in Table 4. Following (Curreli et al. 2025), we adopt two metrics to quantify motion realism: (1) limb-length error, the normalized L_1 distance between predicted and ground-truth limb lengths; and (2) limb-length jitter, the normalized L_1 distance between predicted limb lengths across consecutive frames. Results show that the S-Transformer significantly reduces both metrics, while the feature isolation design further enhances accuracy by preventing occasional keypoint misdetections from corrupting the unified feature representation.

Methods	Modules TPR	Modules FDM	Modules GST	mmBody ADE↓	mmBody FDE↓	mm-Fi ADE↓	mm-Fi FDE↓
M1				0.460	0.487	0.408	0.396
M2	✓			0.455	0.485	0.379	0.359
M3		✓		0.456	0.486	0.337	0.326
M4			✓	0.460	0.485	0.373	0.354
M5	✓	✓		0.448	0.476	0.355	0.327
M6		✓	✓	0.423	0.458	0.325	0.310
M7	✓		✓	0.439	0.474	0.334	0.324
M8	✓	✓	✓	**0.420**	**0.456**	**0.319**	**0.305**

Table 3: Ablation studies of proposed modules on mmBody and mm-Fi. Specifically, M2, M3, and M4 show the efficiency of TPR, FDM, and GST, respectively. M5, M6, and M7 explore the necessity of GST, TPR, and FDM, respectively.

GST	Realism Error↓	Realism Jitter↓
w/o S-Tran.	10.67	7.01
w/ S-Trans.	**9.92**	**6.09**

Feat. Isolation	Accuracy ADE↓	Accuracy FDE↓
no	0.438	0.470
yes	**0.420**	**0.456**

Table 4: Ablation study of the GST module. **Left:** Effect of the S-Transformer on realism metrics, including limb-length error and limb-length jitter. **Right:** Effect of GST's joint-wise isolation design on prediction accuracy.

Conclusion

This paper presents mmPred, the first diffusion-based HMP framework tailored for the mmWave radar modality. To address the challenges posed by noisy and unstable radar signals, we propose a dual-domain historical motion representation as the diffusion condition. Furthermore, the Global Skeleton Transformer backbone enhances motion realism by modeling joint-wise correlations. Extensive experiments demonstrate that mmPred achieves superior robustness and accuracy under adverse conditions compared to existing methods. In the future, cross-domain generalization and the incorporation of raw radar signals warrant further exploration.

Acknowledgements

This work is supported by Ministry of Education (MOE), Singapore, under AcRF TIER 1 Grant RG64/23 and National Research Foundation of Singapore Medium-sized Centre for Advanced Robotics Technology Innovation. This work is jointly supported by MOE Singapore Tier 1 Grant RG83/25, RS36/24 and a Start-up Grant from Nanyang Technological University.

References

An, S.; and Ogras, U. Y. 2021. Mars: mmwave-based assistive rehabilitation system for smart healthcare. *ACM Transactions on Embedded Computing Systems (TECS)*, 20(5s): 1–22.

An, S.; and Ogras, U. Y. 2022. Fast and scalable human pose estimation using mmwave point cloud. In *Proceedings of the 59th ACM/IEEE Design Automation Conference*, 889–894.

Barquero, G.; Escalera, S.; and Palmero, C. 2023. Belfusion: Latent diffusion for behavior-driven human motion prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2317–2327.

Cao, Z.; Simon, T.; Wei, S.-E.; and Sheikh, Y. 2017. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7291–7299.

Chao, Y.-W.; Yang, J.; Price, B.; Cohen, S.; and Deng, J. 2017. Forecasting human dynamics from static images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 548–556.

Chen, A.; Wang, X.; Zhu, S.; Li, Y.; Chen, J.; and Ye, Q. 2022. mmBody benchmark: 3D body reconstruction dataset and analysis for millimeter wave radar. In *Proceedings of the 30th ACM International Conference on Multimedia*, 3501–3510.

Chen, H.; Feng, R.; Wu, S.; Xu, H.; Zhou, F.; and Liu, Z. 2023a. 2D Human pose estimation: A survey. *Multimedia Systems*, 29(5): 3115–3138.

Chen, L.-H.; Zhang, J.; Li, Y.; Pang, Y.; Xia, X.; and Liu, T. 2023b. Humanmac: Masked motion completion for human motion prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9544–9555.

Cui, Q.; and Sun, H. 2021. Towards accurate 3d human motion prediction from incomplete observations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4801–4810.

Curreli, C.; Muhle, D.; Saroha, A.; Ye, Z.; Marin, R.; and Cremers, D. 2025. Nonisotropic Gaussian Diffusion for Realistic 3D Human Motion Prediction. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 1871–1882.

Dang, L.; Nie, Y.; Long, C.; Zhang, Q.; and Li, G. 2022. Diverse human motion prediction via gumbel-softmax sampling from an auxiliary space. In *Proceedings of the 30th ACM international conference on multimedia*, 5162–5171.

Ding, F.; Luo, Z.; Zhao, P.; and Lu, C. X. 2024. milliflow: Scene flow estimation on mmwave radar point cloud for human motion sensing. In *European Conference on Computer Vision*, 202–221. Springer.

Fan, H.; Yang, Y.; and Kankanhalli, M. 2021. Point 4d transformer networks for spatio-temporal modeling in point cloud videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14204–14213.

Fan, J.; Yang, J.; Xu, Y.; and Xie, L. 2024. Diffusion model is a good pose estimator from 3d rf-vision. In *European Conference on Computer Vision*, 1–18. Springer.

Gui, L.-Y.; Zhang, K.; Wang, Y.-X.; Liang, X.; Moura, J. M.; and Veloso, M. 2018. Teaching robots to predict human motion. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 562–567. IEEE.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Ionescu, C.; Papava, D.; Olaru, V.; and Sminchisescu, C. 2013. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7): 1325–1339.

Kocabas, M.; Athanasiou, N.; and Black, M. J. 2020. Vibe: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5253–5263.

Lee, S.-P.; Kini, N. P.; Peng, W.-H.; Ma, C.-W.; and Hwang, J.-N. 2023. Hupr: A benchmark for human pose estimation using millimeter wave radar. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 5715–5724.

Li, M.; Chen, S.; Zhang, Z.; Xie, L.; Tian, Q.; and Zhang, Y. 2022. Skeleton-parted graph scattering networks for 3d human motion prediction. In *European conference on computer vision*, 18–36. Springer.

Martinez, J.; Black, M. J.; and Romero, J. 2017. On human motion prediction using recurrent neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2891–2900.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Prabhakara, A.; Jin, T.; Das, A.; Bhatt, G.; Kumari, L.; Soltanaghai, E.; Bilmes, J.; Kumar, S.; and Rowe, A. 2023. High resolution point clouds from mmwave radar. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 4135–4142. IEEE.

Rahman, M. M.; Yataka, R.; Kato, S.; Wang, P.; Li, P.; Cardace, A.; and Boufounos, P. 2024. MMVR: Millimeter-wave multi-view radar dataset and benchmark for indoor perception. In *European Conference on Computer Vision*, 306–322. Springer.

Saadatnejad, S.; Rasekh, A.; Mofayezi, M.; Medghalchi, Y.; Rajabzadeh, S.; Mordan, T.; and Alahi, A. 2023. A generic diffusion-based approach for 3d human pose prediction in the wild. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 8246–8253. IEEE.

Sun, J.; and Chowdhary, G. 2024. Comusion: Towards consistent stochastic human motion prediction via motion diffusion. In *European Conference on Computer Vision*, 18–36. Springer.

Sun, K.; Xiao, B.; Liu, D.; and Wang, J. 2019. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5693–5703.

Sun, Y.; Huang, Z.; Zhang, H.; Cao, Z.; and Xu, D. 2021. 3DRIMR: 3D reconstruction and imaging via mmWave radar based on deep learning. In *2021 IEEE International Performance, Computing, and Communications Conference (IPCCC)*, 1–8. IEEE.

Tang, Y.; Ma, L.; Liu, W.; and Zheng, W. 2018. Long-term human motion prediction by modeling motion context and enhancing motion dynamic. *arXiv preprint arXiv:1805.02513*.

Tevet, G.; Raab, S.; Gordon, B.; Shafir, Y.; Cohen-Or, D.; and Bermano, A. H. 2022. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*.

Troje, N. F. 2002. Decomposing biological motion: A framework for analysis and synthesis of human gait patterns. *Journal of vision*, 2(5): 2–2.

Waldschmidt, C.; Hasch, J.; and Menzel, W. 2021. Automotive radar—From first efforts to future systems. *IEEE Journal of Microwaves*, 1(1): 135–148.

Wu, E.; and Koike, H. 2018. Real-time human motion forecasting using a RGB camera. In *Proceedings of the 24th ACM Symposium on Virtual Reality Software and Technology*, 1–2.

Xue, H.; Ju, Y.; Miao, C.; Wang, Y.; Wang, S.; Zhang, A.; and Su, L. 2021. mmMesh: Towards 3D real-time dynamic human mesh construction using millimeter-wave. In *Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services*, 269–282.

Yan, X.; Li, H.; Wang, C.; Seo, J.; Zhang, H.; and Wang, H. 2017. Development of ergonomic posture recognition technique based on 2D ordinary camera for construction hazard prevention through view-invariant features in 2D skeleton motion. *Advanced Engineering Informatics*, 34: 152–163.

Yang, J.; Huang, H.; Zhou, Y.; Chen, X.; Xu, Y.; Yuan, S.; Zou, H.; Lu, C. X.; and Xie, L. 2023. MM-Fi: Multi-Modal Non-Intrusive 4D Human Dataset for Versatile Wireless Sensing. *arXiv preprint arXiv:2305.10345*.

Yuan, Y.; and Kitani, K. 2020. Dlow: Diversifying latent flows for diverse human motion prediction. In *European Conference on Computer Vision*, 346–364. Springer.

Zhang, J.; Xi, R.; He, Y.; Sun, Y.; Guo, X.; Wang, W.; Na, X.; Liu, Y.; Shi, Z.; and Gu, T. 2023. A survey of mmWave-based human sensing: Technology, platforms and applications. *IEEE Communications Surveys & Tutorials*.

Zhao, H.; Jiang, L.; Jia, J.; Torr, P. H.; and Koltun, V. 2021. Point transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, 16259–16268.

Zhao, P.; Lu, C. X.; Wang, J.; Chen, C.; Wang, W.; Trigoni, N.; and Markham, A. 2019. mid: Tracking and identifying people with millimeter wave radar. In *2019 15th International Conference on Distributed Computing in Sensor Systems (DCOSS)*, 33–40. IEEE.

Zhong, C.; Hu, L.; Zhang, Z.; Ye, Y.; and Xia, S. 2022. Spatio-temporal gating-adjacency gcn for human motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6447–6456.