Diffusion Model is a Good Pose Estimator from 3D RF-Vision

Junqiao Fan[1], Jianfei Yang[1,2⋆], Yuecong Xu[1], and Lihua Xie[1]

[1] School of Electrical and Electronic Engineering
[2] School of Mechanical and Aerospace Engineering
Nanyang Technological University, Singapore
{fanj0019, jianfei.yang, xuyu0014, elhxie}@ntu.edu.sg

Abstract. Human pose estimation (HPE) from Radio Frequency vision (RF-vision) performs human sensing using RF signals that penetrate obstacles without revealing privacy (e.g., facial information). Recently, mmWave radar has emerged as a promising RF-vision sensor, providing radar point clouds by processing RF signals. However, the mmWave radar has a limited resolution with severe noise, leading to inaccurate and inconsistent human pose estimation. This work proposes mmDiff, a novel diffusion-based pose estimator tailored for noisy radar data. Our approach aims to provide reliable guidance as conditions to diffusion models. Two key challenges are addressed by mmDiff: (1) miss-detection of parts of human bodies, which is addressed by a module that isolates feature extraction from different body parts, and (2) signal inconsistency due to environmental interference, which is tackled by incorporating prior knowledge of body structure and motion. Several modules are designed to achieve these goals, whose features work as the conditions for the subsequent diffusion model, eliminating the miss-detection and instability of HPE based on RF-vision. Extensive experiments demonstrate that mmDiff outperforms existing methods significantly, achieving state-of-the-art performances on public datasets.

1 Introduction

Human pose estimation (HPE) has been a widely-studied computer vision task for predicting coordinates of human keypoints and generating human skeletons [3, 26, 56]. It is a fundamental task for human-centered applications, such as augmented/virtual reality [27, 48, 55], rehabilitation [4, 41], and human-robot interaction [11, 14]. Current HPE solutions mainly rely on RGB(D) cameras [9, 21]. Though demonstrating promising accuracy, camera-based pose estimators have intrinsic limitations under adverse environments, e.g., smoke, low illumination, and occlusion [44]. The privacy issue of cameras also hinders the viability of HPE in medical scenarios, e.g., rehabilitation systems in hospitals [1].

Overcoming limitations of camera-based HPE, Radio-Frequency vision (RF-vision) has attracted surging attention for human sensing. The emerging mmWave

⋆ J. Yang is the project lead.

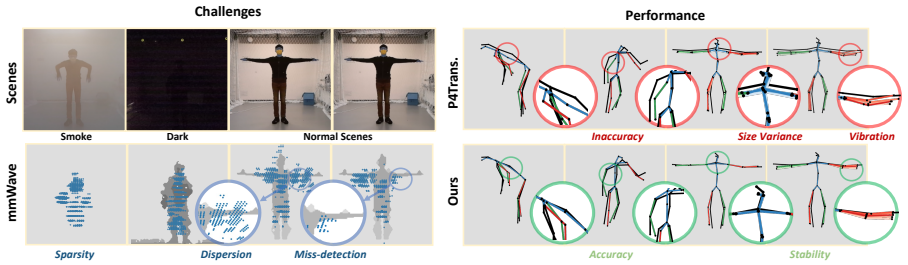


Fig. 1: Left: challenges of mmWave PCs. Right: the performance of existing SOTA (P4Transformer [13]) compared to ours. PC's sparsity and dispersion cause inaccurate spline and shoulder. Inconsistent PCs with occasional miss-detection further cause size variance and pose vibration. mmDiff proposes diffusion-based pose estimation with enhanced accuracy and stability.

radar technology presents a promising and feasible solution for HPE due to its privacy preservation, portability, and energy efficiency [43]. Operated at the frequencies of 30-300 GHz [49], commercial mmWave radars transmit and receive RF signals that penetrate human targets and occlusions. Radar point clouds (PCs) are further extracted as the salient target detection via monitoring signals' characteristic changes [33, 46]. Therefore, the extracted PCs are more robust to adverse environments [51] and reveal little privacy [47], which inspires accurate and privacy-preserving HPE [2, 23, 46, 47]. However, due to the bandwidth and hardware limitations [29], radar PCs are sparse with limited geometric information obtained [2], leading to huge difficulties in achieving HPE. The sparse PCs are noisy in two aspects: (1) mmWave radar has a lower spatial resolution, leading to PC dispersion throughout the target area accompanied by ghost points caused by multi-path effect [29, 40]; (2) signal's specular reflection and interference [5, 6] further cause inconsistent sensing data, leading to occasional miss-detection of human parts.

To deal with sparse and noisy mmWave PCs, existing solutions mainly rely on kinds of data augmentations, e.g., multi-frame aggregation to enhance PC resolution [2, 46]. For the feature extractor, they directly borrow Long Short-Term Memory (LSTM) [46] or transformer-based architectures [6, 47] from existing RGB(D) HPE methods. However, these feature encoders are tailored for visual and language modalities, which struggle to handle noisy and inconsistent radar PCs [30]. As shown in Figure 1 (right), the existing SOTA solution [13] still suffers from pose vibration and severe drift, achieving undesirable performance.

Denoising Diffusion Probabilistic Models (DDPMs) [16, 37], also known as diffusion models, have demonstrated superior performance in various generative tasks, such as image generation and image restoration [15, 17, 57]. Diffusion models perform progressive noise elimination, transferring noisy distribution into desired target distribution [38]. Inspired by such capability, we aim to mitigate the noise of mmWave HPE, which motivates mmDiff, a diffusion-based pose estimator tailored for noisy radar PCs. Different from existing diffusion-based

HPE using RGB(D), HPE using mmWave PCs confronts two key challenges: (1) extracting robust features from noisy PCs where miss-detection of human bodies may happen, and (2) overcoming signal inconsistency for stable HPE. For the first challenge, we propose to isolate the feature extraction for different body joints, so that occasional miss-detection would not affect the feature extraction of detected joints. Extracting features directly from local PCs also improves the feature resolution. For the inconsistency issue, prior knowledge of human body structure and motion can reduce unreasonable cases, achieving consistent feature learning. As the human structure has a size constraint where limb-length should remain constant [8], the limb-length can be additionally estimated to prevent pose variance. Moreover, inspired by human motion continuity [32] that discourages abrupt human behavior changes, historical poses can be leveraged for pose generation refinement, minimizing pose vibration.

To this end, mmDiff first designs a conditional diffusion model capable of injecting radar information as the guiding conditions. Four modules are designed to extract clean and consistent information from radar point clouds: (1) Global radar context is proposed to isolate the globally extracted features for different human joints using a transformer [42], which generates more robust joint-wise features to handle miss-detection. (2) Local radar context is proposed to extract local features around body joints with a local transformer, which performs point-level attention for higher resolution. (3) Structural human limb-length consistency is proposed to extract human limb-length as consistent patterns, which reduce limb-length variance. (4) Temporal motion consistency is proposed to learn smooth motion patterns from historical estimated poses, which avoid pose vibration. Experiments have shown a significant improvement in pose estimation accuracy using mmDiff compared to the state-of-the-art models. Meanwhile, the generated poses demonstrate comparable structural and motion stability, validating the effectiveness of our designed conditional modules.

In summary, our contributions are three-fold. First, we propose a novel diffusion-based HPE framework with sparse and noisy mmWave radar PCs. To the best of our knowledge, mmDiff is the first diffusion-based paradigm for mmWave radar-based HPE. Second, four modules are proposed to extract robust representations from the noisy and inconsistent radar PCs considering global radar context, local radar context, structural limb-length consistency, and temporal motion consistency, used as the conditions to guide the diffusion process. Finally, extensive experiments show our approach achieves state-of-the-art performance on two public datasets: mmBody [6] and mm-Fi [47].

2 Related Work

mmWave Human Pose Estimation. For decades, extensive works [3, 26] have been focused on human pose estimation from RGB(D) images. Though achieving desirable accuracy, the major challenge faced by RGB(D) HPE is the performance drop under adverse environments and with self-occlusion [7]. Recently, commercial mmWave radar has been proven to extract sufficient information for human

body reconstruction [46], bringing the potential for mmWave HPE. Despite methods [22,45] utilizing raw radar signals for mmWave HPE, point clouds-based methods [1,2,6,46,47] become popular for its format uniformity. Particularly, Xue et al. [46] utilizes the LSTM model with anchor-based local encoding to deal with the noisy nature of radar. An et al. [2] integrate point clouds of consecutive frames to handle the point cloud sparsity. Chen et al. [6] and Yang et al. [47] propose transformer-based benchmarks based on their dataset. However, the sparse and noisy radar point clouds still hinder the accuracy of HPE using existing non-parametric regression models.

Diffusion Model for Human Pose Estimation. Diffusion models have been widely applied in image generation, such as image restoration [25], super-resolution [19], and text-to-image synthesis [34]. Since the proposition of the denoising diffusion probabilistic model (DDPM) [16], the diffusion model is extended to a wider range of generative applications, including pose/skeleton generation. DiffPose [15] formulate the 3D pose estimation problem as a pose generation task from low-determinant 2D poses to high-determinant 3D poses. Following DiffPose, Shan et al. [36] proposes to generate multi-hypothesis poses for 3D pose ambiguity and Saadatnejad et al. [35] proposes to predict the human poses in future time frames. Meanwhile, RGB-guided diffusion models for 2D and 3D human pose generation are also achieved [31,54]. Nevertheless, existing methods focus on diffusion-based HPE from stable and informative modality sources, either well-estimated 2D poses or high-resolution RGB images. None of these works investigate how noisy and sparse mmWave radar point clouds are used to guide the pose generation.

3 Methodology

We address the task of 3D human pose estimation (HPE) with noisy and sparse point clouds (PCs) from mmWave radar. At each time frame $t \geq 0$, given the input 6-dimensional radar point cloud $R_t \in \mathbb{R}^{N\times 6}$, where N denotes the number of detected points, our task is to estimate the ground-truth 17-joint human pose $H_t = \{h_t^1, h_t^2, ..., h_t^{17}\} \in \mathbb{R}^{17\times 3}$. Apart from the Cartesian coordinate $\{x, y, z\}$, each detected radar point also contains three attributes $\{v, E, A\}$, where v denotes velocity, E denotes energy, and A denotes amplitude.

To extract clean radar features and overcome mmWave signal inconsistency, we propose mmDiff as a diffusion-based paradigm for mmWave radar-based HPE. The overview of the architecture of mmDiff is presented in Figure 2. We start with an overall description of how human poses are generated by diffusion models using radar information as the guiding conditions in Section 3.1, followed by a detailed illustration of the four modules for extracting robust representations from the noisy and inconsistent radar PCs in Section 3.2 and Section 3.3. We omit the subscript t denoting the time frame in subsequent discussions for clarity.

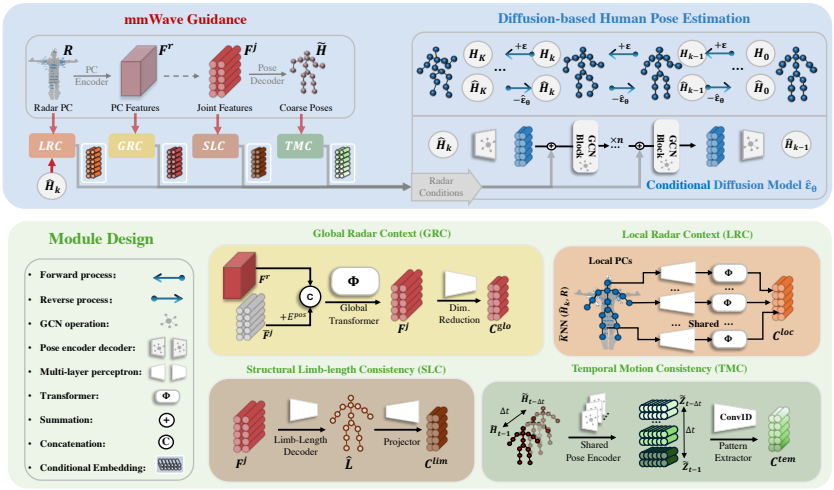


Fig. 2: mmDiff proposes diffusion-based pose estimation with a conditional diffusion model, using mmWave radar information as conditions. $k \in [0..K]$ denotes the diffusion step. Four modules are proposed to extract more reliable guidance, addressing PCs' noise and inconsistency respectively: GRC and LRC first extract more robust global-local radar features, C^{glo} and C^{loc}; SLC and TMC then extract consistent human structure and motion patterns, C^{tem} and C^{lim}.

3.1 Diffusion-based Human Pose Estimation

Diffusion models [16, 38] are a category of probabilistic generative models popular in various generative tasks, e.g., image generation. Given a noisy image from a noisy distribution, diffusion models can generate realistic image samples that match the natural image distribution, through iterative noise removal [19, 25]. Extended to HPE, diffusion models can estimate the distribution of reasonable human poses for realistic pose generation. Particularly with noisy radar modality, miss-detected body joints can be accurately estimated by inferring from the detected ones, and inaccurate joints causing twisted human structures are potentially refined.

The diffusion-based HPE consists of two processes: (1) a forward process gradually generates noisier samples as the training guidance and (2) a reverse process learns to invert the forward diffusion. Modelled by a Markov chain of length K, the forward process starts from the ground truth pose H_0, iteratively samples a noisier pose H_k by adding Gaussian noise $\epsilon \in \mathcal{N}(0, I)$:

$$q\left(H_k \mid H_{k-1}\right) = \mathcal{N}\left(H_k \mid \sqrt{1-\beta_k} H_{k-1}, \beta_k I\right), \tag{1}$$

where $k \in [0..K]$ refers to the diffusion step and β_k refers to the noise scale. On the contrary, the reverse process starts from a noisy pose initialization $\hat{H}_K$ and progressively removes noises until $\hat{H}_0$ is generated. For each step k, a diffusion

model $\hat{\varepsilon}_\theta$ is trained to identify the pose noise $\hat{\varepsilon}_k$ within $\hat{H}_k$, and remove it for more reliable pose $\hat{H}_{k-1}$:

$$\hat{\varepsilon}_k = \hat{\varepsilon}_\theta(\hat{H}_k, k), \tag{2}$$

$$\hat{H}_{k-1} = (1-\beta_k)^{-1}(\hat{H}_k - \beta_k \hat{\varepsilon}_k). \tag{3}$$

To better leverage the context information provided by radar's sensing data, we propose a set of conditions C containing latent feature embeddings extracted from the mmWave modality, to guide each step of the reverse process:

$$\hat{\varepsilon}_k = \hat{\varepsilon}_\theta(\hat{H}_k, C, k). \tag{4}$$

As an implementation, we propose a conditional diffusion model, which injects the radar conditions C in the latent feature space, using a Graph Convolution Network (GCN) [52] as the backbone. The GCN takes the 17-joint human pose $H_k \in \mathbb{R}^{17\times 3}$ as input, which is subsequently encoded into the latent pose embedding $Z_k \in \mathbb{R}^{17\times 96}$ by a GCN encoder, fed into n GCN blocks, and decoded into $\hat{\varepsilon}_k \in \mathbb{R}^{17\times 3}$ by a GCN decoder. To inject radar conditions, we propose to add the conditional embeddings from C to the pose feature Z_k before each GCN block, which serves as extra information for more accurate noise estimation of $\hat{\varepsilon}_k$. To align features, the conditional embeddings are also projected to the $\mathbb{R}^{17\times 96}$ feature space. In the following sections, we discuss in detail how to extract clean and consistent radar features that construct C.

3.2 Global-local Radar Context

Accurate guidance for the diffusion model depends on robust radar feature extraction, which should carefully handle the miss-detection of human bodies. In this section, we first revisit the existing mmWave feature extraction paradigm. Then, we further propose to improve it with two modules: (1) Global Radar Context (GRC) extracting features from overall PCs that can handle miss-detection; and (2) Local Radar Context (LRC) extracting local PC features near body joints for higher resolution.

Revisit mmWave Feature Extraction. Existing mmWave HPE paradigm applies encoder-decoder architectures to encode the radar point clouds (PCs), $R \in \mathbb{R}^{N\times 6}$, into latent feature representations and decode them into human poses. Anchor-based methods [13, 50] are common options for PC encoders, where point anchors are first sampled from the Farthest Point Sampling algorithm and nearby PCs are extracted as the anchor features. As a result, the holistic radar PCs are encoded into PC features $F^r \in \mathbb{R}^{P\times 1024}$, where P indicates the anchor number. To decode the PC features, a Multi-Layer Perceptron (MLP) is commonly applied as a dimension-reducing projection to generate the joint feature $F^j \in \mathbb{R}^{17\times 1024}$, which is further decoded into human poses $\tilde{H} \in \mathbb{R}^{17\times 3}$ by another MLP. However, occasional radar miss-detection causes uncertainty within the PC features F^r, while MLP-based projection can hardly identify miss-detected joints. Additionally, radar low-resolution PCs further impose noises. As a result, the estimated human joints are generally coarse.

Global Radar Context (GRC). To handle occasional miss-detection, GRC is proposed to isolate feature extraction for different body joints, using global information from F^r to construct more robust joint features F^j. From joint features, the diffusion model potentially identifies miss-detected joints and utilizes human prior knowledge for more accurate estimation. We exploit a Global-Transformer Φ^g to facilitate joint-wise feature extraction from the F^r. Following the *cls* token design in ViT [12], we first randomly initialize a trainable joint feature template (with no information) $\bar{F}^j \in \mathbb{R}^{17\times1024}$ and add it with positional embedding $E^{pos} \in \mathbb{R}^{17\times1024}$. Then, the joint feature template is concatenated with the PC feature $[\bar{F}^j, F^r] \in \mathbb{R}^{(17+P)\times1024}$ and fed into Φ^g:

$$F^j, F'^r = \Phi^g(\bar{F}^j, F^r). \tag{5}$$

The output $F^j \in \mathbb{R}^{17\times1024}$ is selected as the joint feature and the rest F'^r is ignored. Within the transformer, deep correlation is captured, not only within F^r but also between F^j and F^r. Each body joint performs individual feature extraction based on their correlation with the PC feature, so that features of detected joints are less affected by other undetected parts. Finally, as the 1024-dim F^j feature space is too sparse for the diffusion latent condition, an MLP-based dimension-reduction function g^g further condenses the extracted information within a $\mathbb{R}^{17\times64}$ conditional embedding:

$$C^{glo} = g^g(F^j). \tag{6}$$

Local Radar Context (LRC). LRC further performs local point-to-point self-attention to extract higher-resolution features from local PCs near body joints. To select local PCs, existing methods [46] utilize static joint anchors from coarse-estimated human poses $\tilde{H}$. However, the local joint features should dynamically reflect the joints' errors at different diffusion steps. Therefore, we propose dynamic joint anchors from intermediate diffusion poses $\hat{H}_k$ for PC selection. With $i \in [1, .., 17]$ and each joint $\hat{h}^i_k$ from the $\hat{H}_k$ as an anchor, the $\bar{K}$-nearest-neighbors ($\bar{K}$NN) algorithm is applied to select $\bar{K}$ nearest points as local PCs, $\bar{R}^i_k \in \mathbb{R}^{\bar{K}\times6}$. Each $\bar{R}^i_k$ is first encoded by a shared MLP g^l into a $\mathbb{R}^{\bar{K}\times64}$ embedding, and then fed into a shared small-scale local transformer Φ^l for point-to-point self-attention. Finally, average pooling is performed to generate $\mathbb{R}^{64}$ embeddings, which are further aggregated (concatenated) for every joint anchor $\tilde{h}^i$ as the conditional embedding:

$$C^{loc} = \bigcup_{i\in[1,..,17]} \Phi^l \circ g^l(\bar{R}^i_k). \tag{7}$$

3.3 Structural-motion Consistent Patterns

Inconsistent radar signals such as occasional miss-detection lead to discontinuous and unstable pose estimation, such as variant limb-length, pose vibration,

or inconsistent error frames. To mitigate such inconsistency, we further extract consistent human patterns based on human structure and motion prior knowledge: (1) Structural Limb-length Consistency (SLC) that extracts limb-length patterns to reduce limb-length variance, and (2) Temporal Motion Consistency (TMC) learns smooth motion patterns from historically estimated human poses.

Structural Limb-Length Consistency (SLC). SLC learns to extract a human-size indicator, the 16 limb-length $\hat{L} = \{\hat{l}^1, ..., \hat{l}^{16}\} \in \mathbb{R}^{16}$, where each limb-length measures the bone length connecting adjacent body joints. The extracted limb-length serves as a structural constraint to reduce the limb-length variance during the pose generation. Similar to decoding a pose, an MLP-based limb decoder g_1^{lim} is first applied to decode the previously extracted global joint feature F^j into predicted limb-length $\hat{L}$. Though with less dimension, the extracted limb-length contains physical meanings and thus is more consistent and stable. To ensure accurate limb-length decoding, a limb loss is designed to guide the training (details in Section 3.4). To further project the estimated limb-length $\hat{L}$ into latent feature space, another MLP-based projector g_2^{lim} then transforms the $\hat{L}$ into the $\mathbb{R}^{96}$ embedding, which is further broadcasted to different joints as the $\mathbb{R}^{17\times96}$ conditional embedding:

$$C^{lim} = g_2^{lim}(\hat{L}) = g_2^{lim} \circ g_1^{lim}(F^j). \tag{8}$$

Temporal Motion Consistency (TMC). Inspired by the fact that human motion is generally stable and consistent [32], multi-frame historical human poses can be utilized to extract the motion patterns in estimating the current pose. Such motion patterns provide temporal constraints for the diffusion model, which avoids error frames and pose vibration. As shown in Figure 2, TMC extracts the latent motion patterns from a sequence of historical-estimated coarse poses $\{\tilde{H}_{t-\Delta t}, ..., \tilde{H}_{t-1}\}$, where Δt is the number of historical frames. Firstly, a shared GCN encoder g_1^{tem} is applied to convert the pose sequence into feature embedding sequence $\{\tilde{Z}_{(t-\Delta t)}, ..., \tilde{Z}_{(t-1)}\} \in \mathbb{R}^{\Delta t\times(17\times96)}$. Then, a 1D-convolution-based pattern extractor g_2^{tem} is applied to extract the motion information along the temporal dimension, which generates a $\mathbb{R}^{17\times96}$ temporal embedding as C^{tem}:

$$C^{tem} = g_2^{tem}\left(\bigcup_{i\in[1..\Delta t]} \tilde{Z}_{(t-i)}\right) = g_2^{tem}\left(\bigcup_{i\in[1..\Delta t]} g_1^{tem}(\tilde{H}_{t-i})\right). \tag{9}$$

The reason for using 1D convolution is two-fold: (1) smooth motion features can be extracted by potentially averaging pose features of historical frames, which avoids pose vibration; and (2) the motion trend of the on-performing actions is potentially extracted to avoid inconsistent error frames. For example, increasing z values of the hand's location are expected when performing the 'raising hand'.

3.4 Overall Learning Objective

As feature extraction from global radar PCs is computation-consuming, we divide the training process into two phases. Phase one facilitates the extraction of the global joint features F^j and coarse estimation of human poses $\tilde{H}$. The GRC, PC encoder (from an off-the-shelf mmWave HPE network), and an MLP-based pose decoder are trained together. The learning objective of the phase one is minimizing the $\mathcal{L}_2$ pose regression loss $\mathcal{L}_{\text{joint}}$:

$$\mathcal{L}_{\text{joint}} = E_{i\sim[1,17]}||h^i - \tilde{h}^i\|_2^2. \tag{10}$$

Then, in phase two, the remaining three conditional modules and the diffusion model are trained together. The extracted F^j and $\tilde{H}$ serve as the input of structural-motion consistency modules, and $\tilde{H}$ initialize $\hat{H}_K$ for the reverse diffusion process. The learning objective is minimizing $\mathcal{L}_{\text{diff}}$, which is the diffusion learning objective following DDPM [16]. To ensure accurate limb-length estimation, a $\mathcal{L}_1$ limb regression loss is further designed and integrated:

$$\begin{aligned}\mathcal{L}_{\text{diff}} = \mathbb{E}_{k\sim[1,T]}\mathbb{E}_{\varepsilon_k\sim\mathcal{N}(0,I)}||\varepsilon_k - \hat{\varepsilon}_\theta(H_k, k, C)\|_2^2 \\ +\lambda * \mathbb{E}_{i\sim[1,16]}|l^i - \hat{l}^i|_1,\end{aligned} \tag{11}$$

where λ is a weighting parameter and each l^i is the ground truth limb-length calculated from ground truth pose H.

4 Experiments

Section 4.1 presents our experiment setting. Section 4.2 demonstrates the overall performance of mmDiff on two public datasets: mmBody [6] and mmFi [47]. Section 4.3 further performs ablation studies to validate the designed modules, followed by analytics of pose stability and module efficiency.

4.1 Experiment Setup

Datasets. mmBody [47] studies the robustness of human sensing with various sensors: RGB, Depth, and mmWave radar. Human skeletons are annotated by the MoCap system. Models are set to train on data collected from 2 standard scenes (Lab1 and Lab2), and tested on 3 basic scenes and 4 adverse scenes (including unseen subjects). Meanwhile, mm-Fi [47] is a larger-scale dataset consisting of 25 activities and 40 subjects using a low-cost mmWave radar (sparser PCs). The skeleton annotations are rather unstable compared to the MoCap annotations, as obtained by RGB using the pretrained HRNet-w48 [39]. We test our methods based on Protocol 1, with all daily activities. Three data-splitting methods are used with a train-test split ratio of 4:1: random split, cross-subject split, and cross-environment split.

Implementation Details. The GCN encoders, GCN blocks, and GCN decoders are designed following [52] and [15]. All MLPs are implemented as (LayerNorm, Linear, Dropout, ReLU, and Linear), and the temporal 1D-convolution extractor g_2^{tem} is implemented as (Conv1D(k=3), Dropout, ReLU, Conv1D(k=3), MaxPool).

Our methods are trained for 100 epochs with a fixed batch size of 1024. The Adam algorithm [20] algorithm is used for optimization, with the learning rate set as $2e-5$, the gradient clip set to 1.0, and the momentum set to 0.9. The length of the forward/reverse diffusion process is set as $K = 25$ with the constant β sampling of 0.001. An average of 5 hypotheses is recorded for fair comparison with non-diffusion methods. We choose Point4D [13] as the GRC's PC encoder for mmBody and PointTransformer [50] for mm-Fi. We set $\bar{K} = 50$ for $\bar{K}$NN algorithm for LRC, but due to insufficient radar points ($N < 100$), the LRC module is neglected for mm-Fi. We further set $\Delta t = 8$ for TMC and $\lambda = 5$ for SLC. The hyper-parameters are obtained empirically, more precise hyper-parameter tuning tricks such as the Bayesian optimization could lead to better results.

Compared Methods. (1) RGB, Depth, and RGB(D) [6] are benchmarks using different modalities. (2) mmWave methods: P4Transformer [13] for mmBody contains Point4D Convolution as the PC encoder and a transformer for self-attention; PointTransformer [50] is designed to handle the sparse PCs in mm-Fi. mmMesh [46] is implemented as an extra mmWave HPE baseline on mmBody. (3) SOTA camera-based HPE methods: camera-based 3D HPE methods generally perform pose lifting from 2D poses to 3D poses. To modify them to perform radar-based HPE, we train the models to perform pose refinements, from coarse poses $\tilde{H}$ to clean 3D poses. PoseFormer [53] is the transformer-based method with temporal-spatial attention. DiffPose [15] is the diffusion-based method using SOTA graph-based GraFormer [52] as backbones.

Evaluation Metric. Two evaluation metrics are adopted following [18]: (1) Mean Per Joint Position Error (MPJPE (mm)): the average joint error between ground truth and prediction (after pelvis alignment), indicating keypoint positional correctness; and (2) Procrustes Analysis MPJPE (PA-MPJPE (mm)): procrustes methods (translation, rotation, and scaling) are performed before error calculation, indicating the overall human pose quality.

4.2 Overall Result

Performance on mmBody. As shown in Table 1, mmWave-based methods have better robustness for cross-domain scenes and adverse environments. Particularly, our proposed mmDiff demonstrates superior results compared to related methods on the mmBody dataset. Compared to the SOTA mmWave HPE method, our method mmDiff(G,L,T,S) outperforms the P4Transformer [13] by 12.8% (MPJPE) and 11.3% (PA-MPJPE). For adverse environments, the improvement is more

Table 1: Quantitative results on mmBody [6], evaluated by MPJPE in white and PA-MPJPE in grey. * indicates diffusion-based methods. G, L, T, and S denote GRC, LRC, TMC, and SLC respectively. Bold is the best.

Methods	Basic Scenes						Adverse Environment								Average	
	Lab1		Lab2		Furnished		Rain		Smoke		Dark		Occlusion			
RGB [6]	74	/	73	/	71	/	80	/	86	/	105	/	/	/	81	/
Depth [6]	55	/	39	/	55	/	86	/	243	/	51	/	/	/	88	/
RGB(D) [6]	58	/	34	/	54	/	95	/	154	/	58	/	/	/	75	/
MM-Mesh [46]	95.1	69.48	87.87	77.3	93.28	73.14	106.9	72.38	106.7	76.52	83.54	64.19	85.55	62.5	94.13	70.79
P4transformer [13]	69.35	54.45	73.40	66.39	75.28	55.77	86.83	65.71	89.82	69.73	73.48	54.52	78.56	57.36	78.10	60.56
PoseFormer [53]	64.53	51.61	70.17	63.26	69.71	51.04	77.49	59.03	84.82	63.57	69.88	50.46	73.52	53.53	72.87	56.07
DiffPose* [15]	66.43	52.56	68.36	65.69	69.78	51.29	77.77	62.60	89.01	69.34	67.27	49.90	74.52	56.20	73.31	58.23
mmDiff(G)*	61.00	49.19	**67.79**	62.45	69.83	51.47	77.39	60.48	81.41	64.44	68.83	49.82	70.64	52.79	70.99	55.80
mmDiff(G,L)*	61.11	49.41	69.06	63.14	68.17	50.60	73.70	58.45	82.26	64.30	66.06	48.56	68.18	50.8	69.79	55.04
mmDiff(G,L,T)*	59.90	**47.81**	68.12	62.02	**66.98**	**48.63**	74.84	**58.13**	80.95	63.44	65.25	**47.26**	68.05	50.4	69.16	53.96
mmDiff(G,L,T,S)*	**59.52**	47.85	69.36	**61.23**	67.15	49.19	**71.03**	58.40	**76.92**	**62.25**	**65.08**	47.47	**67.47**	**49.54**	**68.08**	**53.71**

Table 2: Quantitative results on mm-Fi [47].

Methods	Random		Cross-Subject		Cross-Environment	
	MPJPE	PA-MPJPE	MPJPE	PA-MPJPE	MPJPE	PA-MPJPE
PointTransformer [50]	$73.09_{\pm 2.70}$	$55.60_{\pm 1.40}$	$75.96_{\pm 6.90}$	$58.70_{\pm 4.30}$	$88.28_{\pm 4.50}$	$68.79_{\pm 2.79}$
Diffpose* [15]	$73.44_{\pm 0.29}$	$56.83_{\pm 0.25}$	$70.31_{\pm 0.27}$	$54.12_{\pm 0.31}$	$86.35_{\pm 0.06}$	$66.87_{\pm 0.17}$
mmDiff(G)*	$68.62_{\pm 0.06}$	$53.11_{\pm 0.05}$	$68.46_{\pm 0.06}$	$52.55_{\pm 0.05}$	$85.63_{\pm 0.53}$	$66.43_{\pm 0.28}$
mmDiff(G,S)*	$65.72_{\pm 0.08}$	$50.72_{\pm 0.01}$	$67.18_{\pm 0.18}$	$51.85_{\pm 0.05}$	$83.39_{\pm 0.17}$	$64.61_{\pm 0.40}$
mmDiff(G,S,T)*	$\mathbf{65.26}_{\pm \mathbf{0.11}}$	$\mathbf{50.35}_{\pm \mathbf{0.09}}$	$\mathbf{65.62}_{\pm \mathbf{0.24}}$	$\mathbf{50.23}_{\pm \mathbf{0.24}}$	$\mathbf{82.73}_{\pm \mathbf{0.62}}$	$\mathbf{63.87}_{\pm \mathbf{0.26}}$

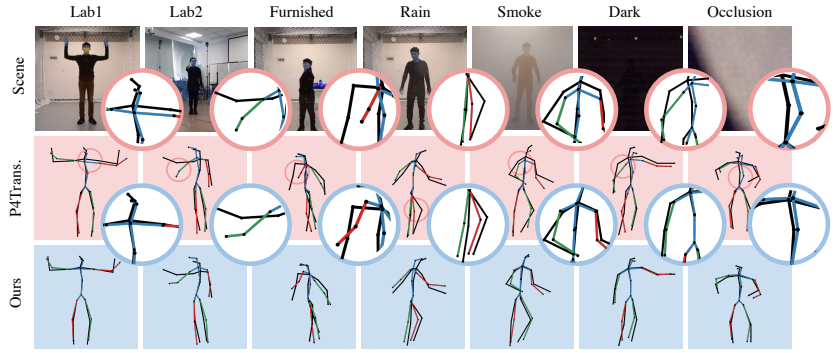


Fig. 3: Qualitative visualization of the estimated poses on mmBody dataset. mmDiff demonstrates higher keypoint accuracy.

significant with 14.7% (MPJPE) and 12.0% (PA-MPJPE), because radar signals get noisier due to the specular reflection and mmDiff has improved noise-handling capability. Compared to Diffpose [15] and PoseFormer [53] that utilize SOTA 3D HPE methods for pose refinement, mmDiff(G,L,T,S) still outperforms by 6.6% (MPJPE) and 4.2% (PA-MPJPE), demonstrating the proposed modules are dedicated to handle noisy and sparse radar modalities. Furthermore, mmDiff enables better mmWave-based performance compared to RGB-based methods under all

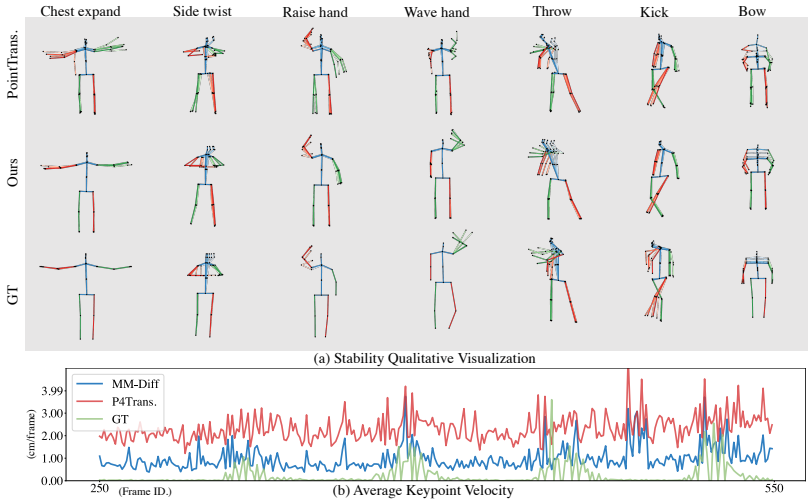


Fig. 4: (a) qualitatively shows the pose motion stability on mm-Fi, by plotting 5 consecutive frames of estimated poses. mmDiff shows more consistent motion patterns (zoom in for details). (b) shows the motion energy levels on mmBody, calculated by the Average Keypoint Velocity (AKV). Lower AKV indicates better stability.

scenes and RGB(D)-based methods under adverse environments. Qualitatively, we observe more accurate human poses as illustrated in Figure 3.

Performance on mm-Fi. As shown in Table 2, the generalizability of our proposed mmDiff is further explored on mm-Fi dataset. We compare the proposed method with the benchmark PointTransformer [50] and the SOTA DiffPose pose refinement [15]. Compared with the benchmark, our proposed method reduces the pose estimation error by 6.29% to 13.61% (MPJPE) and 7.15% to 14.42% (PA-MPJPE). Though the vanilla DiffPose for pose refinement can improve the result, the performance is further improved by 4.19% to 11.14% (MPJPE) and 4.49% to 11.40% (PA-MPJPE) with the integration of the radar context module and the consistency modules, further demonstrating their effectiveness. Moreover, using cross-subject splitting, our method shows comparable results to random splitting. As our model explicitly learns the human structural and motion patterns from the radar signals, these patterns potentially generalize for unseen subjects. Our method also demonstrates cross-domain ability as the performance steadily improves with the integration of different modules.

4.3 Analytics

Ablation Study. As shown in Table 3, we perform the ablation study to verify the effectiveness of each component from the following 4 perspectives. (1) The

vanilla diffusion design without any radar guidance can reduce the joint error by modeling pose distribution and refine deformed poses. (2) The effectiveness of the global-local radar context modules is verified by the performance gain of 4.8% (MPJPE) and 5.5% (PA-MPJPE). Both modules extract clean radar information with a different focus (local radar PCs or global PC features). (3) The effectiveness of the structure-motion consistency modules are verified by the performance gain of 5.3% (MPJPE) and 6.1% (PA-MPJPE). (4) The necessity of each module is further proved, as the elimination of different modules leads to a performance drop. Specifically, the performance drop is more significant when removing the limb-length module (C^{lim}) and temporal motion consistency module (C^{tem}), as the pose instability causes great pose inaccuracy.

Effect of Global-local Radar Context. To demonstrate the effectiveness of joint-wise feature extraction using the joint feature template, we compare our design with an alternative in Table 4: to directly generate PC feature guidance using the transformer without any template. Our designed joint feature guidance significantly outperforms the PC feature guidance, as the PC features are easily affected by the radar's miss-detection. To demonstrate the effectiveness of dynamic anchors of LRC, in Table 4 we compare our design with static anchor design [46] using coarse estimated human poses $\tilde{H}$. Our design has better performance and is more suitable for diffusion-based local PC selection.

Effects of Temporal Motion Consistency. In Figure 4 (a), we observe a mixing of incorrect skeletons within the correct skeleton timeline with PointTransformer [50], such as chest-expanding and hand-waving. With mmDiff, smooth motion patterns are ensured based on human behavioral prior knowledge. Additionally, our proposed mmDiff can mitigate the uncertain locations of legs and arms (caused by low resolution), demonstrating enhanced pose accuracy. Furthermore, for throwing and kicking actions, our method is more stable compared to RGB-extracted ground truth, as camera-based HPE suffers from self-occlusion. In Figure 4 (b), we further apply the Average Keypoint Velocity (AKV) to quantitatively measure the pose stability. AKV is defined as $E_{i=[0..17]}(||J_t^i - J_{t-1}^i||^2)$, which measures the average inter-frame joint moving distance, indicating the motion energy level and pose stability. The proposed mmDiff demonstrates en-

Table 3: Ablation studies of proposed modules on mmBody.

		Diffusion Model		Context Modules			Consistency Modules			Modules Elimination				Overall
Modules	Diffusion		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	Global			✓		✓					✓	✓	✓	✓
	Local				✓	✓				✓		✓	✓	✓
	Limb						✓		✓	✓	✓		✓	✓
	Temporal							✓	✓	✓	✓	✓		✓
Average	MPJPE	78.10	73.31	70.99	70.43	69.79	69.85	71.71	69.46	68.83	68.81	69.45	69.16	**68.08**
	PA-MPJPE	60.56	58.23	55.80	55.45	55.04	54.90	57.06	54.70	54.18	54.58	54.53	53.96	**53.71**

Table 4: Ablation studies of the detailed designs in global-local radar context modules.

Global Radar Context				
Modules	Diffusion	✓	✓	✓
	PC Features F^r		✓	
	Joint Features F^j			✓
Average	MPJPE	73.31	72.50	**70.99**
	PA-MPJPE	58.23	58.03	**55.80**

Local Radar Context				
Modules	Diffusion	✓	✓	✓
	Static Anchors		✓	
	Dynamic Anchors			✓
Average	MPJPE	73.31	71.05	**70.43**
	PA-MPJPE	58.23	56.06	**55.45**

Table 5: Model efficiency evaluated on mmBody. We compare mmDiff's diffusion training in phase two with the benchmark method P4Transformer [13]. Extra computational resources of our designed modules for one diffusion step are illustrated.

Modules	**Input**	**Input Size**	**Latency**	**#Params.**	**GFLOPs.**
P4Transformer [13]	R	$(N, 6)$	40.48 ms	128.00M	43.50
Diffusion model (D)	$\hat{H}_k$	$(17, 3)$	7.59 ms	1.03M	0.03
Global context (G)	F^j	$(17, 64)$	0.36 ms	0.02M	0.01
Local context (L)	$R, \hat{H}_k$	$(N, 6), (17, 3)$	2.00 ms	0.19M	0.40
Motion consistency (T)	$\{\hat{H}_{t-i}\}_{i=1}^{\Delta t}$	$(\Delta t, 17, 3)$	1.74 ms	15.98M	0.19
Limb consistency (S)	F^j	$(17, 64)$	0.16 ms	1.29M	0.02
(D+G+L+T+S)	/	/	11.85 ms	18.51M	0.62

hanced pose stability by minimizing joint vibration and avoiding abrupt pose changes.

Effects of Structural Limb-length Consistency. To validate the effectiveness of the limb-length consistency, we compare mmDiff w/ and w/o the module (C^{lim}) and plot the histograms that indicate the limb-length distribution in Figure 5. The ground truth limb length remains constant for all limbs. We observe both reduced limb-length error (in the estimated arms, legs, and spline) and variance (in forearms and lower legs), indicating enhanced pose accuracy and structural stability. We argue that more accurate limb-length can lead to more accurate

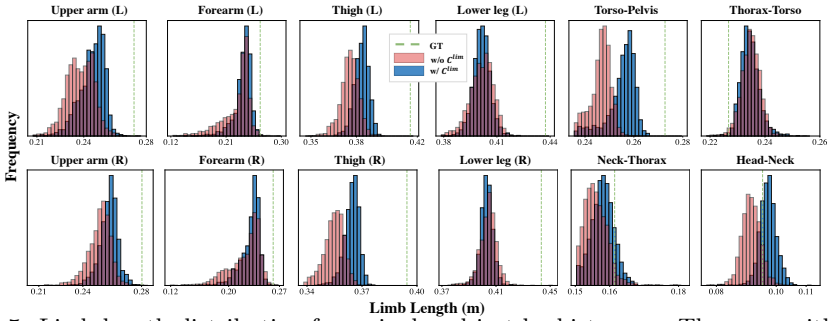


Fig. 5: Limb-length distribution for a single subject by histogram. The error within 5 cm can be treated as correct. With C^{lim}, more accurate limb-length and less variance are observed, as the distribution moves towards the GT and is more concentrated.

pose estimation. Qualitatively in figure 4, the variant height and arm's length for chest expanding and hand raising are mitigated with mmDiff.

Model Efficiency. We provide the efficiency analysis in Table 5, where our proposed modules require substantially little computational resources during phase-two diffusion training. We observe that all designed modules have a latency of less than 2ms, demonstrating outstanding efficiency in supporting the iterative diffusion process. Meanwhile, as reflected by the module's parameters, the model size of our designed modules is relatively small compared to the P4Transformer [13], leading to more stable training convergence. Moreover, the efficient module design requires minimum computation complexity, as reflected by GFLOPs, demonstrating the potential for applications like robotics, the Internet of Things, and edge computing.

5 Conclusion

In this paper, we propose mmDiff as a human pose estimation (HPE) method based on Radio Frequency vision (RF-vision). A conditional diffusion model is designed to generate accurate and stable human poses, conditioning on noisy mmWave radar point clouds. Our proposed modules perform the extraction of guiding features, demonstrating enhanced robustness in handling radar's miss-detection and signal inconsistency. Compared to the state-of-the-art methods, the proposed mmDiff demonstrates better accuracy and stability in mmWave human pose estimation.

References

1. An, S., Li, Y., Ogras, U.: mri: Multi-modal 3d human pose estimation dataset using mmwave, rgb-d, and inertial sensors. Advances in Neural Information Processing Systems **35**, 27414–27426 (2022) 1, 4
2. An, S., Ogras, U.Y.: Fast and scalable human pose estimation using mmwave point cloud. In: Proceedings of the 59th ACM/IEEE Design Automation Conference. pp. 889–894 (2022) 2, 4, 22
3. Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B.: 2d human pose estimation: New benchmark and state of the art analysis. In: Proceedings of the IEEE Conference on computer Vision and Pattern Recognition. pp. 3686–3693 (2014) 1, 3
4. Ar, I., Akgul, Y.S.: A computerized recognition system for the home-based physiotherapy exercises using an rgbd camera. IEEE Transactions on Neural Systems and Rehabilitation Engineering **22**(6), 1160–1171 (2014) 1
5. Bansal, K., Rungta, K., Zhu, S., Bharadia, D.: Pointillism: Accurate 3d bounding box estimation with multi-radars. In: Proceedings of the 18th Conference on Embedded Networked Sensor Systems. pp. 340–353 (2020) 2
6. Chen, A., Wang, X., Zhu, S., Li, Y., Chen, J., Ye, Q.: mmbody benchmark: 3d body reconstruction dataset and analysis for millimeter wave radar. In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 3501–3510 (2022) 2, 3, 4, 9, 10, 11, 21, 22, 24, 27

7. Chen, H., Feng, R., Wu, S., Xu, H., Zhou, F., Liu, Z.: 2d human pose estimation: A survey. Multimedia Systems **29**(5), 3115–3138 (2023) 3
8. Chen, T., Fang, C., Shen, X., Zhu, Y., Chen, Z., Luo, J.: Anatomy-aware 3d human pose estimation with bone-based pose decomposition. IEEE Transactions on Circuits and Systems for Video Technology **32**(1), 198–209 (2021) 3
9. Chen, Y., Wang, Z., Peng, Y., Zhang, Z., Yu, G., Sun, J.: Cascaded pyramid network for multi-person pose estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7103–7112 (2018) 1
10. Defferrard, M., Bresson, X., Vandergheynst, P.: Convolutional neural networks on graphs with fast localized spectral filtering. Advances in neural information processing systems **29** (2016) 23
11. Deng, T., Xie, H., Wang, J., Chen, W.: Long-term visual simultaneous localization and mapping: Using a bayesian persistence filter-based global map prediction. IEEE Robotics & Automation Magazine **30**(1), 36–49 (2023) 1
12. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020) 7
13. Fan, H., Yang, Y., Kankanhalli, M.: 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. pp. 14204–14213 (2021) 2, 6, 10, 11, 14, 15, 24
14. Gao, Q., Liu, J., Ju, Z., Zhang, X.: Dual-hand detection for human–robot interaction by a parallel network based on hand detection and body pose estimation. IEEE Transactions on Industrial Electronics **66**(12), 9663–9672 (2019) 1
15. Gong, J., Foo, L.G., Fan, Z., Ke, Q., Rahmani, H., Liu, J.: Diffpose: Toward more reliable 3d pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13041–13051 (2023) 2, 4, 10, 11, 12, 25
16. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020) 2, 4, 5, 9, 21
17. Hu, M., Wang, Y., Cham, T.J., Yang, J., Suganthan, P.N.: Global context with discrete diffusion in vector quantised modelling for image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11502–11511 (2022) 2
18. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: 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 (2013) 10, 23
19. Kawar, B., Elad, M., Ermon, S., Song, J.: Denoising diffusion restoration models. Advances in Neural Information Processing Systems **35**, 23593–23606 (2022) 4, 5
20. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014) 10
21. Kocabas, M., Athanasiou, N., Black, M.J.: Vibe: Video inference for human body pose and shape estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5253–5263 (2020) 1
22. Lee, S.P., Kini, N.P., Peng, W.H., Ma, C.W., Hwang, J.N.: Hupr: A benchmark for human pose estimation using millimeter wave radar. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 5715–5724 (2023) 4

23. Li, G., Zhang, Z., Yang, H., Pan, J., Chen, D., Zhang, J.: Capturing human pose using mmwave radar. In: 2020 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops). pp. 1–6. IEEE (2020) 2
24. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: Smpl: A skinned multi-person linear model. In: Seminal Graphics Papers: Pushing the Boundaries, Volume 2, pp. 851–866. ACM (2023) 21
25. Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., Van Gool, L.: Repaint: Inpainting using denoising diffusion probabilistic models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11461–11471 (2022) 4, 5
26. Martinez, J., Hossain, R., Romero, J., Little, J.J.: A simple yet effective baseline for 3d human pose estimation. In: Proceedings of the IEEE international conference on computer vision. pp. 2640–2649 (2017) 1, 3
27. Mehta, D., Sridhar, S., Sotnychenko, O., Rhodin, H., Shafiei, M., Seidel, H.P., Xu, W., Casas, D., Theobalt, C.: Vnect: Real-time 3d human pose estimation with a single rgb camera. Acm transactions on graphics (tog) **36**(4), 1–14 (2017) 1
28. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32** (2019) 22
29. Prabhakara, A., Jin, T., Das, A., Bhatt, G., Kumari, L., Soltanaghai, E., Bilmes, J., Kumar, S., Rowe, A.: High resolution point clouds from mmwave radar. In: 2023 IEEE International Conference on Robotics and Automation (ICRA). pp. 4135–4142. IEEE (2023) 2, 22
30. Qi, J., Du, J., Siniscalchi, S.M., Ma, X., Lee, C.H.: Analyzing upper bounds on mean absolute errors for deep neural network-based vector-to-vector regression. IEEE Transactions on Signal Processing **68**, 3411–3422 (2020) 2
31. Qiu, Z., Yang, Q., Wang, J., Wang, X., Xu, C., Fu, D., Yao, K., Han, J., Ding, E., Wang, J.: Learning structure-guided diffusion model for 2d human pose estimation. arXiv preprint arXiv:2306.17074 (2023) 4
32. Ramakrishna, V., Kanade, T., Sheikh, Y.: Tracking human pose by tracking symmetric parts. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3728–3735 (2013) 3, 8
33. Rao, S.: Introduction to mmwave sensing: Fmcw radars. Texas Instruments (TI) mmWave Training Series pp. 1–11 (2017) 2, 21
34. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022) 4
35. Saadatnejad, S., Rasekh, A., Mofayezi, M., Medghalchi, Y., Rajabzadeh, S., Mordan, T., Alahi, A.: A generic diffusion-based approach for 3d human pose prediction in the wild. In: 2023 IEEE International Conference on Robotics and Automation (ICRA). pp. 8246–8253. IEEE (2023) 4
36. Shan, W., Liu, Z., Zhang, X., Wang, Z., Han, K., Wang, S., Ma, S., Gao, W.: Diffusion-based 3d human pose estimation with multi-hypothesis aggregation. arXiv preprint arXiv:2303.11579 (2023) 4
37. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020) 2
38. Song, Y., Ermon, S.: Generative modeling by estimating gradients of the data distribution. Advances in neural information processing systems **32** (2019) 2, 5

39. Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5693–5703 (2019) 9
40. Sun, Y., Huang, Z., Zhang, H., Cao, Z., Xu, D.: 3drimr: 3d reconstruction and imaging via mmwave radar based on deep learning. In: 2021 IEEE International Performance, Computing, and Communications Conference (IPCCC). pp. 1–8. IEEE (2021) 2
41. Tao, T., Yang, X., Xu, J., Wang, W., Zhang, S., Li, M., Xu, G.: Trajectory planning of upper limb rehabilitation robot based on human pose estimation. In: 2020 17th International Conference on Ubiquitous Robots (UR). pp. 333–338. IEEE (2020) 1
42. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017) 3, 24
43. Waldschmidt, C., Hasch, J., Menzel, W.: Automotive radar—from first efforts to future systems. IEEE Journal of Microwaves **1**(1), 135–148 (2021) 2
44. Wang, J., Jin, S., Liu, W., Liu, W., Qian, C., Luo, P.: When human pose estimation meets robustness: Adversarial algorithms and benchmarks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11855–11864 (2021) 1
45. Xue, H., Cao, Q., Miao, C., Ju, Y., Hu, H., Zhang, A., Su, L.: Towards generalized mmwave-based human pose estimation through signal augmentation. In: Proceedings of the 29th Annual International Conference on Mobile Computing and Networking. pp. 1–15 (2023) 4
46. Xue, H., Ju, Y., Miao, C., Wang, Y., Wang, S., Zhang, A., Su, L.: 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. pp. 269–282 (2021) 2, 4, 7, 10, 11, 13, 19, 21, 25
47. Yang, J., Huang, H., Zhou, Y., Chen, X., Xu, Y., Yuan, S., Zou, H., Lu, C.X., Xie, L.: Mm-fi: Multi-modal non-intrusive 4d human dataset for versatile wireless sensing. arXiv preprint arXiv:2305.10345 (2023) 2, 3, 4, 9, 11, 22
48. Yang, J., Zhou, Y., Huang, H., Zou, H., Xie, L.: Metafi: Device-free pose estimation via commodity wifi for metaverse avatar simulation. In: 2022 IEEE 8th World Forum on Internet of Things (WF-IoT). pp. 1–6. IEEE (2022) 1
49. Zhang, J., Xi, R., He, Y., Sun, Y., Guo, X., Wang, W., Na, X., Liu, Y., Shi, Z., Gu, T.: A survey of mmwave-based human sensing: Technology, platforms and applications. IEEE Communications Surveys & Tutorials (2023) 2
50. Zhao, H., Jiang, L., Jia, J., Torr, P.H., Koltun, V.: Point transformer. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 16259–16268 (2021) 6, 10, 11, 12, 13, 24, 30
51. Zhao, M., Li, T., Abu Alsheikh, M., Tian, Y., Zhao, H., Torralba, A., Katabi, D.: Through-wall human pose estimation using radio signals. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7356–7365 (2018) 2
52. Zhao, W., Wang, W., Tian, Y.: Graformer: Graph-oriented transformer for 3d pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20438–20447 (2022) 6, 10, 23
53. Zheng, C., Zhu, S., Mendieta, M., Yang, T., Chen, C., Ding, Z.: 3d human pose estimation with spatial and temporal transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11656–11665 (2021) 10, 11, 25

54. Zhou, J., Zhang, T., Hayder, Z., Petersson, L., Harandi, M.: Diff3dhpe: A diffusion model for 3d human pose estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2092–2102 (2023) 4
55. Zhou, Y., Huang, H., Yuan, S., Zou, H., Xie, L., Yang, J.: Metafi++: Wifi-enabled transformer-based human pose estimation for metaverse avatar simulation. IEEE Internet of Things Journal (2023) 1
56. Zhou, Y., Yang, J., Huang, H., Xie, L.: Adapose: Towards cross-site device-free human pose estimation with commodity wifi. arXiv preprint arXiv:2309.16964 (2023) 1
57. Zhu, Y., Li, Z., Wang, T., He, M., Yao, C.: Conditional text image generation with diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14235–14245 (2023) 2

Appendix

1 mmWave Human Sensing

In Figure 6, we illustrate the utilization of mmWave radar for human sensing, which detects human actions within a range of 3-5 meters from the radar. This process generates mmWave point clouds (PCs). Simultaneously, a keypoint annotation system such as VICON, Mocap, or Cameras is deployed to record ground-truth human poses for reference. The mmWave radar-based human pose estimation refers to training a neural network to estimate human poses using mmWave radar point clouds as input.

For general mmWave human sensing, FMCW (Frequency Modulated Continuous Wave) chirp signals are transmitted and their reflections are received through antenna arrays. These chirp signals are defined by parameters such as start frequency f_c, bandwidth B, and duration T_c. To generate radar PCs [46], range-FFT separates different frequency components f from the IF signals, enabling the extraction of object distances using the formula $R = \frac{cfT_c}{2B}$, where c is the speed of light. Doppler-FFT measures phase changes ω of the IF signals, facilitating the calculation of object velocities using $v = \frac{\lambda\omega}{4\pi T_c}$, where λ is the wavelength of the chirp. Elevation angles φ and azimuth angles θ of the detected objects are determined based on $\varphi = sin^{-1}(\frac{\omega_z}{\pi})$ and $\theta = sin^{-1}(\frac{\omega_x}{cos(\varphi)\pi})$, where ω_x is the phase change between azimuth antennas and corresponding elevation antennas, and ω_z is the phase change of consecutive azimuth antennas. Finally, the Cartesian coordinates (x, y, z) of the detected point clouds are calculated as follows: $x = Rcos(\varphi)sin(\theta)$, $z = Rsin(\varphi)$, and $y = (R^2 - x^2 - z^2)$.

2 The Diffusion Model for Human Pose Generation

Forward process. The forward process involves the gradual sampling of increasingly noisy human poses, resulting in an intermediate distribution of noisy poses that serve as training guidance. Here, $k \in [1..K]$ denotes the diffusion steps, where noisy human poses H_k are sampled by adding noise to the original ground

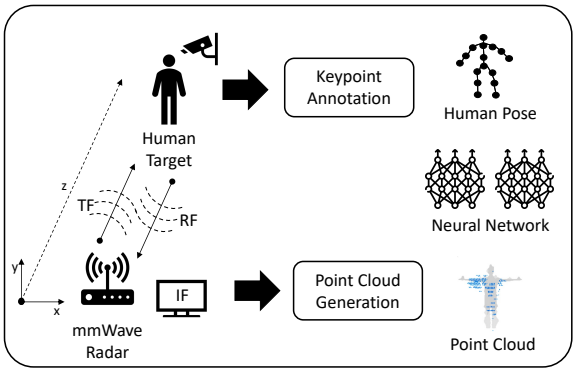


Fig. 6: Workflow of mmWave sensing and mmWave human pose estimation.

truth pose H_0. Following the Markov process, we interatively add noise to the pose H_{k-1} and obtain a noisier version of pose H_k:

$$q\left(H_k \mid H_{k-1}\right) = \mathcal{N}\left(H_k \mid \sqrt{\alpha_k} H_{k-1}, \left(1-\alpha_k\right) I\right), \tag{12}$$

$$H_k = \sqrt{1-\beta_k} H_{k-1} + \beta_k \varepsilon, \tag{13}$$

where β_k denotes the noise scale, $\varepsilon \in \mathcal{N}(0, I)$ denotes the random noise, and $\alpha_k = 1 - \beta_k$. With a small noise scale $\beta_k \in [0.0001, 0.001]$, H_k is approaching to H_{k-1}, which allows us to model both the forward sampling $q(H_k|H_{k-1})$ and the reverse estimation $p_\theta(\hat{H}_{k-1}|H_k, C)$ as Gaussian distributions. To directly sample H_k from the ground truth H_0, Eq. 13 can be rewritten to:

$$H_k = \sqrt{\gamma_k} H_0 + \sqrt{1-\gamma_k}\boldsymbol{\varepsilon}, \tag{14}$$

where $\gamma_k = \prod_{i \in [1..k]} \alpha_k$. Eventually, when $k \to \infty$, H_k approaches pure random noise following Gaussian distribution.

Reverse process. At each iteration k of the reverse process, a cleaner pose $\hat{H}_{k-1}$ is estimated to approximate H_{k-1} (generated by the forward process), given a noisy pose H_k and the conditional set C. The reverse process can be formulated as:

$$p_\theta\left(\hat{H}_{0:K} \mid H_0, C\right) = p\left(H_K\right) \prod_{k=1}^{K} p_\theta\left(\hat{H}_{k-1} \mid H_k, C\right). \tag{15}$$

During training, $H_{1:K}$ are obtained by adding noise to the ground truth H_0, following the forward process. However, during inference, as the ground truth is unavailable, we set $H_{1:K} = \hat{H}_{1:K}$ by iteratively inputting the estimated human poses $\hat{H}_{1:K}$ into the trained diffusion model. As discussed in the main paper, $\hat{H}_K$ is initialized by a coarsely estimated pose $\tilde{H}$ as the starting point of the reverse process.

Table 6: Overview of the datasets used for mmWave human pose estimation.

Dataset:	mmMesh	mmBody	mm-Fi
Radar Type:	AWR1843BOOST mmWave radar from Texas Instruments.	Phoenix mmWave radar from Arbe Robotics.	IWR6843 60-64GHz mmWave radar from Texas Instruments.
Annotations:	Mesh annotated by VICON motion capture system and generated by SMPL.	55 keypoints are annotated by the OptiTrack Mocap system; Mesh is generated by Mosh++ and SMPL-X.	2D keypoints are obtained by HRNet-w48 from two-view infra-red cameras; 3D keypoints are calculated by triangulation.
Point format:	Cartesian: (x, y, z); PC attributes: (range, velocity, energy).	Cartesian: (x, y, z); PC attributes: (velocity, amplitude, energy).	Cartesian: (x, y, z); PC attributes: (velocity, intensity).
Public or not:	No.	Yes.	Yes.
# **of subjects:**	Not mentioned.	20 (10 males, 10 females).	40 (29 males and 11 females)
# **of actions:**	Not mentioned.	100 motions (16 static poses, 9 torso motions, 20 leg motions, 25 arm motions, 3 neck motions, 14 sports motions, 7 daily indoor motions, and 6 kitchen motions).	27 actions (14 daily activities and 13 rehabilitation exercises) for a duration of 30 seconds.
# **of frames:**	Not mentioned.	39892 frames for training and 28048 frames for testing.	133920 frames for training and 38400 frames for testing.
Scenes:	Normal and occlusion.	Lab1, Lab2, Furnished, Poor_lighting, Rain, Smoke, and Occlusion.	Normal, Cross-subject, and Cross-environment.

Model training. We follow DDPM [16] for faster convergence of the diffusion model. Firstly, diffusion step $k \in [1..K]$ and $\varepsilon \in \mathcal{N}(0, I)$ are randomly sampled, and H_k is calculated according to Eq. 14. Then, the diffusion model is trained to approximate $\hat{\varepsilon}$ to ε, rather than directly approximate $\hat{H}_{k-1}$ to H_{k-1}. The learning objective following DDPM is formulated as:

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{k\sim[1,T]}\mathbb{E}_{\varepsilon\sim\mathcal{N}(0,I)}||\varepsilon - \hat{\varepsilon}_\theta(H_k, k, C)||_2^2. \tag{16}$$

3 Evaluated Datasets

As shown in Table 6, we present a comparison of existing datasets for point-cloud-based mmWave human pose estimation (HPE), focusing on several key attributes including mmWave radar sensor type, keypoint annotations, radar point cloud format, dataset size, and the variety of scenes covered in the dataset.

mmMesh [46]. mmMesh is the systematic work proposing mmWave human sensing (mesh reconstruction) with an off-the-shelf commercial mmWave sensor, AWR1843BOOST mmWave from Texas Instruments [33]. It provides a systematic way to annotate human mesh (including human keypoints) with the VICON system and SMPL [24] algorithm. The mmMesh dataset is presented for comparison, as it is not public and only contains normal and occlusion scenarios.

mmBody [6]. mmBody dataset is a public dataset for mesh reconstruction with multi-modal sensors: depth camera, RGB camera, and mmWave radar. Specifically,

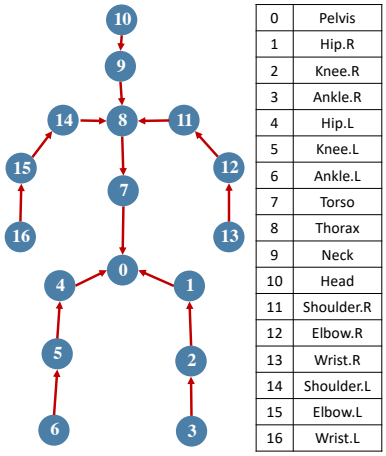
Phonix mmWave radar from Arbe Robotics is chosen as the mmWave sensor, which extracts thousands of radar points for scene detection. Still, the detected radar point clouds are noisy and sparse compared to RGB and depth sensors. The dataset contains daily-life motions, with heterogeneous human motions including motion of the torso, leg, arm, etc. Due to numerous subtle limb motions in the dataset, it is challenging to accurately predict human poses. To evaluate the robustness of mmWave HPE, the dataset includes various cross-domain scenes (lab2 and furnished) and adverse scenes (dark, rain, smoke, and occlusion). Meanwhile, except for lab2 containing seen subjects, all other scenes contain unseen subjects for testing, which is challenging for the model's generalizability. Our experiment is conducted following the mmBody settings [6].

mm-Fi [47]. mm-Fi offers a broader scope of human sensing, including action recognition and HPE, leveraging a variety of multi-modal sensors: RGB(D), LIDAR, WiFi, and mmWave radar. It is a large-scale dataset with 40 subjects participating and over 15k frames for training and testing. The mmWave radar utilized in the dataset is IWR6843 60-64GHz mmWave, which is a low-cost option generating a limited number of radar points. The point cloud format omits redundant range features. Meanwhile, different from the other two datasets, the model is trained in a self-supervised manner, as the human pose annotations are obtained by RGB image using HRNet-w48 [29]. The annotations are rather unstable compared to the motion capture systems. To test the model's generalizability, the dataset also proposes the cross-subject and cross-environment splits. Our experiment follows protocol 1 (P1) to include all daily-life activities and adopt all splitting methods.

4 Detailed Experiment Settings

4.1 Data Preprocessing

Radar point clouds preprocessing. Due to radar sparsity and the occasional miss-detection, we follow [2] to concatenate adjacent frames to enrich the number of points. Specifically, 4 frames are concatenated for mmBody and 5 frames are concatenated for mm-Fi. Further, since mmWave radar point clouds are generated by the targets with salient Doppler velocity, the number of radar points is frame-wise variant. As a result, to enable mini-batch training using PyTorch dataloader [28], we perform zero-padding (null points with 0 values) to guarantee the invariant input tensor shape. For mmBody, we zero-padding the point clouds to 5000 points, while a dynamic padding technique [47] is applied for mm-Fi. Moreover, to handle noisy radar points resulting from environmental reflection and interference, we perform point cloud cropping to select only the points within the region of human activities. For mmBody, the region of human activities is centered at the ground-truth pelvis location, with a region size of (x:$\pm$1.6m, y:$\pm$1.6m, z:$\pm$1.6m). However, for mm-Fi, as point clouds are generated solely by moving targets, cropping is unnecessary.



0	Pelvis
1	Hip.R
2	Knee.R
3	Ankle.R
4	Hip.L
5	Knee.L
6	Ankle.L
7	Torso
8	Thorax
9	Neck
10	Head
11	Shoulder.R
12	Elbow.R
13	Wrist.R
14	Shoulder.L
15	Elbow.L
16	Wrist.L

Fig. 7: Selected keypoints (ID and names) for mmWave human pose estimation. Red arrows indicate the select limbs.

Human pose and limb-length preprocessing. Our ground truth keypoints $H = \{h_1, .., h_{17}\}$ are selected according to Figure 7. Following [18] to construct ground-truth human poses, we perform pose normalization by pelvis alignment: subtracting the pelvis position h_1 from every keypoint $h_{1:17}$ of the skeleton. It's worth noting that the pelvis alignment is missing for the mm-Fi dataset, resulting in a higher Mean Per Joint Position Error (MPJPE). Consequently, we incorporate human pose normalization for a fair comparison. To calculate the ground truth 16 limb-length $L = \{l_1, .., l_{16}\}$, we compute the $\mathcal{L}_2$ distance between two adjacent human keypoints, as selected according to Figure 7.

4.2 Implementation of mmDiff

Conditional diffusion model. The conditional diffusion model adopts a Graph Convolution Network (GCN) architecture inspired by GraphFormer [52] as its backbone. The pose encoder and decoder utilize Chebyshev graph convolution layers [10] to project human poses to 96-dimensional pose embeddings. The GCN block is implemented by stacking a Chebyshev graph convolution layer and a graph attention layer, with skip connection. The GCN backbone consists of 5 GCN blocks. Chebyshev graph convolution layer first projects the 96-dimensional pose embeddings to 96-dimensional hidden feature embeddings. Subsequently, each graph attention layer performs self-attention with 4 attention heads in a 96-dimensional hidden feature space. During training, the GCN backbone applies a dropout rate of 0.25 and an exponential moving average (EMA) rate of 0.999 0.999. The conditional embeddings are added at the start of each GCN block. Notably, as the Graph-Conditional Generation (GRC) model is trained before the diffusion training, the extracted conditional embeddings C^{glo} require an additional projection to align with the human pose feature embeddings. This

alignment is achieved through a Chebyshev graph convolution layer projecting from 64 dimensions to 96 dimensions.

Radar point cloud encoder. Since mmBody and mm-Fi utilize different mmWave radars for sensing, and mm-Fi has sparser point clouds, the design of the radar point cloud (PC) encoder within the Global Radar Context (GRC) model differs between the two datasets. For mmBody, the PC encoder adopts the point 4D convolution following the P4Transformer benchmark [13]. Anchors are selected using farthest point sampling (FPS) with a spatial stride of 32, resulting in the selection of 312 anchors. Ball query is then applied to retrieve nearby radar points within a radius of 0.1 and 32 samples. A temporal convolution with a kernel size of 32 and a stride of 2 is subsequently applied to encode the nearby PCs into a 1024-dimensional PC feature space. On the other hand, for mm-Fi, the radar PC encoder is designed following the PointTransformer approach [50]. It utilizes 5 point-attention blocks with $n_{\text{neighbor}} = 16$ and $d_f = 128$ feature dimensions. After the radar PC feature encoders, the extracted PC features are concatenated with randomly initialized joint feature templates and fed into the global-transformer layers [42].

Global and local transformers. The global transformer consists of $D = 10$ transformer layers, each utilizing $H = 8$ attention heads and a hidden feature dimension of 256. The local lightweight transformer consists of $D = 5$ transformer layers, $H = 8$ attention heads, and a hidden feature dimension of 96. In each attention layer of both transformers, skip connections are applied. Additionally, after the local transformer, each anchor computes the proportion of points that are within a distance threshold of $thre = 0.04$m. This proportion serves as a measure of reliability for the local features. We incorporate it by multiplying the local features with the calculated proportion (which falls in the range $[0, 1]$). This approach helps to weigh the local features based on their distance to nearby points, enhancing the model's ability to focus on more relevant information.

4.3 Implementation of other Compared Methods

The methods using other modalities, e.g., RGB and depth are directly recorded from mmBody [6]. Other methods use mmWave radar point clouds (PCs) as the input modality, which are described in detail:

mmWave-based HPE methods. For a more direct comparison of performance gains achieved by mmDiff, it's important to note that the radar point cloud (PC) encoders of P4Transformer [13] and PointTransformer [50] are the same with mmDiff's radar PC encoder.

In the case of P4Transformer, which serves as the benchmark for mmBody, the PC encoder first utilizes a point4D convolution layer to extract PC features $F^r \in \mathbb{R}^{P \times 1024}$. Subsequently, a global transformer performs self-attention on

these features. Unlike GRC modules, the input to the global transformer is solely the PC features F^r without any joint feature templates $\bar{F}^j$. Following the global transformer, a nonparametric max-pooling layer aggregates $\mathbb{R}^{P\times 1024}$ PC features into $\mathbb{R}^{1024}$. This aggregated feature is then dimensionally reduced into a $\mathbb{R}^{64}$ feature space through an MLP and decoded into $\mathbb{R}^{17*3}$ human poses. On the other hand, PointTransformer, which serves as the benchmark for mm-Fi, employs a point attention PC encoder to extract PC features. Similar to P4Transformer, a global transformer performs self-attention on these features. Then, a max-pooling layer aggregates PC features, followed by two MLPs for dimension reduction and pose decoding. Lastly, mmMesh [46] is implemented following its open-source design without modification.

Camera-based 3D HPE methods. Since camera-based 3D Human Pose Estimation (HPE) methods cannot directly utilize radar point clouds as input, we adopt a different approach. These methods typically involve pose-lifting, where 2D human poses are used as input to predict 3D human poses. Therefore, we fine-tune these models for 3D pose refinement: Given coarsely estimated 3D human poses $\tilde{H}$, the models aim to estimate cleaner 3D human poses $\hat{H}$ by eliminating noise within $\tilde{H}$.

To ensure a fair comparison, the input coarse human poses $\tilde{H}$ are the same for mmDiff and the compared 3D HPE methods. We implement PoseFormer [53] following its open-source design. Similarly, we implement DiffPose [15] using the same GCN diffusion backbone as used in mmDiff. This consistent setup allows for a direct comparison between the performance of mmDiff and these 3D HPE methods in refining 3D human poses.

5 Supplementary Results

5.1 Visualization of Diffusion Process

As shown in Figure 8, we provide the visualization of progressive noise elimination of diffusion-based human pose estimation, where mmDiff performs progressive noise elimination and generates human poses from coarse to fine. The pose is refined from yellow to green. We can observe progressively improved keypoint accuracy during the pose refinement process: For the lab1 scene, the hands' locations are approaching the head with the increment of diffusion steps; For the rain scene, the legs' locations are also corrected to better reflect the walking poses. Meanwhile, we observe the correction of pose deformity in the smoke and dark scenes, as the shoulders' locations are progressively refined to the ground truth location. Furthermore, due to the limited pages in the main paper, we provide the visualization of motion continuity consistency and limb-length distribution of the spine length here. As shown in Figure 9, we observe the correction of erroneous frames w/ temporal motion consistency.

5.2 Visualization of Global Radar Context

To demonstrate the effectiveness in extracting more robust joint-wise features, we perform further visualization of the transformer's attention heatmap using the global transformer in the Global Radar Context (GRC). First, as shown in Figure 10, the global transformer extracts the self-correlation within the PC features F^r and the inter-correlation between F^r and F^j. In Figure 11, we further demonstrate the feature extraction of different joints does not affect each other, as they focus on different parts of the PC features. Such quality facilitates joint-wise feature extraction. Finally, in Figure 12 and 13, we visualize the attention region of different joint features. We observe that the feature extraction of detected joints focuses on the correct-detected PC region, which facilitates more robust feature extraction. Though feature extraction of undetected joints focuses on the wrong part, it does not affect the feature extraction of other detected joints. Therefore, the joint-wise feature extraction is more robust for detected joints.

5.3 Visualization of Local Radar Context

To illustrate the effectiveness of dynamic local PC selection in the Local Radar Context (LRC), we provide a visualization of the local PC selection. We examine the local PC selection during model inference, as shown in Figure 14. The dynamic joint anchors utilize intermediate diffusion poses $\hat{H}_k$, starting from coarsely estimated poses $\tilde{H}$ (as in $k = 25$) and progressively refined ($k = 25 \rightarrow 0$). Though initially, the joint anchor from $\hat{H}_k$ failed to select the upper left local PCs around the human neck, the anchor is progressively refined to the ground truth location. Finally, as $k = 0$, the upper left local PCs are considered for more robust local feature extraction. On the other hand, static joint anchors only incorporate $\tilde{H}$, leading to biased local PC selection.

5.4 Effect of Hyper-parameters

Table 7: Parameter sensitivity analysis of the diffusion steps K. We record joint errors by MPJPE in white and PA-MPJPE in gray. Bold is the best and red is the worst.

K	Basic Scenes						Adverse Environment								Average	
	Lab1		Lab2		Furnished		Rain		Smoke		Poor_lighting		Occlusion			
12	61.3	50.23	71.37	66.08	70.52	51.76	76.41	60.43	78.75	63.86	67.84	49.67	71.22	51.89	71.06	56.27
25	59.52	**47.85**	69.36	**61.23**	67.15	**49.19**	**71.03**	58.40	76.92	**62.25**	65.08	**47.47**	**67.47**	49.54	**68.08**	**53.71**
36	**59.21**	48.68	69	63.42	67.75	49.79	72.38	58.95	**76.41**	62.41	65.54	48.37	69.09	51.12	68.48	54.68
50	60.16	48.83	**66.51**	61.39	67.8	50.18	73.44	**57.18**	80.65	63.04	65.38	48.11	68.11	**49.46**	68.86	54.03
60	59.97	48.65	67.61	62.53	**67.07**	49.32	73.12	57.37	80.02	62.73	**65.01**	47.85	68.34	49.62	68.73	54.01

Table 8: Parameter sensitivity analysis of the β scheduling for the diffusion model. We record joint errors by MPJPE in white and PA-MPJPE in gray. C denotes constant β scheduling, L denotes linear β scheduling. Bold is the best and red is the worst.

β scheduling	Basic Scenes						Adverse Environment								Average	
	Lab1		Lab2		Furnished		Rain		Smoke		Poor_lighting		Occlusion			
C: 0.001	**59.37**	48.66	67.24	60.69	65.83	48.65	72.24	57.77	79.82	63.55	64.34	47.1	64.81	48.57	**67.66**	**53.57**
C: 0.002	61.29	49.37	72.98	61.7	66.87	48.43	70.86	57.77	77.47	63.74	63.72	46.76	71.57	51.45	69.25	54.17
L: [0.0001, 0.001]	59.52	**47.85**	69.36	61.23	67.15	49.19	71.03	58.40	76.92	62.25	65.08	47.47	67.47	49.54	68.08	53.71
L: [0.0001, 0.002]	60.87	48.39	66.79	61.74	66.66	48.63	74.92	57.88	80.62	62.68	65.26	47.61	69.45	49.96	69.22	53.84

Table 9: Parameter sensitivity analysis of the limb loss weighting parameter λ for the structural limb-length consistency module. We record joint errors by MPJPE in white and PA-MPJPE in gray. Bold is the best and red is the worst.

λ	Basic Scenes						Adverse Environment								Average	
	Lab1		Lab2		Furnished		Rain		Smoke		Poor_lighting		Occlusion			
2	59.35	47.99	**67.71**	61.97	**66.99**	49.63	73.04	57.85	78.77	**62.06**	**64.99**	47.80	68.29	49.70	68.45	53.86
5	58.65	**47.66**	68.91	62.25	67.86	50.33	71.42	**57.58**	77.19	62.84	65.50	47.86	67.79	**49.15**	68.19	53.95
8	59.47	48.47	68.07	61.90	67.83	49.69	72.34	57.75	78.58	62.72	65.59	48.26	68.68	51.25	68.65	54.29
10	59.52	47.85	69.36	**61.23**	67.15	**49.19**	**71.03**	58.40	**76.92**	62.25	65.08	**47.47**	67.47	49.54	68.08	**53.71**
15	**58.36**	47.75	68.01	62.65	67.80	49.92	71.16	57.87	78.22	63.37	65.49	47.94	**67.00**	49.66	**68.00**	54.17

Effect of diffusion steps K. As shown in Table 7, we present how the joint error is affected by the number of diffusion steps K on mmBody [6]. When diffusion steps $K < 25$, we observe that the increment of K can improve the performance of mmDiff. As with more iterations, mmDiff potentially can handle noisier human poses. However, with the number of diffusion steps $K > 25$, the performance of mmDiff converges. We argue that the pose noise within the coarsely estimated human poses is well handled with 25 diffusion steps.

Effect of β scheduling. As shown in Table 8, we present how the β scheduling affects the diffusion-based HPE. We observe different β scheduling significantly affects the mmDiff performance, as the selected noise scale should accurately approximate the noise contained by the coarsely estimated human poses. For linear scheduling, if the β range is increased (as in range$[0.0001, 0.002]$), the performance drops significantly as the reverse diffusion process is not stable. We also observe that constant β scheduling also performs well, as long as the noise scale is selected properly.

Effect of the weighting parameter λ for $\mathcal{L}_{\text{diff}}$. As shown in Table 9, we present how the model performs with different weighting parameters λ to integrate the limb loss $\left|L - \hat{L}\right|$ into the diffusion loss $\mathcal{L}_{\text{diff}}$. We observe that our model is not sensitive to the selection of λ, with 68.27 ± 0.24 (MPJPE) statistically lower than the performance without spatial limb consistency, 69.16 (MPJPE). We argue that as long as the limb loss is presented, the model can estimate a subject's limb length with acceptable accuracy.

Table 10: Parameter sensitivity analysis of the historical pose sequence length Δt for the temporal motion consistency module. We record joint errors by MPJPE in white and PA-MPJPE in gray. Bold is the best and red is the worst.

Δt	Basic Scenes						Adverse Environment								Average	
	Lab1		Lab2		Furnished		Rain		Smoke		Poor_lighting		Occlusion			
2	60.86	48.54	69.32	62.82	67.72	49.68	72.94	57.85	81.39	63.72	65.19	48.18	69.25	50.99	69.52	54.54
4	59.59	47.77	68.63	62.75	67.11	49.45	73.05	57.59	79.91	63.12	65.50	47.93	68.70	49.80	68.93	54.06
6	59.52	47.85	69.36	**61.23**	67.15	49.19	**71.03**	58.40	**76.92**	62.25	65.08	47.47	67.47	49.54	**68.08**	53.71
8	**58.44**	47.89	68.79	64.02	68.00	50.44	72.15	58.52	77.18	**61.81**	65.86	48.58	**67.12**	50.19	68.22	54.49
10	58.57	**47.51**	**68.05**	62.59	**66.96**	**49.01**	73.10	**57.47**	78.14	62.46	**64.90**	**47.34**	67.74	**49.23**	68.21	**53.66**

Effect of the sequence length Δt. As shown in Table 10, we present how our approach is affected by the number of adjacent time frames used in the temporal motion consistency module. When the sequence length $\Delta t \leq 6$, we observe that the increment of Δt can improve the performance. As with longer sequence lengths, mmDiff potentially can extract more reliable and robust motion patterns of the subjects. However, as the sequence lengths $K > 6$ keep increasing, the performance of mmDiff begins to drop, especially in adverse environments. As the training of mmBody is conducted on basic scenes, the increment of sequence length tends to overfit the model to basic scenes, causing a performance drop in adverse environments. As a result, $\Delta t = 6$ is the optimum for the mmBody dataset.

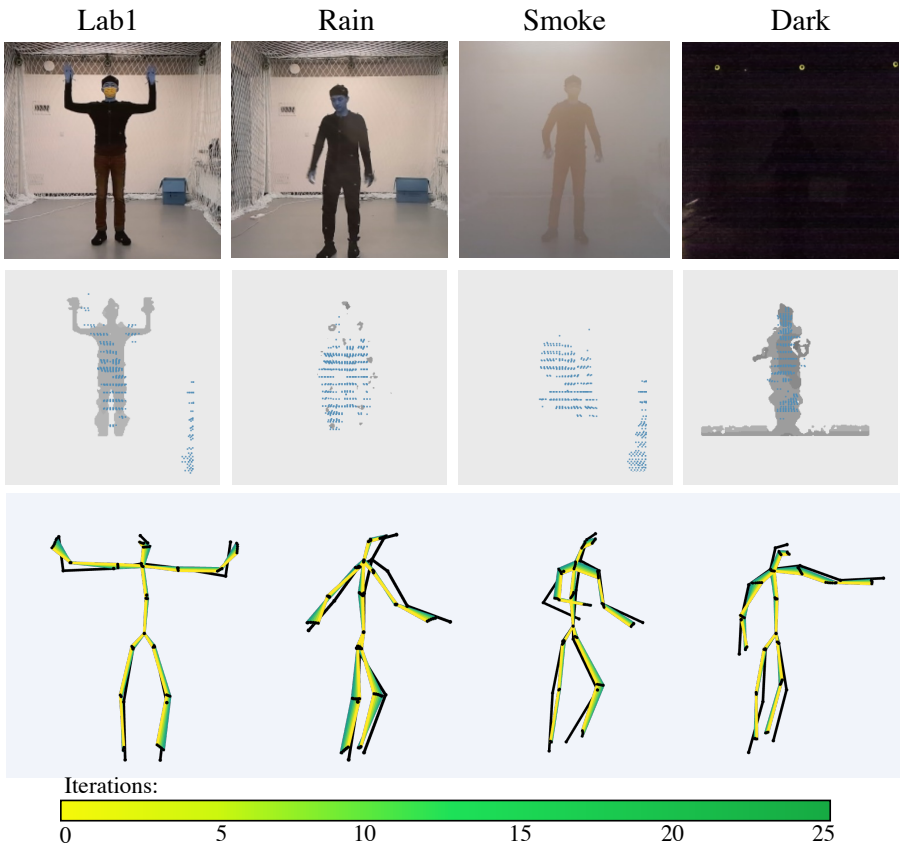


Fig. 8: Visualization of diffusion-based human pose estimation with diffusion 25 steps. We use gradient colors (from yellow to green) to illustrate the refined poses of different iterations. The yellow pose is the initialized coarse human pose, and the green pose is the final refined pose.

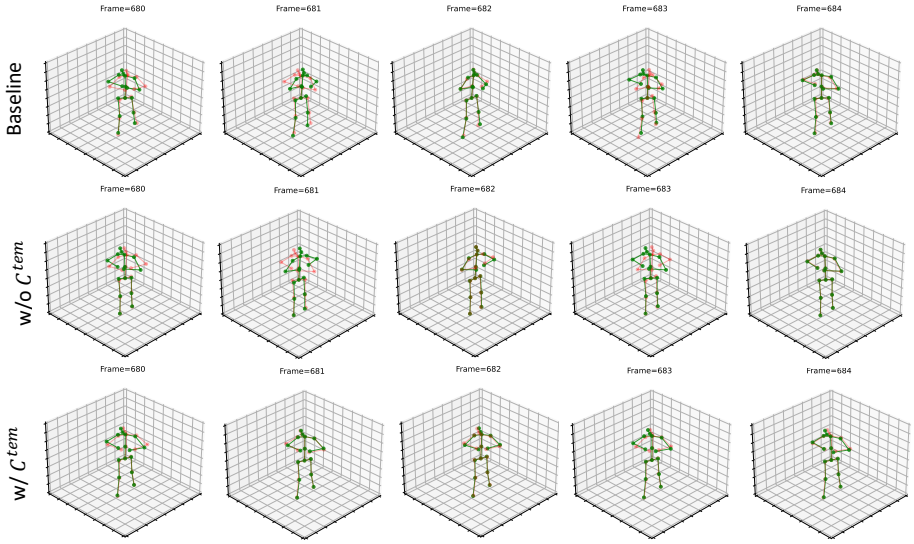


Fig. 9: Visualization of motion continuity on mm-Fi, with PointTransformer baseline [50], ours w/o C^{tem} and ours w/ C^{tem}. Green poses are the ground truth and red poses are the prediction. We can observe occasion erroneous frames are corrected based on smooth motion patterns.

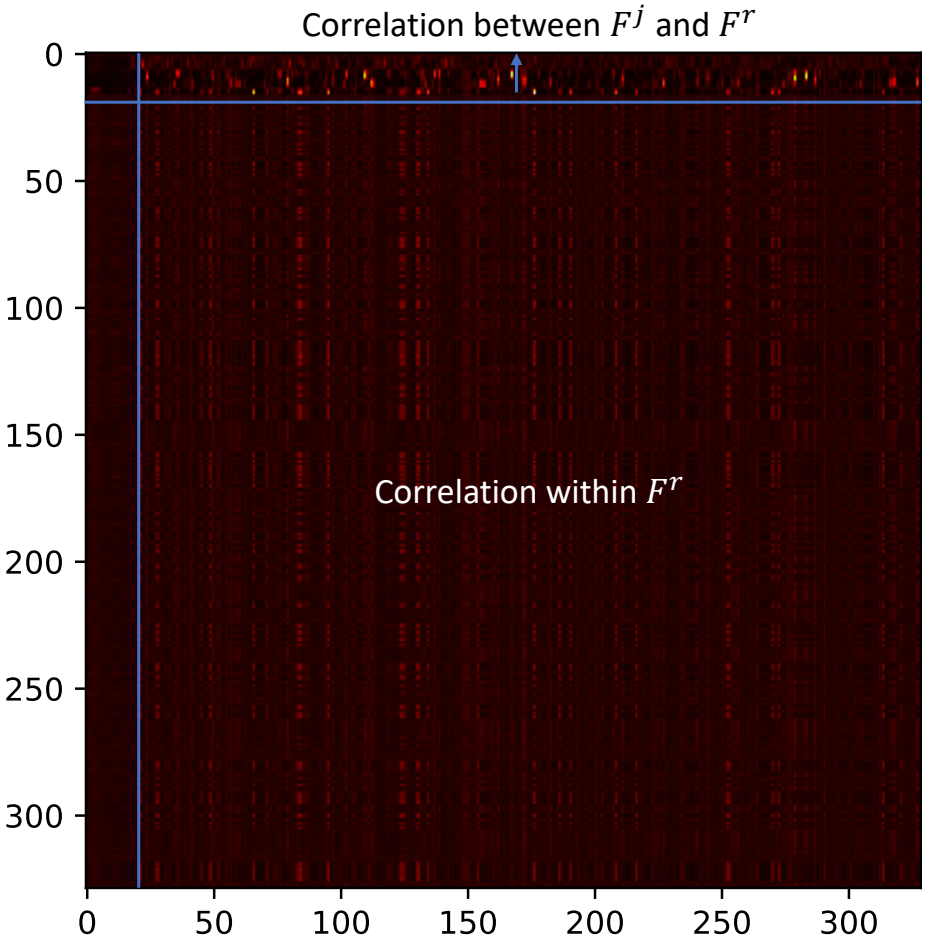


Fig. 10: Visualization of the correlation using the global transformer ϕ^g of GRC tested on mmBody: within PC features F^r, and between joint features F^j and PC features F^r.

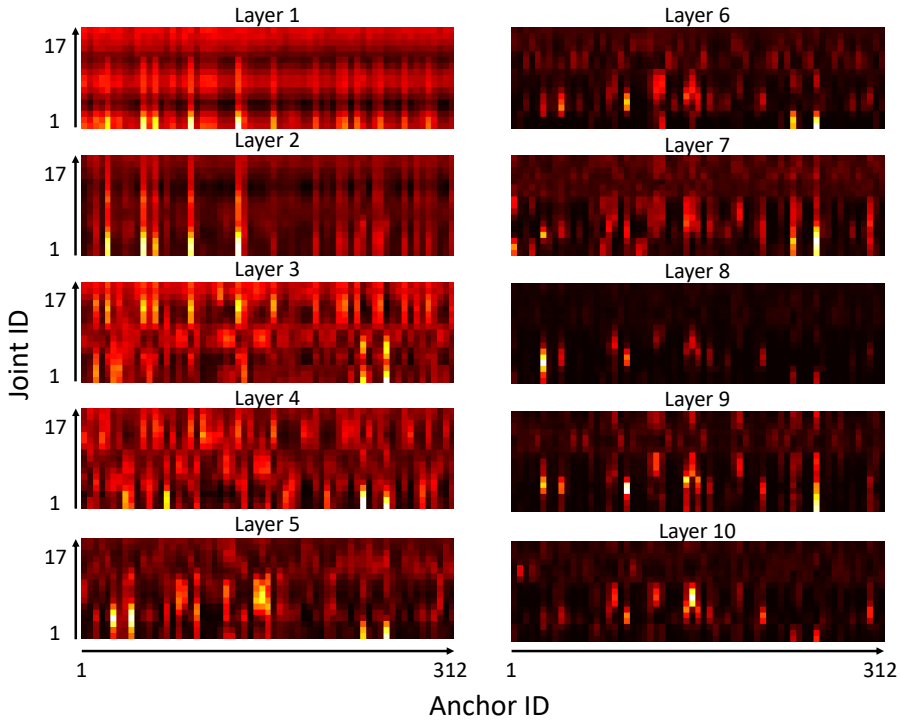


Fig. 11: Visualization of the correlation between joint features F^j and PC features F^r, using the global transformer ϕ^g of GRC tested on mmBody. Feature extraction of different joints depends on individual correlation with the PC feature, which is less affected by other joint features after 5 transformer layers.

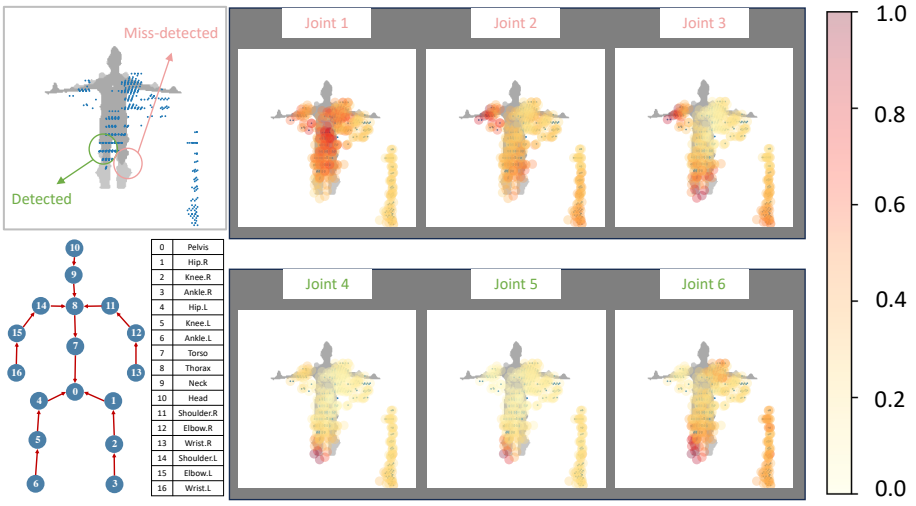


Fig. 12: Visualization of the attention when performing feature extraction of joints 1-6 (legs). Each joint performs individual feature extraction. For detected joints 4-6(left leg), the attention is more concentrated as correctly focuses on the left leg part, extracting more robust features. The feature extraction is more distracted and less reliable for undetected joints 1-3 (right leg) due to miss-detection.

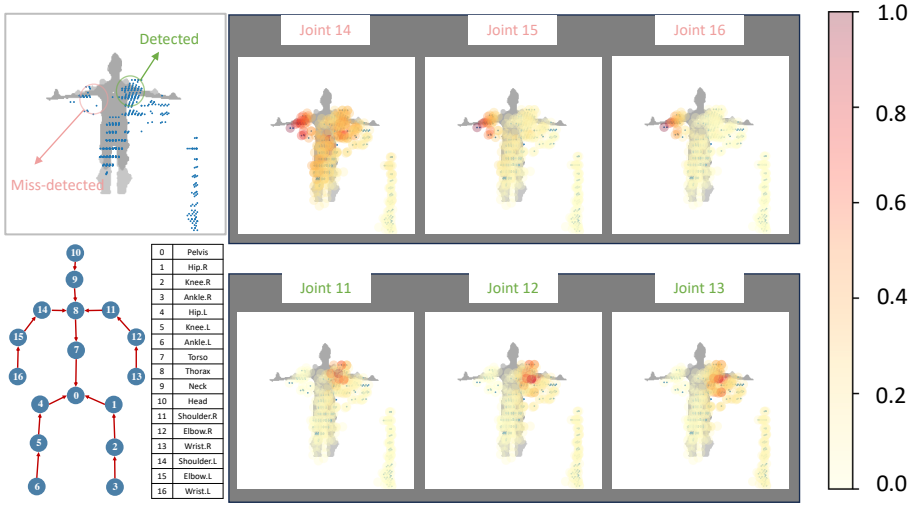


Fig. 13: Visualization of the attention when performing feature extraction of joints 11-16 (arms). As reflected by the attention heatmap, detected joints 11-13(right arm) have more robust features, preventing the influence of undetected joints 14-16 (left arm).

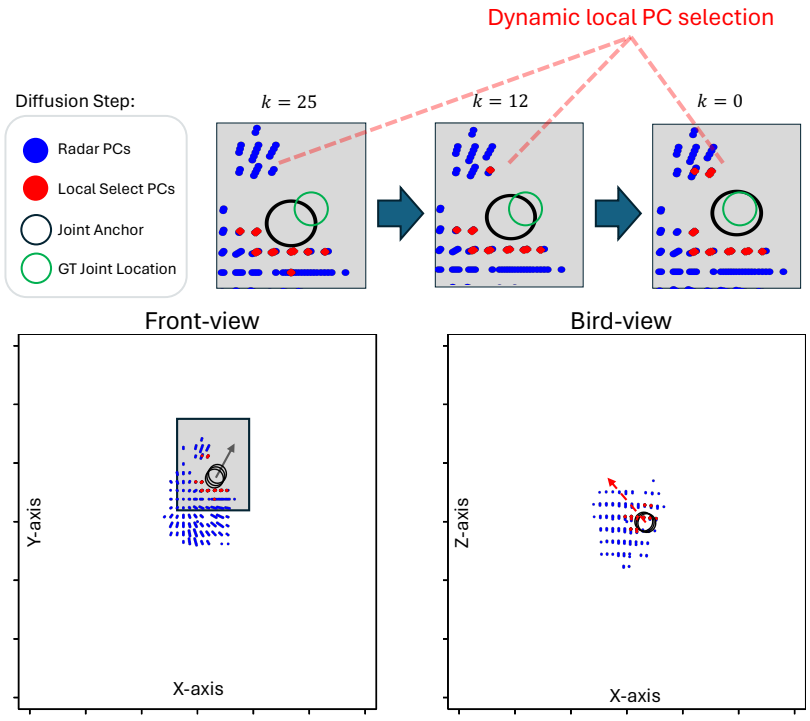


Fig. 14: Visualization of the local PC selection using dynamic joint anchors (right shoulder) during model inference phase. Initially, as $k = 25$, the joint anchor is derived from the coarsely estimated pose $\tilde{H}$, failing to select the upper left local PCs. The dynamic joint-anchor starts from $\tilde{H}$ and is progressively refined with the diffusion steps $k = 25 \rightarrow 0$. Finally, as $k = 0$, the upper left local PCs are selected for more robust local PC features. As the upper left PCs correspond to the human neck, a more robust local feature is extracted considering both shoulder and neck PCs.