PRISM: Human Point Cloud Reconstruction via Skeleton-Guided Diffusion from MmWave Radar

Jiacheng Huang[1*] **Long Tian**[2*] **Yuan Liu**[1†] **Andy W. H. Khong**[1,3]

[1] School of Electrical and Electronic Engineering, Nanyang Technological University
[2] School of Computer Science and Technology, Xidian University
[3] Lee Kong Chian School of Medicine, Nanyang Technological University

jiacheng008@e.ntu.edu.sg tianlong@xidian.edu.cn
{yuan.liu, andykhong}@ntu.edu.sg

Abstract

Millimeter-wave (mmWave) radar is an attractive modality for human sensing, offering all-weather, non-contact, and privacy-preserving perception. However, its inherent sparsity severely limits downstream human-centric understanding, and there is currently no public dataset that provides paired quantitative benchmarks for dense human point cloud reconstruction under occlusion. We present MIST, the first mmWave human point cloud dataset with paired quantitative benchmarks for through-obstacle dense human reconstruction. MIST captures synchronized clear-view and through-obstacle observations of the same subjects performing identical actions, separated by a physical barrier, and provides LiDAR-based 3D ground truth. We further incorporate a three-level occlusion design that converts attenuation severity into a controlled variable, covering fifteen subjects, thirty action classes, and 270 k paired frames. To establish a baseline on MIST and address the limitations of point cloud completion under extreme sparsity and human motion, we propose PRISM, a skeleton-guided conditional latent diffusion framework for reconstructing dense human point clouds from sparse mmWave radar. Three conditioning streams—skeleton joint positions, coarse body geometry, and action class embeddings—guide a Transformer denoiser in the latent space of a point cloud VAE. On MIST, PRISM achieves a COV-CD of 0.362, significantly outperforming the strongest completion baseline (0.113). Notably, PRISM maintains structurally coherent reconstruction even under severe through-obstacle attenuation at approximately 50 points per frame, whereas all baselines collapse to fixed-template outputs.

1 Introduction

Millimeter-wave (mmWave) radar has emerged as an increasingly attractive modality for non-intrusive human sensing, operating reliably under complete darkness and adverse weather without physical contact or line-of-sight constraints [1, 2]. These properties have enabled its adoption across a wide range of body-centric tasks, including human activity recognition [3, 4, 5], skeletal pose estimation [6, 7, 8, 9], and multi-person localization and tracking [10, 11, 12]. Its point cloud output carries no photorealistic record of the subject's appearance, making it particularly well-suited for healthcare monitoring [13, 14] and other privacy-sensitive scenarios [15].

LiDAR and RGB-D cameras produce dense, geometrically accurate reconstructions but cannot penetrate solid obstacles, making them unsuitable for through-wall human monitoring. Although

*Equal contribution.
†Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

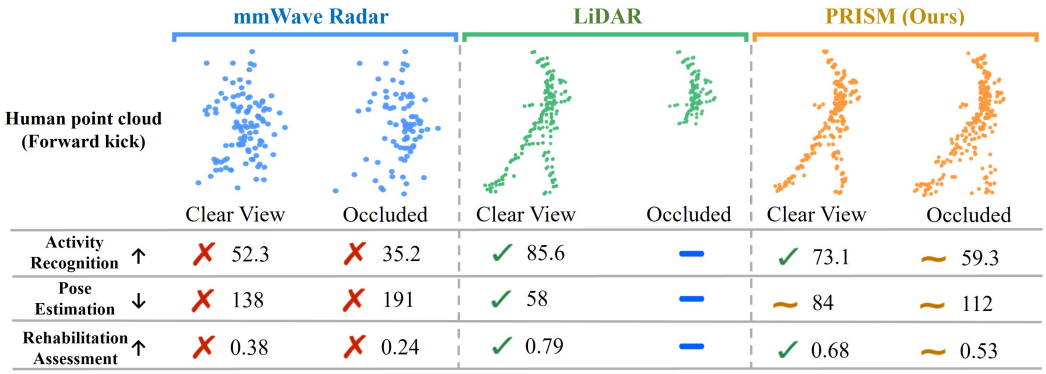


Figure 1: Human point cloud quality across three sensing modalities under clear-view and through-obstacle conditions with a forward kick action in which the kicking leg is partially occluded.

mmWave radar can penetrate non-metallic barriers, its point cloud density remains limited by factors such as finite antenna aperture, bandwidth, signal-to-noise ratio, and the sparsity of human reflections, typically yielding only 50–150 points per frame. To address the sparsity problem, underdetermined reconstruction can become tractable when intrinsic source structure is explicitly exploited through autoregressive temporal modeling [16]. As shown in Figure 1, occlusion further degrades mmWave point cloud quality, while the kicking leg region disappears entirely from the LiDAR scan. This combination of through-obstacle capability and inherent sparsity introduces a dual challenge that remains unaddressed by existing datasets and methods. Notwithstanding the above, all existing public mmWave human datasets [1, 17, 18] provide skeleton or mesh annotations while assuming unobstructed sensor placement. Datasets that include an occlusion scene [19, 20] involves covering the sensor surface with materials rather than placing a physical barrier between the sensor and subject, precluding any simulation of real through-wall deployment and providing no paired clear-view reference. In terms of the approach, general-purpose point cloud completion methods [21, 22, 23] are designed for dense partial observations of rigid synthetic objects and do not model the sequential evolution of body geometry across an action.

We present MIST that allows for the systematic evaluation of through-obstacle mmWave human reconstruction. We facilitate this by pairing clear-view and occluded observations under controlled barrier conditions with synchronized LiDAR providing geometric supervision. To address reconstruction under such extreme sparsity, PRISM leverages radar-aligned joint representations inferred from sparse returns as structural priors within a conditional latent diffusion framework, guiding the denoising process toward temporally consistent and geometrically coherent human point clouds. As illustrated in Figure 1, PRISM reconstructs clear and structured human point clouds under both clear-view and through-obstacle conditions, faithfully recovering the kicking motion and approaching LiDAR reference. PRISM substantially outperforms raw mmWave input across three downstream healthcare monitoring tasks, where activity recognition accuracy improves from 52.3% to 73.1%, pose estimation mean per joint position error (MPJPE) decreases from 138 mm to 84 mm, and rehabilitation assessment correlation increases from 0.38 to 0.68. Our contributions are as follows:

- **MIST** is the first mmWave frequency-modulated continuous wave (FMCW) human point cloud dataset featuring a physical barrier between the sensor and the subject, providing paired clear-view and through-obstacle observations with synchronized overhead LiDAR as dense ground truth. A three-level occlusion design converts attenuation severity into a controlled experimental variable. MIST spans fifteen subjects and thirty action classes, including rehabilitation and daily activities, with synchronized LiDAR, depth, and mmWave streams for multi-modal research.
- **PRISM** is a skeleton-guided conditional diffusion framework that reconstructs dense human point clouds from sparse mmWave radar. PRISM incorporates a skeleton keypoint detector SkelKD that first extracts radar-aligned joint positions before a Coarse Generator produces stable coarse body geometry robust to extreme input sparsity. Together with action class embeddings, these conditioning signals guide a Diffusion Denoiser under a denoising diffusion probabilistic model (DDPM) framework, constraining reconstruction to follow the kinematic evolution of each action.

- **Downstream task benchmarks.** Using the reconstructed sequences as input, we benchmarked PRISM against existing models on the MIST dataset for human action recognition, skeletal pose estimation, and rehabilitation exercise quality assessment. These benchmarks demonstrate that high-quality reconstruction yields substantial and consistent gains in performance across the full spectrum of body analysis required for real-world healthcare monitoring.

2 Related work

mmWave-based human sensing datasets. Existing mmWave human datasets are collected under obstacle-free conditions. MM-Fi [1] provides synchronized multi-modal streams and M4Human [18] acquires high-fidelity mesh annotations via marker-based motion capture—both datasets were collected in the absence of a physical barrier between sensor and subject. While mmBody [19] includes an occlusion scene, the sensor surface is covered with materials with no paired clear-view reference and no simulation of through-wall deployment. This systematic gap leaves through-obstacle deployment without any quantitative benchmark. More broadly, wave-based localization often requires extracting spatial information from measurements distorted by the propagation medium. Dispersion-aware localization on solid surfaces has shown that such distortions can be explicitly modeled to recover source position [24]. In contrast, PRISM operates on sparse radar-derived point clouds and learns structural human priors for dense geometric reconstruction.

Point cloud completion and human body reconstruction. Completion methods such as PCN [21] and PoinTr [22] have been designed for dense partial scans of rigid synthetic objects—under the extreme sparsity and non-rigid articulation of human radar returns, their feature aggregation collapses to semantically empty representations. Generative models such as LION [25] and LDT [26] operate in hierarchical variational autoencoder (VAE) [27] latent spaces with per-step conditioning for expressive 3D generation. However, both approaches are conditioned on global latent codes or dense spatial features extracted from the input rather than structured anatomical priors. While mmPoint [28] recovers dense human point clouds from sparse mmWave signals via deterministic regression, it generates a fixed average-body template under extreme sparsity, leaving action categories indistinguishable across frames. We note that none of the above approaches incorporates joint-level skeletal topology as a structural conditioning signal. In addition, these models are not designed for action-consistent generation from streams of extremely sparse radar observations through physical barriers.

Downstream task analysis. MiliPoint [29] establishes action classification benchmarks on sparse mmWave radar point clouds and temporal methods such as mmChainPose [30] perform skeletal pose estimation by propagating joint geometry cues across frames. Both approaches operate under obstacle-free sensor placement and are evaluated on raw radar returns rather than reconstructed dense clouds. For rehabilitation exercise quality assessment, KIMORE [31] offers kinematic scoring benchmarks derived from RGB-D sequences of patients performing standardised exercises, demonstrating that precise joint trajectory estimates translate to clinically meaningful motion quality scores. No existing benchmark evaluates activity recognition, pose estimation, or rehabilitation exercise quality from point clouds reconstructed under through-obstacle sensor conditions, leaving the downstream utility of through-wall radar sensing without direct experimental validation.

3 The MIST Dataset

All publicly available mmWave human datasets are collected under obstacle-free conditions, leaving a critical evaluation gap for through-obstacle deployment. MIST addresses this by placing a physical barrier between the sensor and a subject, providing the first paired quantitative benchmark for through-obstacle dense human point cloud reconstruction. Three occlusion levels (OLs) are defined by the cumulative thickness of 18 mm wooden boards placed between the right-side radar and the subject, where OL0 corresponds to a clear line-of-sight without any boards, OL1 with one 18 mm board being interposed, and OL2 with two boards being stacked to a total thickness of 36 mm. The 18 mm board material matches the signal attenuation characteristics of a typical indoor partition while allowing reproducible and incremental thickness control. A detailed comparison with existing mmWave human datasets is provided in Appendix A.1.

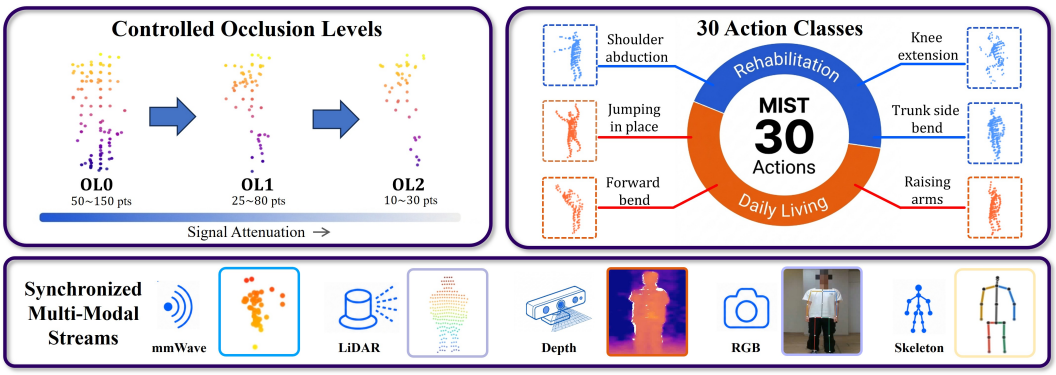


Figure 2: Overview of MIST, comprising three occlusion levels, thirty action classes spanning rehabilitation and daily-living motions, and five sensing modalities providing ground truth.

Sensor configuration. Two TI IWR6843 FMCW radars are mounted at 1.4 m height, one providing a clear-view path for OL0 capture and one operating through the barrier at OL1 or OL2. Collecting both streams simultaneously ensures that every paired frame is captured at an identical subject pose. An overhead Livox Mid-360 LiDAR and a door-side Intel RealSense D435i depth camera (extrinsic-calibrated to a shared world coordinate frame) provide dense point clouds and 17-joint H36M skeleton ground truth [32]. Skeleton annotations are extracted at 30 fps and resampled to 10 fps to match the radar frame rate. Sensor specifications are listed in Appendix A.2.

Occlusion levels and statistics. MIST defines three occlusion levels by cumulative board thickness. OL0 provides 50–150 points per frame and serves as the primary training and evaluation split. OL1, with a single board, attenuates signal returns to 25–80 points per frame, and OL2 further reduces returns to 10–30 points per frame due to signal attenuation from the two stacked boards. The approximately five-fold density reduction from OL0 to OL2 causes FPS-based feature aggregation and all completion baselines to produce degenerated outputs, making OL1 and OL2 a stress test for density-invariant reconstruction. MIST covers fifteen subjects performing thirty action classes across three occlusion levels; each action is performed repeatedly for 30 s in every recording, giving 300 frames at 10 fps and approximately 270 k paired frames in total.

Action categories. The thirty action classes span two functional categories. The rehabilitation category R01 to R14 covers fourteen classes targeting upper and lower limb mobility, balance, and gait assessment relevant to clinical motion screening. The daily-living category D01 to D16 covers sixteen classes of whole-body movements frequent in elderly-care monitoring, including D16 Forward Fall, which directly supports fall detection as a downstream task. Of the thirty classes, eight share equivalent kinematic categories with MM-Fi, enabling direct cross-dataset transfer evaluation, and the complete action list is provided in Appendix A.3.

4 Method

We propose PRISM, a new skeleton-guided multimodal diffusion model, to verify the effectiveness and robustness of human point cloud reconstruction from mmWave radar based on our proposed MIST benchmark for healthcare downstream applications. During training, PRISM accepts multimodal inputs including mmWave, LiDAR, and depth camera, and learns a jointly aligned multimodal latent space for enhanced human point cloud reconstruction. Since direct multimodal fusion can amplify modality-specific undesirable noise [33], PRISM instead assigns complementary roles to each sensing stream, using LiDAR for dense geometric supervision, depth sensing for skeletal supervision, and mmWave as the deployment modality. At test time, PRISM takes the mmWave point cloud and the action label of the recording as input, under various degrees of occlusion.

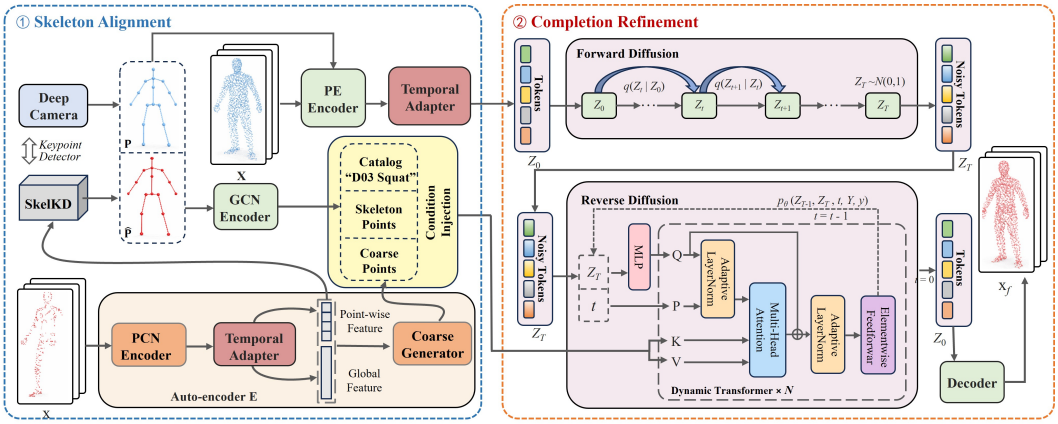


Figure 3: Overview of the proposed PRISM model consisting of a skeleton alignment stage and a completion refinement stage. In the skeleton alignment stage, we extract and align multimodal features from LiDAR point clouds (dense), depth camera skeletons, and mmWave point clouds (sparse). These features serve as clean tokens and conditional information for the completion refinement stage. Notably, we design SkelKD and Temporal Adapter modules to align skeleton predictions based on mmWave point clouds and fuse features across multiple time steps, respectively. In the completion refinement stage, we design a conditional DDPM equipped with clean tokens and conditional information from the previous stage to reconstruct enhanced point clouds, especially when the mmWave point clouds are partially occluded at test time.

4.1 Skeleton Alignment

With reference to Fig. 3, PRISM consists of skeleton alignment stage and completion refinement stage. The skeleton alignment stage primary focuses on extracting, fusing and aligning multimodal features of input sequences from non-occluded LiDAR point clouds $\boldsymbol{x}_{\text{lidar}} \in \mathbb{R}^{t\times m\times 3}$ and deep camera skeletons $\boldsymbol{x}_{\text{deep}} \in \mathbb{R}^{t\times p\times 3}$, as well as partially-occluded mmWave point clouds $\boldsymbol{x}_{\text{mmwave}} \in \mathbb{R}^{t\times n\times 3}$. Here, t denotes the number of frames in the point cloud sequence, m and n are the number of points per frame, and p is the number of skeleton keypoints. In practice, $m > n \gg p$. The temporal-fused LiDAR point cloud features, sampled under the guidance of deep camera skeletons, can be expressed as

$$\boldsymbol{h}^{\text{TA}}_{\text{lidar}} = f_{\boldsymbol{\theta}_{\text{TA}}}(\boldsymbol{h}^{\text{PE}}_{\text{lidar}}), \tag{1}$$

$$\boldsymbol{h}^{\text{PE}}_{\text{lidar}} = f_{\boldsymbol{\theta}_{\text{PE}}}(\boldsymbol{x}_{\text{lidar}}, \boldsymbol{x}_{\text{deep}}), \tag{2}$$

where $\boldsymbol{\theta}_{\text{PE}}$ and $\boldsymbol{\theta}_{\text{TA}}$ denote, respectively, the parameters of the PE encoder [21] and the proposed Temporal Adapter module. We propose SkelKD (see details in Appendix B) to predict skeleton keypoints $\boldsymbol{x}_{\text{skelKD}} \in \mathbb{R}^{t\times p\times 3}$ from the pointwise features $\boldsymbol{h}^{\text{point}}_{\text{mmwave}}$ associated with the mmWave point clouds $\boldsymbol{x}_{\text{mmwave}}$. These features are extracted by a cascaded PCN encoder [34] followed by our proposed Temporal Adapter module as $\boldsymbol{x}_{\text{skelKD}} = f_{\boldsymbol{\theta}_{\text{seklKD}}}(\boldsymbol{h}^{\text{point}}_{\text{mmwave}})$, $\boldsymbol{h}^{\text{point}}_{\text{mmwave}}, \boldsymbol{h}^{\text{global}}_{\text{mmwave}} = f_{\boldsymbol{\theta}_{\text{PCN}}}(\boldsymbol{x}_{\text{mmwave}})$, where $\boldsymbol{\theta}_{\text{skelKD}}$ and $\boldsymbol{\theta}_{\text{PCN}}$ are, respectively, the parameters of PCN encoder and our Temporal Adapter. Meanwhile, we use a graph convolutional network (GCN) encoder [35] parameterized by $\boldsymbol{\theta}_{\text{GCN}}$ to further aggregate the spatial information of the predicted skeleton keypoints $\boldsymbol{x}_{\text{skelKD}}$, producing enhanced skeleton points $\boldsymbol{x}^{\text{enhance}}_{\text{skelKD}} = f_{\boldsymbol{\theta}_{\text{GCN}}}(\boldsymbol{x}_{\text{skelKD}})$ of the same shape as $\boldsymbol{x}_{\text{skelKD}}$. We then use a Coarse Generator [34] to obtain the enhanced (dense but occluded) coarse point clouds given the concatenation of the pointwise features and global features, denoted by

$$\boldsymbol{x}^{\text{dense}}_{\text{mmwave}} = f_{\boldsymbol{\theta}_{\text{CG}}}([\boldsymbol{h}^{\text{point}}_{\text{mmwave}}, \boldsymbol{h}^{\text{global}}_{\text{mmwave}}]), \tag{3}$$

where $\boldsymbol{\theta}_{\text{CG}}$ is the parameters of Coarse Generator. Finally, the condition injection is constructed by concatenating action category $\boldsymbol{y}_{\text{action}}$ (e.g., `D03 Squat`), the enhanced skeleton points $\boldsymbol{x}^{\text{enhance}}_{\text{skelKD}}$, and the enhanced coarse points $\boldsymbol{x}^{\text{dense}}_{\text{mmwave}}$. The action category is the ground-truth label of each recording and is provided at both training and test time; Section 5.3 shows that removing it lowers COV-CD only from 0.362 to 0.351.

Temporal Adapter. Similar to the Temporal Adapter being applied after the PE encoder, the Temporal Adapter is applied after the PCN encoder. Specifically, we obtain temporally fused features $\boldsymbol{h}^{\text{TA}}_{\text{lidar}}$ by

introducing a Transformer encoder layer parameterized by $\boldsymbol{\theta}_{\mathrm{TA}}$ on top of the frame-level features $\boldsymbol{h}_{\mathrm{lidar}}^{\mathrm{PE}} \in \mathbb{R}^{t\times q}$ from the PE Encoder, where $q = m \times d$ and d is the feature dimension. Self-attention is applied along the temporal dimension [36].

SkelKD. SkelKD regresses radar-aligned skeleton points $\boldsymbol{x}_{\mathrm{skelKD}}$ using PointNet [37], a lightweight CNN-based feature extractor, from sparse mmWave point clouds $\boldsymbol{x}_{\mathrm{mmwave}}$ with deep camera supervision only. To avoid nearest-neighbor collapse under occlusion, SkelKD uses p learnable joint query embeddings as cross-attention queries over per-point features extracted by a per-point multi-layer perceptron (MLP), with symmetric max-pooling over all N returns yielding a global silhouette descriptor $\boldsymbol{h}_{\mathrm{global}} \in \mathbb{R}^{B\times 256}$. Each query embedding attends over the full point set via scaled dot-product attention $\boldsymbol{f}_{\mathrm{joint}} = \mathrm{softmax}(QK^{\top}/\sqrt{d})\,V$ with $Q \in \mathbb{R}^{p\times 128}$ being the query and $K, V \in \mathbb{R}^{N\times 128}$ being the key and the value, respectively. Occluded joints therefore spread their response across all visible returns rather than anchoring to a single noisy neighbor. The concatenated joint attention features and broadcast global descriptor are then mapped through a residual MLP regression head [38] to per-joint 3-D coordinates $\boldsymbol{x}_{\mathrm{skelKD}} \in \mathbb{R}^{t\times p\times 3}$.

4.2 Completion Refinement

The Completion Refinement phase in Fig. 3 is responsible for recovering clean point clouds from the conditional injection derived from occluded and sparse mmWave inputs. It consists of a forward diffusion process that progressively adds noise to clean tokens (obtained from non-occluded, dense LiDAR point clouds) to produce noisy tokens, and a backward diffusion process that denoises the noisy tokens using conditional information extracted from the mmWave point clouds.

Forward Diffusion. Given the clean tokens of LiDAR point cloud features $\boldsymbol{z}_0 = \boldsymbol{h}_{\mathrm{lidar}}^{\mathrm{TA}}$, we apply the forward process of denoising diffusion probabilistic models (DDPM) [39, 40] to convert them into Gaussian noise $\boldsymbol{z}_T$ over T time steps as

$$q(\boldsymbol{z}_{1:T}|\boldsymbol{z}_0) = \prod_{t=1}^{T} q(\boldsymbol{z}_t|\boldsymbol{z}_{t-1}), \quad q(\boldsymbol{z}_t|\boldsymbol{z}_{t-1}) = \mathcal{N}(\sqrt{1-\beta_t}\boldsymbol{z}_{t-1}, \beta_t\mathbf{I}), \tag{4}$$

where β_t is the noise level that is typically set as a small constant. The forward process can be recursively derived as $q(\boldsymbol{z}_t|\boldsymbol{z}_0) = \mathcal{N}(\sqrt{\alpha_t}\boldsymbol{z}_0, (1-\alpha_t)\mathbf{I})$, where $\alpha_t = \prod_{i=1}^{t}(1-\beta_i)$.

Backward Diffusion. Given the condition injection $\boldsymbol{c} = \mathrm{EMB}(\boldsymbol{y}_{\mathrm{action}}, \boldsymbol{x}_{\mathrm{skelKD}}^{\mathrm{enhance}}, \boldsymbol{x}_{\mathrm{mmwave}}^{\mathrm{dense}})$ derived from the embeddings of action category, skeleton points, and coarse points as illustrated in Fig. 3, we can gradually denoise from $\boldsymbol{z}_T$ conditioned on $\boldsymbol{c}$ as

$$p_{\boldsymbol{\Theta}_{\mathrm{CR}}}(\boldsymbol{z}_{0:T}) = p_{\boldsymbol{\Theta}_{\mathrm{CR}}}(\boldsymbol{z}_T)\prod_{t=1}^{T} p_{\boldsymbol{\Theta}_{\mathrm{CR}}}(\boldsymbol{z}_{t-1}|\boldsymbol{z}_t),\ p_{\boldsymbol{\Theta}_{\mathrm{CR}}}(\boldsymbol{z}_{t-1}|\boldsymbol{z}_t) = \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{\Theta}_{\mathrm{CR}}}(\boldsymbol{z}_t, \boldsymbol{c}, t), \beta_t\mathbf{I}). \tag{5}$$

Here, $\boldsymbol{\Theta}_{\mathrm{CG}}$ are parameters of the conditional diffusion model in the Completion Refinement stage and $\boldsymbol{\mu}_{\boldsymbol{\Theta}_{\mathrm{CR}}}(\boldsymbol{z}_t, \boldsymbol{c}, t) = \frac{1}{\sqrt{1-\beta_t}}(\boldsymbol{z}_t - \frac{\beta_t}{\sqrt{1-\alpha_t}}\boldsymbol{\epsilon}_{\boldsymbol{\Theta}_{\mathrm{CR}}}(\boldsymbol{z}_t, \boldsymbol{c}, t))$. The forward process provides ground truth signals for training the parameters in the backward process.

Training and Inference: We propose a skeleton-regularized noise prediction objective to train the model parameters $\boldsymbol{\Theta} = \{\boldsymbol{\Theta}_{\mathrm{SA}}, \boldsymbol{\Theta}_{\mathrm{CR}}\}$ as

$$\mathcal{L} = \min_{\boldsymbol{\Theta}} \mathbb{E}_{t, \boldsymbol{z}_0, \boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\mathbf{1})} \|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_{\boldsymbol{\Theta}_{\mathrm{CR}}}\|_2^2 + \lambda\|\boldsymbol{x}_{\mathrm{deep}} - \boldsymbol{x}_{\mathrm{skelKD}}\|_2^2, \tag{6}$$

where $\boldsymbol{\Theta}_{\mathrm{SA}} = \{\boldsymbol{\theta}_{\mathrm{PE}}, \boldsymbol{\theta}_{\mathrm{TA}}, \boldsymbol{\theta}_{\mathrm{GCN}}, \boldsymbol{\theta}_{\mathrm{PCN}}, \boldsymbol{\theta}_{\mathrm{CG}}\}$ denote the parameters in Skeleton Alignment stage. The first term in Eq. 6 aims to recover dense, non-occluded human point clouds from sparse, partially occluded mmWave inputs. The second term aligns the skeletons predicted from mmWave point clouds with those from the depth camera.

5 Experiments

5.1 Experiment Setup

Data was collected within a controlled indoor laboratory with approximately 5 m $\times$ 6 m usable floor area, as illustrated in Figure 4. Subjects performed all actions at 2 to 3 m distance from the sensor

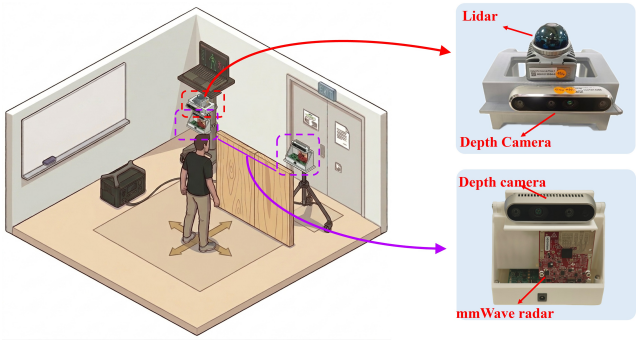


Figure 4: MIST setup showing no board for OL0, one 18 mm board for OL1, and two 18 mm boards for OL2 are placed between the subject and the right-side mmWave radar, while the left-side LiDAR, mmWave radar, and depth camera remain unoccluded to provide ground truth.

Table 1: Per-action point cloud quality on MIST-OL0 (COV-CD↑, MMD-CD↓ scaled $\times 10^{-3}$).

Method	Type	*Squat*		*Wave Hand*		*Kick*		*All*	
		COV	MMD	COV	MMD	COV	MMD	COV	MMD
PCN [21]	Compl.	0.079	9.16	0.088	8.92	0.082	9.04	0.083	9.01
PoinTr [22]	Compl.	0.108	8.37	0.121	8.14	0.114	8.25	0.113	8.26
SnowflakeNet [23]	Compl.	0.071	10.86	0.083	10.61	0.076	10.74	0.079	10.72
LAKe-Net [35]	Compl.	0.080	8.78	0.091	8.57	0.083	8.69	0.085	8.69
mmPoint [28]	Det.	0.090	9.03	0.103	8.79	0.096	8.91	0.097	8.91
LION [25]	Gen.	0.173	7.49	0.192	7.27	0.183	7.38	0.185	7.37
TIGER [43]	Gen.	0.197	7.21	0.219	6.97	0.209	7.11	0.209	7.08
PRISM (ours)	Gen.	0.342	5.19	0.392	4.62	0.357	4.98	0.362	4.83

array. The LiDAR, mmWave radar, and depth camera remained unoccluded throughout, collecting complete unoccluded captures as dense reconstruction ground truth and ensuring that ground truth quality is unaffected by board placement. Three datasets OL0, OL1, and OL2 described in Section 3 and summarized in Table 3 were collected. All sensors were extrinsically-calibrated to a shared global coordinate frame, with calibration parameters held fixed across all subjects and sessions.

Training and evaluation protocol. Subjects S01 to S06 form the training split while S07 and S08 were held-out as evaluation split; the remaining seven of the fifteen subjects are not used in our experiments. All three occlusion levels were evaluated without obstacle-specific adaptation, providing a zero-shot test of density-invariant generalization. Training was performed on four sequential stages of SkelKD, Temporal Adapter, joint optimization of the Diffusion Denoiser with Condition Network and Coarse Generator, and end-to-end fine-tuning via the Adam optimizer [41] with cosine learning-rate decay. The full set of hyperparameters are listed in Appendix B.2. Cross-dataset generalization results, where all models are trained on MIST and evaluated directly on MM-Fi [1] using the eight shared action classes, are reported in Appendix C.1.

5.2 Quantitative results on MIST

Evaluation metrics. We adopted two set-level distribution metrics from [42]. In particular, COV-CD computes the fraction of ground-truth shapes matched by at least one generated sample, rewarding broad coverage of the target distribution. MMD-CD, on the other hand, determines the average distance from each ground-truth shape to its nearest generated counterpart, rewarding distributional fidelity. Results are reported per action and averaged over all thirty action classes.

Results. Table 1 shows that all seven baselines fail to generate action-discriminative reconstructions, though for different reasons. The four deterministic completion backbones cluster tightly between

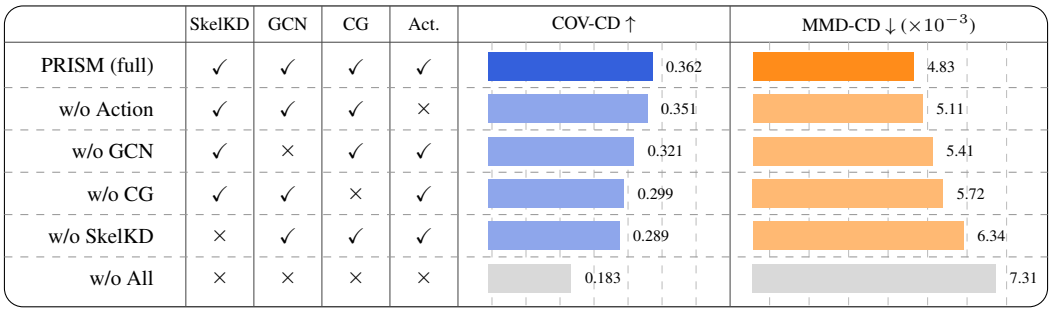


Figure 5: Component ablation on MIST-OL0. COV-CD and MMD-CD for each conditioning module.

COV-CD 0.079 and 0.113, confirming that the ceiling is set by the absence of a body prior rather than by model capacity. SnowflakeNet suffers from lower performance than PCN because its multi-stage farthest-point sampling compounds density sensitivity at each upsampling level under extreme sparsity. mmPoint achieves a COV-CD of 0.097, barely above the completion cluster since single-frame deterministic regression cannot resolve pose ambiguity from a sparse, action-agnostic radar input. LION and TIGER fare noticeably better by sampling from a learned distribution, achieving COV-CD 0.185 and 0.209, respectively. Their remaining gap relative to PRISM reveals the limitation of distribution diversity without structural grounding, as the generated shapes cover the marginal body distribution but are not steered toward any particular action.

We also note that PRISM achieves a COV-CD of 0.362 and an MMD-CD of 4.83×10^{-3}, representing a 3.2 times improvement in coverage over PoinTr and a 41.5% reduction in matching distance, with consistent action discrimination across Squat, Wave Hand, and Kick. Qualitative visualisations of cross-action comparisons, temporal reconstruction, and the reverse diffusion process are provided in Appendix D.

5.3 Ablation study

Figure 5 illustrates the contribution of each conditioning module on the MIST-OL0 validation split. We observe that SkelKD is the dominant component—its removal reduces COV-CD from 0.362 to 0.289 and increases MMD-CD to 6.34, the largest single-module degradation. The Coarse Generator and GCN each contribute meaningfully, reducing COV-CD to 0.299 and 0.321, respectively, when removed, confirming that stable coarse geometry and kinematic-chain propagation are both necessary for full-body coverage. The action label has the smallest individual effect yet remains relevant for actions sharing similar instantaneous joint configurations. Removing all three streams collapses COV-CD to 0.183, on par with unconditional LION, implying that the conditioning stack serves as a discriminator for the reconstruction process. Keeping the full conditioning stack but replacing the diffusion denoiser with a deterministic latent regressor of identical capacity costs a comparable amount of coverage, so neither component is sufficient on its own; this condition-matched comparison is reported in Appendix C.2. Through-obstacle robustness on OL1 and OL2 is reported in Appendix C.3.

5.4 Downstream task evaluation

We evaluate downstream utility across three tasks using a shared PointNet++ backbone [37, 44] with lightweight task-specific heads, comparing raw mmWave input against PoinTr, TIGER, and PRISM across OL0 to OL2. Fall detection instead adopts an existing real-time mmWave detector applied unchanged to each reconstructed stream, with D16 Forward Fall as the positive class. As a streaming module rather than a per-frame classifier, it is scored over whole held-out sequences.

Figure 6 summarises performance across five evaluation axes: HAR accuracy at OL0, fall detection F1, pose estimation quality, rehabilitation assessment correlation, and robustness at OL2. An unoccluded LiDAR oracle forms the outer ring as the modality upper bound. We note from the results that PRISM leads all reconstruction baselines on every axis, and the advantage over raw mmWave widens consistently as occlusion severity increases.

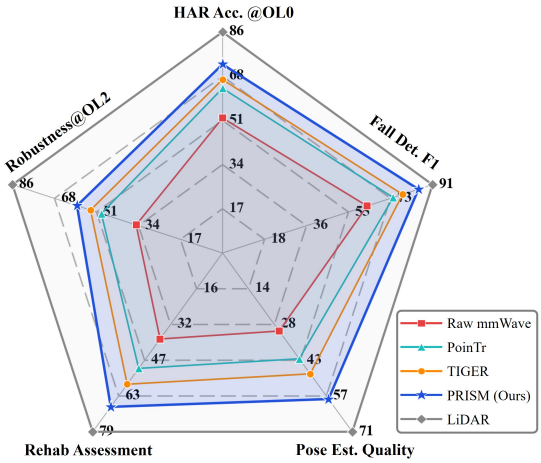


Figure 6: Downstream task performance radar chart across five axes. LiDAR Oracle (grey) defines the outer ring.

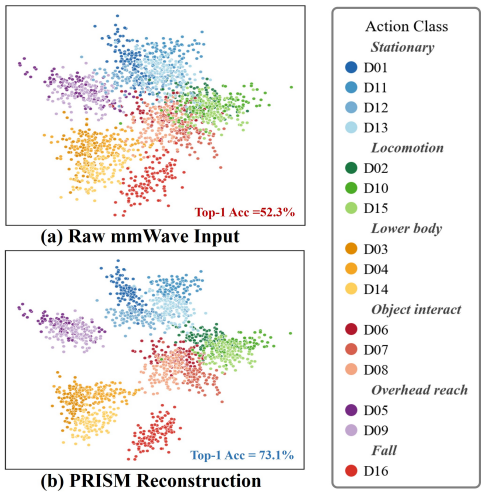


Figure 7: t-SNE of action features on 16 daily-living classes. (a) Raw mmWave input and (b) PRISM reconstruction.

Table 2: Downstream task performance on MIST (test subjects S07–S08).

	HAR Acc (%)↑			Fall F1↑	HPE MPJPE (mm)↓			Rehab r↑		
Input	OL0	OL1	OL2		OL0	OL1	OL2	OL0	OL1	OL2
Raw mmWave	52.3	43.6	35.2	0.628	138	163	191	0.38	0.31	0.24
PoinTr [22]	63.8	57.1	49.4	0.741	116	147	172	0.51	0.44	0.37
TIGER [43]	67.2	60.1	53.8	0.782	104	131	158	0.58	0.51	0.43
PRISM (ours)	**73.1**	**66.8**	**59.3**	**0.851**	**84**	**96**	**112**	**0.68**	**0.61**	**0.53**

Human activity recognition. We use the sixteen daily-living action classes in MIST for this evaluation. Because the PRISM reconstructions are conditioned on the ground-truth action label, which is the quantity this task predicts, the HAR accuracy of PRISM should be read as an upper bound. Results show that PRISM achieves 73.1% top-1 accuracy at OL0—a 20.8 percentage points above raw mmWave and 5.9 points above TIGER. Under heavy occlusion at OL2 the margin widens to 24.1 percentage points, showing that reconstruction disproportionately benefits harder sensing conditions. The t-SNE visualization in Figure 7 corroborates this at the feature level, where raw mmWave roughly separates broad action groups but still confuses similar motion types such as lower-body activities, while PRISM reconstructions form tighter and more separable clusters despite residual coupling between adjacent classes.

Human pose estimation. A 17-joint regression head predicts H36M-convention joint coordinates, with accuracy measured in MPJPE, and full results presented in Table 2. PRISM reduces MPJPE from 138 to 84 mm at OL0 and still achieves 112 mm at OL2, which remains below raw mmWave at clear-view OL0. This result demonstrates that reconstruction compensates effectively for through-obstacle signal attenuation. The MPJPE gap between PRISM and raw mmWave grows from 54 mm at OL0 to 79 mm at OL2.

Rehabilitation exercise quality assessment. Motion quality scores derived from per-joint velocity cosine similarity for classes R01–R14 are compared with LiDAR reference scores by Pearson correlation, with detailed results tabulated in Table 2. As can be seen, PRISM achieves $r = 0.68$ at OL0, surpassing raw mmWave at $r = 0.38$ and PoinTr at $r = 0.51$. At OL2, PRISM retains $r = 0.53$ while raw mmWave decreases to $r = 0.24$, confirming that PRISM supports clinically meaningful exercise monitoring without requiring unobstructed sensor placement.

6 Limitations

MIST was recorded in a single laboratory with 18 mm wooden boards as the only barrier material. Reconstruction behind barriers with different electromagnetic properties, such as metal, reinforced concrete, or laminated glass, is uncharacterized, and our evaluation does not establish zero-shot generalization to new scenes; because radar returns also depend on reflections from walls and furniture, a new scene would likely require recalibration or additional training data. The dataset covers fifteen subjects, of whom only eight are used in the reported experiments, and the participant pool does not span a wide range of body morphologies or age groups. Both MIST and PRISM assume a single subject within the sensing volume; separating overlapping returns from several people would require a different conditioning strategy and data collection protocol. PRISM generates each frame independently, so its reconstructions can jitter over time, and all results come from single training runs without repeated random seeds. With 200 reverse diffusion steps, PRISM takes 0.88 s per frame on one NVIDIA RTX 6000 Ada GPU; reducing the sampler to 10 steps lowers this to 49 ms while COV-CD stays within 5% of the 200-step value. Finally, although mmWave radar records no facial or clothing appearance, it still reveals a person's presence, location, pose, and activity, so through-wall sensing could enable monitoring without the person's knowledge or consent, and deployments should require the consent of the people being monitored.

7 Conclusion

We presented MIST, the first through-obstacle mmWave human benchmark with three paired occlusion levels and synchronized LiDAR ground truth, and PRISM, a skeleton-guided latent diffusion framework that reconstructs dense human point clouds from sparse mmWave radar. PRISM resolves the two failure modes of prior approaches by inferring radar-aligned skeleton joints from sparse returns, pairing them with density-invariant coarse geometry and action embeddings, and injecting this structural prior at every denoising step of a latent diffusion model. On MIST, PRISM achieves COV-CD 0.362 against 0.113 for the strongest completion baseline and maintains action-discriminative reconstruction at OL2 where all deterministic baselines collapse, and downstream evaluations confirm that PRISM reconstructions consistently outperform raw mmWave at every occlusion level, with OL2 outputs still surpassing raw mmWave under clear-view conditions. Together, MIST and PRISM provide a foundation for privacy-preserving, through-obstacle healthcare monitoring using mmWave radar alone.

Acknowledgments and Disclosure of Funding

This work was supported by the National Natural Science Foundation of China (62506277). The authors declare no competing interests.

References

[1] Jianfei Yang, He Huang, Yunjiao Zhou, Xinyan Chen, Yuecong Xu, Shenghai Yuan, Han Zou, Chris Xiaoxuan Lu, and Lihua Xie. MM-Fi: Multi-modal non-intrusive 4D human dataset for versatile wireless sensing. *Advances in Neural Information Processing Systems*, 36:18756–18768, 2023.

[2] Jia Zhang, Rui Xi, Yuan He, Yimiao Sun, Xiuzhen Guo, Weiguo Wang, Xin Na, Yunhao Liu, Zhenguo Shi, and Tao Gu. A survey of mmWave-based human sensing: Technology, platforms and applications. *IEEE Communications Surveys & Tutorials*, 25(4):2052–2087, 2023.

[3] Jiacheng Huang, Honglin Liao, Cunyi Yin, Hao Jiang, Jing Chen, Zhaoke Huang, and Zhiwen Chen. RoPEHAR: A real-time rotary position encoding informer for mmWave-based human activity recognition in substations. *IEEE Internet of Things Journal*, 13(13):28620–28632, 2026.

[4] Akash Deep Singh, Sandeep Singh Sandha, Luis Garcia, and Mani Srivastava. RadHAR: Human activity recognition from point clouds generated through a millimeter-wave radar. In *Proc. 3rd ACM Workshop on Millimeter-wave Networks and Sensing Systems*, pages 51–56, 2019.

[5] Zhuo Chang and Shilong Lou. Efficient human action recognition with fine-grained spatiotemporal feature extraction from millimeter-wave point clouds. *IEEE Sensors Journal*, 2025.

[6] Yingxiao Wu, Zhongmin Jiang, Haocheng Ni, Changlin Mao, Zhiyuan Zhou, Wenxiang Wang, and Jianping Han. mmHPE: Robust multiscale 3D human pose estimation using a single mmWave radar. *IEEE Internet of Things Journal*, 12(1):1032–1046, 2024.

[7] Wen Ding, Zhongping Cao, Jianxiong Zhang, Rihui Chen, Xuemei Guo, and Guoli Wang. Radar-based 3D human skeleton estimation by kinematic constrained learning. *IEEE Sensors Journal*, 21(20):23174–23184, 2021.

[8] Zhongping Cao, Guangyu Mei, Xuemei Guo, and Guoli Wang. VirTeach: mmWave radar point-cloud-based pose estimation with virtual data as a teacher. *IEEE Internet of Things Journal*, 11(10):17615–17628, 2024.

[9] Yujun Wang, Fabian Seguel, Stefan Hägele, Driton Salihu, and Eckehard Steinbach. PoinTS: mmWave Radar-based pointNet and transformer network for human skeleton estimation in privacy-preserving rehabilitation. In *Proc. IEEE EUROCON 2025-21st International Conference on Smart Technologies*, pages 1–6. IEEE, 2025.

[10] Yuexuan Feng, Songchen Dai, Qifei Zhang, Zhao Wang, Xianmin Zhang, and Yulin Zhou. M3pose: Multi-person 3D pose estimation using sparse millimeter-wave radar point clouds. In *Proc. Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pages 504–517. Springer, 2024.

[11] Yuan Liu, Zhi-Wei Tan, Andy W. H. Khong, and Hongwei Liu. An iterative implementation-based approach for joint source localization and association under multipath propagation environments. *IEEE Transactions on Signal Processing*, 71:121–135, 2023.

[12] Yuan Liu, Hongwei Liu, Lu Wang, and Guoan Bi. Target localization in high-coherence multipath environment based on low-rank decomposition and sparse representation. *IEEE Transactions on Geoscience and Remote Sensing*, 58(9):6197–6209, 2020.

[13] Yuan Liu, Xuemei Fu, Ritesh Chandra Tewari, Xingze Wang, and Andy W. H. Khong. Contactless vital sign monitoring for multiple people using a millimeter-wave MIMO radar. In *Proc. ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025.

[14] Xingze Wang, Yuan Liu, Andy W. H. Khong, and Tzu-Yi Hung. Remote multi-person blood pressure monitoring using mmwave radar. In *Proc. ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 21061–21065. IEEE, 2026.

[15] Xingze Wang, Yuan Liu, Long Tian, and Andy W. H. Khong. Spectrogram restoration and classification framework for multi-person through-obstacle human activity recognition. In *Proc. ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 21351–21355. IEEE, 2026.

[16] Benxu Liu, V. G. Reju, and Andy W. H. Khong. A linear source recovery method for underdetermined mixtures of uncorrelated AR-model signals without sparseness. *IEEE Transactions on Signal Processing*, 62(19):4947–4958, Oct. 2014.

[17] Sizhe An, Yin Li, and Umit Ogras. 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.

[18] Junqiao Fan, Yunjiao Zhou, Yizhuo Yang, Xinyuan Cui, Jiarui Zhang, Lihua Xie, Jianfei Yang, Chris Xiaoxuan Lu, and Fangqiang Ding. M4Human: A large-scale multimodal mmWave radar benchmark for human mesh reconstruction. *arXiv preprint arXiv:2512.12378*, 2025.

[19] Anjun Chen, Xiangyu Wang, Shaohao Zhu, Yanxu Li, Jiming Chen, and Qi Ye. mmBody benchmark: 3D body reconstruction dataset and analysis for millimeter wave radar. In *Proc. 30th ACM International Conference on Multimedia*, pages 3501–3510, 2022.

[20] Mingmin Zhao, Tianhong Li, Mohammad Abu Alsheikh, Yonglong Tian, Hang Zhao, Antonio Torralba, and Dina Katabi. Through-wall human pose estimation using radio signals. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 7356–7365, 2018.

[21] Wentao Yuan, Tejas Khot, David Held, Christoph Mertz, and Martial Hebert. PCN: Point completion network. In *Proc. 2018 International Conference on 3D Vision (3DV)*, pages 728–737. IEEE, 2018.

[22] Xumin Yu, Yongming Rao, Ziyi Wang, Zuyan Liu, Jiwen Lu, and Jie Zhou. PoinTr: Diverse point cloud completion with geometry-aware transformers. In *Proc. IEEE/CVF International Conference on Computer Vision*, pages 12498–12507, 2021.

[23] Peng Xiang, Xin Wen, Yu-Shen Liu, Yan-Pei Cao, Pengfei Wan, Wen Zheng, and Zhizhong Han. SnowflakeNet: Point cloud completion by snowflake point deconvolution with skip-transformer. In *Proc. IEEE/CVF International Conference on Computer Vision*, pages 5499–5509, 2021.

[24] Amir Sulaiman, Kirill Poletkin, and Andy W. H. Khong. Source localization in the presence of dispersion for next generation touch interface. In *Proc. of the 2010 International Conference on Cyberworlds*, pages 82–86, 2010.

[25] Xiaohui Zeng, Arash Vahdat, Francis Williams, Zan Gojcic, Or Litany, Sanja Fidler, and Karsten Kreis. LION: Latent point diffusion models for 3D shape generation. In *Advances in Neural Information Processing Systems*, volume 35, pages 10021–10039, 2022.

[26] Junzhong Ji, Runfeng Zhao, and Minglong Lei. Latent diffusion transformer for point cloud generation. *The Visual Computer*, 40(6):3903–3917, 2024.

[27] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[28] Qian Xie, Qianyi Deng, Ta Ying Cheng, Peijun Zhao, Amir Patel, Niki Trigoni, and Andrew Markham. mmPoint: Dense human point cloud generation from mmWave. In *Proc. 34th British Machine Vision Conference BMVC*, pages 194–196, 2023.

[29] Han Cui, Shu Zhong, Jiacheng Wu, Zichao Shen, Naim Dahnoun, and Yiren Zhao. Milipoint: A point cloud dataset for mmWave radar. *Advances in Neural Information Processing Systems*, 36:62713–62726, 2023.

[30] Wenteng Lu, Yinan Chen, and Xiaoxia Huang. mmChainPose: Geometry-aware temporal chaining for robust human pose estimation from mmwave point clouds. *Neurocomputing*, page 132019, 2025.

[31] Marianna Capecci, Maria Gabriella Ceravolo, Francesco Ferracuti, Sabrina Iarlori, Andrea Monteriu, Luca Romeo, and Federica Verdini. The Kimore dataset: Kinematic assessment of movement and clinical scores for remote monitoring of physical rehabilitation. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 27(7):1436–1448, 2019.

[32] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3D human sensing in natural environments. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, 2013.

[33] Rongqing Kenneth Ong and Andy W. H. Khong. Spectrum-based modality representation fusion graph convolutional network for multimodal recommendation. In *Proceedings of the 18th ACM International Conference on Web Search and Data Mining*, pages 773–781, 2025.

[34] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[35] Junshu Tang, Zhijun Gong, Ran Yi, Yuan Xie, and Lizhuang Ma. LAKe-Net: Topology-aware point cloud completion by localizing aligned keypoints. In *Proc. IEEE/CVF Conference on Computer Vision Pattern Recognition*, pages 1726–1735, 2022.

[36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

[37] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. PointNet: Deep learning on point sets for 3D classification and segmentation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 652–660, 2017.

[38] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[39] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing systems*, 33:6840–6851, 2020.

[40] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

[41] Diederik P Kingma and Jimmy Ba. ADAM: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[42] Panos Achlioptas, Olga Diamanti, Ioannis Mitliagkas, and Leonidas Guibas. Learning representations and generative models for 3D point clouds. In *Proc. International Conference on Machine Learning*, pages 40–49. PMLR, 2018.

[43] Zhiyuan Ren, Minchul Kim, Feng Liu, and Xiaoming Liu. Tiger: Time-varying denoising model for 3D point cloud generation with diffusion process. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9462–9471, 2024.

[44] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in Neural Information Processing Systems*, 30, 2017.

[45] Hongfei Xue, Yan Ju, Chenglin Miao, Yijiang Wang, Shiyang Wang, Aidong Zhang, and Lu Su. mmMesh: Towards 3D real-time dynamic human mesh construction using millimeter-wave. In *Proc. 19th Annual International Conference on Mobile Systems, Applications, and Services*, pages 269–282, 2021.

[46] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[47] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

A MIST dataset details

A.1 Comparison with Existing mmWave Datasets

Table 3 compares representative mmWave human sensing datasets along five axes, where *Phys. Barrier* indicates a physical object between sensor and subject, *Paired* indicates simultaneous clear-view and occluded captures of identical poses, *LiDAR GT* indicates synchronized overhead LiDAR as dense 3-D reconstruction ground truth, *Skel/Mesh GT* indicates skeleton or mesh annotations, and *OL* indicates the number of controlled occlusion levels. RF-Pose uses sub-6 GHz radio and produces 2-D skeleton heatmaps rather than a point cloud, and mmBody's occlusion variant covers the sensor aperture rather than placing a physical barrier between sensor and subject.

Obstacle-free benchmarks. The dominant paradigm collects data under clear sensor–subject line-of-sight. mmMesh [45] pioneered human mesh construction from sparse 60 GHz reflections but covered only 20 subjects across 8 actions with no occlusion scenario. mmBody [19] expanded action coverage to 100 categories with paired mesh annotations, and its occlusion variant places absorbent material over the sensor aperture rather than a physical barrier, providing no spatial selectivity and no clear-view reference. mRI [17] pairs mmWave with RGB-D and inertial sensing for 3-D skeleton annotation under obstacle-free conditions. MM-Fi [1] includes synchronized overhead LiDAR as a dense reconstruction reference but collects every session in an open indoor space without any barrier. M4Human [18] scales to 50 action categories with Vicon mesh annotations yet retains the same obstacle-free assumption.

Through-barrier and simulated-occlusion sensing. RF-Pose [20] demonstrated that a physical wall need not block human sensing, using sub-6 GHz radio signals to produce 2-D skeleton heatmaps of subjects behind a wall. However, the sub-60 GHz frequency band, 2-D output format, single fixed wall, and absence of a simultaneous clear-view reference or LiDAR reconstruction target leave mmWave-specific dense point cloud reconstruction under controlled multi-level occlusion unaddressed.

MIST. MIST is the first mmWave dataset to simultaneously satisfy all three requirements. Occlusion levels are defined by cumulative board thickness, where OL0 uses no board, OL1 interposes one 18 mm wooden board, and OL2 stacks two boards totalling 36 mm between the subject and one radar, and a second co-located radar maintains an unobstructed path, so every frame provides a paired clear-view capture at OL0 and through-obstacle captures at OL1 and OL2 of the identical pose. An overhead Livox Mid-360 LiDAR, physically separated from the barrier, provides dense reconstruction ground truth unaffected by board placement across all levels. The three-level occlusion design converts attenuation severity into a controlled experimental variable, enabling the first direct evaluation of density-invariant reconstruction under realistic through-barrier conditions.

Table 3: Comparison of MIST with representative mmWave human sensing datasets.

Dataset	Year	Phys. Barrier	Paired	LiDAR GT	Skel/Mesh GT	OL	#Subj	#Act	#Frames	mmWave pts/frame
RF-Pose [20]	2018	✓	×	×	×	1	∼100	1	–	–
mmMesh [45]	2021	×	×	×	✓ (mesh)	1	20	8	∼3k	∼50
mmBody [19]	2022	×	×	×	✓ (mesh)	1	20	100	∼200k	∼200
mRI [17]	2022	×	×	×	✓ (skel)	1	20	12	∼160k	50–150
MM-Fi [1]	2023	×	×	✓	✓ (skel)	1	40	27	>320k	50–150
M4Human [18]	2024	×	×	×	✓ (mesh)	1	20	50	∼661k	50–150
MIST (ours)	2026	✓	✓	✓	✓ (skel)	3	15	30	∼270k	50–150 (OL0) 25–80 (OL1) 10–30 (OL2)

A.2 Sensor Configuration

Data collection took place in a controlled indoor laboratory with approximately 5 m × 6 m usable floor area, where subjects performed actions at 2–3 m from the door-side sensor array. Two TI IWR6843 FMCW radars are co-located at ∼1.4 m, one maintaining a clear line-of-sight for OL0 and one operating through a wooden barrier of 18 mm for OL1 or two stacked boards totalling 36 mm for OL2, and collecting both streams simultaneously ensures that every paired frame is captured at an identical subject pose. A Livox Mid-360 LiDAR is mounted overhead and extrinsic-calibrated to a shared world coordinate frame, providing 360° dense point clouds as reconstruction ground truth unaffected by board placement. An Intel RealSense D435i depth camera at the door-side mount delivers synchronized RGB-D streams from which H36M-17 skeleton ground truth is extracted at 30 fps and resampled to 10 fps to match the radar frame rate. Table 4 summarises the key sensor parameters.

A.3 Action List

MIST covers 30 action classes across two functional categories. The rehabilitation category includes 14 exercises targeting upper and lower limb mobility, balance, and gait assessment relevant to clinical motion screening. The daily-living category includes 16 activities emphasising whole-body movements that appear frequently in elderly-care monitoring scenarios. Of the 30 actions, 8 share equivalent kinematic categories with MM-Fi, enabling direct cross-dataset transfer evaluation.

A.4 Directory Structure

Figure 8 shows the hierarchical layout of the MIST dataset. The root contains one sub-directory per subject (S01–S15), and each subject directory is further divided by action (R01–R14 and D01–D16).

Table 4: MIST sensor specifications. N/A indicates the parameter is not applicable to that modality.

Parameter	TI IWR6843 $\times 2$	Livox Mid-360	Intel RealSense D435i
Modality	60–64 GHz FMCW	905 nm LiDAR	RGB-D and IMU
Bandwidth	4 GHz	N/A	N/A
Range resolution	4.4 cm	2 cm	N/A
Max range	5 m	40 m	0.1–10 m
Velocity resolution	0.13 m/s	N/A	N/A
Field of view	± 60 az and ± 15 el	360 az and -7 to 52 el	87×58
Antenna	3 Tx $\times$ 4 Rx	N/A	N/A
Frame rate	10 fps	10 fps	30 fps
Color and depth res.	N/A	N/A	640×480
Skeleton convention	N/A	N/A	H36M-17 joints

Table 5: MIST complete action list.

Rehabilitation (R01–R14)				*Daily Living (D01–D16)*			
R01	Shoulder forward flexion	R08	Knee extension	D01	Standing still	D09	Reaching overhead
R02	Shoulder abduction	R09	Ankle dorsiflexion	D02	Walking in place	D10	Full body turn
R03	Elbow flexion extension	R10	Balance left leg	D03	Squat	D11	Waving hand
R04	Bilateral wrist rotation	R11	Balance right leg	D04	Rising from chair	D12	Phone call gesture
R05	Neck rotation	R12	Tandem gait	D05	Raising arms overhead	D13	Clapping
R06	Trunk side bend	R13	Sit to stand	D06	Forward bend	D14	Forward kick
R07	Hip flexion	R14	Lateral weight shift	D07	Picking up object	D15	Jumping in place
				D08	Carrying object	D16	Forward fall

Within each action folder, the paired occlusion design is reflected directly in the directory structure, where one sub-folder pairs OL0 (clear-view) with OL1 (18 mm board), and a second sub-folder pairs OL0 with OL2 (36 mm total), so that every through-obstacle capture has an exact simultaneous clear-view counterpart at the same pose. Each occlusion-level folder stores mmWave radar point clouds (`.npy`), LiDAR point clouds (`.bin`), depth maps (`.npy`), and a single per-sequence skeleton annotation file (`ground_truth_h36m.npy`) providing 17-joint H36M coordinates at 10 fps. RGB frames were recorded only to extract the skeletons and are not released.

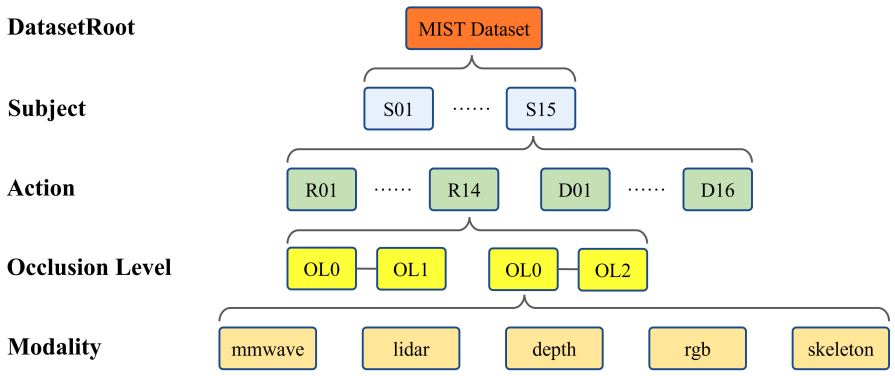


Figure 8: Expanded directory structure of the MIST dataset.

B Architecture and training details

B.1 Module Hyperparameters

Table 6 lists the architecture hyperparameters for each PRISM module. All hidden dimensions, attention head counts, and layer depths are chosen to balance expressive capacity with training feasibility on three NVIDIA A6000 GPUs (48 GB each). The Temporal Adapter output dimension of

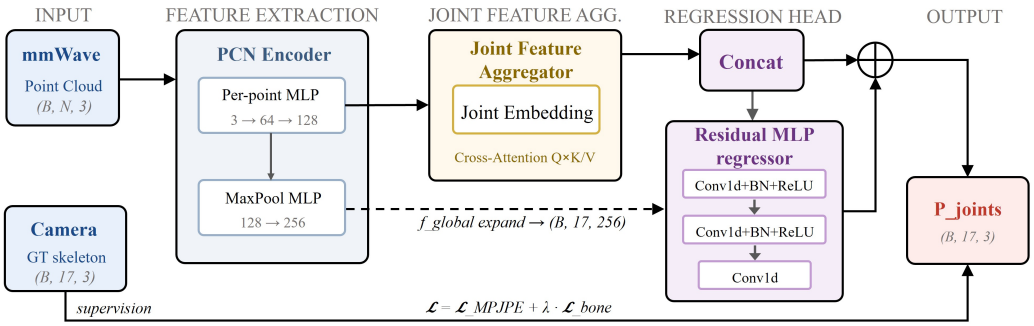


Figure 9: SkelKD architecture in which 17 learnable joint query embeddings attend over encoded radar returns via cross-attention to regress radar-aligned skeleton joints.

8×120 is fixed prior to diffusion training and shared across all conditioning configurations. The structure of SkelKD is shown in 9

Table 6: PRISM module hyperparameters.

Module	Parameter	Value
SkelKD	Conv feature dim (local)	128
	Global pooled feature dim	256
	Joint query embeddings	17×128
Temporal Adapter	Transformer encoder layers	6
	Output feature dim	20 (8×120 total)
	Hidden dim and attention heads	128 and 4
	Output tokens K	256
Coarse Generator	Feature dim	256
	Output coarse points	128
Condition Network	GCN layers and node dim	2 and 256
	Coarse-geometry encoder	Conv1d, 256 channels
	Total K/V tokens	145 (17 skeleton + 128 geometry)
Diffusion Denoiser	Residual blocks	6
	Hidden dim and attention heads	256 and 4
	Conditioning	AdaLN [46] and CFG [47]

B.2 Training Schedule

PRISM is trained in four sequential stages on three NVIDIA A6000 GPUs (48 GB each). Stage 1 pre-trains SkelKD at a learning rate of 10^{-3} with batch size 16 for 5000 epochs (approximately 12 h) and freezes it. Stage 2 trains the Temporal Adapter independently at 2×10^{-4} with batch size 8 for 3000 epochs (approximately 48 h) and freezes it. Stage 3 jointly optimises the Diffusion Denoiser, Condition Network, and Coarse Generator at 2×10^{-4} with batch size 16 for 5000 epochs (approximately 24 h). Stage 4 fine-tunes all components except SkelKD end-to-end at 3×10^{-5} with batch size 16 for 2000 epochs (approximately 8 h). All stages use Adam with cosine learning-rate decay, gradient clipping at 1.0, and 500 warm-up steps, and Stages 3 and 4 additionally apply EMA with decay 0.9999.

C Additional experiments

C.1 Cross-dataset generalization on MM-Fi

To evaluate cross-dataset generalisation, all models are trained on MIST-OL0 and evaluated directly on MM-Fi [1], a large-scale public benchmark comprising 40 subjects performing 27 action classes across 4 environments with synchronised mmWave point clouds and depth-camera skeleton ground

truth under the H36M convention. Evaluation covers the 8 action classes shared between MIST and MM-Fi, namely A03, A12, A13, A17, A19, A22, A26, and A27, using test subjects S08 to S10 from environment E01. Baseline methods are additionally re-trained on MM-Fi subjects S01 to S07 to provide an in-domain upper bound, while PRISM transfers directly from its MIST-trained weights without any MM-Fi-specific fine-tuning.

Table 7 shows that PRISM achieves consistent improvements on MM-Fi comparable to those on MIST-OL0. MMD-CD is 40.6% lower than the best completion baseline PoinTr and COV-CD is 220% higher, confirming that skeleton-guided conditioning transfers across dataset conditions without any target-domain fine-tuning. The 1-NN-CD gap between PRISM and LION narrows compared to MIST, which is expected, given that with 7 training subjects in MM-Fi versus 6 in MIST-OL0, unconditional generative baselines benefit from a richer training distribution, while PRISM's skeleton conditioning still provides the decisive structural advantage. The SnowflakeNet degradation, with COV-CD 0.081 falling below PCN at 0.086, replicates the pattern observed on MIST, confirming that multi-stage farthest-point sampling failure under extreme sparsity is an architectural limitation rather than a dataset artefact.

Table 7: Cross-dataset generalization on MM-Fi across 8 overlapping action classes on test subjects S08 to S10. Methods marked * are re-trained on MM-Fi, while PRISM transfers directly from MIST-trained weights without fine-tuning.

Method	Venue	Type	MMD-CD↓	COV-CD↑	1-NN-CD↓
PCN* [21]	3DV'18	Compl.	0.00887	0.0855	0.9994
PoinTr* [22]	ICCV'21	Compl.	0.00813	0.1172	0.9989
SnowflakeNet* [23]	ICCV'21	Compl.	0.01056	0.0808	0.9998
LAKe-Net* [35]	CVPR'22	Compl.	0.00856	0.0878	0.9993
mmPoint* [28]	BMVC'23	Det.	0.00877	0.0991	0.9995
LION* [25]	NeurIPS'22	Gen.	0.00724	0.1892	0.9813
TIGER* [43]	CVPR'24	Gen.	0.00695	0.2147	0.9776
PRISM (ours)	–	Gen.	0.00483	0.3748	0.9697

Per-action COV-CD on MM-Fi ranges from 0.344 for Arm Raise, which involves subtle distal motion, to 0.412 for Bowing, which involves pronounced full-body displacement, with a macro-average of 0.374 consistent with the aggregate in Table 7, confirming that per-frame skeleton conditioning is the primary driver of action-level discrimination.

C.2 Condition-matched deterministic baseline

The ablations in Section 5.3 remove conditioning while keeping the diffusion denoiser. This appendix reports the complementary experiment: the full conditioning stack is retained and only the denoiser is replaced, which isolates the contribution of diffusion refinement from that of the structural priors.

The deterministic variant keeps SkelKD, the GCN, the Coarse Generator, the Temporal Adapter, the action-label embedding and the frozen decoder unchanged, and reuses the denoiser network at identical capacity. Instead of predicting noise along a reverse trajectory, it receives a constant input at a fixed timestep and regresses the clean latent directly under a squared-error objective, so a single forward pass replaces the 200 sampling steps. The auxiliary coarse Chamfer loss and the optimiser settings follow Section 5.1; only the denoising objective differs. Both systems are evaluated under one protocol, with per-sample normalisation on 2000 held-out frames.

Table 8 shows that the deterministic variant attains the lower paired Chamfer Distance of the two, yet covers less than a third of the reference distribution and is almost perfectly separable from real data under the 1-NN test. This is the signature of regression to the mean: a squared-error objective on the latent drives the model toward the conditional average body, which is close to every reference in aggregate while matching none in detail. Distributional coverage therefore comes from the diffusion process, and paired Chamfer Distance alone is not a sufficient criterion for this task.

Table 8: Condition-matched comparison of the latent decoder. Both systems share the same conditioning stack, decoder and evaluation protocol, and differ only in whether the latent is produced by reverse diffusion or by direct regression.

Latent decoder	COV-CD↑	MMD-CD↓	1-NN-CD↓	Paired CD↓
Deterministic regression	0.134	0.01056	0.970	0.01400
Reverse diffusion (ours)	0.387	0.00663	0.852	0.01634

C.3 Through-obstacle robustness on OL1 and OL2

Table 9 disaggregates the through-obstacle evaluation by occlusion level, complementing the main-text discussion of OL1/OL2 robustness. All methods are trained exclusively on MIST-OL0 (S01–S06) and evaluated zero-shot on OL1 and OL2 with no obstacle-specific adaptation.

At OL1, where a single 18 mm board attenuates returns to 25–80 points per frame, PRISM retains a COV-CD of 0.298, a relative drop of 17.7% from OL0, while all deterministic baselines drop to COV-CD below 0.07, a relative degradation of 40–50%. At OL2, where two stacked 18 mm boards totalling 36 mm reduce returns to 10–30 points per frame, the most severe condition in MIST, PRISM still achieves a COV-CD of 0.214 while baselines collapse below 0.035. The ΔCOV columns confirm that PRISM's relative degradation is consistently two to three times smaller than any baseline across both occlusion levels.

Two design choices jointly account for this resilience. First, the Coarse Generator maps inputs of any density to 128 stable coarse body points, so geometric anchoring remains intact even when FPS-based aggregation degenerates under extreme sparsity. Second, SkelKD's cross-attention aggregation spreads over all visible returns rather than relying on a nearest-neighbor match, so the skeleton conditioning signal degrades gracefully as the return count falls rather than producing a catastrophic feature collapse. Together, these properties allow PRISM to maintain action-discriminative reconstruction at point densities where every deterministic baseline produces a fixed-template output.

Table 9: Through-obstacle robustness on MIST (zero-shot from OL0 training). ΔCOV: relative COV-CD change from OL0, COV/MMD: COV-CD/MMD-CD.

		OL1				OL2			
Method	Type	MMD↓	COV↑	1-NN↓	ΔCOV	MMD↓	COV↑	1-NN↓	ΔCOV
PCN* [21]	Compl.	0.0142	0.051	0.9961	−38.6%	0.0182	0.021	0.9983	−74.7%
PoinTr* [22]	Compl.	0.0126	0.067	0.9949	−40.9%	0.0166	0.031	0.9973	−72.6%
SnowflakeNet* [23]	Compl.	0.0183	0.039	0.9972	−50.6%	0.0225	0.017	0.9990	−78.5%
mmPoint* [28]	Det.	0.0151	0.054	0.9957	−44.3%	0.0191	0.024	0.9979	−75.2%
PRISM (ours)	Gen.	0.0076	0.298	0.9734	−17.7%	0.0102	0.214	0.9768	−40.9%

D Qualitative results

D.1 Cross-action comparison

Figure 10 illustrates the two dominant failure modes of baseline methods. For distal limb actions such as Wave Hand and Kick, all baselines produce a neutral standing shape regardless of pose, as the sparse radar returns from a raised arm or swinging leg are insufficient to activate the nearest-neighbour features that completion backbones rely on, and unconditional generative methods lack the structural prior needed to distinguish these configurations from rest. Squat exposes a different failure mode, as the full-body postural change distributes returns across an unusual spatial extent, causing baselines to average over the observed cloud and yield a blurred, anatomically inconsistent shape. Per-frame skeleton conditioning steers PRISM toward the correct kinematic state in all four actions, producing body shapes that faithfully match the LiDAR ground truth silhouette and limb arrangement.

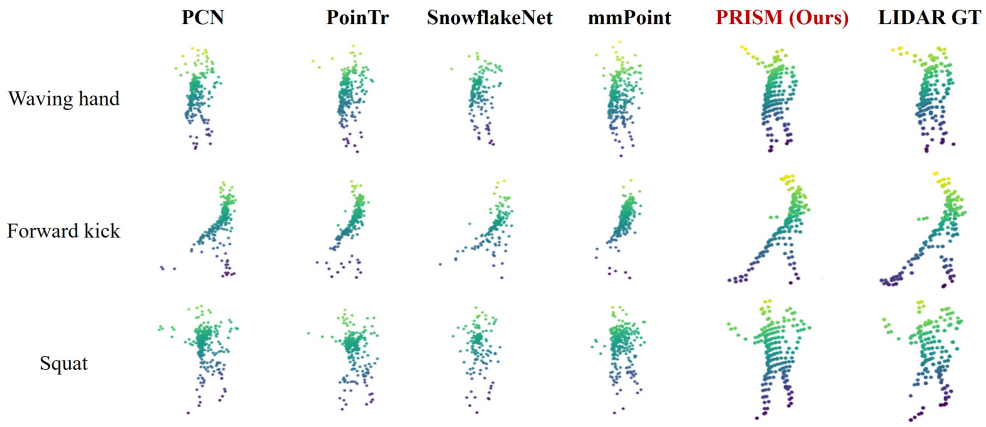


Figure 10: Cross-action qualitative comparison of four actions (rows) versus six methods (columns), with LiDAR ground truth as the rightmost column.

D.2 Latent diffusion process

Figure 11 visualises the step-by-step reverse process in the Temporal Adapter feature space for eight representative MIST actions. Each panel shows the decoded point cloud at a fixed denoising timestep, from pure noise at $t = 1.0$ on the left to the fully denoised output at $t = 0$ on the right. Starting from isotropic Gaussian noise, the Diffusion Denoiser progressively resolves global body structure before refining local pose detail. This coarse-to-fine behaviour is consistent across all 30 MIST action classes. Actions with large limb displacements establish recognisable global body structure as early as $t = 0.6$, as the skeleton K/V tokens inject a strong kinematic prior at every denoising layer. Actions with subtle distal motion maintain the correct global silhouette from $t = 0.6$ but require later timesteps to resolve fine-grained limb configurations.

D.3 Temporal reconstruction across a bowing sequence

Figure 12 compares point cloud reconstruction across three timesteps of a bowing sequence, from an upright starting pose through mid-bow to the fully extended bow. All four baselines produce a near-identical upright silhouette at every timestep regardless of the pronounced trunk flexion between frames, as completion backbones trained on rigid objects have no mechanism to track continuous non-rigid articulation and unconditional generative models converge to the marginal body distribution independent of the sparse radar input. PRISM faithfully tracks the bow arc at each step, progressively deepening the trunk angle while maintaining anatomically plausible limb geometry, because the radar-aligned skeleton joints injected as conditioning tokens at every denoising step anchor the reconstruction to the current kinematic state of the subject. This temporal fidelity emerges from per-frame conditioning alone and requires no explicit across-frame sequence modelling.

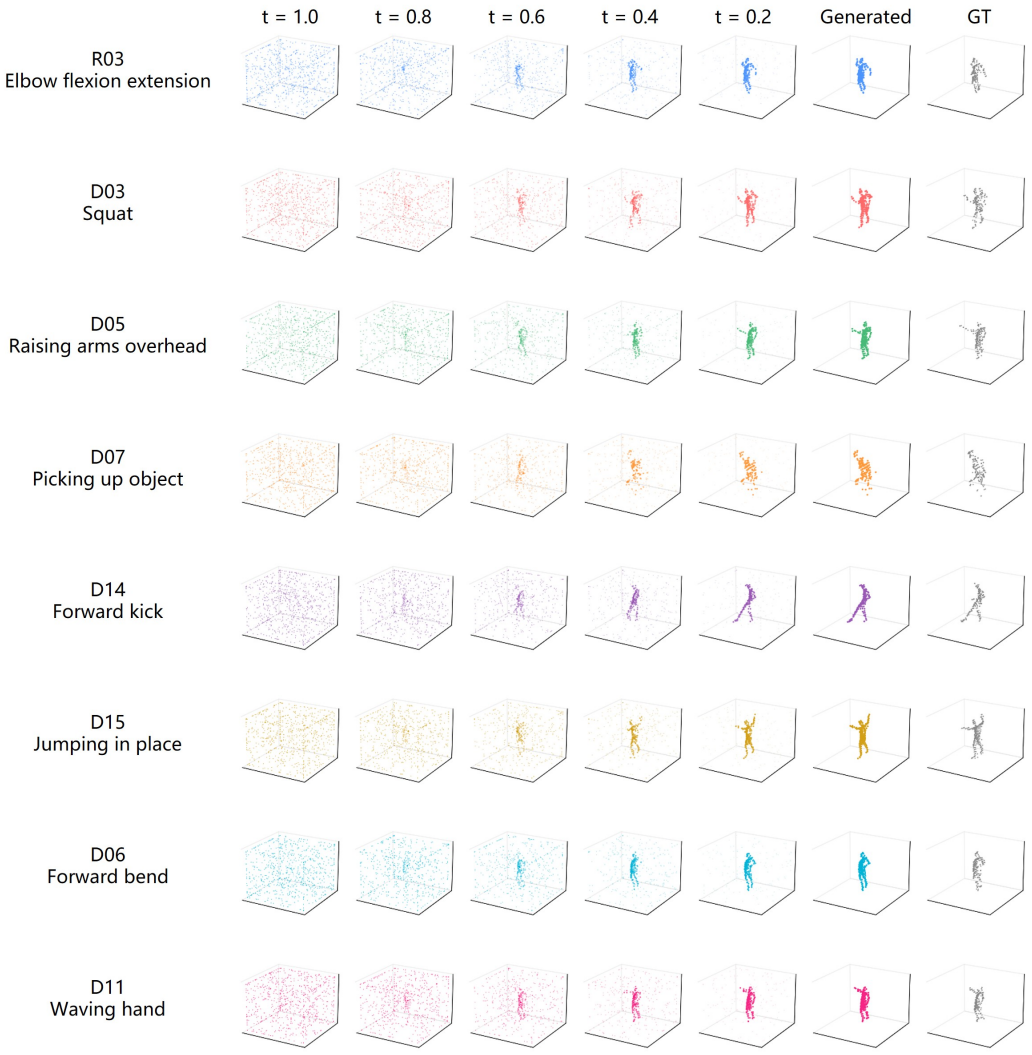


Figure 11: PRISM reverse denoising process for action Forward Bend, subject S01, visualised at six timesteps from $t = 1.0$ to $t = 0$.

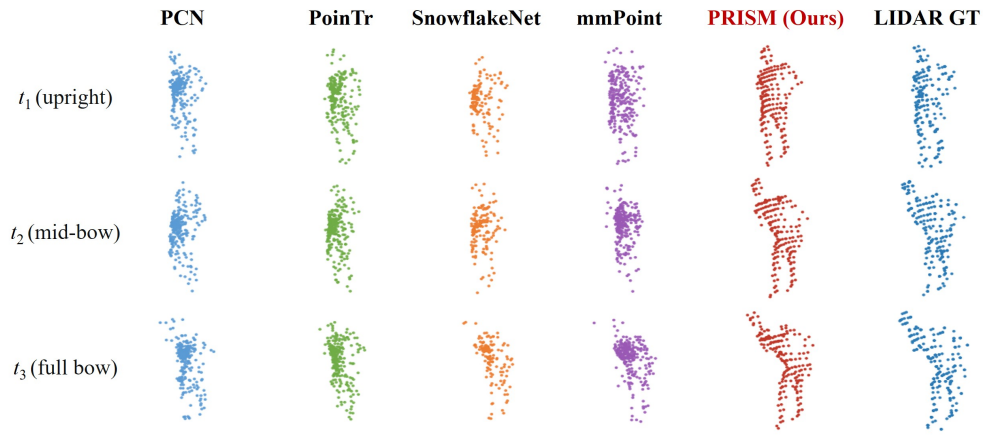


Figure 12: Temporal point cloud reconstruction across three timesteps of a bowing sequence, comparing baseline methods and PRISM against LiDAR ground truth.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and Section 1 state three contributions, the MIST dataset (Section 3), the PRISM framework (Section 4) and the downstream benchmarks (Section 5.4), each supported by Tables 1 and 2. Healthcare monitoring is presented as the motivating application, not as a clinically validated result.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 6 discusses the single barrier material and recording room, which means a new scene would need recalibration, the size of the subject pool, the single-person assumption, the inference latency of 200 diffusion steps, the use of single training runs, and the privacy risks of through-obstacle sensing.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper makes no theoretical claims; its equations define the diffusion process and the training objective.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section 5.1 specifies the subject-disjoint split (six training and two evaluation subjects among the eight used for benchmarking), the four training stages and the evaluation metrics, and Appendices B and B.2 list all architecture hyperparameters and per-stage training settings.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The code and the MIST dataset are not yet public; we plan to release them, without RGB frames, at a later date. Section 5.1 and Appendices B and B.2 give the architecture, data split and training settings needed to reimplement the method.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 5.1 and Appendix B.2 give the data split, optimizer, learning rates, batch sizes, epoch counts, gradient clipping and hardware, and all baselines are retrained on the same split.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

Justification: No error bars are reported because each result comes from a single training run, and one full run takes about 92 wall-clock hours on three NVIDIA A6000 GPUs; Section 6 acknowledges this. The main comparison, COV-CD 0.362 for PRISM against 0.209 for the strongest baseline TIGER, leaves a wide margin.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix B.2 reports the hardware, three NVIDIA A6000 GPUs with 48 GB each, and the wall-clock time of each training stage, about 92 hours in total for PRISM. Baseline retraining and preliminary experiments required additional compute that is not itemised.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work conforms with the NeurIPS Code of Ethics. All participants gave written informed consent under institutional ethics approval, and RGB frames, which could identify participants, will be withheld from any public release.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.

- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Section 1 motivates through-obstacle sensing for healthcare and elderly-care monitoring. Section 6 discusses the negative side: although mmWave radar records no facial or clothing appearance, it still reveals a person's presence, pose and activity, so through-wall use could enable monitoring without the person's knowledge or consent.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The model reconstructs body geometry from radar, and any public release of the dataset will exclude RGB frames; neither belongs to the high-risk categories this question targets, such as pretrained language or image generators or scraped datasets.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The existing assets used are the MM-Fi dataset and the public implementations of the seven baselines, all cited in the paper; their versions and licenses will be listed with the code release.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: MIST is documented in Section 3 and Appendix A, covering the sensors, the thirty actions, the occlusion levels, the directory layout and the file formats, and a README and license will accompany the public release.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: Section 3 and Appendix A describe the protocol each participant followed: thirty scripted actions, each performed repeatedly for 30 s, under clear-view and through-obstacle conditions. All fifteen participants gave written informed consent and were compensated at the local minimum wage, and the participant instructions will be provided with the dataset release.

    Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: Data collection was approved by the Institutional Review Board of Nanyang Technological University. The protocol involves only low-risk everyday and rehabilitation movements, the risks were explained to all participants before they gave written consent, and no vulnerable populations were recruited.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are not part of PRISM or of the experimental pipeline; the method uses only point-cloud encoders, a graph convolutional network, a hierarchical VAE and a Transformer denoiser.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.