RadarLLM: Empowering Large Language Models to Understand Human Motion from Millimeter-Wave Point Cloud Sequence

Zengyuan Lai[1*], Jiarui Yang[1*], Songpengcheng Xia[1*], Lizhou Lin[1], Lan Sun[1], Renwen Wang[2], Jianran Liu[2], Qi Wu[2], Ling Pei[1†]
[1]Shanghai Jiao Tong University
[2]Bytedance Research
{zy.lai, jr.yang, songpengchengxia, ling.pei}@sjtu.edu.cn, {wangrenwen, liujianran}@bytedance.com

Abstract

Millimeter-wave radar offers a privacy-preserving and environment-robust alternative to vision-based sensing, enabling human motion analysis in challenging conditions such as low light, occlusions, rain, or smoke. However, its sparse point clouds pose significant challenges for semantic understanding. We present RadarLLM, the first framework that leverages large language models (LLMs) for human motion understanding from radar signals. RadarLLM introduces two key innovations: (1) a motion-guided radar tokenizer based on our Aggregate VQ-VAE architecture, integrating deformable body templates and masked trajectory modeling to convert spatial-temporal radar sequences into compact semantic tokens; and (2) a radar-aware language model that establishes cross-modal alignment between radar and text in a shared embedding space. To overcome the scarcity of paired radar-text data, we generate a realistic radar-text dataset from motion-text datasets with a physics-aware synthesis pipeline. Extensive experiments on both synthetic and real-world benchmarks show that RadarLLM achieves state-of-the-art performance, enabling robust and interpretable motion understanding under privacy and visibility constraints, even in adverse environments.

Code — https://inowlzy.github.io/RadarLLM/
Extended version — https://arxiv.org/abs/2504.09862

1 Introduction

Human motion understanding is critical in applications such as elderly care, smart home automation, and health monitoring (Lai et al. 2024; Zhang et al. 2024; Xia et al. 2025; Shan et al. 2025). These scenarios require robust and non-intrusive sensing technologies capable of analyzing human activities while preserving privacy (Xia et al. 2024; Song et al. 2025; Xu 2025). However, traditional vision-based systems face significant limitations due to lighting variations, occlusions, and privacy concerns, making them unsuitable for real-world, long-term deployment.

Millimeter-wave (mmWave) radar offers a promising alternative, providing privacy-preserving motion sensing that is robust to poor lighting, occlusions, rain, and smoke, and does not capture visual identity (Ding et al. 2024; Gu et al. 2025). Despite recent advancements in radar-based methods for activity recognition (Meng et al. 2020; Cao et al. 2024) and pose estimation (Yang et al. 2025a), these methods primarily focus on classification or regression tasks, limiting their ability to generate fine-grained motion descriptions.

Meanwhile, large language models (LLMs) show strong capabilities in semantic reasoning across modalities like vision, audio, and motion (Cho et al. 2025; Huang et al. 2024b; Jiang et al. 2023). Inspired by this, we propose RadarLLM, the first framework bridging mmWave radar sensing and language understanding for semantic-rich motion analysis (Figure 1). However, applying LLMs to radar data poses two challenges: (1) sparse, noisy point clouds hinder spatiotemporal modeling; (2) the semantic gap between radar signals and language requires sophisticated cross-modal alignment.

To address these challenges, we introduce two key components: (1) a Motion-guided Radar Tokenizer, based on an Aggregate VQ-VAE, that encodes radar point cloud sequences into discrete semantic tokens via deformable body templates and masked trajectory modeling; (2) a Radar-aware Language Model, which aligns radar tokens with textual representations in a shared embedding space to generate motion descriptions. Moreover, to overcome the lack of paired radar-text data, we propose a physics-aware synthesis pipeline that simulates realistic radar reflections from motion-text datasets, enabling effective training at scale. In summary, our main contributions of this work are as follows:

- We propose RadarLLM, the first LLM-based framework that translates low-level radar point clouds into high-level semantic motion descriptions, pioneering a new paradigm for privacy-preserving motion understanding.
- We introduce a novel Aggregate VQ-VAE-based Motion-Guided Radar Tokenizer. It encodes sparse radar sequences into LLM-compatible semantic tokens, leveraging deformable body templates for structural priors and masked trajectory modeling for dependency learning.
- We develop a physics-aware virtual radar simulator that synthesizes realistic radar-text data from motion-text datasets, effectively bypassing the bottleneck of paired real-world data scarcity for large-scale training.

[*]These authors contributed equally.
[†]Corresponding Author

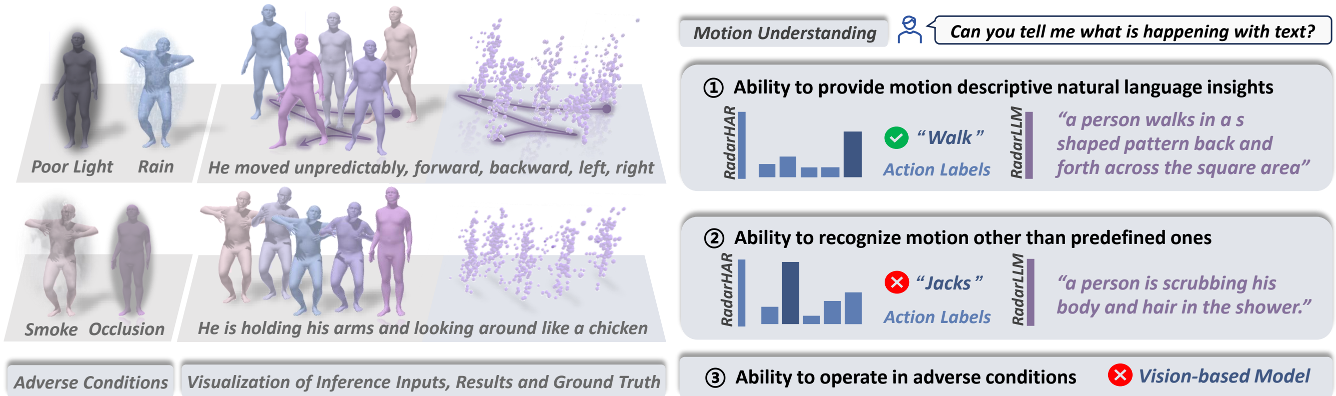


Figure 1: We propose RadarLLM, a LLM-based radar-text human motion understanding framework over traditional action label-based motion recognition in providing descriptive natural language insights, recognizing unconventional motions beyond predefined categories, and operating robustly in adverse conditions (e.g., poor lighting, occlusion, rain, and smoke).

2 Related Works

2.1 Radar-Based Human Motion Understanding

Millimeter-wave radar is robust, privacy-preserving, and effective under adverse conditions (Kong et al. 2024; Saadat and Sur 2024; Yang et al. 2025a). Early methods extracted handcrafted Micro-Doppler features and utilized classifiers such as SVM (Suykens 2001) and RF (Pal 2005), achieving around 91% accuracy on basic actions (Javier and Kim 2014; Smith et al. 2018). Deep-learning architectures, including LSTMs (Shrestha et al. 2020) and dual-stream CNNs (Ding et al. 2022), automated feature learning to reach 94–99% accuracy in controlled conditions (Ding et al. 2022; Kang et al. 2023). Spatial-temporal transformers (ST-PCT (Kang et al. 2023)) and point-based paradigms like milliFlow (Ding et al. 2024) further enhanced generalization. However, these approaches remain limited to low-level gesture classification, lacking semantic interpretation of composite activities (Cao et al. 2024; Yang et al. 2024) — a challenge that remains largely unaddressed.

2.2 Human Motion Understanding with Multimodal LLMs

Despite advances in sensor-based classification using RGB, IMU, and skeleton sequences (Xia et al. 2024; Haresamudram et al. 2025; Lu et al. 2025), such methods are constrained by fixed action sets and limited semantic depth. To overcome these challenges, motion-to-text frameworks have emerged, encoding discrete motion tokens for LLM translation (e.g., MotionGPT (Jiang et al. 2023), PointLLM (Xu et al. 2024), AvatarGPT (Zhou, Wan, and Wang 2024)) and aligning motion with natural language semantics. Building on this paradigm, Mojito extended tokenization to IMU signals (Shan et al. 2025), while vision-language models have integrated visual motion cues and textual descriptions for richer interpretation (Li et al. 2024). However, radar-based LLM reasoning remains largely unexplored, a critical gap our work addresses to provide a privacy-preserving, contactless solution that eliminates the need for wearable sensors while ensuring robust perception, even in adverse scenes.

2.3 mmWave Radar Signal Generation

Large-scale radar–text datasets are essential for training LLM-based motion understanding models. However, existing collections (<9h duration, <40 participants, <27 classes) (Liu et al. 2020; Wang et al. 2021; Yang et al. 2023; Wang et al. 2024) also lack paired textual annotations (Chen et al. 2022; An, Li, and Ogras 2022). To bridge this gap, researchers have turned to synthetic data generation—but efficiency and realism remain at odds: Vid2Dopplerr (Ahuja et al. 2021) delivers speed at the expense of spatial detail, while RF-Genesis (Chen and Zhang 2023) achieves physical fidelity only with prohibitive compute demands. These limitations have driven the emergence of our physics-informed synthesis, which embeds motion-capture priors to strike a balance between kinematic authenticity and scalability. At the same time, vision–language studies (Deng et al. 2023a,b) and lidar-based works (An et al. 2025) demonstrate that such physically grounded synthetic data can closely approximate real-world annotations. Building on these insights, we adopt SMPL-X-text pairs from HumanML3D (Guo et al. 2022) to produce a richly annotated Radar–Text corpus via physics-aware signal synthesis for training RadarLLM.

3 Radar-Text Dataset Preparation

Building an end-to-end mmWave-based motion understanding LLM is fundamentally hindered by the absence of paired radar point clouds and natural language annotations. Existing datasets with language labels focus on RGB and motion sequences, unsuitable for radar-based training. Inspired by virtual radar data generation methods (Deng et al. 2023b; Chen and Zhang 2023), we construct a large-scale virtual radar–text dataset from HumanML3D (Guo et al. 2022), comprising 13,308 SMPL-X motion sequences from AMASS (Mahmood et al. 2019) paired with text annotations. To assess performance in real scenes, we collect a real-world test set covering one normal and four adverse environments.

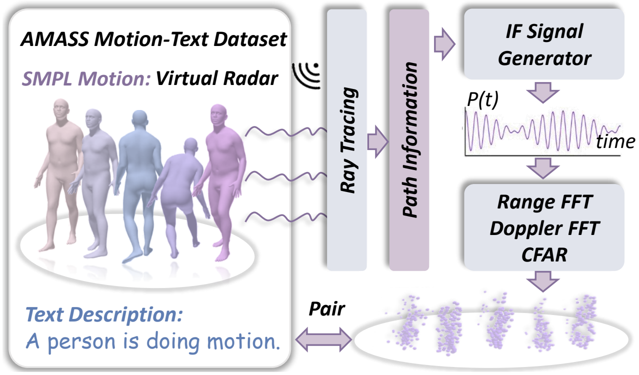


Figure 2: Virtual radar-text data generation pipeline. The Radar-Text dataset is constructed by simulating radar reflections from SMPL-X sequences using ray tracing and signal processing, based on existing motion-text datasets.

3.1 Virtual Data Preparation

To synthesize mmWave radar signals from human motions, we employ a physics-aware synthesis. As illustrated in Figure 2, the resulting point cloud sequences are paired with text descriptions to form our virtual radar–text dataset.

IF Signal Simulation The first step is to simulate the ray paths between the rendered human meshes and the virtual radar antennas. Traditional ray tracing relies on Monte Carlo sampling, which causes a trade-off between accuracy and computational cost. So we adopt the RF adaptive sampling technique (Chen and Zhang 2023) to focus on the human body by edge detection. Then the path information for each ray is accumulated using the Physical Optics Integral (POI) method to obtain the simulated IF signal.

Point Cloud Generation After sampling on the simulated IF signal, we first compute the Range-FFT and Doppler-FFT, followed by applying a static clutter removal algorithm to eliminate background noise by subtracting the average Doppler-FFT heatmap values across all receiving antennas. Instead of using the traditional CFAR algorithm with fixed thresholds, we directly select the 128 points per frame based on intensity from the Doppler-FFT heatmap (Xue et al. 2021), and it ensures a consistent number of point clouds.

3.2 Real Data Preparation

To further evaluate our method in real scenes, we collect a real-world test set of 125 distinct motions from the HumanML3D test split, each repeated three times (375 sequences, 6–9s), recorded with a TI AWR1843BOOST radar and DCA1000EVM board under normal conditions. Another test set of four adverse conditions—rain, smoke, poor lighting, and occlusions—is created by downsampling the public mmwave dataset MMBody's (Chen et al. 2022) point clouds to the 128 highest-intensity points per frame. Point cloud synthesis follows our virtual data pipeline, and text annotations are first generated by MotionGPT conditioned on paired SMPL-X ground truth and manually checked (Wang et al. 2025).

4 Method

4.1 Problem Statement

Our goal is to enable a semantic-rich understanding of human motion from radar point clouds, moving beyond conventional activity classification to natural language interpretation. Given a millimeter-wave radar point cloud sequence $\mathbf{P}_{1:T} = [p_1, \dots, p_T]$, where each frame $p_t \in \mathbb{R}^{N_t \times 4}$ contains the spatiotemporal coordinates (x, y, z, t) of body reflections, we aim to generate a descriptive text sequence $\mathbf{Y} = [y_1, \dots, y_L]$, $y_l \in \mathcal{V}_{\text{text}}$ (predefined WordPieces vocabulary (Song et al. 2020)). Unlike prior methods that predict discrete action labels $\hat{Y}_R$, our approach translates radar observations directly into structured textual descriptions.

To achieve this, we propose RadarLLM with three core modules: (1) constructing a virtual radar–text dataset $\{(\mathbf{P}^{syn}_{1:T}, \mathbf{Y})\}$ by simulating radar reflections from human motions $\mathbf{M}_{1:T}$ via ray tracing; (2) encoding noisy sequences $\mathbf{P}^{syn}_{1:T}$ into semantic tokens $\mathbf{z}_{1:L} = \{\mathbf{z}_i\}_{i=1}^{L}$ through an Aggregate VQ-VAE, where $\mathbf{z}_i \in \{1, \dots, K\}$ denotes a discretized motion code, K is the total numbers of codes forming a codebook, $L = T/r$, and r is the temporal downsampling rate; (3) training a conditional radar-aware language model to generate textual motion descriptions.

4.2 RadarLLM Model Architecture

To bridge radar point clouds and natural language, we introduce ***RadarLLM***, a unified framework that integrates mmWave radar with large language models. As shown in Figure 3, RadarLLM comprises (1) a *motion-guided radar tokenizer* that converts point cloud sequences into discrete semantic tokens, and (2) a *radar-aware language model* trained via multi-modal pre-training and task-specific fine-tuning to align these tokens with textual semantics.

Motion-Guided Radar Tokenizer To extract quantized semantic features from sparse, noisy radar point clouds for the injection into the language model, we compress spatial–temporal patterns into discrete codes via our designed ***Aggregate VQ-VAE***, leveraging human template priors and mask-enhanced spatio–temporal dependency learning, as well as motion semantics guidance. This comprises three stages: template-prior grouping, masked context aggregation, and aggregated quantization, shown in Fig 4.
(1) Template-Prior Grouping. To overcome the inconsistency of point location and counts among frames, we initialize N_g anchors on a deterministic $N_x \times N_y \times N_z$ grid within a bounding-box template to construct temporal associations for each body region around anchors, then aggregate neighborhood points with the SOTA P4Conv encoder $\mathbf{E}$ in (Fan, Yang, and Kankanhalli 2021; Jing et al. 2024; Yang et al. 2025a), yielding $\mathbf{F}_{group} \in \mathbb{R}^{L \times N_g \times C}$.
(2) Masked Context Aggregation. To enhance the learning of the dependencies among body parts, 50% of anchor trajectories are masked to form visible features $\mathbf{F}_{vis} \in \mathbb{R}^{L \times N_g^{vis} \times C}$; a transformer decoder reconstructs $\mathbf{F}_{msk} = D(\mathbf{F}_{vis})$ via cross-attention, and merging with $\mathbf{F}_{vis}$ produces $\mathbf{F}_{all} = [\mathbf{F}_{vis}, \mathbf{F}_{msk}]$, expected to approach motion

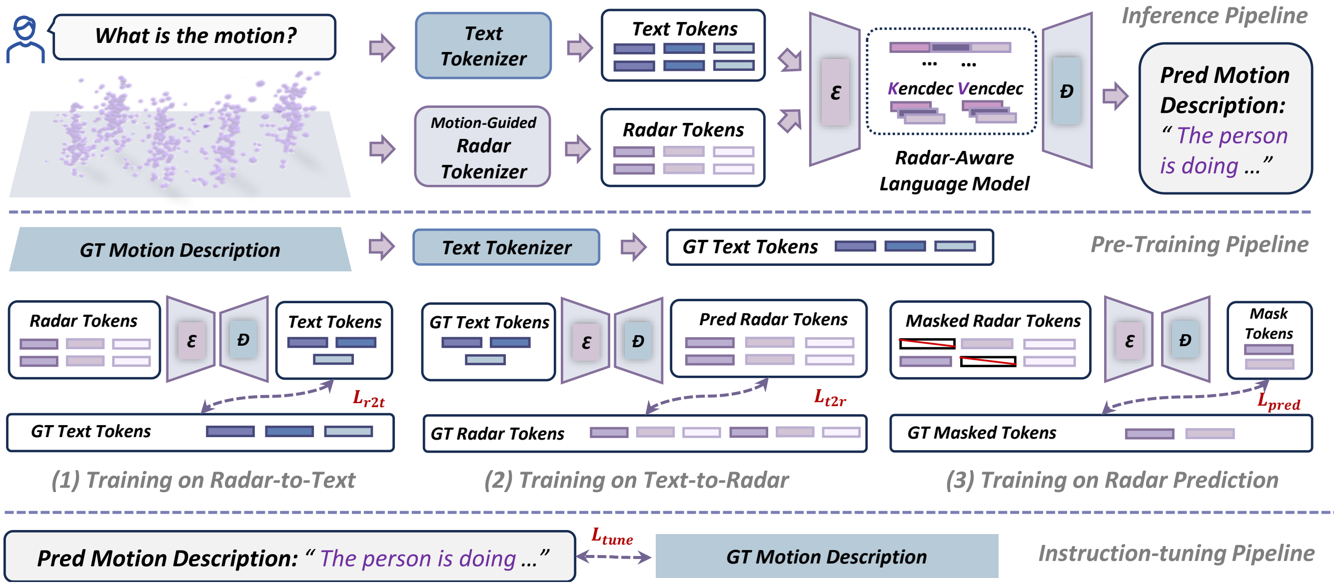


Figure 3: The overview of RadarLLM. We first encode radar point clouds into discrete tokens via a Motion-guided Radar Tokenizer. The Radar-aware Language Model then aligns these tokens with textual representations in a shared embedding space through joint optimization of unsupervised token reconstruction and supervised bidirectional radar-text translation.

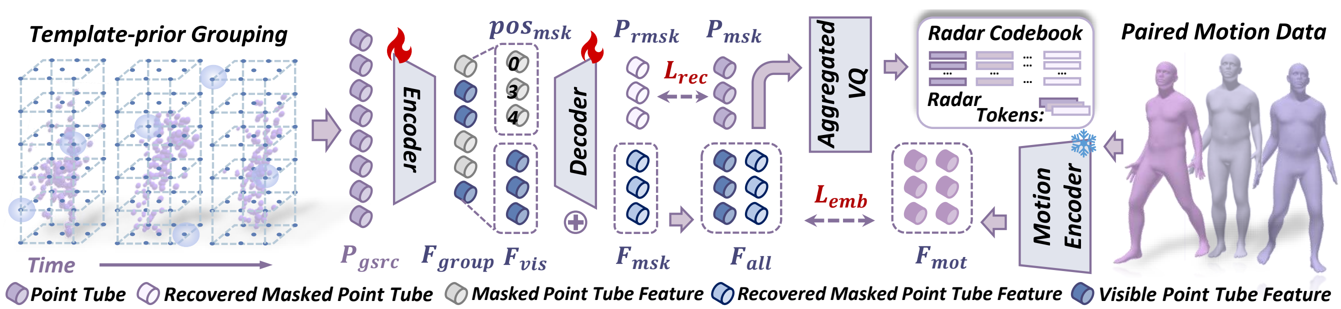


Figure 4: Architecture and training pipeline of motion-guided radar tokenizer. The Motion-guided Radar Tokenizer, built upon our Aggregate VQ-VAE architecture, compresses radar point cloud sequences into discrete semantic tokens through point cloud sequence reconstruction and motion embedding learning.

semantic features extracted from the paired motion by a motion encoder, enriching semantics learned efficiently.

(3) Aggregated Quantization. Aggregated quantization then maps each $\mathbf{F}^t_{all}$ to its nearest code in the trainable codebook $\mathcal{Z} = \{\mathbf{z}_k\}_{k=1}^{K} \subset \mathbb{R}^{512\times512}$ to construct a discrete token sequence aligned with the language model's embedding space, enabling direct cross-modal translation, where

$$\mathbf{z}_t = \arg\min_{\mathbf{z}_k\in\mathcal{Z}} \|\mathbf{F}^t_{all} - \mathbf{z}_k\|_2, \quad t = 1, \ldots, L. \tag{1}$$

An ablation study for the effectiveness of vector quantization can be found in the supplementary file.

(4) Train and Inference Paradigm. In training, we optimize

$$\mathcal{L}_{\text{VQ}} = \mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{emb}} + \mathcal{L}_{\text{commit}}, \tag{2}$$

where

$$\mathcal{L}_{\text{rec}} = \frac{1}{|\mathbf{P}_{msk}|}\sum_{\mathbf{x}\in\mathbf{P}_{msk}} \min_{\mathbf{y}\in\mathbf{P}_{rmsk}} \|\mathbf{x}-\mathbf{y}\|^2, \tag{3}$$

forming the Chamfer Distance in (Shen et al. 2023a,b) to promote learning the spatial-temporal features of point cloud sequence via the sequence reconstruction,

$$\mathcal{L}_{\text{emb}} = \|\mathbf{F}_{all} - \mathbf{F}_{mot}\|_2^2 \tag{4}$$

forming the motion guidance via aligning radar features with motion semantics, which is expected to accelerate feature learning. $\mathbf{F}_{mot}$ is the corresponding motion features.

$$\mathcal{L}_{\text{commit}} = \|\text{sg}[\mathbf{F}_{all}] - \mathbf{z}\|_2^2 + \|\mathbf{F}_{all} - \text{sg}[\mathbf{z}]\|_2^2. \tag{5}$$

forming the commitment loss to stabilize codebook learning and enforce encoder–codebook consistency.

In inference, no masking is applied, allowing robust encoding via learned dependencies.

Radar-Aware Language Model To establish semantic equivalence between continuous radar patterns and discrete text and enable cross-modal reasoning, we tokenize radar clouds into discrete tokens $\mathbf{z}_{1:L}$, map them to indices $\mathbf{s}_{1:L} \in \{1, \ldots, K\}^L$, and merge with word tokens in a unified vocabulary $\mathcal{V} = \mathcal{V}_{\text{text}} \cup \mathcal{V}_{\text{radar}}$ (32,768 WordPieces+K radar tokens+`/som`, `/eom`). The combined input

$$\mathbf{X} = [w_1, \ldots, w_{L_t},\, s_1, \ldots, s_L] \tag{6}$$

is fed into a modified T5 (Jiang et al. 2023): shared embeddings project all tokens to 512 dimensions, cross-modal attention learns joint context, and the decoder autoregressively predicts text $\mathbf{Y} = [y_1, \ldots, y_L]$.

To effectively align radar tokens with textual semantics while ensuring adaptability to diverse motion-to-text instructions, we employ a two-stage training process: a pre-training stage for robust cross-modal representation learning and an instruction-tuning stage for task-specific refinement.

(1) Pre-Training Stage. To learn robust cross-modal representations, we adopt a multi-task training paradigm (Jiang et al. 2023; Zhou, Wan, and Wang 2024).

- *Radar Prediction:* Following the span corruption strategy of T5 (Jiang et al. 2023), we randomly mask 15% of radar tokens and replace them with sentinel tokens. The model predicts original tokens through:

$$\mathcal{L}_{\text{pred}} = -\sum_{i \in \mathcal{M}} \log p(s_i | s_{\mathcal{M}}), \tag{7}$$

  where $\mathcal{M}$ denotes masked positions.
- *Radar→Text*: Encode radar tokens $\mathbf{z}_{1:L}$, decode text $\mathbf{w}_{1:L}$:

$$\mathcal{L}_{\text{r2t}} = -\sum_{t=1}^{L} \log p(w_t | \mathbf{z}_{1:L}, \mathbf{w}_{<t}). \tag{8}$$

- *Text→Radar*: Encode text $\mathbf{w}_{1:L}$, autoregressively generate radar tokens:

$$\mathcal{L}_{\text{t2r}} = -\sum_{t=1}^{L} \log p(z_t | \mathbf{w}_{1:L}, \mathbf{z}_{<t}). \tag{9}$$

The total pretraining loss combines these objectives:

$$\mathcal{L}_{\text{pretrain}} = \lambda_1 \mathcal{L}_{\text{pred}} + \lambda_2 \mathcal{L}_{\text{r2t}} + \lambda_3 \mathcal{L}_{\text{t2r}}, \tag{10}$$

where $\lambda_1, \lambda_2, \lambda_3$ are the hyperparameters for balancing each loss's contribution.

The ablations on language model selection and multi-task training strategy are detailed in the Experiments section.

(2) Instruction-Tuning Stage. To enhance task adaptability, we adopt instruction-aware prompts (e.g. "Describe the motion <Motion_Placeholder>...") concatenated with $\mathbf{z}$ (Xu et al. 2024), and refine alignment via a similarity-based tuning loss $\mathcal{L}_{tune}$ against ground-truth descriptions.

5 Experiments

This section first introduces competing methods and metrics in our experiment (Sec. 5.1). We then present comprehensive comparisons on the radar-to-text task across both virtual and real datasets (Sec. 5.2). Finally, we conduct ablation studies to validate the effectiveness of each module (Sec. 5.3) and evaluate robustness in adverse environments (Sec. 5.4). Additional experiment results and discussions are provided in the supplementary material.

5.1 Experimental Setup

Baselines To our knowledge, RadarLLM is the first end-to-end radar-to-text framework for human motion understanding using mmWave point clouds, with no directly comparable baselines. Following the two-stage evaluation protocol of PointLLM (Xu et al. 2024) and LidarLLM (Yang et al. 2025b), we retrain the real-time 3D human pose estimator mmMesh (Xue et al. 2021) on our radar dataset to convert sparse point clouds into SMPL-X meshes. These SMPL-X meshes are then rendered as videos or converted into skeleton sequences, which serve as inputs to state-of-the-art video- and motion-based text generation models, respectively, including MotionGPT (Jiang et al. 2023), AvatarGPT (Zhou, Wan, and Wang 2024), Video-LLaMA2 (Cheng et al. 2024), Video-ChatGPT (Maaz et al. 2023), Video-LLaVa (Lin et al. 2023), and VTimeLLM (Huang et al. 2024a). Although these general-purpose models are originally trained on substantially larger vision and motion corpora, this unified two-stage setup ensures a fair comparison using identical inputs and underscores the strengths of our end-to-end design.

Datasets As mentioned in Sec. 3, using HumanML3D splits (Sec. 3), we train on virtual train-set and evaluate on virtual/real test-set. Adverse condition evaluation is tested on the mentioned down-sampled MMBody subsets.

Evaluation Metrics Following previous multi-modal text generation works (Jiang et al. 2023; Xu et al. 2024), we use Rouge (Lin 2004), BLEU (Zhang et al. 2019), METEOR (Banerjee and Lavie 2005), Cider (Vedantam, Lawrence Zitnick, and Parikh 2015), BertScore (Zhang et al. 2019) and SimCSE (Gao, Yao, and Chen 2021) to evaluate the quality of generated captions.

5.2 Comparisons on Radar-to-Text

To demonstrate superior cross-modal understanding performance, we conduct the radar-to-text experiments on both virtual and real data. The aforementioned radar-based HPE mmMesh and our proposed RadarLLM are trained fully on the virtual dataset, and the pre-trained motion- and video-based models are adopted to generate the text descriptions. The quantitative results on virtual and real test datasets are shown in Table 1. Our method achieves state-of-the-art performance across all metrics on the virtual test dataset, outperforming the strongest baseline (AvatarGPT) by +20.0% ROUGE-L, +22.1% CIDE, and +128% BLEU-4 improvement, indicating better preservation of motion semantics in textual descriptions. To further evaluate the performance in real scenes, we conduct experiments on collected test data; our method remains best on almost all metrics, demonstrating the superior generalization ability to real data.

Table 2 presents qualitative comparisons of textual descriptions generated from our method and two best baselines. Our method outperforms AvatarGPT and Video-LLaMA2 by capturing fine-grained motion details and contextual semantics more accurately, while baseline methods produce generic or simple descriptions.

Model	Data Domain	ROUGE-1	ROUGE-L	BLEU-1	BLEU-4	METEOR	CIDEr	BERTScore	SimCSE
MotionGPT* (Jiang et al. 2023)	Virtual	31.2	29.4	37.6	5.0	26.1	6.5	82.6	88.9
	Real	28.0	25.6	36.1	2.9	21.9	3.2	80.5	87.2
AvatarGPT* (Zhou, Wan, and Wang 2024)	Virtual	32.2	30.0	36.3	5.0	28.3	6.8	82.4	88.7
	Real	31.0	28.8	38.1	4.2	25.6	5.6	81.4	88.1
Video-LLaMA2* (Cheng et al. 2024)	Virtual	30.2	26.7	35.2	3.6	30.4	4.2	81.0	88.4
	Real	31.4	28.8	38.3	4.3	**28.6**	**7.0**	80.1	88.0
VideoChatGPT* (Maaz et al. 2023)	Virtual	18.6	16.1	19.5	0.8	15.3	1.0	78.5	85.7
	Real	17.8	15.6	19.4	0.1	13.0	1.2	77.6	85.0
Video-LLaVA* (Lin et al. 2023)	Virtual	22.8	19.2	26.7	1.3	19.2	2.1	80.3	87.6
	Real	22.6	19.3	27.2	1.4	17.4	3.5	79.7	87.2
VTimeLLM* (Huang et al. 2024a)	Virtual	19.1	15.8	19.0	0.9	17.7	1.4	79.8	87.4
	Real	21.3	16.9	24.1	0.8	18.0	2.1	79.6	87.3
RadarLLM (Ours)	Virtual	**38.4**	**36.0**	**48.0**	**11.4**	**33.7**	**8.3**	**83.3**	**89.6**
	Real	**31.7**	**28.8**	**44.2**	**5.0**	25.7	4.0	**81.4**	**88.1**

Table 1: Comparison with state-of-the-art methods on virtual and real datasets

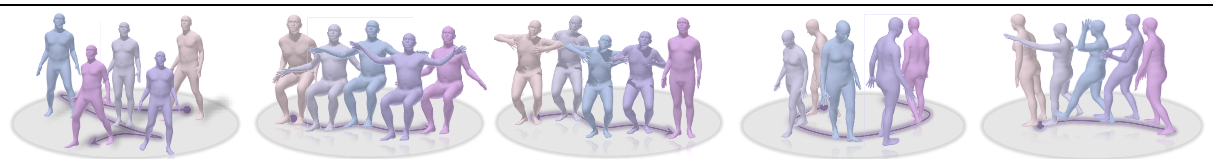
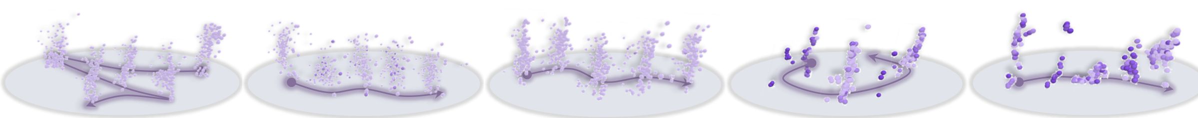
Ground Truth Motions					
Input Radar Point Cloud Sequences					
Real	He moved unpredictably, **forward, backward, left, right**.	He sat, **extended both arms to the sides**, then brought them **down**.	The person is **holding his arms** and looking around **like a chick**.	A person **walks** in a **counter clockwise circle**.	He is **pushed by an unseen force**, and then **recovers** to standing.
AvatarGPT	He stands hands in front, uses his left hand to twist/turn something.	He arms out, looks down, raising/lowering right then left arm.	A person stands, brings arms to chest, twists both counterclockwise.	A person raises their arms and turns their torso to the right.	A person takes a step forward with their right foot.
Video-LLaMA2	A person is walking slowly and then turns around.	A person is walking towards the right side of the screen.	A person is walking forward in a straight line.	A person is shown walking toward the camera, then turns away.	A mannequin is shown from different angles as it moves up and down.
Ours	A person **walks in an S-shape back and forth** across a **square area**.	A person **waves both arms and hands** as if to signal to someone.	A person is **scrubbing his body and hair** in the shower.	A person **walks** in a **counter clockwise circle**.	A person is **pushed by an unseen force**, before they **recover**.

Table 2: Visualization of predicted text descriptions with corresponding motion and radar point cloud sequences. Left three columns: real-world normal environment results; right two columns: synthetic virtual data predictions.

5.3 Ablation Study

Effectiveness of Aggregate VQ-VAE To thoroughly evaluate the architectural contributions of our radar tokenizer, we conduct ablation studies on three key components at the AMASS test dataset. We first replace the template-based anchor grouping mechanism with the vanilla Farthest Point Sampling (FPS), the results in Table 3 show severe performance degradation of 27.3% in ROUGE-1, highlighting the importance of consistent spatial-semantic correspondence. We alter the training objective by reconstructing the full point cloud sequence instead of recovering the masked point tube, following the traditional self-supervised learning strategy in VAE. This change leads to a 23.7% drop in BLEU-4 scores, demonstrating the effectiveness of our masked point tube recovery approach. Finally, by removing the embedding loss, it proves crucial for cross-modal alignment, with its absence leading to 54.2% lower CIDEr scores.

Effectiveness of LLM Selection To assess LLM selection effectiveness and scalability across model sizes, Table 4 demonstrates that T5-Small achieves the fastest inference

Model	ROUGE-1	ROUGE-L	BLEU-1	BLEU-4	METEOR	CIDEr	BERTScore	SimCSE
w/o template-based anchor	27.9	25.7	34.8	3.8	22.5	3.2	81.1	87.8
w/o mask for training	35.0	32.4	43.1	8.7	31.0	**11.3**	83.2	89.5
w/o embedding loss	28.6	26.5	35.4	4.2	23.5	3.8	81.5	88.1
RadarLLM	**38.4**	**36.0**	**48.0**	**11.4**	**33.7**	8.3	**83.3**	**89.6**

Table 3: Ablation study on the Aggregate VQ-VAE components.

LLM Model	Params	FPS↑	Self-BLEU↓	ROUGE-L↑	SimCSE↑
T5-small	60M	**97.0**	**92.2**	36.0	89.6
GPT2-M	355M	72.7	96.2	35.4	89.5
Deepseek-R1	1.8B	53.6	98.4	**37.4**	**89.9**

Table 4: Ablation study on different LLM architectures.

Task	ROUGE-L	BLEU-1	METEOR	BERTScore
R→T	33.0	42.8	31.2	82.5
R→T & T→R	33.0	43.1	31.2	82.5
R→T & R–Pred	33.9	43.2	32.4	82.9
All Tasks	**36.0**	**48.0**	**33.7**	**83.3**

Table 5: Ablation of multi-task training strategy

Adverse Scenes		
Real	A person **brushes something** in the front.	A person **swings his arm and leg in place**.
Ours	A person is **scrubbing a window.**	A person is **jogging on the spot**.
Adverse Scenes		
Real	A person **shakes his hips and arms** while dancing.	A person **steps slightly towards left and front**.
Ours	A person is **scrubbing his body&hair** in the shower.	A person **walks forward** slowly.

Table 6: Visualization of predictions in adverse conditions.

and highest diversity, with only minor semantic degradation. GPT2-Medium strikes a middle ground in speed (FPS) and diversity (Self-BLEU), albeit with somewhat lower semantic precision, while DeepSeek-R1 delivers the best lexical and semantic scores at the expense of throughput and increased repetition. Under limited resources, LoRA tuning on GPT2 and DeepSeek may reinforce frequent token patterns, suggesting that full fine-tuning under sufficient resources can be explored; Thus, T5-Small remains the most balanced choice.

Effectiveness of Multi-Task Training Strategy To evaluate multi-task objectives, we compare four schemes in Table 5: Radar→Text (R→T), R→T + Text→Radar (T→R), R→T + radar-prediction (R–Pred), and all three tasks. Adding T→R yields the arising of BLEU-1 (+0.7%). Incorporating R–Pred improves ROUGE-L by 2.7%, BLEU-1 by 0.9%, METEOR by 3.8%, and BERTScore by 0.5%. Jointly optimizing all objectives boosts ROUGE-L by 9.1%, BLEU-1 by 12.2%, METEOR by 8.0%, and BERTScore by 1.0%, confirming that multi-task learning substantially enhances motion-to-text performance.

5.4 Robustness Under Adverse Environments

To assess the robustness of RadarLLM under adverse environmental conditions, we evaluated it on four corrupted subsets of the MMBody dataset—rain, smoke, poor lighting, and occlusions. ROUGE-L declines by only 14.2% (28.8→24.7) and SimCSE by just 1.4% (88.1→86.9), revealing the robustness in multi-word precision and overall semantics, whereas BLEU-1 and METEOR drop by 46.8% (44.2→23.5) and 22.2% (25.7→20.0), respectively, but mainly influence the single word similarity and fluency slightly. These results confirm that our model preserves semantic coherence to some extent under realistic adverse conditions. Table 6 visualizes example outputs across the four scenarios to show our robustness.

6 Conclusion

In this work, we introduce RadarLLM, the first end-to-end LLM-based framework for semantic motion understanding from mmWave point clouds, using physics-aware signal synthesis to create realistic radar–text pairs for addressing the data scarcity challenge. Experiments on synthetic and real data in various conditions claim state-of-the-art results.

Limitations and Future Work While RadarLLM demonstrates strong performance, several avenues remain for future exploration. First, although we calibrate simulation parameters to our TI AWR1843BOOST setup, evaluating and adapting RadarLLM across diverse radar hardware—with varying point-cloud densities and range/angle resolutions—could further validate its generality. Second, extending our dataset to include environmental context and human–object interactions would enrich scene understanding and broaden applicability to complex real-world scenes.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (NSFC) under Grant 62273229.

References

Ahuja, K.; Jiang, Y.; Goel, M.; and Harrison, C. 2021. Vid2doppler: Synthesizing doppler radar data from videos for training privacy-preserving activity recognition. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–10.

An, S.; Li, Y.; and Ogras, U. 2022. mri: Multi-modal 3d human pose estimation dataset using mmwave, rgb-d, and inertial sensors. *Advances in Neural Information Processing Systems (NeurIPS)*, 35: 27414–27426.

An, X.; Zhao, L.; Gong, C.; Li, J.; and Yang, J. 2025. Pre-training a Density-Aware Pose Transformer for Robust LiDAR-based 3D Human Pose Estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 1755–1763.

Banerjee, S.; and Lavie, A. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 65–72.

Cao, Q.; Xue, H.; Liu, T.; Wang, X.; Wang, H.; Zhang, X.; and Su, L. 2024. mmCLIP: Boosting mmWave-based Zero-shot HAR via Signal-Text Alignment. In *Proceedings of the 22nd ACM Conference on Embedded Networked Sensor Systems*, 184–197.

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

Chen, X.; and Zhang, X. 2023. Rf genesis: Zero-shot generalization of mmwave sensing through simulation-based data synthesis and generative diffusion models. In *Proceedings of the 21st ACM Conference on Embedded Networked Sensor Systems*, 28–42.

Cheng, Z.; Leng, S.; Zhang, H.; Xin, Y.; Li, X.; Chen, G.; Zhu, Y.; Zhang, W.; Luo, Z.; Zhao, D.; et al. 2024. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*.

Cho, C.-H.; Moon, W.; Jun, W.; Jung, M.; and Heo, J.-P. 2025. Ambiguity-Restrained Text-Video Representation Learning for Partially Relevant Video Retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 2500–2508.

Deng, K.; Zhao, D.; Han, Q.; Zhang, Z.; Wang, S.; Zhou, A.; and Ma, H. 2023a. Midas: Generating mmWave radar data from videos for training pervasive and privacy-preserving human sensing tasks. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)*, 7(1): 1–26.

Deng, K.; Zhao, D.; Zhang, Z.; Wang, S.; Zheng, W.; and Ma, H. 2023b. Midas++: generating training data of mmWave radars from videos for privacy-preserving human sensing with mobility. *IEEE Transactions on Mobile Computing (TMC)*, 23(6): 6650–6666.

Ding, C.; Zhang, L.; Chen, H.; Hong, H.; Zhu, X.; and Li, C. 2022. Human motion recognition with spatial-temporal-convLSTM network using dynamic range-doppler frames based on portable FMCW radar. *IEEE Transactions on Microwave Theory and Techniques*, 70(11): 5029–5038.

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

Fan, H.; Yang, Y.; and Kankanhalli, M. 2021. Point 4d transformer networks for spatio-temporal modeling in point cloud videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 14204–14213.

Gao, T.; Yao, X.; and Chen, D. 2021. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*.

Gu, Z.; Ma, J.; Huang, Y.; Wei, H.; Chen, Z.; Zhang, H.; and Hong, W. 2025. HGSFusion: Radar-camera fusion with hybrid generation and synchronization for 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 3185–3193.

Guo, C.; Zou, S.; Zuo, X.; Wang, S.; Ji, W.; Li, X.; and Cheng, L. 2022. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5152–5161.

Haresamudram, H.; Beedu, A.; Rabbi, M.; Saha, S.; Essa, I.; and Ploetz, T. 2025. Limitations in Employing Natural Language Supervision for Sensor-Based Human Activity Recognition-And Ways to Overcome Them. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 273–281.

Huang, B.; Wang, X.; Chen, H.; Song, Z.; and Zhu, W. 2024a. Vtimellm: Empower llm to grasp video moments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 14271–14280.

Huang, R.; Li, M.; Yang, D.; Shi, J.; Chang, X.; Ye, Z.; Wu, Y.; Hong, Z.; Huang, J.; Liu, J.; et al. 2024b. Audiogpt: Understanding and generating speech, music, sound, and talking head. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 23802–23804.

Javier, R. J.; and Kim, Y. 2014. Application of linear predictive coding for human activity classification based on micro-Doppler signatures. *IEEE Geoscience and Remote Sensing Letters*, 11(10): 1831–1834.

Jiang, B.; Chen, X.; Liu, W.; Yu, J.; Yu, G.; and Chen, T. 2023. Motiongpt: Human motion as a foreign language. *Advances in Neural Information Processing Systems (NeurIPS)*, 36: 20067–20079.

Jing, L.; Xue, Y.; Yan, X.; Zheng, C.; Wang, D.; Zhang, R.; Wang, Z.; Fang, H.; Zhao, B.; and Li, Z. 2024. X4d-sceneformer: Enhanced scene understanding on 4d point cloud videos through cross-modal knowledge transfer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 2670–2678.

Kang, L.; Li, Z.; Zhao, X.; Zhao, Z.; and Braun, T. 2023. ST-PCT: Spatial–temporal point cloud transformer for sensing activity based on mmWave. *IEEE Internet of Things Journal*, 11(6): 10979–10991.

Kong, H.; Huang, C.; Yu, J.; and Shen, X. 2024. A survey of mmwave radar-based sensing in autonomous vehicles, smart homes and industry. *IEEE Communications Surveys & Tutorials*.

Lai, Z.; Yang, J.; Xia, S.; Wu, Q.; Sun, Z.; Yu, W.; and Pei, L. 2024. SMART: Scene-motion-aware human action recognition framework for mental disorder group. *IEEE Internet of Things Journal*.

Li, K.; Wang, Y.; He, Y.; Li, Y.; Wang, Y.; Liu, Y.; Wang, Z.; Xu, J.; Chen, G.; Luo, P.; et al. 2024. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 22195–22206.

Lin, B.; Ye, Y.; Zhu, B.; Cui, J.; Ning, M.; Jin, P.; and Yuan, L. 2023. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*.

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, 74–81.

Liu, H.; Wang, Y.; Zhou, A.; He, H.; Wang, W.; Wang, K.; Pan, P.; Lu, Y.; Liu, L.; and Ma, H. 2020. Real-time arm gesture recognition in smart home scenarios via millimeter wave sensing. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*, 4(4): 1–28.

Lu, H.; Chen, J.; Liang, F.; Tan, M.; Zeng, R.; and Hu, X. 2025. Understanding emotional body expressions via large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 1447–1455.

Maaz, M.; Rasheed, H.; Khan, S.; and Khan, F. S. 2023. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*.

Mahmood, N.; Ghorbani, N.; Troje, N. F.; Pons-Moll, G.; and Black, M. J. 2019. AMASS: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 5442–5451.

Meng, Z.; Fu, S.; Yan, J.; Liang, H.; Zhou, A.; Zhu, S.; Ma, H.; Liu, J.; and Yang, N. 2020. Gait recognition for co-existing multiple people using millimeter wave sensing. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 849–856.

Pal, M. 2005. Random forest classifier for remote sensing classification. *International Journal of Remote Rensing*, 26(1): 217–222.

Saadat, M. S.; and Sur, S. 2024. Enabling Coexistence of Indoor Millimeter-Wave Networking and Human Activity Sensing. In *2024 IEEE/ACM Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)*, 37–48. IEEE.

Shan, Z.; He, Y.; Zhao, C.; Du, J.; Zhang, J.; Zhang, Q.; Yu, J.; and Xu, L. 2025. Mojito: LLM-Aided Motion Instructor with Jitter-Reduced Inertial Tokens. *arXiv preprint arXiv:2502.16175*.

Shen, Z.; Sheng, X.; Fan, H.; Wang, L.; Guo, Y.; Liu, Q.; Wen, H.; and Zhou, X. 2023a. Masked spatio-temporal structure prediction for self-supervised learning on point cloud videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 16580–16589.

Shen, Z.; Sheng, X.; Wang, L.; Guo, Y.; Liu, Q.; and Zhou, X. 2023b. Pointcmp: Contrastive mask prediction for self-supervised learning on point cloud videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1212–1222.

Shrestha, A.; Li, H.; Le Kernec, J.; and Fioranelli, F. 2020. Continuous human activity classification from FMCW radar with Bi-LSTM networks. *IEEE Sensors Journal*, 20(22): 13607–13619.

Smith, K. A.; Csech, C.; Murdoch, D.; and Shaker, G. 2018. Gesture recognition using mm-wave sensor for human-car interface. *IEEE Sensors Letters*, 2(2): 1–4.

Song, X.; Salcianu, A.; Song, Y.; Dopson, D.; and Zhou, D. 2020. Fast wordpiece tokenization. *arXiv preprint arXiv:2012.15524*.

Song, Y.; Li, J.; Bian, Y.; and Cai, Z. 2025. Predicting User Behavior in Smart Spaces with LLM-Enhanced Logs and Personalized Prompts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 764–772.

Suykens, J. A. 2001. Support vector machines: a nonlinear modelling and control perspective. *European Journal of Control*, 7(2-3): 311–327.

Vedantam, R.; Lawrence Zitnick, C.; and Parikh, D. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4566–4575.

Wang, F.; Lv, Y.; Zhu, M.; Ding, H.; and Han, J. 2024. Xrf55: A radio frequency dataset for human indoor action analysis. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)*, 8(1): 1–34.

Wang, Y.; Guo, J.; Bai, J.; Yu, R.; He, T.; Tan, X.; Sun, X.; and Bian, J. 2025. Instructavatar: Text-guided emotion and motion control for avatar generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 8132–8140.

Wang, Y.; Liu, H.; Cui, K.; Zhou, A.; Li, W.; and Ma, H. 2021. m-activity: Accurate and real-time human activity recognition via millimeter wave radar. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 8298–8302. IEEE.

Xia, S.; Chu, L.; Pei, L.; Yang, J.; Yu, W.; and Qiu, R. C. 2024. Timestamp-supervised wearable-based activity segmentation and recognition with contrastive learning and order-preserving optimal transport. *IEEE Transactions on Mobile Computing (TMC)*.

Xia, S.; Zhang, Y.; Su, Z.; Zheng, X.; Lv, Z.; Wang, G.; Zhang, Y.; Wu, Q.; Chu, L.; and Pei, L. 2025. EnvPoser: Environment-aware Realistic Human Motion Estimation from Sparse Observations with Uncertainty Modeling. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Xu, K. 2025. AI-Driven Personalized Fall Prevention for Older Adults. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 29610–29612.

Xu, R.; Wang, X.; Wang, T.; Chen, Y.; Pang, J.; and Lin, D. 2024. Pointllm: Empowering large language models to understand point clouds. In *European Conference on Computer Vision (ECCV)*, 131–147. Springer.

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

Yang, J.; Huang, H.; Zhou, Y.; Chen, X.; Xu, Y.; Yuan, S.; Zou, H.; Lu, C. X.; and Xie, L. 2023. Mm-fi: Multi-modal non-intrusive 4d human dataset for versatile wireless sensing. *Advances in Neural Information Processing Systems (NeurIPS)*, 36: 18756–18768.

Yang, J.; Xia, S.; Lai, Z.; Sun, L.; Wu, Q.; Yu, W.; and Pei, L. 2025a. mmDEAR: mmWave Point Cloud Density Enhancement for Accurate Human Body Reconstruction. *IEEE International Conference on Robotics and Automation (ICRA2025)*.

Yang, J.; Xia, S.; Song, Y.; Wu, Q.; and Pei, L. 2024. mmBaT: A Multi-Task Framework for Mmwave-Based Human Body Reconstruction and Translation Prediction. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 8446–8450. IEEE.

Yang, S.; Liu, J.; Zhang, R.; Pan, M.; Guo, Z.; Li, X.; Chen, Z.; Gao, P.; Li, H.; Guo, Y.; et al. 2025b. Lidar-llm: Exploring the potential of large language models for 3d lidar understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 9247–9255.

Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Zhang, Y.; Xia, S.; Chu, L.; Yang, J.; Wu, Q.; and Pei, L. 2024. Dynamic inertial poser (dynaip): Part-based motion dynamics learning for enhanced human pose estimation with sparse inertial sensors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1889–1899.

Zhou, Z.; Wan, Y.; and Wang, B. 2024. Avatargpt: All-in-one framework for motion understanding planning generation and beyond. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1357–1366.