mmExpert: Integrating Large Language Models for Comprehensive mmWave Data Synthesis and Understanding

Yifan Yan[1], Shuai Yang[1], Xiuzhen Guo[1] *, Xiangguang Wang[1], Wei Chow[1], Yuanchao Shu[1], Shibo He[1]
[1]Zhejiang University
Hangzhou, Zhejiang, China

Abstract

Millimeter-wave (mmWave) sensing technology holds significant value in human-centric applications, yet the high costs associated with data acquisition and annotation limit its widespread adoption in our daily lives. Concurrently, the rapid evolution of large language models (LLMs) has opened up opportunities for addressing complex human needs. This paper presents mmExpert, an innovative mmWave understanding framework consisting of a data generation flywheel that leverages LLMs to automate the generation of synthetic mmWave radar datasets for specific application scenarios, thereby training models capable of zero-shot generalization in real-world environments. Extensive experiments demonstrate that the data synthesized by mmExpert significantly enhances the performance of downstream models and facilitates the successful deployment of large language models for mmWave understanding.

CCS Concepts

• **Human-centered computing** → **Ubiquitous and mobile computing**; **Ubiquitous and mobile computing systems and tools**.

Keywords

mmWave Sensing, Data Synthesis, Large Language Models (LLMs)

ACM Reference Format:
Yifan Yan[1], Shuai Yang[1], Xiuzhen Guo[1] *, Xiangguang Wang[1], Wei Chow[1],, Yuanchao Shu[1], Shibo He[1]. 2025. mmExpert: Integrating Large Language Models for Comprehensive mmWave Data Synthesis and Understanding. In *International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing (MobiHoc '25), October 27–30, 2025, Houston, TX, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3704413.3764420

1 Introduction

In recent years, there have been significant advancements in the field of wireless sensing, especially millimeter-wave (mmWave) [14, 24, 29, 31, 32]. The intrinsic non-invasive and privacy-preserving features make mmWave radars well-suited for a broad spectrum of human-centric applications, including healthcare, security, and smart homes [12, 19, 20, 32]. As the sensed mmWave signal contains information on human activities, state-of-the-art methods resort to deep neural models for inference. However, due to limited training datasets, the performance of these methods degrades drastically when encountering unseen test cases (e.g., users and environments).

To address this issue, recent studies have proposed synthetic data generation methods to augment the training data for mmWave radar-based human sensing tasks [2, 5, 28, 33, 34]. These methods leverage the channel model, simulator, or deep learning model to generate training data without requiring data collection experiments involving humans in the loop. As the rapid advancements in large language models (LLMs) show remarkable potential for understanding multi-modal data [27], the latest works have been striving to generate mmWave radar data directly from textual descriptions or prompts [5, 6, 34]. The LLM-based methods show impressive performance in mitigating data scarcity by leveraging the inherent web-scale world knowledge.

Although the LLM has shown its great potential to mitigate data scarcity, the power of the LLM has not been fully unleashed in mmWave sensing. Compared with traditional mmWave sensing tasks like classification or reconstruction, the mmWave understanding task fundamentally differs by requiring open-vocabulary reasoning (Table 1) instead of providing discriminative results. And this reasoning capability is uniquely achievable by LLMs.

In this paper, we present **mmExpert**, a framework that enables text-to-mmWave cross-modal generation and mmWave-to-text sensing data understanding. As shown in Figure 1, mmExpert consists of two modules, i.e., "text-to-mmWave" and "mmWave-to-text". The "text-to-mmWave" module transforms textual descriptions into synthetic mmWave signals by leveraging LLM to automate the generation of synthetic mmWave radar datasets for required application scenarios. Conversely, the "mmWave-to-text" module converts real-world mmWave signals into textual descriptions by utilizing open-vocabulary training and transferring pre-trained models from other domains. mmExpert significantly reduces the reliance on extensive real-world data collection while equipping LLMs with the ability to interpret and understand mmWave sensing data, thereby enabling more robust and scalable mmWave sensing applications.

To achieve this, several non-trivial challenges need to be addressed.

(i) Decoupling real-world data dependency in the training stage. Synthetic data often fails to capture the full characteristics of real-world scenarios, leading to potential discrepancies in model performance when deployed in practical applications. Furthermore,

*Xiuzhen Guo is the corresponding author.

MobiHoc '25, Houston, TX, USA

ACM ISBN 979-8-4007-1353-8/25/10
https://doi.org/10.1145/3704413.3764420

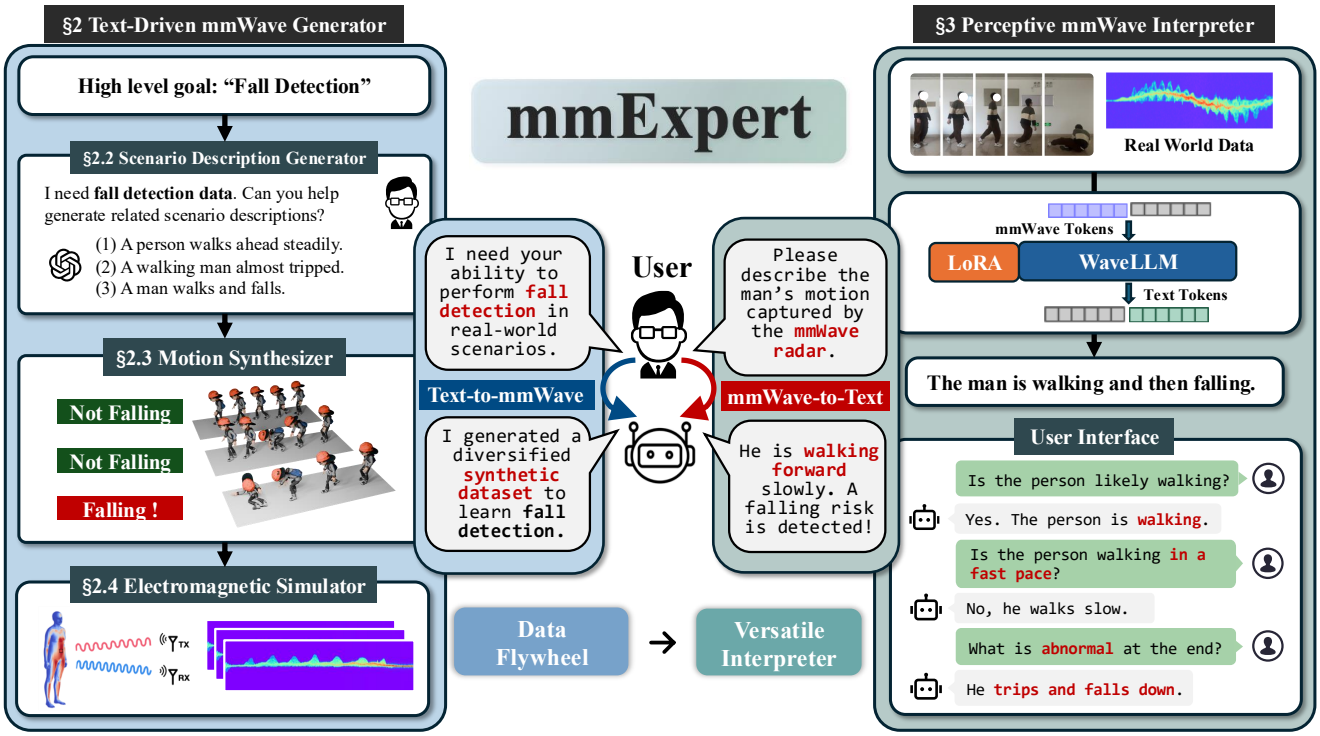


Figure 1: Overview of the mmExpert. mmExpert comprises a data flywheel that generates mmWave-language data hierarchically and a versatile interpreter capable of understanding mmWave data across diverse contexts.

under these constraints, it is not feasible to train Neural Style Transfer (NST) models using any real-world data, and real-world data remains unavailable for the fine-tuning phase. To overcome these limitations, mmExpert employs electromagnetic simulation to model the spatial propagation of mmWave signals, thereby accurately characterizing the physical properties of mmWave. Additionally, domain randomization techniques are utilized to design various random factors within the simulation system, which enhances the sim-to-real transfer without requiring real-world data.

(ii) Data curation pipeline for mmWave-text alignment. Accurate and comprehensive textual annotation for mmWave data is an exceedingly expensive and labour-intensive process. Previous works have struggled to achieve high-quality mmWave-text alignment as their textual descriptions lack diversity and were restricted to a specific format [4]. This limitation hinders the models' ability to integrate mmWave signals with free-form text. We address this challenge in mmExpert by utilizing hierarchical action description prompts that encompass rich textual content to generate detailed action descriptions. The prompts allow mmExpert to accommodate a wide range of actions without being restricted to a predefined set of categories, enhancing the flexibility and scalability of the dataset. Furthermore, mmExpert employs a language-aligned pre-training phase, leveraging contrastive learning to establish a robust correspondence between the mmWave data and text descriptions.

(iii) Comprehensive benchmark of the mmWave interpreter. Evaluating the mmWave Interpreter is particularly challenging due to its open-vocabulary text outputs, which traditional metrics cannot adequately assess. Moreover, the absence of a comprehensive test set further complicates the evaluation process. To address these challenges, we conduct extensive experiments using a diverse array of metrics, including traditional language metrics, machine learning-based metrics, and recent evaluations employing LLM. To better gauge the comprehensive understanding abilities of the interpreter, we also carefully curate a benchmark test set that specifically includes question-answering and captioning, which are pivotal for assessing core language skills. Specifically, this question-answering component utilizes natural language to pose questions that probe for fine-grained details about the observed actions and inquire into the possible intentions behind them.

Our contributions are summarized as follows:

- We propose a scalable and cost-effective framework for generating diverse mmWave signals using world knowledge from LLMs, eliminating the need for real-world data collection, which is often expensive and labor-intensive.
- With the proposed training diagram and encoder design, we not only pioneer the exploration of scaling laws in mmWave signal modeling but also significantly improve performance, surpassing the strongest baseline by 19% in classification tasks and reducing FID by 22% in signal generation.
- We design and train WaveLLM, a multi-modal large language model, exclusively on synthesized data. WaveLLM demonstrates strong zero-shot performance on real-world data, showcasing understanding and reasoning capabilities.

2 Text-Driven mmWave Generator: Text-to-mmWave

2.1 Overview

In this section, we introduce the design of the text-driven mmWave generator, the text-to-mmWave part of mmExpert. It includes four modules. (1) **Scenario Description Generator** (§2.2) produces diverse and contextually accurate motion descriptions with user-defined textual prompts. (2) **Motion Synthesizer** (§2.3) transform these motion descriptions into 3D human motion sequences by utilizing pre-trained motion generation models. (3) **RF-Signal Electromagnetic Simulation** (§2.4) applies physical electromagnetic principles to simulate radar signal interactions with the generated human models, yielding realistic radar data. (4) **Sim-to-Real Domain Randomization** (§2.5) further introduces variability in parameters such as radar angles, body segmentation, antenna pattern, background noise, and nonlinearity scaling to enhance the generalization and robustness of the system in real-world scenarios.

2.2 Scenario Description Generator

mmExpert interprets the user's scenario description and generates a range of diverse motion prompts. mmExpert utilizes world knowledge and the strong reasoning ability of the LLM. This enables mmExpert to recognize not only explicit actions but also to infer implicit behaviors and possible outcomes based on typical situational dynamics. With prompt engineering and its own rich knowledge, the LLM generates accurate descriptions for precise text-to-motion translation and disassembles complex actions into manageable units.

To generate a high-quality text-mmWave paired synthetic dataset, two key considerations must be addressed: accurately defining the characters' actions and ensuring diversity in the descriptions. mmExpert considers the following aspects:

Syntactic and synonym substitution. mmExpert generates motion description prompts using diverse syntactic structures, such as different tenses and inversion patterns, to emphasize aspects of the action while also drawing from a wide range of synonyms to enhance description variety.

Temporal relation of actions. The temporal relationship between actions is crucial for coherent motion descriptions. mmExpert considers the sequence and repetition of actions, ensuring that each action is presented in an appropriate temporal context.

2.3 Motion Synthesizer

We employ an off-the-shelf, text-driven human motion generation model [10] to perform as the motion synthesizer. This model is structured as a sequence-to-sequence architecture, which accepts descriptive text prompts as input and generates corresponding motion sequences. These motion sequences are typically represented as three-dimensional human skeletal keypoints. To transform these keypoint sequences into detailed and realistic human surface representations, we fit them to the SMPL human surface model [3]. This fitting process converts the skeletal data into a comprehensive 3D mesh that accurately depicts the human body's surface.

However, motion data generated in the temporal domain can exhibit potential jitter and frame-to-frame inconsistencies, which may compromise the overall quality and realism of the motion. To mitigate these issues, we apply Gaussian filtering to the generated human surface models along the time dimension. This filtering process smooths out abrupt changes and reduces noise, resulting in more fluid and coherent motion sequences that better capture natural human movements.

2.4 RF Signal Electromagnetic Simulation

For the most commonly used Frequency Modulated Continuous Wave (FMCW) radar, the IF signal is essentially an estimation of the distance and reflection intensity of all reflection points, which can be expressed as

$$x_{\mathrm{IF}}(t) = \sum_{i=1}^{N} a_i e^{j2\pi\tau_i(f_c+St)}, \tag{1}$$

where N is the number of reflection points, f_c is the carrier frequency of the radar signal, S is the frequency sweep rate of the radar signal, a_i is the amplitude of the i-th reflection point, and τ_i is the time delay of the i-th reflection point.

We observe that the amplitude of the IF signal plays a crucial role in ensuring the authenticity of the synthesized mmWave radar data. The amplitude a_i of a single idealized path can be calculated:

$$a_i = \frac{\lambda}{4\pi} \cdot \underbrace{\frac{1}{R^2}}_{\text{Path loss}} \cdot \underbrace{\sqrt{\mathbf{C}_{\mathrm{tx}}\mathbf{C}_{\mathrm{rx}}}}_{\text{Antenna loss}} \cdot \underbrace{\Gamma\sqrt{dA'\mathbf{f}_{\mathrm{s}}}}_{\text{Scattering loss}}, \tag{2}$$

where $\mathbf{C}_{\mathrm{tx}}$ and $\mathbf{C}_{\mathrm{rx}}$ are the gains of the transmitting and receiving antennas after accounting for radar pattern, λ is the wavelength of the millimeter wave signal, R is the distance between the radar and the target, dA' is the projection area of the scattering surface, Γ is the scattering coefficient, and $\mathbf{f}_{\mathrm{s}}$ is the scattering pattern function.

The above Equation 2 inspires us to incorporate the variations in reflection intensity caused by path loss, antenna loss, and scattering loss into the signal simulation:

(1) Antenna loss represents inefficiencies in signal transmission and reception due to imperfections in the antenna design, leading to signal degradation.

(2) Path loss refers to the reduction in signal strength as electromagnetic waves propagate through space, which varies with the distance between the transmitter and the reflector.

(3) Scattering loss arises from the interaction of electromagnetic waves with the detection target, causing the signal to scatter in multiple directions, further reducing the reflected signal strength.

By incorporating these factors into the signal simulation, we can more accurately model the real-world behavior of radar signals and improve the overall realism and reliability of the simulation outcomes.

We adopt the mesh-tracking method to simulate the radar signal instead of sending uniform rays through space. Our approach only targets the visible facets of the human mesh model from the radar's perspective.

To accurately model the micro-Doppler effect, we interpolate the mesh's position between keyframes to derive precise surface velocities. This technique allows us to perform computationally intensive ray-tracing at a lower frame rate (e.g., 20 fps) while still generating high-resolution, millisecond-level micro-Doppler data.

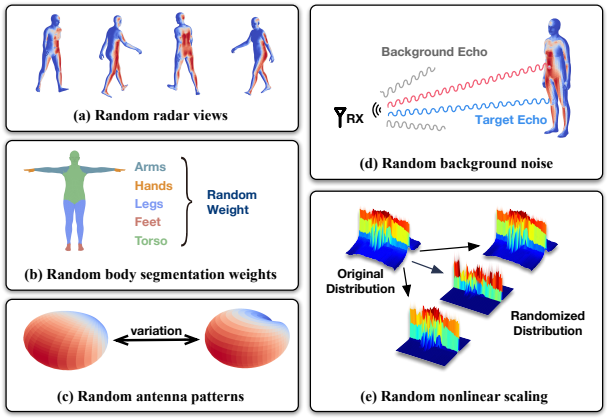


Figure 2: Illustration of domain randomization factors.

2.5 Sim-to-Real Domain Randomization

To overcome the sim-to-real challenge, domain adaptation methods are essential to enable a model to generalize from synthetic data distributions to real-world data distributions. Previous works [2, 5, 7, 33, 34] utilize the neutral style transfer method and the fine-tuning method to improve the sim-to-real transfer performance. However, both methods rely on real-world data, which is costly and time-consuming to collect. In contrast, mmExpert operates under a zero-shot generalization condition, which means that our system does not depend on real-world data during the training stage.

The domain randomization technique [25] introduces sufficient random variation into the simulated training data, allowing the model to learn robust features that are invariant to distribution shifts between the synthetic data and the real-world data. In our implementation, we apply the following domain randomization variations as shown in Figure 2.

(a) Random radar views. We randomly alter the radar view angle to capture radar signals from different perspectives. This variation enables the downstream model to learn invariant features of human motion, regardless of the observation angle.

(b) Random body segmentation weights. The human mesh is divided into segments corresponding to body parts based on anatomical structure. Random weights are assigned to each part to simulate varying reflection intensities, emulating the diverse material properties of the human body.

(c) Random antenna patterns. Due to the complexity of accurately modeling radar antenna characteristics, we randomly adjust the beam width of the simulated radar in both azimuth and elevation within a predefined range. This is achieved by assigning parameters to the antenna pattern function as described in §2.4.

(d) Random background noise. Gaussian white noise is added to the synthesized mmWave radar data to simulate thermal noise inherent in radar systems. Additionally, to emulate static background noise encountered in real-world scenarios, we generate background echoes with random energy intensities during simulation.

(e) Random nonlinearity scaling. In typical data preprocessing, radar signals undergo dB conversion to compress the dynamic range. However, the logarithm operation is extremely sensitive to the input data distribution, which introduces nonlinearity. We apply stochastic nonlinearity scaling to enhance the diversity of the distribution of the synthesized training data.

3 Perceptive mmWave Interpreter: mmWave-to-Text

3.1 Overview

In this section, we introduce the Perceptive mmWave Interpreter, the mmWave-to-Text component of mmExpert. The interpreter is designed to convert mmWave radar signals into meaningful textual descriptions, thereby enhancing downstream perception tasks. The architecture consists of four main modules. (1) **mmWave Signal Feature Extractor** (§3.2) utilizes a transformer-based encoder to capture characteristics of mmWave signals, providing a comprehensive set of features essential for semantic interpretation. (2) **Sentence Feature Extractor** (§3.3) employs a BERT-based model to generate rich semantic embeddings from textual inputs, ensuring effective representation of natural language descriptions. (3) **Language-Aligned Pre-training** (§3.4) adopts a contrastive learning approach inspired by CLIP to align mmWave signal features with corresponding text embeddings, creating a cohesive semantic framework that bridges the gap between radar data and language. (4) **mmWave Interpreter: WaveLLM** (§3.5) integrates the aligned features into a large language model (LLM) framework, enabling the generation of contextually accurate and semantically rich textual descriptions based on mmWave signal inputs.

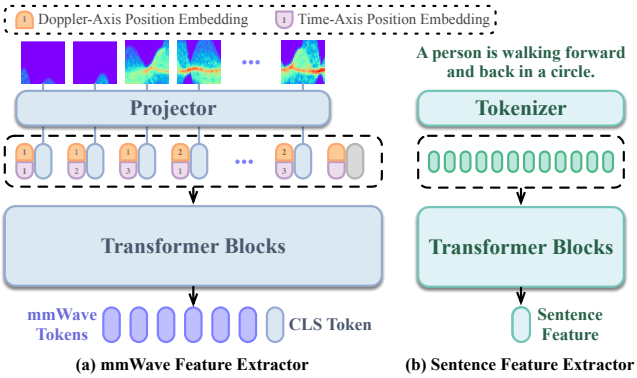
3.2 mmWave Signal Feature Extractor

We leverage a transformer-based encoder to capture temporal information from the micro-Doppler spectrum, which provides a compact and informative representation of human motion. This module processes the input spectrogram data as a 2D matrix $\boldsymbol{I} \in \mathbb{R}^{H \times W}$, where H represents the temporal axis and W corresponds to the Doppler frequency axis, representing the target's radial velocity. The encoder architecture is shown in Figure 3 (a).

To extract meaningful features, the input spectrogram is partitioned into a sequence of flattened 2D patches $x \in \mathbb{R}^{N \times (P^2)}$, with P denoting the patch size and N the number of patches. These patches are then projected into a latent feature space of dimension D. To effectively preserve the contextual information along both the temporal and Doppler axes, we incorporate learnable positional embeddings for each dimension separately.

Next, we utilize a multi-layer transformer encoder to extract effective global features from the input. By incorporating a learnable class token, the encoder aggregates comprehensive information, enabling the module to summarize the overall spectrogram information.

The output of this module includes both a global feature embedding derived from the class token and a sequence of token embeddings $z \in \mathbb{R}^{(N+1) \times D}$. The global feature is aligned with text embeddings during the contrastive learning process (§3.4), while the sequence embeddings retain temporal motion details for subsequent semantic modeling tasks. Together, these features comprehensively represent the temporal characteristics essential for mmWave semantic interpretation.



(a) mmWave Feature Extractor (b) Sentence Feature Extractor

Figure 3: Architecture of the encoders. The mmWave encoder extracts motion features from the signal, and the sentence encoder extracts semantic features from the text.

3.3 Sentence Feature Extractor

The Sentence Feature Extractor is a BERT model [23] meticulously designed to generate rich semantic representations from textual inputs, specifically at the sentence level, as shown in Figure 3 (b). Built upon a transformer-based architecture, it effectively captures intricate contextual dependencies within text through its core, which consists of multiple transformer layers utilizing self-attention mechanisms.

The text extractor begins by embedding input tokens into a high-dimensional feature space, capturing essential syntactic and semantic information. Positional encodings are added to preserve the sequential structure of the text. The processed tokens are then passed through a transformer encoder, similar to the previously described mmWave encoder (§3.2), which aggregates global sentence-level features.

At the final stage, a pooling mechanism is applied to generate a compact sentence-level embedding. Specifically, the model aggregates information from all token representations, effectively summarizing the overall semantic content of the input. This representation serves as the global feature for subsequent alignment with the mmWave-based motion features during the contrastive learning phase (§3.4).

To enhance the capability, the Sentence Feature Extractor is pre-trained on an extensive corpus of diverse natural language tasks. Such robust pre-training enables it to capture the full semantic essence of sentences accurately. Consequently, it can extract rich, context-aware representations that generalize proficiently across various language-understanding tasks.

3.4 Language-Aligned Pre-training

We utilize a contrastive learning approach inspired by the CLIP [22] method to align mmWave signal features with text features. This alignment is essential for integrating mmWave motion features with semantic textual content, ultimately enriching the understanding of scene and event descriptions.

The contrastive learning framework pairs mmWave signal features, extracted via the mmWave Signal Feature Extractor, with sentence embeddings generated by the Sentence Feature Extractor. During training, the model learns to maximize the similarity of corresponding pairs (i.e., the mmWave feature corresponding to a specific textual description) while minimizing the similarity between non-corresponding pairs. This dual-objective approach encourages the model to discern fine-grained relationships between motion and text, bridging the gap between these modalities.

Specifically, we use cosine similarity to evaluate the distance between multi-modal features in the latent space, enabling seamless comparisons between mmWave and textual representations. The model is then trained using the InfoNCE loss [21], defined as:

$$\mathcal{L}_{\text{CLIP}} = -\frac{1}{N}\sum_{i=1}^{N}\log\left(\frac{\exp(\boldsymbol{v}_i \cdot \boldsymbol{t}_i/\tau)}{\sum_{j=1}^{N}\exp(\boldsymbol{v}_i \cdot \boldsymbol{t}_j/\tau) + \sum_{j=1}^{N}\exp(\boldsymbol{t}_i \cdot \boldsymbol{v}_j/\tau)}\right), \quad (3)$$

where $\boldsymbol{v}_i$ and $\boldsymbol{t}_i$ represent the mmWave feature and its corresponding text feature embedding for the i-th sample, and τ is a temperature parameter that scales the logit outputs. This loss function minimizes the distance between positive pairs and maximizes the distance between negative pairs, fostering a robust semantic alignment. By aligning the mmWave features with text features, we enable the model to map the disparate data types into a coherent semantic space.

3.5 mmWave Interpreter: WaveLLM

To further unlock the power of mmWave semantic understanding, we present WaveLLM, a generative model designed to understand complex multi-modal contexts from mmWave signals and language. The model comprises three main components: a pre-trained mmWave feature extractor f_E, detailed in § 3.2, a multi-modal projection layer f_P, and a pre-trained large language model (LLM) backbone f_{LLM}.

To enable WaveLLM to handle both mmWave signal features and language features effectively, we enhance the pre-trained LLM's capabilities. This is accomplished by strategically combining the two types of data. Specifically, we take the features extracted from mmWave signals and align them with corresponding language features. Before these features are combined, a shallow MLP (multi-layer perceptron) based projection layer is used, denoted as f_P. This layer plays a crucial role in converting mmWave features into a form that can be easily integrated with language features. Once the conversion is complete, the model concatenates these aligned mmWave and language features. This combined input is then fed into the LLM, which processes it to understand the multi-modal content. The role of the LLM is to sequentially predict the next token in the output sequence in an auto-regressive manner. As the user inputs a sequence of mmWave signals along with a corresponding question, the WaveLLM responds based on the underlying human motion depicted in the signals.

To preserve the generality of the pre-trained LLM while enabling it to comprehend and reason with mmWave signals, we apply Low-Rank Adaptation (LoRA) [13].

LoRA modifies the forward pass of linear layers in the Transformer without fully fine-tuning them. Instead, it introduces trainable low-rank matrices, altering the forward pass as follows:

$$h = W_0 x + \Delta W x = W_0 x + BAx, \quad (4)$$

where $h \in \mathbb{R}^d$ and $x \in \mathbb{R}^k$ are the output and input of the layer, respectively; here, W_0 denotes the original weight matrix, which

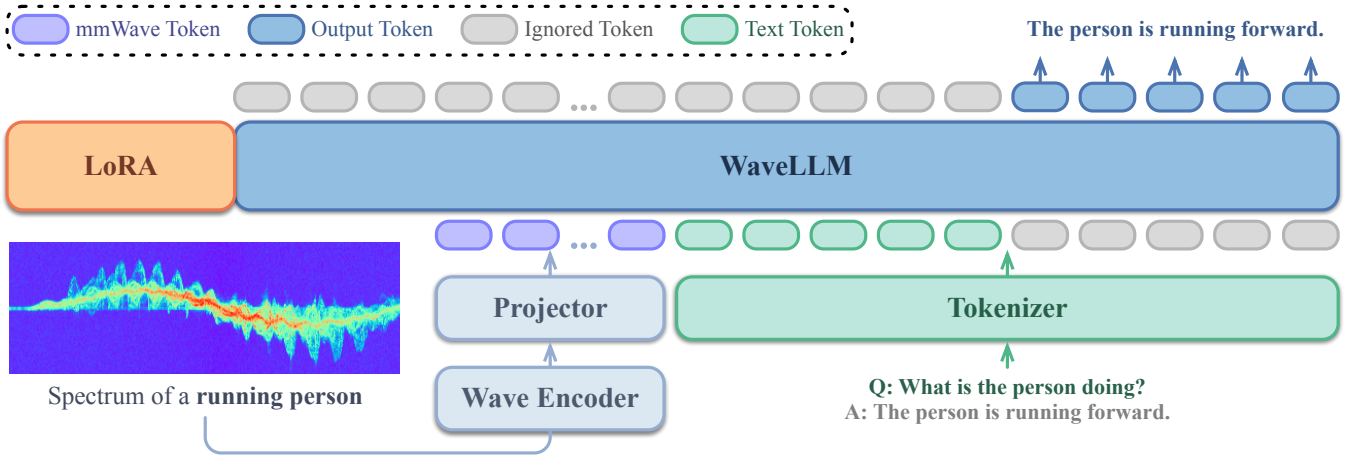


Figure 4: Architecture of WaveLLM. The wave encoder extracts features from the signal, and the projector maps these features to the LLM's embedding space. Leveraging multi-modal context, the LLM generates the answer.

remains frozen during training. The trainable weights are $A \in \mathbb{R}^{r\times k}$ and $B \in \mathbb{R}^{d\times r}$, where $r \ll \min(d, k)$.

By utilizing the efficient design of WaveLLM, we effectively train our multi-modal LLM to offer robust understanding capabilities, adeptly interpreting and reasoning across the mmWave and text data spaces.

4 Implementation Details

4.1 Radar Configuration

The mmWave radar parameters in simulation and real-world data capture systems are synchronized. For the real-world hardware system, we utilize an IWR1843BOOST radar sensor equipped with three transmit antennas, four receive antennas, and a DCA1000EVM evaluation board by Texas Instruments for real-time data capture and streaming. The radar operates at a frequency of 77 GHz and bandwidth of 4 GHz, providing a range resolution of about 4.3 cm. The frame rate of the radar is set to 50fps, and each radar frame consists of 128 chirps. The radar is placed at a height of 1 meter, and the distance between the radar and the human subject ranges from 1 to 5 meters. To fully capture the whole motion of the human body, we put the radar sideways, which reduces the vertical energy attenuation, as illustrated in Figure 5.

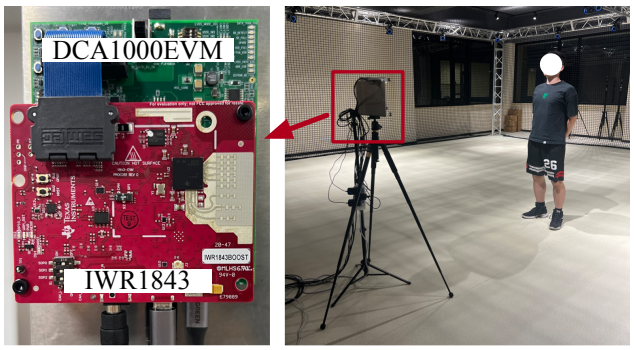
4.2 Data Curation

The training data consists of the HumanML3D [11] dataset and the synthesized mmWave signals generated by the Scenario Description Generator. The HumanML3D dataset is a publicly available 3D human motion dataset that provides diverse motion capture data across various human activities. Meanwhile, HumanML3D provides high-quality text annotations for each motion sequence, which contain detailed descriptions of the motion characteristics. However, it is important to note that the categories included in HumanML3D are limited and do not encompass all complex actions encountered in daily life. Therefore, there remains a need to employ text-to-mmWave synthesis systems to augment this data effectively.

We utilize the Scenario Description Generator introduced in §2.2 to generate description prompts for extended training data generation. These generated prompts are used to guide the motion generation and the mmWave signal synthesis process. Notably, we do not explicitly provide motion category labels during the prompt generation stage for scenario guidance. Instead, the downstream model is required to learn the correlations between open-vocabulary prompts and synthesized mmWave signals; subsequently, it selects the most relevant labels based on incoming mmWave signals during the inference stage.

The real-world evaluation dataset consists of over 1000 real-world mmWave micro-Doppler spectra corresponding to human motion with high-quality human-verified text annotations, including class labels, detailed captions, and question-answerings. The benchmark is able to evaluate the encoder's open vocabulary classification ability and the LLM's multi-modal understanding ability.

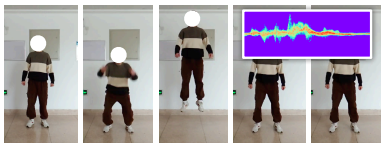
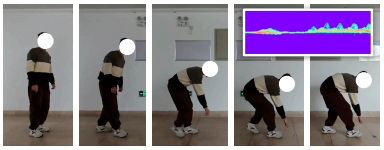
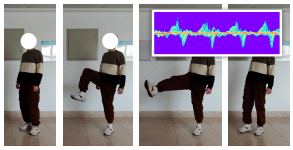
Each spectrum sample is a time-Doppler signal with a duration ranging from 6 to 12 seconds, captured at 50 fps. Our dataset is collected across two scenarios and includes 10 volunteers. It comprises 15 action categories in total: 9 basic actions (e.g., walking, running, bending over) and 6 complex actions, which are combinations of



(a) Radar Placement **(b) Capture Scene**

Figure 5: Data capture system. The collected data is utilized exclusively during the testing phase and is not employed at any stage of training.

Table 1: Real-world deployment demo of WaveLLM. All responses from WaveLLM are derived directly from the multi-modal LLM's inference on mmWave signal.

Scenario			 ×4
Q&A	**Predicted Caption:** The person is jumping up and down in place. **User:** Does the person stand still? **WaveLLM:** No, the person is actively moving and jumping. **User:** Is the person jumping in rapid succession or jumping hard and high? **WaveLLM:** The person is jumping hard and high. **User:** Does the person stop jumping at the end? **WaveLLM:** Yes, the person stops jumping at the end.	**Predicted Caption:** The person bends over, then walks forward. **User:** How many actions are involved according to the wave signal? **WaveLLM:** Two actions, bending down and walking. **User:** What might the person want to do? **WaveLLM:** The person might want to pick something up from the ground. **User:** Does the man walk fast? **WaveLLM:** No, he walks slowly.	**Predicted Caption:** The person kicks in place multiple times. **User:** Is the person likely to do sports? **WaveLLM:** Yes, the person is likely engaging in a sport that involves kicking. **User:** What kind of sport might he be playing? **WaveLLM:** He could be playing soccer or martial arts. **User:** How many kicks does this person perform? **WaveLLM:** The person performs four kicks.

the basic movements. Each action sequence varies in speed and frequency and corresponds to a unique text description.

4.3 Model Training Detail

We use a pre-trained Vision Transformer [9] as the spectrum feature extractor, characterized by a patch size of 16. The model comprises 12 transformer layers, each with 12 attention heads and a hidden size of 768. For the sentence feature extractor, we employ Sentence-BERT [23], a pre-trained model specifically designed to generate semantically meaningful sentence embeddings. The model is based on the BERT architecture and consists of 6 hidden layers, each containing 12 attention heads and a hidden size of 384.

During the CLIP training phase, we set the learning rate to 1×10^{-4}, and the temperature parameter is configured to 0.1. The model is trained for 50 epochs to ensure convergence and optimal performance.

For the mmWave interpreter, we employ a compact yet robust language model backbone, Phi-3 [1], which consists of 3.8 billion parameters. In our primary analysis, we utilize LoRA [13] with a rank of 8, where the trainable weight percentage is 0.43%.

5 Evaluation Results

5.1 Baselines

Baselines for data synthesis. To illustrate the quality of the generated data, we choose the state-of-the-art publicly available text-based mmWave signal synthesis systems as the baseline systems for comparison.

- **RFGen [5]** utilizes a diffusion-based text-to-motion model to generate human mesh models. It proposes a physics-based ray tracing simulator for mmWave signal synthesis. RFGen further adopts another diffusion-based neural style transfer model to overcome the sim-to-real gap.
- **Text2Doppler [34]** utilizes a transformer-based generative model to generate motion from text prompts [10]. It approximates the human body segments as ellipsoids. We use the signal synthesis pipeline in Text2Doppler for comparison.

The classification models are trained using the same configuration, with the only difference being the training data generated by different synthesis systems. While RFGen offers an end-to-end generation pipeline, we directly input the motion description prompts and obtain the synthesized mmWave signals. For Text2Doppler, we use the same human motion data generated by our motion synthesizer module, since only the signal simulator is released.

Baselines for mmWave understanding. To the best of our knowledge, mmExpert is the first framework to integrate large language models (LLMs) for mmWave understanding. For comparison, we use WaveLLM against two types of baselines: non-semantic alignment training methods (Classification) and fundamental models (ResNet).

5.2 Overall Performance

Zero-shot classification performance. The results of the zero-shot generalization classification are presented in Figure 6. The classification model (ViT encoder + MLP decoder), trained using the synthesized mmWave signals from our proposed mmExpert, demonstrates outstanding performance across 15 human motion categories, achieving an impressive average accuracy of 84.6%. To measure the model's ability to identify positive instances while minimizing false negatives, we calculate the F1-score for each category.

Furthermore, we find that certain similar human actions exhibit slightly lower accuracies; nonetheless, mmExpert consistently outperforms baseline methods in these cases. This underscores the effectiveness of mmExpert in producing high-quality mmWave signals for real-world human activity recognition.

We observed that RFGen's ray-tracing strategy based on depth images may cause abrupt changes in pixel-level depth during fast motion, thereby leading to periodic extension in the Doppler dimension and disrupting the time-Doppler signal. Meanwhile, Text2Doppler uses a simplified human model without sim-to-real design. Building on these observations, mmExpert is designed to synthesize high temporal-resolution data by tracking the model's surface and employs a dedicated sim-to-real module to bridge the gap to real-world performance.

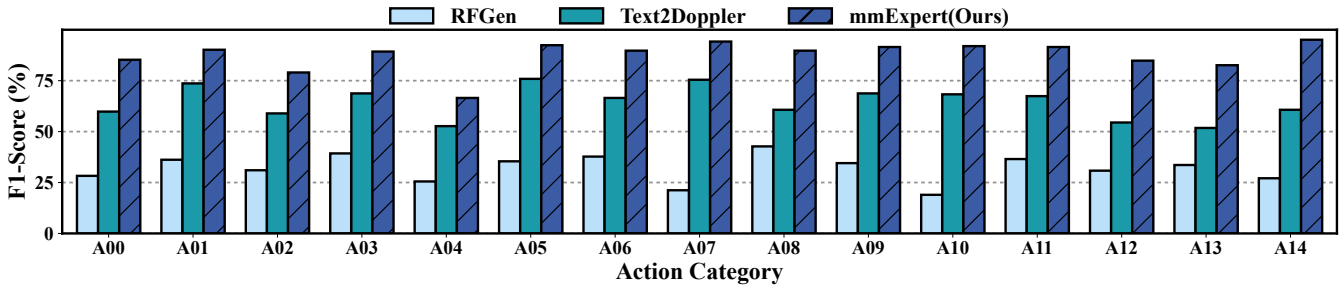


Figure 6: Comparison of F1-scores for each action category with baselines. A00 to A14 refer to different action categories. The results indicate that the RFGen shows poor sim-to-real generalization ability, and our method surpasses Text2Doppler with a consistent performance gain.

Table 2: WaveLLM is evaluated using GPT-based assessments and widely used language metrics. "CLS" represents classification as the training objective, while "CLIP" refers to contrastive learning. "ResNet" and "ViT" denote encoder architectures.

Model	GPT Evaluation			Text-based Evaluation				
	Correctness	Hallucination↓	Precision	SBERT	SimCSE	BLEU-1	ROUGE-L	METEOR
WaveLLM-CLS-ResNet	2073	889	70.0	69.0	67.5	41.5	45.3	21.0
WaveLLM-CLS-ViT	2049	913	69.2	69.4	67.7	42.3	45.7	21.3
WaveLLM-CLIP-ResNet	2105	857	71.1	69.8	68.6	41.9	46.0	21.3
WaveLLM-CLIP-ViT	**2196**	**766**	**74.2**	**72.2**	**71.2**	**46.9**	**49.9**	**23.5**

WaveLLM understanding performance. In this section, we primarily present the optimal configuration of WaveLLM. The test set assesses captioning and question-answering tasks, emphasizing the LLM's comprehension capabilities. We selected details from each description for point-based scoring. As shown in Table 2, ViT consistently outperforms ResNet, and the CLIP training paradigm surpasses classification-based training. WaveLLM with ViT pre-trained using CLIP demonstrates improved accuracy in answer recognition and reduced hallucination, as indicated by GPT evaluation results, while also achieving superior overall language performance across various metrics.

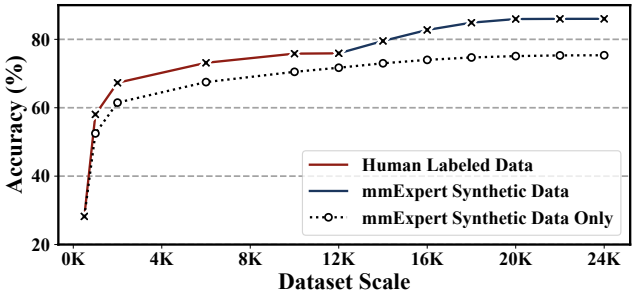


Figure 7: Ablation of the extension dataset scale. When human-labeled data reaches saturation (red), data from mmExpert (blue) further extends the boundary.

5.3 Ablation Studies

Data scaling law. The data scaling law [15, 16] plays a pivotal role in understanding and optimizing the relationship between model performance and the amount of training data. To investigate this, we trained a CLIP classification model using both the human-labeled dataset (i.e., HumanML3D) and the synthesized dataset generated by mmExpert from text descriptions.

As shown in Figure 7, the red curve represents the data scaling law derived from the human-labeled dataset, highlighting the critical trend that increased data generally leads to better performance. Notably, the blue curve showcases the scaling trend extended with our dataset, underscoring its ability to sustain consistent performance gains and further validate the significance of the scaling law in driving progress. This setup allows us to compare the impact of human-curated data versus synthetic data on model performance and to further demonstrate the robustness of the data scaling law.

Impact of scenario-guiding prompts on dataset quality. We evaluate the effect of scenario-guiding prompts on the quality of generated datasets through an ablation study, wherein we modify the prompts used to synthesize mmWave signals. Three configurations of the prompt generator are tested: template-based descriptions, free-form descriptions incorporating diverse syntactic structures and synonym substitutions, and complex descriptions that integrate combinations of temporal actions that highlight temporal relationships. The results of the different designs are shown in Figure 8. It can be observed that free-form descriptions, characterized by diverse syntactic structures and synonyms, significantly outperform template-based descriptions, highlighting the critical role of high-quality descriptions. Additionally, increasing the temporal complexity of the data further enhances performance.

Impact of domain randomization. We conduct ablation studies to evaluate the effects of various domain randomization techniques by individually adding each randomization component. As illustrated in Figure 9, the model achieves the best performance when all randomization factors are used. Notably, stochastic nonlinearity

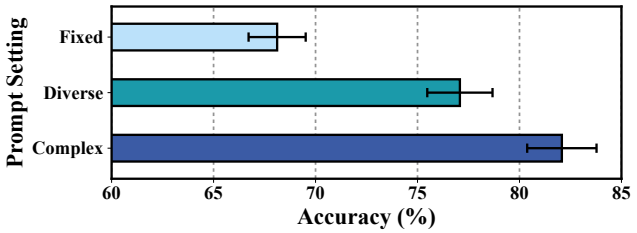


Figure 8: Ablation of the description types. Legend: Fixed – template-based descriptions; Diverse – free-form descriptions; Complex –temporal action combination.

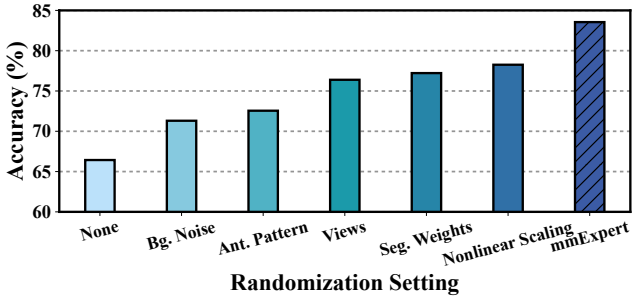


Figure 9: Ablation study of domain randomization factors. We evaluate the final performance by assessing the impact of adding each factor individually. Legend: mmExpert– all randomization factors adopted; Bg. – background; Ant. – antenna; Seg. – segment; None – no randomization factors adopted.

scaling yields the most substantial improvement. Other randomization elements, such as radar view variations, body segmentation weights, and background noise, also lead to a certain increase in performance.

Comparison for LLM and non-LLM models. To evaluate our LLM's classification capability, we adapted it by replacing its final layer with a classification head and compared it against the classification model (ViT encoder + MLP decoder) previously trained on the same data for classification tasks. The results in Table 3 show the fine-tuned LLM classifier achieves higher accuracy. This demonstrates a key advantage: leveraging knowledge transferred from a large pre-trained model enables superior generalization and accuracy, even when fine-tuned on limited task-specific data.

5.4 Robustness to Multipath Reflections

To address concerns about environmental factors, we evaluated mmExpert's performance in an environment with significant multipath reflections caused by nearby metallic surfaces. Our system mitigates these effects using a two-fold strategy. First, range gating isolates signals within the target's specific distance, effectively rejecting delayed multipath signals. Concurrently, by analyzing the time-Doppler signal, we filter out off-axis reflections from static objects by focusing on the target's distinct velocity profile.

This approach proved effective in our evaluation. The system achieved a classification accuracy of 84.6% in an ideal open-space setting, which only slightly decreased to 83.4% when strong reflectors (e.g., walls) were introduced. This minor degradation confirms the robustness of mmExpert in more complex, realistic environments.

Table 3: Performance comparison between our LLM-based classifier and a standard baseline.

Setting	Accuracy (%)
Baseline Classifier (from scratch)	84.6
LLM-based Classifier (fine-tuned)	**85.1**

5.5 System Overhead

During the signal synthesis stage, with 48 GB CUDA memory (1 × Nvidia L20 GPU), the simulation of a 12-second motion (50 fps) sequence and mmWave signal synthesis can be achieved within 1 second. In the LLM inference stage, with 96 GB of CUDA memory (2 × Nvidia L20 GPUs), a single Question & Answer (Q&A) can be completed within 3 seconds.

6 Related Work

6.1 mmWave Signal Synthesis Systems

In recent years, numerous studies have focused on mmWave-based human sensing [32]. However, the lack of large-scale datasets and the high cost of data collection remain significant obstacles to the development of mmWave-based human sensing systems [5, 8, 28]. To overcome these challenges, researchers have proposed to synthesize mmWave radar data from diverse sources, including vision sensors [2, 28, 33] and text descriptions [5, 6, 34].

Generating mmWave radar data from text offers greater flexibility than vision-based approaches, as it is less dependent on data sources. One method uses generative models like GANs or diffusion models [6] to directly synthesize radar data, though it requires substantial real-world data for training to ensure authenticity. Another approach generates human motion sequences from text and synthesizes mmWave data via simulation [5, 34], decoupling the process from data constraints while preserving the physical principles of mmWave radar. RFGenesis [5] pioneers this approach, combining text descriptions with diffusion models to enhance data quality. Similarly, Text2Doppler [34] employs a large corpus to guide human motion generation and subsequent radar synthesis.

6.2 Multi-modal Large Language Models

2D large multimodal models [18, 26] have demonstrated exceptional versatility and capability in tasks such as image captioning, visual question answering, visual grounding, visual dialogue, and even segmentation. Unified under the next-token prediction framework, these models derive significant benefits from large-scale image captioning datasets [17]. Additionally, they often leverage GPT-based methods for data augmentation, achieving remarkable success within the 2D domain. This paradigm has proven both efficient and adaptable when extended to other fields, such as 3D point clouds [27] and LiDAR signal interpretation [30].

To the best of our knowledge, we are the first to extend large multimodal models to the domain of mmWave signal understanding by introducing **WaveLLM**, a multimodal large language model capable of directly interpreting mmWave signals. The primary challenge

in this endeavor lies in the scarcity of paired mmWave-language data. To address this, we propose mmExpert, a versatile data curation pipeline designed to enhance the language alignment of the mmWave encoder and facilitate the fine-tuning of WaveLLM.

7 Conclusion

This paper introduces mmExpert, a novel framework for text-to-mmWave generation and mmWave-to-text understanding, leveraging LLMs to bridge the gap between textual descriptions and mmWave sensing data. mmExpert demonstrates significant advances in scalability, cost-effectiveness, and performance. Leveraging only synthetic data, mmExpert achieves state-of-the-art results in signal generation and classification tasks, showcasing robust zero-shot capabilities on real-world data. This work establishes a foundation for the integration of LLMs in mmWave sensing, paving the way for more versatile and efficient applications.

Acknowledgments

This work is supported by the National Science Fund of China under grant No. 62472379, No. 62394341, No. 62394344, and No. 92467301, Key Research and Development Program of Zhejiang Province No. 2025C01012.

References

[1] Marah Abdin, Jyoti Aneja, Hany Awadalla, and et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219* (2024).

[2] Karan Ahuja, Yue Jiang, Mayank Goel, and Chris Harrison. 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.

[3] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, and et al. 2016. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *ECCV*.

[4] Qiming Cao, Hongfei Xue, Tianci Liu, Xingchen Wang, Haoyu Wang, Xincheng Zhang, and Lu Su. 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.

[5] Xingyu Chen and Xinyu Zhang. 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.

[6] Guoxuan Chi, Zheng Yang, Chenshu Wu, Jingao Xu, Yuchong Gao, Yunhao Liu, and Tony Xiao Han. 2024. RF-diffusion: Radio signal generation via time-frequency diffusion. In *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking*. 77–92.

[7] Kaikai Deng, Dong Zhao, Zihan Zhang, Shuyue Wang, Wenxin Zheng, and Huadong Ma. 2023. Midas++: Generating training data of mmwave radars from videos for privacy-preserving human sensing with mobility. *IEEE Transactions on Mobile Computing* 23, 6 (2023), 6650–6666.

[8] Kaikai Deng, Dong Zhao, Wenxin Zheng, Yue Ling, Kangwen Yin, and Huadong Ma. 2024. G 3 R: Generating Rich and Fine-Grained mmWave Radar Data From 2D Videos for Generalized Gesture Recognition. *IEEE Transactions on Mobile Computing* (2024).

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

[10] Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, and Li Cheng. 2024. Momask: Generative masked modeling of 3d human motions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1900–1910.

[11] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. 2022. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 5152–5161.

[12] Xiuzhen Guo, Long Tan, Tao Chen, Chaojie Gu, Yuanchao Shu, Shibo He, Yuan He, Jiming Chen, and Longfei Shangguan. 2024. Exploring biomagnetism for inclusive vital sign monitoring: Modeling and implementation. In *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking*. 93–107.

[13] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR* 1, 2 (2022), 3.

[14] Haotian Jiang, Jiacheng Zhang, Xiuzhen Guo, and Yuan He. 2021. Sense me on the ride: Accurate mobile sensing over a LoRa backscatter channel. In *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems*. 125–137.

[15] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361* (2020).

[16] Fanqi Lin, Yingdong Hu, Pingyue Sheng, Chuan Wen, Jiacheng You, and Yang Gao. 2024. Data scaling laws in imitation learning for robotic manipulation. *arXiv preprint arXiv:2410.18647* (2024).

[17] Tsung-Yi Lin, Michael Maire, Serge Belongie, and et al. 2014. Microsoft coco: Common objects in context. In *ECCV*.

[18] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems* 36 (2023), 34892–34916.

[19] Haipeng Liu, Yuheng Wang, Anfu Zhou, Hanyue He, Wei Wang, Kunpeng Wang, Peilin Pan, Yixuan Lu, Liang Liu, and Huadong Ma. 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 (2020), 1–28.

[20] Rex Liu, Albara Ah Ramli, Huanle Zhang, Erik Henricson, and Xin Liu. 2021. An overview of human activity recognition using wearable sensors: Healthcare and artificial intelligence. In *International Conference on Internet of Things*. Springer, 1–14.

[21] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* (2018).

[22] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PmLR, 8748–8763.

[23] Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084* (2019).

[24] Yimiao Sun, Yuan He, Jiacheng Zhang, Xin Na, Yande Chen, Weiguo Wang, and Xiuzhen Guo. 2023. BIFROST: Reinventing WiFi signals based on dispersion effect for accurate indoor localization. In *Proceedings of the 21st ACM Conference on Embedded Networked Sensor Systems*. 376–389.

[25] Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. 2017. Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE, 23–30.

[26] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Song XiXuan, et al. 2024. Cogvlm: Visual expert for pretrained language models. *Advances in Neural Information Processing Systems* 37 (2024), 121475–121499.

[27] Runsen Xu, Xiaolong Wang, Tai Wang, Yilun Chen, Jiangmiao Pang, and Dahua Lin. 2024. Pointllm: Empowering large language models to understand point clouds. In *European Conference on Computer Vision*. Springer, 131–147.

[28] Hongfei Xue, Qiming Cao, Chenglin Miao, Yan Ju, Haochen Hu, Aidong Zhang, and Lu Su. 2023. Towards generalized mmwave-based human pose estimation through signal augmentation. In *Proceedings of the 29th Annual International Conference on Mobile Computing and Networking*. 1–15.

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

[30] Senqiao Yang, Jiaming Liu, Renrui Zhang, Mingjie Pan, Ziyu Guo, Xiaoqi Li, Zehui Chen, Peng Gao, Hongsheng Li, Yandong Guo, et al. 2025. Lidar-llm: Exploring the potential of large language models for 3d lidar understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 9247–9255.

[31] Haobo Zhang, Xiangchao Huang, Xiuzhen Guo, Shibo He, Chaojie Gu, Yuanchao Shu, and Jiming Chen. 2025. Terahertz Sensing, Communication, and Networking: A Survey. *IEEE Transactions on Network Science and Engineering* (2025).

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

[33] Xiaotong Zhang, Zhenjiang Li, and Jin Zhang. 2022. Synthesized millimeter-waves for human motion sensing. In *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems*. 377–390.

[34] Yi Zhou, Miguel López-Benítez, Limin Yu, and Yutao Yue. 2024. Text2doppler: Generating radar micro-doppler signatures for human activity recognition via textual descriptions. *IEEE Sensors Letters* (2024).