Towards a mmWave Foundation Model with AI-Generated Infinite Environments

Xinghua Sun Junkai Wu

University of Washington

1 Introduction and Problem Definition

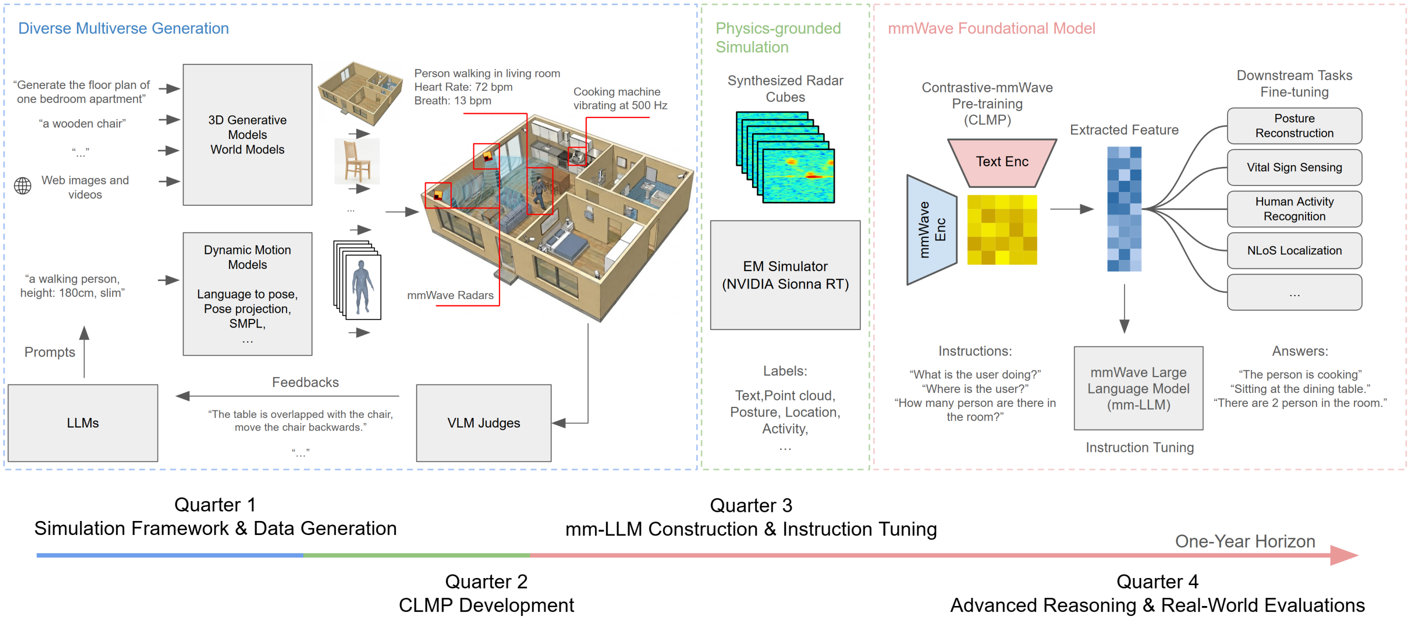

Diverse Multiverse Generation
“Generate the floor plan of one bedroom apartment”
“a wooden chair”
“...”
Web images and videos
3D Generative Models World Models
Person walking in living room
Heart Rate: 72 bpm
Breath: 13 bpm
Cooking machine vibrating at 500 Hz
mmWave Radars
“a walking person, height: 180cm, slim”
Dynamic Motion Models
Language to pose, Pose projection, SMPL, ...
Prompts
LLMs
Feedbacks
“The table is overlapped with the chair, move the chair backwards.”
“...”
VLM Judges
Physics-grounded Simulation
Synthesized Radar Cubes
EM Simulator (NVIDIA Sionna RT)
Labels:
Text,Point cloud, Posture, Location, Activity, ...
mmWave Foundational Model
Contrastive-mmWave Pre-training (CLMP)
Text Enc
mmWave Enc
Extracted Feature
Downstream Tasks Fine-tuning
Posture Reconstruction
Vital Sign Sensing
Human Activity Recognition
NLoS Localization
...
Instructions:
“What is the user doing?”
“Where is the user?”
“How many person are there in the room?”
mmWave Large Language Model (mm-LLM)
Instruction Tuning
Answers:
“The person is cooking”
“Sitting at the dining table.”
“There are 2 person in the room.”
Quarter 1
Simulation Framework & Data Generation
Quarter 2
CLMP Development
Quarter 3
mm-LLM Construction & Instruction Tuning
Quarter 4
Advanced Reasoning & Real-World Evaluations
One-Year Horizon

Figure 1: System overview and one-year horizon of the project. We propose an automated pipeline that generates diverse 3D environments for synthesizing physics-grounded mmWave data. We train a Contrastive Language-mmWave Pre-training (CLMP) foundation model to achieve robust generalizability across diverse sensing tasks/environments/subjects. These aligned representations serve as the perceptual backbone for a mmWave Large Language Model (mm-LLM), enabling the system to understand raw RF signals as language and perform complex interactive reasoning.

Millimeter-wave (mmWave) radar is an emerging sensing modality for the next generation of smart devices, spanning consumer electronics, autonomous robotics, and smart infrastructure. Benefiting from its ability to capture subtle micro-doppler movements, high range resolution, and robustness to lighting conditions, mmWave offers a non-invasive alternative to cameras for smart home applications. Research has demonstrated its versatility across a spectrum of applications, from contactless vital sign monitoring and sleep tracking to fine-grained gesture recognition and occupancy detection. However, a critical bottleneck prevents mmWave sensing from achieving the ubiquity of vision, text, and speech based systems: **the lack of generalization due to data scarcity**.

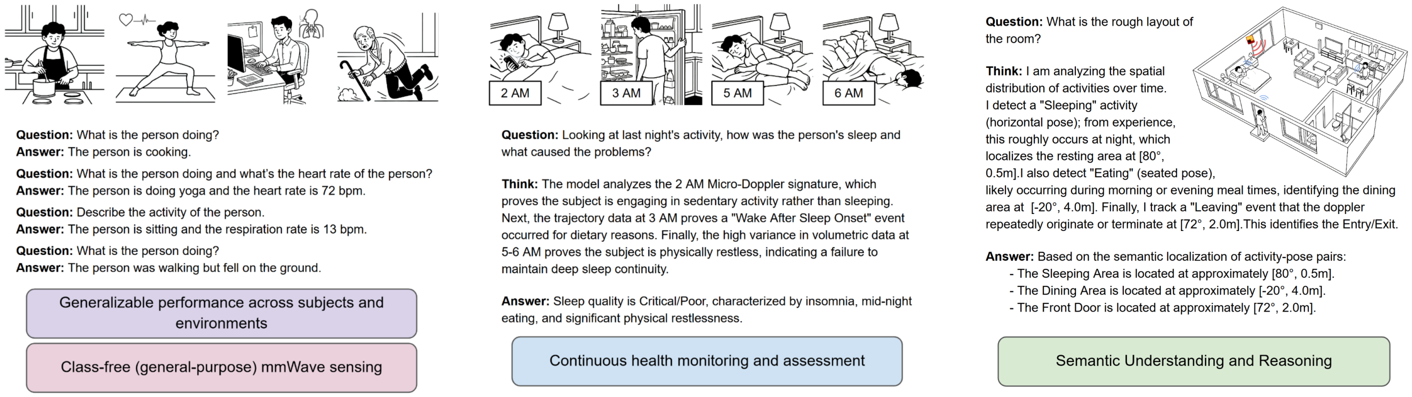
While the field of computer vision (CV) and Natural language processing (NLP) have benefited from the “Scaling Laws” that the model performance increases logarithmically with dataset size [1], RF sensing has lagged behind. Large Language Models (LLMs) are trained on trillions of tokens from the open web, and Vision Transformers are pre-trained on billions of images. In contrast, RF data is difficult to collect, requiring specialized infrastructures, physical deployment in real-world diverse environments, and human subjects [2, 3]. As a result, current mmWave datasets are much smaller, homogeneous, and labeled for narrow, specific tasks. This leads to two fundamental problems that bottlenecks in designing a general purpose foundational models for mmWave sensing systems:

1. **The Generalization Gap:** A model trained on data collected in "Lab A" typically fails when tested in "Living Room B" due to the unique multipath characteristics and the sparse responses of RF signals. The complex reflection, diffraction, and scattering patterns caused by clutter (furniture, walls, static objects) create a "domain shift" that current small-scale models cannot overcome.

2. **The Semantic Gap:** Current mmWave models are "blind" to the context of the environment. They can classify a gesture but cannot reason about the interaction between the user and the surrounding space (e.g., "The user is sitting on the sofa" vs. "The user is falling near the stairs").

Proposed Solution: We propose a fundamental paradigm shift in wireless sensing: transitioning from labor-intensive manual data collection to **AI-driven large-scale synthesis**. Our solution centers on the development of a **general-purpose mmWave foundation**

model pre-trained on an "infinite" synthetic multiverse. By leveraging generative AI to construct physics-consistent 3D environments and simulating physics-grounded electromagnetic wave propagation within them, we generate millions of labeled mmWave data samples paired with rich language descriptions at a significant lower cost of traditional methods. This framework sims to resolve the RF data bottleneck, unlocking robust, generalizable sensing for mmWave devices. By aligning mmWave embeddings with natural language, we move beyond simple detection to enable semantic understanding and logical reasoning, revolutionizing how indoor systems interpret and interact with the physical world.


2 AM
3 AM
5 AM
6 AM
Generalizable performance across subjects and environments
Class-free (general-purpose) mmWave sensing
Continuous health monitoring and assessment
Semantic Understanding and Reasoning

Question: What is the person doing?
Answer: The person is cooking.

Question: What is the person doing and what's the heart rate of the person?
Answer: The person is doing yoga and the heart rate is 72 bpm.

Question: Describe the activity of the person.
Answer: The person is sitting and the respiration rate is 13 bpm.

Question: What is the person doing?
Answer: The person was walking but fell on the ground.

Question: Looking at last night's activity, how was the person's sleep and what caused the problems?

Think: The model analyzes the 2 AM Micro-Doppler signature, which proves the subject is engaging in sedentary activity rather than sleeping. Next, the trajectory data at 3 AM proves a "Wake After Sleep Onset" event occurred for dietary reasons. Finally, the high variance in volumetric data at 5-6 AM proves the subject is physically restless, indicating a failure to maintain deep sleep continuity.

Answer: Sleep quality is Critical/Poor, characterized by insomnia, mid-night eating, and significant physical restlessness.

Question: What is the rough layout of the room?

Think: I am analyzing the spatial distribution of activities over time. I detect a "Sleeping" activity (horizontal pose); from experience, this roughly occurs at night, which localizes the resting area at [80°, 0.5m].I also detect "Eating" (seated pose), likely occurring during morning or evening meal times, identifying the dining area at [-20°, 4.0m]. Finally, I track a "Leaving" event that the doppler repeatedly originate or terminate at [72°, 2.0m].This identifies the Entry/Exit.

Answer: Based on the semantic localization of activity-pose pairs:
- The Sleeping Area is located at approximately [80°, 0.5m].
- The Dining Area is located at approximately [-20°, 4.0m].
- The Front Door is located at approximately [72°, 2.0m].

Figure 2: Unlocking generalizable intelligence in mmWave sensing via mmWave foundation model and LLM alignment.

2 Innovation Proposal and Relation to the State of the Art

2.1 Related Work (Current SoTA)

Automatic scene generation. The paradigm of environment creation has shifted from manual procedural generation to language-guided synthesis, enabling the automated production of diverse, interactive 3D worlds. Holodeck represents the state-of-the-art in this domain, utilizing a LLM to interpret complex user prompts (e.g., "apartment for a researcher with a cat") and generate spatial layout constraints [4]. By querying massive asset libraries like Objaverse [5] and optimizing object placement through constraint solvers, Holodeck produces physically plausible environments. Additionally, the material of the environment (e.g. walls and tables) can be freely edited, making diverse and realistic simulation possible.

Simulation Environments. For simulating mmWave signals, prior methods directly calculate the IF signal with uniformly sampled point clouds from the 3D models [2]. A critical evolution in mmWave simulation is the transition to physics-grounded electromagnetic rendering that models material and geometry, and accurately captures diffraction, multi-reflection, micro-Doppler effects, and etc. The SoTA workflow leverages NVIDIA Sionna to perform high-fidelity ray tracing that captures physics effects and is efficient running on GPUs [6, 7]. This work has also extended to model micro skin displacements enabling the simulation of physiological signals like respiration and heartbeats directly on digital human (SMPL) models [8].

mmWave foundational models. "Towards Foundational Models for Single-Chip Radar" establishes the current benchmark with its Generalizable Radar Transformer (GRT), trained on the massive Rad-1M dataset [6]. By supervising with high-resolution LiDAR point cloud data, GRT successfully recovers 3D occupancy and geometry from the radar doppler map that contains sparse information. However, GRT uses the movement of radar itself to create doppler information of the environment which is different from common radar use cases (e.g., home sensing, health monitoring). Additionally, it lacks semantic information since it's supervised by LiDAR. RadarLLM aligned static radar point clouds with language descriptions [9]. However, it's only trained on simple environments and loses subtle but critical information by discretizing the raw ADC signal to point clouds (e.g., heart beat, micro-doppler).

Multimodal LLMs. Language-supervised contrastive learning (e.g., CLIP [10], CLAP [11]) has established itself as the standard for aligning sensory data with text, enabling robust zero-shot generalization in vision and audio. These aligned embeddings serve as critical conditioning for generative tasks and form the perceptual backbone for the current generation of Multimodal Large Language Models (e.g., GPT-4o [12], Qwen3-VL [13]). By bridging sensory encoders with pre-trained LLMs via lightweight adapters, these systems transcend task-specific "expert" models, offering the ability to solve highly generalized, compositional problems that require complex reasoning and interactive dialogue. Currently, no prior work has aligned mmWave raw data to language on large-scale data.

2.2 Proposed Innovation

The above works have provided the theoretical foundation and engineering tools for building the proposed mmWave foundational model. Our project contains 3 main components: 1) mmWave data generation pipeline, 2) self-supervised mmWave-text pre-training, and 3) mmWave-LLM supervised fine tuning (SFT).

Dynamic Scene Synthesis & Physics-grounded mmWave Simulation. While platforms like Holodeck [4] and ProcTHOR-10K[14] provide diverse, large-scale 3D datasets with editable material properties, they lack dynamic human occupancy. To bridge this gap, we integrate language-driven SMPL models [15] to populate these static environments with moving human agents. To ensure the generated activities are physically and semantically plausible (e.g., preventing a "sleeping" agent from being spawned in a kitchen sink), we utilize a Vision-Language Model (VLM) as a semantic judge to validate agent placement and posture, then provide feedbacks to the initial prompts. Finally, we employ NVIDIA Sionna to perform high-fidelity electromagnetic ray tracing on these populated 4D scenes. By mapping dielectric properties and scattering patterns to different materials, Sionna synthesizes physics-grounded raw ADC data that accurately captures complex wave propagation effects, including multipath, diffraction, and the micro-Doppler signatures of human motion.

Contrastive Language-mmWave Pre-training (CLMP). Leveraging the contrastive learning paradigm mentioned in Sec.2.1, we propose the Contrastive Language-mmWave Pre-training (CLMP) model. Unlike standard approaches limited by scarce real-world RF data, CLMP utilizes our novel simulation framework to generate a potentially infinite stream of synthetic mmWave-text pairs. This ensures dense coverage of the data distribution with precise, multi-granularity captions—ranging from high-level semantics (e.g., "a person walking") to low-level signal characteristics (e.g., "micro-Doppler fluctuations at 30Hz"). The resulting CLMP encoder will serve as a robust backbone for solving complex mmWave challenges, including zero-shot gesture recognition, semantic radar log retrieval, and language-guided clutter suppression.

mmWave Large Language Model (mm-LLM). Following the proven architecture of state-of-the-art open-source models like Qwen3-VL [13] and Audio-Flamingo-2 [16], our mm-LLM will integrate three core components: (1) our proposed CLMP as the domain-specific mmWave encoder, (2) a pre-trained SOTA LLM (e.g., Qwen3 or Llama4 [17]) to provide reasoning capabilities, and (3) a lightweight projector (adapter) to bridge the two. While visual and audio models often struggle to procure high-quality instruction data, we will leverage our simulation framework to generate a massive-scale mmWave Instruction-Following Dataset. By prompting a teacher LLM to generate diverse, complex queries based on our precise ground-truth simulation logs, we can train the mm-LLM via Supervised Fine-Tuning (SFT) to understand RF signals as if they were a language. We expect this model to transcend the limitations of current classifiers, moving from simple tasks like gesture classification or occupancy detection to complex, interactive reasoning tasks, such as "Analyze the subject's gait for early signs of fatigue" or "Analyze the log data and what's a rough layout of the monitored room".

3 One-Year Horizon of the Project

Quarter 1 - Simulation Framework & Data Generation: We will establish the end-to-end pipeline by integrating Holodeck with language-driven SMPL agents, using a VLM judge to ensure semantic consistency (e.g., valid activity placement). Concurrently, we will deploy NVIDIA Sionna to generate "Sim-mmWave-1M," a pilot dataset of physics-grounded ADC data paired with fine-grained text captions.

Quarter 2 - CLMP Development: Focus will shift to training the Contrastive Language-mmWave Pre-training (CLMP) encoder on the pilot dataset. We will optimize the alignment between raw RF signal representations and text embeddings, validating the model's generalized capability through zero-shot benchmarks such as open-vocabulary gesture recognition.

Quarter 3 - mm-LLM Construction & Instruction Tuning: We will assemble the mm-LLM by bridging the frozen CLMP encoder with a state-of-the-art LLM (e.g., Llama4) via a lightweight trainable adapter. To enable reasoning, we will perform Supervised Fine-Tuning (SFT) using a massive, synthetic instruction-following dataset generated by a teacher LLM from our ground-truth simulation logs.

Quarter 4 - Advanced Reasoning & Real-World Evaluations: The final phase will bridge the sim-to-real gap by fine-tuning the model on a small set of real-world RF data using Low-Rank Adaptation (LoRA). We will conclude by evaluating the system on complex, high-level interactive queries (e.g., gait analysis for fatigue, home safety monitoring) and producing a final demonstration.

4 Strength of the Team

Our team has the right mix of skills to execute this full-stack proposal. **Xinghua Sun** anchors the physical and sensing layers with prior expertise in designing mmWave sensing systems. His development of *POLySight* demonstrates deep proficiency with commodity mmWave radars and material sensing, while his *RayTrack* project validates his ability to integrate ray-tracing simulators for signal synthesis. **Junkai Wu** complements this with his expertise in foundation models. His background in contrastive self-supervised learning (*CSSL*) and meta-learning for signal processing provides the theoretical foundation for the radar encoder, while his recent work on fine-tuning LLM agents for interactive navigation ensures the team can execute the complex instruction-tuning phase. To support this work, we have access to the UW Hyak cluster for large-scale model training and the NEWT Lab's radar testbeds for real-world validation.

References

[1] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei, “Scaling laws for neural language models,” *arXiv preprint arXiv:2001.08361*, 2020.

[2] X. Chen and X. Zhang, “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*, ser. SenSys ’23. New York, NY, USA: Association for Computing Machinery, 2024, p. 28–42. [Online]. Available: https://doi.org/10.1145/3625687.3625798

[3] H. Cai, B. Korany, C. R. Karanam, and Y. Mostofi, “Teaching rf to sense without rf training measurements,” *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 4, no. 4, Dec. 2020. [Online]. Available: https://doi.org/10.1145/3432224

[4] Y. Yang, F.-Y. Sun, L. Weihs, E. VanderBilt, A. Herrasti, W. Han, J. Wu, N. Haber, R. Krishna, L. Liu *et al.*, “Holodeck: Language guided generation of 3d embodied ai environments,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 16 227–16 237.

[5] M. Deitke, D. Schwenk, J. Salvador, L. Weihs, O. Michel, E. VanderBilt, L. Schmidt, K. Ehsani, A. Kembhavi, and A. Farhadi, “Objaverse: A universe of annotated 3d objects,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 13 142–13 153.

[6] Z. Zheng, D. Hu, and M. Zhao, “Scalable rf simulation in generative 4d worlds,” *arXiv preprint arXiv:2508.12176*, 2025. [Online]. Available: https://arxiv.org/abs/2508.12176

[7] J. Hoydis, F. Aït Aoudia, S. Cammerer, M. Nimier-David, N. Binder, G. Marcus, and A. Keller, “Sionna rt: Differentiable ray tracing for radio propagation modeling,” in *2023 IEEE Globecom Workshops (GC Wkshps)*. IEEE, 2023, pp. 317–321.

[8] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, *SMPL: A Skinned Multi-Person Linear Model*, 1st ed. New York, NY, USA: Association for Computing Machinery, 2023. [Online]. Available: https://doi.org/10.1145/3596711.3596800

[9] Z. Lai, J. Yang, S. Xia, L. Lin, L. Sun, R. Wang, J. Liu, Q. Wu, and L. Pei, “Radarllm: Empowering large language models to understand human motion from millimeter-wave point cloud sequence,” 2025. [Online]. Available: https://arxiv.org/abs/2504.09862

[10] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, “Learning transferable visual models from natural language supervision,” in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.

[11] B. Elizalde, S. Deshmukh, M. Al Ismail, and H. Wang, “Clap learning audio concepts from natural language supervision,” in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[12] A. Hurst, A. Lerer, A. P. Goucher, A. Perelman, A. Ramesh, A. Clark, A. Ostrow, A. Welihinda, A. Hayes, A. Radford *et al.*, “Gpt-4o system card,” *arXiv preprint arXiv:2410.21276*, 2024.

[13] S. Bai, Y. Cai, R. Chen, K. Chen, X. Chen, Z. Cheng, L. Deng, W. Ding, C. Gao, C. Ge *et al.*, “Qwen3-vl technical report,” *arXiv preprint arXiv:2511.21631*, 2025. [Online]. Available: https://arxiv.org/abs/2511.21631

[14] M. Deitke, E. VanderBilt, A. Herrasti, L. Weihs, K. Ehsani, J. Salvador, W. Han, E. Kolve, A. Kembhavi, and R. Mottaghi, “Procthor: Large-scale embodied ai using procedural generation,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 5982–5994, 2022.

[15] M. Zhang, Z. Cai, L. Pan, F. Hong, X. Guo, L. Yang, and Z. Liu, “Motiondiffuse: Text-driven human motion generation with diffusion model,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 46, no. 6, pp. 4115–4128, 2024.

[16] S. Ghosh, Z. Kong, S. Kumar, S. Sakshi, J. Kim, W. Ping, R. Valle, D. Manocha, and B. Catanzaro, “Audio flamingo 2: An audio-language model with long-audio understanding and expert reasoning abilities,” *arXiv preprint arXiv:2503.03983*, 2025.

[17] Meta Llama Team, “Llama models.” [Online]. Available: https://github.com/meta-llama/llama-models