HybridSim: A Physics–Learning Hybrid Digital Twin for mmWave Human Sensing

Weitao Xiong[1,2] Tianyu Liu[2] Peng Li[2] Kok Chung Chua[1]
Toa Chean Khim[1] Pu Wang[3] Hongfei Xue[3†]

https://weitao-xiong.github.io/HybridSim/

[1]Xiamen University Malaysia
[2]The Hong Kong University of Science and Technology
[3] University of North Carolina at Charlotte
[†]Corresponding Author

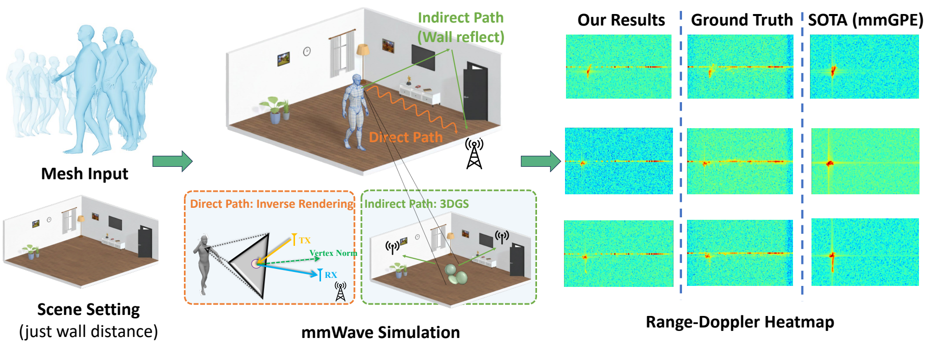

Fig. 1: HybridSim: Dynamic Human Motion mmWave simulation via Physics–Learning & decoupled hybrid rendering. HybridSim synthesizes realistic mmWave signals from dynamic human meshes and indoor scenes by decoupling propagation into direct (inverse rendering-based) and indirect (3D Gaussian Splatting-based) paths. Compared to existing simulators, HybridSim preserves fine-grained kinematic details in Range-Doppler heatmaps, showing high fidelity to the physical ground truth measurements.

Abstract. High-fidelity simulation of mmWave radar signals for dynamic human motion is valuable for developing radar-based human sensing models; yet collecting accurately labeled measurements for a specific deployment site remains expensive. We present HybridSim, a physics–learning hybrid simulator that synthesizes mmWave radar signals from dynamic human meshes under a fixed indoor room configuration, explicitly decoupling propagation into two components. To parameterize the human subject, we use a tri-plane representation to extract human features and a Graph Convolutional Network to stabilize optimization and mitigate gradient instability. The direct signal path is modeled via an inverse-rendering formulation with a microfacet BRDF to capture primary surface reflections. In parallel, the indirect path is approximated by combining 3D Gaussian Splatting with a virtual-receiver geometry to fit and reproduce site-specific multipath interference patterns, achieving substantially lower computational cost than explicit full ray tracing. Experiments in a fixed-room setting show improved agreement with a physically based reference and consistent gains on downstream radar-based human sensing tasks when using HybridSim for site-specific data augmentation.

Keywords: Human Sensing · mmWave Synthesis · Neural Representation

1 Introduction

Millimeter-wave radar provides a robust and cost-effective alternative to vision-based human sensing systems due to its resilience to adverse lighting and preservation of privacy. Developing robust deep learning models for radar perception requires extensive training datasets encompassing diverse human activities. Collecting and annotating such data in physical environments is highly resource-intensive. Consequently, synthesizing high-fidelity radar signals through physical simulation serves as a practical solution to scale training datasets and bridge the domain gap between synthetic and real-world measurements [20, 22].

Existing mmWave simulation methods face distinct challenges when rendering continuous dynamic human motion. Unlike high-density LiDAR, mmWave yield exceedingly sparse spatial measurements. Consequently, radar-based human sensing relies heavily on time-varying micro-Doppler signatures, rather than dense spatial geometry, to infer action. Physics-based ray tracing frameworks provide accurate electromagnetic propagation models but suffer from exponential computational complexity when evaluating higher-order multipath interactions [6, 24]. Furthermore, some ray tracing methods highly demand precise scanned geometries of the environment, making them inconvenient for ubiquitous deployment [27]. Conversely, recent neural representations adapt radiance fields and Gaussian splatting to the radio frequency domain [26, 29]. While these approaches excel at static site-specific channel modeling, they inherently entangle the primary surface scattering of the dynamic human body with complex environmental multipath reflections. This entanglement degrades the synthesis of time-varying micro-Doppler signatures, which are essential for identifying articulated kinematics in single-sensor setups. Other data-driven generation methods and physical augmenters address these efficiency issues but often compromise on coherent multipath interference and physical fidelity [21, 28].

To resolve these limitations, we present HybridSim, a hybrid physics and learning simulator designed to synthesize realistic intermediate-frequency millimeter-wave signals from dynamic human meshes in static indoor environments. We decouple the complex electromagnetic propagation into two distinct and interpretable components. For the direct signal path, we employ an inverse rendering formulation utilizing a microfacet bidirectional reflectance distribution function to accurately capture sharp primary reflections from the subject and the room boundaries. Simultaneously, we model the indirect path by integrating geometrically constrained 3D Gaussian splatting (3DGS) [12] with proxy virtual receivers positioned on the room boundaries. This formulation functions as an efficient multi-scatterer surrogate that learns and reconstructs the site-specific multipath interference patterns at a significantly lower computational cost than explicit multi-bounce ray tracing.

By explicitly disentangling the direct surface scattering from the indirect environmental multipath effects, our framework preserves fine-grained kinematic details within the generated Range-Doppler heatmaps. Extensive evaluations confirm that the synthesized signals maintain high physical fidelity to real-world measurements. Furthermore, using the simulated data for data augmentation significantly improves the classification accuracy of downstream radar-based human activity recognition models under cross-subject and sim-to-real evaluation protocols. The primary contributions of this work include a novel decoupled rendering architecture for efficient dynamic radar synthesis, a unified multi-scatterer method for modeling complex multipath effects, and empirical validation of the simulator in enhancing practical human sensing applications.

2 Related Works

2.1 RF Synthesis

While mmWave systems like mmMesh [22] and M^{4}esh [20] offer robust, privacy-preserving human pose estimation, their generalization to unseen activities is often bottlenecked by the scarcity of comprehensive real-world training data.

To handle these limitations, many approaches have been developed. Data-driven approaches like SynMotion [28] generate radar signals directly from vision-based skeletal data, whereas methods like mmGPE [21] employ physical simulation augmenters to mimic body reflections.

Ray-Tracing RF Simulation. Physics-based simulations rely on geometrical optics and ray tracing to approximate electromagnetic propagation. Foundational work by Yun and Iskander [24] provide a way to modeling these multi-path effects, offering a reasonable explanation of how signals will interact with the environment. And Schüßler et al. [18] extended for large MIMO arrays in automotive contexts. RF-Genesis [6] integrated ray tracing with diffusion models to scale vision-to-RF data generation across diverse environments. While these ray-tracing pipelines achieve high physical fidelity, their computational cost scales exponentially with higher-order multipath interactions, rendering them inefficient for continuous dynamic motion synthesis.

Differentiable RF Rendering. To enable parameter estimation and end-to-end optimization, recent efforts have introduced differentiable RF rendering. Hofmann et al. [10] adapted inverse rendering to near-field radar to reconstruct material properties from measurements. Similarly, InverTwin [3] targets digital twin construction by addressing local non-convexities and boundary discontinuities in RF simulations. However, these methods are primarily designed for inverse problem solving rather than functioning as efficient forward simulators for highly articulated human dynamics.

Neural RF Representations. The inflexibility of monolithic simulators has motivated the development of neural RF representations. Frameworks such as RFCanvas [5] and RFScape [4] fuse visual priors with neural distance fields to build editable channel models. Concurrently, radiance field and 3DGS [12,16] have been adapted for RF propagation. Pioneering this shift, NeRF^2 [29] translates optical neural radiance fields to the RF domain by employing a complex-valued multilayer perceptron to model both amplitude and phase alongside a Friis-based ray tracing model. RF-3DGS [26] reconstructs radio radiance fields from sparse measurements, and GSRF [23] introduces complex-valued Gaussian primitives with directional bases to model amplitude and phase.

While these methods excel at static channel modeling, dynamic human sensing imposes distinct requirements. The simulated signal must preserve the time-varying micro-Doppler signatures of articulated motion while maintaining fidelity to coherent multipath interference. Existing neural methods often entangle these phenomena, making it difficult to isolate the moving target from static environment reflections. Furthermore, some methods rely on distributed multi-sensor setups and will suffer significant performance degradation in practical single-sensor scenarios. HybridSim addresses these gaps by explicitly decoupling the Bidirectional Reflectance Distribution Function (BRDF) based primary scattering from the 3DGS-based multipath reflections.

2.2 Human Modeling

Parametric human body models, such as SCAPE [1], SMPL [14], and SMPL-X [17] represent 3D humans using compact shape and pose parameters applied to a fixed-topology mesh, enabling efficient animation and optimization. In HybridSim, we use SMPL as the mesh template to get dynamic human poses.

3 Method

Overview. HybridSim is a hybrid physics-learning framework that takes dynamic human mesh sequences and static room configurations as inputs and outputs high-fidelity intermediate-frequency (IF) mmWave radar signals. As shown in Fig. 2, we achieve this synthesis by decoupling complex electromagnetic propagation into two interpretable components: a primary direct path and a multipath indirect reflection path.

Specifically, Sec. 3.1 introduces neural feature representations for the human subject and room boundaries alongside dynamic visibility computation. Next, Sec. 3.2 models the direct path using microfacet inverse rendering to capture sharp primary reflections. Sec. 3.3 then details our unified multi-scatterer surrogate model for the indirect path, which approximates complex environmental multipath effects via 3DGS and proxy boundary-geometry. Finally, Sec. 3.4 outlines the Range-Doppler domain loss function for end-to-end optimization.

3.1 Neural Feature Representation and Visibility

To effectively model the complex indoor electromagnetic environment, we parameterize the dynamic human subject and the static scene using distinct neural

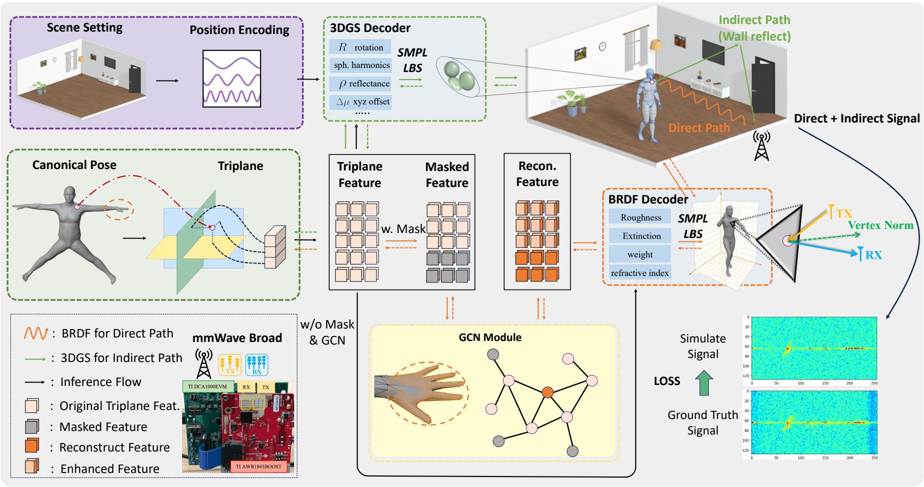

Fig. 2: Pipeline of our framework. We extract human features from a canonical pose using a tri-plane representation and obtain scene features through positional encoding. To model the direct path, the human features are masked and reconstructed via a Graph Convolutional Network to generate spatially regularized embeddings. These enhanced features are processed by a BRDF Decoder to predict material properties including roughness and extinction. Concurrently, for the indirect path, the unmasked Triplane features are fed into a 3DGS Decoder to predict parameters such as rotation, spherical harmonics, reflectance, scaling, and vertex offsets. Both sets of decoded parameters are transformed into the dynamic target pose space using Linear Blend Skinning. Finally, we synthesize the direct and indirect signals, and compute the loss against the ground truth signal.

representations. We extract spatial features for the human target from a canonical SMPL mesh using a tri-plane representation [2, 13]. Mapping the canonical vertices into these three orthogonal feature planes yields a continuous representation that remains fundamentally invariant to articulated poses.

For the static room components, rather than relying on complex scanned geometries, we parametrically construct the environmental boundaries using scalar distance configurations relative to the radar transceiver. Based on these spatial limits, we generate a synthetic, grid-based point cloud for each planar surface, encompassing the walls, floor, and ceiling. Static interior objects such as furniture do not require explicit geometric modeling. Instead, their complex electromagnetic scattering effects are implicitly absorbed and represented by the learnable neural parameters distributed across these proxy boundaries. Then we apply periodic positional encoding to these normalized coordinates to yield the scene feature embeddings, equipping the network to learn high-frequency spatial variations across the simulated environment.

A physically accurate radar simulation inherently requires rigorous handling of self-occlusions and mutual line-of-sight blockages between the subject and the environment. To achieve this, we dynamically determine the visibility of all scatterers using a Hidden Point Removal (HPR) algorithm [15]. The detailed implementation of visibility computation is in Supplementary Material Sec. 6.

3.2 Direct Path Simulation

The direct signal path strictly evaluates the single-bounce (i.e., line-of-sight scattering) radar reflections returning directly to the receiver from the human body and the static room boundaries. By defining this path purely as 1-bounce propagation, we decouple the complex radar signal into a learned amplitude component and a physically derived phase component.

To determine the scattering amplitude, we first pass the extracted neural features through a Graph Convolutional Network (GCN) [25] with stochastic masking, which emulates partial visibility and encourages robust feature recovery. At mmWave frequencies, the human body surface is well-approximated as a piecewise-smooth electromagnetic material field; treating each mesh vertex as an independent scatterer ignores this spatial coherence and can induce non-physical, high-frequency variations in the inferred material parameters. The GCN introduces a soft structural prior via Laplacian-like propagation on the mesh graph, encouraging locally consistent latent material embeddings and improving optimization stability under the high sensitivity of BRDF parameters. In particular, the smoothing effect mitigates gradient concentration on a small subset of vertices, which otherwise can dominate updates and lead to transient memory spikes in the inverse-rendering stage during training. Accordingly, we use the GCN strictly as a training-time regularizer (randomly enabled during training together with stochastic masking) to stabilize convergence and improve generalization to partial visibility. During inference, we directly decode the converged latent features without additional graph propagation, since further message passing may introduce unnecessary spatial smoothing that can suppress legitimate high-frequency scattering changes, while the continuity prior has already been absorbed into the learned latent field through training. Additional analyses and ablations are provided in Supplementary Material Sec. 10.

The human features and encoded scene features are fed into BRDF decoders to predict physical material parameters, including surface roughness, a metallic blending factor, and the complex refractive index. Together with the surface normals of the posed SMPL mesh and static wall patches, these parameters define a unified set of direct-path scatterers $\mathcal{P}_{dir}$. For each scatterer $p \in \mathcal{P}_{dir}$ at time t, we model the scattering amplitude L_p using BRDF formulation with complex Fresnel equations [10]. More details are in Supplementary Material Sec. 7.

While the network predicts the amplitude, the phase is fully determined by propagation physics through the total delay

$$\tau_p(t) = \frac{\|\mathbf{x}_{\text{tx}} - \mathbf{x}_p(t)\| + \|\mathbf{x}_p(t) - \mathbf{x}_{rx}\|}{c}, \tag{1}$$

where c represents the speed of light, $\mathbf{x}_{\text{tx}}$ is the transmitter location, and $\mathbf{x}_{rx}$ is the receiver location. The phase term within the complex signal formulation is driven by the carrier frequency f_0, the chirp slope S, and this propagation delay.

By combining the learned amplitude and the physical phase, the final direct mmWave signal is synthesized as the coherent summation of contributions from all globally visible scatterers:

$$Sim_d(t) = \sum_{p \in \text{vis}} \frac{L_p \cdot \exp\left(j2\pi\left(f_0\tau_p(t) + St\tau_p(t)\right)\right)}{\|\mathbf{x}_{\text{tx}} - \mathbf{x}_p(t)\| \cdot \|\mathbf{x}_p(t) - \mathbf{x}_{rx}\|}. \tag{2}$$

In this equation, $Sim_d(t)$ denotes the aggregated signal at the receiver, and the summation evaluates over the unified subset of visible scattering centers $p \in \text{vis}$ (determined dynamically via HPR) from both the dynamic human mesh and the static room boundaries. Following the equation in mmGPE [21], we omit the negligible residual video phase (RVP) term as well. A fundamental physical distinction between radar signal synthesis and standard optical rendering is that while conventional Bidirectional Reflectance Distribution Function and 3DGS pipelines project 3D properties onto a high-resolution 2D image plane for per-pixel shading and blending, a single radar receiving antenna functions effectively as a spatial point sink (*i.e.*, a single-pixel sensor). Consequently, rather than performing spatial rasterization, the receiver coherently superimposes the complex electromagnetic waves from all visible scattering centers into a unified time-domain sequence. Ultimately, the denominator explicitly accounts for physical free-space attenuation, scaling the signal amplitude inversely with the geometric distances between the transmitter, the scattering elements, and the receiver.

3.3 Indirect Path Simulation: A Unified Multi-Scatterer Surrogate

The indirect path is exclusively dedicated to multi-bounce reflections (≥ 2 bounces), encompassing human-to-wall, floor-to-wall, and wall-to-wall interactions. To strictly prevent double-counting with the single-bounce direct path, explicit multi-bounce ray tracing suffers from exponential computational complexity and is notoriously optimization-unfriendly. Instead of tracking unpredictable discrete rays, we introduce a physically defensible unified multi-scatterer formulation paired with proxy boundary-geometry.

Rather than treating environmental boundaries merely as smooth mirrors, we discretize both the dynamic human mesh and the static room layout (walls, floor, and ceiling) into a unified set of active scattering elements using geometrically constrained 3DGS. Let $\mathcal{P}_{ind} = \mathcal{P}_{human} \cup \mathcal{P}_{env}$ denote this combined set of global scatterers for the indirect path.

Let $\mathcal{K}$ denote the set of room boundaries. We instantiate a set of proxy virtual receivers $\mathbf{x}_{w,k}$ on each boundary $k \in \mathcal{K}$. This design converts the complex, intractable multi-bounce integral into an efficient, differentiable three-segment propagation model: transmitter $\mathbf{x}_{\text{tx}} \rightarrow$ intermediate scatterer $p \in \mathcal{P}_{ind} \rightarrow$ proxy receiver $\mathbf{x}_{w,k} \rightarrow$ physical receiver $\mathbf{x}_{rx}$.

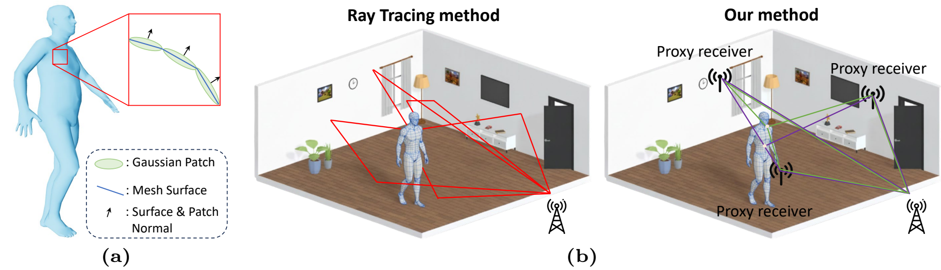

Fig. 3: Illustration of the indirect path simulation using surface-aligned Gaussian patches and proxy wall-geometry. (**a**) The dynamic human mesh is augmented with geometrically constrained Gaussian primitives. To prevent unconstrained volumetric inflation, these scattering elements are flattened into 2D surface patches, with their principal axes aligned to the local tangent plane and compressed along the surface normal. (**b**) Comparison between conventional ray tracing and the proposed proxy wall-geometry approach. While traditional multi-bounce ray tracing (left, red paths) entails a continuous and optimization-unfriendly process, the proposed framework (right) discretizes the multipath environment by instantiating virtual wall receivers on the room boundaries. This formulation simplifies complex indirect reflections into a differentiable three-segment propagation model, as illustrated by the distinct green and purple paths originating from different scattering points on the human subject.

Because the intermediate scatterer p is sampled from the unified set $\mathcal{P}_{ind}$, this formulation acts as a physically-constrained neural representation that explicitly learns the site-specific environmental multipath signatures. Rather than analytically calculating secondary bounces, it encodes the spatial multipath texture into the trainable parameters of the 3DGS. The 3DGS decoders will predict parameters like spherical harmonics, opacity, scaling, and offset for each patch [12], yielding a directional amplitude $SH_{p,k}(\mathbf{x}_p(t), \boldsymbol{\omega}_{p\to w,k})$ evaluated toward the proxy receiver $\mathbf{x}_{w,k}$. To prevent the volumetric inflation typical of unconstrained Gaussians, these scattering elements are formulated as flattened 2D surface patches that strictly adhere to the underlying geometric topology.

The global indirect signal is synthesized by coherently summing the contributions over all proxy boundaries and their respective visible scatterers (excluding self-intersections on the same boundary):

$$Sim_{ind}(t) = \sum_{k\in\mathcal{K}} \sum_{p\in vis(k)} \frac{SH_{p,k} \exp(j2\pi(f_0\tau_{p,k}(t) + St\tau_{p,k}(t)))}{\|\mathbf{x}_{\text{tx}} - \mathbf{x}_p(t)\| \cdot \|\mathbf{x}_p(t) - \mathbf{x}_{w,k}\| \cdot \|\mathbf{x}_{w,k} - \mathbf{x}_{rx}\|} \tag{3}$$

where $\tau_{p,k}(t) = (\|\mathbf{x}_{\text{tx}} - \mathbf{x}_p(t)\| + \|\mathbf{x}_p(t) - \mathbf{x}_{w,k}\| + \|\mathbf{x}_{w,k} - \mathbf{x}_{rx}\|)/c$ defines the proxy propagation delay.

By directly aggregating the multipath energy scattered from both the human body and the environment toward the proxy boundaries, this global summation acts as a robust, optimization-friendly surrogate for multi-bounce ray tracing, preserving complex multipath interference signatures by encoding them as spatial textures, entirely bypassing the prohibitive computational overhead of analytical multi-bounce calculations.

3.4 Loss Function

To bridge the sim-to-real gap prior to frequency transformation, we inject zero-mean complex Gaussian noise into the idealized synthetic signal. The noise standard deviation is dynamically parameterized by a learnable rate multiplied by the signal's mean amplitude (further analyzed in Sec. 4.9).

Directly optimizing the raw time-domain complex signals is highly intractable. Radar absolute phase is extremely sensitive to sub-millimeter errors in kinematic tracking and geometric reconstruction; these inevitable sim-to-real misalignments completely distort the absolute phase, rendering raw complex Mean Squared Error (MSE) unreliable for network supervision [7,8]. Therefore, we adopt an amplitude-only strategy, processing signals into Range-Doppler (RD) amplitude heatmaps via 2D Fast Fourier Transforms (FFT).

Crucially, the proxy-geometry formulation delineates the interface between kinematics-driven path evolution and data-driven electromagnetic effects. While real-world multipath involves complex and often intractable phenomena (e.g., polarization, frequency-dependent absorption, and reflection-dependent phase offsets), our hybrid design assigns complementary roles to geometry and learning. The proxy delay $\tau_{p,k}(t)$ explicitly models the time-varying propagation delay, thereby enforcing that the dominant Doppler evolution is governed by the geometric rate of path-length change ($d\tau/dt$). Under this formulation, static, material-dependent phase offsets affect mainly the relative interference pattern but do not alter the kinematics-induced Doppler trend encoded by $\tau_{p,k}(t)$. In parallel, the neural 3DGS parameters ($SH_{p,k}$) serve as a compact surrogate to absorb unmodeled electromagnetic attenuation and scattering strength. This separation constrains the degrees of freedom of the learned component and empirically reduces the tendency to fit motion-irrelevant heatmap textures, as supported by our ablation study in Sec. 4.6

Furthermore, raw radar reflections exhibit immense dynamic range, where strong primary reflections can monopolize gradients and overshadow vital kinematic details (more detail in Supplementary Material Sec. 8). We balance this optimization penalty by applying a $20\log_{10}(\cdot)$ transformation to convert amplitudes to a decibel (dB) scale.

The framework is optimized end-to-end by minimizing the Mean Squared Error (MSE) between the logarithmic heatmaps:

$$\mathcal{L} = \frac{1}{N}\sum_{n=1}^{N}\left(20\log_{10}\left(|\mathcal{F}_{RD}(Sim)|_n\right) - 20\log_{10}\left(|\mathcal{F}_{RD}(GT)|_n\right)\right)^2, \tag{4}$$

where $\mathcal{F}_{RD}(\cdot)$ denotes the sequential 2D RD FFT operation, and index n iterates over all N bins in the flattened RD heatmap matrix.

4 Experiments

4.1 Implementation Details

We implement our framework using the mmMesh dataset [22], which collects signals via a commercial TI AWR1843BOOST mmWave radar equipped with 3 transmitting (TX) and 4 receiving (RX) antennas. For the training phase, we select eight distinct action sequences performed by a single primary subject. Each sequence comprises approximately 3,000 frames, and we allocate 80% of this data to optimize the network weights. We train the entire framework end-to-end on a single NVIDIA RTX 4090 GPU, which requires a peak memory footprint of approximately 20 GB and takes roughly seven hours. During the inference phase, synthesizing an action takes under 1 second per frame.

4.2 Baselines

The selection of baselines is strictly constrained to approaches that share comparable hardware and signal configurations with the experimental setup. The synthesis of intermediate frequency (IF) and RD heatmaps is inherently coupled with the underlying radar signal model, which encompasses the waveform, center frequency, bandwidth, and chirp schedule, alongside the antenna array configuration, such as transceiver geometry and the arrangement of virtual channels. Many existing mmWave synthesis methods assume divergent transceiver setups or rely on multi-view sensing, rendering the corresponding outputs fundamentally incomparable to the proposed single-radar pipeline. Crucially, the closed-source nature of several contemporary approaches precludes exact reproduction and fair benchmarking. Consequently, the comparative evaluation primarily focuses on mmGPE [21] and RF-Genesis [6], as these frameworks most closely align with the targeted configuration.

4.3 Evaluation Protocol

To assess synthesis quality and cross-subject generalization under a fixed scene, we establish 2 testing settings. For intra-subject synthesis, we evaluate the training subject's remaining 20% frames containing unseen dynamic poses. For cross-subject evaluation, we test the model on a new human subject performing the same eight action categories, each comprising 3,000 frames. This confirms whether the framework overfits the primary subject's body shape or learns transferable physical scattering properties. We quantitatively measure generated RD heatmap fidelity using 3 standard metrics: Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS). All experiments are conducted in the same room layout, and more details about the dataset composition are provided in Supplementary Material Sec. 11.

Table 1: Quantitative Evaluation on the mmMesh Dataset. We report the PSNR, SSIM, and LPIPS metrics for the generated RD heatmaps. The evaluation is divided into two settings: Intra-Subject (testing on the 20% unseen frames of the training subject) and Cross-Subject (testing on 8 actions each with full 3,000 frames of an unseen human in the same scene).

Method	**Unseen Frames (Intra-Subject)**			**Unseen Human (Cross-Subject)**		
	PSNR↑	**SSIM↑**	**LPIPS↓**	**PSNR↑**	**SSIM↑**	**LPIPS↓**
RF-Genesis [6]	20.03	0.219	0.274	20.02	0.218	0.271
mmGPE [21]	20.09	0.221	0.261	20.09	0.220	0.253
HybridSim (Ours)	**22.30**	**0.262**	**0.089**	**22.29**	**0.262**	**0.084**

Table 2: Downstream Task Human Activity Recognition Performance. The mmAP classification network is trained exclusively on synthetic data generated by different simulators and tested directly on real-world ground truth measurements. Higher accuracy on real data indicates a smaller sim-to-real domain gap.

Training Source (Simulator)	**Accuracy (%)↑**	**F1-Score↑**
RF-Genesis	30.68	0.2584
mmGPE	54.22	0.4990
HybridSim (Ours)	**92.07**	**0.9216**

4.4 Main Results: Range-Doppler Synthesis

As shown in Table 1, HybridSim outperforms all baselines in both settings, notably achieving a significant LPIPS reduction that indicates superior structural fidelity. This is visually corroborated in Fig. 4: our framework accurately preserves fine-grained kinematics matching the physical ground truth. Additional qualitative decomposition of the direct path, indirect path, and learned noise contributions is provided in Supplementary Material Sec. 13.

4.5 Downstream Task: Human Activity Recognition

To validate the practical utility of the proposed simulator in bridging the sim-to-real domain gap, we evaluate the performance of the framework on a downstream human activity recognition (HAR) task. We employ the mmAP model [11] as the classification backbone, which utilizes three types of heatmaps (*i.e.*, time-Doppler, time-range, and time-angle) to categorize human activities.

Under a sim-to-real evaluation protocol, we train the classifier exclusively on synthetic data generated for an unseen subject across eight action categories, and then test it directly on corresponding physical ground truth measurements. The inclusion of this downstream evaluation is critical, as standard reconstruction metrics often fail to fully reflect the fidelity of radar cross-section (RCS) dynamics. Metrics like PSNR provide a global measure of pixel-level similarity but suffer from inherent bias due to the sparsity of radar data. Since micro-Doppler signatures

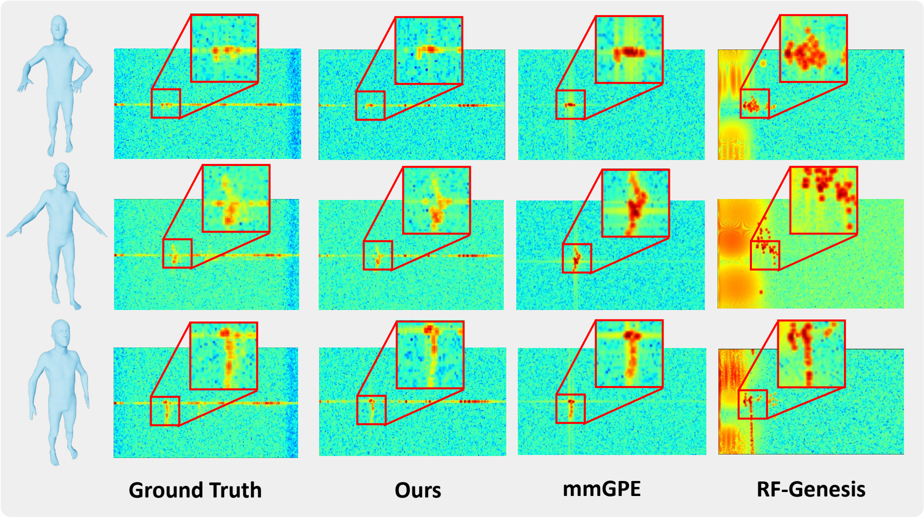

Fig. 4: Qualitative comparison of RD heatmaps. The red boxes provide zoomed-in views of the micro-Doppler signatures generated by articulated human limb movements. Compared to existing baseline simulators (mmGPE and RF-Genesis), HybridSim synthesizes signals that exhibit significantly higher visual fidelity to the physical ground truth, accurately preserving fine-grained kinematic details.

of human motion occupy only a marginal region within the heatmaps, the background noise floor dominates the spatial area. Thus, substantial improvements in synthesizing critical kinematic semantics yield only incremental gains in global PSNR.

The downstream HAR task provides a more representative assessment by heavily penalizing physically unrealistic artifacts and missing micro-Doppler components. As Table 2 shows, the model trained on HybridSim data achieves a recognition accuracy of 92.07%. This constitutes an absolute gain of 37.85% over mmGPE (54.22%), even though HybridSim shows only a modest 10.9% improvement in PSNR (from 20.09 to 22.29) under the cross-subject setting (Table 1). This discrepancy confirms that classification accuracy on real sensor data serves as a more reliable indicator of whether a simulator preserves the essential motion semantics required for robust real-world radar perception.

To further evaluate cross-subject generalization and increase the diversity of the downstream assessment, we additionally conduct HAR evaluation on another held-out human subject. These supplementary results are reported in Supplementary Material Sec. 9 and further confirm the robustness of HybridSim.

4.6 Ablation Study: Impact of Decoupled Hybrid Rendering

To evaluate the necessity of our decoupled physics-learning architecture, we conduct an ablation study focusing on the rendering formulation. We construct

Table 3: Ablation Study on Rendering Formulation. We compare our full decoupled architecture against a Coupled 3DGS variant where both the direct and indirect signal paths are unified under a pure Gaussian Patch representation without using BRDF. The evaluation is conducted on the unseen subject with 8 actions reporting both RD synthesis fidelity and HAR performance.

Method	Synthesis Metrics			Downstream Task (HAR)	
	PSNR↑	SSIM↑	LPIPS↓	Accuracy (%)↑	F1-Score↑
HybridSim (Coupled 3DGS)	19.12	0.252	0.099	85.61	0.8492
HybridSim (w. BRDF)	**22.29**	**0.262**	**0.084**	**92.07**	**0.9216**

a variant termed Coupled 3DGS, where the direct and indirect signal paths are unified and modeled entirely using Gaussian Patches. This configuration removes the BRDF model, forcing the network to implicitly learn all wave-surface interactions using Spherical Harmonics. Consistent with our cross-subject protocol, we evaluate this variant on the RD heatmap synthesis task and the downstream human activity recognition task using the unseen subject.

Quantitative results, summarized in Table 3, show a performance degradation across all metrics when the paths are coupled. Specifically, the PSNR drops significantly from 22.29 dB to 19.12 dB. Because the Gaussian patch relies on low-degree Spherical Harmonics to represent view-dependent amplitude, it struggles to accurately synthesize the sharp, highly directional specular reflections characteristic of the primary direct path. Furthermore, unifying the direct and indirect paths within a single representation forces the network to implicitly resolve complex multipath interference, severely entangling the dynamic human kinematics with the static environmental reflections. This lack of physical decoupling and explicit BRDF constraints heavily impacts downstream performance, with classification accuracy falling from 92.07% to 85.61% and the F1-score decreasing from 0.9216 to 0.8492 compared to our full HybridSim model. This decline confirms that when these distinct physical phenomena are coupled, the network overfits to the training data and fails to capture the intrinsic micro-Doppler semantics required for robust sim-to-real generalization. These findings substantiate that explicitly disentangling primary surface reflections from complex environmental multipath effects is essential for high-fidelity radar simulation.

4.7 Ablation Study: Contribution of Direct and Indirect path

Fig. 10 in the suppl. shows that the direct path carries the dominant human-motion response, while the indirect path contributes weaker scene-dependent multipath reflections. Quantitatively, removing the indirect path reduces HAR performance from 92.07% accuracy / 0.9216 F1-score to 88.96% / 0.8874. This shows that the indirect path is not the main source of human-motion cues, but provides complementary multipath information that improves Sim2Real alignment.

We also add a two-bounce geometric ray-tracing baseline based on ideal specular reflection. In this baseline, the direct path is kept unchanged as ours, and only the 3DGS-based indirect path is replaced. This baseline requires 3.44s/frame, whereas our method requires under 1s/frame, reducing the per-frame runtime by over 70%.. And its HAR accuracy drops to 41.6%, mainly because fixed empirical reflection coefficients cannot model site-specific multipath responses. The resulting spurious high-energy reflections distort RD heatmaps and hurt HAR training, supporting our indirect-path design.

4.8 Ablation Study: Site-specific Second-scene Adaptation.

Since the mmMesh dataset does not contain a second room layout, we repurpose another self-collected dataset. It contains one male with 12 new actions, 8 actions for training and 4 for evaluation. These actions have smaller motion amplitudes than those in main paper (details in Suppl. Sec. 11). HybridSim achieves 22.01 dB PSNR, 0.241 SSIM, 0.108 LPIPS, and 88.68% accuracy / 0.8875 F1-score for HAR. This shows that HybridSim can be adapted to a new room layout. The lower HAR here is partly due to weaker limb-related reflections caused by smaller motion amplitudes, which makes fine-grained action discrimination more difficult.

4.9 Ablation Study: Effect of Noise Floor on Sim-to-Real Alignment

To evaluate the impact of dynamic noise injection on bridging the sim-to-real domain gap, we compare HybridSim variants with different noise-floor strategies under the HAR task. As introduced in Sec. 3.4, idealized physical simulations inherently lack the sensor noise present in real-world measurements. We delegate the approximation of the unpredictable environmental and hardware noise floor to statistical learning. Without injecting synthetic noise, a downstream classifier trained on pristine simulated data tends to overfit to an unrealistically clean background. As shown in Table 4, the variant without any noise injection, denoted as HybridSim (w/o noise), achieves an accuracy of only 72.79% and an F1-score of 0.6790. This significant performance drop highlights the necessity of background noise for robust sim-to-real transfer.

To establish a control variable for comparison, we evaluate a variant injected with a fixed noise floor. It uses the same constant noise level as the baseline methods in Sec. 4.5; details are provided in Supplementary Material Sec. 12. Applying this matching static noise elevates the classification accuracy to 90.51% and the F1-score to 0.9068. Surpassing the baselines under the exact same noise constraint proves that the primary performance driver is the decoupled physics-learning architecture rather than the noise augmentation. The full model with a dynamically optimized noise level achieves the highest accuracy of 92.07% and an F1-score of 0.9216. This progressive sim-to-real alignment is visually evident in Fig. 5. Compared to the unrealistically clean background of the zero-noise simulation, learning the optimal noise rate during training directly aligns the synthetic data distribution with the statistical properties of the physical ground truth, further closing the sim-to-real gap.

Table 4: Quantitative evaluation of noise injection strategies on the downstream HAR task. We compare the classification accuracy and F1-score of models trained on sim data without noise, with a fixed noise, and with learnable noise.

Training Source (Simulator)	Accuracy (%)↑	F1-Score↑
HybridSim (w/o noise)	72.79	0.6790
HybridSim (w. fixed noise)	90.51	0.9068
HybridSim (w. learnable noise)	**92.07**	**0.9216**

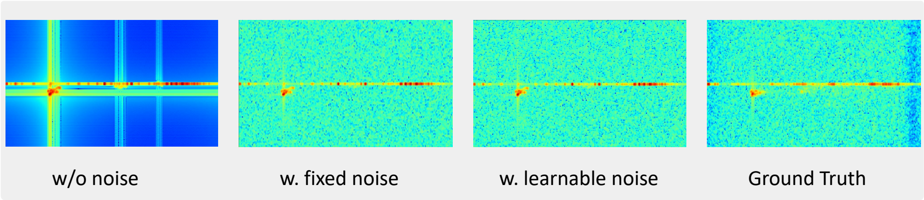

Fig. 5: Visual comparison of Range-Doppler heatmaps generated under different noise injection strategies. From left to right: simulation without noise, simulation with a fixed noise, simulation with a learnable noise, and ground truth. The learnable noise strategy produces a background noise distribution that most closely resembles the real-world sensor measurements.

5 Conclusion & Limitation

In this paper, we present HybridSim, a physics-learning hybrid simulator designed to synthesize high-fidelity mmWave radar RD heatmaps from dynamic human meshes. By explicitly decoupling signal propagation into an inverse-rendering-based direct path and a 3DGS-based indirect path, the framework successfully disentangles primary surface scattering from complex environmental multipath effects. Extensive evaluations, alongside substantial improvements in downstream human activity recognition tasks, demonstrate the superior capability of HybridSim in bridging the sim-to-real domain gap compared to existing baselines.

While HybridSim exhibits robust cross-subject generalization for unseen human targets, the current static scene representation is optimized for specific training environments, requiring parameter fine-tuning for novel indoor layouts. A compelling direction for future research involves integrating generalized neural scene priors to decouple the environmental feature embeddings from specific room geometries. This extension aims to facilitate zero-shot or rapid few-shot adaptation to arbitrary novel environments, thereby further advancing the scalability of synthetic radar data generation for ubiquitous human sensing applications.

Acknowledgements

This work was supported in part by the Xiamen University Malaysia Research Fund Cycle 14/2024 under Grant No. XMUMRF/2024-C14/IECE/0052.

References

1. Anguelov, D., Srinivasan, P., Koller, D., Thrun, S., Rodgers, J., Davis, J.: SCAPE: Shape Completion and Animation of People. Association for Computing Machinery, New York, NY, USA, 1 edn. (2023), https://doi.org/10.1145/3596711.3596797
2. Chan, E.R., Lin, C.Z., Chan, M.A., Nagano, K., Pan, B., de Mello, S., Gallo, O., Guibas, L., Tremblay, J., Khamis, S., Karras, T., Wetzstein, G.: Efficient Geometry-aware 3D Generative Adversarial Networks . In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16102–16112. IEEE Computer Society, Los Alamitos, CA, USA (Jun 2022). https://doi.org/10.1109/CVPR52688.2022.01565
3. Chen, X., Ding, J., Zheng, K., Fang, X., Zhang, X., Lu, C.X., Li, Z.: Invertwin: Solving inverse problems via differentiable radio frequency digital twin (2025), https://arxiv.org/abs/2508.14204
4. Chen, X., Feng, Z., Qian, K., Zhang, X.: Radio frequency ray tracing with neural object representation for enhanced rf modeling. In: 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 21339–21348 (2025). https://doi.org/10.1109/CVPR52734.2025.01988
5. Chen, X., Feng, Z., Sun, K., Qian, K., Zhang, X.: Rfcanvas: Modeling rf channel by fusing visual priors and few-shot rf measurements. In: Proceedings of the 22nd ACM Conference on Embedded Networked Sensor Systems. p. 464–477. SenSys '24, Association for Computing Machinery, New York, NY, USA (2024). https://doi.org/10.1145/3666025.3699351
6. Chen, X., Zhang, X.: 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. p. 28–42. SenSys '23, Association for Computing Machinery, New York, NY, USA (2024). https://doi.org/10.1145/3625687.3625798
7. Chen, X., Zhang, X., Xia, Q., Fang, X., Lu, C.X., Li, Z.: Differentiable radio frequency ray tracing for millimeter-wave sensing (2023), https://arxiv.org/abs/2311.13182
8. Das, R.: Building differentiable simulators for mmwave radar signals. https://www.seas.upenn.edu/~richeek/blog/2024/mmwave-material-extraction/ (2024), accessed: 2026-03-02
9. Heitz, E.: Understanding the masking-shadowing function in microfacet-based brdfs (2014), https://api.semanticscholar.org/CorpusID:123674006
10. Hofmann, N., Wirth, V., Bräunig, J., Ullmann, I., Vossiek, M., Weyrich, T., Stamminger, M.: Inverse rendering of near-field mmwave mimo radar for material reconstruction. IEEE Journal of Microwaves **5**(2), 356–372 (2025). https://doi.org/10.1109/JMW.2025.3535077
11. Joshi, V., Xu, S., Cao, Q., Zhu, Y., Wang, P., Xue, H.: Towards robust mmwave-based human activity recognition using large simulated dataset for model pretraining. In: 2024 IEEE International Conference on Big Data (BigData). pp. 7332–7337 (2024). https://doi.org/10.1109/BigData62323.2024.10825019
12. Kerbl, B., Kopanas, G., Leimkuehler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Trans. Graph. **42**(4) (Jul 2023). https://doi.org/10.1145/3592433
13. Kocabas, M., Chang, J.H.R., Gabriel, J., Tuzel, O., Ranjan, A.: HUGS: Human Gaussian Splats . In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 505–515. IEEE Computer Society, Los Alamitos, CA, USA (Jun 2024). https://doi.org/10.1109/CVPR52733.2024.00055

14. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: Smpl: a skinned multi-person linear model. ACM Trans. Graph. **34**(6) (Nov 2015). https://doi.org/10.1145/2816795.2818013
15. Mehra, R., Tripathi, P., Sheffer, A., Mitra, N.J.: Technical section: Visibility of noisy point cloud data. Comput. Graph. **34**(3), 219–230 (Jun 2010). https://doi.org/10.1016/j.cag.2010.03.002
16. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: representing scenes as neural radiance fields for view synthesis. Commun. ACM **65**(1), 99–106 (Dec 2021). https://doi.org/10.1145/3503250
17. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A., Tzionas, D., Black, M.J.: Expressive Body Capture: 3D Hands, Face, and Body From a Single Image . In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10967–10977. IEEE Computer Society, Los Alamitos, CA, USA (Jun 2019). https://doi.org/10.1109/CVPR.2019.01123
18. Schüßler, C., Hoffmann, M., Bräunig, J., Ullmann, I., Ebelt, R., Vossiek, M.: A realistic radar ray tracing simulator for large mimo-arrays in automotive environments. IEEE Journal of Microwaves **1**(4), 962–974 (2021). https://doi.org/10.1109/JMW.2021.3104722
19. Walter, B., Marschner, S.R., Li, H., Torrance, K.E.: Microfacet models for refraction through rough surfaces. In: Proceedings of the 18th Eurographics Conference on Rendering Techniques. p. 195–206. EGSR'07, Eurographics Association, Goslar, DEU (2007), https://dl.acm.org/doi/10.5555/2383847.2383874
20. Xue, H., Cao, Q., Ju, Y., Hu, H., Wang, H., Zhang, A., Su, L.: M4esh: mmwave-based 3d human mesh construction for multiple subjects. In: Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems. p. 391–406. SenSys '22, Association for Computing Machinery, New York, NY, USA (2023). https://doi.org/10.1145/3560905.3568545
21. Xue, H., Cao, Q., Miao, C., Ju, Y., Hu, H., Zhang, A., Su, L.: Towards generalized mmwave-based human pose estimation through signal augmentation. In: Proceedings of the 29th Annual International Conference on Mobile Computing and Networking. ACM MobiCom '23, Association for Computing Machinery, New York, NY, USA (2023). https://doi.org/10.1145/3570361.3613302
22. Xue, H., Ju, Y., Miao, C., Wang, Y., Wang, S., Zhang, A., Su, L.: mmmesh: towards 3d real-time dynamic human mesh construction using millimeter-wave. In: Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services. p. 269–282. MobiSys '21, Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3458864.3467679
23. Yang, K., Dong, G., Ji, S., Du, W., Srivastava, M.: Gsrf: Complex-valued 3d gaussian splatting for efficient radio-frequency data synthesis (2025), https://arxiv.org/abs/2502.01826
24. Yun, Z., Iskander, M.F.: Ray tracing for radio propagation modeling: Principles and applications. IEEE Access **3**, 1089–1100 (2015). https://doi.org/10.1109/ACCESS.2015.2453991
25. Zhang, H., Lu, G., Zhan, M., Zhang, B.: Semi-supervised classification of graph convolutional networks with laplacian rank constraints. Neural Process. Lett. **54**(4), 2645–2656 (Aug 2022). https://doi.org/10.1007/s11063-020-10404-7
26. Zhang, L., Sun, H., Berweger, S., Gentile, C., Qingyang Hu, R.: Rf-3dgs: Wireless channel modeling with radio radiance field and 3d gaussian splatting. IEEE Transactions on Wireless Communications **25**, 10419–10433 (2026). https://doi.org/10.1109/TWC.2026.3652154

27. Zhang, L., Sun, H., Sun, J., Hu, R.Q.: Wisegrt: Dataset for site-specific indoor radio propagation modeling with 3d segmentation and differentiable ray-tracing: (invited paper). In: 2024 International Conference on Computing, Networking and Communications (ICNC). pp. 744–748 (2024). https://doi.org/10.1109/ICNC59896.2024.10556262
28. Zhang, X., Li, Z., Zhang, J.: Synthesized millimeter-waves for human motion sensing. In: Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems. p. 377–390. SenSys '22, Association for Computing Machinery, New York, NY, USA (2023). https://doi.org/10.1145/3560905.3568542
29. Zhao, X., An, Z., Pan, Q., Yang, L.: Nerf2: Neural radio-frequency radiance fields. In: Proceedings of the 29th Annual International Conference on Mobile Computing and Networking. ACM MobiCom '23, Association for Computing Machinery, New York, NY, USA (2023). https://doi.org/10.1145/3570361.3592527