AIGC for Wireless Sensing: Diffusion-empowered Human Activity Sensing

Ziqi Wang, *Student Member*, IEEE and Shiwen Mao, *Fellow, IEEE*

***Abstract*—Machine learning (ML) for wireless communications and networking requires abundant, high-quality radio frequency (RF) data, yet collecting this data is often challenging and costly. To address this, we propose RF-ACCLDM (Activity Class Conditional Latent Diffusion Model), a framework designed to generate synthetic RF data for human activity sensing. Operating in latent domains, RF-ACCLDM produces RF data conditioned on activity class labels, supporting various RF technologies and modalities, including Radio Frequency Identification (RFID), WiFi Channel State Information (CSI), and Frequency-Modulated Continuous Wave (FMCW) radar. Training of the framework is universal and achieves consistent quality. This approach outperforms plain diffusion on raw RF data in terms of quality, computational efficiency, and scalability. Using the Frechet Inception Distance (FID) metric, we measure and demonstrate the fidelity of the generated data. Through extensive ablation studies, we demonstrate the effects of varying latent dimensions, noise schedules, and training configurations, validating the robustness of RF-ACCLDM. Furthermore, we evaluate the performance of our model in downstream tasks such as RF-based 3D human pose tracking and human activity recognition (HAR), where it can match or even outperform counterparts trained solely on real data. Our approach offers a scalable and cost-effective solution for enhancing ML-based schemes in wireless sensing and communications.**

***Index Terms*—AIGC, Conditional diffusion, Data augmentation, human activity recognition, RF sensing.**

I. Introduction

Machine learning (ML)-based wireless communications and networking have advanced significantly in the past decade [3]. However, the performance of DL-empowered methods is heavily reliant on the availability of vast, high-quality radio frequency (RF) data. Despite initial success, these models generally lack scalability and generalizability due to the constraints during data collection, typically ranging from limited settings or environments to class imbalance, loss, redundancy, and mislabelling [4]. A greater volume and better quality of data are required, along with additional parameters that need to be configured and learned, since the model architecture has become more intricate and sophisticated. In contrast to other domains like natural language processing (NLP) and computer vision (CV), wireless data measurements are naturally complex and noisy from commercial devices (e.g., SX1276 LoRa Connect transceiver or 5300 Wi-Fi network interface card (NIC)). Collecting high-quality RF data is a notable challenge due to its vulnerability to spatial, spectral, and temporal variations. Changes in the environment or transceiver location can significantly alter the data captured, necessitating new data collection for each unique condition. Additionally, the dependency on frequency bands and specific transceiver protocols introduces further complexity, as data features can drastically differ across the spectrum, exemplified by the contrast between 5GHz WiFi and 75 GHz millimeter wave channels. The dynamic nature of wireless channels, which fluctuate over time and activity, further complicates consistent data collection. Consequently, the intricate process of collecting diverse, reliable, and quality RF datasets incurs substantial costs, underscoring the need for innovative solutions in wireless communications research, and poses the first obstacle towards successful scenarios of “ML/AI for wireless”. The challenges related to these obstacles are further supported by [5]. The paper discusses the significant effort and expense required to collect cross-domain RF data, address signal and environmental sensitivity, and ensure effective calibration. Such challenges underscore the value of our approach, as synthetic data generation can alleviate these burdens by reducing the need for extensive, labor-intensive data collection across diverse domains.

Valiant attempts with deep learning (DL) have been made to address these challenges in different fields of wireless sensing. Chao et al. [6] utilizes cycle consistency loss to mitigate the performance degradation caused by unseen test subjects during training for Radio Frequency Identification (RFID)-based 3D pose tracking, while [7] creates a domain-independent body coordinate velocity profile (BVP) to represent the hand motion in the body coordinates, which enables cross-domain gesture recognition. However, they do not tackle the root issue, which is the lack of high-quality data and the efforts required to collect new data whenever a new domain arises. Data augmentation has been used extensively in the field of wireless sensing to overcome this challenge. Zhang et al. [8] applied various transformations, including time stretching and spectrum scaling, to synthesize Channel State Information (CSI) spectrogram. In [9], three operations are leveraged to generate millimeter wave (mmWave) point cloud samples at varying distances, angles, and human motion velocities. The synthesized data are used as copies of original data with transformed features to enhance the DL model performance, instead of serving as new data that can be seamlessly applied to other downstream tasks. The approaches only work for one RF sensing modality (e.g., WiFi CSI spectrogram or mmWave point cloud), and are not tested in cross-modality scenarios.

On the other hand, the rapidly evolving artificial intelligence-generated content (AIGC) concept has ignited a new revolution in Computer Vision (CV) and Natural Language Processing (NLP) with products such as Sora, ChatGPT,

Z. Wang and S. Mao are with the Department of Electrical and Computer Engineering, Auburn University, Auburn, AL 36849-5201 USA. E-mail: zzw0104@auburn.edu, smao@ieee.org.

This work is supported in part by the NSF under Grants CNS-2107190, CNS-2148382, CNS-2319342, and IIS-2306789. This work was presented in part at IEEE ICNC 2024 [1] and IEEE ICC 2024 [2].

and Midjourney, laying the foundation for the emergence of artificial general intelligence (AGI). These applications are primarily utilized in the context of text-to-image generation or text-prompted AI agents, and they typically use transformer and diffusion models as generative backbones. RF sensing data for human activity recognition (HAR) typically involves high-dimensional complex data with time dimensions, similar to CV data (image, video, or audio). Hence cross-framework sensing [10], [11] encourages applying well-established frameworks in the field of CV and NLP to the wireless field. It naturally leads us to inquire whether it is possible to leverage AIGC to solve problems in wireless sensing, especially for generating RF data. To this end, *diffusion model*, also referred to as denoising diffusion probabilistic model [12], provides a promising solution. Diffusion models learn to generate high-quality, diverse samples that closely mimic the distribution of real-world data by simulating a process that meticulously reverses chaos into structured, realistic data by neural networks such as U-Nets [13].

In this paper, we tackle the lack of high-quality and diverse data challenges in RF sensing-based HAR with a diffusion approach. We examine RF sensing data across three data sources—WiFi CSI, FMCW radar, and RFID—each with distinct modalities or feature types relevant to our analysis. Our WiFi CSI data focuses on the phase difference between neighboring antennas as a modality, while FMCW radar data examines range profiles, and RFID data captures phase variation between consecutive phase readings. Each modality offers unique insights and is further detailed in the following sections. We first perform diffusion on raw RFID sensing data with the RFID-ACCDM system [2]. However, the fidelity of the generated RFID data still falls short of optimal, and the system does not offer robustness and scalability to other RF sensing technologies.

To overcome this issue, we take one step further to propose an Activity Class Conditional Latent Diffusion Model (termed RF-ACCLDM), a conditional latent diffusion model (CLDM) capable of generating super-realistic RF sensing data of rich diversity, based on user input of desired human activity class labels. Built upon [1], our system can be generalized to generate WiFi CSI and mmWave radar data of consistent quality by compressing data modalities of varying RF feature dimensions into the same size of latent dimensions. In contrast, RFID-ACCDM is optimized to RFID data of 36 phase variation measurements (features for diffusion models to learn) only. However, WiFi CSI phase difference can have up to 90 features for three receiving antennas, and FMCW radar range profile can have 256 features from a 256-point FFT. To reduce computational expenses and enhance the generative capabilities of diffusion models, we initially train a Recurrent Variational Autoencoder (R-VAE). This approach allows for the sampling of latent representations that capture the temporal dependencies of RF sensing data. Subsequently, we employ a CLDM to refine the training of the diffusion process within these RF latent dimensions.

Moreover, CLDMs offer distinct advantages for generating high-quality RF data, making them more suitable than traditional generative models such as GANs and VAEs. Unlike GANs, which often struggle with mode collapse—a problem where the model generates limited variations of data—CLDMs provide a probabilistic framework that encourages diversity in generated samples. This characteristic is particularly beneficial for RF data, where preserving the variability of real-world signals is crucial for applications like activity recognition and localization. Additionally, CLDMs model data generation as a denoising process, which aligns well with the inherent noisiness of RF signals caused by environmental interference. By learning to reconstruct data from progressively noisier versions, CLDMs can effectively capture the stochastic nature of RF signals and produce robust samples even in high-noise scenarios. Furthermore, unlike VAEs, which often produce blurry or over-smoothed outputs due to the constraints of KL divergence in the latent space, CLDMs operate in a latent space while leveraging diffusion, allowing them to generate high-fidelity, fine-grained details essential for accurately representing RF signal patterns. This quality is vital in RF data generation, where small variations in signal strength, phase, and amplitude can carry important information. Given these considerations, we adopt CLDM as our approach for generating diverse, high-quality RF data, surpassing the limitations of GANs and VAEs in this domain. The generated data holds significant potential for various downstream wireless tasks, particularly in RF sensing applications, such as enhancing the robustness of human activity recognition (HAR) systems in diverse and challenging environments

The main contributions of this study include:

- To the best of our knowledge, this is the first work that harnesses the power of CLDM to generate RF data. The quality of the synthesized data, in terms of accessibility, quantity, fidelity, and diversity, surpasses that of existing methods. More important, the proposed AIGC model is lightweight, only requiring a small amount of real RF training data to be effective.
- We qualitatively demonstrate the performance of RF-ACCLDM through visual comparisons of its synthesized data with ground truth. Furthermore, we quantitatively show that our generated data is of high quality through metrics of Frechet Inception Distance (FID) [14] and diversity.
- The data generated by our RF-ACCLDM model can significantly enhance the efficiency of HAR tasks, eliminating the requirement for domain gap mitigation using additional real RF data. This was validated in our experiments with two representative downstream tasks of HAR with RF sensing, showing that the DL models trained with RF-ACCLDM generated data surpass the performance of those trained with real RF data.
- By utilizing latent representations, a substantial amount of time and computation resources on diffusion training and inference is saved. Furthermore, this approach opens doors for cross-modality sensing, by having one AIGC model to be capable of generating different modalities of RF sensing data such as RFID phase variations, WiFi CSI phase difference, and mmWave radar range profile data.

In summary, we address the following questions with an *AIGC* *for RF sensing approach*: how to alleviate the substantial costs and efforts involved in collecting large-scale RF data, and how to generate diverse, high-quality synthetic RF data that can effectively support downstream tasks such as HAR and 3D human pose tracking.

The remainder of this paper is structured as follows. We first review related work regarding AIGC and RF sensing in Section II. Then Section III illustratively depicts the proposed system design and the training of the latent diffusion models. Section IV details the challenges for designing a unified framework capable of technology-agnostic AIGC-based RF sensing data generation. Section V presents our experimental study. Section VI discuss future work and Section VII summarizes this paper.

II. Related Works

1) RF sensing for HAR: HAR focuses on identifying specific movements or actions of a person, playing a vital role in daily life by providing advanced insights into human behavior. By leveraging existing RF devices, this technology enables the detection of human activities, supporting a wide range of emerging applications such as healthcare monitoring, autonomous driving, and augmented and virtual reality (AR/VR). Its capabilities extend from recognizing large-scale activities, such as daily activity classification and 3D human pose tracking, to detecting fine-scale motions, including vital signs monitoring and hand gesture recognition [15]. Our study centers on the smart implementation of large-scale HAR.

RFID, WiFi, and Frequency-Modulated Continuous Wave (FMCW) radar have been explored for large-scale HAR [16]. RFID primarily involves inexpensive tags and readers for contactless interaction. The ability to categorize tags based on their inherent electrical energy and frequency allows them to be attached to various parts of the human body, serving as an effective wearable sensor for accurate and precise monitoring of activities. Next, FMCW radar works by measuring the distance and velocity of body movement. Differences in chirp frequencies, also known as beat frequencies, can be harnessed to derive distance, velocity, Doppler frequency, and angular information about the detected human body. Additionally, activity-sensitive features, such as micro-Doppler signatures, can be extracted using short-time Fourier transform (STFT) for sensing tasks. Last but not least, WiFi channel state information (CSI) is a metric that describes wireless channel properties and takes into consideration some important factors affecting signal propagation, like environmental attenuation, distance attenuation, and signal scattering. Common and effective ways of sensing the human body are through data modality of amplitude and phase measurements. These three technologies have been extensively utilized for activity recognition [16]–[18] and 3D pose tracking [19]–[21].

2) Lack of AIGC Adaptations For RF Sensing: Over time, Generative Adversarial Networks (GANs), an earlier approach of AIGC technology has been utilized for data augmentation in the RF sensing scene. In [22], a multimodal GAN was designed to synthesize CSI data to tackle problems with environment changes, with the multimodal system being rather complex, consisting of two generators and one classification model. Liao et al. [23] utilized a GAN network based on time-frequency semantics to synthesize various RF signals regarding gesture recognition to deal with class imbalance issues. Our team has also investigated GAN-based data synthesis [24], utilizing an autoencoder-based GAN to generate RFID signals from 3D human pose data. While GAN-based generation offers advantages such as domain adaptation, rapid synthesis, and a well-established framework, achieving high fidelity in the synthesized data remains a challenge. As a result, GAN-generated data is often limited to being a performance enhancer through augmentation, rather than serving as standalone artificial intelligence-generated content (AIGC) data. Furthermore, effective GAN models typically require complex architectures and multimodal systems, which can be particularly challenging to implement for wireless data. The inherent complexity of wireless signals, combined with the difficulty of training GAN models [25], often leads to low-fidelity outputs. Therefore, a simple yet powerful data augmentation strategy is essential for RF sensing applications.

3) Lack of RF Sensing Applications In AIGC: Applications of AIGC leveraging diffusion technologies have predominantly been centered around CV. Initially, diffusion methods were applied to standard CV datasets, achieving groundbreaking advancements in image synthesis, as demonstrated in [26]. This pivotal work, by demonstrating the viability and advantages of diffusion models for creating realistic and diverse images, inspired researchers in other fields to adapt diffusion models to solve their own problems. For instance, The foundational Diffusion Probabilistic Model (DPM) showcased its prowess in medical image segmentation [27], outperforming leading methods in segmentation accuracy. Cao et al. demonstrated the effectiveness of diffusion models in high-frequency spaces, achieving notable success in fast MRI reconstruction in [28]. Furthermore, conditional diffusion models (CDM) extend the capabilities of basic diffusion models by incorporating conditional information such as class labels and texts. These models effectively capture the intricate relationships between the conditions and the generated data, making them highly suitable for tasks requiring both fidelity and specificity, particularly when dealing with continuous and complex data. For instance, conditional Denoising Diffusion Probabilistic Models (DDPM) and conditional Score-based Diffusion models have been employed for generating 3D point clouds from partial scans and for time-series imputation tasks, respectively, showcasing their superiority over traditional models in [29] and [30].

The pursuit of enhancing content generation has led researchers to explore the simpler, lower-dimensional latent space, where diffusion models are hypothesized to perform even better. The introduction of Latent Diffusion Models (LDMs) marked a significant milestone in achieving state-of-the-art image and video synthesis with reduced computational demands [31], [32]. The unprecedented SORA video generation model [33] reinforces our commitment to the potential of LDMs. Vision-based 3D human pose estimation is a direct means of human motion sensing and has achieved prior success using plain diffusion models [34], but limited by the huge computational overhead caused by the inherently messy and complex human movements. Chen et al. in [35] performed conditional diffusion on the motion latent space, which achieved novel fidelity and diversity on extensive human motion generation with greatly reduced cost. Given their robustness across various domains, diffusion models are particularly well-suited for RF sensing tasks, which involve processing multi-dimensional data comprising time frames and RF features.

III. System Design

The proposed system, illustrated in Fig. 1, consists of a two-stage architecture designed for efficient and high-fidelity RF data generation.

In the first stage, a Recurrent Variational Autoencoder (R-VAE) encodes high-dimensional RF data into a compact latent representation, capturing temporal dependencies critical for accurate activity representation. The R-VAE achieves this through its recurrent layers, which effectively model the sequential nature of RF data. It performs two key functions: sampling from the learned latent distribution to generate diverse data representations, and reconstructing RF data from the latent space. This compression step significantly reduces the complexity of RF data, allowing the subsequent CLDM to operate efficiently while retaining activity-relevant features.

In the second stage, the Conditional Latent Diffusion Model (CLDM) operates within the latent space generated by the R-VAE. By conditioning on activity class labels, the CLDM generates class-specific latent representations. During training, the CLDM employs a diffusion process in the latent space, progressively adding and removing noise to learn realistic variations within each activity class. This training approach enhances the model's ability to produce diverse and robust data representations. Once trained, the CLDM can generate synthetic latent representations corresponding to different human activities. These synthetic latents are then decoded by the R-VAE's decoder to reconstruct high-fidelity RF signals, enabling the generation of realistic and diverse RF data tailored to specific activity classes.

A. R-VAE

RF-ACCLDM aims to generate RF data $x^L_{1:N} = \{x^L_n\}^N_{i=1}$ corresponding to human activities, which is 2D time-series data enriched with multiple features. Here, N indicates the number of time frames and L specifies the number of RF features. RF signals, sensitive to nearby movements, exhibit unique cyclical patterns when captured by RF devices, corresponding to various human activity classes. To capture the temporal dependencies inherent in our RF data and accurately sample time-dependent latent vectors, we integrate Long Short-Term Memory (LSTM) units within the encoder and decoder architecture of the Variational Autoencoder (VAE), naming them the LSTM RF encoder ε and LSTM RF decoder ψ, respectively. An LSTM cell has a complex structure with a gating mechanism designed to tackle the vanishing gradient problem typical in standard RNNs, enabling it to remember information for long periods [36]. The LSTM encoder ε compresses the entire sequences of real RF data $x^n_{1:N}$ into a latent vector $z = \varepsilon(x^L_{1:N}) \in R^{1\times\ell}$, where z is a 1D vector of predefined length ℓ determined by the model architecture. At each time step n, the LSTM encoder outputs the hidden state h_n utilizing the input at the current time step n, hidden state at the previous time step n, and cell state at the previous time step n. This is the way the LSTM encoder tries to capture and retain relevant information from the RF data sequence. The cell state serves as the "memory" of the network, carrying information through the sequence of inputs. The final hidden state is fed into two linear transformation modules to estimate the mean μ and log variance σ^2 of the posterior $p(z|x)$. The latent vector z sampled from $p(z|x)$ is then fed into a linear transformation module to output the initial hidden state for the decoder ψ. The initial hidden state and cell state are stacked across the hidden layer depths for tracking short-term and long-term dependencies, respectively. After that, the original RF data can be reconstructed into $\tilde{x}^L_{1:N}$ using the layered states through another linear module. The LSTM encoder ε and decoder ψ can be modeled by $q_\phi(z|x)$ and $p_\theta(x|z)$, respectively. The former approximates the true posterior and the latter represents the likelihood of the complex process of data generation that results in data $\tilde{x}^L_{1:N}$ from z. ϕ and θ are the variational parameters.

The training objective is to minimize the loss function:

$$\begin{aligned}&\min_{\phi,\theta}\ \mathcal{L}_{R-VAE}(\phi,\theta)\\&= -D_{KL}(q_\phi(z|x^L_{1:N})||p_\theta(z)) + \mathbb{E}_{q_\phi(z|x^L_{1:N})}[\log p_\theta(\tilde{x}^L_{1:N}|z)],\end{aligned}$$

where D_{KL} is the Kullback-Leibler (KL) divergence and the reconstruction probability is the Monte Carlo estimation of the log-likelihood $\mathbb{E}_{q_\phi(z|x^L_{1:N})}[\log p_\theta(\tilde{x}^L_{1:N}|z)$ [37]. The former term can be transformed to $-0.5\sum^{\ell}_{l=1}(1+\log(\sigma^2_l)-\mu^2_l-\exp\log(\sigma^2_l))$, and the latter can be trained with mean squared error (MSE) $(x^L_{1:N} - \tilde{x}^L_{1:N})^2$. In each epoch, the total loss is calculated through $\sum^M_{m=1} x_m$ for M RF data with $x_m = x^L_{1:N}$ being the RF data for the mth individual activity.

A standard normal distribution $\mathcal{N}(0,\mathbf{I})$ is utilized as the prior $p_\theta(z)$ of the latent space. To enable back propagation of the latent sampling, a reparameterization trick is executed to approximate z as $z = \mu + \tilde{\sigma}\cdot\epsilon$, where $\tilde{\sigma} = e^{0.5\times\log\sigma^2}$ and ϵ is sampled from a standard normal distribution $\mathcal{N}(0,\mathbf{I})$ with the same shape of the standard deviation $\tilde{\sigma}$. The encoder and decoder are each implemented by a 3-layer LSTM with a hidden size of 1,024. The latent length ℓ of z is set to 256.

B. RF Data Generation with Conditional Latent Diffusion

Denoising diffusion probabilistic models (DDPMs), as introduced by Ho et al. [12], employ a two-phase process involving the gradual application of noise to contaminate data through a "forward diffusion" phase, followed by a "reverse diffusion" phase that systematically eliminates the added noise to generate new data instances. In the forward phase, a fixed-variance scheduler over a T-length Markov chain transforms the data distribution into an isotropic Gaussian distribution. Conversely, the reverse phase employs another T-length Markov chain to undo the Gaussian noise, learning the transitional kernels parametrically modeled by a neural network $\epsilon_\theta(x_t,t)$ such as the U-Net [13].

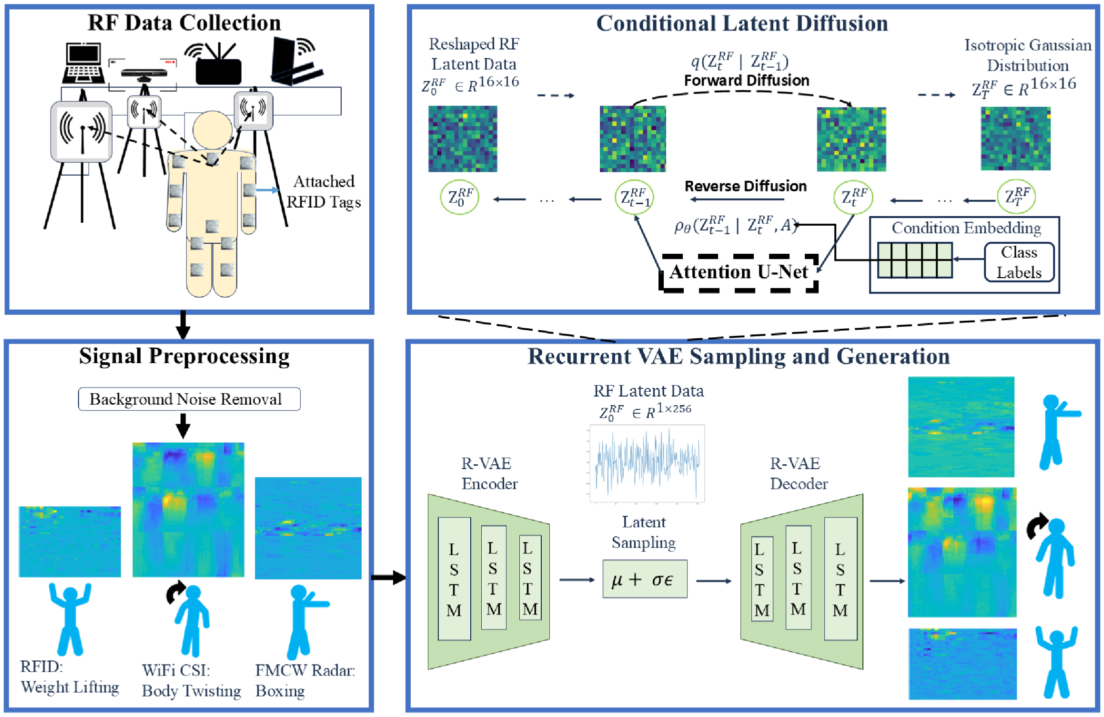


Figure 1. The diagram illustrates the conditional RF data generation process via RF-ACCLDM, starting with data collection and culminating in the production of generative RF data. It also visualizes the forward and reverse diffusion processes occurring within the RF latent space.

However, raw RF data, characterized by intricate, motion-specific features and high-frequency outliers, present great challenges for diffusion models in accurately learning the underlying distribution of the data. The complexity increases with a broader variety of activity classes, making it difficult for a standard DDPM framework with U-Net architecture to generate realistic RF data corresponding to their class labels without incurring significant computational costs.

Here, we introduce our diffusion framework RF-ACCLDM within the condensed and representative RF latent space, i.e., $z \in R^{1\times256}$, to both decrease the computational costs and improve the generative quality. This approach involves initially transforming the latent space into a two-dimensional format of 1×16×16 to accommodate the input requirements of the U-Net. we introduce a streamlined approach to navigate both forward and reverse diffusion processes, utilizing latent vectors denoted as z_t^{RF} for any given timestep t in the noise schedule of LDM. The forward process is conceptualized as a Markov chain of T steps, mathematically described by perturbing the data distribution towards an isotropic Gaussian model:

$$q(z_t^{RF}|z_{t-1}^{RF}) = \mathcal{N}(z_t^{RF}; \sqrt{\alpha_t} z_{t-1}^{RF}, 1-\alpha_t \mathbf{I}),$$
$$q(z_{1:T}^{RF}|z_0^{RF}) = \prod_{t=1}^{T} q(z_t^{RF}|z_{t-1}^{RF}),$$

where the dynamically varying parameter $\alpha_t \in (0,1)$ is crucial for noise scheduling and latent generation sampling. The value of α_t represents the proportion of the original latent retained at each time step. Utilizing the Markov chain, z_t^{RF} can be sampled as $\sqrt{\bar{\alpha}_t} \cdot z_0^{RF} + \sqrt{1-\bar{\alpha}_t} \cdot \epsilon_0$ with $\epsilon_0 \sim \mathcal{N}(0, \mathbf{I})$ and $\bar{\alpha}_t = \prod_{\tau=1}^{t} \alpha_\tau$. Furthermore, $z_0^{RF} = \varepsilon(x_{1:N}^{L})$ is the clean latent vector before noise scheduling of the forward diffusion process, and at the same time, the sampled generative latent vector at the end of the reverse process. Naturally, z_T^{RF} stands for a completely obfuscated latent sample of isotropic Gaussian distribution.

To accommodate generations given a wide array of human activities, from simple (e.g., standing still) to complex ones (e.g., body twisting), we enable conditional latent diffusion by conditioning on activity class labels, denoted as $\mathcal{A}$. We devise a custom reverse diffusion process tailored to the latent space of RF sensing data, mathematically formulated as a Markov chain with transitional kernels parameterized by a U-Net, which predicts the noise to be removed at each step.

$$p_\theta(\mathbf{z}_{t-1}^{RF}|\mathbf{z}_t^{RF}, \mathcal{A}) = \mathcal{N}(\mathbf{z}_{t-1}^{RF}; \boldsymbol{\mu}_\theta(\mathbf{z}_t^{RF}, t \mid \mathcal{A}), \boldsymbol{\Sigma}_\theta(\mathbf{z}_t^{RF}, t \mid \mathcal{A})),$$
$$p_\theta(\mathbf{z}_{0:T}^{RF}|\mathcal{A}) = p(\mathbf{z}_T^{RF})\prod_{t=1}^{T} p_\theta(\mathbf{z}_{t-1}^{RF}|\mathbf{z}_t^{RF}, \mathcal{A}).$$

The following specific parameterization of the transitional kernels for the reverse process $p_\theta(z_{t-1}^{RF} \mid z_t^{RF})$ is considered:

$$\mu_\theta(z_t^{RF}, t \mid \mathcal{A})) = \frac{1}{\sqrt{\alpha_t}}\left(z_t^{RF} - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}(\epsilon - \epsilon_\theta(z_t^{RF}, t \mid \mathcal{A})\right),$$
$$\sigma_\theta(z_t^{RF}, t \mid \mathcal{A}) = \sqrt{\tilde{\beta}_t}\, where\, \tilde{\beta}_t = \begin{cases} \frac{1-\alpha_{t-1}}{1-\alpha_t}\beta_t, & \text{if } t \geq 1 \\ \beta_1, & \text{if } t < 1, \end{cases}$$

where the denoiser U-Net $\epsilon_\theta(z_t^{RF}, t \mid \mathcal{A})$, using activity class labels as the conditional input, estimates the added noise vector ϵ at time step t. μ_θ represents the mean of the Gaussian distribution from which the next state (z_{t-1}^{RF}) is sampled, while σ_θ predicts the variance of the distribution. Together, these parameters guide the reverse diffusion process at each step, providing direction for the denoising operation to iteratively refine the current state, bringing it closer to the original data distribution. Here, $\beta_t = 1 - \alpha_t$ is a small positive constant that controls the amount of noise added at each step of the forward diffusion process. Under this parameterization, the reverse process of our RF-ACCLDM system can be trained by minimizing the following loss function, as shown in [12]:

$$\min_{\theta} \mathcal{L}_{RF-ACCLDM}(\theta) \tag{1}$$
$$= \mathbb{E}_{t,\epsilon\in\mathcal{N}(0,\mathbf{I}),z_0^{RF}\in q(z_0^{RF})} \left\|\left(\epsilon - \epsilon_\theta(z_t^{RF}, t \mid \mathcal{A})\right)\right\|^2,$$

where in every epoch, a timestep t is first randomly chosen for each latent data in the batch followed by backpropagating the loss between the predicted and actual noise once per batch. This is done to introduce randomness into the training process, which can help the model generalize better to potentially improve the diversity and quality of the generated data. During the training of ϵ_θ, the encoder ε is frozen to compress motion into z_0^{RF}. During the reverse diffusion phase, $\epsilon_\theta(z_t^{RF}, t \mid \mathcal{A})$ first predicts $\tilde{z}_0^{RF}$ with T iterative denoising steps. Then ψ reshapes and decodes $\tilde{z}_0^{RF}$ back to RF data for specific human activities, that is $\tilde{x}_{1:N}^{L} = \psi(z) = \psi(\varepsilon(x_{1:N}^{L}))$. For devices with limited computational resources, compressing RF data into latent vectors across various activity classes before initiating the diffusion process can mitigate computational burdens. However, this approach may introduce limitations in terms of the overall scalability and ease of use of the system.

C. U-Net for Denoising

U-Net models are chosen as the trainable denoising function ϵ_θ due to their proficiency in effectively processing and reconstructing noisy latent representations at a given timestep t, allowing for accurate noise prediction. This capability is crucial for the reverse diffusion process to generate new samples. The training objective in each epoch $(\epsilon_\theta - \epsilon)^2$ can be modeled by the MSE function between the predicted noise ϵ_θ and the introduced noise ϵ. To account for the specific timestep t of the diffusion process, sinusoidal positional encodings are employed to enhance the ability of the model to recognize the noise level and timestep, thus improving denoising performance. Additionally, to facilitate diffusion generation conditioned on activity classes, class labels are transformed into dense representations via a multilayer perceptron (MLP) layer. This class embedding is seamlessly incorporated into the U-Net by combining with the positional encoded timestep t, effectively creating a modified timestep $\tilde{t}$. This helps the denoiser learn to adjust its predictions based on the desired output class and the current stage of the reverse diffusion process. The implementation of our U-Net model is shown in Fig. 1.

The U-Net model architecture entails residual blocks and self-attention mechanism. This mechanism enables the model to focus on relevant features across the entire input space, enhancing its ability to capture long-range dependencies and intricate patterns in the latent data. The encoder compresses our reshaped latents $z_0^{RF} \in R^{16\times16}$ to as small as $R^{4\times4}$. A holistic guide on the generative process is provided in Algorithm 1 and Algorithm 2 for the training and sampling procedures of RF-ACCLDM, respectively.

Algorithm 1 Training Procedure of RF-ACCLDM

```
repeat
    $z_0^{RF} \sim q(\varepsilon(x_0^{RF}))$; // Get a batch of latent vectors and labels from datasets.
    $t \sim \text{Uniform}(1, 2, ..., T)$; // Randomly sample a timestep $t$ for each data in the batch
    $z_t^{RF} = \sqrt{\bar{\alpha}_t} z_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$ $(\epsilon \sim \mathcal{N}(0, \mathbf{I}))$;// Add noise to latents at the sampled $t$
    $\tilde{t} = t \bigoplus \tilde{\mathcal{A}}$; // Concatenate the embedded label with $t$
    $\nabla_\theta = \left\|\left(\epsilon - \epsilon_\theta(z_t^{RF}, \tilde{t})\right)\right\|^2$; // Take gradient descent step on loss between added noise and predicted noise
until Convergence
```

Algorithm 2 Sampling Procedure of RF-ACCLDM

```
Input: $\mathcal{A}$
$z_T \sim \mathcal{N}(0, \mathbf{I})$; // Sample a random noise latent for data generation
for $t = T, ..., 1$ do (start the reverse process)
    $s \sim \mathcal{N}(0, \mathbf{I})$ if t > 1, else $s = 0$; // Sample a new random noise tensor
    $z_{t-1}^{RF} = \frac{1}{\sqrt{\alpha_t}}\left(z_t^{RF} - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(z_t^{RF}, t \mid \mathcal{A}\right) + \sigma_\theta(z_t^{RF}, t \mid \mathcal{A})s$; // Class-conditionally generate the next latent vector using $\mu_\theta$ and $\sigma_\theta$ from the current noise tensor
end for
Return $\boldsymbol{x}_{\mathbf{0}}^{\boldsymbol{RF}} = \psi(z_0^{RF})$; // Finally generate a latent vector sampled from the latent distribution and decode it back to reconstructed RF data
```

IV. Challenges In Technology-agnostic Generation

RF sensing-based HAR involves a variety of data modalities, even within the same RF sensing technology [38]. The challenges include fitting diverse RF signal features from various sources into the training of a universal AIGC model, and handling the variability in signal patterns caused by different human activities. To this end, we develop a technology and modality-agnostic AIGC framework, using RFID phase variations, WiFi CSI phase difference, and FMCW radar range profile as representative examples.

1) RFID Platform. In this study, 12 passive RFID tags are attached to the specific joints of a participant. During various poses performed by the participant, three reader antennas interrogate the tags, gathering phase variation data, which represents the change between two consecutive phase readings from the tag responses. This data format effectively captures the detailed movements of the participant's joints:

$$\Delta\phi_{RFID} = \text{mod}\left\{\frac{4\pi(S_t - S_{t-1})f_\alpha}{c}, 2\pi\right\}, \tag{2}$$

in which S_t stands for the tag-to-antenna distance for the tth sampled data on the same channel α, and c is the speed of light. In (2), $(S_t - S_{t-1})$ indicates the shift in the relative distance from the previous sample, making it appropriate for monitoring the tag's movement. RFID phase data is sampled at a frequency of 110 Hz.

2) FMCW Radar Platform. In this study, an off-the-shelf FMCW radar (IWR1843BOOST) is used. The frequency of the intermediate frequency (IF) signal (between transmitting and receiving chirps), which reflects the distance between the radar and an object, is determined by the formula $f_{IF} = S2d/c$, where d is the tag-radar distance and S is the slope of the frequency modulation of the transmitted signal. This relationship allows for the extraction of range profiles, $X[k]$, indicating the strength of reflected signals at varying distances, by applying a 1D Fast Fourier Transform (FFT) on the sampled IF signals, as

$$X[k] = Ae^{j\phi_{IF}} P_{N_s}\left(\frac{2\pi k}{N_s} - \omega_{IF}\right),\ 0 < k \leq N_s, \tag{3}$$

in which ω_{IF} and ϕ_{IF} are the discrete angular frequency and the phase of the IF signal, and $P_N(\omega)$ is the Fourier transform of a square window function of length N. The FMCW radar operates at a sampling rate of 10 Hz for $X[k]$. With a 256-point Range-FFT, the range bin resolution is approximately 0.044 meters. This level of precision is adequate for detecting movements across various human body joints.

3) WiFi Platform. The CSI (5 GHz), a fine-grained feature representation of amplitude and phase across the OFDM WiFi channel, is captured by a commodity WiFi platform. The phase value differences between neighboring antennas for the nth subcarrier are utilized in this paper:

$$\Delta\phi_{CSI} = (\phi_{k,n} - \phi_{(k+1)} \text{ mod }_{k,n}) + \epsilon, \tag{4}$$

where k represents the antenna that collects the phase data, and ϵ stands for the random noise. For each time frame of the activity being captured, 30 subcarrier-level phase information is collected from each of the three antennas to obtain 90 phase difference samples.

4) Remarks. Numerous and varying amounts of measurements are captured simultaneously for different RF techniques, yet their sensitivity to human activity can vary substantially. In the case of RFID technology, the scenario diverges markedly. The sensitivity, as evidenced in the measurements, hinges critically on the positioning of the RFID tags on body joints, given that the phase values received are determined by the tag movements, hence all of the 36 measurements from the 3 antennas are crucial. As for FMCW, signal strength in the approximate range of 1.8 to 2.5 meters (a total of 64 measurements) is more responsive to human movements owing to this being the average distance from the subject to the FMCW radar. Consequently, measurements at this range ought to have a greater influence on the correct extraction of motion features. WiFi CSI, on the other hand, typically requires utilizing all 90 subcarriers-level information for activity sensing, particularly when subsequent feature extraction methods, such as spectrogram generation, are employed. The increasing number of measurements of various values amplifies the difficulty of training AIGC models with respect to memory and time costs.

We experiment with training plain diffusion models on RFID data, arranged in a 3D format of time frames, joint numbers, and antennas, for better learning of RF features across tags and antennas [2]. Suboptimal fidelity is obtained, while the computation time and costs on a readily available GPU GTX 1660 Ti are concerning. With more than 6 classes of RF activity data, the model fails to converge and the generated data suffers from conforming to the correct labels. We further proceed to training WiFi CSI data with the same model. The inputs and outputs of the model have to be adjusted. The computational load and memory requirements increase with larger input dimensions. This is due to the larger spatial dimensions of feature maps in each layer, requiring more computations for convolutions and noise scheduling during forward and backward passes. The inherently messy nature of CSI measurements also renders the training complicated and time-consuming. However, by leveraging RF-ACCLDM, heterogeneous RF sensing data can be used to train a universal AIGC model that is of lightweight and consistently higher quality.

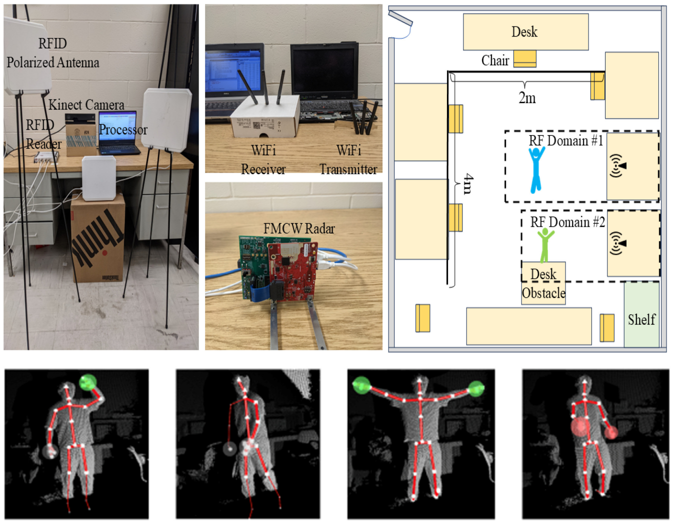


Figure 2. The configuration of the experimental system for RFID, WiFi CSI, and FMCW radar sensing.

V. Experimental Study

A. Prototype System and Dataset Collections

We construct a prototype utilizing a range of representative RF technologies, such as RFID, 5GHz WiFi, and FMCW radar. The RFID setup comprises an off-the-shelf Impinj R420 reader, passive ALN-9634 (HIGG-3) tags, and three S9028PCR polarized antennas. For the WiFi CSI platform, a standard Intel 5300 NIC capable of operating at 5 GHz frequencies is employed. Additionally, an IWR1843 Boost single-chip FMCW mmWave sensor, operating in the 76 to 81 GHz range, is utilized for the mmWave platform. A Lenovo laptop equipped with a GTX 1660 Ti GPU and a workstation with RTX4000 were used for signal processing, as well as for training and inference of the DL models. The configuration of the system is depicted in Fig. 2.

Data is collected by capturing eleven distinct activities performed by a test subject positioned within the detection range of various RF sensing platforms and a Kinect 2.0 device. The activities include: raising the left arm (LA), raising the left leg (LL), drinking (DK), waving up and down (UD), boxing (BX), standing still (ST), twisting (TW), walking (WA), squatting (SQ), kicking (KI), and weight lifting (WL). The test subject repeats the full range of these activities continuously to facilitate consistent data sampling.

For RFID-based data acquisition, twelve passive RFID tags are strategically affixed to the test subject at key joint locations, including the pelvis, neck, left hip, left knee, right hip, right knee, left shoulder, left elbow, left wrist, right shoulder, right elbow, and right wrist. This configuration ensures comprehensive coverage by the three polarized antennas, guaranteeing at least one antenna is always in communication with each tag. Meanwhile, WiFi data collection leverages a transmitter set to the *injection* mode and a receiver in the *monitor* mode, operating at the 5.3GHz frequency band. Furthermore, the FMCW radar with model IWR1843 Boost, is used to create detailed range profiles of the area where the activities of the test subject transpire. Complementarily, the Kinect device concurrently captures visual data, facilitating the pretraining of our models with synchronized RFID-vision datasets, enabling RFID-based 3D pose estimation at 7.5Hz.

Only nine RF data files of 64 time frames are used to train the lightweight RF-ACCLDM system for each RF platform. Six test subjects are involved in the collection of real ground truth RFID and Kinect data. We designate two main data domains within a lab environment of 4m by 2m, illustrated in Fig. 2. Five of the six test subjects are collected under RF data domain #1. Within these five subjects, three are collected under homogeneous settings (i.e., similar body shapes, viewpoints, and locations). The diffusion and DL models for downstream tasks' training data only come from these domains. The rest two subjects are collected under heterogeneous domains compared to the former (slightly different subjects and locations between the test subject and the RF sensing device). Compared to RF domain #1, RF domain #2 includes a shorter test subject, slightly altered movement variations, a different equipment and test subject location, and a new obstacle (desk). This is utilized to test the generalization abilities of our generated data in out-of-domain scenarios. There are roughly a total of 99 minutes of training data (9 minutes for each activity) after applying a sliding window of 3 seconds width with a sliding factor of 1 second. We choose a 1-second sliding factor to create more diversity within the collected data since activities can change moderately in this time window. There are 4.7 minutes of testing data for each activity. These data are used for training and testing the performance of DL models.

The above data collection process empirically shows that the need for specialized hardware, sensitivity to environmental factors, labor-intensive setup, high storage and processing demands, and diverse domain problems all contribute to the substantial costs and efforts involved. This is further backed up by some other related works [23], [39], [40], where multiple RF sensing data sources need to be collected for sensing tasks.

RFID exhibits robust resistance to environmental interference, making it especially reliable for capturing detailed features of human motion. The fidelity of human movement features is more accurately represented through RFID tags affixed to body joints, outperforming other platforms, particularly in environments subject to change and variability. Such resilience allows RFID data to carry a richer array of motion features, enhancing the precision of human activity tracking. Therefore, RFID is selected as the representative RF sensing data to showcase the comprehensive benefits of our data generation approach. We leverage generated RFID data to train models for two distinct downstream tasks in RF sensing: RFID-based 3D pose estimation (a regression task) and Human Activity Recognition (HAR, a classification task). We evaluate the performance of identical DL models trained on generated, mixed (generated data mixed with real data), finetuned (trained with generated data and finetuned on real data), and real data across these tasks, providing a comparative analysis of their efficacy and, more importantly, identifying what the unparallel quality and diversity of our generative model can unlock the road of *AIGC for Wireless*.

B. Diffusion Implementation

The basic setup for training the diffusion models is as follows: A fixed linear schedule β_t is chosen, starting $\beta_1 = 10^{-4}$ and increasing to $\beta_T = 0.02$ over $T = 1,000$ steps for the diffusion training process. Drawing from the principles in [41], a classifier-free guidance approach is adopted for more robust data generation. The model is trained without conditions for a portion of each epoch, specifically 10%. During the sampling phase, a gradual shift from unconditional to conditional generation occurs. This technique significantly bolsters the model's performance in generating class-consistent RF data and improves the overall sample quality. We choose the specific number of training epochs based on the convergence behavior observed during model training, ensuring the stability of both the loss function and generated RF data quality. Using open-sourced visualization tool "Weights and Biases", we can visualize the RF data quality in each training step, and when combined with the knowledge of the training and validation loss, we eventually decide to use 1,200 training epochs. The U-Net architecture is selected due to its proven effectiveness in diffusion models, particularly for image-like data where spatial dependencies need to be captured effectively. Since RF sensing data often carries spatio-temporal patterns, U-Net's encoder-decoder structure allows for multi-scale feature extraction and contributes significantly to the high fidelity of generated data.

C. Quality of Generated Data

1) Direct Visualization:

As discussed in Section IV, RFID data captures fine-grained features that represent joint movement information across three antennas, effectively simulating a 3D plane. A scaled color visualization of each feature value provides a clear and intuitive way to compare differences and similarities between samples generated by different models. Fig. 4 presents a comparison of randomly generated RFID samples from RF-ACCLDM, synthesized samples from RF-RGAN, and real RFID data across nine distinct activities.

The scaled colormap 'Parula' is used, transitioning from dark blue (negative values) to yellow (positive values), with green representing mid-range values. For each time frame, activity features across the three antennas are arranged in the tag order described in Section V-A, with the bottom half representing root and leg joints and the top half representing upper body joints. For instance, in the middle cluster of real RFID samples, the drinking activity exhibits distinctly more prominent features in the top half, while the boxing activity shows clear feature patterns spanning both the lower and upper body joints.

Compared to real samples, the RFID samples synthesized by RFPose-GAN appear less sharp and noticeably blurrier, particularly exhibiting reduced non-linearity across time frames, where feature values tend to cluster unnaturally. Though following the overall pattern of real features, significant discrepancies are evident. In particular, the sample of standing still highlights a key limitation of the GAN model, which struggles with uncertainty when attempting to synthesize messy RFID features from mostly constant vision features, resulting in incomplete outputs that retain substantial portions of vision-specific features. In contrast, RF-ACCLDM-generated samples exhibit far fewer missed details and naturally align with the feature patterns of real data while maintaining diversity. Unlike the large divergences observed in RFPose-GAN-generated squatting and kicking samples, RF-ACCLDM-generated outputs display consistency and fidelity to real features. Our approach achieves superior results in terms of sharpness, contrast, and brightness, effectively addressing the limitations of GAN-based methods. Importantly, our method effectively mitigates the difficulty of training plain diffusion models. In Fig 5, plain diffusion-based RFID-ACCDM can generate data with an overly yellowish tone, indicating significant deviations in value scales. Additionally, the joint-related tag features fail to align well with their real counterparts, reflecting the inherent difficulty of learning complex and noisy RF features with plain diffusion models.

In Fig. 3, we showcase the generation process of the RF signal for the kicking activity. On the very left, the synthetic RF signal is at the initial stage of the denoising process (Gaussian noise). Here, across time frames, the RFID features do not have any distinct pattern of kicking (fluctuations of peak values in the leg regions, i.e., 6 to 18 on the y-axis of the signal values map) besides random patterns across the y-axis. The same can be seen from the 3D surface plot. After some steps, the RF signal still appears chaotic but more fine-grained. As the reverse diffusion process reaches the later stage, the RF signal is gradually refined until the distinctive kicking pattern appears. In the final denoising steps, some of the details are further refined as seen in the far right example. This shows a close resemblance to its real counterpart, shown in the middle clusters of the samples in Fig. 4.

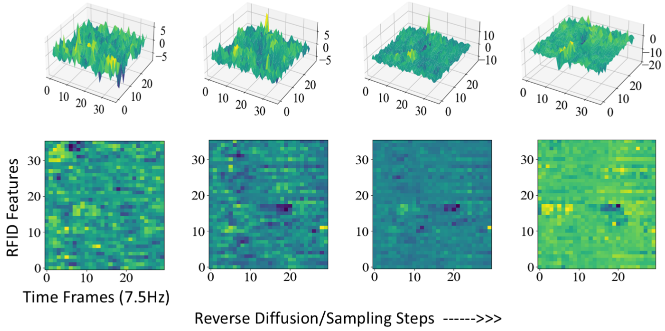


Figure 3. Synthetic RF signal for the kicking activity, shown across different stages of the diffusion generation process. The progression from left to right illustrates the transformation from initial random noise to the final generated signal. From top to bottom, the visualization shifts from surface plots to signal value maps, offering complementary perspectives on the generation process.

2) Quantitative Evaluation: Our proposed model produces high-quality RFID data and ensures significant diversity within the generated data, moving beyond mere replication of the homogeneous features of the training set. While SSIM [42] can provide valuable insights into the perceptual similarity between individual pairs of RF data, its utility is limited in evaluating the performance of such robust generative models. To address this issue and provide a more comprehensive evaluation, we adopt the Frechet Inception Distance (FID), as outlined in [14], to quantitatively assess the similarity in distribution between generated and real RFID data collections. FID evaluates the closeness of feature vectors within a high-dimensional latent space, where a lower FID score denotes closer similarity and, thus, higher fidelity of the generated RFID data to the actual data.

$$\mathbf{FID} = \|\mu - \mu'\|_2^2 + \mathrm{Tr}(\Sigma + \Sigma' - 2\sqrt{\Sigma \times \Sigma'}), \tag{5}$$

where $\mathrm{Tr}(\cdot)$ is the trace linear algebra operation, μ and μ' are the feature-wise means of the extracted feature vectors from real and generated data, respectively, and Σ and Σ' denote the respective covariance matrices.

Furthermore, we utilize the diversity metric, which measures the variance of the high-dimensional feature vectors of RF data across all activity classes, to quantitatively measure the overall diversity of generated data. The diversity metric is defined as:

$$\mathbf{DIV} = \frac{1}{S_d}\sum_{i=1}^{S_d} \left\| f_i - f_i' \right\|_2, \tag{6}$$

in which after choosing two groups of identically sized samples (with size S_d) at random, we compute the overall variance of the RF data. The value of S_d in our experiments is 200.

The FID scores for four representative activities and overall performance are presented in Table I to illustrate the superiority of RF-ACCLDM over other baseline models. The overall FID score of RF-ACCLDM (10.45) is comparable to that of real data (6.22), indicating high fidelity with minimal divergences. In contrast, the other generative models show significantly poorer performance, with FID scores of 25.64 and 50. An FID score of 25.64 reflects moderate divergence from the baseline, suggesting a noticeable decline in data quality, while an FID score of 50 indicates pronounced degradation, with the generated RF data exhibiting substantial differences in visual fidelity compared to real data. For simple activities involving isolated limb movements, such as drinking or waving, RFID-ACCDM performs similarly to RF-ACCLDM. However, for complex activities requiring full-body motion, such as walking and boxing, RFID-ACCDM falls short, though it still significantly outperforms RFPose-GAN.

Using the proposed RF-ACCLDM framework, we demonstrate that data modalities from different RF technologies—such as WiFi CSI phase difference and FMCW radar range profile—can be generated with high fidelity without modifying the model architecture. Unlike RFPose-GAN, whose diversity is limited by the quantity and diversity of the real dataset, RF-ACCLDM generates data with a diversity

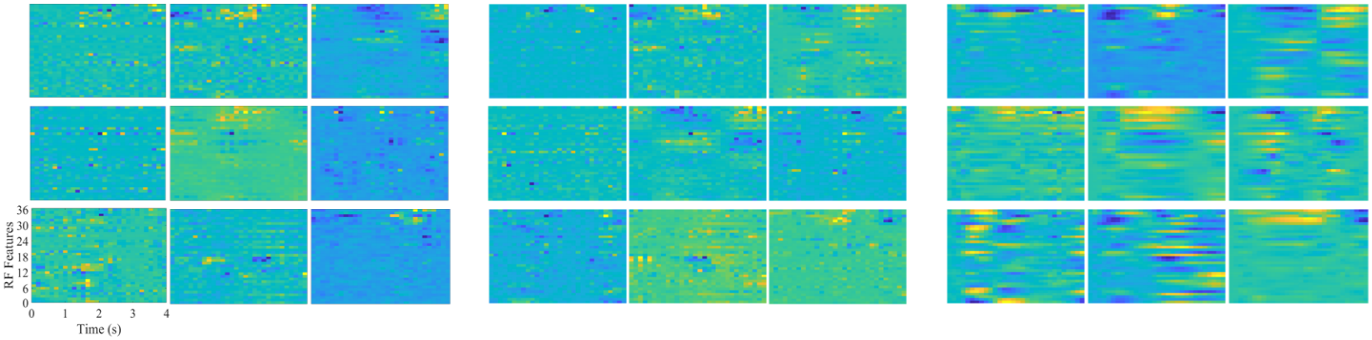


Figure 4. Comparison of data samples from nine activity classes generated by our latent diffusion model with classifier-free guidance (FID: 10.45, left), the training set (FID: 6.22, middle), and the RNN Autoencoder-based RFPose-GAN (FID: 48.89, right). Each cluster represents the following activity classes in a left-to-right and then top-to-bottom order: drinking, waving up and down, boxing, standing still, twisting, walking, squatting, kicking, and weightlifting.

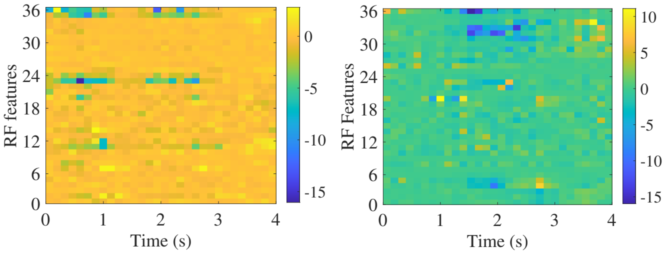


Figure 5. Visual comparisons between the outliers in RFID-ACCDM generated RF data (left) and real data (right) for the activity of waving up and down in the format of 'Parula' Scaled colored images.

Table I
COMPARISON OF SAMPLE QUALITY (MEASURED IN FID) GENERATED BY OUR RF-ACCLDM AGAINST PLAIN DIFFUSION MODELS, AUTOENCODER-BASED RFPOSE-GAN MODELS, AND REAL DATA, EVALUATED ACROSS SELECTED HUMAN ACTIVITIES AND ALL ACTIVITIES

Model	Standing	Waving	Walking	Boxing	Overall
RFPose-GAN	36.18	33.01	44.97	69.56	48.89
RFID-ACCDM	8.79	8.25	20.68	40.54	25.64
RF-ACCLDM	4.56	7.01	3.64	4.84	10.45
Real	5.17	7.36	4.78	4.49	6.22

comparable to a well-curated real dataset, even when trained on a lightweight dataset. As shown in Table II, RFID-ACCDM achieves an unchecked diversity value of 11.10, significantly higher than other models. However, as illustrated in Fig. 5, RFID-ACCDM-generated samples can have FID scores exceeding 100, indicating the presence of outliers that inflate the diversity score. Notably, diversity values are most meaningful when they align closely with those obtained from real RF data.

Table II
COMPARISON OF DIVERSITY SCORES

Model	Diversity score
RFPose-GAN [24]	$9.48^{\pm 0.25}$
RFID-ACCDM	$11.10^{\pm 0.21}$
RF-ACCLDM	$9.16^{\pm 0.31}$
Real	$9.33^{\pm 0.25}$

D. Ablation Studies

To better understand the relationships between training hyperparameters, hardware setup, and generation performance, in-depth ablation studies are conducted, which examine multiple configurations, varying latent dimensions (128, 256, and 512), noise steps (500, 1,000, and 2,000), noise schedules (linear vs. cosine), number of training epochs (400, 800 and 1,200) and GPU hardware (GTX 1660 Ti vs. RTXA4000). Additionally, we measure training and inference time to understand how each setup impacts training and generation speed.

The ablation study results in Table III demonstrate why latent diffusion is necessary: in a resource-limited computing environment (Nvidia GTX 1660 Ti is a laptop-grade GPU, which only has 6 GB VRAM and older-generation architecture. RTX 4000 instead has 16 GB VRAM and newer architecture), plain diffusion on raw RFID consumes 5.51 GB VRAM, nearly the total computing power of GTX 1660 Ti, and it takes 43 hours to train and 26 seconds to infer one sample. However, latent diffusion takes almost half of the training time and memory cost. It only takes 3 seconds to infer a sample, which is much more deployable in real-time applications. In the meantime, the generation fidelity has improved greatly. Plain diffusion on raw CSI and FMCW data fails to start training due to memory overload on GTX 1660 Ti, and still consumes 15.98 and 10.12 GB VRAM on RTX A4000. The number of noise steps has to be reduced to 500 for the diffusion training on raw CSI data to be finished. The generated data quality remains suboptimal, with FID scores of 80.53 and 41.04.

Our findings in this study indicate that the setup with 256 latent dimensions, 1,000 linear noise steps, and 800 epochs achieves the best balance between FID score and GPU memory usage on the RTXA4000. The results highlight that increasing noise steps (e.g., to 2,000) or latent dimensions (e.g., to 512) yields diminishing returns in FID while incurring greater computational costs.

E. Downstream Task I: DL-powered HAR

We train a simple customized 4-layer convolutional neural networks (CNN) model for HAR. The design is identical to that in [1], which consists of four 2D convolutional layers. The last three convolution layers are each followed by a maxpooling2D layer. The convolution output is flattened and fed into a fully connected layer for classification. HAR accuracy indicates the synthetic RF data generated by RF-ACCLDM retains the critical correlations between RF features and activity labels

Table III
ABLATION STUDY RESULTS FOR LATENT DIFFUSION CONFIGURATIONS

Setup	Dimensions	No. Steps	Noise Schedule	Epochs	GPU Model	FID	Training Time and Average Memory Cost	Inference Time
Raw WiFi CSI	64 × 90	500	Linear	1200	RTXA4000	80.53	32 h and 15.98 GB	27 s
WiFi CSI Latents	256	1000	Linear	1200	RTXA4000	16.97	22 h and 4.21 GB	1.7 s
Raw FMCW	64 × 64	1000	Linear	1200	RTXA4000	41.04	26 h and 10.12 GB	13 s
FMCW Latents	256	1000	Linear	1200	RTXA4000	13.67	19 h and 3.53 GB	1.5 s
Raw RFID	**64 × 36**	**1000**	**Linear**	**1200**	**GTX 1660 Ti**	**25.64**	**43 h and 5.51 GB**	**26 s**
RFID Latents	**256**	**1000**	**Linear**	**1200**	**GTX 1660 Ti**	**10.45**	**29 h and 3.12 GB**	**3 s**
RFID 128 Latents	128	1000	Linear	1200	RTXA4000	**13.52**	9 h and 1.04 GB	1.1 s
RFID 512 Latents	512	1000	Linear	1200	RTXA4000	11.81	23 h and 6.6 GB	2 s
RFID 500 steps	256	500	Linear	1200	RTXA4000	11	13 h and 1.28 GB	0.8 s
RFID 2000 steps	256	2000	Linear	1200	GTX 1660 Ti	11.26	54 h and 2.14 GB	4.1 s
RFID Cosine Schedule	256	1000	Cosine	1200	RTXA4000	10.89	18 h and 1.39 GB	1.2 s
RFID 400 Epochs	256	500	Linear	400	RTXA4000	14.65	5 h and 1.41 GB	1.4 s
RFID 800 Epochs	256	1000	Linear	800	RTXA4000	**10.68**	11 h and 1.47 GB	1.4 s

Table IV
COMPARISON OF RECOGNITION ACCURACY FOR 6-CLASS HAR UNDER DIFFERENT TRAINING SCHEMES

	Recognition Accuracy (%)		
Method	Real	Synthetic	Mixed
5-shot real (2 minutes)	52.08	-	-
Limited real (10.5 minutes)	67.82	-	-
Modest real (32 minutes)	83.47	-	-
Sufficient real (64 minutes)	92.63	-	-
RFPose-GAN (64 minutes)	-	64.52	88.04
RFID-ACCDM (64 minutes)	-	88.52	90.69
Our RF-ACCLDM (64 minutes)	-	91.80	93.13

necessary for practical applications. A high HAR accuracy ensures that the RF data conveys meaningful and consistent representations of human activities. Additionally, it indirectly indicates the quality of the underlying CNN model, as robust classification performance requires both high-quality training data and a well-trained model capable of learning intricate patterns within the RF data.

1) Application 1: Synthetic Models Rival Real Models: When the DL model is only trained on generated data but tested on real data, the performance suffers from domain gap. This is particularly apparent with GAN models. However, diffusion models bring unprecedented quality and have made this scenario applicable.

A six-class HAR has been implemented using real, RFPose-GAN, RFID-ACCDM, and RF-ACCLDM models, as summarized in Table IV. The terms real, synthetic, and mixed refer to activity classifier models trained on real data, generated data, or a mixture of a limited amount of real data and synthetic data, respectively. Training with limited real data significantly hampers the performance of CNN classifiers, reducing accuracy to below 70%, while training with a sufficient amount of real data yields the best performance consistently. However, collecting a diverse set of real data remains challenging in the RF data domain. While GAN-based structures can generate data more quickly, their quality is inferior to that of diffusion-based models. For instance, training solely on RFPose-GAN-synthesized data leads to unsatisfactory accuracy at 64.52%. However, mixing a small amount of real data with GAN-generated data can significantly improve the performance of the model to 88.04%. RFPose-GAN-synthesized data do not provide comprehensive coverage as they are prone to overfitting specific modes of the data distribution. Mixing in real data helps fill in the gaps in distribution coverage, enhancing the robustness and generalization of the model. Diffusion and latent diffusion methods demonstrate superior performance when trained on synthetic data alone, achieving accuracies of 88.52% and 91.80%, respectively. The high fidelity of diffusion-generated data results in smaller marginal gains when mixed with limited real data. This is because the generated data already covers a broad range of scenarios and variations present in real datasets, minimizing gaps that real data would otherwise address. For example, mixing RF-ACCLDM-generated data with five shots of real data yields only minor performance gains. Given the difficulty of collecting large-scale datasets, this finding highlights the importance of designing practical applications that effectively leverage the accessibility of real data to optimize synthetic model performance.

To further test model scalability, we extend the evaluation to a nine-class HAR task. Leveraging the increased number of generative classes enabled by latent diffusion, we successfully conduct nine-class HAR even on a resource-constrained GTX 1660 Ti. In contrast, most existing DL works on wireless sensing-based HAR are limited to seven or fewer activity classes, primarily due to the challenges associated with collecting large-scale RF data. We add more complexity to the task by purposefully selecting similar activities, such as weight lifting and waving up and down, or boxing and body twisting. Weight lifting involves coordinated upward and downward motions of both arms, which closely resemble waving. Similarly, boxing includes torso twisting motions, which overlap with body twisting movements involving circulatory shifts of arms and legs.

The training data for each model consists of the same amount of generated data (10.67 minutes per activity class). The nine-class case is more challenging, given the increased number of classes and overlapping features among certain activities. The RF-ACCLDM model achievs an accuracy of 85.82%, falling short of the 90.03% accuracy obtained by the model trained on sufficient real data. Both models face difficulties in distinguishing similar activities; however, RF-ACCLDM struggles more notably in cases such as boxing versus twisting and weightlifting versus waving up and down. This indicates that the domain gap remains a critical issue for RF-ACCLDM-generated data, particularly when applied to tasks with fine-grained activity distinctions. The detailed confusion matrices are presented in Fig. 6.

2) Application 2: Generated Data-based Few-shot Learning Outperforms and Adapts Better Than Real Models: We then proceed to conduct thorough trials using more sophisticated training methods for in-domain and out-of-domain scenarios. The term in-domain denotes the case where the model is trained with data from RF domain #1, and then tested on data from RF domain #1, whereas out-of-domain represents the case where the model is trained with data from RF domain #1, but tested on RF domain #2.

Few-shot Learning is a data-efficient learning strategy that only utilizes a small amount of samples of each category for training. The capabilities of large quantity, high quality, and rich diversity render RF-ACCLDM generated data possible for pretraining a model with comprehensive synthetic knowledge before fine-tuning it with a few shots of real data when deployed in dynamic scenarios. Few-shot learning contributes to wireless-based HAR in practice since it only requires a small number of new samples, but most existing works have only been able to utilize the laboriously collected real data for pretraining their models [43].

Via thorough experiments, we find that this method is far more practical for its stable performance and rapid deployment. We first pretrain the synthetic CNN classifier with 42.7 minutes of RF-ACCLDM generated data per class and then proceed to finetune the model with 3 and 10 shots of real data. As Table V shows, the pretrained synthetic models can be fine-tuned on real data to provide a performance boost for the in-domain case. When fine-tuned on only 3 shots of real data, the recognition accuracy of 88.06% is better than training with solely RF-ACCLDM generated data. When fine-tuned on 10 shots of real data, the accuracy gets a significant boost to a new height of around 95% in just 500 epochs, surpassing the model trained with sufficient real data, which could take 2,000 epochs to reach optimal performance. The plots of training history are shown in Fig. 7, and the comparison highlights that synthetic pretraining combined with limited real data fine-tuning (as in (a)) provides a more balanced learning approach, resulting in a model that generalizes better and experiences less overfitting compared to the model trained on a large real dataset alone. The difference between validation and training loss is as large as 0.5, considerably larger than that of the finetuned model. The detailed performance boost for using the pretraining and finetuning mechanism is shown in the middle and right confusion matrices in Fig. 6, compared to the confusion matrix on the left when the synthetic model is trained with only generative data. For the out-of-domain case, considerable performance enhancements are achieved with a small number of real data from RF domain #2. In contrast, the model would be unfit for practical applications when trained with 3 shots or even 10 shots of real data, as shown in Table V. This proves that the knowledge learned while training on a large set of diverse synthetic data paves the way for model fine-tuning with only a few real samples per class.

Table V
Comparison of Recognition Accuracy for 9-class HAR Using Different Training Schemes Including Fine-tuning with Few-shot Real Data on pretrained models, Training with Few-shot Real Data, and Training with Real Data Guided Latent Diffusion Generated Data in Both In-domain and Out-of-domain Cases

	Recognition Accuracy (%)	
Method	In-domain	out-of-domain
3-shot real	45.17	34.76
10-shot real	60.33	50.89
Sufficient real	90.43	53.65
Trained on RF-ACCLDM generated data from scratch	85.82	69.76
Pretrained and finetuned on 3-shot real	88.06	80.91
Pretrained and finetuned on 10-shot real	95.42	86.81
Latent diffusion using real guidance from test data	82.46	71.32
Latent diffusion using real guidance from training data	88.02	77.51

3) Application 3: Synthetic Models Using Real Guidance: We use the few-shot real samples from training and test datasets as guidance to generate synthetic data in which the few-shot real samples (with artificially added Gaussian noise) replace the random noise at the beginning of the latent diffusion generation to guide the diffusion process. The results can be seen in the last two rows of Table V. This is a convenient way to reduce the dependency on direct access to real data. When leveraging the guidance from real training data, the synthetic model presents improvements (88.02%, compared to 85.82% when trained with data generated from random noise) for in-domain cases and robustness in out-of-domain cases (77%, compared to 69.76% when trained with sufficient real data from RF domain #1). This proves that by using real data as guidance, the diversity of RF-ACCLDM-generated data can be controlled for higher recognition accuracy. However, synthetic data generated using guidance from test data exhibits worse performance in both data domain cases. This is because test data is unseen during the training of the diffusion models, which leads to a higher difficulty for the noise distribution to reverse. This experiment shows that applying real data as guidance to the diffusion process helps reduce the domain gap and can be much more effective at achieving functional performance in out-of-domain cases when there is a lack of real data from this domain.

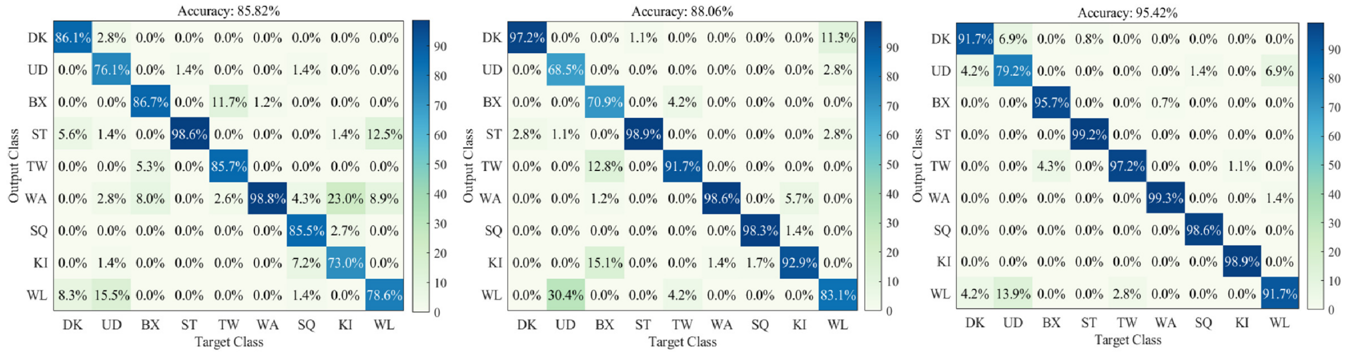
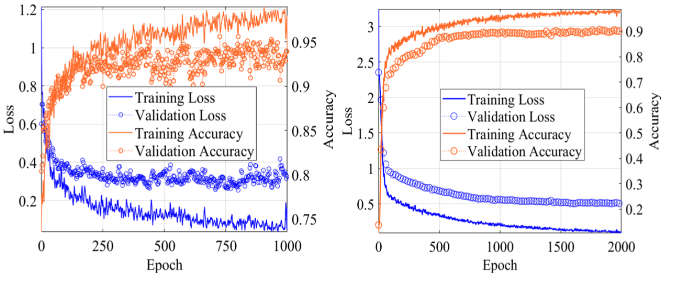


Figure 6. Confusion matrices comparing CNN-based classifier models for the 9-class HAR task: (left) trained on 96 minutes of RF-ACCLDM-generated RFID data, (middle) fine-tuned with 1.8 minutes of real data (3 shots per class) after pretraining on 42.7 minutes per class of RF-ACCLDM-generated data, and (right) fine-tuned with 6 minutes of real data (10 shots per class) using the same synthetic pretraining.



(a) Pretrained + finetuned on 10 shot of real data (b) Sufficient real data (96 minutes)

Figure 7. Traces of the 9-class HAR classifier training and validation across epochs.

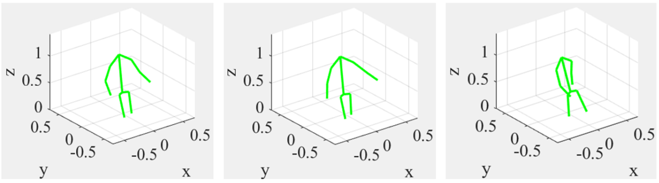


Figure 8. Estimated 3D human pose from RF-ACCLDM generated RFID data with a 0.8-second difference between the three animation video frames presented.

F. Downstream Task II: RFID-based Human Pose Estimation

RFID-based 3D human pose tracking involves using RFID to estimate the three-dimensional position and movements of human subjects. Achieving high accuracy in pose tracking accuracy indicates several key aspects of the RFID signal, including signal strength, consistency, and temporal stability, and serves as a direct and useful indicator of the quality of our generated RFID data.

We train the same deep kinematic neural network used in [16] with real, RFPose-GAN synthesized, and RF-ACCLDM generated data. The one trained on real data will be used as the baseline. When leveraging the model trained on real data, our RF-ACCLDM generated RFID data achieves good 3D human pose estimation as illustrated in Fig. 8. This directly demonstrates that the generated RFID data learns fine-grained movement information, which is then seamlessly mapped to 3D human pose animation. In addition to displaying postures that are realistically human-like, the 3D human poses demonstrate a natural temporal smoothness that closely resembles the real poses captured by Kinect cameras.

However, this application is more tailored to generating new pose data instead of critically alleviating the challenges in RF-based pose estimation. The RFID-Pose network requires supervised training with a one-to-one ratio for RFID and vision-based Kinect Camera data. This incurs an extensive amount of data collection work, including time synchronization between the two data types. The data scarcity also leads to the limitation of poses that can be inferred, while the age of AIGC craves variety in every field. To this end, we creatively find a way to utilize the diverse set of generated RFID data for 3D human pose tracking. We first use a pre-trained RFID-Pose model to estimate synthetic pose data from generated data, and then employ pairs of generated RFID data and estimated pose data for the supervised training of a synthetic RFID-Pose model. When it comes to the RFPose-GAN synthesized RFID data, we pair it with its source data, the simulated pose data, to train another synthetic model. We test and compare the synthetic models with the real model on real data. This validates the quality and practicality of generated data and how well it can be generalized to real data.

For every time frame t, The mean per joint position error (MPJPE) for all 12 joints is computed as follows to evaluate the performance of the RFID-Pose network at estimating 3D human pose from RFID data.

$$MPJPE = \frac{1}{N}\sum_{n=1}^{N}\left\|\hat{P}_n^t - P_n^t\right\|, \tag{7}$$

in which P_n^t denotes the ground truth position sampled by the Kinect device, and $\hat{P}_n^t$ for the estimated position for joint n at time t; and $\|\hat{P}_n^t - P_n^t\|$ stands for the Euclidean distance between the two positions in the 3D space. Presented as the cumulative distribution function (PDF) of MPJPE, the overall performance of the human pose estimation is shown in Fig. 9. Estimating poses for complex activities indicates the boundaries of a practical RFID-Pose system and is typically the weakest aspect compared to simple activities. The synthetic model through RF-ACCLDM achieves a median error of 3.92cm, comparable with the real model, which has a median error of 3.71cm. On the other hand, the synthetic model through RFPose-GAN is inferior with a median error of 4.58cm. It is important to note that, for better visualization, outlier estimations are excluded. Outlier estimations in certain joint positions can be more dominant in synthetic models. The unfiltered median errors for the two synthetic models are

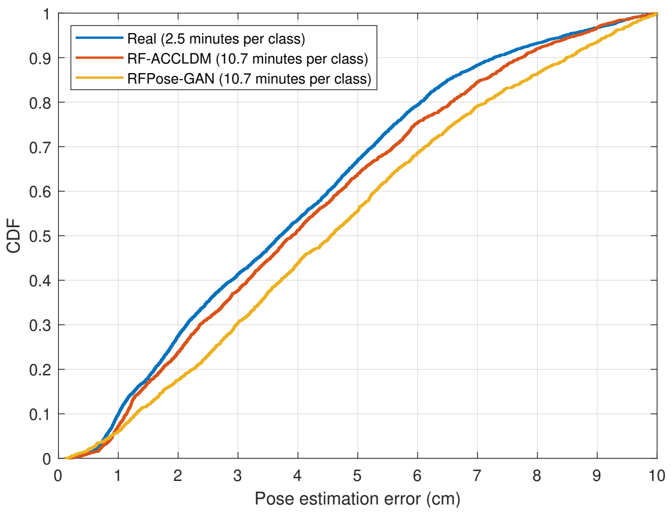


Figure 9. Overall pose estimation performance regarding complex activities in the form of CDF of estimation errors.

4.23cm and 6.1cm, which are gains at levels of 7.9% and 33.2%, respectively. The median errors help explain why the latter synthetic model would achieve inconsistent and foul joint positions at times, while the former one through RF-ACCLDM can estimate consistent and smooth trajectory but slightly deviated joint positions. To achieve this moderate joint estimation performance, the real model requires around 2.5 minutes of data per class, while the two synthetic models require around 10.67 minutes per class. We also note that subject skeletons are key factors of 3D human pose estimations, but we only use one fixed skeleton for the synthetic estimation of 3D human pose from RF-ACCLDM generated data, which limits the abilities of RF-ACCLDM.

VI. Limitations and Future Work

While the RF-ACCLDM framework demonstrates strong performance on small-to-medium-scale RF datasets, its computational cost could escalate with larger datasets or higher-resolution data (e.g., dense WiFi CSI across many antennas or high-resolution radar data). Although latent diffusion mitigates some computational burdens, higher-dimensional or more complex RF data increases the challenges for the VAE, particularly in preserving fine-grained details and variability. RF signals are inherently high-dimensional and sensitive to small variations, which can make accurate compression and reconstruction difficult, potentially leading to mode collapse or reduced diversity in the generated data.

Future work includes improving the fidelity of generated RF data, particularly in complex environments, by enhancing the decoder to handle out-of-distribution latents more effectively. Addressing challenges associated with more complicated activities and interference-prone settings would significantly benefit RF-based human activity recognition, which suffers from a lack of diverse and high-quality data. Incorporating vision data, such as video sequences without privacy and security breaches, alongside RF signals could further enhance the capabilities by enabling multi-modal learning. Finally, optimizing training and inference time will be critical for improving the real-world deployability of the model, and adapting the diffusion framework to new domains remains a promising direction for future exploration.

VII. Conclusions

In this paper, we address the persistent challenge of data scarcity in the wireless sensing field through an AIGC-powered approach utilizing conditional latent diffusion models. The proposed RF-ACCLDM system demonstrates the capability to generate RF sensing data of exceptional quality. This AIGC framework is both efficient and versatile, enabling effective training and application across diverse RF sensing technologies and data modalities. To evaluate the quality of the generated data, we developed a comprehensive metrics system incorporating FID and diversity scores. Beyond data quality assessment, we investigated the utility of the generated data in two representative downstream tasks: HAR and RF-based 3D human pose estimation. These tasks exemplify the practical applications and broader potential of our approach. An in-depth ablation study regarding the quality and efficiency of our data generation was conducted to understand the effects of various data and training configurations. The proposed *AIGC for wireless* framework highlights the effectiveness of diffusion models in wireless sensing, offering a promising solution to critical challenges such as the high cost of RF data collection and the generation of high-fidelity RF data.

References

[1] Z. Wang and S. Mao, “AIGC for RF sensing: The case of RFID-based human activity recognition,” in *Proc. ICNC 2024*, Big Island, HI, Feb. 2024, pp. 1092–1097.

[2] Z. Wang and S. Mao, “AIGC for wireless data: The case of RFID-based human activity recognition,” in *Proc. IEEE ICC 2024*, Denver, CO, June 2024, pp. 4060–4065.

[3] Y. Sun, M. Peng, Y. Zhou, Y. Huang, and S. Mao, “Application of machine learning in wireless networks: Key technologies and open issues,” *IEEE Communications Surveys and Tutorials*, vol. 21, no. 4, pp. 3072–3108, Fourth Quarter 2019.

[4] C. Zhang, P. Patras, and H. Haddadi, “Deep learning in mobile and wireless networking: A survey,” *IEEE Commun. Surv. Tutor.*, vol. 21, no. 3, pp. 2224–2287, Thirdquarter 2019.

[5] Y. Zhang, Y. Zheng, K. Qian, G. Zhang, Y. Liu, C. Wu, and Z. Yang, “Widar3.0: Zero-effort cross-domain gesture recognition with Wi-Fi,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 11, pp. 8671–8688, Nov. 2022.

[6] C. Yang, X. Wang, and S. Mao, “Subject-adaptive skeleton tracking with rfid,” in *2020 16th International Conference on Mobility, Sensing and Networking (MSN)*, 2020, pp. 599–606.

[7] Y. Zheng, Y. Zhang, K. Qian, G. Zhang, Y. Liu, C. Wu, and Z. Yang, “Zero-effort cross-domain gesture recognition with wi-fi,” in *MobiSys ’19*, 2019, p. 313–325.

[8] J. Zhang, F. Wu, B. Wei, Q. Zhang, H. Huang, S. W. Shah, and J. Cheng, “Data augmentation and dense-LSTM for human activity recognition using WiFi signal,” *IEEE Internet of Things J.*, vol. 8, no. 6, pp. 4628–4641, Mar. 2021.

[9] Z. Wang, D. Jiang, B. Sun, and Y. Wang, “A data augmentation method for human activity recognition based on mmwave radar point cloud,” *IEEE Sensors Letters*, vol. 7, no. 5, pp. 1–4, 2023.

[10] C. Li, Z. Cao, and Y. Liu, “Deep AI enabled ubiquitous wireless sensing: A survey,” *ACM Comput. Surv.*, vol. 54, no. 2, mar 2021.

[11] Y. Tian, G. Pan, and M.-S. Alouini, “Applying deep-learning-based computer vision to wireless communications: Methodologies, opportunities, and challenges,” *IEEE Open Journal of the Communications Society*, vol. 2, pp. 132–143, 2021.

[12] J. Ho, A. Jain, and P. Abbeel, “Denoising diffusion probabilistic models,” *arXiv preprint arxiv:2006.11239*, Dec. 2020. [Online]. Available: https://arxiv.org/abs/2006.11239

[13] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Medical Image Comput. Computer-Assisted Intervention 2015*, 2015, pp. 234–241.

[14] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," in *Proc. NIPS 2017*, Long Beach, CA, Dec. 2017, pp. 6629–6640.

[15] S. Tan, Y. Ren, J. Yang, and Y. Chen, "Commodity wifi sensing in ten years: Status, challenges, and opportunities," *IEEE Internet of Things Journal*, vol. 9, no. 18, pp. 17 832–17 843, 2022.

[16] C. Yang, X. Wang, and S. Mao, "RFID-Pose: Vision-aided 3D human pose estimation with RFID," *IEEE Transactions on Reliability*, vol. 70, no. 3, pp. 1218–1231, Sept. 2021.

[17] A. Sengupta, F. Jin, R. Zhang, and S. Cao, "mm-pose: Real-time human skeletal posture estimation using mmwave radars and cnns," *IEEE Sensors Journal*, vol. 20, no. 17, pp. 10 032–10 044, 2020.

[18] W. Jiang, H. Xue, C. Miao, S. Wang, S. Lin, C. Tian, S. Murali, H. Hu, Z. Sun, and L. Su, "Towards 3D human pose construction using WiFi," in *Proc. ACM Mobicom 2020*, London, UK, Apr. 2020, pp. 1–14.

[19] X. Li, Y. Zhang, I. Marsic, A. Sarcevic, and R. S. Burd, "Deep learning for RFID-based activity recognition," in *Proc. ACM SenSys 2016*, Stanford, CA, Nov. 2016, pp. 164–175.

[20] A. D. Singh, S. S. Sandha, L. Garcia, and M. Srivastava, "RadHAR: Human activity recognition from point clouds generated through a millimeter-wave radar," in *Proc. 3rd ACM Workshop Millimeter-Wave Netw. Sensing Syst.*, Los Cabos, MX, Oct. 2019, pp. 51–56.

[21] E. Shalaby, N. ElShennawy, and A. Sarhan, "Utilizing deep learning models in CSI-based human activity recognition," *Springer Neural Comput. Appl.*, vol. 34, no. 1, pp. 5993–6010, Jan. 2022.

[22] D. Wang, J. Yang, W. Cui, L. Xie, and S. Sun, "Multimodal CSI-based human activity recognition using GANs," *IEEE Internet of Things J.*, vol. 8, no. 24, pp. 17 345–17 355, Dec. 2021.

[23] P. Liao, X. Wang, L. An, S. Mao, T. Zhao, and C. Yang, "Tfsemantic: A time-frequency semantic gan framework for imbalanced classification using radio signals," *ACM Trans. Sen. Netw.*, aug 2023.

[24] Z. Wang, C. Yang, and S. Mao, "Data augmentation for RFID-based 3D human pose tracking," in *Proc. IEEE VTC-Fall 2022*, London, UK, Sept. 2022, pp. 1–2.

[25] D. Saxena and J. Cao, "Generative adversarial networks (gans): Challenges, solutions, and future directions," *ACM Comput. Surv.*, vol. 54, no. 3, may 2021.

[26] P. Dhariwal and A. Nichol, "Diffusion models beat GANs on image synthesis," in *Proc. NeurIPS 2021*, Virtual Conference, Dec. 2021, pp. 8780–8794.

[27] J. Wu, R. Fu, H. Fang, Y. Zhang, Y. Yang, H. Xiong, H. Liu, and Y. Xu, "MedSegDiff: Medical image segmentation with diffusion probabilistic model," in *Proc. Medical Imaging with Deep Learning 2023*, Nashville, TN, Mar. 2023.

[28] C. Cao, Z.-X. Cui, S. Liu, H. Zheng, D. Liang, and Y. Zhu, "High-frequency space diffusion models for accelerated MRI," *arXiv preprint arXiv:2208.05481*, Dec. 2022. [Online]. Available: https://arxiv.org/abs/2208.05481

[29] Z. Lyu, Z. Kong, X. XU, L. Pan, and D. Lin, "A conditional point diffusion-refinement paradigm for 3D point cloud completion," in *Proc. ICLR 2022*, Virtual Conference, Apr. 2022, pp. 1–24.

[30] Y. Tashiro, J. Song, Y. Song, and S. Ermon, "CSDI: Conditional score-based diffusion models for probabilistic time series imputation," in *Proc. NeurIPS 2021*, Virtual Conference, Dec. 2021, pp. 1–13.

[31] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF CVPR 2022*, New Orleans, LA, June 2022, pp. 10 684–10 695.

[32] A. Blattmann, R. Rombach, H. Ling, T. Dockhorn, S. W. Kim, S. Fidler, and K. Kreis, "Align your latents: High-resolution video synthesis with latent diffusion models," in *Proc. IEEE/CVF CVPR 2023*, Vancouver, Canada, June 2023, pp. 22 563–22 575.

[33] T. Brooks *et al.*, "Video generation models as world simulators," 2024. [Online]. Available: https://openai.com/research/video-generation-models-as-world-simulators

[34] G. Tevet, S. Raab, B. Gordon, Y. Shafir, D. Cohen-Or, and A. H. Bermano, "Human motion diffusion model," in *Proc. ICLR 2023*, Kigali, Rwanda, May 2023, pp. 1–16.

[35] X. Chen, B. Jiang, W. Liu, Z. Huang, B. Fu, T. Chen, and G. Yu, "Executing your commands via motion diffusion in latent space," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 18 000–18 010.

[36] Y. Hu, A. Huber, and S.-C. Liu, "Overcoming the vanishing gradient problem in plain recurrent networks," 2018. [Online]. Available: https://openreview.net/forum?id=Hyp3i2xRb

[37] D. P. Kingma and M. Welling, *An Introduction to Variational Autoencoders*, 2019.

[38] C. Yang, X. Wang, and S. Mao, "TARF: Technology-agnostic RF sensing for human activity recognition," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 2, pp. 636–647, Feb. 2023.

[39] Z. Wang, C. Yang, and S. Mao, "AIGC for RF-based human activity sensing," *IEEE Internet of Things Journal*, to appear. DOI: 10.1109/JIOT.2024.3482256.

[40] X. Chen and X. Zhang, "RF Genesis: Zero-shot generalization of mmWave sensing through simulation-based data synthesis and generative diffusion models," in *Proc. ACM SenSys'23*, Istanbul, Turkiye, Nov. 2023, pp. 28–42.

[41] J. Ho and T. Salimans, "Classifier-free diffusion guidance," in *Proc. NeurIPS 2021 Workshops*, Virtual Conference, Dec. 2021, pp. 1–8.

[42] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Transa. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[43] J. Yang, X. Chen, H. Zou, C. X. Lu, D. Wang, S. Sun, and L. Xie, "Sensefi: A library and benchmark on deep-learning-empowered wifi human sensing," *Patterns*, vol. 4, no. 3, p. 100703, 2023.

Ziqi Wang [S'23] received the B.S. degree in electrical engineering from Auburn University, Auburn, AL, USA in 2022. He has been pursuing a PhD degree in the department of Electrical and Computer Engineering at Auburn University since 2023. His current research focuses on Artificial Intelligence of Things (AIot) and wireless sensing. He is a recipient of IEEE ICC 2024 NSF student travel grant, and a co-recipient of Best Demo Award of IEEE INFOCOM 2024.

Shiwen Mao [S'99-M'04-SM'09-F'19] received a Ph.D. in Electrical Engineering from Polytechnic University in 2004. He is a Professor and Earle C. Williams Eminent Scholar, and Director of the Wireless Engineering Research and Education Center at Auburn University. Dr. Mao's research interest includes wireless networks, multimedia communications, and smart grid. He is the editor-in-chief of IEEE Transactions on Cognitive Communications and Networking, and Vice President of Technical Activities of IEEE Council on Radio Frequency Identification (CRFID). He received the IEEE ComSoc MMTC Outstanding Researcher Award in 2023, the SEC 2023 Faculty Achievement Award for Auburn, the IEEE ComSoc TC-CSR Distinguished Technical Achievement Award in 2019, the Auburn University Creative Research & Scholarship Award in 2018, and the NSF CAREER Award in 2010, and several service awards from IEEE. He is a co-recipient of the 2022 Best Journal Paper Award of IEEE ComSoc eHealth Technical Committee, the 2021 Best Paper Award of Elsevier/KeAi Digital Communications and Networks Journal, the 2021 IEEE Internet of Things Journal Best Paper Award, the 2021 IEEE Communications Society Outstanding Paper Award, the IEEE Vehicular Technology Society 2020 Jack Neubauer Memorial Award, the 2018 Best Journal Paper Award and the 2017 Best Conference Paper Award from IEEE ComSoc MMTC, and the 2004 IEEE Communications Society Leonard G. Abraham Prize in the Field of Communications Systems. He is a co-recipient of 12 best conference paper/demo awards.