

RF-Diffusion: Radio Signal Generation via Time-Frequency Diffusion

Guoxuan Chi¹, Zheng Yang^{1✉}, Chenshu Wu², Jingao Xu¹, Yuchong Gao³,
Yunhao Liu¹, Tony Xiao Han⁴

¹Tsinghua University ²The University of Hong Kong

³Beijing University of Posts and Telecommunications ⁴Huawei Technologies Co., Ltd
{chiguoxuan, hmilyyz, wucs32, xujingao13, gaoyc01, yunhaoliu}@gmail.com, tony.hanxiao@huawei.com

ABSTRACT

Along with AIGC shines in CV and NLP, its potential in the wireless domain has also emerged in recent years. Yet, existing RF-oriented generative solutions are ill-suited for generating high-quality, time-series RF data due to limited representation capabilities. In this work, inspired by the stellar achievements of the diffusion model in CV and NLP, we adapt it to the RF domain and propose RF-Diffusion. To accommodate the unique characteristics of RF signals, we first introduce a novel *Time-Frequency Diffusion* theory to enhance the original diffusion model, enabling it to tap into the information within the time, frequency, and complex-valued domains of RF signals. On this basis, we propose a *Hierarchical Diffusion Transformer* to translate the theory into a practical generative DNN through elaborated design spanning network architecture, functional block, and complex-valued operator, making RF-Diffusion a versatile solution to generate diverse, high-quality, and time-series RF data. Performance comparison with three prevalent generative models demonstrates the RF-Diffusion's superior performance in synthesizing Wi-Fi and FMCW signals. We also showcase the versatility of RF-Diffusion in boosting Wi-Fi sensing systems and performing channel estimation in 5G networks.

CCS CONCEPTS

• **Human-centered computing** → Ubiquitous and mobile computing; • **Networks** → Mobile networks.

KEYWORDS

RF Signal, Generative Model, Time-Frequency Diffusion, Wireless Sensing, Channel Estimation

ACM Reference Format:

Guoxuan Chi, Zheng Yang, Chenshu Wu, Jingao Xu, Yuchong Gao, Yunhao Liu, Tony Xiao Han. 2024. RF-Diffusion: Radio Signal Generation via Time-Frequency Diffusion. In *International Conference On Mobile Computing And Networking (ACM MobiCom '24)*, September 30–October 4, 2024, Washington D.C., DC, USA. ACM, New York, NY, USA, 16 pages. <https://doi.org/10.1145/3636534.3649348>

1 INTRODUCTION

Artificial intelligence generated content (AIGC) has catalyzed a revolutionary impact in both industrial and academic frontier, birthing a constellation of cutting-edge products founded on deep neural networks (DNNs). Remarkable odysseys include Stable Diffusion [54], Midjourney [43], DALL-E [51] for image creation, and ChatGPT [46] for text generation.

Nowadays, AIGC is gradually knocking on the door of the radio-frequency (RF) domain. Current practice offers initial proof of its potential to boost wireless systems in terms of data augmentation [53], signal denoising [7] and time-series prediction [22]. In downstream tasks like device localization [76], human motion sensing [72], and channel estimation [38], such progress not only enhances system performance but also cuts down the cumbersome ground truth annotation costs for application-layer DNN training.

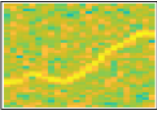
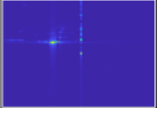
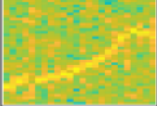
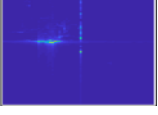
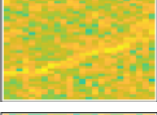

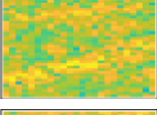
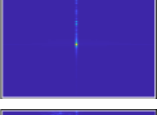
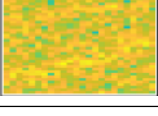
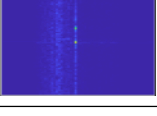
Existing RF data generation models can be broadly divided into two main categories:

• **Environment modeling based generative model.** This approach exploits LiDAR point clouds or video footage to craft a detailed 3D model of the environment. It then employs physical models, like ray tracing [42], to simulate how RF signals interact with surroundings, which eventually aids in forecasting the signals a receiver might capture. However, one notable limitation is the method's insufficient consideration of how the materials and properties of targets can affect RF signal propagation. Additionally, obtaining a 3D model with accuracy compatible with RF signal wavelengths (e.g., 1-10 mm) remains a challenge and will significantly raise system expenses. While the recent study uses the neural radiance field for implicit modeling of RF-complex environments to estimate signal propagation [76], it requires a stationary receiver (Rx), which complicates generating essential time-series data for wireless communication systems or tasks like human motion recognition.

• **Data-driven probabilistic generative model.** Current innovations leverage models like generative adversarial network (GAN) and variational autoencoder (VAE) to augment RF datasets [21]. Essentially, these models learn the distribution within the training data and then generate new RF data

✉ Zheng Yang is the corresponding author. Our project is available [here](#).

Table 1: Illustrative examples.

Methods	Examples		SSIM	
	Wi-Fi	FMCW	Wi-Fi	FMCW
Ground Truth			N/A	N/A
Ours			0.81	0.75
DDPM [25]			0.65	0.58
DCGAN [50]			0.68	0.61
CVAE [61]			0.47	0.4

that follow this distribution. However, these models mainly focus on expanding feature-level distributions and struggle to precisely generate raw RF signals due to their constrained representation capabilities [72]. Additionally, most of them are designed for specific tasks with dedicated loss functions and DNN architectures, limiting their versatility. On the other hand, GAN's training is notoriously fickle due to the tug-of-war between the generator and discriminator [69].

Remark. Albeit inspiring, there still lacks a versatile generative model for generating accurate and time-series raw RF signals suitable for diverse downstream applications.

Recently, *Diffusion Model* has emerged as a luminous star in the visual AIGC cosmos, underpinning a variety of innovative DNNs for a range of prominent image/video applications such as Stable Diffusion, Midjourney, and DALL-E. Compared to the aforementioned generative models, its unique iterative process of noise addition (i.e., noising) and removal (i.e., denoising) allows for precise capture of intricate raw data distributions [70]. Moreover, its training is straightforward and avoids typical problems like mode collapse or convergence troubles, since it doesn't juggle competing parts or require delicate fine-tuning [13].

These compelling advantages inspire us to embrace the diffusion model for synthesizing RF data. However, transferring existing diffusion models [25] to the RF domain faces significant challenges arising from RF signal's unique characteristics beyond images, as summarized below.

(i) *Time series.* RF signals capture dynamic details like target movement and environment/channel changes over time,

unlike static snapshots. Diffusion models designed for single-image generation struggle to synthesize RF signal sequences. (ii) *Frequency domain.* Essential RF details (e.g., Doppler shift, chirp) are embedded in the frequency domain. While recent video diffusion models can create time series, they mainly focus on the spatial domain (e.g., pixel-wise brightness), discarding the rich information in the frequency domain.

(iii) *Number field.* RF data is complex-valued with both amplitude and phase readings. While existing diffusion models only focus on amplitude (e.g., light strength), the phase data can't be ignored due to its crucial role in wireless systems. In summary, while diffusion models hold great promise, there is a need to upgrade current models to suit the unique traits of RF signals and tap into the underlying information in the time series, frequency, and complex-valued domains.

Our Work. We propose RF-Diffusion, the first versatile generative model for RF signals based on **Diffusion** model. To overcome the above challenges, we expand existing denoising-based diffusion model to the time-frequency domain by revisiting its theoretical foundation, overall DNN architecture, and detailed operator design, enabling RF-Diffusion to generate diverse, high-quality, and time-series RF data.

• **Time-Frequency Diffusion Theory.** We first propose the *time-frequency diffusion* (TFD) theory as a novel paradigm to guide diffusion models in extracting and leveraging characteristics of RF signals across both temporal and frequency domains. Specifically, we demonstrate a diffusion model could effectively destruct and restore high-quality RF signals by alternating between adding noise in the time domain and blurring in the frequency domain (§3).

• **Hierarchical Diffusion Transformer Design.** We further re-design the DNNs of existing denoising-based diffusion model to be compatible with TFD. The derived DNN, designated as the *hierarchical diffusion transformer* (HDT), from a top-down perspective, incorporates (i) a hierarchical architecture to fully uncover time-frequency details by decoupling spatio-temporal dimensions of RF data; (ii) attention-based diffusion blocks leveraging enhanced Transformers to extract RF features; and (iii) a complex-valued design to encode both signal strength and phase information. The three key designs work hand-in-hand to enable RF-Diffusion to generate high-quality RF data (§4).

We implement RF-Diffusion and conduct extensive experiments that include synthesis of both Wi-Fi and FMCW signals. To provide a clear understanding of its performance, an intuitive comparison of the time-frequency spectrograms generated by RF-Diffusion and those from related works is presented in Table 1. Evaluation results demonstrate that RF-Diffusion generates RF signals with high fidelity, achieving an average structural similarity of 81% relative to the ground

truth. This performance surpasses prevalent generative models such as DDPM, DCGAN, and CVAE by over 18.6%. We also demonstrate the performance of RF-Diffusion in two case studies: augmented Wi-Fi gesture recognition and 5G FDD channel estimation. By employing RF-Diffusion as a data augmentor, existing wireless gesture recognition systems experience a significant accuracy improvement ranging from 4% to 11%. When applied to the channel estimation task, RF-Diffusion showcases a substantial 5.97 dB improvement in SNR compared to state-of-the-arts.

In summary, this paper makes the following contributions.

- (1) We propose RF-Diffusion, the first generative diffusion model tailored for RF signal. RF-Diffusion is versatile and can be leveraged in a wide spectrum of fundamental wireless tasks such as RF data augmentation, channel estimation, and signal denoising, propelling AIGC to shine in the RF domain.
- (2) We present the *Time-Frequency Diffusion* theory, an advanced evolution beyond traditional denoising-based diffusion methods. The integration of TFD with its bespoke *Hierarchical Diffusion Transformer* (HDT) enables enhanced precision in time-series sampling and a balanced focus on spectral details of the data.
- (3) We fully implement RF-Diffusion. Extensive evaluation results from case studies show RF-Diffusion's efficacy.

Community Contribution. RF-Diffusion's code and pre-trained model are publicly available. Our solution, in part or in all, could provide a collection of tools for both industry and academia to push forward AIGC in RF domain. Moreover, its ability to handle time-series sampling while highlighting the spectral nuances of the data has potential benefits beyond the wireless community, offering value to video, audio processing, and other time-series-dependent modalities.

2 OVERVIEW

We propose RF-Diffusion, a pioneering probabilistic generative model for RF data that leverages the diffusion model framework, as detailed in Fig. 1. At its core, RF-Diffusion aligns with the principle of denoising-based diffusion models by employing a dual-process approach: a forward process of integrating noise into the data, and a reverse process of generating data from noise. However, RF-Diffusion distinguishes itself through two innovative features:

- (i) RF-Diffusion incorporates the proposed Time-Frequency Diffusion (§3) theory to direct each stage of state transition in both forward (i.e., $q(x_t|x_{t-1})$) and reverse (i.e., $p_\theta(x_{t-1}|x_t)$) processes, enabling RF-Diffusion to harness the RF signal information across both the time and frequency domain.
- (ii) RF-Diffusion introduces the Hierarchical Diffusion Transformer (§4), which is a restructured DNN model for the reverse generation process, to align with the Time-Frequency Diffusion theory and the characteristics of RF signals.

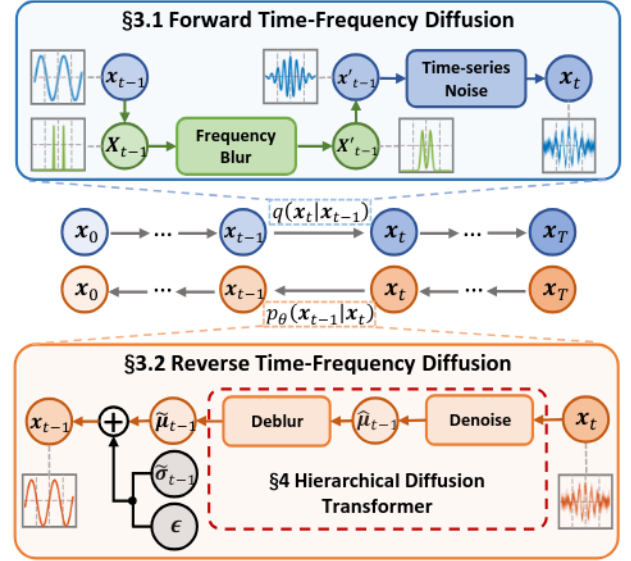


Figure 1: RF-Diffusion overview.

As for the specific data flow, RF-Diffusion gradually introduces Gaussian noise in the time domain and blurs the spectrum in the frequency domain at each stage in the forward direction. As the diffusion step t advances, the original RF signal x_0 diminishes, eventually degrading into noise. In TFD theory, we demonstrate any destructed signal x_t can be restored to its original form x_0 using a parameterized reverse process. Guided by the destruction process alternating in time-frequency domain, the reverse restoration process emphasizes both time-domain amplitude accuracy and frequency-domain continuity to achieve time-frequency high-fidelity signal generation.

In the reverse direction, HDT are served as the parameterized model for learning the restoration process. It decouples the Gaussian noise and the spectral blur, effectively addressing them in the spatial denoise and time-frequency deblur stages, respectively. During its training, HDT takes destructed signal x_t as the model input, and uses the signal of previous diffusion step x_{t-1} to supervise the output. Once trained, RF-Diffusion is capable of iteratively transforming fully degraded noise back into a specific type of signal.

3 TIME-FREQUENCY DIFFUSION

In this section, we introduce the proposed Time-Frequency Diffusion (TFD) process. Unlike prevailing denoising diffusion models, the time-frequency diffusion process comprehensively addresses two potential distortions in wireless signal data: 1) amplitude distortion due to additive Gaussian noise; 2) spectral aliasing resulting from insufficient time resolution. Therefore, the learned reverse process focuses not only on precisely reconstructing the amplitude of individual samples but also on preserving spectral resolution in

time-series signals. In what follows, we first introduce the forward destruction process (§3.1) which jointly eliminates the original data distribution in the time and the frequency domain. On this basis, we describe how to reverse the process (§3.2) and fit it through a parameterized model, which is the basis of our conditional generation (§3.3) task.

3.1 Forward Destruction Process

The time-frequency diffusion model is proposed for the RF signal, which can be treated as the complex-valued time-series data. Therefore, we take the signal as a two-dimensional complex tensor $\mathbf{x} \in \mathbb{C}^{M \times N}$, where M represents the spatial dimension of each sample, while N represents the temporal dimension of the times series.

Given a signal that follows a specific distribution $\mathbf{x}_0 \sim q(\mathbf{x}_0)$, the forward destruction process yields a progression of random variables $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_T$. Each diffusion step in this process disrupts the original distribution from both the time and frequency domains. Specifically, the forward diffusion process from step $t-1$ to t is described as follows:

- **Frequency Blur.** To dissipate the spectral details of the original signal, the Fourier transform $\mathfrak{F}(\cdot)$ is first performed to the temporal dimension. Subsequently, with the predefined Gaussian convolution kernel \mathbf{G}_t , a cyclic convolution $*$ operation is performed on the spectrum, resulting in a blurred spectrum $\mathbf{G}_t * \mathfrak{F}(\mathbf{x}_{t-1})$.
- **Time-series Noise.** To drown out the amplitude details of the signal, complex standard Gaussian noise $\epsilon \sim \mathcal{CN}(0, \mathbf{I})$ is introduced, and a weighted summation is performed with a predefined parameter $\sqrt{\alpha_t}$, where $\alpha_t \in (0, 1)$.

By combining the above two steps, we get:

$$\mathbf{x}_t = \sqrt{\alpha_t} \mathfrak{F}^{-1}(\mathbf{G}_t * \mathfrak{F}(\mathbf{x}_{t-1})) + \sqrt{1 - \alpha_t} \epsilon, \quad (1)$$

where $\mathfrak{F}^{-1}(\cdot)$ indicates the inverse Fourier transform.

To ensure the practical feasibility of the time-frequency diffusion process, it is essential that the transition from \mathbf{x}_0 to \mathbf{x}_t for any given step $t \in [1, T]$ can be executed with an acceptable time complexity, instead of involving an iteration of t steps. To simplify this process, certain advantageous characteristics of the Fourier transform and the Gaussian function are leveraged. Based on the convolution theorem [68], we have $\mathfrak{F}^{-1}(\mathbf{G}_t * \mathfrak{F}(\mathbf{x}_{t-1})) = \mathfrak{F}^{-1}(\mathbf{G}_t) \mathbf{x}_{t-1}$ ¹. Therefore, the operation in Eqn. 1 can be expressed as:

$$\mathbf{x}_t = \sqrt{\alpha_t} \mathbf{g}_t \mathbf{x}_{t-1} + \sqrt{1 - \alpha_t} \epsilon, \quad (2)$$

where $\mathbf{g}_t = \mathfrak{F}^{-1}(\mathbf{G}_t)$ is still a Gaussian kernel, which means the convolution of the signal with the Gaussian kernel in the frequency domain can be equivalently transformed into the multiplication of the signal with another Gaussian kernel in the time domain. For ease of notion, let $\mathbf{y}_t = \sqrt{\alpha_t} \mathbf{g}_t$, and

$\sigma_t = \sqrt{1 - \alpha_t}$, indicating the weight of the signal \mathbf{x}_{t-1} and the standard deviation of the added noise at step t .

Since the forward process is a Markov chain, by recursively applying Eqn. 2 and incorporating with the reparametrization trick [34], the relationship between the original signal \mathbf{x}_0 and the degraded signal \mathbf{x}_t can be obtained:

$$\mathbf{x}_t = \bar{\mathbf{y}}_t \mathbf{x}_0 + \sum_{s=1}^t (\sqrt{1 - \alpha_s} \frac{\bar{\mathbf{y}}_t}{\bar{\mathbf{y}}_s}) \epsilon = \bar{\mathbf{y}}_t \mathbf{x}_0 + \bar{\sigma}_t \epsilon, \quad (3)$$

where $\bar{\mathbf{y}}_t = \prod_{s=1}^t \mathbf{y}_s = \mathbf{y}_t \cdots \mathbf{y}_1$. As α_t and \mathbf{g}_t are predefined hyperparameters corresponding to the noise and blur scheduling strategy, any $\bar{\mathbf{y}}_t$ and $\bar{\sigma}_t$ are constant coefficients, representing the weight of the original signal and the standard deviation of the added noise. Thus, the forward destruction process to any step t can be quickly completed without iteration. Stated in probabilistic terms, essentially \mathbf{x}_t follows an non-isotropic Gaussian distribution conditioned on \mathbf{x}_0 :

$$q(\mathbf{x}_t | \mathbf{x}_0) = \mathcal{CN}(\mathbf{x}_0; \bar{\boldsymbol{\mu}}_t, \bar{\boldsymbol{\sigma}}_t^2 \mathbf{I}), \quad (4)$$

where $\bar{\boldsymbol{\mu}}_t = \bar{\mathbf{y}}_t \mathbf{x}_0$ and $\bar{\sigma}_t = \sum_{s=1}^t (\sqrt{1 - \alpha_s} \frac{\bar{\mathbf{y}}_t}{\bar{\mathbf{y}}_s})$. Specifically, the vector $\bar{\mathbf{y}}_t$ consists of distinct weighting coefficients, each applied multiplicatively across the temporal dimension of the original signal to perform weighting adjustments.

It is proven in Appendix A that as the diffusion step t increases, the original signal is gradually eliminated, and \mathbf{x}_t eventually converges to a closed-form noise distribution:

$$\lim_{T \rightarrow \infty} \mathbf{x}_T = \lim_{T \rightarrow \infty} \sum_{t=1}^T (\sqrt{1 - \alpha_t} \frac{\bar{\mathbf{y}}_T}{\bar{\mathbf{y}}_t}) \epsilon = \lim_{T \rightarrow \infty} \bar{\sigma}_T \epsilon, \quad (5)$$

where $\bar{\sigma}_T = \sum_{t=1}^T (\sqrt{1 - \alpha_t} \frac{\bar{\mathbf{y}}_T}{\bar{\mathbf{y}}_t}) \epsilon$ is determined by predefined noise scheduling strategy in practical implementation.

3.2 Reverse Restoration Process

The restoration process is the reversal of the destruction, which gradually eliminates the noise and restores the original data distribution.

To learn a parameterized distribution $p_\theta(\mathbf{x}_0)$ which approximates the original distribution $q(\mathbf{x}_0)$, an effective approach is to minimize their Kullback-Leibler (KL) divergence:

$$\begin{aligned} \theta &= \arg \min_{\theta} D_{\text{KL}}(q(\mathbf{x}_0) \| p_\theta(\mathbf{x}_0)) \\ &= \arg \min_{\theta} (\mathbb{E}_{q(\mathbf{x}_0)} [-\log p_\theta(\mathbf{x}_0)] + \mathbb{E}_{q(\mathbf{x}_0)} [\log q(\mathbf{x}_0)]) \\ &= \arg \max_{\theta} \mathbb{E}_{q(\mathbf{x}_0)} [\log p_\theta(\mathbf{x}_0)]. \end{aligned} \quad (6)$$

Unfortunately, $q(\mathbf{x}_0)$ is intractable to calculate in general [35, 60], therefore $\mathbb{E}_{q(\mathbf{x}_0)} [\log p_\theta(\mathbf{x}_0)]$ cannot be expressed explicitly. Building on the concepts of prior works [25, 63], we approximate the distribution by maximizing the variational

¹Vector multiplications in this paper default to element-wise products.

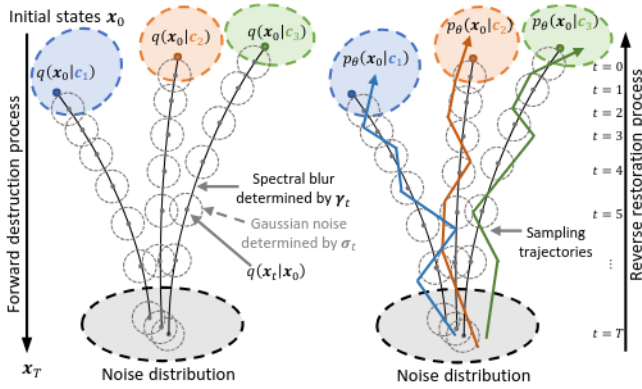


Figure 2: Illustration of the conditional forward and reverse trajectories.

lower bound. As established in [25], the optimization problem in Eqn. 6 can be approximated as:

$$\theta = \arg \min_{\theta} D_{\text{KL}}(q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) \| p_{\theta}(\mathbf{x}_{t-1}|\mathbf{x}_t)), \quad (7)$$

where $q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)$ represents the actual reverse process conditioned on \mathbf{x}_0 , while $p_{\theta}(\mathbf{x}_{t-1}|\mathbf{x}_t)$ denotes the reverse process fitted by our model. Eqn. 7 shows the problem of reconstructing the original data distribution can be transformed into a problem of fitting the reverse process. Rewrite $q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)$ based on the Bayesian theorem (Appendix B), and we prove it follows a Gaussian distribution over \mathbf{x}_{t-1} :

$$q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) \sim \mathcal{CN}(\mathbf{x}_{t-1}; \tilde{\boldsymbol{\mu}}_{t-1}, \tilde{\boldsymbol{\sigma}}_{t-1}^2 \mathbf{I}), \quad (8)$$

$$\tilde{\boldsymbol{\mu}}_{t-1} = \frac{1}{\tilde{\boldsymbol{\sigma}}_t^2} (\boldsymbol{\gamma}_t \tilde{\boldsymbol{\sigma}}_{t-1}^2 \mathbf{x}_t + \tilde{\boldsymbol{\gamma}}_{t-1} \boldsymbol{\sigma}_t^2 \mathbf{x}_0), \quad \tilde{\boldsymbol{\sigma}}_{t-1} = \frac{\tilde{\boldsymbol{\sigma}}_{t-1}}{\tilde{\boldsymbol{\sigma}}_t} \boldsymbol{\sigma}_t.$$

Let's assume that $p_{\theta}(\mathbf{x}_{t-1}|\mathbf{x}_t)$ is a Gaussian Markov process:

$$p_{\theta}(\mathbf{x}_{t-1}|\mathbf{x}_t) \sim \mathcal{CN}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_{\theta}(\mathbf{x}_t), \boldsymbol{\sigma}_{\theta}^2(\mathbf{x}_t) \mathbf{I}). \quad (9)$$

Therefore, the KL divergence of two Gaussian distributions in Eqn. 7 can be simplified as follows:

$$D_{\text{KL}}(q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) \| p_{\theta}(\mathbf{x}_{t-1}|\mathbf{x}_t)) = \mathbb{E}_{q(\mathbf{x}_0)} \left[\frac{1}{2\tilde{\boldsymbol{\sigma}}_t^2} \|\tilde{\boldsymbol{\mu}}_{t-1} - \boldsymbol{\mu}_{\theta}(\mathbf{x}_t)\|^2 \right] + C. \quad (10)$$

In summary, the optimization of the parameterized model $p_{\theta}(\mathbf{x}_{t-1}|\mathbf{x}_t)$ can be achieved by minimizing the mean square error (MSE) between $\boldsymbol{\mu}_{\theta}$ and $\tilde{\boldsymbol{\mu}}_{t-1}$. In other words, if a model can infer the mean value $\tilde{\boldsymbol{\mu}}_{t-1}$ of the previous step from the input \mathbf{x}_t of the current diffusion step, then it is competent for the data generation task.

3.3 Conditional Generation

In most practical applications, the generation process is expected to be guided by the condition label \mathbf{c} , which indicates a specific type of the generated signal (e.g., the signal corresponding to a specific device location or human activity).

Incorporating the conditional generation mechanism into RF-Diffusion offers significant advantages: (i) *Enhanced practicality*. The conditional generation mechanism enables RF-Diffusion system to generate signals of different categories based on various condition combinations. This eliminates the need for training separate models for each signal type, significantly improving the model's utility in practical applications. (ii) *Increased signal diversity*. A well-trained conditional generation model creates diverse samples featuring any conceivable combination of characteristics within the condition-label space of the training dataset, which extends the model's generalizability beyond the initial scope of the training set, ensuring that data augmentation contributes to performance improvements in downstream tasks.

In this context, the condition input \mathbf{c} defines specific scenarios, including various rooms, Tx-Rx deployments, human activity types, and signal bandwidths. This input guides the generation process to produce data that aligns with the conditional distribution $p_{\theta}(\mathbf{x}|\mathbf{c})$. An illustration of the conditional forward and reverse processes is presented in Fig. 2.

Following the conclusion of previous work [13, 27, 61], we directly incorporate the condition \mathbf{c} in both the forward process Eqn. 4 and the reverse process Eqn. 8, and get $q(\mathbf{x}_t|\mathbf{x}_0, \mathbf{c})$ and $q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0, \mathbf{c})$ respectively. Then, by combining Eqn. 7 and Eqn. 10, the optimization can be written as:

$$\begin{aligned} \theta &= \arg \min_{\theta} D_{\text{KL}}(q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0, \mathbf{c}) \| p_{\theta}(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{c})) \\ &= \arg \min_{\theta} \mathbb{E}_{q(\mathbf{x}_0)} [\|\tilde{\boldsymbol{\mu}}_{t-1} - \boldsymbol{\mu}_{\theta}(\mathbf{x}_t(\mathbf{x}_0, t, \boldsymbol{\epsilon}), \mathbf{c})\|^2]. \end{aligned} \quad (11)$$

Algorithm 1 RF-Diffusion Training.

Input: Dataset following $\mathbf{x} \sim q(\mathbf{x})$ with condition \mathbf{c}

Output: Trained model μ_{θ}

- 1: Set hyperparameters T , α_t and \mathbf{g}_t
 - 2: **while** μ_{θ} not converged **do**
 - 3: Sample $\mathbf{x}_0 \sim q(\mathbf{x}_0)$ with condition \mathbf{c} from dataset
 - 4: Sample diffusion step $t \in \text{Uniform}(1, \dots, T)$
 - 5: Sample noise $\boldsymbol{\epsilon} \sim \mathcal{CN}(0, \mathbf{I})$
 - 6: Get $\mathbf{x}_t = \tilde{\boldsymbol{\gamma}}_t \mathbf{x}_0 + \tilde{\boldsymbol{\sigma}}_t \boldsymbol{\epsilon}$ ▷ Eqn. 3
 - 7: Calculate $\tilde{\boldsymbol{\mu}}_{t-1}$ based on \mathbf{x}_0 and \mathbf{x}_t ▷ Eqn. 8
 - 8: Minimize $\|\tilde{\boldsymbol{\mu}}_{t-1} - \boldsymbol{\mu}_{\theta}(\mathbf{x}_t(\mathbf{x}_0, t, \boldsymbol{\epsilon}), \mathbf{c})\|^2$ ▷ Eqn. 11
 - 9: **end while**
-

The training process of the parameterized model used for restoration is summarized Algorithm 1. By incorporating the desired signal type as a conditional input, the trained model can iteratively synthesize the original signal from a sampled noise. The generative process is illustrated in Algorithm 2.

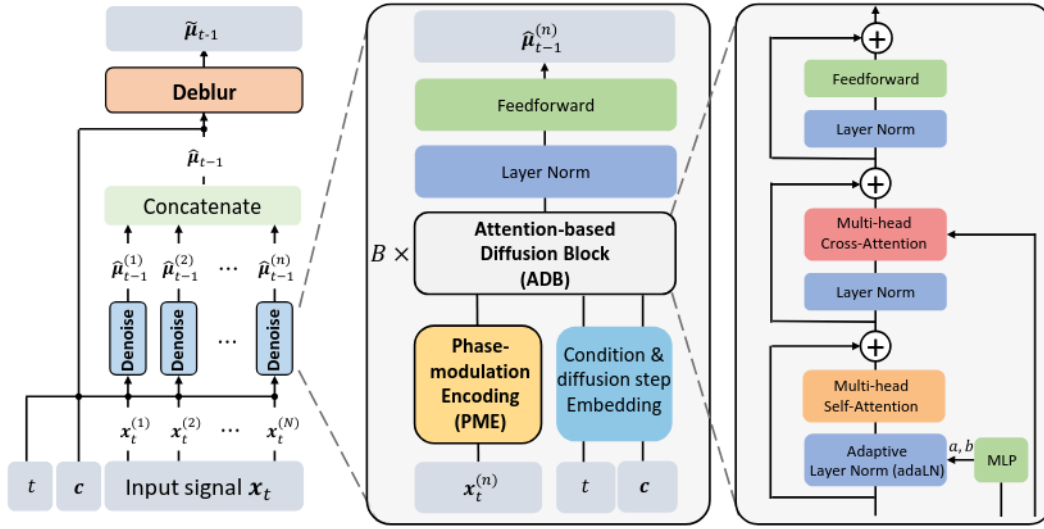


Figure 3: Hierarchical Diffusion Transformer design.

Algorithm 2 RF-Diffusion Sampling.**Input:** Trained model μ_θ , condition c **Output:** Generated sample x_0

- 1: Set hyperparameters T , α_t and g_t
- 2: Sample noise $\epsilon \sim \mathcal{CN}(0, I)$
- 3: Let $x_T = \bar{\sigma}_T \epsilon$ ▷ Eqn. 5
- 4: **for** $t = T, \dots, 1$ **do**
- 5: Get model output $\mu_\theta(x_t, c)$, and let $\sigma_\theta = \bar{\sigma}_{t-1}$
- 6: Sample $x_{t-1} \sim p_\theta(x_{t-1}|x_t)$ with μ_θ and σ_θ , which means let $x_{t-1} = \mu_\theta(x_t, c) + \sigma_\theta \epsilon$ ▷ Eqn. 9
- 7: **end for**
- 8: **return** x_0

4 HIERARCHICAL DIFFUSION TRANSFORMER

To bridge the gap between time-frequency theory and a practical generative model, we introduce a hierarchical diffusion transformer (HDT). Our proposed HDT incorporates many innovative designs, aligning it with the underlying time-frequency diffusion theory and making it adept for RF signal generation. We first introduce the overarching hierarchical design (§4.1), followed by the detailed design of our proposed attention-based diffusion block (ADB) (§4.2). Addressing the challenge of complex-valued signal generation, we extend the core design of the classic transformer block [66] into the complex domain (§4.3). Moreover, we propose phase modulation encoding (PME) (§4.4), a novel positional encoding approach tailored for complex-valued neural networks.

4.1 Hierarchical Architecture

From the top perspective, HDT adopts a hierarchical architecture to efficiently decouples the estimation of non-isotropic

noise. As shown in Fig. 3, HDT is divided into two stages: spatial denoising and time-frequency deblurring.

The diffusion step, denoted as t , is encoded, thereby informing the model about the current input's diffusion level. The conditional vector c undergoes encoding as well. In conjunction with the input $x_t^{(n)}$, these components engage in computations, striving to discern the latent correlation between the input and its pertinent condition.

Our observation is that the non-isotropic noise can be dissected into two components: 1) Independent Gaussian noise ϵ across both the spatial dimension M and the temporal dimension N . 2) Different information and noise weights (i.e., $\bar{y}_t^{(n)}$ and $\bar{\sigma}_t^{(n)}$) along the temporal dimension N . Therefore, by splitting the time-series data into separate samples, we get $x_t^{(n)} = \bar{y}_t^{(n)} x_0^{(n)} + \bar{\sigma}_t^{(n)} \epsilon^{(n)}$. Herein, within each sample, both signal and noise weights remain constant. Therefore, each spatial denoising module processes a single sample $x_t^{(n)}$ of the input sequence independently. During this stage, denoising circumvents the temporal domain weighting induced by spectral blurring, focusing exclusively on the Gaussian noise $\epsilon^{(n)}$ introduced into the original information. This approach resonates with the principles of denoising diffusion [25].

Although the spatial denoising module effectively mitigates the impact of noise ϵ , its individual treatment for each sample disregards the temporal weighting effects originating from spectral blurring. Therefore, the processed results are concatenated as $\hat{\mu} = [\hat{\mu}^{(1)}, \dots, \hat{\mu}^{(N)}]$ and serve as sequence input for the time-frequency deblurring module, aiming to estimate the mean value $\tilde{\mu}_{t-1}$.

4.2 Attention-based Diffusion Block

As shown in Fig. 3, the input data are process by a sequence of transformer blocks in both the denoising and deblur stage.

We introduce an innovative attention-based diffusion block to jointly analyze the noisy input \mathbf{x}_t , condition \mathbf{c} , and step t .

Self-attention for feature extraction. The multi-head self-attention module captures autocorrelation feature from the noisy input and extracts the high-level representations implicit in the signal. Compared to convolutional layers with translation invariance, attention layers are sensitive to the positional information of each sample in the sequence, thus enabling more effective restoration of the original signal.

Cross-attention for conditioning. To enhance the conditional generation capability, RF-Diffusion incorporates a cross-attention module to learn the latent associations between the inputs and their corresponding conditions. This module is designed to directly capture the intricate dynamics between the inputs and specified conditions, thereby improving the diversity and fidelity of generated signals.

Adaptive layer normalization for diffusion embedding. Inspired by the widespread usage of adaptive normalization layer (adaLN) [47] in existing conditional generative models [9, 13], we explore replacing standard layer normalization with adaLN. Rather than directly learn dimension-wise scale a and shift parameters b , we regress them from the t , embedding the diffusion step information into our model.

4.3 Complex-Valued Module Design

In order to work effectively with complex-valued wireless signals, the RF-Diffusion model is designed as a complex-valued neural network. Several adaptations have been made to HDT to facilitate complex-valued operations.

Complex-valued attention module. Two key improvements have been implemented in the dot-product attention mechanism to accommodate complex number computation: 1) The dot product of the query and key vectors is extended to the hermitian inner product, $\mathbf{q}^H \mathbf{k}$, which captures the correlation of two vectors in the complex space. This preserves the effective information of both the real and imaginary parts to the fullest extent. 2) Given that the softmax function operates on real numbers, adjustments have been made to make it compatible with complex vectors. Specifically, softmax is applied to the magnitude of the dot product, while the phase information remains unchanged. This modification maintains the probabilistic interpretation of vector relevance. In mathematical terms, the complex-valued attention computation for complex vectors \mathbf{q} and \mathbf{k} can be expressed as:

$$\text{softmax}(|\mathbf{q}^H \mathbf{k}|) \exp(j \angle(\mathbf{q}^H \mathbf{k})). \quad (12)$$

Complex-valued feed-forward module. Feed-forward module consists of two main operations: linear transformation and non-linear activation. A complex-valued linear transformation can be decomposed into real-valued ones [64]. Specifically, for complex-valued input $\mathbf{x} = \mathbf{x}_r + j\mathbf{x}_i$, the transformation with complex weight $\mathbf{w} = \mathbf{w}_r + j\mathbf{w}_i$ and bias

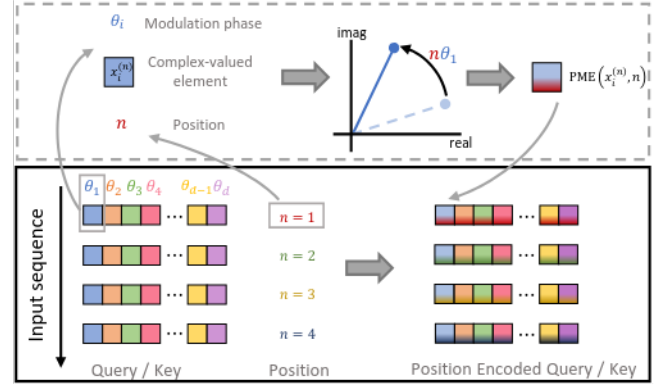


Figure 4: Illustration of phase modulation encoding.

$\mathbf{b} = \mathbf{b}_r + j\mathbf{b}_i$ can be written as follows:

$$\mathbf{w}\mathbf{x} + \mathbf{b} = \begin{bmatrix} \Re(\mathbf{w}\mathbf{x} + \mathbf{b}) \\ \Im(\mathbf{w}\mathbf{x} + \mathbf{b}) \end{bmatrix} = \begin{bmatrix} \mathbf{w}_r & -\mathbf{w}_i \\ \mathbf{w}_r & \mathbf{w}_i \end{bmatrix} \begin{bmatrix} \mathbf{x}_r \\ \mathbf{x}_i \end{bmatrix} + \begin{bmatrix} \mathbf{b}_r \\ \mathbf{b}_i \end{bmatrix}. \quad (13)$$

Furthermore, applying an activation function $g(\cdot)$ to a complex value can be seen as activating the real and imaginary parts separately: $g(\mathbf{x}) = g(\mathbf{x}_r) + jg(\mathbf{x}_i)$.

4.4 Phase Modulation Encoding

Leveraging the attention mechanism, a Transformer network parallelly processes the entire sequence. Yet, it lacks inherent capability to discern the positional information of the input. To address this, we introduce an innovative phase modulation encoding (PME) strategy tailored for complex spaces, serving as the positional encoding scheme for HDT.

As illustrated in Fig. 4, suppose the maximum dimension of each vector in the sequence is d . For the i -th element of the n -th vector in the sequence, PME operates as follows:

$$\text{PME}(\mathbf{x}^{(n)}(i), n) = \mathbf{x}_i^{(n)} \exp(jn\theta_i), \quad (14)$$

where θ_i is given by $\theta_i = 10000^{-\frac{i}{d}}$. This procedure can be conceptualized as a phase modulation process—essentially imparting a specific phase offset to the original data based on the position n in the sequence.

The PME inherently decodes the relative position during computation, establishing its essential role in position encoding. Specifically, when executing the complex-domain Attention operation on the encoded key vector \mathbf{k} and query vector \mathbf{q} , it is equivalent to:

$$\text{PME}(\mathbf{q}, n)^H \text{PME}(\mathbf{k}, m) = \text{PME}(\mathbf{q}^H \mathbf{k}, m - n). \quad (15)$$

Therefore, the relative position information $m - n$ can be derived. This enables our model to learn more proficiently by integrating the positional details of the sequence.

5 IMPLEMENTATION

We implement RF-Diffusion based on PyTorch and train our model on 8 NVIDIA GeForce 3090 GPUs, incorporating essential implementation techniques outlined below.

Exponential moving average. Following common practice in most generative models, we adopt the exponential moving average (EMA) mechanism with a decay rate of 0.999. EMA calculates a sliding average of the model's weights during training, which improves model robustness.

Weight initialization. We zero-initialize each final layer before the residual to accelerate large-scale training [20], and apply Xavier uniform initialization [18] to other layers, which is a standard weight initialization technique in transformer-based models [15].

Hyperparameters. We train our model using AdamW optimizer [33, 39] with an initial learning rate of 1×10^{-3} . A step learning rate scheduler with a decay factor of 0.5 is adopted to improve training efficiency. In the training process, we apply a dropout rate of 0.1 to mitigate overfitting.

Noise scheduling strategy. In our implementation, the rate of data destruction is designed to increase incrementally from a lower to a higher intensity as the diffusion process progresses. This is aimed at achieving a balance between the model complexity and the generation quality [35]. Specifically, we configure the diffusion process with a maximum of $T = 300$ steps. The noise coefficient, $\beta_t = \sqrt{1 - \alpha_t}$, is set to linearly increase from 10^{-4} to 0.03, i.e., $\beta_t = 10^{-4}t$. In parallel, the standard deviation of the Gaussian convolution kernel in the frequency domain, denoted as G_t , is adjusted to linearly escalate from 10^{-3} to 0.3, facilitating the controlled amplification of noise across the diffusion steps.

Data preprocessing. Each signal sequence from the dataset is either interpolated or downsampled to a consistent length of 512. This guarantees uniformity in the model's input length. Prior to input into our model, each sample within the input sequence is normalized by average signal power, which means each sample is divided by the average L2-norm of all the samples in the sequence.

6 EVALUATION

6.1 Experiment Design

6.1.1 Data Collection. As shown in Fig. 5, our dataset comprises wireless signals collected under three distinct scenarios, featuring variations in room selection, device location, and human factors, including their location, orientation, and activities. We compile condition labels for each sequence into a conditional vector \mathbf{c} , guiding both training and sampling phases. Our research evaluates RF-Diffusion's proficiency in producing signals across different modulation modes, focusing on Wi-Fi and FMCW radar signals as two primary types of wireless sensing and communications.

- **Wi-Fi.** We collect Wi-Fi signal based on the commercial NIC IWL5300 working in 5.825 GHz with 40 MHz bandwidth. The transmitter injects Wi-Fi packets to 3 receivers to extract the channel state information (CSI) corresponding to the environment.
- **FMCW.** FMCW signals are recorded using the mmWave radar IWR1443 [29]. This radar device can be placed at either one of two different locations in each scene, working at a frequency band from 77 GHz to 81 GHz.

More than 20,000 Wi-Fi sequences and 13,000 FMCW sequences are collected. Each sequence has an associated condition label indicating the room, device placement, human ID, location, orientation and activity type. All experiments conducted in this paper conform to the IRB policies.

6.1.2 Comparative Methods. We compare RF-Diffusion with three representative data generation model:

- **DDPM** [25]. The denoising diffusion probabilistic model (DDPM) introduces Gaussian noise to original data and subsequently learns to reverse this process, thereby generating raw data from the noise.
- **DCGAN** [50]. The deep convolutional generative adversarial network (DCGAN) stands as a widely recognized GAN. In DCGAN, two models (i.e., generator and discriminator) are simultaneously trained in an adversarial manner. Once trained, the generator can produce data that convincingly bypasses the discriminator's scrutiny.
- **CVAE** [62]. The conditional variational autoencoder (CVAE) learns the Gaussian implicit representation of the data, thereby enabling data generation. This method is widely adopted in both sensing [21] and communication [38] systems to synthesize wireless features.

To adapt them for RF signal, we have re-implemented the model using complex-valued neural networks [64].

6.1.3 Evaluation Metrics. For a comprehensive evaluation, we adopt two metrics, both of which are commonly used in previous research for evaluating data-driven generative models [2, 13, 25, 30, 44]. Recognizing that a definitive "gold standard" for generative models has not been established, these metrics are among the most authoritative.

- **SSIM** [67]: The Structural Similarity Index Measure (SSIM) is a prominent criterion for gauging the similarity between two samples by analyzing their means and covariances. We've adapted SSIM for the complex domain, making it suitable for assessing complex-valued signals.
- **FID** [24]: The Fréchet Inception Distance (FID) evaluates generative models by measuring the Fréchet distance between the high-level features of real and synthesized data. We adopt a pretrained STFNets [73] as the feature extractor to better fit the property of wireless signals.

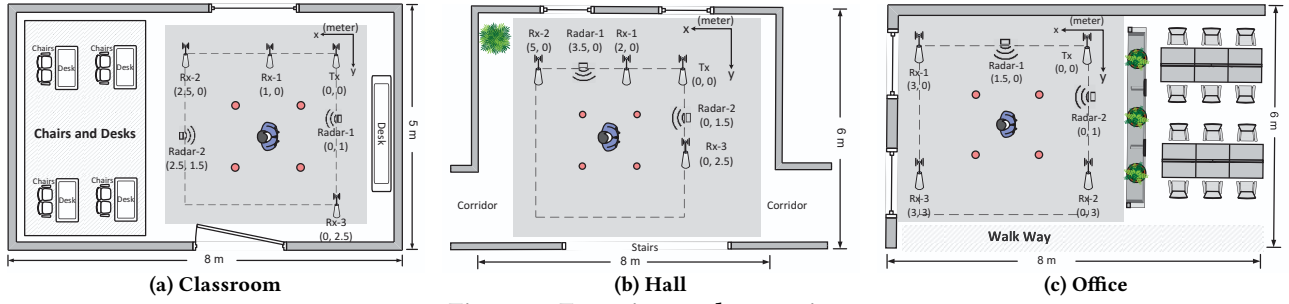


Figure 5: Experimental scenarios.

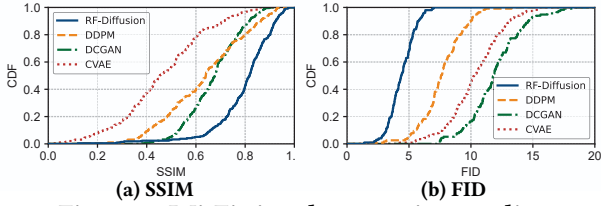


Figure 6: Wi-Fi signal generation quality.

6.2 Overall Generation Quality

The evaluation result for RF-Diffusion on Wi-Fi and FMCW signal are illustrated in Fig. 6 and Fig. 7 respectively. As shown, our proposed RF-Diffusion has proved the superiority over comparative methods on both two metrics.

Specifically, as shown in Fig. 6, RF-Diffusion generates Wi-Fi signal with an average SSIM of 0.81, exceeding DDPM, DCGAN, and CVAE by 25.4%, 18.6% and 71.3% respectively. RF-Diffusion achieves an FID of 4.42, outperforming the above comparative methods by 42.4%, 63.0%, and 57.3%.

RF-Diffusion also outperforms the comparative methods in terms of generating high-fidelity FMCW signals. As shown in Fig. 7, the FMCW signal generated by RF-Diffusion achieves an average SSIM of 0.75 and an average FID of 6.10.

The impressive performance of RF-Diffusion can be attributed to several key factors: 1) Our proposed time-frequency diffusion adopted by RF-Diffusion emphasizes refining the frequency spectrum of the RF signal, thereby preserving finer spectral details in the generated signals, which is difficult to be captured by other methods. 2) Through its iterative generation approach, RF-Diffusion attains precise reconstruction of data details via multi-step approximations, leading to a superior quality of generated data. 3) In contrast to DCGAN, which optimizes two models concurrently, RF-Diffusion's loss function is more streamlined and its training process more stable, ensuring a richer diversity in the generated signal and contributing to a commendable FID score.

6.3 Micro-benchmarks

6.3.1 Impact of Diffusion Methods. To validate the efficacy of our proposed time-frequency diffusion theory, we retained the network model architecture of RF-Diffusion but

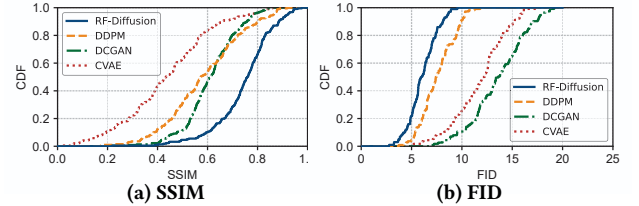


Figure 7: FMCW signal generation quality.

replaced the time-frequency diffusion process with two alternate schemes: 1) Gaussian diffusion, which is similar to DDPM and by only introduces Gaussian noise to the signal amplitude, and 2) blur diffusion which only performs spectral blurring. As depicted in Fig. 8, our time-frequency diffusion theory consistently outperforms both in terms of the SSIM and FID metrics. Specifically, the SSIM values for time-frequency diffusion, Gaussian diffusion, and blur diffusion stand at 0.81, 0.71, and 0.45, respectively. This translates to the time-frequency diffusion offering an SSIM improvement of 13.9% over Gaussian diffusion and a notable 79.2% over blur diffusion. In terms of the FID, the time-frequency diffusion surpasses the other two methods by margins of 41.3% and 83.5%, respectively. The results indicates that the time-frequency diffusion theory successfully incorporates two diffusion methods on orthogonal spaces, and thus achieving complementary benefits.

6.3.2 Impact of Network Design. To demonstrate the advantages of our proposed hierarchical diffusion transformer (HDT), we compare it against: 1) single-stage diffusion transformer (SDT), which is a simplified form of HDT, with only one stage for end-to-end data restoration, and 2) U-Net [55], a popular choice in prevalent diffusion models. As shown in Fig. 9, our proposed HDT outperforms the SDT and U-Net. Specifically, the SSIM for HDT, SDT, and U-Net are 0.81, 0.75, and 0.68 respectively. This indicate that HDT achieves a SSIM boost of 7.7% over SDT and a significant 18.9% increment compared to U-Net. When assessed using the FID metric, HDT continues to lead by margins of 48.2% and 49.3% against SDT and U-Net, respectively. The outstanding performance benefits from the follows aspects: 1) Compared with SDT,

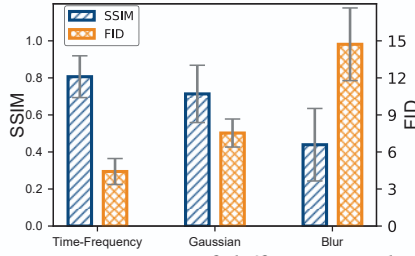


Figure 8: Impact of diffusion method.

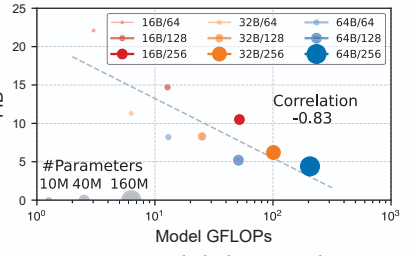
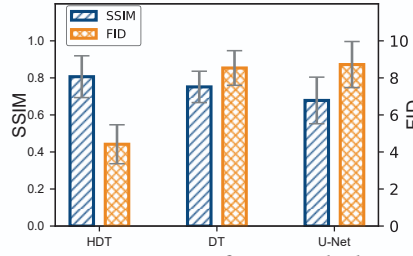


Figure 10: Scalability analysis.

HDT can efficiently decouple the non-isotropic noise introduced in the diffusion process and eliminate it through two sequential stages; 2) Compared with translation-invariant U-Net, HDT's transformer architecture can effectively distinguish the signal characteristics at different times, thereby achieving more accurate signal generation.

6.3.3 Scalability Analysis. Scalability refers to a model's ability to enhance its performance with increasing size, which is critical for large generative models like RF-Diffusion. To verify the scalability of RF-Diffusion, we trained 9 models of different sizes, exploring different numbers of attention-based diffusion blocks (16B, 32B, 64B) and hidden dimensions (64, 128, 256). Fig. 10 illustrates that the FID performance of RF-Diffusion is strongly correlated with model parameters and GFLOPs, indicating that scaling up model parameters and additional model computation is the critical ingredient for better performance. Increasing the model size is anticipated to further enhance RF-Diffusion's performance.

7 CASE STUDY

This section showcases how RF-Diffusion benefits wireless researches in two distinct downstream tasks: Wi-Fi-based gesture recognition and 5G FDD channel estimation.

7.1 Wi-Fi Gesture Recognition

Wireless sensing [12, 17, 71, 74] has emerged as a major research focus. By serving as a data augmentor, RF-Diffusion can boost the performance of existing wireless sensing systems, all while preserving the original model structure without any modifications. In particular, our approach involves initially training RF-Diffusion using a real-world dataset. Subsequently, RF-Diffusion generates synthetic RF signals of the designated type, guided by condition labels. These synthetic samples are then integrated with the original dataset, collectively employed to train the wireless sensing model. Both RF-Diffusion-augmented solution and baseline are fundamentally based on the same real-world dataset, ensuring a fair comparison, as RF-Diffusion itself is trained on this real-world dataset and no extra data is ever involved.

We illustrate this approach through the case of Wi-Fi-based gesture recognition and evaluate the performance

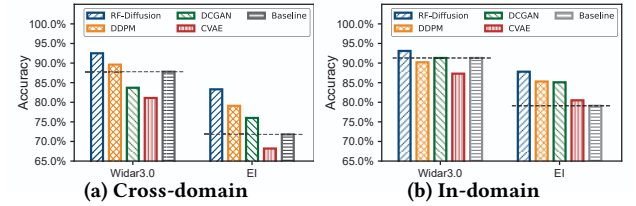


Figure 11: Performance of augmented Wi-Fi sensing.

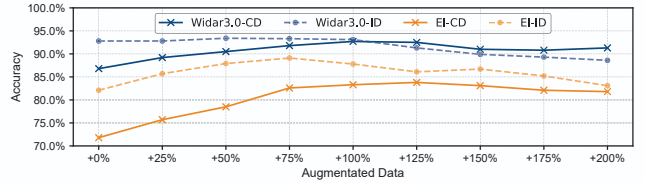


Figure 12: Impact of synthetic data volume.

gains achieved by integrating RF-Diffusion into established gesture recognition models.

7.1.1 Experiment Design. We select two different types of Wi-Fi-based model for a comprehensive evaluation:

- **Widar3.0** [78] is a gesture recognition model founded on physical principles. It initially extracts features from raw signals and subsequently conducts recognition through a deep neural network.
- **EI** [31] is a data-driven end-to-end human activity recognition model that takes raw signal as input.

We utilize the publicly available dataset from Widar3.0 [78] to assess performance. This evaluation encompasses scenarios where RF-Diffusion and comparative methods (§6.1.2) were employed as data augmentors.

7.1.2 Cross-domain Evaluation. We first evaluate the sensing performance when the training and testing set are from different domains (i.e., room, device placement, human location, orientation, etc.), a common scenario in real-world wireless sensing system deployments. We synthesize an equivalent volume of data as the real-world dataset using the pre-trained RF-Diffusion. Subsequently, both synthesized and authentic datasets are used for training. As shown in Fig. 11a, integrating RF-Diffusion brings performance improvements of 4.7% and 11.5% for Widar3.0 and EI, respectively. Integrating DDPM can bring relatively limited performance gains

of only 1.8% and 7.3%, respectively. Additionally, the integration of DCGAN or CVAE may result in a degradation of recognition accuracy due to deviations in the synthetic data distribution from the original data distribution.

Compared with Widar 3.0, the EI model obtains a more significant improvement since: 1) EI is more sensitive to the data volume and data diversity as an end-to-end DNN; 2) the information in the wireless signals generated in a data-driven manner cannot be fully exploited in Widar3.0 when being converted into physical features.

In conclusion, RF-Diffusion enhances the cross-domain performance of wireless sensing systems in two aspects:

- **Enhanced data diversity.** Synthetic training data with higher diversity avoid model overfitting and thus implicitly improves the model's domain generalization ability.
- **Feature distillation.** The generative model RF-Diffusion implicitly imparts its learned signal features to the recognition model through synthetic training data, contributing to improved performance.

7.1.3 In-domain Evaluation. In the in-domain scenario, the training and testing data are from the same domain. As shown in Fig. 11b, the integration of RF-Diffusion yields performance improvements of 1.8% and 8.7% for Widar3.0 and EI, respectively. Overall, compared with the cross-domain scenario, the performance gain in the in-domain case is relatively modest. This is attributed to the limited impact of diverse synthetic training data on enhancing the model's performance within the same domain. For in-domain testing, even with a less diverse synthetic data generated by DCGAN, an obvious performance gain can be achieved.

7.1.4 Impact of Synthesized Data Ratio. We further investigate the impact of the synthesized data ratio used for training to provide more insights. We evaluate Widar 3.0 and EI in both cross-domain (CD) and in-domain (ID) cases.

As shown in the Fig. 12, we introduce varying quantities of synthetic data (from +25% to +200%) to the real-world dataset for joint training of the recognition model. Notably, as the volume of synthetic data increases, the trend in recognition accuracy exhibits an ascent to a peak followed by a decline. Specifically, in the cross-domain case, Widar3.0 reaches the highest accuracy of 92.7% with +100% synthetic data, while EI reaches the highest accuracy of 83.8% with +125% synthetic data. In the in-domain case, Widar3.0 achieved the highest accuracy of 93.4% with +50% synthetic data, while EI achieved the highest accuracy of 89.1% with +75% synthetic data. Drawing from these statistical findings, we deduce the following insights: 1) For most wireless recognition models, judicious incorporation of synthetic data into the training set can effectively enhance model performance. 2) Excessive introduction of synthetic data can potentially shift the training

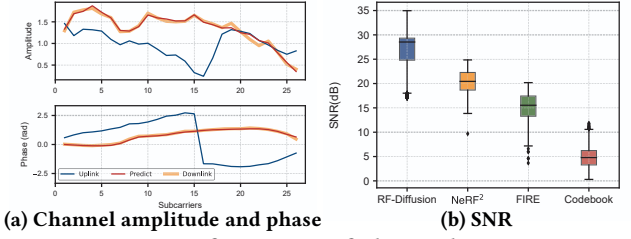


Figure 13: Performance of channel estimation.

data distribution away from the original, consequently diminishing recognition accuracy. 3) The cross-domain scenario requires a greater infusion of synthetic data into the training set to achieve optimal model performance compared to the in-domain scenario. 4) Data-driven end-to-end models (e.g., EI) reap more substantial benefits from data augmentation facilitated by RF-Diffusion.

7.2 5G FDD Channel Estimation

In this section, we discuss how RF-Diffusion enables channel estimation of the Frequency Domain Duplex (FDD) system in 5G, where the uplink and downlink transmissions operate at different frequency bands. Therefore, the principle of reciprocity that two link channels are equal no longer holds [38]. To estimate the downlink channel state, client devices must receive additional symbols from a base station with a massive antenna array and send back the estimated results, causing unsustainable overheads. To solve this problem, substantial research is devoted to predicting the downlink channel by observing the uplink channel state information. For example, FNN [5] and FIRE [38] make use of a fully connected network and a VAE to transfer the estimated CSI from the uplink to the downlink, respectively.

We discover that by employing the uplink CSI as conditional input, RF-Diffusion demonstrates the capacity to estimate downlink channel CSI in a generative manner. Specifically, in RF-Diffusion, the downlink CSI \mathbf{x}_{down} serves as the target data for generation, while the uplink CSI is encoded as the condition \mathbf{c}_{up} and input into the model. The trained RF-Diffusion learns the correlation between \mathbf{c}_{up} and \mathbf{x}_{down} , thereby accomplishing the channel estimation task. This efficacy is rooted in the assumption of shared propagation paths, positing that both link channels are shaped by the same underlying physical environment [28, 65].

7.2.1 Experiment Design. Our evaluation is based on the publicly available dataset Argos [58], which is a real-world MIMO dataset collected in a complex environment with a large number of non-line-of-sight (NLoS) propagation. Each CSI frame contains 52 subcarriers. Similar to previous works [38, 76], we designate the initial 26 subcarriers for the uplink channel, while the remaining 26 are allocated

to the downlink channel. For a comprehensive evaluation, we compare our approach against three different types of channel estimation solutions:

- **NeRF²** [76] implicitly learns the signal propagation environment from the uplink channel state based on a neural network, and then estimate the downlink channel.
- **FIRE** [38] is a channel estimation system based on VAE, which compresses the uplink channel CSI into a latent space representation and further transforms it into a downlink channel estimation.
- **Codebook** [32], commonly used in standard implementations per the 3GPP physical channel standard[1], requires both base stations and clients to maintain a codebook of vectors created using predefined rules. Clients measure the channel locally, select the closest codebook vectors, and send the corresponding indices back to the base station.

7.2.2 Channel Estimation Accuracy. As shown in Fig. 13a, when we input the blue uplink CSI as a condition into the trained RF-Diffusion, the red downlink estimate will be output, which closely aligns with the ground truth downlink channel state. Assessment of channel estimation accuracy employs the Signal-to-Noise Ratio (SNR) metric [38, 76]. This metric gauges the congruence between the estimated downlink channel \mathbf{x}_{est} and the ground truth \mathbf{x}_{down} through the following formulation:

$$\text{SNR} = -10 \log_{10} \left(\frac{\|\mathbf{x}_{\text{down}} - \mathbf{x}_{\text{est}}\|^2}{\|\mathbf{x}_{\text{down}}\|^2} \right). \quad (16)$$

A higher positive SNR corresponds to enhanced proximity between the predicted channel and the ground truth. As shown in Fig. 13b, RF-Diffusion achieves the highest SNR among all comparative methods with an average SNR of 27.01, outperforming NeRF² and FIRE by 34.6% and 77.5% respectively, and achieves more than 5× performance gain compared with the standard implementation based on codebook.

The observed underperformance of NeRF² can be attributed to its treatment of the signal propagation space as a time-invariant system, a characterization that may not hold in practical scenarios. VAE-based FIRE and codebook-based methods fall short in the fine-grained characterization of the underlying distribution of channel states. In contrast, RF-Diffusion adeptly learns the intricate correlation between the uplink and downlink channels, leveraging its robust modeling capacity to achieve highly accurate channel estimation.

8 RELATED WORK

We briefly review the related works in the following.

Diffusion probabilistic models. Diffusion probabilistic models [60, 70] have emerged as a powerful new family of deep generative models with record-breaking performance in many applications [13], including image synthesis, point

cloud completion, and natural language processing, etc. One of the best-known diffusion model is the DDPM [60], which progressively destruct data by injecting gaussian noise, then learn to reverse this process for high-fidelity sample generation. On this basis, DDIM [44] expedites reverse sampling, while LDM [54] conducts diffusion in latent space to curtail computational overhead. The above schemes have been widely used in a wide range of tasks such as image super-resolution [26, 57], inpainting [40], and style transfer [56]. The most recent studies [37, 52] have successfully applied a combination of blurring and additive noise to the image, yielding satisfactory results. Although first proposed for image generation, the diffusion model's versatility extends to other domains including point cloud completion [41, 79], text generation [3, 19], audio synthesis [11, 35], and beyond. In addition, diffusion model has a great potential for multi-modal generation. By integrating pre-trained language model [49], the diffusion models achieve impressive performance in text-to-image [45, 51] and text-to-audio [48] tasks.

RF-Diffusion, in contrast, stands as the pioneering diffusion model tailored for wireless signal generation. It introduces an innovative time-frequency diffusion process, which regulates noise and blurring across two orthogonal domains, thus encompassing both temporal and spectral intricacies of wireless signals. By generating high-fidelity signals, RF-Diffusion benefits wireless applications like Wi-Fi sensing and 5G channel estimation.

Signal generation in wireless systems. Conventional wireless signal generation schemes are mainly based on modeling and simulation. In particular, these methods involve utilizing LiDAR-scanned 3D models, and employing electromagnetic (EM) ray tracing techniques [42] to simulate the distribution of wireless signals. Recent studies [10, 36, 75] have integrated vision-based human reconstruction techniques with signal propagation models, enabling the generation of wireless signals that interact with the human body. Unfortunately, the above schemes fails to model the structure material and physical characteristics, which constraints their performance in real-world applications. The recently proposed NeRF² [76] learns the properties of the signal propagation space based on a deep neural network and then accomplishes the signal generation task. However, NeRF² is limited to specific static scenarios and degrades for dynamic real-world scenarios. RF-EATS [21] and FallDar [72] employ Variational Autoencoders (VAEs) to extract environment-independent features, thereby enhancing the generalizability of wireless sensing models. Additionally, other studies have utilized Generative Adversarial Networks (GANs) to generate Doppler spectrum [16]. Other research endeavors have addressed channel estimation in wireless communication systems using either GANs [6, 14] or VAEs [8, 38]. Nonetheless, due to their limited representation capabilities, solutions

based on GANs and VAEs struggle to faithfully characterize the intrinsic properties of original wireless signals. Consequently, the aforementioned systems are suitable solely for specific tasks, lacking the competence for general-purpose wireless data generation.

In contrast, RF-Diffusion, as a versatile generative model for wireless signals, can proficiently generate fine-grained signals in high-fidelity, even within dynamic scenarios.

9 DISCUSSION AND FUTURE WORK

RF-Diffusion is a pioneering attempt towards diffusion-based RF signal generation, and there is room for continued research in various perspectives.

• **RF-Diffusion for data-driven downstream tasks.** Extensive practices [4, 23, 59, 77] indicate that synthetic data from generative models significantly enhances data-driven downstream tasks. As a conditional generative model, RF-Diffusion effectively captures the representative features and their novel combinations, while randomizing non-essential details. This approach allows for the generation of innovative data samples that extend beyond the initial scope of the dataset, thus improving the generalization ability of downstream models. This paper specifically explores and experiments with applying RF-Diffusion to augment Wi-Fi gesture recognition, demonstrating its potential. However, the applicability of RF-Diffusion extends to any data-driven task in wireless communication and sensing.

• **RF-Diffusion as a simulation tool.** As a probabilistic generative model, RF-Diffusion operates independently of any signal propagation assumptions and does not require pre-modeling of the environment. This flexibility implies that, while RF-Diffusion offers novel opportunities for signal synthesis, it may not achieve the same level of stability and precision as traditional signal simulation tools in all scenarios. RF-Diffusion is not designed to supplant simulation tools but rather to introduce a novel data-driven approach for signal synthesis, which is particularly valuable in complex and dynamic environments, such as indoor spaces with human activity, where accurate modeling poses challenges.

• **Autoregressive signal generation.** RF-Diffusion, a non-autoregressive generative model, processes time series as a unified entity, necessitating downsampling and interpolation for variable-length sequences, which limits its versatility. The advent of autoregressive models like GPT introduces alternative methods for time-series signal generation, which improves adaptability for sequences of differing lengths and enable effective exploration of temporal correlation features.

10 CONCLUSION

This paper presents RF-Diffusion, the pioneering generative diffusion model designed for RF signals. RF-Diffusion excels in generating high-fidelity time-series signals by employing

a novel time-frequency diffusion process. This process captures the intricate characteristics of RF signals across spatial, temporal, and frequency domains. This theoretical framework is then translated into a practical generative model based on the hierarchical diffusion transformer. RF-Diffusion exhibits remarkable versatility. It holds significant potential for essential wireless tasks, ranging from boosting the accuracy of wireless sensing systems, to estimating channel states in communication systems, shedding light on the applications of AIGC in wireless research.

11 ACKNOWLEDGMENTS

We sincerely thank the MobiSense Group, the anonymous reviewers, and our shepherd for their constructive comments and feedback in improving this work. This paper is supported in part by the NSFC under grant No. 62372265, No.62302254, and No. 62222216, and by the Hong Kong RGC ECS under grant 27204522 and RIF under grant R5060-19.

A CONVERGENCE OF FORWARD DESTRUCTION PROCESS

As $T \rightarrow \infty$, the forward process converges to a distribution independent of the original signal. The above proposition is equivalent to the following two conditions: (1) $\lim_{T \rightarrow \infty} \bar{\mu}_T = \mathbf{0}$, (2) $\lim_{T \rightarrow \infty} \bar{\sigma}_T < \infty$. We find a sufficient condition for the above to hold true: all element in $\mathbf{y}_t = \sqrt{\alpha_t} \mathbf{g}_t$ should be less than 1, i.e., $\forall n, y_t^{(n)} < 1$. Under this condition, according to Eqn. 3, $\lim_{T \rightarrow \infty} \bar{\mu}_T = \lim_{T \rightarrow \infty} \bar{\mathbf{y}}_T \mathbf{x}_0 = \mathbf{0}$ holds. Let $\alpha_{\min} = \min(\alpha_t), t \in [1, T]$ and $y_{\max}^{(n)} = \max(y_t^{(n)})$ and $\mathbf{y}_{\max} = (y_{\max}^{(1)}, \dots, y_{\max}^{(N)})$. It can be proven that:

$$\begin{aligned} \lim_{T \rightarrow \infty} \bar{\sigma}_T &= \lim_{T \rightarrow \infty} \sum_{t=1}^T (\sqrt{1 - \alpha_t} \frac{\bar{\mathbf{y}}_T}{\mathbf{y}_t}) \\ &\leq \sqrt{1 - \alpha_{\min}} \lim_{T \rightarrow \infty} \sum_{t=1}^T (y_{\max})^{t-1} = \frac{\sqrt{1 - \alpha_{\min}}}{1 - y_{\max}} < \infty \end{aligned} \quad (17)$$

B REVERSE PROCESS DISTRIBUTION

Based on the Bayes' theorem, we get:

$$\begin{aligned} q(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{x}_0) &= q(\mathbf{x}_t | \mathbf{x}_{t-1}, \mathbf{x}_0) \frac{q(\mathbf{x}_{t-1} | \mathbf{x}_0)}{q(\mathbf{x}_t | \mathbf{x}_0)} \\ &\propto \exp\left(-\frac{1}{2} \left(\frac{(\mathbf{x}_t - \mathbf{y}_t \mathbf{x}_{t-1})^2}{\sigma_t^2} + \frac{(\mathbf{x}_{t-1} - \bar{\mathbf{y}}_{t-1} \mathbf{x}_0)^2}{\bar{\sigma}_{t-1}^2} - \frac{(\mathbf{x}_t - \bar{\mathbf{y}}_t \mathbf{x}_0)^2}{\bar{\sigma}_t^2} \right)\right) \\ &= \exp\left(\left(\frac{\mathbf{y}_t}{\sigma_t}\right)^2 + \left(\frac{1}{\bar{\sigma}_{t-1}}\right)^2 \mathbf{x}_{t-1}^2 - \left(\frac{2\mathbf{y}_t}{\sigma_t^2} \mathbf{x}_t + \frac{2\bar{\mathbf{y}}_{t-1}}{\bar{\sigma}_{t-1}^2} \mathbf{x}_0\right) \mathbf{x}_{t-1} + C(\mathbf{x}_t, \mathbf{x}_0)\right), \end{aligned} \quad (18)$$

in which the recursive relationship is used: $\bar{\sigma}_t^2 = \mathbf{y}_t^2 \bar{\sigma}_{t-1}^2 + \sigma_t^2$, which can be inferred by combining Eqn. 2 and Eqn. 3.

REFERENCES

- [1] 3rd Generation Partnership Project (3GPP). 2023. *5G; NR; Physical channels and modulation*. Technical Specification (TS) 38.211. 3rd Generation Partnership Project (3GPP). Version 17.5.0.
- [2] Zeyuan Allen-Zhu and Yuanzhi Li. 2023. Forward Super-Resolution: How Can GANs Learn Hierarchical Generative Models for Real-World Distributions. In *The Eleventh International Conference on Learning Representations*.
- [3] Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne Van Den Berg. 2021. Structured denoising diffusion models in discrete state-spaces. *Advances in Neural Information Processing Systems* 34 (2021), 17981–17993.
- [4] Shekoofeh Azizi, Simon Kornblith, Chitwan Saharia, Mohammad Norouzi, and David J Fleet. 2023. Synthetic data from diffusion models improves imagenet classification. *arXiv preprint arXiv:2304.08466* (2023).
- [5] Arjun Bakshi, Yifan Mao, Kannan Srinivasan, and Srinivasan Parthasarathy. 2019. Fast and efficient cross band channel prediction using machine learning. In *The 25th Annual International Conference on Mobile Computing and Networking*. 1–16.
- [6] Eren Balevi and Jeffrey G Andrews. 2021. Wideband channel estimation with a generative adversarial network. *IEEE Transactions on Wireless Communications* 20, 5 (2021), 3049–3060.
- [7] Yoshiaki Bando, Kouhei Sekiguchi, and Kazuyoshi Yoshii. 2020. Adaptive Neural Speech Enhancement with a Denoising Variational Autoencoder. In *INTERSPEECH*.
- [8] Michael Baur, Benedikt Fesl, Michael Koller, and Wolfgang Utschick. 2022. Variational autoencoder leveraged mmse channel estimation. In *2022 56th Asilomar Conference on Signals, Systems, and Computers*. IEEE, 527–532.
- [9] Andrew Brock, Jeff Donahue, and Karen Simonyan. 2018. Large Scale GAN Training for High Fidelity Natural Image Synthesis. In *International Conference on Learning Representations*.
- [10] Hong Cai, Belal Korany, Chitra R Karanam, and Yasamin Mostofi. 2020. Teaching rf to sense without rf training measurements. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 4 (2020), 1–22.
- [11] Nanxin Chen, Yu Zhang, Heiga Zen, Ron J Weiss, Mohammad Norouzi, and William Chan. 2021. WaveGrad: Estimating Gradients for Waveform Generation. In *International Conference on Learning Representations*.
- [12] Guoxuan Chi, Zheng Yang, Jingao Xu, Chenshu Wu, Jialin Zhang, Jianzhe Liang, and Yunhao Liu. 2022. Wi-drone: wi-fi-based 6-DoF tracking for indoor drone flight control. In *Proceedings of the 20th Annual International Conference on Mobile Systems, Applications and Services*. 56–68.
- [13] Prafulla Dhariwal and Alexander Nichol. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems* 34 (2021), 8780–8794.
- [14] Akash S Doshi, Manan Gupta, and Jeffrey G Andrews. 2022. Over-the-Air Design of GAN Training for mmWave MIMO Channel Estimation. *IEEE Journal on Selected Areas in Information Theory* 3, 3 (2022), 557–573.
- [15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.
- [16] Baris Erol, Sevgi Z Gurbuz, and Moeness G Amin. 2019. GAN-based synthetic radar micro-Doppler augmentations for improved human activity recognition. In *2019 IEEE Radar Conference (RadarConf)*. IEEE, 1–5.
- [17] Yuchong Gao, Guoxuan Chi, Guidong Zhang, and Zheng Yang. 2023. Wi-Prox: Proximity Estimation of Non-directly Connected Devices via Sim2Real Transfer Learning. In *GLOBECOM 2023-2023 IEEE Global Communications Conference*. IEEE.
- [18] Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 249–256.
- [19] Shansan Gong, Mukai Li, Jiangtao Feng, Zhiyong Wu, and Lingpeng Kong. 2023. DiffuSeq: Sequence to Sequence Text Generation with Diffusion Models. In *The Eleventh International Conference on Learning Representations*.
- [20] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. 2017. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677* (2017).
- [21] Unsoo Ha, Junshan Leng, Alaa Khaddaj, and Fadel Adib. 2020. Food and liquid sensing in practical environments using {RFIDs}. In *17th USENIX Symposium on Networked Systems Design and Implementation (NSDI 20)*. 1083–1100.
- [22] Mutasem Q Hamdan and Khairi A Hamdi. 2020. Variational Auto-encoders application in wireless Vehicle-to-Everything communications. In *2020 IEEE 91st Vehicular Technology Conference (VTC2020-Spring)*. IEEE.
- [23] Ruifei He, Shuyang Sun, Xin Yu, Chuhui Xue, Wenqing Zhang, Philip Torr, Song Bai, and Xiaojuan Qi. 2023. Is synthetic data from generative models ready for image recognition?. In *The Eleventh International Conference on Learning Representations*.
- [24] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *Advances in Neural Information Processing Systems*, Vol. 30. Curran Associates, Inc.
- [25] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems* 33 (2020), 6840–6851.
- [26] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. 2022. Cascaded diffusion models for high fidelity image generation. *The Journal of Machine Learning Research* 23, 1 (2022), 2249–2281.
- [27] Jonathan Ho and Tim Salimans. 2021. Classifier-Free Diffusion Guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*.
- [28] Chongwen Huang, George C Alexandropoulos, Alessio Zappone, Chau Yuen, and Mérouane Debbah. 2019. Deep learning for UL/DL channel calibration in generic massive MIMO systems. In *ICC 2019-2019 IEEE International Conference on Communications (ICC)*. IEEE, 1–6.
- [29] Texas Instruments. 2022. Texas Instruments IWR1443BOOST. <https://www.ti.com/tool/IWR1443BOOST>
- [30] Liming Jiang, Bo Dai, Wayne Wu, and Chen Change Loy. 2021. Focal frequency loss for image reconstruction and synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 13919–13929.
- [31] Wenjun Jiang, Chenglin Miao, Fenglong Ma, Shuochao Yao, Yaqing Wang, Ye Yuan, Hongfei Xue, Chen Song, Xin Ma, Dimitrios Koutsonikolas, et al. 2018. Towards environment independent device free human activity recognition. In *Proceedings of the 24th annual international conference on mobile computing and networking*. 289–304.
- [32] Florian Kaltenberger, David Gesbert, Raymond Knopp, and Marios Kountouris. 2008. Performance of multi-user MIMO precoding with

- limited feedback over measured channels. In *IEEE GLOBECOM 2008-2008 IEEE Global Telecommunications Conference*. IEEE, 1–5.
- [33] D Kinga, Jimmy Ba Adam, et al. 2015. Adam: A method for stochastic optimization. In *International conference on learning representations (ICLR)*, Vol. 5. San Diego, California, 6.
- [34] Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).
- [35] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. 2020. DiffWave: A Versatile Diffusion Model for Audio Synthesis. In *International Conference on Learning Representations*.
- [36] Belal Korany, Chitra R Karanam, Hong Cai, and Yasamin Mostofi. 2019. XModal-ID: Using WiFi for through-wall person identification from candidate video footage. In *The 25th Annual International Conference on Mobile Computing and Networking*. 1–15.
- [37] Sangyun Lee, Hyungjin Chung, Jaehyeon Kim, and Jong Chul Ye. 2022. Progressive Deblurring of Diffusion Models for Coarse-to-Fine Image Synthesis. In *NeurIPS 2022 Workshop on Score-Based Methods*. <https://openreview.net/forum?id=KP8BrpZBbv>
- [38] Zikun Liu, Gagandeep Singh, Chenren Xu, and Deepak Vasishth. 2021. FIRE: enabling reciprocity for FDD MIMO systems. In *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking*. 628–641.
- [39] Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.
- [40] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. 2022. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11461–11471.
- [41] Zhaoyang Lyu, Zhifeng Kong, Xudong Xu, Liang Pan, and Dahua Lin. 2021. A conditional point diffusion-refinement paradigm for 3d point cloud completion. *arXiv preprint arXiv:2112.03530* (2021).
- [42] John W McKown and R Lee Hamilton. 1991. Ray tracing as a design tool for radio networks. *IEEE Network* 5, 6 (1991), 27–30.
- [43] Midjourney. 2023. Midjourney. <https://www.midjourney.com/>
- [44] Alexander Quinn Nichol and Prafulla Dhariwal. 2021. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*. PMLR, 8162–8171.
- [45] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. 2022. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. In *Proceedings of the 39th International Conference on Machine Learning*.
- [46] OpenAI. 2023. GPT-4 Technical Report. *arXiv:2303.08774 [cs.CL]*
- [47] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. 2018. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.
- [48] Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, and Mikhail Kudinov. 2021. Grad-tts: A diffusion probabilistic model for text-to-speech. In *International Conference on Machine Learning*. PMLR, 8599–8608.
- [49] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.
- [50] Alec Radford, Luke Metz, and Soumith Chintala. 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434* (2015).
- [51] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125* 1, 2 (2022), 3.
- [52] Severi Rissanen, Markus Heinonen, and Arno Solin. 2023. Generative Modelling with Inverse Heat Dissipation. In *The Eleventh International Conference on Learning Representations*. <https://openreview.net/forum?id=4PJUBT9f2OI>
- [53] Hamada Rizk, Ahmed Shokry, and Moustafa Youssef. 2019. Effectiveness of data augmentation in cellular-based localization using deep learning. In *2019 IEEE Wireless Communications and Networking Conference (WCNC)*. IEEE.
- [54] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10684–10695.
- [55] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III* 18. Springer, 234–241.
- [56] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. 2022. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*. 1–10.
- [57] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. 2022. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 4 (2022), 4713–4726.
- [58] Clayton Shepard, Jian Ding, Ryan E Guerra, and Lin Zhong. 2016. Understanding real many-antenna MU-MIMO channels. In *2016 50th Asilomar Conference on Signals, Systems and Computers*. IEEE, 461–467.
- [59] C Shivashankar and Shane Miller. 2023. Semantic Data Augmentation With Generative Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 863–873.
- [60] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*. PMLR, 2256–2265.
- [61] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. 2015. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems* 28 (2015).
- [62] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. 2015. Learning Structured Output Representation using Deep Conditional Generative Models. In *Advances in Neural Information Processing Systems*, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett (Eds.), Vol. 28. Curran Associates, Inc.
- [63] Yang Song and Stefano Ermon. 2019. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems* 32 (2019).
- [64] Chihab Trabelsi, Olexa Bilaniuk, Ying Zhang, Dmitriy Serdyuk, Sandeep Subramanian, Joao Felipe Santos, Soroush Mehri, Negar Rostamzadeh, Yoshua Bengio, and Christopher J Pal. 2018. Deep Complex Networks. In *International Conference on Learning Representations*.
- [65] Deepak Vasishth, Swarun Kumar, Hariharan Rahul, and Dina Katabi. 2016. Eliminating channel feedback in next-generation cellular networks. In *Proceedings of the 2016 ACM SIGCOMM Conference*. 398–411.
- [66] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).
- [67] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* 13, 4 (2004), 600–612.
- [68] Eric W Weisstein. 2023. Convolution Theorem. <https://mathworld.wolfram.com/ConvolutionTheorem.html>

- [69] Weihao Xia, Yulun Zhang, Yujia Yang, Jing-Hao Xue, Bolei Zhou, and Ming-Hsuan Yang. 2022. Gan inversion: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022).
- [70] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Yingxia Shao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. 2022. Diffusion models: A comprehensive survey of methods and applications. *arXiv preprint arXiv:2209.00796* (2022).
- [71] Zheng Yang, Yi Zhang, Kun Qian, and Chenshu Wu. 2023. SLNet: A Spectrogram Learning Neural Network for Deep Wireless Sensing. In *20th USENIX Symposium on Networked Systems Design and Implementation (NSDI 23)*. 1221–1236.
- [72] Zheng Yang, Yi Zhang, and Qian Zhang. 2022. Rethinking fall detection with Wi-Fi. *IEEE Transactions on Mobile Computing* (2022).
- [73] Shuochao Yao, Ailing Piao, Wenjun Jiang, Yiran Zhao, Huajie Shao, Shengzhong Liu, Dongxin Liu, Jinyang Li, Tianshi Wang, Shaohan Hu, et al. 2019. Stfnets: Learning sensing signals from the time-frequency perspective with short-time fourier neural networks. In *The World Wide Web Conference*. 2192–2202.
- [74] Guidong Zhang, Guoxuan Chi, Yi Zhang, Xuan Ding, and Zheng Yang. 2023. Push the Limit of Millimeter-wave Radar Localization. *ACM Transactions on Sensor Networks* 19, 3 (2023), 1–21.
- [75] Xiaotong Zhang, Zhenjiang Li, and Jin Zhang. 2022. Synthesized Millimeter-Waves for Human Motion Sensing. In *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems*. 377–390.
- [76] Xiaopeng Zhao, Zhenlin An, Qingrui Pan, and Lei Yang. 2023. NeRF²: Neural Radio-Frequency Radiance Fields. *arXiv preprint arXiv:2305.06118* (2023).
- [77] Chenyu Zheng, Guoqiang Wu, and Chongxuan Li. 2023. Toward Understanding Generative Data Augmentation. *arXiv preprint arXiv:2305.17476* (2023).
- [78] Yue Zheng, Yi Zhang, Kun Qian, Guidong Zhang, Yunhao Liu, Chenshu Wu, and Zheng Yang. 2019. Zero-effort cross-domain gesture recognition with Wi-Fi. In *Proceedings of the ACM MobiSys*.
- [79] Linqi Zhou, Yilun Du, and Jiajun Wu. 2021. 3d shape generation and completion through point-voxel diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5826–5835.