

Wideband Low-complexity High-speed 5G NR Backscatter

Zhenzhe Lin
George Mason University
Fairfax, Virginia, USA
zlin9@gmu.edu

Yoon Chae
University of Texas at Arlington
Arlington, Texas, USA
yoon.chae@uta.edu

Panneer Selvam Santhalingam
Brooklyn College, CUNY
Brooklyn, New York, USA
ps.santhalingam@brooklyn.cuny.edu

Mingyo Jeong
George Mason University
Fairfax, Virginia, USA
mjeong6@gmu.edu

Parth Pathak
George Mason University
Fairfax, Virginia, USA
ppathak@gmu.edu

Abstract

With recent wireless standards such as 5G NR utilizing large bandwidth channels, wideband OFDM backscatter can achieve very high data rates (tens of Mbps) and support applications such as HD video streaming from low-power IoT sensors. However, realizing wideband OFDM backscatter in practice is challenging due to its extremely high complexity of demodulation, its dependence on accurate channel estimation, and the lack of accurate synchronization. In this paper, we present WiNB, a low-complexity, high-speed 5G NR FR2 wideband OFDM backscatter system. At the core of WiNB is a dual-task transformer model that runs on a backscatter receiver and exploits the interdependency between channel estimation and backscatter demodulation to not only improve backscatter bit recovery but also provide tolerance to synchronization errors and significantly lower demodulation complexity. Furthermore, the model performs well across different environments without relying heavily on site-specific measurements. We evaluate WiNB using a 5G NR tag prototype and software radios. WiNB achieves 71 Mbps throughput with a BER of 10^{-4} and a three orders-of-magnitude improvement in OFDM backscatter demodulation over the state of the art.

CCS Concepts

• Computer systems organization → Sensor networks.

Keywords

Backscatter, 5G NR, Channel Estimation, Transformer

ACM Reference Format:

Zhenzhe Lin, Yoon Chae, Panneer Selvam Santhalingam, Mingyo Jeong, and Parth Pathak. 2026. Wideband Low-complexity High-speed 5G NR Backscatter. In *The 24th Annual International Conference on Mobile Systems, Applications and Services (MobiSys '26)*, June 21–25, 2026, Cambridge, United Kingdom. ACM, New York, NY, USA, 14 pages. <https://doi.org/10.1145/3745756.3809221>

1 Introduction

Backscatter communication that leverages ambient signals, commonly known as ambient backscatter, has emerged as a key technology for ultra-low-power Internet of Things (IoT) devices. Here, low-power tags backscatter existing wireless signals (WiFi or cellular) and embed their data by modulating phase, amplitude, and/or frequency, which can be received by commodity devices. This enables low-power tag operations while leveraging existing wireless infrastructure for excitation signals, reducing deployment costs.

The early works [27, 34] on ambient backscatter relied on on-off keying type modulations, which achieved much lower data rates (typically a few Kbps). Recent works [24, 62] have sought to increase data rates by backscattering OFDM signals, which are widely used in WiFi and cellular systems. The approaches presented in [26, 62] embed one backscatter bit per OFDM symbol (symbol-level modulation) via phase-shift keying. While these solutions improve the data rate, the speeds remain limited to hundreds of Kbps. Recent works such as [13, 35, 43] propose to push the throughput limits by changing the phase of each sample (i.e., subcarrier/sample-level) of the OFDM symbol, allowing much higher data rates of 10 Mbps.

While the feasibility of sample-level OFDM backscatter has been demonstrated for WiFi and LTE cellular signals, their data rates remain low (a few Mbps), limiting their support for many practical IoT applications. For example, recent works [16, 23, 53, 56] highlight the need for higher speeds to support applications such as HD video streaming from low-power IoT sensors. Such high-throughput backscatter can be deployed in industry or enterprise-scale private 5G networks, which are becoming increasingly popular in recent times. For example, the tag can be attached to items/goods in a warehouse or equipment in a manufacturing plant for low-power, high-throughput video monitoring or surveillance. Addressing the trade-off between speed and power in backscatter design is extremely challenging, especially within the realm of commodity wireless networks. One potential solution to achieving a high data rate in commodity backscatter is to leverage wideband channels used in recent wireless standards. Latest WiFi standards, such as 802.11be propose to use 320 MHz wide channels for more capacity. Similarly, 5G NR has standardized the use of 100, 200, and even 400 MHz bandwidths for mmWave FR2 bands. Most prior works on commodity backscatter focus on narrow channels (typically 20 MHz) due to the high complexity involved in backscattering wideband signals. While our prior work [10] demonstrated backscattering on wideband 60 GHz 802.11ad channels, it used only



This work is licensed under a Creative Commons Attribution 4.0 International License. *MobiSys '26, Cambridge, United Kingdom*
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2027-7/2026/06
<https://doi.org/10.1145/3745756.3809221>

single-carrier control packets. Other works on mmWave backscatter [5, 9, 10, 29, 36–38, 46] either require dedicated FMCW radars as readers or use custom non-OFDM waveforms for backscatter excitation. Backscattering ambient wideband OFDM signals with sample-level modulation on hundreds of subcarriers to achieve high data rates remains an open problem.

In this paper, we present WiNB, which is a low-complexity, high-speed, wideband backscatter solution for 5G NR OFDM signals. WiNB creates a novel approach with two-fold contributions. First, we carefully understand and outline the challenges involved in wideband OFDM backscatter. Through experiments, we find that applying existing backscatter demodulation techniques to wideband OFDM signals with hundreds of subcarriers creates serious performance bottlenecks in terms of demodulation complexity and BER. We then develop a custom-tailored transformer-based solution to model the interactions between channel estimation and backscatter demodulation. The model provides 71 Mbps with BER as low as 10^{-4} . Second, we address these wideband backscatter problems in the context of 5G NR mmWave FR2 cellular networks that are already widely deployed and are increasingly being adopted as private 5G networks. WiNB addresses protocol-specific challenges to ensure correct backscatter operations within the 5G frames while maintaining the tag's low-power profile.

Challenges. We first experimentally evaluate the performance of existing OFDM backscatter techniques in wideband 5G NR channels. We find three key limitations: (1) Demodulation in sample-level OFDM backscatter requires solving a maximum-likelihood optimization problem. Existing methods for estimating solutions rely on quasi-Newton (QN) or genetic algorithms (GA), which have very high complexity even for a reasonable BER. As the number of subcarriers (i.e., the number of backscatter bits embedded in each OFDM symbol) increases, solving the problem becomes computationally prohibitive: demodulating an OFDM symbol that is a few microseconds long can take tens, if not hundreds, of milliseconds. (2) Conventional OFDM backscatter demodulation techniques perform channel estimation first before demodulating the backscatter bits. This means that any channel estimation error is carried forward to backscatter demodulation, adversely affecting its BER. We find that in wideband backscatter, even a small channel estimation error yields very high BER, diminishing any gains of using wide channels in practice. (3) Existing backscatter demodulation solutions are highly sensitive to any synchronization errors between the transmitted OFDM symbol and the backscatter symbols to be embedded on it. Given that a low-power tag can only perform coarse envelope detection through power detectors, such synchronization errors are unavoidable in practice. The problem is exacerbated when we have more samples per OFDM symbol, resulting in very high BER even for small synchronization errors.

Our approach. We address these challenges by developing a dual-task transformer model that runs on the 5G NR receiver. Our key insight is that both channel estimation and backscatter demodulation are tightly coupled problems. We model this intricate dependency using a transformer with two tasks learned in parallel, each solving its own objective, while better learning and sharing common representative features across the two tasks to improve their performance. Our dual-task transformer achieves a much

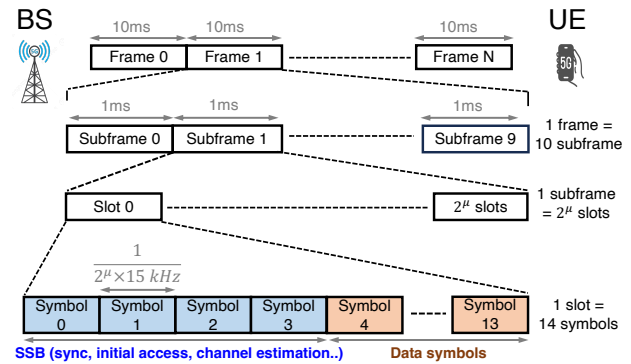


Figure 1: 5G NR frame, subframe, slots, and symbols.

higher demodulation accuracy. It is tolerant of synchronization errors because transformers can model long-term temporal patterns when modulating wideband OFDM symbols with many samples. Furthermore, we develop a custom pruning framework to ensure that inference times are much closer to the actual OFDM symbol durations, making the wideband backscatter system practically viable. Another salient feature is that the model has much lower dependency on environment-specific training and performs well in new environments. Table 1 compares WiNB with state-of-the-art.

Prototyping and evaluation. We thoroughly evaluate WiNB using a 28 GHz 5G NR backscatter prototype with a phased antenna array and a high-speed RF switch that can backscatter wideband signals, along with two 5G NR software radios configured as the transmitter and receiver. We find that WiNB can achieve a BER of 10^{-4} and a throughput of 71 Mbps. Compared to the state-of-the-art solutions, this is a $7\times$ improvement in throughput. Compared to QN and GA algorithms for backscatter demodulation, which take hundreds of ms, WiNB dual-task transformer takes $16\ \mu\text{s}$, which is close to the actual OFDM symbol duration, enabling near real-time demodulation. This is a three orders-of-magnitude improvement in demodulation time per symbol. We also find that our dual-task transformer outperforms single-task alternatives in terms of BER, with its ability to better model channel estimation and backscatter demodulation dependencies. We evaluate the model across different cross-environment testing scenarios and find that it achieves high accuracy even when trained and tested in different environments.

Contributions. The paper's main contributions can be summarized as follows.

- (1) We systematically understand the limitations of existing backscatter techniques when applied to wideband 5G NR OFDM channels capable of achieving high data rates.
- (2) We develop a dual-task transform model that addresses the channel-backscatter dependencies, synchronization, and demodulation complexity challenges of wideband backscatter.
- (3) We prototype 5G NR backscatter tags and evaluate our model to demonstrate its superior performance in terms of BER, throughput, near real-time demodulation, and cross-site generalization.

2 Background

5G NR PHY and frame structure. Fig. 1 illustrates the 5G NR frame structure. A 5G NR frame is 10 ms long and consists of 10

System	Ambient signal band (bandwidth (MHz))	OFDM?	OFDM Demod. Complexity	Modulation level	Throughput (Mbps)	Tolerance to unsynchronization
WiTAG [3]	sub-6 WiFi (20)	Yes	low	Symbol-level	0.004	high
Hitchhike [61]	sub-6 WiFi (20)	No	N/A	Symbol-level	0.3	low
PTL [43]	sub-6 WiFi (20)	Yes	high	Sample-level	10	low
TScatter [35]	sub-6 WiFi (20)	Yes	high	Sample-level	10	medium
LScatter [13]	sub-6 LTE (20)	Yes	high	Sample-level	13	medium
mmComb [10]	mmWave WiFi (2176)	No	N/A	Symbol-level	55	low
WiNB	mmWave 5G NR (100)	Yes	low	Sample-level	71	high

Table 1: Comparison of existing commodity wireless backscatter systems.

subframes (1 ms each). Each subframe contains 2^μ slots, and each slot comprises 14 OFDM symbols. The parameter μ , known as the numerology index where $\mu = \{0, 1, 2, 3, 4\}$, determines the subcarrier spacing (SCS) as $15 \times 2^\mu$ KHz. The numerology directly affects the number of subcarriers in a channel, the OFDM symbol duration, and the number of slots per subframe, as shown in Fig. 1. Increasing the numerology index (μ) widens the SCS and shortens both the OFDM symbol duration and slot duration, thereby increasing the number of slots within a subframe. For example, $\mu = 0$ (15 KHz SCS) yields a $66.67 \mu\text{s}$ OFDM symbol and a 1 ms slot, while $\mu = 4$ (240 KHz SCS) yields a $4.17 \mu\text{s}$ symbol and 16 slots per subframe. Each slot consists of 14 OFDM symbols. To enable UEs to detect the presence of a 5G network and achieve time- and frequency-synchronization, 5G NR transmits a Synchronization Signal Block (SSB), which consists of four consecutive OFDM symbols containing the PSS, PBCH, SSS, and PBCH-DMRS symbol. Similar to WiFi beacons, SSBs are transmitted periodically (typically every 5–20 ms), and their positions within a frame depend on the configured subcarrier spacing, frequency range, and deployment options.

5G NR OFDM sample-level backscatter. A 5G NR OFDM sample-level backscatter system consists of a sender, a backscatter tag, and a receiver, as shown in Fig. 2. The sender is typically a 5G NR base station (BS), and the receiver is a UE.

Sender: 5G NR adopts OFDM modulation for downlink and uplink. After mapping modulation symbols onto respective subcarriers, an Inverse Discrete Fourier Transform (IDFT) generates the discrete-time baseband OFDM samples $x(n) = \text{IDFT}\{X[f]\}$, where $X[f]$ denotes the complex symbol assigned to subcarrier f . This time-domain OFDM signal is then upconverted to the carrier frequency f_c to generate the transmitted passband waveform $S(t) = x(t)e^{j2\pi f_c t}$.

Tag: In existing WiFi-based OFDM backscatter, tags encode data by phase inversion of time-domain OFDM samples (i.e., sample-level modulation in Fig. 2). Similarly, for 5G NR OFDM, the tag can use a square waveform (approximated as the first-order harmonic cosine wave as in [13]) to modulate the phases of the samples. The backscatter tag uses zero phase offset to transmit data '0' and a phase inversion by π to transmit data '1'. To mitigate self-interference, the tag also shifts the carrier frequency (Fig. 2) using the single sideband technique proposed in [61]. This way, the resulting modified backscatter signal becomes

$$B(t) = \begin{cases} S(t)e^{2\pi f_s t} e^{j0}, & \text{tag bit '0'} \\ S(t)e^{2\pi f_s t} e^{j\pi}, & \text{tag bit '1'} \end{cases} \quad (1)$$

where f_s is the frequency shift applied by the tag. This frequency shift causes the backscattered component to appear on an adjacent channel instead of the original channel used by the BS (Ch 1 and Ch

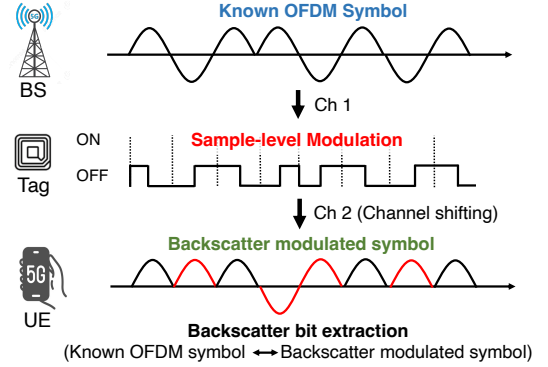


Figure 2: Overview of 5G NR backscatter.

2 in Fig. 2). This naturally suppresses the strong self-interference caused by the direct signal from the BS to the UE, enabling the UE to isolate and demodulate the weak backscatter signal.

Receiver: At the 5G NR receiver, the SSBs are used to estimate the channel, and this estimate is then applied to demodulate the subsequent data symbols. For a receiver operating on an adjacent channel that only receives the backscatter signal, the transmitter must send pre-known data symbols to enable the decoding of unknown tag data. Because OFDM inherently spreads any time-domain modification across all subcarriers, the sample-level phase modulation applied by the tag affects the entire OFDM symbol. This means that decoding the tag's data by just examining individual subcarriers is not possible. Instead, the receiver extracts the tag's encoded phase by comparing the received waveform against a set of reference signals with known symbols, selecting the one with the smallest Euclidean distance (i.e., maximum-likelihood optimization problem)

$$\theta = \arg \min_{\theta} \|Y^* - Y(n)\| \quad (2)$$

where $Y(n)$ is the received subcarrier values that carry the combined effect of the original transmission and the tag's modulation, and Y^* represents the mapped nearest constellation points for $Y(n)$. The objective is to find the tag modulation vector $\theta = [\theta_1, \theta_2, \dots, \theta_N]^T$ that minimizes the discrepancy between the received subcarrier values and the mapped constellation points. Because the search space for θ grows exponentially with the vector length, the exhaustive search is naturally computationally prohibitive. More tractable and efficient iterative optimization algorithms, such as QN and GA are employed in prior work, such as TScatter [35], LScatter [13], and PTL [43] for demodulation.

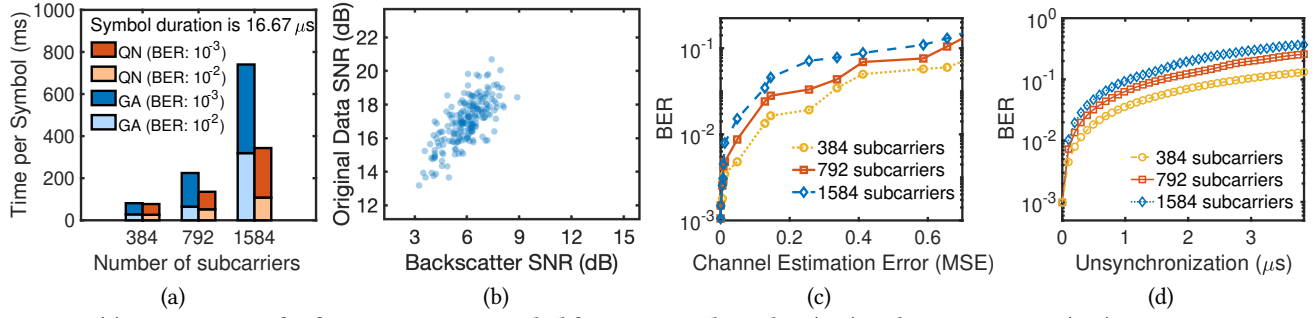


Figure 3: (a) Comparison of inference time per symbol for Genetic Algorithm (GA) and Quasi-Newton (QN) approximation across different number of subcarriers ($N_{\text{sub}} = 384, 792, \text{ and } 1584$); (b) Comparison of original data SNR (Channel 1) and backscatter SNR (Channel 2); (c) BER vs. channel estimation error (MSE) across subcarriers; (d) BER vs. unsynchronization across subcarriers.

3 Motivation

Challenges in wideband OFDM backscatter. Wideband OFDM backscatter systems, such as 5G NR backscatter, face unique challenges in reliably and efficiently decoding backscatter data.

C1 Time overhead. A key challenge in demodulating the sample-level OFDM backscatter is the high computational complexity of solving the optimization problem in Equ. 2. It essentially requires searching for tag data θ (of N bits for N subcarriers) that minimizes the distance between the expected and received data. When the sample-level backscatter is performed on wideband channels, the number of subcarriers also increases (large N), resulting in prohibitively high computational complexity of solving the optimization problem. The search space is 2^N for BPSK (M^N if the backscatter modulation is M -ary PSK), and the complexity of the standard quasi-Newton (QN) estimator (without refactorizing the Hessian) is $O(N^2)$. Prior works [35, 43] have proposed using a genetic algorithm (GA), which still requires multiple generations (G) from a large population size (P) over a large number of iterations, resulting in complexity of at least $O(NGP)$. We apply the standard QN and GA methods to demodulate backscatter samples within an OFDM symbol across different subcarrier settings. For our experiments, we adopt four representative 5G NR FR2 configurations, two for 100 MHz bandwidth (i) 120 KHz SCS, 792 subcarriers, and (ii) 60 KHz, 1584 subcarriers, and two for 50 MHz bandwidth (iii) 120 KHz SCS, 384 subcarriers, and (iv) 60 KHz, 792 subcarriers. The symbol durations vary from 8.33 to 16.67 μs for these configurations. Fig. 3(a) shows the demodulation time for different targeted BER for backscatter data for MacBook M3 (CPU only) and compares it with the OFDM symbol duration (16.67 μs). As we can see, achieving a BER of 10^{-2} for 384 subcarriers requires a minimum decoding time of 53ms for GA, and requires 51ms for QN, which is at least 3000 \times greater than the 16.67 μs symbol duration. When the backscatter tag continuously modulates the 5G NR OFDM symbols, the time overhead can easily overwhelm the UE receiver’s memory and compute. There is clearly a need for a more efficient yet accurate demodulation method in wideband OFDM backscatter.

C2 Impact of channel estimation. Prior works [13, 35, 43] treat channel estimation and backscatter demodulation as sequential tasks where first the channel is estimated using pilot/reference symbols and then equalized symbols are used for backscatter demodulation. This means that even small channel estimation errors

can have a significant impact on backscatter demodulation because the tag signal is orders of magnitude weaker than the OFDM data signal. Fig. 3(b) shows the SNR of OFDM symbols received on Channel 1 (original transmitted data without backscatter) and Channel 2 (original transmitted data with backscatter). We see that the backscatter data has much lower SNR compared to the original data. As a result, minor channel estimation inaccuracies that have a negligible impact on demodulating the original OFDM data can cause a disproportionately large degradation in backscatter demodulation performance. Fig. 3(c) plots the backscatter BER as a function of MSE (mean-squared error) for channel estimation across different subcarrier configurations. Here, the MSE measures the deviation between the estimated and true channel. Since we cannot directly obtain ground truth by measuring the channel, we reconstruct the pilot using the least-squares method and calculate the difference using MSE between the original and reconstructed pilots. The results show that as channel estimation accuracy decreases, the backscatter BER increases, and the increase is steeper for a larger number of subcarriers. The impact of even small channel estimation errors is amplified with more subcarriers (i.e., more backscatter bits per symbol), leading to more erroneous backscatter demodulation.

C3 Synchronization. A 5G NR backscatter tag must detect and avoid modulating the SSB symbols because modulating these symbols would corrupt the channel estimation. Achieving synchronization between the transmitted and backscatter signals at the sample level is extremely challenging, as the tags can only perform coarse detection of incoming signals through envelope detectors. This unsynchronization between the transmitted and tag symbols makes it even more difficult for QN and GA methods to decode the backscatter bits. The cyclic prefix-based solution introduced in [13] can help alleviate the problem, but it sacrifices a significant (over one-third) amount of throughput. Fig. 3(d) shows the BER for 5G NR sample-level OFDM backscatter under varying synchronization errors for different numbers of subcarriers. For these configurations with symbol durations varying from 8 to 16 μs , we find that as the synchronization error increases to 3 μs , the BER increases dramatically to 10^{-1} . Furthermore, in wideband OFDM with an increasing number of subcarriers, BER increases much more with higher levels of unsynchronization. This means that backscatter demodulation should be robust to such synchronization errors, especially with a larger number of subcarriers in wideband OFDM.

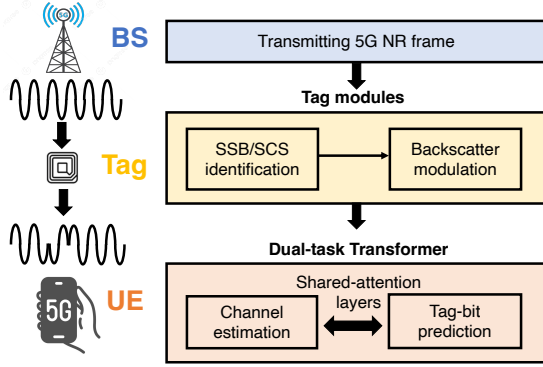


Figure 4: System overview.

4 WiNB Design

4.1 System Overview

Fig. 4 shows an overview of WiNB. An unmodified 5G NR BS transmits 5G NR frames, which are received by the backscatter tag and passed through the SSB/SCS detection and identification module (Sec. 5). The signal is then backscattered using sample-level OFDM modulation. The backscattered signal is received by the UE and fed into a dual-task transformer that produces channel estimates and demodulates backscatter tag bits.

4.2 Demodulation with dual-task transformer

Why dual-task multi-head transformer? Phase variations caused by the backscatter modulation do not remain confined to a single frequency but appear as variations across the entire OFDM symbol due to the IFFT. Models that can capture such temporal variations are, therefore, a natural fit. Given that our problem is somewhat analogous to sequence-to-sequence modeling [6, 51, 52], where the input is frequency-domain symbols and the output is per-subcarrier predictions for channel and tag bits, we choose transformers for this purpose. We propose a novel dual-task transformer architecture. Multi-task learning is widely used for the simultaneous learning of multiple, related prediction tasks. By exploiting the commonalities and differences across related tasks, it has been shown to achieve better generalization across all tasks [8, 45, 49]. It can also bias the model to better learn features that are representative of all tasks while ensuring task-specific performance.

We adopt a dual-task architecture with a shared encoder and two tasks: a *channel estimation* task that estimates the channel, and a *backscatter demodulation* task that demodulates the tag bits. The reason is that channel estimation accuracy and backscatter demodulation quality are tightly coupled. As we saw in Sec. 3, poor channel estimation results in higher BER for backscatter bits. The backscattered OFDM symbols carry the effect of the channel as well as the tag modulation. In practice, the effective channel observed over the OFDM symbols in a backscatter system combines both effects, and the model must learn to separate the tag-induced variation from the underlying propagation characteristics derived from the pilots. By training the network to perform both tasks jointly, the channel head supplies the necessary reference for the backscatter head, enabling more accurate bit recovery. Furthermore, better channel estimates not only improve backscatter demodulation, but

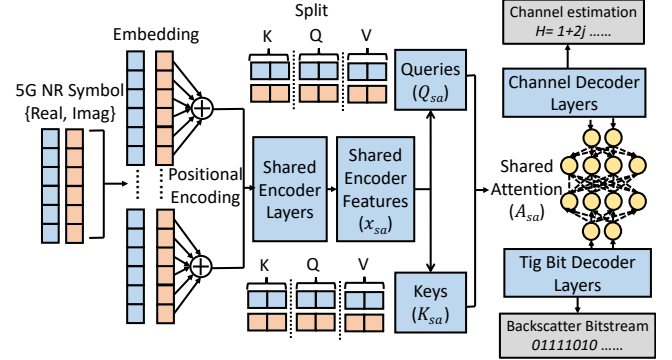


Figure 5: Dual-task transformer architecture

more reliable tag-bit decoding also provides structural feedback that can refine channel predictions.

4.3 Model architecture and training

Input preprocessing of 5G NR signals. Fig. 5 shows our dual-task transformer model. Our model takes a single 5G NR OFDM symbol as input. Each instance consists of N subcarriers, and the complex IQ samples of these subcarriers are stored as real-imaginary pairs $x \in \mathbb{R}^{B \times N \times 2}$ where B is the batch size. Treating each subcarrier as an individual token yields a sequence of length N . Before feeding the data into the transformer, each token is embedded into a d_{model} -dimensional space and combined with a positional encoding that preserves the ordering of subcarriers along the frequency axis. This step produces the input sequence $h' \in \mathbb{R}^{B \times N \times d_{\text{model}}}$ which captures both the raw IQ structure and the frequency-dependent position information of the OFDM symbol.

Shared attention encoder and decoder. Our model begins by encoding the input sequence with a shared transformer encoder $E(\cdot)$, comprising two shared encoder layers, and each layer consists of multi-head self-attention followed by layer norm and fully connected layers. Let H denote the number of attention heads, and let $d_h = d_{\text{model}}/H$ be the dimensionality of each head. The encoder outputs a sequence of contextualized features $x_{\text{sa}} = E(x)$, which serves as the shared representation for both channel estimation and tag-bit demodulation. To enable multitask processing, we construct a decoder $D(\cdot)$ with two task-specific branches. Unlike a standard transformer architecture, where each task computes its own attention maps, our decoder forms a shared attention pattern. Specifically, the decoder first produces a set of shared queries and keys, Q_{sa} and K_{sa} , from the encoder output $x_{\text{sa}} = E(x)$. These are used to compute a shared attention map $A_{\text{sa}} = \text{softmax}\left(\frac{Q_{\text{sa}}K_{\text{sa}}^T}{\sqrt{d_h}}\right)$, where all operations are applied independently across the H heads. The dimensionality d_h appears in the normalization term and corresponds to the per-head feature size. Each task-specific decoder branch supplies its own value projection, allowing both tasks to update their features using the same attention structure while maintaining task-dependent transformations. Shared attention enables the model to focus on a subset of subcarriers important to both tasks, whereas task-specific value paths allow the two tasks to specialize. In the time domain, a single attention map from shared encoder features captures the most informative parts of the sequence. The decoder stacks two such shared-attention blocks with

feed-forward layers to progressively refine the representations before prediction. The channel estimation branch ultimately produces a per-subcarrier estimate of the complex channel coefficient, denoted $\hat{h}_n \in \mathbb{R}^2$, containing its real and imaginary components. The backscatter demodulation branch outputs the predicted probability of the backscatter bit at each subcarrier, denoted $\hat{b}_n \in (0, 1)$. Both predictions are obtained using feed-forward heads on top of the final task-specific decoder representations.

Training loss. The model is trained using a multitask objective that jointly supervises channel estimation and backscatter demodulation. Since both tasks operate on the same sequence of subcarriers, a shared-loss formulation forces the network to learn representations that serve both objectives simultaneously, rather than treating them as independent prediction problems.

Since it is difficult to measure the ground-truth channel, we supervise the channel estimation task to minimize the difference between the transmitted pre-known symbols and the symbols reconstructed from the received signal and the estimated channel. While we can use conventional estimators, such as least-squares (LS), they can yield inaccurate channel estimates, especially at low SNR. Hence, we simply rely on reconstructing the transmitted signal as our way to supervise the task. Let $Y_n^{(i)}$ denote the received complex sample i on subcarrier n , and let X_n be the known transmitted signal. The model outputs the real and imaginary parts of the channel on each subcarrier, $\hat{H}_n^{(i)} = \hat{h}_{n,\text{re}}^{(i)} + j\hat{h}_{n,\text{im}}^{(i)}$. The loss compares the received sample $Y_n^{(i)}$ and the reconstructed sample $X_n\hat{H}_n^{(i)}$ as

$$\mathcal{L}_{\text{CE}} = \frac{1}{BN} \sum_{i=1}^B \sum_{n=1}^N \left(\left| \Re\{Y_n^{(i)} - X_n\hat{H}_n^{(i)}\} \right|^2 + \left| \Im\{Y_n^{(i)} - X_n\hat{H}_n^{(i)}\} \right|^2 \right) \quad (3)$$

For tag bit demodulation, the goal is to estimate the BPSK bit transmitted by the backscatter tag on each subcarrier index. Let $b_n \in \{0, 1\}$ denote the reference bit and \hat{b}_n the predicted probability. We adopt a standard binary cross-entropy loss:

$$\mathcal{L}_{\text{TB}} = -\frac{1}{BN} \sum_{i=1}^B \sum_{n=1}^N \left(b_n^{(i)} \log \hat{b}_n^{(i)} + (1 - b_n^{(i)}) \log (1 - \hat{b}_n^{(i)}) \right) \quad (4)$$

The final training objective is formed by weighting the two task-specific losses, reflecting the fact that channel estimation and bit demodulation contribute differently to the overall learning. After empirical tuning, we assign separate coefficients λ_{CE} and λ_{TB} to balance the two loss terms to the combined loss function $\mathcal{L} = \lambda_{\text{CE}} \mathcal{L}_{\text{CE}} + \lambda_{\text{TB}} \mathcal{L}_{\text{TB}}$.

4.4 Reducing model size and complexity

While the dual-task transformer model can address Challenges C2 (dependency on channel estimation) and C3 (lack of synchronization), the resulting trained model can still be large, still resulting in a high time overhead of demodulation (Challenge C1). Hence, we first analyze the time complexity of our dual-task model and devise solutions to reduce it.

Dual-task transformer complexity. To make the comparison concrete, we estimate the cost of the dual-task transformer using explicit FLOP counts. For a sequence of length N and embedding dimension d_{model} , one self-attention layer requires approximately $4Nd_{\text{model}}^2 + 4N^2d_{\text{model}} + 2N^2$ operations [1, 58] (projection of

Metric	Transformer (Pre-pruning)	Transformer (Post-pruning)	GA
FLOPs	9.6 G	2.5 G	2 T
Memory	6.8 MB	1.3 MB	100 MB
Arithmetic Intensity	1.4K Flops/byte	1.9K Flops/byte	20K Flops/byte

Figure 6: Complexity comparison between the proposed transformer (pre- and post-pruning) and a GA baseline to achieve BER of 10^{-3} for 1584 subcarriers.

queries/keys/values, dot products, softmax, and output projection). The position-wise feed-forward network with hidden dimension d_{ff} contributes about $4Nd_{\text{model}}d_{\text{ff}}$ FLOPs (two linear layers with ReLU). Thus, one encoder layer costs $4Nd_{\text{model}}^2 + 4N^2d_{\text{model}} + 2N^2 + 4Nd_{\text{model}}d_{\text{ff}}$ FLOPs. In addition to the shared encoder, the two shared attention layers in the decoder require twice as many FLOPs as a single self-attention layer and position-wise feed-forward network. The two task-specific decoders are single linear projections with output dimensions d_A (channel estimation) and d_B (tag-bit prediction), which add about $2Nd_{\text{model}}d_A$ and $2Nd_{\text{model}}d_B$ FLOPs, respectively. Overall, the dual-task transformer therefore performs approximately the following number of FLOPs per inference instance:

$$L \left(4Nd_{\text{model}}^2 + 4N^2d_{\text{model}} + 2N^2 + 4Nd_{\text{model}}d_{\text{ff}} \right) + 2Nd_{\text{model}}(d_A + d_B) \quad (5)$$

For our configuration ($N = 1584$, $d_{\text{model}} = 192$, $d_{\text{ff}} = 768$, $L = 4$, $d_A = 2$, $d_B = 2$), this is about 9.6 GFLOPs for the full (pre-pruning) model, shown in the “Transformer (pre-pruning)” column of Fig. 6. For GA baseline, each candidate backscatter bit vector b is evaluated by computing the objective $J(b) = \sum_{n=1}^N |y(n) - \hat{y}(n; b)|^2$ over n subcarriers. For every sample n , this involves roughly 8 FLOPs. With P candidates and N samples, one GA generation requires $8PN$ FLOPs. To reach the same BER (10^{-3}) as our dual-task transformer, the GA needs on average $G = 800\text{K}$ generations. With $N = 1584$ and $P = 200$, this amounts to 2 TFLOPs per OFDM symbol, nearly three orders of magnitude higher than our transformer (Fig. 6). The GA also stores all candidate waveforms across generations, leading to $\mathcal{O}(NPG)$ memory (10^{11} bytes). As a result, its arithmetic intensity is much higher (20 KFLOPs/byte vs. 1.9 KFLOPs/byte), making it significantly more compute-bound and impractical.

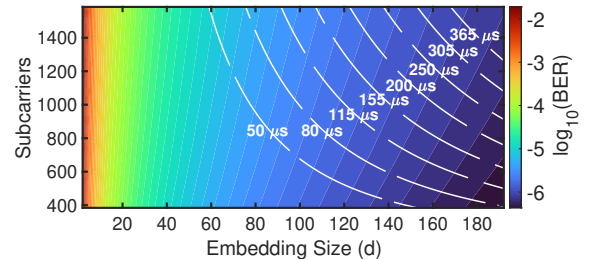


Figure 7: Model accuracy and computational cost (μs) across embedding sizes and subcarriers.

Embedding size. From Equ. 5, the embedding size (d_{model}) is a key contributor to the FLOPs of our dual-task transformer. Fig. 7 illustrates the cost–accuracy trade-off across subcarrier configurations and embedding sizes, showing $\log_{10}(\text{BER})$ with overlaid

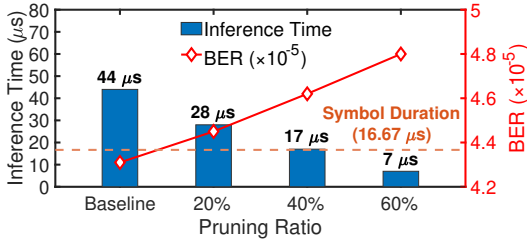


Figure 8: Improving the model efficiency (inference time) of our multi-head shared-attention transformer.

contours of inference latency (μs). Increasing the embedding size improves BER but also increases latency, while subcarrier settings further affect this trade-off. These results guide the selection of the embedding size to balance performance and computational cost for different OFDM configurations.

Pruning. To reduce model complexity and improve inference efficiency, we adopt a structured pruning strategy based on Fisher Information, inspired by [31]. The key idea is to measure each parameter’s importance by its impact on the loss. During a forward-backward pass, squared gradients are accumulated to compute importance scores. After normalization, parameters are ranked, and those below a threshold determined by the pruning ratio are removed. We first fine-tune the full dual-task transformer and compute Fisher information on the calibrated model. We then apply structured pruning by removing the lowest-ranked attention heads and FFN neurons (up to 75%), while preserving the shared encoder and task-specific decoders. Parameters below the pruning threshold are zeroed out without changing the architecture. This post-training pruning retains key weights while reducing computation and memory, with a brief fine-tuning stage to stabilize performance. As shown in Fig. 6, after pruning, our dual-task model requires 2.5 GFLOPs, enabling it to run on edge devices such as a smartphone UE and produce inference time much closer to the actual OFDM symbol duration. Fig. 8 shows how different levels of pruning affect the FLOPs and BER for 1584-subcarrier configuration, where the OFDM symbol duration is $16.67 \mu\text{s}$. The results show that pruning substantially lowers both FLOPs and inference latency, dropping from $44 \mu\text{s}$ (unpruned) to $7 \mu\text{s}$ at 60% pruning with more than a $6\times$ reduction in computational load. Although pruning introduces a small BER increase (on the order of 10^{-5}), the degradation is modest relative to the gains in efficiency.

5 Tag Design and Operations

WiNB tag embeds data onto the 5G NR frame by modulating the phase of the reflected signal. To avoid modifying critical control and channel estimation information, the tag must be synchronized with the transmitted frame by detecting the SSB. However, such synchronization is challenging because 5G NR configurations span a wide range of bandwidths and subcarrier spacings, which in turn alter the SSB symbol duration. The SSB comprises four key elements. The Primary Synchronization Signal (PSS) allows the UE to achieve coarse time synchronization and identify the cell group. The Secondary Synchronization Signal (SSS) completes the cell-ID determination and provides frame alignment. The Physical Broadcast Channel (PBCH) carries the basic system-configuration information needed for initial access. Finally, the PBCH Demodulation Reference

Signal (PBCH-DMRS) provides the channel estimates needed for reliable PBCH decoding. Under the tag’s limited-power constraints, achieving reliable synchronization is difficult. To address this, we propose an efficient SSB detection method using a power detector (an ADL6010 power detector).

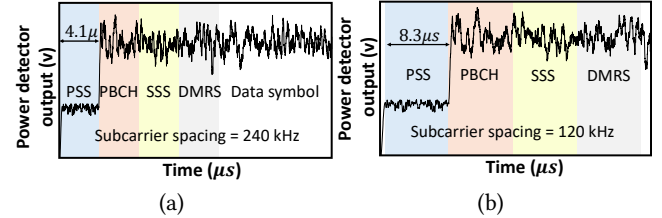


Figure 9: The power detector produces a distinctive PSS pattern whose duration depends on the SCS, enabling SSB detection without modification.

WiNB tag receives a distorted version of the 5G NR signal from the power detector due to the detector’s narrow bandwidth and loss of phase information. Nonetheless, two unique properties of the PSS enable reliable SSB detection: (i) periodicity of the PSS and (ii) a distinctive pattern. As described in Sec. 2, the PSS is transmitted periodically, creating a repeating structure in time. This periodicity provides a predictable temporal reference that the tag can look for even when the waveform is heavily distorted. Unlike other OFDM symbols, the PSS consists of 127 M-sequences whose energy is spread almost uniformly across the entire spectrum, similar to white noise. When this wideband sequence passes through a power detector, it produces a sudden amplitude drop, with a duration that depends on the SCS as shown in Fig. 9. In FR2 systems, SSBs are primarily used for beam management and initial access. According to the 3GPP NR specification [2, 11], the PSS occupies only subcarriers 56 to 182 within the SSB resource grid, while the remaining subcarriers are set to zero or unused for PSS transmission. This enables reliable SSB detection in our case. In practice, a base station can also intentionally leave the data symbols empty during the SSB transmission interval, allowing low-power devices such as IoT tags to detect SSBs.

Our tag detects this drop by monitoring the autocorrelation of the power-detector output, where the magnitude sharply decreases at the PSS boundary. Using the known PSS length, it isolates the full SSB ($\text{PSS}\times 4$). Since SSB content is relatively static, the tag applies cross-correlation across consecutive SSBs to improve reliability. We implement this using an ADL6010 power detector to capture 5G NR waveforms and identify SSB locations, achieving 97% detection accuracy. The resulting processing delay introduces minor timing misalignment, which our transformer handles during backscatter demodulation. Once the SSB is detected, the tag modulates the remaining symbols in the slot. Since the first 4 OFDM symbols are reserved for SSB, the tag uses the remaining 10 for data and performs sample-level modulation by toggling its reflection state at a high switching rate. As a result, the number of modulated samples per second is directly determined by the tag’s switching rate f_{sw} . The theoretical throughput can thus be approximated as $R_{\text{max}} = \frac{s}{t} f_{\text{sw}}$, where s is the number of OFDM symbols used for data modulation, and t is the total number of symbols including SSB.

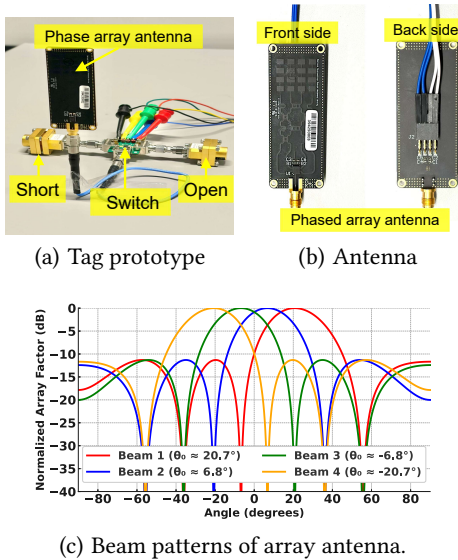


Figure 10: Prototype of our 28 GHz 5G NR backscatter tag.

6 Implementation and Evaluation

6.1 Prototypes, test environments, & model training

5G NR backscatter prototype. We implement 5G NR backscatter tag prototype, manufactured on a Rogers RO4003C RF laminate, as shown in Fig. 10(a). The tag consists of (i) an ADRF5021 SPDT silicon switch with a 100 MHz switching rate, (ii) SWO-34-F1 waveguide short and open terminations connected to the switch to produce 0° and 180° reflection phases for modulation, and (iii) a phased-array antenna with a passive beamforming network (described below). The tag is controlled by a TerasIC Cyclone V FPGA.

We design a linear patch-array antenna to provide high-gain directional transmission and reception at 28 GHz, as illustrated in Fig. 10(b). The antenna comprises a 4×4 Butler beamforming matrix that feeds a 4×4 patch array with an element spacing of 5.687 mm ($\approx 0.53\lambda_0$). Both the Butler matrix and the array are optimized using the finite integration technique (FIT) solver in CST Microwave Studio [14] to achieve a wide impedance bandwidth and desirable radiation characteristics. To achieve low dielectric loss and high antenna radiation efficiency, a 12-mil-thick Rogers Ro4003C RF laminate was used as an RF substrate. The Butler network is interfaced with an Analog Device ADRF5045BCCZN SP4T RF switch to change beam patterns according to the FPGA control. Fig. 10(c) shows the four normalized array factors resulting from the Butler beamformer. The tag antenna (Tx and Rx) gain is 13.6 dBm after counting for the beamformer circuit loss.

BS/UE prototypes. Both BS (Tx) and UE (Rx) use USRP X310 [44] with a SiversIMA EVK02004 RF front-end. For experiments, we select BS, UE, and tag beam combinations that maximize backscatter SNR. The SiversIMA front-end limits EIRP to 47 dBm. We generate 3GPP-compliant 5G NR FR2 waveforms in MATLAB with full SSB implementation and varied numerology/subcarrier settings.

Test environments. We evaluate our system in three indoor and two outdoor settings (Fig. 12) with diverse layouts and clutter levels, resulting in varying multipath conditions. Room 1 is a 12×10 m

open lab, Room 2 an 8×20 m corridor, Room 3 an 8×6 m conference room, Outdoor 1 a 22×14 m roof, and Outdoor 2 a 35×15 m alley. These settings enable evaluation across diverse environments.

Training and Testing Methodology. The model is implemented in PyTorch and trained with AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$). The learning rate starts at 1×10^{-3} and follows a cosine annealing schedule with a short warm-up. We use a batch size of 256 and train for 120 epochs unless otherwise noted. Hyperparameters are tuned on Room 1 and then fixed across all experiments (in-domain, cross-room, and fine-tuning) for fair comparison. The dataset includes measurements from multiple indoor and outdoor environments. Room 1 provides ~ 180 K samples for training, augmented with 20K synthetic samples via random synchronization offsets up to $10\mu\text{s}$. Rooms 2 and 3 each contain ~ 40 K samples for training and testing to evaluate cross-room generalization. Two outdoor environments (Outdoor 1/2) each contribute ~ 200 K samples. In total, the dataset comprises over 700K real measurements. The outdoor data introduce more diverse channel conditions and are used to evaluate cross-environment generalization and mixed (indoor + outdoor) training. We train the model on Nvidia RTX 5090 GPU [41] and test it on different UE platforms (MacBook M3 [4], Nvidia Jetson Orin Nano board [40], Google Pixel 10 smartphone [22]) after pruning and other optimizations. Testing is performed strictly on data that are not used during training or hyperparameter tuning.¹

6.2 Microbenchmarks

Distance, SNR, and BER. We first evaluate WiNB at different Tx-tag-Rx distances in the three test setups. Here, the Tx, Rx, and tag are placed at arbitrary distances and angles. Fig. 11(a) shows the SNR over distance. The Tx, Rx, and tag are placed in a triangle geometry. The Tx-tag and tag-Rx distances are always kept the same, ranging from 1–6 m. The distances in the Fig. 11(a) refer to the total round-trip distance from Tx to tag to Rx. We find that WiNB can maintain a high SNR over distances up to 12m, thanks to the high-gain beamforming phased-array antenna on the tag. Fig. 11(b) shows the backscatter BER over different distances for different subcarrier configurations. We find that WiNB can maintain a low BER of 10^{-4} , which is a significant improvement over previous mmWave backscatter systems [10]. Configurations with fewer subcarriers achieve better BER as fewer backscatter bits are embedded in each OFDM symbol. It is important to note that the FCC maximum allowed EIRP in 5G NR FR2 can reach 75 dBm per 100 MHz [18], depending on deployment conditions. Therefore, future implementations using higher transmit power and/or antenna gain can significantly extend the achievable communication range.

Throughput. Fig. 11(c) shows how these SNR and BER translate into end-to-end throughput across distances. Here, the backscatter tag employs BPSK modulation (1 backscatter bit per OFDM sample) with a 100 MHz switching rate. Furthermore, the tag modulates data symbols to avoid affecting SSB symbols, thereby affecting the observed throughput. We find that our observed throughput T approximately follows $T \approx \frac{10}{14} \times 100 \text{ MHz} (1 - \text{BER})$. WiNB achieves a throughput of 71 Mbps, a $7\times$ improvement over the state-of-the-art OFDM backscatter system [43].

¹<https://github.com/Wireless-IoT-Sensing-Lab/Wideband-5G-NR-Backscatter/>

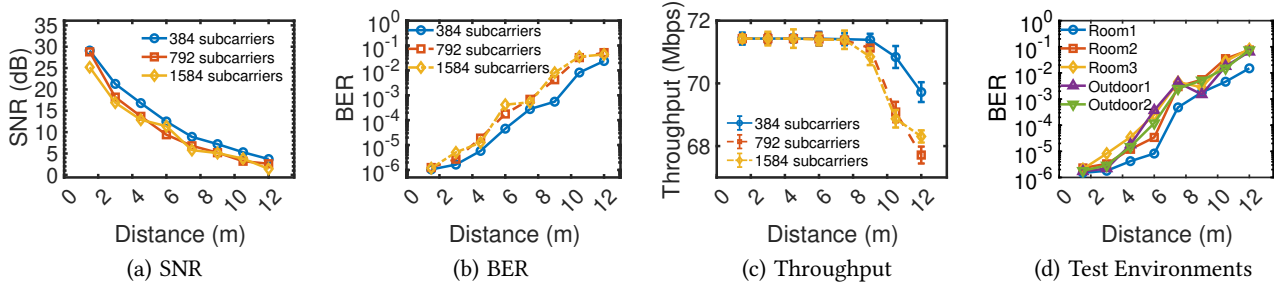


Figure 11: Microbenchmarking WiNB backscatter demodulation at different distances.

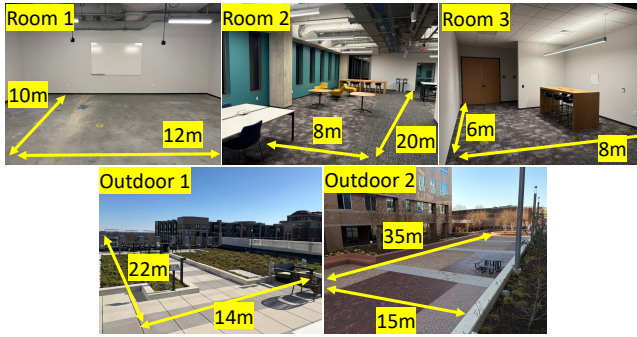


Figure 12: Test environments: Room 1 (lab), Room 2 (corridor), Room 3 (conference room), Outdoor 1 (building roof), and Outdoor 2 (alley)

Different environments. Fig. 11(d) shows the BER across different environments. We evaluate WiNB in multiple real-world settings with varying propagation characteristics, including indoor and outdoor scenarios with different levels of multipath and blockage. We observe that WiNB consistently maintains a low BER across these environments, demonstrating its robustness to environmental variations. For example, Room 1 and Outdoor environments achieve very low BER, especially at shorter distances, due to fewer multipaths. Environments with richer multipath (like Rooms 2 and 3) introduce more channel diversity and self-interference. Even in such cases, our dual-task model can reliably decode the backscatter data. These results suggest that WiNB can operate accurately under diverse real-world conditions.

6.3 Comparison with State-of-the-art

Fig. 13 compares WiNB’s transformer-based demodulation with GA and QN based demodulations. Given that both GA and QN are iterative methods, we first measure the run time for our WiNB demodulation and run the GA and QN for the same duration on the same platform for a fair comparison. Also, standard least-squares and interpolation-based 5G NR channel estimation methods are used for GA and QN. We find that WiNB achieves orders of magnitude lower BER than the other two methods, thanks to its learned function estimator. The difference becomes even more pronounced as the number of subcarriers increases, affecting both channel estimation and the backscatter solution search space.

Fig. 14 compares the throughput and BER of WiNB with three other representative state-of-the-art commodity backscatter systems. mmComb [10] is a mmWave WiFi backscatter solution that

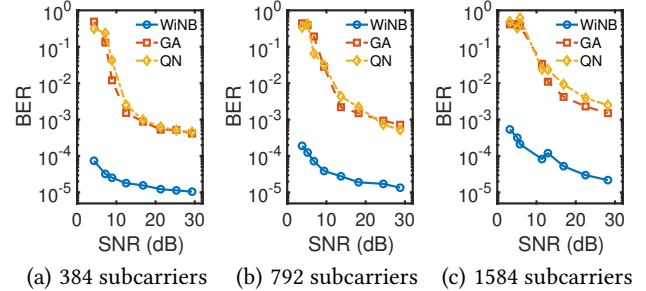


Figure 13: Comparing WiNB with QN and GA based demodulation for different subcarrier configurations.

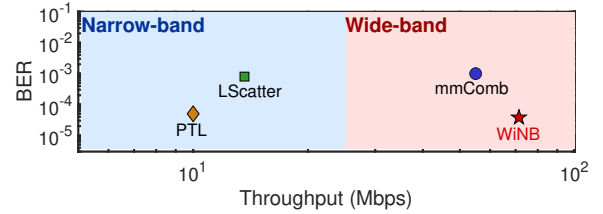


Figure 14: Throughput–BER comparison with state-of-the-art commodity backscatter schemes [10, 13, 43] when the Tx-tag-Rx distance is 7 meters (albeit for different environments and deployment settings used).

achieves a high data rate but a much higher BER. Lscatter [13] leverages the LTE OFDM signal for backscattering. While it achieves a better BER, it is limited to 20 MHz of bandwidth, resulting in much lower throughput. Lastly, PTL [43] is a state-of-the-art WiFi backscatter system. Similar to Lscatter, it achieves lower BER but is limited to 20 MHz channels and 64 subcarriers. Since both Lscatter and PTL use QN and GA based demodulation, their BER increases dramatically when more subcarriers are used.

Demodulation time. Table 2 shows the time for backscatter demodulation of one OFDM symbol with a target BER of 10^{-3} . We run our model on MacBook M3 (CPU only) along with the GA and QN methods. We find that our inference time is as low as $16 \mu\text{s}$, compared to tens of, or even hundreds of, ms for GA and QN methods. This is, on average, a three orders-of-magnitude improvement. In fact, the demodulation time for WiNB is very close to the actual OFDM symbol duration, making it possible to even demodulate the backscatter bits in real-time. To improve the fairness of comparison, we also implement the GA and QN baselines with GPU acceleration using NVIDIA RTX 5090. Specifically, we adopt a JAX-based GPU-accelerated genetic algorithm implementation [15] and

# Subcarriers	WiNB	GA		QN	
		CPU	GPU	CPU	GPU
384	16 μ s	28 ms	9 ms	26 ms	3 ms
792	38 μ s	64 ms	18 ms	52 ms	7 ms
1584	52 μ s	319 ms	75 ms	108 ms	24 ms

Table 2: Demodulation latency per OFDM symbol across different decoding approaches on CPU and GPU platforms.

a GPU implementation of the L-BFGS-B optimizer [19]. While GPU execution reduces their runtime (by 3–8 \times), their latency remains in the tens-of-milliseconds range and is still significantly higher than that of WiNB, due to the algorithms’ higher complexity and memory-access overhead.

We directly deploy our model on two additional platforms: (1) A high-end GPU, NVIDIA RTX 5090 [41], and (2) a Jetson Orin Nano board [40]. Table 3 shows the demodulation times. Additionally, we estimate the inference time on Google Pixel 10 smartphone [22]. Since our model cannot be directly deployed on the phone, we estimate its inference time to be 60–120 μ s using model profiling and latency prediction based on the latency lookup tables [30, 59]. We observe that across all platforms, including edge devices, the backscatter demodulation time per OFDM symbol remains in the tens of μ s, whereas existing methods yield milliseconds of latency. **Impact of unsynchronization and data augmentation.** We now validate our earlier claims that our transformer-based model can better tolerate the unsynchronization between the transmitted OFDM signal and the backscatter signal. Fig. 15(a) shows how increasing the synchronization offset from perfectly aligned samples to up to a 10 μ s mismatch affects the three demodulation methods. We use 1584-subcarrier setting here with symbol duration being 16.67 μ s. This means that at a synchronization error of 5 μ s, approximately 30% of backscatter samples are not aligned with their corresponding OFDM symbol. We find that WiNB’s transformer can tolerate the synchronization errors much better by maintaining the low BER. On the other hand, the BER sharply increases for GA and QN, converging to an unacceptable level. As discussed before,

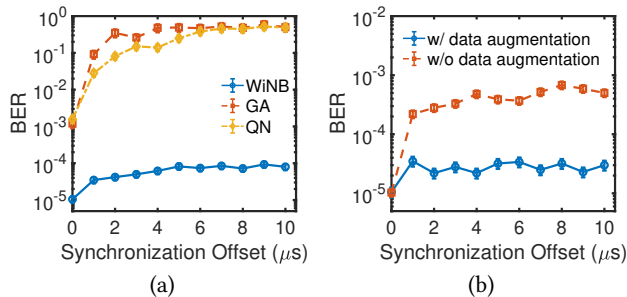


Figure 15: (a) Impact of synchronization offset on BER, and (b) model performance with and without data augmentation through synthetic unsynchronized samples.

our model is trained on both synchronized and unsynchronized instances of backscatter data embedded in OFDM symbols. This not only makes the model robust to synchronization errors during inference but also helps it converge faster during training without requiring extensive measurements. Fig. 15(b) shows the impact of

Platform	Per-symbol Inference Latency
RTX 5090	1 μ s
MacBook M3 (CPU)	16 μ s
Jetson Orin Nano	43 μ s

Table 3: Measured and estimated per-symbol inference latency for the pruned model on different platforms.

this data augmentation on BER with different synchronization offsets. As expected, we find that the model with augmentation from unsynced data helps it to tolerate synchronization offsets better.

6.4 Model complexity and generalization

Need for dual-task learning. Our key insight in this work is that channel estimation and backscatter demodulation are tightly coupled tasks, and jointly resolving them through learning can improve the performance. To validate this claim, we compare our

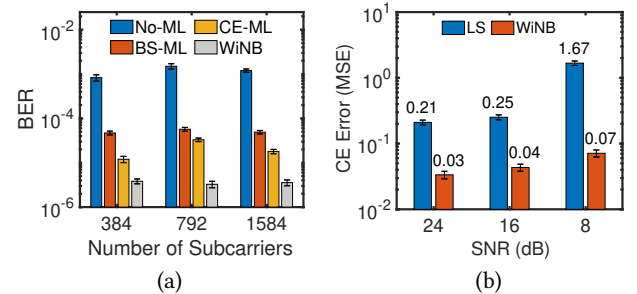


Figure 16: (a) Dual task model comparison with single task models, and (b) channel estimation performance in our dual-task model. Error bars show 95% confidence intervals over repeated measurements.

dual-task learning model with three variations: (1) No-ML: here, both channel estimation and backscatter demodulation are performed in sequence through non-ML methods (least-squares with interpolation for channel estimation and QN for backscatter modulation). This is the baseline we have used in the results discussed so far. (2) BS-ML: Here, the channel estimation is performed through least-squares (no CE task in our model), but the backscatter demodulation is performed through a single-task transformer (trained separately). (3) CE-ML: Here, the channel estimation is performed through a single-task transformer, but once the channel estimate is available, the QN method is used for demodulating the backscatter bits. This ablation helps us understand how much the dual-task model can help improve performance. Fig. 16(a) shows the comparison of four methods for different subcarrier configurations. We observe that the dual-task model consistently outperforms both single-task models, demonstrating its ability to model dependencies between the two tasks via shared attention in the transformer.

We also evaluate how our dual-task model improves channel estimation along with backscatter BER. We compare WiNB’s channel estimation with the default 5G NR standard channel estimation technique based on least squares (LS) and interpolation on DM-RS symbols. Fig. 16(b) compares channel estimation errors. Since we do not have the ground truth channel in our experiments, we compare the original pilot symbols with the received pilot symbols equalized using our estimated channel. This is similar to our loss

Case	Environment	Training	Testing	BER
(1)	Indoor Only	R1	R2	3.2×10^{-4}
(2)		R1 + R3	R2	3.1×10^{-4}
(3)		R1 + R2 + R3	R2	1.2×10^{-4}
(4)	Outdoor Only	O1	O2	2.7×10^{-4}
(5)		O1 + O2	O2	1.5×10^{-4}
(6)	Indoor + Outdoor	R1 + R2 + R3 + O1 + O2	R2	1.1×10^{-4}
(7)		R1 + R2 + R3 + O1 + O2	O2	0.9×10^{-4}

Table 4: BER across indoor, outdoor, and indoor+outdoor environments. Training with both indoor and outdoor data provides the best generalization performance.

function for the channel estimation task in Equ. 3. We find that our transformer’s channel estimation task achieves much better channel estimates than LS. This is in line with results from other ML-based channel estimator works [12, 32, 47] that have shown a better performance than the default methods.

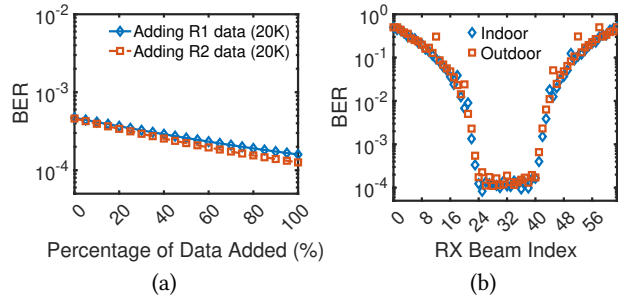


Figure 17: (a) Two training augmentations with testing in Room 3, and (b) Impact of suboptimal beam selection under indoor and outdoor environments.

Generalization across different environments. We evaluate how WiNB’s transformer generalizes across indoor, outdoor, and indoor+outdoor environments. Since the dual-task design separates channel effects from backscatter modulation within each OFDM symbol, it generalizes well across diverse conditions. Table 4 shows three indoor cases (R1–R3), with 200K training and 20K testing samples. Training on R1 and testing on R2 yields BER of 10^{-4} ; adding R3 data (while keeping training size fixed) slightly improves BER due to increased diversity. Adding same-domain data (R2) further reduces BER, though gains are modest, indicating limited reliance on site-specific data. We extend this to outdoor settings (O1, O2) and mixed training. Outdoor-only training achieves comparable BER to indoor cases, while combining indoor and outdoor data yields the best performance, reducing BER across both test settings and improving robustness to diverse propagation conditions, including richer multipath and environmental variations. Fig. 17(a) shows how BER changes when the model is augmented with additional training data. We train the model using Room 1 (160K samples) and Room 2 data (40K samples), and test it on Room 3 samples (20K) while gradually adding more training data in two following ways: (i) Out-of-domain: adding 20K more samples from Room 1, and (ii) In-domain: adding 20K more samples from Room 2. We find that both cases yield almost comparable BER, and the model appears to

be mostly benefiting from having more data (in or out-of-domain) in training as opposed to having more in-domain data. Overall, these results show that WiNB generalizes well not only across different indoor environments but also across indoor and outdoor domains. With newly available diverse channels and backscatter data, it can achieve much lower BER in practice.

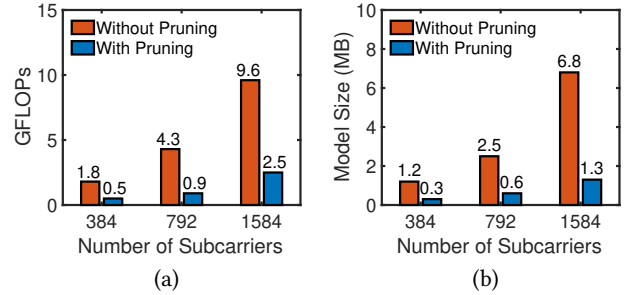


Figure 18: Model complexity: (a) Computational cost, reported as GFLOPs per inference; and (b) Model size, i.e., the memory footprint required for each configuration.

Impact of Mobility and Beam Drift. To evaluate the robustness of WiNB under practical deployment conditions, we conduct experiments by selecting different receive beams to emulate beam misalignment due to UE mobility. Specifically, we sweep over the beam indices defined by the Sivers codebook at the UE, which provides 64 discrete beam directions, with the main lobe angle between adjacent beams being 1.6° , to study how performance varies with beam drift. As shown in Fig. 17(b), we observe that a contiguous range of beam indices achieves low BER (on the order of 10^{-4}) when the UE beam mainlobe still remains under the tag’s transmit antenna beamwidth. Within this region, the performance remains stable even under moderate beam misalignment, indicating robustness to small UE movements. As the selected beam moves away from this region, the BER gradually increases as expected, and eventually approaches a high error rate when the beam no longer aligns with the signal direction.

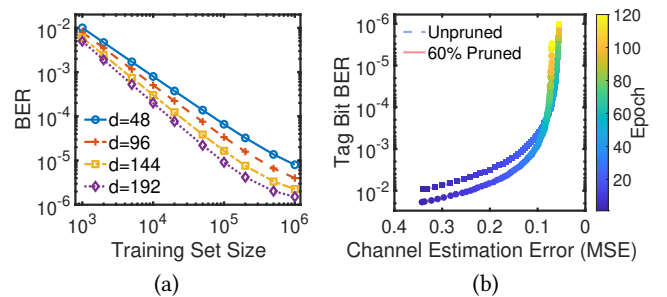


Figure 19: (a) BER under increasing training data for different embedding sizes, and (b) Relationship between channel estimation accuracy and tag-bit BER across training epochs for both baseline and pruned models.

Model complexity and size. To study the scalability of the model in different OFDM configurations, we evaluated computational and memory costs in three subcarrier settings (384, 792, and 1584), using

embedding dimensions of 24, 48, and 192, respectively, to maintain a similar BER. Fig. 18(a) reports the required GFLOPs per symbol (with and without pruning), showing a steady increase as the input dimension grows, rising from roughly 0.5 GFLOPs at 384 subcarriers to about 2.5 GFLOPs at 1584. The corresponding model size, shown in Fig. 18(b), also increases with input dimensionality but remains compact (0.3-1.3 MB). Both FLOP and memory size values observed from Pytorch match with our complexity analysis in Fig. 6 within a 1.4% error margin.

Training size and epochs. To evaluate the efficiency of our multi-head shared-attention transformer, we study how model accuracy and computational cost evolve across different architectural and training configurations. We investigate how the model benefits from increased training data in Fig. 19(a). We find that a much lower BER can, in fact, be achieved if the model is trained with much more data than we have used so far ($\approx 10^6$). This means the model's performance can be further improved by incrementally training it with additional data from diverse environments. To gain insight into the learning dynamics, Figure 19(b) characterizes the relationship between channel-estimation accuracy and tag-bit BER across training epochs, and contrasts the full model with its 60%-pruned variant, illustrating that most performance is preserved even after substantial structural reduction. Together, the proposed architecture achieves high accuracy while remaining highly scalable, and pruning provides an effective mechanism for meeting real-time constraints without compromising performance.

6.5 Tag Power consumption

The WiNB tag consists of four key components: a clock generator, an SSB detector, a backscatter modulator, and an SPDT switch. We evaluate the power consumption of these components using Libero SoC SmartPower [33]. The clock, generated by a chain of serial inverters at 100 MHz, drives both the modulator and the frame detector. It consumes $13.6 \mu W$ of power. The backscatter modulator, which provides the control signals to the SPDT switch for embedding backscatter bits onto the carrier signal, consumes $4 \mu W$. The SSB frame detection requires $139.6 \mu W$ of computational power. The ADRF5021 SPDT switch control used for phase modulation consumes $264 \mu W$, while the ADRF5045 SP4T switch control for beam-pattern consumes $372 \mu W$. We further validate the power consumption of key RF components through measurements using a Monsoon power monitor [48]. Our measurements show that the ADRF5021 switch consumes 1.83 mW in the static state and 2.15 mW during dynamic operation, while the ADL6010 power detector consumes 10.72 mW (static) and 10.98 mW (dynamic). These measurements are higher than the circuit-level estimates, as they include practical overheads such as biasing, RF interfacing, and measurement conditions. Overall, while the measured RF front-end components incur additional power consumption, the total system power consumes 13.59 mW. This is comparable to prior mmWave backscatter systems [36, 42]. In particular, MilBack [36] reports a total power consumption of 18 mW during localization and downlink, and 32 mW during uplink, while BiScatter [42] reports a total system power consumption of 48 mW. These results demonstrate that WiNB maintains a low-power design suitable for practical backscatter deployments.

7 Related Works

Ambient backscatter. Conventional backscatter systems have been studied for more than two decades [20, 25, 39] and are primarily designed for ultra-low-power communication. However, the high cost of deploying dedicated RFID readers [21] has driven the development of ambient backscatter systems, which leverages existing RF signals such as TV [34], WiFi [3, 7, 27, 28, 61], LTE [13], IoT signals [54, 55, 60], and even radar transmissions [5, 50]. Early ambient backscatter efforts primarily focused on sub-6 GHz WiFi signals due to their favorable adaptability. More recently, mmWave backscatter systems [5, 9, 10, 29, 36–38, 46] have received significant recent attention motivated by the large bandwidth and potential for high-speed data rates at the mmWave band. Many of these mmWave backscatter designs [5, 36, 46] employ mmWave FMCW radars as readers, which enable precise distance and Doppler estimation for sensing and localization. However, radar-based readers are not well-suited for high data rate applications.

OFDM-based backscatter. OFDM-based backscatter has emerged as a compelling alternative because OFDM is widely deployed and supports high-throughput communication via its orthogonal subcarriers. Early systems [3, 17, 63] embed only a single bit per OFDM symbol, simplifying demodulation but underutilizing OFDM's frequency-domain resources and limiting throughput. Recent efforts [13, 35, 43, 57] adopt sample-level modulation to fully exploit the subcarrier structure and improve performance. While sample-level OFDM designs improve data rates, they incur high demodulation complexity along with synchronization and channel estimation challenges. In contrast, our dual-task model jointly reduces complexity and resolves synchronization, enabling near real-time backscatter demodulation with compatibility for wide-area 5G NR deployments.

8 Conclusion and Discussion

The paper presents a novel 5G NR OFDM backscatter system that achieves low complexity of backscatter demodulation, low BER, and high throughput. It advances the state of the art by developing a dual-task transformer that can be implemented on a backscatter receiver, thanks to its high efficiency and small memory footprint. Our current design opens up several directions for future exploration. (1) *Applications*: While the high-throughput backscatter design enables applications such as backscatter video streaming, a complete end-to-end system requires additional components, including sensing, video capture, and efficient encoding. Integrating these modules and optimizing the full pipeline for real-time operation remain important directions for future work. (2) *UE Mobility and Beamforming*: Although our results show robustness under moderate mobility and suboptimal beam selection, more dynamic scenarios with fast user movement and beam drift remain to be explored. Future work can investigate adaptive beam tracking and joint optimization of beamforming and backscatter modulation to further improve reliability under mobility. (3) *COTS UEs*: Our current prototype requires access to OFDM IQ samples, which are not always available on COTS UEs, and thus does not represent a typical UE deployment. Extending our current techniques to data and interfaces commonly available on COTS UEs remains an outstanding challenge for practical deployment.

References

- [1] A guide to LLM inference and performance – baseten.co. <https://www.baseten.co/blog/llm-transformer-inference-guide/>. [Accessed 05-12-2025].
- [2] 3GPP. Nr; physical channels and modulation. Technical Report TS 38.211, 3rd Generation Partnership Project (3GPP), 2019. Release 15, Table 7.4.3.1-1.
- [3] Ali Abedi, Farzan Dehbashi, Mohammad Hossein Mazaheri, Omid Abari, and Tim Brecht. Witag: Seamless wifi backscatter communication. In *Proceedings of the Annual conference of the ACM Special Interest Group on Data Communication on the applications, technologies, architectures, and protocols for computer communication*, pages 240–252, 2020.
- [4] Apple. Apple m3 pro chip specifications. <https://support.apple.com/en-us/117736>, 2023. Accessed: 2025-12-05.
- [5] Kang Min Bae, Namjo Ahn, Yoon Chae, Parth Pathak, Sung-Min Sohn, and Song Min Kim. Omniscatter: extreme sensitivity mmwave backscattering using commodity fmcw radar. In *Proceedings of the 20th Annual International Conference on Mobile Systems, Applications and Services*, pages 316–329, 2022.
- [6] Dmzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.
- [7] Dinesh Bharadia, Kiran Raj Joshi, Manikanta Kotaru, and Sachin Katti. Backfi: High throughput wifi backscatter. *ACM SIGCOMM Computer Communication Review*, 45(4):283–296, 2015.
- [8] Rich Caruana. Multitask learning. *Machine Learning*, 1997.
- [9] Yoon Chae, Kang Min Bae, Parth Pathak, and Song Min Kim. On the feasibility of millimeter-wave backscatter using commodity 802.11 ad 60 ghz radios. In *Proceedings of the 14th International Workshop on Wireless Network Testbeds, Experimental evaluation & Characterization*, pages 56–63, 2020.
- [10] Yoon Chae, Zhenzhe Lin, Kang Min Bae, Song Min Kim, and Parth Pathak. mm-comb: High-speed mmwave commodity wifi backscatter. In *21st USENIX Symposium on Networked Systems Design and Implementation (NSDI 24)*, 2024.
- [11] Liang Chen, Xin Zhou, Feifei Chen, Lie-Liang Yang, and Ruizhi Chen. Carrier phase ranging for indoor positioning with 5g nr signals. *IEEE Internet of Things Journal*, 9(13):10908–10919, 2021.
- [12] Zhuolin Chen, Fanglin Gu, and Rui Jiang. Channel estimation method based on transformer in high dynamic environment. In *2020 International Conference on Wireless Communications and Signal Processing (WCSP)*, pages 817–822. IEEE, 2020.
- [13] Zicheng Chi, Xin Liu, Wei Wang, Yao Yao, and Ting Zhu. Leveraging ambient lte traffic for ubiquitous passive communication. In *Proceedings of the Annual conference of the ACM Special Interest Group on Data Communication on the applications, technologies, architectures, and protocols for computer communication*, pages 172–185, 2020.
- [14] Dassault Systèmes. Cst studio suite - electromagnetic field simulation software. <https://www.3ds.com/products/simulia/cst-studio-suite>, 2024. Accessed: 2024-06-24.
- [15] Sigur De Vries, Sander Wessel Keemink, and Marcel Antonius Johannes van Gerven. Kozax: flexible and scalable genetic programming in jax. In *Proceedings of the Genetic and Evolutionary Computation Conference Companion*, pages 603–606, 2025.
- [16] Yuxing Ding, Shanyue Wang, Yachen Mao, Yubo Yan, and Panlong Yang. Videoback: High quality video backscatter with ambient wifi. In *2023 IEEE 29th International Conference on Parallel and Distributed Systems (ICPADS)*, pages 1991–1998. IEEE, 2023.
- [17] Caihui Du, Jihong Yu, Rongrong Zhang, Ju Ren, and Jianping An. Orthcatter: High-throughput in-band {OFDM} backscatter with {Over-the-Air} code division. In *21st USENIX Symposium on Networked Systems Design and Implementation (NSDI 24)*, pages 1301–1314, 2024.
- [18] Federal Communications Commission. 47 cfr part 30: Upper microwave flexible use service. <https://www.ecfr.gov/current/title-47/chapter-I/subchapter-B/part-30>, 2026. Code of Federal Regulations, Title 47, Chapter I, Subchapter B, Part 30.
- [19] Raymond Fei. Gpu implementation of l-bfgs-b. <https://github.com/raymondfeilbfgsb-gpu>. Accessed: 2026.
- [20] Klaus Finkenzeller. *RFID handbook: fundamentals and applications in contactless smart cards, radio frequency identification and near-field communication*. John Wiley & sons, 2010.
- [21] Gary M Gaukler and Ralf W Seifert. Applications of rfid in supply chains. *Trends in supply chain design and management*, pages 29–48, 2007.
- [22] Google. Google pixel 10 – specifications. https://store.google.com/product/pixel_10_specs?hl=en-US, 2025. Accessed: 2025-12-05.
- [23] Xiuzhen Guo, Yuan He, Longfei Shangguan, Yande Chen, Chaojie Gu, Yuanchao Shu, Kyle Jamieson, and Jiming Chen. Mighty: Towards long-range and high-throughput backscatter for drones. *IEEE Transactions on Mobile Computing*, 2024.
- [24] Yuan Hu, Gang Yang, Songbo Fu, Marco Di Renzo, and Mérouane Debbah. Cellscatter: Efficient control and backscatter communication via ambient cellular signals. In *IEEE INFOCOM 2025-IEEE Conference on Computer Communications*, pages 1–10. IEEE, 2025.
- [25] Elisabeth Ilie-Zudor, Zsolt Kemény, Fred Van Blommestein, László Monostori, and André Van Der Meulen. A survey of applications and requirements of unique identification systems and rfid techniques. *Computers in Industry*, 62(3):227–252, 2011.
- [26] Vikram Iyer, Vamsi Talla, Bryce Kellogg, Shyamnath Gollakota, and Joshua Smith. Inter-technology backscatter: Towards internet connectivity for implanted devices. In *Proceedings of the 2016 ACM SIGCOMM Conference*, pages 356–369, 2016.
- [27] Bryce Kellogg, Aaron Parks, Shyamnath Gollakota, Joshua R Smith, and David Wetherall. Wi-fi backscatter: Internet connectivity for rf-powered devices. In *Proceedings of the 2014 ACM Conference on SIGCOMM*, pages 607–618, 2014.
- [28] Bryce Kellogg, Vamsi Talla, Shyamnath Gollakota, and Joshua R Smith. Passive {Wi-Fi}: Bringing low power to {Wi-Fi} transmissions. In *13th USENIX Symposium on Networked Systems Design and Implementation (NSDI 16)*, pages 151–164, 2016.
- [29] John Kimionis, Apostolos Georgiadis, and Manos M Tentzeris. Millimeter-wave backscatter: A quantum leap for gigabit communication, rf sensing, and wearables. In *2017 IEEE MTT-S International Microwave Symposium (IMS)*, pages 812–815. IEEE, 2017.
- [30] Zhenglun Kong, Dongkuan Xu, Zhengang Li, Peiyang Dong, Hao Tang, Yanzhi Wang, and Subhabrata Mukherjee. Autovit: Achieving real-time vision transformers on mobile via latency-aware coarse-to-fine search. *International Journal of Computer Vision*, pages 1–17, 2025.
- [31] Woosuk Kwon, Sehoon Kim, Michael W. Mahoney, Joseph Hassoun, Kurt Keutzer, and Amir Gholami. A fast post-training pruning framework for transformers, 2022.
- [32] Lianjun Li, Hao Chen, Hao-Hsuan Chang, and Lingjia Liu. Deep residual learning meets ofdm channel estimation. *IEEE Wireless Communications Letters*, 9(5):615–618, 2019.
- [33] Libero. Soc v11.8 archive. <https://www.microsemi.com/product-directory/root/5485-libero-soc-v11-8-archive>.
- [34] Vincent Liu, Aaron Parks, Vamsi Talla, Shyamnath Gollakota, David Wetherall, and Joshua R Smith. Ambient backscatter: Wireless communication out of thin air. *ACM SIGCOMM computer communication review*, 43(4):39–50, 2013.
- [35] Xin Liu, Zicheng Chi, Wei Wang, Yao Yao, Pei Hao, and Ting Zhu. Verification and redesign of {OFDM} backscatter. In *18th USENIX symposium on networked systems design and implementation (NSDI 21)*, pages 939–953, 2021.
- [36] Haofan Lu, Mohammad Mazaheri, Reza Rezvani, and Omid Abari. A millimeter wave backscatter network for two-way communication and localization. In *Proceedings of the ACM SIGCOMM 2023 Conference*, pages 49–61, 2023.
- [37] Mohammad Hossein Mazaheri, Alex Chen, and Omid Abari. Millimeter wave backscatter: Toward batteryless wireless networking at gigabit speeds. In *Proceedings of the 19th ACM Workshop on Hot Topics in Networks*, pages 139–145, 2020.
- [38] Mohammad Hossein Mazaheri, Alex Chen, and Omid Abari. Mmtag: A millimeter wave backscatter network. In *Proceedings of the 2021 ACM SIGCOMM 2021 Conference*, pages 463–474, 2021.
- [39] Pavel V Nikitin and KV Seshagiri Rao. Theory and measurement of backscattering from rfid tags. *IEEE Antennas and Propagation Magazine*, 48(6):212–218, 2006.
- [40] NVIDIA. Jetson orin nano developer kit. <https://developer.nvidia.com/blog/nvidia-jetson-orin-nano-developer-kit-gets-a-super-boost/>, 2024. Accessed: 2025-12-05.
- [41] NVIDIA. Nvidia geforce rtx 5090 graphics card. <https://www.nvidia.com/en-us/geforce/graphics-cards/50-series/rtx-5090/>, 2025. Accessed: 2025-12-05.
- [42] Ryu Okubo, Luke Jacobs, Jinhua Wang, Steven Bowers, and Elahe Soltanaghahi. Integrated two-way radar backscatter communication and sensing with low-power iot tags. In *Proceedings of the ACM SIGCOMM 2024 Conference*, ACM SIGCOMM '24, page 327–339, New York, NY, USA, 2024. Association for Computing Machinery.
- [43] Qihui Qin, Kai Chen, Yaxiong Xie, Heng Luo, Dingyi Fang, and Xiaojiang Chen. Pushing the throughput limit of ofdm-based wi-fi backscatter communication. In *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking*, pages 968–983, 2024.
- [44] Ettus Research. USRP X310. <https://www.ettus.com/all-products/x310-kit/>, 2025. Accessed: 2025-11-01.
- [45] Sebastian Ruder. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*, 2017.
- [46] Elahe Soltanaghahi, Akarsh Prabhakara, Artur Balanuta, Matthew Anderson, Jan M Rabaey, Swarun Kumar, and Anthony Rowe. Millimetro: mmwave retro-reflective tags for accurate, long range localization. In *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking*, pages 69–82, 2021.
- [47] Mehran Soltani, Vahid Pourahmadi, Ali Mirzaei, and Hamid Sheikhzadeh. Deep learning-based channel estimation. *IEEE Communications Letters*, 23(4):652–655, 2019.
- [48] Monsoon Solutions. High voltage power monitor, 2023. <https://www.monsoon.com/online-store/High-Voltage-Power-Monitor-p90002590>.

- [49] Trevor Standley et al. Which tasks should be learned together in multi-task learning? In *ICML*, 2020.
- [50] Axel Strobel, Christian Carlowitz, Robert Wolf, Frank Ellinger, and Martin Vossiek. A millimeter-wave low-power active backscatter tag for fmcw radar systems. *IEEE transactions on microwave theory and techniques*, 61(5):1964–1972, 2013.
- [51] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27, 2014.
- [52] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- [53] Shanyue Wang, Yubo Yan, Yachen Mao, Yuxin Ding, Yinghao Zhao, Panlong Yang, and Xiang-Yang Li. Reliable backscatter video streaming with ambient wifi. *ACM Transactions on Internet of Things*.
- [54] Zhaoyuan Xu and Wei Gong. Enabling zigbee backscatter communication in a crowded spectrum. In *2022 IEEE 30th international conference on network protocols (ICNP)*, pages 1–11. IEEE, 2022.
- [55] Zhaoyuan Xu and Wei Gong. Bumblebee: Enabling the vision of pervasive zigbee backscatter communication. In *2023 IEEE international conference on pervasive computing and communications (PerCom)*, pages 252–261. IEEE, 2023.
- [56] Yifan Yang, Longzhi Yuan, Jia Zhao, and Wei Gong. Content-agnostic backscatter from thin air. In *Proceedings of the 20th annual international conference on mobile systems, applications and services*, pages 343–356, 2022.
- [57] Jihong Yu, Caihui Du, Jiahao Liu, Rongrong Zhang, and Shuai Wang. Subcarrier-level ofdm backscatter. In *IEEE INFOCOM 2023-IEEE Conference on Computer Communications*, pages 1–10. IEEE, 2023.
- [58] Zhihang Yuan, Yuzhang Shang, Yang Zhou, Zhen Dong, Zhe Zhou, Chenhao Xue, Bingzhe Wu, Zhikai Li, Qingyi Gu, Yong Jae Lee, et al. Llm inference unveiled: Survey and roofline model insights. *arXiv preprint arXiv:2402.16363*, 2024.
- [59] Li Lina Zhang, Shihao Han, Jianyu Wei, Ningxin Zheng, Ting Cao, Yuqing Yang, and Yunxin Liu. nn-meter: Towards accurate latency prediction of deep-learning model inference on diverse edge devices. In *Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services*, page 81–93, New York, NY, USA, 2021. ACM.
- [60] Maolin Zhang, Jia Zhao, Si Chen, and Wei Gong. Reliable backscatter with commodity ble. In *IEEE INFOCOM 2020-IEEE Conference on Computer Communications*, pages 1291–1299. IEEE, 2020.
- [61] Pengyu Zhang, Dinesh Bharadia, Kiran Joshi, and Sachin Katti. Hitchhike: Practical backscatter using commodity wifi. In *Proceedings of the 14th ACM Conference on Embedded Network Sensor Systems CD-ROM*, pages 259–271, 2016.
- [62] Jia Zhao, Wei Gong, and Jiangchuan Liu. Spatial stream backscatter using commodity wifi. In *Proceedings of the 16th annual international conference on mobile systems, applications, and services*, pages 191–203, 2018.
- [63] Renjie Zhao, Fengyuan Zhu, Yuda Feng, Siyuan Peng, Xiaohua Tian, Hui Yu, and Xinbing Wang. Ofdma-enabled wi-fi backscatter. In *The 25th Annual International Conference on Mobile Computing and Networking*, pages 1–15, 2019.