Scaling Law for Large Wireless Models

Ziheng Liu[1], Jiayi Zhang[1]*, Haoyu Wang[2,3], Bokai Xu[1], Chen Zhang[4], Yiyang Zhu[5], Enyu Shi[1]
[1]School of Electronics and Information Engineering, Beijing Jiaotong University
[2]Institute of Information Science, Beijing Jiaotong University
[3]Visual Intelligence +X International Cooperation Joint Laboratory of MOE
[4]Department of Electrical and Electronic Engineering, The University of Hong Kong
[5]School of Electrical and Electronics Engineering, Nanyang Technological University
{zihengliu, jiayizhang, wanghy23, 20251197, enyushi}@bjtu.edu.cn, czhang6@connect.hku.hk, YIYANG015@e.ntu.edu.sg

Abstract

Emerging from recent advances in foundation models, Large Wireless Models (LWMs) represent a new paradigm of general-purpose intelligence for wireless communications that transcends task-specific engineering. The success of foundation models is critically underpinned by scaling laws, which provide a predictable roadmap for how performance scales with resources. However, established scaling laws from language and vision, charting performance as a power-law of model and dataset sizes, are ill-suited for the wireless domain, as their core formulations cannot model the structured nature of the physical channel. To address this, we propose a novel wireless scaling law that extends the classical formulation by modeling two wireless-native factors: channel heterogeneity and discretization granularity. These two factors reshape scaling behavior via nested linear and power-law relationships, recasting the scaling law's parameters (notably the scaling exponent and irreducible loss) from universal constants into dynamic variables dictated by the physical environment. Our physics-aware formulation reveals two key insights: first, that compute-optimal scaling is not dictated by a fixed model-data ratio but is instead a dynamic function of heterogeneity and granularity, and second, that this dependency is particularly sensitive to granularity, allowing significant performance to be unlocked from existing data simply by refining its resolution. Crucially, this establishes a reliable roadmap for designing powerful yet resource-efficient LWMs, translating theoretical insights into actionable engineering principles. Extensive experiments validate our wireless scaling law, showing a 32.31% prediction accuracy improvement over classical laws in diverse wireless scenarios where they fail.

Introduction

The rapid development of foundation models, exemplified by Large Language Models (LLMs) (Radford et al. 2018; Devlin et al. 2019) and Large Vision Models (LVMs) (Dosovitskiy et al. 2020; Radford et al. 2021), has unlocked a new caliber of general-purpose capabilities. Inspired by this success, a transformative wave is now extending into scientific and engineering domains, most notably in wireless communications, giving rise to Large Wireless Models (LWMs)

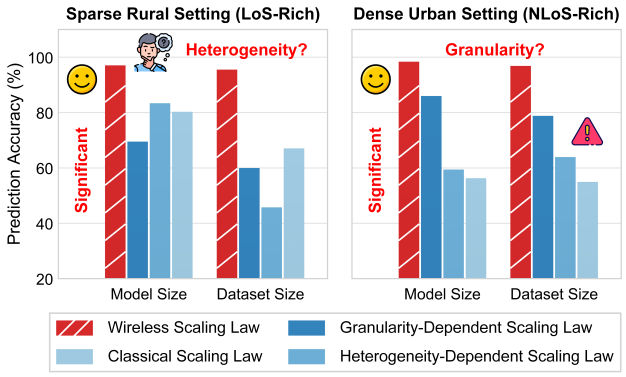

Figure 1: Comparison of prediction accuracy for different scaling laws across diverse wireless scenarios: from simple LoS-rich environments to complex NLoS-rich ones.

(Alikhani, Charan, and Alkhateeb 2024; Shao et al. 2024). This emerging paradigm marks a departure from the long-standing tradition in wireless engineering of designing task-specific algorithms (Chen, Xiong, and Yang 2024; Liu et al. 2024b; Zhang et al. 2023a,b, 2025b). At their core, these models learn a unified, high-dimensional representation directly from raw signals (Alikhani, Charan, and Alkhateeb 2024; Zhang et al. 2024, 2025a). This versatile representation enables strong generalization across diverse wireless scenarios, overcoming the limitations of traditional handcrafted engineering and reducing the need for costly task-specific retraining (Xu et al. 2024; Liu et al. 2025b,c).

Despite this visionary potential, realizing powerful LWMs confronts a formidable obstacle (Alikhani, Charan, and Alkhateeb 2024; Shen et al. 2024; Tong et al. 2025; Jiang et al. 2025): an immense design space spanning model size, dataset size, and compute budget, where brute-force trial and error is both scientifically inefficient and financially prohibitive. In sharp contrast, this very challenge of brute-force experimentation was overcome in domains like language and vision through the discovery of "scaling laws" (Kaplan et al. 2020; Henighan et al. 2020; Hoffmann et al. 2022; Zhai et al. 2022; Cherti et al. 2023). Distilled from large-scale experiments, this principled methodology provides a crucial roadmap by charting how model performance scales

*Corresponding author.

as a function of model and dataset sizes.

Nevertheless, we find that the predictive power of established scaling laws falters in the wireless domain, where their direct application yields systematic deviations from empirical results. This discrepancy stems from a mismatch in the underlying data assumptions. Classical scaling laws, derived from domains like language and vision (Kaplan et al. 2020; Zhai et al. 2022), presuppose data with macro-scale statistical consistency that permits the formulation of unified scaling laws. The wireless domain violates this premise: its signals are not abstract tokens but physical measurements whose statistical distributions are deterministically shaped by the propagation environment (Cover 1999; Tse and Viswanath 2005). For instance, datasets from line-of-sight (LoS)-dominant rural and non-line-of-sight (NLoS)-rich urban scenarios (Xiao et al. 2017; O'shea and Hoydis 2017; Ye et al. 2020) represent not merely distinct semantic contexts but different statistical populations. This suggests that for scaling laws to be effective in the wireless domain, their static constants need to be reconceptualized as dynamic functions of the physical environment.

Motivated by this necessity, we propose a central hypothesis: LWM scaling is not universal, but is instead primarily governed by two fundamental physical axes. The first is channel heterogeneity, which arises from structured statistical variations in the propagation environment (e.g., LoS vs. NLoS). The second is discretization granularity, which defines the model's effective spatial resolution by determining how the channel's physical structure is sampled. Though analogous to concepts like tokenization in language, the choice of granularity in the wireless domain is not merely a computational decision. Instead, it imposes a physical trade-off between capturing fine-grained channel variations and preserving long-range structural coherence. Neglecting these two physical axes in classical scaling laws creates a methodological void, making a physics-aware formulation essential to guide the principled development of LWMs. To validate this hypothesis, we introduce and systematically evaluate a new paradigm termed the Wireless Scaling Law. A comprehensive study, which integrates theoretical analysis with extensive empirical validation, culminates in a novel formulation that fundamentally extends the classical paradigm. By analytically incorporating heterogeneity and granularity, our formulation transforms key scaling parameters, such as the scaling exponent and irreducible loss, from universal constants into dynamic variables governed by the physical environment. Consequently, this scaling law establishes a principled roadmap for LWM development, transforming theoretical insights into actionable engineering principles. Figure 1 shows a comparison of various scaling laws, where our proposed wireless scaling law yields substantially higher prediction accuracy across diverse wireless scenarios. Overall, our key contributions are listed as follows:

- We introduce the Wireless Scaling Law, a novel paradigm for LWM scaling grounded in two wireless-native factors: heterogeneity and granularity. This paradigm reveals a fundamental causal chain: these factors induce nested mathematical relationships that reshape scaling behavior, in turn transforming key scaling parameters from universal constants into dynamic variables.
- We elucidate the physical mechanisms shaping compute-optimal scaling in the wireless domain, demonstrating that it is not a fixed model-data ratio, but rather a dynamic function of heterogeneity and granularity. Crucially, our analysis uncovers this dependency is particularly sensitive to granularity, a finding which not only unlocks significant performance from existing data via resolution refinement, but also establishes a reliable roadmap for designing resource-efficient LWMs.
- Extensive experiments across diverse wireless scenarios validate the generalizability of our wireless scaling law. Our unified formulation demonstrates superior prediction accuracy, outperforming classical laws by 32.31%, as well as models focusing solely on heterogeneity (by 33.83%) or granularity (by 23.38%), proving the critical importance of modeling both factors in tandem.

Related Work

Large Wireless Models. Inspired by the transformative success of foundation models in language and vision (Li, Sun, and Li 2023; Li et al. 2024; Fang et al. 2024; Zhou, Hong, and Wu 2024), the wireless research community is embracing a new paradigm centered on LWMs (Jiang et al. 2024a,b). These efforts can be broadly classified into two main categories. The first focuses on spatial inference (Alikhani, Charan, and Alkhateeb 2024; Ott et al. 2024; Catak, Kuzlu, and Cali 2025), developing models that map complex spatial correlations to optimal real-time actions. The second category specializes in generative temporal modeling (Jiang et al. 2022; Liu et al. 2024a, 2025a,b; Yang et al. 2025), learning underlying temporal dependencies to forecast the evolution of channel states.

Scaling Laws. The discovery of scaling laws provides a predictable roadmap charting how performance scales with key resources like model and dataset sizes, which underpins our foundational understanding of deep models' empirical behavior (Kaplan et al. 2020; Zhai et al. 2022; Chowdhery et al. 2023). Subsequent research has broadened the scope of these principles, establishing more granular scaling laws for specific contexts, including data-constrained settings (Muennighoff et al. 2023), temporal dynamics (Xiong et al. 2024), model distillation (Busbridge et al. 2025), and the use of synthetic data (Fan et al. 2024; Qin et al. 2025). However, this body of work is predominantly rooted in language and vision, rendering the scaling behavior of models under the unique constraints of the wireless domain a conspicuous blind spot. In this paper, we present the first systematic investigation into the scaling behavior of LWMs.

Wireless Scaling Law

This section presents the theoretical foundation of LWMs, beginning with essential preliminaries and subsequently investigating the scaling behavior of LWMs.

Preliminaries of Large Wireless Models

Departing from traditional task-specific models, LWMs (Alikhani, Charan, and Alkhateeb 2024) leverage the expres-

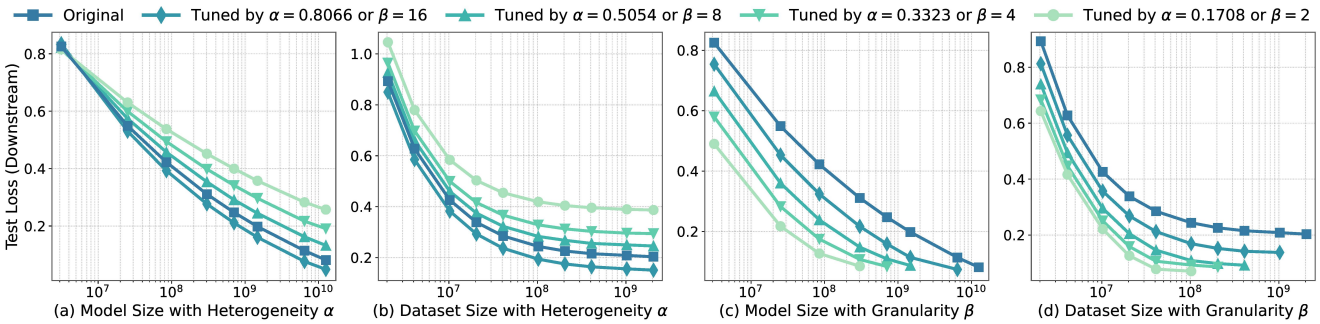

Figure 2: Impact of channel heterogeneity (α) and discretization granularity (β) on the scaling behavior of LWMs.

sive power of foundation models to establish a task-agnostic framework, capable of capturing intricate correlations and generalizing across diverse tasks. For example, consider a set of wireless tasks, each defined by an objective function $g_{0,\text{task}}$ and n_c constraints $g_{c,\text{task}}$ $(1 \leqslant c \leqslant n_c)$, satisfying

$$\max_{\mathbf{o}_{\text{task}}} \quad g_{0,\text{task}}(\mathbf{h}_{\text{task}}, \mathbf{o}_{\text{task}}) \tag{1a}$$

$$\text{s.t.} \quad g_{c,\text{task}}(\mathbf{h}_{\text{task}}, \mathbf{o}_{\text{task}}) \geqslant 0, \forall c, \tag{1b}$$

where $\mathbf{h}_{\text{task}}$ and $\mathbf{o}_{\text{task}}$ represent input features and optimization objectives, respectively. Traditional approaches employ bespoke models $f_{\theta,\text{task}}$ for individual tasks $g_{0,\text{task}}$, a design choice that fundamentally limits their generalization. In stark contrast, LWMs, grounded in a unified paradigm, are intrinsically advantageous for wireless communications, as their pretrained models $\mathcal{F}_\theta$ enable few- or zero-shot generalization, reducing the need for task-specific retraining.

The realization of this paradigm hinges on a potent architectural foundation. Following core LLM design principles, the Transformer, which is an architecture already proven effective in the wireless domain (Shao et al. 2024), is an inherently fitting choice (Vaswani et al. 2017). Specifically, its encoder excels at extracting high-dimensional inputs (e.g., sequences $\mathbf{h} \in \mathbb{C}^d$) into a potent latent representation $\mathbf{z}$, enabling few- and zero-shot generalization to diverse tasks. This encoder-centric design is deliberate: wireless components (e.g., spatial, frequency) lack the strong causal order found in natural language. This distinction shifts the primary objective from sequential prediction to holistic inference, a task for which the encoder's capacity to model global dependencies via self-attention is intrinsically superior.

During pre-training, the input sequences $\mathbf{h}$ are tokenized and transformed into the Transformer's embedding space, allowing LWMs to learn complex interdependencies that form the latent representation $\mathbf{z}$. This involves preprocessing steps such as patch generation, which first separates the inputs $\mathbf{h}$ into the real and imaginary components ($\mathbf{h}^{\text{real}}$ and $\mathbf{h}^{\text{imag}}$) and then partitions each component into $P/2$ s-sized patches $\mathbf{x}_p^{\text{real}}$ and $\mathbf{x}_p^{\text{imag}}$, $p \in \{1, \ldots, P/2\}$, with the input dimension satisfying $d = Ps/2$. Subsequently, a learnable classication (CLS) patch $\mathbf{x}^{\text{cls}} \in \mathbb{R}^s$, is prepended to the sequence of P patches. This patch serves as a global feature extractor for $\mathbf{h}$, resulting in the final input sequences:

$$\mathbf{x} = \{\mathbf{x}^{\text{cls}}, \{\mathbf{x}_p^{\text{real}}\}_{p=1}^{P/2}, \{\mathbf{x}_{p+P/2}^{\text{imag}}\}_{p=1}^{P/2}\} \in \mathbb{R}^{s\times(P+1)}. \tag{2}$$

Furthermore, channel masking is implemented by randomly masking portions of the input sequences (i.e., $\mathbf{x} \to \hat{\mathbf{x}}$), to compel LWMs to learn robust contextual dependencies through self-supervised learning. Thus, LWMs ($\mathcal{F}_\theta : \hat{\mathbf{x}} \to \mathbf{z}$) are trained to reconstruct the original input by minimizing the mean squared error (MSE) between the output of a decoder $\mathcal{F}_{\text{dec}}(\mathbf{z})$ and the original sequences $\mathbf{x}$, computed exclusively over the masked indices $\mathcal{M}$.

Inference. Notably, pre-training LWMs on masked sequences $\hat{\mathbf{x}}$, while effective for learning robust dependencies, creates a detrimental mismatch with the original sequences $\mathbf{x}$ used during inference. This mismatch impairs downstream performance and thus necessitates computing the test loss $\mathcal{L}$ with $\mathcal{M} \subseteq \{1, \ldots, P+1\}$ and $\mathbf{z} = \mathcal{F}_\theta(\mathbf{x})$, as follows:

$$\mathcal{L} = \frac{1}{|\mathcal{M}|} \sum_{p\in\mathcal{M}} \|\mathcal{F}_{\text{dec}}(\mathbf{z})_p - \mathbf{x}_p\|_2^2. \tag{3}$$

This objective is well-aligned with practical tasks like channel state information (CSI) compression (Yin et al. 2022), where the encoder generates a compact representation $\mathbf{z}$ and the decoder's primary goal is accurate CSI reconstruction.

Scaling Laws for Large Wireless Models

The emergence of LWMs, with their remarkable few- and zero-shot generalization, heralds a paradigm shift from task-specific engineering to universal intelligence. This advancement, however, raises a pivotal question for the practical deployment of these models in wireless scenarios: **How does their performance predictably scale with core resources**, particularly **model sizes**, limited by on-device hardware, and **dataset sizes**, constrained by user privacy?

Inspired by the prior scaling laws in language models (Kaplan et al. 2020), which demonstrated that the test loss $\mathcal{L}$ follows a predictable power-law relationship with each of model size N and dataset size D when not bottlenecked by the other, satisfying

$$\mathcal{L}(N, \infty) \propto (1/N^{\alpha_{\text{n}}}), \mathcal{L}(\infty, D) \propto (1/D^{\alpha_{\text{d}}}), \tag{4}$$

where the power-law exponents α_{n} and α_{d} are constants.

While this constant-exponent power-law provides a compelling precedent, its universality to the wireless domain is fundamentally limited. This limitation stems from its inability to account for performance variations induced by two critical properties of the physical medium (see Figure 2):

Scaling Paradigms	Mathematical Formulation of Scaling Behavior
Classical Scaling Law	Model: $\mathcal{L}(N) = -0.241 + (\frac{4901100}{N})^{0.144}$ Dataset: $\mathcal{L}(D) = 0.198 + (\frac{1211190}{D})^{0.692}$
α-Dependent Scaling Law	Model: $\mathcal{L}(N, \alpha) = -0.022\alpha - 0.225 + (\frac{75008\alpha+4853924}{N})^{0.105\alpha+0.076}$ Dataset: $\mathcal{L}(D, \alpha) = -0.359\alpha + 0.429 + (\frac{74246\alpha+1164312}{D})^{-0.100\alpha+0.758}$
β-Dependent Scaling Law	Model: $\mathcal{L}(N, \beta) = -1.311\beta^{0.047} + 1.303 + (\frac{73030\beta^{1.191}+365297}{N})^{0.460\beta^{-0.245}-0.051}$ Dataset: $\mathcal{L}(D, \beta) = 0.226\beta^{0.258} - 0.350 + (\frac{36493\beta^{0.361}+1083745}{D})^{1.003\beta^{0.049}-0.493}$
Wireless Scaling Law (**Ours**)	Model: $\mathcal{L}(N, \alpha, \beta) = 0.012\alpha - 1.662\beta^{0.036} + 1.641 + \frac{(208181\alpha+75190\beta^{1.185}+216948)^{e_1}}{N^{(e_1:0.102\alpha+0.518\beta^{-0.203}-0.179)}}$ Dataset: $\mathcal{L}(D, \alpha, \beta) = -0.344\alpha + 0.277\beta^{0.227} - 0.185 + \frac{(203835\alpha+17430\beta^{0.601}+964826)^{e_2}}{D^{(e_2:-0.098\alpha+1.737\beta^{0.030}-1.166)}}$

Table 1: Comparison of four different scaling paradigms, all evaluated on the DeepMIMO dataset. The classical scaling law serves as a baseline, scaling loss solely with model size N and dataset size D. In contrast, the subsequent scaling laws are physics-aware: the α-dependent and β-dependent models are ablations incorporating only heterogeneity α or granularity β, respectively, while our wireless scaling law provides a unified formulation incorporating both.

- **Channel heterogeneity.** The statistical composition of a wireless channel, particularly the dominance of LoS paths (Ye et al. 2020), profoundly alters model performance, rendering a fixed scaling exponent insufficient to capture the diversity inherent in wireless scenarios.
- **Discretization granularity.** As a critical design choice, the channel's granularity defines the model's effective spatial resolution. Unlike in natural language, where this choice is primarily computational, here it imposes a fundamental physical trade-off between resolving fine-grained details and preserving long-range coherence (Alikhani, Charan, and Alkhateeb 2024), which in turn fundamentally alters the scaling dynamics.

As observed in Figure 2, the impact of these physical factors on LWMs challenges the notion of traditional fixed scaling parameters, indicating they are instead a function of the channel's properties. We therefore propose the **Wireless Scaling Law**: a novel paradigm where the scaling parameters are dynamically modulated by channel heterogeneity and discretization granularity, providing a reliable roadmap for designing resource-efficient LWMs.

Impact of Channel Heterogeneity

Prior scaling laws (Kaplan et al. 2020; Hoffmann et al. 2022) model test loss as a function of model and dataset sizes, a formulation that implicitly relies on a premise of information homogeneity. While this premise holds in domains like natural language, where each data sample can be considered to provide a roughly equivalent amount of learnable information, it is fundamentally challenged in wireless communications. Here, the physical environment induces inherent heterogeneity in the information content across different channel sequences $\mathbf{h}$ (Liu et al. 2024d,c), satisfying

$$\mathbf{h}(\alpha) = \sqrt{\alpha}\mathbf{h}_{\mathrm{LoS}} + \sqrt{1-\alpha}\mathbf{h}_{\mathrm{NLoS}}, \tag{5}$$

where $\alpha \in [0, 1]$ denotes the LoS power fraction, $\mathbf{h}_{\mathrm{LoS}}$ is a rank-one matrix representing the deterministic LoS path, and $\mathbf{h}_{\mathrm{NLoS}}$ is a full-rank random matrix modeling the stochastic NLoS paths (ElMossallamy et al. 2020). Then, considering two datasets of the same size: $\mathcal{D}_1$, comprising channels $\mathbf{h}_1 = \mathbf{h}(\alpha \to 1)$ from a sparse rural area dominated by LoS paths, and $\mathcal{D}_2$, comprising channels $\mathbf{h}_2 = \mathbf{h}(\alpha \to 0)$ from a dense urban area rich in NLoS paths. There are significant differences between these two datasets, as follows:

- **LoS-dominant.** These channels $\mathbf{h}_1$ are characterized by a deterministic main path, while stochastic paths contribute substantially less power. Thus, their matrices are approximately rank-one $\mathrm{rank}(\mathbf{h}_1) \approx 1$ with low information entropy $h(\mathbf{h}_1)$, greatly enhancing predictability.
- **NLoS-dominant.** These channels $\mathbf{h}_2$ exhibit complex and rapidly changing fading patterns caused by numerous reflections and rich scattering. Thus, their matrices are typically near full-rank $\mathrm{rank}(\mathbf{h}_2) \gg \mathrm{rank}(\mathbf{h}_1)$ with higher information entropy $h(\mathbf{h}_2) \gg h(\mathbf{h}_1)$, making it difficult for LWMs to learn intricate dependencies.

In this context, the channel heterogeneity, parameterized by α, governs the spatial complexity $\mathrm{rank}(\mathbf{h}(\alpha))$ and inherent unpredictability $h(\mathbf{h}(\alpha))$ of the channels, both of which are monotonically decreasing functions of α, satisfying

$$\frac{\partial}{\partial\alpha}\mathrm{rank}(\mathbf{h}(\alpha)) < 0, \frac{\partial}{\partial\alpha}h(\mathbf{h}(\alpha)) < 0, \alpha \in [0, 1]. \tag{6}$$

Our analysis, which challenges the flawed assumption of information homogeneity, we reveal that a channel's "learnability" is not constant but is instead dictated by its structural predictability, quantified by α (see Figure 2). This insight leads to a novel performance limit: a channel structure-dependent lower bound on prediction error scaling as

$$\mathcal{L}_{\min}(\alpha) \propto (1 - \alpha), \tag{7}$$

which forms the first part of our **Wireless Scaling Law**.

Impact of Discretization Granularity

Prior scaling laws for language models are typically built upon a standardized pre-tokenized paradigm (Kaplan et al. 2020), which provides a uniform measure of input granularity. This premise of uniformity is untenable in wireless communications, where the choice of granularity is not fixed but is instead dictated by the channel's physical properties.

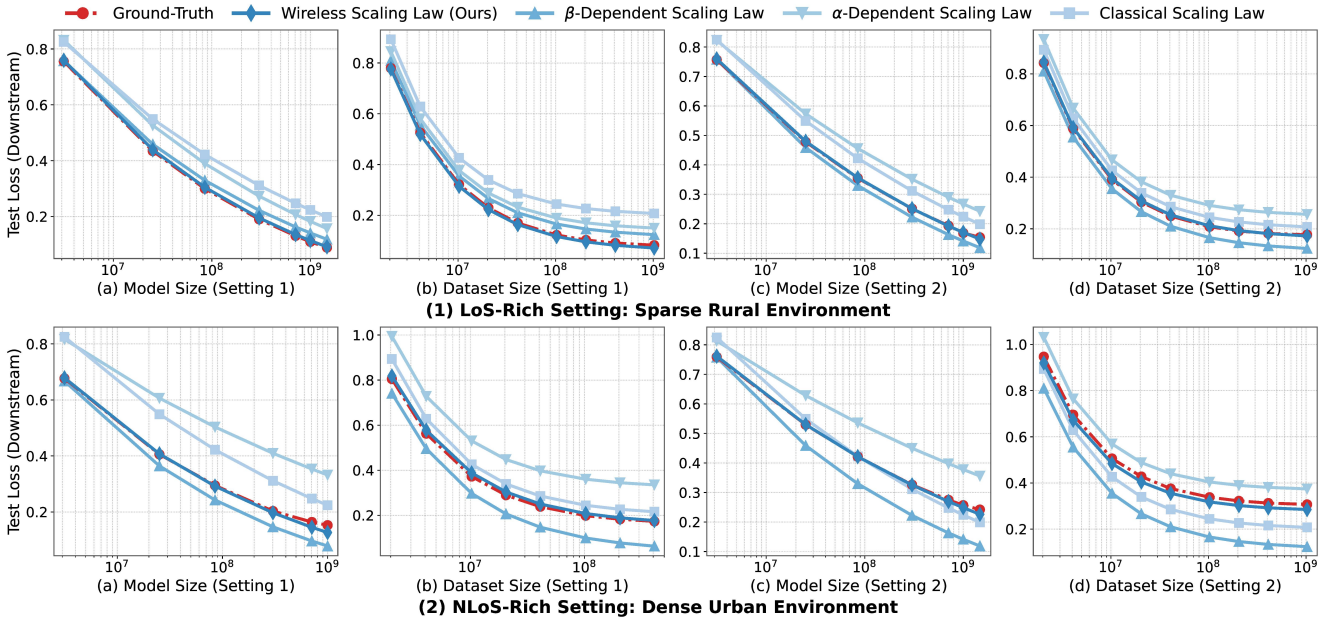

Figure 3: Evaluation of the prediction accuracy for different scaling paradigms against ground-truth performance across diverse wireless scenarios (sparse rural and dense urban). For all paradigms, scaling parameters (see Table 1) are derived from the empirical results of Setting 1 (and other settings) and then applied to predict performance on Setting 2, an unseen configuration.

For instance, a dataset $\mathcal{D}_1$ from a sparse rural area and a dataset $\mathcal{D}_2$ from a dense urban area possess distinct physical properties, thus necessitating different granularities for effective modeling. To formalize this dependency, we decompose the test loss $\mathcal{L}(\alpha, \beta)$, now a function of both heterogeneity α and granularity β, into two competing terms:

$$\mathcal{L}(\alpha, \beta) = \mathcal{L}_{\text{cr}}(\alpha, \beta) + \mathcal{L}_{\text{sr}}(\alpha, \beta), \tag{8}$$

where $\beta \in [1, 2d]$ can be quantified by the patch size s, and the terms $\mathcal{L}_{\text{cr}}(\alpha, \beta)$ and $\mathcal{L}_{\text{sr}}(\alpha, \beta)$ are defined as follows:

- **Coherence reconstruction loss $\mathcal{L}_{\text{cr}}(\alpha, \beta)$.** This component quantifies the difficulty of reconstructing the channel's coherent low-entropy structure, which is primarily driven by spatial LoS paths. A finer granularity (low-β) fragments this global structure into a long sequence of patches, posing a significant challenge for LWMs in aggregating long-range context. Thus, $\mathcal{L}_{\text{cr}}(\alpha, \beta)$ is a monotonically decreasing function of β and is modeled as $\mathcal{L}_{\text{cr}}(\alpha, \beta) \propto \alpha\beta^{-\gamma_1}$, where $\gamma_1 > 0$ is the context aggregation difficulty exponent.
- **Selectivity representation loss $\mathcal{L}_{\text{sr}}(\alpha, \beta)$.** In contrast, this component measures the fidelity loss in capturing the channel's fine-grained high-entropy variations, which arise from frequency-selective NLoS paths. A coarser granularity (high-β) averages out these rapid fluctuations across subcarriers within each patch, causing a direct loss of detailed channel information. Thus, $\mathcal{L}_{\text{sr}}(\alpha, \beta) \propto (1-\alpha)\beta^{\gamma_2}$ is a monotonically increasing function of β with the fidelity decay exponent $\gamma_2 > 0$.

This decomposition reveals a fundamental trade-off between reconstructing global coherence $\mathcal{L}_{\text{cr}}(\alpha, \beta)$ and preserving local fidelity $\mathcal{L}_{\text{sr}}(\alpha, \beta)$. In practice, however, this balance is heavily skewed in wireless scenarios (Yang et al. 2025). Architectures like the Transformer, with their powerful self-attention mechanisms, are exceptionally adept at modeling the highly structured low-entropy patterns of LoS paths. Consequently, the loss $\mathcal{L}_{\text{cr}}(\alpha, \beta)$ becomes negligible in most practical scenarios. This insight permits a crucial simplification, where the total test loss is dominated by $\mathcal{L}_{\text{sr}}(\alpha, \beta)$:

$$\mathcal{L}(\alpha, \beta) \approx \mathcal{L}_{\text{sr}}(\alpha, \beta) \propto \underbrace{\alpha\beta^{-\gamma_1}}_{\approx 0} + (1-\alpha)\beta^{\gamma_2}, \tag{9}$$

which forms the second part of our **Wireless Scaling Law**.

Unified Wireless Scaling Law

Building on the coupled impacts of heterogeneity α and granularity β, we integrate these findings with classical scaling laws for N and D. This unification, guided by (8) and (9), forms our core paradigm. We hypothesize that these physical properties intrinsically modulate the original scaling dynamics rather than merely introducing additive error terms. This leads to our wireless scaling law, where the scaling parameters are themselves functions of α and β:

$$\mathcal{L}(N, D, \alpha, \beta) = L_{\min}(\alpha, \beta) + \frac{A(\alpha, \beta)}{N^{\alpha_{\text{n}}(\alpha, \beta)}} + \frac{B(\alpha, \beta)}{D^{\alpha_{\text{d}}(\alpha, \beta)}}, \tag{10}$$

where the irreducible loss, $L_{\min}(\alpha, \beta)$, is governed by channel unpredictability. Crucially, the coefficients $A(\alpha, \beta)$, $B(\alpha, \beta)$ and exponents $\alpha_{\text{n}}(\alpha, \beta)$, $\alpha_{\text{d}}(\alpha, \beta)$ are no longer universal constants but are instead functions of (α, β). This dynamic physics-aware formulation, which posits that learning difficulty and scaling efficiency are dictated by the physical environment, will be empirically validated below.

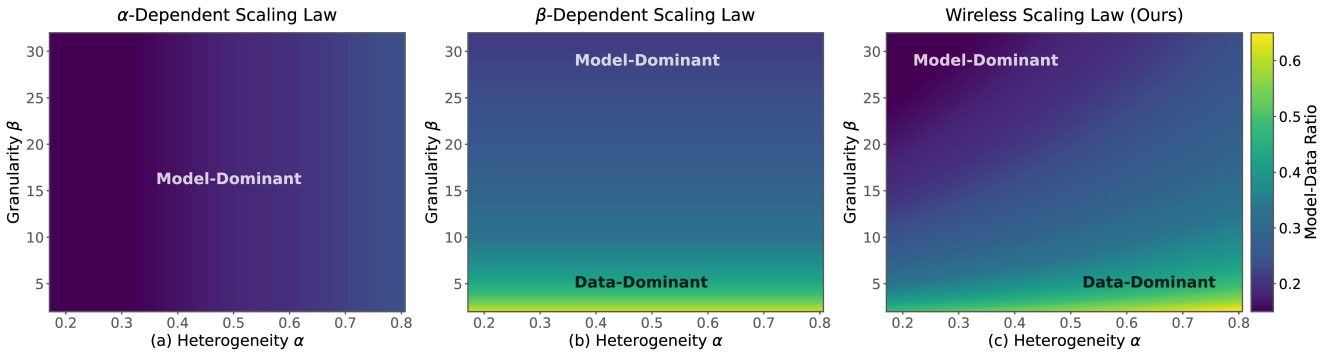

Figure 4: Optimal model-data ratio as a function of channel heterogeneity α and granularity β. In sharp contrast to the fixed ratio N^{208}/D prescribed by classical laws, our wireless scaling law reveals that the optimal ratio is dynamically determined by both α and β. This dependency establishes distinct Model-Dominant (low α, high β) and Data-Dominant (high α, low β) regimes, revealing that a fixed ratio is a suboptimal strategy for designing resource-efficient LWMs.

Experiments

Experimental Setup

In this subsection, we introduce the experiment setup, comprising models, datasets, and training procedures.

Models. We primarily train a wide variety of encoder-centric LWMs (Alikhani, Charan, and Alkhateeb 2024), with the model size spanning more than seven orders of magnitude, from 768 to 13B non-embedding parameters. Notably, the number of non-embedding parameters (Kaplan et al. 2020) is determined by several key hyperparameters: the number of layers, the dimension of the residual stream, the dimension of the feed-forward layer, the dimension of the attention output, and the number of attention heads. Following this pretraining, we fit our wireless scaling law to the empirical test loss $\mathcal{L}$ using non-linear least squares, achieving near-perfect fits with R^2 values exceeding 0.988.

Datasets. Following standard practice in wireless communications where live data access is often infeasible, we train LWMs on the DeepMIMO dataset (Alkhateeb 2019; Alikhani, Charan, and Alkhateeb 2024). As a high-fidelity dataset derived from real-world layouts, it is widely accepted as a representative proxy for validation (Yang et al. 2025). It comprises 20 city scenarios ($\sim$1500 users each) and 3 large-scale scenarios ($>$0.1M effective users each), totaling nearly 1M effective users and 1.96B channel tokens. From this, we reserve 20M channel tokens for the validation set.

Training. Following most recent open-source LWMs (Alikhani, Charan, and Alkhateeb 2024; Liu et al. 2024a, 2025a,b), we adopt the Adam optimizer (Kingma and Ba 2014) with a learning rate of 10^{-4} for all experiments to guarantee the broad applicability of our work. All models are pretrained with 1.94B tokens.

Validating the Impact of Heterogeneity on Scaling

To empirically validate the impact of heterogeneity on the scaling behavior of LWMs, we modify the training dataset used in the original experimental setup, while the validation dataset, models, and training procedures remain unchanged.

Datasets. We partition the original DeepMIMO dataset into several overlapping subsets based on channel heterogeneity α. To maintain consistency with the scale of the original setup, we augment each subset to match the original's size by repeatedly resampling its instances. This procedure preserves the unique statistical properties of each subset.

Following the training procedure outlined above, we systematically evaluate a large range of LWMs with varying channel heterogeneity α, model sizes N and dataset sizes D. Our empirical results (see Figure 3) validate that α significantly impacts the test loss $\mathcal{L}$, which adheres to the foundational power-law relationship with N and D but whose scaling parameters are predictably modulated by α following a linear-law relationship, satisfying

$$\mathcal{L}(N,\alpha) = a_1\alpha + a_2 + \left(\frac{a_3\alpha + a_4}{N}\right)^{a_5\alpha+a_6}, \tag{11a}$$

$$\mathcal{L}(D,\alpha) = b_1\alpha + b_2 + \left(\frac{b_3\alpha + b_4}{D}\right)^{b_5\alpha+b_6}, \tag{11b}$$

where the modulating parameters are given in Table 1.

Predictable performance. We establish a novel physics-aware formulation for LWMs that transcends the traditional classical paradigm. Instead of being dictated by a fixed model-data ratio like $N^{0.208}/D$, we find that optimal generalization performance is fundamentally coupled with heterogeneity α. This paradigm reveals a crucial design trade-off: for any given α, performance improves smoothly only when N and D are scaled in tandem, with predictable diminishing returns when either is constrained.

Our analysis reveals that compute-optimal scaling follows the law $D \propto N^{(0.105\alpha+0.076)/(-0.100\alpha+0.758)}$, where the data-scaling exponent is not fixed, but is instead dictated by α (see Figure 4a). This reflects a fundamental principle: the high-entropy data from complex environments (low-α) makes aggressive model scaling exceptionally data-efficient. The practical impact is profound: **a tenfold model increase** in a NLoS setting (low-α) requires only **doubling the dataset**, whereas the same scaling in a simpler LoS setting (high-α) demands **approximately tripling it**. This finding establishes a clear guideline: for maximum data efficiency, prioritize model scaling in rich-scattering environments.

Setting	Models	CR	L/N	BP-16	BP-32
Rural-1	Raw	0.171	0.348	0.233	0.201
	LWM	**0.847**	**0.955**	**0.364**	**0.372**
Rural-2	Raw	0.153	0.301	0.208	0.173
	LWM	**0.837**	**0.891**	**0.357**	**0.368**
Urban-1	Raw	0.123	0.338	0.203	0.188
	LWM	**0.828**	**0.906**	**0.355**	**0.361**
Urban-2	Raw	0.118	0.459	0.189	0.181
	LWM	**0.823**	**0.886**	**0.351**	**0.354**

Table 2: Performance (F1-score) evaluation of representations generated by the LWM (302M parameters) against a raw signal baseline. The comparison covers a suite of downstream tasks, including channel reconstruction (CR), LoS/N-LoS classification (L/N), and 16/64-beam prediction (BP).

Validating the Impact of Granularity on Scaling

Conditioned on a fixed wireless environment (a constant α), we create multiple versions of the original training dataset, each characterized by a distinct granularity β, to validate the impact of granularity on the scaling behavior of LWMs.

Our empirical results (see Figure 3), holding α constant, validates that β also critically shapes the test loss $\mathcal{L}$, which adheres to the same foundational power-law but whose scaling parameters are further and precisely refined by β following a pow-law relationship, as captured by:

$$\mathcal{L}(N,\beta) = a_1\beta^{a_2} + a_3 + \left(\frac{a_4\beta^{a_5}+a_6}{N}\right)^{a_7\beta^{a_8}+a_9}, \quad (12a)$$

$$\mathcal{L}(D,\beta) = b_1\beta^{b_2} + b_3 + \left(\frac{b_4\beta^{b_5}+b_6}{D}\right)^{b_7\beta^{b_8}+b_9}, \quad (12b)$$

where the modulating parameters are given in Table 1, standing in sharp contrast to their α-dependent counterparts, and whose robustness is confirmed by a $\pm 20\%$ parameter perturbation test, where our law's prediction error ($<5.3\%$) far surpasses the classical law's ($>14.2\%$).

Predictable performance. Our analysis of granularity β reveals a second equally scaling principle for LWMs: optimal generalization performance is not merely dependent on total dataset size D, but is critically determined by its effective resolution β. Crucially, employing channel-aware granularity acts as a powerful form of implicit regularization, compelling the model to learn features invariant to fading dynamics and thereby lowering its irreducible error. Figure 4b visualizes this principle, demonstrating that as granularity β increases, the optimal strategy shifts from a data-dominant to a model-dominant regime. The practical impact of this channel-regularized approach is a profound gain in data efficiency: achieving a given target error requires **over 90% less training data** compared to coarse-grained aggregation. This discovery unlocks a powerful path to resource efficiency, offering a transformative advantage in data-scarce settings by extracting untapped performance from existing data at no extra cost.

Validating the Unified Wireless Scaling Law

Building upon our findings from heterogeneity and granularity, we further validate the unified wireless scaling law by evaluating a comprehensive suite of LWMs under the simultaneous variation of α and β.

Our empirical results, presented in Figure 3, reveal a remarkably consistent scaling behavior. This unified view represents a decisive advance beyond two prior paradigms: classical scaling laws, which are built upon static scaling parameters, and our preliminary formulations from α or β. The key scaling parameters are dynamic, governed by a composite function that synergistically integrates a linear dependency on α with a power-law dependency on β. This synthesis culminates in our unified wireless scaling law:

$$\mathcal{L}(N,\alpha,\beta) = a_{1,2,3,4} + \left(\frac{a_{5,6,7,8}}{N}\right)^{a_{9,10,11,12}}, \quad (13a)$$

$$\mathcal{L}(D,\alpha,\beta) = b_{1,2,3,4} + \left(\frac{b_{5,6,7,8}}{D}\right)^{b_{9,10,11,12}}, \quad (13b)$$

where scaling parameters are defined as dynamic functions of α and β via $a_{i,j,k,l} = a_i\alpha + a_j\beta^{a_k} + a_l$ and $b_{i,j,k,l} = b_i\alpha + b_j\beta^{b_k} + b_l$ (see Table 1). Demonstrating high fidelity across a wide spectrum of channel conditions, particularly in unseen propagation environments (see Figure 1), our formulation outperforms classical scaling laws by 32.31%, the α-dependent model by 33.83%, and the β-dependent model by 23.38% in prediction accuracy.

Crucially, our wireless scaling law provides a core principle for designing powerful (see Table 2) yet resource-efficient LWMs (see Figure 4c), revealing that the optimal strategy is a dynamic trade-off between α and β, determining precisely **when to prioritize model scaling over data acquisition**. For instance, in rich-scattering (low-α) settings, it is more efficient to invest in larger models (model-dominant). Conversely, in low-granularity (low-β) settings that capture fine-grained details, larger datasets are required (data-dominant). This demonstrates that applying a fixed model-data ratio across diverse wireless scenarios is a sub-optimal strategy, hindering the design of efficient LWMs.

Conclusion

This paper introduces a novel wireless scaling law for LWMs, moving beyond the traditional focus on model and dataset sizes to demonstrate that scaling behavior is governed by two wireless-native factors: heterogeneity and granularity. Our paradigm reveals a fundamental principle: while foundational power-laws of scale persist, their key scaling parameters are not universal constants but are dynamically dictated by these physical properties. This is achieved through a composite structure that integrates a linear dependency on heterogeneity with a power-law dependency on granularity, improving prediction accuracy by up to 32.31%, over classical scaling laws. Notably, our wireless scaling law provides a reliable roadmap for designing resource-efficient LWMs. It redefines compute-optimal scaling as a dynamic function of the two wireless-native factors and, crucially, reveals a profound sensitivity to granularity, a property that allows significant performance gains to be unlocked from existing data simply by refining its resolution.

Acknowledgments

This work is supported by the Fundamental Research Funds for the Central Universities under Grant No.2024YJS138 & 2025JBZX037, in part by National Natural Science Foundation of China under Grant 62471027, in part by ZTE Industry-University-Institute Cooperation Funds under Grant No. IA20250115003-PO0001.

References

Alikhani, S.; Charan, G.; and Alkhateeb, A. 2024. Large wireless model (LWM): A foundation model for wireless channels. *arXiv preprint arXiv:2411.08872*.

Alkhateeb, A. 2019. DeepMIMO: A generic deep learning dataset for millimeter wave and massive MIMO applications. *arXiv preprint arXiv:1902.06435*.

Busbridge, D.; Shidani, A.; Weers, F.; Ramapuram, J.; Littwin, E.; and Webb, R. 2025. Distillation scaling laws. *arXiv preprint arXiv:2502.08606*.

Catak, F. O.; Kuzlu, M.; and Cali, U. 2025. BERT4MIMO: A foundation model using bert architecture for massive mimo channel state information prediction. *arXiv preprint arXiv:2501.01802*.

Chen, X.; Xiong, Y.; and Yang, K. 2024. Robust beamforming for downlink multi-cell systems: a bilevel optimization perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 7969–7977.

Cherti, M.; Beaumont, R.; Wightman, R.; Wortsman, M.; Ilharco, G.; Gordon, C.; Schuhmann, C.; Schmidt, L.; and Jitsev, J. 2023. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2818–2829.

Chowdhery, A.; Narang, S.; Devlin, J.; Bosma, M.; Mishra, G.; Roberts, A.; Barham, P.; Chung, H. W.; Sutton, C.; Gehrmann, S.; et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240): 1–113.

Cover, T. M. 1999. *Elements of information theory*. John Wiley & Sons.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.

ElMossallamy, M. A.; Zhang, H.; Song, L.; Seddik, K. G.; Han, Z.; and Li, G. Y. 2020. Reconfigurable intelligent surfaces for wireless communications: Principles, challenges, and opportunities. *IEEE Transactions on Cognitive Communications and Networking*, 6(3): 990–1002.

Fan, L.; Chen, K.; Krishnan, D.; Katabi, D.; Isola, P.; and Tian, Y. 2024. Scaling laws of synthetic images for model training... for now. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7382–7392.

Fang, M.; Deng, S.; Zhang, Y.; Shi, Z.; Chen, L.; Pechenizkiy, M.; and Wang, J. 2024. Large language models are neurosymbolic reasoners. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 17985–17993.

Henighan, T.; Kaplan, J.; Katz, M.; Chen, M.; Hesse, C.; Jackson, J.; Jun, H.; Brown, T. B.; Dhariwal, P.; Gray, S.; et al. 2020. Scaling laws for autoregressive generative modeling. *arXiv preprint arXiv:2010.14701*.

Hoffmann, J.; Borgeaud, S.; Mensch, A.; Buchatskaya, E.; Cai, T.; Rutherford, E.; de Las Casas, D.; Hendricks, L. A.; Welbl, J.; Clark, A.; et al. 2022. Training compute-optimal large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, 30016–30030.

Jiang, F.; Pan, C.; Dong, L.; Wang, K.; Debbah, M.; Niyato, D.; and Han, Z. 2025. A comprehensive survey of large AI models for future communications: Foundations, applications and challenges. *arXiv preprint arXiv:2505.03556*.

Jiang, F.; Peng, Y.; Dong, L.; Wang, K.; Yang, K.; Pan, C.; Niyato, D.; and Dobre, O. A. 2024a. Large language model enhanced multi-agent systems for 6G communications. *IEEE Wireless Communications*.

Jiang, F.; Peng, Y.; Dong, L.; Wang, K.; Yang, K.; Pan, C.; and You, X. 2024b. Large AI model-based semantic communications. *IEEE Wireless Communications*, 31(3): 68–75.

Jiang, H.; Cui, M.; Ng, D. W. K.; and Dai, L. 2022. Accurate channel prediction based on transformer: Making mobility negligible. *IEEE Journal on Selected Areas in Communications*, 40(9): 2717–2732.

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

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Li, S.; Sun, L.; and Li, Q. 2023. CLIP-ReID: exploiting vision-language model for image re-identification without concrete text labels. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 1405–1413.

Li, Z.; Huang, J.; Liu, J.; Zhu, F.; Zhao, E.; Dodds, W.; Velingker, N.; Alur, R.; and Naik, M. 2024. Relational programming with foundational models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 10635–10644.

Liu, B.; Gao, S.; Liu, X.; Cheng, X.; and Yang, L. 2025a. WiFo: Wireless foundation model for channel prediction. *Science China Information Sciences*, 68(6): 162302.

Liu, B.; Liu, X.; Gao, S.; Cheng, X.; and Yang, L. 2024a. LLM4CP: Adapting large language models for channel prediction. *Journal of Communications and Information Networks*, 9(2): 113–125.

Liu, S.; Li, X.; Mao, Z.; Liu, P.; and Huang, Y. 2024b. Model-driven deep neural network for enhanced AoA estimation using 5G gNB. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 214–221.

Liu, X.; Gao, S.; Liu, B.; Cheng, X.; and Yang, L. 2025b. LLM4WM: Adapting LLM for wireless multi-tasking. *IEEE Transactions on Machine Learning in Communications and Networking*.

Liu, Z.; Zhang, J.; Shi, E.; Liu, Z.; Niyato, D.; Ai, B.; and Shen, X. 2024c. Graph neural network meets multi-agent reinforcement learning: Fundamentals, applications, and future directions. *IEEE Wireless Communications*, 31(6): 39–47.

Liu, Z.; Zhang, J.; Shi, E.; Zhu, Y.; Ng, D. W. K.; and Ai, B. 2024d. Cooperative multi-target positioning for cell-free massive MIMO with multi-agent reinforcement learning. *IEEE Transactions on Wireless Communications*.

Liu, Z.; Zhang, J.; Zhu, Y.; Shi, E.; and Ai, B. 2025c. Robust multidimensional graph neural networks for signal processing in wireless communications with edge-graph information bottleneck. *IEEE Transactions on Signal Processing*.

Muennighoff, N.; Rush, A.; Barak, B.; Le Scao, T.; Tazi, N.; Piktus, A.; Pyysalo, S.; Wolf, T.; and Raffel, C. A. 2023. Scaling data-constrained language models. *Advances in Neural Information Processing Systems*, 36: 50358–50376.

O'shea, T.; and Hoydis, J. 2017. An introduction to deep learning for the physical layer. *IEEE Transactions on Cognitive Communications and Networking*, 3(4): 563–575.

Ott, J.; Pirkl, J.; Stahlke, M.; Feigl, T.; and Mutschler, C. 2024. Radio foundation models: Pre-training transformers for 5G-based indoor localization. In *2024 14th International Conference on Indoor Positioning and Indoor Navigation (IPIN)*, 1–6. IEEE.

Qin, Z.; Dong, Q.; Zhang, X.; Dong, L.; Huang, X.; Yang, Z.; Khademi, M.; Zhang, D.; Awadalla, H. H.; Fung, Y. R.; et al. 2025. Scaling laws of synthetic data for language models. *arXiv preprint arXiv:2503.19551*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 8748–8763. PMLR.

Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I.; et al. 2018. Improving language understanding by generative pre-training.

Shao, J.; Tong, J.; Wu, Q.; Guo, W.; Li, Z.; Lin, Z.; and Zhang, J. 2024. Wirelessllm: Empowering large language models towards wireless intelligence. *arXiv preprint arXiv:2405.17053*.

Shen, Y.; Shao, J.; Zhang, X.; Lin, Z.; Pan, H.; Li, D.; Zhang, J.; and Letaief, K. B. 2024. Large language models empowered autonomous edge AI for connected intelligence. *IEEE Communications Magazine*, 62(10): 140–146.

Tong, J.; Guo, W.; Shao, J.; Wu, Q.; Li, Z.; Lin, Z.; and Zhang, J. 2025. Wirelessagent: Large language model agents for intelligent wireless networks. *arXiv preprint arXiv:2505.01074*.

Tse, D.; and Viswanath, P. 2005. *Fundamentals of wireless communication*. Cambridge University Press.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Xiao, M.; Mumtaz, S.; Huang, Y.; Dai, L.; Li, Y.; Matthaiou, M.; Karagiannidis, G. K.; Björnson, E.; Yang, K.; Ghosh, A.; et al. 2017. Millimeter wave communications for future mobile networks. *IEEE Journal on Selected Areas in Communications*, 35(9): 1909–1935.

Xiong, Y.; Chen, X.; Ye, X.; Chen, H.; Lin, Z.; Lian, H.; Su, Z.; Huang, W.; Niu, J.; Han, J.; et al. 2024. Temporal scaling law for large language models. *arXiv preprint arXiv:2404.17785*.

Xu, S.; Thomas, C. K.; Hashash, O.; Muralidhar, N.; Saad, W.; and Ramakrishnan, N. 2024. Large multi-modal models (LMMs) as universal foundation models for AI-native wireless systems. *IEEE Network*.

Yang, T.; Zhang, P.; Zheng, M.; Shi, Y.; Jing, L.; Huang, J.; and Li, N. 2025. WirelessGPT: A generative pre-trained multi-task learning framework for wireless communication. *IEEE Network*.

Ye, H.; Liang, L.; Li, G. Y.; and Juang, B.-H. 2020. Deep learning-based end-to-end wireless communication systems with conditional GANs as unknown channels. *IEEE Transactions on Wireless Communications*, 19(5): 3133–3143.

Yin, Z.; Xu, W.; Xie, R.; Zhang, S.; Ng, D. W. K.; and You, X. 2022. Deep CSI compression for massive MIMO: A self-information model-driven neural network. *IEEE Transactions on Wireless Communications*, 21(10): 8872–8886.

Zhai, X.; Kolesnikov, A.; Houlsby, N.; and Beyer, L. 2022. Scaling vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12104–12113.

Zhang, C.; Bu, W.; Ren, Z.; Liu, Z.; Wu, Y.-C.; and Wong, N. 2025a. Nonparametric Teaching for Graph Property Learners. In *ICML*.

Zhang, C.; Cao, X.; Liu, W.; Tsang, I.; and Kwok, J. 2023a. Nonparametric Iterative Machine Teaching. In *ICML*.

Zhang, C.; Cao, X.; Liu, W.; Tsang, I.; and Kwok, J. 2023b. Nonparametric Teaching for Multiple Learners. In *NeurIPS*.

Zhang, C.; Luo, S. T. S.; Li, J. C. L.; Wu, Y.-C.; and Wong, N. 2024. Nonparametric Teaching of Implicit Neural Representations. In *ICML*.

Zhang, J.; Liu, Z.; Zhu, Y.; Shi, E.; Xu, B.; Yuen, C.; Niyato, D.; Debbah, M.; Jin, S.; Ai, B.; et al. 2025b. Multi-agent reinforcement learning in wireless distributed networks for 6G. *arXiv preprint arXiv:2502.05812*.

Zhou, G.; Hong, Y.; and Wu, Q. 2024. NavGPT: Explicit reasoning in vision-and-language navigation with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 7641–7649.