In-Context Radio Map Estimation via Ripple Autoregressive Modeling

Yuanzhe Peng
Department of ECE
University of Florida
pengy1@ufl.edu

Jie Xu
Department of ECE
University of Florida
jie.xu@ufl.edu

Abstract

Accurate radio map estimation is critical for wireless applications such as coverage planning, localization, and network deployment. However, most existing methods follow a supervised learning mindset, designing various U-Net-based model architectures or loss functions that rely on costly labeled data and delicate model training. Inspired by the remarkable generalization ability of large language models, we are the first to formulate radio map estimation as an in-context learning (ICL) problem, leveraging a pretrained large autoregressive vision model (LAVM) to predict the radio map for a new transmitter (Tx) position, prompted by a few input-output demonstrations without requiring model updates. We propose RIPPLE, a novel ICL framework that integrates visual tokenization with a ripple autoregressive modeling strategy to explicitly capture the causal structure of wireless signal propagation from the Tx outward. Furthermore, we introduce a two-stage generation strategy for coarse-to-fine prediction to better model non-line-of-sight (NLoS) propagation effects. Extensive experiments demonstrate that RIPPLE outperforms ICL baselines, highlighting its effectiveness and generalizability in radio map estimation.

1 Introduction

Accurate radio map estimation is critical for wireless applications such as coverage planning, localization, and network deployment [1]. A radio map provides spatially resolved estimates of pathloss (PL), which quantifies signal attenuation in decibels (dB) throughout the environment for a given transmitter (Tx) location. In this paper, we focus on radio map estimation in indoor environments, which is particularly challenging due to complex propagation effects such as reflection, transmission, and diffraction. Traditional *model-driven* approaches, including the log-distance PL model and 3GPP channel models, offer physical interpretability but often fail to generalize across diverse and irregular indoor layouts [2]. More recently, *data-driven* deep learning methods have improved estimation performance by formulating the task as a supervised learning problem, similar to image-to-image translation from scene map input to the corresponding PL map output. Most existing deep learning-based methods focus on three main directions: (1) designing various encoder-decoder architectures inspired by U-Net and implemented using CNNs or vision transformers (ViTs) [3–8]; (2) customizing task-specific loss functions [9–11]; (3) adopting various feature fusion techniques or applying extensive data augmentation [5, 12]. All of these methods fall under the supervised learning paradigm and aim to improve the performance of trained models for radio map estimation. Despite these advances, two fundamental **limitations** remain under the supervised learning mindset. First, collecting labeled PL maps is costly, requiring computationally expensive ray tracing or labor intensive field measurements. Second, even with sufficient data, new deployments still require delicate model training.

Recent progress in large language models (LLMs), such as GPTs, has demonstrated remarkable generalization across diverse tasks by leveraging only a few examples. This has led to the emergence

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: AI and ML for Next-Generation Wireless Communications and Networking (AI4NextG).

of in-context learning (ICL), a novel paradigm in which models perform new tasks by conditioning on a small set of input-output examples, commonly referred to as *prompts*, without requiring model updates [13, 14]. Motivated by these developments, we present the first attempt to explore the feasibility of applying the ICL paradigm to radio map estimation, thereby eliminating the need for costly labeled data and delicate model training.

However, adapting ICL to radio map estimation introduces two unique **challenges** that do not arise in conventional language or vision tasks. First, unlike natural language inputs, which consist of discrete word tokens, visual inputs are continuous and spatially structured. This makes visual tokenization nontrivial but necessary for leveraging transformer-based models. Second, most existing methods use a raster scan order for visual autoregressive modeling [15], which fails to capture the inherent spatial causality in radio propagation. Specifically, *wireless signals propagate directionally, governed by physical laws, expanding outward from the transmitter (Tx) like a ripple.* This directional expansion reflects the causal structure that should be considered in autoregressive modeling, where each token prediction is conditioned on causal relationships defined by previously generated tokens.

To address these challenges, we propose RIPPLE, a novel ICL framework that integrates visual tokenization with a ripple autoregressive modeling strategy for radio map estimation. The ripple order proceeds outward from the Tx, explicitly capturing the causal nature of signal propagation. Given a few prompts, RIPPLE leverages a pretrained large autoregressive vision model (LAVM) to predict the radio map for a new Tx position within the same building layout. Furthermore, we introduce a two-stage generation strategy for coarse-to-fine prediction, in which the second stage incorporates the coarse estimate from the first stage into the prompt set for self-refinement, enabling the model to leverage global context and better capture non-line-of-sight (NLoS) propagation effects.

Contributions. (1) We formulate radio map estimation as an ICL problem without requiring model updates. (2) We propose RIPPLE, a novel ICL framework that adopts ripple autoregressive modeling to explicitly capture the causal structure of signal propagation from the Tx outward, along with a two-stage generation strategy that effectively captures NLoS effects. (3) Extensive experiments show that RIPPLE outperforms three baselines, demonstrating its effectiveness in radio map estimation.

2 Related Work

Radio Map Estimation. Radio map estimation has evolved through two major stages. The first is *model-driven*, where traditional approaches rely on analytical or physics-based models such as the log-distance PL model, standardized 3GPP channel models, and ray-tracing simulations [2]. While these methods offer physical interpretability, they often fail to generalize across diverse and irregular indoor layouts. The second is *data-driven*, inspired by the success of deep learning, where recent methods formulate radio map estimation as a supervised learning task in which neural networks predict radio maps from multi-channel scene inputs [3, 5, 9, 12]. These models typically adopt encoder-decoder architectures inspired by U-Net and implemented using CNNs or ViTs [4, 8]. However, whether through model-architectural modifications, loss function design, or improved feature fusion, they still rely on costly labeled data and delicate model training under the supervised learning mindset.

In-Context Learning. Recent advances in LLMs such as GPTs have demonstrated remarkable generalization across diverse tasks, leading to the emergence of ICL, a paradigm enabling models to perform new tasks prompted by few input-output demonstrations without model updates [16]. While ICL was first introduced in the language domain, recent works have extended it to the visual domain [17], where images are tokenized using vector quantization (VQ)-based techniques such as VQ-VAE [18, 19] or VQ-GAN [20], followed by next-token prediction [15]. These methods have shown promising performance on vision tasks such as semantic segmentation, image inpainting [21], and other unseen tasks. However, applying ICL to radio map estimation remains challenging, as the directional and causal propagation of wireless signals from the Tx outward, governed by physical laws, is poorly captured by the raster-scan ordering commonly used in visual autoregressive modeling.

3 Problem Formulation

We consider the problem of indoor radio map estimation, which aims to predict a spatially resolved PL map for a given Tx location within a building layout. The PL value at each point quantifies signal attenuation in dB. We discretize each building layout into a 2D spatial grid of size $H \times W$, represented by a coordinate set $\mathcal{X} = \{x_i\}_{i=1}^{N}$, where each x_i denotes the i-th location and $N = H \cdot W$ is the total number of grid points. Given a Tx position $x_{\mathrm{Tx}} \in \mathcal{X}$, the goal is to estimate the PL value $\hat{y}_{x_i}$ at

every x_i, resulting in the complete predicted radio map: $\hat{\mathbf{y}} = \{\hat{y}_{x_1}, \ldots, \hat{y}_{x_N}\} \in \mathbb{R}^{H \times W}$. The input is $\mathbf{I} \in \mathbb{R}^{H \times W \times C}$, where C is the number of channels. In this paper, the input involves three channels: $\mathbf{I}^{\mathrm{D}}_{x_i}$ is the distance from x_i to the Tx (in meters); $\mathbf{I}^{\mathrm{T}}_{x_i}$ is the transmittance at x_i (in dB); and $\mathbf{I}^{\mathrm{R}}_{x_i}$ is the reflectance at x_i (in dB). Notably, $\mathbf{I}^{\mathrm{T}}$ and $\mathbf{I}^{\mathrm{R}}$ are physically meaningful only at interfaces and serve as prior knowledge for a given building layout with fixed materials, regardless of Tx positions.

Conventional Formulation. Most existing work formulates radio map estimation as a supervised learning problem: a model f_θ is trained on a labeled dataset to learn a mapping from the input scene to the output PL map, i.e., $\hat{\mathbf{y}} = f_\theta(\mathbf{I})$. The model parameters θ are typically optimized by minimizing loss between predicted and ground-truth maps: $\min_\theta \mathcal{L}(\hat{\mathbf{y}}, \mathbf{y})$, where $\mathcal{L}$ denotes a task-specific loss. The model is usually implemented using a U-Net-based architecture with CNNs or ViTs. However, the design of different model architectures or loss functions *still follows the same supervised learning mindset*, which inherently depends on costly labeled data and delicate model training.

Our ICL Formulation. Instead of training a task-specific model, we leverage a pretrained LAVM to predict the radio map for a new Tx position within the same building layout by "learning" from a few input-output demonstrations, referred to as *prompts*. Let $\mathcal{P} = \{(\mathbf{I}_k, \mathbf{y}_k)\}_{k=1}^K$ denote a set of prompts (with a small K in practice), where each $\mathbf{I}_k$ is a scene input associated with a known Tx location $x^k_{\mathrm{Tx}} \in \mathcal{X}$, and $\mathbf{y}_k$ is the corresponding PL map. Given a test input (i.e., query) $\mathbf{I}_q$ with an arbitrary Tx position x^q_{Tx} from the same layout, the LAVM predicts $\hat{\mathbf{y}}_q$ by leveraging in-context information (i.e., implicit input-output mappings) from prompts $\mathcal{P}$:

$$\hat{\mathbf{y}}_q = f_\tau(\mathcal{P}, \mathbf{I}_q), \quad (1)$$

where f_τ represents a pretrained LAVM with frozen parameters τ. This formulation not only shows promising generalization without requiring model updates but also aligns with practical constraints, as most large models (e.g., GPTs), although capable of handling diverse tasks, are accessible only via fixed APIs and cannot be retrained by end users for specific tasks.

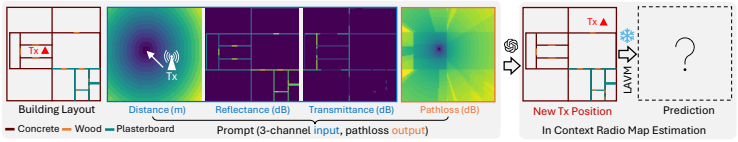

Figure 1: In-context radio map estimation for arbitrary Tx positions.

4 Methodology

Prompt Construction. Since only a limited number of costly ray-tracing simulations can be performed for the query layout to construct prompts, we aim to select K representative Tx positions that are spatially well distributed across the building layout. Let $\mathcal{X}_{\mathrm{Tx}} = \{x_n\}_{n=1}^N$ denote all candidate Tx positions for a given layout. We adopt K-means clustering to partition $\mathcal{X}_{\mathrm{Tx}}$ into K clusters $\{\mathcal{C}_k\}_{k=1}^K$ with corresponding centroids $\{\mu_k\}_{k=1}^K$, where each cluster $\mathcal{C}_k$ contains Tx positions assigned to centroid μ_k. For each cluster $\mathcal{C}_k$, we select the position closest to its centroid: $x^k_{\mathrm{Tx}} = \arg\min_{x_n \in \mathcal{C}_k} \|x_n - \mu_k\|_2$. This yields K spatially diverse Tx positions $\mathcal{K}_{\mathrm{Tx}} = \{x^1_{\mathrm{Tx}}, \ldots, x^K_{\mathrm{Tx}}\}$, where ray-tracing simulation is performed to construct the prompt set $\mathcal{P} = \{(\mathbf{I}_k, \mathbf{y}_k)\}_{k=1}^K$.

Visual Tokenization. After constructing the prompts $\mathcal{P}$ via limited ray-tracing, we tokenize the visual input into discrete tokens for transformer processing. Specifically, we use a pretrained VQ-GAN as a tokenizer with an encoder $E(\cdot)$ and a quantizer $Q(\cdot)$. The encoder projects the input into a latent tensor $\hat{\mathbf{z}}_k = E(\mathbf{I}_k) \in \mathbb{R}^{h \times w \times d}$, where $h = H/f$, $w = W/f$, d is the embedding dimension, and f is the downsampling factor imposed by the VQ-GAN. Each $\hat{\mathbf{z}}^{i,j}_k$ is quantized to the nearest codeword in a discrete codebook $\mathcal{Z} = \{\mathbf{z}_v\}_{v=1}^V \subset \mathbb{R}^d$, defined as $\mathbf{z}^{i,j}_k = Q(\hat{\mathbf{z}}^{i,j}_k) = \arg\min_{\mathbf{z}_v \in \mathcal{Z}} \|\hat{\mathbf{z}}^{i,j}_k - \mathbf{z}_v\|$. This process produces a token grid $\mathbf{z}_k \in \{1, \ldots, V\}^{h \times w}$, with each entry indexing a codeword in $\mathcal{Z}$. We then perform tokenization on the inputs and outputs of the prompts, obtaining $\{\mathbf{z}^{\mathrm{in}}_k\}$ and $\{\mathbf{z}^{\mathrm{out}}_k\}$.

Radial Ripple Ordering. In contrast to the raster scan order commonly used in visual autoregressive modeling, which fails to capture the causal structure of signal propagation in radio map estimation, we propose a novel radial ripple order that emulates the outward propagation of waves from a Tx, as shown in Fig. 2. By sorting tokens based on their latent distance to the Tx, this ordering reflects the time-like nature of wireless propagation and enables the sequential autoregressive model to learn spatial causality. Based on the previous tokenization, we flatten the token grids into informative token sequences. We define the latent coordinate set for each input k as $\tilde{\mathcal{X}}_k = \{\tilde{x}_{k,t}\}_{t=1}^T$, where $\tilde{x}_{k,t}$ denotes the spatial coordinate in the latent space corresponding to the t-th token, and $T = h \cdot w$.

First, in order to impose a radial ripple structure, we calculate the distance from each latent token position $\tilde{x}_{k,t}$ to the Tx position: $r_{k,t} = \|\tilde{x}_{k,t} - x_{\text{Tx}}^k\|_2$. This radial distance serves as a proxy for the arrival time of the wavefront at $\tilde{x}_{k,t}$. We then divide the token grid into L concentric ripple layers based on a set of increasing radial thresholds $\{r_\ell\}_{\ell=1}^L$, with $r_0 = 0$ and r_L chosen to span the full latent space. Each ripple layer groups tokens that lie within a specific radial shell: $\mathcal{R}_k^\ell = \{\tilde{x}_{k,t} \in \tilde{\mathcal{X}}_k \mid r_{\ell-1} < \|\tilde{x}_{k,t} - x_{\text{Tx}}^k\|_2 \le r_\ell\},\ \ell \in [L]$. Notably, when the Tx is off-center, some ripples may form incomplete rings, but the resulting sequence still reflects spatial causality, similar to how real ripples disrupted by a shoreline can still propagate.

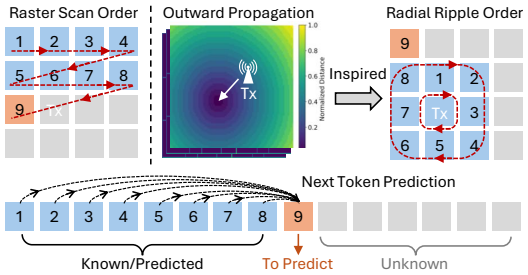

Figure 2: Inspired by the causal structure of signal propagation from the Tx outward, we propose a radial ripple order for autoregressive modeling.

Second, to associate environmental context with radio propagation behavior, we construct interleaved token pairs within each ripple layer by alternating input and output tokens. Specifically, for each token position $\tilde{x}_{\ell,j} \in \mathcal{R}_k^\ell$, where $j = 1, \ldots, R_k^\ell$ and $R_k^\ell = |\mathcal{R}_k^\ell|$, we define the interleaved token pair as $\mathbf{s}_k^{\ell,j} = [z_k^{\text{in}}(\tilde{x}_{\ell,j}),\ z_k^{\text{out}}(\tilde{x}_{\ell,j})]$. Thus, the complete token sequence for the ℓ-th ripple layer is $\mathbf{s}_k^\ell = [z_k^{\text{in}}(\tilde{x}_{\ell,1}), z_k^{\text{out}}(\tilde{x}_{\ell,1}), \ldots, z_k^{\text{in}}(\tilde{x}_{\ell,R_k^\ell}), z_k^{\text{out}}(\tilde{x}_{\ell,R_k^\ell})]$. We then concatenate all ripple layers in ascending order of their distance to the Tx to construct the k-th prompt sequence $\mathbf{s}_k = [\mathbf{s}_k^1, \mathbf{s}_k^2, \ldots, \mathbf{s}_k^L]$. Finally, we construct the prompt sequence $\mathcal{P}^*$ by concatenating the ripple-ordered token sequences from all prompts: $\mathcal{P}^* = [\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_K]$.

The radial ripple ordering and the interleaved input-output construction offer several advantages. First, they impose a spatially *causal structure* that mimics the physical propagation of wireless signals from the Tx outward. Second, they preserve local alignment between scene inputs and corresponding PL responses, similar in spirit to the *cross-attention* mechanism used in NLP translation tasks.

Two-Stage Generation. With the token sequence constructed via radial ripple ordering that explicitly captures the causal structure of propagation, we perform next-token prediction using a causal transformer. Indoor environments pose challenges due to phenomena such as multi-path reflection, diffraction, and NLoS effects. To better model such propagation and overcome the limitations of unidirectional autoregressive prediction, we propose a two-stage autoregressive generation strategy.

Stage I: Causal Autoregressive Prediction. Given the query input $\mathbf{I}_q$ with the Tx position x_{Tx}^q, we perform visual tokenization and obtain $\mathbf{z}_q^{\text{in}}$. We define $\tilde{\mathcal{X}}_q = \{\tilde{x}_{q,t}\}_{t=1}^T$, where each $\tilde{x}_{q,t}$ corresponds to the position of the t-th token. We compute the radial distance from each latent token position to the Tx as $r_{q,t} = \|\tilde{x}_{q,t} - x_{\text{Tx}}^q\|_2$, and partition the grid into L ripple layers using a set of radial thresholds $\{r_\ell\}_{\ell=1}^L$ and $\mathcal{R}_q^\ell = \{\tilde{x}_{q,t} \in \tilde{\mathcal{X}}_q \mid r_{\ell-1} < \|\tilde{x}_{q,t} - x_{\text{Tx}}^q\|_2 \le r_\ell\}_{\ell \in [L]}$. We then concatenate all L concentric ripple layers in ascending order, i.e., $\mathcal{R}_q = [\mathcal{R}_q^1, \mathcal{R}_q^2, \ldots, \mathcal{R}_q^L]$, and prepare the incomplete token sequence for subsequent prediction. Common implementation details such as beginning-of-sequence `[BOS]`, end-of-sequence `[EOS]`, and separator tokens `[SEP]` are omitted below. Given a pretrained LAVM f_τ with frozen parameters τ, we perform next-token prediction autoregressively following the radial ripple order, capturing the sequential dependencies in signal propagation. This corresponds to a conditional factorization of the likelihood over the predicted token sequence: $p(\mathbf{z}_q^{(\text{I})} \mid \mathcal{P}^*, \mathbf{z}_q^{\text{in}}) = \prod_{t=1}^T p(z_{q,t}^{(\text{I})} \mid z_{q,1}^{(\text{I})}, \ldots, z_{q,t-1}^{(\text{I})}, \mathcal{P}^*, \mathbf{z}_q^{\text{in}})$.

Stage II: Refinement with Self-Context. To mitigate the limitations of the unidirectional nature of the autoregressive model, we introduce a refinement stage that improves predictions using the self-context. We construct a self-refinement prompt by interleaving the input and predicted tokens from Stage I: $\mathcal{P}_q^* = [z_{q,1}^{\text{in}}, z_{q,1}^{(\text{I})}, \ldots, z_{q,T}^{\text{in}}, z_{q,T}^{(\text{I})}]$. We concatenate $\mathcal{P}_q^*$ with the original prompt sequence to form a self-refinement context that captures dependencies from NLoS effects: $\mathcal{P}_{\text{ref}}^* = [\mathcal{P}^*, \mathcal{P}_q^*]$. We then perform next-token prediction via the following conditional factorization of the likelihood:

$$p\left(\mathbf{z}_q^{(\text{II})} \mid \mathcal{P}_{\text{ref}}^*, \mathbf{z}_q^{\text{in}}\right) = \prod_{t=1}^T p\left(z_{q,t}^{(\text{II})} \mid z_{q,<t}^{(\text{II})}, \mathcal{P}_{\text{ref}}^*, \mathbf{z}_q^{\text{in}}\right), \tag{2}$$

where $\mathbf{z}_q^{(\text{II})}$ reassembles a discrete token grid of dimension $\{1, \ldots, V\}^{h \times w}$. The token indices are mapped to codewords in the codebook $\mathcal{Z}$ via lookup $Z(\cdot)$, and then fed into the VQ-GAN decoder

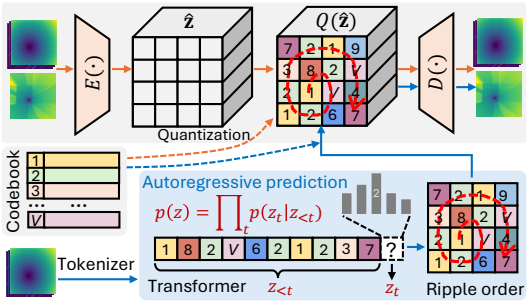

Figure 3: RIPPLE consists of two key modules: a *tokenizer* for visual input tokenization and a *transformer* for ripple autoregressive prediction (blue).

$D(\cdot)$ to reconstruct the predicted radio map:

$$\hat{\mathbf{y}}_q = D(Z(\mathbf{z}_q^{(\mathrm{II})})). \tag{3}$$

This two-stage autoregressive generation enables coarse-to-fine prediction, where Stage I models the causal structure of wireless signal propagation from the Tx outward. Stage II leverages the self-refinement context to improve predictions by capturing complex NLoS effects with a global view, thereby mitigating the limitations of causal masking and the unidirectional nature of visual autoregressive modeling.

The overall process is analogous to how students solve problems: they first learn from several example questions with solutions (prompts), then attempt an initial solution for a new problem (Stage I), and finally refine it after gaining a global understanding (Stage II). Moreover, we provide theoretical background from recent studies to explain why ICL performs well. ICL has been demonstrated to be equivalent to gradient-based optimization; specifically, performing K steps of gradient descent on a training set $\mathcal{D}^{\text{train}}$ is equivalent to applying a Transformer with K linear self-attention layers, when $\mathcal{D}^{\text{train}}$ is provided as in-context data $\mathcal{D}^{\text{context}}$ [22]. Similar theorems are established in [23].

5 Experiments

Dataset. We use a publicly available indoor radio map dataset generated by ray tracing [24]. Each sample is a unique combination of building layout (25 total), frequency band (3 total), and antenna radiation pattern (5 total), with Tx positions ranging from 50 to 80 in each layout. The input includes reflectance, transmittance, and distance to the Tx, and the output is the corresponding PL map.

Baselines. We compare RIPPLE against three baselines following the ICL principle without requiring model updates. *Baseline 1:* Randomly select K prompts and apply raster-scan order in autoregressive modeling; *Baseline 2:* Select K informative prompts and apply raster-scan order in autoregressive modeling; *Baseline 3:* Randomly select K prompts and apply ripple order in autoregressive modeling.

Implementation and Metrics. We implement our framework using `PyTorch 2.6` and conduct all experiments on NVIDIA B200 GPUs. We adopt pre-trained VQ-GAN [25] and LLaMA models [26] for visual tokenization and autoregressive prediction, respectively. Since the visual inputs consist of the multi-channel physical features that differ from natural images, we pretrain the VQ-GAN to improve tokenization and reconstruction quality prior to ICL on 20 building layouts (B1-B20). Although no input-output mapping is learned, we use only five other layouts (B21-B25) for testing. The VQ-GAN model (with downsampling factor f=16) tokenizes each resized image (256×256 pixels) into 256 discrete tokens using an 8192-entry codebook with dimension 64. For sequence modeling, we use the LLaMA-7B model released by LVM [21], pretrained on 50 diverse datasets and 420B tokens covering various visual tasks and demonstrating strong generalization to unseen scenarios. We adopt the *root mean squared error (RMSE)* (↓ denotes lower is better) as the evaluation metric. We report the mean RMSE and standard deviation over five independent runs.

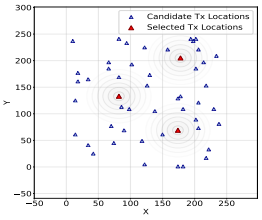

Figure 4: Selected Tx.

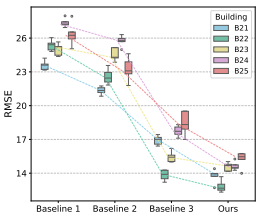

Figure 5: RMSE (↓).

Results and Analysis. We conduct experiments where prompts are constructed from selected representative Tx positions (K=3 by default), and the remaining positions serve as queries for evaluation. As shown in Fig. 4, the selected Tx positions (red) are well distributed, providing informative context (e.g., signal transmission and reflection at different positions) for queries with arbitrary Tx positions (blue). As shown in Fig. 5, our method achieves better performance (lower RMSE ↓) than all baselines across various building layouts. This improvement is attributed not only to the selection of K informative prompts but also to the design of a radial ripple order for autoregressive modeling, which more effectively captures the causal nature of wireless signal propagation from the Tx outward. Note that when the building layout is relatively simple (e.g., B22),

or the Tx position of the query is spatially close to that of the prompt, prediction performance may improve. However, the Tx position of the query is unknown during prompt construction, which aligns with real-world deployment. We also present a visual comparison of predictions in Fig. 6. Compared to raster-scan autoregressive modeling (Baselines 1 and 2), which is widely used in visual autoregressive tasks, our ripple autoregressive modeling predicts radio maps that match the target more closely. This improvement is intuitive, as RIPPLE not only learns implicit input-output mappings from prompts but also captures the causal structure of signal propagation from the Tx outward.

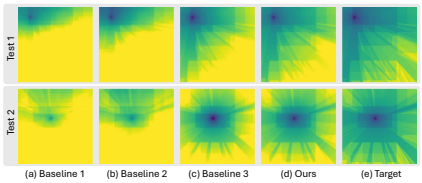

Figure 6: Visual comparison.

To further analyze error distributions, instead of evaluating the predicted radio map using RMSE, we compute the region-wise $\mathrm{RMSE}_{\mathcal{R}} = \sqrt{1/|\mathcal{R}| \sum_{i \in \mathcal{R}} (y_i - \hat{y}_i)^2}$ based on the distance from the Tx. Specifically, we define three circular regions centered at the Tx, with increasing radii: $r \leq 0.3R$, $r \leq 0.6R$, and $r \leq 0.9R$, where R is the radius of the smallest enclosing circle of the radio map. As shown in Fig. 7, our method (and Baseline 3, which also uses ripple autoregressive modeling) achieves a low RMSE in the inner region, which gradually increases toward the outer region. This improvement over Baselines 1 and 2 is intuitive because ripple autoregressive modeling better captures the causal effect of signal propagation from the Tx outward. We also identify future work in addressing outer-region errors, as these regions often involve complex NLoS effects that are difficult to capture due to the unidirectional nature of autoregressive modeling.

Comparison of Error Distribution	(Region-wise) $\mathrm{RMSE}_{\mathcal{R}}$		
	$r \leq 0.3R$	$r \leq 0.6R$	$r \leq 0.9R$
Baseline 1	13.82 ± 1.95	16.82 ± 2.27	18.97 ± 3.04
Baseline 2	13.16 ± 1.74	15.63 ± 2.16	18.42 ± 2.85
Baseline 3	3.85 ± 0.57	9.34 ± 1.62	13.35 ± 1.97
Ours	3.17 ± 0.48	8.45 ± 1.31	10.94 ± 1.64

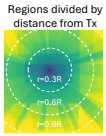

Figure 7: Comparison of $\mathrm{RMSE}_{\mathcal{R}}$.

To analyze the impact of the number of prompts, we vary K and present both RMSE and visual comparisons. Here, K is kept relatively small due to LLaMA's input length constraint and the few-shot nature of ICL. As shown in Fig. 8, increasing the number of prompts improves performance. This is intuitive, as more prompts in ICL provide richer implicit input-output mappings, analogous to having more training samples or rounds in supervised learning. Moreover, since prompts are selected based on spatial diversity rather than randomly, the quality and variety of in-context information are improved, helping the model assign a higher probability to the correct next token.

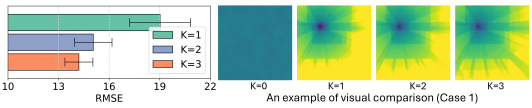

Figure 8: Impact of number of prompts K.

In addition to the above experiments, other factors may also impact results. (1) The number of tokens per image. Given the fixed input length constraint, more prompts can be included by reducing the number of tokens per image, which introduces a trade-off between prompt quantity and token utility. (2) The arrangement of prompts; e.g., different spatial configurations of demonstrations can influence how well the model captures contextual dependencies. (3) The configurations of pretrained models, such as the downsampling factor f, the codebook size V, and the parameter size of LLaMA. However, these adjustments are orthogonal to our main contribution, which lies in ripple autoregressive modeling for capturing the causal structure of signal propagation from the Tx outward.

Inspiration for Future Work. Our work opens several promising directions for exploration. First, while RIPPLE achieves strong performance with token-based autoregressive modeling, more advanced strategies such as scale-based autoregressive modeling could be explored to further mitigate long-range error accumulation. Second, extending RIPPLE to dynamically changing environments (e.g., human motion, furniture reconfiguration, or varying signal conditions) would benefit from temporal adaptation techniques such as online prompt updating or streaming inference. Third, applying RIPPLE within collaborative or distributed learning frameworks (e.g., federated or edge-assisted ICL) could address privacy constraints and reduce communication overhead in multi-agent wireless systems.

6 Conclusion

In conclusion, we are the first to formulate radio map estimation as an ICL problem without requiring model updates, and we propose RIPPLE, a novel framework that integrates visual tokenization with ripple autoregressive modeling to explicitly capture the causal structure of signal propagation from the transmitter outward. We further introduce a two-stage generation strategy for coarse-to-fine prediction to better model NLoS propagation. Extensive experiments demonstrate that RIPPLE outperforms three ICL baselines, highlighting its effectiveness and generalizability in radio map estimation.

Acknowledgments

The work of Yuanzhe Peng and Jie Xu is partially supported by NSF under grant 2515982.

References

[1] S. Zhang, B. Choi, F. Ouyang, and Z. Ding, "Physics-Inspired Machine Learning for Radiomap Estimation: Integration of Radio Propagation Models and Artificial Intelligence," *IEEE communications magazine*, vol. 62, no. 8, pp. 155–161, 2024.

[2] B. Feng, M. Zheng, W. Liang, and L. Zhang, "A Recent Survey on Radio Map Estimation Methods for Wireless Networks," *Electronics*, vol. 14, no. 8, p. 1564, 2025.

[3] R. Levie, Ç. Yapar, G. Kutyniok, and G. Caire, "RadioUNet: Fast Radio Map Estimation with Convolutional Neural Networks," *IEEE Transactions on Wireless Communications*, vol. 20, no. 6, pp. 4001–4015, 2021.

[4] S. Bakirtzis, K. Qiu, J. Chen, H. Song, J. Zhang, and I. Wassell, "Rigorous Indoor Wireless Communication System Simulations With Deep Learning-Based Radio Propagation Models," *IEEE Journal on Multiscale and Multiphysics Computational Techniques*, 2024.

[5] B. Feng, M. Zheng, W. Liang, and L. Zhang, "IPP-Net: A Generalizable Deep Neural Network Model for Indoor Pathloss Radio Map Prediction," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–2.

[6] S. Bakirtzis, J. Chen, K. Qiu, J. Zhang, and I. Wassell, "EM DeepRay: An Expedient, Generalizable, and Realistic Data-Driven Indoor Propagation Model," *IEEE Transactions on Antennas and Propagation*, vol. 70, no. 6, pp. 4140–4154, 2022.

[7] X. Li, R. Liu, S. Xu, S. G. Razul, and C. Yuen, "TransPathNet: A Novel Two-Stage Framework for Indoor Radio Map Prediction," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–2.

[8] P. Q. Viet and D. Romero, "Spatial Transformers for Radio Map Estimation," *arXiv preprint arXiv:2411.01211*, 2024.

[9] W. Lu, Z. Lu, J. Yan, and S. Gao, "SIP2Net: Situational-Aware Indoor Pathloss-Map Prediction Network for Radio Map Generation," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–2.

[10] M. Wu, M. Skocaj, and M. Boban, "Enhancing Convolutional Models for Indoor Radio Mapping via Ray Marching," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–2.

[11] C. T. Cisse, O. Baala, V. Guillet, F. Spies, and A. Caminada, "Generalizable Indoor Path Loss Prediction," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–2.

[12] J.-H. Lee and A. F. Molisch, "A Scalable and Generalizable Pathloss Map Prediction," *IEEE Transactions on Wireless Communications*, 2024.

[13] S. Sia, D. Mueller, and K. Duh, "Where does In-context Learning Happen in Large Language Models?" *Advances in Neural Information Processing Systems*, vol. 37, pp. 32 761–32 786, 2024.

[14] X. Wang, W. Wang, Y. Cao, C. Shen, and T. Huang, "Images Speak in Images: A Generalist Painter for In-Context Visual Learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 6830–6839.

[15] J. Xiong, G. Liu, L. Huang, C. Wu, T. Wu, Y. Mu, Y. Yao, H. Shen, Z. Wan, J. Huang *et al.*, "Autoregressive Models in Vision: A Survey," *Transactions on Machine Learning Research*, 2025.

[16] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing," *ACM computing surveys*, vol. 55, no. 9, pp. 1–35, 2023.

[17] Z. Hao, J. Guo, C. Wang, Y. Tang, H. Wu, H. Hu, K. Han, and C. Xu, "Data-efficient Large Vision Models through Sequential Autoregression," in *Proceedings of the 41st International Conference on Machine Learning*, 2024, pp. 17 572–17 596.

[18] A. Van Den Oord, O. Vinyals *et al.*, "Neural Discrete Representation Learning," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[19] A. Razavi, A. Van den Oord, and O. Vinyals, "Generating Diverse High-Fidelity Images with VQ-VAE-2," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[20] P. Esser, R. Rombach, and B. Ommer, "Taming Transformers for High-Resolution Image Synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12 873–12 883.

[21] Y. Bai, X. Geng, K. Mangalam, A. Bar, A. L. Yuille, T. Darrell, J. Malik, and A. A. Efros, "Sequential Modeling Enables Scalable Learning for Large Vision Models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 22 861–22 872.

[22] J. Von Oswald, E. Niklasson, E. Randazzo, J. Sacramento, A. Mordvintsev, A. Zhmoginov, and M. Vladymyrov, "Transformers Learn In-Context by Gradient Descent," in *International Conference on Machine Learning*. PMLR, 2023, pp. 35 151–35 174.

[23] M. E. Sander, R. Giryes, T. Suzuki, M. Blondel, and G. Peyré, "How do Transformers Perform In-Context Autoregressive Learning?" in *Proceedings of the 41st International Conference on Machine Learning*, 2024, pp. 43 235–43 254.

[24] S. Bakirtzis, Çagkan Yapar, K. Qui, I. Wassell, and J. Zhang, "Indoor Radio Map Dataset," *IEEE Dataport*, 2024.

[25] H. Chang, H. Zhang, J. Barber, A. Maschinot, J. Lezama, L. Jiang, M.-H. Yang, K. P. Murphy, W. T. Freeman, M. Rubinstein *et al.*, "Muse: Text-To-Image Generation via Masked Generative Transformers," in *International Conference on Machine Learning*. PMLR, 2023, pp. 4055–4075.

[26] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar *et al.*, "LLaMA: Open and Efficient Foundation Language Models," *arXiv preprint arXiv:2302.13971*, 2023.