HOLOROPE: HOLOGRAPHIC ROPE FOR HIERARCHICAL WIRELESS FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Wireless foundation models (WFMs) extend the pretraining paradigm of foundation models to wireless communications, learning high-quality representations from large-scale channel data that serve diverse downstream wireless tasks. The input to WFMs is typically channel state information (CSI), which has two axes that index antennas and subcarriers and sample the propagation channel in space and in frequency, respectively. Adjacent subcarriers in wireless data typically lie within the coherence bandwidth, and we exploit this property by introducing Asymmetric Hybrid-Resolution Attention (AHRA), a hierarchical attention design that better suits wireless signals. AHRA aggregates local signals through pooling, after which prevailing positional encodings assign each new token a single fixed position from the original grid. However, this assignment is limiting for wireless signals, which are sensitive to positional inductive bias. By constructing the response difference between different ways of computing the positional encoding, we decompose it into the main factors that drive the positional difference. On this basis, we propose Holographic Rotary Position Embedding (HoloRoPE), which scales regional keys with learnable gains based on the geometric and content factors. Extensive experiments on large-scale wireless datasets show that both AHRA and HoloRoPE improve performance over the baseline models.

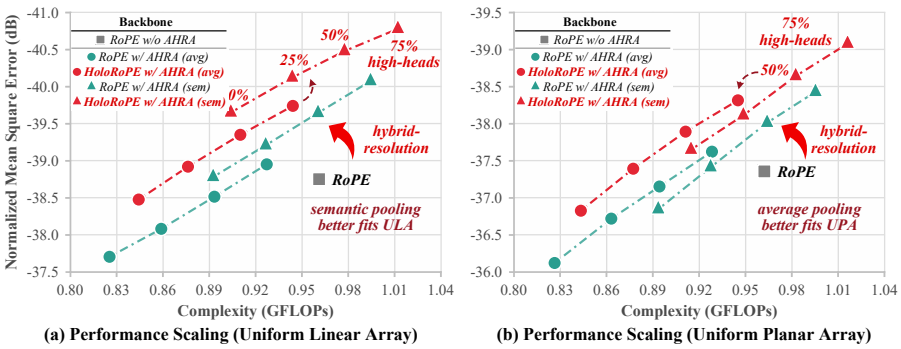
1 INTRODUCTION

Catalyzed by the success of 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), wireless foundation models (WFMs) extend the pretraining paradigm to wireless communications, learning transferable representations from large-scale channel data for downstream tasks ranging from channel prediction to localization (Alikhani et al., 2024; Liu et al., 2025; Pan et al., 2025). Their input, channel state information (CSI), has two axes that index antennas and subcarriers, sampling the propagation channel in space and in frequency, respectively. The relation between two CSI patches depends on their relative offset rather than their absolute positions. Rotary Position Embedding (RoPE) (Su et al., 2023) encodes relative positions through coordinate-dependent rotations of queries and keys (Barbero et al., 2025) and is therefore a natural choice of positional encoding for WFMs.

In wireless data, adjacent subcarriers often lie within the coherence bandwidth and share path-level structure, so aggregating neighboring features reinforces this structure while averaging out independent perturbations such as estimation noise (Tse & Viswanath, 2005; van de Beek et al., 1995). Deep fades caused by destructive multipath interference, however, are local and would be obscured if every key were aggregated. Hierarchical model architectures are therefore better suited to WFMs. To this end, we propose Asymmetric Hybrid-Resolution Attention (AHRA), in which all heads keep queries at native resolution, a subset of heads pools neighboring keys and values following spatial-reduction attention (Wang et al., 2021; Pan et al., 2022), and the remaining heads stay at native resolution. As shown in Figure 1, such hybrid-resolution configurations achieve lower channel extrapolation error than full-resolution attention at comparable complexity.

However, pooling also changes what a key represents. A regional key summarizes patches at different coordinates, yet standard RoPE rotates it at a single representative coordinate. The distinct offsets between a query and the constituent patches are thus replaced by one common offset, so the relative-position structure that makes RoPE suitable for WFMs is only partially retained for regional keys.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) Performance Scaling (Uniform Linear Array) (b) Performance Scaling (Uniform Planar Array)

Figure 1: Performance–complexity trade-offs under (a) ULA and (b) UPA configurations. RoPE and HoloRoPE are evaluated with average (avg) and semantic (sem) pooling, alongside RoPE without AHRA. Percentages denote the proportion of native-resolution heads.

Existing extensions of rotary encoding rescale (Sun et al., 2023), extend (Hua et al., 2025; Liu et al., 2026a), or adapt (Movahedi et al., 2026; Zhang et al., 2026) the rotations, but they encode each key at a single position. Since wireless data are sensitive to positional encoding, how a key that summarizes a region should be encoded remains an open question.

To address this question, we analyze the rotary response of a regional key within AHRA. By examining the difference between two positional encodings, the common-coordinate response and the constituent-coordinate reference, we identify its main factors, namely a geometric factor determined by the constituent coordinates and the rotary frequency and a content factor that couples within-region feature variation with positional phases. Building on this analysis, we propose *Holographic RoPE (HoloRoPE)*, which scales each rotated component of a regional key by a bounded, learnable gain. The gain follows the quadratic rotary-frequency shape of the geometric factor, with a strength predicted from within-region feature dispersion and a regional scale predicted from pooled content, and it preserves the encoded phases and thus the relative-position structure of RoPE. As Figure 1 shows, HoloRoPE improves upon RoPE under every head allocation and for both pooling operators. We summarize our contributions below.

- Addressing redundancy among neighboring subcarriers and localized deep fades, we propose AHRA, a hybrid-resolution attention for WFMs that aggregates neighboring keys and values in a subset of heads while keeping queries and the remaining heads at native resolution.
- Addressing the limitations of existing positional encodings after aggregation, we compare the common-coordinate response with the constituent-coordinate reference and analyze the geometric and content factors that shape the positional encoding of regional keys. Building on this analysis, we propose HoloRoPE, which applies bounded, content-conditioned, frequency-dependent gains to regional keys while preserving their encoded phases.
- Experiments on DeepMIMO (Alkhateeb, 2019) and Sionna (Hoydis et al., 2022) across multiple tasks, antenna configurations, and model sizes show that AHRA improves upon full-resolution attention and HoloRoPE outperforms corresponding rotary baselines, with its strongest variant leading all reported metrics.

2 Methodology

2.1 Preliminaries

Wireless Foundation Models (WFMs). Building on the pretraining paradigm established in language and vision, WFMs learn transferable representations from large-scale wireless data. Their input is channel state information (CSI). As with images, each input channel of CSI (e.g., its real or imaginary part) can be represented as a two-dimensional array. However, whereas image data have the same properties along both axes, CSI does not: as shown in Figure 2, its two axes correspond to antennas and subcarriers, with the antenna axis sampling the wavefront in space and the subcarrier axis sampling the channel response in frequency.

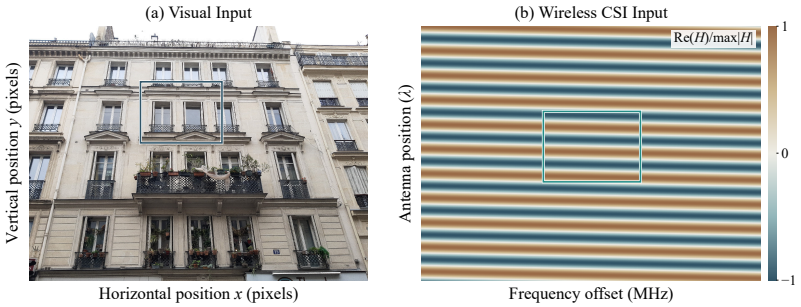


Figure 2: Comparison of visual and wireless inputs. (a) An image is indexed by horizontal and vertical pixel positions. (b) CSI is indexed by antenna position and subcarrier frequency offset, with color representing $\mathrm{Re}(H)/\max|H|$. Boxes indicate illustrative patch regions.

WFMs learn representations from unlabeled CSI through self-supervised pretraining, such as masked channel reconstruction, which, analogous to masked image modeling, recovers the coefficients of randomly masked patches (Alikhani et al., 2024; Liu et al., 2025). The pretrained backbone is then adapted to downstream tasks through task-specific heads, with the backbone frozen or fine-tuned. Our backbone and pretraining objective are detailed in Appendix A.1.

Rotary Position Embedding (RoPE). Without positional cues, unmasked self-attention is permutation equivariant (Yun et al., 2020). Originally introduced for language modeling and later extended to vision (Su et al., 2023; Heo et al., 2024), RoPE encodes relative positional information through coordinate-dependent rotations of queries and keys. Specifically, for a fixed head and coordinate axis, pairing the d projected real channels gives $\mathbf{q}_m, \mathbf{k}_n \in \mathbb{C}^{d/2}$, where d is even and m, n index native patches at coordinates p_m, p_n. For component $j \in \{1, \ldots, d/2\}$ with rotary frequency θ_j,

$$\widetilde{q}_{m,j} = q_{m,j} e^{\mathrm{i} p_m \theta_j}, \quad \widetilde{k}_{n,j} = k_{n,j} e^{\mathrm{i} p_n \theta_j}. \tag{1}$$

Here, complex notation represents paired feature channels, and θ_j is a rotary rather than a subcarrier frequency. This axis contributes to the unscaled attention score as

$$\mathcal{A}_{m,n} = \mathrm{Re}\left[\sum_{j=1}^{d/2} q_{m,j} k^*_{n,j} e^{\mathrm{i}(p_m - p_n)\theta_j}\right]. \tag{2}$$

Contributions from all coordinate blocks are summed before standard attention scaling and softmax. The positional factor depends on relative displacement $p_m - p_n$, while each rotation preserves its feature pair's magnitude. These properties provide a geometric inductive bias for WFMs, encoding relations between channel patches without rescaling their feature-pair magnitudes. Since the relation between two CSI patches depends on their relative offset along each axis rather than their absolute positions, RoPE is naturally suited to WFMs.

2.2 Asymmetric Hybrid-Resolution Attention

In wireless channels, adjacent subcarriers are typically spaced much more closely than the coherence bandwidth, the frequency range over which the channel stays nearly constant, so neighboring entries share a common path-level structure (Tse & Viswanath, 2005). The principle of coherent combining suggests that local aggregation can reinforce shared structure while reducing independent perturbations, such as estimation noise. Exploiting channel correlation is well established in OFDM channel estimation (van de Beek et al., 1995). Aggregating neighboring features into regional keys can thus improve performance by summarizing the path-level structure relevant to channel extrapolation and beam prediction. However, destructive interference among propagation paths can produce deep fades, and using regional keys in all heads may obscure these fine-grained variations.

Motivated by these properties, we propose Asymmetric Hybrid-Resolution Attention (AHRA), which aggregates neighboring features in selected heads, such as pairs of adjacent subcarriers in our main setting, and keeps the remaining heads at native resolution. Consistent with this motivation, we pool only along the subcarrier axis, using pairs of adjacent subcarriers in our main setting.

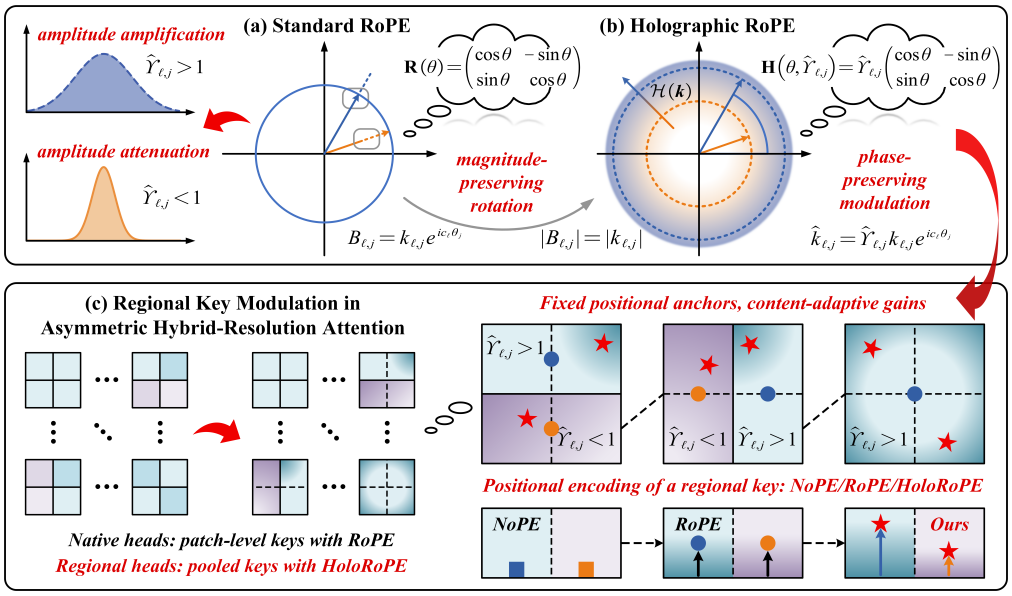


Figure 3: Implementation of HoloRoPE. (a) Standard RoPE rotates regional key components at a representative coordinate without changing their magnitudes. (b) HoloRoPE modulates these magnitudes with learned positive gains, preserving encoded phases. (c) AHRA retains native queries and combines patch-level keys with RoPE and pooled keys with HoloRoPE across heads.

Specifically, we partition the H attention heads into disjoint native and regional groups, $\mathcal{H}_N$ and $\mathcal{H}_R$. Native heads form keys and values from individual patch features, as in standard attention. Regional heads form them from pooled features, following spatial-reduction attention (Wang et al., 2021; Pan et al., 2022). Assigning different key–value resolutions to different heads resembles the multi-rate head design of Ren et al. (2022). Both groups keep queries at native resolution and attend globally over their respective key sets, so each patch interacts with individual patches through native heads and with regional summaries through regional heads.

Let $\mathbf{x}_m \in \mathbb{R}^{d_{\text{model}}}$ denote the feature of native patch m. Each region Ω_ℓ contains neighboring patches whose features form the rows of $\mathbf{X}_{\Omega_\ell}$. Regional heads apply pooling and head-specific projections:

$$\mathbf{z}_\ell = \mathcal{N}_P\left(\mathcal{P}(\mathbf{X}_{\Omega_\ell})\right), \quad \mathbf{k}_\ell^{(h)} = \mathbf{W}_K^{(h)} \mathbf{z}_\ell, \quad \mathbf{v}_\ell^{(h)} = \mathbf{W}_V^{(h)} \mathbf{z}_\ell, \quad h \in \mathcal{H}_R. \tag{3}$$

Here, $\mathcal{P}$ is the pooling operator, such as average or learned semantic pooling, and $\mathcal{N}_P$ denotes normalization after pooling. Native heads instead project keys and values directly from $\mathbf{x}_n$. Both groups form queries as $\mathbf{q}_m^{(h)} = \mathbf{W}_Q^{(h)} \mathbf{x}_m$.

After projection, we apply standard RoPE to queries and keys. Queries and native keys are encoded at their patch coordinates, whereas each regional key is encoded at a representative coordinate c_ℓ along each axis. Let $\mathcal{I}_h$ index native patches for $h \in \mathcal{H}_N$ and regions for $h \in \mathcal{H}_R$. With the encoded query and key vectors containing all coordinate blocks, each head computes

$$s_{m,u}^{(h)} = \frac{1}{\sqrt{d_h}} \operatorname{Re}\left[\sum_{j=1}^{d_h/2} \tilde{q}_{m,j}^{(h)} \left(\tilde{k}_{u,j}^{(h)}\right)^*\right], \quad \mathbf{o}_m^{(h)} = \sum_{u \in \mathcal{I}_h} \frac{\exp(s_{m,u}^{(h)})}{\sum_{u' \in \mathcal{I}_h} \exp(s_{m,u'}^{(h)})} \mathbf{v}_u^{(h)}, \tag{4}$$

where d_h is the real query and key dimension of head h, and j indexes complex pairs across coordinate blocks. Values remain unrotated, and head outputs are concatenated and linearly projected to the model width. Head allocation controls the capacity assigned to each granularity and attention cost. An architectural comparison and implementation details are provided in Appendix A.2.

2.3 Holographic Rotary Position Embedding

Standard RoPE treats a pooled regional key in the same way as an individual patch: it rotates the key at a single representative coordinate c_ℓ, effectively representing the region as a point. However,

RoPE is suited to WFMs since the relation between CSI patches depends on their relative offsets along each axis (Section 2.1). Once a region is collapsed to a single coordinate, the distinct offsets $p_m - p_r$ of its constituents at coordinates p_r relative to the query are all replaced by the common offset $p_m - c_\ell$. Since WFMs are sensitive to this relative-position inductive bias, it is necessary to examine how the rotary encoding of a regional key should reflect the spatial structure within the region. Holding the constituent feature contributions fixed, two response modes can be distinguished: the common-coordinate response, obtained by assigning a shared representative coordinate, and the constituent-coordinate reference, obtained by retaining the individual coordinates.

Common-Coordinate Response. For a fixed regional head and coordinate axis, standard RoPE encodes the pooled key at the representative coordinate c_ℓ:

$$B_{\ell,j} = k_{\ell,j} e^{\mathrm{i} c_\ell \theta_j}. \tag{5}$$

This common-coordinate rotation preserves the magnitude of each aggregated key component and positions the region as a whole at c_ℓ (Figure 3a). Its interaction with a query at p_m therefore carries a positional factor determined by the common offset $p_m - c_\ell$.

Constituent-Coordinate Reference. The regional representation combines contributions from patches at coordinates p_r, each with displacement $p_m - p_r$ relative to the query. To examine their positional effects, we express the pooled key as $\mathbf{k}_\ell = \sum_{r\in\Omega_\ell} \boldsymbol{\kappa}_r + \boldsymbol{\kappa}_b$. Here, $\boldsymbol{\kappa}_r$ incorporates the patch's pooling transformation, normalization coefficients evaluated on the pooled feature, and key projection, while $\boldsymbol{\kappa}_b$ collects the bias contribution. Appendix A.3 provides the construction. Keeping these contributions fixed, we retain their individual coordinates in the rotary encoding:

$$T_{\ell,j} = \kappa_{b,j} e^{\mathrm{i} c_\ell \theta_j} + \sum_{r\in\Omega_\ell} \kappa_{r,j} e^{\mathrm{i} p_r \theta_j}. \tag{6}$$

The bias remains encoded at c_ℓ, as in $B_{\ell,j}$. This reference retains the constituent coordinates while aggregating the same feature contributions into a single regional key.

Response Difference. We construct $T_{\ell,j} - B_{\ell,j}$ to analyze which terms dominate the positional difference between the two responses. Relative to the common rotation in $B_{\ell,j}$, each patch contribution in $T_{\ell,j}$ receives an additional phase offset $(p_r - c_\ell)\theta_j$. These offsets act on individual contributions before summation, so their effect depends on how feature content is distributed across the region (Figure 3c, right). Subtracting the common-coordinate response isolates the resulting change:

$$T_{\ell,j} - B_{\ell,j} = e^{\mathrm{i} c_\ell \theta_j} \sum_{r\in\Omega_\ell} \kappa_{r,j} \left[e^{\mathrm{i}(p_r - c_\ell)\theta_j} - 1 \right]. \tag{7}$$

The coordinate offsets and rotary frequency determine the phase factors, while the local feature contributions determine how these factors combine. Regional geometry alone therefore does not generally determine the response difference. To distinguish these effects, we decompose the difference into a common geometric factor acting on the summed contributions and a residual capturing the interaction between contribution variations and positional phases.

Specifically, let $K_\ell = |\Omega_\ell|$ and define the coordinate baseline $\Phi_\ell(\theta)$ for $\theta \in \mathbb{R}$ and the contribution-weighted residual $C_{\ell,j}$ at each rotary frequency θ_j:

$$\Phi_\ell(\theta) = \frac{1}{K_\ell} \sum_{r\in\Omega_\ell} e^{\mathrm{i}(p_r - c_\ell)\theta}, \quad C_{\ell,j} = \sum_{r\in\Omega_\ell} \kappa_{r,j} \left[e^{\mathrm{i}(p_r - c_\ell)\theta_j} - \Phi_\ell(\theta_j) \right]. \tag{8}$$

Here, Φ_ℓ is the mean positional phase factor over the region. Its uniform weights define an analytical baseline, while the actual pooling transformations remain incorporated in $\kappa_{r,j}$. Adding and subtracting $\Phi_\ell(\theta_j)$ within each summand yields the decomposition

$$T_{\ell,j} - B_{\ell,j} = e^{\mathrm{i} c_\ell \theta_j} \left[\big(\Phi_\ell(\theta_j) - 1\big)\big(k_{\ell,j} - \kappa_{b,j}\big) + C_{\ell,j} \right], \tag{9}$$

where the first term applies the common coordinate factor $\Phi_\ell(\theta_j) - 1$ to the summed patch contributions, while $C_{\ell,j}$ captures the interaction between variations in those contributions and their positional phases. In particular, $C_{\ell,j}$ vanishes when the patch contributions are identical. Appendix A.3 provides the derivation and a bound on the residual.

Holographic Modulation. This decomposition identifies two aspects of regional rotary encoding: the frequency dependence of the coordinate factor $\Phi_\ell(\theta)$ and the content–position coupling captured by $C_{\ell,j}$. These observations motivate ***HoloRoPE***, a content-adaptive approach to regional rotary encoding. The analysis determines which quantities the modulation should depend on and how it should vary with rotary frequency, whereas the gain values and their direction are learned from the task. We examine these aspects in turn to develop its modulation structure.

We first consider the magnitude of the coordinate baseline. As the rotary frequency grows from zero, the phase factors of constituents at different coordinates spread apart, and $|\Phi_\ell|$ falls below one. For a fixed finite region, the coordinate baseline satisfies

$$-\log|\Phi_\ell(\theta)| = \tfrac{1}{2}\sigma_\ell^2\theta^2 + O(\theta^4), \quad \theta \to 0, \tag{10}$$

where σ_ℓ^2 is the coordinate variance under the uniform weights defining Φ_ℓ. The proof is given in Appendix A.4. The leading term separates a spatial scale from a quadratic dependence on rotary frequency. This local behavior motivates using θ_j^2 as a frequency basis for amplitude modulation, rather than approximating the full response difference. We adopt amplitude modulation to adjust regional key magnitudes while retaining the representative-coordinate phases. Retaining these phases keeps the positional factor of each regional key determined by $p_m - c_\ell$ alone, so the relative-position structure that makes RoPE suitable for WFMs (Section 2.1) is preserved.

The frequency basis alone, however, does not account for the content–position coupling captured by $C_{\ell,j}$. To incorporate this dependence, we condition the modulation on regional features. The pooled representation $\mathbf{z}_\ell$ summarizes regional content but need not retain the variation among its constituent features. We therefore supplement it with a local feature-dispersion statistic s_ℓ. The feature-deviation component of $C_{\ell,j}$ is bounded in proportion to both $\|s_\ell\|$ and $\sqrt{1-|\Phi_\ell(\theta_j)|^2}$ (Appendix A.5), where the geometric factor has leading behavior $\sigma_\ell|\theta_j|$ at low frequency. The content-dependent part of the response difference thus also grows with rotary frequency, at a scale set by the dispersion, which motivates letting s_ℓ determine the strength of the θ_j^2 term.

Together, $\mathbf{z}_\ell$ and s_ℓ provide the conditioning information for learning the frequency-dependent modulation strength and a common regional scale:

$$\nu_\ell = f_\nu(s_\ell), \quad a_\ell = f_a(\mathbf{z}_\ell, s_\ell), \tag{11}$$

where $\nu_\ell > 0$ controls the frequency-dependent factor and a_ℓ is a bounded positive regional gate. The statistic and mappings are specified in Appendix A.5. Here, ν_ℓ is a learned content-dependent scale, rather than an estimate of the coordinate variance σ_ℓ^2.

Combining the quadratic basis with the learned scale, we define $u_{\ell,j} = \frac{1}{2}\lambda\nu_\ell\theta_j^2$, where $\lambda \in \mathbb{R}$ is learned for each layer and regional head. The coordinate baseline alone implies attenuation, since $|\Phi_\ell| \le 1$, whereas content–position coupling can make $|T_{\ell,j}|$ exceed $|B_{\ell,j}|$ (Appendix A.6); we therefore leave the sign of λ to be learned. We then use the positive bounded spectral gain $G_{\ell,j} = 1 + \tanh(u_{\ell,j})$, allowing attenuation or amplification according to the sign of λ (Figure 3a, left). The regional gate a_ℓ controls the common scale across rotary components, yielding

$$\widehat{\Upsilon}_{\ell,j} = a_\ell G_{\ell,j}, \quad \widehat{k}_{\ell,j} = \widehat{\Upsilon}_{\ell,j} B_{\ell,j} = \widehat{\Upsilon}_{\ell,j} k_{\ell,j} e^{\mathrm{i}c_\ell\theta_j}. \tag{12}$$

The positive gain modulates component magnitudes without changing their encoded phases (Figure 3b) and is learned through the task objective, not by matching the constituent-coordinate response. HoloRoPE operates on regional keys within AHRA: pooling, assigned coordinates, and head allocation remain unchanged, queries and native keys retain RoPE, and values remain unrotated.

3 Experiments

3.1 Experimental Setups

Models and Training Settings. We use an encoder-based WFM (Alikhani et al., 2024; Liu et al., 2026b) with 3M, 25M, 85M, 302M, or 708M parameters, evaluating complex-valued variants with matched head counts (EH) or cache sizes (EC). Training uses AdamW with weight decay 0.05 and peak learning rate 5×10^{-4}. We report NMSE (dB) for channel extrapolation (CExt) and estimation (CEst), and accuracy (%) for LoS/NLoS classification (L/N) and beam prediction (BPred). Appendix B.1 provides detailed descriptions of the downstream tasks and evaluation settings.

Table 1: Benchmark results on DeepMIMO and Sionna with ULA and UPA. In models with AHRA, keys and values are pooled in a subset of attention heads. Bold and underlined values indicate the best and second-best results for each metric within each dataset.

Method	Uniform Linear Array					Uniform Planar Array				
	CExt† NMSE↓	CEst† NMSE↓	L/N¶ Acc (2)↑	BPred¶ Acc (64)↑	BPred¶ Acc (128)↑	CExt† NMSE↓	CEst† NMSE↓	L/N¶ Acc (2)↑	BPred¶ Acc (64)↑	BPred¶ Acc (128)↑
DeepMIMO (Deterministic Ray-Tracing) – *w/o AHRA*										
RoPE	-38.86	-17.79	95.37	65.14	64.73	-37.32	-18.57	95.28	60.57	60.07
DeepMIMO (Deterministic Ray-Tracing) – *w/ AHRA*										
Learned	-16.53	-15.13	94.32	61.54	60.67	-25.93	-16.50	96.20	55.32	55.06
Sinusoidal	-17.23	-15.09	95.89	62.85	62.45	-28.42	-16.31	96.26	51.79	57.45
NoPE	-17.07	-14.90	95.82	63.47	63.33	-27.20	-16.37	96.30	60.49	56.86
ALiBi	-31.36	-16.05	94.60	67.28	61.63	-28.81	-18.11	95.90	60.31	60.92
FoPE	-38.77	-19.43	95.77	64.93	60.41	-37.27	-19.75	95.85	64.67	63.47
RoPE	-40.12	-18.99	95.84	66.21	66.90	-38.37	-19.74	95.84	61.15	60.38
RoPE++$_{EH}$	-38.49	-18.02	93.14	66.92	63.33	-35.85	-19.28	96.71	59.64	60.48
RoPE++$_{EC}$	-40.76	-19.25	93.08	67.87	64.65	-38.75	<u>-20.34</u>	<u>98.66</u>	61.85	60.46
HoloRoPE	<u>-41.14</u>	<u>-19.59</u>	<u>97.19</u>	<u>73.89</u>	70.31	<u>-39.19</u>	-20.24	98.63	<u>67.35</u>	<u>65.30</u>
HoloRoPE++$_{EH}$	-39.42	-18.86	95.99	72.02	**71.40**	-36.78	-19.71	97.37	63.73	65.05
HoloRoPE++$_{EC}$	**-41.79**	**-19.81**	**97.38**	**77.60**	<u>70.96</u>	**-39.54**	**-20.76**	**99.21**	**68.19**	**66.18**
Sionna (Stochastic 3GPP-Compliant) – *w/o AHRA*										
RoPE	-14.94	-16.04	95.61	67.13	65.34	-15.55	-13.21	95.33	59.83	58.71
Sionna (Stochastic 3GPP-Compliant) – *w/ AHRA*										
Learned	-14.54	-13.88	94.34	63.28	61.15	-15.01	-12.81	96.20	61.35	60.08
Sinusoidal	-14.55	-13.87	94.35	62.18	61.93	-15.43	-12.78	95.93	61.87	59.95
NoPE	-15.19	-12.65	95.67	63.61	65.41	-15.18	-11.68	95.85	62.93	58.73
ALiBi	-15.03	-14.08	95.67	63.28	66.51	-15.77	-12.82	94.73	62.97	62.04
FoPE	-16.12	-17.37	95.70	69.83	66.55	-15.90	-14.19	96.01	60.21	57.10
RoPE	-15.61	-17.13	95.83	67.55	65.57	-16.18	-13.97	95.73	60.58	59.18
RoPE++$_{EH}$	-14.56	-16.46	94.61	62.49	60.64	-15.24	-13.53	95.93	60.11	59.47
RoPE++$_{EC}$	-16.18	-17.06	94.49	67.25	61.10	-16.61	-14.13	94.32	62.98	62.31
HoloRoPE	<u>-16.65</u>	<u>-17.73</u>	<u>98.65</u>	<u>77.08</u>	**74.29**	<u>-16.96</u>	<u>-14.42</u>	97.30	<u>65.67</u>	<u>65.45</u>
HoloRoPE++$_{EH}$	-15.64	-17.18	97.30	71.21	70.90	-15.98	-14.02	<u>98.63</u>	64.02	63.29
HoloRoPE++$_{EC}$	**-17.15**	**-17.99**	**98.68**	**77.76**	<u>72.16</u>	**-17.32**	**-14.63**	**98.67**	**67.82**	**66.31**

Datasets and Baselines. We pretrain on nearly 1B channel tokens generated using DeepMIMO (Alkhateeb, 2019) for site-specific ray tracing and Sionna (Hoydis et al., 2022) for stochastic 3GPP-compliant channel simulation. The data cover urban microcell and rural macrocell environments, uniform linear arrays (ULAs) and uniform planar arrays (UPAs), and varying channel dimensions. We compare HoloRoPE and its complex-valued variants (HoloRoPE++) against learned embeddings (Alikhani et al., 2024; Zhou et al., 2025), sinusoidal encodings (Zhou et al., 2024; Liu et al., 2024), ALiBi (Press et al., 2022), RoPE (Su et al., 2023), RoPE++ (Liu et al., 2026a), FoPE (Hua et al., 2025), and NoPE (Kazemnejad et al., 2023).

Performance Analysis. Table 1 shows that HoloRoPE and its complex-valued variants outperform corresponding rotary baselines across tasks, propagation environments, and antenna configurations, supporting regional response modulation in both real- and complex-valued models. The variants offer different performance–resource trade-offs: HoloRoPE++$_{\mathrm{EH}}$ improves upon its matched-head rotary baseline while requiring less KV storage and fewer QKV projection parameters than the real-valued counterpart. HoloRoPE++$_{\mathrm{EC}}$ improves predictive performance under a matched KV storage budget, while base HoloRoPE improves upon standard RoPE without a complex-valued extension.

Pooling Resolution. Both methods perform best with 1×2 antenna–subcarrier pooling (Figure 4). Coarser pooling generally degrades performance, while HoloRoPE retains an advantage over RoPE across the reduced-resolution configurations shown. These results favor moderate regional aggregation over progressively coarser pooling.

3.2 Ablation Studies

We conduct targeted ablation studies to examine the design choices underlying HoloRoPE. Specifically, we isolate the individual and joint contributions of the spectral factor and regional gate, and assess the effects of pooling operators and hybrid-resolution configurations for HoloRoPE and its complex-valued variants. Unless otherwise specified, all non-ablated settings follow Section 3.1.

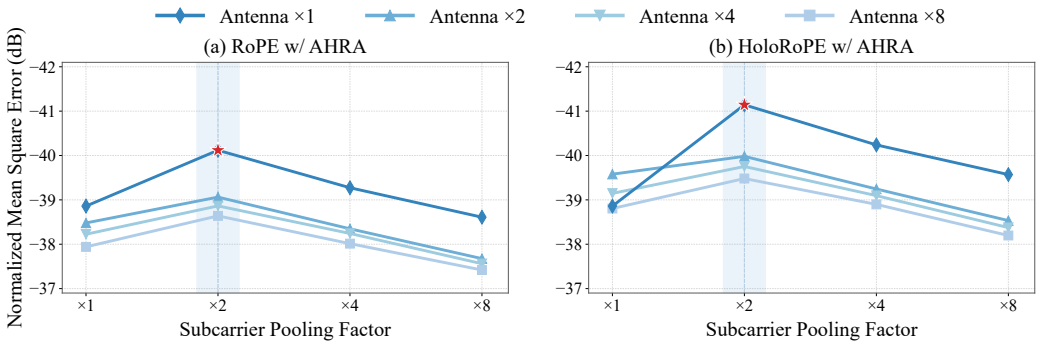


Figure 4: Effect of antenna and subcarrier pooling factors on channel extrapolation: (a) RoPE with AHRA and (b) HoloRoPE with AHRA. Red stars mark the best configuration for each method.

Table 2: Ablation of the spectral factor $G_{\ell,j}$ and regional gate a_ℓ on DeepMIMO under ULA and UPA. These components modulate gains across rotary frequencies and the common amplitude scale within each region, respectively. Bold and underlined values indicate the best and second-best results for each metric within each rotary family.

Method	Components		Uniform Linear Array				Uniform Planar Array			
	Spectral Factor	Regional Gate	CExt* NMSE↓	CEst* NMSE↓	L/N¶ Acc (2)↑	BPred¶ Acc (128)↑	CExt* NMSE↓	CEst* NMSE↓	L/N¶ Acc (2)↑	BPred¶ Acc (128)↑
Real-Valued RoPE (Vanilla)										
RoPE	✘	✘	-40.12	-18.99	95.84	66.90	-38.37	-19.74	95.84	60.38
RoPE-Spec	✘	✔	-40.96	-19.29	96.76	70.08	-38.99	-20.10	97.49	62.32
RoPE-Gate	✔	✘	-40.63	-19.47	96.37	68.24	-38.79	-19.68	97.63	63.11
HoloRoPE	✔	✔	**-41.14**	**-19.59**	**97.19**	**70.31**	**-39.19**	**-20.24**	**98.63**	**65.30**
Complex-Valued RoPE++$_{EH}$ (Equal Head Count)										
RoPE++$_{EH}$	✘	✘	-38.49	-18.02	93.14	63.33	-35.85	-19.28	96.71	60.48
RoPE++$_{EH}$-Spec	✘	✔	-39.25	-18.57	95.22	64.95	-36.55	-19.50	96.85	62.61
RoPE++$_{EH}$-Gate	✔	✘	-38.96	-18.64	94.86	70.24	-36.42	-19.41	97.02	61.20
HoloRoPE++$_{EH}$	✔	✔	**-39.42**	**-18.86**	**95.99**	**71.40**	**-36.78**	**-19.71**	**97.37**	**65.05**
Complex-Valued RoPE++$_{EC}$ (Equal Cache Size)										
RoPE++$_{EC}$	✘	✘	-40.76	-19.25	93.08	64.65	-38.75	-20.34	98.66	60.46
RoPE++$_{EC}$-Spec	✘	✔	-41.56	-19.68	96.78	70.49	-39.27	-20.53	98.82	63.05
RoPE++$_{EC}$-Gate	✔	✘	-41.36	-19.69	96.24	67.89	-39.14	-20.45	98.97	64.37
HoloRoPE++$_{EC}$	✔	✔	**-41.79**	**-19.81**	**97.38**	**70.96**	**-39.54**	**-20.76**	**99.21**	**66.18**

Complementary Roles of Spectral and Regional Modulation. Table 2 examines the individual and joint contributions of the spectral factor $G_{\ell,j}$ and regional gate a_ℓ. The spectral factor uses local feature dispersion to adjust relative gains across rotary frequencies, whereas the regional gate combines dispersion with pooled content to control their common amplitude scale. Both components generally improve upon the corresponding rotary baselines when used independently, although their relative benefits vary across tasks and antenna configurations. Combining the two components yields further improvements over the single-component variants in Table 2. These results support combining frequency-dependent gain shaping with regional amplitude control to enable content-conditioned rotary response modulation in HoloRoPE.

Effectiveness across Pooling Operators. To examine whether HoloRoPE's benefits depend on the aggregation rule, we evaluate average pooling with uniform window weights and parametric semantic pooling with a learned mapping. As shown in Table 3, HoloRoPE and its complex-valued variants outperform their corresponding rotary baselines under both operators and antenna configurations. Notably, these gains persist with parametric semantic pooling, suggesting that learned aggregation does not eliminate the benefit of explicit regional response modulation. The two operations serve complementary roles: pooling constructs regional representations, whereas HoloRoPE uses local feature dispersion and pooled content to modulate their rotary responses through bounded amplitude gains. Additional ablation results are provided in Appendix B.2.

Table 3: Evaluation of HoloRoPE and its complex-valued variants across pooling operators on DeepMIMO with ULA and UPA. Each HoloRoPE variant is compared with its corresponding rotary baseline using the same pooling operator. Response modulation denotes the joint use of the spectral factor $G_{\ell,j}$ and regional gate a_ℓ. Bold values indicate the better result in each matched pair.

Method	Configuration	Uniform Linear Array				Uniform Planar Array			
	Response Modulation $a_\ell G_{\ell,j}$	CExt* NMSE↓	CEst* NMSE↓	BPred¶ Acc (64)↑	BPred¶ Acc (128)↑	CExt* NMSE↓	CEst* NMSE↓	BPred¶ Acc (64)↑	BPred¶ Acc (128)↑
Average Pooling									
RoPE	✘	-38.94	-17.85	65.41	65.37	-37.56	-19.02	60.34	60.29
HoloRoPE	✔	**-40.07**	**-18.47**	**72.13**	**69.28**	**-38.54**	**-19.36**	**66.28**	**64.51**
RoPE++$_{\mathrm{EH}}$	✘	-37.47	-16.99	65.69	62.17	-35.02	-18.37	61.09	60.36
HoloRoPE++$_{\mathrm{EH}}$	✔	**-38.76**	**-17.82**	**71.15**	**70.43**	**-35.89**	**-18.84**	**62.55**	**61.37**
RoPE++$_{\mathrm{EC}}$	✘	-39.55	-18.31	66.51	63.93	-38.02	-19.42	60.03	60.29
HoloRoPE++$_{\mathrm{EC}}$	✔	**-40.73**	**-18.76**	**76.07**	**69.52**	**-38.88**	**-19.93**	**67.21**	**65.09**
Parametric Semantic Pooling									
RoPE	✘	-40.12	-18.99	66.21	66.90	-38.37	-19.74	61.15	60.38
HoloRoPE	✔	**-41.14**	**-19.59**	**73.89**	**70.31**	**-39.19**	**-20.24**	**67.35**	**65.30**
RoPE++$_{\mathrm{EH}}$	✘	-38.49	-18.02	66.92	63.33	-35.85	-19.28	59.64	60.48
HoloRoPE++$_{\mathrm{EH}}$	✔	**-39.42**	**-18.86**	**72.02**	**71.40**	**-36.78**	**-19.71**	**63.73**	**65.05**
RoPE++$_{\mathrm{EC}}$	✘	-40.76	-19.25	67.87	64.65	-38.75	-20.34	61.85	60.46
HoloRoPE++$_{\mathrm{EC}}$	✔	**-41.79**	**-19.81**	**77.60**	**70.96**	**-39.54**	**-20.76**	**68.19**	**66.18**

4 Related Works

Wireless Foundation Models. Building on advances in language (Radford et al., 2018; Devlin et al., 2019) and vision (Dosovitskiy et al., 2020; Radford et al., 2021), wireless foundation models (WFMs) seek to learn transferable representations from large-scale wireless data. Existing studies explore self-supervised wireless representation learning for communication and sensing (Alikhani et al., 2024), wireless localization (Pan et al., 2025), and hybird modeling for channel prediction (Liu et al., 2025). Scaling studies examine model capacity, channel heterogeneity, and input granularity (Liu et al., 2026b). Alongside these advances, the multiscale structure of wireless channels motivates hierarchical representations that combine fine-grained features with regional summaries obtained through regional aggregation (Wang et al., 2021; Ren et al., 2022). Building on these designs, we study hybrid attention for wireless modeling and positional encoding for regional representations.

Positional Embeddings. Rotary Position Embedding (RoPE) encodes relative positional relationships through coordinate-dependent rotations of queries and keys (Su et al., 2023), with extensions to two-dimensional vision representations (Heo et al., 2024). Subsequent work extends rotary encoding through position-dependent scaling for length extrapolation in XPos (Sun et al., 2023), fourier-based representations in FoPE (Hua et al., 2025), and both real and imaginary components of rotary interactions in RoPE++ (Liu et al., 2026a). Content-adaptive approaches further condition rotations on input features, as in selective RoPE (Movahedi et al., 2026), or adapt rotary frequencies using channel statistics, as in adaptive RoPE (Zhang et al., 2026). Our work instead focuses on the rotary response mismatch arising when features aggregated from multiple positions are encoded at a single regional coordinate. HoloRoPE addresses this mismatch with content-conditioned amplitude modulation while preserving the phase of each encoded component.

5 Conclusion

We study hierarchical wireless foundation models and their positional encoding. Since adjacent subcarriers typically lie within the coherence bandwidth, we propose AHRA, which pools neighboring keys and values in a subset of attention heads. After aggregation, standard RoPE has limitations for wireless data that are sensitive to positional inductive bias. From the difference between the common-coordinate response and the constituent-coordinate reference, we identify geometric and content factors and build HoloRoPE on them, applying learnable gains conditioned on regional content and local feature dispersion to regional key components. Large-scale experiments on DeepMIMO and Sionna show that AHRA consistently improves upon full-resolution attention and that HoloRoPE consistently improves upon the corresponding rotary baselines. Extending the analysis to coordinate choice and to additional axes such as time is left for future work.

AI USE STATEMENT

We used AI tools to assist with code writing and with language polishing of the manuscript. We analyzed and discussed experimental results with AI tools during the research, but AI tools did not directly contribute to the proposal of the paper's key ideas. We reviewed all AI-assisted text and take full responsibility for the final content of this paper, including its text and claims.

REFERENCES

Sadjad Alikhani, Gouranga Charan, and Ahmed Alkhateeb. Large wireless model (LWM): A foundation model for wireless channels. *arXiv preprint arXiv:2411.08872*, 2024.

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

Federico Barbero, Alex Vitvitskyi, Christos Perivolaropoulos, Razvan Pascanu, and Petar Veličković. Round and round we go! what makes rotary positional encodings useful? In *International Conference on Learning Representations*, 2025.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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)*, pp. 4171–4186, 2019.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16×16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020.

Byeongho Heo, Song Park, Dongyoon Han, and Sangdoo Yun. Rotary position embedding for vision transformer. In *European Conference on Computer Vision*, pp. 289–305, 2024.

Jakob Hoydis, Sebastian Cammerer, Fayçal Ait Aoudia, Avinash Vem, Nikolaus Binder, Guillermo Marcus, and Alexander Keller. Sionna: An open-source library for next-generation physical layer research. *arXiv preprint arXiv:2203.11854*, 2022.

Ermo Hua, Che Jiang, Xingtai Lv, Kaiyan Zhang, Youbang Sun, Yuchen Fan, Xuekai Zhu, Biqing Qi, Ning Ding, and Bowen Zhou. Fourier position embedding: Enhancing attention's periodic extension for length generalization. In *Proceedings of the 42nd International Conference on Machine Learning*, Proceedings of Machine Learning Research, 2025.

Amirhossein Kazemnejad, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Payel Das, and Siva Reddy. The impact of positional encoding on length generalization in transformers. In *Advances in Neural Information Processing Systems*, 2023.

Boxun Liu, Xuanyu Liu, Shijian Gao, Xiang Cheng, and Liuqing Yang. LLM4CP: Adapting large language models for channel prediction. *Journal of Communications and Information Networks*, 9 (2):113–125, 2024.

Boxun Liu, Shijian Gao, Xuanyu Liu, Xiang Cheng, and Liuqing Yang. WiFo: Wireless foundation model for channel prediction. *Science China Information Sciences*, 68(6):162302, 2025.

Xiaoran Liu, Yuerong Song, Zhigeng Liu, Zengfeng Huang, Qipeng Guo, Zhaoxiang Liu, Shiguo Lian, Ziwei He, and Xipeng Qiu. Beyond real: Imaginary extension of rotary position embeddings for long-context LLMs. In *International Conference on Learning Representations*, 2026a.

Ziheng Liu, Jiayi Zhang, Haoyu Wang, Bokai Xu, Chen Zhang, Yiyang Zhu, and Enyu Shi. Scaling law for large wireless models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 24017–24025, 2026b.

Sajad Movahedi, Timur Carstensen, Arshia Afzal, Frank Hutter, Antonio Orvieto, and Volkan Cevher. Selective rotary position embedding. In *International Conference on Learning Representations*, 2026.

Guangjin Pan, Kaixuan Huang, Hui Chen, Shunqing Zhang, Christian Häger, and Henk Wymeersch. Large wireless localization model (LWLM): A foundation model for positioning in 6G networks. *arXiv preprint arXiv:2505.10134*, 2025.

Zizheng Pan, Jianfei Cai, and Bohan Zhuang. Fast vision transformers with HiLo attention. In *Advances in neural information processing systems*, volume 35, pp. 14541–14554, 2022.

Ofir Press, Noah A. Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. In *International Conference on Learning Representations*, 2022.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. In *OpenAI blog*, 2018.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763, 2021.

Sucheng Ren, Daquan Zhou, Shengfeng He, Jiashi Feng, and Xinchao Wang. Shunted self-attention via multi-scale token aggregation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10853–10862, 2022.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *arXiv preprint arXiv:2104.09864*, 2023.

Yutao Sun, Li Dong, Barun Patra, Shuming Ma, Shaohan Huang, Alon Benhaim, Vishrav Chaudhary, Xia Song, and Furu Wei. A length-extrapolatable transformer. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pp. 14590–14604, 2023.

David Tse and Pramod Viswanath. *Fundamentals of wireless communication*, volume 1. Cambridge New York, NY, 2005.

Jan-Jaap van de Beek, Ove Edfors, Magnus Sandell, Sarah Kate Wilson, and Per Ola Börjesson. On channel estimation in OFDM systems. In *Proceedings of the IEEE 45th Vehicular Technology Conference*, volume 2, pp. 815–819, 1995.

Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *IEEE/CVF international conference on computer vision*, pp. 548–558. IEEE, 2021.

Chulhee Yun, Srinadh Bhojanapalli, Ankit Singh Rawat, Sashank J Reddi, and Sanjiv Kumar. Are transformers universal approximators of sequence-to-sequence functions? In *International Conference on Learning Representations*, 2020.

Chenyu Zhang, Xinchen Lyu, Chenshan Ren, Yanzhao Hou, Xuefei Zhang, Shuhan Liu, and Qimei Cui. Adaptive 3D-RoPE: Physics-aligned rotary positional encoding for wireless foundation models. *arXiv preprint arXiv:2605.00968*, 2026.

Binggui Zhou, Xi Yang, Shaodan Ma, Feifei Gao, and Guanghua Yang. Pay less but get more: A dual-attention-based channel estimation network for massive MIMO systems with low-density pilots. *IEEE Transactions on Wireless Communications*, 23(6):6061–6076, 2024.

Binggui Zhou, Xi Yang, Shaodan Ma, Feifei Gao, and Guanghua Yang. Low-overhead channel estimation via 3d extrapolation for tdd mmwave massive mimo systems under high-mobility scenarios. *IEEE Transactions on Wireless Communications*, 24(4):2797–2813, 2025.

A TECHNICAL APPENDICES

A.1 DETAILED DESCRIPTION OF WIRELESS FOUNDATION MODELS

A wireless foundation model (WFM) uses a shared pretrained backbone to support tasks with different observations, outputs, and objectives (Alikhani et al., 2024; Liu et al., 2025). We instantiate the backbone as an encoder-only Transformer and pretrain it by masked channel reconstruction. The resulting representations are adapted to downstream tasks, whose objectives and physical constraints remain task-specific. For example, a task with input $\mathbf{h}_{\text{task}}$ and output $\mathbf{o}_{\text{task}}$ may be written as

$$\begin{aligned} \max_{\mathbf{o}_{\text{task}}} \quad & g_{0,\text{task}}(\mathbf{h}_{\text{task}}, \mathbf{o}_{\text{task}}) \\ \text{s.t.} \quad & g_{c,\text{task}}(\mathbf{h}_{\text{task}}, \mathbf{o}_{\text{task}}) \geq 0, \quad c = 1, \ldots, n_c. \end{aligned} \tag{13}$$

This formulation describes task-specific decision making supported by the pretrained representations, rather than an additional training objective for HoloRoPE.

Each CSI input is partitioned along the antenna and subcarrier dimensions using the same patch layout for its real and imaginary components. Their samples within patch r are collected in $\boldsymbol{\xi}_r \in \mathbb{R}^{d_{\text{patch}}}$ and jointly embedded as $\mathbf{e}_r = \mathcal{E}(\boldsymbol{\xi}_r) \in \mathbb{R}^D$, where $D =$ and $\mathcal{E}$ is shared across patches. The layout determines the coordinates used by rotary encoding. When a classification (CLS) token is used, its learned embedding is prepended:

$$\mathbf{Y}^{(0)} = [\mathbf{e}_{\text{cls}}, \mathbf{e}_0, \ldots, \mathbf{e}_{N-1}]^\top \in \mathbb{R}^{(N+1)\times D}. \tag{14}$$

Here the embeddings are stacked as rows, and N counts CSI patches only. The complex notation used to analyze RoPE pairs projected real feature channels; it does not identify rotary phases with physical propagation phases or rotary frequencies with subcarrier frequencies.

A nonempty set of patches $\mathcal{M}$ is randomly masked to obtain $\widetilde{\mathbf{h}}$ while retaining the original patch layout. The backbone, including tokenization and embedding, produces $\mathbf{Y}^{\text{out}} = \mathcal{F}_\vartheta(\widetilde{\mathbf{h}})$. A reconstruction decoder predicts the masked patch contents:

$$\mathcal{L}_{\text{pre}} = \frac{1}{|\mathcal{M}|} \sum_{r\in\mathcal{M}} \left\| \mathcal{F}_{\text{dec}}(\mathbf{Y}^{\text{out}})_r - \boldsymbol{\xi}_r \right\|_2^2. \tag{15}$$

The decoder index refers to the original patch, independently of the leading CLS entry, which is not a reconstruction target. Bidirectional attention provides the available context for this objective.

After pretraining, the encoder supplies contextual features to the appropriate task-specific output module. Patch-level representations support channel reconstruction or prediction, while a sequence-level representation can support classification. The downstream objective determines which outputs are supervised and how the pretrained model is adapted. HoloRoPE changes the positional encoding of regional keys within this backbone, rather than introducing a separate pretraining target or changing the downstream task definition.

A.2 ASYMMETRIC HYBRID-RESOLUTION ATTENTION: COMPUTATION AND COMPLEXITY

Figure 5 compares three attention architectures. Vanilla multi head attention uses native-resolution keys and values in all heads, whereas spatial-reduction attention uses regional keys and values throughout. AHRA combines these granularities across head groups while retaining native-resolution queries. All three architectures attend globally over their respective key sets, differing in key/value granularity rather than attention masks.

Regional representations are shared across regional heads, while key/value projections and HoloRoPE modulation outputs are head-specific. When a CLS token is present, native heads include it in their query, key, and value sequences without applying rotary encoding to its query or key. Regional heads exclude CLS from pooling and attention, and their outputs are padded with a zero CLS entry before head concatenation. The CLS residual path remains unchanged.

A.3 POOLING CONTRIBUTIONS AND COORDINATE-ASSIGNMENT RESPONSES

Fix a regional head and a coordinate axis, and suppress the head index. A region Ω_ℓ contains K_ℓ native patches with features $\mathbf{x}_r \in \mathbb{R}^D$. All sums below are taken over $r \in \Omega$. The pooling operators

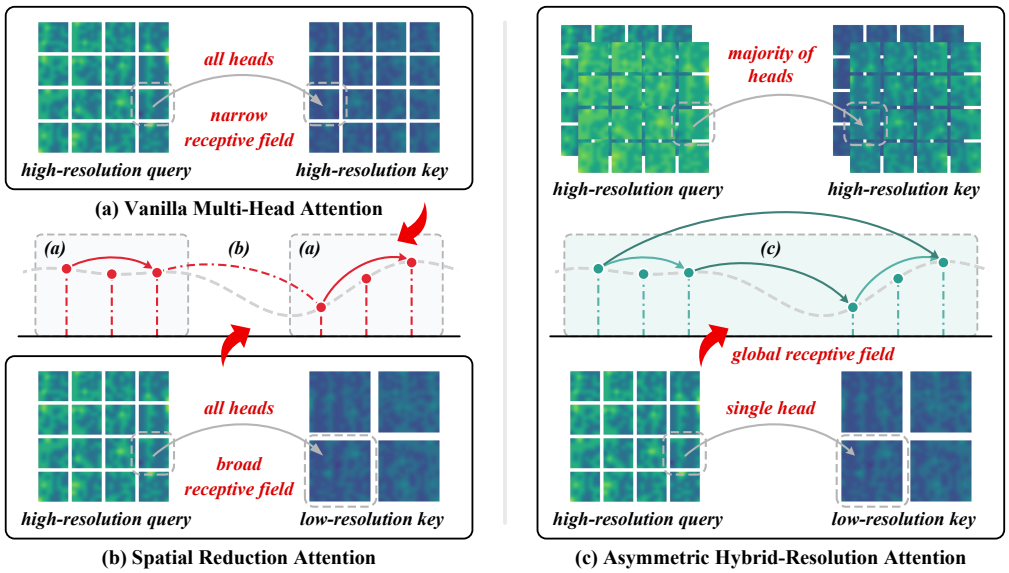


Figure 5: Comparison of attention architectures. (a) Multi-head attention uses native-resolution keys and values in all heads. (b) Spatial-reduction attention uses regional keys and values in all heads. (c) AHRA combines native and regional head groups. Queries retain native resolution in all designs.

considered here admit the pointwise representation

$$\mathbf{y}_\ell = \mathcal{P}(\mathbf{X}_{\Omega_\ell}) = \sum_{r \in \Omega_\ell} \mathbf{A}_r \mathbf{x}_r + \boldsymbol{b}. \tag{16}$$

where $\mathbf{A}_r$ is the transformation assigned to patch r by the pooling rule and $\boldsymbol{b}$ is the pooling bias. Specifically, average pooling corresponds to $\mathbf{A}_r = \mathbf{I}/K_\ell$, while $\mathbf{A}_r = \mathbf{U}\mathbf{E}_r$ for the learned semantic pooling used here, with $\boldsymbol{b} = 0$ in both cases. Here, $\mathbf{U}$ is the window projection and $\mathbf{E}_r$ inserts $\mathbf{x}_r$ at the corresponding positions in the flattened window. This construction also covers convolutional pooling, with $\mathbf{A}_r$ given by the kernel matrix at offset r and $\boldsymbol{b}$ by the bias. More generally, input-dependent $\mathbf{A}_r$ are permitted when specified by the pooling rule and evaluated on the same $\mathbf{X}_{\Omega_\ell}$.

For pooled RMS normalization, write

$$\mathcal{N}_P(\mathbf{y}_\ell) = \rho_\ell \boldsymbol{\Gamma} \mathbf{y}_\ell, \quad \rho_\ell = \left(D^{-1}\|\mathbf{y}_\ell\|_2^2 + \epsilon\right)^{-1/2}, \tag{17}$$

where $\epsilon > 0$ and $\boldsymbol{\Gamma}$ is the learned diagonal gain. With a bias-free key projection, define

$$\mathbf{F}_\ell = \rho_\ell \mathbf{W}_K \boldsymbol{\Gamma}, \quad \boldsymbol{\kappa}_r = \mathbf{F}_\ell \mathbf{A}_r \mathbf{x}_r, \quad \boldsymbol{\kappa}_b = \mathbf{F}_\ell \boldsymbol{b}. \tag{18}$$

Distributing the common matrix $\mathbf{F}_\ell$ over the pooling sum gives $\mathbf{k}_\ell = \sum_r \boldsymbol{\kappa}_r + \boldsymbol{\kappa}_b$. This is an identity at the given input, not a claim that RMS normalization or input-dependent pooling is globally linear. In particular, the normalization coefficients are not recomputed separately for each contribution.

Let p_r be the coordinate of patch r and c_ℓ the representative coordinate assigned to the region. The implementation uses the window origin for c_ℓ; the identities below allow any fixed representative coordinate. After pairing real channels, the two responses are

$$B_{\ell,j} = \left(\kappa_{b,j} + \sum_r \kappa_{r,j}\right) e^{\mathrm{i}c_\ell\theta_j}, \quad T_{\ell,j} = \kappa_{b,j} e^{\mathrm{i}c_\ell\theta_j} + \sum_r \kappa_{r,j} e^{\mathrm{i}p_r\theta_j}. \tag{19}$$

With the pooling matrices, normalization coefficients, and key projection held fixed. Thus, $T_{\ell,j}$ changes only their coordinate assignment: it is neither the output of an unpooled attention layer nor a target assumed to yield better downstream performance. Both constructions still produce one key per region. The bias is encoded at c_ℓ in both responses and cancels in their difference:

$$T_{\ell,j} - B_{\ell,j} = \sum_r \kappa_{r,j}\left(e^{\mathrm{i}p_r\theta_j} - e^{\mathrm{i}c_\ell\theta_j}\right) = e^{\mathrm{i}c_\ell\theta_j} \sum_r \kappa_{r,j}\left[e^{\mathrm{i}(p_r - c_\ell)\theta_j} - 1\right], \tag{20}$$

Write $w_{r,j} = e^{\mathrm{i}(p_r - c_\ell)\theta_j}$ and define $\Phi_\ell(\theta_j) = K_\ell^{-1} \sum_r w_{r,j}$ and $C_{\ell,j} = \sum_r \kappa_{r,j}[w_{r,j} - \Phi_\ell(\theta_j)]$. Adding and subtracting the mean phase factor yields

$$T_{\ell,j} - B_{\ell,j} = e^{\mathrm{i}c_\ell\theta_j} \left[\big(\Phi_\ell(\theta_j) - 1\big)\big(k_{\ell,j} - \kappa_{b,j}\big) + C_{\ell,j} \right]. \tag{21}$$

The uniform weights in Φ_ℓ define an analytical coordinate baseline; they do not replace the actual pooling transformations contained in $\kappa_{r,j}$.

Let $\bar{\kappa}_j = K_\ell^{-1} \sum_r \kappa_{r,j}$. Because $\sum_r [w_{r,j} - \Phi_\ell(\theta_j)] = 0$, the residual can equivalently be written as

$$C_{\ell,j} = \sum_r (\kappa_{r,j} - \bar{\kappa}_j)[w_{r,j} - \Phi_\ell(\theta_j)]. \tag{22}$$

Using $|w_{r,j}| = 1$ gives

$$\sum_r |w_{r,j} - \Phi_\ell(\theta_j)|^2 = K_\ell\big(1 - |\Phi_\ell(\theta_j)|^2\big). \tag{23}$$

Cauchy–Schwarz therefore implies

$$|C_{\ell,j}|^2 \le K_\ell\big(1 - |\Phi_\ell(\theta_j)|^2\big) \sum_r |\kappa_{r,j} - \bar{\kappa}_j|^2. \tag{24}$$

Identical patch contributions make $C_{\ell,j} = 0$, but the converse need not hold because distinct terms may cancel. The bound separates dispersion in positional phases from variation in projected contributions. It does not order the task performance of $T_{\ell,j}$ and $B_{\ell,j}$. Their effect on a fixed query's attention logits is discussed in Appendix A.2.

A.4 Coordinate Geometry and the Quadratic Frequency Basis

For a fixed region, define its coordinate mean and variance under the uniform weights used by Φ_ℓ:

$$\mu_\ell = \frac{1}{K_\ell} \sum_r p_r, \quad \sigma_\ell^2 = \frac{1}{K_\ell} \sum_r (p_r - \mu_\ell)^2. \tag{25}$$

Repeated coordinates are counted once for each corresponding patch. Centering at μ_ℓ gives

$$\Phi_\ell(\theta) = e^{\mathrm{i}(\mu_\ell - c_\ell)\theta} \phi_\ell(\theta), \quad \phi_\ell(\theta) = \frac{1}{K_\ell} \sum_r e^{\mathrm{i}(p_r - \mu_\ell)\theta}. \tag{26}$$

The representative coordinate affects the overall phase but not $|\Phi_\ell|$. Centering is only an analytical reparameterization, so keys remain encoded at the assigned coordinate c_ℓ.

The squared magnitude has the expansion

$$\begin{aligned} |\Phi_\ell(\theta)|^2 &= \frac{1}{K_\ell^2} \sum_{r,t \in \Omega_\ell} e^{\mathrm{i}(p_r - p_t)\theta} = \frac{1}{K_\ell^2} \sum_{r,t \in \Omega_\ell} \cos\left((p_r - p_t)\theta\right) \\ &= 1 - \frac{\theta^2}{2K_\ell^2} \sum_{r,t \in \Omega_\ell} (p_r - p_t)^2 + O(\theta^4) = 1 - \sigma_\ell^2 \theta^2 + O(\theta^4). \end{aligned} \tag{27}$$

The imaginary terms cancel between ordered pairs, and $\sum_{r,t} (p_r - p_t)^2 = 2K_\ell^2 \sigma_\ell^2$. Since $\Phi_\ell(0) = 1$, its magnitude is nonzero in a neighborhood of zero. Taking the logarithm establishes

$$-\log|\Phi_\ell(\theta)| = -\frac{1}{2} \log\left(1 - \sigma_\ell^2\theta^2 + O(\theta^4)\right) = \frac{1}{2}\sigma_\ell^2\theta^2 + O(\theta^4), \quad \theta \to 0, \tag{28}$$

The limit is taken for a fixed region, with the remainder depending on its coordinate moments. This expansion characterizes the coordinate baseline rather than the full complex response difference.

A non-asymptotic bound follows from $1 - \cos t \le t^2/2$:

$$0 \le 1 - |\Phi_\ell(\theta)|^2 = \frac{1}{K_\ell^2} \sum_{r,t \in \Omega_\ell} \left[1 - \cos\left((p_r - p_t)\theta\right)\right] \le \min\left\{1, \sigma_\ell^2\theta^2\right\}. \tag{29}$$

This bounds the phase-dispersion factor in Equation (24) at any real frequency.

For a regular window with equally frequent coordinates $c_\ell, \dots, c_\ell + L_\ell - 1$ along the selected axis,

$$\Phi_\ell(\theta) = \frac{1}{L_\ell} \sum_{t=0}^{L_\ell - 1} e^{\mathrm{i}t\theta} = e^{\mathrm{i}(L_\ell - 1)\theta/2} D_{L_\ell}(\theta), \quad D_{L_\ell}(\theta) = \frac{\sin(L_\ell \theta/2)}{L_\ell \sin(\theta/2)}. \tag{30}$$

The quotient is extended by continuity where its denominator vanishes. Here L_ℓ counts distinct coordinates, not all patches in the window. Its coordinate variance is $(L_\ell^2 - 1)/12$. For $L_\ell \geq 2$, D_{L_ℓ} is positive on $|\theta| < 2\pi/L_\ell$ and zero at the endpoints. Consequently, the relevant local scale is the within-window phase spread $|\theta|(L_\ell - 1)$, not the numerical frequency alone.

These results motivate a quadratic frequency basis for learned magnitude modulation. They do not prescribe a reciprocal gain $1/|\Phi_\ell|$, which is undefined at zeros of Φ_ℓ, or identify either response as optimal. Even in the special case $C_{\ell,j} = 0$ and $\kappa_{b,j} = 0$, the identity $T_{\ell,j} = \Phi_\ell(\theta_j) B_{\ell,j}$ does not imply phase agreement between the two responses. HoloRoPE retains the representative-coordinate phase and learns its magnitude modulation through the task objective.

A.5 Feature Dispersion and Properties of the Bounded Modulator

The residual in Equation (22) depends on projected contributions, whereas the modulation is predicted from features available before key projection. We relate these quantities without assuming that a pooled feature uniquely specifies its constituents.

For $K_\ell > 1$, let $\bar{\mathbf{x}}_\ell = K_\ell^{-1} \sum_r \mathbf{x}_r$ and define the channelwise sample standard deviation

$$[\mathbf{s}_\ell]_t = \left[\frac{1}{K_\ell - 1} \sum_r (x_{r,t} - \bar{x}_{\ell,t})^2 \right]^{1/2}, \quad t = 1, \dots, D. \tag{31}$$

The local dispersion statistic denoted by s_ℓ in the main text is this vector $\mathbf{s}_\ell$, rather than the coordinate variance σ_ℓ^2. It satisfies

$$\sum_r \|\mathbf{x}_r - \bar{\mathbf{x}}_\ell\|_2^2 = (K_\ell - 1)\|\mathbf{s}_\ell\|_2^2. \tag{32}$$

The sample-standard-deviation formula assumes at least two patches; a singleton window requires an explicit implementation convention rather than division by zero.

Let $\bar{\mathbf{A}} = K_\ell^{-1} \sum_r \mathbf{A}_r$ and retain the matrices evaluated at the fixed input of Appendix A.3. Centering both contributions and phase factors gives $C_{\ell,j} = C_{\ell,j}^{\mathrm{feat}} + C_{\ell,j}^{\mathrm{pool}}$, satisfying

$$C_{\ell,j}^{\mathrm{feat}} = \sum_r [\mathbf{F}_\ell \mathbf{A}_r (\mathbf{x}_r - \bar{\mathbf{x}}_\ell)]_j [w_{r,j} - \Phi_\ell(\theta_j)], \tag{33}$$

and

$$C_{\ell,j}^{\mathrm{pool}} = \sum_r [\mathbf{F}_\ell (\mathbf{A}_r - \bar{\mathbf{A}}) \bar{\mathbf{x}}_\ell]_j [w_{r,j} - \Phi_\ell(\theta_j)]. \tag{34}$$

The subscript j extracts the corresponding complex pair from the projected real vector. To verify the identity, expand $\mathbf{x}_r = (\mathbf{x}_r - \bar{\mathbf{x}}_\ell) + \bar{\mathbf{x}}_\ell$ in $C_{\ell,j}$ and subtract the common term $\mathbf{F}_\ell \bar{\mathbf{A}} \bar{\mathbf{x}}_\ell$, whose product with the centered phase factors sums to zero.

Using spectral matrix norms and $M_A = \max_r \|\mathbf{A}_r\|_2$, Cauchy–Schwarz yields

$$\begin{aligned} |C_{\ell,j}^{\mathrm{feat}}| &\leq \|\mathbf{F}_\ell\|_2 \sqrt{K_\ell (1 - |\Phi_\ell(\theta_j)|^2)} \left[\sum_r \|\mathbf{A}_r (\mathbf{x}_r - \bar{\mathbf{x}}_\ell)\|_2^2 \right]^{1/2} \\ &\leq \|\mathbf{F}_\ell\|_2 M_A \sqrt{K_\ell (K_\ell - 1)(1 - |\Phi_\ell(\theta_j)|^2)}\, \|\mathbf{s}_\ell\|_2, \end{aligned} \tag{35}$$

and

$$|C_{\ell,j}^{\mathrm{pool}}| \leq \|\mathbf{F}_\ell\|_2 \sqrt{K_\ell (1 - |\Phi_\ell(\theta_j)|^2)} \left[\sum_r \|(\mathbf{A}_r - \bar{\mathbf{A}}) \bar{\mathbf{x}}_\ell\|_2^2 \right]^{1/2}. \tag{36}$$

Average pooling has identical $\mathbf{A}_r$, so $C_{\ell,j}^{\mathrm{pool}} = 0$. Offset-dependent learned pooling need not have this property: identical input features can still produce different projected contributions. Thus, dispersion motivates additional conditioning information but is not a sufficient statistic for the full residual.

This distinction can also be seen by exchanging two features in a window under average pooling. Swapping $\mathbf{x}_+$ and $\mathbf{x}_-$ between coordinates p_1 and p_2 leaves the pooled feature, its normalization, and $\mathbf{s}_\ell$ unchanged. Nevertheless, the constituent-coordinate responses differ by

$$T_{\ell,j}^{\text{before}} - T_{\ell,j}^{\text{after}} = \frac{1}{K_\ell}[\mathbf{F}_\ell(\mathbf{x}_+ - \mathbf{x}_-)]_j \left(e^{ip_1\theta_j} - e^{ip_2\theta_j}\right), \tag{37}$$

which is nonzero in general. A predictor using only $(\mathbf{z}_\ell, \mathbf{s}_\ell)$ therefore cannot reconstruct every $T_{\ell,j}$, even though these descriptors can be useful for task-trained modulation. This example concerns the regional interface under average pooling; it does not assert the same invariance for offset-dependent learned pooling or for the full network.

For each layer, the predictors are implemented as

$$\nu_\ell^{(h)} = \text{Softplus}(\mathbf{W}_\nu^{(h)}\mathbf{s}_\ell), \quad a_\ell^{(h)} = 2\,\text{sigmoid}\left(\mathbf{W}_2^{(h)}\,\text{GELU}(\mathbf{W}_1[\mathbf{z}_\ell; \mathbf{s}_\ell])\right). \tag{38}$$

The linear mappings are bias-free. The hidden mapping $\mathbf{W}_1$ is shared across regional heads within a layer, while the output rows of $\mathbf{W}_\nu$ and $\mathbf{W}_2$ provide head-specific scalars. Thus $\nu_\ell^{(h)} > 0$ is a learned feature-dependent spectral scale, not an estimate of σ_ℓ^2, and $a_\ell^{(h)} \in (0, 2)$ is a common regional gate. Softplus does not impose an upper bound on $\nu_\ell^{(h)}$.

Suppressing the head index again, HoloRoPE combines these scalars as

$$u_{\ell,j} = \tfrac{1}{2}\lambda\nu_\ell\theta_j^2, \quad G_{\ell,j} = 1 + \tanh(u_{\ell,j}), \quad \widehat{\Upsilon}_{\ell,j} = a_\ell G_{\ell,j}, \quad \widehat{k}_{\ell,j} = \widehat{\Upsilon}_{\ell,j} B_{\ell,j}. \tag{39}$$

The parameter $\lambda \in \mathbb{R}$ is learned per layer and regional head. For a given region and head, ν_ℓ, a_ℓ, and λ are shared across coordinate blocks, with the componentwise spectral factor determined by the rotary frequency. For finite inputs, $G_{\ell,j} \in (0, 2)$ and $\widehat{\Upsilon}_{\ell,j} \in (0, 4)$. At nonzero frequencies, $G_{\ell,j} - 1$ has the same sign as λ. The total gain depends on a_ℓ and need not lie on the same side of one.

The identity $1 + \tanh u = e^u / \cosh u$ gives

$$\log \widehat{\Upsilon}_{\ell,j} = \log a_\ell + u_{\ell,j} - \log\cosh(u_{\ell,j}). \tag{40}$$

Holding the input and learned parameters fixed, the associated frequency profile satisfies

$$\log \widehat{\Upsilon}_\ell(\theta) = \log a_\ell + \tfrac{1}{2}\lambda\nu_\ell\theta^2 + O(\theta^4), \quad \theta \to 0. \tag{41}$$

Consequently, the bounded parameterization retains the quadratic leading term motivated by the coordinate baseline without evaluating or inverting it. The gate sets a frequency-independent regional scale, while component gain ratios depend only on their spectral factors.

For a nonzero encoded component, positivity implies

$$|\widehat{k}_{\ell,j}| = \widehat{\Upsilon}_{\ell,j}|B_{\ell,j}|, \quad \arg \widehat{k}_{\ell,j} = \arg B_{\ell,j} \pmod{2\pi}. \tag{42}$$

This is a componentwise property, not phase preservation for a sum of differently modulated components. Pooling, coordinate assignment, and values remain unchanged. The predictors and spectral parameters are optimized via the task objective. Neither $T_{\ell,j}$ nor $C_{\ell,j}$ is required as a training target.

A.6 Analysis of Positive Amplitude Modulation

HoloRoPE restricts its modulation to positive real gains. We characterize this restriction using the fixed responses $B_{\ell,j}$ and $T_{\ell,j}$ of Appendix A.3. The distances below are analytical diagnostics, not the training loss or a ranking of the two coordinate assignments.

For one component and a scalar $g > 0$, define $d(g) = |gB - T|^2$. Expansion gives

$$d(g) = g^2|B|^2 - 2g\,\text{Re}(BT^*) + |T|^2, \tag{43}$$

and hence

$$d(g) - d(1) = (g - 1)\left[(g + 1)|B|^2 - 2\,\text{Re}(BT^*)\right]. \tag{44}$$

Thus, increasing magnitude alone does not necessarily bring an encoded component closer to T. If $B \neq 0$, minimizing this diagnostic over the closed half-line $g \geq 0$ gives

$$g_\star = \max\left\{0, \frac{\text{Re}(BT^*)}{|B|^2}\right\}. \tag{45}$$

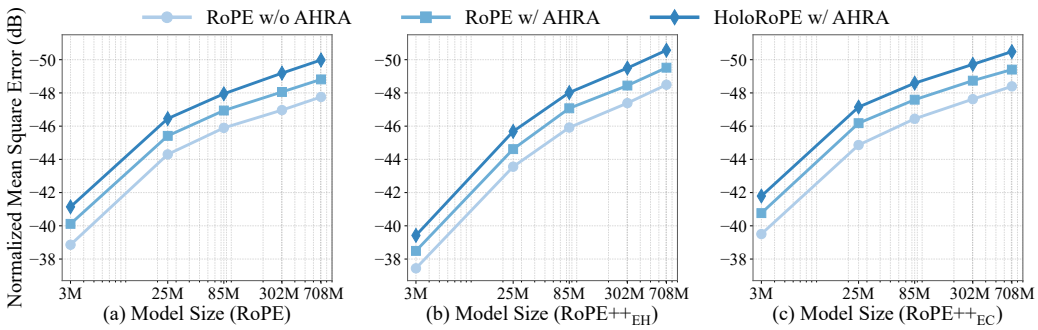


Figure 6: Model scaling on DeepMIMO under the ULA configuration in real-valued, equal-head-count (EH), and equal-cache-size (EC) settings.

The value $g_\star = 0$ is a boundary limit for positive gains. HoloRoPE uses bounded, shared, feature-conditioned gains, so this independent scalar optimum need not be realizable by its parameterization.

A positive scalar maps $B \neq 0$ to T exactly only if T/B is positive real and lies in the admissible gain range. If $B = 0$, any finite multiplicative gain leaves it zero. These facts distinguish amplitude modulation from a general complex-valued transformation. They also explain why the response comparison motivates a restricted design rather than a general reconstruction procedure.

Including the pooling bias, the response decomposition gives

$$\begin{aligned} |gB_{\ell,j} - T_{\ell,j}| &= |(g - \Phi_\ell(\theta_j))(k_{\ell,j} - \kappa_{b,j}) + (g-1)\kappa_{b,j} - C_{\ell,j}| \\ &\leq |g - \Phi_\ell(\theta_j)|\,|k_{\ell,j} - \kappa_{b,j}| + |g-1|\,|\kappa_{b,j}| + |C_{\ell,j}|. \end{aligned} \tag{46}$$

The bias cancels in $T_{\ell,j} - B_{\ell,j}$ but not generally in $gB_{\ell,j} - T_{\ell,j}$.

A two-patch example makes the phase restriction explicit. Set $p_1 = c_\ell = 0$, $p_2 = 1$, $\kappa_b = 0$, and $\kappa_1 = A/2$, $\kappa_2 = Ae^{-\mathrm{i}\theta}/2$ for $A \neq 0$ and $0 < \theta < \pi$. Then $T = A$ and $B = Ae^{-\mathrm{i}\theta/2}\cos(\theta/2)$, so $|B| < |T|$, but

$$|gB - T|^2 - |B - T|^2 = (g-1)^2|A|^2\cos^2(\theta/2) \geq 0. \tag{47}$$

Here the unmodulated magnitude is already closest to T along the admissible ray. This example neither favors a particular task model nor rules out useful learned gains; it shows that a magnitude deficit alone cannot justify a universal amplification rule.

B Experimental Appendices

B.1 Downstream Tasks and Evaluation Metrics

We organize downstream evaluation into two settings, assessing channel reconstruction and the use of channel representations for sensing and prediction.

Setting 1: Physical-Layer Reconstruction. This setting includes *Channel Extrapolation* (CExt), which predicts channels at unobserved subcarriers or antenna ports, and *Channel Estimation* (CEst), which reconstructs CSI from noisy observations. Both tasks are evaluated using normalized mean squared error (NMSE) in dB, with lower values indicating better performance.

Setting 2: Contextual Sensing and Prediction. This setting includes *LoS/NLoS Classification* (L/N), which distinguishes line-of-sight from non-line-of-sight propagation, and *Beam Prediction* (BPred), which selects a beam based on channel observations. Both tasks are evaluated by accuracy (%), with higher values indicating better performance.

B.2 Model Scale and Supplementary Ablations

We evaluate models with 3M, 25M, 85M, 302M, and 708M parameters under real-valued, equal-head-count (EH), and equal-cache-size (EC) settings. EH and EC match head counts and KV storage budgets, respectively, without matching all architectural or computational quantities.

RoPE variants with and without spatial reduction assess the effect of regional aggregation. Corresponding RoPE and HoloRoPE variants with spatial reduction assess positional encoding within the same attention structure.

Figure 6 shows that HoloRoPE consistently achieves lower NMSE than its corresponding rotary baselines across all five model scales and three comparison settings. Performance improves with increasing model capacity, while the advantage of HoloRoPE persists from compact 3M models to larger 708M models. These results suggest that regional rotary modulation complements model scaling rather than benefiting only a particular capacity or rotary formulation.