LatentWave: JEPA Pretraining for Wireless Foundation Models

Ahmed Mohamed, Ahmed Aboulfotouh, and Hatem Abou-Zeid
Department of Electrical and Software Engineering, University of Calgary, Canada

***Abstract*—Wireless foundation models have emerged as a promising alternative to building separate models for each wireless task. However, existing approaches rely on masked input reconstruction, which can bias representations toward low-level signal details. In this paper, we propose LatentWave, a wireless foundation model pretrained using a Joint-Embedding Predictive Architecture (JEPA) on diverse wireless spectrograms and channel state information (CSI). By predicting masked regions in latent space, LatentWave learns representations that are more transferable out of the box across diverse downstream tasks. The proposed architecture employs per-channel patch embeddings with stochastic channel sampling during pretraining, allowing it to process variable antenna counts and improving usability across heterogeneous wireless configurations. We evaluate LatentWave on four downstream tasks: RF signal classification, 5G NR positioning, beam prediction, and LoS/NLoS classification, comparing against a masked-modeling baseline (WavesFM) pretrained on the same data. Additionally, we show that the masking geometry introduces a task-dependent inductive bias: frequency masking strongly favors channel-related tasks such as positioning and beam prediction, while region masking better preserves discriminability for signal classification.**

I. Introduction

Conventional wireless AI systems require building a separate model for each task and environment, creating a growing collection of specialized pipelines [1] that are expensive to develop, deploy, and maintain. Foundation models and self-supervised learning [2] offer a compelling alternative: by learning general patterns from diverse, unlabeled data, a single pretrained model can be rapidly adapted to new downstream tasks, replacing many specialized systems with one general and adaptable architecture. In the context of network softwarization, such an architecture can be viewed as a reusable software intelligence layer that maps heterogeneous wireless measurements into task-agnostic representations. Rather than deploying separate AI models for each wireless service, the Wireless Foundation Model (WFM) provides a shared substrate that can support multiple downstream functions such as sensing, localization, and beam management.

Recent advances have been made toward this paradigm in the wireless domain, including the works in [3]–[7]. The dominant pretraining strategy for WFMs has been masked modeling with pixel-space reconstruction. However, this approach has an important limitation - by requiring the model to reconstruct raw input values, the model is forced to dedicate representational capacity to low-level spectral details, noise textures, and fine-grained amplitude variations that may carry little discriminative value for downstream tasks. Results indicate that WFMs pretrained with pixel-space masking and reconstruction require either fine-tuning of several last layers or parameter-efficient adaptation to achieve competitive performance on tasks that differ substantially from the pretraining distributions. This suggests that the frozen representations alone do not fully capture the high-level semantic structure needed for diverse wireless tasks.

The Joint-Embedding Predictive Architecture (JEPA) framework [8] offers an alternative - instead of reconstructing missing pixels, JEPA trains a model to predict the representations of *masked regions* in a learned latent space. Because the prediction target is itself a learned abstraction, the model is encouraged to capture higher-level semantic features while discarding irrelevant low-level variation. However, the effectiveness of such an approach has not been explored for wireless modalities and tasks that leverage CSI and spectrograms.

In this paper, we present **LatentWave**, a WFM pretrained with JEPA on a diverse dataset of wireless spectrograms and CSI. We evaluate the learned representations on four downstream tasks spanning communication, sensing, and positioning. Our main contributions are as follows:

- We propose **LatentWave**, the first JEPA-based WFM designed to support diverse downstream wireless tasks involving both spectrogram and CSI inputs. The architecture employs a per-channel patch embedding that allows processing of variable antenna counts, and a stochastic channel sampling strategy during pretraining that exposes the model to varying antenna configurations, enabling generalization across heterogeneous wireless setups.
- We conduct a systematic comparison of four masking strategies: region, frequency, time, and random with varying numbers of masks. We show that the masking geometry introduces a task-dependent inductive bias: frequency masking favors channel-related tasks such as positioning and beam prediction, while region and time masking better preserve temporal discriminability needed for signal classification. No single strategy dominates all tasks.

II. Related Work

Several recent works have explored the development of wireless foundation models. WavesFM [3] demonstrated that masked modeling can learn broadly transferable representations from image-like wireless modalities, including spectrograms and CSI, jointly supporting sensing, localization, and communication tasks. Other notable efforts have proposed foun-

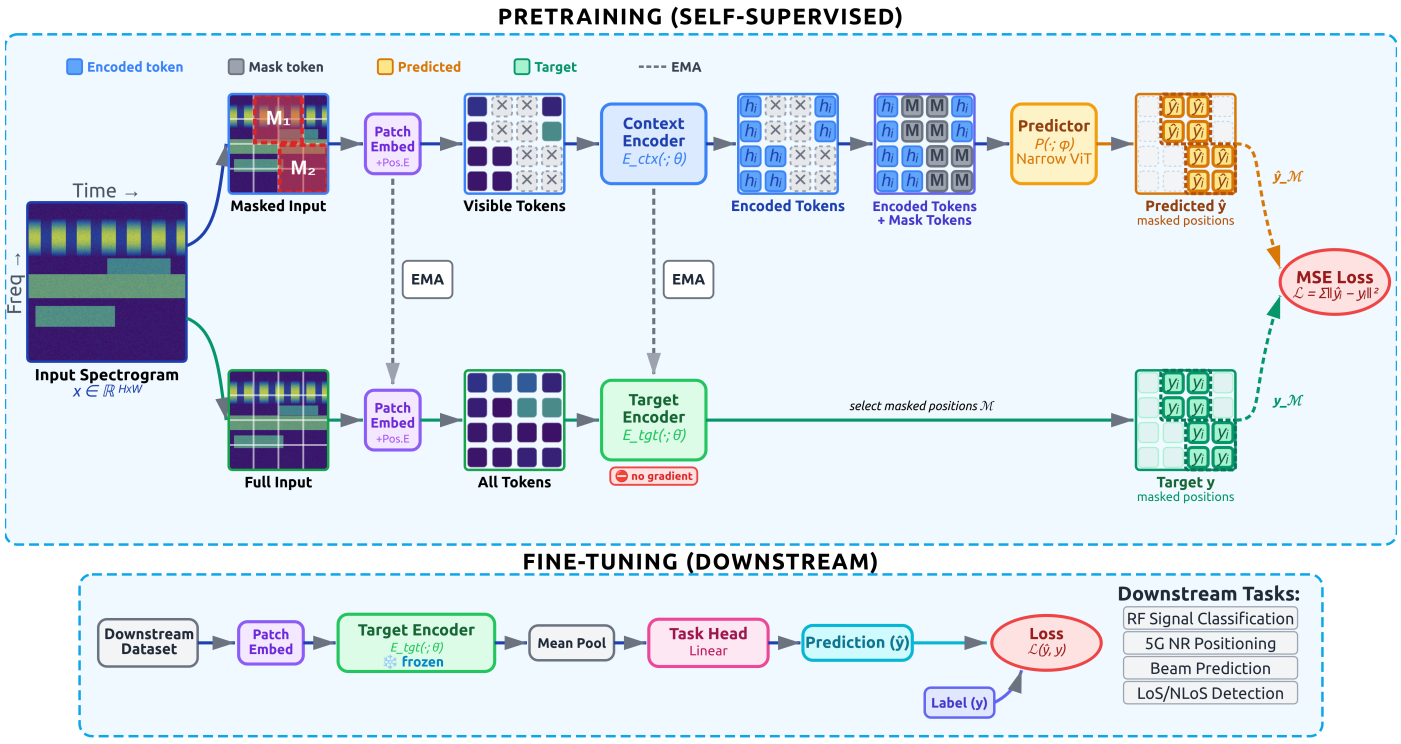

Fig. 1: Latent-WFM architecture. Pretraining (top): a JEPA-based self-supervised approach predicts latent representations of masked spectrogram patches. Fine-tuning (bottom): the pretrained encoder is adapted to downstream wireless tasks.

dation models for CSI feature extraction, using masked modeling as the pretraining objective with evaluation on line-of-sight/non-line-of-sight classification, and beam prediction [4]. Contrastive learning has also been explored as an alternative self-supervised strategy: CSI2Vec [7] employs a triplet-based contrastive objective and validates its representations on positioning and channel charting, while IQFM [5] learns general-purpose features from raw in-phase and quadrature (IQ) signals for modulation classification, RF device fingerprinting, and angle-of-arrival estimation, demonstrating strong sample efficiency in few-shot settings. Multi-modal approaches [6] that jointly handle spectrograms, CSI, and IQ signals have also been investigated.

More recently, WirelessJEPA [9] applied CNNs within a latent prediction framework to learn representations of IQ signals outperforming contrastive approaches on multiple downstream tasks. However, this work is limited to I/Q signals in both pretraining and downstream. More recently, the work in [10] proposed a JEPA-based approach to model the temporal evolution of CSI. Their results show strong long-horizon prediction, suggesting that JEPA-based learning can capture meaningful temporal wireless dynamics in the form of CSI embeddings. While this work is valuable for temporal CSI representation learning and prediction, it does not focus on learning general-purpose features for wireless foundation models that would enable a broad and diverse set of downstream wireless tasks.

III. LatentWave Methodology

This section presents the LatentWave methodology, including the architecture and masking strategies, training algorithms, and datasets for pretraining and downstream tasks.

A. Proposed Pretraining Framework and Architecture

Our foundation model is pretrained on a diverse, unlabeled dataset of wireless signal spectrograms and CSI using the JEPA framework [8]. The core principle is to learn representations by predicting in a learned latent space rather than reconstructing raw input data, as shown in Fig. 1. JEPA is a self-supervised framework that operates on masked inputs through three interacting components: a *context encoder*, a *target encoder*, and a *predictor*. Given an input, a subset of its content is masked. The context encoder processes only the visible (unmasked) portions and maps them to latent representations. The target encoder, which shares the same architecture, processes the complete input and produces latent targets. The predictor then forecasts the target encoder's representations of the masked regions using only the context encoder's output. Predictions occur in the latent space rather than in the raw input space. This process is illustrated in detail in [8].

We now describe our instantiation of this framework for wireless data. Let $x \in \mathbb{R}^{C\times H\times W}$ denote input signals with C antennas, H frequency bins and W time steps. Each antenna channel $x^{(c)} \in \mathbb{R}^{H\times W}$, $c = 1,\ldots,C$, is independently divided

into a grid of $\frac{H}{p_h} \times \frac{W}{p_w}$ non-overlapping patches of size $p_h \times p_w$. A shared linear projection maps each single-channel patch to a D-dimensional embedding. This yields a total of $N = C \times \frac{H}{p_h} \times \frac{W}{p_w}$ patch tokens $\mathbf{z} = [z_1, z_2, \ldots, z_N] \in \mathbb{R}^{N \times D}$. Because the projection operates on individual channels rather than across all C channels simultaneously, the patch embedding layer is decoupled from the number of input antennas, allowing the same model to ingest inputs with an arbitrary number of channels without architectural modification. Sinusoidal positional embeddings are added to each token to encode its spatial position within the signal grid.

To improve robustness against varying channel configurations, we employ a stochastic channel sampling strategy during pretraining. For each training sample with C available channels, we uniformly draw a random integer $C' \sim \mathcal{U}\{1, C\}$ and retain a randomly selected subset of C' channels, discarding the rest. The effective token sequence length thus varies as $N = C' \times \frac{H}{p_h} \times \frac{W}{p_w}$ across iterations, exposing the model to diverse antenna configurations.

During pretraining, we adopt a multi-block masking strategy in which several contiguous blocks spanning meaningful time–frequency-antenna regions are masked (rather than random individual patches). This encourages the model to learn high-level semantic structure in order to predict the missing regions, as local interpolation alone is insufficient. The masked patch indices form the set $\mathcal{M} \subset \{1, \ldots, N\}$, and the complementary visible set is defined as $\mathcal{V} = \{1, \ldots, N\} \setminus \mathcal{M}$. Full details of the masking strategy and its hyperparameters are provided in Section III-B.

Context Encoder. The context encoder $E_{\text{cxt}}(\,\cdot\,;\theta)$ is a standard Vision Transformer (ViT) [11]. It receives only the visible patch tokens $\{z_i\}_{i \in \mathcal{V}}$ and produces their latent representations:

$$h = E_{\text{cxt}}\big(\{z_i\}_{i \in \mathcal{V}};\, \theta\big) \in \mathbb{R}^{|\mathcal{V}| \times D}. \tag{1}$$

By processing only the unmasked subset, the context encoder is encouraged to build contextual representations from incomplete observations.

Target Encoder. The target encoder $E_{\text{tgt}}(\,\cdot\,;\bar{\theta})$ shares the same ViT architecture as the context encoder but processes the *complete* sequence of patch tokens (over the C' sampled channels) to produce the latent targets:

$$y = E_{\text{tgt}}\big(\{z_i\}_{i=1}^{N};\, \bar{\theta}\big) \in \mathbb{R}^{N \times D}. \tag{2}$$

No gradients flow through the target encoder. Instead, its parameters are updated as an Exponential Moving Average (EMA) of the context encoder weights:

$$\bar{\theta}_t = \tau\, \bar{\theta}_{t-1} + (1 - \tau)\, \theta_t, \tag{3}$$

where $\tau \in (0, 1)$ is the momentum coefficient. This asymmetric update rule ensures the target encoder evolves smoothly to provide stable training targets, preventing representation collapse.

Predictor. The predictor $P(\,\cdot\,;\phi)$ is a smaller, narrower ViT. It takes the context encoder output h and forecasts the latent representations corresponding to the masked positions:

$$\hat{y} = P\big(h;\, \phi\big) \in \mathbb{R}^{|\mathcal{M}| \times D}. \tag{4}$$

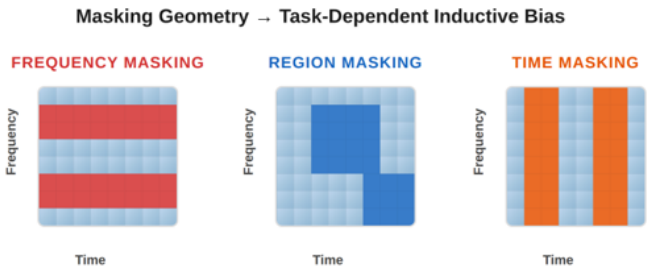

Fig. 2: JEPA mask geometries to create different wireless inductive biases.

The predictor's limited capacity is deliberate: it forces the context encoder to produce maximally informative representations rather than offloading the prediction task to the predictor.

The model is trained end-to-end to minimize the mean squared error between the predicted and target representations over the masked positions:

$$\mathcal{L} = \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} \big\|\hat{y}_i - y_i\big\|_2^2, \tag{5}$$

with gradients flowing only through the context encoder parameters θ and the predictor parameters ϕ. The complete pretraining procedure is summarized in Algorithm 1.

LatentWave Foundation Model. After pretraining, the target encoder is extracted as the foundation model for downstream tasks. Because it processed complete spectrograms throughout pretraining without any masking, its representations are directly aligned with the full inputs encountered during fine-tuning, unlike the context encoder, which was trained exclusively on partial observations.

The context and target encoders are ViT models with 8 transformer layers, 8 attention heads, embedding dimension D=256, and patch size 16×16, totaling approximately 6.4M parameters. The predictor is a smaller ViT with 4 layers, 4 attention heads, embedding dimension 128, and approximately 880K parameters. Pretraining runs for 240 epochs with batch size 256, using a cosine learning rate schedule from 10^{-6} to a peak of 10^{-3} with 12 warmup epochs, weight decay of 0.04, and EMA momentum annealed from 0.996 to 1.0. For masking, the target scale is sampled from $[0.15, 0.2]$ with aspect ratio in $[0.75, 1.5]$, and the context scale from $[0.85, 1.0]$.

B. Masking Strategy

The masking strategy defines the prediction problem faced during pretraining: what information is hidden, what context remains, and which structures the model must infer. We investigate four strategies, each imposing a distinct inductive bias on the learned representation. In all variants, the context visible to the encoder is obtained by removing the target masks and sampling a context block of scale $s_c \in [s_{c,\min}, s_{c,\max}]$ from the remaining patches. In all cases, masking a patch at a given time–frequency position removes the corresponding tokens across all sampled channels, ensuring a consistent prediction target in the spatial domain.

Algorithm 1: LatentWave JEPA Pretraining

Require: Dataset $\mathcal{D}$, epochs E, total steps T, EMA bounds $[\tau_{\min}, \tau_{\max}]$

▷ *Initialization*
Initialize context encoder $E_{cxt}(\theta)$ and predictor $P(\phi)$
$E_{tgt}(\bar{\theta}) \leftarrow \text{copy}(E_{cxt}(\theta))$ {Initialize target encoder}
Disable gradient tracking for E_{tgt} {EMA updated.}
$t \leftarrow 0$ {Global step counter}
for $e = 1$ **to** E **do**
  for each mini-batch $x \in \mathcal{D}$ **do**
    $t \leftarrow t + 1$
    ▷ **Multi-block masking**
    Sample N_m target masks $\{M_t^{(i)}\}_{i=1}^{N_m}$
    $M_c \leftarrow \text{SampleContext}(x,\ \{M_t^{(i)}\})$ {Context mask excluding target regions}
    $x_{vis} \leftarrow \text{Apply}(x,\ M_c)$ {Extract visible patches via context mask}
    ▷ **Context encoding and prediction**
    $h \leftarrow E_{cxt}(x_{vis};\ \theta)$ {Encode visible patches}
    $\hat{y} \leftarrow P(h,\ \text{pos}(\{M_t^{(i)}\});\ \phi)$ {Predict targets with positional info}
    ▷ **Target representation**
    $y_{\text{full}} \leftarrow E_{tgt}(x;\ \bar{\theta})$ {Target encoder on full input}
    $y_{norm} \leftarrow \text{LayerNorm}(y_{\text{full}})$ {Normalize}
    $y^{(i)} \leftarrow \text{Extract}(y_{norm},\ M_t^{(i)})$ for $i = 1, \ldots, N_m$ {Extract target representations}
    ▷ **Loss computation and optimization**
    $\mathcal{L} \leftarrow \frac{1}{N_m} \sum_{i=1}^{N_m} \text{MSE}\big(\hat{y}^{(i)},\ y^{(i)}\big)$ {Average loss over masks}
    Update θ, ϕ via gradient descent on $\mathcal{L}$
    ▷ **EMA and schedule updates**
    $\tau \leftarrow \tau_{\min} + \frac{t}{T}\,(\tau_{\max} - \tau_{\min})$ {Linear EMA schedule}
    $\bar{\theta} \leftarrow \tau\,\bar{\theta} + (1 - \tau)\,\theta$ {EMA update of target encoder}
    Update learning rate and weight decay {Cosine schedules}
  end for
end for

Region masking. Following I-JEPA [8], N_m two-dimensional masks are generated by sampling a target scale $s \in [s_{\min}, s_{\max}]$ and an aspect ratio $a \in [a_{\min}, a_{\max}]$, which together determine the mask dimensions along the frequency and time axes. Each mask is placed at a uniformly random position within the spectrogram. Masks may overlap with one another but not with the context region. This encourages the model to capture joint spectro-temporal structure, such as modulation patterns and bandwidth occupancy.

Frequency masking. Each of the N_m masks spans the *entire time axis* while covering a contiguous subset of frequency bins.

TABLE I: Pretraining Hyperparameters

Hyperparameter	**Value**
Encoder Architecture	
Input size ($H \times W$)	224×224
Patch size ($p_h \times p_w$)	16×16
Embedding dimension (D)	256
Transformer layers	8
Attention heads	8
Parameters	$\approx$ 6.4M
Predictor Architecture	
Embedding dimension	128
Transformer layers	2
Attention heads	4
Parameters	$\approx$ 490K
Masking Strategy	
Number of target masks (N_m)	2
Target mask scale ($[s_{\min},\ s_{\max}]$)	[0.15, 0.2]
Aspect ratio ($[a_{\min},\ a_{\max}]$)	[0.75, 1.5]
Context mask scale ($[s_{c,\min},\ s_{c,\max}]$)	[0.85, 1.0]
Minimum visible patches	10
Optimization	
Epochs	3000
Warmup epochs	150
Batch size	128
Initial learning rate	1×10^{-6}
Reference learning rate	1×10^{-3}
Final learning rate	1×10^{-6}
Initial weight decay	0.04
Final weight decay	0.4
EMA momentum ($[\tau_{\min},\ \tau_{\max}]$)	[0.996, 1.0]

The model must reconstruct complete spectral bands from the remaining frequency content, imposing a bias toward learning inter-frequency dependencies and spectral envelope structure.

Time masking. Each of the N_m masks spans the *entire frequency axis* while covering a contiguous subset of time steps. The model must predict full spectral snapshots at hidden time instants, encouraging the capture of temporal dynamics such as burst timing and temporal correlations across frames.

Random masking. As in WavesFM [3], individual patches are sampled uniformly at random without any spatial contiguity constraint. This provides maximal spatial diversity but allows local interpolation from neighboring patches to partially solve the prediction task.

For the region, frequency, and time strategies, we vary $N_m \in \{2, 3, 4, 5, 6\}$ to control the total masked area. A systematic comparison of all strategies and mask counts is presented in Section IV-C.

C. Pretraining Data

We pretrain on four unlabeled datasets spanning spectrograms and CSI. The first is a privately collected spectrogram dataset that contains 3200 WiFi, LTE, Bluetooth, 5G-NR, and ISM signals captured over the air using Software-Defined Radios (SDRs) at various sub-6GHz center frequencies with sampling rates of 10–60 MHz. The second is an RF fingerprinting spectrogram dataset comprising approximately 6000 single-channel samples [12]. The third is a 5G NR indoor CSI dataset with approximately 15 000 five-channel

TABLE II: Downstream performance under linear probing. Classification tasks report mean per-class accuracy (%, ↑); positioning reports mean positioning error (m, ↓). Results are mean ± std over three seeds. Best self-supervised result in **bold**.

Task	LatentWave (Region)	LatentWave (Freq.)	WavesFM	Supervised
RF Signal Classification	**80.9 ± 2.94**	66.1 ± 3.10	80.3 ± 0.73	86.5 ± 1.69
5G NR Positioning (m)	2.54 ± 0.009	**2.32 ± 0.023**	2.77 ± 0.037	0.71 ± 0.001
Beam Prediction	51.6 ± 0.35	**63.1 ± 0.45**	51.2 ± 0.36	88.9 ± 0.50
LoS/NLoS Classification	92.9 ± 0.57	**93.4 ± 0.26**	93.4 ± 0.54	95.9 ± 0.27

samples [13]. The fourth is a WiFi CSI dataset with 840 three-channel samples [14]. Together, the pretraining corpus totals roughly 25 000 samples with varying numbers of channels. All spectrogram samples are converted to decibel scale as $20\log_{10}(\cdot)$. Every sample, whether spectrogram or CSI, is then resized to 224×224 and normalized using per-channel mean and standard deviation computed over each pretraining dataset.

D. Baseline and Downstream Evaluation

We use WavesFM [3] as the primary baseline, since it was pretrained with masked modeling. We reproduced it on the same pretraining datasets and evaluated on the same downstream tasks. This controlled comparison isolates the effect of the JEPA pretraining objective on representation quality.

We evaluate on four downstream tasks: RF signal classification using CommRad RF [15], 5G NR positioning using 5G CSI (the outdoor scenario, distinct from the indoor scenario used during pretraining), beam prediction, and LoS/NLoS classification both using DeepMIMO [16]. These tasks span communication, sensing, and positioning domains, with several distributions and modalities that differ from the pretraining data. For each task, we evaluate using linear probing where the encoder is frozen, and a single linear head is trained on the mean-pooled patch representations for LatentWave and the CLS token for WavesFM.

IV. Results

We now present and discuss the results. Section IV-A details the evaluation metrics and implementation specifics. Section IV-B compares LatentWave against supervised and self-supervised baselines under linear probing. Section IV-C ablates the masking strategies to justify the design choices.

A. Experimental Setup

For all downstream tasks, we split the data into 80% training and 20% test sets. We report mean per-class accuracy for all classification tasks, namely RF signal classification, beam prediction, and LOS/NLOS classification, as it accounts for class imbalance more faithfully than overall accuracy. Given C classes, the mean per-class accuracy is defined as

$$\mathrm{Acc}_{\mathrm{mpc}} = \frac{1}{K}\sum_{k=1}^{K}\frac{n_k^{\mathrm{correct}}}{n_k}, \tag{6}$$

where n_k is the total number of test samples belonging to class k and n_k^{correct} is the number of correctly classified samples in that class. For the positioning task, we report the mean positioning error, defined as the average Euclidean distance between the predicted and ground-truth coordinates:

$$\mathrm{MPE} = \frac{1}{N}\sum_{i=1}^{N}\left\|\hat{\mathbf{p}}_i - \mathbf{p}_i\right\|_2, \tag{7}$$

where N is the number of test samples, $\hat{\mathbf{p}}_i \in \mathbb{R}^3$ is the predicted position, and $\mathbf{p}_i \in \mathbb{R}^3$ is the ground-truth position of the i-th sample.

For all tasks, the frozen encoder embeddings are first standardized to zero mean and unit variance using statistics computed on the training split. A logistic regression classifier is used for classification tasks, and linear regression for positioning. Main results are reported as the mean and standard deviation over three random seeds, controlling the train/test split.

B. Main Results

Table II compares the downstream performance of all methods under linear probing. LatentWave with region masking performs on par with WavesFM across all tasks, with marginal gains on RF signal classification and positioning, indicating that predicting in latent space matches pixel-level reconstruction in representation quality.

Switching to frequency masking yields a markedly different profile: beam prediction improves by over 11 percentage points and positioning error decreases from 2.54 m to 2.32 m, but RF signal classification drops substantially from 80.9% to 66.1%. This suggests that frequency masking encourages representations that capture spatial and spectral structure beneficial for channel-related tasks, at the cost of temporal patterns needed to discriminate signal types. LoS/NLoS classification remains largely unaffected across all strategies.

Despite using only a frozen encoder with a linear head, the self-supervised methods closely approach the supervised upper bound on RF classification and LoS/NLoS classification, with the remaining gap on beam prediction and positioning suggesting clear benefit from scaling pretraining data or exploring lightweight fine-tuning.

TABLE III: Masking strategy ablation. The effective masking ratio for each configuration is shown in the second header row. Evaluation metrics are identical to table II. Best result per task in **bold**.

	Region					**Frequency**					**Time**					**Rand.**
N_m	2	3	4	5	6	2	3	4	5	6	2	3	4	5	6	–
Mask ratio	29%	43%	57%	71%	80%	29%	43%	57%	71%	80%	29%	43%	57%	71%	80%	75%
RF Signal Classif.	73.3	73.1	74.4	79.7	**83.5**	74.6	73.9	68.9	68.0	68.3	76.0	78.2	77.7	80.2	81.5	71.2
5G NR Positioning (m)	3.19	2.71	2.30	2.45	2.55	3.20	2.95	2.42	2.36	**2.32**	3.43	3.54	3.43	3.18	2.84	2.90
Beam Prediction	37.2	40.7	46.5	48.6	51.6	44.6	45.6	49.8	59.5	**63.5**	33.6	38.5	37.4	38.7	44.6	39.2

C. Masking Ablations

Table III ablates the masking strategy and the number of masks N_m across three downstream tasks. The effective masking ratio, i.e., the average proportion of patches hidden from the context encoder, increases with N_m from 29% at N_m=2 to 80% at N_m=6.

Increasing N_m consistently improves beam prediction across all structured strategies. RF signal classification follows the opposite trend under frequency masking, degrading as more spectral bands are hidden, whereas region and time masking both improve with more masks. This confirms that frequency masking biases representations toward inter-frequency structure at the expense of temporal discriminability needed for signal type recognition. Positioning error generally decreases with more masks for frequency and time strategies, while region masking achieves its best positioning result at N_m=4 before slightly degrading.

Random masking at a 75% ratio performs competitively with low-mask-count structured strategies but is consistently outperformed by higher mask counts, confirming that contiguous masking with sufficient coverage yields stronger representations. Overall, no single strategy dominates all tasks, motivating the fusion approach in Table II and the task-specific strategy selection discussed in Section IV-B.

V. Conclusion

This paper presented LatentWave, a JEPA-based wireless foundation model that learns by predicting masked regions in latent space. The architecture decouples the encoder from a fixed antenna count through per-channel patch embedding and stochastic channel sampling, enabling a single pretrained model to handle both spectrogram and CSI inputs with varying numbers of channels. We also showed that the masking geometry introduces a task-dependent inductive bias: frequency masking favors channel-related tasks such as positioning and beam prediction, while region masking better preserves discriminability for signal classification. These findings position masking strategy design as a promising direction for building more transferable wireless foundation models.

References

[1] A. Alhammadi, I. Shayea, A. A. El-Saleh, M. H. Azmi, Z. H. Ismail, L. Kouhalvandi, S. A. Saad, and S. El Kafhali, "Artificial intelligence in 6G wireless networks: Opportunities, Applications, and Challenges," *Int. J. Intell. Syst.*, 2024.

[2] Z. Yang, H. Du, D. Niyato, X. Wang, Y. Zhou, L. Feng, F. Zhou, W. Li, and X. Qiu, "Revolutionizing wireless networks with self-supervised learning: A Pathway to Intelligent Communications," *arXiv preprint arXiv:2406.06872*, 2024.

[3] A. Aboulfotouh, E. Mohammed, and H. Abou-Zeid, "6G WavesFM: A foundation model for sensing, communication, and localization," *IEEE Open J. Commun. Soc.*, vol. 6, 2025.

[4] T. Yang, P. Zhang, M. Zheng, Y. Shi, L. Jing, J. Huang, and N. Li, "WirelessGPT: A generative pre-trained multi-task learning framework for wireless communication," *IEEE Network*, vol. 39, no. 5, pp. 58–65, 2025.

[5] O. Mashaal and H. Abou-Zeid, "IQFM—A wireless foundation model for I/Q streams in AI-Native 6G," *IEEE Open J. Commun. Soc.*, vol. 7, pp. 1426–1441, 2026.

[6] A. Aboulfotouh and H. Abou-Zeid, "Multimodal wireless foundation models," *arXiv preprint arXiv:2511.15162*, 2025.

[7] V. Palhares, S. Taner, and C. Studer, "CSI2Vec: Towards a universal CSI feature representation for positioning and channel charting," *arXiv preprint arXiv:2506.05237*, 2025.

[8] M. Assran, Q. Duval, I. Misra, P. Bojanowski, P. Vincent, M. Rabbat, Y. LeCun, and N. Ballas, "Self-supervised learning from images with a joint-embedding predictive architecture," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 15619–15629, June 2023.

[9] V. Chu, O. Mashaal, and H. Abou-Zeid, "WirelessJEPA: A multi-antenna foundation model using spatio-temporal wireless latent predictions," *arXiv preprint arXiv:2601.20190*, 2026.

[10] S. Naoumi, M. Bennis, and M. Chafii, "Structured latent dynamics in wireless CSI via homomorphic world models," *arXiv preprint arXiv:2603.20048*, 2026.

[11] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021.

[12] G. Reus-Muns, D. Jaisinghani, K. Sankhe, and K. Chowdhury, "Trust in 5G open RANs through machine learning: RF fingerprinting on the POWDER PAWR platform," in *Proc. IEEE Glob. Commun. Conf. (GLOBECOM)*, 2020.

[13] J. Yang, X. Chen, H. Zou, D. Wang, Q. Xu, and L. Xie, "EfficientFi: Toward large-scale lightweight WiFi sensing via CSI compression," *IEEE Internet Things J.*, vol. 9, no. 13, pp. 13086–13095, 2022.

[14] D. Wang, J. Yang, W. Cui, L. Xie, and S. Sun, "Multimodal CSI-based human activity recognition using GANs," *IEEE Internet of Things Journal*, vol. 8, no. 24, pp. 17345–17355, 2021.

[15] M. Zahid, "CommRad RF: A dataset of communication radio signals for detection, identification and classification," *Zenodo*, 2024.

[16] A. Alkhateeb, "DeepMIMO: A generic deep learning dataset for millimeter wave and massive MIMO applications," in *Proc. Inf. Theory Appl. Workshop (ITA)*, (San Diego, CA), pp. 1–8, Feb 2019.