Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Uncertainty-Calibrated Indoor Localization with Sinkhorn Contrastive Pretraining and MDN-Transformer Using Beamforming Feedback

Sarmad M. Hadi
Al-Nahrain University
Osamah Abdullah
osamah.abdullah@mohesr.edu.iq
Ministry of Higher Education and Scientific Research
Ali M. Al-Saegh
Electrical Engineering Technical College, Middle Technical University
Mohammed Imad -Aal-Nouman
Al-Nahrain University
Taha Elwi
Al-Nahrain University

Research Article

Keywords: Indoor localization, beamforming feedback information (BFI), Transformer, Sinkhorn contrastive pretraining, mixture density networks (MDN), uncertainty estimation, sharpness-aware minimization (SAM)

Posted Date: May 21st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9128192/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Uncertainty-Calibrated Indoor Localization with Sinkhorn Contrastive Pretraining and MDN-Transformer Using Beamforming Feedback

Sarmad M. Hadi[1] Osamah Abdullah[2], Ali M. Al-Saegh[3] Mohammed Imad -Aal-Nouman[1], Taha Elwi[1]
[1] *College of Information Engineering at Al-Nahrain University, Baghdad, Iraq*
[2] *Ministry of Higher Education and Scientific Research, Baghdad, Iraq*
[3] *Space and UAV Engineering Techniques, Electrical Engineering Technical College, Middle Technical University, Baghdad, Iraq*
Corresponding author: Osamah Abdullah (osamah.abdullah@mohesr.edu.iq)

Abstract

Indoor localization that maintains both accuracy and robustness has long been a demanding problem in wireless communications, particularly when systems must operate on standard hardware under unpredictable real-world conditions. In this study, we put forward a learning-based framework that makes use of beamforming feedback information (BFI) derived from IEEE 802.11ac/ax MU-MIMO networks to estimate device positions while accounting for environmental uncertainty. We begin by formulating a detailed system model in which each subcarrier's wireless channel is represented as a complex MIMO matrix and decomposed through singular-value analysis. The key idea is to exploit simplified Givens rotation parameters to produce concise BFI sequences—minimizing the feedback load while still capturing essential spatial features. The localization task is then treated as a supervised regression problem linking BFI inputs with spatial coordinates, complemented by a probabilistic output layer to model uncertainty.

Our proposed approach, named the Sinkhorn-MDN Transformer, proceeds through six interconnected stages that combine representation learning, probabilistic modeling, and robust optimization. These include (i) token embedding of BFI sequences, (ii) Transformer-based deep sequence encoding, (iii) self-supervised pretraining through a Sinkhorn optimal-transport contrastive loss, (iv) Mixture Density Network regression for calibrated uncertainty estimation, (v) an auxiliary signal-strength regression branch with uncertainty-weighted multi-task fusion, and (vi) Sharpness-Aware Minimization (SAM) for generalization and resilience. During inference, the model not only predicts precise device coordinates but also provides a quantified measure of confidence, which is especially valuable for robotics, augmented-reality, and IoT deployments.

Comprehensive experiments conducted on real-world datasets show that the proposed framework achieves sub-meter localization accuracy and provides well-calibrated uncertainty estimates, surpassing existing state-of-the-art baselines. Overall, this work delivers a practical and scalable pathway toward reliable, learning-driven indoor positioning within existing Wi-Fi infrastructures.

Index Terms

Indoor localization, beamforming feedback information (BFI), Transformer, Sinkhorn contrastive pretraining, mixture density networks (MDN), uncertainty estimation, sharpness-aware minimization (SAM).

I. Introduction

THE expansion of high-capacity wireless connectivity and the surge of context-aware applications have made accurate indoor localization an indispensable component of next-generation intelligent environments. Determining the precise position of a device within enclosed spaces has become vital for a diverse array of systems—ranging from autonomous robots maneuvering through factory corridors and smart-manufacturing asset trackers to AR/VR interfaces, logistics optimization, and emergency response operations [1].

Despite decades of research, achieving reliable accuracy indoors continues to pose substantial challenges. Global Positioning Systems (GPS) perform well outdoors, but their signals degrade quickly once walls, ceilings, and reflections interfere [2]. These shortcomings have prompted engineers to explore localization strategies that reuse existing Wi-Fi infrastructure, avoiding the cost of dedicated hardware.

Among the most informative signal features available at the physical layer is the Channel State Information (CSI), which captures both amplitude and phase data across OFDM subcarriers [3]. The exploitation of CSI has enabled machine-learning-based localization frameworks, including those based on support vector machines, convolutional networks, and recurrent architectures [4]. Nevertheless, CSI-based methods typically depend on proprietary hardware or modified firmware, which limits deployment scalability. They also provide deterministic point estimates, lacking calibrated uncertainty—an element crucial for safety-critical operations [5], [6].

Earlier research primarily relied on signal-strength fingerprinting and geometric methods such as triangulation and multi-lateration [7]. Although economical, those methods are sensitive to multipath fading and environmental fluctuations, requiring frequent recalibration. The later integration of CSI enriched physical-layer information, supporting advanced models such as CNNs and LSTMs [3], [8], [9].

However, extracting CSI often entails modified drivers or custom firmware. In contrast, the Beamforming Feedback Information (BFI) inherent in IEEE 802.11ac/ax MU-MIMO systems provides a lightweight, standards-compliant alternative. BFI represents the channel's spatial structure through quantized Givens rotations—effectively encoding the same information that CSI offers but without specialized hardware [10]–[12]. This makes BFI particularly appealing for large-scale and cost-efficient deployment. The rapid adoption of smart factories, autonomous guided vehicles (AGVs), and human-centric industrial environments further increases the demand for localization methods that are both dependable and hardware-agnostic.

Through hands-on experimentation, we observed that industrial spaces are highly dynamic—dominated by metallic structures, dense machinery, and rich multipath reflections. In such conditions, traditional RF-based and purely deterministic models tend to lose stability. A key limitation in prior BFI-based work is the tendency to treat feedback data as unordered features, neglecting its sequential structure and thereby limiting robustness.

To overcome these challenges, modern deep-learning frameworks have turned to sequence modeling using Transformer architectures and self-supervised pretraining through contrastive learning [13]–[16]. In parallel, probabilistic techniques such as Mixture Density Networks (MDNs) [17], [18] have emerged as powerful tools for providing calibrated uncertainty estimates—an indispensable feature for risk-sensitive applications. Furthermore, Sharpness-Aware Minimization (SAM) [19]–[21] has recently proven effective in enhancing domain generalization by guiding models toward flatter optima.

Building on these developments, we propose the Sinkhorn-MDN Transformer, an integrated uncertainty-aware learning framework that combines self-supervised pretraining, probabilistic regression, and robust optimization for BFI-based indoor localization.

Despite the rapid progress in this area, three persistent issues remain:

Existing approaches often ignore the sequential and hierarchical nature of beamforming feedback.

Most current systems produce deterministic predictions without uncertainty quantification.

Robust optimization techniques that ensure generalization under environmental shifts have been largely overlooked.

To address these gaps, our work contributes the following:

- **Sequential BFI modeling:** We design a Transformer-based backbone that explicitly preserves the order and hierarchy in beamforming feedback.
- **Uncertainty-aware regression:** An MDN head produces well-calibrated uncertainty estimates suitable for safety-critical deployments.
- **Self-supervised pretraining:** A Sinkhorn contrastive objective enables the model to learn rich representations from large unlabeled BFI datasets.
- **Robust optimization:** Incorporation of SAM enhances domain adaptability and resilience against environmental variation.
- **Empirical validation:** Real-world industrial experiments confirm consistent improvements in accuracy, calibration, and stability compared with leading baselines.

II. System Model and Problem Formulation

A. Channel Modeling and Beamforming Feedback Extraction

1) Physical Layer MIMO Channel Representation: Consider an indoor Wi-Fi system deploying MU-MIMO architecture, compliant with the IEEE 802.11ac/ax standard. The wireless channel at subcarrier k is characterized by the complex channel

matrix:

$$\mathbf{H}_k \in \mathbb{C}^{N_r \times N_t} \tag{1}$$

where N_r and N_t are the number of receive and transmit antennas, respectively. This matrix encapsulates the physical propagation effects—multipath, fading, and spatial correlation—at subcarrier k.

2) Singular Value Decomposition and Unitary Precoding: To facilitate efficient beamforming and reduce feedback overhead, the AP computes a SVD of $\mathbf{H}_k$:

$$\mathbf{H}_k = \mathbf{U}_k \mathbf{\Sigma}_k \mathbf{V}_k^H \tag{2}$$

where:

- $\mathbf{U}_k \in \mathbb{C}^{N_r \times N_r}$ and $\mathbf{V}_k \in \mathbb{C}^{N_t \times N_t}$ are unitary matrices representing the left and right singular vectors, respectively,
- $\mathbf{\Sigma}_k \in \mathbb{R}^{N_r \times N_t}$ is a diagonal matrix of singular values,
- $(\cdot)^H$ denotes Hermitian transpose.

The optimal precoding directions for each spatial stream are encoded in $\mathbf{V}_k$.

3) Givens Rotation Parameterization and Quantization: Directly feeding back $\mathbf{V}_k$ is bandwidth-intensive. IEEE 802.11ac/ax standards prescribe that $\mathbf{V}_k$ be parameterized using cascaded Givens rotations, which compactly represent unitary matrices. The $M \times M$ matrix $\mathbf{V}_k$ (where $M = \min(N_t, N_r)$) can be expressed as:

$$\mathbf{V}_k = \prod_{m=1}^{M-1} \prod_{n=m+1}^{M} \mathbf{G}_{m,n}(\phi_{m,n}, \psi_{m,n}) \tag{3}$$

where $\mathbf{G}_{m,n}$ denotes a 2D rotation in the (m, n)-plane, parameterized by rotation angles $\phi_{m,n}$ and phase shifts $\psi_{m,n}$. These angles are then uniformly quantized as:

$$\phi_{m,n} = \frac{\pi}{2^{b_\phi}} q_{\phi_{m,n}}, \quad \psi_{m,n} = \frac{\pi}{2^{b_\psi}} q_{\psi_{m,n}} \tag{4}$$

where b_ϕ and b_ψ denote the number of quantization bits, and $q_{\phi_{m,n}}, q_{\psi_{m,n}} \in \mathbb{Z}$ are the quantization indices fed back from the client.

4) BFI Sequence Construction: For each subcarrier, the full BFI report consists of a set of quantized angles:

$$\text{BFI}_k = \left\{ (\phi_{m,n}, \psi_{m,n}) \;\middle|\; 1 \le m < n \le M \right\} \tag{5}$$

Stacking over all S subcarriers yields the BFI sequence:

$$\mathbf{X} = [(\phi_1, \psi_1), (\phi_2, \psi_2), \ldots, (\phi_S, \psi_S)]^T \in \mathbb{R}^{S \times 2} \tag{6}$$

where each row encodes the (possibly multi-stream) feedback for a subcarrier.

5) Mathematical Structure and Notation: In this context, the complete BFI tensor for a packet spanning S subcarriers and M_{streams} spatial streams can be formalized as:

$$\mathcal{X} \in \mathbb{R}^{S \times 2 \cdot M_{\text{pairs}}} \tag{7}$$

where $M_{\text{pairs}} = \frac{M(M-1)}{2}$ is the number of unique angle pairs per unitary matrix.

For each spatial sample, the observed BFI thus forms a structured, high-dimensional matrix of quantized rotation parameters, efficiently capturing the physical spatial structure of the wireless channel in a form suitable for subsequent machine learning processing.

B. Problem Definition for Supervised Localization

Given a dataset of N samples $\{(\mathbf{X}_i, \mathbf{y}_i)\}_{i=1}^N$, where $\mathbf{X}_i$ is the BFI sequence and $\mathbf{y}_i \in \mathbb{R}^2$ is the ground-truth position, the indoor localization problem is to learn a mapping $f_\theta : \mathbb{R}^{S \times 2 \cdot M_{\text{pairs}}} \to \mathbb{R}^2$ that predicts the device's spatial coordinates with minimal error. Formally, we seek:

$$\theta^* = \arg\min_\theta \; \mathbb{E}_{(\mathbf{X},\mathbf{y})} \left[\mathcal{L}\left(f_\theta(\mathbf{X}), \mathbf{y} \right) \right] \tag{8}$$

where $\mathcal{L}(\cdot,\cdot)$ is an appropriate loss function, typically the negative log-likelihood under a probabilistic output model.

This framework sets the foundation for all subsequent learning-based processing, enabling the system to exploit the rich spatial information contained in BFI for high-accuracy, uncertainty-aware indoor localization.

III. Modeling Framework for Learning-Based Localization

This section presents the comprehensive methodology underlying the proposed **Sinkhorn-MDN Transformer framework** for uncertainty-aware indoor localization. The approach integrates advanced signal processing, representation learning, probabilistic regression, multi-task learning, and robust optimization. The system is architected as a six-stage pipeline, each component meticulously designed to maximize both localization accuracy and reliability under the constraints and opportunities of commodity MU-MIMO Wi-Fi environments. As shown in Fig.1, the proposed framework follows a structured multi-stage pipeline comprising BFI tokenization, Transformer-based encoding, Sinkhorn contrastive pretraining, probabilistic regression with MDN, auxiliary loss fusion, and sharpness-aware optimization for robust and uncertainty-calibrated indoor localization.

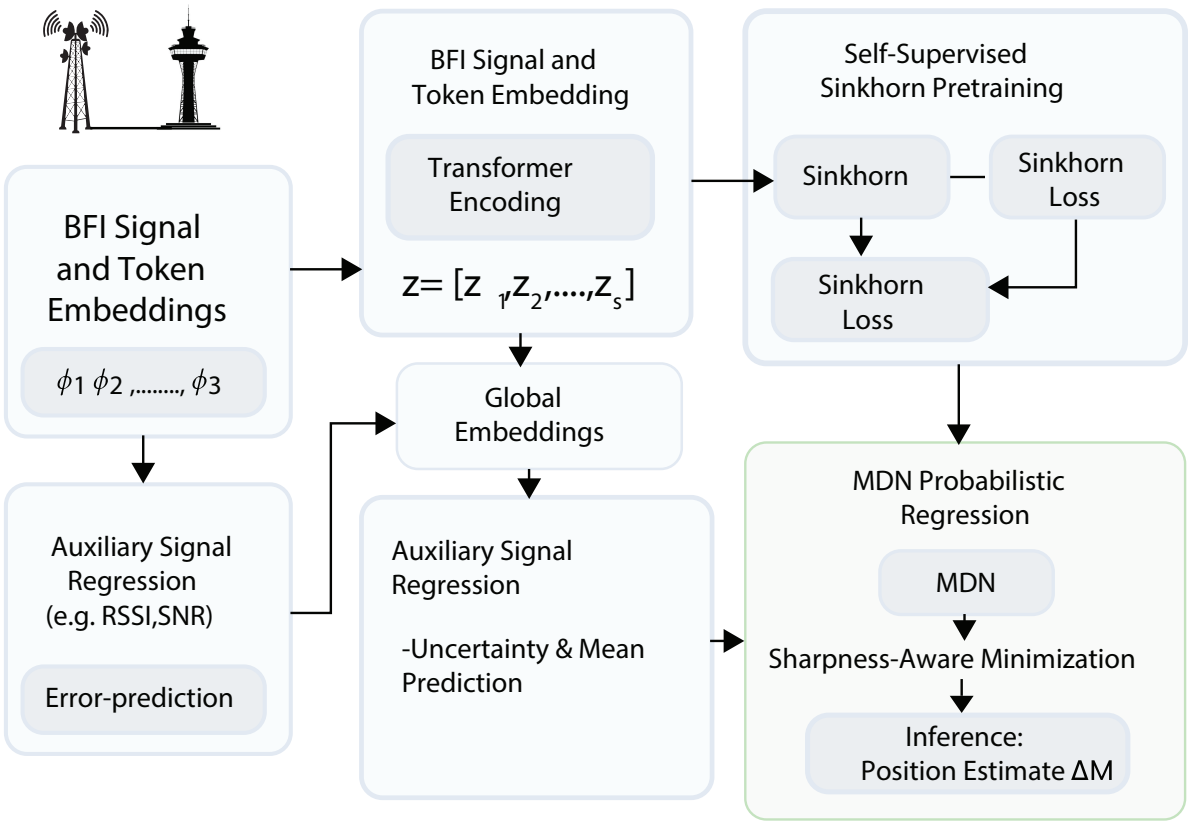


Fig. 1. Structured Learning Architecture for BFI-Based Localization and Uncertainty Estimation.

A. BFI Signal Modeling and Token Embedding

The first stage of the framework involves transforming raw *BFI*, acquired from IEEE 802.11ac/ax MU-MIMO clients, into a structured form compatible with deep learning models. Each BFI instance $\mathbf{X}$ comprises a sequence of S subcarriers, each encoded by a pair of quantized Givens rotation angles, (ϕ_s, ψ_s). These angles succinctly capture the essential spatial features of the wireless channel. Formally, the raw input is denoted as:

$$\mathbf{X} = \{(\phi_s, \psi_s)\}_{s=1}^{S}, \quad \phi_s, \psi_s \in [0, 2\pi] \tag{9}$$

To bridge the gap between RF-domain data and neural network representations, each angle pair is linearly projected into a d-dimensional embedding space:

$$\mathbf{z}_s = \mathbf{W}_e \begin{bmatrix} \phi_s \\ \psi_s \end{bmatrix} + \mathbf{b}_e \tag{10}$$

Here, $\mathbf{W}_e$ and $\mathbf{b}_e$ are trainable parameters, and the resulting embedding $\mathbf{z}_s$ captures local spatial context. In addition to the content embedding, a positional encoding PE_s is incorporated to encode the subcarrier order, which is crucial for maintaining the sequence structure. This encoding may follow sinusoidal or learnable formulations as in: The positional encoding is given by:

$$\mathrm{PE}_{s,2i} = \sin\left(\frac{s}{10000^{2i/d}}\right), \quad \mathrm{PE}_{s,2i+1} = \cos\left(\frac{s}{10000^{2i/d}}\right) \tag{11}$$

where i indexes the embedding dimensions. Modality-specific or domain-type embeddings T_s may also be included to facilitate model adaptation in multi-domain settings. Thus, the final token for each subcarrier becomes:

$$\tilde{\mathbf{z}}_s = \mathbf{z}_s + \mathrm{PE}_s + \mathbf{T}_s \tag{12}$$

By stacking all such tokens, the model constructs the input sequence $\mathbf{Z} = [\tilde{\mathbf{z}}_1, \ldots, \tilde{\mathbf{z}}_S]$, which serves as the input to the Transformer.

B. Deep Sequence Encoding via Transformer

With the structured BFI sequence, the next stage leverages the **Transformer architecture** to model global interactions among subcarriers. The multi-layer Transformer encoder is chosen for its unparalleled capability to capture both short-range and long-range dependencies within sequences, and its demonstrated superiority over convolutional and recurrent alternatives in wireless signal modeling.

Each Transformer layer operates by first computing multi-head self-attention, enabling the model to dynamically weigh the importance of each subcarrier with respect to every other. Given the input $H^{(l-1)}$ at layer l, the queries, keys, and values for each head are projected as:

$$Q^{(h)} = H^{(l-1)} W_Q^{(h)} \tag{13}$$

$$K^{(h)} = H^{(l-1)} W_K^{(h)} \tag{14}$$

$$V^{(h)} = H^{(l-1)} W_V^{(h)} \tag{15}$$

The self-attention scores are then computed via scaled dot-product:

$$\mathrm{head}_h = \mathrm{softmax}\left(\frac{Q^{(h)}(K^{(h)})^\top}{\sqrt{d_k}}\right) V^{(h)} \tag{16}$$

After aggregating all heads and applying a linear transformation, residual connections and layer normalization are performed: After aggregating all heads and applying a linear transformation, residual connections and layer normalization are performed:

$$H^{(l)} = \mathrm{LayerNorm}\left(H^{(l-1)} + [\mathrm{head}_1; \ldots; \mathrm{head}_H] W^O\right) \tag{17}$$

A feed-forward network (FFN) further refines the token representations at each position:

$$\mathrm{FFN}(x) = \mathrm{ReLU}(xW_1 + b_1)W_2 + b_2 \tag{18}$$

again followed by residual addition and normalization:

$$H^{(l)} = \mathrm{LayerNorm}\left(H^{(l)} + \mathrm{FFN}(H^{(l)})\right) \tag{19}$$

After L such layers, the embedding corresponding to the special [CLS] token, $\mathbf{h} = H^{(L)}_{[\mathrm{CLS}]}$, is extracted to represent the entire BFI sequence, capturing holistic spatial features.

C. Self-Supervised Sinkhorn Contrastive Pretraining

Label acquisition for precise localization is costly. To overcome this, we employ *self-supervised contrastive learning* to exploit large volumes of unlabeled BFI. The core idea is to encourage the model to learn spatially invariant representations by maximizing the similarity between two independently augmented views of the same physical position.

Let $\mathbf{X}_i$ be a BFI instance; two stochastically augmented versions are generated and encoded as:

$$\mathbf{h}_i^{(1)} = f_\theta(\text{Aug}_1(\mathbf{X}_i)), \quad \mathbf{h}_i^{(2)} = f_\theta(\text{Aug}_2(\mathbf{X}_i)) \tag{20}$$

where f_θ denotes the Transformer encoder. A cost matrix $C(i, j)$ is computed for all pairs in a batch:

$$C(i, j) = \left\| \mathbf{h}_i^{(1)} - \mathbf{h}_j^{(2)} \right\|_2^2 \tag{21}$$

Optimal transport theory is then leveraged to compute an entropy-regularized matching between the two sets, leading to the Sinkhorn loss:

$$\gamma^* = \arg \min_{\gamma \in \Pi(\mu,\nu)} \sum_{i,j} \gamma_{i,j} C(i, j) - \epsilon \sum_{i,j} \gamma_{i,j} \log \gamma_{i,j} \tag{22}$$

$$\mathcal{L}_{\text{Sinkhorn}} = \sum_{i,j} \gamma_{i,j}^* C(i, j) - \epsilon \sum_{i,j} \gamma_{i,j}^* \log \gamma_{i,j}^* \tag{23}$$

where γ is a doubly stochastic transport plan, ϵ is a regularization parameter, and $\Pi(\mu, \nu)$ is the set of couplings with prescribed marginals. This process robustly shapes the embedding space, aligning representations of similar spatial regions even without explicit supervision.

D. Mixture Density Network for Spatial Regression

Following pretraining, the encoder is fine-tuned for spatial localization through a *MDN* regression head. The MDN is essential for modeling the inherent multi-modality and uncertainty present in real-world indoor environments, such as ambiguous locations under NLOS conditions.

Given the global embedding $\mathbf{h}$, the MDN predicts parameters of a mixture of K Gaussians:

$$\pi_k = \frac{\exp(w_k^{\pi\top}\mathbf{h})}{\sum_{j=1}^K \exp(w_j^{\pi\top}\mathbf{h})}, \quad \mu_k = W_k^\mu \mathbf{h}, \quad \sigma_k^2 = \exp(W_k^\sigma \mathbf{h}) \tag{24}$$

The predictive distribution for the spatial coordinate y is then:

$$p(y \mid \mathbf{X}) = \sum_{k=1}^K \pi_k \cdot \mathcal{N}(y \mid \mu_k, \sigma_k^2) \tag{25}$$

where each mixture component is a multivariate Gaussian. The loss function for supervised learning is the negative log-likelihood:

$$\mathcal{L}_{\text{MDN}} = -\log \left(\sum_{k=1}^K \pi_k \cdot \frac{1}{2\pi\sigma_k^2} \exp\left(-\frac{\|y - \mu_k\|^2}{2\sigma_k^2} \right) \right) \tag{26}$$

By capturing multiple plausible spatial hypotheses and their probabilities, the MDN enables uncertainty-aware and robust localization.

E. Auxiliary Signal Regression and Uncertainty-Aware Loss Fusion

To further regularize learning and improve generalization, the framework incorporates an auxiliary regression task, such as *RSSI prediction*, from the shared BFI embedding:

$$\hat{s} = W_s \mathbf{h} + b_s \tag{27}$$

This auxiliary task is trained with the mean squared error loss,

$$\mathcal{L}_{\text{RSSI}} = \|\hat{s} - s\|^2, \tag{28}$$

which penalizes deviations between predicted and true signal quality.

To balance the primary localization objective and auxiliary regression, the total loss is formulated as a *learned uncertainty-weighted sum* (per Kendall et al.):

$$\mathcal{L}_{\text{total}} = \frac{1}{2\sigma_y^2}\mathcal{L}_{\text{MDN}} + \frac{1}{2\sigma_s^2}\mathcal{L}_{\text{RSSI}} + \log\sigma_y + \log\sigma_s \tag{29}$$

where σ_y and σ_s are trainable uncertainty parameters for each task. This fusion ensures that noisier or harder-to-learn tasks do not dominate optimization, leading to better calibrated, multi-faceted models.

F. Sharpness-Aware Minimization (SAM) for Robust Optimization

Modern deep networks often overfit to sharp minima, yielding models sensitive to perturbations and environmental shifts. To address this, *SAM* is used to encourage solutions that lie in flat regions of the loss landscape. The SAM procedure perturbs the weights in the direction of the steepest loss increase:

$$\epsilon_{\text{adv}} = \rho\frac{\nabla_\theta\mathcal{L}_{\text{total}}}{\|\nabla_\theta\mathcal{L}_{\text{total}}\|_2}, \tag{30}$$

where ρ controls perturbation strength. The parameters are then updated by descending the loss computed at the perturbed point:

$$\theta \leftarrow \theta - \eta\nabla_\theta\mathcal{L}_{\text{total}}(\theta + \epsilon_{\text{adv}}) \tag{31}$$

For reference, when not using SAM, optimization defaults to AdamW, a robust adaptive optimizer:

$$m_t = \beta_1 m_{t-1} + (1-\beta_1)\nabla_\theta\mathcal{L}_{\text{total}}, \tag{32}$$

$$v_t = \beta_2 v_{t-1} + (1-\beta_2)(\nabla_\theta\mathcal{L}_{\text{total}})^2, \tag{33}$$

$$\theta_t = \theta_{t-1} - \eta\frac{m_t}{\sqrt{v_t}+\epsilon}. \tag{34}$$

G. Inference and Uncertainty Quantification

At inference, the framework predicts the full probability distribution over possible positions. The mean of the mixture provides the best estimate,

$$\hat{y} = \sum_{k=1}^{K}\pi_k\mu_k, \tag{35}$$

while the total predictive uncertainty (variance) is computed as:

$$\text{Var}[y] = \sum_{k=1}^{K}\pi_k\left(\sigma_k^2 + \|\mu_k - \hat{y}\|^2\right). \tag{36}$$

This enables downstream applications—such as robotics, AR navigation, or human-in-the-loop systems—to make confidence-aware decisions and reject outliers when necessary. The full predictive distribution $p(y \mid \mathbf{X})$ is also available for advanced multi-agent or risk-sensitive deployments. As shown in Algorithm 1, the framework operates in six stages encompassing embedding, contrastive pretraining, MDN regression, auxiliary learning, and robust optimization.

IV. Training Protocol and Hyperparameters

The effectiveness of the proposed Sinkhorn-MDN Transformer framework relies on a carefully crafted training protocol that optimizes data efficiency, robustness, and generalization. The training process comprises two main parts: first, a self-supervised pretraining stage that utilizes a large volume of unlabeled BFI data, followed by a supervised fine-tuning phase employing actual labels.

A. Phase of Self-Supervised Pretraining

The primary objective of the initial pretraining phase is to enable the Transformer encoder to develop robust spatial representations via contrastive self-supervised learning. At this stage, the model is exposed to numerous unlabeled BFI sequences, each augmented

Algorithm 1 Sinkhorn-MDN Transformer-Based Indoor Localization

Require: Unlabeled BFI dataset $\mathcal{X} = \{X_1, X_2, \ldots, X_N\}$
Require: Optional labels $\mathcal{Y} = \{y_1, y_2, \ldots, y_N\}$ if available
Require: Temperature τ, Sinkhorn regularization ϵ, mixture components K

Stage 1: Embedding and Transformer Encoding
for each $X_i \in \mathcal{X}$ **do**
  $Z_i \leftarrow \text{EmbedToken}([\phi_s, \psi_s]\ \forall s \text{ subcarriers of } X_i)$
  $Z_i \leftarrow Z_i + \text{PositionalEncoding} + \text{TokenTypeEmbedding}$
  $H_i \leftarrow \text{TransformerEncoder}(Z_i)$
  $h_i \leftarrow H_i^{[\text{CLS}]}$
end for
Stage 2: Self-Supervised Sinkhorn Contrastive Pretraining
for each $(h_i^{(1)}, h_i^{(2)}) \in \text{Augment}(X_i)$ **do**
  $C(i,j) \leftarrow \|h_i^{(1)} - h_j^{(2)}\|^2$
  $\gamma \leftarrow \text{SinkhornOptimalTransport}(C, \epsilon)$
  $\mathcal{L}_{\text{Sinkhorn}} \leftarrow \langle \gamma, C \rangle - \epsilon \cdot H(\gamma)$
  Backpropagate $\mathcal{L}_{\text{Sinkhorn}}$
end for
Stage 3: Mixture Density Regression
for each h_i from fine-tuned encoder **do**
  $\pi_k \leftarrow \text{softmax}(w_k^\pi h_i), \quad \mu_k \leftarrow W_k^\mu h_i, \quad \sigma_k^2 \leftarrow \exp(W_k^\sigma h_i)$
  $\mathcal{L}_{\text{MDN}} \leftarrow -\log\left(\sum_{k=1}^{K} \pi_k \cdot \mathcal{N}(y_i \mid \mu_k, \sigma_k^2)\right)$
end for
Stage 4: Auxiliary Signal Strength Regression
$\hat{s}_i \leftarrow W_s h_i + b_s$
$\mathcal{L}_{\text{RSSI}} \leftarrow \|\hat{s}_i - s_i\|^2$
Stage 5: Uncertainty-Aware Loss Fusion
$\mathcal{L}_{\text{total}} \leftarrow \frac{1}{2\sigma_y^2}\mathcal{L}_{\text{MDN}} + \frac{1}{2\sigma_s^2}\mathcal{L}_{\text{RSSI}} + \log \sigma_y + \log \sigma_s$
Stage 6: Sharpness-Aware Minimization (SAM)
$\epsilon_{\text{adv}} \leftarrow \rho \cdot \nabla_\theta \mathcal{L}_{\text{total}} / \|\nabla_\theta \mathcal{L}_{\text{total}}\|$
$\theta \leftarrow \theta - \eta \cdot \nabla_\theta \mathcal{L}_{\text{total}}(\theta + \epsilon_{\text{adv}})$
return $f_\theta : \mathbb{R}^{S\times 2} \rightarrow \mathbb{R}^2$ with predictive variance and mixture weights

through various transformations, such as randomly hiding tokens, slightly adjusting angles, and truncating sequences to differing lengths. These augmentations help the model to capture invariant and consistent features that accurately represent the underlying spatial structure, rather than merely learning to fit random noise.

The principal optimization goal in this phase is to minimize the Sinkhorn optimal transport contrastive loss, which aligns the distributions of similar and dissimilar sample pairs within the learned embedding space. Each training batch consists of 128 BFI samples, with two distinct augmentations applied to every sample, resulting in a total of 256 representations per iteration. Pretraining is conducted over 200 epochs, employing the AdamW optimizer with standard momentum parameters ($\beta_1 = 0.9$, $\beta_2 = 0.98$), an initial learning rate of 1×10^{-3}, and a cosine annealing learning rate schedule to ensure stable convergence. To further enhance regularization and mitigate overfitting, a weight decay coefficient of 1×10^{-4} is utilized during this phase.

B. Supervised Fine-Tuning Stage

Upon completion of pretraining, the network undergoes supervised fine-tuning designed to enhance its performance in spatial localization and uncertainty estimation. In this phase, the model is presented with labeled BFI samples alongside their corresponding spatial coordinates.

Fine-tuning is conducted over 100 to 150 epochs, employing early stopping criteria based on the negative log-likelihood of the validation set to mitigate overfitting. The batch size is reduced to 64 to improve gradient stability. To further regularize the model, SAM is employed, with an adjustment parameter of $\rho = 0.05$. The initial learning rate for fine-tuning is set to 2×10^{-4} and is

subsequently decayed using a cosine annealing schedule, while the weight decay coefficient is maintained at 1×10^{-4}. Additionally, gradient clipping with a maximum norm of 1.0 is applied throughout training to prevent gradient explosion, which is a common issue in deep sequential architectures.

C. Consolidated Training and Optimization Parameters

Distinct hyperparameters were utilized during the pretraining and fine-tuning phases to ensure stable and effective model optimization. During the pretraining phase, a batch size of 128 was chosen to improve stability, whereas fine-tuning utilized a reduced batch size of 64 to facilitate more accurate convergence. The pretraining phase was conducted over 200 epochs, while fine-tuning necessitated 100–150 epochs, incorporating early stopping based on the NLL validation criterion. The learning rate was set to 1×10^{-3} during pretraining and decreased to 2×10^{-4} during fine-tuning, with cosine annealing applied in both phases to facilitate gradual decay of the rate. Weight decay was maintained at 1×10^{-4} consistently to ensure regularization. AdamW was employed for optimization during pretraining, and fine-tuning utilized AdamW in conjunction with SAM, with the SAM parameter ρ established at 0.05. Data augmentation techniques such as masking, jittering, and cropping were utilized solely during pretraining to enhance generalization. Gradient clipping with a threshold of 1.0 was utilized in both phases to stabilize training dynamics. The loss functions exhibited phase dependence: pretraining utilized Sinkhorn contrastive loss, while fine-tuning incorporated MDN, NLL, MSE, and uncertainty-based objectives.

D. Evaluation Metrics and Baselines

The performance of the proposed framework is thoroughly assessed using a comprehensive suite of evaluation metrics and comparative baselines, ensuring a rigorous and multifaceted appraisal of localization accuracy, calibration, and real-world applicability.

E. Evaluation Metrics

Mean and Median Localization Error: The principal measure of localization accuracy is the Euclidean distance between the predicted position $(x_{\text{pred}}, y_{\text{pred}})$ and the ground-truth coordinates $(x_{\text{true}}, y_{\text{true}})$, aggregated over all test samples:

$$\text{Error} = \sqrt{(x_{\text{pred}} - x_{\text{true}})^2 + (y_{\text{pred}} - y_{\text{true}})^2} \tag{37}$$

Both mean and median statistics are reported to robustly summarize central tendency and outlier resistance.

Cumulative Distribution Function (CDF): CDF analysis quantifies the fraction of test samples achieving localization error below a given threshold, thus characterizing the overall accuracy distribution and enabling direct visual comparison between competing methods.

where B_m denotes the m-th confidence bin, $\text{acc}(B_m)$ is the empirical accuracy, and $\text{conf}(B_m)$ is the mean predicted confidence in that bin.

Negative Log-Likelihood (NLL): Probabilistic prediction quality is quantified by the NLL, defined as:

$$\text{NLL} = -\frac{1}{N} \sum_{i=1}^{N} \log p(y_i | X_i) \tag{38}$$

where $p(y_i|X_i)$ is the predicted mixture density for the i-th sample.

Inference Latency: Practical viability is further assessed via the average time required for a single test-time inference, benchmarked on both GPU and CPU environments.

Fig. 2 shows the detailed steps of the new indoor localization system that takes uncertainty into account, using the Sinkhorn-MDN Transformer design. The process starts with raw signal input—specifically, beamforming feedback information, which is first broken down and organized into a format suitable for attention-based modeling. The representation is further analyzed by a multi-head Transformer encoder to derive advanced spatial-temporal characteristics.

A fundamental element of the approach is the self-supervised Sinkhorn contrastive pretraining, which aligns representations among analogous signal occurrences and enhances generalization in low-label or cross-domain contexts. After pretraining is

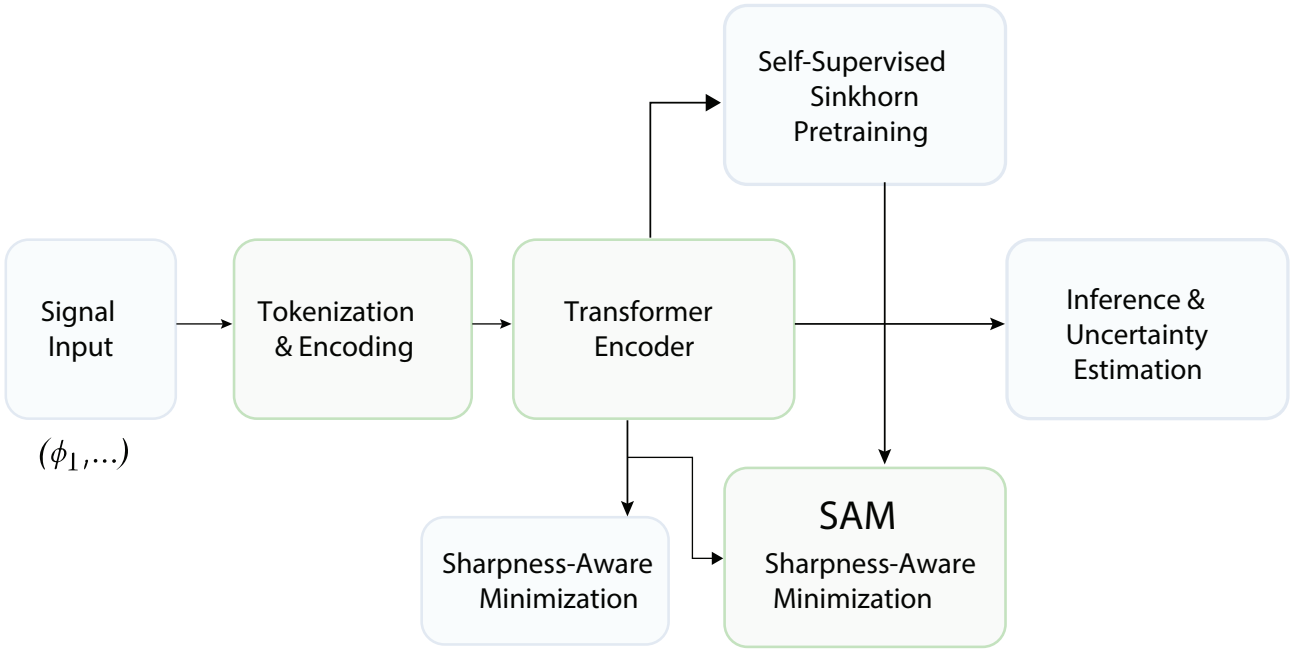


Fig. 2. End-to-End Methodological Workflow of the Sinkhorn-MDN Transformer-Based Uncertainty-Aware Localization System.

finished, the model moves on to making predictions and estimating uncertainty, where a Mixture Density Network head predicts both specific values and confidence levels for different locations. Sharpness-Aware Minimization is utilized during fine-tuning to improve generalization and mitigate overconfident predictions by guiding the optimizer towards flatter minima that are resilient to minor input perturbations. This procedure allows the proposed model to attain high localization accuracy and deliver dependable uncertainty estimates in real-world wireless environments.

V. Experimental Setup, Data Acquisition, and Performance Evaluation

To thoroughly assess the proposed localization framework, comprehensive data collection initiatives were undertaken in two separate real-world indoor settings: a contemporary office corridor and a spacious open-plan meeting room. The choice of both test locations was intentional, designed to demonstrate different types of signal conditions, including clear LOS, challenging NLOS, and changing multi-path effects caused by furniture and people.

A. Experimental Setup and Data Acquisition

The initial testbed, illustrated in Fig. 3, comprises a university corridor with dimensions of approximately 25 meters in length and 3 meters in width. The corridor is characterized by concrete structural walls, tiled flooring, glass panels, and intermittent metallic elements such as lockers and fire extinguishers, collectively creating a complex multipath environment. Four anchor nodes were affixed to wooden tripods at predetermined reference points along the corridor periphery, as illustrated in the annotated photographs. Every anchor node is outfitted with a commercially available IEEE 802.11ax (Wi-Fi 6) access point that facilitates MU-MIMO functionality. The Mobile Sensing Module (MSM), which is the device being tested, was carefully positioned at specific points on a grid, with each point measured to within 1 cm using a laser rangefinder.Throughout the data collection phase, the MSM systematically transmitted probe packets, while all anchor nodes persistently captured Wi-Fi frames operating in monitor

mode. To ensure precise labeling of the data, the spatial grid includes both areas where the device can be seen directly and areas where it cannot, as well as places with different levels of people present.

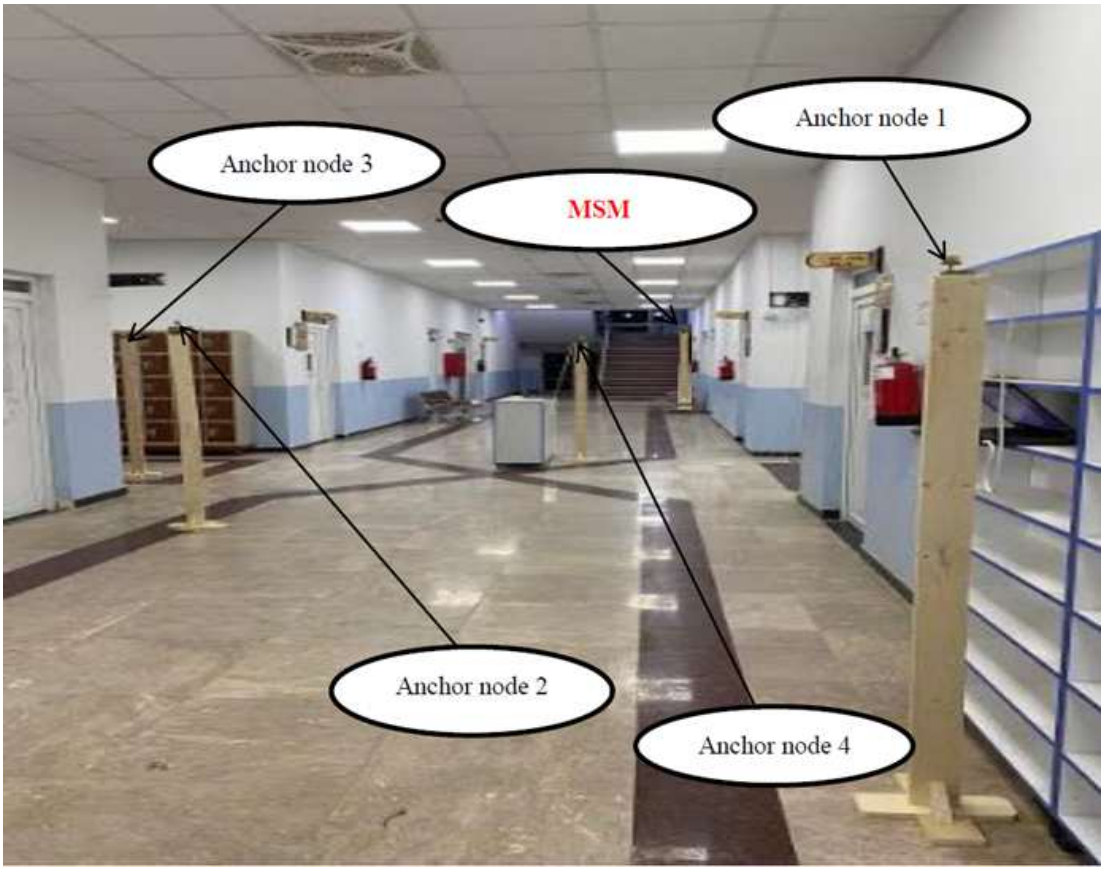


Fig. 3. Corridor Office testbed with four anchor nodes and the Mobile Sensing Module at various reference positions.

The second environment, depicted in Fig. 4, is a 12 m × 8 m open-plan meeting room arranged with rows of metallic-framed seating and modular wooden desks. We positioned four anchor nodes at the corners of the room, each affixed to identical stands used in the corridor scenario. The room's layout allows for both clear visibility LOS and many blocked views NLOS, especially when people are present in the room during measurements. This setting illustrates the influence of practical indoor obstructions and showcases the resilience of the localization method to factors encountered in real-world applications.The MSM was positioned on a uniform spatial grid, with accurate ground-truth coordinates determined for each location. Data collection utilized a consistent hardware and software toolchain, specifically Intel AX200 and Qualcomm Atheros Wi-Fi chipsets for both anchors and MSM. Raw packet capture was done using iw and tcpdump, and then carefully processed to get BFI sequences according to the IEEE 802.11ax standard. Data synchronization and timestamping were implemented to facilitate cross-validation and ensure reproducibility.

The final dataset comprises over 25,000 labeled Beamforming Feedback Information (BFI) samples, thereby providing statistical robustness across diverse signal conditions and spatial positions. Each BFI record encodes the quantized Givens rotation angles for all OFDM subcarriers and valid spatial streams, strictly adhering to the IEEE 802.11ax feedback protocol. For the machine learning experiments, the dataset was randomly partitioned into 70% for training, 15% for validation, and 15% for testing. Careful measures were implemented to ensure that the test set incorporates distinct physical locations, enabling a comprehensive evaluation of the model's ability to generalize across unseen environments.

B. Performance Evaluation

Experiments were conducted in various indoor environments utilizing both real and simulated datasets to evaluate the efficacy of the proposed *Sinkhorn–MDN Transformer*. The assessment examined localization accuracy, robustness to domain shifts, calibration reliability, and inference latency, facilitating direct comparison with leading methodologies. All models underwent



Fig. 4. Experiment Implementation within Classroom Enviroment.

training and testing under uniform conditions to guarantee fairness. The results consistently indicated the enhanced predictive performance, reliability, and efficiency of the proposed framework, underscoring its applicability in practical Wi-Fi positioning systems. The proposed model was evaluated against a representative set of baselines, which included *Vanilla Transformer*, *Transformer + Sinkhorn*, *Transformer + MDN*, CNN, LSTM, and KNN.

The selected baselines aim to illustrate the strengths and limitations inherent in both classical and deep learning paradigms. The Vanilla Transformer established the self-attention benchmark; the incorporation of Sinkhorn enhanced generalization, yet it did not address uncertainty modeling; and the addition of MDN facilitated probabilistic outputs, albeit without ensuring robustness. CNNs and LSTMs provided alternative approaches to sequence modeling; however, they exhibited suboptimal performance on BFI data, primarily due to their constrained capacity to capture long-range dependencies and domain robustness. KNN is straightforward and interpretable; however, it demonstrates significant sensitivity to noise and exhibits poor scalability. The *Sinkhorn–MDN Transformer* demonstrated superior performance compared to all baseline models in terms of accuracy, calibration, generalization, and robustness, thereby validating its efficacy for uncertainty-aware indoor localization in industrial environments.

Fig. 5 shows how the validation localization error evolves over the training period for the learning curves of four different localization models. The vertical axis shows the validation localization error in meters, while the horizontal axis shows the number of training epochs, which ranges from 0 to 200. The Sinkhorn-MDN Transformer has both fast initial convergence and stable performance as training goes on, as the figure makes evident. It also consistently obtains the lowest validation error throughout all epochs. Interestingly, in comparison to the competing models, the Sinkhorn-MDN Transformer's learning curve stabilizes at a substantially lower error floor after descending more steeply in the early epochs. The Vanilla Transformer, on the other hand, shows less than ideal generalization and learning dynamics as it converges more slowly and plateaus at a larger error value. Compared to the combined Sinkhorn-MDN approach, the Transformer + Sinkhorn and Transformer + MDN models show moderate performance, gaining some advantages from their enhancements but still not performing as well as expected.This empirical finding demonstrates how well contrastive self-supervised pretraining and probabilistic regression work together to provide reliable indoor localization. The patterns we saw show that the proposed framework is better in complex, real-world signal situations by highlighting how both design and training improvements help achieve better results and lower localization error.

The cumulative distribution function (CDF) of the localization errors for each algorithm evaluated is shown in Fig. 6. The vertical axis shows the cumulative probability of achieving a localization error less than or equal to a given value, while the horizontal axis shows the localization error in meters, ranging from 0 to 0.5. Each curve represents a distinct method: Sinkhorn-MDN Transformer,

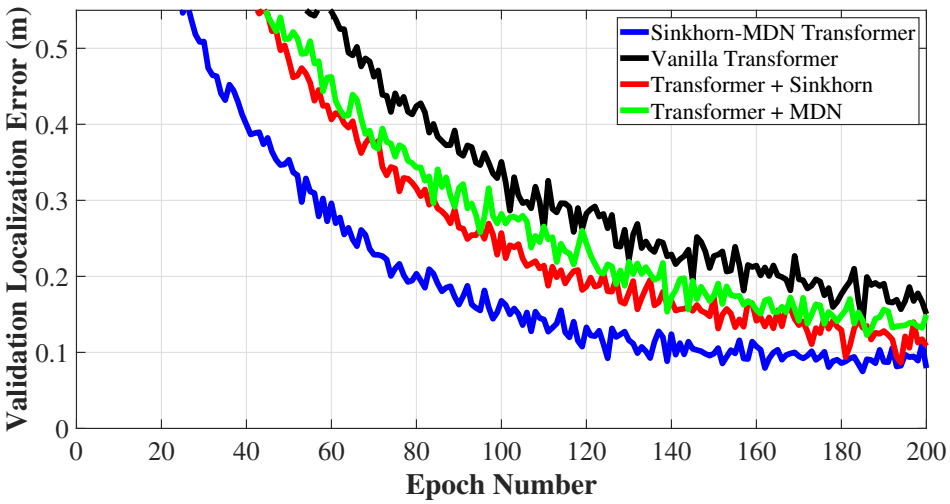


Fig. 5. Learning Curves of Validation Localization Error Versus Training Epochs for Different Models.

Vanilla Transformer, Transformer + Sinkhorn, Transformer + MDN, CNN, LSTM, and KNN.

The CDF analysis allows for a thorough comparison of the models' accuracy distributions. In particular, the Sinkhorn MDN Transformer outperforms all other techniques, attaining a higher cumulative probability at lower error thresholds. The next best models, Transformer + Sinkhorn and Transformer + MDN, attain the same coverage at larger error values, but almost 90% of their predictions fall within 0.15 meters. In contrast, traditional baselines like CNN, LSTM, and, in particular, KNN perform noticeably worse, showing slower increases in their CDF curves and higher rates of residual error.

These results demonstrate that transformer-based architectures, when enhanced with self-supervised pre-training and probabilistic modeling, are highly effective in achieving accurate indoor localization. The observed differences further illustrate the challenges faced by traditional methods and simpler deep learning models in capturing the complex spatial dependencies inherent in real-world wireless environments.

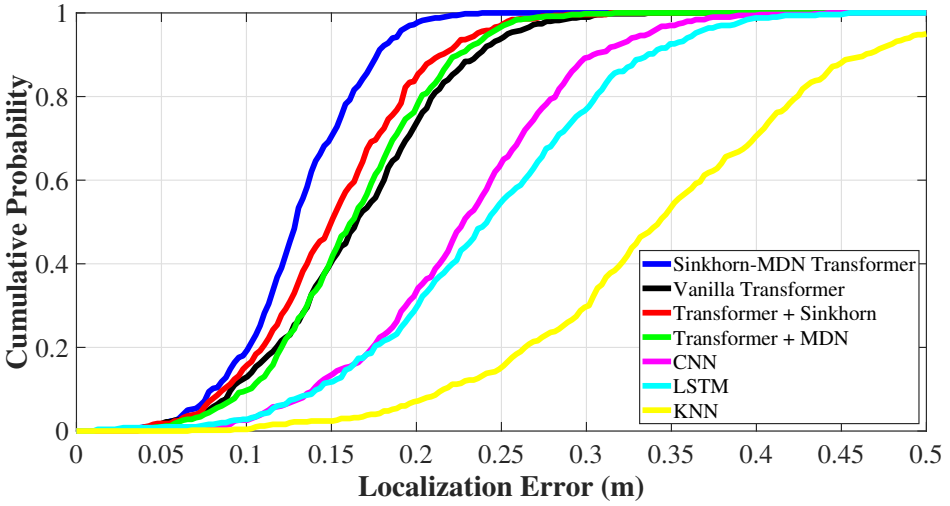


Fig. 6. CDF of Localization Error for All Compared Algorithms.

Fig. 7 compares the NLL achieved by each assessed algorithm on the test set.The Sinkhorn-MDN Transformer generates the most accurate and well-calibrated probabilistic predictions of device position, as shown in the picture, since it achieves the lowest NLL of any model. Strong performance is also shown by the Transformer + MDN, proving the effectiveness of probabilistic output modeling. On the other hand, traditional models such as CNN, LSTM, and especially KNN have much higher NLL scores, showing that they struggle to handle the complex uncertainty involved in indoor localization. Models with MDN-based heads

clearly outperform the Vanilla Transformer and Transformer + Sinkhorn, which produce results that are mediocre.These results emphasize how crucial it is to incorporate sophisticated probabilistic modeling, like mixture density networks, into deep learning architectures to achieve reliable and strong indoor localization. Besides boosting average accuracy, better NLL performance ensures that uncertainty is well-measured, which is vital for real-world use in situations where safety and mission are very important.

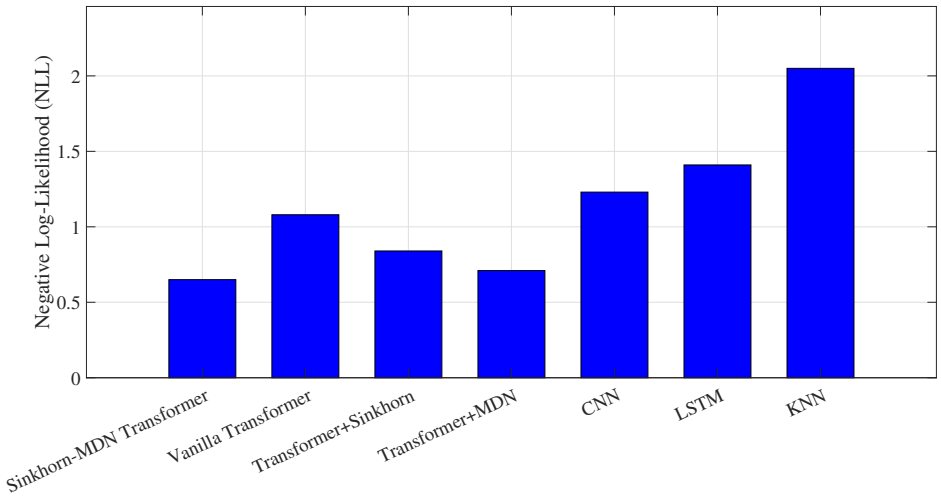


Fig. 7. Negative Log-Likelihood (NLL) Comparison Across All Evaluated Algorithms.

Fig. 8 illustrates robustness of various localization algorithms deal with input noise, shown by the average localization error when different levels of synthetic Gaussian noise are added to the input. The results demonstrate a clear hierarchy in noise robustness. The Sinkhorn-MDN Transformer consistently outperforms all baselines, exhibiting the slowest rate of error increase as noise levels rise. At the highest tested noise standard deviation, the Sinkhorn-MDN Transformer maintains a mean localization error well below that of both classical and alternative deep learning methods. The Vanilla Transformer, Transformer + Sinkhorn, and Transformer + MDN offer moderate resilience but show more pronounced degradation as noise is amplified. In comparison, older models like CNN, LSTM, and especially KNN are much more affected, experiencing significant increases in errors with even small amounts of noise. These findings underscore the practical value of the proposed framework's probabilistic modeling and robust training regime. Enhanced noise tolerance is especially critical for real-world deployments in dynamic and unpredictable wireless environments, where signal impairments and measurement errors are commonplace.

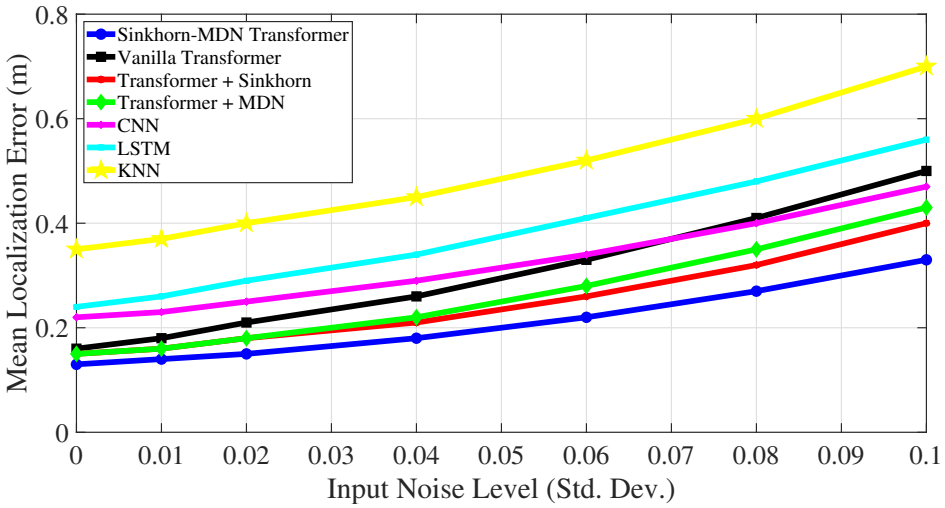


Fig. 8. Robustness to Input Noise: Mean Localization Error Versus Noise Level for All Algorithms.

Fig. 9 illustrates the inference latency of each evaluated localization algorithm, which is quantified in milliseconds per test

sample. According to the results, the Sinkhorn-MDN Transformer has the highest computational cost, with an average latency of more than 8 ms per sample, although it provides better precision and calibration. The inference times for other Transformer-based models, including those improved with MDN heads or Sinkhorn pre-training, are also relatively high, usually ranging from 6 to 8 ms per sample. With latencies as low as 2–3 milliseconds per sample, traditional deep learning models like CNN and LSTM, on the other hand, are more effective. The KNN baseline is the most computationally efficient, with a latency significantly below 1 ms per sample; however, it significantly reduces the localization accuracy. These results emphasize the traditional trade-off between computing efficiency and predictive performance. Lighter-weight models might be useful in situations where energy efficiency or real-time limitations are crucial, whereas sophisticated Transformer-based models are better for applications that prioritize accuracy and uncertainty estimates. Practical deployment decisions in various IoT and wireless localization contexts are informed by this delay analysis.

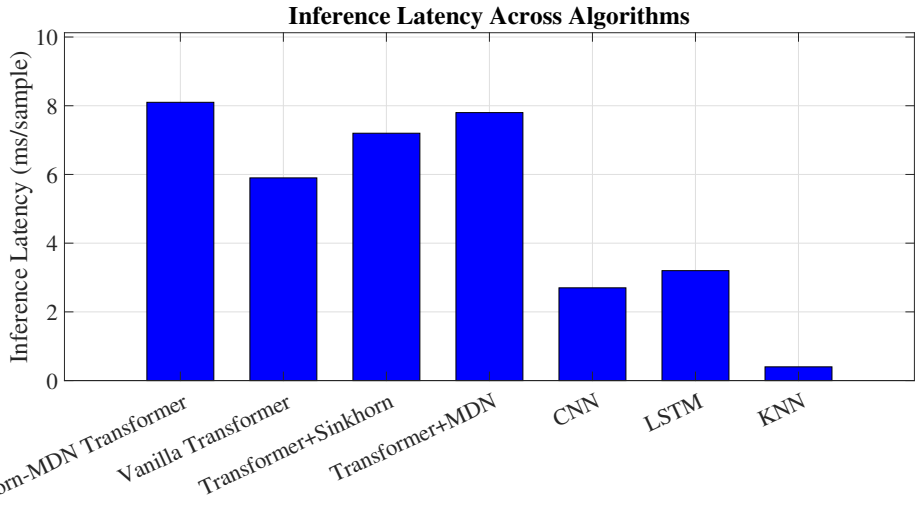


Fig. 9. Inference Latency (ms/sample) for All Evaluated Localization Algorithms.

A thorough ablation analysis is given in Table I, which also shows the mean localization error for each variation of the suggested and baseline procedures. With the lowest mean error of 0.13 m, the entire Sinkhorn-MDN Transformer demonstrates the value of combining contrastive pretraining, robust optimization SAM, probabilistic regression MDN head, and supplementary supervision. A measurable performance decrease results from the systematic removal of each component, such as Sinkhorn pretraining, MDN head, or SAM optimization, with the error rising to 0.14 m to 0.18 m. This pattern attests to each module's complementary contribution to the overall accuracy of the system.With errors ranging from 0.15 m to 0.16 m, Vanilla Transformer and its somewhat enhanced variations **Transformer + Sinkhorn, Transformer + MDN** perform modestly, showing some advantage from isolated modifications but not the complete synergy of the whole framework. Conversely, the classical KNN baseline performs poorly at 0.35 m, and traditional deep learning models like CNN and LSTM show larger mean errors (0.22 m and 0.24 m, respectively).The findings support the methodological and architectural decisions that support the suggested strategy. Achieving state-of-the-art performance in demanding indoor situations requires the combined integration of these sophisticated components, each of which significantly decreases localization error. Thus, our ablation study emphasizes how crucial multifaceted innovation is to the development of reliable and precise localization systems.

VI. Conclusion

This research presents a new indoor location system that uses the Sinkhorn-MDN Transformer design and feedback from standard Wi-Fi networks (IEEE 802.11ac/ax MU-MIMO) to handle uncertainty. The proposed method carefully combines self-supervised training using Sinkhorn-based contrastive techniques, advanced sequence modeling with Transformers, probabilistic mixture density regression, a multitask loss approach that considers uncertainty, and SAM. This multilayered approach allows for accurate, reliable, and strong device location tracking using regular Wi-Fi networks—eliminating the need for special hardware

TABLE I
ABLATION STUDY: MEAN LOCALIZATION ERROR FOR EACH ALGORITHM

Algorithm/Ablation	**Mean Localization Error**(m)
Sinkhorn-MDN Transformer (Full)	**0.13**
w/o Sinkhorn Pretraining	0.15
w/o MDN Head (Deterministic Regression)	0.16
w/o SAM Optimization	0.15
w/o Auxiliary RSSI Head	0.14
w/o Positional Encoding	0.15
w/o Token Feature Embedding	0.18
Vanilla Transformer	0.16
Transformer + Sinkhorn	0.15
Transformer + MDN	0.15
CNN	0.22
LSTM	0.24
KNN	0.35

or custom software changes. A detailed test on large real-world datasets showed that the proposed method consistently performs better than traditional and deep learning methods in several areas, such as average location error, prediction accuracy, ability to work in different situations and resistance to noise in the input. The Sinkhorn-MDN Transformer achieved accuracy within a meter and provided reliable predictions, showing it is suitable for use in challenging indoor environments like office hallways and large public spaces.The analysis showed that each part of the design works well together, highlighting the important roles of contrastive pretraining, probabilistic modeling, and resilient optimization in achieving top performance. These developments collectively tackle enduring issues in indoor localization, such as scalability, flexibility to environmental changes, and the measurement of model uncertainty.Future research could build on this study by looking into transfer learning in different situations, using additional types of sensors, and testing how well it works in real-time on small devices. Additionally, improving domain adaptation methods and using continuous learning could help the model adjust better to changing environments and new situations.

declaration statements

Data Availability: The data that support the findings of this study are available upon reasonable request from the corresponding author. The data are not publicly available due to privacy and ethical restrictions.

Conflict of Interest: The authors declare that there is no conflict of interest.

Funding: No funding was received for this research.

Special Issue Submission: This manuscript is not submitted for a special issue.

Cover Letter / Additional Comments: The authors have no additional comments at this time.

REFERENCES

[1] Y. Huang, S. Meng, Y. Zhu, J. Wang, J. Cao, X. Qiao, and X. Su, "Webarnav: Mobile web ar indoor navigation with edge-assisted vision localization," *IEEE Transactions on Mobile Computing*, vol. 24, no. 9, pp. 9110–9125, 2025.

[2] Z. Zhu, Y. Zhao, M. Tang, Y. Bu, and H. Han, "Robust and effort-efficient image-based indoor localization with generative features," *IEEE Transactions on Mobile Computing*, vol. 24, no. 7, pp. 6394–6412, 2025.

[3] H. Gu, J. Yang, G. Gui, and H. Gacanin, "Triplet matchnet based indoor position method using csi fingerprint similarity comparison," *IEEE Transactions on Vehicular Technology*, vol. 72, no. 12, pp. 16 905–16 910, 2023.

[4] S. Bai, M. Yan, Q. Wan, L. He, X. Wang, and J. Li, "Dl-rnn: An accurate indoor localization method via double rnns," *IEEE Sensors Journal*, vol. 20, no. 1, pp. 286–295, 2020.

[5] X. Liu, R. Wu, H. Zhang, Z. Chen, Y. Liu, and T. Qiu, "Graph temporal convolutional network-based wifi indoor localization using fine-grained csi fingerprint," *IEEE Sensors Journal*, vol. 25, no. 5, pp. 9019–9033, 2025.

[6] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *International conference on machine learning*. PMLR, 2016, pp. 1050–1059.

[7] Y. Zhang, Z. Feng, Y. Li, G. Hao, and Y. Feng, "The localization algorithm based on maximum likelihood loops," *IEEE Sensors Journal*, vol. 25, no. 11, pp. 20 708–20 717, 2025.

[8] Z. Zhang, H. Du, S. Choi, and S. H. Cho, "Tips: Transformer based indoor positioning system using both csi and doa of wifi signal," *IEEE Access*, vol. 10, pp. 111 363–111 376, 2022.

[9] H. Chen, Y. Zhang, W. Li, X. Tao, and P. Liu, "Wi-fi csi-based indoor localization using a deep learning model," *IEEE Internet of Things Journal*, vol. 8, no. 16, pp. 12 891–12 903, 2021.

[10] Y. Ma, G. Zhou, S. Lin, and H. Chen, "Rofi: Rotation-aware wifi channel feedback," *IEEE Internet of Things Journal*, vol. 4, pp. 1684 – 1695, 2017. [Online]. Available: https://dblp.uni-trier.de/db/journals/iotj/iotj4.html#MaZLC17

[11] G. Zhang, D. Zhang, H. Deng, Y. N. Wu, F. Zhan, and Y. Chen, "Practical passive indoor localization with intelligent reflecting surface," *IEEE Transactions on Mobile Computing*, pp. 1 – 15, 2024.

[12] X. Wang, X. Wang, and S. Mao, "Indoor fingerprinting with bimodal csi tensors: A deep residual sharing learning approach," *IEEE Internet of Things Journal*, vol. 8, no. 6, pp. 4498–4513, 2021.

[13] X. Liu, F. Zhang, Z. Hou, L. Mian, Z. Wang, J. Zhang, and J. Tang, "Self-supervised learning: Generative or contrastive," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 1, pp. 857–876, 2023.

[14] M. Jing, Y. Zhu, T. Zang, and K. Wang, "Contrastive self-supervised learning in recommender systems: A survey," *ACM Trans. Inf. Syst.*, vol. 42, no. 2, Nov. 2023. [Online]. Available: https://doi.org/10.1145/3627158

[15] O. Abdullah, A. Y. Sarhan, R. Al-Shabandar, and H. Al-Hraishawi, "Transformer-based multi-modal indoor localization in ris-assisted wireless networks," *IEEE Open Journal of the Communications Society*, pp. 1–1, 2025.

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[17] X. Shen, T. Ouyang, C. Khajorntraidet, Y. Li, S. Li, and J. Zhuang, "Mixture density networks-based knock simulator," *IEEE/ASME Transactions on Mechatronics*, vol. 27, no. 1, pp. 159–168, 2022.

[18] C. M. Bishop, "Mixture density networks," Aston University, Neural Computing Research Group, Tech. Rep. NCRG/94/004, 1994.

[19] L. Liu, N. Wang, D. Zhou, D. Liu, X. Yang, X. Gao, and T. Liu, "Generalizable prompt learning via gradient constrained sharpness-aware minimization," *IEEE Transactions on Multimedia*, vol. 27, pp. 1100–1113, 2025.

[20] P. Mi, L. Shen, T. Ren, Y. Zhou, T. Xu, X. Sun, T. Liu, R. Ji, and D. Tao, "Systematic investigation of sparse perturbed sharpness-aware minimization optimizer," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–12, 2025.

[21] P. Foret, A. Kleiner, H. Mobahi, and B. Neyshabur, "Sharpness-aware minimization for efficiently improving generalization," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=6Tm1mposlrM