Physics-Informed Map-Conditioned Transformers for Probabilistic UE Localization

Coarse-to-Fine Mixture Posteriors, Differentiable Radio-Map Likelihoods, and Test-Time Refinement

Nir Tzur

January 7, 2026

Abstract

Localizing user equipment (UE) from sparse, irregular radio measurements is an ill-posed inverse problem: multiple spatial hypotheses can explain the same measurements, and purely data-driven predictors can violate basic propagation constraints. We present a map-conditioned Transformer that amortizes Bayesian inference over UE position while remaining tightly coupled to a physics simulator through a differentiable radio-map likelihood. The model (i) encodes measurement sequences as sets of timestamped, multi-layer features, (ii) encodes multi-channel environmental context (radio maps and semantic GIS layers) via a vision Transformer, (iii) fuses both modalities with cross-attention, and (iv) outputs a coarse-to-fine top-K Gaussian mixture posterior over 2D position. Training combines a proper probabilistic objective (coarse cell cross-entropy + mixture negative log-likelihood) with a physics-consistency term computed by differentiable bilinear sampling of precomputed radio maps. At inference, the learned posterior provides calibrated uncertainty and supports optional MAP refinement by gradient descent on an energy combining the network density and the physics likelihood. We detail the model end-to-end from first principles, including the induced invariances, the probabilistic semantics of the heads, and the differentiable physics coupling.

1 Introduction

UE localization from sparse radio measurements under complex urban propagation is central to digital twins, network planning, and context-aware communications. Yet it poses two fundamental difficulties.

First, the inverse map from measurements to position is multi-valued: different locations can generate similar received power/quality patterns due to shadowing, multipath, and limited sampling. Second, the forward physics is constrained (e.g., occlusion by buildings, diffraction/reflection structure), and purely data-driven regressors can produce predictions that are not physically plausible.

This work treats localization as probabilistic inference with explicit environmental conditioning. We assume access to a multi-channel map context $\mathcal{M}$ (semantic GIS layers and/or precomputed radio fields) and a variable-length measurement sequence $\mathcal{X}$. We learn an amortized posterior $p_\theta(\mathbf{y} \mid \mathcal{X}, \mathcal{M})$ over UE position $\mathbf{y} \in \mathbb{R}^2$. To capture multi-modality at scale, we represent the posterior as a coarse-to-fine top-K Gaussian mixture tied to a coarse spatial grid. To enforce physics consistency, we add a differentiable radio-map likelihood term: predicted positions must align with precomputed radio-map features when sampled via bilinear interpolation. Finally, we expose a principled refinement mechanism: when needed, we optimize a MAP objective combining the learned posterior and the physics likelihood.

Contributions.

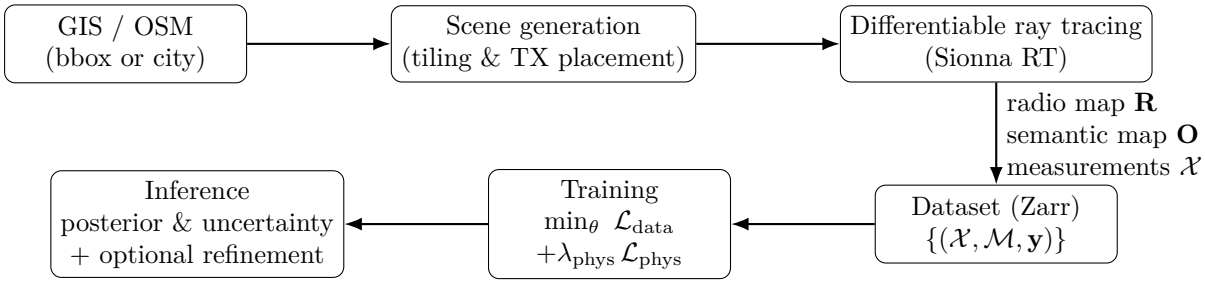

Figure 1: End-to-end pipeline: GIS/OSM → scenes and deployments → differentiable RT → dataset → training → probabilistic inference with optional refinement.

- Map-conditioned Transformer posterior. A dual-encoder architecture (Transformer-based radio encoder with `[CLS]` pooling + map ViT with optional E(2)-equivariance) fused by multi-query cross-attention.
- Coarse-to-fine mixture density head. A hierarchical discretize-then-refine posterior $p(\mathbf{y}) = \sum_c \pi_c \mathcal{N}(\mathbf{y}; \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$ with a top-K truncation that preserves multi-modality while controlling compute.
- Differentiable physics coupling. A physics-consistency loss derived as a negative log-likelihood under a measurement noise model, computed via differentiable sampling of precomputed radio fields.
- MAP refinement. An inference-time energy minimization that improves low-confidence predictions while remaining consistent with the learned posterior.

2 Related Work

Differentiable radio propagation. Differentiable ray tracing enables gradients of propagation quantities with respect to scene and system parameters, supporting inverse problems and end-to-end optimization. Sionna RT provides a differentiable ray tracer for radio propagation modeling and radio-map computation [6, 7]. RayLoc reformulates localization as an inverse ray-tracing problem with fully differentiable simulation [8]. Unlike fully differentiable localization pipelines that invoke ray tracing during inference, we amortize inference with a neural posterior conditioned on precomputed maps, and optionally apply local physics refinement when needed. WiNeRT explores neural surrogates for wireless ray tracing, aiming for fast differentiable signal rendering [9].

Radio map estimation and learning with maps. CNN-based estimators such as RadioUNet learn to predict pathloss/radio maps from urban context [10]. Transformer-based estimators have recently shown strong performance and favorable inductive biases for sparse spatial observations; STORM is an attention-based estimator for radio map estimation and active sensing [11], and TransPathNet combines transformer feature extraction with multiscale decoding for indoor pathloss mapping [12]. Physics-informed neural networks have also been proposed for radio environment mapping by embedding PDE residuals as soft constraints [13, 14].

Set-structured and probabilistic prediction. Measurement sequences are irregular and naturally set-valued; Deep Sets characterizes permutation-invariant set functions [3], and Set Transformer provides an attention-based framework for permutation-invariant/equivariant learning on sets [4]. For multi-modal regression, mixture density networks (MDNs) parameterize conditional densities via mixtures [5]. Our method combines these ideas: set-structured radio encoding, map-conditioned attention, and a mixture posterior with an explicit physics likelihood term.

3 Problem Formulation from First Principles

3.1 Forward model and inverse objective

Let $\mathcal{E}$ denote the environment (geometry, materials, terrain) and $\mathcal{T}$ the transmitter deployment (sites, antenna patterns, carrier frequency, bandwidth, etc.). For a UE location $\mathbf{y} \in \mathbb{R}^2$ (2D ground plane for simplicity), a physics simulator induces a *forward operator*

$$\Phi_{\mathcal{E},\mathcal{T}} : \mathbb{R}^2 \to \mathbb{R}^{C_r}, \qquad \mathbf{r}(\mathbf{y}) = \Phi_{\mathcal{E},\mathcal{T}}(\mathbf{y}), \tag{1}$$

where $\mathbf{r}(\mathbf{y})$ is a vector of radio features (e.g., path gain, SNR, SINR, throughput, BLER, ToA/AoA when available). In practice, we precompute $\Phi_{\mathcal{E},\mathcal{T}}$ on a grid to obtain a *radio map* $\mathbf{R} \in \mathbb{R}^{C_r \times H \times W}$ and sample it continuously at arbitrary $\mathbf{y}$ via differentiable interpolation (Sec. 5.3).

We observe a sparse sequence of measurement events $\mathcal{X}$ collected over time and potentially across cells/beams. Abstractly, for each event $t \in \{1, \dots, T\}$,

$$\mathbf{x}_t = (\underbrace{c_t, b_t}_{\text{IDs}}, \underbrace{\tau_t}_{\text{time}}, \underbrace{\mathbf{f}_t}_{\text{features}}), \qquad \mathbf{f}_t \in \mathbb{R}^d, \tag{2}$$

where c_t and b_t denote categorical identifiers (cell/beam) and $\mathbf{f}_t$ concatenates multi-layer features (ray-tracing statistics, PHY metrics, MAC metrics, etc.). The inverse goal is to infer $\mathbf{y}$ given $\mathcal{X}$ and map context $\mathcal{M}$:

$$\text{infer } \mathbf{y} \quad \text{from} \quad (\mathcal{X}, \mathcal{M}), \tag{3}$$

where $\mathcal{M}$ may include a semantic map $\mathbf{O} \in \mathbb{R}^{C_o \times H \times W}$ and the radio map $\mathbf{R}$ (or a subset of channels).

A principled approach is Bayesian inference with a posterior $p(\mathbf{y} \mid \mathcal{X}, \mathcal{M})$. We learn an amortized approximation $p_\theta(\mathbf{y} \mid \mathcal{X}, \mathcal{M})$ with explicit uncertainty.

3.2 Coordinate system and discretization

Let the physical extent be the rectangle $\Omega = [x_{\min}, x_{\max}] \times [y_{\min}, y_{\max}]$ in meters. We discretize Ω into a $G \times G$ grid of cells $\{\Omega_c\}_{c=1}^{G^2}$ with cell size $s_x = (x_{\max} - x_{\min})/G$ and $s_y = (y_{\max} - y_{\min})/G$. Let $\mathbf{c}_c \in \mathbb{R}^2$ be the center of cell Ω_c. Define the cell-index function $c(\mathbf{y})$ as the unique index such that $\mathbf{y} \in \Omega_{c(\mathbf{y})}$, and denote its center by $\mathbf{c}(\mathbf{y}) \triangleq \mathbf{c}_{c(\mathbf{y})}$. We will use a latent cell variable $C \in \{1, \dots, G^2\}$ to build a hierarchical mixture posterior (Sec. 4.4).

4 Map-Conditioned Transformer Posterior

We now specify the model $p_\theta(\mathbf{y} \mid \mathcal{X}, \mathcal{M})$ from first principles: representation of inputs, invariances, and the probabilistic head.

4.1 Radio measurement tokenization and invariances

Each measurement event $\mathbf{x}_t = (c_t, b_t, \tau_t, \mathbf{f}_t)$ is mapped to a token $\mathbf{h}_t \in \mathbb{R}^{d_r}$ via learned embeddings and linear projections:

$$\mathbf{h}_t = W\Big[\underbrace{e_c(c_t)}_{\text{cell ID}} \,\|\, \underbrace{e_b(b_t)}_{\text{beam ID}} \,\|\, \underbrace{\psi(\tau_t; \boldsymbol{\tau})}_{\text{time}} \,\|\, \underbrace{P\,\mathbf{f}_t}_{\text{features}}\Big], \tag{4}$$

where $\|$ denotes concatenation, e_c, e_b are learned embeddings, P projects continuous features (RT, PHY, MAC measurements), and ψ is a time embedding. In our implementation, ψ uses a per-sequence *normalized relative time axis*:

$$\delta_t = \tau_t - \tau_1, \qquad \bar{\delta}_t = \frac{\delta_t}{\max_s \delta_s + \varepsilon}, \qquad \psi(\tau_t; \boldsymbol{\tau}) = \mathrm{PE}(\bar{\delta}_t), \tag{5}$$

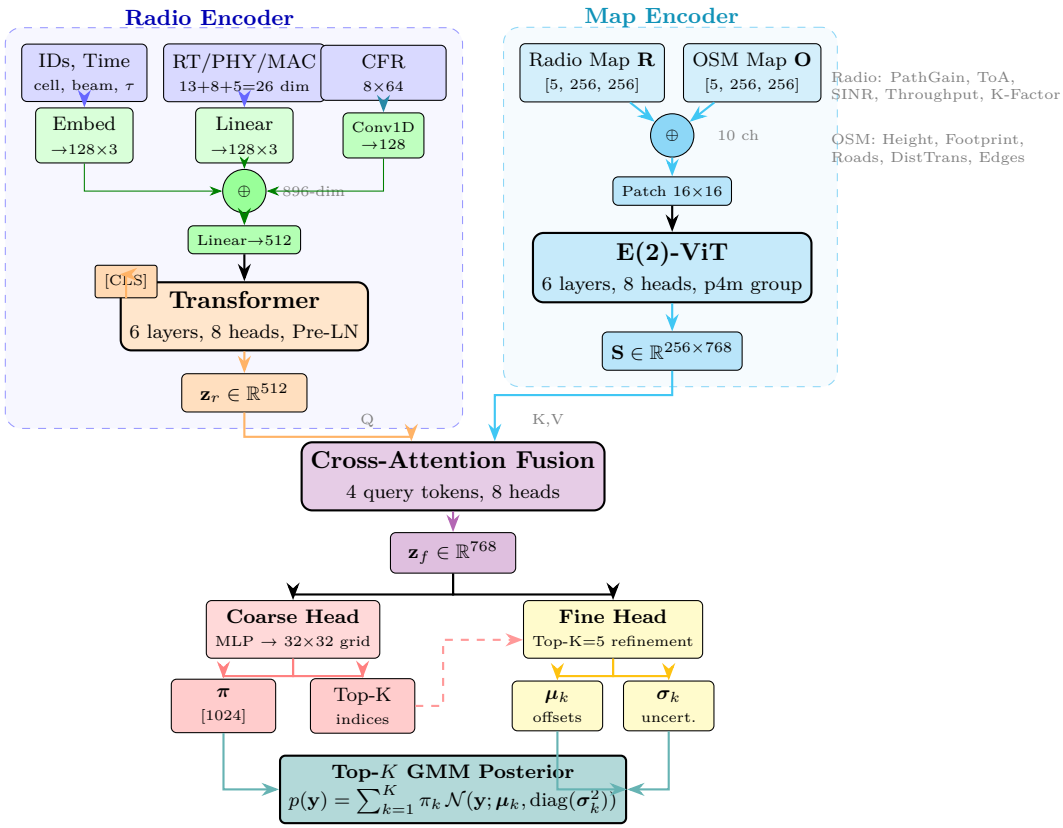

Figure 2: **Model architecture.** The Radio Encoder tokenizes multi-layer measurements (RT, PHY, MAC features + optional CFR) via learned embeddings and projections, processes them with a 6-layer Transformer, and extracts a global embedding $\mathbf{z}_r$ via a [CLS] token. The Map Encoder fuses radio and OSM maps (10 channels total) through patch embedding and an E(2)-equivariant ViT, producing spatial tokens $\mathbf{S}$. Cross-attention fusion attends radio queries to map keys/values with 4 learnable query tokens. The Coarse Head predicts a 32×32 cell probability distribution, and the Fine Head outputs sub-cell offsets and uncertainties for the Top-K cells, yielding a calibrated Gaussian mixture posterior over 2D position.

where PE is a sinusoidal lookup table (as in standard Transformers [1]).

A standard Transformer encoder with self-attention processes the token set $\{\mathbf{h}_t\}_{t=1}^{T}$ (with padding masks for variable-length sequences) and outputs contextualized embeddings; we use a learnable `[CLS]` token prepended to the sequence to pool information into a global representation $\mathbf{z}_r \in \mathbb{R}^{d_r}$.

Implementation note: Transformer vs. Set Transformer. While the Set Transformer architecture [4] provides explicit set-processing primitives (Induced Set Attention Blocks, Pooling by Multihead Attention), we employ a standard Transformer encoder with `[CLS]` pooling for simplicity and compatibility with existing tooling. Both architectures achieve permutation invariance (via the `[CLS]` aggregation for Transformers, or explicit set functions for Set Transformer), but the standard Transformer offers more straightforward integration with pre-trained weights and is computationally well-optimized. The choice can be configured in the model YAML.

Proposition 1 (Time affine invariance of the embedding)**.** *Assume timestamps transform as* $\tau'_t = a\,\tau_t + b$ *with* $a > 0$*. Then the normalized relative time coordinates* $\bar{\delta}_t$ *in Eq. (5) are invariant:* $\bar{\delta}'_t = \bar{\delta}_t$ *for all* t *(up to the stabilizer* ε*), and therefore* $\psi(\tau'_t; \boldsymbol{\tau}') = \psi(\tau_t; \boldsymbol{\tau})$*.*

This invariance makes the encoder robust to absolute clock shifts and changes in sampling rate, focusing attention on the *relative temporal structure* of measurements.

Proposition 2 (Permutation invariance via set encoding)**.** *If the encoder does not use index-based positional encodings (only token-local features such as Eq. (4)), then the Transformer is permutation* equivariant *in its token outputs, and the pooled* `[CLS]` *representation is permutation* invariant *with respect to the order of measurements.*

This aligns with the theory of set functions [3] and attention-based set models [4]. In practice, this property is crucial because measurement sequences are irregular and may arrive unordered or with missing entries.

4.2 Map encoder: radio + semantic fields

We represent map context as a multi-channel tensor $\mathbf{M} \in \mathbb{R}^{(C_r+C_o)\times H\times W}$ by early fusion of radio-map channels $\mathbf{R}$ and semantic channels $\mathbf{O}$. Our implementation provides two encoder options, configurable via the `use_e2_equivariant` flag:

Standard Vision Transformer (default). The default encoder is a standard Vision Transformer (ViT) [2]. The input $\mathbf{M}$ is partitioned into non-overlapping patches, projected to embeddings via a convolutional layer, and processed by a stack of Transformer encoder layers with learned positional embeddings. A prepended `[CLS]` token aggregates global context, and the spatial patch embeddings form the token sequence $\mathbf{S} \in \mathbb{R}^{N\times d_m}$ passed to the fusion stage. This encoder is computationally efficient and provides strong baselines, but is *not* equivariant to rotations or reflections of the input map.

E(2)-Equivariant Vision Transformer (optional). For applications where rotation/reflection invariance is important (e.g., datasets with diverse map orientations), we provide an optional E(2)-Equivariant ViT architecture based on GE-ViT [15]. This encoder explicitly preserves rotational and reflectional symmetries under the discrete group $G = p4m$ (the symmetry group of the square, $|G| = 8$). The input map $\mathbf{M}$ is first partitioned into patches, then mapped from the planar domain to the group domain via a *lifting self-attention* layer:

$$\mathbf{Z}^{(0)} = \mathrm{LiftAttn}(\mathbf{M}) \in \mathbb{R}^{d_m\times|G|\times H'\times W'}, \tag{6}$$

which lifts features from $\mathbb{R}^2$ to the semidirect product $\mathbb{R}^2 \rtimes G$. Subsequent layers employ *group-equivariant self-attention*, ensuring that rotations of the input map result in corresponding permutations of the feature fibers. Finally, spatial tokens $\mathbf{S}$ are obtained by flattening the group and spatial dimensions, preserving local equivariant structure for the fusion stage. This option is enabled by setting `use_e2_equivariant: true` in the model configuration but incurs additional computational cost.

4.3 Cross-attention fusion

We fuse radio and map modalities using *multi-query cross-attention*, where multiple learnable query tokens attend to map spatial tokens, conditioned on the radio embedding. Specifically, we maintain K_q learnable query tokens $\{\mathbf{q}_i\}_{i=1}^{K_q}$, which are combined with the radio embedding $\mathbf{z}_r$ to form radio-conditioned queries:

$$\tilde{\mathbf{q}}_i = \mathbf{q}_i + W_{\text{radio}}\,\mathbf{z}_r, \qquad i = 1,\dots,K_q. \tag{7}$$

These queries attend to the map tokens $\mathbf{S}$ via multi-head cross-attention:

$$\mathbf{z}_i = \underbrace{\mathrm{MHA}\big(Q = \tilde{\mathbf{q}}_i W_Q,\ K = \mathbf{S}W_K,\ V = \mathbf{S}W_V\big)}_{\text{cross-attention}} \in \mathbb{R}^{d_f}, \qquad i = 1,\dots,K_q. \tag{8}$$

The K_q output embeddings are concatenated and projected to yield the fused representation:

$$\mathbf{z}_f = W_{\text{agg}} \left[\mathbf{z}_1 \,\|\, \cdots \,\|\, \mathbf{z}_{K_q}\right] \in \mathbb{R}^{d_f}. \tag{9}$$

A residual MLP refines $\mathbf{z}_f$ (standard Transformer block structure [1]). This multi-query design (with $K_q = 4$ by default) allows the fusion module to capture multiple aspects of the radio-map relationship simultaneously, providing richer context for the downstream heads than single-query attention.

4.4 Coarse-to-fine mixture posterior

Hierarchical latent variable view. We introduce a discrete latent cell variable $C \in \{1, \ldots, G^2\}$ and factor the posterior as

$$p_\theta(\mathbf{y} \mid \mathcal{X}, \mathcal{M}) = \sum_{c=1}^{G^2} p_\theta(C = c \mid \mathcal{X}, \mathcal{M})\; p_\theta(\mathbf{y} \mid C = c, \mathcal{X}, \mathcal{M}). \tag{10}$$

The coarse head outputs logits over cells and a categorical distribution

$$\pi_c \;=\; p_\theta(C = c \mid \mathcal{X}, \mathcal{M}) = \text{softmax}(\mathbf{u})_c, \qquad \mathbf{u} = \text{MLP}_{\text{coarse}}(\mathbf{z}_f). \tag{11}$$

Cell-conditioned continuous refinement. For each cell c, the fine head parameterizes a local Gaussian over positions centered at the cell center $\mathbf{c}_c$:

$$p_\theta(\mathbf{y} \mid C = c, \mathcal{X}, \mathcal{M}) = \mathcal{N}\Big(\mathbf{y};\; \mathbf{c}_c + \boldsymbol{\mu}_c,\; \text{diag}(\boldsymbol{\sigma}_c^2)\Big)\,, \tag{12}$$

with $(\boldsymbol{\mu}_c, \boldsymbol{\sigma}_c) = \text{Head}_{\text{fine}}(\mathbf{z}_f,\, e_{\text{cell}}(c))$ and e_{cell} a learned embedding of the cell index (a "local chart" for each coarse region). This *cell-conditioned* parameterization is essential: without $e_{\text{cell}}(c)$, the fine head cannot represent spatially varying offsets/uncertainties tied to the coarse hypothesis.

Top-K truncation. Evaluating all G^2 components can be expensive. We therefore retain only the top-K most likely cells $\{c_k\}_{k=1}^K$ under $\boldsymbol{\pi}$ and renormalize:

$$\hat{\pi}_k = \frac{\pi_{c_k}}{\sum_{j=1}^K \pi_{c_j}}, \qquad k = 1, \ldots, K. \tag{13}$$

The resulting approximation is the top-K mixture posterior

$$p_{\theta,K}(\mathbf{y} \mid \mathcal{X}, \mathcal{M}) = \sum_{k=1}^{K} \hat{\pi}_k\; \mathcal{N}\Big(\mathbf{y}; \mathbf{c}_{c_k} + \boldsymbol{\mu}_{c_k}, \text{diag}(\boldsymbol{\sigma}_{c_k}^2)\Big)\,. \tag{14}$$

This truncation preserves multi-modality while controlling computation; see Appendix A for a quantitative discussion of the approximation error.

Uncertainty decomposition. The mixture posterior supports a principled decomposition of predictive uncertainty:

$$\mathbb{E}[\mathbf{y}] = \sum_{k=1}^{K} \hat{\pi}_k\, \boldsymbol{m}_k, \qquad \boldsymbol{m}_k = \mathbf{c}_{c_k} + \boldsymbol{\mu}_{c_k}, \tag{15}$$

$$\text{Cov}[\mathbf{y}] = \sum_{k=1}^{K} \hat{\pi}_k \Big(\boldsymbol{\Sigma}_k + (\boldsymbol{m}_k - \mathbb{E}[\mathbf{y}])(\boldsymbol{m}_k - \mathbb{E}[\mathbf{y}])^\top\Big), \quad \boldsymbol{\Sigma}_k = \text{diag}(\boldsymbol{\sigma}_{c_k}^2), \tag{16}$$

where the first term is within-component (heteroscedastic) uncertainty and the second captures between-component multi-modality.

5 Learning Objective and Physics Coupling

Training must (i) concentrate probability mass on the correct region of Ω, (ii) learn calibrated uncertainty and multi-modality, and (iii) respect the propagation constraints encoded by the radio maps. We therefore combine a *coarse* classification term, a *fine* mixture negative log-likelihood, and a *physics* consistency term. We abbreviate the supervised data term as $\mathcal{L}_{\text{data}} \triangleq \lambda_{\text{coarse}} \mathcal{L}_{\text{coarse}} + \lambda_{\text{fine}} \mathcal{L}_{\text{fine}}$.

5.1 Coarse supervision on the latent cell

Given a ground-truth position $\mathbf{y}^\star$, define the corresponding cell index

$$c^\star \triangleq c(\mathbf{y}^\star) \in \{1, \dots, G^2\}. \tag{17}$$

The coarse head outputs $\boldsymbol{\pi}$ (Eq. (11)). We apply the standard cross-entropy / negative log-likelihood

$$\mathcal{L}_{\text{coarse}}(\theta) = -\log \pi_{c^\star}. \tag{18}$$

This term is particularly important with top-K routing: if the true cell is not in the selected set, gradients from the fine loss alone may be weak or absent for the coarse logits. Eq. (18) ensures the gating distribution learns to place mass near the correct region.

5.2 Fine top-K mixture negative log-likelihood

Given the top-K posterior $p_{\theta,K}$ (Eq. (14)), the proper scoring rule for density prediction is the negative log-likelihood

$$\mathcal{L}_{\text{fine}}(\theta) = -\log p_{\theta,K}(\mathbf{y}^\star \mid \mathcal{X}, \mathcal{M}) = -\operatorname{LogSumExp}_{k=1}^{K} \left(\log \hat{\pi}_k + \log \mathcal{N}(\mathbf{y}^\star; \boldsymbol{m}_k, \boldsymbol{\Sigma}_k)\right). \tag{19}$$

Equation (19) is *not* equal to the commonly used surrogate $\sum_k \hat{\pi}_k \left(-\log \mathcal{N}(\mathbf{y}^\star; \boldsymbol{m}_k, \boldsymbol{\Sigma}_k)\right)$. By Jensen's inequality, the surrogate upper-bounds the true mixture NLL and tends to encourage *all* components to explain every sample, reducing useful multi-modality. In contrast, Eq. (19) rewards allocating probability mass to *at least one* well-placed component, which is the correct likelihood for mixture density networks [5].

5.3 Differentiable radio-map lookup

To couple predictions to physics, we require a differentiable map $\mathbf{y} \mapsto \mathbf{r}(\mathbf{y})$ extracted from the precomputed radio map $\mathbf{R}$. Let $\mathbf{R}$ be defined over a grid in Ω with resolution Δ meters/pixel. We define sample$(\mathbf{R}, \mathbf{y})$ as bilinear interpolation of each channel at location $\mathbf{y}$. In implementation, this is realized by mapping $\mathbf{y}$ to normalized coordinates in $[-1, 1]^2$ and applying a differentiable sampler (e.g., `grid_sample`).

Formally, for channel $j \in \{1, \dots, C_r\}$,

$$r_j(\mathbf{y}) = \sum_{p \in \mathcal{N}(\mathbf{y})} w_p(\mathbf{y})\ R_j[p], \tag{20}$$

where $\mathcal{N}(\mathbf{y})$ are the four neighboring grid points and $w_p(\mathbf{y})$ are bilinear weights. This map is piecewise differentiable in $\mathbf{y}$, enabling gradient-based refinement (Sec. 6.2).

5.4 Physics-consistency loss as a likelihood term

Let $\mathbf{m} \in \mathbb{R}^{C_{\text{phys}}}$ denote a vector of observed features extracted from $\mathcal{X}$ that correspond to channels in the radio map (e.g., {path gain, SNR, SINR, throughput, BLER}). Let $\mathbf{r}(\mathbf{y}) \in \mathbb{R}^{C_{\text{phys}}}$ denote the corresponding radio-map features sampled at $\mathbf{y}$. A simple measurement model is

$$\mathbf{m} = \mathbf{r}(\mathbf{y}) + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \operatorname{diag}(\boldsymbol{\sigma}_{\text{phys}}^2)). \tag{21}$$

The negative log-likelihood (up to constants) yields a weighted squared residual:

$$-\log p(\mathbf{m} \mid \mathbf{y}, \mathbf{R}) \;\propto\; \sum_{j=1}^{C_{\text{phys}}} \frac{(m_j - r_j(\mathbf{y}))^2}{\sigma^2_{\text{phys},j}} \;=\; \left\| W_{\text{phys}}(\mathbf{m} - \mathbf{r}(\mathbf{y})) \right\|_2^2, \tag{22}$$

with $W_{\text{phys}} = \text{diag}(\sigma^{-1}_{\text{phys},j})$. This motivates the physics loss

$$\mathcal{L}_{\text{phys}}(\mathbf{y}) = \left\| W_{\text{phys}}(\mathbf{m} - \mathbf{r}(\mathbf{y})) \right\|_2^2, \tag{23}$$

optionally replaced by a robust penalty (e.g., Huber) to reduce sensitivity to outliers or systematic simulator–measurement mismatch.

Implementation note: feature normalization. Normalizing each feature channel by its empirical mean and standard deviation (as done in our code) is equivalent to using a diagonal covariance estimate in Eq. (21) and improves conditioning of the refinement gradients.

5.5 Total objective and optimization

Let θ denote all neural parameters. We minimize

$$\min_{\theta} \; \lambda_{\text{coarse}}\, \mathbb{E}[\mathcal{L}_{\text{coarse}}(\theta)] + \lambda_{\text{fine}}\, \mathbb{E}[\mathcal{L}_{\text{fine}}(\theta)] + \lambda_{\text{phys}}\, \mathbb{E}\big[\mathcal{L}_{\text{phys}}(\hat{\mathbf{y}}_\theta(\mathcal{X}, \mathcal{M}))\big]\,, \tag{24}$$

where $\hat{\mathbf{y}}_\theta$ is a point estimate extracted from the posterior (Sec. 6) and expectations are over the training distribution of $(\mathcal{X}, \mathcal{M}, \mathbf{y}^\star)$. The weights $\lambda_{\text{coarse}}, \lambda_{\text{fine}}, \lambda_{\text{phys}}$ control the relative influence of gating accuracy, probabilistic calibration, and physical consistency.

6 Inference and MAP Refinement

6.1 Point prediction

From the mixture posterior $p_{\theta,K}$ we can extract:

- **MAP component mean:** $\hat{\mathbf{y}}_{\text{MAP}} = \boldsymbol{m}_{k^\star}$ with $k^\star = \arg\max_k \hat{\pi}_k$.
- **Posterior mean:** $\hat{\mathbf{y}}_{\text{mean}} = \mathbb{E}[\mathbf{y}]$ (Eq. (16)).

The MAP mean is robust for multi-modal distributions (it chooses a single hypothesis), while the posterior mean can be biased toward "in-between" regions when hypotheses are far apart.

6.2 MAP refinement via energy minimization

The physics loss enables test-time refinement. We define an energy over $\mathbf{y}$ that can combine the learned posterior density and the physics likelihood:

$$\mathcal{E}(\mathbf{y}) = \alpha\Big(-\log p_{\theta,K}(\mathbf{y} \mid \mathcal{X}, \mathcal{M}) \Big) + \beta\, \mathcal{L}_{\text{phys}}(\mathbf{y}), \tag{25}$$

where $\alpha \geq 0$ and $\beta \geq 0$ control the trust in the network posterior and the physics term, respectively. The special case $\alpha = 0$ corresponds to *physics-only* refinement (the setting used in our current refinement module), while $\alpha > 0$ yields a composite MAP objective that keeps refined solutions in high-density regions of the learned posterior.

Refinement is most useful when the posterior is diffuse (low confidence) or when small corrections are needed to satisfy propagation constraints. In practice, refinement can be applied conditionally (e.g., only when $\max_k \hat{\pi}_k$ falls below a threshold). Refining multiple candidate hypotheses $\{\boldsymbol{m}_k\}_{k=1}^K$ and selecting the one with the lowest physics loss is a simple extension that further improves robustness when the posterior is strongly multi-modal.

Algorithm 1 Optional test-time refinement (per sample)

Require: posterior parameters $\{\hat{\pi}_k, \boldsymbol{m}_k, \boldsymbol{\Sigma}_k\}_{k=1}^{K}$, observed features $\mathbf{m}$, radio map $\mathbf{R}$

initialize $\mathbf{y}^{(0)} \leftarrow \boldsymbol{m}_{k^\star}$ where $k^\star = \arg\max_k \hat{\pi}_k$

for $i = 0, 1, \ldots, S-1$ **do**

  $\mathbf{g}^{(i)} \leftarrow \nabla_{\mathbf{y}} \mathcal{E}(\mathbf{y}^{(i)})$ using Eq. (25) and differentiable lookup

  $\mathbf{y}^{(i+1)} \leftarrow \Pi_\Omega\left(\mathbf{y}^{(i)} - \eta\, \mathbf{g}^{(i)}\right)$ ▷ step size η, project to extent

end for

return refined position $\mathbf{y}^{(S)}$

7 Data Generation Pipeline and Experimental Protocol

7.1 Synthetic data generation with differentiable ray tracing

We generate datasets by (i) constructing 3D scenes from GIS/OSM data, (ii) placing transmitter sites according to configurable strategies, and (iii) running differentiable ray tracing to compute dense radio maps and sample sparse measurement sequences. Sionna RT provides GPU-accelerated differentiable ray tracing for radio propagation and supports radio-map computation [6, 7]. The resulting dataset is stored in a chunked array format (Zarr) to support large-scale training.

7.2 Inputs and channels

We use a multi-channel map tensor with C_r radio channels and C_o semantic channels. Example semantic channels include building footprint/height and terrain/road layers; example radio channels include path gain and quality metrics (SNR/SINR/throughput/BLER). The measurement sequence $\mathcal{X}$ includes categorical identifiers (cell/beam) and multi-layer features (RT/PHY/MAC). Missing or padded events are handled by an attention mask.

7.3 Metrics and ablations

We recommend reporting (i) median/percentile Euclidean error in meters, (ii) NLL (Eq. (19)) to assess uncertainty quality, (iii) calibration curves (optional), and (iv) refinement gain (improvement after Algorithm 1). Key ablations include removing map conditioning, removing physics loss, removing top-K mixture (single-mode regression), and removing cell-conditioned fine embeddings.

8 Experimental Results

We demonstrate the model's ability to leverage both radio propagation physics and semantic map context for probabilistic localization. Figure 3 shows the input modalities: the radio map captures signal propagation patterns (shadowing, reflections), while the semantic map provides building geometry constraints.

Figure 4 visualizes the model's predicted posterior density as a Gaussian mixture. The heatmap reveals multi-modal ambiguity in challenging scenarios (e.g., signal reflections from multiple buildings), demonstrating that the model correctly quantifies localization uncertainty rather than producing overconfident point estimates.

9 Conclusion and Limitations

We presented a physics-informed, map-conditioned Transformer for UE localization that outputs a multi-modal posterior over position, coupled to a differentiable radio-map likelihood. The

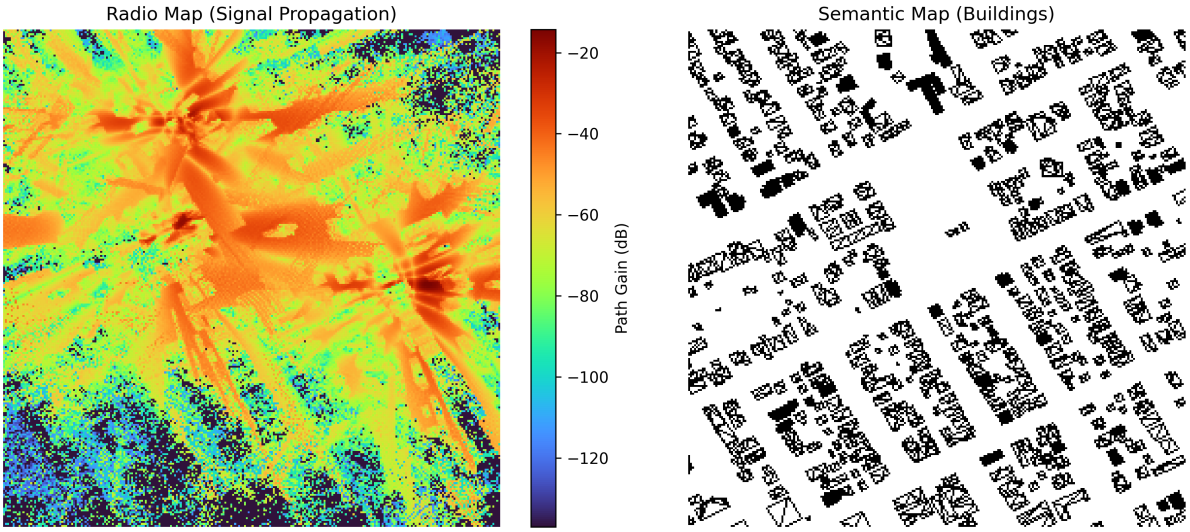

Figure 3: Input modalities for map-conditioned localization: (Left) Radio map showing path gain distribution from ray tracing. (Right) Semantic map with building footprints that constrain feasible UE positions and explain signal occlusions.

model combines amortized inference (fast prediction, uncertainty) with physics-based consistency (regularization and optional refinement). The modular architecture supports both standard Vision Transformer and E(2)-equivariant encoders for map processing, allowing practitioners to trade off computational cost for rotational invariance as needed.

Limitations and future work. Our current formulation relies on precomputed radio maps and therefore inherits simulator bias; bridging the sim-to-real gap (e.g., by learning residual corrections or jointly learning material parameters as in differentiable RT) is an important direction. While we provide an optional E(2)-equivariant map encoder, evaluating its benefits on diverse real-world datasets with varying map orientations remains future work. Additional structured priors and stronger relative positional attention mechanisms are also promising extensions.

A Top-K truncation: approximation discussion

Let $p(\mathbf{y}) = \sum_{c=1}^{G^2} \pi_c \, \varphi_c(\mathbf{y})$ be the full mixture with component densities φ_c and weights π_c. Let $\mathcal{K}$ be the set of top-K indices and $\pi_{\mathcal{K}} = \sum_{c \in \mathcal{K}} \pi_c$. The truncated-renormalized mixture is $p_K(\mathbf{y}) = \sum_{c \in \mathcal{K}} (\pi_c / \pi_{\mathcal{K}}) \varphi_c(\mathbf{y})$. The lost mass is $1 - \pi_{\mathcal{K}}$. If $\pi_{\mathcal{K}} \approx 1$, then p_K is close to p in total variation whenever the omitted components do not carry significant mass near $\mathbf{y}^\star$. Empirically, we find that modest K (e.g., $K = 5$) captures multi-modality while keeping inference efficient.

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[2] 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 *International Conference on Learning Representations (ICLR)*, 2021.

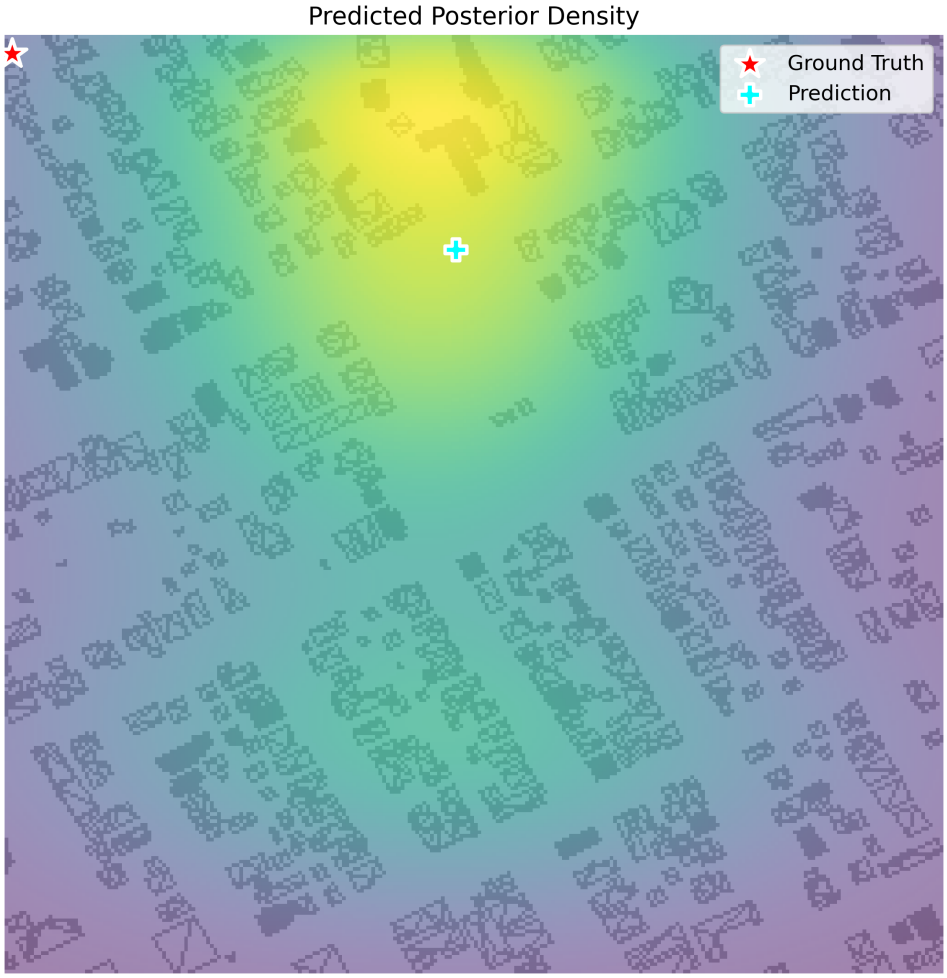

Figure 4: Predicted posterior density visualization: The GMM heatmap (viridis colormap) shows probability mass concentrated in plausible regions consistent with both radio measurements and map geometry. Red star indicates ground truth; cyan marker shows MAP estimate. The multi-modal structure captures inherent ambiguity in the inverse problem.

[3] M. Zaheer, S. Kottur, S. Ravanbakhsh, B. Póczos, R. Salakhutdinov, and A. Smola. Deep sets. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[4] J. Lee, Y. Lee, J. Kim, A. R. Kosiorek, S. Choi, and Y. W. Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *International Conference on Machine Learning (ICML)*, 2019.

[5] C. M. Bishop. Mixture density networks. Technical Report NCRG/94/004, Aston University, 1994.

[6] J. Hoydis, F. Aït Aoudia, S. Cammerer, M. Nimier-David, N. Binder, G. Marcus, and A. Keller. Sionna RT: Differentiable ray tracing for radio propagation modeling. *arXiv:2303.11103*, 2023. Also presented at IEEE Globecom Workshops (GC Wkshps), 2023.

[7] F. Aït Aoudia, J. Hoydis, M. Nimier-David, S. Cammerer, and A. Keller. Sionna RT: Technical report. *arXiv:2504.21719*, 2025. NVIDIA technical report.

[8] X. Han, T. Zheng, T. X. Han, and J. Luo. RayLoc: Wireless indoor localization via fully differentiable ray-tracing. *arXiv:2501.17881*, 2025.

[9] T. Orekondy, P. Kumar, S. Kadambi, H. Ye, J. Soriaga, and A. Behboodi. WiNeRT: Towards neural ray tracing for wireless channel modelling and differentiable simulations. In *International Conference on Learning Representations (ICLR)*, 2023.

[10] R. Levie, Ç. Yapar, G. Kutyniok, and G. Caire. RadioUNet: Fast radio map estimation with convolutional neural networks. *IEEE Transactions on Wireless Communications*, 20(6):4001–4017, 2021.

[11] P. Q. Viet and D. Romero. Spatial transformers for radio map estimation. *arXiv:2411.01211*, 2024.

[12] X. Li, R. Liu, S. Xu, S. G. Razul, and C. Yuen. TransPathNet: A novel two-stage framework for indoor radio map prediction. *arXiv:2501.16023*, 2025.

[13] M. Shahid, K. Das, H. Ushaq, H. Zhang, J. Song, D. Qiao, S. Babu, Y. Guan, Z. Zhu, and A. Ahmed. ReVeal: A physics-informed neural network for high-fidelity radio environment mapping. *arXiv:2502.19646*, 2025.

[14] M. Shahid, K. Das, H. Ushaq, H. Zhang, J. Song, D. Qiao, S. Babu, Y. Guan, Z. Zhu, and A. Ahmad. ReVeal-MT: A physics-informed neural network for multi-transmitter radio environment mapping. *arXiv:2512.04100*, 2025.

[15] M. Xu, et al. E(2)-equivariant vision transformer. *arXiv preprint*, 2023.

Table 1: Model architecture and training hyperparameters. We provide two map encoder options: a standard ViT (default) and an E(2)-equivariant ViT for rotation/reflection invariance. The radio encoder uses a standard Transformer with `[CLS]` pooling for permutation-invariant sequence encoding.

Component	**Specification**	**Intuition / Rationale**
Data & Context		
Map resolution	256×256	Covers $\sim$1 km^2 urban area at $\sim$4 m/px precision.
Coarse grid	32×32	1024 hypotheses; balances classification difficulty with spatial resolution.
Max sequence	$T = 20$	Sufficient to capture movement trends without high memory cost.
Neural Architecture		
Radio Encoder	$d_r = 256$, $L = 4$, $H = 8$ `[CLS]` pooling	**Standard Transformer:** Processes radio token sets with self-attention. `[CLS]` token provides permutation-invariant pooling.
Map Encoder	**Default:** Standard ViT $d_m = 384$, $L = 6$, $H = 6$ `patch_size`=16	**Efficient:** Standard ViT is computationally efficient and performs well when map orientations are consistent.
(Optional)	E2-ViT, `group`=$p4m$ $\lvert G\rvert = 8$	**Equivariant:** E2-ViT (lifting + group attn) ensures rotation/reflection invariance; higher compute cost.
Fusion	$d_f = 256$, $H = 8$ $K_q = 4$ query tokens	**Multi-query:** Multiple learnable queries attend to map tokens, conditioned on radio features.
Positional Heads	Coarse: MLP$\rightarrow G^2$ Fine: 2D pos. enc., GMM	**Hierarchical:** Coarse classification over grid cells; fine head outputs cell-conditioned Gaussian means/variances.
Training		
Mixture size	Top-$K = 5$	Covers primary ambiguous modes (e.g., street canyon reflections) while suppressing noise.
Loss weights	$\lambda_{\text{coarse}} = 0.5$, $\lambda_{\text{fine}} = 1.0$	Fine NLL is primary; coarse ensures correct cell is in top-K.