Attention-based Multi-Anchor Direct-Path Detection for Robust TDoA Localization

Ali Sourchap, Omid Esrafilian, and David Gesbert
Communication Systems Department, EURECOM, Sophia Antipolis, France
Emails: {ali.sourchap, omid.esrafilian, david.gesbert}@eurecom.fr

***Abstract*—Accurate indoor localization with time-difference-of-arrival (TDoA) requires reliable direct-path detection in each anchor's channel impulse response (CIR). In multipath-rich environments, reflected components can dominate the direct path, causing peak detectors to select a biased arrival. Most non-line-of-sight (NLoS) identification and mitigation approaches treat this as a per-anchor problem and operate on scalar range measurements, leaving cross-anchor structure unexploited. We propose a joint processing approach using an attention-based neural network operating on the full multi-anchor CIR matrix. For each anchor, the network predicts a direct-path placement distribution and a visibility score, which directly parameterize a residual-weighted TDoA framework without requiring hard NLoS exclusion. Supervised by channel-derived soft labels, the model captures spatial features rather than memorizing environment-specific geometry. Evaluations on ray-traced indoor scenes demonstrate that the proposed system achieves $91.96\,\%$ direct-path detection accuracy, $95.68\,\%$ LoS/NLoS classification accuracy, and a 2.33 m mean localization error, outperforming all evaluated baselines in mean localization error. Furthermore, the model retains strong localization performance in an entirely unseen environment, whereas fingerprinting and range-based baselines degrade substantially.**

I. Introduction

Wireless localization has attracted ongoing research attention as the demand for accurate positioning continues to grow across diverse environments. Among available methods, time-based positioning estimates the user location from signal propagation delays measured at multiple anchors, using Time-of-Arrival (ToA) or Time Difference of Arrival (TDoA) [1]. ToA requires tight transmitter–receiver synchronization, which is hard to guarantee and introduces clock-offset bias. TDoA avoids this by differencing arrival times across transmitters, canceling the unknown transmit instant, and offering a more practical ranging modality. In both cases, the key parameter is the delay of the direct path, i.e., the shortest propagation path between transmitter and receiver. This delay is obtained from the channel impulse response (CIR), which characterizes the received signal over time. Under line-of-sight (LoS) conditions, the direct path typically corresponds to the first significant peak in the CIR and can be reliably detected with standard first-path estimators [1].

This work was supported by the HUAWEI France Chair on Future Wireless Networks at EURECOM and the NVIDIA Academic Grant Program. The work of David Gesbert was also supported by the 3IA Côte d'Azur Artificial Intelligence Interdisciplinary Project funded by the French ANR.

In some scenarios, such as indoor environments, obstructions introduce non-line-of-sight (NLoS) conditions where the signal reaches the receiver via longer paths, leading to a positive bias in range estimates [2]. Classifying CIR observations as LoS or NLoS, known as *identification methods*, has been studied extensively [1]. These approaches typically extract statistical features from the CIR or received signal and apply classical or learning-based classifiers [1]. Deep learning models have extended this by learning representations directly from raw CIR data [1], [3], [4].

Mitigation methods address how to incorporate measurements with varying levels of uncertainty into localization [1]. In some approaches, NLoS and uncertain links are preferred to be discarded to avoid potential range bias, but this exclusion can remove useful geometric constraints [1], [5]. There are also approaches that try to retain all measurements and address NLoS effects at the localization stage, including bias-correction, robust optimization and filtering techniques that model uncertainty or use the temporal structure [1].

One of the main approaches in this direction is residual weighting, which assigns different weights to anchors in the localization optimization based on their measurement characteristics or estimated uncertainties [1]. Depending on the signal model, these weights can be derived from measurement error statistics, distance-based metrics, or likelihood scores. They can also be computed from channel parameters such as path loss [6]. Moreover, weights can also be derived from heuristic or functional approaches that use measurement behavior and characteristics as a reliability proxy [7].

Most approaches operate on scalar range measurements produced by peak or first-path detectors, attempting to mitigate uncertain estimates at the localization stage. However, accurately identifying the direct path in multipath-rich CIRs remains challenging, especially when a strong reflected component arrives with higher power and the direct path is weakened by partial obstruction. Several studies have addressed this issue by exploring improved first-path detection or waveform-based ranging methods that analyze the full CIR structure to identify the earliest arrival component [1], [8], [9].

Learning-based approaches have emerged as effective solutions for this problem. For example, [9] utilizes convolutional neural networks to identify the direct-path component directly from raw CIR data. A recent work, [5], proposes Neural Network architectures that jointly process the coupled CIRs

of an anchor pair to estimate TDoA, rather than estimating each anchor's range independently. However, each CIR pair is still treated in isolation, without access to observations from other anchors. As a result, cross-anchor correlations and spatial information that could improve direct-path detection are not exploited, either in the direct-path estimation process or measurement weight assignment.

Another research direction bypasses ranging estimation entirely and uses neural networks to map CIR observations directly to position coordinates, commonly known as fingerprinting [10], [11]. While these models can achieve high positioning accuracy, the learned representations are site-specific, tied to the geometry and multipath characteristics of the training environment. As a result, such systems often struggle to generalize to unseen deployments.

In CIR-based localization, Transformer-based models have shown strong performance, particularly in ultra-wideband (UWB) systems that provide high-resolution CIRs [12], as well as in fingerprinting formulations [10], [11]. Attention mechanisms are well suited to this setting, as they capture dependencies across the entire delay axis, allowing the model to better exploit the rich multipath structure of the CIR.

Despite extensive prior work, jointly analyzing the full multi-anchor CIR matrix to exploit cross-anchor relationships for direct-path detection and LoS/NLoS classification remains unaddressed. We tackle this gap by processing the complete multi-anchor CIR measurements together, rather than operating on individual or paired anchor observations [5], [9].

Analyzing all anchor CIRs jointly enables the system to exploit cross-anchor spatial correlations in the received signals, which is particularly important for resolving direct-path ambiguities when strong reflections dominate weakened direct paths. To fully leverage this structure, the system employs an attention-based model, which can capture global dependencies across the entire multi-anchor CIR input, unlike convolutional approaches limited to local receptive fields.

The proposed system is a two-head neural network that takes the full CIR matrix as input and produces two per-anchor outputs: (i) a *direct-path placement distribution*, representing the estimated location of the direct-path component across delay taps, and (ii) a *reliability scalar*, quantifying the likelihood that a distinct direct path is present at the predicted arrival.

These outputs address both key objectives in the literature: the placement distribution enables direct-path-based LoS/NLoS identification, while the reliability scalar provides per-anchor confidence for soft mitigation.

Unlike fingerprinting, the model is trained on channel-level targets rather than position coordinates. This grounds the learned representations in channel-level and cross-anchor features rather than environment-specific geometry, improving generalization to unseen deployments.

The main contributions of this work are:

- An attention-based architecture that jointly processes the full multi-anchor CIR matrix for direct-path detection, producing per-anchor outputs via a Cross-Covariance Image Transformer (XCiT) backbone [13], resulting in linear complexity with respect to the CIR length.
- A residual weighting strategy for TDoA localization that mitigates NLoS effects by weighting anchor contributions based on direct-path detection reliability, aiming to retain NLoS measurements rather than discarding them.
- A supervised training strategy using constructed channel-level labels, grounding the model in cross-anchor geometric and physical signal features to enhance generalization in new environments.

II. System Model

We consider a positioning system comprising M transmitters (e.g., base stations), referred to as anchors, located at known positions $\{\mathbf{a}_m \in \mathbb{R}^3\}_{m=1}^{M}$, and a single target device, with unknown position $\mathbf{p} \in \mathbb{R}^3$, to be localized. For TDoA processing, one anchor is designated as the reference, denoted by $\mathbf{a}_{\text{ref}}$, which requires at least $M \geq 3$ anchors for 2D localization. The system operates over a bandwidth B Hz with subcarrier spacing Δf and N_{sc} subcarriers, yielding a sampling rate $f_s = N_{\text{sc}}\Delta f$ and corresponding time resolution $T_s = 1/f_s$. Given a maximum excess delay $\tau_{\max}$ set by the scene geometry, the CIR is sampled with period T_s, yielding $N_t = \lceil \tau_{\max}/T_s \rceil$ discrete delay taps indexed by $n \in \{1, \ldots, N_t\}$.

The channel between the target and the m-th anchor is modeled as a superposition of L propagation paths, represented in discrete time as:

$$h_m[n] = \alpha_{m,0}\,\delta[n - n_{m,0}] + \sum_{\ell=1}^{L-1} \alpha_{m,\ell}\,\delta[n - n_{m,\ell}], \quad (1)$$

where $\alpha_{m,\ell} \in \mathbb{C}$ is the complex gain of the ℓ-th path, and $n_{m,\ell} \in \{0, 1, \ldots, N_t\}$ is its discrete delay index, obtained by discretizing the physical delay as $n_{m,\ell} = \lfloor \tau_{m,\ell}/T_s \rceil$. The term $\ell = 0$ denotes the direct path, while $\ell \geq 1$ represents NLoS multipath components. The discrete-time CIR vector is then given by $\mathbf{h}_m = [h_m[1], \ldots, h_m[N_t]]^\top \in \mathbb{C}^{N_t}$, and $h_m[n]$ denotes its n-th element. Under LoS conditions, the direct-path component $\alpha_{m,0}$ is the strongest in the CIR and can be reliably detected by standard first-path estimators; in practice, however, this dominance is not guaranteed.

We denote the magnitude CIRs from all M anchors as a matrix $\mathbf{H} = [|\mathbf{h}_m|]_{m=1}^{M} \in \mathbb{R}^{M \times N_t}$. Using a conventional peak detection method [2], [5], the direct-path delay index is estimated as $\widehat{n}_{m,0}$. The corresponding range is then given by $\widehat{r}_m = cT_s\widehat{n}_{m,0}$, where c denotes the speed of light. By selecting a reference anchor, the TDoA measurement for anchor m is then $\widehat{r}_m - \widehat{r}_{\text{ref}}$. Under ideal LoS conditions, and tight synchronization between anchors, the TDoA error can be modeled as an independent zero-mean Gaussian random variable with variance σ_m^2 [7].

Under this assumption, the Maximum Likelihood Estimate of the target position reduces to the Weighted Least Squares (WLS) problem [2]:

$$\widehat{\mathbf{p}} = \arg\min_{\mathbf{p}} \sum_{m \neq \text{ref}} \omega_m \big(\|\mathbf{p} - \mathbf{a}_m\| - \|\mathbf{p} - \mathbf{a}_{\text{ref}}\| - \widehat{r}_m + \widehat{r}_{\text{ref}}\big)^2, \quad (2)$$

where ω_m is the weight associated with anchor m. To achieve the best linear unbiased estimator, the optimal weights are $\omega_m = 1/\sigma_m^2$, i.e., measurements with smaller error variance are assigned larger weights [7], [14].

In practice, however, NLoS propagations introduce positive bias and non-zero-mean ranging errors, which violate the WLS optimality conditions. Moreover, since the true error distributions are generally unknown, the variances σ_m^2 are not available, and the optimal $\omega_m = 1/\sigma_m^2$ cannot be set. A common strategy is therefore to retain the same WLS *formulation* but to employ heuristic weighting functions, where the weights are chosen to be inversely related to the ranging residuals rather than to the exact error variances [1], [6].

Another challenge in TDoA-based localization is selecting the reference anchor. Common heuristics choose the anchor with the strongest received signal, highest SNR, or most reliable LoS characteristics [5]. However, these methods risk selecting an NLoS anchor as the reference, which can bias all TDoA measurements and degrade localization performance.

Section III addresses the challenges of anchor weighting and reference selection by introducing a model that jointly estimates the direct path and its reliability for each anchor's CIR. These range estimates and reliability measures are then integrated into the localization framework.

III. Proposed Solution

Rather than relying on LoS/NLoS classification and conventional peak-based detection, we use a neural network to infer direct-path information directly from the full multi-anchor CIR matrix $\mathbf{H} \in \mathbb{R}^{M \times N_t}$. For each anchor $m = 1, \ldots, M$, the network maps $\mathbf{H}$ to two per-anchor outputs: a *placement distribution* $\widehat{\mathbf{q}}_m \in [0,1]^{N_t}$, representing a soft profile over CIR tap indices that indicates the likely location of the direct path, constrained as a probability distribution $\sum_n \widehat{q}_{m,n} = 1$, and a *visibility score* $\widehat{v}_m \in [0,1]$, capturing the prominence and detectability of the direct-path arrival through the fraction of channel energy concentrated around estimated tap $\widehat{n}_{m,0}$.

This formulation reflects two practical considerations. First, in multipath-rich environments, the direct-path delay cannot always be assigned to a single tap with high confidence; representing it as a distribution over CIR tap indices explicitly captures this uncertainty. Second, the direct path is not equally detectable across anchors; $\widehat{v}_m$ serves as a per-anchor reliability measure, quantifying the prominence of the direct-path arrival and directly indicating the likelihood that a distinct direct-path component is present at the expected delay tap. Together, these outputs provide, for each anchor, an estimate of the direct-path location and its reliability.

To train the neural network, we first construct a carefully designed dataset with per-anchor ground-truth annotations. The model is then trained in a supervised manner to predict the placement distribution $\widehat{\mathbf{q}}_m \in [0,1]^{N_t}$ and the visibility score $\widehat{v}_m \in [0,1]$. In the following, we describe the dataset construction process and the training procedure in detail.

A. Supervised Target Construction

Assuming the direct-path index $n_{m,0}$ is available for all anchors, we explicitly construct two per-anchor supervision targets from it: $\mathbf{q}_m$ and v_m. This approach contrasts with binary LoS/NLoS class labels and instead generates continuous soft labels directly from the CIR structure, framing the task as a supervised regression problem.

Direct-path placement distribution. Since the model outputs a distribution over delay taps, we construct the training targets to mirror this structure. The target $\mathbf{q}_m \in [0,1]^{N_t}$ is defined as a narrow discrete probability distribution over the CIR tap indices, satisfying $\sum_n q_{m,n} = 1$ and centered at the direct-path tap $n_{m,0}$. We generate this target by placing a fixed symmetric template (e.g., a sharp triangular pulse with a support smaller than N_t) at $n_{m,0}$ and zero-padding it to length N_t. This formulation addresses the inherent mismatch between the continuous direct-path delay and the discretely sampled CIR. In practice, the physical signal arrival rarely aligns with the sampling resolution. By using a soft distribution rather than a strict single-index label, we assign probability mass to adjacent taps, which take into account quantization error and encourages stable direct-path estimation.

Visibility score. The scalar variable v_m quantifies how clearly the direct-path component stands out relative to surrounding multipath and how reliably it can be detected given $n_{m,0}$.

We compute this in three steps. First, to suppress noise and emphasize propagation paths, the squared CIR magnitude is smoothed by cross-correlating it with a short symmetric template $g[n]$: $s_m[n] = \big(|\mathbf{h}_m|^2 \star g\big)[n]$. Here, $g[n]$ (e.g., a triangular pulse) acts as a matched filter to enhance structured energy over the noise floor. Next, we extract the set of local peaks $\mathcal{P}_m$ from $s_m[n]$ and compute the energy of each peak $k \in \mathcal{P}_m$ over its prominence support:

$$E_k = \sum_{n=l_k}^{r_k} s_m[n], \tag{3}$$

where l_k, r_k denote the left and right boundaries of the peak.

Finally, given a window of half-width Δ around $n_{m,0}$ and monotonically decaying weights $\{\Gamma_d\}_{d=0}^{\Delta}$ (indexed by the offset $d = |k - n_{m,0}|$), the visibility score is defined as

$$v_m = \frac{\sum_{k \in \mathcal{P}_m,\, |k-n_{m,0}| \le \Delta} \Gamma_{|k-n_{m,0}|} E_k}{\sum_{k \in \mathcal{P}_m} E_k}. \tag{4}$$

The decaying weights penalize energy farther from $n_{m,0}$, so $v_m \to 1$ when the direct-path component is dominant and well-isolated, and $v_m \to 0$ when it is weak or masked by multipath. Together, $\mathbf{q}_m$ and v_m provide complementary supervision, enabling the model to learn direct-path location and reliability. The construction pipeline is shown in Figure 2.

B. TDoA Localization via Placement and Visibility

The predicted outputs $\widehat{\mathbf{q}}_m$ and $\widehat{v}_m$ can directly be used to embody the three components of the WLS formulation: reference anchor selection, per-anchor range estimation, and reliability weighting. The reference anchor is selected as the one with highest predicted visibility score,

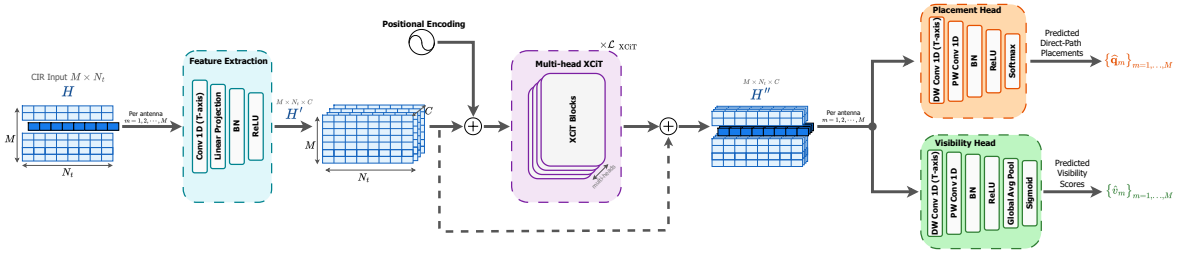

Fig. 1. The proposed Transformer-based neural network architecture. The model processes the full multi-anchor CIR matrix $\mathbf{H}$ jointly, with dual output heads producing visibility scores $\widehat{v}_m$ and placement distributions $\widehat{\mathbf{q}}_m$.

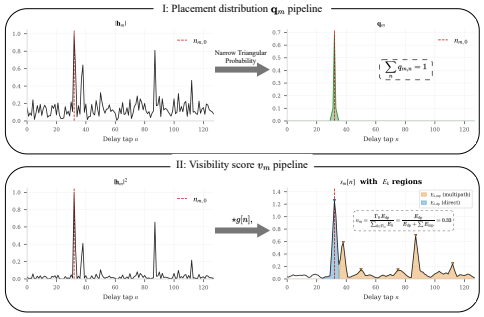

Fig. 2. Per-anchor supervision targets construction: placement distribution $\mathbf{q}_m$ centered at direct-path index $n_{m,0}$ (top), and visibility score v_m derived via cross-correlation (4) (bottom).

ref $= \arg\max_{m\in\{1,\ldots,M\}} \widehat{v}_m$. This criterion naturally favors anchors whose direct-path component is strongly dominant, which in practice corresponds to reliable LoS links and avoids ranging bias in TDoA. Subsequently, for each anchor m, we formulate the estimated range as a direct function of $\widehat{\mathbf{q}}_m$. By extracting the mode of the placement distribution to serve as the direct-path delay index, we obtain:

$$\widehat{r}_m = c\,T_s\,\widehat{n}_{m,0}, \quad \widehat{n}_{m,0} \triangleq \arg\max_{n\in\{1,\ldots,N_t\}} \widehat{q}_{m,n}. \tag{5}$$

Finally, to account for the reliability of the estimated ranges in the WLS formulation (2), we define the weights directly as $\omega_m = \widehat{v}_m$. The motivation is that $\widehat{v}_m$ encodes the reliability of the estimated direct-path arrival for anchor m, making it a natural per-anchor confidence measure: anchors with higher visibility scores have greater influence on the position estimate, while those with lower visibility may have experienced more obstruction or attenuation and are thus automatically down-weighted. With these assignments, the WLS position estimate becomes:

$$\widehat{\mathbf{p}} = \arg\min_{\mathbf{p}} \sum_{m\neq\text{ref}} \widehat{v}_m \big(\|\mathbf{p}-\mathbf{a}_m\| - \|\mathbf{p}-\mathbf{a}_{\text{ref}}\| - \widehat{r}_m + \widehat{r}_{\text{ref}}\big)^2. \tag{6}$$

C. Proposed Neural Network Architecture

To obtain the visibility scores and placement distributions, we propose a multi-task architecture (illustrated in Figure 1) that processes $\mathbf{H}$ as input and jointly outputs $\{\widehat{v}_m\}_{m=1}^M$ and $\{\widehat{\mathbf{q}}_m\}_{m=1}^M$. Since these two estimates are correlated, unifying them within a single multi-task system enables the shared backbone to extract richer intra-anchor and cross-anchor features than two separate models. The neural network processes data through three main stages: feature extraction, cross-anchor attention, and dual output heads. Moreover, a residual connection [15] is employed to preserve per-anchor indexing and feature dimensionality through to the output heads.

Feature extraction maps $\mathbf{H} \in \mathbb{R}^{M\times N_t}$ to a higher-dimensional representation $\mathbf{H}' \in \mathbb{R}^{M\times N_t\times C}$. For each anchor m, the CIR magnitude $|\mathbf{h}_m|$ is normalized by its maximum value to yield the scale-invariant sequence $\bar{\mathbf{h}}_m$, which is then processed independently using a shared 1D convolution with C_{stem} filters. The resulting features are projected to C channels via a linear transformation, followed by batch normalization and ReLU, yielding per-anchor feature sequences in $\mathbb{R}^{N_t\times C}$.

For the cross-anchor attention stage, we employ a multi-head XCiT architecture [13], which computes cross-covariance attention along the channel dimension rather than across token positions. Treating $\mathbf{H}'$ as a sequence of $N = M\cdot N_t$ tokens, standard self-attention scales as $\mathcal{O}(N^2)$, which is prohibitive for large M or N_t. XCiT addresses this by using a $C\times C$ cross-covariance matrix, reducing complexity to $\mathcal{O}(NC^2)$; since $C \ll N$, this yields linear scaling with N. We stack $\mathcal{L}_{\text{XCiT}}$ such attention layers to enhance feature extraction. Sinusoidal positional encodings [16] along the anchor and delay dimensions are summed into a composite encoding and added to $\mathbf{H}'$ before the XCiT blocks. The output is then combined with $\mathbf{H}'$ via a residual connection, producing the enriched tensor $\mathbf{H}'' \in \mathbb{R}^{M\times N_t\times C}$ with cross-anchor context.

From $\mathbf{H}''$, two independent output heads predict $\widehat{\mathbf{q}}_m$ and $\widehat{v}_m$ for each anchor. Each head applies its own lightweight depthwise-separable 1D convolution [17] (a per-channel depthwise filter followed by a pointwise projection), keeping the parameter count low so that the XCiT backbone carries the primary computational burden.

Visibility Estimator: The per-anchor features are compressed via global average pooling, then passed through batch normalization, ReLU, dropout, and a linear–sigmoid projection to produce $\widehat{v}_m \in [0,1]$.

Placement Estimator: Unlike the visibility head, the placement estimator retains full temporal resolution: per-tap features are passed through batch normalization, ReLU, dropout, and a linear projection, and softmax over the delay axis yields the distribution $\widehat{\mathbf{q}}_m \in [0,1]^{N_t}$.

Unlike standard convolutional approaches, the attention-based architecture captures long-range dependencies across both the delay axis and anchor dimension, promoting gener-

alization to new environments. Candidate direct-path taps are evaluated using all anchor observations, enabling the model to leverage geometric consistency while learning features related to signal morphology and multipath occlusion. In new scenes or changed layouts, decisions remain grounded in cross-anchor relationships and per-anchor statistics captured by attention across delay and anchor dimensions.

At training, we minimize a joint loss defined over the predicted $\widehat{v}_m$ and $\widehat{\mathbf{q}}_m$. For the visibility head, we use the Symmetric Absolute Relative Error (SARE), $|(\widehat{v}_m - v_m)|/(|\widehat{v}_m| + |v_m|)$, which is bounded and scale-invariant, and for the placement head, we use the Total Variation Distance (TVD), $\sum_n |\widehat{q}_{m,n} - q_{m,n}|$, measuring the discrepancy between probability distributions over delay taps. The overall objective combines these two terms with a weighting parameter $\lambda \in [0, 1]$:

$$\mathcal{L}_{\text{total}} = \sum_{m=1}^{M} \left(\lambda \frac{|\widehat{v}_m - v_m|}{|\widehat{v}_m| + |v_m|} + (1-\lambda) \sum_{n=1}^{N_t} |\widehat{q}_{m,n} - q_{m,n}| \right). \quad (7)$$

IV. Experiments

To construct supervision targets for training, we generate multi-anchor CIRs using the Sionna ray-tracing simulator [18]. We model indoor environments of 72×130 m with 10 m-high walls. Eight training scenes share the same outer boundaries but differ in internal layout and obstacle materials, producing diverse multipath conditions. The system uses a 100 MHz bandwidth, a 122.88 MHz sampling rate, and $N_t = 128$ delay taps.

In each scene, $M = 4$ anchors are placed, one per wall at random coordinates. User locations are sampled at 1.5 m height on a grid with spacing $\delta_d = 0.4$ m; points with fewer than two LoS anchors are discarded. The collected CIRs are split $80/20$ for training and validation. Figure 3 shows a 3D view of a simulated indoor environment, illustrating the scene layout, random anchor placement, and their orientations toward the center. A ninth scene, unseen during training, is reserved exclusively for generalization testing.

We augment the simulated CIRs to narrow the gap to real-world conditions. Complex Gaussian noise is added across all taps, followed by log-normal power fluctuations to emulate small-scale fading [19]. We also attenuate the strongest multipath tap or the direct path at random, simulating shadowing and dropout [2]. These perturbations reduce overfitting and spatial leakage, forcing the network to learn robust features.

The supervision targets $\mathbf{q}_m$ and v_m are constructed as described in Section III, with a triangular template of width 5 taps for the placement distribution and visibility parameters $\Delta = 3$, $\Gamma = [1, 0.6, 0.3, 0.1]$. The network uses $C = 32$ channels, $\mathcal{L}_{\text{XCiT}} = 5$ XCiT layers, 8 attention heads, and $\lambda = 0.5$. Our model contains ~ 70k learnable parameters.

To evaluate our proposed method, we compare it against four baselines: (1) a *VAE classifier* [3] and (2) a *CNN-LSTM* [4], both supervised binary LoS/NLoS classifiers that perform TDoA localization using anchors classified as LoS (requiring at least three); (3) *ANN-TDoA* [5], which clusters anchors into LoS/NLoS in an unsupervised manner to define ranging weights and estimates TDoA for each anchor pair relative to the reference using a learning-based model; and (4) *CIR-TF* [10], an attention-based fingerprinting method which maps $\mathbf{H}$ directly to coordinates. For this experiment, this network is configured with approximately 819k parameters due to the substantial scaling of vanilla transformers [16], and exhibits clear overfitting to the test dataset. All models are trained for 100 epochs.

TABLE I
LoS/NLoS and Direct-Path Detection Accuracy.

Model	Seen Scenes (%)		Unseen Scene (%)	
	LoS / NLoS	Direct-Path	LoS / NLoS	Direct-Path
Our model	**95.68**	**91.96**	87.50	**85.47**
VAE classifier	79.18	N/A	77.16	N/A
CNN-LSTM	92.44	N/A	**88.98**	N/A
ANN-TDoA (Unsupervised)	58.73	N/A	56.69	N/A

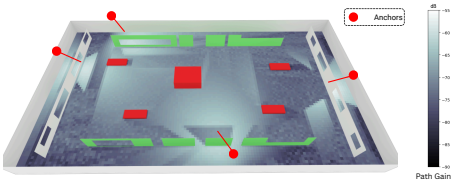

Fig. 3. Top-down view of a simulated scene showing internal layout, anchor placement, and the path-gain map at 1.5 m height (the maximum gain across all anchors).

We first evaluate the model's ability to locate the direct path and, as a by-product, classify LoS conditions. An anchor is inferred to be LoS if the estimated tap $\widehat{n}_{m,0} = \arg\max_n \widehat{q}_{m,n}$ matches the maximum of $|\mathbf{h}_m|$. Table I reports direct-path detection and LoS classification accuracy compared to the VAE, CNN-LSTM, and unsupervised method used at ANN-TDoA, which output binary LoS/NLoS labels.

The unsupervised ANN-TDoA performs substantially worse, as LoS and NLoS CIRs are not well separated in its feature space. While our approach leverages channel-level supervision, ANN-TDoA relies on unsupervised clustering, underscoring the performance gains from explicit direct-path supervision, even though ANN-TDoA has access to nominal pairwise TDoA labels during its own supervised training. On seen scenes, our model achieves 95.68% LoS accuracy and outperforms all baselines. On the unseen scene, it maintains 87.50% LoS accuracy and 85.47% direct-path detection. In both settings, the LoS classification is competitive with the CNN-LSTM, while additionally providing direct-path localization, which the classifiers do not.

To assess how direct-path detection translates to positioning, we evaluate localization error on samples where each model identifies at least three usable anchors (LoS for the classifiers, direct-path for ours; ANN-TDoA and CIR-TF use all samples). We report results on the full dataset and two subsets: *direct-path localizable* (a direct path exists at three or more anchors, regardless of whether it is the strongest tap) and *LoS-localizable* (at least three true LoS anchors).

Table II reports the mean (μ) and 95$^{\text{th}}$-percentile (P95) localization errors. On the seen scenes, our model achieves a 2.33 m mean error, slightly outperforming CIR-TF (2.75 m),

TABLE II
LOCALIZATION ERROR STATISTICS: MEAN (μ) AND 95$^{\text{TH}}$ PERCENTILE (P95) IN METERS

Model	Seen Scenes (Validation)							Unseen Scene (Test Only)						
	% Local.*	All Samples		Direct-path Loc.		LoS Loc.		% Local.*	All Samples		Direct-path Loc.		LoS Loc.	
	%	μ	P95	μ	P95	μ	P95	%	μ	P95	μ	P95	μ	P95
Our model	87.51	**2.33**	8.75	**2.16**	7.45	**1.41**	**3.52**	90.35	**4.49**	**20.79**	**4.34**	**20.16**	**2.99**	**12.88**
VAE classifier	65.93	16.35	61.82	15.17	59.20	10.55	48.61	73.19	10.06	42.03	9.91	41.19	5.99	31.38
CNN-LSTM	60.98	10.08	49.18	9.58	48.24	8.33	45.20	72.73	7.69	36.71	7.47	36.00	5.56	30.71
ANN-TDoA	100.00	12.98	36.21	12.73	35.97	9.80	28.59	100.00	13.15	44.54	13.00	44.40	10.29	36.84
CIR-TF	100.00	2.75	**6.77**	2.62	**6.23**	2.06	4.78	100.00	15.61	41.28	15.54	40.52	14.53	35.28

*Percent of localizable samples. For supervised classifiers and our method, error statistics include only samples meeting three-anchor criterion.

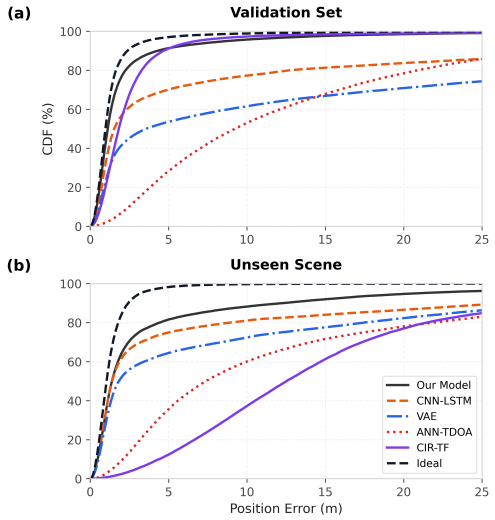

Fig. 4. Cumulative Distribution Function (CDF) of the localization error on (a) the validation set, and (b) the unseen scene.

which attains a tighter P95 (6.77 m vs. 8.75 m) but requires over 11× more parameters (819k vs. 70k). Both methods comfortably surpass ANN-TDoA and the LoS/NLoS classification pipelines. The CDF in Figure 4(a) confirms these bounds. By relying on direct-path detection rather than strict LoS classification, our model also localizes roughly 25% more samples than the VAE and CNN-LSTM.

The generalization advantage of our approach is clearest on the unseen scene. CIR-TF collapses to 15.61 m mean error and 41.28 m P95 (Table II, Figure 4(b)), while our model retains 4.49 m mean error, substantially outperforming the best baseline (CNN-LSTM at 7.69 m). It continues to localize about 27% more samples than the VAE and CNN-LSTM. These results confirm that jointly estimating per-anchor direct-path placement and visibility from the full CIR matrix yields consistent gains in both accuracy and environmental reliability over per-anchor and coordinate-supervised alternatives.

V. CONCLUSION

We presented a joint attention-based framework that extracts per-anchor direct-path placement and visibility from the full multi-anchor CIR matrix. By capturing cross-anchor spatial features and feeding soft-label predictions into a residual-weighted TDoA solver, the method enables robust localization without hard NLoS exclusion. Experiments show our approach improves both direct-path detection and localization accuracy over existing baselines, and the system maintains strong performance in an unseen environment where fingerprinting and range-based methods degrade substantially.

The current evaluation relies on ray-traced channels; validation on hardware-measured CIRs is needed to confirm real-world applicability. Future work will prioritize training on measured data and extending the framework to address practical challenges such as unsynchronized anchors and possible intermittent measurement misses due to anchor outages.

REFERENCES

[1] R. E. Nkrow *et al.*, "NLOS Identification and Mitigation for Time-based Indoor Localization Systems: Survey and Future Research Directions," *ACM Computing Surveys*, Oct. 2024.

[2] İ. Güvenç and C.-C. Chong, "A Survey on ToA Based Wireless Localization and NLOS Mitigation Techniques," *IEEE Communications Surveys & Tutorials*, 2009.

[3] M. Stahlke *et al.*, "Estimating ToA Reliability With Variational Autoencoders," *IEEE Sensors Journal*, 2022.

[4] C. Jiang *et al.*, "UWB NLOS/LOS Classification Using Deep Learning Method," *IEEE Communications Letters*, 2020.

[5] A. Kirmaz *et al.*, "Accurate time-based positioning using deep learning with anchor selection," *IEEE Access*, 2025.

[6] T. Yang *et al.*, "A survey of recent indoor localization scenarios and methodologies," *Sensors*, 2021.

[7] P. Wu *et al.*, "Time difference of arrival (tdoa) localization combining weighted least squares and firefly algorithm," *Sensors*, 2019.

[8] A. G. Ferreira *et al.*, "Feature selection for real-time NLOS identification and mitigation for body-mounted UWB transceivers," *IEEE Transactions on Instrumentation and Measurement*, 2021.

[9] T. Feigl *et al.*, "Robust toa-estimation using convolutional neural networks on randomized channel models," in *2021 International Conference on Indoor Positioning and Indoor Navigation (IPIN)*, 2021.

[10] D. Coppens *et al.*, "Beyond Convolutions: Transformer Networks for Improved UWB CIR-based Fingerprinting," 2024.

[11] W. Li *et al.*, "LoT: A Transformer-Based Approach Based on Channel State Information for Indoor Localization," *IEEE Sensors Journal*, Nov. 2023.

[12] J. Ott *et al.*, "Estimating Multipath Component Delays With Transformer Models," *IEEE Journal of Indoor and Seamless Positioning and Navigation*, 2024.

[13] A. El-Nouby *et al.*, "XCiT: Cross-Covariance Image Transformers," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[14] P. Tarrío *et al.*, "Weighted least squares techniques for improved received signal strength based localization," *Sensors*, 2011.

[15] K. He *et al.*, "Deep Residual Learning for Image Recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[16] A. Vaswani *et al.*, "Attention Is All You Need," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[17] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," 2017.

[18] J. Hoydis *et al.*, "Sionna rt: Differentiable ray tracing for radio propagation modeling," 2023.

[19] A. F. Molisch *et al.*, "A Comprehensive Standardized Model for Ultrawideband Propagation Channels," *IEEE Transactions on Antennas and Propagation*, 2006.