DRL-based Position-Aided Beam-Management in RIS-Enabled Indoor Environments

Iacovos Ioannou[‡†], Marios Raspopoulos[§], Michael Georgiades[¶],
Christophoros Christophorou[†§], Ala' Khalifeh [‖], Vasos Vassiliou[*†]
[‡] Department of Computer Science, Philips University,
[†] CYENS - Centre of Excellence, Cyprus
[*] Department of Computer Science, University of Cyprus
[§] UCLAN Cyprus,
[¶]Department of Computer Science, Neapolis University & Infostrada Communications, Cyprus,
[‖]German Jordanian University, Amman, Jordan

***Abstract*—Reconfigurable Intelligent Surfaces (RIS) have emerged as a transformative technology for next-generation wireless networks, enabling programmable control of electromagnetic wave propagation through arrays of passive reflecting elements. Beyond their well-established role in enhancing communication capacity and coverage, RIS panels offer exceptional opportunities for high-precision indoor positioning by creating additional controllable propagation paths that can be exploited for geometric localization. This paper presents a deep reinforcement learning (DRL) framework for joint position estimation and adaptive beam management in multi-RIS indoor environments that leverages time-division RIS operation combined with learned signal processing to achieve accurate user equipment (UE) localization in challenging non-line-of-sight (NLOS) conditions. The proposed architecture comprises three synergistic components: i) a deep neural network (DNN) position estimator that maps time-difference-of-arrival and received signal strength features directly to UE coordinates, implicitly learning to handle NLOS bias and clock synchronization errors from augmented training data; ii) a Deep Q-Network (DQN) that learns an adaptive RIS anchor selection policy to identify and exclude NLOS-corrupted measurements; and iii) a hybrid geometric refinement stage where the DNN-predicted position initializes Levenberg-Marquardt optimization on the DQN-selected anchor subset. The framework is evaluated through extensive Monte Carlo simulations in a representative 60×40 m indoor environment with four strategically deployed 256-element RIS panels operating at 3.5 GHz with 100 MHz bandwidth. Comparative analysis against joint Levenberg-Marquardt estimation, standard TDoA processing, robust TDoA with iteratively reweighted least squares, and K-Nearest Neighbors fingerprinting demonstrates that the proposed DQN-DNN-LM pipeline achieves a median positioning accuracy of 1.41 m, representing a 27.4% improvement over the best geometric baseline while providing increasingly significant advantages under high NLOS contamination. Comprehensive parametric studies examining the impact of signal-to-noise ratio, RIS array size, and NLOS probability reveal the fundamental scaling behavior and practical design trade-offs governing DRL-aided RIS positioning systems.**

***Index Terms*—Reconfigurable Intelligent Surfaces (RIS), indoor positioning, deep reinforcement learning, Deep Q-Network, deep neural network, Time-Difference-of-Arrival, 5G New Radio, beam management.**

I. Introduction

Indoor positioning remains one of the most challenging requirements for next-generation wireless networks, where Global Navigation Satellite System (GNSS) signals are severely attenuated or entirely unavailable, and traditional positioning methods face significant obstacles from multipath propagation, signal attenuation, and non-line-of-sight (NLOS) conditions [1]. With the proliferation of location-aware services across commercial, industrial, healthcare, and emergency response environments, achieving sub-meter positioning accuracy has become a key enabler for 5G and beyond systems [2]. Applications ranging from asset tracking in warehouses, autonomous navigation of mobile robots, augmented reality guidance, and emergency personnel localization all demand reliable indoor positioning with accuracy that conventional Wi-Fi or Bluetooth-based systems struggle to consistently provide.

Reconfigurable Intelligent Surfaces (RIS) have recently gained transformative attention as a means of intelligently controlling the wireless propagation environment through programmable electromagnetic wave manipulation [3]. A RIS consists of a large number of low-cost, passive reflecting elements, each capable of introducing a controllable phase shift to impinging electromagnetic waves, arranged in a planar array structure. Unlike active relay nodes, RIS panels require no radio-frequency (RF) chains, power amplifiers, or analog-to-digital converters, resulting in dramatically reduced hardware complexity, power consumption, and deployment cost. While the majority of RIS research has focused on enhancing communication capacity, coverage extension, and interference management [4], the potential of RIS for high-accuracy positioning has only recently begun to be systematically explored [5], [6].

The fundamental value proposition of RIS for positioning stems from the creation of additional controllable propagation paths that introduce geometric diversity essential for trilateration and multilateration [7]. In conventional anchor-based positioning, each base station or access point serves as a geometric reference whose position is known a priori. Deploying sufficient anchors with favorable spatial distribution is often prohibitively expensive in terms of infrastructure, power, and backhaul connectivity. RIS panels address this challenge by serving as *virtual anchors* that redirect existing signals along precisely controlled paths without requiring active transmis-

sion capability. By strategically deploying multiple RIS panels on the walls of an indoor environment and operating them in a time-division multiplexed fashion, the positioning system can obtain unambiguous range measurements through each RIS-reflected path, exploiting the known bistatic geometry of the UE-RIS-gNB signal configuration.

However, practical RIS-aided positioning faces several critical challenges that must be addressed for reliable operation. The indoor multipath environment causes time-of-arrival (ToA) estimates to be corrupted by secondary reflections that bias the measured propagation delay. NLOS conditions, in which the direct UE-RIS path is obstructed by furniture, walls, or human bodies, introduce positive range bias that can severely degrade position estimation accuracy if not properly handled. The unknown clock offset between the UE transmitter and the gNB receiver creates a common bias in all range measurements that must be either estimated jointly or eliminated through differential processing. Additionally, the geometric dilution of precision (GDOP) varies spatially across the coverage area, causing position-dependent accuracy variations that affect system reliability. While existing geometric approaches such as robust least squares with Huber loss functions and subset selection [8] provide principled outlier mitigation, they suffer from sensitivity to local minima in the non-convex cost function, require careful initialization heuristics, and, most importantly, cannot learn from accumulated positioning experience to improve their NLOS handling over time.

Deep reinforcement learning (DRL) has recently demonstrated remarkable success in sequential decision-making problems across wireless communications, including resource allocation, power control, and beam management [9]. The ability of DRL agents to learn optimal policies through interaction with complex environments without requiring explicit mathematical models of the underlying dynamics makes them natural candidates for the RIS anchor selection problem, where the agent must learn to identify and exclude NLOS-corrupted measurements based on observable signal features. Furthermore, deep neural networks (DNNs) have shown considerable promise for direct position estimation from wireless measurements [10], learning the nonlinear mapping between measurement features and spatial coordinates in a data-driven manner that implicitly accounts for propagation effects that are difficult to model analytically. However, the joint application of DRL-based anchor selection and DNN-based position estimation to RIS-aided indoor positioning remains largely unexplored.

This paper makes the following specific contributions:

1) A deep neural network position estimator trained on 30,000 augmented samples spanning diverse SNR and NLOS conditions, incorporating geometric feature priors derived from TDoA measurements, power-weighted centroids, and pairwise range differences, that directly maps measurement features to UE coordinates while implicitly learning to mitigate NLOS bias and clock synchronization errors.
2) A Deep Q-Network for adaptive RIS anchor selection that learns, through interaction with the positioning environment over 5,000 training episodes with experience replay, which anchor subset minimizes positioning error for each measurement scenario, effectively learning to identify and exclude NLOS-corrupted anchors without explicit NLOS detection or channel characterization.
3) A hybrid DQN-DNN-LM positioning pipeline that combines learned anchor selection, neural position estimation, and geometric Levenberg-Marquardt refinement into a three-stage closed-loop system, achieving 27.4% improvement over the best geometric baseline under practical NLOS conditions.
4) Extensive Monte Carlo simulation analysis comparing seven positioning algorithms across varying SNR conditions, RIS array configurations, and NLOS contamination levels, establishing fundamental performance bounds, identifying the NLOS crossover threshold above which the DRL approach dominates geometric methods, and providing practical design guidelines.

The remainder of this paper is organized as follows. Section II describes the system architecture, signal model, and measurement framework. Section III presents the proposed DRL-based methodology in detail, including dataset generation, feature engineering, network architectures, and hyperparameter optimization. Section IV summarizes the experimental setup, the compared baselines, and the simulation results with in-depth analysis. Section V concludes the paper with key findings and future research directions.

II. System Model

A. Architecture, Signal Model, and RIS Operation

The proposed positioning system operates in an indoor environment measuring 60×40 m, representative of large commercial facilities such as shopping malls, warehouses, exhibition halls, or factory floors where GPS signals are unavailable and traditional positioning methods face significant challenges due to rich multipath propagation. The system deploys four RIS panels strategically positioned on the room walls to maximize spatial diversity and geometric coverage. Specifically, RIS-1 is mounted at $(0, 20)$ m on the left wall midpoint, RIS-2 at $(60, 20)$ m on the right wall midpoint, RIS-3 at $(30, 0)$ m at the bottom wall midpoint, and RIS-4 at $(30, 40)$ m at the top wall midpoint. This symmetric four-anchor configuration ensures that user equipment (UE) anywhere within the coverage area maintains favorable geometric diversity with respect to at least three RIS panels simultaneously, providing the angular spread essential for accurate two-dimensional position determination.

A single next-generation NodeB (gNB) base station equipped with an $N_{\text{rx}} = 8$ element uniform linear receive antenna array is positioned near the room corner at coordinates $(2, 2)$ m. The gNB serves as the central measurement collection and processing point, receiving uplink pilot signals that have been reflected and phase-shifted by the RIS panels. Each RIS panel comprises $N_{\text{total}} = 256$ programmable reflecting

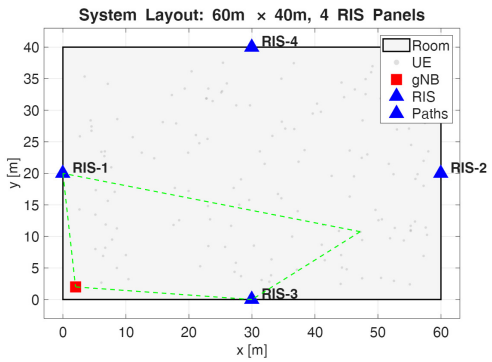


Fig. 1: System layout showing the 60×40 m indoor environment with four symmetrically deployed RIS panels on the walls, gNB receiver at position $(2, 2)$ m, sample random UE test positions (gray dots), and example bistatic signal paths (dashed green arrows) through RIS-1 and RIS-3 demonstrating the UE→RIS→gNB measurement geometry.

elements arranged in a 16×16 planar array. To maintain computational feasibility for real-time phase optimization, the effective number of simultaneously controllable elements is limited to $N = 64$ through element grouping, where each group of four adjacent elements shares a common phase configuration. The complete system layout, including representative UE test positions and example bistatic signal paths, is illustrated in Fig. 1.

The system operates at carrier frequency $f_c = 3.5$ GHz with signal bandwidth $B = 100$ MHz, corresponding to 273 resource blocks with 30 kHz subcarrier spacing as specified in the 5G NR numerology [11]. The fundamental operating principle exploits the bistatic geometry formed by the UE-RIS-gNB signal path, where the total propagation distance through the k-th RIS panel is given by:

$$r_k = d_{\mathrm{UE},k} + d_{k,\mathrm{gNB}} = \|\boldsymbol{p}_{\mathrm{UE}} - \boldsymbol{p}_k\| + \|\boldsymbol{p}_k - \boldsymbol{p}_{\mathrm{gNB}}\| \quad (1)$$

where $\boldsymbol{p}_{\mathrm{UE}} = [x, y]^T$ denotes the unknown UE position, $\boldsymbol{p}_k$ is the known position of the k-th RIS panel, and $\boldsymbol{p}_{\mathrm{gNB}}$ is the known gNB position. The cascaded UE-RIS-gNB channel is modeled using the 3GPP CDL-D profile [12] with Rician K-factor of 13.3 dB and RMS delay spread $\sigma_\tau = 30$ ns.

The system implements coherent RIS phase alignment where the optimal phase for the m-th element is $\phi_m^\star = -\angle(h_m^{\text{UE-RIS}}) - \angle(h_m^{\text{RIS-gNB}})$, yielding coherent combining gain that scales as N^2 in received power. Time-division multiplexing separates measurements from different RIS panels, with only one panel active per measurement slot, producing a complete set of four bistatic range measurements $\{\hat{r}_k\}_{k=1}^4$ and four received signal power measurements $\{P_k\}_{k=1}^4$.

B. Measurement Model with Impairments

The measured bistatic range through the k-th RIS path incorporates several realistic impairments:

$$\hat{r}_k = r_k + b + n_k + \beta_k \quad (2)$$

where b represents an unknown common clock bias (± 5 m equivalent range), $n_k \sim \mathcal{N}(0, \sigma_r^2)$ is additive Gaussian ranging noise, and β_k is a non-negative NLOS bias term. The ranging noise standard deviation follows the Cramér-Rao lower bound (CRLB) for time-delay estimation [13]:

$$\sigma_r = \frac{c}{B\sqrt{2 \cdot \mathrm{SNR}}} \quad (3)$$

where c is the speed of light. At the baseline configuration of $B = 100$ MHz and SNR$= 30$ dB, this yields $\sigma_r \approx 0.067$ m, indicating that the wideband 5G NR waveform provides inherently high ranging precision under favorable conditions. The NLOS bias is modeled as an exponentially distributed positive random variable ($\beta_k \sim \mathrm{Exp}(\mu_\beta)$ with mean $\mu_\beta = 4$ m) that occurs independently on each RIS panel with probability $p_{\mathrm{NLOS}} = 0.25$.

III. Proposed Methodology

This section presents the complete proposed DRL-based positioning methodology, encompassing the training dataset construction, feature engineering pipeline, deep neural network and Deep Q-Network architectures, hyperparameter optimization strategy, and the integrated three-stage positioning algorithm.

A. Dataset Generation and Augmentation Strategy

A fundamental challenge in training neural networks for positioning is ensuring that the learned model generalizes across the diverse operating conditions encountered in practical indoor deployments, where both the signal-to-noise ratio and the NLOS contamination level vary significantly depending on the UE location, the time of day, and the presence of moving obstacles. To address this challenge, the proposed framework employs a systematic data augmentation strategy that generates 30,000 training samples spanning a structured grid of environmental conditions.

The training data is generated by uniformly sampling UE positions across the 60×40 m room interior (with 2 m margin from walls) and computing the corresponding bistatic range measurements according to the analytical model in Eq. (2). Rather than generating all samples at a single operating point, the augmentation strategy distributes samples across five SNR levels (20, 25, 30, 35, and 40 dB) and five NLOS probability levels (0, 0.1, 0.2, 0.3, and 0.4), producing 1,200 samples per condition and ensuring balanced representation of both clean and corrupted measurements. This structured augmentation exposes the DNN to the full range of noise levels and NLOS contamination patterns that it will encounter during deployment, preventing overfitting to any single operating configuration. A 10% hold-out validation set is reserved for early stopping and model selection.

For DQN training, a complementary strategy generates episodes under continuously randomized conditions, with the SNR uniformly sampled from the interval $[20, 40]$ dB and the NLOS probability uniformly sampled from $[0, 0.4]$ at the start of each episode. This continuous randomization, as opposed to the structured grid used for DNN training, ensures that the

DQN encounters a smooth continuum of difficulty levels and develops a robust anchor selection policy that transfers to any combination of SNR and NLOS within the training range.

B. Feature Engineering

The design of the input feature vector is critical to the DNN's positioning accuracy, as the features must simultaneously eliminate the unknown clock bias, preserve geometric position information, and provide implicit indicators of NLOS contamination. The proposed 30-dimensional feature vector incorporates six categories of carefully engineered features.

The first category consists of time-difference-of-arrival features computed by selecting the RIS panel with the strongest received power as the reference anchor and forming the four range differences $\Delta r_k = \hat{r}_k - \hat{r}_{k^\star}$. Because the unknown clock bias b appears identically in all range measurements, the differencing operation cancels it exactly, providing the DNN with clock-bias-free inputs that encode the relative distances between the UE and the four RIS panels. The second category comprises normalized power features $P_k - \bar{P}$ and relative power features $P_k - \max_j P_j$, which encode distance-dependent signal attenuation information and highlight which anchors receive the weakest signals.

The third category introduces geometric prior features that exploit the known and fixed RIS-to-gNB distances $d_{k,\mathrm{gNB}}$ to convert the TDoA measurements into estimated UE-to-RIS distance differences via $\hat{d}_k - \hat{d}_{\mathrm{ref}} = \Delta r_k - (d_{k,\mathrm{gNB}} - d_{\mathrm{ref,gNB}})$, providing the network with distance-domain information that relates more directly to position than raw range differences. The fourth category consists of all six pairwise range differences $\hat{r}_i - \hat{r}_j$ for $i < j$, which provide redundant geometric constraints that help the DNN resolve positional ambiguities. The fifth category is a power-weighted centroid direction vector computed as $\boldsymbol{c} = \sum_k w_k \boldsymbol{p}_k / \sum_k w_k$ where $w_k = 10^{P_k/10}$, which provides an approximate bearing estimate toward the UE based on signal strength alone. The sixth and final category comprises consistency features, specifically the variance and range of the estimated distance differences, which serve as implicit NLOS indicators since a single corrupted measurement will cause the distance estimates to become inconsistent, producing elevated variance.

C. DNN Position Estimator

The deep neural network position estimator directly learns the mapping from the 30-dimensional feature vector to the UE position coordinates $\hat{\boldsymbol{p}} = [\hat{x}, \hat{y}]^T$ through supervised training. The architecture is a four-layer fully connected feedforward network with layer dimensions $30 \to 256 \to 128 \to 64 \to 2$, employing LeakyReLU activation functions with negative slope $\alpha = 0.01$ in all hidden layers and a linear activation in the output layer. The output coordinates are normalized to the range $[0, 1]$ by dividing by the room dimensions, ensuring that both position components contribute equally to the training loss regardless of the room aspect ratio.

The LeakyReLU activation is preferred over standard ReLU to prevent dead neuron problems that can arise during training when large gradient updates push neuron pre-activations permanently into the negative regime. The progressively narrowing layer widths (256, 128, 64) implement an information bottleneck that forces the network to learn increasingly abstract representations of the position-relevant features, discarding noise and NLOS-induced artifacts in the intermediate layers while preserving the geometric information needed for accurate coordinate prediction.

Training is performed using the Adam optimizer [14] with initial learning rate 10^{-3}, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and mini-batch size 256. To prevent overfitting and improve generalization, two regularization mechanisms are employed: Gaussian noise injection with standard deviation 0.02 is applied to input features during training, simulating measurement variability and forcing the network to learn robust representations; and learning rate decay by a factor of 0.5 is applied at epochs 100 and 150 to enable fine-grained convergence in the later stages of training. The network is trained for a maximum of 200 epochs, with the model achieving the lowest validation error retained as the final estimator.

D. Deep Q-Network for Adaptive RIS Selection

The Deep Q-Network addresses the anchor selection problem through reinforcement learning, where the agent learns to identify which subset of RIS panels should be used for position estimation based on the current measurement characteristics. This is formulated as a single-step Markov decision process in which the state captures the complete measurement context, the action determines the anchor configuration, and the reward reflects the resulting positioning accuracy.

The state representation is a 32-dimensional vector formed by concatenating the 30 normalized measurement features with the DNN-estimated position coordinates $(\hat{x}/W, \hat{y}/H)$, providing the agent with both the raw measurement information and the current best position hypothesis. Including the DNN estimate in the state enables the DQN to assess whether the initial estimate appears geometrically consistent with the measurements, which can indicate the presence of NLOS corruption. The action space comprises five discrete actions corresponding to four 3-of-4 anchor subsets (excluding RIS-1, RIS-2, RIS-3, or RIS-4 respectively) and one use-all-four action. The reward signal is the negative Euclidean positioning error $R = -\|\hat{\boldsymbol{p}} - \boldsymbol{p}_{\mathrm{true}}\|$ computed after the complete DNN re-estimation and LM refinement pipeline executes on the selected anchor subset, directly driving the agent to learn anchor selections that minimize the end-to-end positioning error.

A critical implementation detail is the measurement masking mechanism that translates the DQN's discrete action into a modified feature vector for the DNN. When the DQN excludes an anchor, the corresponding range measurement is replaced with the mean of the retained measurements and the power is suppressed by 20 dB, ensuring that the excluded anchor contributes negligible information to the reconstructed feature vector. This masking approach allows the same DNN architecture to be used regardless of which anchor subset is

selected, avoiding the need for separate networks for each subset configuration.

The DQN architecture is a three-layer network with dimensions $32 \to 64 \to 32 \to 5$ and LeakyReLU activations, outputting five Q-values corresponding to the five possible actions. Training employs experience replay with a circular buffer of 10,000 transitions and mini-batch size 64, which decorrelates sequential training samples and improves learning stability. The ϵ-greedy exploration strategy begins with $\epsilon = 1.0$ (fully random actions) and decays by a factor of 0.997 per episode until reaching $\epsilon_{\min} = 0.05$, at which point the agent primarily exploits the learned policy while maintaining a small exploration probability. Training proceeds for 5,000 episodes under diverse environmental conditions as described in Section IIIA.

E. Hyperparameter Optimization

The hyperparameters governing both the DNN and DQN were selected through systematic evaluation on validation data, with the following key design choices motivated by empirical performance analysis. The DNN width of 256 neurons in the first hidden layer was chosen as the narrowest configuration that achieved validation error below 1.5 m on the augmented dataset, as wider networks (512 neurons) provided marginal improvement at significantly increased computational cost while narrower networks (128 neurons) exhibited validation errors exceeding 1.8 m. The training data volume of 30,000 samples represents the point of diminishing returns, as increasing to 50,000 samples reduced validation error by less than 3% while doubling the data generation and training time. The noise injection magnitude of $\sigma = 0.02$ was selected as the largest value that did not degrade training convergence, with larger values ($\sigma > 0.05$) causing training instability and smaller values ($\sigma < 0.01$) providing insufficient regularization.

For the DQN, the replay buffer size of 10,000 and the exploration decay rate of 0.997 were selected to ensure that the agent explores sufficiently before exploitation dominates, with the ϵ value reaching its minimum around episode 1,000. The 5,000-episode training budget was determined by monitoring the moving-average reward, which stabilized after approximately 4,000 episodes. The choice of five discrete actions (four 3-of-4 subsets plus use-all) rather than finer-grained subset combinations was motivated by the observation that with four anchors and at most one NLOS-corrupted measurement per trial, identifying and excluding a single anchor captures the dominant failure mode while maintaining a tractable action space for the DQN.

F. Proposed DQN–DNN–LM Positioning Method

The complete three-stage DQN–DNN–LM pipeline is formalized below and summarized in Algorithm 1.

The proposed positioning framework combines deep learning, reinforcement learning, and nonlinear geometric refinement in a hybrid pipeline. Given measured ranges $\{\hat{r}_k\}_{k=1}^4$ and received powers $\{P_k\}_{k=1}^4$, the first step is to construct the input feature vector

$$\boldsymbol{x} = \mathcal{F}(\{\hat{r}_k\}, \{P_k\}), \tag{4}$$

where $\mathcal{F}(\cdot)$ denotes the feature extraction function producing the 30-dimensional input representation. The extracted features are then normalized as

$$\tilde{\boldsymbol{x}} = \frac{\boldsymbol{x} - \boldsymbol{\mu}}{\boldsymbol{\sigma}}, \tag{5}$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ are the feature-wise mean and standard deviation computed from the training dataset. This normalized vector is passed to the DNN to obtain a coarse UE position estimate

$$\hat{\boldsymbol{p}}_0 = f_\theta(\tilde{\boldsymbol{x}}) \odot [W, H]^T, \tag{6}$$

where $f_\theta(\cdot)$ denotes the trained DNN and W and H are the room dimensions used to rescale the normalized network output to physical coordinates.

The coarse position estimate is then incorporated into the DQN state

$$\boldsymbol{s} = [\tilde{\boldsymbol{x}}, \hat{x}_0/W, \hat{y}_0/H], \tag{7}$$

which augments the normalized measurement features with the normalized coarse coordinate estimate. Based on this state, the DQN selects the anchor-selection action

$$a^* = \arg\max_a Q_\phi(\boldsymbol{s}, a), \tag{8}$$

where $Q_\phi(\boldsymbol{s}, a)$ is the learned action-value function. The chosen action determines the anchor subset $\mathcal{S}_{a^*}$ retained for refinement. Anchors not selected by the DQN are masked by replacing their range measurements with the mean of the retained ranges and suppressing their received powers by 20 dB. Using the masked measurements, a new feature vector is formed as

$$\boldsymbol{x}' = \mathcal{F}(\{\hat{r}'_k\}, \{P'_k\}), \tag{9}$$

and normalized as

$$\tilde{\boldsymbol{x}}' = \frac{\boldsymbol{x}' - \boldsymbol{\mu}}{\boldsymbol{\sigma}}. \tag{10}$$

The same DNN is then reused to produce a refined estimate

$$\hat{\boldsymbol{p}}_1 = f_\theta(\tilde{\boldsymbol{x}}') \odot [W, H]^T, \tag{11}$$

which serves as the initialization for the final nonlinear refinement stage.

To eliminate the unknown common clock bias, a TDoA formulation is adopted on the selected subset. The reference anchor is chosen as

$$k_{\text{ref}} = \arg\max_{k \in \mathcal{S}_{a^*}} P_k, \tag{12}$$

that is, the anchor with the highest received power within the selected subset. The corresponding TDoA measurements are then computed as

$$\Delta r_k = \hat{r}_k - \hat{r}_{k_{\text{ref}}}, \qquad \forall k \in \mathcal{S}_{a^*} \setminus \{k_{\text{ref}}\}. \tag{13}$$

Algorithm 1 DQN-Assisted Hybrid DNN–LM Positioning

Input: Measured ranges $\{\hat{r}_k\}_{k=1}^4$, received powers $\{P_k\}_{k=1}^4$, trained DNN f_θ, trained DQN Q_ϕ
Output: Estimated UE position $\hat{\boldsymbol{p}}$
$\boldsymbol{x} \leftarrow \mathcal{F}(\{\hat{r}_k\}, \{P_k\})$ Construct 30-dimensional feature vector
$\tilde{\boldsymbol{x}} \leftarrow (\boldsymbol{x} - \boldsymbol{\mu})/\boldsymbol{\sigma}$ Normalize features
$\hat{\boldsymbol{p}}_0 \leftarrow f_\theta(\tilde{\boldsymbol{x}}) \odot [W, H]^T$ DNN coarse estimate
$\boldsymbol{s} \leftarrow [\tilde{\boldsymbol{x}}, \hat{x}_0/W, \hat{y}_0/H]$ Construct DQN state
$a^* \leftarrow \arg\max_a Q_\phi(\boldsymbol{s}, a)$ Select action
Determine anchor subset $\mathcal{S}_{a^*}$ from a^*
Mask excluded anchors: replace excluded ranges by mean retained range and suppress excluded powers by 20 dB
$\boldsymbol{x}' \leftarrow \mathcal{F}(\{\hat{r}'_k\}, \{P'_k\})$ Reconstruct feature vector
$\tilde{\boldsymbol{x}}' \leftarrow (\boldsymbol{x}' - \boldsymbol{\mu})/\boldsymbol{\sigma}$ Normalize masked features
$\hat{\boldsymbol{p}}_1 \leftarrow f_\theta(\tilde{\boldsymbol{x}}') \odot [W, H]^T$ DNN refined estimate
$k_{\mathrm{ref}} \leftarrow \arg\max_{k \in \mathcal{S}_{a^*}} P_k$ Select reference anchor
$\Delta r_k \leftarrow \hat{r}_k - \hat{r}_{k_{\mathrm{ref}}},\ \forall k \in \mathcal{S}_{a^*} \setminus \{k_{\mathrm{ref}}\}$ Compute TDoA
$\hat{\boldsymbol{p}} \leftarrow \hat{\boldsymbol{p}}_1$ Initialize LM
repeat
  Update $\hat{\boldsymbol{p}}$ by one LM iteration using TDoA residuals over $\mathcal{S}_{a^*}$
until $\|\hat{\boldsymbol{p}}^{(i+1)} - \hat{\boldsymbol{p}}^{(i)}\| < \epsilon$
Clamp $\hat{\boldsymbol{p}}$ within $[0, W] \times [0, H]$
return $\hat{\boldsymbol{p}}$

Using these differential measurements, the final UE position is estimated by solving the nonlinear least-squares problem

$$\hat{\boldsymbol{p}} = \arg\min_{\boldsymbol{p}} \sum_{k \in \mathcal{S}_{a^*} \setminus \{k_{\mathrm{ref}}\}} \left(\Delta r_k - [\|\boldsymbol{p} - \boldsymbol{a}_k\| - \|\boldsymbol{p} - \boldsymbol{a}_{k_{\mathrm{ref}}}\|]\right)^2, \tag{14}$$

where $\boldsymbol{a}_k$ denotes the known position of anchor k. This optimization is solved iteratively using the Levenberg–Marquardt algorithm initialized at $\hat{\boldsymbol{p}}_1$. After convergence, the resulting estimate is clamped to the feasible room area.

The complete procedure is summarized in Algorithm 1.

IV. Simulation Results and Analysis

A. Experimental Setup and Compared Methods

The positioning framework is evaluated through comprehensive Monte Carlo simulations with 500 random UE positions uniformly distributed throughout the room interior, each independently evaluated over $N_{\mathrm{MC}} = 20$ trials with different random noise and NLOS realizations. The complete set of simulation parameters is summarized in Table I.

The proposed framework is compared against four representative baselines spanning geometric, robust-statistical, and fingerprinting-based positioning.

The Joint Levenberg-Marquardt (LM) algorithm estimates the UE position $\boldsymbol{p} = [x, y]^T$ and the unknown clock bias b jointly by minimizing the weighted sum of squared range residuals $\sum_{k=1}^4 w_k(\hat{r}_k - \|\boldsymbol{p} - \boldsymbol{p}_k\| - d_{k,\mathrm{gNB}} - b)^2$, where $w_k = 10^{P_k/10}/\sum_j 10^{P_j/10}$ are power-derived weights [15]. Multi-start initialization from five candidate positions mitigates the risk of convergence to local minima.

TABLE I: Simulation Parameters

Parameter	Value
Room dimensions	60×40 m
Number of RIS panels	4
RIS elements (total / effective)	256 / 64
gNB receive antennas (N_{rx})	8
Carrier frequency (f_c) / Bandwidth (B)	3.5 GHz / 100 MHz
Channel model / K-factor	3GPP CDL-D / 13.3 dB
Baseline SNR	30 dB
NLOS probability (p_{NLOS}) / bias mean	0.25 / 4 m
Clock bias range	± 5 m equiv.
KNN grid spacing / neighbors	5 m / K=3
DNN architecture	30→256→128→64→2
DNN training samples / epochs	30,000 / 200
DQN architecture	32→64→32→5
DQN episodes / replay buffer	5,000 / 10,000
UE test positions / MC trials	500 / 20

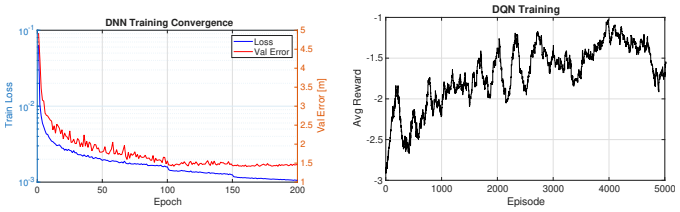
The Time-Difference-of-Arrival (TDoA) baseline eliminates the unknown clock bias through differential processing by selecting the strongest RIS panel as reference and computing range differences $\Delta r_k = \hat{r}_k - \hat{r}_{k^\star}$, which cancels the common bias exactly. The resulting two-unknown system is solved via LM optimization with multi-start initialization. The reduction from three unknowns to two improves the geometric conditioning of the estimation problem when only four anchors are available.

The K-Nearest Neighbors (KNN) fingerprinting algorithm operates in the received signal strength domain using an offline calibration database constructed on a $\Delta = 5$ m grid. During online positioning, the measured RSSI vector is compared against all database entries, and the UE position is estimated as the centroid of the $K = 3$ nearest reference positions. While this approach requires no geometric modeling and naturally handles arbitrary propagation environments, its accuracy is fundamentally limited by the fingerprint grid resolution.

The Robust TDoA algorithm combines TDoA processing with Iteratively Reweighted Least Squares (IRLS) optimization using the Huber loss function and exhaustive best-3-of-4 subset selection [8], [16]. The Huber loss transitions between quadratic and linear regimes at the threshold $1.345\hat{\sigma}$, where $\hat{\sigma}$ is estimated via the Median Absolute Deviation, effectively down-weighting corrupted measurements while preserving smooth convergence. The best-subset mechanism evaluates all $\binom{4}{3} = 4$ possible three-anchor subsets and selects the solution achieving the lowest capped full-set residual, enabling identification and exclusion of the single most problematic measurement.

B. Training Convergence

Fig. 2a presents the DNN training convergence over 200 epochs. The training loss decreases by approximately two orders of magnitude within the first 50 epochs and continues to improve gradually, reaching a final value of approximately 10^{-3}. The validation error stabilizes around 1.4 m after epoch 120, with the gap between training loss and validation error



(a) DNN training convergence. (b) DQN reward convergence.

Fig. 2: Training convergence of the learning components used in the proposed framework. Both the supervised estimator and the reinforcement-learning agent converge stably under the adopted training settings.

remaining small throughout, confirming adequate generalization without significant overfitting. This result is attributable to the combination of augmented training data, Gaussian noise injection, and early stopping.

Fig. 2b shows the DQN training reward curve smoothed with a 200-episode moving average. The average reward improves from approximately -2.8 in the early exploration phase to approximately -1.4 by episode 5,000, representing a 50% reduction in average positioning error through learned anchor selection. The reward exhibits notable variance in the later training stages due to the continuously randomized episode conditions, but the overall upward trend confirms that the agent progressively learns to identify and exclude NLOS-corrupted anchors.

C. Overall Positioning Accuracy and Error Distribution

Table II summarizes the positioning accuracy metrics for all seven algorithms under the baseline configuration. The proposed DQN-DNN-LM pipeline achieves the best overall performance with a mean error of 1.528 m and a median error of 1.410 m, representing a 27.4% improvement in mean error over the Robust TDoA baseline (2.106 m mean, 2.026 m median), which is the strongest geometric competitor. The DNN-only estimator also outperforms all geometric baselines with 1.581 m mean error, confirming that the learned feature representation and augmented training effectively mitigate NLOS bias that degrades conventional methods. Among the geometric approaches, the Robust TDoA achieves the best performance with 2.026 m median, followed by TDoA baseline (2.142 m) and Joint LM (2.238 m). The KNN fingerprinting approach yields significantly higher errors (6.533 m mean, 6.507 m median), confirming that the coarse 5 m grid resolution is the dominant accuracy-limiting factor.

An important observation concerns the Hybrid DNN+LM approach, which uses the DNN estimate to initialize LM on the full four-anchor set without DQN-based anchor selection. This approach performs slightly worse than plain TDoA (2.098 m vs. 2.142 m median), indicating that LM refinement without anchor selection can actually degrade the DNN estimate when NLOS-corrupted measurements are included, as the geometric solver converges toward a biased solution that fits the corrupted measurement. This observation directly motivates the DQN anchor selection stage, since excluding the suspected NLOS anchor before LM refinement allows the geometric solver to operate on clean measurements and converge to the correct solution. In terms of median positioning error, the proposed method improves by 37.0% over Joint LM, 34.2% over TDoA Baseline, 30.4% over Robust TDoA, and 7.2% over DNN Only, while also outperforming KNN by a wide margin.

TABLE II: Positioning Accuracy at SNR = 30 dB, $p_{\mathrm{NLOS}} = 25\%$ (meters)

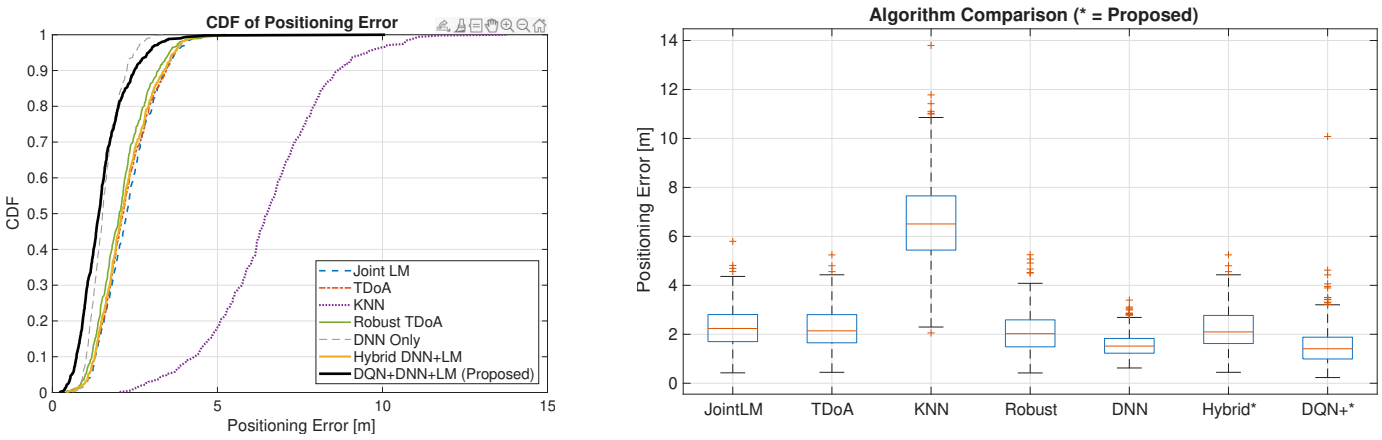
Algorithm	Mean	Median	90th%	95th%
Joint LM	2.309	2.238	3.425	3.745
TDoA Baseline	2.253	2.142	3.453	3.719
KNN (K=3)	6.533	6.507	8.647	9.560
Robust TDoA (IRLS)	2.106	2.026	3.219	3.549
DNN Only	1.581	1.519	2.255	2.473
Hybrid DNN+LM	2.216	2.098	3.386	3.706
DQN-DNN-LM	**1.528**	**1.410**	**2.489**	**2.916**

Fig. 3a presents the cumulative distribution function (CDF) of the positioning error across all 500 UE positions. The proposed DQN-DNN-LM achieves approximately 80% probability of sub-2 m positioning, compared to approximately 55% for the geometric baselines and only 5% for KNN. The DNN-only curve closely tracks the proposed method for errors below 2 m but diverges slightly in the upper tail, where the DQN's anchor selection provides additional outlier resilience by excluding NLOS-corrupted measurements before LM refinement. The KNN CDF is clearly separated from all other methods, with its steep rise around 5 to 7 m reflecting the quantization effect of the 5 m grid spacing.

Fig. 3b provides a box-plot comparison of the error distributions. The DNN-based methods, namely DNN-only and DQN-DNN-LM, exhibit the most compact distributions with the lowest medians, with interquartile ranges of approximately 0.8 to 1.5 m. The geometric baselines show wider distributions with medians around 2.0 to 2.2 m, while the KNN box is centered around 6.5 m with an IQR spanning approximately 5 to 8 m.

D. Impact of SNR, RIS Array Size, and NLOS Probability

Fig. 4a presents the median positioning error as a function of SNR from 10 to 40 dB, with NLOS disabled to isolate the effect of ranging noise. The geometric algorithms and the hybrid methods exhibit a clear monotonically decreasing relationship with SNR, with errors dropping from approximately 0.8 m at 10 dB to below 0.03 m at 40 dB. On the semi-logarithmic scale this relationship appears approximately linear, consistent with the CRLB prediction of $\sigma_r \propto 1/\sqrt{\mathrm{SNR}}$ from Eq. (3), confirming that the positioning accuracy is fundamentally noise-limited in the absence of NLOS bias. The DNN-only estimator exhibits a distinct behavior, with its error plateauing at approximately 0.4 to 0.6 m for SNR values exceeding 25 dB, reflecting the irreducible generalization error of the learned regression model. This learning floor arises because the DNN was trained on samples with noise levels



(a) CDF of positioning error.

(b) Box-plot comparison of error distributions.

Fig. 3: Statistical comparison of positioning accuracy at SNR = 30 dB with $p_{\mathrm{NLOS}} = 0.25$. The proposed DQN-DNN-LM consistently exhibits the lowest errors and the most compact distribution among the compared methods.

spanning 20 to 40 dB and its output represents a statistical average over these conditions. The hybrid and DQN-DNN-LM methods do not suffer from this floor because their LM refinement stage exploits the exact measurement equations, achieving CRLB-consistent performance when the DNN provides a sufficiently good initialization. The KNN algorithm remains at approximately 5 to 6 m across the entire SNR range, confirming that its accuracy is grid-resolution-limited rather than SNR-limited.

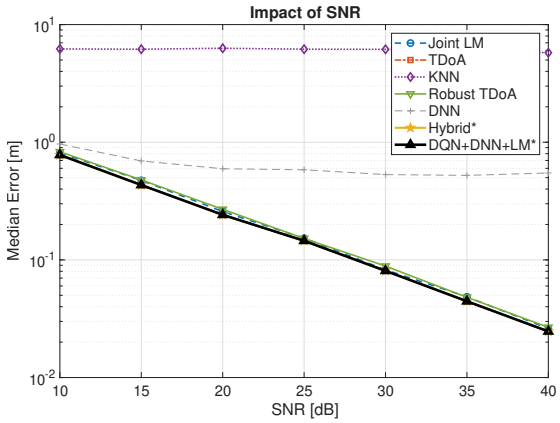
Fig. 4b examines the effect of RIS array size by varying the number of effective elements from $N = 16$ to $N = 256$, again without NLOS. The geometric and hybrid methods converge to sub-5-centimeter accuracy at $N = 256$ due to the N^2 coherent combining gain, while the DNN-only estimator exhibits its characteristic learning floor at approximately 0.4 m independent of array size, which is a consequence of being trained on data with fixed $N = 64$. This has important design implications: for deployments where NLOS is rare and high SNR is achievable through large arrays, geometric methods suffice; however, for practical environments where NLOS contamination is unavoidable, the DRL-based approach provides robust performance whose advantages stem from NLOS mitigation rather than noise reduction.

Fig. 5 presents the most practically relevant parametric study, showing the median positioning error as a function of NLOS probability from 0 to 50%. In the ideal LOS scenario, the geometric methods achieve approximately 0.05 to 0.10 m median error consistent with the CRLB, while the DNN-based methods start at their learning floor of approximately 0.5 m. A critical crossover occurs around $p_{\mathrm{NLOS}} \approx 8\%$, below which geometric methods provide superior accuracy due to their CRLB-optimal noise handling and above which the proposed DQN-DNN-LM consistently outperforms all baselines due to its learned NLOS resilience. As the NLOS probability increases to 50%, the geometric baselines degrade substantially, with Joint LM and TDoA reaching approximately 3.4 m and Robust TDoA reaching 3.1 m, demonstrating that even principled outlier mitigation through Huber loss and subset selection provides limited protection against frequent NLOS contamination. In contrast, the proposed DQN-DNN-LM degrades more gracefully to approximately 2.6 m, maintaining 26% lower error than the Robust TDoA at the highest contamination level.

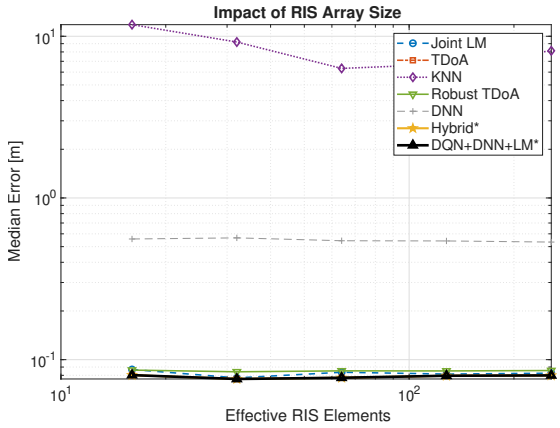
The DNN-only estimator surpasses the geometric baselines by approximately $p_{\mathrm{NLOS}} = 8\%$ and maintains lower error throughout the high-NLOS regime, confirming that the augmented training, which exposed the DNN to diverse NLOS conditions during learning, enables it to implicitly develop NLOS-resilient features that extract position information from the subset of clean measurements without explicit outlier detection. The DQN-DNN-LM further improves upon DNN-only by actively selecting the best anchor subset through the learned policy, providing approximately 0.2 to 0.5 m additional improvement at high NLOS probabilities.

E. Position Error Bound Analysis

Fig. 6 compares the theoretical Position Error Bound (PEB) derived from the Cramér-Rao bound with the achieved positioning error of the proposed algorithm across the room area. The spatial error distribution shows that the lowest errors, below 1.5 m, are concentrated in the room interior within the convex hull formed by the four RIS panel positions, where the UE benefits from favorable angular diversity and the DNN has dense training coverage. Errors gradually increase toward the room boundaries and corners, particularly near the walls adjacent to the RIS panels, where the geometric constraints from nearby panels become nearly parallel. The spatial error pattern remains more uniform than would be expected from purely geometric methods, as the DNN's learned regression smooths out the sharp GDOP-dependent variations that characterize LM-based estimators. Quantitatively, the PEB ranges from 0.067 m in the room center to 0.087 m near the walls, while the achieved DQN-DNN-LM error ranges from approximately 0.5 m in the best interior positions to 4 to 5 m near boundaries.



(a) Median error versus SNR.



(b) Median error versus effective RIS elements.

Fig. 4: Sensitivity of the compared methods to LOS operating conditions. Geometric and hybrid methods follow the expected SNR and array-gain scaling, while the DNN-only estimator exhibits a learning floor at high signal quality.

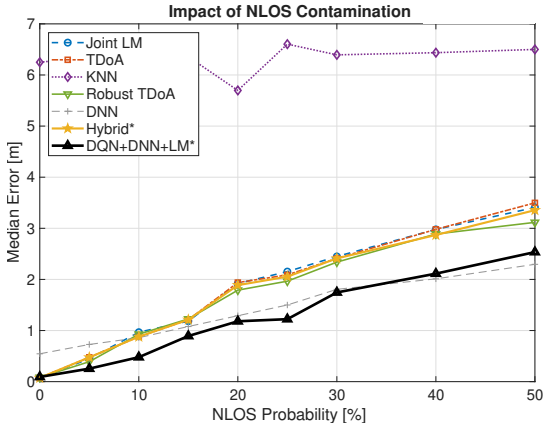


Fig. 5: Median positioning error versus NLOS probability (0 to 50%). The proposed DQN-DNN-LM shows the most gradual error degradation, with increasing advantage over all baselines as NLOS contamination grows.

The large gap between the theoretical bound and the achieved error is dominated by NLOS bias rather than fundamental geometric limitations, confirming that NLOS mitigation, rather than improved noise handling or geometric optimization, represents the primary opportunity for positioning accuracy improvement in practical indoor environments. This validates the DRL-based approach's focus on learned NLOS resilience as the most impactful design choice.

F. Summary of Key Findings

The comprehensive simulation analysis reveals several key findings for practical DRL-aided RIS positioning system design. The proposed DQN-DNN-LM pipeline consistently outperforms all baselines under NLOS conditions, with the advantage growing from approximately 7% at low NLOS probability to 26% at 50% NLOS, which is precisely the most challenging and practically relevant operating regime for indoor positioning. A clear crossover threshold exists at approximately 8% NLOS probability, below which geometric methods are preferable due to their CRLB-optimal performance and above which the DRL approach dominates. The three-stage hybrid architecture proves essential because the DNN alone provides robust initialization that avoids LM local minima, the DQN provides intelligent anchor selection that excludes NLOS measurements, and the LM refinement provides geometric precision that exploits exact measurement equations. The augmented training strategy, which spans diverse SNR and NLOS conditions, is critical for DNN generalization, while the DQN's experience replay and diverse episode conditions ensure that the learned anchor selection policy transfers to test conditions not seen during training. Finally, the range-based algorithms, both geometric and DRL-enhanced, outperform KNN fingerprinting by approximately one order of magnitude in positioning accuracy, confirming that exploitation of wideband 5G NR timing precision fundamentally outperforms power-based methods for RIS-aided indoor positioning. Overall, the results indicate that the proposed method is most beneficial in moderate-to-high NLOS regimes, where learned anchor selection and DNN initialization improve robustness without sacrificing the benefits of geometric refinement.

V. Conclusions and Future Work

This paper presented a deep reinforcement learning framework for position-aided beam management in multi-RIS indoor environments. The proposed DQN-DNN-LM pipeline combines a deep neural network position estimator trained on 30,000 augmented samples with geometric feature priors, a Deep Q-Network that learns adaptive RIS anchor selection through 5,000 episodes of experience replay, and a Levenberg-Marquardt geometric refinement stage that exploits the DQN-selected anchor subset for precise position recovery. Extensive Monte Carlo simulations in a representative $60 \times 40\,\mathrm{m}$ indoor environment with four 256-element RIS panels at 3.5 GHz demonstrated that the proposed pipeline achieves a median positioning accuracy of 1.41 m at 30 dB SNR with

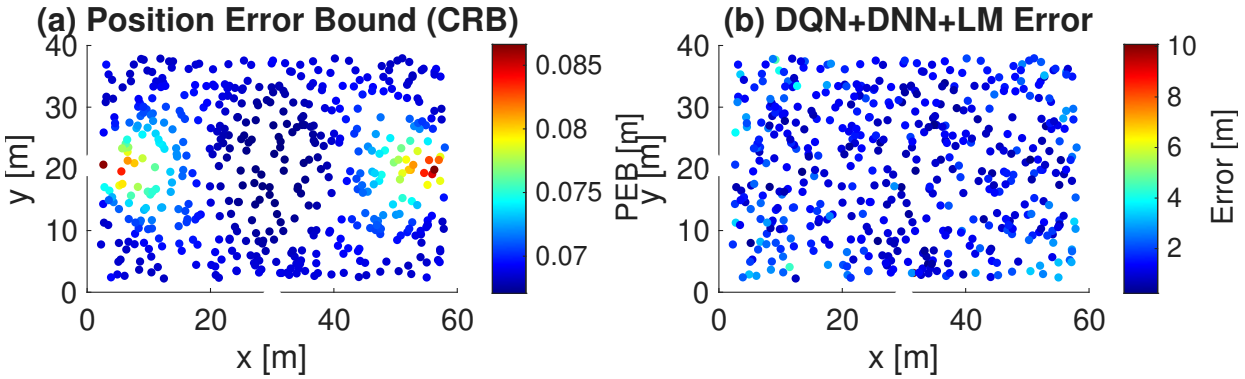


Fig. 6: Comparison of the theoretical Position Error Bound (PEB) from the Cramér-Rao bound (left) and the achieved DQN-DNN-LM positioning error (right) across the $60 \times 40\,\text{m}$ room. The PEB represents the fundamental noise-limited accuracy achievable under LOS conditions, while the achieved error is dominated by residual NLOS bias.

25% NLOS probability, representing a 27.4% improvement over the best geometric baseline. Comprehensive parametric studies confirmed four fundamental results: geometric methods achieve CRLB-optimal sub-centimeter accuracy in LOS conditions and should be preferred when NLOS probability is below approximately 8%; above this crossover the DRL-based approach provides increasingly significant advantages reaching 26% lower error at 50% NLOS probability; the three-stage pipeline architecture is essential as the DNN provides robust initialization, the DQN provides learned anchor selection, and the LM provides geometric precision; and the N^2 coherent combining gain of RIS enables order-of-magnitude noise-limited accuracy improvements through larger arrays, although NLOS mitigation remains the dominant accuracy-limiting factor in practical indoor environments.

Future work will extend the framework to three-dimensional positioning by incorporating elevation information from multi-panel vertical aperture, investigate online DNN and DQN adaptation for time-varying environments through continual learning, explore joint optimization of RIS phase configuration and positioning through multi-agent reinforcement learning, and pursue experimental validation using hardware RIS prototypes and software-defined radio platforms to bridge the gap between simulation and practical deployment.

Acknowledgment

This work was co-funded by the European Union under the programme of social cohesion "THALIA 2021-2027," through the Research and Innovation Foundation (Project: EXCELLENCE/0524/0218).

References

[1] J. A. del Peral-Rosado, R. Raulefs, J. A. López-Salcedo, and G. Seco-Granados, "Survey of cellular mobile radio localization methods: From 1G to 5G," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 2, pp. 1124–1148, 2018.

[2] R. Keating, M. Säily, J. Hulkkonen, and J. Karjalainen, "Overview of positioning in 5G new radio," in *2019 16th International Symposium on Wireless Communication Systems (ISWCS)*. Oulu, Finland: IEEE, Aug. 2019, pp. 320–324.

[3] M. Di Renzo *et al.*, "Smart radio environments empowered by reconfigurable intelligent surfaces: How it works, state of research, and the road ahead," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 11, pp. 2450–2525, Nov. 2020.

[4] Q. Wu and R. Zhang, "Intelligent reflecting surface enhanced wireless network via joint active and passive beamforming," *IEEE Trans. Wireless Commun.*, vol. 18, no. 11, pp. 5394–5409, Nov. 2019.

[5] H. Wymeersch, J. He, B. Denis, A. Clemente, and M. Juntti, "Radio localization and mapping with reconfigurable intelligent surfaces: Challenges, opportunities, and research directions," *IEEE Veh. Technol. Mag.*, vol. 15, no. 4, pp. 52–61, Dec. 2020.

[6] A. Elzanaty, A. Guerra, F. Guidi, and M.-S. Alouini, "Reconfigurable intelligent surfaces for localization: Position and orientation estimation through curvature of wavefront," *IEEE Commun. Mag.*, vol. 59, no. 11, pp. 14–20, Nov. 2021.

[7] Z. Abu-Shaban, K. Keykhosravi, M. F. Keskin, G. C. Alexandropoulos, G. Seco-Granados, and H. Wymeersch, "Near-field localization with a reconfigurable intelligent surface acting as lens," in *Proc. IEEE ICC*, 2021, pp. 1–6.

[8] R. Zong and Z. Huang, "Robust positioning based on iteratively reweighted least squares for NLOS environments," *IEEE Commun. Lett.*, vol. 26, no. 3, pp. 654–658, Mar. 2022.

[9] N. C. Luong *et al.*, "Applications of deep reinforcement learning in communications and networking: A survey," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 4, pp. 3133–3174, 2019.

[10] X. Wang, L. Gao, S. Mao, and S. Pandey, "CSI-based fingerprinting for indoor localization: A deep learning approach," *IEEE Trans. Veh. Technol.*, vol. 66, no. 1, pp. 763–776, Jan. 2017.

[11] E. Dahlman, S. Parkvall, and J. Sköld, *5G NR: The Next Generation Wireless Access Technology*, 2nd ed. Academic Press, 2020.

[12] 3GPP, "Study on channel model for frequencies from 0.5 to 100 GHz," 3GPP, Technical Report TR 38.901 V16.1.0, Nov. 2020, release 16.

[13] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Upper Saddle River, NJ, USA: Prentice-Hall, 1993.

[14] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, San Diego, CA, USA, 2015.

[15] K. Levenberg, "A method for the solution of certain non-linear problems in least squares," *Quart. Appl. Math.*, vol. 2, no. 2, pp. 164–168, 1944.

[16] P. J. Huber, *Robust Statistics*. Hoboken, NJ, USA: Wiley, 2004.