

DEEP REINFORCEMENT LEARNING FOR DYNAMIC SENSING AND COMMUNICATIONS

Abolfazl Zakeri¹, Nhan Thanh Nguyen¹, Ahmed Alkhateeb², and Markku Juntti¹

¹ Centre for Wireless Communications – Radio Technologies (CWC-RT), University of Oulu, Finland

Email: {abolfazl.zakeri, nhan.nguyen, markku.juntti}@oulu.fi

² School of Electrical, Computer, and Energy Engineering, Arizona State University, USA

Email: alkhateeb@asu.edu

ABSTRACT

Environmental sensing can significantly enhance mmWave communications by assisting beam training, yet its benefits must be balanced against the associated sensing costs. To this end, we propose a unified machine learning framework that dynamically determines when to sense and leverages sensory data for beam prediction. Specifically, we formulate a joint sensing and beamforming problem that maximizes the average signal-to-noise ratio under an average sensing budget. Lyapunov optimization is employed to enforce the sensing constraint, while a deep Q-Network determines the sensing slots. A pretrained deep neural network then maps the sensing data to optimal beams in the codebook. Simulations based on the real-world DeepSense dataset demonstrate that the proposed approach substantially reduces sensing overhead while maintaining satisfactory communications performance.

Index Terms—Multimodal sensing and communications, deep Q-Network, machine learning, beam prediction

1. INTRODUCTION

Beam training is essential for reliable performance in wireless systems, particularly at millimeter-wave (mmWave) frequencies, where links are highly sensitive to path loss and blockages. Recently, multimodal sensing-aided communications, which leverage sensory data such as visual, LiDAR, and radar measurements, have attracted growing interest [1–3]. By enhancing environmental perception and situational awareness, multimodal sensing helps reduce beam training overhead [4, 5] and improve beam alignment [6]. These benefits are especially valuable in high-mobility settings where proactive line-of-sight (LoS) prediction and beam selection are critical for sustaining reliable connectivity.

Recent studies in multimodal sensing-aided communications [4, 6–16] have primarily focused on beam prediction. Patel *et al.* [4] showed that sensor-aided deep learning enables efficient channel estimation for multi-user mmWave systems, achieving interference-free beamforming and improved spectral efficiency. Jiang *et al.* [13] demonstrated that LiDAR-aided machine learning can predict and track beams in real

vehicular settings with low training overhead, a result later extended to multimodal prediction [7, 14]. These works show that multimodal sensing not only facilitates beam training but also improves beamforming performance.

Nonetheless, most prior works overlook the optimization of the sensing process itself, often assuming that sensory data is continuously available. While such assumptions yield useful insights into how different modalities can aid beam training, they leave open a fundamental question: *When should sensing be performed, and which modality should be employed to efficiently sustain communication performance?* Addressing this question requires explicitly modeling the tradeoff between the sensing rate (i.e., how often to sense) and the sensing modality (i.e., which modality to employ) versus the resulting communications performance. This, in turn, necessitates algorithms that support *dynamic and adaptive joint* sensing and communications design.

In our previous work [11], we introduced a constrained sensing and dynamic beamforming approach. That approach requires an exhaustive search for sensing decisions at each slot and relies solely on position-based channel modeling. To overcome these limitations, in this work, we propose a two-module learning framework: a Deep Q-Network (DQN) for sensing decisions and a deep neural network (DNN) for beam prediction (see Fig. 1). The objective is to dynamically determine both when to sense and which beam to select to maximize average signal-to-noise-ratio (SNR), subject to an average sensing constraint. Specifically, we first pre-train the DNN on available data with optimal beam labels, then apply Lyapunov queue stability to enforce the sensing constraint, thereby transforming the problem into a Markov decision process solved by the DQN in conjunction with the pretrained DNN. Preliminary results on the DeepSense position dataset show that our framework can reduce sensing frequency by nearly 50% while maintaining beam prediction accuracy comparable to always sensing in every slot.

2. SYSTEM MODEL AND PROBLEM FORMULATION

We consider a downlink mmWave communications system including a base station (BS) and a single-antenna mobile user

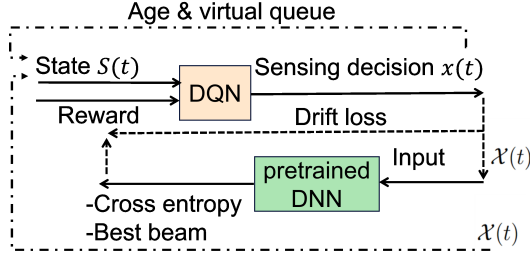


Fig. 1. A schematic of the proposed two-module machine learning method for dynamic sensing and beam prediction

equipment (UE). The BS is equipped with N antennas and a sensing unit. At time slot $t = 1, 2, \dots$, the BS employs analog beamforming vector $\mathbf{w}(t) \in \mathcal{W}$ for signal transmission, where $\mathcal{W} = \{\mathbf{w}_1, \dots, \mathbf{w}_M\}$ denotes the codebook of M candidate beamforming vectors with $\|\mathbf{w}_m\|_2^2 = 1, \forall m$.

Denote by $\mathbf{h}(t) \in \mathbb{C}^{N \times 1}$ the channel between the BS and the UE at time slot t . Furthermore, let $s(t) \in \mathbb{C}$ be the transmit data symbol from the BS to the UE, $\mathbb{E}\{|s(t)|^2\} = P$ with P being the transmit power. The received signal at the UE is then given by

$$y(t) = \mathbf{h}^H(t)\mathbf{w}(t)s(t) + n(t), \quad (1)$$

where $n(t) \in \mathbb{C}$ is additive white Gaussian noise (AWGN) following the distribution $\mathcal{CN}(0, \sigma^2)$, with σ^2 denoting the noise variance at the UE's receiver.

Problem Formulation: The primary task in multimodal sensing for communications is to exploit sensory data (e.g., position, images, LiDAR) to select the best beam from M candidates that maximizes the UE's received SNR, i.e., $|\mathbf{h}^H(t)\mathbf{w}(t)|^2$. This is typically done by training a DNN on collected sensory data with corresponding best-beam labels. In this paper, we extend this task to a dynamic *joint sensing and beam prediction* problem.

Let $x(t) \in \{0, 1\}$ denote the sensing decision at slot t , where $x(t) = 1$ indicates that sensing is performed and $x(t) = 0$ otherwise. Performing sensing provides fresh data for beam prediction but incurs a cost c , which accounts for, e.g., power consumption and processing overhead of data acquisition. Since using recent data generally improves beam prediction accuracy, this naturally leads to a tradeoff between sensing frequency (i.e., data acquisition rate) and communication performance. We formulate this tradeoff as the following optimization problem:

$$\begin{aligned} & \underset{\{m(t), x(t)\}_{t=1,2,\dots}}{\text{maximize}} && \limsup_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T \mathbb{E}[|\mathbf{h}^H(t)\mathbf{w}_{m(t)}|^2] \quad (2a) \\ & \text{subject to} && \limsup_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T \mathbb{E}[cx(t)] \leq c^{\max}, \quad (2b) \end{aligned}$$

where $m(t) \in \{1, \dots, M\}$ is the selected beam index, $x(t) \in \{0, 1\}$ is the sensing decision, and c^{\max} denotes the sensing

cost budget. The implicit effect of $x(t)$ lies in determining whether *fresh* sensory data is available for beam selection.

The main challenge in solving problem (2) is that the instantaneous channel information $\mathbf{h}(t)$ is not available. As an alternative, multimodal sensory data combined with machine learning (ML) can be leveraged to learn the best beam selections. To this end, we propose a two-module ML framework (see Fig. 1): a DQN for sensing decisions and a DNN for beam prediction, as detailed in the next section.

3. PROPOSED JOINT SENSING AND BEAM PREDICTION APPROACH

As illustrated in Fig. 1, our framework consists of two ML modules: (i) a DQN that learns over time the optimal sensing actions, and (ii) a DNN that predicts the best beam given the available sensory data. The DNN is first trained offline using sensory data samples from all time slots. Then, the DQN is trained with the pretrained DNN as its beam prediction module. It is important to note that the DQN's sensing decisions must satisfy the sensing cost constraint in (2b).

To enforce the average constraint (2b), we adopt the Lyapunov queue stability framework [17, Ch. 2]. Let $Q(t)$ denote the virtual queue associated with constraint (2b) at slot t , whose evolution is given by

$$Q(t+1) = \max \left[Q(t) + x(t) - \frac{c^{\max}}{c}, 0 \right]. \quad (3)$$

Here, $Q(t)$ evolves as a queue with arrival rate $x(t)$ and service rate $\frac{c^{\max}}{c}$. According to [17, Ch. 2], the time-average constraint (2b) is satisfied if the queue is *strongly stable*, i.e., $\limsup_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T \mathbb{E}\{Q(t)\} < \infty$.

We now define the Lyapunov function and its drift to facilitate the stability analysis. Let $L(Q(t)) = \frac{1}{2}Q^2(t)$ denote the quadratic Lyapunov function [17, Ch. 3]. Stabilization of the virtual queue can be achieved by minimizing the expected change of this function across slots [17, Ch. 3]. The one-slot conditional Lyapunov drift, denoted by $\Delta(t)$, is defined as the expected change in the Lyapunov function over one slot, conditioned on the current system state $S(t)$, i.e.,

$$\Delta(t) \triangleq \mathbb{E}[L(Q(t+1)) - L(Q(t)) | S(t)]. \quad (4)$$

Having the drift is defined, next, we explain the cost function used for the training of the DQN module in Fig. 1. We have used a similar idea in [18] and apply a drift-plus-penalty notion [17] to define the cost function for each state $S(t)$ and action $x(t)$, denoted by $C(S(t), x(t))$. Thus, the objective of DQN can be expressed as

$$J(\pi) = \mathbb{E} \left[\sum_{t=0}^{\infty} \gamma^t C(S(t), x(t)) \right], \quad 0 \leq \gamma < 1. \quad (5)$$

where π denotes the policy, i.e., a mapping from each state

$S(t)$ to the binary sensing action $x(t)$, determined by the DQN via the $\arg\max$ operator on its output layer. Moreover, γ is the discount factor. The expectation is taken over the random trajectories induced by policy π , due to (possibly) stochastic action selection and environment dynamics.

DQN Training Cost Function: The immediate cost function for DQN, denoted as $C(S(t), x(t))$, consists of two components: (1) the *beam prediction loss* associated with the DNN module, since the DQN's sensing decision determines the DNN input and thereby affects beam prediction accuracy; and (2) the penalty associated with the average sensing constraint (2b), captured by the drift term $L(Q(t+1)) - L(Q(t))$. For the beam prediction loss, let the softmax output of the DNN be given by $\Pr\{m^*(t) = i\} = p_i(t)$, where

$$p_i(t) = \text{softmax}(z_i(t)) \triangleq \frac{e^{z_i(t)}}{\sum_{j=1}^M e^{z_j(t)}}, i = 1, \dots, M, \quad (6)$$

and $z_i(t)$ is the i -th logit (i.e., raw output of the model). Ideally, we desire $p_i(t)$ to be a delta function, i.e., $p_i(t) = \delta(i - m^*(t))$, where $m^*(t)$ is the optimal beam index at time t . Accordingly, the beam prediction loss is defined using cross-entropy:

$$\mathcal{L}(t) \triangleq - \sum_{i=1}^M \delta(i - m^*(t)) \log(p_i(t)) = -\log(p_{m^*(t)}(t)). \quad (7)$$

The immediate cost function is then given by

$$C(S(t), x(t)) \triangleq V\mathcal{L}(t) + (L(Q(t+1)) - L(Q(t))), \quad (8)$$

where V is a non-negative parameter chosen to desirably adjust a trade-off between the size of the virtual queue and the beam prediction accuracy.

DQN State $S(t)$: Let $\mathcal{X}(t)$ denote the *input* to the DNN module in Fig. 1, e.g., a 2D position concatenated with RGB image features. Let $\mathcal{X}_{\text{curr}}(t)$ denote the *current* data at slot t (e.g., current position and RGB image). Then,

$$\mathcal{X}(t) = \begin{cases} \mathcal{X}_{\text{curr}}(t), & \text{if } x(t) = 1, \\ \mathcal{X}_{\text{old}}(t), & \text{if } x(t) = 0, \end{cases} \quad (9)$$

where $\mathcal{X}_{\text{old}}(t) = \mathcal{X}_{\text{curr}}(t')$ for the latest t' such that $x(t') = 1$. The DQN state is then defined as

$$S(t) \triangleq (\mathcal{X}(t), Q(t), \theta(t)), \quad (10)$$

where $\theta(t)$ is the *age of the most recent sample*, evolving as

$$\theta(t+1) = \begin{cases} 1, & \text{if } x(t) = 1, \\ \theta(t) + 1, & \text{if } x(t) = 0. \end{cases} \quad (11)$$

In Section 4, we numerically evaluate the impact of in-

Algorithm 1: Proposed Joint Sensing and Beam Prediction Scheme

```

/* Initialization */
1 Set system parameters  $V$ , sensing limit, and initialize DNN
  and DQN parameters;
/* Step (1): Train DNN for beam
  prediction */
2 for each epoch do
3   Sample a mini-batch from the dataset;
4   Compute predictions and cross-entropy loss;
5   Update DNN parameters via backpropagation;
/* Step (2): Train DQN for sensing
  decisions */
6 for each episode do
7   Initialize environment and state  $S(0)$ ;
8   for each time step  $t = 1, 2, \dots, T$  do
9     Choose action  $x(t)$  using the  $\epsilon$ -greedy policy;
10    Execute  $x(t)$ : update  $Q(t+1)$  by (3),  $\mathcal{X}(t)$  by (9),
      and age  $\theta(t+1)$  by (11), then obtain  $S(t+1)$ ;
11    Use  $\mathcal{X}(t)$  and the trained DNN (Step 1) to
      compute the cross-entropy loss in (7);
12    Compute the cost function in (8);
13    Store transition  $(S(t), x(t), C(t), S(t+1))$  in
      replay memory;
14    Sample a mini-batch from replay memory and
      update DQN parameters via Q-learning;
15    Set  $S(t) \leftarrow S(t+1)$ ;

/* Output */
16 Return the trained DNN and DQN for real-time
  inference (Fig. 1);

```

cluding the age of information $\theta(t)$ in the DQN state (see Fig. 4). The implementation details of both the DQN and DNN modules are provided in the next section.

We summarize our method in Alg. 1. After initialization of system parameters and neural network models, loading a dataset, two main steps are performed. In Step (1), the DNN is trained in a supervised manner using labeled beam data, where mini-batches are sampled, predictions are generated, and parameters are updated via cross-entropy backpropagation. In Step (2), the DQN is trained: at each time step, an action is selected using the ϵ -greedy strategy, i.e., choosing a random action with probability ϵ for exploration or the best action with probability $1 - \epsilon$ for exploitation. The chosen action updates the system state and the DQN cost, while transitions are stored in replay memory and sampled to update the Q-network. Finally, the trained DNN and DQN are returned for dynamic (online) sensing and beam prediction.

4. NUMERICAL RESULTS AND DISCUSSIONS

This section presents the simulation results to evaluate the performance of the proposed method (see Fig. 1).¹ For com-

¹The source code is available at https://github.com/AZakeri94/DQN_ML_Dynamic_Sensing-Communication.git

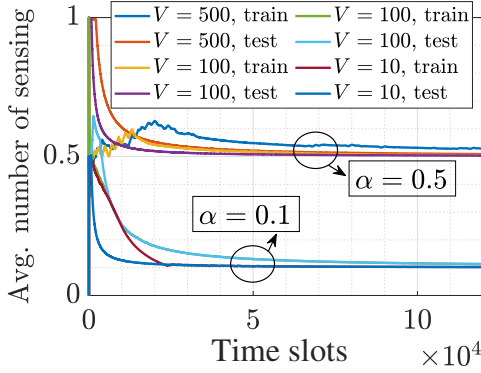


Fig. 2. Satisfaction of the average constraint (2b) by the proposed DQN method for different V and α

parison, we consider two benchmarks. The first is *randomized sensing*, where the sensing decision $x(t)$ is chosen randomly while still satisfying the constraint in (2b). The second is the *without age* case, where the age of samples is excluded from the DQN state $S(t)$. Performance is evaluated using Top- k accuracy, defined as the probability that the optimal beam lies within the top- k predicted beams.

For the sensing modality, we use position data from the DeepSense dataset (Scenario 5) [1]. The beam prediction DNN module follows the pretrained model in [7], implemented as a two-layered multilayer perceptron. The input dimension matches the feature size, with each hidden layer containing 1024 neurons and ReLU activation at each layer. Training is performed using the Adam optimizer with a learning rate of 0.01 and a batch size of 32.

For the DQN module, we employ a three-layer fully connected network with 128 neurons in each hidden layer. The input dimension matches the DQN state size, and the output dimension corresponds to the action space. Training is performed with a discount factor of 0.99999, a learning rate of 0.001 using the Adam optimizer, a batch size of 64, and a replay memory of 50,000. The model is trained for 300 epochs, each consisting of 400 iterations.

We first demonstrate in Fig. 2 that the proposed DQN-based sensing module satisfies the average constraint in (2b) for different values of the normalized sensing budget $\alpha \triangleq \frac{c^{\max}}{c}$ with $\alpha \in [0, 1]$. The results confirm that the cost term introduced in (7) effectively enforces the average constraint. It is further observed that larger values of V yield slower convergence to the constraint limit α .

Fig. 3 shows the Top-1, Top-2, and Top-3 beam prediction accuracies for different methods, including the full-sensing case where sensing is performed at every time slot. Two main observations arise: (i) reducing the sensing frequency by 50% leads to almost no loss in beam prediction accuracy, and (ii) incorporating sensing data age into the DQN state yields a notable performance gain. The latter underscores the importance of the age of information [19] in improving remote in-

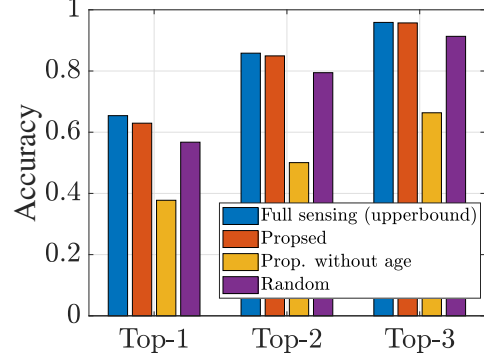


Fig. 3. Accuracy comparison between different methods for the normalized sensing budget $\alpha = 0.5$ and $V = 100$

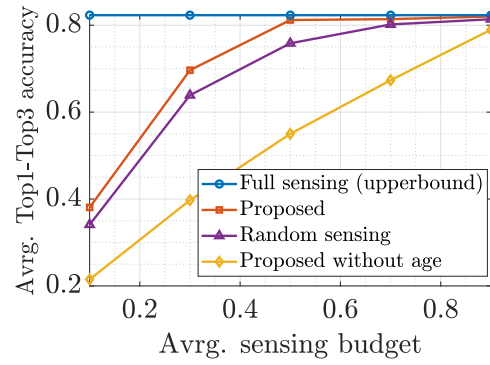


Fig. 4. Average Top1-Top3 accuracy vs. the sensing budget.

ference for machine learning-driven communication tasks.

Fig. 4 compares the average accuracy, defined as $(\text{Top-1} + \text{Top-2} + \text{Top-3})/3$, of different sensing methods versus the sensing budget. The results also highlight the benefit of incorporating the age of the last sensing data into the DQN state, showing consistent gains over randomized sensing. Moreover, our method converges to the full-sensing performance when $\alpha \geq 0.5$, corresponding to a 50% reduction in sensing cost.

5. CONCLUSIONS

We proposed a dynamic sensing-aided communication framework that combines Lyapunov optimization with deep reinforcement learning for beam prediction under an average sensing cost budget. The framework integrates a DNN for beam prediction with a DQN for sensing decisions, where the DQN state incorporates the virtual queue, the DNN input, and the age of the most recent sensing data. A customized cost function was designed by combining cross-entropy with the Lyapunov drift. Evaluations on the DeepSense Scenario 5 dataset showed that the proposed method reduces sensing cost by nearly 50% without sacrificing beam prediction accuracy, and that incorporating sensing data age yields significant gains. Future work will extend the approach to general multi-modal sensing scenarios to further assess its effectiveness.

6. ACKNOWLEDGMENT

This work was supported by the Research Council of Finland through the 6G Flagship Program (Grant No. 369116), project DIRECTION (Grant No. 354901), project DYNAMICS (Grant No. 367702), and project S6GRAN (Grant No. 370561); supported in part by CHIST-ERA through the project PASSIONATE (Grant No. 359817); and in part by the HORIZON-JU-SNS-2023 project INSTINCT.

7. REFERENCES

- [1] A. Alkhateeb, G. Charan, T. Osman, A. Hredzak, J. Morais, U. Demirhan, and N. Srinivas, "Deepsense 6G: A large-scale real-world multi-modal sensing and communication dataset," *IEEE Commun. Mag.*, vol. 61, no. 9, pp. 122–128, Sep. 2023.
- [2] A. Ali, N. Gonzalez-Prelcic, R. W. Heath, and A. Ghosh, "Leveraging sensing at the infrastructure for mmWave communication," *IEEE Commun. Mag.*, vol. 58, no. 7, pp. 84–89, Jul. 2020.
- [3] J. Gu, B. Salehi, D. Roy, and K. R. Chowdhury, "Multimodality in mmWave MIMO beam selection using deep learning: Datasets and challenges," *IEEE Commun. Mag.*, vol. 60, no. 11, pp. 36–41, 2022.
- [4] K. Patel and R. W. Heath, "Harnessing multimodal sensing for multi-user beamforming in mmWave systems," *IEEE Trans. Wireless Commun.*, vol. 23, no. 12, pp. 18725–18739, Dec. 2024.
- [5] S. Imran, G. Charan, and A. Alkhateeb, "Environment semantic communication: Enabling distributed sensing aided networks," *IEEE Open J. Commun. Soc.*, vol. 5, pp. 7767–7786, Dec. 2024.
- [6] M. B. Mollah, H. Wang, M. A. Karim, and H. Fang, "Multi-modality sensing in mmWave beamforming for connected vehicles using deep learning," *IEEE Trans. on Cogn. Commun. Netw.*, Early Access, 2025.
- [7] G. Charan, T. Osman, A. Hredzak, N. Thawdar, and A. Alkhateeb, "Vision-position multi-modal beam prediction using real millimeter wave datasets," in *Proc. IEEE Wireless Commun. and Networking Conf.*, pp. 2727–2731, Austin, TX, USA, Apr. 2022.
- [8] A. Oliveira, D. Suzuki, S. Bastos, I. Correa, and A. Klautau, "Machine learning-based mmwave MIMO beam tracking in V2I scenarios: Algorithms and datasets," in *Proc. IEEE Latin-American Conf. on Commun. (LATINCOM)*, pp. 1–5, Medellin, Colombia, Dec. 2024.
- [9] Y. Cui, J. Nie, X. Cao, T. Yu, J. Zou, J. Mu, and X. Jing, "Sensing-assisted high reliable communication: A transformer-based beamforming approach," *IEEE J. Sel. Topics Signal Process.*, vol. 18, no. 5, pp. 782–795, Jul. 2024.
- [10] B. Salehihi Kouei, *Leveraging Deep Learning on Multimodal Sensor Data for Wireless Communication: From mmWave Beamforming to Digital Twins*. PhD thesis, Northeastern University, 2024.
- [11] A. Zakeri, N. T. Nguyen, A. Alkhateeb, and M. Juntti, "Constrained multimodal sensing-aided communications: A dynamic beamforming design," in *Proc. IEEE Global Commun. Conf.*, Accepted, 2025. Available at: <https://arxiv.org/abs/2505.10015>.
- [12] C. Zheng, J. He, C. G. Kang, G. Cai, Z. Yu, and M. Debbah, "M2BeamLLM: Multimodal sensing-empowered mmWave beam prediction with large language models," *arXiv preprint arXiv:2506.14532*, Jun. 2025.
- [13] S. Jiang, G. Charan, and A. Alkhateeb, "LiDAR aided future beam prediction in real-world millimeter Wave V2I communications," *IEEE Commun. Lett.*, vol. 12, no. 2, pp. 212–216, Feb. 2023.
- [14] B. Shi, M. Li, M.-M. Zhao, M. Lei, and L. Li, "Multimodal deep learning empowered millimeter-Wave beam prediction," in *Proc. IEEE Veh. Technol. Conf.*, pp. 1–6, Singapore, Jun. 2024.
- [15] K. Zhang, W. Yu, H. He, S. Song, J. Zhang, and K. B. Letaief, "Multimodal deep learning-empowered beam prediction in future THz ISAC systems," *arXiv preprint arXiv:2505.02381*, May 2025.
- [16] Y. M. Park, Y. K. Tun, W. Saad, and C. S. Hong, "Resource-efficient beam prediction in mmWave communications with multimodal realistic simulation framework," *arXiv preprint arXiv:2504.05187*, Apr. 2025.
- [17] M. J. Neely, *Stochastic network optimization with application to communication and queueing systems*. Synth. Lectures Commun. Netw., vol. 3, no. 1, pp. 1–211, Jan. 2010.
- [18] A. Zakeri, M. Moltafet, M. Leinonen, and M. Codreanu, "Minimizing the AoI in resource-constrained multi-source relaying systems: Dynamic and learning-based scheduling," *IEEE Trans. Wireless Commun.*, vol. 23, no. 1, pp. 450–466, Jan. 2024.
- [19] S. Kaul, R. Yates, and M. Gruteser, "Real-time status: How often should one update?," in *Proc. IEEE Int. Conf. on Computer Commun.*, pp. 2731–2735, Orlando, FL, USA, Mar. 2012.