

Meta-Hierarchical Reinforcement Learning based Beamforming for Near-field Multi-user Communications

Yang Chen, Saba Al-Rubaye, *Life Senior Member, IEEE*, Antonios Tsourdos, *Member, IEEE*, Hongyu Li, *Member, IEEE*, Xu Shi, *Member, IEEE*, Zhuangkun Wei, Zeyu Liu, *Member, IEEE*, Lawrence Baker, Colin Gillingham

Abstract—Extremely Large Antenna Array Systems (ELAAs) promise to provide ultra-high data rates for multi-user 6G Communications, leveraging massive antenna elements to enhance spatial diversity and spectral efficiency. Unfortunately, considering the propagation of spherical waves in ELAAs, near-field beamforming suffers from substantial performance degradation, diminished spatial focusing precision, considerable computational overhead, and increased system costs. To address these challenges, this paper develops an intelligent hybrid beamforming scheme to empower near-field multi-user THz communications with meta-hierarchical reinforcement learning (MHRL). On the one hand, a novel codebook is designed with Hierarchical Deep Q Network (H-DQN), dramatically releasing the potential of near-field beamforming in ELAAs. To explain the availability of our codebook, the quantization bound is analyzed along with the corresponding computational complexity. On the other hand, we further develop a meta-beamforming framework by integrating our intelligent codebook to meta-reinforcement learning (MRL)-based hybrid beamformer, accelerating the CSI acquisition for multi-user communication. For deep investigation, the rate loss of our scheme is calculated against the ideal continuous beamformer, which reveals that the quantization error has been properly bounded. Simulation results show that the proposed algorithm can efficiently manage the near-field beam with higher beamforming gain, and the proposed beamforming framework exhibits improved spectral efficiency, as well as increased robustness to ELAAs.

Index Terms—Near-field, Codebook Design, Multi-user THz Communication, Meta-Hierarchical Reinforcement Learning.

I. INTRODUCTION

THE high frequency band [1], such as Millimeter Wave (mmWave) and Tera-Hertz (THz), is poised to address the escalating demands for ultra-high-speed data transmission in next-generation wireless communication [2]. However, high-frequency bands inherently suffer from substantial path loss

[3], calling for the development of innovative mitigation techniques. Extremely Large Antenna Array Systems (ELAAs) offer a promising solution by exploiting the shorter wavelength of high-frequency signals to achieve ultra-dense antenna element configurations [4]. As a result, the integration of THz communication with ELAAs is anticipated to be a potential candidate for future wireless networks [5], [6].

ELAAs primarily increases the aperture size, which in turn significantly expands the spatial region of near-field, owing to the increase of the Rayleigh distance [7]. The Rayleigh distance, which distinguishes the boundary between near-field and far-field, is proportional to the square of the array aperture normalized by the wavelength. Therefore, the range of near-field can be extended to several hundreds of meters in the THz communications of ELAAs, occupying substantial area of the typical cell [8]. In the far-field scenario, the wireless channel is held by a planar wave, where the array response vector is a linear function of the antenna index [9]. In contrast, a near-field channel has to be established as a spherical wave, where the array response vector varies nonlinearly with the antenna index [10]. Consequently, the enlarged range and the unique characteristics of the near-field channel motivates us to conduct advanced transmitter design which could better adapt to the propagation environment and thus provide enhanced performance [11].

A. Prior Works

Owing to the enlarged dimension of the antenna array, which further identifies unique physical characteristics in the near-field, it is necessary to redesign lots of transmission strategies, such as beamforming [12], beam training [13], and beam tracking [14]. Currently, near-field beamforming is attracting more and more attention due to its growing potentials in both single-user and multi-user communications.

Early research on near-field beamforming primarily focused on single-user scenarios due to their relatively straightforward system models and lower computational complexity. For instance, Wang et al. [15] proposed a closed-form solution for true-time delayers (TTDs) based hybrid beamforming to mitigate the spatial-wideband effect in single-user based near-field communication. However, the inherent high peak-to-average power ratio (PAPR) will significantly degrade the transmission efficiency, particularly at high frequencies. To further reduce the hardware cost and energy consumption of large-scale antenna arrays at high frequencies, holographic metasurface

This work is supported by Future Aviation Security Solutions Industrial PHD Partnerships (FASS IPP) under grant P14868.

Yang Chen, Saba Al-Rubaye, Antonios Tsourdos, and Zeyu Liu are with the Faculty of Engineering and Applied Sciences (FEAS), Cranfield University, Milton Keynes, MK43 0AL, UK (e-mail: {yang.chen, s.alrubaye, a.tsourdos, zeyu.liu}@cranfield.ac.uk)

Hongyu Li is with the Internet of Things Thrust, The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, Guangdong 511400, China (e-mail: hongyuli@hkust-gz.edu.cn).

Xu Shi is with the Department of Electronic Engineering, Tsinghua University, Beijing 100084, China (e-mail: 13671193792@163.com)

Zhuangkun Wei is with the Department of Computing, Imperial College London, London, SW7 2BX, UK (e-mail: zhuangkun.wei@imperial.ac.uk)

Lawrence Baker and Colin Gillingham are with NCC Group, Manchester, M3 3AQ, UK (e-mail: {lawrence.baker, colin.gillingham}@nccgroup.com)

antennas (HMAs) emerged as a promising solution. Li et al. [16] derived an optimal closed-form digital precoder as well as analog HMA weighting matrices to minimize the total transmit power for each HMAs while maintaining the signal to interference plus noise ratio (SINR) constraint of single-user scenario. Although these methods achieve sharp beam focusing in the near-field, the closed-form digital precoder is not suitable for the future dynamic radio environment. In response, dynamic metasurface antennas (DMAs) is proposed to adapt their configuration to the real-time radio environment. For example, Kim et al. [17] developed a closed-form solution to optimize DMAs configuration in the single-user multiple-input single-output (MISO) systems, maintaining a robust signal-to-noise ratio (SNR) gain throughout the predefined range. However, the optimal solution only considered the propagation region in the far-field instead of near-field.

Building on single-user methodologies, subsequent research extended beamforming frameworks to multi-user scenarios, aiming to maximize system capacity while mitigating inter-user interference. Firstly, Liu et al. [18] developed a near-field equivalent model to design multifocus array with augmented accuracy. Nevertheless, this model can only design near-field multifocus (NFMF) antenna array with specified amplitude and phase features. Furthermore, Zhang et al. [19] proposed efficient beam focusing optimization algorithms on sum-rate maximization for the fully digital and hybrid beamforming architecture, and dynamic metasurface antenna. Their work provides a new degree of freedom (DoF) to mitigate interference in both angle and distance domains, but this optimal strategy remained highly dependent on the architecture of antenna array and precoder. In [20], Wu et al. extended the concept of the new DoF to location division multiple access (LDMA) by designing the spherical-domain codebook for the near-field multiple access system, exploiting more spatial resources in the distance domain. However, LDMA lacks the transition design for the boundary of near-far field. Regarding the far-to-near successive interference cancellation order, Zuo et al. [21] extended the deployment of spherical-domain codebook to near-field non-orthogonal multiple access systems (NF-NOMA), offering higher flexibility for multiple access transmission. Nonetheless, most beamforming solutions are derived from traditional model-based optimization methods, which significantly limit the efficiency and scalability.

With the evolution of machine learning, deep learning (DL) and reinforcement learning (RL) are becoming powerful tools to proceed the large-dimension research of 6G communications, especially imperfect CSI-based beamforming in ELAAs [22], [23]. For example, Yuan et al. [24] developed a meta-learning based adaptive algorithm to avoid the mismatch issue between training and test channel during beamforming. To implement near-field channel estimation for hybrid beamforming, Jang et al. [25] proposed a neural network-aided joint optimization of the beamformer and localization function. However, most deep learning architectures are not suitable for the acquisition of instantaneous channel state information (CSI) due to poor adaptability to fast-varying near-field channels [8]. Inspired from LDMA [20], Monemi et al. [26] developed CSI-independent TD3-DRL optimizer to

realize smart spot-beamforming (SBF) in the Fresnel zone. Nevertheless, traditional DRL algorithms are constrained in ELAAs due to inefficient exploration strategy.

B. Our Contributions

Inspired by the potential of hierarchical reinforcement learning (HRL) to address complex decision-making challenges [27], this article aims to map the instantaneous CSI of near-field THz communication in ELAAs with the Hierarchical Deep Q Network (H-DQN) [28]. With regard to mismatched distribution between codebook design and hybrid beamformer, H-DQN is upgraded to Meta H-DQN (MH-DQN) to jointly optimize codebook design and interference canceling in multi-user communications. The motivation and principal benefits of employing MH-DQN in this paper are:

- **Scalability:** The hierarchical structure allows efficient handling of the large-scale CSI inherent in ELAA-based THz systems.
- **Adaptability:** Meta-learning enables the agent to quickly adapt to environment changes (e.g., user mobility, new scatterers, NLoS conditions) without retraining from scratch.
- **Joint Optimization:** MH-DQN supports joint optimization of codebook selection and interference management for multi-user communications, outperforming conventional model-based or single-level RL approaches.

On account of these insights, the main contributions of this paper are summarized as follows:

- An intelligent codebook is designed for efficient near-field beamforming. In detail, this codebook captures the near-field characteristic of angle and distance to achieve an equivalent two-dimensional optimization. In that case, H-DQN, as a goal-conditioned HRL, is deployed to adaptively explore the desired angle optimizer and distance optimizer, substantially enhancing the availability of codebook in the near-field channels with ELAAs.
- A comprehensive analysis of the quantization bounds of the proposed codebook is conducted to improve beamforming precision in ELAAs. By minimizing quantization errors through adaptive temporal abstraction, the proposed quantization bound ensures the accuracy and efficiency of real-time beam alignment. Moreover, the derived computational complexity demonstrates the superior scalability and robustness of the proposed codebook in diverse near-field channels.
- With the intention of applying our intelligent codebook to hybrid beamforming, we propose a novel MH-DQN framework which combines HRL with Meta-Learning as a two-step method. Specifically, meta-train is first conducted for multi-user codebook design as the initial step. Next, we construct a three-level HRL architecture to optimize spatial multiplexing, at the last two layers of which meta-test is deployed to accelerate near-field beamforming adaptation.
- Under this MH-DQN framework, the rate loss of the proposed beamforming strategy is derived from the difference between the perfect infinite-resolution codebook and the proposed codebook. Consequently, compared to

previous research [29]–[31], the above rate loss can be bounded to a much lower value, theoretically defending the accuracy of the proposed beamforming framework.

C. Organization and Notation

The remaining contexts are organized as follows. Section. II introduces the hybrid beamforming system and the near-field channel model. Section. III develops HRL based codebook design and then calculates the theoretical range of codebook quantization. Section. IV extends the proposed codebook to the meta-framework and completes the corresponding performance analysis. Section. V offers simulation results. Section. VI draws the conclusions.

Notations: Boldface letters represent the vectors and matrices. $\mathcal{CN}(\cdot, \cdot)$ means complex normal distribution. $(\cdot)^*$, $\|\cdot\|$, $(\cdot)^T$, $(\cdot)^H$, $(\cdot)^{-1}$ represent desired value, Euclidean norm of vector, transpose, conjugate transpose, and inverse of matrix respectively. $|\cdot|$ means the norm of its complex argument. $\mathbb{I}\{\cdot\}$, $E[\cdot]$, and $\Gamma(\cdot)$ denote indicator function, expectation, and gamma function. $O(\cdot)$ and $\Omega(\cdot)$ represent asymptotic maximum and minimum growth.

II. SYSTEM MODEL

A. System Model and Problem Formulation

In this paper, we consider a narrowband downlink system consisting of a multi-antenna base station (BS), and U single-antenna users. Specifically, the BS applies a hybrid beamforming architecture, where there are N_{RF} RF chains connected to the N_{BS} -antenna ELAA via N_{BS} phase shifters with $U \leq N_{RF} \leq N_{BS}$, as illustrated in Fig. 1. For simplicity, we assume $N_{RF} = U$ all through the paper. The received signal is formulated for all UEs at downlink as

$$\mathbf{y}_D = \mathbf{H}\mathbf{F}_A\mathbf{F}_D\mathbf{s} + \mathbf{n}, \quad (1)$$

where $\mathbf{y}_D = [y_1, y_2, \dots, y_U]^T$ represents the received signal from all UEs, $\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_U]^H \in \mathbb{C}^{U \times N_{BS}}$ denotes channel matrix between BS and UEs. During each transmission, the signal vector $\mathbf{s} \in \mathbb{C}^{U \times 1}$ satisfying $E[\mathbf{s}\mathbf{s}^H] = \frac{P}{U}\mathbf{I}$ is propagated in the desired direction or area with the support of the digital beamformer $\mathbf{F}_D = [\mathbf{f}_{D,1}, \dots, \mathbf{f}_{D,U}] \in \mathbb{C}^{N_{RF} \times U}$ and the analog beamformer $\mathbf{F}_A \in \mathbb{C}^{N_{BS} \times N_{RF}}$. Consequently, $\mathbf{n} \sim \mathcal{CN}(0, \sigma_N^2 \mathbf{I}_U)$ denotes the additive white Gaussian noise (AWGN) with variance σ_N^2 .

For performance evaluation, the total spectral efficiency of our system is calculated as

$$R = \sum_{i=1}^U R_i = \sum_{i=1}^U \log_2 \left(1 + \frac{\frac{P}{U} |\mathbf{h}_i^H \mathbf{F}_A \mathbf{f}_{D,i}|^2}{\sigma_N^2 + \sum_{l \neq i} \frac{P}{U} |\mathbf{h}_i^H \mathbf{F}_A \mathbf{f}_{D,l}|^2} \right), \quad (2)$$

where P represents the total power consumption of BS, $\mathbf{f}_{D,i}$ is the i^{th} column of $\mathbf{F}_D$.

In this work, our main objective is to jointly design the digital and analog beamformer to maximize the spectral efficiency (SE) in (2). This leads to the following problem, where $\mathcal{F}_A$ is the index set from non-zero entries of $\mathbf{F}_A$.

$$\begin{aligned} \text{(P1): } & \max_{\mathbf{F}_A, \mathbf{F}_D} R \\ \text{s.t. } & \|\mathbf{F}_A \mathbf{f}_{D,u}\|^2 = 1, \quad \forall u \\ & |[\mathbf{F}_A]_{i,j}|^2 = N_{BS}^{-1}, \quad \forall (i, j) \in \mathcal{F}_A \end{aligned} \quad (3)$$

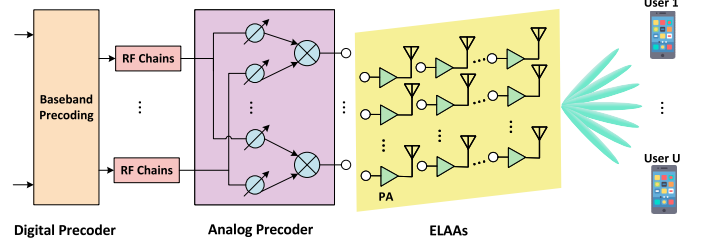


Fig. 1. Hybrid Beamforming in ELAAs.

B. Near-Field Channel Model

The wireless channel can be exactly classified as near-field model and a far-field model, which are identified with the Rayleigh distance $r_{RD} = \frac{2D^2}{\lambda}$, where λ and D denote the wavelength and antenna aperture, respectively. Within the Rayleigh distance, the area will be covered by a near-field channel, where the radiation pattern is shaped with spherical wavefront. In contrast, the far-field channel model will take over the area out of Rayleigh distance, assuming the direction of propagation as a planar wave. In the THz-MIMO 6G communications, the Rayleigh distance has been substantially expanded due to the extremely large carrier frequency and the massive antennas. As a result, the radiation pattern cannot be regarded totally as the planar-wave in ELAAs anymore. To characterize the near-field ELAAs, the wireless channel is modeled with the assumption of scatter points as

$$\begin{aligned} \mathbf{h}_u^{near} = & \sqrt{N_{BS}} \alpha_{LoS} \mathbf{b}(r_0, \theta_0, \phi_0) \\ & + \sqrt{\frac{N_{BS}}{L}} \sum_{l=1}^L \alpha_{l,NLoS} \mathbf{b}(r_l, \theta_l, \phi_l), \end{aligned} \quad (4)$$

where $\mathbf{b}(r_0, \theta_0, \phi_0)$ and $\mathbf{b}(r_l, \theta_l, \phi_l)$ denote the beamforming vector in the line-of-sight (LoS) path and the nonline-of-sight (NLoS) path, respectively. $\alpha_{LoS} = \sqrt{\frac{\kappa}{\kappa+1}}$ means the complex gain for the LoS path, while that gain for the NLoS path is expressed as $\alpha_{l,NLoS} \sim \mathcal{CN}(0, \sigma_l^2)$ with $\sigma_l = \sqrt{\frac{1}{\kappa+1}}$. The Rician factor κ means the power ratio between LoS and NLoS path.

For generalization, BS will firstly configure uniform planar array (UPA) as Fig. 2, where the near-field beamforming vector can be calculated as

$$\begin{aligned} \mathbf{b}(r_0, \theta_0, \phi_0) = & \frac{1}{\sqrt{N_{BS}}} \left[e^{-\frac{2\pi j}{\lambda} (r_0^{(-\bar{N})} - r_0)}, \dots, \right. \\ & \left. e^{-\frac{2\pi j}{\lambda} (r_0^{(n)} - r_0)}, \dots, e^{-\frac{2\pi j}{\lambda} (r_0^{(\bar{N})} - r_0)} \right]^T, \end{aligned} \quad (5)$$

where $\bar{N} = (\bar{N}_1, \bar{N}_2)$, $N_{BS} = N_1 \times N_2$, $\bar{N}_1$ and $\bar{N}_2$ are defined as the maximum element of UPA, along with $\bar{N}_i = \frac{N_i-1}{2}$. $r_0^{(n)}$ represents the distance between UE and the (n_1, n_2) -element of UPA, while r_0 represents the distance between UE and the center of UPA. In terms of the spherical-wave propagation model, $r_0^{(n)}$ can be rewritten as

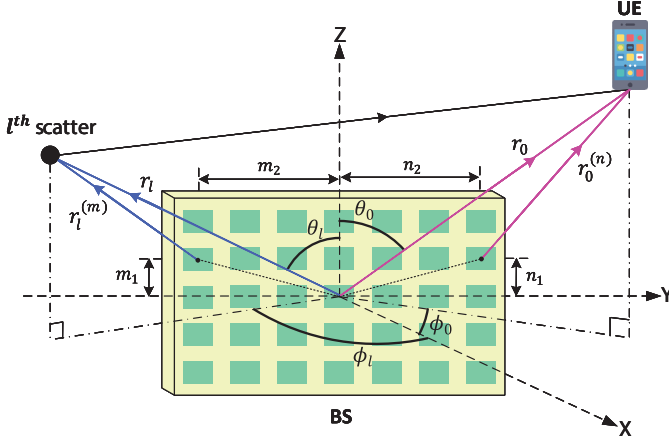


Fig. 2. Near-field Channel Model in UPA System.

$$\begin{aligned}
 r_0^{(n)} &= r_0^{(n_1, n_2)} \\
 &= \sqrt{(r_0 \sin \theta_0 \cos \phi_0)^2 + (r_0 \cos \theta_0 - n_1 d)^2} \\
 &\quad + (r_0 \sin \theta_0 \sin \phi_0 - n_2 d)^2} \\
 &\approx r_0 - n_1 d \cos \theta_0 - n_2 d \sin \theta_0 \sin \phi_0 \\
 &\quad + \frac{n_1^2 d^2}{2r_0} (1 - \cos^2 \theta_0) + \frac{n_2^2 d^2}{2r_0} (1 - \sin^2 \theta_0 \sin^2 \phi_0) \\
 &\quad - \frac{n_1 n_2 d^2 \cos \theta_0 \sin \theta_0 \sin \phi_0}{r_0},
 \end{aligned} \tag{6}$$

where θ_0 and ϕ_0 denote the elevation angle and the azimuth angle of UE, and d is the spacing between the adjacent antenna elements. Likewise, in Fig. 2, r_l denotes the distance between the l^{th} scatterer and the array center, instead $r_l^{(m)}$ denotes the distance between the l^{th} scatterer and the m -array element. Moreover, θ_l means the elevation angle of the l^{th} scatterer, and $r_l^{(m)}$, $b(r_l, \theta_l, \phi_l)$ can be calculated with the similar method as (6) and (5), respectively. Furthermore, this scenario will be simplified to the uniform linear array (ULA) system in Section. III as a one-dimensional special case of the UPA.

The subsequent objective of the codebook design is to optimize the analog beamforming vector by maximizing the beamforming gain based on the instantaneous channel modeled as (4). This leads to the following optimization problem, where $\mathbf{f}_u$ means the codeword for the u^{th} user from the codebook $\mathcal{F}$, and $\mathbf{f}_{u,n}$ denotes the codeword element for the n^{th} antenna of the u^{th} user.

$$\begin{aligned}
 \text{(P2): } & \max_{\mathbf{f}_u} |(\mathbf{h}_u^{near})^H \mathbf{f}_u|^2 \\
 \text{s.t. } & \mathbf{f}_u \in \mathcal{F} \\
 & |\mathbf{f}_{u,n}|^2 = N_{BS}^{-1}, \quad \forall n
 \end{aligned} \tag{7}$$

III. INTELLIGENT CODEBOOK-BASED BEAMFORMING

In this section, we propose a HRL framework to enable intelligent codebook-based beamforming for near-field ELAAs. Traditional model-based optimization methods typically relying on linear regression to update codebooks, which have to simplify or ignore the bilinear quadratic terms $\sin \theta \sin \phi$ and $\cos^2 \theta$ in (6) for the reduction of computational complexity, severely limit the potential capacity of the higher-dimensional array systems. In contrast, HRL benefits the high-dimensional codebook from enhanced adaptability and the

ability to explore the nonlinear relationship between distance, azimuth, and elevation, offering a more robust solution for nonconvex optimization, especially in UPA codebook-based beamforming.

The rest of this section is organized as follows. Firstly, we review the H-DQN framework and the corresponding update policy. Furthermore, we define the system model in Section. II as Semi-Markov Decision Process (S-MDP) [32], where the H-DQN framework is deployed to optimize an intelligent codebook. Next, the update policy of that intelligent codebook is presented along with the quantization order. Consequently, the computation complexity of this new framework is analyzed to evaluate the performance theoretically.

A. Hierarchical Deep Q Network

H-DQN [28] is the integration of HRL with Deep Q Network (DQN), aiming at resolving the sparse and delayed feedback of S-MDP. The framework of H-DQN is made up of two levels for decision-making, including meta-controller and controller. Meta-controller estimates the current subgoal, while controller takes actions to achieve that subgoal. Correspondingly, high-level Markov Decision Process (MDP) is defined as $(\mathcal{S}_{hi}, \mathcal{G}, \mathbf{P}_{hi}, \mathbf{R}_{hi}, \mathcal{T}_{hi}, \mathbf{H}_{hi})$ in meta-controller, while low-level MDP is defined as $(\mathcal{S}_{lo} \times \mathcal{G}, \mathcal{A}, \mathbf{P}_{lo}, \mathbf{R}_{lo}, \mathcal{T}_{lo}, \mathbf{H}_{lo})$ in controller. In detail, $\mathcal{G}$ and $\mathcal{A}$ mean the set of subgoals and actions with $g \in \mathcal{G}$ and $a \in \mathcal{A}$, respectively. Considering $s_{lo} \in \mathcal{S}_{lo}$ and $s_{hi} \in \mathcal{S}_{hi}$, the equivalent state $s \in \mathcal{S}$ of the original MDP can be described as a tuple $(s_{lo}, s_{hi}) \in \mathcal{S}_{lo} \times \mathcal{S}_{hi}$. Since the next state of low-level $s'_{lo} \in \mathcal{S}_{lo}$, the low-level transition $\mathcal{T}_{lo} : (s_{lo}, a, g) \rightarrow s'_{lo}$ is accomplished with the probability $\mathbf{P}_{lo}$ and the policy $\pi_{lo} : \mathcal{S}_{lo} \times \mathcal{G} \rightarrow \mathcal{A}$. Analogously, the high-level transition $\mathcal{T}_{hi} : (s_{hi}, g) \rightarrow s'_{hi}$ is achieved with the probability $\mathbf{P}_{hi}$ and the policy $\pi_{hi} : \mathcal{S}_{hi} \rightarrow \mathcal{G}$. As the evolution of DQN, H-DQN estimates Q value function from intrinsic reward $\varrho_{lo} \in \mathbf{R}_{lo}$ and extrinsic reward $\varrho_{hi} \in \mathbf{R}_{hi}$. The function $Q_{lo} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times \mathcal{G} \rightarrow \mathbb{R}$ denotes the low-level Q function as (8), while the function $Q_{hi} : (\mathcal{S} \times \mathcal{A} \times \mathcal{S})^K \times \mathcal{G} \rightarrow \mathbb{R}$ denotes the high-level Q function from the temporal abstraction of K -step low-level MDP as (9).

$$\begin{aligned}
 Q_{lo}(s, a, g) &= \max_{\pi_{lo}} E \left[\sum_{i=t}^{\infty} \gamma^{i-t} \varrho_{i,lo} \mid s_t = s, a_t = a, g_t = g, \pi_{lo} \right] \\
 &= \max_{\pi_{lo}} E \left[\varrho_{t,lo} + \gamma \max_{a_{t+1}} Q_{lo}(s_{t+1}, a_{t+1}, g) \mid s_t = s, \right. \\
 &\quad \left. a_t = a, g_t = g, \pi_{lo} \right]
 \end{aligned} \tag{8}$$

$$\begin{aligned}
 Q_{hi}(s, g) &= \max_{\pi_{hi}} E \left[\sum_{i=t}^{t+K} \varrho_{i,hi} + \gamma \max_{g'} Q_{hi}(s_{t+K}, g') \mid \right. \\
 &\quad \left. s_t = s, g_t = g, \pi_{hi} \right]
 \end{aligned} \tag{9}$$

where γ means the discount factor, s_t , a_t , and g_t denote the current state, action, and subgoal. Meanwhile, $\varrho_{i,lo}$ and $\varrho_{i,hi}$ mean the current intrinsic reward and the extrinsic reward, respectively. Given the current subgoal, K represents the number of iterations before the controller suspends, and g' denotes the goal of state s_{t+K} .

Owing to the current neural network parameters $\omega_{lo,i}$ and $\omega_{hi,i}$, the above Q function $Q_{lo}(s, g)$ and $Q_{hi}(s, a, g)$ can be approximated as the Q Network $Q_{lo}(s, g, \omega_{lo,i})$ and $Q_{hi}(s, a, \omega_{hi,i}, g)$. Furthermore, these neural networks are optimized by the minimization of loss function $L_{lo}(\omega_{lo,i})$ and $L_{hi}(\omega_{hi,i})$ as

$$L_{lo}(\omega_{lo,i}) = E_{(s,a,g,\varrho_{lo},s') \sim \mathcal{D}_{lo}} \left[\left(\varrho_{lo} - Q_{lo}(s, a, \omega_{lo,i}, g) + \gamma \max_{a'} Q_{lo}(s', a', \omega_{lo,i-1}, g) \right)^2 \right], \quad (10)$$

$$L_{hi}(\omega_{hi,i}) = E_{(s,g,\varrho_{hi},s') \sim \mathcal{D}_{hi}} \left[\left(\varrho_{hi} - Q_{hi}(s, g, \omega_{hi,i}) + \gamma \max_{g'} Q_{hi}(s', g', \omega_{hi,i-1}) \right)^2 \right], \quad (11)$$

where $\mathcal{D}_{lo}$ and $\mathcal{D}_{hi}$ mean the experience memory of the controller and meta-controller, and i represents the number of training iterations.

B. Hierarchical MDP Formulation for Near-field ELAAs

In UPA system, the near-field codebook is responsible for deducing the optimal codewords of several fields, such as the distance, elevation angle, and azimuth angle of the beamforming vector. However, this deduction is inherently complex due to the extremely large searching space, the complicated interaction between near-field channels and codebook, and the non-stationary nature of wireless communications. Moreover, the beam radiation is dominated by spherical wavefronts in the near-field, for which the near-field propagation is highly sensitive to the tiny deviation of distance and angle from the beam pattern. As a consequence, conventional flat MDP has to be optimized directly across a high-dimensional state-action space, bringing about inefficient exploration and long convergence time.

To smoothly enable codebook-based beamforming in near-field ELAAs, we propose to formulate the near-field codebook design as a novel Hierarchical MDP (H-MDP), which is a particular type of S-MDP involving high-level MDP and low-level MDP. The hierarchical framework decomposes the objective of near-field analog beamforming into several hints or subgoals, which release computational intractability and increase the efficiency of codebook design.

1) High-level MDP Formulation

High-level MDP plays the role of meta-controller, determining which dimension of the near-field codebook to optimize, including distance and angle. For the given beam pattern, this temporal abstraction will focus the limited exploration on the more crucial aspect of beamforming. For instance, the optimization of distance may be prioritized to multipath interference, while the optimal angle may be more significant for aligning the beam precisely towards the target receiver. The specific high-level MDP is formulated as follows.

Option Space: The option $\mathcal{S}_{hi}$ is the term for the primitive space of H-MDP, representing the extrinsic Channel State Information (CSI) as $\mathcal{S}_{hi} = \{r^*, \theta^*, \phi^*, \hat{h}_r, \hat{h}_i\}$, where $\hat{h}_r$

and $\hat{h}_i$ denote the real and imaginary parts of the abstract channel model $\hat{h}$, and r^*, θ^* , and ϕ^* mean the optimal distance, elevation angle, and azimuth angle to compose the best temporal beamforming vector.

Goal Space: The goal $g \in \mathcal{G}$ estimates the temporal dimension of codebook design from the binary goal space $\mathcal{G} = \{0, 1, 2\}$. Particularly, $g = 0$ implies the corresponding controller will concentrate on the design of optimal distance, while $g = 1$ and $g = 2$ imply the temporal optimization will focus on the design of elevation angle and azimuth angle separately in the low-level MDP.

Extrinsic Reward: According to the extracted value from the above goal space, we can formulate the extrinsic reward $\varrho_{i,h}$ as (12), where $\hat{r}_i, \hat{\theta}_i$, and $\hat{\phi}_i$ denote the expected codeword for distance, elevation angle, and azimuth angle of the i^{th} iteration.

$$\varrho_{i,h} = \begin{cases} \mathbb{I} \left\{ \left| \hat{h}^H \mathbf{b}(\hat{r}_i, \theta^*, \phi^*) \right| > \left| \hat{h}^H \mathbf{b}(r^*, \theta^*, \phi^*) \right| \right\}, & g = 0 \\ \mathbb{I} \left\{ \left| \hat{h}^H \mathbf{b}(r^*, \hat{\theta}_i, \phi^*) \right| > \left| \hat{h}^H \mathbf{b}(r^*, \theta^*, \phi^*) \right| \right\}, & g = 1 \\ \mathbb{I} \left\{ \left| \hat{h}^H \mathbf{b}(r^*, \theta^*, \hat{\phi}_i) \right| > \left| \hat{h}^H \mathbf{b}(r^*, \theta^*, \phi^*) \right| \right\}, & g = 2 \end{cases} \quad (12)$$

2) Low-level MDP Formulation

Once the high-level MDP has been completed, the low-level MDP will be triggered to optimize the chosen dimension of codebook. For example, if meta-controller has desired to upgrade the pattern of distance field, the distance controller will execute fine-grained action to adjust the design of distance field with the maximization of the intrinsic reward. Therefore, the hierarchical decomposition simplifies the 3-D optimization to an equivalent two-step optimization, enabling fine-control over the near-field beam pattern. The specific low-level MDP is formulated as follows.

State Space: Compared to the state of original MDP, the state space $\mathcal{S}_l$ is a kind of goal-conditioned area in low-level MDP, associated to the goal space of high-level MDP. With regard to the near-field beamforming, the state space is identified to the intrinsic CSI as $\mathcal{S}_{lo} = \{\hat{h}_r, \hat{h}_i\}$.

Action Space: To bundle with meta-controller, the low-level agent has been classified as distance-controller, elevation-controller, and azimuth-controller. Correspondingly, the action spaces for distance, elevation angle, and azimuth angle are derived as $\mathcal{A}_r = \{0, 1, \dots, N_{Q_r}\}$, $\mathcal{A}_\theta = \{0, 1, \dots, N_{Q_\theta}\}$, and $\mathcal{A}_\phi = \{0, 1, \dots, N_{Q_\phi}\}$, where N_{Q_r} , N_{Q_θ} , and N_{Q_ϕ} denote the quantization of distance, elevation angle, and azimuth angle respectively.

Intrinsic Reward: Correspondingly, the intrinsic reward is also designed for distance-controller, elevation-controller, and azimuth-controller as (13), (15), and (17) where $\varrho_{i,r}$ is the intrinsic distance reward, and $\varrho_{i,\theta}$, $\varrho_{i,\phi}$ are the intrinsic elevation reward and the intrinsic azimuth reward. Given the constraint of distance and angle as $r \in [r_{\min}, r_{\max}]$, $\theta \in [\theta_{\min}, \theta_{\max}]$, and $\phi \in [\phi_{\min}, \phi_{\max}]$, the predicted distance, elevation angle, and azimuth angle are calculated by (14), (16), and (18), where $a_r \in \mathcal{A}_r$, $a_\theta \in \mathcal{A}_\theta$, and $a_\phi \in \mathcal{A}_\phi$. In addition, the optimal distance, elevation angle, and azimuth angle are determined from memories, satisfying $(r^*, \theta^*, \phi^*) = \operatorname{argmax} \left| \hat{h}^H \mathbf{b}(r, \theta, \phi) \right|^2$.

$$\varrho_{i,r} = \left| \hat{\mathbf{h}}^H \mathbf{b}(\hat{r}_i, \theta^*, \phi^*) \right|^2 \quad (13)$$

$$\hat{r}_i = r_{\min} + a_r \frac{r_{\max} - r_{\min}}{N_{Q_r}} \quad (14)$$

$$\varrho_{i,\theta} = \left| \hat{\mathbf{h}}^H \mathbf{b}(r^*, \hat{\theta}_i, \phi^*) \right|^2 \quad (15)$$

$$\hat{\theta}_i = \theta_{\min} + a_\theta \frac{\theta_{\max} - \theta_{\min}}{N_{Q_\theta}} \quad (16)$$

$$\varrho_{i,\phi} = \left| \hat{\mathbf{h}}^H \mathbf{b}(r^*, \theta^*, \hat{\phi}_i) \right|^2 \quad (17)$$

$$\hat{\phi}_i = \phi_{\min} + a_\phi \frac{\phi_{\max} - \phi_{\min}}{N_{Q_\phi}} \quad (18)$$

C. Intelligent Codebook Design

In terms of H-MDP from Section. III-B, we propose a novel hierarchical intelligent codebook as Fig. 3 to solve the joint assignment of spatial parameters of $\mathbf{f}_u$ in (7), such as distance and angle. The whole End-to-End (E2E) architecture [33] is divided into the channel estimation network [34] and the intelligent codebook. Initially, BS collect the environmental state $\mathcal{S}_t = \{\mathbf{y}_W^t, \bar{\mathbf{h}}_W^t, \mathbf{s}, \mathbf{h}_p^t\}$, where $\mathbf{y}_W^t$ and $\bar{\mathbf{h}}_W^t$ denote the pilot signal of the uplink and the prediction of the channel separately from the sliding window with time slot W . In addition, $\mathbf{s}$ means pilot sequence, and $\mathbf{h}_p^t$ is the feedback CSI from the mentioned sliding window. The channel estimation network is then implemented to predict the CSI of the downlink as $\hat{\mathbf{h}}$ with the environmental state observed at BS. Given the estimated channel $\hat{\mathbf{h}}$ and the best temporal codeword pair (r^*, θ^*, ϕ^*) , the beamforming module will calculate the extrinsic reward for high-level MDP as (12). With the reception of the predicted CSI, intelligent codebook will play the role of beamforming dictionary, where the optimal codewords will polish up the temporal beamforming strategy.

Remark 1 (Benefit of the Proposed E2E Architecture): Although conventional codebook-based beamforming indeed reduces system overhead by avoiding explicit CSI measurement, they are inherently limited in dynamic or near-field ELAAs, leading to suboptimal beam selections and frequently repeated beam searches. In contrast, the proposed E2E reinforcement learning framework strategically combines low-complex partial implicit CSI with a lightweight intelligent adaptive codebook, which significantly enhances overall beamforming performance along with high efficiency and robustness against environmental dynamics and channel imperfections. Therefore, this hybrid approach effectively balances overhead and operational performance in practical ELAAs.

Under the H-MDP framework, the hierarchical decision-making process has been launched to dynamically search for the beamforming vector from multiple dimensions simultaneously instead of looking up a static table. Aroused from the H-DQN structure [28], the intelligent codebook is established with meta-controller, distance controller, elevation controller, and azimuth controller. To tighten the solution space, the meta-controller is built up with the DQN-based abstraction network, which determines the priority of codeword optimization as the goal $\mathcal{G}$. Referring to (9), the Q-value function can be customized as (19) in the meta-controller, where $\hat{\mathbf{h}}_t$ denotes the predicted CSI at the t^{th} timestep. As a result, the learning

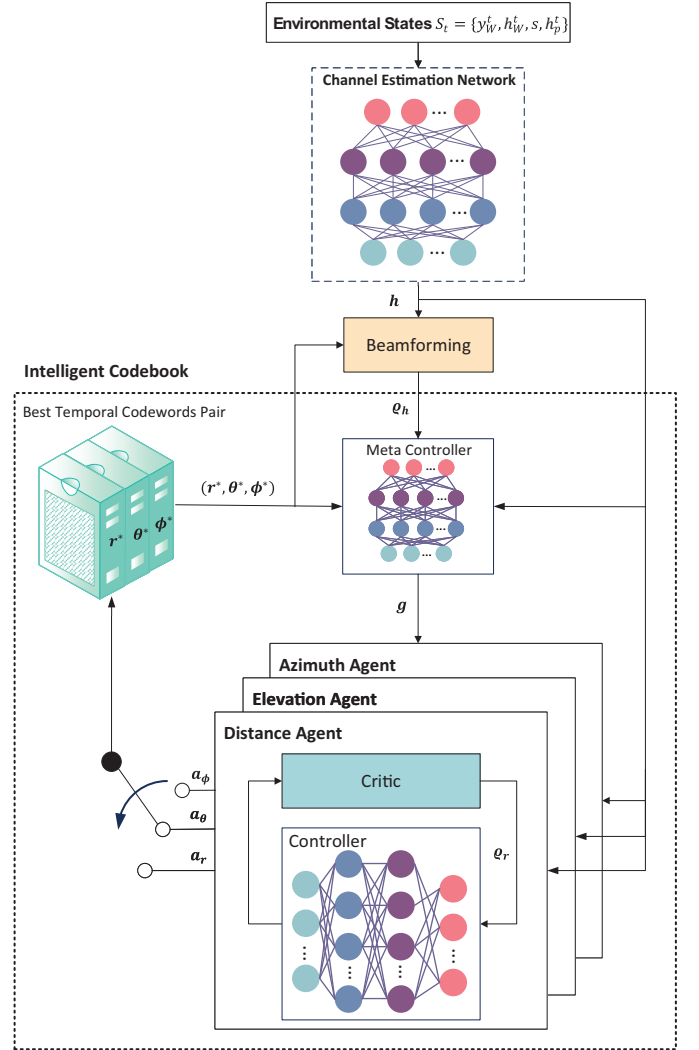


Fig. 3. Hierarchical Intelligent Codebook Design.

parameter ω_{hi} of meta-controller can be updated with the loss function of $Q_{hi}(\hat{\mathbf{h}}, g)$ as (20) and (21), where $\hat{\mathbf{h}}$ and $\hat{\mathbf{h}}'$ mean the current and the next predicted CSI separately.

$$Q_{hi}(\hat{\mathbf{h}}, g) = \max_{\pi_{hi}} E \left[\sum_{i=t}^{t+K} \varrho_{i,hi} + \gamma \max_{g'} Q_{hi}(\hat{\mathbf{h}}_{t+K}, g') \mid \hat{\mathbf{h}}_t = \hat{\mathbf{h}}, g_t = g, \pi_{hi} \right] \quad (19)$$

$$L_{hi}(\omega_{hi,i}) = E_{(\hat{\mathbf{h}}, g, \varrho_{hi}, \hat{\mathbf{h}}') \sim \mathcal{D}_{hi}} \left[\left(\varrho_{hi} - Q_{hi}(\hat{\mathbf{h}}, g, \omega_{hi,i}) \right)^2 + \gamma \max_{g'} Q_{hi}(\hat{\mathbf{h}}', g', \omega_{hi,i-1}) \right] \quad (20)$$

$$\omega_{hi,i+1} = \omega_{hi,i} - \alpha_{hi} \nabla L_{hi}(\omega_{hi,i}) \quad (21)$$

Once the meta-controller has selected a goal, the corresponding controller will take over to achieve that goal. In the context of codebook-based beamforming, the low-level MDP is set up as distance controller, elevation controller, and azimuth controller with the Q-value function as (22), (23), and (24), where a_r , a_θ , and a_ϕ represent the index of codeword for

distance, elevation angle, and azimuth angle, and $a_{t,r}$, $a_{t,\theta}$, and $a_{t,\phi}$ denote the index of them at the t^{th} timestep. Similar with (8), π_r , π_θ , and π_ϕ denote the policy of distance controller, elevation controller, and azimuth controller.

$$Q_r(\hat{\mathbf{h}}, a_r, g) = \max_{\pi_r} E \left[\varrho_{t,r} + \gamma \max_{a_{t+1,r}} Q_{lo}(\hat{\mathbf{h}}_{t+1}, a_{t+1,r}, g) \mid \hat{\mathbf{h}}_t = \hat{\mathbf{h}}, a_{t,r} = a_r, g_t = g, \pi_r \right] \quad (22)$$

$$Q_\theta(\hat{\mathbf{h}}, a_\theta, g) = \max_{\pi_\theta} E \left[\varrho_{t,\theta} + \gamma \max_{a_{t+1,\theta}} Q_{lo}(\hat{\mathbf{h}}_{t+1}, a_{t+1,\theta}, g) \mid \hat{\mathbf{h}}_t = \hat{\mathbf{h}}, a_{t,\theta} = a_\theta, g_t = g, \pi_\theta \right] \quad (23)$$

$$Q_\phi(\hat{\mathbf{h}}, a_\phi, g) = \max_{\pi_\phi} E \left[\varrho_{t,\phi} + \gamma \max_{a_{t+1,\phi}} Q_{lo}(\hat{\mathbf{h}}_{t+1}, a_{t+1,\phi}, g) \mid \hat{\mathbf{h}}_t = \hat{\mathbf{h}}, a_{t,\phi} = a_\phi, g_t = g, \pi_\phi \right] \quad (24)$$

Given the policy of distance controller, elevation controller, and azimuth controller, their critic loss function will be derived as (25), (26), and (27) separately, by the feedback of which the learning parameters of each controller can be efficiently updated. $\omega_{r,i}$, $\omega_{\theta,i}$, and $\omega_{\phi,i}$ denote the learning parameters of distance controller, elevation controller, and azimuth controller at the i^{th} iteration, and a'_r , a'_θ , a'_ϕ denote the predicted index of codewords for distance, elevation angle, and azimuth angle.

$$L_r(\omega_{r,i}) = E_{(\hat{\mathbf{h}}, a_r, g, \varrho_r, \hat{\mathbf{h}}') \sim \mathcal{D}_r} \left[\left(y_r - Q_r(\hat{\mathbf{h}}, a_r, \omega_{r,i}, g) \right)^2 \right] \\ y_r = \varrho_r + \gamma \max_{a'_r} Q_r(\hat{\mathbf{h}}', a'_r, \omega_{r,i-1}, g) \quad (25)$$

$$L_\theta(\omega_{\theta,i}) = E_{(\hat{\mathbf{h}}, a_\theta, g, \varrho_\theta, \hat{\mathbf{h}}') \sim \mathcal{D}_\theta} \left[\left(y_\theta - Q_\theta(\hat{\mathbf{h}}, a_\theta, \omega_{\theta,i}, g) \right)^2 \right] \\ y_\theta = \varrho_\theta + \gamma \max_{a'_\theta} Q_\theta(\hat{\mathbf{h}}', a'_\theta, \omega_{\theta,i-1}, g) \quad (26)$$

$$L_\phi(\omega_{\phi,i}) = E_{(\hat{\mathbf{h}}, a_\phi, g, \varrho_\phi, \hat{\mathbf{h}}') \sim \mathcal{D}_\phi} \left[\left(y_\phi - Q_\phi(\hat{\mathbf{h}}, a_\phi, \omega_{\phi,i}, g) \right)^2 \right] \\ y_\phi = \varrho_\phi + \gamma \max_{a'_\phi} Q_\phi(\hat{\mathbf{h}}', a'_\phi, \omega_{\phi,i-1}, g) \quad (27)$$

The detailed design of intelligent codebook has been listed as Algorithm. 1, where we explore each DQN agent with ϵ -greedy function [35].

Remark 2 (Specialization to ULA Case): The proposed H-DQN-based codebook design framework is presented for the general case of UPA, accounting for distance, azimuth and elevation domains. If we consider the simplification to ULA scenario, the system model and the codebook design procedure can be straightforwardly adapted. In particular, the elevation angle and its quantization are omitted, and the action space reduces from the tuple (r, θ, ϕ) to (r, ϕ) . Correspondingly, the array manifold and the beamforming vector are described only by the distance and azimuth domains, and the HRL agent optimizes beamforming gain over this reduced parameter

space. This demonstrates the flexibility of the proposed method to accommodate different array architectures by appropriately configuring the state and action representations.

Algorithm 1 Intelligent Codebook Design Procedure

Initialization: experience $\{\mathcal{D}_{hi}, \mathcal{D}_r, \mathcal{D}_\theta, \mathcal{D}_\phi\}$, and parameters $\{\omega_{hi}, \omega_r, \omega_\theta, \omega_\phi\}$ for meta-controller, distance controller, elevation controller, and azimuth controller, along with exploration probability $\epsilon = 1$.

for $i = 1$ **to** N **do**

Obtain the current and next CSI $\hat{\mathbf{h}}$ and $\hat{\mathbf{h}}'$.

$g \leftarrow \epsilon$ -greedy($\hat{\mathbf{h}}, \mathcal{G}, \epsilon, Q_{hi}$)

for $t = 1$ **to** H **do**

$R_{hi} \leftarrow 0, \hat{\mathbf{h}}_0 \leftarrow \hat{\mathbf{h}}$.

while goal g not reached **do**

if $g = 0$ **then**

$a_r \leftarrow \epsilon$ -greedy($\{\hat{\mathbf{h}}, g\}, \mathcal{A}_r, \epsilon, Q_r$)

Execute a_r to obtain extrinsic reward ϱ_h and distance reward ϱ_r .

Store transition $(\{\hat{\mathbf{h}}, g\}, a_r, \varrho_r, \{\hat{\mathbf{h}}', g\})$ in $\mathcal{D}_r$

Calculate $L_r(\omega_{r,t-1})$ as (25) and update distance agent by: $\omega_{r,t} \leftarrow \omega_{r,t-1} - \alpha_r \nabla L_r(\omega_{r,t-1})$

else if $g = 1$ **then**

$a_\theta \leftarrow \epsilon$ -greedy($\{\hat{\mathbf{h}}, g\}, \mathcal{A}_\theta, \epsilon, Q_\theta$)

Execute a_θ to obtain extrinsic reward ϱ_h and angle reward ϱ_θ .

Store transition $(\{\hat{\mathbf{h}}, g\}, a_\theta, \varrho_\theta, \{\hat{\mathbf{h}}', g\})$ in $\mathcal{D}_\theta$

Calculate $L_\theta(\omega_{\theta,t-1})$ as (26) and update angle agent by: $\omega_{\theta,t} \leftarrow \omega_{\theta,t-1} - \alpha_\theta \nabla L_\theta(\omega_{\theta,t-1})$

else

$a_\phi \leftarrow \epsilon$ -greedy($\{\hat{\mathbf{h}}, g\}, \mathcal{A}_\phi, \epsilon, Q_\phi$)

Execute a_ϕ to obtain extrinsic reward ϱ_h and angle reward ϱ_ϕ .

Store transition $(\{\hat{\mathbf{h}}, g\}, a_\phi, \varrho_\phi, \{\hat{\mathbf{h}}', g\})$ in $\mathcal{D}_\phi$

Calculate $L_\phi(\omega_{\phi,t-1})$ as (27) and update angle agent by: $\omega_{\phi,t} \leftarrow \omega_{\phi,t-1} - \alpha_\phi \nabla L_\phi(\omega_{\phi,t-1})$

end if

$R_{hi} \leftarrow R_{hi} + \varrho_{hi}; \hat{\mathbf{h}} \leftarrow \hat{\mathbf{h}}'$.

Initialize the next CSI $\hat{\mathbf{h}}'$

end while

Store transition $(\hat{\mathbf{h}}_0, g, R_{hi}, \hat{\mathbf{h}}')$ in $\mathcal{D}_{hi}$

Update ω_{hi} of meta-controller by (20) and (21)

$g \leftarrow \epsilon$ -greedy($\hat{\mathbf{h}}, \mathcal{G}, \epsilon, Q_{hi}$)

end for

Anneal ϵ

end for

D. Analyses of Quantization Bound and Computational Complexity

To further analyze the property of Algorithm. 1, we first investigate the quantization bound of the proposed codebook, and then calculate all computational complexity of Algorithm. 1.

1) Quantization Bound

Considering episodic fixed-Horizon reinforcement learning in Algorithm. 1, the sample complexity of them has been bounded in [36]. In Section. III-C, it has been decomposed

by H-MDP [37] with the deployment in near-field THz beamforming. As a result, Theorem. 1 is proposed to update the quantization bound of intelligent codebook design.

Theorem 1 (Quantization Bound of Intelligent Codebook): Given any beamforming vector $\mathbf{b}(r_{Q_1}, \theta_{Q_2}, \phi_{Q_3})$ from the intelligent codebook of Algorithm. 1 with $Q_i = 0, 1, \dots, N_{Q_i} - 1$, the quantized fraction of the i^{th} codeword N_{Q_i} is bounded by (28) with the PAC guarantee for (ϵ, δ) -accurate.

$$\frac{H_{hi}^2 \ln(\frac{1}{\delta + c_{1,hi}})}{c_{2,l} H_{lo}^2 \ln(\frac{1}{\delta})} \leq N_{Q_i} \leq \frac{c_{2,hi} H_{hi}^2 \ln(\frac{1}{\delta})}{H_{lo}^2 \ln(\frac{1}{\delta + c_{1,lo}})} \quad (28)$$

where H_{hi} and H_{lo} denote the high-level and low-level horizons of timesteps, so $H = H_{hi} H_{lo}$ in Algorithm. 1. In addition, $c_{1,hi}$ and $c_{1,lo}$ are positive constants along with $c_{2,hi}, c_{2,lo} \in (0, N_{BS}]$.

Proof: detailed in Appendix. A. ■

Remark 3 (Analysis on Theorem. 1): The performance of quantization exhibits an asymptotic independence on the number of available channel models. Instead, with regard to the PAC guarantee (ϵ, δ) , the quantized fraction is subject to probability δ rather than accuracy ϵ . Furthermore, the quantization bound is significantly influenced by the ratio of the timestep horizon between the high-level and low-level MDP, which highlights the role of temporal abstraction in enhancing quantization efficiency.

2) Computational Complexity

From Appendix. B, the runtime complexity of Algorithm. 1 is yielded to $\max\{TC_1, TC_2\}$, where

$$TC_1 = \min \left\{ \tilde{O} \left(\frac{c_{2,hi}^2 N_{BS}^2 |\mathcal{G}|^2 H_{hi}^3 H_{lo}}{\epsilon^2} \ln \frac{1}{\delta} \right), \tilde{O} \left(\frac{c_{2,hi} c_{2,lo} N_{BS}^2 |\mathcal{G}|^2 N_{Q_i} H_{hi}^3 H_{lo}}{\epsilon^2} \ln \frac{1}{\delta} \right) \right\},$$

and

$$TC_2 = \min \left\{ \tilde{O} \left(\frac{c_{2,lo}^2 N_{BS}^2 |\mathcal{G}|^2 N_{Q_i}^2 H_{hi}^3 H_{lo}}{\epsilon^2} \ln \frac{1}{\delta} \right), \tilde{O} \left(\frac{c_{2,hi} c_{2,lo} N_{BS}^2 |\mathcal{G}|^2 N_{Q_i} H_{hi}^3 H_{lo}}{\epsilon^2} \ln \frac{1}{\delta} \right) \right\}.$$

Compared to the time complexity of the model-based optimization method for hierarchical codebook design in [38], Algorithm. 1 exhibits lower computational complexity when $\frac{\lambda N_{BS}^2}{r_{\min}} > \max\{TC_3, TC_4\}$, where $r_{\min}$ denotes the minimum BS serving distance, $TC_3 = \min \left\{ \frac{c_{2,hi}^2 H_{hi}^2 \ln(1/\delta)}{\epsilon^2}, \frac{c_{2,hi} c_{2,lo} N_{Q_i} H_{lo}^2 \ln(1/\delta)}{\epsilon^2} \right\}$, and $TC_4 = \min \left\{ \frac{c_{2,lo}^2 N_{Q_i}^2 H_{lo}^2 \ln(1/\delta)}{\epsilon^2}, \frac{c_{2,hi} c_{2,lo} N_{Q_i} H_{hi}^2 \ln(1/\delta)}{\epsilon^2} \right\}$. Due to the flexibility of the hyperparameter in TC_3 and TC_4 , the condition above is easy to satisfied.

In contrast, the space complexity is dependent on the size of minibatch $\mathcal{B}$ and the architecture of neural network. Therefore, the total space complexity has been derived as $O[N_{BS} |H_{hi}| |\mathcal{G}| + (N_{BS} + |\mathcal{G}|) H_{lo} N_{Q_i} + |\mathcal{B}| (N_{BS} + |\mathcal{G}| + N_{Q_i})]$. According to [36], the equivalent space complexity of DQN can be derived as $O[(H_{hi} H_{lo} N_{Q_i}^2 + |\mathcal{B}|) N_{BS} + |\mathcal{B}| N_{Q_i}^2]$,

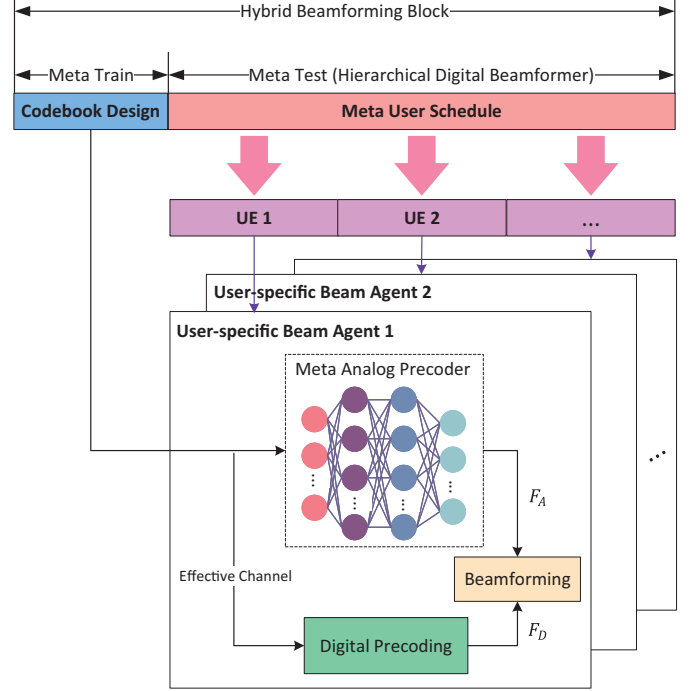


Fig. 4. Meta Hybrid Beamforming Procedure.

which is obviously outperformed by the space complexity of the proposed algorithm. The detailed calculation has been provided in Appendix. B.

IV. META HYBRID BEAMFORMING FOR MULTI-USER THZ SYSTEM

In this section, we extend the proposed beamforming in Section. III-C to the framework of Meta Hybrid Beamforming (MHB) for multiuser communications, which complete beam management and interference cancellation from two separate timescales. First, we outline the whole procedure of MHB, then meta-analog beamformer and hierarchical digital beamformer are developed accordingly. Consequently, the performance improvement of MHB is analysed for multi-user THz systems.

A. Procedure of MHB Scheme

For the sake of the stationary CSI distribution in multi-user communications, meta-reinforcement learning (MRL) [39] has been deployed to two-timescale hybrid beamforming [40], [41] as Fig. 4. The hybrid beamforming block is split into meta-analog beamforming (MAB) and hierarchical digital beamforming (HDB) respectively, where the first is responsible for meta-train, and the latter serves as meta-test. As the heading of MHB, MAB is devoted to designing the meta-intelligent codebook as analog beamformer, where the effective channel can be determined to conduct digital beamforming during meta-test. In contrast, as the successive timeslot of MHB, the user-specific beam agent is optimized with HRL during HDB. The detailed scheme will be elaborated on in the following subsection.

B. Meta Analog Beamformer

Advanced to Section. III-C, we combine intelligent codebook design with meta-learning to acquire the general channel knowledge map (CKM) [42]. The meta task is defined as the

set of predicted channel model $\mathcal{T}_{\hat{H}} = \{\hat{\mathbf{h}}_1, \hat{\mathbf{h}}_2, \dots, \hat{\mathbf{h}}_U\}$ from the feedback of all UEs. For user-specific feedback, the HMDP will be redesigned across all UEs with the loss function as (29)(30)(31)(32), where $\hat{\mathbf{h}}_i$ is randomly extracted from $\mathcal{T}_{\hat{H}}$.

$$L_{T_i}(\omega_h) = E_{(\hat{\mathbf{h}}_i, g, \varrho_h, \hat{\mathbf{h}}'_i) \sim \mathcal{D}_h} \left[\left(\varrho_h - Q_h(\hat{\mathbf{h}}_i, g, \omega_{h,t}) + \gamma \max_{g'} Q_h(\hat{\mathbf{h}}'_i, g', \omega_{h,t-1}) \right)^2 \right] \quad (29)$$

$$L_{T_i}(\omega_{r,t}) = E_{(\hat{\mathbf{h}}_i, a_r, g, \varrho_r, \hat{\mathbf{h}}'_i) \sim \mathcal{D}_r} \left[\left(y_r - Q_r(\hat{\mathbf{h}}_i, a_r, \omega_{r,t}, g) \right)^2 \right] \quad (30)$$

$$L_{T_i}(\omega_{\theta,t}) = E_{(\hat{\mathbf{h}}_i, a_{\theta}, g, \varrho_{\theta}, \hat{\mathbf{h}}'_i) \sim \mathcal{D}_{\theta}} \left[\left(y_{\theta} - Q_{\theta}(\hat{\mathbf{h}}_i, a_{\theta}, \omega_{\theta,t}, g) \right)^2 \right] \quad (31)$$

$$L_{T_i}(\omega_{\phi,t}) = E_{(\hat{\mathbf{h}}_i, a_{\phi}, g, \varrho_{\phi}, \hat{\mathbf{h}}'_i) \sim \mathcal{D}_{\phi}} \left[\left(y_{\phi} - Q_{\phi}(\hat{\mathbf{h}}_i, a_{\phi}, \omega_{\phi,t}, g) \right)^2 \right] \quad (32)$$

Motivated by Model-Agnostic Meta-Learning (MAML) [39], the design pattern of intelligent codebook is upgraded to MAB in Algorithm. 2.

Algorithm 2 Meta Analog Beamforming Scheme

Initialization: set of predicted channel model $\mathcal{T}_{\hat{H}} = \{\hat{\mathbf{h}}_1, \hat{\mathbf{h}}_2, \dots, \hat{\mathbf{h}}_U\}$, and parameters $\{\omega_h, \omega_r, \omega_{\theta}, \omega_{\phi}\}$ for meta-controller, distance controller, and angle controller, learning rate $\alpha_h, \alpha_r, \alpha_{\theta}, \alpha_{\phi}, \beta_h, \beta_r, \beta_{\theta}, \beta_{\phi}$.

for $i = 1$ **to** N **do**

 Sample $T_i \sim \mathcal{T}_{\hat{H}}$.

for all T_i **do**

 Calculate $L_{T_i}(\omega_h)$ and goal g with (29) and Algorithm. 1.

$\omega'_{h_i} \leftarrow \omega_{h_i} - \alpha_h \nabla_{\omega_{h_i}} L_{T_i}(\omega_{h_i})$

while goal g not reached **do**

 Evaluate $L_{T_i}(\omega_r), L_{T_i}(\omega_{\theta}), L_{T_i}(\omega_{\phi})$ with respect to fixed horizons by (30)(31)(32) and Algorithm. 1.

$\omega'_{r_i} \leftarrow \omega_{r_i} - \alpha_r \nabla_{\omega_{r_i}} L_{T_i}(\omega_{r_i})$

$\omega'_{\theta_i} \leftarrow \omega_{\theta_i} - \alpha_{\theta} \nabla_{\omega_{\theta_i}} L_{T_i}(\omega_{\theta_i})$

$\omega'_{\phi_i} \leftarrow \omega_{\phi_i} - \alpha_{\phi} \nabla_{\omega_{\phi_i}} L_{T_i}(\omega_{\phi_i})$

end while

end for

$\omega_h \leftarrow \omega_h - \beta_h \nabla_{\omega_h} \sum_{T_i \sim \mathcal{T}_{\hat{H}}} L_{T_i}(\omega'_{h_i})$

$\omega_r \leftarrow \omega_r - \beta_r \nabla_{\omega_r} \sum_{T_i \sim \mathcal{T}_{\hat{H}}} L_{T_i}(\omega'_{r_i})$

$\omega_{\theta} \leftarrow \omega_{\theta} - \beta_{\theta} \nabla_{\omega_{\theta}} \sum_{T_i \sim \mathcal{T}_{\hat{H}}} L_{T_i}(\omega'_{\theta_i})$

$\omega_{\phi} \leftarrow \omega_{\phi} - \beta_{\phi} \nabla_{\omega_{\phi}} \sum_{T_i \sim \mathcal{T}_{\hat{H}}} L_{T_i}(\omega'_{\phi_i})$

end for

C. Hierarchical Digital Beamforming

As illustrated in Section. IV-B, the MAB of the u^{th} user can be defined as $\mathbf{w}_u = \mathbf{b}(r_u^*, \theta_u^*, \phi_u^*)$ with the optimal distance and angle $r_u^*, \theta_u^*, \phi_u^*$ from the meta-test of Algorithm. 2. Then, BS can extend the analog beamforming to all UEs as $\mathbf{F}_A = \{\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_U\}$, so that the effective channel model can be established as

$$\bar{\mathbf{h}}_u = \mathbf{F}_A^H \hat{\mathbf{h}}_u. \quad (33)$$

Through the effective channel, BS is capable of formulating the digital beamforming as (34) by zero-forcing (ZF) [43], where $\bar{\mathbf{H}}_T = [\bar{\mathbf{h}}_1, \bar{\mathbf{h}}_2, \dots, \bar{\mathbf{h}}_U]$ is defined as the effective channel of all UEs. The diagonal matrix $\mathbf{\Lambda}$ denotes the power allocation for all UEs with $\|\mathbf{F}_A \mathbf{f}_{D,u}\|^2 = 1$, where $\mathbf{f}_{D,u}$ denotes the u^{th} column of $\mathbf{F}_D$.

$$\mathbf{F}_D = \bar{\mathbf{H}}_T^H (\bar{\mathbf{H}}_T \bar{\mathbf{H}}_T^H)^{-1} \mathbf{\Lambda} \quad (34)$$

To simultaneously optimize digital beamforming for all UEs under hybrid constraints, the hierarchical framework as in Fig. 4 constructs user-specific beam agents whose actions are coordinated through the joint optimization of a common digital precoder at each time slot, ensuring that inter-user coupling is coherently managed by ZF objectives rather than by independent learning. In light of the main objective (3), we have updated the MDP¹ by defining the user-specific rewards, their Q-function, and loss function as (35)(36)(37) to manage the update of controller for each user, where $\hat{\mathbf{F}}_A$ denotes the estimated analog beamforming with the meta-initialization from Algorithm. 2, and $\hat{\mathbf{f}}_{D,i}$ can be derived from $\hat{\mathbf{F}}_A$ with (34). Under this circumstance, Meta User Scheduler can be updated by

$$\varrho_{uh} = \sum_{i=1}^U \hat{R}_i = \sum_{i=1}^U \log_2 \left(1 + \frac{\frac{P}{U} |\mathbf{h}_i^H \hat{\mathbf{F}}_A \hat{\mathbf{f}}_{D,i}|^2}{\sigma_N^2 + \sum_{l \neq i} \frac{P}{U} |\mathbf{h}_i^H \hat{\mathbf{F}}_A \hat{\mathbf{f}}_{D,l}|^2} \right), \quad (35)$$

$$Q_{uh}(\hat{\mathbf{H}}_T, u) = \max_{\pi_{uh}} E \left[\sum_{i=t}^{t+KM} \varrho_{i,uh} + \gamma \max_{u'} Q_{uh}(\hat{\mathbf{H}}_{T_t+KM}, u') \mid \hat{\mathbf{H}}_{T_t} = \hat{\mathbf{H}}_T, u_t = u, \pi_{uh} \right], \quad (36)$$

$$L_{uh}(\omega_{uh,i}) = E_{(\hat{\mathbf{H}}_T, u, \varrho_{uh}, \hat{\mathbf{H}}'_T) \sim \mathcal{D}_{uh}} \left[\left(\varrho_{uh} - Q_{uh}(\hat{\mathbf{H}}_T, u, \omega_{uh,i}) + \gamma \max_{u'} Q_{uh}(\hat{\mathbf{H}}'_T, u', \omega_{uh,i-1}) \right)^2 \right], \quad (37)$$

$$\omega_{uh,i+1} = \omega_{uh,i} - \alpha_{uh} \nabla_{\omega_{uh,i}} L_{uh}(\omega_{uh,i}). \quad (38)$$

Accordingly, HDB is devised as Algorithm. 3 to accelerate the efficiency of meta-test for MHB.

D. Performance Analysis of MHB with Multi-user Systems

Since LoS is dominant in THz systems, the performance of MHB will be evaluated with single-path channel. With the assumption of perfect channel knowledge, the lower bound of achievable rate is figured out with Infinite-Resolution Codebook (IRC) in the first place. In addition, the performance gap is further investigated between MAB and IRC, ensuring that MAB is approximate to the asymptotical solution of IRC.

¹It is worth noting that the MDP components of the MAB and HDB are fundamentally identical, since both rely on the same state features (CSI, user location), action space (beam selection from the codebook), reward metrics (SE or beamforming gain), and channel-driven transitions. The main difference is more structural. Particularly, MAB employs meta-reinforcement learning to accelerate codebook adaptation across tasks, whereas HDB introduces an additional ultra-high-level MDP to coordinate between analog and digital domains, thereby extending the hierarchy without altering the underlying MDP definition.

Algorithm 3 Hierarchical Digital Beamforming Scheme

Initialization: user group $\mathcal{G}_{uh} = \{0, 1, \dots, U\}$, experience $\{\mathcal{D}_{uh}, \mathcal{D}_h, \mathcal{D}_r, \mathcal{D}_\theta, \mathcal{D}_\phi\}$, and parameters $\{\omega_{uh}, \omega_h, \omega_r, \omega_\theta, \omega_\phi\}$ for user scheduler, meta-controller, distance controller, and angle controller, along with exploration probability $\epsilon = 1$, current and next CSI $\hat{H}$ and $\hat{H}'$ for all UEs. $\{\omega_h, \omega_r, \omega_\theta, \omega_\phi\}$ is transferred from Algorithm. 2.

for $i = 1$ **to** N **do**

$u \leftarrow \epsilon\text{-greedy}(\hat{H}_T, \mathcal{G}_{uh}, \epsilon, Q_{uh})$

for $j = 1$ **to** M **do**

$R_{uh} \leftarrow 0, \hat{H}_{T_0} \leftarrow \hat{H}_T.$

$g \leftarrow \epsilon\text{-greedy}(\hat{H}_u, \mathcal{G}, \epsilon, Q_h)$

for $t = 1$ **to** H **do**

$R_h \leftarrow 0, \hat{h}_0 \leftarrow \hat{h}_u.$

while goal g not reached **do**

Update $\omega_r, \omega_\theta, \omega_\phi$, then obtain user-specific reward ϱ_{uh} , extrinsic reward ϱ_h and intrinsic reward $\varrho_r, \varrho_\theta, \varrho_\phi$ as Algorithm. 1.

$R_h \leftarrow R_h + \varrho_h; \hat{h}_u \leftarrow \hat{h}_u'.$

Initialize the next CSI of the u^{th} UE $\hat{h}_u'$

end while

Store transition $(\hat{h}_0, g, R_h, \hat{h}_u')$ in $\mathcal{D}_h$

Update ω_h of meta-controller by (20)(21)

$R_{uh} \leftarrow R_{uh} + \varrho_{uh}; g \leftarrow \epsilon\text{-greedy}(\hat{h}_u, \mathcal{G}, \epsilon, Q_h)$

end for

Store transition $(\hat{h}_0, u, R_{uh}, \hat{h}_u')$ in $\mathcal{D}_{uh}$

Update ω_{uh} of user scheduler by (37)(38)

$u \leftarrow \epsilon\text{-greedy}(\hat{H}_T, \mathcal{G}_{uh}, \epsilon, Q_{uh})$

end for

Anneal ϵ

end for

1) Lower Bound of Asymptotic Capacity

For simplicity, the single-path channel is defined as (39) in the UPA, where α_u denotes the complex gain of the LoS path for each UE, and $\mathbf{b}(r_u, \theta_u, \phi_u)$ means the beamforming vector of the u^{th} user in the near-field.

$$\mathbf{h}_u = \sqrt{N_{BS}} \alpha_u \mathbf{b}(r_u, \theta_u, \phi_u) \quad (39)$$

Since the perfect CSI and IRC have been assumed, the analog beamformer is equivalent to the set of beamforming vectors for all UEs as $\mathbf{F}_A = \mathbf{B} = [\mathbf{b}(r_1, \theta_1, \phi_1), \dots, \mathbf{b}(r_U, \theta_U, \phi_U)]$. Meanwhile, ZF is applied to design the digital beamformer as (34). Under these circumstances, the lower bound of achievable rate can be derived as Theorem 2 for each UE.

Theorem 2 (Lower Bound of Achievable Rate): Given the single-path channel model is adopted for all UEs with perfect CSI, the achievable rate can be lower bounded as (40) for each UE, where $G(\mathbf{B}) = 4 \left(\frac{\sigma_{\max}^2(\mathbf{B})}{\sigma_{\min}^2(\mathbf{B})} + \frac{\sigma_{\min}^2(\mathbf{B})}{\sigma_{\max}^2(\mathbf{B})} + 2 \right)^{-1}$ with minimum and maximum singular values $\sigma_{\min}^2(\mathbf{B})$ and $\sigma_{\max}^2(\mathbf{B})$, and $\frac{P}{\sigma_N^2}$ denotes the signal-to-noise ratio (SNR) of each UE.

$$R_u \geq \log_2 \left(1 + \frac{PN_{BS}|\alpha_u|^2 G(\mathbf{B})}{\sigma_N^2 U} \right) \quad (40)$$

Proof: detailed in Appendix. C. ■

2) Rate Loss over MHB

As claimed by Proposition 4 of [29] and Theorem. 2, $\lim_{N \rightarrow \infty} R_u = R_u^*$ is achieved by $R_u^* = \log_2 \left(1 + \frac{PN_{BS}|\alpha_u|^2}{\sigma_N^2 U} \right)$. On that account, the performance of MHB can be defended with the lower rate loss between IRC and MAB as Theorem. 3.

Theorem 3 (Upper Bound of Rate Loss over MAB): Recalling R_u is the achievable rate for each UE from IRC, the averaged rate loss over MAB is defined as $\overline{\Delta R_u} = E[R_u - R_u^M]$, where R_u^M denotes the rate achieved by MAB for each UE. Given N_Q is the quantization of intelligent codebook, $\overline{\Delta R_u}$ is upper bounded by (41), where $\bar{\alpha} = E[\alpha_u]$.

$$\overline{\Delta R_u} \leq \log_2 \left[\frac{1 + \frac{P}{U\sigma_N^2} N_{BS} \bar{\alpha} \left(1 + \frac{U-1}{N} \right) N_Q^{-\frac{2}{U-1}}}{\min_{\mathbf{f}_i \in \mathcal{F}} \max_{\mathbf{f}_j \in \mathcal{F}} |\mathbf{f}_i^H \mathbf{f}_j|^2} \right] \quad (41)$$

Proof: detailed in Appendix. D. ■

Remark 4 (Analysis on Theorem. 3): Let $\overline{\mu_{\mathcal{F}}} = \min_{\mathbf{f}_i \in \mathcal{F}} \max_{\mathbf{f}_j \in \mathcal{F}} |\mathbf{f}_i^H \mathbf{f}_j|^2$, it is obvious that the constraint of $\overline{\mu_{\mathcal{F}}}$ depends on the design pattern of MAB. Additionally, since more N_Q results in lower $\overline{\Delta R_u}$, the rate loss over MAB also relies on the quantization size of codebook. Fortunately, the delicate MHB releases the potential performance of the large-scale codebook in THz ELAAs, so the rate loss can be perfectly bounded.

V. SIMULATION RESULTS

In this section, the performance of MHB is evaluated along with their hierarchical intelligent codebook through simulations. For a comprehensive assessment, the uniform UE distribution is considered for both the ULA and UPA systems.

A. Uniform Circular Distribution Scenario

The scenario is assumed as a single-cell THz communication, where one BS with ELAAs serves multiple single-antenna UEs. Moreover, we set the carrier frequency as 150 GHz, and each array at BS is spaced with half-wavelength. On the one hand, the ULA system is implemented under the 256-element. All UEs are randomly chosen in an area with the radius range $[4, 100]\text{m}^2$, and angle range $[-\frac{\pi}{3}, \frac{\pi}{3}]$. On the other hand, the UPA system is applied with the 128×32 -element layout. In this case, we assume that UEs are randomly distributed with $r \in [4, 100]$, $\theta \in [\frac{\pi}{3}, \frac{2\pi}{3}]$, and $\phi \in [-\frac{\pi}{6}, \frac{\pi}{6}]$. Furthermore, the multi-path channel is defined with the number of NLoS path $L = 3$, and the number of UEs is set to $U = 5$.

B. Hyper-parameters Setup

As the general strategy of hyper-parameter tuning, we initially clarify the neural network parameters of the proposed algorithms, and then determine the exploration strategy for ULA and UPA separately. Specifically, each DQN consists of four fully connected hidden layers, each of which contains

²Throughout this paper, the channel model is defined with a unified spherical wavefront for all UE locations, including those that exceed the Rayleigh distance. Although conventional wisdom suggests that the planar wave approximation becomes valid beyond the Rayleigh distance, the spherical wavefront inherently generalizes both channel model in the near-field and far-field. Consequently, this approach ensures consistent modeling and enables seamless analysis across both propagation regimes.

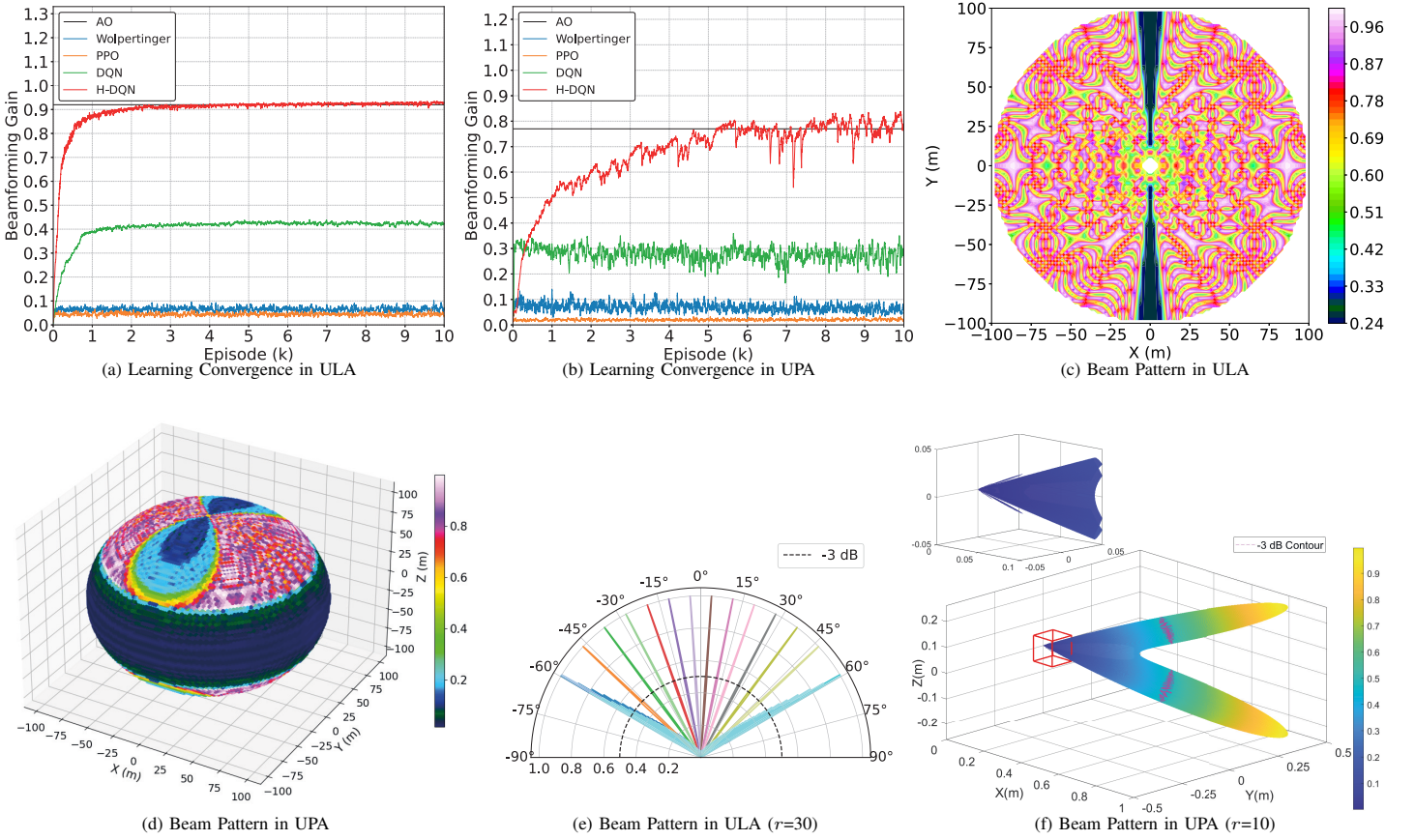


Fig. 5. Performance Analyses of HDQN-based Codebook

256 neurons. The networks are trained over a horizon of $H = 500$ steps using the Adam optimizer along with ϵ -greedy exploration. The experience replay memory is configured with a size of 10,000 samples, and a mini-batch size of $\mathcal{B} = 64$ is utilized during training. To achieve smooth learning of H-DQN, learning rate of the high-level is determined as a tenth of that of the low-level. For the ULA configuration, the quantization fraction, denoted as N_{Q_i} with $i = [0, 1]$, is set to 1000. The initial learning rate of the low-level for ULA training is 3×10^{-4} , with a decay rate of 1×10^{-2} . In contrast, for the UPA setup, the quantization fraction is increased to 2000 with $i = [0, 1, 2]$, while the initial learning rate of the low-level is adjusted to 5×10^{-4} , with a decay rate of 1×10^{-3} .

C. Beam Learning Results of Near-field Intelligent Codebook

The performance of proposed codebook is evaluated from two aspects, including convergence efficiency and beam pattern. For each aspect, the results are derived from ULA and UPA separately to verify their robustness to ELAAs. Both Fig. 5a and Fig. 5b indicate that the HRL algorithm extremely empowers the traditional DQN codebook through the approximate beamforming gain with the proposed alternating optimization (AO) [38]. More importantly, regarding Fig. 5a and Fig. 5b, the proposed H-DQN algorithm exhibits much less gain loss than DQN [44], PPO [45], Wolpertinger [46], and AO, thus demonstrating the best robustness of near-field beamforming under high-dimension array systems. Furthermore, the proposed codebook is deployed to THz-aided ELAAs for beam pattern learning. In ULA systems, the twin 512-beam is

scheduled to direct to the specific area with the angle range $[-\frac{\pi}{3}, \frac{\pi}{3}] \cup [\frac{2\pi}{3}, \frac{4\pi}{3}]$. Fig. 5c indicates most desired area has been covered with THz communications, which unleashes the potential of ELAAs in ULA system. To further investigate the magnitude of sidelobe, given $r = 30$, we uniformly extract 16 codewords from the proposed ULA codebook for evaluation. Fig. 5e indicates that the beamforming gain of the most sidelobe is ignoreable or lower than the -3dB level in $[-\frac{\pi}{3}, \frac{\pi}{3}]$. In UPA systems, the twin 128×32 -beam is propagated to the partial globe area with the angle range $[-\frac{\pi}{3}, \frac{\pi}{3}] \cup [\frac{2\pi}{3}, \frac{4\pi}{3}]$ and the azimuth range $[-\frac{\pi}{6}, \frac{\pi}{6}] \cup [\frac{5\pi}{6}, \frac{7\pi}{6}]$. Fig. 5d shows each beam has illuminated their central propagation area, but inter-beam interference still remains for future research. To verify whether the sidelobe is acceptable, we choose the optimal codeword from the 16 uniform codewords with $r = 10$. Fig. 5f indicates that the main lobes are all focused separately above or below the XoY plane in the predefined area, while the magnitude of all sidelobes is much lower than the contour of -3dB.

D. Performance Comparison between MHB and Other Frameworks

To verify the performance of MHB, we first evaluate the beamforming gain, and then validate the spectral efficiency (SE) with different factors. The environment is defined as ULA systems in Section. III-C, which is a special case of UPA. In Fig. 6a, MHB is evaluated under 5 UEs and only LoS path in predefined areas, which indicates that the beamforming gain of all UEs can achieve convergence within 20 episodes instead of 4000 episodes in Fig. 5a. Meanwhile, all UEs have

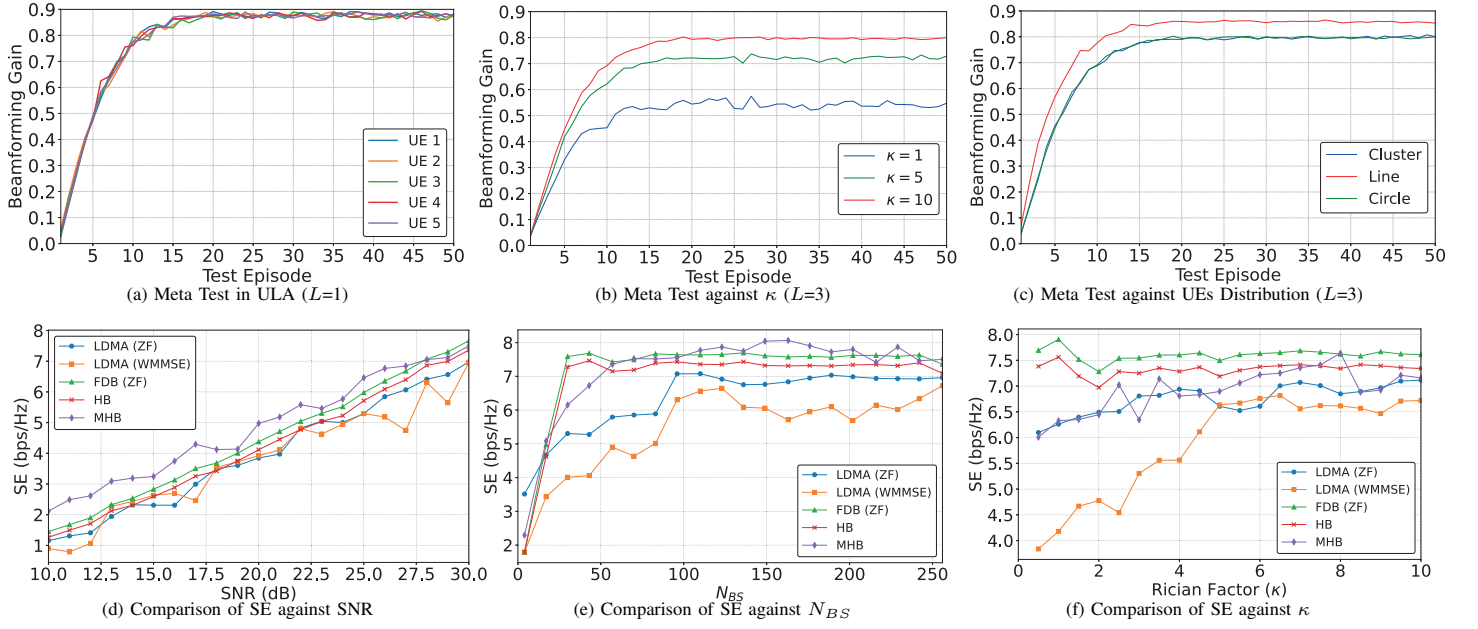


Fig. 6. Performance Evaluation of MHB

acquired the comparative beamforming gain against single-user communications. To further validate the robustness of the proposed meta-train in Algorithm 2, we extend the propagation path to $L = 3$, and then evaluate their beamforming gain against various Rician factor and UEs distribution separately. In Fig. 6b, even though Rician factor reduced from 10 to 5, the rate of gain loss can be still manipulated within 10%. In Fig. 6c, the beamforming gain is measured in various UE distribution scenarios, including cluster hotspots, line uniform distribution, and circle uniform distribution. The numerical results demonstrate that the beamforming gain will be kept over 0.8, which is acceptable for ELAAs. From Fig. 6d to Fig. 6f, it is obvious that fully digital beamforming (FDB) and hybrid beamforming (HB) play the role of a benchmark but partially suffer from dimension disasters in practice. With regard to LDMA [20], ZF is a more robust candidate than WMMSE [47]. In Fig. 6d, the performance of all algorithms is consistently increasing with the inflation of SNR, where the performance gap between HB and codebook-based beamforming is negligible. Instead, Fig. 6e shows that MHB exhibits the same robustness to SE as HB and FDB when N_{BS} exceeds 50, while outperforming the SE of LDMA. For the Rician factor κ , Fig. 6f implies that MHB tends to approximate the SE of HB and exceeds the performance of LDMA with $\kappa \leq 8$. In all, MHB has achieved a trade-off between accuracy and efficiency even with ELAAs.

VI. CONCLUSION

In this paper, the H-DQN codebook is developed to release the potential of large-scale near-field beamforming in both ULA and UPA systems, compensating for imperfect channel estimation in most near-field areas. The quantization bound and computational complexity are derived to enable a high PAC-guarantee of beamforming gain in THz-aided communication. Moreover, MHB is further proposed to accelerate multi-user near-field beamforming. Then, the performance im-

provement of MHB is analyzed against the IRC and quantized codebook to investigate the related impact factor in THz ELAAs. Simulation results demonstrate the superiority of H-DQN codebook and MHB framework, respectively, remaining the further research for the dead zone or attenuation point at the fresnel zone boundary. For future exploration, we intend to upgrade the current intelligent codebook to a large language model (LLM)-based CKM, increasing the flexibility and availability of beamforming, especially at the transition zone of THz-aided 6G communication.

APPENDIX A PROOF OF THEOREM. 1

To derive the quantization bound for the intelligent codebook, we first simplify the codebook design to a general H-MDP framework, the sample complexity of which allows us to limit the quantization of the intelligent codebook. Then, the quantization bound of codewords is determined to guide the design of intelligent codebook.

A. Sample Complexity of General H-MDP

Given the episodic finite-horizon MDP $\mathcal{M}_0 = \langle \mathcal{S}_0, \mathcal{A}_0, R_0, P_0, \mathcal{T}_0, H_0 \rangle$, we decompose them into a hierarchical set $(\mathcal{M}_{hi}, \mathcal{M}_{lo})$ as general H-MDP. According to Section. III-B, the high-level MDPs are defined as $\mathcal{M}_{hi} = \langle \mathcal{S}_{hi}, \mathcal{G}, R_{hi}, P_{hi}, \mathcal{T}_{hi}, H_{hi} \rangle$, while the low-level MDPs are denoted as $\mathcal{M}_{lo} = \langle \mathcal{S}_{lo} \times \mathcal{G}, \mathcal{A}, R_{lo}, P_{lo}, \mathcal{T}_{lo}, H_{lo} \rangle$. For the original MDP $\mathcal{M}_0$, Lemma. 1 has been proposed to constrain the quantization limit for $\mathcal{A}_0$ in [36].

Lemma 1 (PAC-Bound of Sample Complexity): There exist positive constants ϵ_0 , and δ_0 , such that any RL algorithm can achieve the PAC guarantee for (ϵ, δ) -accurate with $\epsilon \in (0, \epsilon_0)$ and $\delta \in (0, \delta_0)$. Then, the sample complexity is bounded by

$$E[N_0] \in \left[\Omega \left(\frac{|\mathcal{S}_0 \times \mathcal{A}_0| H_0^2}{\epsilon^2} \ln \left(\frac{1}{\delta + c_1} \right) \right), \right. \\ \left. \tilde{O} \left(\frac{c_2 |\mathcal{S}_0 \times \mathcal{A}_0| H_0^2}{\epsilon^2} \ln \frac{1}{\delta} \right) \right], \quad (42)$$

where $E[N_0]$ denotes the sample complexity of any RL algorithm, c_1 is a positive constant, and $c_2 \in [0, |\mathcal{S}_0|]$.

In H-MDP, (ϵ, δ) -accurate obliges the hierarchical policy set (π_{hi}, π_{lo}) to satisfy $|Q_0^{\pi_{lo}, \pi_{hi}} - Q_0^{\pi_{lo}, \pi_{hi}}| \leq \epsilon$, where $Q_0^{\pi_{lo}, \pi_{hi}}$ means the Q-value of the optimal hierarchical policy in the original MDP. In reference to the triangle inequality, $|Q_0^{\pi_{lo}, \pi_{hi}} - Q_0^{\pi_{lo}, \pi_{hi}}| \leq \epsilon$ can be ensured by $|Q_0^{\pi_{lo}, \pi_{hi}} - Q_0^{\pi_{lo}, \pi_{hi}}| + |Q_0^{\pi_{lo}, \pi_{hi}} - Q_0^{\pi_{lo}, \pi_{hi}}| \leq \epsilon$. Therefore, the PAC guarantee is decomposed as high-level version $|Q_0^{\pi_{lo}, \pi_{hi}} - Q_0^{\pi_{lo}, \pi_{hi}}| \leq \frac{\epsilon}{2}$ and low-level version $|Q_0^{\pi_{lo}, \pi_{hi}} - Q_0^{\pi_{lo}, \pi_{hi}}| \leq \frac{\epsilon}{2}$.

1) High-level PAC Guarantee

For the reason that π_{lo} is stationary in $|Q_0^{\pi_{lo}, \pi_{hi}} - Q_0^{\pi_{lo}, \pi_{hi}}| \leq \frac{\epsilon}{2}$, high-level PAC guarantee is equivalent to an independent PAC bound for high-level MDP. Therefore, Lemma. 1 can be adjusted to bound the sample complexity of high-level MDP $E[N_{hi}]$ as

$$E[N_{hi}] \in \left[\Omega \left(\frac{|\mathcal{S}_{hi} \times \mathcal{G}| H_{hi}^2}{\epsilon^2} \ln \left(\frac{1}{\delta + c_{1,hi}} \right) \right), \right. \\ \left. \tilde{O} \left(\frac{c_{2,hi} |\mathcal{S}_{hi} \times \mathcal{G}| H_{hi}^2}{\epsilon^2} \ln \frac{1}{\delta} \right) \right], \quad (43)$$

where $c_{1,hi}$ is a positive constant, $c_{2,hi} \in [0, |\mathcal{S}_{hi}|]$, and the total sample complexity $E[N]$ complies with $E[N] \subseteq E[N_{hi}]$.

2) Low-level PAC Guarantee

Similar to Section. A-A1, low-level PAC guarantee is identical to an independent PAC bound for low-level MDP. Consequently, Lemma. 1 can be modified to limit the sample complexity of low-level MDP $E[N_{lo}]$ as

$$E[N_{lo}] \in \left[\Omega \left(\frac{|\mathcal{S}_{lo} \times \mathcal{G} \times \mathcal{A}| H_{lo}^2}{\epsilon^2} \ln \left(\frac{1}{\delta + c_{1,lo}} \right) \right), \right. \\ \left. \tilde{O} \left(\frac{c_{2,lo} |\mathcal{S}_{lo} \times \mathcal{G} \times \mathcal{A}| H_{lo}^2}{\epsilon^2} \ln \frac{1}{\delta} \right) \right], \quad (44)$$

where $c_{1,lo}$ is a positive constant, $c_{2,lo} \in [0, |\mathcal{S}_{lo}|]$, and the total sample complexity $E[N]$ complies with $E[N] \subseteq E[N_{lo}]$.

Hence, the total sample complexity $E[N]$ is defined as the intersection of the corresponding high-level and low-level one. In detail, the upper and lower total PAC bounds are

$$E[N_U] = \min \left\{ \tilde{O} \left(\frac{c_{2,hi} |\mathcal{S}_{hi} \times \mathcal{G}| H_{hi}^2}{\epsilon^2} \ln \frac{1}{\delta} \right), \right. \\ \left. \tilde{O} \left(\frac{c_{2,lo} |\mathcal{S}_{lo} \times \mathcal{G} \times \mathcal{A}| H_{lo}^2}{\epsilon^2} \ln \frac{1}{\delta} \right) \right\}, \quad (45)$$

$$E[N_L] = \max \left\{ \Omega \left(\frac{|\mathcal{S}_{hi} \times \mathcal{G}| H_{hi}^2}{\epsilon^2} \ln \left(\frac{1}{\delta + c_{1,hi}} \right) \right), \right. \\ \left. \Omega \left(\frac{|\mathcal{S}_{lo} \times \mathcal{G} \times \mathcal{A}| H_{lo}^2}{\epsilon^2} \ln \left(\frac{1}{\delta + c_{1,lo}} \right) \right) \right\}. \quad (46)$$

B. Quantization Bound of Intelligent Codebook Design

According to H-MDP for near-field ELAAs in Section. III-C, the quantized fraction N_{Q_i} is exactly the low-level action space $|\mathcal{A}|$. The boundary of N_{Q_i} is derived as (47) by solving the related inequality from (45) and (46).

$$\begin{aligned} & \left\{ \frac{|\mathcal{S}_{lo}| |\mathcal{G}| |\mathcal{A}| H_{lo}^2}{\epsilon^2} \ln \left(\frac{1}{\delta + c_{1,lo}} \right) \leq \frac{c_{2,hi} |\mathcal{S}_{hi}| |\mathcal{G}| H_{hi}^2}{\epsilon^2} \ln \left(\frac{1}{\delta} \right) \right. \\ & \left. \frac{|\mathcal{S}_{hi}| |\mathcal{G}| H_{hi}^2}{\epsilon^2} \ln \left(\frac{1}{\delta + c_{1,hi}} \right) \leq \frac{c_{2,lo} |\mathcal{S}_{lo}| |\mathcal{G}| |\mathcal{A}| H_{lo}^2}{\epsilon^2} \ln \left(\frac{1}{\delta} \right) \right\} \\ & \Rightarrow \frac{|\mathcal{S}_{hi}| H_{hi}^2 \ln \left(\frac{1}{\delta + c_{1,hi}} \right)}{c_{2,lo} |\mathcal{S}_{lo}| H_{lo}^2 \ln \left(\frac{1}{\delta} \right)} \leq N_{Q_i} \leq \frac{c_{2,hi} |\mathcal{S}_{hi}| H_{hi}^2 \ln \left(\frac{1}{\delta} \right)}{|\mathcal{S}_{lo}| H_{lo}^2 \ln \left(\frac{1}{\delta + c_{1,lo}} \right)} \end{aligned} \quad (47)$$

Due to $\mathcal{S}_{hi}$ and $\mathcal{S}_{lo}$ both containing the temporal estimated CSI, $|\mathcal{S}_h| = |\mathcal{S}_l| = N_{BS}$ is obviously achieved. As a result, the constraint of N_{Q_i} can be simplified as (48).

$$\frac{H_{hi}^2 \ln \left(\frac{1}{\delta + c_{1,hi}} \right)}{c_{2,lo} H_{lo}^2 \ln \left(\frac{1}{\delta} \right)} \leq N_{Q_i} \leq \frac{c_{2,hi} H_{hi}^2 \ln \left(\frac{1}{\delta} \right)}{H_{lo}^2 \ln \left(\frac{1}{\delta + c_{1,lo}} \right)} \quad (48)$$

APPENDIX B

DERIVATION OF COMPUTATIONAL COMPLEXITY FOR ALGORITHM. 1

A. Runtime Complexity

Owing to the H-MDP structure, we intend to calculate the runtime complexity of Algorithm. 1 from the update of the high-level and low-level Q function separately. In high-level MDP, the runtime cost is essentially from sampling and policy updates. Referring to the runtime for sampling [36], meta-controller can be modeled with Upper-confidence fixed-horizon RL (UCFH), so the runtime cost is yielded to $O(H_{hi} E[N_U])$ for sampling. Each calling of Q-Network costs runtime as $O(H_{hi} N_{BS} |\mathcal{G}|^{c_{2,hi}})$, then the total runtime for policy update is derived as

$$\begin{aligned} & O(E[N_U] H_{hi} N_{BS} |\mathcal{G}|^{c_{2,hi}}) \\ & = \min \left\{ \tilde{O} \left(\frac{c_{2,hi}^2 N_{BS}^2 |\mathcal{G}|^2 H_{hi}^3}{\epsilon^2} \ln \frac{1}{\delta} \right), \right. \\ & \quad \left. \tilde{O} \left(\frac{c_{2,hi} c_{2,lo} N_{BS}^2 |\mathcal{G}|^2 N_{Q_i} H_{lo}^2 H_{hi}}{\epsilon^2} \ln \frac{1}{\delta} \right) \right\}. \end{aligned} \quad (49)$$

According to manifold gradient optimization in matrix computation, the time complexity is derived as $O \left[H N_{BS}^2 \left(\left\lceil \frac{\lambda N_{BS}^2}{\delta r_{\min}} \right\rceil - 1 - \left\lceil \log_2 \frac{\lambda N_{BS}^2}{\delta r_{\min}} \right\rceil \right) \right]$ in model-based Hierarchical Codebook Design [38], where $H = H_{hi} H_{lo}$, and $r_{\min}$ denotes the minimum BS serving distance. As a result, if the proposed algorithm exhibits fewer time-complexities, we have to achieve

$$TC_1 < O \left[HN_{BS}^2 \left(\left\lceil \frac{\lambda N_{BS}^2}{\delta r_{\min}} \right\rceil - 1 - \left\lfloor \log_2 \frac{\lambda N_{BS}^2}{\delta r_{\min}} \right\rfloor \right) \right] \text{ and } \\ TC_2 < O \left[HN_{BS}^2 \left(\left\lceil \frac{\lambda N_{BS}^2}{\delta r_{\min}} \right\rceil - 1 - \left\lfloor \log_2 \frac{\lambda N_{BS}^2}{\delta r_{\min}} \right\rfloor \right) \right].$$

B. Space Complexity

Similarly, we can explore the space complexity from the high-level and low-level MDP. In high-level agent, the storage includes neural network space and experience replay buffer, where the first one occupies $O(N_{BS}H_{hi}|\mathcal{G}|)$, and the other holds $O(|\mathcal{B}|(N_{BS} + |\mathcal{G}|))$. As a result, the total space is up to $O(N_{BS}H_{hi}|\mathcal{G}| + |\mathcal{B}|(N_{BS} + |\mathcal{G}|))$. Correspondingly, the total space holds $O((N_{BS} + |\mathcal{G}|)H_{lo}N_{Q_i} + |\mathcal{B}|(N_{BS} + |\mathcal{G}| + N_{Q_i}))$ for low-level agent. In summary, the total space complexity is calculated as $O(N_{BS}H_{hi}|\mathcal{G}| + (N_{BS} + |\mathcal{G}|)H_{lo}N_{Q_i} + |\mathcal{B}|(N_{BS} + |\mathcal{G}| + N_{Q_i}))$. According to [36], the space complexity of DQN is $O(|\mathcal{S}|H|\mathcal{A}| + |\mathcal{B}|(|\mathcal{S}| + |\mathcal{A}|))$. In ELAAs and H-MDP framework, given $|\mathcal{S}| = N_{BS}$, $|\mathcal{A}| = N_{Q_i}^2$, and $H = H_{hi}H_{lo}$, the equivalent space complexity of DQN is updated as $O((H_{hi}H_{lo}N_{Q_i}^2 + |\mathcal{B}|)N_{BS} + |\mathcal{B}|N_{Q_i}^2)$.

APPENDIX C

PROOF OF THEOREM. 2

The achievable rate is defined as (50) for each UE.

$$R_u = \log_2 \left(1 + \frac{P}{U\sigma_N^2} \left| \bar{\mathbf{h}}_u^H \mathbf{f}_{D,u} \right|^2 \right) \quad (50)$$

According to (33) and (34), $\left| \bar{\mathbf{h}}_u^H \mathbf{f}_{D,u} \right|^2$ can be simplified to (51). Therefore, (52) can be derived by the combination of (50) and (51).

$$\left| \bar{\mathbf{h}}_u^H \mathbf{f}_{D,u} \right|^2 = \left| \sqrt{N_{BS}} \alpha_u \mathbf{F}_A^H \mathbf{B}_u \mathbf{f}_{D,u} \right|^2 \\ = N_{BS} |\alpha_u|^2 [\mathbf{B}^H \mathbf{B}]_{u,u} \quad (51)$$

$$R_u = \log_2 \left(1 + \frac{P}{U\sigma_N^2} N_{BS} |\alpha_u|^2 [\mathbf{B}^H \mathbf{B}]_{u,u} \right) \quad (52)$$

As stated by Lemma 3 of [29], the upper bound of $[(\mathbf{B}^H \mathbf{B})^{-1}]_{u,u}$ satisfies (53), where $\lambda_{\max}(\mathbf{B}^H \mathbf{B})$ and $\lambda_{\min}(\mathbf{B}^H \mathbf{B})$ denote the maximum and minimum eigenvalue of $[\mathbf{B}^H \mathbf{B}]$. In terms of (53), (52) is exactly limited to (40).

$$[(\mathbf{B}^H \mathbf{B})^{-1}]_{u,u} \leq \frac{1}{4[\mathbf{B}^H \mathbf{B}]_{u,u}} \left(\frac{\lambda_{\max}(\mathbf{B}^H \mathbf{B})}{\lambda_{\min}(\mathbf{B}^H \mathbf{B})} + \frac{\lambda_{\min}(\mathbf{B}^H \mathbf{B})}{\lambda_{\max}(\mathbf{B}^H \mathbf{B})} + 2 \right) \quad (53)$$

APPENDIX D

PROOF OF THEOREM. 3

Given the intelligent codebook $\mathcal{F}$ from MAB, the optimal beamforming vector is defined as $\mathbf{b}(\hat{r}_u, \hat{\theta}_u)$ for each UE, where $\hat{r}_u$ and $\hat{\theta}_u$ are the quantized distance and angle from $\mathcal{F}$ respectively. Furthermore, based on $\mu(\hat{\mathbf{B}}_u, \mathbf{B}_u) = \mathbf{b}^H(\hat{r}_u, \hat{\theta}_u) \mathbf{b}(r_u, \theta_u)$, the effective channel of each UE can be rewritten as (54), where $\beta_{u,n}^Q = \frac{\mathbf{b}^H(r_u, \theta_u) \mathbf{b}(\hat{r}_n, \hat{\theta}_n)}{\mu^H(\hat{\mathbf{B}}_u, \mathbf{B}_u)}$. In addition, $\tilde{\mathbf{h}}_u^Q = \frac{\bar{\mathbf{h}}_u^Q}{\|\bar{\mathbf{h}}_u^Q\|}$ is defined to normalize the effective channel.

$$\tilde{\mathbf{h}}_u^Q = \sqrt{N_{BS}} \alpha_u \mu^H(\hat{\mathbf{B}}_u, \mathbf{B}_u) [\beta_{u,1}^Q, \dots, \beta_{u,U}^Q] \quad (54)$$

Referring to [29], the rate loss $\overline{\Delta R_u}$ can be rewritten as (55), where $\bar{\mathbf{h}}_u^H = \sqrt{N_{BS}} \alpha_u [\beta_{u,1}, \dots, \beta_{u,U}]$ with $\beta_{u,n} = \frac{\mathbf{b}^H(r_u, \theta_u) \mathbf{b}(r_n, \theta_n)}{\mu^H(\mathbf{B}_u, \mathbf{B}_u)}$.

$$\overline{\Delta R_u} = E \left[\log_2 \left(1 + \frac{P \left| \bar{\mathbf{h}}_u^H \mathbf{f}_{D,u} \right|^2}{U\sigma_N^2} \right) \right] \\ - E \left[\log_2 \left(1 + \frac{\frac{P}{U\sigma_N^2} \left| \bar{\mathbf{h}}_u^Q \hat{\mathbf{f}}_{D,u} \right|^2}{\frac{P}{U\sigma_N^2} \sum_{n \neq u}^U \left| \bar{\mathbf{h}}_u^Q \hat{\mathbf{f}}_{D,n} \right|^2 + 1} \right) \right] \\ \leq E \left[\log_2 \left(\frac{\|\bar{\mathbf{h}}_u^H\|^2}{\|\bar{\mathbf{h}}_u^{QH}\|^2} \right) \right] + E \left[\log_2 \left(\frac{\|\tilde{\mathbf{h}}_u^H\|^2}{\|\tilde{\mathbf{h}}_u^{QH}\|^2} \right) \right] \\ + E \left[\log_2 \left(1 + \frac{P}{U\sigma_N^2} \sum_{n \neq u}^U \left| \bar{\mathbf{h}}_u^Q \hat{\mathbf{f}}_{D,n} \right|^2 \right) \right] \quad (55)$$

Considering $A_1 = E \left[\log_2 \left(\frac{1}{|\mu(\hat{\mathbf{B}}_u, \mathbf{B}_u)|} \right) \right]$ and $A_2 = E \left[\log_2 \left(\frac{1 + \sum_{n \neq u}^U |\beta_{u,n}|^2}{1 + \sum_{n \neq u}^U |\beta_{u,n}^Q|^2} \right) \right]$, (55) can be simplified as $\overline{\Delta R_u} \leq E \left[\log_2 \frac{1 + \frac{P}{U\sigma_N^2} \sum_{n \neq u}^U \left| \bar{\mathbf{h}}_u^Q \hat{\mathbf{f}}_{D,n} \right|^2}{|\mu(\hat{\mathbf{B}}_u, \mathbf{B}_u)|} \right]$. In terms of $|\mu(\hat{\mathbf{B}}_u, \mathbf{B}_u)| \geq \min_{\mathbf{f}_u \in \mathcal{F}} \max_{\mathbf{f}_n \in \mathcal{F}} |\mathbf{f}_u^H \mathbf{f}_n|$ and Jensen's inequality, $\overline{\Delta R_u}$ can be rewritten as

$$\overline{\Delta R_u} \leq \log_2 \left[\frac{1 + \frac{P(U-1)}{U\sigma_N^2} E \left[\|\bar{\mathbf{h}}_u^Q\|^2 \right] E \left[\left| \tilde{\mathbf{h}}_u^Q \hat{\mathbf{f}}_{D,n} \right|^2 \right]}{\min_{\mathbf{f}_u \in \mathcal{F}} \max_{\mathbf{f}_n \in \mathcal{F}} |\mathbf{f}_u^H \mathbf{f}_n|} \right] \\ = \log_2 \left[\frac{1 + \frac{PN_{BS}(U-1)}{U\sigma_N^2} \left(1 + \frac{U-1}{N_{BS}} \right) E \left[\left| \tilde{\mathbf{h}}_u^Q \hat{\mathbf{f}}_{D,n} \right|^2 \right]}{\min_{\mathbf{f}_u \in \mathcal{F}} \max_{\mathbf{f}_n \in \mathcal{F}} |\mathbf{f}_u^H \mathbf{f}_n|} \right] \quad (56)$$

According to Lemma. 2 of [30], we have $E \left[\left| \tilde{\mathbf{h}}_u^Q \hat{\mathbf{f}}_{D,n} \right|^2 \right] \leq E \left[\sin^2 \langle \tilde{\mathbf{h}}_u^Q, \hat{\mathbf{h}}_u^Q \rangle \right]$. Since the quantized beamforming vectors are independent, the quantization error of MAB $\sin^2 \langle \tilde{\mathbf{h}}_u^Q, \hat{\mathbf{h}}_u^Q \rangle$ is the minimum of N_Q^2 independent beta random variables with size $(U-1, 1)$. As claimed by [31], we get $E \left[\sin^2 \left(\tilde{\mathbf{h}}_u^Q, \hat{\mathbf{h}}_u^Q \right) \right] = N_Q^2 \beta \left(N_Q^2, \frac{U}{U-1} \right)$.

For $U = 2$, we have $N_Q^2 \beta \left(N_Q^2, 2 \right) = \frac{N_Q^2 \Gamma(N_Q^2) \Gamma(2)}{\Gamma(N_Q^2+2)} = \frac{(N_Q^2)!}{(N_Q^2+1)!} \leq N_Q^{-2}$. In contrast, for $U > 2$, we have $N_Q^2 \beta \left(N_Q^2, \frac{U}{U-1} \right) \leq \frac{N_Q^2 \Gamma(N_Q^2)}{\Gamma(N_Q^2+1+\frac{1}{U-1})} = \frac{\Gamma(N_Q^2+1)}{\Gamma(N_Q^2+1+\frac{1}{U-1})}$. Leveraging Kershaw's inequality for gamma function [48], we fur-

then get $\frac{\Gamma(N_Q^2+1)}{\Gamma(N_Q^2+1+\frac{1}{U-1})} < \left(N_Q^2 + \frac{U}{2(U-1)}\right)^{-\frac{1}{U-1}} < N_Q^{-\frac{1}{U-1}}$. In all, (41) is derived with any U .

REFERENCES

- [1] R. W. Heath, N. González-Prelcic, S. Rangan, W. Roh, and A. M. Sayeed, "An Overview of Signal Processing Techniques for Millimeter Wave MIMO Systems," *IEEE Journal of Selected Topics in Signal Processing*, vol. 10, no. 3, pp. 436–453, 2016.
- [2] R. Zhong, X. Mu, Y. Zhang, M. Jaber, and Y. Liu, "Mobile Edge Generation: A New Era to 6G," *IEEE Network*, vol. 38, no. 5, pp. 47–55, 2024.
- [3] M. Erdem, A. Saleem, O. Gurbuz, C. Ecemis, and A. Saeed, "A Simple Analytical Model for THz Band Path Loss," *IEEE Communications Letters*, vol. 27, no. 3, pp. 996–1000, 2023.
- [4] E. D. Carvalho, A. Ali, A. Amiri, M. Angelichinoski, and R. W. Heath, "Non-Stationarities in Extra-Large-Scale Massive MIMO," *IEEE Wireless Communications*, vol. 27, no. 4, pp. 74–80, 2020.
- [5] Y. Chen, S. Al-Rubaye, A. Tsourdos, K.-F. Chu, Z. Wei, L. Baker, and C. Gillingham, "Federated Deep Reinforcement Learning-Based Intelligent Surface Configuration in 6G Secure Airport Networks," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–17, 2024.
- [6] R. Zhong, X. Liu, Y. Liu, Y. Chen, and Z. Han, "Mobile Reconfigurable Intelligent Surfaces for NOMA Networks: Federated Learning Approaches," *IEEE Transactions on Wireless Communications*, vol. 21, no. 11, pp. 10020–10034, 2022.
- [7] M. Cui, Z. Wu, Y. Lu, X. Wei, and L. Dai, "Near-Field MIMO Communications for 6G: Fundamentals, Challenges, Potentials, and Future Directions," *IEEE Communications Magazine*, vol. 61, no. 1, pp. 40–46, 2023.
- [8] H. Lu, Y. Zeng, C. You, Y. Han, J. Zhang, Z. Wang, Z. Dong, S. Jin, C.-X. Wang, T. Jiang, X. You, and R. Zhang, "A Tutorial on Near-Field XL-MIMO Communications Toward 6G," *IEEE Communications Surveys & Tutorials*, vol. 26, no. 4, pp. 2213–2257, 2024.
- [9] X. Li, Z. Dong, Y. Zeng, S. Jin, and R. Zhang, "Multi-User Modular XL-MIMO Communications: Near-Field Beam Focusing Pattern and User Grouping," *IEEE Transactions on Wireless Communications*, vol. 23, no. 10, pp. 13 766–13 781, 2024.
- [10] H. Lei, J. Zhang, Z. Wang, B. Ai, and E. Bjornson, "Near-Field User Localization and Channel Estimation for XL-MIMO Systems: Fundamentals, Recent Advances, and Outlooks," *IEEE Wireless Communications*, pp. 1–9, 2025.
- [11] X. Shi, J. Wang, X. Wang, C. You, and J. Song, "Double-Sided Near-Field XL-MIMO: Beamfocusing Codeword Selection and Channel Estimation," *IEEE Transactions on Communications*, vol. 73, no. 5, pp. 3441–3455, 2025.
- [12] C. Zhang, X. Wang, D. Li, R. Jiang, and Y. Xu, "Codebook Design for Beamforming in Near-Field Cylindrical Antenna Array Systems," in *ICC 2024 - IEEE International Conference on Communications*, 2024, pp. 2348–2353.
- [13] Y. Xie, B. Ning, L. Li, and Z. Chen, "Near-Field Beam Training in THz Communications: The Merits of Uniform Circular Array," *IEEE Wireless Communications Letters*, vol. 12, no. 4, pp. 575–579, 2023.
- [14] K. Chen, C. Qi, C.-X. Wang, and G. Y. Li, "Beam Training and Tracking for Extremely Large-Scale MIMO Communications," *IEEE Transactions on Wireless Communications*, vol. 23, no. 5, pp. 5048–5062, 2024.
- [15] Z. Wang, X. Mu, Y. Liu, and R. Schober, "TTD Configurations for Near-Field Beamforming: Parallel, Serial, or Hybrid?" *IEEE Transactions on Communications*, vol. 72, no. 6, pp. 3783–3799, 2024.
- [16] Y. Li, S. Gong, H. Liu, C. Xing, N. Zhao, and X. Wang, "Near-Field Beamforming Optimization for Holographic XL-MIMO Multiuser Systems," *IEEE Transactions on Communications*, vol. 72, no. 4, pp. 2309–2323, 2024.
- [17] S. F. Kimaryo and K. Lee, "Downlink Beamforming for Dynamic Metasurface Antennas," *IEEE Transactions on Wireless Communications*, vol. 22, no. 7, pp. 4745–4755, 2023.
- [18] R. Liu and K. Wu, "Antenna Array for Amplitude and Phase Specified Near-Field Multifocus," *IEEE Transactions on Antennas and Propagation*, vol. 67, no. 5, pp. 3140–3150, 2019.
- [19] H. Zhang, N. Shlezinger, F. Guidi, D. Dardari, M. F. Imani, and Y. C. Eldar, "Beam Focusing for Near-Field Multiuser MIMO Communications," *IEEE Transactions on Wireless Communications*, vol. 21, no. 9, pp. 7476–7490, 2022.
- [20] Z. Wu and L. Dai, "Multiple Access for Near-Field Communications: SDMA or LDMA?" *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 6, pp. 1918–1935, 2023.
- [21] J. Zuo, X. Mu, and Y. Liu, "Near-Field Non-Orthogonal Multiple Access Communications," in *GLOBECOM 2023 - 2023 IEEE Global Communications Conference*, 2023, pp. 649–654.
- [22] W. Zhao, C. Han, H.-J. Song, and E. Björnson, "DNN based Two-stage Compensation Algorithm for THz Hybrid Beamforming with imperfect Hardware," 2024. [Online]. Available: <https://arxiv.org/abs/2411.14699>
- [23] R. Zhong, Y. Liu, X. Mu, Y. Chen, X. Wang, and L. Hanzo, "Hybrid Reinforcement Learning for STAR-RISs: A Coupled Phase-Shift Model Based Beamformer," *IEEE Journal on Selected Areas in Communications*, vol. 40, no. 9, pp. 2556–2569, 2022.
- [24] Y. Yuan, G. Zheng, K.-K. Wong, B. Ottersten, and Z.-Q. Luo, "Transfer Learning and Meta Learning-Based Fast Downlink Beamforming Adaptation," *IEEE Transactions on Wireless Communications*, vol. 20, no. 3, pp. 1742–1755, 2021.
- [25] S. Jang and C. Lee, "Neural Network-Aided Near-Field Channel Estimation for Hybrid Beamforming Systems," *IEEE Transactions on Communications*, vol. 72, no. 11, pp. 6768–6782, 2024.
- [26] M. Monemi, M. A. Fallah, M. Rasti, and M. Latva-Aho, "6G Fresnel Spot Beamfocusing using Large-Scale Metasurfaces: A Distributed DRL-Based Approach," *IEEE Transactions on Mobile Computing*, vol. 23, no. 12, pp. 11 670–11 684, 2024.
- [27] P.-L. Bacon, J. Harb, and D. Precup, "The option-critic architecture," in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, ser. AAAI'17. AAAI Press, 2017, p. 1726–1734.
- [28] T. D. Kulkarni, K. R. Narasimhan, A. Saeedi, and J. B. Tenenbaum, "Hierarchical deep reinforcement learning: integrating temporal abstraction and intrinsic motivation," in *Proceedings of the 30th International Conference on Neural Information Processing Systems*, ser. NIPS'16. Red Hook, NY, USA: Curran Associates Inc., 2016, p. 3682–3690.
- [29] A. Alkhateeb, G. Leus, and R. W. Heath, "Limited Feedback Hybrid Precoding for Multi-User Millimeter Wave Systems," *IEEE Transactions on Wireless Communications*, vol. 14, no. 11, pp. 6481–6494, 2015.
- [30] N. Jindal, "MIMO Broadcast Channels With Finite-Rate Feedback," *IEEE Transactions on Information Theory*, vol. 52, no. 11, pp. 5045–5060, 2006.
- [31] C. K. Au-yeung and D. J. Love, "On the performance of random vector quantization limited feedback beamforming in a MISO system," *IEEE Transactions on Wireless Communications*, vol. 6, no. 2, pp. 458–462, 2007.
- [32] M. Li, L. Zhao, and H. Liang, "An SMDP-Based Prioritized Channel Allocation Scheme in Cognitive Enabled Vehicular Ad Hoc Networks," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 9, pp. 7925–7933, 2017.
- [33] M. Chu, A. Liu, V. K. N. Lau, C. Jiang, and T. Yang, "Deep Reinforcement Learning Based End-to-End Multiuser Channel Prediction and Beamforming," *IEEE Transactions on Wireless Communications*, vol. 21, no. 12, pp. 10 271–10 285, 2022.
- [34] S. Bhattacharya and A. K. Gupta, "Deep Learning for THz Channel Estimation and Beamforming Prediction via Sub-6GHz Channel," in *2022 IEEE International Conference on Signal Processing and Communications (SPCOM)*, 2022, pp. 1–5.
- [35] C. Dann, Y. Mansour, M. Mohri, A. Sekhari, and K. Sridharan, "Guarantees for Epsilon-Greedy Reinforcement Learning with Function Approximation," in *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, ser. Proceedings of Machine Learning Research, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvári, G. Niu, and S. Sabato, Eds., vol. 162. PMLR, 2022, pp. 4666–4689. [Online]. Available: <https://proceedings.mlr.press/v162/dann22a.html>
- [36] C. Dann and E. Brunskill, "Sample complexity of episodic fixed-horizon reinforcement learning," in *Proceedings of the 29th International Conference on Neural Information Processing Systems - Volume 2*, ser. NIPS'15. Cambridge, MA, USA: MIT Press, 2015, p. 2818–2826.
- [37] A. Robert, C. Pike-Burke, and A. A. Faisal, "Sample Complexity of Goal-Conditioned Hierarchical Reinforcement Learning," in *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., vol. 36. Curran Associates, Inc., 2023, pp. 62 696–62 712. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2023/file/c5ed2c8acda8c3716b1b6f9c6c713aaa-Paper-Conference.pdf
- [38] X. Shi, J. Wang, Z. Sun, and J. Song, "Spatial-Chirp Codebook-Based Hierarchical Beam Training for Extremely Large-Scale Massive MIMO," *IEEE Transactions on Wireless Communications*, vol. 23, no. 4, pp. 2824–2838, 2024.
- [39] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proceedings of the 34th Interna-*

tional Conference on Machine Learning - Volume 70, ser. ICML'17. JMLR.org, 2017, p. 1126–1135.

- [40] B. Lin, A. Liu, M. Lei, and H. Zhou, "Low-Complexity Two-Timescale Hybrid Precoding for mmWave Massive MIMO: A Group-and-Codebook Based Approach," *IEEE Transactions on Wireless Communications*, vol. 23, no. 7, pp. 7263–7277, 2024.
- [41] R. Cerna Loli and B. Clerckx, "A Meta-Learning-Based Precoder Optimization Framework for Rate-Splitting Multiple Access," *IEEE Wireless Communications Letters*, vol. 13, no. 2, pp. 347–351, 2024.
- [42] D. Wu, Y. Zeng, S. Jin, and R. Zhang, "Environment-Aware Hybrid Beamforming by Leveraging Channel Knowledge Map," *IEEE Transactions on Wireless Communications*, vol. 23, no. 5, pp. 4990–5005, 2024.
- [43] H. Yu, H. D. Tuan, E. Dutkiewicz, H. V. Poor, and L. Hanzo, "Regularized Zero-Forcing Aided Hybrid Beamforming for Millimeter-Wave Multiuser MIMO Systems," *IEEE Transactions on Wireless Communications*, vol. 22, no. 5, pp. 3280–3295, 2023.
- [44] X. Ye and L. Fu, "Joint Codebook Selection and UE Scheduling for Unlicensed MmWave NR-U/WiGig Coexistence Based on Deep Reinforcement Learning," *IEEE Transactions on Mobile Computing*, vol. 23, no. 9, pp. 8919–8934, 2024.
- [45] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal Policy Optimization Algorithms," 2017. [Online]. Available: <https://arxiv.org/abs/1707.06347>
- [46] Y. Zhang, M. Alrabeiah, and A. Alkhateeb, "Reinforcement Learning of Beam Codebooks in Millimeter Wave and Terahertz MIMO Systems," *IEEE Transactions on Communications*, vol. 70, no. 2, pp. 904–919, 2022.
- [47] J. Shi, A.-A. Lu, W. Zhong, X. Gao, and G. Y. Li, "Robust WMMSE Precoder With Deep Learning Design for Massive MIMO," *IEEE Transactions on Communications*, vol. 71, no. 7, pp. 3963–3976, 2023.
- [48] D. Kershaw, "Some Extensions of W. Gautschi's Inequalities for the Gamma Function," *Mathematics of Computation*, vol. 41, no. 164, pp. 607–611, 1983. [Online]. Available: <http://www.jstor.org/stable/2007697>



Yang Chen received the B.Eng. degree in Transportation from Civil Aviation University of China, the M.S. degree in Aerospace from Cranfield University, UK. He is currently pursuing the Ph.D. degree with Advanced Connectivity & System Integration Research Group, at Faculty of engineering and applied sciences Cranfield University. His current research interests include 6G connectivity & security and machine learning.



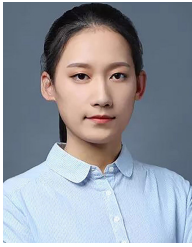
Saba Al-Rubaye is Professor Chair in Telecommunications and Autonomous Systems at Cranfield University, also member of expert group at the UKTIN-UK Telecoms Innovation Network. She leads the development of 6G technologies and their applications in future communication networks, particularly in support of space and air mobility. Prof. Al-Rubaye contributed to the development of industry standards, including IEEE P1932.1 for License/Unlicensed Interoperability and IEEE P1920.2 for Vehicle-to-Vehicle Communications for

Unmanned Aircraft Systems. Prof. Al-Rubaye has been actively involved in various high-profile projects funded by Innovate UK, Research England, EPSRC, the Department for Transport, the Natural Sciences and Engineering Research Council of Canada (NSERC), and the U.S. government. She contributed to establishing the hardware-in-the-loop testbed for smart grid communications and system integration at Quanta's Sustainable Technology Integration Laboratory (QT-STIL) in Toronto, Canada. These initiatives have played a key role in fast-tracking the adoption and integration of emerging technologies into future communication networks. Her research has been widely recognized, winning best technical paper awards at the IEEE Vehicular Technology conferences in 2011 and 2015, also the best paper awards at the IEEE DASC conference in 2020, 2021, 2024 respectively and Best Paper awarded at ICNS Conference in 2025. Prof. Al-Rubaye is registered as a Chartered Engineer (CEng), MIET, and Life Member of IEEE.



Antonios Tsourdos is a Professor of Autonomous Systems and Control with Cranfield University. He was appointed Head of the Autonomous Systems Group, Cranfield Defence and Security 2007- 2012, Head of the Centre of Autonomous and Cyber-Physical Systems, School of Aerospace, Transport and Manufacturing 2012 - 2024, Director of Research - Aerospace, Transport and Manufacturing in 2015 and Director of AI, Robotics and Space Theme, Faculty of Engineering and Applied Science in 2024. He was member of the BSI Future Flight Standards

Hub (2023-2024) and the AIAA Advancing Aerial Mobility Certification Task Force (2023-2024). Professor Tsourdos is a Fellow of the Royal Aeronautical Society, he has served (2017-2023) as chair of the IFAC Technical Committee on Aerospace Control. He was co-investigator of more than 10 UKRI Future Flight Challenge projects as well as the EPSRC funded research programme CASCADE (Complex Autonomous aircraft Systems Configuration, Analysis and Design Exploratory)

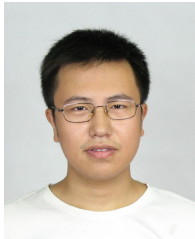


Hongyu Li is currently an assistant professor with the Internet of Things, Information Hub at the Hong Kong University of Science and Technology (Guangzhou). She received the B.S. degree in electronics and information engineering and the M.S. degree in information and communication engineering from Dalian University of Technology, Dalian, China, in 2018 and 2021, respectively, and the Ph.D. degree with the Department of Electrical and Electronic Engineering, Imperial College London, U.K., in 2025. She was a visiting Ph.D. student with the

Laboratory of Signals and Systems (L2S) of Paris-Saclay University, CNRS, CentraleSupélec, Paris, France, in 2023. Her current research include the broad areas of signal processing, wireless communications, multiple access, reconfigurable intelligent surfaces, as well as reconfigurable antenna systems.



Colin Gillingham is an Associate Director with 25 years of experience of working in the cyber security industry. Colin has managed and led technical teams engaged in security focused programmes, reviews and assessments and engages across industry on several initiatives to support the growth of the sector. This includes working with industry and academia on the progression of cyber specialisations and career paths.



Xu Shi received the B.Eng. and Ph.D. degrees in electrical engineering both from Tsinghua University, Beijing, China, in 2019 and 2024, respectively. He is working as a research associate at the Department of Electronic Engineering, Tsinghua University. His research interests include millimeter wave, RIS, extremely large-scale MIMO, channel knowledge map, and underwater acoustic communication.



Zhuangkun Wei received his PhD from the School of Engineering, University of Warwick, UK. He was working as a Research Fellow funded by EPSRC Trustworthy Autonomous System in Security node (TAS-S), and is a visiting Research Academic at Imperial College London, UK, working on the EPSRC-funded Communications Hub for Empowering Distributed cloud computing Applications and Research (CHEDDAR) project. His research interests cover artificial intelligence, 6G Non-Terrestrial Networking (NTN), and UAV control.



Zeyu Liu is a Ph.D. researcher at Cranfield University. Before that, he received MSc degree in Aerospace Vehicle Design at Cranfield University, UK, and BEng degree at Xidian University, China. His current research interests include machine learning, system reliability design, fault diagnosis, and fault-tolerant control and their applications to aerospace, robotic, and autonomous systems.



Lawrence Baker is a Chartered Engineer with 20 years of experience of working in industry. He has an extensive background in the transport and security sector, most notably in aerospace, rail and automotive. Lawrence has a strong interest in systems security engineering, understanding both the technical qualities of products and the organisational processes to develop and operate them. He maintains a detailed understanding of industry regulation, standards and best practice for securing transportation systems, engaging with key stakeholders and standards bodies.

Lawrence has also collaborated with industry and academia on numerous security research topics.



DSPACE

<https://dspace.org/>

Meta-hierarchical reinforcement learning based beamforming for near-field multi-user communications

Chen, Yang; Al-Rubaye, Saba; Tsourdos, Antonios; Li, Hongyu; Shi, Xu ...

2026

Institute of Electrical and Electronics Engineers (IEEE)

<https://doi.org/10.1109/twc.2025.3618804>

Chen Y, Al-Rubaye S, Tsourdos A, et al., (2026) Meta-hierarchical reinforcement learning based beamforming for near-field multi-user communications. IEEE Transactions on Wireless Communications, Volume 25, 2026, pp. 5446-5462