This article has been accepted for publication in IEEE Transactions on Wireless Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TWC.2024.3402267

Generalized RIS Tile Exclusion Strategy for Indoor mmWave Channels Under Concept Drift

Zi-Yang Wu, *Member, IEEE*, and Muhammad Ismail, *Senior Member, IEEE*

***Abstract*—Reconfigurable intelligent surfaces (RIS) experience considerable control overhead, particularly problematic in mobile multi-user programmable wireless environments (PWEs). This paper presents a strategy that dynamically excludes shadowed RIS tiles from supporting non-line-of-sight mmWave communications, reducing overheads and enhancing RIS usage efficiency. Spatio-temporal variations caused by crowd mobility necessitate the dynamic adaption of this strategy. The primary challenge lies in identifying optimal occasions for updating RIS tile exclusion decisions, which must strike a balance between improving achievable channel gain (performance) and power consumption as well as controlling overhead (cost) associated with decision updates. Given the absence of a general mmWave channel model, this paper applies deep reinforcement learning (DRL) for managing the RIS exclusion strategy. DRL caters to the susceptibility of mmWave channels to concept drift, where spatio-temporal variations in crowd mobility alter the probability distribution of RIS channel gains and outages. To counter concept drift and generate a universal RIS exclusion strategy for any indoor mmWave environment, we propose an adaptive exclusion update mechanism powered by DRL. This mechanism utilizes hierarchical decision decomposition, reward signal embedding, and a fusion of concept drift and temporal features due to crowd mobility, enabling efficient adaptation to environmental changes. Extensive cross-validation confirms the agent's impressive generalization ability, directly applicable to varied environments. This adaptive mechanism, despite containing only $2,300$ learnable parameters, achieves more than a two-fold increase in efficiency relative to static exclusion timing methods. Furthermore, decision execution, based on a low-cost RIS controller, only takes a few tens of nanoseconds, showcasing practicality and efficiency.**

***Index Terms*—Millimeter wave, programmable wireless environments, reconfigurable intelligent surfaces, mobility, resource allocation, deep reinforcement learning, concept drift, domain generalization.**

I. Introduction

THE next-generation wireless networks are envisioned to have programmable channels that can achieve the desired quality-of-service (QoS) for novel air interfaces such as mmWave [1]. This programmable wireless environment (PWE) can be achieved through the adoption of reconfigurable intelligent surfaces (RISs) that operate in a near-passive manner [2], [3]. RIS tiles are formed with several thin, inexpensive antennas or meta-surface arrays installed on walls that can reshape or reflect incident waves to improve signal quality while consuming minimal power [1], [4]. In practice, the control optimization problem of RIS is highly nonconvex and nonlinear [5], thus deep reinforcement learning (DRL) is usually adopted to seek an optimal strategy for blockage-aware handover [6], full-duplex security enhancement [7], [8], beamforming design [9], [10], non-orthogonal multiple access support [11], and edge computing [12]. The DRL algorithms in [13] provided RIS system with strong resistance against jamming and showed fast and smooth convergence when training. In [14], it is found that the decoupling of the continuous RIS configurations helps to learn faster in DRL. All these control strategies require real-time mmWave band RIS channel measurement and control signaling.

Recent studies underscore how RIS control overheads, signaling, and power usage can restrict the potential benefits of PWEs. In response, we propose strategies for minimizing these overheads, principally by excluding shadowed, ineffective RIS tiles. However, dynamic environments instigate concept drift in RIS channel features due to user mobility and RIS spatial distribution, necessitating timely updates of our RIS exclusion strategy. Despite enhanced channel gain, frequent updates increase signaling and power costs. This work thus highlights the need for balancing overhead minimization and performance optimization in PWEs amid significant crowd mobility.

This paper introduces a novel adaptive strategy that effectively responds to such concept drifts and updates the RIS exclusion decisions to strike a balance between improving achievable channel gain (performance) and power consumption as well as controlling overhead (cost). By employing a minimal statistical window and strategically determining the optimal time instant for decision updates, our objective is to design a decision-making system that exhibits adaptability to various situations. Through this adaptive strategy, we aim to achieve superior RIS usage and control efficiency compared to exhaustive methods in most cases, and potentially exceed their capabilities.

A. Related Research and Challenges

1) Crowd mobility: Recent research studies have highlighted that the RIS control overheads can limit the expected benefits of PWEs [15], [16]. These overheads scale with the number of RIS tiles [16]. The presence of indoor mobile users introduces time-varying shadows on RIS tiles, leading to dynamic large-scale outage areas and posing significant challenges for RIS management. However, existing studies have

Z.-Y. Wu is with the College of Information Science and Engineering, Northeastern University, Shenyang 110819, China (email: {wuziyang}@ise.neu.edu.cn). M. Ismail is with the Department of Computer Science, Tennessee Tech University, Cookeville, TN 38505 USA (email: mismail@tntech.edu). The work of M. Ismail is supported by the National Science Foundation (NSF) Award 2210251. The work of Z.-Y. Wu is supported by the National Natural Science Foundation of China (NSFC) under Grant 62103088, and China Postdoctoral Science Foundation Award 2021M700723 and 2023T160087, and the Fundamental Research Funds for the Central Universities under Grant N2304015 and 2023-MSBA-076.

This article has been accepted for publication in IEEE Transactions on Wireless Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TWC.2024.3402267

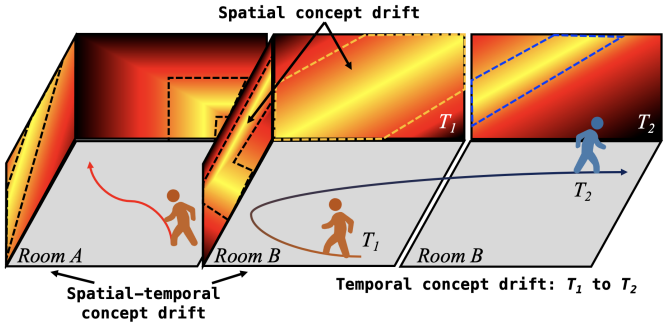


Fig. 1. There are three folds of concept drift in an indoor mobile RIS system. Here, we illustrate two rooms, in which the users are performing different trajectories due to the different layouts. To enhance the efficiency of RIS tile utilization, we exclude and deactivate some shadowed tiles from beam reflection. As shown in each wall, the tiles outside the dashed-line confined regions are excluded. In Room B, the user enters at time T_1 and is located at another position at time T_2, which induces the channel gain of RIS tiles on each wall to vary over time, and hence the exclusion area needs to change along the mobility from the yellow dash-line the blue one. We conclude this evolution of the RIS channel gain distribution on the wall as the temporal concept drifts. Extrapolating with the same logic, we find the spatial concept drifts between different walls in a room, and also the spatial-temporal concept drifts between the walls in Room A and Room B. The existence of these mobility-induced concept drifts hinders the effectiveness of machine learning-based strategies.

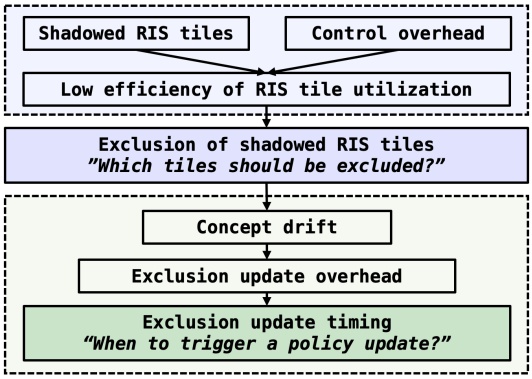


Fig. 2. The motivation and logic of this work. The shadowed RIS tiles are excluded to reduce the required control overheads, where the first problem is which tiles should be excluded given the estimation of RIS tile channel gains. Because of the concept drifts caused by mobility, the exclusion strategy should be updated over time. However, the aggressive updating raises the control overhead, thus, the second problem is on which time slot the exclusion decision should be updated such that the utilization of RIS tiles is enhanced while the control overheads are suppressed. The key is to find a good time to balance the adaptation to concept drift and the decision update overhead.

overlooked the impact of mobility on RIS channels. Previous research, such as [17] and [18], has shown that in indoor scenarios, shadowing effects, channel gain statistics, and line-of-sight (LoS) outage probabilities do not follow single-peak or general probability distributions. In our recent work [19], we proposed efficient strategies to reduce these overheads in indoor PWEs by excluding the ineffective RIS tiles shadowed due to user mobility and thus do not significantly improve QoS. In dynamic environments with user mobility, there will be concept drift in the RIS channel features, and the RIS exclusion strategy should be dynamic where decisions are updated over time. The update of exclusion improves the performance in terms of achievable channel gain, while frequent updates cost aggressive signaling overhead and power consumption. As shown in Fig. 1, we will demonstrate three types of concept drift in this work, which means the channel statistical feature varies in three different scales.

The occurrence of concept drift in RIS channels can be attributed to two prominent factors. Firstly, the wide spatial distribution of RIS tiles results in significantly varied statistical characteristics across different locations. Secondly, due to considerable indoor mobility, user equipment (UE) experiences notable displacements with respect to RIS tiles within brief periods, which precipitates further temporal concept drift in the distribution of RIS channels. The statistical analyses provided in this paper substantiate these factors. Updating the RIS exclusion strategy at a low rate causes performance degradation due to the drift of channel statistics under mobility. However, the dynamic decision update adds extra overhead to refresh the exclusion strategy. In the layouts considered in this work, updating the RIS exclusion in every frame (100 ms) would result in a control signal data rate of at least 60 kbit/s. Therefore, additional efforts are required to minimize overheads and maximize the benefits of PWEs in the presence of crowd mobility.

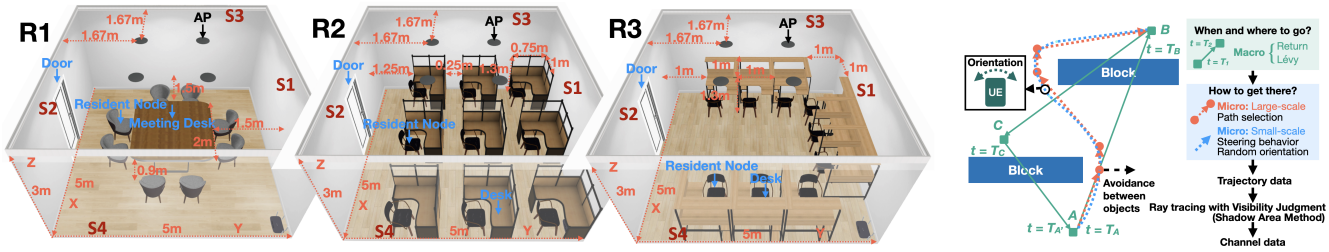
2) Strategy's generalization ability: In DRL, bootstrapping is a crucial aspect of the actor-critic framework, relying on the Markov property of interactions between the environment and the agent. Actor-critic enables policy gradient learning and thus empowers the DRL agents for continuous action output. A good Markov property ensures that the next state solely depends on the current state and action, which is measured as the *first memory order of a Markov chain* [20].

However, in the presence of crowd mobility, ensuring that the current state contains sufficient information to predict the future without relying on historical memories becomes challenging for RIS channels. Our previous works [21] and [22] demonstrated that utilizing long short-term memory (LSTM) neural networks to reconstruct the state sequence representation enhances the DRL performance in dynamic environments. This is because, in indoor mobility scenarios, the observed states possess a relatively high memory order, which means the next state depends on several previous states and actions. This makes the bootstrapping introduce significant biases in policy value estimation [23]. Then, the LSTM leverages a long channel history to predict the state one step further, which reduces the memory order, therefore the value estimation bias can be mitigated.

However, this method requires additional training and hardware implementation complexity due to the involvement of recurrent neural networks. Therefore, we need to find a more efficient method for the improvement of Markov property, which should be integrated inside the agents with low-cost hardware.

Moreover, the environment layout and the user density are dynamic in the real world, which causes a generalization problem for learning-based methods. DRL suffers from poor generalization ability owing to its sparse and delayed rewards as well as high-dimensional state-action spaces [24]. Most of

This article has been accepted for publication in IEEE Transactions on Wireless Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TWC.2024.3402267



(a) Room layouts (b) Mobility generator

Fig. 3. (a) Illustration of the room layouts. Mobile users, holding UEs that connect with APs via mmWave communications, are considered in these scenarios. The room layouts present $1,500$ RIS tiles on each wall to enable an indoor PWE [19]. (b) An illustration of our adopted mobility generation procedures [17], [19].

the recent research using DRL to optimize wireless systems trained and validated the agent in the same environment for a fixed layout and fixed user density. This raises serious questions on whether the DRL methods need refinements every time encountering environmental changes. Recent solutions incorporating domain knowledge with transfer learning [25] or meta-learning techniques [26] include more learnable components in agents and significantly increase the complexity of hardware implementation. Hence, there is a need for an efficient DRL algorithm that generalizes to the underlying concept drift in mobile PWEs.

B. Contributions

In summary, the motivation and logic of this work are illustrated in Fig. 2. We exclude the shadowed RIS tiles to reduce the required control overheads. The first problem is which tiles should be excluded given the estimation of RIS tile channel gains. Moreover, due to the concept drifts caused by mobility, the exclusion strategy should be updated over time. A more frequent update means better tracking of mobility features. However, the aggressive updating raises the control overhead, thus, the second problem is on which time slot the exclusion decision should be updated such that the utilization of RIS tiles is enhanced while the control overheads are suppressed. The key is to find a good time to balance the adaptation to concept drift and the decision update overhead. This is an NP-hard integer programming problem.

The DRL strategy proposed in this paper allows the agent to determine the optimal time instant for dynamically excluding RIS tiles, effectively balancing performance improvement and re-exclusion cost. This is achieved with minimal observational data, utilizing only channel gain samples from a single time slot. Once trained, the agent can readily adapt to any environmental changes without transfer learning. This paper makes several important contributions in constructing a dynamic exclusion of RIS tiles that can generalize to different environments, which can be summarized as follows:

- **Hierarchical decision decomposition**: This paper tackles the dynamic exclusion of RIS tiles by decomposing the problem into two parts. The first part utilizes the DRL agent to determine the optimal time instant for updating the exclusion strategy. The second part focuses on selecting the tiles to exclude, which is proven to be a convex problem, enabling the optimal solution to be found efficiently via a single search. This decomposition approach ensures optimality in each decision and allows the agent to be implemented in low-cost hardware with a minimal neural network size.
- **Reward signal shaping for learning efficiency improvement**: To address the problem of balancing exclusion gains and costs in different environments, this paper explores how to translate the optimization objective into a reward signal. We introduce reward signal embedding, which maps the gains and costs to a bounded scalar space. The shaping of reward signals has been mathematically and experimentally proven to enhance the trainability of DRL.
- **Fusion of concept drift and mobility**: This paper addresses the violation of the Markov assumption in DRL by the fusion of channel drift events and mobility temporal features. The proposed fusion state enhances the Markovian property of the state sequence and reduces bias in policy value estimation, leading to improved stability during DRL training. It also allows the prediction of concept drifts with simple observations across various layouts.
- **High generalization ability**: Through extensive cross-validations on generalization ability, the proposed DRL strategy is proved to generalize to any environment based on training on any wall in any layout. The channel gain enhancement per strategy update of the proposed strategy is nearly twice as high as the fixed update time instant benchmark.
- **Trade-off between complexity and performance**: This work explores the most suitable neural network size for the agent as well as the most efficient DRL training framework. The validated optimal agent requires only about 2,304 learnable parameters. The execution for a single decision takes 5 ns on field-programmable gate arrays (FPGA), and 15 μs on reduced instruction-set computer (RISC) chips.

The rest of this paper is organized as follows. Section II presents the network model and concept drift in mobile RIS channels. Section III formulates the problems on the dynamic exclusion of RIS tiles. Section IV describes the proposed DRL strategy. The numerical experiment results and the associated discussion pertaining to the implementation of

This article has been accepted for publication in IEEE Transactions on Wireless Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TWC.2024.3402267

the dynamic RIS tile exclusion strategies are presented in Section V. Conclusions are presented in Section VI.

II. Evidence of Concept Drift Caused by Crowd Mobility

First, we introduce an indoor environment covered with the RIS tiles operating at 28 GHz mmWave bands. Three typical indoor scenarios are used for functional verification and performance testing. Then, a brief discussion of the method of generating mobile channel data is also provided. Finally, the concept drift in the RIS channel under indoor crowd mobility is examined for conventional channel modeling methods, and the crucial challenge led by this drift is conveyed.

A. Indoor Programmable Wireless Environment

Three types of indoor layouts are considered:

- R1: There is only one meeting desk at the room center as illustrated in Fig. 3(a). The desk size is 2m×2m×0.9m.
- R2: There are nine desks with separators at a height of 1.5m in total, which are evenly located in the room as illustrated in Fig. 3(a). The desk size is 0.75m×1m.
- R3: There are nine desks (with the same size as in R2) located along the sides of the wall, leaving the center of the room empty, as illustrated in Fig. 3(a).

For all layouts, the room size is 5m $\times$ 5m $\times$ 3m, and there is only one entrance centrally located at (2.5m, 0m). There are four mmWave access points (APs) distributed evenly on the ceiling plane.

RIS tiles are deployed over the four walls of the room. To express the locations of the RIS tiles, each wall is described as a grid with each square representing a location of one RIS tile. At the 28 GHz mmWave band, a meta-surface-based RIS tile dimension is expected to be roughly 10cm×10cm square [27] such that the far-field condition holds for each tile. Hence, in the layout shown in Fig. 3(a), it is expected to have more than $1,500$ RIS tile locations per wall.

B. Indoor Practical Mobility Generation

One to eight users will sequentially enter the room and move around following the indoor mobility generator proposed in [19], allowing for an examination of crowd mobility effects. At high-frequency bands, the wireless channel is influenced by the relative displacement between the transmitter and receiver, making it dependent on UE trajectories, fading, and the spatial relationship with non-transparent objects. Our recent work [17] introduced a realistic indoor mobility model that captures human mobility at macro and micro scales. The macro-scale determines the departure time and destination point based on a time-homogeneous semi-Markov renewal process, encompassing return regularity and bounded Lévy-walk behavior, as shown in Fig. 3(b). Meanwhile, the micro-scale incorporates details such as the shortest path, steering behavior, and UE orientation during movement. The steering behavior influences interactions among users and their surroundings, impacting UE orientation. To validate the mobility model proposed in [17], synthetic user trajectory data were compared with measurements obtained using the Phyphox application.

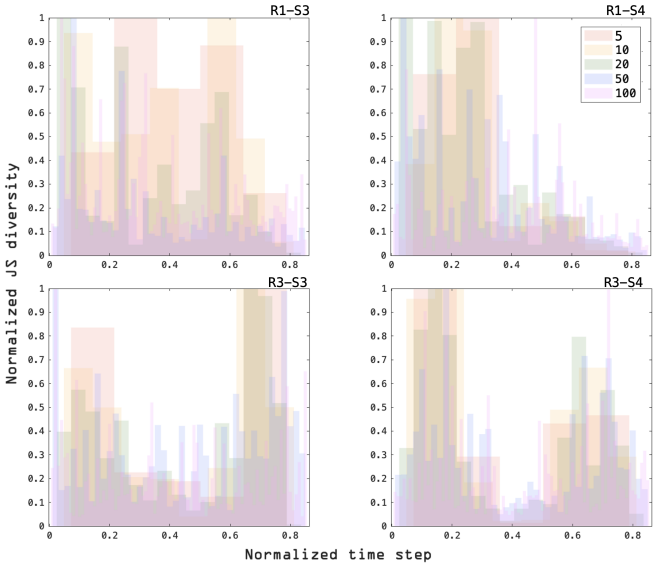


Fig. 4. The Jensen-Shannon (JS) divergence of RIS channel gain distribution between adjacent statistical windows under mobility over time. Five statistic window update frequencies are considered as 5, 10, 20, 50, and 100. The JS divergence is normalized over time. The results at walls S3 and S4 in the room layouts R1 and R3 are illustrated.

C. Indoor mmWave Channel Gain

Each RIS tile is designed in a small size such that a far-field path-loss feature could be applied [4]. The performance of RIS as a reflector for receiving uplink (UL) and downlink (DL) signals is identical within each separate band. Thus, for simplicity, we assume that the antenna parameters of AP and UE are the same. Consequently, DL (*AP→ RIS → UE*) and UL (*UE→ RIS → AP*) exhibit very similar channel gains at time slot t as

$$\mathbf{G}(t) = \mathbf{G}^{\text{AP-RIS}}(t)\boldsymbol{\Theta}(t)\mathbf{G}^{\text{RIS-UE}}(t), \tag{1}$$

where $\boldsymbol{\Theta}$ describes the phase and reflection coefficient control matrix of the RIS. Since the size of each RIS tile satisfies the far-field condition, essentially, the beam of each RIS tile can be independently controlled, or in other words, the control matrix of each tile is independent. In this work, we assume that the control matrix of each tile can be well-designed so that the beam from each RIS tile tracks UE as a reflector, and hence we can focus on the channel gain $G(n,t)$ of each tile n. The channel gain consists of both LoS and non-LoS (NLoS) components as $\mathbf{G} = \mathbf{G}_{\text{LoS}} + \mathbf{G}_{\text{NLoS}}$. For the reflected path from an RIS tile to the receiver (UE or AP), it is assumed that the NLoS component in the reflected link $\mathbf{G}_{\text{NLoS}}^{\text{RIS}\rightarrow\text{UE/AP}}$ can be ignored compared to the LoS component $\mathbf{G}_{\text{LoS}}^{\text{RIS}\rightarrow\text{UE/AP}}$ in an indoor environment [28]. However, in the link from AP/UE to RIS, the NLoS component $\mathbf{G}_{\text{NLoS}}^{\text{UE/AP}\rightarrow\text{RIS}}$ cannot be ignored as $\mathbf{G}_{\text{NLoS}}^{\text{UE/AP}\rightarrow\text{RIS}}$ would be in the same order of magnitude of $\mathbf{G}_{\text{LoS}}^{\text{UE/AP}\rightarrow\text{RIS}}$ [28]. Furthermore, once an LoS or sub-reflected ray is judged as blocked by any object (desks, another user, or the user's own body), the term $\mathbf{G}_{\text{LoS}} = 0$, and the only reception belongs to $\mathbf{G}_{\text{NLoS}}$. The shadow area method [19] is adopted as an efficient judgment of LoS blockages.

We note that examining the AP-UE link does not help us

understand the pattern of RIS tiles being blocked. Our recent work also indicates that the mobility of the UE and the user's body significantly contributes to the LoS outages in the AP-RIS-UE link [17], [19]. Therefore, when considering which RIS tiles are unnecessary, we focus on the AP-RIS-UE link.

D. Concept Drift in RIS Channel

There are three levels of concept drift in the context of RIS channels with crowd mobility. The first level is temporal concept drift, where mobility evolves over time, resulting in significant variations in channel statistics across different time periods. This fact can be observed in any sub-figure of Fig. 4, where the Jensen-Shannon (JS) divergence of RIS channel gain distribution between adjacent statistical windows changes over time. A high divergence means serious concept drift. The second level is spatial concept drift, where RIS channel evolution patterns on different walls within the same environment exhibit noticeable concept drift. This is due to the uneven distribution caused by mobility constraints imposed by the indoor layout. It can be seen in Fig. 4 through the comparison of any two walls in the same room layout. The third level is spatio-temporal concept drift, which arises from combining different room layouts, resulting in varying spatial mobility constraints. Additionally, the differences in room function also lead to variations in mobility distribution over time. As a result, the spatio-temporal characteristics of the RIS channels exhibit significant concept drift as shown in Fig. 4. Therefore, data-driven methods are essential to determine whether to update strategies at each moment in order to address the drift in channel characteristics. This challenging task cannot be easily achieved through traditional model-driven approaches alone.

III. Problem Statement

We formulate the exclusion strategy of shadowed RIS tiles balancing the QoS and the efficient utilization of RIS against the concept drift in the channel due to mobility. The focus is on maximizing the RIS channel gains over time among the included RIS tiles. The expectation of the RIS channel gain at tile n over time among all the UEs u and APs a can be expressed as

$$\Gamma(n,t) = \mathbb{E}_{u,a}\left[G(t,n|a,u)\right], \tag{2}$$

which is estimated in the first time slot of each frame. Within each frame, large-scale fading caused by mobility and the associated concept drift can be ignored since the duration of each frame T is shorter than the coherence time of indoor mobility [17]. We set each frame to be 100 ms, consisting of 10 time slots with an interval of 10 ms between each time slot. Therefore, the channel gain expectation of each frame is estimated based on the first 10 ms of data averaged over time. Then, for a given tile n, the estimated average channel gain over A APs and U UEs for a frame starting at time t with one single time slot τ data is

$$\begin{aligned}\Gamma(n,t) &= \mathbb{E}_{t,u,a}\left[G(t,n|a,u)\right], \\ &= \frac{1}{TAU}\sum_{a}\sum_{u}\int_{t}^{t+T} G(t,n|a,u)\,dt, \\ &\qquad (T \ll \text{mobility coherent time}), \\ &\simeq \frac{1}{\tau AU}\sum_{a}\sum_{u}\int_{t}^{t+\tau} G(t,n|a,u)\,dt.\end{aligned} \tag{3}$$

A. Problem Definition of RIS Tile Exclusion

The optimization aims to balance the enhancement in channel gain averages due to the RIS tiles exclusion and the utilization efficiency of the RIS tiles.

Problem 1. *The effective RIS tile channel gain enhancement metric $\eta(t)$ measured at the first time slot t of each frame is*

$$\max_{\rho(n,t)\in\{0,1\}} \eta(t) = \sum_{\mathcal{N}} \rho(n,t)\frac{\Gamma(n,t) - \mathbb{E}_n\left[\Gamma(n,t)\right]}{\mathbb{E}_n\left[\Gamma(n,t)\right]}, \tag{4}$$

where $\rho(n,t)$ is a binary decision variable such that $\rho(n,t) = 0$ means that RIS tile n is not involved in communication support (i.e., excluded), otherwise $\rho(n,t) = 1$.

$\mathbb{E}_n\left[\Gamma(n,t)\right]$ means the averaged channel gain over all the RIS tiles at time slot t, which is irrelevant to the exclusion $\rho(n,t)$. The second term $\sum_{\mathcal{N}} \rho(n,t)\mathbb{E}_n\left[\Gamma(n,t)\right]$ corresponds to the average channel gain among all the tiles without the adoption of the exclusion strategy and the first term $\sum_{\mathcal{N}} \rho(n,t)\Gamma(n,t)$ indicates the average channel gain when the exclusion strategy is adopted. The order of magnitude of $\Gamma(n,t)$ will be different in various environments, hence, we need to consider a ratio of channel gain enhancement by considering $\frac{1}{\mathbb{E}_n[\Gamma(n,t)]}$.

However, we cannot update the exclusion decision at each time slot or frame as it costs aggressive control signaling, which hinders the very motivation of tile exclusions due to the cost of exclusion decision update into Problem 1. Therefore, we break down the problem into two parts: one is to find the optimal exclusion when the decision is updated as defined in Problem 2, and the other is when to trigger such decision updates as defined in Problem 3.

Problem 2. *A global policy is pursued using a threshold on the channel gain expectation, i.e., $\Gamma_{th}(t) \in \Re \cap [\min\Gamma(n,t), \max\Gamma(n,t)]$. Then, Problem 1 can be re-written as*

$$\max_{\Gamma_{th}(t)} \quad \eta\left(t|\Gamma_{th}(t)\right) = \tag{5}$$

$$\sum_{\mathcal{N}} \mathbf{1}\left(\Gamma_{th}(t) < \Gamma(n,t)\right)\frac{\Gamma(n,t) - \mathbb{E}_n\left[\Gamma(n,t)\right]}{\mathbb{E}_n\left[\Gamma(n,t)\right]},$$

$$\textit{s.t.} \qquad \Gamma_{th}(t) \in [\min\Gamma(n,t), \max\Gamma(n,t)]. \tag{6}$$

The convexity and the optimal solution to the objective function (5) are obtained as follows.

Theorem 1. *The optimal policy for the decision variable on tile exclusion at each decision-making occasion is given by*

$$\rho(n,t) = \mathbf{1}\left(\mathbb{E}_n\left[\Gamma(n,t)\right] < \Gamma(n,t)\right), \tag{7}$$

which means that the optimal strategy is to exclude the RIS tiles whose channel gain is less than the average channel gain among all tiles.

Proof. Let the tile index n be ordered such that $\Gamma(n,t)$ is monotonically decreasing. Therefore, it yields

$$\frac{\partial}{\partial \Gamma_{\text{th}}(t)}\eta = \frac{\partial}{\partial \Gamma_{\text{th}}(t)}\left(\sum_{n=1}^{n_{\text{th}}} \frac{\Gamma(n,t)}{\mathbb{E}_n\left[\Gamma(n,t)\right]} - n_{\text{th}}\right), \tag{8}$$

where n_{th} is the RIS tile index of which the channel gain expectation is the closest to $\Gamma_{\text{th}}(t)$ as

$$n_{\text{th}}(t) = \arg\min_{n\in\mathcal{N}} \|\Gamma_{\text{th}}(t) - \Gamma(n,t)\|. \tag{9}$$

Since $\Gamma(n,t)$ is monotonically decreasing with respect to n, the increment of $n_{\text{th}}(t)$ is inversely proportional to the increment of $\Gamma_{\text{th}}(t)$. Therefore, the following holds true

$$\begin{aligned}(8) \implies & \frac{\partial}{\partial \Gamma_{\text{th}}(t)}\eta(t|\Gamma_{\text{th}}(t)) \\ & \propto \sum_{n=1}^{n_{\text{th}}-1}\left(\frac{\Gamma(n,t)}{\mathbb{E}_n\left[\Gamma(n,t)\right]}\right) - (n_{\text{th}}-1) \\ & \quad - \left[\sum_{n=1}^{n_{\text{th}}}\left(\frac{\Gamma(n,t)}{\mathbb{E}_n\left[\Gamma(n,t)\right]}\right) - n_{\text{th}}\right] \\ & = -\frac{\Gamma_{\text{th}}(t)}{\mathbb{E}_n\left[\Gamma(n,t)\right]} + 1.\end{aligned} \tag{10}$$

Then, the objective function (5) is proven to be concave since

$$\begin{aligned}\frac{\partial^2}{\partial\left(\Gamma_{\text{th}}(t)\right)^2}\eta &= \frac{\partial}{\partial \Gamma_{\text{th}}(t)}\left(-\frac{\Gamma_{\text{th}}(t)}{\mathbb{E}_n\left[\Gamma(n,t)\right]} + 1\right) \\ &= -\frac{1}{\mathbb{E}_n\left[\Gamma(n,t)\right]} < 0.\end{aligned} \tag{11}$$

Meanwhile, (10) has one and only one zero point, which is also the optimal solution to Problem 2, i.e.,

$$\Gamma_{\text{th}}(t) = \mathbb{E}_n\left[\Gamma(n,t)\right]. \tag{12}$$

□

The complexity of reaching this optimal strategy is $O(N)$. After demonstrating the optimal choices for each decision occasion, we have to optimize the time for decision updates. Please also note that the solution in (12) implies we need to estimate the average of RIS tile channel gain, which can be obtained via the sparse sensors across the RIS tiles. We only need to collect sensor readings from a few RIS tiles and use the average to make exclusion decisions [5]. Hence, the accurate channel impulse measurement is not required. Meanwhile, the AP can also help to estimate this average channel gain through the channel estimation in the link of AP-RIS-UE according to our definition of channel gain in (1). The RIS channel measurement is reused for beamforming, UE tracking, and our proposed adaptive exclusion process.

Problem 3. *The performance metric $\mathcal{J}$ considers the costs incurred due to decision updates as*

$$\max_{\kappa(t)\in\{0,1\},\rho(n,t)} \mathcal{J}(\kappa,\rho) = \sum_t \eta(t|\rho(t)) - c(t|\kappa(t)) \tag{13}$$

where $\kappa(t_0) = 1$ indicates that at time slot t_0, $\rho(n,t_0)$ is updated following $\Gamma_{th}(t_0) = \mathbb{E}_n\left[\Gamma(n,t_0)\right]$, otherwise $\rho(n,t)$ is not updated from the previous decision; $c(t|\kappa(t))$ denotes the cost of decision update, defined by the number of update times.

The cost function can be measured using different methods, such as signal throughput or switching energy consumption, all of which exhibit a positive correlation with the number of updates. Hence, to avoid inconsistencies in energy evaluations across diverse systems, we simplify the approach in this paper and consider the number of updates as the loss without loss of generality. In this scheme, the cost signal for each frame assigns a value of -1 if the decision is made to update the policy; otherwise, it is zero. Under this setup, the exclusion region on RIS would remain unchanged during the interval of $\kappa(t) = 0$. The definition of $c(t|\kappa(t))$ should take into account the trade-off between the update cost and η. This problem is non-linear and non-convex, which is addressed by a novel approach proposed in the next section.

As aforementioned, our goal is to balance the number of included RIS tiles and the performance that can be achieved, hence there is a channel capacity loss after excluding the shadowed tiles. The upper bound of the loss is described in Property 1.

Property 1. *After exclusions of shadowed tiles on the surface covered by N RIS tiles, the channel capacity would be reduced by approximately $\log_2 \frac{N}{\sum_{\mathcal{N}}\rho(n,t)+\eta(t)}$ (bit/s/Hz).*

Proof. According to the definition of η in (4), we have

$$\begin{aligned}&\eta(t) = \frac{1}{\mathbb{E}_n\left[\Gamma(n,t)\right]}\sum_{\mathcal{N}}\rho(n,t)\Gamma(n,t) - \sum_{\mathcal{N}}\rho(n,t), \\ &\implies \sum_{\mathcal{N}}\rho(n,t)\Gamma(n,t) = \mathbb{E}_n\left[\Gamma(n,t)\right]\left(\eta(t) + \sum_{\mathcal{N}}\rho(n,t)\right),\end{aligned} \tag{14}$$

where $\sum_{\mathcal{N}}\rho(n,t)\Gamma(n,t)$ means the achievable total channel gain after exclusions, and $\sum_{\mathcal{N}}\rho(n,t)$ is the number of the included RIS tiles. Assume a unit transmitting power at the AP side, and the noise density of the receiver is n_0. The achievable total channel of all the N RIS tiles without exclusions as $\sum_{\mathcal{N}}\mathbb{E}_n\left[\Gamma(n,t)\right]$, the channel capacity loss is

$$\begin{aligned}C_{\text{loss}} &= \log_2\left(1 + \frac{N\mathbb{E}_n\left[\Gamma(n,t)\right]}{n_0}\right) \\ &\quad - \log_2\left(1 + \frac{\left(\sum_{\mathcal{N}}\rho(n,t) + \eta(t)\right)\mathbb{E}_n\left[\Gamma(n,t)\right]}{n_0}\right) \\ &= \log_2\left(\frac{n_0 + N\mathbb{E}_n\left[\Gamma(n,t)\right]}{n_0 + \left(\sum_{\mathcal{N}}\rho(n,t) + \eta(t)\right)\mathbb{E}_n\left[\Gamma(n,t)\right]}\right).\end{aligned} \tag{15}$$

Considering $n_0 << N\mathbb{E}_n\left[\Gamma(n,t)\right]$, we obtain

$$C_{\text{loss}} < \log_2 \frac{N}{\sum_{\mathcal{N}}\rho(n,t) + \eta(t)} \quad \text{(bit/s/Hz)}. \tag{16}$$

□

A higher η might cause a very low capacity loss, however, the number of activated RIS tiles $\sum_{\mathcal{N}} \rho(n,t)$ is not linear with respect to η. Therefore, if the exclusion strategy could optimize the update time instant, the sacrifice in capacity would be low. This point will be justified in the numerical experimentation.

IV. Proposed Strategy for Updating the RIS Exclusion Decision

In the Problem 3, the optimal solution of decision variable ρ has been proved in Theorem 1, whereas the decision update solution of κ involves a highly non-linear and non-convex optimization. The decision update strategy should learn to adapt to the time-vary channel and make new decisions once the concept drift is serious. We propose an adaptive decision update strategy that can adapt to scenario changes, by utilizing a minimal statistical window and optimal decision update time instant. This adaptive strategy is capable of achieving, and in many cases even surpassing the performance of exhaustive methods.

A. Observation of Concept Drift in RIS Channels

The concept drift in RIS channel statistics can be defined by the offset of current adopted decision $\Gamma_{\text{th}}(t)$ from the actual RIS sensed average channel gain $\mathbb{E}_n[\Gamma(n,t)]$ as $\Gamma_{\text{th}}(t)/\mathbb{E}_n[\Gamma(n,t)]-1$. Note that it equals to $-\partial\eta/\partial\Gamma_{\text{th}}$, which means how the decision change would affect performance. In the observation, what we need to characterize is how much the performance will decrease when the decision remains unchanged, and it is equivalent to the first derivative $-\partial\eta/\partial\Gamma_{\text{th}}$. Thus, the observation of the concept drift is valued by

$$\delta(t) = \left(-\frac{\partial\eta(\Gamma_{\text{th}}(t))}{\partial\Gamma_{\text{th}}(t)}\right)^2 = \left(1 - \frac{\Gamma_{\text{th}}(t)}{\mathbb{E}_n[\Gamma(n,t)]}\right)^2, \quad (17)$$

where the squaring operation allows the agent to disregard the direction of concept drift. $\mathbb{E}_n[\Gamma(n,t)]$ is measured at the first time slot of each frame, and the observation is updated after the measurement of RIS channels.

B. Reward Signal Embedding

When it comes to DRL-based adaptive exclusion, the priority is to fine-tune the reward signal since it implies the learning objective of agents. η requires mapping to a bounded scalar space to accommodate different changes in various scenarios. This calls for the reward shaping [29], which is a feasible way to accelerate the convergence of training without changing the original optimization objective. Since the aim is to overcome concept drift, and η itself drifts with mobility, it is essential to embed η into a stable and dense state space. To achieve this, we employ the *embedding* technique, which involves coupling η and c by embedding the two signals returned by RIS into the reward space using a dense forward neural layer. The problem concerning the agent reward $r(t)$ is defined as follows in this paper.

Problem 4. *Reformulating Problem 3 for Reinforcement Learning with shaped reward:*

$$\begin{aligned}\max_{\kappa(t)} \mathcal{J}(\kappa) &= \max_{\kappa(t)} \sum_t r(t) \\ &= \max_{\kappa(t)} \left\{ \sum_t w_k g\left(w_e \eta(t|\Gamma_{th}(t))\right) - \sum_t \kappa(t) \right\},\end{aligned} \quad (18)$$

where the weight for η is denoted as w_e, and the logistic activation function is represented by $g(\cdot)$ to encounter the impact of the disturbance caused by mobility. w_k denotes the weight of performance gain from decision updating strategy compared to its cost.

When an update is triggered, the new ρ is set according to the currently perceived channel gain mean, thereby reducing the drift value of the current policy. As a result, η will increase, with its specific value change determined by the environment. Sometimes η can increase significantly, overshadowing the presence of c, causing the agent to prefer continuous updates. In certain environments, the change in η might be too small compared with the value of c, leading the agent to rarely update. These are ill-conditioned strategies. We propose using activation functions to shape and restrict the reward to avoid these ill-conditioned strategies. The logistic activation we use is a commonly adopted nonlinear activation function, which keeps the reward signal derivable and its gradient concerning exclusion strategy follows the original objective, which is proved in Theorem 2. To simplify the training process and enable effective gradient back-propagation, we intentionally set the value of w_{e} in a way that ensures η is mostly distributed within the non-saturated region of the logistic function.

Theorem 2. *The reward shaping process does not change the objective of the original objective. The embedded η in the reward signal is designed to align with the gradient direction of the exclusion decision $\Gamma_{th}(t)$.*

Proof. We notice that

$$\frac{\partial\mathcal{J}(\kappa,\rho)}{\partial\rho(n,t)} \propto \left.\frac{\partial r(t)}{\partial\rho(n,t)}\right|_{\kappa(t)=1} \propto \left.\frac{\partial r(t)}{\partial\Gamma_{\text{th}}(t)}\right|_{\kappa(t)=1}. \quad (19)$$

When the exclusion decision is updated, $\kappa(t)=1$, however, the contribution of the update cost is irrelevant to specific $\Gamma_{\text{th}}(t)$. Hence we have

$$\begin{aligned}\left.\frac{\partial r(t)}{\partial\Gamma_{\text{th}}(t)}\right|_{\kappa(t)=1} &= \frac{\partial r(t)}{\partial\eta(t)}\frac{\partial\eta(t)}{\partial\Gamma_{\text{th}}(t)} \\ &= \underbrace{\frac{w_{\text{k}} w_{\text{e}} e^{-w_{\text{e}}\eta(\Gamma_{\text{th}}(t))}}{(1+e^{-w_{\text{e}}\eta(\Gamma_{\text{th}}(t))})^2}}_{>0}\left(1-\frac{\Gamma_{\text{th}}(t)}{\mathbb{E}_n[\Gamma(n,t)]}\right),\end{aligned} \quad (20)$$

and

$$\left.\frac{\partial^2 r(t)}{\partial \Gamma_{\text{th}}(t)^2}\right|_{\kappa(t)=1} = \frac{\partial}{\partial \Gamma_{\text{th}}(t)}\left(\frac{\partial r(t)}{\partial \eta(t)}\frac{\partial \eta(t)}{\partial \Gamma_{\text{th}}(t)}\right) = \underbrace{\frac{2w_{\text{k}}w_{\text{e}}^2 e^{-2w_{\text{e}}\eta(\Gamma_{\text{th}}(t))}}{(1+e^{-w_{\text{e}}\eta(\Gamma_{\text{th}}(t))})^3}}_{>0}\left(\frac{\Gamma_{\text{th}}(t)}{\mathbb{E}_n\left[\Gamma(n,t)\right]}-1\right). \tag{21}$$

Therefore when $\Gamma_{\text{th}}(t) > \mathbb{E}_n\left[\Gamma(n,t)\right]$, $\mathcal{J}$ monotonically decreases, when $\Gamma_{\text{th}}(t) < \mathbb{E}_n\left[\Gamma(n,t)\right]$, $\mathcal{J}$ monotonically increases, and when $\Gamma_{\text{th}}(t) = \mathbb{E}_n\left[\Gamma(n,t)\right]$, $\mathcal{J}$ reaches a stationary point, which is also the maximum of $r(t)$ as well as $\mathcal{J}(\rho|\kappa=1)$. □

This ensures that the training objective of the DRL agent continues to encompass the original Problem 2 while simultaneously addressing the cost-related concern. By maintaining consistency in the gradient direction, the agent is able to strike a balance between considering the original problem and mitigating the impact of costs. The utilization of the logistic activation function causes η to have smaller values compared to w_{k}. As a result, when the decision to update the policy is made, the corresponding reward is reduced to approximately $w_{\text{k}}-1$, effectively offsetting the increase in η for that frame. This mechanism ensures a balance between the gain in η and the policy update reward. Excessive and frequent updates result in negative rewards, nullifying the benefits brought by the increase in η through re-exclusions. Similarly, untimely or too few updates do not yield substantial gains in η, leading to lower rewards as well.

C. Exclusion Update with Parameterized Stochastic Policy

First of all, consider the composition of the action space of the exclusion decision update. To minimize the complexity of the decision update actor network and enhance its ability to generalize, the decision update actor does not differentiate between distinct walls. Specifically, during training, the actor solely relies on data from a single wall, but during testing and validation, it is evaluated across different walls and layouts. This methodology enables the actor to acquire a generalized decision update policy that can be effectively applied to diverse layouts.

The decision update action can be generated in two ways, i.e., deterministic and probabilistic outputs. The deterministic action gives a direct indication of re-exclusion at t. The probabilistic outputs $K(t)$ offer the likelihood of reconsidering the RIS exclusion policy, rather than providing binary decisions. Such probabilistic outputs are enabled by actor-critic DRL frameworks. Once the probability reaches a threshold $P_{\kappa,\text{th}}$, the agent generates an update indication $\kappa=1$. Usually, this soft decision makes the agents more suitable for a concept drift task.

To adapt to concept drift, we enable the RIS agent to produce parameterized outputs, specifically the mean $\mu(t)$ and standard deviation $\sigma(t)$ of a probability distribution. In this paper, we create a continuous Gaussian update actor for this goal since it is the continuous distribution with the maximum entropy, which accelerates and stabilizes the training process. By parameterizing the output, the agent can effectively capture and adapt to the changing dynamics and uncertainties associated with concept drift. More specifically, the update action is sampled from a parameterized Gaussian probability distribution, allowing for the implementation of a stochastic policy as

$$\kappa(t) = \mathbf{1}\left(K(t) \sim \mathcal{N}(\mu(t), \sigma(t)^2) \geq P_{\kappa,\text{th}}\right), \tag{22}$$

where $\mu(t)$ and $\sigma(t)$ are the parameterized output of the actor. This means that the re-exclusion actions have a certain level of randomness or variability, influenced by the parameters of the Gaussian distribution. By adjusting these parameters, the actor can control the exploration and exploitation trade-off in the training stage.

D. Incorporating Mobility Feature and Markov Feature Enhancement

Bootstrapping in DRL relies on the Markov property of agent-environment interactions. However, when facing crowd mobility perturbations, ensuring sufficient information in the current state becomes challenging for RIS channels without relying on past states. By representing observations with lower memory order but better Markov property, we can reduce value estimation biases induced by bootstrapping in DRL. In this section, we aim to enhance the Markov property of the channel observation representation while minimizing the size of the needed neural network in DRL. This requires the use of auxiliary signals related to temporal features associated with mobility. Considering the semi-Markov property of macro-scale mobility characteristics, especially the periodicity of returns and the time correlation of Lévy walks, we project the channel observations onto the mobility timer to fuse the two processes as

$$\mathbf{s}(t) = \max\left(\mathbf{0}, \mathbf{W}_s\left[\delta(t), t/t_{\max}\right]^\top + \mathbf{b}_s\right) \tag{23}$$

where $\mathbf{W}_s$ and $\mathbf{b}_s$ correspond to the weight and bias of this state fusion layer optimized by DRL, $\max(\cdot)$ denotes the ReLU activation, $\delta(t)$ is the channel drift observation according to (17), t is the timer, and $t_{\max}$ is an upper limit of the timer t, which is set as the possible longest duration of the trajectory in this room.

This approach enhances the Markov property by considering the observation of channel drift as a key component of the state space. The channel drift is influenced by both the exclusion update actions and the inherent fluctuations of the channel. However, the fluctuations of the channel itself are independent of the exclusion update actions and are therefore irrelevant to the decision-making process we are interested in. To illustrate this, consider constructing a state transition matrix for channel drift. We would observe that the probability values vary significantly over time. When the exclusion is not updated, the Markov order of observation state arises. However, after executing an exclusion update action, the drift diminishes considerably, making this component closely related to the action. To enhance the Markov property, we project the observation of drift onto the underlying temporal features of mobility. This linkage, as described in Theorem 3, strengthens

the Markov property of the observation state, improving its ability to capture relevant information for decision-making.

Lemma 1. *Define the order ω of a Markov chain $\mathbf{s}(t)$ as*

$$\begin{aligned}\omega(\mathbf{s}(t)) &= \min_{\omega}(\omega : P\{\mathbf{s}(t) \mid \mathbf{s}(0), \ldots, \mathbf{s}(t-T)\} \\ &= P\{\mathbf{s}(t) \mid \mathbf{s}(t-\omega T), \ldots, \mathbf{s}(t-T)\}).\end{aligned} \tag{24}$$

Consider a one-order process $\{t\}$ and a high-order process $\{\delta\}$ with $\omega(\{t\}) = 1$ and $\omega(\{\delta\}) > 1$, respectively. The process fusion operation (23) *projects $\{\delta\}$ onto a timer, which is equivalent to having a marginal distribution $\delta_t = \{\delta : P\{\delta, t\} > 0\}$, where $P\{\delta, t\}$ is the joint probability that the channel drift is δ at time t. As such the fused state $\mathbf{s}(t) = \{(\delta_t, t)\}$ adheres to*

$$\begin{aligned}&P\{(\delta_t, t) \mid ((\delta_0, 0), \ldots, (\delta_{t-T}, t-T)\} \\ &\quad = P\{(\delta_t, t) \mid (\delta_{t-\omega T}, t-\omega T), \ldots, (\delta_{t-T}, {}_{t-T})\} \\ &\quad = P\{(\delta_t, t) \mid (\delta_{t-T}, t-T)\},\end{aligned} \tag{25}$$

which yields $\omega(\mathbf{s}(t)) = 1$.

Theorem 3. *The DRL sample $< \mathbf{s}(t), \kappa(t), r(t), \mathbf{s}(t+T) >$ belongs to a Markov decision reward process with an order of 1.*

Proof. When $\kappa(t) = 1$, the input of (23) at the next frame $t+T$ would be $(0, (t+T)/t_{\max})$ with probability of 1. Otherwise, the state would step into $(\delta(t+T), (t+T)/t_{\max})$. The transient from $\delta(t)$ to $\delta(t+T)$ varies during mobility, thus its order of Markov chain would be high as $\omega(\delta(t)) > 1$. However, the timer is stable and is a Markov process with the order of 1, therefore, the coupled state space has a Markov order of 1 as $\omega(\mathbf{s}(t)) = 1$ according to Lemma 1. □

Please note that the projection of the channel drift process onto a mobility timer is optimized through the DRL with (23).

E. Policy Gradient of Exclusion Decision and Training Dynamics

The system framework is briefly shown in Fig. 5. The actor-critic architecture enables continuous action space with policy gradient in temporal-differential learning, where the actor network Θ outputs action $K(t|\Theta)$ based on the input observation $\mathbf{s}(t)$, and the critic network φ evaluates the value $V(\mathbf{s}(t)|\varphi)$ of the input observation when training. As for the renew actor network Θ in Fig. 5, to ensure the optimization in the policy space, rather than solely in the parameter space [30], it is essential to relate the policy gradient towards maximizing the policy advantage function [31] as

$$\max_{\Theta'} \quad \mathbb{E}_t \left[\min\left(\mathcal{A}(t), \mathcal{B}(t)\right) - \alpha \mathcal{H}(\mathcal{N}(\mu_{\Theta'}(t), \sigma_{\Theta'}(t)))\right], \tag{26}$$

where Θ' denotes the new actor network after one-step training, α represents the positive weight of parametric Gaussian policy entropy for balancing the exploitation and exploration. Due to the involvement of trust-region optimization [30] and maximum-entropy policy, the training framework we adopted is similar to proximal policy optimizations (PPO) [31]. Given z as the discount factor for rewards, w for the credit assignment of returns, and ϵ as the clip factor, we list the following components for the policy gradient [31]:

Surrogate advantage:

$$\mathcal{A}(t) = \frac{K(t|\Theta')}{K(t|\Theta)} A(t), \tag{27}$$

Clipped advantage:

$$\mathcal{B}(t) = \max\left(\min\left(\frac{K(t|\Theta')}{K(t|\Theta)}, 1+\epsilon\right), 1-\epsilon\right) A(t), \tag{28}$$

Generalized advantage:

$$A(t) = \sum_{j=0}^{(t/T)-1} (zw)^j D(t+jT|\Theta, \varphi), \tag{29}$$

Temporal difference error:

$$D(t|\Theta, \varphi) = r(t|\Theta) + zV\left(\mathbf{s}(t+T)|\varphi\right) - V\left(\mathbf{s}(t)|\varphi\right), \tag{30}$$

Policy entropy:

$$\mathcal{H}(\mathcal{N}(\mu_{\Theta'}(t), \sigma_{\Theta'}(t))) = \frac{1}{2}\ln(2\pi e \sigma^2_{\Theta'}). \tag{31}$$

The generalized advantage is used to reduce the variance in gradient estimation with zw as the eligibility trace [32]. As shown in Fig. 5, the actor network captures both the channel drift and mobility timer, which are then fused by a fully connected layer (FCL) with a dimension of $2\times N_{\text{fusion}}$ and feed-forward to an FCL with a dimension of $N_{\text{fusion}} \times N_{\text{fusion}}$. Next, these hidden features are mapped to the mean and standard deviation of the Gaussian policy via two different FCLs with a dimension of $1\times N_{\text{fusion}}$. The activation functions of the neural outputs are ReLU, apart from the mean value activated by $\tanh$.

As for the critic network φ in Fig. 5, the training objective is minimizing the value prediction error from real reward samples, which corresponds to

$$\min_{\varphi} \quad \mathbb{E}_t \|D(t|\Theta, \varphi)\|_2. \tag{32}$$

As shown in Fig. 5, the critic network has a similar structure as the actor network but outputs the policy value estimating the expectation of upcoming returns. Both the current state and the next state values are estimated. They are then used to generate generalized advantage with reward samples.

Note that to ensure that the samples used for generating the gradient are independent and identically distributed, we adopt the experience reply technique [24], where the state-action-reward-next state samples are buffered and randomly chosen for generating the gradient.

V. Results and Discussion

In this section, we will first examine how the fused state space enhances the Markovian property of the observed channel drift sequence, thereby making the state representation more compact and efficient. Then, we will focus on finding an appropriate neural network size and selecting the lowest complexity neural network from an acceptable set of performances, which facilitates deployment on FPGA or RISC hardware. The most important aspect is to test how different training scenarios affect performance and whether the agent

This article has been accepted for publication in IEEE Transactions on Wireless Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TWC.2024.3402267

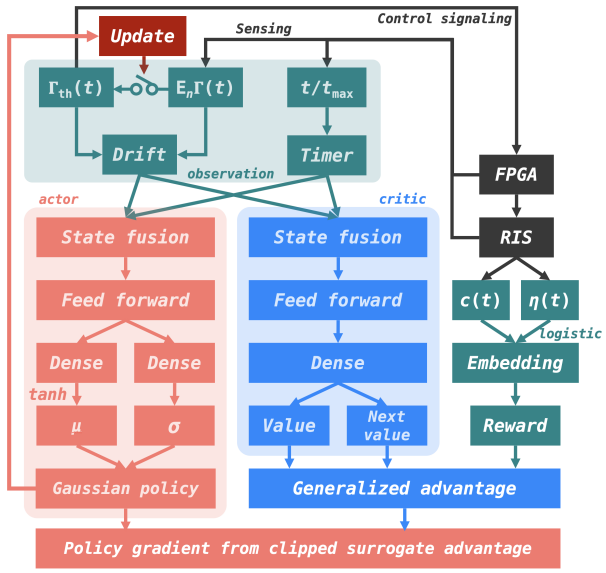


Fig. 5. The training framework and the system signaling diagram. This illustration omits the experience reply in the training stage, where the state-action-reward-next state samples are buffered and randomly chosen for generating a gradient. The red blocks denote the decision update policy actor, and the blue blocks depict the policy value estimator for training. The green blocks confine the observation and reward signals. Γ_{th} is the tile exclusion threshold in terms of the channel gain. $\mathbb{E}_n\Gamma_{\text{th}}$ accounts for the average channel gain over all RIS tile sensors. The drift is measured by (17). t is the mobility timer normalized by the maximum trajectory duration $t_{\max}$. The drift signal and timer are fed into the state fusion layer in the agent to capture a dense Markov observation space. The reward signal is reshaped by a non-linearly mapping from the decision update cost c and the effective channel gain enhancement metric η due to decision updates. Unless otherwise stated, all neurons are fully connected and activated with ReLU.

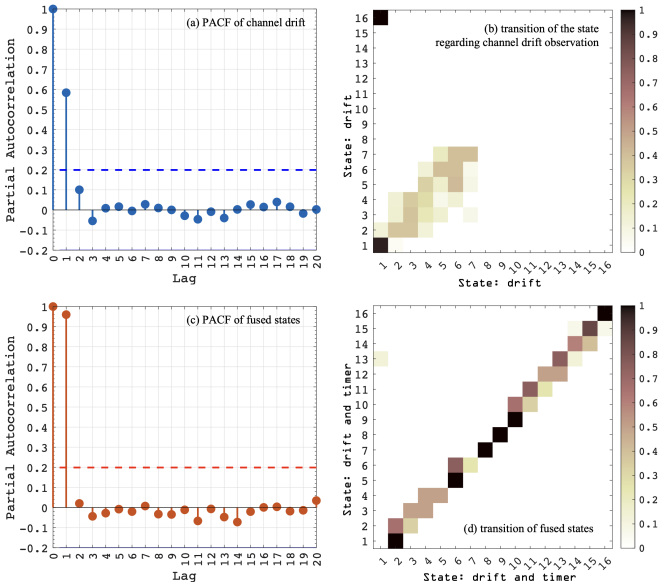


Fig. 6. (a) and (c) show the partial auto-correlation function (PACF) of the original channel drift space and the fused state space, where the PACF at a lag of 1 corresponds to the fitness of the one-order Markov process. (b) and (d) illustrate the transition probability among states, where the continuous states are categorized into 16 levels for statistics. A good Markov property for DRL means that the next state is only determined by the current state and action. When reflected in the illustrations of (a) and (c), this property means the PACF with a lag of 1 should approach 1, and the PACF with higher lags should be as small as possible. When it comes to the transition among states as illustrated in (d), most of the states transit to the adjacent states, making the prediction of channel drift more accurate. In (b), the transition among states is more random, and some of the states are hardly visited, which hinders the learnability under such state sequences.

can generalize to different scenarios. We will also compare the fixed exclusion update interval method with other state-of-the-art DRL methods.

A. Training and Hyper-parameter Setup

The training framework of PPO and our policy optimization method are not sensitive to hyper-parameter selection. We choose to start with the recommended setups [31] and fine-tune them to adapt to our learning method. The setup of hyper-parameters for agents includes: $z = 0.99$, $w = 0.95$, $\epsilon = 0.2$, $\alpha = 0.01$, the batch size is 128, and the experience horizon is 512. As for the actor and critic networks, the initial learning rate is 0.01, the ADAM optimizer is leveraged, the denominator offset is 1×10^{-8}, the gradient decay factor is 0.9, the squared gradient decay factor is 0.999, and the ℓ_2 regularization is adopted with a factor of 0.0001.

Regarding the training environment adhering to practical mobility, we determine the following parameters based on real-world measurements. The first key parameter is the displacement exponent which indicates the power-law distribution of destination selection. According to real-world measurements, this exponent was found to be approximately 0.5. As for the sojourn duration exponent featuring the power-law distribution of residential interval decision, we assume it to be 1 since this work only studies mobility impacts, and the detailed sojourn behavior is thus irrelevant. UE rotation is usually transferred into the spherical coordinate system with the polar angle and azimuth angle. The recent statistics [17] show that the polar angle subjects to Laplace distribution while sitting with a mean of 45.11 and a standard deviation of 7.84, and it follows Gaussian when walking with a mean of 31.79 and a standard deviation of 7.61.

There are 100 UE mobility trajectories used for training in each layout, where the four walls in a room share the same UE trajectories but generate their respective RIS channel data. Another 100 UE mobility trajectories are utilized for validation in each layout. Each trajectory is closed because there is only one entrance and exit.

B. Markovian State Space

Ensuring a Markov decision process proves valuable in enhancing learning efficiency. As depicted in Fig. 6(a), the original observed drift signal at a specific time exhibits a correlation of approximately 0.58 with the adjacent previous observation and maintains a correlation of about 0.1 with the observation lagged with 2 steps. This suggests a substantial Markov order when viewed from another perspective, as illustrated by the state transition matrix in Fig. 6(b). The estimation of the order is based on the number of non-zero elements in the eigenvectors of the state transition matrix, taking into account the discretization of the continuous state space for such analysis. Regardless of the number of levels into which we divide the state space, the order is approximately

This article has been accepted for publication in IEEE Transactions on Wireless Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TWC.2024.3402267

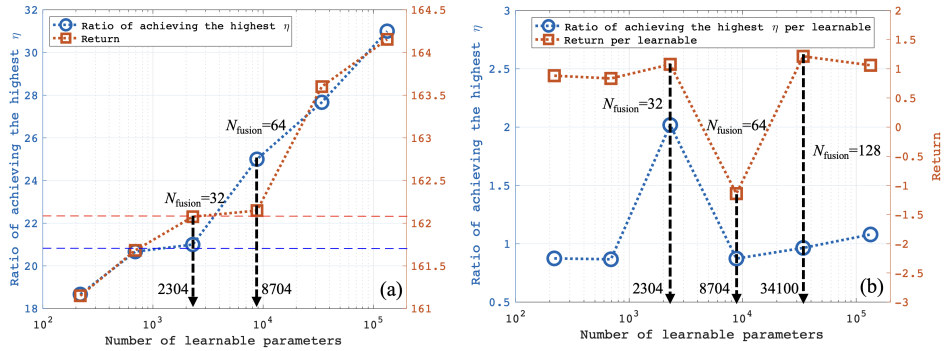


Fig. 7. Different numbers of learnable parameters (including weights and biases) are tested in the validation environments among all the walls and layouts. (a) The ratio of each number of learnable parameters achieving the best performance among the validation environment sets, and their corresponding returns, i.e. the cumulative reward along a trajectory. (b) The performance contributed by each learnable parameter, where the number of learnable parameters is normalized.

30% of the total number of states, indicating the existence of long-range dependencies between states. Consequently, in the face of such a state space, the agent requires a generalized advantage estimator to capture the long-term dependencies of state transitions, albeit at the expense of reducing the sample efficiency of learning. This trade-off is necessary to effectively handle the complexities introduced by the long-term dependencies and ensure accurate estimation of the advantages associated with different actions and state transitions.

The application of (23) yields a significant impact, as depicted in Fig. 6(c) and (d). The partial auto-correlation function (PACF) of the fused state signal with the adjacent previous time step improves to 0.97, indicating a strong correlation. At the same time, the correlation with signals from earlier time steps decreases to below 0.04. This transformation signifies that the fused state signal now possesses strict Markovian properties, where the current state contains all the necessary information for decision-making without reliance on distant past states. To ensure rigor, we further examine the order of the Markov matrix. Fig. 6(d) illustrates more stable and dense correlations between states, and regardless of the level of quantization applied to the state space, the order of the transition matrix remains at 1. This indicates that the fused state signal, incorporating the mobility timer and channel drift observations, effectively enhances the Markovian properties of the state representation. In subsequent experiments, the improved learning performance resulting from this enhanced Markovian property will be clearly demonstrated, showcasing the practical benefits of the proposed approach.

C. Number of Learnable Parameters

We aim to strike a balance between complexity and performance to avoid imposing an excessive computational burden on the energy-efficient RIS system while still achieving performance improvements. Hence, we will focus on finding an appropriate neural network size in a DRL agent, i.e., the number of learnable parameters, with the lowest complexity from a set of acceptable performance options to facilitate deployment on FPGA or hardware with a reduced instruction set.

We keep the structure unchanged, only the number of neurons is adjustable. Therefore, the parameter counts in an agent correspond to the number of neurons in the state fusion layer N_{fusion}, ranging from 2^3 to 2^8 for the test, while the number of layers in the forward neural network remains unchanged. The computational cost of the actor network approximates to $O(8N_{\text{fusion}} + 2N_{\text{fusion}}^2)$. The actor network will be leveraged for validation after training. The critic network has a similar structure but is only used for training, and its computational cost also approximates to $O(8N_{\text{fusion}} + 2N_{\text{fusion}}^2)$. The room layouts R1 to R3, and each four walls in the rooms are used for the validations.

Generally, the number of learnable parameters in a neural network is positively correlated with performance, as confirmed by Fig. 7. From Fig. 7(a), we observe that for every 4% improvement in the percentage of agents achieving the best performance among all validation environments, the number of learnable parameters needs to increase by approximately 4 times. Such a cost is significant when considering hardware deployment. Fig. 7(b) provides more insights into the contribution of each neuron to performance improvement. To ensure comparability across a wide range of neuron counts, we normalize the values. It can be observed that after exceeding $34,100$ learnable parameters, the increase in the number of neurons no longer justifies the performance improvement. In contrast, smaller neural networks with fewer learnable parameters, such as $2,304$, are much more efficient.

In Fig. 7(b), the performance metric is divided by the number of learnable parameters. The performances both $N_{\text{fusion}} = 32$ and $N_{\text{fusion}} = 64$ is similar are shown in Fig. 7(a), indicating that both agents have learned a comparable pattern for policy update. However, $N_{\text{fusion}} = 64$ seems to have reached the capacity limit at this neural network scale. To further enhance performance, it is necessary to expand the neural network scale, meaning increasing the number of learnable parameters. Therefore, in Fig. 7(b), we see that the performance contributed by each learnable parameter of $N_{\text{fusion}} = 32$ reaches a peak, while $N_{\text{fusion}} = 64$ shows a decline due to the lack of further performance improvements. Once the number of parameters is increased beyond the performance bottleneck, the performance contributed by each agent neuron decline

This article has been accepted for publication in IEEE Transactions on Wireless Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TWC.2024.3402267

becomes irredeemable.

In Fig. 7(b) (originally Fig. 5(b)), the performance metric is divided by the number of learnable parameters. The performances of both $N_{\text{fusion}} = 32$ and $N_{\text{fusion}} = 64$ are similar as shown in Fig. 7(a), indicating that both have learned a comparable pattern for policy update. However, $N_{\text{fusion}} = 64$ seems to have reached the capacity limit at this neural network scale. To further enhance performance, it is necessary to expand the neural network scale, meaning increasing the number of learnable parameters. Therefore, in Fig. 7(b), we see that the performance contributed by each learnable parameter of $N_{\text{fusion}} = 32$ reaches a peak, while $N_{\text{fusion}} = 64$ shows a decline due to the lack of further performance improvements. Once the number of parameters is increased beyond the performance bottleneck, the performance contributed by each agent neuron decline becomes irredeemable.

Considering these observations, we choose to have 32 neurons in the state fusion layer, which corresponds to approximately $2,304$ learnable parameters. The computational complexity of the deployed execution network on hardware approximates $O(2304)$, requiring a floating-point computational requirement of no less than $4,608$ FLOPs. With optimized designs, computations of this scale can be performed in less than 5 ns on platforms like JETSON TX2 or an FPGA with 28 nm process technology, while it may require at least 15 μs on a Raspberry Pi. However, these computations still fall within the interval of a time slot as defined, ensuring real-time performance, thus the requirement of our agent on hardware has been reduced significantly.

D. Prediction Ability on Significant Channel Fluctuations

Next, we provide a detailed analysis of the performance of the neural network with a suitable scale for hardware deployment during the validation process. Fig. 8 illustrates an example of the drift, decision update probability, and cumulative reward of the actor network at various scales during validation on wall S4 of layout R1.

In Fig. 8(a), it is evident that after the agent with $N_{\text{fusion}} = 32$ triggers a decision update, the observed drift stays at a low level for a long duration, which reduces the demand of update times. This can be attributed to its excellent learned decision update time instant, as depicted in Fig. 8(b). The agent triggers re-exclusion when the update probability is above 0.5. In this trajectory, the agent with $N_{\text{fusion}} = 32$ chooses to update the decision only three times, which is fewer compared to other agents, yet achieves the highest cumulative reward, as shown in Fig. 8(c). The agent with $N_{\text{fusion}} = 8$ performs the worst, while the agent with $N_{\text{fusion}} = 32$ achieves the highest return.

In Fig. 8(d), the Jensen-Shannon (JS) divergence between adjacent statistical time windows of different lengths is used to characterize the degree of change in the distribution of RIS channel gains between consecutive time intervals. The agent with $N_{\text{fusion}} = 32$ tends to choose to update decisions within the time window of a high JS divergence. The two prominent peaks in Fig. 8(d) correspond to transitions in the mobility phase, where the UEs reach their first destinations and start to wander in the room, or prepare to leave the room

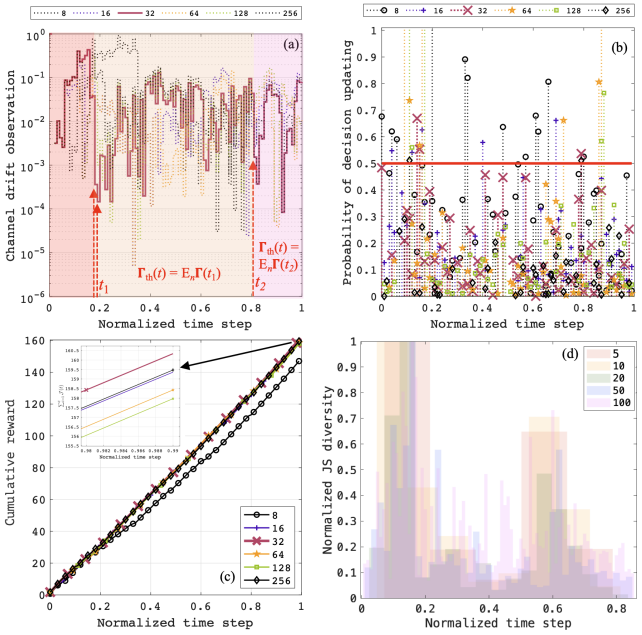


Fig. 8. The state, action, and return of the trained DRL agents with various learnable parameter numbers validating in a trajectory of layout R1 on wall S4. The state of the channel drift observation is shown in (a). The action of the probability of updating the exclusion decision is shown in (b). The cumulative reward at each time step is shown in (c). (d) demonstrates the JS divergence of RIS channel gain distribution between adjacent decision update intervals. In (a) to (c), the neuron numbers of the state fusion layer are tested from 8 to 256. In (d), five statistic window update frequencies are considered as 5, 10, 20, 50, and 100. The JS divergence is normalized over time.

[17]. However, such concept drift in terms of JS divergence in each time interval cannot be predicted in advance precisely, as it requires extensive data statistics and analysis. Any slight changes in the layout or UE number fluctuations would impact these statistics. This indicates that the DRL agents have learned how to perceive or even predict the occurrence of significant fluctuations in the channel evolution pattern.

Therefore, the algorithms that overly rely on data statistics are inherently limited in their ability to counteract concept drift in highly dynamic environments. However, our proposed method, which eliminates excessive reliance on models and statistics, can overcome these challenges, as we will demonstrate in the following analysis.

E. Generalization Ability

The paper focuses on the generalization capability of the proposed strategy, which is a challenging problem in DRL. Fig. 9(a) illustrates the performance of the proposed strategy in all training-validation environment combinations. Each layout consists of four walls, resulting in a total of 12 sets of environments used for training and validation. Fig. 9(a) shows the equivalent η considering the cost of decision updates as $\hat{\eta} = -\frac{1}{w_e} \ln\left(\frac{w_k t_{\max}}{\sum_t \mathcal{J}(t)} - 1\right)$ based on (18), which allows for comparison with the fixed decision update timing method.

Overall, the agents do not achieve higher rewards when the training and validation environments are the same, as there is no significant increase in η along the diagonal. However, the agents trained in the R1-S2 environment appear

This article has been accepted for publication in IEEE Transactions on Wireless Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TWC.2024.3402267

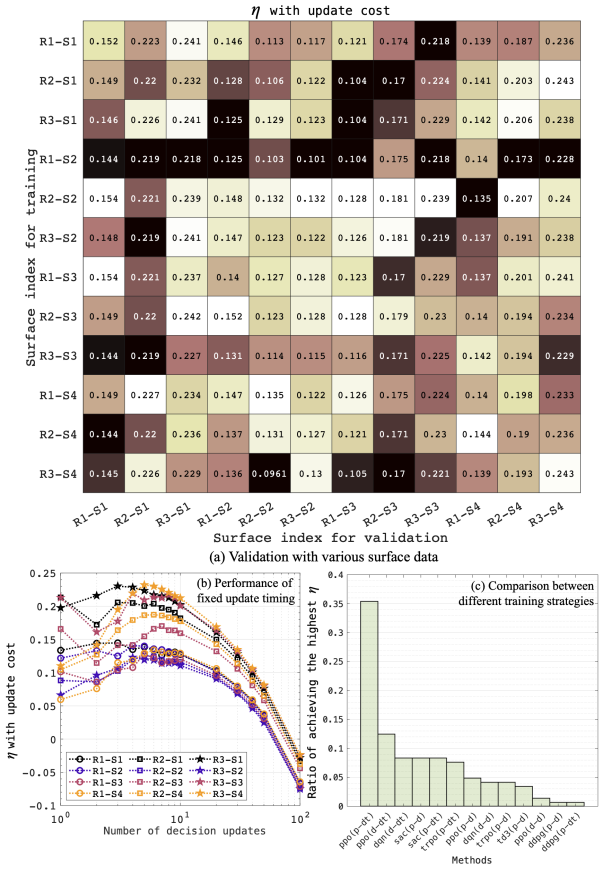


Fig. 9. (a) The validation across all the layouts and walls, where the channel data are labeled with {layout index}-{wall index}. (b) The performance of the fixed decision update timing strategy under each wall of each layout. (c) The ratio of each DRL training strategy achieving the highest return, where the state and action pairs are encoded as {p (probabilistic output) or d (deterministic output)}-{d (drift) t (if mobility timer is fused)}. In (a) and (b), the performance is measured in terms of η converted with the cost of re-exclusions. 100 traces are used for the validations in each room.

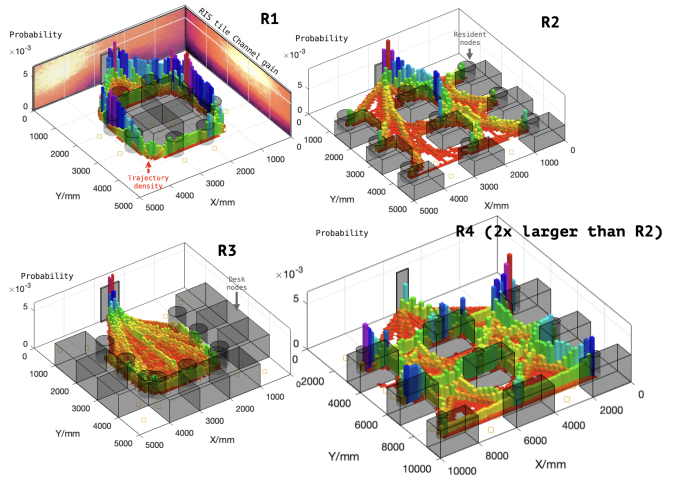


Fig. 10. The spatial distribution of the user trajectories in the validation dataset, where the frequency of each unit area being visited is calculated. In addition, the spatial distribution of the RIS tile channel gain on the wall is also demonstrated in R1, where the darker area has a lower average channel gain, and the brighter area has a higher. R4 is an office room with nine desks similar to R2 but with a four times larger size. The trajectory distribution significantly varies across R1 to R4, which brings the spatial concept drift. Nevertheless, the proposed DRL algorithm formulates a proper update timing strategy that generalizes to all these walls with various sizes and layouts.

to perform poorly in all validation environments, including when validated in the same training environment itself. R1-S2 corresponds to the wall with the entrance in the meeting room layout, and this RIS channel evolution patterns under different mobility trajectories exhibit significant variance, leading to poor performance of the learned strategies. It is important to note that not all walls with an entrance are suitable for training the agents. For example, training on the wall R2-S2 yields good performance in most of the validation environments. Even in the exceptional cases of R1-S2 and R2-S2, their advantages or disadvantages are relative and limited. For instance, in the actual validation on wall R1-S1, the performance difference between the agents trained on R1-S2 and R2-S2 is only 0.01, which is a small gap. Therefore, the agents are not sensitive to the differences between the training and deployment environments, which is a crucial result aligned with our motivation to learn to generalize to any environment and overcome concept drift caused by mobility. The optimal solution for non-learning methods could be updating decisions whenever the JS divergence is high, which is, however, infeasible to obtain in practice since the statistic requires long-duration channel sensing, in addition, the update time instant needs a delicate JS threshold.

It is worth noting that the task the agents face is challenging, as they need to select the optimal re-exclusion time instant with minimal observed information. A good decision update time instant can greatly reduce the drift of $\delta(t)$, decrease the subsequent re-exclusion probability, and ultimately reduce the overall decision update cost, which cannot be achieved by non-learning methods.

Fig. 9(b) presents the results of the non-learning method, i.e., the fixed timing method, applied through an exhaustive search. In most environments, the best performance is achieved when the update frequency is between 4 and 6 updates per trajectory. More importantly, the proposed learning method achieves the best performance with only 2 re-exclusions, while the non-learning method requires 5 exclusion updates. This highlights the advantage of the proposed learning method in identifying the optimal renewal time instant. When examining the η gains achieved by each re-exclusion, they can be estimated by $\eta(t)/\sum\kappa(t)$. Compared to the non-learning method, the DRL strategy proposed in this paper shows an improvement of 72.87%, 155.90%, and 112.59% in the ratio of η to the number of decision updates in R1 to R3, respectively. This improvement is significant, emphasizing the agent's ability to evaluate concept drift risks and identify the optimal update time instant, as well as to generalize to any scenario without the need for transfer learning. More importantly, we validate that the achievable capacity loss is very low after exclusions with our proposed DRL strategy. From validation layout R1 to R3, the capacity losses of the exclusive search method are 1.0659 bit/s/Hz, 1.0569 bit/s/Hz, and 1.0405 bit/s/Hz, respectively. The capacity losses of the proposed DRL method are 0.7013 bit/s/Hz, 0.6850 bit/s/Hz,

This article has been accepted for publication in IEEE Transactions on Wireless Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TWC.2024.3402267

and 0.6804 bit/s/Hz, respectively. This means we can save about 65% capacity loss using the proposed DRL method compared to the exclusive search method.

We also conduct a deeper analysis of the outcomes, incorporating 2 updates in the validation results of the proposed DRL method and taking a comparison with the exhaustive search method. Without considering the number of updates $\sum \kappa(t)$, the average η performance enhancement of the proposed DRL method reaches 43.67%, with 35.19% in R1, 54.10% in R2, and 41.72% in R3 over the exhaustive search method. This result provides an understanding of why the update timing is crucial. When the proposed DRL method results in a policy with fewer updates, it implies that the DRL agent has learned to make better decisions regarding the update time instants, which are aligned more closely with the occurrence of concept drift. Consequently, this strategic timing allows the DRL approach to secure a higher η performance. In contrast, the exhaustive search method, even when the number of updates is limited to match that of the DRL approach, struggles to achieve a high reward since it cannot perceive the occurrence of concept drift and precisely trigger the updates accordingly.

Compared to heuristic algorithms, such as the non-dominated sorting genetic algorithm II (NSGA-II), the proposed DRL method is inherently better equipped to address dynamic, model-free, nonlinear, non-convex NP-hard integer programming challenges within a vast policy-search space. In this work, the search space for the exclusion update policy grows exponentially to $2^{t_{\max}/T}$, where $t_{\max}/T$ means the total number of frames in the mobile trajectory. This exponential growth renders heuristic algorithms less capable of effectively exploring the solution space, as they tend to get trapped in local optima. More importantly, heuristic algorithms are optimized for a specific indoor layout, which significantly limits their ability to generalize across varying rooms or layouts. In contrast, the proposed DRL method is not constrained by the vast policy spaces and the variability in the environment since it is designed to overcome concept drift on a large spatiotemporal scale.

The advantage of our generalization ability remains valid when the size of the room changes. We validate this by deploying an agent trained in R2 directly in a room four times larger, R4 $(10\text{m} \times 10\text{m})$, where the number of tiles in R4 is double, as shown in Fig. 10. We present the spatial distribution probability of UE trajectories in different environments. The considerable differences between various layouts demonstrate the challenges in achieving a highly generalizable strategy. In R4, compared to the non-learning method, the DRL strategy proposed in this paper shows an improvement of 125.77% in the ratio of η to the number of decision updates. This still represents a substantial enhancement, confirming the advantage of the generalization ability of DRL even when room sizes change. The reason why it can be generalized to such entirely different sizes is that the state-action space we have defined is completely decoupled from the layout sizes and highly abstracted. In rooms of any size, the average value of RIS channel gain can be directly calculated, which is precisely what we need. In addition, different room shapes can lead to significant changes in tile channel evolution, which could greatly alter the underlying Markov process. Nevertheless, the agents trained in this study could potentially serve as pre-trained models for fine-tuning in these significantly different environments.

We compared various state-of-the-art DRL training frameworks, including PPO, TRPO, SAC, deep Q-network (DQN) [24], deep deterministic policy gradient (DDPG) [33], and twin delayed DDPG (TD3) [23]. The proposed DRL agent is not limited to training with the PPO framework alone. In fact, any DRL training framework that allows for stochastic policy optimization can be applied, but their generalization abilities are different. It should be noted that DQN includes a max operation during policy evaluation, making it unable to directly output probabilistic decisions. Additionally, only PPO and SAC can incorporate the maximum entropy method, which is important for capturing the uncertainty and exploration in the policy. Fig. 9(c) demonstrates the effectiveness of the proposed training framework (PPO) used in this paper. Among the state-of-the-art DRL frameworks, the proposed strategy has the highest ratio to gain the highest return in all the validations. Moreover, in the scenarios where the proposed strategy does not achieve the highest return, its disadvantage from the highest value is trivial.

Remark: The framework proposed in this study has been tested on three commonly used layouts, as mentioned in [1], [34], [35]. However, it is important to note that the proposed framework remains applicable to different layouts by excluding idle RIS tiles and reducing controlling overhead. The method presented in this paper is not limited by spatio-temporal dimensions, as the mobility constraints of the RIS channel exhibit scale-free statistics, as characterized in [36].

VI. Conclusions and Future Directions

This paper addresses the problem of dynamic RIS tile exclusion operating in a mmWave band with DRL. The methodology employs a dual-part strategy: a DRL-based decision on timing updates to minimize redundant configuration switching, and a convex problem approach for selecting tiles to exclude, ensuring efficiency and optimality. The introduction of a state fusion method improves the Markovian property of the agent-environment interaction, stabilizing the DRL training. Furthermore, the fused state enables the agent to predict significant concept drift even with simple observations. The validated generalizable DRL agent requires only $2,304$ learnable parameters, enhancing execution efficiency on FPGA and RISC chips to nearly twice that of traditional fixed-timing methods. It can be trained using the environment of any wall in any layout and directly deployed in other environments.

In the future, we aim to further improve the perception of concept drift and enhance generalization capabilities in complex mobility environments. It is possible that a neural network trained in a simple layout could serve as a pre-trained model, which can be transferred to more complex environments, thereby reducing the need for detailed environmental modeling. Our proposed framework can also be extended to outdoor settings, allowing the learned strategies to deactivate the RIS systems where there is strong LoS coverage or a long-term absence of users.

REFERENCES

[1] P. Del Hougne, M. Fink, and G. Lerosey, "Optimally diverse communication channels in disordered environments with tuned randomness," *Nature Electronics*, vol. 2, no. 1, pp. 36–41, 2019.

[2] T. V. Chien, H. Q. Ngo, S. Chatzinotas, and B. Ottersten, "Reconfigurable intelligent surface-assisted massive mimo: Favorable propagation, channel hardening, and rank deficiency," *IEEE Signal Proc. Mag.*, vol. 39, no. 3, pp. 97–104, 2022.

[3] X. G. Zhang, Y. L. Sun, B. Zhu, W. X. Jiang, Q. Yu, H. W. Tian, C.-W. Qiu, Z. Zhang, and T. J. Cui, "A metasurface-based light-to-microwave transmitter for hybrid wireless communications," *Light Sci. Appl*, vol. 11, no. 1, pp. 1–10, 2022.

[4] W. Tang, M. Z. Chen, X. Chen, J. Y. Dai, Y. Han, M. Di Renzo, Y. Zeng, S. Jin, Q. Cheng, and T. J. Cui, "Wireless communications with reconfigurable intelligent surface: Path loss modeling and experimental measurement," *IEEE Transactions on Wireless Communications*, vol. 20, no. 1, pp. 421–439, 2021.

[5] R. Hashemi, S. Ali, N. H. Mahmood, and M. Latva-Aho, "Deep reinforcement learning for practical phase-shift optimization in ris-aided miso urllc systems," *IEEE Internet Things J.*, vol. 10, no. 10, pp. 8931–8943, 2023.

[6] L. Jiao, P. Wang, A. Alipour-Fanid, H. Zeng, and K. Zeng, "Enabling efficient blockage-aware handover in ris-assisted mmwave cellular networks," *IEEE Transactions on Wireless Communications*, vol. 21, no. 4, pp. 2243–2257, 2022.

[7] Z. Peng, Z. Zhang, L. Kong, C. Pan, L. Li, and J. Wang, "Deep reinforcement learning for ris-aided multiuser full-duplex secure communications with hardware impairments," *IEEE Internet Things J.*, vol. 9, no. 21, pp. 21121–21135, 2022.

[8] R. Saleem, W. Ni, M. Ikram, and A. Jamalipour, "Deep-reinforcement-learning-driven secrecy design for intelligent-reflecting-surface-based 6g-iot networks," *IEEE Internet Things J.*, vol. 10, no. 10, pp. 8812–8824, 2023.

[9] C. Huang, Z. Yang, G. C. Alexandropoulos, K. Xiong, L. Wei, C. Yuen, Z. Zhang, and M. Debbah, "Multi-hop ris-empowered terahertz communications: A drl-based hybrid beamforming design," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 6, pp. 1663–1677, 2021.

[10] R. Huang and V. W. S. Wong, "Joint user scheduling, phase shift control, and beamforming optimization in intelligent reflecting surface-aided systems," *IEEE Transactions on Wireless Communications*, vol. 21, no. 9, pp. 7521–7535, 2022.

[11] J. Zhao, L. Yu, K. Cai, Y. Zhu, and Z. Han, "Ris-aided ground-aerial noma communications: A distributionally robust drl approach," *IEEE Journal on Selected Areas in Communications*, vol. 40, no. 4, pp. 1287–1301, 2022.

[12] T. Zhang, H. Wen, Y. Jiang, and J. Tang, "Deep-reinforcement-learning-based irs for cooperative jamming networks under edge computing," *IEEE Internet Things J*, vol. 10, no. 10, pp. 8996–9006, 2023.

[13] S. Hu, X. Yuan, W. Ni, X. Wang, and A. Jamalipour, "Ris-assisted jamming rejection and path planning for uav-borne iot platform: A new deep reinforcement learning framework," *IEEE Internet of Things Journal*, vol. 10, no. 22, pp. 20162–20173, 2023.

[14] X. Yuan, S. Hu, W. Ni, R. P. Liu, and X. Wang, "Joint user, channel, modulation-coding selection, and ris configuration for jamming resistance in multiuser ofdma systems," *IEEE Transactions on Communications*, vol. 71, no. 3, pp. 1631–1645, 2023.

[15] S. Abadal *et al.*, "Computing and communications for the software-defined metamaterial paradigm: A context analysis," *IEEE Access*, vol. 5, pp. 6225–6235, 2017.

[16] X. Liu, Y. Liu, Y. Chen, and H. V. Poor, "Ris enhanced massive non-orthogonal multiple access networks: Deployment and passive beamforming design," *IEEE J. SEL. AREA COMM.*, vol. 39, no. 4, pp. 1057–1071, 2021.

[17] Z. Y. Wu, M. Ismail, J. Kong, E. Serpedin, and J. Wang, "Channel characterization and realization of mobile optical wireless communications," *IEEE Trans. on Commu.*, vol. 68, no. 10, pp. 6426–6439, 2020.

[18] Z.-Y. Wu, M. Ismail, E. Serpedin, and J. Wang, "Artificial intelligence for smart resource management in multi-user mobile heterogeneous rf-light networks," *IEEE Wireless Communications*, vol. 28, no. 4, pp. 152–158, 2021.

[19] Z.-Y. Wu, M. Ismail, and J. Wang, "Efficient exclusion strategy of shadowed ris in dynamic indoor programmable wireless environments," *IEEE Transactions on Wireless Communications*, vol. 23, no. 2, pp. 994–1007, 2024.

[20] O. E. Williams, L. Lacasa, A. P. Millán, and V. Latora, "The shape of memory in temporal networks," *Nature Communications*, vol. 13, no. 1, p. 499, 2022.

[21] Z.-Y. Wu, M. Ismail, E. Serpedin, and J. Wang, "Efficient prediction of link outage in mobile optical wireless communications," *IEEE Transactions on Wireless Communications*, vol. 20, no. 2, pp. 882–896, 2021.

[22] Z.-Y. Wu, M. Ismail, E. Serpedin, and J. Wang, "Data-driven link assignment with qos guarantee in mobile rf-optical hetnet of things," *IEEE Internet Things J.*, vol. 7, no. 6, pp. 5088–5102, 2020.

[23] S. Fujimoto, H. Hoof, and D. Meger, "Addressing function approximation error in actor-critic methods," in *International conference on machine learning*, pp. 1587–1596, PMLR, 2018.

[24] V. Mnih *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[25] E. Chalmers, E. B. Contreras, B. Robertson, A. Luczak, and A. Gruber, "Learning to predict consequences as a method of knowledge transfer in reinforcement learning," *IEEE Trans. on Neur. Net. and Lear.*, vol. 29, no. 6, pp. 2259–2270, 2018.

[26] Z. Xu, X. Chen, and L. Cao, "Fast task adaptation based on the combination of model-based and gradient-based meta learning," *IEEE Transactions on Cybernetics*, vol. 52, no. 6, pp. 5209–5218, 2022.

[27] C. Liaskos, A. Tsioliaridou, S. Nie, A. Pitsillides, S. Ioannidis, and I. F. Akyildiz, "On the network-layer modeling and configuration of programmable wireless environments," *IEEE/ACM Trans. on Netw.*, vol. 27, no. 4, pp. 1696–1713, 2019.

[28] E. Basar, I. Yildirim, and F. Kilinc, "Indoor and outdoor physical channel modeling and efficient positioning for reconfigurable intelligent surfaces in mmwave bands," *IEEE Trans. on Commun.*, vol. 69, no. 12, pp. 8600–8611, 2021.

[29] A. Y. Ng, D. Harada, and S. J. Russell, "Policy invariance under reward transformations: Theory and application to reward shaping," in *Proceedings of the Sixteenth International Conference on Machine Learning*, ICML '99, (San Francisco, CA, USA), p. 278–287, Morgan Kaufmann Publishers Inc., 1999.

[30] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz, "Trust region policy optimization," in *International conference on machine learning*, pp. 1889–1897, PMLR, 2015.

[31] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[32] J. Schulman, P. Moritz, S. Levine, M. I. Jordan, and P. Abbeel, "High-dimensional continuous control using generalized advantage estimation," in *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016.

[33] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," *arXiv preprint arXiv:1509.02971*, 2015.

[34] S. Priebe and T. Kurner, "Stochastic modeling of thz indoor radio channels," *IEEE Trans. on Wirel. Commun.*, vol. 12, no. 9, pp. 4445–4455, 2013.

[35] K. Lee, H. Park, and J. R. Barry, "Indoor channel characteristics for visible light communications," *IEEE Communications Letters*, vol. 15, no. 2, pp. 217–219, 2011.

[36] D. Brockmann, L. Hufnagel, and T. Geisel, "The scaling laws of human travel," *Nature*, vol. 439, no. 7075, pp. 462–465, 2006.