

MODE: A Mixture-of-Decision-Experts Reinforcement Learning Architecture for High-Dimensional Coupled Wireless Optimization

Shiyi Lin, Hongyang Du



Abstract—Next-generation wireless networks have garnered notable attention due to their ability to deliver ultra-reliable connectivity, extreme data rates, and low-latency services, significantly impacting applications ranging from immersive communications to intelligent transportation. However, achieving these ambitious targets is hindered by tightly coupled design variables across multiple layers of the system, give rise to high-dimensional, nonconvex optimization problems that are difficult to solve efficiently in real time. To address this challenge and maximize spectral efficiency, we propose a novel learning-based framework that integrates Mixture-of-Experts (MoE) with deep reinforcement learning (DRL). We evaluate the proposed architecture against a heuristic baseline with low inference complexity and a conventional MLP-based deep reinforcement learning (DRL) approach. Numerical results show that our method achieves higher spectral efficiency than the heuristic baseline and delivers both faster convergence and superior spectral efficiency compared to the MLP-based approach, demonstrating both efficiency and robustness.

Index Terms—Generative AI (GAI), networks

1 INTRODUCTION

Recently, wireless networks have been rapidly evolving towards sixth-generation (6G) systems, driven by demands for ubiquitous connectivity, extreme data rates, low-latency services, and dense device deployment [1]. These trends introduce three interrelated characteristics. (i) Ultra-dense and heterogeneous infrastructure emerges, encompassing many base stations, reconfigurable intelligent surfaces(RIS) [2], unmanned aerial vehicles(UAVs) [3] and edge nodes [4]; (ii) massive device connectivity with diverse quality-of-service(QoS) requirements must be supported [5]; and (iii) stringent performance targets are imposed, including high spectral and energy efficiency together with low latency [6]. Collectively, these characteristics significantly increase the dimensionality and nonlinearity of system models, thereby turning classical design tasks into large-scale, high-dimensional, and nonconvex optimization problems. These challenges manifest as tightly coupled decision problems. For example, antenna placement, beamforming and power allocation interacting through the physical channel and receiver processing across many wireless applications, where the optimization of one component strongly affects

others. Meanwhile, practical deployments further demand real-time or near real-time solutions, pushing the .

To fully realize the performance benefits of next-generation wireless systems, these coupled problems must be solved efficiently and robustly under realistic modeling uncertainty. Prior work has pursued model-based and learning-based approaches. For example, a hierarchical three-stage beam training scheme achieves low-overhead antenna position and beam alignment by using multi-resolution codebooks and coarse-to-fine search, but it depends on pre-computed codebooks and still requires exhaustive sampling of candidate positions within each codebook level, which limits adaptability in highly dynamic mobile environments where channel geometry and line-of-sight conditions change rapidly [7]. Another effective class of methods leverages fractional programming together with block coordinate descent to transform certain ratio-type objectives into tractable subproblems and to alternate closed-form updates [8], [9]. These approaches deliver strong performance for joint power control, beamforming, and scheduling, yet they rely on iterative subproblem solves and often require grid or combinatorial searches when antenna placement or discrete configuration decisions are involved, and their performance and runtime can be sensitive to initialization. As a result, they are not always suitable for real-time adaptation in large-scale heterogeneous deployments. Learning-based approaches such as Deep Reinforcement Learning (DRL) circumvent explicit modeling by learning policies from interaction, and therefore have been applied to joint beamforming, power control, and interference coordination [10]. Nevertheless, DRL exhibits its own limitations, where many of its methods are sample-inefficient, struggle with very large or mixed discrete–continuous action spaces, and exhibit unstable training when the control problem decomposes into multiple interacting subtasks. This contrast highlights a clear gap, that is, neither purely model-based nor vanilla end-to-end learning fully balances scalability, data efficiency, specialization across subtasks, and stable training in tightly coupled wireless control problems.

Taken together, these limitations motivate the following research question:

How can we obtain data-efficient, scalable policies that solve deeply coupled wireless optimization problems while avoiding interference between heterogeneous subtasks?

Addressing this question calls for architectures that (i) decompose complex decisions into specialized subtasks, (ii) increase representational capacity in a controlled way, and (iii) maintain stable learning for continuous, multimodal control. Mixture-of-experts architectures are a natural candidate because they distribute capacity across specialized experts via learned routing, which reduces cross-task interference and enables more focused representation learning [11], [12]. Dense activation strategies that route to many or all experts improve specialization and reduce routing overlap at the cost of increased computation per example. This trade-off, that is, higher compute for stronger specialization and more robust learning, matches our goal of solving tightly-coupled wireless optimization tasks where representational fidelity and stable training matter.

To satisfy the requirements above, we propose a MoE driven decision aided DRL framework tailored for tightly-coupled wireless optimization problems that arise in multiple application domains. Specifically, MoDE rests on three design principles. (i) Task decomposition. The MoE-actor is organized in layers that reflect natural groupings of the joint action such as position selection, power allocation, and beamforming decoding. Each layer produces the layer-specific action component and is modeled by its own MoE block, which reduces the dimensionality and cross-talk facing any single module and clarifies learning targets for experts. (ii) Expert specialization. Rather than uniformly assigning the same number of experts to every layer, MoDE allows heterogeneous expert counts and internal segmentation: high-dimensional or more complex layers receive more experts, while simpler layers use fewer. We also design shared experts across layers to capture common structure. This allocation improves representational fidelity and stabilizes learning in composite action spaces. (iii) Adaptive routing and dense activation. MoDE permits routing policies that activate multiple experts when subtasks interact strongly, trading extra computation for improved representational fidelity and stable learning. Algorithmically, MoDE is unified and agnostic to deployment. The training loop is end-to-end and implemented with an off-policy, entropy-regularized backbone so that continuous, multimodal control benefits from robust exploration and sample reuse. At deployment time MoDE is flexible because it can run entirely on a single platform or be partitioned across devices.

The main contributions of this paper are summarized as follows.

- We propose the MoDE, a general MoE decision-making architecture for reinforcement learning. MoDE factorizes multi-layer, high-dimensional, and mixed discrete-continuous control into staged expert modules, enabling specialization on tightly-coupled decision blocks while remaining algorithm- and deployment-agnostic
- We realize the MoDE idea within an off-policy, entropy-regularized RL framework by integrating a layered MoE actor with Soft Actor-Critic (SAC). The resulting design supports heterogeneous expert allo-

cation, fine-grained expert segmentation and shared experts, adaptive dense routing, plus regularizers to prevent expert collapse. We further provide concrete, scenario-specific algorithmic instantiations for three representative wireless problems in pinching antenna system(PASS), Multi-User Multiple-Input Multiple-Output (MU-MIMO), and Intelligent reflecting surface (IRS)-aided MIMO.

- We perform extensive experiments that validate MoDE’s effectiveness. Compared to heuristic baselines and standard MLP-based RL, MoDE attains consistently higher spectral efficiency, faster convergence and improved sample efficiency after sufficient training in our testbed scenarios. We also show that allocating more experts to higher-dimensional layers yields additional gains over uniform expert allocation.

2 RELATED WORK

This section focuses on three canonical, tightly-coupled wireless optimization scenarios to clarify the concrete technical challenges that motivate MoDE.

2.1 Pinching Antenna positions and beamforming design in PASS.

The placement of pinching elements and the digital beamformer jointly determine the baseband-equivalent channel matrix entries, including their phases [13]. Activating different pinching elements modifies propagation paths and line-of-sight components and therefore changes per-stream channel gains and phase alignments [14], [15]. Consequently, a beamformer that provides coherent combining and high array gain for one antenna configuration may suffer phase misalignment or reduced gain under another configuration. Hence, jointly optimizing antenna placement and beamformer design generally yields better spectral efficiency and energy efficiency than optimizing them separately.

2.2 Power allocation and beamforming in MU-MIMO.

In MU-MIMO scenarios, system performance is shaped primarily by how transmit power is allocated across streams and by the selected beamforming vectors [16], [17]. Adjusting the power assigned to a stream directly changes the received signal strengths and alters the interference experienced by other users [18]. Modifying beamforming vectors changes the spatial distribution of radiated energy and thus reshapes per-user channel gains and interference geometry [19], [20]. Because power allocation and beamforming influence each other, an adjustment on one side typically changes the optimal choice on the other and the combined problem is high-dimensional and nonconvex. Therefore, treating power allocation and beamforming as a joint design problem yields better opportunities to balance throughput fairness and energy efficiency.

2.3 IRS phase and beamforming design in IRS-aided MIMO.

The configuration of IRS phase shifts and transmit beamforming collectively shapes the end-to-end amplitude and

phase response of wireless channels [2]. On one hand, tuning the IRS phase shifts modifies the composite channel response by applying a multiplicative effect to signals along reflected paths. Meanwhile, adapting the transmit beamforming alters the spatial excitation of the propagation environment, influencing how signals combine after reflection [21], [22]. Due to the fact that IRS phases and transmitter beamforming interact multiplicatively and that IRS elements have unit modulus constraints, changes on one side change the best choice on the other and the joint problem is strongly coupled and nonconvex [23], [24]. As a result, joint optimization of IRS phases and transmitter beamforming is required to reach favorable trade-offs in spectral efficiency and robustness.

3 MoE-AIDED DRL

In this section, we propose the MoE-aided DRL method. Specifically, we first discuss the MoE architecture, and then we explain how to use the agent as evaluators to feedback the sum spectral efficiency as the reward for DRL algorithms. The proposed MoE-SAC algorithm is designed as a general policy network suitable for the three representative coupled-decision problems outlined in Section 1. For clarity of exposition, we present the structure using the PASS case as a running example, while noting that the same principles apply to the other two scenarios.

3.1 MoEActor Architecture

Our policy network is a two-stage MoE that jointly predicts antenna “pinching” positions and beamforming latent codes. As shown in Figure 1, the MoEActor consists of:

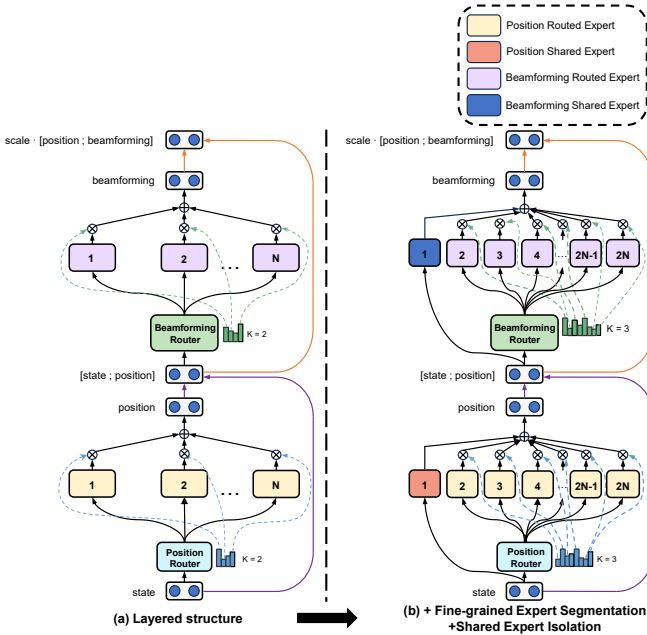


Fig. 1. Illustration of MoE for PASS. Subfigure (a) showcases the layered MoE structure. Subfigure (b) illustrates the fine-grained expert segmentation strategy and the integration of the shared expert isolation strategy.

3.1.1 Layered MoE Architecture with different number of experts

In a conventional MoE actor, all experts receive the same state input and are trained jointly to output actions for a single decision step. However, direct application creates two main issues in the application of PASS:

- Coupled learning difficulty. Position and beamforming are interdependent, but their optimal policies require different feature sensitivities. Training a single-layer MoE to handle both leads to noisy gradient updates, as position-related learning interferes with beamforming-related learning.
- Slow convergence. When the action space is large and multi-stage, the MoE router must simultaneously learn to route inputs for both tasks, which increases exploration cost and delays expert specialization.

If we separate the decision process into multiple MoE layers, each layer can specialize in its own subtask. This “layered” design transforms the learning problem from one large entangled decision into two simpler sequential decisions. This hierarchical decomposition reduces training variance and accelerates convergence while maintaining the computational efficiency of standard MoE and achieving higher specialization without interference from unrelated subtasks. In pursuit of the goal, we design a Layered MoE Actor consisting of two sequential MoE stages:

- Position MoE Layer: Routes the state to position experts, producing an optimized antenna position.
- Beamforming MoE Layer: Takes both the state and the chosen position as input, which can be expressed as:

$$\tilde{s} = [s; p] \in \mathbb{R}^{d_{\text{obs}} + d_{\text{pos}}}, \quad (1)$$

where we enrich the state with the intermediate action p , forming the augmented input and d_{obs} represents the dimensionality of the original observation vector while d_{pos} is the dimensionality of the intermediate position action p . Then we can route the enriched state \tilde{s} to beamforming experts for final signal optimization.

Crucially, we assign different numbers of experts to the two MoE layers. The position output is low dimensional and its mapping from state to antenna coordinates is relatively simple, while the beamforming mapping is high dimensional and highly nonlinear. Allocating more experts and finer grained segmentation to the beamforming layer increases representational capacity where it is most needed and enables distinct experts to specialize on diverse channel configurations. At the same time, assigning a smaller expert set to the position layer avoids unnecessary overparameterization and reduces routing overhead for the simpler subtask.

3.1.2 Fine-Grained Expert Segmentation and Shared Experts Isolation

In conventional MoE with limited experts, a single expert often has to capture highly diverse patterns of knowledge. This forces the expert parameters to capture incompatible

patterns, reducing their ability to optimize for any one pattern[12]. To address this, one strategy is to split an expert into smaller subexperts and activate more of them simultaneously. In this way, diverse knowledge can be distributed across specialized subspaces, allowing each subexpert to focus on a narrower scope of representation learning [25], [26].

However, diversity alone is not sufficient. In many routing strategies, different experts inadvertently learn overlapping common knowledge, leading to redundancy and inefficient parameter utilization. This overlap reduces the capacity available for capturing unique, specialized features. To mitigate this, we introduce dedicated shared experts that always receive input. These shared experts act as repositories of common knowledge, thereby freeing the fine-grained experts to focus exclusively on specialized features [26]. Together, segmentation and isolation create a complementary mechanism: segmentation enhances specialization, while shared experts centralize generalization.

Building on the Layered MoE Actor architecture described above, we implement two categories of experts: position experts and beamforming experts. Each category contains eight fine-grained experts and one shared expert. Every expert is parameterized as a two-hidden-layer fully connected network with separate output heads for the mean and log standard deviation of a Gaussian distribution. The feedforward computation can be expressed as:

$$h_1 = \text{ReLU}(W_1 x + b_1), \quad W_1 \in \mathbb{R}^{d_h \times d_{\text{in}}}, \quad (2)$$

$$h_2 = \text{ReLU}(W_2 h_1 + b_2), \quad W_2 \in \mathbb{R}^{d_h \times d_h}, \quad (3)$$

$$\mu = W_\mu h_2 + b_\mu, \quad W_\mu \in \mathbb{R}^{d_{\text{out}} \times d_h}, \quad (4)$$

$$\log \sigma = W_\sigma h_2 + b_\sigma, \quad W_\sigma \in \mathbb{R}^{d_{\text{out}} \times d_h}, \quad (5)$$

where $x \in \mathbb{R}^{d_{\text{in}}}$ is the input feature vector, d_{in} is the input dimension, d_h is the hidden dimension, and d_{out} is the output dimension. The parameters $W_1, W_2, W_\mu, W_\sigma$ are trainable weight matrices, and $b_1, b_2, b_\mu, b_\sigma$ are their associated bias vectors. The outputs $\mu \in \mathbb{R}^{d_{\text{out}}}$ and $\log \sigma \in \mathbb{R}^{d_{\text{out}}}$ represent the mean and log standard deviation of the Gaussian distribution, respectively.

The output of one expert feedforward network is:

$$\text{ExpertFFN}(x) = (\mu, \log \sigma).$$

Hence, Fine-Grained Experts outputs and Shared expert outputs can be expressed as:

$$(\mu_i^{\text{pos}}, \log \sigma_i^{\text{pos}}) = E_i^{\text{pos}}(s), \quad i = 1 \dots n, \quad (6)$$

$$(\mu_j^{\text{bf}}, \log \sigma_j^{\text{bf}}) = E_j^{\text{bf}}(\tilde{s}), \quad j = 1 \dots m, \quad (7)$$

$$(\mu_k^{\text{pos}}, \log \sigma_k^{\text{pos}}) = E_k^{\text{pos}}(s), \quad (8)$$

$$(\mu_k^{\text{bf}}, \log \sigma_k^{\text{bf}}) = E_k^{\text{bf}}(\tilde{s}), \quad (9)$$

where n and m denote the segmentation factors for position and beamforming experts, respectively. Specifically, each position expert FFN is split into n sub-experts by reducing the hidden dimension of each FFN to $\frac{1}{n}$ of its original size. To maintain constant computational cost, the number of simultaneously activated sub-experts is increased to n times

the original. An analogous segmentation is applied to the beamforming experts with factor m .

Each MoE layer employs a lightweight router that maps the layer input x to per-expert gating logits. The logits are converted to a normalized weight vector by a softmax:

$$\mathbf{w} = \frac{\exp((R\mathbf{x} + \mathbf{r})_i)}{\sum_{j=1}^E \exp((R\mathbf{x} + \mathbf{r})_j)} \in \mathbb{R}^E,$$

where $R \in \mathbb{R}^{E \times d_{\text{in}}}$ and $\mathbf{r} \in \mathbb{R}^E$ are the router parameters, d_{in} is the input dimension, and E is the number of fine-grained experts in that layer. The resulting weights satisfy $w_i \geq 0$ and $\sum_{i=1}^E w_i = 1$ for each sample. Using the per-sample weights w_i , the mixed distribution parameters for the position layer are computed by adding the shared expert output and the weighted sum of fine experts:

$$\mu^{\text{pos}}(s) = \mu_{\text{shared}}^{\text{pos}}(s) + \sum_{i=1}^{n_{\text{fine}}} w_i^{\text{pos}}(s) \mu_i^{\text{pos}}(s), \quad (10)$$

$$\log \sigma^{\text{pos}}(s) = \log \sigma_{\text{shared}}^{\text{pos}}(s) + \sum_{i=1}^{n_{\text{fine}}} w_i^{\text{pos}}(s) \log \sigma_i^{\text{pos}}(s). \quad (11)$$

We sample the intermediate position action:

$$p = \tanh(\mu^{\text{pos}}(s) + \sigma^{\text{pos}}(s) \cdot \epsilon), \quad \epsilon \sim \mathcal{N}(0, I). \quad (12)$$

Here, $p \in \mathbb{R}^{d_{\text{pos}}}$ represents the antenna positioning decision.

For the beamforming control stage, we enrich the state with the intermediate action p , forming the augmented input as(1). The beamforming action distribution parameters are computed similarly:

$$\mu^{\text{bf}}(\tilde{s}) = \mu_{\text{shared}}^{\text{bf}}(\tilde{s}) + \sum_{j=1}^{n_{\text{fine}}} w_j^{\text{bf}}(\tilde{s}) \mu_j^{\text{bf}}(\tilde{s}), \quad (13)$$

$$\log \sigma^{\text{bf}}(\tilde{s}) = \log \sigma_{\text{shared}}^{\text{bf}}(\tilde{s}) + \sum_{j=1}^{n_{\text{fine}}} w_j^{\text{bf}}(\tilde{s}) \log \sigma_j^{\text{bf}}(\tilde{s}). \quad (14)$$

We then sample the beamforming action:

$$b = \tanh(\mu^{\text{bf}}(\tilde{s}) + \sigma^{\text{bf}}(\tilde{s}) \cdot \epsilon'), \quad \epsilon' \sim \mathcal{N}(0, I). \quad (15)$$

Here, $b \in \mathbb{R}^{d_{\text{bf}}}$ represents the beamforming coefficients.

The final action vector \mathbf{a} is the concatenation of position and beamforming actions:

$$\mathbf{a} = \text{scale} \cdot [p; b] + \text{bias}. \quad (16)$$

3.1.3 Load Balance Consideration

Automatically learned routing strategies may encounter the issue of load imbalance, which manifests two notable defects. Firstly, there is a risk of routing collapse. The router repeatedly selects only a small subset of experts for most tokens, preventing other experts from sufficient training [26]. Secondly, persistent load imbalance prevents fair utilization of all experts, which can lead to unequal convergence rates and wasted model capacity.

Diversity loss. To mitigate routing collapse, we introduce a diversity loss that encourages the router to maintain high-entropy routing distributions. This ensures that tokens

are spread more evenly across experts rather than collapsing onto a few dominant ones:

$$\bar{H}_{\text{pos}} = \frac{1}{B} \sum_{b=1}^B \left[- \sum_{i=1}^E w_{b,i}^{\text{pos}} \log w_{b,i}^{\text{pos}} \right], \quad (17)$$

$$\bar{H}_{\text{bf}} = \frac{1}{B} \sum_{b=1}^B \left[- \sum_{i=1}^E w_{b,i}^{\text{bf}} \log w_{b,i}^{\text{bf}} \right], \quad (18)$$

$$\mathcal{L}_{\text{diversity}} = -(\bar{H}_{\text{pos}} + \bar{H}_{\text{bf}}), \quad (19)$$

where B represents the batch size, E represents the number of experts in a router and $w_{b,i}^{\text{pos}}$ is normalized routing weight of position MoE for sample b and expert i while $w_{b,i}^{\text{bf}}$ is normalized routing weight of beamforming MoE for sample b and expert i .

This term is minimized when the router assigns weights more evenly within each sample, thus reducing the chance of collapse to one or two experts.

Balance loss. In addition, we define an expert-level balance loss that directly regulates the long-term average usage of experts across a batch. This term penalizes disproportionate routing and promotes equalized expert utilization, as expressed by:

$$\mu_i^{\text{pos}} = \frac{1}{B} \sum_{b=1}^B w_{b,i}^{\text{pos}}, \quad \mu_i^{\text{bf}} = \frac{1}{B} \sum_{b=1}^B w_{b,i}^{\text{bf}}, \quad (20)$$

$$\mathcal{L}_{\text{balance}} = \frac{1}{E} \sum_{i=1}^E (\mu_i^{\text{pos}})^2 + \frac{1}{E} \sum_{i=1}^E (\mu_i^{\text{bf}})^2. \quad (21)$$

This term encourages average usage to be spread across experts. This loss reaches its minimum when all experts are equally utilized on average, i.e., when $\mu_i = \frac{1}{E}$ for every expert.

We combine diversity loss and balance loss into a single expert loss term that softly guides the router towards a balanced yet specialized routing, with $\lambda = 10^{-3}$ in our experiments, balancing the stronger anti-collapse effect with a softer global balancing term.

$$\mathcal{L}_{\text{expert}} = \underbrace{\mathcal{L}_{\text{diversity}}}_{\text{anti-collapse}} + \lambda \underbrace{\mathcal{L}_{\text{balance}}}_{\text{load equalization}}. \quad (22)$$

3.2 DRL with SAC feedback Framework

SAC is explicitly formulated for reinforcement-learning tasks with continuous actions and employs entropy-regularized policy optimization to promote robust exploration, while its off-policy training reuses prior experience to improve data efficiency[27]. These properties make SAC a strong fit for continuous placement control under time-varying radio environments, enabling adaptive, sample-efficient updates[28]. For these reasons we adopt SAC as the RL backbone and integrate it with a layered MoE actor in this work to provide large representational capacity while preserving training stability. The general algorithm for implementing SAC is shown as Algorithm 1. Specifically, the management model initializes with parameters θ, ϕ, ψ and is trained through interactions with K generative agents that simulate user feedback. Training continues for E episodes; each episode e starts from an initial state s

and iterates until a terminal condition is met. A terminal condition indicates episode completion and may correspond to task success, a maximum step budget, or a failure event. At each decision step, agents sample an action a from the policy $\pi_{\theta}(s)$ and receive rewards formed by assessments, where each Agent_k outputs a user-centric utility, that is, per-step spectral efficiency. State transitions (s, a, r, s') are stored in an off-policy replay buffer and sampled to perform gradient updates of the policy and critics. The agent learns by minimizing an entropy-regularized objective, which promotes stable learning and robust exploration in continuous action spaces. Through iterative training over E episodes, the management model progressively refines its decision-making capability, leveraging simulated user feedback to converge toward a robustly trained management model.

Algorithm 1 Deep Reinforcement Learning using Soft Actor-Critic

Initialize: Actor-critic networks with parameters θ, ϕ, ψ ; replay buffer \mathcal{B} ; SAC empowered generative agents K to simulate K users

Output: Trained actor-critic networks θ, ϕ, ψ

```

1 for each episode  $e = 1, 2, \dots, E$  do
2   Initialize state  $s$  while  $s$  is not terminal do
3     Generate action  $a \leftarrow \pi_{\theta}(s)$ 
4     Obtain reward  $r \leftarrow \sum_{k=1}^K \text{Agent}_k(s, a)$ 
5     Observe next state  $s'$ 
6     Store transition  $(s, a, r, s')$  in replay buffer  $\mathcal{B}$ 
7     Sample a minibatch of transitions from  $\mathcal{B}$ 
8     Update critic parameters  $\phi, \psi$  by minimizing the soft
       Bellman residual
9     Update policy parameters  $\theta$  via gradient ascent on
       the expected soft value
10    Adjust temperature parameter  $\alpha$  to match the target
       entropy
11     $s \leftarrow s'$ 
```

4 CASES

In this section we describe the three representative application scenarios and explain how the proposed layered MoE actor and the SAC backbone can be applied in each case. We perform experiments for the pinching-antenna placement and beamforming problem and outline how the same architecture generalizes to MU-MIMO and IRS-aided-MIMO.

4.1 PASS

In this section we evaluate the proposed MoE-aided SAC agent on the pinching-antenna placement and beamforming problem and then show its effectiveness.

4.1.1 User-centric Spectral-Efficiency Maximization Problem

We consider a downlink multi-user scenario with N parallel waveguides at the base station and K single-antenna users deployed in a square region of side length D . Each waveguide carries M pinching antennas positioned along

its length; the x-coordinate of the m -th antenna on waveguide n is denoted $x_{m,n}^p \in [0, D]$ and the waveguide height is d .

The objective is user-centric in the sense that the agent seeks to maximize an aggregate utility of the instantaneous per-user spectral efficiencies. Let $R_k(s, a)$ denote the instantaneous spectral efficiency of user k under state s and action a . The optimization objective at each decision step is

$$\max_a \sum_{k=1}^K R_k(s, a). \quad (23)$$

The physical propagation model used in the environment follows the structure in PASS. The complex channel from all pinching antennas to user k is collected in the row vector $\mathbf{h}_k^H \in \mathbb{C}^{1 \times NM}$, whose entries are small-scale line-of-sight gains. For the m -th antenna on waveguide n the entry is

$$h_{n,m,k} = \frac{\sqrt{\eta} \exp(-j \frac{2\pi}{\lambda} \|\psi_k - \psi_{m,n}^p\|)}{\|\psi_k - \psi_{m,n}^p\|},$$

where $\psi_k = (x_k, y_k, 0)$ is the user location, $\psi_{m,n}^p = (x_{m,n}^p, y_n^p, d)$ is the antenna location, $\lambda = \frac{c}{f_c}$ is the free-space wavelength, and $\eta = \frac{c}{2\pi f_c}$ is the propagation constant used in our implementation. The feed-point phase shifts along each waveguide are collected into a per-waveguide vector \mathbf{g}_n , and the block-diagonal pinching matrix $\mathbf{G} = \text{blockdiag}(\mathbf{g}_1, \dots, \mathbf{g}_N)$ captures the per-antenna phase reweighting induced by the waveguide feed and the specific pinching positions.

Given a baseband precoder matrix $\mathbf{W} = [\mathbf{w}_1, \dots, \mathbf{w}_K] \in \mathbb{C}^{NM \times K}$ and transmitted symbols x_k , the signal transmitted toward user k is $\mathbf{s}_k = \mathbf{G} \mathbf{w}_k x_k$. The received scalar signal at user k is therefore

$$y_k = \mathbf{h}_k^H \mathbf{s}_k + \sum_{j \neq k} \mathbf{h}_k^H \mathbf{s}_j + n_k,$$

where $n_k \sim \mathcal{CN}(0, \sigma_0^2)$ is complex Gaussian noise with variance σ_0^2 .

From this model the instantaneous signal-to-interference-plus-noise ratio (SINR) for user k is

$$\text{SINR}_k(\Phi^p, \mathbf{W}) = \frac{|\mathbf{h}_k^H \mathbf{G} \mathbf{w}_k|^2}{\sum_{j \neq k} |\mathbf{h}_k^H \mathbf{G} \mathbf{w}_j|^2 + \sigma_0^2}, \quad (24)$$

and the per-user spectral efficiency is

$$R_k(\Phi^p, \mathbf{W}) = \log_2(1 + \text{SINR}_k(\Phi^p, \mathbf{W})). \quad (25)$$

The agent's learning objective is to maximize an aggregate user utility based on the instantaneous spectral efficiencies. The per-step spectral efficiencies is written as

$$\text{SE}_t = \sum_{k=1}^K R_k(\Phi_t^p, \mathbf{W}_t), \quad (26)$$

where $\text{Agent}_k(\cdot)$ is a per-user utility function.

4.1.2 Beamforming design

The beamforming latent produced by the actor, denoted $z_w \in \mathbb{R}^L$, is decoded by a small parametric PowerAllocator network into two sets of parameters: a per-user positive uplink factor vector $\boldsymbol{\lambda} = [\lambda_1, \dots, \lambda_K]^T$ and a normalized

downlink power fraction vector $\mathbf{p} = [p_1, \dots, p_K]^T$ with $\sum_k p_k = 1$. In the implementation the mapping is

$$\boldsymbol{\lambda} = \text{softplus}(f_\theta(z_w)) + \epsilon, \quad \mathbf{p} = \text{softmax}(g_\theta(z_w)), \quad (27)$$

where f_θ and g_θ are the linear heads of the PowerAllocator network and $\epsilon > 0$ is a small constant for numerical stability. The PowerAllocator architecture and activation choices match the implementation in the environment.

Given the equivalent channel $\tilde{\mathbf{H}} = \mathbf{H} \mathbf{G} \in \mathbb{C}^{K \times NM}$ that already incorporates the pinching-phase matrix \mathbf{G} , the implemented precoder follows a regularized linear structure

$$\mathbf{W} = \tilde{\mathbf{H}}^H (\Lambda \tilde{\mathbf{H}} \tilde{\mathbf{H}}^H + \sigma^2 \mathbf{I}_K)^{-1} \mathbf{P}^{1/2}, \quad (28)$$

where $\Lambda = \text{diag}(\boldsymbol{\lambda})$ and $\mathbf{P}^{1/2} = \text{diag}(\sqrt{p_1}, \dots, \sqrt{p_K})$. The inverse is implemented as a direct matrix inverse with fallback to a pseudo-inverse when necessary to avoid numerical singularities. Finally, \mathbf{W} is normalized by its Frobenius norm to satisfy the total transmit-power constraint:

$$\mathbf{W}' = \frac{\mathbf{W}}{\|\mathbf{W}\|_F}. \quad (29)$$

With \mathbf{W} and $\tilde{\mathbf{H}}$ computed, we evaluate the SINR of each user using (24), in the implementation the term $\mathbf{G} \mathbf{w}_k$ replaced by the corresponding columns of $\tilde{\mathbf{H}}^H \mathbf{W}$, and the resulting per-user rates determine the scalar reward defined in subsection 4.1.2.

4.1.3 MoE-DRL for PASS optimization

To address the non-convex challenge outlined in section 1, we design a layered MoE actor composed of two sequential stages. The first stage proposes pinching positions for the antennas, and the second produces beamforming latent codes. Below we detail the SAC state space, action space, and reward function:

- **State s :** The state is designed to include the environment and configuration variables relevant to antenna placement and beamforming. Concretely, the state contains user positions, current pinching positions, and the complex channel matrix represented by its real and imaginary parts. We therefore write the state space as

$$\mathcal{S} = \{\{\mathbf{k}_i\}, \{\mathbf{p}_i\}, \{r_i\}\}, \quad (30)$$

where user positions, stacked as $(x_1, y_1, \dots, x_K, y_K)$, length $2K$; current pinching positions flattened over waveguides and antennas, length NM ; the complex channel state \mathbf{H} represented by its real and imaginary parts and flattened to length $2KMN$.

- **Action a :** The action space \mathcal{A} is defined as the optimization process of pinching antenna position together with beamforming matrices \mathbf{W} , which concatenates normalized initial per-waveguide positions, relative inter-antenna increments, and the beamforming latent vector z_w . The continuous action $a_t \in \mathbb{R}^A$ is partitioned as

$$a_t = [\mathbf{x}_1^{(n)}, \Delta^{(n,m)}, z_w],$$

where $\mathbf{x}_1^{(n)} \in \mathbb{R}^N$ contains normalized first-antenna positions per waveguide, in the agent output range

$[-1, 1]$ and mapped to physical coordinates by $x_{1,n}^p = (x_{1,n}^{(n)} + 1)D/2$; $\Delta^{(n,m)} \in \mathbb{R}^{N(M-1)}$ are normalized relative increments for subsequent antennas on each waveguide; each normalized delta is mapped to a feasible incremental distance that enforces the minimum spacing Δ_{\min} and ensures the last antenna remains within $[0, D]$; $z_w \in \mathbb{R}^L$ is the beamforming latent vector decoded by a small PowerAllocator network into per-user uplink factors $\{\lambda_k\} > 0$ and normalized downlink power fractions $\{p_k\}$ with $\sum_k p_k = 1$.

- **Reward r :** The reward function takes into account the objective function, that is, the sum spectral efficiency R_k . It is designed as aggregating per-user utilities, and In the environment implementation used for training and evaluation, we take the per-episode step reward to be proportional to the sum spectral efficiency with a small improvement bonus, namely

$$r_t = 10 \left(\text{SE}_t + 0.1(\text{SE}_t - \text{SE}_{t-1}) \right), \quad (31)$$

where SE_t is defined in (26) and SE_{t-1} denotes the previous step value; the increment term is omitted at the first step; the multiplicative factor is a numeric scaling chosen for training stability.

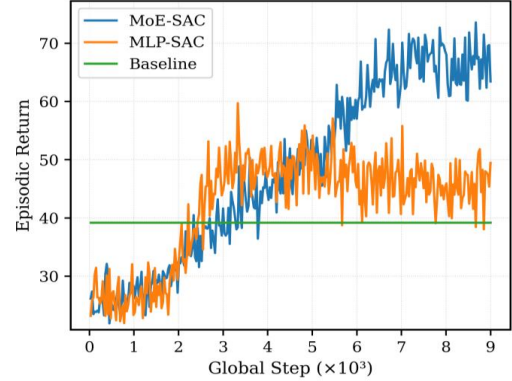
4.1.4 Experimental Results

We evaluate the proposed MoE-SAC agent on the joint pinching-antenna placement and beamforming task. Consider a PASS with N waveguides, each equipped with M pinching antennas. A total of K users are uniformly distributed in a $100 \times 100 \text{ m}^2$ square region. The waveguide height is set to $d = 3 \text{ m}$, the carrier frequency is $f_c = 28 \text{ GHz}$, the effective refractive index is $n_{\text{eff}} = 1.4$, and minimum inter-antenna spacing is $\Delta_{\min} = \lambda_g$. Without loss of generality, We consider multiple training scenarios to demonstrate robustness across system sizes.

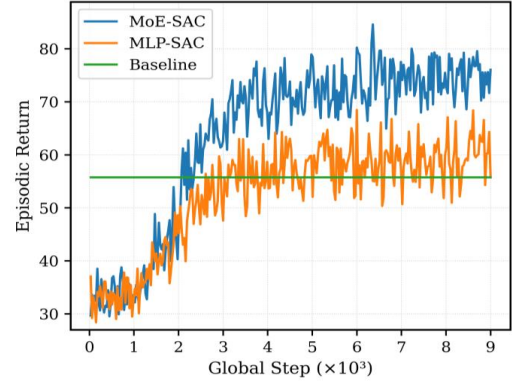
Figure 2 plots the cumulative episodic return versus training iterations for the MoE-SAC agent, MLP-SAC agent and a deterministic baseline under identical random seeds and hyperparameters.. In the baseline, the first antenna on each waveguide is aligned with the nearest user in the x -direction, the remaining antennas are placed at the minimum spacing, and the transmit beamforming is implemented using zero-forcing (ZF). It plots the scenario where $N = 2, M = 2$ and $K = 4$ and the scenario using $N = 7, M = 3$ and $K = 13$. In both scenarios the MoE-SAC agent attains higher final episodic return than the MLP SAC agent, and it surpasses the deterministic baseline after sufficient training.

Figure 3 focuses on the effect of expert allocation across the two MoE layers and displays the learning curves for these two allocations together with the deterministic baseline. Here we fix the nominal expert count to four and compare two layerwise allocations. Configuration A uses fine grained segmentation factor one for the position layer and factor four for the beamforming layer, which yields four experts for the position stage and sixteen fine grained experts for the beamforming stage. Configuration B uses segmentation factor two for both layers, which yields eight experts in each layer. The results show that allocating more experts to the beamforming layer produces higher

converged episodic return than the symmetric allocation. The symmetric allocation still improves slightly upon the deterministic baseline, but it is outperformed by the beamforming heavy configuration.



(a) $N = 2, M = 2, K = 4$



(b) $N = 7, M = 3, K = 13$

Fig. 2. Comparison of average return versus training steps for MoE-SAC, MLP-SAC and the heuristic baseline in PASS.

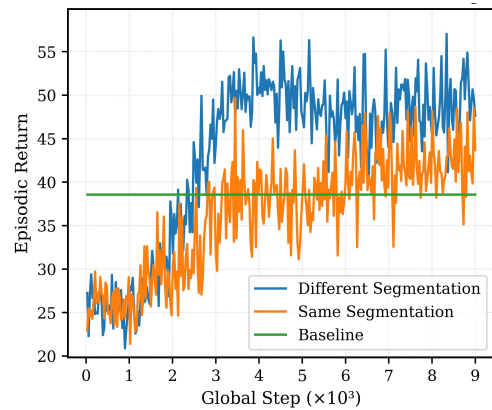


Fig. 3. Average return versus training steps for MoE with different configuration of experts.

4.2 MU-MIMO Power Allocation and Beamforming Design

4.2.1 MoE-DRL for PASS Optimization

We consider a downlink multi-user MIMO system where a base station equipped with N_t transmit antennas serves

K single-antenna users. The instantaneous channel from the base station to user k is denoted $\mathbf{h}_k \in \mathbb{C}^{N_t}$, with the aggregate channel matrix $\mathbf{H} = [\mathbf{h}_1, \dots, \mathbf{h}_K]$. The joint optimization of power allocation and beamforming parameters directly determines the achievable signal-to-interference-plus-noise ratios across users.

The objective is to maximize a network utility function defined over instantaneous user rates. Let $R_k(s, a)$ represent the spectral efficiency of user k under state s and action a . We formulate the weighted sum-rate maximization problem as the equation (23).

To capture the distinct aspects of power allocation and beamforming design, the actor employs a layered mixture of experts. The first layer processes the observation through experts that output user power allocation factors of dimension K . The second layer concatenates the original observation with these power factors and routes the augmented input to beamforming experts that generate latent representations for precoding of dimension 24. This hierarchical structure ensures that power allocation decisions guide the subsequent beamforming optimization in an integrated manner.

In our soft actor-critic implementation, the state encompasses current channel measurements \mathbf{H} , user geographical positions, and antenna array geometry. The continuous action vector encodes both beamforming parameters and relaxed representations of discrete configuration choices, maintaining end-to-end differentiability. The reward function directly corresponds to the objective in Equation (31), with entropy regularization in SAC encouraging exploration across diverse configuration-beamforming combinations and enhancing robustness in multimodal optimization landscapes.

4.2.2 Experimental Results

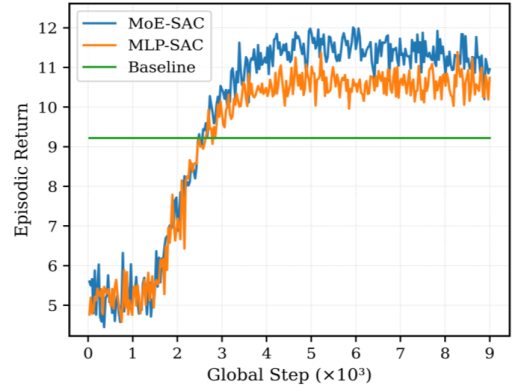
We evaluate the proposed MoE-SAC agent on the joint power allocation and beamforming optimization task in MU-MIMO systems. The base station is equipped with N transmit antennas serving K single-antenna users uniformly distributed in a $D \times D$ square meter area. The system operates at a carrier frequency of $f_c = 28$ GHz with a path loss exponent of 2.5. We consider different training scenarios to demonstrate robustness across different system configurations.

Figure 4 presents the average return versus training steps for the MoE-SAC agent, MLP-SAC agent, and a heuristic baseline under identical random seeds and hyperparameters. The baseline employs MMSE precoding combined with water-filling power allocation. The results show when $N = 7$, $K = 13$, $D = 50$ and when $N = 11$, $K = 18$, $D = 40$, the MoE-SAC agent demonstrates faster convergence speed and achieves a higher final performance compared to the MLP-SAC agent, with both significantly outperforming the baseline after sufficient training.

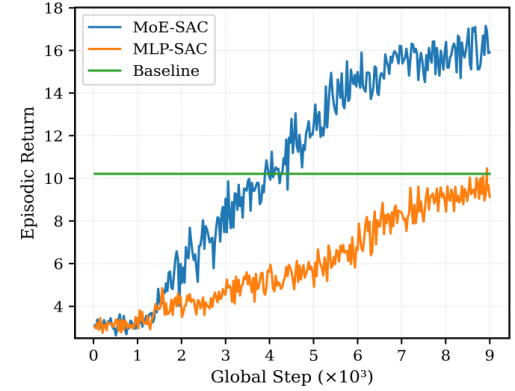
4.3 IRS-Aided MIMO Joint IRS Phase and Beamforming Optimization

4.3.1 MoE-DRL for IRS Optimization

We consider an IRS-assisted downlink MIMO system where an access point (AP) equipped with M transmit antennas serves K single-antenna users via both a direct link and an



(a) $N = 11$, $K = 18$, $D = 40$



(b) $N = 7$, $K = 13$, $D = 50$

Fig. 4. Comparison of average return versus training steps for MoE-SAC, MLP-SAC and the heuristic baseline in MU-MIMO.

intelligent reflecting surface (IRS) with N passive reflecting elements. Let $\mathbf{G} \in \mathbb{C}^{N \times M}$ denote the AP to IRS channel, $\mathbf{d}_k \in \mathbb{C}^{M \times 1}$ the direct AP to user k channel, and $\mathbf{r}_k \in \mathbb{C}^{N \times 1}$ the IRS to user k channel. $\Phi = \text{diag}(e^{j\theta_1}, \dots, e^{j\theta_N})$ denotes the IRS diagonal phase-shift matrix, the effective channel from the AP to user k , which is a column vector of length M , is given by

$$\mathbf{h}_k^H = \mathbf{r}_k^H \Phi \mathbf{G} + \mathbf{d}_k^H.$$

The joint optimization of IRS phase $\{\theta_1, \theta_2, \dots, \theta_N\}$ and the AP precoder $\mathbf{W} \in \mathbb{C}^{M \times K}$ determines the end-to-end SINR and hence the achievable spectral efficiencies.

The objective is to maximize the network sum-rate over instantaneous user rates. Let $R_k(s, a)$ represent the spectral efficiency of user k under state s and action a . We formulate the weighted sum-rate maximization problem as the equation(23).

To handle the mixed nature of IRS phase adjustments and continuous beamforming design, the actor employs a layered mixture of experts architecture. The first layer processes the environmental observation through specialized experts that generate IRS phase configurations, outputting both sine and cosine components of phase shifts with dimension $2N$. The second layer concatenates the original observation with these phase parameters and routes the augmented input to beamforming experts that produce latent representations for precoder optimization of dimension 24. This hierarchical decomposition ensures coordinated

optimization where IRS phase adjustments establish favorable propagation conditions that the beamforming module subsequently exploits.

In our soft actor-critic implementation, the state incorporates previous IRS phase shifts, current effective channel measurements \mathbf{H}_k , and user positions. The continuous action vector encodes both IRS phase parameters and beamforming latent variables, maintaining full differentiability throughout the optimization pipeline. The reward function directly corresponds to the objective in Equation (31), with entropy regularization in SAC promoting exploration across diverse IRS-beamforming configurations and enhancing performance in complex multimodal optimization landscapes.

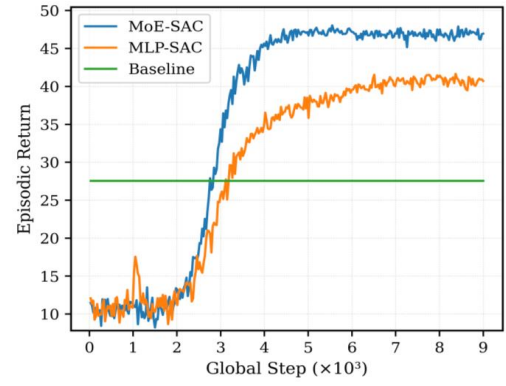
4.3.2 Experimental Results

We evaluate the proposed MoE-SAC agent on the joint IRS phase and beamforming optimization task in IRS-aided downlink MISO systems. The system configuration includes an access point with M transmit antennas, an intelligent reflecting surface with N passive elements, and K single-antenna users uniformly distributed in a $30 \times 30 \text{ m}^2$ square region. The system operates at a carrier frequency of $f_c = 3.5 \text{ GHz}$ with path loss exponents of 3.0 for direct links and 2 for IRS-reflected paths. The reference path loss is set to $P_{lo} = -20 \text{ dB}$ at $d_o = 1 \text{ m}$, with an additional IRS gain of 10 dB enhancing the reflected signals. Two environment configurations are considered to demonstrate robustness across scales.

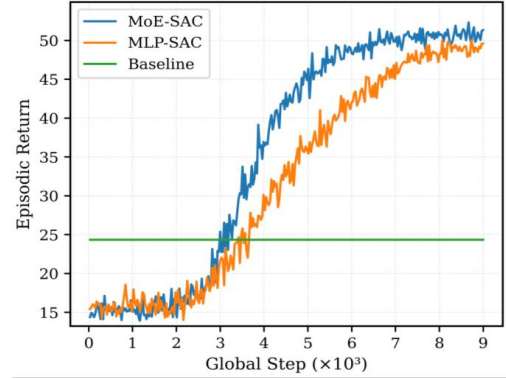
Figure 5 present the average return versus training steps for the MoE-SAC agent, MLP-SAC agent, and an alternating optimization baseline under identical random seeds and hyperparameters. The baseline operates on local copies of environment channels and alternately updates the MMSE precoder and per-element IRS phases via closed-form phase updates. Figure 5 corresponds to the scenarios with $N = 24$, $M = 2$, $K = 3$, and with $N = 36$, $M = 4$, and $K = 6$. The results show that the MoE-SAC agent demonstrates faster convergence speed and achieves higher final performance compared to the MLP-SAC agent. In both configurations, the proposed MoE-SAC and MLP-SAC agents significantly outperform the alternating optimization baseline after sufficient training.

5 CONCLUSION

We proposed a novel approach to enhancing user spectral efficiency in next-generation wireless networks, focusing on joint physical-layer configuration and resource allocation. Our primary contribution is the MoE-SAC algorithm, designed to address the tightly coupled optimization challenges in high-dimensional, non-convex decision spaces. By employing an MoE-aided agent, our framework provides real-time, sample-efficient feedback on SE, reflecting both antenna placement and beamforming adaptation, which naturally extends to other coupled wireless problems where the layered MoE assigns specialized experts to complementary subtasks such as power allocation and precoding in MU-MIMO and IRS phase and precoding in IRS-aided-MIMO. Our methodology demonstrates the potential of combining MoE architectures with DRL to enhance adaptive



(a) $N = 24, M = 2, K = 3$



(b) $N = 36, M = 4, K = 6$

Fig. 5. Comparison of average return versus training steps for MoE-SAC, MLP-SAC and the heuristic baseline in IRS-aided-MIMO.

resource allocation, paving the way for real-time resource management in large-scale, heterogeneous 6G wireless networks.

REFERENCES

- [1] H. F. Alhashimi, M. N. Hindia, K. Dimiyati, E. B. Hanafi, N. Safie, F. Qamar, K. Azrin, and Q. N. Nguyen, "A survey on resource management for 6g heterogeneous networks: current research, future trends, and challenges," *Electronics*, vol. 12, no. 3, p. 647, 2023.
- [2] Q. Wu and R. Zhang, "Intelligent reflecting surface enhanced wireless network via joint active and passive beamforming," *IEEE transactions on wireless communications*, vol. 18, no. 11, pp. 5394–5409, 2019.
- [3] Y. Zeng, R. Zhang, and T. J. Lim, "Wireless communications with unmanned aerial vehicles: Opportunities and challenges," *IEEE Communications magazine*, vol. 54, no. 5, pp. 36–42, 2016.
- [4] Y. Mao, C. You, J. Zhang, K. Huang, and K. B. Letaief, "A survey on mobile edge computing: The communication perspective," *IEEE communications surveys & tutorials*, vol. 19, no. 4, pp. 2322–2358, 2017.
- [5] F. Guo, F. R. Yu, H. Zhang, X. Li, H. Ji, and V. C. Leung, "Enabling massive iot toward 6g: A comprehensive survey," *IEEE Internet of Things Journal*, vol. 8, no. 15, pp. 11 891–11 915, 2021.
- [6] W. Saad, M. Bennis, and M. Chen, "A vision of 6g wireless systems: Applications, trends, technologies, and open research problems," *IEEE network*, vol. 34, no. 3, pp. 134–142, 2019.
- [7] Y. Lu, Z. Zhang, and L. Dai, "Hierarchical beam training for extremely large-scale mimo: From far-field to near-field," *IEEE Transactions on Communications*, vol. 72, no. 4, pp. 2247–2259, 2023.
- [8] K. Shen, W. Yu, L. Zhao, and D. P. Palomar, "Optimization of mimo device-to-device networks via matrix fractional programming: A minorization-maximization approach," *IEEE/ACM Transactions on Networking*, vol. 27, no. 5, pp. 2164–2177, 2019.

- [9] K. Shen and W. Yu, "Fractional programming for communication systems—part i: Power control and beamforming," *IEEE Transactions on Signal Processing*, vol. 66, no. 10, pp. 2616–2630, 2018.
- [10] N. C. Luong, D. T. Hoang, S. Gong, D. Niyato, P. Wang, Y.-C. Liang, and D. I. Kim, "Applications of deep reinforcement learning in communications and networking: A survey," *IEEE communications surveys & tutorials*, vol. 21, no. 4, pp. 3133–3174, 2019.
- [11] W. Cai, J. Jiang, F. Wang, J. Tang, S. Kim, and J. Huang, "A survey on mixture of experts in large language models," *IEEE Transactions on Knowledge and Data Engineering*, 2025.
- [12] N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton, and J. Dean, "Outrageously large neural networks: The sparsely-gated mixture-of-experts layer," *arXiv preprint arXiv:1701.06538*, 2017.
- [13] A. Bereyhi, S. Asaad, C. Ouyang, Z. Ding, and H. V. Poor, "Downlink beamforming with pinching-antenna assisted mimo systems," *arXiv preprint arXiv:2502.01590*, 2025.
- [14] Z. Ding, R. Schober, and H. V. Poor, "Flexible-antenna systems: A pinching-antenna perspective," *IEEE Transactions on Communications*, 2025.
- [15] X. Xu, X. Mu, Z. Wang, Y. Liu, and A. Nallanathan, "Pinching-antenna systems (pass): Power radiation model and optimal beamforming design," *arXiv preprint arXiv:2505.00218*, 2025.
- [16] Q. Shi, M. Razaviyayn, Z.-Q. Luo, and C. He, "An iteratively weighted mmse approach to distributed sum-utility maximization for a mimo interfering broadcast channel," *IEEE Transactions on Signal Processing*, vol. 59, no. 9, pp. 4331–4340, 2011.
- [17] T. Yoo and A. Goldsmith, "On the optimality of multiantenna broadcast scheduling using zero-forcing beamforming," *IEEE Journal on selected areas in communications*, vol. 24, no. 3, pp. 528–541, 2006.
- [18] C. Li, X. Wang, L. Yang, and W.-P. Zhu, "A joint source and relay power allocation scheme for a class of mimo relay systems," *IEEE Transactions on Signal Processing*, vol. 57, no. 12, pp. 4852–4860, 2009.
- [19] E. Ali, M. Ismail, R. Nordin, and N. F. Abdulah, "Beamforming techniques for massive mimo systems in 5g: overview, classification, and trends for future research," *Frontiers of Information Technology & Electronic Engineering*, vol. 18, no. 6, pp. 753–772, 2017.
- [20] F. W. Vook, A. Ghosh, and T. A. Thomas, "Mimo and beamforming solutions for 5g technology," in *2014 IEEE MTT-S International Microwave Symposium (IMS2014)*. IEEE, 2014, pp. 1–4.
- [21] P. Wang, J. Fang, X. Yuan, Z. Chen, and H. Li, "Intelligent reflecting surface-assisted millimeter wave communications: Joint active and passive precoding design," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 12, pp. 14 960–14 973, 2020.
- [22] H. Xie, J. Xu, and Y.-F. Liu, "Max-min fairness in 5g-aided multi-cell mimo systems with joint transmit and reflective beamforming," *IEEE Transactions on Wireless Communications*, vol. 20, no. 2, pp. 1379–1393, 2020.
- [23] X. Li, J. Fang, F. Gao, and H. Li, "Joint active and passive beamforming for intelligent reflecting surface-assisted massive mimo systems," *arXiv preprint arXiv:1912.00728*, 2019.
- [24] C. Huang, A. Zappone, G. C. Alexandropoulos, M. Debbah, and C. Yuen, "Reconfigurable intelligent surfaces for energy efficiency in wireless communication," *IEEE transactions on wireless communications*, vol. 18, no. 8, pp. 4157–4170, 2019.
- [25] W. Fedus, B. Zoph, and N. Shazeer, "Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity," *Journal of Machine Learning Research*, vol. 23, no. 120, pp. 1–39, 2022.
- [26] D. Dai, C. Deng, C. Zhao, R. Xu, H. Gao, D. Chen, J. Li, W. Zeng, X. Yu, Y. Wu *et al.*, "Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models," *arXiv preprint arXiv:2401.06066*, 2024.
- [27] M. Haklidić and H. Temeltaş, "Guided soft actor critic: A guided deep reinforcement learning approach for partially observable markov decision processes," *IEEE Access*, vol. 9, pp. 159 672–159 683, 2021.
- [28] Q. Wang, K. Feng, X. Li, and S. Jin, "Precodernet: Hybrid beamforming for millimeter wave systems with deep reinforcement learning," *IEEE Wireless Communications Letters*, vol. 9, no. 10, pp. 1677–1681, 2020.