

Mixture of Experts for Decentralized Generative AI and Reinforcement Learning in Wireless Networks: A Comprehensive Survey

Yunting Xu¹, Jiacheng Wang¹, *Member, IEEE*, Ruichen Zhang¹, *Member, IEEE*, Changyuan Zhao¹, *Graduate Student Member, IEEE*, Dusit Niyato², *Fellow, IEEE*, Jiawen Kang¹, *Senior Member, IEEE*, Zehui Xiong¹, *Senior Member, IEEE*, Bo Qian³, *Member, IEEE*, Haibo Zhou¹, *Senior Member, IEEE*, Shiwen Mao¹, *Fellow, IEEE*, Abbas Jamalipour⁴, *Fellow, IEEE*, Xuemin Shen¹, *Fellow, IEEE*, and Dong In Kim⁵, *Life Fellow, IEEE*

Abstract—Mixture of Experts (MoE) has emerged as a promising paradigm for scaling model capacity while preserving computational efficiency, particularly in large-scale machine learning architectures such as large language models (LLMs). Recent advances in MoE have facilitated its adoption in wireless networks to address the increasing complexity and heterogeneity

of modern communication systems. This paper presents a comprehensive survey of the MoE framework in wireless networks, highlighting its potential in optimizing resource efficiency, improving scalability, and enhancing adaptability across diverse network tasks. We first introduce the fundamental concepts of MoE, including various gating mechanisms and the integration with generative AI (GenAI) and reinforcement learning (RL). Subsequently, we discuss the extensive applications of MoE across critical wireless communication scenarios, such as vehicular networks, unmanned aerial vehicles (UAVs), satellite communications, heterogeneous networks, integrated sensing and communication (ISAC), and mobile edge networks. Furthermore, key applications in channel prediction, physical layer signal processing, radio resource management, network optimization, and security are thoroughly examined. Additionally, we present a detailed overview of open-source datasets that are widely used in MoE-based models to support diverse machine learning tasks. Finally, this survey identifies crucial future research directions for MoE, emphasizing the importance of advanced training techniques, resource-aware gating strategies, and deeper integration with emerging 6G technologies.

Received 27 April 2025; revised 25 August 2025 and 8 October 2025; accepted 12 November 2025. Date of publication 8 December 2025; date of current version 7 January 2026. This work was supported in part by the Seatrium New Energy Laboratory, Singapore Ministry of Education (MOE) Tier 1 under Grant RT5/23 and Grant RG24/24; in part by Nanyang Technological University (NTU) Centre for Computational Technologies in Finance (NTU-CCTF); in part by the Research Innovation and Enterprise (RIE) 2025 Industry Alignment Fund-Industry Collaboration Projects (IAF-ICP) administered by the Agency for Science, Technology and Research (A*STAR) under Award I2301E0026; in part by the Ministry of Science and ICT (MSIT), South Korea, under the ICT Creative Consilience Program under Grant IITP-2020-0-01821; in part by the ITRC Support Program supervised by the Institute for ICT Planning and Evaluation (IITP) under Grant IITP-2023-RS-2023-00258639; in part by the National Natural Science Foundation of China (NSFC) under Grant 62271244 and Grant 62572132; and in part by the JSPS Grant-in-Aid for Early-Career Scientists under Grant JP25K21195. (*Corresponding author: Haibo Zhou.*)

Yunting Xu, Jiacheng Wang, Ruichen Zhang, Changyuan Zhao, and Dusit Niyato are with the College of Computing and Data Science, Nanyang Technological University, Singapore 639798 (e-mail: yunting.xu@ntu.edu.sg; jiacheng.wang@ntu.edu.sg; ruichen.zhang@ntu.edu.sg; zhao0441@e.ntu.edu.sg; dniyato@ntu.edu.sg).

Jiawen Kang is with the School of Automation, Guangdong University of Technology, Guangzhou 510006, China (e-mail: kavinkang@gdut.edu.cn).

Zehui Xiong is with the School of Electronics, Electrical Engineering and Computer Science (EECS), Queen's University Belfast, BT7 1NN Belfast, U.K. (e-mail: z.xiong@qub.ac.uk).

Bo Qian is with the Information Systems Architecture Science Research Division, National Institute of Informatics, Tokyo 101-8430, Japan (e-mail: boqian@ieee.org).

Haibo Zhou is with the School of Electronic Science and Engineering, Nanjing University, Nanjing 210023, China (e-mail: haibozhou@nju.edu.cn).

Shiwen Mao is with the Department of Electrical and Computer Engineering, Auburn University, Auburn, AL 36849 USA (e-mail: smao@ieee.org).

Abbas Jamalipour is with the School of Electrical and Computer Engineering, The University of Sydney, NSW 2006, Australia, and also with the Graduate School of Information Sciences, Tohoku University, Sendai 980-8577, Japan (e-mail: a.jamalipour@ieee.org).

Xuemin Shen is with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada (e-mail: sshen@uwaterloo.ca).

Dong In Kim is with the Department of Electrical and Computer Engineering, Sungkyunkwan University, Suwon 16419, South Korea (e-mail: dongin@skku.edu).

Digital Object Identifier 10.1109/COMST.2025.3641591

Index Terms—Mixtures of experts, wireless networks, generative AI, reinforcement learning, large language model.

I. INTRODUCTION

A. Background

IN RECENT years, the integration of artificial intelligence (AI) and machine learning technology in wireless communication systems has undergone substantial development, notably characterized by the deployment of large generative AI (GenAI) models like Generative Pre-trained Transformer (GPT) [1], [2], [3]. As the complexity of evolutionary mobile networks escalates and the demand for enhanced wireless performance intensifies, the utilization of large GenAI models has shown significant potential in improving system efficiency and optimizing overall wireless performance [4], [5], [6], [7]. These GenAI models with billions of parameters possess the ability to process tremendous amounts of wireless data and extract intricate patterns of network behavior, which become instrumental in advancing technologies such as channel state information (CSI) prediction [8], orthogonal frequency division multiplexing (OFDM) symbol decoding [9], and transmit power optimization [10]. The employment of GenAI models

has effectively addressed the increasing heterogeneity and dynamism of modern networks, facilitating more efficient resource management, capacity expansion, and enhancement of user experiences [11], [12], [13].

However, despite the substantial potential of GenAI models, their large-scale parameter size presents significant challenges for practical deployment in wireless environments. Prominent examples include LLaMA from Meta, which comprises approximately 65 billion parameters [14], PaLM from Google with 540 billion parameters [15], and GPT-4, which exceeds trillions of parameters [16]. Such immense sizes necessitate considerable computational and storage resources, making it difficult for resource-constrained wireless infrastructures such as access points (APs), edge servers, and mobile devices to directly support models of this magnitude [17], [18], [19], [20]. To alleviate the challenge of resource limitations while guaranteeing computational accuracy, distributed deployment and parallel computing strategies have been widely investigated for large-scale GenAI models [21], [22], [23]. Nonetheless, the distributed design often lacks cooperation among devices, leading to limited scalability and inefficiency [24], [25], [26]. Addressing these constraints necessitates a unified framework that combines the strengths of diverse paradigms, emphasizing localized specialization and cooperative mechanisms, especially in heterogeneous and dynamic wireless environments.

B. Motivation

The framework of mixture of experts (MoE) has emerged as an effective paradigm to address diverse computational demands and resource constraints for wireless communication systems [27]. The fundamentals of MoE are to decompose a large model into multiple expert modules, each functioning as an independent sub-network specialized for specific inputs [41]. The expert modules are dynamically managed by a gating mechanism that selectively activates only a subset of experts based on the characteristics of the input data, rather than activating all modules simultaneously. Compared with traditional AI, GenAI models require more computational resources for training due to their complex mission objectives and inference procedures. State-of-the-art GenAI models, such as DeepSeek [69], GLaM [48], and Switch Transformer [50], have effectively utilized MoE frameworks to expand model capacity while minimizing computational costs. The dynamic expert selection and sparse activation mechanism of MoE can effectively mitigate computational complexity and improve energy efficiency for GenAI models, enabling resource-constrained wireless infrastructures to enhance scalability and reduce resource consumption in complex wireless task [42].

Another significant advantage of MoE architecture lies in its ability to perform distributed computing and enable collaborative strategies over local experts [94]. The decentralized implementation of MoE experts enables independent processing, inference, and fine-tuning of local data, enhancing specialization by effectively capturing the unique characteristics of personal input [69]. Through using gating mechanisms, the collaborative strategies inherent in the MoE architecture empower edge nodes and wireless devices to collaborate in

dynamic and adaptive manner [47]. Moreover, these collaborative weighting and activation mechanisms present extensibility for multimodal and multi-task processing [125], [126], [127]. By integrating multimodal data in wireless environments, such as signal strength, antenna gain, and channel state, MoE can significantly enhance network perception and optimization capabilities based on a shared expert pool to accommodate the complex wireless modalities [55]. For multi-task processing in heterogeneous wireless networks, different tasks often exhibit interdependencies, where tasks such as power control [128] and beamforming [129] aim to enhance signal-to-noise ratio (SNR), and tasks such as spectrum [130] and channel allocation [131] may be constrained by limited resource. With the dynamic expert selection and task-specific delegation, MoE provides a robust solution to the challenges of task conflicts and relationship modeling, achieving significant improvements in terms of efficiency and adaptability for wireless tasks that require simultaneous optimization of multiple metrics [132].

C. Comparisons With Related Surveys and Contributions

The framework of MoE has demonstrated substantial potential for optimizing wireless resources, improving energy efficiency, and supporting heterogeneous network management. This paper aims to provide a comprehensive survey on the fundamentals of MoE, along with its applications in GenAI and a broad range of wireless scenarios and technologies. Table I presents an in-depth comparison of related surveys, emphasizing MoE, GenAI, and their implementation in wireless networks. These surveys have primarily focused on summarizing the evolution and optimization of MoE. For instance, the authors in [27] reviewed the historical development of MoE, covering fundamental concepts, statistical properties, and practical applications across various domains. The authors in [28] investigated MoE's core principles, including expert specialization and routing mechanisms with a focus on the field of deep learning. The authors in [29] classified MoE frameworks into mixture of implicitly localized experts (MILE) and mixture of explicitly localized experts (MELE), discussing their applications in regression and classification tasks. Furthermore, the authors in [30] explored the optimization of MoE across system and hardware levels, investigating techniques such as expert selection, load balancing, and distributed computing strategies. For GenAI, existing studies have contributed to enhancing the efficiency of large models with the framework of MoE. For example, the authors in [31] explored the transformative impact of MoE and multimodal learning on GenAI, discussing the role of MoE in improving training and load balancing for GenAI tasks. The authors in [32] and [33] reviewed training techniques for large language models (LLMs), such as fine-tuning, load balancing strategies, gradient routing optimization, and efficient expert pruning.

Additionally, several studies have investigated the integration of GenAI and wireless communication networks. The authors in [34] surveyed the application of GenAI in physical layer communications, underscoring its role in improving channel estimation, equalization, signal classification, and emerging technologies such as reconfigurable intelligent surface (RIS) and beamforming. Similarly, the authors in [35]

TABLE I
SUMMARY OF RELATED SURVEY

| Ref. | Description | MoE | GenAI | Wireless Networks |
|-------------|--|-----|-------|-------------------|
| [27] | Reviews on the two-decade evolution of MoE, covering foundational concepts, key advancements, and applications in model selection and expert structure. | ✓ | ✗ | ✗ |
| [28] | Traces the development of MoE, highlighting key advances and its role in deep learning, with a focus on expert specialization and optimization techniques. | ✓ | ✗ | ✗ |
| [29] | Categorizes MoE models into mixture of implicitly localised experts (MILE) and mixture of explicitly localised experts (MELE), discussing their advantages, challenges, and applications in regression and classification tasks. | ✓ | ✗ | ✗ |
| [30] | Explores optimization techniques for MoE inference, focusing on architectural, system, and hardware-level approaches to improve efficiency and scalability. | ✓ | ✗ | ✗ |
| [31] | Investigates the transformative impact of MoE and multimodal learning on GenAI models, focusing on the advanced methodologies of MoE in enhancing GenAI. | ✓ | ✓ | ✗ |
| [32] | Introduces the MoE framework, categorizes MoE models, and explores the applications of different gating mechanisms in LLMs. | ✓ | ✓ | ✗ |
| [33] | Reviews techniques for enhancing the efficiency of LLMs, focusing on model-centric, data-centric, and framework-centric methods to reduce resource demands. | ✓ | ✓ | ✗ |
| [34] | Investigates the applications of GenAI in physical layer communications, such as channel estimation and signal classification, and compares GenAI's benefits over traditional AI models. | ✗ | ✓ | ✓ |
| [35] | Provides a comprehensive history and survey of AI-generated content (AIGC), focusing on GenAI models such as GANs and VAEs, and discusses their potential applications in wireless communication networks. | ✗ | ✓ | ✓ |
| [21] | Surveys the use of LLMs in telecommunications, including network optimization, traffic management, and security, as well as the integration of GenAI in sixth generation mobile networks. | ✗ | ✓ | ✓ |
| [36] | Explores the integration of MoE and GenAI in the Internet of Vehicles (IoVs), focusing on traffic management, autonomous driving, and collaborative decision-making. | ✗ | ✓ | IoV |
| This survey | Focuses on the efficient training and deployment of MoE frameworks with advanced GenAI approaches for various wireless network scenarios and wireless technologies. | ✓ | ✓ | ✓ |

offered a review of AI-generated content (AIGC), exploring GenAI models and their potential applications in wireless communication networks, while the authors in [21] discussed the implementation of LLMs into telecommunications, focusing on network optimization, traffic management, and security in sixth generation (6G) mobile networks. Furthermore, the authors in [36] incorporated MoE and GenAI models in the Internet of Vehicles (IoV), addressing challenges in traffic management, autonomous driving, and collaborative decision-making for vehicular networks.

However, existing studies lack sufficient investigation into the integrated potential of MoE and advanced GenAI technologies within the broader context of wireless communication systems. This paper fills this gap by comprehensively exploring the diverse application scenarios of MoE in wireless networks, such as vehicular networks [84], [85], unmanned aerial vehicle (UAV) networks [86], [87], [88], satellite networks [89], [90], and other fundamental 6G scenarios [91], [92], [93], [94]. Through in-depth analysis, we demonstrate the capabilities of MoE in solving complex wireless tasks, such as channel estimation [95], [96], [97], physical layer signal processing [61], [98], [99], [100], radio resource allocation [101], [102], [103], traffic management [104], [105], [106], distributed edge computing [107], [108], [109], and network security [110], [111], [112]. The dynamic expert selection and sparse activation mechanisms of MoE have been developed to improve resource utilization, enhance computational efficiency, and ensure adaptability in heterogeneous

wireless environments. The main contributions are summarized as follows.

- We provide a foundational overview of MoE methodologies, covering classical MoE models as well as their integration with GenAI and reinforcement learning (RL), highlighting MoE's significant capabilities in addressing complex computational demands and optimizing resource utilization.
- We illustrate the advantages of applying MoE frameworks in heterogeneous and dynamic wireless systems by extensively examining diverse scenarios such as vehicular networks, UAV networks, satellite communications, and emerging 6G scenarios.
- We thoroughly investigate innovative applications of MoE in wireless technologies, including channel estimation, physical layer signal processing, radio resource allocation, traffic optimization, distributed computing, and network security, demonstrating MoE's versatility and effectiveness across a broad range of complex tasks.
- We present a detailed case study that demonstrates the performance advantages of MoE, along with a comprehensive review of publicly available datasets that are widely utilized in MoE-based models to facilitate further research and experimentation in wireless networking.

The rest of this paper is organized as follows. Section II provides an overview of MoE, including its components, gating mechanisms, and the integration with GenAI and RL. Section III presents a comprehensive investigation of MoE for

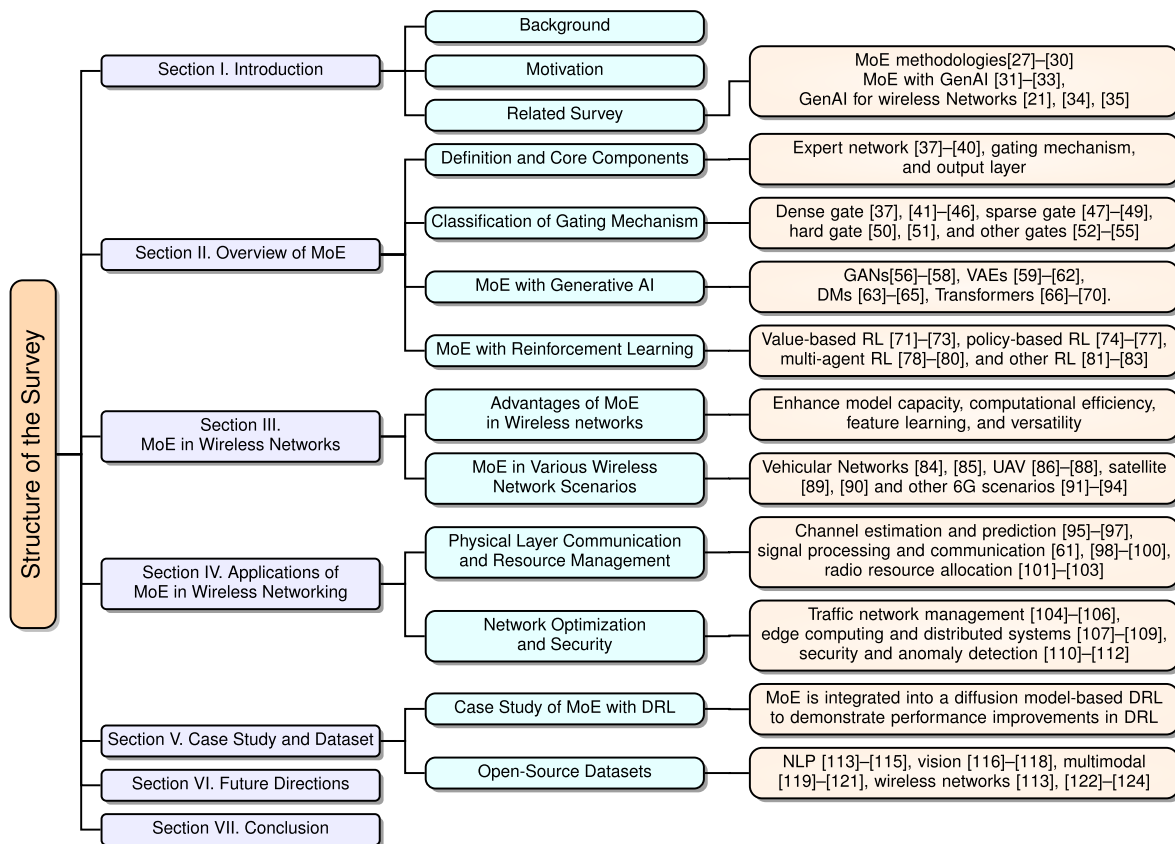


Fig. 1. The structure of this survey.

diverse wireless networks. Section IV focuses on the practical applications of MoE in wireless technologies. In Section V, we provide a case study on the integration of MoE with an RL task and summarize open-source datasets that support MoE research. Section VI discusses the future research directions. Finally, the paper is concluded in Section VII. The overall structure of this survey is illustrated in Figure 1.

II. OVERVIEW OF MIXTURE OF EXPERTS

Initially introduced in the early 1990s to overcome the limitations of single-architecture models [41], MoE has undergone substantial advancements in modularity and composition, allowing for integration with diverse expert networks, sophisticated gating mechanisms, and recent advanced GenAI models and RL methods. These developments have established MoE as a foundational technique across various domains, including natural language processing (NLP), computer vision, multimodal learning, and wireless networks. Figure 2 presents a comprehensive timeline that traces the evolution of MoE expert networks, gating strategies, and the combination with GenAI, alongside the growing breadth of applications. Building on this foundation, this section presents the definition and core components of MoE, provides an overview of gating mechanisms, and examines its integration with GenAI models and RL methods.

A. Definition and Core Components of MoE

MoE is a machine learning framework that incorporates multiple specialized models, referred to as experts, each

handling a subset of input data within the training process [29]. As illustrated in Figure 3 (a) and (b), compared with classical deep neural networks (DNNs) that directly process an entire dataset, the MoE framework leverages a gating network to allocate dynamically training samples to the most relevant experts [133]. The core components of MoE consist of the expert network, gating scheme, and output layer.

- **Expert Network:** MoE decomposes a model into smaller, more manageable subnetworks, enabling each expert to specialize in distinct data distributions or feature representations. The expert of MoE is typically a trainable network with variations in architecture, depth, or hyperparameters, ranging from conventional machine learning models to advanced DNNs. Examples of expert networks include support vector machines (SVMs) [37], multilayer perceptrons (MLPs) [134], self-attention heads [38], and hybrid structures that integrate feedforward networks (FNNs) with self-attention heads [39]. In more complex scenarios, entire language models can also serve as the expert network within the MoE framework [40].
- **Gating Mechanism:** A gating mechanism is responsible for selecting appropriate experts to deal with input samples. Typically implemented as a lightweight model, a gating network computes a probability distribution over the available experts, determining which experts should be activated for model training and inference, thereby reducing computational overhead. Through joint training with expert networks, the gating network continuously improves expert selection, enhancing task specialization and overall model efficiency. Various gating mechanisms

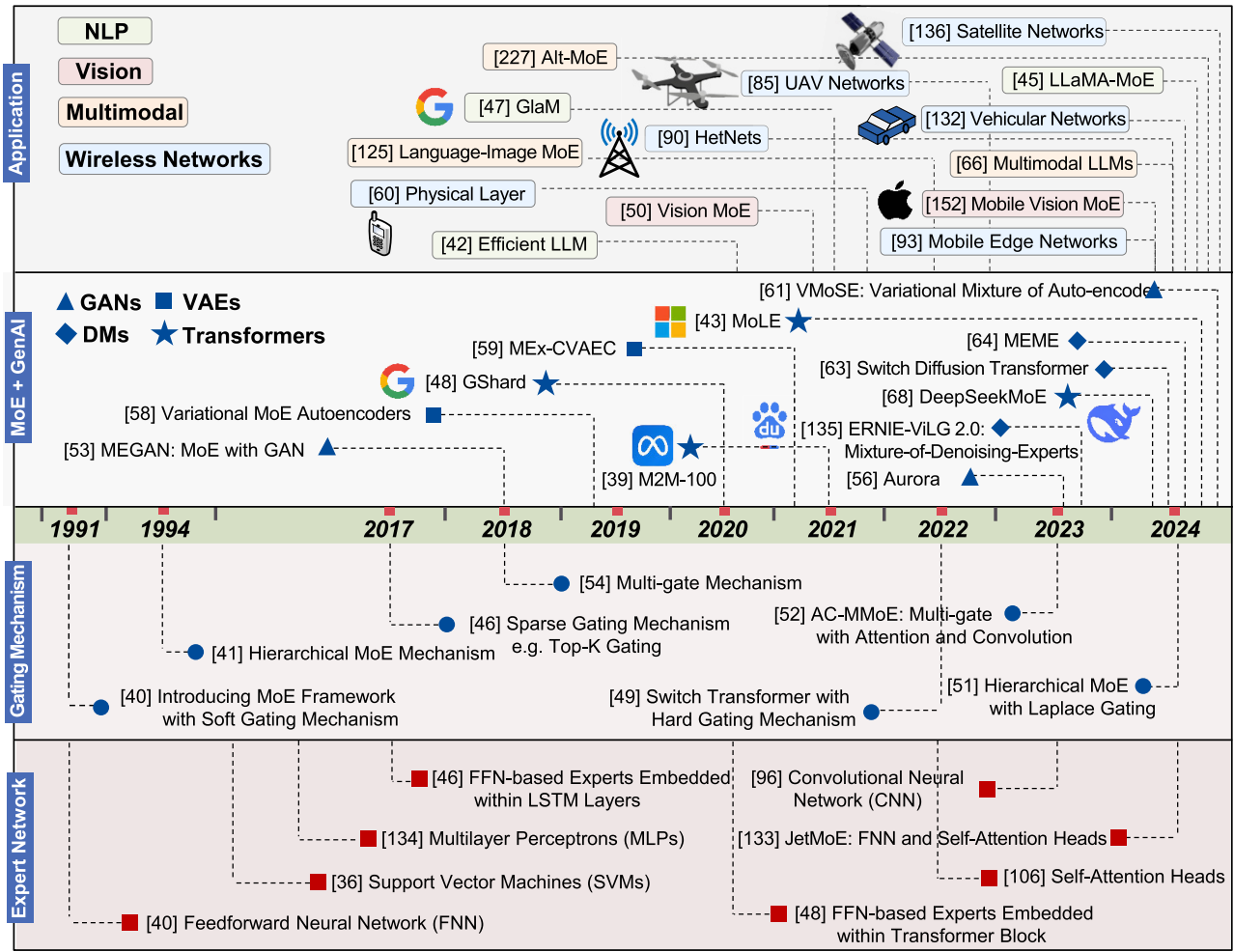


Fig. 2. A comprehensive timeline illustrating the evolution of MoE models, structured across four dimensions: expert networks, gating mechanisms, MoE integrated with GenAI, and MoE applications. The timeline presents diverse expert networks such as FNNs, SVMs, MLPs, CNNs, and self-attention heads, along with the development of dense, sparse, hard, and other variant gating mechanisms. The framework of MoE is widely integrated with GenAI models, such as GANs, VAEs, DMs, and Transformers. Recent applications span NLP, computer vision, multimodal tasks, and wireless networks, demonstrating the versatility of MoE in modern AI and communication systems.

have been proposed to optimize expert selection, including dense, sparse, and hard gating, as well as more advanced schemes such as hierarchical gating and multi-gate mechanisms [32].

- **Output Layer:** The output layer aggregates the predictions from activated experts and computes the final output as their combination. The most common approach is a weighted sum, where the weights are determined by the gating network to appropriately scale expert contributions based on their relevance to the input [135]. While this method facilitates smooth integration, more advanced implementations incorporate additional approaches to enhance the output quality. For instance, normalization techniques [136] and residual connections [137] are employed to regulate the scale of expert outputs, preventing any single expert from dominating the aggregation process.

Compared to single architectures designed for uniform data, the expert-structured MoE framework improves specialization and training efficiency for diverse data or tasks, particularly in domains requiring high model capacity and adaptability [138].

B. Classification of Gating Mechanism

The gating mechanism plays a pivotal role in MoE frameworks, serving as the decision maker that dynamically activates the most appropriate experts to process a specific input. We classify the gating mechanism into four types, including dense gating, which assigns weights to all experts; sparse gating, which selects only a subset of experts for enhancing efficiency; hard gating, deterministically allocating each input to one expert; and variant strategies, designed to enhance flexibility and specialization in expert selection.

1) *Dense Gating Mechanism:* The dense gating mechanism activates all expert networks by assigning probability weights at each iteration. These weights are typically computed using a softmax function, producing a probability distribution over all available experts. The final output is derived as a weighted sum of all expert outputs, ensuring that each expert contributes to the final prediction. Formally, for an MoE framework consisting of N experts $\{E_1, \dots, E_N\}$, the output of the dense MoE layer is given by

$$\mathcal{M}_{\text{dense}}(\mathbf{x}; \Theta, \{\theta_n\}_{n=1}^N) = \sum_{n=1}^N G(\mathbf{x}; \Theta)_n E_n(\mathbf{x}; \theta_n), \quad (1)$$

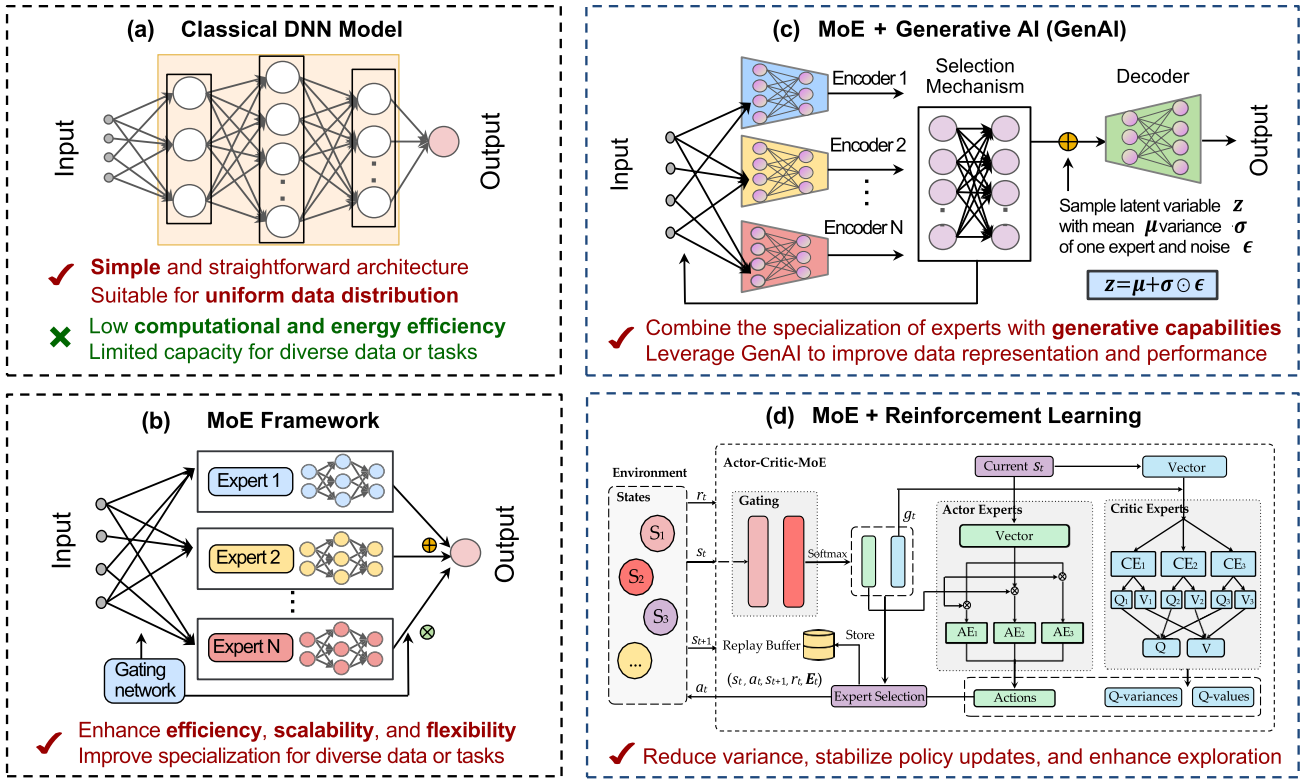


Fig. 3. Illustration of three learning architectures. (a) **Classical DNN model**: A straightforward architecture for uniform data distribution but limited by low computational and energy efficiency. (b) **MoE framework**: Incorporate multiple expert networks coordinated by a gating mechanism for diverse data or tasks. (c) **MoE with GenAI**: Combine the specialization of MoE with generative capability, illustrated by an example of MoE integrated with multiple encoder experts of variational autoencoder (VAE) [61]. (d) **MoE with RL**: Employ multiple actor experts (AEs) and critic experts (CEs) to enhance the training performance of the actor-critic (AC) algorithm [73].

where \mathbf{x} denotes the input data and $E_n(\mathbf{x}; \theta_n)$ is the output of the n -th expert parameterized by θ_n . $G(\mathbf{x}; \Theta)_n$ represents the gating weight for the n -th expert, with Θ denoted as the parameters of the gating network. If the softmax function is used to compute the gating weight, $G(\mathbf{x}; \Theta)_n$ is expressed as

$$G(\mathbf{x}; \Theta)_n \triangleq \text{Softmax}(G(\mathbf{x}; \Theta))_n = \frac{\exp(G(\mathbf{x}; \Theta)_n)}{\sum_{i=1}^N \exp(G(\mathbf{x}; \Theta)_i)}. \quad (2)$$

Dense gating has been adopted in multiple early investigations [37], [41], [42], and has more recently been applied to large AI models such as LoRAMoE [43], MoLE [44], and DS-MoE [45]. Since the dense gating mechanism activates all experts for each input instance, they have proved effective in tasks demanding comprehensive expert engagement, especially under high uncertainty or diverse feature distributions [46]. However, every expert receives gradients during back-propagation, causing the total computational cost and memory usage to scale linearly with the number of experts N . While this scheme simplifies expert selection and enhances model expressiveness, it usually incurs substantial inter-expert communication overhead and computational costs, making it impractical for large-scale applications. [139].

2) *Sparse Gating Mechanism*: Sparse gating mechanism uses the Top-K expert selection approach to limit the number of activated experts per input to $K \ll N$, which tremendously reduces the computational overhead [47]. The gating weight of the n -th expert $G(\mathbf{x}; \Theta)_n$ using the Top-K approach is

computed as

$$G(\mathbf{x}; \Theta)_n = \text{Softmax}(\text{Top-K}(G(\mathbf{x}; \Theta)))_n, \quad \text{Top-K}(G(\mathbf{x}; \Theta))_i = \begin{cases} G(\mathbf{x}; \Theta)_i, & G(\mathbf{x}; \Theta)_i \text{ is among the} \\ & \text{top-K values of } G(\mathbf{x}; \Theta), \\ -\infty, & \text{otherwise.} \end{cases} \quad (3)$$

Based on the sparse activation mechanism, only the selected experts participate in forward and backward passes, thereby reducing the training floating point operations (FLOPs) and parameter updates by an approximate factor of N/K . Examples of sparse gating implementations include GShard [49] and GLaM [48], which significantly improve efficiency by activating a small subset of experts. Empirical results of sparse gating from GShard demonstrate more than $2\times$ improvements in training throughput and substantial reductions in memory consumption compared to dense gating mechanism. However, these gains come at the cost of a slight reduction in model capacity and the introduction of a load-balancing loss, which is required to ensure uniform expert utilization in sparse setups.

3) *Hard Gating Mechanism*: Unlike soft and sparse gating, the hard gating mechanism deterministically selects a single expert per input based on the highest gating probability. The output weights of the hard gating mechanism are expressed as

$$G(\mathbf{x}; \Theta)_n = \begin{cases} 1, & n = \arg \max_i G(\mathbf{x}; \Theta)_i, \\ 0, & \text{otherwise.} \end{cases} \quad (4)$$

Consequently, the final output of MoE layer is provided by the output of the highest-weighted expert as

$$\begin{aligned} \mathcal{M}_{\text{hard}}(\mathbf{x}; \Theta, \{\theta_n\}_{n=1}^N) &= E_{i^*}(\mathbf{x}; \theta_{i^*}), \\ i^* &= \arg \max_i G(\mathbf{x}; \Theta)_i. \end{aligned} \quad (5)$$

The hard gating mechanism substantially reduces computational overhead, making it particularly suitable for resource-constrained environments. Despite its simplicity, the hard gating mechanism, as utilized in Switch Transformer [50] and DeepSpeed-MoE [140], has demonstrated competitive performance even when activating only one expert per sample. Nonetheless, the hard gating mechanism may be confronted with the challenge of imbalanced expert utilization, where certain experts may be over-utilized while others remain underused, potentially limiting the model's capacity utilization and expressive generalization.

4) *Other Variant Gating Mechanisms:* Beyond conventional gating mechanisms, several advanced and specialized gating strategies have been developed to further optimize expert selection. Multi-gate MoE (MMoE) frameworks employ multiple gating networks to handle different tasks, enabling task-specific routing that enhances model specialization [53], [55]. The expert networks of MMoE are shared across all tasks, while each task is assigned an independent gating network that separately determines the optimal weighted combination of experts. For an MMoE framework with T gates, the t -th output of the MoE layer is computed as

$$\mathcal{M}_{\text{multi-gate}}^t(\mathbf{x}; \Theta_t, \{\theta_n\}_{n=1}^N) = \sum_{n=1}^N G_t(\mathbf{x}; \Theta_t)_n E_n(\mathbf{x}; \theta_n). \quad (6)$$

This framework enables efficient parameter sharing among tasks while allowing task-specific specialization through dedicated gating mechanisms. Meanwhile, hierarchical gating, as a variant of the gating mechanism, introduces multi-level decision-making, where experts are partitioned into subgroups, and the gating functions are performed in multiple stages to refine expert selection [52], [54]. The output of a two-level hierarchical gating mechanism is expressed as

$$\begin{aligned} \mathcal{W}_{\text{hier}}(\mathbf{x}; \Theta^{\text{high}}, \{\Theta_m^{\text{low}}\}_{m=1}^M, \{\theta_{m,n}\}_{n=1}^{N_m}) \\ = \sum_{m=1}^M G^{\text{high}}(\mathbf{x}; \Theta^{\text{high}})_m \sum_{n=1}^{N_m} G_m^{\text{low}}(\mathbf{x}; \Theta_m^{\text{low}})_n E_{m,n}(\mathbf{x}; \theta_{m,n}), \end{aligned} \quad (7)$$

where the high-level gating $G^{\text{high}}(\mathbf{x}; \Theta^{\text{high}})$ first selects expert group m from M available groups for the input \mathbf{x} . Within the selected group m , the low-level gating $G_m^{\text{low}}(\mathbf{x}; \Theta_m^{\text{low}})$ further weights the expert output among N_m experts, with the n -th expert output denoted as $E_{m,n}(\mathbf{x}; \theta_{m,n})$. The hierarchical gating mechanism is well-suited for tasks requiring multi-level decision-making and is particularly effective in wireless network scenarios with strong task correlations.

Mitigation of Expert Collapse and Overfitting: Although gating mechanisms effectively regulate expert activation and

manage computational resource allocation, issues such as expert collapse and model overfitting persist and require further mitigation, particularly under limited training data. Expert collapse occurs when a small number of experts dominate training, leading to under-utilization of other experts, thereby diminishing model diversity and capacity. Several strategies have been proposed to address the challenge of expert collapse. Load-balancing losses, such as auxiliary entropy-based losses, encourage a balanced distribution of expert activations [51]. For example, a router Z-loss is introduced to enable a more uniform output proportion of the gating network [141]. Given B input data samples, the Z-loss is formulated as

$$L_Z = \frac{1}{B} \sum_{b=1}^B \left(\log \sum_{n=1}^N \exp(G(\mathbf{x}_b; \Theta)_n) \right)^2, \quad (8)$$

where \mathbf{x}_b is the b -th input data and the loss function helps the $G(\mathbf{x}_b; \Theta)$ values converge towards a more balanced distribution. The load-balancing strategies effectively mitigate expert collapse by penalizing excessive reliance on individual experts, thereby enhancing the stability of model training. Additionally, overfitting problem arises when experts overly adapt to limited training data, reducing their ability to generalize effectively. Regularization methods such as dropout, early stopping, and weight decay are commonly integrated into MoE frameworks to prevent overfitting [142]. More recently, knowledge distillation [143] and expert pruning techniques [144] have emerged as effective approaches to maintain expert diversity, especially under data-scarce conditions. These mechanisms collectively enhance the robustness of MoE, ensuring better generalization across diverse tasks and environments.

Lessons Learned: Gating mechanisms serve as a critical component of the MoE framework, responsible for regulating expert activation and managing the allocation of computational resources. These mechanisms differ significantly in their trade-offs among computational efficiency, representational capacity, and system complexity. Specifically, dense gating activates all experts for every input, ensuring full expert participation and rich expressiveness, but at the cost of substantial computational and memory overhead. Sparse gating, such as Top-k selection, reduces training FLOPs and memory consumption by activating only a subset of experts per input, which is widely adopted in large-scale models such as GShard and GLaM. However, it usually incurs a slight reduction in model capacity and requires effective load-balancing strategies to mitigate expert underutilization. Hard gating deterministically selects a single expert for each input, achieving maximal computational efficiency and making it suitable for resource-constrained scenarios. Nonetheless, hard gating introduces high gradient variance and potential expert imbalance that impairs the stability of the training process. Advanced gating strategies, such as multi-gate and hierarchical gating, support task-specific or multi-level expert selection, enabling finer specialization and better adaptation to heterogeneous inputs, but also increase architectural and training complexity. Overall, each gating strategy offers distinct advantages depending on the target deployment scenarios and resource constraints. As such, the design and deployment of gating mechanisms

TABLE II

SUMMARY OF GENERATIVE AI MODEL WITH MIXTURE OF EXPERTS INTEGRATION. **NOTE:** RED CIRCLES INDICATE THE LIMITATIONS OF CONVENTIONAL GenAI AND GREEN CIRCLES HIGHLIGHT THE ADVANTAGES OF MoE

| Model | Ref. | Task | Insight | GenAI Limits & MoE Advantages |
|------------------|------|----------------------------|--|---|
| GANs | [57] | Language Generation | Improve representation and coherence via expert generators and feature alignment | <ul style="list-style-type: none"> ● Mode collapse, limited sample diversity and difficulties in multimodal distributions ● Multi GAN generators improve generation diversity and captures distinct data modes ● Generator selection enhances semantic alignment and generation efficiency |
| | [57] | Image Generation | Use gating network to dynamically select specialized generators | |
| | [58] | Text-to-image Generation | Apply sparse gating and FNN experts for semantically aligned feature synthesis | |
| VAEs | [60] | Anomaly Detection | Use convolutional VAEs for manifold-specific modeling | <ul style="list-style-type: none"> ● Struggle with complex representation in the latent space ● Enable expert specialization through multiple encoding and decoding modules ● Enhance model expressiveness by mitigating latent discontinuities |
| | [61] | Channel Coding | Multi encoders to handle representation discontinuities under bandwidth limits | |
| | [62] | Multi-modal Recommendation | Fuse modality-specific experts based on uncertainty-aware stochastic sampling | |
| Diffusion Models | [63] | Text-to-image Generation | Train specialized denoisers for different denoising intervals to align text and image fidelity | <ul style="list-style-type: none"> ● Fixed parameters across denoising steps hinder adaptability ● Employ specialized experts for different denoising stages, adapting to distinct noise characteristics at each time step ● Sparse experts reduce inference cost while maintaining generation quality |
| | [64] | Image Generation | Assign timestep intervals to pruned experts for reducing inference overhead | |
| | [65] | Image Generation | Employ timestep-aware MoE for adaptive routing and improved convergence | |
| Transformers | [68] | Language Generation | Construct attention and MLP-based MoE for instructed LLMs with post-training | <ul style="list-style-type: none"> ● Substantial computational requirements and limited adaptability to diverse modalities ● Replace FFNs with MoE layers to reduce computation and maintain model capacity ● Enable experts to handle inputs across language, vision, and multimodal domain |
| | [69] | Language Generation | Employ expert segmentation and shared experts to maximize specialization | |
| | [70] | Multimodal Modeling | Combine modality-specific MoE in vision and language tasks | |

are pivotal for scaling MoE frameworks to meet the diverse requirements of modern machine learning applications, particularly in dynamic and resource-sensitive environments such as wireless networks.

C. MoE With Generative AI

The employment of MoE framework in GenAI models has garnered increasing attention, leveraging MoE experts with generative capabilities to enhance data synthesis, representation learning, and adaptive processing. Recent advancements have demonstrated the effectiveness of MoE in improving the performance of GenAI models, including generative adversarial networks (GANs), variational autoencoders (VAEs), diffusion models (DMs), and Transformer-based architectures [79]. This section explores the integration of MoE and GenAI, highlighting key advancements in optimizing generative performance across diverse applications. Figure 3(c) illustrates an MoE-enhanced GenAI framework based on VAE, and a summary of related approaches is provided in Table II.

1) *Generative Adversarial Networks:* GANs are widely used for high-quality image synthesis, consisting of a generator that synthesizes realistic data samples and a discriminator that differentiates between real and synthetic inputs [145]. Despite their effectiveness in generation capability, conventional GANs often suffer from mode collapse, limited sample diversity, and difficulties in modeling complex multimodal distributions [146]. The integration of MoE into GANs provides a scalable solution by incorporating multiple specialized generators or expert modules within the generator [147]. The

gating network in MoE dynamically selects the most relevant expert, improving mode diversity, mitigating mode collapse, and enhancing generalization across heterogeneous data distributions. Recent advancements have explored the integration of MoE framework into GANs to enhance generative modeling across different domains. In the context of language generation, MoE-GAN [56] incorporates multiple expert generators for language feature statistics alignment, which offers fine-grained learning fidelity and improves the diversity and coherence of generated text. Additionally, MEGAN [57] applies MoE to image generation by employing a set of generators and a gating network adaptively selecting suitable generators for each input, thereby capturing distinct data modalities. Extending this concept to text-to-image synthesis, a sparse MoE-based GAN framework is introduced to leverage both textual embeddings and latent variables for specialized feature generation, thus providing more specialized and semantically aligned images [58].

2) *Variational Autoencoders:* VAEs are probabilistic generative models that encode input data into a latent distribution and reconstruct data from a sampled latent variables [59]. While the encoding-decoding structure of VAEs is effective in data compression and generation, conventional VAEs often struggle with capturing complex representation in the latent space. Exploiting MoE in VAEs enables specialized latent space modeling by dynamically allocating data samples to different encoder or decoder experts, improving reconstruction capability and representation diversity. For instance, in anomaly detection, employing a set of convolutional VAEs as experts, together with a convolutional gating network to

select experts for modeling manifold-specific patterns, has been shown to improve detection accuracy [60]. In communication scenarios, an introduction of multiple VAE encoders to mitigate representation discontinuities under bandwidth constraints has led to improved robustness across varying channel conditions [61]. Furthermore, in multi-modal recommendation, a mixture of variational stochastic auto-encoders [62] is leveraged to select and fuse modality-specific latent experts, which effectively enhances generalization for different modalities under incomplete input conditions.

3) *Diffusion Models*: DMs comprise a forward process, where random noise is gradually added to the original data, and a reverse process that reconstructs structured data through a multi-step denoising process [148]. Conventional DMs use a single set of parameters across all denoising steps, limiting their adaptability to different noise levels [149]. MoE-based diffusion models overcome this constraint by employing specialized experts for different denoising stages based on distinct noise characteristics [150]. Additionally, MoE enhances computational efficiency through sparse expert activation, reducing the need for full-network inference at every denoising step [151]. Recent developments have explored the integration of MoE into diffusion models to improve both generation quality and computational efficiency. For instance, eDiff-I, a text-to-image diffusion model proposed by NVIDIA, has introduced an ensemble of expert denoisers for different generation stages, thereby improving text alignment and image fidelity in text-to-image synthesis [63]. Meanwhile, DiffPruning [64] reduces the inference cost of image generation by assigning clustered timestep intervals to pruned expert networks, while Switch-DiT [65] integrates timestep-specific characteristics within sparse MoE layer to accelerate convergence.

4) *Transformers*: Transformers have emerged as the prevalent architecture in GenAI models, demonstrating remarkable capabilities in sequential modeling through self-attention mechanisms and encoder-decoder architectures [152]. However, as the model sizes continue to grow, the computational demands of Transformers increase substantially, posing challenges for efficient training and deployment, particularly in LLM applications [153]. A widely adopted solution is to replace the standard feedforward networks (FFNs) in Transformer blocks with MoE layers, where a gating mechanism dynamically selects a sparse subset of smaller FNN for each input token, thereby reducing computation while maintaining model capacity [66]. Moreover, MoE also facilitates expert specialization by enabling different experts to handle distinct input patterns or tasks, enhancing performance across language, vision, and multimodal domains [67]. The effectiveness of MoE-based Transformers has been validated in recent models such as LLaMA-MoE [68] for scalable language modeling, DeepSeek-MoE [69] for cross-layer expert allocation, and MoME [70] for flexible multimodal integration.

Lessons Learned: The integration of MoE into GenAI models has demonstrated significant improvements in computational efficiency, representation learning, and adaptive specialization throughout the generative process. A key insight is that generative models often involve highly

structured procedures, such as multi-step denoising [63], [64], [65], sequential decision-making [154], [155], or modality-conditioned synthesis [58], [67]. In these settings, MoE allows GenAI models to activate only the most relevant experts at each stage, enabling fine-grained refinement of the generation process and adaptive utilization of model capacity. The resulting specialization improves fidelity and stability without increasing computational complexity, particularly in models designed to handle heterogeneous data types. Furthermore, with the emergence of new generative paradigms such as Generative Flow Networks (GFlowNets) [156], MoE provides a scalable and modular framework that supports expert specialization, enhances generation quality, and addresses the increasing demands of advanced GenAI models.

D. MoE With Reinforcement Learning

An integration of MoE with RL has demonstrated significant potential in enhancing value function estimation, policy learning, and decision-making efficiency. Figure 3(d) illustrates the integration of multiple actor and critic experts within a deep RL (DRL) framework. The following explores the MoE framework across various RL paradigms, including value-based, policy-based, and multi-agent RL, as well as its applications in other RL approaches. A summary of RL methods with MoE integration is presented in Table III.

1) *MoE for Value-Based RL*: In state or state-action value-based RL methods, such as Temporal Difference learning [157], Q-learning [158], and Deep Q-Networks (DQN) [159], as well as the critic component in actor-critic (AC) frameworks including Deep Deterministic Policy Gradient (DDPG) [160] and Twin Delayed DDPG (TD3) [161], accurate value estimation is essential for effective policy optimization. However, these conventional methods face several challenges, such as value overestimation bias, limited generalization across dynamic environments, and difficulties in adapting to non-stationary conditions. MoE has been introduced as an advanced value function network, leveraging multiple specialized expert networks to approximate the value function under diverse conditions [71]. By incorporating a gating network, MoE adaptively selects the most suitable expert, thereby enhancing estimation accuracy, mitigating overestimation bias, and improving learning stability [72]. For instance, TD3 reduces the overestimation bias of DDPG by maintaining two Q-value estimators and selecting the minimum value during policy updates. MoE extends this approach by introducing multiple value function experts, allowing the model to dynamically adjust to different environmental conditions and state distributions [73]. This multi-critic networks enabled by MoE not only improves robustness but also facilitates better decision-making in complex reinforcement learning settings.

2) *MoE for Policy-Based RL*: For policy-based RL methods and the actor component in AC frameworks, such as Reinforce [162], Proximal Policy Optimization (PPO) [163], and Soft Actor-Critic (SAC) [164], policy networks effectively learn decision-making strategies that adapt to diverse environmental states. However, these methods face several challenges, including inefficient policy exploration, high variance in policy gradients, and limited adaptability in optimizing policies

TABLE III

SUMMARY OF REINFORCEMENT LEARNING METHODS WITH MIXTURE OF EXPERTS INTEGRATION. **NOTE:** RED CIRCLES INDICATE THE LIMITATIONS OF CONVENTIONAL RL AND GREEN CIRCLES HIGHLIGHT THE ADVANTAGES OF MoE

| Methodology | Ref. | Algorithm | Insight | RL Limits & MoE Advantages |
|-------------------|------|-----------------------|---|---|
| Value-based RL | [71] | DQN | Improve Q-value estimation via expert selection in multiple Q-networks | <ul style="list-style-type: none"> ● Overestimation bias, limited generalization, non-stationarity issues ● Employ multi-value networks for more accurate estimation and stable training |
| | [72] | Double DQN (DDQN) | Enhance dynamic decision-making with weighted DDQN experts | |
| Policy-based RL | [74] | AC | Adaptive actor and critic network selection in dynamic environments | <ul style="list-style-type: none"> ● Inefficient policy exploration, high policy gradient variance, and limited adaptability in complex policy scope ● MoE is used for multiple actors, while critic can be either single or multiple ● Reduce variance in policy gradients, stabilize policy updates, and improve policy exploration efficiency |
| | [75] | PPO | Enhance PPO with MoE-based actors for demand-aware capacity planning | |
| | [76] | DDPG | Mixture Gaussian actors while retaining a standard single critic | |
| | [77] | SAC | Employ probabilistic MoE in Actor while retaining a standard SAC critic | |
| Multi-Agent RL | [78] | Multi-agent AC (MAAC) | Use an MoE framework to dynamically optimize each agent's policy | <ul style="list-style-type: none"> ● Non-stationary agent behaviors, limited scalability, and inefficient coordination ● MoE acts as a unified multi-agent structure or be used within each agent ● Mitigate non-stationarity, achieve high adaptability, and improve coordination |
| | [79] | MAPPO | Dynamic policy network selection for multi-agent coordination | |
| | [80] | MAPPO | Integrate mixture of policy networks within each agent | |
| Other RL Approach | [81] | HRL | Employ a high-level gating network to activate low-level experts for sub-tasks such as pushing, grasping, and inserting | <ul style="list-style-type: none"> ● Inefficient hierarchical policy exploration ● Improve multi-tier coordination and decision-making ● Limited adaptability to novel distributions ● Improve generalization across environments ● Task interference and negative transfer ● Enhance specialization for distinct tasks |
| | [82] | Meta-RL | Mixture of representative experts to enhance generalization for unseen tasks | |
| | [83] | Multi-task RL | Apply attention-based experts for adaptive policy learning across diverse tasks | |

under complex environmental or reward landscapes [165]. To address these limitations, MoE provides an adaptive policy representation structure, where multiple expert networks learn distinct strategies for different environmental conditions [166]. Through gating mechanisms, MoE selects the most appropriate expert for distinct state distributions, effectively reducing variance in policy gradients and improving stability during policy updates. Moreover, by distributing policy learning across multiple experts, MoE encourages diverse policy exploration, enabling the discovery of more effective strategies and preventing premature convergence to suboptimal policies. MoE has been integrated into the actor-network of PPO [75], DDPG [76], and SAC [77], where a probabilistic gating mechanism dynamically assigns different environmental conditions or state distributions to specific actor experts. This enables some actor networks to focus on high-risk exploratory strategies while others refine exploitation-based policies, thereby reducing gradient variance, stabilizing policy updates, and improving exploration efficiency.

3) *MoE for Multi-Agent RL:* MoE significantly enhances multi-agent RL by improving coordination strategies, scalability, and adaptability in systems involving multiple interacting agents [78]. One of the primary challenges in multi-agent RL is the non-stationary nature of the joint action decisions, where dynamically changing agent behaviors influence the learning process [167]. MoE mitigates this issue by allowing all agents to share the same set of policy networks while selecting specialized experts for different interaction patterns [168]. For example, in the proposed MAPPO-MoE algorithm [79], rather than relying on a single agent policy, each agent utilizes a

set of different agent policy networks, where a probabilistic gating mechanism dynamically selects the most suitable experts based on the current state and interaction context. This approach employs centralized training and decentralized execution, facilitating adaptive decision-making for diverse agent behaviors while enhancing exploration efficiency by promoting diverse strategy selection. Furthermore, MoE enhances multi-agent RL scalability by integrating multiple expert networks within each agent, enabling robust coordination and adaptive policies for the decision-making of multi-agent systems [80].

4) *MoE for Other RL Approaches:* MoE can be further integrated into various advanced RL paradigms, such as hierarchical RL (HRL) [81], Meta-RL [91], and Multi-task RL [169]. In HRL, MoE facilitates hierarchical decision-making by selecting sub-policies operating at different decision levels, where higher-level experts focus on strategic planning and long-term decisions, while lower-level experts handle fine-grained control and real-time adaptation [170], [171]. In Meta-RL, MoE enables efficient knowledge transfer to new conditions through a set of pre-trained experts, expediting learning in non-stationary environments. By dynamically selecting experts with relevant prior knowledge, MoE enhances sample efficiency and accelerates policy adaptation, making it suitable for applications where rapid learning is essential for the tasks [82]. Additionally, in multi-task RL, the MoE framework allows for efficient specialization across different tasks by allocating task-specific expert networks to different distinct domains, ensuring that each expert optimizes a specific objective [83]. While leveraging shared representations for all tasks, MoE-based multi-task RL not only mitigates

task interference but also facilitates stable learning across potentially conflicting task distributions.

Lessons Learned: Integrating MoE into RL methods has led to significant improvements, including reducing overestimation bias in value-based RL [71], [72], [73], high variance in policy gradients [74], [75], [76], [77], and non-stationarity in multi-agent systems [78], [79], [80]. Additionally, MoE facilitates scalability in hierarchical [81] and multi-task [83] RL, enabling structured decision-making and shared generalization across related tasks [82]. The capability of MoE combined with different RL paradigms can be further extended to address the challenges in heterogeneous wireless environments. For example, diverse wireless technologies such as cellular [91] and UAV access [172] can benefit from robust policy adaptation and scalable decision-making with MoE-based RL approaches. However, while MoE offers substantial benefits, it introduces challenges such as increased computational overhead, expert collapse where a few experts dominate learning while others remain underutilized, and the difficulty of designing adaptive gating mechanisms. Addressing these issues necessitates further investigations into efficient training schemes, adaptive expert allocation strategies, and scalable architectures that preserve diversity and stability across varied RL learning tasks.

III. SCENARIOS OF MIXTURE OF EXPERTS IN WIRELESS NETWORKS

With the ability to enhance model capacity and computational efficiency, MoE has been widely adopted in wireless networks. This section highlights the advantages of MoE in wireless networks and explores its applications across various communication scenarios.

A. Advantages of MoE in Wireless Networks

In wireless communications, the ongoing evolution of wireless network technologies has brought about significant improvements in network capacity and data transmission rates [173]. However, with the rapid proliferation of wireless devices and the expanding deployment of wireless infrastructures, wireless networks are becoming increasingly heterogeneous and complicated, posing substantial challenges for efficient resource allocation and network management [174]. Additionally, network components such as edge servers and APs are typically resource-constrained, facing critical limitations to support a wide range of wireless applications [175]. The adoption of MoE frameworks in wireless communications provides substantial advantages.

- **Increases model capacity:** Traditional machine learning models struggle to efficiently capture the diverse characteristics of wireless communication environments. MoE extends model capacity by introducing multiple expert modules that specialize in different aspects of wireless communication, such as adaptive modulation, interference cancellation, and predictive channel estimation. This modular approach allows for a more effective representation of complex wireless features [70].
- **Enhances computational efficiency:** Wireless networks usually operate under stringent resource constraints, such

as limited bandwidth, power, and storage capacity. The expert selection mechanism in MoE ensures that only a subset of expert networks is engaged for each task, beneficial for resource-constrained environments such as mobile edge networks and IoT deployments, where lightweight yet high-performance models are essential for real-time decision-making and low-latency processing [44].

- **Enables dynamic feature learning:** Wireless networks exhibit high dynamism, characterized by continuously changing channel conditions, user mobility, and interference patterns. MoE employs multiple experts that can adapt to varying network conditions while maintaining high representation capability [176]. The adaptability of MoE enables dynamic feature learning in applications such as spectrum allocation, beamforming optimization, and interference management, which require real-time responsiveness to fluctuations in network conditions.
- **Versatility in multi-task and multimodal specialization:** MoE experts are specialized to address specific wireless communication tasks or adapt to varying operating conditions [177]. For example, specific expert networks are tailored for high- or low-SNR regimes in adaptive modulation classification (AMC), thereby optimizing their performance for each sub-task and significantly improving accuracy over general-purpose models [178]. Additionally, experts are designed to handle different input modalities, such as radar, LiDAR, or visual data [179]. Each expert focuses exclusively on its designated modality, achieving superior performance compared to models that process mixed-modal data without differentiation. Furthermore, MoE experts can be specialized based on general wireless network characteristics, such as network density, topology, propagation conditions, and traffic patterns [88], which enables the MoE framework to adapt to heterogeneous network environments and facilitate more efficient context-aware decision-making [180].

B. MoE in Wireless Network Scenario

As illustrated in Figure 4, MoE has been integrated into various wireless communication scenarios, including vehicular networks, UAV networks, satellite networks, heterogeneous networks (HetNets), integrated sensing and communications (ISAC), and mobile edge networks. The following section explores the application of MoE in these wireless scenarios, with a summary of the integration provided in Table IV.

1) *Vehicular Networks:* Vehicular networks have become a crucial component of modern transportation systems, enabling real-time data exchange between vehicles and roadside units (RSUs) to enhance traffic efficiency and safety [181], [182]. However, the highly dynamic nature of vehicular environments, characterized by rapid topology changes, heterogeneous driver behaviors, and strict latency requirements, poses significant challenges for intelligent decision-making systems [183], [184]. MoE has emerged as an effective approach to address these challenges by dynamically selecting specialized expert networks tailored to different driving scenarios. In the urban road scenario, the authors in [84] leverage

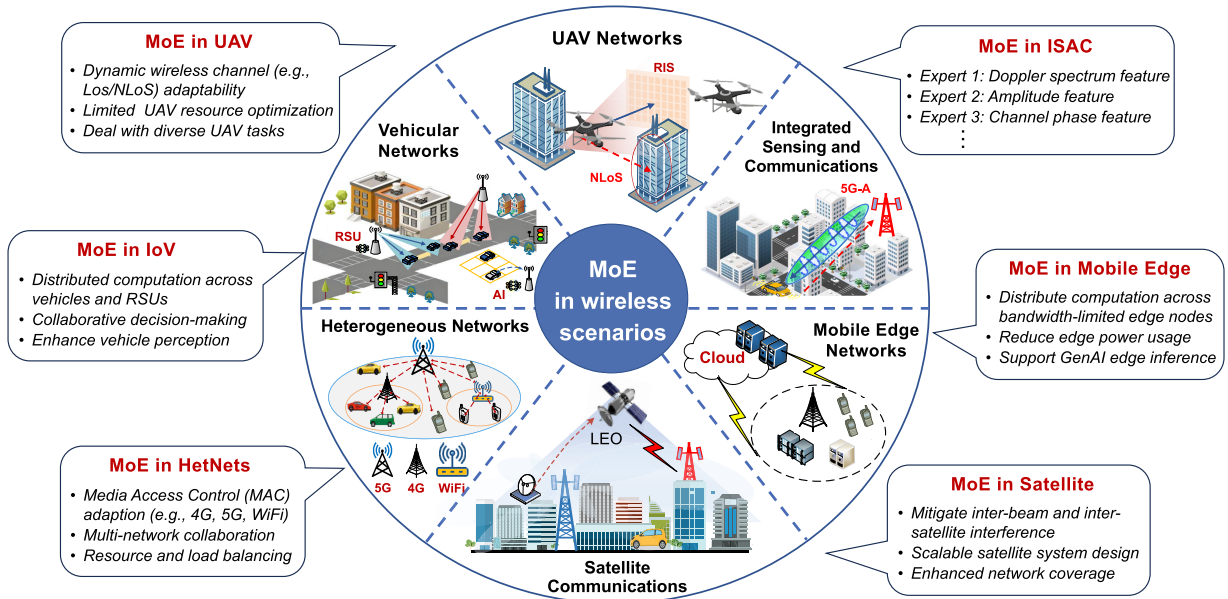


Fig. 4. Illustration of MoE integration across diverse wireless network scenarios, including vehicular networks, UAV networks, heterogeneous networks (HetNets), satellite networks, integrated sensing and communications (ISAC), and mobile edge networks.

a multi-gate MoE (MMoE) to process vehicle telematics data with RSU-vehicle collaboration, allowing for more accurate driver behavior prediction in resource-constrained vehicles. The proposed model achieves an accuracy of 98%, which outperforms traditional centralized models without MoE by 4%. In addition, an MoE-enhanced SAC framework incorporates heuristic experts and learning-based experts [85]. Through dynamically activating these two models, MoE can effectively mitigate collision risks and ensure smooth lane transitions, leading to a 13.75% improvement in average driving speed compared to a conventional DRL method with continuous action spaces.

2) *UAV Networks:* UAVs are increasingly employed across a wide range of applications, including aerial surveillance, disaster response, and communication relay [185]. Nonetheless, UAV networks may face significant challenges, such as real-time state estimation, limited computational resources, and adaptive trajectory optimization in uncertain environments [186], [187]. MoE enables UAV networks to improve real-time control, trajectory planning, and task execution by employing expert models adapted to specific flight conditions [172], [188]. For example, To improve UAV position estimations using a single Kalman filter (KF), the authors in [86] leverage multiple KF experts based on real-time flight conditions, ensuring resilience against sensor noise and enabling UAVs to achieve a 20% reduction in localization errors. Additionally, given the high computational cost of large-scale AI models, MoE-based model partitioning has been proposed to allow UAVs to load only the most relevant experts dynamically, reducing GPU memory usage by 86.4% while maintaining 98.2% UAV-onboard inference accuracy [87]. MoE can be further employed to effectively adapt to complex aerial environments such as line-of-sight (LoS) and non-line-of-sight (NLoS) aerial-ground radio propagation. For instance, a reconfigurable intelligent surface (RIS)-aided UAV network

is introduced in [88], where MoE's expert selection mechanism enables fast adaptation to diverse communication conditions, leading to a 17% improvement in UAV trajectory utility evaluated by the cumulative quality of wireless links along the UAV path.

3) *Satellite Networks:* Satellite networks serve as a fundamental component of global communication systems, particularly for providing connectivity in remote and under-served areas [189]. However, managing satellite communication resources presents significant challenges due to the heterogeneity of satellite constellations, the dynamic nature of orbital movement, and the need for adaptive power allocation and beamforming [190], [191]. To address these challenges, the authors in [89] enhance transmission efficiency by selecting specialized MoE experts for multi-task processing in Low Earth Orbit (LEO) and Geostationary Earth Orbit (GEO) satellite networks, including power control, beamforming, and interference mitigation. Through integrating MoE-based PPO frameworks, satellite networks achieve a 23.4% increase in throughput while reducing transmission interference by 17.8% compared to traditional RL-based approaches. Besides, MoE has proven highly effective in onboard satellite data processing. Traditional processing methods require extensive off-line computational resources, whereas the authors in [90] leverage MoE-optimized neural networks to enable near-real-time satellite data processing. The integration of MoE with significantly reduce latency while maintaining high accuracy and minimal memory usage, offering a promising way for the deployment of next-generation spaceborne AI applications.

4) *Heterogeneous Networks:* HetNets are characterized by the coexistence of multiple communication technologies, such as 5G, WiFi, and IoT networks [192], [193], [194]. Efficient allocation of radio resources and dynamic network access control are critical in these networks due to interference management challenges and network congestion [195], [196].

TABLE IV

SUMMARY OF WIRELESS NETWORK SCENARIOS WITH MIXTURE OF EXPERTS INTEGRATION. **NOTE:** LIGHT BLUE CIRCLES INDICATE THE MoE METHOD, GREEN CHECKMARKS AND RED CROSSES REPRESENT THE ADVANTAGES AND CHALLENGES OF MoE

| Scenario | Ref. | Network Component | Optimization Variable | MoE Advantages & Challenges |
|----------------------|------|--|--|--|
| Vehicular Networks | [84] | <ul style="list-style-type: none"> Connected vehicles Roadside units Cloud server | Vehicle speed, location, acceleration, and transmission overhead | <ul style="list-style-type: none"> Employ MMoE to learn distinct vehicle behaviors Achieves 98% RSU-aided behavior prediction (+4pp) Limited computation and storage resource in vehicle |
| | [85] | <ul style="list-style-type: none"> Connected vehicles Vehicle controller V2I communication | Steering angle, acceleration, and lane changing decision | <ul style="list-style-type: none"> Different experts for speed control and lane selection Improves average speed by 13.75% with zero collision MoE switching requires reliable real-time vehicle data |
| UAV Networks | [86] | <ul style="list-style-type: none"> UAVs UAV onboard sensor Kalman filters | UAV position, angular velocity, flying velocity, and attitude | <ul style="list-style-type: none"> Hierarchical MoE for multi UAV states modeling Achieve a 20% reduction in trajectory localization errors Complex model switching under dynamic UAV conditions |
| | [87] | <ul style="list-style-type: none"> UAVs Edge servers Ground base station | UAV-Edge association, expert model selection, UAV memory capacity, and download delay | <ul style="list-style-type: none"> GNN-based MoE to learn UAV-server selection strategy Reduce 86.4% UAV-onboard GPU memory usage and achieve 98.2% UAV-onboard inference accuracy Communication latency and storage limitation in UAV |
| | [88] | <ul style="list-style-type: none"> UAVs RIS Ground UEs Eavesdropper | UAV and UE trajectories, UAV active beamforming, RIS passive beamforming | <ul style="list-style-type: none"> Employ MoE for multi UAV-RIS tasks Improve 17% UAV trajectory utility evaluated by the cumulative quality of wireless links along the UAV path Uncertainty of UAV, UE, and eavesdropper trajectories |
| Satellite Networks | [89] | <ul style="list-style-type: none"> LEO satellite GEO satellite Ground UEs Ground station | Spectrum and channel resource, private and common transit beamforming, energy efficiency, message rate | <ul style="list-style-type: none"> MoE-PPO for spectrum, channel, beamforming tasks Optimize satellite communication throughput by 23.4% and reduce transmission delay by 17.8% Continuously evolving and highly dynamic satellite network properties |
| HetNets | [91] | <ul style="list-style-type: none"> Access points WiFi systems Mobile devices | Multi-access, channel collision probability, network throughput, and fairness | <ul style="list-style-type: none"> Meta-RL based MoE for multi-access control modeling Improve network spectrum efficiency by 21.5% Real-time adjustment in accessing heterogeneous 5G, WiFi, and IoT networks |
| ISAC | [92] | <ul style="list-style-type: none"> IoT Devices Sensing modules Datasets | CSI amplitude, phase, Doppler, and target detection accuracy | <ul style="list-style-type: none"> MoE experts to characterize different CSI feature Achieve 18% improvement in multi-target detection accuracy Simultaneously perform sensing and communication tasks |
| Mobile Edge Networks | [94] | <ul style="list-style-type: none"> Edge servers Base stations Mobile devices | Task offloading strategy, data transfer and computation time | <ul style="list-style-type: none"> Employ experts to match task type and server capability Reduce test loss in generalization errors by 15.4% Varied MEC features of server availability and capability |

MoE has been successfully applied to multiple access control (MAC) protocols to optimize spectrum access decisions dynamically across diverse environments. For example, the authors in [91] leverage Meta-RL to adjust MAC policies based on real-time network conditions. An MoE-enhanced encoder architecture is established to allow fine-grained access task representation, leading to faster convergence and improved generalization across varying spectrum scenarios. Simulation results demonstrate that MoE-based MAC improves spectrum efficiency by up to 21.5% compared to DRL approaches without MoE. MoE-driven HetNets achieve robust spectrum sharing, and higher network throughput, providing a promising solution for next-generation heterogeneous wireless access technology.

5) *Integrated Sensing and Communications*: ISAC has emerged as one of the fundamental technologies in 6G wireless networks, where sensing and communication functionalities are integrated to optimize network intelligence and environmental awareness [197]. MoE enhances ISAC performance by activating distributed ISAC experts for parallel execution of multiple sensing and communication tasks. For instance, the authors in [92] leverage different MoE experts for beamforming, spectrum allocation, and objective tracking tasks, achieving 18% improvement in multi-target detection accuracy while enhancing the efficiency of spectrum allocation

and signal processing. In addition, by leveraging MoE-based multi-modal learning, such as characterizing channels with Doppler effect, amplitude, and phase features, ISAC systems can efficiently fuse diverse data sources and ensure real-time adaptability to changing network conditions. This capability is especially advantageous in environments with rapidly fluctuating network conditions, such as smart cities and industrial automation.

6) *Mobile Edge Networks*: Mobile edge computing (MEC) plays a critical role in low-latency and high-efficiency wireless communications, where computational resources are deployed closer to end-users to reduce the reliance on centralized cloud computing [198], [199]. MoE improves MEC efficiency by dynamically offloading computational workloads via base stations (BSs), optimizing both inference and training tasks among mobile devices and edge servers [200], [201]. Through an adaptive gating network, MoE can select edge experts for each incoming task, ensuring efficient task routing and load balancing [94]. The employment of MoE in mobile edge networks mitigates the limitations of traditional task offloading strategies, which often lead to high generalization errors and inefficient resource utilization, reducing test loss by up to 15.4% compared to conventional models. Furthermore, MoE improves real-time decision-making by enabling edge devices to dynamically adjust model complexity based on

energy constraints and communication bandwidth, striking an optimal balance between computational efficiency and task performance [107].

C. Deployment Challenges and Limitations in Wireless Communications

1) *Resource and Deployment Constraints:* While MoE architectures offer promising scalability and efficiency across diverse wireless scenarios, their practical deployment in wireless communication systems remains challenging. Real-world wireless environments, such as vehicular networks, UAV-assisted aerial communication, and LEO satellite systems, impose stringent constraints on computational and wireless resources. These constraints limit the applicability of monolithic deep models and pose unique challenges to deploying MoE in a distributed wireless setting.

- **Computational Limitation and Heterogeneity:** In wireless edge scenarios, devices such as RSUs, UAVs, and LEO satellites are typically equipped with limited onboard processing capabilities, memory storage, and energy consumption. These limitations fundamentally constrain the feasibility of deploying large-scale MoE models in real-world wireless environments. Unlike cloud servers with sufficient computational resources to support intensive parallel processing, edge devices often struggle to support even moderately sized inference workloads. Moreover, the computational heterogeneity across different edge nodes, ranging from low-power microcontrollers to advanced AI accelerators, further complicates consistent MoE deployment and performance optimization in wireless systems [202].
- **Communication Resource Limitation:** In decentralized wireless networks, experts may be distributed across multiple edge nodes with heterogeneous capabilities and variable connectivity. Coordinated MoE execution under such conditions requires efficient model partitioning, inter-node communication, and synchronization. However, limited spectrum, backhaul capacity, and transmission power can severely degrade collaboration performance among MoE experts. These constraints introduce new challenges in expert placement, load balancing, and communication overhead. For example, dynamic wireless links may cause fluctuating delays and partial expert failures, thereby complicating expert routing and increasing inference uncertainty [108].
- **Trade-offs Between Model Performance and Resource Efficiency:** The expert activation mechanism of MoE provides flexibility to control the number of active experts during inference. However, practical deployment requires a delicate balance between model accuracy and system resource usage. Over-activating experts generally leads to higher prediction performance but incurs increased latency, memory access, and energy consumption, potentially offsetting the overall gains in model efficiency. Conversely, aggressive sparsity may compromise task performance in wireless systems due to insufficient expert capacity to adapt to rapid variations in dynamic wireless

environments [203]. Therefore, it remains challenging for adaptive expert scheduling strategies to meet application-specific QoS requirements, such as task priority, channel conditions, or residual battery levels, while also satisfying practical system resource constraints.

2) *Privacy and Security Challenges in Multi-Agent MoE Systems:* The deployment of MoE in distributed wireless environments, such as vehicular networks, multi-UAV coordination, and distributed ISAC, inherently aligns with the architectural paradigm of multi-agent AI systems, where autonomous agents exchange intermediate representations and engage in collaborative inference. However, the openness and decentralization of the MoE framework increase the system's exposure to security vulnerabilities, undermine trustworthy collaboration, and pose significant challenges to the resilience of distributed decision-making under dynamic network conditions.

- **Privacy Leakage through Expert Sharing:** In collaborative multi-agent MoE setups, expert modules may be offloaded, shared, or replicated across heterogeneous nodes (e.g., vehicle-to-vehicle or UAV-to-UAV), raising privacy concerns in systems composed of multiple expert agents [204]. Even without direct data exchange, privacy leakage can occur through latent features and activation patterns that implicitly encode sensitive user behaviors or location traces. Recent studies have demonstrated that collaborative inference and activation inversion attacks can reconstruct user attributes or usage histories, while gradient-based leakage further exposes training-time identities [205]. These risks are aggravated in multi-agent MoE due to frequent expert migration and routing synchronization across agents. To mitigate such threats, privacy-preserving communication and federated MoE frameworks can be leveraged, integrating differential privacy, encrypted aggregation, and split learning to protect shared representations while maintaining model accuracy [206]. Incorporating these mechanisms is essential for constructing trustworthy and privacy-aware multi-agent MoE systems tailored to wireless edge networks.
- **Trust Mechanisms for Multi-Agent MoE:** In decentralized multi-agent MoE deployments, reliable inter-agent trust serves as a critical foundation for secure and coordinated expert collaboration. Traditional approaches such as identity verification are often insufficient in dynamic, decentralized settings where agent behaviors are heterogeneous and unpredictable [207]. To establish operational trust between agents, mechanisms such as reputation-based expert scoring [208] to evaluate historical performance, policy-consistent access control [209] to prevent unauthorized expert invocation, and tamper-evident audit trails [210] to record expert selection and routing can serve as effective means for secure MOE expert collaboration. Consistent with trust requirements in edge intelligence-based MoE, blockchain-assisted ledgers can record model versions and inter-agent transactions, while dynamic trust scores guide expert selection and

task delegation under intermittent connectivity [211]. Moreover, multi-agent trust frameworks, such as trust inference and attack detection, offer generalizable patterns that can be seamlessly integrated into MoE expert selection and coordination across heterogeneous agents [212].

- **Assurance and Resilience in Trustworthy Multi-Agent MoE:** Ensuring assurance and resilience in multi-agent MoE systems remains a fundamental and challenging task, requiring that expert routing and inter-agent collaboration be verifiable, explainable, and robust against both system-level failures and adversarial threats. Within the trustworthy assurance layer, providing interpretable explanations and calibrated uncertainty estimates for MoE gating and routing strategies is essential to enable traceable decision-making and support risk management in large-scale, decentralized environments [213]. Recent research suggests that combining interpretability with uncertainty quantification can enhance trust calibration, particularly when multiple agents jointly select experts under dynamic communication and resource constraints [214]. Additionally, extending assurance to system-level operation introduces additional complexity, particularly in wireless edge deployments where achieving runtime integrity becomes a critical challenge. Ensuring trust and resilience across edge agents requires not only remote attestation to verify device integrity but also redundancy-aware task allocation and dynamic trust reconfiguration to sustain reliable cooperation [215]. Ultimately, trustworthy and resilient operations form the foundation for multi-agent MoE systems, enabling stable performance under dynamic and uncertain environments.

Lessons Learned: The integration of MoE into wireless communication systems has demonstrated promising potential in improving adaptability, computational efficiency, and intelligent decision-making across diverse wireless scenarios. Through leveraging sparse expert activation and dynamic selection mechanisms, MoE enables efficient task execution under wireless edge environments, such as vehicular [71], [72], UAV [86], [87], [88], and satellite networks [89], [90]. Furthermore, MoE supports multi-modal data fusion and adaptive resource allocation, which are essential for emerging paradigms such as multi-access control [91], ISAC [92], and mobile edge computing [93], [94]. However, the practical deployment of MoE in wireless networks remains constrained by several challenges, such as increased computational overhead and imbalanced expert utilization under dynamic and uncertain environments. These challenges become especially critical in large-scale, resource-constrained wireless edges, where bandwidth, energy, and storage resources are severely restricted. Future research requires a focus on developing lightweight and generalizable MoE architectures, improving balanced expert activation, and designing resource-aware gating mechanisms that jointly optimize quality, latency, and energy efficiency for next-generation wireless systems.

IV. APPLICATIONS OF MIXTURES OF EXPERTS IN WIRELESS NETWORKING

In this section, we explore key applications of MoE in physical layer communication and resource management, including channel prediction and estimation, physical layer signal processing, radio resource allocation, as well as higher-level network optimization, including traffic management, distributed computing, and network security. A detailed summary of these applications is presented in Tables V and VI.

A. Physical Layer Communication and Resource Management

1) *Channel Prediction and Estimation:* In wireless communication, channel prediction and estimation are critical tasks for ensuring optimal system performance, especially in complex environments involving multiple users and channel conditions [216]. Traditional channel prediction methods typically rely on prior channel models and limited historical data [217]. However, these methods may struggle in dynamic environments with rapidly changing channel conditions [218]. The key advantage of the MoE model lies in its ability to employ multiple experts specialized in different aspects of the channel, which can significantly enhance the accuracy of estimation and prediction under complex channel environments [219].

For example, the authors in [95] address the critical challenge of adapting to fast time-varying CSI in multi-path fading environments by introducing a self-organizing map (SOM)-based MoE framework. Each MoE expert module is implemented using a functional link neural network (FLNN), which offers a computationally efficient alternative to conventional DNN models while maintaining robust approximation capabilities [220]. The FLNN-based experts serve as non-linear channel predictors, effectively capturing the complex temporal variations of Rayleigh fading channels. Additionally, a radial basis function (RBF) network is utilized as the gating mechanism, dynamically assigning weights to different experts based on the extracted local channel characteristics. Through extensive simulations in an OFDMA system with a 5-tap Rayleigh fading channel and a maximum Doppler spread of 40 Hz, the proposed predictor demonstrates its capability to forecast CSI five steps ahead (5 ms prediction horizon). This predictive capability significantly enhances adaptive modulation schemes by mitigating the impact of outdated CSI. In addition, simulation results indicate that the MoE-based predictor achieves a 19.2% improvement in channel prediction accuracy, while simultaneously maintaining a low bit error rate (BER) under real-time prediction constraints.

Building on the strengths of hybrid approaches, the authors in [96] introduce an MoE-based framework for channel gain (CG) estimation, integrating both location-based (LocB) and location-free (LocF) predictive models. As illustrated in Figure 5(a), the LocB expert utilizes estimated locations of transmitter and receiver as input, employing the kernel ridge regression (KRR) method [221] with a distance-based loss to infer CG variations. In contrast, the LocF expert extracts pilot signal features, such as time-of-arrival (ToA) from the channel

TABLE V

SUMMARY OF MoE APPLICATIONS IN PHYSICAL LAYER COMMUNICATION AND RESOURCE MANAGEMENT. **NOTE:** LIGHT BLUE CIRCLES INDICATE THE MoE METHOD, GREEN CHECKMARKS AND RED CROSSES REPRESENT THE ADVANTAGES AND CHALLENGES OF MoE

| Wireless Technology | Ref. | Wireless Task | MoE Framework | MoE Advanvatages & Challenges |
|-------------------------------------|-------------|----------------------------------|---|--|
| Channel Prediction and Estimation | [95] | Channel prediction | Functional link neural network (FLNN)-based MoE experts | <ul style="list-style-type: none"> ● Employ FLNN experts to capture channel dynamics ✓ Maintain low BER under real-time prediction constraints ✗ Suboptimal expert allocation due to SOM limitations |
| | [96] | Channel gain estimation | Location-based (LocB) and location-free (LocF) experts | <ul style="list-style-type: none"> ● Utilize MoE to combine LocB and LocF estimations ✓ Improve channel gain mapping accuracy under uncertainty ✗ Gating function optimization may suffer from local minima ✗ Generalization to new environments remains uncertain |
| | [97] | Radio map estimation | Transfer learning based MoE | <ul style="list-style-type: none"> ● Use transfer learning to fine-tune LocB and LoC experts ✓ Reduces data demand via knowledge transfer ✗ Effectiveness of MoE adaption depends on similarity between source and new environments |
| Signal Processing and Communication | [98] | Neural receiver design | CNN-based experts with hard gating mechanism | <ul style="list-style-type: none"> ● Select specialized SNR experts for received signal ✓ Improve block error rate in SIMO systems ✗ Hard gating may cause expert selection discontinuity |
| | [99], [100] | RF power amplifier linearization | MLP-based experts with sparse gating mechanism | <ul style="list-style-type: none"> ● Use Top-K to select relevant experts for signal modeling ✓ Improve power efficiency and maintain signal quality ✗ Unbalanced expert training due to sparse gating |
| | [61] | Joint source-channel coding | VAE encoder-based experts with dense gating mechanism | <ul style="list-style-type: none"> ● Employ different encoder experts and reconstruct signal ✓ Improve adaptation for varying channel conditions ✗ High training complexity due to dense gating |
| Radio Resource Allocation | [101] | Power control | MLP-based MoE experts and gating network | <ul style="list-style-type: none"> ● Select experts for different noise levels ✓ Enable decentralized power control without retraining ✗ Decentralized architecture may impact global efficiency |
| | [102] | Bandwidth allocation | Knowledge-based Dynamic neural network (DyNN) | <ul style="list-style-type: none"> ● Adjust model capacity based on resource constraints ✓ Balance computation and transmission delay ✗ Offline knowledge base may have limited generalization |
| | [103] | Network task optimization | LLM-based gating network | <ul style="list-style-type: none"> ● LLM-based MoE for different optimization tasks ✓ Reduce overhead while improving decision efficiency ✗ High complexity and inference latency of LLM |

impulse response. The KRR method is subsequently employed for CG gain estimation in environments where accurate positioning is unreliable due to multipath propagation and NLoS effects. The key advantage of the MoE-based approach lies in its adaptive gating mechanism, which dynamically assigns weights to the LocB and LocF experts based on localization uncertainty. The gating mechanism is optimized using block-coordinate minimization (BCM) [222], which iteratively updates the gating function weights and expert parameters to improve estimation accuracy while mitigating overfitting risks. Experimental results demonstrate that this hybrid MoE framework achieves a 35% reduction in location-based estimation errors compared to conventional LocB-only approaches, while significantly improving CG mapping accuracy in NLoS conditions.

Unlike the study in [95], which employs a simple FLNN structure that may struggle to capture complex channel characteristics, and the study in [96], which relies on BCM for gating optimization but risks suboptimal expert selection due to local minima, the authors in [97] enhance expert selection and channel estimation by leveraging DNN-based LocB and Loc experts, along with a DNN-based gating network. In addition, to further extend its capabilities to the new wireless environment, the authors employ a transfer learning approach for wireless environment adaptation. As depicted in Figure 5(b), when it comes to a new wireless environment, the authors either fine-tune a pre-trained MoE model via

transfer learning or directly train a new network without model transfer. A CNN module is employed to extract features and compute the similarity between the source and new environments, determining the amount of data required for fine-tuning the pre-trained MoE network. Through the transfer learning scheme, the proposed approach requires only 20%-40% of additional data to retrain the prediction model for target wireless environments. Besides, the MoE-based approach outperforms individual LocB and Loc models in mean squared error (MSE) by 26%-47%, demonstrating significant effectiveness in improving radio map estimation accuracy under dynamically changing wireless environments.

2) *Physical Layer Signal Processing and Communication:* Traditional wireless transceivers employ predetermined signal processing modules encompassing channel coding, modulation, equalization, and demodulation, as depicted in Figure 6(a). While these techniques are foundational to modern communication systems, their rigid structure limit adaptability to dynamic wireless environments and fluctuating channel conditions [99]. Neural network-based transceivers have emerged as a transformative alternative, offering enhanced signal processing, channel modeling, and error correction capabilities [223]. However, the increasing model complexity and computational demands of neural transceivers pose substantial deployment challenges [224]. MoE addresses these challenges by employing multiple experts and dynamically activating specialized experts, significantly reducing parameter overhead

TABLE VI

SUMMARY OF MoE APPLICATIONS IN NETWORK OPTIMIZATION AND SECURITY **NOTE:** LIGHT BLUE CIRCLES INDICATE THE MoE METHOD, GREEN CHECKMARKS AND RED CROSSES REPRESENT THE ADVANTAGES AND CHALLENGES OF MoE

| Wireless Technology | Ref. | Wireless Task | MoE Framework | MoE Advantages & Challenges |
|--|-------|------------------------------------|--|--|
| Traffic Management | [104] | Traffic state prediction | GNN-based experts with MLP-based gating network | <ul style="list-style-type: none"> ● Use entropy-based loss to improve expert specialization ✓ Enhance spatial-temporal relationship modeling ✗ High computational cost due to multiple graph experts |
| | [105] | Traffic congestion prediction | Attention experts with hierarchical gating network | <ul style="list-style-type: none"> ● Employ multimodal MoE for spatial-temporal dependency ✓ Enhance robustness by trend and periodic experts ✗ Real-time inference efficiency needs improvement |
| | [106] | Vehicle trajectory prediction | MLP-based MoE in Transformer block | <ul style="list-style-type: none"> ● Employ MoE within Transformer for trajectory prediction ✓ Reduce error accumulation in non-autoregressive decoding ✗ High cost due to Transformer and MoE integration |
| Edge Computing and Distributed Systems | [107] | Multi visual task inference | MLP-based experts and MLP-based gating network | <ul style="list-style-type: none"> ● MoE training at individual devices for multi-vision tasks ✓ Reduces memory overhead and energy consumption ✗ Lacks support for distributed coordination among devices |
| | [108] | Wireless distributed LLM inference | FNN-based experts and MLP-based gating network | <ul style="list-style-type: none"> ● Gating network deployed at BS while experts at devices ✓ Achieves collaborative inference on wireless devices ✗ Lack storage management and offloading strategy |
| | [109] | Memory-aware LLM inference | FNN-based experts with sparse gating mechanism | <ul style="list-style-type: none"> ● Top-2 experts based on storage hierarchies across devices ✓ Improve memory usage and reduce inference latency ✗ LLM inference sensitive to wireless link quality |
| Security and Anomaly Detection | [110] | Pilot spoofing attack detection | CNN-based experts with dense gating mechanism | <ul style="list-style-type: none"> ● CNN-based experts to process angular channel fingerprint ✓ Achieve high detection accuracy in low-SNR regime ✗ Performance degrades under severe fingerprint overlap |
| | [111] | Network Intrusion detection | CNN-based experts with sparse gating mechanism | <ul style="list-style-type: none"> ● Top-32 among 128 experts for handling extracted features ✓ Improves detection accuracy across diverse attack types ✗ Requires extensive labeled data and sensitive to data quality |
| | [112] | Network Anomaly Detection | MLP-based experts with sparse gating mechanism | <ul style="list-style-type: none"> ● MoE-based diffusion model for wireless anomaly detection ✓ Improve accuracy through denoising expert specialization ✗ Limited evaluation on real-world wireless datasets |

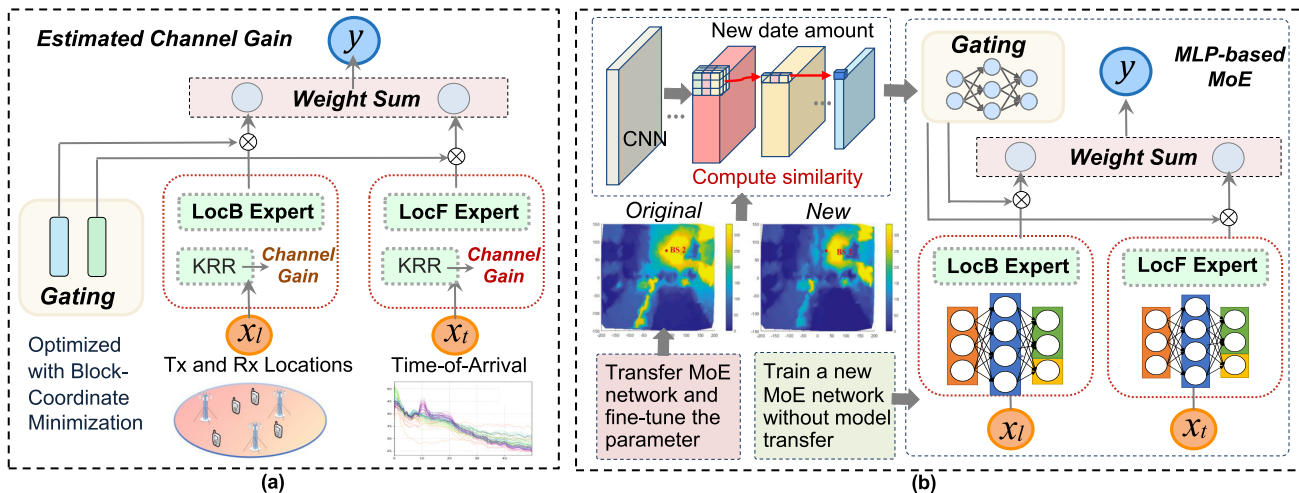


Fig. 5. Illustration of channel estimation using MoE. (a) The MoE framework in [96] combines a LocB expert that uses the transmitter’s and receiver’s locations and a LocF expert based on time-of-arrival features, weighted by a gating network optimized through block-coordinate minimization. (b) Exploit transfer learning to fine-tune a pre-trained MoE model or directly train a new MoE network without model transfer [97]. For the transfer learning scheme, a CNN module is employed to compute the similarity between the source and new environments, determining the amount of data required for fine-tuning the pre-trained MoE.

while improving adaptability to diverse communication conditions.

In the realm of neural transmitter, MoE has demonstrated significant efficacy in radio frequency (RF) power amplifier (PA) linearization by effectively mitigating nonlinear distortion effects. Specifically, the authors in [100] propose a sparsely gated MoE neural network (MENN) framework, which integrates a real-valued time-delay neural network (RVTDNN) as expert models, coupled with a top-K gating mechanism to

optimize computational efficiency. The experts of the proposed MENN are trained to capture distinct nonlinear and memory effects of PAs. The gating network is designed as a fully connected neural network, taking the instantaneous signal envelope as input and determining the optimal expert combination via a softmax-based probabilistic weighting function. When applied to 100 MHz and 120 MHz 5G NR OFDM signals, the proposed approach achieves significant reductions in modeling error and adjacent channel leakage ratio compared

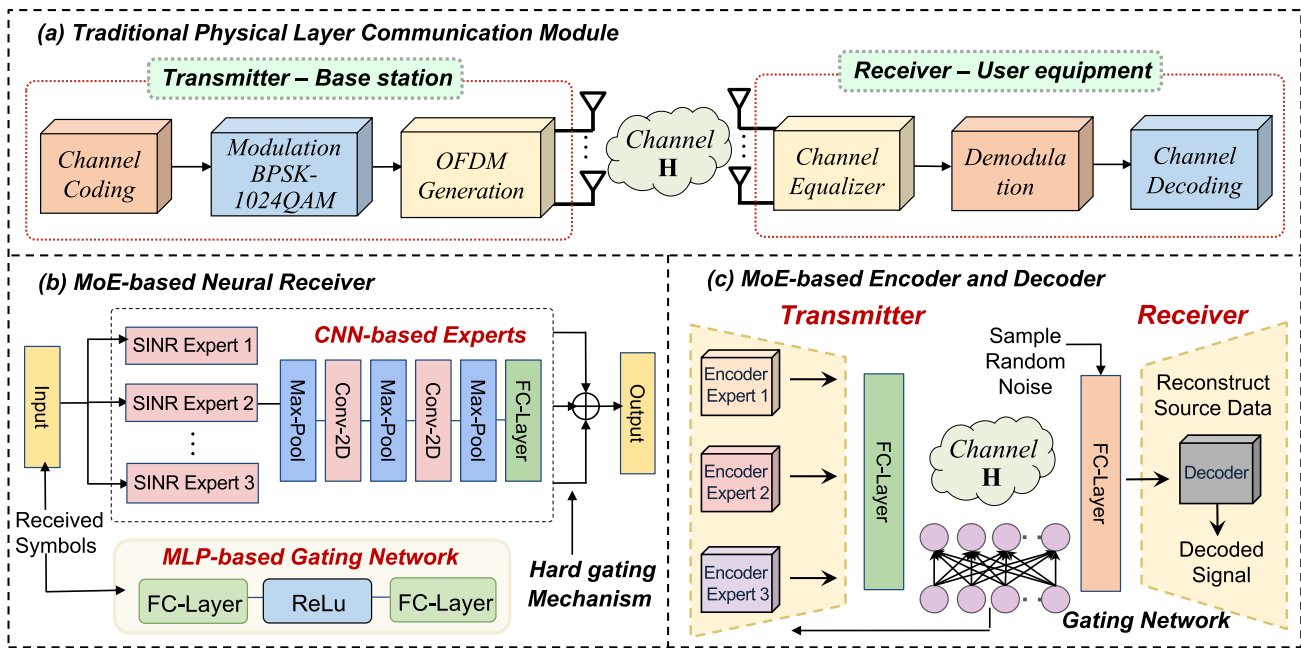


Fig. 6. Illustration of physical layer transmission for wireless communication. (a) The traditional physical layer modules, consisting of channel coding, modulation, OFDM generation, equalization, demodulation, and decoding. (b) The MoE-based neural receiver employs a hard gating network to select CNN-based experts, each specializing in different SINR conditions [98]. (c) The MoE-based joint source-channel coding framework employs multiple encoder experts and an MLP gating network to adapt encoding strategies to varying source characteristics [61].

to conventional NN-based methods. In addition, the MENN reduces half run-time complexity while maintaining superior linearization performance, demonstrating its practical viability for high-throughput real-time implementations.

Different from the study in [100] that focuses on enhancing the transmitter-side PA linearization, the authors in [98] address the receiver-side processing in neural transceivers. As illustrated in Figure 6(b), an MoE-based adaptive neural (MEAN) receiver is introduced for received signal decoding in single-input multiple-output (SIMO) systems. Unlike conventional static neural receivers, which deploy a neural network to handle all channel conditions, the MEAN architecture adopts a dynamic neural network paradigm, leveraging a hard-gated MoE framework to optimize both computational efficiency and adaptability. The hard gating mechanism dynamically activates a CNN-based expert, which is specialized for distinct SINR conditions. Instead of using dense gating mechanisms, the hard gating expert selection ensures that only a single expert is active at any given time. This design significantly reduces runtime computational complexity by up to 50% compared to static neural receivers, while maintaining competitive demodulation accuracy. Additionally, MEAN surpasses conventional least-squares (LS)-based estimation techniques, especially in moderate-to-high SINR regimes, underscoring the efficacy of MoE-driven methods in next-generation wireless receivers.

Extending beyond separate optimizations of the transmitter or receiver, the study in [61] integrates MoE into a unified end-to-end transceiver framework, addressing both encoding and decoding through joint source-channel coding (JSCC). As depicted in Figure 6(c), the authors in [61] introduce an MoE-based JSCC framework that leverages a mixture of variational autoencoders (MoVAE) to enhance adaptive signal encoding and reconstruction in additive white Gaussian noise

(AWGN) channels. Unlike traditional parametric encoding approaches, the proposed system employs multiple specialized VAE-based encoders to handle distinct regions of the source space, along with a universal decoder that ensures robust signal reconstruction. To dynamically assign encoders based on input characteristics, the framework uses an MLP-based gating network with an exponential loss function for encouraging encoder specialization while preventing mode collapse. Experimental results show that MoVAE achieves state-of-the-art rate-distortion performance, surpassing power-constrained vector quantization schemes, with a performance gap of less than 1dB from the Shannon limit in various bandwidth usage scenarios. The incorporation of an MLP-based gating network allows adaptive encoder selection based on real-time channel conditions, leading to a 20% improvement in peak signal-to-noise ratio (PSNR) compared to single-VAE methods.

3) *Radio Resource Allocation*: Efficient radio resource allocation is fundamental to optimizing wireless network performance, encompassing key tasks such as power, spectrum, and bandwidth allocation [225], [226], [227]. In next-generation wireless networks, achieving optimal resource distribution is challenging due to dynamic channel conditions, diverse user requirements, and computational constraints [228], [229], [230]. Traditional methods, such as iterative convex optimization or static machine learning approaches, often suffer from rigid structure, limited adaptability, and high computational overhead, rendering them unsuitable for rapidly changing environments [231].

Recent advancements in resource management have utilized MoE to tackle the challenges of transmit power management. In [101], the authors propose a deep MoE (DMoE) framework for decentralized power control, addressing the challenge of adapting to time-varying CSI uncertainty without requiring

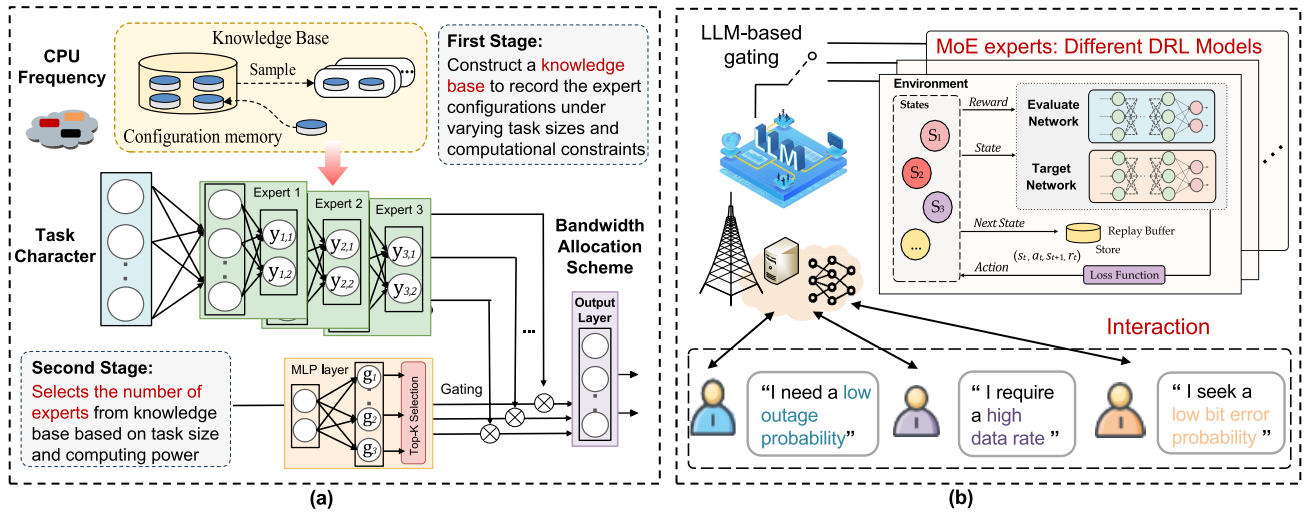


Fig. 7. Illustration of radio resource allocation and management using MoE. (a) The knowledge-assisted two-stage framework for bandwidth allocation [102]. In the first stage, a knowledge base is constructed by training models with different numbers of MoE experts. In the second stage, the system dynamically selects the number of experts based on task size and available computing power, ensuring efficient bandwidth allocation. (b) The LLM-based MoE framework in [103] dynamically assigns DRL experts for network task optimization based on the quality of service (QoS) requirements, such as minimizing outage probability, maximizing data rate, or reducing bit error probability.

frequent model retraining. Unlike conventional data-driven power control schemes, which assume a fixed noise level of CSI feedback and require costly retraining when noise statistics change, the proposed DMoE enables each expert to specialize in a different feedback noise regime. A gating network, implemented as a lightweight DNN, dynamically selects the most suitable expert based on the estimated CSI uncertainty level, effectively adapting to varying noise statistics. This selective expert learning strategy ensures robust power allocation policies while significantly reducing retraining overhead. Experimental results over a Rayleigh fading channel with multi-user interference show that DMoE achieves a 12% higher sum-rate than conventional DNNs, particularly under dynamically varying feedback noise conditions. Compared with weighted minimum mean square error (WMMSE)-based approaches, DMoE improves sum-rate by approximately 20% while significantly reducing computational complexity, which shows the promising potential for real-time distributed resource allocation in next-generation wireless networks.

While the study in [101] leverages MoE to enhance power control adaptability, it employs a fixed network architecture that does not dynamically adjust to varying computational constraints and resource availability. To address this limitation, the study in [102] extends MoE-based decision-making by incorporating a knowledge-assisted dynamic neural network (DyNN) for bandwidth resource management. As depicted in Figure 7(a), the maximum number of MoE experts is investigated in relation to performance, revealing that the optimal width of DyNN varies with both task size and CPU frequency. Instead of using static networks with fixed maximum experts, DyNN dynamically adjusts its network width through a two-stage optimization. In the first stage, a knowledge base is constructed by training models with different maximum MoE expert numbers. The system analyzes their performance under varying tasks and computational constraints (e.g., CPU frequency) and records the optimal expert configurations. In the second stage, the maximum number of experts is dynamically

selected based on the knowledge base, considering the current task and available computing power. Simulation results demonstrate that increasing the maximum number of MoE experts improves bandwidth allocation performance, but at the cost of higher computational complexity. For instance, using only 1 expert achieves 93.6% of the full 30-expert model performance, with less than 4% of the total operation costs, while using 5 experts achieves over 98% performance. In addition, by leveraging prior knowledge for adapting network width and computational resources, DyNN realizes a 47.1% reduction in service delay for small tasks and 12.8%–20.6% reduction for larger tasks compared to static neural networks.

However, knowledge-based expert selection may be inherently limited in scenarios with unpredictable network situations, where offline knowledge may not generalize to unseen tasks. To further enhance adaptability, the authors in [103] propose an LLM-enabled MoE framework that eliminates the need for predefined expert mappings. Through leveraging the contextual reasoning capabilities of LLMs to infer user requirements, LLMs themselves can be employed as intelligent gating mechanisms to select experts and assign expert weights accordingly. As illustrated in Figure 7(b), a key application is the utility maximization of a network service provider (NSP), where users exhibit varying quality of service (QoS) demands, such as low outage probability (OP) for uninterrupted voice calls or high data rates for streaming applications. The LLM-based MoE first interprets textual user requests, translating them into formal optimization objectives. It then selects the most relevant DRL experts, such as OP minimization experts or data rate maximization experts, and synthesizes their outputs to determine an optimal transmit power allocation strategy. Unlike traditional gating networks, LLMs can obtain human intent from natural language, enabling better alignment between user requests and optimization goals in diverse user-centric tasks. Compared to single DRL-based power control, the LLM-based MoE approach achieves over 40% reduction in computational cost, resulting in a 15%

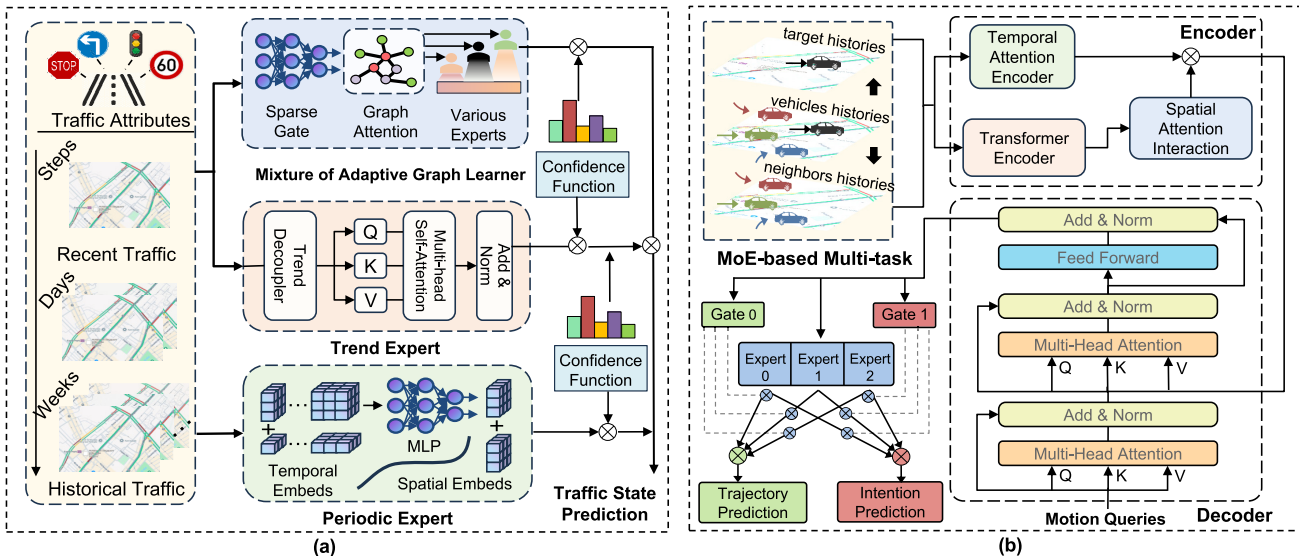


Fig. 8. (a) The CP-MoE framework in [105] introduces a hierarchical gating structure to model distinct temporal patterns, where experts are designed for stable traffic trends and periodic variations. (b) The TAME framework in [106] addresses simultaneous trajectory forecasting and driver intention inference. A non-autoregressive Transformer with temporal and spatial attention encoders is utilized to capture intra-vehicle and inter-vehicle dependencies, along with an MoE-based decoder for multi-modal trajectory generation and driver behavior prediction.

revenue gain for the NSP under varying network loads and user demands.

B. Network Optimization and Security

1) *Traffic Management*: Traffic management involves intricate spatial-temporal dependencies, rapidly evolving vehicular environments, and stringent requirements for real-time decision-making [232], [233], [234]. MoE frameworks can effectively utilize specialized expert models for distinct tasks such as traffic flow prediction, time series analysis, and driving motion estimation, thereby enhancing performance and adaptability in complex traffic optimization scenarios [235]. To capture the complex spatial structures of road networks and the temporal dependencies in traffic dynamics, the authors in [104] propose an MoE-based spatial-temporal graph convolutional network (STGCN) for traffic state prediction. The proposed approach integrates multiple graph neural network (GNN)-based experts, with a neural gating network to dynamically assign expert weights based on real-time traffic inputs. Unlike traditional deep learning models that rely on a single traffic predictor, STGCN effectively learns region-specific traffic patterns through different experts. To prevent expert overfitting and encourage specialization, the authors introduce a novel entropy-based loss function, which ensures that different experts focus on distinct regions of the input space rather than all experts contributing equally to every prediction. Experiments conducted on 228 road segments with traffic flow data demonstrate that, compared to single GNN model, STGCN more effectively captures diverse spatial-temporal traffic patterns, resulting in a 12.3% reduction in mean absolute error (MAE) for 15-minute forecasts, 10.6% for 30-minute forecasts, and 8.8% for 45-minute forecasts.

However, while the study in [104] utilizes the MoE framework to enhance traffic state prediction, it does not explicitly differentiate between long-term stable and

short-term non-recurring fluctuations in urban traffic, limiting its adaptability to varying congestion scenarios. To address this challenge, the authors in [105] propose a congestion prediction MoE (CP-MoE) framework, which extends conventional MoE by introducing a hierarchical gating structure that explicitly models different congestion patterns. As depicted in Figure 8(a), distinct experts of the CP-MoE framework specialize in learning stable traffic trends (e.g., recurring peak-hour congestion) and periodic variations (e.g., congestion caused by specific events), allowing for more refined temporal modeling and improving adaptability. A hierarchical integration with confidence functions dynamically selects and fuses experts based on traffic stability, enhancing the model's interpretability through different expert contributions. Extensive experiments on real-world traffic datasets validate the effectiveness of CP-MoE, achieving a 3.9% improvement in accuracy and a 5.7% improvement in congestion forecast tasks. These results shed light on the applicability of MoE in spatial-temporal pattern modeling for wireless-enabled intelligent transportation systems and broader vehicular network scenarios.

Unlike the previous studies in [104] and [105], which focus on modeling temporal patterns and solely predicting traffic flow dynamics, the authors in [106] simultaneously address trajectory forecasting and driver intention inference for multi-task processing in autonomous driving. To handle diverse driving scenarios and uncertainty in multi-modal trajectory and intention prediction, this study proposes a non-autoregressive Transformer with attention-based MoE (TAME) framework. As illustrated in Figure 8(b), this framework consists of a temporal attention encoder that captures intra-vehicle and inter-vehicle dependencies and a spatial-attention interaction encoder that refines vehicle interaction representations. Besides, a multi-task prediction decoder is employed based on MoE structure to generate multi-modal trajectory predictions while simultaneously inferring driver intentions,

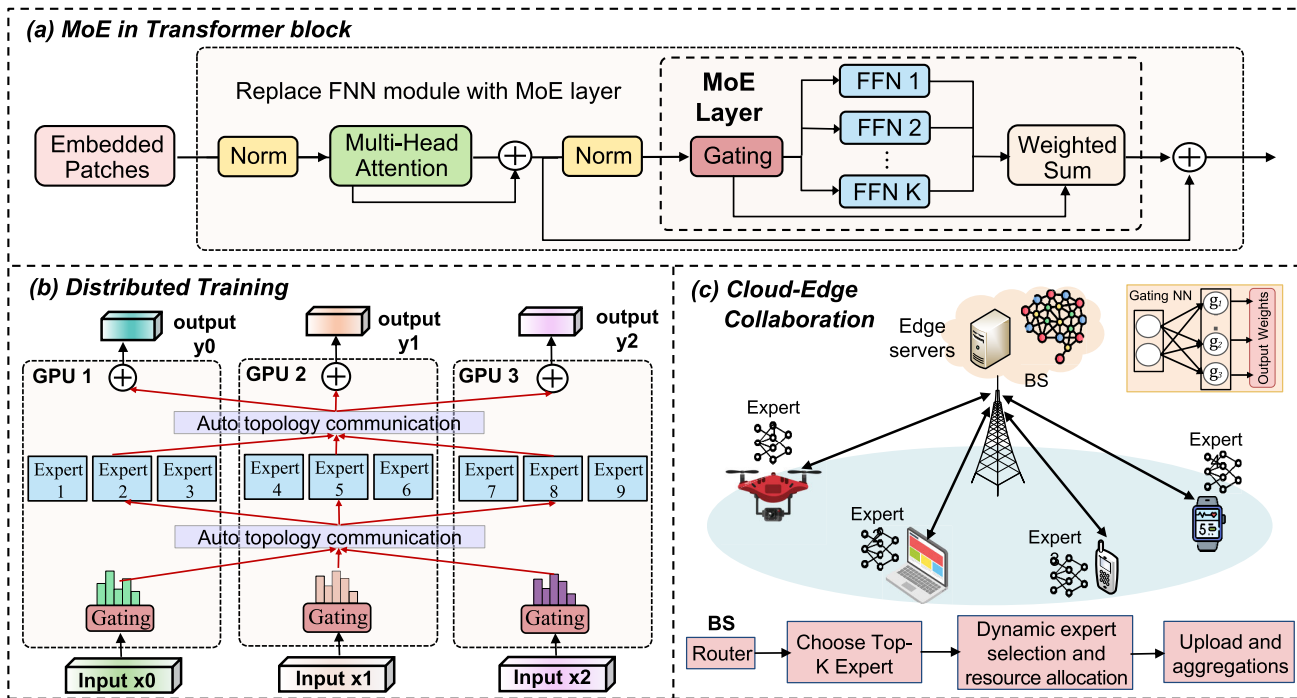


Fig. 9. Illustration of MoE-based frameworks for the distributed and collaborative computing paradigm. (a) MoE integration in Transformer blocks replaces traditional FFN with multiple smaller expert sub-networks, coordinated by a coordinated by a gating network to activate a sparse subset of experts [49]. (b) FlexMoE [236] enables distributed training across multiple GPUs by dynamically assigning expert sub-networks to heterogeneous devices. (c) WDMoE architecture [108] supports collaborative inference of LLMs across cloud-edge systems. Attention modules and a gating network are executed at the BS, while lightweight expert networks are distributed among edge devices.

which strengthens the model’s representational capacity across diverse motion patterns and predictive tasks. Experimental evaluations demonstrate that the proposed approach outperforms state-of-the-art motion prediction models, achieving a 28.6% reduction in average displacement error. Additionally, the approach exhibits superior long-term prediction accuracy, with a 68% improvement in root mean square error (RMSE) for 4-5 second predictions, showcasing its capability in handling complex, long-horizon motion forecasting tasks.

2) *Edge Computing and Distributed Systems*: MoE frameworks provide an effective paradigm for parallel and collaborative processing, thereby facilitating their application in edge computing scenarios that are constrained by limited resources and specialized computational demands [237], [238], [239]. As illustrated in Figure 9(a), MoE was initially applied to large-scale Transformer models by replacing the original FFN layers with multiple smaller expert sub-networks [49]. With their inherent modularity and parallelism, the frameworks of MoE have enabled large-scale training and inference across multiple devices [240]. For instance, FlexMoE [236], a distributed MoE training framework illustrated in Figure 9(b), supports expert parallelism across multiple GPUs of heterogeneous devices based on workload distribution, demonstrating significant scalability of MoE frameworks for distributed model training.

Building upon the development in distributed training, MoE has been explored for inference optimization on edge devices with limited computational resources. To enable efficient multi-task visual inference on resource-constrained edge platforms, the authors in [107] propose Edge-MoE, a Vision Transformer (ViT)-based architecture that leverages an MoE

mechanism with task-specific experts for edge deployment. Edge-MoE integrates multiple MLP-based expert networks, each specialized for different vision tasks such as depth estimation and semantic segmentation. A compact MLP-based gating network is employed to assign expert weights and selectively activate a subset of experts based on each task and input token, significantly reducing memory usage and computational overhead. In addition, to support real-time inference on low-memory edge inference, the authors incorporate techniques including attention reordering [241], expert-level pipelining [242], and approximate activation functions [243], which collectively reduce memory overhead and computational latency. Experiments on autonomous driving benchmarks demonstrate that Edge-MoE achieves up to $18.8\times$ reduction in inference latency and over $4\times$ improvement in energy efficiency compared to traditional ViT baselines. These enhancements highlight the effectiveness of Edge-MoE in supporting real-time, resource-efficient inference in wireless edge scenarios.

However, Edge-MoE primarily focuses on computational optimization at individual edge devices, without addressing the broader challenges of distributed coordination and communication constraints in wireless environments. To enable scalable and low-latency inference of LLMs over wireless edge networks, the authors in [108] propose a wireless distributed MoE (WDMoE) architecture that facilitates collaborative execution of MoE-based LLMs across a BS-equipped server and multiple mobile devices. As depicted in the architecture of Figure 9(c), the computationally intensive attention modules and the MLP-based gating network for server-device collaboration are deployed at the BS, while the lightweight

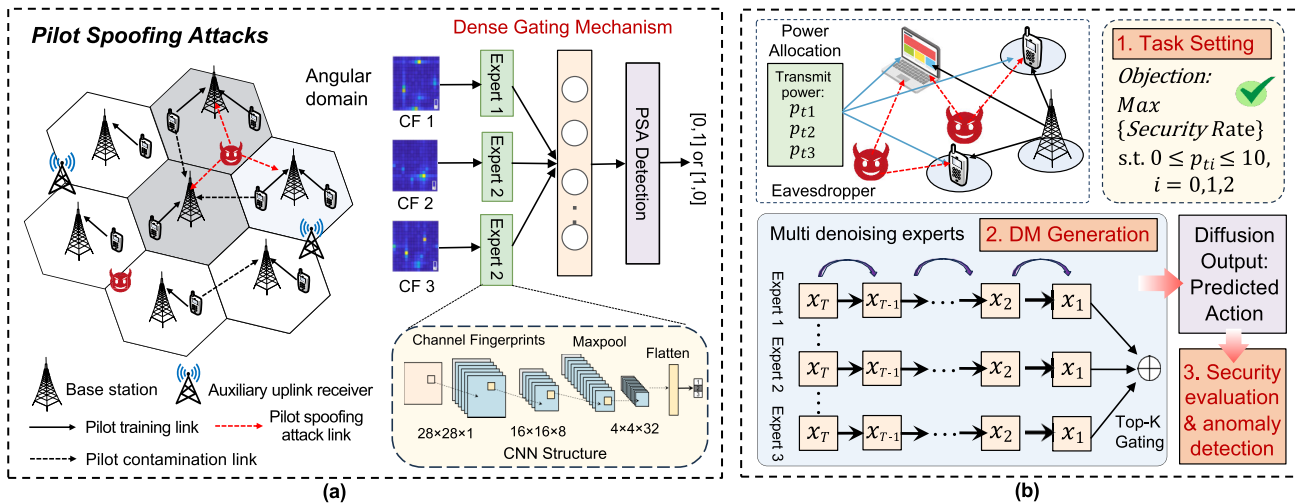


Fig. 10. Illustration of MoE-based frameworks for wireless security and anomaly detection. (a) The Distributed MoE-based Neural Network (D-MoENN) [110] addresses pilot spoofing attacks (PSAs) in multi-cell massive MIMO systems by leveraging angular-domain channel fingerprints (CFs) collected from multiple auxiliary uplink receivers. (b) A generative MoE-based framework combines modular denoising experts with a diffusion model for signal-level anomaly detection, enabling adaptive denoising process across varying received signals [112].

FNN-based expert sub-networks are distributed among mobile devices. This decomposition leverages the modularity and computational parallelism of MoE frameworks, enabling efficient utilization of distributed and heterogeneous wireless edge resources in terms of both computation and communication. Unlike conventional cloud-based LLM inference, WDMoE leverages heterogeneous wireless link qualities and device capabilities to enable dynamic expert selection and bandwidth-aware collaboration. To coordinate this process, an optimization framework is employed to jointly maximize expert utility and minimize overall inference latency under wireless constraints. Simulations are conducted to compare with the Mixtral deployment model [244], which distributes experts across devices but lacks coordinated expert selection and bandwidth optimization. The results show that WDMoE achieves up to a 45.75% reduction in latency on the Physical Interaction Question Answering (PIQA) dataset [245], while maintaining high accuracy across multiple NLP benchmarks.

While WDMoE effectively enables collaborative inference across wireless edge devices, it does not explicitly incorporate fine-grained storage management or adaptive offloading strategies, both of which are critical for efficient deployment in bandwidth and storage-constrained wireless environments. To address these challenges, the authors in [109] propose MedMixtral $8 \times 7B$, a fine-tuned MoE model designed for resource-efficient collaboration between wireless edge nodes and local servers. MedMixtral introduces a memory-aware inference offloading strategy that leverages heterogeneous storage hierarchies across devices, such as GPU memory, system memory, and local disk storage, to balance memory utilization and reduce inference latency. The model adopts a sparse MoE architecture, where 2 out of 8 FNN-based experts are activated for each token and executed on edge devices. This Top-2 gating mechanism, implemented with a single linear projection followed by a softmax function at the BS, computes expert weights based on input representations and system status. Experimental evaluations on real-world question and answer (Q&A) datasets [246] show that MedMixtral reduces

memory consumption by up to 45.1% and effectively improves inference latency by 32% through the proposed offloading strategy, which allows experts to be flexibly distributed across edge devices and BS based on runtime resource availability.

3) *Security and Anomaly Detection*: Ensuring security and detecting anomalies in communication networks is increasingly challenging due to the dynamic nature of threats and the complexity of modern wireless systems. MoE frameworks offer a powerful approach to enhance security measures such as detection accuracy and secrecy rates by dynamically activating specialized experts to capture heterogeneous wireless characteristics [247]. For example, the authors in [110] propose a Distributed MoE-based Neural Network (D-MoENN) framework to address the severe security threats posed by pilot spoofing attacks (PSAs) in multi-cell massive MIMO systems. Traditional detection methods, such as hypothesis testing [248] and angular-domain fingerprinting [249], often suffer from poor performance under low-SNR conditions and lack scalability in distributed settings. D-MoENN leverages spatial diversity from multiple auxiliary uplink receivers, each extracting angular-domain channel fingerprints (CFs) that capture the angle-of-arrival (AoA) characteristics of both legitimate and spoofed signals. As depicted in Figure 10(a), these CFs are processed by CNN-based local expert models, trained to estimate the number of users sharing a pilot sequence. A hidden gating network at the central BS aggregates the expert outputs via learned soft weights and produce a global prediction of the pilot reuse level used for PSA detection. Simulation results show that the proposed D-MoENN achieves up to 90% higher detection accuracy and 10% improvement under low-SNR conditions (e.g., -10 dB).

While D-MoENN in [110] focuses on wireless anomaly detection and secure transmission, network-level intrusion detection remains a critical component of 5G and beyond wireless infrastructures. To handle heterogeneous network traffic and evolving threat models, the authors in [111] propose a sparsely-gated MoE-CNN framework for intelligent and adaptive intrusion detection. The proposed model first reshapes

one-dimensional network flow features into 6×13 matrices, enabling spatial feature extraction through a four-layer CNN module. The resulting representation vector is then passed to an MoE layer composed of 128 fully connected experts, where a gating network dynamically activates the top-32 experts per input. To improve generalization and avoid expert collapse, the framework introduces load balancing regularization for encouraging equitable weight assignment and uniform data allocation among experts. Evaluated on the real-world 5G-NIDD dataset [250], which includes eight types of attacks collected from an operational 5G network, MoE-CNN achieves an overall intrusion detection accuracy of 99.96%, outperforming strong baselines such as the CNN-LSTM approach. These results provide empirical evidence for the MoE framework in addressing diverse network attack types.

Complementing the discriminative approaches used in D-MoENN [110] and MoE-CNN [111], generative modeling techniques have been adopted to address wireless threats such as pilot spoofing and jamming in a more flexible and data-driven manner. However, while recent GenAI models have demonstrated potential in reconstructing perturbed signals and identifying anomalies, they often suffer from limited adaptability to heterogeneous attack scenarios and high computational overhead due to monolithic inference structures. To overcome these limitations, the authors in [112] propose a modular generative framework that integrates the MoE architecture with a diffusion-based denoising process for signal-level anomaly detection. In the proposed design illustrated in Figure 10(b), a sparse gating mechanism dynamically activates a subset of specialized denoising expert networks based on the perturbation characteristics of the received signal. The activated experts enable the detection of both known and unseen anomalies under diverse SNR and channel conditions, guiding the diffusion model toward accurate signal reconstruction. Experimental results demonstrate an average accuracy improvement of 12.3% over the traditional diffusion model baseline, underscoring the effectiveness of combining MoE with generative modeling to achieve adaptive and robust wireless network security.

Lessons Learned: The integration of MoE into wireless systems and technologies demonstrates layer-specific strengths and functional alignment across different wireless protocol stacks. In the foundational physical-layer communication domain, which includes channel prediction [95], [96], [97], signal processing [61], [98], [99], [100], and radio resource management [101], [102], [103], MoE supports model specialization under heterogeneous and time-varying conditions, significantly improving both predictive performance and computational efficiency. In the system and network-layer domain, covering traffic management [104], [105], [106], distributed computing [107], [108], [109], and network security [110], [111], [112], MoE enables communication-aware expert activation and scalable multi-task inference, which satisfies the growing demand for low-latency and adaptive decision-making in complex wireless environments. These developments underscore MoE's potential in modularity, adaptability, and resource-aware design. Nevertheless, ensuring expert diversity, avoiding mode collapse, and developing

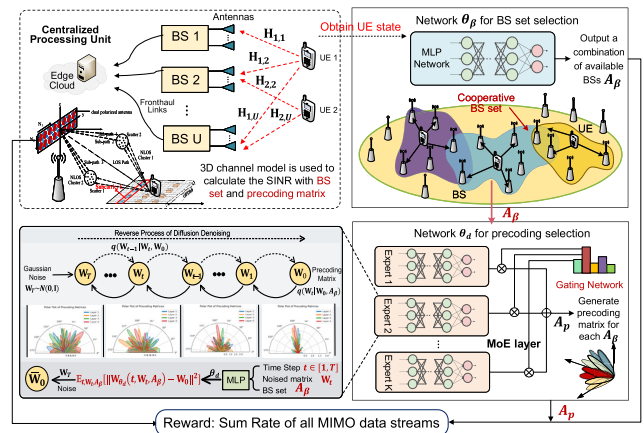


Fig. 11. Multi-BS reception scenario and the proposed DRL network.

lightweight, robust gating mechanisms remain key challenges. A principled exploration of these design aspects is envisioned to be essential for realizing intelligent, flexible, and resilient wireless infrastructures in future 6G and beyond systems.

V. CASE STUDY AND DATASETS

In this section, we first present a case study that integrates MoE into a diffusion model-based DRL framework, demonstrating the performance enhancements of MoE for wireless network optimization. Subsequently, we provide an overview of publicly available datasets that are widely adopted in MoE-based models to support diverse machine learning tasks.

A. Case Study

In modern wireless networks, multi-BS coordinated reception has emerged as a promising access technology to enhance received signal strength and improve spectral efficiency [251]. However, achieving the potential gains remains challenging due to the substantial computation overhead associated with jointly selecting cooperative BS and precoding schemes for each user equipment (UE) [128]. Consequently, we leverage DRL to obtain an optimal cooperative BS set and precoding matrix for multi-BS reception under MIMO transmission.

1) *Problem Description:* As depicted in Figure 11, we investigate the multi-BS reception scenario, where both the BS and UE are equipped with multiple antennas to enable spatial multiplexing and support simultaneous transmission of multiple data streams. The BS receive antennas are assumed to be a cross-polarized uniform planar array (UPA), and the UE received antennas are assumed to be a single-polarized uniform linear array (ULA). The wireless signal propagation is characterized by the three-dimensional (3D) channel model that accounts for both horizontal and vertical propagation effects, thereby providing a more accurate representation of real-world wireless environments. For multi-BS coordinated reception, each UE must select a cooperative BS set from the available BSs and choose a precoding matrix W from a predefined codebook. This joint selection process is computationally prohibitive, particularly in large-scale networks with dense BS deployments and high-dimensional codebooks.

TABLE VII
COMPARISON OF NETWORK ARCHITECTURES

| Method | MoE | Network Architecture |
|-----------------|-----------------|----------------------|
| Top-1 Expert | 4 total experts | 20 → 26 → 32 |
| Top-2 Experts | 4 total experts | 20 → 26 → 32 |
| Baseline | w/o MoE | 20 → 26 → 32 |
| Deeper Baseline | w/o MoE | 20→23→26→29→32 |

2) *Solution Design*: We propose a DRL framework that employs two neural networks to independently generate the cooperative BS set and the precoding scheme as in Figure 11, thereby reducing computational complexity and enabling efficient multi-BS coordinated reception. Note that once the cooperative BS set is decided, we can iterate all the satisfied precoding matrices in the codebook to obtain an optimal precoding matrix that maximizes the SINR of all data streams. The optimal precoding matrix is regarded as expert knowledge and approximated using a diffusion model, which allows the DRL framework to generalize under varying BS cooperation schemes [252]. Nonetheless, despite the effectiveness of the diffusion model, BS selection remains a computationally intensive task, as the size of the search space increases combinatorially with the number of candidate coordinated BSs [251]. To enhance the model capacity, MoE layers are integrated into the denoising phase of the diffusion process, where a sparse gating mechanism activates appropriate experts to perform adaptive denoising and generate the precoding matrix. Specifically, the state, action, and reward function are defined as follows.

- **State**: To achieve multi-BS coordinated reception for each UE, the horizontal and vertical coordinates loc_x, loc_y are used as the state information for the DRL framework.
- **Action**: The action space of the DRL framework comprises two components for each UE, namely the selection of a cooperative BS set \mathcal{A}_B and the determination of a precoding matrix \mathcal{A}_P . For \mathcal{A}_B , we use the binomial coefficient $\binom{U}{N_i}$ to indicate the actions of selecting N_i BSs from a total of U available BSs for UE i . For \mathcal{A}_P , the precoding matrix is generated by the MoE-based diffusion model, which is designed to adapt to different BS selections and varying data stream conditions.
- **Reward**: The reward function is defined as the sum of post-equalized communication rates across all transmit data streams, serving as a key performance metric to evaluate the effectiveness of the set of selected cooperative BSs and the precoding matrix.

3) *Simulation Setting and Results*: To investigate the impact of network architecture on the DRL training performance, we evaluate four models comprising two MoE-based architectures and two baselines without MoE. The architectural specifications are summarized in Table VII. Specifically, both the MoE-based Top-1 and Top-2 models incorporate 4 candidate MLP experts to perform the diffusion denoising process. Each expert consists of three hidden layers with dimensions 20, 26, and 32, respectively. A single-layer gating network outputs the selection probabilities for the 4 experts via a softmax function, from which either one (Top-1) or two

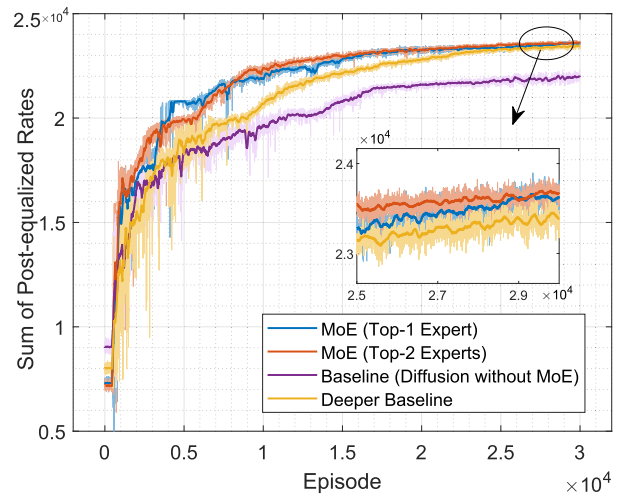


Fig. 12. Training performance of different network architectures.

(Top-2) experts are dynamically selected during each forward pass. In contrast, the other two models do not incorporate expert modules. The standard baseline adopts the same three-layer MLP architecture as the MoE models, whereas the deeper baseline increases the depth to five layers with hidden dimensions of 20, 23, 26, 29, and 32. This deeper configuration is designed to validate whether increased network depth can offset the lack of expert-based specialization.

Figure 12 presents the training curves of the compared models. The simulation environment is based on the Outdoor 1 (O1) scenario provided by the DeepMIMO dataset [123], where a total of 18 BSs are available for coordinated multi-BS reception. It is evident from Figure 12 that both MoE-based architectures achieve faster convergence and higher final reward compared to the other two baselines without MoE modules. The Top-2 Expert model slightly outperforms the Top-1 Expert configuration, demonstrating the benefit of activating more experts in improving learning performance. Additionally, for the two baseline models without MoE, the deeper baseline with increased network depth exhibits a noticeable performance improvement compared to the standard baseline. However, it still lags behind both the Top-1 and Top-2 Expert models in terms of overall performance. Moreover, its training performance exhibits more pronounced fluctuations during the early training stages, suggesting a slower convergence attributed to the larger network scale. These simulation results highlight the advantage of incorporating MoE to enhance the model representation capacity while reducing the computational overhead associated with DRL-based network optimization.

To further validate the effectiveness of MoE in wireless decision-making, we conducted physical-layer simulations to evaluate the throughput performance (measured in Mbit/s) of the MoE-based DRL framework under multi-BS cooperation scenarios. The simulation pipeline includes channel coding, modulation, channel equalization, demodulation, and decoding, which offers an end-to-end evaluation of real-world transmission performance.

As illustrated in Fig. 13, the MoE-based models with Top-1 and Top-2 expert selection consistently outperform both

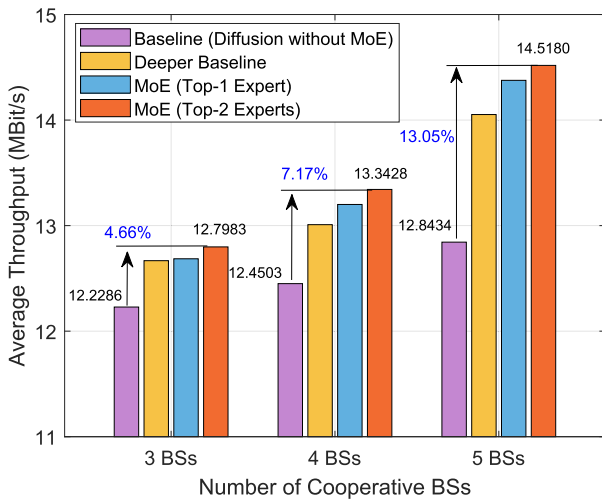


Fig. 13. Average Throughput of different network architectures.

baselines in terms of average throughput across all cooperative BS configurations ranging from 3 to 5 BSs. It is worth noting that the number of possible cooperative combinations increases combinatorially with the number of BSs, with a complexity of approximately $\mathcal{O}(2^n)$ for n cooperating BSs. As the cooperation complexity increases, the performance gains of MoE become more pronounced, rising from 4.66% with 3 BSs to 13.05% with 5 BSs, demonstrating MoE's enhanced effectiveness in more complex network scenarios. These results highlight the capability of MoE to manage high-dimensional decision spaces through scalable and adaptive expert selection, particularly in cooperative transmission tasks.

B. Open-Source Datasets for MoE

High-quality data plays a crucial role in training large-scale AI models and evaluating their performance. In this subsection, we provide an overview of the publicly available datasets that are utilized by MoE-based models in the fields of NLP, computer vision, multimodal, and wireless networks. A summary of these datasets is presented in Table VIII.

1) *Datasets for Natural Language Processing*: In the domain of NLP, datasets such as GLUE [113] and C4 [114] have been widely used within different MoE architectures, e.g., LiteMoE [253] and MoE-based LLM [254], to address diverse language understanding and text generation tasks. GLUE serves as a comprehensive NLP benchmark, supporting sentiment analysis, sentence similarity, and natural language inference, providing a robust ground for various linguistic model generalization and evaluation. Meanwhile, C4, referred to as the Colossal Clean Crawled Corpus, delivers an extensive collection of high-quality text curated from the Common Crawl project.¹ The hundreds of gigabytes of text from C4 and the diverse language tasks on GLUE allow MoE models to learn specialized expert sub-networks, each proficient in distinct aspects of language representation and usage. Additionally, domain-specific datasets such as TeleQnA [115], comprising 10,000 Q&A instances in telecommunications,

facilitate the integration of MoE experts to support compositional reasoning and effective domain adaptation [255].

2) *Dataset for Computer Vision*: Computer vision datasets such as ImageNet [116], nuScenes [117], and DOTA [118] have been extensively utilized by MoE-based models for large-scale image classification and object detection across diverse scenarios. ImageNet serves as a foundational dataset for generic visual recognition, providing 3.2 million labeled images organized by the WordNet hierarchy [256] across a broad spectrum of object categories. This large-scale, structured dataset enables MoE-based models to learn diverse hierarchical features in both supervised and transfer learning settings [257]. Meanwhile, nuScenes supports 3D detection and tracking using multimodal inputs such as cameras, lidars, and radars. The rich annotations and diverse sensor modalities on nuScenes facilitate the training of specialized experts for distinct perception tasks such as autonomous driving [118]. Furthermore, designed for object detection in aerial scenes, the DOTA dataset provides high-resolution aerial imagery with substantial variations in scale, orientation, and scene complexity. Its fine-grained, arbitrarily oriented bounding boxes offer a rigorous benchmark for evaluating advanced MoE-based detection in aerial and remote sensing tasks [258].

3) *Dataset for Multimodal*: In the domain of multimodal research, particularly in image-text and video-text generation tasks, datasets such as MS COCO [119], LAION-5B [120], and HowTo100M [121] have been widely adopted to align visual and textual modalities. Specifically, MS COCO provides 330,000 images, each annotated with five human-written captions, which is commonly applied for fine-grained analysis of visual scenes in tandem with language annotations [126]. LAION-5B is an open-source large-scale dataset containing 5.85 billion CLIP-filtered image-text pairs [135] across multiple languages, offering unprecedented scale and diversity for training and evaluating multimodal models in tasks such as retrieval, captioning, and zero-shot classification [260]. Meanwhile, HowTo100M contains 1.22 million narrated instructional videos with 136 million CLIP-caption pairs and automatic speech recognition (ASR) transcripts, enabling multimodal learning for video-text alignment across 23,000+ task types. Multimodal datasets often present challenges in aligning modalities with different sequence lengths, such as text-image or text-video pairs [263]. MoE-based models have demonstrated the ability to perform robust cross-modal matching strategies through specialized experts trained on these multimodal datasets, enabling effective handling of text-image or text-video alignment within a unified framework [261].

4) *Dataset for Wireless Networks*: Datasets for wireless networks such as RadioML [122], DeepMIMO [123], 5G-NIDD [250], and CitySim [124] serve diverse wireless scenarios ranging from physical-layer signal processing to system-level intrusion detection and vehicular networking. In particular, RadioML offers labeled in-phase/quadrature (IQ) signal recordings, which are utilized by the MoE-based automatic modulation classification model to capture distinctive signal features under varying SINR scenarios [178]. Meanwhile, DeepMIMO, built on accurate ray-tracing simulations,

¹<https://commoncrawl.org>

TABLE VIII
OPEN-SOURCE DATASET UTILIZED BY MoE IN NLP, VISION, MULTIMODAL, AND WIRELESS NETWORKS

| Task Domain | Dataset | Description | Composition | Application |
|----------------------------------|------------------------|--|---|---|
| NLP | | | | |
| Language Processing | GLUE [113] | A language understanding benchmark across tasks such as sentiment analysis, paraphrase detection, and natural language inference | 857,000 samples of sentences or sentence pairs and 9 NLP tasks | LiteMoE [253] |
| | C4 [114] | A large-scale corpus derived from web pages and filtered for cleaned content | 750GB of clean English text | MC-MoE LLM [254] |
| Question and Answer | TeleQnA [115] | Knowledge of LLM in lexicon, research overview, publications, standards overview, and specifications | 10,000 questions distributed across 5 categories | ETR-MoE [255] |
| Computer Vision | | | | |
| Common Image | ImageNet [116] | Large-scale image dataset organized by the WordNet hierarchy for visual recognition tasks | 3.2 million images, 5,247 categories, 500-1,000 images per category | Hard MoE [257] |
| Vehicular Image | nuScenes [117] | Autonomous driving dataset including camera, lidar, and radar, annotated for 3D object detection and tracking | 1.4 million images, 400,000 LiDAR sweeps, 1.3 million radar frames, 23 object classes | LiMoE [259] |
| Aerial Image | DOTA [118] | A large-scale aerial image dataset with oriented object annotations for real-world remote sensing applications | 2806 images, 188,000 objects, 15 classes | SM3Det [258] |
| Multimodal | | | | |
| Image-Text | MS COCO [119] | Natural images with human-written captions and detailed visual annotations including detection boxes, masks, and key points | 330,000 images, 5 captions per image, and annotations for 80 object categories | LIMOE [126] |
| | LAION-5B [120] | A large-scale dataset includes image-text pairs in multiple languages, filtered based on CLIP similarity scores | 5.85 billion image-text pairs | RAPHAEL [260] |
| Vedio-Text | HowTo100M [121] | Video-text data and ASR transcripts collected from narrated YouTube instructional videos | 1.22 million videos, 136 million clip-caption pairs, 23,000+ task types | Automatic speech recognition [261] |
| Wireless Networks | | | | |
| Physical Layer Signal Processing | RadioML 2018.01a [122] | Open dataset for automatic modulation classification in wireless signals | 24 modulation schemes, 26 SNR levels, 1024-sample IQ data per instance | MOE-AMC [178] |
| Wireless Channel | DeepMIMO [123] | A configurable ray-tracing-based dataset for deep learning in mmWave and massive MIMO systems | 12 base stations, 497,931 users, 1 blockage surface, and 2 reflector surfaces | MoE-based mmWave beam selection [262] |
| 5G Network | 5G NIDD [113] | A labeled intrusion detection dataset collected from a real 5G test network, supporting ML-based security research | 1.2 million flows, 9 attack types, and 1 benign traffic | MoE-based intrusion detection [111] |
| Vehicular Network | CitySim [124] | A high-precision vehicle trajectory dataset over diverse road scenarios, designed for intelligent Internet of Vehicle research. | 25,000+ vehicle trajectories covering highway, urban, and intersection scenes | MoE-based vehicle trajectory prediction [181] |

provides realistic mmWave and massive MIMO channels tailored to beam selection, channel estimation, and large-antenna array optimizations [262]. For network security, 5G-NIDD presents comprehensive network intrusion detection logs with fully labeled traffic traces derived from operational real-world 5G testbeds. This dataset allows MoE-based intrusion detection systems to differentiate between benign connections, newly emerging threats, and zero-day exploits, reducing false alarms and improving detection agility [111]. Additionally, the CitySim dataset focuses on connected and autonomous vehicles, delivering drone-captured vehicular trajectories that include complex mobility patterns, thereby supporting safety-oriented and MoE-based wireless research for connected and intelligent transportation systems [181].

Datasets for wireless networks such as RadioML [122], DeepMIMO [123], 5G-NIDD [250], CitySim [124], W ViWi-Drone [264] and DroneRF [265] serve diverse wireless scenarios ranging from physical-layer signal processing to system-level intrusion detection, vehicular networking, and

UAV-assisted communications. In particular, RadioML offers labeled in-phase/quadrature (IQ) signal recordings, which are utilized by MoE-based automatic modulation classification (AMC) models to capture distinctive signal features under varying SINR scenarios [178]. DeepMIMO, built on accurate ray-tracing simulations, provides realistic mmWave and massive MIMO channels tailored to beam selection, channel estimation, and large-antenna array optimizations [262]. For network security, 5G-NIDD presents comprehensive intrusion detection logs with fully labeled traffic traces derived from operational real-world 5G testbeds. This dataset allows MoE-based systems to differentiate between benign connections, newly emerging threats, and zero-day exploits, thus reducing false alarms and improving detection agility [111]. CitySim focuses on connected and autonomous vehicles, delivering drone-captured vehicular trajectories with complex mobility patterns to support safety-oriented and MoE-based research for intelligent transportation systems [181]. Additionally, ViWi-Drone and DroneRF provide realistic air-to-ground

and UAV-to-UAV channel measurements, enabling MoE-based models to address dynamic beam prediction, link adaptation, and mobility-aware resource allocation in low-altitude network scenarios [266].

VI. FUTURE RESEARCH DIRECTIONS

MoE has demonstrated promising advances across various wireless scenarios. Nonetheless, several open research directions remain to be explored for fully utilizing its capabilities. This section outlines key future directions, categorized into improving existing solutions and exploring new methodologies to extend MoE's capabilities in intelligent wireless systems.

A. Improving Existing Solutions

1) *Lightweight and Resource-Efficient MoE Architectures:*

While MoE frameworks offer substantial capacity and adaptability, the intensive computation and memory demands remain challenging for edge devices, such as IoT nodes and UAVs that are power or storage constrained. To overcome these limitations, parameter-reduction strategies, such as expert pruning [267], quantization-aware training [268], and knowledge distillation [269], present promising solutions to selectively retain critical expert network weights, thereby reducing the computational overhead. In addition, collaborative model execution across edge devices and cloud servers offers a compelling direction, where lightweight expert modules are deployed locally for low-latency processing, and more complex components are offloaded to centralized infrastructure.

2) *Dynamic Gating Mechanisms:* Dynamic gating represents a promising direction for enhancing the adaptability and efficiency of MoE frameworks in wireless communication systems. Future developments may focus on context-aware gating, where expert selection is conditioned on environmental variables such as user location, mobility patterns, spectrum availability, and signal-to-noise ratio, enabling adaptive prioritization of experts under congestion or energy constraints [270]. Reinforcement learning-based gating can further optimize expert activation by dynamically adjusting policies according to real-time feedback from network performance metrics such as throughput, latency, and energy consumption [271]. In addition, multi-task and multi-modal adaptive gating can facilitate shared experts across heterogeneous tasks and modalities while maintaining specialization [272]. For example, in ISAC scenarios, expert participation can be adapted depending on whether the task involves spectrum allocation, beamforming, or target detection. Finally, meta-gating and other self-evolving architectures, empowered by meta-learning, allow the gating policy itself to adapt efficiently to new operational contexts without retraining from scratch [66]. Collectively, these techniques can yield more intelligent and responsive MoE systems, particularly under the stringent performance and adaptability requirements anticipated in 6G and beyond.

3) *Communication-Efficient MoE in Distributed Wireless Networks:* Real-time processing is a critical requirement for distributed MoE systems deployed in time-sensitive applications such as vehicular networks and ISAC systems.

However, the inherent latency of inter-agent coordination, expert selection, and output aggregation poses significant challenges to meeting stringent timing requirements. Emerging 6G technologies offer promising solutions to mitigate these constraints. Terahertz (THz) communications and extremely-large scale MIMO (XL-MIMO) provide ultra-high data rates and ultra-low transmission delays, enabling near-instantaneous exchange of gating signals and expert outputs. Cell-free massive MIMO eliminates the need for frequent handovers and supports seamless expert migration across distributed nodes, thereby reducing coordination delays in highly dynamic network topologies. Furthermore, advanced ISAC-enhancement techniques, such as joint beamforming for sensing and communication, wideband sensing, and predictive channel estimation, improve real-time responsiveness by delivering rapid environmental awareness and enabling context-driven expert activation. Leveraging these technologies allows distributed MoE frameworks to reduce expert selection latency, accelerate inter-agent communication, and deliver inference results within the strict quality requirements of 6G-enabled mission-critical services.

4) *Multimodal MoE for Integrated Wireless Networks:*

While MoE frameworks have exhibited remarkable performance in multimodal learning tasks involving image-text or video-text data, their application in wireless networks has predominantly remained within unimodal scenarios, typically confined to tasks such as radio frequency signal processing [178]. By incorporating auxiliary data sources such as environmental imagery, inertial sensor measurements, or high-level semantic user intents, expert networks can be designed to extract complementary features across distinct data patterns. In parallel, adaptive gating mechanisms can enable expert selection based on different modality availability, thereby supporting more accurate channel estimation, adaptive beamforming, and interference suppression. Wireless multimodal synthesis holds significant potential to enhance situational awareness and advance the development of communication networks.

B. Exploring New Methodologies

1) *Cross-Layer Coordination and Optimization:* Current utilization of MoE in wireless protocol stacks conventionally focuses on isolated protocol layers, such as physical-layer communication or MAC-layer access control. A promising research direction lies in developing cross-layer MoE architectures that jointly optimize different layer protocol tasks, including link adaptation, congestion control, traffic scheduling, and routing strategy, within a unified decision-making framework. Such cross-layer integration facilitates holistic optimization and ensures that decisions made at individual layers align with global system goals, enhancing overall performance such as spectral efficiency, latency reduction, and reliability.

2) *Hardware Acceleration and Deployment of MoE:* Beyond the domain of algorithmic design, advancing MoE for wireless applications requires addressing challenges related to hardware compatibility and system-level integration. Exploring co-optimized hardware architectures, such as AI accelerators with embedded MoE gating logic [107]

or CPU/GPU design pipelines tailored for sparse computation [69], provides efficient support for expert selection and conditional MoE execution. Scalable deployment of MoE under such conditions demands intelligent runtime orchestration frameworks capable of adapting to heterogeneous hardware capabilities, dynamic network topologies, and evolving wireless infrastructures.

3) Novel Paradigms in IoT, Digital Twins, and Beyond:

As wireless networks evolve toward emerging paradigms such as digital twins, cognitive IoT, and zero-touch network management [273], MoE frameworks are envisioned to serve as a fundamental approach for dynamic adaptability and low-latency responsiveness. By tailoring gating mechanisms to contextual factors such as user mobility, service-level agreements, and energy-efficiency constraints, next-generation wireless systems can more effectively coordinate heterogeneous operational objectives with minimal resource overhead, thereby advancing toward context-aware and data-driven network intelligence.

VII. CONCLUSION

This survey has comprehensively reviewed the integration of MoE in wireless networks, demonstrating its substantial advantage in enhancing adaptability and improving efficiency across various wireless scenarios. By systematically exploring fundamental MoE methodologies, including various gating mechanisms and their integration with GenAI and RL, the survey has extensively explored applications across key wireless domains such as vehicular networks, UAVs, satellite networks, HetNets, ISAC, and mobile edge computing. Furthermore, critical wireless network tasks, including physical layer communications, radio resource management, network optimization, and security, have been thoroughly discussed. A case study integrating MoE into a diffusion-based DRL framework has revealed its empirical advantages in wireless optimization tasks. Additionally, an overview of open-source datasets has been presented to support ongoing MoE-related experimentation. Future directions should focus on lightweight architectures, balanced expert allocation, and integration with emerging paradigms, facilitating MoE as a pivotal component for next-generation wireless systems.

REFERENCES

- [1] Z. Chen, Z. Zhang, and Z. Yang, "Big AI models for 6G wireless networks: Opportunities, challenges, and research directions," *IEEE Wireless Commun.*, vol. 31, no. 5, pp. 164–172, Oct. 2024.
- [2] J. Achiam et al., "GPT-4 technical report," 2023, *arXiv:2303.08774*.
- [3] Y. Sun et al., "LOSEC: Local semantic capture empowered large time series model for IoT-enabled data centers," *IEEE Internet Things J.*, vol. 12, no. 10, pp. 13144–13156, May 2025.
- [4] O. Friha, M. A. Ferrag, B. Kantarci, B. Cakmak, A. Ozgun, and N. Ghoulmi-Zine, "LLM-based edge intelligence: A comprehensive survey on architectures, applications, security and trustworthiness," *IEEE Open J. Commun. Soc.*, vol. 5, pp. 5799–5856, 2024.
- [5] Y. Wang, Z. Gao, D. Zheng, S. Chen, D. Gündüz, and H. V. Poor, "Transformer-empowered 6G intelligent networks: From massive MIMO processing to semantic communication," *IEEE Wireless Commun.*, vol. 30, no. 6, pp. 127–135, Dec. 2023.
- [6] T. Zhang et al., "Handover-free multi-connectivity mobility management for downlink FD-RAN: A hierarchical DRL-based approach," *IEEE Trans. Cognit. Commun. Netw.*, vol. 11, no. 2, pp. 1281–1296, Apr. 2025.
- [7] M. Xu et al., "Unleashing the power of edge-cloud generative AI in mobile networks: A survey of AIGC services," *IEEE Commun. Surveys Tuts.*, vol. 26, no. 2, pp. 1127–1170, 2nd Quart., 2024.
- [8] B. Liu, X. Liu, S. Gao, X. Cheng, and L. Yang, "LLM4CP: Adapting large language models for channel prediction," 2024, *arXiv:2406.14440*.
- [9] V. T. Kunde, V. Rajagopalan, C. S. K. Valmееkam, K. Narayanan, S. Shakkottai, D. Kalathil, and J.-F. Chamberland, "Transformers are provably optimal in-context estimators for wireless communications," 2023, *arXiv:2311.00226*.
- [10] W. Lee and J. Park, "LLM-empowered resource allocation in wireless communications systems," 2024, *arXiv:2408.02944*.
- [11] L. Bariah, Q. Zhao, H. Zou, Y. Tian, F. Bader, and M. Debbah, "Large generative AI models for telecom: The next big thing?" *IEEE Commun. Mag.*, vol. 62, no. 11, pp. 84–90, Nov. 2024.
- [12] M. Xu et al., "When large language model agents meet 6G networks: Perception, grounding, and alignment," *IEEE Wireless Commun.*, vol. 31, no. 6, pp. 63–71, Dec. 2024.
- [13] K. Qiu, S. Bakirtzis, I. Wassell, H. Song, J. Zhang, and K. Wang, "Large language model-based wireless network design," *IEEE Wireless Commun. Lett.*, vol. 13, no. 12, pp. 3340–3344, Dec. 2024.
- [14] H. Touvron et al., "LLaMA: Open and efficient foundation language models," 2023, *arXiv:2302.13971*.
- [15] A. Chowdhery et al., "PaLM: Scaling language modeling with pathways," *J. Mach. Learn. Res.*, vol. 24, no. 240, pp. 1–113, 2023.
- [16] J. A. Baktash and M. Dawodi, "GPT-4: A review on advancements and opportunities in natural language processing," 2023, *arXiv:2305.03195*.
- [17] G. Qu, Q. Chen, W. Wei, Z. Lin, X. Chen, and K. Huang, "Mobile edge intelligence for large language models: A contemporary survey," 2024, *arXiv:2407.18921*.
- [18] M. U. Hadi, q. a. tashi, R. Qureshi, A. Shah, a. muneer, M. Irfan, A. Zafar, M. B. Shaikh, N. Akhtar, J. Wu, and S. Mirjalili, "Large language models: A comprehensive survey of its applications, challenges, limitations, and future prospects," *TechRxiv Preprint*, Nov. 2023. [Online]. Available: <http://dx.doi.org/10.36227/techrxiv.23589741.v4>
- [19] J. Stojkovic, E. Choukse, C. Zhang, I. Goiri, and J. Torrellas, "Towards greener LLMs: Bringing energy-efficiency to the forefront of LLM inference," 2024, *arXiv:2403.20306*.
- [20] H. Zhou et al., "Large language models for wireless networks: An overview from the prompt engineering perspective," 2024, *arXiv:2411.04136*.
- [21] H. Zhou et al., "Large language model (LLM) for telecommunications: A comprehensive survey on principles, key techniques, and opportunities," *IEEE Commun. Surveys Tuts.*, vol. 27, no. 3, pp. 1955–2005, 4th Quart., 2025.
- [22] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4510–4520.
- [23] X. Jiao et al., "TinyBERT: Distilling BERT for natural language understanding," 2019, *arXiv:1909.10351*.
- [24] Y. Shen et al., "Large language models empowered autonomous edge AI for connected intelligence," *IEEE Commun. Mag.*, vol. 62, no. 10, pp. 140–146, Oct. 2024.
- [25] Z. Lin et al., "Efficient parallel split learning over resource-constrained wireless edge networks," *IEEE Trans. Mobile Comput.*, vol. 23, no. 10, pp. 9224–9239, Oct. 2024.
- [26] T. Schuster et al., "Confident adaptive language modeling," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2022, pp. 17456–17472.
- [27] S. E. Yuksel, J. N. Wilson, and P. D. Gader, "Twenty years of mixture of experts," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 8, pp. 1177–1193, Aug. 2012.
- [28] A. Vats, R. Raja, V. Jain, and A. Chadha, "The evolution of mixture of experts: A survey from basics to breakthroughs," *Preprints*, Aug. 2024. [Online]. Available: <https://doi.org/10.20944/preprints202408.0583.v2>
- [29] S. Masoudnia and R. Ebrahimipour, "Mixture of experts: A literature survey," *Artif. Intell. Rev.*, vol. 42, no. 2, pp. 275–293, Aug. 2014.
- [30] J. Liu et al., "A survey on inference optimization techniques for mixture of experts models," 2024, *arXiv:2412.14219*.
- [31] T. R. McIntosh, T. Susnjak, T. Liu, P. Watters, and M. N. Halgamuge, "From Google Gemini to OpenAI Q*(Q-star): A survey of reshaping the generative artificial intelligence (AI) research landscape," 2023, *arXiv:2312.10868*.
- [32] W. Cai, J. Jiang, F. Wang, J. Tang, S. Kim, and J. Huang, "A survey on mixture of experts in large language models," 2024, *arXiv:2407.06204*.

- [33] Z. Wan et al., "Efficient large language models: A survey," 2023, *arXiv:2312.03863*.
- [34] N. Van Huynh et al., "Generative AI for physical layer communications: A survey," *IEEE Trans. Cognit. Commun. Netw.*, vol. 10, no. 3, pp. 706–728, Jun. 2024.
- [35] Y. Cao et al., "A comprehensive survey of AI-generated content (AIGC): A history of generative AI from GAN to ChatGPT," 2023, *arXiv:2303.04226*.
- [36] M. Xu et al., "Integration of mixture of experts and multimodal generative AI in Internet of Vehicles: A survey," 2024, *arXiv:2404.16356*.
- [37] R. Collobert, S. Bengio, and Y. Bengio, "A parallel mixture of SVMs for very large scale problems," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 14, 2002, pp. 1105–1114.
- [38] X. Zhang, Y. Shen, Z. Huang, J. Zhou, W. Rong, and Z. Xiong, "Mixture of attention heads: Selecting attention heads per token," 2022, *arXiv:2210.05144*.
- [39] Y. Shen, Z. Guo, T. Cai, and Z. Qin, "JetMoE: Reaching Llama2 performance with 0.1M dollars," 2024, *arXiv:2404.07413*.
- [40] M. Li et al., "Branch-Train-merge: Embarrassingly parallel training of expert language models," 2022, *arXiv:2208.03306*.
- [41] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton, "Adaptive mixtures of local experts," *Neural Comput.*, vol. 3, no. 1, pp. 79–87, Mar. 1991.
- [42] M. I. Jordan and R. A. Jacobs, "Hierarchical mixtures of experts and the EM algorithm," *Neural Comput.*, vol. 6, no. 2, pp. 181–214, Mar. 1994.
- [43] S. Dou et al., "LoRAMoE: Alleviate world knowledge forgetting in large language models via MoE-style plugin," 2023, *arXiv:2312.09979*.
- [44] X. Wu, S. Huang, and F. Wei, "Mixture of LoRA experts," 2024, *arXiv:2404.13628*.
- [45] B. Pan et al., "Dense training, sparse inference: Rethinking training of mixture-of-experts language models," 2024, *arXiv:2404.05567*.
- [46] D. Eigen, M. Ranzato, and I. Sutskever, "Learning factored representations in a deep mixture of experts," 2013, *arXiv:1312.4314*.
- [47] N. Shazeer et al., "Outrageously large neural networks: The sparsely-gated mixture-of-experts layer," 2017, *arXiv:1701.06538*.
- [48] N. Du et al., "GLaM: Efficient scaling of language models with mixture-of-experts," in *Proc. Int. Conf. Mach. Learn.*, vol. 162, 2022, pp. 5547–5569.
- [49] D. Lepikhin et al., "GShard: Scaling giant models with conditional computation and automatic sharding," 2020, *arXiv:2006.16668*.
- [50] W. Fedus, B. Zoph, and N. Shazeer, "Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity," *J. Mach. Learn. Res.*, vol. 23, no. 120, pp. 1–39, 2022.
- [51] C. Riquelme et al., "Scaling vision with sparse mixture of experts," in *Proc. NIPS*, 2021, pp. 8583–8595.
- [52] H. Nguyen, X. Han, C. Harris, S. Saria, and N. Ho, "On expert estimation in hierarchical mixture of experts: Beyond softmax gating functions," 2024, *arXiv:2410.02935*.
- [53] K. Li and J. Xu, "AC-MMOE: A multi-gate mixture-of-experts model based on attention and convolution," *Proc. Comput. Sci.*, vol. 222, pp. 187–196, Jan. 2023.
- [54] W. Li et al., "Hierarchical mixture of experts: Generalizable learning for high-level synthesis," 2024, *arXiv:2410.19225*.
- [55] J. Ma, Z. Zhao, X. Yi, J. Chen, L. Hong, and E. H. Chi, "Modeling task relationships in multi-task learning with multi-gate mixture-of-experts," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2018, pp. 1930–1939.
- [56] Y. Chai, Q. Yin, and J. Zhang, "Improved training of mixture-of-experts language GANs," in *Proc. ICASSP - IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2023, pp. 1–5.
- [57] D. Keetae Park, S. Yoo, H. Bahng, J. Choo, and N. Park, "MEGAN: Mixture of experts of generative adversarial networks for multimodal image generation," 2018, *arXiv:1805.02481*.
- [58] J. Zhu, C. Yang, K. Zheng, Y. Xu, Z. Shi, and Y. Shen, "Exploring sparse MoE in GANs for text-conditioned image synthesis," 2023, *arXiv:2309.03904*.
- [59] Y. Shi, N. Siddharth, B. Paige, and P. H. S. Torr, "Variational mixture-of-experts autoencoders for multi-modal deep generative models," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 32, 2019, pp. 15692–15703.
- [60] Q. Yu, M. S. Kavitha, and T. Kurita, "Mixture of experts with convolutional and variational autoencoders for anomaly detection," *Appl. Intell.*, vol. 51, no. 6, pp. 3241–3254, Jun. 2021.
- [61] Y. M. Saidutta, A. Abdi, and F. Fekri, "Joint source-channel coding over additive noise analog channels using mixture of variational autoencoders," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 7, pp. 2000–2013, Jul. 2021.
- [62] J. Yi and Z. Chen, "Variational mixture of stochastic experts auto-encoder for multi-modal recommendation," *IEEE Trans. Multimedia*, vol. 26, pp. 8941–8954, 2024.
- [63] Y. Balaji et al., "EDiff-I: Text-to-image diffusion models with an ensemble of expert denoisers," 2022, *arXiv:2211.01324*.
- [64] A. Ganjdanesh, Y. Kang, Y. Liu, H. Zhang, Z. Lin, and H. Huang, "Mixture of efficient diffusion experts through automatic interval and sub-network selection," in *Proc. Eur. Conf. Comput. Vis.*, 2024, pp. 54–71.
- [65] B. Park, H. Go, J. Y. Kim, S. Woo, S. Ham, and C. Kim, "Switch diffusion transformer: Synergizing denoising tasks with sparse mixture-of-experts," in *Proc. Eur. Conf. Comput. Vis.*, 2024, pp. 461–477.
- [66] Z. Tao, Z. Chi, L. Gu, Y. Wang, Y. Yu, and J. Tang, "Meta-DMoE: Adapting to domain shift by meta-distillation from mixture-of-experts," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 22243–22257.
- [67] Y. Li et al., "Uni-MoE: Scaling unified multimodal LLMs with mixture of experts," 2024, *arXiv:2405.11273*.
- [68] X. Qu, D. Dong, X. Hu, T. Zhu, W. Sun, and Y. Cheng, "LLaMA-MoE v2: Exploring sparsity of LLaMA from perspective of mixture-of-experts with post-training," 2024, *arXiv:2411.15708*.
- [69] D. Dai et al., "DeepSeekMoE: Towards ultimate expert specialization in mixture-of-experts language models," 2024, *arXiv:2401.06066*.
- [70] M. Artetxe et al., "Efficient large scale language modeling with mixtures of experts," 2021, *arXiv:2112.10684*.
- [71] J. Obando-Ceron et al., "Mixtures of experts unlock parameter scaling for deep RL," 2024, *arXiv:2402.08609*.
- [72] Y. Tao and J. Doe, "Double deep Q-learning in opponent modeling," 2022, *arXiv:2211.15384*.
- [73] Z. Zheng et al., "Self-supervised mixture-of-experts by uncertainty estimation," in *Proc. AAAI Conf. Artif. Intell.*, 2019, vol. 33, no. 1, pp. 5933–5940.
- [74] M. Khamassi, L. Martinet, and A. Guillot, "Combining self-organizing maps with mixtures of experts: Application to an actor-critic model of reinforcement learning in the Basal Ganglia," in *Proc. Int. Conf. Simulation Adapt. Behav. (SAB)*. Cham, Switzerland: Springer, 2006, pp. 394–405.
- [75] T. Danket, S. Tanachutiwat, and V. Rungreunganun, "A mixture-of-experts approach to production capacity planning for diverse demand patterns via deep reinforcement learning," *Eng. Access*, vol. 10, no. 2, pp. 154–165, 2024.
- [76] S. Dey and G. Sharon, "Comparing deterministic and soft policy gradients for optimizing Gaussian mixture actors," *Trans. Mach. Learn. Res.*, Dec. 2024. [Online]. Available: <https://openreview.net/forum?id=qS9pPu8ODt>
- [77] J. Ren, Y. Li, Z. Ding, W. Pan, and H. Dong, "Probabilistic mixture-of-experts for efficient deep reinforcement learning," 2021, *arXiv:2104.09122*.
- [78] L. Meng, X. Zhang, D. Xing, and B. Xu, "A new pre-training paradigm for offline multi-agent reinforcement learning with suboptimal data," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2024, pp. 7520–7524.
- [79] R. Zhang et al., "Optimizing generative AI networking: A dual perspective with multi-agent systems and mixture of experts," 2024, *arXiv:2405.12472*.
- [80] H. Nguyen, B. Pham, H. Du, S. Thudumu, R. Vasa, and K. Mouzakis, "CSAOT: Cooperative multi-agent system for active object tracking," 2025, *arXiv:2501.13994*.
- [81] E. Triantafyllidis, F. Acero, Z. Liu, and Z. Li, "Hybrid hierarchical learning for solving complex sequential tasks using the robotic manipulation network ROMAN," *Nature Mach. Intell.*, vol. 5, no. 9, pp. 991–1005, Sep. 2023.
- [82] Q. Wang and H. Van Hoof, "Learning expressive meta-representations with mixture of expert neural processes," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 35, 2022, pp. 26242–26255.
- [83] G. Cheng, L. Dong, W. Cai, and C. Sun, "Multi-task reinforcement learning with attention-based mixture of experts," *IEEE Robot. Autom. Lett.*, vol. 8, no. 6, pp. 3812–3819, Jun. 2023.
- [84] J. Vyas, Bhumika, D. Das, and S. Chaudhury, "Federated learning based driver recommendation for next generation transportation system," *Expert Syst. Appl.*, vol. 225, Sep. 2023, Art. no. 119951.

- [85] F. Yao, C. Sun, B. Lu, B. Wang, and H. Yu, "Mixture of experts framework based on soft actor-critic algorithm for highway decision-making of connected and automated vehicles," *Chin. J. Mech. Eng.*, vol. 38, no. 1, p. 1, Jan. 2025.
- [86] E. Kawamura, D. Azimov, J. S. Allen, and C. Ippolito, "Hierarchical mixture of experts for autonomous unmanned aerial vehicles utilizing thrust models and acoustics," *Robot. Auto. Syst.*, vol. 162, Apr. 2023, Art. no. 104369.
- [87] F. Chen, P. Li, S. Pan, L. Zhong, and J. Deng, "Giant could be tiny: Efficient inference of giant models on resource-constrained UAVs," *IEEE Internet Things J.*, vol. 11, no. 12, pp. 21170–21179, Jun. 2024.
- [88] Y. J. Wong, M.-L. Tham, B.-H. Kwan, and A. Iqbal, "Addressing environmental stochasticity in reconfigurable intelligent surface aided unmanned aerial vehicle networks: Multi-task deep reinforcement learning based optimization for physical layer security," *Internet Things*, vol. 27, Oct. 2024, Art. no. 101270.
- [89] R. Zhang et al., "Generative AI agents with large language model for satellite networks via a mixture of experts transmission," *IEEE J. Sel. Areas Commun.*, vol. 42, no. 12, pp. 3581–3596, Dec. 2024.
- [90] D. G. Loyola, "Combining neural networks for the near-real-time processing of satellite data," in *Proc. 1st Int. IEEE Symp. Intell. Syst.*, vol. 1, Aug. 2002, pp. 233–237.
- [91] Z. Liu, X. Wang, C. Feng, X. Sun, W. Zhan, and X. Chen, "Meta-reinforcement learning with mixture of experts for generalizable multi access in heterogeneous wireless networks," 2024, *arXiv:2412.03850*.
- [92] J. Wang et al., "Optimizing 6G integrated sensing and communications (ISAC) via expert networks," 2024, *arXiv:2406.00408*.
- [93] X. Liu, T. Ratnarajah, M. Sellathurai, and Y. C. Eldar, "Multimodal learning for integrated sensing and communication networks," in *Proc. 32nd Eur. Signal Process. Conf. (EUSIPCO)*, Aug. 2024, pp. 1177–1181.
- [94] H. Li and L. Duan, "Theory of mixture-of-experts for mobile edge computing," 2024, *arXiv:2412.15690*.
- [95] H. M. S. B. Senevirathna and K. Yamashita, "Channel prediction for OFDMA using mixtures of experts," *Int. J. Knowl.-based Intell. Eng. Syst.*, vol. 10, no. 3, pp. 193–200, May 2006.
- [96] L. M. Lopez-Ramos, Y. Teganya, B. Beferull-Lozano, and S.-J. Kim, "Channel gain cartography via mixture of experts," in *Proc. IEEE Global Commun. Conf.*, Dec. 2020, pp. 1–7.
- [97] R. K. Jaiswal, M. Elnourani, S. Deshmukh, and B. Beferull-Lozano, "Leveraging transfer learning for radio map estimation via mixture of experts," *IEEE Trans. Cognit. Commun. Netw.*, early access, May 15, 2025, doi: [10.1109/TCCN.2025.3570469](https://doi.org/10.1109/TCCN.2025.3570469).
- [98] B. Van Bolderik, V. Menkovski, S. Heemstra, and M. D. Gomony, "MEAN: Mixture-of-experts based neural receiver," in *Proc. IFIP/IEEE 32nd Int. Conf. Very Large Scale Integr. (VLSI-SoC)*, Oct. 2024, pp. 1–4.
- [99] A. Fischer-Bühner, A. Brihuega, L. Anttila, M. D. Gomony, and M. Valkama, "Mixture of experts neural network for modeling of power amplifiers," in *IEEE MTT-S Int. Microw. Symp. Dig.*, Jun. 2022, pp. 510–513.
- [100] A. Fischer-Bühner et al., "Sparsely gated mixture of experts neural network for linearization of RF power amplifiers," *IEEE Trans. Microw. Theory Techn.*, vol. 72, no. 7, pp. 4367–4382, Jul. 2024.
- [101] M. Zecchin, D. Gesbert, and M. Kountouris, "Team deep mixture of experts for distributed power control," in *Proc. IEEE 21st Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, May 2020, pp. 1–5.
- [102] L. Ma, N. Cheng, X. Wang, R. Sun, and N. Lu, "On-demand resource management for 6G wireless networks using knowledge-assisted dynamic neural networks," in *Proc. IEEE Int. Conf. Commun.*, Seoul, South Korea, May 2022, pp. 1–6.
- [103] H. Du et al., "Mixture of experts for intelligent networks: A large language model-enabled approach," in *Proc. Int. Wireless Commun. Mobile Comput. (IWCMC)*, May 2024, pp. 531–536.
- [104] R. Chattopadhyay and C.-K. Tham, "Mixture of experts based model integration for traffic state prediction," in *Proc. IEEE 95th Veh. Technol. Conf. (VTC-Spring)*, Jun. 2022, pp. 1–7.
- [105] W. Jiang, J. Han, H. Liu, T. Tao, N. Tan, and H. Xiong, "Interpretable cascading mixture-of-experts for urban traffic congestion prediction," in *Proc. 30th ACM SIGKDD Conf. Knowl. Discovery Data Mining*, Aug. 2024, pp. 5206–5217.
- [106] H. Jiang, C. Hu, Y. Niu, B. Yang, H. Chen, and X. Zhang, "Hybrid attention-based multi-task vehicle motion prediction using non-autoregressive transformer and mixture of experts," *IEEE Trans. Intell. Vehicles*, early access, Dec. 26, 2024, doi: [10.1109/TIV.2024.3523318](https://doi.org/10.1109/TIV.2024.3523318).
- [107] R. Sarkar, H. Liang, Z. Fan, Z. Wang, and C. Hao, "Edge-MoE: Memory-efficient multi-task vision transformer architecture with task-level sparsity via mixture-of-experts," in *Proc. IEEE/ACM Int. Conf. Comput. Aided Design (ICCAD)*, Oct. 2023, pp. 01–09.
- [108] N. Xue et al., "WDMoE: Wireless distributed large language models with mixture of experts," 2024, *arXiv:2405.03131*.
- [109] X. Yuan, W. Kong, Z. Luo, and M. Xu, "Efficient inference offloading for mixture-of-experts large language models in Internet of Medical Things," *Electronics*, vol. 13, no. 11, p. 2077, May 2024.
- [110] X. Wang, Z. Zheng, Z. Fei, Z. Han, and Y. Huang, "Fighting against active eavesdropper: Distributed pilot spoofing attack detection and secure coordinated transmission in multi-cell massive MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 23, no. 9, pp. 11184–11198, Sep. 2024.
- [111] L. Ilias, G. Doukas, V. Lamprou, C. Ntanos, and D. Askounis, "Convolutional neural networks and mixture of experts for intrusion detection in 5G networks and beyond," 2024, *arXiv:2412.03483*.
- [112] C. Zhao et al., "Enhancing physical layer communication security through generative AI with mixture of experts," 2024, *arXiv:2405.04198*.
- [113] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, "GLUE: A multi-task benchmark and analysis platform for natural language understanding," 2018, *arXiv:1804.07461*.
- [114] C. Raffel et al., "Exploring the limits of transfer learning with a unified text-to-text transformer," *J. Mach. Learn. Res.*, vol. 21, no. 140, pp. 1–67, 2019.
- [115] A. Maatouk, F. Ayed, N. Piovesan, A. De Domenico, M. Debbah, and Z.-Q. Luo, "TeleQnA: A benchmark dataset to assess large language models telecommunications knowledge," 2023, *arXiv:2310.15051*.
- [116] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
- [117] H. Caesar et al., "nuScenes: A multimodal dataset for autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11621–11631.
- [118] G.-S. Xia et al., "DOTA: A large-scale dataset for object detection in aerial images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3974–3983.
- [119] T. Lin et al., "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.
- [120] C. Schuhmann et al., "LAION-5B: An open large-scale dataset for training next generation image-text models," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2022, pp. 25278–25294.
- [121] A. Miech, D. Zhukov, J.-B. Alayrac, M. Tapaswi, I. sLaptev, and J. Sivic, "HowTo100M: Learning a text-video embedding by watching hundred million narrated video clips," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2630–2640.
- [122] T. J. O'Shea, T. Roy, and T. C. Clancy, "Over-the-Air deep learning based radio signal classification," *IEEE J. Sel. Topics Signal Process.*, vol. 12, no. 1, pp. 168–179, Feb. 2018.
- [123] A. Alkhateeb, "DeepMIMO: A generic deep learning dataset for millimeter wave and massive MIMO applications," 2019, *arXiv:1902.06435*.
- [124] O. Zheng, M. Abdel-Aty, L. Yue, A. Abdelraouf, Z. Wang, and N. Mahmoud, "CitySim: A drone-based vehicle trajectory dataset for safety-oriented research and digital twins," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 2678, no. 4, pp. 606–621, Apr. 2024.
- [125] M. Xu et al., "A survey of resource-efficient LLM and multimodal foundation models," 2024, *arXiv:2401.08092*.
- [126] B. Mustafa, C. Riquelme, J. Puigcerver, R. Jenatton, and N. Houlsby, "Multimodal contrastive learning with LIMoE: The language-image mixture of experts," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2022, pp. 9564–9576.
- [127] R. Zhang et al., "Interactive AI with retrieval-augmented generation for next generation networking," *IEEE Netw.*, vol. 38, no. 6, pp. 414–424, Nov. 2024.
- [128] Y. Xu, B. Qian, K. Yu, T. Ma, L. Zhao, and H. Zhou, "Federated learning over fully-decoupled RAN architecture for two-tier computing acceleration," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 3, pp. 789–801, Mar. 2023.
- [129] X. Zhang, X. Qin, Z. Zhang, L. X. Cai, H. Zhou, and W. Zhuang, "RIS-aided MIMO downlink transmission for ultradense LEO satellite-terrestrial networks," *IEEE Internet Things J.*, vol. 12, no. 11, pp. 15304–15318, Jun. 2025.

- [130] B. Qian et al., "Leveraging dynamic Stackelberg pricing game for multi-mode spectrum sharing in 5G-VANET," *IEEE Trans. Veh. Technol.*, vol. 69, no. 6, pp. 6374–6387, Jun. 2020.
- [131] D. Zhang, M. Piao, T. Zhang, C. Chen, and H. Zhu, "New algorithm of multi-strategy channel allocation for edge computing," *AEU-Int. J. Electron. Commun.*, vol. 126, Nov. 2020, Art. no. 153372.
- [132] H. Hazimeh et al., "DSelect-k: Differentiable selection in the mixture of experts with applications to multi-task learning," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 34, 2021, pp. 29335–29347.
- [133] C. M. Bishop and M. Svensen, "Bayesian hierarchical mixtures of experts," 2012, *arXiv:1212.2447*.
- [134] R. Ebrahimpour, E. Kabir, and M. R. Yousefi, "Face detection using mixture of MLP experts," *Neural Process. Lett.*, vol. 26, no. 1, pp. 69–82, Jul. 2007.
- [135] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, vol. 139, 2021, pp. 8748–8763.
- [136] T. Wei et al., "Skywork-MoE: A deep dive into training techniques for mixture-of-experts language models," 2024, *arXiv:2406.06563*.
- [137] L. Wu, M. Liu, Y. Chen, D. Chen, X. Dai, and L. Yuan, "Residual mixture of experts," 2022, *arXiv:2204.09636*.
- [138] Y. Zhou et al., "Mixture-of-experts with expert choice routing," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2022, pp. 7103–7114.
- [139] S. Wang, Z. Chen, B. Li, K. He, M. Zhang, and J. Wang, "Scaling laws across model architectures: A comparative analysis of dense and MoE models in large language models," 2024, *arXiv:2410.05661*.
- [140] S. Rajbhandari et al., "DeepSpeed-MoE: Advancing mixture-of-experts inference and training to power next-generation AI scale," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 18332–18346.
- [141] B. Zoph et al., "ST-MoE: Designing stable and transferable sparse expert models," 2022, *arXiv:2202.08906*.
- [142] S. Shen et al., "Mixture-of-experts meets instruction tuning: A winning combination for large language models," 2023, *arXiv:2305.14705*.
- [143] F. Shu et al., "LLaVA-MoD: Making LLaVA tiny via MoE knowledge distillation," 2024, *arXiv:2408.15881*.
- [144] X. Hou, J. Wang, J. Du, C. Jiang, Y. Ren, and D. Niyato, "Lightweight federated learning over wireless edge networks," *IEEE Trans. Mobile Comput.*, vol. 25, no. 1, pp. 300–312, Jan. 2026.
- [145] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, "Generative adversarial networks: An overview," *IEEE Signal Process. Mag.*, vol. 35, no. 1, pp. 53–65, Jan. 2018.
- [146] S. Gurumurthy, R. K. Sarvadevabhatla, and R. V. Babu, "DeLiGAN: Generative adversarial networks for diverse and limited data," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 166–174.
- [147] Q. Hoang, T. D. Nguyen, T. Le, and D. Phung, "Multi-generator generative adversarial nets," 2017, *arXiv:1708.02556*.
- [148] T. Wu et al., "CDDM: Channel denoising diffusion models for wireless communications," in *Proc. IEEE Global Commun. Conf.*, Dec. 2023, pp. 7429–7434.
- [149] Z. Feng et al., "ERNIE-ViLG 2.0: Improving text-to-image diffusion model with knowledge-enhanced mixture-of-denoising-experts," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 10135–10145.
- [150] Y. Lee, J. Kim, H. Go, M. Jeong, S. Oh, and S. Choi, "Multi-architecture multi-expert diffusion models," in *Proc. AAAI Conf. Artif. Intell.*, 2024, vol. 38, no. 12, pp. 13427–13436.
- [151] J. Luo et al., "Staleness-centric optimizations for parallel diffusion MoE inference," 2024, *arXiv:2411.16786*.
- [152] E. Daxberger et al., "Mobile V-MoEs: Scaling down vision transformers via sparse mixture-of-experts," 2023, *arXiv:2309.04354*.
- [153] H. Xi, Z. Shao, D. A. Hensher, J. D. Nelson, H. Chen, and K. Wijayarathna, "A multi-task transformer with mixture-of-experts for personalized pe-riodic predictions of individual travel behavior in multimodal public transport," *Transp. Res. C, Emerg. Technol.*, vol. 179, Oct. 2025, Art. no. 105287. [Online]. Available: 10.1016/j.trc.2025.105287
- [154] Z. Song, O. Simeone, and B. Rajendran, "Neuromorphic in-context learning for energy-efficient MIMO symbol detection," in *Proc. IEEE 25th Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, Sep. 2024, pp. 1–5.
- [155] M. Zecchin, K. Yu, and O. Simeone, "Cell-free multi-user MIMO equalization via in-context learning," in *Proc. IEEE 25th Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, Sep. 2024, pp. 646–650.
- [156] F. Khoramnejad and E. Hossain, "Generative AI for the optimization of next-generation wireless networks: Basics, state-of-the-art, and open challenges," *IEEE Commun. Surveys Tuts.*, early access, Jan. 28, 2025, doi: 10.1109/COMST.2025.3535554.
- [157] G. Tesauro et al., "Temporal difference learning and TD-Gammon," *Commun. ACM*, vol. 38, no. 3, pp. 58–68, 1995.
- [158] C. J. Watkins and P. Dayan, "Q-learning," *Mach. Learn.*, vol. 8, nos. 3–4, pp. 279–292, 1992.
- [159] V. Mnih et al., "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.
- [160] T. P. Lillicrap et al., "Continuous control with deep reinforcement learning," 2015, *arXiv:1509.02971*.
- [161] S. Fujimoto, H. van Hoof, and D. Meger, "Addressing function approximation error in actor-critic methods," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 1587–1596.
- [162] R. J. Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," *Mach. Learn.*, vol. 8, nos. 3–4, pp. 229–256, May 1992.
- [163] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," 2017, *arXiv:1707.06347*.
- [164] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 1861–1870.
- [165] C. Liu and Y. Wang, "MDPo: Offline reinforcement learning based on mixture density policy network," in *Proc. Int. Conf. Generative Artif. Intell. Inf. Secur.*, May 2024, pp. 132–137.
- [166] D. Li, X. Li, J. Wang, and P. Li, "Video recommendation with multi-gate mixture of experts soft actor critic," in *Proc. 43rd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2020, pp. 1553–1556.
- [167] I. P. Gomes, C. Prenebida, and D. F. Wolf, "Multi-agent interaction-aware behavior intention prediction using graph mixture of experts attention network on urban roads," *Expert Syst. Appl.*, vol. 270, Apr. 2025, Art. no. 126485.
- [168] J. Cui et al., "Chatlaw: A multi-agent collaborative legal assistant with knowledge graph enhanced mixture-of-experts large language model," 2023, *arXiv:2306.16092*.
- [169] T. Willi, J. Obando-Ceron, J. Foerster, K. Dziugaite, and P. Samuel Castro, "Mixture of experts in a mixture of RL settings," 2024, *arXiv:2406.18420*.
- [170] D. Esteban, L. Roza, and D. G. Caldwell, "Hierarchical reinforcement learning for concurrent discovery of compound and composable policies," in *Proc. IEEE/RISJ Int. Conf. Intell. Robots Syst. (IROS)*, Nov. 2019, pp. 1818–1825.
- [171] Y. Chow, A. Tulepbergenov, O. Nachum, M. Ryu, M. Ghavamzadeh, and C. Boutilier, "A mixture-of-expert approach to RL-based dialogue management," 2022, *arXiv:2206.00059*.
- [172] C. Zhao et al., "Generative AI-enabled wireless communications for robust low-altitude economy networking," 2025, *arXiv:2502.18118*.
- [173] R. Zhang, K. Xiong, Y. Lu, P. Fan, D. W. K. Ng, and K. B. Letaief, "Energy efficiency maximization in RIS-assisted SWIPT networks with RSMAs: A PPO-based approach," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 5, pp. 1413–1430, May 2023.
- [174] A. A. Puspitasari, T. T. An, M. H. Alsharif, and B. M. Lee, "Emerging technologies for 6G communication networks: Machine learning approaches," *Sensors*, vol. 23, no. 18, p. 7709, Sep. 2023.
- [175] R. Zhang, K. Xiong, Y. Lu, B. Gao, P. Fan, and K. B. Letaief, "Joint coordinated beamforming and power splitting ratio optimization in MU-MISO SWIPT-enabled HetNets: A multi-agent DDQN-based approach," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 2, pp. 677–693, Feb. 2022.
- [176] A. Fan et al., "Beyond english-centric multilingual machine translation," *J. Mach. Learn. Res.*, vol. 22, no. 107, pp. 1–48, 2021.
- [177] H. Lei et al., "Alt-MoE: Multimodal alignment via alternating optimization of multi-directional moe with unimodal models," 2024, *arXiv:2409.05929*.
- [178] J. Gao, Q. Cao, and Y. Chen, "MoE-AMC: Enhancing automatic modulation classification performance using mixture-of-experts," 2023, *arXiv:2312.02298*.
- [179] Z. Zong et al., "MoVA: Adapting mixture of vision experts to multimodal context," 2024, *arXiv:2404.13046*.
- [180] F. Souza, J. Mendes, and R. Araújo, "A regularized mixture of linear experts for quality prediction in multimode and multiphase industrial processes," *Appl. Sci.*, vol. 11, no. 5, p. 2040, Feb. 2021.
- [181] R. Yuan, M. Abdel-Aty, Q. Xiang, Z. Wang, and X. Gu, "A temporal multi-gate mixture-of-experts approach for vehicle trajectory and driving intention prediction," *IEEE Trans. Intell. Vehicles*, vol. 9, no. 1, pp. 1204–1216, Jan. 2024.

- [182] H. Wu, H. Zhou, J. Zhao, Y. Xu, B. Qian, and X. Shen, "Deep learning enabled fine-grained path planning for connected vehicular networks," *IEEE Trans. Veh. Technol.*, vol. 71, no. 10, pp. 10303–10315, Oct. 2022.
- [183] Y. Xu, H. Zhou, T. Ma, J. Zhao, B. Qian, and X. Shen, "Leveraging multiagent learning for automated vehicles scheduling at unsignalized intersections," *IEEE Internet Things J.*, vol. 8, no. 14, pp. 11427–11439, Jul. 2021.
- [184] G. Zhao et al., "When autonomous vehicles meet accidents: A DT-enabled post-accident maintenance scheme," *IEEE Internet Things J.*, vol. 10, no. 24, pp. 22233–22245, Dec. 2023.
- [185] C. Stöcker, R. Bennett, F. Nex, M. Gerke, and J. Zevenbergen, "Review of the current state of UAV regulations," *Remote Sens.*, vol. 9, no. 5, p. 459, May 2017.
- [186] L. Gupta, R. Jain, and G. Vaszkun, "Survey of important issues in UAV communication networks," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 2, pp. 1123–1152, 2nd Quart., 2015.
- [187] J. Zheng, J. Zhang, and B. Ai, "UAV communications with WPT-aided cell-free massive MIMO systems," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 10, pp. 3114–3128, Oct. 2021.
- [188] B. Fraser, A. Perrusquía, D. Panagiotakopoulos, and W. Guo, "A deep mixture of experts network for drone trajectory intent classification and prediction using non-cooperative radar data," in *Proc. IEEE Symp. Ser. Comput. Intell. (SSCI)*, Dec. 2023, pp. 1–6.
- [189] X. Qin, T. Ma, Z. Tang, X. Zhang, H. Zhou, and L. Zhao, "Service-aware resource orchestration in ultra-dense LEO satellite-terrestrial integrated 6G: A service function chain approach," *IEEE Trans. Wireless Commun.*, vol. 22, no. 9, pp. 6003–6017, Sep. 2023.
- [190] X. Zhang et al., "Cybertwin-assisted mode selection in ultra-dense LEO integrated satellite-terrestrial network," *J. Commun. Inf. Netw.*, vol. 7, no. 4, pp. 360–374, Dec. 2022.
- [191] T. Ma, B. Qian, X. Qin, X. Liu, H. Zhou, and L. Zhao, "Satellite-terrestrial integrated 6G: An ultra-dense LEO networking management architecture," *IEEE Wireless Commun.*, vol. 31, no. 1, pp. 62–69, Feb. 2024.
- [192] B. Agarwal, M. A. Togou, M. Marco, and G.-M. Muntean, "A comprehensive survey on radio resource management in 5G HetNets: Current solutions, future trends and open issues," *IEEE Commun. Surveys Tuts.*, vol. 24, no. 4, pp. 2495–2534, 4th Quart., 2022.
- [193] R. Zhang, K. Xiong, Y. Lu, D. W. K. Ng, P. Fan, and K. B. Letaief, "SWIPT-enabled cell-free massive MIMO-NOMA networks: A machine learning-based approach," *IEEE Trans. Wireless Commun.*, vol. 23, no. 7, pp. 6701–6718, Jul. 2024.
- [194] Z. Liu, L. Huang, Z. Gao, M. Luo, S. Hosseinipour, and H. Dai, "GA-DRL: Graph neural network-augmented deep reinforcement learning for DAG task scheduling over dynamic vehicular clouds," *IEEE Trans. Netw. Service Manage.*, vol. 21, no. 4, pp. 4226–4242, Aug. 2024.
- [195] Y. Hui et al., "Digital twins enabled on-demand matching for multi-task federated learning in HetVNs," *IEEE Trans. Veh. Technol.*, vol. 72, no. 2, pp. 2352–2364, Feb. 2023.
- [196] L. Jiao, J. Zhao, Y. Xu, T. Zhang, H. Zhou, and D. Zhao, "Performance analysis for downlink transmission in multiconnectivity cellular V2X networks," *IEEE Internet Things J.*, vol. 11, no. 7, pp. 11812–11824, Apr. 2024.
- [197] X. Zhu et al., "Enabling intelligent connectivity: A survey of secure ISAC in 6G networks," *IEEE Commun. Surveys Tuts.*, vol. 27, no. 2, pp. 748–781, 2nd Quart., 2025.
- [198] M. Mehrabi, D. You, V. Latzko, H. Salah, M. Reisslein, and F. H. P. Fitzek, "Device-enhanced MEC: Multi-access edge computing (MEC) aided by end device computation and caching: A survey," *IEEE Access*, vol. 7, pp. 166079–166108, 2019.
- [199] Y. Lin et al., "A model training mechanism based on onchain and offchain collaboration for edge computing," in *Proc. IEEE Int. Conf. Commun.*, Jun. 2021, pp. 1–6.
- [200] R. Yi, L. Guo, S. Wei, A. Zhou, S. Wang, and M. Xu, "EdgeMoE: Empowering sparse large language models on mobile devices," 2023, *arXiv:2308.14352*.
- [201] Y. Xu, H. Zhou, J. Chen, T. Ma, and S. Shen, "Cybertwin assisted wireless asynchronous federated learning mechanism for edge computing," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2021, pp. 1–6.
- [202] X. Wang et al., "Empowering edge intelligence: A comprehensive survey on on-device AI models," *ACM Comput. Surv.*, vol. 57, no. 9, pp. 1–39, Sep. 2025.
- [203] Y. Wang and J. Zhao, "A unified and resource-aware framework for adaptive inference acceleration on edge and embedded platforms," *Electronics*, vol. 14, no. 11, p. 2188, May 2025.
- [204] Y. Zhang et al., "A survey on privacy in graph neural networks: Attacks, preservation, and applications," *IEEE Trans. Knowl. Data Eng.*, vol. 36, no. 12, pp. 7497–7515, Dec. 2024.
- [205] Y. Su, N. Yan, Y. Deng, and R. Schober, "PWC-MoE: Privacy-aware wireless collaborative mixture of experts," 2025, *arXiv:2505.08719*.
- [206] X. Hou, J. Wang, Z. Zhang, J. Wang, L. Liu, and Y. Ren, "Split federated learning for UAV-enabled integrated sensing, computation, and communication," 2025, *arXiv:2504.01443*.
- [207] C. Ma et al., "Trusted AI in multiagent systems: An overview of privacy and security for distributed learning," *Proc. IEEE*, vol. 111, no. 9, pp. 1097–1132, Sep. 2023.
- [208] N. M. Al-Maslmani, M. Abdallah, and B. S. Ciftler, "Reputation-aware multi-agent DRL for secure hierarchical federated learning in IoT," *IEEE Open J. Commun. Soc.*, vol. 4, pp. 1274–1284, 2023.
- [209] L. Liu, C. Huang, D. Zhu, D. Liu, J. Ni, and X. Shen, "Enabling efficient and distributed access control for pervasive edge computing services," *IEEE Trans. Mobile Comput.*, vol. 23, no. 12, pp. 11342–11356, Dec. 2024.
- [210] X. Wang, A. Shankar, K. Li, B. D. Parameshachari, and J. Lv, "Blockchain-enabled decentralized edge intelligence for trustworthy 6G consumer electronics," *IEEE Trans. Consum. Electron.*, vol. 70, no. 1, pp. 1214–1225, Feb. 2024.
- [211] W. Zhu et al., "When MoE meets blockchain: A trustworthy distributed framework of large models," 2025, *arXiv:2509.12141*.
- [212] X. Wang, B. Wang, Y. Wu, Z. Ning, S. Guo, and F. R. Yu, "A survey on trustworthy edge intelligence: From security and reliability to transparency and sustainability," *IEEE Commun. Surveys Tuts.*, vol. 27, no. 3, pp. 1729–1757, 3rd Quart., 2025.
- [213] Y. Jiang, V. S. F. Garnot, K. Schindler, and J. D. Wegner, "Uncertainty voting ensemble for imbalanced deep regression," in *DAGM German Conference on Pattern Recognition*. Springer, 2024, pp. 329–343.
- [214] J. He et al., "Toward mixture-of-experts enabled trustworthy semantic communication for 6G networks," *IEEE Netw.*, vol. 39, no. 5, pp. 221–230, Sep. 2025.
- [215] Z. Aljabri and J. H. Abawajy, "Permutation-based firmware remote attestation for Internet-of-Things edge-based network," *IEEE Syst. J.*, vol. 19, no. 2, pp. 346–357, Jun. 2025.
- [216] B. Fesl, N. Turan, M. Joham, and W. Utschick, "Learning a Gaussian mixture model from imperfect training data for robust channel estimation," *IEEE Wireless Commun. Lett.*, vol. 12, no. 6, pp. 1066–1070, Jun. 2023.
- [217] Y. Liu, Z. Tan, H. Hu, L. J. Cimini, and G. Y. Li, "Channel estimation for OFDM," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 4, pp. 1891–1908, 4th Quart., 2014.
- [218] H. M. S. B. Senevirathna, K. Yamashitha, and H. Lin, "Self organizing map based channel prediction for OFDMA," in *Proc. IEEE Int. Symp. Circuits Syst.*, Mar. 2005, pp. 2506–2509.
- [219] L. Haihan, Y. Li, S. Zhou, and W. Jing, "A novel method to obtain CSI based on Gaussian mixture model and expectation maximization," in *Proc. 8th Int. Conf. Wireless Commun. Signal Process. (WCSP)*, Oct. 2016, pp. 1–5.
- [220] S. Dehuri and S.-B. Cho, "A comprehensive survey on functional link neural networks and an adaptive PSO–BP learning for CFLNN," *Neural Comput. Appl.*, vol. 19, no. 2, pp. 187–205, Mar. 2010.
- [221] V. Vovk, "Kernel ridge regression," in *Empirical Inference: Festschrift Honor Vladimir N. Vapnik*. Berlin, Germany: Springer, 2013, pp. 105–116.
- [222] Y. Xu and W. Yin, "A globally convergent algorithm for nonconvex optimization based on block coordinate update," *J. Sci. Comput.*, vol. 72, no. 2, pp. 700–734, Aug. 2017.
- [223] F. A. Aoudia and J. Hoydis, "End-to-end learning for OFDM: From neural receivers to pilotless communication," *IEEE Trans. Wireless Commun.*, vol. 21, no. 2, pp. 1049–1063, Feb. 2022.
- [224] J. Hoydis et al., "Sionna: An open-source library for next-generation physical layer research," 2022, *arXiv:2203.11854*.
- [225] B. Qian et al., "Enabling fully-decoupled radio access with elastic resource allocation," *IEEE Trans. Cognit. Commun. Netw.*, vol. 9, no. 4, pp. 1025–1040, Aug. 2023.
- [226] J. Liu, J. Chen, Z. Liu, and H. Zhou, "Enabling feedback-free MIMO transmission for FD-RAN: A data-driven approach," *IEEE Trans. Mobile Comput.*, vol. 24, no. 3, pp. 2437–2454, Mar. 2025.

- [227] R. Zhang et al., "Generative AI for space-air-ground integrated networks," *IEEE Wireless Commun.*, vol. 31, no. 6, pp. 10–20, Dec. 2024.
- [228] B. Lai et al., "Resource-efficient generative mobile edge networks in 6G era: Fundamentals, framework and case study," *IEEE Wireless Commun.*, vol. 31, no. 4, pp. 66–74, Aug. 2024.
- [229] K. He, L. He, L. Fan, Y. Deng, G. K. Karagiannidis, and A. Nallanathan, "Learning-based signal detection for MIMO systems with unknown noise statistics," *IEEE Trans. Commun.*, vol. 69, no. 5, pp. 3025–3038, May 2021.
- [230] J. Xue, K. Yu, T. Zhang, H. Zhou, L. Zhao, and X. Shen, "Cooperative deep reinforcement learning enabled power allocation for packet duplication URLLC in multi-connectivity vehicular networks," *IEEE Trans. Mobile Comput.*, vol. 23, no. 8, pp. 8143–8157, Aug. 2024.
- [231] H. Du et al., "The age of generative AI and AI-generated everything," *IEEE Netw.*, vol. 38, no. 6, pp. 501–512, Nov. 2024.
- [232] Z. Ke, H. Duan, and S. Qian, "Interpretable mixture of experts for time series prediction under recurrent and non-recurrent conditions," 2024, *arXiv:2409.03282*.
- [233] Q. Sun et al., "Generalizing motion planners with mixture of experts for autonomous driving," 2024, *arXiv:2410.15774*.
- [234] J. Xue et al., "Sparse mobile crowdsensing for cost-effective traffic state estimation with spatio-temporal transformer graph neural network," *IEEE Internet Things J.*, vol. 11, no. 9, pp. 16227–16242, May 2024.
- [235] V. John et al., "Estimation of steering angle and collision avoidance for automated driving using deep mixture of experts," *IEEE Trans. Intell. Vehicles*, vol. 3, no. 4, pp. 571–584, Dec. 2018.
- [236] X. Nie et al., "FlexMoE: Scaling large-scale sparse pre-trained model training via dynamic device placement," *Proc. ACM Manage. Data*, vol. 1, no. 1, pp. 1–19, May 2023.
- [237] Z. Chen, Q. Sun, N. Li, X. Li, Y. Wang, and I. Chih-Lin, "Enabling mobile AI agent in 6G era: Architecture and key technologies," *IEEE Netw.*, vol. 38, no. 5, pp. 66–75, Sep. 2024.
- [238] R. Zhang et al., "Toward democratized generative AI in next-generation mobile edge networks," *IEEE Netw.*, vol. 39, no. 6, pp. 251–260, Nov. 2025.
- [239] L. Ma et al., "Dynamic neural network-based resource management for mobile edge computing in 6G networks," *IEEE Trans. Cognit. Commun. Netw.*, vol. 10, no. 3, pp. 953–967, Jun. 2024.
- [240] G. Liu et al., "Fusion of mixture of experts and generative artificial intelligence in mobile edge metaverse," 2024, *arXiv:2404.03321*.
- [241] T. Yang et al., "DTQAtten: Leveraging dynamic token-based quantization for efficient attention architecture," in *Proc. Design, Autom. Test Eur. Conf. Exhib. (DATE)*, Mar. 2022, pp. 700–705.
- [242] Y. Qian et al., "EPS-MoE: Expert pipeline scheduler for cost-efficient MoE inference," 2024, *arXiv:2410.12247*.
- [243] B. Li, Y. Qin, B. Yuan, and D. J. Lilja, "Neural network classifiers using stochastic computing with a hardware-oriented approximate activation function," in *Proc. IEEE Int. Conf. Comput. Design (ICCD)*, Nov. 2017, pp. 97–104.
- [244] A. Q. Jiang et al., "Mixtral of experts," 2024, *arXiv:2401.04088*.
- [245] Y. Bisk et al., "PIQA: Reasoning about physical commonsense in natural language," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 5, 2020, pp. 7432–7439.
- [246] Y. Li, Z. Li, K. Zhang, R. Dan, S. Jiang, and Y. Zhang, "ChatDoctor: A medical chat model fine-tuned on a large language model meta-AI (LLaMA) using medical domain knowledge," *Cureus*, vol. 15, no. 6, pp. 1–12, Jun. 2023.
- [247] C. Zhao et al., "Generative AI for secure physical layer communications: A survey," *IEEE Trans. Cognit. Commun. Netw.*, vol. 11, no. 1, pp. 3–26, Feb. 2025.
- [248] W. Wang, N. Cheng, K. C. Teh, X. Lin, W. Zhuang, and X. Shen, "On countermeasures of pilot spoofing attack in massive MIMO systems: A double channel training based approach," *IEEE Trans. Veh. Technol.*, vol. 68, no. 7, pp. 6697–6708, Jul. 2019.
- [249] N. Wang, L. Jiao, A. Alipour-Fanid, M. Dabaghchian, and K. Zeng, "Pilot contamination attack detection for NOMA in 5G mm-wave massive MIMO networks," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 1363–1378, 2020.
- [250] S. Samarakoon et al., "5G-NIDD: A comprehensive network intrusion detection dataset generated over 5G wireless network," 2022, *arXiv:2212.01298*.
- [251] Y. Xu et al., "Fully-decoupled RAN for feedback-free multi-base station transmission in MIMO-OFDM system," *IEEE J. Sel. Areas Commun.*, vol. 43, no. 3, pp. 780–794, Mar. 2025.
- [252] H. Du et al., "Enhancing deep reinforcement learning: A tutorial on generative diffusion models in network optimization," *IEEE Commun. Surveys Tuts.*, vol. 26, no. 4, pp. 2611–2646, 2024.
- [253] Y. Zhuang, Z. Zheng, F. Wu, and G. Chen, "LiteMoE: Customizing on-device LLM serving via proxy submodel tuning," in *Proc. 22nd ACM Conf. Embedded Networked Sensor Syst.*, Nov. 2024, pp. 521–534.
- [254] W. Huang et al., "Mixture compressor for mixture-of-experts LLMs gains more," 2024, *arXiv:2410.06270*.
- [255] J. Li et al., "Expert-token resonance MoE: Bidirectional routing with efficiency affinity-driven active selection," 2024, *arXiv:2406.00023*.
- [256] C. Fellbaum, "WordNet," in *Theory and Applications of Ontology: Computer Applications*. Cham, Switzerland: Springer, 2010, pp. 231–243.
- [257] S. Gross, M. Ranzato, and A. Szlam, "Hard mixtures of experts for large scale weakly supervised vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5085–5093.
- [258] Y. Li et al., "SM3Det: A unified model for multi-modal remote sensing object detection," 2024, *arXiv:2412.20665*.
- [259] X. Xu, L. Kong, H. Shuai, L. Pan, Z. Liu, and Q. Liu, "LiMoE: Mixture of LiDAR representation learners from automotive scenes," 2025, *arXiv:2501.04004*.
- [260] Z. Xue et al., "RAPHAEL: Text-to-image generation via large mixture of diffusion paths," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2023, pp. 41693–41706.
- [261] Y. Wu, Y. Peng, Y. Lu, X. Chang, R. Song, and S. Watanabe, "Robust audiovisual speech recognition models with mixture-of-experts," in *Proc. IEEE Spoken Lang. Technol. Workshop (SLT)*, Dec. 2024, pp. 43–48.
- [262] M. Isaksson, F. Vannella, D. Sandberg, and R. Cöster, "mmWave beam selection in analog beamforming using personalized federated learning," in *Proc. IEEE Future Netw. World Forum (FNWF)*, Nov. 2023, pp. 1–6.
- [263] X. Yang, S. Feng, D. Wang, and Y. Zhang, "Image-text multimodal emotion classification via multi-view attentional network," *IEEE Trans. Multimedia*, vol. 23, pp. 4014–4026, 2021.
- [264] M. Alrabeiah, A. Hredzak, Z. Liu, and A. Alkhateeb, "ViWi: A deep learning dataset framework for vision-aided wireless communications," in *Proc. IEEE 91st Veh. Technol. Conf. (VTC-Spring)*, May 2020, pp. 1–5.
- [265] M. S. Allahham, M. F. Al-Sa'd, A. Al-Ali, A. Mohamed, T. Khattab, and A. Erbad, "DroneRF dataset: A dataset of drones for RF-based detection, classification and identification," *Data Brief*, vol. 26, Oct. 2019, Art. no. 104313.
- [266] Y. Xu et al., "Enhancing wireless networks for IoT with large vision models: Foundations and applications," 2025, *arXiv:2508.00583*.
- [267] J. Kim, S. Chang, and N. Kwak, "PQK: Model compression via pruning, quantization, and knowledge distillation," 2021, *arXiv:2106.14681*.
- [268] X. Huang, Z. Liu, S.-Y. Liu, and K.-T. Cheng, "Efficient and robust quantization-aware training via adaptive coreset selection," 2023, *arXiv:2306.07215*.
- [269] J. Kim, Y. Bhalgat, J. Lee, C. Patel, and N. Kwak, "QKD: Quantization-aware knowledge distillation," 2019, *arXiv:1911.12491*.
- [270] Q. Song, S. Jing, S. Zhang, S. Zhang, and C. Huang, "Mixture-of-experts for distributed edge computing with channel-aware gating function," 2025, *arXiv:2504.00819*.
- [271] W. Wu et al., "Mixture-of-experts meets in-context reinforcement learning," 2025, *arXiv:2506.05426*.
- [272] Y. Guo, Z. Cheng, X. Tang, Z. Tu, and T. Lin, "Dynamic mixture of experts: An auto-tuning approach for efficient transformer models," 2024, *arXiv:2405.14297*.
- [273] C. Benzaid and T. Taleb, "AI-driven zero touch network and service management in 5G and beyond: Challenges and research directions," *IEEE Netw.*, vol. 34, no. 2, pp. 186–194, Mar. 2020.



Yunting Xu received the Ph.D. degree from Nanjing University in 2024. He is currently a Research Fellow with the School of Computer Science and Engineering, Nanyang Technological University, Singapore. He mainly focuses on artificial intelligence and networking optimization in the field of emerging wireless networks.



Jiacheng Wang (Member, IEEE) received the M.S. and Ph.D. degrees from the School of Communication and Information Engineering, Chongqing University of Posts and Telecommunications, in 2018 and 2022, respectively. From 2021 to 2022, he was a Visiting Researcher with the College of Computing and Data Science, Nanyang Technological University, Singapore, where he is currently a Post-Doctoral Research Fellow. He has published more than 40 papers, including IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS,

IEEE TRANSACTIONS ON MOBILE COMPUTING, IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, IEEE TRANSACTIONS ON COGNITIVE COMMUNICATIONS AND NETWORKING, IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, IEEE COMMUNICATIONS SURVEYS AND TUTORIALS, IEEE WIRELESS COMMUNICATIONS, *IEEE Network*, IEEE WIRELESS COMMUNICATIONS LETTERS, IEEE GLOBECOM, IEEE ICC, and IEEE WCNC. His research interests include generative AI, integrated sensing and communications, network optimization, and edge intelligence. He has received the IEEE ICC 2025 Best Paper Award and he was a Guest Editor of IEEE TRANSACTIONS ON COGNITIVE COMMUNICATIONS AND NETWORKING, *Wireless Communications*, IEEE OPEN JOURNAL OF THE COMMUNICATIONS SOCIETY, *IEEE Internet of Things Magazine*, and IEEE NETWORKING LETTERS.



Ruichen Zhang (Member, IEEE) received the B.E. degree from Henan University (HENU), China, in 2018, and the Ph.D. degree from Beijing Jiaotong University (BJTU), China, in 2023. In 2024, he was a Visiting Scholar with the College of Information and Communication Engineering, Sungkyunkwan University, Suwon, South Korea. He is currently a Post-Doctoral Research Fellow with the College of Computing and Data Science, Nanyang Technological University (NTU), Singapore. His research interests include agentic AI, LLM-empowered network-

working, reinforcement learning-enabled wireless communications, generative AI models, and heterogeneous networks.



Changyuan Zhao (Graduate Student Member, IEEE) received the B.Sc. degree in computing and information science from the University of Science and Technology of China, Hefei, China, in 2020, and the M.A.Eng. degree in computer science from the Institute of Software, CAS, Beijing, China, in 2023. He is currently pursuing the Ph.D. degree with the College of Computing and Data Science, Nanyang Technological University, Singapore. His research interests include generative AI, communication security, and resource allocation.



Dusit Niyato (Fellow, IEEE) received the B.Eng. degree from the King Mongkut's Institute of Technology Ladkrabang (KMITL), Thailand, and the Ph.D. degree in electrical and computer engineering from the University of Manitoba, Canada. He is currently a Professor with the College of Computing and Data Science, Nanyang Technological University, Singapore. His research interests include mobile generative AI, edge intelligence, decentralized machine learning, and incentive mechanism design.



Jiawen Kang (Senior Member, IEEE) received the M.S. and Ph.D. degrees from Guangdong University of Technology, China, in 2015 and 2018, respectively. He was a Post-Doctoral Researcher with Nanyang Technological University, Singapore, from 2018 to 2021. He is currently a Full Professor at Guangdong University of Technology. His main research interests include blockchain, metaverse, and AIGC in wireless communications and networking.



Zehui Xiong (Senior Member, IEEE) received the Ph.D. degree from Nanyang Technological University (NTU). He was a Visiting Scholar with Princeton University and the University of Waterloo. He is currently a Full Professor with the School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast, U.K. Prior to that, he was with Singapore University of Technology and Design and NTU. Recognized as a Clarivate Highly Cited Researcher, he has published over 250 peer-reviewed research articles in leading journals,

with numerous best paper awards from international flagship conferences. His honors include the IEEE Asia-Pacific Outstanding Young Researcher Award, the IEEE VTS Early Career Award, the IEEE Early Career Award for Excellence in Scalable Computing, the IEEE Technical Committee on Blockchain and Distributed Ledger Technologies Early Career Award, the IEEE Internet Technical Committee Early Achievement Award, the IEEE TCSVC Rising Star Award, the IEEE TCI Rising Star Award, the IEEE TCCLD Rising Star Award, the IEEE ComSoc Outstanding Paper Award, the IEEE Best Land Transport Paper Award, the IEEE Asia-Pacific Outstanding Paper Award, the IEEE CSIM Technical Committee Best Journal Paper Award, the IEEE SPCC Technical Committee Best Paper Award, and the IEEE Big Data Best Influential Conference Paper Award. Featured in Forbes Asia 30U30, he serves as an editor for many leading journals and the chair for numerous international conferences.



Bo Qian (Member, IEEE) received the B.S. and M.S. degrees from the College of Mathematics, Sichuan University, Chengdu, China, in 2015 and 2018, respectively, and the Ph.D. degree in information and communication engineering from Nanjing University, Nanjing, China, in 2022. From 2022 to 2024, he was a Post-Doctoral Fellow with the Peng Cheng Laboratory, Shenzhen, China. He is currently an Assistant Professor (Special Appointment) with the Information Systems Architecture Science Research Division, National

Institute of Informatics, Tokyo, Japan. His main research interests focus on cellular RAN, SAGIN, and the industrial internet. He was a recipient of the Best Paper Award at IEEE VTC2020-Fall.



Haibo Zhou (Senior Member, IEEE) received the Ph.D. degree in information and communication engineering from Shanghai Jiao Tong University, Shanghai, China, in 2014. From 2014 to 2017, he was a Post-Doctoral Fellow with the Broadband Communications Research Group, Department of Electrical and Computer Engineering, University of Waterloo. He is currently a Full Professor with the School of Electronic Science and Engineering, Nanjing University, Nanjing, China. He was a recipient of the 2019 IEEE ComSoc Asia-Pacific Outstanding

Young Researcher Award, the 2023 IEEE ComSoc WTC Outstanding Young Researcher Award, the 2023-2024 IEEE ComSoc Distinguished Lecturer, and the 2023-2025 IEEE VTS Distinguished Lecturer. He served as the Track/Symposium Co-Chair for IEEE/CIC ICC 2019, IEEE VTC-Fall 2020, IEEE VTC-Fall 2021, WCSP 2022, IEEE GLOBECOM 2022, IEEE ICC 2024, and IEEE GLOBECOM 2024. He is an Associate Editor of IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, IEEE INTERNET OF THINGS JOURNAL, *IEEE Network Magazine*, and *Journal of Communications and Information Networks*. His research interests include resource management and protocol design in B5G/6G networks, vehicular ad hoc networks, and space-air-ground integrated networks.



Shiwen Mao (Fellow, IEEE) received the Ph.D. degree in electrical engineering from Polytechnic University in 2004. He is currently a Professor and an Earle C. Williams Eminent Scholar and the Director of the Wireless Engineering Research and Education Center at Auburn University. His research interests include wireless networks, multimedia communications, and smart grid. He received the IEEE ComSoc MMTIC Outstanding Researcher Award in 2023, the SEC Faculty Achievement Award for Auburn in 2023, the IEEE ComSoc TCSR Distinguished Technical Achievement Award in 2019, Auburn University

Creative Research and Scholarship Award in 2018, the NSF CAREER Award in 2010, and multiple IEEE service awards. He was also a co-recipient of several best paper and best journal paper awards, including the 2022 IEEE ComSoc eHealth Technical Committee Best Journal Paper Award, the 2021 Elsevier/KeAi Digital Communications and Networks Best Paper Award, the 2021 IEEE Internet of Things Journal Best Paper Award, the 2021 IEEE Communications Society Outstanding Paper Award, the 2020 IEEE Vehicular Technology Society Jack Neubauer Memorial Award, the 2018 IEEE ComSoc MMTC Best Journal Paper Award, the 2017 IEEE ComSoc MMTC Best Conference Paper Award, and the 2004 IEEE Communications Society Leonard G. Abraham Prize in the field of communications systems. He was a co-recipient of the best paper/demo awards of 12 conferences. He is the Editor-in-Chief of IEEE TRANSACTIONS ON COGNITIVE COMMUNICATIONS AND NETWORKING, the Member-at-Large of the Board of Governors of the IEEE Communications Society, and the Vice President of Technical Activities of the IEEE Council on Radio Frequency Identification (CRFID).



Abbas Jamalipour (Fellow, IEEE) received the Ph.D. degree in electrical engineering from Nagoya University, Nagoya, Japan, in 1996. He is currently a Professor of ubiquitous mobile networking with The University of Sydney, Camperdown, NSW, Australia. He has authored nine technical books, eleven book chapters, over 650 technical articles, and five patents, all in wireless communications and networking. He has been an Elected Member of the Board of Governors of the IEEE Vehicular Technology Society since 2014. He is a fellow of

the Institute of Electrical, Information, and Communication Engineers, the Institution of Engineers Australia, and the International Artificial Intelligence Industry Alliance, an ACM Professional Member, and an IEEE Distinguished Speaker. He was a recipient of a number of prestigious awards, including 15 best paper awards. He was the Editor-in-Chief of IEEE WIRELESS COMMUNICATIONS, the Vice President-Conferences, and a member of the Board of Governors of the IEEE Communications Society. He is the Editor-in-Chief of IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY. He was the President of the IEEE Vehicular Technology Society from 2020 to 2021. Previously, he held the positions of the Executive Vice President and the Editor-in-Chief of the VTS Mobile World.



Xuemin Shen (Fellow, IEEE) received the Ph.D. degree in electrical engineering from Rutgers University, New Brunswick, NJ, USA, in 1990. He is currently a University Professor with the Department of Electrical and Computer Engineering, University of Waterloo, Canada. His research focuses on network resource management, wireless network security, the Internet of Things, AI for networks, and vehicular networks. He is a registered Professional Engineer in ON, Canada, an Engineering Institute of Canada Fellow, a Canadian Academy of Engineering

Fellow, a Royal Society of Canada Fellow, a Chinese Academy of Engineering Foreign Member, an International Fellow of the Engineering Academy of Japan, and a Distinguished Lecturer of the IEEE Vehicular Technology Society and the Communications Society. He received the “West Lake Friendship Award” in Zhejiang in 2023, the President’s Excellence in Research from

the University of Waterloo in 2022, Canadian Award for Telecommunications Research from Canadian Society of Information Theory (CSIT) in 2021, the R.A. Fessenden Award from IEEE, Canada, in 2019, the Award of Merit from the Federation of Chinese Canadian Professionals (Ontario) in 2019, the James Evans Avant Garde Award from the IEEE Vehicular Technology Society in 2018, the Joseph LoCicero Award and the Education Award from the IEEE Communications Society (ComSoc) in 2015 and 2017, respectively, and the Technical Recognition Award from the Wireless Communications Technical Committee in 2019 and the AHSN Technical Committee in 2013. He has also received the Excellent Graduate Supervision Award from the University of Waterloo in 2006 and the Premier’s Research Excellence Award (PREA) in ON, in 2003. He is the past President of the IEEE Communications Society. He was the Vice President of Technical and Educational Activities, the Vice President of Publications, the Member-at-Large of the Board of Governors, the Chair of the Distinguished Lecturer Selection Committee, and a member of the IEEE Fellow Selection Committee of the ComSoc. He served as the Editor-in-Chief for IEEE INTERNET OF THINGS JOURNAL, *IEEE Network*, and *IET Communications*.



Dong In Kim (Life Fellow, IEEE) received the Ph.D. degree in electrical engineering from the University of Southern California, Los Angeles, CA, USA, in 1990. He was a Tenured Professor with the School of Engineering Science, Simon Fraser University, Burnaby, BC, Canada. He is currently a Distinguished Professor with the College of Information and Communication Engineering, Sungkyunkwan University, Suwon, South Korea. He is a fellow of Korean Academy of Science and Technology and a Life Member of the National

Academy of Engineering of Korea. He was the first recipient of the NRF of Korea Engineering Research Center (ERC) in Wireless Communications for RF Energy Harvesting from 2014 to 2021. He received several research awards, including the 2023 IEEE ComSoc Best Survey Paper Award and the 2022 IEEE Best Land Transportation Paper Award. He was selected as the 2019 recipient of the IEEE ComSoc Joseph LoCicero Award for Exemplary Service to Publications. He has been listed as a 2020 (Cross-Field)/2022 (Cross-Field)/2025 (Computer Science) Highly Cited Researcher (HCR) by Clarivate Analytics. He was the General Chair of IEEE ICC 2022, Seoul. From 2001 to 2024, he served as an Editor, the Editor-at-Large, and an Area Editor for Wireless Communications I of IEEE TRANSACTIONS ON COMMUNICATIONS. From 2002 to 2011, he served as an Editor and a Founding Area Editor of Cross-Layer Design and Optimization for IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS. From 2008 to 2011, he served as the Co-Editor-in-Chief for IEEE/KICS JOURNAL OF COMMUNICATIONS AND NETWORKS. He served as the Founding Editor-in-Chief for IEEE WIRELESS COMMUNICATIONS LETTERS from 2012 to 2015.