LLM-Based Port Selection and Beamforming for Multiuser MISO with Fluid Antenna Systems

Wei Guo, *Graduate Student Member, IEEE*, Kai Liang, *Member, IEEE*, Gan Zheng, *Fellow, IEEE*, Xiaoli Chu, *Senior Member, IEEE*, Guorong Zhou, *Member, IEEE*, Kai-Kit Wong, *Fellow, IEEE*, and Chan-Byoung Chae, *Fellow, IEEE*

***Abstract*—Fluid antenna systems (FAS) introduce additional spatial degrees of freedom (DoF) by dynamically adjusting antenna positions. When integrated with multiple-input multiple-output (MIMO) technologies (forming MIMO-FAS), this capability significantly enhances wireless communication performance. However, the joint optimization of high-dimensional port selection and beamforming in MIMO-FAS presents a challenging non-convex combinatorial problem. In this paper, we propose a novel large language model (LLM)-based intelligent framework for jointly optimizing port selection and beamforming in multiuser MIMO-FAS systems. The objective is to maximize the sum rate under base station (BS) power and port activation constraints. Departing from the conventional two-stage approaches, where port selection and beamforming are handled sequentially, we adopt a parallel output strategy that simultaneously determines port indices and beamforming coefficients, leveraging the multi-task learning capabilities of LLMs. To enhance efficiency, we incorporate low-rank adaptation (LoRA) for fine-tuning pre-trained LLMs, significantly reducing training cost while maintaining generalization performance. Also, we employ Gumbel-Sinkhorn stochastic relaxation to make discrete port selection differentiable, enabling end-to-end optimization. Numerical results demonstrate that the proposed method outperforms state-of-the-art techniques in terms of sum rate, validating the effectiveness of the LLM-driven joint optimization approach.**

***Index Terms*—Fluid antenna system, large language model (LLM), multiuser MISO, port selection, LoRA fine-tuning.**

I. Introduction

MULTIPLE-input multiple-output (MIMO) is the most celebrated mobile technology in the modern history of wireless communications. The current fifth generation (5G) is relying on massive multiuser MIMO in the physical layer [1], [2], [3]. Going forward, an extra-large version of MIMO in the form of a cell-free network architecture is anticipated [4]. Understandably, the power of MIMO comes from the number of antennas at the base station (BS). With or without the cell-free setup, nonetheless, we cannot expect that the number of antennas continues to increase without bound. In fact, recent discussion for the sixth generation (6G) seems to suggest that each BS has only 32 antennas, less than that in 5G, contrary to most of the early predictions that a BS in 6G would have hundreds or even thousands of antennas. With not so massive MIMO at the BS but still aiming for more ambitious goals, a new degree of freedom (DoF) needs to be sought [5], [6].

This has motivated the effort of the fluid antenna system (FAS), which promotes an integrated reconfigurable physical-layer system exploiting shape-flexible position-reconfigurable antenna systems to empower the physical layer with more DoF [7], [8]. First introduced by Wong *et al.* in [9], [10], FAS is also encouraged by the recent advancements in reconfigurable antenna technologies such as movable arrays [11], [12], liquid antennas [13], pixel-based antennas [14], and metamaterials-based antennas [15], [16]. Since the emergence of FAS, there have been discussion about the potential of utilizing artificial intelligence (AI) [17] and large language model (LLM) [18] in FAS system optimization. Moreover, Lu *et al.* in [19] offered an explanation of FAS from an electromagnetic perspective. Wu *et al.* in [20] categorized various technologies for implementing FAS and emphasized their strengths and characteristics.

With position reconfigurability, FAS has been demonstrated to have tremendous spatial diversity given a fixed space [21], [22]. For instance, with a space of $0.5\lambda \times 0.5\lambda$ (with λ being the carrier wavelength) at both ends, a fixed MIMO channel would have a diversity order of 16[1] but the diversity order for FAS-assisted MIMO systems (MIMO-FAS) reaches 169, a tenfold increase.

However, the capability of FAS originates from the possibility of finding the optimal antenna positions (known as port selection) in conjunction with beamforming [23] which is an non-deterministic polynomial-time hard (NP-hard) combinatorial optimization problem. Traditional optimization approaches typically address this problem by solving partially relaxed convex approximations, employing techniques such as semidefinite relaxation (SDR) [24] or successive convex approximation (SCA). Another commonly used approach is alternating optimization (AO) [25], which iteratively optimizes beamforming for fixed port selections, and vice versa, until convergence. Nevertheless, these optimization methods entail high computational complexity, often requiring repeated solutions of large-scale convex optimization problems or exhaustive search subproblems. Moreover, even after applying convex relaxation methods, the obtained solutions remain suboptimal without guaranteed optimality. Consequently, when the num-

(Corresponding authors: K. Liang; G. Zheng.)

W. Guo, K. Liang and G. Zhou are with the School of Telecommunications Engineering, Xidian University, Xi'an, 710071, China (e-mail: wguoone@163.com; kliang@xidian.edu.cn; guor_zhou@163.com).

G. Zheng is with the School of Engineering, University of Warwick, Coventry, CV4 7AL, United Kingdom (e-mail: gan.zheng@warwick.ac.uk).

X. Chu is with the Department of Electronic and Electrical Engineering, University of Sheffield, Sheffield, S10 2TN, United Kingdom (e-mail: x.chu@sheffield.ac.uk).

K. K. Wong is affiliated with the Department of Electronic and Electrical Engineering, University College London, Torrington Place, WC1E 7JE, United Kingdom (email: kai-kit.wong@ucl.ac.uk) and he is also affiliated with Yonsei Frontier Lab, Yonsei University, Seoul, Korea.

C.-B. Chae is with the School of Integrated Technology, Yonsei University, Seoul, 03722, Korea (e-mail: cbchae@yonsei.ac.kr).

[1]This assumes a rich-scattering environment.

ber of ports becomes moderately large, the computational cost of these algorithms grows dramatically.

Alternatively, learning-based methods can make decisions via one-shot model inference, hence potentially lowering online computation overhead while preserving performance. For example, the authors in [26] investigated an over-the-air federated learning system where fluid antennas (FA) are integrated at the access point and proposed an long short-term memory (LSTM)-based algorithm to jointly optimize and simultaneously output the FA positions, beamforming vectors, and user selection strategy, aiming to maximize the number of selected users. However, these learning-based approaches generally require a substantial amount of offline optimization solutions serving as labeled training data. Obtaining these training labels inherently constitutes an NP-hard problem, effectively equivalent to performing exhaustive offline searches or repeatedly applying computationally intensive algorithms. Even when reinforcement learning [27] is employed to directly approximate optimal decisions through unsupervised training with the system rate serving as a reward, the training process still requires extensive trial-and-error exploration in simulated environments. Moreover, analogous to AO algorithm, learning-based approaches often utilize sequential two-stage methods to tackle the port selection and beamforming optimization problems [28]. Overall, iterative AO or any two-stage methods, however, tend to ignore the inherent coupling between port selection and beamforming and naively compromise the two sets of variables when optimizing.

Clearly, a more efficient solution is needed moving forward, and LLMs seem well positioned to provide it. In recent years, LLMs have achieved remarkable success in natural language processing, information retrieval, and related domains [29], [30]. The introduction of LLMs into physical layer design has also attracted much attention. For instance, building on the expressive power of LLMs and leveraging fine-tuning of pre-trained networks, [31] proposed a method to predict future downlink channel state information (CSI) sequences from historical uplink CSI sequences. Pre-trained on extensive textual corpora, LLMs possess impressive capabilities including generalization, expressive, and multi-task learning. Unlike traditional algorithms that rely on mechanical searches within combinatorial spaces, LLMs can effectively extract latent patterns and perform inference from high-dimensional complex space. This offers novel insights into overcoming the joint optimization challenges in MIMO-FAS using LLMs. In this paper, we propose an LLM-based port selection and beamforming method for multiuser multiple-input single-output system assisted by FAS (MISO-FAS). The contributions of this paper are summarized as follows:

- First, we propose an LLM-based end-to-end joint optimization framework to maximize the sum rate for a multiuser MISO-FAS in the downlink with the BS transmission power and port activation constraints. The framework consists of a preprocessing module, a backbone network, and a post-processing module. The preprocessing module utilizes multi-head attention and positional embedding to align the complex CSI data format with that of the pre-trained LLM and extracts initial feature representations. The backbone network leverages the powerful generalization and expressive capabilities of the pre-trained LLM to improve CSI feature extraction and generate richer feature representations, while the post-processing module utilizes a model-based optimal beamforming solution structure [32] to optimize the port selection and the beamforming power allocation. Overall, this framework achieves end-to-end optimization, mapping input CSI directly to explicit beamforming outputs and discrete port indices learned from differentiable soft outputs and hardened during inference.
- We propose a parallel output mechanism to alleviate the potential performance degradation caused by traditional two-stage methods that select port first and then perform beamforming. By enabling the pre-trained LLM to output both the selected port indices and the power allocation factors for beamforming derivation in parallel, thereby fully exploiting the LLM's inherent expressive power and multi-task learning capability to enable joint optimization of port selection and beamforming.
- Also, we incorporate the low-rank adaptation (LoRA) [33] fine-tuning technology in the backbone network, which freezes the original parameters of the pre-trained LLM and introduces a small number of trainable low-rank matrices, thereby significantly reducing computational resource requirements without sacrificing performance. Furthermore, to address the issue of non-differentiability arising from discrete decisions in port selection and the strong reliance of commonly adopted pointer networks [34] on labeled data, we employ a stochastic relaxation technique based on the Gumbel-Sinkhorn distribution [35]. This maps discrete decisions onto differentiable operations, enabling the entire framework to be trained end-to-end via gradient descent optimization algorithms in an unsupervised manner, while eliminating the dependence on labeled data.
- Numerical results demonstrate that our proposed framework outperforms existing deep learning (DL)-based convolutional neural network (CNN) [36] and transformer [37] approaches in terms of downlink sum rate, and also achieves a performance gain compared to the LLM-based two-stage method. Moreover, millisecond-level inference latency also demonstrates the efficiency of the proposed method.

The rest of this paper is organized as follows. In Section II, we review related work on port selection and beamforming in FAS. In Section III, we introduce the system model and problem formulation. Section IV then presents the proposed LLM-based optimization framework. Section V provides numerical results, and finally, we conclude this paper in Section VI.

Notations—Matrices and vectors are represented in upper and lower case bold letters, respectively. The superscripts $()^{\dagger}$, $()^{T}$ and $()^{-1}$ represent the conjugate transpose, the transpose and inverse of a matrix or vector, respectively. $|\cdot|$ and $\|\cdot\|$ stand for absolute value and Euclidean norm, respectively. Also, $\Re(\cdot)$ and $\Im(\cdot)$ denote the real and imaginary parts of

a complex variable, respectively. $\binom{N}{n}$ denotes the number of combinations of n elements selected from N elements. $\mathrm{vec}(\cdot)$ is the vectorization operator. $\mathbf{I}_N$ denotes an $N \times N$ identity matrix. Moreover, $x \sim \mathcal{CN}(0, \sigma^2)$ denotes a complex Gaussian random variable with zero mean and covariance σ^2.

II. Related Work

In this section, we provide a detailed review and summary of existing literature on port selection in MIMO-FAS, joint port selection and beamforming design in MIMO-FAS, and LLMs for port/beamforming design and optimization. Table I summarizes the comparison of related work in terms of focus, methods, and limitations.

A. Port Selection in MIMO-FAS

FAS enhance the achievable capacity of MIMO system by introducing additional spatial DoF through dynamic selection among multiple potential antenna ports. Several attempts have been made to address the joint transmit and receive port selection problem in MIMO-FAS (a.k.a. fluid MIMO). In [38], based on a capacity upper bound, two schemes, namely joint convex relaxation (JCR) and a reduced exhaustive search (RES), as well as JCR and AO, were devised. More recently, the authors of [39] introduced a probabilistic, learning-based scheme within the cross-entropy optimization (CEO) framework to solve the port selection problem for fluid MIMO systems. In [40], the authors proposed a method based on least squares regression to estimate channel parameters in multi-ray millimeter-wave FAS, effectively reducing the number of required channel estimations. The authors of [41] derived the outage probability of FAS utilizing the maximum ratio combining (MRC) via Laplace transform. They also provided lower bound and asymptotic expressions for the outage probability, indicating that FAS can exploit substantial spatial diversity. Moreover, the authors in [42] employed an AO algorithm to jointly optimize the time scheduling and port activation of wireless energy transfer and wireless information transfer, and analyzed the overall performance of a wireless powered communication network utilizing FA. In [43], the authors formulated the optimal control problem involving BS transmit power and user FA positioning as a mean-field game (MFG) and used a finite-difference-based iterative algorithm to maximize energy efficiency.

Additionally, some literature has explored the optimization of port selection using intelligent learning. For example, the number of signal-to-noise ratio (SNR) observations for port selection can be reduced using machine learning (ML) with only partial CSI [44]. Besides, building on multi-armed bandit learning, the authors of [45] proposed an online learning-induced port selection method for FAS, which can learn the optimal port selection by interacting with different channel dynamics. By leveraging the spatial and temporal channel correlations over the ports, [46] presented a LSTM-based learning approach to estimate and predict the port CSI for fast port selection. In [47], the authors proposed employing conditional generative adversarial networks (cGANs) to facilitate FAS port selection, leveraging the correlation among the ports to generate channel gains for ports without observations. To mitigate the computational burden arising from FAS, the authors in [48] proposed a deep unfolding neural network method based on the block successive upper bound minimization (BSUM) algorithm, which accelerates the optimization and reduces computational latency. In [49], the authors proposed a resilient decentralized reinforcement learning (RL)-based approach for opportunistic fluid antenna multiple access (O-FAMA), designed to jointly optimize user scheduling and optimal port selection for maximizing the network sum rate. In [50], the authors proposed an attention-enhanced recurrent multi-agent reinforcement learning framework, which jointly optimizes the trajectories of all active unmanned aerial vehicles (UAVs) and the FA ports of passive UAVs, aimed at minimizing the average positioning error of the target UAV. A recent study [51] formulated the FAS port selection as a multi-label classification problem, employing liquid neural networks (LNNs) to predict optimal ports in emerging fluid antenna multiple access scenarios. The above results indicate great promises for learning methods on port selection but so far these studies have not considered port selection together with beamforming. This oversight inevitably limits the performance of MIMO-FAS considerably.

B. Joint Port Selection and Beamforming Design in MIMO-FAS

There are some efforts addressing the joint port selection and beamforming exists. In [52], an AO algorithm based on the penalty method and SCA was developed to minimize the total transmission power of a MISO-FAS downlink communication system. Similarly, [53] applied FAS in near-field communication scenarios, in which an AO algorithm was used to iteratively solve the subproblems of antenna position and beamforming for maximizing the energy efficiency. In [54], the authors proposed a block coordinate ascent (BCA)-based algorithm to jointly optimize beamforming and FA positioning in FA-aided downlink multiuser MIMO systems, and utilized a decentralized baseband processing (DBP) architecture for distributed implementation. The authors in [55] discussed the importance of integrating FAS and active reconfigurable intelligent surfaces (ARIS) and proposed an AO algorithm to optimize both the BS transmit beamforming and the positioning of FA. In [56], the authors investigated an energy efficient wireless communication system based on fluid antenna relay (FAR) and proposed an AO algorithm to jointly optimize the FAR and FA positions, power control, and beamforming design for enhancing the system's energy efficiency. In [57], the authors proposed an uplink two-timescale transmission scheme for MU-MIMO-FAS, where antenna positions and beamforming are optimized based on statistical and rapidly varying instantaneous CSI, respectively. By considering the port switching delay and energy consumption in FAS-assisted integrated data and energy transfer (IDET) systems, the authors in [58] alternately optimized port selection and beamforming vectors to maximize the short-term and long-term wireless energy transfer efficiency. In [59], the authors proposed an alternating

TABLE I
COMPARISON OF RELATED WORK IN FOCUS, METHODS AND LIMITATIONS

Ref.	Year	Focus	Methods	Limitations
[38]	2024	Transmit and receive antenna port selection in FAS	JCR&RES, JCR&AO	High computational complexity, optimality loss
[39]	2025	Joint transmit and receive port selection in Fluid-MIMO	Probability learning, CEO	Multiple iterations, slow convergence
[40]	2023	Multi-ray mmWave FAS	Least squares regression	Simplify channel model, performance degradation
[41]	2024	Outage probability of FAS	MRC, Laplace transform	Asymptotic approximation
[44]	2022	Port selection in FAS	ML, Smart, 'Predict and Optimize' (SPO)	Analytical approximation
[45]	2024	Port selection for FA in dynamic channel environment	Multi-armed bandit learning	Exploration-Exploitation trade-off
[46]	2023	Temporal and spatial correlation for FAS	LSTM	Reobservations
[47]	2025	Port selection in FAS	cGANs	Label dependency
[48]	2025	FA-enhanced vehicular communication	BSUM, deep unfolding	High complexity, slow convergence
[51]	2025	Fluid antenna multiple access	LNNs	Unstable training, hyperparameter optimization
[52]	2024	Antenna positioning and beamforming design for multiuser MISO-FAS	AO, penalty method and SCA	High computational complexity, optimality loss
[53]	2025	Joint beamforming and antenna design for Near-Field FAS	AO	Optimality loss, ignore coupling
[54]	2025	Joint beamforming and antenna position optimization for multiuser MIMO-FAS	BCA, DBP, fractional programming	High complexity, iterative optimization
[58]	2025	Port selection and beamforming design for IDET assisted by FAS	AO, SDR, constrained soft actor critic	Iterative optimization, optimality loss
[59]	2025	FAS assisted semantic communication	Alternating algorithm	High computational complexity, optimality loss
[60]	2025	Joint optimize beamforming and port selection for FAS-enabled ISAC	Sparse optimization, convex approximation	Iterative optimization, optimality loss
[61]	2025	Joint antenna positioning and beamforming design for FAS-enabled ISAC	AO, SDR, SCA, $\mathcal{S}$-procedure	High complexity, iterative optimization
[62]	2025	Joint on-off selection and beamforming for FRIS	CEO	High computational cost, local optima
[65]	2025	Joint antenna positioning and beamforming for Multiuser MISO-FAS	GNN, two-stage	Ignore coupling, local optima
[66]	2025	Hardware-software co-design for optimize beamforming and port selection in FAS	GNN, RPS	Model limitations, oversmoothing
[67]	2024	FAS Liberating multiuser MIMO for ISAC	DRL, LSTM	Extensive trial-and-error exploration
[71]	2024	LLM Enabled multi-task physical layer	LLM, LoRA	Large number of parameters
[72]	2025	LLM for wireless multi-task	LLM, mixture of experts LoRA	Large number of parameters
[73]	2025	LLM-empowered near-field communications for LAE	LLM	Large number of parameters, memory requirements
[74]	2025	Beam prediction based on LLMs	LLM, PaP	Difficult to maintain
[78]	2025	LLM-based channel prediction for OTFS-enabled satellite FAS links	LLM, LoRA	Large number of parameters
[79]	2025	Port prediction for FA based on LLMs	LLM, LoRA	Large number of parameters

algorithm to address the rate maximization problem in FAS-assisted semantic communication systems by jointly optimizing beamforming, port selection, and semantic compression rate. Additionally, the authors in [60] proposed an iterative algorithm based on sparse optimization, convex approximation, and penalty methods to address the minimum transmit power problem in integrated sensing and communication (ISAC) systems enabled by FAS, while satisfying both communication and sensing requirements. Focusing on FA-enhanced ISAC systems with perfect and imperfect CSI, the authors in [61] respectively developed two iterative AO algorithms (SDR and SCA, $\mathcal{S}$-procedure and SCA) to formulate the dual-functional beamforming and FA position problems. Furthermore, in [62], the authors proposed an iterative algorithm based on the CEO framework for jointly optimizing the element selection and discrete phase shifts in fluid reconfigurable intelligent surface (FRIS). In [63], the authors investigated UAV-assisted ISAC system employing FA, where an AO iterative algorithm was adopted to jointly optimize the activation port selection of the UAV-mounted FA and the communication beamforming vector. In [64], the authors proposed a framework for joint precoding and port selection in FAS, designed to enhance the physical layer security of ISAC systems.

But the above mentioned SCA and AO algorithms come at a high price of computational complexity and are often trapped in local optima or experience slow convergence. A few studies use neural networks to figure this out. The authors in [65] hence proposed a two-stage graph neural network (GNN) that separately generates antenna positions and beamforming vectors, targeting utility maximization in MISO-FAS. In [66], the authors proposed a hardware and software co-optimizing method for practical FAS, combining GNN with random port selection (RPS) to jointly optimize port selection and beamforming, and designed an instruction-driven deep learning accelerator based on field-programmable gate arrays to minimize inference latency. Furthermore, [67] adopted a deep reinforcement learning (DRL) framework for tackling the joint optimization of port selection and precoding matrices in multiuser MIMO-FAS ISAC systems with sensing constraints. In [68], the authors proposed a block coordinate descent (BCD) integrated with DRL-based intelligent antenna positioning approach for an FAS-assisted ISAC system, which addresses the joint optimization of beamforming and antenna positioning and balances sensing and communication performance. The authors in [69] proposed an FAS realized using a reconfigurable holographic surface (RHS), where element activation enables different antenna positions for each codeword, to perform codebook design and beam training based transmission. In [70], the authors proposed a model-based multi-agent reinforcement learning algorithm, which employs a distributed framework to solve the precoding and port selection problems of FAMA systems in multi-cell scenarios.

C. LLMs for Port/Beamforming Design and Optimization

With the recent advancements in generative artificial intelligence techniques, LLMs have begun demonstrating promising potential in wireless communications. In [71], the authors introduced a multi-task LLM framework for physical layer communications capable of simultaneously handling multiuser precoding, signal detection, and channel prediction. The authors in [72] proposed LLM4WM, a multi-task fine-tuning framework based on LLM, tailored specifically to wireless channel modeling tasks. By leveraging diverse multi-task datasets, LLM4WM can simultaneously perform multiple wireless channel-associated tasks, such as channel estimation, channel prediction, localization enhancement, and beam management. In [73], the authors applied LLM to maximize spectral efficiency for near-field communication scenarios in the low-altitude economy (LAE). By designing specialized adapters and fine-tuning the pre-trained Generative Pre-trained Transformer 2 (GPT-2) model, they effectively distinguished far-field and near-field users, achieving joint optimization of precoding and power allocation. Furthermore, in [74], the authors proposed utilizing LLMs for beam prediction, employing the prompt-as-prefix (PaP) technique to enrich context and enhance the model's understanding and reasoning capabilities regarding wireless data. In [75], the authors proposed a masked language model-based deep learning method for FAS CSI extrapolation, which infers the complete port CSI by exploiting incomplete CSI together with positional information encoding. In [76], the authors proposed a framework that integrating LLMs and convex-based optimization methods for joint user association and transmit beamforming optimization in multi-BS ISAC systems. The authors in [77] proposed a novel framework, BERT4Beam, based on bidirectional encoder representations from transformers (BERT), which formulates the beamforming optimization problem as a token-level sequence learning task and adopts task-specific pre-training and fine-tuning strategies.

Though research on applying LLMs to FAS remains limited, several initial efforts have emerged. In [78], the authors proposed a FAS-LLM architecture for predicting future channel states in orthogonal time-frequency-space (OTFS) satellite downlinks equipped with FAS. In [79], a port-LLM model was devised to address user equipment mobility challenges by strategically relocating FAS across available ports, thereby maintaining a relatively stable channel state over time. Although initial attempts have been made to apply LLMs to FAS, the joint optimization problem of port selection and beamforming in multiuser MISO-FAS scenarios utilizing LLM has remains an open area of research.

III. System Model and Problem Formulation

As shown in Fig. 1, we consider a multiuser MISO-FAS downlink system, in which the BS is equipped with a two-dimensional (2D) FAS with N_{rf} RF-chains to serve K users each with a single fixed-position antenna (FPA), also referred to as multiuser Tx-MISO-FAS [8]. The FAS at the BS has a total of $N = N_x \times N_y$ ports, which are uniformly distributed over a 2D surface with an area $W = W_x\lambda \times W_y\lambda$, where λ is the wavelength of the carrier signal. The N_i ports are evenly distributed along a straight line of length $W_i\lambda$ for $i \in \{x, y\}$. In this multiuser Tx-MISO-FAS scenario, we assume that the BS activates only n out of N ports to serve K users. First, the

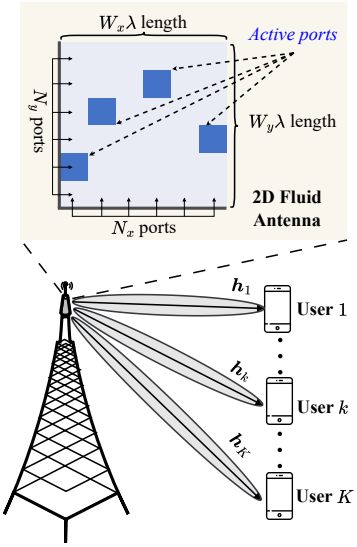

Fig. 1. The multiuser MISO-FAS downlink communication system.

total number of ports N in FAS is typically much larger than the number of available RF chains N_{rf}[8], and each activated port requires an independent controllable RF path. Therefore, the number of activated ports must satisfy $n \leq N_{rf} \ll N$. Second, because FAS ports are densely distributed and often exhibit strong spatial correlation and mutual coupling [22], [44], indiscriminately activating all ports may result in an ill-conditioned effective channel, as well as increased hardware complexity and power consumption [14], [80]. Moreover, since we aim to simultaneously transmit K independent data streams over the same time-frequency resources and ensure their separability, we set $n \geq K$ [8], [67].

Considering a rich scattering environment [22], we employ Jake's moded [81] to characterize the spatial correlation between any ports. Specifically, the spatial correlation between the (n_x, n_y)-th port and the $(\tilde{n}_x, \tilde{n}_y)$-th port is given by

$$J_{(n_x,n_y),(\tilde{n}_x,\tilde{n}_y)} = j_0\left(2\pi\sqrt{\left(\frac{|n_x-\tilde{n}_x|}{N_x-1}W_x\right)^2+\left(\frac{|n_y-\tilde{n}_y|}{N_y-1}W_y\right)^2}\right), \quad (1)$$

where $j_0(\cdot)$ denotes the zero-order spherical Bessel function. To simplify the notations, we order the port indices from the top-left to the bottom-right, mapping the 2D port coordinates into one-dimensional (1D) index. Under this convention, the (n_x, n_y)-th port can be mapped to the new index

$$l_{(n_x,n_y)} = (n_y-1)N_x + n_x, \quad (2)$$

where $l_{(n_x,n_y)} \in \{1,\ldots,N\}$. Consequently, we have the spatial correlation matrix $\boldsymbol{J}$ as

$$\boldsymbol{J} = \begin{bmatrix} J_{1,1} & J_{1,2} & \cdots & J_{1,N} \\ J_{2,1} & J_{2,2} & \cdots & J_{2,N} \\ \vdots & \vdots & \ddots & \vdots \\ J_{N,1} & J_{N,2} & \cdots & J_{N,N} \end{bmatrix}, \quad (3)$$

where $J_{i,j}$ is the spatial correlation between the i-th and j-th port after mapping. Noting that $\boldsymbol{J}$ is symmetric (i.e., $J_{i,j} = J_{j,i}$), (3) can be decomposed into

$$\boldsymbol{J} = \boldsymbol{F}\boldsymbol{\Lambda}\boldsymbol{F}^\dagger, \quad (4)$$

where $\boldsymbol{F} \in \mathbb{C}^{N\times N}$ is the unitary matrix whose columns are the eigenvectors of $\boldsymbol{J}$, and $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_N)$ is the diagonal eigenvalue matrix of $\boldsymbol{J}$ in descending order, i.e., $\lambda_1 > \lambda_2 > \cdots > \lambda_N$.

Given $\boldsymbol{F}$ and $\boldsymbol{\Lambda}$, the complex channel from the BS to the k-th user can be modeled as

$$\boldsymbol{h}_k^\dagger = \sqrt{\varrho_k}\boldsymbol{g}_k^\dagger\sqrt{\boldsymbol{\Lambda}}^\dagger\boldsymbol{F}^\dagger, \quad (5)$$

where ϱ_k denotes the large-scale path loss for user k, and $\boldsymbol{g}_k \in \mathbb{C}^{N\times 1}$ is the small-scale fading channel vector whose entry is independent and follows the circularly symmetric complex Gaussian (CSCG) distribution, i.e., $\mathcal{CN}(0,1)$. Accordingly, the overall channel matrix can be represented as $\boldsymbol{H}_K = [\boldsymbol{h}_1^\dagger, \ldots, \boldsymbol{h}_K^\dagger]^T \in \mathbb{C}^{K\times N}$. Supposing that n distinct ports are activated, the received signal of the k-th user is

$$z_k = \boldsymbol{h}_k^\dagger\boldsymbol{\mathcal{A}}_n\boldsymbol{C}\boldsymbol{s} + \zeta_k, \quad (6)$$

where $\boldsymbol{\mathcal{A}}_n = [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n]$ denotes the activated port matrix and $\boldsymbol{a}_l$ is the standard basis vector (i.e., $\boldsymbol{a}_l \in \{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_N\}$ and $\boldsymbol{e}_i$ is the i-th column of $\mathbf{I}_N$), $\boldsymbol{C} = [\boldsymbol{c}_1, \ldots, \boldsymbol{c}_K] \in \mathbb{C}^{n\times K}$ denotes the beamforming matrix, $\boldsymbol{s} = [s_1, \ldots, s_K]^T \in \mathbb{C}^{K\times 1}$ denotes the transmitted symbols for all users with $\mathbb{E}\{|s_k|^2\} = 1$, and $\zeta_k \sim \mathcal{CN}(0, \sigma^2)$ is the additive noise with zero mean and variance σ^2. Then the signal-to-interference-plus-noise ratio (SINR) of the k-th user is given by

$$\gamma_k = \frac{\left|\boldsymbol{h}_k^\dagger\boldsymbol{\mathcal{A}}_n\boldsymbol{c}_k\right|^2}{\sum_{j\neq k}\left|\boldsymbol{h}_k^\dagger\boldsymbol{\mathcal{A}}_n\boldsymbol{c}_j\right|^2 + \sigma^2}. \quad (7)$$

Our goal is to maximize the achievable sum rate of multiuser Tx-MISO-FAS under the BS transmission power budget and the port activation constraint by optimizing the port selection matrix $\boldsymbol{\mathcal{A}}_n$ and the beamforming matrix $\boldsymbol{C}$. Therefore, the optimization problem can be formulated as

$$\max_{\boldsymbol{\mathcal{A}}_n, \boldsymbol{C}} \quad \sum_{k=1}^{K}\log_2(1+\gamma_k) \quad (8a)$$

$$\text{s.t.} \quad \boldsymbol{\mathcal{A}}_n(:,l) \in \{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_N\}, \quad (8b)$$

$$\boldsymbol{\mathcal{A}}_n(:,l) \neq \boldsymbol{\mathcal{A}}_n(:,m), \forall l \neq m, \quad (8c)$$

$$\sum_{k=1}^{K}\|\boldsymbol{c}_k\|^2 \leq P_{\max}, \quad (8d)$$

where (8b) represents that the l-th column in the activation port matrix $\boldsymbol{\mathcal{A}}_n$ is one of the columns in the $N\times N$ identity matrix, (8c) ensures the activated ports are distinct, and (8d) denotes the power constraint of BS with the maximum power $P_{\max}$. Clearly, (8) is a nonconvex optimization problem that cannot be solved in closed form because i) port selection is an NP-hard combinatorial optimization problem. Employing mathematical optimization techniques, such as exhaustive search, requires solving $\binom{N}{n}$ nonconvex problems. As the number of ports, N, increases, the computational complexity and solution

time will increase dramatically; ii) The beamforming matrix $\boldsymbol{C}$ and the selected ports $\boldsymbol{A}_n$ are intricately coupled, making the joint optimization even more complicated. Specifically, $\boldsymbol{C}$ can only be designed at the selected ports, while the port selection variables and beamforming vectors jointly determine the effective channel, appearing in a multiplicative form within the objective function. To address these challenges, we exploit the approximation capability of neural networks. In particular, motivated by the powerful modeling and expressive capacity of LLM in learning complex nonlinear mappings, we propose an LLM-based learning framework for joint port selection and beamforming design.

IV. Joint Optimization of Port Selection and Beamforming Learning Framework

Here, we propose an LLM-based joint optimization of port selection and beamforming learning framework in multiuser Tx-MISO-FAS. As shown in Fig. 2, the overall learning framework includes preprocessing module, backbone network, and post-processing module. Next, we describe each component in detail and then outline our overall learning strategy.

A. Preprocessing Module

The preprocessing module is employed to align the complex CSI data with the text-based pre-trained LLM input format while simultaneously extracting initial feature representations from the raw CSI data. We assume that the BS has access to the $N \times K$ instantaneous CSI across all N ports and K users, which serves as the unified input for port selection and beamforming. Below, we present the key parts of the preprocessing module in the order of CSI processing.

1) Fully connected layer: Consider that neural network frameworks (such as PyTorch or TensorFlow) generally require real-valued inputs while CSI data is complex-valued, we first decompose the complex CSI matrix $\boldsymbol{H}_K \in \mathbb{C}^{K\times N}$ into its real part $\Re(\boldsymbol{H})$ and imaginary part $\Im(\boldsymbol{H})$ and convert them into $\boldsymbol{H}_t \in \mathbb{R}^{2\times K\times N}$. Then, we apply the vectorization operator to flatten both $\Re(\boldsymbol{H}_t)$ and $\Im(\boldsymbol{H}_t)$ into vectors, denoted as

$$\begin{cases} \boldsymbol{h}_{\text{real}} = \text{vec}(\Re(\boldsymbol{H}_t)), \\ \boldsymbol{h}_{\text{imag}} = \text{vec}(\Im(\boldsymbol{H}_t)), \end{cases} \tag{9}$$

resulting in $\boldsymbol{h}_{\text{real}}, \boldsymbol{h}_{\text{imag}} \in \mathbb{R}^{KN}$. Next, we employ a fully connected (FC) layer (i.e., FC_1) to project $\boldsymbol{h}_{\text{real}}$, $\boldsymbol{h}_{\text{imag}}$ into the feature space (with dimension d_{mha}) of the multi-head attention layer and rearrange them into $\boldsymbol{H}_{r_{fc1}}, \boldsymbol{H}_{i_{fc1}} \in \mathbb{R}^{K\times d_{mha}}$, respectively. This transformation facilitates initial feature extraction by the multi-head attention mechanism [37]. After processing through the multi-head attention layer, we merge and rearrange the real and imaginary components into the unified representation $\boldsymbol{h}_{mer} \in \mathbb{R}^{2Kd_{mha}}$, which is then fed into FC_2 to effectively map the multi-head attention outputs to the feature dimension (d_{llm}) of the pre-trained LLM. Concurrently, the output of FC_2 is reshaped as $\boldsymbol{H}_{fc2} \in \mathbb{R}^{Nn\times d_{llm}}$, ensuring alignment with the pre-trained LLM input format.

2) Multi-head attention: To enhance the model's representation capability for CSI, we utilize multi-head attention [37] to simultaneously capture and learn distinct features from multiple attention heads, see Fig. 3. Specifically, before being fed into the pre-trained LLM, the CSI is modeled through the multi-head attention mechanism, allowing each head to learn statistical correlations and geometric structures across port and user dimensions. Meanwhile, the CSI is mapped to a feature space consistent with that of the pre-trained LLM, achieving effective feature extraction and dimensional alignment. Recall that within our proposed framework, the multi-head attention layer receives inputs $\boldsymbol{H}_{r_{fc1}}$ and $\boldsymbol{H}_{i_{fc1}}$. Taking $\boldsymbol{H}_{r_{fc1}}$ as an example ($\boldsymbol{H}_{i_{fc1}}$ undergoes identical processing), we first employ linear layers to transform $\boldsymbol{H}_{r_{fc1}}$ into *Query* ($\boldsymbol{Q}$), *Key* ($\boldsymbol{K}$), and *Value* ($\boldsymbol{V}$) matrices, as

$$\begin{cases} \boldsymbol{Q}_{r_m} = \boldsymbol{H}_{r_{fc1}}\boldsymbol{U}_m^Q, \\ \boldsymbol{K}_{r_m} = \boldsymbol{H}_{r_{fc1}}\boldsymbol{U}_m^K, \\ \boldsymbol{V}_{r_m} = \boldsymbol{H}_{r_{fc1}}\boldsymbol{U}_m^V, \end{cases} \tag{10}$$

where $\boldsymbol{U}_m^Q, \boldsymbol{U}_m^K, \boldsymbol{U}_m^V \in \mathbb{R}^{d_{mha}\times d_h}$ denotes the transformation matrix for the m-th attention head with $m \in \{1, \ldots, M\}$, and $d_h = d_{mha}/M$ denotes the dimension of each head.

For the m-th attention head, we first compute the scaled dot-product attention scores between $\boldsymbol{Q}_{r_m}$ and $\boldsymbol{K}_{r_m}$, followed by applying the softmax function to produce normalized attention weights. These attention weights are then multiplied with the corresponding value matrix $\boldsymbol{V}_{r_m}$ to yield the final weighted sum of features. Mathematically, the complete operation for the m-th attention head can be expressed as

$$\text{Attention}_m(\boldsymbol{Q}_{r_m}, \boldsymbol{K}_{r_m}, \boldsymbol{V}_{r_m}) = \text{softmax}\left(\frac{\boldsymbol{Q}_{r_m}\boldsymbol{K}_{r_m}^T}{\sqrt{d_h}}\right)\boldsymbol{V}_{r_m}, \tag{11}$$

where the scaling factor $\sqrt{d_h}$ is utilized to alleviate gradient instability issues such as explosion or vanishing. After processing across all heads, the results are concatenated and mapped through a linear transformation, as

$$\begin{aligned} &\text{MultiHead}(\boldsymbol{Q_r}, \boldsymbol{K_r}, \boldsymbol{V_r}) = \\ &\quad \text{Concat}\left(\text{Attention}_1(\boldsymbol{Q}_{r_1}, \boldsymbol{K}_{r_1}, \boldsymbol{V}_{r_1}), \ldots,\right. \\ &\quad \left.\ldots, \text{Attention}_M(\boldsymbol{Q}_{r_M}, \boldsymbol{K}_{r_M}, \boldsymbol{V}_{r_M})\right)\boldsymbol{U}^O, \end{aligned} \tag{12}$$

where $\boldsymbol{U}^O \in \mathbb{R}^{Md_h\times d_{mha}}$ denotes a learnable weight matrix. By processing multiple attention heads in parallel, the multi-head attention mechanism can effectively learn representations from diverse subspaces within the CSI data.

3) Positional embedding: Following the initial feature extraction from the CSI data, we incorporate positional embedding [37] to provide sequential information to the pre-trained LLM. The positional embedding is mathematically represented as

$$\begin{cases} \boldsymbol{H}_{pe}(i, 2j) = \sin\left(\dfrac{i}{10000^{2j/d_{llm}}}\right), \\ \boldsymbol{H}_{pe}(i, 2j+1) = \cos\left(\dfrac{i}{10000^{2j/d_{llm}}}\right), \end{cases} \tag{13}$$

where i indicates the position within the input sequence, j represents the dimension index within the embedding, and resulting in $\boldsymbol{H}_{pe} \in \mathbb{R}^{Nn\times d_{llm}}$. Note that this positional embedding does not contain any learnable parameters. Sub-

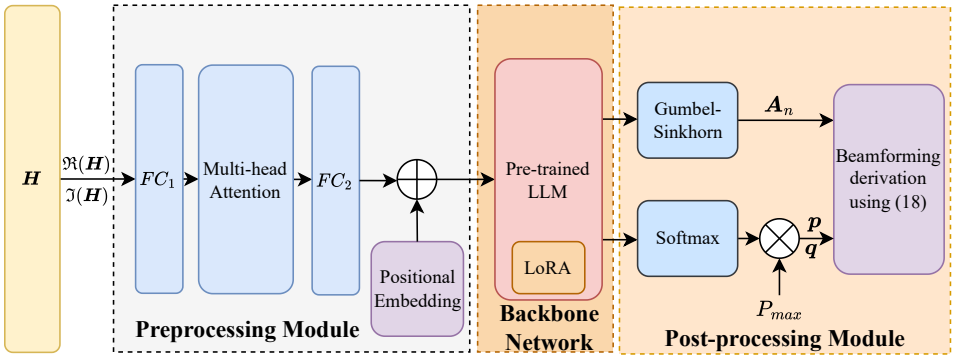

Fig. 2. The overall learning framework for joint optimization of port selection and beamforming.

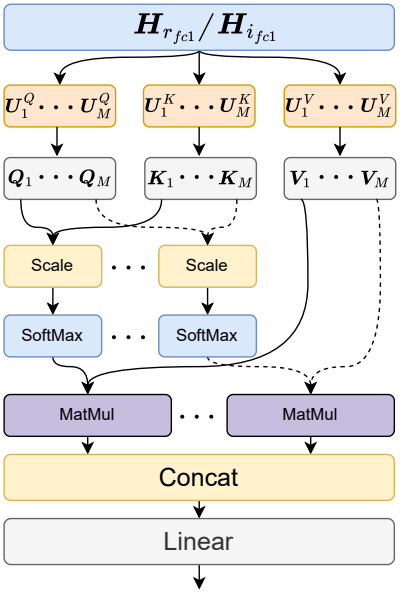

Fig. 3. The multi-head attention mechanism.

sequently, the positional embedding $\boldsymbol{H}_{pe}$ is added to the feature embeddings $\boldsymbol{H}_{fc2}$, yielding the final input embedding $\boldsymbol{H}_{em} \in \mathbb{R}^{Nn \times d_{llm}}$ for the pre-trained LLM, given by

$$\boldsymbol{H}_{em} = \boldsymbol{H}_{fc2} + \boldsymbol{H}_{pe}. \tag{14}$$

This ensures that the pre-trained LLM receives embeddings containing both extracted feature information and positional information necessary for effectively processing CSI data.

B. Backbone Network

Following the generation of input embeddings, GPT-2 [82] is employed as the backbone pre-trained LLM architecture, as depicted in Fig. 4. GPT-2 is an autoregressive language model based on the transformer decoder structure, comprising an embedding layer followed by N_L stacked transformer decoder blocks. Each transformer decoder block consists of a multi-head masked attention, layer normalization, and a position-wise feed forward network (FFN) composed of two FC layers. Layer normalization is utilized to stabilize the training process and accelerate convergence. The multi-head masked attention mechanism is analogous to the previously discussed multi-head attention, but it incorporates mask attention [83] to enforce the autoregressive characteristic (i.e., each position can only rely on information from itself and preceding positions) of the GPT-2. The stacked multi-head masked attention layers leverage historical states and preprocessing feature information to output the port selection matrix and power allocation factors in parallel, enabling decision inference and joint optimization. The FFN is used to provide nonlinear transformation capabilities, represented as

$$\mathrm{FFN}(\boldsymbol{x}) = \mathrm{ReLU}(\boldsymbol{x}\boldsymbol{U}_1 + \boldsymbol{b}_1)\boldsymbol{U}_2 + \boldsymbol{b}_2, \tag{15}$$

where $\boldsymbol{U}_1, \boldsymbol{U}_2$ represent learnable weight matrices, $\boldsymbol{b}_1, \boldsymbol{b}_2$ are corresponding bias vectors, and $\mathrm{ReLU}(\cdot)$ denotes the Rectified Linear Unit activation function.

To fully leverage the generalization capabilities of GPT-2, we perform fine-tuning to jointly address the tasks of port selection and beamforming optimization. Given GPT-2's substantial parameter scale, full parameter fine-tuning incurs significant computational and memory overhead. Hence, we use the LoRA technique [33] to effectively reduce computational costs and GPU memory consumption during fine-tuning. LoRA operates by freezing parameters of the pre-trained LLM and introducing low-rank decompositions to simulate the parameter updates. Specifically, we apply LoRA to the $\boldsymbol{Q}_t$ and $\boldsymbol{V}_t$ matrices of the t-th head within GPT-2's multi-head masked attention layers, which can be expressed as

$$\begin{cases} \mathrm{LoRA}(\boldsymbol{Q}_t) = \boldsymbol{L}_t^Q + \boldsymbol{B}_t^Q \boldsymbol{A}_t^Q, \\ \mathrm{LoRA}(\boldsymbol{V}_t) = \boldsymbol{L}_t^V + \boldsymbol{B}_t^V \boldsymbol{A}_t^V, \end{cases} \tag{16}$$

where $\boldsymbol{L}_t^Q, \boldsymbol{L}_t^V$ are the original frozen weight matrices for $\boldsymbol{Q}_t$ and $\boldsymbol{V}_t$, respectively. The matrices $\boldsymbol{A}^Q, \boldsymbol{A}^V \in \mathbb{R}^{r \times d_{llm}}$ and $\boldsymbol{B}^Q, \boldsymbol{B}^V \in \mathbb{R}^{d_{llm} \times r}$ represent the low-rank decomposition of the weight update matrix, with $r \ll d_{llm}$. Additionally, $\boldsymbol{A}^Q$ and $\boldsymbol{A}^V$ are initialized from a Gaussian distribution, while $\boldsymbol{B}^Q$ and $\boldsymbol{B}^V$ are initialized to zeros, ensuring that the network is equivalent to the pre-trained model at initialization, i.e., the weight update matrix is initially zero. The gradients

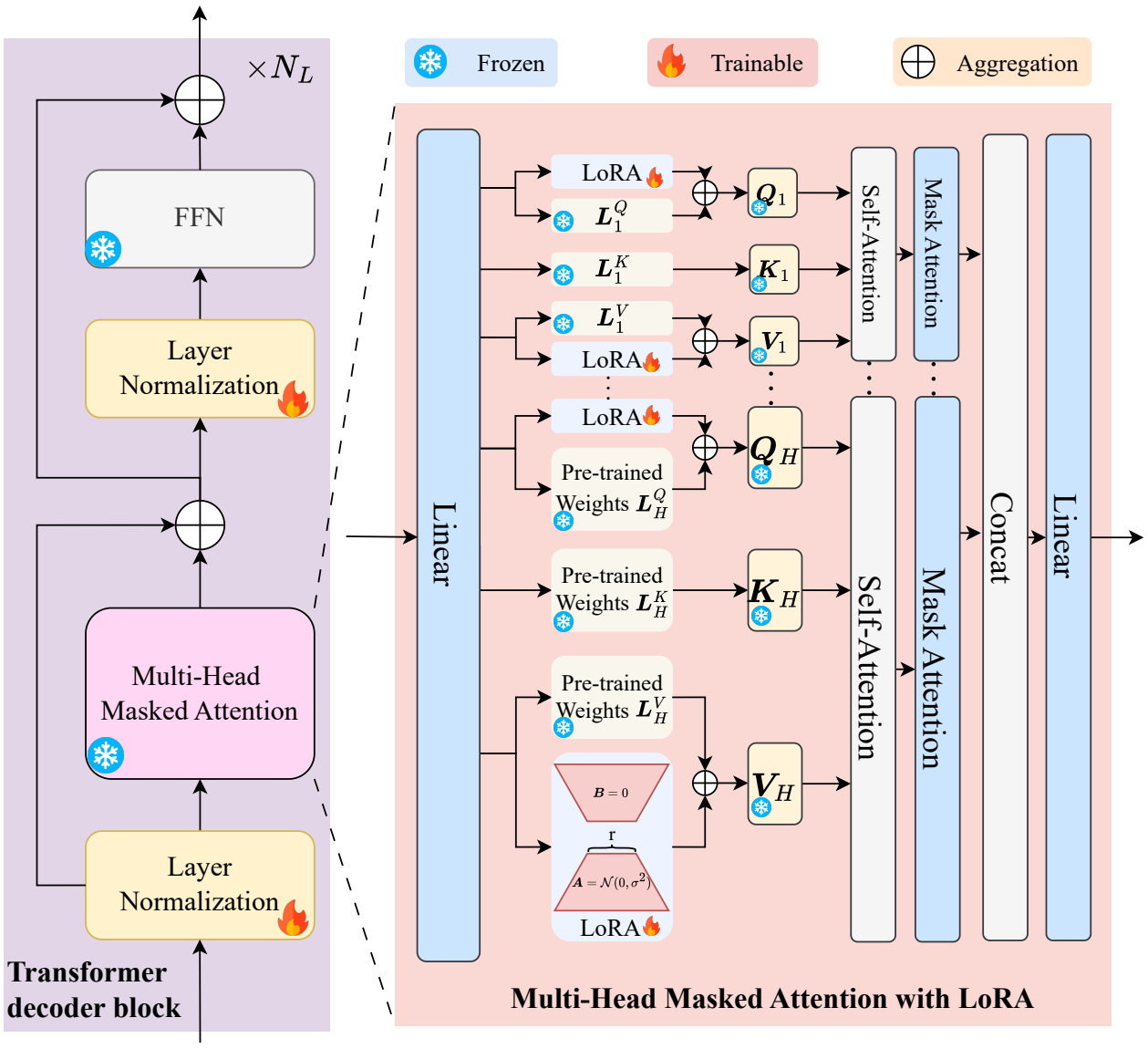

Fig. 4. The backbone network architecture implemented with GPT-2.

of all pre-trained parameters are disabled (set to False) and excluded from the optimizer's parameter groups, with only $\boldsymbol{A}^Q, \boldsymbol{B}^Q, \boldsymbol{A}^V, \boldsymbol{B}^V$ included in the optimization process. Consequently, during backpropagation, only $\boldsymbol{A}$ and $\boldsymbol{B}$ receive gradient updates, while the pre-trained weights remain fixed.

During fine-tuning of the pre-trained LLM, we retain its inherent absolute positional embeddings and freeze the positional embedding parameters within the backbone network. Only the parameters associated with the LoRA matrices $(\boldsymbol{A}, \boldsymbol{B})$, as well as Layer Normalization and aggregation, are updated. Moreover, the parameters in multi-head masked attention and FFN layers remain frozen, thereby preserving the pre-trained knowledge and significantly reducing computational demands without compromising model performance.

After processing the preprocessed input embeddings $\boldsymbol{H}_{em}$ through the backbone GPT-2 network with LoRA fine-tuning, the resulting output is given by

$$\boldsymbol{H}_{llm} = \mathrm{GPT2}_{\mathrm{LoRA}}(\boldsymbol{H}_{em}), \tag{17}$$

where $\mathrm{GPT2}_{\mathrm{LoRA}}(\cdot)$ denotes the LoRA-fine-tuned GPT-2 backbone network, and the output tensor $\boldsymbol{H}_{llm} \in \mathbb{R}^{Nn \times d_{llm}}$ matches the expected dimensionality for subsequent processing. Although GPT-2 is adopted here as the backbone model, alternative LLMs, such as LLaMA [84] or DeepSeek [85], may also be considered. However, the choice of backbone model should carefully balance the trade-off between computational training cost and model performance.

C. Post-Processing Module

To fully exploit the general knowledge of the pre-trained LLM, we output the feature representations processed by the pre-trained LLM in parallel, enabling simultaneous optimization of both port selection and beamforming. Furthermore, the structure of the optimal beamforming solution [32] is utilized to derive the beamforming vector, as

$$\boldsymbol{c}_k^* = \sqrt{p_k}\frac{\left(\mathbf{I}_N + \sum_{i=1}^{n}\frac{q_i}{\sigma^2}\boldsymbol{h}_i\boldsymbol{h}_i^\dagger\right)^{-1}\boldsymbol{h}_k}{\left\|\left(\mathbf{I}_N + \sum_{i=1}^{n}\frac{q_i}{\sigma^2}\boldsymbol{h}_i\boldsymbol{h}_i^\dagger\right)^{-1}\boldsymbol{h}_k\right\|}, \forall k, \tag{18}$$

where $\boldsymbol{h}_i$ denotes the CSI of the activated port, and $\sum_k^K p_k = \sum_k^K q_k = P_{\max}$ satisfies the power constraint. We also use $\boldsymbol{p} = [p_1, \ldots, p_K]^T$ to denote the power allocation vector, while $\boldsymbol{q} = [q_1, \ldots, q_K]^T$ determines the beamforming direction. Specifically, we apply two separate linear projection layers to map the output $\boldsymbol{H}_{llm}$ to the port selection vector $\boldsymbol{x}_{port} \in \mathbb{R}^{Nn}$ and the power allocation vector $\boldsymbol{x}_{power} \in \mathbb{R}^{2K}$. The vector $\boldsymbol{x}_{port}$ is then reshaped into $\boldsymbol{X}_{port} \in \mathbb{R}^{n\times N}$, while a sigmoid activation function is applied to $\boldsymbol{x}_{power}$.

Given that port selection is inherently a combinatorial optimization problem involving discrete variables, traditional discrete sampling methods are non-differentiable and therefore incompatible with gradient-based learning in neural networks. To overcome this challenge, we adopt the Gumbel-Sinkhorn technique [35], which enables differentiable approximation of discrete sampling, allowing end-to-end training via backpropagation. To be specific, we first inject Gumbel noise into the rearranged port selection logits $\boldsymbol{X}_{port} \in \mathbb{R}^{n\times N}$, given by

$$\begin{cases} \boldsymbol{X}_{gum} = (\boldsymbol{X}_{port} + \text{Gumbel}) \sim \boldsymbol{\mathcal{G}}.\boldsymbol{\mathcal{M}}., \\ \text{Gumbel} = -\log_e(-\log_e(\varsigma)),\ \varsigma \sim U(0,1), \end{cases} \tag{19}$$

where $\boldsymbol{\mathcal{G}}.\boldsymbol{\mathcal{M}}.$ denotes the Gumbel-Matching distribution, and $\varsigma \sim U(0,1)$ is a variable following the uniform distribution. The above perturbed matrix is then divided by a temperature parameter τ and passed through a softmax function to obtain a relaxed approximation of the categorical distribution

$$\boldsymbol{X}_s = \text{softmax}\left(\frac{\boldsymbol{X}_{gum}}{\tau}\right), \tag{20}$$

where τ controls the degree of relaxation, i.e., higher values yield smoother distributions, while lower values lead to sharper approximations of the discrete outcomes. Subsequently, the relaxed matrix $\boldsymbol{X}_s$ is transformed into a doubly stochastic matrix $\boldsymbol{X}_{sink} \in \mathbb{R}^{n\times N}$ that follows the Gumbel-Sinkhorn distribution via the Sinkhorn operator, which performs iterative row and column normalization. We then execute the argmax operation on $\boldsymbol{X}_{sink}$ along its column dimension to produce a hard output of the activated ports. Note that we only use this hard output during inference. During training, we retain the continuous and differentiable $\boldsymbol{X}_{sink}$ produced by the Gumbel-Sinkhorn procedure to enable gradient backpropagation. The entire computation pipeline during training consists of differentiable operators—including Gumbel noise injection, temperature scaling, and Sinkhorn iterations—thus ensuring end-to-end differentiability.

To the best of our knowledge, the Gumbel-Sinkhorn approach has not previously been employed in neural network architectures designed for communication optimization problems. In this work, we introduce the Gumbel-Sinkhorn method into our proposed framework for port selection. This allows us to convert the discrete combinatorial optimization problem into a differentiable continuous optimization over doubly stochastic matrices. By carefully controlling the temperature parameter τ, we can achieve a gradual annealing from initial soft probabilistic decisions to approximately hard discrete selections. This mechanism ensures that the entire framework remains differentiable, facilitating end-to-end training via standard backpropagation. Compared to existing approaches, such as deep Q-networks or reinforcement learning (RL) employed in binary reconfigurable intelligent surface phase optimization [86], [87], the Gumbel-Sinkhorn method seamlessly integrates with conventional backpropagation workflows, eliminating the need for additional policy gradient computations or RL-specific algorithms. Furthermore, RL-based methods often experience considerable training variability due to reliance on exploration strategies such as the ε-greedy approach, while the Gumbel-Sinkhorn-based method exhibits significantly smoother and more stable training behavior.

After obtaining the power vector $\boldsymbol{x}_{power}$, we rearrange it into $\boldsymbol{X}_p \in \mathbb{R}^{2\times K}$, and then apply row-wise softmax and scale by the maximum transmission power $P_{\max}$ to ensure both $\boldsymbol{p}$ and $\boldsymbol{q}$ satisfy the power constraint. Finally, given the CSI $\boldsymbol{h}_n$ of the activated ports and the power allocation factors $\boldsymbol{p}, \boldsymbol{q}$, we derive the beamforming vector according to (18). This end-to-end framework thus accepts raw CSI as input and outputs the beamforming vector that jointly optimizes discrete port selection and continuous power allocation.

D. Overall Learning Strategy

Given the substantial number of parameters in LLM, we adopt a mini-batch gradient descent strategy [88] during training to prevent GPU memory overflow and ensure computational stability. By reasonably controlling the batch size, an effective balance between computational efficiency and memory utilization can be achieved. Moreover, owing to the NP-hard and nonconvex characteristics of the joint optimization problem in (8), obtaining extensive labeled optimal solutions is computationally prohibitive. Therefore, we adopt an unsupervised learning manner, directly employing the system sum rate as our loss function, formally defined as

$$\mathcal{L} = -\frac{1}{B}\sum_{b=1}^{B}\sum_{k=1}^{K}\log_2\left(1 + \gamma_k\left(\boldsymbol{\mathcal{A}}_n^{(b)}, \boldsymbol{c}_k^{(b)}\right)\right), \tag{21}$$

where B represents the mini-batch size, and $\boldsymbol{\mathcal{A}}_n^{(b)}$ and $\boldsymbol{c}_k^{(b)}$ represent the port selection matrix and beamforming vector of the b-th batch, respectively.

In each training epoch, we randomly sample a mini-batch of size B from the training set and perform one forward and backward propagation process as follows. The raw CSI data $\boldsymbol{H}^{(b)}$ is fed into the preprocessing module to output the initial feature representations $\boldsymbol{H}_{em}^{(b)}$. Subsequently, the backbone network outputs port selection $\boldsymbol{\mathcal{A}}_n^{(b)}$ and power allocation

vectors $\boldsymbol{p}^{(b)}, \boldsymbol{q}^{(b)}$ in parallel. Then, within the post-processing module, the beamforming vectors $\boldsymbol{c}_k^{(b)}$ are derived jointly from $\boldsymbol{\mathcal{A}}_n^{(b)}$, $\boldsymbol{p}^{(b)}$ and $\boldsymbol{q}^{(b)}$ using (18). For all trainable parameters, encompassing parameters within the preprocessing module, LoRA low-rank adaptation matrices and layer normalization, we use the Adam optimizer [89] for updates. During the testing stage, we load the trained model parameters and perform forward inference on an independent test dataset using the same mini-batch procedure without computing gradients, and take the average system sum rate as the evaluation metric.

V. Numerical Results

In this section, we present extensive computer simulations to evaluate the performance of our proposed LLM-based joint port selection and beamforming method.

A. Parameter Settings and Benchmarks

Following [67], we consider a 2D FAS deployed at the BS, with $N_x = N_y$ and $W_x = W_y$. Unless otherwise specified, the multiuser Tx-MISO-FAS parameters are set as follows [90]. We set the number of users $K = 3$, power budget $P_{\max} = 20$ dBm, noise power spectral density of -174 dBm/Hz, bandwidth of 10 MHz, carrier frequency $f_c = 2$ GHz, the number of ports $N = 4 \times 4$, the number of active ports $n = 4$, and antenna area $W = 2\lambda \times 2\lambda$. In addition, the path loss between each user and the BS is modeled as $128.1 + 37.6 \log_{10}(d)$, where d is in kilometers. We have fixed $d = 0.2$ km as the default setting if not stated otherwise.

All experiments are performed on a workstation equipped with an Intel i9-14900K CPU and an NVIDIA GeForce RTX 4090 GPU, and are implemented using the PyTorch DL framework with Python 3.12. In the preprocessing module, the multi-head attention mechanism uses an embedding dimension of $d_{mha} = 768$, and $M = 8$ attention heads. The backbone network employs a lightweight version of the GPT-2 model with $d_{llm} = 768$ feature dimensions, utilizing only the initial $N_L = 6$ layers of the pre-trained GPT-2 architecture. The rank of the LoRA low-rank matrices is set to $r = 4$ [79]. In the post-processing module, we apply exponential annealing to the temperature coefficient τ in Gumbel-Sinkhorn, updating τ at each training epoch according to $\tau = \max(0.1, 0.95^{epoch})$, where $epoch$ denotes the number of current epoch. The number of Sinkhorn iterations is set to 10. Additionally, the learning rate of Adam optimizer is 0.000001.

We generated $10,000$ channel data samples for the training set and $1,000$ samples for the test set. 20% of the training data was reserved as a validation set. Meanwhile, the early stopping mechanism was employed to monitor validation performance, terminating training if no improvement was observed for 10 consecutive epochs. The model was trained for up to 200 epochs with a batch size of 100.

The proposed method is evaluated by the sum rate across all users and compared against the following benchmark methods.

- **Random:** We perform port selection and beamforming optimization in a sequential manner. First, a random function directly generates the port selection results. Then, a multi-layer perceptron (MLP) maps the feature representation to the dimensions of the power allocation vector, and ultimately the beamforming vectors are derived.
- **CNN [36]:** Similarly, we sequentially optimize port selection and beamforming. For both port selection and beamforming, we first apply a 2D convolutional layer to extract feature information, and then apply a FC layer to map the features to the corresponding output dimensions.
- **LLM-sequential:** Port selection and beamforming are also performed sequentially. We first use a pre-trained LLM to output the port selection. Then, we use a 2D convolutional layer and a FC layer to produce the power allocation vector, and finally derive the beamforming vectors.
- **Transformer [37]:** In this case, we replace the backbone network in our proposed framework with a transformer model consisting of an 8-head multi-head attention module, with parameter settings identical to those of the multi-head attention in the preprocessing module.

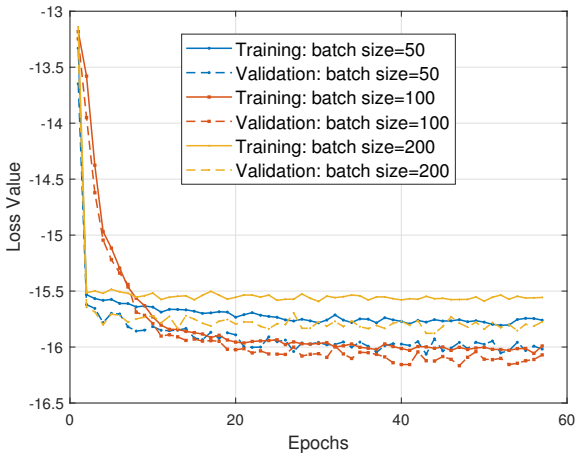

Fig. 5. Convergence performance for different batch size.

B. Performance Comparison

We first evaluated the convergence performance of the proposed method under varying batch sizes (bs), as illustrated in Fig. 5. Both the training and validation curves decrease rapidly during the first 5-10 epochs, after which the loss values gradually stabilize. Throughout training, the gap between the two curves remains small and evolves synchronously, with the validation loss continuously improving and then leveling off. The sharp decline in the early stage corresponds to the model beginning to learn coarse-grained patterns of port selection and power allocation. As the Gumbel-Sinkhorn temperature is annealed, the loss decreases slowly and converges to a stable value, indicating that the selected port set and beamforming factors have largely stabilized. Furthermore, the training curve with $bs = 100$ maintains the lowest loss from about the 10-th epoch onward and achieves the best final convergence performance, outperforming both $bs = 50$ and $bs = 200$. Although $bs = 50$ consumes the least GPU memory, it requires more epochs to reach a comparable loss level and shows slightly higher instability. Conversely, $bs = 200$ incurs higher memory

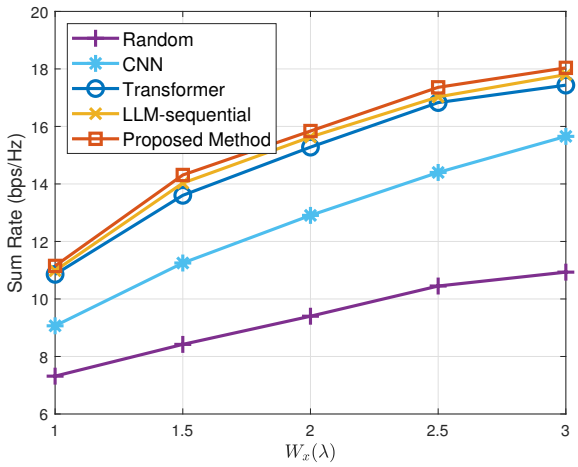

Fig. 6. The sum rate versus the array size W_x.

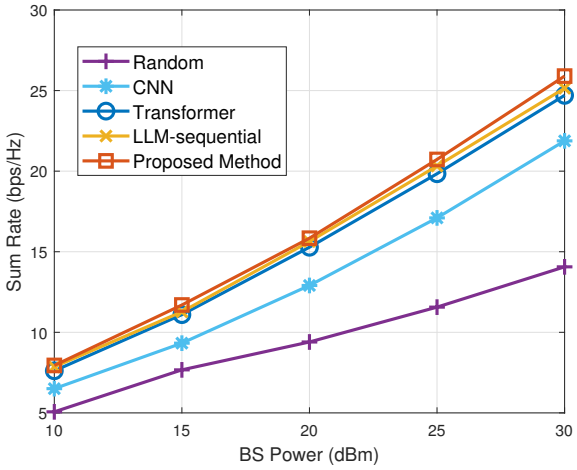

Fig. 7. The sum rate versus the BS power $P_{\max}$.

consumption and exhibits marginally worse loss performance. Therefore, we adopt $bs = 100$ as a balanced choice between computational efficiency and memory utilization.

In Fig. 6, we illustrate the sum rate versus the array size of W_x for different methods. As can be observed, the sum rate of all methods increases with a larger array size W_x, since a larger W_x reduces channel correlation and therefore enhances the spatial diversity gain. Notably, our proposed method consistently achieves the highest sum rate across all evaluated values of W_x. Specifically, compared to random port selection, optimizing port position provides additional spatial diversity benefits, leading to significant performance improvements. Additionally, conventional DL architectures such as CNN and Transformer exhibit limited capability in effectively extracting meaningful features for the high-dimensional port selection problem. In contrast, the enhanced representational capacity of LLM substantially alleviates this limitation. Moreover, serial methods perform sequential optimizations, i.e., first port selection and then perform beamforming, which might yield locally optimal solutions in each step, but fail to achieve global optimality for the overall system performance. Our proposed parallel output method, however, capitalizes on the multi-task learning capabilities of LLM to simultaneously generate the components (port selection and beamforming parameters) required to maximize the overall sum rate objective, resulting in improved performance gains. For example, at $W_x = 2\lambda$, the sum rate of our proposed method outperforms the Random, CNN, Transformer, and LLM-sequential methods by 6.44 bps/Hz, 2.93 bps/Hz, 0.56 bps/Hz, and 0.21 bps/Hz, respectively. By fully leveraging the powerful feature representation and multi-task learning capabilities of LLMs, our proposed parallel optimization method consistently delivers superior performance improvements across various array sizes.

Next, Fig. 7 provides the sum rate results versus the BS power. As the transmission power at the BS increases, all methods exhibit enhancement in the sum rate. This improvement arises because higher transmission power strengthens the received signal at the users, thus augmenting the overall channel capacity. Utilizing the multi-task parallel capability of LLM for jointly optimizing port selection and beamforming decisions, our proposed method achieves the highest sum rate performance under various power conditions. When $P_{\max} = 30$ dBm, our proposed method achieves sum rate improvements of 11.83 bps/Hz, 4.01 bps/Hz, 1.2 bps/Hz and 0.75 bps/Hz compared to the Random, CNN, Transformer and LLM-sequential methods, respectively. The reasons are twofold. First, the rich representation capability of the pre-trained LLM effectively maps complex CSI into high-dimensional features, overcoming the limitations of traditional deep neural networks, which struggle to adequately extract features in high-dimensional combinational spaces. Additionally, compared to the conventional sequential two-stage approaches, our proposed method simultaneously executes port selection and beamforming within the pre-trained LLM framework. This parallelization allows the method to simultaneously attend to the global CSI information relevant to both port selection and power allocation, fully capturing their coupling relationship and thus achieving more coordinated decisions.

Then, we fixed a randomly selected channel sample from the test set and presented, in Table II, the selected port sets for different numbers of activated ports across various methods. The port indices follow the mapping rule defined in the paper. The Random method is excluded since its port selection is randomly assigned rather than obtained through learning or optimization. For each method producing a port set $\boldsymbol{\mathcal{A}}_n$ at a given n, all methods employ the same model-based optimal beamforming structure (as given in equation (18)) to derive the beamforming vectors. In our experiments, the CNN and LLM-sequential methods perform port selection and beamforming derivation sequentially, using the same network structure for power factor $(\boldsymbol{p}, \boldsymbol{q})$ learning. By contrast, both the Transformer and the proposed method conduct parallel joint optimization of port selection and beamforming, sharing an identical power factor learning network structure. Hence, the performance differences between the CNN/LLM-sequential group and the Transformer/proposed group mainly stem from the port selection quality rather than differences in beamforming design. Furthermore, the difference between the LLM-sequential and the proposed method reflects the advantage of our parallel output framework. Finally, by performing forward inference

TABLE II
THE SELECTED PORT SETS FOR DIFFERENT METHODS

No. of n / Methods	$n=3$	$n=4$	$n=5$	$n=6$	$n=7$
CNN	{2,8,11}	{2,3,5,10}	{1,5,12,14,15}	{4,8,10,13,15,16}	{2,3,10,11,12,13,15}
Transformer	{3,9,16}	{4,5,14,16}	{3,5,6,12,13}	{1,3,9,10,12,15}	{2,4,5,9,12,15,16}
LLM-sequential	{3,12,13}	{4,5,13,16}	{1,3,10,12,15}	{2,3,6,9,12,14}	{1,4,6,9,11,14,16}
Proposed Method	{4,9,16}	{1,3,13,16}	{1,4,6,12,14}	{2,4,5,7,12,14}	{1,4,7,9,11,14,16}

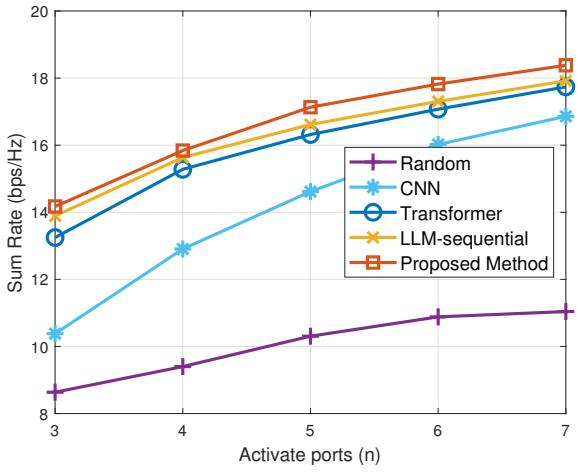

Fig. 8. The sum rate versus the number of activated ports n.

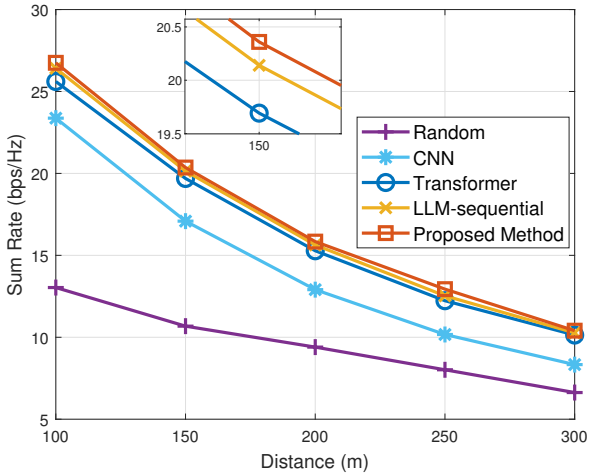

Fig. 9. The sum rate versus the distance between the user and the BS.

over all test samples and averaging the results, we obtain the sum rate performance shown in Fig. 8. These results clearly indicate that the observed performance improvement of the proposed method originates from its superior port selection and parallel joint optimization capability, rather than solely from the beamforming optimization component.

Fig. 8 demonstrates the sum rate versus the number of activated ports n. It is evident that the sum rate for all considered methods significantly increases with n. This improvement arises because activating more ports provides additional spatial DoF, thus enlarging the overall system capacity. For instance, at $n = 6$, the proposed method achieves a sum rate gain of approximately 6.93 bps/Hz over the Random benchmark, and also demonstrates notable improvements compared with other benchmark schemes. This can be explained by recognizing that the proposed parallel output strategy enables joint optimization of port selection and power allocation factors in a single inference, avoiding the suboptimality of traditional sequential decision-making. This effectively exploits the close coupling between the two, thereby fully leveraging the advantages of LLM multi-task learning and high-dimensional feature representation, thus achieving improved sum rate performance.

Fig. 9 shows the sum rate versus the distance between the user and the BS. With increasing distance, the signal path loss intensifies, reducing the signal strength and degrading the sum rate for all methods. But the proposed method consistently achieves the best performance among the compared methods. First, the Random method indicates the worst performance due to its lack of adaptability and optimization capacity regarding channel variations. Also, although the CNN method slightly outperforms the Random approach, its limited ability to extract spatial features still results in relatively poor performance. Leveraging the self-attention mechanism's capability to capture variations in channel characteristics induced by changes in the distance between users and the BS, the Transformer method exhibits considerable improvement over CNN. However, its overall performance is not as good as that of the LLM approach. With superior feature extraction and decision-making capabilities, the LLM-sequential method outperforms CNN and Transformer overall. Sequential optimization restricts the method's ability to fully exploit the coupling relationship between port selection and power allocation, leading to inferior performance compared to the proposed parallel approach across all evaluated distances. For instance, when $d = 150$ m, the proposed method demonstrates performance gains of 9.68 bps/Hz, 3.27 bps/Hz, 0.67 bps/Hz, and 0.22 bps/Hz over the Random, CNN, Transformer, and LLM-sequential methods, respectively. Our proposed method jointly outputs port selection and power allocation factors in a parallel manner, effectively capturing channel characteristic variations induced by changing distances and thereby achieving better optimization.

Now, Fig. 10 shows the sum rate performance with respect to the increasing number of ports. The sum rate of all methods improves as the number of antenna ports increases. This is because more antenna ports provide richer spatial DoF, enabling the system to perform spatial multiplexing and interference suppression more effectively. The results reveal

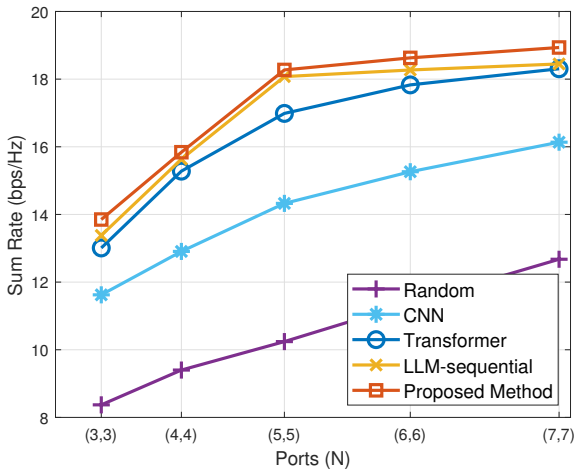

Fig. 10. The sum rate versus the number of ports N.

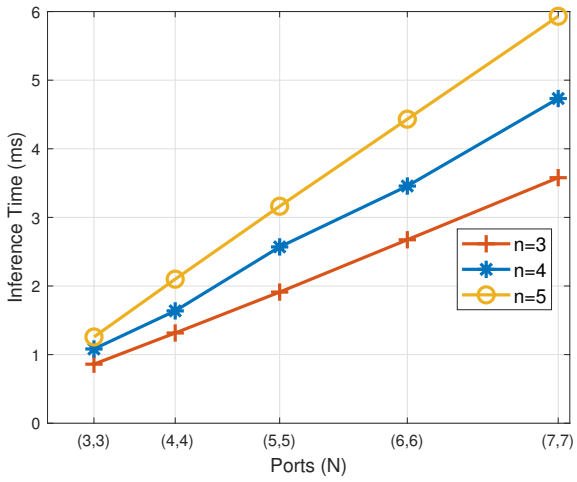

Fig. 11. The inference time versus the number of ports N for different numbers of activated ports n.

that the proposed method consistently achieves the highest performance among all compared methods. When $N = 36$, i.e., (6,6), the sum rate of the proposed method surpasses other methods by approximately $0.36 - 7.25$ bps/Hz. Specifically, random port selection fails to take advantage of the increased DoF, resulting in its worst performance. Although CNN and Transformer methods offer improvements in spatial information extraction, these models remain limited under scenarios involving larger numbers of ports. Leveraging the capabilities of pre-trained LLMs, the LLM-based methods achieve improved performance. However, the LLM-sequential approach restricts complete joint optimization of port selection and beamforming, resulting in performance that consistently lags behind the proposed parallel method. The proposed parallel method enables simultaneous optimization of both port selection and beamforming using the generalization and multi-task capabilities of the LLM, in contrast to the LLM-sequential method which employs LLM only for optimizing port selection. This parallel output framework facilitates shared utilization of the LLM's underlying feature representations for both port selection and beamforming components. As such, both components can be simultaneously optimized within the same feature space. Through joint training, the mutual reinforcement between these components effectively captures complex CSI data relationships, enabling the proposed method to better exploit spatial DoF and improving sum rate performance.

Finally, Fig. 11 presents the inference time of the proposed method with respect to the number of ports, N, and the number of activated ports, n. As either the number of ports or active ports increases, the data dimensionality and computational complexity correspondingly rise, resulting in increased inference time. When the antenna port size expands from $(3, 3)$ to $(7, 7)$, the inference time increases approximately linearly. This demonstrates that, although more ports incur greater computational demands, the computational complexity of the proposed approach rises steadily rather than exponentially with the array size. For instance, when the total number of ports is expanded to 49, the inference time remains under 6 ms even with 5 active ports. Therefore, the proposed method can efficiently handle complex wireless communication scenarios with very low inference latency in practical deployments, effectively meeting the stringent real-time resource optimization requirements of future wireless networks.

VI. Conclusion

This paper proposed a novel LLM-based parallel learning framework for jointly optimizing port selection and beamforming in multiuser Tx-MISO-FAS downlink communication system. Our goal is to maximize the sum rate while satisfying the BS transmission power and the port activation constraints. Unlink traditional sequential two-stage methods, i.e., first port selection then beamforming, we fully leveraged the powerful multi-task learning and high-dimensional mapping capabilities of LLM to achieve simultaneous parallel optimization of port selection indices and beamforming power factors. Specifically, we utilized the GPT-2 model as the backbone network for pre-trained LLM and applied the LoRA technique to fine-tune it, thus preserving its general representation capabilities while reducing training computational overhead. Furthermore, we employed a Gumbel-Sinkhorn-based stochastic relaxation to convert discrete port selection into continuous, differentiable operations, enabling the entire framework to support end-to-end training via gradient descent. Numerical results indicated that the proposed method outperforms existing state-of-the-art benchmarks in terms of sum rate. Additionally, the lower inference latency also showed the efficiency of the proposed method. Future research will explore exploiting spatial correlations among FAS ports for channel reconstruction, thereby reducing CSI acquisition and feedback costs. On the other hand, considering channel estimation errors and hardware non-idealities in real-world environments, robust joint optimization under imperfect CSI is also of importance.

References

[1] K. N. R. S. V. Prasad, E. Hossain, and V. K. Bhargava, "Energy efficiency in massive MIMO-based 5G networks: Opportunities and challenges," *IEEE Trans. Wireless Commun.*, vol. 24, no. 3, pp. 86–94, Jun. 2017.

[2] R. M. Dreifuerst and R. W. Heath, "Massive MIMO in 5G: How beamforming, codebooks, and feedback enable larger arrays," *IEEE Commun. Mag.*, vol. 61, no. 12, pp. 18–23, Dec. 2023.

[3] D. Astely, P. Von Butovitsch, S. Faxér and E. Larsson, "Meeting 5G network requirements with Massive MIMO," *Ericsson Technol. Rev.*, vol. 2022, no. 1, pp. 2–11, Feb. 2022.

[4] H. Q. Ngo, A. Ashikhmin, H. Yang, E. G. Larsson and T. L. Marzetta, "Cell-free massive MIMO versus small cells," *IEEE Trans. Wireless Commun.*, vol. 16, no. 3, pp. 1834–1850, Mar. 2017.

[5] F. Tariq *et al.*, "A speculative study on 6G," *IEEE Wireless Commun.*, vol. 27, no. 4, pp. 118–125, Aug. 2020.

[6] C.-X. Wang *et al.*, "On the road to 6G: Visions, requirements, key technologies, and testbeds," *IEEE Commun. Surveys Tuts.*, vol. 25, no. 2, pp. 905–974, 2nd Quarter. 2023.

[7] K. K. Wong, K.-F. Tong, Y. Shen, Y. Chen, and Y. Zhang, "Bruce Lee-inspired fluid antenna system: Six research topics and the potentials for 6G," *Frontiers Commun. Netw.*, vol. 3, Art. no. 853416, Mar. 2022.

[8] W. K. New *et al.*, "A tutorial on fluid antenna system for 6G networks: Encompassing communication theory, optimization methods and hardware designs," *IEEE Commun. Surv. Tuts.*, early access, doi:10.1109/COMST.2024.3498855, 2024.

[9] K. K. Wong, A. Shojaeifard, K. F. Tong, and Y. Zhang, "Performance limits of fluid antenna systems," *IEEE Commun. Lett.*, vol. 24, no. 11, pp. 2469–2472, Nov. 2020.

[10] K. K. Wong, A. Shojaeifard, K. F. Tong, and Y. Zhang, "Fluid antenna systems," *IEEE Trans. Wireless Commun.*, vol. 20, no. 3, pp. 1950–1962, Mar. 2021.

[11] S. Basbug, "Design and synthesis of antenna array with movable elements along semicircular paths," *IEEE Antennas Wireless Propag. Lett.*, vol. 16, pp. 3059–3062, Oct. 2017.

[12] L. Zhu, and K. K. Wong, "Historical review of fluid antennas and movable antennas," *arXiv preprint*, arXiv:2401.02362v2, Jan. 2024.

[13] Y. Shen *et al.*, "Design and implementation of mmWave surface wave enabled fluid antennas and experimental results for fluid antenna multiple access," *arXiv preprint*, arXiv:2405.09663, May. 2024.

[14] J. Zhang *et al.*, "A novel pixel-based reconfigurable antenna applied in fluid antenna systems with high switching speed," *IEEE Open J. Antennas & Propag.*, vol. 6, no. 1, pp. 212-228, Feb. 2025.

[15] B. Liu, K.-F. Tong, K. K. Wong, C.-B. Chae, and H. Wong, "Programmable meta-fluid antenna for spatial multiplexing in fast fluctuating radio channels," *Optics Express*, vol. 33, no. 13, pp. 28898–28915, 2025.

[16] B. Liu, K. F. Tong, K. K. Wong, C.-B. Chae, and H. Wong, "Be water, my antennas: Riding on radio wave fluctuation in nature for spatial multiplexing using programmable meta-fluid antenna," *arXiv preprint*, arXiv:2502.04693, 2025.

[17] C. Wang *et al.*, "AI-empowered fluid antenna systems: Opportunities, challenges, and future directions," *IEEE Wireless Commun.*, vol. 31, no. 5, pp. 34–41, Oct. 2024.

[18] C. Wang, K. K. Wong, Z. Li, L. Jin, and C.-B. Chae, "Large language model empowered design of fluid antenna systems: Challenges, frameworks, and case studies for 6G," to appear in *IEEE Wireless Commun.*, 2025.

[19] W.-J. Lu *et al.*, "Fluid antennas: Reshaping intrinsic properties for flexible radiation characteristics in intelligent wireless networks," *IEEE Commun. Mag.*, vol. 63, no. 5, pp. 40–45, May 2025.

[20] T. Wu *et al.*, "Fluid antenna systems enabling 6G: Principles, applications, and research directions," *arXiv preprint*, arXiv:2412.03839, Dec. 2024.

[21] W. K. New, K. K. Wong, H. Xu, K. F. Tong and C. B. Chae, "Fluid antenna system: New insights on outage probability and diversity gain," *IEEE Trans. Wireless Commun.*, vol. 23, no. 1, pp. 128–140, Jan. 2024.

[22] W. K. New, K. K. Wong, H. Xu, K.-F. Tong and C.-B. Chae, "An information-theoretic characterization of MIMO-FAS: Optimization, diversity-multiplexing tradeoff and q-outage capacity," *IEEE Trans. Wireless Commun.*, vol. 23, no. 6, pp. 5541–5556, Jun. 2024.

[23] K. K. Wong, K.-F. Tong, and C.-B. Chae, "Fluid antenna system–Part-II: Research opportunities," *IEEE Commun. Lett.*, vol. 27, no. 8, pp. 1924–1928, Aug. 2023.

[24] L. Zhang, H. Yang, Y. Zhao, and J. Hu, "Joint port selection and beamforming design for fluid antenna assisted integrated data and energy transfer," *IEEE Wireless Commun. Lett.*, vol. 13, no. 7, pp. 1833–1837, Jul. 2024.

[25] C. Zhang, Y. Xu, S. Peng, X. Guo, X. Ou, H. Hong, D. He, and W. Zhang, "Fluid antenna-aided robust secure transmission for RSMA-ISAC systems," *arXiv preprint*, arXiv:2503.05515, Mar. 2025.

[26] M. Ahmadzadeh *et al.*, "AI-based fluid antenna design for client selection in over-the-air federated learning," *IEEE Internet of Things J.*, vol. 12, no. 20, pp. 42549–42558, Oct. 2025.

[27] H. Gu, L. Zhao, Z. Han, G. Zheng, and S. Song, "AI-Enhanced cloudedge-terminal collaborative network: Survey, applications, and future directions," *IEEE Commun. Surveys Tuts.*, vol. 26, no. 2, pp. 1322–1385, doi:10.1109/COMST.2023.3338153., 2024.

[28] C. Weng, Y. Chen, L. Zhu, and Y. Wang, "Learning-based joint beamforming and antenna movement design for movable antenna systems," *IEEE Wireless Commun. Lett.*, vol. 13, no. 8, pp. 2120–2124, Aug. 2024.

[29] Y. Yao *et al.*, "A survey on large language model (LLM) security and privacy: The good, the bad, and the ugly," *High-Confidence Comput.*, p. 100211, 2024.

[30] M. U. Hadi *et al.*, "Large language models: A comprehensive survey of its applications, challenges, limitations, and future prospects," *Tech. Rep.*, 2023.

[31] B. Liu, X. Liu, S. Gao, X. Cheng, and L. Yang, "LLM4CP: Adapting large language models for channel prediction," *J. Commun. Inf. Netw.*, vol. 9, no. 2, pp. 113–125, Jun. 2024.

[32] E. Björnson, M. Bengtsson and B. Ottersten, "Optimal multiuser transmit beamforming: A difficult problem with a simple solution structure [Lecture Notes]," *IEEE Sig. Process. Mag.*, vol. 31, no. 4, pp. 142–148, Jul. 2014.

[33] E. J. Hu *et al.*, "LoRA: Low-rank adaptation of large language models," in *Proc. Int. Conf. Learning Represent.*, 2022.

[34] O. Vinyals, M. Fortunato, and N. Jaitly, "Pointer networks," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS).*, Red Hook, NY, USA: Curran Associates, pp. 1–9, 2015.

[35] G. Mena, D. Belanger, S. Linderman, and J. Snoek, "Learning latent permutations with Gumbel-Sinkhorn networks," *Proc. Int. Conf. Learning Represent.*, 2018.

[36] K. O'Shea and R. Nash, "An introduction to convolutional neural networks," *arXiv preprint*, arXiv:1511.08458, 2015.

[37] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Conf. Neural Inf. Process. Syst.*, vol. 30, pp. 1–15, 2017.

[38] C. N. Efrem and I. Krikidis, "Transmit and receive antenna port selection for channel capacity maximization in fluid-MIMO systems," *IEEE Wireless Commun. Lett.*, vol. 13, no. 11, pp. 3202–3206, Nov. 2024.

[39] J.-C. Chen, T. L. Cheng, K. K. Wong and H. Shin, "Improved joint transmit and receive port selection for capacity maximization in fluid-MIMO systems," *IEEE Wireless Commun. Lett.*, vol. 14, no. 6, pp. 1693–1697, Jun. 2025.

[40] R. Wang, Y. Chen, Y. Hou, K.-K. Wong, and X. Tao, "Estimation of channel parameters for port selection in millimeter-wave fluid antenna systems," in *Proc. IEEE/CIC Int. Conf. Commun. China (ICCC Workshops)*, pp. 1-6, Aug. 2023.

[41] X. Lai, T. Wu, J. Yao, C. Pan, M. Elkashlan, and K.-K. Wong, "On performance of fluid antenna system using maximum ratio combining," *IEEE Commun. Lett.*, vol. 28, no. 2, pp. 402–406, Feb. 2024.

[42] T. Mao *et al.*, "Joint time scheduling and port activation design for fluid antenna-empowered wireless powered communication networks," *IEEE Internet Things J.*, vol. 12, no. 13, pp. 22904–22914, Jul. 2025.

[43] Y. Chen *et al.*, "Energy-efficiency optimization for slow fluid antenna multiple access using mean-field game," *IEEE Wireless Commun. Lett.*, vol. 13, no. 4, pp. 915–918, Apr. 2024.

[44] Z. Chai, K. K. Wong, K.-F. Tong, Y. Chen, and Y. Zhang, "Port selection for fluid antenna systems," *IEEE Commun. Lett.*, vol. 26, no. 5, pp. 1180–1184, May. 2022.

[45] J. Zou, S. Sun, and C. Wang, "Online learning-induced port selection for fluid antenna in dynamic channel environment," *IEEE Wireless Commun. Lett.*, vol. 13, no. 2, pp. 313–317, Feb. 2024.

[46] S. Zhang *et al.*, "Fast port selection using temporal and spatial correlation for fluid antenna systems," in *Proc. IEEE Stat. Signal Process. Workshop (SSP)*, pp. 95–99, Hanoi, Vietnam, 2-5 Jul. 2023.

[47] M. Eskandari, A. G. Burr, K. Cumanan and K. -K. Wong, "Port selection for fluid antenna systems via conditional generative adversarial networks," *2025 Joint Eur. Conf. Netw. Commun. & 6G Summit (EuCNC/6G Summit)*, Poznan, Poland, June. 2025.

[48] B. Feng, C. Feng, K. -K. Wong and T. Q. S. Quek, "Deep unfolding neural networks for fluid antenna-enhanced vehicular communication," *IEEE Trans. Veh. Technol.*, early access, doi:10.1109/TVT.2025.3559786., 2025.

[49] N. Waqar *et al.*, "Opportunistic fluid antenna multiple access via team-inspired reinforcement learning," *IEEE Trans. Wireless Commun.*, vol. 23, no. 9, pp. 12068–12083, Sep. 2024.

[50] X. Xu *et al.*, "Transformer based collaborative reinforcement learning for fluid antenna system (FAS)-enabled 3D UAV positioning," *arXiv preprint*, arXiv:2507.09094, Jul. 2025.

[51] Alvim, Pedro D *et al.*, "LNN-powered fluid antenna multiple access.," *arXiv preprint*, arXiv:2507.08821, Jun. 2025.

[52] H. Qin *et al.*, "Antenna positioning and beamforming design for fluid antenna-assisted multi-user downlink communications," *IEEE Wireless Commun. Lett.*, vol. 13, no. 4, pp. 1073–1077, Apr. 2024.

[53] Y. Chen *et al.*, "Joint beamforming and antenna design for near-field fluid antenna system," *IEEE Wireless Commun. Lett.*, vol. 14, no. 2, pp. 415–419, Feb. 2025.

[54] T. Liao, W. Guo, H. He, S. Song, J. Zhang, and K. B. Letaief, "Joint beamforming and antenna position optimization for fluid antennaassisted MU-MIMO networks," *arXiv preprint*, arXiv:2503.04040, Mar. 2025.

[55] J. Yao, L. Zhou, T. Wu, M. Jin, C. Huang, and C. Yuen, "FAS vs. ARIS: Which is more important for FAS-ARIS communication systems?," *arXiv preprint*, arXiv:2408.09067, Aug. 2024.

[56] R. Xu *et al.*, "Energy efficient fluid antenna relay (FAR)-assisted wireless communications," *IEEE J. Select. Areas Commun.*, early access, Oct. 2025.

[57] L. Hu *et al.*, "Two-Timescale design for fluid antenna enhanced multiuser MIMO system with imperfect CSI," in *Proc. IEEE 101st Veh. Technol. Conf. (VTC2025-Spring)*, pp. 1–5, Jun. 2025.

[58] L. Zhang *et al.*, "Energy-Efficient port selection and beamforming design for integrated data and energy transfer assisted by fluid antennas," *arXiv preprint*, arXiv:2503.04147, Mar. 2025.

[59] S. Liang *et al.*, "Rate maximization for fluid antenna system assisted semantic communication," *arXiv preprint*, arXiv:2506.22943, Jun. 2025.

[60] J. Zou *et al.*, "Shifting the ISAC trade-off with fluid antenna systems," *IEEE Wireless Commun. Lett.*, vol. 13, no. 12, pp. 3479-3483, Dec. 2024.

[61] T. Hao *et al.*, "Fluid-Antenna enhanced ISAC: joint antenna positioning and dual-functional beamforming design under perfect and imperfect CSI," *IEEE Trans. Veh. Technol.*, early access, doi:10.1109/TVT.2025.3574072, 2025.

[62] H. Xiao *et al.*, "Fluid reconfigurable intelligent surfaces: Joint on-off selection and beamforming with discrete phase shifts," *IEEE Wireless Commun. Lett.*, early access, doi:10.1109/LWC.2025.3587070, 2025.

[63] X. Yang *et al.*, "Rate maximization for UAV-assisted ISAC system with fluid antennas," in *Proc. IEEE/CIC Int. Conf. Commun. China (ICCC Workshops)*, pp. 1-5, Aug. 2025.

[64] A. Salem *et al.*, "Secure ISAC with fluid antenna systems: joint precoding and port selection," *arXiv preprint*, arXiv:2509.26572, Sep. 2025.

[65] C. He *et al.*, "Graph neural network enabled fluid antenna systems: A two-stage approach," *IEEE Trans. Veh. Technol.*, doi:10.1109/TVT.2025.3570319, 2025.

[66] S. Xu *et al.*, "Toward practical fluid antenna systems: co-optimizing hardware and software for port selection and beamforming," *arXiv preprint*, arXiv:2507.14035, Jul. 2025.

[67] C. Wang *et al.*, "Fluid antenna system liberating multiuser MIMO for ISAC via deep reinforcement learning," *IEEE Trans. Wireless Commun.*, vol. 23, no. 9, pp. 10879–10894, Mar. 2024.

[68] S. Yang *et al.*, "Towards intelligent antenna positioning: leveraging DRL for FAS-aided ISAC systems," *IEEE Internet of Things J.*, vol. 12, no. 16, pp. 34615–34618, Aug. 2025.

[69] S. Zhang *et al.*, "Fluid antenna systems enabled by reconfigurable holographic surfaces: beamforming design and experimental validation," *IEEE J. Select. Areas Commun.*, early access, Oct. 2025.

[70] G. Li *et al.*, "Model-based multi-agent reinforcement learning for joint port and precoding optimization in multi-cell fluid antenna system," in *Proc. IEEE 101st Veh. Technol. Conf. (VTC2025-Spring)*, pp. 1–7, Jun. 2025.

[71] T. Zheng and L. Dai, "Large language model enabled multi-task physical layer network," *arXiv preprint*, arXiv:2412.20772v2, Mar. 2024.

[72] X. Liu, S. Gao, B. Liu, X. Cheng and L. Yang, "LLM4WM: Adapting LLM for wireless multi-tasking," *IEEE Trans. Mach. Learn. Commun. Netw.*, early access, doi:10.1109/TMLCN.2025.3585845, 2025.

[73] Z. Xu, T. Zheng and L. Dai, "LLM-Empowered near-field communications for low-altitude economy," *IEEE Trans. Commun.*, early access, doi:10.1109/TCOMM.2025.3581051, 2025.

[74] Y. Sheng, K. Huang, L. Liang, P. Liu, S. Jin, and G. Y. Li, "Beam prediction based on large language models," *IEEE Wireless Commun. Lett.*, vol. 14, no. 5, pp. 1406-1410, May. 2025.

[75] X. Wu *et al.*, "Channel state information extrapolation in fluid antenna systems based on masked language model," in *2024 IEEE International Conference on Communications Workshops (ICC Workshops)*, pp. 1383–1388. Jun. 2024.

[76] H. Li *et al.*, "Joint user association and beamforming design for ISAC networks with large language models," *IEEE Open J. Commun. Soc.*, vol. 6, pp. 7620–7644, Sept. 2025.

[77] Y. Li *et al.*, "BERT4beam: Large AI model enabled generalized beamforming optimization," *arXiv preprint*, arXiv:2509.11056, Sept. 2025.

[78] H. Yang, S. Lambotharan, and M. Derakhshani, "FAS-LLM: Large language model-based channel prediction for OTFS-enabled satellite FAS links," *arXiv preprint*, arXiv:2505.09751, May. 2025.

[79] Y. Zhang *et al.*, "Port-LLM: A port prediction method for fluid antenna based on large language models," *arXiv preprint*, arXiv:2502.09857, Feb. 2025.

[80] X. Gao, O. Edfors, F. Tufvesson, and E. G. Larsson, "Massive MIMO in real propagation environments: Do all antennas contribute equally?" *IEEE Trans. Commun.*, vol. 63, no. 11, pp. 3917-3928, Nov. 2015.

[81] W. C. Jakes and D. C. Cox, *Microwave mobile communications*. Wiley-IEEE press, 1994.

[82] A. Radford *et al.*, "Language models are unsupervised multitask learners," *OpenAI Blog*, vol. 1, no. 8, p. 9, 2019.

[83] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 1290–1299, Jun. 2022.

[84] H. Touvron *et al.*, "LLaMA: Open and efficient foundation language models," *arXiv preprint*, arXiv:2302.13971, 2023.

[85] D. Guo *et al.*, "DeepSeek-r1: Incentivizing reasoning capability in LLMs via reinforcement learning," *arXiv preprint*, arXiv:2501.12948, 2025.

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

[87] K. Stylianopoulos and G. C. Alexandropoulos, "Online ROS configuration learning for arbitrary large numbers of 1-bit phase resolution elements," in *Proc IEEE Int. Workshop Sig. Process. Adv. Wireless Commun. (SPAWC)*, pp. 1–5, Jul. 2022.

[88] S. Khirirat, H. R. Feyzmahdavian, and M. Johansson, "Mini-batch gradient descent: Faster convergence under data sparsity," in *Proc. IEEE 56th Annu. Conf. Decis. Control (CDC)*, pp. 2880–2887, Dec. 2017.

[89] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, pp. 1–15, 2015.

[90] Y. Xu, Y. Chen, Y. Hou, K.-K. Wong, Q. Cui, and X. Tao, "Energy efficiency maximization under delay-outage probability constraints using fluid antenna systems," in *2023 IEEE Statistical Signal Processing Workshop (SSP)*, pp. 105-109, 2023.