Received 2 March 2025; revised 16 April 2025; accepted 27 April 2025. Date of publication 6 May 2025; date of current version 3 June 2025.
Digital Object Identifier 10.1109/OJCOMS.2025.3567495

Large Sequence Model for MIMO Equalization in Fully Decoupled Radio Access Network

KAI YU [1], HAIBO ZHOU [1] (Senior Member, IEEE), YUNTING XU [1], ZONGXI LIU [1], HONGYANG DU [2] (Member, IEEE), AND XUEMIN SHEN [3] (Fellow, IEEE)

(Special Issue on Generative AI and Large Language Models Enhanced 6G Wireless Communication and Sensing)

[1]School of Electronic Science and Engineering, Nanjing University, Nanjing 210023, China
[2]Department of Electrical and Electronic Engineering, The University of Hong Kong, Hong Kong, SAR, China
[3]Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada

CORRESPONDING AUTHOR: H. ZHOU (e-mail: haibozhou@nju.edu.cn).

This work was supported in part by the National Natural Science Foundation Original Exploration Project of China under Grant 62250004; in part by the National Natural Science Foundation of China under Grant 62271244; in part by the Natural Science Fund for Distinguished Young Scholars of Jiangsu Province under Grant BK20220067; and in part by the High-Level Innovation and Entrepreneurship Talent Introduction Program Team of Jiangsu Province under Grant JSSCTD202202.

ABSTRACT Fully decoupled RAN (FD-RAN) aims to improve network performance by decoupling the hardware of base stations (BSs) and enabling flexible cooperation, making it a promising architecture for next-generation wireless networks. With the emergence of artificial intelligence, FD-RAN provides an opportunity to integrate physical-layer signal processing with neural network models. However, conventional deep learning-based multi-input multi-output (MIMO) equalization methods often rely on extensive offline training under fixed channel conditions, resulting in limited generalization to unseen wireless environments. Motivated by the strong generalization ability demonstrated by in-context learning (ICL) in natural language processing, we extend ICL to cooperative MIMO equalization in the FD-RAN framework. In this setup, geographical location information is incorporated as side information to enhance inference accuracy. Lightweight Transformer encoders are deployed at resource-constrained BSs to compress received signals, which are then forwarded to a central unit where a large decoder-only Transformer, adapted from GPT-2, performs equalization. The components are jointly trained to capture channel characteristics effectively. We further evaluate the generalization capability of large models by comparing the proposed ICL-based equalizer against meta-learning baselines. Experimental results show that our method achieves over 29% improvement in normalized mean square error under 8-bit and 12-bit fronthaul constraints compared to an unquantized LMMSE baseline. Moreover, as the pretraining dataset size increases, the ICL-based equalizer consistently outperforms meta-learning approaches, underscoring its scalability and potential for deployment in large-scale, data-driven wireless systems.

INDEX TERMS MIMO equalization, transformer, in-context learning.

I. INTRODUCTION

ARTIFICIAL intelligence-enabled radio access networks (AI-RAN) are increasingly recognized as an indispensable component of next-generation wireless networks [1]. These networks harness the power of AI to optimize resource allocation, network management, and user experience, marking a significant departure from traditional approaches. As the demand for more efficient and scalable wireless systems grows, AI-driven solutions have become crucial in handling the complexity of modern communication environments. Transformer-based sequence models have demonstrated remarkable capabilities across various domains, particularly in natural language processing (NLP), where their ability to perform in-context learning (ICL) has significantly improved adaptation and generalization [2]. Unlike traditional deep learning models that require task-specific training, large Transformer models can infer from contextual information and generalize to new tasks without explicit parameter updates [3]. This powerful feature of ICL has sparked significant interest in applying these models beyond NLP, extending their potential to fields such as wireless communications [4], [5], which increase the adaptability to dynamic conditions.

MIMO equalization is a fundamental problem in wireless communication, aiming to mitigate channel distortion and interference to ensure reliable data transmission. Traditional equalization techniques, such as minimum mean square error (LMMSE) and maximum likelihood detection, require substantial prior knowledge of channel statistics [6]. Recent advances in deep learning have introduced data-driven equalization approaches, including supervised learning and meta-learning [7], [8], which enhance adaptability to dynamic channel conditions. However, these methods heavily rely on labeled training data and can struggle with generalization in highly heterogeneous and dynamic wireless environments, where the channel properties fluctuate rapidly. A promising new paradigm in next-generation wireless networks is the fully-decoupled RAN (FD-RAN) [9], [10], [11], [12]. Unlike conventional RAN architectures that tightly couple uplink and downlink operations, FD-RAN allows for their independent optimization, offering enhanced flexibility in resource allocation, interference management, and network scalability. One of the key advantages of FD-RAN is its ability to leverage spatial location information for network parameter adaptation. By incorporating geographic information (GI), FD-RAN can dynamically predict and optimize transmission power, beamforming strategies, improving network efficiency and performance under varying propagation conditions.

This paper investigates the application of ICL based on GI in multi-user MIMO (MU-MIMO) systems. A preliminary version of this work was presented in [13], where we focused on MU setup. Specifically, we consider the scenario where the uplink BS to central unit (CU) fronthaul capacity is limited. In this case, the analog signals received at the uplink BS are first quantized into digital signals with a fixed number of bits and then transmitted to the CU for MU equalization. Traditional quantization equalizers typically rely on the Bussgang decomposition-based LMMSE algorithm, where the received signals are passed through the Bussgang uniform quantizer [14], [15], [16]. In contrast, we propose a novel equalization framework that combines lightweight tiny-Transformer-based encoders deployed at uplink BSs with a centralized decoder-only Transformer decoder at the CU, incorporating GI as part of the input. Specifically, the tiny-Transformer is used for uplink data compression, thus reducing the fronthaul capacity requirements, while the decoder-only Transformer is for MU equalization. For a fair comparison, we still adopt the same Bussgang uniform quantizer. To investigate the model's generalization capability, we define *tasks* as specific user distributions, training the model on existing user drops and testing it on new user drops. Unlike traditional deep learning-based methods that rely on fine-tuning (e.g., meta-learning-based methods [17]), our approach utilizes prompt-based reasoning based on GI as well as pilot information for data symbol inference, thus eliminating the need for additional parameter updates when testing on new user drop data. To evaluate the effectiveness of this approach, we compare our ICL-based equalizer with meta-learning baselines based on model-agnostic meta-learning [18]. The main contributions of this paper are summarized as follows:

- We propose the concept of *tasks* within a representative equalization framework. Through comparative analysis of training and testing datasets from both ICL and meta-learning baselines, this methodology enables comprehensive evaluation of model performance across task variations and adaptability to novel data scenarios. Furthermore, we conduct a parametric study examining the impact of pre-training data volume on equalization effectiveness, revealing through numerical simulations that ICL demonstrates superior MSE performance over meta-learning approaches with increasing pre-training data scales. This empirical evidence substantiates the enhanced generalization potential inherent in large-scale model architectures.
- We propose a novel joint compression-equalization framework based on an encoder-decoder architecture, specifically designed for resource-constrained wireless communication systems. The framework employs tiny Transformers as encoders, strategically deployed at each uplink BS with limited computational resources, to efficiently compress received analog signals. The decoding process is implemented through a large decoder-only Transformer architecture, and we leverage the decoder-only GPT-2 model at the CU. This architecture processes quantized compressed signals from multiple uplink BSs through concatenation before equalization. Notably, our framework incorporates an innovative approach by integrating user geographic location information as auxiliary input during the equalization process.
- We performed a comprehensive investigation into the impact of encoder compression dimensionality on system performance, specifically comparing two configurations: E2_GI_ICL, where the encoder outputs a 2-dimensional representation, and E1_GI_ICL, where the encoder outputs a 1-dimensional representation. These proposed configurations were evaluated against multiple baseline systems: the conventional Bussgang uniform quantization-based LMMSE equalizer (BLMMSE), the standard LMMSE equalizer without quantization, and the GI_ICL equalizer without encoder implementation. Simulation results demonstrate that the proposed encoder-integrated ICL algorithm achieves superior MSE performance across various fronthaul capacity scenarios. Particularly noteworthy is the performance breakthrough of E1_GI_ICL, which

surpasses conventional LMMSE performance at 8-bit fronthaul capacity, marking a significant advancement for low-capacity fronthaul equalization systems.

II. RELATED WORKS

A. ADVANCES IN MIMO EQUALIZATION

MIMO systems are crucial for beyond-5G and 6G wireless networks, aiming to provide uniform service quality by utilizing distributed access points to serve users jointly. Peng et al. studied uplink resource allocation for CF massive MIMO-enabled URLLC in smart factories under finite blocklength and imperfect CSI constraints [19], and further investigated CF massive MIMO URLLC systems with pilot reuse, proposing joint pilot and power allocation to maximize user admission [20]. Björnson and Sanguinetti highlighted the significance of centralized MMSE processing in maximizing spectral efficiency and minimizing fronthaul signaling, thereby outperforming conventional cellular MIMO and small cell networks [21]. Kassam et al. proposed a centralized hybrid analog-digital equalizer for millimeter-wave cell-free MIMO, achieving near-digital performance with reduced complexity and power consumption [22]. Additionally, He et al. developed a distributed expectation propagation detector, which improves bit-error rates and system performance under imperfect channel state information conditions [23]. Kassam et al. [24] introduced a distributed hybrid analog/digital equalization scheme for cooperative cell-free networks. In the context of digital band processing-based massive MIMO systems, Zhao et al. [25] proposed decentralized LMMSE equalization techniques to mitigate the impact of colored noise in both star and daisy chain digital band processing architectures. Their work emphasizes reducing the communication overhead while maintaining high detection performance through dimensionality reduction and block coordinate descent optimization. To further address the trade-off between centralized and distributed processing, Hong et al. [26] proposed a group-joint MMSE complementary-based distributed uplink scheme. Their work bridges the gap between centralized and fully distributed MMSE equalization through a novel group-sliced approach, distributing computational and backhaul signaling overhead across the network. Low-resolution digital-to-analog converters (DACs) and quantized signal processing have gained significant attention in massive MU-MIMO systems to reduce power consumption and hardware complexity. Several studies have investigated the impact of low-resolution DACs and quantizers on the system performance. Jacobsson et al. [27] analyzed the impact of low-resolution DACs in a massive MU-MIMO downlink scenario, considering the practically relevant case of oversampling DACs. Using Bussgang's theorem, they derived analytical expressions for the uncoded bit error rate under linear precoding schemes, such as zero-forcing. Chen [28] addressed the equalization problem in massive MU-MIMO uplink systems under low-resolution quantization. The study proposed a finite-alphabet MMSE equalizer formulated as an optimization problem on a Riemannian manifold. The proposed Riemannian manifold optimization algorithm improved the performance of a 2-bit quantized equalizer while maintaining similar computational complexity as existing approaches. Liu et al. [29] introduced the second-order Hermite expansion model to analyze low-resolution quantized MIMO detection. Unlike existing linear approximation methods, their approach characterizes quantization distortion through the second-order Hermite kernel while preserving the first-order signal kernel.

B. DEEP LEARNING FOR MIMO EQUALIZATION

Deep learning has been widely applied to enhance the detection performance in massive MIMO systems. Several works have explored the application of deep neural networks (DNNs) in MIMO detection, focusing on performance improvement and complexity reduction. Albreem et al. [30] provided an extensive analysis of deep learning-based MIMO detection techniques, categorizing different architectures and discussing their trade-offs in terms of performance and computational complexity. Hu et al. [31] investigated the interpretability of deep MIMO detection methods. They compared data-driven deep detectors with model-driven approaches based on traditional detection algorithms. Raviv et al. proposed an online meta-learning framework for DNN-based receivers, demonstrating substantial performance gains in rapidly changing channels [32]. They further explored adaptive and flexible AI solutions for deep receivers, emphasizing efficient data acquisition and robust training algorithms [32]. Zhang et al. developed a meta-learning-based framework for MIMO detectors, showcasing enhanced convergence and robustness. These advancements underscore the transformative impact of meta-learning in optimizing receiver design for modern wireless communication systems [8]. Khobahi et al. [33] proposed LoRD-Net, a deep detection network designed to recover information symbols from one-bit quantized measurements. Their approach utilizes deep unfolding of first-order optimization iterations to develop a model-aware data-driven architecture. Notably, LoRD-Net does not require prior knowledge of the channel matrix, leveraging a two-stage training method to enhance performance. In the context of uplink massive MIMO systems, Lin et al. [34] developed a deep learning-based linear MMSE precoder that improves system performance under one-bit analog-to-digital converter (ADC) constraints. Their method unfolds the projected gradient descent algorithm into a deep neural network, treating step sizes as trainable parameters. Simulation results indicate that this approach outperforms conventional projected gradient descent algorithms while maintaining lower complexity. For channel estimation in millimeter-wave MIMO systems, Doshi and Andrews [35] formulated the problem as an inverse reconstruction task using deep generative networks. Their method optimizes the input vector of a pre-trained generative model to maximize a correlation-based loss function. The proposed deep generative prior significantly outperforms the

Bernoulli-Gaussian approximative message passing method in Line-of-Sight (LOS) scenarios and achieves comparable performance in non-LOS conditions.

C. AI-DRIVEN MIMO: GENERATIVE AI & TRANSFORMER

Generative AI plays a crucial role in wireless networks by optimizing resource allocation, enhancing channel estimation, improving MIMO equalization, and enabling intelligent cooperative computing [36], [37], [38]. The success of Transformers in various domains has inspired their adoption in wireless communication, particularly for enhancing signal processing, channel estimation, and interference management. Zayat et al. [39] introduced Transformer-masked autoencoders for efficient data compression and complexity reduction in advanced wireless networks. Wang et al. [40] developed the WIR-Transformer to recognize wireless interference with reduced computational overhead. Singh et al. [41] proposed a graph Transformer for channel estimation in intelligent reflection surface-aided systems, improving accuracy while lowering training costs. Liu et al. [42] demonstrated that a transformer-based denoising network could enhance angle-of-arrival estimation in non-LOS environments. Rajagopalan et al. [43] further explored in-context estimation with Transformers, achieving near-optimal performance with minimal data. These works collectively highlight the adaptability and efficiency of Transformers in wireless communication tasks. Beyond conventional Transformer applications, recent research has leveraged large language models (LLMs), which are built on Transformer architectures, to optimize MIMO systems. Sun et al. [44] introduced an AirFL-based LoRA framework for fine-tuning LLMs in 6G networks, utilizing over-the-air computation (AirComp) for efficient model aggregation in MIMO orthogonal frequency division multiplexing systems. Liu et al. [45] proposed LLM4CP, a pre-trained LLM-based approach for channel prediction in massive MIMO, demonstrating superior generalization capabilities and lower computational costs. Sheng et al. [46] developed UADFormer, a Transformer-based deep learning model for user activity detection in cell-free massive MIMO, integrating sparse attention and transfer learning for enhanced efficiency.

However, existing studies have not fully explored the ICL capabilities of transformers in wireless communication. In particular, the potential of decoder-only Transformers for MIMO equalization remains underexplored, and there is a lack of research on encoder-decoder-based joint training architectures for communication tasks. To bridge this gap, we investigate the role of decoder-only Transformer-based ICL in MIMO equalization and propose a novel encoder-decoder framework to enhance its effectiveness.

Notations: In this paper, $(\cdot)^{-1}$, $(\cdot)^{*}$, $(\cdot)^{H}$, and $(\cdot)^{T}$ indicate the inverse, conjugate, conjugate transpose, and regular transpose of a matrix, respectively. I_N refers to the identity matrix of size $N \times N$, while $\|\cdot\|$ stands for the Euclidean norm. The expression $x \sim \mathcal{CN}(0, \Sigma)$ describes a complex Gaussian random vector with zero mean and covariance Σ. $\mathbb{E}$ signifies the expected value, and $\mathrm{tr}\{\cdot\}$ stands for the trace of a matrix. The block diagonal matrix $\mathrm{diag}\{A_1, \ldots, A_L\}$ contains $A_1, \ldots, A_L$ along its diagonal.

III. SYSTEM MODEL

We consider a collaborative equalization scenario in FD-RAN, including L geographically distributed uplink BSs as well as a CU, where each uplink BS is equipped with N antennas. Each uplink BS connects to the CU via a limited capacity fronthaul connection. A block-fading environment with τ-length blocks is considered in this work, including $\tau_p < \tau$ channel uses as pilots, and the following $\tau_d = \tau - \tau_p$ channel uses for data transmission. There are K single-antenna users (UE) in the network and the channel between uplink BS l and UE k is denoted as $h_{kl} \in \mathbb{C}^N$. Taking into account a correlated Rayleigh fading channel, each channel instance can be generated by drawing from the correlated Rayleigh fading distribution

$$h_{kl} \sim \mathcal{CN}(0, R_{kl}) \in \mathbb{C}^N, \tag{1}$$

where $R_{kl} \in \mathbb{C}^{N\times N}$ denotes the spatial correlation matrix. $\boldsymbol{\beta}_k = [\beta_{k1}, \ldots, \beta_{kL}]$ denotes the collection of the large-scale fading between UE k and BS l, where $\beta_{kl} = \mathrm{tr}(R_{kl})/N$ includes pathloss and shadowing.

A. PILOT AND DATA TRANSMISSION WITHOUT QUANTIZATION

For the uplink pilot training, we assume the mutually orthogonal τ_p-length pilot sequence $\{\psi_1, \ldots, \psi_{\tau_p}\}$, where $\sqrt{\tau_p}\psi_{p_k} \in \mathbb{C}^{\tau_p\times 1}$ with the normalized energy $\|\psi_{p_k}\|^2 = 1$. Let $p_k \in \{1, \ldots, \tau_p\}$ denote the pilot assigned to user k, and $\mathcal{P}_k$ refers to the subset of users that use the same pilot as user k. The assigned transmitted pilot sequences of all UEs are $\boldsymbol{\Psi} = [\psi_{p_1}, \ldots, \psi_{p_K}] \in \mathbb{C}^{\tau_p\times K}$, the l-th BS receives

$$\mathbf{Y}_l^p = \sqrt{\rho_i\tau_p}\sum_{i=1}^{K} h_{il}\psi_{p_i}^{\mathrm{T}} + N_l^p, \tag{2}$$

where ρ_i denotes the transmit power of pilot ψ_{p_i}, $N_l^p \in \mathbb{C}^{N\times\tau_p}$ represents complex Gaussian noise at the receiver with independent and identically distributed (i.i.d) elements (i.e., $[N_l^p]_{i,j} \sim \mathcal{CN}(0, \sigma^2)$). As the number of users often exceeds the available number of pilot sequences (i.e., $K > \tau_p$), multiple users may use the same pilot sequence simultaneously, inevitably leading to pilot contamination.

The CU estimates the channels using MMSE based on received pilot signals and channel statistics from BSs. These estimates are independently computed for each BS without compromising optimality. Subsequently, for each user k, the CU aggregates these results to form a comprehensive channel estimate [21]

$$\hat{\mathbf{h}}_k = \begin{bmatrix} \hat{h}_{k1} \\ \vdots \\ \hat{h}_{kL} \end{bmatrix} \sim \mathcal{CN}\Big(0, \rho_k\mathbf{R}_k\Phi_k^{-1}\mathbf{R}_k\Big), \tag{3}$$

TABLE 1. The optimal step Size and distortion of a uniform quantizer with Bussgang decomposition [47].

α	Δ_{opt}	$\sigma^2_{\tilde{e}}$	$\tilde{a}$
1	1.596	0.2313	0.6366
2	0.9957	0.10472	0.88115
3	0.586	0.036037	0.96256
4	0.3352	0.011409	0.98845
5	0.1881	0.003482	0.996505
6	0.1041	0.001389	0.99896
7	0.0568	0.000342	0.99969
8	0.0307	0.0000876	0.999912

where $\mathbf{R}_k = \text{diag}(R_{k1}, \ldots, R_{kL}) \in \mathbb{C}^{LN\times LN}$, the inverse of the correlation matrix $\mathbf{\Phi}_k$ can be expressed as $\mathbf{\Phi}_k^{-1} = \text{diag}(\mathbf{\Phi}_{p_k1}^{-1}, \ldots, \mathbf{\Phi}_{p_kL}^{-1})$, and each $\mathbf{\Phi}_{p_k,l}$ this matrix in terms of the received signal $y^p_{p_k,l}$ can be denoted by

$$\mathbf{\Phi}_{p_k,l} = \mathbb{E}\left\{y^p_{p_k,l}\left(y^p_{p_k,l}\right)^{\text{H}}\right\} = \sum_{i\in\mathcal{P}_k} \tau_p\rho_i R_{il} + I_N, \tag{4}$$

where $y^p_{p_k,l} = \mathbf{Y}^p_l\psi^*_{p_k} \in \mathbb{C}^N$ denotes the correlates of the received signal $\mathbf{Y}^p_l$ with the associated pilot signal ψ_{p_k}.

The estimation error for the channel is given by $\mathbf{h}_k = \mathbf{h}_k - \hat{\mathbf{h}}_k \sim \mathcal{CN}(0, \mathbf{C}_k)$, where $\mathbf{C}_k = \text{diag}(\mathbf{C}_{k1}, \ldots, \mathbf{C}_{kL})$.

The estimate $\hat{h}_{kl}$ and estimation error $\check{h}_{kl} = h_{kl} - \hat{h}_{kl}$ are i.i.d $\check{h}_{kl} \sim \mathcal{CN}(0, C_{kl})$ with [21]

$$C_{kl} = \mathbb{E}\left\{\check{h}_{kl}\check{h}^{\text{H}}_{kl}\right\} = R_{kl} - \rho_k\tau_p R_{kl}\mathbf{\Phi}^{-1}_{p_k,l}R_{kl}. \tag{5}$$

During the uplink data transmission phase, all UEs transmit their signals to the BSs. The signal sent by the kth UE is represented as $x_k = \sqrt{\rho_k}s_k$, where $s_k \in \mathcal{S}_k$ is the transmitted symbol drawn from the constellation $\mathcal{S}_k$, with $\mathbb{E}\{|s_k|^2\} = 1$. Here, q_k denotes the transmit power. The received signal at the lth BS, denoted by the $N \times 1$ vector $\mathbf{Y}^d_l$, is given by

$$\mathbf{Y}^d_l = \sqrt{\rho_i}\sum_{k=1}^{K} h_{kl}s^T_k + N^d_l, \tag{6}$$

where $N^d_l \in \mathbb{C}^N$ represents the noise at the l-th BS, with $[N^d_l]_{i,j} \sim \mathcal{CN}(0, \sigma^2)$, following a circularly symmetric complex Gaussian distribution.

B. PILOT AND DATA TRANSMISSION WITH QUANTIZATION

Each receiver at one BS leverages two identical ADCs in each radio-frequency chain to quantify the in-phase and quadrature signals separately. Suppose that the ADCs at BS l use b_l bits for quantization. Thus, $\mathbf{Y}^p_l$ is quantized via

$$\tilde{\mathbf{Y}}^p_l = \mathcal{Q}_c\big(\mathbf{Y}^p_l\big) = \mathcal{Q}\big(\Re(\mathbf{Y}^p_l)\big) + j\mathcal{Q}\big(\Im(\mathbf{Y}^p_l)\big), \tag{7}$$

where, $\mathcal{Q}_c$ denotes the element-wise complex-valued quantizer including two real-valued quantizer $\mathcal{Q}$. To be specific, the midrise uniform quantizer is applied [47], and each element y of $\mathbf{Y}^p_l$ is quantized as follows

$$\mathcal{Q}(y) = \begin{cases} -\frac{V-1}{2}\Delta, & y \le -\left(\frac{V}{2}+1\right)\Delta, \\ \left(v+\frac{1}{2}\right)\Delta, & v\Delta \le y \le (v+1)\ \Delta, \\ & v = -\frac{V}{2}+1, \ldots, \frac{V}{2}-2, \\ \frac{V-1}{2}\Delta, & y \ge \left(\frac{V}{2}-1\right)\Delta, \end{cases} \tag{8}$$

with Δ denotes the step size of the quantizer with $V = 2^b$, where b denotes the quantization bits. The optimal step size can be solved by maximizing the signal-to-distortion noise ratio, which can be found in Table 1 [47].

To approximate the nonlinear effect of quantization, we refer the well-known Bussgang decomposition theory [48], the quantized version of the received pilot signal at the lth BS is given by

$$\begin{aligned}\tilde{\mathbf{Y}}^p_l &= \tilde{a}\mathbf{Y}^p_l + \mathbf{E}^p_l \\ &= \tilde{a}\sqrt{\tau_p\rho_i}\sum_{i=1}^{K} h_{il}\psi^{\text{T}}_{p_i} + \tilde{a}N^p_l + \mathbf{E}^p_l,\end{aligned} \tag{9}$$

where $\mathbf{E}^p_l$ denotes the quantization distortion caused by the ADCs at BS l during the stage of uplink pilot transmission.

Next, CU can get $\tilde{y}^p_{kl}$ of user k at BS l, by exploiting the pilot sequence $\psi^*_{p_k}$ to correlate the received quantized signal $\tilde{\mathbf{Y}}^p_l$ as

$$\begin{aligned}\tilde{y}^p_{kl} &= \tilde{\mathbf{Y}}^p_l\boldsymbol{\psi}^*_{p_k} \\ &= \tilde{a}\sqrt{\tau_p\rho_k}h_{kl} + \tilde{a}\sqrt{\tau_p\rho_i}\sum_{i=1,i\neq k}^{K} h_{il}\boldsymbol{\psi}^{\text{T}}_{p_i}\boldsymbol{\psi}^*_{p_k} \\ &\quad + \tilde{a}\dot{n}^p_{kl} + \dot{e}^p_{kl},\end{aligned} \tag{10}$$

where, $\dot{\mathbf{n}}^p_{kl} \triangleq \mathbf{N}^p_l\psi^*_{p_k}$, and $\dot{\mathbf{e}}^p_{lk} \triangleq \mathbf{E}^p_l\psi^*_{p_k}$. Therefore, the LMMSE estimate of $\mathbf{h}_{kl}$ given $\tilde{\mathbf{y}}^p_{kl}$ is

$$\tilde{\mathbf{h}}_{kl} = \mathbb{E}\{\mathbf{h}_{kl}{}^{\text{H}}\tilde{\mathbf{y}}^p_{kl}\}\left(\mathbb{E}\left\{\left(\tilde{\mathbf{y}}^p_{kl}\right)^{\text{H}}\tilde{\mathbf{y}}^p_{kl}\right\}\right)^{-1}\tilde{\mathbf{y}}^p_{kl} = c_{kl}\tilde{\mathbf{y}}^p_{kl}, \tag{11}$$

where

$$c_{kl} = \frac{\tilde{a}\sqrt{\tau_p\rho_k}\beta_{lk}}{\tilde{a}^2\left(\tau_p\rho_i\sum_{i=1,i\neq k}^{K}\beta_{il}\left|\psi^{\text{T}}_{p_k}\psi^*_{p_i}\right|^2+1\right)+\sigma^2_{\tilde{e}}\left(\rho_i\sum_{i=1}^{K}\beta_{il}+1\right)}. \tag{12}$$

The mean-square of the n-th element of $\tilde{\mathbf{h}}_{lk}$ denoted by γ_{kl}, and is given by

$$\begin{aligned}\gamma_{kl} &\triangleq \mathbb{E}\left\{\left|\left[\tilde{\mathbf{h}}_{kl}\right]_n\right|^2\right\} \\ &= \frac{\tilde{a}^2\tau_p\rho_k\beta^2_{kl}}{\tilde{a}^2\left(\tau_p\rho_i\sum_{i=1,i\neq k}^{K}\beta_{il}\left|\psi^{\text{T}}_{p_k}\psi^*_{p_i}\right|^2+1\right)+\sigma^2_{\tilde{e}}\left(\rho_i\sum_{i=1}^{K}\beta_{il}+1\right)}.\end{aligned} \tag{13}$$

The detailed proof can be seen in [49, Lemma 5].

The received signal at the l-th BS from all UEs, i.e., $\mathbf{Y}^d_l$ is given by Eq. (6). Exploiting the same quantizer, the quantized received data $\mathbf{Y}^d_l$ is given by

$$\tilde{\mathbf{Y}}^d_l = \mathcal{Q}_c\left(\mathbf{Y}^d_l\right) = \mathcal{Q}\left(\Re(\mathbf{Y}^d_l)\right) + j\mathcal{Q}(\Im\left(\mathbf{Y}^d_l\right). \tag{14}$$

Notably, the same quantizer used for pilot signal quantization is applied here for the data signal quantization, ensuring uniformity in quantization handling across transmission phases. This approach allows the system to maintain consistent quantization distortion characteristics in both pilot and data stages, which simplifies the design and analysis of quantization effects on channel estimation and subsequent data decoding.

C. CENTRALIZED EQUALIZATION

Let $\tilde{v}_k \in \mathbb{C}^{LN}$ be a detector matrix that depends on the estimated channel information at the receiver, i.e., $\tilde{h}_{kl}, \forall l, k$. We assume $\tilde{v}_k = [v_{k1}^{\mathrm{T}}, \ldots, v_{kL}^{\mathrm{T}}]^{\mathrm{T}}$ refers to the k-th column of the detector matrix $\tilde{V}$, with $\tilde{v}_{kl} \in \mathbb{C}^{N\times 1}$. The received signal for the k-th user after using the detector at the CU is given by

$$\tilde{s}_k = v_k^{\mathrm{H}}\left[\left(\tilde{\mathbf{Y}}_1^d\right)^{\mathrm{T}}, \ldots, \left(\tilde{\mathbf{Y}}_L^d\right)^{\mathrm{T}}\right]^{\mathrm{T}}, \tag{15}$$

where $\tilde{\mathbf{Y}}_l^d$ is defined in Eq. (14).

Equalization with Quantization: Quantization introduces additional uncertainty in the channel estimates, which can lead to performance degradation. We account for the effects of quantization on the channel estimates, a critical consideration in practical communication systems. The centralized MMSE equalizer is defined as

$$\tilde{v}_k = \left(\tilde{\alpha}^2 \rho_i \sum_{i=1}^{K} \tilde{h}_i \tilde{h}_i^{\mathrm{H}} + R^{\mathcal{Q}}\right)^{-1} \tilde{h}_k, \tag{16}$$

where $\tilde{h}_k = [\tilde{h}_{k1}^T, \ldots, \tilde{h}_{kL}^T]^T$, and the covariance matrix in terms of quantization $R^{\mathcal{Q}}$ is denoted as

$$\mathbf{R}^{\mathcal{Q}} = \rho_i \sum_{i=1}^{K} q_i \mathbf{W}_i^{\mathcal{Q}} + \mathbf{I}_{LN}^{\mathcal{Q}} + \mathbf{F}^{\mathcal{Q}}, \tag{17}$$

$$\mathbf{W}_i^{\mathcal{Q}} = \mathbf{S}_i^{\mathcal{Q}} - \mathbf{T}_i^{\mathcal{Q}}, \tag{18}$$

$$\mathbf{S}_i^{\mathcal{Q}} = \left(1 + \frac{\sigma_{\tilde{e},B}^2}{\tilde{\alpha}^2}\right) \mathrm{diag}(\mathrm{rep}(\beta_{i1}, N), \ldots, \mathrm{rep}(\beta_{iL}, N)), \tag{19}$$

$$\mathbf{T}_i^{\mathcal{Q}} = \mathrm{diag}(\mathrm{rep}(\gamma_{i1}, N), \ldots, \mathrm{rep}(\gamma_{iL}, N)), \tag{20}$$

$$\mathbf{F}^{\mathcal{Q}} = \frac{\sigma_{\tilde{e}}^2}{\tilde{\alpha}^2} \mathbf{I}_{LN}. \tag{21}$$

Equalization without Quantization: In this case, we consider the scenario where quantization effects are negligible, allowing for idealized channel estimation. The MMSE combining vector is given by

$$\mathbf{v}_k = \rho_k \left(\sum_{i=1}^{K} \rho_i \left(\hat{\mathbf{h}}_i \hat{\mathbf{h}}_i^{\mathrm{H}} + \mathbf{C}_i\right) + \sigma^2 \mathbf{I}_{LN}\right)^{-1} \hat{h}_k. \tag{22}$$

Eq. (22) emphasizes the use of precise channel state information, facilitating optimal signal combining. The MMSE approach minimizes the mean squared error between the estimated signal and the actual transmitted signal, effectively enhancing the quality of the received signal.

IV. TASK DEFINITION AND META-LEARNING BASELINE

This section defines the equalization task, outlines the data preparation process, introduces a meta-learning-based baseline using MAML for collaborative equalization, and compares it with ICL in terms of training and inference procedures.

A. TASK DEFINITION AND DATA PREPARATION

Unlike prior works [4], [50], [51], where each *task* corresponds to channel realizations sampled from a fixed channel covariance matrix, we propose a more dynamic definition. This definition accounts for variations in user distribution, spatial correlation, and small-scale fading effects. Specifically, we define *task* $\mathcal{T}_i$ as the coherence blocks generated from the i-th user drop, with corresponding channel realizations $\{h_1, \ldots, h_K\}$, where $h_k = [h_{k1}, \ldots, h_{kL}]$ is drawn from spatial correlation matrices $\{R_1, \ldots, R_K\}$, and $R_k = \mathrm{diag}(R_{k1}, \ldots, R_{kL})$.

By denoting $\mathcal{T}^{\mathrm{re}}$ as the number of channel realizations, the data in *task* $\mathcal{T}_i$ can be represented as

$$\mathcal{D}_{\mathcal{T}_i} = \sum_{j=1}^{\mathcal{T}^{\mathrm{re}}} \{\tilde{\mathbf{Y}}^p, \mathbf{\Psi}, \tilde{\mathbf{Y}}^d, s\}_j. \tag{23}$$

The data in each *task* $\mathcal{T}_i$ is represented as tuples $\{\tilde{\mathbf{Y}}^p, \mathbf{\Psi}, \tilde{\mathbf{Y}}^d, s\}$, sampled during i-th user drop. Besides, $\mathbf{\Psi}$ and $s = [s_1, \ldots, s_K]$ represent the transmit pilot and data, respectively, while $\tilde{\mathbf{Y}}^p$ and $\tilde{\mathbf{Y}}^d$ denote the received pilot and data that experience the same channel. Additionally, $\tilde{\mathbf{Y}}^p = [\tilde{\mathbf{Y}}_1^p, \ldots, \tilde{\mathbf{Y}}_L^p]$ and $\tilde{\mathbf{Y}}^d = [\tilde{\mathbf{Y}}_1^d, \ldots, \tilde{\mathbf{Y}}_L^d]$ represent the concatenation of quantized received pilot signals and data signals, respectively. In corresponding the number of *tasks* is denoted as $\mathcal{T}^{\mathrm{dr}}$ and the total number of data is represented as $\mathcal{D} = \sum_{i=1}^{\mathcal{T}^{\mathrm{dr}}} \mathcal{D}_{\mathcal{T}_i}$. For meta-training and testing, in *task* $\mathcal{T}_i$, we define $\mathcal{D}_{\mathcal{T}_i}^{\mathrm{train}} = \sum_{j=1}^{\mathcal{T}_i^{\mathrm{re}}} \{\tilde{\mathbf{Y}}^p, \mathbf{\Psi}\}_j$ for model training, and $\mathcal{D}_{\mathcal{T}_i}^{\mathrm{test}} = \sum_{j=1}^{\mathcal{T}_i^{\mathrm{re}}} \{\tilde{\mathbf{Y}}^d, s\}_j$ for calculating the model testing loss.

The detailed data preparation process for each *task* is delineated in Algorithm 1. For every *task* $\mathcal{T}_i$, the number of users, denoted as K, is randomly selected from a predefined set $\{1, \ldots, K\}$. Each user is assigned random locations within the network coverage area, simulating realistic deployment scenarios. During each channel realization step j, channel state information is generated, to calculate further the training data $\mathcal{D}_{\mathcal{T}_i}^{\mathrm{train}}$ and testing data $\mathcal{D}_{\mathcal{T}_i}^{\mathrm{test}}$. This methodology ensures that each *task* encapsulates the randomness inherent in user positioning and signal reception, yielding a robust and diverse dataset for both training the meta-learning and ICL algorithm.

B. META-LEARNING FOR COLLABORATIVE EQUALIZATION

We use the MAML framework [18] for fast adaptation to new equalization tasks. MAML optimizes a set of shared model parameters, θ, such that minimal updates are required to adapt the model to new, unseen tasks, leveraging two

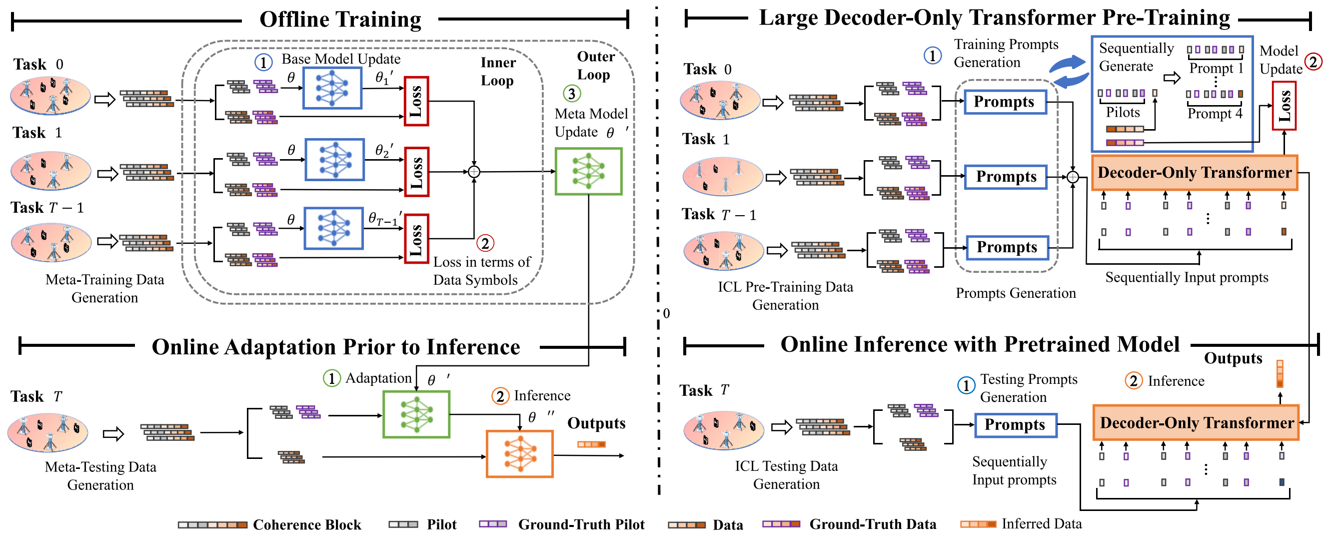

FIGURE 1. Meta-learning-based equalization is illustrated on the left, while in-context learning-based equalization is depicted on the right.

Algorithm 1 Data Preparation

Initialize dataset $\mathcal{D}$
for each task $\mathcal{T}_i$, where $i = 1, \ldots, \mathcal{T}^{\text{dr}}$ **do**
  Initialize task-specific dataset $\mathcal{D}_{\mathcal{T}_i}$
  Randomly select the number of users $k \in \{1, \ldots, K\}$
  // Generate random user drops
  for each channel realization step $j = 1, \ldots, \mathcal{T}_i^{\text{re}}$ **do**
    Each user randomly selects a pilot sequence $\{\boldsymbol{\Phi}\}_j$ and data symbols $\{s\}_j$
    Generate channel realizations $\{h_1, \ldots, h_K\}_j$
    Obtain quantized received pilot signals $\{\tilde{Y}^p\}_j$ and data symbols $\{\tilde{Y}^d\}_j$.
    Append to task data: $\mathcal{D}_{\mathcal{T}_i} := \mathcal{D}_{\mathcal{T}_i} \cup \{\tilde{Y}^p, \boldsymbol{\Psi}, \tilde{Y}^d, s\}_j$
  end for
  Add task data to the main dataset: $\mathcal{D} := \mathcal{D} \cup \mathcal{D}_{\mathcal{T}_i}$
end for

main processes: the inner-loop (task-specific adaptation) and outer-loop (meta-optimization).

Inner-Loop (Task-Specific Update): For each task $\mathcal{T}_i$, the model begins with shared parameters θ and performs gradient descent steps on the task-specific loss function $\mathcal{L}_{\mathcal{T}_i}^{\text{train}}(\theta) = \sum_{j=1}^{\mathcal{T}_i^{\text{re}}} |\{\tilde{\mathbf{Y}}_\theta^p\}_j - \{\boldsymbol{\Psi}\}_j|$; $\tilde{\mathbf{Y}}_\theta^p$ denotes the output of the network θ given $\tilde{\mathbf{Y}}^p$, computed over the dataset $\mathcal{D}_{\mathcal{T}_i}^{\text{train}}$. The task-specific parameters, θ_i', are updated as follows:

$$\theta_i' = \theta - \eta \nabla_\theta \mathcal{L}_{\mathcal{T}_i}^{\text{train}}(\theta) \tag{24}$$

where η is the inner-loop learning rate. The resulting task-adapted parameters, θ_i', are optimized to minimize the equalization error for the current task.

Outer-Loop (Meta-Optimization): After obtaining θ_i' for each task, we evaluate the adapted model on a new dataset (the query set) $\mathcal{D}_{\mathcal{T}_i}^{\text{test}}$. The meta-objective is to minimize the accumulated loss $\mathcal{L}_{\mathcal{T}_i}^{\text{test}}(\theta') = \sum_{j=1}^{\mathcal{T}_i^{\text{re}}} |\{\tilde{\mathbf{Y}}_{\theta'}^d\}_j - \{s\}_j|$, where $\tilde{\mathbf{Y}}_{\theta'}^d$ denotes the output of the network θ' given $\tilde{\mathbf{Y}}^d$ across all tasks, updating the shared parameters θ' to improve generalization across tasks. The outer-loop optimization is given by:

$$\theta'' \leftarrow \theta - \lambda \nabla_\theta \sum_{i=1}^{\mathcal{T}^{\text{dr}}} \mathcal{L}_{\mathcal{T}_i}^{\text{test}}(\theta_i') \tag{25}$$

where λ is the outer-loop learning rate. This step ensures that θ is optimized to provide a strong initialization, enabling rapid adaptation to new equalization tasks with minimal data.

Algorithm 2 Meta-Learning for Collaborative Equalization

// Meta Training stage:
Input: Data $\mathcal{D}$, inner-loop learning rate η, outer-loop learning rate λ
Initialize model parameters θ
for outer-loop update **do**
  for each task $\mathcal{T}_i$ **do**
    for inner-loop update **do**
      Update task-adapted parameters according to Eq. (24)
    end for
    Calculate the testing loss $\mathcal{L}_{\mathcal{T}_i}^{\text{test}}(\theta')$
  end for
  Update θ using meta-gradient according to Eq. (25)
end for
Store parameters θ''
// Inference stage:
for each unseen task **do**
  Generate data $\{\tilde{Y}_l^p, \boldsymbol{\Psi}, \tilde{Y}_l^d\}$
  Do adaptation on θ'' given $\{\tilde{Y}_l^p, \boldsymbol{\Psi}\}$, like Eq.(24) and inference $\tilde{Y}_l^d$ with updated parameters
end for

As shown in Algorithm 2, by learning a meta-model, the model can leverage pilot data $\mathcal{D}^{\text{train}}$ to perform rapid

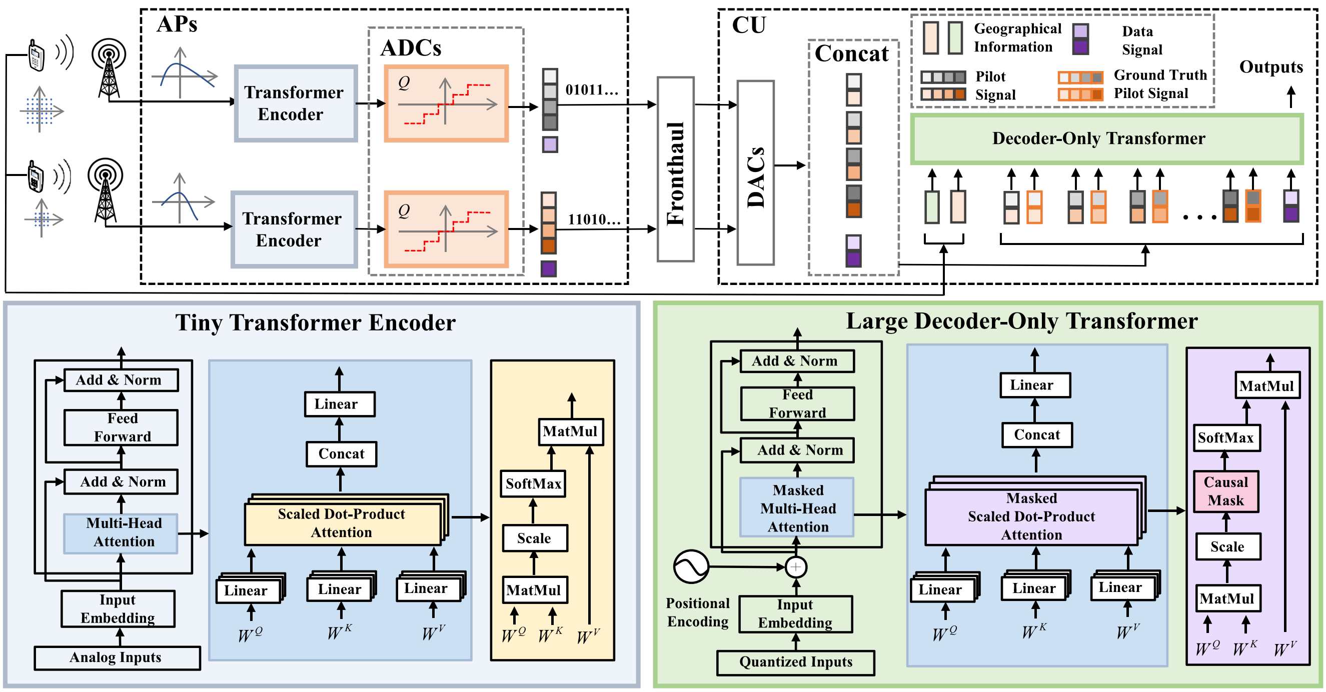

FIGURE 2. Joint Transformer-based signal processing: tiny encoder Transformers at BSs and a large decoder-only Transformer at the CU.

adaptation and inference in new, unseen environments (i.e., with new user distributions). This approach is more efficient than training from scratch, as it avoids the time-consuming process of starting from zero for each new environment.

C. ICL VERSUS META LEARNING ON EQUALIZATION

In Fig. 1, the left and right part shows the collaborative equalization pipelines based on meta learning and ICL, respectively. As offline supervised learning algorithms, both utilize the same dataset for model training and testing, but they differ in implementation details. The training and inference phases for each are analyzed as follows: *Training Phase*: As shown in the upper left of Fig. 1, for Meta Learning-based equalization, given identical data, meta learning (top left of Fig. 1) and ICL (top right of Fig. 1) respectively perform meta training and pretraining using the coherence block data generated. Meta larning involves three steps: first, for each task, the received pilot data and the ground truth of the pilot data, $\mathcal{D}_{\mathcal{T}_i}^{\text{train}}$, are used to update the initial model parameters θ, resulting in task-specific network parameters θ_i'; next, for each task, with the updated parameters θ_i', the received symbol data are input, and a loss is computed based on the model output and the received data, $\mathcal{D}_{\mathcal{T}_i}^{\text{test}}$; finally, the network is updated based on the aggregated loss to obtain the meta network θ''. For ICL pretraining, as illustrated in the top right of Fig. 1, the coherence blocks in the received data are processed such that one coherence block is sequentially split into multiple prompts, with each prompt's data equal to the number of data symbols in that coherence block.

Inference Phase: As depicted in the lower left of Fig. 1, meta learning-based equalization requires an adaptation of the meta network θ'' using pilot data and its ground truth, resulting in a new network parameters, and subsequently performing inference on the received symbol data using the updated parameters. The ICL inference process (bottom right of Fig. 1) involves generating prompts based on the received coherence blocks and directly inferring using the pretrained prompts and model parameters. The detailed ICL process is provided in the subsequent sections.

V. TRANSFORMER-ENCODER-BASED ICL FOR MIMO EQUALIZATION

As described in Fig. 2, in this work, we propose an advanced equalization framework leveraging a joint Transformer encoder and decoder-only Transformer decoder architecture for collaborative equalization. This framework begins with multiple BSs collecting analog signals, which are subsequently quantized through ADCs before being processed by a Transformer encoder. The encoded features are then transmitted via the fronthaul to the CU, where they are aggregated with additional contextual information, including pilot signals and geographical information. These combined inputs are then decoded by a decoder-only Transformer, which is designed to perform equalization through masked multi-head attention, ensuring causality by allowing each output to depend only on current and past inputs.

A. TRANSFORMER BASED ENCODER

As illustrated in the left side of Fig. 2, the Transformer-based encoder takes the sequence of received pilot and data signals as input, represented as

$$\left(\mathbf{Y}_l^p, \mathbf{Y}_l^d\right) = \left(\left[\mathbf{Y}_l^p\right]_1, \ldots, \left[\hat{\mathbf{Y}}_l^p\right]_i, \ldots, \left[\mathbf{Y}_l^p\right]_{\tau_p}, \mathbf{Y}_l^d\right) \in \mathbb{C}^{N\times(\tau_p+1)}, \quad (26)$$

which consists of the pilot $\mathbf{Y}_l^p$ and data $\mathbf{Y}_l^d$ received by BS l.

The first step is linear embedding operation $E(\cdot)$, which maps each received symbol to a vector of dimension D_e. Before that, the complex received symbols $[\mathbf{Y}_l^p]_i, \mathbf{Y}_l^d, \in \mathbb{C}^N$ are mapped into the corresponding real-valued vectors $[\hat{\mathbf{Y}}_l^p]_i, \in \mathbb{R}^{2N}$ by concatenating real and imaginary components as

$$\left[\hat{\mathbf{Y}}_l^p\right]_i = \left[\Re\left([\mathbf{Y}_l^p]_i\right), \Im\left([\mathbf{Y}_l^p]_i\right)\right], \quad (27)$$

and the same operation is applied to the received vectors $\mathbf{Y}_l^d$ to obtain the real-valued vectors $\hat{\mathbf{Y}}_l^d \in \mathbb{R}^{2N}$. In the following, the notations $[\hat{\mathbf{Y}}_l^p]_i$ and $\hat{\mathbf{Y}}_l^d$ refers to $2N \times 1$ real-valued vectors. The post-embedding sequence $\mathbf{E}_l = E(\hat{\mathbf{Y}}_l^p, \hat{\mathbf{Y}}_l^d) \in \mathbb{R}^{D_e\times(\tau_p+1)}$ is obtained by multiplying each vector, i.e., pilot $\hat{Y}_l^p$ or data $\hat{Y}_l^d$ with a trainable embedding matrix $M_e \in \mathbb{R}^{D_e\times 2N}$, as

$$\mathbf{E}_l = E\left(\mathbf{Y}_l^p, \mathbf{Y}_l^d\right) = E\left(\left[\mathbf{Y}_l^p\right]_1, \ldots, \left[\hat{\mathbf{Y}}_l^p\right]_i, \ldots, \left[\mathbf{Y}_l^p\right]_{\tau_p}, \mathbf{Y}_l^d\right), \quad (28)$$

where each column of matrix $\mathbf{E}_l$ corresponds to a *token*.

The embedded sequence $\mathbf{E}_l$ undergoes multiple iterations of a multi-head attention mechanism across C layers. Each layer c produces a sequence of τ_p+1 transformed tokens $\mathbf{E}_l^c$ for $c = 1, \ldots, C$. The c-th layer takes the sequence of tokens $\mathbf{E}_l^{c-1}$ from the preceding layer as its input, with $\mathbf{E}_l^0 = \mathbf{E}_l$. It applies H attention heads, aggregated to form the input $\mathbf{E}_l^c$ for the subsequent layer. In the final layer, the output $\mathbf{E}_l^C$ is quantized and transmitted to the CU via wired fronthaul.

In each layer c, the attention mechanism involves $3H$ trainable weight matrices: key matrices $\mathbf{W}_h^K \in \mathbb{R}^{D_\omega\times D_e}$, query matrices $\mathbf{W}_h^Q \in \mathbb{R}^{D_\omega\times D_e}$, and value matrices $\mathbf{W}_h^V \in \mathbb{R}^{D_v\times D_e}$ for all heads $h = 1, \ldots, H$. Here $D_e = D_v = D_e/H$. For each head h in layer c, the softmax self-attention generates $\tau_p + 1$ modified tokens as follows

$$B_h^l = \mathbf{W}_h^V \mathbf{E}_l^{c-1} \cdot \mathrm{softmax}\left(\frac{\left(\mathbf{W}_h^K \mathbf{E}_l^{c-1}\right)^T \left(\mathbf{W}_h^Q \mathbf{E}_l^{c-1}\right)}{\sqrt{D_k}}\right), \quad (29)$$

where the softmax function is applied column-wise. The outputs from all heads are concatenated token-wise and linearly projected into a sequence of $\tau_p + 1$ tokens, each with dimension D_e as

$$A^c = \left(\mathbf{W}^O\right)^T \left[B_1^c, \ldots, B_h^c, \ldots, B_H^c\right], \quad (30)$$

where $\mathbf{W}^O \in \mathbb{R}^{HD_v\times D_e}$ is another trainable weight matrix. Finally, each token is processed in parallel through a two-layer feed-forward neural network with residual connections to produce the output

$$\mathbf{E}_l^c = \mathbf{W}_1 \delta\left(\mathbf{W}_2 \beta(A^c + \mathbf{E}_l^{c-1})\right) + A^c + \mathbf{E}_l^{c-1}, \quad (31)$$

where $\mathbf{W}_1 \in \mathbb{R}^{D_e\times D_f}$ and $\mathbf{W}_2 \in \mathbb{R}^{D_f\times D_e}$ are trainable weight matrices, D_f denotes the number of hidden neurons, $\beta(\cdot)$ is the layer normalization operation, and $\delta(\cdot)$ is the Gaussian Error Linear Unit (GeLU) activation function.

The attention output of the final layer, denoted as E_l^C, is subjected to a trainable linear transformation. This linear layer is designed to have an output dimension D_o, which is less than $2N$ (i.e., $D_o < 2N$), such that $\mathbf{E}_l^C \in \mathbb{R}^{D_o\times(\tau_p+1)}$. This linear layer primarily serves for data compression, reducing the dimensionality from $2N$ to a more compact dimension D_o.

B. ICL EQUALIZATION WITH DECODER-ONLY TRANSFORMER

Let $\tilde{\mathbf{E}}_l^C$ represent the quantized output of the compressed signal $\mathbf{E}_l^C$ from BS l, using the same quantizer. In this way, we can denote the concatenation of the i-th symbol, where $i = 1, \ldots, \tau_p + 1$, from all BSs as

$$\left[\tilde{\mathbf{E}}^C\right]_i = \left[[\tilde{\mathbf{E}}_1^C]_i, \ldots, [\tilde{\mathbf{E}}_l^C]_i, \ldots, [\tilde{\mathbf{E}}_L^C]_i\right], \quad (32)$$

where $[\tilde{\mathbf{E}}^C]_i \in \mathbb{R}^{LD_o}$, denotes the i-th, $i < \tau_p+1$, compressed and quantized pilots from BS l. Furthermore, and the last symbol $[\tilde{\mathbf{E}}_l^C]_{\tau_p+1}$ is compressed and quantized data. Accordingly, the concatenation of the i-th ground truth pilot symbol of UE k is denoted as

$$\left[\hat{\mathbf{\Psi}}\right]_i = \left[[\hat{\psi}_{p_1}]_i, \ldots, [\hat{\psi}_{p_k}]_i, \ldots, [\hat{\psi}_{p_K}]_i\right], \quad (33)$$

where $[\hat{\mathbf{\Psi}}]_i \in \mathbb{R}^{2K}$. Here, $\hat{\psi}_{p_k} = [\Re(\psi_{p_k}), \Im(\psi_{p_k})] \in \mathbb{R}^{2\times\tau_p}$ represents the concatenation of real and imaginary components of ψ_{p_k}.

Geographical Information-Aware Equalization: Let (x_k, y_k, z_k) denote the location of the k-th UE and (x_l, y_l, z_l) denote the location of the l-th uplink BS. The AoA between UE k and uplink BS l is given by

$$\alpha_{k,l} = \tan^{-1}\left(\frac{y_k - y_l}{x_k - x_l}\right).$$

Thus, we define the AoA matrix as $\boldsymbol{\alpha} \in \mathbb{R}^{K\times L}$, where $[\boldsymbol{\alpha}]_{k,l} = \alpha_{k,l}$. To incorporate geographical information into

$$E(\mathbf{P}) = \left(M_e\boldsymbol{\alpha}_k, \ldots, M_e\boldsymbol{\alpha}_{j\in\mathcal{P}_k}, M_e\left[\tilde{\mathbf{E}}^C\right]_1, M_e\left[\hat{\mathbf{\Psi}}\right]_1, \ldots, M_e\left[\tilde{\mathbf{E}}^C\right]_i, M_e\left[\hat{\mathbf{\Psi}}\right]_i, \ldots, M_e\left[\tilde{\mathbf{E}}^C\right]_{\tau_p}, M_e\left[\hat{\mathbf{\Psi}}\right]_{\tau_p}, M_e\left[\tilde{\mathbf{E}}^C\right]_{\tau_p+1}\right), \quad (35)$$

the prompt, we utilize the AoA vector $\boldsymbol{\alpha}_k \in \mathbb{R}^L$. Specifically, we integrate both $\boldsymbol{\alpha}_k$ and $\boldsymbol{\alpha}_j, j \in \mathcal{P}_k$, into the prompt formulation as follows:

$$\mathbf{P} = \Big(\boldsymbol{\alpha}_k, \ldots, \boldsymbol{\alpha}_{j\in\mathcal{P}_k}, \Big[\tilde{\mathbf{E}}^C\Big]_1, \Big[\hat{\boldsymbol{\Psi}}\Big]_1, \ldots, \Big[\tilde{\mathbf{E}}^C\Big]_i, \Big[\hat{\boldsymbol{\Psi}}\Big]_i, \ldots, \Big[\tilde{\mathbf{E}}^C\Big]_{\tau_p}, \Big[\hat{\boldsymbol{\Psi}}\Big]_{\tau_p}, \Big[\tilde{\mathbf{E}}^C\Big]_{\tau_p+1}\Big). \tag{34}$$

Since the dimensions of large-scale information and the pilot information are different, zero-padding is used to keep a common dimension $D_s = \max\{LD_o, 2K, L\}$ before embedding.

As illustrated in Eq. (35), shown at the bottom of the previous page, the post-embedding sequence $E(\mathbf{P}) \in \mathbb{R}^{D_s\times(|\mathcal{P}_k|+2\tau_p+1)}$, where $|\mathcal{P}_k|$ denotes the cardinality of the set $\mathcal{P}_k$, is obtained by multiplying each vector with a trainable embedding matrix $M_e \in \mathbb{R}^{D_e\times D_s}$.

The positional encoding is added to the input embeddings to provide the inhrent sequential information. The positional encoding vector is generally added to each embedded token $E(\mathbf{P})$ as

$$\mathbf{E}^{\text{pos}} = E(\mathbf{P}) + \mathbf{P}_{\text{pos}}, \tag{36}$$

where $\mathbf{P}_{\text{pos}} \in \mathbb{R}^{D_s\times(|\mathcal{P}_k|+2\tau_p+1)}$ is the positional encoding matrix, which encodes the positions of each token in the sequence using a sine and cosine function, given by [3]

$$\begin{aligned}\mathbf{P}_{\text{pos}}(i, 2j) &= \sin\left(\frac{i}{10000^{2j/D_s}}\right),\\ \mathbf{P}_{\text{pos}}(i, 2j+1) &= \cos\left(\frac{i}{10000^{2j/D_s}}\right),\end{aligned} \tag{37}$$

where i denotes the position index of the token in the sequence, and j refers to the dimension index of the encoding.

Let M' be the causal mask matrix, which is a lower triangular matrix. For a sequence $\mathbf{E}^{\text{pos}}$ with a length of $|\mathcal{P}_k| + 2\tau_p + 1$, the mask matrix $M' \in \mathbb{R}^{(|\mathcal{P}_k|+2\tau_p+1)\ \times(|\mathcal{P}_k|+2\tau_p+1)}$ is defined as

$$M'_{ij} = \begin{cases} 0, & \text{if } i \geq j \\ -\infty, & \text{if } i < j \end{cases} \tag{38}$$

The input sequence to the decoder-only Transformer is $\mathbf{E}^{\text{pos}}$. In contrast to the compressed encoder, we denote H', C' as the heads and layers for the decoder-only transformer. This mask is added to the attention logits before applying the softmax function, ensuring that future tokens are not considered during attention. The masked attention formula for each head h' becomes

$$B_{h'} = \mathbf{W}_{h'}^V\tilde{\mathbf{E}}^{c'-1}\cdot\text{softmax}\left(\frac{\left(\mathbf{W}_{h'}^K\tilde{\mathbf{E}}^{c'-1}\right)^T\left(\mathbf{W}_{h'}^Q\tilde{\mathbf{E}}^{c'-1}\right)+M}{\sqrt{D_k}}\right). \tag{39}$$

By applying the same concatenation operation as in Eq. (30) with a trainable matrix $\mathbf{W}^O \in \mathbb{R}^{HD_v\times D_e}$, we obtain $A^{c'}$. Next, we use the residual connection as in Eq. (31). Finally, for the last layer, the output is given by $\tilde{\mathbf{E}}^{C'} \in \mathbb{R}^{D_o\times(|\mathcal{P}_k|+2\tau_p+1)}$.

Algorithm 3 Joint Encoder-Decoder Pre-Training for Collaborative Equalization

// Pre-Training Stage
Input: Training dataset $\mathcal{D}$
for $i = 1, 2, \ldots,$ *training steps* **do**
  for $b = 1, 2, \ldots,$ *batch size* **do**
    Sample a batch of training data from $\mathcal{D}$;
    Each uplink BS l receives the pre-compressed sequence E_l from Eq. (28);
    Each uplink BS l receives the post-compressed sequence E_l^c from Eq. (31);
    The CU quantizes the received input $\tilde{E}_l^C$ using the same quantizer in Eq. (8);
    Construct the prompt matrix P as described in Eq. (34), incorporating positional encoding from Eq. (36);
    Pass P through a central large-scale Transformer with a causal mask (Eq. (38)) to obtain the final inference output $E^{C'}$;
  end for
  Update network parameters $\boldsymbol{\varepsilon}$ according to Eq. (40);
end for
// Inference Stage
Given a new user drop, collect coherence blocks as examples along with the data symbols to be inferred;
Construct the prompt following the same procedure as in the training stage;
Obtain the predicted data symbols $E^{C'}$.

C. JOINT ENCODER-DECODER PRE-TRAINING

As illustrated in Algorithm (3), the proposed MIMO equalization framework leverages a joint encoder-decoder pre-training strategy to enhance generalization by capturing diverse contextual information in wireless environments. This pre-training phase is crucial as it enables the model to effectively integrate both the compact feature representations extracted by the Transformer encoder and the contextual prompts utilized by the decoder-only Transformer.

As illustrated in Fig. 2, by adopting this joint encoder-decoder approach, the model gains a comprehensive understanding of the communication environment. During pre-training, the joint architecture is optimized using a large corpus of training data, where pairs of transmitted and received signals serve as learning samples. The model is trained to minimize the reconstruction error, which quantifies the discrepancy between the predicted equalized signals and the actual transmitted signals. The training objective is formulated as

$$\mathcal{L}_{\mathcal{T}_i}^{\text{ICL}}(\boldsymbol{\varepsilon}) = \sum_{i=1}^{B}\sum_{j=1}^{\mathcal{T}_i^{\text{re}}}\left|\{\tilde{\mathbf{E}}_{\boldsymbol{\varepsilon}}^{C'}\}_j - \{s\}_j\right| \tag{40}$$

where $\mathcal{L}$ denotes the loss function, B represents the number of tasks in a batch, and $\mathcal{T}_j^{\text{re}}$ refers to the number of realizations per task. Here, $\tilde{\mathbf{E}}_{\boldsymbol{\varepsilon}}^{C'}$ denotes the equalized output,

TABLE 2. Comparative analysis of MAML and ICL performance with varying pre-trained data volumes.

Tasks: $\mathcal{T}^{\rm dr}$		2^2		2^4		2^6		2^8		2^{10}	
Methods		MAML	ICL	MAML	ICL	MAML	ICL	MAML	ICL	MAML	ICL
Num of Realizations: $\mathcal{T}^{\rm re}$	2^2	**0.84**	1.21	**0.85**	1.19	**0.77**	0.95	0.76	**0.70**	0.75	**0.52**
	2^6	**0.79**	1.11	0.72	**0.69**	0.71	**0.53**	0.68	**0.24**	0.68	**0.24**
	2^{10}	0.73	**0.61**	0.71	**0.34**	0.69	**0.26**	0.70	**0.17**	0.68	**0.15**

while s represents the corresponding ground truth transmitted signal.

The parameter $\boldsymbol{\varepsilon}$ encapsulates the entire network, including the decoder and all encoder components. The optimization process leverages the straight-through estimator to enable gradient propagation through the quantized representations, ensuring efficient backpropagation despite the discrete nature of certain operations within the encoder. This approach allows the model to effectively learn from compressed input representations while preserving the robustness of gradient-based optimization.

VI. SIMULATION RESULTS

We considered a 1-kilometer square simulation area, where four access points are symmetrically placed, each equipped with two receiving antennas. The receiver noise power is set to -50 dBm by default. In each drop, the number of users randomly varies from 1 to 4. We considered a coherent transmission mode, where the pilot sequence length is 8 and the inquiry data length is 1. In this coherent block, both pilot and data symbols experience the same channel. Additionally, we incorporated modulation diversity, where each data symbol is randomly modulated using one of the following schemes: {2PSK, 8PSK, 4QAM, 16QAM, 64QAM}.

The default decoder-only Transformer model follows a GPT-2 architecture with an embedding dimension of 256, representing the size of the input embeddings fed into the model. The model utilizes multi-head self-attention with 8 heads to capture diverse patterns and dependencies within the input sequence. It consists of 4 Transformer layers, each contributing to the depth and complexity of the representation learning. To prevent overfitting, a configurable dropout rate is used, with the default set to 0, meaning no dropout is applied during training. The model is optimized with a learning rate of 1×10^{-4}, ensuring gradual adjustments to model weights. Training is conducted in batches of size 1024, enabling efficient processing of large datasets and stable gradient updates throughout the training process.

By default, ICL training is performed on 2^{13} tasks with 2^{10} channel realizations, and the model is evaluated on 2^{10} unseen tasks. This ensures a rigorous assessment of generalization performance, demonstrating the model's ability to adapt to previously unseen task distributions.

A. PERFORMANCE COMPARISON BETWEEN MAML AND ICL

As illustrated in Table 2, we aim to compare the performance of MAML and ICL as the number of tasks and the number of channel realizations per task vary. For MAML, we use a two-layer Transformer with embedding dim is 256 with 4 heads. Data generated in the Table 2 in under 3bits quantization case, and the MAML method performs better on small-scale tasks (such as 2^2 and 2^4), with significantly lower MSE compared to ICL. This indicates that MAML is more effective at fitting data and reducing errors when the tasks are simpler. However, as the task scale increases, the performance of the ICL method gradually surpasses that of MAML, particularly for tasks of size 2^6, 2^8, and 2^{10}, where ICL's MSE is significantly lower than that of MAML. In particular, for the 2^{10} task, ICL shows superior performance across multiple realizations.Specifically, for the 2^{10} task, ICL achieves the lowest MSE, far outperforming MAML.

B. PERFORMANCE COMPARISON AMONG ICL, BLMMSE, AND LMMSE

In Fig. 4, we consider the case with pilot contamination, where the tested MSE values were evaluated over an signal-to-noise ratio (SNR) range from 5 dB to 35 dB. We compared four algorithms: LMMSE (LMMSE without quantization), BLMMSE (LMMSE with quantization), GI_ICL (ICL-based equalization with geographic information), E2_GI_ICL (encoder-based ICL with an encoder output dimension of 2), and E1_GI_ICL (encoder-based ICL with an encoder output dimension of 1). In this figure, we use fronthaul capacity as the independent variable and MSE as the performance metric. It can be observed that as the fronthaul capacity increases, GI_ICL, E2_GI_ICL, and E1_GI_ICL gradually improve in performance and eventually match or surpass LMMSE when the fronthaul capacity reaches 12 bits. Notably, E1_GI_ICL outperforms all other methods, exceeding the performance of LMMSE with just 8 bits of fronthaul capacity. In fact, E1_GI_ICL achieves over a 29% improvement in MSE compared to LMMSE at both 8-bit and 12-bit fronthaul capacity.

C. PERFORMANCE COMPARISON IN TERMS OF DIFFERENT FRONTHAUL CAPACITIES

In Fig. 3, we compare the MSE performance of the proposed algorithms under different fronthaul capacities, specifically 4, 8, and 12 bits, in the presence of pilot contamination. It can be observed that regardless of the fronthaul capacity, the proposed methods do not surpass LMMSE in the low-SNR region. However, in the high-SNR region, their performance consistently exceeds that of BLMMSE and, in certain cases, outperforms LMMSE. When the fronthaul capacity is 4 bits, the proposed algorithms achieve better

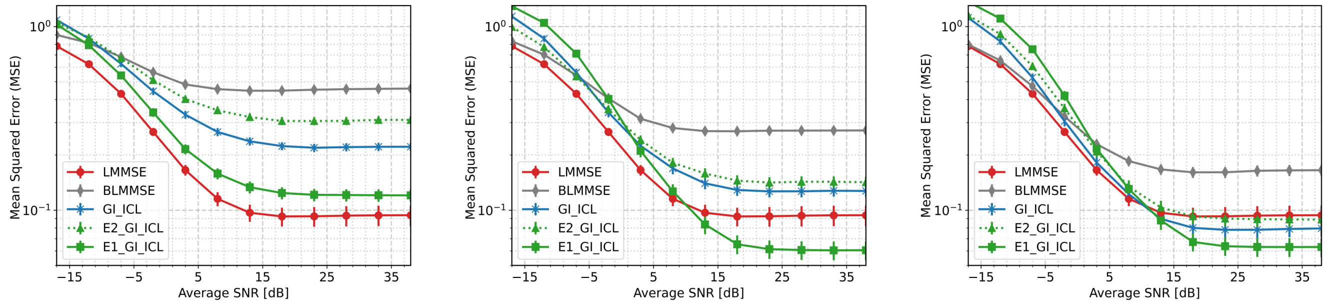

FIGURE 3. Comparison of fronthaul capacities at different bit rates: (a) 4bits, (b) 8bits, (c) 12bits.

performance than BLMMSE but do not surpass LMMSE across the entire tested SNR range. However, when the fronthaul capacity increases to 8 bits and 12 bits, E1_GI_ICL outperforms LMMSE in high-SNR scenarios. Notably, at a fronthaul capacity of 12 bits, all three proposed algorithms: E1_GI_ICL, E2_GI_ICL, and GI_ICL, consistently surpass LMMSE, demonstrating their effectiveness under different fronthaul capacity constraints. These results highlight the adaptability of the proposed methods, showcasing their ability to achieve competitive performance across varying fronthaul capacities, particularly in high-SNR conditions.

D. PERFORMANCE UNDER ORTHOGONAL PILOTS IN TERMS OF DIFFERENT SNR

In Fig. 5, we compare the MSE performance of various schemes under orthogonal pilot allocation, with a fronthaul capacity of 8 bits across different SNR ranges. From this figure, it is evident that at low SNR, the proposed algorithms initially perform worse than both LMMSE and BLMMSE. However, as the SNR increases, their performance gradually surpasses that of BLMMSE, demonstrating their effectiveness in higher-SNR scenarios. Notably, E1_GI_ICL exhibits outstanding performance at high SNR levels, significantly outperforming BLMMSE, which is particularly impressive. Furthermore, it can be observed that all the proposed algorithms outperform BLMMSE in the high-SNR region. This indicates that the proposed approaches remain highly effective even under orthogonal pilot allocation, reinforcing their robustness in handling diverse channel conditions.

E. PERFORMANCE UNDER NON-ORTHOGONAL PILOTS IN TERMS OF DIFFERENT SNR

Fig. 6 illustrates the variation of MSE as the pilot reuse factor increases under a fronthaul capacity of 8 bits. The pilot reuse factor indicates the number of users sharing the same pilot. From this figure, it can be observed that in both low and high pilot reuse scenarios, the proposed algorithms do not surpass LMMSE in the low-SNR region. However, as SNR increases, their performance improves significantly, eventually outperforming LMMSE in the high-SNR region. Focusing on the high-SNR regime, when the pilot reuse factor is 1, only E1_GI_ICL surpasses LMMSE. However, as the pilot reuse factor increases to 2 and 3, all three proposed algorithmsachieve performance that exceeds LMMSE.

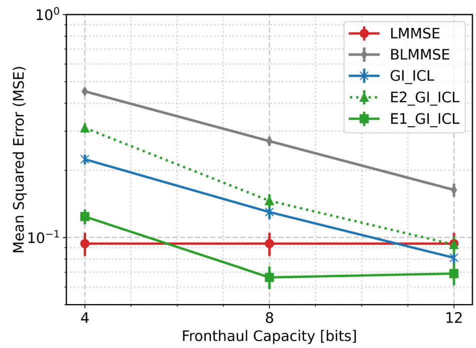

FIGURE 4. The MSE in terms of different fronthaul capacities.

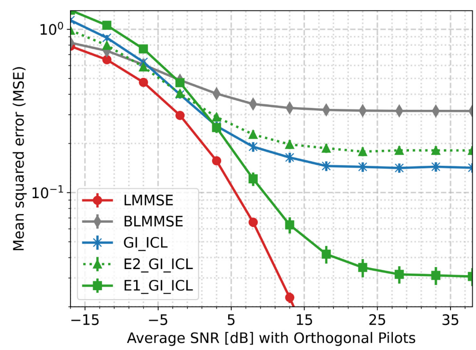

FIGURE 5. Test results of the network under orthogonal pilots allocation with a backhaul capacity of 8 bits.

VII. CONCLUSION

In this paper, we have investigated the integration of ICL into cooperative MIMO equalization in the FD-RAN architecture. A novel joint compression-equalization framework based on an encoder-decoder architecture has been proposed. Through extensive simulations, we have shown that our encoder integrated ICL framework can significantly outperform conventional LMMSE and BLMMSE equalizers under limited fronthaul capacity constraints. These results have demonstrated the effectiveness of ICL-based methods in capacity-limited communication scenarios. Furthermore, the integration of geographic location information has enhanced equalization accuracy, especially in multi-user scenarios. These findings underscore the transformative potential of ICL and largescale Transformer models in next-generation

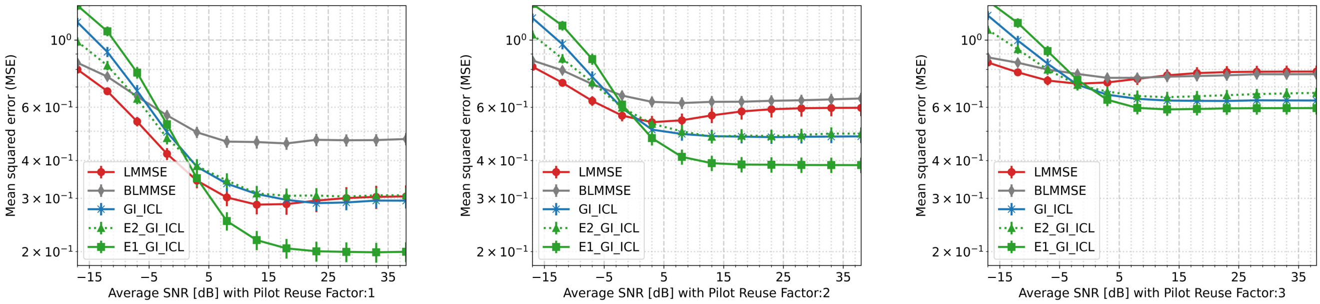

FIGURE 6. **Test results of the network under varying pilot contamination conditions with a backhaul capacity of 8 bits.**

wireless systems. In the future work, we will focus on optimizing computational efficiency and reducing inference latency to make the proposed framework more practical for real-time deployment. Additionally, we will plan to extend the framework's applicability to other critical wireless communication tasks, such as channel estimation, beamforming, and interference management, to further explore its potential in large-scale, data-driven wireless systems.

REFERENCES

[1] K. Alam et al., "A comprehensive tutorial and survey of O-RAN: Exploring slicing-aware architecture, deployment options, use cases, and challenges," 2024, *arXiv:2405.03555*.
[2] Q. Dong et al., "A survey on in-context learning," 2024, *arXiv:2301.00234*.
[3] A. Vaswani et al., "Attention is all you need," 2023, *arXiv:1706.03762*.
[4] M. Zecchin, K. Yu, and O. Simeone, "Cell-free multi-user MIMO Equalization via in-context learning," in *Proc. IEEE 25th Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, 2024, pp. 646–650.
[5] Z. Song, M. Zecchin, B. Rajendran, and O. Simeone, "In-context learned equalization in cell-free massive MIMO via state-space models," in *Proc. IEEE Int. Conf. Mach. Learn. Commun. Netw. (ICMLCN)*, 2025, pp. 1–6.
[6] H. Q. Ngo, L.-N. Tran, T. Q. Duong, M. Matthaiou, and E. G. Larsson, "On the total energy efficiency of cell-free massive MIMO," *IEEE Trans. Green Commun. Netw.*, vol. 2, no. 1, pp. 25–39, Mar. 2018.
[7] T.-V. Nguyen, S. Nassirpour, I. Atzeni, A. Tolli, A. L. Swindlehurst, and D. H. N. Nguyen, "MIMO detection with spatial sigma-delta ADCs: A variational Bayesian approach," 2024, *arXiv:2410.03891*.
[8] J. Zhang, Y. He, Y.-W. Li, C.-K. Wen, and S. Jin, "Meta learning-based MIMO detectors: Design, simulation, and experimental test," *IEEE Trans. Wireless Commun.*, vol. 20, no. 2, pp. 1122–1137, Feb. 2021.
[9] Q. Yu et al., "A fully-decoupled RAN architecture for 6G inspired by neurotransmission," *J. Commun. Inf. Netw.*, vol. 4, no. 4, pp. 15–23, Dec. 2019.
[10] K. Yu et al., "Fully-decoupled radio access networks: A flexible downlink multi-connectivity and dynamic resource cooperation framework," *IEEE Trans. Wireless Commun.*, vol. 22, no. 6, pp. 4202–4214, Jun. 2023.
[11] Y. Xu et al., "Fully-decoupled RAN for feedback-free multi-base station transmission in MIMO-OFDM system," *IEEE J. Sel. Areas Commun.*, vol. 43, no. 3, pp. 780–794, Mar. 2025.
[12] Z. Liu et al., "Leveraging self-supervised learning for MIMO-OFDM channel representation and generation," 2024, *arXiv:2407.07702*.
[13] M. Zecchin, K. Yu, and O. Simeone, "In-context learning for MIMO equalization using transformer-based sequence models," in *Proc. IEEE Int. Conf. Commun. Workshops (ICC Workshops)*, 2024, pp. 1573–1578.
[14] M. Bashar et al., "Exploiting deep learning in limited-fronthaul cell-free massive MIMO uplink," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 8, pp. 1678–1697, Aug. 2020.
[15] B. Fesl, N. Turan, B. Böck, and W. Utschick, "Channel estimation for quantized systems based on conditionally gaussian latent models," *IEEE Trans. Signal Process.*, vol. 72, pp. 1475–1490, Feb. 2024.
[16] M. Bashar, K. Cumanan, A. G. Burr, H. Q. Ngo, E. G. Larsson, and P. Xiao, "Energy efficiency of the cell-free massive MIMO uplink with optimal uniform quantization," *IEEE Trans. Green Commun. Netw.*, vol. 3, no. 4, pp. 971–987, Dec. 2019.
[17] T. Hospedales, A. Antoniou, P. Micaelli, and A. Storkey, "Meta-learning in neural networks: A survey," 2020, *arXiv:2004.05439*.
[18] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," 2017, *arXiv:1703.03400*.
[19] Q. Peng, H. Ren, C. Pan, N. Liu, and M. Elkashlan, "Resource allocation for uplink cell-free massive MIMO enabled URLLC in a smart factory," *IEEE Trans. Commun.*, vol. 71, no. 1, pp. 553–568, Jan. 2023.
[20] Q. Peng, H. Ren, M. Dong, M. Elkashlan, K.-K. Wong, and L. Hanzo, "Resource allocation for cell-free massive MIMO-aided URLLC systems relying on pilot sharing," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 7, pp. 2193–2207, Jul. 2023.
[21] E. Björnson and L. Sanguinetti, "Making cell-free massive MIMO competitive with MMSE processing and Centralized implementation," *IEEE Trans. Wireless Commun.*, vol. 19, no. 1, pp. 77–90, Jan. 2020.
[22] J. Kassam, D. Castanheira, A. Silva, R. Dinis, and A. Gameiro, "Centralized hybrid equalization for cell free mMIMO mmWave based systems," in *Proc. 13th Int. Symp. Commun. Syst., Netw. Digit. Signal Process. (CSNDSP)*, 2022, pp. 727–731.
[23] H. He, H. Wang, X. Yu, J. Zhang, S. H. Song, and K. B. Letaief, "Distributed expectation propagation detection for cell-free massive MIMO," in *Proc. IEEE Glob. Commun. Conf. (GLOBECOM)*, 2021, pp. 1–6.
[24] J. Kassam, D. Castanheira, A. Silva, R. Dinis, and A. Gameiro, "Distributed hybrid Equalization for cooperative millimeter-wave cell-free massive MIMO," *IEEE Trans. Commun.*, vol. 70, no. 8, pp. 5300–5316, Aug. 2022.
[25] X. Zhao, M. Li, B. Wang, E. Song, T.-H. Chang, and Q. Shi, "Efficient LMMSE Equalization for massive MIMO systems under Decentralized baseband processing architecture," *IEEE J. Sel. Areas Commun.*, vol. 43, no. 3, pp. 736–751, Mar. 2025.
[26] Z. Hong, T. Li, C. Li, D. Wang, and X. You, "Group-joint MMSE complementary-based distributed uplink for cell-free massive MIMO," *IEEE Trans. Wireless Commun.*, vol. 23, no. 10, pp. 13648–13663, Oct. 2024.
[27] S. Jacobsson, G. Durisi, M. Coldrey, and C. Studer, "Linear precoding with low-resolution DACs for massive MU-MIMO-OFDM downlink," *IEEE Trans. Wireless Commun.*, vol. 18, no. 3, pp. 1595–1609, Mar. 2019.
[28] J.-C. Chen, "Low-resolution MMSE Equalization for massive MU-MIMO systems," *IEEE Trans. Veh. Technol.*, vol. 72, no. 6, pp. 8164–8169, Jun. 2023.
[29] L. Liu, Y. Ma, and N. Yi, "Hermite expansion model and LMMSE analysis for low-resolution quantized MIMO detection," *IEEE Trans. Signal Process.*, vol. 69, pp. 5313–5328, Sep. 2021.
[30] M. A. Albreem, A. H. Alhabbash, S. Shahabuddin, and M. Juntti, "Deep learning for massive MIMO uplink detectors," *IEEE Commun. Surveys Tuts.*, vol. 24, no. 1, pp. 741–766, 1st Quart., 2022.

[31] Q. Hu, F. Gao, H. Zhang, G. Y. Li, and Z. Xu, “Understanding deep MIMO detection,” *IEEE Trans. Wireless Commun.*, vol. 22, no. 12, pp. 9626–9639, Dec. 2023.
[32] T. Raviv, S. Park, O. Simeone, Y. C. Eldar, and N. Shlezinger, “Online meta-learning for hybrid model-based deep receivers,” *IEEE Trans. Wireless Commun.*, vol. 22, no. 10, pp. 6415–6431, Oct. 2023.
[33] S. Khobahi, N. Shlezinger, M. Soltanalian, and Y. C. Eldar, “LoRD-Net: Unfolded deep detection network with low-resolution receivers,” *IEEE Trans. Signal Process.*, vol. 69, pp. 5651–5664, Oct. 2021.
[34] Q. Lin, H. Shen, and C. Zhao, “Learning linear MMSE precoder for uplink massive MIMO systems with one-bit ADCs,” *IEEE Wireless Commun. Lett.*, vol. 11, no. 10, pp. 2235–2239, Oct. 2022.
[35] A. Doshi and J. G. Andrews, “One-bit mmWave MIMO channel estimation using deep generative networks,” *IEEE Wireless Commun. Lett.*, vol. 12, no. 9, pp. 1593–1597, Sep. 2023.
[36] J. Wang et al., “Generative AI for integrated sensing and communication: Insights from the physical layer perspective,” *IEEE Wireless Commun.*, vol. 31, no. 5, pp. 246–255, Oct. 2024.
[37] J. Wang et al., “Generative AI based secure wireless sensing for ISAC networks,” 2024, *arXiv:2408.11398*.
[38] J. Wang et al., “Optimizing 6G integrated sensing and communications (ISAC) via expert networks,” 2024, *arXiv:2406.00408*.
[39] A. Zayat, M. A. Hasabelnaby, M. Obeed, and A. Chaaban, “Transformer masked autoencoders for next-generation wireless communications: Architecture and opportunities,” 2024, *arXiv:2401.06274*.
[40] P. Wang, Y. Cheng, B. Dong, R. Hu, and S. Li, “WIR-transformer: Using transformers for wireless interference recognition,” *IEEE Wireless Commun. Lett.*, vol. 11, no. 12, pp. 2472–2476, Dec. 2022.
[41] S. Singh, A. Trivedi, and D. Saxena, “Channel estimation for intelligent reflecting surface aided communication via graph transformer,” *IEEE Trans. Green Commun. Netw.*, vol. 8, no. 2, pp. 756–766, Jun. 2024.
[42] J. Liu, T. Wang, Y. Li, C. Li, Y. Wang, and Y. Shen, “A transformer-based signal denoising network for AoA estimation in NLoS environments,” *IEEE Commun. Lett.*, vol. 26, no. 10, pp. 2336–2339, Oct. 2022.
[43] V. Rajagopalan et al., “Transformers are efficient in-context estimators for wireless communication,” 2023, *arxiv:2311.00226*.
[44] H. Sun, H. Tian, W. Ni, J. Zheng, D. Niyato, and P. Zhang, “Federated low-rank adaptation for large models fine-tuning over wireless networks,” *IEEE Trans. Wireless Commun.*, vol. 24, no. 1, pp. 659–675, Jan. 2025.
[45] 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.
[46] Z. Sheng, P. Zhu, and X. You, “UADFormer: A transformer-based deep learning method for user activity detection in cell-free massive MIMO systems,” *IEEE Trans. Wireless Commun.*, vol. 23, no. 12, pp. 18516–18531, Dec. 2024.
[47] M. Bashar, K. Cumanan, A. G. Burr, H. Q. Ngo, M. Debbah, and P. Xiao, “Max–min rate of cell-free massive MIMO uplink with optimal uniform quantization,” *IEEE Trans. Commun.*, vol. 67, no. 10, pp. 6796–6815, Oct. 2019.
[48] P. Zillmann, “Relationship between two distortion measures for memoryless nonlinear systems,” *IEEE Signal Process. Lett.*, vol. 17, no. 11, pp. 917–920, Nov. 2010.
[49] M. Bashar et al., “Uplink spectral and energy efficiency of cell-free massive MIMO with optimal uniform quantization,” *IEEE Trans. Commun.*, vol. 69, no. 1, pp. 223–245, Jan. 2021.
[50] Z. Song, O. Simeone, and B. Rajendran, “Neuromorphic in-context learning for energy-efficient MIMO symbol detection,” in *Proc. IEEE 25th Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, 2024, pp. 1–5.
[51] Z. Song, Y. Ma, C. You, H. Yuan, J. Peng, and Y. Gao, “Transformer-based adaptive OFDM MIMO Equalization in intelligence-native RAN,” in *Proc. IEEE/CIC Int. Conf. Commun. China (ICCC Workshops)*, 2024, pp. 179–184.