Optimal Power Allocation and AI Receiver Design for Superimposed DMRS and Data Transmission

Sha Hu and Zhongwang Fu

Abstract

In this paper, we consider transmissions with superimposed (SI) demodulation-reference-symbol (DMRS) and data in orthogonal frequency-division multiplexing (OFDM) based multiple-input multiple-output (MIMO) systems. First, we derive an analytical framework to characterize the iterative behaviour between the mean-square errors (MSEs) of channel estimation (CE) and MIMO detection (MD) within an iterative CE and detection (ICED) process. This framework is subsequently utilized to optimize power allocation and pilot patterns between the DMRS and data symbols for SI-DMRS transmission. Second, we design an artificial intelligence (AI) based receiver built upon Transformer encoders for SI-DMRS transmissions, which incorporates an iterative CE and detection (ICED) structure. Simulation results demonstrate that the proposed AI-ICED receiver, combined with SI-DMRS, effectively increases spectral efficiency (SE) compared to conventional systems using non-overlapped DMRS and data symbols.

I. Introduction

Driven by evolutions towards the sixth-generation radio (6GR) networks, artificial intelligence (AI) and deep neural-network (NN) driven paradigm shifts have became a broad consensus [2]–[7]. Future wireless networks are steadily transitioning towards AI-native designs, including AI transceiver designs for multiple-input multiple-output (MIMO) and orthogonal frequency-division multiplexing (OFDM) systems. Unlike a conventional receiver design that relies on a cascade of independently optimized modules, AI-based receiver can unify multiple core functionalities such as channel estimation (CE) and MIMO detection (MD) [8]–[16] to harvest joint processing gains.

Although classical algorithms such as linear minimum mean square error (LMMSE) estimator and quasi-maximum likelihood detector (MLD) [17], [18] demonstrate good performance, exploring AI design is still of interest. Firstly, the inference process naturally aligns with contemporary parallel computing hardware, offering a path to reduce computational latency. Secondly, a unified NN architecture can mitigate

The authors work with Nyquist Research Center, Huawei Technologies Sweden AB, Sweden, and Zhongwang Fu was an intern during the time of this work [1]. Email: {hu.sha, zhongwang.fu}@huawei.com.

information barriers among different modules, achieving joint processing gains through latent feature exchanges, thereby demonstrating a potential to surpass conventional baselines in challenge scenarios.

A direction that AI could be beneficial in 6GR system is to optimize the overheads of pilots to increase spectral efficiency (SE). A classical dilemma is that to attain a good CE accuracy, more pilots should be transmitted but then the SE is decreased, and vice-visa, with fewer pilots the CE is not good enough and may degrade the data detection performance. In current fifth-generation new-radio (5G-NR) system, demodulation reference signal (DMRS) as pilots are transmitted on stand-alone resource elements (RE) to avoid interfering data transmissions. To increase SE, superimposed DMRS (SI-DMRS) has been proposed [19]–[21] which directly transmits DMRS overlapped with data symbols. This yields zero overheads from pilots, but the challenge is to attain an accurate CE from SI-DMRS and data[1]. While reducing pilot overheads or adopting superimposed DMRS (SI-DMRS) directly increases SE, it degrades channel state information (CSI) accuracy, which compromises symbol detection. Although attention-based NN architectures have been proposed to enhance CE under limited pilots [10], questions of how to optimize the balance between DMRS overheads and data detection, and design an AI receiver that is capable of reliable data recovery with a minimal pilot overheads, remain inadequately addressed.

Although conventional iterative CE and detection (ICED) [22], [23] can be applied, the two tasks of CE and MD on REs with SI-DMRS are entangled with each other, and it is challenging to achieve satisfying performances for both. However, this type of problem is perfectly for AI to attack, which learns from training dataset to approach the performance bound of joint CE and MD that is infeasible with conventional methods. On the other hand, with the rapid iteration of large foundation models centred on Transformers (e.g., Qwen, DeepSeek, Kimi [24]–[26]), in-context learning (ICL) has attracted widespread attention [27]–[29]. ICL essentially provides an unprecedented few-shot adaptation paradigm. Mapping this to communications systems, a few pilot signals and their RE-locations can serve as contextual information for interpreting the channel state information (CSI) and data-detecting of the received MIMO-OFDM grids.

Motivated by challenges with SI-DMRS and inspirations from ICL advances, we propose an AI-ICED receiver architecture for resolving CE and MD with SI-DMRS transmissions in MIMO-OFDM systems, which leverages advanced Transformer encoders to provide a unified design for diverse transmission schemes that achieves near-optimal performance. The architecture adopts an unfolded iterative cascade with a Transformer-encoder based backbone that incorporates Virtual Width Networks (VWN) [30], Tensor Product Attention (TPA) [31], and Mixture-of-Experts (MoE) [32] layers. Between iteration stages,

[1]Note that in our framework introduced later, the current DMRS patterns in 5G-NR can be seen as special cases of SI-DMRS by setting the power allocation of data symbols to zero on REs with SI-DMRS.

a differentiable physics-guided feature construction (PGFC) module converts intermediate predictions into explicit physical features and are forwarded to subsequent stages.

The main contributions of this paper are summarized as follows:

- **Theoretical analysis of power and DMRS allocations:** We have derived the MSEs for CE and MD within the ICED structure, proving the existence of an equilibrium state across iterations. Consequently, determining optimal power and DMRS resource allocations for general settings becomes straightforward. Furthermore, we incorporated mutual information effective signal-to-noise ratio mapping (MIESM) [42] to effectively combine the two RE subsets for data detections, with and without superimposed DMRS, to evaluate final decoding performance, thereby enhancing the practical applicability of the proposed optimizations.
- **Transformer unified AI-ICED receiver design:** We have designed an AI-ICED receiver using Transformer encoder blocks as its core backbone. The AI receiver processes the received signal, DMRS symbols, and a pilot mask to output symbol likelihoods at each stage. In the final stage, bit log-likelihood ratios (LLRs) are computed and passed to the decoder to evaluate block error rate (BLER) performance. The key features of the proposed AI-ICED receiver are twofold: joint CE and MD at each iteration stage, and a physics-guided feature construction (PGFC) mechanism between iterations that enables progressive and joint refinements.
- **Thorough simulation results and benchmarks:** We have evaluated the proposed AI-ICED receiver against optimal MLD and genie-CSI-aided benchmarks. The results quantify the effectiveness of AI receiver under superimposed DMRS (SI-DMRS) schemes, demonstrating near-optimal MD performance with genie-CSI. Furthermore, the proposed receiver converges rapidly within 2 to 3 iterations and achieves substantial SE gains by mitigating DMRS overhead compared to a standard 5G-NR baseline.

The remainder of this paper is organized as follows. Sec. II introduces the MIMO-OFDM system model with SI-DMRS, and a theoretical framework for analysing the MSEs of CE and MD in conventional ICED process is developed. Sec. III investigated the optimal power allocation of SI-DMRS based on the analytical framework established in Sec. II. Sec. IV presents the proposed AI-ICED receiver design in detail. Sec. V provides the simulation results, and Sec VI concludes the paper.

Notations:: The Hermitian of a matrix $\boldsymbol{A}$ is denoted as $\boldsymbol{A}^{\dagger}$, and the identity matrix is denoted as $\boldsymbol{I}$. The expectation operator is $\mathbb{E}\{\cdot\}$. The variables B, N_{r}, N_{t}, N_{sym}, and N_{sc} denote the batch size, number of receive antennas, number of transmit antennas, number of OFDM symbols within one subframe that carries data (excludes OFDM symbols carrying control channel), and number of subcarriers for the considered bandwidth, respectively. The variables N and M denotes the number of REs that only carries data symbols,

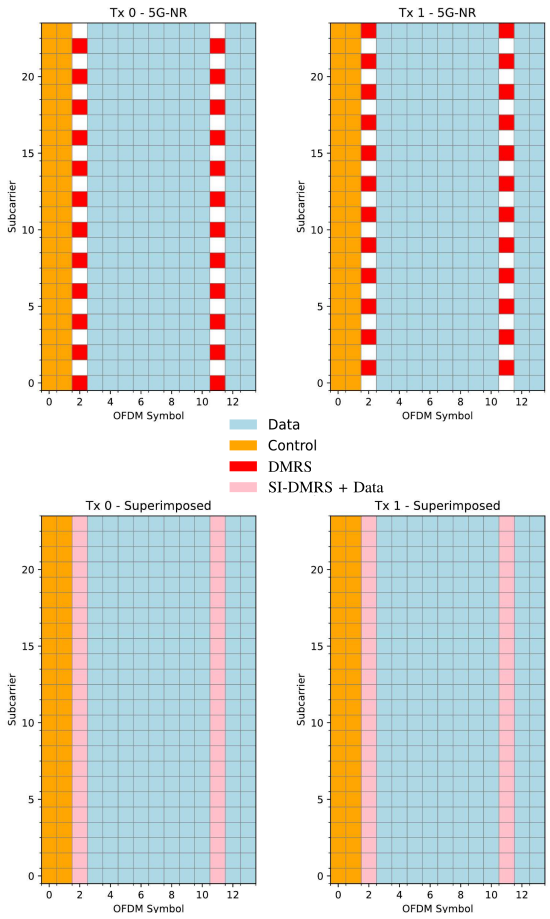

Fig. 1. Example patterns of DMRS and data transmission in standard 5G-NR (top) versus with SI-DMRS (bottom). The bandwidth is 2 physical resource block (2PRB) with $N_{\text{sc}} = 24$, and two OFDM symbols are used for control channel such that $N_{\text{sym}} = 12$ for the SI-DMRS case. In both cases, the transmit power per RE is normalized to unity. Under the SI-DMRS scheme, superimposing additional DMRS symbols onto data introduces no pilot overhead, however, because total transmit power is split between the DMRS and data components, both CE and MD become more challenging

and the number of REs carrying both SI-DMRS and data, respectively. Furthermore, $L = N_{\text{sym}} N_{\text{sc}} = N + M$ denote the input sequence length for AI receiver for a CE and MD occurrence. The power allcoation factor between SI-DMRS and data is denoted as b, and the power factor for data without SI-DMRS is set to a, both $0 < a, b \leq 1$. Furthermore, d_{model} is model-size in Transformer encoder blocks, and d_{ff} is (feed-forward network) FFN dimension.

II. System Model and ICED Process

A. Received Signal Model

We consider a MIMO-OFDM system consisting of N_{t} transmit-antennas (Tx) and N_{r} receiving-antennas (Rx). The transmitted signals are structured into a two-dimensional (2D) time-frequency resource grid, where transmission time interval (TTI) comprises N_{sym} OFDM symbols in the time domain and N_{sc} subcarriers in the frequency domain [33]. Let $\boldsymbol{x}_{n,k}$ denote the transmit symbol vector mapped to the

nth OFDM symbol and the kth subcarrier, with its entries drawn from an M-ary quadrature amplitude modulation (M-QAM) constellation.

The corresponding received signal vector $\boldsymbol{y}_{n,k} \in \mathbb{C}^{N_\mathrm{r} \times 1}$ without DMRS is

$$\boldsymbol{y}_{n,k} = \sqrt{\frac{1}{N_\mathrm{t}}} \boldsymbol{H}_{n,k} \boldsymbol{x}_{n,k} + \boldsymbol{w}_{n,k}, \tag{1}$$

where $\boldsymbol{H}_{n,k} \in \mathbb{C}^{N_\mathrm{r} \times N_\mathrm{t}}$ is the complex-valued frequency-domain MIMO channel matrix, and $\boldsymbol{w}_{n,k} \in \mathbb{C}^{N_\mathrm{r} \times 1}$ represents the additive white Gaussian noise (AWGN) at the receiver. The entries of $\boldsymbol{w}_{n,k}$ are independent and identically distributed (i.i.d.) complex Gaussian random variables distributed as $\boldsymbol{w}_{n,k} \sim \mathcal{CN}(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$, where σ^2 denotes noise denotes the noise variance per receive antenna. For simplicity, we assume no spatial correlation, meaning the entries of $\boldsymbol{H}_{n,k}$ are also i.i.d. complex Gaussian random variables satisfying $\mathbb{E}\{\boldsymbol{H}_{n,k} \boldsymbol{H}_{n,k}^\dagger\} = N_\mathrm{t} \boldsymbol{I}$. Furthermore, the data vectors are transmitted with unit-power such that $\mathbb{E}\{\boldsymbol{x}_{n,k} \boldsymbol{x}_{n,k}^\dagger\} = \boldsymbol{I}$, and the received signal-to-noise ratio (SNR) is defined as $\mathrm{SNR} = 1/\sigma^2$. By stacking all REs within one subframe, the received signal tensor $\boldsymbol{Y} \in \mathbb{C}^{N_\mathrm{r} \times N_\mathrm{sym} \times N_\mathrm{sc}}$ and the corresponding channel tensor $\boldsymbol{H} \in \mathbb{C}^{N_\mathrm{r} \times N_\mathrm{t} \times N_\mathrm{sym} \times N_\mathrm{sc}}$ constitute the complete observation space for the subsequent processing tasks.

To establish a framework for the iterative behaviours in ICED, we define some parameters as follows. Within one subframe, N REs are allocated for data transmission without DMRS, and M REs are allocated for SI-DMRS and data transmission. The total transmit power within one subframe is normalized to $N{+}M$. Further, denote $\mu = \frac{N}{M}$ and $k = 1 + \mu(1-a)$.

The power allocation is structured as follows:

- Data symbol vectors transmitted on REs without DMRS are with a power factor a.
- SI-DMRS symbols are transmitted with power bk.
- Data symbol vectors with SI-DMRS are transmitted with power $(1-b)k$.

Accordingly, the total power is fixed since $aN + kM = N + M$. The purpose of introduce the parameter a is to analyse the potential benefits of borrowing powers from data symbols to boost the SNR on REs with SI-DMRS, thereby the overall detection performance can be improved.

B. CE with SI-DMRS

On REs containing SI-DMRS, after removing the estimated data component, the received signal model for estimating the channel vector $\boldsymbol{h}$ for each Tx reads

$$\boldsymbol{y}_1 = \sqrt{bk} \boldsymbol{h} s_{\mathrm{DMRS}} + \sqrt{\frac{(1-b)k}{N_\mathrm{t}}} (\boldsymbol{H}\boldsymbol{x} - \tilde{\boldsymbol{H}}\tilde{\boldsymbol{x}}) + \boldsymbol{w}, \tag{2}$$

where $\tilde{\boldsymbol{H}}$ and $\tilde{\boldsymbol{x}}$ are the estimated MIMO channel and data symbols, respectively, and s is the transmitted DMRS symbol for a given Tx[2].

Denote the CE error matrix as $\boldsymbol{\Delta H} = \boldsymbol{H} - \tilde{\boldsymbol{H}}$, with $\boldsymbol{\Delta h}$ denoting the column corresponding to a single transmit antenna. Likewise, define the detection error vector as $\boldsymbol{\Delta x} = \boldsymbol{x} - \tilde{\boldsymbol{x}}$. Assuming the CE-MSE and MD-MSE from the previous iteration are respectively defined as

$$\mathbb{E}\{\boldsymbol{\Delta h}(\boldsymbol{\Delta h})^{\dagger}\} = p\boldsymbol{I}, \;\; 0 \geq p \leq 1, \tag{3}$$

$$\mathbb{E}\{\boldsymbol{\Delta x}(\boldsymbol{\Delta x})^{\dagger}\} = q\boldsymbol{I}, \;\; 0 \geq d <\leq 1. \tag{4}$$

In the next CE step in an ICED process, the update CE-MSE based on (2) can be computed as (see Appendix-A)

$$\mathbb{E}\{\boldsymbol{\Delta h}(\boldsymbol{\Delta h})^{\dagger}\} = \frac{k(1-b)(p+q-pq)+\sigma^2}{kb+k(1-b)(p+q-pq)+\sigma^2}\boldsymbol{I}. \tag{5}$$

C. CE De-noising Among Pilots

Note that the CE-MSE in (5) is evaluated for a single RE, and typically an LMMSE filter is subsequently applied to de-noise the to the CE over multiple REs across multiple REs carrying DMRS. In this case, the refined channel estimate is given by

$$\boldsymbol{h}_{\text{LMMSE}} = \boldsymbol{R}_{\text{h}}(\boldsymbol{R}_{\text{h}} + \tilde{\sigma}^2\boldsymbol{I})^{-1}\bar{\boldsymbol{h}}, \tag{6}$$

where $\boldsymbol{R}_{\text{h}}$is the time-frequency channel correlation matrix for each Tx–Rx pair, the vector $\bar{\boldsymbol{h}}$ comprises the CE on M/N_{t} REs for each link, and $\tilde{\sigma}^2$ is the effective noise power that incorporates both the noise and the residual data interference components. This formulation enables a maximum coherent processing gain of M/N_{t}.

Although the practical gain can be analysed depends on $\boldsymbol{R}_{\text{h}}$, for simplicity we assume that $\mathbb{E}\{\boldsymbol{\Delta h}(\boldsymbol{\Delta h})^{\dagger}\}$ is decreased by a factor $\frac{\gamma N_{\text{t}}}{M}$, where $\gamma \geq 1$ accounts for any loss relative to the maximum coherent gain. Consequently, after de-noising, the CE-MSE is modelled as $p\boldsymbol{I}$, where

$$p = \left(\frac{\gamma N_{\text{t}}}{M}\right)\frac{k(1-b)(p+q-pq)+\sigma^2}{kb+k(1-b)(p+q-pq)+\sigma^2}. \tag{7}$$

[2]We consider the case where DMRS for different Tx are non-overlapped and transmitted on separate REs. We also evaluated direct DMRS superimposition across Tx on the same RE (without orthogonal cover codes (OCC) as in 5G-NR), but found no noticeable performance gains, as it introduces spatial interference and reduces the effective DMRS power per Tx.

D. MD with SI-DMRS

After CE is obtained and de-noised, after removing the DMRS component, the received signal model for detecting $\boldsymbol{x}$ is

$$\begin{aligned}\boldsymbol{y}_2 &= \sqrt{bk}\boldsymbol{\Delta h} s_{\text{DMRS}} + \sqrt{\frac{(1-b)k}{N_{\text{t}}}}(\tilde{\boldsymbol{H}} + \boldsymbol{\Delta H})\boldsymbol{x} + \boldsymbol{w} \\ &= \sqrt{\frac{(1-b)k}{N_{\text{t}}}}\tilde{\boldsymbol{H}}\boldsymbol{x} + \sqrt{\frac{(1-b)k}{N_{\text{t}}}}\boldsymbol{\Delta H}\boldsymbol{x} + \sqrt{bk}\boldsymbol{\Delta h} s_{\text{DMRS}} + \boldsymbol{w}.\end{aligned} \tag{8}$$

For analytical tractability, the MD-MSE $\mathbb{E}\{\boldsymbol{\Delta x}(\boldsymbol{\Delta x})^{\dagger}\}$ based on (8) can be approximated as $q\boldsymbol{I}$ (see Appendix-B), where

$$q = \frac{kp + \sigma^2}{k(1-b+pb) + \sigma^2}. \tag{9}$$

As seen, equations (7) and (9) define the equilibrium state of CE-MSE and MD-MSE. This is formalized in Property 1, with its proof provided in Appendix C.

Property 1: There exists a stable equilibrium state for CE-MSE and MD-MSE within the ICED framework. Furthermore, the minimum CE-MSE solution p $(0 \leq p \leq 1)$ can be obtained as the root of a third-order polynomial equation:

$$Ap^3 + Bp^2 + Cp + D = 0, \tag{10}$$

where $u = 1-b$, $v = \frac{\sigma^2}{k}$, and $\rho = \frac{\gamma N_{\text{t}}}{M}$, and

$$A = u^2, \tag{11}$$

$$B = -((\rho+2)u^2 + (1-u)(1+v)), \tag{12}$$

$$C = \rho(u(1+u) + (1-u)v) - (u(1-u) + (1+u)v + v^2), \tag{13}$$

$$D = \rho v(2u+v). \tag{14}$$

Property 1 provides a theoretical framework to analyse the iterative behaviour of CE and MD on REs with SI-DMRS for general configurations of $(N, M, a, b, N_{\text{t}}, \gamma, \sigma^2)$, enabling the converged CE-MSE and MD-MSE to be determined directly by solving the equilibrium equations. However, to analyse the overall decoding performance across the entire subframe, we must also account for the MD-MSE on REs without DMRS.

E. MD on REs without SI-DMRS

Once the iterative steps between CE and MD on REs with SI-DMRS have converged, the CE $\tilde{\boldsymbol{H}}$ achieves high accuracy and can be utilized for data-detection on the remaining REs[3]. On REs without SI-DMRS, the received signal model for detecting $\boldsymbol{x}$ reads

$$\begin{aligned}\boldsymbol{y}_3&=\sqrt{\frac{a}{N_{\mathrm{t}}}}(\tilde{\boldsymbol{H}}+\boldsymbol{\Delta H})\boldsymbol{x}+\boldsymbol{w}\\&=\sqrt{\frac{a}{N_{\mathrm{t}}}}\tilde{\boldsymbol{H}}\boldsymbol{x}+\sqrt{\frac{a}{N_{\mathrm{t}}}}\boldsymbol{\Delta H}\boldsymbol{x}+\boldsymbol{w}.\end{aligned}\tag{15}$$

In this case, the MD-MSE as approximated as $t\boldsymbol{I}$ (see Appendix-D), where

$$t=\frac{ap+\sigma^2}{a(1-b)+\sigma^2}.\tag{16}$$

III. Optimal Power Allocations for SI-DMRS Transmissions

With the CE-MSE and MD-MSE for the ICED process derived above, we now seek to address the following fundamental question: ***How can we optimize the power allocation factors*** (a,b) ***to maximize the final performance?***

A. MIESM

Before addressing this question, we first introduce the mutual information effective SNR mapping (MIESM) to quantify the combined MD-MSE. The basic principles are as follows:

- Convert the MD-MSE of the two subsets into effective SNR [**?**] values:

$$\theta_1=\frac{1}{q}-1,\quad \theta_2=\frac{1}{t}-1.\tag{17}$$

- Map SNR to MI using a constellation-constrained MI function:

$$I_i=f_{\mathrm{MI}}\left(\frac{\theta_i}{\beta}\right),\tag{18}$$

  where $f_{\mathrm{MI}}(\cdot)$ is the MI mapping for the chosen modulation and the parameter β is a calibration factor. Typical functions of $f_{\mathrm{MI}}(\cdot)$ are shown in Appendix-E.

- Compute the weighted average MI:

$$I_{\mathrm{avg}}=\frac{MI_1+NI_2}{M+N}.\tag{19}$$

- Map the average MI back to an equivalent SNR:

$$\theta_{\mathrm{eff}}=\beta\, f_{\mathrm{MI}}^{-1}(I_{\mathrm{avg}}),\tag{20}$$

At last, the combined MD-MSE is calculated as $1/(1+\theta_{\mathrm{eff}})$.

[3]In AI-based receivers, it is feasible and beneficial to detect data symbols across all REs and leverage all data estimates to further enhance CE by fully exploiting the correlations between the data and channel. This methodology is adopted in our AI receiver design, which is elaborated later.

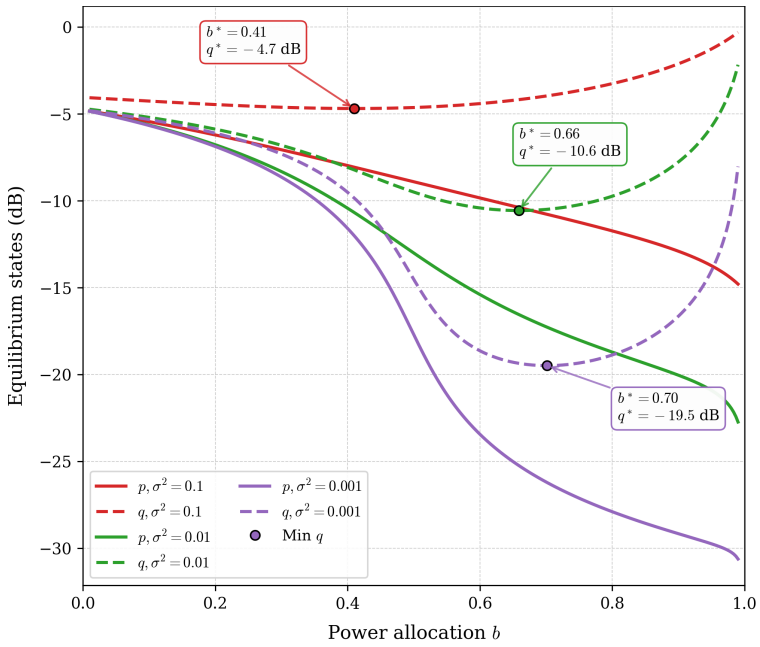

Fig. 2. The equilibrium state of CE-MSE and MD-MSE based on Property 1 with ($a=1$, $\gamma=2$, $N_t=4$, $N=264$, $M=24$).

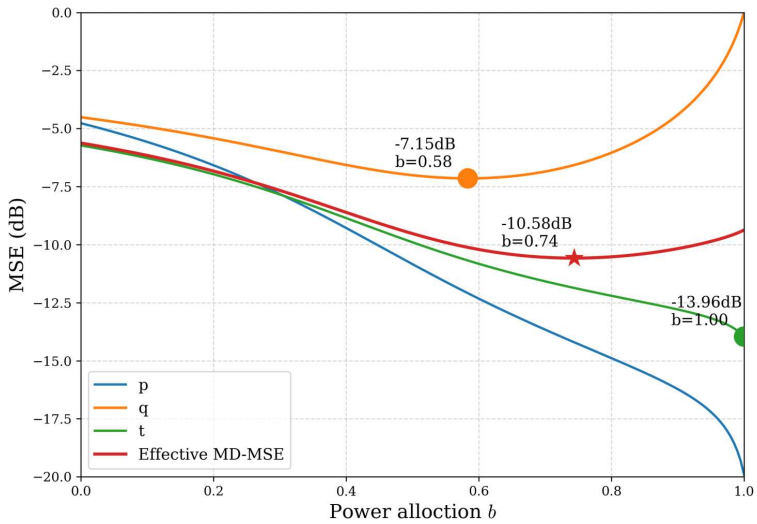

Fig. 3. The effective MD-MSE based on MIESM mapping of p and t with ($a=1$, $\gamma=2$, $N_t=4$, $N=264$, $M=24$) and $\sigma^2=-15$dB.

B. Optimal Power Allocations

From Property 1, the equilibrium state of (p, q) and t can be directly computed for different power allocation strategies. Subsequently, the effective SNR θ_{eff} is obtained from (q, t). Thus, the goal of optimal power allocation is to maximize θ_{eff}. This provides a clean analytical approach to assess the performance of SI-DMRS, with several illustrative examples shown below.

In Fig. 2, the values of p and q are illustrated for different noise powers ($\sigma^2 = -10$dB, -20dB, and -30dB) with $a=1$ (i.e., without power-boosting from REs carrying exclusively data symbols). As observed, there exists an optimal value of b-the power allocation factor for DMRS-that minimizes q (the

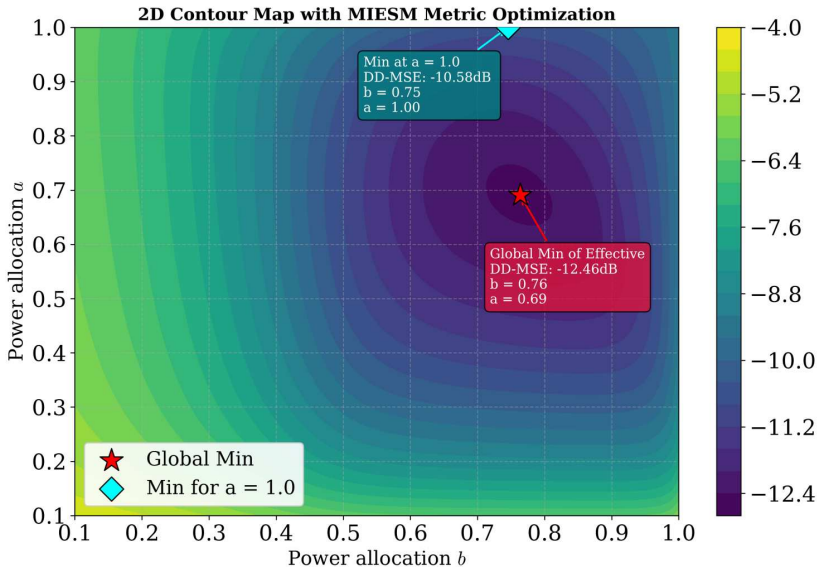

Fig. 4. 2D contour of effective MD-MSE for joint power allocations over (a, b) with ($\gamma = 2$, $N_{\mathrm{t}} = 4$, $N = 264$, $M = 24$) and $\sigma^2 = -15$dB.

MD-MSE), whereas the CE-MSE continues to decrease as b increases. Furthermore, as the SNR increases, the optimal b also increases. This behaviour occurs because, at higher SNRs, the CE-MSE becomes the bottleneck affecting the MD-MSE.

In Fig. 3, the effective MD-MSE is presented using $\beta = 1$ in (18) and $\sigma^2 = -15$dB, which aligns more closely with overall decoding performance. As seen, the effective MD-MSE-derived via MIESM from (q, t)-is minimized at a higher value of b compared to the case that considers q in isolation. Furthermore, examining the curve for t indicates that the MD-MSE of data symbols on REs without SI-DMRS requires higher CE accuracy to achieve improved error performance. Note that in both Fig. 2 and Fig. 3, the data power allocation factor is fixed to $a = 1$.

Fig. 4 illustrates the joint optimization over both a and b. As demonstrated, the effective MD-MSE is reduced from -10.42dB to -12.2dB when a is decreased from 1 to 0.67. This highlights the significant performance benefits of jointly optimizing power allocation across REs[4].

C. Is Superimposing More DMRS with Data Beneficial?

Another interesting question is ***whether we should superimpose more DMRS with data***. In the preceding figures, the number of DMRS is set to $M = 24$, with each Tx occupying 6 REs out of a total grid of 240 REs.In Fig. 5, we evaluated different configurations of M (24, 48, and 96) and plotted the minimum effective MD-MSE. For each configuration, we also tested different factors ($\gamma = 2$, 4, and 6) representing varying de-noising gains. As shown, for a constant $\gamma = 2$, increasing M from 24 to 96

[4]A potential drawback of this approach, however, is an increase in the peak-to-average-power ratio (PAPR) for the MIMO-OFDM system.

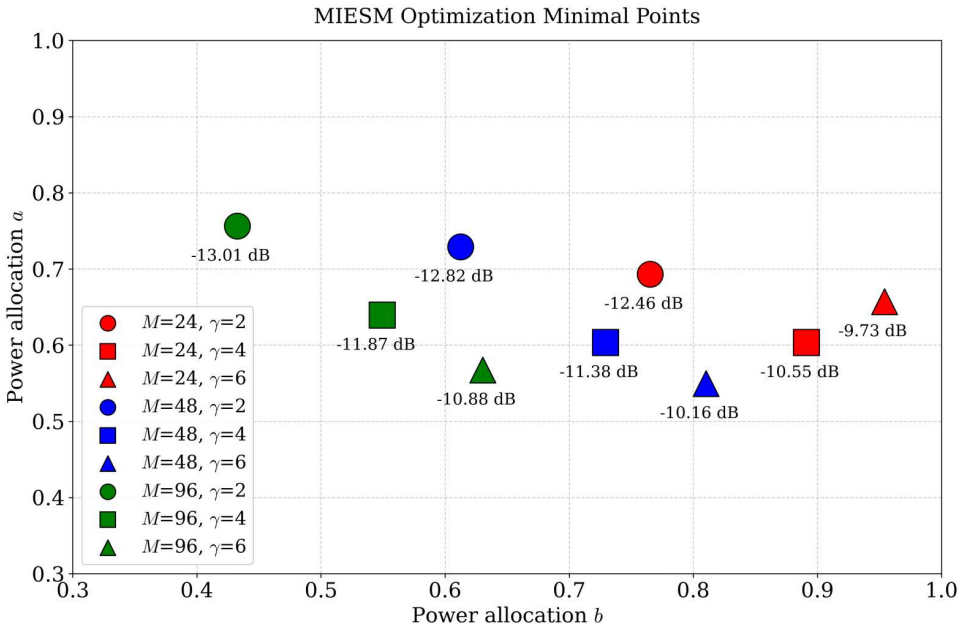

Fig. 5. The minimal effective MIESM with different configurations of M, i.e., for superimposing more DMRS with data. The gains are limited even under the idea assumption that the de-nosing gain of CE linearly increases in the number of DMRS.

yields gains. However, keeping γ constant implies that the de-noising gain increases linearly with M. When compared to the extreme case where the de-noising gain remains unchanged (i.e., when γ/M is constant), the configuration with $M=24$ and $\gamma=2$ outperforms the $M=48$ and $\gamma=4$ configuration by 1.55dB. Therefore, it is not always beneficial to superimpose more DMRS with data, and the optimal choice depends heavily on the de-noising gains and operational system parameters. Nevertheless, this analytical framework provides a valuable perspective for analysing general cases to optimize SI-DMRS.

IV. The Proposed Transformer based AI-ICED Receiver Design

With the theoretical framework established, we next present the AI-ICED receiver design, which harnesses the spectral efficiency (SE) gains enabled by SI-DMRS transmission. The overall architecture is illustrated in Fig. 6. The AI receiver processes three tensor inputs: the received signal tensor $\mathbf{Y}$, the DMRS symbols $\mathrm{s_p}$, and the DMRS coordinate mask Ω_p. By treating the DMRS information as contextual prompts, the AI architecture establishes a direct mapping from the observation space to the symbol space. Meanwhile, the AI receiver applies an ICED design, and at iteration stage i, it updates the CE and the a posteriori probabilities of symbol vectors, $p^i(\boldsymbol{x} \mid \boldsymbol{Y}, \tilde{\boldsymbol{H}}^{(i-1)}, \tilde{\boldsymbol{x}}^{(i-1)})$, by capturing the dependencies between the intermediate CE and soft symbol estimates from the previous iteration. Ultimately, the logits generated by the AI-ICED receiver are used to calculate bit log-likelihood ratios (LLRs), which are then fed into the channel decoder. As depicted in Fig. 6, the overall design is structured as a cascaded loop where the inputs at each stage are processed through three primary modules: the tokenizer, the ICED module, and the physics-guided feature construction (PGFC) module between iteration stages.

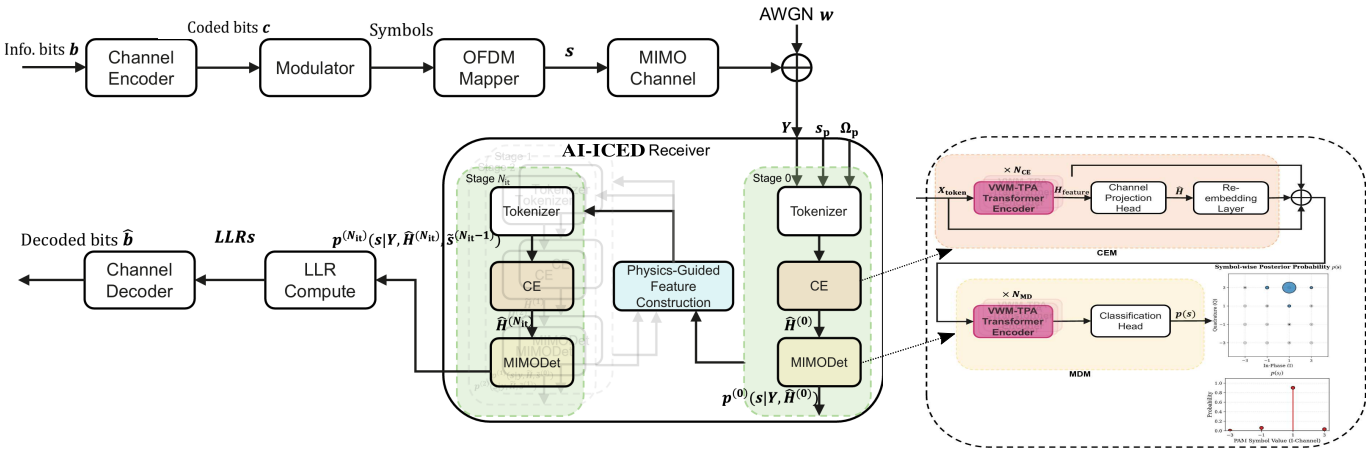

Fig. 6. The structure of the MIMO-OFDM system and the proposed AI-ICED receiver, along with details regarding the NN construction and configurations provided in Appendix F through I and Table III.

A. *Tokenization*

This module serves as a tensor pre-processor that performs a spatial flattening operation, concatenates the tensors along their feature axes, and applies a linear projection. This step maps the physical tensors into a unified d_{virtual}-dimensional token sequence. Note that all defined complex-valued tensors include a batch dimension B, with the real and imaginary components decoupled and stacked along a dedicated feature axis of size two. The module receives the real-valued observation tensor $\boldsymbol{Y}_{\mathbb{R}} \in \mathbb{R}^{B\times N_{\text{r}}\times N_{\text{sym}}\times N_{\text{sc}}\times 2}$, the symbol tensor $\boldsymbol{s}_{\mathbb{R}} \in \mathbb{R}^{B\times N_{\text{t}}\times N_{\text{sym}}\times N_{\text{sc}}\times 2}$, and the binary mask $\boldsymbol{\Omega}_{\text{p}} \in \{0,1\}^{B\times N_{\text{t}}\times N_{\text{sym}}\times N_{\text{sc}}}$ indicating where an SI-DMRS is transmitted.

The raw feature vector at a specific RE coordinate (n,k) is constructed by concatenating the flattened antennas and the real-imaginary components, and the mapping is formulated as

$$\boldsymbol{x}_{\text{raw}}^{(i)}[n,k] = \text{Concat}\left(\boldsymbol{\Phi}^{(i-1)}[n,k], \tilde{\boldsymbol{s}}^{(i-1)}[n,k], \boldsymbol{\Omega}_{\text{p}}[n,k]\right), \tag{21}$$

where the symbol vector is reconstructed according to the SI-DMRS scheme and power allocations as

$$\tilde{\boldsymbol{s}}_{n,k}^{(i)} = \begin{cases} \sqrt{\rho}\, s_{\text{DMRS}} + \sqrt{1-b}\, \tilde{\boldsymbol{x}}_{n,k}^{(i-1)}, & \text{if } \boldsymbol{\Omega}_{\text{p}}(n,k) = 1 \\ a\tilde{\boldsymbol{x}}_{n,k}^{(i-1)}, & \text{Otherwise} \end{cases}. \tag{22}$$

At initial stage ($i=0$), the tensor $\boldsymbol{\Phi}^{(-1)}[n,k]$ is initialized by $\boldsymbol{Y}_{\mathbb{R}}[n,k]$, and $\tilde{\boldsymbol{x}}^{(-1)}[n,k]=\boldsymbol{0}$ such that $\tilde{\boldsymbol{s}}_{n,k}^{(0)}$ only contains known DMRS. The dimension of $\boldsymbol{x}_{\text{raw}}^{(0)}[n,k]$ is $D_{\text{in}}^{(0)}=2N_{\text{r}}+3N_{\text{t}}$. In subsequent stages, the dimension $D_{\text{in}}^{(i)}$ expands to accommodate more features generated by the PGFC module.

A projection matrix $\mathbf{W}^{(i)} \in \mathbb{R}^{D_{\text{in}}^{(i)}\times d_{\text{virtual}}}$ is applied to map the raw feature vector into a d_{virtual}-dimensional latent space as $\boldsymbol{x}_{\text{raw}}^{(i)}[n,k]\mathbf{W}^{(i)}$. By aggregating all projected vectors across L REs, the tokenizer outputs a sequence $\boldsymbol{X}_{\text{token}}^{(i)} \in \mathbb{R}^{B\times L\times d_{\text{virtual}}}$ as input to the next ICED module. This tokenization strategy closely aligns with the physical characteristics of MIMO-OFDM systems. By flattening the

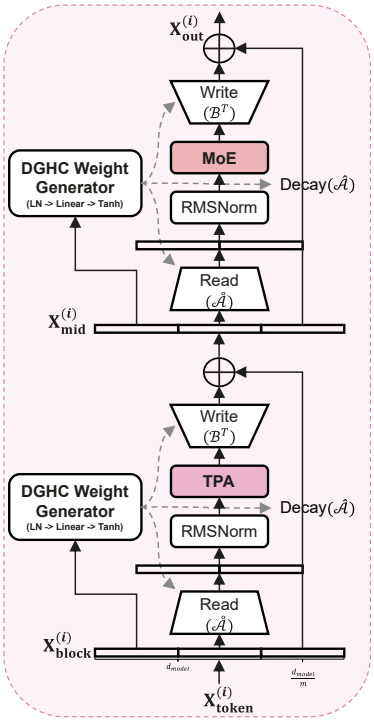

Fig. 7. The VWN-TPA-MoE architecture enhanced Transformer encoder design.

2D time-frequency grids into a sequence of length L, the self-attention mechanism is augmented with Rotary Position Embedding (RoPE) [36] and effectively captures the local coherence of MIMO channels. Furthermore, by collapsing spatial antennas into a single d_{virtual}-dimensional feature vector per token, the architecture delegates the resolution of spatial correlations to the network. This structural design enables the attention layers to perform MD internally within the high-dimensional latent representation.

B. ICED Backbone

The ICED operating on the tokenized sequence is designed as an unfolded cascaded architecture that performs CE and MD. Rather than treating them as disjoint modules, the AI-ICED design unifies them into a fused pipeline, as illustrated in Fig. 6. The Transformer encoder architecture incorporates with the latest VWN, TPA and MoE techniques to optimize feature representation.

The prompt $\mathbf{X}_{\text{token}}^{(i)}$ received from the tokenizer is firstly processed by N_{CE} stacked customized Transformer encoder blocks that incorporates VWN, TPA, and MOE. These specialized encoder blocks iteratively resolve the time-frequency correlations and spatial interference, yielding a channel latent representation $\boldsymbol{H}_{\text{feature}}^{(i)} \in \mathbb{R}^{B\times L\times d_{\text{virtual}}}$. To reconstruct the MIMO channel, a projection head (a linear fully-connected layer) $\mathbf{W}_{\text{CE}}^{(i)}$ maps $\mathbf{H}_{\text{feature}}^{(i)}$ back to physical antenna dimensions

$$\hat{\boldsymbol{H}}_{\text{flat}}^{(i)} = \boldsymbol{H}_{\text{feature}}^{(i)} \boldsymbol{W}_{\text{CE}}^{(i)} \in \mathbb{R}^{B\times L\times (2N_{\text{r}}N_{\text{t}})}. \tag{23}$$

It is subsequently reshaped to the real-valued CE tensor $\hat{\boldsymbol{H}}^{(i)} \in \mathbb{R}^{B\times N_{\text{r}}\times N_{\text{t}}\times N_{\text{sym}}\times N_{\text{sc}}\times 2}$.

Instead of solely using $\hat{\mathbf{H}}_{\text{flat}}$, the input to MD module is formulated via an additive residual bridge

$$\boldsymbol{Z}_{\text{dd}}^{(i)} = \boldsymbol{X}_{\text{token}}^{(i)} + \boldsymbol{H}_{\text{feature}}^{(i)} + \hat{\boldsymbol{H}}_{\text{flat}}^{(i)} \boldsymbol{W}_{\text{re-embed}}^{(i)}, \tag{24}$$

where $\boldsymbol{W}_{\text{re-embed}}^{(i)}$ acts as a re-embedding layer that projects the CE back into the high-dimensional latent space. This design linearly aggregates the original input prompt $\mathbf{X}_{\text{token}}^{(i)}$, the uncompressed high-dimensional channel memory $\mathbf{H}_{\text{feature}}^{(i)}$, ensuring maximum information retention[5].

The fused sequence $\mathbf{Z}_{\text{dd}}^{(i)}$ is subsequently processed by N_{MD} transformer encoder blocks, which apply the same VWN-TPA-MOE design to detect the symbols. Finally, a classification head maps the output features into logits for the real and imaginary pulse amplitude modulation (PAM) symbols:

$$\boldsymbol{Z}_{\text{PAM}}^{(i)} = \text{Encoder}_{\text{MD}}\big(\boldsymbol{Z}_{\text{dd}}^{(i)}\big) \boldsymbol{W}_{\text{MD}}^{(i)}, \tag{25}$$

where $\boldsymbol{W}_{\text{MD}}^{(i)}$ is the projection weight, and $\sqrt{M}$ represents the PAM alphabet size converted from the M-QAM modulation. The output tensor $\boldsymbol{Z}_{\text{PAM}}^{(i)} \in \mathbb{R}^{B \times L \times (2N_{\text{t}}\sqrt{M})}$ encompasses the real and imaginary logits, $\boldsymbol{Z}_{\text{Re}}^{(i)}$ and $\mathbf{Z}_{\text{Im}}^{(i)}$, which are utilized for loss computation. By applying standard softmax normalization, these logits are mapped to the probability tensor $\hat{\boldsymbol{P}}^{(i)}$.

C. Physics Guided Feature Construction (PGFC)

While the VWN-TPA-MoE encoders provide a powerful computational engine, a purely data-driven approach faces limitations in modeling the high-dimensional state space of MIMO-OFDM detection. Rather than propagating raw outputs across refinement stages, PGFC functions as a differentiable bridge. It constructs communication-theoretic tensors derived from current channel and symbol estimates and supplies the subsequent VWN-TPA encoder with structured features. This relieves the network from learning fundamental spatial projections strictly from data, thereby accelerating convergence and improving generalization across varying channel conditions.

Using the probability tensor $\hat{\boldsymbol{P}}^{(i)}$ output from the ICED module, the soft estimates $\tilde{\boldsymbol{x}}^{(i)}$ of the transmitted data can be computed. Based on these estimates, several features are assembled for the next stage: an interference cancellation (IC) result, $\boldsymbol{R}^{(i)} = \boldsymbol{Y} - \hat{\boldsymbol{H}}^{(i)} \tilde{\boldsymbol{S}}^{(i)}$; matched-filter (MF) outputs, $\boldsymbol{Z}_{\text{MF}}^{(i)} = (\hat{\boldsymbol{H}}^{(i)})^{\mathsf{H}} \boldsymbol{Y}$ and $\boldsymbol{R}_{\text{MF}}^{(i)} = (\hat{\boldsymbol{H}}^{(i)})^{\mathsf{H}} \boldsymbol{R}^{(i)}$; and the Gram matrix $\boldsymbol{G}^{(i)} = (\hat{\boldsymbol{H}}^{(i)})^{\mathsf{H}} \hat{\boldsymbol{H}}^{(i)}$. After converting them into their real-valued versions, the tensor $\boldsymbol{\Phi}^{(i)}$ in (21) sent to the tokenizer is augmented as follows:

$$\boldsymbol{\Phi}^{(i)} = \text{Concat}\big(\boldsymbol{Y}_{\mathbb{R}},\ \boldsymbol{R}_{\mathbb{R}}^{(i)},\ \boldsymbol{Z}_{\text{MF},\mathbb{R}}^{(i)},\ \boldsymbol{G}_{\mathbb{R}}^{(i)},\ \boldsymbol{R}_{\text{MF},\mathbb{R}}^{(i)}\big). \tag{26}$$

[5]However, the purpose of channel projection head and re-embedding is just to obtain the estimate of $\boldsymbol{H}$ and measure the CE-MSE for training loss optimization during training, once the NN is trained, these two modules can be combined together as one in inference stage.

Combined with the updated soft state $\tilde{\mathbf{s}}^{(i)}$ and the pilot mask $\mathbf{\Omega}_{\mathrm{p}}$ are forwarded to the tokenizer for the next ICED stage[6].

D. Training Loss Design

To simultaneously optimize CE and MD, the training loss at iteration stage i balances these objectives using a hyper-parameter $\beta_1 \in [0,1]$:

$$\mathcal{L}_{\text{stage}}^{(i)} = \beta_1 \mathcal{L}_{\boldsymbol{H}}^{(i)} + (1-\beta_1)\mathcal{L}_{\boldsymbol{S}}^{(i)}, \tag{27}$$

where the detection loss is measured via cross-entropy as

$$\mathcal{L}_{\boldsymbol{S}}^{(i)} = \mathrm{CrossEntropy}\big(\boldsymbol{Z}_{\text{PAM}}^{(i)}, \boldsymbol{Z}_{\text{PAM,Label}}\big), \tag{28}$$

and the CE loss is computed using the MSE:

$$\mathcal{L}_{\boldsymbol{H}}^{(i)} = \mathrm{MSE}\big(\hat{\boldsymbol{H}}^{(i)}, \mathbf{H}_{\text{Label}}\big). \tag{29}$$

To further mitigate vanishing gradients across the unfolded cascade, deep supervision injects gradients into all intermediate stages. Controlled by a weighting factor $\beta_2 \in [0,1]$, the total loss over $N_{\mathrm{it}}{+}1$ stages is defined as

$$\mathcal{L}_{\text{total}} = (1-\beta_2)\mathcal{L}_{\text{stage}}^{(N_{\mathrm{it}})} + \frac{\beta_2}{N_{\mathrm{it}}} \sum_{i=0}^{N_{\mathrm{it}}-1} \mathcal{L}_{\text{stage}}^{(i)}. \tag{30}$$

E. Inference Complexity

A critical challenge for AI-based receivers is inference complexity. The per-stage computational complexity is primarily bounded by attention-score computation, low-rank TPA projections, and MoE operations. As the unfolded receiver iterates over $N_{\mathrm{it}}+1$ stages, the total inference cost scales linearly with the number of applied stages .To manage this complexity, the architecture incorporates three key design strategies: VWN expands the latent memory capacity to $d_{\text{virtual}}{=}(n/m)d_{\text{model}}$, but restricts dense nonlinear operations to m out of n virtual slots. This effectively decouples representational capacity from active computational width. TPA factorizes the Query, Key, and Value projections into low-rank tensor products. Given a query rank r_q and a shared key/value rank r, the dense projection complexity drops from $\mathcal{O}(Ld_{\text{model}}^2)$ to $\mathcal{O}(Ld_{\text{model}}(r_q{+}2r))$. Meanwhile, the $\mathcal{O}(L^2 d_{\text{model}})$ attention-score term is retained to capture global time-frequency dependencies. The MoE feed-forward block expands parameter capacity via expert specialization. By activating exactly N_s shared experts and K_r routed experts per token, the active computation maintains FLOP equivalence with a standard dense Transformer FFN. A list of the complexity is summarized in Table I in Appendix I.

[6]A stop-gradient operation is applied to $\tilde{\boldsymbol{s}}^{(i)}$ to isolate the backward pass within each iterative stage, stabilizing the deep supervision dynamics across the unfolded cascade

V. Numerical Results

Next we analyse the performance of the AI-ICED receiver under SI-DMRS and 5G-NR settings. The configurations and channel models are aligned with 3GPP specifications and are summarized in Table II, and the hyper-parameter configurations for the AI-ICED receiver construction are summarized in Table 3, respectively. The basic configurations are the same as depicted in Fig. 1, where we assume a bandwidth of two PRBs, and two OFDM symbols are used for SI-DMRS transmissions (so $M=48$), regardless of N_{t}, and $N=240$. So the maximal SE gain can be obtained is 20%. Performance is characterized in terms of CE-MSE, BLER, and throughput for different SNRs and code-rates. For all SI-DMRS configurations, the DMRS pattern adopts the layout shown in Fig. 1. Based on the theoretical results in Sec. III, which are also validated by our tests, the power allocation factor is set to $b=0.8$, assigning a larger proportion of the transmit power to the pilot symbols. For a fair comparison with the 5G-NR baseline, no power boosting is applied (i.e., $a=1$).

The multipath fading channels are modelled as Extended Typical Urban (ETU70) [34]. For evaluation, the generated datasets comprise 800,000 frames for the 2×2 MIMO setup, alongside 200,000 subframes for the 4×4 MIMO configuration. All datasets are partitioned into 81% for training, 9% for validation, and 10% for testing. Finally, the block error rate (BLER) and throughput of the proposed architecture are benchmarked against classical baselines including conventional LMMSE and the optimal log-map based maximum likelihood detector (MLD). Furthermore, to decouple the errors induced by CE from detection, a genie-aided variant of the proposed architecture is tested. It bypasses the estimated channel tensor $\hat{\boldsymbol{H}}_{\text{flat}}$ and directly injects the ground-truth $\boldsymbol{H}$ into the latent residual fusion bridge, which reflects the MIMO detection performance under genie-aided channel state information (CSI).

Note that the same AI-ICED design is also applied to the 5G-NR baseline simulations using separate training sessions, which corresponds to a special case of SI-DMRS where $b=1$. Consequently, the overarching design principles and AI receiver architecture apply equally to standard 5G-NR configurations.

A. CE-MSE and BLER Convergences of the Proposed AI-ICED Receiver

The CE-MSE and BLER performance of the proposed AI-ICED receiver under a 4×4 MIMO and 16-QAM modulation (4-PAM) are presented in Fig. 8 and Fig. 9, respectively, using an LDPC code-rate of $2/3$. The AI-ICED receiver is trained at an SNR of 16dB, but evaluated across a wide SNR range. As shown, both the CE-MSE and BLER converge rapidly within a single iteration. With two additional iterations, the performance gain is approximately 0.5dB in SNR for both metrics, highlighting the effectiveness of the iterative AI-ICED design. Furthermore, as illustrated in Fig. 9, the AI-ICED receiver approaches the performance of the optimal MLD provided with genie-aided CSI.

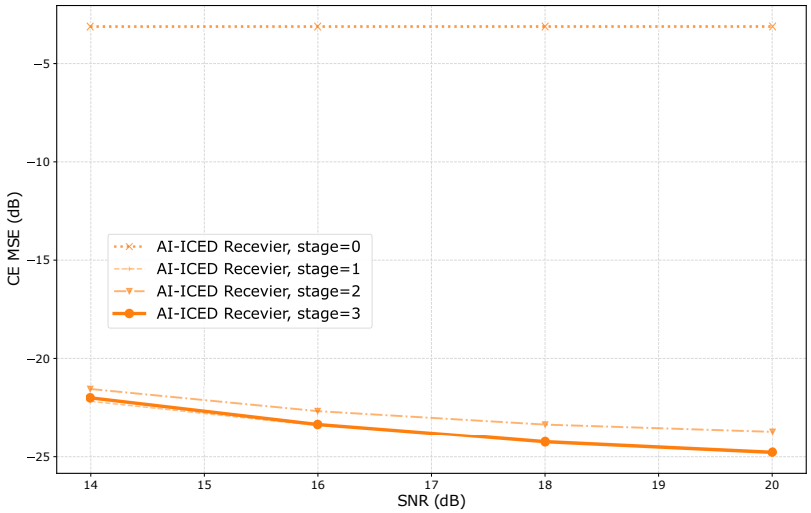

Fig. 8. CE-MSE over iteration stages for 4×4 MIMO and 16QAM modulation.

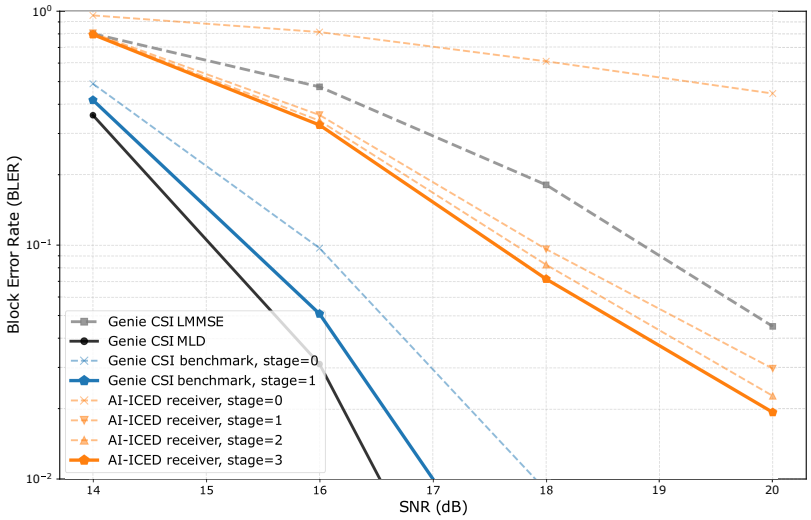

Fig. 9. BLER performance with the same configurations in Fig. 8.

In Fig. 10, the CE-MSE across a wide SNR range is presented, along with the corresponding required operational SNR for different code-rates. Although the AI-ICED receiver trained at a fixed SNR can generalize to adjacent SNRs, train-test SNR mismatch prevents optimal detection over the entire operating range. To handle this mismatch, dedicated models are trained at regular SNR intervals to capture variations and maintain optimality. As shown, due to the interference introduced between DMRS and data symbols under SI-DMRS, there is a CE-MSE degradation of approximately 2–4 dB compared to the 5G-NR baseline. However, the CE-MSE error remains insignificant compared to the noise power. Furthermore, this performance loss is well justified by the increased SE and higher throughput achieved by the AI-ICED receiver as shown later.

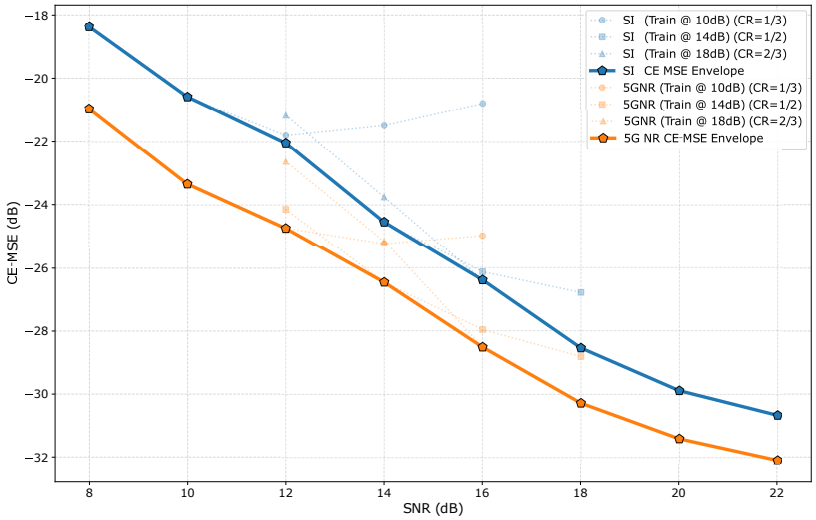

Fig. 10. CE-MSE under 2×2 MIMO and 16-QAM across a wide SNR range.

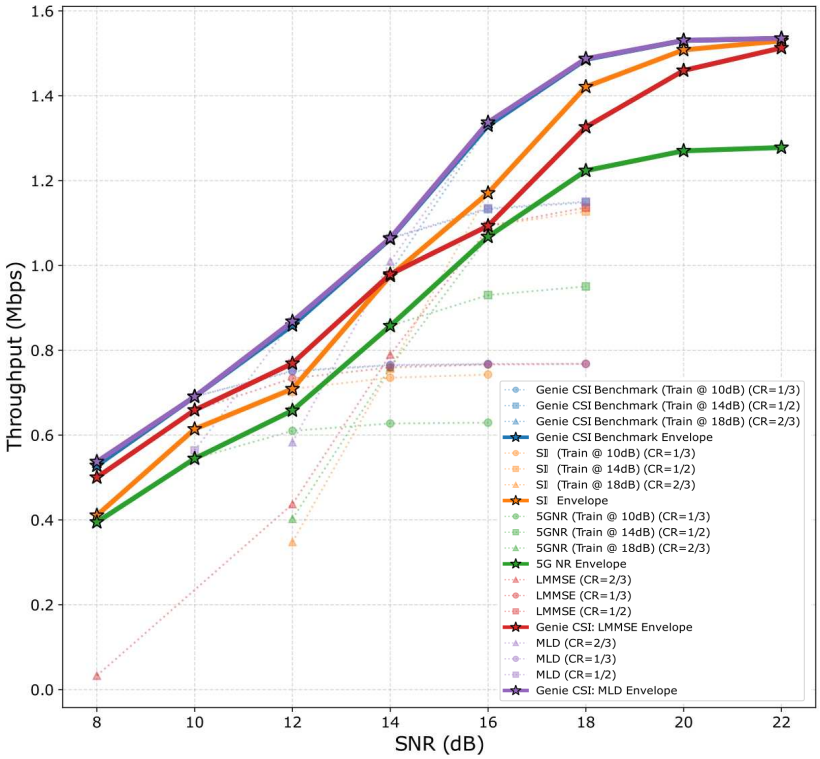

Fig. 11. Throughput envelopes under the 2×2 and 16-QAM modulation.

B. Throughput Increments

Although both the CE-MSE and BLER can be worse than 5G-NR baselines without superimposed DMRS, the primary advantage of SI-DMRS is the increased SE resulting from its zero DMRS overhead. The throughput envelopes are illustrated in Fig. 11 and Fig. 12 for the 2×2 and 4×4 cases for LDPC code-rates (1/3, 1/2, 2/3) corresponding to low, medium, and high code-rates, respectively, The proposed AI-ICED receiver is trained at SNR points (10dB, 14dB, 16dB) to cover different SNR range and code-

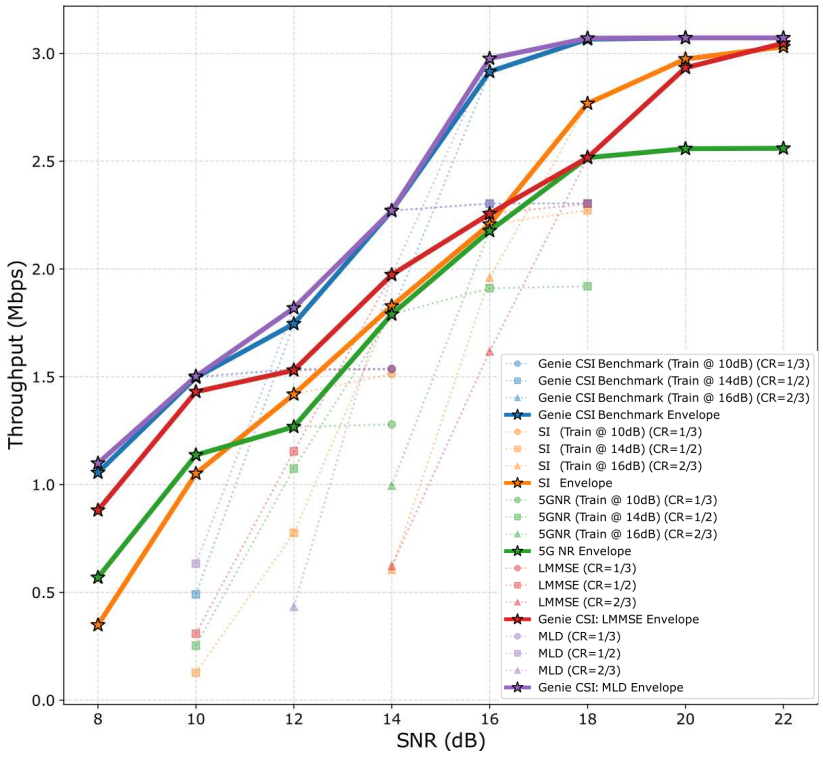

Fig. 12. Throughput envelopes under the 4×4 and 16-QAM modulation.

rates. As shown, the throughputs achieved using SI-DMRS (represented by the yellow curves) combined with the AI-ICED receiver consistently outperform the 5G-NR baseline (the green curves), even though both utilize the proposed AI-ICED receivers. Notably, as the SNR increases, the throughput of the 5G-NR approach saturates due to DMRS overheads, whereas transmissions utilizing SI-DMRS achieve full throughput.

Conversely, when using genie CSI, the AI-ICED receiver simplifies to an AI-based MIMO detector, performing (illustrated by the blue curves) nearly as well as the performance bound, namely the genie-CSI-aided MLD (the purple curves). This highlights the effectiveness of AI-ICED in approaching optimal detector performance with genie CSI. Additionally, compared against the baseline genie-CSI-aided LMMSE detector (the red curves), the proposed AI-ICED receiver demonstrates superior performance in middle-to-high SNR regions where detector capability is critical.

VI. Summary

We have introduced a Transformer encoder based AI-ICED receiver that overcomes DMRS overheads in MIMO-OFDM systems. By integrating the latest VWN, TPA, and MoE architectures, the proposed AI-ICED receiver design reformulates CE and MD as a joint contextual learning process, utilizing a PGFC bridge and an iterative unfolded cascade that is highly resilient to pilot sparsity. With the AI-ICED receiver, both the CE-MSE and BLER converge rapidly within a few iterations, yielding significant

performance gains. Furthermore, it achieves detection performance close to that of the optimal MLD when given genie CSI input. Validated by throughput envelopes, the SI-DMRS scheme, combined with the AI-ICED receiver, delivers higher throughput and SE compared to conventional 5G-NR baselines with non-superimposed DMRS and data transmission.

A key question regarding the SI-DMRS scheme involves the power allocation between DMRS and data symbols, as CE and MD are entangled. We have derived the CE-MSE and MD-MSE in approximated closed forms and developed an analytical framework to directly evaluate the final equilibrium states under an ICED process, eliminating the need for numerical simulations. Studies indicate that allocating a power factor of around 0.8 to DMRS and 0.2 to data provides an effective design. Furthermore, we have extended the analysis to cases where the power on REs with SI-DMRS can be boosted by borrowing power from remaining REs, yielding substantial additional gains. To evaluate the final decoding performance, we have adopted MIESM to combine MD-MSEs from the two sets of REs: those containing SI-DMRS and those without.

APPENDICES

A. Derivations of CE-MSE

Letting

$$\boldsymbol{z}_1 = \sqrt{\frac{(1-b)k}{N_{\mathrm{t}}}}(\boldsymbol{H}\boldsymbol{x} - \tilde{\boldsymbol{H}}\tilde{\boldsymbol{x}}) + \boldsymbol{w}, \tag{31}$$

and since the estimation error $\boldsymbol{\Delta H}$ is orthogonal to $\tilde{\boldsymbol{H}}$ from the orthogonality principle, it holds that $\mathbb{E}\{\boldsymbol{\Delta H}\tilde{\boldsymbol{H}}^{\dagger}\} = \mathbf{0}$. Similarily, $\mathbb{E}\{\boldsymbol{\Delta x}\tilde{\boldsymbol{x}}^{\dagger}\} = \mathbf{0}$. Hence, it holds that

$$\begin{aligned}\mathbb{E}\{\boldsymbol{z}_1\boldsymbol{z}_1^{\dagger}\} \;=\; & \frac{(1-b)k}{N_{\mathrm{t}}}\Big(\mathbb{E}\{\boldsymbol{H}\boldsymbol{H}^{\dagger}\}\mathbb{E}\{\Delta\boldsymbol{x}\Delta\boldsymbol{x}^{\dagger}\} \\ & + \mathbb{E}\{\Delta\boldsymbol{H}\Delta\boldsymbol{H}^{\dagger}\}\mathbb{E}\{\boldsymbol{x}\boldsymbol{x}^{\dagger}\} \\ & - \mathbb{E}\{\Delta\boldsymbol{H}\Delta\boldsymbol{H}^{\dagger}\}\mathbb{E}\{\Delta\boldsymbol{x}\Delta\boldsymbol{x}^{\dagger}\}\Big) + \sigma^2\boldsymbol{I} \\ =\; & \big(k(1-b)(p+q-pq) + \sigma^2\big)\boldsymbol{I},\end{aligned} \tag{32}$$

where

$$\mathbb{E}\{\Delta\boldsymbol{H}\Delta\boldsymbol{H}^{\dagger}\}\} = N_{\mathrm{t}}\mathbb{E}\{\Delta\boldsymbol{h}\Delta\boldsymbol{h}^{\dagger}\}\} = N_{\mathrm{t}}p\boldsymbol{I} \tag{33}$$

$$\mathbb{E}\{\Delta\boldsymbol{x}\Delta\boldsymbol{x}^{\dagger}\}\} = q\boldsymbol{I}. \tag{34}$$

With the LMMSE estimator, the CE-MSE equals

$$\begin{aligned}\mathbb{E}\{\boldsymbol{\Delta h}\boldsymbol{\Delta h}^{\dagger}\} &= \left(bk^{\dagger}\mathbb{E}\{\boldsymbol{z}_1\boldsymbol{z}_1^{\dagger}\}^{-1}+\boldsymbol{I}\right)^{-1} \\ &= \left(\frac{bk}{k(1-b)(p+q-pq)+\sigma^2}+1\right)^{-1}\boldsymbol{I} \\ &= \frac{k(1-b)(p+q-pq)+\sigma^2}{bk+k(1-b)(p+q-pq)+\sigma^2}\boldsymbol{I}. \end{aligned} \tag{35}$$

B. Derivations of MD-MSE

Similar, letting

$$\boldsymbol{z}_2 = \sqrt{\frac{(1-b)k}{N_{\mathrm{t}}}}\boldsymbol{\Delta H}\boldsymbol{x}+\sqrt{bk}\boldsymbol{\Delta h}s+\boldsymbol{w}, \tag{36}$$

it holds that

$$\begin{aligned}\mathbb{E}\{\boldsymbol{z}_2\boldsymbol{z}_2^{\dagger}\} &= \frac{(1-b)k}{N_{\mathrm{t}}}\mathbb{E}\{\Delta\boldsymbol{H}\Delta\boldsymbol{H}^{\dagger}\}\mathbb{E}\{\boldsymbol{x}\boldsymbol{x}^{\dagger}\} \\ &\quad +bk\mathbb{E}\{\Delta\boldsymbol{h}\Delta\boldsymbol{h}^{\dagger}\}+\sigma^2\boldsymbol{I} \\ &= (1-b)k\mathbb{E}\{\Delta\boldsymbol{h}\Delta\boldsymbol{h}^{\dagger}\}+bk\mathbb{E}\{\Delta\boldsymbol{h}\Delta\boldsymbol{h}^{\dagger}\}+\sigma^2\boldsymbol{I} \\ &= (kp+\sigma^2)\boldsymbol{I}. \end{aligned} \tag{37}$$

With an LMMSE estimator, the MD-MSE equals

$$\begin{aligned}\mathbb{E}\{\boldsymbol{\Delta x}\boldsymbol{\Delta x}^{\dagger}\} &= \left(\frac{(1-b)k}{N_{\mathrm{t}}}\tilde{\boldsymbol{H}}^{\dagger}\mathbb{E}\{\boldsymbol{z}_2\boldsymbol{z}_2^{\dagger}\}^{-1}\tilde{\boldsymbol{H}}+\boldsymbol{I}\right)^{-1} \\ &= \left(\frac{(1-b)k}{N_{\mathrm{t}}(kp+\sigma^2)}\tilde{\boldsymbol{H}}^{\dagger}\tilde{\boldsymbol{H}}+\boldsymbol{I}\right)^{-1}. \end{aligned} \tag{38}$$

Note that an instantaneous MD-MSE depends on the MIMO channel realization $\boldsymbol{H}$, although the expectation has been taken over $\boldsymbol{x}$, $\boldsymbol{w}$ and $\boldsymbol{\Delta h}$. By Jensen's inequality, the ergodic MD-MSE with taking expectation over $\boldsymbol{H}$ holds that

$$\mathbb{E}_{\boldsymbol{H}}\{\mathbb{E}\{\boldsymbol{\Delta x}\boldsymbol{\Delta x}^{\dagger}\}\} \geq \left(\frac{(1-b)k}{N_{\mathrm{t}}(kp+\sigma^2)}\mathbb{E}\{\tilde{\boldsymbol{H}}^{\dagger}\tilde{\boldsymbol{H}}\}+\boldsymbol{I}\right)^{-1}. \tag{39}$$

Noting that

$$\mathbb{E}\{\boldsymbol{H}^{\dagger}\boldsymbol{H}\} = \mathbb{E}\{\boldsymbol{H}^{\dagger}\boldsymbol{H}\}-\mathbb{E}\{\boldsymbol{\Delta H}^{\dagger}\boldsymbol{\Delta H}\} = N_{\mathrm{t}}(1-q)\boldsymbol{I}, \tag{40}$$

the ergodic MD-MSE is bounded as

$$\mathbb{E}_{\boldsymbol{H}}\{\mathbb{E}\{\boldsymbol{\Delta x}\boldsymbol{\Delta x}^{\dagger}\}\} \geq \frac{kp+\sigma^2}{k(1-b+bp)+\sigma^2}\boldsymbol{I}. \tag{41}$$

In order to analyse the general iterative behaviour, we approximate $\mathbb{E}\{\boldsymbol{\Delta x}\boldsymbol{\Delta x}^{\dagger}\}$ by its bound, and assuming

$$\mathbb{E}\{\boldsymbol{\Delta x}\boldsymbol{\Delta x}^{\dagger}\} \approx \frac{kp+\sigma^2}{k(1-b+bp)+\sigma^2}\boldsymbol{I}. \tag{42}$$

C. Proof of Property 1

Note that the equilibrium state of CE-MSE and MD-MSE is defined by the equations

$$p = \frac{\gamma N_{\mathrm{t}}}{M} \frac{k(1-b)(p+q-pq)+\sigma^2}{kb+k(1-b)(p+q-pq)+\sigma^2}, \tag{43}$$

$$q = \frac{kp+\sigma^2}{k(1-b+pb)+\sigma^2}. \tag{44}$$

With substitutions $u = 1-b$, $v = \frac{\sigma^2}{k}$, and $\rho = \frac{\gamma N_t}{M}$, it holds that

$$p+q-pq = \frac{p(u+1)-up^2+v}{u+pb+v}. \tag{45}$$

Inserting it into the p-equation and collecting powers of p yields the cubic polynomial

$$Ap^3 + Bp^2 + Cp + D = 0, \tag{46}$$

with coefficients A, B, C, D defined in Property 1.

D. Derivations of MD-MSE on REs without DMRS

Letting

$$\boldsymbol{z}_3 = \sqrt{\frac{a}{N_{\mathrm{t}}}} \boldsymbol{\Delta H} \boldsymbol{x} + \boldsymbol{w}, \tag{47}$$

it holds that

$$\begin{aligned} \mathbb{E}\{\boldsymbol{z}_3 \boldsymbol{z}_3^{\dagger}\} &= \frac{a}{N_{\mathrm{t}}} \mathbb{E}\{\Delta \boldsymbol{H} \Delta \boldsymbol{H}^{\dagger}\} \mathbb{E}\{\boldsymbol{x}\boldsymbol{x}^{\dagger}\} + \sigma^2 \boldsymbol{I} \\ &= (ap+\sigma^2)\boldsymbol{I}. \end{aligned} \tag{48}$$

With an LMMSE estimator, the MD-MSE equals

$$\begin{aligned} \mathbb{E}\{\boldsymbol{\Delta x} \boldsymbol{\Delta x}^{\dagger}\} &= \left(\frac{a}{N_{\mathrm{t}}} \tilde{\boldsymbol{H}}^{\dagger} \mathbb{E}\{\boldsymbol{z}_3 \boldsymbol{z}_3^{\dagger}\}^{-1} \tilde{\boldsymbol{H}} + \boldsymbol{I} \right)^{-1} \\ &= \left(\frac{a}{N_{\mathrm{t}}(ap+\sigma^2)} \tilde{\boldsymbol{H}}^{\dagger} \tilde{\boldsymbol{H}} + \boldsymbol{I} \right)^{-1}. \end{aligned} \tag{49}$$

Following the same argumentations in Appendix-B, we approximate it as

$$\mathbb{E}\{\boldsymbol{\Delta x} \boldsymbol{\Delta x}^{\dagger}\} \approx \frac{ap+\sigma^2}{a(1+p)+\sigma^2}. \tag{50}$$

E. MIESM

For a modulation order M (number of bits carried by each constellation symbol), the constellation constrained MI mapping function $f(\theta)$ is commonly approximated as

$$f(\gamma) \approx \log_2(M) \left(1 - \sum_{k=1}^{K} a_k e^{-b_k \theta}\right), \tag{51}$$

where

$$\sum_{k=1}^{K} a_k = 1. \tag{52}$$

For the widely used exponential fit with $K=3$, the paramters are listed in Table III.

F. VWN

Standard Transformer encoders use a fixed hidden dimension (d_{model}) across their layers. Increasing this dimension to handle complex MIMO-OFDM wireless channels causes a quadratic rise in computing cost. As shown in Fig. 7, VWN separate memory capacity from computational width. The input token $\boldsymbol{X}_{\text{token}}^{(i)}$ is expanded to a larger dimension $d_{\text{virtual}} = (n/m) \cdot d_{\text{model}}$ (where $n > m$) and reshaped into n blocks:

$$\boldsymbol{X}_{\text{block}}^{(i)} \in \mathbb{R}^{B \times L \times n \times d_{\text{b}}}$$

Here, the block dimension is $d_{\text{b}} = d_{\text{model}}/m$.

Inside each VWN layer, heavy non-linear operations $\mathcal{F}(\cdot)$ (such as TPA or MoE) are restricted to a narrow m-slot computational subspace. Dynamic generalized hyper-connections (DGHC) [30] manage this interaction. A lightweight linear projection and Tanh-activation on the normalized input dynamically generate a write matrix $\mathcal{B} \in \mathbb{R}^{m \times n}$ and a joint transformation matrix $\mathcal{A} \in \mathbb{R}^{(m+n) \times n}$. Horizontally partitioning $\mathcal{A}$ produces the read component $\mathring{\mathcal{A}} \in \mathbb{R}^{m \times n}$ and the decay residual $\hat{\mathcal{A}} \in \mathbb{R}^{n \times n}$. The core read-compute-write cycle is executed as

$$\mathbf{X}_{\text{out}}^{(i)} = \mathcal{B}^{\top} \mathcal{F}\big(\mathring{\mathcal{A}} \mathbf{X}_{\text{block}}^{(i)}\big) + \hat{\mathcal{A}} \mathbf{X}_{\text{block}}^{(i)}. \tag{53}$$

It outlines the data flow: the network tracks complex multi-path residuals across the n-dimensional highway, while isolating expensive dense computations to the efficient m-dimensional subspace through dynamic gating.

G. TPA

While the VWN architecture manages macroscopic memory capacity, each layer's core process relies on TPA. Standard Multi-Head Attention (MHA) [35] uses monolithic parameter matrices, limiting its inductive bias. TPA addresses this by factorizing queries, keys, and values into sums of contextual tensor products [31], improving parameter efficiency and injecting physical priors.

Contextual Factorization: Given the intermediate state $\mathbf{X} \in \mathbb{R}^{B\times L\times d_{\text{model}}}$ inside a VWN computational slot, TPA decomposes query, key, and value generation into independent latent factor maps. Let h denote the number of attention heads, $d_{\text{h}} = d_{\text{model}}/h$ the per-head dimension, and $r_{\text{q}}, r_{\text{k}}$ the rank constraints for queries and keys/values, respectively. For queries, two linear projections produce head-specific mixing weights $\mathbf{A}_{\text{Q}} \in \mathbb{R}^{B\times L\times h\times r_{\text{q}}}$ and a shared feature basis $\mathbf{B}_{\text{Q}} \in \mathbb{R}^{B\times L\times r_{\text{q}}\times d_{\text{h}}}$ such that

$$\mathbf{A}_{\text{Q}}(\mathbf{X}) \;=\; \mathbf{X}\mathbf{W}_{\text{A}}^{\text{Q}}, \tag{54}$$

$$\mathbf{B}_{\text{Q}}(\mathbf{X}) \;=\; \mathbf{X}\mathbf{W}_{\text{B}}^{\text{Q}}, \tag{55}$$

where $\mathbf{W}_{\text{A}}^{\text{Q}} \in \mathbb{R}^{d_{\text{model}}\times(h\cdot r_{\text{q}})}$ and $\mathbf{W}_{\text{B}}^{\text{Q}} \in \mathbb{R}^{d_{\text{model}}\times(r_{\text{q}}\cdot d_{\text{h}})}$ are factor weight matrices. Analogous projections with rank r_{k} yield $(\mathbf{A}_{\text{K}}, \mathbf{B}_{\text{K}})$ and $(\mathbf{A}_{\text{V}}, \mathbf{B}_{\text{V}})$.

Physical Adaptation with 2D-RoPE and Rank Scaling: Standard 1D rotary position embeddings [36] fail to preserve the 2D coherence patterns of MIMO-OFDM channels. To fix this, we treat the sequence axis as a 2D physical grid and apply RoPE_{2D} to the feature bases before tensor contraction:

$$\tilde{\mathbf{B}}_{\text{Q}} \;=\; \text{RoPE}_{\text{2D}}\big(\mathbf{B}_{\text{Q}}\big), \tag{56}$$

$$\tilde{\mathbf{B}}_{\text{K}} \;=\; \text{RoPE}_{\text{2D}}\big(\mathbf{B}_{\text{K}}\big). \tag{57}$$

The full query tensor $\mathbf{Q} \in \mathbb{R}^{B\times L\times h\times d_{\text{h}}}$ is reconstituted via tensor contraction

$$\mathbf{Q} = \mathbf{A}_{\text{Q}}(\mathbf{X})\tilde{\mathbf{B}}_{\text{Q}}. \tag{58}$$

For the i-th head at the t-th token, this represents a sum of outer products $\mathbf{Q}_t^{(i)} = \sum_{j=1}^{r_{\text{q}}} \mathbf{a}_{t,j}^{(i)} \otimes \mathbf{b}_{t,j}$.

Scaled Dot-Product Attention: Using the factorized tensors $\mathbf{Q}, \mathbf{K}, \mathbf{V} \in \mathbb{R}^{B\times L\times h\times d_{\text{h}}}$, attention is computed per head as

$$\text{head}_i = \text{Softmax}\left(\frac{\mathbf{Q}_i\mathbf{K}_i^{\top}}{\sqrt{d_{\text{h}}}}\right)\mathbf{V}_i, \tag{59}$$

where $\mathbf{Q}_i, \mathbf{K}_i, \mathbf{V}_i \in \mathbb{R}^{B\times L\times d_{\text{h}}}$ denote the slices along the head dimension. The outputs are concatenated and projected via $\mathbf{W}_{\text{O}} \in \mathbb{R}^{(h\cdot d_{\text{h}})\times d_{\text{model}}}$ as

$$\text{TPA}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Concat}\big(\text{head}_1, \ldots, \text{head}_h\big)\,\mathbf{W}_{\text{O}}. \tag{60}$$

This output completes the TPA layer before routing to the subsequent MoE block.

H. MoE

After the TPA process, the outputs pass through a MoE feed-forward block adapted from the DeepSeek-MoE framework [24]. The expert pool is divided into K_s always-active shared experts and K_r selectively routed experts, both utilizing a SwiGLU-based feed-forward structure [37]. For an input token $\boldsymbol{x} \in \mathbb{R}^{d_{\text{model}}}$, the MoE forward pass is given by

$$\mathbf{y} = \sum_{k=1}^{K_s} \text{MLP}_s^{(k)}(\mathbf{x}) + \sum_{k \in \mathcal{K}} g_k(\mathbf{x})\, \text{MLP}_r^{(k)}(\mathbf{x}), \tag{61}$$

where $\mathcal{K}$ represents the set of top-K_r experts selected by a lightweight gating network, with gating scores computed via a linear projection followed by a Sigmoid activation $\mathbf{s} = \text{Sigmoid}(\mathbf{x}\mathbf{W}_{\text{route}})$, and the top-$K_r$ entries are normalized to yield the final routing weights:

$$g_k(\mathbf{x}) = \frac{s_k}{\sum_{k \in \mathcal{K}} s_k + \epsilon}, \quad k \in \mathcal{K}, \tag{62}$$

where ϵ is a small constant ensuring numerical stability.

I. NN Configuration of the AI-ICED Receiver

The proposed AI-ICED receiver is configured using the hyperparameters listed in Table IV. The VWN chassis is configured with $m = 2$ memory slots and $n = 3$ active computational slots per layer, producing an expanded virtual dimension $d_{\text{virtual}} = (n/m) \cdot d_{\text{model}} = 768$. The TPA sub-layer operates with a query rank $r_q = 16$ and a key/value rank $r_k = 16$. The MoE sub-layer employs $N_s = 1$ shared expert, $N_r = 4$ routed experts, and $K_r = 2$ active experts per token. The per-expert hidden dimension scaled by a factor of $(8/3)/(K_r + N_s)$ to maintain equivalence with a dense SwiGLU FFN [37]. To effectively optimize the deep cascaded Transformer backbone, a hybrid optimization strategy is employed: Multi-dimensional weight matrices are updated using the momentum-based Muon optimizer [38]; while one-dimensional parameters, such as layer normalizations and biases, are routed to AdamW [39]. The learning rate follows a linear warmup at first 10% steps, followed by cosine annealing down to a minimum ratio of 0.01 of the peak value [40], [41].

REFERENCES

[1] Z. Fu, "Deep in-context learning (ICL) for wireless communications," Master's thesis, Dept. of electrical and information technology, Lund University, Lund, Sweden, Jun. 2026.

[2] X. Wang, L. Lu, Q. Li, Q. Sun, N. Shi, Z. Chen, and T. Sun, "A task-driven design approach for 6G AI-native architecture," *Engineering*, vol. 56, no. 1, pp. 87–103, 2026.

[3] Y. Wu *et al.*, "A comprehensive review of AI-native 6G: Integrating semantic communications, reconfigurable intelligent surfaces, and edge intelligence for next-generation connectivity," *Front. Commun. Netw.*, vol. 5, p. 1655410, 2024.

[4] Z. Qin, H. Ye, G. Y. Li, and B.-H. F. Juang, "Deep learning in physical layer communications," *IEEE Wireless Commun.*, vol. 26, no. 2, pp. 93–99, Apr. 2019.

[5] J. Sajid *et al.*, "Empowering embodied AI in 6G networks: Architecture, enablers, and open challenges," *arXiv preprint arXiv:2606.20592*, 2026.

[6] T. O'Shea and J. Hoydis, "An introduction to deep learning for the physical layer," *IEEE Trans. Cogn. Commun. Netw.*, vol. 3, no. 4, pp. 563–575, Dec. 2017.

[7] X. Lin, "An overview of the 3GPP study on artificial intelligence for 5G new radio," *arXiv preprint arXiv:2308.05315*, 2023.

[8] X. Qin, S. Hu, J. Zhang, J. Qian, and H. Wang, "AI receiver design with deep learning based channel estimation and MIMO detection," in *Proc. IEEE PIMRC*, 2024, pp. 1–7.

[9] X. Li, X. Zhou, Y. Cao, J. Zhang, C.-K. Wen, X. Li, and S. Jin, "Learning-aided iterative receiver for superimposed pilots: Design and experimental evaluation," *arXiv preprint arXiv:2507.10074*, 2025.

[10] X. Qin and S. Hu, "Dual-attention based 3D channel estimation," *arXiv preprint arXiv:2604.01769*, 2026.

[11] S. Cammerer *et al.*, "A neural receiver for 5G NR multi-user MIMO," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Kuala Lumpur, Malaysia, 2023, pp. 329–334.

[12] S. Hu, "Invariant transformation and resampling based epistemic-uncertainty reduction," *arXiv preprint arXiv:2602.23315*, 2026.

[13] F. B. Saghezchi, M. Pourghasemian, B. Ding, A. Abdi, B. Lee, and A. Baron, "AI-native radio transceiver signal processing for next-generation mobile communication systems," *IEICE Trans. Commun.*, vol. E109-B, no. 4, pp. 555–572, Apr. 2026.

[14] H. He, C.-K. Wen, S. Jin, and G. Y. Li, "A model-driven deep learning network for MIMO detection," in *Proc. IEEE Glob. Conf. Signal Inf. Process. (GlobalSIP)*, Anaheim, CA, USA, 2018, pp. 584–588.

[15] A. Mazumdar, C. N. Manchon, O. E. Barbu, and R. O. Adeogun, "Comparative evaluation of model based deep learning receivers in coded MIMO systems," in *Proc. IEEE Veh. Technol. Conf. (VTC-Fall)*, 2024, pp. 1–7.

[16] M. -H. Hsieh and C. -H. Wei, "Channel estimation for OFDM systems based on comb-type pilot arrangement in frequency selective fading channels," *IEEE Trans. Consum. Electron.*, vol. 44, no. 1, pp. 217–225, Feb. 1998.

[17] D. Wubben, R. Bohnke, V. Kuhn, and K.-D. Kammeyer, "MMSE-based lattice-reduction for near-ML detection of MIMO systems," in *Proc. ITG Workshop Smart Antennas*, 2004, pp. 106–113.

[18] S. Hu and F. Rusek, "A soft-output MIMO detector with achievable information rate based partial marginalization," *IEEE Trans. Signal Process.*, vol. 65, no. 6, pp. 1622–1637, Mar. 2017.

[19] K. Upadhya, S. A. Vorobyov, and M. Vehkapëä, "Superimposed pilots are superior for mitigating pilot contamination in massive MIMO," *IEEE Trans. Signal Process.*, vol. 65, no. 11, pp. 2917–2932, Jun. 2017.

[20] S. Rezaie, M. Honkala, D. Korpi, D. C. Melgarejo, T. Izydorczyk, D. Gold, and O.-E. Barbu, "Superimposed DMRS for spectrally efficient 6G uplink multi-user OFDM: Classical vs AI/ML receivers," *arXiv preprint arXiv:2506.20248*, 2025.

[21] X. Li, X. Zhou, J. Zhang, C.-K. Wen, and S. Jin, "AI-driven iterative receiver for superimposed pilot Schemes in MIMO-OFDM systems," in *Proc. IEEE WCNC*, 2025, pp. 1–6.

[22] K. Ying *et al.*, "Conditional diffusion model-driven sassive MIMO iterative detection," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Glasgow, United Kingdom, 2026, pp. 1–6.

[23] M. Abuthinien, S. Chen, and L. Hanzo, "Semi-blind joint maximum likelihood channel estimation and data detection for MIMO systems," *IEEE Signal Process. Lett.*, vol. 15, pp. 202–205, 2008.

[24] DeepSeek-AI *et al.*, "DeepSeek-V3.2: pushing the frontier of open large language models," *arXiv preprint arXiv:2512.02556*, 2025.

[25] A. Yang *et al.*, "Qwen3 technical report," *arXiv preprint arXiv:2505.09388*, 2025.

[26] Kimi Team *et al.*, "Kimi K2.5: Visual Agentic Intelligence," *arXiv preprint arXiv:2602.02276*, 2026.

[27] Z. Song, M. Zecchin, B. Rajendran, and O. Simeone, "Turbo-ICL: In-context learning-based turbo equalization," *arXiv preprint arXiv:2505.06175*, 2025.

[28] M. Shanmugam, S. Balaraman, and K. Ranganathan, "Improving channel equalization in cell-free MIMO networks using reconfigurable intelligent surfaces and in-context learning," *Ann. Telecommun.*, vol. 81, pp. 259–273, 2026.

[29] 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,

[30] Seed *et al.*, "Virtual Width Networks," *arXiv preprint arXiv:2511.11238*, 2025.

[31] Y. Zhang *et al.*, "Tensor product attention is all you need," *arXiv preprint arXiv:2501.06425*, 2026.

[32] D. Dai *et al.*, "DeepSeekMoE: Towards ultimate expert specialization in mixture-of-experts language models," *arXiv preprint arXiv:2401.06066*, 2024.

[33] 3GPP, "5G; NR; Physical channels and modulation," 3GPP Technical Specification TS 38.211, V18.6.0, Apr. 2025.

[34] 3GPP, "Evolved Universal Terrestrial Radio Access (E-UTRA); User Equipment (UE) radio transmission and reception," 3GPP Technical Specification TS 36.101, V18.9.0, Apr. 2025.

[35] A. Vaswani *et al.*, "Attention is all you need," in *Advances Neural Inf. Process. Syst. (NeurIPS)*, vol. 30, 2017, pp. 5998–6008.

[36] J. Su *et al.*, "Roformer: Enhanced transformer with rotary position embedding," *Neurocomputing*, vol. 568, p. 127063, 2024.

[37] N. Shazeer, "GLU variants improve Transformer," *arXiv preprint arXiv:2002.05202*, 2020.

[38] K. Jordan *et al.*, "Muon: An optimizer for hidden layers in neural networks," 2024. [Online] Available: *https://kellerjordan.github.io/posts/muon/*.

[39] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2019.

[40] P. Goyal *et al.*, "Accurate, large Minibatch SGD: Training ImageNet in 1 Hour," *arXiv preprint arXiv:1706.02677*, 2017.

[41] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," *arXiv preprint arXiv:1608.03983*, 2017.

[42] J. Yang, H. Zhao, W. Wang, and C. Zhang, "An effective SINR mapping models for 256QAM in LTE-Advanced system," in *IEEE Annu. Int. Symp. Pers., Indoor, Mobile Radio Commun. (PIMRC)*, Washington, DC, USA, 2014, pp. 343–347.

[43] N. Kim, Y. Lee, and H. Park, "Performance analysis of MIMO system with linear MMSE receiver," *IEEE Trans. Wireless Commun.*, vol. 7, no. 11, pp. 4474–4478, Nov. 2008.

TABLE I

DOMINANT INFERENCE COMPLEXITY PER STAGE.

Operation	**Dense Transformer block**	**Proposed backbone**
Latent memory width	d_{model}	$d_{\text{virtual}} = (n/m)d_{\text{model}}$
Q/K/V projections	$\mathcal{O}(Ld_{\text{model}}^2)$	$\mathcal{O}(Ld_{\text{model}}(r_q + 2r))$
FFN computation	$\mathcal{O}(Ld_{\text{model}}d_{\text{ff}})$	$\mathcal{O}(Ld_{\text{model}}d_{\text{ff}})$
Attention score	$\mathcal{O}(L^2 d_{\text{model}})$	$\mathcal{O}(L^2 d_{\text{model}})$

TABLE II

SYSTEM PARAMETERS.

Symbol	**Description**	**Value**
$N_{\text{r}} \times N_{\text{t}}$	MIMO size	2×2, 4×4
$\mathcal{A}$	Modulation scheme	16QAM, 64QAM
N_{sym}	Number of OFDM symbols carrying data	12
N_{sc}	Active subcarriers	24
N	Number of REs carrying only data	240
M	Number of REs carrying both SI-DMRS and data	48
L	Token lenght for the AI-ICED receiver	288
Codes	LDPC	Code-rates $1/3$, $1/2$, $2/3$
(a, b)	Power allocation factors	$(a = 1, b = 0.8)$
Channels	3GPP channel models	ETU-70Hz, EPA-5Hz

TABLE III

THREE-TERM EXPONENTIAL MI MAPPING PARAMETERS.

Modulation	a_1	b_1	a_2	b_2	a_3	b_3
QPSK	0.9810	1.0420	0.0190	4.2500	0.0000	0.0000
16-QAM	0.6053	0.2845	0.3541	1.2562	0.0406	5.3784
64-QAM	0.4076	0.0717	0.3951	0.3556	0.1973	1.8315

TABLE IV

HYPERPARAMETER CONFIGURATION OF THE PROPOSED AI-ICED RECEIVER

Symbol	**Description**	**Value**
	Transformer Backbone	
d_{model}	Latent dimension	512
N_{head}	Number of attention heads	8
m/n	VWN memory / compute slots	$2/3$
d_{virtual}	Virtual memory width $(n/m) \cdot d_{\text{model}}$	768
r_q/r	TPA query rank / key-value rank	$16/16$
N_s/N_r	Shared / routed experts (MoE)	$1/4$
K_r	Active routed experts per token	2
Expert factor	Per-expert hidden factor $(8/3)/(K_r + N_s)$	$8/9$
Dropout	Dropout probability	0.1
	Cascade Architecture	
$N_{\text{CE}}^{(0)}/N_{\text{MD}}^{(0)}$	Bootstrap CE / MD encoder layers	$3/3$
$N_{\text{CE}}^{(i)}/N_{\text{MD}}^{(i)}$	Refinement CE / MD encoder layers	$6/6$
N_{it}	Number of ICED iteration	1
	Training Protocol	
Epochs	Total training epochs	100
Batch size	Samples per batch	16
Optimizer	Hybrid	Muon + AdamW
LR (Muon)	Peak learning-rate for multi-dimensional weights	2×10^{-3}
LR (AdamW)	Peak learning-rate for embeddings, norms, biases	3×10^{-4}
Weight decay	ℓ_2 regularization penalty	0.01
β_1	Loss balance between bit cross-entropy and CE-MSE	0.1
β_2	Auxiliary stage weight	0.3