Deep In-Context Learning (ICL) for Wireless Communications

ZHONGWANG FU
MASTER´S THESIS
DEPARTMENT OF ELECTRICAL AND INFORMATION TECHNOLOGY
FACULTY OF ENGINEERING | LTH | LUND UNIVERSITY

Deep In-Context Learning (ICL) for Wireless Communications

Zhongwang Fu

Department of Electrical and Information Technology
Lund University

Advisor: Juan Vidal Alegría
Sha Hu †
Jonathan Lindberg †

Examiner: Michael Lentmaier

June 10, 2026

† This thesis is completed as an intern project at Nyquist Research Center, Huawei Technologies Sweden AB. Sha Hu and Jonathan Lindberg are industrial supervisors.

Printed in Sweden
Tryckeriet i E-huset, Lund

Abstract

In multiple-input multiple-output orthogonal frequency-division multiplexing (MIMO-OFDM) systems, the spectral overhead required by orthogonal Demodulation Reference Signals (DMRS) limits overall system capacity. Traditional linear estimators experience severe degradation under high pilot sparsity or non-orthogonal superposition. To address these limitations, this thesis proposes a Physics-Guided In-Context Learning (ICL) receiver that jointly performs channel estimation (CE) and MIMO detection. By formulating reference signals as contextual prompts through a tokenization scheme, the proposed artificial intelligence (AI) receiver circumvents the strict requirement of classical spatial orthogonality.

The architecture adopts an unfolded iterative receiver design, where a Transformer-based backbone with Virtual Width Networks (VWN), Tensor Product Attention (TPA), and Mixture-of-Experts (MoE) layers performs joint representation learning for CE and detection. A differentiable Physics-Guided Feature Construction (PGFC) bridge converts intermediate outputs into soft symbol estimates and explicit physical features, which are then fed into subsequent stages for refinement. Additionally, a novel resampling strategy is adopted during inference to improve detection reliability without requiring any network retraining.

The proposed architecture is evaluated through link-level simulations adopting standard 3rd Generation Partnership Project (3GPP) fading channels (e.g., EPA and ETU) across various MIMO-OFDM configurations and modulations. Our results indicate that the proposed architecture approaches Maximum Likelihood Detection (MLD) in baseline configurations. The receiver maintains detection capability under sparse pilot conditions where standard 5G-NR baselines become intractable, and it mitigates pilot-data self-interference in superimposed DMRS scenarios. By achieving reliable detection with reduced pilot overhead, the framework translates DMRS overhead savings into throughput improvements, presenting a scalable AI receiver architecture for spectrally efficient future wireless networks.

Popular Science Summary

Modern wireless communication systems need to transmit more and more data through limited radio spectrum. One important challenge is that a wireless receiver must first understand how the radio channel has changed the transmitted signal before it can recover the data. To do this, current systems send known reference signals, often called pilots. These pilots help the receiver estimate the channel, but they also consume time-frequency resources that could otherwise be used to transmit user data.

This thesis studies how artificial intelligence can help reduce this pilot overhead in multiple-antenna wireless systems. Instead of treating channel estimation and data detection as two separate steps, the proposed receiver learns to perform them jointly. The pilot signals are used as contextual information, allowing the neural network to infer the channel condition and recover the transmitted data within a unified model.

The work investigates several pilot designs. In current 5G systems, pilot signals from different antennas are kept separated to avoid interference. This is reliable, but it requires more pilot resources. In the proposed in-context learning based scheme, different antennas are allowed to transmit pilots over the same resource elements. Although this creates interference, the AI receiver can learn to interpret the overlapping pilots as useful context. The results show that this can reduce the need for strictly orthogonal pilot patterns while maintaining good detection performance.

The thesis also studies superimposed pilot transmission, where pilot and data symbols are placed on the same resource elements. This removes dedicated pilot overhead, but it also makes the detection problem harder because pilots and data interfere with each other. The results show that the proposed receiver can mitigate part of this interference. In favorable signal-to-noise and coding conditions, the saved pilot resources can lead to higher throughput.

Overall, this thesis shows that combining physical communication knowledge with deep learning can make wireless receivers more flexible. The proposed receiver does not simply replace traditional signal processing with a black-box model; instead, it uses the structure of the wireless system to guide the learning process. This provides a possible direction for future wireless systems that need to use spectrum more efficiently while supporting increasingly complex transmission schemes.

Acknowledgements

I would like to express my sincere gratitude to Huawei Nyquist Research Center for providing me with the opportunity and resources to conduct this Master's thesis. This work would not have been possible without the support, technical guidance, and research environment provided during my thesis project.

I am especially grateful to my industry advisors, Sha Hu and Jonathan Lindberg, for their continuous guidance throughout the project. Their advice and discussions were valuable in shaping the research direction, improving the technical implementation, interpreting the experimental results, and refining the presentation of the thesis.

I would particularly like to thank Juan Vidal Alegría for his careful review of the manuscript, constructive suggestions, and generous help with both technical questions and university-related matters during the thesis process. I would also like to thank Michael Lentmaier for serving as the examiner of this thesis and for providing academic supervision and feedback from the university side.

Finally, I am deeply grateful to my family for their constant support, encouragement, and understanding throughout my Master's studies in Sweden.

Declaration of Generative AI Use

Generative AI and language tools, including ChatGPT, Gemini, and DeepL, were used during the writing and editing stage of this thesis as auxiliary support. Their use was limited to language polishing, improving clarity and readability, rephrasing technical descriptions, and assisting with LaTeX formatting.

The technical content, research design, simulation setup, implementation, numerical results, figures, analysis, and final conclusions were developed and verified by the author. Generative AI was not used as an independent source of scientific evidence, nor was it used to generate experimental results or replace the author's own judgement. All AI-assisted text was reviewed, edited, and approved by the author, who takes full responsibility for the final content of the thesis.

Table of Contents

List of Abbreviations

3GPP	3rd Generation Partnership Project
5G-NR	5th Generation New Radio
6G	Sixth Generation
AI	Artificial Intelligence
AWGN	Additive White Gaussian Noise
BCE	Binary Cross-Entropy
BER	Bit Error Rate
BLER	Block Error Rate
CE	Channel Estimation
CEM	Channel Estimation Module
CSI	Channel State Information
DGHC	Dynamic Generalized Hyper-Connection
DL	Deep Learning
DMRS	Demodulation Reference Signal
E2E	End-to-End
EPA	Extended Pedestrian A channel model
ETU	Extended Typical Urban channel model
FEC	Forward Error Correction
FFN	Feed-Forward Network
FLOP	Floating-Point Operation
GPU	Graphics Processing Unit
IAI	Inter-Antenna Interference
IC	Interference Cancellation
ICL	In-Context Learning
i.i.d.	Independent and Identically Distributed
I/Q	In-phase and Quadrature
KL	Kullback-Leibler
LDPC	Low-Density Parity-Check
LLM	Large Language Model
LLR	Log-Likelihood Ratio
LMMSE	Linear Minimum Mean Square Error
Log-MAP	Logarithmic Maximum A Posteriori
LR	Learning Rate

LS	Least Squares
LSB	Least Significant Bit
MAP	Maximum A Posteriori
MDM	MIMO Detection Module
MF	Matched Filter
MF-IC	Matched-Filter Interference Cancellation
MHA	Multi-Head Attention
MIMO	Multiple-Input Multiple-Output
ML	Maximum Likelihood
MLD	Maximum Likelihood Detection
MLP	Multi-Layer Perceptron
MMSE	Minimum Mean Square Error
MoE	Mixture of Experts
MSE	Mean Squared Error
NN	Neural Network
NLP	Natural Language Processing
OCC	Orthogonal Cover Code
OFDM	Orthogonal Frequency-Division Multiplexing
PAM	Pulse-Amplitude Modulation
PAPR	Peak-to-Average Power Ratio
PGFC	Physics-Guided Feature Construction
PHY	Physical Layer
QAM	Quadrature Amplitude Modulation
QPSK	Quadrature Phase-Shift Keying
QR	QR Decomposition
RE	Resource Element
RoPE	Rotary Position Embedding
SD	Sphere Decoding
SE	Spectral Efficiency
SER	Symbol Error Rate
SI	Superimposed
SINR	Signal-to-Interference-plus-Noise Ratio
SNR	Signal-to-Noise Ratio
SOTA	State of the Art
Stg	Stage
SVD	Singular Value Decomposition
SwiGLU	Swish-Gated Linear Unit
TPA	Tensor Product Attention
TTI	Transmission Time Interval
VWN	Virtual Width Network

Chapter 1

Introduction

1.1 Background

Driven by the evolution towards sixth generation (6G) networks, the Artificial Intelligence (AI)-driven paradigm shift within physical layer (PHY) design has become a broad consensus in both academia and industry [1]. As standardization bodies such as the 3rd Generation Partnership Project (3GPP) continue to promote the integration of AI into air interface technologies [2], future wireless networks are steadily transitioning towards an AI-native architecture. Against this backdrop, end-to-end (E2E) AI receivers for massive multiple-input multiple-output (MIMO) and orthogonal frequency-division multiplexing (OFDM) systems are gradually emerging. Unlike traditional signal processing chains that rely on a cascade of disjoint and independently optimized modules, deep neural network-based architectures aim to unify core functionalities such as channel estimation (CE) and symbol detection within a jointly optimized latent space [3–5].

While advanced classical algorithms such as linear minimum mean square error (LMMSE) estimation, sphere decoding (SD), and K-Best search [6, 7] demonstrate robust performance that closely approaches theoretical limits in conventional scenarios, exploring deep learning (DL) models for the PHY remains of significant engineering importance. First, the inference process of neural networks (NNs) heavily relies on matrix and tensor operations, which naturally aligns with the development trajectory of contemporary parallel computing hardware, offering a new path to break through computational latency bottlenecks. Second, a unified neural architecture can break down the information barriers between traditional CE and detection, achieving significant joint processing gains through implicit feature passing, thereby demonstrating the potential to surpass conventional baselines in extreme scenarios of nonlinear hardware impairments or complex interference.

In recent years, major breakthroughs in natural language processing (NLP) have inspired the design of communication systems. With the rapid iteration of large foundation models centered on Transformers (e.g., Qwen, DeepSeek, Kimi [8–10]), their demonstrated capability of In-Context Learning (ICL) has attracted widespread attention. ICL essentially provides an unprecedented few-shot adaptation paradigm. Mapping this to wireless communications, the few pilot signals acquired at the receiver can be regarded as a prompt that activates the

network's prior knowledge. By leveraging this mechanism, pre-trained models can effectively extract the dynamic characteristics of time-varying channels and data recovery without any parameter fine-tuning, opening up a new dimension for the design of next-generation adaptive receivers.

1.2 Motivation

DL and ICL provide new approaches for wireless receiver design. However, their practical deployment faces three main challenges. First, existing models are rarely benchmarked against theoretical optimal limits. Second, the trade-off between pilot overhead and spectral efficiency (SE) requires receivers to operate reliably with minimal pilots. Third, current architectures lack universality and require structural redesigns for different pilot configurations.

First, a rigorous assessment of neural receivers must use the Maximum Likelihood Detector (MLD) bound as a reference. The MLD provides the fundamental performance limit for MIMO symbol detection, yet the majority of existing studies evaluate deep receivers and ICL equalizers only against sub-optimal linear baselines such as LMMSE [11–13]. Relying solely on practical references obscures the remaining gap to optimal detection. For instance, while recent ICL methods can outperform LMMSE even when the latter has perfect CSI, earlier model-driven receivers like DetNet and OAMPNet2 often fail to sustain their performance in fully coded MIMO systems [14]. To address this, the recently proposed AI receiver in [3] sets a new state-of-the-art (SOTA) benchmark, demonstrating the best available performance by approaching the MLD bound. Therefore, before tackling severe channel impairments, it is essential to first establish an expressive backbone capable of matching this near-MLD capability in idealized settings.

Second, beyond establishing a robust baseline, modern communication systems face a trade-off between demodulation reference signal (DMRS) overhead and SE. While reducing pilot density or adopting superimposed DMRS (SI DMRS) directly increases SE, it degrades channel state information (CSI) accuracy, which compromises symbol detection [15–17]. Although dedicated attention-based architectures have been proposed to enhance CE under limited pilots [18], how to optimally balance DMRS overhead and symbol detection performance remains inadequately addressed. This highlights the critical need for highly expressive AI receivers capable of reliable data recovery with minimal pilot context.

Third, existing solutions lack architectural universality. Although recent models have adapted baseline architectures designed for orthogonal grids (e.g., DeepRx) to support SI DMRS [15, 16] and developed ICL-based equalizers [5, 19], these models remain highly specialized. Currently, transitioning from standard orthogonal schemes to non-orthogonal or SI DMRS configurations alters the interference structure. Such shifts require redesigning the network to handle the resulting pilot-pilot or pilot-data interference. Therefore, a unified framework that can handle any pilot configuration without structural changes is still missing.

Motivated by these challenges, there is a compelling need to design a novel, versatile AI receiver architecture. Drawing structural inspiration from SOTA large language models (LLMs), this architecture leverages advanced Transformer

components to provide a unified solution for diverse transmission schemes. By establishing a robust non-linear backbone that achieves near-MLD performance, our ICL receiver can effectively break the rigid trade-off between pilot overhead and system reliability, expanding the spectral efficiency boundaries of next-generation MIMO systems.

1.3 Main Contributions

In response to the challenges of architectural fragmentation, pilot overhead trade-offs, and sub-optimal benchmarking in current PHY designs, this thesis proposes a **Unified In-Context Learning (ICL) Receiver architecture**. Drawing structural inspiration from SOTA LLMs, the proposed receiver leverages the ICL paradigm to perform joint CE and MIMO detection within a cascaded network. The main contributions of this work are summarized as follows:

- **A Unified ICL Receiver Framework:** We introduce a ICL receiver that operates independently of specific pilot configurations. Unlike existing solutions requiring dedicated interference cancellation modules or dynamic input dimensions, the proposed architecture processes both orthogonal allocations and non-orthogonal or superimposed pilot schemes within Transformer backbones. By treating pilot signals as contextual prompts, the framework adapts to varying pilot patterns and densities without architectural modifications.
- **Establishing a Near-MLD Performance Baseline:** To validate the representational capacity of the proposed architecture, we benchmark its performance against theoretical limits. We demonstrate that under Genie CSI, the ICL receiver achieves an accuracy that approaches the MLD. Furthermore, in joint CE and detection scenarios, the proposed model outperforms the Genie-CSI-aided LMMSE, demonstrating robust detection capabilities across diverse evaluated configurations.
- **Improving the Trade-off Between Pilot Overhead and SE:** We demonstrate that the architecture sustains reliable joint CE and MIMO detection under varying pilot allocations and non-orthogonalization. By operating with reduced DMRS overhead and exploiting the zero-overhead property of SI DMRS, more Resource Elements (REs) are preserved for data transmission, thereby directly improving the throughput and overall SE.

1.4 Thesis Organization

The remainder of this thesis is organized as follows.

Chapter 2 introduces the mathematical model of MIMO-OFDM system. It details mathematical representations of the three distinct DMRS transmission schemes evaluated in this work: standard 5G-NR orthogonal pilots, non-orthogonal ICL DMRS, and SI DMRS.

Chapter 3 details the proposed Physics-Guided ICL receiver architecture. It describes how the network jointly performs CE and MIMO detection. Specifically, the network is organized as an unfolded iterative cascade that tightly couples explicit physical signal models with a data-driven Transformer backbone. Key architectural innovations discussed include Universal Tokenization, an advanced Transformer backbone featuring latent fusion, Physics-Guided Feature Construction (PGFC), and the iterative unfolded cascade mechanism.

Chapter 4 establishes the training framework and outlines the methodological evolution of detection enhancement strategies. It formulates the composite loss functions and training strategies, including physics-informed data augmentations. The chapter also documents the exploration of various auxiliary detection enhancements, ultimately highlighting Resampling as the most robust strategy for closing the performance gap to the theoretical Maximum Likelihood (ML) bound.

Chapter 5 presents the experimental results. After defining the simulation setup and theoretical baseline algorithms, it provides a detailed gain analysis of the iterative mechanism. The chapter further investigates the impact of pilot sparsity on traditional 5G-NR versus ICL paradigms and rigorously characterizes the ICL receiver's performance, robustness, and throughput under SI DMRS.

Chapter 6 summarizes the main conclusions derived from this thesis, discusses current limitations of the proposed architecture, and outlines potential directions for future research.

Chapter 2

System Model and Transmission Schemes

2.1 System Model

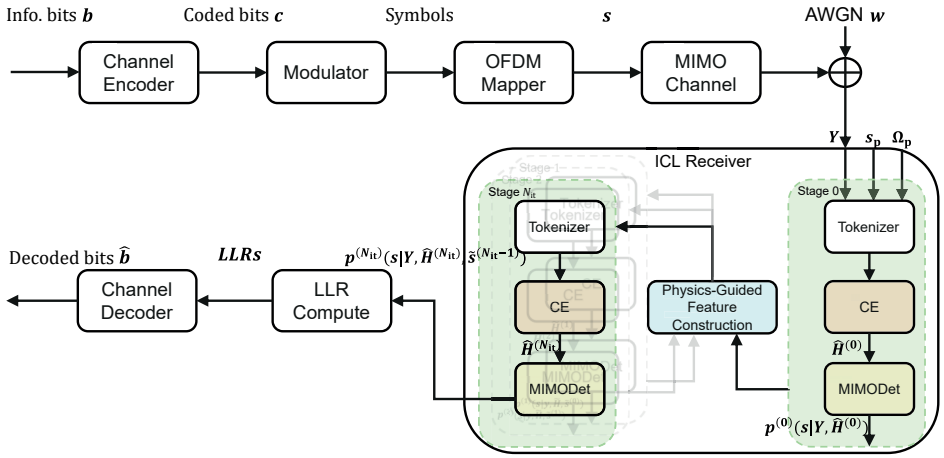


Figure 2.1: Block diagram of the MIMO-OFDM system and the ICL receiver structure.

We consider a MIMO communication system employing OFDM. The system consists of a transmitter equipped with N_{tx} antennas and a receiver equipped with N_{rx} antennas, where typically $N_{\mathrm{rx}} \geq N_{\mathrm{tx}}$. The overall architecture of the proposed $N_{\mathrm{rx}} \times N_{\mathrm{tx}}$ MIMO-OFDM system is illustrated in Fig. 2.1.

2.1.1 Transmitter Pipeline and RE Mapping

Following a standard coded spatial multiplexing pipeline, the information source at the transmitter generates a binary information bit sequence $\mathbf{b} \in \{0,1\}^K$, where K is the number of information bits. These bits are first processed by a channel

encoder, producing a coded bit sequence $\mathbf{c} = f_{\text{enc}}(\mathbf{b}) \in \{0,1\}^N$, where N denotes the total number of coded bits.

The coded bits are then mapped to complex-valued constellation symbols according to an M-ary quadrature amplitude modulation (M-QAM) scheme. Specifically, the modulation mapping $f_{\text{map}} : \{0,1\}^{\log_2 M} \to \mathcal{A}$ adheres to the Gray-coded M-QAM rules defined in the 3GPP 5G-NR standard TS 38.211 [20]. The mapped symbols are normalized by scaling factors (e.g., $1/\sqrt{42}$ for 64-QAM) to maintain a unit average power constraint.

The transmitted signals are structured into a 2D time-frequency resource grid. Each transmission time interval (TTI) consists of N_{sym} OFDM symbols in the time domain and N_{sc} subcarriers in the frequency domain. A specific RE in this grid is indexed by the coordinate (n,k), where $n \in \{0,\ldots,N_{\text{sym}}-1\}$ and $k \in \{0,\ldots,N_{\text{sc}}-1\}$.

The REs in the time-frequency grid are partitioned into three functional sets defined by spatial-temporal boolean masks: the pilot set Ω_{p}, the control set Ω_{c}, and the data set Ω_{d}. The specific allocation of these REs varies depending on the chosen DMRS scheme, which will be detailed in Section 2.2.

2.1.2 Received Signal Formulation

Let $\tilde{\mathbf{s}}_{n,k} \in \mathbb{C}^{N_{\text{tx}}\times 1}$ denote the unscaled transmit symbol vector mapped to the n-th OFDM symbol and the k-th subcarrier across all N_{tx} antennas. As established in the modulation stage, all non-zero entries in $\tilde{\mathbf{s}}_{n,k}$ possess unit average power. To maintain a constant total transmit power constraint across the array regardless of the pilot pattern or spatial multiplexing vacancy, a spatial power scaling is applied. The final normalized transmitted symbol vector $\mathbf{s}_{n,k} \in \mathbb{C}^{N_{\text{tx}}\times 1}$ is given by:

$$\mathbf{s}_{n,k} = \frac{1}{\sqrt{N_{\text{act}}(n,k)}}\tilde{\mathbf{s}}_{n,k}, \tag{2.1}$$

where $N_{\text{act}}(n,k) \in [1, N_{\text{tx}}]$ represents the number of active transmit antennas at RE (n,k). This scaling ensures that the expected total transmit power per RE remains normalized, i.e., $\mathbb{E}[\|\mathbf{s}_{n,k}\|^2] = 1$.

For a specific RE (n,k), the received signal vector $\mathbf{y}_{n,k} \in \mathbb{C}^{N_{\text{rx}}\times 1}$ is modeled as:

$$\mathbf{y}_{n,k} = \mathbf{H}_{n,k}\mathbf{s}_{n,k} + \mathbf{w}_{n,k}, \tag{2.2}$$

where $\mathbf{H}_{n,k} \in \mathbb{C}^{N_{\text{rx}}\times N_{\text{tx}}}$ is the frequency-domain complex channel response matrix at symbol n and subcarrier k. It is assumed that the channel is normalized such that the expected power of each channel coefficient is unity, i.e., $\mathbb{E}[|[\mathbf{H}_{n,k}]_{i,j}|^2] = 1$. The term $\mathbf{w}_{n,k} \in \mathbb{C}^{N_{\text{rx}}\times 1}$ represents the additive white Gaussian noise (AWGN) vector at the receiver array. The entries of $\mathbf{w}_{n,k}$ are independent and identically distributed (i.i.d.) circularly symmetric complex Gaussian random variables, $\mathbf{w}_{n,k} \sim \mathcal{CN}(\mathbf{0}, \sigma_w^2\mathbf{I}_{N_{\text{rx}}})$, where σ_w^2 is the noise variance per receive antenna.

Given the unit-power normalizations of both the transmit symbols and the channel matrix, the received signal-to-noise ratio (SNR) of the system is defined as $\text{SNR} = 1/\sigma_w^2$. Stacking the spatial dimensions across the entire time-frequency grid, the overall received signal tensor $\mathbf{Y} \in \mathbb{C}^{N_{\text{rx}}\times N_{\text{sym}}\times N_{\text{sc}}}$ and the corresponding

channel tensor $\mathbf{H} \in \mathbb{C}^{N_{\mathrm{rx}} \times N_{\mathrm{tx}} \times N_{\mathrm{sym}} \times N_{\mathrm{sc}}}$ constitute the complete observation space for the subsequent neural network-based detection tasks.

2.1.3 ICL Receiver Design

As illustrated in Fig. 2.1, the ICL receiver replaces the conventional disjoint CE and MIMO detection modules with a unified architecture.

To perform joint CE and MIMO detection, the ICL receiver takes three physical inputs processed by the tokenizer: the received frequency-domain signal tensor $\mathbf{Y}$, the predefined pilot symbols $\mathbf{s}_{\mathrm{p}}$ (i.e., the transmit symbols $\mathbf{s}_{n,k}$ for $(n,k) \in \Omega_{\mathrm{p}}$), and the pilot coordinate mask Ω_{p}. By treating the pilot information as contextual prompts, the architecture establishes a direct mapping from the observation space to the symbol space.

Following multi-stage iterative processing, the output of the ICL receiver at the final iteration N_{it} is a set of soft symbol detections. Mathematically, this is formulated as the conditional a posteriori probabilities $p^{(N_{\mathrm{it}})}(\mathbf{s} \mid \mathbf{Y}, \tilde{\mathbf{H}}^{(N_{\mathrm{it}})}, \tilde{\mathbf{s}}^{(N_{\mathrm{it}}-1)})$, which captures the dependence on the intermediate channel estimates and the soft symbol feedback from the preceding layer.

From an implementation perspective, the ICL receiver outputs these probability distributions as unnormalized logits. These logits are then passed to the log-likelihood ratio (LLR) computation module to produce bit-level LLRs. Finally, the LLR sequence is fed into the channel decoder for forward error correction (FEC), yielding the estimated information bit sequence $\hat{\mathbf{b}}$ and completing the receiver pipeline.

The detailed architectural design of the ICL receiver, including the input tensor transformations (tokenization) and the iterative mechanisms, is presented in Chapter 3.

2.2 DMRS Transmission Schemes

In this section, we formulate the mathematical representations for the three distinct DMRS transmission schemes evaluated in this work. To formalize the element-wise signal construction across all configurations, let $d_{n,k}^{(t)} \in \mathcal{A}$ denote the complex-valued data symbol mapped to the t-th transmit antenna. Correspondingly, let $p_{n,k}^{(t)}$ denote the scalar pilot sequence symbol. Throughout our evaluations, $p_{n,k}^{(t)}$ is drawn from the standard pseudo-random QPSK alphabet $\{\pm\frac{1}{\sqrt{2}} \pm j\frac{1}{\sqrt{2}}\}$ to align with conventional peak-to-average power ratio (PAPR) constraints.

As shown in Fig. 2.2, the distinction between the standard 5G-NR and the ICL DMRS schemes mainly lies in their spatial resource allocation. To accommodate varying constraints of pilot overhead, we define a spectrum of pilot sparsity configurations ranging from I1 (densest) to I12 (sparsest). For each sparsity level, the schemes are matched in terms of pilot overhead, while the exact pilot locations follow the corresponding allocation patterns. Standard 5G-NR requires pilot REs to be orthogonal across transmit antennas. Consequently, for a given sparsity configuration, the number of pilot REs allocated per antenna is diluted as the

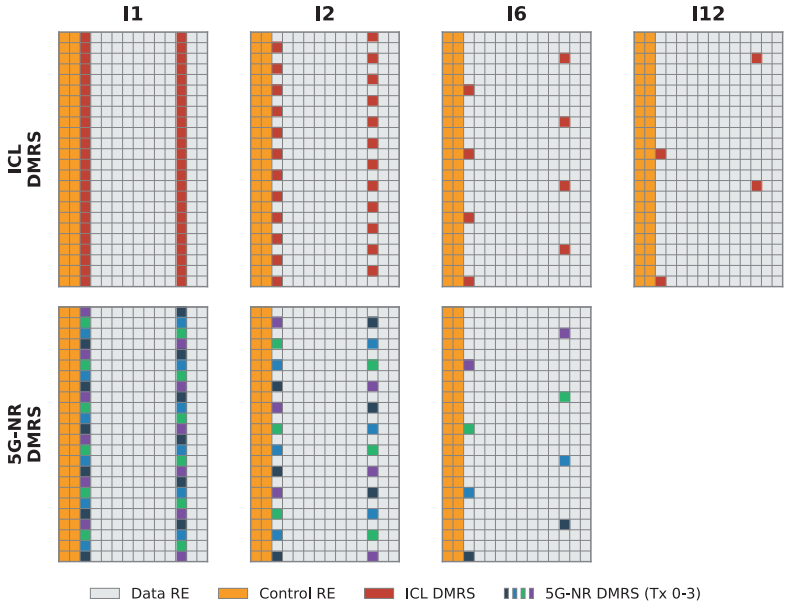


Figure 2.2: Time-frequency RE grids illustrating the pilot sparsity configurations (l1, l2, l6, and l12) for a 4×4 MIMO setup.

MIMO array scales up. In contrast, the ICL DMRS scheme utilizes non-orthogonal pilots, allowing all transmit antennas to simultaneously transmit over the same active pilot REs.

2.2.1 Orthogonal 5G-NR DMRS

In conventional 5G-NR transmission, pilot and data symbols are separated in the time-frequency grid. The global sets of REs allocated for pilots (Ω_{p}) and data (Ω_{d}) are mutually exclusive, satisfying $\Omega_{\mathrm{p}} \cap \Omega_{\mathrm{d}} = \emptyset$. To ensure that classical CE algorithms can acquire accurate CSI without spatial interference, the reference signals transmitted from different antennas must be orthogonal.

While the 5G-NR standard supports achieving this spatial orthogonality through Orthogonal Cover Codes (OCCs), the baseline evaluated in this project adopts the disjoint mapping configuration. Let $\Omega_{\mathrm{p}}^{(t)}$ denote the specific subset of pilot coordinates assigned to the t-th transmit antenna, where the pilot mask is defined as $\Omega_{\mathrm{p}} = \bigcup_{t=1}^{N_{\mathrm{tx}}} \Omega_{\mathrm{p}}^{(t)}$. Spatial isolation is guaranteed by ensuring these subsets are non-overlapping:

$$\Omega_{\mathrm{p}}^{(t)} \cap \Omega_{\mathrm{p}}^{(t')} = \emptyset, \quad \forall t \neq t'. \tag{2.3}$$

Under this disjoint paradigm, when the t-th antenna transmits the reference symbol $p_{n,k}^{(t)}$ at its allocated RE $(n, k) \in \Omega_{\mathrm{p}}^{(t)}$, all other transmit antennas must remain silent at that specific coordinate. Consequently, the unscaled transmit

signal $\tilde{s}_{n,k}^{(t)}$ for the t-th antenna is formulated as:

$$\tilde{s}_{n,k}^{(t)} = \begin{cases} p_{n,k}^{(t)}, & \text{if } (n,k) \in \Omega_{\mathrm{p}}^{(t)} \\ 0, & \text{if } (n,k) \in \Omega_{\mathrm{p}}^{(t')} \text{ for } t' \neq t \\ d_{n,k}^{(t)}, & \text{if } (n,k) \in \Omega_{\mathrm{d}} \end{cases} . \tag{2.4}$$

2.2.2 Non-Orthogonal ICL DMRS

Unlike the 5G-NR baseline that enforces strict orthogonality, the non-orthogonal scheme allows all N_{tx} transmit antennas to simultaneously transmit independent reference symbols over the same time-frequency REs without applying orthogonal sequences. This lack of orthogonality introduces spatial pilot-to-pilot interference. In this thesis, this configuration is referred to as the ICL DMRS scheme to reflect its processing mechanism: rather than relying on an explicit interference cancellation procedure, the proposed ICL receiver is trained in a data-driven manner to interpret the overlapping signals as contextual prompts for joint CE and symbol detection.

The time-frequency orthogonality between pilots and data ($\Omega_{\mathrm{p}} \cap \Omega_{\mathrm{d}} = \emptyset$) is still maintained, identical to standard 5G-NR. However, because spatial isolation is abandoned, all transmit antennas share the exact same pilot mask Ω_{p}. Consequently, the unscaled transmit signal $\tilde{s}_{n,k}^{(t)}$ for the t-th antenna is formulated as:

$$\tilde{s}_{n,k}^{(t)} = \begin{cases} p_{n,k}^{(t)}, & \text{if } (n,k) \in \Omega_{\mathrm{p}} \\ d_{n,k}^{(t)}, & \text{if } (n,k) \in \Omega_{\mathrm{d}} \end{cases} . \tag{2.5}$$

2.2.3 SI DMRS

SI DMRS overlaps the pilots directly onto the data [15–17]. In this scheme, data symbols still span the available grid, and the pilot mask is configured as a subset of it, such that $\Omega_{\mathrm{p}} \subseteq \Omega_{\mathrm{d}}$.

To maintain spatial orthogonality among the superimposed pilot components across multiple antennas, this scheme restricts the pilot coordinates to sparse, disjoint layouts ($\Omega_{\mathrm{p}}^{(t)} \cap \Omega_{\mathrm{p}}^{(t')} = \emptyset$), structurally identical to the standard 5G-NR baseline. Data symbols are then superimposed over these specific pilot REs.

For the resource elements within the t-th antenna's pilot mask $(n,k) \in \Omega_{\mathrm{p}}^{(t)}$, the final unscaled transmit signal is constructed as a linear superposition of the pilot symbol $p_{n,k}^{(t)}$ and the data symbol $d_{n,k}^{(t)}$. Controlled by a power allocation factor $\rho \in [0,1]$, it is formulated as:

$$\tilde{s}_{n,k}^{(t)} = \begin{cases} \sqrt{\rho}\, p_{n,k}^{(t)} + \sqrt{1-\rho}\, d_{n,k}^{(t)}, & \text{if } (n,k) \in \Omega_{\mathrm{p}}^{(t)} \\ d_{n,k}^{(t)}, & \text{if } (n,k) \in \Omega_{\mathrm{d}} \setminus \Omega_{\mathrm{p}}^{(t)} \end{cases} . \tag{2.6}$$

While this superposition guarantees zero additional time-frequency overhead, it introduces pilot-data interference. At the receiver side, the term $\sqrt{1-\rho}\, d_{n,k}^{(t)}$ acts as self-interference during CE at the pilot-carrying REs, whereas $\sqrt{\rho}\, p_{n,k}^{(t)}$

is superimposed on the data symbol and affects symbol detection. Resolving this coupled interference presents a core challenge for the proposed ICL receiver, addressing a critical scenario where classical disjoint estimators experience severe performance degradation [15].

Chapter 3

A Proposed Unified ICL Receiver Architecture

This chapter details the proposed Physics-Guided ICL receiver. By embedding the physical structure of MIMO-OFDM systems into a Transformer-based backbone, the architecture jointly performs CE and MIMO detection. The network is organized as an unfolded iterative cascade that tightly couples explicit physical signal models with a data-driven Transformer backbone.

The overarching design is driven by four key architectural contributions, which define the structural flow of this chapter:

- **Universal Physical Tokenization:** To achieve compatibility across diverse DMRS paradigms, a generalized tokenization block is introduced. By utilizing an invariant binary pilot occupancy mask, the tokenizer unifies distinct transmission schemes, the real-valued received observation tensor, and the pilot symbol tensor into a multi-dimensional latent format.
- **Transformer Backbone with Latent Fusion:** Inspired by recent LLMs [8–10], the core learning engine is customized for PHY signals. It integrates a Channel Estimation Module (CEM) and a MIMO Detection Module (MDM), whose encoder blocks incorporate Virtual Width Networks (VWN) [21], Tensor Product Attention (TPA) [22], and Mixture-of-Experts (MoE) [9, 23]. A latent residual fusion bridge combines the high-dimensional channel feature into the original prompt, eliminating information bottlenecks between the CEM and MDM stages.
- **Physics-Guided Feature Construction (PGFC):** To accelerate convergence and complement the data-driven backbone with structured physical knowledge, a differentiable PGFC bridge is introduced between cascaded stages. Rather than propagating raw logits, PGFC explicitly synthesizes communication-theoretic tensors, including prior-constrained soft symbol estimates, Interference Cancellation (IC) residuals, and Matched Filter (MF) responses. This provides the subsequent stages with physically localized error gradients that accelerate training and improve generalization.
- **Iterative Unfolded Cascade:** The architecture operates as a deep cascaded loop. Starting from an initial bootstrap stage, the bundled feature streams,

comprising the augmented observation features, the intermediate soft symbol estimates, and the pilot mask, are forwarded into successive refinement blocks. This recurrent-like unfolding progressively resolves complex multi-antenna interference, yielding performance gains over conventional single-shot neural detectors.

The remainder of this chapter is organized as follows. Section 3.2 presents the universal tokenization strategy. Section 3.3 details the macro-architecture of the joint CEM and MDM pipeline. Section 3.4 formulates the PGFC mechanism. Section 3.3.2 breaks down the micro-architectural designs of the VWN, TPA, and MoE sub-layers.

3.1 Architecture Overview

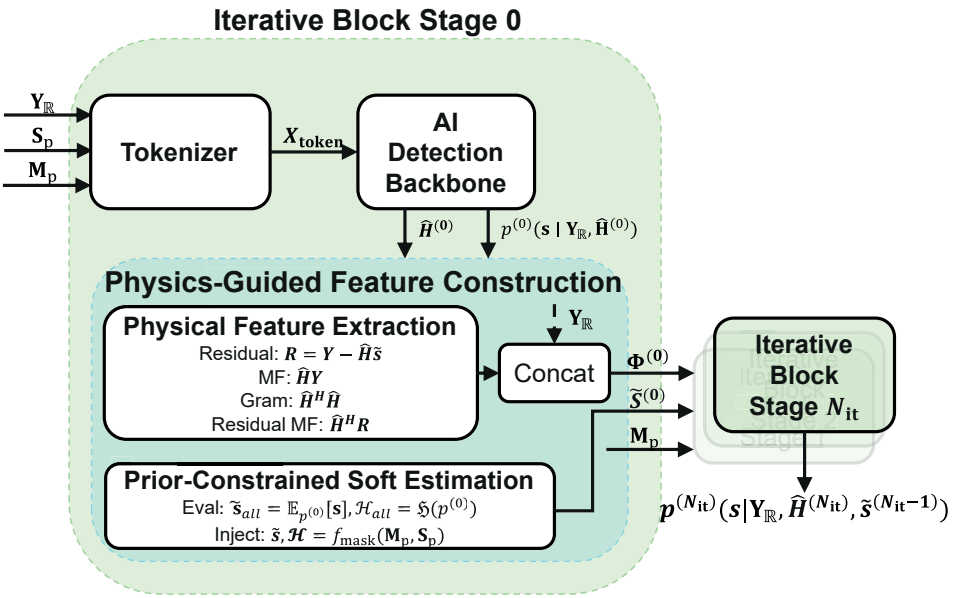


Figure 3.1: Overall architecture of the proposed ICL receiver.

The overall architecture is illustrated in Fig. 3.1. The receiver is structured as an unfolded cascaded loop, which progressively cancels interference and jointly optimizes channel inference and symbol recovery.

The data flow within the initial bootstrap block (Stage 0) proceeds through three primary processing modules:

Tokenizer: This module serves as the tensor pre-processor for each computational stage. During the initial bootstrap block (Stage 0), it receives the real-valued observation tensor $\mathbf{Y}_{\mathbb{R}}$, the pilot symbol tensor $\mathbf{S}_{\mathrm{p}}$, and the binary pilot mask $\mathbf{M}_{\mathrm{p}}$ (which are mapped from the physical signal components $\mathbf{Y}$, s_{p}, and Ω_{p} defined in Section 2.1, as detailed in Section 3.2). In subsequent iterative stages ($i > 0$), the

initial observation and pilot tensors are replaced by the physics-guided feature streams generated from the previous stage. Regardless of the varying input feature dimensions, the tokenizer executes a spatial flattening operation, concatenates the tensors along their feature axes, and applies linear projection. This step maps the physical tensors into a unified d_{virtual}-dimensional token sequence.

AI Detection Backbone: Operating on the tokenized sequence, this Transformer-based backbone jointly optimizes a CEM and a MDM. For the initial bootstrap block (Stage 0), the forward propagation yields the intermediate frequency-domain CE $\hat{\mathbf{H}}^{(0)}$ alongside the preliminary symbol-wise posterior probabilities $p^{(0)}(\mathbf{s} \mid \mathbf{Y}_{\mathbb{R}}, \hat{\mathbf{H}}^{(0)})$.

Physics-Guided Feature Construction (PGFC): This module transforms the intermediate neural outputs into structured physical feature streams. The network's real-valued outputs are first mapped back to the complex domain $\mathbb{C}$. As illustrated in the PGFC architecture, the operation is partitioned into two steps:

- *Prior-Constrained Soft Estimation*: It first computes the complex soft symbol expectations and their corresponding uncertainty measure, represented by the Shannon entropy, from the backbone's posterior probabilities. The known pilot values are then injected according to the pilot mask $\mathbf{M}_{\text{p}}$. This forms the composite soft transmission state $\mathbf{S}_{\text{comp}}^{(0)}$, which is mapped to the real domain and concatenated with the entropy tensor to produce the updated soft state tensor $\tilde{\mathbf{S}}^{(0)}$ for the next stage.

- *Physical Feature Extraction*: Using the intermediate channel estimate $\hat{\mathbf{H}}^{(0)}$, the composite soft symbol estimate, and the received signal $\mathbf{Y}$, the module computes several communication-theoretic feature tensors, including the soft signal residual $\mathbf{R}^{(0)}$, the matched-filter response $\mathbf{Z}_{\text{MF}}^{(0)}$, the Gram matrix $\mathbf{G}^{(0)}$, and the residual matched-filter response $\mathbf{R}_{\text{MF}}^{(0)}$. These feature tensors are decoupled into their real and imaginary components and concatenated with the raw observation $\mathbf{Y}_{\mathbb{R}}$ along the feature axis to form the augmented observation tensor $\mathbf{\Phi}^{(0)}$.

Upon completion of the PGFC operations, the bundled feature streams, comprising the augmented observation tensor $\mathbf{\Phi}^{(0)}$, the updated soft state $\tilde{\mathbf{S}}^{(0)}$, and the invariant pilot mask $\mathbf{M}_{\text{p}}$, are forwarded into the subsequent iterative block (Stage 1). This cascaded refinement mechanism ensures that the i-th iteration strictly outputs the progressively optimized posterior $p^{(i)}(\mathbf{s} \mid \mathbf{Y}_{\mathbb{R}}, \hat{\mathbf{H}}^{(i)}, \tilde{\mathbf{S}}^{(i-1)})$, thereby achieving robust detection under complex interference.

This unfolded progression is repeated across N_{it} stages. Ultimately, the architecture yields the highly refined a posteriori probabilities of the transmitted symbols at the final stage, ready for downstream channel decoding.

3.2 Input Representation and Tokenization

Input Tensor Definitions

Let B, N_{rx}, N_{tx}, N_{sym}, and N_{sc} denote the batch size, number of receive antennas, number of transmit antennas, number of OFDM symbols, and number of subcarriers, respectively. To interface the physical signal models with the AI network, the complex-valued tensors defined in Section 2.1 are extended with a batch dimension B and mapped to the real domain. Specifically, all complex-valued quantities are decoupled into their real and imaginary parts and stacked along a dedicated feature axis of size 2.

The receiver operates on three core input tensors, whose spatial configurations are illustrated in Fig. 3.2 for representative DMRS schemes:

1. **Received observation tensor:** $\mathbf{Y}_{\mathbb{R}} \in \mathbb{R}^{B\times N_{\mathrm{rx}}\times N_{\mathrm{sym}}\times N_{\mathrm{sc}}\times 2}$, derived from the physical complex tensor $\mathbf{Y}$.
2. **Pilot symbol tensor:** $\mathbf{S}_{\mathrm{p}} \in \mathbb{R}^{B\times N_{\mathrm{tx}}\times N_{\mathrm{sym}}\times N_{\mathrm{sc}}\times 2}$, containing the decoupled real and imaginary components of the reference signals.
3. **Pilot occupancy mask:** $\mathbf{M}_{\mathrm{p}} \in \{0,1\}^{B\times N_{\mathrm{tx}}\times N_{\mathrm{sym}}\times N_{\mathrm{sc}}}$, a binary indicator tensor mapped from the antenna-specific pilot coordinate sets $\Omega_{\mathrm{p}}^{(t)}$, where entries are set to 1 for active pilot locations and 0 otherwise.

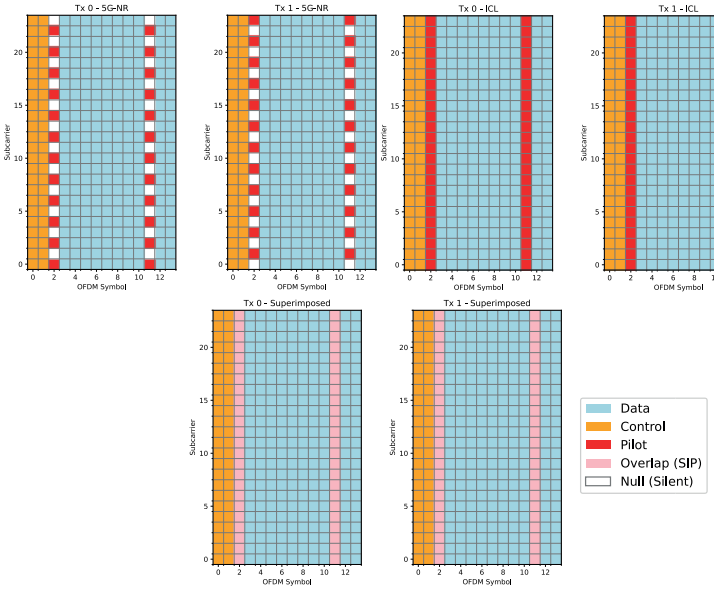


Figure 3.2: Visualization of the tensor input structures (pilot masks $\mathbf{M}_{\mathrm{p}}$ and symbols $\mathbf{S}_{\mathrm{p}}$) for a 2×2 MIMO system under the three evaluated transmission schemes: 5G-NR DMRS, ICL DMRS, and SI DMRS.

As established in Section 2.2, the pilot coordinate sets $\Omega_{\mathrm{p}}^{(t)}$ define the time-frequency pilot allocations for each transmit antenna. To represent this structural information within the AI backbone, they are mapped to the binary mask $\mathbf{M}_{\mathrm{p}}$. For the t-th transmit antenna at resource element (n, k), the binary entry is defined as $[\mathbf{M}_{\mathrm{p}}]_{b,t,n,k} = 1$ if $(n, k) \in \Omega_{\mathrm{p}}^{(t)}$, and 0 otherwise.

Consequently, the pilot symbol tensor $\mathbf{S}_{\mathrm{p}}$ is constrained by this mask, taking non-zero values only at indices where $[\mathbf{M}_{\mathrm{p}}]_{b,t,n,k} = 1$. This ensures that $\mathbf{S}_{\mathrm{p}}$ preserves the sparsity and spatial allocation pattern of the underlying DMRS scheme. The binary mask $\mathbf{M}_{\mathrm{p}}$ unifies different DMRS transmission schemes, including 5G-NR, ICL, and SI DMRS, into a standardized tensor format. This tokenization maps the signal models to the input space of the Transformer framework.

Tokenization and Latent Projection

To interface the multi-dimensional PHY tensors with the Transformer backbone, the tokenization process separates the time-frequency domain from the spatial-complex domain. For a given computational block i, the raw feature vector at a specific RE coordinate (n, k) is constructed by concatenating the flattened spatial antennas and their real-imaginary components. The mapping is formulated as a stage-dependent function:

$$\mathbf{x}_{\mathrm{raw}}^{(i)}[n,k] = \begin{cases} \mathrm{Concat}\left(\mathbf{Y}_{\mathbb{R}}[n,k],\ \mathbf{S}_{\mathrm{p}}[n,k],\ \mathbf{M}_{\mathrm{p}}[n,k]\right), & \text{if } i = 0, \\ \mathrm{Concat}\left(\mathbf{\Phi}^{(i-1)}[n,k],\ \tilde{\mathbf{S}}^{(i-1)}[n,k],\ \mathbf{M}_{\mathrm{p}}[n,k]\right), & \text{if } i \geq 1. \end{cases} \tag{3.1}$$

For the initial bootstrap stage ($i = 0$), the resulting raw feature dimension is $D_{\mathrm{in}}^{(0)} = 2N_{\mathrm{rx}}+3N_{\mathrm{tx}}$. In subsequent iterative stages ($i \geq 1$), the input dimension $D_{\mathrm{in}}^{(i)}$ expands to accommodate the physical statistics embedded within the augmented observation tensor $\mathbf{\Phi}^{(i-1)}$. To handle this stage-dependent input dimension, each block employs an independent learnable projection matrix $\mathbf{W}^{(i)} \in \mathbb{R}^{D_{\mathrm{in}}^{(i)} \times d_{\mathrm{virtual}}}$ to map the raw feature vector at each RE into the d_{virtual}-dimensional latent space.

By aggregating these projected vectors across all $L = N_{\mathrm{sym}} \times N_{\mathrm{sc}}$ grid locations, the tokenizer outputs a standardized token sequence:

$$\mathbf{X}_{\mathrm{token}}^{(i)} \in \mathbb{R}^{B \times L \times d_{\mathrm{virtual}}}, \tag{3.2}$$

which serves as the input to the AI Detection Backbone for stage i.

This dimension-mapping strategy aligns with the physical characteristics of MIMO-OFDM transmissions. By flattening the two-dimensional time-frequency grid into a 1D sequence of length L, the self-attention mechanism, augmented with Rotary Position Embedding (RoPE) [24], captures the local coherence of the fading channels. Furthermore, by collapsing the spatial antennas and their complex-valued components 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 MIMO interference cancellation internally within the high-dimensional latent representation.

3.3 AI Detection Backbone

The core backbone operating on the tokenized sequence is designed as an unfolded cascaded architecture that jointly performs CE and data recovery. To address the complexity of the MIMO-OFDM PHY without losing vital gradient information across deep layers, we define a joint detection pipeline unified by a latent feature fusion bridge.

3.3.1 Joint CE and MIMO Detection Pipeline

Rather than treating CE and MIMO detection as disjoint modular networks, the backbone unifies them into a sequentially fused pipeline. The structure of the backbone is illustrated in Fig. 3.3. The macro-architecture proceeds in two sequential phases:

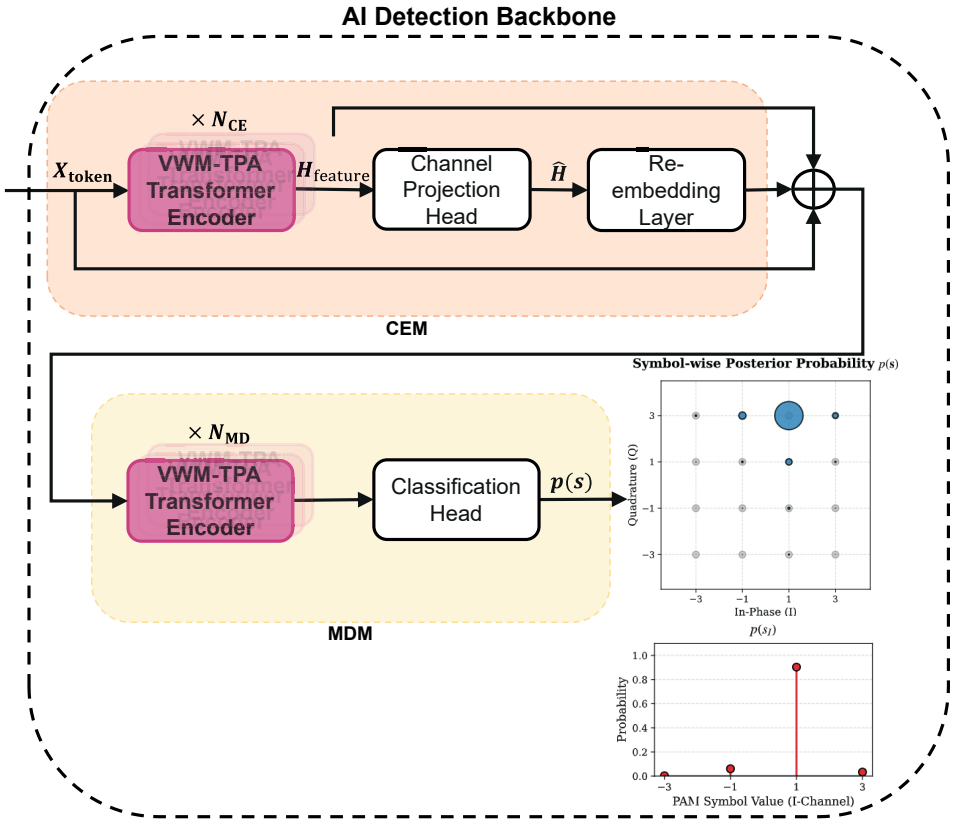


Figure 3.3: Architecture of the AI Detection Backbone.

1) Channel Estimation Module (CEM): Receiving the standardized prompt $\mathbf{X}_{\text{token}}^{(i)}$ directly from the tokenizer, the sequence is processed by N_{CE} stacked custom Transformer encoder blocks, incorporating VWN and TPA. As detailed in Sec-

tion 3.3.2, these specialized VWN-TPA blocks iteratively resolve the local time-frequency correlations and spatial interference, yielding an uncompressed stage-specific channel latent representation denoted as $\mathbf{H}_{\text{feature}}^{(i)} \in \mathbb{R}^{B\times L\times d_{\text{virtual}}}$.

To reconstruct the physical channel state, a dedicated Channel Projection Head (a linear fully-connected layer) maps $\mathbf{H}_{\text{feature}}^{(i)}$ back to the physical antenna dimensions:

$$\hat{\mathbf{H}}_{\text{flat}}^{(i)} = \mathbf{H}_{\text{feature}}^{(i)}\mathbf{W}_{\text{CE}}^{(i)} \quad \in \mathbb{R}^{B\times L\times(2N_{\text{rx}}N_{\text{tx}})}, \tag{3.3}$$

where $\mathbf{W}_{\text{CE}}^{(i)}$ is the stage-specific learnable weight matrix. By decoupling the real and imaginary components along a dedicated feature axis, $\hat{\mathbf{H}}_{\text{flat}}^{(i)}$ is subsequently reshaped to yield the explicit frequency-domain, real-valued channel estimate tensor $\hat{\mathbf{H}}^{(i)} \in \mathbb{R}^{B\times N_{\text{rx}}\times N_{\text{tx}}\times N_{\text{sym}}\times N_{\text{sc}}\times 2}$.

2) Latent Fusion and MIMO Detection Module (MDM): To ensure the signal equalization phase benefits from the original contextual prompt and the channel feature, a latent fusion mechanism is employed. Instead of solely relying on the explicit estimate $\hat{\mathbf{H}}_{\text{flat}}^{(i)}$, the input to the MDM encoders is formulated via an additive residual bridge:

$$\mathbf{Z}_{\text{eq}}^{(i)} = \mathbf{X}_{\text{token}}^{(i)} + \mathbf{H}_{\text{feature}}^{(i)} + \hat{\mathbf{H}}_{\text{flat}}^{(i)}\mathbf{W}_{\text{re-embed}}^{(i)}, \tag{3.4}$$

where $\mathbf{W}_{\text{re-embed}}^{(i)}$ acts as a stage-specific Re-embedding Layer that projects the explicit channel estimate back into the high-dimensional virtual latent space. This specific design linearly aggregates the original input prompt ($\mathbf{X}_{\text{token}}^{(i)}$), the uncompressed high-dimensional channel memory ($\mathbf{H}_{\text{feature}}^{(i)}$), and the re-embedded physical channel estimate, ensuring maximum information retention.

The fused sequence $\mathbf{Z}_{\text{eq}}^{(i)}$ is then processed by N_{MD} equalizer blocks (VWN-TPA encoders) to resolve the residual spatial interference. Finally, a Classification Head maps the equalizer's output features into independent real and imaginary Pulse Amplitude Modulation (PAM) logits:

$$\mathbf{Z}_{\text{PAM}}^{(i)} = \text{Encoder}_{\text{MD}}\big(\mathbf{Z}_{\text{eq}}^{(i)}\big)\mathbf{W}_{\text{MD}}^{(i)} \quad \in \mathbb{R}^{B\times L\times(2N_{\text{tx}}\sqrt{M})}, \tag{3.5}$$

where $\mathbf{W}_{\text{MD}}^{(i)}$ is the stage-specific projection weight, and $\sqrt{M}$ represents the PAM alphabet size per quadrature dimension. The output tensor $\mathbf{Z}_{\text{PAM}}^{(i)}$ encompasses the unnormalized logits ($\mathbf{Z}_{\Re}^{(i)}$ and $\mathbf{Z}_{\Im}^{(i)}$) utilized for loss computation and inference resampling. By applying a standard Softmax normalization, these logits are mapped to the stage-specific symbol-wise probability tensor $\hat{\mathbf{P}}^{(i)}$, providing the refined soft information required by the downstream stages or outer-loop channel decoders.

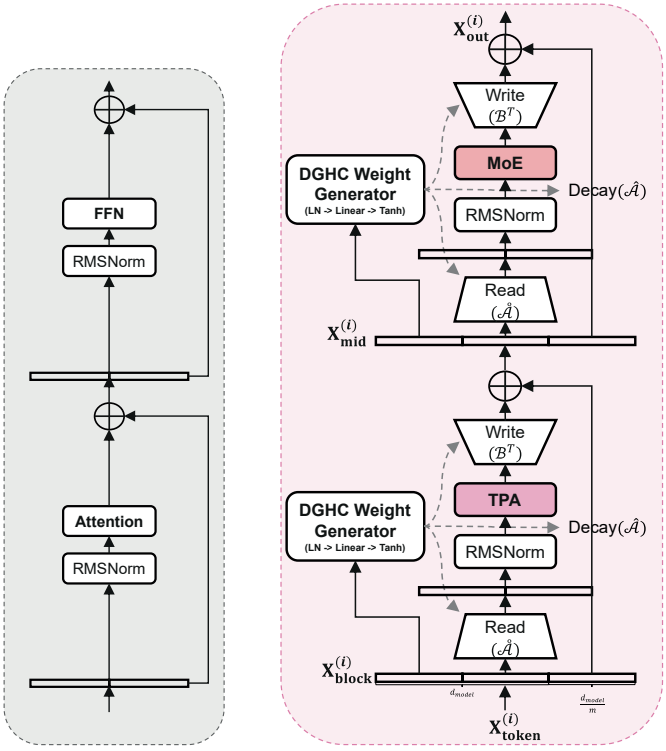
Genie-Aided Perfect CSI Benchmark: To evaluate the isolated equalization capability of the MDM and establish a theoretical upper bound for detection performance, we construct a Genie-Aided Benchmark variant. In this architectural variant, the estimated channel $\hat{\mathbf{H}}_{\text{flat}}^{(i)}$ in the fusion bridge is completely bypassed. Instead, the ground-truth physical channel tensor $\mathbf{H}$ is flattened and injected into the Re-embedding Layer. Thus, the residual bridge in Eq. (3.4) is modified to fuse the true channel realization, isolating the equalizer blocks from CE errors.

3.3.2 Micro-Architecture: Virtual Width Networks and Tensor Product Attention

The macro-architecture coordinates the iterative signal recovery process, whereas the micro-architecture defines feature extraction within each stage. Standard Transformer encoders [25], typically consisting of Multi-Head Attention (MHA) and Feed-Forward Networks (FFN), present a scaling trade-off when applied to high-dimensional MIMO-OFDM signals. Increasing the hidden dimension aids in resolving spatial interference but leads to parameter growth and training sensitivity; conversely, restricted dimensions limit the detection capacity.

To address these design constraints, the proposed architecture adopts VWN [21], incorporating TPA [22] and MoE [9, 23] modules to optimize feature representation efficiency.

Virtual Width Network (VWN)



(a) Vanilla Transformer. **(b)** VWN-TPA encoder.

Figure 3.4: Structural comparison of the encoder blocks.

As illustrated in Fig. 3.4(a), standard Transformer encoders utilize a fixed hidden dimension d_{model} across sequential residual blocks. Scaling this dimension

to resolve the high-dimensional spatial interference of MIMO-OFDM channels leads to a quadratic increase in computational complexity.

The VWN module decouples memory capacity from computational width. The standardized input $\mathbf{X}_{\text{token}}^{(i)}$ is projected to an expanded dimension $d_{\text{virtual}} = (n/m) \cdot d_{\text{model}}$, where $n > m$. To structure this high-dimensional representation, the latent tensor is reshaped into n blocks, yielding the format:

$$\mathbf{X}_{\text{block}}^{(i)} \in \mathbb{R}^{B \times L \times n \times d_{\text{b}}}, \tag{3.6}$$

where the block dimension is defined as $d_{\text{b}} = d_{\text{model}}/m$.

Within each VWN layer, heavy non-linear operations $\mathcal{F}(\cdot)$ (i.e., TPA or MoE) are confined to the narrower m-slot computational subspace. The interaction is governed by Dynamic Generalized Hyper-Connections (DGHC) [21]. 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 $\mathcal{A} \in \mathbb{R}^{(m+n) \times n}$. Partitioning $\mathcal{A}$ horizontally yields 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:

$$\mathbf{X}_{\text{out}}^{(i)} = \mathcal{B}^{\top} \mathcal{F}\big(\mathring{\mathcal{A}} \mathbf{X}_{\text{block}}^{(i)}\big) + \hat{\mathcal{A}} \mathbf{X}_{\text{block}}^{(i)}. \tag{3.7}$$

Equation (3.7) encapsulates the micro-architectural flow: the network tracks complex multi-path residuals in the n-dimensional highway, while isolating expensive dense computations to the efficient m-dimensional subspace via dynamic gating.

Tensor Product Attention (TPA)

While the VWN architecture governs the macroscopic memory capacity, the core interference cancellation within each computational slot is realized via TPA. Standard MHA [25] projects sequences using monolithic parameter matrices, which limits the inductive bias for the MIMO-OFDM PHY. TPA mitigates this constraint by factorizing queries, keys, and values into sums of contextual tensor products [22]. This low-rank parametrization improves parameter efficiency and enables the injection of physical priors into the factorization pathway.

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 $\mathbf{B}_{\text{Q}} \in \mathbb{R}^{B \times L \times r_{\text{q}} \times d_{\text{h}}}$:

$$\mathbf{A}_{\text{Q}}(\mathbf{X}) = \mathbf{X}\mathbf{W}_{\text{A}}^{\text{Q}}, \tag{3.8}$$

$$\mathbf{B}_{\text{Q}}(\mathbf{X}) = \mathbf{X}\mathbf{W}_{\text{B}}^{\text{Q}}, \tag{3.9}$$

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: 2D RoPE and Rank Scaling. Standard 1D rotary position embeddings [24] map time-frequency grids into linear sequences and do not preserve the two-dimensional coherence patterns of MIMO-OFDM channels. To address this, we define the sequence axis as a 2D physical grid and apply $\mathrm{RoPE}_{2\mathrm{D}}$ to the feature bases before tensor contraction:

$$\tilde{\mathbf{B}}_{\mathrm{Q}} = \mathrm{RoPE}_{2\mathrm{D}}\big(\mathbf{B}_{\mathrm{Q}}\big), \qquad \tilde{\mathbf{B}}_{\mathrm{K}} = \mathrm{RoPE}_{2\mathrm{D}}\big(\mathbf{B}_{\mathrm{K}}\big). \tag{3.10}$$

This endows the latent subspaces with spatial awareness along the time and frequency axes. The full query tensor $\mathbf{Q} \in \mathbb{R}^{B\times L\times h\times d_{\mathrm{h}}}$ is reconstituted via tensor contraction:

$$\mathbf{Q} = \mathbf{A}_{\mathrm{Q}}(\mathbf{X})\tilde{\mathbf{B}}_{\mathrm{Q}}. \tag{3.11}$$

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

In the canonical TPA formulation, the tensor product is scaled by $1/r_{\mathrm{q}}$ to manage variance. We omit this internal scaling. In the PHY, the magnitude of a tensor product reflects the energy levels of wireless interference. Preserving this dynamic range allows the network to distinguish between distinct channel states.

Scaled Dot-Product Attention. With contextual factorization established, query, key, and value tensors $\mathbf{Q}, \mathbf{K}, \mathbf{V} \in \mathbb{R}^{B\times L\times h\times d_{\mathrm{h}}}$ are processed via standard attention. For each head $i \in \{1, \ldots, h\}$, let $\mathbf{Q}_i, \mathbf{K}_i, \mathbf{V}_i \in \mathbb{R}^{B\times L\times d_{\mathrm{h}}}$ denote the slices along the head dimension. The attention for head i is computed as:

$$\mathrm{head}_i = \mathrm{Softmax}\left(\frac{\mathbf{Q}_i\mathbf{K}_i^{\top}}{\sqrt{d_{\mathrm{h}}}}\right)\mathbf{V}_i. \tag{3.12}$$

The outputs of all h heads are concatenated and projected via the output weight matrix $\mathbf{W}_{\mathrm{O}} \in \mathbb{R}^{(h\cdot d_{\mathrm{h}})\times d_{\mathrm{model}}}$:

$$\mathrm{TPA}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathrm{Concat}\big(\mathrm{head}_1, \ldots, \mathrm{head}_h\big)\,\mathbf{W}_{\mathrm{O}}. \tag{3.13}$$

The resulting tensor completes the TPA sub-layer and is routed to the subsequent MoE block.

Mixture-of-Experts (MoE)

Following the TPA sub-layer, the token representations are processed by a MoE feed-forward block, adapted from the DeepSeekMoE framework [9, 23]. The expert pool is partitioned into N_s always-active shared experts and N_r selectively activated routed experts, all employing the same SwiGLU-based feed-forward structure [26]. For an input token $\mathbf{x} \in \mathbb{R}^{d_{\mathrm{model}}}$, the MoE forward pass is given by

$$\mathbf{y} = \sum_{j=1}^{N_s} \mathrm{MLP}_s^{(j)}(\mathbf{x}) + \sum_{k\in\mathcal{K}} g_k(\mathbf{x})\,\mathrm{MLP}_r^{(k)}(\mathbf{x}), \tag{3.14}$$

where $\mathcal{K}$ is the set of top-K_r experts selected by a lightweight gating network. The gating scores are computed via a linear projection followed by Sigmoid activation,

$\mathbf{s} = \text{Sigmoid}(\mathbf{x}\mathbf{W}_{\text{route}})$, and the top-$K_r$ entries are normalized to produce the routing weights

$$g_k(\mathbf{x}) = \frac{s_k}{\sum_{j\in\mathcal{K}} s_j + \epsilon}, \quad k \in \mathcal{K}, \tag{3.15}$$

with ϵ ensuring numerical stability.

Two simplifications distinguish our design from its NLP counterparts. First, the expert pool is kept deliberately compact: empirical evaluation shows that only a small number of routed experts are necessary, reflecting the inherently structured nature of MIMO-OFDM interference patterns. Second, all auxiliary load-balancing losses are removed. In language tasks, these losses prevent routing collapse across semantically diverse tokens; however in the PHY, channel impairments such as deep fades and interference that a few specialized experts can handle without forced uniform load distribution. Removing the balancing bias allows the routing to directly reflect the conditional activation mapping of the physical environment.

3.4 Physics-Guided Feature Construction (PGFC)

While the VWN-TPA architecture provides a powerful computational engine, a purely data-driven approach faces limitations in modeling the high-dimensional state space of MIMO-OFDM detection. To improve convergence and incorporate physical domain knowledge, we propose the PGFC mechanism. Rather than propagating raw neural outputs across refinement stages, the PGFC mechanism functions as a differentiable bridge. It constructs communication-theoretic tensors, including IC residuals and MF outputs, derived from current channel and symbol estimates.

3.4.1 Prior-Constrained Soft Estimation

At the end of the i-th iteration, the backbone outputs the CE $\hat{\mathbf{H}}^{(i)}$ and classification logits. Let $\mathbf{p}^{(i)} \in [0,1]^{\sqrt{M}}$ denote the softmax probability distribution over the PAM constellation set $\mathcal{C}$ for a given spatial stream. The expected complex soft symbol tensor $\tilde{\mathbf{S}}_{\text{c}}^{(i)}$ and its normalized entropy $\mathcal{H}^{(i)}$ are computed as

$$\tilde{\mathbf{S}}_{\text{c}}^{(i)} = \sum_{c\in\mathcal{C}} \mathbf{p}_c^{(i)} \cdot c, \tag{3.16}$$

$$\mathcal{H}^{(i)} = -\frac{1}{\log_2(\sqrt{M})} \sum_{c\in\mathcal{C}} \mathbf{p}_c^{(i)} \log_2(\mathbf{p}_c^{(i)} + \epsilon), \tag{3.17}$$

where $\mathbf{p}_c^{(i)}$ denotes the probability associated with constellation point c, and ϵ prevents numerical instability. The entropy provides an explicit per-RE confidence measure.

To construct the updated soft state tensor $\tilde{\mathbf{S}}^{(i)}$ for the next stage, the complex estimates are first mapped to their real-valued representation $\mathbf{S}_{\text{soft}}^{(i)}$ by stacking the real and imaginary components. These soft estimates are combined with the known pilot values $\mathbf{S}_{\text{p}}$ via the universal pilot mask $\mathbf{M}_{\text{p}}$, as defined in Section 3.2.

To ensure dimensional consistency, we form a real-valued composite transmission symbol $\mathbf{S}_{\text{comp}}^{(i)} \in \mathbb{R}^{B\times N_{\text{tx}}\times N_{\text{sym}}\times N_{\text{sc}}\times 2}$, which is subsequently concatenated with the entropy tensor. Two injection strategies are supported:
Orthogonal Pilot Injection: Pilots and data occupy disjoint REs. The composite symbol is formed as

$$\mathbf{S}_{\text{comp}}^{(i)} = \mathbf{M}_{\text{p}} \odot \mathbf{S}_{\text{p}} + (\mathbf{1} - \mathbf{M}_{\text{p}}) \odot \mathbf{S}_{\text{soft}}^{(i)}. \tag{3.18}$$

The updated soft state is then assembled as $\tilde{\mathbf{S}}^{(i)} = \left[\, \mathbf{S}_{\text{comp}}^{(i)},\ (\mathbf{1} - \mathbf{M}_{\text{p}}) \odot \mathcal{H}^{(i)} \,\right]$. The entropy is masked to zero at pilot positions, reflecting the fact that deterministic pilots carry no estimation uncertainty.
Superimposed Pilot Injection: Data spans all REs, and pilots are superimposed on the subset defined by $\mathbf{M}_{\text{p}}$ with a power allocation factor $\rho \in (0, 1)$. The composite symbol is formulated as

$$\mathbf{S}_{\text{comp}}^{(i)} = \mathbf{M}_{\text{p}} \odot \left(\sqrt{\rho}\,\mathbf{S}_{\text{p}} + \sqrt{1-\rho}\,\mathbf{S}_{\text{soft}}^{(i)}\right) + (\mathbf{1} - \mathbf{M}_{\text{p}}) \odot \mathbf{S}_{\text{soft}}^{(i)}. \tag{3.19}$$

Here, the soft data symbol dynamically backs off its amplitude to $\sqrt{1-\rho}$ at pilot-carrying REs while transmitting at full power elsewhere. The soft state is subsequently assembled as $\tilde{\mathbf{S}}^{(i)} = \left[\, \mathbf{S}_{\text{comp}}^{(i)},\ \mathcal{H}^{(i)} \,\right]$. Unlike the orthogonal case, the entropy is continuously cascaded across all REs to guide interference cancellation under the coupled pilot-data interference.

Finally, a stop-gradient operation is applied to $\tilde{\mathbf{S}}^{(i)}$ to isolate the backward pass within each iterative stage, stabilizing the deep supervision dynamics across the unfolded cascade.

3.4.2 Physical Feature Extraction

To compute the classical linear detection primitives, the PGFC module operates in the complex domain. Let $\mathbf{S}_{\text{c,comp}}^{(i)} \in \mathbb{C}^{B\times N_{\text{tx}}\times N_{\text{sym}}\times N_{\text{sc}}}$ denote the complex-valued equivalent of the composite symbol tensor constructed in Section 3.4.1. Given the complex CE $\hat{\mathbf{H}}^{(i)}$ and the physical observation $\mathbf{Y}$, the module synthesizes four physical feature tensors:

1. **Soft Signal Residual (IC):** An IC step reconstructs the expected received signal and subtracts it from the actual observation:

   $$\mathbf{R}^{(i)} = \mathbf{Y} - \hat{\mathbf{H}}^{(i)}\mathbf{S}_{\text{c,comp}}^{(i)}. \tag{3.20}$$

   This residual isolates the unresolved interference and noise by explicitly removing the explained signal energy.

2. **MF Response:** The observation tensor is projected onto the estimated channel subspace to maximize the per-stream signal-to-noise ratio:

   $$\mathbf{Z}_{\text{MF}}^{(i)} = (\hat{\mathbf{H}}^{(i)})^{\mathsf{H}}\mathbf{Y}, \tag{3.21}$$

   where $(\cdot)^{\mathsf{H}}$ denotes the Hermitian transpose.

3. **Gram Matrix:** The spatial Gram matrix explicitly quantifies the inter-antenna interference (IAI) caused by non-orthogonal channel vectors:

$$\mathbf{G}^{(i)} = (\hat{\mathbf{H}}^{(i)})^{\mathsf{H}}\hat{\mathbf{H}}^{(i)}. \tag{3.22}$$

4. **Residual MF Response (MF-IC):** The interference residual is projected back into the spatial stream domain, providing a localized error gradient for the subsequent equalizer:

$$\mathbf{R}_{\mathrm{MF}}^{(i)} = (\hat{\mathbf{H}}^{(i)})^{\mathsf{H}}\mathbf{R}^{(i)}. \tag{3.23}$$

To interface with the AI Detection Backbone, these complex-valued tensors are decoupled into their real and imaginary components. Let $(\cdot)_{\mathbb{R}}$ denote this real-domain mapping along the feature axis. The components are concatenated with the raw real-valued observation $\mathbf{Y}_{\mathbb{R}}$ to form the augmented observation tensor:

$$\mathbf{\Phi}^{(i)} = \mathrm{Concat}\big(\mathbf{Y}_{\mathbb{R}},\ \mathbf{R}_{\mathbb{R}}^{(i)},\ \mathbf{Z}_{\mathrm{MF},\mathbb{R}}^{(i)},\ \mathbf{G}_{\mathbb{R}}^{(i)},\ \mathbf{R}_{\mathrm{MF},\mathbb{R}}^{(i)}\big). \tag{3.24}$$

By embedding these communication-theoretic operations into the forward pass, the PGFC module supplies the subsequent VWN-TPA encoder with a physically structured feature set. This relieves the network from learning fundamental spatial projections strictly from data, thereby accelerating convergence and improving generalization across varying channel conditions.

Finally, the augmented observation tensor $\mathbf{\Phi}^{(i)}$, the updated soft state $\tilde{\mathbf{S}}^{(i)}$, and the invariant pilot mask $\mathbf{M}_{\mathrm{p}}$ are forwarded to the tokenizer of the next iterative stage, closing the refinement loop.

Chapter 4

Training Framework and Detection Enhancement Strategies

This chapter describes the training procedure for the proposed receiver and evaluates several methods for improving its detection performance. First, the joint CE and symbol detection objective is introduced, together with deep supervision across the iterative stages and the structured channel augmentations used during training. The chapter then investigates three detection enhancement methods: soft Log-MAP supervision generated by a K-best detector, an additional LLR refinement head, and inference-time Invariant-Transformation Resampling. In the evaluated setting, the first two methods provide only limited improvements, whereas resampling yields a clearer performance gain without modifying or retraining the trained receiver.

4.1 Loss Function and Training Strategy

This section formulates the composite loss and the supervision mechanism utilized to train the iterative architecture. The training strategy also incorporates physics-informed data augmentation, which is detailed at the conclusion of this section.

4.1.1 Joint Decomposed Loss

To optimize both CE and symbol detection simultaneously while managing computational complexity, the M-QAM symbol detection is decomposed into independent in-phase and quadrature $\sqrt{M}$-PAM classification tasks. For the i-th refinement stage, let $\mathbf{Z}_{\Re}^{(i)}, \mathbf{Z}_{\Im}^{(i)} \in \mathbb{R}^{B\times N_{\mathrm{tx}}\times\sqrt{M}\times N_{\mathrm{sym}}\times N_{\mathrm{sc}}}$ denote the unnormalized logits for the real and imaginary PAM components. Let $\mathbf{T}_{\Re}, \mathbf{T}_{\Im} \in \{0,\dots,\sqrt{M}-1\}^{B\times N_{\mathrm{tx}}\times N_{\mathrm{sym}}\times N_{\mathrm{sc}}}$ denote the corresponding ground-truth PAM target indices.

The symbol detection loss is computed exclusively over the data-carrying resource elements. Let $\mathbf{M}_{\mathrm{d}} \in \{0,1\}^{B\times N_{\mathrm{tx}}\times N_{\mathrm{sym}}\times N_{\mathrm{sc}}}$ define the binary data mask, where the entries are 1 at data-carrying locations and 0 otherwise. The spatially

masked cross-entropy loss for stage i is formulated using the Hadamard product:

$$\mathcal{L}_{\mathrm{S}}^{(i)} = \frac{1}{\|\mathbf{M}_{\mathrm{d}}\|_1} \sum \Big(\mathbf{M}_{\mathrm{d}} \odot \big[\mathrm{CE}\big(\mathbf{Z}_{\Re}^{(i)}, \mathbf{T}_{\Re}\big) + \mathrm{CE}\big(\mathbf{Z}_{\Im}^{(i)}, \mathbf{T}_{\Im}\big) \big] \Big), \tag{4.1}$$

where $\mathrm{CE}(\cdot,\cdot)$ denotes the unreduced element-wise cross-entropy operator, $\sum$ aggregates the values across all tensor dimensions, and the L1 norm $\|\mathbf{M}_{\mathrm{d}}\|_1$ acts as the normalization factor, representing the total number of active data elements within the batch.

The CE loss computes the mean squared error (MSE) between the estimated and ground-truth complex channel tensors:

$$\mathcal{L}_{\mathrm{H}}^{(i)} = \mathrm{MSE}\big(\hat{\mathbf{H}}^{(i)}, \mathbf{H}\big). \tag{4.2}$$

The stage-wise composite loss balances the two objectives via a hyperparameter $\beta \in [0,1]$:

$$\mathcal{L}_{\mathrm{stage}}^{(i)} = \beta \mathcal{L}_{\mathrm{H}}^{(i)} + (1-\beta)\mathcal{L}_{\mathrm{S}}^{(i)}. \tag{4.3}$$

To mitigate vanishing gradients across the unfolded cascade, we employ deep supervision that injects gradients into all intermediate stages. Let $N_{\mathrm{it}} \geq 1$ denote the total number of refinement stages, yielding $N_{\mathrm{it}}+1$ detection stages in total (one bootstrap stage followed by N_{it} refinement stages). The total training loss is a weighted combination of the final refinement stage loss and the auxiliary losses from all preceding stages:

$$\mathcal{L}_{\mathrm{total}} = (1-\alpha)\mathcal{L}_{\mathrm{stage}}^{(N_{\mathrm{it}})} + \frac{\alpha}{N_{\mathrm{it}}} \sum_{i=0}^{N_{\mathrm{it}}-1} \mathcal{L}_{\mathrm{stage}}^{(i)}, \tag{4.4}$$

where $\alpha \in [0,1]$ controls the strength of intermediate supervision, and $i=0$ corresponds to the bootstrap stage. For the default configuration with $N_{\mathrm{it}}=1$, the total loss simplifies to $\mathcal{L}_{\mathrm{total}} = (1-\alpha)\mathcal{L}_{\mathrm{stage}}^{(1)} + \alpha\mathcal{L}_{\mathrm{stage}}^{(0)}$.

4.1.2 Physics-Informed Data Augmentation

To improve generalization and prevent overfitting to the specific channel, we apply structured data augmentations to the complex channel tensor $\mathbf{H}$. These transformations are stochastically composed via a randomized pipeline during training.

1. Spatial Invariance and Symmetry. To enforce the inductive bias that MIMO channel statistics remain invariant to arbitrary antenna indexing, we apply:

- *Antenna Permutation:* Receive and transmit antenna indices are independently shuffled to prevent the memorization of fixed spatial orderings.
- *Complex Conjugation:* The channel is probabilistically replaced by its complex conjugate, $\mathbf{H}_{\mathrm{aug}} = \mathbf{H}^*$. This inverts the phase signature, generating symmetric channel realizations to expand the training support.

2. Spatial Subspace Perturbation. To diversify the spatial correlation properties of the training samples, exactly one of the following transformations is randomly applied per forward pass:

- *Random Unitary Transformation:* A random unitary matrix $\mathbf{U}_{\mathrm{rx}}$, derived via the QR decomposition of a Gaussian matrix, is applied to the receive side ($\mathbf{H}_{\mathrm{aug}} = \mathbf{U}_{\mathrm{rx}}\mathbf{H}$) to rotate the spatial subspace.

- *Singular Value Decomposition (SVD)-Based Perturbation:* Using the SVD of the time-frequency-averaged channel, $\bar{\mathbf{H}} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathsf{H}}$, uniform jitter is injected into the singular values: $\mathbf{\Sigma}_{\mathrm{aug}} = \mathbf{\Sigma} \odot (\mathbf{1} + \mathbf{\Delta})$, where $\mathbf{\Delta} \sim \mathcal{U}(-\delta_{\max}, \delta_{\max})$, with $\delta_{\max}$ representing a configurable maximum singular value jitter. These augmented singular values are used to reconstruct the channel, which exhibits an altered condition number, expanding the diversity of spatial multiplexing scenarios encountered during training.

- *SVD Unitary Projection:* The projection $\mathbf{H}_{\mathrm{aug}} = \mathbf{U}^{\mathsf{H}}\mathbf{H}$ exposes the model to an alternative spatial coordinate system and reduces its dependence on a fixed receive-side basis.

3. Cross-Sample Channel Mixing. To construct diverse channels, we generate additional channel samples by mixing antenna slices from two independently sampled channel tensors. Specifically, a random subset of receive or transmit antenna slices in the current channel is replaced by the corresponding slices from another channel sample in the mini-batch. This operation increases the diversity of training channels and reduces overfitting to the simulated dataset.

4. Stochastic Noise Injection. To prevent the network from overfitting to the idealized channel distributions, we apply noise regularization to the channel:

- *Phase Jitter:* $\mathbf{H}_{\mathrm{aug}} = \mathbf{H} \odot \exp(\mathrm{j}\mathbf{\Theta})$, where the phase noise tensor is uniformly distributed as $\mathbf{\Theta} \sim \mathcal{U}(-\theta_{\max}, \theta_{\max})$, with $\theta_{\max}$ representing a configurable maximum phase deviation. This introduces random continuous phase rotations to the channel coefficients.

- *Channel Perturbation:* An independent additive complex Gaussian noise tensor $\mathbf{N}_{\mathrm{H}}$ is injected directly into the channel prior to signal synthesis: $\mathbf{H}_{\mathrm{aug}} = \mathbf{H} + \mathbf{N}_{\mathrm{H}}$. The perturbation tensor is distributed as $\mathbf{N}_{\mathrm{H}} \sim \mathcal{CN}(\mathbf{0}, \sigma_{\mathrm{p}}^2\mathbf{I})$, where the variance is defined as $\sigma_{\mathrm{p}}^2 = \rho_{\mathrm{p}}\mathbb{E}[\|\mathbf{H}\|^2]$ and ρ_{p} is a configurable perturbation factor. By decoupling the channel uncertainty from the operating SNR, this additive noise encourages the network to learn robust decision boundaries rather than overfitting exact channel values.

4.2 Exploration of Detection Enhancement Techniques

During architectural development, several strategies were investigated to further improve the detection accuracy of the ICL receiver. This section documents three directions, two of which yielded marginal gains while the third, Invariant-Transformation Resampling, provided a consistent and effective improvement. For each approach, we present the technical formulation, empirical observations, and an analysis of its limitations or effectiveness.

4.2.1 Knowledge Distillation via Soft Log-MAP Supervision

We generate soft detection targets using a K-best sphere decoder with a sufficiently large value of K. Under the evaluated MIMO configuration, the generated LLRs closely match those obtained from Log-MAP MLD. These LLRs are then used as soft supervision signals for training the ICL receiver. The evaluation is performed under the Genie-Aided ideal-CSI setting, so that the comparison focuses on the equalization performance of the model.

Let $P_{\mathrm{MAP}}(b_k)$ denote the posterior probability of the k-th bit generated by the Log-MAP approximation. The target LLR is formulated as:

$$\mathrm{LLR}_k^{\mathrm{target}} = \ln \frac{P_{\mathrm{MAP}}(b_k = 0)}{P_{\mathrm{MAP}}(b_k = 1)}. \tag{4.5}$$

These LLRs are subsequently mapped to soft probability targets $p_k = P_{\mathrm{MAP}}(b_k = 1)$ via the sigmoid function $\sigma(\cdot)$:

$$p_k = \frac{1}{1 + \exp(\mathrm{LLR}_k^{\mathrm{target}})} = \sigma(-\mathrm{LLR}_k^{\mathrm{target}}). \tag{4.6}$$

To train the neural network to replicate this theoretical distribution, we replace the standard hard-label cross-entropy with a soft binary cross-entropy (BCE) loss. Let $\hat{p}_k = \sigma(-\mathrm{LLR}_k^{\mathrm{pred}})$ denote the bit probability predicted by the neural network. The soft BCE loss is given by:

$$\mathcal{L}_{\mathrm{soft\_BCE}} = -\sum_k \Big[p_k \ln(\hat{p}_k) + (1 - p_k) \ln(1 - \hat{p}_k) \Big]. \tag{4.7}$$

Mathematically, this optimization objective is equivalent to minimizing the Kullback-Leibler (KL) divergence between the MAP target distribution and the neural network's predicted distribution. Since the entropy of the target distribution $H(P_{\mathrm{MAP}})$ is constant with respect to the network parameters θ, we have:

$$\nabla_\theta \mathcal{L}_{\mathrm{soft\_BCE}} = \nabla_\theta D_{\mathrm{KL}}\big(P_{\mathrm{MAP}} \parallel P_{\mathrm{NN}}\big) + \nabla_\theta H(P_{\mathrm{MAP}}) = \nabla_\theta D_{\mathrm{KL}}\big(P_{\mathrm{MAP}} \parallel P_{\mathrm{NN}}\big). \tag{4.8}$$

In the experiments, soft-target training produced small but consistent improvements over hard-label training. Under soft supervision, the Block Error Rate (BLER), Symbol Error Rate (SER), and uncoded Bit Error Rate (BER) were all slightly reduced.

We also compare the hard decisions of the AI receiver with those of the MAP-based reference detector. With standard cross-entropy training, some samples were correctly detected by the AI receiver while the MAP-based reference decision was incorrect. These cases indicate a mismatch between the learned decision rule and the reference detector, rather than a systematic advantage of the AI receiver. After training with soft targets, this mismatch ratio decreased. At the same time, the fraction of samples where both detectors made the same decision increased slightly.

These results suggest that soft supervision helps align the AI receiver more closely with the MAP-based reference detector. However, the improvement re-

mains limited, indicating that the VWN-TPA model trained with standard hard-label cross-entropy already learns decision regions that are largely consistent with the reference detector.

4.2.2 LLR Refinement Head

To further mitigate residual detection errors, a lightweight post-processing module was designed to refine the baseline logits by exploiting physical features. For each transmit antenna, a feature tensor $\mathbf{F}_{\text{refine}}$ is constructed from the baseline LLR predictions, the corresponding soft symbol estimates $\mathbf{S}_{\text{soft}}$, the matched filter response $\mathbf{Z}_{\text{MF}} = \mathbf{H}^{\text{H}}\mathbf{Y}$, the residual projection $\mathbf{R}_{\text{MF}} = \mathbf{H}^{\text{H}}(\mathbf{Y} - \mathbf{H}\mathbf{S}_{\text{soft}})$, the spatial Gram matrix $\mathbf{G} = \mathbf{H}^{\text{H}}\mathbf{H}$, and the localized SNR.

These features are processed by a lightweight TPA Transformer encoder. The output projection layer maps the refined representations to unnormalized energy values E_s for each PAM constellation point s. The refined LLR for the k-th bit is computed via the Log-Sum-Exp operation:

$$\text{LLR}_{\text{refine}}^{(k)} = \ln\left(\sum_{s\in\mathcal{S}_0^{(k)}} \exp(E_s)\right) - \ln\left(\sum_{s\in\mathcal{S}_1^{(k)}} \exp(E_s)\right), \tag{4.9}$$

where $\mathcal{S}_0^{(k)}$ and $\mathcal{S}_1^{(k)}$ denote the PAM constellation subsets for which the k-th bit is 0 and 1, respectively. The final LLR is obtained by blending with the baseline prediction:

$$\text{LLR}_{\text{final}} = \gamma\text{LLR}_{\text{base}} + (1-\gamma)\text{LLR}_{\text{refine}}, \tag{4.10}$$

where $\gamma \in [0, 1]$ is a predefined blending hyperparameter. We set $\gamma = 0.5$ in our default training and evaluation configuration.

In the experiments, the refinement head provided only marginal BLER improvement. A symbol-level comparison showed that the module corrected some errors made by the baseline model, but also introduced errors on some symbols that were originally detected correctly. As a result, the overall gain remained limited after blending.

This result suggests that the baseline VWN-TPA model already captures most of the information that can be exploited by this additional refinement head. The remaining errors are not easily corrected by a shallow post-processing module based on the selected auxiliary features. Therefore, the following analysis focuses on inference-time strategies that use the trained model directly, rather than adding an extra refinement stage.

4.2.3 Invariant-Transformation Resampling Strategies

The two preceding approaches share a practical limitation: whether by modifying the training objective or by appending an additional parametric module, both yield only marginal improvements. This suggests that reshaping the learned decision rule or adding a shallow refinement stage provides limited additional gains beyond the VWN-TPA backbone.

We consider an inference-time ensemble based on invariant transformations. The trained AI receiver processes multiple transformed versions of the received signal and channel representation. These transformations preserve the underlying detection problem while changing the input representation seen by the model. Therefore, the corresponding outputs can be mapped back to the original constellation space and aggregated. Since the inference errors across these equivalent representations are not perfectly correlated, averaging the aligned outputs can reduce residual inference errors. We refer to this strategy as Invariant-Transformation Resampling [27].

Let $\mathbf{Z}_{\Re}, \mathbf{Z}_{\Im} \in \mathbb{R}^{B\times N_{\mathrm{tx}}\times\sqrt{M}\times N_{\mathrm{sym}}\times N_{\mathrm{sc}}}$ denote the classification logits generated by the baseline inference function $F(\mathbf{Y}, \mathbf{S}_{\mathrm{p}}, \mathbf{H}, \mathbf{M}_{\mathrm{p}})$. Querying the model with transformed inputs yields raw logits $\mathbf{Z}'_{\Re}$ and $\mathbf{Z}'_{\Im}$, which must be aligned back to the original target domain. Let $\mathrm{Rev}(\cdot)$ denote the reversal of the $\sqrt{M}$-PAM logit bin index order. We define four distinct signal symmetries:
1. Negative (Point Symmetry). A global sign inversion yields $-\mathbf{Y} = \mathbf{H}(-\mathbf{s}) - \mathbf{N}$. The aligned logits are recovered by reversing both PAM bins:

$$\mathbf{Z}_{\Re,\mathrm{neg}} = \mathrm{Rev}(\mathbf{Z}'_{\Re}), \quad \mathbf{Z}_{\Im,\mathrm{neg}} = \mathrm{Rev}(\mathbf{Z}'_{\Im}), \tag{4.11}$$

where $\mathbf{Z}'_{\Re}, \mathbf{Z}'_{\Im} = F(-\mathbf{Y}, -\mathbf{S}_{\mathrm{p}}, \mathbf{H}, \mathbf{M}_{\mathrm{p}})$.
2. Conjugate (I/Q Symmetry). The I/Q swap $\mathbf{s} \to \mathrm{j}\mathbf{s}^*$ yields $\mathrm{j}\mathbf{Y}^* = \mathbf{H}^*(\mathrm{j}\mathbf{s}^*) + \mathrm{j}\mathbf{N}^*$. The aligned logits are restored by swapping the real and imaginary logit branches:

$$\mathbf{Z}_{\Re,\mathrm{conj}} = \mathbf{Z}'_{\Im}, \quad \mathbf{Z}_{\Im,\mathrm{conj}} = \mathbf{Z}'_{\Re}, \tag{4.12}$$

where $\mathbf{Z}'_{\Re}, \mathbf{Z}'_{\Im} = F(\mathrm{j}\mathbf{Y}^*, \mathrm{j}\mathbf{S}^*_{\mathrm{p}}, \mathbf{H}^*, \mathbf{M}_{\mathrm{p}})$.
3. Unitary (Spatial Basis Rotation). A random unitary receive matrix $\mathbf{U}$ yields $\mathbf{UY} = (\mathbf{UH})\mathbf{s} + \mathbf{UN}$. This rotates the spatial basis without altering the transmitted symbols; thus, no inverse logit mapping is required:

$$\mathbf{Z}_{\Re,\mathrm{uni}} = \mathbf{Z}'_{\Re}, \quad \mathbf{Z}_{\Im,\mathrm{uni}} = \mathbf{Z}'_{\Im}, \tag{4.13}$$

where $\mathbf{Z}'_{\Re}, \mathbf{Z}'_{\Im} = F(\mathbf{UY}, \mathbf{S}_{\mathrm{p}}, \mathbf{UH}, \mathbf{M}_{\mathrm{p}})$.
4. Rot90 (Quadrature Rotation). A 90° phase shift $\mathbf{s} \to \mathrm{j}\mathbf{s}$ yields $\mathrm{j}\mathbf{Y} = \mathbf{H}(\mathrm{j}\mathbf{s}) + \mathrm{j}\mathbf{N}$. To reverse this mapping in the logit space, the branches are crossed and inverted:

$$\mathbf{Z}_{\Re,\mathrm{rot}} = \mathbf{Z}'_{\Im}, \quad \mathbf{Z}_{\Im,\mathrm{rot}} = \mathrm{Rev}(\mathbf{Z}'_{\Re}), \tag{4.14}$$

where $\mathbf{Z}'_{\Re}, \mathbf{Z}'_{\Im} = F(\mathrm{j}\mathbf{Y}, \mathrm{j}\mathbf{S}_{\mathrm{p}}, \mathbf{H}, \mathbf{M}_{\mathrm{p}})$.
Inference Aggregation. During evaluation, the baseline prediction $\mathbf{Z}_{\mathrm{ori}}$ and the selected resampling branches are processed in parallel. The final ensembled logits are obtained by averaging across the active paths $\mathcal{S} \subseteq \{\mathrm{ori}, \mathrm{neg}, \mathrm{conj}, \mathrm{uni}, \mathrm{rot}\}$, where $\mathcal{S}$ represents the specific subset of branches selected for a given ensemble configuration:

$$\mathbf{Z}_{c,\mathrm{final}} = \frac{1}{|\mathcal{S}|}\sum_{k\in\mathcal{S}} \mathbf{Z}_{c,k}, \quad \text{for } c \in \{\Re, \Im\}. \tag{4.15}$$

The final softmax probability distributions $\hat{\mathbf{P}}_{\Re}$ and $\hat{\mathbf{P}}_{\Im}$ are then computed from these ensembled logits.

Chapter 5
Experimental Evaluation and Results

This chapter evaluates the proposed ICL receiver framework. In addition to benchmarking the architecture against classical signal processing baselines and theoretical bounds, this chapter analyzes the performance of the joint CE and MIMO detection across different configurations. These evaluations are conducted under various spatial multiplexing dimensions, modulation orders, pilot patterns, and channels.

The remainder of this chapter is organized as follows. Section 5.1 details the simulation setup, including the 3GPP-aligned PHY parameters, channel models, and baseline definitions. Section 5.2 analyzes the performance gain of the iterative mechanism across 5G-NR and superimposed pilot schemes. Section 5.3 evaluates the inference-time performance improvements provided by the resampling strategies under ideal and practical CSI conditions. Section 5.4 investigates the impact of pilot sparsity on detection accuracy and throughput, comparing the proposed ICL pilot scheme against traditional 5G-NR allocations. Finally, Section 5.5 characterizes the performance of the non-orthogonal SI DMRS scheme, evaluating its CE accuracy andthroughput envelopes across multiple coding rates.

5.1 Simulation Setup and Baseline Definitions

5.1.1 System Parameters and Channel Models

To evaluate the performance of the proposed ICL receiver under 5G-NR MIMO frameworks, the PHY configurations and channel models are established in alignment with 3GPP specifications, as summarized in Table 5.1.

The simulation environment is implemented using the PyTorch deep learning library and accelerated on an NVIDIA RTX 6000 Ada Generation GPU. The multipath fading channel is modeled using the standardized Extended Typical Urban (ETU70) and Extended Pedestrian A (EPA5) [28] profiles under spatially uncorrelated conditions. For evaluation, the generated datasets comprise 800,000 (ETU70) and 640,000 (EPA5) frames for the 2×2 configurations, alongside 200,000 (ETU70) and 160,000 (EPA5) frames for the 4×4 configurations. All datasets are

partitioned into 81% training, 9% validation, and 10% testing sets. We benchmark the BLER and throughput of our proposed architecture against classical baselines (e.g., Unbiased LMMSE and exact Log-MAP MLD) across both 2×2 and 4×4 MIMO configurations, coupled with multi-level modulation schemes and different Low-Density Parity-Check (LDPC) code rates.

Table 5.1: Summary of system parameters and channel models.

Symbol	Description	Value
	MIMO-OFDM Configuration	
$N_{\mathrm{rx}} \times N_{\mathrm{tx}}$	Spatial configuration	2×2, 4×4
$\mathcal{A}$	Modulation scheme	16-QAM, 64-QAM
N_{sym}	OFDM symbols per frame	14
N_{sc}	Active subcarriers per frame	24
Code	Channel coding scheme	LDPC
R	Code rate	$1/3$, $1/2$, $2/3$
ρ	SI DMRS power allocation factor	0.8
	Channel Model	
Profile	Multipath fading profile	ETU, EPA
f_D	Maximum Doppler shift	70 Hz, 5 Hz
Spatial correlation	Transmit / receive correlation	Uncorrelated

5.1.2 Model Configuration and Training Protocol

The proposed receiver is instantiated with the hyperparameters summarized in Table 5.2. The VWN chassis is configured with $m = 2$ memory slots and $n = 3$ active computational slots per layer, yielding 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 $r = 16$. The MoE sub-layer employs $N_s = 1$ shared expert, $N_r = 4$ routed experts, and $K_r = 2$ active experts per token, with the per-expert hidden dimension scaled by a factor of $(8/3)/(K_r + N_s)$ to maintain FLOP equivalence with a dense SwiGLU feed-forward network [26].

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 [29], while one-dimensional parameters, including layer normalizations and biases, are routed to AdamW [30]. The learning rate follows a linear warmup at first 10% steps followed by cosine annealing to a minimum ratio of 0.01 of the peak value [31, 32]. Physics-informed data augmentation, as described in Section 4.1.1, is applied throughout training.

5.1.3 Baseline Definitions

To establish performance benchmarks, the proposed ICL receiver is evaluated against classical signal processing limits and theoretical upper bounds. All baseline

Table 5.2: Default hyperparameter configuration of the proposed ICL 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 $(8/3)/(K_r + N_s)$	$8/9 \approx 0.889$
Dropout	Dropout probability	0.1
	Cascade Architecture	
$N_{CE}^{(0)}$ / $N_{MD}^{(0)}$	Bootstrap encoder / equalizer layers	3 / 3
$N_{CE}^{(i)}$ / $N_{MD}^{(i)}$	Refinement encoder / equalizer layers	6 / 6
N_{it}	Number of refinement stages (default)	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	Loss balance: CE vs. MSE for $\mathbf{H}$	0.1
α	Deep supervision: auxiliary stage weight	0.3

algorithms are supplied with ground-truth CSI and output soft LLRs compatible with downstream soft-decision decoding.

1. **Unbiased LMMSE Baseline.** This classical linear detector serves as the industrial performance benchmark. Given the true channel matrix $\mathbf{H}$ and the signal-to-noise ratio SNR, the noise variance per receive antenna is computed as $\sigma_n^2 = 1/\text{SNR}$, assuming unit average transmit power. The LMMSE filtering matrix is

$$\mathbf{W} = (\mathbf{H}^H\mathbf{H} + \sigma_n^2\mathbf{I})^{-1}\mathbf{H}^H. \tag{5.1}$$

To prevent the biased amplitude shrinkage inherent to standard MMSE filtering, the equivalent channel gain is extracted from the inverse term as

$$\boldsymbol{\mu} = \operatorname{diag}\big(\mathbf{I} - \sigma_n^2(\mathbf{H}^H\mathbf{H} + \sigma_n^2\mathbf{I})^{-1}\big). \tag{5.2}$$

The unbiased symbol estimates are then obtained as $\hat{\mathbf{s}}_{\text{unb}} = (\mathbf{W}\mathbf{y}) \oslash \boldsymbol{\mu}$, with per-stream effective noise variance

$$\sigma_{\text{eff}}^2 = \sigma_n^2 \cdot \operatorname{diag}\big((\mathbf{H}^H\mathbf{H} + \sigma_n^2\mathbf{I})^{-1}\big) \oslash \boldsymbol{\mu}. \tag{5.3}$$

Soft Log-MAP LLRs are extracted from the unbiased estimates as follows. For each spatial stream, the distance from the unbiased estimate $\hat{s}$ to each

QAM constellation point $c \in \mathcal{A}$ is evaluated under the per-stream effective noise variance:

$$d(c) = \frac{|\hat{s} - c|^2}{\sigma_{\text{eff}}^2}, \quad \forall c \in \mathcal{A}. \tag{5.4}$$

The LLR for the b-th bit is then obtained via the stabilized Log-Sum-Exp operation over the constellation subsets $\mathcal{A}_0^{(b)}$ and $\mathcal{A}_1^{(b)}$ where the b-th bit is 0 and 1, respectively:

$$LLR_b = \ln\left(\sum_{c \in \mathcal{A}_0^{(b)}} \exp\big(-d(c)\big)\right) - \ln\left(\sum_{c \in \mathcal{A}_1^{(b)}} \exp\big(-d(c)\big)\right). \tag{5.5}$$

This unbiased formulation ensures that the soft information quality is not degraded by residual amplitude distortion, establishing a stronger linear baseline than conventional biased LMMSE receivers.

2. **Log-MAP MLD Baseline.** To establish the theoretical performance ceiling, an exact Maximum Likelihood Detection (MLD) baseline is implemented. This detector exhaustively evaluates every candidate symbol vector in the transmitted lattice $\mathbb{X} = \mathcal{A}^{N_{\text{tx}}}$ by computing the full distance metric:

$$d(\mathbf{x}) = \frac{\|\mathbf{y} - \mathbf{H}\mathbf{x}\|^2}{\sigma_n^2}, \quad \forall \mathbf{x} \in \mathbb{X}. \tag{5.6}$$

The same Log-Sum-Exp marginalization defined in (5.5) is applied, with the scalar constellation $\mathcal{A}$ and scalar distance $d(c)$ replaced by the vector lattice $\mathbb{X}$ and the vector distance $d(\mathbf{x})$. The resulting LLRs are information-theoretically optimal, strictly bounding the maximum achievable detection performance for the given MIMO configuration.

3. **K-Best Sphere Decoder Baseline.** For high-order configurations such as 4×4 64-QAM, exhaustive MLD requires evaluating $|\mathbb{X}| = 64^4 \approx 1.68 \times 10^7$ candidate vectors per resource element, rendering it prohibitively slow for large-scale simulation. A K-Best sphere decoder is therefore adopted as a computationally feasible alternative. Given the QR decomposition $\mathbf{H} = \mathbf{Q}\mathbf{R}$ and the projected observation $\tilde{\mathbf{y}} = \mathbf{Q}^H\mathbf{y}$, the search proceeds from layer N_{tx} down to 1, maintaining at most K surviving paths. At layer ℓ, the accumulated partial distance for the p-th surviving path is updated as

$$d_\ell^{(p)} = d_{\ell+1}^{(p)} + \frac{1}{\sigma_n^2}\left\|\tilde{y}_\ell - R_{\ell,\ell}x_\ell^{(p)} - \sum_{j=\ell+1}^{N_{\text{tx}}} R_{\ell,j}x_j^{(p)}\right\|^2, \tag{5.7}$$

where $x_\ell^{(p)} \in \mathcal{A}$. The $K \cdot |\mathcal{A}|$ expanded candidates are pruned to the best K at each layer. The surviving K paths form a reduced candidate lattice $\mathbb{L} = \{\mathbf{x}_1, \ldots, \mathbf{x}_K\}$, and the Log-Sum-Exp operation of (5.5) is applied with $\mathbb{X}$ replaced by $\mathbb{L}$. Under the evaluated 4×4 64-QAM configuration with $K = 256$, the K-Best detector produces detection metrics indistinguishable from those of the exact MLD baseline, thereby establishing the same theoretical upper bound at a fraction of the computational cost.

4. **Genie-Aided ICL Benchmark (Ideal CSI).** To decouple the errors induced by the CEM from the intrinsic equalization capacity of the MDM, we configure a Genie-Aided variant of the proposed architecture. As formulated in Section 3.3, this benchmark bypasses the estimated channel tensor $\hat{\mathbf{H}}_{\text{flat}}$ and directly injects the perfect ground-truth channel $\mathbf{H}$ into the latent residual fusion bridge, isolating the pure equalization gain of the VWN-TPA backbone.

5.2 Gain Analysis of the Iterative Mechanism across Pilot Schemes

This section quantifies the performance improvement delivered by the iterative cascade under 4×4 16-QAM configuration with LDPC coding at rate $2/3$ over an ETU70 fading channel.

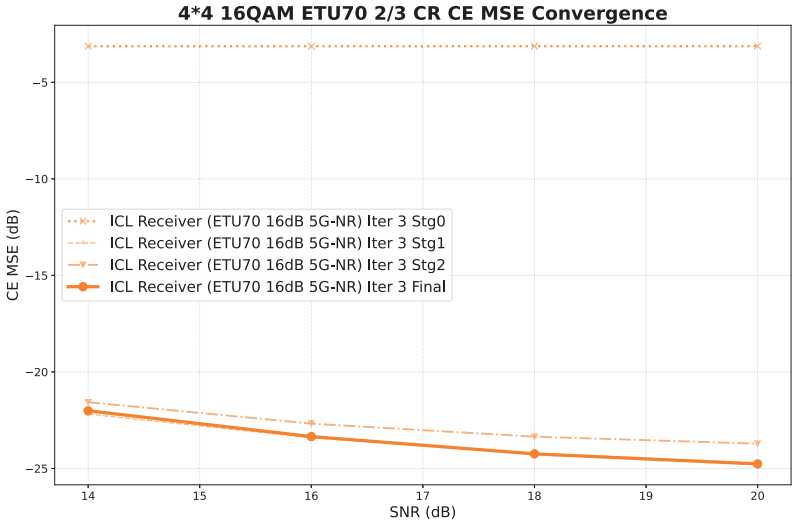
5.2.1 5G-NR DMRS Scheme

Figure 5.1 reports the stage-wise CE loss and BLER under the traditional 5G-NR orthogonal pilot scheme. Although the proposed ICL receiver is optimized at a fixed training SNR of 16 dB, the evaluations demonstrate its exceptional generalization capability across the entire 14–20 dB regime. The BLER curves (Fig. 5.1b) exhibit a waterfall descent as the SNR increases, proving that the architecture effectively learns the underlying physical channel manifold rather than merely overfitting to the noise distribution of the specific training point.

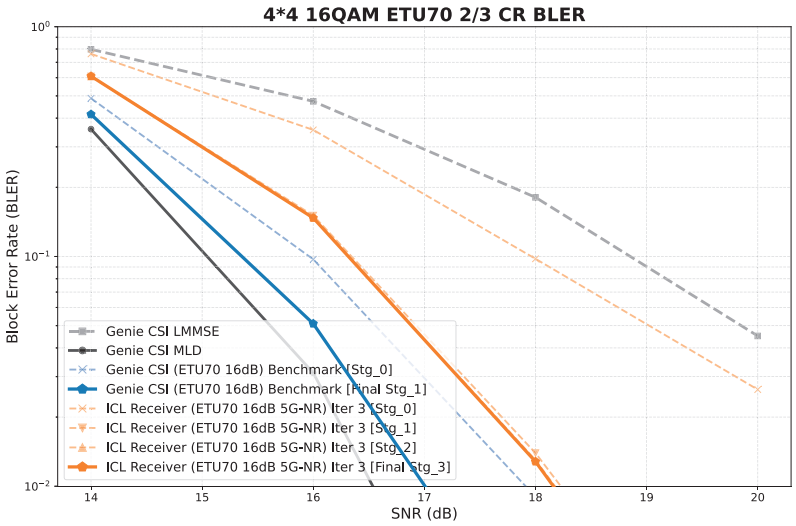
Bootstrap superiority. Even before any iterative refinement, the bootstrap stage (Stg0) already outperforms the unbiased LMMSE baseline equipped with perfect CSI. At a target BLER of 10^{-1}, the SNR gap between Stg0 and LMMSE is approximately 0.85 dB. This confirms that the global contextual representations learned by the Transformer backbone provide a more reliable prior for joint CE and detection than linear filtering alone.

Dominant iterative gain. The most substantial performance leap occurs during the first refinement stage (Stg1). From a PHY perspective, the received signal can be modeled as $\mathbf{y} = (\hat{\mathbf{H}} + \Delta\mathbf{H})\mathbf{s} + \mathbf{n}$. The effective noise variance during detection is thus dominated by the interference term $\mathbb{E}[\|\Delta\mathbf{H}\mathbf{s}\|^2] + \sigma_n^2$, where the CE MSE directly dictates the interference power $\mathbb{E}[\|\Delta\mathbf{H}\|^2]$. As shown in Fig. 5.1a, at an SNR of 16 dB, the initial CE MSE at Stg0 is approximately -3.1 dB. After the first iteration, the CE MSE decreases to -23.4 dB, corresponding to an huge error reduction of more than 20 dB. Since the normalized thermal noise power at 16 dB SNR is -16 dB, the residual CE MSE after Stg1 is about 7.4 dB below the thermal noise power. Thus, the CE-induced interference term is greatly reduced and no longer dominates the effective detection noise. This explains why the first refinement stage provides the largest iterative gain.

This CE MSE reduction translates into corresponding decoding improvements, as shown in Fig. 5.1b. At a target BLER threshold of 10^{-1}, the SNR gap between Stg0 and Stg1 is approximately 1.7 dB. Furthermore, the iterative mechanism also



(a) Stage-wise CE MSE.



(b) BLER vs. SNR across iterative stages.

Figure 5.1: Stage-wise CE MSE and BLER performance of the iterative mechanism under the 5G-NR DMRS scheme (4×4 16-QAM, LDPC rate 2/3, ETU70).

provides equalization benefits independent of the CE process: under the ideal CSI benchmark, the performance gap between Stg0 and Stg1 is approximately 0.7 dB. Comparing the practical joint receiver at Stg1 to the MLD bound, the overall performance gap is reduced to 1.3 dB.

Convergence profile. Subsequent stages (Stg2 and Stg3) yield minimal additional gains, with the BLER curves virtually overlapping. This rapid convergence demonstrates that the vast majority of achievable iterative benefit is extracted within a single refinement block ($N_{\mathrm{it}} = 1$). Consequently, the architecture saturates its performance at Stg1, which limits inference latency without compromising detection accuracy.

5.2.2 SI DMRS Scheme

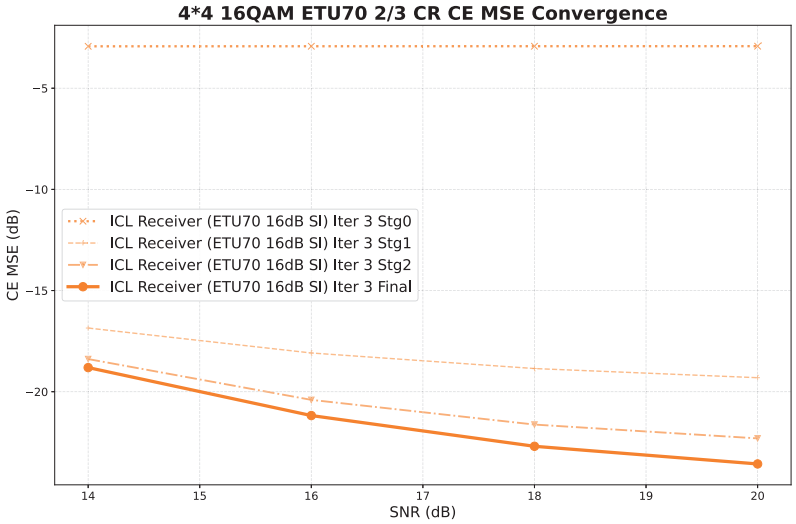
Figure 5.2 presents the corresponding results under the SI DMRS scheme with a pilot power allocation factor of $\rho = 0.8$.

Interference resolution and iterative gain. In the evaluated SI DMRS scheme, the transmit power is split between pilots ($\rho = 0.8$) and data ($1 - \rho = 0.2$). This non-orthogonal superposition introduces a dual challenge: coupled pilot-data interference and reduced power allocation for the data symbols. Consequently, the bootstrap stage (Stg0) experiences performance degradation, failing to reach the 10^{-1} BLER target within the evaluated SNR regime. The iterative mechanism mitigates this limitation. At 16 dB, the initial CE MSE at Stg0 is approximately -2.9 dB. After the first refinement stage (Stg1), the MSE decreases to -18.1 dB, representing an error reduction of over 15 dB. Although the lack of Stg0 convergence prevents the calculation of an initial SNR gap, the BLER curves demonstrate that Stg1 establishes reliable decoding for the SI DMRS scheme.

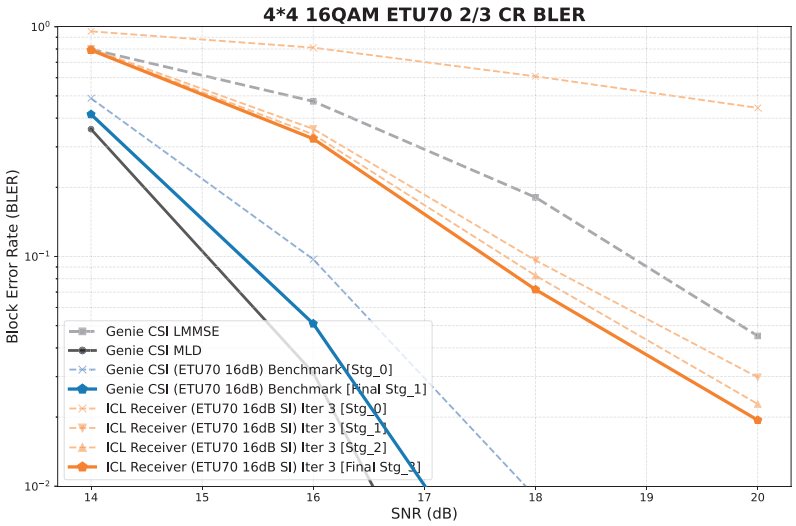
Extended refinement and residual gap. Unlike the 5G-NR scheme that saturates after the first iteration, subsequent stages (Stg2 and Stg3) in the SI DMRS configuration provide continuous refinement to process the coupled interference, yielding an additional decoding gain of approximately 0.5 dB by Stg3. A residual performance gap of 2.5 dB remains between the fully refined Stg3 and the MLD bound. This gap reflects the fundamental Signal-to-Interference-plus-Noise Ratio (SINR) penalty inherent to non-orthogonal transmission. It indicates that while the ICL architecture resolves the self-interference, the overall performance remains mathematically constrained by the reduced data transmit power and the superimposed interference floor.

5.3 Performance Gain of Resampling Strategies

This section evaluates the inference-time performance gain delivered by the Resampling strategies introduced in Section 4.2.3. The evaluation is conducted under a 4×4 64-QAM configuration with LDPC coding at rate $4/5$ over an ETU70 fading channel. Detailed numerical analysis is performed on the Genie-Aided ideal-CSI benchmark to isolate the resampling gain from CE errors. The same strategy is also validated on the practical joint CE and MIMO detection ICL receiver.



(a) Stage-wise CE MSE.



(b) BLER vs. SNR across iterative stages.

Figure 5.2: Stage-wise CE MSE and BLER performance of the iterative mechanism under the SI DMRS scheme (4×4 16-QAM, $\rho = 0.8$, LDPC rate 2/3, ETU70).

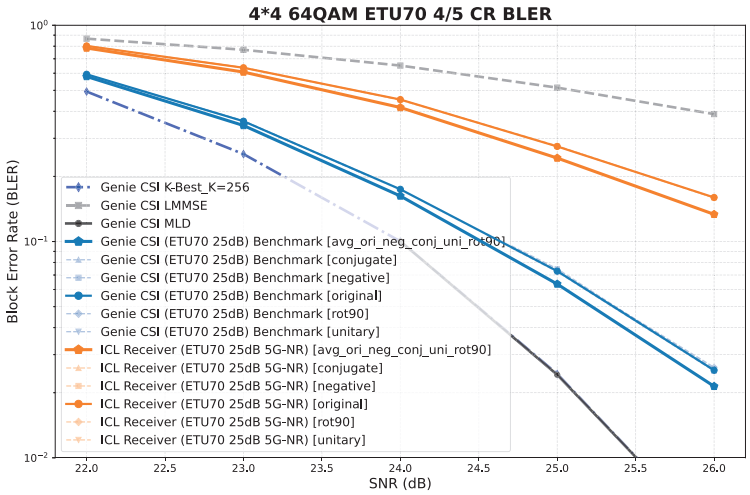


Figure 5.3: BLER performance of resampling strategies compared to the original baseline (4×4 64-QAM, LDPC rate 4/5, ETU70).

Figure 5.3 reports the BLER performance of the original baseline alongside the distinct resampling configurations. Consistent with the mathematical equivalence of the applied transformations, the individual symmetric branches (i.e., negative, conjugate, unitary, and rot90) yield BLER curves that closely align with the baseline prediction, for both the Genie-Aided benchmark and the CE and MIMO detection receiver. This performance indicates that the end-to-end AI receiver maintains structural equivariance with respect to the fundamental symmetries of the MIMO channel.

However, the non-linear nature of the neural parameterization introduces localized approximation variances. This epistemic uncertainty is most pronounced near the dense decision boundaries of the 64-QAM constellation, where the Euclidean distance between adjacent symbols is minimal. When querying the model with symmetrically transformed inputs, these boundary-level prediction variances manifest in statistically independent directions. By aggregating these multiple views in the logit domain, the multi-branch ensemble effectively suppresses this residual uncertainty, thereby translating into a net decoding gain.

Training-free ensemble gain. Table 5.3 summarizes the decoding performance of the ensemble configurations. As shown in the lower section, while these resampling operations are mathematically isomorphic and would yield identical outputs in classical algorithms, they present numerically distinct input tensors to the AI receiver. The non-linear parameterization of the network reacts to these distinct tensors with independent prediction variances. For example, processing the conjugate tensor introduces 77 new block errors but simultaneously corrects 76 distinct blocks that the baseline failed to decode. Logit aggregation leverages this independence to average out the epistemic uncertainty and retain correct decisions. Consequently, combining all five branches maximizes this correction

capability, reducing the BLER from 6.98×10^{-2} to 6.21×10^{-2} at 25 dB.

Table 5.3: Performance leaderboard of spatial ensemble combinations via logit aggregation (25 dB, 4×4 64-QAM, LDPC rate 4/5, ETU70).

Configuration	Corrected (+)	Introduced (-)	Net Gain	BLER
Ensemble Combinations				
O+N+C+U+R	159	5	**154**	$\mathbf{6.21 \times 10^{-2}}$
O+N+C+R	158	8	150	6.23×10^{-2}
O+N+C+U	140	4	136	6.30×10^{-2}
O+C+U+R	141	6	135	6.30×10^{-2}
O+N+R	139	4	135	6.30×10^{-2}
O+N+U+R	139	5	134	6.31×10^{-2}
O+C+R	137	7	130	6.33×10^{-2}
O+C+U	122	3	119	6.38×10^{-2}
O+U+R	120	3	117	6.39×10^{-2}
O+C	120	8	112	6.42×10^{-2}
O+R	113	7	106	6.45×10^{-2}
O+N+U	106	9	97	6.49×10^{-2}
O+N	96	15	81	6.57×10^{-2}
O+U	39	13	26	6.85×10^{-2}
Standalone Transformations				
Baseline (O)	0	0	0	6.98×10^{-2}
N	73	70	3	6.97×10^{-2}
C	76	77	-1	6.99×10^{-2}
U	41	39	2	6.97×10^{-2}
R	76	70	6	6.95×10^{-2}

O = original, N = negative, C = conjugate, U = unitary, R = rot90.

Error correction dynamics. The exceptionally high ratio of corrected to introduced errors confirms that the ensemble reliably suppresses epistemic uncertainty without destabilizing already well-classified symbols. A deep symbol-level diagnostic analysis reveals that over 56% of these corrections are concentrated in the two least significant bits (LSBs, i.e., b_4 and b_5). Under the standard 64-QAM Gray mapping, these LSBs dictate the finest granularity of the symbol's spatial position, effectively distinguishing between the closest adjacent constellation points. Consequently, they possess the densest decision boundaries and the weakest Euclidean distance protection, making them inherently the most vulnerable to minute spatial perturbations. By averaging the logits across symmetric views, the ensemble effectively smooths the decision boundaries for these highly ambiguous adjacent points.

Approaching theoretical limits. The absolute gain achieved by resampling reflects the fact that the baseline VWN-TPA model is already highly converged. With the full invariant-transformation ensemble, the performance gap to the MLD bound is narrowed to approximately 0.5 dB. This confirms that the proposed architecture, coupled with physical symmetry constraints, achieves near-optimal

equalization capacity. Furthermore, this entire gain is obtained strictly at inference time without any additional training or parameterized modules, making resampling a highly practical enhancement for scenarios where maximizing BLER reliability at high SNR justifies the supplementary forward passes.

5.4 Impact of Pilot Sparsity on Traditional 5G-NR and ICL Pilot Schemes

This section investigates the trade-off between CE accuracy and throughput under varying degrees of pilot sparsity. We benchmark the proposed ICL pilot scheme against the traditional 5G-NR orthogonal allocations across different spatial dimensions and multipath delay profiles.

5.4.1 Pilot Pattern Configurations and Overhead

As illustrated in Fig. 2.2, we define a spectrum of pilot configurations ranging from I1 (densest) to I12 (sparsest) to evaluate the system under varying constraints of total pilot overhead. The plotted allocations correspond to the implemented patterns used in the experiments. For each sparsity level, the comparison is matched by pilot overhead rather than by identical time-frequency coordinates. The main distinction between the two evaluated schemes is their spatial allocation. Traditional 5G-NR schemes assign disjoint pilot REs across transmit antennas, whereas the proposed ICL scheme allows all transmit antennas to simultaneously transmit over the same active pilot REs. This shifts the PHY challenge from strict resource isolation to interference decoupling. The 2×2 configurations follow the same sparsity scaling logic, with the traditional 5G-NR pilot REs orthogonally distributed across two spatial streams.

5.4.2 Evaluation on 4×4 16-QAM ETU70 Configuration

The primary evaluation establishes the system's baseline performance under the ETU70 propagation environment. The ICL receivers share the same architecture but are trained separately under different pilot schemes. All models are trained on the ETU70 channel at an SNR of 16 dB. During evaluation, the LDPC code rate is fixed to $2/3$. The resulting CE MSE, BLER, and throughput are presented in Figs. 5.4, 5.5, and 5.6, respectively.

Interference decoupling under reduced overhead. The evaluation demonstrates the capability of the ICL receiver to resolve non-orthogonal pilot superpositions. Both the I2 ICL and I2 5G-NR configurations reduce the pilot overhead by 50% relative to the densest I1 pattern. While the 5G-NR scheme avoids interference by maintaining orthogonal resource grids, the ICL scheme superimposes the pilots, which fundamentally introduces mutual interference. However, the model effectively decouples these overlapping signals, resulting in only a marginal penalty in CE MSE. Consequently, the decoding performance of I2 ICL remains highly competitive, and the throughput gap between the non-orthogonal ICL and

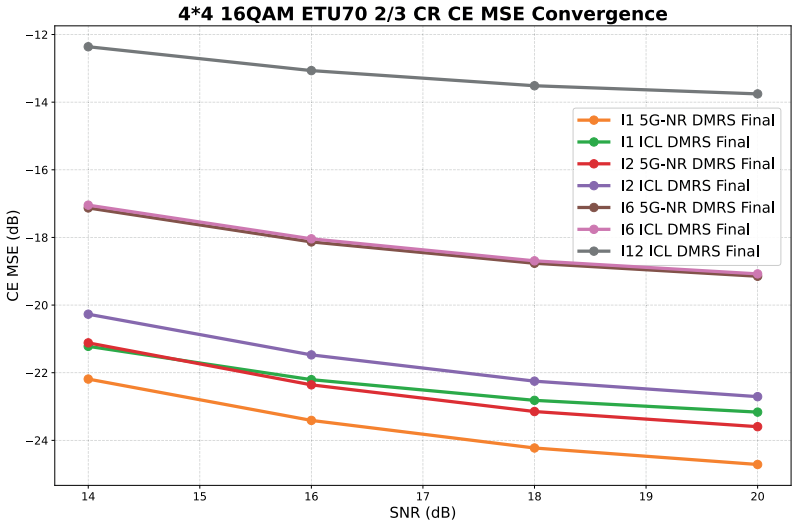


Figure 5.4: CE MSE vs. SNR under the 4×4 16-QAM ETU70 configuration.

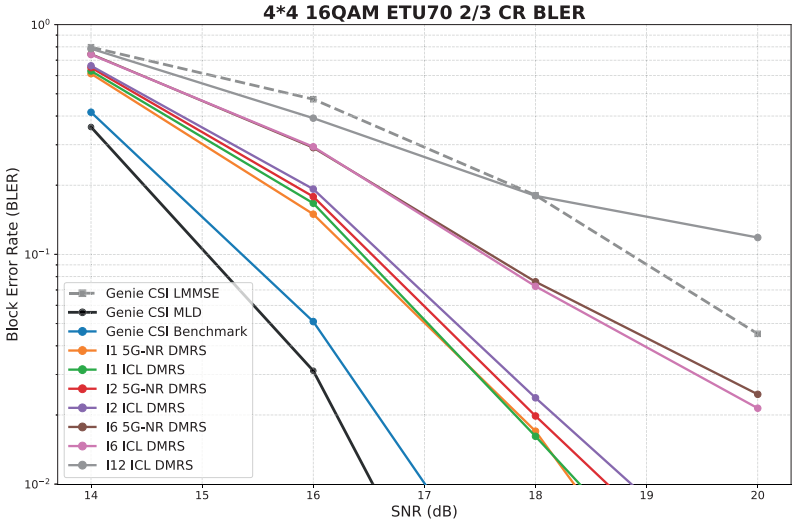


Figure 5.5: BLER vs. SNR under the 4×4 16-QAM ETU70 configuration.

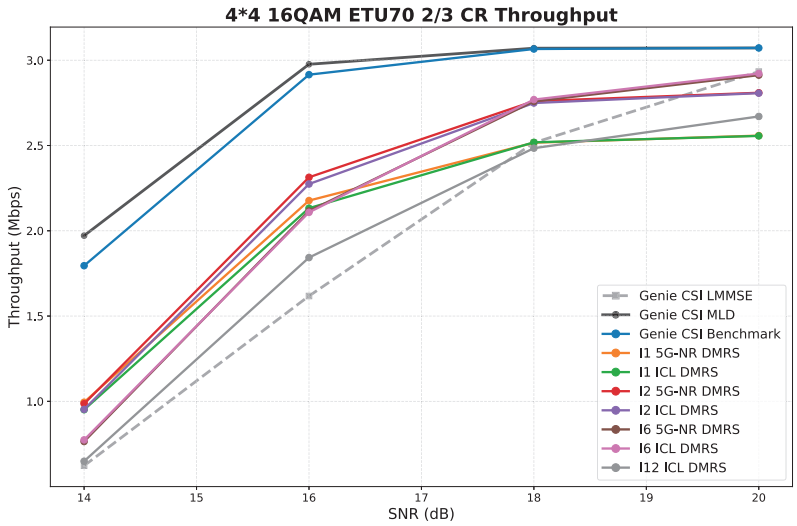


Figure 5.6: Throughput vs. SNR under the 4×4 16-QAM ETU70 configuration.

the orthogonal 5G-NR is minimal. This confirms the model's robust interference cancellation capability under constrained resource allocations.

Sparsity scaling and throughput crossover. As pilot density decreases further to the I6 configuration, the CE MSE degrades to approximately -18.0 dB at 16 dB SNR, causing a visible decoding penalty. In the low-to-medium SNR regime (14–18 dB), this decoding degradation dominates, causing the throughput of I6 to strictly lag behind the denser I1 and I2 configurations. However, as the channel conditions improve beyond 18 dB, the high spectral efficiency resulting from the drastic reduction in pilot overhead begins to dictate system performance. As depicted in Fig. 5.6, the throughput curve of I6 exhibits a clear crossover point around 18 dB, eventually surpassing the I2 configurations and demonstrating a higher capacity ceiling in the high-SNR regime.

Extreme sparsity vs. ideal CSI baseline. The performance of the extremely sparse I12 configuration highlights the robustness of the non-linear AI architecture. While I12 incurs a substantial absolute BLER penalty compared to denser pilot patterns, the I12 ICL receiver continues to outperform the mathematical Genie-Aided LMMSE baseline (which operates with perfect, ideal CSI) in the low-to-medium SNR regime. This indicates that the global contextual representations learned by the Transformer backbone can effectively mitigate severe channel undersampling. Even with minimal and non-orthogonal pilot symbols, the data-driven model extracts sufficient structural information to achieve better equalization than traditional linear methods equipped with ideal channel knowledge.

5.4.3 Evaluation on 2×2 16-QAM ETU70 Configuration

To verify the spatial scalability of the non-orthogonal representations, the evaluation is extended to a lower-dimensional MIMO regime. Compared to the 4×4 setup, the 2×2 channel matrix has fewer spatial streams, which reduces the estimation and interpolation burden on the receiver. The ICL receivers share the same architecture but are trained separately under different pilot schemes. All models in this configuration are trained on the ETU70 channel at an SNR of 14 dB. During evaluation, the LDPC code rate is fixed to $2/3$. The corresponding performance metrics are reported in Figs. 5.7, 5.8, and 5.9.

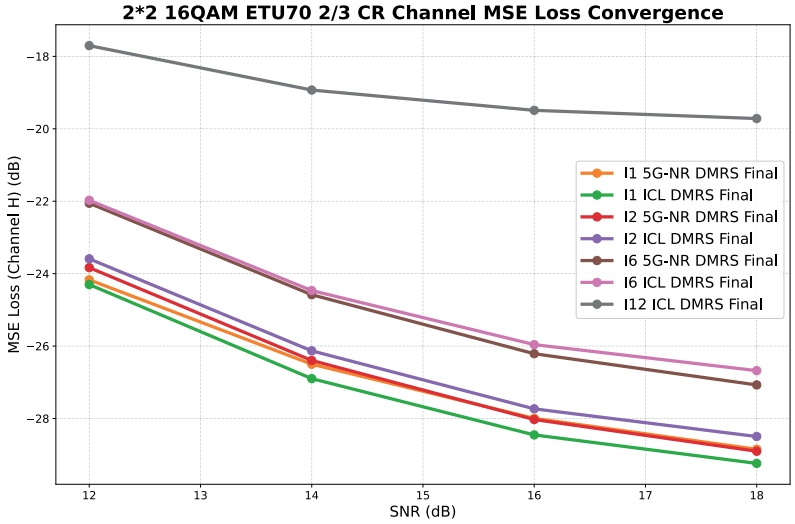


Figure 5.7: CE MSE vs. SNR under the 2×2 16-QAM ETU70 configuration.

High tolerance to sparsity scaling (I1 to I6). It shows that the 2×2 configuration exhibits resilience to pilot reduction. From the densest I1 pattern down to the highly sparse I6 pattern, the CE MSE remains significantly small relative to the noise floor. Across these configurations, the non-orthogonal ICL scheme closely tracks the decoding performance of the orthogonal 5G-NR baselines, confirming that the interference decoupling mechanism operates reliably in lower-dimensional spatial multiplexing.

Throughput scaling and operational limits. Because the decoding penalty introduced by the I6 configuration is small in the 2×2 scenario, the SE gains translate into system performance improvements at lower SNRs. Unlike the 4×4 setup, where the throughput of I6 surpasses denser configurations only in the high-SNR regime, the 2×2 I6 configuration maintains a higher throughput across a broader SNR range. It achieves improved SE without a substantial trade-off in link reliability. At the highest evaluated sparsity level (I12), the degradation in CE MSE becomes prominent. While this translates into a BLER penalty, the ICL receiver maintains stable convergence, defining the operational boundary for pilot

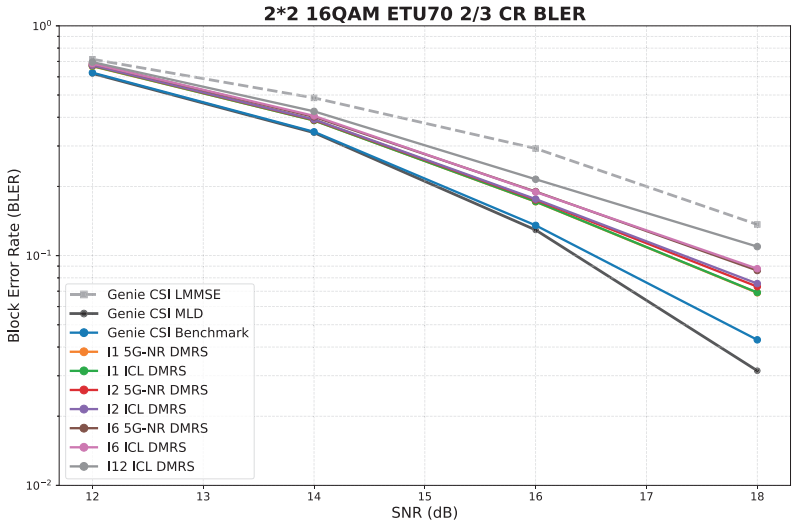


Figure 5.8: BLER vs. SNR under the 2×2 16-QAM ETU70 configuration.

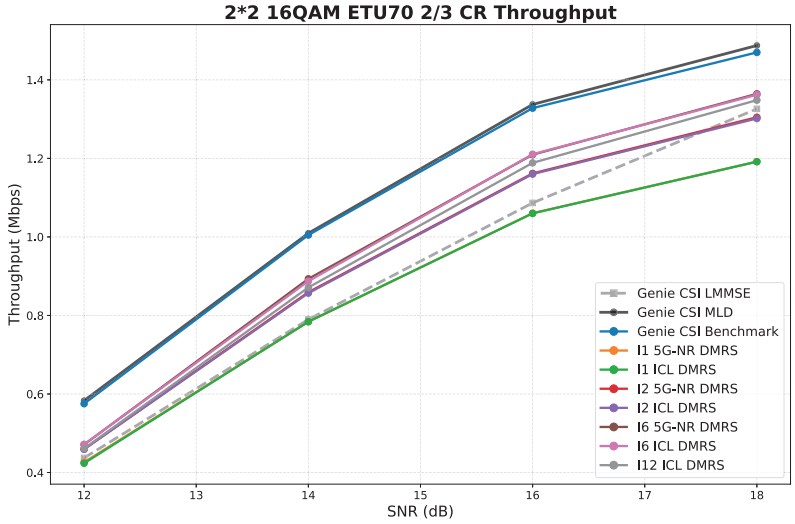


Figure 5.9: Throughput vs. SNR under the 2×2 16-QAM ETU70 configuration.

compression in this low-dimensional MIMO setup.

5.4.4 Evaluation on 4×4 16-QAM EPA5 Configuration

To verify the robustness of the sparse ICL patterns across different channel models, the system is further evaluated over the EPA5 channel. The corresponding models are trained at a fixed SNR of 16 dB. The resulting performance metrics are shown in Figs. 5.10, 5.11, and 5.12.

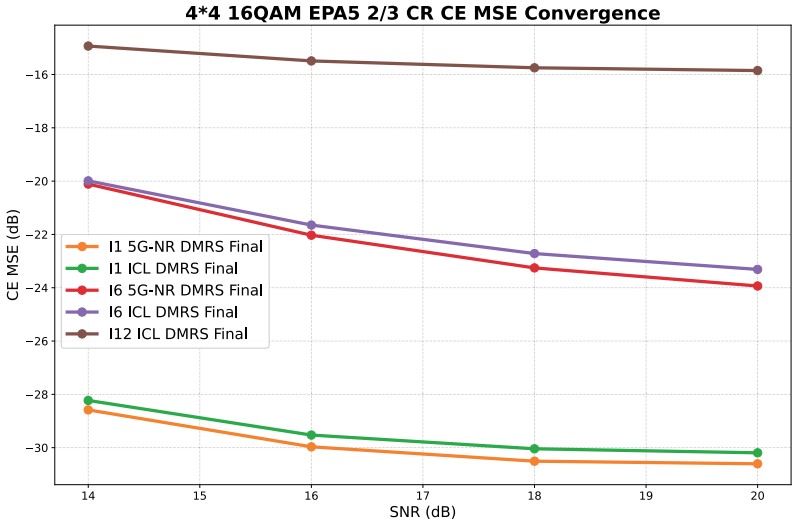


Figure 5.10: CE MSE vs. SNR under the 4×4 16-QAM EPA5 configuration.

The observations from the ETU70 evaluation are preserved under the EPA5 channel, with consistent trends across the I1, I6, and I12 configurations. The primary difference is the absolute CE MSE, which is lower under EPA5 due to reduced frequency selectivity. This reduction translates into improved BLER across all pilot densities, while the relative performance remains unchanged: the ICL receiver aligns with the 5G-NR baseline, the I6 throughput crossover persists, and the I12 configuration incurs a BLER penalty while still surpassing the ideal-CSI LMMSE baseline. These results indicate that the interference decoupling capability of the network generalizes across different multipath profiles.

5.5 Performance Characterization of SI DMRS

This section evaluates the SI DMRS scheme against the 5G-NR baseline. Performance is characterized in terms of CE accuracy and throughput across varying SNRs and code rates. For all SI DMRS configurations, the pilot pattern adopts the dense I1 layout (Fig. 3.2), with all transmit antennas superimposed over the shared REs. The power allocation factor is set to $\rho = 0.8$, assigning a larger proportion

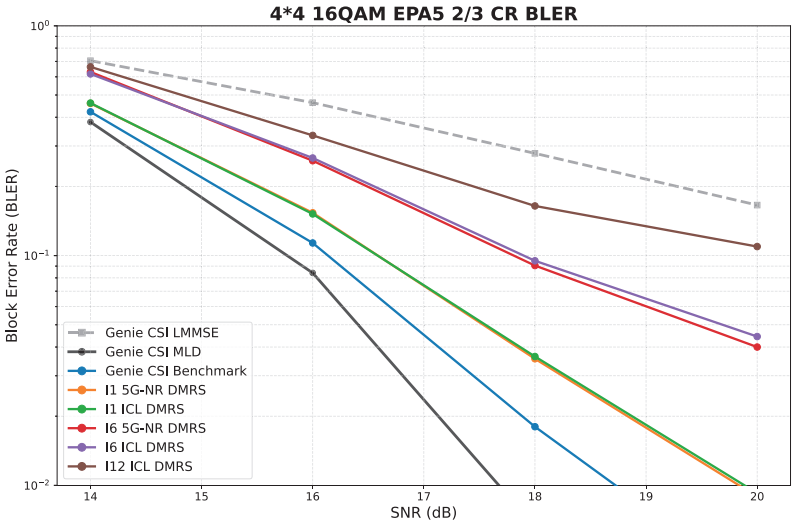


Figure 5.11: BLER vs. SNR under the 4×4 16-QAM EPA5 configuration.

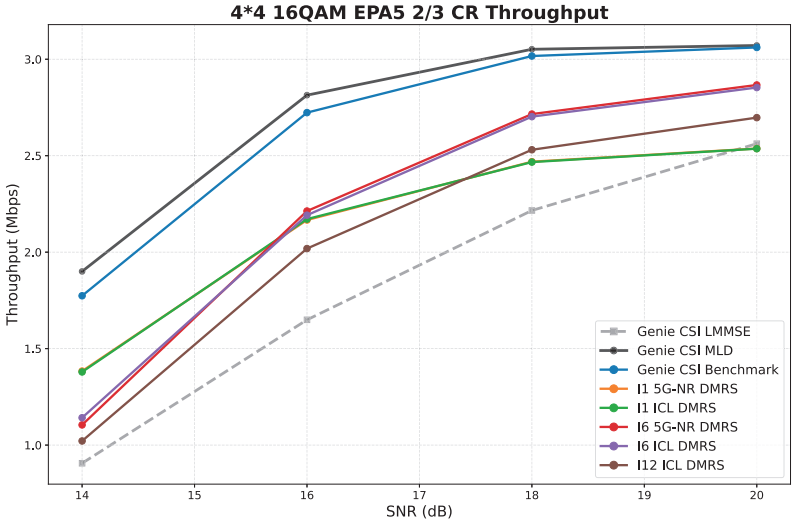


Figure 5.12: Throughput vs. SNR under the 4×4 16-QAM EPA5 configuration.

of the transmit power to pilot symbols. The evaluation is conducted over the ETU70 channel, using multiple LDPC code rates (1/3, 1/2, and 2/3) to construct the performance envelopes.

5.5.1 Evaluation on 2×2 16-QAM Configuration

We first evaluate the SI DMRS scheme under the 2×2 spatial multiplexing regime to establish its viability. The corresponding CE MSE and throughput envelopes are presented in Figs. 5.13 and 5.14.

To characterize the performance bounds, these results are plotted as envelopes. Although the ICL receiver trained at a fixed SNR can generalize across varying noise levels, the 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 in the power domain. The presented envelopes trace the maximum achievable performance across these SNR-specific models and multiple code rates.

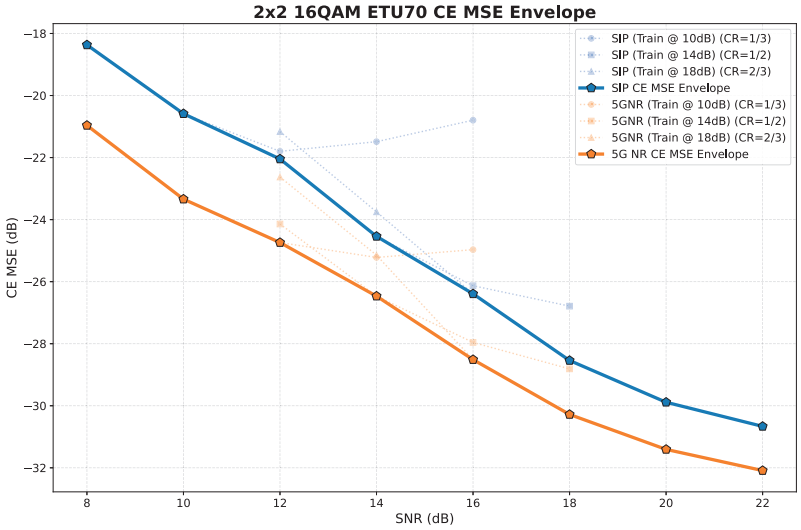


Figure 5.13: CE MSE vs. SNR under the 2×2 16-QAM ETU70 configuration across multiple coding rates.

Power splitting and throughput scaling. By abandoning orthogonal isolation, the SI DMRS scheme introduces pilot-to-data interference and power splitting, which reduces the transmit energy for both CE and data detection. As shown in Fig. 5.13, this dual penalty leads to a CE MSE degradation compared to the 5G-NR baseline. However, in the 2×2 setup, the CE error remains bounded relative to the operating SNR. While the combined effect of the estimation penalty and reduced data power causes the receiver to perform lower than the LMMSE baseline in the low-SNR regime, the interference is effectively resolved by the AI architecture. Because the PHY utilizes all resource elements for data transmission, SI DMRS yields a higher throughput than 5G-NR. As depicted in Fig. 5.14, the SI DMRS

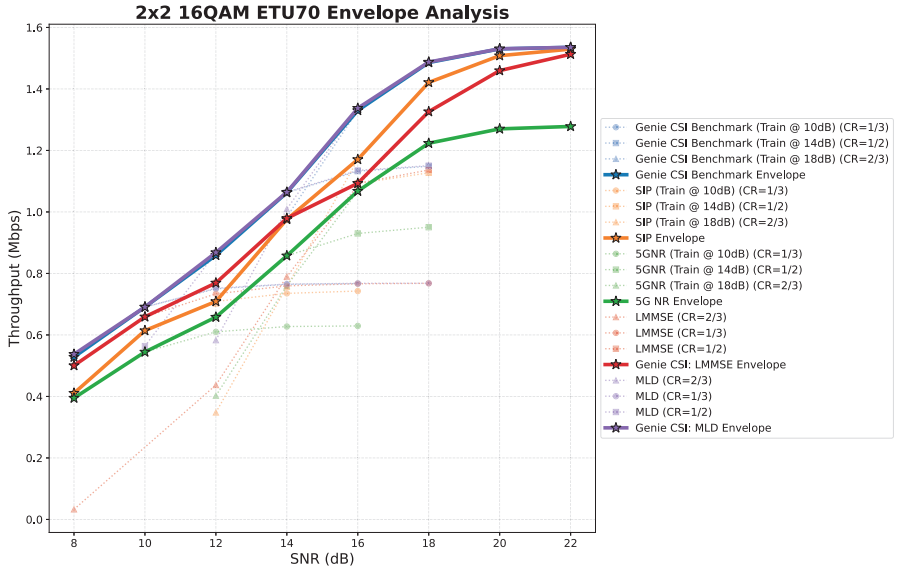


Figure 5.14: Throughput Envelope vs. SNR under the 2×2 16-QAM ETU70 configuration across multiple coding rates.

throughput envelope converges to a higher capacity limit, capitalizing on its zero time-frequency pilot overhead.

5.5.2 Evaluation on 4×4 16-QAM Configuration

The evaluation is extended to the 4×4 spatial multiplexing configuration. While this higher-order setup increases the density of coupled spatial interference, it scales the overhead reduction achieved by the SI DMRS.

Severe estimation penalty and code-rate-dependent resilience. As depicted in Fig. 5.15, the superposition of four independent spatial streams severely contaminates the initial channel observations. Unlike the 2×2 scenario, the CE MSE loss relative to the operating SNR is substantial. At high SNR levels, the CE MSE degrades to a point where it is roughly equivalent to the SNR, indicating a interference-limited estimation phase. This compounded error translates into a general SNR penalty of approximately 1.5 dB at a target BLER of 10^{-1} compared to the orthogonal 5G-NR baseline.

The corresponding BLER evaluations (Figs. 5.16 to 5.18) reveal a code-rate-dependent decoding resilience. In the low-SNR, low-rate regime (Fig. 5.16), the severe interference prevents the SI DMRS model from matching the ideal Genie-CSI LMMSE baseline. As the code rate increases to 1/2 (Fig. 5.17), the orthogonal 5G-NR ICL receiver successfully surpasses the LMMSE, whereas the SI DMRS scheme still struggles to close the gap. However, under high-SNR and high-rate conditions (code rate=2/3, Fig. 5.18), the SI DMRS receiver not only stabilizes but surpasses the ideal Genie-Aided LMMSE baseline, proving its capacity to unravel

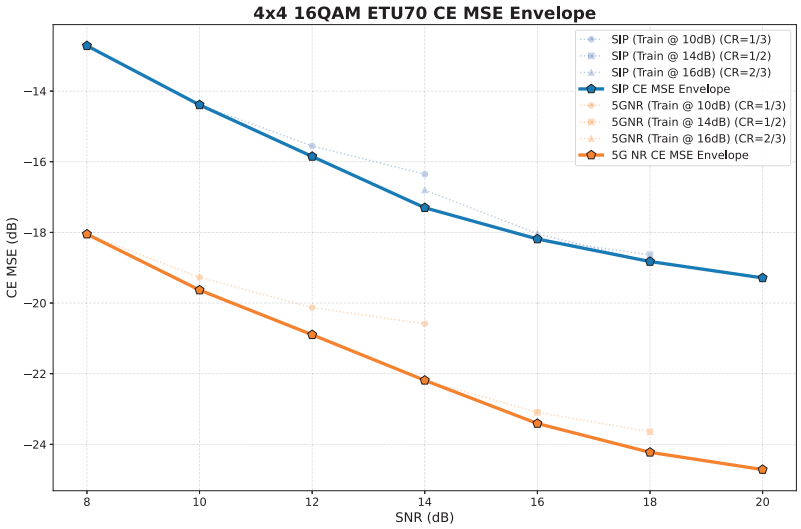


Figure 5.15: CE MSE vs. SNR under the 4×4 16-QAM ETU70 configuration across multiple coding rates.

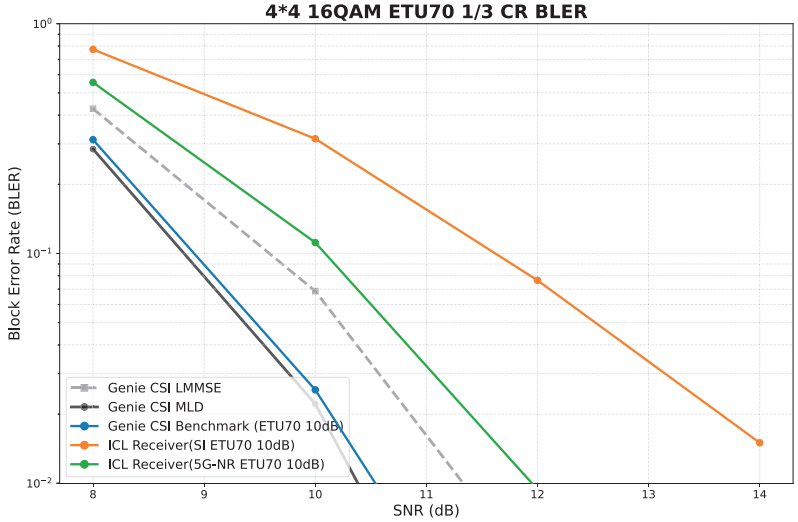


Figure 5.16: BLER vs. SNR under the 4×4 16-QAM ETU70 configuration with LDPC code rate 1/3.

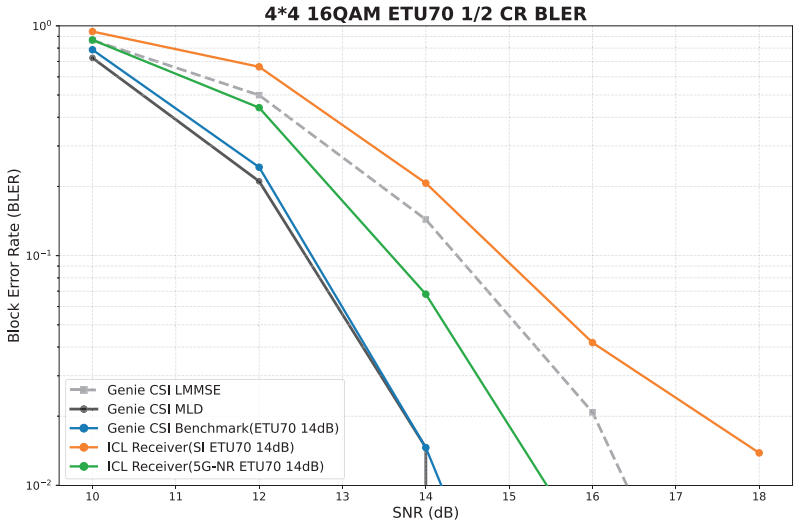


Figure 5.17: BLER vs. SNR under the 4×4 16-QAM ETU70 configuration with LDPC code rate 1/2.

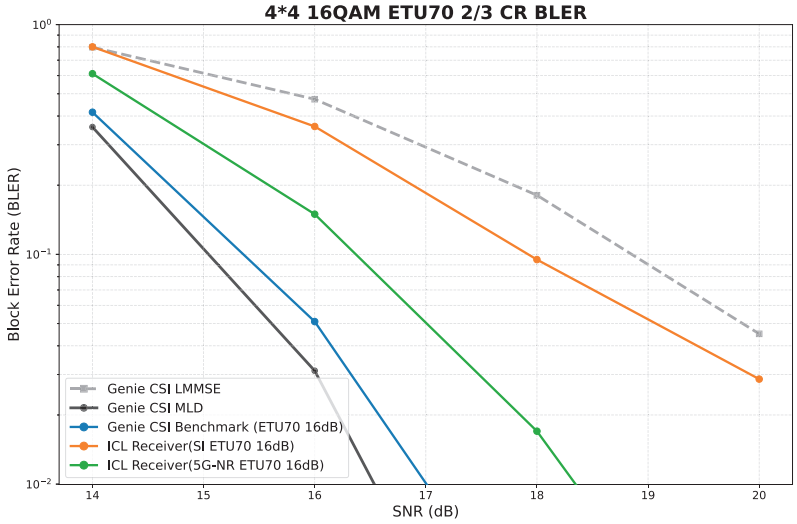


Figure 5.18: Throughput envelope vs. SNR under the 4×4 16-QAM ETU70 configuration across multiple coding rates.

spatial aliasing when sufficient signal power is available.

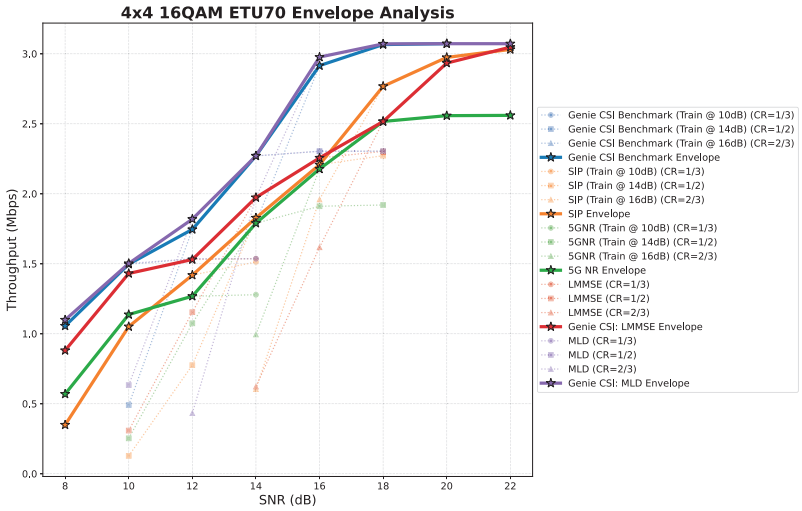


Figure 5.19: Throughput Envelope vs. SNR for the 4×4 16-QAM ETU70 configuration across multiple coding rates.

Throughput trade-offs and capacity limits. The performance is characterized by the throughput envelope in Fig. 5.19. Due to a larger BLER penalty (~ 1.5 dB) compared to the 2×2 configuration, the throughput gain over the 5G-NR baseline requires a higher SNR to manifest. In the low-SNR regime, the SI DMRS throughput is lower than that of the orthogonal schemes. However, as the operating point transitions to the medium-to-high SNR regime and the network mitigates the spatial interference, the increased data RE allocation outweighs the detection penalty, yielding a net throughput improvement. Furthermore, in the high-SNR regime with the highest code rate (2/3), the SI DMRS converges to a higher capacity limit. By eliminating the time-frequency pilot overhead, it achieves a peak SE exceeding that of conventional orthogonal allocations.

Chapter 6
Conclusion and Future Work

6.1 Conclusion

This thesis investigated the limitations of pilot overhead and rigid spatial orthogonality in MIMO-OFDM systems, proposing a Physics-Guided ICL receiver to address these capacity bottlenecks. By unifying classical communication-theoretic models with a Transformer-based architecture incorporating VWN, TPA, and MoE, the proposed framework successfully reformulates CE and symbol detection as a joint contextual learning process.

The evaluations demonstrated that the proposed architecture achieves detection performance approaching the MLD bound under baseline configurations. The comparison between the orthogonal 5G-NR pilot scheme and the non-orthogonal ICL pilot scheme reveals two main properties of the receiver. First, with the PGFC bridge and the iterative unfolded cascade, the receiver is less sensitive to pilot sparsity. When the pilot density is reduced, the initial channel estimate becomes less accurate, but the joint CE-detection process can partly compensate for this loss during data detection. Second, the ICL receiver does not require spatially isolated pilots. In the evaluated setting, the non-orthogonal ICL pilot scheme achieves comparable performance to the orthogonal 5G-NR scheme, suggesting that overlapping reference signals can be used as contextual information by the model. In addition, the inference-time Invariant-Transformation Resampling strategy improves the high-SNR performance without retraining the network, suggesting that the ICL receiver maintains consistency across mathematically equivalent input-output representations.

The evaluation of the SI DMRS paradigm demonstrates the viability of the proposed ICL receiver in improving SE. By superimposing pilot and data symbols, the scheme requires splitting the transmit power. This introduces a dual physical challenge: coupled spatial self-interference and reduced effective transmit energy for both CE and data detection. In lower-order spatial multiplexing configurations, the CE error remains bounded, and the elimination of pilot overhead translates into direct throughput improvements over dense 5G-NR schemes. In higher-order configurations, the expanded channel dimension exacerbates both the estimation error and the power-splitting penalty. Nevertheless, under some SNR and code rate conditions, the zero time-frequency pilot overhead of SI DMRS provides sufficient SE gain to overcome these constraints, leading to throughput

improvement.

Overall, this research demonstrates that integrating physical priors with deep learning enables robust joint CE and MIMO detection under flexible pilot configurations, providing a spectrally efficient receiver design that overcomes the overhead and orthogonality limitations of classical MIMO systems.

6.2 Limitations and Future Work

While the proposed ICL receiver demonstrates robust interference decoupling and near-optimal equalization, several architectural limitations and open challenges remain.

Scalability to Massive MIMO. The current evaluation validates the architecture up to a 4×4 spatial multiplexing order, where both orthogonal 5G-NR and non-orthogonal ICL schemes maintain comparable performance. However, as the system scales to Massive MIMO regimes, the requirement for spatial orthogonality exhausts available pilot resources, bottlenecking payload capacity. While the non-orthogonal ICL scheme resolves this overhead limitation by sharing pilot coordinates, scaling the current Transformer backbone to massive arrays introduces a dimensionality challenge due to the quadratic complexity of global self-attention. A critical future direction is to integrate linear-complexity or structured sparse attention mechanisms to efficiently process high-dimensional spatial streams without prohibitive computational overhead.

End-to-end optimization of pilot allocation. In this study, the ICL receiver is trained and evaluated using fixed, predefined pilot patterns (e.g., I1 to I12 configurations). An extension is to integrate the pilot placement into the end-to-end learning process. By formulating the pilot mask as a differentiable parameter during training, the network could autonomously discover optimal time-frequency grids. This data-driven approach would allow the system to adapt pilot positions to specific multipath delay profiles or SNR regimes, potentially yielding higher throughput than static human-designed configurations.

Data-driven discovery of physical features. The current PGFC module computes a fixed set of communication-theoretic features: IC residuals, MF responses, and Gram matrix entries. These features are derived from classical signal processing primitives and are sufficient for linear and mildly non-linear interference cancellation. A future direction is to investigate whether the set of physical features can itself be meta-optimized during training, either by making the feature selection differentiable or by introducing a learned feature generation head that complements the hand-crafted PGFC tensors. This would allow the network to move beyond classical linear-algebraic primitives and autonomously discover higher-order physical structures that are not easily expressible through fixed analytical formulas, further narrowing the gap between model-driven and data-driven receiver design.

Chapter 7

Comments

References

[1] T. O'Shea and J. Hoydis, "An introduction to deep learning for the physical layer," *IEEE Transactions on Cognitive Communications and Networking*, vol. 3, no. 4, pp. 563–575, 2017.

[2] X. Lin, *An overview of the 3GPP study on artificial intelligence for 5G New Radio*, 2023. arXiv: 2308.05315 [cs.NI].

[3] X. Qin, S. Hu, J. Zhang, J. Qian, and H. Wang, "AI receiver design with deep learning based channel estimation and MIMO detection," in *2024 IEEE 35th International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, 2024, pp. 1–7.

[4] S. Cammerer, F. A. Aoudia, J. Hoydis, A. Oeldemann, A. Roessler, T. Mayer, and A. Keller, *A neural receiver for 5G NR multi-user MIMO*, 2023. arXiv: 2312.02601 [cs.IT].

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

[6] D. Wubben, R. Bohnke, V. Kuhn, and K.-D. Kammeyer, "MMSE-based lattice-reduction for near-ML detection of MIMO systems," in *ITG Workshop on Smart Antennas (IEEE Cat. No.04EX802)*, 2004, pp. 106–113.

[7] D. Wubben, R. Bohnke, V. Kuhn, and K.-D. Kammeyer, "MMSE extension of V-BLAST based on sorted QR decomposition," in *2003 IEEE 58th Vehicular Technology Conference. VTC 2003-Fall (IEEE Cat. No.03CH37484)*, vol. 1, 2003, 508–512 Vol.1.

[8] A. Yang et al., *Qwen3 technical report*, 2025. arXiv: 2505.09388 [cs.CL].

[9] DeepSeek-AI et al., *DeepSeek-V3.2: Pushing the frontier of open large language models*, 2025. arXiv: 2512.02556 [cs.CL].

[10] K. Team et al., *Kimi K2.5: Visual agentic intelligence*, 2026. arXiv: 2602.02276 [cs.CL].

[11] N. Samuel, T. Diskin, and A. Wiesel, "Learning to detect," *IEEE Transactions on Signal Processing*, vol. 67, no. 10, pp. 2554–2564, May 2019.

[12] N. Samuel, T. Diskin, and A. Wiesel, *Deep MIMO detection*, 2017. arXiv: 1706.01151 [stat.ML].

[13] Z. Qin, H. Ye, G. Y. Li, and B.-H. F. Juang, *Deep learning in physical layer communications*, 2019. arXiv: 1807.11713 [cs.IT].

[14] 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 *2024 IEEE 100th Vehicular Technology Conference (VTC2024-Fall)*, 2024, pp. 1–7.

[15] 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*, 2025. arXiv: 2506.20248 [eess.SP].

[16] 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*, 2025. arXiv: 2507.10074 [cs.IT].

[17] 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 *2025 IEEE Wireless Communications and Networking Conference (WCNC)*, 2025, pp. 1–6.

[18] X. Qin and S. Hu, *Dual-attention based 3D channel estimation*, 2026. arXiv: 2604.01769 [cs.LG].

[19] M. Zecchin, K. Yu, and O. Simeone, *In-context learning for MIMO equalization using transformer-based sequence models*, 2024. arXiv: 2311.06101 [cs.IT].

[20] ETSI, "5G; NR; physical channels and modulation," European Telecommunications Standards Institute (ETSI), TS 138 211 V18.6.0, Apr. 2025.

[21] Seed et al., *Virtual width networks*, 2025. arXiv: 2511.11238 [cs.LG].

[22] Y. Zhang, Y. Liu, H. Yuan, Z. Qin, Y. Yuan, Q. Gu, and A. C.-C. Yao, *Tensor product attention is all you need*, 2026. arXiv: 2501.06425 [cs.CL].

[23] D. Dai, C. Deng, C. Zhao, R. X. Xu, H. Gao, D. Chen, J. Li, W. Zeng, X. Yu, Y. Wu, Z. Xie, Y. K. Li, P. Huang, F. Luo, C. Ruan, Z. Sui, and W. Liang, *DeepSeekMoE: Towards ultimate expert specialization in mixture-of-experts language models*, 2024. arXiv: 2401.06066 [cs.CL].

[24] J. Su, M. Ahmed, Y. Lu, S. Pan, W. Bo, and Y. Liu, "Roformer: Enhanced transformer with rotary position embedding," *Neurocomputing*, vol. 568, p. 127 063, 2024.

[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, *Attention is all you need*, 2023. arXiv: 1706.03762 [cs.CL].

[26] N. Shazeer, *GLU variants improve transformer*, 2020. arXiv: 2002.05202 [cs.LG].

[27] S. Hu, *Invariant transformation and resampling based epistemic-uncertainty reduction*, 2026. arXiv: 2602.23315 [cs.AI].

[28] 3GPP, "Evolved universal terrestrial radio access (E-UTRA); user equipment (UE) radio transmission and reception," 3rd Generation Partnership Project (3GPP), TS 36.101, 2023.

[29] K. Jordan, Y. Jin, V. Boza, J. You, F. Cesista, L. Newhouse, and J. Bernstein, *Muon: An optimizer for hidden layers in neural networks*, [Online]. Available: https://kellerjordan.github.io/posts/muon/, 2024.

[30] I. Loshchilov and F. Hutter, *Decoupled weight decay regularization*, 2019. arXiv: 1711.05101 [cs.LG].

[31] P. Goyal, P. Dollár, R. Girshick, P. Noordhuis, L. Wesolowski, A. Kyrola, A. Tulloch, Y. Jia, and K. He, *Accurate, large minibatch SGD: Training ImageNet in 1 hour*, 2018. arXiv: 1706.02677 [cs.CV].

[32] I. Loshchilov and F. Hutter, *SGDR: Stochastic gradient descent with warm restarts*, 2017. arXiv: 1608.03983 [cs.LG].

Printed by Tryckeriet i E-huset, Lund 2026

LUND
UNIVERSITY

Series of Master´s theses
Department of Electrical and Information Technology
LU/LTH-EIT 2026-1147
http://www.eit.lth.se