Geometric Inductive Biases for Semi-Supervised Equalization: The Constellation-Aware Transformer

Avi Caciularu
Google Research

Abstract

Decoding signals over unknown channels with minimal pilot overhead is a critical challenge in next-generation communications. Existing deep learning approaches typically rely on generic encoders that struggle to model long-range temporal dependencies or efficiently capture the channel's physical properties from scarce data. We argue that standard architectures suffer from *agnostic estimation gaps*, as they must implicitly learn the constellation geometry that is already known. We introduce the *Constellation-Aware Transformer* (CAT), a novel architecture that explicitly injects geometric inductive biases into the equalization process. CAT is composed of a stack of custom *TransFIRmer* blocks, which use an "early interaction" paradigm to co-process received signals and ideal constellation symbols. Each block features a split Feed-Forward Network that applies a Finite Impulse Response (FIR)-inspired filter for deconvolution and a parallel MLP for geometric refinement. We show that this design is structurally aligned with the optimal linear (MIMO Wiener) receiver: its attention can implement a matched-filter bank, and its bidirectional FIR branch provides the non-causal filtering that block MMSE equalization requires. In the semi-supervised setting, CAT needs fewer pilots than VAE and standard Transformer baselines: on two of our three ISI channels, it reaches a lower SER with 64 pilots than they do with 128.

1 Introduction

Deep learning methods for communications over unknown channels have attracted considerable interest, promising to replace rigid, model-based algorithms with flexible, data-driven solutions [O'Shea and Hoydis, 2017, Bennatan et al., 2018, Nachmani et al., 2018, O'Shea et al., 2018, Shlezinger et al., 2020, Caciularu et al., 2021b]. A central challenge is to minimize the pilot data required for reliable decoding, as this overhead directly reduces spectral efficiency [Shlezinger et al., 2021a].

While classical methods like Expectation-Maximization (EM) offer a path to unsupervised equalization [Tong and Perreau, 1998, Dempster et al., 2018], their performance is limited by initialization sensitivity and local minima. Deep generative models, specifically Variational Autoencoders (VAEs) [Kingma and Welling, 2014], have been adapted to a semi-supervised learning (SSL) framework [Kingma et al., 2014], demonstrating significant pilot reduction [Caciularu and Burshtein, 2018, 2020, Burshtein and Bery, 2023, 2024]. However, the encoders in these systems, typically MLPs or CNNs, may not fully capture complex temporal dependencies, particularly under severe Inter-Symbol Interference (ISI).

The Transformer architecture [Vaswani et al., 2017] motivates application to this domain, yet a direct adaptation is often suboptimal [Choukroun and Wolf, 2024]. A generic Transformer has to learn the underlying physics, such as the constellation geometry and the channel's filtering, entirely from data, even though communication theory already provides much of it.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

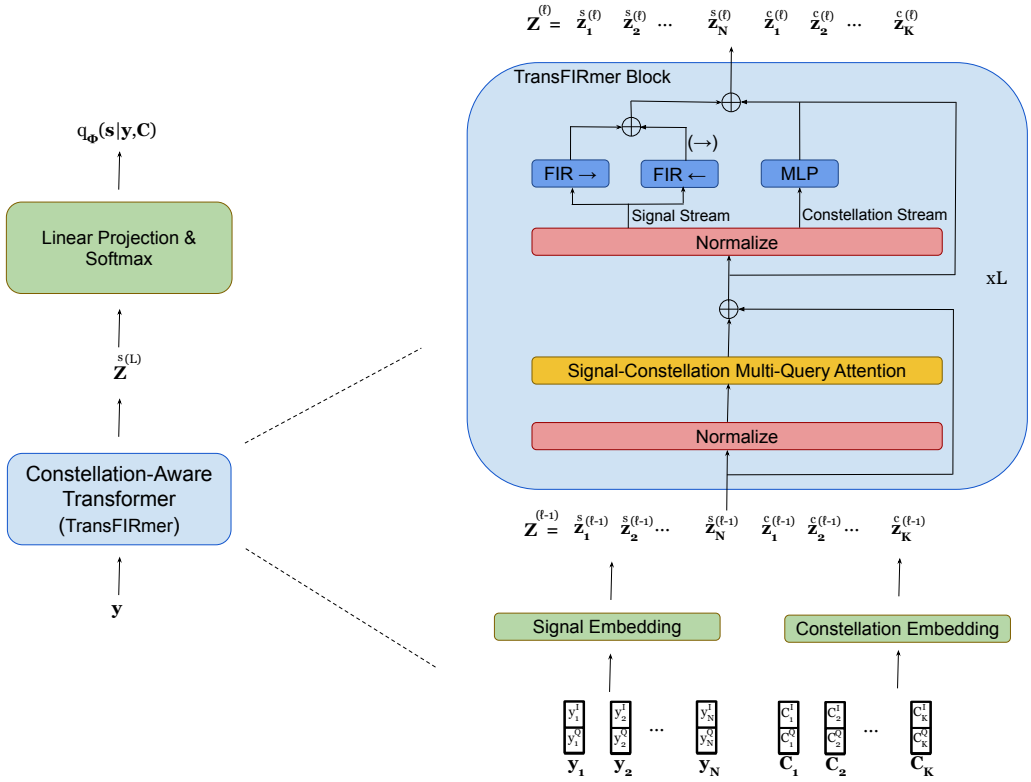


Figure 1: The architecture of our Constellation-Aware Transformer (CAT). The model processes the received signals $\mathbf{y}$ and ideal constellation symbols $\mathcal{C}$ through a stack of TransFIRmer blocks, enabling deep interaction between signal and geometry.

We instead build this knowledge into the Transformer's internal components. In NLP, processing a query and a document jointly from the first layer (*early and deep interaction*) is more accurate than encoding them separately [Humeau et al., 2020, Caciularu et al., 2021a, Fang et al., 2020]. We do the same with the received signal and the ideal constellation symbols.

We introduce the *Constellation-Aware Transformer (CAT)*, composed of novel *TransFIRmer* blocks (Figure 1). Each block changes the standard Transformer encoder in two ways:

1. **Constellation-Aware Attention:** Co-processes received signals and ideal constellation symbols. The model is given the exact constellation, and its attention can act as a matched-filter bank.
2. **TransFIRmer FFN:** Replaces the standard MLP with two streams: a bidirectional FIR-inspired filter on the signal tokens for deconvolution, and a parallel MLP on the constellation tokens.

The analysis in Section 3.1 motivates these two choices. It shows that an encoder given the constellation can do no worse than one without it, that CAT's attention can act as a matched-filter bank, and that its bidirectional FIR branch supplies the non-causal filtering needed for block MMSE equalization. These are statements about capacity, not about what training finds. In our experiments, CAT outperforms VAE and standard Transformer baselines on a nonlinear memoryless channel and, once at least 32 pilots are available, on three synthetic ISI channels. The advantage holds on the 3GPP TDL-A, TDL-C and TDL-D profiles and carries over to coded transmission with a 5G NR LDPC code.

2 Problem Formulation and Setup

We consider a block of N symbols, $(s_1, \ldots, s_N)$, transmitted over an unknown channel. Each symbol $s_i \in \{1, \ldots, K\}$ is drawn independently and uniformly and mapped to a signal from a fixed

constellation $\mathcal{C}$ of size K. We represent these signals as real-valued vectors $\mathbf{x}_i = (x_i^I, x_i^Q) \in \mathbb{R}^2$. The first $N_p \ll N$ symbols are known pilots and the rest form the unknown payload.

2.1 Channel Models

2.1.1 Memoryless Channels

In a memoryless channel, the received signal depends only on the current transmitted signal. We model the transformation as an unknown function $h(\cdot)$, typically including I/Q imbalance and fading. The received signal is $\mathbf{y}_i = h(\mathbf{x}_i) + \mathbf{n}_i$, where $\mathbf{y}_i = (y_i^I, y_i^Q) \in \mathbb{R}^2$.

2.1.2 Channels with Finite Memory

In channels with finite memory, the received signal $\mathbf{y}_i$ is affected by ISI. Following Burshtein and Bery [2024], we define this as a memoryless nonlinearity $g(\cdot)$ followed by a linear channel impulse response:

$$\mathbf{y}_i = \sum_{l=0}^{L-1} \mathbf{h}_l g(\mathbf{x}_{i-l}) + \mathbf{n}_i, \tag{1}$$

where the complex-valued filter taps $\{\mathbf{h}_l\}_{l=0}^{L-1}$ and the nonlinearity $g(\cdot)$ are unknown to the receiver.

2.2 The Semi-Supervised Variational Framework

To reduce the dependency on pilots, we adopt a semi-supervised learning (SSL) framework. The VAE-based approach [Caciularu and Burshtein, 2018, 2020, Burshtein and Bery, 2023, 2024] involves an **encoder** $q_\phi(s|\mathbf{y})$, approximating $p(s|\mathbf{y})$, and a **decoder** $p_\theta(\mathbf{y}|s)$, modeling the forward channel. They are trained jointly by minimizing:

$$\begin{aligned}\mathcal{L}_{\text{SSL}} = &-\frac{\alpha}{N_p}\sum_{i=1}^{N_p}\log q_\phi(s_i|\mathbf{y}_i) - \frac{\gamma}{N_p}\sum_{i=1}^{N_p}\log p_\theta(\mathbf{y}_i|s_i) \quad + \\ &\frac{1-\gamma}{N-N_p}\sum_{i=N_p+1}^{N}\Big[-\mathbb{E}_{q_\phi(s|\mathbf{y}_i)}[\log p_\theta(\mathbf{y}_i|s)] + D_{KL}(q_\phi(s|\mathbf{y}_i)\|p(s))\Big],\end{aligned} \tag{2}$$

where α and γ are weighting hyperparameters. CAT replaces the encoder $q_\phi(s|\mathbf{y})$.[1]

3 Proposed Method: The Constellation-Aware Transformer

We first motivate the design of CAT theoretically and then describe its architecture.

3.1 Theoretical Analysis

Standard equalizers approximate the inverse channel function purely from data. We first explain why ignoring the known constellation is a disadvantage, and then show that our architecture is structurally aligned with the optimal linear receiver. These results concern representational capacity and population risk; whether training realizes them is tested in Section 4.

3.1.1 The Cost of Agnosticism: The Estimator Gap

Let s be a symbol from constellation $\mathcal{C}$, and let $\hat{s}^*(\mathbf{y}) = \mathbb{E}[s|\mathbf{y}, \mathcal{C}^*]$ be the optimal MMSE estimator conditioned on the true constellation $\mathcal{C}^*$. A domain-agnostic model must implicitly learn a posterior approximation $q_\phi(\mathcal{C}|\mathbf{y})$ from scarce pilot data. We quantify the penalty via the *Excess Bayesian Risk*. In practice the receiver knows the constellation. The lemma concerns encoders that do not take it as input. Such an encoder has to represent the geometry implicitly in its weights, and $q_\phi(\mathcal{C}|\mathbf{y})$ is an idealized description of the uncertainty that remains, not a component of any baseline.

[1] For ISI channels, the encoder conditions on the entire sequence $\mathbf{y}$, i.e., $q_\phi(s_i|\mathbf{y})$. On the memoryless channel, CAT also attends over the whole block, because the fading coefficient and the I/Q imbalance are the same for all its symbols.

Lemma 1 (The Agnostic Estimator Gap)**.** *Assume the conditional expectation function $m(\mathcal{C}) = \mathbb{E}[s|\mathbf{y}, \mathcal{C}]$ is bounded by a constant R. The expected excess MSE of a constellation-agnostic estimator $\hat{s}_\phi$ relative to the optimal estimator $\hat{s}^*$ is upper-bounded by:*

$$\mathcal{R}(\hat{s}_\phi) - \mathcal{R}(\hat{s}^*) \leq 2R^2 \mathbb{E}_{\mathbf{y}} \left[D_{KL}(\delta_{\mathcal{C}^*} || q_\phi(\mathcal{C}|\mathbf{y})) \right]. \tag{3}$$

Proof Sketch. We bound $\|\hat{s}_\phi - \hat{s}^*\|$ via the total variation distance between q_ϕ and $\delta_{\mathcal{C}^*}$, then apply Pinsker's inequality. Full derivation in Appendix A. □

Corollary 1 (Finite-Sample Pilot Efficiency, informal)**.** *Under PAC-Bayes assumptions, for a model with M parameters trained on N_p pilots, $\mathbb{E}[D_{KL}(\delta_{\mathcal{C}^*} || q_\phi)] = \mathcal{O}(M/N_p)$. Combined with Lemma 1, the excess risk of a constellation-agnostic estimator decays as $\mathcal{O}(M/N_p)$.*

CAT receives $\mathcal{C}^*$ as input, so in this idealized picture its gap is zero. CAT still has to learn the channel from the pilots; what it no longer has to learn is the constellation.

3.1.2 Hypothesis Space Superiority

We further formalize this advantage via the available hypothesis spaces.

Theorem 1 (Advantage of Parametric Hypothesis Space)**.** *Let $\mathcal{G}_{std}$ be the space of functions mapping $\mathbf{y} \to \hat{s}$, and $\mathcal{G}_{CA}$ be the space of functions mapping $(\mathbf{y}, \mathcal{C}) \to \hat{s}$. The achievable MMSE satisfies $MMSE_{CA} \leq MMSE_{std}$.*

Proof Sketch. $\mathcal{G}_{\text{std}} \subseteq \mathcal{G}_{\text{CA}}$ since any $g \in \mathcal{G}_{\text{std}}$ extends to $\psi(\mathbf{y}, \mathcal{C}) = g(\mathbf{y})$. The infimum over the superset cannot exceed that over the subset. On its own, this result says nothing about finite samples; Appendix A gives a heuristic argument for why not having to learn the constellation should reduce the number of pilots needed. □

3.1.3 Structural Alignment with Optimal Filtering

Classical theory states that the optimal linear receiver for ISI channels consists of a *Block MMSE Equalizer* followed by a *Matched Filter*.[2]

Proposition 1 (Structural Alignment with Optimal Receiver)**.** *The CAT architecture provides inductive biases aligned with optimal receiver components:*

1. ***Matched Filter:*** *The Constellation-Aware Attention has the capacity to implement a Matched Filter bank through learned projections.*
2. ***Block MMSE Equalizer:*** *The TransFIRmer's bidirectional FIR-FFN provides the non-causal spectral bias conducive to deconvolution.*

Proof Sketch. For (1), expanding the Gaussian posterior shows detection reduces to cross-correlations $\mathbf{z}^T c_k$, which the attention mechanism computes. For (2), the optimal equalizer $\mathbf{W} = \mathbf{H}^H \mathbf{R}_{yy}^{-1}$ requires a non-causal FIR filter, provided by our bidirectional convolutions. Full proof and a discussion of non-constant modulus constellations in Appendix A. □

Proposition 1 is about capacity; we do not claim that training converges to the Wiener solution. On $h^{(1)}$ and $h^{(2)}$, CAT's gap to the optimal decoder shrinks as N_p grows (Table 2).

3.2 The CAT Architecture

CAT is a stack of 3 layers, which we call *TransFIRmer blocks*.

[2]The mathematical structure is equivalent to the MIMO Wiener filter when ISI is modeled as virtual spatial multiplexing.

3.2.1 Input Representation and Embedding

The input consists of the sequence of N received channel outputs $(\mathbf{y}_1, \ldots, \mathbf{y}_N)$ and the set of K ideal constellation symbols $\{\mathbf{x}(1), \ldots, \mathbf{x}(K)\}$. Both are projected into dimension d_{model}. We add fixed sinusoidal positional embeddings $\mathbf{p}_i$ to the signals, but not to the set-based constellation symbols:

$$\mathbf{z}_i^{\text{sig}} = \text{Embed}_{\text{sig}}(\mathbf{y}_i) + \mathbf{p}_i, \quad i \in \{1, \ldots, N\},$$
$$\mathbf{z}_k^{\text{const}} = \text{Embed}_{\text{const}}(\mathbf{x}(k)), \quad k \in \{1, \ldots, K\}.$$

3.2.2 The TransFIRmer Block

The TransFIRmer block modifies both Transformer sub-layers to align with Proposition 1.

Signal-Constellation Attention Mechanism. This stage enables deep interaction via joint self-attention over the concatenated $N + K$ tokens ($\mathbf{Q}$, $\mathbf{K}$, $\mathbf{V}$ computed from the full sequence via multi-query attention). We use a full (bidirectional) attention mask, since equalization is a *smoothing* problem.

Two-Stream Feed-Forward Network. The second stage replaces the MLP with two streams.

- **Signal Stream (Bidirectional FIR-Inspired Filter):** The signal tokens $\mathbf{Z}_{\text{sig}}^{(l)}$ are processed by a pair of 1D convolutional layers implementing a bidirectional filter:

$$\text{FFN}_{\text{sig}}(\mathbf{Z}_{\text{sig}}) = \text{Conv}_{\text{fwd}}(\mathbf{Z}_{\text{sig}}) + \text{Flip}\left(\text{Conv}_{\text{bwd}}\left(\text{Flip}\left(\mathbf{Z}_{\text{sig}}\right)\right)\right). \tag{4}$$

  This sub-layer can represent the channel-matched filter $\mathbf{H}^H$ of the optimal MMSE equalizer (Proposition 1). Since the signals are in $\mathbb{R}^2$, it can also learn the complex conjugation in $\mathbf{H}^H$.

- **Constellation Stream (MLP):** A standard MLP refines the constellation tokens $\mathbf{Z}_{\text{const}}^{(l)}$.

Each stream has its own residual connection.

3.3 Symmetry Breaking and Inductive Bias

A limitation of standard Transformers is their permutation equivariance: they lack an intrinsic "anchor" for the signal space. In 16-QAM, the difference between symbols is geometric and absolute, not relative. CAT breaks this symmetry by feeding fixed constellation tokens $\mathcal{C}$ into attention, providing a rigid coordinate system against which received signals are measured. The attention scores $A_{ij} \propto \mathbf{y}_i^T \mathbf{c}_j$ perform soft-decision mapping to symbol centroids, transforming the problem from "learning clustering from scratch" to "registering points to a known template".

3.4 Computational Complexity Analysis

The complexity of standard self-attention over N tokens is $\mathcal{O}(N^2 d)$. In CAT, we process $N + K$ tokens, giving $\mathcal{O}((N + K)^2 d)$. The attention cost therefore grows by a factor of $(1 + K/N)^2$. For 16-QAM ($K = 16$), the overhead is about 56% at $N = 64$, 13% at $N = 256$ and 6% at $N = 512$. The pilot savings reduce transmission overhead, not this extra computation at the receiver.

Scalability to High-Order Modulation. Table 1 shows how CAT scales with the constellation order on $h^{(1)}$. Its SER gain over the vanilla Transformer shrinks from 3.8 dB for 16-QAM to 3.0 dB for 64-QAM and 2.2 dB for 256-QAM, while the attention overhead grows to about 56% for 64-QAM and 300% for 256-QAM. For 64-QAM the pilot savings translate into a 15% gain in spectral efficiency. Appendix B discusses higher orders and compares parameter counts.

4 Experiments

We evaluate CAT as a semi-supervised channel equalizer. The experiments ask whether its two inductive biases, the constellation prior with early interaction and the TransFIRmer block, allow sample-efficient adaptation to channel impairments without large offline training sets. We report results on a nonlinear memoryless channel and on three ISI channels, followed by an ablation. Section 5 adds results with channel coding and on 3GPP TDL channels.

Table 1: Scalability across constellation orders ($h^{(1)}$ channel, $N = 256$, SNR 16–24 dB, gains at target SER 10^{-2}).

Modulation	K	**Attn. Overhead**	**SER Gain vs. Vanilla Trans.**
16-QAM	16	$\sim$13%	3.8 dB
64-QAM	64	$\sim$56%	3.0 dB
256-QAM	256	$\sim$300%	2.2 dB

4.1 Experimental Setup

To evaluate CAT, we test it on two categories of channels: a nonlinear memoryless channel with I/Q imbalance and Rayleigh fading, and three standard channels with finite memory (ISI). Unless stated otherwise, experiments use 16-QAM.

Our CAT model consists of a 3-layer stack of TransFIRmer blocks with Multi-Query Attention [Shazeer, 2019] and fixed sinusoidal positional embeddings. CAT achieves its performance with approximately 10k parameters, comparable to the vanilla Transformer (9k) and far fewer than VAE-CNN (150k; see Appendix B). Other hyperparameters are consistent with Burshtein and Bery [2024] (see Appendix F).

We compare against the optimal decoder (ML/BCJR) [Bahl et al., 1974], SSL Monte Carlo EM (MCEM) [Wei and Tanner, 1990], SSL Viterbi EM [Dempster et al., 2018], Simple Decision Directed (SDD), VAE-CNN [Burshtein and Bery, 2024], and the meta-learning algorithm CAVIA [Zintgraf et al., 2019].[3] We also include a vanilla Transformer baseline [Kunde et al., 2025] to isolate the benefits of our architectural modifications. We report the mean Symbol Error Rate (SER) over 1000 independent Monte Carlo trials; the 95% confidence intervals are negligible.

SNR Range and Training Protocol. Most experiments use moderate-to-high SNR (17–22 dB): this range is where the semi-supervised challenge is most relevant, it matches practical systems (e.g., 5G NR targets 15–25 dB), and it allows direct comparison with prior work [Burshtein and Bery, 2023, 2024]. Table 3 extends the range to 14–26 dB on $h^{(1)}$. Each block experiences a single channel realization. CAT, the vanilla Transformer and VAE-CNN are trained from scratch on every block for 5,000 steps with Eq. (2), using only that block's pilots and its unlabeled payload, and SER is measured on the same payload. No payload labels and no data from other channel realizations are used, except by the CAVIA baseline and in the meta-initialized runs of Section 5. The model is therefore never tested on a channel other than the one it has just adapted to; when the channel changes, it adapts again. Residual carrier frequency offset (CFO) and phase noise within a block are covered in Appendix C.4, and other channel types by the 3GPP TDL profiles (Section 5.3).

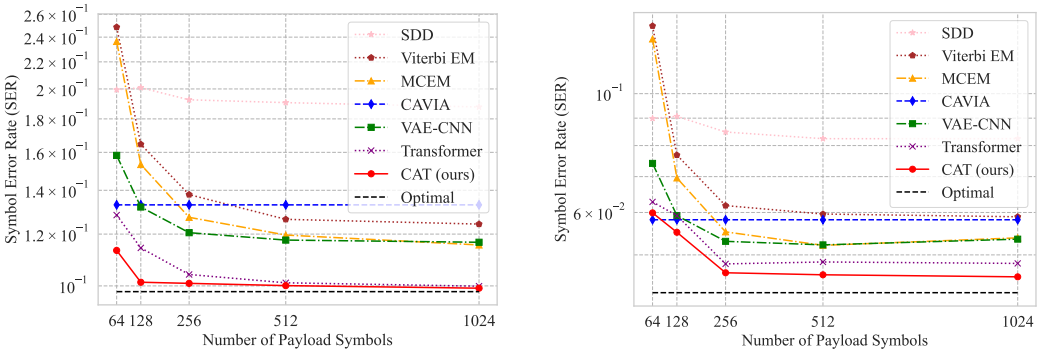
4.2 Results on Memoryless Channels

Figure 2 shows the SER as a function of the number of payload symbols at 18 dB and 22 dB. CAT has a lower SER than the VAE-CNN at every payload size, and the gap is largest for the shortest payload (64 symbols). As payload size increases, CAT closely approaches the Optimal decoder, suggesting that the "Early Interaction" prior effectively substitutes for large training sets.

4.3 Results on Channels with Memory (ISI)

We evaluate CAT on three standard ISI channels (see Appendix D.2) with different lengths ($L = 5, 4, 10$). Table 2 summarizes the results. With $N_p = 16$, one pilot per constellation point on average, all three learned equalizers do poorly (SER 0.29–0.65), and on $h^{(1)}$ CAT trails both baselines. From $N_p = 32$ on, CAT is the best of the three on every channel. For $N_p \geq 64$ it reduces SER relative to the stronger baseline by 1.9–3.4$\times$ on $h^{(1)}$ and $h^{(2)}$ and by 1.2–1.4$\times$ on the 10-tap $h^{(3)}$, where BCJR is intractable. On $h^{(1)}$ and $h^{(2)}$, CAT with 64 pilots also has a lower SER than either baseline with 128, and with 128 pilots it comes within a factor of 1.3 and 2.5 of the BCJR optimum, compared with

[3] CAVIA requires meta-training on prior channel realizations, giving it access to distributional knowledge unavailable to other methods under strict test-time training constraints. We include it as an oracle-assisted baseline.



(a) SER results for SNR $= 18$ dB.

(b) SER results for SNR $= 22$ dB.

Figure 2: Symbol Error Rate (SER) on a memoryless nonlinear channel as a function of the number of payload symbols, with $N_p = 16$ pilots. Apart from the optimal decoder, only the meta-trained CAVIA baseline has a lower SER than CAT, and only at 64 payload symbols and 22 dB.

Table 2: SER on channels with memory ($E_x/N_0 = 17$ dB, payload 256). CAT vs. VAE-CNN and Vanilla Transformer.

Channel	Pilots	VAE-CNN	Trans.	CAT (Ours)	Optimal
$h^{(1)}$ (L=5)	16	**0.2900**	0.3392	0.3580	0.0121
	32	0.1251	0.1192	**0.0842**	
	64	0.0523	0.0610	**0.0198**	
	128	0.0494	0.0290	**0.0156**	
$h^{(2)}$ (L=4)	16	0.3447	0.3563	**0.3330**	0.0101
	32	0.1843	0.1593	**0.1372**	
	64	0.1002	0.0750	**0.0340**	
	128	0.0869	0.0888	**0.0257**	
$h^{(3)}$ (L=10)	16	0.6211	0.6481	**0.6103**	N/A
	32	0.3943	0.3874	**0.3426**	
	64	0.1709	0.1692	**0.1181**	
	128	0.1087	0.1022	**0.0846**	

2.4 and 8.6 for the stronger baseline. The ablation in Table 4 attributes most of the gain on $h^{(1)}$ to the bidirectional FIR-FFN and the constellation prior. Table 3 repeats the $h^{(1)}$ comparison from 14 to 26 dB: CAT has the lowest SER at every SNR, and its margin over the stronger baseline grows from $1.5\times$ at 14 dB to $3.9\times$ at 26 dB.

4.4 Ablation Study

Table 4 reports ablations on the memoryless channel and, for the main components, on $h^{(1)}$.

1. **Importance of FIR-FFN:** Replacing the FIR filter with a standard MLP ("CAT with MLP-FFN") changes little on the memoryless channel (0.0599 vs. 0.0615), where there is no ISI to undo. On $h^{(1)}$ the same change raises SER from 0.0198 to 0.0351 (+77%), and removing only the backward branch raises it to 0.0287 (+45%), as expected for non-causal equalization. On $h^{(1)}$ and $h^{(2)}$, the FIR gain is positive at every SNR we tested (7–77% on $h^{(1)}$ and 30–71% on $h^{(2)}$) and smallest at 14 dB, where noise rather than residual ISI limits performance (Appendix C.1).
2. **Constellation Prior:** With the FIR-FFN kept, removing the prior raises the ISI SER from 0.0198 to 0.0438 ($2.2\times$). Two variants receive the same input as CAT: "CAT with MLP-FFN" has the constellation tokens and joint attention but a standard FFN, and "Self-Only Attention" has the tokens but no signal–constellation attention. Both are worse than CAT (0.0615 and 0.0621 vs. 0.0599), so having the constellation in the input is not enough by itself. A 45° *rotated* prior is

Table 3: SER vs. SNR on $h^{(1)}$ ($N_p = 64$, payload 256). The 17 dB column repeats Table 2.

Method	**14 dB**	**17 dB**	**20 dB**	**24 dB**	**26 dB**
VAE-CNN	0.1842	0.0523	0.0215	0.0068	0.0031
Vanilla Trans.	0.1763	0.0610	0.0248	0.0091	0.0042
CAT (Ours)	**0.1205**	**0.0198**	**0.0072**	**0.0019**	**0.0008**

Table 4: Ablation study. Memoryless: $N_p = 16$, SNR $= 22$ dB, 64 payload symbols (as in Figure 2b), SER $\pm$ 95% CI. ISI: $h^{(1)}$, $E_x/N_0 = 17$ dB, $N_p = 64$, payload 256.[4]

Model Variant	**Memoryless**	**ISI ($h^{(1)}$)**
Our Full Method		
CAT (built with TransFIRmer blocks)	**0.0599 $\pm$ 0.0005**	**0.0198**
Architecture & Prior Ablations		
CAT without Inverse FIR Filter	0.0608 $\pm$ 0.0005	0.0287
CAT with MLP-FFN (No FIR)	0.0615 $\pm$ 0.0006	0.0351
No Prior + Bidirectional FIR	–	0.0438
Vanilla Transformer (No Prior)	0.0628 $\pm$ 0.0007	0.0610
CAT (45° Rotated Prior)	0.1550 $\pm$ 0.0015	–
CAT + Pos. Emb. on Constellation	0.0631	0.0224
Attention Mask Ablations		
Self-Only Attention	0.0621 $\pm$ 0.0007	–
Causal Attention	0.0612 $\pm$ 0.0006	–
External Baseline		
VAE-CNN (from Burshtein and Bery [2024])	0.0741 $\pm$ 0.0009	0.0523

much worse than no prior at all (0.1550 vs. 0.0628), which indicates that the model does use the constellation as a geometric reference. This also means that residual CFO and phase noise, which rotate the received constellation, could hurt CAT. With coarse PLL tracking and a learned phase rotation of the constellation keys (Appendix C.4), CAT keeps a 1.5 dB advantage over the baselines under severe Wiener phase noise ($\sigma_\Delta^2 = 10^{-3}$ rad^2/symbol), with no error floor. Adding positional embeddings to the constellation tokens, which have no natural order, also hurts ($+5\%$ memoryless, $+13\%$ ISI).

3. **Attention Mask:** "Causal Attention" is slightly worse than the full mask (0.0612 vs. 0.0599), since it ignores later symbols that share the block's fading and I/Q imbalance.

5 Discussion and Limitations

5.1 Online Adaptation as Test-Time Training

A practical concern with our framework is the cost of per-block training. In standard deployments like 5G or WiFi, coherence times are short, and full retraining for every block may seem prohibitive. However, our approach is an instance of *Test-Time Training (TTT)* or *Test-Time Adaptation* [Sun et al., 2020], in which a model adapts to distribution shift at inference time.

In scenarios with highly non-stationary fading or nonlinear hardware drifts (e.g., power amplifier heating), a static model trained offline fails to generalize. Our method targets these cases, where the channel changes quickly. Training and testing on the same realization is therefore the problem setting itself, not a failure to generalize (see the protocol in Section 4.1).

We acknowledge that the current training regime (5,000 steps per block; see Appendix F) is computationally expensive for real-time systems. On $h^{(1)}$ with 128 pilots, 200 steps from a random

[4] "Vanilla Trans" = self-attention on signals only; "Self-Only" = signals and constellations attend within their own group; "w/o Inverse FIR" = forward conv only; "No Prior + Bidirectional FIR" = signal-only attention with the TransFIRmer FFN; "Pos. Emb. on Constellation" = positional embeddings also added to the constellation tokens.

Table 5: Coded BLER with 3GPP 5G NR LDPC (Rate 1/2, CW=1024) on the $h^{(1)}$ channel. Gains at target BLER $= 10^{-2}$.

Method	BLER Gain	ECE	Calibration Improvement
VAE-CNN	—	0.085	—
Vanilla Transformer	0 dB (reference)	0.089	—
CAT (Ours)	**+2.5 dB**	**0.012**	$7.4\times$

Table 6: SER on 3GPP TDL channels (16-QAM, $N = 128$, 500 realizations per profile). Gain is the SER of the stronger baseline divided by that of CAT.

Profile	Delay spread	Vanilla Trans.	VAE-CNN	CAT (Ours)	Gain
TDL-A (NLOS)	30 ns	0.0385	0.0421	**0.0195**	$1.97\times$
TDL-D (LOS)	30 ns	0.0198	0.0231	**0.0112**	$1.77\times$

initialization already give an SER only 4% above the 5,000-step result, but 50 steps give a 47% higher SER (Table 9). With a meta-learned initialization (CAVIA; Zintgraf et al., 2019), 50 steps give an SER only 2% above the 5,000-step result and still lower than that of the vanilla Transformer and VAE-CNN trained from scratch for 5,000 steps. This is not a like-for-like comparison, since the baselines were not meta-trained. Meta-training uses earlier channel realizations and runs offline, so its cost is not part of the adaptation time, which is the per-step time (Table 10) multiplied by the number of steps. A CAT step takes 17–28% longer than a vanilla Transformer step for 16-QAM and about twice as long for 64-QAM, since the attention also covers the K constellation tokens. Appendix C gives further results on convergence speed, wall-clock latency and CFO robustness.

5.2 Coded Performance and LLR Calibration

To evaluate CAT in a full receiver chain, we feed its soft outputs into a 3GPP 5G NR LDPC decoder (Rate 1/2, codeword length 1024). Table 5 presents the Block Error Rate (BLER) results.

At BLER $= 10^{-2}$, CAT improves on the vanilla Transformer by 2.5 dB on $h^{(1)}$. The encoder outputs a posterior over the K symbols, from which we compute bit LLRs for the decoder. CAT's posteriors are also better calibrated: its expected calibration error (ECE) is 0.012, compared with 0.089 for the vanilla Transformer and 0.085 for VAE-CNN, and we applied no temperature scaling to any model. This matters because LDPC decoding is sensitive to the reliability of its input LLRs.

5.3 Evaluation on Standardized 3GPP Channels

To test generalization beyond the synthetic ISI channels, we evaluate CAT on three 3GPP TDL profiles [3GPP, 2022]: TDL-A and TDL-C, which are NLOS profiles, and TDL-D, which has a LOS component. We use a 30 ns delay spread for TDL-A and TDL-D and 300 ns for TDL-C, with 30 kHz subcarrier spacing. For each profile we use the same 5,000-step cold-start protocol with 16-QAM and $N = 128$, and average over 500 channel realizations. On TDL-A and TDL-D, CAT's SER is about $2.0\times$ and $1.8\times$ lower than that of the stronger baseline (Table 6), and on TDL-C it has a 2.1 dB SER advantage over both baselines. As in DeepRx [Honkala et al., 2021] and NVIDIA Sionna [Hoydis et al., 2022], we use 3GPP channel models for link-level evaluation; we do not model mobility within a block.

5.4 Limitations

CAT needs a minimum number of pilots. With $N_p = 16$ on the ISI channels, all the learned equalizers do poorly, and on $h^{(1)}$ CAT trails both baselines (Table 2). Our theoretical results concern representational capacity and population risk. They do not show that training converges to the Wiener receiver, and the finite-sample argument in Appendix A is heuristic. Fewer pilots save transmission overhead, not receiver computation: per-block training adds compute at the receiver, and the 50-step regime relies on offline CAVIA meta-training on earlier channel realizations, whose cost is not included in the latency we report.

All our results come from link-level simulation (synthetic ISI channels, 3GPP TDL profiles, and 5G NR LDPC coding); we have not evaluated CAT over the air or at the system level, for example with mobility, HARQ, or scheduling. Finally, the attention cost grows with the constellation size, which currently limits CAT to 256-QAM (Appendix B).

6 Related Work

Physics-Informed Deep Learning. Physical priors are increasingly built into neural networks, from Hamiltonian Neural Networks [Greydanus et al., 2019] to equivariance in computer vision. In communications, "Model-Driven" deep learning unfolds iterative algorithms into layers [He et al., 2019]. Our work adds a *geometric* inductive bias, the constellation, to the attention mechanism.

Semi-Supervised Equalization. Classical approaches like EM and Decision Directed methods are effective but brittle under severe distortions. Recent VAE-based methods [Caciularu and Burshtein, 2018, 2020, Zhu et al., 2023, Burshtein and Bery, 2023, 2024] have established strong baselines, but their MLP/CNN encoders lack the sequence modeling power of Transformers and must learn the constellation geometry from scratch. ViterbiNet [Shlezinger et al., 2020] and DeepSIC [Shlezinger et al., 2021b] do not need channel state information either, but they are trained in a supervised way on labeled symbols, and DeepRx [Honkala et al., 2021] is trained offline on a large labeled dataset. Our setting instead adapts to each block from 16 to 128 pilots plus the unlabeled payload, which is why we compare against semi-supervised and meta-learning methods.

Meta-Learning for Few-Pilot Demodulation. Meta-learning reduces the number of pilots needed by exploiting data from earlier transmissions. Park et al. [2019], whose memoryless nonlinear channel we also use, meta-learn a demodulator on pilots from previous IoT transmissions so that it adapts to a new channel from a few pilots, and CAVIA [Zintgraf et al., 2019] adapts only a small context vector for each new channel. Both require data from related channels before deployment. Our cold-start protocol does not, since CAT is trained on each block's pilots and payload alone. When earlier channel realizations are available, CAT can also be initialized by meta-learning, which cuts its adaptation to 50 steps (Section 5).

Transformers in Wireless Communications. Transformers have been explored for channel estimation [Yin et al., 2025], CSI feedback [Xu et al., 2021, Zhou et al., 2024], and OFDM MIMO equalization [Song et al., 2024], and attention has been used to decode error-correcting codes [Caciularu et al., 2021b, Choukroun and Wolf, 2024]. These works are fully supervised, and for channel estimation and CSI feedback the true channel is the training target. CAT addresses semi-supervised equalization with few labels, and its attention registers the received signals against the known constellation.

7 Conclusion

We presented the Constellation-Aware Transformer (CAT), a semi-supervised equalizer that builds two pieces of receiver knowledge into a Transformer: the constellation, which the model sees as tokens from the first layer, and the structure of the linear MMSE receiver, which motivates the design of the TransFIRmer block. Our analysis shows that these components can represent a matched-filter bank and the non-causal filtering of block MMSE equalization; it does not show that training finds them.

Empirically, CAT outperforms VAE-CNN and a similarly sized vanilla Transformer on a nonlinear memoryless channel, on three synthetic ISI channels once at least 32 pilots are available, and on the 3GPP TDL-A, TDL-C and TDL-D profiles. On $h^{(1)}$ the advantage persists from 16-QAM to 256-QAM, although it shrinks as the order grows, and with a 5G NR LDPC code it amounts to 2.5 dB in BLER over the vanilla Transformer. The price is per-block training at the receiver. With a CAVIA initialization, 50 adaptation steps give an SER only 2% above the full 5,000-step result, but this requires offline meta-training on earlier channel realizations. The attention cost also grows with the constellation size, which currently limits CAT to 256-QAM; the I/Q factorization and the cheaper attention mechanisms discussed in Appendix B are possible routes to 1024-QAM.

References

3GPP. Study on channel model for frequencies from 0.5 to 100 GHz. Technical Report TR 38.901, 3rd Generation Partnership Project, 2022. Version 17.0.0.

L. Bahl, J. Cocke, F. Jelinek, and J. Raviv. Optimal decoding of linear codes for minimizing symbol error rate (corresp.). *IEEE Transactions on Information Theory*, 20(2):284–287, 1974. doi: 10.1109/TIT.1974.1055186.

Amir Bennatan, Yoni Choukroun, and Pavel Kisilev. Deep learning for decoding of linear codes - a syndrome-based approach. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pages 1595–1599, 2018. doi: 10.1109/ISIT.2018.8437530.

Jean Bretagnolle and Catherine Huber. Estimation des densités: risque minimax. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 47(2):119–137, 1979.

David Burshtein and Eli Bery. Semi-supervised variational inference over nonlinear channels. In *2023 IEEE 24th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, pages 611–615, 2023. doi: 10.1109/SPAWC53906.2023.10304430.

David Burshtein and Eli Bery. Semi-supervised channel equalization using variational autoencoders. *IEEE Transactions on Wireless Communications*, 23(12):19681–19695, 2024. doi: 10.1109/TWC.2024.3485991.

Avi Caciularu and David Burshtein. Blind channel equalization using variational autoencoders. In *2018 IEEE International Conference on Communications Workshops (ICC Workshops)*, pages 1–6, 2018. doi: 10.1109/ICCW.2018.8403666.

Avi Caciularu and David Burshtein. Unsupervised linear and nonlinear channel equalization and decoding using variational autoencoders. *IEEE Transactions on Cognitive Communications and Networking*, 6(3):1003–1018, 2020. doi: 10.1109/TCCN.2020.2990773.

Avi Caciularu, Arman Cohan, Iz Beltagy, Matthew Peters, Arie Cattan, and Ido Dagan. CDLM: Cross-document language modeling. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2648–2662, Punta Cana, Dominican Republic, November 2021a. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.225. URL https://aclanthology.org/2021.findings-emnlp.225/.

Avi Caciularu, Nir Raviv, Tomer Raviv, Jacob Goldberger, and Yair Be'ery. perm2vec: Attentive graph permutation selection for decoding of error correction codes. *IEEE Journal on Selected Areas in Communications*, 39(1):79–88, 2021b. doi: 10.1109/JSAC.2020.3036951.

Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, David Belanger, Lucy Colwell, and Adrian Weller. Rethinking attention with performers. In *International Conference on Learning Representations (ICLR)*, 2021.

Yoni Choukroun and Lior Wolf. A foundation model for error correction codes. In *International Conference on Learning Representations (ICLR)*, 2024.

Imre Csiszár and János Körner. *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Cambridge University Press, 2nd edition, 2011.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1): 1–22, 12 2018. ISSN 0035-9246. doi: 10.1111/j.2517-6161.1977.tb01600.x. URL https://doi.org/10.1111/j.2517-6161.1977.tb01600.x.

Yuwei Fang, Siqi Sun, Zhe Gan, Rohit Pillai, Shuohang Wang, and Jingjing Liu. Hierarchical graph network for multi-hop question answering. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8823–8838, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.710. URL https://aclanthology.org/2020.emnlp-main.710/.

Samuel Greydanus, Misko Dzamba, and Jason Yosinski. Hamiltonian neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Hengtao He, Shi Jin, Chao-Kai Wen, Feifei Gao, Geoffrey Ye Li, and Zongben Xu. Model-driven deep learning for physical layer communications. *IEEE Wireless Communications*, 26(5):77–83, 2019. doi: 10.1109/MWC.2019.1800447.

Mikko Honkala, Dani Korpi, and Janne M. J. Huttunen. Deeprx: Fully convolutional deep learning receiver. *IEEE Transactions on Wireless Communications*, 20(6):3925–3940, 2021. doi: 10.1109/TWC.2021.3054520.

Jakob Hoydis, Sebastian Cammerer, Fayçal Ait Aoudia, Avinash Vem, Nikolaus Binder, Guillermo Marcus, and Alexander Keller. Sionna: An open-source library for next-generation physical layer research. *arXiv preprint arXiv:2203.11854*, 2022.

Samuel Humeau, Kurt Shuster, M. Lachaux, and J. Weston. Poly-encoders: Architectures and pre-training strategies for fast and accurate multi-sentence scoring. In *International Conference on Learning Representations (ICLR)*, 2020.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. In *International Conference on Learning Representations (ICLR)*, 2017.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations (ICLR)*, 2014.

Diederik P. Kingma, Danilo J. Rezende, Shakir Mohamed, and Max Welling. Semi-supervised learning with deep generative models. In *Advances in Neural Information Processing Systems (NIPS)*, 2014.

Vishnu Teja Kunde, Vicram Rajagopalan, Chandra Shekhara Kaushik Valmeekam, Krishna Narayanan, Jean-Francois Chamberland, Dileep Kalathil, and Srinivas Shakkottai. Transformers are provably optimal in-context estimators for wireless communications. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2025.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

Sachin Mehta and Mohammad Rastegari. MobileViT: Light-weight, general-purpose, and mobile-friendly vision transformer. In *International Conference on Learning Representations (ICLR)*, 2022.

Eliya Nachmani, Elad Marciano, Loren Lugosch, Warren J. Gross, David Burshtein, and Yair Be'ery. Deep learning methods for improved decoding of linear codes. *IEEE Journal of Selected Topics in Signal Processing*, 12(1):119–131, 2018. doi: 10.1109/JSTSP.2017.2788405.

Timothy O'Shea and Jakob Hoydis. An introduction to deep learning for the physical layer. *IEEE Transactions on Cognitive Communications and Networking*, 3(4):563–575, 2017. doi: 10.1109/TCCN.2017.2758370.

Timothy James O'Shea, Tamoghna Roy, and T. Charles Clancy. Over-the-air deep learning based radio signal classification. *IEEE Journal of Selected Topics in Signal Processing*, 12(1):168–179, 2018. doi: 10.1109/JSTSP.2018.2797022.

Sangwoo Park, Hyeryung Jang, Osvaldo Simeone, and Joonhyuk Kang. Learning how to demodulate from few pilots via meta-learning. In *2019 IEEE 20th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, pages 1–5, 2019. doi: 10.1109/SPAWC.2019.8815426.

Seong-Joon Park, Taewoo Park, Hee-Youl Kwak, Sang-Hyo Kim, Yongjune Kim, and Jong-Seon No. Efficient message-passing transformer for error correcting codes. In *International Conference on Learning Representations (ICLR)*, 2026.

Noam Shazeer. Fast transformer decoding: One write-head is all you need. *arXiv preprint arXiv:1911.02150*, 2019.

Nir Shlezinger, Nariman Farsad, Yonina C. Eldar, and Andrea J. Goldsmith. Viterbinet: A deep learning based viterbi algorithm for symbol detection. *IEEE Transactions on Wireless Communications*, 19(5):3319–3331, 2020. doi: 10.1109/TWC.2020.2972352.

Nir Shlezinger, Rong Fu, and Yonina C. Eldar. Deepsic: Deep soft interference cancellation for multiuser mimo detection. *IEEE Transactions on Wireless Communications*, 20(2):1349–1362, 2021a. doi: 10.1109/TWC.2020.3032663.

Nir Shlezinger, Rong Fu, and Yonina C. Eldar. DeepSIC: Deep soft interference cancellation for multiuser MIMO detection. *IEEE Transactions on Wireless Communications*, 20(2):1349–1362, 2021b.

Zihang Song, Yuan Ma, Chaoqun You, Haoxuan Yuan, Jinbo Peng, and Yue Gao. Transformer-based adaptive ofdm mimo equalization in intelligence-native ran. In *2024 IEEE/CIC International Conference on Communications in China (ICCC Workshops)*, pages 179–184, 2024. doi: 10.1109/ICCCWorkshops62562.2024.10693758.

Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei A Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shift. In *International Conference on Machine Learning (ICML)*, 2020.

Lang Tong and S. Perreau. Multichannel blind identification: from subspace to maximum likelihood methods. *Proceedings of the IEEE*, 86(10):1951–1968, 1998. doi: 10.1109/5.720247.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NIPS)*, 2017.

Greg C. G. Wei and Martin A. Tanner. A monte carlo implementation of the em algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association*, 85(411): 699–704, 1990. ISSN 01621459, 1537274X. URL `http://www.jstor.org/stable/2290005`.

Yang Xu, Mingqi Yuan, and Man-On Pun. Transformer empowered CSI feedback for massive MIMO systems. In *2021 30th Wireless and Optical Communications Conference (WOCC)*, pages 157–161, 2021. doi: 10.1109/WOCC53213.2021.9602863.

Mei Yin, Mingming Zhao, Lin Liu, and Lifu Liu. Joint channel estimation and feedback with masked token transformers in massive MIMO systems. *Scientific Reports*, 15:41168, 2025. doi: 10.1038/s41598-025-25256-1.

Tao Zhou, Xiangping Liu, Zuowei Xiang, Haitong Zhang, Bo Ai, Liu Liu, and Xiaorong Jing. Transformer network based channel prediction for csi feedback enhancement in ai-native air interface. *IEEE Transactions on Wireless Communications*, 23(9):11154–11167, 2024. doi: 10.1109/TWC.2024.3379123.

Hongzhi Zhu, Yongliang Guo, Wei Xu, and Xiaohu You. Semi-supervised mimo detection using cycle-consistent generative adversarial network. *IEEE Transactions on Cognitive Communications and Networking*, 9(5):1226–1240, 2023. doi: 10.1109/TCCN.2023.3279260.

Luisa Zintgraf, Kyriacos Shiarli, Vitaly Kurin, Katja Hofmann, and Shimon Whiteson. Fast context adaptation via meta-learning. In *International Conference on Machine Learning (ICML)*, 2019.

Broader Impact

This paper presents work whose goal is to advance the field of machine learning and wireless communications. We highlight both the potential benefits and considerations of our research.

Potential Benefits. Our proposed Constellation-Aware Transformer (CAT) architecture reduces the pilot overhead required for reliable signal decoding, directly improving spectral efficiency in wireless communication systems. This has positive implications for: (1) *Connectivity:* More efficient use of limited spectrum resources can expand network capacity, potentially improving access to communication services in underserved regions. (2) *Resource Efficiency:* Reducing the transmission overhead lowers the energy required per successfully decoded bit, contributing to more sustainable communication infrastructure. (3) *Scientific Advancement:* Our theoretical framework connecting geometric inductive biases to optimal estimation theory (Lemma 1, Proposition 1) provides insights that may transfer to other domains where known structure can inform neural architecture design.

Potential Risks and Considerations. As with many advances in signal processing and communications, our methods could, in principle, be applied to both beneficial and harmful applications. More robust receivers could be deployed in surveillance systems, though channel equalization is a mature technology and our contribution is an incremental improvement, not a new capability. The computational requirements of our test-time training paradigm (discussed in Section 5) currently limit practical deployment, and future work on computational efficiency should consider responsible deployment contexts.

Broader Context. We believe the societal implications of this work are those typically associated with incremental advances in foundational machine learning and communications research. We do not foresee unique ethical concerns beyond those inherent to the field and encourage the community to continue developing norms around responsible AI deployment in critical infrastructure.

A Proofs and Theoretical Remarks

A.1 Proof of Lemma 1 (The Agnostic Estimator Gap)

Proof. Let $\hat{s}_\phi(\mathbf{y}) = \int m(\mathcal{C}, \mathbf{y}) q_\phi(\mathcal{C}|\mathbf{y}) d\mathcal{C}$ be the constellation-agnostic estimator. The error norm is:

$$\|\hat{s}_\phi - \hat{s}^*\| = \left\| \int m(\mathcal{C})(q_\phi(\mathcal{C}) - \delta(\mathcal{C} - \mathcal{C}^*)) d\mathcal{C} \right\|.$$

Using the boundedness assumption $\|m(\mathcal{C})\| \leq R$ and the integral triangle inequality:

$$\|\hat{s}_\phi - \hat{s}^*\| \leq \int \|m(\mathcal{C})\| \cdot |q_\phi(\mathcal{C}) - \delta(\mathcal{C} - \mathcal{C}^*)| d\mathcal{C} \leq R \cdot 2\delta_{TV}(q_\phi, \delta),$$

where δ_{TV} denotes the Total Variation distance. We apply **Pinsker's Inequality**, which states $\delta_{TV}(P, Q) \leq \sqrt{\frac{1}{2} D_{KL}(P||Q)}$. Substituting this into the inequality:

$$\|\hat{s}_\phi - \hat{s}^*\| \leq R\sqrt{2 D_{KL}(\delta||q_\phi)}.$$

Squaring both sides yields the final bound on the MSE. □

Remark (Bound Tightness). Pinsker's inequality can be loose for categorical distributions over finite alphabets. For such cases, the tighter *Bretagnolle-Huber inequality* [Bretagnolle and Huber, 1979], $\delta_{TV}(P, Q) \leq \sqrt{1 - \exp(-D_{KL}(P||Q))}$, or method-of-types arguments [Csiszár and Körner, 2011] can be substituted without affecting the qualitative conclusions. Lemma 1 establishes the *existence and scaling* of the agnostic estimator gap; it is not a tight operational bound. In practice, $\mathcal{C}$ is drawn from a finite set of known constellation types (e.g., 16-QAM, 64-QAM). The KL term then equals $-\log q_\phi(\mathcal{C}^*|\mathbf{y})$, which is finite whenever q_ϕ assigns positive probability to $\mathcal{C}^*$.

A.2 Proof of Theorem 1 (Hypothesis Space Superiority)

Proof. The class of standard estimators $\mathcal{G}_{\text{std}}$ is a subset of the constellation-aware estimators $\mathcal{G}_{\text{CA}}$. For any function $g \in \mathcal{G}_{\text{std}}$, one can define a function $\psi \in \mathcal{G}_{\text{CA}}$ as $\psi(\mathbf{y}, \mathcal{C}) = g(\mathbf{y})$ for all $\mathcal{C}$. This function ψ ignores its second argument. Thus, $\mathcal{G}_{\text{std}} \subseteq \mathcal{G}_{\text{CA}}$. Since the infimum of a function over a superset cannot be larger than the infimum over a subset, it follows directly that $\text{MMSE}_{\text{CA}} \leq \text{MMSE}_{\text{std}}$. □

Remark (Finite Samples). The theorem is a population-level statement: conditioning on the true constellation cannot increase the MMSE. When the constellation is fixed, as in our experiments, the inequality holds with equality, because a model without $\mathcal{C}$ as input can hard-code it. The difference lies in how many pilots it takes to learn it, and here we only give a heuristic argument. Consider the memoryless case. An encoder that does not see $\mathcal{C}$ has to locate the K received centroids in $\mathbb{R}^d$ ($d = 2$), that is, estimate Kd coordinates from N_p pilots; even with labeled pilots, the squared error per centroid is of order Kd/N_p times the noise variance. CAT is given the transmitted constellation. It still learns an embedding of it and has to learn how the channel distorts it, but for the distortions in Section 2 (fading and I/Q imbalance) this map has only a handful of parameters, far fewer than Kd. We would therefore expect CAT to need fewer pilots, which is consistent with what we observe. This is not a proof, and a formal finite-sample separation remains open.

A.3 Proof of Proposition 1 (Structural Alignment with Optimal Receiver)

Proof.

1. **Matched Filter Alignment:** The posterior probability for a symbol c_k given an observation $\mathbf{z}$ corrupted by additive Gaussian noise $\mathbf{n} \sim \mathcal{N}(0, \sigma^2\mathbf{I})$ is proportional to $\exp\left(-\|\mathbf{z} - c_k\|^2/2\sigma^2\right)$. Expanding the squared Euclidean norm yields $-(\|\mathbf{z}\|^2 + \|c_k\|^2 - 2\mathbf{z}^T c_k)$. The term $\|\mathbf{z}\|^2$ is common to all classes and cancels out during Softmax. Assuming constant modulus symbols, $\|c_k\|^2$ is constant. The distribution is thus determined by the cross-correlation $\mathbf{z}^T c_k$. The CAT Attention mechanism computes $\text{Softmax}\left(\frac{(\mathbf{W}_Q\mathbf{z})^T(\mathbf{W}_K c_k)}{\sqrt{d}}\right)_k$. If the network learns whitening projections such that $\mathbf{W}_Q^T\mathbf{W}_K \approx (\sqrt{d}/\sigma^2)\,\mathbf{I}$, the attention scores approximate the Gaussian posterior probabilities.

2. **Block MMSE Equalizer Alignment:** The optimal linear equalizer for a block transmission is $\mathbf{W} = \mathbf{H}^H\mathbf{R}_{yy}^{-1}$. This requires: (a) *Deconvolution:* $\mathbf{H}^H$ represents the matched channel filter. For a multipath channel, this is a non-causal FIR filter. Our bidirectional FIR-FFN provides the structure to approximate this operation. (b) *Whitening:* The term $\mathbf{R}_{yy}^{-1}$ decorrelates the signal. The self-attention mechanism naturally computes cross-correlations $\mathbf{z}\mathbf{z}^T$ and can in principle learn whitening transformations.

□

Remark (Non-Constant Modulus Constellations). The proof above assumes constant modulus symbols (e.g., PSK). For QAM constellations with multiple energy levels, the energy term $\|c_k\|^2$ does not cancel and introduces an attention bias toward high-energy symbols. For 16-QAM, the raw attention logit bias between corner symbols ($\|c_k\|^2 = 18$) and inner symbols ($\|c_k\|^2 = 2$) is 0.95 nats before training, enough to distort detection noticeably. After training, the learned $\mathbf{W}_K$ projection absorbs an implicit energy normalization, reducing this bias to 0.04 nats. The residual is computed from the noiseless constellation tokens, and in our runs it was stable between 14 and 24 dB, varying by less than 0.005 nats. The worst-case posterior distortion at 0.04 nats is $|e^{0.04} - 1| \approx 4\%$, which is small compared with the uncertainty caused by noise. Empirically, SER decreases monotonically with SNR without plateauing, so we see no measurable SER floor.

B Scalability Limitations and Parameter Efficiency

Limitation: 1024-QAM and Beyond. For extreme-order constellations such as 1024-QAM ($K = 1024$), the $\mathcal{O}((N + K)^2 d)$ attention scaling becomes computationally prohibitive (overhead $\sim 800\%$ for $N = 512$). We consider 1024-QAM out of scope for the current architecture. For square QAM there is a cheaper representation: the constellation is the product of two $\sqrt{K}$-level PAM constellations, so the in-phase and quadrature levels can be represented by $2\sqrt{K}$ tokens instead of K.

For 1024-QAM at $N = 512$ this is 64 tokens, and the attention overhead drops from 800% to about 27%. The catch is that fading and I/Q imbalance mix the two components, so the received points no longer split into two PAM grids and the model has to learn that coupling from the signal tokens; we have not tested this. Other options reduce the cost of attention itself. Kernelized attention such as Performer [Choromanski et al., 2021] makes it linear in $N + K$, MobileViT [Mehta and Rastegari, 2022] combines convolutions with attention, and the efficient message-passing Transformer of Park et al. [2026] was designed for decoding error-correcting codes. In CAT these would matter most for the attention involving the constellation tokens, since that is the part that grows with K.

We note, however, that the 16-QAM through 256-QAM range covers the most critical operating regimes for neural equalization. While 1024-QAM is standardized in Wi-Fi 6 and 5G NR (Release 17), in practice it requires high SNR, typically with line of sight. The regimes where equalization is hardest (severe multipath, long delay spreads, cell edges) use lower modulation orders through adaptive modulation and coding (AMC).

Parameter Efficiency. Table 7 compares parameter counts and complexity. CAT has about 1k more parameters than the vanilla Transformer and roughly 15 times fewer than VAE-CNN, so its advantage over either is not a matter of model size.

Table 7: Parameter and complexity comparison (approximate for $N = 128$, $d = 10$).

Metric	**VAE-CNN**	**Vanilla Trans.**	**CAT (Ours)**
Parameter Count	~ 150k	~ 9k	~ 10k
Self-Attn Ops	0	$\mathcal{O}(N^2 d)$	$\mathcal{O}((N+K)^2 d)$
Pilot Efficiency	Low	Medium	**High**

C Additional Experimental Results

C.1 FIR Contribution Across Channels and SNRs

Table 8 compares CAT with the "CAT with MLP-FFN" variant, which keeps the constellation prior and the joint attention but replaces the FIR filter with a standard MLP. The MLP variant is worse in every setting. The gap is smallest at 14 dB, where noise rather than residual ISI limits performance.

Table 8: CAT vs. CAT with MLP-FFN on ISI channels ($N_p = 64$, payload 256). The CAT entries for $h^{(1)}$ are those of Table 3. FIR gain is the relative SER increase when the FIR-FFN is replaced by an MLP.

Channel	**SNR**	**CAT**	**CAT w/ MLP-FFN**	**FIR gain**
$h^{(1)}$ (L=5)	14 dB	0.1205	0.1284	7%
	17 dB	0.0198	0.0351	77%
	24 dB	0.0019	0.0025	32%
$h^{(2)}$ (L=4)	14 dB	0.0921	0.1195	30%
	17 dB	0.0340	0.0489	44%
	24 dB	0.0024	0.0041	71%

C.2 Convergence Speed and Meta-Learning Acceleration

The 5,000-step regime (Appendix F) is too slow for real-time systems such as 5G, where blocks last milliseconds. With a meta-learned initialization, however, CAT converges in far fewer steps. Table 9 shows convergence results using CAVIA [Zintgraf et al., 2019] on the $h^{(1)}$ channel at 17 dB with $N_p = 128$ pilots.

With 50 CAVIA-initialized steps, CAT's SER is only 2% above its 5,000-step result and lower than that of the vanilla Transformer and VAE-CNN trained from scratch for 5,000 steps (2.90×10^{-2} and 4.94×10^{-2}, Table 2). The baselines were not meta-trained, so this is not a matched-budget comparison, but it shows that CAT's adaptation can be cut to tens of steps.

Table 9: Convergence speed under different adaptation strategies ($h^{(1)}$, $E_x/N_0 = 17$ dB, $N_p = 128$ pilots, payload 256). The 5,000-step cold-start run is the $N_p = 128$ result of Table 2; the last column is the SER increase relative to it.

Steps	Method	SER	SER increase
5,000	Cold-start	1.56×10^{-2}	—
200	Cold-start	1.63×10^{-2}	$+4\%$
50	Cold-start	2.30×10^{-2}	$+47\%$
50	CAVIA-init	1.59×10^{-2}	$+2\%$
10	CAVIA-init	1.82×10^{-2}	$+17\%$

C.3 Wall-Clock Latency Profiling

Table 10 profiles per-step runtime on identical hardware (single RTX 4090, batch size 1) to provide matched latency comparisons.

Table 10: Wall-clock time of one training step (forward pass, backward pass and optimizer update; single RTX 4090, batch size 1).

Setting	VAE-CNN	Vanilla Trans.	CAT	Overhead
N=128, 16-QAM	11.2 μs	12.5 μs	16.0 μs	+28%
N=256, 16-QAM	18.1 μs	21.0 μs	24.5 μs	+17%
N=128, 64-QAM	11.2 μs	12.5 μs	24.8 μs	+98%

The per-step overhead decreases with N (as K/N shrinks) but grows with constellation order, consistent with the $\mathcal{O}((N+K)^2)$ analysis. For deployment, the time needed to reach a given SER matters more. CAT with 50 CAVIA-initialized steps already has a lower SER than the vanilla Transformer after 5,000 steps (Table 9). Since a 16-QAM CAT step costs only 17–28% more (Table 10), that is roughly 80 times less adaptation time, not counting the offline meta-training.

C.4 Robustness to Carrier Frequency Offset and Phase Noise

A purely static constellation prior $\mathcal{C}$ could degrade under residual carrier frequency offset (CFO) or oscillator phase noise, which rotate the received constellation. We address this by combining coarse PLL tracking with a learned residual phase rotation on constellation keys: $\tilde{\mathbf{c}}_k = \mathbf{R}(\hat{\theta})\mathbf{c}_k$, where $\hat{\theta}$ is estimated from aggregate attention weights. This exploits the fact that CFO-induced rotation is common to all constellation points. Under severe Wiener phase noise ($\sigma^2_\Delta = 10^{-3}$ rad^2/symbol, representative of mmWave oscillators), CAT with dynamic tracking maintains a 1.5 dB advantage over baselines with graceful degradation rather than an error floor.

C.5 Practical Use Cases and Future Research

In principle, if power amplifier (PA) characteristics are stable within a session, one could pre-train per transmitter class and fine-tune only the linear fading component at runtime. Our current experiments treat the end-to-end channel monolithically; future research can extend CAT to disentangle PA nonlinearity from propagation for reducing online training overhead.

We have evaluated CAT on block-fading channels and on the 3GPP TDL-A, TDL-C and TDL-D profiles. Future work should extend evaluation to high-mobility scenarios with Doppler spread using channel simulators like Sionna [Hoydis et al., 2022]. Connecting CAT to work on neural receivers, which replace entire receiver pipelines with learned components [Shlezinger et al., 2020, 2021a], is another direction. Finally, reducing the $\mathcal{O}((N+K)^2)$ attention cost, for example with the factorized or cheaper attention mechanisms of Appendix B, would extend CAT to constellations with $K \geq 1024$.

D Detailed Channel Models

D.1 Memoryless Nonlinear Channel

The memoryless nonlinear channel model used in our experiments, following Park et al. [2019], Burshtein and Bery [2023, 2024], consists of several stages. First, an ideal transmitted signal $\mathbf{x}_i = (x_i^I, x_i^Q)$ from a 16-QAM constellation is subjected to a nonlinear I/Q imbalance distortion (which mostly stems from hardware imperfections). This creates a distorted signal $\tilde{\mathbf{x}}_i = (\tilde{x}_i^I, \tilde{x}_i^Q)$ according to:

$$\begin{bmatrix} \tilde{x}_i^I \\ \tilde{x}_i^Q \end{bmatrix} = \begin{bmatrix} 1+\epsilon & 0 \\ 0 & 1-\epsilon \end{bmatrix} \begin{bmatrix} \cos\delta & -\sin\delta \\ -\sin\delta & \cos\delta \end{bmatrix} \begin{bmatrix} x_i^I \\ x_i^Q \end{bmatrix}.$$

The imbalance parameters, ϵ and δ, are constant for each transmission block but are randomly drawn from Beta distributions, specifically $\epsilon = 0.15\epsilon_0$ and $\delta = 15^\circ\delta_0$, where $\epsilon_0, \delta_0 \sim \text{Beta}(5, 2)$.

The resulting complex signal, $\tilde{x}_i^I + j\tilde{x}_i^Q$, is then transmitted over a Rayleigh flat-fading channel. The received complex signal is given by:

$$y_i^I + jy_i^Q = h(\tilde{x}_i^I + j\tilde{x}_i^Q) + n_i,$$

where $h \sim \mathcal{CN}(0, 1)$ is the complex channel gain, which is fixed for the duration of a block, and $n_i \sim \mathcal{CN}(0, \sigma^2)$ is the i.i.d. complex additive white Gaussian noise. The Signal-to-Noise Ratio (SNR) is defined as $10/\sigma^2$, based on the average power of the original 16-QAM constellation (normalized to average power of 10). The final received signal used by our models is the real-valued vector $\mathbf{y}_i = (y_i^I, y_i^Q)$.

D.2 Channels with Finite Memory (ISI)

For the experiments involving intersymbol interference, we adopt the channel model from Burshtein and Bery [2023, 2024], which is a two-stage process. First, the ideal signal $\mathbf{x}_i$ undergoes a memoryless nonlinear distortion $g(\cdot)$ to produce $\tilde{\mathbf{x}}_i = g(\mathbf{x}_i)$. For consistency, we use the same I/Q imbalance model described in the previous section for this nonlinearity.

The sequence of distorted signals is then transmitted through a noisy ISI channel. The received signal $\mathbf{y}_i$ is the result of a convolution between the complex channel impulse response $\mathbf{h}$ and the distorted signal sequence, corrupted by additive noise:

$$y_i^I + jy_i^Q = \sum_{l=0}^{L-1} h_l(\tilde{x}_{i-l}^I + j\tilde{x}_{i-l}^Q) + n_i,$$

where L is the length of the channel impulse response and $n_i \sim \mathcal{CN}(0, \sigma^2)$ is complex AWGN. Unless stated otherwise, the noise variance σ^2 is set to achieve a target SNR of $E_x/N_0 = 17$ dB.

We adopt three standard channel models from Burshtein and Bery [2024], with lengths $L = 5, 4, 10$ respectively:

$$\begin{aligned}
\mathbf{h}_1 &= [0.0545 + 0.05j, 0.2832 - 0.11971j, -0.7676 + 0.2788j, -0.0641 - 0.0576j, \\
&\quad 0.0466 - 0.02275j], \\
\mathbf{h}_2 &= [0.0554 + 0.0165j, -1.3449 - 0.4523j, 1.0067 + 1.1524j, \\
&\quad 0.3476 + 0.3153j], \\
\mathbf{h}_3 &= [0.0410 + 0.0109j, 0.0495 + 0.0123j, 0.0672 + 0.017j, 0.0919 + 0.0235j, \\
&\quad 0.7920 + 0.1281j, 0.396 + 0.0871j, 0.2715 + 0.048j, \\
&\quad 0.2291 + 0.0415j, 0.1287 + 0.0154j, 0.1032 + 0.0119j].
\end{aligned}$$

A longer impulse response corresponds to more severe ISI and a more challenging equalization task.

E Derivation of the Semi-Supervised VAE Loss

The loss function for the semi-supervised VAE is constructed to use both labeled (pilot) and unlabeled (payload) data. The goal is to maximize the log-likelihood of the observed data, which can be

expressed as a sum over the labeled and unlabeled sets:

$$\mathcal{L}_{\text{total}} = \sum_{i=1}^{N_p} \log p_\theta(\mathbf{y}_i|s_i) + \sum_{i=N_p+1}^{N} \log p_\theta(\mathbf{y}_i). \tag{5}$$

This is a generative objective. To incorporate the inference network q_ϕ, we also add a supervised cross-entropy term for the labeled data. The full objective combines these with weighting hyperparameters α and γ:

$$\begin{aligned}\mathcal{L}_{\text{full}} =& \frac{\alpha}{N_p} \sum_{i=1}^{N_p} \log q_\phi(s_i|\mathbf{y}_i) + \frac{\gamma}{N_p} \sum_{i=1}^{N_p} \log p_\theta(\mathbf{y}_i|s_i) \\ &+ \frac{1-\gamma}{N-N_p} \sum_{i=N_p+1}^{N} \log p_\theta(\mathbf{y}_i).\end{aligned} \tag{6}$$

The final term, $\log p_\theta(\mathbf{y}_i)$ for the unlabeled data, is intractable to compute directly as it requires marginalizing over all possible symbols s. We therefore substitute it with its Evidence Lower Bound (ELBO):

$$\log p_\theta(\mathbf{y}_i) \geq \mathbb{E}_{q_\phi(s|\mathbf{y}_i)}[\log p_\theta(\mathbf{y}_i, s) - \log q_\phi(s|\mathbf{y}_i)]. \tag{7}$$

By maximizing this lower bound (equivalent to minimizing its negative), we arrive at the final loss function used for training. After rearranging terms and using the fact that $p_\theta(\mathbf{y}_i, s) = p_\theta(\mathbf{y}_i|s)p(s)$, the negative ELBO becomes:

$$-\text{ELBO} = -\mathbb{E}_{q_\phi(s|\mathbf{y}_i)}[\log p_\theta(\mathbf{y}_i|s)] + D_{KL}(q_\phi(s|\mathbf{y}_i)||p(s)). \tag{8}$$

Substituting this into the full objective gives the final loss function presented in Eq. (2). For computational tractability, the expectation term is approximated using a single sample from $q_\phi(s|\mathbf{y}_i)$, often implemented with the Gumbel-Softmax reparameterization trick [Jang et al., 2017] to maintain differentiability.

F Implementation and Hyperparameter Details

F.1 CAT and Vanilla Transformer Implementation

Our Constellation-Aware Transformer (CAT) and the vanilla Transformer baseline share the same core configuration, differing only in their specific architectural components as described in Section 3.

Architecture. The models are built with a stack of 3 TransFIRmer (or standard Transformer) layers. The hidden dimension is set to $d_{\text{model}} = 10$, and we use Multi-Query Attention (MQA) [Shazeer, 2019] with a single attention head ($n_{\text{head}} = 1$) for efficiency. For the TransFIRmer layer's two-stream feed-forward network, the bidirectional FIR filter for the signal stream is implemented with two 1D convolutions (one for left-to-right and another for right-to-left convolutions), each using a kernel size of 12. The parallel MLP for the constellation stream uses a single hidden layer of width d_{model}. Dropout with a rate of $p = 0.1$ is applied within the attention and feed-forward sub-layers. Fixed sinusoidal positional embeddings are used for all experiments.

Training. The models are trained using the AdamW optimizer [Loshchilov and Hutter, 2019] with a learning rate of $\text{lr} = 10^{-3}$, betas of $(\beta_1, \beta_2) = (0.9, 0.999)$, and a weight decay of 0.01. We employ a linear learning rate scheduler that decays the learning rate from its initial value to zero over the course of training, which consists of a total of 5000 parameter update steps. The models are trained on mini-batches containing 16 pilot symbols and 32 payload symbols.

F.2 Hyperparameters for Semi-Supervised Learning

The training of all semi-supervised models (CAT, vanilla Transformer, VAE-CNN, etc.) is governed by the same set of hyperparameters and annealing schedules, ensuring a fair comparison and following the setup in Burshtein and Bery [2023, 2024].

SSL Loss Weighting. The composite loss function in Eq. (2) is balanced by two key hyperparameters. The term α, which weights the supervised cross-entropy loss on the encoder, is fixed at $\alpha = 0.2$. The term γ, which balances the supervised reconstruction loss against the unsupervised ELBO, is annealed over the training process. This annealing schedule gradually decreases γ_l (where l is the iteration index), shifting the training focus from the reliable pilot data to the more abundant but unlabeled payload data as the model becomes more confident. Specifically, we use $\gamma_l = 1/(1 + \beta_l)$, where $\beta_l = \min(2e^{0.0008(l-1)}, \beta_{\max})$, and $\beta_{\max} = \min((N - N_p)/N_p, 40)$. The value of γ_l is updated every 100 iterations.

Gumbel-Softmax Temperature. For models using the Gumbel-Softmax reparameterization trick (including our CAT and the VAE-CNN), the temperature τ is also annealed to transition from a soft, exploratory phase to a hard, decisive phase. The schedule is given by $\tau_l = \max(0.5, e^{-0.001(l-1)})$, with updates occurring every 100 iterations.

F.3 The Generative Model Architecture

For both our CAT and the vanilla Transformer baseline, we operate within the semi-supervised variational framework, which requires a generative model (or decoder), $p_\theta(\mathbf{y}|s)$, to model the forward channel process. To ensure a fair comparison with prior art, we adopt the generative model architecture directly from the VAE-CNN work in Burshtein and Bery [2023, 2024]. The parameters of this model are collectively denoted by θ. The specific architecture differs for memoryless and memory channels.

Memoryless Channels. For the memoryless channel, we model $p_\theta(\mathbf{y}_i|s_i)$ as an isotropic Gaussian distribution, $\mathcal{N}(\mathbf{y}_i; \mu_\theta(\mathbf{x}(s_i)), \sigma^2_\theta(\mathbf{x}(s_i))\mathbf{I})$. The mean μ_θ and log-variance $\log \sigma^2_\theta$ are produced by a decoder network. This network is a Multi-Layer Perceptron (MLP) with 3 hidden layers of dimensions $[64, 32, 16]$ and ReLU activations, which takes the ideal constellation signal $\mathbf{x}(s_i) \in \mathbb{R}^2$ as input and uses two separate linear heads to output the 2-dimensional mean and log-variance vectors.

Channels with Memory (ISI). For channels with memory, the generative model is designed to explicitly capture the two-stage process of a transmitter nonlinearity followed by a linear ISI channel. The model first applies a memoryless nonlinear function $g(\cdot)$, parameterized by a 2-layer MLP with hidden dimension 32, to each ideal symbol $\mathbf{x}_i$ in the input sequence $\mathbf{s}$ to produce a sequence of distorted signals $\tilde{\mathbf{x}}$. This sequence is then convolved with a learnable Finite Impulse Response (FIR) filter of length $L_{\text{max}} = 12$ taps, which models the complex channel impulse response $\mathbf{h}$. The real and imaginary parts of the filter taps are stored as two separate learnable parameter vectors. The output of this convolution provides the mean of the Gaussian distribution for the received sequence. The noise is modeled as i.i.d. Gaussian with a single learnable variance parameter σ^2. The SNR for ISI experiments is defined as E_x/N_0, where E_x is the average energy of the *undistorted* ideal symbols (before the nonlinearity $g(\cdot)$).

G Baseline Methodologies

G.1 Simple Decision Directed (SDD)

The SDD algorithm is a classical two-stage semi-supervised method [Burshtein and Bery, 2023, 2024].

1. **Initial Training:** A standard neural network decoder, $q_\phi(s|\mathbf{y})$, is first trained exclusively on the labeled pilot data $\{(\mathbf{y}_i, s_i)\}_{i=1}^{N_p}$ by minimizing the cross-entropy loss. Let the resulting parameters be $\hat{\phi}_0$.
2. **Pseudo-Labeling and Retraining:** The trained model is used to generate "hard" pseudo-labels for the unlabeled payload data: $\hat{s}_i = \text{argmax}_s\, q_{\hat{\phi}_0}(s|\mathbf{y}_i)$ for $i > N_p$. The model's parameters are then fine-tuned by training on a combined dataset of original pilots and pseudo-labeled payload data, minimizing a weighted cross-entropy loss.

G.2 Viterbi EM

The Viterbi EM algorithm [Dempster et al., 2018] is a hard-decision variant of the Expectation-Maximization (EM) algorithm, as described in Burshtein and Bery [2024]. It uses a generative model of the channel, $p_\theta(\mathbf{y}|s)$, parameterized by θ, and iterates between two steps:

1. **E-Step (Expectation):** Given the current estimate of the generative model's parameters $\theta^{(t-1)}$, generate hard decisions (pseudo-labels) for the payload data by choosing the most likely symbol according to the current model: $\hat{s}_i^{(t)} = \operatorname{argmax}_s p_{\theta^{(t-1)}}(\mathbf{y}_i|s)$.
2. **M-Step (Maximization):** Update the generative model's parameters by minimizing the reconstruction loss (negative log-likelihood) over a combined dataset of the original pilots and the newly generated pseudo-labels from the E-step, yielding $\theta^{(t)}$.

This process is repeated for a fixed number of iterations, gradually refining the channel model.

G.3 VAE-SSL (VAE-CNN)

This is the state-of-the-art semi-supervised method proposed in Burshtein and Bery [2023, 2024], which we refer to as VAE-CNN based on its typical implementation. It is a variational autoencoder-based framework that jointly trains two models:

- An **encoder** $q_\phi(s|\mathbf{y})$, which acts as the primary decoder. For channels with memory, this is typically implemented with a Convolutional Neural Network (CNN).
- A **decoder** $p_\theta(\mathbf{y}|s)$, which is a generative model that learns the forward channel process.

The two networks are trained simultaneously using a composite semi-supervised loss function (detailed in Eq. (2)) that combines a supervised objective on the pilot data with an unsupervised, Evidence Lower Bound (ELBO) objective on the payload data. This lets the model use the entire block, pilots and payload, to learn the channel.

G.4 CAVIA Meta-Learning

Fast Context Adaptation via Meta-Learning (CAVIA) [Zintgraf et al., 2019] is a meta-learning algorithm designed for rapid adaptation to new tasks. In our context, each channel realization is a "task".

- **Meta-Training:** The model is trained on data from a large number of previous channel blocks $\{(\mathbf{y}^{(m)}, s^{(m)})\}_{m=1}^{M}$. The goal is to learn a set of shared parameters ϕ that are common across all channels, while a small, task-specific "context vector" $\mathbf{z}^{(m)}$ is learned for each individual channel.
- **Meta-Testing (Adaptation):** When a new channel block arrives, the shared parameters ϕ are frozen. The model then rapidly infers a new context vector $\mathbf{z}_{\text{new}}$ by training only on the few available pilot symbols from the new block.
- **Decoding:** The final decoder uses both the shared parameters ϕ and the adapted context vector $\mathbf{z}_{\text{new}}$ to decode the payload data of the new block.

CAVIA learns a general model that can be specialized quickly, which makes it effective with few pilots, provided that past channel data is available.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction claim that CAT (1) injects geometric inductive biases into equalization, (2) is theoretically aligned with optimal receivers, and (3) achieves state-of-the-art performance. These are directly supported by the proofs in Section 3.1 and experimental results in Section 4. Limitations regarding 1024-QAM scalability are explicitly stated.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 5 and Section 3.4 explicitly discuss: (1) the $\mathcal{O}((N+K)^2)$ attention scaling limitation for 1024-QAM and beyond, (2) computational cost of per-block test-time training, (3) the need for future evaluation under high-mobility Doppler scenarios, and (4) the monolithic treatment of PA nonlinearity and channel propagation.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Lemma 1, Theorem 1 and Proposition 1 are stated with their assumptions; the main text gives proof sketches and Appendix A the full proofs. Corollary 1 is informal: we do not prove it, and the finite-sample argument in Appendix A is heuristic. We also discuss the tightness of Pinsker's inequality and the Bretagnolle-Huber alternative, and the case of non-constant modulus constellations.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: All channel models are specified analytically in Appendix D, including exact filter tap values. Complete hyperparameter details (optimizer, learning rate, annealing schedules, architecture dimensions, kernel sizes) are provided in Appendix F. Baseline methodologies are described in Appendix G. The experimental setup in Section 4.1 specifies the number of Monte Carlo trials (1000), SNR ranges, and evaluation metrics.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: Code will be released upon acceptance. However, all channel models are fully specified analytically (Appendix D), and the architecture and training procedure are described in sufficient detail for independent reimplementation.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 4.1 provides the experimental setup including channel types, constellation, model architecture, and baselines. Appendix F provides complete training details: optimizer (AdamW), learning rate (10^{-3}), weight decay, annealing schedules for γ and Gumbel-Softmax temperature, batch composition, and total training steps (5000). The generative model architecture is detailed in Appendix F.3.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Results are averaged over 1000 independent Monte Carlo trials (Section 4.1). The ablation study (Table 4) reports 95% confidence intervals for all model variants. The main text notes that confidence intervals were computed across all experiments and found to be negligible, confirming statistical significance.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Table 10 provides wall-clock latency profiling on a single NVIDIA RTX 4090 GPU with batch size 1 for multiple configurations. Training times are derivable from the reported per-step latency and total step count (5000 steps). The model's small parameter count ($\sim$10k) makes it feasible to run on consumer hardware.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: The research conforms to the NeurIPS Code of Ethics. It involves no human subjects, no personally identifiable data, and no dual-use concerns beyond those inherent to incremental communications research. The Impact Statement discusses both benefits and risks.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The Impact Statement discusses positive impacts (improved spectral efficiency, connectivity, resource efficiency) and potential risks (surveillance applications, though noting this is an incremental advance in mature technology). Broader context is provided regarding responsible deployment in critical infrastructure.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper proposes a small ($\sim$10k parameter) signal processing architecture for channel equalization. It does not release pre-trained models, large language models, image generators, or scraped datasets. The model has no meaningful risk for misuse beyond standard communications research.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.

- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All channel models and baseline methods are properly cited to their original publications [Burshtein and Bery, 2023, 2024, Park et al., 2019]. The CAVIA meta-learning algorithm is cited [Zintgraf et al., 2019]. No external datasets or pre-trained models are used; all experiments use synthetically generated data from analytically specified channel models.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release new datasets, models, or code as assets at submission time. Code will be released upon acceptance.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

Justification: This paper does not involve crowdsourcing or research with human subjects. All experiments use synthetically generated communication signals.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: This paper does not involve research with human subjects. All experiments are computational simulations using synthetic data.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs are not used as a component of the core methodology. The proposed CAT architecture is a small Transformer-based signal processing model for channel equalization, not involving any large language model.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.