无线先验对多用户多天线预编码学习计算与能耗开销的影响

丛鹏宇[①] 韩圣千[①] 邓鸣宇[①] 刘胜杰[①] 杨晨阳[①] 沈嵩辉[②]

[①](北京航空航天大学电子信息工程学院 北京 100191)

[②](中国移动通信有限公司研究院 北京 100053)

摘 要：随着移动通信系统中天线阵列规模的持续增大，多用户多输入多输出(Multiuser Multi-Input Multi-Output, MU-MIMO)预编码在显著提高系统性能的同时，也带来了高计算复杂度与高能耗等挑战。本文针对MU-MIMO预编码策略学习问题，研究在神经网络结构设计中引入无线策略数学先验(简称“无线先验”)知识对深度学习计算和能耗的影响。为此，首先分析MU-MIMO预编码策略的多维联合置换等变性和不变性，在此基础上构建一种注意力图神经网络(Attention-based Graph Neural Network, AGNN)；然后，基于3GPP标准信道数据集开展仿真与硬件实测，对AGNN与传统数值算法和Transformer模型进行比较。研究结果表明，AGNN在频谱效率、计算复杂度和推理能耗方面均优于仅满足一维置换等变性的Transformer模型，表明充分利用无线先验对高效学习的重要影响。同时，随着天线规模与带宽的扩大，传统数值算法的计算复杂度和能耗增长速度明显高于性能增益；而AGNN能够在提高系统性能的同时，显著降低推理时间与推理能耗，表明融入无线先验的学习方法是未来6G预编码优化的有效技术路径。

关键词：6G；预编码；无线先验；图神经网络；计算能耗

文献标识码：A **文章编号**：1009-5896(2026)00-0001-12

DOI: 10.11999/JEIT260388 **CSTR**: 32379.14.JEIT260388

1 引言

多用户多输入多输出(Multiuser Multi-Input Multi-Output, MU-MIMO)技术是提高系统频谱效率的重要手段[1,2]。传统基于数值优化的MU-MIMO预编码与用户配对算法，如迫零块对角化(zero forcing block diagonalization, ZFBD)和贪婪配对算法等，虽然性能优异，但其计算复杂度随基站天线数和用户数的增加呈幂次增长，导致推理能耗与推理时间快速上升。面向未来第六代移动通信(6G)系统对超高速率、海量连接以及绿色可持续发展的需求，MU-MIMO预编码优化面临严峻的计算复杂度与能耗挑战[1]。

近年来，深度学习方法被广泛用于无线通信的复杂资源优问题化[3,4]。这类方法通过采用“离线训练、在线推理”的范式，能够大幅度降低在线推理阶段的计算复杂度和时间。然而，现有工作在评估基于深度学习的预编码策略(即从信道到预编码的映射)复杂度时，通常把训练和推理时长或浮点运算数(FLOPs)作为评估指标，缺少对于训练与推理过程在实际硬件平台上所消耗能量的直接实测验证[5-8]。

深度学习模型的训练和推理复杂度与网络结构设计密切相关[5]。若采用未引入任何无线策略数学先验(简称“无线先验”)的全连接神经网络学习预编码策略，训练复杂度高且性能欠佳[6]。网络结构与策略数学性质不匹配同样会导致此类问题，例如采用满足平移等变性的卷积神经网络学习满足置换等变性的预编码策略[6]。MU-MIMO预编码策略本身具有复杂的多维置换等变性[7]。为了在降低训练复杂度的同时提高学习性能和泛化能力，在设计网络结构时应合理利用无线先验[8]。例如，在自然语言处理等领域表现卓越的Transformer模型具备一维置换等变性，因此能够利用用户的置换等变性[8]。

深度神经网络的能耗开销通常受网络规模、硬件架构以及数据访问模式等多种因素影响。文献[9]针对超大规模MIMO预编码问题，分析了深度学习模型的计算能耗、权重及激活函数的输入/输出在内存与计算组件间传输的能耗等，并据此构建了理论能耗模型。尽管该模型对能耗来源进行了细分，但其准确性缺乏实测数据验证。文献[10]指出，对于大规模神经网络，频繁的片外内存访问是主要能耗来源。然而，基于无线先验设计的神经网络通常规模较小，其内存访问能耗的重要性仍待进一步明确。文献[11]提出了一种基于晶体管操作数的能耗模型，把深度学习中的运算映射为晶体管级计算任务，并通过在GPU和CPU平台上的实测数据建立能耗与FLOPs和晶体管操作数之间的拟合关系。然而，该模型未考虑数据访问能耗，且模型参数分别针对全连接网络和卷积神经网络独立拟合，不同网络类型之间拟合参数差异存在较大差异。因此，对于Transformer和GNN等新兴网络架构，该模型难以对能耗进行准确的评估。

收稿日期：2026-xx-xx；改回日期：2026-07-15；网络出版：2026-07-25

本文对不同数值与人工智能(Artificial Intelligence, AI)预编码算法的计算复杂度及能耗开销开展实际测量，在此基础上研究在网络结构设计中引入无线先验的影响。本文的主要贡献如下：

(1)揭示了MU-MIMO预编码策略在基站天线维和用户维满足的置换等变性、以及在用户天线维满足的置换不变性，在此基础上构建了基于用户间注意力机制的多维联合置换等变图神经网络(Attention-based Graph Neural Network, AGNN)，并将其扩展至宽带通信场景。

(2)针对现有预编码策略学习在复杂度评估中忽略计算能耗的问题，搭建了面向GPU、CPU和DRAM的组件级能耗测量平台，实现了不同通信场景、系统规模及运行阶段下，多种数值优化和学习算法的组件级功耗和能耗评估。

(3)仿真与测试结果表明，仅利用一维置换等变性的Transformer参数规模庞大，计算复杂度与能耗开销很高，且频谱效率低于传统数值算法；融入多维置换等变先验的AGNN在频谱效率超过传统数值算法的同时，显著降低了推理时间与能耗；随着天线规模和带宽的扩大，传统数值算法的计算复杂度和能耗的增长明显快于系统性能的提高，而AGNN能够以更低的推理时间和能耗有效应对大规模系统的资源优化挑战。

2 系统模型

考虑下行多用户MU-MIMO系统，其中基站装有N_{T}根发射天线，服务K个用户，每个用户有N_{R}根接收天线。第k个用户的接收信号可以表示为：

$$\boldsymbol{y}_k=\boldsymbol{H}_k\boldsymbol{V}_k^{\mathrm{H}}\boldsymbol{s}_k+\boldsymbol{H}_k\sum_{j=1,j\neq k}^{K}\boldsymbol{V}_j^{\mathrm{H}}\boldsymbol{s}_j+\boldsymbol{n}_k \tag{1}$$

其中，$\boldsymbol{H}_k,\boldsymbol{V}_k\in\mathbb{C}^{N_{\mathrm{R}}\times N_{\mathrm{T}}}$分别表示第$k$个用户的信道矩阵和预编码矩阵，$\boldsymbol{s}_k,\boldsymbol{n}_k\in\mathbb{C}^{N_{\mathrm{R}}\times 1}$分别表示第$k$个用户的发射信号和接收噪声。为了便于表述，把所有用户的信道矩阵记为$\boldsymbol{H}=\left[\boldsymbol{H}_1^{\mathrm{T}},\cdots,\boldsymbol{H}_K^{\mathrm{T}}\right]^{\mathrm{T}}\in\mathbb{C}^{N_{\mathrm{R}}K\times N_{\mathrm{T}}}$，把所有用户的预编码矩阵记为$\boldsymbol{V}=\left[\boldsymbol{V}_1^{\mathrm{T}},\cdots,\boldsymbol{V}_K^{\mathrm{T}}\right]^{\mathrm{T}}\in\mathbb{C}^{N_{\mathrm{R}}K\times N_{\mathrm{T}}}$。

不失一般的，考虑如下面向和速率最大化的预编码优化问题：

$$\begin{aligned}\max_{\boldsymbol{V}}R=\sum_{k=1}^{K}R_k=\sum_{k=1}^{K}\log_2\Bigg|&\boldsymbol{I}_{N_{\mathrm{R}}}+\boldsymbol{H}_k\boldsymbol{V}_k^{\mathrm{H}}\boldsymbol{V}_k\boldsymbol{H}_k^{\mathrm{H}}\\&\cdot\left(\sum_{j=1,j\neq k}^{K}\boldsymbol{H}_k\boldsymbol{V}_j^{\mathrm{H}}\boldsymbol{V}_j\boldsymbol{H}_k^{\mathrm{H}}+\sigma^2\boldsymbol{I}_{N_{\mathrm{R}}}\right)^{-1}\Bigg|\\ \text{s.t. }&||\boldsymbol{V}||^2\leq P_{\max}\end{aligned} \tag{2}$$

其中，$||\cdot||$表示Frobenius范数，$P_{\max}$为基站最大发射功率，σ^2为噪声方差。

预编码策略为从所有用户的信道矩阵到最优预编码矩阵的映射，记为：

$$\boldsymbol{V}=f(\boldsymbol{H}) \tag{3}$$

深度学习方法通过构建并训练神经网络对预编码策略$f(\cdot)$进行拟合，从而在推理阶段在给定信道输入时直接输出对应的预编码。

3 基于无线先验的预编码学习方法

3.1 MU-MIMO预编码策略的多维联合置换性质

式(3)中的预编码策略$f(\cdot)$满足多维联合置换等变与不变性。具体而言，当用户顺序、用户天线顺序和基站天线顺序发生置换时，信道矩阵$\boldsymbol{H}$的行和列会发生相应置换，最优预编码矩阵需要作出相应的行和列置换或保持不变，才能保证式(2)中的优化目标不变。下面分别讨论三类置换性质。

(1)**用户置换**：当用户顺序发生由置换矩阵$\boldsymbol{\Pi}_K\in\mathbb{R}^{K\times K}$表示的置换时，信道矩阵$\boldsymbol{H}$变为$(\boldsymbol{\Pi}_K\otimes\boldsymbol{I}_{N_{\mathrm{R}}})\boldsymbol{H}$。若最优预编码矩阵$\boldsymbol{V}$同样发生对应的置换，记为$(\boldsymbol{\Pi}_K\otimes\boldsymbol{I}_{N_{\mathrm{R}}})\boldsymbol{V}$，则式(2)中的目标函数保持不变，其中$\otimes$表示Kronecker积。

(2)**用户天线置换**：当第k个用户的天线发生由矩阵$\boldsymbol{\Pi}_{N_{\mathrm{R}},k}\in\mathbb{R}^{N_{\mathrm{R}}\times N_{\mathrm{R}}}$表示的置换时，信道矩阵$\boldsymbol{H}$转变为$\mathrm{diag}\left(\boldsymbol{\Pi}_{N_{\mathrm{R}},1},\cdots,\boldsymbol{\Pi}_{N_{\mathrm{R}},K}\right)\boldsymbol{H}$。若最优预编码矩阵不作改变，则式(2)中第$k$个用户的速率可以表示为：

$$\begin{aligned}&\log_2\Bigg|\boldsymbol{I}_{N_{\mathrm{R}}}+\boldsymbol{\Pi}_{N_{\mathrm{R}},k}\boldsymbol{H}_k\boldsymbol{V}_k^{\mathrm{H}}\boldsymbol{V}_k\boldsymbol{H}_k^{\mathrm{H}}\boldsymbol{\Pi}_{N_{\mathrm{R}},k}^{\mathrm{H}}\\&\quad\cdot\left(\sum_{j=1,j\neq k}^{K}\boldsymbol{\Pi}_{N_{\mathrm{R}},k}\boldsymbol{H}_k\boldsymbol{V}_j^{\mathrm{H}}\boldsymbol{V}_j\boldsymbol{H}_k^{\mathrm{H}}\boldsymbol{\Pi}_{N_{\mathrm{R}},k}^{\mathrm{H}}+\sigma^2\boldsymbol{I}_{N_{\mathrm{R}}}\right)^{-1}\Bigg|\\&=\log_2\Bigg|\boldsymbol{I}_{N_{\mathrm{R}}}+\boldsymbol{H}_k\boldsymbol{V}_k^{\mathrm{H}}\boldsymbol{V}_k\boldsymbol{H}_k^{\mathrm{H}}\\&\quad\cdot\left(\sum_{j=1,j\neq k}^{K}\boldsymbol{H}_k\boldsymbol{V}_j^{\mathrm{H}}\boldsymbol{V}_j\boldsymbol{H}_k^{\mathrm{H}}+\sigma^2\boldsymbol{I}_{N_{\mathrm{R}}}\right)^{-1}\Bigg|=R_k\end{aligned} \tag{4}$$

可见，用户天线顺序的置换不改变用户速率，因此最优预编码矩阵对用户天线顺序置换保持不变。

(3) **基站天线置换**：当基站天线发生由置换矩阵$\boldsymbol{\Pi}_{N_{\mathrm{T}}}\in\mathbb{R}^{N_{\mathrm{T}}\times N_{\mathrm{T}}}$表示的置换时，信道矩阵$\boldsymbol{H}$变为$\boldsymbol{H}\boldsymbol{\Pi}_{N_{\mathrm{T}}}$。若预编码矩阵相应变为$\boldsymbol{V}\boldsymbol{\Pi}_{N_{\mathrm{T}}}$，则式(2)中第$k$个用户的速率变为：

$$\begin{aligned}&\log_2\left|\boldsymbol{I}_{N_{\mathrm{R}}}+\boldsymbol{H}_k\boldsymbol{\Pi}_{N_{\mathrm{T}}}\boldsymbol{\Pi}_{N_{\mathrm{T}}}^{\mathrm{H}}\boldsymbol{V}_k^{\mathrm{H}}\boldsymbol{V}_k\boldsymbol{\Pi}_{N_{\mathrm{T}}}\boldsymbol{\Pi}_{N_{\mathrm{T}}}^{\mathrm{H}}\boldsymbol{H}_k^{\mathrm{H}}\right.\\&\cdot\left(\sum_{j=1,j\neq k}^{K}\boldsymbol{H}_k\boldsymbol{\Pi}_{N_{\mathrm{T}}}\boldsymbol{\Pi}_{N_{\mathrm{T}}}^{\mathrm{H}}\boldsymbol{V}_j^{\mathrm{H}}\boldsymbol{V}_j\boldsymbol{\Pi}_{N_{\mathrm{T}}}\boldsymbol{\Pi}_{N_{\mathrm{T}}}^{\mathrm{H}}\boldsymbol{H}_k^{\mathrm{H}}\right.\\&\left.\left.+\sigma^2\boldsymbol{I}_{N_{\mathrm{R}}}\right)^{-1}\right|\\&=\log_2\left|\boldsymbol{I}_{N_{\mathrm{R}}}+\boldsymbol{H}_k\boldsymbol{V}_k^{\mathrm{H}}\boldsymbol{V}_k\boldsymbol{H}_k^{\mathrm{H}}\right.\\&\left.\cdot\left(\sum_{j=1,j\neq k}^{K}\boldsymbol{H}_k\boldsymbol{V}_j^{\mathrm{H}}\boldsymbol{V}_j\boldsymbol{H}_k^{\mathrm{H}}+\sigma^2\boldsymbol{I}_{N_{\mathrm{R}}}\right)^{-1}\right|=R_k\end{aligned}\tag{5}$$

可见，预编码矩阵相应置换时，式(2)中的目标函数保持不变。

根据以上的分析，当信道矩阵在用户、用户天线和基站天线三个集合发生置换时，信道矩阵由$\boldsymbol{H}$变为$\mathrm{diag}\left(\boldsymbol{\Pi}_{N_{\mathrm{R}},1},\cdots,\boldsymbol{\Pi}_{N_{\mathrm{R}},K}\right)\left(\boldsymbol{\Pi}_K\otimes\boldsymbol{I}_{N_{\mathrm{R}}}\right)\boldsymbol{H}\boldsymbol{\Pi}_{N_{\mathrm{T}}}$；相应地，最优预编码矩阵由$\boldsymbol{V}$变为$\left(\boldsymbol{\Pi}_K\otimes\boldsymbol{I}_{N_{\mathrm{R}}}\right)\boldsymbol{V}\boldsymbol{\Pi}_{N_{\mathrm{T}}}$。可见，最优预编码策略$f(\cdot)$满足如下的多维联合置换等变与不变性：

$$\begin{aligned}\left(\boldsymbol{\Pi}_K\otimes\boldsymbol{I}_{N_{\mathrm{R}}}\right)\boldsymbol{V}\boldsymbol{\Pi}_{N_{\mathrm{T}}}=&f\left(\mathrm{diag}\left(\boldsymbol{\Pi}_{N_{\mathrm{R}},1},\cdots,\boldsymbol{\Pi}_{N_{\mathrm{R}},K}\right)\right.\\&\left.\cdot\left(\boldsymbol{\Pi}_K\otimes\boldsymbol{I}_{N_{\mathrm{R}}}\right)\boldsymbol{H}\boldsymbol{\Pi}_{N_{\mathrm{T}}}\right)\end{aligned}\tag{6}$$

3.2 满足多维联合置换等变与不变性的AGNN

为了利用策略的置换性质，已有工作提出了多种类型的GNN，如EGNN[6]、GAT[12]、模型GNN[13]、梯度GNN[14]等。然而，相关方法往往考虑用户仅配备单天线，因此难以适配多天线用户场景。文献[8]给出两种基于Transformer架构的预编码学习方法，其中把用户作为词元(token)，把用户的信道矩阵作为词元特征。尽管这种方法可以用于学习MU-MIMO预编码，但模型仅对用户维度满足一维置换等变性。

合理利用策略的置换性质，可在网络设计中引入归纳偏置，通过参数共享压缩假设空间。Transformer虽然利用了用户维置换等变性，相比传统全连接或卷积神经网络有一定增益，但未充分利用基站和用户天线维度的置换等变和不变性，未能最大化压缩假设空间。为了充分利用MU-MIMO预编码策略的性质，构建由用户、用户天线、基站天线三类顶点构成的超图，其中三类任意不同顶点之间均有一条超边。

本节设计满足多维联合置换等变与不变性的GNN。参照文献[7]中的多维GNN设计框架，在更新GNN的边特征时，只需聚合邻接超边特征。由此可得GNN第l层的更新方程为：

$$\begin{aligned}\boldsymbol{x}_{n_{\mathrm{R}},k,n_{\mathrm{T}}}^{(l)}=\phi^{(l)}\Bigg\{&\boldsymbol{x}_{n_{\mathrm{R}},k,n_{\mathrm{T}}}^{(l-1)}\boldsymbol{P}_1^{(l)}+\left(\frac{1}{N_{\mathrm{T}}}\sum_{n'_{\mathrm{T}}=1}^{N_{\mathrm{T}}}\boldsymbol{x}_{n_{\mathrm{R}},k,n'_{\mathrm{T}}}^{(l-1)}\right)\\&\boldsymbol{P}_2^{(l)}+\left(\frac{1}{N_{\mathrm{R}}}\sum_{n'_{\mathrm{R}}=1}^{N_{\mathrm{R}}}\boldsymbol{x}_{n'_{\mathrm{R}},k,n_{\mathrm{T}}}^{(l-1)}\right)\boldsymbol{P}_3^{(l)}\\&+\left(\frac{1}{K}\sum_{k'=1}^{K}\boldsymbol{x}_{n_{\mathrm{R}},k',n_{\mathrm{T}}}^{(l-1)}\right)\boldsymbol{P}_4^{(l)}\Bigg\}\end{aligned}\tag{7}$$

其中$\boldsymbol{P}_i^{(l)}\in\mathbb{R}^{n_{\mathrm{f}}^{(l-1)}\times n_{\mathrm{f}}^{(l)}}\ (i=1,2,\cdots,4)$为可训练权重矩阵，$\boldsymbol{x}_{n_{\mathrm{R}},k,n_{\mathrm{T}}}^{(l)}\in\mathbb{R}^{1\times n_{\mathrm{f}}^{(l)}}$为由用户天线顶点$n_{\mathrm{R}}$、用户顶点$k$和基站天线顶点$n_{\mathrm{T}}$构成的超边对应的隐藏特征向量，$n_{\mathrm{f}}^{(l)}$为第$l$层输出的隐藏表示数，$\phi^{(l)}(\cdot)$为激活函数。

在式(7)中，边特征更新默认所有邻边贡献相同。但在MU-MIMO预编码中，多用户间干扰是性能瓶颈，而用户间信道相关性能够有效反映用户间干扰程度。因此，采用用户间信道相关值来计算注意力系数，对邻边特征加权聚合。引入用户间注意力机制后的更新方程如下：

$$\begin{aligned}\boldsymbol{x}_{n_{\mathrm{R}},k,n_{\mathrm{T}}}^{(l)}=\phi^{(l)}\Bigg\{&\boldsymbol{x}_{n_{\mathrm{R}},k,n_{\mathrm{T}}}^{(l-1)}\boldsymbol{P}_1^{(l)}+\left(\frac{1}{N_{\mathrm{T}}}\sum_{n'_{\mathrm{T}}=1}^{N_{\mathrm{T}}}\boldsymbol{x}_{n_{\mathrm{R}},k,n'_{\mathrm{T}}}^{(l-1)}\right)\\&\boldsymbol{P}_2^{(l)}+\left[\frac{1}{N_{\mathrm{R}}}\sum_{n'_{\mathrm{R}}=1}^{N_{\mathrm{R}}}\boldsymbol{\alpha}_{kn'_{\mathrm{R}}\to kn_{\mathrm{R}}}^{(l)}\odot\left(\boldsymbol{x}_{n'_{\mathrm{R}},k,n_{\mathrm{T}}}^{(l-1)}\boldsymbol{P}_3^{(l)}\right)\right]\boldsymbol{Q}_1^{(l)}\\&+\left[\frac{1}{(K-1)N_{\mathrm{R}}}\sum_{k'=1,k'\neq k}^{K}\sum_{n'_{\mathrm{R}}=1}^{N_{\mathrm{R}}}\boldsymbol{\alpha}_{k'n'_{\mathrm{R}}\to kn_{\mathrm{R}}}^{(l)}\right.\\&\left.\odot\left(\boldsymbol{x}_{n'_{\mathrm{R}},k',n_{\mathrm{T}}}^{(l-1)}\boldsymbol{P}_4^{(l)}\right)\right]\boldsymbol{Q}_2^{(l)}\Bigg\}\\\boldsymbol{\alpha}_{k'n'_{\mathrm{R}}\to kn_{\mathrm{R}}}^{(l)}=\tanh\Bigg[&\frac{1}{N_{\mathrm{T}}}\sum_{n_{\mathrm{T}}=1}^{N_{\mathrm{T}}}\left(\boldsymbol{x}_{n'_{\mathrm{R}},k',n_{\mathrm{T}}}^{(l-1)}\boldsymbol{P}_5^{(l)}\right)\\&\odot\left(\boldsymbol{x}_{n_{\mathrm{R}},k,n_{\mathrm{T}}}^{(l-1)}\boldsymbol{P}_6^{(l)}\right)\Bigg]\end{aligned}\tag{8}$$

其中$\boldsymbol{\alpha}_{k'n'_{\mathrm{R}}\to kn_{\mathrm{R}}}^{(l)}\in\mathbb{R}^{1\times n_{\mathrm{f}}^{(l)}}$为第$k'$个用户第$n'_{\mathrm{R}}$根天线边特征向第$k$个用户第$n_{\mathrm{R}}$根天线边特征聚合的注意力系数向量，用于度量用户之间的相关性，$\odot$表示Hadamard积，$\boldsymbol{P}_i^{(l)}\in\mathbb{R}^{n_{\mathrm{f}}^{(l-1)}\times n_{\mathrm{f}}^{(l)}}\ (i=1,\cdots,6)$和$\boldsymbol{Q}_1^{(l)},\boldsymbol{Q}_2^{(l)}\in\mathbb{R}^{n_{\mathrm{f}}^{(l)}\times n_{\mathrm{f}}^{(l)}}$为可训练权重矩阵。基于式(8)的网络结构称为注意力图神经网络(AGNN)。

AGNN的第一层输入特征向量$\boldsymbol{x}_{n_{\mathrm{R}},k,n_{\mathrm{T}}}^{(0)}$由第$k$个用户的第$n_{\mathrm{R}}$根天线到基站第$n_{\mathrm{T}}$根天线的信道系数$h_{n_{\mathrm{R}},k,n_{\mathrm{T}}}$构成，表示为$\boldsymbol{x}_{n_{\mathrm{R}},k,n_{\mathrm{T}}}^{(0)}=\left[\mathrm{Re}\left(h_{n_{\mathrm{R}},k,n_{\mathrm{T}}}\right),\mathrm{Im}\left(h_{n_{\mathrm{R}},k,n_{\mathrm{T}}}\right)\right]\in\mathbb{R}^{1\times 2}$。AGNN最后一层的输出特征向量$\boldsymbol{x}_{n_{\mathrm{R}},k,n_{\mathrm{T}}}^{(L)}$用于构成第$k$个用户的第$n_{\mathrm{R}}$根天线到基站的第$n_{\mathrm{T}}$根天线的预编码系数$v_{n_{\mathrm{R}},k,n_{\mathrm{T}}}$，即$\boldsymbol{x}_{n_{\mathrm{R}},k,n_{\mathrm{T}}}^{(L)}=\left[\mathrm{Re}\left(v_{n_{\mathrm{R}},k,n_{\mathrm{T}}}\right),\mathrm{Im}\left(v_{n_{\mathrm{R}},k,n_{\mathrm{T}}}\right)\right]\in\mathbb{R}^{1\times 2}$，其

中L为AGNN总层数。为了满足发射功率约束$||\boldsymbol{V}||^2 \le P_{\max}$，并考虑到最大化和速率的预编码需使用全部发射功率，AGNN的最后一层采用激活函数 $\boldsymbol{X}^{(L)} = \sqrt{P_{\max}}\widetilde{\boldsymbol{X}}^{(L)} / \left|\left|\widetilde{\boldsymbol{X}}^{(L)}\right|\right|^2$，其中$\widetilde{\boldsymbol{X}}^{(L)}$, $\boldsymbol{X}^{(L)} \in \mathbb{C}^{N_{\mathrm{R}}\times K\times N_{\mathrm{T}}\times 2}$分别为该层激活函数的输入和输出。

可以验证，采用式(8)所示更新公式构建的AGNN能够满足式(6)定义的多维联合置换等变与不变性。

3.3 AGNN的推理FLOPs

在AGNN中，第l层根据式(8) 更新边特征，涉及矩阵向量乘法、累加、Hadamard积、tanh等基本运算。其中，单次加法、乘法和指数运算所需FLOPs分别为1、α_{m}和α_{e}；对于$1\times n1\times n$维向量与$n\times m$维矩阵相乘，共需mn次乘法和$m(n-1)$ $m\,(n-1)$次加法，对应的FLOPs为$(\alpha_{\mathrm{m}}+1)mn-m$ $[(\alpha_{\mathrm{m}}+1)\,mn-m]$；单次tanh函数计算包含一次指数运算、两次乘法和两次加法，其FLOPs为$(\alpha_{\mathrm{e}}+2\alpha_{\mathrm{m}}+2)(\alpha_{\mathrm{e}}+2\alpha_{\mathrm{m}}+2)$。通过逐层累加式(8)各项基础运算的开销，可得整个AGNN在推理阶段的总FLOPs如下：

$$\begin{aligned} M_{\mathrm{AGNN}}^{\mathrm{MIMO}} = \sum_{l=1}^{L} (\alpha_{\mathrm{m}}+1)\left[2N_{\mathrm{T}}N_{\mathrm{R}}Kn_{\mathrm{f}}^{l2} + (5N_{\mathrm{T}}+1)\right. \\ \cdot N_{\mathrm{R}}Kn_{\mathrm{f}}^{l}n_{\mathrm{f}}^{l-1} + (2K+1)\,N_{\mathrm{T}}N_{\mathrm{R}}^{2}Kn_{\mathrm{f}}^{l}\big] \\ + \left[\alpha_{\mathrm{e}}N_{\mathrm{R}}^{2}K^{2} + (3\alpha_{\mathrm{m}}+1)\,N_{\mathrm{R}}^{2}K^{2} + (2\alpha_{\mathrm{m}}-1)\right. \\ \cdot N_{\mathrm{T}}N_{\mathrm{R}}K\big]\,n_{\mathrm{f}}^{l} + (\alpha_{\mathrm{m}}+N_{\mathrm{T}}-1)\,N_{\mathrm{R}}Kn_{\mathrm{f}}^{l-1} \end{aligned} \tag{9}$$

其中，LL是AGNN的层数。由于式(8)中激活函数$\phi^{(l)}(\cdot)$的计算量在总FLOPs中占比很小，故在式(9)中忽略。

从式(9)可知，AGNN的推理复杂度满足$O(N_{\mathrm{R}}^2K^2)$，随着用户数和用户天线数呈二次方增长。

3.4 策略扩展：学习宽带预编码

在宽带系统中，逐子载波计算预编码的计算复杂度高，而相邻子载波信道相关性较强。因此，在实际应用中常把多个子载波划分为一个子带，统一计算一个预编码矩阵。此时，预编码策略可以表示为：$\boldsymbol{V} = f(\boldsymbol{H}_1, \boldsymbol{H}_2, \cdots, \boldsymbol{H}_C)$，其中$\boldsymbol{H}_c$表示子带内第$c$个子载波上的所有用户信道矩阵，$C$表示子带大小。

与窄带预编码相比，宽带预编码策略的输入包含多子载波的用户信道矩阵，因此需要在AGNN输入层对网络结构进行扩展。具体而言，AGNN的第一层输入特征向量扩展为：

$$\begin{aligned} \boldsymbol{x}_{n_{\mathrm{R}},k,n_{\mathrm{T}}}^{\mathrm{WB},(0)} = [&\mathrm{Re}\left(h_{n_{\mathrm{R}},k,n_{\mathrm{T}},1}\right),\mathrm{Im}\left(h_{n_{\mathrm{R}},k,n_{\mathrm{T}},1}\right),\cdots, \\ &\mathrm{Re}\left(h_{n_{\mathrm{R}},k,n_{\mathrm{T}},C}\right),\mathrm{Im}\left(h_{n_{\mathrm{R}},k,n_{\mathrm{T}},C}\right)] \in \mathbb{R}^{1\times 2C} \end{aligned} \tag{10}$$

其中$h_{n_{\mathrm{R}},k,n_{\mathrm{T}},c}$表示在第$c$个子载波上第$k$个用户的第$n_{\mathrm{R}}$根天线到基站第$n_{\mathrm{T}}$根天线的信道。

为了让子带预编码适用于子带中的所有子载波，通常基于信道的频域相关性划分子带(如按相干带宽划分)以保证同一子带内各子载波的信道差异较小。当系统带宽增加时，子带大小C不变，而全频带内的子带数将相应增加。式(10)采用拼接方式扩展输入特征，适用于C较小且子带内信道相关性较强的场景。

4 计算能耗测量方法

为了准确评估不同预编码算法的计算能耗和功耗，分别独立采集GPU、CPU和DRAM的能耗或功耗数据，进而得到算法运行所需的总能耗和总功耗。需说明，本文测量结果只包含计算硬件本身的能耗，不包含散热系统、电源转换等基础设施的额外能耗。下面介绍测试的硬件平台、测量工具及测试流程。

4.1 测试平台

测试在一台高性能服务器上完成。表1汇总了主要硬件配置及GPU、CPU、DRAM的功耗预算和闲置功耗。GPU功耗预算采用显卡总功耗(Total Graphics Power, TGP)衡量，CPU 与DRAM以热设计功率(Thermal Design Power, TDP)衡量。CPU的TDP是指所有物理核激活并以基础频率持续工作时的平均功率。GPU的TGP和CPU的TDP数据均来自厂商官方公开信息，而DRAM因主流厂商通常不公开TDP，其参考值由第三方专业硬件检测工具AIDA64获取。

闲置功耗指硬件设备在无计算负载的空闲状态下维持运行所需的最小基础功耗。其具体测量方法

表 1 测试平台硬件设置

硬件组件	型号名称 × 数量	存储容量	TGP 或TDP	闲置功耗
GPU	NVIDIA GeForce RTX 4090 D GPU×1	24 GB (显存)	425 W	6.76 W
CPU	Intel Xeon Gold 5515+ CPU×1	—	165 W	34.11 W
DRAM	Samsung DDR5 R-ECC 32 GB 内存×2 SK Hynix DDR5 R-ECC 32 GB 内存×2	128 GB	30.5 W	5.31 W

注：全文各表中“—”表示该指标不适用。

将在下一小节说明。由于功耗采样依赖操作系统中运行的软件测量工具，因此测试中所界定的“空闲状态”是在严格关闭所有非必要后台进程后、仅保留基础系统服务以及测量工具运行所需的相对纯净环境。

在测量过程中，测试平台搭载的GPU和CPU分别来自NVIDIA与Intel，因此选用两家厂商提供的底层硬件监控工具NVSMI(采集GPU功耗)和Intel PCM(采集CPU与DRAM的能耗数据)。两款工具均支持多个性能监控指标，所支持的最小硬件级采样间隔为1 ms。

4.2 测试流程

表2总结了测试流程的具体步骤。

5 仿真与测量结果

本节比较MU-MISO(Multi-Input Single-Output)和MU-MIMO 场景下不同预编码算法的计算复杂度和能耗开销。所比较的算法包括：

• **ZFBD+Greedy**：采用贪婪算法进行串序空间用户配对，每次迭代以ZFBD算法计算预编码并根据和速率选择最优用户[15]。如第3.4节所述，为了让子带内的所有子载波共享相同预编码，基于所有子载波的信道矩阵计算子带等效信道(详见算法1)。把子带等效信道输入ZFBD算法，即可得到宽带预编码。

• **AGNN**：采用第3.2和3.4 节构建的AGNN学习子带预编码。

• **MDGNN**：在AGNN的基础上移除用户间注意力机制，用于评估注意力机制的有效性。

• **Transformer适配一**：标准Transformer含“编码器-解码器”结构，其中编码器用于提取输入序列特征，解码器用于生成输出序列[16]。预编码学习任务不涉及序列生成，故移除解码器；由于预编码策略不依赖词元顺序，因此移除用于表征序列顺序的位置编码。该变体网络称为Transformer适配一。

• **Transformer适配二**：文献[8]指出，Transformer的注意力计算中采用softmax归一化会破坏对用户间干扰的表征；同时，层归一化并不会带来性能提升。移除这两部分组件可以在提升学习性能的同时降低计算复杂度，对应的网络称为Transformer适配二。

上述所有网络超参数见表3。文献[8]进一步指出，Transformer的词元设计明显影响预编码学习

表 2　GPU、CPU和DRAM能耗测试流程

1) 测试准备阶段

正式测量前终止所有非必需后台进程，独立启动两个命令行终端，分别运行NVSMI和Intel PCM，采样间隔100ms并保持连续采样。工具开启后，服务器空闲运行10 分钟，使各硬件组件进入稳定的空闲水平(GPU和CPU均进入深度休眠)。此阶段数据不参与后续能耗计算。

2) 测量服务器闲置功耗

服务器继续保持空闲状态运行10分钟。该阶段内由NVSMI与Intel PCM采集的数据用于计算GPU、CPU和DRAM 的闲置功耗，计算结果为10分钟内平均值。

3) 测量算法运行的能耗与功耗

启动PyCharm等待加载完毕，运行目标程序。为确保测量期间CPU和GPU稳定，先对所有样本执行一轮预热推理(数据不参与最终统计)。程序记录算法起止时间戳，以划定功耗/能耗采样数据的统计区间。不同类型算法的测量方式如下：

● **传统数值算法**：先把所有测试数据加载至DRAM，再记录预编码计算过程的起止时间戳。

● **AI算法**：先把数据加载至DRAM并构建网络模型；若在GPU上进行计算，则需再把数据和模型参数从DRAM迁移至显存。之后记录时间戳，具体方式如下：

–**训练阶段**：记录模型整个训练周期的起止时间戳；

–**推理阶段**：测试样本需逐个输入网络，记录完成所有测试样本推理的起止时间戳。

表 3　不同场景下各神经网络的超参数

系统配置	$N=32$, 带宽51RB			$N=128$, 带宽273RB	
MIMO技术	MU-MISO			MU-MIMO	MU-MIMO
学习算法	Transformer适配一/二	AGNN	MDGNN	AGNN	AGNN
网络规模	2*[2 560]	5*[32]	5*[32]	4*[32]	6*[16]
批大小	256	256	256	64	64
学习率	0.000 1	0.000 1	0.000 1	0.001	0.001

注1：网络规模m*[n]表示隐藏层数为m，隐藏表示数为n。

注2：表中各模型超参数均基于不同场景下的实际训练收敛性能调优确定。

性能，把用户作为词元远优于把天线作为词元。因此，两种Transformer均采用用户词元，并把每个用户的信道矩阵作为词元特征，其维度为$N_{\mathrm{T}}N_{\mathrm{R}}$。该构造满足用户维度的一维置换等变性[8]。Transformer的注意力机制需计算K个词元间的注意力系数，计算复杂度为$O(K^2N_{\mathrm{R}}N_{\mathrm{T}})$。对比第3.3节给出的AGNN复杂度，考虑到基站天线数N_{T}大于用户天线数N_{R}，因此Transformer的复杂度随系统服务用户数的增长速度更快。

本节结果基于中国移动研究院系统级仿真平台生成的信道数据集，符合3GPP TR 38.901标准[17]。为了比较系统规模对预编码算法计算开销的影响，考虑“小规模”和“大规模”两种系统配置，其载频分别为4 GHz和6 GHz，子载波间隔30 kHz，带宽20 MHz和100 MHz，对应最大资源块(Resource Block, RB)数分别为51和273(每个RB 包含12个子载波)[18]。信道基于3GPP TR 38.901中定义的UMa场景(19个基站，57个扇区)生成，包含大、小尺度信道。基站间距200 m，发射功率46 dBm，天线高度25 m；基站天线数N分别配置为32和128；波束指向方位角、机械下倾角和电子下倾角分别为0°、0°和9°，天线方向图中水平和垂直波束宽度(–3 dB)分别为65°和90°，天线增益为6 dBi。每个扇区内随机散布$K=10$个室外用户，终端高度参照3GPP TR36.873的3D-UMa[19]。每个用户配备$M=4$根全向接收天线(增益0 dBi)。训练数据集包含40,000个样本，测试集包含1,000个样本。评估中默认基站已知理想信道信息，涉及非理想信道信息的场景将另行说明。

算法 1 计算第k个用户的子带等效信道

输入： 第k个用户在子带内C个子载波上的信道矩阵$\boldsymbol{H}_{k,c}$。

输出： 第k个用户的子带等效信道$\widetilde{\boldsymbol{H}}_k$。

1. 计算子带信道相关矩阵
$\boldsymbol{R}_k=1/C\sum_{c=1}^{C}\boldsymbol{H}_{k,c}^{\mathrm{H}}\boldsymbol{H}_{k,c}\in\mathbb{C}^{N_{\mathrm{T}}\times N_{\mathrm{T}}}$；
2. 对$\boldsymbol{R}_k$进行特征分解$\boldsymbol{R}_k=\boldsymbol{U}_k\boldsymbol{\Sigma}_k\boldsymbol{U}_k^{\mathrm{H}}$，得到特征值和特征向量；
3. 从$\boldsymbol{\Sigma}_k$中选取最大的N_{R}个特征值σ_i及对应的特征向量；
4. 计算等效信道矩阵$\widetilde{\boldsymbol{H}}_k=\left[\begin{array}{ccc}\sqrt{\sigma_1}\boldsymbol{u}_1 & \cdots & \sqrt{\sigma_{N_{\mathrm{R}}}}\boldsymbol{u}_{N_{\mathrm{R}}}\end{array}\right]^{\mathrm{H}}$

5.1 MU-MISO场景

针对小规模MU-MISO传输场景，表4给出了不同预编码算法的频谱效率、训练复杂度(包括训练时间和模型参数量)、以及推理复杂度(包括推理FLOPs和推理时间)，其中基站服务$K=8$个单天线用户。从表中可见，仅利用一维置换等变性的两种Transformer的频谱效率低于数值算法ZFBD+Greedy，而所提出的AGNN能够达到更高的频谱效率。

参考第3.3节的推导方式得到各算法的推理FLOPs。在推理FLOPs方面，两种Transformer仍高于ZFBD+Greedy, AGNN则降低了约一个数量级。在推理时间方面，三种学习算法的推理时间均明显短于数值算法。其中，AGNN因利用多维置换等变和不变性，大幅度减少模型参数量，CPU推理时间明显低于两种Transformer；而Transformer得益于大语言模型工程中的较多底层硬件与算子优化，GPU运行时间比AGNN更短。在训练阶段，AGNN有效利用预编码策略的多维置换性质，使其模型参数量远低于基于一维置换等变性的两种Transformer，从而训练时间更短。

表5给出了各算法在训练和推理阶段的功耗与能耗。数值算法仅支持CPU推理，学习算法则在CPU和GPU上均有推理结果。从表中的CPU推理结果可见，数值算法具有较低的推理功耗；但由于其推理时间较长(对应表4的推理时间)，导致整体推理能耗最高。

就推理能耗与功耗而言，与两种Transformer相比，AGNN在CPU和GPU上的结果均更低，且AGNN在CPU上的结果低于其在GPU上的结果。这是因为AGNN网络规模较小，难以充分利用GPU的并行计算能力，因此CPU在该任务上更具优势。在训练阶段，与两种Transformer相比，AGNN通过利用多维置换性质进一步压缩了网络规模，从而降低训练阶段的能耗与功耗。

表 4 小规模MU-MISO场景下的频谱效率、训练和推理复杂度比较

预编码算法	ZFBD+Greedy	Transformer适配一	Transformer适配二	AGNN	MDGNN
频谱效率(bps/Hz)	24.6	17.0	18.2	25.8	17.6
推理FLOPs	6.7×10^8	7.5×10^9	5.6×10^9	1.3×10^7	2.6×10^6
CPU推理时间 (ms)	320	21.3	14.5	4.02	2.61
GPU推理时间 (ms)	—	2.07	2.45	5.22	3.94
训练时间 (s)	—	2,715	1,186	970	839
模型参数量	—	1.1×10^8	7.0×10^7	2.5×10^4	1.5×10^4

由于缺少用户间注意力机制，MDGNN的频谱效率相较AGNN下降31.8%，且低于Transformer适配二，验证了所设计的用户间注意力机制的必要性和有效性。

综合表4和表5可以看出，利用预编码策略的多维置换等变性和不变性设计的神经网络结构，能够在提升频谱效率的同时，降低训练和推理的计算复杂度、功耗和能耗。由于两种Transformer和MDGNN在频谱效率指标上表现较差，后续在MU-MIMO场景中不再评估这些方法。

5.2 MU-MIMO场景

针对小规模和大规模MU-MIMO场景，表6给出了不同预编码算法的频谱效率、训练复杂度和推理复杂度，其中基站服务$K = 10$个多天线用户，每个用户配备$M = 4$根接收天线。表7给出了不同预编码算法在训练和推理阶段的能耗与功耗。通过比较大规模场景下AGNN与ZFBD+Greedy算法的频谱效率、推理FLOPs、推理时间、推理能耗与推理功耗，我们可以得到与第5.1节类似的结论，故此处不再赘述。需要说明的是，表6中测得的15.1 ms是采用单个GPU同时为273个RB计算预编码时的整体耗时。考虑到不同子带的预编码计算相互独立，实际部署时可通过增加硬件资源实现并行化(作为参考，我们测试得到单GPU计算单个子带预

表 5　小规模MU-MISO场景下训练与推理阶段的能耗与功耗比较

预编码算法	硬件组件	训练阶段		推理阶段(CPU)		推理阶段(GPU)	
		能耗 (J)	功耗 (W)	能耗 (mJ)	功耗 (W)	能耗 (mJ)	功耗 (W)
ZFBD+Greedy	CPU	—	—	1.2×10^4	36.4	—	—
	DRAM	—	—	1,750	5.46	—	—
	合计	—	—	**1.3×10^4**	**41.9**	—	—
Transformer适配一	GPU	9.1×10^5	334	—	—	537	259
	CPU	1.2×10^5	44.5	2,715	127	151	73.1
	DRAM	1.6×10^4	5.93	222	10.4	11.8	5.71
	合计	**1.0×10^6**	**384**	**2,937**	**138**	**700**	**338**
Transformer适配二	GPU	4.0×10^5	333	—	—	520	212
	CPU	5.5×10^4	46.8	**1,830**	126	186	75.7
	DRAM	6,975	5.88	141	9.79	14.3	5.84
	合计	**4.6×10^5**	**386**	**1,972**	**136**	**719**	**294**
MDGNN	GPU	4.4×10^4	51.9	—	—	83.2	21.1
	CPU	6.0×10^4	71.9	233	89.4	271.4	68.8
	DRAM	4,920	5.86	14.6	5.60	22.2	5.63
	合计	**1.1×10^5**	**130**	**247.8**	**95.0**	**377**	**95.6**
AGNN	GPU	1.1×10^5	109	—	—	130	24.8
	CPU	7.0×10^4	72.1	373	92.9	345	66.0
	DRAM	5,618	5.79	22.2	5.53	29.0	5.56
	合计	**1.8×10^5**	**187**	**395**	**98.4**	**503**	**96.4**

表 6　MU-MIMO场景下的频谱效率、训练和推理复杂度比较

系统规模	$N = 32$, 带宽51RB	$N = 128$, 带宽273RB	
预编码算法	ZFBD+Greedy	ZFBD+Greedy	AGNN
频谱效率 (bps/Hz)	53.9	118	124
和速率 (Gbps)	1.08	11.8	12.4
推理FLOPs	9.9×10^7	7.9×10^9	1.9×10^8
CPU推理时间 (ms)	444	9,431	124
GPU推理时间 (ms)	—	—	15.1
训练时间 (s)	—	—	3,331
模型参数量	—	—	1.1×10^4

编码的推理时间仅为3.0 ms)。此外，根据我们的测试结果，采用ONNX/TensorRT等加速框架后，GPU上的推理时间可下降至原本运行时间的1/10左右，此时推理时间能够达到亚毫秒级。如果进一步考虑低比特量化等技术，推理时间还能进一步降低。

与数值方法相比，AGNN需要额外的训练开销。根据表7中的能耗数据，图1给出了大规模场景下AGNN的训练与推理总能耗与ZFBD+Greedy算法的推理能耗随推理次数增加的变化趋势，其中AGNN采用GPU推理。可以看出，当推理次数超过1,300次后(由于预编码推理需求为毫秒级，因而1,300次推理对应的实际时间仅为秒级)，AGNN的训练与推理总能耗将低于ZFBD+Greedy，且ZFBD+Greedy的能耗增长速度显著高于AGNN。

为了分析在未来6G大规模预编码场景中引入AI的优势，基于表6和表7的评估数据，表8进一步给出了大规模系统下AGNN与ZFBD+Greedy算法相对于小规模系统下ZFBD+Greedy的指标倍数，其中AGNN采用GPU推理。可以看出，当两个系统均采用ZFBD+Greedy时，大规模系统的和速率提高10.9倍，但高维矩阵运算导致其推理FLOPs达到79.0倍，推理时间和能耗分别增长至21.3倍和38.6倍。可见，其计算开销的增长速度明显高于和速率的提高速度，这将难以满足未来6G系统绿色与可持续发展的目标。相比之下，当大规模系统采用AGNN时，和速率提高11.5倍，而推理FLOPs仅微增1.89倍，推理时间和推理能耗分别仅为0.03倍和0.20倍。这个结果表明，为充分利用未来6G系统的大规模天线和带宽资源来提高系统容量，充分利用预编码策略的先验知识设计的AI方法是降低推理开销的有效途径。

上述评估中假设基站已知理想信道信息，但实际系统中信道信息的获取会受到多种非理想因素的影响[20]。为此，我们进一步考虑基站已知非理想信道信息的场景，基站采用最小二乘算法进行信道估计。同时，考虑用户移动导致的信道过时问题，信道数据集符合3GPP TR 38.901标准协议，用户移动速度为60 km/h，信道过时时长为8 ms。AGNN的输入为过时信道估计结果，网络规模扩展为6*[16]，其余超参数与表3一致。在此条件下重新评估表6、表7和表8中的各项指标，结果分别如表9、表10和表11所示。

由表9～表11可见，在非理想信道条件下，数

表 7 MU-MIMO场景下训练与推理阶段的能耗与功耗比较

预编码算法	硬件组件	训练阶段		推理阶段(CPU)		推理阶段(GPU)	
		能耗 (J)	功耗 (W)	能耗 (mJ)	功耗 (W)	能耗 (mJ)	功耗 (W)
小规模ZFBD+Greedy	CPU	—	—	2.2×10^4	49.9	—	—
	DRAM	—	—	2,408	5.43	—	—
	合计	—	—	$\mathbf{2.5\times10^4}$	**55.4**	—	—
大规模ZFBD+Greedy	CPU	—	—	8.9×10^5	94.9	—	—
	DRAM	—	—	5.3×10^4	5.64	—	—
	合计	—	—	$\mathbf{9.5\times10^5}$	**101**	—	—
大规模AGNN	GPU	9.0×10^5	271	—	—	3,795	251
	CPU	2.3×10^5	68.8	1.5×10^4	117	1,060	70.2
	DRAM	2.1×10^4	6.22	1,871	15.1	95.8	6.35
	合计	$\mathbf{1.2\times10^6}$	**346**	$\mathbf{1.6\times10^4}$	**132**	**4,950**	**328**

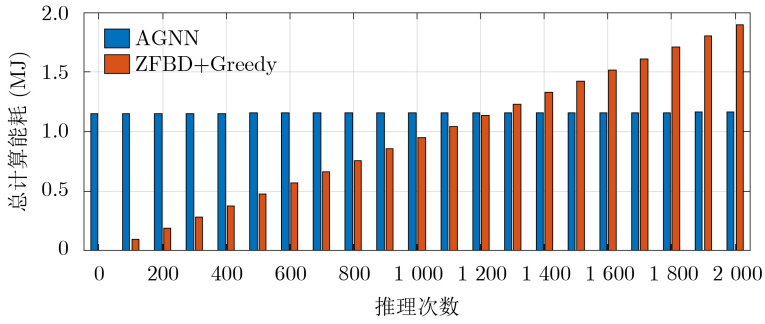

图 1 大规模MU-MIMO场景下的总计算能耗

表 8　大规模场景下各算法相对于小规模场景ZFBD+Greedy的对比

预编码算法	ZFBD+Greedy	AGNN
和速率	10.9倍	11.5倍
推理FLOPs	79.0倍	1.89倍
推理时间	21.3倍	0.03倍
推理能耗	38.6倍	0.20倍

表 9　非理想信道下MU-MIMO场景的频谱效率、训练和推理复杂度比较

系统规模	$N=32$, 带宽51RB	$N=128$, 带宽273RB	
预编码算法	ZFBD+Greedy	ZFBD+Greedy	AGNN
频谱效率 (bps/Hz)	44.6	81.6	90.9
和速率 (Gbps)	0.892	8.16	9.09
推理FLOPs	9.9×10^{7}	7.9×10^{9}	2.8×10^{8}
CPU推理时间 (ms)	444	9,431	260
GPU推理时间 (ms)	—	—	15.1
训练时间 (s)	—	—	31.75
模型参数量	—	—	3.6×10^{5}

表 10　非理想信道下MU-MIMO场景训练与推理阶段的能耗与功耗比较

预编码算法	硬件组件	训练阶段		推理阶段(CPU)		推理阶段(GPU)	
		能耗 (J)	功耗 (W)	能耗 (mJ)	功耗 (W)	能耗 (mJ)	功耗 (W)
小规模ZFBD+Greedy	CPU	—	—	2.2×10^{4}	49.9	—	—
	DRAM	—	—	2,408	5.43	—	—
	合计	—	—	$\mathbf{2.5\times10^{4}}$	**55.4**	—	—
大规模ZFBD+Greedy	CPU	—	—	8.9×10^{5}	94.9	—	—
	DRAM	—	—	5.3×10^{4}	5.64	—	—
	合计	—	—	$\mathbf{9.5\times10^{5}}$	**101**	—	—
大规模AGNN	GPU	6.0×10^{5}	264	—	—	7,625	240
	CPU	2.0×10^{5}	66.8	3.0×10^{4}	113	2,125	68.4
	DRAM	1.5×10^{4}	6.48	4120	16.1	215.5	6.79
	合计	$\mathbf{1.2\times10^{6}}$	**338**	$\mathbf{3.4\times10^{4}}$	**129**	**9,966**	**315**

表 11　非理想信道下大规模场景各算法相对于小规模场景ZFBD+Greedy的对比

预编码算法	ZFBD+Greedy	AGNN
和速率	9.1倍	10.2倍
推理FLOPs	79.0倍	2.83倍
推理时间	21.3倍	0.03倍
推理能耗	38.6倍	0.40倍

值算法和AGNN的频谱效率均有所下降，但两类算法在和速率、推理FLOPs、推理时间和推理能耗等方面的相对关系未发生明显变化。

6　结论

本文针对下行MU-MIMO预编码优化问题，研究了利用无线先验对深度学习的计算复杂度和计算能耗的影响。为此，首先构建了一种满足多维置换等变性与不变性的注意力图神经网络(AGNN)，搭建了硬件能耗测试平台，对各类算法的能耗开销进行了测量与比较。研究结果表明，在AI模型设计方面，若不充分利用无线先验直接采用Transformer等通用结构来学习预编码，则不仅会导致学习性能低于传统数值算法，而且由于模型参数量庞大带来很高的计算开销。与之相比，基于多维置换性质设计的AGNN能够在达到更高频谱效率的同时，大幅度降低训练时间和能耗、模型参数量、推理时间和能耗，以及推理FLOPs等开销。随着系统规模(如天线数和带宽)增大，若仍采用传统数值算法，则所需计算和能耗的增长速度将明显快于系统

和速率的提高倍数；而融入无线先验的AGNN能够在不增加推理时间和推理能耗的前提下解决大规模系统的资源优化问题，有望成为未来6G预编码优化的重要使能技术。

参 考 文 献

[1] WANG Chengxiang, YOU Xiaohu, GAO Xiqi, *et al*. On the road to 6G: Visions, requirements, key technologies, and testbeds[J]. *IEEE Communications Surveys & Tutorials*, 2023, 25(2): 905–974. doi: 10.1109/COMST.2023.3249835.

[2] 沈璐瑶, 周星光, 许子乐, 等. 无蜂窝大规模MIMO系统中下行短包传输的叠加导频功率分配[J]. 电子与信息学报, 2026, 48(1): 57–66. doi: 10.11999/JEIT250655.

SHEN Luyao, ZHOU Xingguang, XU Zile, *et al*. Power allocation for downlink short packet transmission with superimposed pilots in cell-free massive MIMO[J]. *Journal of Electronics & Information Technology*, 2026, 48(1): 57–66. doi: 10.11999/JEIT250655.

[3] 霍如, 吕科呈, 黄韬. 车联网中路径预测驱动的任务切分与计算资源分配方法[J]. 电子与信息学报, 2025, 47(10): 3658–3669. doi: 10.11999/JEIT250135.

HUO Ru, LV Kecheng, and HUANG Tao. Task segmentation and computing resource allocation method driven by path prediction in internet of vehicles[J]. *Journal of Electronics & Information Technology*, 2025, 47(10): 3658–3669. doi: 10.11999/JEIT250135.

[4] 邢智童, 李云, 吴广富, 等. 智能反射面辅助的无线通信系统波束赋形及智能反射面相移技术综述[J]. 电子与信息学报, 2026, 48(2): 697–712. doi: 10.11999/JEIT250790.

XING Zhitong, LI Yun, WU Guangfu, *et al*. A review on phase rotation and beamforming scheme for intelligent reflecting surface assisted wireless communication systems[J]. *Journal of Electronics & Information Technology*, 2026, 48(2): 697–712. doi: 10.11999/JEIT250790.

[5] HUANG Zhaohui, WANG Zhaocheng, and HAN Zhu. Graph neural network empowered wireless communications: Fundamentals, state-of-the-art, challenges, and opportunities[J]. *IEEE Wireless Communications*, 2025, 32(5): 162–168. doi: 10.1109/MWC.002.2400327.

[6] ZHAO Baichuan, GUO Jia, and YANG Chenyang. Understanding the performance of learning precoding policies with graph and convolutional neural networks[J]. *IEEE Transactions on Communications*, 2024, 72(9): 5657–5673. doi: 10.1109/TCOMM.2024.3388506.

[7] LIU Shengjie, GUO Jia, and YANG Chenyang. Multidimensional graph neural networks for wireless communications[J]. *IEEE Transactions on Wireless Communications*, 2024, 23(4): 3057–3073. doi: 10.1109/TWC.2023.3305124.

[8] DUAN Yuxuan, GUO Jia, and YANG Chenyang. Learning precoding in multi-user multi-antenna systems: Transformer or graph transformer?[J]. *IEEE Transactions on Wireless Communications*, 2026, 25: 6284–6300. doi: 10.1109/TWC.2025.3624290.

[9] KASALAEE G, KARKAN A H, FRIGON J F, *et al*. Compression of site-specific deep neural networks for massive MIMO precoding[C]. 2025 IEEE International Conference on Machine Learning for Communication and Networking, Barcelona, Spain, 2025: 1–6. doi: 10.1109/ICMLCN64995.2025.11140523.

[10] HAN Song, POOL J, TRAN J, *et al*. Learning both weights and connections for efficient neural networks[C]. NIPS'15: Proceedings of the 29th International Conference on Neural Information Processing Systems - Volume 1, Montreal, Canada, 2015: 1135–1143.

[11] LI Chen, TSOURDOS A, and GUO Weisi. A transistor operations model for deep learning energy consumption scaling law[J]. *IEEE Transactions on Artificial Intelligence*, 2024, 5(1): 192–204. doi: 10.1109/TAI.2022.3229280.

[12] VELIČKOVIĆ P, CASANOVA A, LIÒ P, *et al*. Graph attention networks[C]. 6th International Conference on Learning Representations ICLR 2018 Conference Track Proceedings, Vancouver, Canada, 2018. doi: 10.17863/CAM.48429. **(查阅网上资料,未找到本条文献出版地信息,请确认)**.

[13] GUO Jia and YANG Chenyang. A model-based GNN for learning precoding[J]. *IEEE Transactions on Wireless Communications*, 2024, 23(7): 6983–6999. doi: 10.1109/TWC.2023.3336911.

[14] ZHANG Lin, HAN Shengqian, and YANG Chenyang. Gradient-driven graph neural networks for learning digital and hybrid precoder[J]. *IEEE Transactions on Communications*, 2026, 74: 706–722. doi: 10.1109/TCOMM.2025.3628767.

[15] SUN Liang and MCKAY M R. Eigen-based transceivers for the MIMO broadcast channel with semi-orthogonal user selection[J]. *IEEE Transactions on Signal Processing*, 2010, 58(10): 5246–5261. doi: 10.1109/TSP.2010.2053709.

[16] VASWANI A, SHAZEER N, PARMAR N, *et al*. Attention is all you need[C]. NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, USA, 2017: 6000–6010.

[17] 3GPP. TR 138. 901 Study on channel model for frequencies from 0.5 to 100 GHz[S]. Cedex: ETSI, 2025.

[18] 3GPP. TS 38. 101-1 User Equipment (UE) radio transmission and reception; Part 1: Range 1 standalone[S]. Cedex: ETSI, 2025.

[19] 3GPP. TR 36. 873 Study on 3D channel model for LTE[S]. Cedex: ETSI, 2017. **(查阅网上资料, 未找到本条文献年份信息, 请确认)**.

[20] ZHU Fenghao, WANG Xinquan, HUANG Chongwen, *et al*. Robust beamforming for RIS-aided communications: Gradient-based manifold meta learning[J]. *IEEE Transactions on Wireless Communications*, 2024, 23(11): 15945–15956. doi: 10.1109/TWC.2024.3435023.

丛鹏宇：男，博士生，研究方向为智能通信.
韩圣千：男，副教授，研究方向为智能通信、移动通信等.
邓鸣宇：男，博士生，研究方向为智能通信.
刘胜杰：男，博士生，研究方向为智能通信.
杨晨阳：女，教授，研究方向为智能通信、移动通信等.
沈嵩辉：男，工程师，研究方向为移动通信物理层技术.

责任编辑：余　蓉

Impact of Wireless Priors on the Computation and Energy Cost of MU-MIMO Precoding Learning

CONG Pengyu[①]　HAN Shengqian[①]　DENG Mingyu[①]　LIU Shengjie[①]
YANG Chenyang[①]　SHEN Songhui[②]

[①](*School of Electronic Information Engineering*, *Beihang University*, *Beijing* 100191, *China*)
[②](*China Mobile Research Institute*, *Beijing* 100053, *China*)

Abstract:

Objective　This paper studies the learning of downlink MU-MIMO (Multiuser Multi-Input Multi-Output) precoding policy from the perspective of computational complexity and energy consumption. Traditional numerical optimization achieves strong performance but incurs rapidly growing computational complexity with the number of base-station's antennas and served users, leading to high inference latency and inference energy consumption. In recent years, deep learning has been widely used to reduce online computational cost, yet existing evaluations typically rely on training/inference time or FLOPs and lack direct energy/power measurements. More importantly, the computational cost of a deep model is tightly coupled with the network architecture, which should be designed to effectively exploit the prior knowledge of the precoding policy. The objective of this paper is thus to develop a network architecture that matches the multi-dimensional permutational properties of the policy, and to analyze how policy priors influence the complexity and energy through comprehensive hardware-based energy/power measurements and simulations.

Methods　We formulate the MU-MIMO precoding policy as a mapping from multi-user channel information to the optimal precoding matrix under a transmit power constraint. The optimal policy exhibits multi-dimensional joint permutation equivariance and invariance with respect to user indices, receive-antenna indices, and base-station antenna indices. To exploit this prior, we propose an Attention-based Graph Neural Network (AGNN) built on a hypergraph structure, whose update and aggregation procedures are designed to satisfy the required equivariance/invariance properties. An attention mechanism modeling inter-user interference is introduced to improve generalization to varying number of users. For broadband precoding, the input layer aggregates multi-subcarrier channel information into an expanded input representation. To quantify energy cost, we build a mixed CPU/GPU measurement framework that collects energy and power for GPU, CPU, and DRAM during training and inference. Simulations use 3GPP TR 38.901 UMa channel data with different antenna array sizes and bandwidth settings. We compare the proposed AGNN against numerical baselines (ZFBD+Greedy pairing) and two Transformer-based architectures, where only one-dimensional permutation properties are satisfied.

Results and Discussions　: The paper reveals two main findings. First, partial exploitation of policy priors leads to poor performance and high cost. In the MU-MISO scenario, the Transformer variants with one-dimensional permutation equivariance yield lower spectral efficiencies than the numerical baseline ZFBD+Greedy and much larger model sizes and inference FLOPs than AGNN. In contrast, AGNN, designed to match the multi-dimensional permutation properties, achieves higher spectral efficiency while reducing inference FLOPs by about one order of magnitude. Hardware measurements further show that AGNN reduces inference energy and

power on both CPU and GPU. Second, in the MU-MIMO scenario with small- and large-scale settings, ZFBD+Greedy increases the system sum rate to 10.9 times, but raise inference FLOPs to 436.4 times, inference time to 20.8 times, and inference energy to 40.8 times. Conversely, AGNN increases the sum rate to 11.5 times while inference FLOPs rise to only 5.3 times; inference time and inference energy become dramatically smaller (0.03 times and 0.22 times, respectively). These results suggest that matching the precoding policy's multi-dimensional permutation priors is an effective way to reduce the computational cost, including FLOPs, latency, and energy consumption, in large-scale 6G MU-MIMO systems.

Conclusions This paper investigated how policy priors affect computational complexity and energy consumption in MU-MIMO precoding learning. By analyzing the multi-dimensional joint permutation equivariance/invariance properties of the optimal precoding policy, we designed an attention-based GNN (AGNN) that matches these properties. A hardware-aware measurement platform was built to obtain direct energy/power measurements for training and inference on CPU, GPU, and DRAM components. Simulations on 3GPP TR 38.901 channel datasets showed that Transformer architectures satisfying only one-dimensional permutation properties can lead to both inferior spectral efficiency and significantly higher computational/energy overhead. In contrast, AGNN achieves higher spectral efficiency while substantially reducing inference FLOPs, inference time, inference energy, and training complexity. As system size increases, traditional numerical methods incur rapidly growing energy costs relative to rate gains, whereas the proposed policy-prior learning method maintains low inference time and energy consumption. Overall, exploiting the prior knowledge of MU-MIMO precoding policy for network architecture design is a key enabler for energy- and complexity-efficient high-dimensional precoding optimization in future 6G networks.

Key words: 6G; Precoding; Policy priors; Graph neural networks; Computational energy consumption