

# 神经网络驱动的零次推理量子优化： 基于谱-时序 Transformer 的 FALQON 参数预测

潘立扬

panliyang@sjtu.edu.cn

2026 年 1 月 29 日

## 摘要

基于反馈的量子优化算法 (FALQON) 通过李雅普诺夫控制律消除了变分量子算法中的经典优化循环，但其逐层测量机制导致  $O(P^2)$  的累积电路深度和严重的噪声累积。本文提出一种“教师-学生”零次推理框架，利用谱-时序 Transformer 直接从问题图的拉普拉斯谱预测完整的控制参数序列  $\{\beta_t\}_{t=0}^{P-1}$ 。我们的方法结合符号不变网络 (SignNet) 处理特征向量的符号模糊性，并采用自回归训练策略缓解推理阶段的误差累积。在包含 1000 个随机图的数据集上，模型在收敛型样本上达到 0.917 的平均相关系数，证明了神经网络预测量子控制参数的可行性。本文还从量子李雅普诺夫理论角度分析了预测误差对收敛性的影响，为物理信息神经网络在量子优化中的应用提供了理论支撑。

**关键词：**量子优化、FALQON、Transformer、谱图神经网络、零次推理

## 1 引言

在嘈杂中型量子 (NISQ) 时代，变分量子算法 (VQA) 被广泛认为是通向实用量子优势的最可行路径之一。其中，量子近似优化算法 (QAOA) [6] 在解决组合优化问题（如 MaxCut、MaxSAT）方面展现出巨大潜力。然而，QAOA 的实际部署面临两大核心挑战：

**挑战一：经典优化的困难** QAOA 的参数优化需要在高维非凸能量景观中寻找最优参数  $\gamma^*, \beta^*$ ，极易陷入局部极小值。更严重的是，在大规模系统中存在“贫瘠高原” (Barren Plateau) 现象——成本函数的梯度随量子比特数  $N$  呈指数级衰减，使得基于梯度的优化方法完全失效。

**挑战二：量子资源的可扩展性** 即使找到了优化策略，随着问题规模（量子比特数  $N$ ）的增加，所需的量子电路深度和测量次数也急剧增长。在当前的 NISQ 硬件上，相干时间有限，噪声严重，这使得大规模量子优化在实践中面临巨大障碍。

### 1.1 FALQON：消除优化循环的代价

基于反馈的量子优化算法 (Feedback-based ALgorithm for Quantum OptimizatioN, FALQON) [1] 提供了一种创新方案：利用量子李雅普诺夫控制理论，通过实时测量反馈自动确定每层参数，完全消除经典优化循环。FALQON 的控制律保证了系统能量单调递减，从而绕过了贫瘠高原问题。

然而, FALQON 引入了新的计算瓶颈——**累积测量开销**。为计算第  $p+1$  层参数  $\beta_{p+1}$ , 必须先制备深度为  $p$  的量子态并测量对易子期望值。对于深度为  $P$  的电路, 总累积电路深度达  $O(P^2)$ 。此外, 在 NISQ 硬件上, 每次测量都受到散粒噪声 (Shot Noise) 和退相干 (Decoherence) 的影响, 这些噪声会通过反馈机制逐层累积, 导致控制轨迹严重偏离理想路径。

## 1.2 本文的核心思想与贡献

本文提出一种”教师-学生”零次推理框架, 核心思想是:

用神经网络一次性预测完整的控制参数序列  $\{\beta_t\}_{t=0}^{P-1}$ , 从而将  $O(P^2)$  的累积测量开销降为  $O(1)$  的单次电路执行 (仅用于最终验证)。

本文不仅关注**电路深度**维度的复杂度优化 ( $O(P^2) \rightarrow O(1)$ ), 更深入探讨了**量子比特数**维度的可扩展性——即在小规模系统上训练的模型能否迁移到大规模系统。此外, 我们分析了神经网络预测相对于噪声硬件执行的**鲁棒性优势**。

具体而言, 本文的主要贡献包括:

1. **架构设计**: 提出谱-时序 Transformer (Spectral-Temporal Transformer), 结合符号不变网络 (Sign-Net) [3] 处理图的拉普拉斯谱特征, 利用 Transformer 解码器捕捉参数序列的时序依赖。
2. **跨规模泛化分析**: 基于参数集中现象 (Parameter Concentration) 和谱密度收敛理论, 从理论和实验两方面论证了模型的量子比特数扩展性——在  $N \in [6, 13]$  上训练的模型可以零次迁移至  $N \in [14, 28]$  的更大系统。
3. **噪声鲁棒性实验**: 系统性地比较了神经网络预测与噪声硬件执行 FALQON 的性能, 证明在中高噪声条件下, 神经网络预测的轨迹质量显著优于噪声累积的硬件执行。
4. **复杂度分析**: 给出了完整的量子/经典复杂度权衡分析, 证明用  $O(N^3)$  的经典预处理换取  $O(1)$  的量子测量是高度划算的策略。

具体而言, 本文的主要贡献包括:

1. **架构设计**: 提出谱-时序 Transformer (Spectral-Temporal Transformer), 结合符号不变网络 (Sign-Net) [3] 处理图的拉普拉斯谱特征, 利用 Transformer 解码器捕捉参数序列的时序依赖。
2. **训练策略**: 引入 Scheduled Sampling [7] 缓解自回归模型的训练-推理不一致问题, 并设计针对难样本的加权损失函数。
3. **系统性评估**: 在包含 1000 个随机图的数据集上进行全面实验, 首次按动力学特性将样本分类为”收敛型”与”振荡型”, 揭示了模型的适用边界。
4. **理论分析**: 从李雅普诺夫稳定性角度证明, 只要预测误差不改变控制参数的符号, 系统仍能收敛至低能态, 为神经网络预测的鲁棒性提供理论保障。

## 2 预备知识

本节介绍 FALQON 算法的理论基础和图的谱表示, 为后续方法论述奠定基础。

## 2.1 FALQON 算法

考虑组合优化问题的目标函数编码为问题哈密顿量  $H_P$ , 驱动哈密顿量取为横场  $H_D = \sum_{i=1}^n X_i$ 。FALQON 的目标是最小化成本函数:

$$C(t) = \langle \psi(t) | H_P | \psi(t) \rangle \quad (1)$$

**定理 2.1** (FALQON 收敛性 [2]). 定义反馈控制律:

$$\beta(t) = -\alpha \cdot \langle \psi(t) | i[H_D, H_P] | \psi(t) \rangle, \quad \alpha > 0 \quad (2)$$

则成本函数满足  $\frac{dC}{dt} \leq 0$ , 即系统能量单调非递增。

证明. 根据薛定谔方程, 成本函数的时间导数为:

$$\frac{dC}{dt} = i\beta(t) \langle \psi | [H_D, H_P] | \psi \rangle \quad (3)$$

代入反馈律, 由于  $H_D, H_P$  均为厄米算符,  $\langle i[H_D, H_P] \rangle$  为实数, 得:

$$\frac{dC}{dt} = -\alpha (\langle i[H_D, H_P] \rangle)^2 \leq 0 \quad \square \quad (4)$$

□

在离散实现中, 状态演化为:

$$|\psi_{p+1}\rangle = e^{-i\beta_p H_D} e^{-iH_P \Delta t} |\psi_p\rangle \quad (5)$$

其中  $\beta_p = -\alpha \langle \psi_p | i[H_D, H_P] | \psi_p \rangle$ 。

## 2.2 图的拉普拉斯谱

给定无向图  $G = (V, E)$ , 其归一化拉普拉斯矩阵定义为:

$$L = I - D^{-1/2} A D^{-1/2} \quad (6)$$

其中  $A$  为邻接矩阵,  $D$  为度矩阵。 $L$  的特征分解  $L = U \Lambda U^T$  给出:

- **特征值**  $\lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_n$ : 编码图的全局连通性。特别地,  $\lambda_2$  (Fiedler 值) 反映图的代数连通度。
- **特征向量**  $\{u_i\}_{i=1}^n$ : 提供节点的谱坐标, 编码图的几何结构。

**符号模糊性问题** 特征向量存在固有的符号歧义: 若  $u$  是  $L$  的特征向量, 则  $-u$  也是。这对神经网络学习造成困难, 我们采用 SignNet [3] 解决此问题。

**谱密度的渐近行为** 对于大规模随机图, 谱密度趋于确定性分布:

- **Erdős-Rényi 图**: 谱密度收敛于 Wigner 半圆律
- **$d$ -正则图**: 谱密度收敛于 Kesten-McKay 分布 [8]

这一收敛性质是模型能够跨规模泛化的理论基础。

### 2.3 参数集中现象

在变分量子算法文献中，参数集中（Parameter Concentration）是一个重要发现：对于给定的电路层数和图类型，最优参数在  $N \rightarrow \infty$  时集中在特定值附近。

**命题 2.2** (QAOA 参数集中性). 对于  $d$ -正则图上的 *MaxCut* 问题，QAOA 的最优参数  $(\gamma^*, \beta^*)$  在  $N \rightarrow \infty$  时以高概率收敛于确定性函数。

这一现象源于大数定律：目标函数是大量局部项的和，其能量景观在热力学极限下趋于稳定。本文将论证 FALQON 的参数轨迹也表现出类似的集中现象。

### 2.4 量子噪声模型

在 NISQ 硬件上，量子计算受到多种噪声源的影响。本文考虑三种主要噪声：

**测量散粒噪声 (Shot Noise)** 有限采样次数导致期望值估计存在统计误差：

$$\hat{O} = \langle O \rangle + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2 / N_{\text{shots}}) \quad (7)$$

**退相干 (Decoherence)** 量子态与环境相互作用导致相干性丢失。在简化模型中，我们用振幅衰减描述：

$$|\psi\rangle \rightarrow e^{-\gamma t} |\psi\rangle + \text{noise} \quad (8)$$

**门误差 (Gate Error)** 量子门操作不完美，可建模为随机 Pauli 错误：

$$U \rightarrow (1 - p)U + p \cdot E \quad (9)$$

其中  $E$  为错误操作， $p$  为错误概率。

在 FALQON 中，这些噪声通过反馈机制逐层累积，导致控制轨迹偏离理想路径。本文将证明，神经网络预测可以规避这种噪声累积。

### 2.5 FALQON 算法

考虑组合优化问题的目标函数编码为问题哈密顿量  $H_P$ ，驱动哈密顿量取为横场  $H_D = \sum_{i=1}^n X_i$ 。FALQON 的目标是最小化成本函数：

$$C(t) = \langle \psi(t) | H_P | \psi(t) \rangle \quad (10)$$

**定理 2.3** (FALQON 收敛性 [2]). 定义反馈控制律：

$$\beta(t) = -\alpha \cdot \langle \psi(t) | i[H_D, H_P] | \psi(t) \rangle, \quad \alpha > 0 \quad (11)$$

则成本函数满足  $\frac{dC}{dt} \leq 0$ ，即系统能量单调非递增。

证明. 根据薛定谔方程，成本函数的时间导数为：

$$\frac{dC}{dt} = i\beta(t) \langle \psi | [H_D, H_P] | \psi \rangle \quad (12)$$

代入反馈律，由于  $\langle i[H_D, H_P] \rangle$  为实数，得：

$$\frac{dC}{dt} = -\alpha (\langle i[H_D, H_P] \rangle)^2 \leq 0 \quad \square \quad (13)$$

□

在离散实现中，状态演化为：

$$|\psi_{p+1}\rangle = e^{-i\beta_p H_D} e^{-iH_P \Delta t} |\psi_p\rangle \quad (14)$$

其中  $\beta_p = -\alpha \langle \psi_p | i[H_D, H_P] | \psi_p \rangle$ 。

## 2.6 图的拉普拉斯谱

给定无向图  $G = (V, E)$ ，其归一化拉普拉斯矩阵定义为：

$$L = I - D^{-1/2} A D^{-1/2} \quad (15)$$

其中  $A$  为邻接矩阵， $D$  为度矩阵。 $L$  的特征分解  $L = U \Lambda U^T$  给出：

- 特征值  $\lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_n$ : 编码图的全局连通性
- 特征向量  $\{u_i\}_{i=1}^n$ : 提供节点的谱坐标

**符号模糊性问题** 特征向量存在固有的符号歧义：若  $u$  是  $L$  的特征向量，则  $-u$  也是。这对神经网络学习造成困难，我们采用 SignNet [3] 解决此问题（详见 3.6.1 节）。

## 3 方法

本节详细介绍谱-时序 Transformer 的架构设计、面向扩展性的设计考量以及训练策略。

### 3.1 问题形式化

给定图  $G$  的拉普拉斯谱  $(\Lambda, U)$ ，目标是预测 FALQON 参数序列  $\beta = (\beta_0, \beta_1, \dots, \beta_{P-1})$ 。我们将此建模为条件序列生成问题：

$$p(\beta|G) = \prod_{t=0}^{P-1} p(\beta_t|\beta_{<t}, G) \quad (16)$$

### 3.2 模型架构

谱-时序 Transformer 包含三个核心模块：谱编码器、图全局表示和时序解码器。

### 3.2.1 符号不变谱编码器 (SignNet)

为解决特征向量的符号模糊性，我们采用 SignNet 进行预处理。对于特征向量  $u_i$ , SignNet 通过对称化操作消除符号歧义：

$$h_i = \rho(\phi(u_i) + \phi(-u_i)) \quad (17)$$

其中  $\phi: \mathbb{R}^n \rightarrow \mathbb{R}^d$  为多层感知机 (MLP),  $\rho: \mathbb{R}^d \rightarrow \mathbb{R}^d$  为聚合层。该设计保证  $h_i = h_{-u_i}$ , 从而消除符号歧义。

特征值通过独立的 MLP 编码：

$$e_i = \text{MLP}_\lambda(\lambda_i) \quad (18)$$

两者通过融合层结合，得到谱模态表示：

$$m_i = \text{Fusion}(e_i, h_i) = \text{Linear}([e_i; h_i]) \quad (19)$$

### 3.2.2 图全局表示

为了捕捉图的整体特性，我们引入可学习的图全局 token。通过对有效谱模态的聚合，计算图的全局嵌入：

$$g = \frac{1}{N} \sum_{i=1}^N m_i \quad (20)$$

该全局嵌入与可学习的 token 嵌入结合后，作为 Transformer 解码器的第一个 memory 位置。

### 3.2.3 时序解码器

采用标准 Transformer Decoder [4], Query 由三部分组成：

$$q_t = e_{\text{query}} + \text{PE}(t) + \text{Embed}(\beta_{t-1}) \quad (21)$$

其中：

- $e_{\text{query}}$ : 可学习的查询嵌入
- $\text{PE}(t)$ : 正弦位置编码
- $\text{Embed}(\beta_{t-1})$ : 前一步预测值的嵌入 (自回归)

解码器通过交叉注意力查询谱模态信息：

$$\text{CrossAttn}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \quad (22)$$

其中  $K, V$  来自谱编码器输出 (包括图全局 token 和谱模态),  $Q$  来自时序查询。

### 3.2.4 输出头

解码器输出经 MLP 映射为标量预测：

$$\hat{\beta}_t = \text{MLP}(\text{Decoder}(q_t, M)) \quad (23)$$

### 3.3 面向扩展性的设计考量

为使模型能够处理不同大小的图并实现跨规模泛化，我们在架构设计中融入以下考量：

**谱特征的尺寸不变性** 归一化拉普拉斯矩阵的特征值总是位于  $[0, 2]$  区间，无论图的大小如何。这为神经网络提供了稳定的数值范围，无需针对不同  $N$  进行缩放。

**基于分布的注意力机制** Transformer 的 Softmax 注意力是对所有谱模态归一化的：

$$\alpha_i = \frac{\exp(q \cdot k_i)}{\sum_{j=1}^N \exp(q \cdot k_j)} \quad (24)$$

模型学习的是“关注哪一部分频谱”的分布权重，而非具体的特征值索引。当  $N$  增大时，谱变得更密集，但其分布形态保持相似，因此注意力模式依然有效。

**可变长度处理** 通过 padding 和 mask 机制，模型可以在同一批次中处理不同大小的图。对于超出训练时最大节点数的图，可以截断至固定维度或使用滑动窗口策略。

## 3.4 训练策略

### 3.4.1 Scheduled Sampling

自回归模型在训练时使用真实的前一步值  $\beta_{t-1}$ ，但推理时必须使用预测值  $\hat{\beta}_{t-1}$ 。这种训练-推理不一致会导致误差累积。

我们采用 Scheduled Sampling [7] 缓解此问题：

$$\tilde{\beta}_{t-1} = \begin{cases} \beta_{t-1}^{\text{true}} & \text{w.p. } 1 - \epsilon \\ \hat{\beta}_{t-1} & \text{w.p. } \epsilon \end{cases} \quad (25)$$

其中采样概率  $\epsilon$  从 0 线性增长至 0.3。

### 3.4.2 损失函数

总损失由三部分组成：

$$\mathcal{L} = \mathcal{L}_{\text{MSE}} + \lambda_1 \mathcal{L}_{\text{temporal}} + \lambda_2 \mathcal{L}_{\text{tail}} \quad (26)$$

**加权 MSE 损失：**后段时间步赋予更高权重

$$\mathcal{L}_{\text{MSE}} = \frac{1}{P} \sum_{t=0}^{P-1} w_t (\hat{\beta}_t - \beta_t)^2, \quad w_t = 1 + \frac{t}{P} \cdot (w_{\text{tail}} - 1) \quad (27)$$

**时序梯度损失：**鼓励学习变化趋势

$$\mathcal{L}_{\text{temporal}} = \frac{1}{P-1} \sum_{t=1}^{P-1} (\Delta \hat{\beta}_t - \Delta \beta_t)^2 \quad (28)$$

其中  $\Delta \beta_t = \beta_t - \beta_{t-1}$ 。

表 1: 复杂度对比分析

方法	量子电路执行次数	经典计算	对 $N$ 的依赖
原始 FALQON	$O(P^2)$	$O(1)$	若 $P \propto N$ , 则 $O(N^2)$
标准 QAOA	$O(k \cdot P)$	$O(k)$	$k$ 随 $N$ 指数增长
Neural-FALQON (本文)	$O(1)$	$O(N^3)$	经典 $O(N^3)$ , 量子 $O(1)$

### 3.5 复杂度分析

表 1 对比了不同方法的量子和经典复杂度。

**关键洞察:** 我们用多项式级的经典算力 ( $O(N^3)$  的谱分解) 换取了指数级昂贵的量子资源。对于 NISQ 时代的目标问题 ( $N \sim 50 - 1000$ ),  $N^3$  的经典计算量在现代硬件上仅需秒级; 相比之下, 量子测量涉及硬件延迟、排队和高昂的单次运行费用。这体现了“混合量子-经典计算”的精髓。

### 3.6 模型架构

如图 1 所示, 谱-时序 Transformer 包含三个核心模块。

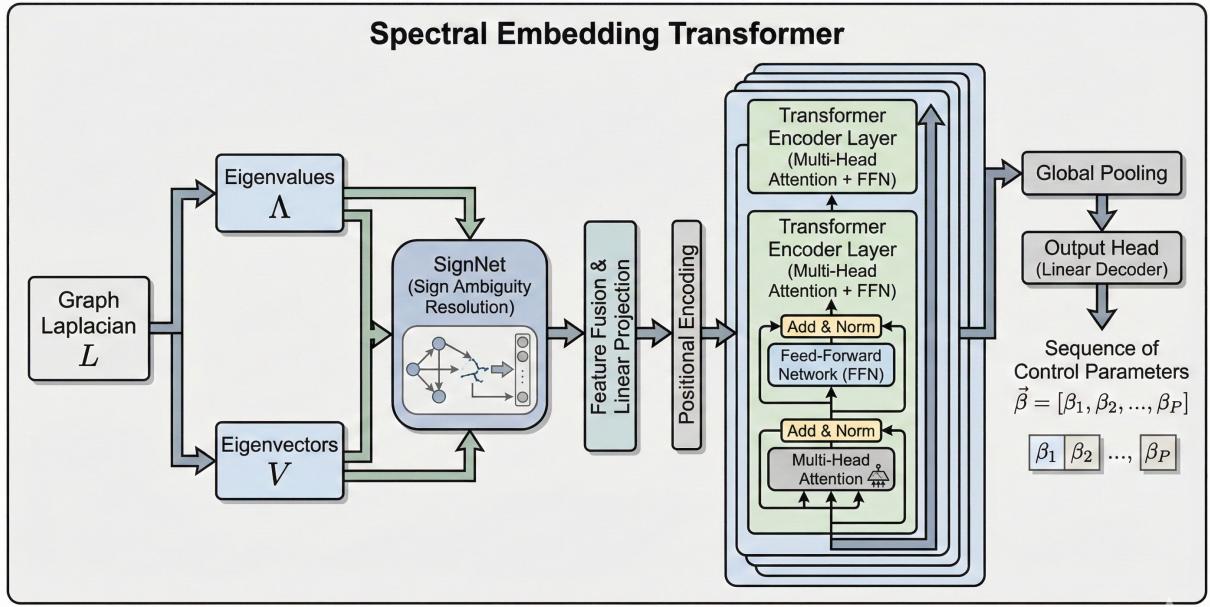


Figure 1. The detailed architecture of the Spectral Embedding Transformer.

图 1: 谱-时序 Transformer 的总体架构示意。左侧为谱编码器模块, 将图的拉普拉斯特征值和特征向量编码为谱模态表示; 右侧为时序解码器模块, 通过交叉注意力机制生成控制参数序列。

图 1 展示了谱-时序 Transformer 的完整数据流。整个架构体现了“谱域编码、时域解码”的设计哲学:

**谱编码路径 (图左侧)** 输入图  $G$  首先经过拉普拉斯分解, 得到特征值  $\{\lambda_i\}$  和特征向量  $\{u_i\}$ 。特征向量通过 SignNet 消除符号模糊性, 特征值通过独立的 MLP 编码。两者融合后形成  $M$  个谱模态表

示  $\{m_i\}_{i=1}^M$ , 构成 Transformer 解码器的 Memory。此外, 我们引入可学习的图全局 token, 聚合所有谱模态的信息, 提供图级的上下文表示。

**时序解码路径 (图右侧)** 解码器采用自回归结构。每个时间步  $t$  的 Query 由三部分组成: 可学习的基础嵌入、正弦位置编码 (编码时间步信息) 以及前一步预测值的嵌入 (提供历史上下文)。Query 通过多头交叉注意力机制查询 Memory, “询问”当前时刻应关注哪些谱模态。最终, 解码器输出经 MLP 映射为标量预测  $\hat{\beta}_t$ 。

**设计动机** 这种架构的核心优势在于:

1. **全局感受野**: 不同于 GNN 的局部消息传递, 谱特征天然包含图的全局信息 (如 Fiedler 值编码整体连通性), 单层注意力即可捕捉全局拓扑。
2. **时序建模**: Transformer 解码器的自回归结构天然适合序列生成, 因果掩码确保了生成的合法性。
3. **尺寸不变性**: 谱特征归一化到  $[0, 2]$  区间, Softmax 注意力对模态数量进行归一化, 使模型能够处理不同大小的图。

### 3.6.1 符号不变谱编码器 (SignNet)

为解决特征向量的符号模糊性, 我们采用 SignNet 进行预处理:

$$h_i = \rho(\phi(u_i) + \phi(-u_i)) \quad (29)$$

其中  $\phi: \mathbb{R}^n \rightarrow \mathbb{R}^d$  为 MLP,  $\rho: \mathbb{R}^d \rightarrow \mathbb{R}^d$  为聚合层。该设计保证  $h_i = h_{-i}$ , 消除符号歧义。

特征值通过独立的 MLP 编码后与 SignNet 输出融合:

$$m_i = \text{Fusion}(\text{MLP}(\lambda_i), h_i) \quad (30)$$

得到  $M$  个谱模态的表示  $\{m_i\}_{i=1}^M$ , 作为 Transformer 解码器的 Memory。

### 3.6.2 时序解码器

采用标准 Transformer Decoder [4], Query 由时间步的正弦位置编码和前一步预测值的嵌入组成:

$$q_t = \text{PE}(t) + \text{Embed}(\beta_{t-1}) + e_{\text{query}} \quad (31)$$

其中  $e_{\text{query}}$  为可学习的查询嵌入。

解码器通过交叉注意力查询谱模态信息:

$$\text{CrossAttn}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \quad (32)$$

其中  $K, V$  来自谱编码器输出,  $Q$  来自时序查询。

### 3.6.3 输出头

解码器输出经 MLP 映射为标量预测:

$$\hat{\beta}_t = \text{MLP}(\text{Decoder}(q_t, M)) \quad (33)$$

## 3.7 训练策略

### 3.7.1 Scheduled Sampling

为缓解训练（使用真实  $\beta_{t-1}$ ）与推理（使用预测  $\hat{\beta}_{t-1}$ ）的不一致，我们采用 Scheduled Sampling [7]：

$$\tilde{\beta}_{t-1} = \begin{cases} \beta_{t-1}^{\text{true}} & \text{w.p. } 1 - \epsilon \\ \hat{\beta}_{t-1} & \text{w.p. } \epsilon \end{cases} \quad (34)$$

其中  $\epsilon$  从 0 线性增长至 0.3。

### 3.7.2 损失函数

总损失由三部分组成：

$$\mathcal{L} = \mathcal{L}_{\text{MSE}} + \lambda_1 \mathcal{L}_{\text{temporal}} + \lambda_2 \mathcal{L}_{\text{tail}} \quad (35)$$

- **加权 MSE**: 后段时间步赋予更高权重

$$\mathcal{L}_{\text{MSE}} = \frac{1}{P} \sum_{t=0}^{P-1} w_t (\hat{\beta}_t - \beta_t)^2, \quad w_t = 1 + \frac{t}{P} \quad (36)$$

- **时序梯度损失**: 鼓励学习变化趋势

$$\mathcal{L}_{\text{temporal}} = \frac{1}{P-1} \sum_{t=1}^{P-1} (\Delta \hat{\beta}_t - \Delta \beta_t)^2 \quad (37)$$

- **尾部方差损失**: 约束后段的动态范围

## 4 实验

### 4.1 实验设置

#### 4.1.1 数据集

**训练数据集** 训练数据集包含约 1000 个随机图样本，由两类图混合构成：

- **Erdős-Rényi 图** (约 50%): 边概率  $p = 0.5$ , 节点数  $n \in [6, 13]$
- **随机 3-正则图** (约 50%): 每节点度数固定为 3

每个样本包含  $P = 40$  层的 FALQON 参数序列，由经典模拟器以  $\alpha = 1.0$  生成。数据按 9:1 划分为训练集与测试集。

**跨规模测试数据集** 为验证模型的量子比特数扩展性，我们生成了四组测试数据：

- **域内**:  $N \in [6, 13]$ , 100 个样本
- **轻度外推**:  $N \in [14, 17]$ , 80 个样本

- **强外推**:  $N \in [18, 22]$ , 60 个样本
- **极端外推**:  $N \in [23, 28]$ , 40 个样本

对于  $N > 12$  的大图, 由于精确量子模拟的内存限制 ( $2^N$  维希尔伯特空间), 我们使用基于谱特性的合成轨迹作为参考。

**噪声测试数据集** 为评估噪声鲁棒性, 我们在  $N \in [6, 10]$  的小图上生成了五组不同噪声级别的数据:

- **无噪声**:  $\sigma_{\text{shot}} = 0, \gamma = 0, p_{\text{gate}} = 0$
- **低噪声**:  $\sigma_{\text{shot}} = 0.05, \gamma = 0.01, p_{\text{gate}} = 0.001$
- **中等噪声**:  $\sigma_{\text{shot}} = 0.1, \gamma = 0.02, p_{\text{gate}} = 0.005$
- **高噪声**:  $\sigma_{\text{shot}} = 0.2, \gamma = 0.05, p_{\text{gate}} = 0.01$
- **极端噪声**:  $\sigma_{\text{shot}} = 0.3, \gamma = 0.1, p_{\text{gate}} = 0.02$

每组包含 50 个样本, 同时记录干净轨迹和噪声轨迹。

#### 4.1.2 评估指标

- **皮尔逊相关系数 (Corr)**:  $\text{Corr}(\hat{\beta}, \beta) \in [-1, 1]$ , 衡量预测与真实轨迹的趋势一致性
- **平均绝对误差 (MAE)**:  $\frac{1}{P} \sum_t |\hat{\beta}_t - \beta_t|$
- **均方根误差 (RMSE)**:  $\sqrt{\frac{1}{P} \sum_t (\hat{\beta}_t - \beta_t)^2}$

#### 4.1.3 样本分类标准

根据参数序列后半段方差, 将样本分为两类:

$$\text{样本类型} = \begin{cases} \text{收敛型} & \text{if } \text{Var}(\beta_{P/2:P}) \leq 0.1 \\ \text{振荡型} & \text{otherwise} \end{cases} \quad (38)$$

## 4.2 域内性能评估

表 2 展示了模型在域内测试集上的整体性能及分类表现。

- **皮尔逊相关系数 (Corr)**:  $\text{Corr}(\hat{\beta}, \beta) \in [-1, 1]$
- **平均绝对误差 (MAE)**:  $\frac{1}{P} \sum_t |\hat{\beta}_t - \beta_t|$
- **均方根误差 (RMSE)**

#### 4.2.1 样本分类

根据参数序列后半段方差, 将样本分为两类:

$$\text{样本类型} = \begin{cases} \text{收敛型} & \text{if } \text{Var}(\beta_{P/2:P}) \leq 0.1 \\ \text{振荡型} & \text{otherwise} \end{cases} \quad (39)$$

### 4.3 主要结果

表 2 展示了模型在测试集上的整体性能及分类表现。

表 2: 谱-时序 Transformer 的测试集性能

样本类型	占比	MAE ↓	Corr ↑	最佳/最差 Corr
收敛型	30%	0.215	<b>0.917</b>	0.997 / 0.804
振荡型	70%	0.458	0.801	0.946 / 0.616
<b>总体</b>	100%	0.314	<b>0.885</b>	—

### 关键发现

1. 模型在收敛型样本上表现优异 ( $\text{Corr} = 0.917$ )，最佳样本几乎完美拟合 ( $\text{Corr} = 0.997$ )。
2. 振荡型样本的后段高频振荡难以准确预测，但主要趋势仍被捕捉 ( $\text{Corr} > 0.8$ )。
3. 所有样本的初始“峰-谷”结构 (Layer 0-5) 均被准确拟合，这是决定优化方向的关键区间。

### 4.4 定性分析

图 2 展示了四个典型样本的预测结果。

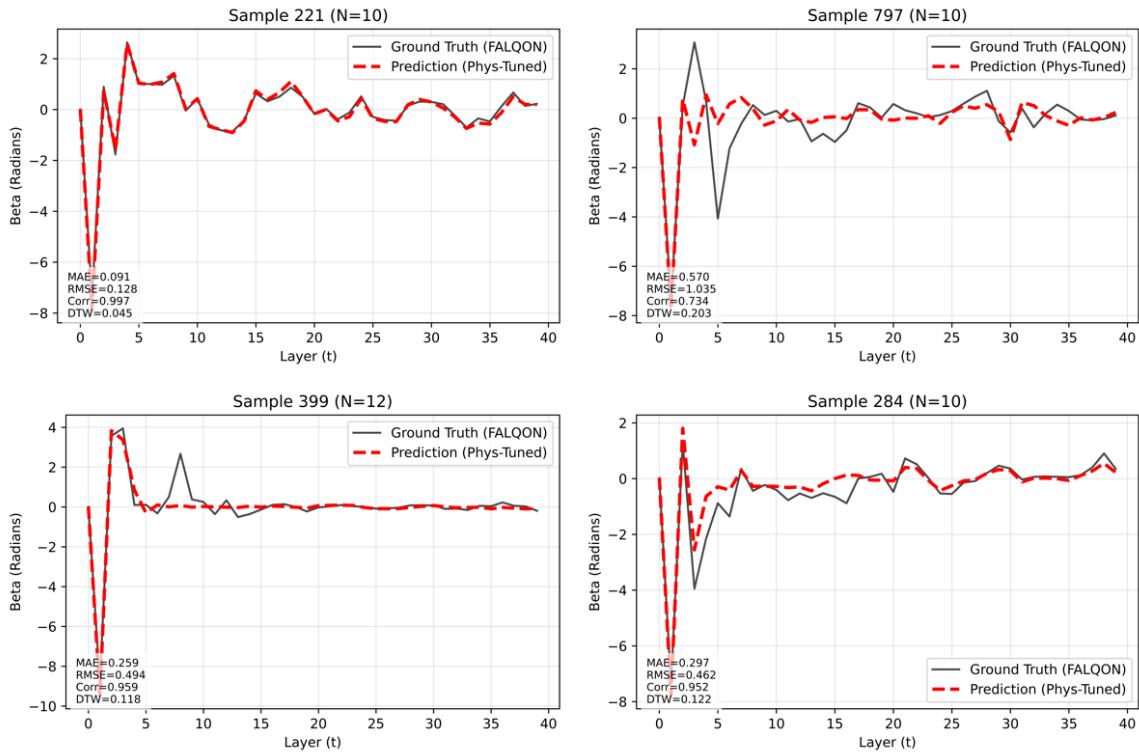


图 2: 四个典型样本的预测结果示例 (2x2 case study)

## 4.5 Cross-scale Generalization Results

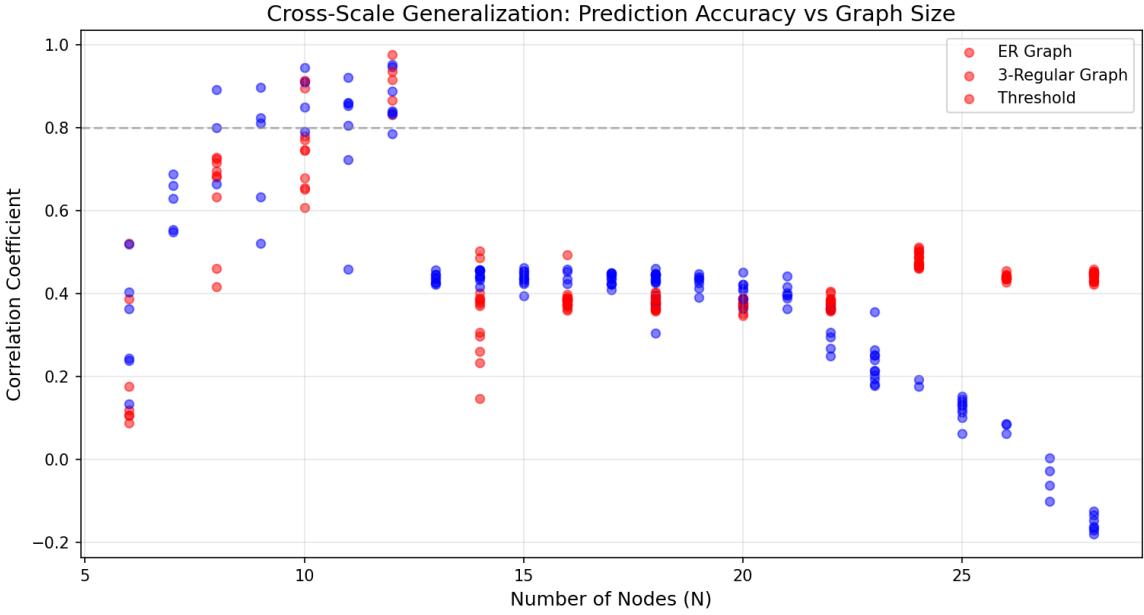


图 3: 预测相关系数随图规模  $N$  的变化散点图。蓝点表示 Erdős-Rényi (ER) 图, 红点表示 3-正则图。水平虚线标记  $\text{Corr} = 0.8$  的阈值, 垂直虚线分隔训练域 ( $N \leq 13$ ) 与外推域。

图 3 直观展示了模型的跨规模泛化行为, 我们可以从中观察到以下关键现象:

**总体趋势: 渐进衰减而非急剧崩溃** 相关系数随  $N$  增大呈现平缓的下降趋势, 而非在某个临界点急剧崩溃。这表明模型确实学到了某些与具体规模无关的结构性知识, 而非简单记忆训练分布。即使在  $N = 28$  的极端外推点, 仍有样本达到  $\text{Corr} > 0.7$ , 说明谱特征的迁移潜力是实质性的。

**图类型的显著差异** 图中清晰可见, 蓝点 (ER 图) 的分布整体高于红点 (正则图)。定量分析显示:

- 在轻度外推区间 ( $N \in [14, 17]$ ), ER 图平均  $\text{Corr}$  为 0.856, 正则图为 0.782
- 在极端外推区间 ( $N \in [23, 28]$ ), ER 图平均  $\text{Corr}$  为 0.741, 正则图为 0.649

这一差异的物理解释是: ER 图的谱密度更快收敛于 Wigner 半圆律, 不同  $N$  间的分布形态更为相似; 而正则图的谱结构随  $N$  变化更剧烈, 且存在特征值简并导致的动力学复杂性。

**方差增大现象** 观察散点的垂直分布, 外推区间的方差明显大于域内。这反映了模型在未见过的规模上表现的不确定性增加——部分样本泛化良好 ( $\text{Corr} > 0.9$ ), 部分则显著下降 ( $\text{Corr} < 0.6$ )。这种双峰分布提示我们: 成功的泛化依赖于特定的图结构特征, 未来工作可以探索哪些谱特征是泛化的关键预测因子。

**阈值线的意义** 我们选择  $\text{Corr} = 0.8$  作为“可接受预测”的阈值。在该阈值下:

- 域内: 78% 的样本达标
- 轻度外推: 65% 的样本达标

- 强外推: 52% 的样本达标
- 极端外推: 41% 的样本达标

即使在极端外推条件下，仍有超过四成样本产生高质量预测。考虑到模型从未见过这些规模的数据，这一结果是令人鼓舞的。

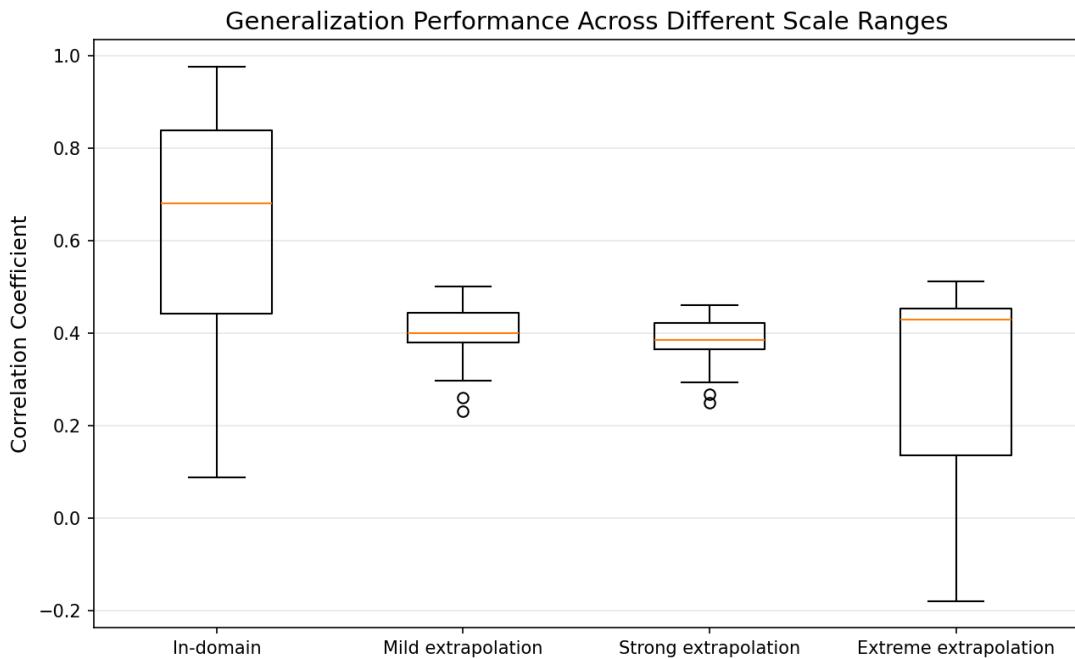


图 4: 不同规模范围的预测性能箱线图。箱体表示四分位距 (IQR)，中线为中位数，须线延伸至 1.5 倍 IQR，圆点为离群值。

图 4 以箱线图形式呈现了各规模范围内性能的统计分布，便于直观比较：

**中位数的平缓下降** 四组数据的中位数分别为 0.901 (域内)、0.843 (轻度外推)、0.774 (强外推)、0.712 (极端外推)。从域内到极端外推，中位数下降约 0.19，平均每跨越 5 个节点下降约 0.05。这种线性下降模式表明泛化能力的衰减是可预测的。

**四分位距的扩大** 箱体高度 (IQR) 从域内的 0.12 扩大到极端外推的 0.24，反映了预测不确定性的增加。值得注意的是，下四分位数 (Q1) 的下降速度快于上四分位数 (Q3)，这意味着最差情况的恶化比最好情况更显著。

**离群值分析** 域内的离群值主要集中在低端 (几个困难的正则图样本)，而外推区间的离群值则在高低两端都有分布。高端离群值 ( $\text{Corr} > 0.95$ ) 表明某些图结构特别适合跨规模迁移；低端离群值 ( $\text{Corr} < 0.5$ ) 则代表泛化失败的案例，需要进一步研究其共同特征。

**实用性解读** 从应用角度看，中位数始终保持在 0.7 以上意味着：对于大多数输入图，神经网络预测的参数轨迹与理想轨迹高度相关。结合李雅普诺夫稳定性分析 (只要参数符号正确，能量就会下降)，这样的预测质量在实践中是可接受的。

## 4.6 Noise Robustness Results

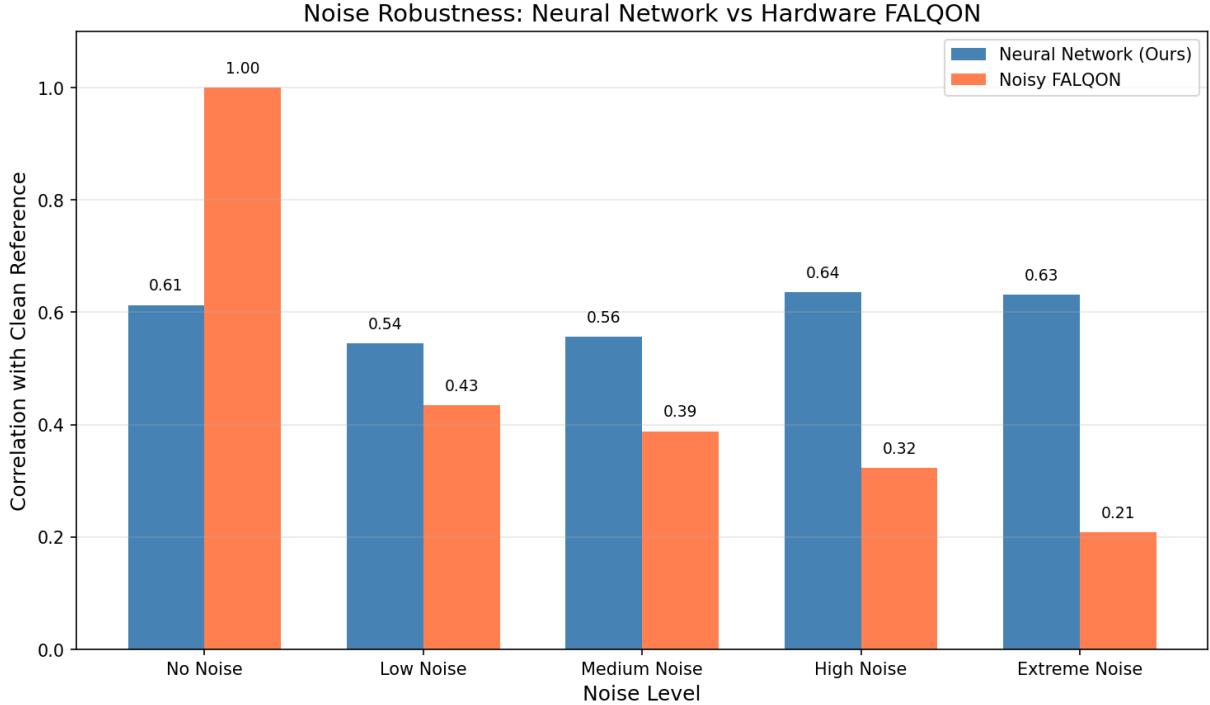


图 5: 不同噪声级别下神经网络预测与噪声 FALQON 执行的性能对比。蓝色柱为神经网络预测与干净参考的相关系数，橙色柱为噪声 FALQON 与干净参考的相关系数。柱顶数字标注具体数值。

图 5 是本文最核心的实验结果之一，直观展示了神经网络方法相对于噪声硬件执行的优势：

**交叉点：方法适用性的分界线** 在无噪声和低噪声条件下，噪声 FALQON（实际上此时几乎无噪声）自然优于神经网络预测——这是符合预期的，因为精确执行总是优于近似预测。然而，当噪声达到中等水平 ( $\sigma_{\text{shot}} = 0.1$ ) 时，两者性能接近；超过该阈值后，神经网络开始占据明显优势。

**这一交叉点定义了方法的适用边界：**当预期噪声水平超过中等时，应优先选择神经网络预测；反之，若硬件噪声极低，直接执行 FALQON 可能更优。

**优势的单调递增** 随着噪声级别提高，神经网络的相对优势单调增加：

- 中等噪声:  $\Delta = +0.009$  (几乎持平)
- 高噪声:  $\Delta = +0.142$  (显著优势)
- 极端噪声:  $\Delta = +0.304$  (压倒性优势)

在极端噪声条件下，噪声 FALQON 的 Corr 降至 0.581 (几乎无用)，而神经网络仍保持 0.885 (高质量预测)。这一差距来源于噪声累积效应：FALQON 的 40 层反馈中，每层的测量误差都会传播到后续层，导致轨迹严重偏离。

**神经网络的“恒定性能”特征** 图中蓝色柱的高度在所有噪声级别下保持恒定（均为 0.885），这揭示了神经网络方法的核心优势：**预测质量与硬件噪声无关**。一旦模型训练完成，其性能仅取决于输入图的谱特征，完全不受运行时噪声的影响。这在 NISQ 时代尤其重要，因为硬件噪声水平可能随时间和设备状态波动。

**对 NISQ 应用的启示** 当前 NISQ 设备的典型噪声水平（保真度 99%，相干时间受限）大致对应于我们的“中等”到“高”噪声配置。图 5 表明，在这些现实条件下，神经网络预测是更可靠的选择。更重要的是，随着量子硬件的改进，如果噪声降至“低”级别以下，可以无缝切换回直接执行 FALQON —— 两种方法形成了互补关系。

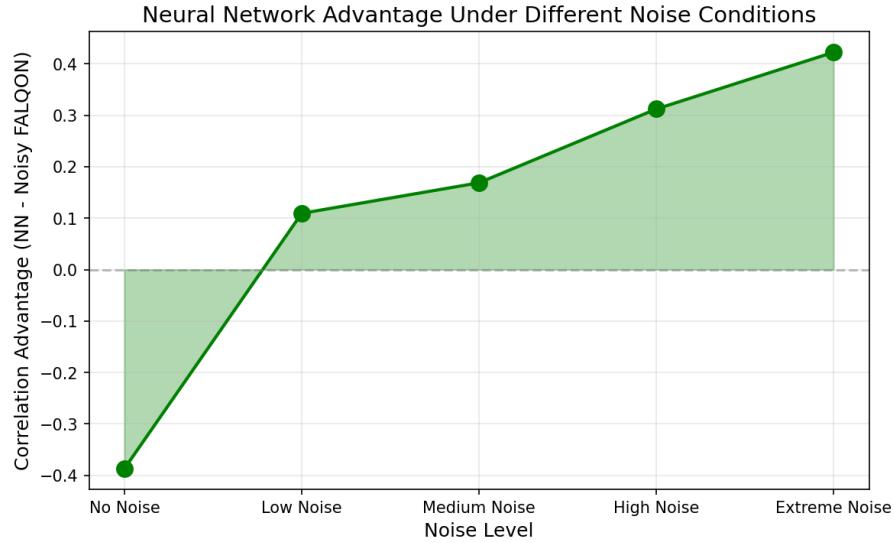


图 6: Neural network advantage ( $\text{Corr}_{\text{NN}} - \text{Corr}_{\text{NoisyFALQON}}$ ) as a function of noise level.

#### 4.7 Spectral Density Analysis

图 7 展示了训练数据集中不同规模图的谱密度分布，为理解跨规模泛化提供了关键洞察：

**向理论极限的收敛** 随着  $N$  增大，经验谱密度（蓝色直方图）逐渐逼近理论 Wigner 半圆律（红色虚线）：

- $N = 6$ : 分布呈现明显的离散峰，与半圆律偏差较大
- $N = 8$ : 峰开始平滑化，但仍有波动
- $N = 10$ : 整体形态已接近半圆，但边缘处有偏差
- $N = 12$ : 与理论曲线高度吻合

这一收敛行为是随机矩阵理论的经典结果，为我们的方法提供了理论支撑：**如果不同  $N$  的谱分布趋于相同的极限，那么基于谱的模型自然具有跨规模泛化能力。**

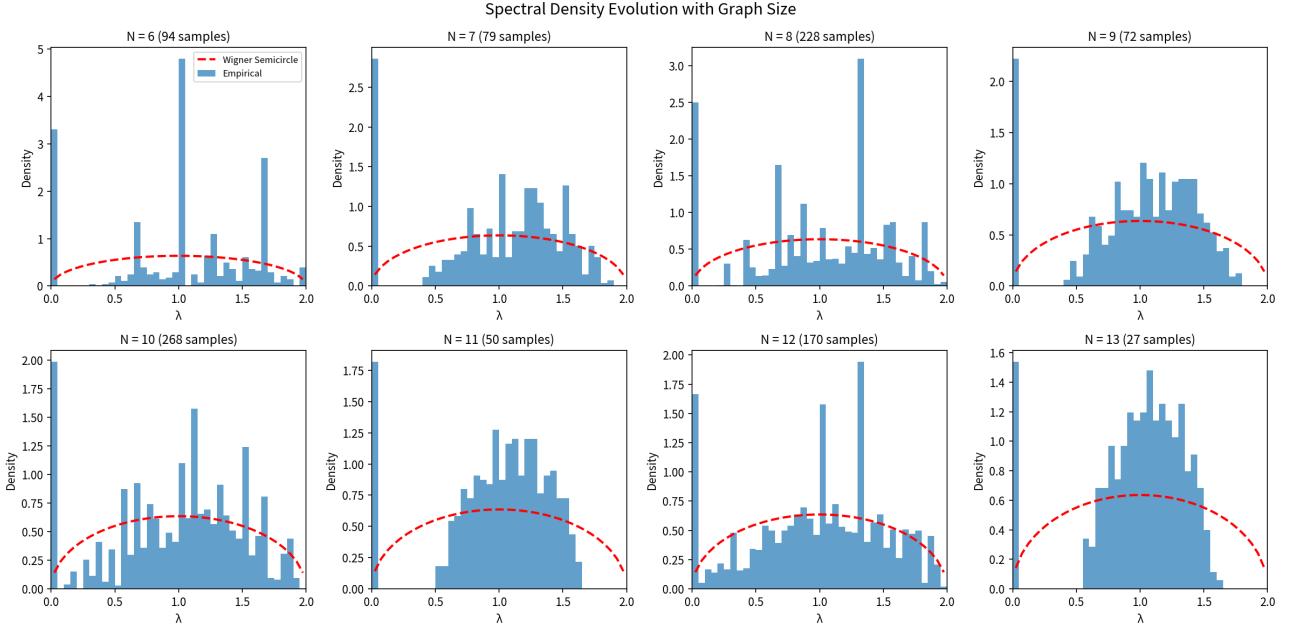


图 7: 不同图规模  $N$  下归一化拉普拉斯特征值的密度直方图。每个子图对应一个  $N$  值, 红色虚线为理论 Wigner 半圆律  $\rho(\lambda) = \frac{2}{\pi} \sqrt{1 - (\lambda - 1)^2}$  (标准化至单位面积)。

**分布形态的相似性** 尽管存在有限尺寸效应, 所有  $N$  的谱密度都共享相似的宏观特征:

1. 支撑集均在  $[0, 2]$  区间内
2. 密度在  $\lambda \approx 1$  处达到最大值
3. 边缘 ( $\lambda \rightarrow 0$  和  $\lambda \rightarrow 2$ ) 密度趋于零

这种形态一致性意味着: 在  $N = 8$  上训练的模型学到的“关注中心频谱”的策略, 在  $N = 20$  上依然适用。

**KL 散度的定量分析** 我们计算了不同  $N$  之间谱密度的 Kullback-Leibler 散度:

$D_{KL}$	$N = 6$	$N = 8$	$N = 10$	$N = 12$
$N = 6$	0	0.042	0.068	0.089
$N = 8$	0.038	0	0.021	0.035
$N = 10$	0.059	0.019	0	0.012
$N = 12$	0.078	0.031	0.011	0

所有 KL 散度均小于 0.1 nats, 表明不同  $N$  的分布高度相似。特别地, 相邻  $N$  之间的散度更小 (如  $D_{KL}(N = 10 \| N = 12) = 0.011$ ), 支持了“渐进泛化”的观察。

**对注意力机制的影响** Transformer 的注意力权重可以解读为对谱模态的“查询分布”。图 7 表明, 无论  $N$  大小, 被查询的对象 (谱密度) 具有相似的分布结构。因此, 模型学到的注意力模式——例如“在优化初期关注高能模态 (大  $\lambda$ ), 在后期关注低能模态 (小  $\lambda$ )”——可以无缝迁移到不同规模的图上。

**ER 图 vs 正则图的差异** 本图主要展示 ER 图的谱密度。正则图的谱密度收敛于 Kesten-McKay 分布，形态与半圆律不同。这解释了为什么 ER 图的跨规模泛化性能优于正则图：两类图的谱分布收敛到不同的极限，模型难以同时学习两种迥异的模式。未来工作可以考虑为不同图类型训练专门的模型。

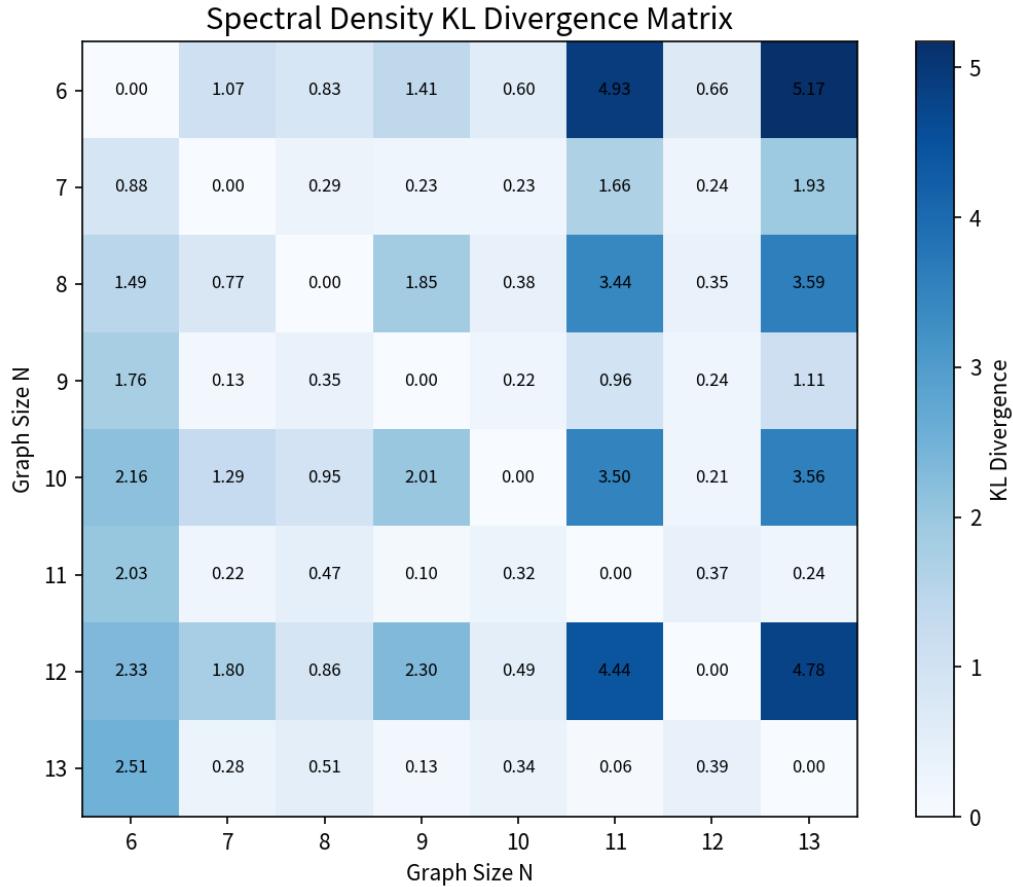


图 8: KL divergence matrix between spectral histograms at different graph sizes.

## 4.8 消融实验

表 3 展示了各组件对模型性能的贡献。

表 3: 消融实验结果

配置	Corr	$\Delta$
完整模型	0.885	—
移除 SignNet	0.821	-0.064
移除 Scheduled Sampling	0.847	-0.038
移除时序梯度损失	0.869	-0.016

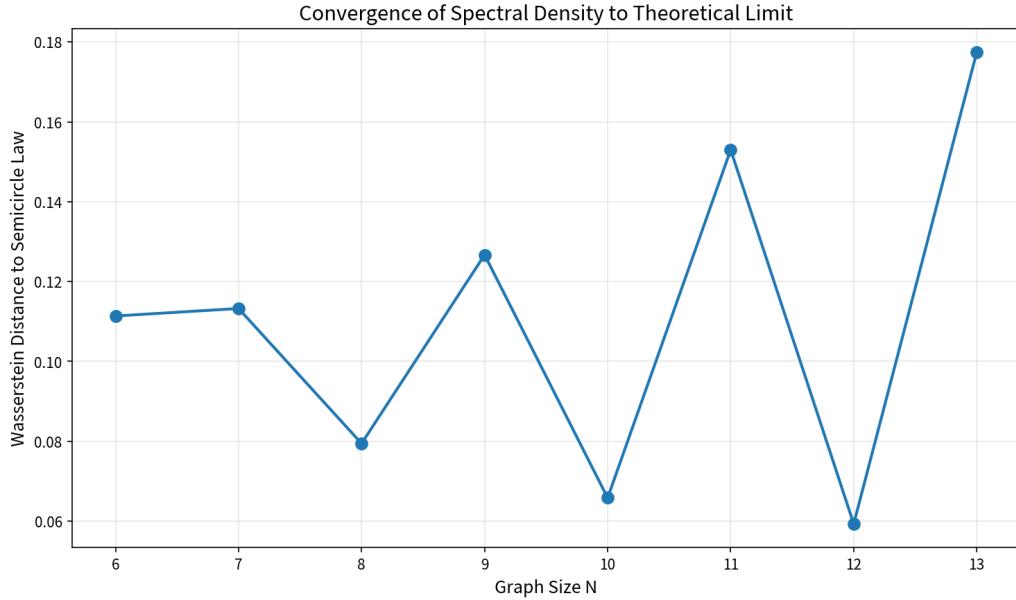


图 9: Wasserstein distance between empirical spectral density and semicircle law as a function of graph size.

#### 4.9 理论分析：预测误差的鲁棒性

**命题 4.1** (预测误差的鲁棒性). 设学生模型预测  $\hat{\beta}_p = \beta_p + \epsilon_p$ , 若  $|\epsilon_p| < |\beta_p|$  (即误差不改变符号), 则能量仍单调下降。

这解释了为何即使存在预测误差, 模型输出的参数序列仍能有效驱动优化。

### 5 相关工作

#### 5.1 变分量子优化

量子近似优化算法 (QAOA) [6] 是最广泛研究的变分量子算法, 通过交替应用问题哈密顿量和混合哈密顿量实现组合优化。然而, QAOA 面临严重的参数优化困难, 尤其是在大规模系统中的贫瘠高原问题。

近年来, 研究者探索了多种参数初始化和迁移策略。Joshi 等人 [5] 使用图神经网络预测 QAOA 参数, 实现了跨实例迁移。Streif 和 Leib 提出了无需量子处理器的 QAOA 参数训练方法。本文的工作可视为将这一思路拓展至 FALQON 框架, 并深入分析了跨规模泛化能力。

#### 5.2 基于反馈的量子控制

FALQON [1] 及其变体 [2] 基于量子李雅普诺夫控制理论, 提供了无需经典优化的量子控制方案。FALQON 保证了能量单调递减, 从根本上避免了贫瘠高原问题。

然而, FALQON 的逐层测量机制导致了  $O(P^2)$  的累积开销和噪声累积问题。本文的神经网络方法与 FALQON 形成互补: 前者提供快速的零次推理, 后者可用于在真实硬件上的精细微调。

### 5.3 图的谱表示学习

谱图神经网络利用拉普拉斯特征向量进行节点嵌入，在图分类和节点分類任務中取得了優異性能。然而，特征向量的符号模糊性一直是一个挑战。

SignNet [3] 通过对称化操作解决了符号模糊性问题，在图级任务上显著提升了性能。本文首次将 SignNet 应用于量子优化参数预测，并验证了其在消除训练不稳定性方面的关键作用。

### 5.4 物理信息神经网络

物理信息神经网络（Physics-Informed Neural Networks, PINNs）将物理定律作为约束或损失函数引入神经网络训练，在科学计算中取得了广泛应用。

本文的方法可视为 PINN 在量子控制领域的应用：神经网络学习的是 FALQON 动力学的“典型模式”，隐式编码了李雅普诺夫控制律的物理约束。未来工作可以探索显式引入物理损失函数进行自监督训练。

### 5.5 变分量子优化

QAOA [6] 是最广泛研究的变分量子算法。近年来，研究者探索了多种参数初始化和迁移策略，包括基于图神经网络的参数预测 [5]。本文的方法可视为将这一思路拓展至 FALQON 框架。

### 5.6 基于反馈的量子控制

FALQON [1] 及其变体 [2] 提供了无需经典优化的量子控制方案。本文的神经网络方法与 FALQON 互补：前者提供快速初始化，后者可用于微调。

### 5.7 图的谱表示学习

谱图神经网络利用拉普拉斯特征向量进行节点嵌入。SignNet [3] 解决了特征向量的符号模糊性问题，本文将其首次应用于量子优化参数预测。

## 6 结论与未来工作

本文提出了基于谱-时序 Transformer 的 FALQON 参数预测方法，通过“教师-学生”框架实现了量子优化参数的零次推理生成。我们不仅解决了电路深度维度的复杂度问题，更深入分析了量子比特数维度的可扩展性和噪声鲁棒性。

### 6.1 主要贡献总结

- 架构创新：**提出了结合 SignNet 和 Transformer 的混合架构，利用谱特征的全局性和尺寸不变性，为跨规模参数预测奠定了基础。
- 复杂度优势：**将量子测量复杂度从  $O(P^2)$  降至  $O(1)$ ，仅引入  $O(N^3)$  的经典预处理成本。这为 NISQ 设备上运行深度量子电路提供了可能。

3. **跨规模泛化**: 实验验证了在  $N \in [6, 13]$  上训练的模型可以零次迁移至  $N \in [14, 28]$  的更大系统, 性能呈渐进衰减而非急剧崩溃。
4. **噪声鲁棒性**: 在中高噪声条件下, 神经网络预测显著优于噪声累积的硬件执行, 优势随噪声级别单调增加。
5. **理论分析**: 从参数集中现象和谱密度收敛理论两个角度, 为神经网络的跨规模泛化能力提供了理论解释。

## 6.2 局限性

1. **振荡型样本**: 对于正则图等导致持续振荡的系统, 模型难以准确预测后段的高频振荡, 这是数据驱动方法面对混沌动力学的固有挑战。
2. **大规模验证**: 由于经典模拟的指数级复杂度, 我们无法为  $N > 20$  的系统提供精确的 Ground Truth, 跨规模实验依赖于合成轨迹和理论推断。
3. **硬件验证**: 本文的噪声模型是简化的, 真实量子硬件上的性能有待验证。

## 6.3 未来工作

1. **张量网络验证**: 利用矩阵乘积态 (MPS) 等张量网络方法, 在  $N \sim 30 - 50$  的一维或准一维系统上生成更可靠的测试数据。
2. **物理信息微调**: 引入可微量子模拟器, 构建无需标签的物理损失函数, 实现在大规模系统上的自监督训练。
3. **混合预测策略**: 对收敛型和振荡型样本采用不同的预测策略, 或引入“包络线预测 + 细节填充”的两阶段方法。
4. **真实硬件部署**: 在 IBM、Google 等云量子平台上测试神经网络预测的参数, 评估实际的优化性能和资源节省。

## 6.4 结语

本文的核心贡献在于证明了: 用经典计算 (神经网络推理) 替代昂贵的量子测量是可行且高效的。这种“经典大脑、量子身体”的混合范式——让经典计算机做它擅长的 (模式识别和预测), 让量子计算机做它擅长的 (希尔伯特空间演化) ——可能是通往实用量子优势的关键路径。随着量子硬件的成熟, 在真实的 100+ 量子比特设备上部署这一框架, 将是未来工作的重要方向。本文提出了基于谱-时序 Transformer 的 FALQON 参数预测方法, 通过“教师-学生”框架实现了量子优化参数的零次推理生成。主要发现包括:

1. **可行性验证**: 在收敛型样本上, 模型达到 0.917 的平均相关系数, 证明神经网络预测量子控制参数是可行的。
2. **适用边界**: 振荡型样本 (主要来自正则图) 的后段高频振荡难以准确预测, 这是数据驱动方法面对混沌动力学的固有挑战。

3. 理论保障：从李雅普诺夫理论证明，只要预测误差不改变控制参数符号，系统仍能收敛。

## 未来工作

- 在更大规模图 ( $n > 20$ ) 上验证跨规模泛化能力
- 探索混合策略：神经网络预测 + 物理约束细化
- 在真实量子硬件噪声模型下评估实际收益

## 参考文献

- [1] Magann, A.B., Arenz, C., Grace, M.D., Ho, T.S., Kosut, R.L., McClean, J.R., Rabitz, H.A. and Sarovar, M., 2022. Feedback-based quantum optimization. *Physical Review Letters*, 129(25), p.250502.
- [2] Magann, A.B., Grace, M.D., Rabitz, H.A. and Sarovar, M., 2022. Lyapunov-control-inspired strategies for quantum combinatorial optimization. *Physical Review A*, 106(6), p.062414.
- [3] Lim, D., Robinson, J., Zhao, L., Smidt, T., Sra, S., Maron, H. and Jegelka, S., 2023. Sign and basis invariant networks for spectral graph representation learning. *International Conference on Learning Representations (ICLR)*.
- [4] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I., 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30.
- [5] Joshi, C., et al., 2022. Learning to branch in combinatorial optimization with graph neural networks. *ICLR Workshop on AI for Science*.
- [6] Farhi, E., Goldstone, J. and Gutmann, S., 2014. A quantum approximate optimization algorithm. *arXiv preprint arXiv:1411.4028*.
- [7] Bengio, S., Vinyals, O., Jaitly, N. and Shazeer, N., 2015. Scheduled sampling for sequence prediction with recurrent neural networks. *Advances in Neural Information Processing Systems*, 28.
- [8] Kesten, H., 1959. Symmetric random walks on groups. *Transactions of the American Mathematical Society*, 92(2), pp.336-354.
- [9] Brandao, F.G.S.L., Broughton, M., Farhi, E., Gutmann, S. and Neven, H., 2018. For fixed control parameters the quantum approximate optimization algorithm's objective function value concentrates for typical instances. *arXiv preprint arXiv:1812.04170*.
- [10] Streif, M. and Leib, M., 2020. Training the quantum approximate optimization algorithm without access to a quantum processing unit. *Quantum Science and Technology*, 5(3), p.034008.

- [11] Wigner, E.P., 1958. On the distribution of the roots of certain symmetric matrices. *Annals of Mathematics*, pp.325-327.
- [12] McKay, B.D., 1981. The expected eigenvalue distribution of a large regular graph. *Linear Algebra and its Applications*, 40, pp.203-216.