深圳市大数据研究院
Shenzhen Research Institute of Big Data

White Paper on
Scientific Research Achievements of
Shenzhen Research Institute of Big Data

深圳市大数据研究院
科研成果白皮书

深圳市大数据研究院

Shenzhen Research Institute of Big Data

序言 / PREFACE

科技发展史，本质上是一部人类不断突破认知边界、携手共进的历史
科技突破，是代际传承的接力、文明互鉴的交响
也是人类面对未知时最坚定的探索宣言
当科技以前所未有的深度与广度重塑世界图景时
当我们叩问“科技为何、科技何为”时
答案始终清晰
立足现实土壤、扎根生活实践
科技
让每一项探索都紧扣时代脉搏，每一次突破都回应真实需求

我们聚焦AI赋能通信网络，让信息传输突破带宽与时延的瓶颈，为万物互联筑牢“数字底座”
我们深耕算力基础设施，为国家打造“超级大脑”，让海量数据的计算与存储成为可能
我们推动垂直领域应用，让人工智能走进车间、诊室，在具体场景中释放技术红利
我们探索运筹优化与供应链，用算法化解复杂系统的资源调配难题，让“流动的中国”更高效、更有韧性
我们攻坚科学计算与工业软件，打破核心技术“卡脖子”枷锁，让“中国智造”走向高质量发展

科技的价值，从来不在“纸上谈兵”，而在“落地生根”

深圳市大数据研究院坚持每一项研究以重大问题为导向
坚持让创新成果跨越实验室边界
坚持每一项研究都紧贴人民所盼、国家所需、社会所急
以科技之力推动世界进步、全球发展

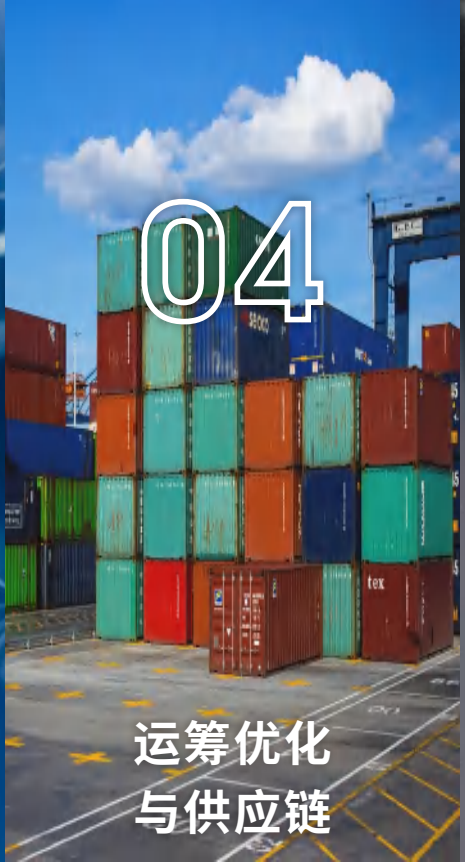
谨以此序，致敬科技发展征程中每一次探索。

罗智泉

中国工程院外籍院士
加拿大皇家科学院院士
深圳市大数据研究院院长

CATALOG

03 AI算力基础设施与垂直领域应用



04 运筹优化与供应链

05 科学计算与工业软件

AI-Empowered Network Communication

AI赋能网络通信

深圳市大数据研究院
Shenzhen Research Institute of Big Data

ACADEMIC ACCUMULATION
学术积累

学习优化理论与方法及其在5G网络中的应用

01 项目背景

第五代移动通信（5G）网络的规模部署为实现更高速率、更低时延和更广连接提供了技术基础。然而，要充分发挥5G硬件设备的性能潜力，使其在各种复杂多变的现实通信场景下均能达到最优服务效果，大规模网络参数的智能配置（即网络优化）是亟待解决的核心挑战。

该挑战源于5G网络固有的复杂性与规模

- 用户连接数量庞大且动态变化
- 可调节的网络参数（如发射功率、波束方向、小区切换阈值、资源分配策略等）维度极高
- 所承载的服务类型繁多且对网络性能（速率、时延、可靠性）要求各异
- 传统基于严重简化理论模型的优化，无法捕捉真实环境中密集城区、室内外混合、高速移动、电磁异构等复杂特征

面对这种超大规模、超高维度、强非线性、动态实时的复杂优化问题，传统网络优化方法已显现出根本性局限。

这类方法主要依赖

- 人工经验的规则和启发性设置
- 有限的实地路测（Drive Test）数据进行统计分析
- 基于严重简化的网络模型进行理论推导

其结果往往是

- 参数配置过程效率低下、成本高昂
- 难以捕捉网络全局状态和参数的相互关联
- 高度依赖特定工程师的技能
- 无法实现复杂环境和大规模网络下的全局或近似最优性能

工业界

通过改进数据采集手段（如众测、无人机测试）或流程协作平台（如区块链调度），提升了数据量和协作效率，但未能实质性地变革“依赖专家经验人工决策”这一核心环节。全球性能评估持续反映出显著的优化空间。

学术界

近年尝试引入基于数据驱动的智能方法，例如：

- 使用深度置信网络等模型优化特定小区域流量
- 应用贝叶斯方法或“多臂老虎机”算法进行发射功率分配
- 采用干扰模型、马尔可夫决策过程或模拟退火优化特定目标（如覆盖、能耗、加权速率）

然而，这些研究普遍集中于单个通信链路、单小区或少量小区构成的“小规模”或“理想化”场景模型。当将这些模型和算法应用于包含成千上万个相互干扰、参数耦合的网络节点和复杂物理环境的全网级优化问题时，其面临重大局限：

- **模型可扩展性差：**小规模模型的假设和结构难以扩展到大网络
- **优化算法效率不足：**计算复杂度可能随网络规模指数增长，无法满足实时或准实时要求
- **机理适配性欠缺：**黑盒模型可能忽略关键的通信物理规律，导致决策物理不可实现或效果不佳
- **泛化能力有限：**在一种场景或特定规模下有效的模型/算法，在变化的环境或更大网络中可能表现显著下降

因此，如何在复杂物理规则约束下，构建能高效求解超大规模、超高维度网络参数配置优化问题的新理论、模型与算法，是该领域当前面临的核心科学瓶颈。其本质在于超越人工经验和局部优化的范畴，发展能够紧密结合底层通信物理机理、充分利用网络数据资源、具备高效全局寻优能力并能适应复杂动态环境的新型智能化优化范式。

本项目开展面向大规模复杂通信网络的智能优化与建模基础理论与关键技术研究，旨在突破现有方法在规模、效率、机理融合性以及求解稳定性等方面的根本限制。

02 创新内容

破局关键：为网络优化搭建一个兼具物理可解释性与数据适应性的“物理与数据双驱”基座模型。

本项目聚焦于5G网络优化的核心问题，围绕网络优化建模与传统优化方法设计展开系统性研究，提出了一系列具有原创性的理论与方法。

在网络环境感知建模方面

项目组依托现网真实数据，结合稀疏信号处理技术，强化对信道传播环境的空时频三维认知，提出空时频三维智能协同感知框架，并创新性地构建了信道模型的多维角度功率谱统计模型，实现了在地化信道精准统计建模。

在网络系统覆盖性能建模方面

项目组提出了基于图注意力神经网络的“数据—模型”双驱动覆盖建模方法，在低采样率条件下显著降低重建误差，有效提升覆盖预测精度。

在大规模网络参数优化方面

项目组以网络覆盖、能源效率及用户体验为核心目标，综合考虑网络环境、用户行为特征、4G/5G协作模式及资源分配策略等多因素，利用零阶优化等方法，构建适应高维度的网络结构参数、用户接入参数及系统资源参数的动态优化方案，并研发了多种基于人工智能的新型优化算法，推动实现网络参数的快速与高效优化。

03 应用场景

网络优化

网络性能建模

04 重要成果

项目在5G网络优化领域取得了多项标志性成果。在网络环境建模方面，提出的空时频三维智能协同感知框架及多维角度功率谱统计模型成果已发表于通信领域顶级期刊IEEE TWC，获得国际学术认可。

在覆盖性能建模方面

所研发的基于图注意力神经网络的覆盖建模方法，在仿真数据集5%稀疏采样率下，十个地区的覆盖电平重建RMSE达到2.8384dB，相较于现有最优方法误差降低17%；在江西移动低空现网路测数据中，RSRP覆盖建模的MAPE为6.98%，比现有最优算法降低9.4%，显著满足运营商对低空覆盖建模的高质量需求。

在大规模网络参数优化方面

形成了“基于图神经网络算法展开的5G网络结构参数优化”“基于多智能体强化学习的5G接入参数优化”等创新性算法，并成功构建面向5G现网优化的高精度仿真平台。经平台与算法联合测试，预计地面平均谱效从931bits/RB提升至978.59 bits/RB，提升5.04%，边缘用户谱效预计提升10.12%，现网实测低空覆盖率提升15.79%，有效验证了所提方法的优越性与可行性。

图1 实测低空覆盖率（> -70dm）：优化前16.94%（左），优化后32.73%（右）

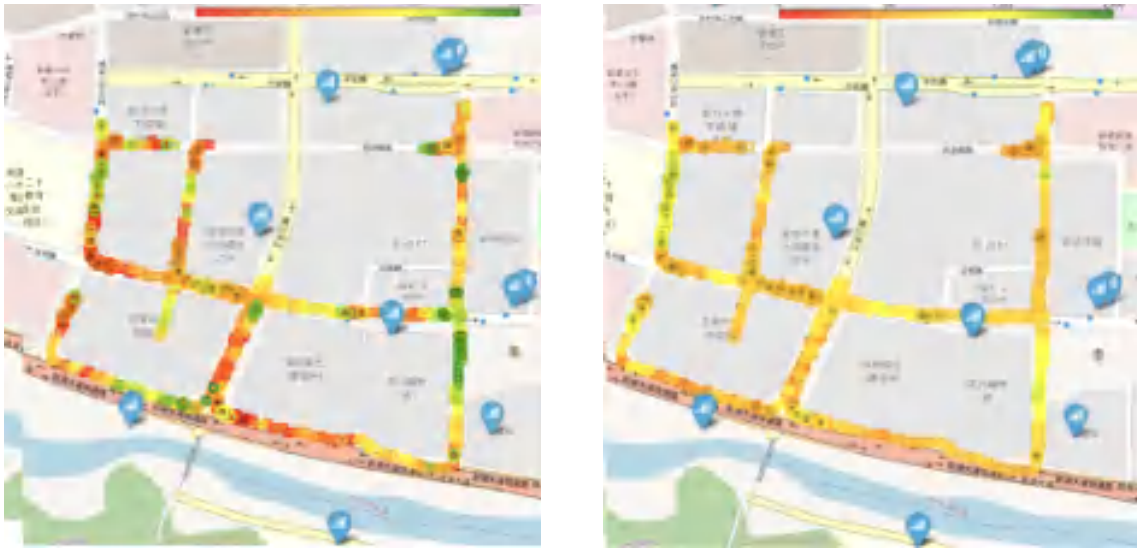


图2 预测地面网络平均谱效：优化前931 bits/RB（左），优化后978.59 bits/RB（右）

移动网络结构优化的数学理论与优化方法

01 项目背景

当前移动网络优化面临以下核心挑战

- **数据局限性：**依赖路测数据和蜂窝网架构，难以应对突发业务和边缘覆盖问题
- **静态优化模式：**传统方法基于离线分析，无法实现动态、智能化的网络调整
- **高维复杂性：**网络参数（如波束、信道、用户流量）维度高，优化难度大

项目旨在通过数学建模与智能算法，推动移动网络从“局部感知-静态优化”向“广域感知-动态自主优化”的范式升级

02 创新内容

核心研究内容

- **高维数据空间建模：**解决MR数据缺失问题，提出波束空间离散化与栅格定位算法（定位误差<6dB）
- **空时性能建模：**基于生成式模型（图神经网络+神经辐射场），预测用户流量、信道状态和频谱效率（预测误差降低20%）
- **黑盒约束优化：**开发高维梯度预测、降维分解和在线优化算法，提升求解效率30%
- **网络结构优化验证：**设计基站协同参数优化方案（如CoMP技术），搭建三小区硬件验证平台

技术创新点

- **数据层面：**首创波束空间离散化框架，实现高维MR数据补全与映射
- **算法层面：**结合生成式AI（如扩散模型）与数学优化理论，突破传统均值预测局限
- **应用层面：**提出“准在线优化”模式，支持动态环境下的实时决策

03 应用场景

网络覆盖增强	网络节能降耗	智能运维
场景：解决边缘用户信号弱、干扰的大问题（如密集城区、室内死角）。 **方案：**通过基站分簇与协同传输（CoMP），提升边缘用户频谱效率15%。	**场景：**5G基站高能耗问题（占运营商总电费30%以上）。 **方案：**动态调整基站工作模式与功率参数，降低全网能耗8%。	**场景：**网络故障预测与参数自优化。 **方案：**基于生成模型的性能预测，实现故障提前预警和资源自动调配。

TECHNOLOGICAL BREAKTHROUGH
技术攻关

多模态信息融合混合波束赋形CSI重构

01 项目背景

在大规模MIMO系统中，数字波束赋形虽能提供最高增益，但因所需射频链路数量与天线数相当，导致成本大幅上升。为降低复杂度和成本，混合波束赋形（HBF）结合模拟与数字波束赋形，能够在减少射频链路的同时接近纯数字波束赋形的性能。然而，HBF的实施依赖于完整的模拟通道信道状态信息（CSI），这在U6G全连接HBF场景中尤为困难，因数字通道数远少于模拟通道数。为此，需利用SRS测量、PMI测量及信道图谱等多模态信息，通过高效融合手段实现精准的全信道重构，从而提升下行链路的频谱效率。

02 创新内容

本项目提出了一种面向U6G的混合波束赋形（HBF）系统中多模态信息融合的信道状态信息（CSI）重构技术，主要创新点包括：

精准测量与重构的协同优化

通过结合信道的时间相关性、空域稀疏性及低秩性，设计自适应HBF模拟权，提升SRS和PMI测量的精度。

多源信息融合机制

融合SRS测量、PMI反馈及信道图谱（如角度-时延谱）等多模态信息，提升全信道估计精度。

高效CSI重构算法

在数字通道远少于模拟通道的受限条件下，实现高精度信道重构，逼近全数字波束赋形（DBF）性能。

面向低移动性的U6G场景优化

针对低速用户（0.5-3km/h）设计时域联合估计与卡尔曼滤波增强机制，提升CSI稳定性与跟踪能力。

03 应用场景

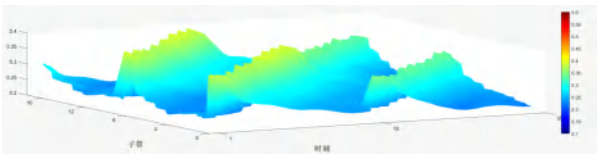
该技术适用于U6G频段（如6.7GHz）的高频大规模MIMO系统，特别是下行单小区多用户MIMO（MU-MIMO）场景，典型配置为：

基站端：480模拟通道 + 32~48（典型42）数字通道的HBF架构

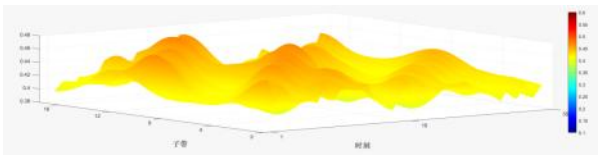
用户端：每用户8接收天线（8Rx）

测量机制：SRS子带跳频周期为40ms，共17跳频点，PMI反馈采用3GPP R16/R18码本

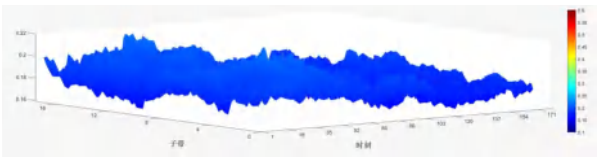
目标性能：相较传统时分扫描与卡尔曼滤波基线方法，信道重构相关性平均提升0.2以上，并逼近全数字波束赋形谱效的90%。适用于未来高密度、高性能的5G-Advanced及6G无线通信系统，特别是在高带宽、高频率、大规模天线部署的场景中。



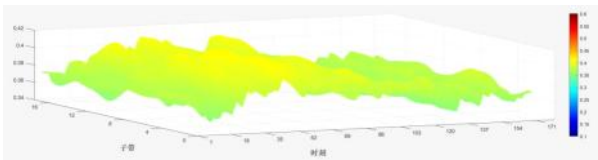
256数字通道在不同子带不同时刻下采用最小二乘法估计信道，取得平均信道相关性0.2683



256数字通道在不同子带不同时刻下采用优化方案估计信道，取得平均信道相关性0.4250，较基线提升58.4%



48数字通道在不同子带不同时刻下采用最小二乘法估计信道，取得平均信道相关性0.1847



48数字通道在不同子带不同时刻下采用优化方案估计信道，取得平均信道相关性0.3801，较基线提升105.8%

04 发明专利（已授权）

- 一种基于批标准化层参数修正联邦学习的图像分类方法
- 一种基于异构混合数据的全局梯度双追踪分布式方法

分布式空地协同微弱电磁目标感知及原型系统验证

01 项目背景

通过在无人机等异构平台上部署无源感知节点，可综合利用分布式多天线积累增益与空间分集增益，实现空间频谱态势感知，提升对微弱电磁目标的探测能力。实现信号级协同处理的核心挑战在于节点间的高精度同步、远距离高速无线数据传输等技术突破。针对复杂电磁环境下的信号微弱感知挑战，亟需发展高稳健性的分布式信号处理技术。为此，在系统研制方面，本项目重点突破小型化机载电子载荷设计、高灵敏度信号侦收与处理、高精度时空同步、远距离宽带无线数据传输及系统电磁兼容性设计等关键技术。在信号处理理论方法方面，以尽可能降低各处理环节信息损失为原则，开展鲁棒融合、构型优化、直接定稳等多项信号处理方法创新，综合实现空间增益、时间增益、相干处理增益累积，全面提升系统的侦察处理能力。

02 痛点问题

微弱信号分布式感知难题

微弱目标信号在复杂电磁环境下往往掩盖在噪声与背景干扰中，单节点捕获的信噪比极低，往往小于-10dB，传统信号处理方法难以有效提取微弱信号。如何通过分布式多节点信号级协同，挖掘微弱信号时-空-频多维增益获取是最大限度地提升微弱目标信号的探测能力的基础性问题。

通信受限下的分布式联合检测定位难题

分布式融合依赖节点之间低误码率的稳定通信链路进行数据交互，与实际面临的的极端通传条件严重失配。由于通信带宽受限或较高误码率，节点间的数据交互能力大幅降低，导致传统联合检测定位方法失效。如何在有限数据交互和高误码率的条件下，实现稳健分布式联合检测定位是亟需突破的关键难题。

检测定位跟踪一体化工程难点

分布式空地协同系统中的不同节点需要在动态环境下同时完成目标的检测、定位与跟踪任务，分布式节点的异构性（如无人机与地面节点的差异）和系统资源限制要求设计轻量化的高效协同处理算法。如何工程化实现跨节点的全流程协同一体化处理，是实现分布式信号鲁棒融合工程化落地的关键。

03 关键技术

分布式融合感知理论

针对分布式节点信号特征构建多维信号物理模型，从时空频多维推导结构化协方差估计器，对多维信号统计量进行充分累积，实现对微弱信号的精准捕获。通过大量实测验证了理论推导结构化协方差检测器的微弱信号捕获能力，为后续系统级研发奠定理论基础。

分布式鲁棒融合感知方法

针对数据高误码通信链路条件，通过大量实测数据分析，揭示了数据误码带来的重尾分布效应。基于结构化协方差，构建了隐变量表示的t分布鲁棒联合检测定位统计量，克服数据误码和缺失带来的检测定位方法失效难点。进一步，针对部分通传链路失效的极端条件，构建了结构化矩阵补全检测器，解决部分节点数据无交互下的鲁棒检测定位。

检测定位跟踪一体化处理

针对分布式空地协同系统的检测定位与跟踪一体化处理需求，自研高精度时频同步技术,构建基于异构节点协同的轻量化处理框架，实现了节点间信号检测、数据融合与轨迹跟踪的全流程一体化处理。重点突破小型化机载电子载荷设计、高灵敏度信号侦收与处理、高精度时空同步、远距离宽带无线数据传输及系统电磁兼容性设计等关键技术，实现复杂环境下微弱目标高效感知的工程化落地验证。

04 系统简介

系统概况

系统包含5个地面感知节点，3个空基感知节点（无人机搭载），3靶标节点（无人机搭载）和自组网数传分系统。系统支持灵活分布式部署与构型优化，通过自组网无线数传将各节点采集信号在显控中心融合，实现高灵敏度检测和高精度定位跟踪。相比于单节点，系统在6节点条件下融合检测能力可有效提升12dB以上，联合检测定位精度相比于传统定位方法可提升30%以上。

图1 分布式空地协同微弱电磁目标感知原型系统

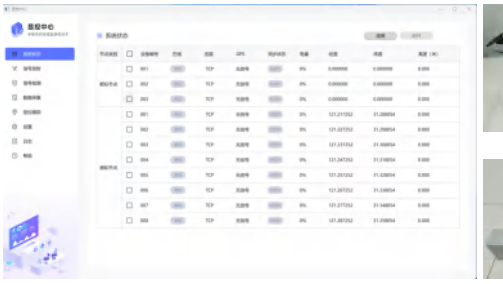


图2 显控中心界面

图3 无人机搭载

图4 感知节点

系统部署及能力情况

在地面和空中部署分布式感知节点进行频谱信号采集，如下图所示，通过自组网无线数传将采集信号回传至显控中心进行融合处理，实现靶标无人机的检测定位与跟踪。

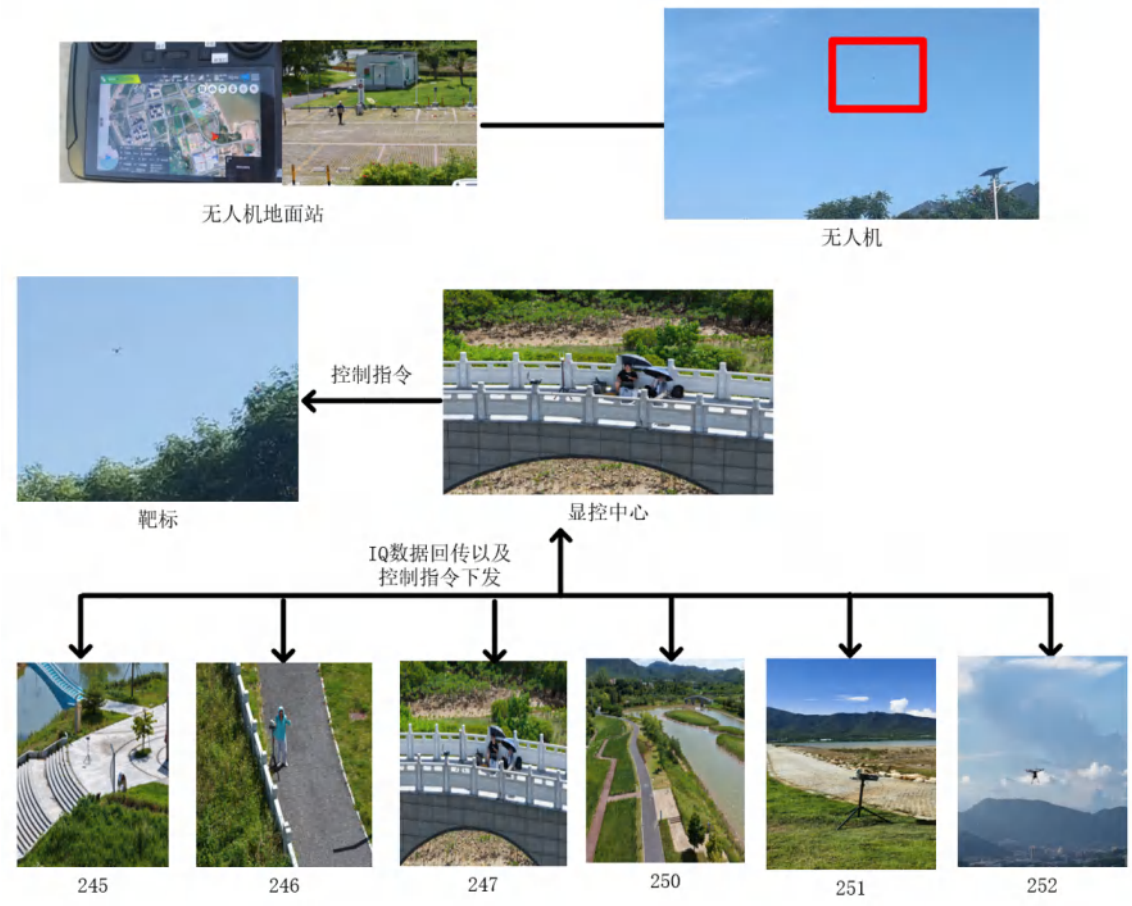


图5 外场部署测试分布图

通过信号级空地协同融合，可实现稳定的靶标联合检测定位，不同位置的联合检测定位结果如下图所示，定位精度小于15米（200kHz窄带辐射信号）。

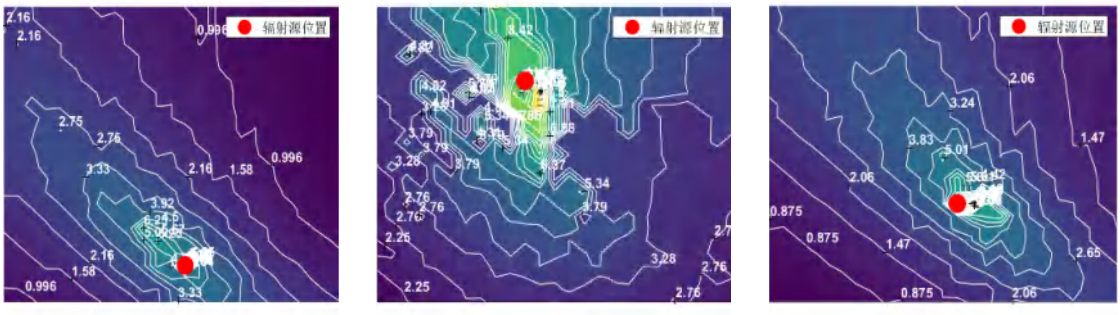
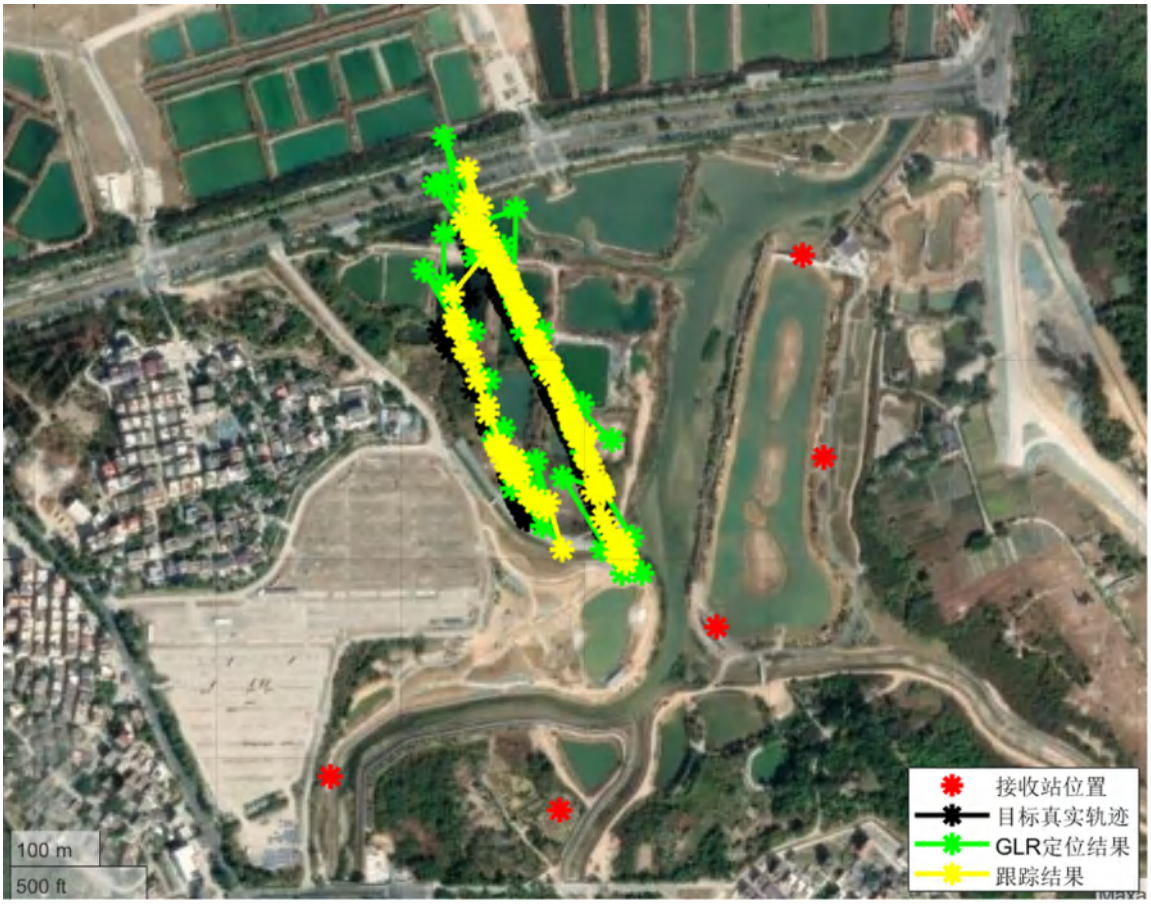


图6 3帧数据联合检测定位结果图

显控中心持续进行空地信号采集与协同融合，实现对靶标的持续跟踪，结果如下图。



基于盲波束赋形算法的智能反射面技术

01 项目背景

作为6G的研究热点技术之一，智能反射面（Intelligent Reflecting Surface, IRS）可以通过控制平面上大量低成本的无源反射元件的相位智能地重构无线传播环境，显著地提升无线通信网络的性能。因此，IRS提供了突破网络结构限制进一步提升网络覆盖和容量的新思路。此前学术界提出的大部分超表面技术需要基于信道状态信息的估计，不适宜在现有的5G网络中实现。

智能反射面技术面临的挑战

对于当前的sub-6GHz频段而言，真实信道中多径成分非常丰富，难以准确地估计主要能量的来波方向。

对于多径丰富的场景，IRS发挥作用需要与基站交互信道CSI信息，进而要设计独立的信息传输链路以及IRS侧具备通信信号接收和处理的相应功能，当前网络架构与协议并不支持。

引入信道交互以及IRS接收机能力后，对应IRS成本大幅抬升，若与基站成本相当，则难以大规模推广部署。

02 创新内容

为了应对上述挑战，基于数据驱动的思想，罗智泉教授带领的联合研发团队创新性地提出了基于盲波束赋形的反射面相位控制技术。具体而言，该项技术不需要获取信道状态即可通过自学习建立场景化模型，得到IRS阵列在当前场景下的最优相位组合。理论证明，在处理二元相位选择$\Phi=\{0,\pi\}$时，以上算法可以渐进达到全局最优；而在处理多元相位选择$\Phi=\{0,\pi/K,2\pi/K,...,(K-1)\pi/K\}$，$K>2$时，以上算法可以渐进达到全局最优的75%以上。

03 应用效果

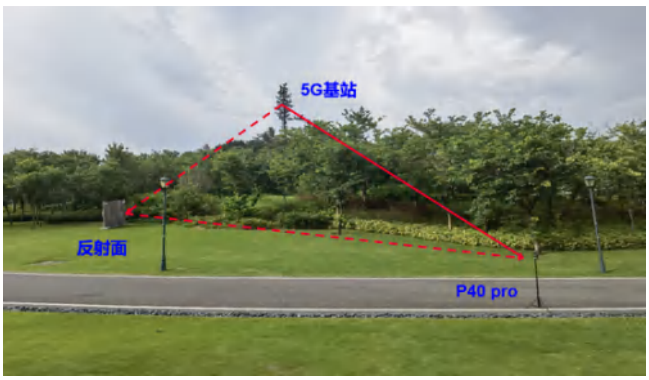
针对上述技术，罗教授带领的联合研发团队已经完成了基于真实5G网络环境（现网5G 2.6GHz频段）的测试验证，证实了基于盲波束赋形的IRS技术能够为5G网络带来两方面的性能提升：

室内5G网络覆盖补盲以及SINR提升

在室内典型的弱覆盖场景下，能有效改善区域弱覆盖，覆盖增益14dB，SINR增益12dB，速率提升200%。

室外5G网络RANK提升

在室外的空旷场景下，多径较少，RANK较低。通过引入IRS，RANK提升一阶，速率增益为50%。



CUSTOMER CASES
客户案例

低空网络覆盖预测与规划关键技术研发

低空 网络覆盖 覆盖预测 无线网络规划

01 项目背景

2024年，“低空经济”首次被写入政府工作报告，明确其作为“新增长引擎”的战略地位。江西省积极响应国家号召，正加速布局低空经济产业，规划建设低空经济产业园，并推动低空技术在物流、农林、旅游、应急、城市管理等领域的应用。这为通信基础设施提出了明确且迫切的需求。尽管4G/5G网络已实现对地面98%以上人口的覆盖，但在面向低空（通常指海拔1000米以下）场景时，现有网络面临严峻挑战。江西移动围绕低空智能网联技术、低空经济产业发展、新质生产力赋能低空经济场景应用等关键议题进行全面部署。

02 行业问题

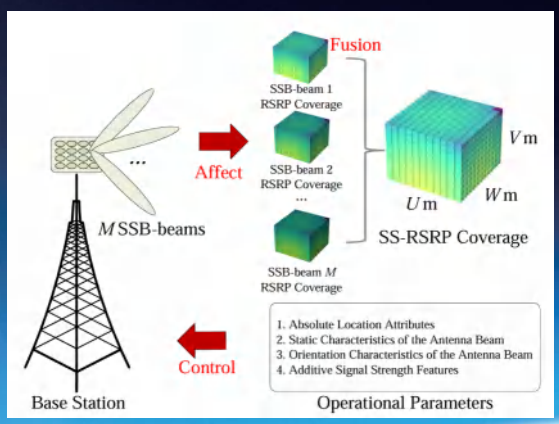
低空产业需求逐步放大，低空网络建设初期的网络信息服务能力无法满足需求。同时，低空网络覆盖预测和规划方案缺乏积累，支撑手段不足。

·江西移动·

电信运营商

03 解决方案

- 基于某个区域的低空无人机采集，训练一个模型，实现任意基站工参到任意区域低空覆盖的预测。
- 通过在江西南昌采集低空路测数据，进行算法模型训练
- 在江西赣州，实现150米低空覆盖结果预测
- 测试方案：模型输入赣州基站工参，能预测150米的覆盖，并且在真实部署后，通过在赣州进行新一轮路测RSRP验证
- 实现：MAE 6.26dB，MAPE 7.69%
- 支撑赣州完成签订1项低空网络类合同，实现合同收入数百万元。



江西赣州实现150米低空覆盖结果预测

测试方案

150米范畴

模型输入赣州基站工参，能预测覆盖

真实部署后

RSRP验证

通过在赣州进行新一轮路测

实现

MAE 6.26dB
MAPE 7.69%

支撑赣州

完成1项低空网络类合同

合同收入数百万元

面向空地一体化融合的无线网络覆盖和性能优化关键技术研发

低空　网络覆盖　覆盖预测　无线网络规划

01 项目背景

2024年，“低空经济”首次被写入政府工作报告，明确其作为“新增长引擎”的战略地位。江西省积极响应国家号召，正加速布局低空经济产业，规划建设低空经济产业园，并推动低空技术在物流、农林、旅游、应急、城市管理等领域的应用。这为通信基础设施提出了明确且迫切的需求。尽管4G/5G网络已实现对地面98%以上人口的覆盖，但在面向低空（通常指海拔1000米以下）场景时，现有网络面临严峻挑战。江西移动围绕低空智能网联技术、低空经济产业发展、新质生产力赋能低空经济场景应用等关键议题进行全面部署。

02 行业问题

低空经济蓬勃发展，需要加强低空的无线覆盖和性能提升，增加设备的方案将带来高昂的成本支出，因此提出一种复用地面基站的方案，在基本不损害地面覆盖和性能的前提下，通过调整现有基站的工参，达到提升低空无线覆盖和性能的目标。

·江西移动·

电信运营商

03 解决方案

采用现网数据与灰盒模型联合建模，通过模型的最优化寻优，在现网中进行测试对比。

先基于实测数据进行多波束统计信道建模，然后利用所建信道预测基站参数调整后的各小区参考信号接收功率（RSRP）和信干噪比（SINR）；然后基于深度学习模型进行谱效预测，以捕捉现实复杂网络指标的随机行为；最后是基于零阶优化算法的谱效优化，以解决梯度难以计算的优化问题，对基站参数进行优化。这三个阶段模块的输入输出相互串联，以达到最优的现实网络谱效，

优化后，牺牲地面覆盖率1.4%的同时，低空覆盖率提升15.79%，谱效提升10%

下图为低空RSRP覆盖实测结果：

优化后

牺牲地面覆盖率仅

1.4%

提升低空覆盖率

15.79%

谱效提升

10%

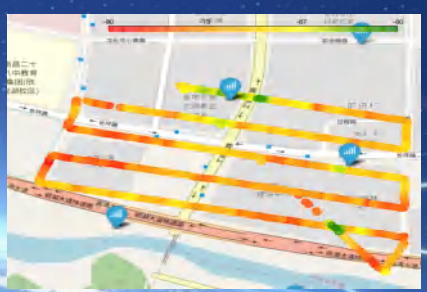


图1 天线参数优化前

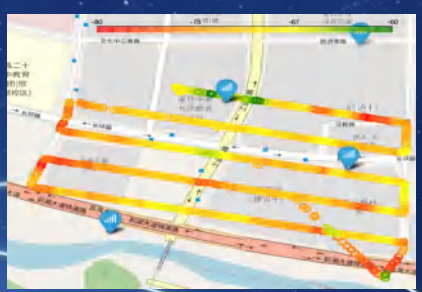


图2 天线参数出后

AI-Powered Energy Grids

AI赋能能源网络

深圳市大数据研究院
Shenzhen Research Institute of Big Data

ACADEMIC ACCUMULATION
学术积累

应对不确定性的跨区域氢储能规划与调度分层优化

01 项目背景

随着风电、光伏等可再生能源的持续增长并大规模接入电网，由其波动性与间歇性所引发的时空供需失衡问题日益突出。跨区域氢能储存系统（Hydrogen Energy Storage Systems, HESS）因能够在能源富余时储氢、需求高峰时放氢，实现区域间的能量调剂与平衡，正逐渐成为增强电力系统灵活性和稳定性的关键手段。然而，跨区域氢储能的大规模推广仍面临多重障碍：

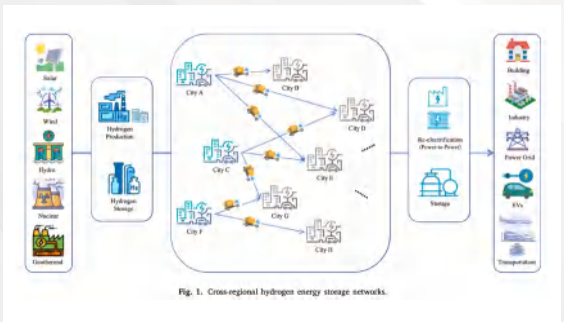


Fig. 1. Cross-regional hydrogen energy storage networks.

成本高昂。氢气的制备、压缩、储存及跨区域运输均需高额投入，在规模化尚未成熟的情况下，单位成本难以有效降低。

供需不确定性强。可再生能源输出与各地负荷呈现高度波动和区域差异，使跨区域能量传输计划复杂化。

优化求解难度大。跨区域调度涉及长周期、大规模耦合决策，传统随机优化与分布式鲁棒优化虽能处理不确定性，但在面对上百甚至上千城市时往往面临求解效率低、可扩展性不足的问题。

在此背景下，亟需一种能够兼顾不确定性处理能力、求解效率与大规模扩展能力的新型优化框架。本文正是围绕这一需求提出了创新性解决方案。通过引入动态传输定价机制并构建分层优化架构，显著简化跨区域能量协调难度，为 HESS 的规划与调度提供了一种更高效、更鲁棒且具备大规模可扩展性的优化方法。

02 核心内容

分层优化框架与动态传输定价机制

论文构建了一个由战略规划层与运行调度层组成的两层优化框架，用于统筹跨区域氢储能系统（HESS）的规划与调度。上层负责氢储能设施的选址与容量配置等投资性决策（涉及整数变量），而下层则在上层布局给定的情况下，优化各城市在24小时周期内的充放电和跨区域能量传输（仅包含连续变量）。这种主从式结构使长期投资与短期运行策略紧密协调，增强了系统整体的可控性与灵活性。

本研究的核心创新之一是动态传输定价机制的引入。该机制通过为区域间能量传输赋予动态成本，量化跨区域输能的难度，使能量流动可由价格信号自动调节：高传输价格抑制跨区流动，而低价格则鼓励在区域间进行能量互济。与传统依赖整数变量描述“是否启用传输路径”的方法不同，动态传输定价将这些组合决策隐含在连续成本参数中，使下层调度模型完全摆脱0–1变量的依赖，从而显著降低计算复杂度。此外，该机制从经济学角度刻画了不同区域间能量交换的边际代价，使跨区交易决策更加准确合理。通过这一机制，分层优化框架在保留鲁棒性与灵活性的同时，极大简化了跨区域协调的建模难度，使能量流动能够在价格驱动下高效调整。

战略规划层：模拟退火优化选址

在战略规划层中，目标是确定氢储能设施的最优部署位置及容量，使总成本最小化。该问题属于典型的组合优化，变量维度高且易陷入局部最优。为此，论文采用了模拟退火（Simulated Annealing, SA）作为上层求解算法。SA通过模拟金属退火过程实现对解空间的全局搜索。算法以随机生成的初始选址方案为基础，通过对设施布置进行小扰动来生成新方案，并依照Metropolis准则决定是否接受：若新方案降低成本则直接接受；若成本增加，则以一定概率接受，从而在搜索初期保持充分探索能力。随着迭代推进，温度逐步降低，算法接受劣解的概率下降，最终向全局最优附近收敛。本文发现，得益于SA的全局跳跃能力，上层能够有效避免陷入局部最优，获得质量较高的选址方案，为下层调度奠定了优良的基础结构。

运行调度层：最小费用流优化调度

在运行调度层，需根据上层确定的站点与容量，优化各城市在24小时内的购电量、充放电行为以及跨区域能量输送。传统方法采用线性规划（LP）建模，可精确处理需求平衡、电价变化及储能操作约束。然而随着城市数量和时间维度上升，LP的规模迅速增大，求解效率难以满足大规模场景需求。为提升求解效率，论文将LP调度模型等价转换为最小费用流（Cost Flow Transformation, CFT）问题。该方法将“城市–时间”结构映射为一个有向网络：每个节点表示城市在特定时段的状态，节点间的有向边表示可能的能量流动路径，其费用由动态传输定价与电价构成，容量由输电能力与储能功率限制决定。在此框架下，调度问题被转化为寻找满足节点供需平衡的最小费用流，从而可利用专用网络流算法高效求解。CFT的引入使得调度层在保持最优性的前提下，极大提升了求解速度与可扩展性。特别是在大规模城市网络中，CFT相较LP展现出数量级的效率提升，并与动态传输定价机制自然融合，使跨区域能量调配更加高效可靠。

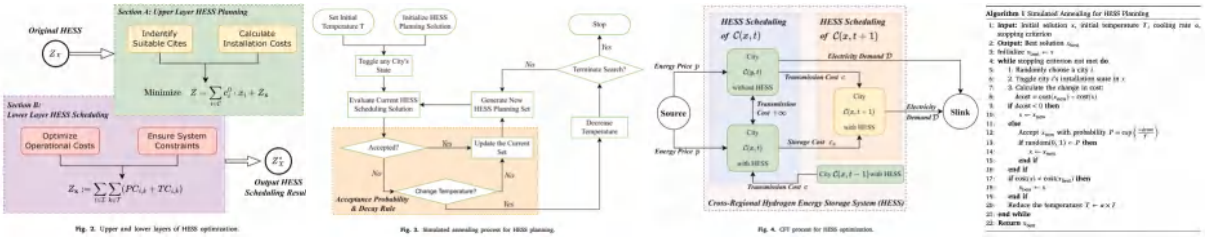


Fig. 2. Upper and lower layers of HESS optimization.

Fig. 3. Simulated annealing process for HESS planning.

Fig. 4. CFT process for HESS optimization.

03 结果

通过数值实验，论文系统验证了所提出分层优化方法在收敛性能与计算效率上的显著优势。实验从收敛特性与求解效率两个维度展开，对比了采用CFT与传统LP作为下层求解器时的性能差异。

首先，在一个包含50个城市的测试网络上，分别以CFT与LP为下层求解方式运行优化算法，并比较其迭代收敛过程。结果显示，两种方式均在约100次迭代后趋于稳定，表明所提出的分层优化框架具备可靠的收敛性。更重要的是，CFT在整个迭代过程中始终表现出更快的目标成本下降速度：尽管两者最终收敛至相同水平，但CFT能以更少时间逼近最优解，在相同迭代限时内取得更低成本，进一步证明其在求解效率上的优势。

接下来，论文深入分析了模拟退火与CFT联合求解时的搜索轨迹。目标成本随迭代的曲线总体呈持续下降趋势，但在局部阶段会出现小幅上升，这正是模拟退火在高温阶段接受次优解以跳出局部最优陷阱的体现。随着温度降低，解逐渐收敛并在全局最优附近小幅波动，表现出典型的“先探索、后收敛”的特性，说明分层优化框架在全局搜索能力与收敛稳定性之间取得了良好平衡。

最后，在计算效率方面，论文对比了CFT与LP在不同网络规模（从10到1000城市）下完成单次调度求解所需的时间。结果极具说服力：CFT的求解时间仅从0.01秒增长到约1.84秒，而LP则从0.02秒急剧攀升至2238.4秒。在最大规模的1000城市网络中，CFT的速度比LP快了超过三个数量级。虽然两者最终获得的最优调度结果一致，但CFT的高扩展性与高效性显著降低了求解成本，使分层优化方法能够实现在大规模跨区域氢储能场景中的实际部署。

综合来看，实验结果明确表明：所提出的动态传输定价+分层优化+CFT的整体架构不仅在求解质量上可靠，而且在大规模系统中具有极强的计算优势，为跨区域氢储能系统的实际应用提供了切实可行的技术路径。

Table 1
Computation time for single scheduling process.

Number of cities	CFT (s)	LP (s)
10	0.01	0.02
50	0.17	0.86
100	0.62	5.94
1000	1.84	2238.40

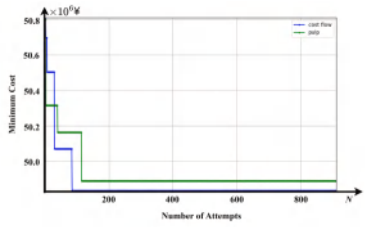


Fig. 5. Cost comparison: CFT vs. LP.

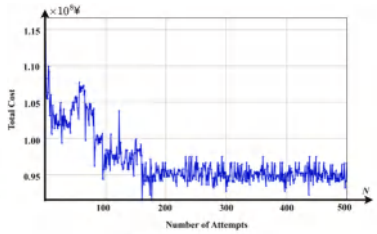


Fig. 6. Annealing steps with CFT.

04 未来展望

这项研究为跨区域氢储能的规划与调度提供了高效鲁棒的新思路，不仅优化了城市间的能源调配决策，还提升了都市能源系统的韧性和可持续性。未来，该框架有望应用于实际能源互联网的构建。例如，将其与车网互动（V2G）技术相结合，可以进一步提高系统调峰的灵活性，降低储能运营成本。同时，研究团队也提出，后续可以考虑引入更加丰富的随机因素，如可再生能源输出和负荷的时空相关不确定性，以及多能源融合（电-氢-热等综合能源系统）的情景，对模型进行扩展与验证。这些展望表明，该分层优化方法在未来能源智慧调度领域有着广阔的应用前景，为实现大规模可再生能源消纳和区域协同供能提供了有力支撑。

TECHNOLOGICAL BREAKTHROUGH
技术攻关

面向新型电力系统的虚拟电厂关键技术与示范

01 项目背景

在“双碳”战略与电力体制深改的交汇期

我国电力系统正从以装机规模为中心的“量的扩张”，转向以消纳能力与系统灵活性为核心的“质的跃迁”，虚拟电厂成为连接分布式资源与市场机制的中枢神经。过去一年，国内可再生能源新增装机高速跃升，但西部、北部等资源地仍存在显著弃风弃光，2025年上半年部分省份的新能源利用不充分已引发广泛关注，这直接倒逼调度侧、用户侧与市场侧的协同机制升级。政策层面，国家发改委推动省级现货市场转入正式运行，强调以市场出清约束调度与结算的实时闭环，为虚拟电厂参与多品类交易与辅助服务提供制度空间。同时，发改委、能源局、国家数据局联合发布的新型电力系统行动方案（2024–2027）指向灵活调节能力建设、需求侧响应规模化与数据要素入网，虚拟电厂被明确为重要组织形式 。国际上，欧盟FENIX等项目早已验证“聚合—协同—市场化”的工程可行性，VPP之于高比例新能源系统，已从理念跃升为支撑性基础设施。虚拟电厂正快速成为贯穿感知、优化、执行与结算全链路的系统工程，对我国能源安全、价格稳定、产业升级与民生保障具有基础性意义与牵引性价值。

02 创新内容

为了应对上述虚拟电厂在高比例新能源系统下的调度挑战，团队创新性的提出“DIASF市场动态信息感知自适应调度框架”，在VPP理论与部署框架层面均做出创新。

理论上

团队证明了在合理条件下，虚拟电厂两阶段与单阶段模型在不同随机场景下的模型逼近程度，并刻画了两者在弱化等价性要求的条件下最优值差的可计算表达，实现以偏差补偿的方式管理滚动误差。

方法上

团队构造基于变量冻结和可变窗口的滚动机制，在每一窗口开始前先行更新短期预测与场景集，冻结上窗决策的关键变量，随后以单阶段模型重优化并在实时端执行，形成“预测—优化—执行—校正”的闭环流程。

工程上

团队引入源荷交互理念作为消纳调峰的结构化接口，以中央空调等可转移负荷为代表，将其运行逻辑抽象为“开停状态—必开—必停—累计最小运行时长”三类统一约束，并将其与风光出力、储能功率与市场购电在功率平衡上显式耦合，实现以负荷吸纳波动、以窗口抑制滚动误差的柔性化调度。

在这一体系内，价格信号既是成本度量也是反馈控制量，负荷的时序可转移性与储能的时移能力形成互补，而滚动窗口的冻结与解冻操作则为算时与连续性提供制度化保障。

03 应用场景

在用户侧

虚拟电厂可面向园区、商办综合体、轨交枢纽与数据中心构建柔性负荷、储能、与分布式光伏的协同单元，通过日内现货、辅助服务与需求响应市场的组合交易获取收益，政策与市场的同步推进提供了现实抓手；现货转正与统一市场体系建设为“边出清边执行”的快循环提供制度基础，而车网互动的规模化试点将以V2G/V1G的形式把电动汽车单元纳入聚合体，使千家万户的小储能真正成为系统级调峰资源。

在电网侧

虚拟电厂可以作为无形电厂参与跨区调剂，配合特高压外送曲线进行本地削峰填谷，直接对冲弃风弃光与新能源波动带来的短时供需错配，契合国家对提升需求响应能力与系统灵活性的阶段性目标。

在国际对标层面

FENIX等工程实践已经展示聚合、协同、市场化的可复制性，为我国在高渗透率新能源下的工程落地提供参照系与对比标尺。研发团队提出的上述“价格可感+负荷可转+算时可控”的体系，既满足监管可审计、企业可落地、终端可参与的三元要求，也为不同地区在不同资源禀赋与不同市场成熟度下的差异化部署留下弹性空间。

04 重要成果

在多日连续运行的长序列仿真中，我们对所提框架开展了三类系统验证。

其一，在“完美预测”的控制场景中，单阶段与两阶段在可再生利用与市场购电两项指标上 92% 的时段完全一致，最大偏差分别严格受限于 0.12% 与 0.4%，验证了理论等价在工程条件下的稳健可达性，为从“两阶段”迁移至“单阶段”提供实践依据。

其二，在峰谷错配显著、滚动误差典型的极端工况下，纯滚动模型的可再生利用率约 83.16%，而引入源荷交互的增强模型与完整DIASF框架分别达到约 91.00% 与 97.25%，充分说明负荷可转移+窗口可冻结的双重机制，能将新能源出力的时间不确定性转化为可管理的优化变量。

其三，在 16 天、215 个算例的综合测试中，DIASF 在 4/6/12 小时窗口下均表现出显著的成本、消纳、算时多方面的综合优势，在保持成本估计与最优解偏差不超过 0.3% 的同时，将平均求解时间控制在百秒量级，部分模块对比基线降低综合运行成本 11.6%～31.1%，新能源利用率整体提升 5.75% 左右。

以上结果在山东风光—负荷—气象典型数据与基于上海电力市场的价格框架下均得到一致呈现，显示出良好的区域可迁移性与市场适配性。

参考文献

[1] Reuters. China’s renewable capacity soars but utilisation lags, data show. 2025-08-05. Available at:https://www.reuters.com/sustainability/climate-energy/chinas-renewable-capacity-soars-utilisation-lags-data-show-2025-08-05/

[2] 国家发展改革委. 关于加快推动全国统一电力市场体系建设的通知. 2023-11-01. 可获取: https://www.ndrc.gov.cn/xxgk/zcfb/tz/202311/t20231101_1361704.html

[3] Climate Cooperation. NDRC, NEA and NDA issue Action Plan on Power System Transformation (2024–2027). Available at: https://climatecooperation.cn/climate/ndrc-nea-and-nda-issue-action-plan-on-power-system-transformation-2024-2027/

[4] Reuters. China’s state planner outlines power system upgrade priorities. 2024-08-06. Available at: https://www.reuters.com/world/china/chinas-state-planner-outlines-power-system-upgrade-priorities-2024-08-06/

[5] European Commission CORDIS. FENIX Project Final Report. Available at: https://cordis.europa.eu/project/id/518272/reporting

[6] 国家发展改革委. 关于推进新能源汽车与电网融合互动的实施意见. 2024-01-04. 可获取: https://www.ndrc.gov.cn/xxgk/zcfb/tz/202401/t20240104_1363096.html

面向电力负荷极端波动场景的混沌振荡预测模型

01 项目背景

在能源、电力、金融等重要领域，预测技术承担着运行调度、风险防控与决策支持的重要使命。然而，对于复杂预测场景的行业数据往往具有高度一致的复杂特征并面对挑战：

- **强非线性：**新能源出力波动、市场情绪变化等因素，使得输入与输出关系难以用线性模型刻画
- **分布随时间快速漂移：**电力市场价格在不同季节的分布差异显著，金融市场在不同政策周期下的波动结构完全不同
- **小概率极端事件主导误差：**如局部电网突发故障、金融市场闪崩，单次事件就可能改变整体预测精度
- **短周期高频信号与噪声交织：**高频交易数据中的秒级波动、电力系统的尖峰负荷等都会掩盖长期趋势

在这些现实场景中，无论是传统统计方法（如GARCH）还是主流深度学习模型（如Transformer、LSTM），在遭遇突发性扰动时都容易出现预测偏差陡增、趋势判断失准的情况。同时，现有方法往往计算量大、依赖海量数据才能稳定运行，导致模型在新领域或数据有限的场景中难以迁移和复用。面向这些领域的高难度预测任务，本项目研发了可跨领域迁移、能够在极端波动条件下保持稳定精度的时序预测模型——混沌振荡变换器网络（COTN）。COTN可在不同类型的数据中快速适配，无需为单一领域重构架构，尤其适合负荷与价格断崖式变化、行情闪崩、流动性骤降等突发性强的场景。此外，COTN针对上述共性难题，通过结构简化与动态激活机制，提升了在小数据集和不同领域间的迁移能力，使得模型在能源、金融等多个高波动系统中均能快速部署并保持高精度预测。

02 创新内容

针对上述共性难题，COTN通过模拟神经系统对外界刺激的动态调节机制，实现了在剧烈波动场景下的稳健预测。核心包括：

混沌动态激活

引入Lee振荡器，将每次激活过程视为一个动态演化系统，捕捉高波动数据中的峰值混沌响应。

入门控融合

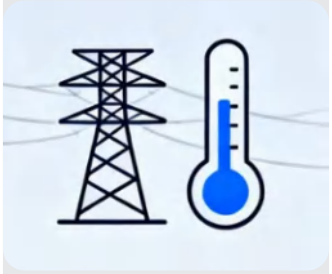
在稳定期偏向平滑激活（GELU），在波动期偏向混沌响应，实现模型对状态变化的自适应切换。

异常感知与隔离

结合独特设计注意力稀释（DAT）与异常结构隔离（ASM）模块，降低注意力机制的复杂度到O(log N)大幅度降低计算成本，将极端值的影响从主干趋势建模中剥离，避免误差扩散。

03 应用场景

电力负荷与价格预测

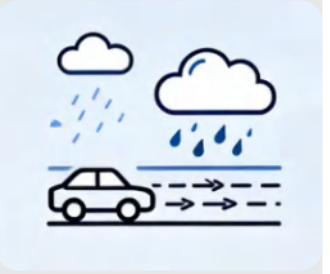


应对新能源冲击、气温骤变等极端工况

高频金融行情预测

处理交易暂停、价格闪崩、流动性骤降等异常

其他高波动时序预测



如交通流断面突变、气象极端事件

04 验证结果

在中国省级电力市场负荷数据与A股分钟级行情数据上的测试表明，COTN在极端波动期较Informer和标准Transformer平均降低**6%–9%的均方误差，最高降低13%，相较GARCH模型误差降低最高达40%**；在包含突发事件的区间，预测结果波动更小，且收敛速度提高约**42%**（两阶段预训练策略）。跨领域实验验证了该模型可直接迁移到其他具备类似数据特性的预测任务中，并在多类极端波动场景下保持稳定表现。

CUSTOMER CASES
客户案例

城市轨道交通网络节能优化

城市轨道交通 | 节能优化 | 运行图调整 | 调度优化求解器

01 项目背景

青岛地铁作为青岛市重要的公共交通骨干力量，近年来随着城市的快速发展以及人口的不断涌入，线路规模持续扩大，目前已有多条线路投入运营，日均客流量高达100万余人次。在城市发展对绿色环保要求日益提高的大背景下，青岛地铁积极响应国家节能减排政策，致力于在保障高效、安全运营的同时，降低能源消耗，提升运营的经济性与可持续性。然而，随着线路增多和客流量的波动变化，如何在复杂的运营环境下实现节能目标成为了一项亟待解决的挑战。

02 行业问题

在城市轨道交通运营中，运行图是列车运行的基础框架，它对列车的牵引系统能耗有着根本性的决定作用。传统节能策略往往侧重于车辆设备本身的技术改进或简单的运行模式调整，却严重忽视了运行控制这一关键环节对节能效果产生的显著影响。具体而言，地铁运营商面临着以下几方面问题：

客流时空分布不均

不同线路、不同时段的客流量差异巨大，传统的固定运行图难以精准匹配客流变化，导致列车在部分时段空驶率较高，能源浪费严重。

·青岛地铁·

城市轨道交通

运行控制缺乏灵活性

现有的运行控制模式相对固化，无法根据不同时段客流情况、线路状况等因素进行动态调整，使得列车运行过程中的能耗未能达到最优状态。

节能策略协同性不足

以往的节能措施多是各自为政，缺乏从整体运营层面进行的系统性规划与协同，难以充分发挥各项节能手段的综合效益。

基于以上问题，地铁运营商需要一种创新的解决方案，能够在不影响现有服务体验和不增加硬件投入的前提下，通过优化运行图和运行控制策略，实现显著节能。

03 解决方案

为帮助青岛地铁解决节能难题，引入城市轨道交通调度优化求解器。该求解器基于大数据、人工智能和运筹学算法的智能决策支持系统，能够对列车运营、乘务计划、车辆维护等多个环节进行全方位的优化。具体解决方案如下：

精准客流预测

通过收集和分析历史客流数据、交通信息、气象数据等多源数据，运用深度学习算法构建精准的客流预测模型。该模型能够提前预测不同线路、不同时段的客流量变化情况，为运行图的动态调整提供可靠依据。

运行图优化

据客流预测结果，调度优化求解器能够生成最优的列车运行图。在客流高峰时段，增加列车开行频率，缩短发车间隔，满足乘客出行需求；在客流低谷时段，合理减少列车数量，降低空驶率，从而有效节约能源。同时，求解器还能够根据线路状况、列车运行状态等实时信息，对运行图进行动态调整，确保列车运行的高效性和稳定性。

智能运行控制

调度优化求解器与列车自动控制系统（ATC）、列车自动驾驶系统（ATO）等紧密集成，实现对列车运行的智能化控制。通过调整列车的运行速度、启停时间等参数，在保证列车准点运行和乘客舒适度的前提下，最大限度地降低牵引能耗。例如，在列车进站停车时，求解器可以根据站台客流情况和后续运行计划，精确控制列车的制动和启动，减少不必要的能量损耗。

协同节能策略制定

从城市轨道交通运营的全局出发，综合考虑列车运行、供电系统、通风空调等多个子系统的运行特性，制定协同节能策略。例如，根据列车运行时刻表，合理安排供电系统的功率输出，避免不必要的电能浪费；根据车站客流情况，智能调节通风空调系统的运行模式，降低能源消耗。

模拟仿真与持续优化

利用调度优化求解器的模拟仿真功能，对不同的运行方案和节能策略进行预先评估和测试，提前发现潜在问题并进行优化调整。在实际运营过程中，持续收集运行数据，对求解器的算法和模型进行迭代优化，不断提升节能效果和运营效率。

04 落地成效

通过在青岛地铁4号线的实际应用，该节能优化解决方案取得了显著的成果：

节能效果突出

在不影响乘客服务体验、无需额外硬件投入且不改变现有运营方式的情况下，成功实现了**牵引能耗下降5%**的节能目标。这一成果不仅为青岛地铁带来了可观的经济效益，也为其在绿色低碳发展方面树立了良好的示范。

运营效率提升

精准的客流预测和高效的运行图优化，使得列车运行更加贴合客流需求，减少了列车空驶时间和乘客候车时间，提高了线路的整体运输能力和服务质量。

智能化水平提高

城市轨道交通调度优化求解器的应用，推动了青岛地铁运营管理的智能化升级。通过实时数据监测和智能决策支持，运营人员能够更加科学、高效地进行调度指挥，提升了应对突发情况和复杂运营场景的能力。

可持续发展贡献

该解决方案的成功实施，为青岛地铁在节能减排、环境保护方面做出了积极贡献，符合国家绿色交通发展战略，有助于提升城市的可持续发展水平。

AI Computing Infrastructure and Vertical Applications

深圳市大数据研究院
Shenzhen Research Institute of Big Data

AI算力基础设施与垂直领域应用

ACADEMIC ACCUMULATION
学术积累

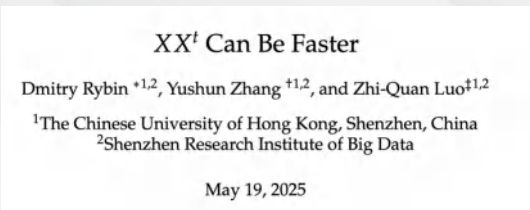
XX^T Can Be Faster

01 项目背景

矩阵乘法优化堪称计算机科学领域的"珠穆朗玛峰"。自1969年Strassen算法横空出世以来，这个充满组合爆炸可能性的数学迷宫就持续考验着人类智慧的边界。Google DeepMind为此专门投入四年心血，先后推出AlphaTensor、AlphaEvolve等机器学习系统来攻克这一难题。这就像短跑运动员将百米纪录从9.58秒推进到9.57秒——每个0.01秒的突破背后，都是对计算理论极限的重新定义。

XX^T（矩阵乘以自身的转置）这类特殊的矩阵乘法广泛存在于各类数据科学的实际应用中，实际应用包括：

- 5G与自动驾驶定制芯片设计
- 线性回归与数据分析
- 大语言模型训练算法（Muon、SOAP）


XX^t Can Be Faster

Dmitry Rybin *[1,2], Yushun Zhang *[1,2], and Zhi-Quan Luo‡[1,2]

[1]The Chinese University of Hong Kong, Shenzhen, China
[2]Shenzhen Research Institute of Big Data

May 19, 2025

XX^T这类操作每分钟在全球执行数万亿次，假如能减少该操作的计算量，对能耗开销可以带来相当可观的节省。令人惊讶的是，相比于普适的矩阵乘法AB，研究者对于XX^T这类的特殊的矩阵乘法的关注少之甚少。Google DeepMind 的AlphaTensor，AlphaEvolve 探索了带有特殊结构的AB矩阵乘法，但他们尚未汇报任何关于XX^T的结果。

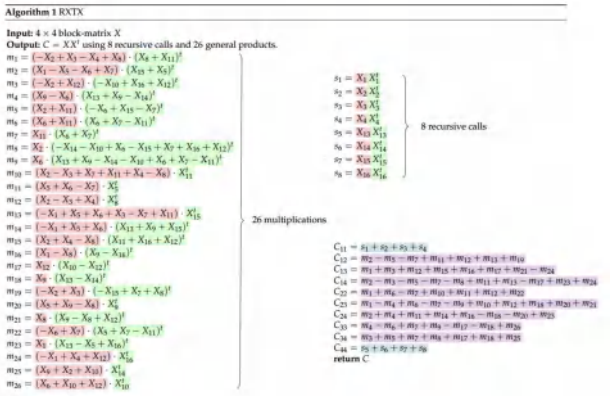
通过观察XX^T运算的特殊结构，该团队发现XX^T的计算确实存在加速空间。

02 核心内容

在AI技术的辅助下，研究团队发掘了新算法（RXTX），以让XX^T这一常见的底层操作减少5%的运算量，这可以进一步转换成节省5%的能耗以及时间（特别的，能耗开销主要由乘法运算数量决定）。值得一提的是，RXTX的5%加速不仅对超大规模矩阵成立，对小规模矩阵也成立，比如：RXTX对4x4矩阵X仅需34次乘法运算。此前最先进的Strassen算法需要38次乘法（减少10%运算量）。

Algorithm	Previous State-of-the-Art for XX^t	RXTX
Illustration in matrix form	$\begin{pmatrix} A^t & C^t \\ B^t & D^t \end{pmatrix}$ $\begin{pmatrix} A & B \\ C & D \end{pmatrix}$ $\begin{pmatrix} AA^t+BB^t & AC^t+BD^t \\ * & CC^t+DD^t \end{pmatrix}$	
Recursive expression	$S(n) = 4S(n/2) + 2M(n/2)$	$R(n) = 8R(n/4) + 26M(n/4)$
Asymptotic speedup (# multiplications for $n \to \infty$)	$S(n) \sim \frac{7}{3}M(n)$	$R(n) \sim \frac{26}{41}M(n)$ (5% ↓)
Non-asymptotic speedup (# multiplications for $n = 4$)	38	34 (10% ↓)

Table 1: Comparison between the proposed algorithm RXTX and previous State-of-the-Art (SotA) algorithm for computing XX^t for $X \in \mathbb{R}^{n\times m}$, $n, m \geq 4$. RXTX is based on recursive 4×4 block matrix multiplication. It uses 8 recursive calls and 26 general products. The previous SotA uses 16 recursive calls and 24 general products. $R(n), S(n), M(n)$ - are the number of multiplications performed by RXTX, previous SotA, and Strassen algorithm respectively for $X \in \mathbb{R}^{n\times n}$. The asymptotic constant of RXTX, $\frac{26}{41} \approx 0.6341$, is approximately 5% smaller than that of the previous state-of-the-art, $\frac{7}{3} \approx 0.6666$.



乘法运算量复杂度分析

研究团队对乘法运算量的复杂度进行了分析。分析结果表明，RXTX的渐进常数26/41≈0.63，较先前最优值2/3≈0.66降低5%。

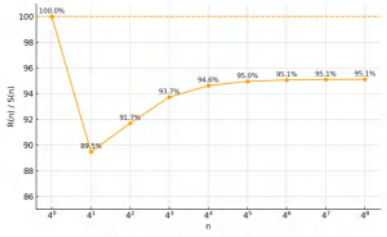
2.1 Number of multiplications

Theorem 1. The number of multiplications for RXTX:

$$R(n) = \frac{26}{41}M(n) + \frac{15}{41}n^{3/2} = \frac{26}{41}n^{\log_2 7} + \frac{15}{41}n^{3/2}.$$

The number of multiplications for recursive Strassen:

$$S(n) = \frac{2}{3}M(n) + \frac{1}{3}n^2 = \frac{2}{3}n^{\log_2 7} + \frac{1}{3}n^2.$$



Ratio of $R(n)$ to $S(n)$.

总运算量（乘法+加法）复杂度分析

研究团队进一步提供了总运算量（乘法+加法）的复杂度分析。分析结果表明，当n≥256时，RXTX的总加法与乘法次数也少于现有最优方案，且渐进意义下约有5%的稳定提升。

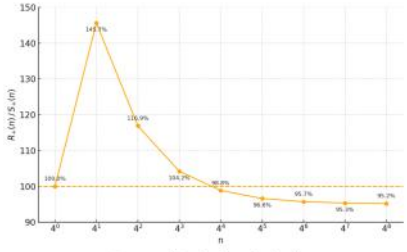
2.2 Total number of operations

Theorem 2. Total number of additions and multiplications for RXTX:

$$R_+(n) = \frac{156}{41}n^{\log_2 7} - \frac{615}{164}n^2 + \frac{155}{164}n^{3/2}.$$

Total number of additions and multiplications for recursive Strassen:

$$S_+(n) = 4n^{\log_2 7} - \frac{7}{4}n^2\log_2 n - 3n^2.$$



Ratio of $R_+(n)$ to $S_+(n)$.

03 核心技术

该方法属于基于神经网络的大邻域搜索方法框架：

- 利用强化学习策略生成候选双线性乘积
- 构建组合问题一（MILP-A）：将目标表达式构建为候选乘积的线性组合
- 构建组合问题二（MILP-B）：筛选能完整表达XX^T结果的最小乘积集

这是 DeepMind 的 AlphaTensor 方法的一种变体--通过使用组合求解器，行动空间被缩小了一百万倍。以下为作者提供的2*2矩阵的简单例子：

3.2 Example: matrix times transpose algorithm search for 2-by-2 matrix

Consider the example for 2×2 matrix X. We want to perform the computation of XX^t:

$$\begin{pmatrix} x_1 & x_2 \\ x_3 & x_4 \end{pmatrix} \cdot \begin{pmatrix} x_1 & x_3 \\ x_2 & x_4 \end{pmatrix} = \begin{pmatrix} x_1^2 + x_2^2 & x_1x_3 + x_2x_4 \\ x_1x_3 + x_2x_4 & x_3^2 + x_4^2 \end{pmatrix}$$

We identify 3 target expressions

$$T = \{x_1^2 + x_2^2, x_3^2 + x_4^2, x_1x_3 + x_2x_4\}.$$

We randomly sample thousands of products $p_1, ..., p_m$, each one given by

$$\left(\sum_{i=1}^{4} \alpha_i x_i\right) \cdot \left(\sum_{j=1}^{4} \beta_j x_j\right)$$

with $\alpha_i, \beta_j \in \{-1, 0, +1\}$ chosen by RL policy π_θ. MILP-A enumerates ways to write target expressions from T as linear combinations of sampled products $\sum \gamma_i p_i$. MILP-B selects minimal number of sampled products such that every target expression can be obtained as their linear combination. Key observation is that MILP-A and MILP-B are rapidly solvable with solvers like Gurobi [Gurobi Optimization, LLC, 2024].

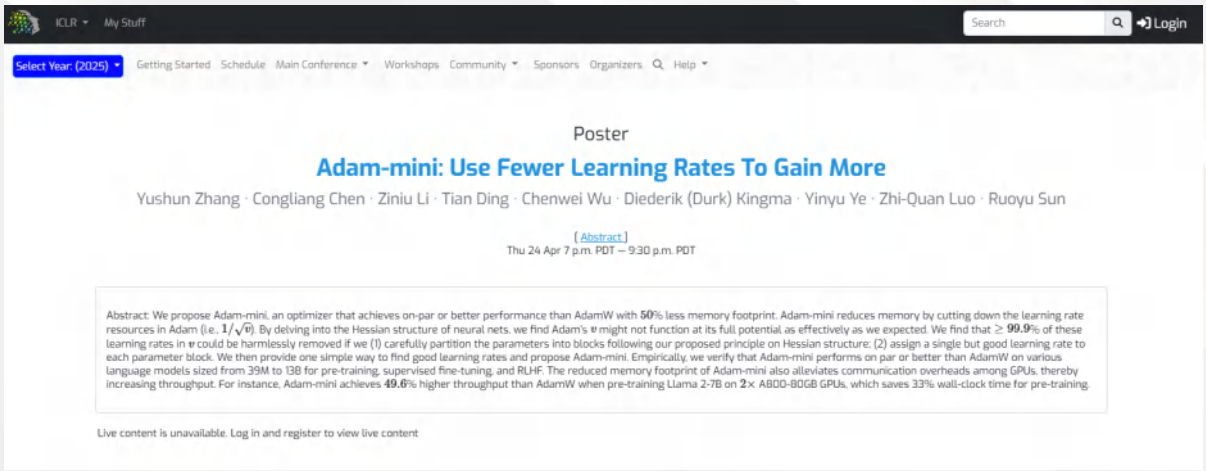
04 总结

本文针对XX^T这类特殊矩阵乘法提出了创新性加速方法，通过引入AI方法设计出新型算法“RXTX”，成功实现了总运算量5%的优化。这一突破不仅从理论上拓展了人类对计算复杂度边界的认识，也为相关领域的算法优化提供了新的研究范式。

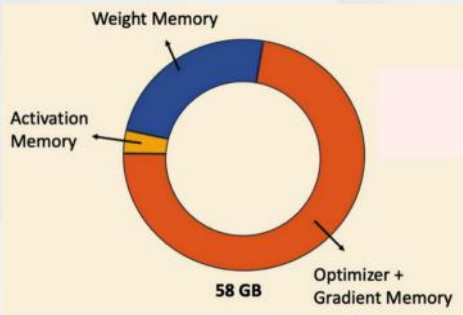
鉴于XX^T矩阵在多个学科领域的基础性作用，本研究成果有望为实际应用场景带来显著的能耗优化。然而，新算法的工程化应用仍面临硬件适配和内存管理等关键挑战，其产业化落地尚需学术界与工业界的持续协同攻关。要实现新算法的全方面落地，仍然面临诸多挑战，可谓任重道远。

Adam-mini: Use Fewer Learning Rates To Gain More

01 项目背景



当前，大语言模型的广泛应用使得训练大型模型的内存开销成为研究者和企业面临的一大挑战。这主要归因于大语言模型广泛采用的 Adam 优化器（Adaptive Moment Estimation），其尽管性能卓越，但内存需求却极为高昂。具体而言，Adam 优化器在训练过程中需要额外存储两个状态变量：一阶动量 m 和二阶动量 v（即每个参数需要存储两组变量），从而显著增加了内存负担。



举例而言，训练一个 Llama 2-7B（70亿参数量）模型的内存占用大致如下：

- 权重参数：28 GB（7B X 4 byte）
- Adam 状态变量：58 GB（约2 x 7B X 4 byte）
- 其他：激活值和梯度缓存等

如此高的内存需求对硬件造成了巨大的压力，即便是顶级的 A100-80GB 显卡也难以直接承载，必须依赖内存分片或CPU卸载技术来分担负载。然而，这些额外的处理方式不仅显著增加了训练过程的复杂性，还加重了通信开销。

02 核心内容

关键发现：Hessian 块异质性

在轻量级的Adam之前，我们需要先理解为什么现在的Transformer架构需要用Adam训练，再尝试在其基础上做改进。我们注意到前人发现了神经网络Hessian 有近似块对角结构（如图 2 所示）。

受这一现象的启发，我们考虑分析不同块对应的 Hessian阵的特征谱（即：该矩阵所有特征值的分布）。我们于 2024 年发现了 Transformer 架构中存在的 Hessian 块异质性（Block Heterogeneity）现象，并发现这一性质和Adam密切相关。相关研究成果已发表于 NeurIPS 2024 。

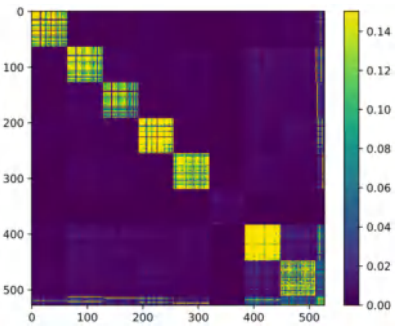


图 2：Hessian 近似块对角结构

Definition [Hessian Block Heterogeneity]

Hessian 块异质性指的是，神经网络的 Hessian 矩阵在不同参数块之间的特征谱具有显著差异的现象。

如图 3(左) 所示，以 CNN 架构的 VGG16 为例，不同块（卷积层）的 Hessian 谱较为相似。然而，图 3(右) 表明，在 Transformer 架构的 BERT 中，不同块的 Hessian 谱差异很大 (比如特征值范围差异可超200倍)。

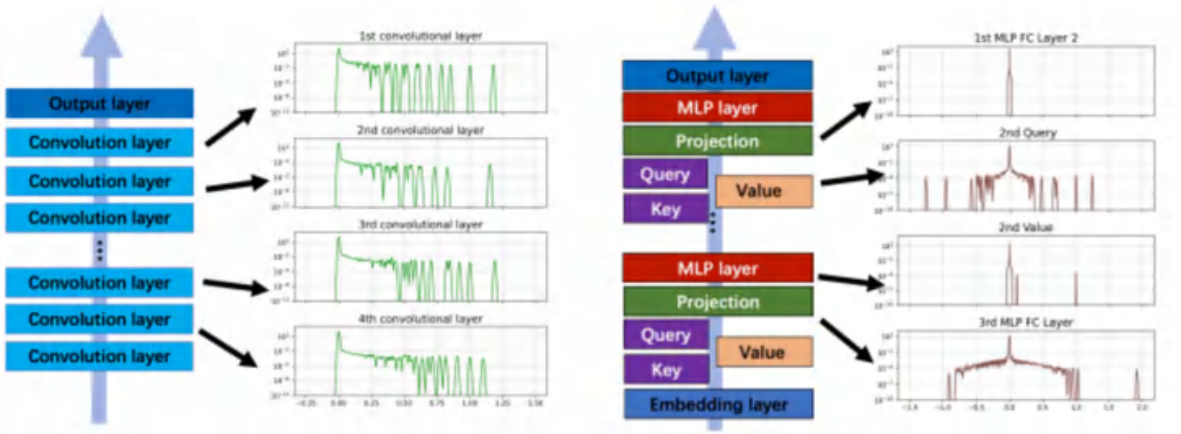


图 3：VGG16（CNN）和 BERT（Transformer）在初始化时的分块 Hessian 谱

众所周知，在 CNN 模型中，SGD 与 Adam 性能相当；但是在 Transformer 模型中，SGD 的性能则显著逊色于 Adam。我们推测，这种性能差异在一定程度上源于 Hessian 块异质性。在大量真实任务和合成的异质性任务上进行的实验表明，SGD 的性能会随着块异质性程度的增加而下降，从而进一步支持了这一推测。

具体来讲，我们将每个参数块的 Hessian 谱归一化为概率分布，并采用 JS distance 来度量分布之间的差异，从而定量描述块异质性的程度。如表 1 和图 3 所示，我们对若干模型在初始化时的 JS distance 进行了计算，得出了以下发现：

图 3 的趋势表明：

- JS distance 越大，块异质性程度越高，SGD 优化器的性能相较于 Adam 越差；
- Transformer 架构主导的模型，其块异质性程度是 CNN 的数百倍。

表 1. 初始化时，成对参数块 Hessian 谱之间的 JS 距离（JS）

Model	ResNet18	VGG16	GPT2 (pretrained)	MLP-mixer	BERT	GPT2	ViT-base
JS^0	0.10	0.09	18.84	34.90	53.38	83.23	286.41

我们进一步将块异质性和优化器的表现构建了联系。块异质性意味着不同的参数块有不同的优化需求，使用单一学习率的SGD难以适应多样优化需求，因此我们认为块异质性正是Adam在Transformer架构中优于 SGD 的关键原因。

这一发现在本文的主题Adam-mini优化器的设计过程中起到了关键作用。

03 成果

- **轻量优化器: Adam-mini**

上文我们的发现表明Adam能满足Transformer 架构中不同参数块的Hessian矩阵的不同的优化需求，这说明了Adam是“充分的”，但还不足以说明它是“必要”的。我们在此基础上开展了进一步的消融实验，我们发现：在同一个参数块内，Adam目前的“独立分配学习率”可能并没有发挥出应有的功效。具体原因涉及到计算数学中一个悠久的话题，即关于diagonal-preconditioner的局限性分析，在此省略。总结来说，我们认为Adam目前的设计可能存在冗余。

基于上述动机和观察，我们对Adam的冗余部分进行了“瘦身”，并提出了Adam-mini优化器，相关研究成果已发表于ICLR 2025会议论文[2]，其核心设计思想如下：

基于Hessian结构（如图4）分块：对于Transformer架构，query和key按head分块；value、attn.proj和MLP按输出神经元分块；embed_layer和output_layer按token分块。

简化学习率：取子块b梯度的均方值为唯一学习率，且在块内共享。

$$v_b = (1-\beta_2) \times \text{mean}(g_b \odot g_b) + \beta_2 \times v_b, \quad b = 1, \ldots, B.$$

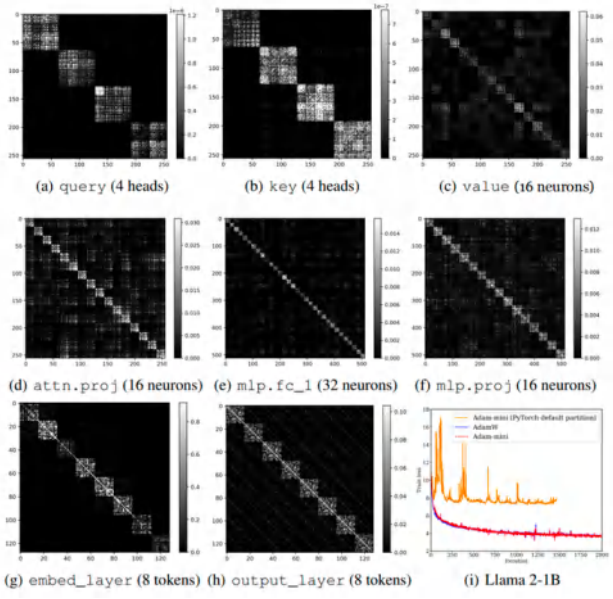


图 4. 小型 Transformer 中不同参数块的 Hessian

简单的理论分析表明，Adam-mini可以无损地移除超过99.9%的二阶动量v，从而为Adam优化器节省接近50%的内存。

● 实验验证

我们从多方面对 Adam-mini 进行了实验评估，实验语言模型的预训练和微调任务，以下是主要结论：

大幅降低内存开销

在Llama系列（图5）模型的预训练中，Adam-mini相比AdamW减少了50%的优化器状态内存开销。与此同时，Adam-mini在多项指标上都与AdamW的表现持平，在某些任务中甚至略优。

提升吞吐量

如图7所示，由于减少了通信开销和内存占用，Adam-mini在预训练Llama 2-7B模型时吞吐量提升了49.6%，训练时间缩短了33%。

无需额外调参

Adam-mini用AdamW原版超参数即可达到一样性能，可实现算法直接迁移。

可扩展到更大的模型

我们进行了大量Scaling law的实验（图6）并发现：对于不同的尺度的模型，Adam-mini的训练曲线都与AdamW高度重合。这为Adam-mini在更大规模实验中的可扩展性和有效性提供了证据。

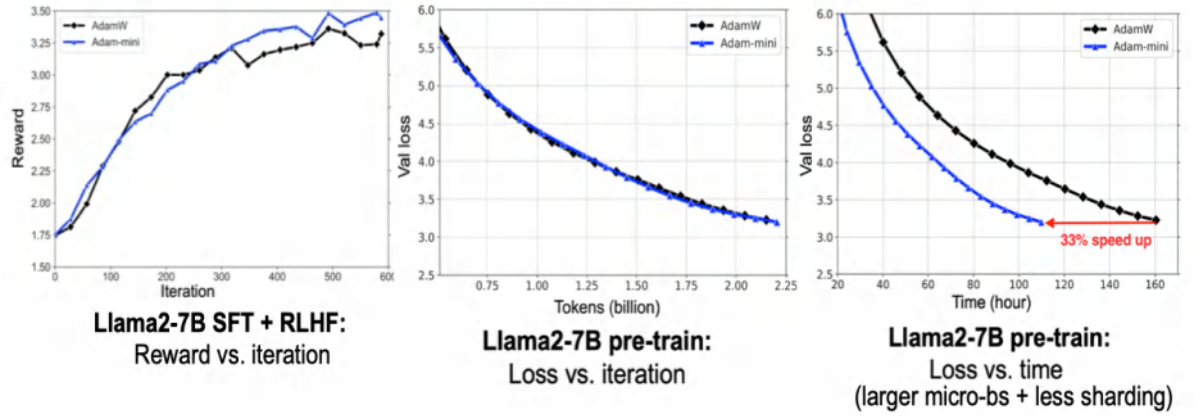


图 5. Llama2-7B 中 Adam-mini 与 AdamW 的训练损失曲线

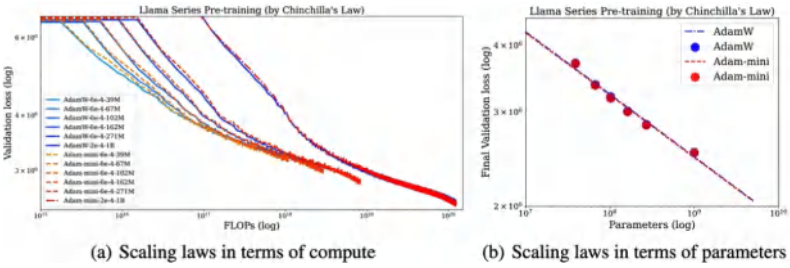


图 6. Llama 系列中 Adam-mini 与 AdamW 的scaling law曲线基本重合

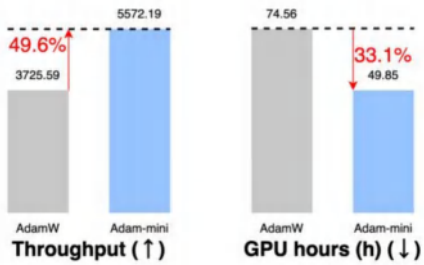


图 7. Adam-mini 与 AdamW 的吞吐量比较

04 未来展望

Adam-mini的延伸: GaLore-mini

近年来，低秩方法已成为内存优化的一个有效方案，其中的代表作是GaLore。我们注意到GaLore和Adam-mini的技术思路在本质上是正交的，因此融合两者的核心思想，提出了全新的优化器GaLore-mini（被NeurIPS 2024 Workshop接收）。

GaLore-mini通过利用梯度的低秩特性，将梯度投影到低维空间来更新优化器状态，同时结合Adam-mini的策略，显著减少了优化器的存储需求，从而极大程度降低了内存开销。相较于AdamW，GaLore-mini在不损失性能的情况下实现了81%的内存节省。

面向深度学习非凸优化的Oscillator Torch

01 项目背景

- 传统主流激活函数如ReLU、GELU等，核心作用是提供非线性变换，支持神经网络的表达能力。然而，这类静态且简单的非线性映射在面对高维参数空间中“峡谷效应”、鞍点激增及宽谷路径不确定性时，难以有效缓解训练中的震荡与收敛困境。

- 值得注意的是，业界多项优秀模型（如线性注意力、线性RNN及Mamba等）已证明，非线性表达能力并非仅能靠激活函数实现。但现有解决方案或在精度与效率间妥协，或依赖特定场景假设，难以实现普适性突破（Swish，Gelu）。这并非组件本身的缺陷，而是高维复杂场景对模型核心组件提出了超越“非线性变换”的新需求。

- 基于此，本项目构建了一套基于震荡范式（Oscillation Paradigm）的动态激活机制库，专注于提升高维非凸优化中神经网络的训练效率和性能表现。本项目提出的震荡范式，通过动态振荡机制替代传统静态激活函数，理论上不仅具备激活函数的通用近似能力，还显著提升了模型逃逸鞍点的概率与速度，成功突破了激活函数的功能边界，为神经网络优化带来了全新的思路和解决方案。

02 创新内容

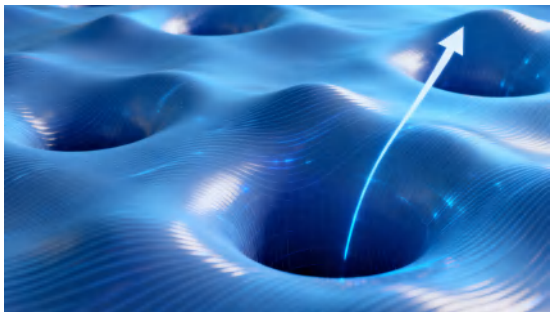
- 系统梳理并总结了2002年以来物理学及深度学习领域关于动力学方程和神经震荡的研究，推导，实验，提出震荡范式的理论框架及多种动态振荡机制。

- 采用变分自编码器（VAE）进行预拟合及工程优化，令模型复杂度降至O(m*n)，与ReLU等传统激活函数的计算复杂度相当。

- 针对高维非凸优化问题，震荡范式通过引入动态振荡过程，使模型能自适应调整非线性响应频率，有效捕捉低至0.1%的细微特征波动，显著缓解训练中的震荡与鞍点困扰。

基于PyTorch框架，开发并发布震荡范式深度学习库——Oscillator Torch，支持低侵入性调用，仅需一行代码即可完成震荡激活机制的集成，极大降低应用门槛，且具备良好的跨领域迁移能力。

03 应用场景



梯度陷阱逃逸

深度学习网络训练

04 重要成果

震荡范式PyTorch库凭借简便集成与多场景兼容性，不仅在理论上突破了传统激活函数的限制，有效提升高维非凸优化中的训练稳定性和收敛效率，也为金融、能源、计算机视觉等实际应用领域带来了切实可行的性能提升和优化方案。具体而言，

通过严格数学证明，验证震荡范式具备通用近似能力，并能显著提高逃逸梯度陷阱的速率和概率。

将离散震荡范式表达式推导为连续随机微分方程，显著增加可解释性，并进一步验证理论正确。

大量实证分析显示，震荡范式库在多领域、多任务中均展现出优异性能：

- 在面对峡谷问题及鞍点问题时，逃逸速度显著高于Gelu，Swish等，收敛速度提高 2.4倍。
- 在能源与金融时间序列预测任务中，均方误差（MSE）较主流模型降低超过13%；模型拟合判定系数（R^2）增幅92%；Transformer模型的MSE降幅最高达40.2%。
- 在计算机视觉领域，ImageNet与Cifar-100数据集中，针对部分基线模型，产生最高2%-5%的收敛精度提升，计算效率与Relu相当。

TECHNOLOGICAL BREAKTHROUGH
技术攻关

硬件亲和的算法与算子自动发现技术

01 项目背景

随着大模型训练与部署需求的爆发式增长，GPU/NPU计算集群成为人工智能发展的核心算力支撑。作为连接上层AI模型与底层硬件的关键“桥梁”，编译技术、底层算法与算子库的水平直接决定了算力的释放效率。国际上以英伟达CUDA生态为代表的体系经过多年积累，形成了成熟完善的底层算法与算子体系，成为其在AI计算芯片领域保持领先的关键壁垒。

相比之下，我国在算法与算子生态的完整性与成熟度方面仍存在显著差距。当前研发模式普遍依赖专家经验进行算子设计与调优，存在周期长、成本高、难以规模化复制等问题，已成为制约国产芯片算力性能发挥的主要瓶颈。如何摆脱“经验驱动”的研发模式，实现算子与算法的自动化生成与性能最优化，成为构建自主可控算力生态的核心挑战。

02 创新内容

为了应对以上挑战，深圳市大数据研究院围绕“硬件亲和算法与算子自动发现”主题，创新性融合数学优化与机器学习方法，系统性重构传统算子研发范式，建立了可度量、可搜索、可优化的自动化算法与算子设计体系。

该体系通过刻画目标硬件的结构特征（如指令级并行度、内存带宽、层级存储结构、数值精度等），将时延、吞吐、能耗与精度等多目标综合建模为优化问题，并引入强化学习与混合整数线性规划（MILP）求解框架，实现算法与算子的自动发现与最优解生成。

03 应用场景

服务华为昇腾为代表的国产算力，以及国产算力的需求方

该场景主要面向国产芯片厂商、服务器厂商、算力中心和云厂商等提供算子优化服务。随着国产算力市场的逐步崛起，特别是以昇腾为代表的AI算力平台需求日益增长，算子优化成为提升硬件性能的关键环节。通过提供基于AI+数学的算子自动生成技术，研究院及其孵化企业“智子芯元”帮助这些需求方提升算子适配性，优化算力的计算效率。对于国产芯片厂商，智能化的算子优化工具能够显著提高芯片的市场竞争力，缩短算子开发周期，同时帮助其在AI应用和高性能计算等领域中占据领先地位。

服务AI应用场景方，覆盖端侧设备、手机等智能终端的计算加速需求

在AI应用场景中，涵盖了具身智能、自动驾驶、医药发现、科学计算等领域的计算加速需求。这些领域对计算性能有着极高的要求，而智能终端如端侧设备和手机等也需要强大的计算支持。研究院及其孵化企业“智子芯元”通过为垂直领域提供特化的高性能算子解决方案，帮助客户解决计算瓶颈，实现AI算法的高效执行。例如，在自动驾驶领域，自动化工具能够通过优化场景特化算子，提升实时计算性能，保障自动驾驶系统的高效运行与安全性。

服务全球多类型、多架构芯片，推动算子自动生成与算法自动发现成为计算芯片工程开发主流范式

随着全球AI技术的快速发展，越来越多不同类型和架构的芯片需求涌现，算子自动生成与算法自动发现技术可以为这些芯片厂商提供定制化的算子优化方案。通过与硬件厂商的合作，推动算子自动生成与算法自动发现范式成为计算芯片的主流开发方式。这不仅能大幅提升芯片的计算效率，还能使其在不同场景中表现出色，为未来的计算芯片设计开辟了新的方向。算子与算法自动发现技术也为新一代芯片设计提供了有力的支持，特别是在合作中推动硬件与算法的深度融合，加速全球范围内AI算力平台的技术进步。

04 重要成果与影响

目前研究院团队已成功研发出“矩阵乘法自动搜索技术”和“数学函数算子自动生成技术”两项核心成果。

矩阵乘法自动搜索技术：针对XX^T类矩阵乘法问题，创新构建了强化学习引导、混合整数线性规划（MILP）求解的两级自动发现流程，将XX^T运算复杂度降低约5%，实现了五十年来该方向的首次性能突破，并在Intel CPU上相较于国际先进BLAS库取得约9%的速度提升。团队成功将矩阵乘法自动搜索流程拓展至Transformer模型核心算子，将自注意力计算复杂度降低10%。

数学函数算子自动生成技术：通过结合大模型Agent与进化搜索，实现数学函数算子的智能生成，单个算子研发周期由半个月以上缩短至2—3天，且生成的新算子性能匹配或超越人类设计算子。针对exp、sigmoid等基础非线性函数，新算子可在精度不变的前提下，将底层操作数减少50%，大幅降低计算时延；针对lgamma、贝塞尔等复杂函数，智能生成可帮助昇腾平台上的算子精度提升2–3个数量级，性能接近CUDA水平。部分自动生成的数学函数算子已成功落入下一代昇腾芯片算子库。

两项技术显著提升了算子设计效率与硬件算力利用率，为构建自主可控、国际先进的国产算力生态提供了重要支撑，并创造了显著的商业化价值。

大模型异构推理的智能调度系统

01 项目背景

随着开源基础模型的广泛普及，人工智能应用进入规模化落地攻坚阶段，同时由于核心算力自主可控的需求，国产算力低成本部署是推动其快速渗透、实现产业赋能的核心前提与关键支撑。开源生态大幅降低了模型开发门槛，催生了多元化应用需求，但算力投入过高、部署周期冗长等问题，严重制约了技术成果向实际生产力的转化。在此背景下，大模型推理过程中的算力适配及效率优化，已成为突破落地瓶颈、支撑行业快速发展的核心命题。当前国产算力集群普遍呈现多元异构架构特征，集成不同型号神经网络处理器（NPU）及存量中央处理器（CPU）资源，在适配开源模型轻量化部署需求的同时，面临着异构硬件协同效率低下、闲置CPU资源利用率不足等关键问题。

传统调度方案存在显著技术短板，难以适配上述复杂应用场景，具体难点体现在三方面：

- 其一，异构资源协同能力薄弱，缺乏对不同型号NPU硬件特性、算力阈值的精准感知机制，无法实现差异化资源分配，且存量闲置CPU资源未得到有效激活，形成算力浪费与资源“孤岛”现象；
- 其二，服务质量管控能力不足，传统调度策略缺乏明确的服务级目标（SLO）导向，难以依据业务优先级动态调配算力资源，导致各类请求的延迟、吞吐量等指标需求无法得到充分满足，SLO达成率偏低；
- 其三，推理过程优化程度不足，缺乏对请求输出长度的预判机制，且在键值缓存（KVCache）管理、任务卸载等关键环节缺乏针对性技术方案，进一步加剧了资源占用与调度冗余，制约了开源模型快速部署的效率。

02 创新内容

针对上述技术难点，本项目设计了一种面向大模型推理的智能调度架构，通过任务异构（多种业务场景&不同SLO要求）、模型异构（不同大小的模型，包括稠密/稀疏MoE）和硬件异构（CPU、NPU、核显GPU、独显GPU）等多种异构策略与软硬件协同优化，实现异构算力高效利用与服务质量精准管控。结合面向推理调度的输出长度预测模型（包括数值、区间、排序结果的预测）和多层级卸载调度策略（包括请求级卸载、decode卸载、attention卸载、专家卸载、KVCache卸载）等优化技术，实现了不同型号NPU与闲置CPU资源的深度挖潜，在显著提升整体算力利用率的同时，优化了各类请求的SLO达成效果，为开源基础模型低成本、快速部署的国产人工智能应用提供了核心技术支撑。

03 应用场景

云端侧

以云端中心节点集成化硬件集群为核心载体，构建多品牌国产NPU与CPU异构融合的集中式算力池，依托SLO感知调度、异构感知跨节点调度、键值缓存（KVCache）集群化智能卸载等技术策略，实现开源大模型的云端集群部署。其中SLO感知调度技术可精准识别金融风控、智慧教育、医疗辅助诊断等不同业务的时延、吞吐量SLO需求，动态分配异构算力资源，解决云端多业务混合部署下高优先级任务SLO保障不足的痛点；同时支持多终端远程接入与弹性算力扩容，在保障推理服务高可用性与低延迟的同时，显著提升云端算力资源的全局利用率，达成部署成本、服务规模与资源效益的三重优化。

边缘侧

聚焦边缘端低时延、低功耗需求，适配边缘节点异构硬件（分散式低功耗NPU/CPU），结合多层级卸载调度策略（包括请求级卸载、decode卸载、attention卸载、专家卸载、KVCache卸载），输出长度预测优化资源占用与推理时延，落地于智慧体育、企业应用等场景，实现开源模型边缘侧高效部署。

端侧

面向手机、NAS网络存储设备、智能穿戴设备等低算力终端设备，聚焦极致轻量化推理与本地算力按需调度需求，适配终端内置的异构计算单元（如手机SoC集成NPU、NAS嵌入式微算力模块），通过算子深度优化、并行机制优化以及模型量化压缩等技术路径，突破端侧算力不足的瓶颈。支撑如体育角、手机AI应用等场景，实现开源大模型的端侧原生部署，兼顾低算力消耗与推理响应速度，保障用户隐私数据的本地安全处理。



04 重要成果

在大模型边缘轻量化部署场景下，通过异构算力协同调度实现核心性能指标的显著突破：在保障模型推理精度的前提下，大模型推理吞吐率提升15%，高效的支撑边缘端高并发推理请求；核心服务等级目标（SLO）中关于响应时延、服务可用性的关键指标满足率达到75%，大幅提升服务的稳定性与可靠性；通过算力资源的精细化调度与冗余算力消减，CPU利用率提升25%，有效缓解边缘设备算力资源紧张的痛点；针对hareGPT数据集开展专项优化，在长文本生成、多轮对话续写等复杂场景的输出长度预测任务中，准确率稳定达到90%以上，为生成内容的长度可控性与任务调度提供了坚实保障。

AceGPT：Localizing Large Language Models in Arabic

01 项目背景

该项目的初步目标是开发一个理解和尊重阿拉伯文化和价值观的大语言模型，从而更好地满足阿拉伯语社区多样化和特定的文化和价值需求。最终目标是创建一个多语言大语言模型AceGPT，支持阿拉伯语、中文和英语。

图1.深圳市领导携代表团访问沙特，见证AceGPT 发布，并为KAUST-SRIBD联合实验室揭牌

FINANCIAL TIMES

COMPANIES TECH MARKETS CLIMATE OPINION LEX WORK & CAREERS LIFE & ARTS HTSI

Artificial intelligence + Add to myFT

Professor Jinchao Xu, an American-Chinese mathematician at Saudi Arabia's King Abdullah University of Science and Technology (Kaust), has launched AceGPT, an Arabic-focused large language model, in collaboration with the Chinese University of Hong Kong, Shenzhen (CUHK-SZ), and the Shenzhen Research Institute of Big Data.

The move is part of Saudi Arabia's efforts to lead the regional development of AI technology, building large supercomputers and rolling out LLMs, the technology that underpins generative AI systems such as chatbots. Along with the United Arab Emirates, the Gulf powerhouse is seeking to compete with AI companies and create bespoke models for Arabic speakers.

图2.AceGPT引领中东LLM研发

02 创新内容

本地化预训练：采用480B阿拉伯语数据进行继续预训练，确保模型对阿拉伯语的深刻理解和精准表达。

本地化指令微调：通过包含500万条对话数据的本地化指令集进行监督式微调（SFT），模型能够生成符合阿拉伯语使 用者习惯和期望的响应。

本地化RLAIF：训练了一个基于11,000条本地化偏好数据的奖励模型，以优化模型在阿拉伯语环境中的表现和适应性。

评估：结合了自动评估和人类评估。自动评估采用GPT-4作为裁判，确保评估的客观性和准确性。人类评估则由11位阿拉伯语母语者参与，以确保模型在实际应用中的自然性和流畅性。

03 应用场景

依托AceGPT的强大能力，致力于在教育、旅游、医疗等关键垂直领域开发定制化应用，推动人工智能技术的实际应用落地。

04 重要成果

开源

- AceGPT包含7B、8B、13B、32B、70B等不同规模的模型；它是最好的开源阿拉伯语大模型，同时支持中文和英文
- 在多个基准测试中（下表），显著超过阿联酋开发的Jais-30B超过了GPT-3.5（1750亿参数量），接近GPT4

	Arabic Datasets								English Datasets			Chinese Datasets		
	Avg.	Arabic MMLU (Ours)	Arabic MMLU (MBZUAI)	ARC	BoolQ	EXAMs	ACVA_clean_f1	ACVA_all_f1	Avg.	MMLU	RACE	Avg.	CMMLU	CEval
AceGPT-70B	**73.99**	**64.26**	**72.50**	**85.53**	**82.66**	**56.99**	78.61	77.38	**83.69**	**78.98**	88.39	67.56	68.03	67.09
AceGPT-32B	70.63	57.12	68.70	78.07	77.22	52.89	**81.36**	**79.03**	82.86	74.43	**91.28**	**77.11**	**76.10**	**78.11**
AceGPT-13B	63.42	47.33	61.70	63.99	69.33	48.37	76.90	76.37	69.68	60.43	78.93	43.32	44.00	42.64
AceGPT-8B	66.69	54.45	62.21	72.44	71.65	52.98	76.54	76.55	75.68	67.33	84.02	52.25	51.68	52.82
Jais-30B	57.84	35.68	62.36	51.02	76.30	32.24	73.63	73.66	57.03	59.65	54.40	31.51	25.91	37.10
GPT-3.5 (175B)	62.44	46.07	57.72	60.24	76.12	45.63	74.45	76.88	74.70	69.10	80.30	53.20	53.90	52.50
GPT-4 (1.7T?)	75.78	65.04	72.50	85.67	85.99	57.76	84.06	79.43	87.00	83.00	91.00	70.45	71.00	69.90

AceGPT显著超越Jais（MBZUAI，阿联酋）

AceGPT demo:
https://chat.acegpt.org/

生成式人工智能：华佗GPT-中文医疗大语言模型

01 项目背景

- 通用大模型在医疗领域知识不足、存在幻觉、标注成本高和评测困难
- 医疗资源不均衡、医疗隐私影响国计民生与国家安全
- 华佗GPT是基于“医疗咨询”前瞻性布局且自主研发的大模型，涉及约TB级别的医疗文本，堪称国内最大、来源最丰富的中文医疗数据
- 作为国内首个可在学术界和产业界应用的中文医学基础大模型，其参数规模接近到百亿级别
- 在医疗咨询（含病人培训、健康建议、初步治疗建议、分诊、心理诊断与治疗等）及情感陪伴等领域，提供线上交互性对话服务

02 创新内容

大模型数理基础：
Re Max算法、Outcome-supervision Value Model（OVM）以及对Adam算法的新理论

大模型的应用：
华佗、Ace、凤凰；
跨语言的医疗视觉-语言预训练

大模型评估平台和标准化：
医疗评测平台CMB和多模态模型评估平台MLLM bench

03 应用场景

- **智能问答服务：**满足居民健康、康复等方面的咨询服务，提供权威的医疗知识
- **智能分诊导诊：**帮助医院实现患者精准就医
- **智能医助：**帮助医生实现预问诊，形成疾病图谱
- **智能科普宣教：**个性化推送医学文章与知识
- **家庭医生：**基层医生AI健康教育咨询助手

04 重要成果

2022.9	着手部署医疗GPT的工作，购买10台A100服务器，构造了Huatuo数据集，包括爬取在线医疗和对话数据、知识图谱、维基百科、医疗长文本等
2022.10	华佗GPT的雏形出来，当时的模型还不够大，只有1亿的参数
2022.11	OpenAI发布ChatGPT
2023.2	大数据研究院在中华医院信息网络大会（CHINC）发布华佗GPT，这是首个中文医疗大模型
2023.5	经过临床医生测评结果显示，华佗GPT中文医疗场景超过了ChatGPT
2023.10	华佗GPT是首个全面通过药剂师、执业医师等多个医疗资格考试的大模型
2024.1	启动多模态GPT和多语言GPT，目前已有内部试用版本
2024.7	深圳市龙岗区区域平台12家公立医院部署华佗GPT实现区域预分诊和预问诊项目启动，为全国首个支持利用大模型开展多家多类型医院的应用落地
2024.11	龙岗区全部12家医院（包括三甲医院、二级医院、中医院、妇幼保健院、骨科医院、耳鼻喉医院）区域平台和微信公众号部署上线，日前使用人次近10万，覆盖龙岗区500万人口，并安排预问诊项目上线
2025.5	龙岗区AI家庭医生系统，该系统深度依托华佗GPT前沿技术，以小程序的形式为居民提供全方位的“数智健康管家”服务，在龙岗区所有区属公立医院社康全面上线，共覆盖12家医院183家社康中心
2025.7	通过大模型技术结合社康和家医知识库，精准回答患者咨询问题，降低医护工作量，并利用知识图谱和RAG技术确保回答准确性超过98%，支持单一入口识别患者意图，自动纠正输入错误，并提供语音和文本交互

05 成果展示

华佗GPT导诊　　华佗GPT

HuatuoEvidence 循证医学多智能体系统

01 项目背景

当前医疗大模型普遍存在“幻觉”与“不可控”问题，导致生成内容缺乏准确性、可解释性差，难以满足临床循证需求。尤其在门诊场景中，医生需花费大量时间梳理患者信息，包括人工问诊、翻查、口述等，易遗漏历史诊疗细节，影响诊疗效率与质量。

02 核心内容

本项目构建了一套精准可溯源的多智能体循证医学知识服务系统，核心技术创新包括：

全维指南分层检索（Precision RAG）

构建结构化多模态检索策略，综合匹配标题、年份、段落、图表等元数据，提升检索精准度。

TraceRL 强化学习对齐机制

引入量表化建模，将临床指南的定性评价转化为可微的定量函数，实现生成内容与循证标准的对齐。

阶段式门控推理机制（Staged Gating）

采用“规划-执行-检查”架构，设立关键节点进行验证，防止错误累积。

全链路证据溯源

每个推理步骤必须标注证据来源（如指南、图表、年份），实现从“黑盒”到“白盒”的转变。

03 应用场景

门诊接诊辅助	临床决策支持	医疗知识管理
系统自动整合患者历史诊疗数据、导诊信息，生成带溯源的信息小结。	为医生提供精准、可溯源的诊疗建议，减少“幻觉”风险，提升诊断质量。	支持医院构建结构化、可检索的循证知识库，辅助科研与教学。

04 重要成果

- 已在龙岗区人民医院实现产品落地，基于患者近5次就诊数据，有效提取关键信息，显著提升门诊效率。
- 帮助医生在1分钟内掌握核心信息，问询时长压缩80%。
- 作为“华佗智医”与“主动健康”项目的核心技术支撑，具备申报知识产权潜力。
- 突破医疗大模型“幻觉”与“不可控”难题，构建“精准循证、逻辑透明”的行业技术壁垒。

行政执法文书生成大模型

01 项目背景

在推进全面依法治国、建设法治政府的背景下，行政执法规范化对文书制作的准确性、高效性提出更高要求。当前基层执法人员常面临文书种类多、填写复杂等问题，易出现格式不规范、内容遗漏等情况，影响执法效率与公信力。随着人工智能技术发展，利用AI赋能行政执法成为可能。项目基于行政执法数据要素，构建行政执法文书生成大模型智能体，实现执法文书自动、规范、高效生成，助力提升行政执法智能化水平，推动法治政府与科技融合创新。

02 创新内容

围绕行政执法领域的现实需求，探索基于大语言模型的智能解决方案，辅助一线执法人员快速、精准、规范办案。目前基于数百万份行政执法文书，完成了第一阶段的产品化开发和大模型训练测试。

03 重要成果

以行政处罚为例，执法人员在执法终端上，选择被执法对象证照、现场照片、执法人员证件等图片，现场制作文书。

AI智能识别和提取图片中当事人信息、执法现场、证据、执法人员行为等数据要素。

行政执法文书生成大模型基于所提取的数据要素，自动生成行政处罚现场勘验笔录、案件处理审批表、立案登记审批表、通知书、决定书等十余份执法文书。

政务服务垂直大模型开发应用与测评标准机制

01 项目背景

在数字中国战略纵深推进、国家治理体系和治理能力现代化加速演进的背景下，政务服务智能化已成为城市治理向精细化、高效化跃升的核心引擎。大语言模型作为人工智能领域的革命性技术，为破解政务服务长期存在的政策传递堵点、诉求响应痛点、服务供给断点提供了突破性可能。然而，当前政务场景的大模型应用面临瓶颈：一方面，通用型大模型对政务领域政策专业性、场景复杂性的适配度不足，难以精准匹配不同城市、不同层级的治理需求；另一方面，统一的政务大模型测评与标准体系尚未建立，导致在技术选型、能力迭代上缺乏科学依据，既制约了政务服务数字化的整体效能，也增加了重复建设、低效投入的风险。在此背景下，政务服务垂直大语言模型开发应用与测评标准机制的构建，成为推动全国政务服务从数字化向智能化跨越的关键支撑，是筑牢数字政府根基的必然要求。

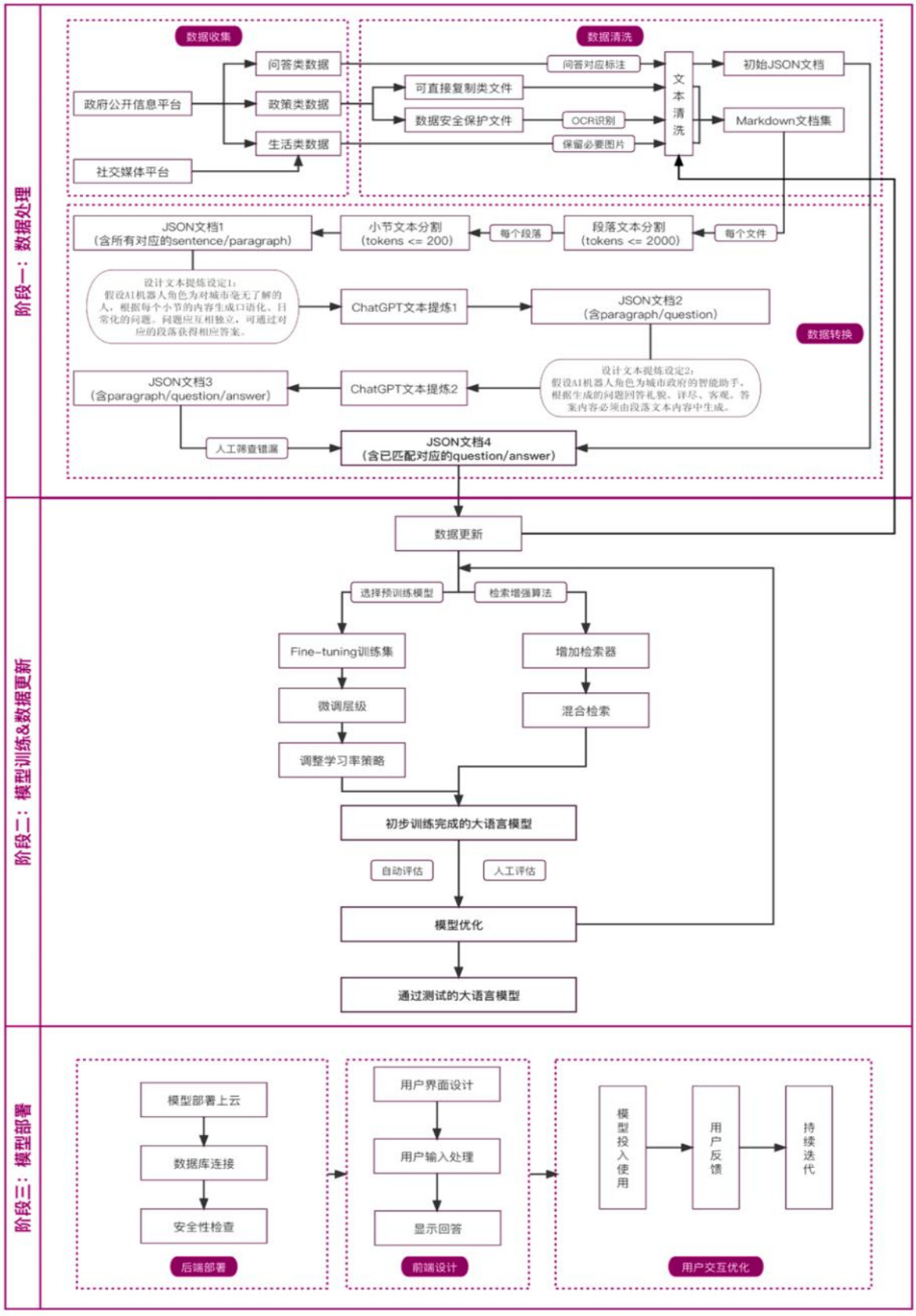
目前，政务服务智能升级已成为全国城市治理的共同命题，无论是超大型城市的复杂治理场景，还是中小城市的民生服务需求，都亟需以专业化技术底座、科学测评机制、全流程标准体系驱动，破解政策理解难、协同效率低、服务精准性不足、测评不统一、标准规范缺位等共性问题。政务服务垂直大模型开发应用与测评标准机制项目，通过垂直大语言模型开发提供即拿即用的智能政务技术支撑，依托专业化测评机制明确技术迭代方向，借助政务通用人工智能标准规范划定统一标尺，三者形成闭环，共同推动政务大模型技术规范高效迭代，进而实现政务服务质量的全域提升，为城市治理现代化提供可复制、可推广的核心技术与制度范式。

02 创新内容

为破解政务服务领域大语言模型场景适配性不足、测评与标准不统一的核心痛点，深圳市大数据研究院联合政务主管部门，推进政务服务垂直大语言模型开发应用与测评标准机制建设，以技术研发与标准构建为核心链路，形成政务智能服务升级的系统性解决方案，核心创新点如下：

政务垂直大语言模型技术研发

以通用大语言模型为底座，结合微调与检索增强技术，深度融入政府职能部门政策公告、便民服务指南等多源政务文本数据，覆盖跨层级、跨区域政策文件信息；研发“阿深”人才资讯大语言模型原型，具备语义理解、智能推理及跨区政策整合能力，可实现政策速配、办事流程智能指引、多轮互动问答等功能，为政府用户与市民提供统一、高效的政务信息获取渠道，有效解决政策碎片化、查询效率低、理解成本高的难题。


阶段一：数据处理
数据收集
问答类数据
政府公开信息平台
政策类数据
生活类数据
社交媒体平台
数据清洗
问答对应标注
可直接复制类文件
数据安全保护文件
OCR识别
保留必要图片
文本清洗
初始JSON文档
Markdown文档集
JSON文档1
（含所有对应的sentence/paragraph）
小节文本分割
（tokens <= 200）
每个段落
段落文本分割
（tokens <= 2000）
每个文件
设计文本提炼设定1：
假设AI机器人角色为对城市毫无了解的人，根据每个小节的内容生成口语化、日常化的问题。问题应互相独立，可通过对应的段落获得相应答案。
ChatGPT文本提炼1
JSON文档2
（含paragraph/question）
数据转换
设计文本提炼设定2：
假设AI机器人角色为城市政府的智能助手，根据生成的问题回答礼貌、详尽、客观。答案内容必须由段落文本内容中生成。
ChatGPT文本提炼2
JSON文档3
（含paragraph/question/answer）
人工筛查错漏
JSON文档4
（含已匹配对应的question/answer）
阶段二：模型训练&数据更新
数据更新
选择预训练模型
检索增强算法
Fine-tuning训练集
微调层级
调整学习率策略
增加检索器
混合检索
初步训练完成的大语言模型
自动评估
人工评估
模型优化
通过测试的大语言模型
阶段三：模型部署
模型部署上云
数据库连接
安全性检查
后端部署
用户界面设计
用户输入处理
显示回答
前端设计
模型投入使用
用户反馈
持续迭代
用户交互优化

政务大模型多维度测评体系构建

针对政务垂直领域大模型应用的新兴特性，设计目标层、准则层、指标层递阶层次分析评价模型；围绕政策理解准确率、内容安全性、交互友好度等核心维度，细化形成可量化的测评指标，并针对民生诉求分拨处理、智能打标、安全审核等典型场景，制定差异化测评方案；通过该体系可精准诊断大模型政务服务效果，科学评比不同模型的优劣差异，为政府部门选型决策、企业技术优化迭代提供专业指导。

政务通用人工智能标准规范编制

- 支撑统一算法管理系统、公共算法库与模型库建设，创新构建接入、运行、管理全流程政务算法标准化体系
- 接入环节明确算法镜像接入、容器微服务运行的技术规范，统一接口参数、数据格式标准
- 运行环节整合安全隐私、系统稳定性、服务质量与响应时效规范，形成监测、预警、处置闭环
- 管理环节融合算法效果评估、可解释性、决策公正性标准，确保算法决策透明公平

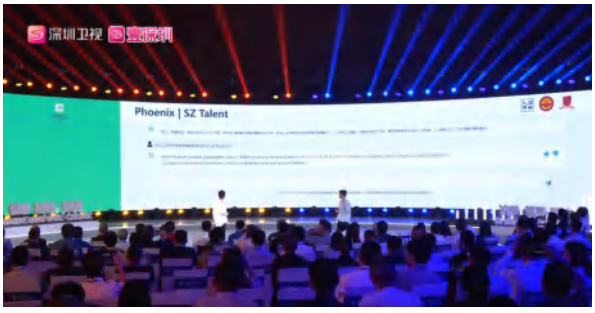
03 应用场景

政府部门政务服务效能提升	市民与企业政务咨询办理	政务大模型研发与迭代优化
为政务服务中心、12345热线等部门提供政务垂直大语言模型，支持跨区域政策整合、智能问答与办事流程指引，解决政策信息碎片化问题，提升群众咨询响应效率与导办精准度。	为来深人才、企业办事人员等群体提供政策速配与智能服务，通过大模型快速匹配区级人才补贴、企业开办流程等个性化需求，简化政策查询与申报环节，降低政务服务获取门槛。	为政务大模型研发企业提供多维度测评与标准规范体系，通过指标体系层次分析与规范统一，评估模型在政策理解、民生诉求处理等场景的性能，指导企业针对性优化模型功能，推动政务大模型适配政务服务实际需求。

04 重要成果

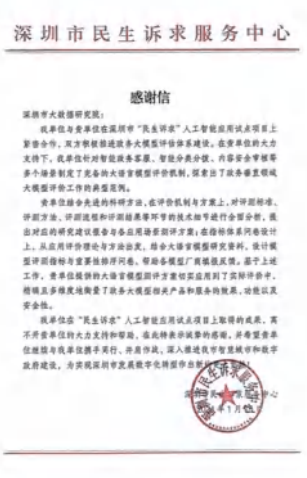
- **政务垂直大模型开发成果**

联合深圳市人才工作局，开发出“阿深”人才资讯大语言模型原型，实现政策速配、办事流程指导、多轮智能问答等功能；该原型已亮相深圳市招才引智活动、全球创新人才论坛，获深圳晚报、网易、腾讯等多家媒体报道。



- **政务大模型测评方案成果**

联合深圳市政务服务和数据管理局，完成政务服务大模型评测标准、评测方法、评测流程的系统分析，设计出覆盖民生诉求处理、政策速配等场景的指标体系与测评方案；该测评方案获参与民生诉求大模型开发企业的广泛征询，应用于实际评价工作，同时获深圳市民生诉求服务中心的感谢信认可。


深圳市民生诉求服务中心

感谢信

- **算法模型标准规范编制成果**

面向深圳市政务通用人工智能算法管理系统，编制算法接口、数据格式、安全和隐私、算法效果评估、算法可解释性要求、系统可用性与稳定性、智能算法与决策公正性、服务质量与响应时效等标准规范，支撑主流厂商算法开放式接入。

社会矛盾纠纷化解的智能辅助系统

01 项目背景

随着我国社会经济快速发展，社会矛盾纠纷呈多元化、复杂化趋势，传统纠纷化解方式面临效率不足、专业资源分布不均等挑战。人工智能技术的突破为社会治理现代化提供了新路径。项目依托法治大数据资源，构建集案件分析、裁判预测、法律检索、调解建议生成于一体的智能辅助工具，旨在提升社会矛盾纠纷化解的效率与精准度，助力构建共建共治共享的社会治理格局。

02 创新内容

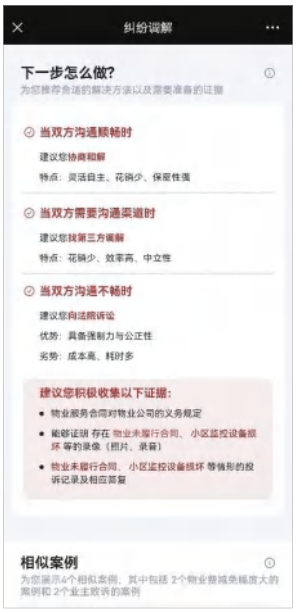
为此，深圳市大数据研究院依托**1.3亿**司法裁判文书的**大数据分析、文本识别、大模型摘要**等AI技术，基于矛盾纠纷具体情形，为当事人提供量化数据分析报告，弥合双方信息差，促进双方达成和解，避免进入诉讼程序，**助力诉源治理和社会安定**。

03 重要成果

选择纠纷情形
50多个物业纠纷问题

法院是否认可?
海量裁判文书大数据

法院会怎么判?
分析胜诉概率和成本



建议怎么做?
和解?调解?诉讼?

相似案例
提供精准案例参考

相关法律知识
相关度最高的法条

根据当事人信息，我们从海量司法裁判文书中精准定位最相似的案件，为当事人提供重要的参考信息，包括**法院可能判决结果、相似案例情形、相关法条、行动建议**等，辅助当事人科学决策。

CUSTOMER CASES
客户案例

华佗智能导诊

智慧医疗 智能导诊 智慧问诊 中文医学大模型

01 项目背景

深圳市龙岗区12家公立医院日常接诊量大，患者就诊需求多样，传统人工分诊压力大、效率有限，亟需通过智能化手段提升导诊效率与就诊体验。

02 行业问题

分诊压力大

医护人员数量有限，面对大量患者咨询难以快速精准分诊。

就诊效率待提升

患者对科室、症状对应关系不清晰，易挂错号、重复排队。

服务时间局限

人工导诊受工作时间限制，无法提供24小时咨询服务。

·龙岗区卫生健康局·

医疗卫生、公立医院体系

03 解决方案

部署“华佗GPT”智能导诊系统，作为国内首个可应用于学术与产业界的中文医学基础大模型，其参数量接近百亿级别，具备强大的医学知识理解与多轮对话能力。系统通过自然语言交互，为患者提供症状分析、科室推荐、就诊指引等服务，实现高效、精准的智能分诊。

有效减轻人工导诊压力
提升患者就诊效率和满意度

截止2025年12月，累计服务超

75万人次

总交互次数超过

140万次

AI家庭医生助手智能体

智慧医疗 AI家医 基层医疗 大模型 多智能体 健康管理 慢病管理

01 项目背景

为贯彻落实国家《“健康中国2030”规划纲要》及《新一代人工智能发展规划》，龙岗区积极推动医疗体系智能化转型，旨在通过大模型与多智能体技术，重构传统医疗服务模式，为辖区内居民提供更公平、高效、精准的智慧医疗服务。

02 行业问题

基层医疗资源不均衡

优质医疗资源集中在医院，社康中心服务能力与效率有待提升。

健康管理连续性不足

居民健康档案分散，缺乏全周期、个性化的动态健康管理方案。

医患沟通与效率瓶颈

医生工作负荷重，居民日常健康咨询、报告解读等需求难以得到及时响应。

03 解决方案

部署“龙小康”AI家医助手智能体医疗体系，以“模型即服务（MaaS）”为核心理念，打造区域级基层AI医疗新基建。

混合基座模型

采用LLMmarket+医疗垂域大模型架构，兼顾通用性与医疗精度。

多智能体集群

融合RAG、MCP、LLM Model Routing等技术，构建医疗智能体矩阵，深度融合知识库与业务场景。

八大核心功能

通过“华佗AI健康助手”提供症状问询、报告解读、药盒识别、慢病咨询、中医养生、就诊客服、健康科普、用药咨询等服务。

·龙岗区卫生健康局·

医疗卫生、公立医院体系、基层医疗

动态健康管理

智能分析居民全维度健康数据，生成定制化、长期性的“一人一策”健康方案。

图“龙小康”健康咨询对话界面

04 重要成果

项目已**在龙岗区所有区属公立医院社康全面上线，覆盖12家医院、183家社康中心**。截止2025年底，**龙岗家医小程序注册用户超40万人，累计AI交互次数约13万次**，有效实现了“让数据多跑路，让居民少跑腿”，显著提升了基层医疗服务效率与居民健康管理水平。

政务AI全链路方案：
垂直大模型开发应用、测评与标准规范

数字政务 垂直大模型 测评体系 标准化建设

01 项目背景

大语言模型为政务服务AI升级提供突破性可能，但当前通用大模型适配政务领域专业属性与复杂场景的能力不足，且缺乏统一的测评机制与标准规范，导致技术选型依据缺失、效能受限及重复建设风险，在此背景下，覆盖开发应用、测评、标准规范的全链路设计，有助于系统性解决政务大模型落地痛点。

02 行业问题

应用层面

通用模型与人才服务、民生诉求等垂直场景业务规则适配不足，跨部门数据壁垒导致训练语料质量不高，碎片化部署系统造成服务标准不一，这些潜在风险阻碍政务服务精准化、协同化目标实现。

测评层面

理论研究的测评指标多聚焦实验室静态数据，缺乏贴合政务实战场景的量化标准，且不同厂商模型性能缺乏统一评估基准，影响选型决策与模型迭代优化的科学性。

标准层面

算法接口、数据格式等基础规范不统一易抬高厂商技术接入适配成本，安全隐私防护等规范缺失存在合规风险，标准更新滞后于技术发展将制约政务AI规模化落地。

03 解决方案

政务垂域大模型技术攻坚

依托通用大语言模型基座，全面整合并沉淀政府政策公告等各类政务文本资源，联合打造垂域大语言模型原型。该原型凝练语义理解、智能推理、跨区域政策整合核心能力，可实现精准匹配、智能导办、互动答疑等实用功能，为政府部门与市民搭建政务信息高效交互的统一入口。

·深圳市政务服务和数据管理局、深圳市人才工作局·
·深圳市民生诉求服务中心·深圳市大数据资源管理中心·

政府机关、政务服务、社会治理、公共管理

政务多维度测评体系搭建

构建目标、准则、指标三级递阶层次分析评价框架，围绕多维度拆解形成可落地的具体测评指标，并针对智能分类分拨、智能标签标注、内容安全审核等代表性业务，设计场景化的测评实施路径，提供模型测评量化参考与实操指引。

政务通用AI标准规范制定

以赋能政务算法统一管理、公共算法库与模型库规范化运营为核心目标，在接入侧，界定统一的接口参数传输协议与数据交换格式，打通跨厂商、跨系统的算法接入通道；在运行侧，建立数据安全与隐私保护、服务效能管控、响应速度保障的一体化规范；在管理侧，纳入算法效果量化评估标准、算法可解释性要求、决策公平性审核规则，为政务AI规范应用夯实制度基础。

04 落地成效

1.政务AI应用方面，人才资讯大模型“阿深”紧扣深圳双招双引部署，亮相深圳市委、深圳市人民政府主办的高规格论坛，获多家知名媒体报道，放大深圳引才品牌影响力。

2.政务AI测评方面，针对民生诉求大模型构建全流程测评体系，**助力首批7个成熟场景落地**，经测评校准通过后，相关大模型成功面向实际业务开放调用，测评方案为模型筑牢性能与适配根基，获深圳市民生诉求服务中心的感谢信。

3.政务AI标准规范方面，从接口、数据、安全等维度制定统一标准，涵盖视频分析、文字识别等百余类算法，支撑主流厂商AI算法上架管理，被**深圳市政务通用人工智能算法管理系统采纳**，推动政务AI规范化落地。

Operations Optimization and Supply Chain

运筹优化与供应链

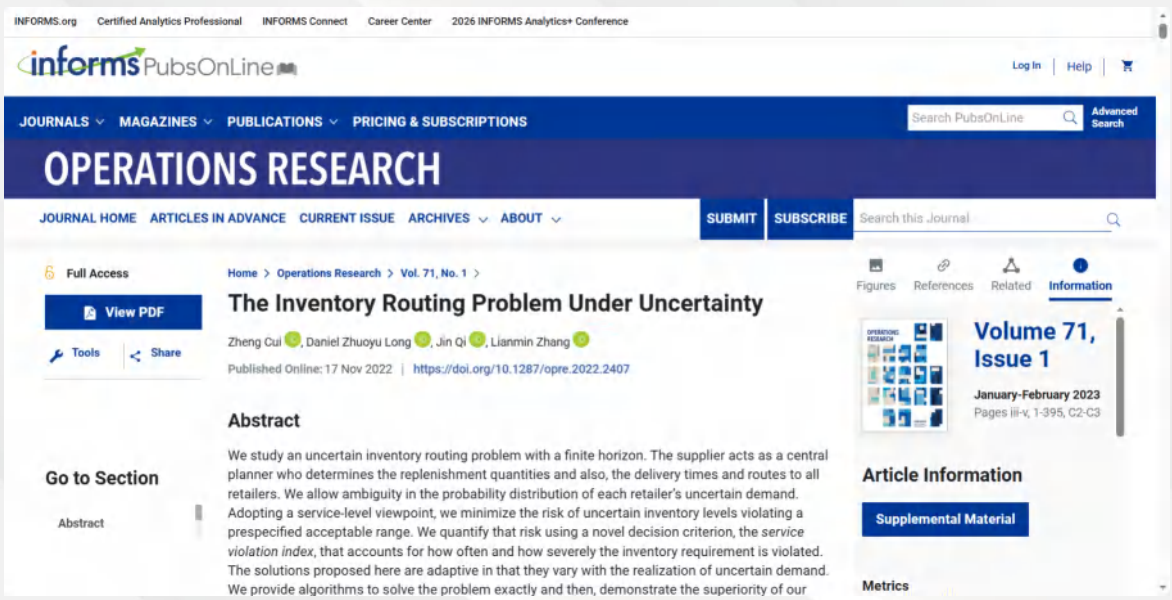
深圳市大数据研究院
Shenzhen Research Institute of Big Data

ACADEMIC ACCUMULATION
学术积累

不确定性下的库存路径问题

01 研究背景

- 库存路径问题是供应链管理中一个关键问题，涉及供应商在管理多个零售商库存的同时决定最优配送路径。
- 传统IRP研究多基于确定性需求，而现实中需求往往具有不确定性。
- 现有研究在处理不确定性时多假设需求分布已知，使用随机规划或马尔可夫决策过程，但实际中分布信息往往不完全。
- 传统鲁棒优化方法仅考虑最坏情况，忽略了频率信息，导致结果过于保守。
- 此外，现有研究多采用期望成本最小化为目标，但成本参数（尤其是缺货成本）难以准确估计，且风险中性视角忽略了决策者的风险厌恶。



02 核心内容

问题建模

- 考虑一个有限周期、单供应商多零售商的IRP问题，需求分布存在模糊性。
- 供应商作为中央决策者，决定补货量、配送时间和路径。

服务视角与风险度量

- 提出以服务水平为目标，要求库存水平维持在预设区间内。
- 引入服务违反指数（Service Violation Index, SVI）作为风险度量标准，综合考虑违反频率与严重程度。
- SVI基于凸风险度量构建，可扩展为多种形式（如基于CVaR或一般效用函数）。

分布鲁棒优化框架

- 采用基于场景的模糊集描述不确定需求，包含边界、均值、平均绝对偏差和相关性信息。
- 使用线性决策规则（LDR）使补货决策随历史需求实现自适应调整。
- 证明了传统Order-Up-to（OU）策略是LDR的一种特例。

算法与求解

- 将问题转化为混合整数线性规划，并利用最大访问节点集和有效路径削减技术提升求解效率。
- 提出Benders分解算法以处理大规模问题。

03 重要成果

数值实验基于真实油品公司数据和合成数据，比较了六种方法：Cost-R（鲁棒成本最小化）；MVP（最小化违反概率）；SVI-RS（带场景信息的SVI）；SVI-R（无场景信息的SVI）；SVI-S（随机SVI）；SVI-OU（基于OU策略的SVI）；

主要结论：

- SVI类方法在控制库存违反风险方面显著优于传统方法。
- SVI-RS在多数情况下表现最佳，说明场景信息的价值。
- SVI-R优于SVI-OU，说明线性决策规则比固定策略更灵活。

模型在109个节点、30个周期的大规模问题上仍具可解性，且求解速度快于Cost-R方法。

04 未来展望

算法改进	自适应路径决策	模型扩展
·开发更高效的算法或启发式方法，如利用经典IRP算法或设计有效不等式。 ·探索是否可将车辆路径问题中的风险厌恶算法（如Adulyasak & Jaillet, 2016）引入IRP。	·当前路径决策为非自适应，未来可研究二元决策规则使访问决策也具备适应性。 ·尽管计算挑战大，但可尝试新建模方式并测试其性能。	·可引入更多运营特征，如生产约束、多车辆、容量限制等。 ·进一步研究不同效用函数对SVI的影响及其在实际中的应用。

级联正交矩阵线性方程组的稀疏解的交替分裂算法

01 项目背景

对一般大规模线性逆问题开发高效通用算法极具挑战性。本文针对一类结构化的大规模线性逆问题问题设计定制化的高效算法技术，其中的矩阵由一系列正交矩阵串联形成。该结构在信号/图像恢复、分布式压缩感知和传感器网络中有重要应用。这种问题结构的核心价值在于低相干性，对鲁棒信号恢复至关重要。然而，现有的凸优化、匹配追踪和阈值迭代等方法将矩阵视为整体，未能利用矩阵的内部结构。压缩采样理论表明，算法性能与测量矩阵结构及其特性（如相干性、受限等距性等）密切相关，仿真亦证实矩阵结构直接影响信号恢复算法的实际性能。因此，亟需开发能充分挖掘矩阵结构并保持低计算成本的高效算法。


Home → SIAM Journal on Matrix Analysis and Applications → Vol. 46, Iss. 4 (2025) → 10.1137/24M1707341

Previous Article　　Next Article

FULL ACCESS

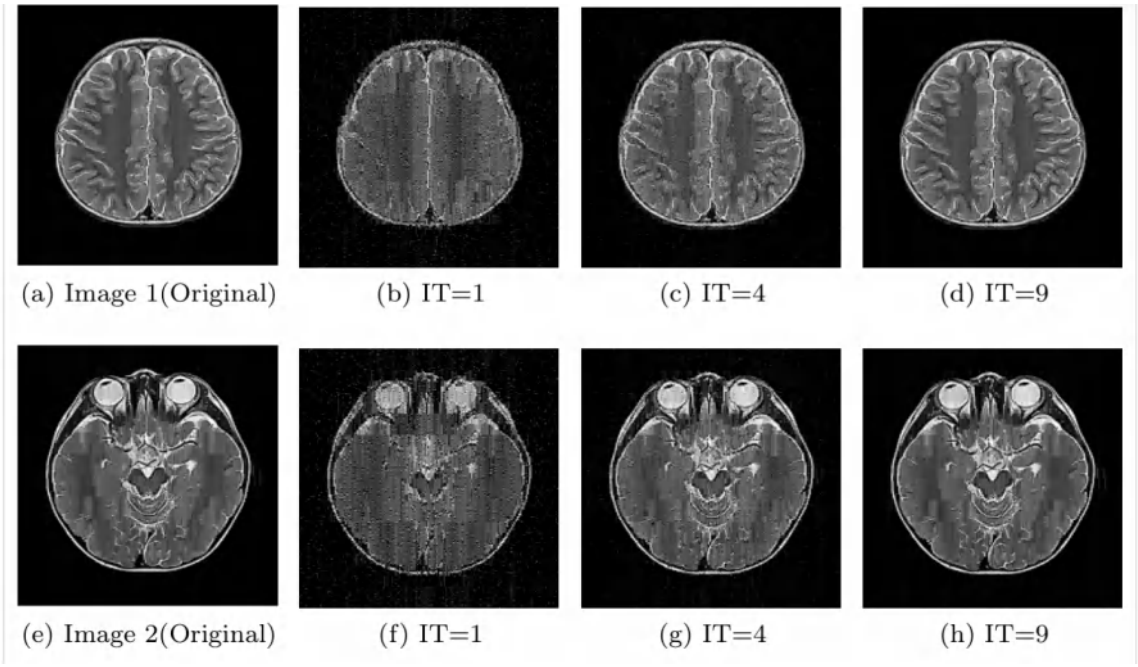
Splitting Alternating Algorithms for Sparse Solutions of Linear Systems with Concatenated Orthogonal Matrices

Authors: Yun-Bin Zhao and Zhong-Feng Sun　AUTHORS INFO & AFFILIATIONS

https://doi.org/10.1137/24M1707341

02 核心内容

首先，针对级联正交矩阵结构的大规模线性逆问题设计了双块（TSAA）和多块（MSAA）分裂交替迭代算法。其目的是将大规模线性逆问题分解为多个耦合子系统进行交替迭代求解，以超低迭代成本完成对问题的高效求解；其次，对算法进行严格理论分析，利用矩阵的相干性值及问题解的稀疏度作为条件，分析算法的理论有效性和收敛性；然后，通过数值模拟实验研究了算法性能优势，该实验表明算法求解问题的成功率和收敛速度显著优于现有的算法技术（包括凸优化方法，正交匹配追踪，及FISTA等主流方法）；最后，将算法应用于医学MRI图像重建，检验算法对实际信号恢复与重建的效能，并与传统算法进行了比较，发现了算法的几个独特性能和计算优势。



(a) Image 1(Original) (b) IT=1 (c) IT=4 (d) IT=9

(e) Image 2(Original) (f) IT=1 (g) IT=4 (h) IT=9

03 重要成果

充分利用问题的正交级联结构，成功设计了首个分裂融合计算方法，实现了线性逆问题的稳定和快速求解，算法仅需矩阵-向量乘积与降维正交投影，计算成本极低；在相互关联性条件下，首次严格证明了算法产生的迭代点列全局收敛到问题的唯一最稀疏解，并给出了算法的线性收敛速率；所提出的算法避免了使用传统的凸优化算法求解大规模问题时所面临的维数灾难，克服了传统阈值型算法“对参数的依赖性和敏感性”的劣势。数值模拟验证了所提出的算法通常只需要极少迭代次数（通常几步迭代）就能够高质量重构稀疏图像，展现出了算法优异的实际图像处理潜力。

04 未来展望

这项研究为线性逆问题的研究展现了一个新的研究视角——深入挖掘问题的结构是获得问题高效求解的可行方向；这个算法设计思想有助于启发大规模线性逆问题的针对性算法设计，有助于推动问题的求解规模迈入大规模计算的最终新时代。

TECHNOLOGICAL BREAKTHROUGH
技术攻关

通用优化求解器：仙鹏求解器

01 项目背景

全球优化求解器市场已趋于成熟且格局稳定

Gurobi、CPLEX、Xpress长期领先达十余年，上述三家求解器厂商均来自于欧美国家。求解器头部厂商持续投入推动LP/MIP技术的迭代，通常每5年实现在算法层可以实现一个数量级的提速。尽管存在开源MIP求解器，但在速度、稳健性与工程成熟度上普遍不及商业产品，求解MIP/MINLP常慢一个量级以上。国产求解器的研发最近3-5年开始起步（如杉数、华为、阿里），在LP/MIP已经取得了不错的进展，但在建模表达、并行能力、算法优化与产品完备度方面仍与国际领先水平存在差距。

线性规划（Linear Programming，简称LP）

线性规划是最优化问题的一个重要分支，在物流、航空、生产制造、能源和金融等领域有着广泛的应用。此外，线性规划还是混合整数规划求解器必不可少的重要模块之一。世界前三大商业线性规划求解器是Guribi、CPLEX和Xpress，目前XOPT的求解性能和CPLEX和Xpress的求解性能相当，与Gurobi相比依然存在着一定的差距。近年来，Gurobi持续优化其单纯形算法与内点法算法，虽已退出国际权威测评榜单，但其线性规划求解器仍被公认为处于全球领先地位；而CPLEX自2022年开发团队解散后已基本停止主要功能更新。在Mitelmann教授维护的权威评测榜单中，杉数科技的COPT求解器已超越CPLEX跃居全球第二，华为天筹求解器紧随COPT位列第三，彰显了国产求解器在线性规划问题上求解的竞争力。

混合整数线性规划求解器梯队划分

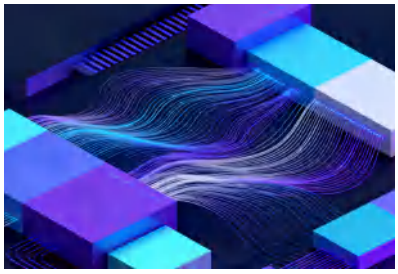
相较之下，XOPT在求解性能方面与上述知名商业求解器仍存在可量化的差距。Gurobi在混合整数线性规划领域占有绝对的优势，短期之内其他商业求解器难以撼动其在混合整数线性规划领域的地位。杉数去年发布了最新的COPT7.0版本，相比其上一个版本6.5版其在混合整数线性规划领域的求解器速度有了很大的进步，从测试榜单上来看已经超越了CPLEX，在国产求解器中排名第一。综合来看可以认为在混合整数线性规划领域，Gurobi处于第一梯队；CPLEX、COPT和华为天筹求解器处于第二梯队；Xpress和XOPT求解器处于第三梯队。

02 创新内容

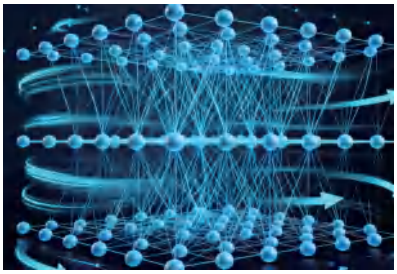
为突破国产求解器软件发展瓶颈，深圳市大数据研究院牵头研发“仙鹏求解器”，核心创新点如下：

创新点1：稳健且并行的LP内核协同

通过改进pivot/扰动策略与退化处理，显著降低单纯形迭代次数并提升数值稳定性；引入Shift-Cost/Shift-Bound等机制进一步稳态收敛。结合Left-looking、Supernode、Multifrontal等LU技术与并行Cholesky，加速关键分解环节。采用同质自对偶（HSD）框架的内点法，与单纯形形成互补的混合求解路径，在不同规模与条件数下自适应切换。



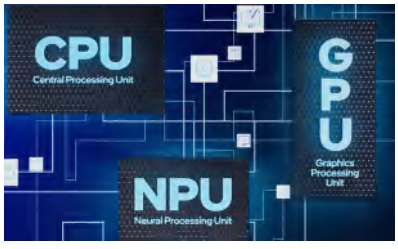
创新点2：结构感知的通用 MILP 框架



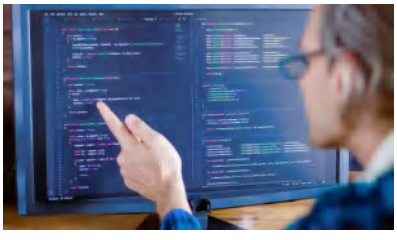
基于分支定界，强化预处理（界限收缩、系数缩紧、对称识别）以压缩搜索空间；以凸包逼近+可行域分解提升下界质量与树搜索效率。融合SAT技术的LP-free启发式，在不依赖LP的情况下快速生成高质量可行解，再用LP/MILP-based启发式迭代改良。研发了基于CPU多线程并行的分支定界算法加速框架，采用了动态负载均衡技术，实现了CPU多线程并行对混合整数线性规划求解的加速。

创新点3：先进算力融合与国产生态适配

面向GPU/NPU深度优化一阶与Objective-Splitting算法，包含Max-QP Newton算子与向量规约在内的关键算子进行内核化与流水并行。在算法层面重构步长/预条件与同步策略，降低通信与访存瓶颈。全面适配鲲鹏、海光、龙芯及欧拉、麒麟、凝思等平台，提供从依赖编译到运行验证的端到端的适配方案。



创新点4：可编程建模与可观测性的产品化提升



提供覆盖C++/Python/Java的一致性API，并无缝对接PuLP、Pyomo等主流建模库，降低集成门槛。开放更细粒度的参数与中止条件控制，支持不同算法策略的快速切换与调优。内置上/下界演化与割/节点统计的可视化监控，配套可复现的示例与文档，形成“即插即用、可解释、可运维”的工程化产品形态。

03 应用场景

生产计划与排程（制造业）

典型问题：批量/轮班主生产计划（MPS）、车间排程（JSP/FJSP/HFSP）、换型与产能约束同步。
建模要点：用二进制变量表示作业在机器上的指派与顺序，时间索引或不相容约束保证容量与优先关系；成本/交期/加班作为目标。
求解器作用：预处理压缩规模、割平面收紧松弛、启发式与MIPStart快速找首解；滚动计划中用上期解作为热启动，可以实现显著的提速。

供应链网络设计与库存优化

典型问题：选址-定容（设施选址/仓网优化）、多级库存与物料流（多期网络流+固定费用）。
建模要点：设施启用/容量为0-1决策，流量守恒与运输容量线性约束，固定费用和运费/库存持有费构成目标。
求解器作用：通过强不等式（固定费用流、覆盖类割）和对称性处理提升下界；可结合列/割生成或Benders分解求大规模实例。

车辆路径规划问题(PDPTW)

典型问题：干线/支线配送、时窗约束、容量限制、多仓、多车型。
建模要点：路径选择（集合划分）或弧变量模型，约束包含子环消除、时间推进与装载容量。
求解器作用：分支割（SECs、容量割、强不等式）与列生成（SPPRC子问题）结合；用启发式得初解、切割提升下界，支持时效性优化。

人员排班与班表（Crew Scheduling/Rostering）

典型问题：满足班次覆盖、技能匹配、劳动法规与休息规则，同时兼顾公平与偏好。
建模要点：集合覆盖/分配型0-1模型，软约束以惩罚项进入目标；可分解为列生成（把可行轮班当列）。
求解器作用：自动处理大量逻辑与不等式、软硬约束混合；列生成+分支定价或直接MILP都可，IIS功能帮助定位不可行原因。

电力系统优化（机组组合问题）

典型问题：机组启停、最小开停机、爬坡、出力上下限与网络潮流（线性化DCOPF）；成本最小或排放/备用约束。
建模要点：二进制表示启停，半连续或PWL表达出力与分段成本，网络为线性流约束。
求解器作用：在强formulation下以分支割求解，PWL/半连续/指示约束由引擎专门加速；滚动计划用MIPStart/VarHint稳定秒级出解。

04 重要成果

自主知识产权的仙鹏求解器

- 自2021年起实现100+核心算法，超过50万行C\C++代码，完成从算法库到通用优化求解器产品化的自主研发闭环
- 线性规划（LP）求解性能已可与欧美主流商业软件媲美，支撑工程级稳定与可扩展应用
- 混合整数线性规划（MILP）在速度与稳健性上整体优于最好的开源求解器，显著缩短求解时间
- 在国际第三方平台（Mittelmann/PLATO）测试中，XOPT进入全球前几位序列，具备国际竞争力。XOPT已开放免费下载，便于科研与工业用户快速验证与集成

仙鹏求解器产学研合作落地成果

- 在科研方面承担两项国家重点研发计划“大规模问题的学习优化算法”课题，推进XOPT在通信网络优化中的方法与应用创新
- “大规模混合整数规划的自适应优化算法与软件”，开展求解器前沿算法理论探索
- 在产业化方面与南航、中广核、华为、中航信、广铁集团、美团/顺丰、国药集团等建立合作，推动在航旅、能源、轨道交通、供应链等场景的规模化应用

中华人民共和国科学技术部 **国家重点研发计划**

1：“学习优化理论与方法及其在5G网络中的应用”
2：“复杂系统的通用优化模型、理论与算法及其应用”

HUAWEI **华为**

1：“基于学习优化的大规模混合整数优化技术研究”
2：“大规模优化问题的fix-and-optimize求解技术研究”

中国南方航空 CHINA SOUTHERN 宇器科技 **南方航空、杭州宇器**

1：“收益管理算法和系统架构研究”
2：“应用于制造场景的物料需求计划的优化算法研究”

面向大数据和优化算法的智能应急决策支持系统

01 项目背景

为了推动应急管理建设，推进铁路应急工作专业化、智能化、智慧化，实现铁路安全发展，深圳市大数据研究院蔡小强教授团队与中国铁路广州局集团有限公司合作，研发了一款基于大数据和优化算法的应急决策支持系统，该项目荣获广州局集团科技进步二等奖。

02 创新内容

该项目是中国铁路广州局集团2021年第二批科研项目计划中的科研专项。针对广铁集团实际应急调度场景的痛点，蔡小强教授团队与广铁集团信息技术所建立了紧密的产学研合作，最终开发出了一套行之有效的解决方案。该方案主要聚焦铁路防洪应急场景，旨在通过数据驱动的模型和算法，有效地调配有限的应急资源，以达到在最短响应时间、运用最低成本进行智慧防洪。

03 应用场景

该模型可应用于广铁集团下辖省市间洪涝灾害的事前防控及事后救助，能够提升铁路应急物资的运输效率并降低其部署成本。

04 重要成果

真实铁路网络

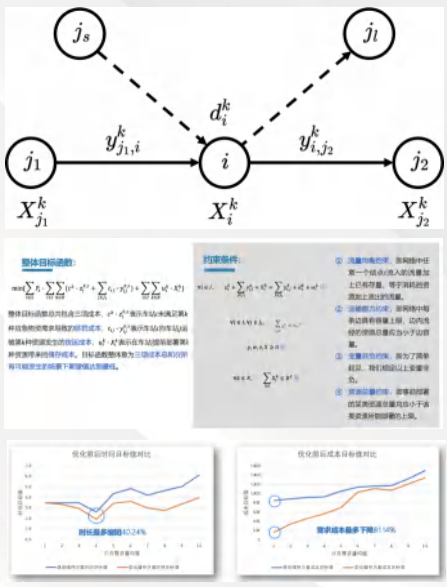
基于真实铁路线路及火车行驶规划，构建应急物资“供给点-途径站-需求点”网络模型。

随机优化模型

针对不确定性的洪涝事件，使用随机优化技术对应急资源进行最优配置。

模型表现情况

通过随机优化模型，该系统在时间和成本方面都取得了显著的改进，最多能节省超过40.24%的响应时间和降低81.14%的成本。



基于NPU的不确定需求下带时间窗取送货路径优化系统

01 项目背景

在现代物流配送系统中，取送货路径规划（Pickup and Delivery Problem, PDP）是核心优化问题之一。传统的确定性PDPTWVRP模型假设所有客户需求、服务时间和配送数量在规划阶段已知且固定。然而，实际运营中面临诸多不确定因素：1.客户订单量、取货/送货需求量存在随机性；2.实时订单、临时取消、需求调整频繁发生；3.传统随机规划方法需要考虑大量不确定场景，导致计算资源消耗巨大。本项目针对上述挑战，提出基于数据驱动的场景约简方法，在保证解的质量情况下同时大幅降低计算复杂度。

02 创新内容

创新点一：需求不确定性建模框架

尝试将客户需求量不确定性引入到模型中

构建基于历史数据的概率分布模型

创新点二：机器学习驱动的智能场景约简

利用机器学习方法，如变分自编码器(Variatioal Autoencoders, VAE)提取场景数据的低维隐变量表示

通过聚类算法，如K-means聚类得到代表性场景

创新点三：两阶段随机规划与ML集成架构

第一阶段：基于ML的代表场景选择

第二阶段：通过NPU-启发式算法对模型进行求解

03 应用场景

PDPTWVRP问题可面向外卖、生鲜、同城快递等即时配送场景，通过机器学习提取需求波动的代表性场景（早晚高峰、节假日、恶劣天气），构建配送路径规划方案。帮助系统在订单量剧烈波动（午高峰需求可达平峰3-5倍）的情况下，提前预留运力缓冲并动态调整路线，显著降低超时率与空驶成本，同时支持突发订单的快速重规划，帮助配送系统成为可预测、可调度的柔性资源池。

04 重要成果

学术论文产出

在学术期刊上发表《Machine Learning-Enhanced Scenario Reduction for Stochastic Pickup and Delivery Problem with Time Windows》

算法与软件产出

针对机器学习模型，保存相关训练数据集，并对模型权重进行记录；针对启发式算法部分，开源相关核心代码。

专利与知识授权

进行相关软著“一种基于NPU的不确定需求物流路径规划方法及系统”的申请。

CUSTOMER CASES
客户案例

药物配送智能调度软件

医药科技研发 医药工业制造 医药商贸流通 医药卫生健康

01 客户背景

中国医药集团有限公司（简称“国药集团”）是国内医药行业的龙头企业，业务覆盖医药研发、生产、流通及卫生健康服务全链条，承担着全国范围内药品（含冷链、特殊药品）的仓储与配送任务。作为医药商贸流通的核心主体，其配送网络需高效覆盖医院、药店等终端，对配送时效性与合规性要求极高。

02 行业问题

订单体量巨大：

药品SKU超几十万种，每日配送订单量庞大，传统人工或常规算法难以快速生成最优配送方案；

配送规则复杂：

药品类型多样（包括冷藏药品、冷冻药品、特殊管制药品等），不同品类对温控、运输条件、配送时段有严格限制；

时间要求苛刻：

部分药品需严格按指定时间送达指定地点（如急救药品、门诊用药），延迟可能影响医疗安全；

配送成本压力：

大规模配送需平衡车辆调度、路线规划与燃油消耗，降低整体运营成本是关键诉求。

·中国医药集团有限公司·

药物配送、智能调度、路径优化

03 解决方案

深圳市大数据研究院联合国药集团开发「运输管理系统软件」，通过以下技术路径解决问题：

数据整合：接入实时运单数据（订单量、药品类型、配送时效要求）与车辆数据（车型、载重、温控能力、当前位置）；

智能建模：基于数学模型将配送问题转化为“车辆路径规划问题（VRP）”，综合考虑订单优先级、药品特性、交通约束等变量；真实地图数据：通过标准的地图API数据可以将自然语言描述的地理位置转化为可供算法使用的经纬度信息，以便于实现对车辆的位置和送货地位置以及他们之间距离的精确计算；

优化求解：通过优化求解器输出最优配送方案（车辆路径、装载顺序、配送时段），实现：

·效率提升：最小化配送车辆数量与总行驶距离，降低空驶率；

·合规保障：自动匹配药品温控要求与车辆资质，确保特殊药品运输合规；

·时效精准：严格满足“给定时间给定地点送达”的严苛要求。

在全国的药物配送框架中

行驶路程4679km~3800km
减少

18.8%

订单延误率

14.6%~0%

全年预计实现经济成本节约

2000万元

Scientific Computing and Industrial Software

科学计算与工业软件

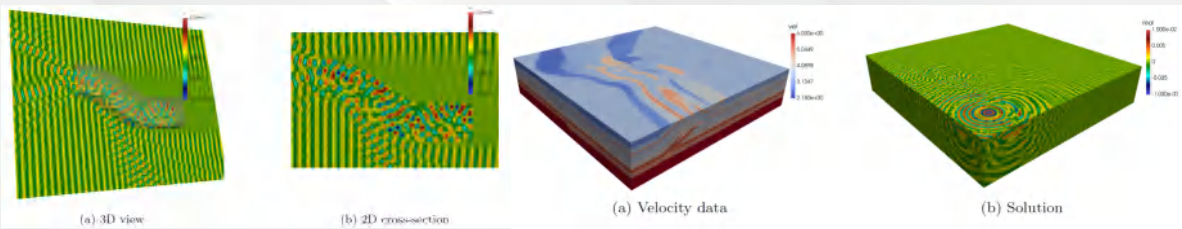
深圳市大数据研究院
Shenzhen Research Institute of Big Data

ACADEMIC ACCUMULATION
学术积累

高效并行Schwarz方法求解高波数亥姆霍兹方程

01 项目背景

- 在声学、电磁学、地震勘探等众多科学与工程领域，高波数亥姆霍兹方程是描述时谐波动现象的核心数学模型，其求解过程面临着严峻挑战。该方程求解会产生大规模、病态、不定且复值的线性方程组，随着波数k（与频率正相关）的增大，方程的数值求解难度呈指数级上升。
- 传统数值方法在处理高波数问题时，存在诸多瓶颈。一方面，为避免“污染误差”，网格尺寸需随波数增大而急剧减小，导致计算量呈$O(k^d)$（d为问题维度）增长，单机计算难以承载；另一方面，现有并行求解方法（如传统区域分解法）常因边界条件处理不当、通信开销过大或收敛性依赖于不切实际的参数设置（如过宽的完美匹配层PML区域），无法在高波数场景下实现高效、稳健的求解。例如，固定宽度的PML和重叠层在子域数量增加时，会导致子域尺寸与扩展区域宽度相当，使方法失去实用性，亟需突破现有技术瓶颈，开发适用于高波数亥姆霍兹方程的高效并行求解框架。



02 项目简介

本项目聚焦高波数亥姆霍兹方程的并行数值求解，以限制加法Schwarz方法（RAS）为基础，结合完美匹配层（PML）技术，提出了一种具有快速收敛特性的实用并行求解算法（RAS-PML改进方法）。项目核心目标是解决高波数场景下方程求解的“大规模计算”与“高效收敛”双重难题，实现算法在并行环境下的强可扩展性、宽场景适应性（如变介质、不同网格精度）及低通信开销。

项目技术路线围绕“问题建模-算法改进-数值验证-性能优化”展开：

- 首先，基于时谐波动方程推导高波数亥姆霍兹方程的弱形式，并通过PML技术将无界计算域转化为有界域
- 其次，针对传统RAS-PML方法的缺陷，提出关键改进
- 最后，通过大量数值实验验证算法的收敛性、效率与稳健性，并优化并行计算资源配置

03 创新内容

本项目在算法设计与工程实现层面提出三大核心创新，突破传统方法的局限性：

阻抗与PML边界条件融合

针对单一边界条件的不足，将阻抗与PML边界条件结合应用，在PML区域内通过复坐标拉伸实现PML条件，同时叠加阻抗条件。该方案虽非“精确无反射条件”，但实际计算中可近似为无反射效果，在PML较薄时仍能保持收敛，提升算法稳健性。

PML与重叠层对数缩放策略

传统方法中固定宽度的PML和重叠层在子域数量增加时会失效，本项目提出PML与重叠层的网格点数随子域数量呈对数增长，避免“扩展区域吞噬子域”的问题，确保迭代次数随波数线性增长，而非指数增长。

重叠区域残差通信优化

通过分析发现，算法迭代更新中残差仅在子域重叠区域非零，因此仅传递重叠区域的残差数据，大幅降低并行通信开销，使计算时间主要集中于局部求解，提升并行效率。

04 应用场景

本项目提出的并行RAS-PML方法可广泛应用于需求解高波数波动问题的科学与工程领域，具体场景包括：

声学工程：高频率声学模拟

适用于航空发动机噪声预测、超声诊断、建筑声学设计等场景，可高效求解高频率声学亥姆霍兹方程，为噪声控制、声学设计提供数值支撑。

电磁学：高频电磁散射与辐射

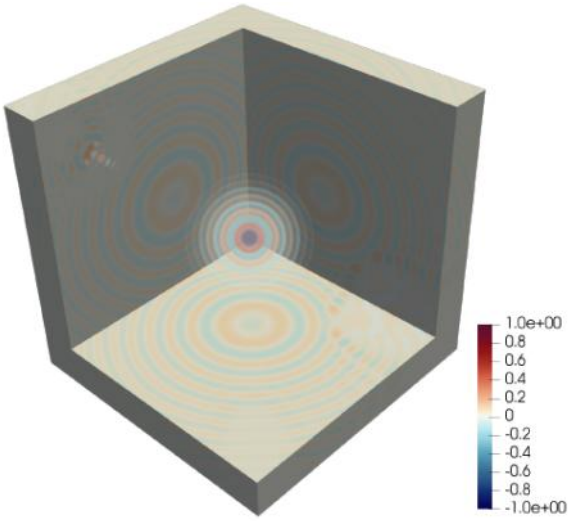
可用于雷达探测、无线通信、天线设计等领域，模拟电磁波在复杂介质中的传播，为雷达目标识别、天线性能优化提供技术支持。

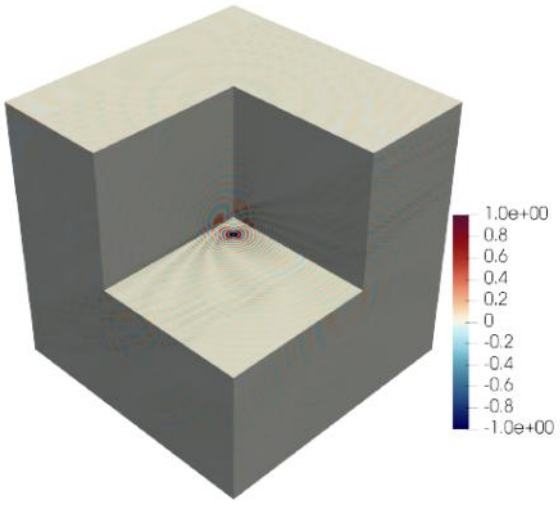
地震勘探：深部地质高分辨率成像

能处理地下非均质介质场景，准确模拟地震波传播路径，为地震数据反演、油气藏定位提供可靠的正演模拟结果。

并行计算软件：大规模科学计算工具集成

可集成至开源或商业科学计算软件（如PETSc、COMSOL）中，适配超算平台，处理千万级网格规模的高波数问题。





05 重要成果

理论成果：确立算法收敛性与复杂度边界

收敛性保证：算法的固定点迭代收敛速度随波数k增大而指数提升，误差可通过迭代次数控制至任意精度。

复杂度优化：算法并行度达$O(k^d)$（d为问题维度），增长迭代次数随k呈线性增长，总计算量增长显著优于传统方法。

数值实验成果：验证算法效率与稳健性

常数介质场景：当波数对应频率从300增至9600（增大32倍），子域数从4增至4096时，迭代次数与总运行时间均线性增长，“总时间/频率”比值保持稳定。

变介质场景：在“球介质”“层状介质”等非均质场景中，迭代次数虽略有增加，但仍保持线性增长趋势，证明算法对非均质介质的适应性。

网格鲁棒性：在不同精度网格中，迭代次数差异小，算法不受网格精度适度降低的影响，可平衡精度与计算成本。

并行性能成果：实现强可扩展性

强 scalability验证：固定问题规模下，子域数在合理范围内增加时，加速比接近线性；确定最优并行配置，兼顾速度与资源效率。

工程实现：基于有限元离散与局部求解器，实现4核到4096核的并行部署，支持千万级网格的高波数问题求解，内存占用控制在合理范围。

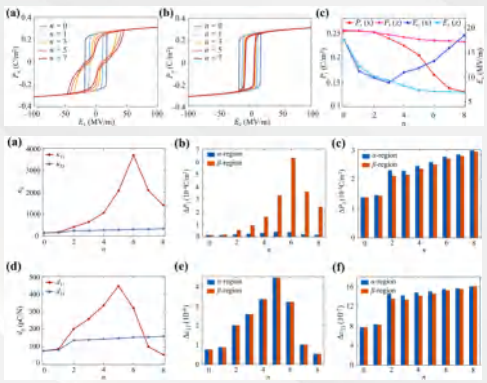
基于相场模拟的位错诱导钛酸钡单晶中巨介电及压电响应研究

01 项目背景

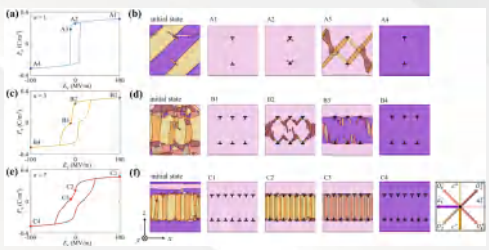
铁电材料因其独特的机电转换能力，在存储器、传感器和微机电系统等电子器件中具有广泛应用。随着市场对高灵敏度、高可靠性电子设备需求的不断增长，开发高性能铁电材料成为研究重点。除了传统的化学改性（如构建相界、元素掺杂）方法外，缺陷工程，特别是利用位错来调控材料微观结构和性能，展现出巨大潜力。位错不仅伴随有应力场，其带电核心和屏蔽空间电荷层还能产生独特的机电特性，为畴结构设计和性能调控提供了新机遇。然而，位错影响铁电材料性能的微观机制尚不明确，需要深入研究。

02 核心内容

本研究旨在系统探究位错偶极子密度对钛酸钡单晶畴结构及其介电、压电性能的影响机制。研究团队结合了相场模拟和热力学分析两种方法。相场模拟通过求解时间相关的金兹堡-朗道方程，模拟了在不同位错密度下，铁电畴在外加电场中的演化过程。热力学分析则通过构建吉布斯自由能模型，将位错引起的复杂应力分布简化为平均应力，从而定性揭示了性能变化的本质原因。研究重点关注了在钙钛矿铁电体中最易激活的〈100〉位错。



畴结构与电滞回线变化



模拟结果表明，畴结构强烈依赖于位错偶极子的密度。当密度较低时，畴结构主要由初始随机状态决定；当密度较高时，应力场导致畴结构出现明显的空间分异：拉伸应力集中的α区域稳定存在a^+/a^-畴，而压缩应力集中的β区域则倾向于形成c^+/c^-畴。这种差异直接影响了电滞回线的形态。当沿x方向施加电场时，出现了类似于反铁电体的双电滞回线，这是由于畴的翻转过程从单一步骤变为多步骤进行，涉及中间亚稳相的形成。而沿z方向施加电场时，回线形状基本保持不变，但面积随位错密度增加而减小。

性能增强与机制分析

研究发现了介电和压电性能的巨大增强。当位错偶极子数量为6时，相对介电常数κ_{11}和压电系数d_{11}分别达到峰值3693和445pC/N，相较于不含位错的情况，提升幅度高达2416%和478%。这种非单调的性能变化（先增强后减弱）归因于位错的双重角色：在中等密度下，位错应力场诱导β区域形成高响应的c畴，极大地提升了整体性能；而在高密度下，过强的压缩应力对畴翻转产生了钉扎效应，反而抑制了性能。热力学分析证实了位错应力可以改变体系的能量势垒，诱导局部相变，从而为性能调控提供了理论解释。

其他影响因素探讨

研究还探讨了挠曲电效应和不同位错类型的影响。分析表明，位错产生的应变梯度所导致的挠曲电效应局限于位错核心区域，对整体的畴结构和性能影响微弱。此外，与〈100〉位错相比，1/2〈110〉位错虽然增加了畴壁数量，但并未诱导出独特的双电滞回线，其对性能的调控能力相对较弱。这凸显了位错类型和构型在缺陷工程中的重要性。

03 重要成果

本研究通过相场模拟与热力学分析相结合的方法，系统探究了位错偶极子密度对钛酸钡（BTO）单晶性能的调控作用，结果表明：BTO单晶的稳定畴结构具有显著空间依赖性，α区（拉伸应力）富集a_1^+/a_1^-畴变体，β区（压缩应力）富集c^+/c^-畴变体，x方向电场下因多步畴切换形成类反铁电的双P-E滞后回线，z方向电场下滞后回线形状基本不变仅面积调整；介电与压电性能呈现各向异性响应，相对介电常数κ_{11}峰值达3693（较无位错偶极子提升2416%），压电系数d_{11}峰值达445pC/N（提升478%），二者随位错偶极子密度呈非单调变化，而κ_{33}与d_{33}则单调递增，且α区与β区对κ_{33}、d_{11}和 d_{33}的贡献相当，κ_{11}的巨幅增强主要源于β区c^+/c^-畴变体的高响应；极化与矫顽场特性依赖于电场方向，x方向电场下剩余极化P_r显著下降、矫顽场E_c呈“先降后升”非单调变化，z方向电场下二者均单调下降；挠曲电效应仅局限于位错核心区域，对整体性能及核心结论影响可忽略，而1/2〈110〉位错未产生双滞后回线，性能调控效果与〈100〉位错存在显著差异，上述结果均源于位错偶极子影响区的相变及应力场调控作用。

04 未来展望

本工作通过相场模拟和热力学分析，深入揭示了通过调控位错偶极子密度来显著增强铁电材料性能的微观机制。研究证明，位错工程是一种无需化学改性的有效性能优化策略。该研究为通过缺陷工程设计高性能铁电材料和器件提供了重要的理论指导。未来研究方向可以包括探索更复杂的位错构型（如位错网络），研究位错与其他类型缺陷的协同效应，并通过实验验证模拟预测的结果，以推动其在先进功能器件中的应用。

TECHNOLOGICAL BREAKTHROUGH
技术攻关

面向大型作业现场的智慧安监系统

01 项目背景

大型作业现场（如电力、石化、基建、制造等）普遍存在作业环境复杂、风险点多、人力监管难覆盖等问题，传统安全监管方式效率低、响应慢，难以应对动态变化的安全隐患。为提升大型作业现场的安全管理水平，本项目基于人工智能与物联网技术，构建了一套智能化、全覆盖、实时响应的智慧安监系统，实现对多类风险行为的自动识别与预警。

02 创新内容

本系统是一款基于人工智能技术面向各类大型作业现场的智能安全监管平台，具备以下创新点：

全场景适配

支持多种作业类型（高空、起重、有限空间等），可根据不同现场需求灵活配置算法与规则，实现跨行业、跨场景的智能监管。

多模态技术融合

集成目标检测、多目标跟踪、步态识别、人脸识别、ReID等技术，实现对人员、设备、环境的全方位感知与行为分析。

高精度算法引擎

基于YOLOv5、DeepSORT、GaitSet等模型进行优化，提升小目标检测、遮挡处理、跨镜跟踪等能力，识别准确率超行业标准5%~20%。

低代码平台化设计

支持快速部署与算法迭代，提供可视化配置界面，用户可自定义风险规则与响应策略，降低使用门槛。

边缘-云协同架构

支持模型压缩与边缘计算，实现低延迟实时分析，同时具备云端数据汇聚与大数据分析能力。

03 应用场景

本系统适用于以下典型大型作业场景：

电力行业：
变电站检修、输电塔施工、发电机吊装等

石化行业：
罐区作业、管线巡检、动火作业监管

基建施工：
桥梁建设、隧道施工、高空作业管理

制造车间：
大型设备安装、流水线安全行为监控

应急管理：
人员聚集预警、疏散通道监测、危险区域管控

系统已在全国多个大型现场（如抽水蓄能电站等）完成部署与验证，具备良好的通用性与实用性。

04 重要成果

技术指标

支持18类以上风险行为的识别，平均准确率超85%，部分场景如安全帽佩戴识别达97%，区域闯入识别达94%。

应用成效

在多个抽水蓄能电站等现场实现违规行为实时告警，提升安全管理效率30%以上。

学术与知识产权

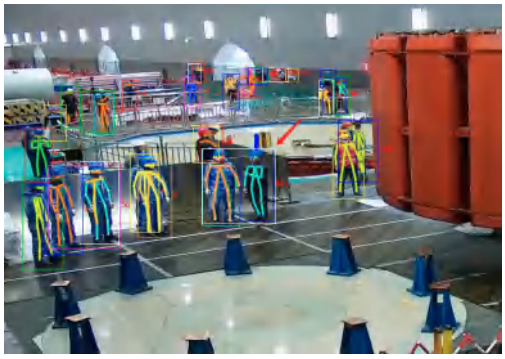
发表EI论文3篇，获软件著作权1项，并在HID2022国际步态识别比赛中荣获第三名。

系统兼容性

支持多品牌摄像头接入，具备RTSP、ONVIF等协议兼容能力，可快速对接现有安防系统。

可扩展性

提供API接口与数据看板，支持与MES、ERP、智慧工地平台等系统集成。



生态化开源工业软件开发框架

01 项目背景

当前，工业软件领域的计算机辅助设计、仿真、制造等技术（统称CAX技术）是高端制造业数字化转型的核心支撑，但国内CAX市场长期被国际主流商业软件（如ANSYS、CATIA、SiemensNX等）垄断，国产CAX软件在功能完整性、算法效率及生态协同方面存在显著差距——一方面，商业软件因涉及核心技术机密，仅开放有限二次开发接口，无法满足企业定制化需求，且存在“卡脖子”风险；另一方面，国内缺乏统一的开源开发工具、专业社区及标准化测试平台，导致CAX技术研发成果难以快速转化，开发者协作效率低下，进一步制约了国产CAX软件的迭代与创新。

从行业现状来看，全球CAX市场规模庞大（2022年CAD软件市场规模103.8亿美元、CAE软件91.63亿美元），但国内市场份额占比低且国产化率不足，多数高校、企业仍依赖国外软件。同时，国家“十四五”规划明确提出“支持数字技术开源社区发展”，深圳市“20+8”产业布局也将工业软件列为重点发展领域，亟需构建自主可控、生态化的CAX开源体系，填补国内在CAX开源工具链、社区协作及测试标准方面的空白。

02 创新内容

为突破国产CAX软件发展瓶颈，深圳市大数据研究院牵头研发“生态化开源工业软件开发框架（OpenCAXPlus）”，构建“开发工具包（SDK）-中文开源社区-标准化测试平台”三位一体的开源体系，核心创新点如下：

跨平台开源开发工具链创新

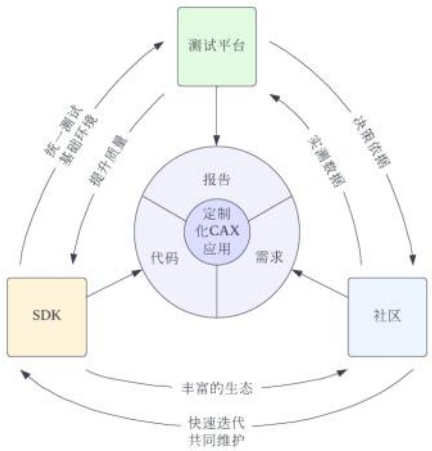
基于Golang开发OpenCAXPlusCLI工具；SDK框架中定义标准化数据/函数/文件接口，支持Windows、Linux等多操作系统及国产芯片适配；通过CMake实现跨平台自动编译，集成30余个开源软件开发包，提供“项目初始化-依赖安装-编译打包-测试发布”全流程工具支持，大幅降低CAX软件开发门槛。

专业化中文开源社区架构设计

采用Vue3+Gin前后端分离架构，打造集“论坛交流-项目展示-知识沉淀-会员管理”于一体的CAX+中文社区，支持敏感词过滤、权限分级管理及全局搜索功能；创新设计“SDK-社区-测试平台”联动机制，开发者可通过 SDK 命令行工具直接登录社区发布项目、提交测试任务，实现研发协作闭环。

容器化标准化测试平台构建

基于Linux集群与容器化技术，搭建支持跨平台性能评估的CAX测试平台，计划建立开放标准算例库（覆盖结构分析、流体仿真等常用工业场景）；通过IPHash负载均衡解决集群会话保持问题，实现用户应用权限控制与资源隔离，为CAX软件提供“自动化测试-性能统计-结果可视化”全流程服务，提升代码质量与用户信任度。

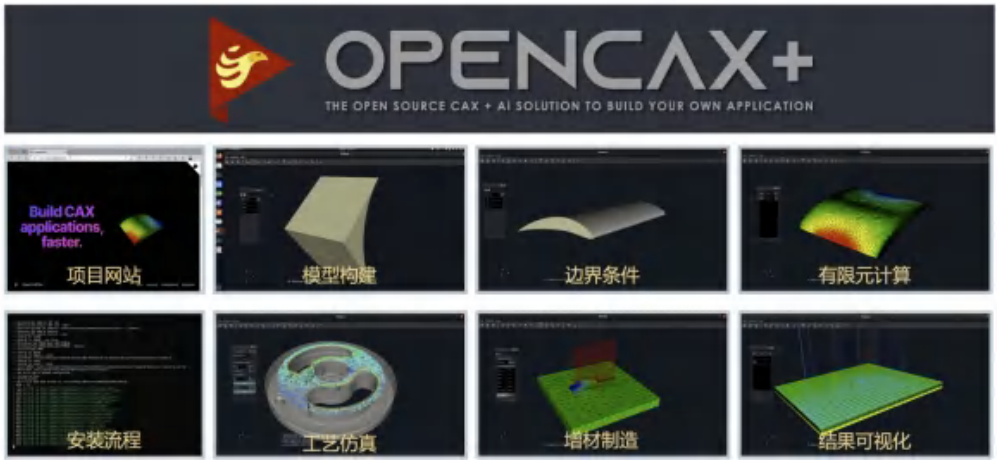


03 应用场景

高校 / 科研机构 CAX 技术研发	中小企业工业软件定制	国产工业软件生态建设
为材料、机械、航空航天等领域的科研团队提供开源开发工具与测试平台，支持快速验证新型算法（如多物理场耦合、AI驱动仿真等），加速科研成果转化。	降低中小企业定制化仿真开发的成本，通过社区共享插件与模板项目，快速开发符合行业需求的定制化软件（如零部件设计工具、产线仿真系统等）。	为国产CAX软件厂商提供开源生态支撑，助力厂商基于OpenCAXPlus SDK拓展功能模块，推动“自主可控CAX软件”从技术研发到产业落地的全链条创新。

04 重要成果

开发工具成果：实现Ubuntu系统“一键安装”，集成Gmsh、OpenCascade等30余款核心组件，支持复合材料设计、油藏模拟等场景，已取得一项软件著作权。计划在2026年正式发布V1版本。



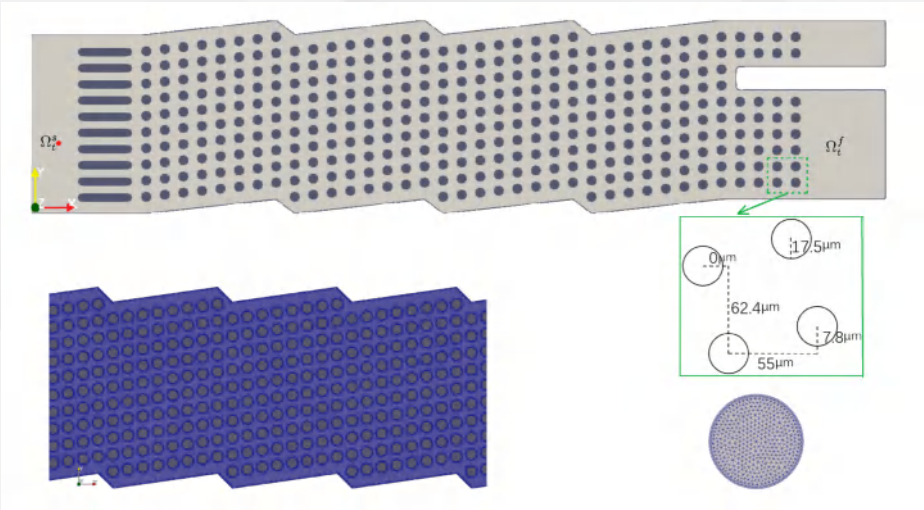
社区与测试平台成果：依靠其他开源社区建设项目和相关算力后台管理系统的开发经验，已完成开源社区和测试平台的初步设计，计划在2027年初正式上线。

低雷诺数流动的区域形状稳健迭代求解器

01 项目背景

低雷诺数流动（如斯托克斯流动）的数值模拟是流体力学关键课题，其控制方程为斯托克斯方程，需满足入口速度固定、出口应力平衡、壁面与孔边界速度为零的边界条件，且采用Taylor-Hood元（速度P_2阶、压力P_1阶）离散。

传统求解方法存在明显缺陷：离散后方程组呈鞍点结构，“质量矩阵近似Schur补”求解器在复杂域形（如含大量孔/柱体区域）中稳定性差；几何多重网格法在粗网格层难以处理大规模系统，亟需兼顾域形稳健性与计算效率的迭代求解器。



02 项目简介

本项目针对低雷诺数流动求解难题，提出几何-代数多重网格求解器（Geo-Algb-Multigrid）与全代数多重网格求解器两类方案。核心思路是将斯托克斯方程离散为块鞍点系统，通过Schur补矩阵近似处理关键计算瓶颈，结合多重网格层级策略，分别依托几何算子或全代数算子实现高效迭代求解，规避传统方法在复杂域形与大规模计算中的局限。

03 创新内容

几何-代数融合多重网格架构

设计Geo-Algb-Multigrid求解器，采用分层处理策略：细网格层用质量矩阵近似Schur补作松弛方法，结合几何算子传递速度、压力信息；粗网格层通过最小二乘近似处理大规模近达西系统，同时支持区域分解法（DDM）作为细网格松弛方法，简化实现且保障耗时性能。

全代数多重网格求解器

为摆脱几何网格依赖，提出全代数方案：通过递归构造代数算子替代几何算子，生成粗网格算子与延拓、限制算子，细网格层用ILU作松弛方法，适配非结构化或难几何加密的网格场景，提升应用灵活性。

高效迭代框架

两类求解器均采用V循环迭代流程：当前网格层松弛迭代→残差传递至粗网格层→粗网格求解误差并回传修正→再次松弛优化，确保迭代次数稳定，且支持并行计算，适配大规模模拟。

04 应用场景

多孔结构流动模拟	复杂域形流动分析
适用于微流控芯片、多孔介质过滤等含大量孔/柱体的场景，可高效模拟流体在复杂多孔区域的流动，输出速度场与压力场，为器件设计提供支撑。	能处理拉伸域、不规则边界域等场景，如工业管道局部变形段、生物血管网络（低雷诺数血流），解决传统方法稳定性问题。

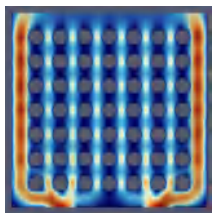
大规模并行计算
支持多处理器并行（最高测试64核），可应用于大型多孔反应器等工程场景，适配大规模网格计算需求。

05 重要成果

2D场景性能优势

强可扩展性：固定24×24孔与网格规模，处理器数从1增至64时，迭代求解器首次求解时间从56.55s降至2.97s，迭代次数稳定，远超直接求解器效率。

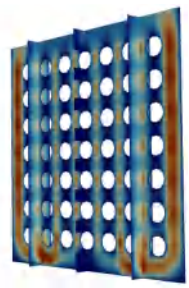
弱可扩展性：孔数量从6×6增至48×48，64核下迭代求解器首次求解时间仅从4.78s增至9.54s，直接求解器则增至63.10s，且大尺寸场景下直接求解器失效。



3D场景性能验证

强可扩展性：固定10×10柱体，处理器数从4增至64时，迭代求解器首次求解时间从43.34s降至6.36s，优势显著。

弱可扩展性：柱体数量从3×3增至32×32，64核下迭代求解器仍能稳定求解，直接求解器在大尺寸场景失效。



CUSTOMER CASES
客户案例

国家电网新源集团大型吊装作业风险管控

AI智慧安监 | 多模态风险识别 | 边缘云协同架构 | 跨行业实时预警 | 低代码配置平台

01 项目背景

国家电网新源集团（全称：国家电网新源控股有限公司）是国家电网公司旗下的核心专业子公司，成立于1994年，专注于抽水蓄能电站的规划、投资、建设与运营管理，是中国最大的抽水蓄能开发运营商。管理全国超40座大型抽水蓄能电站，总装机容量逾3000万千瓦，占全国抽水蓄能总装机容量的80%以上。单个项目现场日均作业人员超200人，涉及变电站检修、大型机组吊装、地下洞室施工等高风险场景，日均高危作业点位50+，覆盖山区、峡谷等复杂地理环境。

02 行业问题

大型作业现场普遍存在以下痛点，导致安全隐患频发、管理成本高企：

监管覆盖不足

作业现场环境复杂（如高空、有限空间），风险点分散（单现场超100个），人力巡检仅能覆盖30%~40%区域，盲区易引发事故。

响应效率低下

传统方式依赖人工上报，隐患识别延迟超30分钟，难以应对动态风险（如人员闯入危险区、设备异常）。

技术适配性差

单一摄像头监控无法识别复合行为（如未系安全带+高空作业），跨场景规则配置繁琐。

成本压力突出

人力监管成本占安全预算40%以上，且事故率年均增长5%，亟需降本增效。

行业数据佐证：据应急管理部统计，2022年大型作业现场事故中，70%源于监管延迟或覆盖遗漏。

·国家电网新源集团·

能源电力、大型基建

03 解决方案

部署“智慧安监系统”，基于AI与IoT构建“端-边-云”一体化平台：

端侧智能感知

接入多品牌摄像头（支持RTSP/ONVIF协议），通过边缘计算设备（如NVIDIA Jetson）实时运行优化模型（YOLOv5+DeepSORT），实现人员/设备/环境的毫秒级分析。

边云协同处理

边缘端处理低延迟任务（如安全帽识别、区域闯入预警），云端汇聚数据并训练模型（GaitSet步态识别增强遮挡处理），支持跨镜跟踪与大数据分析。

低代码配置平台

提供可视化界面，客户可自定义18+类风险规则（如“高空作业未系安全带”），10分钟内完成场景适配（如石化动火作业规则库）。

系统集成能力

通过API对接MES/ERP系统，实现告警自动推送至指挥中心，并联动智慧工地平台触发应急响应。

04 落地成效

技术成果

支持18类以上风险行为识别，平均准确率超**85%**

应用效益

实现实时告警和快速响应，提升安全管理效率**30%+**

预期效益

降低事故率，减少人工成本，提高合规性，并可通过**API**集成到现有系统

显式动力学新型结构化任意拉格朗日欧拉求解器研发

S-ALE 求解器 | 流固耦合 | 异构并行（CPU/GPU） | 大变形仿真 | 大规模网格计算

01 项目背景

在全球科技革命与产业变革背景下，高性能计算仿真软件是关键领域“工业母机”，直接关乎国家产业链安全与核心竞争力。当前，以LS-DYNA为代表的进口软件形成技术垄断，我国高端仿真高度依赖进口，面临成本高、技术封锁、数据安全等风险，成为产业升级瓶颈。

我国向“制造强国”转型，战略性产业对仿真软件需求激增，但国产软件在结构化ALE核心技术上存在短板。本项目研发对标LS-DYNA的自主可控新型结构化网格ALE软件，将打破国际垄断，填补技术空白，为高端制造业提供高效精准工具，带动产业链协同创新，强化我国在该领域的国际话语权，为国家战略安全提供技术支撑。

02 行业问题

在我国高端制造业向“制造强国”转型的关键阶段，航空航天、国防、汽车等战略性产业对大变形、流固耦合类仿真需求日益迫切，但行业面临多重核心痛点：

技术依赖与自主化短板

国际主流软件（如LS-DYNA）垄断S-ALE核心技术，国内多数企业缺乏自主可控的求解工具，长期依赖进口导致研发成本高、数据安全存忧，且关键场景面临技术封锁，制约产业自主创新。

性能瓶颈难以突破

传统ALE算法无法满足千万级以上单元的大规模仿真需求，求解规模仅停留在百万级别；高速极端工况下易出现单元负体积、时间步长为零等稳定性问题，狭缝、薄壁结构仿真中物质泄露频发，无法支撑高端装备精准设计。

效率与精度失衡

现有技术并行运算效率低，难以适配上亿单元级分析场景；前处理流程复杂，网格数据处理逻辑不合理导致计算速度慢，且压力曲线峰值等关键指标精度与商业软件或实验结果偏差较大，影响产品研发迭代效率。

适配性与扩展性不足

缺乏CPU-GPU异构并行等多元计算方式，难以适配HPC集群、GPU服务器等不同计算平台；算法架构僵化，无法灵活应对油箱液体晃动、飞行器溅落、弹药穿透等多样化工程场景，难以满足高端制造业个性化、复杂化仿真需求。

·A公司·
工业软件（显式动力学仿真软件）

这些问题本质是核心算法与软件工具的技术鸿沟，不仅推高了高端制造业研发成本、延长了产品周期，更削弱了我国产业在全球市场的核心竞争力，成为制约我国高端制造业向高精尖突破的关键瓶颈。

03 解决方案

解决方案：创新突破核心技术难关

针对高端制造业大变形、流固耦合仿真的技术依赖与性能瓶颈，依托三大核心创新攻克关键难点：

创新混合网格架构

全局采用结构化网格提升计算效率，流固界面附近引入局部非结构化动态贴体网格，实现高精度匹配。

异构并行技术

研发CPU-GPU异构并行框架，集成负载均衡与精度优化，支撑数亿单元级仿真，突破效率与规模瓶颈。

核心算法优化

优化界面重构、流固耦合离散等算法，解决高速极端工况稳定性问题，提升仿真精度。

通过模块化分工与标准化集成，形成完整技术链条。

04 预期效益

赋能高端制造，强化核心竞争力

1.技术突破：填补国内S-ALE核心技术空白，构建自主可控的高性能仿真算法体系，打破进口软件垄断；

2.性能提升：求解规模**从百万级别跃升至数亿级别，计算效率提升20%-40%，典型工况精度偏差≤10%**；

3.产业赋能：为航空航天、国防、汽车等高端制造业提供高效仿真工具，降低研发成本、缩短迭代周期；

4.生态支撑：带动高性能计算、数值算法等领域创新，培养专业人才，强化我国高端工业软件国际竞争力。

深圳市大数据研究院
Shenzhen Research Institute of Big Data

WHITE PAPER ON
SCIENTIFIC RESEARCH ACHIEVEMENTS
科研成果白皮书

合作联系方式：0755-84273753

官网：https://www.sribd.com/

地址：深圳市龙岗区龙翔大道1001号