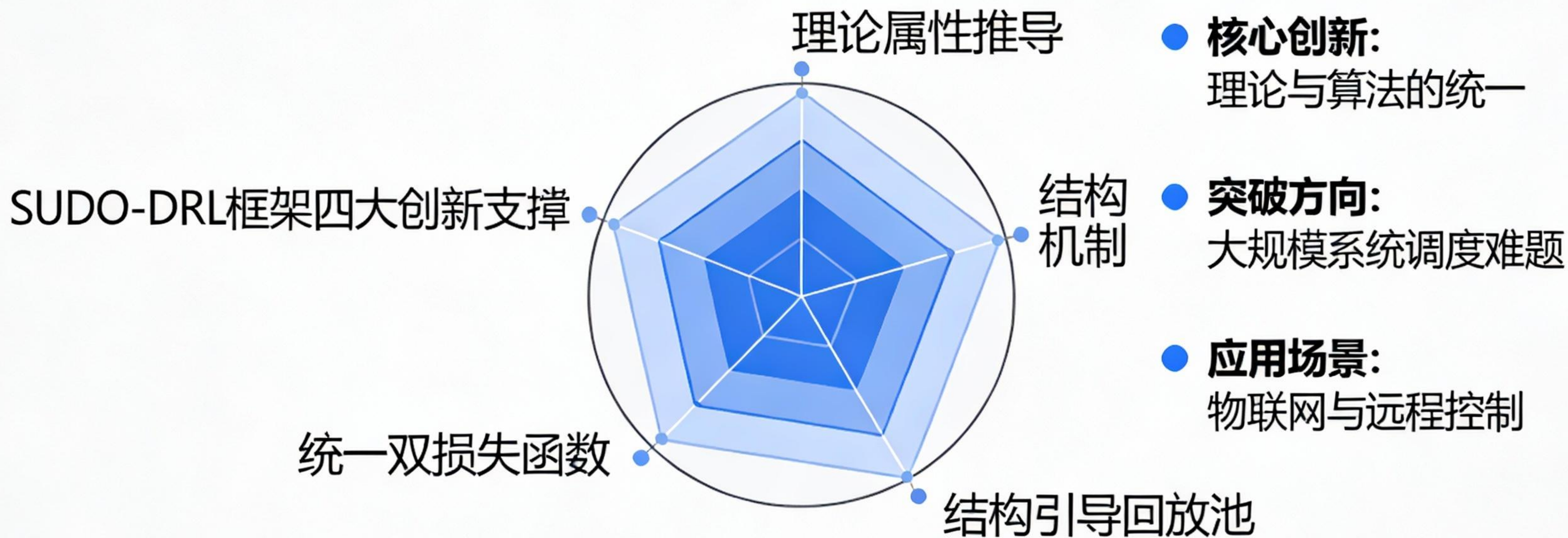


面向目标的传输调度： 基于结构引导的统一双策略深度强化学习方法



作者：陈佳政，刘万春*（通讯作者）

作者：陈佳政，刘万春*（通讯作者）

研究背景与问题定义: 面向目标的传输调度

MORANDI
16:9 校tland

在多设备多信道无线通信系统中， N 个边缘设备共享 M 个有限衰落信道 ($M < N$)。研究的核心问题是寻找最优调度策略 π ，以最小化所有设备的长期期望总成本。在远程状态估计场景下，这等同于最小化与信息年龄 (AoI) 相关的估计状态均方误差 (MSE)

关键挑战



1. 高维状态空间:
必须实时跟踪每个设备的AoI和信道状态，复杂度随设备数量呈指数级增长



2. 庞大的动作空间: 对于 N 设备 M 信道, 动作数为 $N! / (N-M)!$ 。在 10 设备 5 信道场景下, 动作数已达 30,240 个, 传统MDP求解器完全失效



3. 目标转向:
从单纯的传输比特转向满足特定应用需求 (Remote State Estimation)



现有方法的局限性与研究空白



MORANDI
10笔以内简文

1. 传统方法：启发式方法

(如Whittle's Index)

虽快但非最优；动态规划
(值迭代/策略迭代) 面临
“维度灾害”



2. 标准DRL算法：

- 离策 (Off-Policy) 如DQN/DDPG:

虽样本效率高，但在大型动
态系统中训练极不稳定，40
设备场景下无法收敛。

- 同策 (On-Policy)

如PPO/TRPO:

稳定性好但样本效率极低，数
据利用率差，易陷入局部最优



3. 理论结合缺失：现有研究多

采用“暴力优化”，未深入挖掘AoI
状态最优值函数的数学性质（如单
调性、凸性），导致在大规模系
统（20-40个传感器）中性能
严重下降或不收敛



4 SUDO-DRL框架概述: 理论与算法的统一

混合DRL框架



1. 理论属性推导:

证明V函数单调性与渐近凸性



2. 结构评分机制:

转化为CM (单调性)、CC (凸性)、AM (策略单调性) 评分



3. 统一双损失函数:

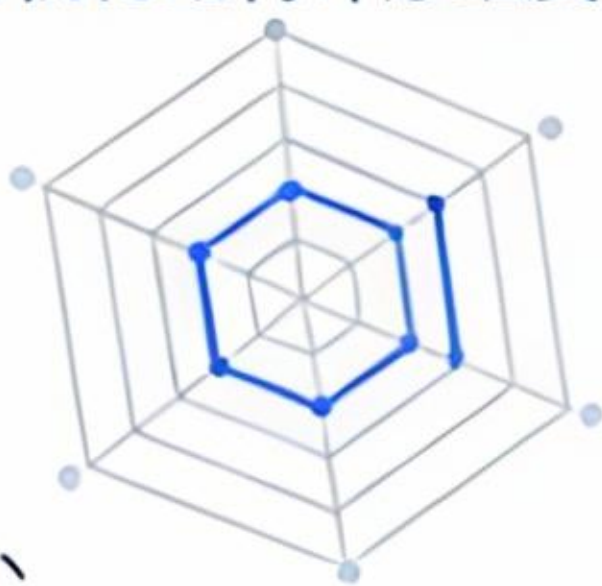
通过权重 β_0 , β_2 桥接在线与离线更新



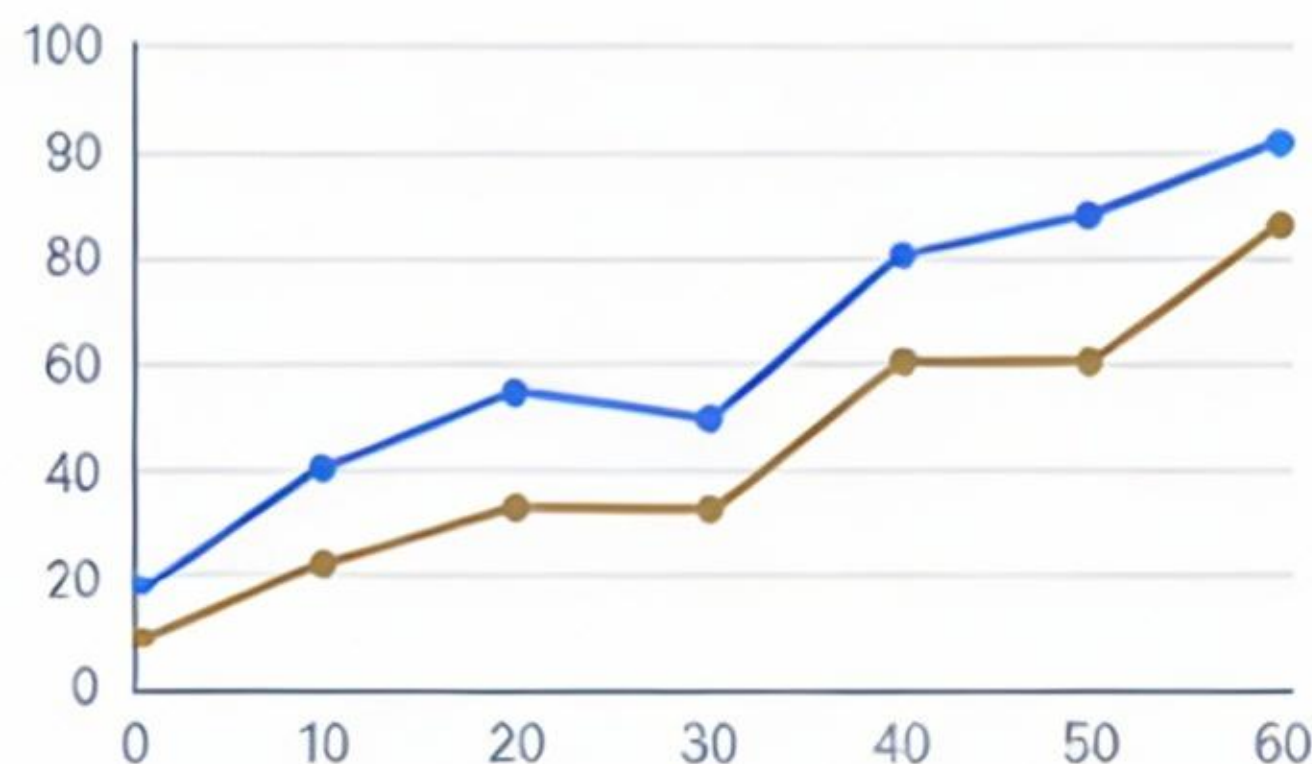
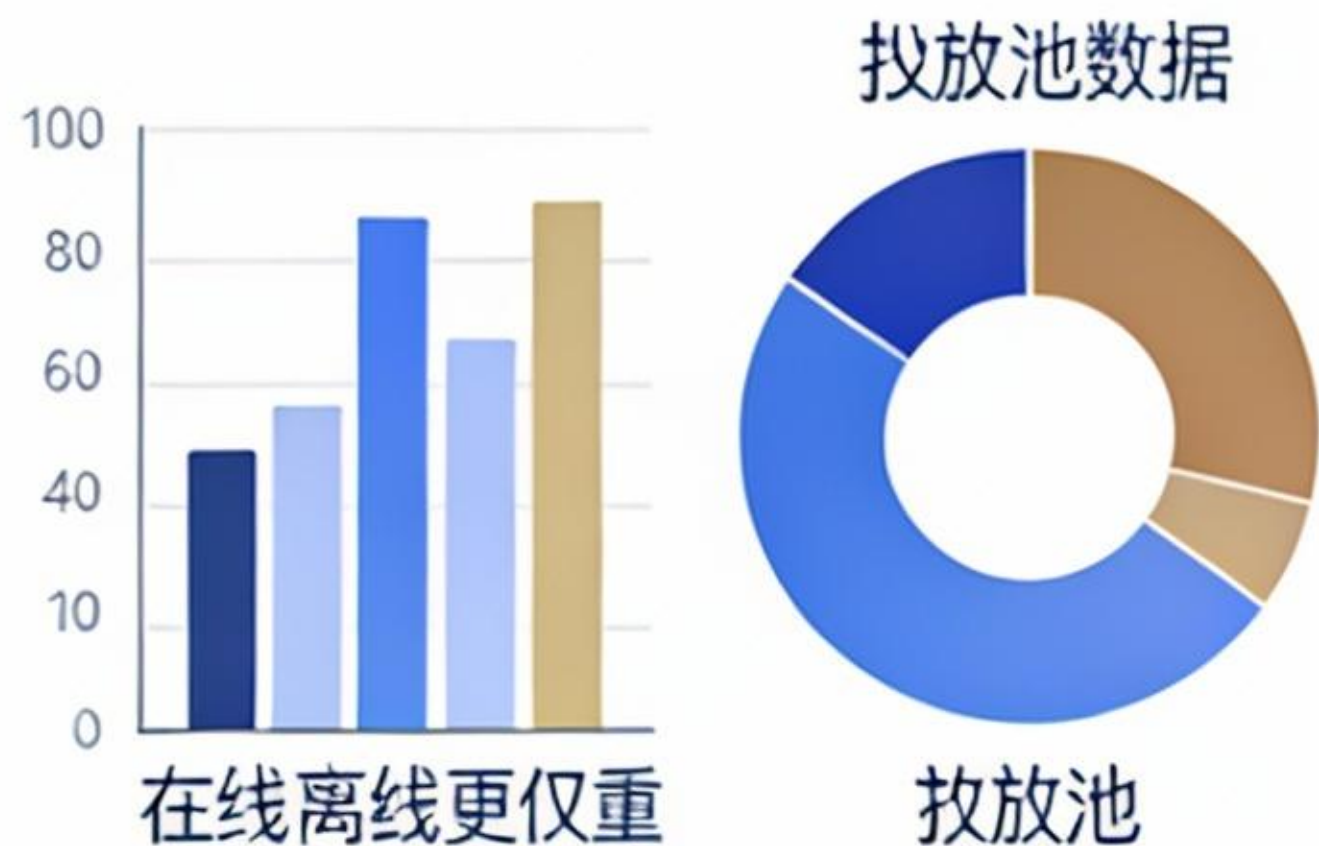
4. 结构引导回放池:

基于评分选择性存储与优先级采样

可视化结构评分维度



训练窗图





数学证明

1. V函数单调性：价值函数随AoI和信道状态增加单调递增
2. 渐近凸性：传输调度领域首次证明V函数随AoI状态渐近凸性
3. 渐近凸性：传输调度领域首次证明V函数随AoI状态渐近凸性
3. 策略单调性：最优策略关于信道状态具单调性
4. 渐近贪婪结构：共址设备AoI极大时最优策略倾向强制调度（为预训练提供理论依据）

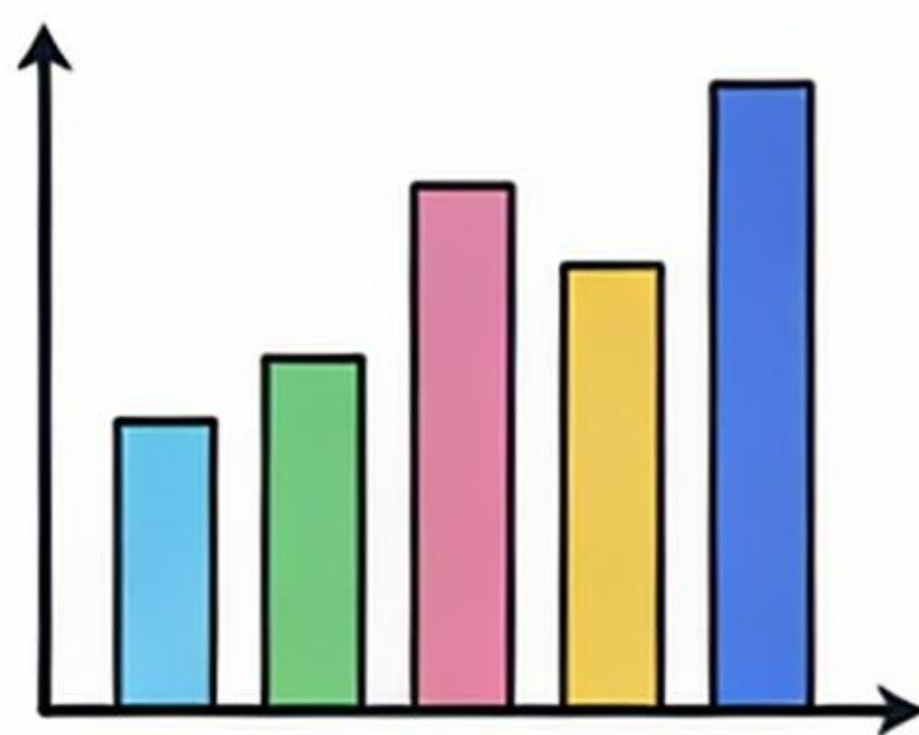
$$\min_{\pi} \lim_{T \rightarrow \infty} \mathbb{E}^{\pi} \left[\sum_{t=1}^T \sum_{n=1}^N \gamma^t c_n(\delta_{n,t}) \right]$$

其中 $\delta_{n,t}$ 为设备n在时间t的AoI

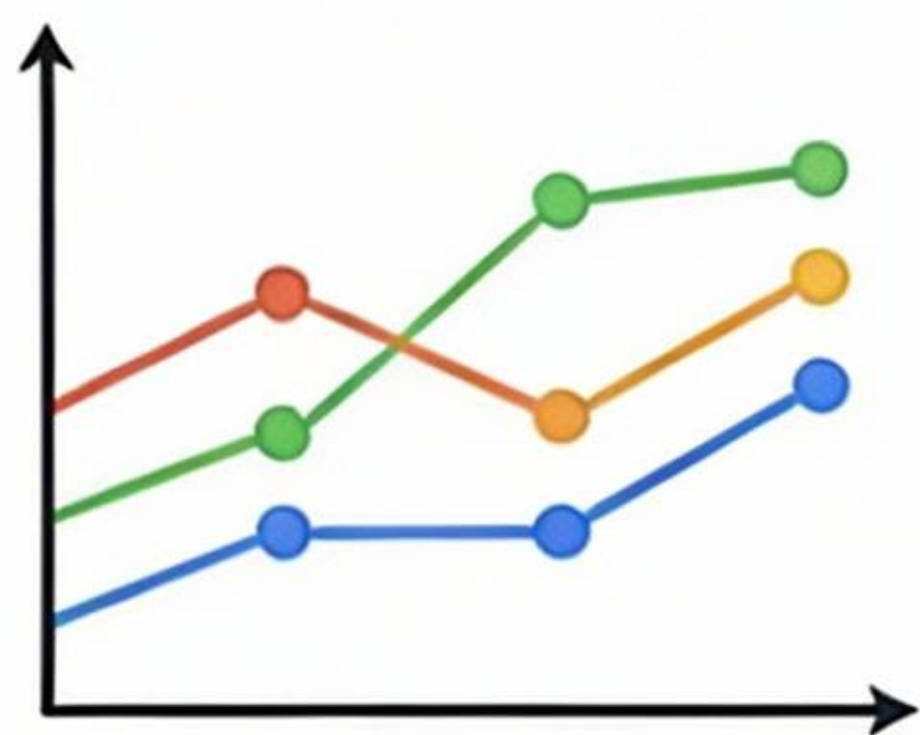


1. CM Score (Critic-Monotonicity) : 评估Critic网络预测的V值是否随AoI增加。
若违反, 则产生惩罚项:

$$\hat{V}_{AoI} = \max(0, v(\mathbf{s}; \boldsymbol{\nu}) - v(\hat{\mathbf{s}}_{(n)}; \boldsymbol{\nu}))$$



+



+



+



2. CC Score (Critic-Convexity) : 评估V函数的二阶差分是否符合凸性:

$$\check{V}_{AoI} = \max(0, 2v(\mathbf{s}; \boldsymbol{\nu}) - (v(\hat{\mathbf{s}}_{(n)}; \boldsymbol{\nu}) + v(\hat{\mathbf{s}}_{(n)}; \boldsymbol{\nu})))$$



3. AM Score (Actor-Monotonicity) :
评估Actor网络的动作选择是否符合信道单调性逻辑

Critic统一损失



$$L_{\text{SUDO}}(\boldsymbol{\nu}) = L_{\text{On}}(\boldsymbol{\nu}) + \beta_1 L_{\text{Off}}(\boldsymbol{\nu})$$

+

+ Structural Penalties

L_{On} 处理当前轨迹TD误差,

L_{Off} 处理回放池历史数据TD误差

Actor统一损失



$$L_{\text{SUDO}}(\boldsymbol{\varphi}) = L_{\text{On}}(\boldsymbol{\varphi}) + \beta_2 L_{\text{Off}}(\boldsymbol{\varphi})$$

结合PPO裁剪损失 (稳定性)

与SAC风格离线策略梯度 (效率)

● 回放池管理与结构引导预训练

MORANDI

选择性存储：系统仅存储结构评分

(CM,CC,AM) 超过历史平均水平的轨迹，
确保回放池充满“高质量”样本



优先级采样：采样概率

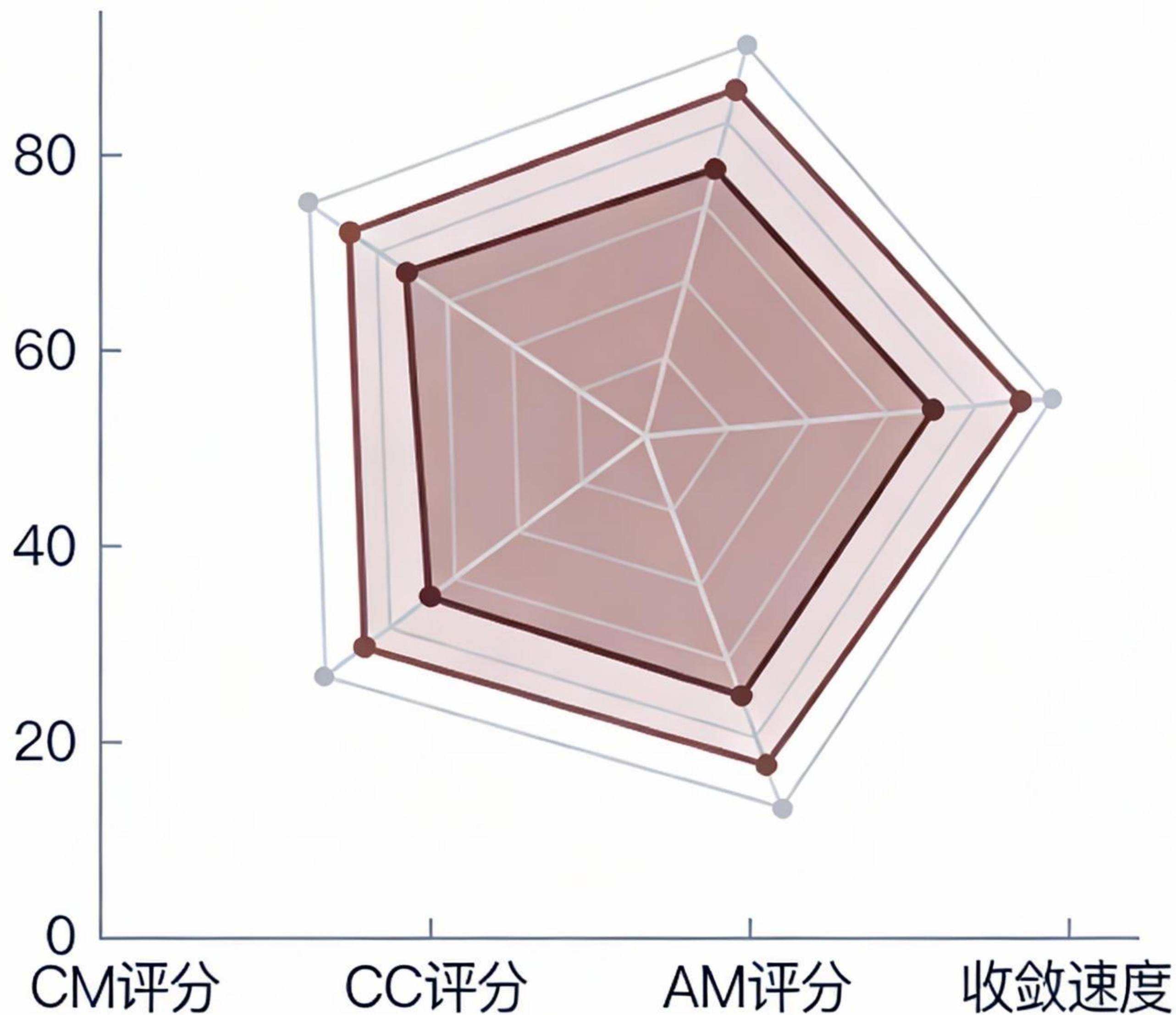
$$P_b \triangleq \frac{p_b \cdot \rho^b}{\sum (p_b \cdot \rho^b)}$$



渐近贪婪预训练：训练初期（前
10 × N 回合），利用定理5贪婪结构指导
动作选择，收敛速度提升约40%



100分



● 实验设置：远程状态估计仿真

MORANDI

仿真环境基于远程状态估计系统：规模覆盖从10设备5信道扩展至40设备20信道（超大规模）；信道模型为i.i.d.

信道模型为 i.d.块衰落信道，量化为5级状态，丢包率范围0.01至0.2；

神经网络中Actor和Critic均为3层隐藏层的全连接网络；

训练参数为10,000训练回合，每连接网络；

训练参数为10,000训练回合，

每回合128步， $\gamma=0.99$ ，

学习率Critic=0.001， $\gamma=0.99$ ，

学习率Critic=0.001、

Actor=0.0001；

硬件支持NVIDIA RTX 3060Ti GPU；

评估指标侧重平均总MSE成本
和收敛速度



Hyperparameters	Unified loss weight, ϵ	Value
Discount factor, γ	0.99	0.2
Clipping parameter, ϵ	β_1, β_2	0.9
Replay buffer, R	0.9	200
Pre-training episodes	$10 \times$	$\times N$





1. 小规模场景(10, 5): SUDO-DRL成本为85.52，接近专门优化的SE-DDPG(77.14)，远优于PPO(119.63)



2. 大规模场景(30, 15)及以上:
所有离策算法 (DDPG变体)均宣告失败(无法收敛)。
SUDO-DRL不仅成功收敛，且性能比PPO提升了30%-40%



3. 鲁棒性：在不同参数设置下(Para. 1-16)，SUDO-DRL始终保持最低的MSE成本，证明了其在复杂、动态无线环境下的稳健性

跨规模性能比较			
Scale (N, M)	PPO	SUDO-DRL	DDPG Variants
(10, 5)	119.63	85.52	77.14-89.26
(20, 10)	569.96	370.63	357.14 (Partial Fail)
(30, 15)	900.71	518.03	Failed (—)
(40, 20)	1291.54	994.80	Failed (—)

1. 结构引导的作用:

SUDO-DRL在200回合内达到100分凸性评分 (CC Score), PPO徘徊80分以下, 结构引导强制模型学习正确物理规律



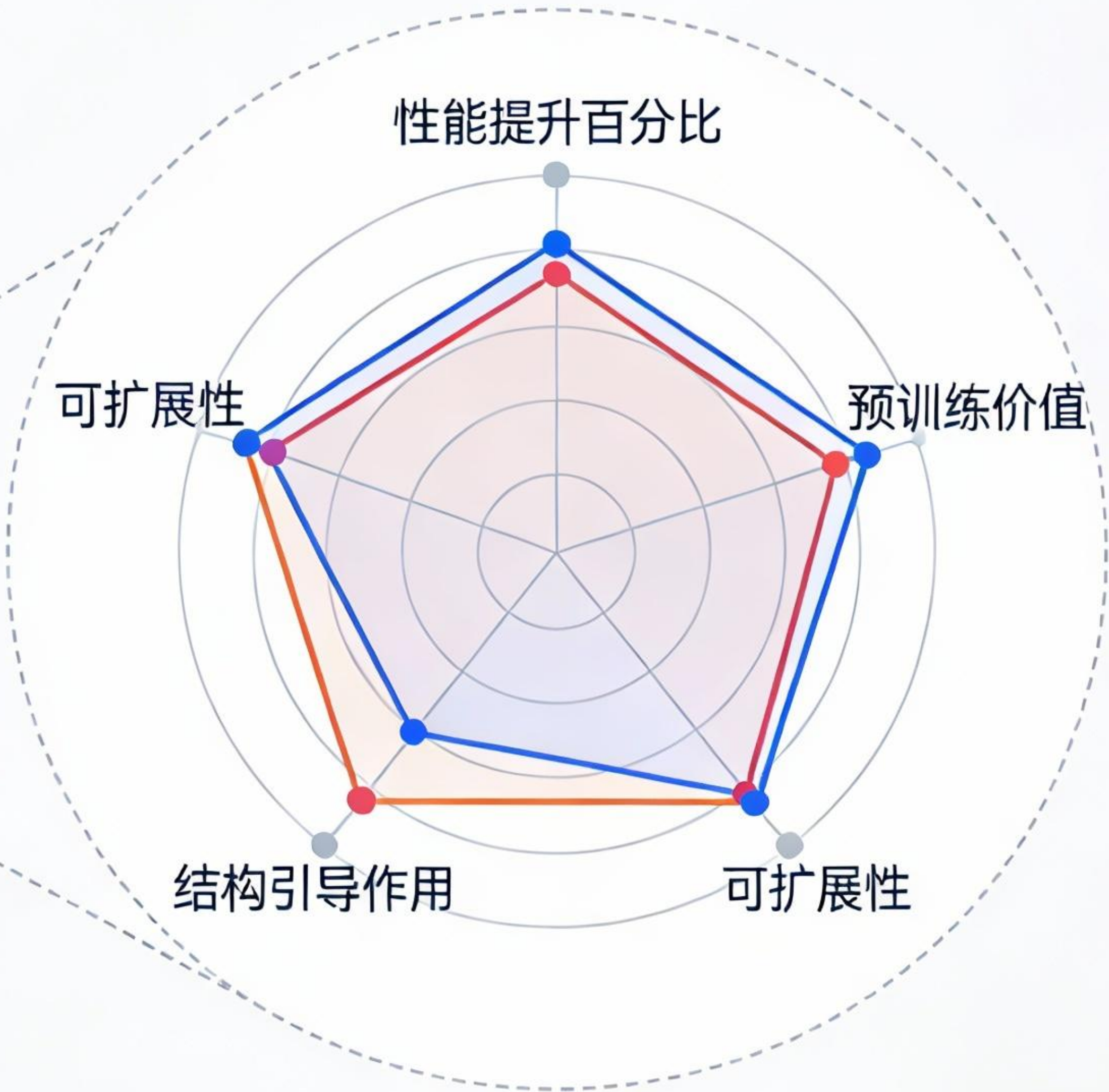
2. 预训练的价值:

全功能版本比无预训练版本收敛速度快40%, 预训练使初始MSE成本从800+降至700以下




3. 可扩展性突破:

SUDO-DRL是首个在40设备20信道规模下保持高性能且稳定收敛的DRL调度算法, 填补大规模目标导向通信技术空白



消融实验结果对比

结论与贡献总结

1. **理论突破：**首次证明AoI调度中价值函数的渐近凸性，利用单调性构建约束学习框架
2. **算法创新：**提出统一双策架构，兼顾PPO稳定性与SAC效率 
3. **性能飞跃：**相比PPO，系统性能提升25%-45%，收敛速度提升40%
4. **大规模适用性：**解决40台设备场景维度灾难，实现理论最优与计算可行平衡

未来研究：探索框架在多跳网络或异构业务流中的应用

