中国科学 : 信息科学 2026 年 第 56 卷 第 10 期: 2608–2623
SCIENTIA SINICA Informationis
面向 6G 的通感算可重构无线网络专题 • 论文

基于世界模型辅助强化学习的 RIS 自适应波束对准方法设计

张宗灏[1], 幸锋[1,2], 王梓帆[1], 秦浩文[1], 徐澍[3,1], 李春国[1*], 杨绿溪[1]

1. 东南大学移动通信全国重点实验室, 南京 210096
2. 中国移动通信集团云南有限公司融合创新中心, 昆明 650100
3. 南京理工大学电子工程与光电技术学院, 南京 210094
* 通信作者. E-mail: chunguoli@seu.edu.cn

收稿日期: 2026–04–22; 修回日期: 2026–08–13; 接受日期: 2026–09–08; 网络出版日期: 2026–10–12

国家自然科学基金 (批准号: U25B2007, 6260010313)、国家重点研发项目 (批准号: 2024YFC3807900)、中央高校基本科研业务费专项资金 (批准号: 30926010527) 和江苏前沿引领基础研究重大项目 (批准号: BK20222001) 资助

摘要 针对可重构智能表面 (reconfigurable intelligent surface, RIS) 可主动重构无线传播环境、提升毫米波通信覆盖范围和链路可靠性的特点, 本文研究非完美信道状态信息 (channel state information, CSI) 下的大规模 RIS 自适应波束对准问题. 然而, 现有方法通常依赖预定义码本搜索或在线强化学习探索, 在大规模 RIS 场景下面临导频开销高、冷启动探索慢和二维相位变量耦合强等问题. 因此, 本文提出一种基于世界模型先验引导与深度强化学习在线微调的两阶段波束对准方法 (world-model-assisted twin delayed deep deterministic policy gradient, WM-TD3). 首先, 利用 RIS 级联信道的空间结构特性, 构建水平与垂直维度独立波束搜索模型, 将高维相位优化问题转化为低维优化问题. 其次, 为缓解纯强化学习缺乏先验导致的随机探索和初期不稳定问题, 设计认知世界模型从宽波束幅相观测中学习潜在信道状态, 并输出初始波束先验和奖励估计基线. 进一步地, 构建解耦双智能体 (decoupled dual-agent TD3, DDA-TD3) 微调机制, 由水平和垂直智能体分别优化对应相位变量, 并通过共享差分奖励围绕端到端信噪比 (signal-to-noise ratio, SNR) 目标协同更新. 仿真结果表明, 与分层搜索和纯 TD3 方法相比, 所提方法能够以较低搜索开销实现更快波束对准和更稳定的接收 SNR 性能.

关键词 可重构智能表面, 毫米波通信, 波束对准, 世界模型, 深度强化学习, 双延迟深度确定性策略梯度 (TD3)

1 引言

随着移动通信设备的快速增长, 复杂的通信环境和海量连接需求要求多种通信技术深度融合以提升系统性能, 这使得易于与现有系统结合的新型技术成为了第六代 (6th generation, 6G) 及未来无线通信的主要研究方向 [1].

为支撑 6G 愿景下海量业务对超高吞吐量的需求, 向毫米波 (millimeter wave, mmWave) 频段拓展频谱资源已成为提升系统容量的核心技术路径 [2], 毫米波通信具备极宽的可用带宽, 能够显著提升数据传输速率. 然而,

引用格式: 张宗灏, 幸锋, 王梓帆, 等. 基于世界模型辅助强化学习的 RIS 自适应波束对准方法设计. 中国科学: 信息科学, 2026, 56: 2608–2623, doi: 10.1360/SSI-2026-0195
Zhang Z H, Xing F, Wang Z F, et al. Design of a world-model-assisted reinforcement learning method for RIS adaptive beam alignment. Sci Sin Inform, 2026, 56: 2608–2623, doi: 10.1360/SSI-2026-0195

© 2026《中国科学》杂志社

在实际部署中, 毫米波固有的高频特性使其面临严峻的自由空间路径损耗. 为弥补这种严重的传播衰落, 系统通常依赖大规模天线阵列进行定向波束成形, 通过窄波束集中能量以提升链路质量. 与此同时, 毫米波的衍射能力较弱, 对物理障碍物的遮挡极为敏感, 极易引发视距 (line of sight, LoS) 链路中断, 导致通信稳定性显著下降.

为实现波束对准, 现有毫米波波束训练方法通常依托预定义离散码本对空间角域进行定向扫描, 以确定收发端初始波束成形方向 [3]. 在高频大规模阵列系统中, 码本设计与信道估计仍是降低训练开销的重要方向. 文献 [4] 针对太赫兹超大规模多输入多输出 (multiple-input multiple-output, MIMO) 系统近场宽带场景研究了蜂窝码本设计与信道估计问题, 表明随着阵列规模和带宽增加, 码本搜索需进一步考虑近场效应和宽带特性. 为避免全维穷举扫描带来的高导频时延, 多分辨率分层码本、自适应训练序列和子阵列并发多波束等方法被相继提出 [5,6], 在训练开销与链路增益之间取得了一定折中.

上述基于码本设计的波束训练方法虽取得了显著进展, 但仅靠收发端优化仍无法从根本上解决毫米波因物理遮挡而导致通信中断的问题. 在高频通信中, 路径损耗、物理遮挡和用户移动都会增加波束训练与追踪难度. 文献 [7] 针对毫米波低空无人机场景讨论了通感一体化波束训练与追踪问题, 也说明高动态场景下快速波束对准对链路可靠性具有重要意义. 近年来, 可重构智能超表面 (reconfigurable intelligent surface, RIS) 因能够主动调控无线传播环境而受到广泛关注 [8]. RIS 通过调整大量无源反射单元的相位响应, 可为被遮挡用户建立虚拟反射链路, 从而提升毫米波通信的覆盖范围和链路质量 [9]. 但 RIS 相位配置通常受到离散相位约束、级联信道耦合和非凸目标函数共同影响, 优化难度较高 [10].

现有 RIS 相位优化方法多以理想信道状态信息 (channel state information, CSI) 为前提. 在 RIS 辅助无线供电通信网络等场景中, 文献 [11,12] 分别通过分式规划、连续凸近似与半正定松弛等经典优化工具, 实现了资源分配与能效提升. 然而, 实际系统中的信道估计 (channel estimation, CE) 开销、噪声干扰和硬件非理想性使完美 CSI 难以获得; 同时, 大规模 RIS 相位配置具有高维、离散和非凸特性, 在缺乏精确 CSI 时传统穷举搜索复杂度高、训练开销大.

针对上述问题, 现有研究提出了多种降低训练开销的机制 [13,14]. 文献 [13] 采用半无源反馈架构, 通过局部相位信息辅助获取加速波束锁定; 文献 [14] 则基于方向图合成构建多分辨率分级码本, 以缓解全维穷举搜索开销. 然而, 此类方法仍主要依赖静态码本或预设训练结构, 在动态遮挡、低信噪比 (signal-to-noise ratio, SNR) 和复杂多径环境下适应性有限; 早期层级误判和并发波束相干误差还可能进一步导致波束对准性能下降.

为克服静态码本适应性不足的问题, 数据驱动方法为波束对准提供了新的思路. 近年来, 深度学习 (deep learning, DL) 已广泛用于阵列天线自适应波束形成、CSI 反馈压缩和波束预测等任务 [15]. 现有 DL 方法通常利用离线数据训练神经网络, 实现端到端波束预测或分级波束选择 [16~19]. 与此不同, 深度强化学习 (deep reinforcement learning, DRL) 方法将波束搜索或反射相位配置建模为智能体与无线环境之间的在线序贯决策过程, 通过接收功率、SNR 或速率等反馈信号持续更新策略. 相关研究已将强化学习 (reinforcement learning, RL) 用于毫米波 MIMO 波束码本学习、RIS 辅助毫米波波束成形码本设计、毫米波初始波束对准以及 RIS 辅助鲁棒波束成形等问题 [20~23], 验证了 DRL 在非理想信道和动态环境下的自适应优化潜力.

此外, 多智能体强化学习也被用于 RIS 辅助复杂网络的联合优化. 文献 [24] 采用多智能体近端策略优化算法 (multi-agent proximal policy optimization, MAPPO) 联合优化多 RIS 辅助车联网边缘计算网络中的边缘服务选择、RIS 部署位置和反射矩阵配置. 研究表明, 强化学习能够为 RIS 辅助通信中的非凸优化问题提供数据驱动求解思路. 然而, 现有工作多侧重码本学习、波束成形或网络资源联合优化, 对大规模 RIS 波束对准中的冷启动探索开销和二维相位耦合问题仍缺乏专门处理. 在无精确 CSI 条件下, 缺乏先验的智能体往往需要在高维相位空间中大量随机探索, 导致训练开销高、初期性能不稳定和收敛速度慢 [25]. 因此, 低导频开销下兼具先验引导、维度解耦和在线自适应能力的 RIS 波束对准策略仍有待研究.

从实现的过程机理来看, 基于码本的分层搜索方法先利用宽波束在大角域内粗定位, 再逐层展开较窄波束完成精细搜索. 该类方法结构清晰、硬件实现简单且训练开销可由码本层数预先控制, 但其搜索路径由静态码本决定, 对复杂多径干涉的适应性有限, 且早期层级的误判会在后续窄波束搜索中持续放大. 纯强化学习方法则将波束对准建模为智能体与无线环境之间的序贯决策过程. 深度强化学习已被应用于不确定环境下的复杂序贯决

策优化问题 [26], 从而在动态状态空间中实现自适应决策. 然而, 在面对大规模 RIS 场景中, 智能体在缺乏先验时往往需要长时间随机探索, 容易出现冷启动、样本效率低和收敛不稳定等问题.

针对上述的难题, 本文将 “实现低开销波束扫描” 的工程难题转化为未知 CSI 前提下最大化平均接收性能的数学优化问题, 并创新性地提出了一种 “基于世界模型先验引导的在线强化学习微调” 的两阶段自适应波束对准算法. 本文的主要贡献总结如下.

(1) 构建了严格的 RIS 级联空间解耦模型: 针对毫米波 Sale-Valenzuela (S-V) 多径模型, 本文将用户端的精确接收信号解耦为水平波束增益与垂直波束增益的克罗内克积, 将高维波束搜索问题转化为低维结构可分离优化问题.

(2) 提出了基于非对称双流残差网络 (residual network, ResNet) 的认知世界模型: 为缓解纯 DRL 冷启动探索开销大以及端到端波束预测方法泛化能力受限的问题, 本文构建了具备物理感知能力的认知世界模型. 该模型通过宽波束幅相观测分别提取能量与相位特征, 并利用多任务约束学习潜在信道状态, 从而为在线强化学习阶段提供初始波束和奖励先验.

(3) 设计了解耦双智能体 (decoupled dual-agent twin delayed deep deterministic policy gradient, DDA-TD3) 协同微调算法: 基于物理空间的解耦特性, 本文提出了 DDA-TD3 架构. 水平与垂直维度的智能体并行输出连续动作进行局部微调, 并基于共享的整体 SNR 增益进行协同更新, 彻底打破了高维动作空间带来的维度灾难, 实现了在时变非凸信道中的快速自适应锁定.

本文剩余部分的结构安排如下: 第 2 节介绍了包含级联信道解耦推导在内的系统与信道模型, 并进行了优化问题建模; 第 3 节详细阐述了所提出的认知世界模型先验引导与 DDA-TD3 协同微调算法; 第 4 节给出了相应的仿真实验与结果分析; 最后, 第 5 节对全文进行了总结.

注意. 在本文中, 标量用普通斜体字母表示, 列向量和矩阵分别用粗体小写和粗体大写字母表示. $\mathbb{C}^{M\times N}$ 表示 $M\times N$ 维的复数矩阵空间. $(\cdot)^{\mathrm{T}}$ 和 $(\cdot)^{\mathrm{H}}$ 分别表示矩阵的转置和共轭转置. $\otimes$ 表示克罗内克积 (Kronecker product). $\mathrm{Diag}\,(x)$ 表示以向量的元素为主对角线元素的对角矩阵. $\|x\|_2$ 表示向量 x 的 ℓ_2 范数. $\angle x$ 表示复数 x 的相位. $\mathbb{E}\,(\cdot)$ 表示统计数学期望.

2　系统模型与数学建模

2.1　系统模型

本文考虑如图 1 所示的由基站 (base station, BS)、RIS 和单天线移动用户设备 (user equipment, UE) 组成的 RIS 辅助的下行毫米波多输入单输出 (multiple-input single-output, MISO) 通信系统. BS 与 UE 之间的直达链路因物理障碍物遮挡而中断, 用户侧接收信号主要经由 BS-RIS-UE 级联反射链路获得. 其中, BS-RIS 链路由信道矩阵 $\mathbf{G}$ 表示, RIS-UE 链路由信道向量 $\mathbf{h}$ 表示. RIS 控制器根据所提世界模型先验引导与深度强化学习在线微调的两阶段波束对准方法 (world-model-assisted TD3, WM-TD3) 生成反射相位配置, 并通过调节 RIS 单元相位实现二维反射波束对准.

该系统主要配备 M 根天线的 BS, 一个具有 N 个反射单元的无源 RIS, 以及一个单天线移动用户. 其中 BS 采用均匀线性阵列 (uniform linear array, ULA); RIS 在布局上由 $N = N_x \times N_y$ 个单元均匀平面阵列 (uniform planar array, UPA) 构成, 其中 N_x 和 N_y 分别为水平方向和垂直方向的单元数量, 单元间距设为半波长 $\frac{\lambda}{2}$. 如图 1 所示, BS 与用户之间的直达链路因物理障碍物遮挡而中断, 通信完全依赖于 RIS 提供的虚拟反射路径, 并且 BS 与 RIS 两者之间的信道具有稳定的 LoS 特性.

本文假设系统的信道满足块衰落特性, 即信道参数在一个相干时间块中保持准静态. 在下行通信过程中, BS 发送归一化的基带数据符号 $s\in\mathbb{C}$ (满足 $\mathbb{E}\left[|s|^2\right]=1$). 因此, 为简化波束训练的复杂度, 本文假设 BS 采用固定的定向波束成形矢量 $\mathbf{f}\in\mathbb{C}^{M\times 1}$ (满足 $\|\mathbf{f}\|_2^2=1$) 对准 RIS. 经由 RIS 的无源反射后, UE 接收的基带信号 y 可表

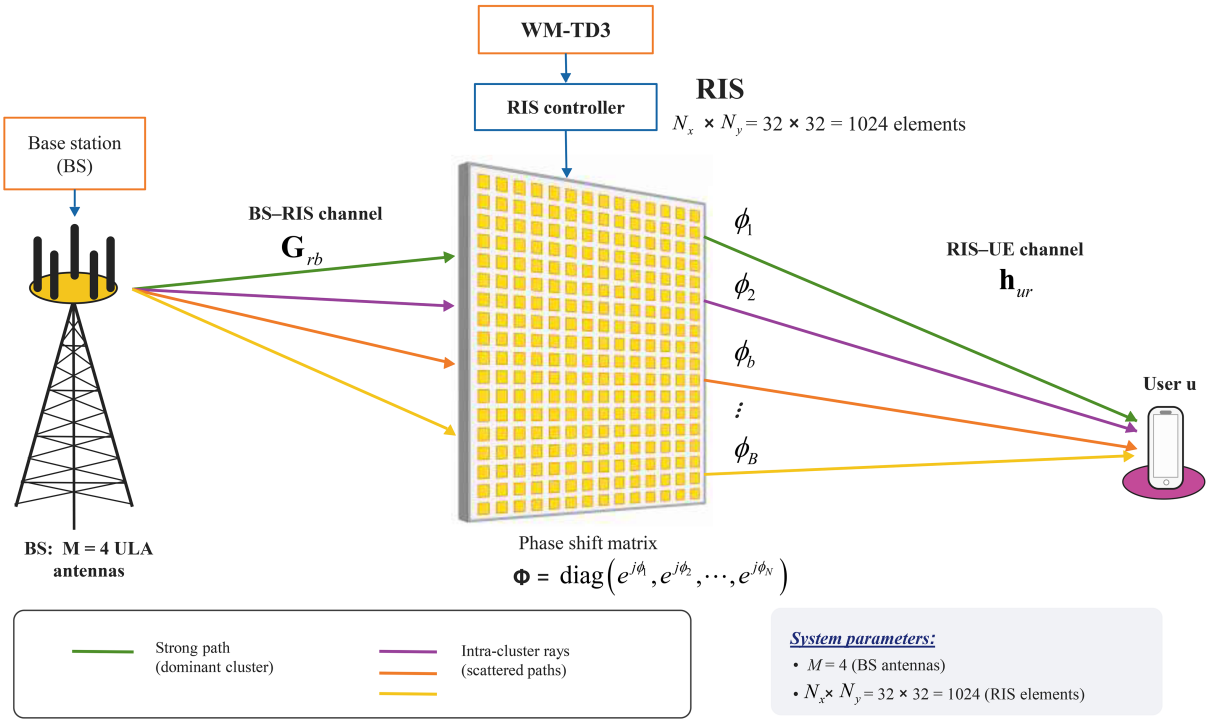

图 1　(网络版彩图) **RIS 辅助毫米波二维波束对齐仿真场景示意图.**

Figure 1　(Color online) Simulation scenario of RIS-assisted 2D beam alignment.

示为

$$y = \sqrt{\rho}\mathbf{h}^{\mathrm{H}}\boldsymbol{\Phi}\mathbf{G}\mathbf{f}s + n, \tag{1}$$

其中, ρ 为归一化发射功率; $\mathbf{G} \in \mathbb{C}^{N\times M}$ 表示 BS 到 RIS 的信道矩阵; $\mathbf{h} \in \mathbb{C}^{N}$ 表示 RIS 到 UE 的信道向量; $n \sim \mathcal{CN}\left(0, \sigma^2\right)$ 为 UE 接收端的加性高斯白噪声 (additive white Gaussian noise, AWGN). RIS 的无源相位矩阵定义为对角阵, 其表达式为

$$\boldsymbol{\Phi} = \mathrm{diag}\left(\mathbf{v}\right) \in \mathbb{C}^{N\times N}, \tag{2}$$

其中, $\mathbf{v} = \left[\mathrm{e}^{\mathrm{j}\phi_1}, \mathrm{e}^{\mathrm{j}\phi_2}, \cdots, \mathrm{e}^{\mathrm{j}\phi_N}\right]^{\mathrm{T}}$ 为反射系数矢量. 考虑到实际硬件的实现复杂度与量化限制, 每个反射单元的相位 ϕ_n 仅能取自有限的离散相位集合:

$$\phi_n \in \mathcal{F} = \left\{0, \frac{2\pi}{2^B}, \cdots, \frac{2\pi\left(2^B-1\right)}{2^B}\right\}, \ \forall n \in \{1, \cdots, N\}, \tag{3}$$

其中 B 为控制每个 RIS 单元相位的量化比特数.

2.2　级联信道模型

在毫米波通信系统中, 考虑到毫米波频段极高的自由空间路径损耗以及稀疏的有效散射体, 信道在空间角域呈现出显著的稀疏性. 因此, 本文采用了 Saleh-Valenzuela (S-V) 几何信道模型来刻画系统的级联通信链路.

在实际的部署中, BS-RIS 链路的信道矩阵 $\mathbf{G} \in \mathbb{C}^{N\times M}$ 建模为

$$\mathbf{G} = \sqrt{\frac{MN}{L_G}} \sum_{l=1}^{L_G} \alpha_l \mathbf{a}_R\left(\theta_l^{\mathrm{AR}}, \varphi_l^{\mathrm{AR}}\right) \mathbf{a}_B^{\mathrm{H}}\left(\theta_l^{\mathrm{AD}}, \varphi_l^{\mathrm{AD}}\right). \tag{4}$$

同理, 假设 RIS-UE 链路中存在 L_h 条散射路径, 其信道向量 $\mathbf{h} \in \mathbb{C}^{N\times 1}$ 可建模为

$$\mathbf{h} = \sqrt{\frac{N}{L_h}} \sum_{k=1}^{L_h} \beta_k \mathbf{a}_R\left(\theta_k^D, \varphi_k^D\right), \tag{5}$$

其中, L_G 与 L_h 为散射簇的数量; α_l 与 β_k 分别表示传播路径的复增益, 其大小受收发端物理距离引起的大尺度路径损耗以及阴影衰落等因素影响. $\mathbf{a}_R(\cdot)$ 与 $\mathbf{a}_B(\cdot)$ 分别表示 RIS 和 BS 的阵列响应矢量. $(\theta_l^{\mathrm{AR}}, \varphi_l^{\mathrm{AR}})$ 表示第 l 条 BS-RIS 路径在 RIS 端的到达角, $(\theta_l^{\mathrm{AD}}, \varphi_l^{\mathrm{AD}})$ 表示该路径在 BS 端的离开角, $(\theta_k^D, \varphi_k^D)$ 表示第 k 条 RIS-UE 路径在 RIS 端的离开角.

考虑到无源 RIS 的物理结构, 系统端到端的级联信道矩阵可定义为 $\mathbf{h}_e = \mathrm{diag}\left(\mathbf{h}^{\mathrm{H}}\right)\mathbf{G} \in \mathbb{C}^{N\times M}$. 将上述 S-V 多径模型代入其中, 可得到级联信道的表达式:

$$\mathbf{h}_e = \sqrt{\frac{MN^2}{L_G L_h}} \sum_{l=1}^{L_G}\sum_{k=1}^{L_h} \alpha_l \beta_k^* \Big[\mathrm{diag}\big(\mathbf{a}_R^{\mathrm{H}}(\theta_k^D, \varphi_k^D)\big)\mathbf{a}_R(\theta_l^{\mathrm{AR}}, \varphi_l^{\mathrm{AR}})\Big]\mathbf{a}_B^{\mathrm{H}}(\theta_l^{\mathrm{AD}}, \varphi_l^{\mathrm{AD}}), \tag{6}$$

其中, 阵列响应矢量 $\mathbf{a}_R$ 的元素均由复指数 $\mathrm{e}^{\mathrm{j}(\cdot)}$ 构成. 利用对角阵乘法的特性, 入射路径与反射路径的空间相位在 RIS 处发生线性叠加. 因此, RIS 级联空间响应项的等效阵列响应矢量为

$$\mathrm{diag}\left(\mathbf{a}_R^{\mathrm{H}}\left(\theta_k^D, \varphi_k^D\right)\right)\mathbf{a}_R\left(\theta_l^{\mathrm{AR}}, \varphi_l^{\mathrm{AR}}\right) = \mathbf{a}_R\left(\widetilde{\theta}_{l,k}, \widetilde{\varphi}_{l,k}\right), \tag{7}$$

其中, $\widetilde{\theta}_{l,k}$ 与 $\widetilde{\varphi}_{l,k}$ 并非实际传播路径中的物理入射角或出射角, 而是由第 l 条 BS-RIS 路径与第 k 条 RIS-UE 路径在 RIS 阵面上的空间频率差共同决定的等效角域参数. 由于 RIS 阵列响应由复指数相位项构成, 该式等价于将两段链路在 RIS 面上的相位进行逐阵元相减, 从而形成一个新的二维复指数序列. 因此可以在 UPA 阵列坐标下表示为水平一维响应与垂直一维响应的 Kronecker 积 (以 $\otimes$ 表示). 基于这一等效空间频率表示, 级联阵列响应可进一步解耦为

$$\mathbf{a}_R\left(\widetilde{\theta}_{l,k}, \widetilde{\varphi}_{l,k}\right) = \mathbf{a}_x\left(\widetilde{\theta}_{l,k}, \widetilde{\varphi}_{l,k}\right) \otimes \mathbf{a}_y\left(\widetilde{\varphi}_{l,k}\right). \tag{8}$$

上述解耦方式建立在规则的 UPA 阵列相应模型之上. 本文假设的 RIS 阵元等间距排布, 且阵元间互耦及极化失配等非理想因素可忽略或已通过校准补偿. 在此条件下, RIS 面上的级联相位响应可表示为由入射与反射空间频率差决定的二维复指数序列, 因此可进一步表示为水平和垂直一维阵列响应的 Kronecker 积. 此建模方式适用于典型毫米波 RIS 中远场窄带通信场景, 对于近场传播、宽带波束斜视等非理想场景, 本文所采用的解耦模型可视为主导空间结构下的近似表示, 并由后续在线微调阶段进一步补偿模型失配.

基于上述公式, 若将 RIS 的相位矢量同样配置为张量积形式 (即 $\mathbf{v} = \mathbf{v}_x \otimes \mathbf{v}_y$), 我们将上述解耦结果代入接收信号中, UE 的精确接收信号 y 可推导为如下严格相等的表达式:

$$y = \left\{\sum_{l=1}^{L_G}\sum_{k=1}^{L_h} \eta_{l,k} \underbrace{\left[\mathbf{v}_x^{\mathrm{T}}\mathbf{a}_x\left(\widetilde{\theta}_{l,k}, \widetilde{\varphi}_{l,k}\right)\right]}_{\text{水平维度增益}} \underbrace{\left[\mathbf{v}_y^{\mathrm{T}}\mathbf{a}_y\left(\widetilde{\varphi}_{l,k}\right)\right]}_{\text{垂直维度增益}}\right\} s + n, \tag{9}$$

其中, $\eta_{l,k} = \sqrt{\frac{\rho MN^2}{L_G L_h}}\alpha_l \beta_k^* \mathbf{a}_B^{\mathrm{H}}\left(\theta_l^{\mathrm{AD}}, \varphi_l^{\mathrm{AD}}\right)\mathbf{f}$.

2.3　优化问题建模

UE 的接收 SNR 可定义为

$$\mathrm{SNR}_u = \frac{1}{\sigma^2}\left|\sum_{l=1}^{L_G}\sum_{k=1}^{L_h} \eta_{l,k}\left[\mathbf{v}_x^{\mathrm{T}}\mathbf{a}_x\left(\widetilde{\theta}_{l,k}, \widetilde{\varphi}_{l,k}\right)\right]\left[\mathbf{v}_y^{\mathrm{T}}\mathbf{a}_y\left(\widetilde{\varphi}_{l,k}\right)\right]\right|^2. \tag{10}$$

在毫米波与 RIS 通信系统中, 为了降低硬件控制开销与波束对准的复杂度, RIS 通常依赖于预先设计的反射波束码本. 因此, 在未知级联 CSI 的前提下, 设计出能够适应当前传播环境的最优独立维度反射码本 $\mathcal{B}_x$ 和 $\mathcal{B}_y$, 并从中找出能使平均接收波束对准性能最大化的最优相位码字. 对应的数学优化问题可表示为

$$\max_{\mathbf{v}_x\in\mathcal{B}_x\,\mathbf{v}_y\in\mathcal{B}_y} \quad \left|\sum_{l=1}^{L_G}\sum_{k=1}^{L_h}\eta_{l,k}\left[\mathbf{v}_x^{\mathrm{T}}\mathbf{a}_x\left(\widetilde{\theta}_{l,k},\widetilde{\varphi}_{l,k}\right)\right]\left[\mathbf{v}_y^{\mathrm{T}}\mathbf{a}_y\left(\widetilde{\varphi}_{l,k}\right)\right]\right|^2 \tag{11a}$$

$$\text{s.t.} \quad [\mathbf{v}_x]_n\in\mathcal{V}, \quad \forall n\in\{1,\ldots,N_x\}, \tag{11b}$$

$$[\mathbf{v}_y]_m\in\mathcal{V}, \quad \forall m\in\{1,\ldots,N_y\}, \tag{11c}$$

$$|\mathcal{B}_x|=B_x, \qquad |\mathcal{B}_y|=B_y, \tag{11d}$$

其中, n 和 m 分别代表水平与垂直维度中的反射单元索引. 集合 $\mathcal{V}$ 通常被设计为预定义的静态码本结构 (如包含等间隔空间角度的离散傅里叶变换 (discrete fourier transform, DFT) 码本), 以匹配底层硬件的移相器精度; B_x 和 B_y 分别对应预设的水平与垂直码本大小.

式 (11a) 旨在未知精确 CSI 条件下, 从水平和垂直两个维度的反射码本中选择合适码字, 使得 RIS 相位补偿后的级联链路复增益尽可能相干叠加, 从而提高 UE 接收的 SNR. 约束 (11b) 和 (11c) 则表示实际的 RIS 移相器的有限分辨率特性; 约束 (11d) 则限定水平与垂直方向的候选码本规模, 反映了导频训练开销和搜索时延受限的工程约束.

因此, 在上述最优化问题 (P1) 中, 原本高维耦合的收发端联合优化转变为在限制维度下的 RIS 码本选择问题.

然而, 直接求解联合优化问题 (P1) 也面临严峻挑战: (1) 目标函数中多径的复数干涉叠加导致优化变量 $\mathbf{v}_x$ 与 $\mathbf{v}_y$ 存在严重相位耦合, 该问题演变为高度非凸的 NP-hard 组合优化难题; (2) 由于硬件限制、测量开销以及动态信道条件, 获取完美的 CSI 既计算成本高又操作困难. 为克服这些限制, 3GPP 标准通常采用波束跟踪, 通过周期性地扫描预定义的一组面向不同方向的波束 (如遍历整个离散傅里叶变换码本) 来定位目标. 然而, 这种穷尽的波束扫描会产生大量的导频开销和训练延迟, 使其在移动目标检测及复杂多变的时变物理传播拓扑场景中极不实际.

3　世界模型辅助的强化学习算法设计

针对问题 (P1) 存在的强相位耦合、NP-hard 优化以及 CSI 获取困难等挑战, 近年来已有诸多前沿研究引入无模型强化学习机制, 尝试通过智能体与环境之间的交互在缺乏显式信道先验的 “黑盒” 条件下, 自适应地学习并映射出最优波束成形策略. 然而, 传统的纯强化学习算法在应用于大规模 RIS 系统时, 往往面临 “冷启动问题”: 由于缺乏 CSI, 智能体在训练初期被迫在呈指数级膨胀的动作空间中进行盲目的随机试错. 这种无导向的探索机制不仅交互试错成本极高、收敛极其缓慢, 而且在崎岖的多径非凸优化流形中极易落入局部最优解, 导致系统无法满足实际通信场景对波束快速、可靠对准的严苛要求.

为应对上述挑战, 打破传统强化学习在未知高维环境下的探索瓶颈, 本文提出一种基于世界模型先验引导与解耦双智能体协同微调的两阶段波束对准方法. 利用世界模型对物理信道演进规律的隐式表征与泛化能力, 为智能体提供初始物理策略先验, 从而直接跨越强化学习初期的低效盲目试错阶段; 随后再依托强化学习在线交互的自适应寻优优势, 进行高精度探索.

通过两个递进的阶段实现高效寻优: (1) 离线预训练阶段, 从少量的探测波束观测中提取信道环境的联合幅相潜特征, 实现对最优波束区间的快速预判与问题降维; (2) 在线学习与微调阶段, 以离线训练预判为基准点, 利用强化学习算法在连续动作空间内进行交互式微调, 逐步消除量化误差并精确锁定最优波束. 具体机制将在下文详细阐述.

3.1　离线阶段: 基于非对称双流 ResNet 的信道环境认知世界模型构建

为克服在线强化学习在大规模 RIS 波束搜索中的冷启动问题, 本文构建认知世界模型对未知信道环境进行离线表征学习. 经典的世界模型核心在于理解物理规律 (预测观测) 与评估环境价值 (预测奖励). 因此, 本文所构建的世界模型并非仅将宽波束观测直接映射为最优波束索引的黑盒分类器, 而是通过幅相观测编码、观测重构、奖励预测和策略先验生成的多任务联合学习, 构建能够表征级联信道隐式结构的潜变量 $\mathbf{z}$. 通过本阶段, 世界模型输出的不仅是初始动作建议, 还包括可供在线阶段利用的物理状态表征.

(1) **联合幅度相位特征提取:** 为避免全空间扫描带来的庞大开销, 系统在初始阶段引入基于二分法构建的分层多分辨率码本. 为了充分捕获信道的宏观多径轮廓, 系统依次选用码本顶层 (第 1 层, 包含 2 个全域宽波束) 与次层 (第 2 层, 包含 4 个子域宽波束), 通过这两层共计 $K=6$ 个多分辨率探测波束对未知信道进行空间维度的非穷举物理 “采样”. 为保证空间物理拓扑信息, 本模型摒弃了仅提取 SNR 的做法, 设计了 18 维的幅度相位量和提取机制.

首先, 提取差分功率特征:

$$\xi_k \;=\; 10\log_{10}\left(\frac{1}{T}\sum_{t=1}^{\mathrm{T}}|y_{k,t}|^2\right)-P_{\mathrm{noise}}^{dB},\quad k\in\{1,\cdots,6\}, \tag{12}$$

其中, $P_{\mathrm{noise}}^{dB}=10\log_{10}\left(\sigma^2\right)$ 为转化为 dB 域的接收机热噪声功率基底. 随后, 选取基准天线接收信号, 并以能量最强的波束 k^* 为基准进行相位差分对齐, 还原了相对空间角度拓扑:

$$\widetilde{\theta}_k \;=\; \angle\left(y_{k,t_0}\right)-\angle\left(y_{k^*,t_0}\right), \tag{13}$$

其中, t_0 表示用于相位比较的参考导频观测索引, 本文取 $t_0=1$. 之后, 利用欧拉公式展开, 构建 6 维能量流 $\mathbf{o}_{\mathrm{mag}}$ 与 12 维相位流 $\mathbf{o}_{\mathrm{phase}}$:

$$\mathbf{o}_{\mathrm{mag}} \;=\; \left[\xi_1,\cdots,\xi_6\right]^{\mathrm{T}}, \tag{14}$$

$$\mathbf{o}_{\mathrm{phase}} \;=\; \left[\cos\widetilde{\theta}_1,\cdots\cos\widetilde{\theta}_6,\sin\widetilde{\theta}_1,\cdots,\sin\widetilde{\theta}_6\right]^{\mathrm{T}}. \tag{15}$$

最后, 包含幅度与连续空间相位的联合幅相观测状态矢量 $\mathbf{o}=\left[\mathbf{o}_{\mathrm{mag}},\mathbf{o}_{\mathrm{phase}}\right]^{\mathrm{T}}$.

(2) **面向异构物理流形的非对称双流宽残差表征架构:** 鉴于提取的 18 维联合特征中, 差分幅度 (对数 dB 域) 与空间相位 (线性正余弦域) 呈现出截然不同的物理量纲与流形分布规律, 若在输入层直接进行全连接映射, 极易导致异构特征之间相互干扰与 “表征塌缩”. 为此, 本模型在网络的浅层采用解耦的非对称双流架构.

- **物理流形解耦与宽残差提取:** 输入特征被解耦为能量流 $\mathbf{o}_{\mathrm{mag}}$ 与相位流 $\mathbf{o}_{\mathrm{phase}}$, 并分别馈入平行的宽残差提取. 残差块中引入层归一化机制对方差剧烈起伏进行平抑, 同时, 采用高斯误差线性单元替代 ReLU 激活函数, 旨在为连续相位映射保留平滑的高阶非线性梯度.
- **“压缩 – 激励” 通道注意力机制:** 在获取了深层独立特征 $\mathbf{h}_{\mathrm{mag}}$ 与 $\mathbf{h}_{\mathrm{phase}}$ 后, 网络将其按通道维度进行级联, 并在融合特征中显式注入高斯随机扰动 $\boldsymbol{\epsilon}\sim\mathcal{N}\left(0,\sigma_f^2\mathbf{I}\right)$, 引入该高斯扰动机制旨在增强网络对极低 SNR 下相位估计误差的鲁棒性, 作为一种变分正则化手段, 促使模型学习具有容错能力的潜在信道表示. 随后, 级联特征馈入 “压缩 – 激励” 通道注意力模块, 自适应地评估并重加权能量与相位特征在当前信道环境下的非线性重要度, 等效地模拟了幅度与相位在接收天线处的复数多径相干叠加过程. 最终, 网络输出物理潜变量 $\mathbf{z}$, 该潜变量深度表征了全局的能量与相位映射关系, 为后续波束策略的生成及环境价值预估提供了核心特征先验.

(3) **三重认知预测头与联合损失函数:** 为使本 “世界模型” 具备认知能力, 在潜变量 $\mathbf{z}$ 上设计了三个独立的预测头.

- **策略先验头:** 潜变量线性分别映射水平与垂直第 3 层目标波束码本的动作空间, 输出预测 8 个子波束置信度的逻辑向量 **logits**, 旨在为智能体提供预测的水平与垂直初始波束指向 $(\widehat{\boldsymbol{\phi}}_{x,t}^{(0)},\widehat{\boldsymbol{\phi}}_{y,t}^{(0)})$ 的物理先验分布, 加快后续的在线探索.

• **观测重构头:** 基于潜变量输出重构特征 $\widehat{\mathbf{o}}$, 以复现原始的 18 维联合幅相观测矢量. 该机制作为本模型的正则化约束, 旨在促使潜变量 $\mathbf{z}$ 真实隐式编码信道, 有效规避了模型陷入传统黑盒模型易对离散动作标签产生严重过拟合的风险.

• **环境价值预估头:** 输出连续标量 $\widehat{\gamma}_{\max}$, 用于评估当前信道状态下理论可达的最优 SNR. 该预测不仅反映了此时物理环境的通信质量上限, 亦为在线强化学习在未知环境下的奖励预期提供基线.

世界模型的离线训练需要通过联合优化三个预测头的损失来实现. 对于策略先验头, 由于毫米波信道的空间连续性, 本文并未采用只奖励单一最优波束的传统 One-hot 硬标签交叉熵, 而是设计了融合能量软分布 P_{energy} 与角度空间容差 P_{spatial} 的混合物理交叉熵损失 $\mathcal{L}_{\text{policy}}$, 具体而言, 能量软分布利用真实测量的全部波束增益 g_i 结合温度系数 τ 构造, 赋予高能量波束成比例的探索概率:

$$P_{\text{energy},i} = \frac{\exp\left[\left(g_i - g_{\max}\right)/\tau\right]}{\sum_{j=1}^{8} \exp\left[\left(g_j - g_{\max}\right)/\tau\right]}, \quad i = 1, \ldots, 8. \tag{16}$$

空间容差分布则以最优波束的物理索引 i^* 为中心, 构造方差为 σ_s^2 的高斯拓扑分布, 以合理分配相邻波束的探索置信度:

$$P_{\text{spatial},i} = \frac{\exp\left[-\frac{(i-i^\star)^2}{2\sigma_s^2}\right]}{\sum_{j=1}^{8} \exp\left[-\frac{(j-i^\star)^2}{2\sigma_s^2}\right]}, \quad i = 1, \ldots, 8. \tag{17}$$

最终的策略先验损失函数定义为硬标签交叉熵与上述软标签库尔贝克 – 莱布勒 (Kullback-Leibler, KL) 散度的加权和:

$$\mathcal{L}_{\text{policy}} = \beta CE\left(\text{logits},\ i^*\right) + (1-\beta)\,\text{KL}\left(\alpha P_{\text{energy}} + (1-\alpha)\,P_{\text{spatial}} || \text{Softmax}\left(\text{logits}\right)\right). \tag{18}$$

结合均方误差, 世界模型的整体多任务认知损失函数定义为

$$\mathcal{L} = \mathcal{L}_{\text{policy}} + \lambda_{\text{recon}} \left\|\mathbf{o} - \widehat{\mathbf{o}}\right\|_2^2 + \lambda_{\text{reward}} \left\|\gamma_{\max}^{\text{true}} - \widehat{\gamma}_{\max}\right\|_2^2, \tag{19}$$

其中, λ_{recon} 与 λ_{reward} 为多任务平衡权重. 经过预训练, 该模型不仅可以输出初始波束的第三层先验 $\widehat{\boldsymbol{\phi}}_i^{(0)}$, $(i \in \{x, y\})$, 更能为在线强化学习提供物理潜变量 $\mathbf{z}$, 赋予系统对物理信道环境的感知, 从而有效突破传统强化学习面对高维原始观测量的盲目搜索瓶颈.

因此, 本阶段的认知世界模型核心作用并非单一的波束标签预测, 而是通过多任务约束学习具有物理含义的信道潜变量 $\mathbf{z}$, 并将该潜变量及其对应的策略先验传递给在线强化学习阶段, 从而实现离线信道认知与在线策略微调之间的衔接.

世界模型的离线预训练具体流程如算法 1 所示. 在具体的训练超参数设置方面, 本文设定总训练轮数为 E_{wm}, 并选取适中的批次大小 B_{wm} 以在单次参数迭代中兼顾计算效率与梯度评估的无偏性; 同时, 训练引入自适应矩估计优化器, 并设定初始学习率 η 以平滑控制网络梯度的下降步长. 系统首先利用式 (12)∼(15) 提取观测特征 $\mathbf{o}$, 并将其连同最优的物理分布存入离线数据集 $\mathcal{D}_{\text{off}}$ 中. 随后, 通过小批量抽取数据并计算式 (19) 的联合损失函数, 利用梯度下降更新世界模型网络参数 Θ_{wm}. 该网络在离线训练中能够单调收敛, 并最终输出捕获了物理环境先验的最优模型参数 Θ_{wm}^*.

3.2　在线阶段: 世界模型辅助下的并行双智能体 TD3 自适应协同微调

经过离线预训练后, 世界模型可由少量宽波束观测提取当前信道潜在状态表示 $\mathbf{z}$, 并生成水平与垂直方向的初始波束先验. 其输出本质上仅为系统提供了一个高质量的初始可行解, 面对具有多径复数干涉叠加且极不规则的实际物理流形, 该先验能够显著缩减在线搜索范围, 但仍会引入不可忽视的对准残差.

Algorithm 1 Offline training of the cognitive world model.

Input: $E_{\mathrm{wm}}, B_{\mathrm{wm}}, \eta, \mathcal{P}_{\mathrm{probe}}$;
Initialize Θ_{wm} and offline dataset $\mathcal{D}_{\mathrm{off}}$;
for each channel realization **do**
  Given $\mathcal{P}_{\mathrm{probe}}$, obtain $\mathbf{o}$ via (12)$\sim$(15);
  Given $\mathbf{o}$, obtain i_x^*, i_y^*, $\gamma_{\max}^{\mathrm{true}}$;
  Given $\mathbf{o}$, i_x^*, i_y^*, $\gamma_{\max}^{\mathrm{true}}$, update $\mathcal{D}_{\mathrm{off}}$;
end for
for $e = 1, 2, \cdots, E_{\mathrm{wm}}$ **do**
  Given $\mathcal{D}_{\mathrm{off}}$, obtain mini-batch $\mathcal{B}$;
  Given $\mathbf{o}$, obtain $\mathbf{z}$, $\mathbf{p}$, $\widehat{\mathbf{o}}$, and $\widehat{\gamma}_{\max}$ via ResNet;
  Given $\bar{\gamma}_x$, $\bar{\gamma}_y$, i_x^*, and i_y^*, obtain $\mathcal{L}_{\mathrm{wm}}$ via (19);
  Given $\mathcal{L}_{\mathrm{wm}}$, update $\Theta_{\mathrm{wm}} \leftarrow \Theta_{\mathrm{wm}} - \eta \nabla_{\Theta_{\mathrm{wm}}} \mathcal{L}_{\mathrm{wm}}$;
end for
Output: Θ_{wm}^*.

为此, 在世界模型的先验邻域内引入在线强化学习微调机制, 即将波束的持续跟踪建模为马尔可夫决策过程 (Markov decision process, MDP). 在解决波束赋形优化问题时, 本文采用了将物理阵列的空间解耦彻底贯彻至算法架构层面, 提出了 DDA-TD3 架构, 为避免在二维波束空间中进行指数级复杂度的联合搜索, 从而将搜索复杂度由 $\mathcal{O}(N^2)$ 降低至 $\mathcal{O}(2N)$, 使其能够通过在线闭环微调快速锁定并维持全局二维最优波束.

(1) **DRL 智能体交互要素建模:** 在深入探讨 TD3 架构的核心之前, 首先系统性地定义适用于水平与垂直子问题 (即维度 $i \in \{x, y\}$) 的 DRL 智能体通用交互要素. 在任意时隙 t 时, 维度 $i \in \{x, y\}$ 的智能体交互要素定义如下.

• **状态向量** ($\mathbf{s}_{i,t}$): 表示智能体在维度 i 上第 t 时隙的状态. 有别于传统强化学习单一的反馈信号, 本智能体的状态输入拼接了世界模型输出的高密度物理潜变量 $\mathbf{z}_t$ 以及该维度专属的硬件历史时序记忆 (即前一时刻的相位配置与整体波束增益). 即状态向量定义为

$$\mathbf{s}_{i,t} = \left[\mathbf{z}_t^{\mathrm{T}}, \boldsymbol{\phi}_{i,t-1}^{\mathrm{T}}, \gamma_{t-1}\right]^{\mathrm{T}}, \tag{20}$$

其中, 当前状态 $\mathbf{s}_{i,t}$ 直接决定了 t 时隙对应维度的初始波束基准. 该解耦复合状态减弱了传统 MDP 在处理高维通信环境时固有的表征缺陷.

• **动作向量** ($\mathbf{a}_{i,t}$): 表示维度 i 上第 t 时隙的动作输出, 其反映了对状态 $\mathbf{s}_{i,t}$ 中相位的连续调整. 具体而言, Actor 网络输出对应维度的连续预动作 $\mathbf{a}_{i,t} = \Delta\boldsymbol{\phi}_{i,t}$. 由于预动作不一定符合底层硬件可用的相位量化等级, 系统将其以世界模型输出的该维度初始先验相位 $\widehat{\boldsymbol{\phi}}_{i,t}^{(0)}$ 为中心进行物理限幅与叠加:

$$\boldsymbol{\phi}_{i,t} = \widehat{\boldsymbol{\phi}}_{i,t}^{(0)} + \Delta_{\max} \cdot \mathbf{a}_{i,t}, \quad i \in \{x,\ y\}. \tag{21}$$

随后, 引入量化器, 通过计算最小欧氏距离将上述连续相位就近映射至 RIS 的可用离散位集合 $\mathcal{F}$ 中, 合并为二维反射阵列后联合下发. 该动作向量决定了对应维度智能体的下一轮次状态演进, 即 $\mathbf{a}_{i,t} \to \mathbf{s}_{i,t+1}$.

• **共享差分奖励函数** (r_t): 尽管 x 与 y 维度的寻优决策过程被完全解耦, 但由于 2.2 小节已证明了二维阵列的接收增益是水平增益与垂直增益的严格克罗内克积, 因此, 两个维度的智能体共享同一个基于整体波束成形增益 γ_t 评估的瞬时奖励. 在第 t 时隙, 根据式 (10), 将该相位配置对应的线性瞬时接收 SNR 定义为 $\gamma_t \triangleq \mathrm{SNR}_u(\mathbf{v}_t)$, γ_t 为线性域 SNR, 则两个智能体共享以下差分奖励:

$$r_t = \log_{10}(1 + \gamma_t) - \log_{10}(1 + \gamma_{t-1}). \tag{22}$$

本文并未直接采用瞬时 SNR 作为奖励, 而是采用相邻时隙 SNR 对数值的增量形式. $\log_{10}(1+\gamma_t)$ 与链路可达速率具有单调对应关系, 能够反映相位调整带来的通信性能变化; 并且差分形式更加关注当前动作相对于上

一时隙的性能改进, 可兼容客观因素造成的奖励尺度偏移. 当相位配置带来有效 SNR 提升时, 智能体获得正向奖励, 反之则受到惩罚.

(2) **DDA-TD3 核心网络架构与协同机制:** 在连续动作空间的深度强化学习中, 智能体 Actor 的迭代优化高度依赖于 Critic 网络对动作价值函数 (Q 函数) 的精准评估. 对于维度 $i \in \{x,\ y\}$, 其真实的 Q 函数 $Q_{\pi_i}\left(\mathbf{s}_{i,t},\ \mathbf{a}_{i,t}\right)$ 的严格数学定义为: 在当前状态 $\mathbf{s}_{i,t}$ 下执行特定动作 $\mathbf{a}_{i,t}$, 并在此后持续遵循策略 π_i 所能获得的期望累积折扣奖励:

$$Q_{\pi_i}\left(\mathbf{s}_{i,t}, \mathbf{a}_{i,t}\right) = \mathbb{E}_{\pi_i}\left[\sum_{k=0}^{\infty} \lambda^k r_{t+k} | \mathbf{s}_{i,t}, \mathbf{a}_{i,t}\right], \tag{23}$$

其中, $\lambda \in (0,1)$ 为平衡近期与远期回报的折扣因子. 在本文的多径波束对准场景中, 该函数并非单纯衡量单步相位微调所引发的瞬时 SNR 波动, 而是深度量化了在当前信道宏观潜特征的前提下, 执行特定连续相位动作后, 系统在整个相干时间块内预期累积的长期全局 SNR 折扣增益潜力. Q 价值更高的相位动作意味着该调整轨迹能够有效规避多径干涉形成的 "局部次优陷阱", 从而指引波束平滑且稳定地收敛至全局最优方向.

然而, 高维连续相位控制中深度确定性策略梯度 (deep deterministic policy gradient, DDPG) 或近端策略优化 (proximal policy optimization, PPO) 易出现价值过估计和收敛不稳定的问题, 本文采用 TD3 架构, 并结合二维 RIS 阵列结构设计双智能体优化机制. 其中, H-Agent ($i = x$) 与 V-Agent ($i = y$) 分别负责水平和垂直相位变量, 二者各自配置独立 Actor 网络 $\pi_{\psi,i}$, 并采用双 Critic 网络 $Q_{\theta_1,i}$ 和 $Q_{\theta_2,i}$ 进行截断 Q 值评估, 交互样本存入对应经验回放池 $\mathcal{D}_i$. 对于智能体, 其利用 TD3 的三项增强机制如下.

• **截断双 Q 评估机制:** 为了抑制多径波动引发的过估计, 各智能体计算其目标 Q 价值 $y_{i,k}$ 时, 均提取专属双目标 Critic 的最小值:

$$y_{i,k} = r_k + \lambda \min_{j=1,2} Q_{\theta_j',i}\left(\mathbf{s}_{i,k+1}, \mathbf{a}_{i,k+1}^{'}\right), \tag{24}$$

其中, r_k 为第 k 步的即时反馈奖励; $Q_{\theta_j',i}$ 分别表示智能体 i 的两个相互独立的目标 Critic 网络. 该公式通过选取双目标网络预测值中的下界 (最小值) 来计算目标价值, 从算法源头上阻断了多径非凸环境中由量化误差与 SNR 突变引起的 Q 值正向累积虚高, 确保了对未知环境长期潜能的稳健、保守评估.

随后, 各自的 Critic 网络的参数基于 MSE 独立更新:

$$\mathcal{L}\left(\theta_{j,i}\right) = \frac{1}{N_B}\sum_{k=1}^{N_B}\left(y_{i,k} - Q_{\theta_{j,i}}\left(\mathbf{s}_{i,k}, \mathbf{a}_{i,k}\right)\right)^2, \tag{25}$$

其中, N_B 为经验回放池抽取的微批次样本大小; $Q_{\theta_j,i}$ 为当前待更新的 Critic 网络预测值. 该损失函数最小化当前动作评估偏离稳健目标价值的均方差, 驱动 Critic 网络精准拟合高维相位空间下真实的累积波束成形增益上限.

• **目标策略平滑:** 为防止过拟合于偶尔产生的虚假波束尖峰, 在评估目标动作 $\mathbf{a}_{i,k+1}^{'}$ 时, 独立添加受截断的探索噪声 ϵ_i:

$$\mathbf{a}_{i,k+1}^{'} = \pi_{\psi',i}\left(\mathbf{s}_{i,k+1}\right) + \epsilon_i, \tag{26}$$

其中, $\pi_{\psi',i}$ 表示智能体 i 的目标 Actor 网络输出的确定性波束决策动作; ϵ_i 为服从特定分布 (如高斯分布) 且经过上下界截断处理的平滑噪声. 该引入机制促使水平与垂直两个维度的智能体在评估未来策略时, 不盲目轻信某一尖锐的极值点, 而是在其附近的局部 SNR 流形中平滑出稳定的优化路径, 有效防止策略对非目标散射路径引发的虚假窄波束尖峰产生严重过拟合.

• **延迟目标网络更新:** 各智能体的 Actor 网络参数采用延迟机制独立更新:

$$\nabla_{\psi,i} J\left(\psi_i\right) \approx \frac{1}{N_B}\sum_{k=1}^{N_B} \nabla_{\mathbf{a}} Q_{\theta_{1,i}}\left(\mathbf{s}_{i,k}, \mathbf{a}\right)|_{\mathbf{a}=\pi_{\psi,i}(\mathbf{s}_{i,k})} \nabla_{\psi,i} \pi_{\psi,i}\left(\mathbf{s}_{i,k}\right), \tag{27}$$

其中, $J(\psi_i)$ 为智能体 i 的策略优化目标函数, ψ_i 为当前待优化的 Actor 网络参数. 该公式利用确定性策略梯度定理, 通过表现最优的 Critic 网络 $Q_{\theta_{1,i}}$ 提供的价值梯度方向, 来引导 Actor 网络产生更高收益的波束动作. 只有当 Critic 网络的价值评估函数经过多次迭代、收敛得足够稳定时, 系统才依据其指导来更新 Actor 的波束决策. 这极大地避免了在复杂时变信道下 "错误评估指导错误动作" 的恶性循环, 确保了寻优轨迹的平滑收敛.

本文所提出的 WM-TD3 方法应理解为一种由离线世界模型先验初始化, 再由在线 DDA-TD3 根据实时 SNR 反馈逐步修正的启发式近似优化算法. 其优势主要体现在降低冷启动探索开销、提高在线收敛效率和增强对信道扰动的适应能力, 而非保证对原始非凸组合问题达到严格全局最优.

DDA-TD3 协同微调的具体执行流程如算法 2 所示. 定义 $T_{\max}$ 为最大交互时隙数 (即微调步数), τ_{soft} 为目标网络的软更新系数, d 为 Actor 策略的延迟更新频率. 解耦的双智能体 DDA-TD3 首先以算法 1 预判的锚点 $\left(\widehat{\boldsymbol{\phi}}_x^{(0)}, \widehat{\boldsymbol{\phi}}_y^{(0)}\right)$ 作为探索基准. 在每个时隙生成动作时, 智能体并行输出叠加了探索噪声 $\epsilon_{\text{explore}}$ 的连续预动作, 并利用式 (14) 的最大微调步长 $\Delta_{\max}$ 进行映射, 随后通过量化器将其就近转化为满足硬件约束的离散相位 $\mathbf{v}_{i,t}$. 最后, 两个智能体基于共享差分奖励和经验回放分别更新 Critic 与 Actor, 并通过截断双 Q 评估和延迟策略更新提高训练稳定性, 从而实现对水平和垂直相位变量的协同微调.

Algorithm 2 Online decoupled dual-agent TD3 (DDA-TD3).

Input: $\mathcal{P}_{\text{probe}}$, $\Delta_{\max}$, Θ_{wm}^*, $\mathcal{T}_{\max}$, d, τ_{soft}, σ_{target}, c;
Given $\mathcal{P}_{\text{probe}}$, obtain $\mathbf{o}$ via (12)∼(15);
Given $\mathbf{z}$ and Θ_{wm}^*, obtain $\left(\widehat{\phi}_x^{(0)}, \widehat{\phi}_y^{(0)}\right)$;
Quantize and deploy the initial reflection coefficient vector $\mathbf{v}_0 = \mathbf{v}_{x,0} \otimes \mathbf{v}_{y,0}$, and measure γ_0;
Set $\gamma_{\text{best}} \leftarrow \gamma_0$ and $\mathbf{v}^\star \leftarrow \mathbf{v}_0$;
Initialize $\pi_{\psi,i}$, $Q_{\theta_{j,i}}$, target networks, and $\mathcal{D}_i$ for $i \in \{x, y\}$;
for $t = 1, \cdots, T_{\max}$ **do**
  Given $\mathbf{s}_{i,t}$, obtain $\mathbf{a}_{i,t}$ via $\pi_{\psi,i}$, $i \in \{x, y\}$;
  Given $\mathbf{a}_{i,t}$, $\epsilon_{\text{explore}}$, and $\widehat{\phi}_i^{(0)}$, obtain $\mathbf{v}_{i,t}$ via (21) and quantizer;
  Deploy $\mathbf{v}_t = \mathbf{v}_{x,t} \otimes \mathbf{v}_{y,t}$, and obtain γ_t;
  Given γ_t, obtain r_t via (22);
  Given $\mathbf{s}_{i,t}$, $\mathbf{a}_{i,t}$, r_t, and $\mathbf{s}_{i,t+1}$, update $\mathcal{D}_i$, $i \in \{x, y\}$;
  for $i \in \{x, y\}$ in parallel **do**
    Given $\mathcal{D}_i$, obtain mini-batch samples;
    Given mini-batch samples, obtain $y_{i,k}$ via (24);
    Given $y_{i,k}$, update $Q_{\theta_{1,i}}$ and $Q_{\theta_{2,i}}$ via (25);
    if t mod $d = 0$ **then**
      Given $Q_{\theta_{1,i}}$, update $\pi_{\psi,i}$ via (27);
      Soft update target networks with τ_{soft};
    end if
  end for
end for
Output: $\mathbf{v}^*$.

4　仿真结果与分析

本节通过一系列数值仿真实验, 全面评估本文所提出的 "基于世界模型先验引导与解耦双智能体协同微调的两阶段波束对准方法" 在 RIS 辅助毫米波通信系统中的性能, 并深入剖析其在寻优开销、收敛速度以及抗噪鲁棒性等方面的核心优势.

4.1　仿真参数与对比方法设置

仿真采用 Saleh-Valenzuela 几何多径信道模型, 载波频率为 28 GHz, BS 配置 $M = 4$ 根 ULA 天线, RIS 采

表 1　网络训练超参数设置.

Table 1　Network training hyperparameters.

Offline pre-training (world model)		Online fine-tuning (DDA-TD3)	
Parameter	Value	Parameter	Value
Dataset size	50000 samples	Channel topologies/ total episodes	2000/20000
Optimizer/learning rate schedule	AdamW/cosine annealing	Max interaction steps per episode	100
Initial learning rate/epochs	10^{-3}/500	Replay buffer capacity/ mini-batch size	10^6/256
Batch size	512	Actor/critic learning rate	$10^{-4}/10^{-4}$
		Discount factor (λ)/soft target update (τ_{soft})	0.99/0.005
		Exploration noise (ϵ)	0.5 linearly decayed to 0.05 (first 6000 eps)

表 2　不同波束对准方法的复杂度与搜索开销对比.

Table 2　Complexity and search overhead comparison of different beam alignment methods.

Method	Search space	Online overhead	Main characteristic
Exhaustive search	B_xB_y	$O(B_xB_y)$	Highest alignment accuracy but prohibitive pilot overhead
Hierarchical search	Layer-wise codebook	$O(B_x+B_y)$	Low implementation complexity but sensitive to early-stage misalignment
Pure TD3	High-dimensional continuous phase space	$O(TC_{\text{NN}})$	Adaptive online optimization but suffers from cold-start exploration
WM-TD3	Decoupled prior-guided local search	$O(K+T_{\max}C_{\text{NN}})$	World-model prior reduces blind exploration and improves search efficiency

用 $N=N_xN_y=32\times32=1024$ 单元 UPA, 阵元间距为 $\frac{\lambda}{2}$. 设置 2 个散射簇, 每簇包含 3 条射线, 测试环境 SNR 为 10 dB.

离线训练集由 S-V 几何信道模型生成. 每个样本对应一个随机信道拓扑, 包括用户位置、散射簇角度、路径增益和阴影衰落等参数. 在该拓扑下, 通过有限宽波束探测得到幅相观测向量, 并通过仿真可获得的全码本搜索或高精度搜索结果生成策略先验标签和最优 SNR 标签. 因完美 CSI 或穷举搜索仅用于离线仿真标注, 不参与在线波束对准过程. 在实际部署中, 该数据也可由射线追踪、历史波束训练日志或少量校准测量构建, 属于一次性或低频更新成本.

本文中信道拓扑数量是由用户位置、散射簇数量、角度分布、路径增益和遮挡状态共同确定的宏观传播几何结构. 训练轮次指在给定拓扑环境下进行的一次完整波束对准交互过程, 包括初始观测、连续相位微调、奖励反馈和终止判决. 一个拓扑环境可生成多个训练轮次, 以模拟同一条件下初始观测变化.

关于网络训练, 本实验分为世界模型离线预训练与 DDA-TD3 在线微调两个阶段. DDA-TD3 中 Actor 与 Critic 网络均采用多层感知机架构. 具体的网络训练超参数设置如表 1 所示.

为进一步说明不同方法在搜索空间、在线测量开销与计算复杂度方面的差异, 本文对所提 WM-TD3 方法及两类基线算法进行复杂度对比, 如表 2 所示. 其中, B_x 和 B_y 分别表示水平与垂直维度的码本规模, T 表示在线微调过程中的最大交互步数, C_{NN} 表示一次神经网络前向或参数更新所需的计算开销.

在仿真分析中, 本文将所提的 WM-TD3 方法与以下两类进行对比, 分别是: 基于传统强化学习的无先验方法 (Pure TD3 算法) 和基于分层搜索的方法. 基于传统强化学习的无先验方法, 不包含世界模型的离线特征提取与高置信度认知先验策略, 智能体直接在 1024 维的高维连续相位空间中进行探索交互, 用于直观验证物理先验引导的重要性; 基于分层搜索的方法是目前最经典、最常用的机制, 其利用多分辨率码本进行逐层扫描, 在传

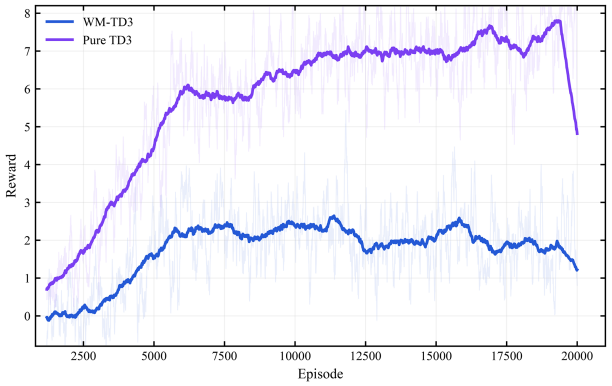

图 2　(网络版彩图) 算法在训练过程中的平均累积奖励对比.

Figure 2　(Color online) Comparison of average cumulative rewards over the training process.

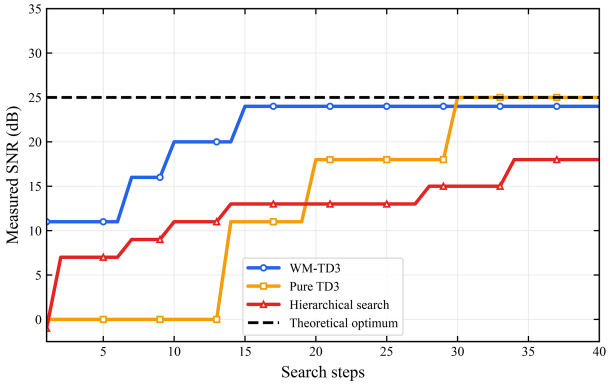

图 3　(网络版彩图) 不同波束搜索方法的单次 SNR 轨迹对比.

Figure 3　(Color online) Single-episode SNR trajectory comparison of different beam search methods.

统波束训练的理论分析与工程部署中具有一定的理论指导意义.

4.2　仿真结果

图 2 展示了不同方法在训练过程中的平均累积差分奖励. 因式 (22) 定义的奖励反映相邻时隙 SNR 对数值的增量, 而非最终接收 SNR. 因此, WM-TD3 的累积奖励低于 Pure TD3 并不表示其对准性能较差: 世界模型提供的初始波束已位于较高 SNR 区域, 在线微调所获得的剩余增量较小; Pure TD3 则需要从较低质量的初始状态开始进行较长时间的探索, 因而累积了更多正向差分奖励. TD3 方法由于缺乏环境先验而在高维动作空间中进行漫长的盲目试错, 累积大量微小正向差分奖励导致总收益高, 且带来了较高对准开销, 本文的 WM-TD3 方法跨越了盲目试错阶段, 仅需极少的交互步数即可完成精准微调, 进一步验证了世界模型先验引导在帮助系统有效规避维度灾难、实现快速寻优中的关键性与重要性.

图 3 进一步给出了单次波束对准过程中的 SNR 演变轨迹. WM-TD3 在搜索初期即获得较高 SNR, 并以较少的交互步数逼近理论最优值; Pure TD3 因缺少先验引导而收敛较慢, 分层搜索则受静态码本分辨率限制而停留在局部次优区域. 上述结果表明, 世界模型先验能够压缩在线搜索空间并提高波束对准效率.

图 4 展示了成功锁定最优波束所需的平均搜索步数随训练轮次的变化. 传统分层搜索在图中表现为固定在 40 步的水平红线, 这是因为该基线采用水平与垂直维度串行解耦搜索, 且两个维度均按照预设码本层级完成固定流程. 在单个维度第 1 层扫描 2 个宽波束并保留 2 个候选, 第 2∼5 层每层扫描 4 个候选波束, 因此单维度测量步数为 20 步, 水平和垂直两个维度共需 40 步. 该结果说明分层搜索的开销主要由静态码本结构决定, 不随当前信道状态自适应变化; 虽然其实现简单、开销可控, 但缺乏提前终止能力, 在有利信道场景下难以进一步降低搜索步数.

相比之下, 数据驱动的强化学习方法均表现出随着训练推进而逐步降低搜索步数的自适应优化趋势. 然而, Pure TD3 方法虽然能够在训练后期将搜索步数降低至约 33 步, 但由于其缺乏有效先验引导, 需要在高维搜索空间中进行较大范围的试探与调整, 因此其搜索效率提升相对有限.

本文的 WM-TD3 方法则从训练初期便表现出更高的搜索效率. 其初始搜索步数约为 27.5 步, 已明显低于 Pure TD3 方法的后期收敛水平, 并在训练后期进一步稳定收敛至约 24 步. 该结果表明, 所提出方法能够显著降低搜索过程中的交互开销. 其原因便在于, 对世界模型提供的先验信息, 智能体能够在更小的有效搜索空间内完成策略优化, 进而减少无效试探并提高搜索效率.

在实际的无线部署中, SNR 的恶化往往会导致波束训练算法发生误判. 图 5 展示了不同环境 SNR 条件下各方法平均搜索步数的对比结果. 仿真结果表明, 传统分层搜索算法在所有 SNR 条件下均表现为固定的 40 步.

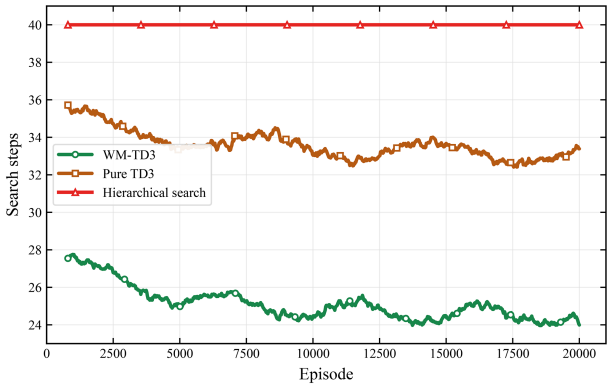

图 4　(网络版彩图) 搜索步数随训练轮次的演进对比.

Figure 4　(Color online) Comparison of the evolution of search steps during training.

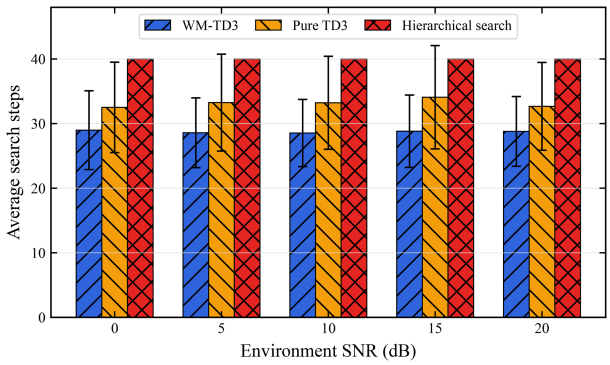

图 5　(网络版彩图) 不同 SNR 环境下的平均搜索步数对比

Figure 5　(Color online) Comparison of average number of search steps under different SNR environments.

这是因为该方法采用预定义的层级码本结构完成逐层波束搜索, 其搜索层数、候选波束数以及测量流程均由算法结构预先设定, 不具备基于环境反馈自适应调整搜索步数的能力. 因此, 其搜索开销在不同环境条件下保持不变. 同时, 在低 SNR 场景中, 分层方法主要依赖能量测量结果进行候选波束判决, 噪声扰动可能导致层级选择错误并逐层累积, 从而限制其搜索性能.

Pure TD3 方法则表现出一定的自适应优化能力. 在较高 SNR 条件下, 其平均搜索步数能够维持在 32~33 步左右; 然而, 当环境 SNR 降低至 0 和 5 dB 时, 其平均搜索步数增加. 这表明, 在低 SNR 场景下, 纯交互式强化学习方法更容易受到噪声污染的影响, 导致奖励反馈和状态观测的不确定性增加, 进而削弱策略学习的稳定性, 并造成搜索开销上升.

相比之下, 本文提出的 WM-TD3 方法在不同环境 SNR 条件下始终保持较低且稳定的搜索步数, 其平均值整体维持在约 24 步附近, 且波动范围显著小于 Pure TD3 方法. 该结果表明, 所提出方法在低 SNR 环境中具有更强的鲁棒性和更稳定的搜索效率. 其原因在于, 世界模型在离线训练阶段通过非对称双流宽残差表征结构, 并结合随机扰动增强, 能够从含噪观测中提取更加稳健的物理特征表示, 从而为在线强化学习提供可靠的先验信息, 因此能够在不同环境 SNR 条件下均保持较低的搜索开销.

5　总结

本文针对未知 CSI 条件下大规模 RIS 辅助毫米波通信系统中的二维波束搜索问题, 提出了一种 WM-TD3 方法. 该方法为高维波束搜索降维提供理论依据. 在此基础上, 构建非对称双流残差网络作为认知世界模型, 从宽波束幅相观测中学习潜在信道特征, 并为在线搜索提供初始波束和奖励先验. 同时, 设计解耦双智能体 TD3 微调机制, 使水平和垂直智能体在共享差分奖励引导下并行优化相位变量, 从而降低高维动作空间带来的交互开销和训练不稳定性.

仿真结果表明, 所提 WM-TD3 方法能够有效缓解传统强化学习在高维相位搜索中的冷启动困难和探索冗余问题, 将平均波束搜索步数降低至约 24 步, 并在低 SNR 环境下保持较好的搜索稳定性和抗噪性能. 与纯强化学习方法和传统分层搜索方法相比, WM-TD3 在搜索效率、训练稳定性和环境鲁棒性方面均表现出更优性能, 可为大规模动态 RIS 辅助通信系统提供一种有效的智能波束对准方案.

参考文献

1　Saad W, Bennis M, Chen M. A vision of 6G wireless systems: applications, trends, technologies, and open research problems. IEEE Netw, 2020, 34: 134–142

2 Heath R W, Gonzalez-Prelcic N, Rangan S, et al. An overview of signal processing techniques for millimeter wave MIMO systems. IEEE J Sel Top Signal Process, 2016, 10: 436–453

3 Song J, Choi J, Love D J. Common codebook millimeter wave beam design: designing beams for both sounding and communication with uniform planar arrays. IEEE Trans Commun, 2017, 65: 1859–1872

4 Liang Y, Wang J Q, Li B L, et al. Near-field wideband cellular codebook design and channel estimation in terahertz extremely large-scale MIMO. J Commun, 2026, 47: 106–118 [梁彦, 王俊清, 李宝龙, 等. 太赫兹超大规模 MIMO 中的近场宽带蜂窝码本设计与信道估计. 通信学报, 2026, 47: 106–118]

5 Noh S, Zoltowski M D, Love D J. Multi-resolution codebook and adaptive beamforming sequence design for millimeter wave beam alignment. IEEE Trans Wireless Commun, 2017, 16: 5689–5701

6 Zhang R, Zhang H, Xu W, et al. A codebook based simultaneous beam training for mmWave multi-user MIMO systems with split structures. In: Proceedings of 2018 IEEE Global Communications Conference (GLOBECOM), 2018

7 Xu Y, Li X Y, Shen J Y, et al. Research on mmWave low-altitude UAV ISAC beam training and tracking technology. J Data Acquis Process, 2025, 40: 56–71 [徐媛, 李心怡, 沈嘉宇, 等. 毫米波低空无人机通感波束训练与追踪技术研究. 数据采集与处理, 2025, 40: 56–71]

8 Cui T J, Qi M Q, Wan X, et al. Coding metamaterials, digital metamaterials and programmable metamaterials. Light Sci Appl, 2014, 3: e218

9 Wu Q, Zhang S, Zheng B, et al. Intelligent reflecting surface-aided wireless communications: a tutorial. IEEE Trans Commun, 2021, 69: 3313–3351

10 Li B, Liu W, Xie W. Secure transmission design for RIS-aided symbiotic radio networks: a DRL approach. Digital Commun Netws, 2024, 10: 1566–1575

11 Guo Y, Guo Y X, Liu G, et al. Beamforming and resource allocation for reconfigurable intelligent surface assisted wireless powered communication network. Sci Sin Inform, 2026, 56: 203–215 [郭漪, 郭颖欣, 刘刚, 等. 智能超表面辅助无线供电通信网络的波束成型与资源分配方案. 中国科学: 信息科学, 2026, 56: 203–215]

12 Xia H, Xue Q, Liu Y, et al. Intelligent angle map-based beam alignment for RIS-aided mmWave communication networks. IEEE Trans Netw Sci Eng, 2026, 13: 5833–5850

13 Singh C, Singh K, Liu K H. Fast beam training for RIS-assisted uplink communication. 2021. ArXiv:2107.14138

14 Wang J, Tang W, Jin S, et al. Hierarchical codebook-based beam training for RIS-assisted mmWave communication systems. IEEE Trans Commun, 2023, 71: 3650–3662

15 Xu Z, Pan Z H, Wang N, et al. Review on deep learning based adaptive beamforming for array antennas. J Data Acquis Process, 2025, 40: 1382–1411 [许峥, 潘子豪, 王宁, 等. 基于深度学习的阵列天线自适应波束形成研究综述. 数据采集与处理, 2025, 40: 1382–1411]

16 Yang J, Zhu W, Tao M. Deep learning for hierarchical beam alignment in mmWave communication systems. In: Proceedings of IEEE Global Communications Conference, 2022

17 Ren S, Shen K, Zhang Y, et al. Configuring intelligent reflecting surface with performance guarantees: blind beamforming. IEEE Trans Wireless Commun, 2023, 22: 3355–3370

18 Hua M, Wu Q. Joint dynamic passive beamforming and resource allocation for IRS-aided full-duplex WPCN. IEEE Trans Wireless Commun, 2022, 21: 4829–4843

19 Li J, Zhu X. Sum-rate optimization methods and analysis for reconfigurable intelligent surface aided communication system. Digital Commun Netws, 2025, 11: 1421–1435

20 Zhang Y, Alrabeiah M, Alkhateeb A. Reinforcement learning of beam codebooks in millimeter wave and terahertz MIMO systems. IEEE Trans Commun, 2022, 70: 904–919

21 Abdallah A, Celik A, Mansour M M, et al. Deep reinforcement learning based beamforming codebook design for RIS-aided mmWave systems. In: Proceedings of 2023 IEEE 20th Consumer Communications & Networking Conference (CCNC), 2023. 1020–1026

22 Tandler D, Dorner S, Gauger M, et al. Deep reinforcement learning for mmWave initial beam alignment. In: Proceedings of the 26th International ITG Workshop on Smart Antennas and 13th Conference on Systems, Communications, and Coding, 2023. 1–6

23 Lin J, Zhou Y, Dong X, et al. Deep reinforcement learning for robust beamforming in IRS-assisted wireless communications. In: Proceedings of IEEE Global Communications Conference, 2020. 1–6

24 Ning X, Zeng M, Hua M, et al. Multiple reconfigurable intelligent surfaces aided vehicular edge computing networks: a MAPPO-based approach. IEEE Trans Veh Technol, 2024, 73: 17496–17509

25 Puspitasari A A, Lee B M. A survey on reinforcement learning for reconfigurable intelligent surfaces in wireless communications. Sensors, 2023, 23: 2554

26 Wu X Q, Yan X F, Wei M Q, et al. Deep reinforcement learning model for job shop scheduling problems with uncertainty. J Data Acquis Process, 2024, 39: 1517–1531 [吴新泉, 燕雪峰, 魏明强, 等. 基于深度强化学习的不确定作业车间调度方法. 数据采集与处理, 2024, 39: 1517–1531]

Design of a world-model-assisted reinforcement learning method for RIS adaptive beam alignment

Zonghao ZHANG[1], Feng XING[1,2], Zifan WANG[1], Haowen QIN[1], Shu XU[3,1], Chunguo LI[1*] & Luxi YANG[1]

1. *National Mobile Communications Research Laboratory, Southeast University, Nanjing 210096, China*
2. *Fusion Innovation Center, China Mobile Communications Group Yunnan Company Limited, Kunming 650100, China*
3. *School of Electronic and Optical Engineering, Nanjing University of Science and Technology, Nanjing 210094, China*
* Corresponding author. E-mail: chunguoli@seu.edu.cn

Abstract This paper investigates adaptive beam alignment for large-scale reconfigurable intelligent surface (RIS)-aided millimeter-wave communication systems under imperfect channel state information (CSI). Existing RIS beam alignment methods mainly rely on predefined codebook search or online reinforcement learning exploration, which suffer from high pilot overhead, slow cold-start exploration, and strong coupling between two-dimensional phase variables. To address these problems, a two-stage beam alignment method, termed world-model-assisted twin delayed deep deterministic policy gradient (WM-TD3), is proposed with world-model prior guidance and online deep reinforcement learning fine-tuning. First, the spatial structure of the RIS cascaded channel is exploited to decompose the high-dimensional phase optimization problem into horizontal and vertical low-dimensional subproblems. Then, a cognitive world model learns latent channel states from wide-beam amplitude and phase observations, and provides initial beam priors and reward-estimation baselines for online learning. Furthermore, a decoupled dual-agent TD3 (DDA-TD3) mechanism fine-tunes the horizontal and vertical phase variables through a shared differential reward related to the end-to-end signal-to-noise ratio (SNR). Simulation results show that WM-TD3 achieves faster beam alignment and more stable received SNR performance with lower search overhead than hierarchical search and pure TD3.

Keywords reconfigurable intelligent surface, millimeter-wave communication, beam alignment, world model, deep reinforcement learning, twin delayed deep deterministic policy gradient (TD3)