具身智能驱动的多 AMR 通信与控制协同优化研究

罗如瑜[1]，高天润[1]，王嘉诚[2]，田辉[1]，张平[1]
（1. 北京邮电大学网络与交换技术全国重点实验室，北京 100876；2. 新加坡南洋理工大学计算与数据科学学院，新加坡 639798）

摘　要：为满足 6G 工业场景中多自主移动机器人（AMR）在高频、多阶段任务下的通信与控制协同需求，研究了一种融合通信增强与路径协同感知的具身智能系统。该系统由具身交互层、通信协同层与智能控制层构成，分别负责局部环境感知、下行链路增强与调度，以及集中式策略学习与在线更新。为克服传统分离式方法在动态场景依赖先验建图、通信与控制割裂等局限，引入非正交多址接入（NOMA）与多天线波束赋形耦合机制，并将任务调度、路径规划与通信资源分配统一建模为“任务–控制–通信”的联合优化问题。针对高维耦合与动态时变环境，在三层协同架构上采用分层深度强化学习方法，并行收集、集中训练和分布执行。仿真结果表明，在无环境先验条件下，所提架构能快速构建可达性与信道估计图，并在多种障碍物设置下保持较高任务完成率与通信速率，有效提升系统鲁棒性与资源利用效率。

关键词：具身智能；自主移动机器人；非正交多址接入；深度强化学习；轨迹控制
中图分类号：TN929.5
文献标志码：A
DOI：10.11959/j.issn.1000–436x.2025182

Embodied intelligence-driven collaborative optimization of communication and control in multi-AMR systems

LUO Ruyu[1], GAO Tianrun[1], WANG Jiacheng[2], TIAN Hui[1], ZHANG Ping[1]
1.State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China
2.College of Computing and Data Science (CCDS), Nanyang Technological University, Singapore 639798, Singapore

Abstract: To meet communication-control co-design requirements of multiple autonomous mobile robots (AMR) performing high-frequency, multi-stage tasks in 6G industrial scenarios, an embodied intelligence system was developed to fuse communication enhancement and trajectory-aware perception. The system was composed of three layers, including embodied interaction, communication collaboration, and intelligent control, which were responsible for local sensing, downlink enhancement and scheduling, and centralized policy learning with online updates, respectively. To overcome the limitations of traditional separated approaches that rely on prior maps and separate communication from control, non-orthogonal multiple access (NOMA) was combined with multi-antenna beamforming. Meanwhile, task scheduling, path planning, and communication resource allocation were jointly formulated as a unified task-control-communication problem. To handle the high-dimensional coupling and environmental dynamics, a hierarchical deep reinforcement learning (DRL) framework with parallel sampling, centralized training and distributed execution was employed. Simulation results show that the proposed architecture rapidly constructs accurate accessibility and channel estimation maps without prior knowledge, and maintains high task completion rates and communication performance across diverse obstacle settings, demonstrating its robustness and efficiency.

Keywords: embodied intelligence, AMR, NOMA, DRL, trajectory control

收稿日期：2025–08–05；修回日期：2025–10–17
通信作者：田辉，tianhui@bupt.edu.cn
基金项目：国家自然科学基金资助项目（No.62071068）；北京邮电大学博士生创新基金资助项目（No.CX2023144）
Foundation Items：The National Natural Science Foundation of China (No.62071068), BUPT Excellent Ph.D. Students Foundation (No.CX2023144)

0 引言

随着具身智能范式的持续演进，智能体通过与环境的交互不断增强任务感知、决策演化与自适应控制能力，为复杂动态场景下的高自主多智能体系统提供了理论与技术支撑[1-2]。自主移动机器人（AMR, autonomous mobile robot）作为具身智能的重要实现形式，已在制造、仓储、物流等典型工业场景中广泛部署，尤其在承担高频次和并发性强的协同搬运任务中展现出关键作用[3]。不同于以固定导引与预设路线为主的自动导引车，AMR面向半/非结构化环境运行，依托自建图与多传感器实现自主定位与动态避障，可在未知地图与时变约束下在线重规划，因而其决策问题呈现出局部可观、强不确定与跨层耦合的特征。在室内工厂环境中，障碍物空间分布往往非均匀，且存在按节拍往返的非任务实体，如穿梭车、堆垛机等，造成通道阶段性占用与遮挡[4]。作业空间的可达区域随时间在开放/封闭之间切换，并叠加其他机器人即时占用导致明显的时变与波动，从而对路径规划与调度提出了更严格的实时性与鲁棒性要求[5]。同时，任务到达的高并发性与不确定性进一步加剧了系统中感知、规划与控制的耦合性，提升了对泛在协同与低时延控制的要求[6-7]。因此，亟须面向6G工业互联网需求的具身智能驱动多AMR系统，在复杂动态环境中实现对高频搬运任务的稳定支撑与智能响应。

为支撑高频搬运任务，多AMR系统需协同解决任务分配、轨迹规划与通信调度三大核心问题。在任务分配方面，文献[8]提出了基于李雅普诺夫优化的联合任务调度和资源分配策略，以降低控制时延并提升系统能效。在轨迹规划方面，文献[9]结合智能反射面提出了面向城市建筑遮挡环境的单无人机路径优化方法，以提升信道覆盖能力。文献[10]则从单机器人避障拓展至多机器人间的动态协同控制。在通信调度方面，文献[11]探讨了非正交多址接入（NOMA, non-orthogonal multiple access）技术在移动边缘计算中的卸载能量最小化问题。尽管上述研究均取得进展，但主流方法多采用分离建模与独立优化，难以刻画任务–控制–通信之间的状态依赖与策略耦合[8-11]，在多障碍、多任务和多AMR协同场景下难以保证端到端性能与系统自动化[11]。近年来，通信—控制协同的探索逐步涌现[6,12-13]。具体地，文献[10]设计了静态工业场景下的多机器人轨迹规划机制，在可达性约束下兼顾通信时延需求。文献[13]在车联网编队控制中，将通信时延作为主要扰动，采用动态拓扑与次优解析搜索方法，实现稳定性与车距的权衡。然而，室内多AMR场景的未知动态障碍会诱发链路跳变，叠加任务随机到达带来的非平稳性与局部观测限制，需在“任务–控制–通信”之间处理混合可行域，并通过跨层特征共享实现在线协同与端到端性能提升。

近年来，深度强化学习（DRL, deep reinforcement learning）在多机器人高维动态决策中展现出良好的自适应与泛化潜力[14]。在轨迹控制方面，DRL算法能实现机器人间的路径协同与避障优化，但多依赖静态或弱动态环境假设，缺乏对障碍演化的显式建模与在线适应机制[15-16]。在通信资源分配方面，文献[17-18]通过构建双时间尺度或分层结构的DRL框架，优化功率分配、任务卸载与传输调度，但在连续任务、频繁路径切换与链路时变的多AMR系统中面临时延与稳定性挑战[19]。此外，直接将交互经验输入策略网络，缺乏状态建模与安全机制，易在高动态场景下导致策略失效与性能退化[20]。因此，迫切需要面向“任务–控制–通信”耦合的联合感知与分层策略学习机制，以提升DRL算法在多机器人系统中的决策鲁棒性与执行可靠性。

针对上述挑战，本文面向高频搬运任务与多AMR系统通信需求，构建融合任务调度、轨迹控制与通信资源分配的一体化优化模型，引入功率域NOMA与多天线波束赋形机制，建模任务生成调度、译码顺序与路径关联特性。不同于“拆分式”独立求解，本文设计了“具身交互–通信协同–智能控制”三层架构，采用集中训练和分层执行的多智能体DRL，并通过虚拟地图在线更新弥补局部可观缺失，实现协同策略学习与闭环控制。本文主要贡献如下。

1) 面向动态工业搬运场景，构建多AMR组成的具身智能系统，引入多天线NOMA机制以增强下行通信性能。在此基础上，系统化建模任务分配、轨迹规划、波束设计与功率控制的耦合关系，提出“任务–控制–通信”联合优化问题，以最大化任务完成数和总通信速率。

2) 将联合优化问题转化为马尔可夫决策过程，

设计具身交互、通信协同与智能控制三层协同架构。进一步提出集中训练和分布执行的多智能体DRL框架，实现环境感知、策略决策与模型演化的闭环体系，支撑复杂环境下的在线适应与多智能体协同。

3) 仿真结果表明，所提框架在无先验条件下具备快速建图与路径调整能力，在相同时间内拥有更高任务完成数，所用NOMA机制在链路性能上显著优于对比算法，验证了本文方法的有效性与鲁棒性。

1　系统模型与问题构建

1.1　系统模型

考虑典型的工业制造场景，作业环境包含固定墙柱、静态障碍物和动态障碍物，任务执行空间因此具有较高不确定性。本文构建一个由K台自主搬运机器人与一个配备L根天线的接入点组成的多AMR协同系统，如图1所示。整个任务调度过程在离散时隙集合$\mathcal{T}=\{1,2,\cdots,T\}$上持续推进，总搬运机器人集合记为$\mathcal{K}=\{1,\cdots,k,\cdots,K\}$。每台机器人执行搬运任务的过程中与环境交互，接入点负责下行链路增强与调度，云端负责集中训练与参数下发。在时隙t，搬运机器人k的空间位置表示为$\boldsymbol{q}_k(t)=(x_k(t),y_k(t))$，并以固定速度$v$移动。为刻画障碍物，本文将坐标$(x,y)$的时变占用拆分为：静态障碍物集合$A_{\mathrm{I},t}(x,y)$、周期性移动实体集合$A_{\mathrm{II},t}(x,y)$和非周期性移动机器人集合$\boldsymbol{Q}_t(x,y)=\{\boldsymbol{q}_k(t)|\forall k\in\mathcal{K}\}$。考虑安全距离$r_{\mathrm{safe}}$与欧氏邻域$\mathcal{B}\triangleq\{u\in\mathbb{R}^2:\|u\|_2\leqslant r_{\mathrm{safe}}\}$，定义$A_t(x,y)=(A_{\mathrm{I},t}\oplus\mathcal{B})\cup(A_{\mathrm{II},t}\oplus\mathcal{B})\cup(\boldsymbol{Q}_t\oplus\mathcal{B})$为时变坐标占用集合。据此，为进一步刻画任务空间中不同时隙的通信状态，当且仅当$(x,y)\notin A_t(x,y)$时，可达性函数$\alpha_t(x,y)=1$；否则$\alpha_t(x,y)=0$。

系统在每个时隙t以固定概率生成新任务，并维护任务集合$\mathcal{M}(t)=\{1,2,\cdots,M(t)\}$。定义任务完成指示变量$\delta_m(t)\in\{0,1\}$，其中$\delta_m(t)=1$表示任务$m$在时隙$t$前已完成，否则$\delta_m(t)=0$。在每个时隙$t$，系统已完成任务集合记为$\mathcal{M}_{\mathrm{done}}(t)=\{m|\delta_m(t)=1\}$，未完成任务集合记为$\mathcal{M}_{\mathrm{remain}}(t)=\{m|\delta_m(t)=0\}$。此外，定义二值变量$\beta_{k,m}(t)\in\{0,1\}$，当且仅当任务$m$在时隙$t$被分配给搬运机器人$k$执行时$\beta_{k,m}(t)=1$。为保证搬运机器人–搬运任务分配唯一性，分配约束写为$\sum_{m=1}^{M}\beta_{k,m}(t)\leqslant 1$，$\sum_{k=1}^{K}\beta_{k,m}(t)\leqslant 1$。任务具有最大完成时限$T_{\mathrm{task}}$，当搬运机器人$k$从当前位置$\boldsymbol{q}_k(t)$出发先到达任务起点$\boldsymbol{q}_m^{\mathrm{start}}$再将目标搬运至终点$\boldsymbol{q}_m^{\mathrm{end}}$的最短可行路径时长不超过剩余时限时，认为任务已完成并更新$\delta_m(t)$。即满足$\boldsymbol{q}_k(t')=\boldsymbol{q}_m^{\mathrm{start}},\boldsymbol{q}_k(t)=\boldsymbol{q}_m^{\mathrm{end}},t-t'\geqslant T_{\mathrm{task}}$，且当$\beta_{k,m}(t'')=1,\forall t'\leqslant t''<t$时，判定搬运机器人$k$已完成任务$m$，并更新任务完成指标为$\delta_m(t)=1$。

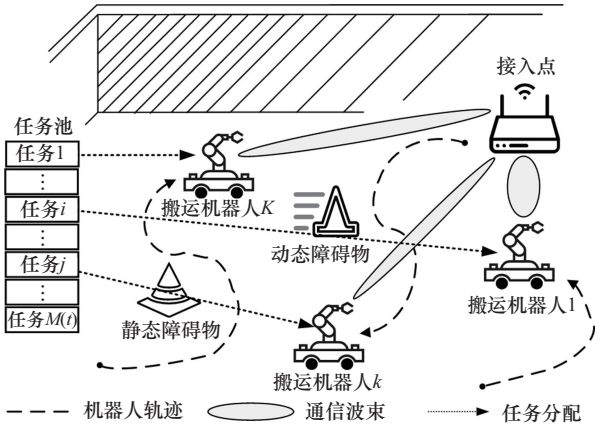

图1　工业场景下多AMR协同控制与通信网络

在整个搬运任务过程中，接入点通过下行无线链路向K台搬运机器人同时传输控制信号与任务数据。为描述搬运机器人与接入点之间的无线传输特性，考虑路径损耗和小尺度衰落的共同影响，在每个时隙t，搬运机器人k与接入点之间的传播路径损耗（单位为dB）可表示为

$$L_k(t)=\mathrm{PL}(t)-20\lg(h_k(t)),\forall k,t \tag{1}$$

本文采用基于轴对齐碰撞箱的碰撞检测算法[21]对搬运机器人k当前位置与接入点之间的直线路径进行遮挡判断，以识别搬运机器人处于视距（LoS, line-of-sight）或非视距（NLoS, non-line-of-sight）信道。基于遮挡判断，在LoS信道下小尺度衰落$h_k(t)$服从莱斯分布；在NLoS信道下退化为瑞利分布。

为描述室内遮挡与路径损耗，本文采用基于LoS/NLoS分段的对数路径损耗模型[22]，定义为

$$\mathrm{PL}_k(t)=\begin{cases}A_{\mathrm{LoS}}+B_{\mathrm{LoS}}\lg(d_{k,A}(t)),\mathrm{LoS}\\A_{\mathrm{NLoS}}+B_{\mathrm{NLoS}}\lg(d_{k,A}(t)),\mathrm{NLoS}\end{cases},\forall k,t \tag{2}$$

其中，$d_{k,A}(t)$表示搬运机器人k与接入点之间的欧

氏距离。参数$A_{\mathrm{LoS/\,NLoS}}$表示与载频相关的频率依赖截距项（包含单位距离参考自由空间损耗及场景附加常数），$B_{\mathrm{LoS/\,NLoS}}$与路径损耗指数对应，反映不同遮挡状态下损耗随距离增长的速率。在时隙t，搬运机器人k与接入点之间的信道向量$\boldsymbol{h}_k(t)$可表示为

$$\boldsymbol{h}_k(t)=\beta_k(t)(\boldsymbol{\alpha}_k(t))^{\mathrm{H}},\forall k,t \tag{3}$$

其中，$\beta_k(t)=10^{-\frac{L_k(t)}{20}}$为路径损耗因子，即式(2)中路径损耗$L_k(t)$由分贝域到线性域的转换。此外，定义$\boldsymbol{\alpha}_k(t)=[1,\mathrm{e}^{\mathrm{j}(\frac{2\pi}{\lambda})d\cos\theta_k(t)},\cdots,\mathrm{e}^{\mathrm{j}(L-1)(\frac{2\pi}{\lambda})d\cos\theta_k(t)}]^{\mathrm{T}}$为均匀线阵的阵列响应，其中$\theta_k(t)$为接入点指向搬运机器人$k$的出射角，$\lambda$为载波波长，$d$为阵列间距。

为支持多台机器人在有限时频资源下的同时频高效通信，本文引入功率域NOMA技术。接入点采用功率叠加技术在同一时频段内并行发送多路机器人信号。搬运机器人通过串行干扰消除（SIC, successive interference cancellation）技术完成解码。在任意时隙t，搬运机器人k的接收信号为

$$\boldsymbol{y}_k(t)=(\boldsymbol{h}_k(t))^{\mathrm{H}}\sum_{j=1}^{K}\sqrt{p_j(t)}\,\boldsymbol{w}_j(t)s_j(t)+\boldsymbol{n}_k(t),\forall k,t \tag{4}$$

其中，对于搬运机器人j，$p_j(t)$表示接入点分配的发射功率，$\boldsymbol{w}_j(t)$表示接入点发送的波束方向向量，$s_j(t)$表示其发射功率归一化后的发射符号，$\boldsymbol{n}_k(t)\sim\mathcal{CN}(0,\sigma^2)$表示复高斯白噪声。在任意时隙$t$，接入点的总发射功率不得超过最大功率$P_{\max}$，即需满足发射功率约束$\sum_{k=1}^{K}p_k(t)\leqslant P_{\max}$。

在功率域NOMA系统中，为保证所有搬运机器人能够成功执行SIC操作，接入点需合理分配搬运机器人的发射功率，以保证解码顺序的正确性和信号的可分离性。具体而言，解码顺序按照等效信道增益从高到低排列，即等效信道增益强的机器人优先被解码，后续机器人继续解码剩余信号。令搬运机器人k的等效信道增益为$g_k(t)=|(\boldsymbol{h}_k(t))^{\mathrm{H}}\boldsymbol{w}_k(t)|^2$。不失一般性，假设在时隙$t$系统的等效信道增益满足$g_K(t)\leqslant\cdots\leqslant g_2(t)\leqslant g_1(t)$，则搬运机器人$k$在解码时将采用SIC技术，依次消除等效信道增益较弱的搬运机器人$\{(k+1),\cdots,(K-1),K\}$干扰。此外，为保证搬运机器人k的信号可在搬运机器人$(k-1)$处被成功检测并消除干扰，传输功率需满足式(6)所示约束。

$$p_k(t)\sqrt{\hat{g}_{k-1,k}(t)}-\sum_{j=1}^{k-1}p_j(t)\sqrt{\hat{g}_{k-1,j}(t)}\geqslant\rho_{\min},\forall k\geqslant 2,t \tag{5}$$

其中，$\hat{g}_{k-1,j}(t)=|(\boldsymbol{h}_{k-1}(t))^{\mathrm{H}}\boldsymbol{w}_j(t)|^2$为机器人$(k-1)$对机器人$j$信号的等效接收增益。实现最小可检测功率差所需的下限可表示为$\rho_{\min}=P_{\mathrm{tol}}N_0B$，其中$P_{\mathrm{tol}}$为信号判别所需的最小功率间隔，$N_0$为噪声功率谱密度，$B$为系统带宽。综上所述，在时隙$t$搬运机器人$k$与接入点之间的可达下行速率可表达为

$$R_k(t)=\mathrm{lb}\left(1+\frac{p_k(t)g_{k,k}(t)}{\sum_{j=1}^{k-1}p_j(t)g_{k,k}(t)+|\boldsymbol{n}_k(t)|^2}\right),\forall k,t \tag{6}$$

此外，所有搬运机器人k在任意时刻t都需要满足最低通信速率约束$R_k(t)\geqslant R_{\min}$。

1.2 目标优化问题

本文遵循“任务–控制–通信”协同思想，联合设计任务分配$\boldsymbol{\beta}(t)=\{\beta_{k,m}(t)|\forall k\in\mathcal{K},m\in\mathcal{M}(t)\}$、多机器人路径规划$\boldsymbol{Q}(t)=\{\boldsymbol{q}_k(t)|\forall k\in\mathcal{K}\}$、NOMA下行发射功率分配$\boldsymbol{p}(t)=\{p_k(t)|\forall k\in\mathcal{K}\}$与接入点发射波束矩阵$\boldsymbol{W}(t)=\{\boldsymbol{w}_k(t)|\forall k\in\mathcal{K}\}$，旨在最大化总任务完成数$\mathcal{M}_{\mathrm{done}}(T)$和总通信速率$\sum_{k=1}^{K}R_k(t)$，以实现任务高效分配、多台机器人轨迹控制与通信资源分配的协同优化。优化问题如式(7)所示。

$$\begin{aligned}
&\mathrm{var}\{\boldsymbol{\beta}(t),\boldsymbol{Q}(t),\boldsymbol{p}(t),\boldsymbol{W}(t)\}\\
&\mathcal{P}_1:\ \max\ T\\
&\mathcal{P}_2:\ \max\Big(\sum_{k=1}^{K}R_k(t)\Big)\\
&\mathrm{C}_1:\ \beta_{k,m}(t)\in\{0,1\},\delta_m(t)\in\{0,1\}\\
&\mathrm{C}_2:\ \sum_m\beta_{k,m}(t)\leqslant 1,\ \sum_k\beta_{k,m}(t)\leqslant 1\\
&\mathrm{C}_3:\ \mathcal{M}(t+1)=\mathcal{M}(t)\setminus\{m|\delta_m(t)=1\}\\
&\mathrm{C}_4:\ \mathcal{M}(0)=\mathcal{M},\mathcal{M}(T-1)\neq\varnothing,\mathcal{M}(T)=\varnothing\\
&\mathrm{C}_5:\ \boldsymbol{q}_k(t)\in\{(x,y)|\alpha_{x,y}(t)=1\}\\
&\mathrm{C}_6:\ \sum_{k=1}^{K}\|\boldsymbol{w}_k(t)\|^2\leqslant 1\\
&\mathrm{C}_7:\ \sum_{k=1}^{K}p_k(t)\leqslant P_{\max}\\
&\mathrm{C}_8:\ R_k(t)\geqslant R_{\min}\\
&\mathrm{C}_9:\ p_k(t)>0
\end{aligned} \tag{7}$$

其中，优化目标$\mathcal{P}_1$与$\mathcal{P}_2$分别对应最大化总任务完成数和最大化总通信速率的系统指标。约束C_1为SIC译码的最小功率差要求。约束C_2与C_3确保同一机器人在同一时隙t最多执行一个任务，且每个任务在系统生命周期内仅需完成一次。约束C_4用于保证任务调度的完整性与终止条件。约束C_5定义了机器人的可达性状态，受障碍物与其他机器人影响。约束C_6为波束限制，约束C_8为最小通信速率限制，约束C_7和C_9规定了系统总发射功率的最大阈值和功率分配的非负性。

然而，该优化问题属于典型的非凸混合整数规划问题，具有高度计算复杂性。一方面，路径规划变量、任务分配状态、功率控制与波束设计之间存在强耦合关系，导致目标函数难以分解，且变量维度高、相关性强。另一方面，系统通信速率$R_k(t)$同时受多机器人位置、动态遮挡状态和小尺度信道衰落的共同影响，进而增加了系统的动态性与不确定性，使传统静态优化方法难以实时适应变化并求得全局最优解。因此，本文提出了一种具身智能学习框架，用于解决高维混合优化变量带来的挑战。

2　架构设计

2.1　算法机理描述

在动态变化的工业互联网环境中，多机器人协同不仅面临路径遮挡、信道不稳定、任务动态变化等挑战，还需要具备高精度路径规划、低时延通信与策略自适应更新能力。为此，本文提出了一种面向6G智简网络的多机器人通信–路径–任务联合优化算法架构。整体框架如图2所示，由具身交互层、通信协同层和智能控制层构成，3层分别对应环境感知、协同决策与策略演化功能。系统通过具身智能体与环境的闭环交互，实现任务完成、路径适应和通信增强的全局协同，共同构建“感知–决策–演化”的具身智能闭环体系。

首先，具身交互层由多台机器人组成，每台机器人都具备局部感知与自主路径执行能力。机器人根据云端下发的策略模型进行路径规划，同时在任务执行过程中记录障碍物坐标、碰撞事件和经过路径点等关键状态信息，并将条件触发轨迹与信道信息上报至云端，作为虚拟地图实时修正依据。其次，通信协同层通过多天线接入点实现波束赋形与NOMA联合控制，保障多机器人的实时下行通信质量与译码顺序管理。最后，智能控制层部署于云端服务器，具备高性能计算与集中训练能力，将基于其他两层的实时反馈构建并更新虚拟可达性地图与信道估计地图，并以此为基础进行智能策略训练。为缩短无效移动时间，本文设计智能控制层根据距离优先原则，将当前空闲任务分配至就近空闲搬运机器人。任务将以一定概率不断随机生成。同时，智能控制层将设计通信协同层的波束及解码顺序。

为实现“任务–控制–通信”在算法层面的内生协同，本文将联合目标形式化为“双目标–双马尔可夫链（MDP, Markov decision process）对齐+跨层耦合”的机制框架，依据信息结构与可观测性将任务侧与通信侧分别建模，并通过通信可行性门控与跨层特征共享实现目标间联动，而非单一加权和拼接。具体地，与双优化目标$\mathcal{P}_1$与$\mathcal{P}_2$对应，本

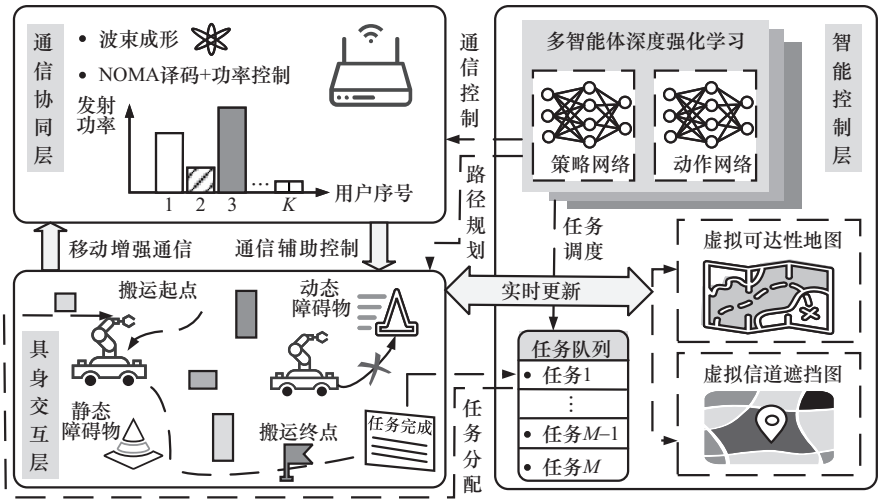

图2　融合路径规划、通信调度与任务分配的三层智能协同架构

文部署多智能体近端策略优化（MAPPO, multi-agent proximal policy optimization）算法对齐“任务完成数”，双延迟深度确定性策略梯度（TD3, twin delayed deep deterministic policy gradient）算法对齐“总通信速率”。联合性通过以下三点实现：① 以最低速率与SIC的硬门控/大惩罚将通信可行性内嵌到任务执行；② TD3决策对应的速率可行性基于MAPPO中的最小信道遮挡设计，且MAPPO的轨迹热度与任务分配作为TD3状态以优先保障关键链路；③ 训练期交替更新、执行期同步决策，避免单一加权和在不同工况下调参不稳。此外，为确保智能控制层与具身交互层的策略参数同步，仅在本地策略无法完成任务（超时/碰撞超阈/适配不足）时按需重训并版本化分发。新版本仅在触发条件满足时生效，未达条件沿用旧版，云端保留最近版本以便快速回退。

2.2 智能控制层的半分布式策略训练机制

为实现多搬运机器人系统在动态复杂环境中的通信–路径–任务联合优化，本文构建了“智能控制层训练、其余两层执行”的分层式DRL框架。智能控制层作为全系统的策略中枢，承担全局策略建模、集中训练与在线演化等关键功能，是系统“感知–决策–演化”闭环的核心组成部分。在每个决策周期内，智能控制层汇聚来自具身交互层的路径感知数据与链路反馈信息，并动态更新虚拟地图，以提升DRL算法对未知障碍与复杂环境的适应能力。具身交互层中机器人等本地设备端主要执行前向推理，避免网络训练带来的高算力负担和安全风险。

本文中MDP统一记为$\mathcal{M}=<\mathcal{S},\mathcal{A},\mathcal{P},\mathcal{R}>$，其中$\mathcal{S}$为状态空间，$\mathcal{A}$为动作空间，$\mathcal{P}$为转移概率，且$\mathcal{R}$为回报函数。考虑到式(7)的高维复杂性与分布异构性，本文采用两层分解：将原问题拆分为分布式MDP轨迹控制子问题与全局MDP任务分配与通信子问题，分别对应机器人本体的局部决策与云端的统一调度。该解耦方式是系统可行解决路径之一，但该框架不局限于该结构，具备进一步扩展至多层多尺度结构的潜力。

2.2.1 基于虚拟地图的多智能体轨迹规划策略

为避免时间尺度混淆，本文显式区分两类时间索引：系统实时时隙t，指真实系统运行过程中的离散时刻（任务生成、机器人上报、策略下发与本地执行均发生在该时间轴上）；智能控制层训练轮次τ，指智能控制层本地机器人模型执行失败后，基于虚拟地图快照进行的多轮仿真交互与策略优化过程（MAPPO的迭代轮次）。

1) 虚拟地图更新

在路径控制策略学习过程中，智能控制层依托云端算力，基于动态构建栅格步长为Δ_S的虚拟可达性地图与虚拟信道遮挡图，部署MAPPO算法进行训练。具体地说，上述虚拟图均被划分为长宽Δ_S的网格。同时，在智能控制层初始化时，本文考虑最困难的情况，即未知任何障碍物，所有位置为$\rho(\boldsymbol{q})=1$。机器人在执行过程中记录碰撞位置、碰撞时间点以及链路质量判定，并采用事件驱动的方式上报：① 下行平均速率连续5 s低于最小保障阈值即上报；② 当前移动任务完成后必上报；③ 新碰撞次数超过阈值Δ_{colli}；④ 当前任务执行超过时间限制T_{task}。

一旦机器人k触发上述事件驱动上报，则智能控制层会执行以下流程。首先，设本次任务执行过程中的碰撞发生时间集合为$\mathcal{T}_{\text{col}}^k$，上传所有碰撞位置$\boldsymbol{q}\in\{\boldsymbol{q}_k(t)|t\in\mathcal{T}_{\text{col}}^k\}$，则可达性估计函数更新为

$$\rho(\boldsymbol{q})\leftarrow\min(0.99,\rho(\boldsymbol{q})+\eta_{\text{I}}),\forall\boldsymbol{q} \tag{8}$$

其中，$\eta_{\text{I}}\in(0,1)$为碰撞惩罚因子。其次，在本次任务执行过程中，任务执行路径为$\mathcal{Q}^k=\{\boldsymbol{q}_k(t)|t\in[t_{\text{s}}^m,t_{\text{e}}^m]\}$，其中$t_{\text{s}}^m$和$t_{\text{e}}^m$分别为机器人执行任务$m$的起始时间和结束时间。对于机器人$k$执行过程中的其他位置$\boldsymbol{q}\in\{\boldsymbol{q}_k(t)|t\in[t_{\text{s}}^m,t_{\text{e}}^m],t\notin\mathcal{T}_{\text{col}}^k\}$，可达性估计函数更新为$\rho(\boldsymbol{q})\leftarrow\max(0,\rho(\boldsymbol{q})-\eta_{\text{II}})$，其中$\eta_{\text{II}}\in(0,\eta_{\text{I}})$为通过概率因子，保证智能控制层对动态障碍物的感知。最后，上传机器人清空本地轨迹记录。与不可直接全面观测的真实环境可达性$\alpha_t(x,y)$不同，可达性估计函数$\rho(\boldsymbol{q})$是智能控制层基于观测结果给出的估计，二者之间的近似度随观测密度增加而提升。

考虑到LoS/NLoS状态在可获得性、时空稳定性及对通信质量的重要性方面相较瞬时信道指标更具优势，本文采用该状态作为轨迹规划的辅助信息，该假设在未来可由环境感知或信道特征提取模块实现，具有一定通用性。基于上传轨迹中LoS/NLoS信道状态测量，虚拟信道遮挡图更新为

$$h(\boldsymbol{q})\leftarrow(1-\chi)h(\boldsymbol{q})+\chi\varpi(\boldsymbol{q}),\forall\boldsymbol{q}\in\mathcal{Q}^k \tag{9}$$

其中，$\chi \in (0,1)$为更新步长。如果当前轨迹为LoS链路，则$\varpi(\boldsymbol{q}) = \max\big(h(\boldsymbol{q}) - \delta,0\big)$；反之，如果当前轨迹为NLoS链路，则$\varpi(\boldsymbol{q}) = \min\big(h(\boldsymbol{q}) + \delta,0\big)$，即$h(\boldsymbol{q})$每次最多向1（或向0）推进$\chi\delta \in (0,1)$，直至大于1或小于0。若需要更细粒度和更平滑的地图演化，可适当减小χ和δ。初始虚拟信道遮挡图均置0.5。基于所有机器人的不定期路径与碰撞反馈，智能控制层将在已有模型失效时，以更新后的地图作为输入，重训练或微调得到最新策略模型，构成“环境交互–反馈采集–策略演化”的闭环流程。

2) 分布式部分可观测MDP构建

当机器人k在实时时隙t上报其当前位置与轨迹、链路等信息后，智能控制层在该时隙形成的虚拟可达性地图$\rho(\boldsymbol{q})$和虚拟信道遮挡图$h(\boldsymbol{q})$中的静态快照上，构造虚拟智能体k，并建立以上传时刻t的状态$\boldsymbol{s}_k^t = [\boldsymbol{q}_k(t),\boldsymbol{q}_k^{\text{target}}]$为起点的分布式部分可观测MDP，即$\mathcal{M}_k =< \mathcal{S}_k,\mathcal{A}_k,\mathcal{P}_k,\mathcal{R}_k >$。如果机器人$k$在当前时隙$t$正在执行任务$m$且机器人未到达任务$m$的起点时，则任务目标$\boldsymbol{q}_k^{\text{target}}$为任务$m$的起点$\boldsymbol{q}_m^{\text{start}}$，否则为任务$m$的终点$\boldsymbol{q}_m^{\text{end}}$。定义该分布式部分可观测MDP的状态空间、动作空间与奖励函数如下。

状态空间$\mathcal{S}_k$：在任意训练轮次τ，智能体k的当前状态由机器人位置和任务目标$\boldsymbol{q}_k^{\text{target}}$组成，定义为$\boldsymbol{s}_k^\tau = [\boldsymbol{q}_k(\tau),\boldsymbol{q}_k^{\text{target}}] \in \mathcal{S}_k$，其中$\boldsymbol{q}_k(\tau)$为虚拟智能体$k$在训练轮次$\tau$时虚拟地图的位置。

动作空间$\mathcal{A}_k$：基于当前轮次位置$\boldsymbol{s}_k^\tau$，虚拟智能体k选择移动动作$a_k^\tau = [\Delta\boldsymbol{q}_k(\tau)] \in \mathcal{A}_k$，并到达下一状态$\boldsymbol{q}_k(\tau+1)$。

奖励函数$\mathcal{R}_k$：执行动作a_k^τ后，每个虚拟智能体通过与环境的持续交互，获取奖励r_k^τ以调整虚拟智能体移动方向，使机器人在保证通信质量的前提下高效完成搬运任务。当$\beta_{k,m}(t) = 1$时，虚拟智能体k在训练轮次τ时的奖励函数r_k^τ为

$$r_k^\tau = -\alpha_{\text{I}} d_k^{\text{target}}(\tau+1) + \alpha_{\text{II}} r_k^{\text{goal}}(\tau) + \alpha_{\text{III}} r_k^{\text{LoS}}(\tau) - r_k^{\text{collision}}(\tau),\forall k,\tau \tag{10}$$

其中，$d_k^{\text{target}}(\tau)$表示虚拟智能体$k$在训练轮次$\tau$的位置与任务目标$\boldsymbol{q}_k^{\text{target}}$之间的距离。当虚拟智能体$k$到达任务目标$\boldsymbol{q}_k^{\text{target}}$时，到达奖励$r_k^{\text{goal}}(\tau) = 1$，否则$r_k^{\text{goal}}(\tau) = 0$。为体现信道遮挡的影响并避免通信死区，设置$r_k^{\text{LoS}}(\tau) = 1 - h(\boldsymbol{q}_k(\tau+1))$。此外，虚拟智能体$k$在无碰撞情况下，令$r_k^{\text{collision}}(t) = 0$；一旦该虚拟智能体与障碍物或其他虚拟智能体发生碰撞，则设置$r_k^{\text{collision}}(\tau) = -1 + \alpha_{\text{I}} d_k^{\text{target}}(\tau+1) - \alpha_{\text{II}} r_k^{\text{goal}}(\tau) - \alpha_{\text{III}} r_k^{\text{LoS}}(\tau)$。常数$\alpha_{\text{I}},\alpha_{\text{II}},\alpha_{\text{III}} > 0$为加权系数，用于将各分量归一化，并遵循“安全–到达–距离–通信”的层次化优先级。分量对应的参数越大，代表该分量在该任务中的重要性占比越大。

3) 基于MAPPO的轨迹规划算法

为基于分布式MDP解决轨迹控制子问题，本文将基于MAPPO训练路径规划策略，并采用智能控制层训练和具身交互层分布执行的高效协同控制体系。具体地说，在训练轮次τ，每个虚拟智能体k训练策略网络$\pi_k(a_k^t|s_k^t;\theta_k^\pi)$与状态值函数网络$V_k(s_k^t;\theta_k^V) \approx \mathbb{E}_{\pi_k}[\sum_{i=0}^{\infty}\gamma^i r_k^{t+i}|s_k^t]$，其中$\theta_k^\pi$和$\theta_k^V$分别为对应策略网络和值函数网络的参数，$\gamma \in (0,1)$为折扣因子。对于虚拟智能体$k$，策略网络$\pi_k(a_k^\tau|s_k^\tau;\theta_{k,\tau}^\pi)$表示在当前状态$s_k^\tau$下，选择动作$a_k^\tau$的概率。为防止策略变化过快引发性能波动，MAPPO使用截断的重要性采样比值，定义策略网络的损失函数为

$$\mathcal{L}^{\text{clip}}(\theta_{k,\tau}^\pi) = \mathbb{E}_\tau[\min(r_k^\tau\theta_{k,\tau}^\pi)A_k^\tau, \text{clip}(r_k^\tau(\theta_{k,\tau}^\pi),1-\varepsilon,1+\varepsilon)A_k^\tau],\forall k,\tau \tag{11}$$

其中，$r_k^\tau(\theta_{k,\tau}^\pi) = \pi_k(a_k^\tau|s_k^\tau;\theta_{k,\tau}^\pi)(\pi_k(a_k^\tau|s_k^\tau;\tilde{\theta}_{k,\tau}^\pi))^{-1}$，$\tilde{\theta}_{k,\tau}^\pi$为旧策略网络参数。此外，剪裁函数$\text{clip}(x,1-\varepsilon,1+\varepsilon) = \min\{\max(x,1-\varepsilon),1+\varepsilon\}$，即令$x$截断在$[1-\varepsilon,1+\varepsilon]$内。

在每轮MAPPO策略更新期间，当新旧策略网络的散度$\text{KL}(\pi_k(a_k^\tau|s_k^\tau;\theta_{k,\tau}^\pi),\pi_k(a_k^\tau|s_k^\tau;\tilde{\theta}_{k,\tau}^\pi))$过大时，提前终止该轮更新。设截断阈值为$\varepsilon$，优势函数估计$A_k^\tau$基于时间差分误差采样。在第$\tau$轮训练中，每个虚拟智能体$k$先冻结旧策略网络参数$\tilde{\theta}_{k,\tau}^\pi$，并在同一环境副本并行收集一条联合轨迹。每个虚拟智能体k得到各自的样本集合后，进行多轮小批量更新。为降低估计方差，优势向量需做零均值和单位方差标准化。此外，策略网络参数根据$\theta_{k,\tau}^\pi \leftarrow \theta_{k,\tau}^\pi + \alpha_{k,\tau}^\pi\nabla\mathcal{L}^{\text{clip}}(\theta_{k,\tau}^\pi)$进行更新，其中$\alpha_{k,\tau}^\pi$为策略网络学习率。策略网络通过最小化均方误差损失进行更新，损失函数定义为

$$\mathcal{L}^V(\theta_{k,\tau}^V) = \mathbb{E}_\tau[(V_k(s_k^\tau;\theta_{k,\tau}^V) - \hat{R}_k^\tau)^2],\ \forall k,\tau \tag{12}$$

其中，$\hat{R}_k^\tau$为累计折扣奖励。更新后策略网络$\pi_k(a_k^\tau|s_k^\tau;\theta_{k,\tau}^\pi)$将被同步下发至虚拟智能体$k$。

2.2.2 基于TD3的资源分配策略

1) 全局MDP构建

基于全局状态信息，在通信调度层建立$\mathcal{M}_{\text{global}}=<\mathcal{S}_{\text{global}},\mathcal{A}_{\text{global}},\mathcal{P}_{\text{global}},\mathcal{R}_{\text{global}}>$，以保障无线通信的可靠性。定义该全局MDP构建如下。

状态空间$\mathcal{S}_{\text{global}}$：通信智能体采用所有机器人位置信息作为全局状态信息$s^t=[(\boldsymbol{q}_k(t))_k]\in\mathcal{S}_{\text{global}}$。

动作空间$\mathcal{A}_{\text{global}}$：基于全局状态$s^t$，通信智能体设计接入点发射功率与下行波束赋形作为联合动作$a^t=\{\boldsymbol{p}(t),\boldsymbol{W}(t)\}\in\mathcal{A}_{\text{global}}$。

奖励函数$\mathcal{R}_{\text{global}}$：执行联合动作$a^t$后，通信智能体将基于当前时隙最小通信速率计算通信奖励r^t，定义为

$$r^t=\alpha_{\text{IV}}\sum_{k=1}^{K}r_k^{\text{commu}}(t)-r_{\min}^{\text{commu}}(t),\forall t \tag{13}$$

其中，$\alpha_{\text{IV}}\in(0,1)$为加权系数。对应式(7)中最小通信约束（$\text{C}_8$），当存在机器人的通信速率小于$R_{\min}$时，设置惩罚项$r_{\min}^{\text{commu}}(t)<0$；当所有机器人的通信速率均大于$R_{\min}$时，则$r_{\min}^{\text{commu}}(t)=0$。为该通信调度策略通过与环境的动态交互，自适应调整波束方向与NOMA功率分配，以有效缓解链路遮挡和多用户干扰问题。此外，如果式(7)中的SIC译码约束无法满足，则该机器人的接收信号与前序用户信号之间的功率差不足，译码器无法将其正确区分并逐级消除干扰，即当前及后续用户均无法成功译码，通信速率为零。

2) 基于TD3的资源分配算法

对应于$\mathcal{M}_{\text{global}}=<\mathcal{S}_{\text{global}},\mathcal{A}_{\text{global}},\mathcal{P}_{\text{global}},\mathcal{R}_{\text{global}}>$，在多移动机器人协同系统中，为了减少资源空闲、提升任务响应效率与调度紧凑性，本文引入一种基于距离最短匹配的任务分配策略。距离$D_{m,k}$定义为在当前虚拟可达性地图上由A^*算法得到的最短可行路径距离，栅格步长为Δ_{S}，采用8邻域连接，障碍与不可达栅格视为不可通行。具体而言，系统为每个待分配任务按距离$D_{m,k}$对候选空闲机器人升序排序，任务按序向最近的空闲机器人发起匹配申请。在任意时刻t，空闲机器人k仅保留其收到的距离最近的任务匹配申请，并拒绝其余申请。被拒绝的任务继续向其名单上的下一位空闲机器人申请匹配。当没有空闲机器人或任务全部完成时算法结束。为避免频繁重分配，距离以固定时隙刷新，并设置最小重分配间隔，并列距离时，以机器人编号打破平局。

为训练通信调度策略，本文基于TD3处理动作连续性与策略学习稳定性问题，智能控制层采用TD3算法对通信资源分配进行训练。在时隙t，基于当前联合状态s^t，TD3采用策略网络$\pi(a_{\text{comm}}^t|s^t;\theta^\pi)$生成联合通信动作$a_{\text{comm}}^t=\{\boldsymbol{p}(t),\boldsymbol{W}(t)\}$。为缓解过估计问题，TD3采用时延更新与双重估计方法，即部署2个Q网络$Q_{\text{I}}(s^t,a_{\text{comm}}^t;\theta^{Q_{\text{I}}})$和$Q_{\text{II}}(s^t,a_{\text{comm}}^t;\theta^{Q_{\text{II}}})$用于精准估计状态动作值函数。对于经验池中$e^j=(s^j,a_{\text{comm}}^t,r^j,s^{j+1})$，Q网络的损失函数定义为$\mathcal{L}^{Q_i}(\theta^{Q_i})=\mathbb{E}_{e^j}[(Q_i(s^j,a_{\text{comm}}^t;\theta^{Q_i})-y)^2],i\in\{\text{I,II}\}$。最小值约束目标定义为

$$y=r^j+\gamma\min_i Q_i(s^{j+1},\pi(s^{j+1};\theta^{\pi,-});\theta^{Q_i,-}) \tag{14}$$

其中，$i\in\{\text{I,II}\}$。策略网络的更新延迟进行，优化目标是最大化策略下的Q值，即最小化损失函数$\mathcal{L}^\pi(\theta^\pi)=\mathbb{E}_i[Q_{\text{I}}(s^i,\pi(s^i;\theta^\pi);\theta^{Q_{\text{I}}})]$，该目标反映了策略网络在状态联合$s^i$下应选择使Q网络给出最大回报的动作。TD3的策略网络更新满足$\theta_{t+1}^{Q_i}=\theta_t^{Q_i}+\hat{\alpha}_t^\pi\nabla_\theta J(\theta_t^{Q_i})$，其中策略梯度$\nabla_\theta J(\theta_t^{Q_i})=\nabla_\pi\pi(a_{\text{comm}}^t|s^t;\theta^\pi)\nabla_a Q_{\text{I}}(s^t,a_{\text{comm}}^t;\theta^{Q_{\text{I}}})$包含Q网络对动作的梯度与动作对参数的梯度乘积，且TD3用$Q_{\text{I}}(s^t,a_{\text{comm}}^t;\theta^{Q_{\text{I}}})$做策略梯度。此外，本文目标网络均采用软更新$\theta^-\leftarrow\varphi\theta+(1-\varphi)\theta^-$，其中$\varphi$为软更新系数。训练完成后，智能控制层将基于状态$s^t$，从目标策略网络中计算下一个状态下的通信设计$a_{\text{comm}}^t$，并驱动通信协同层执行具体链路控制操作。

2.3 整体架构流程与复杂度分析

2.3.1 架构流程

算法1和算法2给出了融合路径规划、通信调度与任务分配的三层智能协同训练算法伪代码，主要分为层间交互和实时决策2个部分。具体而言，算法1中层间交互部分描述系统中三层结构（具身交互层、通信协同层和智能控制层）在每个调度周期的交互过程，主要实现任务分配、策略执行与反馈采集的闭环运行。算法2是每个时间步智能算法做决策的流程，用于获得路径控制与通信调度策略，构成系统的学习交互模块。

算法1 层间交互流程

输入 待调度任务集合$\mathcal{M}(t)$，空闲机器人集合$\mathcal{K}_{\text{free}}(t)=\{k|\sum_m\beta_{k,m}(t)=0\}$，初始虚拟可达性

地图$\rho(\boldsymbol{q})$，虚拟信道遮挡图$h(\boldsymbol{q})$

输出　更新后的虚拟可达性地图$\rho(\boldsymbol{q})$，虚拟信道遮挡图$h(\boldsymbol{q})$，控制变量$\boldsymbol{\beta}(t),\boldsymbol{Q}(t),\boldsymbol{p}(t),\boldsymbol{W}(t)$

1) 对每个空闲机器人$k \in \mathcal{K}_{\text{free}}$，智能控制层指派起点距离最近的待调度任务$m^* = \arg\min_{m \in \mathcal{M}(t)} \|\boldsymbol{q}_k(t) - \boldsymbol{q}_m^{\text{start}}\|$，并更新空闲机器人集合$\mathcal{K}_{\text{free}}(t) \leftarrow \mathcal{K}_{\text{free}}(t)\backslash\{k\}$、待调度任务集合$\mathcal{M}(t) \leftarrow \mathcal{M}(t)\backslash\{m^*\}$和二值调度变量$\beta_{k,m^*}(t) = 1$（智能控制层→具身交互层）。

2) 根据当前全局状态s^t，由智能控制层TD3策略网络输出联合通信动作a_{comm}^t，并下发给通信协同层执行（智能控制层→通信协同层）。

3) 每个机器人k接收智能控制层中的MAPPO路径的分布策略网络参数$\pi_k(a_k^t|s_k^t;\theta_k^\pi)$。具身交互层根据其本地状态$s_k^t$选择移动动作$a_k^t$（智能控制层→具身交互层）。

4) 当机器人k达到碰撞次数阈值/到达规定起点开始执行任务/完成当前搬运任务后，将上传以下参数至智能控制层：当前轨迹路径点集合、通信信道状态测量、碰撞位置集合（具身交互层→智能控制层）；上传后清空本地轨迹存储。

5) 智能控制层一旦收到具身交互层上传的新轨迹，立即基于式(13)和式(14)更新虚拟可达性地图$\rho(\boldsymbol{q})$和虚拟信道遮挡图$h(\boldsymbol{q})$（智能控制层）。

6) 智能控制层基于最新$\rho(\boldsymbol{q})$和$h(\boldsymbol{q})$，测试现有机器人k的策略网络能否在规定时间到达规定任务起点或终点。如果无法到达，将基于现有MAPPO进行重训练，直到能成功到达后，下发更新后的策略网络至机器人k，记录当前时隙为t_k^{start}（智能控制层→具身交互层）。

算法2　实时决策流程

输入　实时机器人位置，当前任务分配结果，空闲任务集合

输出　搬运任务分配，多机器人路径规划，NOMA下行发射功率分配与接入点发射波束矩阵

1)for $t = 1, 2, \cdots, T$ do
2)　智能控制层基于当前全局状态s^t，输出通信动作$a_{\text{comm}}^t = \text{clip}(\mathcal{N}(0,\sigma),-c,c) + \pi^-(s^t;\theta^\pi)$
3)　到达状态s^{t+1}，获得回报奖励r^t
4)　从TD3经验池采样小批次$\left(s^i,a^i,s^{i+1},r^i\right)_i$
5)　更新TD3中策略网络和双Q网络
6)　if 机器人k空闲 do
7)　　机器人k选择任务起点距离最近的任务
8)　end if
9)　if 机器人k连续执行任务m时间超过T_{task}do
10)　　机器人k放弃该任务并上传本地轨迹数据
11)　　置机器人k空闲
12)　end if
13)　if智能控制层指派机器人k执行任务m do
14)　　if机器人k未到达任务m起点/终点do
15)　　机器人k根据本地$\pi_k(a_k^t|s_k^t;\theta_k^\pi)$选择$a_k^t$
16)　　end if
17)　end if
18)end for

2.3.2　算法收敛性分析

1) 单周期任务分配

在任务分配策略中，每个任务对每个机器人最多申请一次匹配，因此任务匹配算法至多进行$|\mathcal{M}_{\text{remain}}(t)||\mathcal{K}_{\text{free}}(t)|$次匹配操作。反证法易证出，该算法终止时得到的匹配结果$\boldsymbol{\beta}(t)$是稳定的，不存在阻挡对。因为若存在阻挡对，则任务在更早时刻已向该机器人发起申请匹配且被接纳，或由更近任务占据。此外，为度量匹配结果质量，定义任务匹配代价$c_{k,m}(t) = d(\boldsymbol{q}_k(t),\boldsymbol{q}_m^{\text{start}})$，即机器人$k$到空闲任务$m$起点的距离。基于匹配结果$\boldsymbol{\beta}(t)$，对于任意两对已匹配任务–机器人组合$(k,m),(k',m')$，如果满足以下条件：① $\max(c_{k,m}(t),c_{k',m'}(t)) \leqslant \max(c_{k,m'}(t),\ c_{k',m}(t))$；② $c_{k,m}(t) + c_{k',m'}(t) \leqslant c_{k,m'}(t) + c_{k',m}(t)$，则交换任务不会改进。记最优瓶颈值$B^* = \min_{\boldsymbol{\beta}} \max_{(k,m) \in \boldsymbol{\beta}} c_{k,m}$，则最近匹配代价$\max_{(k,m) \in \boldsymbol{\beta}(t)} c_{k,m}(t) \leqslant 2B^*$，即最坏配对的代价不超过理论最优的2倍。

2) 基于DRL的轨迹规划和通信资源分配

在多AMR轨迹规划中，本文提出的MAPPO学习率序列满足$\alpha_{k,\tau}^\pi > 0, \sum_\tau \alpha_{k,\tau}^\pi = \infty, \sum_\tau (\alpha_{k,\tau}^\pi)^2 < \infty$、$\alpha_{k,\tau}^Q > 0$，$\sum_\tau \alpha_{k,\tau}^Q = \infty, \sum_\tau (\alpha_{k,\tau}^Q)^2 < \infty$，且两时间尺符合$\dfrac{\alpha_{k,\tau}^\pi}{\alpha_{k,\tau}^Q} \to 0$。假设存在$\Delta_V(\tau) \geqslant 0$使$\sup_s \|V_{\text{approx}}(s) - V_k(s;\tilde{\theta}_{k,\tau}^V)\| \leqslant \Delta_V(\tau)$，当值函数网络充分拟合时，则$\Delta_V(\tau) \to 0$。

当τ足够大时，策略网络几乎必然收敛到ε–级一阶驻点，且任一聚点θ^*满足$\|\nabla J(\theta^*)\| \leqslant C\ln(1+\varepsilon)$，其中$C$与优势估计相关。

在通信资源分配算法方面，类似上述分段光滑、有界梯度、目标网络缓变和两时间尺度步长假设，若噪声收缩、目标网络固定，当t趋于无穷时，则Q网络拟合误差$\mathbb{E}[(y_t - Q(s^t,a^t_{\text{comm}};\theta^{Q_i}))^2]\to 0$。当策略网络$\pi$在时隙$t$达稳态时，存在有界偏差$\eta_t$，使$\mathbb{E}[\nabla_\theta J(\theta_t^{Q_i})|\mathcal{F}_t] = \nabla_\theta J(\theta_t^{Q_i}) + \eta_t$，其中$\mathcal{F}_t$为到第$t$时隙为止“已知信息”的$\sigma$–代数。进一步，若目标网络滞后按同阶衰减，则$\eta_t \to 0$，策略网络收敛到一阶驻点$\nabla_\theta J(\theta^{Q_i,*}) = 0$。综合上述收敛性分析，在冻结–更新交替与特征共享为有界扰动的设定下，按随机分块坐标下降理论，整体迭代收敛到分块驻点。

2.3.3 *算法复杂度分析*

1) 时间复杂度

在任务分配阶段，对每个任务遍历全部空闲机器人，采用最小距离匹配，时间复杂度最大为$\mathcal{O}(KM)$。当机器人k进行路径推理时，以3层全连接网络为例，则前向一次，策略网络的时间复杂度约为$\mathcal{O}(|\mathcal{S}_k|H_{\text{PPO}} + {H_{\text{PPO}}}^2 + H_{\text{PPO}}|\mathcal{A}_k|) \approx \mathcal{O}({H_{\text{PPO}}}^2)$，其中$H_{\text{PPO}}$为MAPPO中每个分布式智能体的每层隐藏维度。当进行通信规划时，TD3策略网络前向推理一次的时间复杂度为$\mathcal{O}({H_{\text{TD3}}}^2)$。此外，假设每个虚拟智能体的策略网络和值函数网络分别有J和I层，每层宽度为$u_{\pi,j}$和$u_{Q,i}$，N_π和N_Q为参与更新的策略网络与值函数/Q网络的数量，则在智能控制层每个智能体的前向推理与反向传播的时间复杂度为$O\left(N_\pi\sum_{j=0}^{J-1}u_{\pi,j}u_{\pi,j+1} + N_Q\sum_{i=0}^{I-1}u_{Q,i}u_{Q,i+1}\right)$。综上，设计的算法时间复杂度最大为

$$O((J+I)(3KT_{\text{task}}{u_{\text{PPO}}}^2 + 6T{u_{\text{TD3}}}^2) + K{H_{\text{PPO}}}^2 + KM + {H_{\text{TD3}}}^2) \tag{15}$$

其中，u为对应网络的最大层宽度。

2) 空间复杂度

对于MAPPO，每个虚拟智能体存储轨迹样本数量为$|\mathcal{D}_k|$，则空间复杂度为$O\left(\sum_{k=1}^{K}|\mathcal{D}_k|\right)$。同理，TD3的空间复杂度为$O(|\mathcal{D}|)$，其中$|\mathcal{D}|$为TD3的经验池容量，则空间复杂度约为$O\left(\sum_{k=1}^{K}|\mathcal{D}_k| + |\mathcal{D}| + (J+I)(1.5KT_{\text{task}}{u_{\text{PPO}}}^2 + 3T{u_{\text{TD3}}}^2)\right)$。

3 仿真分析

3.1 实验设置

为验证所提多机器人路径规划与通信调度的智能协同框架的有效性，本文构建了一个正方形室内工厂场景，包含3类障碍物实体：静态障碍物，即多个大小不一的不可移动物体；周期性移动设备，按给定规律往返，造成通道周期性开放与封闭；非周期移动机器人，在执行任务的过程中穿行并临时占道。因此，区域的可达性随时间动态变化。除非特殊说明，仿真默认采用$K=8$台移动机器人，场地为100 m × 100 m的栅格化环境，AP位于场地中心。考虑3.5 GHz室内通信信道[23]，所有机器人初始随机分布，将任务到达建模为以时隙为单位的Bernoulli到达。在算法方面，本文采用紧凑的多层MLP对齐常见基线并控制计算开销。设置奖励参数$\alpha_{\text{IV}} = \dfrac{0.1}{K}$，且$\alpha_{\text{I}} = \dfrac{1}{2d_{\max}}$，其中$d_{\max}$为地图的最大对角线距离。在硬件与软件环境方面，GPU平台采用NVIDIA RTX 4090（24 GB），CPU平台为AMD Ryzen 7 PRO 4750U（32 GB）。其他系统仿真参数如表1所示。

表1 系统仿真参数

参数	数值
噪声功率密度N_0/(dBm·Hz^{-1})	−174
带宽B/MHz	20
频率依赖截距项$A_{\text{LoS/NLoS}}$	43.685/57.882
损耗随距离增长的速率$B_{\text{LoS/NLoS}}$	16.9/25.5
最大碰撞阈值Δ_{colli}	10
任务执行时间限制T_{task}	300
折扣因子γ	0.99
TD3经验池容量	100 000
单PPO经验池容量	50 000
到达重要性系数α_{II}	0.5
最低通信速率要求系数α_{III}	0.01

作为对比基线，本文选取了两类算法进行比较：一是随机策略，即机器人在无路径规划和通信优化策略指导下，随机选择动作执行移动与通信；二是纯分布式PPO算法，该算法由各个机器人独立在本地训练，不依赖云端虚拟地图建模与集中式预训练过程，仅利用局部感知信息进行策略更新，代表典型的弱感知和弱协同的策略。

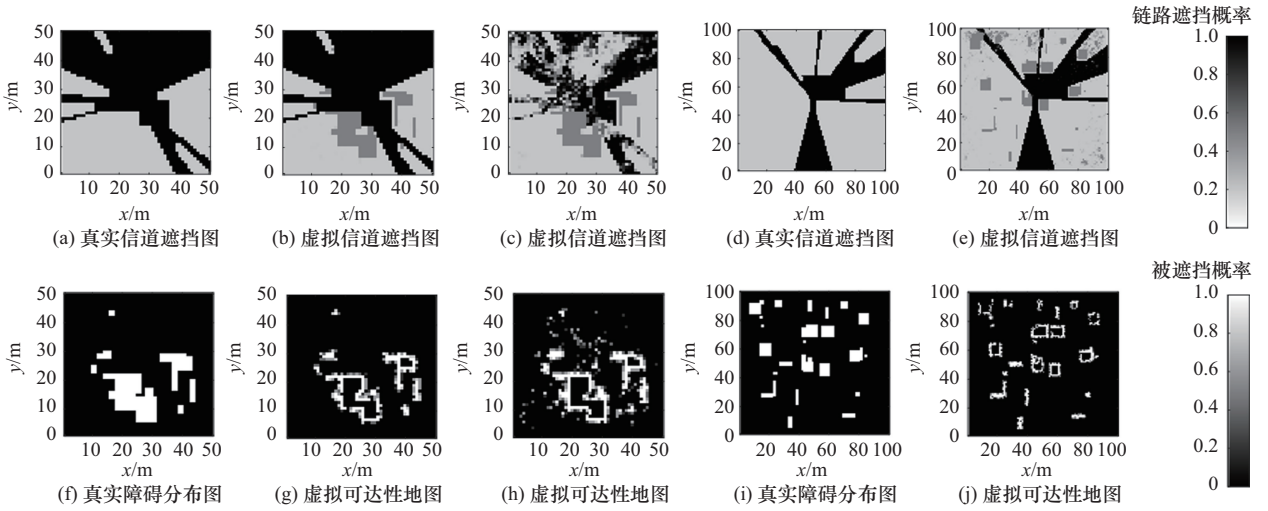

图3　不同环境设置下真实环境与虚拟环境热力图对比

3.2　结果分析

基于上述仿真实验设置，图3直观地展示了所提框架对复杂未知环境的建模与感知能力。可以看出，机器人在进行搬运任务的过程中，智能控制层能够在较短时间内学习并实时更新与真实环境高度一致的虚拟可达性地图和虚拟信道遮挡图。以50 m×50 m小尺度室内场景为例，图3(a)和图3(f)展示了仅包含静态障碍物时的真实信道分布和障碍物分布情况，对应的虚拟图如图3(b)和图3(g)所示，其可视纹理由Δ_S决定。在t=3 000时，系统已成功重构出主要遮挡区域与不可达区域，说明策略具备快速建图能力。进一步，在引入周期性动态障碍物的场景中，图3(c)和图3(h)表明系统不仅保持了对静态障碍的识别能力，还能够在动态轨迹附近以一定遮挡概率对其进行标定，体现了系统对动态场景的适应性与识别能力。对于100 m×100 m的大尺度场景，图3(d)、图3(i)、图3(e)和图3(j)展示了系统在t=20 000后仍能构建出与真实环境较为接近的估计图，但其收敛速度与整体精度相对小场景更慢，表明在大场景下环境感知与图更新的复杂度更高。

图4展示了在不同场景设置下信道相似度与可达性相似度随时隙t的演化。其中，前者通过计算虚拟与真实信道图的网格均方误差所构建，后者以真实障碍边缘与虚拟轮廓的重合度衡量。本文将两类相似度与单位时间窗口的任务完成数移动平均对齐，当三者进入窄幅稳定带时视为收敛。从图中可以看出，所提框架在初始无任何环境先验知识的前提下，在100 m×100 m的大场景中，由于探索空间更大且障碍物部署结构更复杂，收敛较50 m×50 m的小场景需要更长时间。在相同尺寸下引入动态障碍，相似度平台降低且振荡加剧，表明系统需持续更新估计并存在一定滞后。此外，信道相似度的收敛更加平滑，因为LoS/NLoS信道结构具有更强几何规律，而可达性相似度受障碍分布与运动不确定性影响波动更大。综上，所提框架在初始无任何环境先验知识的前提下，能够通过机器人在搬运任务的探索过程中，逐步构建出对实际障碍物分布和NLoS/LoS信道遮挡结构的有效估计。

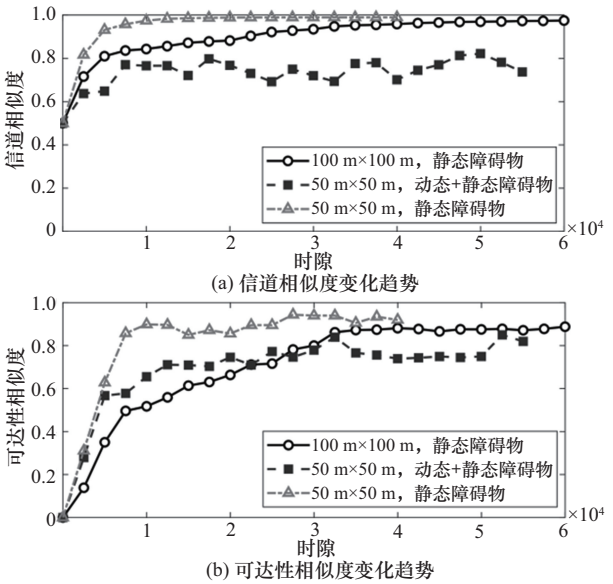

图4　信道与可达性相似度趋势对比

图5展示了不同算法在多种环境设置下的完成任务数随时隙演化规律。所有算法在初始阶段均无

任何环境先验知识，因此前期完成任务数较少。仅含静态障碍物时，场景越大收敛越慢但完成任务数仍可保持较高水平，小场景因探索空间更小而更快收敛。当在50 m×50 m场景中引入动态障碍物后，环境从近似平稳转为非平稳，障碍运动与LoS/NLoS频繁切换使优势估计方差增大、策略更新更保守，收敛速度明显下降。作为对比方案，随机策略在多障碍环境下几乎无法有效完成任务；纯分布式PPO算法受限于弱感知与本地探索，学习速度与任务完成数均劣于本文算法。对应的，本文算法引入基于虚拟地图更新的集中式训练，显著提升收敛速度与完成任务数；一旦移除虚拟地图更新，训练将出现长期振荡且难以收敛，表明该模块对时变可行域尤为关键。在多障碍环境下安全性方面，截至t=40 000，纯分布式PPO算法相比随机策略的碰撞数减少了39.65%，本文提出的协同架构减少了89.42%。在计算代价方面，GPU平台单时隙推理时延约20 ms、峰值显存约1.5 GB，训练耗时数秒至几十秒；轻量级CPU平台推理同为毫秒级、训练耗时几十秒至数分钟，内存占用为数百MB。按2.3.3节的单时隙理论推理量换算，实测吞吐与理论量级一致。综上，所提框架在动态遮挡与任务并发条件下兼具更快收敛、更高完成率与较低推理/训练开销，具备工程可行性与鲁棒性。

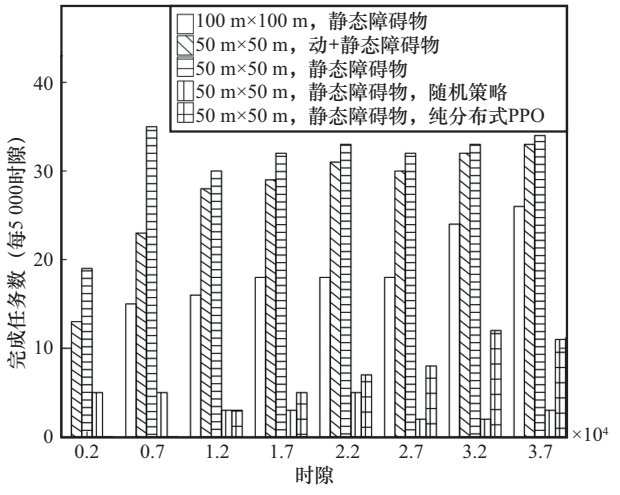

图5　不同算法与多样环境设置下完成任务数对比

图6展示了在不同用户数K、天线数L以及接入方式（NOMA/OMA）设置下，系统总数据速率随最大发射功率$P_{\max}$的趋势。从图6可以看到，随着总发射功率$P_{\max}$的提升，所有方案的系统总数据速率均呈上升趋势。同时，在相同功率条件下，增大天线数L显著提升了系统总数据速率，体现出多天线系统在波束赋形与空间复用方面的优势。在固定天线数下，机器人数量K的增加会导致单位机器人分配的资源减少，从而引起系统总数据速率增幅趋缓。此外，在相同机器人数K、天线数L与功率约束下，NOMA方案始终优于OMA方案，拥有更高的总数据速率。因此，在多机器人、固定天线数和功率受限的条件下，NOMA机制的引入能够有效提升系统通信性能。

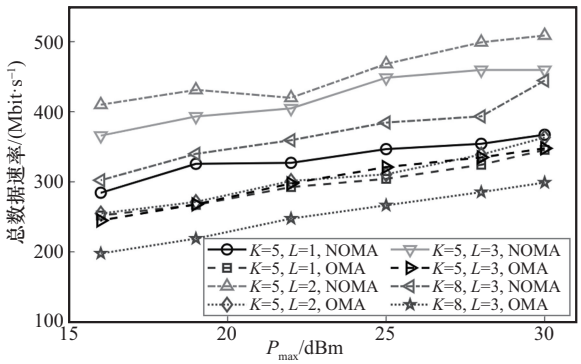

图6　不同环境设置下总数据速率对比

4　结束语

本文以多机器人系统在6G工业智简网络中协同完成连续多阶段搬运任务为研究背景，提出了一种融合通信增强与路径协同感知的具身智能系统。围绕任务调度效率与通信稳定性双目标，构建了任务–路径–通信联合优化问题，并构建了“具身交互–通信协同–智能控制”三层协同架构，以实现系统分层管理。结合功率域NOMA下行传输策略与多天线波束控制机制，本文架构采用集中训练和分布执行并结合虚拟地图的在线更新，实现了在未知时变环境下的快速感知与策略闭环演化。实验结果验证了本文方法在多障碍动态场景下具有更高的任务完成率和通信速率。未来工作将进一步探索在多任务优先级控制、异构终端协同以及大规模AMR集群中的算法可扩展性和跨域泛化能力。

参考文献：

[1] TALLAT R, HAWBANI A, WANG X F, et al. Navigating industry 5.0: a survey of key enabling technologies, trends, challenges, and opportunities[J]. IEEE Communications Surveys & Tutorials, 2024, 26(2): 1080-1126.

[2] DUAN J F, YU S, TAN H L, et al. A survey of embodied AI: from simulators to research tasks[J]. IEEE Transactions on Emerging Topics in Computational Intelligence, 2022, 6(2): 230-244.

[3] YANG H J, QIN Z Q, XIA Y Q, et al. Digital twin-based autonomous navigation and control of omnidirectional mobile robots[J]. IEEE Transactions on Vehicular Technology, 2025, 74(4): 5687-5697.
[4] ALIREZAZADEH S, ALEXANDRE L A. A survey on task allocation and scheduling in robotic network systems[J]. IEEE Internet of Things Journal, 2025, 12(2): 1484-1508.
[5] NIU X, YU C, JIN H. CRSM: computation reloading driven by spatial-temporal mobility in edge-assisted automated industrial cyber-physical systems[J]. IEEE Transactions on Industrial Informatics, 2022, 18(12): 9283-9291.
[6] 田辉, 贺硕, 林尚静, 等. 工业互联网感知通信控制协同融合技术研究综述[J]. 通信学报, 2021, 42(10): 211-221.
TIAN H, HE S, LIN S J, et al. Survey on cooperative fusion technologies with perception, communication and control coupled in industrial Internet[J]. Journal on Communications, 2021, 42(10): 211-221.
[7] 李竟博, 马礼, 李阳, 等. 感传算协同工业互联网优化设计[J]. 通信学报, 2023, 44(6): 12-22.
LI J B, MA L, LI Y, et al. Optimized design of sensing transmission and computing collaborative industrial Internet[J]. Journal on Communications, 2023, 44(6): 12-22.
[8] ZHOU F Q, FENG L, KADOCH M, et al. Multiagent RL aided task offloading and resource management in Wi-Fi 6 and 5G coexisting industrial wireless environment[J]. IEEE Transactions on Industrial Informatics, 2022, 18(5): 2923-2933.
[9] LUO R Y, TIAN H, NI W L, et al. Deep reinforcement learning enables joint trajectory and communication in Internet of robotic things[J]. IEEE Transactions on Wireless Communications, 2024, 23(12): 18154-18168.
[10] FAN X K, LIU M, CHEN Y L, et al. RIS-assisted UAV for fresh data collection in 3D urban environments: a deep reinforcement learning approach[J]. IEEE Transactions on Vehicular Technology, 2023, 72(1): 632-647.
[11] HUANG W Q, DING Z G. New insight for multi-user hybrid NOMA offloading strategies in MEC networks[J]. IEEE Transactions on Vehicular Technology, 2024, 73(2): 2918-2923.
[12] 齐俏, 陈晓明. 面向边缘智能网络的通-感-算融合：架构、挑战和展望[J]. 移动通信, 2024, 48(3): 40-46.
QI Q, CHEN X M. Integrated sensing, communication, and computing for edge intelligent[J]. Mobile Communications, 2024, 48(3): 40-46.
[13] LIU J H, ZHOU Y Q, LIU L. Communication delay-aware cooperative adaptive cruise control with dynamic network topologies: a convergence of communication and control[J]. Digital Communications and Networks, 2025, 11(1): 191-199.
[14] CHAI R Q, NIU H L, CARRASCO J, et al. Design and experimental validation of deep reinforcement learning-based fast trajectory planning and control for mobile robot in unknown environment[J]. IEEE Transactions on Neural Networks and Learning Systems, 2024, 35(4): 5778-5792.
[15] LUO R Y, NI W L, TIAN H, et al. Joint trajectory and radio resource optimization for autonomous mobile robots exploiting multi-agent reinforcement learning[J]. IEEE Transactions on Communications, 2023, 71(9): 5244-5258.
[16] ZHAO H Y, GUO Y N, LI X D, et al. Hierarchical control framework for path planning of mobile robots in dynamic environments through global guidance and reinforcement learning[J]. IEEE Internet of Things Journal, 2025, 12(1): 309-333.
[17] LIU Q Q, ZHANG H X, ZHANG X, et al. Joint service caching, communication and computing resource allocation in collaborative MEC systems: a DRL-based two-timescale approach[J]. IEEE Transactions on Wireless Communications, 2024, 23(10): 15493-15506.
[18] ZHAO T T, LI F, HE L J. DRL-based joint resource allocation and device orchestration for hierarchical federated learning in NOMA-enabled industrial IoT[J]. IEEE Transactions on Industrial Informatics, 2023, 19(6): 7468-7479.
[19] WANG H, ZHANG H J, LIU X N, et al. Joint UAV placement optimization, resource allocation, and computation offloading for THz band: a DRL approach[J]. IEEE Transactions on Wireless Communications, 2023, 22(7): 4890-4900.
[20] WU S, CHEN N, WEN G H, et al. Virtual network embedding for task offloading in IIoT: a DRL-assisted federated learning scheme[J]. IEEE Transactions on Industrial Informatics, 2024, 20(4): 6814-6824.
[21] ZHANG X Y, KIM Y J. Interactive collision detection for deformable models using streaming AABBs[J]. IEEE Transactions on Visualization and Computer Graphics, 2007, 13(2): 318-329.
[22] ZHONG R K, LIU X, LIU Y W, et al. Mobile reconfigurable intelligent surfaces for NOMA networks: federated learning approaches[J]. IEEE Transactions on Wireless Communications, 2022, 21(11): 10020-10034.
[23] LUO R Y, NI W L, TIAN H, et al. Federated deep reinforcement learning for RIS-assisted indoor multi-robot communication systems[J]. IEEE Transactions on Vehicular Technology, 2022, 71(11): 12321-12326.

[作者简介]

罗如瑜（1999– ），女，甘肃白银人，北京邮电大学博士生，主要研究方向为多智能体深度强化学习、多机器人轨迹规划、无线通信资源分配等。

高天润（1993– ），男，河南郑州人，北京邮电大学博士生，主要研究方向为联邦学习、多模态医学大模型。

王嘉诚（1992– ），男，重庆人，博士，新加坡南洋理工大学研究员，主要研究方向为低空无线网络、通感一体化网络、生成式AI、无线资源管理。

田辉（1963– ），女，河南郑州人，博士，北京邮电大学教授、博士生导师，主要研究方向为无线资源管理、智能边缘计算、多智能体协同等。

张平（1959– ），男，陕西汉中人，中国工程院院士，北京邮电大学教授、博士生导师，主要研究方向为语义通信和语用达意网络。