面向协作式多智能体强化学习的交互式蒸馏

Minwoo Cho、Batuhan Altundas 和 Matthew Gombolay 机器人与智能机器研究所，佐治亚理工学院 {mcho318, batuhan}@gatech.edu, matthew.gombolay@cc.gatech.edu

arXiv:2601.05407v1 [cs.LG] 8 Jan 2026

摘要—知识蒸馏（KD）通过采用集中式教师来训练分散式学生，有望加速多智能体强化学习（MARL），但仍面临若干关键瓶颈。具体而言，存在以下挑战：（1）在复杂领域中合成高性能教学策略；（2）教师必须在分布外（OOD）状态下进行推理时面临困难；（3）分散式学生与集中式教师的观测空间不匹配。为解决这些局限，我们提出 HINT（分层交互式教师迁移），这是一个面向集中式训练、分散式执行设定下 MARL 的新型知识蒸馏框架。通过利用分层强化学习，HINT 能够提供可扩展且高性能的教师。我们的关键创新——伪离策略强化学习——使教师策略能够利用教师和学生的经验进行更新，从而改善对分布外状态的适应能力。HINT 还采用基于性能的筛选，仅保留与结果相关的指导信息，减少观测不匹配。我们在具有挑战性的协作领域中评估 HINT（例如，用于资源分配的 FireCommander 和用于战术作战的 MARINE）。在这些基准测试中，HINT 的表现优于基线方法，成功率提升了 60% 至 165%。

一、引言

多智能体强化学习（MARL）已在各种协作任务中得到广泛研究，包括搜救行动 [25, 4],海上补给 [7,2], 和仓库管理 [17, 31], ，在这些任务中，智能体必须协调行动以实现共同目标。随着任务复杂度和团队规模的增加，人们对能够实现有效协调、同时不牺牲单个智能体自主性的方法的需求日益增长。尽管已有许多进展，但由于部分可观测性、非平稳动态以及有效信用分配等挑战，部署多智能体强化学习仍非易事。

为应对这些挑战，集中式训练、分散式执行（CTDE）已成为多智能体强化学习中的主流范式。CTDE 在训练阶段利用全局状态信息或联合观测来稳定学习，同时通过限制智能体在执行阶段仅使用局部观测来保持可扩展性。基于这一框架，许多基于CTDE的算法已在协作基准任务中取得成功 [26, 33, 42, 18]。在此基础上，人们提出了许多基于通信的方法 [34, 32, 30] ，使智能体能够共享信息、推断共享上下文并适应动态环境，从而增强协调能力。尽管这一扩展显著提升了协作能力，却也使训练更加复杂：智能体必须同时优化决策和通信，这会增加对噪声信号的敏感性以及策略梯度方差 [10]。为克服

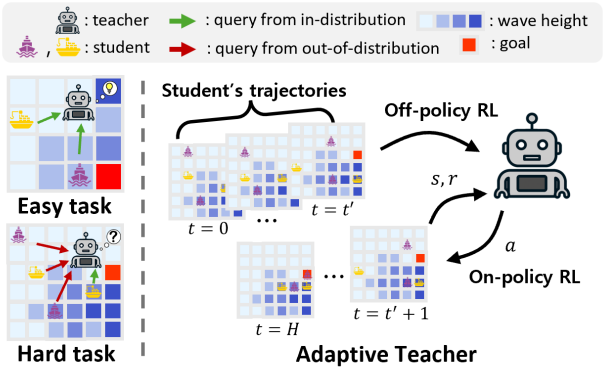

图 1. **弥合教师与学生之间的分布差距。**随着任务复杂度增加，离线训练的教师对于偏离训练分布的学生轨迹会提供不可靠的指导；我们通过自适应优化来弥合这一差距。（示例环境：MARINE）

这些局限性，我们提出了一种新颖框架，将层级监督与交互式知识蒸馏相结合，以最低的协调开销实现稳健的多智能体强化学习。

近年来，知识蒸馏（KD） [13], 作为一种颇具前景的替代方案应运而生：它让较简单的学生模型通过模仿较大型教师模型的输出进行学习，从而缓解多智能体强化学习中的不稳定训练信号和信用分配问题 [43, 3, 14, 21]。在这一范式中，能够访问全局状态的集中式教师通过最小化双方策略分布之间的差异来监督学生策略。然而，蒸馏方法面临着关键瓶颈。随着任务复杂度增加，构建可靠且性能出色的教师变得愈发困难。对于多智能体寻路等结构化任务，已有最优求解器 [5, 19],；但在多数场景中，教师仍依赖基于多智能体强化学习的教师训练，而这种方法在复杂、动态的领域中依然脆弱且样本效率低。此外，在此类环境中，即使是能力出众的教师也经常会遇到分布外的学生状态，导致示范不一致或质量低下（见图 1、第 IV 节）。最后，教师与学生观测之间的不对称会降低监督质量——教师必须动态调整指导，以适应学生上下文。

为解决这些局限性，我们提出 HINT（分层交互式教师迁移），这是一种用于有效合作式多智能体强化学习的新型知识蒸馏框架（图 2）。我们首先通过分层强化学习（RL）预训练一个集中式教师，将决策分解为两个层级，以提升可扩展性和性能。随后，去中心化

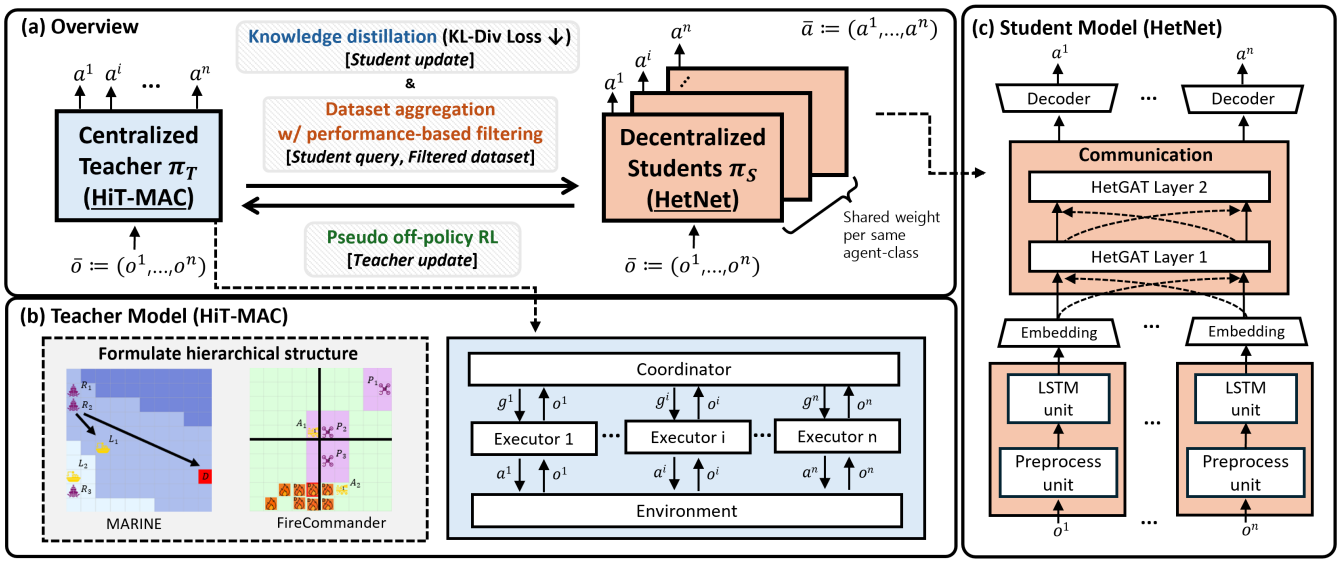

图 2. **HINT 概述**。 (a) 集中式教师通过三种机制指导去中心化学生：通过知识蒸馏更新学生，通过伪离策略强化学习改进教师，以及通过数据集聚合和基于性能的筛选响应学生查询并构建高质量数据集。(b) 教师采用分层运作方式，由高层协调器根据特定任务的层级结构向低层执行器分配子目标。该结构通过将战略决策与战术决策解耦，实现时间抽象。(c) 每个学生都包含一个预处理单元、一个用于时间抽象的 LSTM 编码器，以及一个用于选择行动的解码器；HetGAT层则支持智能体间通信。

学生策略通过知识蒸馏进行训练，并辅以在线专家查询，使训练数据与学生的测试时分布紧密对齐。HINT 的一个关键特点是伪离策略强化学习，教师在此过程中会利用自身和学生的轨迹进行更新。这能增强教师对学生行为的认知，并使其在分布外状态下提供更丰富的指导。基于性能的筛选进一步提升了数据集质量，只保留与结果相关的示范，从而减少观测不匹配。我们在两个具有挑战性的环境中测试 HINT（例如，用于资源分配的 FireCommander (FC) [28]，以及用于战术作战的 MARINE [12]），其中的随机性和时变动态带来了额外的训练难题。实验表明，HINT 的表现始终优于基线方法，验证了其在复杂领域中的鲁棒性。**我们的贡献如下：**

- 我们采用分层强化学习设计教师，以管理协调开销。这种结构提高了样本效率和教师效能，使蒸馏过程切实可行。
- 我们提出伪离策略强化学习和基于性能的筛选，使教师能够适应学生的行为并提供高质量的示范。这降低了对预言机级教师的依赖，而以往的知识蒸馏方法均假定存在这类教师。
- 我们验证了 HINT 始终优于具有竞争力的集中式训练、分散式执行和知识蒸馏基线方法，成功率提升了 60% 至 165%，展现出更强的协调能力和鲁棒性。

II. 相关工作

A. 合作式多智能体强化学习中的集中式训练、分散式执行

合作式多智能体强化学习广泛采用集中式训练、分散式执行（CTDE），通过在训练期间利用集中式价值函数来降低非平稳性带来的不稳定性，同时支持去中心化执行。这些方法分为离策略 [26, 33, 15, 38,22] 和同策略两类 [11, 42, 20, 18],，在样本效率与学习稳定性之间寻求平衡。例如，QMIX [26] 和 QTRAN [33]使用经验回放缓冲区来提升可扩展性，而多智能体近端策略优化（MAPPO） [42] 和 DAE [20] 则依靠信赖域更新来缓解梯度不稳定问题。

集中式训练、分散式执行中一个长期存在的根本挑战，是如何根据智能体对团队奖励的个体贡献，公平地分配信用。为解决信用分配问题，研究者提出了多种策略：价值分解（QMIX [26], QTRAN [33]）、基于注意力的评论家网络（MAAC [15]），以及夏普利值（SQDDPG [38]）或反事实基线（COMA [11]）等显式公平性估计器。HAPPO [18] 和 DAE [20] 等较新的方法，则通过分解联合优势或利用差分奖励，进一步改善梯度流。

为补充集中式训练，并在去中心化执行期间实现自适应协调，一些集中式训练、分散式执行方法引入了通信机制。这些机制包括连续通信通道（CommNet [34]）、条件激活（IC3Net [32]）和基于注意力的消息传递（TarMAC [6], HetNet [28, 30]），使智能体能够共享上下文线索，并优先处理相关交互。

尽管取得了这些进展，集中式训练、分散式执行方法仍然由于共享价值分解和智能体间协调中的根本架构限制而难以实现稳定训练。为克服这些挑战，我们采用集中式分层强化学习框架，对复杂性进行分解，并实现从粗到细的时间抽象（第 III-A 节），同时通过知识蒸馏实现去中心化执行（第 III-B 节和第 V-A 节）。

B. 采用知识蒸馏的多智能体学习

近年来，知识蒸馏已成为一种颇具前景的方法，用于将协作策略从集中式智能体迁移至去中心化智能体。知识蒸馏的主要优势之一，是借助能够访问全局状态或联合观测的教师模型提供稳定的训练信号，从而隐式地应对信用分配和非平稳性等挑战。然而，这一优势通常以存在一位能够持续提供最优指导的类似预言机的教师为前提。

蒸馏策略因以下方面而异：（1）蒸馏什么（例如，价值 [43, 44] 与策略 [21]）；（2）由谁来教（例如，规划者 [5, 19],人类 [29, 41]）；（3）如何处理交互（例如，单次 [23, 36]与迭代式 [21, 14]）。然而，这些方法往往依赖静态的、预言机级别的监督，因此在开放式环境中较为脆弱。我们的实证发现表明，在复杂、动态的协作任务中，尤其是在教师可能遇到陌生状态时（第 IV 节），这一最优性假设并不成立。为克服这一局限，我们通过分层教师实现自适应监督，使其基于学生行为持续改进自身策略（第 V-B 节），并采用基于性能的筛选，有选择地蒸馏高质量示范（第 V-C 节）。

III. 预备知识

A. 分层目标导向多智能体协作

我们采用分层目标导向多智能体协作（HiT-MAC）[40] 作为集中式教师，因为它在动态、复杂的环境中表现稳健。具体而言，HiT-MAC 通过分层智能体-目标结构促进高效的多智能体训练。如图 2b 所示，高层协调器根据联合观测分配子目标，而每个智能体则充当低层执行器，根据自身的局部观测和分配的子目标选择原始行动。

在技术层面上，HiT-MAC 采用集中式训练与集中式执行（CTCE）范式，通过对联合观测运用自注意力机制 [37] 来捕捉智能体间依赖关系，并评估智能体-目标对的相对重要性。该设计通过沙普利值近似有效地进行信用分配，从而量化每个智能体-目标对对整体性能的贡献。借助这一机制，HiT-MAC 不仅解决了集中式训练、分散式执行方法中具有挑战性的信用分配问题，还成为分散式学生可靠且性能出色的教师。在实证测试中，HiT-MAC 在具有挑战性的基准场景中取得了超过 80% 的成功率，优于受测的集中式训练、分散式执行和知识蒸馏基线。为适应智能体的异质性，我们在原始设计的基础上进行了扩展，加入类别专属编码器，以处理不同的观测模态和信息访问权限。更多设计细节见附录 B。

B. 异质策略网络（*HetNet*）

为支持异构智能体之间的去中心化执行，我们采用 HetNet [30, 28] 作为学生模型。HetNet 专为支持去中心化协调而设计，通过对智能体特定的观测与动作模态进行建模来实现这一目标。

图 2c 展示了每个智能体如何通过预处理模块处理其局部观测，随后输入长短期记忆网络 [16],，通过保留相关历史信息来缓解部分可观测性。为实现结构化通信，我们采用异构图注意力（HetGAT）层，根据智能体类型及其关系上下文分配注意力权重。该机制使智能体能够选择性地关注相关队友，并共享特定于任务的信息。最后，解码器将局部特征与聚合后的图消息相结合，以生成行动概率。更多架构细节请参见附录 B。

C. 教师策略与学生策略的记法

教师策略 π_T 定义在联合观测$\bar{o}=(o^1,\ ..,\ o^n)$和联合动作 $\bar{a}=(a^1,\ ..,\ a^n)$上，表示为

$$\pi_T(\bar{a}|\bar{o})=\sum_{\bar{g}}\pi_T^H(\bar{g}|\bar{o})\prod_{i=1}^{n}\pi_T^{L_i}(a^i|o^i,g^i) \tag{1}$$

其中，π_T^H是一个高层策略，根据联合观测 $\bar{o}$分配子目标 $\bar{g}=(g^1,\ ..,\ g^n)$；$\pi_T^{L_i}$是智能体 i的低层策略，利用其观测 o^i 和分配的子目标 g^i选择行动 a^i 。总体教师策略 π_T 和价值函数 V^{π_T} 分别由 θ 和ψ参数化。

联合学生策略 π_S 定义在联合观测 $\bar{o}=(o^1,\ ..,\ o^n)$和联合动作 $\bar{a}=(a^1,\ ..,\ a^n)$上，表示为

$$\pi_S(\bar{a}\mid\bar{o})=\prod_{i=1}^{n}\pi_{S_i}(a^i\mid o^i) \tag{2}$$

其中，每个π_{S_i} 都是一种去中心化策略，它根据局部观测 o^i 选择 a^i 。策略 π_S 由 ϕ参数化。

IV. 扩展环境时的学生状态分布

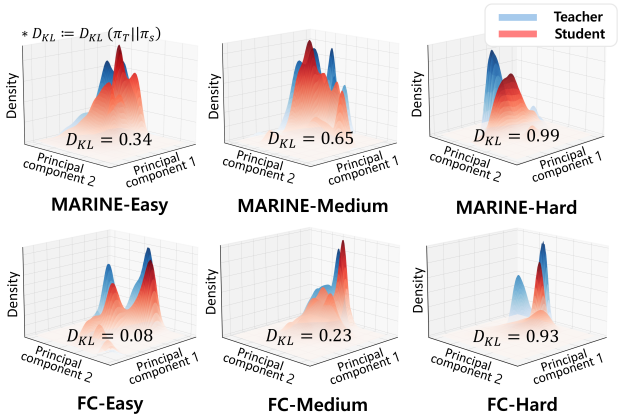

图 3. MARINE 和 FireCommander 中教师（蓝色）与学生（红色）的状态分布投影到共享潜在空间后的对比。每种设置对应成功率为 10–30% 的学生滚动轨迹。随着任务复杂度增加，学生与教师分布之间的差距（以 KL 散度衡量）逐渐扩大，这表明教师越来越多地接触到分布外（OOD）状态。

本节研究学生状态分布如何随着环境规模扩大和团队规模增长而变化。我们聚焦于成功率为 10–30% 的学生滚动轨迹，并将其轨迹与教师演示进行比较，方法是将两者投影到共享的潜在空间中（见图 3）。此处使用主成分分析（PCA） [1]，以可视化共同潜在空间中的差异。在较小规模的设置中，学生分布通常与教师分布一致，这表明教师仍可能提供高质量的示范。然而，在更复杂、智能体团队规模更大的环境中，学生会出现偏离（例如，KL 散度更高），并产生更多分布外状态，从而削弱教师提供持续且最优指导的能力。为应对这一挑战，我们提出了两个关键组成部分：伪离策略强化学习和基于性能的筛选（第 V-B 和 V-C 节）。其有效性将在第 VI-B 和 VI-C 节中评估。[1]

V. 方法

本节提出 HINT，这是一种交互式知识蒸馏框架，旨在支持稳健的多智能体学习（见图 2a）。HINT 包含三个组成部分：（1）知识蒸馏，（2）伪离策略强化学习，以及（3）基于性能的筛选。其核心是教师利用学生轨迹改进自身策略，从而提供更有价值的指导。与此同时，学生执行自身策略，并有选择地向教师查询以获得改进后的行动，确保只保留与结果相关的反馈。这个循环持续进行，直至学生策略收敛。请注意，HINT 在所有子模块中均采用多线程结构，以提高计算效率。

A. 知识蒸馏

与以往的多智能体强化学习知识蒸馏方法相比，我们借助集中式分层教师提供了更高效、更稳健的示范，因为它能够系统地分解决策过程，从而提高复杂多智能体任务的可处理性。由于教师可以访问所有智能体的观测并推断全局上下文，因此能够有效解决智能体之间的信用分配问题。然而，为了实现智能体的去中心化执行，必须将知识从教师传递给学生。整体流程总结于算法 1。请注意，虽然我们遵循典型的知识蒸馏流程，但我们的方法有所不同：数据集 $\mathcal{D}$ 会通过自适应教师持续扩充，从而为学生提供更丰富的指导。

我们的策略蒸馏目标通过将 KL 散度与熵正则化项相结合，最小化教师策略 π_T 与学生策略 π_S 之间的差异，以鼓励模仿和策略多样性。损失由可调系数 α加权，如公式 3 所示。

$$\mathcal{L}_\phi = \mathbb{E}_{(\bar{o}_t,\bar{a}_t)\sim\pi_T}\Big[\big(\log\pi_T(\bar{a}_t \mid \bar{o}_t) - \log\pi_S(\bar{a}_t \mid \bar{o}_t)\big) + \alpha\mathcal{H}(\pi_S(\cdot \mid \bar{o}_t))\Big] \tag{3}$$

[1]次优示范的更多示例以及图 3 所示流程详见附录 C，环境详情见第 VI-A 节和附录 A。

算法 1： 知识蒸馏

输入： 数据集 $\mathcal{D} = \{\tau = \bar{o}_t\ \bar{a}_t\ {}_{t=1}^{H}\}$　　$\mathcal{B}(,)$，缓冲区 **初**

1 **始化** ϕ 作为 π_S 的参数2 **对于** τ，在 $\mathcal{D}$ **执行**3 **对于** $t=1$ **到** H **执行**4 计算 $\log\pi_S(\bar{a}_t|\bar{o}_t)$ 和 $\mathcal{H}(\pi_S(\cdot \mid \bar{o}_t))$，并存入 $\mathcal{B}$；5 **如果** $\mathcal{B}$ 已满**则** 6 使用公式 3 计算损失 $\mathcal{L}_\phi$；// 使用熵正则化的 KL 散度

7 $\phi \leftarrow \phi - \lambda_\phi\nabla_\phi\mathcal{L}_\phi$；8 清空缓冲区 $\mathcal{B}$；

算法 2： 伪离策略强化学习

输入： 教师 π_T，学生 π_S，缓冲区 $\mathcal{B}$，回合 N_{pseudo}

1 冻结 π_S；
2 **对于** $e=1$ **到** N_{pseudo}，**执行**
3　采样切换点 $t' \sim$ 均匀分布(1，H)；
4　**对** $t=0$ **到** H **执行**
5　　**如果** $t \le t'$ **则**
6　　　$\bar{a}_t \sim \pi_S(\cdot \mid \bar{o}_t)$；　// 学生进行探索
7　　**else**
8　　　$\bar{a}_t \sim \pi_T(\cdot \mid \bar{o}_t)$；　// 教师恢复
9　　将 $(\bar{o}_t, \bar{a}_t)$ 存入 $\mathcal{B}$；
10　　**如果** $\mathcal{B}$ 包含 n 步数 **则**
11　　　计算$v_t^{\pi_T}$ (式 4)；　// 校正离策略
12　　　计算损失 $\mathcal{L}_\psi$, J_θ (式 5、6)；
13　　　$\psi \leftarrow \psi - \lambda_\psi\nabla_\psi\mathcal{L}_\psi$, $\theta \leftarrow \theta + \lambda_\theta\nabla_\theta J_\theta$；// 更新教师
14　　　清除 $\mathcal{B}$；

为确保训练稳定，样本会存储在回放缓冲区中，并通过基于批次的梯度优化进行更新。我们通过实验证实，这种形式在测试的基准测试中产生了稳定的梯度（见附录 F）。

B. 伪离策略强化学习

我们的伪离策略强化学习为教师与学生如何在多智能体学习中共同演化提供了新的视角。整体流程如算法 2 所示。我们不再将学生数据和教师数据分开处理，而是将二者结合在单条轨迹中：学生首先进行探索，随后由教师完成路径（第 5—8 行）。这种联合展开结构（图 4）弥合了教师与学生之间的差距，使教师不仅能够学习该做什么，还能学习如何从学生的错误中恢复（第 11—13 行）。这种方法造就了更具韧性的教师，并通过更丰富、结合上下文的指导促进学生学习——尤其是在复杂场景中，学生经常偏离训练分布，仅依靠反应式纠正往往不足以应对这些情况。

为了有效利用这些混合轨迹，我们需要一种稳定的方法，使用离策略数据更新教师。为此，我们采用了 IMPALA [8]，中的 V-trace 方法，该方法已被证明能够通过有界的异策略校正实现稳定收敛。这样，我们便可以根据学生和教师轨迹的混合数据估计经过校正的价值目标$v_t^{\pi_T}$（第 11 行）。

$$v_t^{\pi_T} = V^{\pi_T}(\bar{o}_t) + \sum_{j=t}^{t+n-1}\gamma^{j-t}\big(\prod_{i=t}^{j-1}c_i\big)\delta_j V \tag{4}$$

其中 $\delta_j V = \rho_j\left(r_j + \gamma V^{\pi_T}(\bar{o}_{j+1}) - V^{\pi_T}(\bar{o}_j)\right)$ 是由截断的重要性权重缩放的时间差分误差，

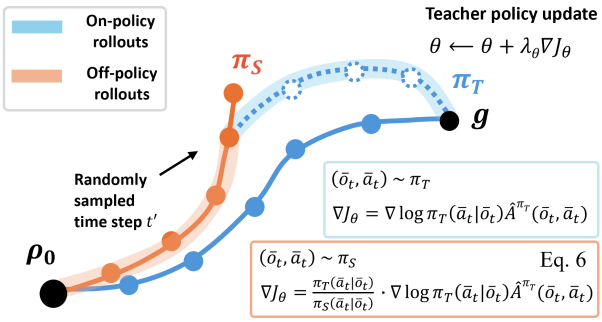

图 4. 教师策略（π_T）通过其同策略轨迹（蓝色阴影）以及来自学生（π_S）的离策略轨迹（橙色阴影）进行优化，并通过重要性采样校正策略梯度。

$\rho_j = \min\left(1, \frac{\pi_T(\bar{a}_j|\bar{o}_j)}{\mu(\bar{a}_j|\bar{o}_j)}\right)$ 和 $c_i = \min\left(1, \frac{\pi_T(\bar{a}_i|\bar{o}_i))}{\mu(\bar{a}_i|\bar{o}_i))}\right)$。其中，$\mu$ 表示行为策略，可以是 π_T 或 π_S。请注意，尽管我们使用了未经修改的 V-trace，但我们的创新之处在于首次将其应用于解决多智能体蒸馏设置中的教师-学生分布不匹配问题。

我们通过最小化其预测值与公式 5（第 12 行）中的校正 V-trace目标之间的平方误差，来训练教师的价值函数 $V^{\pi_T}(\bar{o}_t)$。即使学生轨迹与教师策略存在显著偏差，这一校正后的价值目标 $v_t^{\pi_T}$ 也有助于稳定训练，否则可能导致训练不稳定。

$$\mathcal{L}_\psi = \mathbb{E}_{(\bar{o}_t, \bar{a}_t) \sim \{\pi_S, \pi_T\}} \left[\left(V^{\pi_T}(\bar{o}_t) - v_t^{\pi_T} \right)^2 \right] \tag{5}$$

同样，策略更新融合了来自教师和学生的轨迹。对于学生生成的样本，我们使用重要性加权进行离策略校正，而教师生成的样本仍保持同策略。请注意，$\hat{}$ 优势估计 $A^{\pi_T}(\bar{o}_t, \bar{a}_t)$ 使用 V-trace目标计算得出，即 $r_t + \gamma v_{t+1}^{\pi_T} - V^{\pi_T}(\bar{o}_t)$。最终的策略目标如公式 6（第 12 行）所示，其中，结合增强型优势估计的离策略校正可提供更稳健、稳定的训练信号，避免受到离策略漂移的误导。

$$J_\theta = \mathbb{E}_{(\bar{o}_t, \bar{a}_t) \sim \pi_S} \left[\frac{\pi_T(\bar{a}_t|\bar{o}_t)}{\pi_S(\bar{a}_t|\bar{o}_t)} \log \pi_T(\bar{a}_t|\bar{o}_t) \hat{A}^{\pi_T}(\bar{o}_t, \bar{a}_t) \right] + \mathbb{E}_{(\bar{o}_t, \bar{a}_t) \sim \pi_T} \left[\log \pi_T(\bar{a}_t|\bar{o}_t) \hat{A}^{\pi_T}(\bar{o}_t, \bar{a}_t) \right] \tag{6}$$

式 5 和式 6 中的目标允许利用同策略和离策略数据对教师进行自适应优化（第 10 行）。这种混合学习方案在保持策略一致性的同时增强了泛化能力。在这一自适应阶段，教师的低层策略保持不变，因为它已在足够多样的状态分布上完成预训练。这一设计选择简化了优化过程，并将自适应重点放在战略目标分配上。

C. 基于性能筛选的 *DAgger*

我们的方法基于标准 DAgger（数据集聚合） [27]，通过在不断扩大的观测-动作对数据集上迭代训练学生策略来解决分布偏移问题。我们通过在环境中运行学生策略，并在其访问的状态下查询专家策略，来收集这些观测-动作对。在这个聚合后的

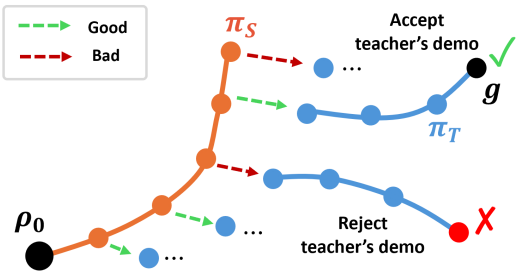

图 5. 数据集聚合过程中应用基于性能的筛选。接收高质量的教师演示（绿色，√），拒绝次优演示（红色，）。

算法 3： 基于性能筛选的 DAgger

输入： 教师 π_T，学生 π_S，数据集 $\mathcal{D}$，总回合数 N_{query}

1 **对于** $e = 1$ **到** N_{query}，**执行**
2 初始化环境，设置轨迹 $\tau \leftarrow []$; // 学生进行探索
3 **对于** $t = 1$ **到** H，**执行**
4 执行学生动作 $\bar{a}_t \sim \pi_S(\cdot|\bar{o}_t)$;
5 查询教师动作 $\bar{a}_t^* \sim \pi_T(\cdot|\bar{o}_t)$; // 教师进行指导
6 模拟教师策略 π_T，从 $(\bar{o}_t, \bar{a}_t^*)$ 开始，直到终止状态 $\bar{o}_H$;
7 **如果**终止状态 $\bar{o}_H$ 满足成功条件 **则**
8 将 $(\bar{o}_t, \bar{a}_t^*)$ 追加到轨迹 τ; // 基于性能的筛选
9 将 τ 添加到数据集 $\mathcal{D}$; // 聚合数据集

数据集上训练，使学生能够从与测试时所遇数据分布相近的数据中学习。然而，DAgger 通常假设可以访问可靠的教师。在实践中，尤其是在开放式环境中（图 3），我们的分层教师策略在交互式查询期间可能会遇到不熟悉或未见过的状态，导致示范不一致，并使集中式教师与去中心化学生之间的观测不匹配不断累积。

为应对这一挑战，我们提出了一种基于性能的过滤机制，在数据聚合过程中运行，以提升专家动作的质量。整体流程如算法 3 所述。如图 5 所示，当学生在每个时间步向教师查询时（第 5 行），我们会从该时刻开始模拟教师的轨迹，直至回合结束（第 6 行）。假设由此得到的终止状态满足任务特定成功标准（第 7 行），那么我们就将最初查询到的行动视为良好行动的代理（绿色），并将其纳入数据集（第 8 行和第 9 行）。否则，就舍弃该专家动作。尽管教师为最优行为提供了启发式代理，但我们的机制通过模拟其长期后果，增加了第二层评估，以判断这一代理是否可信。

这种领域无关过滤方法使数据聚合能够更好地抵御带有噪声或次优的示范。通过有选择地纳入经验证与成功结果相关联的专家动作，学生策略得以在质量更高的数据上训练。此外，为了稳定训练并适应教师策略不断演变的特点，我们保留固定数量的初始示范，并定期采样近期交互。虽然这一策略类似于带有课程偏置的经验回放，但我们的方法独特地融合了静态专家监督与自适应反馈，既提供可靠的

基础，也提供与学生当前能力相匹配的新鲜指导。

VI. 结果与讨论

本节通过与基线方法进行比较（第 VI-B 节），评估 HINT 的性能，并通过消融研究验证各个子模块的贡献（第 VI-C 节）。详细的训练设置，以及完整的基准测试结果和消融研究结果，分别见附录 D、E 和 F。

A. 环境

我们选取了两个高度动态、部分可观测且具有异质性的多智能体领域。这些领域体现了现实世界中的挑战，需要及时且有效的协调，因此非常适合用于测试 HINT。性能通过两个关键指标衡量：成功率，即成功完成的回合所占比例；以及平均步数，即完成每项任务所需的步数。有关环境的更多详情，请参阅附录 A。

MARINE [12]: 一个海上物流环境，路径规划智能体在动态海况中航行，而后勤智能体则提供海上加油。该环境使用 WaveWatchIII 预报数据 [35], ，将天气诱发的不确定性引入规划与协调之中。路径规划智能体必须在燃料耗尽前抵达目的地，因此需要及时与后勤智能体会合。环境分为三个难度等级：简单（5×5，2*R*/1*L*）、中等（10×10，3*R*/2*L*）和困难（20×20，6*R*/4*L*），其中*R* 和*L* 分别表示路径规划智能体和后勤智能体。

FireCommander（FC） [28]: 一种基于网格的野火环境，灵感来自 FARSITE [9], ——这是一款广泛使用的模拟器，通过微分方程对火灾的时空行为进行建模。在 FC 中，行动智能体负责扑灭火灾，但不具备感知能力；感知智能体则监测火势蔓延，但无法扑灭火灾。这种角色分工促进了智能体在受动态火灾传播影响的随机环境中协作。该环境设有三个难度等级：简单（5×5，2*A*/1*P*）、中等（10×10，3*A*/2*P*）和困难（21×21，6*A*/4*P*），其中 *A* 和 *P* 分别表示行动智能体和感知智能体。

B. 基准测试

1) 在线集中训练、分散执行基线：我们将所提出的方法与具有竞争力的在线集中训练、分散执行方法进行比较，包括两种不使用通信的方法（多智能体近端策略优化 [42]和 HAPPO [18]）、四种使用通信的方法（TarMAC [6],IC3Net [32], Comm- Net [34], 和 HetNet [30]）。为确保公平，所有基线方法均使用额外的时间步进行训练，以匹配训练教师策略时所使用的时间步数。为进一步验证交互式蒸馏的有效性，我们在困难任务上将 HINT 与采用热启动的强化学习方法进行比较。这里的热启动是指通过行为克隆（行为克隆）使用我们的教师策略初始化强化学习基线，而不是从头开始训练。根据表 I，HAPPO 和 HetNet 分别是 MARINE 和 FireCommander 中表现强劲的基线方法。因此，我们分别使用教师策略对这些方法进行 25% 和 50% 的热启动初始化。

表 I 显示，在 MARINE 和 FC 的中等及困难任务中，我们提出的方法始终优于其他 CTDE 基线。具体而言，较高的成功率与较少的步数表明，我们的方法能够训练出既可靠又高效的智能体，即使在具有挑战性的场景中也是如此。此外，在 MARINE 困难任务中，HINT 相较于采用热启动的 HAPPO 取得了显著提升。我们推测，MARINE 的终止条件——路径规划智能体燃料耗尽——会阻碍探索，并使智能体容易陷入局部最优，即使采用热启动也是如此。在 FireCommander 困难任务中，HINT 的表现比采用热启动的 HetNet 高出 20%–45%，同时标准差也小得多，体现出训练过程更加稳定。

这些改进可归因于我们框架中的结构化通信。然而，与 Het- Net 和其他热启动变体相比，后者采用了与 HINT 相同的学生策略，或使用教师演示进行初始化；通过比较，我们可以看出，稳定的训练信号和自适应教师指导如何促进最终性能。这种信息丰富的指导来自我们的集中式分层教师，以及由伪离策略强化学习和性能筛选数据指导的自适应机制。这些组成部分共同支持在部分可观测环境和分布外状态下进行稳健的策略学习，同时减少教师与学生之间的观测不匹配。请注意，为确保比较公平，所有方法均使用相同数量的样本进行训练。完整的比较结果和训练曲线请参见附录 E。

2) 在线知识蒸馏基线：在此，我们将 HINT 与三种在线知识蒸馏（KD）方法进行比较：

1. CTDS [43] - 该方法采用能够访问全局信息的集中式教师，并通过模仿学习训练分散式学生。2. IGM-DA [14] - 与 CTDS 类似，但集中式教师使用学生收集的数据进行训练，这些数据还融合了全局信息；随后通过 DAgger 将知识蒸馏回学生。3. PTDE [3]- 引入全局信息个性化（GIP）模块作为教师网络，将面向智能体个性化的全局知识蒸馏到智能体的局部信息中。

所有 KD 基线方法均针对基于价值的多智能体强化学习设计。因此，我们为教师和学生采用 QMiX [26] 作为基础策略，因为它在原始知识蒸馏研究中取得了最佳性能。为缓解教师与学生网络协同演化所导致的收敛问题，我们对 KD 基线方法的训练时间步数设置得比 CTDE 基线更多。

图 6 显示，在 MARINE 和 FireCommander 环境中，HINT 的表现均优于其他 KD 基线。在 FireCommander 中，由于 QMIX 缺少通信模块，KD 基线难以取得良好性能。尽管如此，CTDS 和 IGM-DA 教师即使能够访问全局

表 I MARINE 和 FireCommander（FC）在线 CTDE 基准测试的定量结果，报告为三个随机种子的均值（±标准差）。此处，*WS* 表示使用我们的教师策略热启动强化学习。

方法	MARINE-中等（10x10，5 个智能体）		MARINE-Hard（20x20，10 个智能体）		FC-Medium（10x10，5 个智能体）		FC-Hard（21x21，10 个智能体）	
	成功率（%）↑	步数↓	成功率（%）↑	步数↓	成功率（%）↑	步数↓	成功率（%）↑	步数↓
多智能体近端策略优化	40.67 ± 51.39	69.23 ± 37.31	0.67 ± 1.15	198.88 ± 1.94	6.00 ± 3.46	96.33 ± 1.19	4.00 ± 0.00	202.89 ± 0.86
TarMAC	94.00 ± 10.39	**25.84 ± 14.77**	0.00 ± 0.00	200.0 ± 0.00	5.33 ± 5.03	96.89 ± 3.97	2.67 ± 3.06	207.51 ± 2.48
IC3Net	88.00 ± 10.58	34.97 ± 17.99	0.00 ± 0.00	200.0 ± 0.00	6.67 ± 6.11	97.07 ± 2.79	2.67 ± 3.06	207.08 ± 2.68
CommNet	0.00 ± 0.00	100.00 ± 0.00	0.00 ± 0.00	200.0 ± 0.00	3.33 ± 3.06	97.81 ± 1.92	1.33 ± 2.31	207.96 ± 1.77
HAPPO	98.67 ± 2.31	27.81 ± 0.46	0.00 ± 0.00	200.0 ± 0.00	0.67 ± 1.15	99.35 ± 1.13	0.67 ± 1.15	208.61 ± 2.40
→ 25% 工作站	-	-	24.00 ± 3.46	169.94 ± 9.25	-	-	-	-
→ 50% 工作站	-	-	38.00 ± 6.93	154.73 ± 8.30	-	-	-	-
HetNet	61.33 ± 53.27	60.33 ± 34.64	0.00 ± 0.00	200.0 ± 0.00	82.00 ± 14.00	49.43 ± 16.00	19.33 ± 7.02	177.93 ± 11.91
→ 25% 工作站	-	-	-	-	-	-	35.33 ± 15.28	152.35 ± 18.78
→ 50% 工作站	-	-	-	-	-	-	42.67 ± 38.28	139.01 ± 62.69
HINT（我们的方法）	**98.67 ± 1.15**	26.49 ± 4.92	**53.33 ± 9.02**	**143.28 ± 4.47**	**84.00 ± 2.00**	**47.95 ± 3.73**	**51.33 ± 5.03**	**136.40 ± 3.86**

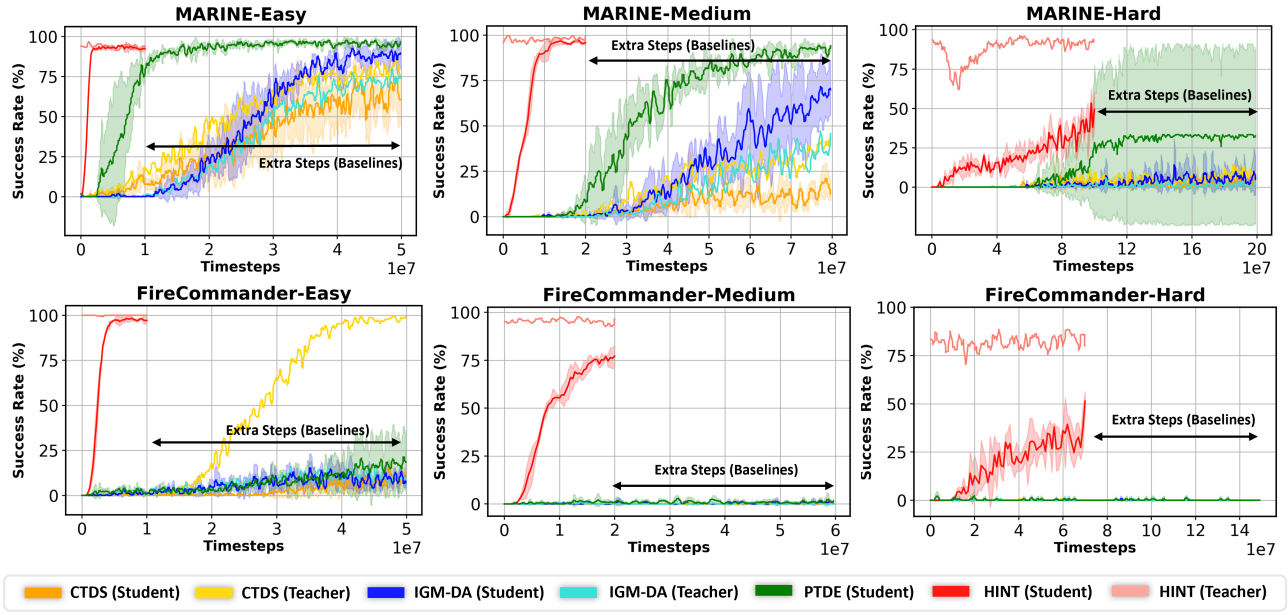

图 6. 在线知识蒸馏基准中 MARINE 和 FireCommander 的学习曲线，报告的是三个随机种子和三种难度设置下的均值（± 标准差）。HINT 在两个领域中始终优于所有基线方法。

信息，其表现仍然不佳，这凸显了我们的分层教师在更具挑战性的设置下的鲁棒性。（由于 PTDE 中的教师充当辅助模块而非策略，因此其性能未在图 6 中展示。）在 MARINE 中，只有 PTDE 在简单和中等任务上的性能能够与 HINT 相媲美，但在困难任务上，HINT 的表现超过 PTDE 60%。与第 VI-B1 节类似，HINT 的成功归因于其结构化通信，但这些结果也支持了自适应分层教师在我们整个流程中的重要性。完整的定量结果请参见附录 E。

C. 消融实验

1）关键模块的作用：为评估 HINT 中各关键组件的影响，我们进行了消融实验，将完整模型与三个变体进行比较：一个移除了伪离策略强化学习，一个移除了基于性能的筛选，还有一个同时移除了这两个组件，使方法实际上退化为标准 DAgger。在训练过程中，我们根据成功率和平均步数评估教师策略与学生策略。对于教师策略，我们还报告次优示范率，该指标用于衡量指导质量：即教师响应学生查询后生成的轨迹，从查询点开始是否能带来成功结果。

图7显示，完整模型始终优于所有消融变体，在 MARINE-Hard 和 FC-Hard 任务上均取得了显著提升。尤其是，与其他变体相比，我们的完整模型具有更低的次优演示率，从而使学生的表现更佳。请注意，在 MARINE-Hard条件下，教师的初始表现可能会下降，因为它需要适应各种新的陌生学生查询。然而，这会减少次优演示，进而提供更具信息量的指导。综合来看，这些发现证实，HINT 的适应能力和筛选机制对于实现稳健、高质量的策略学习至关重要。完整的消融结果请见附录 F。

2）学生结构：为了更好地理解学生模型中不同架构组件的贡献，我们在 MARINE-中等和 FC-中等任务上进行了消融研究。我们提出的学生模型 HetNet 集成了两个关键模块：用于应对部分可观测性的长短期记忆网络，以及

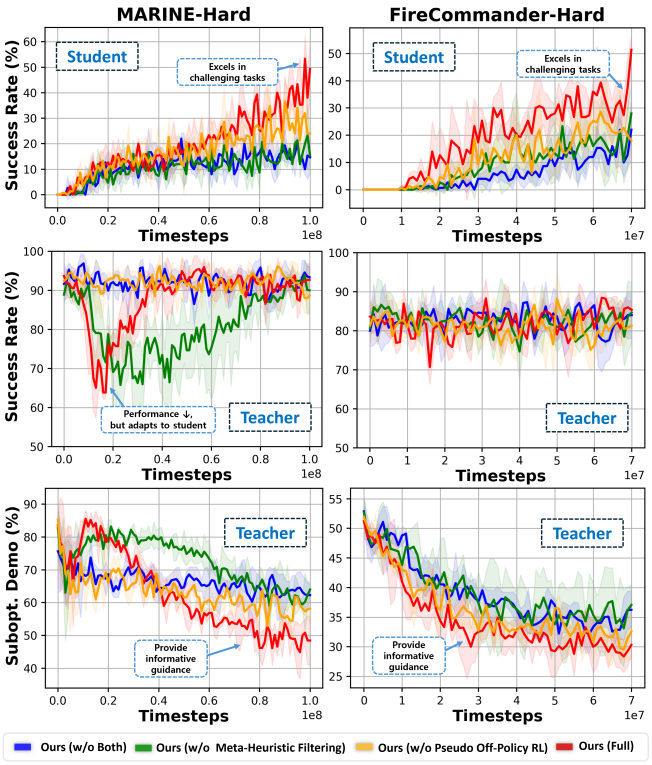

图7. MARINE-Hard和FC-Hard上关键模块的消融结果。每一行分别对应学生的成功率、教师的成功率和教师的次优性演示率，报告的是三个随机种子下的均值（± 标准差）。

用于结构化智能体间通信的 HetGAT 机制。我们将 HetNet 与两个消融变体进行了比较：

- LSTM-Only：采用LSTM的学生模型（即无通信）
- LSTM+GNN：采用LSTM和图神经网络的学生模型 [19]

如表II所示，HetNet的表现始终优于两种消融基线方法，在FC-Medium任务上的提升最大——与LSTM-Only相比，成功率提高了84个百分点；与LSTM+GNN相比，则提高了18个百分点。在这一环境中，由于行动智能体的感知能力有限，结构化通信至关重要。LSTM+GNN与HetNet之间的性能差距凸显了异构注意力在我们架构中的价值。

3) 敏感性分析：我们分析了FC-Medium任务中的四个关键超参数：α（学生的熵系数）、lr（学习率）、N_{pseudo}（每个训练周期用于教师训练的回合数）以及 N_{query}（每个训练周期用于专家查询的回合数）。如表 III 所示，HINT在广泛的设置范围内保持稳定（成功率变化范围为-2.3%至 +6.3%），较高的取值通常能通过增加训练数据和正则化带来适度的提升。总体而言，这些结果凸显了HINT的鲁棒性，并阐明了各个超参数对性能的影响。

D. 局限性与未来工作

由于我们的教师策略遵循 CTCE范式，其可扩展性从根本上受到限制。尽管如此，我们选择

表II MARINE-中等和 FC-Medium 上学生结构的消融结果，以三个随机种子的均值（±标准差）报告。

方法	MARINE-中等		FC-Medium	
	成功比率（%）↑	步数步数 ↓	成功比率（%）↑	步数采纳 ↓
我们的方法（长短期记忆网络）	92.67 ± 3.06	36.28 ± 6.31	0.00 ± 0.00	100.00 ± 0.00
我们的方法（长短期记忆网络 + 图神经网络）	98.00 ± 2.00	27.80 ± 0.97	66.00 ± 8.72	56.91 ± 6.17
我们的方法（HetNet）	**98.67 ± 1.15**	**26.49 ± 4.92**	**84.00 ± 2.00**	**47.95 ± 3.73**

表 III

FC-Medium 中不同超参数下的性能对比，报告为三个随机种子的均值（±标准差）。

超参数	α			lr		
	0.01（我们的方法）	0.02	0.05	1e-4（我们的方法）	2e-4	5e-4
成功率（%）↑	84.00 ± 2.00	82.00 ± 8.00	**89.33 ± 4.62**	84.00 ± 2.00	86.00 ± 2.00	**89.33 ± 4.62**
步数耗时 ↓	47.95 ± 3.73	47.27 ± 3.95	**44.05 ± 5.39**	47.95 ± 3.73	44.62 ± 2.24	**44.05 ± 5.39**

超参数	N_{pseudo}			N_{query}		
	10	20（我们的方法）	40	50（我们的方法）	75	100
成功比率（%）↑	84.67 ± 2.31	84.00 ± 2.00	**85.33 ± 7.57**	84.00 ± 2.00	85.33 ± 7.02	**86.67 ± 8.08**
步数已采取 ↓	48.35 ± 5.98	47.95 ± 3.73	**46.88 ± 6.41**	47.95 ± 3.73	47.51 ± 7.36	**44.66 ± 7.91**

一个完全集中的教师，因其性能出色，并尝试通过分层强化学习提升可扩展性。在实践中，HINT 可扩展至 20×20 个网格环境，最多支持 10 个智能体；规模看似不大，但在联合学习通信协议时仍颇具挑战。此外，与其他集中式训练、分散式执行和知识蒸馏基线相比，我们的框架包含更多样的模块，可能导致更高的计算成本。不过，我们通过在整个流程中采用多线程结构缓解了这一问题，并发现总体计算时间与有竞争力的基线方法相当（例如，在 MARINE-Hard 上，HINT：146 小时，PTDE：161 小时；在 FC-Hard 上，HINT：90 小时，HetNet：179 小时）。更多细节见附录 G。最后，HINT 依赖人类专业知识来定义恰当的层级结构。尽管这种设计提供了实用的灵活性且易于实现，但完全端到端的训练可能会带来更具适应性的行为，并展现涌现能力。未来的工作可以探索通过基于宏动作的技能发现实现层级构建自动化，并开发基于提取技能、可在智能体之间异步运行的信用分配机制。

VII. 结论

我们提出了 HINT，这是一种稳健的多智能体学习框架，利用基于分层教师的自适应知识蒸馏。通过结合伪离策略强化学习和基于性能的筛选，我们的方法能够在动态、复杂的协作任务中实现自适应且稳健的知识迁移。实证评估表明，即使环境复杂度和团队规模不断增加，HINT 仍能始终保持稳健的性能。我们的研究结果凸显了应对教师策略次优性的至关重要性，并为未来自适应、稳健的多智能体系统研究铺平了道路。

致谢

本研究得到了海军研究实验室资助，资助项目编号为 N00173-21-1-G009 和 N00173-25-1-0050；此外还得到了 MIT 林肯实验室的支持。

参考文献

[1] Hervé Abdi 和 Lynne J Williams。主成分分析。*Wiley* 跨学科评论：计算统计学，2(4):433–459，2010。[2] Gerald G Brown、Walter C DeGrange、Wilson L Price 和 Anton A Rowe。为美国海军安排战斗后勤部队海上补给。《海军研究：后勤》，64(8):677–693，2017。[3] Yiqun Chen、Hangyu Mao、Jiaxin Mao、Shiguang Wu、Tianle Zhang、Bin Zhang、Wei Yang 和 Hongxing Chang。PTDE：通过蒸馏执行实现多智能体强化学习的个性化训练。*arXiv* 预印本 *arXiv:2210.08872*，2022。[4] John R Cooper。基于势场理论的最优多智能体搜索与救援。发表于*AIAA Scitech 2020* 论坛，第 0879 页，2020。[5] Mehul Damani、Zhiyao Luo、Emerson Wenzel 和 Guillaume Sartoretti。Primal _2：通过强化学习与模仿实现终身多智能体路径规划。*IEEE*《机器人学与自动化快报》，6(2):2666–2673，2021。[6] Abhishek Das、Théophile Gervet、Joshua Romoff、Dhruv Batra、Devi Parikh、Mike Rabbat 和 Joelle Pineau。TarMAC：定向多智能体通信。发表于国际机器学习会议，第 1538–1546 页。PMLR，2019。[7]Jared R Deiter。海上补给规划器的统计敏感性分析。硕士论文，海军研究生院，2022。[8] Lasse Espeholt、Hubert Soyer、Remi Munos、Karen Simonyan、Vlad Mnih、Tom Ward、Yotam Doron、Vlad Firoiu、Tim Harley、Iain Dunning 等。IMPALA：采用重要性加权演员-学习器架构的可扩展分布式深度强化学习。发表于国际机器学习会议，第 1407–1416 页。PMLR，2018。[9] Mark Arnold Finney。*FARSITE*，火灾区域模拟器——模型开发与评估。第 4 号。美国农业部森林局落基山研究站，1998。[10] Jakob Foerster、Ioannis Alexandros Assael、Nando De Freitas 和 Shimon Whiteson。通过深度多智能体强化学习学习通信。神经信息处理系统进展，29，2016。[11] Jakob Foerster、Gregory Farquhar、Triantafyllos Afouras、Nantas Nardelli 和 Shimon Whiteson。反事实多智能体策略梯度。发表于*AAAI* 人工智能会议论文集，第 32 卷，2018。[12] Byeolyi Han、Minwoo Cho、Letian Chen、Rohan Paleja、Zixuan Wu、Sean Ye、Esmaeil Seraj、David Sidoti 和 Matthew Gombolay。学习多智能体协同

海上补给。*IEEE*《机器人学与自动化快报》，2024。[13] Geoffrey Hinton、Oriol Vinyals 和 Jeff Dean。蒸馏神经网络中的知识。*arXiv* 预印本 *arXiv:1503.02531*，2015。[14] Yitian Hong、Yaochu Jin 和 Yang Tang。重新思考协作式多智能体强化学习中的个体全局最大值。神经信息处理系统进展，35:32438–32449，2022。[15] Shariq Iqbal 和 Fei Sha。用于多智能体强化学习的演员-注意力-评论家。发表于国际机器学习会议，第 2961–2970 页。PMLR，2019。[16] Rafal Jozefowicz、Wojciech Zaremba 和 Ilya Sutskever。循环网络架构的实证探索。发表于国际机器学习会议，第 2342–2350 页。PMLR，2015。[17] Aleksandar Krnjaic、Raul D Steleac、Jonathan D Thomas、Georgios Papoudakis、Lukas Schäfer、Andrew Wing Keung To、Kuan-Ho Lao、Murat Cubuktepe、Matthew Haley、Peter Börsting 等。面向机器人与人类协作者仓储物流的可扩展多智能体强化学习。发表于*2024*年*IEEE/RSJ*智能机器人与系统国际会议（*IROS*），第 677–684 页。IEEE，2024。[18] Jakub Grudzien Kuba、Ruiqing Chen、Muning Wen、Ying Wen、Fanglei Sun、Jun Wang 和 Yaodong Yang。多智能体强化学习中的信赖域策略优化。发表于国际学习表征会议，2022。[19]Qingbiao Li、Fernando Gama、Alejandro Ribeiro 和 Amanda Prorok。用于去中心化多机器人路径规划的图神经网络。发表于*2020*年*IEEE/RSJ*智能机器人与系统国际会议（*IROS*），第 11785–11792 页。IEEE，2020。[20] Yueheng Li、Guangming Xie 和 Zongqing Lu。用于多智能体策略梯度的差分优势估计。发表于国际机器学习会议，第 13066–13085 页。PMLR，2022。[21]Alex Tong Lin、Mark Debord、Katia Estabridis、Gary Hewer、Guido Montufar 和 Stanley Osher。通过模仿集中式控制器实现去中心化多智能体。发表于数学与科学机器学习，第 619–651 页。PMLR，2022。[22] Ryan Lowe、Yi I Wu、Aviv Tamar、Jean Harb、OpenAI、Pieter Abbeel 和 Igor Mordatch。混合协作竞争环境中的多智能体演员-评论家。神经信息处理系统进展，30，2017。[23] Linghui Meng、Muning Wen、Chenyang Le、Xiyun Li、Dengpeng Xing、Weinan Zhang、Ying Wen、Haifeng Zhang、Jun Wang、Yaodong Yang 等。离线预训练多智能体决策变换器。机器智能研究，20(2):233–248，2023。[24]Volodymyr Mnih、Adria Puigdomenech Badia、Mehdi Mirza、Alex Graves、Timothy Lillicrap、Tim Harley、David Silver

深度强化学习方法。载于国际机器学习会议，第1928–1937页。PMLR，2016。[25] Aowabin Rahman、Arnab Bhattacharya、Thiagarajan Ramachandran、Sayak Mukherjee、Himanshu Sharma、Ted Fujimoto和Samrat Chatterjee。Adversar：通过多智能体强化学习进行对抗性搜索与救援。载于*2022*年*IEEE*国土安全技术国际研讨会 *(HST)*，第1–7页。IEEE，2022。[26] Tabish Rashid、Mikayel Samvelyan、Christian Schroeder de Witt、Gregory Farquhar、Jakob Foerster和Shimon Whiteson。深度多智能体强化学习的单调价值函数分解。机器学习研究期刊，21(178)：1–51，2020。[27] Stéphane Ross、Geoffrey Gordon和Drew Bagnell。将模仿学习和结构化预测转化为无悔在线学习。载于第十四届人工智能与统计国际会议论文集，第627–635页。JMLR研讨会与会议论文集，2011。[28]Esmaeil Seraj、Zheyuan Wang、Rohan Paleja、Daniel Martin、Matthew Sklar、Anirudh Patel和Matthew Gombolay。学习高效多样化通信，以实现协作异构团队。载于自主智能体与多智能体系统国际会议论文集，第1173–1182页，2022。[29]Esmaeil Seraj、Jerry Xiong、Mariah Schrum和Matthew Gombolay。用于机器人团队人类训练的混合主动式多智能体学徒学习。神经信息处理系统进展，36：35426–35440，2023。[30] Esmaeil Seraj、Rohan Paleja、Luis Pimentel、Kin Man Lee、Zheyuan Wang、Daniel Martin、Matthew Sklar、John Zhang、Zahi Kakish和Matthew Gombolay。用于复合机器人团队通信与协调的异构策略网络。*IEEE*机器人学汇刊，2024。[31] Yi Shen、Benjamin McClosky、Joseph W Durham和Michael M Zavlanos。面向大型机器人仓库分拣中心资源分配的多智能体强化学习。载于*2023*年第*62*届*IEEE*决策与控制会议 *(CDC)*，第7137–7143页。IEEE，2023。[32]Amanpreet Singh、Tushar Jain和Sainbayar Sukhbaatar。在多智能体合作与竞争任务中学习如何实现大规模通信。载于国际学习表征会议论文集，2018。[33] Kyunghwan Son、Daewoo Kim、Wan Ju Kang、David Earl Hostallero和Yung Yi。Qtran：通过变换学习分解，以实现协作多智能体强化学习。载于国际机器学习会议，第5887–5896页。PMLR，2019。[34] Sainbayar Sukhbaatar、Rob Fergus等。通过反向传播学习多智能体通信。神经信息处理系统进展，29，2016。[35]Hendrik L. Tolman、Bhavani Balasubramaniyan、Lawrence D. Burroughs、Dmitry V. Chalikov、Yung Y. Chao、Hsuan S. Chen 和 Vera M. Gerald。NCEP 风生海洋表面波模型的开发与实现。《天气与预报》，17(2): 311 – 333，2002。[36] Wei-Cheng Tseng、Tsun-Hsuan Johnson Wang、Yen- Chen Lin 和 Phillip Isola。基于知识蒸馏的离线多智能体强化学习。神经信息处理系统进展，35:226–237，2022。[37]Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N Gomez、Łukasz Kaiser 和 Illia Polosukhin。Attention Is All You Need。神经信息处理系统进展，30，2017。[38] Jianhong Wang、Yuan Zhang、Tae-Kyun Kim 和 Yunjie Gu。Shapley Q值：一种解决全局奖励博弈的局部奖励方法。载于*AAAI* 人工智能会议论文集，第34卷，7285–7292页，2020。[39] Stanisław W.eglarczyk。核密度估计及其应用。载于*ITM*会议网络，第23卷，文章编号00037。EDP Sciences，2018。[40] Jing Xu、Fangwei Zhong 和 Yizhou Wang。学习多智能体协调以增强定向传感器网络中的目标覆盖。神经信息处理系统进展，33:10053–10064，2020。[41] Pei Xu 和 Ioannis Karamouzas。运用知识蒸馏实现受人类启发的多智能体导航。载于*2021*年*IEEE/RSJ*智能机器人与系统国际会议 *(IROS)*，8105–8112页。IEEE，2021。[42] Chao Yu、Akash Velu、Eugene Vinitsky、Jiaxuan Gao、Yu Wang、Alexandre Bayen 和 Yi Wu。近端策略优化在合作式多智能体博弈中的惊人效果。载于神经信息处理系统进展，数据集与基准专题，2022。[43] Jian Zhao、Xunhan Hu、Mingyu Yang、Wengang Zhou、Jiangcheng Zhu 和 Houqiang Li。CTDS：用于多智能体强化学习的集中式教师与去中心化学生。*IEEE*游戏汇刊，16(1):140–150，2022。[44] Yang Zhou、Siying Wang、Wenyu Chen、Ruoning Zhang、Zhitong Zhao 和 Zixuan Zhang。用于多智能体强化学习的双重蒸馏网络。*arXiv* 预印本 *arXiv:2502.03125*，2025。

附录 A 环境详情

A. MARINE

一种海上物流环境，其中路径规划智能体在动态海洋条件下航行，而后勤智能体则提供海上加油支持（见图8）。该环境使用 WaveWatchIII 预报数据，使规划与协调面临天气诱发的不确定性。路径规划智能体必须在燃料耗尽前抵达目的地，因为燃料耗尽会导致回合终止。

路径规划智能体和后勤智能体都在 $n \times n$ 部分可观测性下运行，因此能够感知附近的波高和智能体位置。它们共享一个由五种基本移动动作组成的动作空间：向上、向下、向左、向右和停留。除停留外，每次移动都会消耗一个单位的燃料。当路径规划智能体和后勤智能体占据同一网格单元时，就会进行加油。加油效率由四级量化的波高决定：在最低等级时不会进行加油，而在最高等级时，加油量最高可达容量的 50%。

奖励结构旨在鼓励协调并抑制不活动。两类智能体在每个时间步都会受到小额惩罚；如果路径规划智能体耗尽燃料，则会受到更大的惩罚。路径规划智能体成功抵达目的地后可获得奖励，并会根据其与目标的接近程度受到惩罚。后勤智能体获得的奖励与其转移的燃料量成正比。其他环境规范见表 IV。

B. FireCommander（FC）

一种受 FARSITE 模型启发的基于网格的野火环境，其中行动智能体可以扑灭火灾，但不具备感知能力；感知智能体则监测火势蔓延，但无法扑灭火灾（见图 9）。该环境涉及部分可观测性下的协作以及受风影响的动态火灾传播。行动智能体只能扑灭感知智能体发现的火灾。

感知智能体在 $n \times n$ 部分可观测性下运行，能够感知附近火灾和智能体的位置，并共享由五种基本移动动作组成的动作空间：上、下、左、右和停留。行动智能体缺乏视觉能力，但可以获取自身的全局位置；它们共享这五种基本移动动作，同时还拥有一个用于扑灭火灾的特殊行动。在本研究中，为加快对火灾传播的监测，引入了火灾概率信息的熵。

奖励结构鼓励合作，并惩罚不活动。感知智能体和行动智能体都会根据活跃火灾的数量受到少量惩罚。行动智能体成功扑灭火灾后会获得奖励。其他环境规范见表 IV。

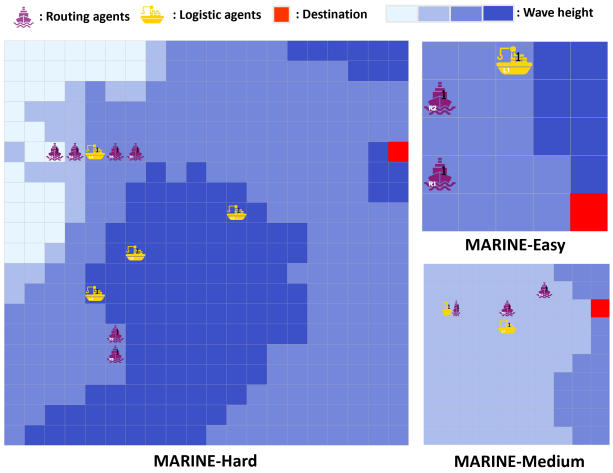

图 8. MARINE 在不同难度级别（简单、中等、困难）下的示意图。

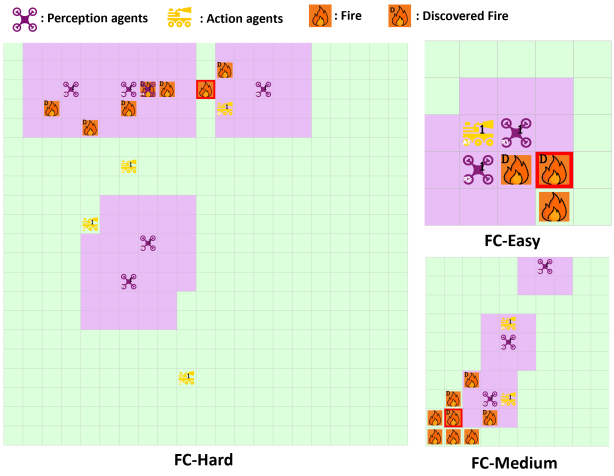

图 9. 不同难度级别（简单、中等、困难）下的 FireCommander（FC）示意图。

附录 B 教师与学生设计详情

A. 教师模型：*HiT-MAC*

1) 构建层级结构：我们展示了 MARINE 和 FC 中分层结构的构建方式（见图10）。在这两个领域中，我们通过采用类别专属编码器来处理异构观测和子目标。为清楚地区分分配给不同智能体类型的子目标，我们使用智能体的类型字符后接一个数字，例如 R_1 或 P_1，而不使用 g_i 之类的通用表示法。

首先，在 MARINE 领域中，路径规划智能体旨在到达后勤智能体或其目的地，而后勤智能体则专注于支持路径规划智能体。路径规划智能体会根据其子目标选择行为，例如，在燃料水平充足时直接前往目的地（例如，$R_2 = (0, 0, 1)$），或在燃料不足时寻求后勤智能体加油（例如，$R_1 = (1, 0, 0)$）。与此同时，后勤智能体可能会优先为燃料不足的路径规划智能体加油（例如，$L_2 = (1, 0, 0)$），或同时支持多个路径规划智能体

表 IV MARINE 和 FireCommander（FC）的环境配置详情。

领域	场景	维度	FOV	# 路径规划 /感知	# 后勤 /行动	# 目的地 /初始火灾	最大步数	初始燃料 （用于路径规划）	子区域大小 （用于教师）
MARINE	Easy	5x5	3x3	2	1	1	50	5	-
	中等	10x10	3x3	3	2	1	100	10	-
	Hard	20x20	5x5	6	4	1	200	20	-
FC	Easy	5x5	3x3	2	1	1	50	-	5x5
	中等	10x10	3x3	3	2	1	100	-	5x5
	Hard	21x21	5x5	6	4	1	210	-	7x7

表 V

预训练集　　MARINE 和 FireCommander（FC）中的教师更新。H 和L 分别代表高层级和低层级　　L.

领域	场景	时间步 （高/低）	学习 Rate	熵 系数	折扣 因子	优化器	k 步数（H）	线程数 (H/L)	训练时间 (H/L)
MARINE	Easy	0.8/0.1e7	5e-4	0.01	0.9	Adam	1	16/10	2.61/0.15 小时
	中等	1.0/0.5e7	5e-4	0.01	0.9	Adam	3	16/10	3.97/0.89 小时
	Hard	3.4/0.6e7	5e-4	0.01	0.9	Adam	5	16/10	11.04/1.06 小时
FC	Easy	0.25/0.15e7	5e-4	0.01	0.9	Adam	1	16/10	0.81/0.37 小时
	中等	0.3/0.5e7	5e-4	0.01	0.9	Adam	3	16/10	1.81/1.44 小时
	Hard	0.5/1.3e7	5e-4	0.01	0.9	Adam	5	16/10	4.24/2.22 小时

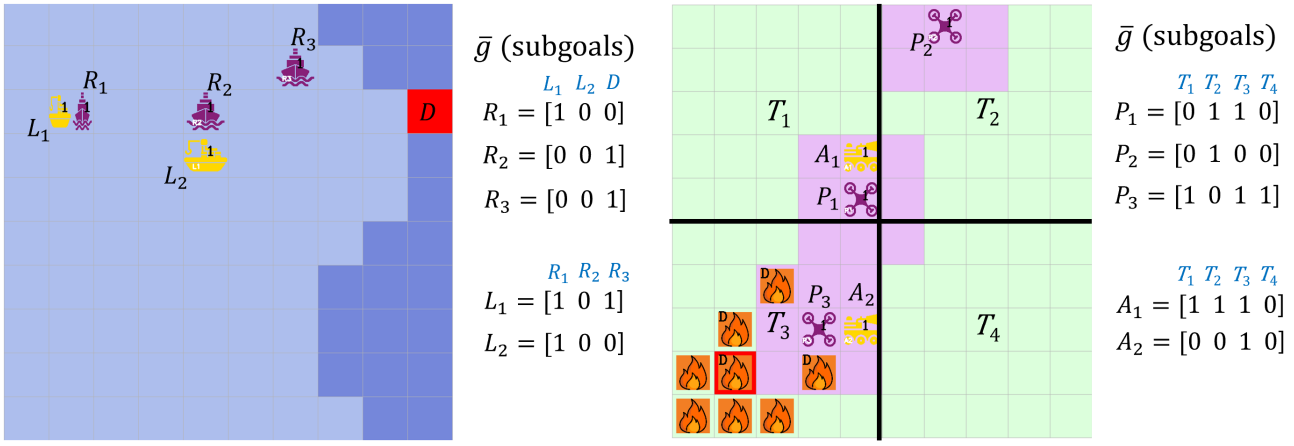

图10. MARINE和FC（中等场景）中层级结构的示例。在MARINE中，R、L和D 分别表示路径规划智能体、后勤智能体和目的地。在FC中，P、A和T 分别表示感知智能体、行动智能体和子区域。

（例如，$L_1 = (1, 0, 1)$）。所有智能体类型共享相同的智能体—目标观测结构，其中包括智能体和目标的索引、智能体或目标的燃料水平、到目标的距离，以及局部波高信息。

接下来，在FC领域中，环境被划分为多个子区域，这些子区域作为感知智能体和行动智能体的共享目标，分别用于监测和灭火任务。智能体可以广泛覆盖多个子区域（例如，$P_3 = (1, 0, 1, 1)$或 $A_1 = (1, 1, 1, 0)$），也可以专注于特定区域（例如，$P_2 = (0, 1, 0, 0)$或 $A_2 = (0, 0, 1, 0)$）。智能体-目标观测通常包含智能体及其分配目标的索引。此外，感知智能体会接收子区域内火灾概率分布中的最大熵位置，而行动智能体会接收最近的已发现火灾位置以及该子区域内已发现火灾的密度。

2）训练设置：在训练过程中，我们采用分层强化学习中常用的两阶段策略，以提高稳定性并效率。首先，独立训练低层策略，使其能够有效执行基础任务。然后，训练高层策略来协调这些行为，避免尚未成熟的低层技能受到干扰。请注意，高层策略每隔 k 步与低层策略交互一次，以降低计算开销。在低层训练期间，我们会随机设置智能体位置和环境因素，例如火灾的数量和位置，以促进泛化。在此阶段，训练主要使智能体能够抵达分配目标；对于 FC领域中的行动智能体，还会因扑灭火灾而获得额外奖励。请注意，教师的训练基于 A3C框架 [24]，并采用多个线程来提高计算效率。教师的预训练设置详情请参阅表 V，对应的学习曲线请参阅图11。

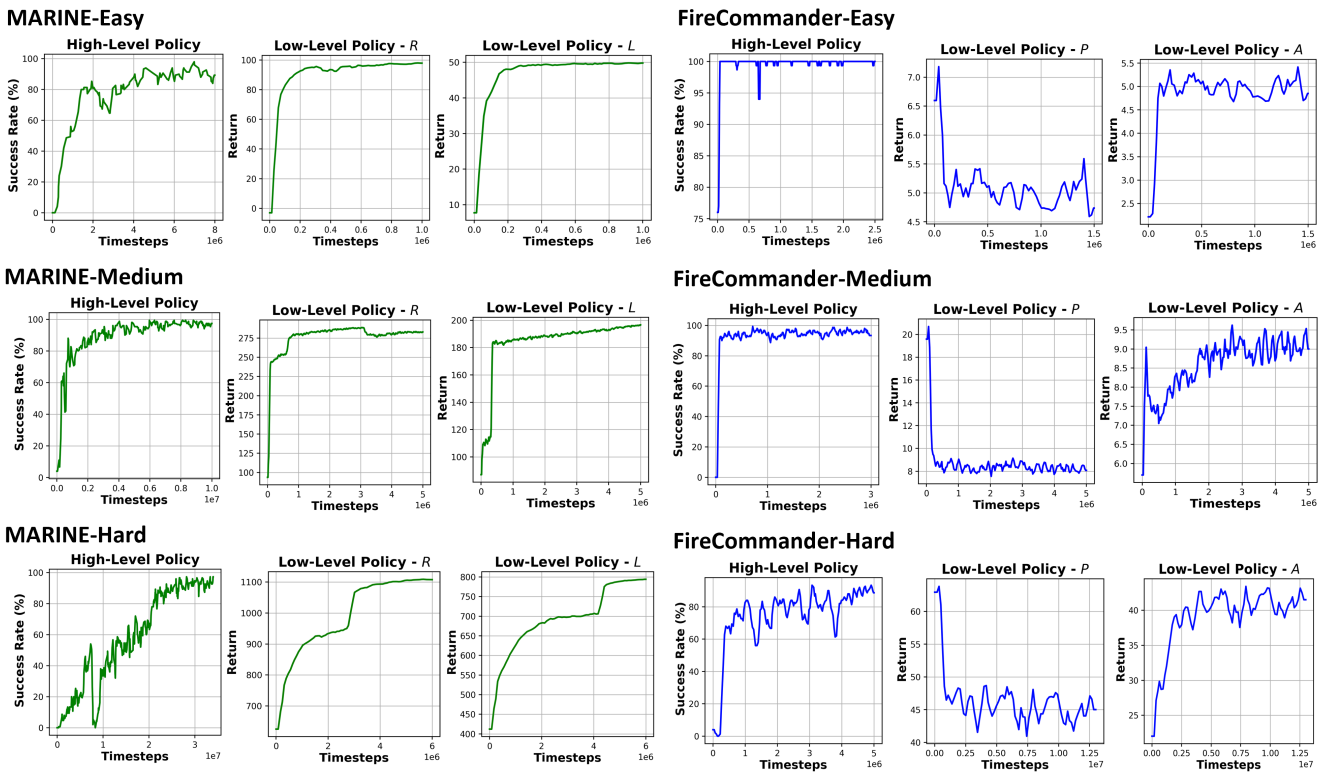

图11. MARINE 和 FireCommander 中教师策略的学习曲线。在 MARINE 中，*R* 和*L* 分别表示路径规划智能体和后勤智能体。在 FireCommander 中，*P*和*A*分别表示感知智能体和行动智能体。

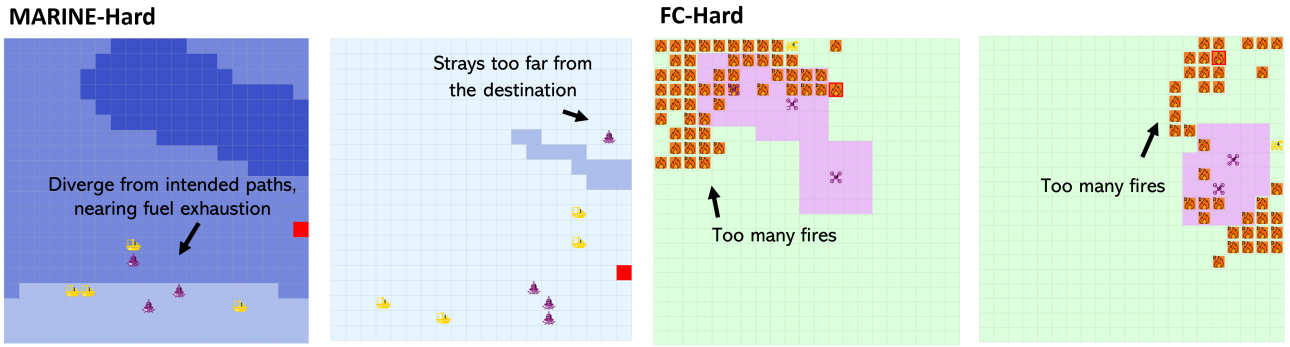

图12. 当学生状态偏离训练分布时，教师可能提供次优示范的示例；尤其是在环境和团队规模扩大时（例如 MARINE-Hard、FC-Hard）。

B. 学生模型：*HetNet*

我们指出，学生模型采用了最新版 HetNet [30]，因为它支持张量化观测结构，能够适应不同的团队配置和环境规模，具有良好的可扩展性。例如，如果智能体的视野为 3×3，就可以通过结构化张量形式捕获附近的智能体位置，以及波高或火灾位置等环境特征。预处理单元使用卷积神经网络（CNN）层处理这些空间数据。为支持异构智能体，预处理单元、长短期记忆网络模块和解码器采用类别专属权重，同类型的智能体共享这些权重以提高效率。此外，我们加入了两个 HetGAT层，每层采用双头多多头注意力，以增强通信表达能力和训练稳定性。

附录 C 备份：环境扩展时的学生状态分布

为比较图3中学生与教师的状态分布，我们从 HiT-MAC 的自注意力模块中提取全局上下文特征，并通过主成分分析（PCA）将其投影到共同的潜在空间中。随后，我们应用核密度估计（KDE） [39] 来估计其分布。对于 MARINE 和 FireCommander，我们分别从简单、中等和困难场景下的教师与学生滚动轨迹中采样 5,000、10,000 和 15,000 个状态。

表 VI MARINE 的训练设置。*E*、*M* 和*H* 分别对应简单、中等和困难任务。

方法	时间步 (*E, M, H*)	学习 Rate	折扣 因子	熵 系数	优化器	批量大小	轨迹数 在数据集 (*E, M, H*)	# 专家查询的回合数 专家查询 每个训练周期 (*E, M, H*)	教师训练的回合数 教师训练 每个训练周期 (*E、M、H*)	线程数 (*E、M、H*)	训练时间 (*E, M, H*)
多智能体运输策略优化	1.9/3.5/14e7	1e-4	0.97	0.01	Adam	200步	-	-	-	10/20/20	11.57/8.69/58.34 小时
TarMAC	1.9/3.5/14e7	1e-4	0.97	0.01	RMSprop	200步	-	-	-	10/20/20	4.74/14.67/54.25 小时
IC3Net	1.9/3.5/14e7	1e-4	0.97	0.01	RMSprop	200步	-	-	-	10/20/20	6.46/13.21/69.80 小时
CommNet	1.9/3.5/14e7	1e-4	0.97	0.01	RMSprop	200步	-	-	-	10/20/20	4.37/13.91/67.90 小时
HAPPO	1.9/3.5/14e7	1e-4	0.97	0.01	Adam	200步	-	-	-	10/20/20	13.27/29.14/313.19 小时
HetNet	1.9/3.5/14e7	1e-4	0.97	0.01	Adam	200步	-	-	-	10/20/20	21.40/32.12/283.17小时
CTDS	5/8/20e7	1e-4	0.97	0.01	Adam	10 个回合	-	-	-	10/20/20	31.12/27.20/305.12小时
IGM-DA	5/8/20e7	1e-4	0.97	0.01	Adam	10 个回合	-	-	-	10/20/20	28.29/54.45/329.29小时
PTDE	5/8/20e7	1e-4	0.97	0.01	Adam	10 个回合	-	-	-	10/20/20	63.86/31.42/230.08小时
HINT（我们的方法）	1/2/10e7	1e-4	-	0.01	Adam	200步	2000/3000/6000	20/100/200	10/150/600	10/20/20	11.46/12.18/133.90 小时

表 VII FireCommander 的训练设置。*E*、*M* 和*H* 分别对应简单、中等和困难任务

S.

方法	时间步 (*E, M, H*)	学习 Rate	折扣 因子	熵 系数	优化器	批量大小	# 轨迹 在数据集中 (*E, M, H*)	# 专家查询所需的 专家查询 每个训练周期 (*E、M、H*)	# 回合数： 教师训练 每个训练周期 (*E、M、H*)	线程数 (*E, M, H*)	训练时间 (*E, M, H*)
多智能体运输策略优化	1.4/2.8/8.8e7	1e-4	0.97	0.01	Adam	200步	-	-	-	10/20/20	3.62/12.83/70.17 小时
TarMAC	1.4/2.8/8.8e7	1e-4	0.97	0.01	RMSprop	200步	-	-	-	10/20/20	5.58/13.76/42.98 小时
IC3Net	1.4/2.8/8.8e7	1e-4	0.97	0.01	RMSprop	200步	-	-	-	10/20/20	5.06/13.14/68.75 小时
CommNet	1.4/2.8/8.8e7	1e-4	0.97	0.01	RMSprop	200步	-	-	-	10/20/20	4.87/12.35/59.49 小时
HAPPO	1.4/2.8/8.8e7	1e-4	0.97	0.01	Adam	200步	-	-	-	10/20/20	3.98/14.3/172.31 小时
HetNet	1.4/2.8/8.8e7	1e-4	0.97	0.01	Adam	200步	-	-	-	10/20/20	11.80/33.07/178.78 小时
CTDS	5/6/15e7	1e-4	0.97	0.01	Adam	10 个回合	-	-	-	10/20/20	29.11/13.52/224.29 小时
IGM-DA	5/6/15e7	1e-4	0.97	0.01	Adam	10 个回合	-	-	-	10/20/20	24.00/18.43/248.90 小时
PTDE	5/6/15e7	1e-4	0.97	0.01	Adam	10 个回合	-	-	-	10/20/20	28.21/21.04/125.54 小时
HINT（我们的方法）	1/2/7e7	1e-4	-	0.01	Adam	200步	1000/2000/4000	20/50/100	10/20/40	10/20/20	7.76/10.06/83.12 小时

图12展示了学生轨迹偏离训练分布、导致我们的分层教师提供次优示范的示例。首先，在 MARINE-Hard环境中，路径规划智能体经常大幅偏离前往目的地或加油点的预定路线，并在与后勤智能体会合前耗尽燃料，只能原地等待后勤智能体到达。教师能够在较小的环境中有效地提供纠正性指导，引导后勤智能体前往受困的路径规划智能体处。然而，随着环境规模扩大，这类场景对教师而言愈发陌生，因为预训练的低层策略使学生在训练期间无法大幅偏离路线。因此，教师往往默认采取闲置行为，而不是主动协调加油操作；有时甚至会引导燃料不足的学生继续前往目的地。

同样，在 FC-Hard 场景中，学生策略的表现往往不如教师策略，导致火灾数量大幅增加；而教师在训练期间很少遇到这种情况。在较小的环境中，火灾通常不会蔓延得那么广，因此这类情况较少见。由于教师策略通常没有接触过这种不断升级的情况，因此难以作出有效应对。缺乏这类经验会导致行动智能体之间的协调效率低下，使它们聚集在特定区域，而不是有效地分散到多个子区域。

附录D 训练设置

本节将介绍用于基准评估的实验设置，并提供若干补充结果。我们的方法在多台机器上进行训练，包括配备 AMD EPYC 7513 中央处理器和 NVIDIA A40 图形处理器的系统、配备 Intel i9-13980HX 中央处理器和 RTX 4070 笔记本电脑图形处理器的系统、配备 Intel i9-13900 中央处理器和 RTX 3070 图形处理器的系统、配备 Intel Ultra 9 285K 中央处理器和 RTX 5080 图形处理器的系统，以及配备 Intel i7-12650H 中央处理器和 RTX 3060 笔记本电脑图形处理器的系统。为确保可复现性，所有实验均使用完全相同的 conda 环境，并在所有系统中采用一致的库版本。评估时，我们在 50 个回合上测试结果，并取三个不同随机种子的平均值。不同环境配置下的训练设置见表 VI 和表 VII。请注意，CTDE 基线和 HINT 在每次梯度更新中使用 200 个样本作为批量大小，而 KD 基线将回合作为批次单位，每批使用 10 个回合。在 MARINE-简单中，KD 基线平均每回合有 15 步，而在其他所有设置中，其平均每回合至少经历 20 步。因此，与 HINT 使用的 200 个时间步相比，KD 基线使用更多样本进行训练。

附录 E 补充材料：基准测试

本节展示完整的基准测试结果。表 IX 和表 VIII 报告了 MARINE 和 FireCommander（FC）上集中式训练、分散式执行和知识蒸馏的基准测试结果。图 13 展示了在线 CTDE 基准测试中 MARINE 和 FireCommander 的学习曲线。

附录 F 补充材料：消融实验

关于 HINT 中关键模块的作用，我们在表 X 和图 14 中给出了完整的消融结果。表 X 报告了 MARINE 和 FireCommander 在 HINT

表 VIII FireCommander（FC）在在线集中训练、分散执行及知识蒸馏基准上的完整定量结果，报告为三个随机种子的均值（± 标准差）。工作站 表示使用我们的教师策略对强化学习进行热启动。

方法	FC-Easy（5×5，3 个智能体）		FC-Medium（10×10，5 个智能体）		FC-Hard（21×21，10 个智能体）	
	成功率（%）↑	步数 ↓	成功率（%）↑	步数 ↓	成功率（%）↑	步数 ↓
多智能体近端策略优化	32.67 ± 9.45	42.47 ± 1.90	6.00 ± 3.46	96.33 ± 1.19	4.00 ± 0.00	202.89 ± 0.86
TarMAC	42.67 ± 9.02	42.20 ± 0.11	5.33 ± 5.03	96.89 ± 3.97	2.67 ± 3.06	207.51 ± 2.48
IC3Net	26.67 ± 3.06	43.03 ± 0.46	6.67 ± 6.11	97.07 ± 2.79	2.67 ± 3.06	207.08 ± 2.68
CommNet	68.67 ± 6.11	34.44 ± 1.93	3.33 ± 3.06	97.81 ± 1.92	1.33 ± 2.31	207.96 ± 1.77
HAPPO	1.33 ± 1.15	49.36 ± 0.55	0.67 ± 1.15	99.35 ± 1.13	0.67 ± 1.15	208.61 ± 2.40
HetNet	94.00 ± 2.00	22.06 ± 3.02	82.00 ± 14.00	49.43 ± 16.00	19.33 ± 7.02	177.93 ± 11.91
→ 25% 工作站	-	-	-	-	35.33 ± 15.28	152.35 ± 18.78
→ 50% 工作站	-	-	-	-	42.67 ± 38.28	139.01 ± 62.69
CTDS	18.89 ± 7.95	45.13 ± 3.04	1.67 ± 2.89	98.49 ± 2.61	0.83 ± 1.44	208.38 ± 2.80
IGM-DA	16.22 ± 6.74	44.77 ± 2.13	2.78 ± 1.27	98.01 ± 2.09	1.11 ± 1.27	207.80 ± 2.59
PTDE	24.00 ± 20.30	44.17 ± 4.51	3.61 ± 3.37	97.56 ± 3.22	2.71 ± 2.01	206.41 ± 4.25
HINT（我们的方法）	**99.33 ± 1.15**	**15.69 ± 1.15**	**84.00 ± 2.00**	**47.95 ± 3.73**	**51.33 ± 5.03**	**136.40 ± 3.86**

表 IX 在线集中训练、分散执行及知识蒸馏基准中 MARINE 的完整定量结果，以三个随机种子的均值（± 标准差）报告。*WS* 表示使用我们的教师策略对强化学习进行热启动。

方法	MARINE-简单（5x5，3 个智能体）		MARINE-中等（10x10，5 个智能体）		MARINE-Hard（20x20，10 个智能体）	
	成功率（%）↑	步数 ↓	成功率（%）↑	步数 ↓	成功率（%）↑	步数 ↓
多智能体近端策略优化	**99.33 ± 1.15**	9.99 ± 0.38	40.67 ± 51.39	69.23 ± 37.31	0.67 ± 1.15	198.88 ± 1.94
TarMAC	97.33 ± 3.06	9.43 ± 2.40	94.00 ± 10.39	**25.84 ± 14.77**	0.00 ± 0.00	200.0 ± 0.00
IC3Net	98.67 ± 1.15	**9.10 ± 2.39**	88.00 ± 10.58	34.97 ± 17.99	0.00 ± 0.00	200.0 ± 0.00
CommNet	98.00 ± 2.00	11.09 ± 1.88	0.00 ± 0.00	100.00 ± 0.00	0.00 ± 0.00	200.0 ± 0.00
HAPPO	97.33 ± 3.06	10.82 ± 1.04	98.67 ± 2.31	27.81 ± 0.46	0.00 ± 0.00	200.0 ± 0.00
→ 25% 工作站	-	-	-	-	24.00 ± 3.46	169.94 ± 9.25
→ 50% 工作站	-	-	-	-	38.00 ± 6.93	154.73 ± 8.30
HetNet	85.33 ± 14.19	17.66 ± 5.02	61.33 ± 53.27	60.33 ± 34.64	0.00 ± 0.00	200.0 ± 0.00
CTDS	76.00 ± 13.38	22.24 ± 1.69	23.06 ± 25.23	85.90 ± 11.16	7.22 ± 11.10	191.54 ± 13.06
IGM-DA	93.33 ± 3.06	13.79 ± 1.42	80.00 ± 11.46	42.95 ± 7.40	14.17 ± 16.65	181.36 ± 21.99
PTDE	98.22 ± 0.38	9.69 ± 1.47	96.11 ± 1.73	32.68 ± 1.01	33.33 ± 57.74	147.05 ± 91.71
HINT（我们的方法）	98.67 ± 1.15	11.39 ± 1.21	**98.67 ± 1.15**	26.49 ± 4.92	**53.33 ± 9.02**	**143.28 ± 4.47**

及其三个变体下的定量结果，而图 14 展示了学生和教师在所有环境设置下的训练进度。结果表明，具有自适应能力的教师能够有效帮助学生在更具挑战性的领域中进行训练，例如 MARINE-Hard、FC-Medium 和 FC-Hard。

我们在图 15 中使用行为克隆（BC）考察批量大小对训练稳定性的影响。BC 使用与 HINT 相同的学生策略和相同的教师示范。为了隔离每次更新的影响，我们按训练迭代绘制各项指标，其中每次迭代对应一次批次更新。在 MARINE-Hard 和 FC-Hard 中，较大的批量大小都能带来更稳定、更有效的学习。

附录 G 训练时间比较

如第 VI-D 节所述，HINT 融合了伪离策略强化学习、基于性能的筛选、数据集聚合以及基于 HetGAT 的通信等多种模块，这可能会带来更高的计算成本。在此，我们比较 HINT 与有竞争力的基线方法（如 HAPPO、HetNet 和 PTDE）在 MARINE-Hard 和 FC-Hard 任务上的性能和训练时间。请注意，HINT 的总训练时间包含预训练时间。对于 PTDE，其训练时间经过插值，以便与 CTDE 基线的时间步数相匹配，因为 PTDE 的训练时间步数远多于 CTDE 基线方法和 HINT。表 XI 显示，与 HAPPO 和 HetNet 相比，HINT 的训练时间更短，同时性能显著提升。此外，HINT 的训练时间与 PTDE 相当且性能有显著提升。

为了探究各个关键模块对 HINT 训练时间的影响，我们比较了 HINT 与其变体的训练时间，如表 XII 所示。大多数变体在数据集聚合期间通过完整回合执行教师策略，以跟踪次优示范率（图 14），这使得难以隔离基于性能的筛选的确切开销。作为替代，我们将完整模型与两个特定变体进行比较：一个仅包含通信（例如行为克隆，BC），另一个不包含伪离策略强化学习。通过这项比较，我们揭示了交互式蒸馏模块（伪离策略强化学习、基于性能的筛选和数据集聚合）的整体效果，以及伪离策略强化学习的独立效果。

总体而言，完整模型的训练时间约为行为克隆的两倍，这表明交互式蒸馏模块显著增加了训练开销。然而，行为克隆的性能会下降 40–70%，因此，尽管交互式蒸馏会增加成本，采用它仍是合理的。将完整模型与不包含伪离策略强化学习的版本进行比较，二者的训练时间相近。这表明，大部分开销来自数据集聚合和基于性能的筛选，而非伪离策略强化学习本身。这些发现凸显了我们需要更复杂的推理模块来评估教师演示的质量，而不是依赖资源密集型的基于性能的筛选。此类改进有望在保持强劲性能的同时，大幅降低计算成本。

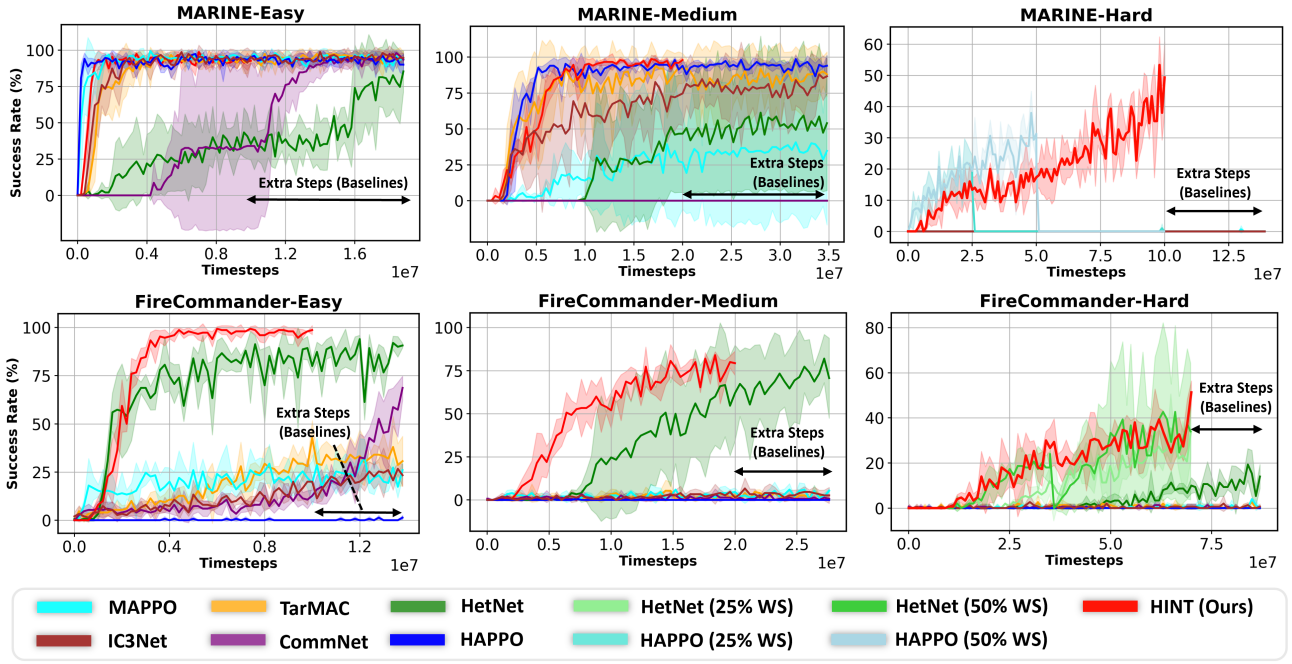

图 13. MARINE 和 FireCommander 在在线 CTDE 基准测试中的学习曲线，结果为三个随机种子和三种难度设置下的均值（± 标准差）。HINT 在两个领域中始终优于所有基线方法。

表 X 消融实验中 MARINE 和 FireCommander（FC）的完整定量结果，报告为三个随机种子下的均值（±标准差）。

方法	MARINE-简单（5x5，3 个智能体）		MARINE-中等（10x10，5 个智能体）		MARINE-Hard（20x20，10 个智能体）	
	成功率（%）↑	步数 ↓	成功率（%）↑	步数 ↓	成功率（%）↑	步数 ↓
我们的方法（两者均不包含）	97.33 ± 1.15	12.23 ± 0.91	96.67 ± 2.31	27.78 ± 3.76	22.00 ± 2.00	171.67 ± 2.42
我们的方法（不含基于性能的筛选）	97.33 ± 1.15	11.95 ± 0.29	**99.33 ± 1.15**	26.81 ± 2.26	23.33 ± 13.32	168.49 ± 15.68
我们的方法（不含伪离策略强化学习）	98.00 ± 0.00	**10.61 ± 1.92**	98.00 ± 2.00	**25.67 ± 0.75**	36.67 ± 10.26	156.71 ± 9.70
我们的方法（完整模型）	**98.67 ± 1.15**	11.39 ± 1.21	98.67 ± 1.15	26.49 ± 4.92	**53.33 ± 9.02**	**143.28 ± 4.47**

方法	FC-Easy（5x5，3 个智能体）		FC-Medium（10x10，5 个智能体）		FC-Hard（21x21，10 个智能体）	
	成功率（%）↑	步数 ↓	成功率（%）↑	步数 ↓	成功率（%）↑	步数 ↓
我们的方法（两者均不包含）	98.67 ± 1.15	17.87 ± 4.24	73.33 ± 7.57	53.09 ± 6.10	22.00 ± 13.11	178.41 ± 16.72
我们的方法（不含基于性能的筛选）	**99.33 ± 1.15**	16.19 ± 1.48	77.33 ± 7.57	50.56 ± 8.06	28.00 ± 10.58	174.15 ± 12.53
我们的方法（不含伪离策略强化学习）	**99.33 ± 1.15**	16.97 ± 2.29	76.67 ± 6.43	50.25 ± 7.11	28.67 ± 1.15	173.09 ± 1.61
我们的方法（完整模型）	**99.33 ± 1.15**	**15.69 ± 1.15**	**84.00 ± 2.00**	**47.95 ± 3.73**	**51.33 ± 5.03**	**136.40 ± 3.86**

表 XI MARINE-Hard 和 FC-Hard 基准测试的性能和训练时间比较。对于 HINT，预训练时间计入总训练时间。PTDE 的训练时间经过插值处理，以匹配集中式训练、分散式执行基线的时间步。

方法	MARINE-Hard				FireCommander-Hard			
	成功率 (%) ↑	Δ (%)	训练时间（小时）↓	Δ (%)	成功率（%）↑	Δ (%)	训练时间（小时）↓	Δ (%)
HAPPO	0.00	-100%	313.19	+83%	0.67	-99%	172.31	+92%
HetNet	0.00	-100%	283.17	+65%	19.33	-62%	178.78	+100%
PTDE	33.33	-37%	161.06	+10%	2.71	-94%	73.65	-18%
HINT	**53.33**	—	**146.00**	—	**51.33**	—	**89.58**	—

表 XII HINT 提示模块对 MARINE-Hard 和 FC-Hard 任务的性能与训练时间的影响。

方法	MARINE-Hard				FireCommander-Hard			
	成功率（%）↑	Δ (%)	训练时间（小时）↓	Δ (%)	成功率（%）↑	Δ (%)	训练时间（小时）↓	Δ (%)
HINT（仅使用通信；行为克隆）	15.33	-71%	65.46	-55%	28.67	-44%	42.93	-52%
HINT（不含伪离策略强化学习）	36.67	-31%	149.43	+2%	28.67	-44%	76.02	-15%
HINT（完整模型）	**53.33**	—	**146.00**	—	**51.33**	—	**89.58**	—

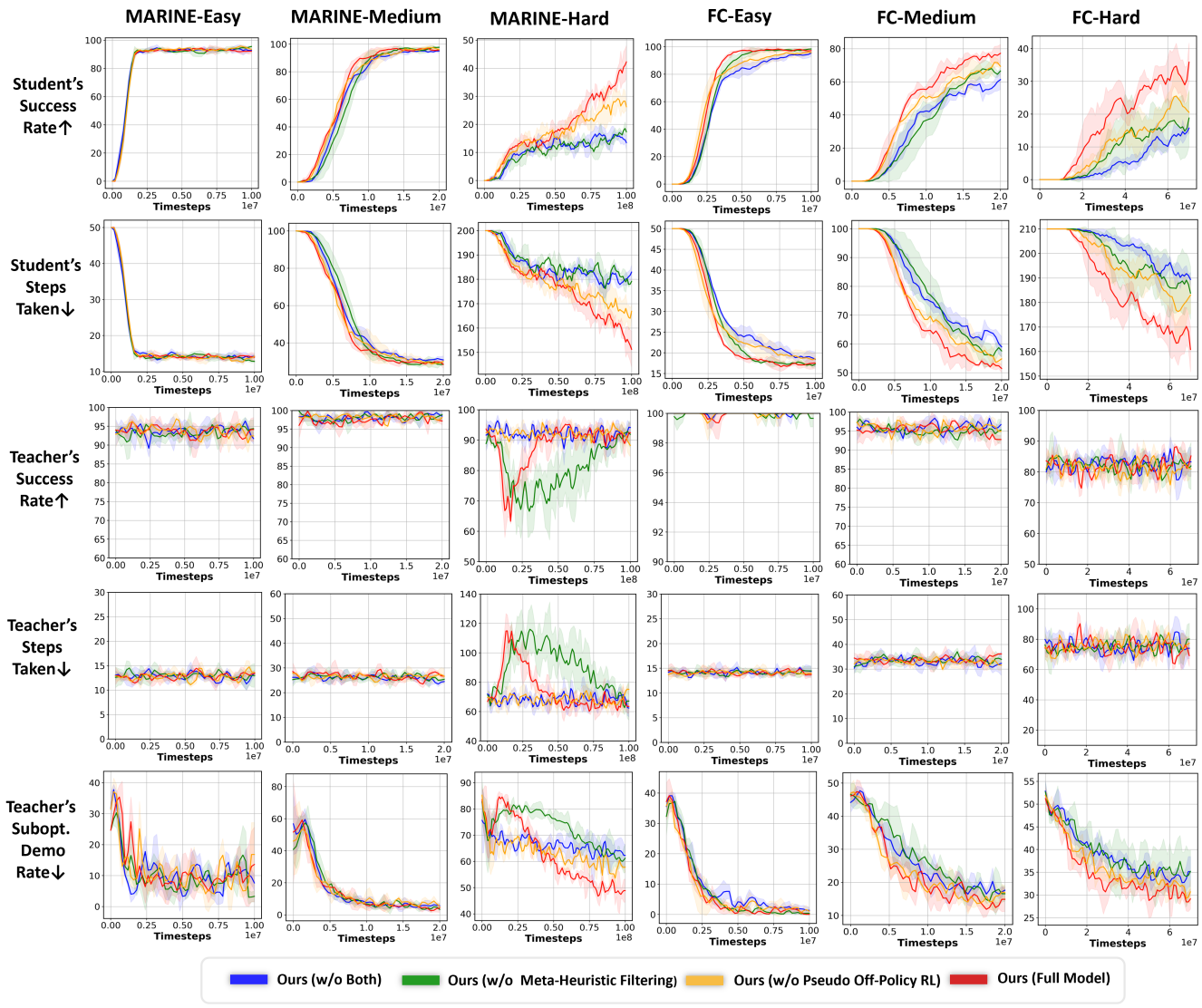

图 14. MARINE 和 FireCommander（FC）上的完整定性消融结果。每列代表一种领域难度设置。每行展示一个评估指标随训练时间步变化的情况（横轴）。曲线为三个随机种子的均值，阴影区域表示标准差（滑动窗口长度=5）。

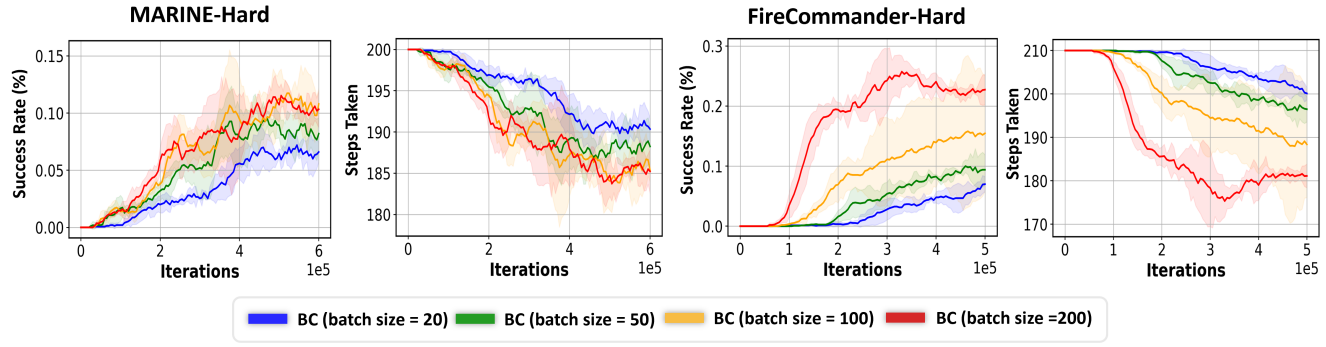

图 15. MARINE-Hard 和 FireCommander-Hard 上不同批量大小下的行为克隆结果，以均值（± 标准差）表示。每次迭代对应使用一个数据批次执行的一次更新步骤。