arXiv:2611.32645v1 [cs.LG] 24 Nov 2026

TeamFlowRL：面向多智能体协调的去中心化流策略在线策略优化

Guowei Zou[1],Beiwen Zhang[1],Haitao Wang[1], Hong Hong[1], Zhiquan Chen[1], Hejun Wu [1]

[1]中山大学

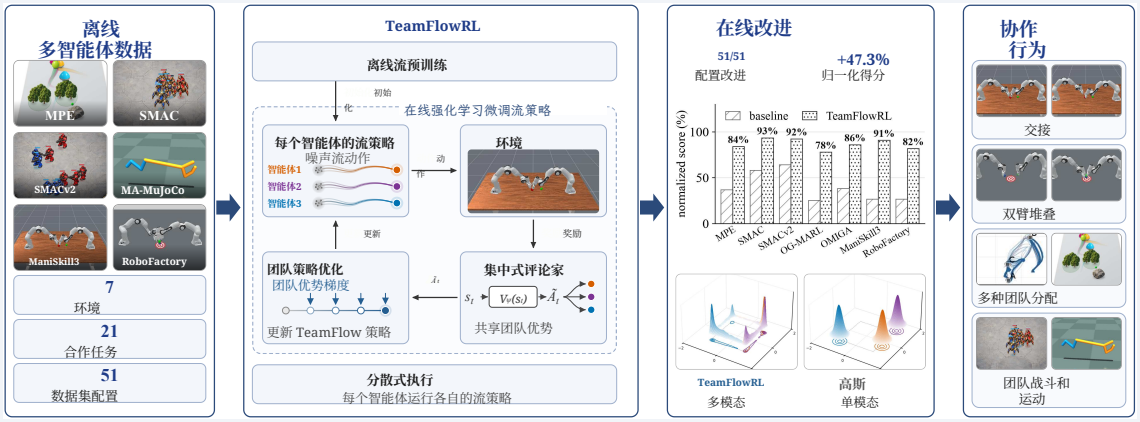

图 1 TeamFlowRL 概览。左：7 个环境、21 个合作任务和 51 种数据集配置。中：预训练流策略通过集中式评论家和团队策略优化进行在线微调，随后每个智能体根据局部输入执行自身的流策略。右：全部 51 种配置均得到改进（分数按任务归一化）。流动作具有多模态特性，而高斯基线（行为克隆+MAPPO）是单模态的，智能体能够学习协作行为。

项目主页： https://teamflowrl.github.io/**代码：**
https://github.com/teamflowrl/TeamFlowRL**模型：**
https://huggingface.co/teamflowrl/TeamFlowRL

摘要

复杂合作任务需要协调一致的多智能体行为。离线生成策略能够从轨迹中学习表达力丰富的协作行为，但仍受固定数据集的限制。在线强化学习可以通过与环境交互来提升这些策略。然而，微调流策略颇具挑战，因为动作似然难以计算。我们提出 TeamFlowRL，这是一种两阶段框架，先离线预先训练多智能体流策略，再在线对其进行微调。我们基于去中心化流策略的完整去噪轨迹制定团队策略优化，从而避免难以处理的动作似然。由此得到的目标函数使用共享团队优势对每个智能体的轨迹比率加权，并分别裁剪每个比率，使局部策略更新与共同团队回报保持一致。在七个仿真组的 51 组配对配置中，TeamFlowRL 提升了每一种离线初始化的表现，归一化得分平均提高 47.3%。此外，在 6 项双手任务中，与高斯策略微调相比，TeamFlowRL 在探索噪声下的平均成功率高出 33.1%。

1 引言

协作式多智能体任务（如机器人操作和协同控制）要求智能体协调各自的决策。多智能体强化学习（MARL）通过互动来学习这种协作，

[14, 23, 32]，而离线生成策略则从固定数据集中学习具有丰富表达能力的动作或轨迹分布 [9, 34]。然而，离线策略仍受限于可用数据的质量和覆盖范围 [10]。

在线强化学习可以通过环境交互和团队奖励进一步提升协作。在集中式训练和分散式执行框架下，现有策略梯度方法通常对离散动作使用分类分布，对连续动作使用单峰高斯分布 [4, 32]。这类参数化方式无法为同一观测表示彼此分离的动作模态，而离线流策略则能够建模更丰富的多模态行为。因此，可以利用在线团队反馈对表达能力强的离线流策略进行微调。

将单智能体流策略方法扩展到多智能体流策略面临两大严峻困难。首先，流策略的动作概率无法进行解析计算。PPO类方法需要该概率来计算新旧策略在同一动作上的比率，以便进行策略更新和改进 [27]。由于流策略的许多去噪轨迹都会到达同一个动作，其对数似然无法解析计算；对于多智能体而言，计算上也不可行。其次，即使假设能够得到该比率，现有多智能体方法（例如 MAPPO）的团队更新方式也无法再直接应用于流策略。在 MAPPO 中，每个智能体根据自身新旧策略的比率，并结合共享团队优势来更新策略 [32]。而在流策略中，许多去噪轨迹都可能到达同一个动作，团队优势仍然评估最终动作。因此，比率与团队优势的乘积仅对最终动作有意义，而对单条轨迹没有意义。即使某种方案使用轨迹之间的比率，而非动作之间的比率，也无法确定该方案是否优化了团队目标，因为轨迹与最终动作之间并非一一对应，而是多对一的关系。

为应对这些挑战，我们提出了 TeamFlowRL，这是一种用于在线微调多智能体离线训练流策略的框架。图 1 展示了该框架。我们通过具有可处理似然的随机去噪轨迹来表示每个智能体的策略。随后，我们构建了一个团队目标，通过来自集中式评论家的共享优势，将这些局部轨迹联系起来。每个智能体的完整轨迹比率分别进行裁剪，而不是裁剪所有智能体比率的乘积。在具有精确团队优势的情况下，该代理目标与采集策略处联合代理目标的梯度一致（经过归一化后），详见附录 A.4。这将去中心化流更新与团队回报优化联系起来。

我们在五个多智能体基准套件和两个双手操作模拟器上的 51 个任务--数据集配置中评估了 TeamFlowRL。在线微调在全部 51 个配置中均优于离线初始化，归一化得分平均提升 47.3%。我们还将 TeamFlowRL 与高斯策略微调进行了比较。行为分析表明，微调后的流策略能够捕捉多模态动作分布，并表现出多样的协作行为。在六项双手任务中，在探索噪声下，TeamFlowRL 的平均成功率比行为克隆+MAPPO 高 33.1%。

我们的贡献主要有三点：

- 据我们所知，TeamFlowRL 是首个用于对离线预训练的多智能体流策略进行在线强化学习微调的框架。
- 我们针对去中心化去噪轨迹制定了团队策略目标，将共享团队优势与对每个智能体的完整轨迹比率分别进行裁剪相结合。在使用精确优势时，该代理目标在采集策略处的梯度与联合代理目标的梯度一致（仅相差归一化因子）。
- 在 51 种配置下，TeamFlowRL 始终能够改进离线初始化，并在多模态动作表达能力和在线适应效率方面优于高斯策略。

2 相关工作

离线多智能体策略学习。 离线多智能体强化学习从固定数据集中学习合作策略，使用行为克隆(行为克隆) [20] 或基于价值和保守目标的方法，例如 TD3+行为克隆、保守Q学习和批量约束深度Q学习 [6-

表1 与现有方法的比较。

方法	策略表示	策略优化	在线	多智能体
DPPO [24]	扩散	基于去噪转移的近端策略优化	✓	---
ReinFlow [33]	Flow	基于去噪轨迹的近端策略优化	✓	---
FPO [15]	Flow	流匹配损失比	✓	---
MAPPO [32]	分类 / 高斯	对每个智能体的动作应用近端策略优化	✓	✓
MADiff [34]	扩散	回报条件化	×	✓
DOM2 [12]	扩散	Q学习 + 行为克隆	×	✓
MAC-Flow [9]	从流中蒸馏出的执行器	价值引导蒸馏	×	✓
TeamFlowRL（我们的方法）	**预训练流生成器**	**对每个智能体的去噪-ing 轨迹**	✓	✓

-8]，OMAR 和 OMIGA 提供了其他离线多智能体强化学习变体 [17, 31]。生成式方法进一步提升了表达能力：MADiff、DOM2 和 DoF 使用扩散进行轨迹或去中心化策略学习 [11, 12, 34]，，而 MAC-Flow 学习联合流模型，并蒸馏去中心化执行器，以生成单步动作 [9]。所有方法仍受离线数据质量和覆盖范围的限制。

在线多智能体强化学习。 在线 MARL 通过与环境交互来学习协作，使用集中式评论家 [4, 14, 18]，价值分解 [23, 28, 30]，或 MAPPO 的集中式价值估计，并结合智能体级近端策略优化更新 [32]。这些方法通常采用分类或单峰高斯策略，从而限制了多模态动作表示。

离线预训练和在线微调。 在线强化学习可以通过优化预训练生成式策略的去噪过程来改进这些策略。DPPO 将近端策略优化应用于扩散转移 [24]，，而 ReinFlow 和 FPO 则利用随机去噪轨迹或流匹配损失 [15, 33]为流策略构造比率。相关工作研究了生成式机器人策略 [35, 36] 和视觉-语言-动作模型（VLA）[1, 19]，的微调，但这些方法仍然面向单智能体，未解决去中心化多智能体流策略中的团队级信用分配问题。

3 多智能体流策略优化

TeamFlowRL 分两个阶段学习用于合作任务的去中心化策略。图 2 展示了完整流程。首先，利用离线输入-动作对初始化共享流策略。在在线微调阶段，我们将流生成器转化为针对完整去噪轨迹的随机策略，从而使每个智能体采样轨迹的似然变得可计算。随后，集中式评论家估计共享团队优势，该优势与智能体级轨迹似然比相乘，形成近端策略优化目标，并联合更新共享的速度网络和噪声网络。在部署时，每个智能体利用局部信息，通过微调后的速度网络独立生成其动作。

3.1 合作问题表述

我们将合作任务建模为一个去中心化部分可观测马尔可夫决策过程，简称 Dec-POMDP，$\mathrm{I}, \mathrm{S}, \{\mathrm{A}^i\}_{i=1}^N, P, R, \{\mathrm{O}^i\}_{i=1}^N, O, \gamma$，其中 $\mathrm{I} = \{1, \ldots, N\}$ 是智能体集合，S 是全局状态空间，A^i 和 O^i 分别是局部动作空间和观测空间，P 是状态转移分布，R是共同奖励函数，O 是观测分布，$\gamma \in (0, 1)$ 是折扣因子。在环境步 t，智能体 i 使用其局部信息 h_t^i选择动作 a_t^i，所有智能体都会获得共享奖励 r_t。局部信息 h_t^i 仅包含智能体 i 可获取的特征。目标是学习联合动作策略 $\boldsymbol{\pi}$，以最大化期望折扣团队回报，$\mathbb{E}_{\boldsymbol{\pi},P}[\sum_{t\geq 0}\gamma^t r_t]$。

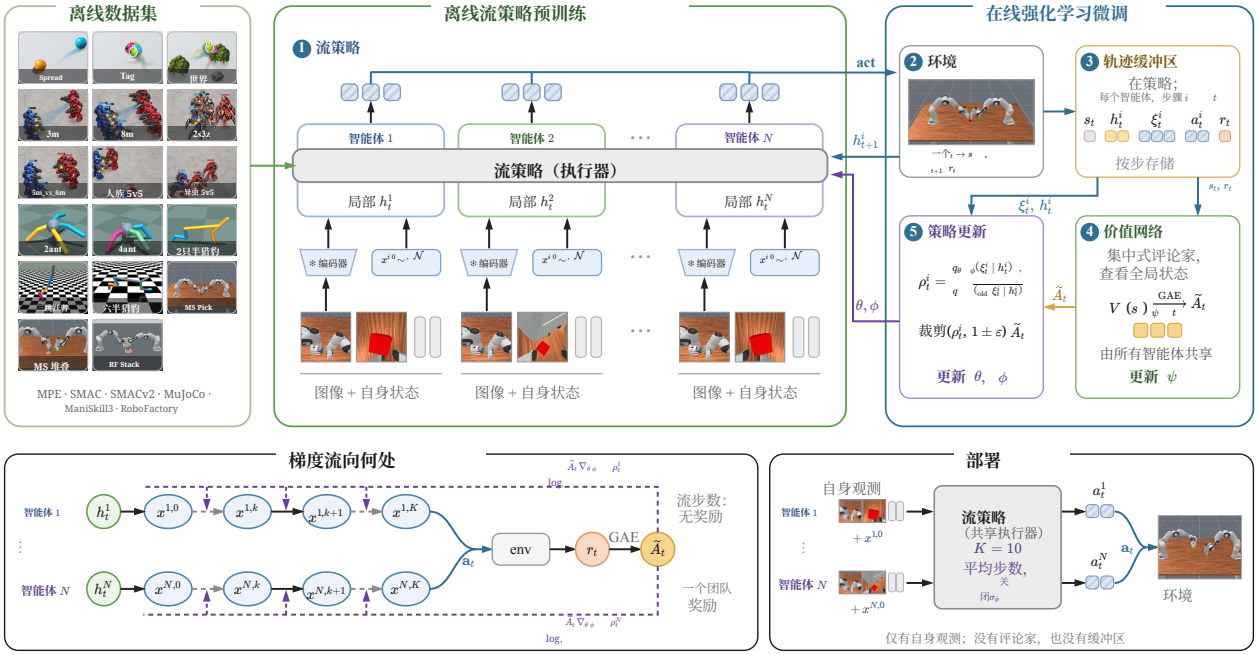

图 2 TeamFlowRL 将离线流预训练（左上和中上）与在线强化学习微调（右上）相结合。左下展示了共享团队优势如何沿每条已存储的去噪轨迹引导更新。右下展示了不含转移噪声的分散式执行。

3.2 离线流初始化

我们使用离线输入-动作对 $(h^i,\ a^i) \in \mathcal{D}$初始化共享速度网络 v_θ。在此处以及单条去噪轨迹中，我们省略环境时间 t。

条件流匹配 [13] 训练 v_θ 以预测从高斯噪声到数据动作的线性插值过程中的速度：

$$x_\tau^i = (1-\tau)z^i + \tau a^i, \qquad z^i \sim \mathcal{N}(0, I), \tag{1}$$

$$\mathcal{L}_{\mathrm{FM}}(\theta) = \mathbb{E}_{\mathcal{D},z,\tau,i}\left[\|v_\theta(h^i, x_\tau^i, \tau) - (a^i - z^i)\|_2^2\right], \tag{2}$$

其中 $\tau \sim \mathcal{U}[0,1]$ 表示流时间。对于离散动作空间，a^i 在公式 1 和 2 中表示独热动作向量。初始化后的策略通过从具有 v_θ 的高斯潜变量开始，使用 K 个欧拉步进行积分来生成动作。

3.3 去噪轨迹上的随机策略

给定局部输入和初始潜变量，欧拉流轨迹是确定性的，无法为近端策略优化提供可处理的转移似然。按照 ReinFlow 的做法，我们在每次去噪转移中加入高斯噪声，从而得到完整轨迹的可处理概率。

在每个环境步，智能体 i采样 $x^{i,\,0} \sim \mathcal{N}(0,\ I)$，并对$k = 0,\ \ldots,\ K-1$执行以下随机更新：

$$\begin{aligned} x^{i,k+1} &= \mu_\theta(h^i, x^{i,k}, k) + \sigma_\phi(h^i, x^{i,k}, k) \odot \epsilon^{i,k}, \\ \mu_\theta(h^i, x^{i,k}, k) &= x^{i,k} + K^{-1} v_\theta(h^i, x^{i,k}, k/K), \end{aligned} \tag{3}$$

具有独立的 $\epsilon^{i,\,k} \sim \mathcal{N}(0,\ I)$ 和逐坐标乘积 $\odot$。均值保留预训练欧拉更新，而共享噪声网络则学习有界的正标准差 σ_ϕ。

因此，完整的去噪轨迹 $\xi^i = (x^{i,0}, \ldots, x^{i,K})$ 具有联合概率 ity

$$q_{\theta,\phi}(\xi^i \mid h^i) = p_0(x^{i,0}) \cdot \prod_{k=0}^{K-1} \mathcal{N}\big(x^{i,k+1} \mid \mu_\theta^{i,k}, \mathrm{diag}((\sigma_\phi^{i,k})^2)\big), \tag{4}$$

其中 p_0 是标准高斯概率，而 $\mu^i_{\theta}{}^k$ 、 $\sigma^i_{\phi}{}^k$ 是方程 3 中两个函数的简写。环境只接收 $a^i = g(x^{i,K})$，其中 g 会对连续动作进行裁剪，或为离散动作选择最大的合法坐标。这种轨迹似然为策略优化提供了一种易于计算的替代方案，可替代动作边缘分布。

给定局部输入后，各智能体独立采样其初始潜变量和转移噪声。尽管采用参数共享，联合轨迹采样策略仍然可以因子分解。我们用 $M_{\theta,\phi}$ 表示这一联合轨迹策略（并用 $q_{\theta,\phi}$ 表示单个智能体的策略），团队目标为

$$\begin{aligned} M_{\theta,\phi}(\boldsymbol{\xi}_t \mid \mathbf{h}_t) &= \prod_{i=1}^{N} q_{\theta,\phi}(\xi_t^i \mid h_t^i), \\ J(\theta,\phi) &= \mathbb{E}_{M_{\theta,\phi}}\Big[\sum_{t\geq 0} \gamma^t r_t\Big]. \end{aligned} \tag{5}$$

这里， $\boldsymbol{\xi}_t = (\xi_t^1, \ldots, \xi_t^N)$ 表示联合去噪轨迹，而 $\mathbf{h}_t = (h_t^1, \ldots, h_t^N)$ 表示演员输入的元组。

3.4 团队优势与策略更新

剩下的一步是将单个轨迹似然与团队回报联系起来。方程 5中的因子分解给出

$$\nabla_{\theta,\phi} \log M_{\theta,\phi}(\boldsymbol{\xi}_t \mid \mathbf{h}_t) = \sum_{i=1}^{N} \nabla_{\theta,\phi} \log q_{\theta,\phi}(\xi_t^i \mid h_t^i). \tag{6}$$

团队策略梯度将该总和乘以一个共同的团队优势。我们采用与 MAPPO 相同的集中式优势估计和 PPO裁剪，并为每条去噪轨迹使用一个似然比。

集中式评论家 $V_\psi(s_t)$ 根据全局状态估计团队回报。GAE [26] 为每次环境转移给出一个优势$\widetilde{}$。我们在整个轨迹采样上对其进行归一化，并将相同的团队优势 A_t 分配给每个处于活动状态的智能体。评论家目标为 $R_t \widehat{=} A_t + \widehat{V_{\psi_{\text{old}}}}(s_t)$，我们使用标准平方误差目标将 V_ψ 拟合到该固定目标。

每个智能体的更新都使用其完整存储的去噪轨迹：

$$\rho_t^i = \frac{q_{\theta,\phi}(\xi_t^i \mid h_t^i)}{q_{\text{old}}(\xi_t^i \mid h_t^i)}, \qquad Z = \max\Big(1, \sum_{t,i} m_t^i\Big), \tag{7}$$

其中 q_{old} 是滚动采样策略， $m_t^i \in \{0,1\}$ 表示具有动作选择的智能体。对数比率对所有去噪步数和动作坐标上的转移对数概率求和。固定初始概率 p_0会相互抵消。

裁剪所有智能体比率的乘积，会使裁剪取决于它们的综合变化。我们改为分别裁剪每个智能体的完整轨迹比率，同时保留共享团队优势，并对轨迹采样中的活跃样本取平均：

$$L_{\text{clip}} = \frac{1}{Z} \sum_{t,i} m_t^i \min\Big(\rho_t^i \widetilde{A}_t,\ \mathrm{clip}(\rho_t^i, 1-\varepsilon, 1+\varepsilon)\widetilde{A}_t\Big), \tag{8}$$

其中 ε 是裁剪阈值。在团队优势精确时，两个代理目标在采集策略处具有相同的一阶方向，至多相差一个共同缩放因子 $1/Z$。附录 A 给出了噪声参数化、与动作边缘分布的形式化联系，以及对此更新的理论分析。

我们最小化 $-L_{\text{clip}}$ 同时固定已存储的去噪轨迹。梯度更新 θ 和 ϕ通过条件均值和标准差。附录B 给出了 GAE 递归、边界处理和实现细节。

表 2 展示了 MPE 和 OG-MARL MuJoCo（连续控制）以及 SMAC（离散控制）上的结果，其中方法为行，任务—数据集对为列。本地复现结果标记为 r ，与已发布的 MAC-Flow 配置不具有直接可比性。OMIGA MuJoCo 和 SMACv2 的结果见表 7。阴影标示每列中的最佳值和次佳值。来源和符号表示见附录 C.1。

MPE	**Spread**				**Tag**			**世界**		
	专家	中等	M-R[b]	随机	专家	中等	随机	专家	中等	随机
MADiff-D[a] [34](2024年)	95.0 ±5.3	64.9 ±7.7	30.3 ±2.5	6.9 ±3.1	120.9 ±14.6米	77.2 ±10.4米	3.2 ±4.0米	122.6 ±14.4米	~~123.5 ±4.5~~米	2.0 ±3.0米
DOM2 [12] （2026年）	**131.32 ±4.82**	55.77 ±7.06	46.02 ±10.81	**49.86 ±7.28**	138.75 ±12.02	84.29 ±25.36	112.18 ±30.21	123.17 ±19.81	105.79 ±27.11	54.23 ±16.57
MAC-Flow [9](2026年)	101.7 ±10.9	**80.1 ±20.6**	**50.4 ±33.2**	31.1 ±6.8	119.97r ±2.26	80.47r ±1.77	72.68r ±3.20	122.57r ±2.91	89.82r ±2.25	94.43r ±3.72
TeamFlowRL（不使用在线强化学习）	72.87 ±4.85	27.13 ±0.93	7.89 ±4.79	−1.79 ±3.45	96.60 ±7.31	58.62 ±6.47	−0.40 ±0.99	109.23 ±13.01	61.69 ±6.60	0.18 ±1.60
TeamFlowRL	110.75† ±3.76	71.13† ±8.31	31.32† ±7.54	24.93† ±2.31	**150.13† ±5.83**	**145.13† ±7.88**	**156.74† ±4.45**	**176.11† ±16.62**	**155.60† ±4.71**	**169.15† ±16.77**

OG-MARL	**2ant**			**2只半猎豹**			**4ant**		
	Good	中等	Poor	Good	中等	Poor	Good	中等	Poor
MADiff-D[a] [34](2024年)	2946 ±77	1211 ±69	946 ±66	8246 ±342	2207 ±23	759 ±18	3080 ±38	1649 ±100	1295 ±57
OMSD [21] (2026年)	2714 ±555	1372 ±107	1213 ±212	8619 ±418	2660 ±125	866 ±78	2844 ±152	1942 ±293	1477 ±192
MAC-Flowr [9](2026年)	1628.46r ±1.72	1077.94r ±42.96	912.21r ±23.51	8569.00r ±42.59	2477.22r ±4.11	937.09r ±8.23	1750.85r ±29.29	1254.43r ±35.45	1010.19r ±39.19
TeamFlowRL（不使用在线强化学习）	527.15 ±134.37	956.05 ±50.61	535.21 ±6.77	6204.54 ±43.82	1616.70 ±33.86	544.34 ±98.30	1554.84 ±41.54	1263.13 ±42.48	515.18 ±73.05
TeamFlowRL	**3453.45 ±224.71**	**3000.60 ±53.38**	**3283.65 ±95.53**	**10559.68 ±167.82**	**4568.23 ±93.63**	**1414.51 ±13.46**	**3769.81 ±36.64**	**3124.54 ±7.83**	**2957.69 ±74.82**

SMACv1	**3m**			**8m**			**2s3z**			**5m6m**		
	Good	中等	Poor	Good	中等	Poor	Good	中等	Poor	Good	中等	Poor
扩散行为克隆 [2] (2023年)	19.5 ±0.5	13.3 ±0.7	4.2 ±0.2	19.4 ±0.5	18.6 ±0.6	4.8 ±0.2	18.0 ±1.0	13.4 ±1.4	6.2 ±1.2	16.8 ±2.3	12.5 ±2.1	8.0 ±1.0
流行为克隆 [13] (2023年)	**20.0 ±0.0**	14.7 ±1.5	4.5 ±0.1	19.5 ±0.2	18.2 ±0.8	4.9 ±0.1	19.5 ±0.1	15.1 ±2.0	6.9 ±0.8	14.7 ±2.1	12.8 ±0.8	7.7 ±0.8
MADiff-D[a] [34](2024年)	19.3 ±0.5	16.4 ±2.6	10.3 ±6.1	18.9 ±1.1	16.8 ±1.6	9.8 ±0.9	15.9 ±1.2	15.6 ±0.3	8.5 ±1.3	16.5 ±2.8	15.2 ±2.6	8.9 ±1.3
DoF [11] (2025年)	19.8 ±0.2	18.6 ±1.2	10.9 ±1.1	19.6 ±0.3	18.6 ±0.8	12.0 ±1.2	18.5 ±0.8	18.1 ±0.9	10.0 ±1.1	17.7 ±1.1	16.2 ±0.9	10.8 ±0.3
MAC-Flow [9] (2026年)	19.8 ±0.2	18.0 ±3.2	10.6 ±2.2	19.7 ±0.3	**19.4 ±0.6**	11.5 ±0.8	19.5 ±0.5	17.6 ±0.6	8.5 ±0.6	**18.6 ±3.5**	15.6 ±1.3	9.8 ±2.1
TeamFlowRL（无在线强化学习）	15.85 ±1.28	10.12 ±1.11	4.26 ±0.13	17.87 ±0.44	11.19 ±1.26	5.20 ±0.11	18.28 ±0.54	12.11 ±0.54	7.12 ±0.35	15.50 ±0.38	11.57 ±0.74	8.01 ±0.47
TeamFlowRL	19.80 ±0.35	**19.81 ±0.34**	**18.63 ±0.89**	**20.00 ±0.00**	18.89 ±0.64	**18.40 ±0.34**	**19.84 ±0.15**	**20.00 ±0.00**	**16.46 ±0.67**	17.43 ±0.77	**19.02 ±0.94**	**12.85 ±0.45**

[a] *MADiff-D* 是去中心化 *MADiff* 变体，与分散式执行相匹配。 [b]*Medium-Replay*（*M-R*）表示数据集质量。所列基线方法在三个方块之间有所不同，因为每种方法仅支持或报告部分连续和离散基准测试。

部署。 每个智能体采样一个新的高斯潜变量，并应用 K 公式 3 中的均值更新，仅使用 h_t^i。训练使用转移噪声，而部署会移除该噪声，既不需要噪声网络，也不需要集中式评论家。初始潜变量使执行策略保持随机性。

4 实验

我们在标准多智能体基准测试和协作式双臂操作任务上评估 TeamFlowRL。实验衡量了在不同任务和数据质量下相较于离线初始化的改进，考察了微调策略的动作分布和协作行为，将 TeamFlowRL 与高斯策略进行比较，并测试了轨迹比率的主要实现选择。

4.1 实验设置

基准测试。 我们评估了五组多智能体基准测试：多智能体粒子环境（MPE）中的协作导航和追捕、OG-MARL MuJoCo [5], OMIGAMuJoCo [31], SMAC [25],和 SMACv2 [3]。它们包含 45 种任务与数据集组合。我们还使用 ManiSkill 中的 TwoRobotPickCube-v1 和 TwoRobotStackCube-v1 评估模拟双臂操作 3 [29] ， 以及 RoboFactory 中的 TwoRobotsStackCube [22], ； 每项任务均配备两条 Franka Panda 机械臂，并将每条机械臂视为一个智能体。我们同时使用状态和相机观测，并纳入两种堆叠

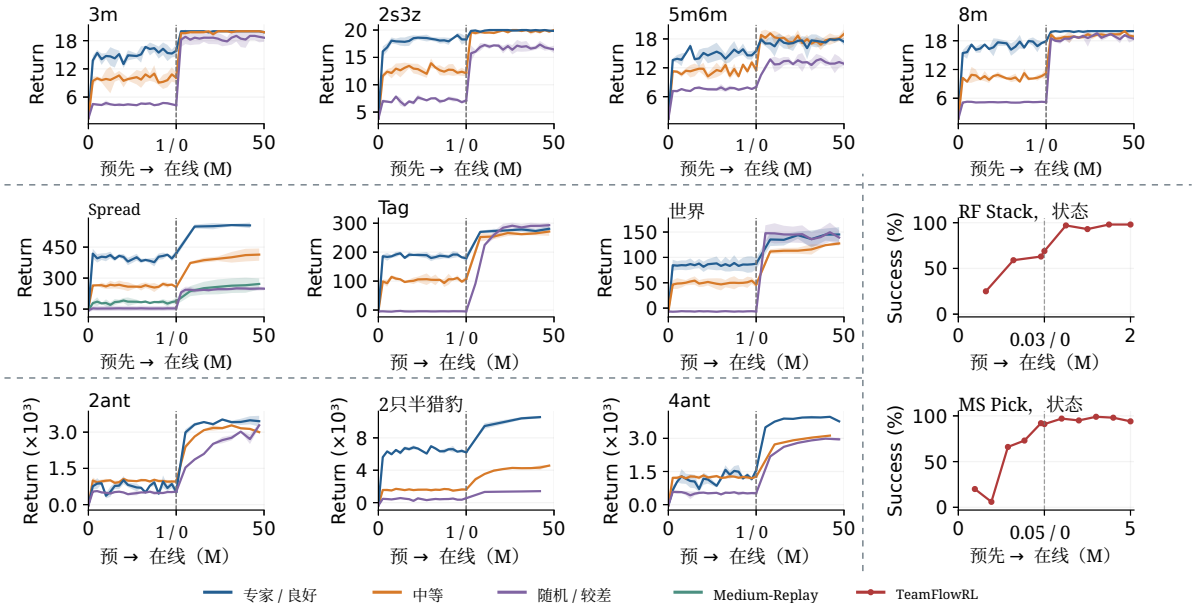

图3 SMAC、MPE、OG-MARL MuJoCo 和机器人任务上的学习曲线。横轴显示预训练更新次数，随后是以百万计的在线环境步数，刻度 1 / 0 标志着在线训练的开始。颜色表示离线数据质量。机器人面板展示了 TeamFlowRL 在部署规则下的成功率。附录 C中提供了其他曲线和报告细节。

任务，其中两条机械臂必须在结束时释放方块并移开。完整的任务设置见附录 C，双臂配置和预训练数据见附录 D.1。

基线方法。 我们将 TeamFlowRL 与基于扩散模型和流模型的代表性方法进行比较，包括 DOM2、MADiff、DoF、扩散行为克隆、流行为克隆和 MAC-Flow。由于并非所有方法都支持每个基准测试，不同方法在 MPE、OG-MARL 和 SMAC 上的覆盖范围有所不同。对于双手操作，我们还在匹配的任务定义、观测、奖励和在线交互预算下，评估了 MAPPO 和 BC+MAPPO。已发表的 RoboFactory 结果采用了不同的视觉运动模仿学习协议，因此无法直接进行比较。

评估协议。 所有基准测试结果均采用统一的 5000 万环境步数在线预算。对于每项任务，机器人方法采用相匹配的在线交互预算。机器人主要比较采用随机执行，并保留训练期间使用的探索噪声。所有机器人结果均取三个独立训练随机种子的平均值，阴影区域表示评估回合上的 95% 置信区间。完整的报告约定、评估详情，以及各方法部署规则下的结果见附录 C 和 D。

4.2 主要结果

相较于离线初始化的提升。 在线微调在全部 51 种配置中均提升了离线流策略。在全部 45 种基准测试配置中，最终平均回报平均比离线初始化高出任务最佳在线回报的 43.2%。表2 报告了主要基准测试结果。在全部六项机器人任务中，如表 3所示，随机执行下的最终成功率平均比预训练初始化高出 61.2%。我们在图 1 中将基准测试回报按最佳在线回报进行归一化，并以百分比表示探索噪声下的机器人成功率。七个组别和全部 51 种配置的归一化得分平均提升了 47.3%。

离线数据质量的影响。 图 4显示，各种数据质量下都能获得增益，而且随着离线数据质量的提升，增益呈线性缩小。在 15 种 SMAC 和 MPE 配置中，数据集均值每提高 1%

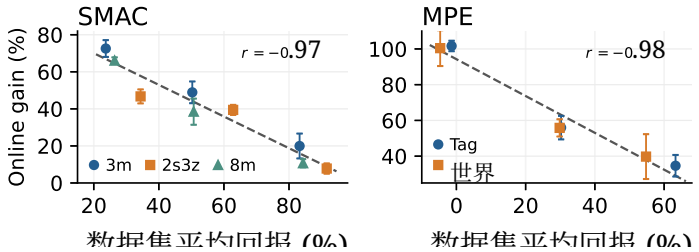

图 4 在线增益与离线数据集平均回报，二者均以任务最佳在线回报的百分比表示。虚线为线性拟合，误差条表示一个标准差。

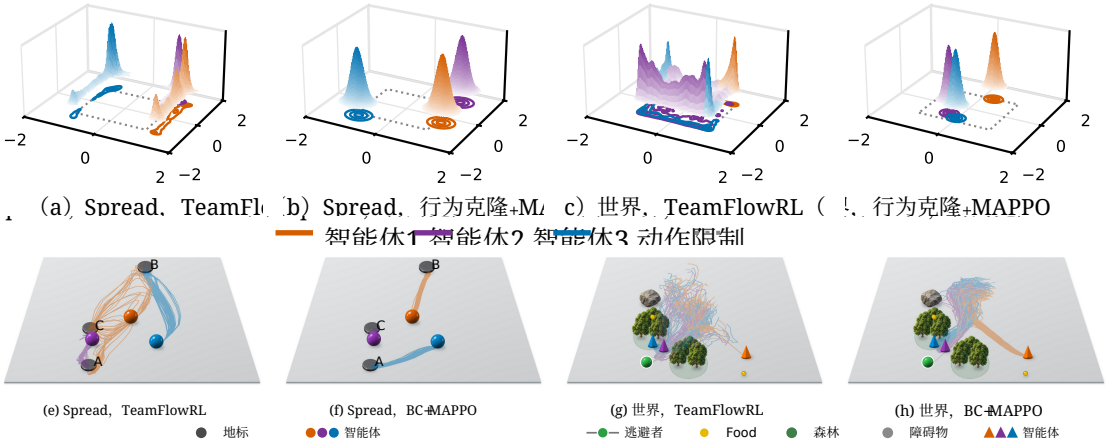

图5 TeamFlowRL 与行为克隆的动作分布和 MPE 轨迹+MAPPO。（a）--（d）展示了一个 Spread 场景和一个 World 场景中，三个智能体在第一步的动作分布，颜色用于区分智能体。（e）--（h）展示了每种方法从同一初始场景进行的 64 次 MPE rollout 的路径，智能体颜色保持一致。附录 C.3中提供了更多诊断细节。

回报，在线增益便降低 0.93%，皮尔逊相关系数为 $r = -0.98$。在 3m、2s3z、8m、Tag 和 World 中，由优质或专家数据微调得到的策略，其平均回报超过了相应数据集的均值。在 3m、8m、Tag 和 World 中，由较差或随机数据微调得到的策略，其最终回报与最佳最终回报的差距不超过 10%。更多学习曲线和机器人回合见附录 C 和 D。

4.3 策略表达能力

动作分布。 图 5（a）--（d）比较了根据相同局部输入生成的动作。在所展示的场景中，微调后的流策略产生了相互分离的动作模态，而行为克隆+MAPPO 用单个高斯分布表示每个智能体的动作。因此，微调后的流策略能够表示高斯策略无法表示的动作分布。

协作行为。 在 Spread 中，三个智能体必须覆盖三个地标。我们将地标标记为 A、B 和 C，并将分配方案写成智能体1、智能体2 和智能体3 最终分别到达的地标，因此 CAB 表示智能体1 前往 C，智能体2 前往 A，智能体3 前往 B。当每个地标都有一个距离其 0.1 以内的智能体时，一次轨迹采样即覆盖所有地标。在一个固定初始场景中进行的 64 次随机滚动执行里，TeamFlowRL 通过两种不同的分配方案覆盖了所有地标。如图 5(e)--(f)所示，它在 35 次轨迹采样中使用 CAB，在 28 次中使用 ACB。剩余的一次轨迹采样有一个地标未被覆盖。我们将其与行为克隆+MAPPO 进行比较；后者在相同数据上预先训练高斯策略，并使用 MAPPO 对其进行微调。行为克隆+MAPPO 在 63 次轨迹采样中覆盖了这些地标，但始终使用唯一的分配方案 BCA。在 World 中，图 5(g)--(h)显示，从固定初始场景出发，TeamFlowRL 遵循的追捕路径也比行为克隆+MAPPO 更多样。因此，微调后的流策略既能提供相当的任务覆盖率，也能带来更高的行为多样性

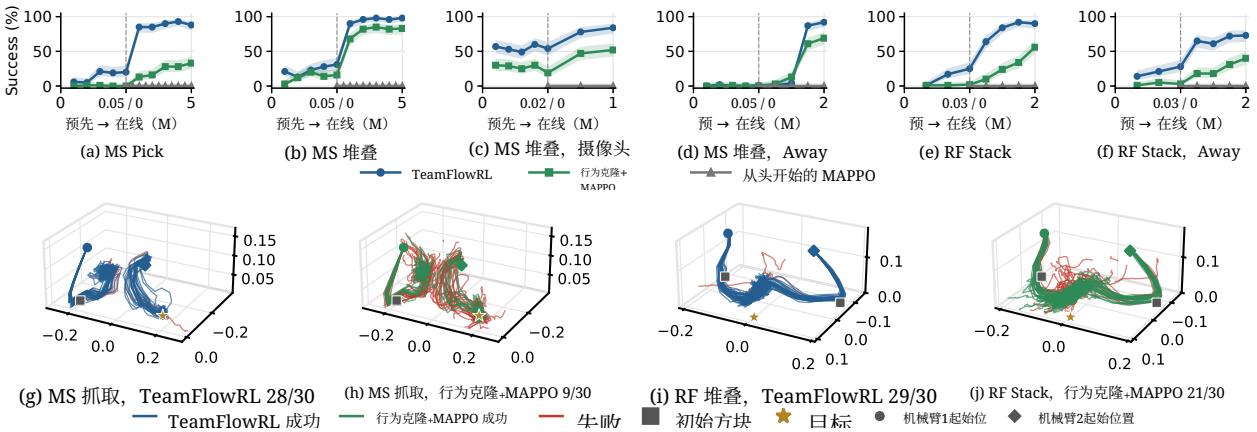

图 6 多机械臂操作中的训练。(a)--(f) 从预训练到在线微调阶段，在探索噪声下的成功率。竖线标记在线训练的开始位置，色带表示评估回合上的 95% 区间。(g)--(j) 在带有探索噪声的一个初始场景中，两个机械臂的工具中心点路径，其中红线表示失败。

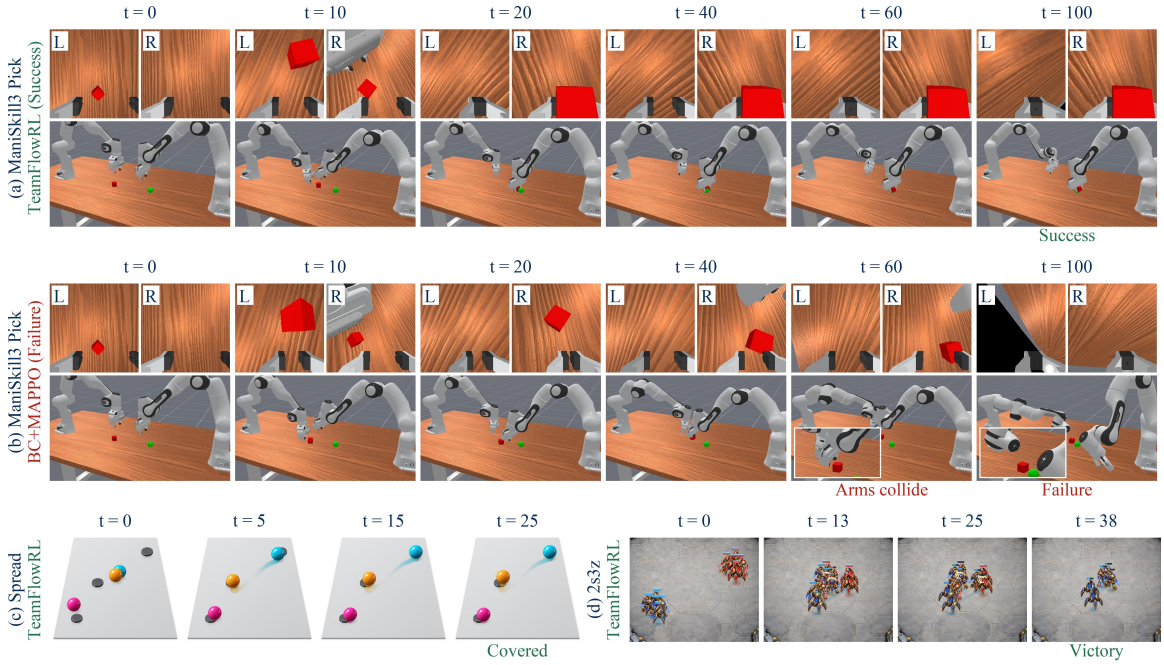

图 7 记录的回合。(a)--(b) 在探索噪声下，TeamFlowRL 和 BC+MAPPO 来自同一个 ManiSkill3 Pick场景，每帧上方显示左侧（L）和右侧（R）腕部视角；(b) 中的插图放大显示了机械臂碰撞以及遗留在桌面上的立方体。(c)--(d) TeamFlowRL 在 Spread 和 SMAC 2s3z 中的表现。

，优于单峰高斯基线。附录 C 和D中提供了更多行为和机器人轨迹诊断结果。

4.4 与高斯策略的比较

我们将 TeamFlowRL 与 BC+MAPPO 进行比较：后者在相同数据上预训练一步高斯策略，并使用 MAPPO 对其进行微调；同时还与从头开始训练的 MAPPO 进行比较。

表 3 机器人在随机执行下的成功率；粗体标示最终结果中的最佳值。

Task	MAPPO	行为克隆+MAPPO		TeamFlowRL	
	已训练	预训练	微调	预训练	微调
MS Pick	0.0%	0%	34.4%	20%	**91.6%**
MS 堆叠	0.0%	16%	82.6%	31%	**95.8%**
MS 堆叠、摄像头	0.0%	19%	52.0%	54%	**84.0%**
MS 堆叠，Away	0.0%	0%	69.0%	1%	**91.6%**
RF Stack	0.0%	2%	49.8%	25%	**90.4%**
RF Stack，Away	0.0%	3%	40.0%	28%	**73.0%**
平均值	0.0%	6.7%	54.6%	26.5%	**87.7%**

MS/RF：ManiSkill3/RoboFactory；cam.：相机；away：手部远离。

在线适应。 图 6（a）--（f）展示了六个机器人任务的训练曲线。经过 100 万步在线训练后，带有探索噪声的 TeamFlowRL 在六个任务上的平均成功率为 68.0%，而 BC+MAPPO 为 31.3%。表 3显示，从头开始训练的 MAPPO 在训练结束时未能成功完成六个任务中的任何一个。这一比较表明，与高斯微调和从头开始训练 MAPPO 相比，离线流预训练能够显著提升在线适应能力。

探索噪声下的性能。 在线训练会使用探索噪声执行每个策略，因此我们评估这些执行的成功率。在训练结束时，TeamFlowRL 在全部六项任务中的成功率均高于行为克隆+MAPPO，平均高出 33.1%，如表 3所示。对于每项任务中一个固定的初始状态，TeamFlowRL 的执行成功率为 95%，而行为克隆+MAPPO 为 50%，如图 6（g）--（j）所示。图 7（a）--（b）展示了其中一个回合：TeamFlowRL 将立方体举到目标位置，而行为克隆+MAPPO 将其留在桌上。总体而言，当在在线训练期间使用的相同随机执行规则下评估策略时，TeamFlowRL 仍保持其优势。这些主要结果评估的是随机训练策略。各方法部署规则下的其他结果见附录 D。

4.5 消融研究

我们在三种具有代表性的配置上，将完整去噪轨迹比率与逐步裁剪、比率平均与钳制进行了比较。在三种配置中的两种配置上，轨迹比率取得了最佳最终性能，并且从未明显劣于其他方法。这一结果支持将完整轨迹比率作为默认更新规则，同时详细曲线和配置结果见附录C.2。

5 结论

我们提出了 TeamFlowRL，这是一种两阶段框架，先对多智能体流策略进行离线预训练，再通过在线交互对其进行微调。TeamFlowRL 在去中心化去噪轨迹上进行团队策略优化，使用共享团队优势，并分别对每个智能体的完整轨迹比率进行裁剪，使局部更新与共同的团队回报保持一致。在七个仿真组的 51 种配置中，在线微调提升了每一种离线初始化的性能，归一化得分平均提升 47.3%。微调后的流策略还能够表示多模态动作分布和多样化的协作行为；在六项双手任务中，与高斯策略微调相比，其在探索噪声下的平均成功率提高了 33.1%。

局限性。 我们的操作评估仅限于模拟环境中的双臂立方体操作。只有两个任务使用了相机观测，且我们未在实体机器人上进行评估。我们使用紧凑型、特定于任务的策略网络，尚未使用大型视觉-语言-动作模型评估 TeamFlowRL。最后，不同任务中离线数据质量对最终在线性能的影响各不相同。如图 3所示，我们尚未弄清为何在某些任务中，较低质量的初始化方案能够赶上较高质量的初始化方案，而在其他任务中却不能。

大语言模型使用声明

在本研究中，我们使用了大语言模型助手（Claude 和 Codex）。它们用于润色英文写作，并协助实现该方法和绘图代码。作者设计了该方法和实验，验证了所有代码、结果和证明，并对论文内容承担全部责任。

参考文献

[1] Kang Chen、Zhihao Liu、Tonghe Zhang、Zhen Guo、Si Xu、Hao Lin、Hongzhi Zang、Xiang Li、Quanlu Zhang、Zhaofei Yu、Guoliang Fan、Tiejun Huang、Yu Wang 和 Chao Yu。 π_{RL}：面向基于流的视觉-语言-动作模型的在线强化学习微调。arXiv 预印本 arXiv:2510.25889，2025年。URL https://arxiv.org/abs/2510.25889。[2] Cheng Chi、Siyuan Feng、Yilun Du、Zhenjia Xu、Eric Cousineau、Benjamin Burchfiel 和 Shuran Song。扩散策略：通过动作扩散进行视觉运动策略学习。载于《机器人学：科学与系统》（RSS），2023年。doi：10.15607/RSS.2023.XIX.026。[3] Benjamin Ellis、Jonathan Cook、Skander Moalla、Mikayel Samvelyan、Mingfei Sun、Anuj Mahajan、Jakob N. Foerster 和 Shimon Whiteson。SMACv2：面向协作式多智能体强化学习的改进基准测试。载于《神经信息处理系统进展》（NeurIPS），第36卷，37593，2023年 37567 10 52202 075280 1634 页 -- 。doi：。/ - 。[4] Jakob N. Foerster、Gregory Farquhar、Triantafyllos Afouras、Nantas Nardelli 和 Shimon Whiteson。反事实多智能体策略梯度。载于 AAAI人工智能大会（AAAI），第32卷，第2974--2982页，2018年。doi：10.1609/aaai.v32i1.11794。[5] Claude Formanek、Asad Jeewa、Jonathan Shock 和 Arnu Pretorius。脱离网格的多智能体强化学习：离线多智能体强化学习的数据集与基线方法。载于国际自主智能体与多智能体系统会议（AAMAS），第2442--2444页，2023年。doi：10.65109/vmil8420。[6] Scott Fujimoto 和 Shixiang Shane Gu。离线强化学习的极简方法。载于《神经信息处理系统进展》（NeurIPS），第20132--20145页，2021年。URL https://arxiv.org/abs/2106. 06860。[7] Scott Fujimoto、David Meger 和 Doina Precup。无需探索的离策略深度强化学习。载于国际机器学习会议（ICML），《机器学习研究论文集》（PMLR）第97卷，第2052--2062页，2019年。[8] Aviral Kumar、Aurick Zhou、George Tucker 和 Sergey Levine。用于离线强化学习的保守Q学习。载于《神经信息处理系统进展》（NeurIPS），第33卷，2020年。URL https://proceedings.neurips.cc/paper/2020/hash/0d2b2061826a5df3221116a5085a6052-Abstract.html。[9] Dongsu Lee、Daehee Lee 和 Amy Zhang。通过流匹配实现多智能体协调。载于国际学习表征会议（ICLR），2026年。URL https://arxiv.org/abs/2511.05005。[10] Sergey Levine、Aviral Kumar、George Tucker 和 Justin Fu。离线强化学习：教程、综述与开放问题展望。arXiv 预印本 arXiv:2005.01643，2020年。URL https://arxiv.org/abs/2005.01643。[11] Chao Li、Ziwei Deng、Chenxing Lin、Wenqi Chen、Yongquan Fu、Weiquan Liu、Chenglu Wen、Cheng Wang 和 Siqi Shen。DoF：用于离线多智能体强化学习的扩散分解框架。载于国际学习表征会议（ICLR），2025年。[12] Zhuoran Li、Ling Pan、Jiatai Huang 和 Longbo Huang。利用扩散提升离线多智能体强化学习的泛化能力与数据效率。《机器学习研究汇刊》（TMLR），2026年。URL https: //openreview.net/forum?id=GKuCKSJKvl。[13] Yaron Lipman、Ricky T. Q. Chen、Heli Ben-Hamu、Maximilian Nickel 和 Matthew Le。用于生成建模的流匹配。载于国际学习表征会议（ICLR），2023年。URL https: //openreview.net/forum?id=PqvMRDCJT9t。

[14] Ryan Lowe、Yi Wu、Aviv Tamar、Jean Harb、Pieter Abbeel 和 Igor Mordatch。用于混合合作-竞争环境的多智能体演员-评论家。载于《神经信息处理系统进展》（NeurIPS），第30卷，第6379--6390页，2017年。[15]David McAllister、Songwei Ge、Brent Yi、Chung Min Kim、Ethan Weber、Hongsuk Choi、Haiwen Feng 和 Angjoo Kanazawa。流匹配策略梯度。载于国际学习表征会议（ICLR），2026年。URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/ 3d43cc5692bf68944ee7cd31b97d0c11-Abstract-Conference.html。[16] Maxime Oquab、Timothée Darcet、Théo Moutakanni、Huy Vo、Marc Szafraniec、Vasil Khalidov、Pierre Fernandez、Daniel Haziza、Francisco Massa、Alaaeldin El-Nouby 等。DINOv2：无需监督地学习鲁棒视觉特征。《机器学习研究汇刊》（TMLR），2024年。[17] Ling Pan、Longbo Huang、Tengyu Ma 和 Huazhe Xu。保守性下制定更优计划：通过演员校正实现离线多智能体强化学习。载于国际机器学习会议（ICML），《机器学习研究论文集》（PMLR）第162卷，第17221--17237页，2022年。URL https://proceedings.mlr.press/v162/pan22a.html。[18] Bei Peng、Tabish Rashid、Christian A. Schroeder de Witt、Pierre-Alexandre Kamienny、Philip H. S. Torr、Wendelin Böhmer 和 Shimon Whiteson。FACMAC：因子化多智能体集中式策略梯度。载于《神经信息处理系统进展》（NeurIPS），第12208--12221页，2021年。URL https: //proceedings.neurips.cc/paper/2021/hash/65b9eea6e1cc6bb9f0cd2a47751a186f-Abstract.html。[19] Physical Intelligence、Ali Amin、Raichelle Aniceto、Ashwin Balakrishna、Kevin Black、Ken Conley、Grace Con- nors、James Darpinian、Karan Dhabalia、Jared DiCarlo、Danny Driess、Michael Equi、Adnan Esmail、Yunhao Fang、Chelsea Finn、Catherine Glossop、Thomas Godden、Ivan Goryachev、Lachy Groom、Hunter Hancock、Karol Hausman、Gashon Hussein、Brian Ichter、Szymon Jakubczak、Rowan Jen、Tim Jones、Ben Katz、Liyiming Ke、Chandra Kuchi、Marinda Lamb、Devin LeBlanc、Sergey Levine、Adrian Li-Bell、Yao Lu、Vishnu Mano、Mohith Mothukuri、Suraj Nair、Karl Pertsch、Allen Z. Ren、Charvi Sharma、Lucy Xiaoyang Shi、Laura Smith、Jost Tobias Springenberg、Kyle Stachowicz、Will Stoeckle、Alex Swerdlow、James Tanner、Marcel Torne、Quan Vuong、Anna Walling、Haohuan Wang、Blake Williams、Sukwon Yoo、Lili Yu、Ury Zhilinsky 和 Zhiyuan Zhou。 $\pi^*_{0.6}$：从经验中学习的视觉-语言-动作模型。载于《机器人学：科学与系统》（RSS），2026年。doi：10.15607/RSS.2026.XXII.087。[20] Dean A. Pomerleau。用于自主导航的人工神经网络高效训练。《神经计算》，3(1)：88--97，1991年。[21] Dan Qiao、Wenhao Li、Shanchao Yang、Hongyuan Zha 和 Baoxiang Wang。通过序列分数分解实现离线多智能体强化学习。载于国际机器学习会议（ICML），《机器学习研究论文集》（PMLR）第306卷，2026年。URL https://arxiv.org/abs/2505.05968。[22] Yiran Qin、Li Kang、Xiufeng Song、Zhenfei Yin、Xiaohong Liu、Xihui Liu、Ruimao Zhang 和 Lei Bai。Robo-Factory：在组合约束下探索具身智能体协作。载于 IEEE/CVF国际计算机视觉会议（ICCV），第10075--10085页，2025年。doi：10.1109/ICCV51701.2025.00939。[23] Tabish Rashid、Mikayel Samvelyan、Christian Schroeder de Witt、Gregory Farquhar、Jakob Foerster 和 Shimon Whiteson。QMIX：用于深度多智能体强化学习的单调价值函数分解。载于国际机器学习会议（ICML），第4292--4301页，2018年。URL https://proceedings.mlr.press/v80/rashid18a.html。[24] Allen Z. Ren、Justin Lidard、Lars L. Ankile、Anthony Simeonov、Pulkit Agrawal、Anirudha Majumdar、Benjamin Burchfiel、Hongkai Dai 和 Max Simchowitz。扩散策略策略优化。载于国际学习表征会议（ICLR），2025年。[25] Mikayel Samvelyan、Tabish Rashid、Christian Schroeder de Witt、Gregory Farquhar、Nantas Nardelli、Tim G. J. Rudner、Chia-Man Hung、Philip H. S. Torr、Jakob Foerster 和 Shimon Whiteson。星际争霸多智能体挑战赛。载于国际自主智能体与多智能体系统会议（AAMAS），第2186-- 2188页，2019年。doi：10.65109/lvzz5205。[26] John Schulman、Philipp Moritz、Sergey Levine、Michael Jordan 和 Pieter Abbeel。利用广义优势估计进行高维连续控制。载于国际学习表征会议（ICLR），2016年。

[27] John Schulman、Filip Wolski、Prafulla Dhariwal、Alec Radford 和 Oleg Klimov。近端策略优化算法。arXiv 预印本 arXiv:1707.06347，2017年。网址：https://arxiv.org/abs/1707.06347。[28] Peter Sunehag、Guy Lever、Audrunas Gruslys、Wojciech Marian Czarnecki、Vinicius Zambaldi、Max Jaderberg、Marc Lanctot、Nicolas Sonnerat、Joel Z. Leibo、Karl Tuyls 和 Thore Graepel。基于团队奖励的协作式多智能体学习价值分解网络。载于国际自主智能体与多智能体系统会议（AAMAS），第 2085--2087 页，2018年。doi：10.65109/jsrc7365。[29] Stone Tao、Fanbo Xiang、Arth Shukla、Yuzhe Qin、Xander Hinrichsen、Xiaodi Yuan、Chen Bao、Xinsong Lin、Yulin Liu、Tse-kai Chan 等。ManiSkill3：面向可泛化具身人工智能的图形处理器并行机器人仿真与渲染。arXiv 预印本 arXiv:2410.00425，2024年。网址：https://arxiv.org/abs/2410.00425。[30] Jianhao Wang、Zhizhou Ren、Terry Liu、Yang Yu 和 Chongjie Zhang。QPLEX：双路对决式多智能体Q学习。载于国际学习表征会议（ICLR），2021年。[31] Xiangsen Wang、Haoran Xu、Yinan Zheng 和 Xianyuan Zhan。通过隐式全局到局部价值正则化进行离线多智能体强化学习。载于神经信息处理系统进展（NeurIPS），第 36 卷，第 52413--52429 页，2023年。doi：10.52202/075280-2282。[32] Chao Yu、Akash Velu、Eugene Vinitsky、Jiaxuan Gao、Yu Wang、Alexandre Bayen 和 Yi Wu。PPO 在协作式多智能体博弈中的惊人有效性。载于神经信息处理系统进展（NeurIPS），第 35 卷，第 24611--24624 页，2022年。doi：10.52202/068431-1787。[33] Tonghe Zhang、Chao Yu、Sichang Su 和 Yu Wang。Reinflow：使用在线强化学习微调流匹配策略。载于神经信息处理系统进展（NeurIPS），第 38 卷，第 117701--117738 页，2025年。doi：10.52202/085713-3547。[34] Zhengbang Zhu、Minghuan Liu、Liyuan Mao、Bingyi Kang、Minkai Xu、Yong Yu、Stefano Ermon 和 Weinan Zhang。MADiff：基于扩散模型的离线多智能体学习。载于神经信息处理系统进展（NeurIPS），2024年。[35] Guowei Zou、Weibing Li、Hejun Wu、Yukun Qian、Yuhang Wang 和 Haitao Wang。D^2PPO：采用色散损失的扩散策略策略优化。载于 AAAI人工智能大会（AAAI），第 40 卷，第 18891--18899 页，2026 年。doi：10.1609/aaai.v40i22.38959。网址：https://ojs.aaai.org/index.php/AAAI/article/ download/38959。[36] Guowei Zou、Haitao Wang、Hejun Wu、Yukun Qian、Yuhang Wang 和 Weibing Li。OGPO：面向实时机器人控制的单步生成式策略优化。载于 ACM国际多媒体会议（ACM MM），2026年。网址：https://arxiv.org/abs/2601.20701。

附录目录

方法细节与理论分析 **15**A.1 记号与索引约定 .15A.2 离线流目标与网络 .15A.3 去噪轨迹似然 .16A.4 理论分析 .17A.5 与高斯基线的比较 .19

B 实现细节 **19**
B.1 输入、动作映射与更新控制 .19B.2 在线微调设置.20B.3 有界修复 .21

C 基准实验 **21**
C.1 实验协议与报告约定 .21C.2 轨迹比消融实验 .23C.3 行为诊断详情 .23

D 机器人操作实验 **24**
D.1 机器人操作设置.24D.2 部署规则下的成功率 .24D.3 部署规则下的轨迹 .25D.4 探索噪声下的双手堆叠 .25D.5 基于相机观测的双手操作 .25D.6 学习释放与撤回 .27

A 方法细节与理论分析

A.1 符号表示与索引约定

环境时间 t、流时间 τ、去噪步骤 k和智能体索引 i 描述不同的操作。智能体索引为上标，环境时间为下标，例如 h_t^i。在一条去噪轨迹中会省略时间索引，因此 $x^{i,\,k}$ 表示智能体 i 在 k 的 $K=10$ 个去噪步骤之后的变量。去噪步骤k 对应于流时间 $\tau = k/K$，此时会在公式 3中计算速度网络。下标 τ，例如 x_τ^i，始终表示流时间，而非环境时间。粗体符号汇总了所有 N 智能体。表 4列出了这些符号。

表 4 训练、部署与分析中使用的符号。

符号	含义
索引 ，时钟 ，以及局部信息	
i, N	智能体索引和受控智能体数量（$N = \lvert\mathcal{I}\rvert$）。
t, T	环境步和回合时域。
k, K	去噪步骤和流步数，其中 $K = 10$。
τ	潜在噪声与数据之间的连续插值时间。
$h_t^i,\ s_t$	局部执行器输入和集中式评论家输入。
$a^i,\ r_t$	执行的局部动作和共享团队奖励。
流初始化和去噪轨迹	
z, I	标准高斯潜变量和单位协方差矩阵。
x_τ	插值离线流输入。
v_θ, θ	条件速度场及其共享参数。
$\mathcal{L}_{\mathrm{FM}}$	条件流匹配目标。
$x^{i,k}, \xi^i$	中间去噪变量和完整存储轨迹（$x^{i,0}, \ldots, x^{i,K}$）。
μ_θ	条件转移均值。
$f_\phi, \sigma_\phi, \phi$	噪声网络、条件转移标准差及其参数。
$\sigma_{\min}, \sigma_{\max}$	条件标准差的上下界。
$\epsilon^{i,k}$	独立标准高斯转移噪声。
$p_0, q_{\theta,\phi}, M_{\theta,\phi}$	初始潜在概率、单智能体轨迹概率和因子化联合轨迹策略。
$g, \pi_{\theta,\phi}, U$	动作变换、诱导动作分布和可测动作集。
b	修正版中的转移均值上界。
团队优势与在线优化	
J, γ	期望折扣团队回报与折扣因子。
$V_\psi,\ \psi$	集中式评论家及其参数。
$u_t,\ d_t$	真实终止指示器与回合边界指示器。
$\delta_t,\ \lambda, \widehat{A}_t$	时间差分残差、GAE迹系数和团队优势估计。
$\ell_{\theta,\ \phi},\ \ell_{\mathrm{old}}$	当前策略和采集策略下的条件转移对数概率之和
策略 .	
ρ_t^i, ε	去噪轨迹的每个智能体的似然比和PPO裁剪阈值。
$m_t^i,\ L_{\mathrm{clip}}$	演员活动掩码和裁剪PPO目标。
$\mathbb{E},\ D_{\mathrm{KL}},$ 裁剪	期望、Kullback–Leibler散度和裁剪算子。
报告的分数	
$R,\ R_{\mathrm{random}},\ R_{\mathrm{expert}}$	MPE 归一化的原始回合回报和参考回报。

A.2 离线流目标和网络

对于数据集对 (h^i, a^i)、潜变量 $z^i \sim \mathcal{N}(0, I)$ 以及 $\tau \sim \mathcal{U}[0,1],$，线性插值 $x_\tau^i = (1-\tau)z^i + \tau a^i$ 具有条件速度

$$u(x_\tau^i \mid z^i, a^i) = \frac{dx_\tau^i}{d\tau} = a^i - z^i, \tag{9}$$

且条件流匹配将网络回归到该速度上，从而得到公式 2。对于真实边缘速度 $u^\star(x,\tau \mid h) = \mathbb{E}[a - z \mid x_\tau = x, h]$，回归目标满足

$$\begin{aligned}&\mathbb{E}\|v_\theta(h, x_\tau, \tau) - (a - z)\|^2 \\ &\quad = \mathbb{E}\|v_\theta(h, x_\tau, \tau) - u^\star(x_\tau, \tau \mid h)\|^2 + C,\end{aligned} \tag{10}$$

其中 C 不依赖于 θ。这是一个标准论证：条件损失与对 $u^\star$进行回归具有相同的极小值。离线教师通过 K 确定性欧拉步生成一个动作，

$$x^{i,k+1} = x^{i,k} + K^{-1} v_\theta(h^i, x^{i,k}, k/K), \tag{11}$$

对应 $k = 0$、.....、$K - 1$，它是方程 3 的转移均值。速度网络是一个多层感知机，在 h^i、$x^{i,k}$ 的拼接结果上使用 GELU 激活函数，并使用 k/K 或 k/K的正弦特征。网络宽度和时间特征从离线检查点导入。噪声网络 f_ϕ 是一个独立的感知机，包含一个宽度为 64 的隐藏层，并在 (h^i、$x^{i,k}$、k/K) 上使用 tanh 激活函数。其输出层的权重初始为零，偏置为常数，因此每个坐标的初始值都是配置的标准差：

$$\begin{aligned}\sigma_\phi(h^i, x^{i,k}, k) = \sigma_{\min} + (\sigma_{\max} - \sigma_{\min}) \\ \cdot \operatorname{sigmoid}\big(f_\phi(h^i, x^{i,k}, k/K)\big),\end{aligned} \tag{12}$$

其中 $\sigma_{\min} = 0.01$、$\sigma_{\max} = 0.5$，初始值为 0.1。

A.3 去噪轨迹似然

给定 $x^{i,k}$，公式 3 中的均值 $\mu_\theta^{i,k}$ 和标准差 $\sigma_\phi^{i,k}$ 是固定的，而 $x^{i,k+1} = \mu_\theta^{i,k} + \sigma_\phi^{i,k} \odot \epsilon^{i,k}$ 是 $\epsilon^{i,k} \sim \mathrm{N}(0, I)$ 的仿射函数。标准高斯向量的仿射函数服从高斯分布，因此 $x^{i,k+1} \mid x^{i,k} \sim \mathrm{N}\big(\mu_\theta^{i,k}, \operatorname{diag}((\sigma_\phi^{i,k})^2)\big)$。由于每一步仅通过 $x^{i,k}$ 依赖于更早的步数，公式 4 中的轨迹概率就是这些状态转移的乘积。每个状态转移在 d动作坐标上都是对角高斯分布，因此

$$\begin{aligned}&\log \mathcal{N}\big(x^{i,k+1} \mid \mu_\theta^{i,k}, \operatorname{diag}((\sigma_\phi^{i,k})^2)\big) \\ &\quad = -\sum_{c=1}^{d} \Big[\frac{(x_c^{i,k+1} - \mu_{\theta,c}^{i,k})^2}{2(\sigma_{\phi,c}^{i,k})^2} + \log \sigma_{\phi,c}^{i,k} + \tfrac{1}{2}\log 2\pi\Big],\end{aligned} \tag{13}$$

存储的对数似然是所有步数和坐标上的求和，

$$\ell_{\theta,\phi}(\xi^i \mid h^i) = \sum_{k=0}^{K-1} \log \mathcal{N}\big(x^{i,k+1} \mid \mu_\theta^{i,k}, \operatorname{diag}((\sigma_\phi^{i,k})^2)\big). \tag{14}$$

公式 7中的比率为

$$\rho_t^i = \exp\big(\ell_{\theta,\phi}(\xi_t^i \mid h_t^i) - \ell_{\text{old}}(\xi_t^i \mid h_t^i)\big).$$

存储的去噪轨迹保持不变，同时在其前缀处重新计算 μ 和 σ 。最终变量的边缘概率表示为

$$q_{\theta,\phi}^K(x^{i,K} \mid h^i) = \int q_{\theta,\phi}(\xi^i \mid h^i)\, dx^{i,0} \cdots dx^{i,K-1}, \tag{15}$$

该值无法直接计算。执行动作分布是其在 g 下的推前分布；对于连续控制，在裁剪边界处存在点质量。

命题 1 (轨迹比与边缘比)。 固定 h^i，并令 $q = q_{\theta,\phi}(\cdot \mid h^i)$ 和 q_{old} 为式 4 所示的两个去噪过程的联合概率，且其标准差均严格为正。那么，对于几乎处处的 $x^{i,K}$，

$$\mathbb{E}_{\xi^i \sim q_{\text{old}}}\left[\frac{q(\xi^i)}{q_{\text{old}}(\xi^i)} \,\middle|\, x^{i,K}\right] = \frac{q^K(x^{i,K})}{q_{\text{old}}^K(x^{i,K})}. \tag{16}$$

此外，如果可以将微分运算移入积分号内，那么

$$\nabla \log q^K(x^{i,K}) = \mathbb{E}_{\xi^i \sim q}\big[\nabla \log q(\xi^i) \,\big|\, x^{i,K}\big]. \tag{17}$$

证明。具有正方差的高斯转移使得 q 和 q_{old} 处处为正。记 $y = (x^{i,0}, \ldots, x^{i,K-1})$。在 q_{old} 下，给定 $x^{i,K}$ 时，y 的条件概率为 $q_{\text{old}}(\xi^i)/q^K_{\text{old}}(x^{i,K})$，因此方程16左侧等于

$$\begin{aligned}\int \frac{q_{\text{old}}(\xi^i)}{q^K_{\text{old}}(x^{i,K})} \frac{q(\xi^i)}{q_{\text{old}}(\xi^i)}\, dy &= \frac{1}{q^K_{\text{old}}(x^{i,K})} \int q(\xi^i)\, dy \\ &= \frac{q^K(x^{i,K})}{q^K_{\text{old}}(x^{i,K})}.\end{aligned}$$

对于方程17，

$$\nabla q^K(x^{i,K}) = \int \nabla q(\xi^i)\, dy = \int q(\xi^i)\, \nabla \log q(\xi^i)\, dy.$$

除以 $q^K(x^{i,K})$，并利用 q 下 y 的条件概率即可得到结论 . □

该命题解释了为什么轨迹比率可以作为难以处理的动作边缘分布比率的替代量：在给定最终变量的条件下，它是边缘比的无偏估计，其期望得分等于边缘得分。根据全方差定律，轨迹比率的无条件方差至少与边缘比率的方差相同，因为中间变量会增加条件随机性。

A.4 理论分析

该分析涉及带有转移噪声的联合采样策略，

$$M_{\theta,\phi}(\boldsymbol{\xi} \mid \chi) = \prod_i q_{\theta,\phi}(\xi^i \mid h^i),$$

其中，χ 是包含联合历史的马尔可夫信息状态，且每个 h^i 都是 χ的固定投影。环境仅能看到一个 $= (g(x^{1,K}), \ldots, g(x^{N,K}))$。假设 $0 < \gamma < 1$、奖励有界、初始分布和转移分布与参数独立、每个历史对应固定的合法掩码，并满足交换微分与期望所需的正则性条件。令 d_M 表示 χ的归一化折扣占用分布，令 $A^M(\chi, \boldsymbol{\xi})$表示精确的团队优势，该优势仅通过一个依赖于 $\boldsymbol{\xi}$。以下陈述涉及带有转移噪声的训练策略。

命题2（固定动作映射下的原始去噪轨迹）。 采样策略的期望回报梯度可以用轨迹得分 $\nabla \log q_{\theta,\phi}(\xi^i \mid h^i)$ 表示。对于在相同历史下定义于去噪轨迹上的两个策略 q、p，

$$D_{\text{KL}}(g_{\#}q^K \| g_{\#}p^K) \le D_{\text{KL}}(q^K \| p^K) \le D_{\text{KL}}(q \| p)$$

只要右侧有限。

证明。 对于一次回报泛函为$\mathcal{R}(g(x^{i,K}))$且与参数独立的决策，

$$\nabla \mathbb{E}_{\xi \sim q}\big[\mathcal{R}(g(x^{i,K}))\big] = \int \mathcal{R}(g(x^{i,K}))\, \nabla q(\xi)\, d\xi = \mathbb{E}_q\big[\mathcal{R}\, \nabla \log q(\xi)\big].$$

将其应用于有限轨迹并取折扣极限，即可得到得分形式。这两个不等式都源于相对熵链式法则。联合分布的 KL 散度等于边缘分布的 KL 散度加上条件分布的期望 KL 散度，后者非负。第一个不等式将其应用于 $(x^{i,K}, g(x^{i,K}))$，第二个则应用于 $(\xi^i, x^{i,K})$。

□

命题 3 (因子化联合变化)。 在固定的信息状态下，对于定义在去噪轨迹 $M = \prod_i q_i$ 和 $M' = \prod_i p_i$ 上的乘积策略，且各自的 *KL* 散度有限，

$$D_{\mathrm{KL}}(M\|M') = \sum_i D_{\mathrm{KL}}(q_i\|p_i), \qquad \prod_i \rho^i = \frac{M'}{M}.$$

证明。

$$D_{\mathrm{KL}}(M\|M') = \mathbb{E}_{\boldsymbol{\xi}\sim M}\Big[\sum_i \log \frac{q_i(\xi^i)}{p_i(\xi^i)}\Big] = \sum_i \mathbb{E}_{\xi^i\sim q_i}\Big[\log \frac{q_i}{p_i}\Big]$$

因为每一项仅取决于 ξ^i，其在 M 下的边缘分布为 q_i。比率恒等式本身就是该因式分解。

□

该因式分解利用了智能体之间相互独立的潜变量和转移噪声。共享参数不会影响这一点。对活跃智能体的平均 KL 散度有界为 δ，则在该历史下，联合 KL 散度至多为 $N\delta$。

命题 4 （共享参数下的团队策略梯度）。 在所述假设下，

$$\begin{aligned}\nabla J = \frac{1}{1-\gamma}\mathbb{E}_{\chi\sim d_M,\boldsymbol{\xi}\sim M}\Big[A^M(\chi,\boldsymbol{\xi}) \\ \cdot \sum_{i=1}^{N} \nabla \log q_{\theta,\phi}(\xi^i \mid h^i)\Big],\end{aligned} \tag{18}$$

其中，梯度是相对于 (θ,ϕ) 求取的。

证明。 有限轨迹的对数似然包含 $\sum_t \sum_i \log q_{\theta,\ \phi}(\xi_t^i \mid h_t^i)$。环境项不依赖于参数。利用得分恒等式，通过条件期望移除每个得分之前的奖励，将剩余回报替换为动作价值，并减去不依赖于动作的状态价值，即可在折扣极限下得到公式 18 。在共享参数的情况下，乘积法则会对所有智能体的得分求和。

□

因此，共同的团队优势与去中心化采样兼容。实现将A^M 替换为由学习得到的评论家计算的归一化广义优势估计，并对滚动采样样本进行均匀平均。

命题 5 （智能体比率的一阶一致性）。 固定采集参数，令 A^{old} 表示相应的优势，并在采集占用分布和策略下取期望，定义为

$$\begin{aligned}S_{\mathrm{agent}} &= \mathbb{E}\Big[A^{\mathrm{old}}\sum_i(\rho^i - 1)\Big], \\ S_{\mathrm{joint}} &= \mathbb{E}\Big[A^{\mathrm{old}}\Big(\prod_i \rho^i - 1\Big)\Big].\end{aligned} \tag{19}$$

在采集参数处，二者均为零，其梯度等于 $(1-\gamma)\,\nabla J$，并且将每个 $\rho^i A^{\mathrm{old}}$替换为其裁剪对应项不会改变梯度。

证明。 在采集参数处，每个比率都等于 1，因此

$$\nabla \prod_i \rho^i = \sum_i \nabla \rho^i = \sum_i \nabla \log q_{\theta,\phi}(\xi^i \mid h^i).$$

代入定义并利用命题 4即可得到梯度恒等式。裁剪区间的内部包含 1，因此裁剪项与未裁剪项在某个邻域内相等，并且具有相同的导数。

□

这种一致性是局部的：在一次迭代的第一次更新之后，比率的求和与乘积会有所不同。根据命题1，轨迹比率与动作边缘分布比率之间也具有相同的局部一致性。

A.5 与高斯基线的比较

基线行为克隆（BC）+MAPPO 使用 MAPPO 对一步高斯策略进行微调。表5 列出了这两个优化问题的差异所在。在上述分析中，对于任意智能体策略乘积成立的陈述，针对的是去噪轨迹上的策略。命题1 以及命题2的轨迹形式，对于一步高斯策略并无对应结论，因为其动作似然可以直接计算。一步基线策略在部署时将其潜变量固定为零，而 TeamFlowRL 在训练和部署期间的每个步骤都会采样一个新的潜变量。

表5 TeamFlowRL 与高斯基线行为克隆+MAPPO。行为克隆的数值+MAPPO 遵循其在机器人任务中的配置，详见表8。

	TeamFlowRL	行为克隆+MAPPO
初始策略	十步流教师 v_θ 来自流匹配	单步学生网络，潜变量固定为 zero
采样分布	由 $K = 10$高斯转移组成的马尔可夫链-移，带有可学习噪声 $\sigma_\phi(h^i, x^{i,k}, k) \in [0.01, 0.5]$	围绕学生网络输出的高斯分布，其标准差在各项之间共享且经过学习
存储样本	完成去噪轨迹 $\xi_t^i = (x_t^{i,0}, \ldots, x_t^{i,K})$	状态和智能体裁剪前的原始动作
似然	K 转移对数概率之和（公式-4）。无法直接计算动作边缘分布-出	原始动作的高斯似然
比率	每个智能体一个轨迹比率（公式7）	每个智能体一个动作比率
裁剪	每个智能体的比率，默认值为 $\varepsilon = 0.2$，com-在对数空间中计算	机器人任务中的每个智能体比率，$\varepsilon = 0.05$
正则化	报告的运行结果中无（$\beta = 0$，无熵项）	参考 KL 系数为 0.01，熵系-在机器人任务中高效达到 0.001
更新幅度控制	在 K 上求和的条件轨迹KL状态转移，阈值 0.02，可选的反向跟踪	动作的解析高斯KL，阈值旧值 0.02
部署	随机初始潜变量，确定性欧拉步，因此动作保持随机性	潜变量固定为零时的均值动作

B 实现细节

B.1 输入、动作映射和更新控制

输入与初始化。 执行器仅以其局部输入为条件。连续控制使用局部观测特征，并采用环境适配器提供的恒等编码。对于星际争霸，执行器以预训练编码器生成的局部循环特征为条件，该编码器在在线优化期间保持冻结。在线执行器从离线检查点导入教师速度网络。所有报告的配置均使用 $K = 10$ 个去噪步骤，并在方程中使用 22。

动作映射。 执行器会在进行任何动作变换之前存储采样的去噪轨迹。对于连续控制，执行动作是 $a^i = \text{clip}(x^{i,K}, -1, 1)$；对于星际争霸，则是 $x^{i,K}$ 中的最大合法坐标。

有效轨迹比率。 有三项实现要求确保方程7中的比率有效。首先，每条去噪轨迹都从方程 3 本身的含噪转移链中采样，并在采集时存储其对数似然。更新时会重新计算同一条已存储轨迹的似然，因此 $\rho_t^i = 1$ 在采集参数下成立，我们会在每次更新时进行检查。其次，$\log \rho_t^i$ 是所有 K 步数和动作坐标上的转移对数概率的精确总和，不进行平均或截断。第三，每个智能体都有

自己的比率，并针对共享团队优势进行裁剪。存储的轨迹已从计算图中分离，因此梯度只会通过存储前缀处的条件均值和标准差传入速度网络和噪声网络。

更新幅度与执行。 梯度裁剪和条件轨迹KL检查会限制更新幅度：一旦轨迹KL超过阈值，本次迭代剩余的演员更新便会跳过；此外，某些任务还会使用KL回溯，详见附录 B.2。有界转移均值定义了一种独立的修复变体，详见附录 B.2。训练使用带有转移噪声的样本，部署则遵循均值更新。

B.2 在线微调设置

使用采集阶段的评论家参数 ψ_{old}，实现会对每条包含 128 步的轨迹采样反向计算广义优势估计，

$$\begin{aligned}\delta_t &= r_t + \gamma(1-u_t)V_{\psi_{\text{old}}}(s_{t+1}^{+}) - V_{\psi_{\text{old}}}(s_t),\\ \widehat{A}_t &= \delta_t + \gamma\lambda(1-d_t)\widehat{A}_{t+1}, \qquad \widehat{R}_t = \widehat{A}_t + V_{\psi_{\text{old}}}(s_t),\end{aligned} \tag{20}$$

其中 s_{t+1}^{+} 表示任何重置发生前的后继状态， $\gamma = 0.99$，且 $\lambda = 0.95$。时间限制下有 $u_t = 0$ 和 $d_t = 1$，因此会从最终状态引导估值，但截断轨迹。价值目标在归一化前生成。每条轨迹的演员优势仅归一化一次，使用总体标准差，然后复制给每个智能体：

$$\widetilde{A}_t = \frac{\widehat{A}_t - \overline{A}}{\max\big(\text{std}(\widehat{A}), 10^{-8}\big)}, \qquad \widetilde{A}_t^{i} = \widetilde{A}_t \text{ for all } i. \tag{21}$$

执行器和评论家最小化

$$\begin{aligned}\mathcal{L}_{\text{actor}} &= -L_{\text{clip}},\\ \mathcal{L}_V &= \tfrac{1}{2}\,\mathbb{E}_t\big[(V_\psi(s_t) - \widehat{R}_t)^2\big],\end{aligned} \tag{22}$$

其中回报目标 $\widehat{R}_t$ 保持不变。实现会在对数空间中计算公式 8，并且仅裁剪由优势符号所选定的一侧，

$$\begin{aligned}&\min\big(\rho\widetilde{A}, \text{clip}(\rho, 1-\varepsilon, 1+\varepsilon)\widetilde{A}\big)\\ &\quad= \begin{cases} e^{\min(\log\rho,\, \log(1+\varepsilon))}\,\widetilde{A}, & \widetilde{A} \ge 0,\\ e^{\max(\log\rho,\, \log(1-\varepsilon))}\,\widetilde{A}, & \widetilde{A} < 0,\end{cases}\end{aligned} \tag{23}$$

这是一个恒等式。对于 $\widetilde{A} \ge 0$， $\rho\widetilde{A}$ 与 $\text{clip}(\rho)\widetilde{A}$ 的最小值为 $\min(\rho,\ 1+\varepsilon)\widetilde{A}$。对于 $\widetilde{A} < 0$，则为 $\max(\rho,\ 1-\varepsilon)\widetilde{A}$。更新幅度通过相对于采集策略的条件轨迹KL来衡量，

$$\begin{aligned}D_t^{i} = \sum_{k=0}^{K-1} D_{\text{KL}}\big(&\mathcal{N}(\mu_{\text{old}}^{i,k}, \text{diag}((\sigma_{\text{old}}^{i,k})^2))\\ &\|\,\mathcal{N}(\mu_\theta^{i,k}, \text{diag}((\sigma_\phi^{i,k})^2))\big),\end{aligned} \tag{24}$$

在存储的前缀上进行评估，并按权重 m_t^{i} 求平均。根据相对熵链式法则，其在来自采集策略的前缀上的期望等于轨迹 KL $D_{\text{KL}}(q_{\text{old}}\|q_{\theta,\ \phi})$。不启用回溯时，每执行一次执行器小批次步骤，都会将该小批次的值与目标 0.02 进行比较。若超过目标，则计算整个轨迹采样的值。若该值也超过 0.02，则停止本次迭代中剩余的执行器更新，但继续进行评论家更新。每轮次结束时也会检查轨迹采样的值。触发停止条件的步骤会予以保留。启用回溯时，只有轨迹采样的值不超过目标，步骤才会被接受。否则，将恢复参数和 Adam 状态，并以减半后的学习率重试该步骤，最多重试配置指定的次数。第一次迭代仅更新评论家。四种配置采用针对任务的优化设置：2-HalfCheetah Good、5m_vs_6m 中等、Tag

Random 和 Spread Random。它们保留十步生成器和完整轨迹比率，相关设置详见发布的代码。

集中式评论家是一个多层感知机，包含两个各有 256 个单元的隐藏层，并使用双曲正切激活。条件噪声网络包含一个有 64 个单元的隐藏层，并使用双曲正切激活。生成器架构从源检查点导入。源检查点还决定了星际争霸的局部循环编码器。编码器保持冻结，以便在回放似然时一致地使用存储的局部特征。

B.3 有界修复

有界修复将 $b\tanh(\mu/b)$ 应用于方程 μ_θ 中的转移均值，其中 $b = 10$。高斯噪声在此变换之后添加。两个 OMIGA 配置采用了这一结构性变更。6 个 HalfCheetah 专家和 3 个 Hopper 专家运行各完成 5000 万步，并见表 7 和图 8。它们的离线参考值是同一批教师在第零步时的原始评估结果。

C 基准实验

C.1 实验协议与报告约定

表 6给出了默认配置，每项任务的配置均见发布的代码。在线步数预算按联合环境转移次数计算。

表 6 在线微调设置。在线预算按联合环境转移计数，每个联合环境步对应一次转移。

基准测试	演员-评论器		γ	λ	Clip	KL	参考熵	训练轮数	小批次	环境	轨迹采样	梯度	预热惩罚	σ_0		执行器输入	评论家输入
SMAC, SMACv2	2e-5	3e-4	0.99	0.95	0.2	0.02	0	0	4	256	8	128	0.5	第1次迭代	off	0.1 冻结	状态
MPE、MA-MuJoCo	2e-5	3e-4	0.99	0.95	0.2	0.02	0	0	4	256	40	128	0.5	第1次迭代	off	0.1 循环特征局部观测	联合观测或状态

执行器、评论家：学习率。 γ：折扣。 λ：广义优势估计参数。裁剪：近端策略优化裁剪。KL：轨迹 KL 阈值。参考、熵：参考系数和熵系数。小批次：小批量大小。环境：并行环境。轨迹采样：展开长度。梯度：梯度范数裁剪。预热：评论家预热。惩罚：参考惩罚。 σ_0：初始高斯转移噪声。

正文中的表格 2报告了 MPE、OG-MARL 和 SMAC，表格 7报告了 OMIGA 和 SMACv2。每个离线源都提供了一个经过一百万次更新预训练的教师。Team-FlowRL 的所有基准测试条目共用 5000 万环境步数的在线预算。机器人任务各自保留 100 万至 500 万步的预算。6-HalfCheetah 和 3-Hopper 的 OMIGA Expert 条目在 5000 万步时采用有界修复，并标记为 ‡。学习曲线还包括两个 SMACv2 Good 数据集。

MPE 面板报告归一化分数。对于原始回报 R，得分为

$$100\frac{R - R_{\text{random}}}{R_{\text{expert}} - R_{\text{random}}}.$$

其余面板报告原始团队回报。文中的学习曲线和配对比较始终使用原始回报。

已发布的基线值保留其来源协议。SMAC 和 SMACv2 基线遵循 MAC-Flow [9],表 1。MPE 基线遵循 MAC-Flow 表 2，但标记为$_{\text{m}}$ 的条目除外，这些条目来自 MADiff [34], 表 1 和 DOM2 [12], 表 1。DOM2 使用其自行收集的数据集。其中的条目将 DOM2 表 1 中的原始回报，依据 DOM2 附录（第 22 页）给出的专家和随机参考回报，转换为归一化分数。DOM2 Spread 平均值由我们根据其四个 Spread 条目计算得出。OG-MARL 基线遵循 MADiff 表 1，OMSD 条目遵循 OMSD 表 2，OMIGA 基线遵循 MAC-Flow 表 2。

标记为 r 的条目是离线 MAC-Flow 学生模型的局部复现。其设置详见发布的代码。表 7 中标记为 r 的三个 SMACv2 Replay 条目，来自用于在这些任务上初始化 TeamFlowRL 的 MAC-Flow 预训练运行。MADiff-D 使用分散式执行，而 MADiff- C 使用集中式执行。已发布的平均值保留其来源覆盖范围，但 OMIGA 除外；OMIGA 的基线平均值对报告的九种任务与数据集组合进行等权重计算。TeamFlowRL Spread 平均值对四个数据集进行等权重计算，其 SMACv2 平均值对三个 Replay 任务进行等权重计算。浅绿色底纹和粗体表示所显示的最大均值。浅灰色底纹和下划线表示

表 7 OMIGA MuJoCo（连续控制）和 SMACv2（离散控制）上的已发表结果与 TeamFlowRL 结果，方法为行，任务和数据集对为列。符号表示遵循表 2。SMACv2 Good 数据集由我们自行收集，并将予以释放。已发布基线仅报告 Replay，因此，当 TeamFlowRL 在 Good 数据集上的结果超过同一任务的所有已发布基线时，该结果以阴影和粗体标示。MAC-Flow^{r} 是我们在局部复现的结果。

OMIGA	HalfCheetah		Hopper			Ant				Avg.
	专家	M-E	专家	M-R	M-E	专家	中等	M-R	M-E	
MACQL [8] (2020年)	4589.5 ±98.5	4738.2 ±181.1	3359.1 ±513.8	31.4 ±15.2	2751.8 ±123.3	2082.4 ±21.7	1033.9 ±66.4	434.6 ±108.3	1800.2 ±21.5	2313.5
MATD3BC [6] (2021年)	4401.6 ±169.1	3518.1 ±381.0	3309.9 ±4.5	269.7 ±41.8	2904.3 ±477.4	2046.9 ±17.1	1422.6 ±21.1	995.2 ±52.8	1636.1 ±96.0	2278.3
OMAR [17] (2022年)	−206.7 ±161.1	−253.8 ±63.9	2.4 ±1.5	3.3 ±3.2	1.4 ±0.9	312.5 ±297.5	−1710.0 ±1589.0	−2014.2 ±844.7	−2992.8 ±7.0	−762.0
OMIGA [31] (2023年)	3383.6 ±552.7	2948.5 ±518.9	859.6 ±709.5	774.2 ±494.3	709.0 ±595.7	2055.5 ±1.6	1418.4 ±5.4	1105.1 ±88.9	1720.3 ±110.6	1663.8
MADiff-D [34] (2024年)	4711.4 ±213.6	4410.9 ±836.8	2853.3 ±593.8	936.1 ±574.0	2810.4 ±723.2	2060.0 ±10.3	1428.4 ±14.7	1294.5 ±360.2	1740.2 ±158.9	2471.69
MAC-Flow [9] (2026年)	4650.0 ±271.6	5139.9 ±84.1	**3592.1** ±8.9	1166.3 ±451.9	**2988.3** ±480.2	2060.2 ±20.0	1432.4 ±17.8	1498.4 ±20.3	2053.3 ±20.4	2731.2
TeamFlowRL（不含在线强化学习）	3477.40 ±246.94	3021.08 ±128.55	2896.91 ±48.37	403.84 ±37.51	900.09 ±64.63	2048.93 ±6.48	1420.26 ±3.02	1092.11 ±19.17	1768.56 ±63.84	1892.13
TeamFlowRL	**6562.96**‡ ±346.62	**6828.97** ±91.78	3059.66‡ ±116.38	**3287.96** ±138.76	2055.42 ±102.72	**3867.80** ±27.18	**3874.97** ±35.42	**3761.87** ±23.21	**5193.92** ±108.55	**4277.06**

SMACv2	人族 5v5		异虫 5对5		人族 10对10	Avg.
	回放	Good	回放	Good	回放	
行为克隆 [20] (1991年)	7.3 ±1.0	--	6.8 ±0.6	--	7.4 ±0.5	7.2
MABCQ [7] (2019年)	13.8 ±4.4	--	10.3 ±1.2	--	12.7 ±2.0	12.3
MACQL [8] (2020年)	11.8 ±0.9	--	10.3 ±3.4	--	11.8 ±2.0	11.3
扩散行为克隆 [2] (2023年)	9.3 ±0.9	--	8.1 ±1.7	--	5.5 ±1.5	7.6
流行为克隆 [13] (2023年)	8.3 ±1.9	--	4.6 ±0.5	--	5.8 ±1.7	6.2
MADiff-D [34] (2024年)	13.3 ±1.8	--	10.2 ±1.1	--	**13.8** ±1.3	12.4
MAC-Flow [9] (2026年)	**16.6** ±4.3	--	9.8 ±1.5	--	13.0 ±4.7	**13.1**
MAC-Flow^{r}	9.70 ±1.40	--	7.01 ±1.27	--	6.03 ±0.57	7.58
TeamFlowRL（无在线强化学习）	7.22 ±0.74	14.22 ±0.45	6.24 ±0.56	12.42 ±0.37	4.77 ±0.94	6.08
TeamFlowRL	12.86 ±0.78	15.17 ±0.81	**10.86** ±1.18	**14.17** ±1.00	10.16 ±0.96	11.29 ±0.93

‡有界修复：转 移均值被限制为 10·tanh(mean/10)，这是为保证数值稳定性而进行的结构性变更。SMACv 2
平均值是对 th ree Replay 任务。

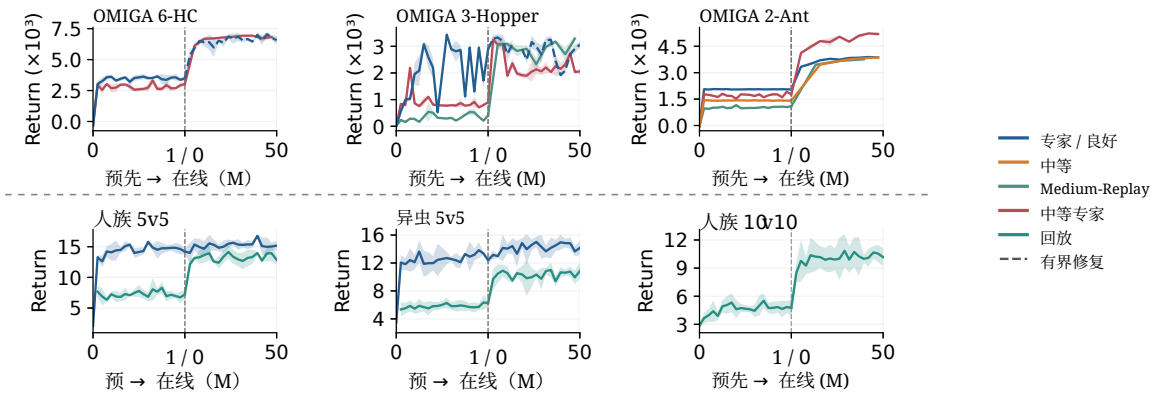

图 8 OMIGA MuJoCo 和 SMACv2 上的学习曲线，绘制方式与图 3相同。虚线表示采用有界修复的运行结果，6-HC 表示 6-HalfCheetah。

第二大的不同均值。四舍五入后相同的均值共享排名。M-R 和 M-E 分别表示 Medium-Replay 和 Medium-Expert。在主表中，匕首符号表示归一化 MPE 分数。

评估详情。 离线流比较采用与 TeamFlowRL 相同的评估规则。在 MPE 中，Medium-Replay 和 Random Spread 初始化方案也有所改进，但最终低于专家。表 2 报告了 MPE 的所有组合，采用已发表参考文献使用的归一化得分尺度。

表 8 采用状态观测的机器人实验配置。

		预训练					在线微调												
	方法	初始化	预先更新	演示数据	划分	批次	序列	步数	环境	回合长度	执行器	评论家	训练轮数	裁剪	KL 停止	参考 KL	熵	小批次	轨迹采样	执行
MS3	TeamFlowRL 10 步	50k	983/1007	90/10	128	20	5M	64	100	2e-5	3e-4	4	--	0.02	--	--	--	6400	随机 z
	行为克隆+MAPPO 单步	50k	983/1007	90/10	128	20	5M	64	100	2e-5	3e-4	4	0.05	0.02	0.01	0.001	128	6400	mean
RF	TeamFlowRL 10 步	30k	817/1024	735/82	--	--	2M	64	100	2e-5	3e-4	4	--	0.02	--	--	--	6400	--
	行为克隆+MAPPO 单步	30k	817/1024	735/82	--	--	2M	64	100	2e-5	3e-4	4	--	0.02	--	--	--	6400	mean

MS3：ManiSkill3。RF：RoboFactory。初始化：初始策略。执行器、评论家：学习率。预先更新：预训练更新。划分：训练/验证。序列：预训练序列长度。步数：在线环境步数。回合长度：回合长度。参考 KL：参考 KL 系数。熵：熵系数。小批次：在线小批次。执行：执行规则。两种方法在每个基准测试中共享预训练运行和演示数据划分。ManiSkill3 的演示数据按抓取/堆叠列出，RoboFactory 使用专家回合中成功的 817 个（共 1,024 个），并使用 5 万次更新中的 3 万次处的预训练检查点。TeamFlowRL 的执行规则会采样随机初始潜变量，且不加入转移噪声。轨迹批次大小按联合环境转移计数。

C.2 轨迹比率消融

我们每次只改变轨迹比率的一种设定，其他所有设置均保持默认配置不变，并使用 5,000 万环境步数的预算。我们将每个智能体在完整去噪轨迹上的轨迹比率与两种替代构造进行比较：分别裁剪每个去噪步骤的比率，就像 DPPO 对单智能体去噪转移所做的那样 [24]，；以及先对各步和各坐标的轨迹对数概率取均值，再将其钳制到 $[-5, 2]$。表 9 报告了 5,000 万步时的回报，以及最后 500 万步的平均回报。

表 9 TeamFlowRL 去噪轨迹比率的变体。最终：训练结束时的回报。最终最后 10%：训练最后 10% 的平均回报。

	Spread Medium		Spread M-R		2-Ant Poor	
比率	最终	最终最后 10%	最终	最后 10%	最终	最后 10%
每个智能体的完整轨迹（默认）	449.6	448.9	300.3	305.4	3195	3162
每个去噪步骤裁剪	401.5	406.2	301.2	301.8	3195	2971
平均后截断	441.2	434.7	255.5	263.2	3035	3052

分别裁剪每个去噪步骤会使 Spread Medium 中的回报降低约 43，使 2-Ant Poor 中的后期均值降低约 190，并且对 Spread Medium-Replay 没有影响。对数概率进行平均并截断后变化很小：尽管在训练结束时，约 74% 的样本受到截断的影响，但可学习转移噪声仍使足够多的样本处于截断范围内。当使用固定且更小的转移噪声时，平均对数概率一开始就高于上限，导致大多数梯度消失，因此截断只有与这种噪声结合使用时才会产生不利影响。

C.3 行为诊断详情

该诊断使用 TeamFlowRL 和行为克隆+MAPPO 的专家检查点，以及从头开始训练的 MAPPO 检查点。随机高斯基线从其学习到的动作分布中进行采样。

诊断使用来自每个固定初始场景的64次 rollout。分配标签描述最近地标；完全覆盖则要求每个地标都在指定距离阈值内。

D 机器人操作实验

D.1 机器人操作设置

ManiSkill3 任务。在 TwoRobotPickCube 中，立方体初始时只有左臂能够触及，目标位置则只有右臂能够触及，因此两条机械臂必须相互传递立方体，才能将其抬至目标位置。当立方体距离目标不超过 0.025 米且右臂静止时，回合即告成功。在 TwoRobotStackCube 中，每条机械臂只能触及两个立方体中的一个。右臂将自己的立方体放入机器人之间中线上的目标区域，左臂再将第二个立方体堆叠在其上。当堆叠的立方体位于目标区域内且两条机械臂都没有抓着立方体时，回合即告成功。每条机械臂都能观测自身的关节位置和速度、自身的末端执行器位姿、立方体的位姿、目标位置、这些物体之间的相对位置、剩余时间以及智能体指示器，但无法观测另一条机械臂。物体和目标的位姿属于特权信息，因为实体机器人必须通过传感器估计这些信息。因此，状态结果衡量的是已知物体位姿条件下的协同控制，而附录 D.5则将这些位姿替换为相机观测。

预训练与在线训练。 离线预训练使用 Man-iSkill3 发布的所有成功演示数据，其中 TwoRobotPickCube 有 983 条，TwoRobotStackCube 有 1007 条，并按 90% 用于训练、10% 用于验证进行划分。两种方法都从同一次预训练运行开始，共进行 50,000 次更新；在此过程中，冻结的循环编码器会汇总每条机械臂的观测历史。TeamFlowRL 从十步流策略开始。BC+MAPPO 从流策略蒸馏得到的单步策略开始，并将潜变量固定为零，因此输出为其高斯策略的均值。在线微调运行 500 万环境步数，使用 64 个并行环境，每个回合持续 100 个环境步数。

RoboFactory。 在 RoboFactory 的 TwoRobotsStackCube 中，每条机械臂都会接收自身的联合状态和末端执行器状态、立方体位姿、目标区域以及智能体指示器，但不会接收另一条机械臂的信息。每个动作都会设定一个环境步内的关节位置变化，每个回合持续 100 步。训练使用分阶段稠密奖励，我们针对到达、抓取、抬升、对齐和放置进行定义，并将其归一化至最高为 1；成功与否则依据 RoboFactory 的评估函数判定。演示数据来自一个使用稠密奖励训练的近端策略优化专家。预训练使用随附的脚本化运动规划演示数据，RoboFactory 的闭环成功率为 0%。两种方法均从相同的预训练检查点开始。在线微调使用 ManiSkill 状态实验的优化设置，执行 200 万环境步数。

释放与撤回。 主表和 RoboFactory 学习曲线报告的是双手撤离成功，这还要求双机械臂释放方块并撤回，同时堆叠能够自行保持直立。附录D.6 定义了这一变体。

D.2 部署规则下的成功

图10 展示了两种方法在其部署规则下的成功情况，从预训练到在线微调阶段，分别在机器人任务中，每个数据点评估100个回合。TeamFlowRL会采样新的初始 al

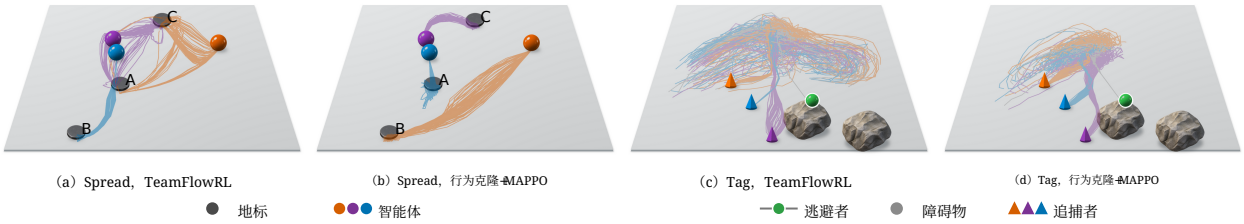

图9 按照图5所示绘制的 MPE rollout，每种方法从一个初始场景生成64次 rollout。（a）--（b）来自第二个初始场景的 Spread。（c）--（d）Tag，其中彩色路径表示三名追捕者，灰色路径表示逃避者。

潜变量且不添加转移噪声，而行为克隆+MAPPO执行其高斯均值。在在线训练结束时，两种方法在每项任务中的成功率差异均不超过5%。尽管最终成功率较为接近，TeamFlowRL仍保留了流策略的多模态动作分布，因为每次执行都会采样新的初始潜变量；而行为克隆+MAPPO执行单一的均值动作。附录 D.3 展示了执行路径由此产生的变化。在表 3的随机执行设置下，TeamFlowRL在六项任务上的平均成功率为87.7%，而行为克隆+MAPPO为54.6%。

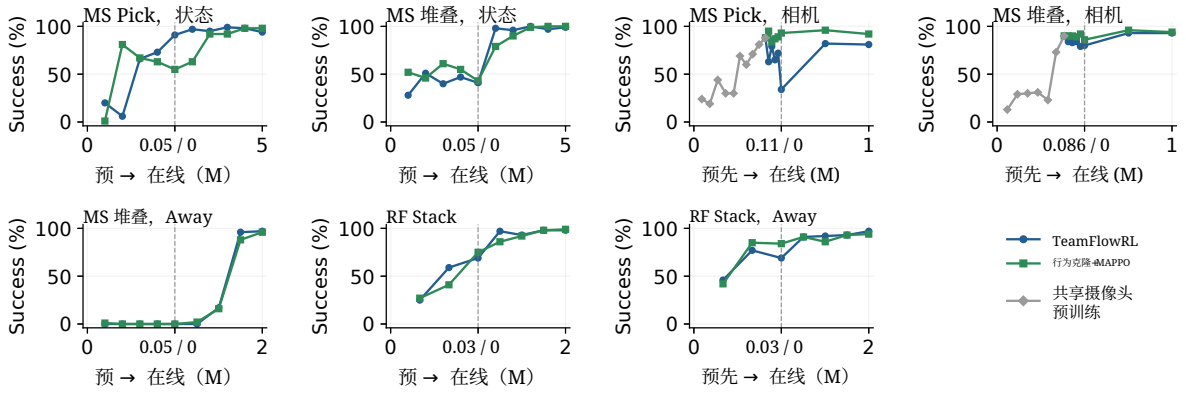

图 10 从预训练到在线微调期间，各方法部署规则下的成功率，每个数据点对应 100个回合。横轴的绘制方式与图 3相同，在线训练开始处标出了预训练长度。在相机面板中，灰色数据点表示两种方法共享的 Flow预训练和 DAgger；随后，每种方法都在冻结编码器上拟合各自的动作头。“Away”表示手部移开变体。

D.3 部署规则下的轨迹

图 11 展示了在各自部署规则下运行每种策略时，两个机械臂的工具中心点路径；每种策略均从一个初始场景运行 31 次。TeamFlowRL 在每次运行中都会抽取新的初始潜变量，且不添加转移噪声，而 BC+MAPPO 执行其高斯均值。为区分模拟器造成的变化，我们还让 TeamFlowRL 在全部 31 次运行中使用相同的潜变量。我们将路径的离散程度定义为：在前 40 个控制步以及两个机械臂上，工具中心点相对于其跨运行均值的均方根距离；随后在四个 MS Pick 场景和八个 RF Stack 场景上取平均值。在 MS Pick 中，TeamFlowRL 的离散程度为 1.6 cm，使用相同潜变量时为 0.4 cm，BC+MAPPO 为 0.3 cm。在 RF Stack 中，三者分别为 0.9 cm、0.2 cm 和 0.3 cm。TeamFlowRL 的潜变量采样使执行路径产生的变化达到仅由模拟器造成的变化的数倍，而 BC+MAPPO 则沿着单一路径运行。变化始终围绕一条路径保持在厘米级别。每个组中至少有 98% 的运行取得成功。

D.4 探索噪声下的双臂堆叠

图 12 展示了 ManiSkill3 中的 TwoRobotStackCube，其布局如图 7所示。TeamFlowRL 和行为克隆+MAPPO 从相同场景开始执行，且均加入了探索噪声。在该场景中，Team- FlowRL 在 30 次运行中有 93% 的成功率，而行为克隆+MAPPO 的成功率为 77%。在图示的 TeamFlowRL 运行中，右臂将绿色立方体放入目标区域，左臂将蓝色立方体堆叠在其上。在图示的行为克隆+MAPPO 运行中，机械臂将两个立方体都移入目标区域，但最终两个立方体并排放置，并未在最后一步完成堆叠。

D.5 使用相机观测进行双臂操作

图 13 总结了 TeamFlowRL 的相机策略是如何获得的。每条机械臂都会接收由两条机械臂共享的全局视图和其自身的腕部视图，这些视图由冻结的 DINOv2 模型编码 [16], , 同时还输入其本体感知状态、目标位置、剩余时间和智能体指示器。立方体位姿

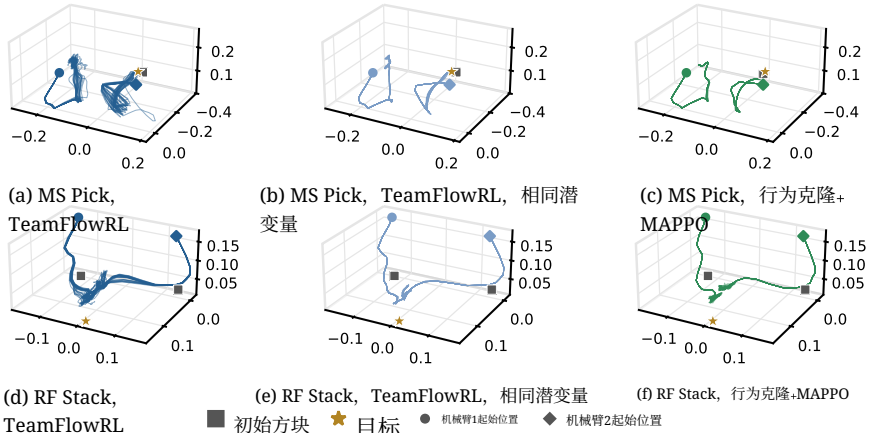

图 11 在部署规则下的工具中心点路径，每个任务从一个初始场景运行 31 个回合。(a)、(d) TeamFlowRL 在每个回合使用新的初始潜变量。(b)、(e) TeamFlowRL 在所有回合中使用相同的潜变量。(c)、(f) 行为克隆+MAPPO。每一行内的坐标轴共用。

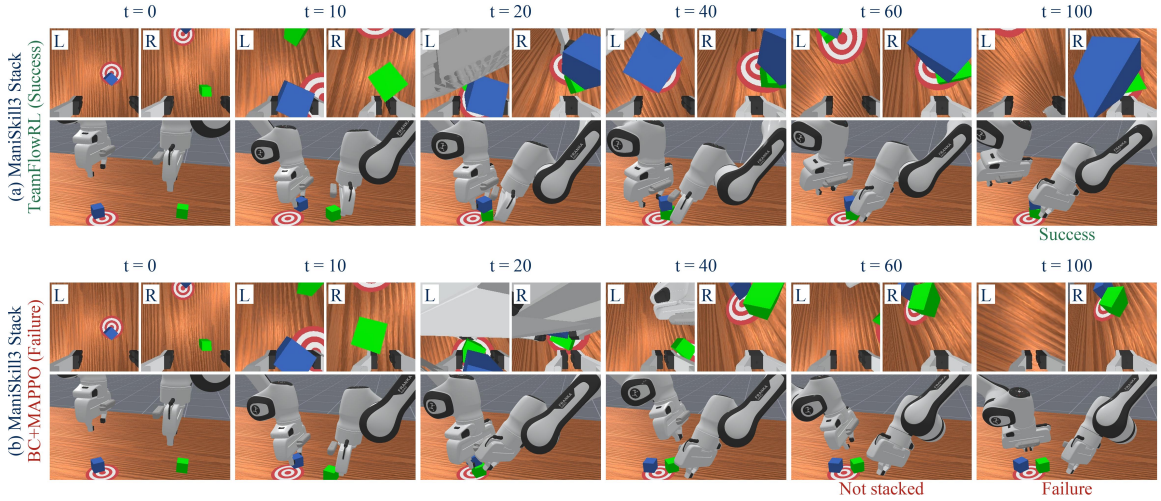

图 12 在 ManiSkill3 中记录的 TwoRobotStackCube 回合。(a)–(b) TeamFlowRL 和行为克隆+MAPPO 在探索噪声下从相同场景开始运行，每帧上方显示左侧 (L) 和右侧 (R) 的腕部视角。

仅对状态教师和集中式评论家可用。状态教师通过 DAgger 训练相机策略。仅在离线演示数据上预训练的相机策略，在部署规则下于 TwoRobotPickCube 中的成功率达到 22%，在 TwoRobotStackCube 中达到 23%；这是因为其自身的轨迹采样很快就会到达演示数据未覆盖的状态。在 DAgger 中，相机策略控制机械臂，状态教师为每个访问状态提供标签，而相机策略则使用所有轮次的数据进行训练。DAgger 策略使用单步动作头，而 TeamFlowRL 优化的是十步去噪轨迹的似然。因此，我们冻结该策略的编码器，并在 DAgger 数据上使用条件流匹配拟合十步流头，将教师动作作为目标。在线微调保持编码器冻结，并使用状态实验的优化设置运行一百万个环境步数。其余设置见发布的代码。

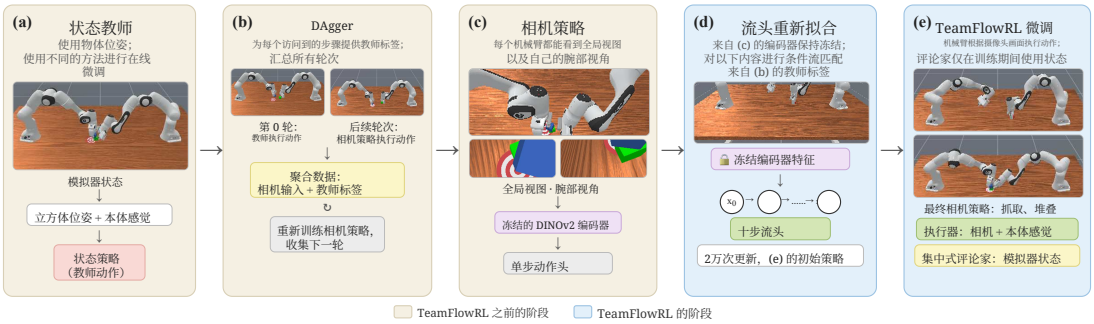

图13 通过 DAgger 使用相机观测训练 TeamFlowRL，包括在冻结编码器上拟合流头，以及进行在线微调。蓝色面板标示 TeamFlowRL 的阶段，米色面板标示之前的阶段。

D.6 学习释放与撤回

ManiSkill 和 RoboFactory 中双臂堆叠任务的手部移开变体要求机械臂释放方块、撤回，并让完成后的堆叠保持直立。当原始任务成功，且在回合结束时两条机械臂都没有接触方块、双手也都远离方块时，该回合即为成功。

在 ManiSkill 中，训练从预训练堆叠策略开始，并增加一个用于撤回的奖励项。在 RoboFactory 中，原始演示数据不包含撤回，因此流策略会先在包含撤回的新演示数据上进行预训练。阈值和其余设置详见发布的代码。