Learning Human-Robot Collaboration via Heterogeneous-Agent Lyapunov Policy Optimization

Hao Zhang[1,2] Yaru Niu[2] Yikai Wang[2] Ding Zhao[2,*] H. Eric Tseng[1,*]

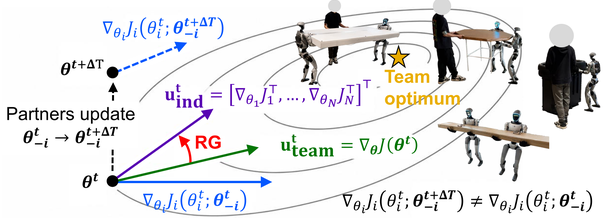
[1]ETAIC Lab, University of Texas at Arlington [2]Safe AI Lab, Carnegie Mellon University *Corresponding Authors

ICML International Conference On Machine Learning

Oral Presentation

Motivation

Scripted human proxies lead to OOD performance collapse in human-robot collaboration (HRC), motivating multi-agent reinforcement learning (MARL). However, MARL suffers from non-stationarity and a structural rationality gap (RG), as decentralized updates deviate from team-optimal paths.



Problem Formulation

We model HRC as a decentralized POMDP $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, P, R, \gamma, N, \mathcal{O}, Z \rangle$. To maximize the global return $J(\boldsymbol{\theta})$, heterogeneous agents utilize decoupled policies $\pi_{\theta_i}(a_{i,t}|o_{i,t})$, operating within a joint parameter $[\theta_1^\top, \dots, \theta_N^\top]^\top \in \mathbb{R}^{\mathbb{D}}$.

Misalignment in Decentralized MARL

Decoupled Learning Dynamics – Interaction between local agent intentions and the global team objective drives learning in CTDE:

- **Independent Rationality Field $\boldsymbol{u}_{\mathbf{ind}}$ –** Assuming other agents are stationary:

$$\mathbf{u}_{\mathbf{ind}}(\boldsymbol{\theta}) \triangleq [\nabla_{\theta_1} J_1^\top, \dots, \nabla_{\theta_N} J_N^\top]^\top \in \mathbb{R}^{\mathbb{D}}$$

- **Team Rationality Field $\boldsymbol{u}_{\mathbf{team}}$ –** True ascent direction of the global team return $J(\boldsymbol{\theta})$:

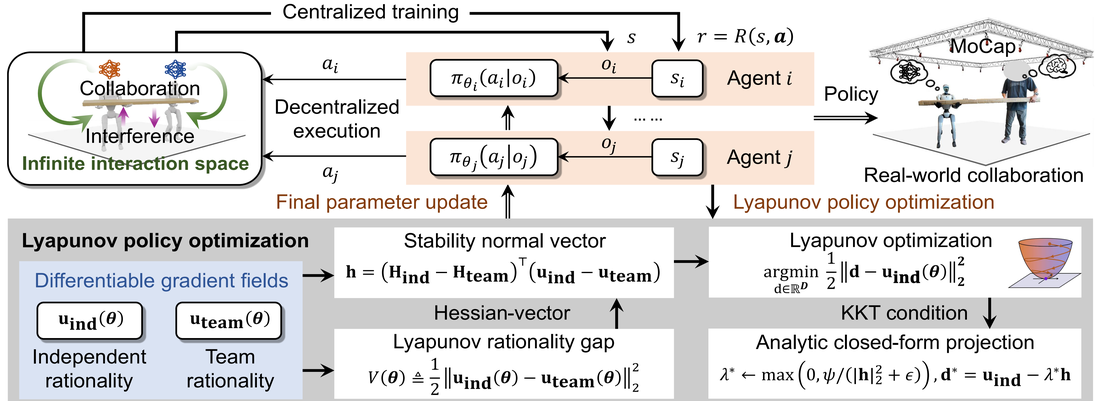
$$\mathbf{u}_{\mathbf{team}}(\boldsymbol{\theta}) \triangleq \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \left[\frac{\partial J}{\partial \theta_1}^\top, \dots, \frac{\partial J}{\partial \theta_N}^\top\right]^\top \in \mathbb{R}^{\mathbb{D}}$$

Contact
Dr. Hao Zhang, Postdoc Researcher
Carnegie Mellon University
Email: haoz4@andrew.cmu.edu

Project Website

HALO Methodology

We propose **heterogeneous-agent Lyapunov policy optimization (HALO)**, a framework that stabilizes decentralized MARL by enforcing Lyapunov-based contraction in policy-parameter space. Unlike Lyapunov-based safe RL, which targets state/trajectory constraints in constrained Markov decision processes, HALO uses Lyapunov certification to stabilize decentralized policy learning.



RG & Lyapunov Objective

The Rationality Gap $V(\boldsymbol{\theta})$ – Defined via a Lyapunov function measuring the mismatch:

$$V(\boldsymbol{\theta}) \triangleq \frac{1}{2} \|\mathbf{u}_{\mathbf{ind}}(\boldsymbol{\theta}) - \mathbf{u}_{\mathbf{team}}(\boldsymbol{\theta})\|_2^2$$

Objective – Enforce a Lyapunov dissipation constraint for asymptotic contraction:

$$\langle \nabla_{\boldsymbol{\theta}} V, \mathbf{d} \rangle \leq -\sigma V(\boldsymbol{\theta}), \quad \sigma > 0$$

Stability-Constrained Projection

Quadratic Program (QP) – Minimum-norm projection onto stable half-space:

$$\min_{\mathbf{d}} \frac{1}{2} \|\mathbf{d} - \mathbf{u}_{\mathbf{ind}}(\boldsymbol{\theta}_k)\|_2^2 \quad \text{s.t.} \quad \langle \nabla_{\boldsymbol{\theta}} V(\boldsymbol{\theta}_k), \mathbf{d} \rangle \leq -\sigma V(\boldsymbol{\theta}_k)$$

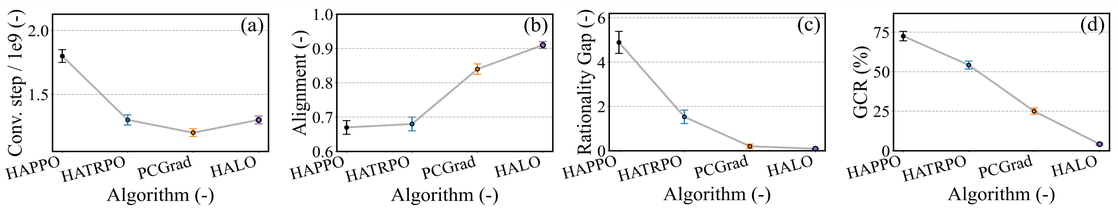
Exact Analytic Solution (KKT) – Here, $\mathbf{h} \triangleq \nabla_{\boldsymbol{\theta}} V(\boldsymbol{\theta}_k)$:

$$\mathbf{d}^* = \mathbf{u}_{\mathbf{ind}} - \max\left(0, \frac{\langle \mathbf{h}, \mathbf{u}_{\mathbf{ind}} \rangle + \sigma V}{\|\mathbf{h}\|_2^2 + \epsilon}\right) \mathbf{h}$$

Scalability via Hessian-Vector Product

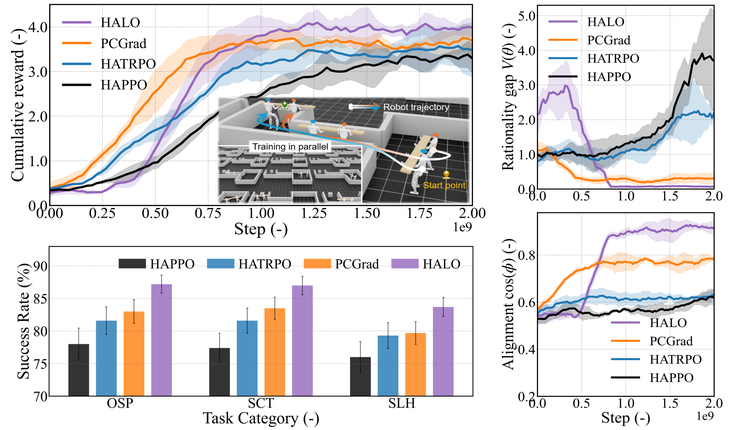
The Bottleneck: Computation of $\mathbf{h} = (\mathbf{H}_{\mathbf{ind}} - \mathbf{H}_{\mathbf{team}})^\top (\mathbf{u}_{\mathbf{ind}} - \mathbf{u}_{\mathbf{team}})$ requires $\mathcal{O}(D^2)$ Hessian.

HALO Solution: Bypasses matrix materialization via double back-propagation (create_graph=True), executing a cheap Hessian-vector product (HVP).

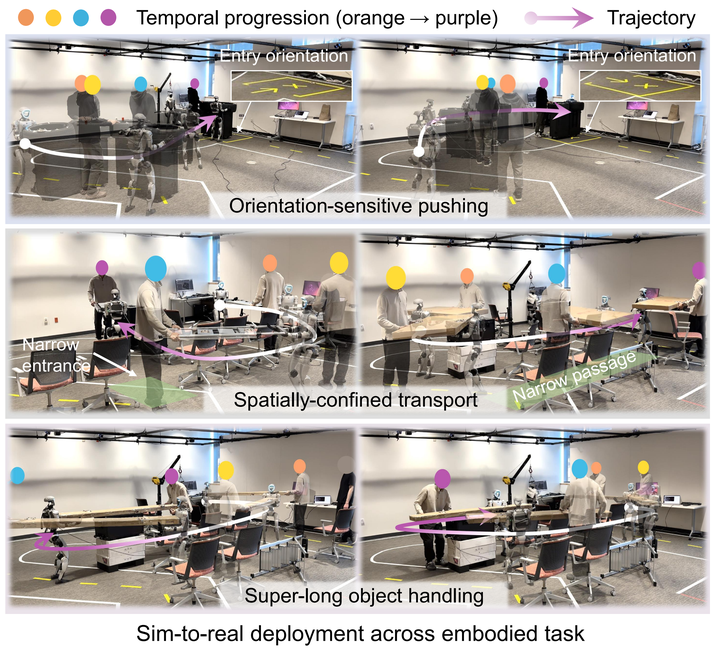


Experiments and Results

Extensive simulations and real-world robot experiments show that this certified stability improves generalization and robustness in collaborative corner case.



Simulation benchmarks and learning dynamics analysis



Sim-to-real deployment across embodied task