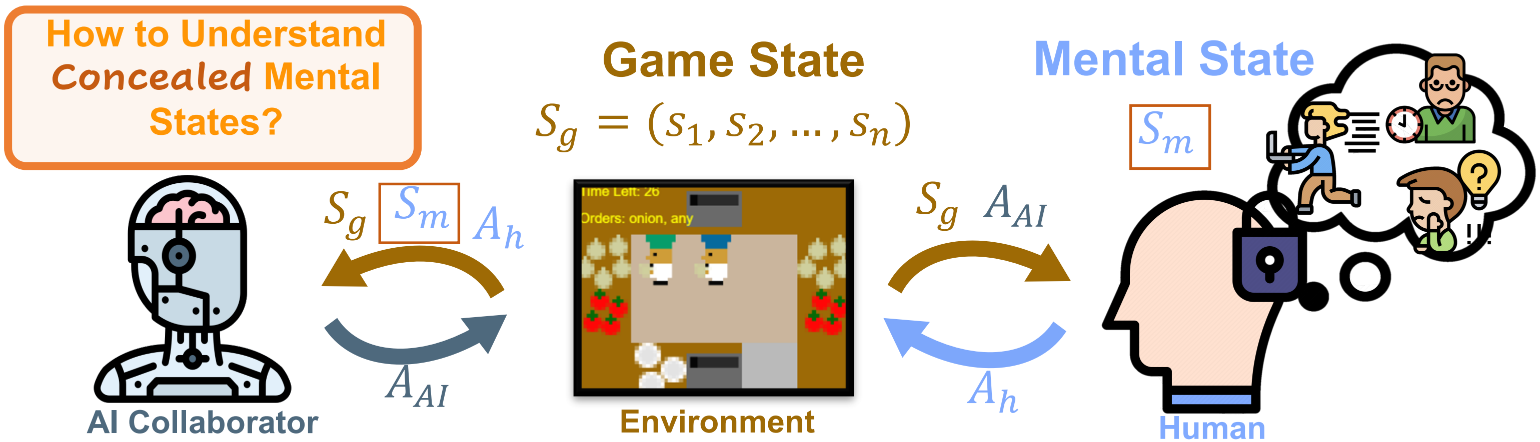
Beyond Single Stationary Policies: Meta-Task Players as Naturally Superior Collaborators

Haoming Wang[*,1], Zhaoming Tian[*,1], Yunpeng Song[1]
Xiangliang Zhang[2], Zhongmin Cai[†,1]

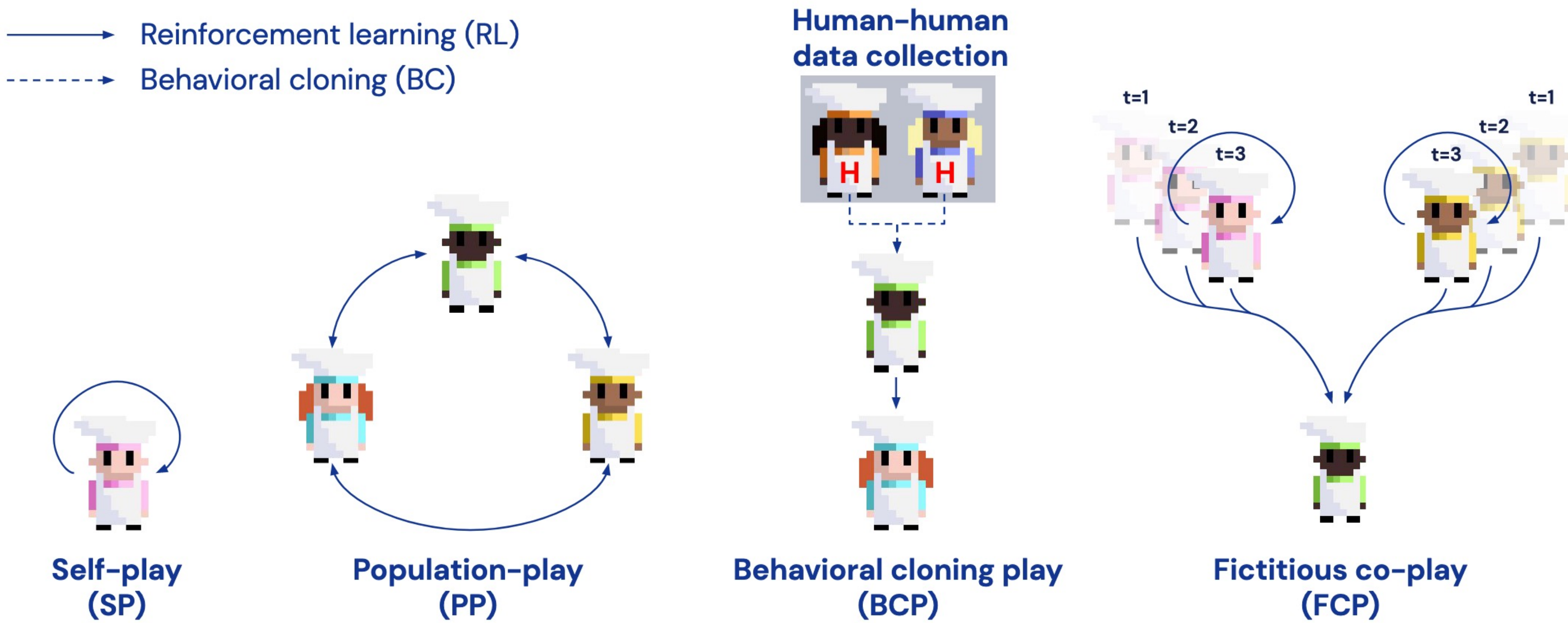
[1]Xi'an Jiaotong University
[2]University of Notre Dame

Non-stationarity of Human Behavior



- The distribution of human behavior, influenced by mental models, is non-stationary, manifesting in various levels of initiative and different collaborative strategies
- For human, the probability distribution $P(A|S_t)$ of action A given an environmental state S_t changes over time, reflecting different mental states
- Such non-stationarity poses a significant challenge in training collaborative agents, as it requires strategies that can adapt to the unpredictable nature of human behavior

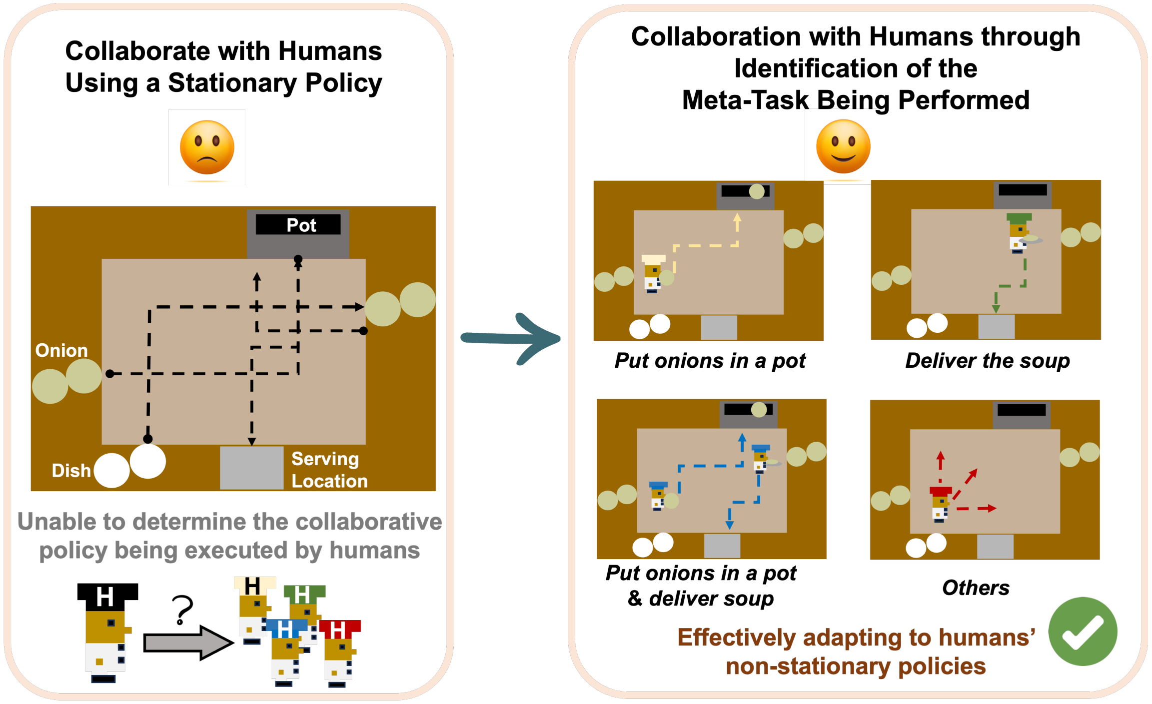
How to Collaborate with Non-stationary Humans?



- Recent works develop collaborative agents through two workflows: using human data (i.e., BCP) OR without human data (i.e., SP, PP and FCP[1])
- They are essentially policy networks following a stationary distribution, thus making it difficult to cope with non-stationary human dynamics
- How to collaborate with non-stationary humans efficiently?

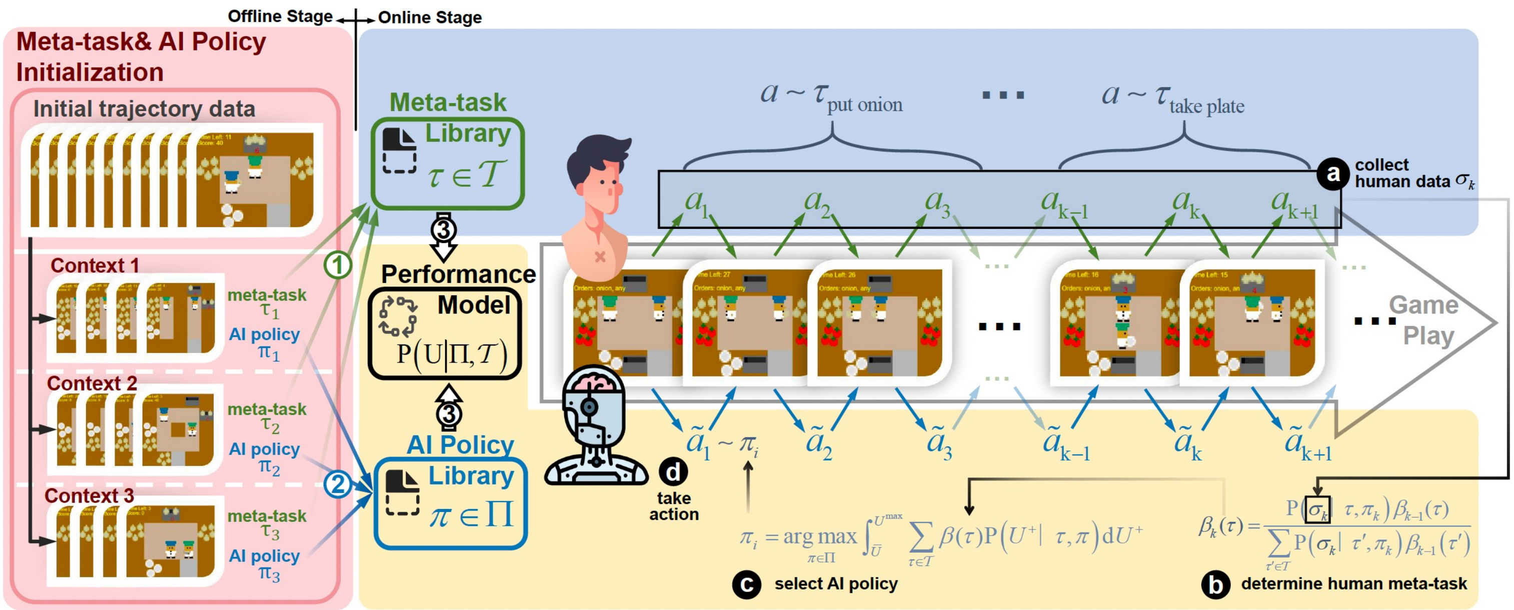
[1] Collaborating with Humans without Human Data. In 35th Conference on Neural Information Processing Systems (*NeurIPS* 2021).

Our Insight: Collaborating through Identification of Meta-tasks



- We discern that despite the inherent diversity in human behaviors, the underlying meta-tasks within specific collaborative contexts tend to be strikingly similar
- Our approach focuses on identifying the meta-tasks underlying human decision-making and trains collaborators to match these meta-tasks in a one-to-one manner
- For example, in the multi-player cooking game *Overcooked*, meta-task set includes {*place onions in pot*, *deliver soup*, *place onions in pot & deliver soup*, *others*}

Overview of Our CBPR Framework



- **Offline Training Stage:** (1) constructing meta-task models. (2) developing cooperative policies for each meta-task. (3) establishing a performance model by evaluating each meta-task and AI policy pair
- **Online Collaboration Stage:** (a) gathering current human data. (b) determining the current meta-task undertaken by the human. (c) selecting the most suitable AI policy. (d) the AI collaborator executes actions according to the selected policy

Theory Analysis of CBPR: *Collaboration Convergence*

- We formulate human-AI collaborative process as a Non-Stationary MDP (NS-MDP)[1]. In this process, the non-stationarity, can be mitigated by decomposing the entire non-stationary decision process into several stationary ones. *Each stationary MDP corresponds to a specific meta-task executed by the human*

THEOREM 1 (Collaboration Convergence of CBPR Agent). *Let $H_i := \{S_i^j, \pi_{hu,i}(S_i^j), R^j\}_{j=0}^{\infty}$ be a trajectory collected from a single stationary MDP M_i within the overall NS-MDP $\{M_i\}_{i=1}^{\infty}$ under the human meta-task policy $\pi_{hu,i}$. Denote $\mathcal{D} := \{(i, H_i) : i \in [1, k]\}$ as a random variable representing a set of trajectories observed prior to the most recently completed stationary MDP M_k. Given $\mathcal{D}$, the response policy of CBPR agent could almost sure converge when interacting with a human partner, even when the human's policy is non-stationary.*

- **Assumptions:** Within each stationary MDP M_{i}, the human policy $\pi_{\{hu,i\}}: S \to \Delta(A)$ is assumed to be stationary, although it may exhibit variations across different stationary MDPs
- We proved that CBPR policy could sure converge when collaborating with human partner, even when the human's policy is non-stationary. This convergence encompasses two parts:

1) CBPR agent identifies the evolving human behavior policy and iteratively updates its belief, converging asymptotically to the underlying true policy

2) As the belief converges, the CBPR algorithm stabilizes, converging to a fixed response policy

[1] Towards Safe Policy Improvement for Non-Stationary MDPs. In 34th Conference on Neural Information Processing Systems (*NeurIPS* 2020).

Theory Analysis of CBPR: *Collaboration Optimality*

THEOREM 2 (Collaboration Optimality of CBPR Agent). *Denoting* CBPR *for CBPR algorithm, let* $\rho(\pi, m) := \mathbb{E}[\int_{\bar{U}}^{U^{\max}} \mathrm{P}(U^+ \mid \tau(m), \pi)\, \mathrm{d}U^+]$ *be the expected return of exploiting AI policy* π *with human meta-task policy* $\tau(m)$ *in MDP* M_m*. Given a positive integer* k *and a set of trajectories* $\mathcal{D}$ *observed prior to the MDP* M_k*, it follows that for any subsequent stationary MDP* $M_{k+\delta}$*, we have:*

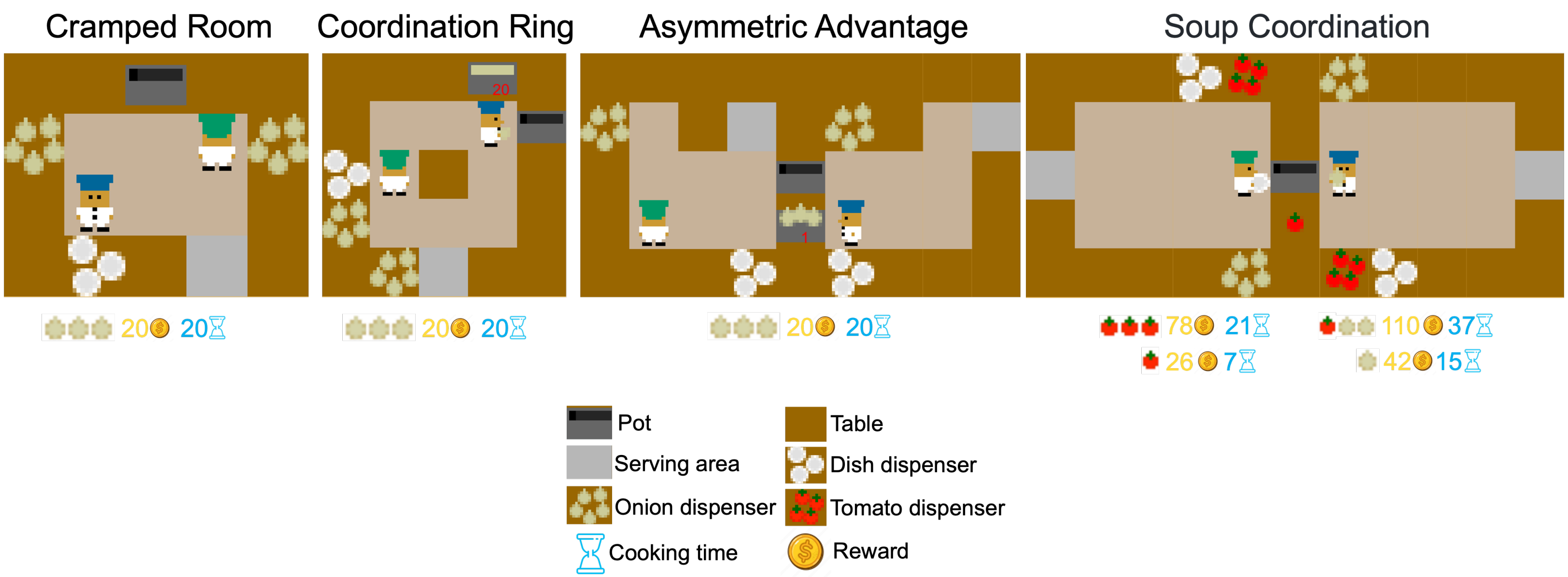
$$\Pr\Big(\rho(\mathrm{CBPR}(\mathcal{D}), k+\delta) \geq \rho(\pi_k^\star, k+\delta)\Big) \to 1 \tag{9}$$

when $k \to \infty$*, where* $\pi_k^\star$ *is the optimal response policy for human meta-task policy at MDP* M_k*.*

- **Assumption:** human policy library and AI policy library encompass all possible human meta-task policies and their corresponding best AI response policies. Although this assumption is too strong to be fully met in practice, optimal performance can still be ensured by augmenting both the human and AI policy libraries with a set of *primitive policies*[1]
- We proved that the collaboration policy generated by CBPR Agent is better than any possible stationary response policy in the long run

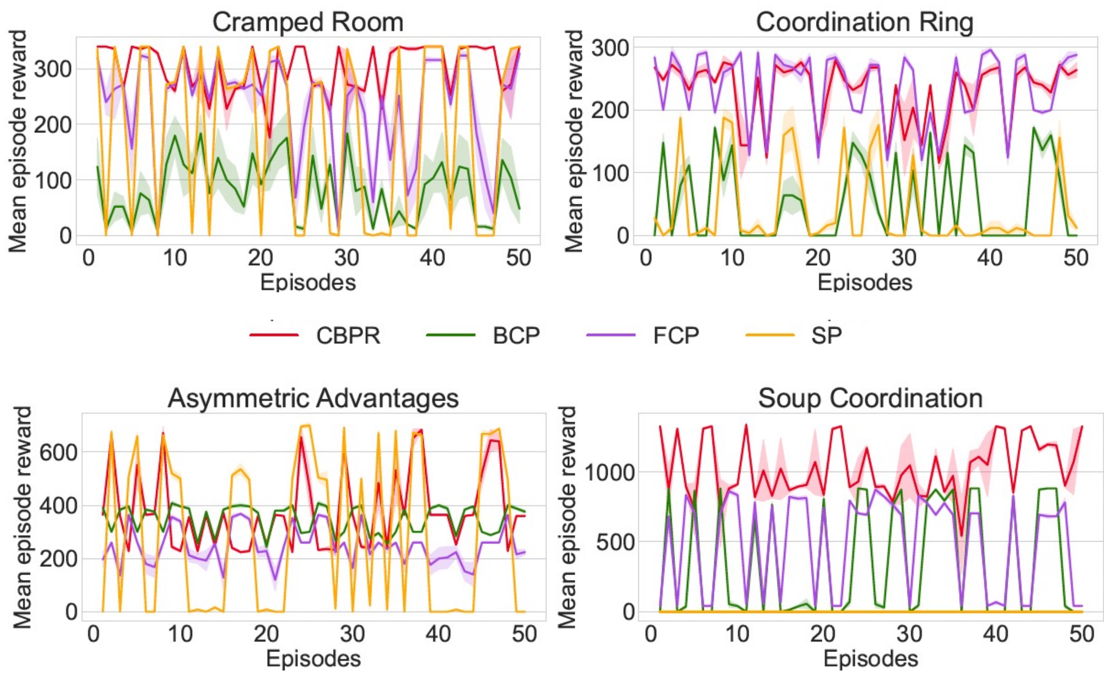
[1] Context-Aware Policy Reuse. In Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (*AAMAS* 2019).

Experimental Setups



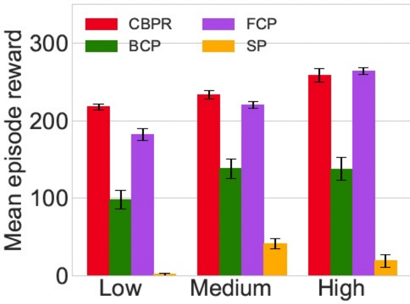
- **Q1:** When interacting with non-stationary agents who switch their strategies, can CBPR outperform established baselines?
- **Q2:** When interacting with non-stationary agents of various collaboration skills, can CBPR surpass other baselines?
- **Q3:** Can CBPR exceed the performance of other baselines in collaboration with real humans?
- **Q4:** How do hyperparameters and number of predefined meta-tasks influence the collaborative performance of CBPR agents?

Results: Collaborating with Rule-based Agents under Dynamic Policy Switching

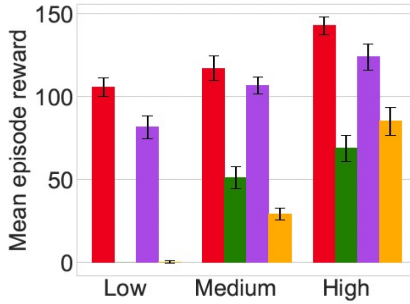


- CBPR consistently outperformed the baseline methods in the majority of cases
- FCP and SP agents experience greater fluctuations in episodic rewards, primarily due to their inability to effectively collaborate with all agents

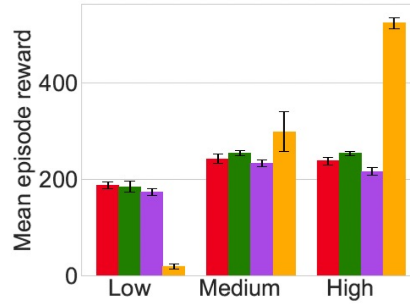
Results: Collaborating with Partners of Various Collaboration Skills



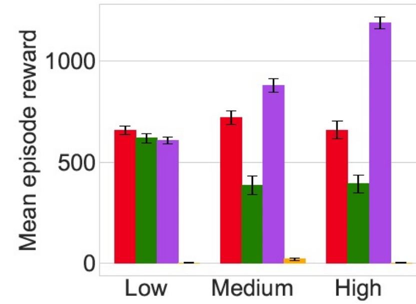
Cramped Rm.



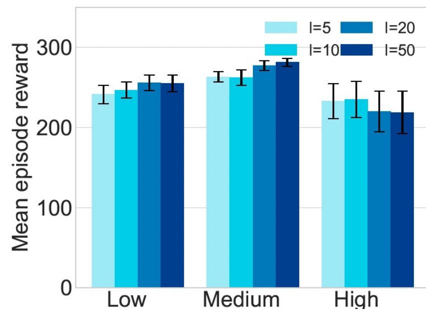
Coord. Ring



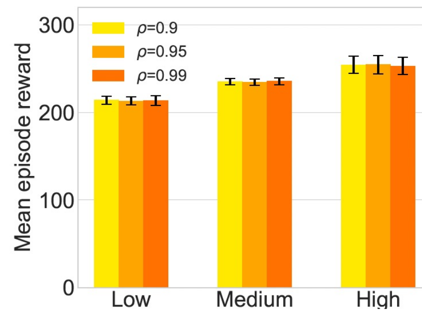
Asymm. Adv.



Soup Coord.



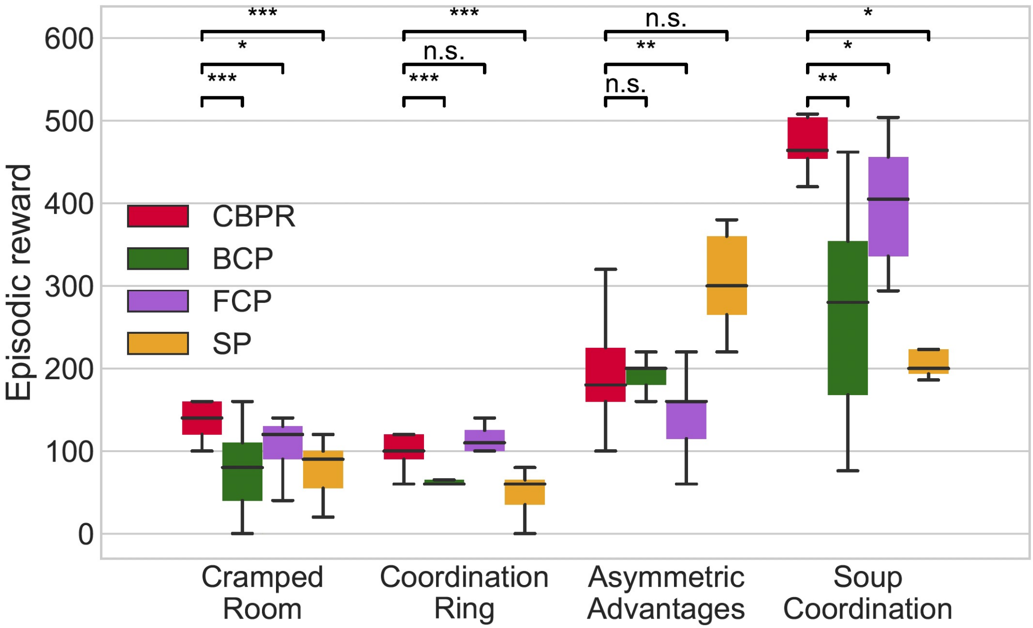
Ablations of l



Ablations of ρ

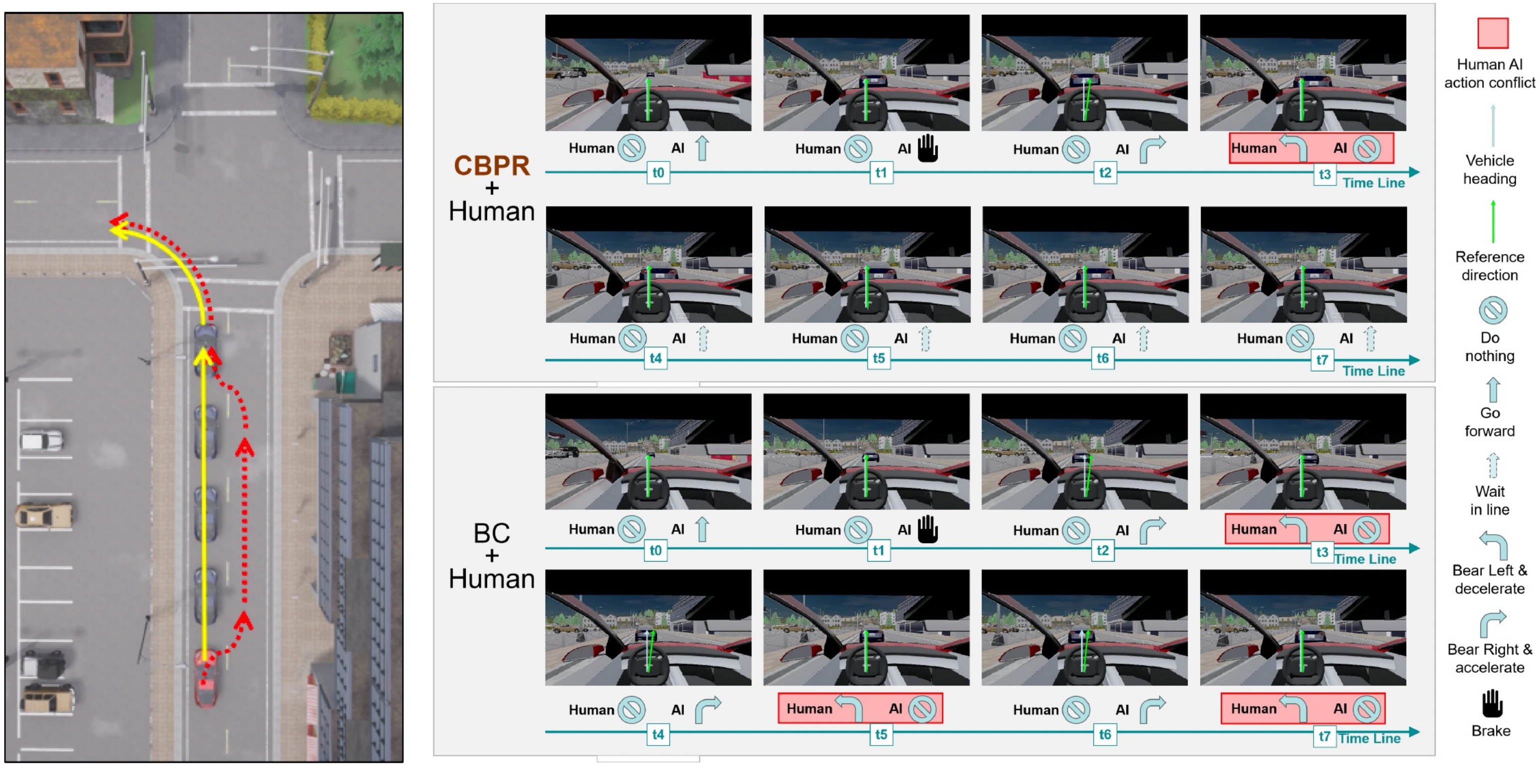
- CBPR consistently achieved higher mean episode rewards than FCP, particularly when collaborating with lower-skilled partners
- CBPR with large l performed well when collaborating with partners using low and medium skill levels, variations in ρ have little impact on the reward

Results: Collaborating with Real Humans



- In most comparisons, CBPR displays significant higher reward according to the one-sided Mann-Whitney U test

Limitations and Future Work



- How to model meta-tasks and establish meta-task library automatically based on human trajectories and collaborative task contexts ?
- How does CBPR performs in real-world domains such as power grid dispatching and autonomous driving ?

西安交通大學

XI'AN JIAOTONG UNIVERSITY

Thanks

Source Code

Paper Link

Wechat