Improving Human-AI Coordination Through Online Adversarial Training and Generative Models

Paresh Chaudhary, Yancheng Liang, Daphne Chen, Simon Du, Natasha Jaques

Human-AI Coordination

The Challenge

Coordination is a critical AI task: optimizing agents for enhanced human experiences.

GOAL: Train a **Cooperative Agent (Cooperator)** to coordinate and adapt to humans in real time.

Key Assumptions:

- Humans are diverse
- They have preferences and biases

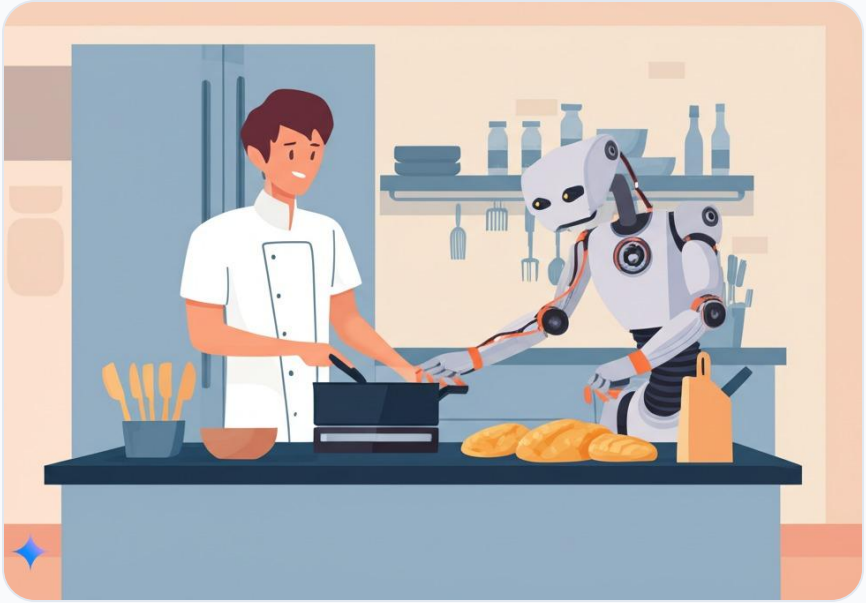


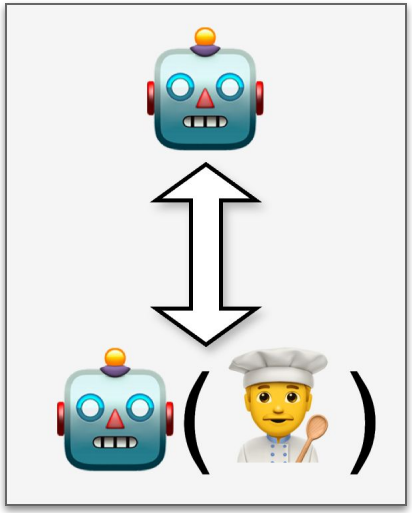
*image generated by Gemini

Cooperator: 

Training Strategies

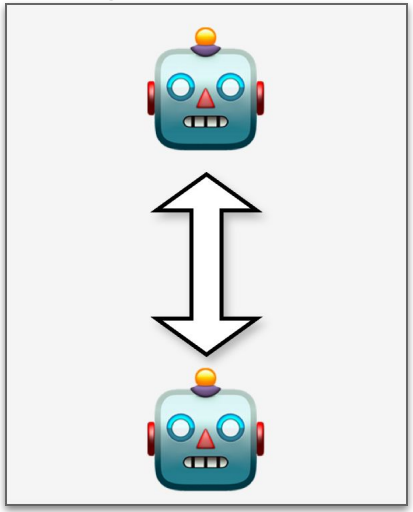
Behavior Cloning

- Brittle policies
- Limited human data



Self-Play

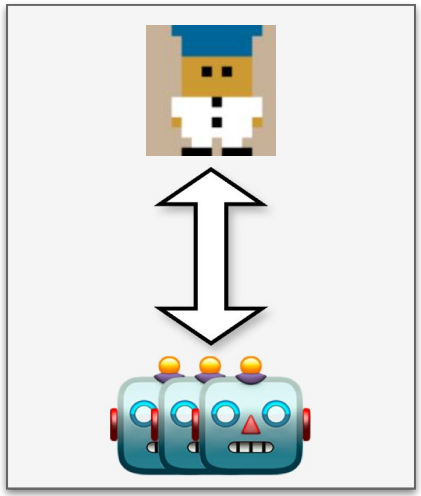
- Rigid policies
- Does not generalize to humans



Training Strategies

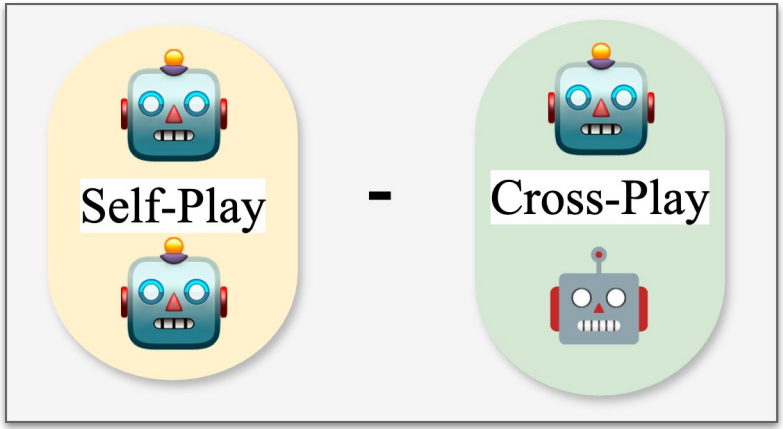
Population of Agent

- Rigid policies
- Fails to cover all scenarios

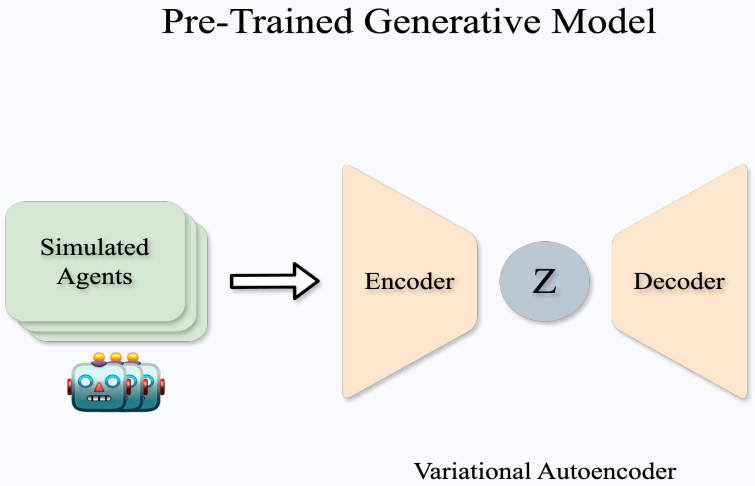


Self-Play & Cross-Play

- Converges to few conventions
- Learns to sabotage the game



Learning A Generative Model On Agent Policies

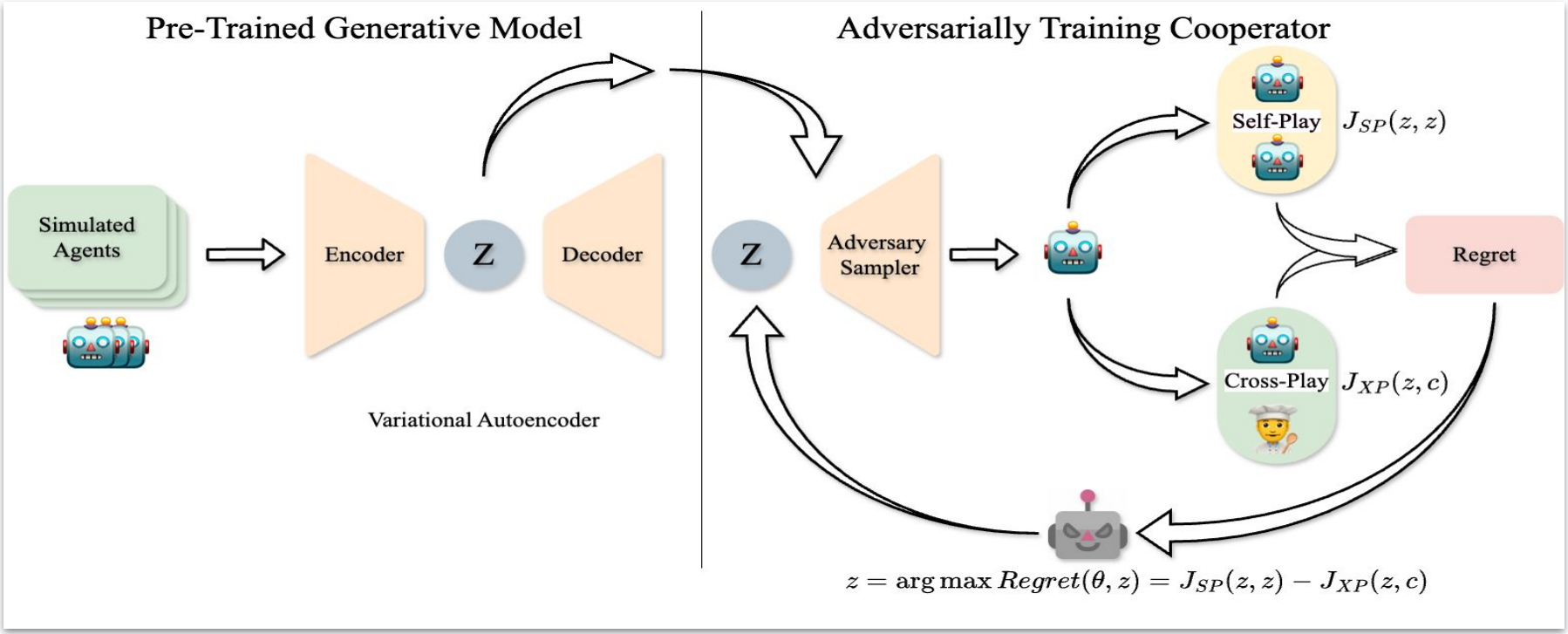


Training Process

1. Train a VAE on simulated Agent policies
2. Freeze the weights of model
3. Sample diverse human-like cooperative agents using standard normal sampling
4. Train the Cooperator agent

Liang et al. (2024)

GOAT: **G**enerative **O**nline **A**dversarial **T**raining

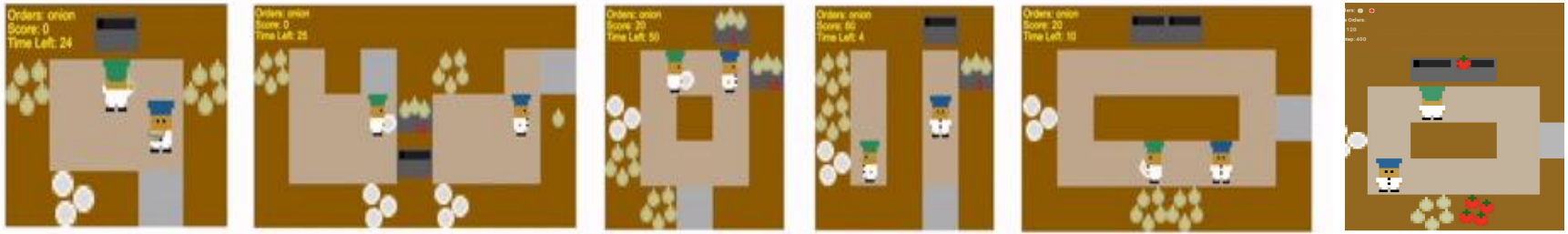


GOAT: **G**enerative **O**nline **A**dversarial **T**raining

Algorithm 1: GOAT

Training VAE on trajectory dataset $D = \tau_{ii=1}^{N}$ where $\tau = s_i, a_i, r_i, s'^{N}_{ii=1}$
Initialize: VAE Decoder $p(a_t|z', \tau 0..t-1; \theta)$, Adversary π_A, Cooperator π_C
while not converged **do**
  Sample $z \sim \mathcal{N}(0, I)$
  Map $z' = \pi_A(z)$
  Generate $\pi_P = p(a_t|z', \tau 0..t-1; \theta)$
  Collect Partner self-play returns $J(\pi_P, \pi_P)$
  Collect Partner-Cooperator returns $J(\pi_P, \pi_C)$
  Compute REGRET $= J(\pi_P, \pi_P) - J(\pi_P, \pi_C)$
  Train Adversary Policy (π_A) with RL to maximize REGRET
  Train Cooperator Policy (π_C) with RL to minimize REGRET
end while

Evaluation Benchmark: Overcooked



To test our method against prior works, we use the **Overcooked** benchmark.

Key Benefits:

- Tests coordination at varying levels of complexity
- Standardized comparison with prior state-of-the-art
- Dynamic multi-agent environments

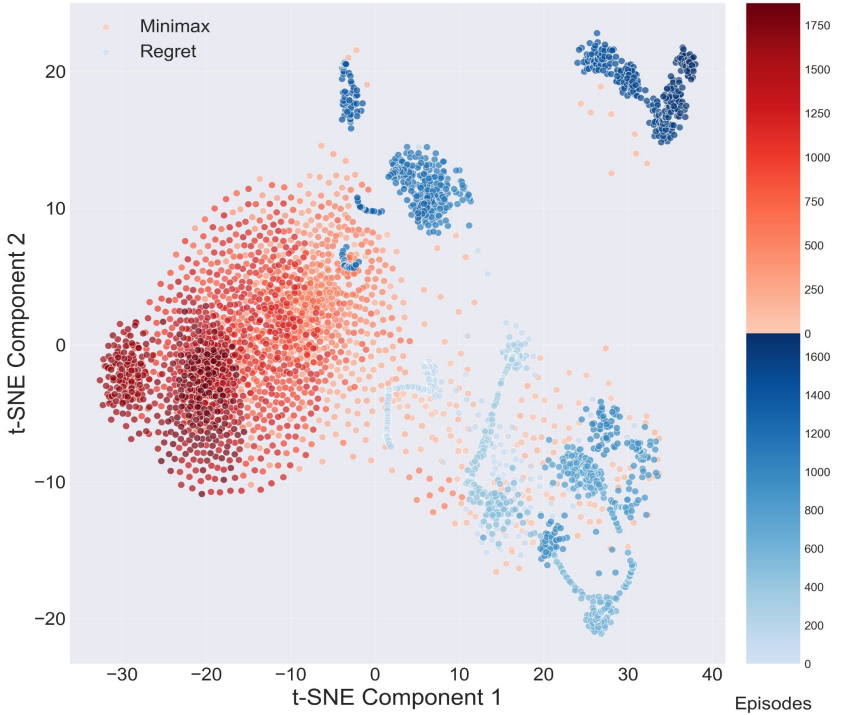
Minimax vs. Regret

Minimax

Stuck in a region where policies do not contribute towards training.

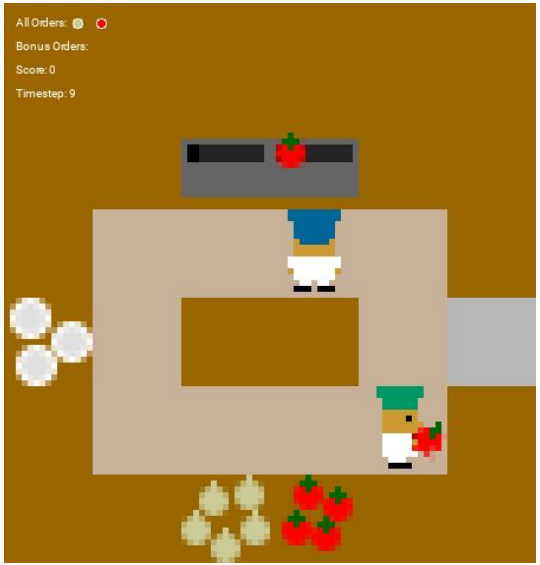
Regret

Allows sampling cooperative policies and natural exploration.



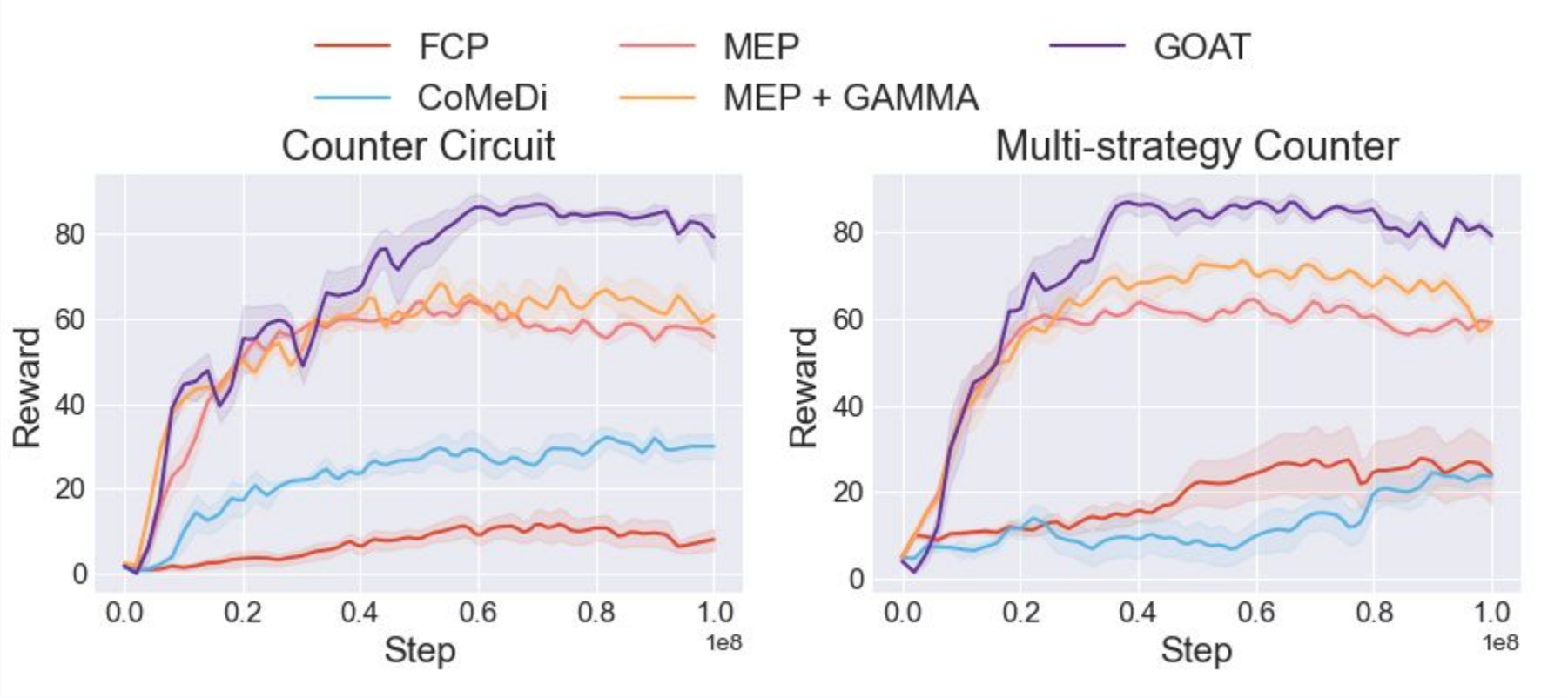
Human Evaluation: Challenging Layouts

Counter Circuit

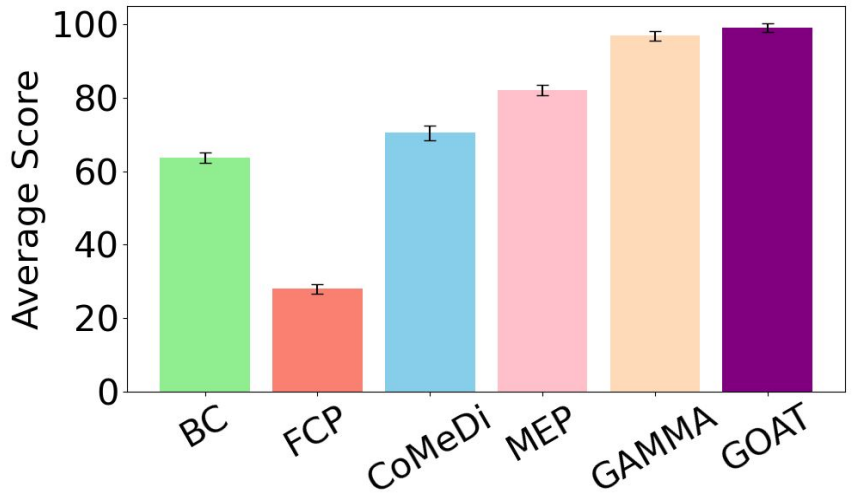


Multi-Strategy Counter

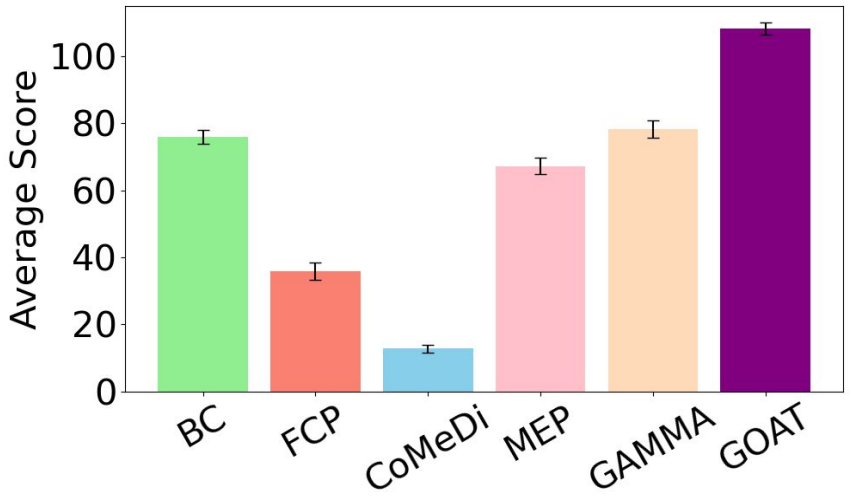
GOAT Against Human Proxy



GOAT Performance vs. Real Humans



Counter Circuit



Multi-Strategy Counter

Conclusion

No Sabotage Behavior

With a regret-based approach, the cooperator agent avoids engaging in sabotage.

Strategic Training

Enables a natural curriculum and more effective strategic training sessions.

Seamless Coordination

GOAT improves coordination and adapts to humans seamlessly in real-time.

Improved Performance

Enhances performance in harder overcooked layouts against real human players.

Thank You!

Improving Human-AI Coordination Through Online Adversarial Training and Generative Models

Paresh Chaudhary, Yancheng Liang, Daphne Chen, Simon Du, Natasha Jaques

pareshrc@cs.washington.edu