Opponent Modeling with In-context Search

Yuheng Jing[1,2], **Bingyun Liu**[1,2], **Kai Li**[1,2,†], **Yifan Zang**[1,2],
Haobo Fu[6], **Qiang Fu**[6], **Junliang Xing**[5], **Jian Cheng**[1,3,4,†]

1 Institute of Automation, Chinese Academy of Sciences
2 School of Artificial Intelligence, University of Chinese Academy of Sciences
3 School of Future Technology, University of Chinese Academy of Sciences
4 AiRiA 5 Tsinghua University 6 Tencent AI Lab

Background

Opponent Modeling (OM) enhances decision-making in multi-agent environments by modeling the behaviors, goals, and etc. of ***other agents*** *(including adversaries and teammates, collectively termed as* ***opponents****).*

Existing OM approaches fall into two categories:

1. Pretraining-Focused Approaches (PFAs): Train models to generalize across diverse opponent policies.

2. Testing-Focused Approaches (TFAs): Update pretrained models at testing time to adapt to unknown opponents through finetuning.

Background

Opponent Modeling (OM) enhances decision-making in multi-agent environments by modeling the behaviors, goals, and etc. of ***other agents*** *(including adversaries and teammates, collectively termed as **opponents**).*

Existing OM approaches fall into two categories:

1. Pretraining-Focused Approaches (PFAs): Train models to generalize across diverse opponent policies.

2. Testing-Focused Approaches (TFAs): Update pretrained models at testing time to adapt to unknown opponents through finetuning.

Problems with Current Approaches:

- **PFAs**: **Limited Generalization** to unknown opponents due to a lack of theoretical analyses and minimal testing-stage adjustments.
- **TFAs**: **Performance Instability** from reliance on small datasets for testing-time gradient updates and sensitivity to hyperparameters.

A unified approach is needed to tackle these challenges.

Contributions

We introduce **Opponent Modeling with In-context Search (OMIS)**, a novel OM approach which leverages **In-Context Learning (ICL)** and **Decision-Time Search (DTS)** to tackle the existing challenges.

The Core Motivation of OMIS:

- We encode the idea of ‘*think before you act*’ through DTS.
- We make DTS *possible* and essentially *adaptive* through ICL.

Contributions

We introduce **Opponent Modeling with In-context Search (OMIS)**, a novel OM approach which leverages **In-Context Learning (ICL)** and **Decision-Time Search (DTS)** to tackle the existing challenges.

The Core Motivation of OMIS:

- We encode the idea of '*think before you act*' through DTS.
- We make DTS *possible* and essentially *adaptive* through ICL.

The Core Innovations of OMIS:

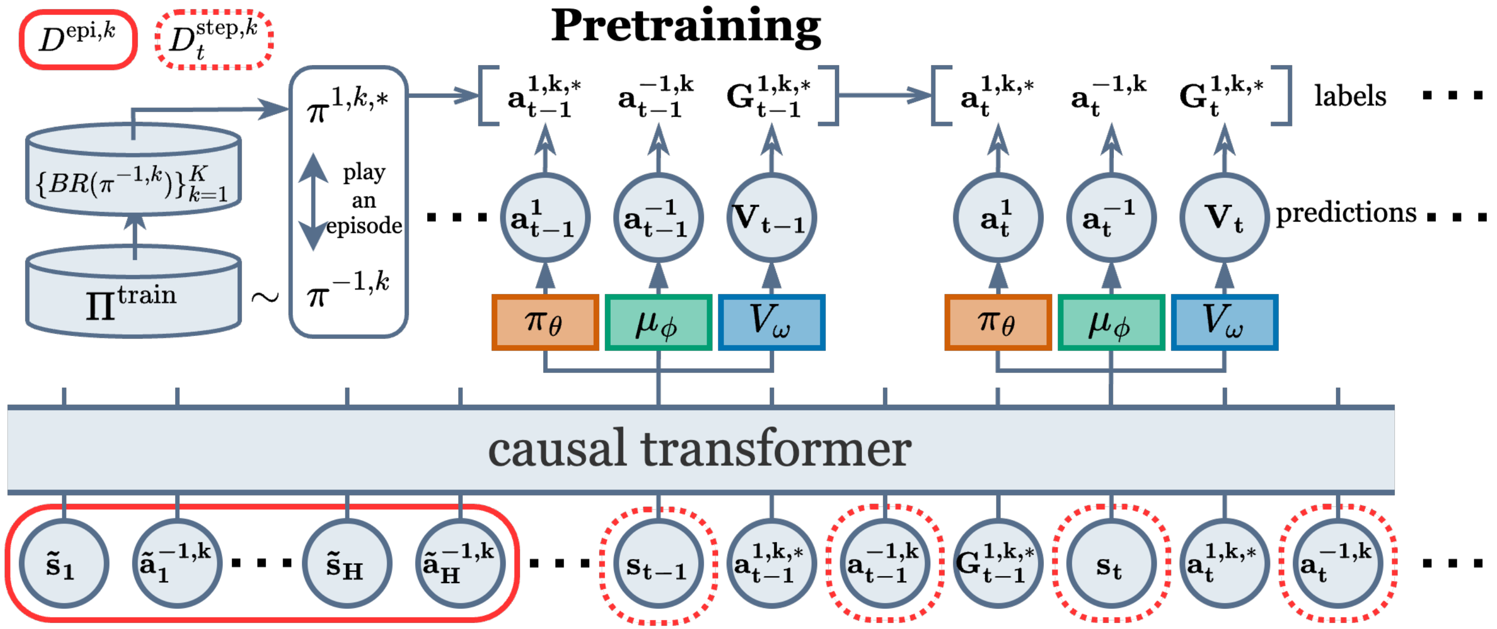
- A *theoretically grounded* **DTS mechanism** to refine the original policy in real-time rather than gradient updates ***to avoid performance instability*** *(main problems of TFAs)*.
- A *theoretically grounded* **ICL-pretrained Transformer** with three in-context components ***to enable DTS and enhance generalization*** *(main problems of PFAs)*:

  (1) **Actor**: Learn to respond appropriately to the current opponent.

  (2) **Opponent Imitator**: Predict the actions of the current opponent.

  (3) **Critic**: Evaluate the values of the states during DTS.

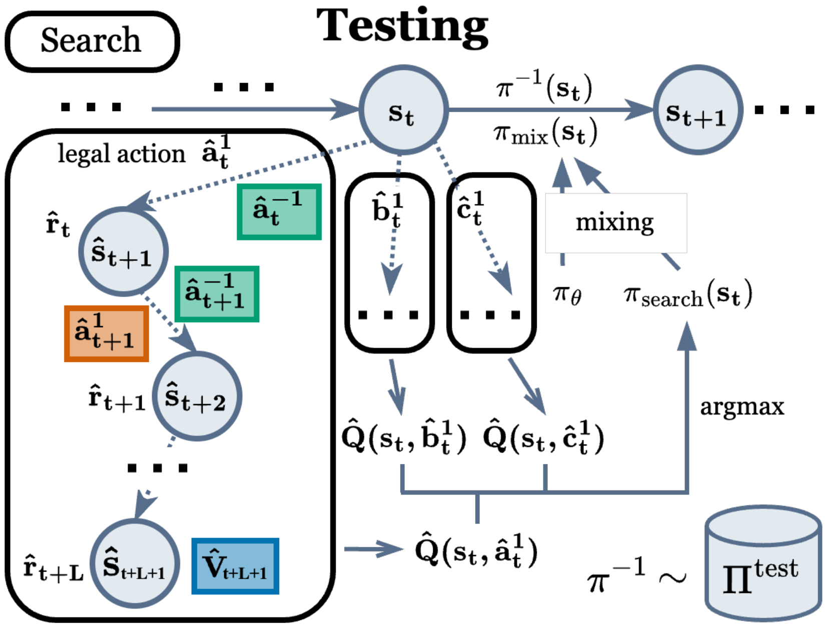
Methodology



$D^{\text{epi},k}$
$D_t^{\text{step},k}$
Pretraining
$\{BR(\pi^{-1,k})\}_{k=1}^{K}$
Π^{train}
$\sim$
$\pi^{1,k,*}$
play an episode
$\pi^{-1,k}$
$[\mathbf{a}_{t-1}^{1,k,*}\ \mathbf{a}_{t-1}^{-1,k}\ \mathbf{G}_{t-1}^{1,k,*}]$
$[\mathbf{a}_{t}^{1,k,*}\ \mathbf{a}_{t}^{-1,k}\ \mathbf{G}_{t}^{1,k,*}]$
labels
$\mathbf{a}_{t-1}^{1}$ $\mathbf{a}_{t-1}^{-1}$ $\mathbf{V}_{t-1}$
$\mathbf{a}_{t}^{1}$ $\mathbf{a}_{t}^{-1}$ $\mathbf{V}_{t}$
predictions
π_θ μ_ϕ V_ω
π_θ μ_ϕ V_ω
causal transformer
$\tilde{\mathbf{s}}_1$ $\tilde{\mathbf{a}}_1^{-1,k}$ $\tilde{\mathbf{s}}_H$ $\tilde{\mathbf{a}}_H^{-1,k}$
$\mathbf{s}_{t-1}$ $\mathbf{a}_{t-1}^{1,k,*}$ $\mathbf{a}_{t-1}^{-1,k}$ $\mathbf{G}_{t-1}^{1,k,*}$
$\mathbf{s}_t$ $\mathbf{a}_t^{1,k,*}$ $\mathbf{a}_t^{-1,k}$

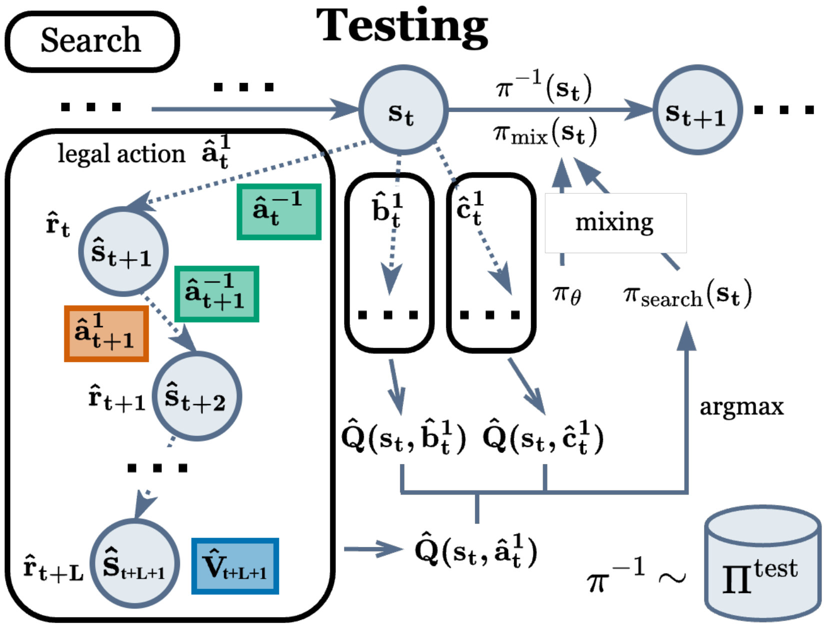
The Pretraining Procedure and Architecture of OMIS. The pretraining steps are as follows: (1) Train *Best Responses* (BRs) against all policies in the *training set of opponent policies* Π^{train}. (2) Continuously sample opponent policy from Π^{train} and collect training data by playing against it using its BR. (3) Train a Transformer model using ICL-based supervised learning, where the model consists of *three in-context components*.

Methodology



The Testing Procedure of OMIS. During testing, OMIS refines the *actor* π_θ through DTS at each timestep. The DTS steps are as follows: (1) Do multiple L-step rollouts for each legal action, where the *actor* π_θ and the *opponent imitator* μ_ϕ are used to simulate actions for the self-agent and opponent, respectively. The *critic* V_ω is used to estimate the value of final search states.

Methodology


Search
Testing
$\mathbf{s_t}$
$\pi^{-1}(\mathbf{s_t})$
$\pi_{\text{mix}}(\mathbf{s_t})$
$\mathbf{s_{t+1}}$
legal action $\hat{\mathbf{a}}_t^1$
$\hat{\mathbf{a}}_t^{-1}$
$\hat{\mathbf{r}}_t$
$\hat{\mathbf{s}}_{t+1}$
$\hat{\mathbf{a}}_{t+1}^{-1}$
$\hat{\mathbf{a}}_{t+1}^{1}$
$\hat{\mathbf{r}}_{t+1}$
$\hat{\mathbf{s}}_{t+2}$
$\hat{\mathbf{r}}_{t+L}$
$\hat{\mathbf{S}}_{t+L+1}$
$\hat{\mathbf{V}}_{t+L+1}$
$\hat{\mathbf{b}}_t^1$
$\hat{\mathbf{c}}_t^1$
mixing
π_θ
$\pi_{\text{search}}(\mathbf{s_t})$
argmax
$\hat{\mathbf{Q}}(\mathbf{s_t}, \hat{\mathbf{b}}_t^1)$
$\hat{\mathbf{Q}}(\mathbf{s_t}, \hat{\mathbf{c}}_t^1)$
$\hat{\mathbf{Q}}(\mathbf{s_t}, \hat{\mathbf{a}}_t^1)$
$\pi^{-1} \sim$
Π^{test}

The DTS steps are as follows: (1) Do multiple L-step rollouts … (2) Estimate the *DTS value* $\hat{Q}$ for all legal actions, and the *search policy* π_{search} selects the legal action with the maximum $\hat{Q}$. (3) Use *mixing technique* to trade-off between the *search policy* π_{search} and the *actor* π_θ to choose the real action to be executed.

Theoretical Results

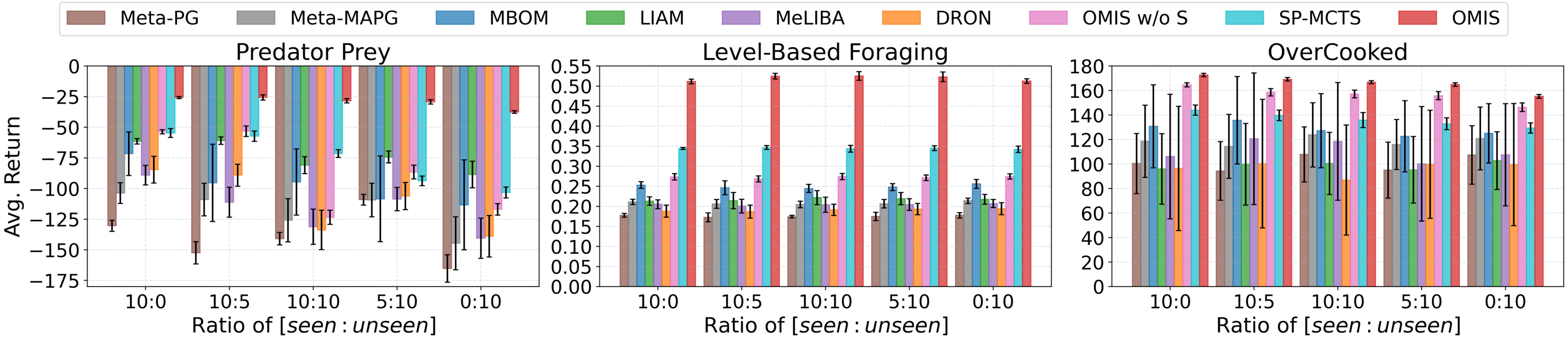
- **Generalization Guarantee of OMIS w/o DTS** (Supported by **Lemma 4.1** and **Theorem 4.2**): OMIS w/o DTS can accurately recognize *seen opponent policies (policies in the training set of opponent policies* Π^{train}*)* and recognize *unseen opponent policies (policies not in* Π^{train}*)* as the most familiar seen ones with the measurement of Kullback-Leibler Divergence.

Theoretical Results

- **Generalization Guarantee of OMIS w/o DTS** (Supported by **Lemma 4.1** and **Theorem 4.2**): OMIS w/o DTS can accurately recognize *seen opponent policies (policies in the training set of opponent policies Π^{train})* and recognize *unseen opponent policies (policies not in Π^{train})* as the most familiar seen ones with the measurement of Kullback-Leibler Divergence.

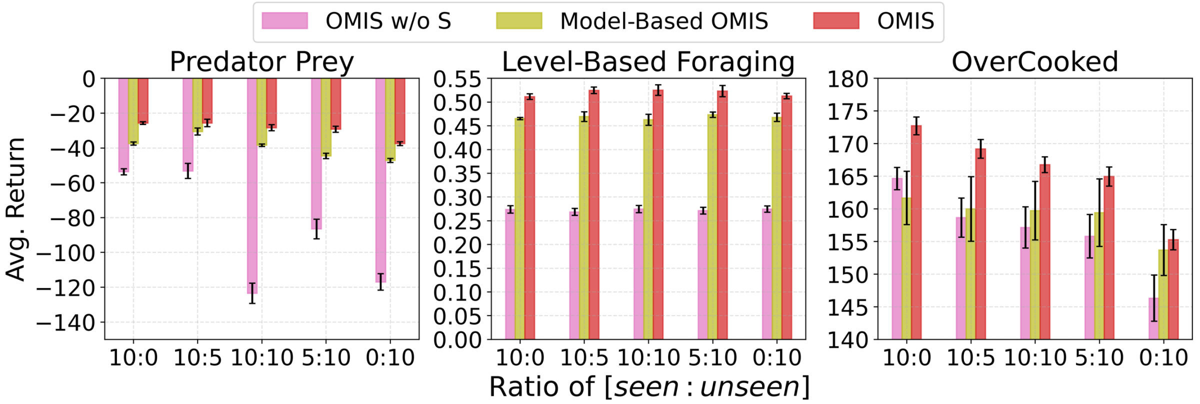
- **Policy Improvement of OMIS w/ DTS** (Supported by **Theorem 4.3**): OMIS w/ DTS avoids any gradient updates and theoretically provides improvement guarantees.

Experimental Results



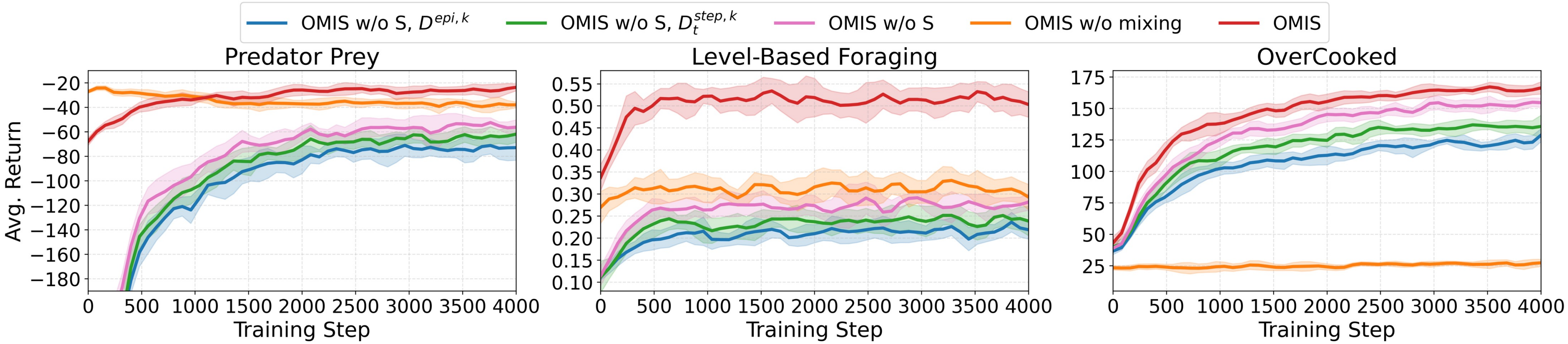
Main Results: OMIS effectively outperforms *PFAs*, *TFAs*, and *DTS-based* OM baselines when testing against *unknown non-stationary opponent agents* in competitive, cooperative and mixed environments.

Experimental Results



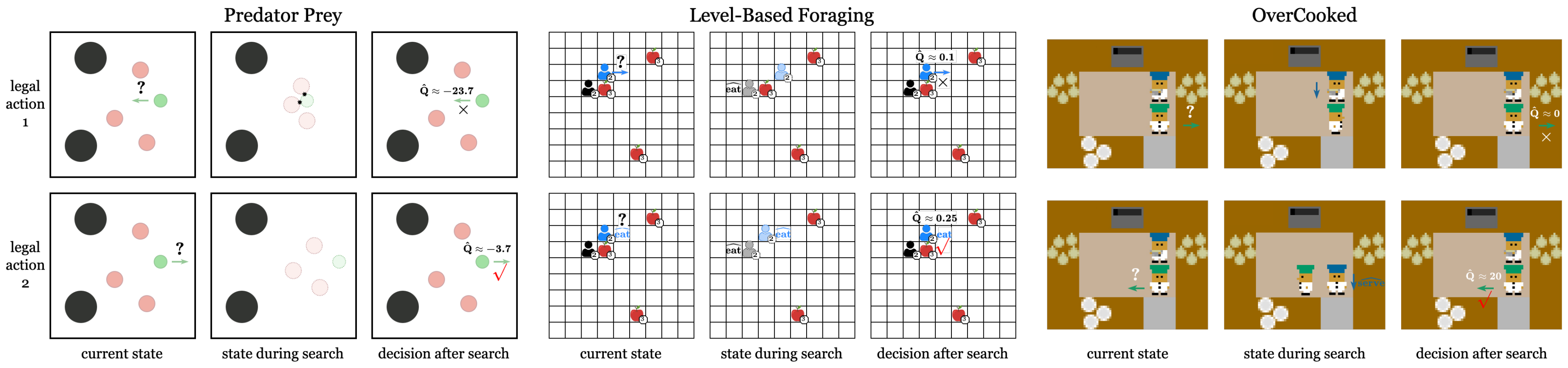
Ablations: OMIS can work effectively when the *transition dynamics* are *unknown* and *learned* instead.

Experimental Results



Ablations: Each designed component of OMIS plays a positive role in boosting the performance.

Experimental Results



Visualizations: When playing against opponents with a previously unseen policy, the DTS of OMIS promptly evaluates each legal action, predicts the opponent's actions and the state values during the DTS process, and ultimately selects the most advantageous action for the self-agent.

CASIA

NEURAL INFORMATION
PROCESSING SYSTEMS

Thank You!

Project Website

Paper

Code