Offline Opponent Modeling with Truncated Q-driven Instant Policy Refinement

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

1 C^{2}DL, 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

Offline Opponent Modeling (OOM)

OOM aims to learn an agent that can dynamically adapt to opponents using only pre-collected, **offline datasets**. This paradigm enhances *practicality* and *efficiency* by removing the dependency on online interaction with the environment and opponents during learning stages.

Background

Offline Opponent Modeling (OOM)

OOM aims to learn an agent that can dynamically adapt to opponents using only pre-collected, **offline datasets**. This paradigm enhances *practicality* and *efficiency* by removing the dependency on online interaction with the environment and opponents during learning stages.

The Problem with Suboptimal Data

Previous OOM work assumes datasets are **optimal** (i.e., the agent plays a Best Response). This is often unrealistic, as real-world data is frequently **suboptimal**. When trained on suboptimal data, the performance of existing OOM algorithms deteriorates dramatically.

Contributions

Main Challenges

- Learning a *workable Q-function* in OOM is highly challenging. Key issues include:

  (1) **Complexity**: The added dimensions and complexity of modeling opponents' actions.

  (2) **Non-stationarity**: The unreliability of Q-estimates as opponents can switch policies during testing.

- Standard *Offline Conservative Learning* (OCL) is ineffective for OOM due to severe **distributional shifts** between offline training and testing with unseen opponents.

Contributions

Main Challenges

- Learning a *workable Q-function* in OOM is highly challenging. Key issues include:
  (1) **Complexity**: The added dimensions and complexity of modeling opponents' actions.
  (2) **Non-stationarity**: The unreliability of Q-estimates as opponents can switch policies during testing.
- Standard *Offline Conservative Learning* (OCL) is ineffective for OOM due to severe **distributional shifts** between offline training and testing with unseen opponents.

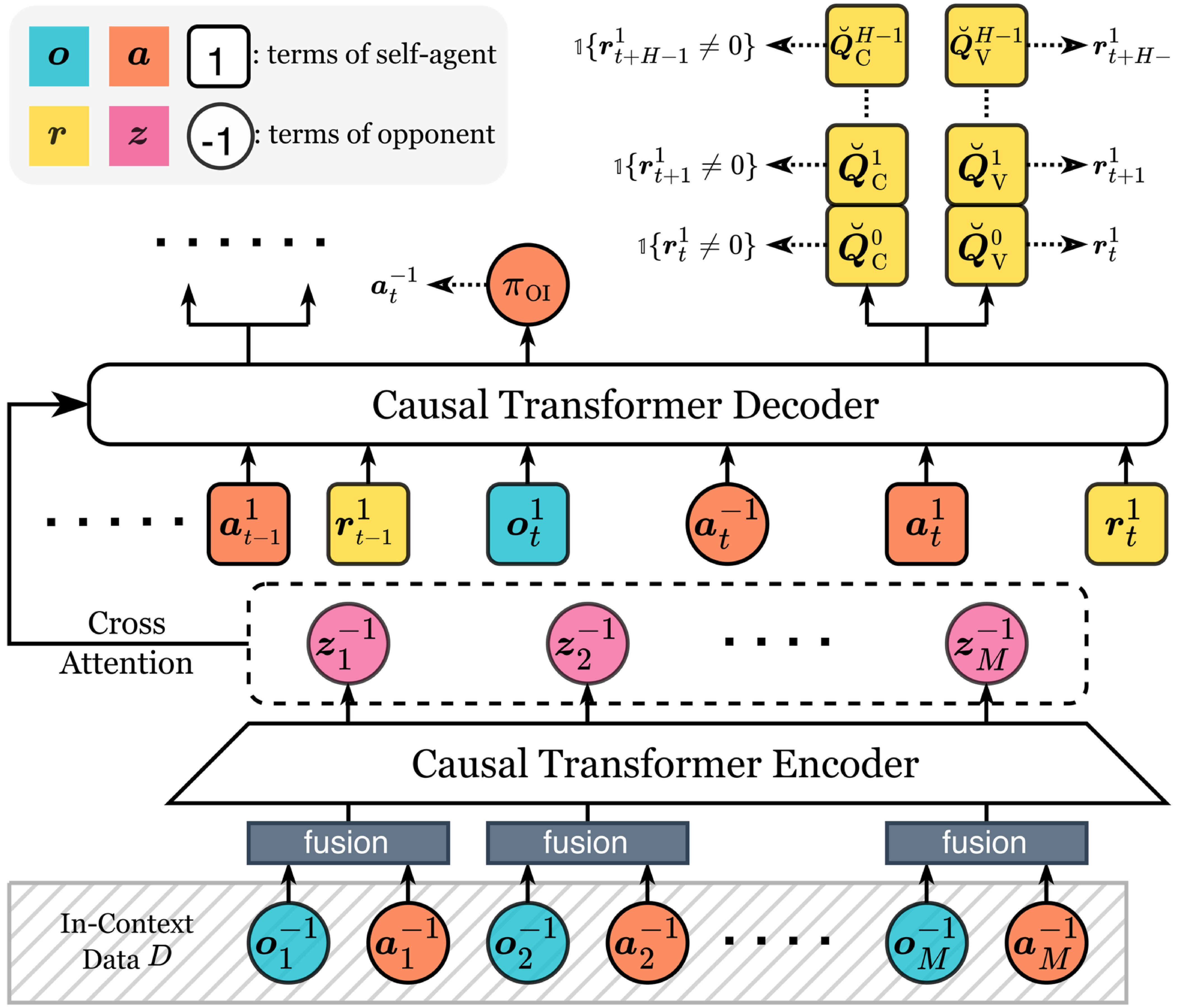
Our Solutions

- Propose **Truncated Q-driven Instant Policy Refinement (TIPR)**, a simple, plug-and-play framework to handle suboptimal datasets in OOM.
- Introduce **Truncated Q**, a horizon-truncated action-value function, and **Instant Policy Refinement (IPR)** for test-time policy improvement.
- Provide theoretical justification for Truncated Q via No Maximization Bias probability analysis.

Methodology

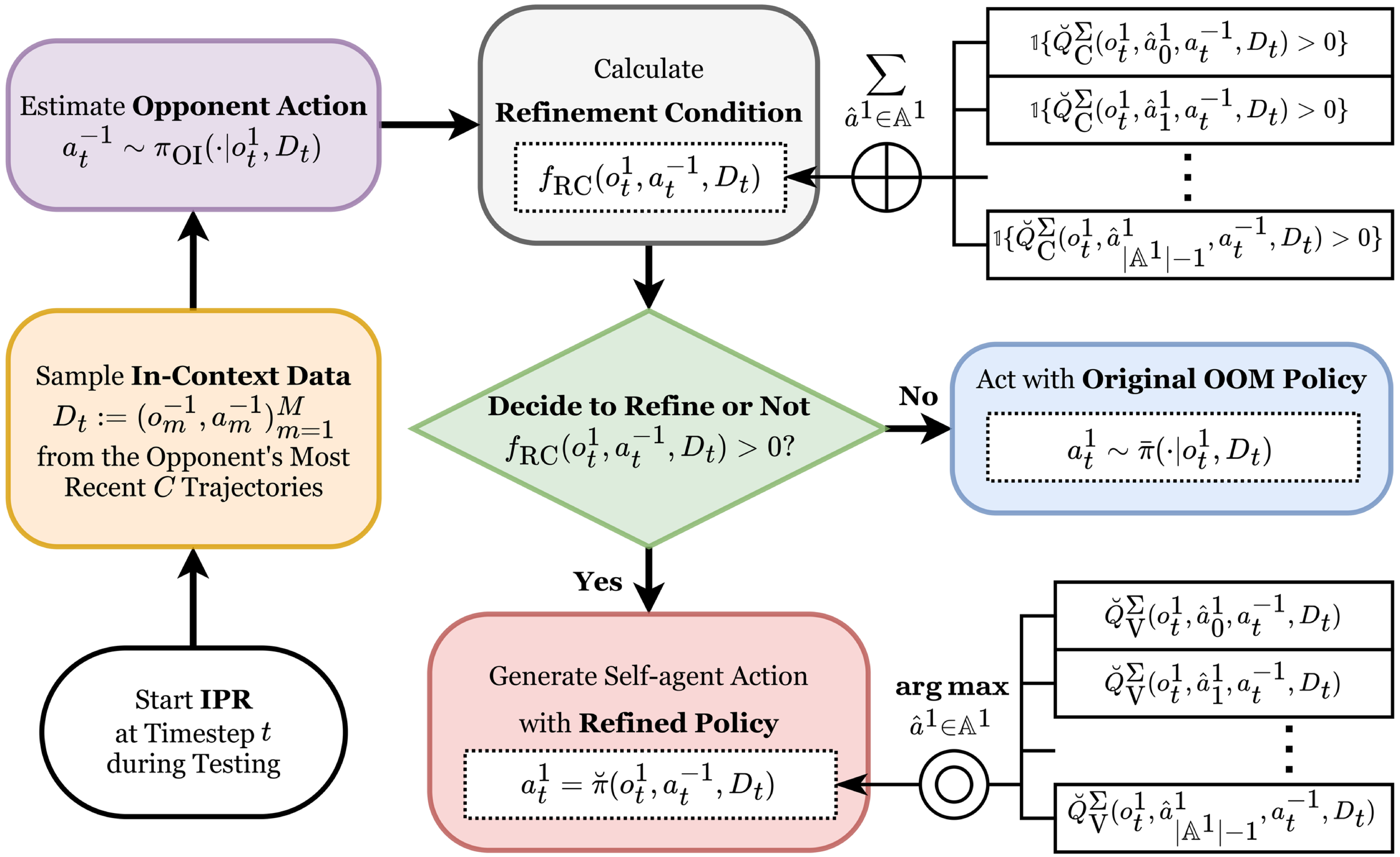
TIPR is a plug-and-play framework that adds two steps to existing OOM algorithms: (1) **Truncated Q Training**: Learn a horizon-truncated, in-context Q-function from the offline dataset. (2) **Instant Policy Refinement (IPR)**: Use the Truncated Q at test-time to decide when and how to refine the agent's policy.

Methodology



Truncated Q is designed to be more learnable and reliable: (1) **Truncated Horizon**: It estimates returns over a shorter, fixed horizon H to reduce cumulative error and learning difficulty. (2) **In-Context Conditioning**: It conditions on opponent data (D) to provide reliable estimates even when opponents are non-stationary.

Methodology



During testing, IPR decides whether to refine the policy at each step: (1) It calculates a **Refinement Condition (RC)** based on Truncated Q's estimated confidence. (2) If the RC is met, IPR generates a refined action by maximizing Truncated Q's estimated value. (3) Otherwise, it defaults to the original OOM policy's action.

Theoretical Results

We justify Truncated Q by analyzing the **No Maximization Bias (NMB) Probability** $y(\mathrm{h}) \coloneqq P(\arg\max_{a^1} \breve{Q}_\mathrm{h} = \arg\max_{a^1} \mathbb{E}\breve{G}_T)$, which is the probability that the learned Q-function selects the truly optimal action.

Theoretical Results

We justify Truncated Q by analyzing the **No Maximization Bias (NMB) Probability** $y(\mathrm{h}) \coloneqq P(\arg\max_{a^1} \breve{Q}_{\mathrm{h}} = \arg\max_{a^1} \mathbb{E}\breve{G}_T)$, which is the probability that the learned Q-function selects the truly optimal action.

This probability is lower-bounded by $y(\mathrm{h}) \geq f(\mathrm{h})g(\mathrm{h})$.

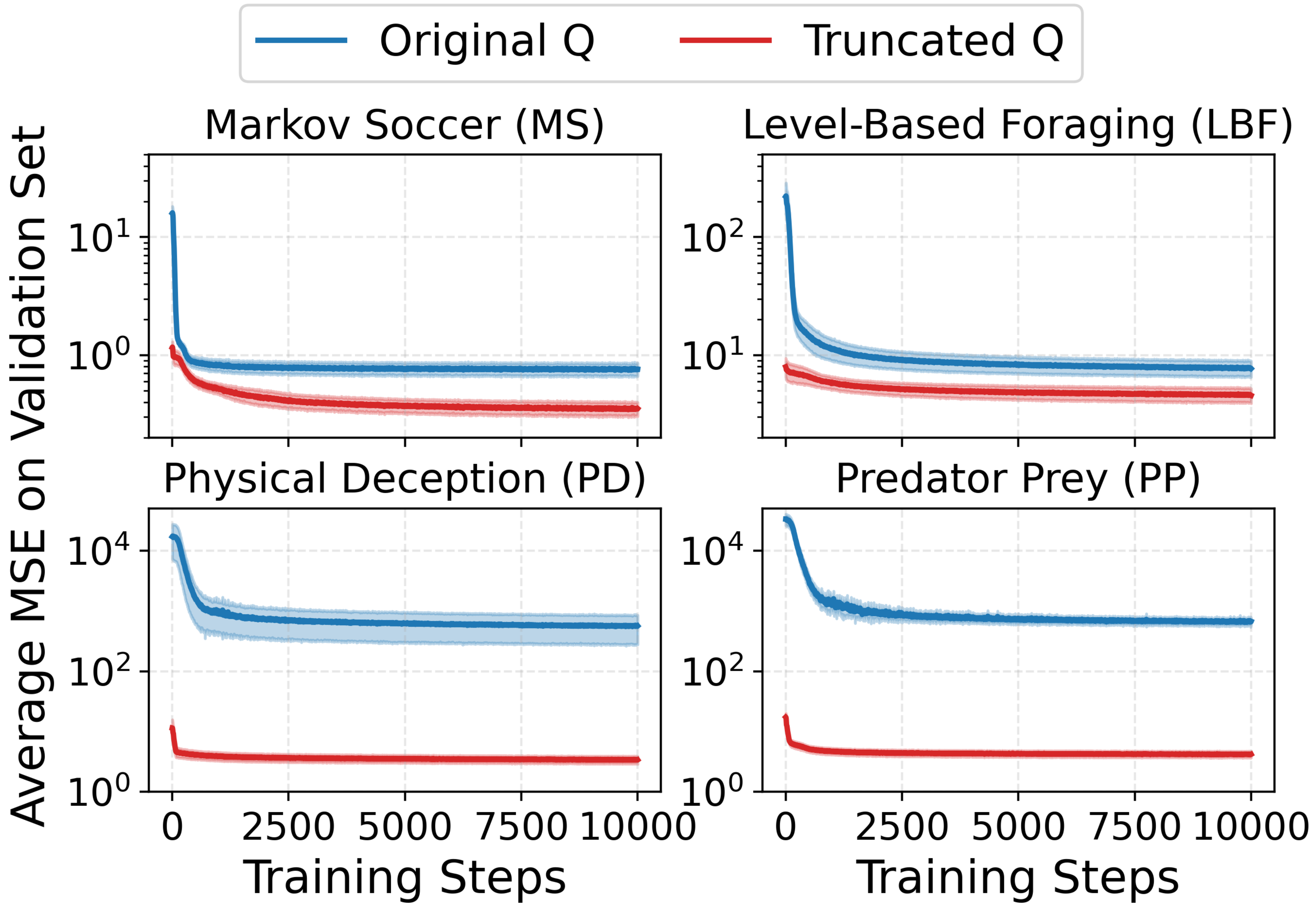
- $f(\mathrm{h})$: Empirical Risk NMB Probability, decided by *model's fitting ability* (↓ as horizon h ↑).
- $g(\mathrm{h})$: Natural NMB Probability, related to *environment's reward structure* (↑ as h ↑).

This shows a **trade-off**, implying **an optimal truncated horizon $\mathbf{h}^* \in [\mathbf{1}, \boldsymbol{T}]$ guarantees to exist** that maximizes the bound.

Experimental Results

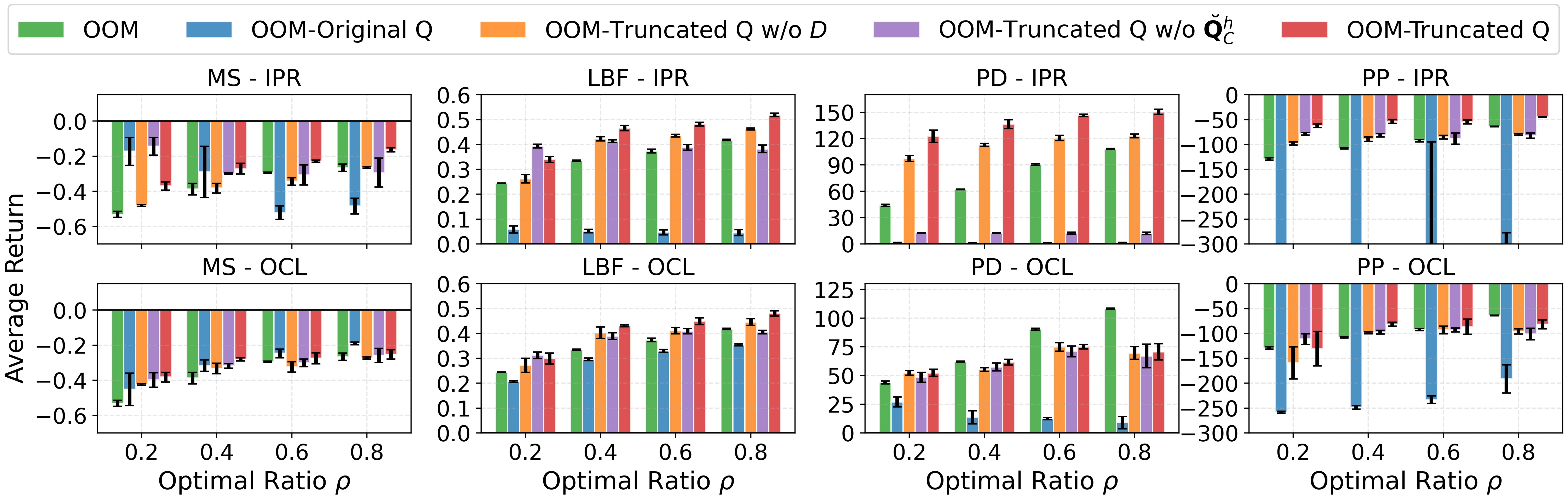
Main Results: (1) When pretrained on suboptimal data, all OOM baselines suffer significant performance loss. (2) TIPR provides stable and considerable improvements across all tested algorithms and dataset qualities.

Experimental Results



Ablation I: Shortening the horizon over which Q-function estimates the expected return can significantly reduce the learning difficulty.

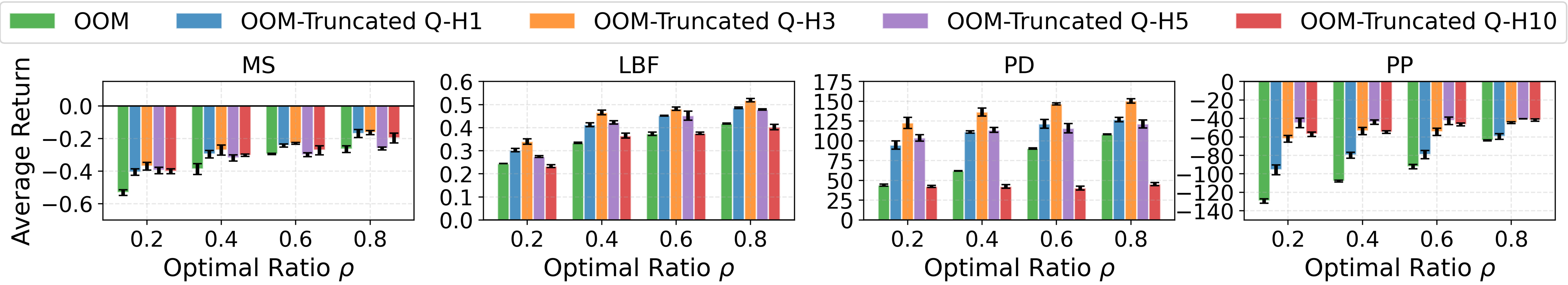
Experimental Results



Ablation II: (1) Our IPR method is more effective than standard OCL, which can degrade performance due to distributional shifts. (2) Using Truncated Q leads to better policy improvement than using a full-horizon Original Q, which often fails catastrophically.

Experimental Results

ICML
International Conference
On Machine Learning



Ablation III: The choice of the horizon H is a tunable parameter. An optimal H exists for different environments; making H too large H can be detrimental, approaching the poor performance of the Original Q.

Thank You for Watching!