arXiv

GitHub

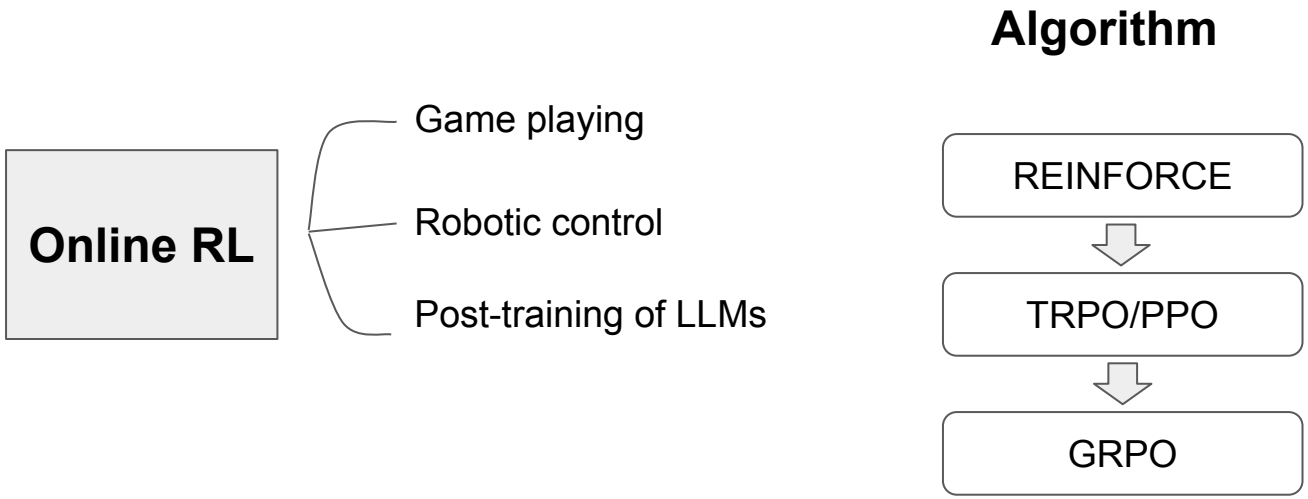
A Snapshot of Influence: A Local Data Attribution Framework for Online Reinforcement Learning

Yuzheng Hu*, Fan Wu*, Haotian Ye, David Forsyth, James Zou, Nan Jiang, Jiaqi W. Ma†, and Han Zhao†

*Equal contribution †Equal advising

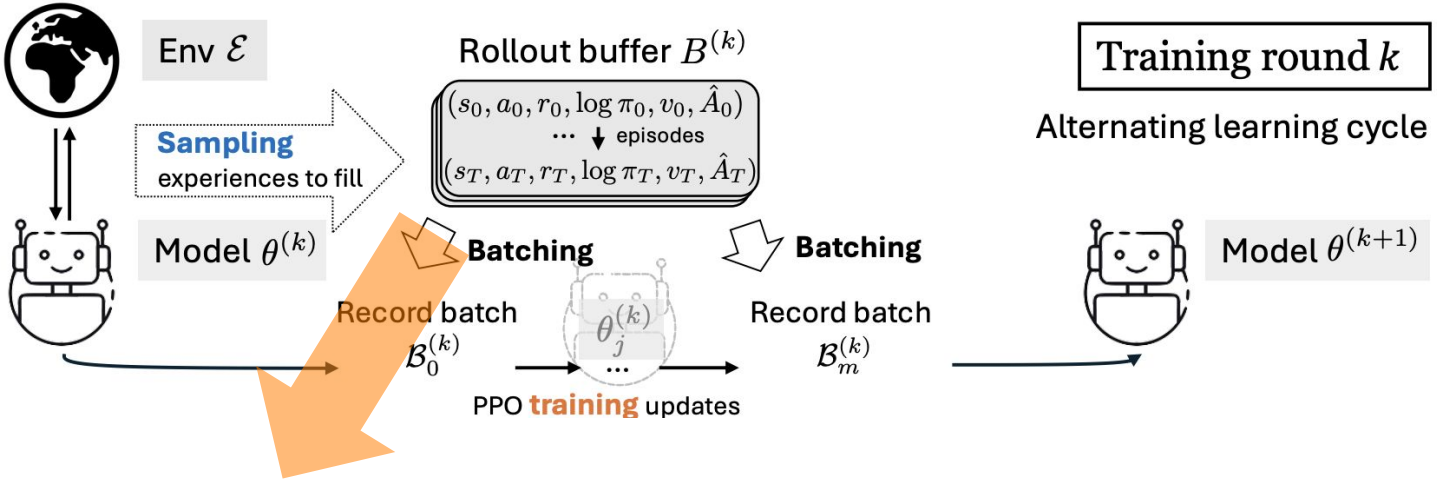
NeurIPS 2025 **oral**

Motivation



Can better data, not just better algorithms, improve online RL?

Background of online RL



Self-generated data: likely imperfect

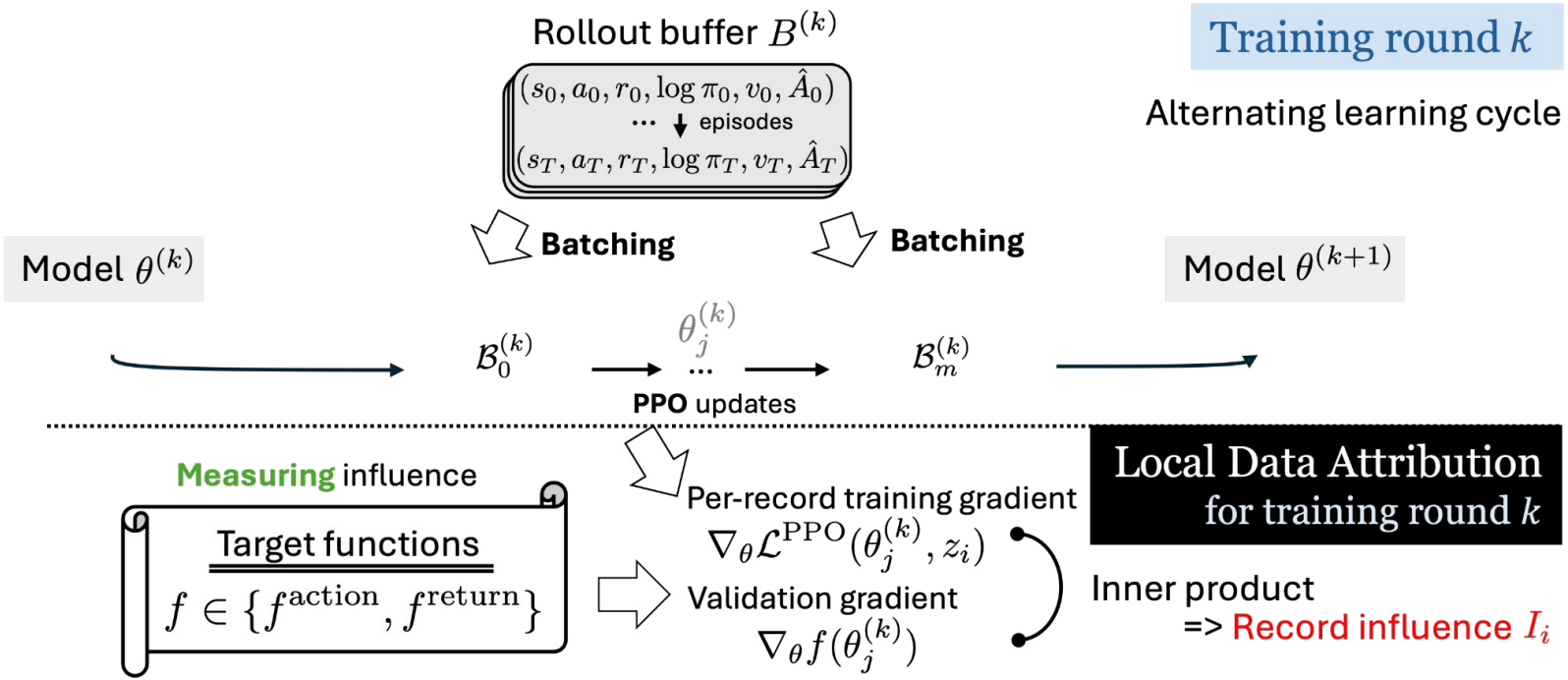
Data attribution

➔ Can we build a principled understanding of data in online RL?

Proposed framework

*"For the agent's **current stage of training**, which experience is most helpful or harmful?"*

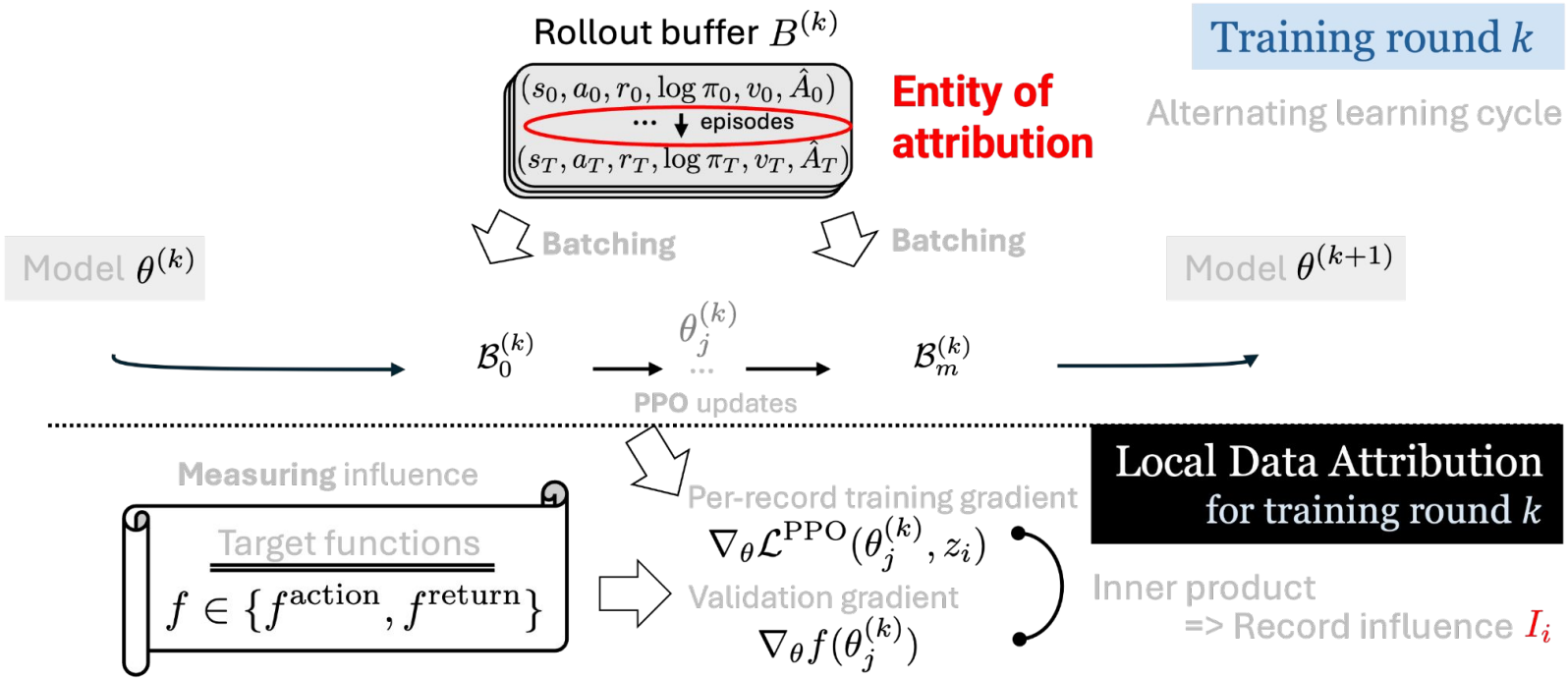
The first data attribution framework for online RL



Proposed framework

*"For the agent's **current stage of training**, which experience is most helpful or harmful?"*

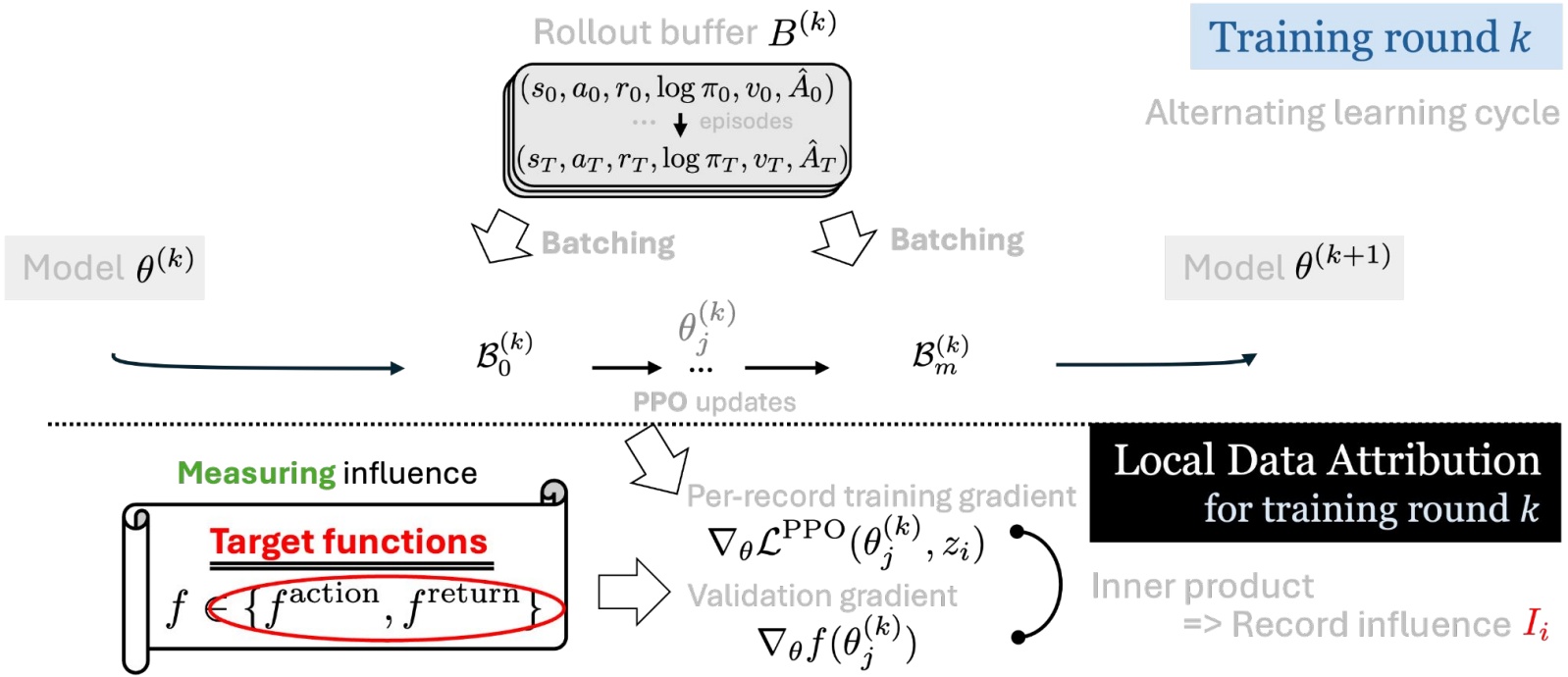
The first data attribution framework for online RL



Proposed framework

*"For the agent's **current stage of training**, which experience is most helpful or harmful?"*

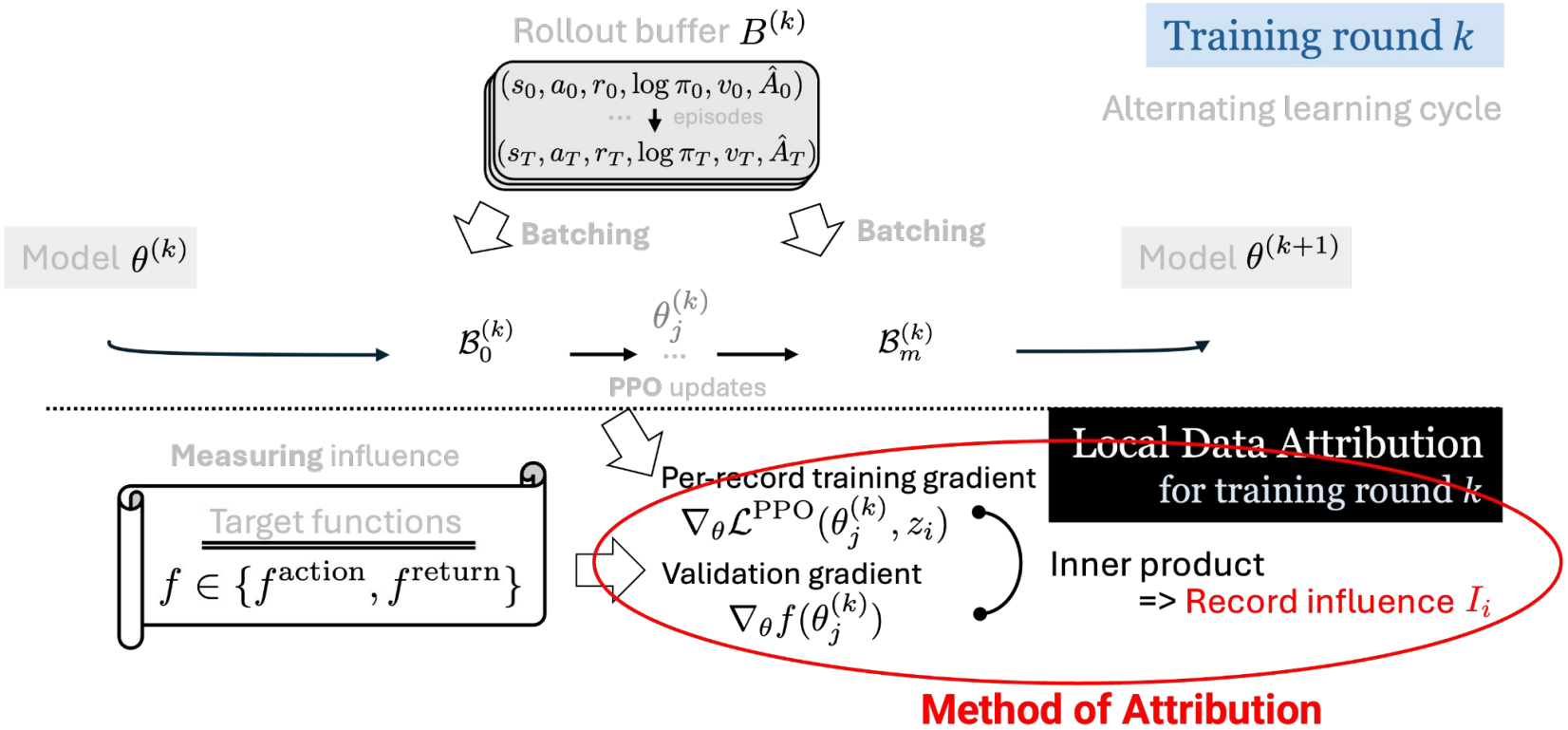
The first data attribution framework for online RL



Proposed framework

*"For the agent's **current stage of training**, which experience is most helpful or harmful?"*

The first data attribution framework for online RL



Interpretation of influence scores

Gradient of the target function represents the direction of *progress*

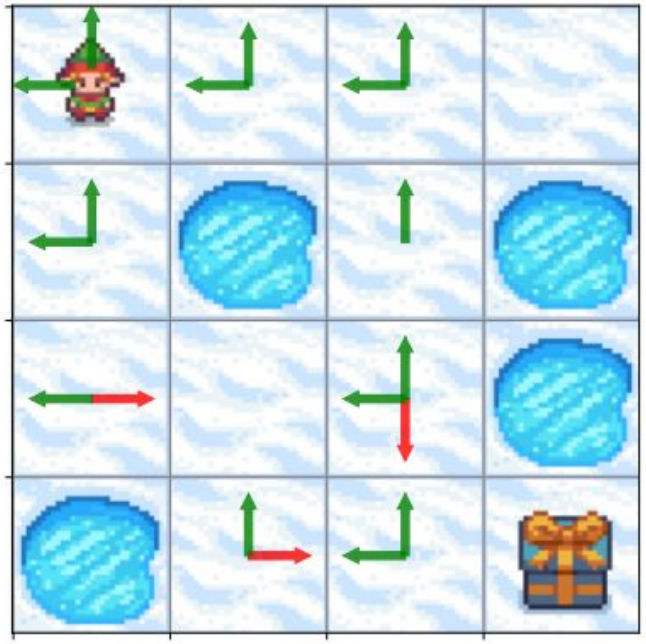
- If a sample's loss gradient aligns with this direction, it's **beneficial** for improving this target function
- If it points in the opposite direction, it's **harmful**

Diagnosis of bottom records

Bottom records feature

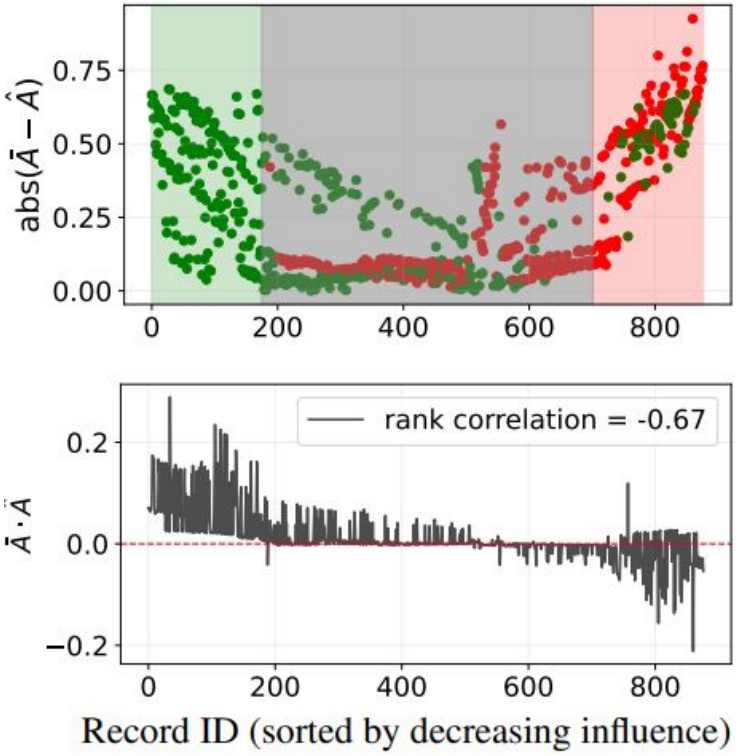
inaccurate advantage estimate

(a) harmful records in `FrozenLake`



Red arrow: neg adv estimate
Green arrow: pos adv estimate

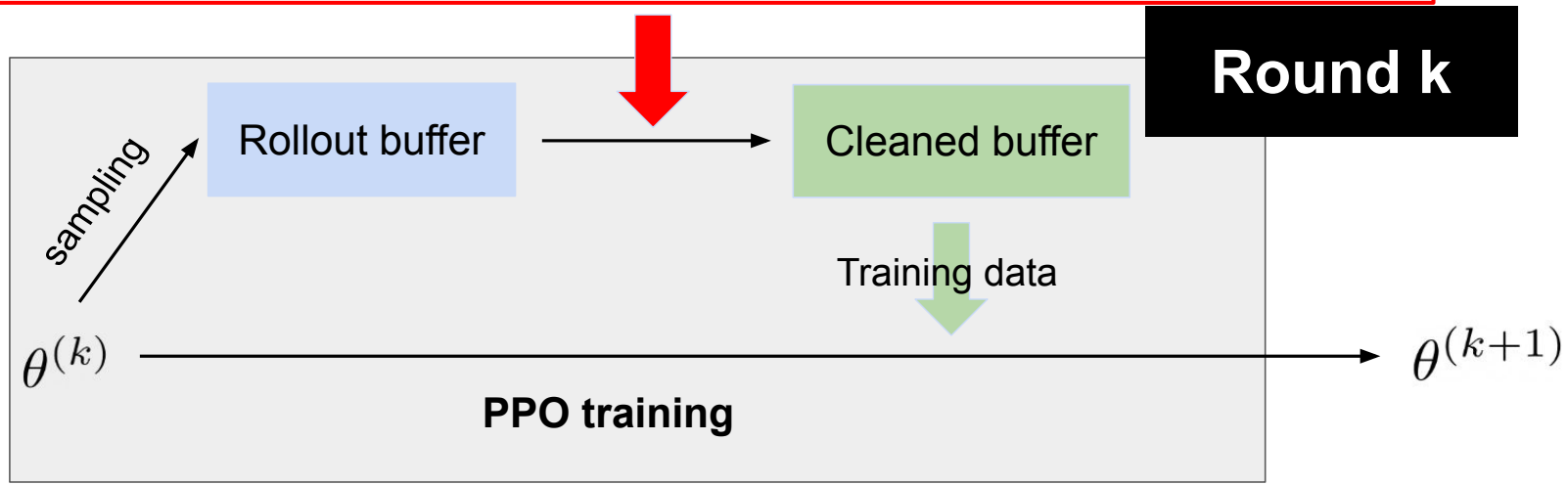
(c-d) analysis in `FrozenLake`



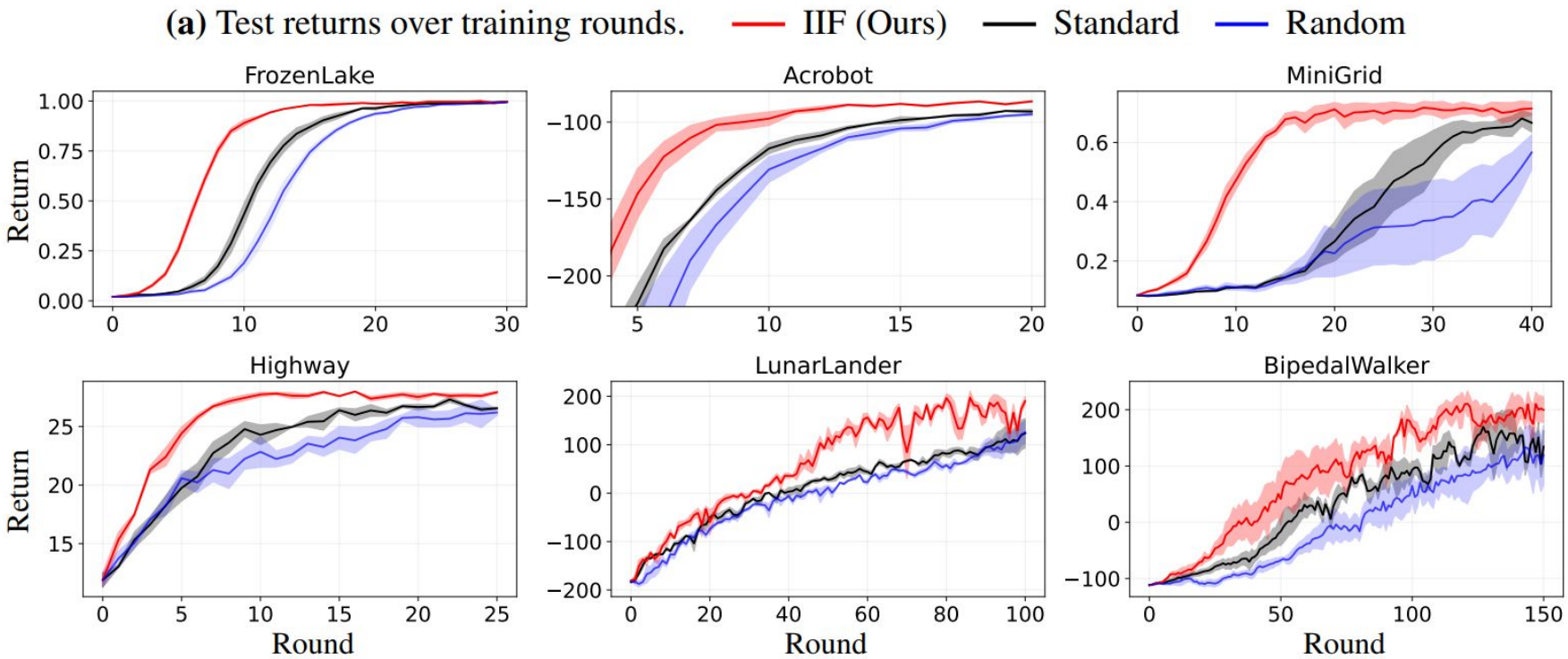
$\bar{A}$: MC estimate
Red point: diff sign
Green point: same

IIF: Iterative Influence-Based Filtering

1. **Attribution**: Assign a score to each record in the rollout buffer
2. **Filtering**: Remove bottom p% of the records with a negative score

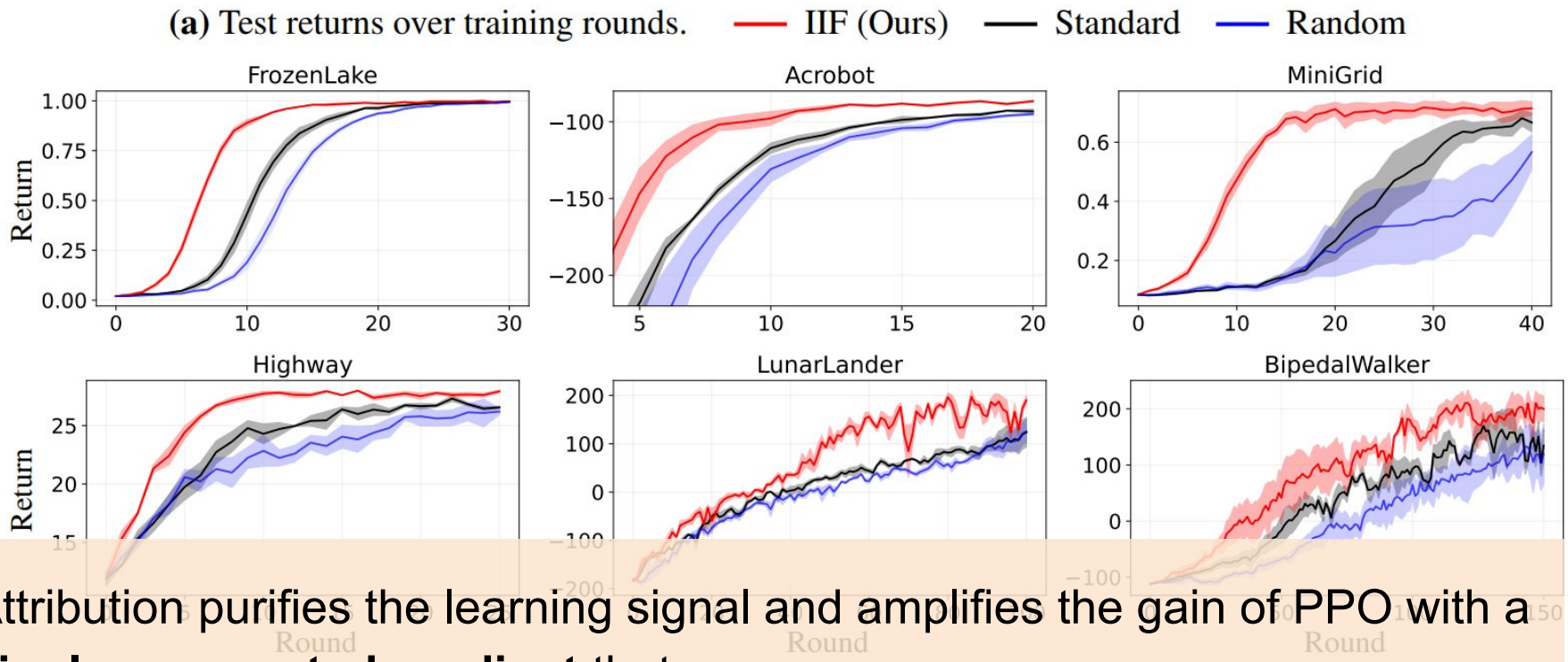


Results (traditional RL)



Reduced (per-round) runtime, improved sample efficiency, higher returns

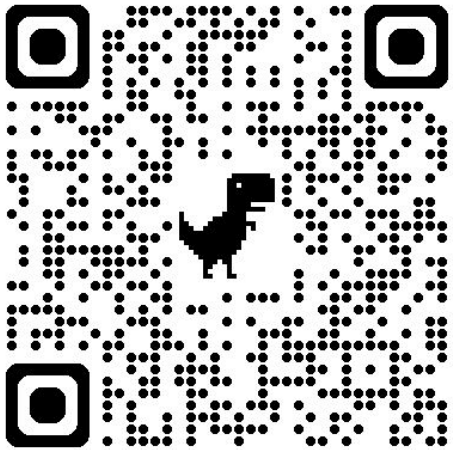
Results (traditional RL)



Attribution purifies the learning signal and amplifies the gain of PPO with a **single aggregated gradient** that measures progress

Thanks!

arXiv



GitHub