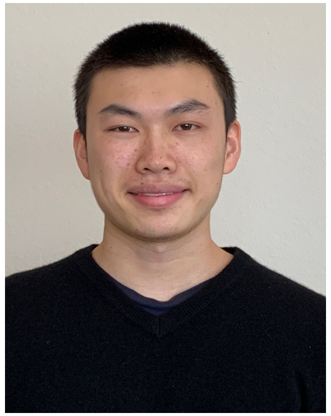
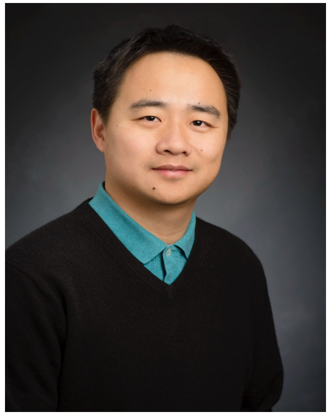
Hindsight Foresight Relabeling for Meta-Reinforcement Learning



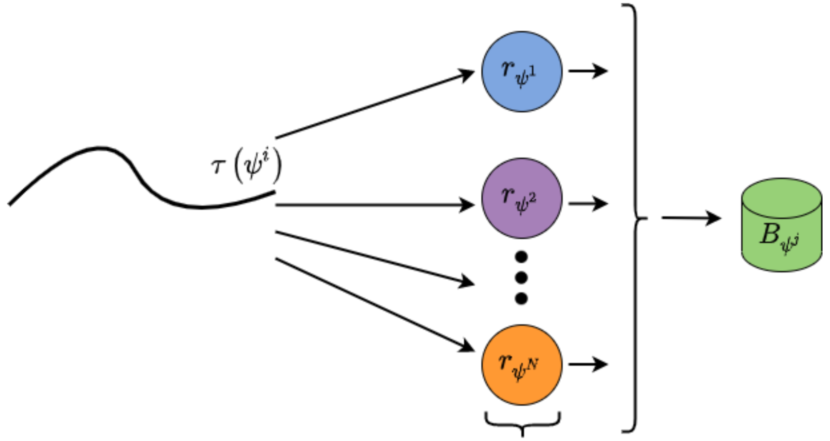


Michael Wan
Jian Peng
Tanmay Gangwani

International Conference on Learning Representations 2022

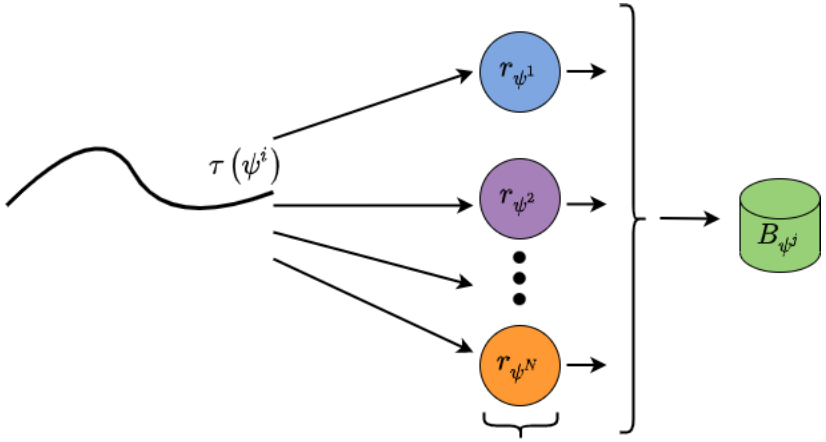
Motivation

- Prior work in multi-task RL has successfully shared data among tasks through <u>reward relabeling</u>
  - Assumption: Transition dynamics same across tasks, reward functions differ
- Meta-RL also involves training on a distribution of tasks
  - So we should also be able to share data between tasks in Meta-RL where transition dynamics remain the same across tasks



Prior Work in Relabeling for Multi-Task RL

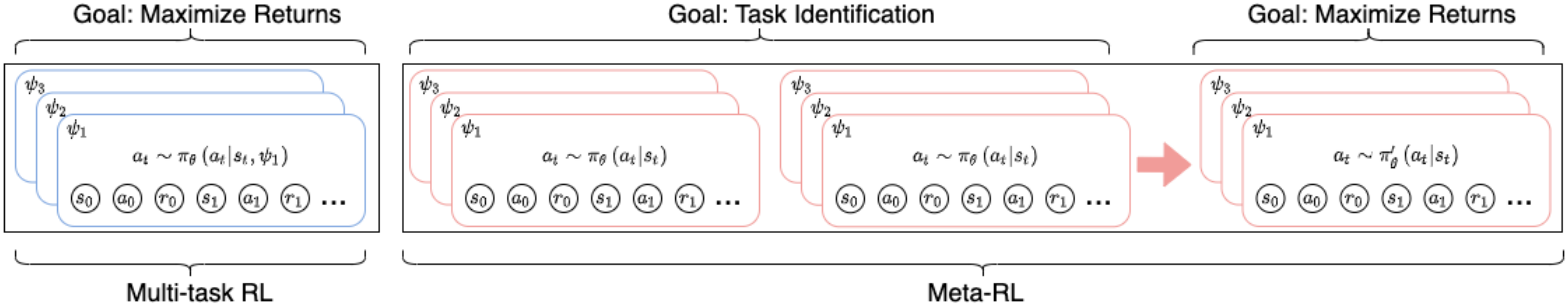
- Prior work [1, 2] has relabeled trajectories based on total return achieved
- Trajectory τ collected for task ψ^i can be used to learn task ψ^j if return of τ under ψ^j is high
  - τ is then relabeled using r_{ψ^j} and added to task buffer B_{ψ^j}



[1] Rewriting History with Inverse RL: Hindsight Inference for Policy Improvement, Eysenbach et al.
[2] Generalized Hindsight for Reinforcement Learning, Li et al.

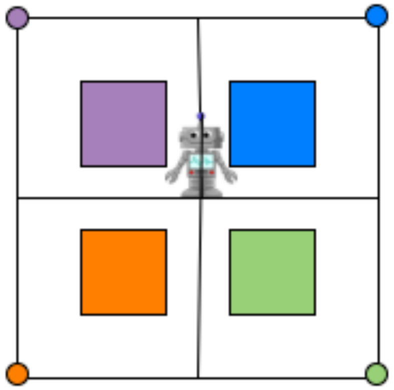
Meta-RL vs Multi-Task RL

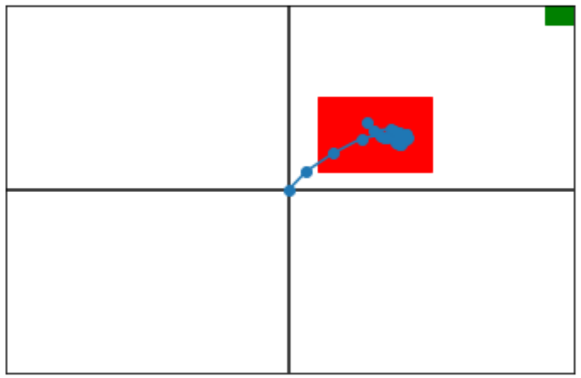
- Multi-Task RL: given task ψ, RL agent seeks to maximize its returns
  - $max_{\theta}\mathbb{E}_{\psi \sim p(\psi), s_t, a_t \sim \pi_\theta}[\sum_{t=1}^{\infty}\gamma^{t-1} r_\psi(s_t, a_t)]$
- Meta-RL: agent must first identify task, then maximize returns
  - Adaptation procedure f_ϕ, post-adaptation policy $\pi'(\theta, \phi)$
  - $max_{\theta,\phi}\mathbb{E}_{\psi \sim p(\psi), s_t, a_t \sim \pi'(\theta,\phi)}[\sum_{t=1}^{\infty}\gamma^{t-1} r_\psi(s_t, a_t)]; \pi'(\theta, \phi) = f_\phi(\pi_\theta, \tau_{pre}, r_\psi)$
- Relabeling based on returns may be sub-optimal for Meta-RL

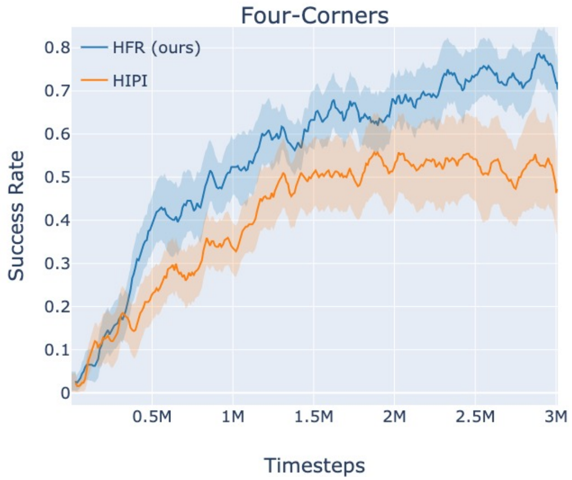


A Didactic Example

- Four-Corners Environment
  - Goal locations in each of the four corners correspond to tasks
  - For each goal, there is a section in the corresponding quadrant that gives large negative reward
- Consider a trajectory that hovers over the blue square
  - Clearly useful for task identification for the blue task
  - Highly negative return under the blue task, so relabeling methods based on returns will not assign the trajectory to the blue (correct) task
  - **HFR** correctly assigns the trajectory to the blue task







Relabeling for Meta-RL (HFR)

- We use **HFR** to relabel based on *post-adaptation returns* (utility)
  - Aligns with Meta-RL objective
  - $U_\psi(\tau) = \mathbb{E}_{s_t, a_t \sim \pi'(\theta,\phi)}[\sum_{t=1}^{\infty} \gamma^{t-1} r_\psi(s_t, a_t)]; \pi'(\theta, \phi) = f_\phi(\pi_\theta, \tau, r_\psi)$
- *Hindsight*: Relabel trajectory using reward functions for different training tasks
- *Foresight*: Compute expected post-adaptation returns (utility) after using the relabeled trajectory for adaptation for different tasks

Relabeling for Meta-RL (HFR)

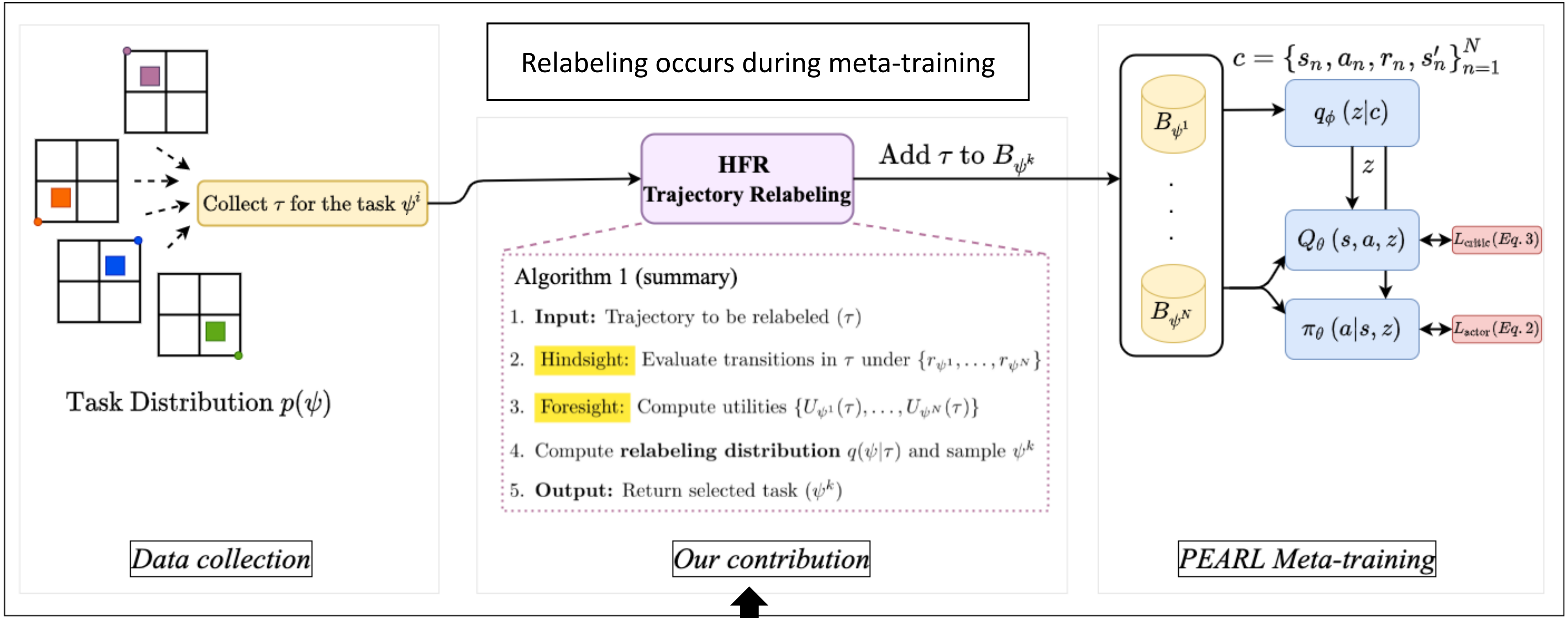
- Trajectory τ more likely to be assigned to tasks for which it has higher (normalized) utility
  - $q(\psi \mid \tau) \propto p(\psi) e^{U_\psi(\tau) - logZ(\psi)}$
  - Similar to prior work on relabeling in multi-task RL [1]
- Sampling trajectories to compute post-adaptation returns is expensive
  - We approximate utility using Q function:

$$U_\psi(\tau) = \mathbb{E}_{s_1 \sim p(s_1), a_1 \sim \pi'(\cdot \mid s_1)}[Q_\psi^{\pi'}(s_1, a_1)]$$

[1] Rewriting History with Inverse RL: Hindsight Inference for Policy Improvement, Eysenbach et al.

Relabeling for Meta-RL (HFR)

HFR can be incorporated into any off-policy Meta-RL algorithm; we use PEARL [1]

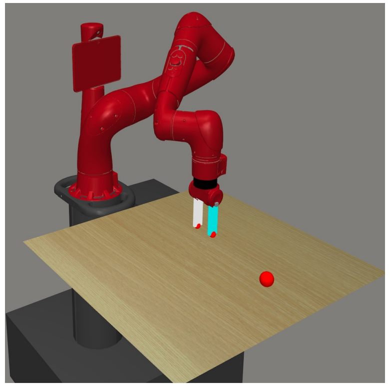
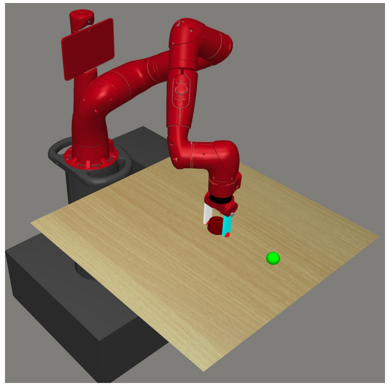
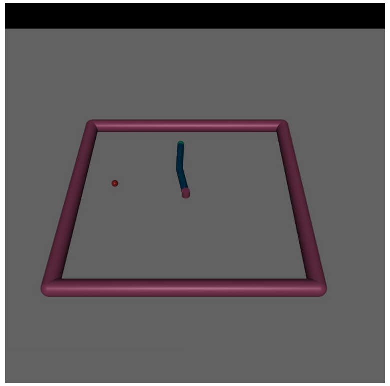
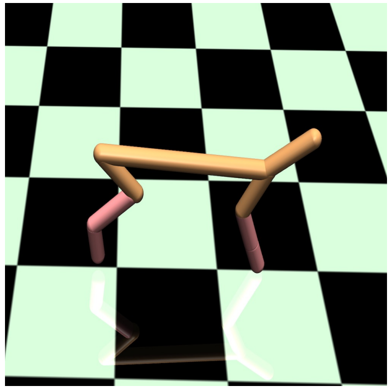
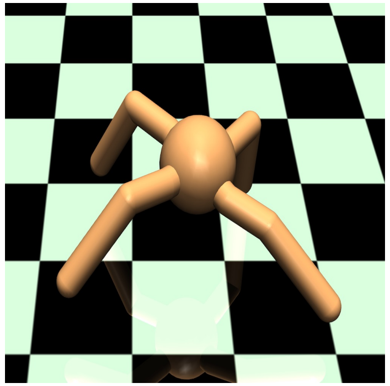


Relabeling is independent of data collection and update rules of underlying Meta-RL algorithm

[1] *Efficient Off-Policy Meta-Reinforcement Learning via Probabilistic Context Variables, Rakelly et al.*

Experiments

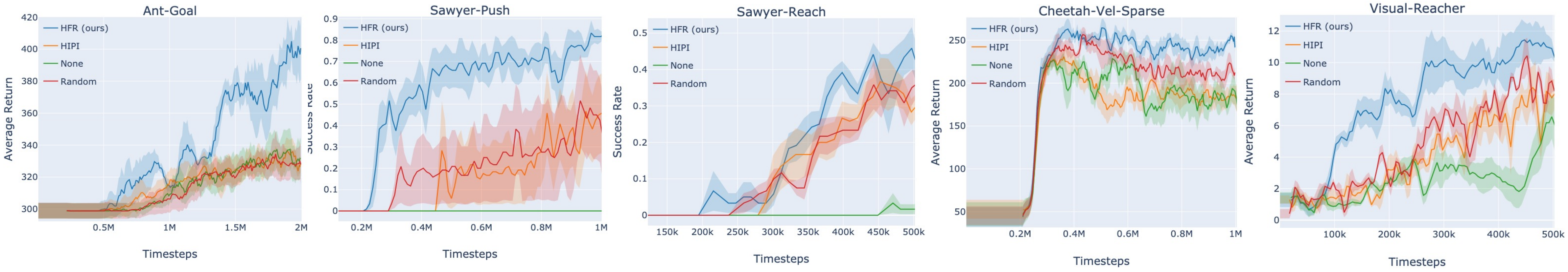
- Evaluate on sparse and dense reward manipulation and locomotion tasks
  - Tasks involve low-dimensional state vectors (Visual-Reacher environment uses high-dimensional images)
- Methods:
  - **HFR (ours)**
  - HIPI [1] (trajectory return value-based relabeling algorithm)
  - Random (assign trajectories to random tasks)
  - None (PEARL with no relabeling)

[1] Rewriting History with Inverse RL: Hindsight Inference for Policy Improvement, Eysenbach et al.

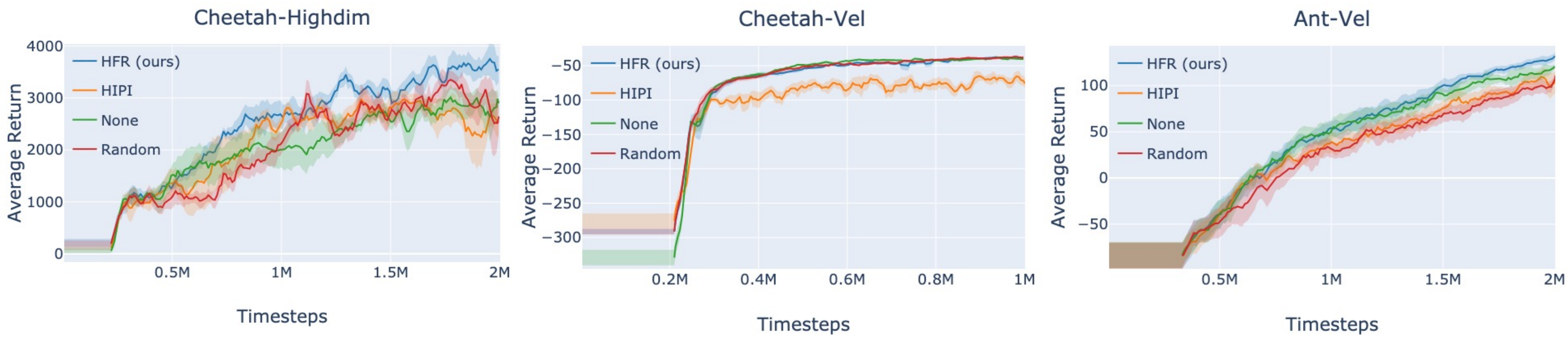
Results

- Sparse Reward Tasks
  - Hard due to lack of reward signal
  - HFR outperforms baselines, suggesting relabeling can mitigate the need for elaborate exploration



Results

- Dense Reward Tasks
  - Benefit of HFR is less pronounced
  - Exploration not as critical due to dense reward



Summary

- We present **HFR**, a trajectory relabeling method for Meta-RL
- Previous Multi-task RL relabeling methods relabel based on trajectory returns; we relabel based on *post-adaptation returns*
    - Relabeling based on post-adaptation returns aligns with Meta-RL objective
- HFR leads to improved performance on a variety of Meta-RL tasks
- To our knowledge, HFR is first approach for data sharing during meta-training phase for Meta-RL