NEURAL INFORMATION PROCESSING SYSTEMS

Off-Dynamics Reinforcement Learning via Domain Adaptation and Reward Augmented Imitation

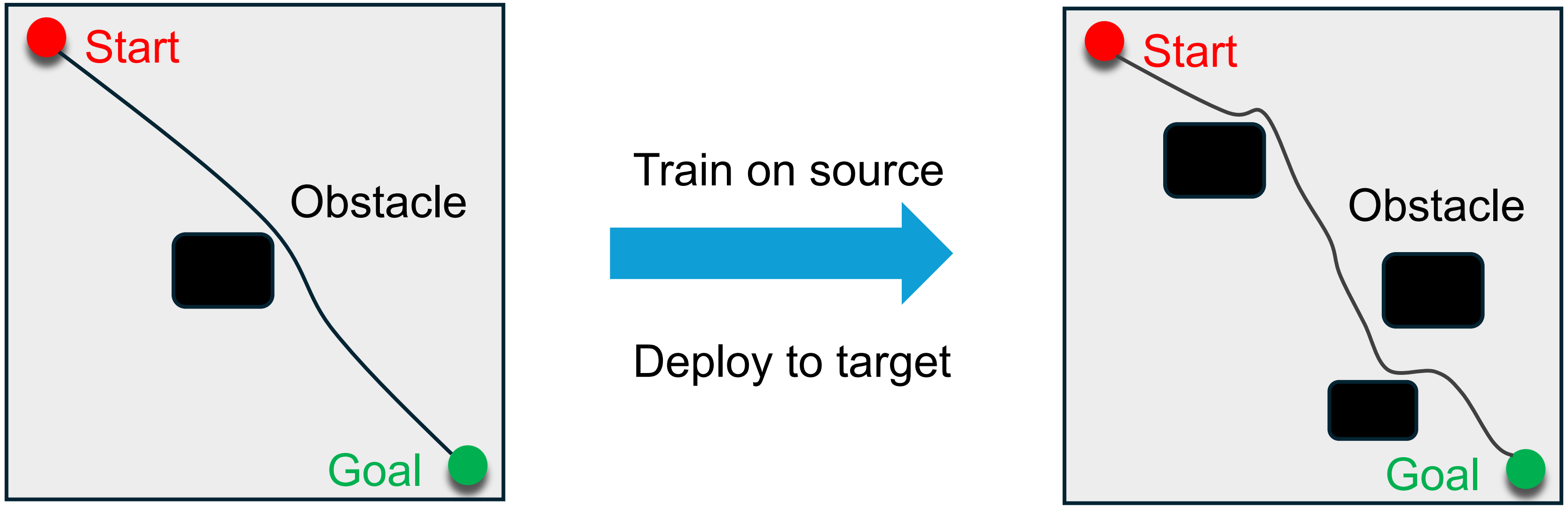
Yihong Guo, Yixuan Wang, Yuanyuan Shi, Pan Xu, Anqi Liu

JOHNS HOPKINS UNIVERSITY

UC San Diego

Duke UNIVERSITY

Off-dynamics Reinforcement Learning

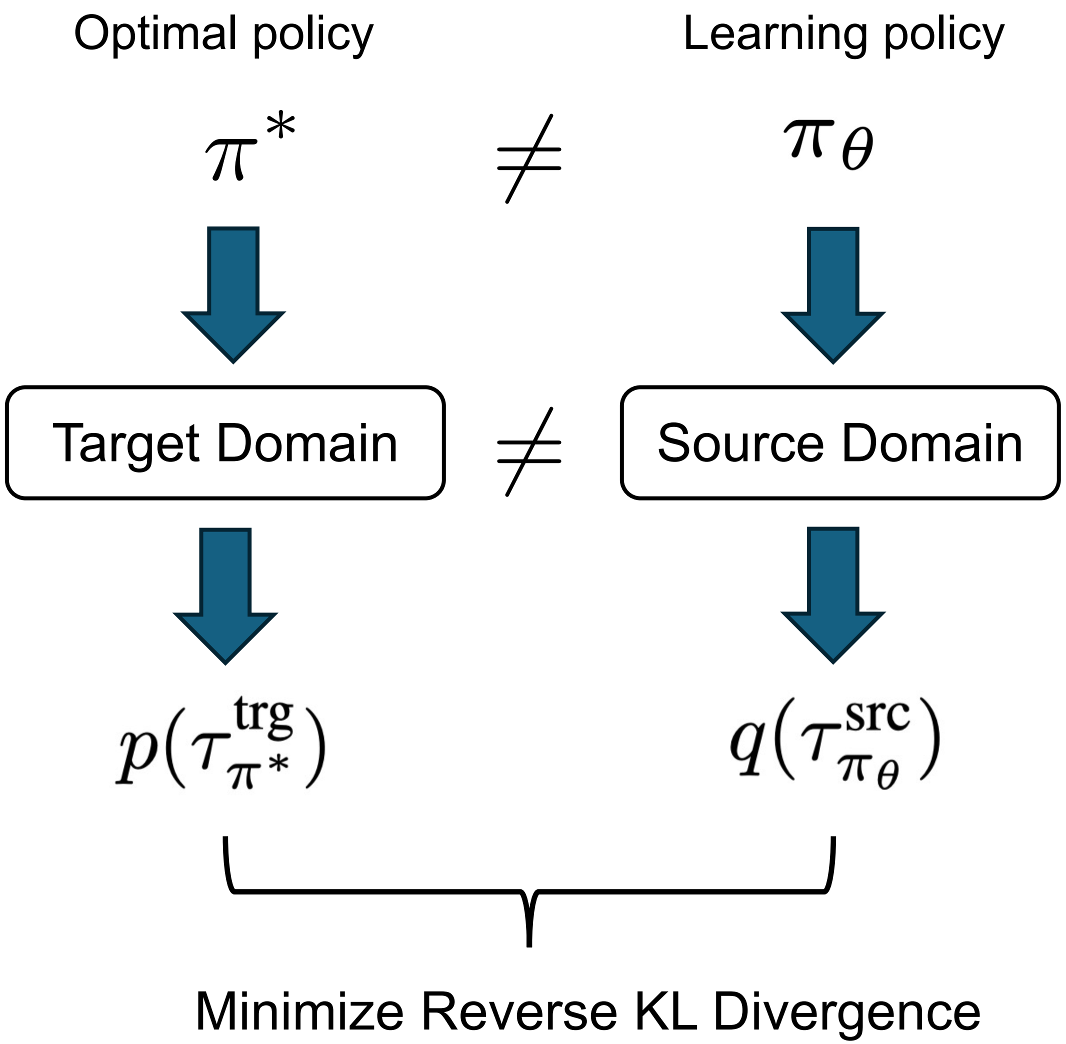


Source domain (simulation)

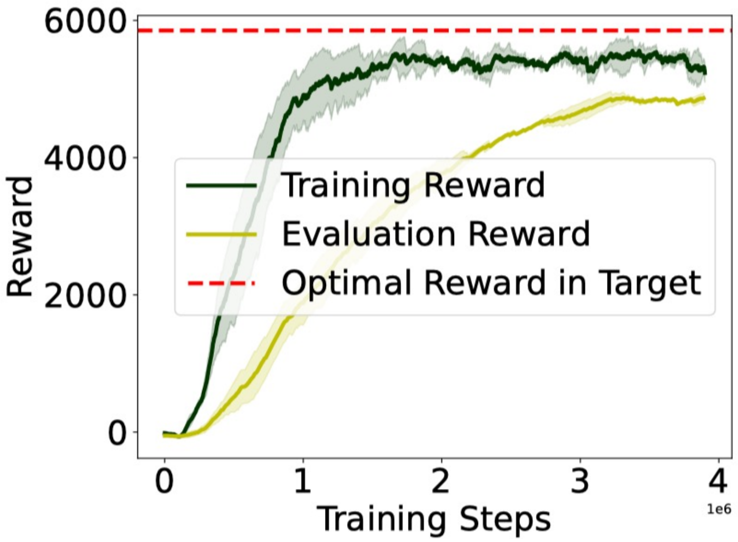
Target domain (real-world)

- Different transition probability: $P_{src}(s_{t+1}|s_t, a_t) \neq P_{trg}(s_{t+1}|s_t, a_t)$
- Same state, action space and reward function.

Limitation of existing reward shaping methods DARC [1]

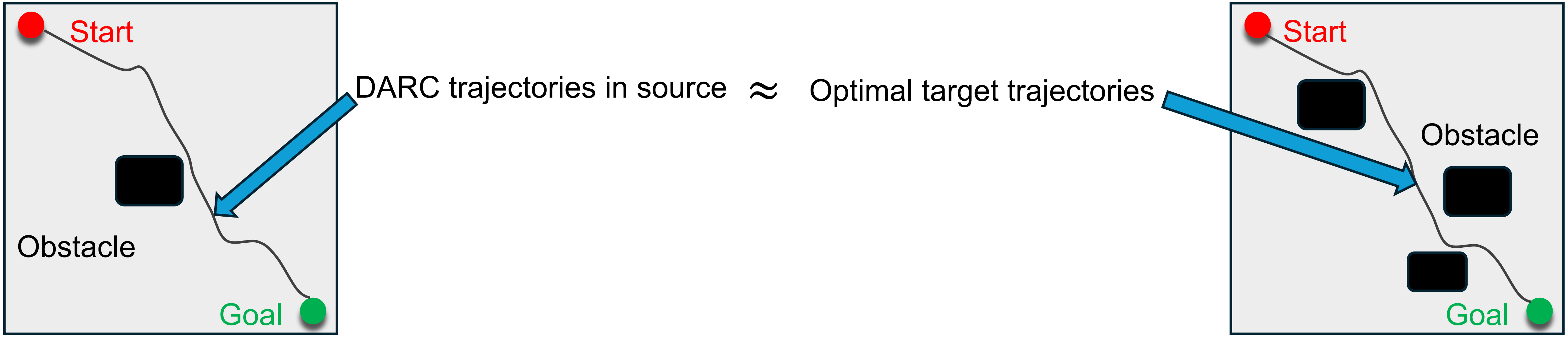


- DARC policy is suboptimal in the target domain.
- DARC relies on the assumption of target optimal policy performs well in the source domain.



[1] Eysenbach B, Asawa S, Chaudhari S, et al. Off-dynamics reinforcement learning: Training for transfer with domain classifiers[J]. arXiv preprint arXiv:2006.13916, 2020.

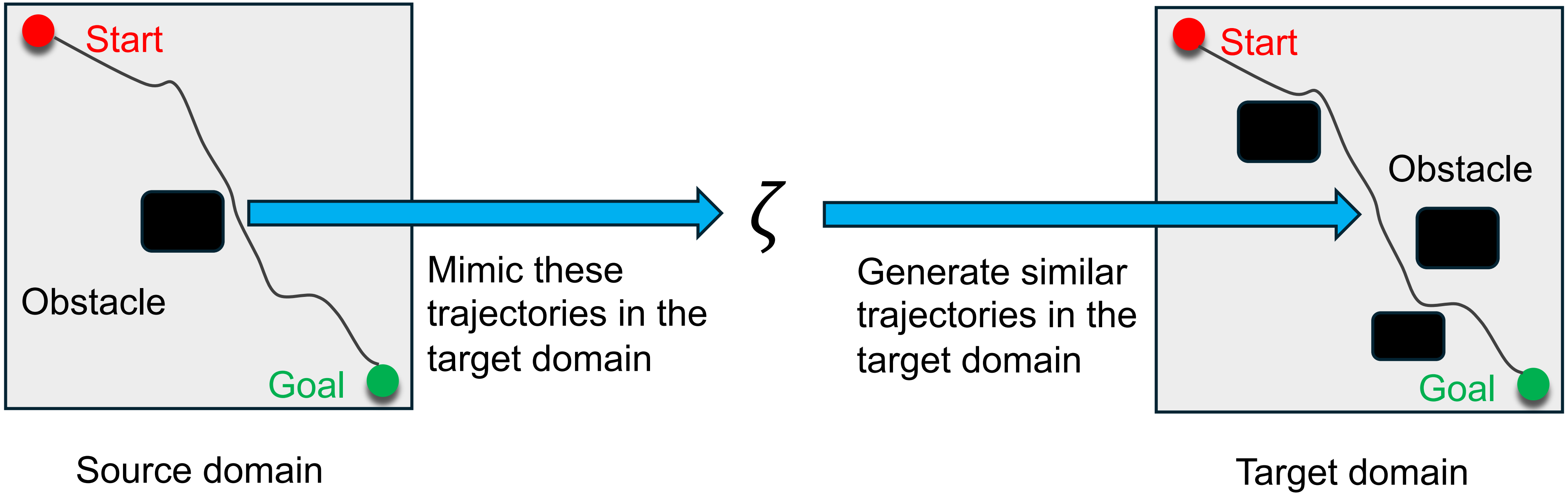
Domain Adaptation and Reward Augmented Imitation Learning (DARAIL)



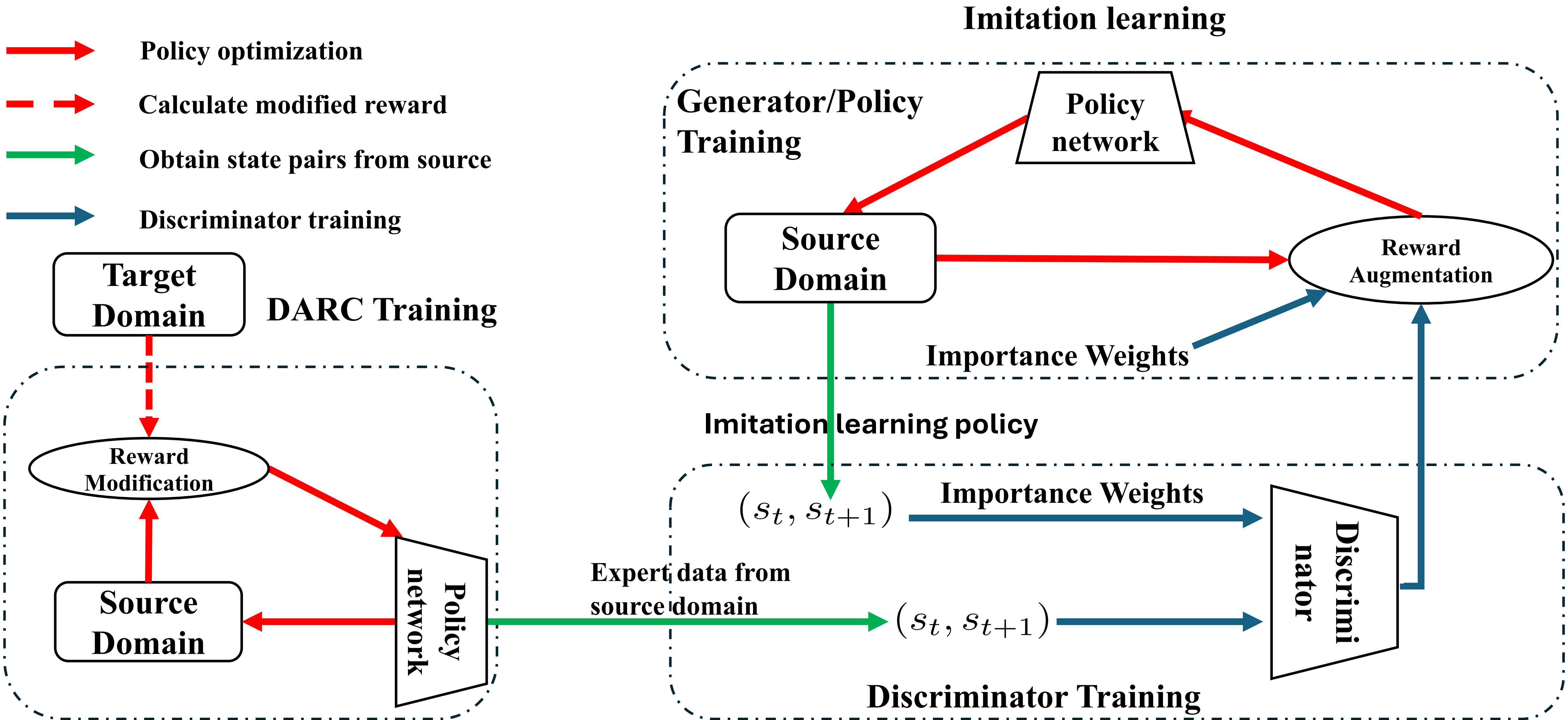
Source domain

Target domain

Domain Adaptation and Reward Augmented Imitation Learning (DARAIL)



Domain Adaptation and Reward Augmented Imitation Learning (DARAIL)



Experiments

- Experiment on Mujoco: HalfCheetah, Ant, Walker2d and Reacher
- Dynamics shift:
  - Broken environment: set the 0-th index action to 0.
  - Modifying parameters: set the scale of gravity/density from 1.0 to 0.5/1.5 in the target domain.
- Evaluate with mean reward and standard deviation among multiple runs.

Comparison with DARC

- DARC evaluation worse than DARC training.
- DARAIL outperforms DARC evaluation results.

Table 1: Comparison of DARAIL with DARC, broken source environment.

	DARC Evaluation	DARC Training	Optimal in Target	DARAIL
HalfCheetah	4133 ± 828	6995 ± 30	8543 ± 230	7067 ± 176
Ant	4280 ± 33	5197 ± 155	6183 ± 348	5357 ± 79
Walker2d	2669 ± 788	3896 ± 523	3899 ± 214	4366 ± 434
Reacher	-26.3 ± 3.3	-11.2 ± 2.9	-7.2 ± 1.2	-13.7 ± 0.9

Table 2: Comparison of DARAIL with DARC, 1.5 gravity.

	DARC Evaluation	DARC Training	Optimal in Target	DARAIL
HalfCheetah	653 ± 142	4897 ± 653	6894 ± 491	4093 ± 1021
Ant	1587 ± 594	2170 ± 258	5320 ± 429	3472 ± 771
Walker2d	257 ± 28	4130 ± 689	4254 ± 345	4409 ± 401
Reacher	-55.3 ± 10.3	-17.2 ± 3.8	-8.3 ± 1.3	-9.5 ± 0.22

DARAIL outperforms other baselines

Table 3: Comparison of DARAIL with baselines in off-dynamics RL, broken source environment.

	DAIL	IS-R	IS-ACL	MBPO	MATL	GARAT	DARAIL
HalfCheetah	6402 ± 362	6007 ± 863	6934 ± 231	4323 ± 7	1538 ± 616	5877 ± 382	**7067** ± 176
Ant	3239 ± 395	1463 ± 1055	2753 ± 94	2445 ± 13	2006 ± 17	3380 ± 268	**5357** ± 79
Walker2d	2330 ± 156	3092 ± 434	3881 ± 269	1012 ± 41	250 ± 5	3296 ± 284	**4366** ± 434
Reacher	-13.9 ± 1.1	-17.6 ± 0.25	-14.1 ± 0.16	-14.3 ± 2	-30 ± 10	-14.7 ± 2.6	**-13.7** ± 0.9

Table 4: Comparison of DARAIL with baselines in off-dynamics RL, 1.5 gravity.

	DAIL	IS-R	IS-ACL	MBPO	MATL	GARAT	DARAIL
HalfCheetah	2666 ± 2037	2718 ± 1978	3576 ± 312	619 ± 311	337 ± 205	3825 ± 437	**4093** ± 1021
Ant	990 ± 251	1712 ± 393	2396 ± 573	989 ± 13	1376 ± 466	1961 ± 115	**3472** ± 771
Walker2d	525 ± 142	1543 ± 604	1369 ± 705	870 ± 451	1419 ± 489	630 ± 230	**4409** ± 401
Reacher	-16.5 ± 1.1	-14.6 ± 0.8	-47.4 ± 8.3	-18.3 ± 0.9	-17.6 ± 0.7	-16.7 ± 0.3	**-9.5** ± 0.22

Summary

- Propose DARAIL for off-dynamics RL.
- Recognize the limitations of DARC and other works with same reward shaping method.
- Propose imitation learning with augmented reward estimator to address the limitation of DARC.
- Propose an error bound with a relaxed assumption.