UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

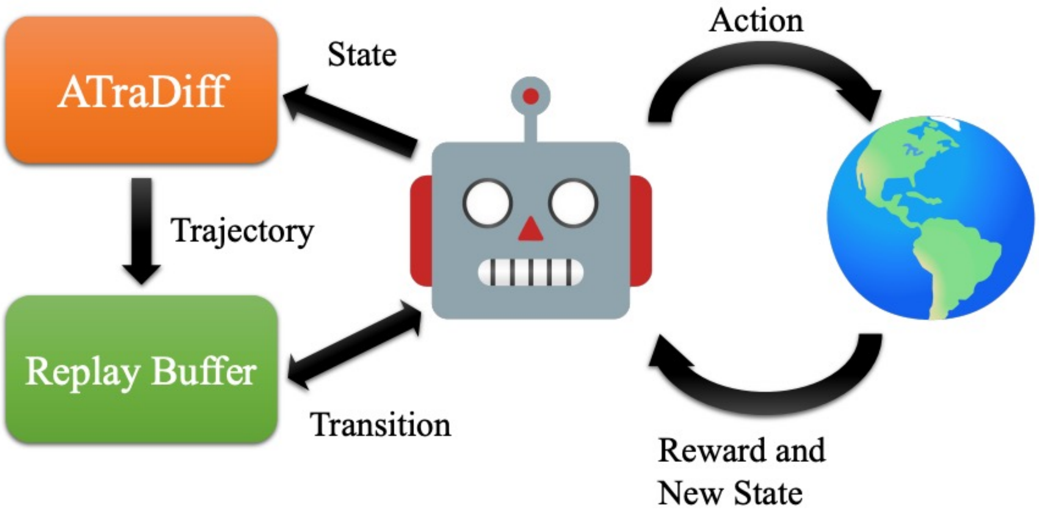
ATraDiff: Accelerating Online Reinforcement Learning with Imaginary Trajectories

Qianlan Yang, Yu-xiong Wang

Motivation

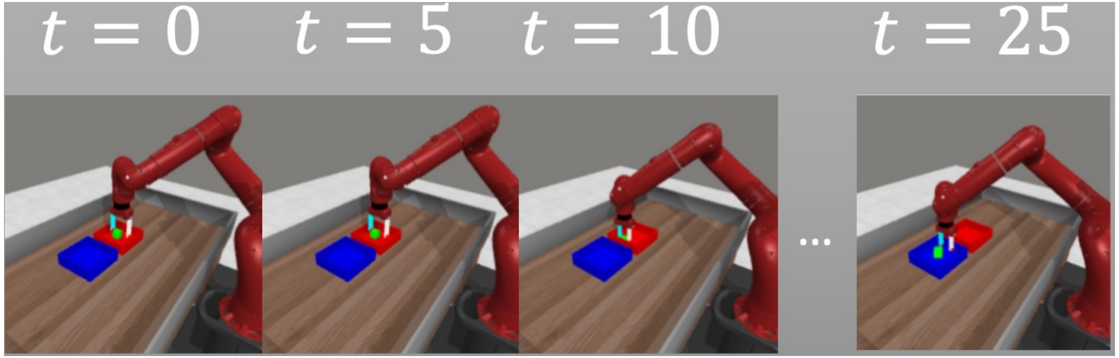
- Sample efficiency is important in Online RL.
- Key problem:

Can we harness modern generative models trained on offline data and synthesize useful data that facilitate online RL?



Why Diffusion models?

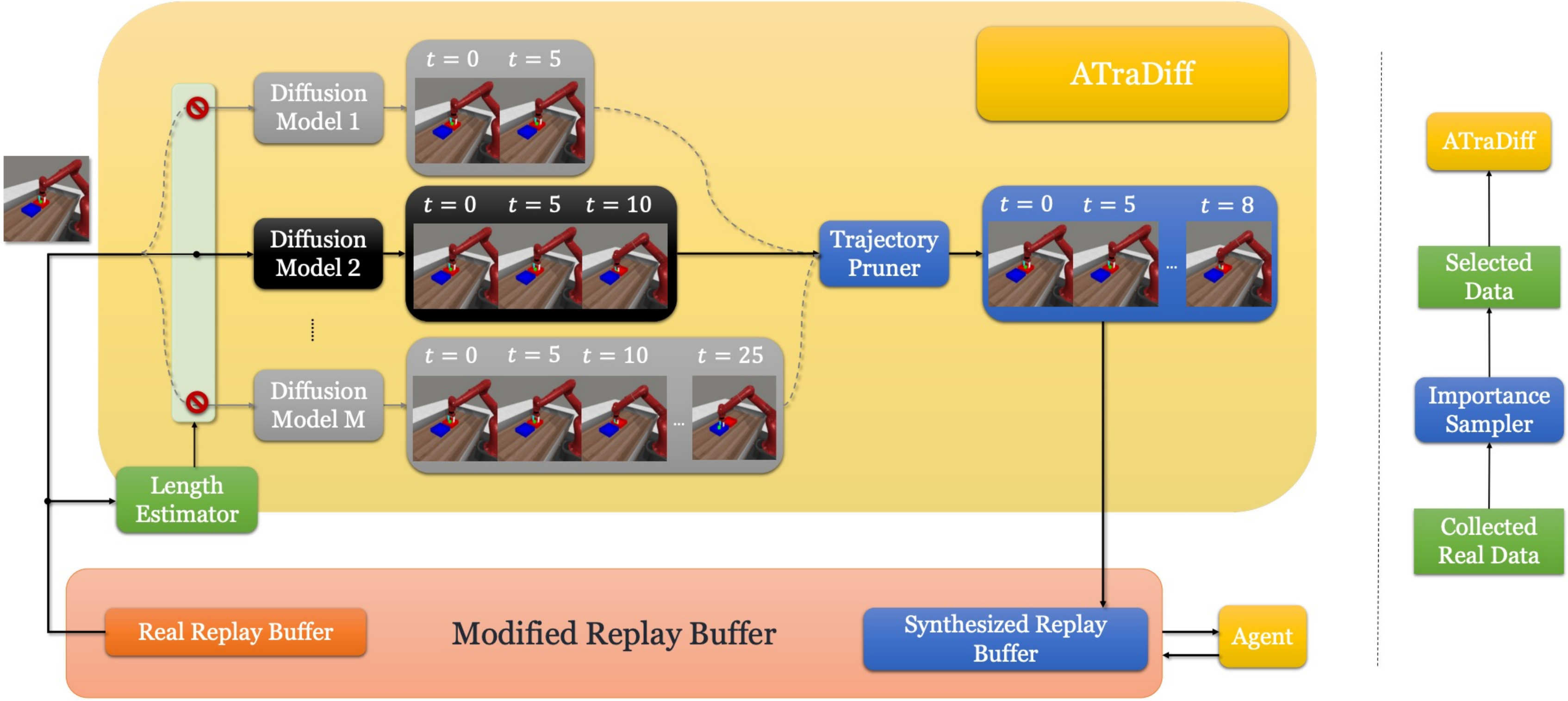
- Diffusion models have shown impressive capabilities in data synthesis across vison and language applications.
- Previous works in Reinforcement Learning focused on generation of transitions instead of trajectories.



Main contribution

- We propose ATraDiff, a novel diffusion-based approach that generate full synthetic trajectories.
- ATraDiff seamlessly integrates with a wide spectrum of RL methods.
- We introduce a simple yet effective coarse-to-precise strategy that ensures generation with flexible lengths.
- We devise an online adaptation mechanism

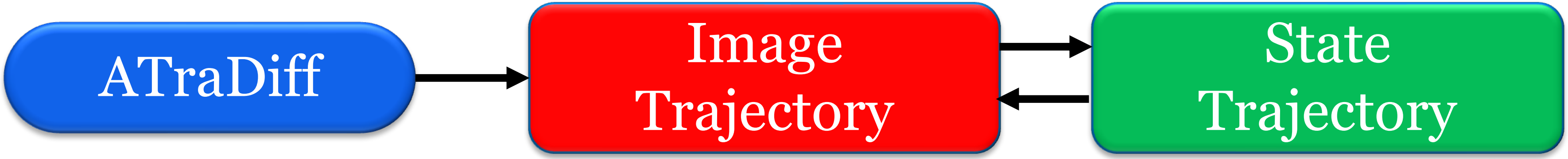
Framework



$t = 0$ $t = 5$
ATraDiff
Diffusion Model 1
Diffusion Model 2
$t = 0$ $t = 5$ $t = 10$
Trajectory Pruner
$t = 0$ $t = 5$ $t = 8$
Diffusion Model M
$t = 0$ $t = 5$ $t = 10$ $t = 25$
Length Estimator
Real Replay Buffer
Modified Replay Buffer
Synthesized Replay Buffer
Agent
ATraDiff
Selected Data
Importance Sampler
Collected Real Data

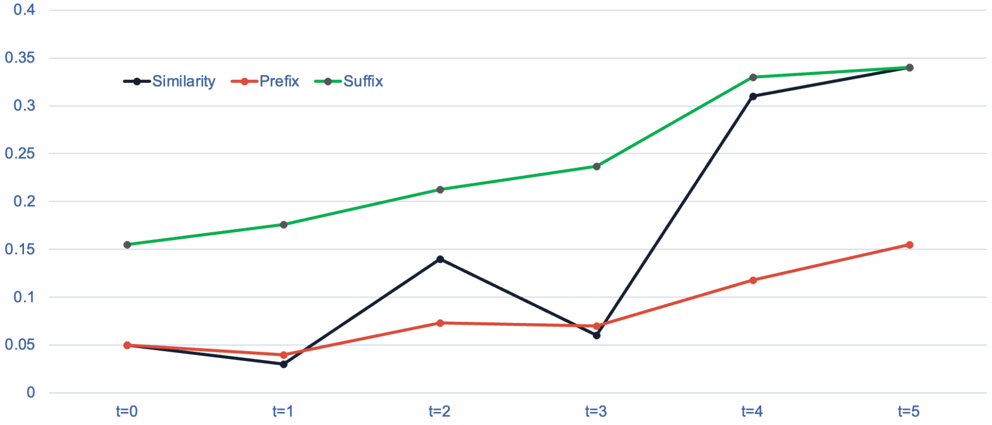
Method: Adaptive Trajectory Diffuser

- Supports generation of both state-level and image-level trajectories, while the image-level generation achieves higher performance.
- An end-to-end state decoder and encoder, use to convert the trajectories from state-level representation and image-level representation.



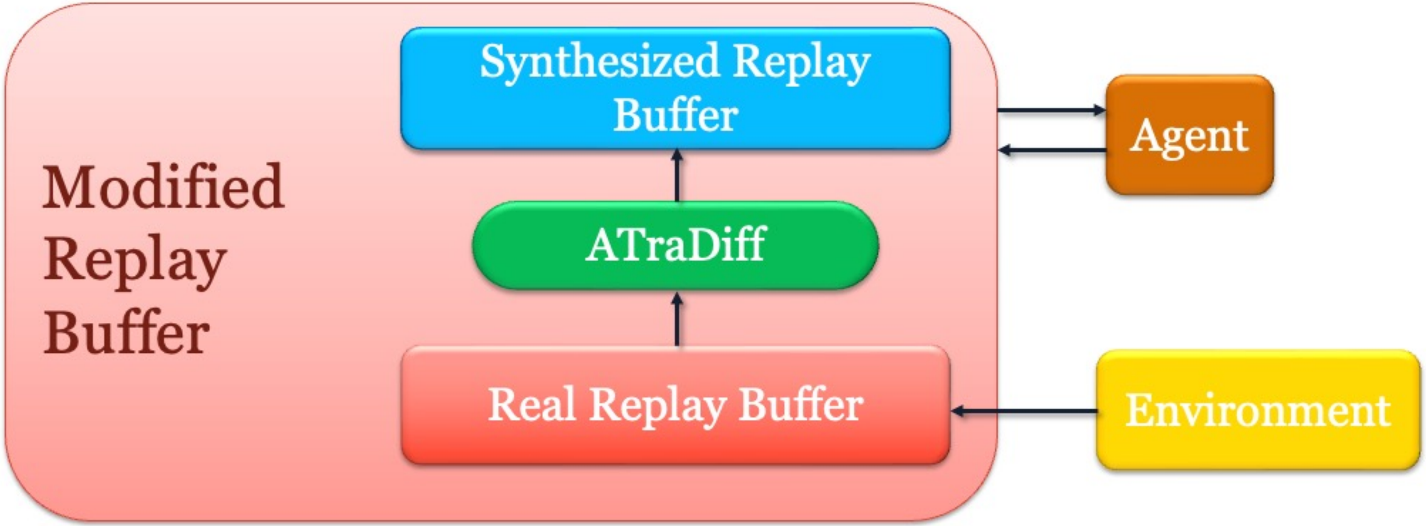
Method: Prune Strategy

- A *coarse-to-precise* strategy is used to generate trajectories with arbitrary flexible lengths.
- Length estimator first estimate the required length, then select a generator.
- A prune algorithm is used to cut the trajectory to a precise length



Method: Deployment

- ATraDiff can be seamlessly applied to accelerate *any* online RL algorithm with a replay buffer.

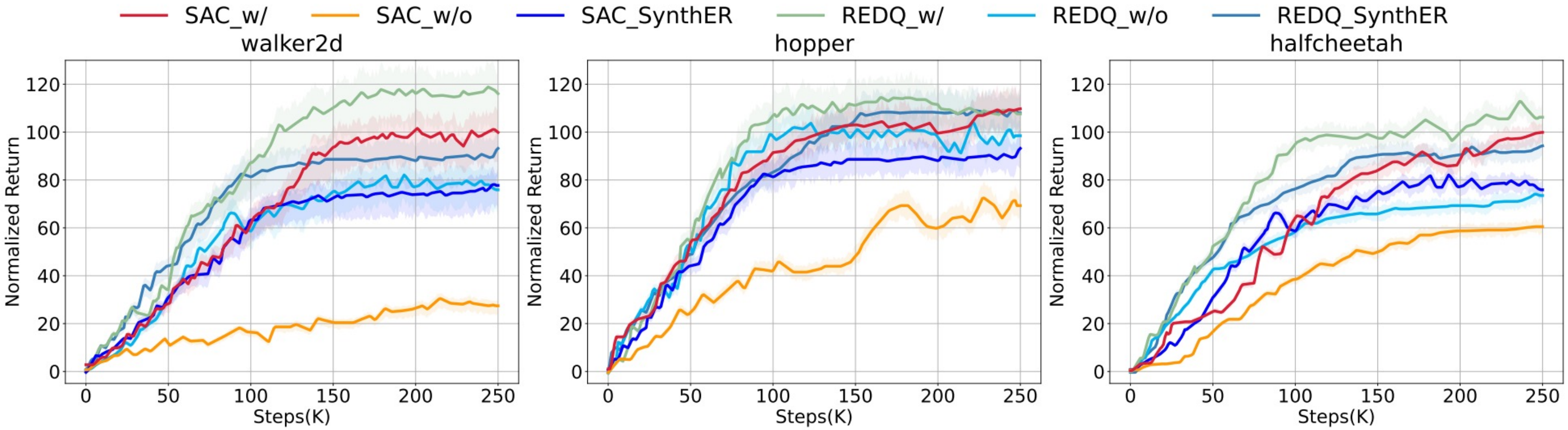


Method: Online Adaptation

- ATraDiff is periodically updated on the online-collected real transitions.
- An indicator used to measure the importance of samples for the online adaption.
- A pick-up strategy to choose samples from the real replay buffer.

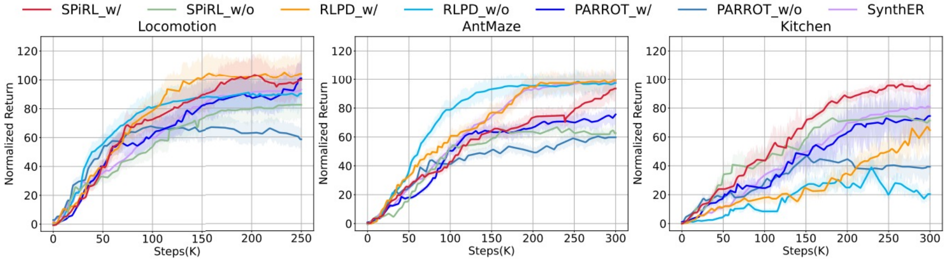
Online Reinforcement Learning

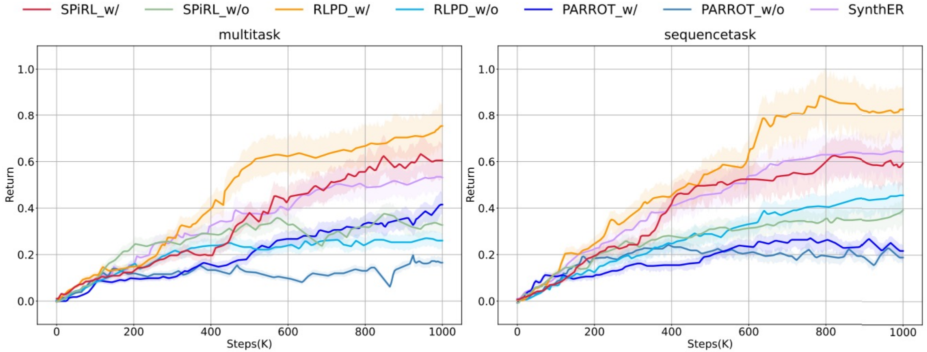
ATraDiff consistently improve the performance of online RL method across different environments.



Offline-to-Online Reinforcement Learning

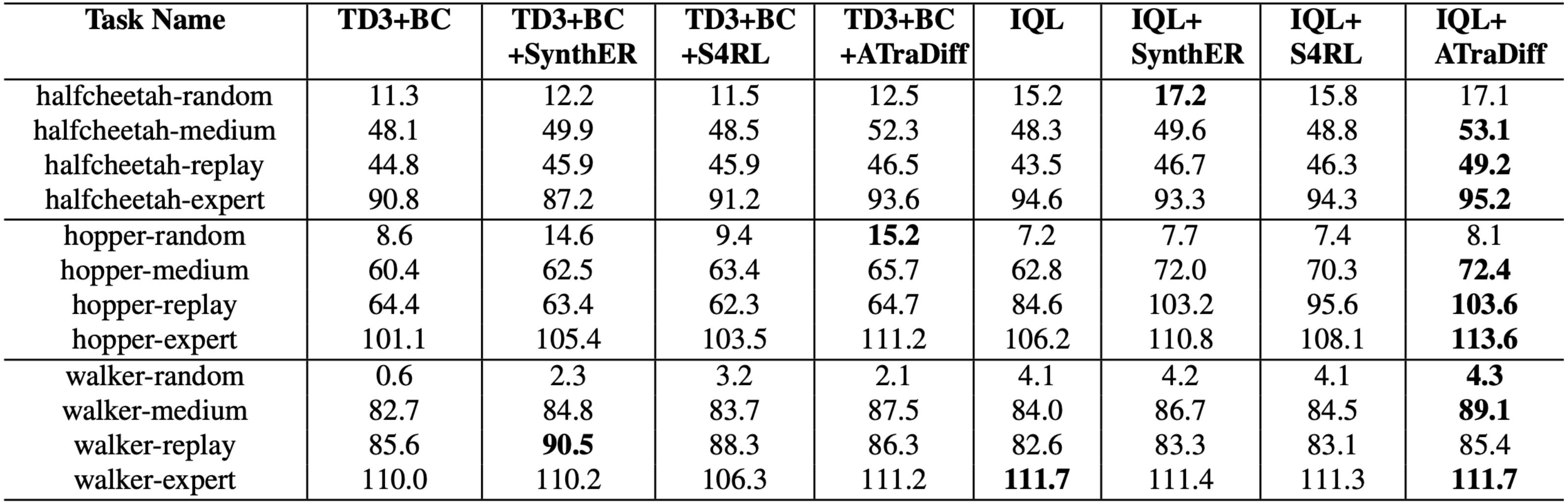
ATraDiff can further boost the performance of offline-to-online RL baselines across different environments, especially in complicated tasks.





Offline Reinforcement Learning

ATraDiff can also improve the performance of offline RL methods.



Task Name	TD3+BC	TD3+BC +SynthER	TD3+BC +S4RL	TD3+BC +ATraDiff	IQL	IQL+ SynthER	IQL+ S4RL	IQL+ ATraDiff
halfcheetah-random	11.3	12.2	11.5	12.5	15.2	**17.2**	15.8	17.1
halfcheetah-medium	48.1	49.9	48.5	52.3	48.3	49.6	48.8	**53.1**
halfcheetah-replay	44.8	45.9	45.9	46.5	43.5	46.7	46.3	**49.2**
halfcheetah-expert	90.8	87.2	91.2	93.6	94.6	93.3	94.3	**95.2**
hopper-random	8.6	14.6	9.4	**15.2**	7.2	7.7	7.4	8.1
hopper-medium	60.4	62.5	63.4	65.7	62.8	72.0	70.3	**72.4**
hopper-replay	64.4	63.4	62.3	64.7	84.6	103.2	95.6	**103.6**
hopper-expert	101.1	105.4	103.5	111.2	106.2	110.8	108.1	**113.6**
walker-random	0.6	2.3	3.2	2.1	4.1	4.2	4.1	**4.3**
walker-medium	82.7	84.8	83.7	87.5	84.0	86.7	84.5	**89.1**
walker-replay	85.6	**90.5**	88.3	86.3	82.6	83.3	83.1	85.4
walker-expert	110.0	110.2	106.3	111.2	**111.7**	111.4	111.3	**111.7**

Main ablation study results

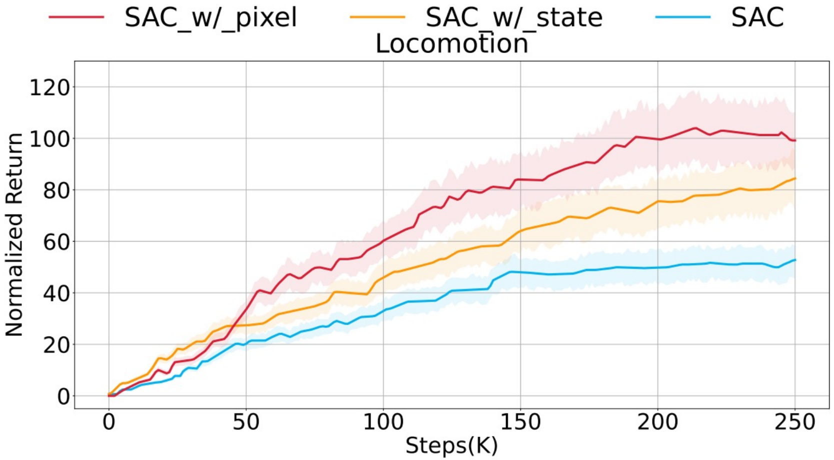
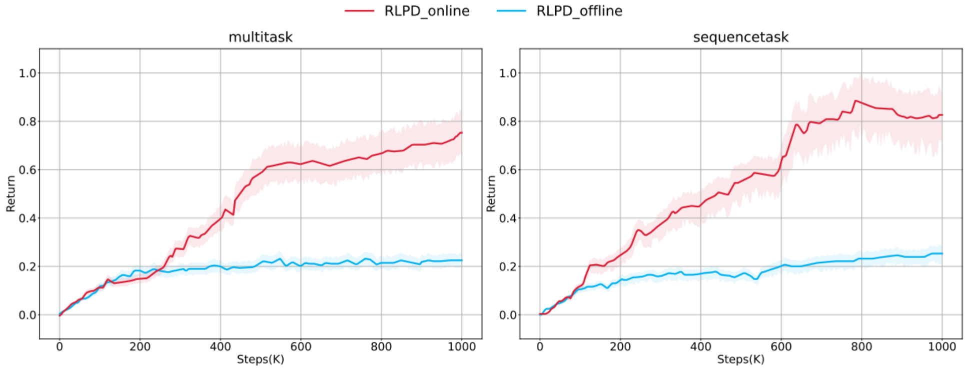


Image generation v.s. State generation



Online v.s. Offline

Thanks for Listening

Project Page