A. Hyperparameter Setting and Environment Description

The hyperparameters for simple PPO, PPO-AMBER, and DISC are summarized in Table A.1, and the dimensions of state and action spaces of Open AI GYM tasks are given in Table A.2.

	PPO	PPO-AMBER	DISC
Clipping factor (ϵ)	0.2	0.2	0.4
Horizon (N)	2048	2048	2048
Discount factor (γ)	0.99	0.99	0.99
TD parameter (λ)	0.95	0.95	0.95
Epoch	10	10	10
Gradient steps per epoch	32	32	32
Mini-batch size (M)	64	variable	variable
Optimizer	Adam	Adam	Adam
Learning rate (β)	$\max(0.0001,\ \text{Anneal}(0.0003,\ 0))$		
Policy distribution	Independent Gaussian distribution		
Policy and value network	Feed forward network with 2 hidden layers of size 64 and $\tanh(\cdot)$ activation		
Batch inclusion parameter (ϵ_b)	·	0.1, 0.2	0.1
Replay length (L)	·	64	64
IS target parameter (J_{targ})	·	·	0.0001
Initial IS weighting factor (α_{IS})	·	·	1

Table A.1: Hyper-parameter setting of PPO, PPO-AMBER, and DISC

Mujoco	State dim.	Action dim.
Ant-v1	111	8
HalfCheetah-v1	17	6
Hopper-v1	11	3
Humanoid-v1	376	17
HumanoidStandup-v1	376	17
Walker2d-v1	17	6
Box2d	State dim.	Action dim.
BipedalWalker-v2	24	4
BipedalWalkerHardcore-v2	24	4

Table A.2: The dimensions of state and action spaces of OpenAI GYM continuous control tasks

B. Ablation Study on More Tasks

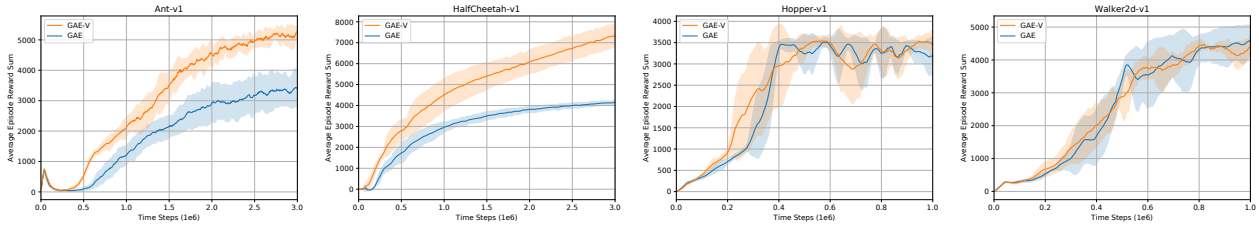

Figure B.1: GAE-V versus GAE

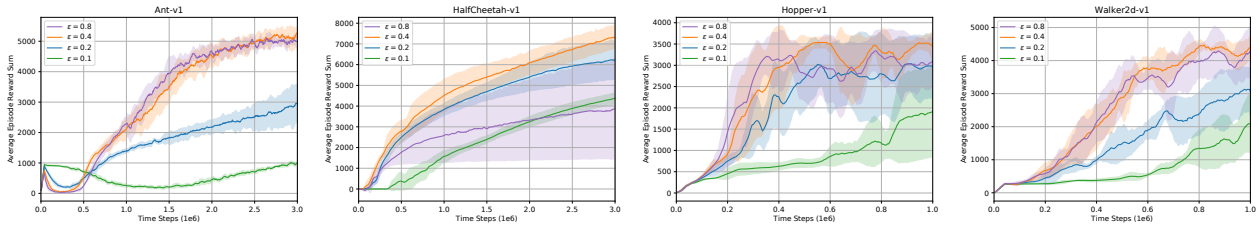

Figure B.2: Impact of the clipping factor ϵ

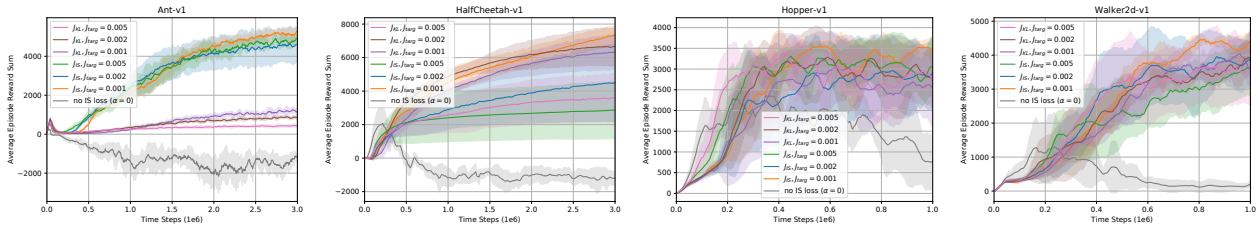

Figure B.3: Impact of the IS loss term J_{IS} target factor J_{targ}

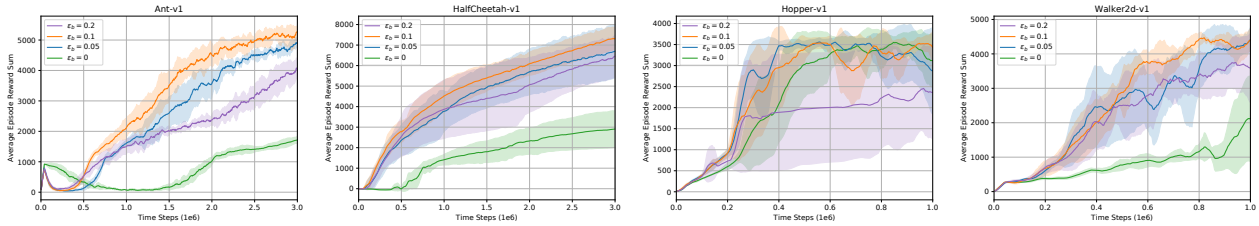

Figure B.4: Impact of the batch inclusion parameter ϵ_b

C. Learning Curves of DISC and Other State-of-the-Art RL Algorithms

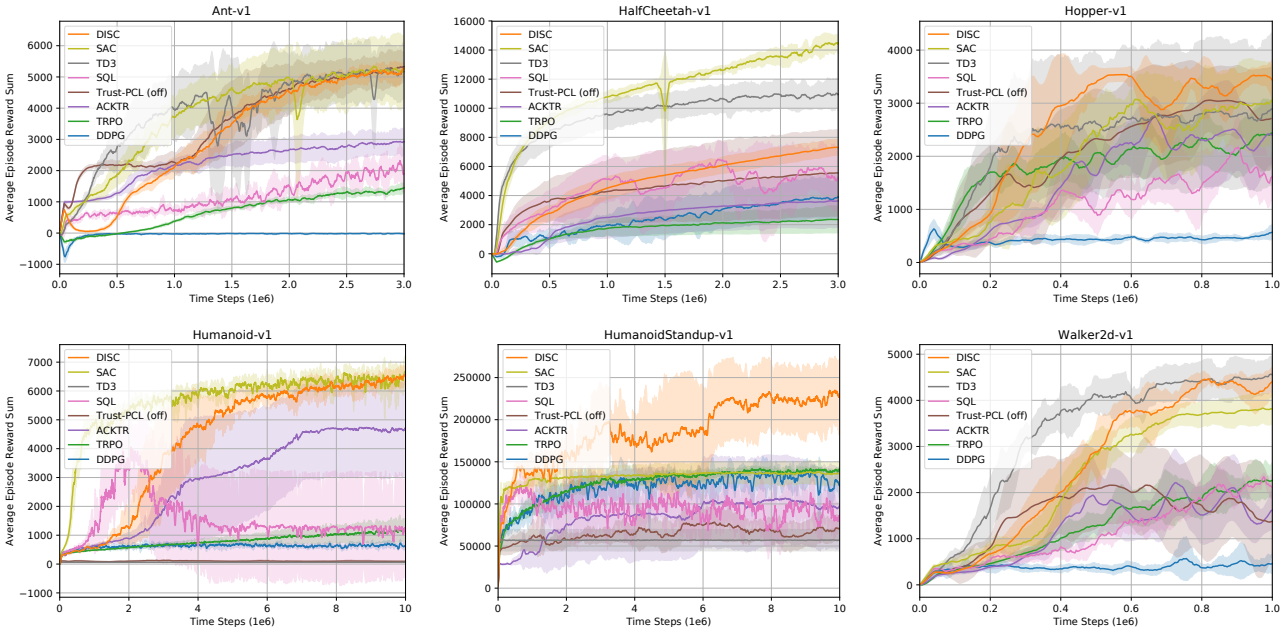

Figure C.1: Learning Curves of DISC and Other State-of-the-Art RL Algorithms on Mujoco tasks