Appendix A: Experimental Details and Hyperparameter

1. DeepMind Control Suite. DeepMind Control Suite is a collection of continuous control tasks that involve the manipulation of high-dimensional systems [16]. The objective of the agent is to learn how to effectively control the environment in order to achieve specific goals. Our experiments focused on nine distinct environments, which are detailed in Table 2. For a comprehensive listing of the hyperparameters utilized in our research, please refer to Table 1.

Hyperparameters	Value
# of ensemble agents	2
Training steps	1×10^6
Discount factor	0.99
Initial collection steps	5000
Minibatch size	1024
Optimizer (all)	Adam
Optimizer (all) : learning rate	0.0001 `humanoid-run` 0.0003 otherwise
Networks (all) : activation	ReLU
Networks (all) : n. hidden layers	2
Networks (all) : hidden units	1024
Initial Temperature	1
Replay Buffer Size	1×10^6
Updates per step (Replay Ratio)	(1, 2, 4)
Target network update period	1
τ	0.005
Reset Interval (gradient steps)	4×10^5
β (action select coefficient)	50

Table 1: Hyperparameters for RDE+SAC on DeepMind Control Suite.

Environment	Task
`acrobot`	`swingup`
`cheetah`	`run`
`finger`	`turn_hard`
`fish`	`swim`
`hopper`	`hop`
`humanoid`	`run`
`quadruped`	`run`
`swimmer`	`swimmer15`
`walker`	`run`

Table 2: The tasks of DMC

2. Atari 100k. Atari 100k is a benchmark that tests an agent's abilities by allowing it to interact with 100k environment steps (equivalent to 400k frames with a frameskip of 4) in 26 Atari games [4]. Each game has different mechanics, providing a diverse evaluation of the agent's capabilities. The benchmark imposes a restriction of 100k actions per environment, which roughly corresponds to 2 hours of human gameplay. The hyperparameters and values are listed in Table 3 and Table 4.

Reset Interval	Environments
4×10^4	`Assault, Asterix, Battle Zone, Boxing, Crazy CLimber, Freeway, Frostbite, Krull, Ms Pacman, Qbert, Road Runner, Seaquest, Up N Down`
8×10^4	`Alien, Amidar, Bank Heist, Breakout, Chopper Command, Demon Attack, Gopher, Hero, Jamesbond, Kangaroo, Kung Fu Master, Pong, Private Eye`

Table 3: Reset Interval in terms of the gradient step for each environment of Atari-100k

Hyperparameters	Value
# of ensemble agents	2
Gray-scaling	True
Observation down-sampling	84×84
Frames stacked	4
Action repetitions	4
Reward clipping	[-1, 1]
Terminal on loss of life	True
Max gradient norm	10
Replay periode every	1 step
Training steps	1×10^5
Discount factor	0.99
Initial collection steps	1×10^4
Minibatch size	32
Optimizer	Adam
Optimizer : learning rate	0.0001
Q network : channels	32, 64, 64
Q network : filter size	8×8, 4×4, 3×3
Q network : stride	4, 2, 1
Q network : activation	ReLU
Q network : hidden units	512
Replay Buffer Size	1×10^5
Updates per step (Replay Ratio)	(1, 2, 4)
Target network update period	1
Exploration	ϵ-greedy
ϵ-decay	1×10^4
τ	0.005
β (action select coefficient)	50
Reset depth	last 2 layers: `Amidar, Bank Heist Freeway, Frostbite, Hero` last 1 layer: otherwise

Table 4: Hyperparameters for RDE+DQN on Atari 100k.

3. MiniGrid. MiniGrid is a collection of goal-oriented, 2D grid-world environments [5]. Each interaction step results in the agent receiving a sparse reward denoted as R_1, which is subject to a small decrement. In this paper, we have set the reward value as $R_1 = 10$. We considered 5 tasks in MiniGrid: `FourRooms`, `SimpleCrossingS9N1`, `LavaCrossingS9N1`, `SimpleCrossingS9N1`, and `GoToDoor-8x8`. We now provide the details of each environment.

FourRooms This environment is designed with four rooms and comprises of a single agent and a green goal. At the start of each episode, both the agent and the green goal are randomly placed within the four rooms. The goal of the agent is to navigate through the environment and ultimately reach the green goal.

SimpleCrossingS9N1, LavaCrossingS9N1 This environment is designed with two rooms that are blocked by obstacles, such as lava (for LavaCrossing) and walls (for SimpleCrossing). The objective for the agent is to successfully reach a goal while avoiding these obstacles. In LavaCrossing, the episode comes to an end if the agent collides with the obstacle, whereas in SimpleCrossing, the episode continues despite the collision.

GoToDoor-8x8 This environment is designed with a single room, four doors, and a single mission text string. The string provides instructions on which door the agent should reach.

LavaGapS7 This environment is designed with a single room, a strip of lava, and a green goal. The objective for the agent is to successfully reach the goal while avoiding the lava.

We provide the hyperparameters used in MiniGrid as follows.

Hyperparameters	Value
ϵ	$0.9 \rightarrow 0.05$
$\epsilon-$decay time step	10^5
target update period	10^3
Replay buffer size	5×10^5
Mini-batch size	256
Optimizer	RMSProp
Learning rate	0.0001
The maxmimum number of steps	100
Reset Interval (gradient steps)	2×10^5 (GoToDoor, LavaCrossing, LavaGap) 1×10^5 (FourRooms, SimpleCrossing)
β (action select coefficient)	50

Table 5: Hyperparameters in MiniGrid.

Appendix B: Experimental Results

We provide the entire results of DQN, SR+DQN, and RDE+DQN on Atari 100k and Minigird in Table 6 and Table 7, respectively. We report the per-environment learning curves of Minigird in Fig. 9. The learning curves of DMC are provided in Fig. 7 and Fig. 8.

RR	1			2			4		
Game	DQN	SR+DQN	RDE+DQN	DQN	SR+DQN	RDE+DQN	DQN	SR+DQN	RDE+DQN
Alien	423.2	512.4	414.4	596.6	506.4	502.4	414.0	**639.6**	610.0
Amidar	46.8	43.2	47.8	54.6	58.2	**68.2**	31.6	66.4	55.2
Assault	438.8	354.5	409.1	409.9	369.2	431.8	372.1	455.3	**462.0**
Asterix	418.0	352.0	426.0	394.0	482.0	**612.0**	306.0	470.0	590.0
Bank Heist	14.0	13.6	16.8	23.2	27.2	21.2	14.4	26.8	**33.6**
Battle Zone	4040.0	3360.0	4040.0	2120.0	4520.0	7880.0	3840.0	7000.0	**8240.0**
Boxing	1.4	-7.7	-2.6	0.8	-0.6	4.2	**5.1**	3.6	1.9
Breakout	16.1	6.7	16.1	19.6	15.0	21.2	**23.8**	20.8	19.5
Chopper Command	828.0	836.0	760.0	**1324.0**	1120.0	1000.0	1080.0	1024.0	1044.0
Crazy Climber	12472.0	16240.0	22556.0	22100.0	22216.0	25784.0	16028.0	25072.0	**56324.0**
Demon Attack	490.8	166.8	324.6	**1492.4**	184.4	652.4	1088.4	355.6	284.8
Freeway	15.1	7.2	4.0	10.9	6.4	16.8	14.8	7.9	**21.2**
Frostbite	233.6	158.4	197.6	206.4	206.8	348.0	116.4	264.4	**271.6**
Gopher	225.6	388.0	381.6	460.8	577.6	535.2	434.0	**876.8**	868.0
Hero	621.6	738.4	1698.6	1068.8	2725.6	2819.4	754.0	3073.2	**3564.2**
Jamesbond	68.0	70.0	126.0	**178.0**	50.0	138.0	140.0	92.0	78.0
Kangaroo	168.0	72.0	104.0	160.0	128.0	168.0	48.0	104.0	**176.0**
Krull	1905.2	2262.4	**5325.6**	2637.5	2460.4	1854.0	2533.6	3144.0	3374.0
Kung Fu Master	8264.0	5908.0	7256.0	6244.0	8216.0	7524.0	6008.0	7996.0	**8284.0**
Ms Pacman	790.0	769.6	609.6	907.2	832.0	831.6	868.4	954.8	**1223.2**
Pong	-20.7	-20.7	-20.8	-19.1	-18.8	-19.5	**-14.4**	-17.0	-18.8
Private Eye	20.0	2.1	44.0	44.0	-69.1	64.0	0.0	40.0	**83.5**
Qbert	457.0	388.0	415.0	497.0	489.0	436.0	615.0	467.0	**941.0**
Road Runner	2288.0	1840.0	1468.0	2288.0	1940.0	**3488.0**	1680.0	1684.0	2132.0
Seaquest	292.0	222.4	243.2	207.2	240.0	337.6	216.0	341.6	**372.0**
Up N Down	1396.8	1068.8	1734.4	1756.0	1769.2	**2258.0**	1662.4	1472.4	1503.6
IQM	1.000	0.852	0.985	1.186	1.024	1.384	1.027	1.282	1.381
Mean	1.000	0.602	1.074	1.314	1.005	1.632	1.159	1.527	1.831

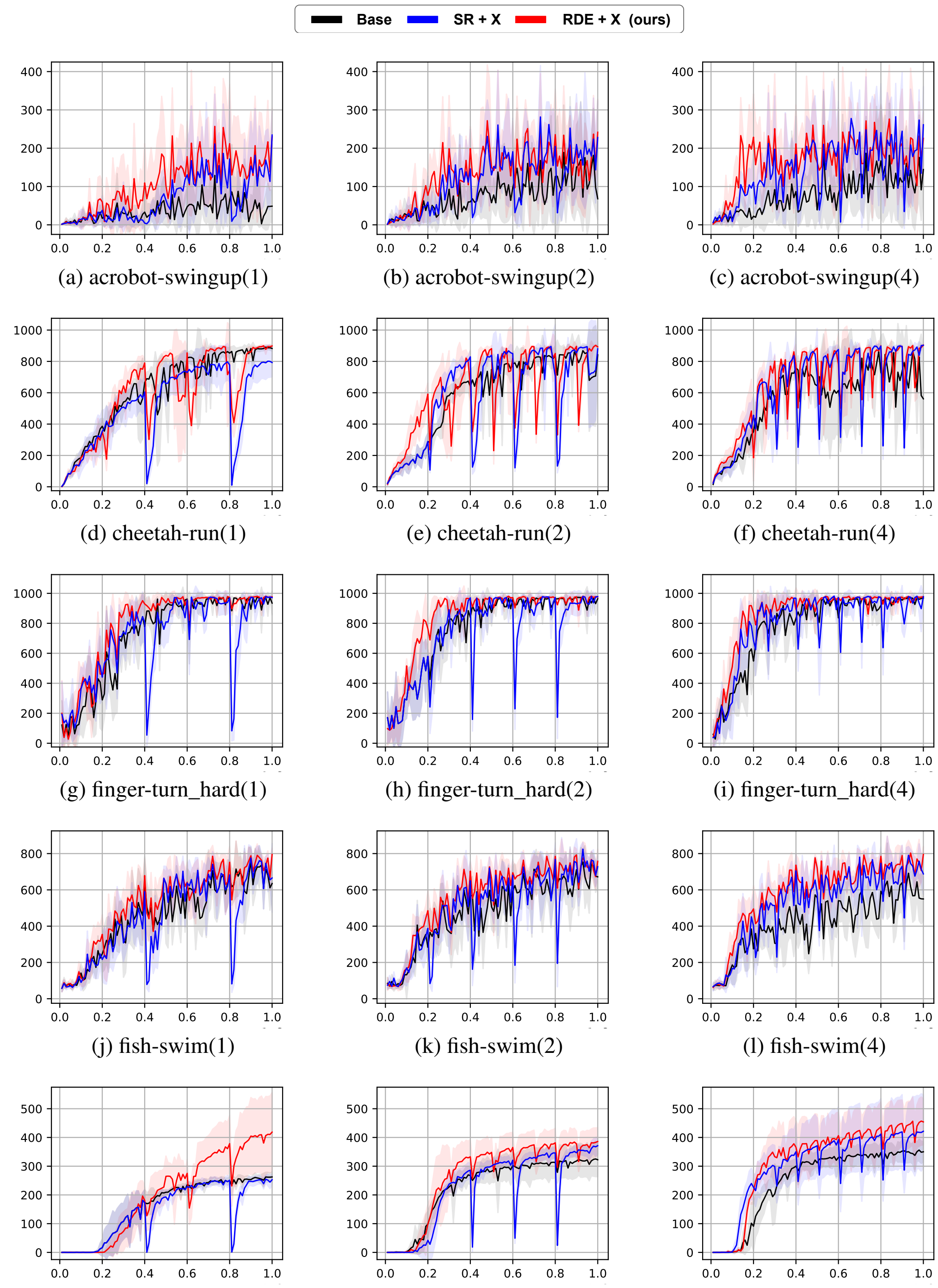
Table 6: Results on Atari-100k

RR	0.5			1			2		
Game	DQN	SR+DQN	RDE+DQN	DQN	SR+DQN	RDE+DQN	DQN	SR+DQN	RDE+DQN
GoToDoor-8x8	0.709	0.71	0.911	0.544	0.659	**0.944**	0.159	0.684	0.929
LavaCrossing	0.035	0	**0.248**	0.016	0.02	0.215	0.029	0	0.237
SimpleCrossingS9N1	0	0.012	**0.186**	0	0.013	0.159	0	0.014	0.137
FourRooms	0.002	0.03	0.148	0	0.072	**0.159**	d 0	0.034	0.155
LavaGapS7	0.747	0.655	0.793	0.761	0.729	**0.793**	0.674	0.706	0.791

Table 7: Results on MiniGrid

Appendix C: Reset depth for Continuous Environments

In Section 4.3, we investigate the impact of reset depth in Minigird environment. In order to demonstrate the effect in continuous tasks, we compare the performance of RDE+DQN using two different reset depth: *reset-1*, which only resets the last layer, and *reset-all*, which entails a complete reset of all layers in the DeepMind Control Suite. As shown in Figure 10, we observe that for several tasks such as `cheetah-run, finger-turn_hard, hopper-hop, swimmer-swimmer15, walker-run`, `reset-1` exhibits comparable performance to *reset-all*. However, for the remaining tasks, namely `acrobot-swingup, fish-swim, humanoid-run, quadruped-run`, *reset-1* demonstrates inferior performance compared to *reset-all*. Furthermore, in the cases of `fish-swim, humanoid-run, swimmer-swimmer15`, the performance of *reset-1* deteriorates with increasing replay ratio, suggesting that shallower levels of resetting render it more susceptible to primacy bias.



(a) acrobot-swingup(1) (b) acrobot-swingup(2) (c) acrobot-swingup(4)

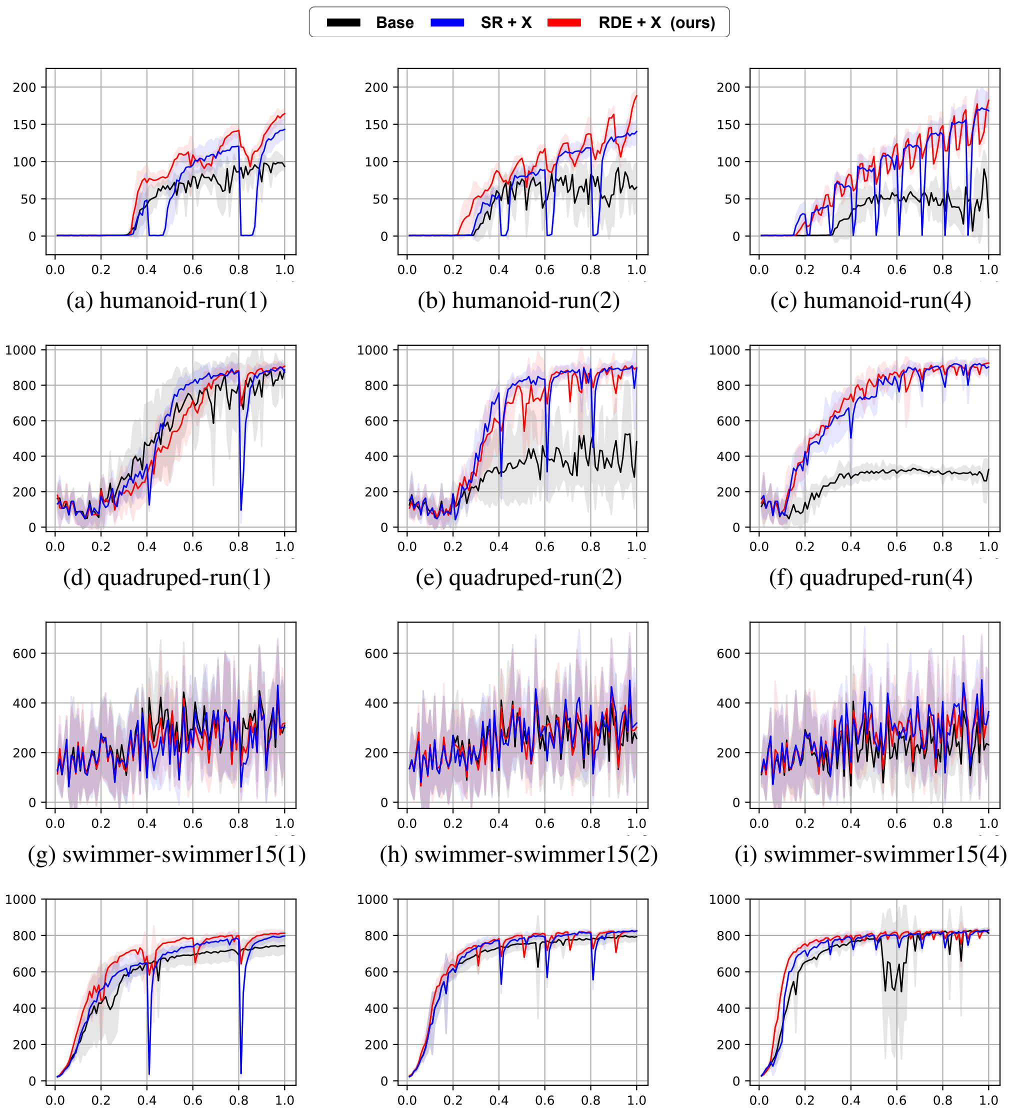
(d) cheetah-run(1) (e) cheetah-run(2) (f) cheetah-run(4)

(g) finger-turn_hard(1) (h) finger-turn_hard(2) (i) finger-turn_hard(4)

(j) fish-swim(1) (k) fish-swim(2) (l) fish-swim(4)

(m) hopper-hop(1) (n) hopper-hop(2) (o) hopper-hop(4)

Figure 7: Per-environment performance in DMC with varying replay ratio values. Note that the number in parentheses indicates the replay ratio. The scale of the x-axis is 10^6. Performances are averaged over 5 seeds.



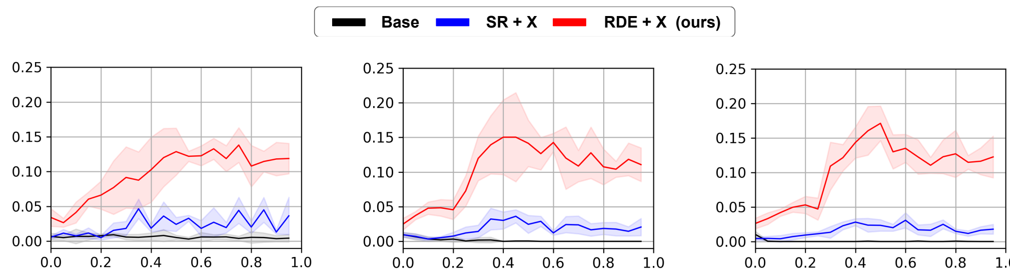
(a) humanoid-run(1) (b) humanoid-run(2) (c) humanoid-run(4)

(d) quadruped-run(1) (e) quadruped-run(2) (f) quadruped-run(4)

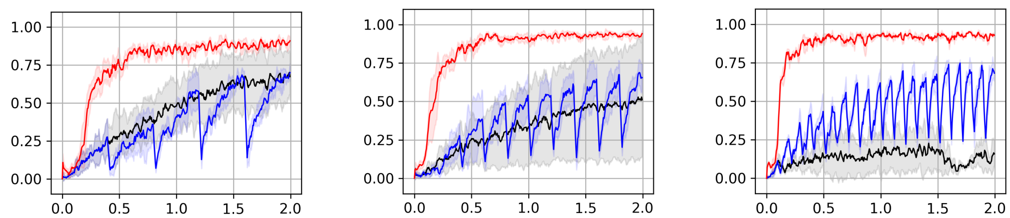
(g) swimmer-swimmer15(1) (h) swimmer-swimmer15(2) (i) swimmer-swimmer15(4)

(j) walker-run(1) (k) walker-run(2) (l) walker-run(4)

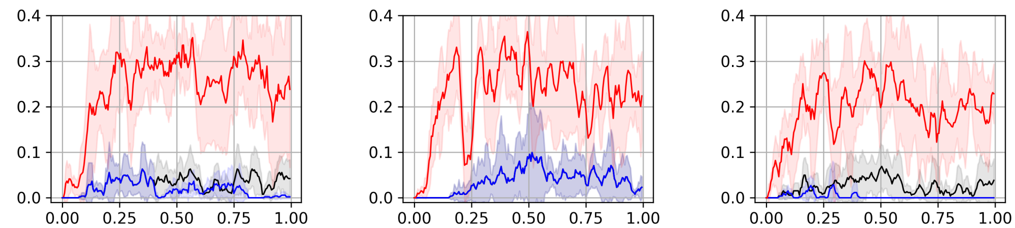
Figure 8: Per-environment performance in DMC with varying replay ratio values. Note that the number in parentheses indicates the replay ratio. The scale of the x-axis is 10^6. Performances are averaged over 5 seeds.



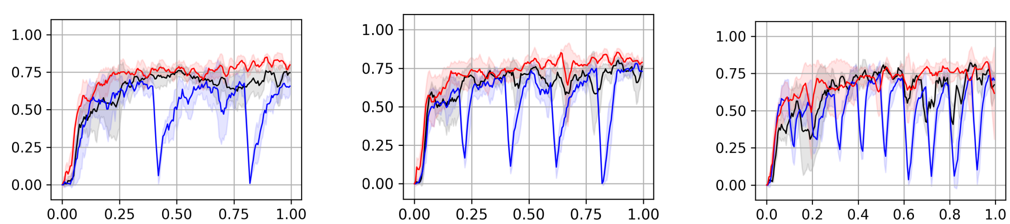
(a) FourRoom(0.5) (b) FourRoom(1) (c) FourRoom(2)



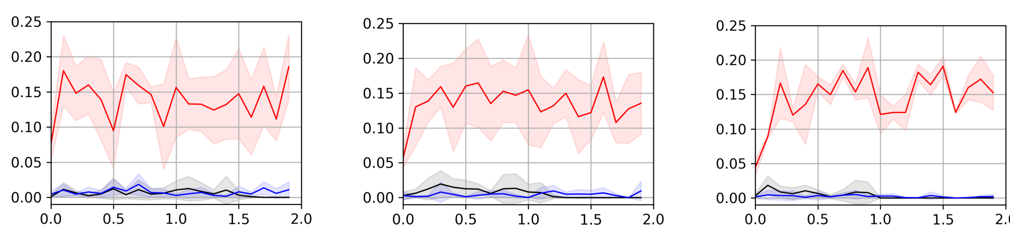
(d) Gotodoor(0.5) (e) Gotodoor(1) (f) Gotodoor(2)



(g) Lavacrossing9(0.5) (h) Lavacrossing9(1) (i) Lavacrossing9(2)

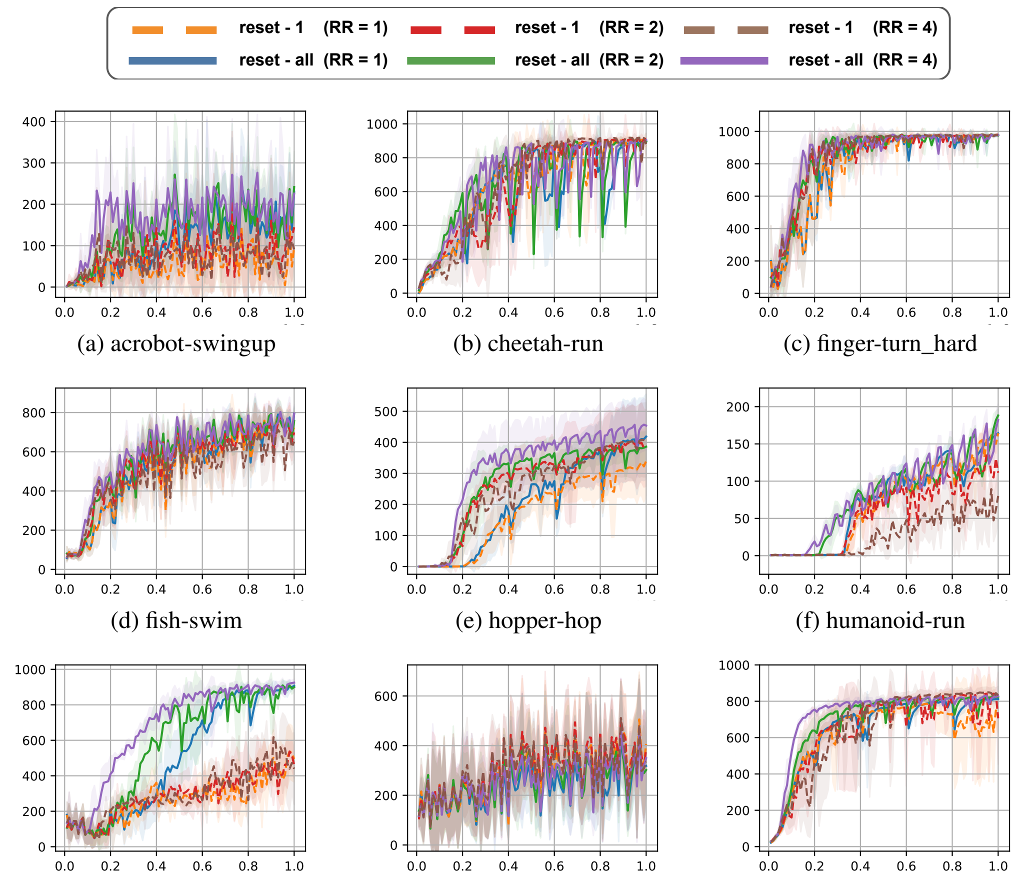


(j) Lavagap7(0.5) (k) Lavagap7(1) (l) Lavagap7(2)



(m) Simplecrossing(0.5) (n) Simplecrossing(1) (o) Simplecrossing(2)

Figure 9: Per-environment performance in minigird with varying replay ratio values. Note that the number in parentheses indicates the replay ratio. The scale of the x-axis is 10^6. Performances are averaged over 5 seeds.



(a) acrobot-swingup (b) cheetah-run (c) finger-turn_hard

(d) fish-swim (e) hopper-hop (f) humanoid-run

(g) quadruped-run (h) swimmer-swimmer15 (i) walker-run

Figure 10: Per-environment performance in DMC with varying replay ratio values. Note that the number in parentheses indicates the replay ratio. The scale of the x-axis is 10^6. Performances are averaged over 5 seeds.