A Experimental Details

In order to generate the results presented in Table 2 Table 3 and Table 4, we conducted a hyperparameter search and selected the best results from the final evaluations for each dataset. Our algorithm was implemented using JAX for the D4RL benchmark. For V-D4RL, we implement our approach using PyTorch adopting the TD3+BC implementation from Clean Offline RL (Tarasov et al., 2022). The experiments were conducted on V100 and A100 GPUs.

Gym-MuJoCo and Adroit tasks. Our study utilized the latest version of the datasets – v2 for Gym-MuJoCo and v1 for Adroit. The agents were trained for one million steps and evaluated over ten episodes.

For ReBRAC, we fine-tuned the β_1 parameter for the actor, which was selected from $0.001, 0.01, 0.05, 0.1$. Similarly, the β_2 parameter for the critic was selected from a range of $0, 0.001, 0.01, 0.1, 0.5$. The selected best parameters for each dataset are reported in Table 9.

For TD3+BC here and in the AntMaze domain, we use the same grid used in ReBRAC for actor regularization parameter α and add the default value of 0.4.

For IQL here and in the AntMaze domain, we selected β value from a range of $0.5, 1, 3, 6, 10$ and IQL τ value from a range of $0.5, 0.7, 0.9, 0.95$. We used the implementation from Clean Offline RL (Tarasov et al., 2022) and kept other parameters unchanged.

For SAC-RND in Adroit domain we tune β_1 (actor parameter) in the range of $0.5, 1.0, 2.5, 5.0, 10.0$ and β_2 (critic parameter) in the range of $0.01, 0.1, 1.0, 5.0, 10.0$.

AntMaze tasks. In our work, we utilized v2 of the datasets. It's worth noting that previous studies have reported results using v0 datasets, which were found to contain numerous issues[1]. Each agent was trained for 1 million steps and evaluated over 100 episodes. Following Chen et al. (2022a), we modified the reward function by multiplying it by 100.

For ReBRAC, the β_1 (actor) and β_2 (critic) hyperparameters were carefully selected from the respective ranges of $0.0005, 0.001, 0.002, 0.003$ and $0, 0.0001, 0.0005, 0.001$. In addition, the actor and critic learning rates were optimized from $0.0001, 0.0002, 0.0003, 0.0005$ and $0.0003, 0.0005, 0.001$, respectively. The optimal hyperparameters for each dataset are presented in Table 9.

We also modified the γ value for ReBRAC when addressing these tasks, driven by the following motivation. The length of the episodes in AntMaze can be as long as 1000 steps, while the reward is sparse and can only be obtained at the end of the episode. As a result, the discount for the reward with the default γ can be as low as $0.99^{1000} = 4 \cdot 10^{-5}$, which is extremely low for signal propagation, even when multiplying the reward by 100. By increasing γ to 0.999, the minimum discount value becomes $0.999^{1000} = 0.36$, which is more favorable for signal propagation.

V-D4RL. We used single-task datasets without distraction with a resolution of 84 × 84 pixels. For ReBRAC β_1 (actor) parameter was selected from the range of $\{0.03, 0.1, 0.3, 1.0\}$ and β_2 (critic) parameter from the range of $\{0.0, 0.001, 0.005, 0.01, 0.1\}$.

---

[1] https://github.com/Farama-Foundation/D4RL/issues/77

B Hyperparameters

B.1 ReBRAC

Table 8: ReBRAC's general hyperparameters.

Parameter	**Value**
optimizer	Adam Kingma & Ba (2014)
batch size	1024 on Gym-MuJoCo, 256 on other
learning rate (all networks)	1e-3 on Gym-MuJoCo, 3e-4 on Adroit and V-D4RL, 1e-4 on Antmaze
tau (τ)	5e-3
hidden dim (all networks)	256
num hidden layers (all networks)	3
gamma (γ)	0.999 on AntMaze, 0.99 on other
nonlinearity	ReLU

Table 9: ReBRAC's best hyperparameters used in D4RL benchmark.

Task Name	β_1 (actor)	β_2 (critic)
halfcheetah-random	0.001	0.1
halfcheetah-medium	0.001	0.01
halfcheetah-expert	0.01	0.01
halfcheetah-medium-expert	0.01	0.1
halfcheetah-medium-replay	0.01	0.001
halfcheetah-full-replay	0.001	0.1
hopper-random	0.001	0.01
hopper-medium	0.01	0.001
hopper-expert	0.1	0.001
hopper-medium-expert	0.1	0.01
hopper-medium-replay	0.05	0.5
hopper-full-replay	0.01	0.01
walker2d-random	0.01	0.0
walker2d-medium	0.05	0.1
walker2d-expert	0.01	0.5
walker2d-medium-expert	0.01	0.01
walker2d-medium-replay	0.05	0.01
walker2d-full-replay	0.01	0.01
antmaze-umaze	0.003	0.002
antmaze-umaze-diverse	0.003	0.001
antmaze-medium-play	0.001	0.0005
antmaze-medium-diverse	0.001	0.0
antmaze-large-play	0.002	0.001
antmaze-large-diverse	0.002	0.002
pen-human	0.1	0.5
pen-cloned	0.05	0.5
pen-expert	0.01	0.01
door-human	0.1	0.1
door-cloned	0.01	0.1
door-expert	0.05	0.01
hammer-human	0.01	0.5
hammer-cloned	0.1	0.5
hammer-expert	0.01	0.01
relocate-human	0.1	0.01
relocate-cloned	0.1	0.01
relocate-expert	0.05	0.01

Table 10: ReBRAC's best hyperparameters used in V-D4RL benchmark.

Task Name	β_1 (actor)	β_2 (critic)
walker-walk-random	0.03	0.1
walker-walk-medium	0.03	0.005
walker-walk-expert	0.1	0.01
walker-walk-medium-expert	0.3	0.005
walker-walk-medium-replay	0.3	0.01
cheetah-run-random	0.1	0.01
cheetah-run-medium	0.1	0.1
cheetah-run-expert	0.01	0.01
cheetah-run-medium-expert	1.0	0.001
cheetah-run-medium-replay	0.03	0.1
humanoid-walk-random	1.0	0.01
humanoid-walk-medium	1.0	0.005
humanoid-walk-expert	1.0	0.1
humanoid-walk-medium-expert	1.0	0.005
humanoid-walk-medium-replay	1.0	0.001

B.2 IQL

Table 11: IQL's best hyperparameters used in D4RL benchmark.

Task Name	β	IQL τ
halfcheetah-random	3.0	0.95
halfcheetah-medium	3.0	0.95
halfcheetah-expert	6.0	0.9
halfcheetah-medium-expert	3.0	0.7
halfcheetah-medium-replay	3.0	0.95
halfcheetah-full-replay	1.0	0.7
hopper-random	1.0	0.95
hopper-medium	3.0	0.7
hopper-expert	3.0	0.5
hopper-medium-expert	6.0	0.7
hopper-medium-replay	6.0	0.7
hopper-full-replay	10.0	0.9
walker2d-random	0.5	0.9
walker2d-medium	6.0	0.5
walker2d-expert	6.0	0.9
walker2d-medium-expert	1.0	0.5
walker2d-medium-replay	0.5	0.7
walker2d-full-replay	1.0	0.7
antmaze-umaze	10.0	0.7
antmaze-umaze-diverse	10.0	0.95
antmaze-medium-play	6.0	0.9
antmaze-medium-diverse	6.0	0.9
antmaze-large-play	10.0	0.9
antmaze-large-diverse	6.0	0.9
pen-human	1.0	0.95
pen-cloned	10.0	0.9
pen-expert	10.0	0.8
door-human	0.5	0.9
door-cloned	6.0	0.7
door-expert	0.5	0.7
hammer-human	3.0	0.9
hammer-cloned	6.0	0.7
hammer-expert	0.5	0.95
relocate-human	1.0	0.95
relocate-cloned	6.0	0.9
relocate-expert	10.0	0.9

B.3 TD3+BC

Table 12: TD3+BC's best hyperparameters used in D4RL benchmark.

Task Name	α
halfcheetah-random	0.001
halfcheetah-medium	0.01
halfcheetah-expert	0.4
halfcheetah-medium-expert	0.1
halfcheetah-medium-replay	0.05
halfcheetah-full-replay	0.01
hopper-random	0.4
hopper-medium	0.05
hopper-expert	0.1
hopper-medium-expert	0.1
hopper-medium-replay	0.4
hopper-full-replay	0.01
walker2d-random	0.001
walker2d-medium	0.4
walker2d-expert	0.05
walker2d-medium-expert	0.1
walker2d-medium-replay	0.1
walker2d-full-replay	0.1
antmaze-umaze	0.4
antmaze-umaze-diverse	0.4
antmaze-medium-play	0.003
antmaze-medium-diverse	0.003
antmaze-large-play	0.003
antmaze-large-diverse	0.003
pen-human	0.1
pen-cloned	0.4
pen-expert	0.4
door-human	0.1
door-cloned	0.4
door-expert	0.1
hammer-human	0.4
hammer-cloned	0.4
hammer-expert	0.4
relocate-human	0.1
relocate-cloned	0.1
relocate-expert	0.4

B.4 SAC-RND

Table 13: SAC-RND's best hyperparameters used in D4RL Adroit tasks.

Task Name	β_1 (actor)	β_2 (critic)
pen-human	1.0	10.0
pen-cloned	2.5	0.01
pen-expert	10.0	5.0
door-human	5.0	0.01
door-cloned	5.0	1.0
door-expert	10.0	10.0
hammer-human	10.0	0.01
hammer-cloned	1.0	1.0
hammer-expert	2.5	10.0
relocate-human	5.0	0.01
relocate-cloned	5.0	1.0
relocate-expert	10.0	10.0

C Comparison to Ensemble-based Methods

Comparison of ReBRAC with the ensemble-based methods is presented in Table 14, Table 15, and Table 16. We add the following ensemble-based methods: RORL for each domain (Yang et al., 2022), SAC-N/EDAC (An et al., 2021) for the Gym-MuJoCo and Adroit tasks[2] and MSG (Ghasemipour et al., 2022) for AntMaze tasks[3]. The mean-wise best results among algorithms are highlighted with **bold**, and the second-best performance is underlined. Our approach, ReBRAC, shows competitive results on the Gym-MuJoCo datasets. On AntMaze tasks, ReBRAC achieves state-of-the-art results among ensemble-free algorithms and a good score compared to ensemble-based algorithms. And on Adroit tasks, our approach outperforms both families of algorithms.

Table 14: ReBRAC evaluation on the Gym domain. We report the final normalized score averaged over 10 unseen training seeds on v2 datasets. CQL, SAC-N and EDAC scores are taken from An et al. (2021). RORL scores are taken from Yang et al. (2022).

	Ensemble-free				Ensemble-based			
Task Name	**TD3+BC**	**IQL**	**CQL**	**SAC-RND**	**SAC-N**	**EDAC**	**RORL**	**ReBRAC, our**
halfcheetah-random	30.9 ± 0.4	19.5 ± 0.8	**31.1** ± 3.5	27.6 ± 2.1	28.0 ± 0.9	28.4 ± 1.0	28.5 ± 0.8	29.5 ± 1.5
halfcheetah-medium	54.7 ± 0.9	50.0 ± 0.2	46.9 ± 0.4	66.4 ± 1.4	67.5 ± 1.2	65.9 ± 0.6	**66.8** ± 0.7	65.6 ± 1.0
halfcheetah-expert	93.4 ± 0.4	95.5 ± 2.1	97.3 ± 1.1	102.6 ± 4.2	105.2 ± 2.6	**106.8** ± 3.4	105.2 ± 0.7	**105.9** ± 1.7
halfcheetah-medium-expert	89.1 ± 5.6	92.7 ± 2.8	95.0 ± 1.4	**108.1** ± 1.5	107.1 ± 2.0	106.3 ± 1.9	107.8 ± 1.1	101.1 ± 5.2
halfcheetah-medium-replay	45.0 ± 1.1	42.1 ± 3.6	45.3 ± 0.3	51.2 ± 3.2	**63.9** ± 0.8	61.3 ± 1.9	61.9 ± 1.5	51.0 ± 0.8
halfcheetah-full-replay	75.0 ± 2.5	75.0 ± 0.7	76.9 ± 0.9	81.2 ± 1.3	84.5 ± 1.2	**84.6** ± 0.9	-	82.1 ± 1.1
hopper-random	8.5 ± 0.6	10.1 ± 5.9	5.3 ± 0.6	19.6 ± 12.4	31.3 ± 0.0	25.3 ± 10.4	**31.4** ± 0.1	8.1 ± 2.4
hopper-medium	60.9 ± 7.6	65.2 ± 4.2	61.9 ± 6.4	91.1 ± 10.1	100.3 ± 0.3	101.6 ± 0.6	**104.8** ± 0.1	102.0 ± 1.0
hopper-expert	109.6 ± 3.7	108.8 ± 3.1	106.5 ± 9.1	109.8 ± 0.5	110.3 ± 0.3	110.1 ± 0.1	**112.8** ± 0.2	100.1 ± 8.3
hopper-medium-expert	87.8 ± 10.5	85.5 ± 29.7	96.9 ± 15.1	109.8 ± 0.6	110.1 ± 0.3	110.7 ± 0.1	**112.7** ± 0.2	107.0 ± 6.4
hopper-medium-replay	55.1 ± 31.7	89.6 ± 13.2	86.3 ± 7.3	97.2 ± 9.0	101.8 ± 0.5	101.0 ± 0.5	**102.8** ± 0.5	98.1 ± 5.3
hopper-full-replay	97.9 ± 17.5	104.4 ± 10.8	101.9 ± 0.6	**107.4** ± 0.8	102.9 ± 0.3	105.4 ± 0.7	-	107.1 ± 0.4
walker2d-random	2.0 ± 3.6	11.3 ± 7.0	5.1 ± 1.7	18.7 ± 6.9	**21.7** ± 0.0	16.6 ± 7.0	21.4 ± 0.2	18.1 ± 4.5
walker2d-medium	77.7 ± 2.9	80.7 ± 3.4	79.5 ± 3.2	92.7 ± 1.2	87.9 ± 0.2	92.5 ± 0.8	**102.4** ± 1.4	82.5 ± 3.6
walker2d-expert	110.0 ± 0.6	96.9 ± 32.3	109.3 ± 0.1	104.5 ± 22.8	107.4 ± 2.4	115.1 ± 1.9	**115.4** ± 0.5	112.3 ± 0.2
walker2d-medium-expert	110.4 ± 0.6	112.1 ± 0.5	109.1 ± 0.2	104.6 ± 11.2	116.7 ± 0.4	114.7 ± 0.9	**121.2** ± 1.5	111.6 ± 0.3
walker2d-medium-replay	68.0 ± 19.2	75.4 ± 9.3	76.8 ± 10.0	89.4 ± 3.8	78.7 ± 0.7	87.1 ± 2.4	**90.4** ± 0.5	77.3 ± 7.9
walker2d-full-replay	90.3 ± 5.4	97.5 ± 1.4	94.2 ± 1.9	**105.3** ± 3.2	94.6 ± 0.5	99.8 ± 0.7	-	102.2 ± 1.7
Average w/o full-replay	66.8	70.1	70.1	79.5	82.4	**82.9**	**85.7**	78.0
Average	70.3	72.9	73.6	82.6	84.4	**85.2**	-	81.2

---

[2] SAC-N and EDAC score 0 on medium and large AntMaze tasks (Tarasov et al., 2022).

[3] MSG numerical results are not available for Gym-MuJoCo tasks and Adroit tasks were not benchmarked.

Table 15: ReBRAC evaluation on AntMaze domain. We report the final normalized score averaged over 10 unseen training seeds on v2 datasets. CQL scores are taken from Ghasemipour et al. (2022). RORL scores are taken from Yang et al. (2022).

	Ensemble-free				Ensemble-based		
Task Name	**TD3+BC**	**IQL**	**CQL**	**SAC-RND**	**RORL**	**MSG**	**ReBRAC, our**
antmaze-umaze	66.3 ± 6.2	83.3 ± 4.5	74.0	97.0 ± 1.5	97.7 ± 1.9	**97.9** ± 1.3	97.8 ± 1.0
antmaze-umaze-diverse	53.8 ± 8.5	70.6 ± 3.7	84.0	66.0 ± 25.0	**90.7** ± 2.9	79.3 ± 3.0	88.3 ± 13.0
antmaze-medium-play	26.5 ± 18.4	64.6 ± 4.9	61.2	38.5 ± 29.4	76.3 ± 2.5	**85.9** ± 3.9	84.0 ± 4.2
antmaze-medium-diverse	25.9 ± 15.3	61.7 ± 6.1	53.7	74.7 ± 10.7	69.3 ± 3.3	**84.6** ± 5.2	76.3 ± 13.5
antmaze-large-play	0.0 ± 0.0	42.5 ± 6.5	15.8	43.9 ± 29.2	16.3 ± 11.1	**64.3** ± 12.7	60.4 ± 26.1
antmaze-large-diverse	0.0 ± 0.0	27.6 ± 7.8	14.9	45.7 ± 28.5	41.0 ± 10.7	**71.3** ± 5.3	54.4 ± 25.1
Average	28.7	58.3	50.6	60.9	65.2	**80.5**	76.8

Table 16: ReBRAC evaluation on Adroit domain. We report the final normalized score averaged over 10 unseen training seeds on v1 datasets. BC, CQL, EDAC and RORL scores are taken from Yang et al. (2022).

	Ensemble-free					Ensemble-based		
Task Name	**BC**	**TD3+BC**	**IQL**	**CQL**	**SAC-RND**	**RORL**	**EDAC**	**ReBRAC, our**
pen-human	34.4	81.8 ± 14.9	81.5 ± 17.5	37.5	5.6 ± 5.8	33.7 ± 7.6	51.2 ± 8.6	**103.5** ± 14.1
pen-cloned	56.9	61.4 ± 19.3	77.2 ± 17.7	39.2	2.5 ± 6.1	35.7 ± 35.7	68.2 ± 7.3	**91.8** ± 21.7
pen-expert	85.1	146.0 ± 7.3	133.6 ± 16.0	107.0	45.4 ± 22.9	130.3 ± 4.2	122.8 ± 14.1	**154.1** ± 5.4
door-human	0.5	-0.1 ± 0.0	3.1 ± 2.0	9.9	0.0 ± 0.0	3.7 ± 0.7	**10.7** ± 6.8	0.0 ± 0.1
door-cloned	-0.1	0.1 ± 0.6	0.8 ± 1.0	0.4	0.2 ± 0.8	-0.1 ± 0.1	**9.6** ± 8.3	1.1 ± 2.6
door-expert	34.9	84.6 ± 44.5	**105.3** ± 2.8	101.5	73.6 ± 26.7	104.9 ± 0.9	-0.3 ± 0.1	104.6 ± 2.4
hammer-human	1.5	0.4 ± 0.4	2.5 ± 1.9	**4.4**	-0.1 ± 0.1	2.3 ± 2.3	0.8 ± 0.4	0.2 ± 0.2
hammer-cloned	0.8	0.8 ± 0.7	1.1 ± 0.5	2.1	0.1 ± 0.4	1.7 ± 1.7	0.3 ± 0.0	**6.7** ± 3.7
hammer-expert	125.6	117.0 ± 30.9	129.6 ± 0.5	86.7	24.8 ± 39.4	132.2 ± 0.7	0.2 ± 0.0	**133.8** ± 0.7
relocate-human	0.0	-0.2 ± 0.0	0.1 ± 0.1	**0.2**	0.0 ± 0.0	0.0 ± 0.0	0.1 ± 0.1	0.0 ± 0.0
relocate-cloned	-0.1	-0.1 ± 0.1	0.2 ± 0.4	-0.1	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	**0.9** ± 1.6
relocate-expert	101.3	**107.3** ± 1.6	106.5 ± 2.5	95.0	3.4 ± 4.5	47.8 ± 13.5	-0.3 ± 0.0	106.6 ± 3.2
Average w/o expert	11.7	18.0	20.8	11.7	1.0	9.6	17.4	**25.5**
Average	36.7	49.9	53.4	40.3	12.9	41.0	21.9	**58.6**

D Computational costs

Table 17: Computational costs for algorithms in Table 2.

Algorithm	**Number of runs**	**Approximate hours per run**
TD3+BC, tuning	360	0.3
IQL, tuning	1440	1.8
ReBRAC, tuning	1440	0.4
TD3+BC, eval	180	0.2
IQL, eval	180	1.8
SAC-RND, eval	180	1.8
ReBRAC, eval	180	0.4
Sum	3960	4032.0

Table 18: Computational costs for algorithms in Table 3 and Table 15.

Algorithm	**Number of runs**	**Approximate hours per run**
TD3+BC, tuning	96	0.5
IQL, tuning	480	2.1
ReBRAC, tuning	384	0.6
TD3+BC, eval	60	0.5
IQL, eval	60	2.0
SAC-RND, eval	60	2.9
MSG, eval	60	5.1
ReBRAC, eval	60	0.4
Sum	1260	1940.4

Table 19: Computational costs for algorithms in Table 4.

Algorithm	**Number of runs**	**Approximate hours per run**
TD3+BC, tuning	240	0.3
IQL, tuning	960	1.8
SAC-RND, tuning	1200	1.1
ReBRAC, tuning	960	0.3
TD3+BC, eval	120	0.2
IQL, eval	120	1.9
SAC-RND, eval	120	1.1
ReBRAC, eval	120	0.3
Sum	3840	3828.0

Table 20: Computational costs for algorithms in Table 5.

Algorithm	**Number of runs**	**Approximate hours per run**
ReBRAC, tuning	600	10.6
ReBRAC, eval	75	10.5
Sum	675	7147.5

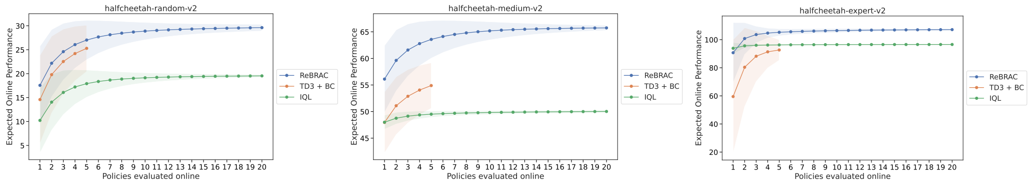
Table 21: Computational costs for algorithms in Table 6 and Figure 2.

Algorithm	**Number of runs**	**Approximate hours per run**
ReBRAC, ablations eval	1104	1.4
Sum	1104	1545.6

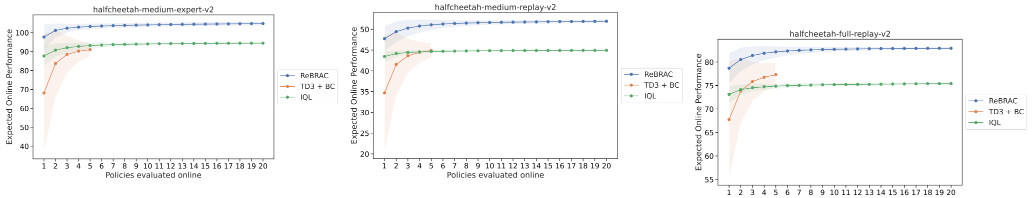
E Expected Online Performance

Table 22: TD3+BC, IQL and ReBRAC Expected Online Performance under uniform policy selection on HalfCheetah tasks.

	random			medium			expert			medium-expert			medium-replay			full-replay		
Policies	**TD3+BC**	**IQL**	**ReBRAC**	**TD3+BC**	**IQL**	**ReBRAC**	**TD3+BC**	**IQL**	**ReBRAC**	**TD3+BC**	**IQL**	**ReBRAC**	**TD3+BC**	**IQL**	**ReBRAC**	**TD3+BC**	**IQL**	**ReBRAC**
1	14.6 ± 9.3	10.2 ± 6.8	**17.6 ± 8.2**	48.0 ± 5.8	48.0 ± 1.3	**56.1 ± 6.3**	59.5 ± 40.5	**93.9 ± 4.2**	90.7 ± 21.5	68.1 ± 31.4	87.7 ± 5.5	**97.7 ± 6.8**	34.7 ± 14.2	43.4 ± 1.3	**47.7 ± 3.0**	67.7 ± 12.5	73.1 ± 1.9	**78.7 ± 3.3**
2	19.8 ± 8.0	14.1 ± 5.9	**22.2 ± 7.0**	51.1 ± 5.4	48.8 ± 1.1	**59.6 ± 5.8**	80.4 ± 28.1	95.6 ± 1.5	**100.8 ± 11.4**	83.7 ± 18.1	90.8 ± 3.7	**101.2 ± 3.8**	41.5 ± 7.7	44.2 ± 0.8	**49.4 ± 2.6**	73.8 ± 7.0	74.1 ± 1.1	**80.5 ± 2.6**
3	22.5 ± 6.8	16.1 ± 4.6	**24.6 ± 5.9**	52.9 ± 5.0	49.1 ± 0.9	**61.6 ± 4.9**	88.2 ± 18.1	96.0 ± 0.8	**103.6 ± 6.0**	88.5 ± 10.1	92.0 ± 2.7	**102.4 ± 2.4**	43.6 ± 4.2	44.4 ± 0.6	**50.3 ± 2.1**	75.8 ± 4.2	74.5 ± 0.8	**81.4 ± 2.0**
4	24.2 ± 5.7	17.2 ± 3.6	**26.1 ± 4.9**	54.0 ± 4.7	49.4 ± 0.7	**62.8 ± 4.2**	91.3 ± 11.5	96.2 ± 0.5	**104.7 ± 3.5**	90.3 ± 5.9	92.7 ± 2.2	**102.9 ± 1.9**	44.5 ± 2.6	44.6 ± 0.4	**50.8 ± 1.6**	76.7 ± 3.1	74.7 ± 0.7	**81.9 ± 1.6**
5	25.3 ± 4.8	17.9 ± 2.8	**27.0 ± 4.1**	54.9 ± 4.3	49.5 ± 0.6	**63.6 ± 3.5**	92.7 ± 7.3	96.3 ± 0.4	**105.2 ± 2.5**	91.0 ± 3.6	93.2 ± 1.8	**103.3 ± 1.6**	44.9 ± 1.8	44.6 ± 0.4	**51.1 ± 1.3**	77.3 ± 2.5	74.9 ± 0.6	**82.2 ± 1.2**
6	-	18.4 ± 2.3	**27.7 ± 3.4**	-	49.6 ± 0.5	**64.1 ± 3.0**	-	96.3 ± 0.3	**105.6 ± 2.1**	-	93.5 ± 1.5	**103.5 ± 1.5**	-	44.7 ± 0.3	**51.3 ± 1.1**	-	75.0 ± 0.5	**82.3 ± 1.0**
7	-	18.7 ± 1.8	**28.1 ± 2.9**	-	49.7 ± 0.5	**64.5 ± 2.6**	-	96.4 ± 0.3	**105.9 ± 1.9**	-	93.7 ± 1.3	**103.7 ± 1.4**	-	44.8 ± 0.3	**51.4 ± 0.9**	-	75.0 ± 0.5	**82.5 ± 0.8**
8	-	18.9 ± 1.5	**28.5 ± 2.5**	-	49.7 ± 0.4	**64.8 ± 2.3**	-	96.4 ± 0.3	**106.1 ± 1.7**	-	93.8 ± 1.1	**103.9 ± 1.4**	-	44.8 ± 0.2	**51.5 ± 0.7**	-	75.1 ± 0.5	**82.6 ± 0.7**
9	-	19.0 ± 1.3	**28.7 ± 2.1**	-	49.8 ± 0.4	**65.0 ± 2.0**	-	96.4 ± 0.2	**106.3 ± 1.6**	-	93.9 ± 1.0	**104.0 ± 1.3**	-	44.8 ± 0.2	**51.6 ± 0.7**	-	75.1 ± 0.4	**82.6 ± 0.6**
10	-	19.1 ± 1.1	**28.9 ± 1.8**	-	49.8 ± 0.4	**65.2 ± 1.7**	-	96.5 ± 0.2	**106.4 ± 1.5**	-	94.0 ± 0.9	**104.1 ± 1.3**	-	44.8 ± 0.2	**51.6 ± 0.6**	-	75.2 ± 0.4	**82.7 ± 0.6**
11	-	19.2 ± 0.9	**29.0 ± 1.6**	-	49.9 ± 0.4	**65.3 ± 1.5**	-	96.5 ± 0.2	**106.6 ± 1.3**	-	94.1 ± 0.8	**104.2 ± 1.3**	-	44.8 ± 0.2	**51.7 ± 0.6**	-	75.2 ± 0.4	**82.7 ± 0.5**
12	-	19.3 ± 0.8	**29.2 ± 1.4**	-	49.9 ± 0.3	**65.4 ± 1.3**	-	96.5 ± 0.2	**106.7 ± 1.2**	-	94.2 ± 0.7	**104.3 ± 1.3**	-	44.9 ± 0.2	**51.7 ± 0.5**	-	75.2 ± 0.4	**82.8 ± 0.4**
13	-	19.3 ± 0.7	**29.2 ± 1.2**	-	49.9 ± 0.3	**65.5 ± 1.1**	-	96.5 ± 0.2	**106.8 ± 1.1**	-	94.2 ± 0.7	**104.4 ± 1.3**	-	44.9 ± 0.2	**51.7 ± 0.5**	-	75.3 ± 0.4	**82.8 ± 0.4**
14	-	19.4 ± 0.6	**29.3 ± 1.1**	-	49.9 ± 0.3	**65.5 ± 1.0**	-	96.5 ± 0.2	**106.8 ± 1.1**	-	94.3 ± 0.6	**104.4 ± 1.2**	-	44.9 ± 0.1	**51.8 ± 0.5**	-	75.3 ± 0.3	**82.8 ± 0.4**
15	-	19.4 ± 0.5	**29.4 ± 1.0**	-	49.9 ± 0.3	**65.6 ± 0.9**	-	96.5 ± 0.2	**106.9 ± 1.0**	-	94.3 ± 0.6	**104.5 ± 1.2**	-	44.9 ± 0.1	**51.8 ± 0.5**	-	75.3 ± 0.3	**82.9 ± 0.3**
16	-	19.5 ± 0.5	**29.4 ± 0.9**	-	50.0 ± 0.3	**65.6 ± 0.8**	-	96.5 ± 0.1	**107.0 ± 0.9**	-	94.4 ± 0.5	**104.6 ± 1.2**	-	44.9 ± 0.1	**51.8 ± 0.5**	-	75.3 ± 0.3	**82.9 ± 0.3**
17	-	19.5 ± 0.4	**29.5 ± 0.9**	-	50.0 ± 0.3	**65.7 ± 0.7**	-	96.6 ± 0.1	**107.0 ± 0.8**	-	94.4 ± 0.5	**104.6 ± 1.2**	-	44.9 ± 0.1	**51.9 ± 0.5**	-	75.3 ± 0.3	**82.9 ± 0.3**
18	-	19.5 ± 0.4	**29.5 ± 0.8**	-	50.0 ± 0.3	**65.7 ± 0.6**	-	96.6 ± 0.1	**107.1 ± 0.8**	-	94.4 ± 0.5	**104.7 ± 1.2**	-	44.9 ± 0.1	**51.9 ± 0.5**	-	75.4 ± 0.3	**82.9 ± 0.2**
19	-	19.5 ± 0.4	**29.6 ± 0.8**	-	50.0 ± 0.3	**65.7 ± 0.5**	-	96.6 ± 0.1	**107.1 ± 0.7**	-	94.4 ± 0.5	**104.7 ± 1.1**	-	44.9 ± 0.1	**51.9 ± 0.4**	-	75.4 ± 0.3	**82.9 ± 0.2**
20	-	19.5 ± 0.3	**29.6 ± 0.7**	-	50.0 ± 0.3	**65.7 ± 0.5**	-	96.6 ± 0.1	**107.1 ± 0.7**	-	94.5 ± 0.4	**104.8 ± 1.1**	-	44.9 ± 0.1	**51.9 ± 0.4**	-	75.4 ± 0.3	**82.9 ± 0.2**



(a) halfcheetah-random EOP. (b) halfcheetah-medium EOP. (c) halfcheetah-expert EOP.

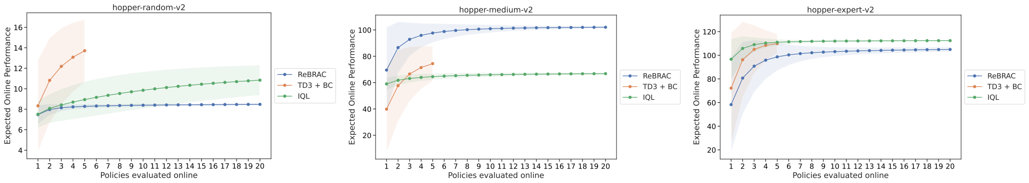


(d) halfcheetah-medium-expert EOP. (e) halfcheetah-medium-replay EOP. (f) halfcheetah-full-replay EOP.

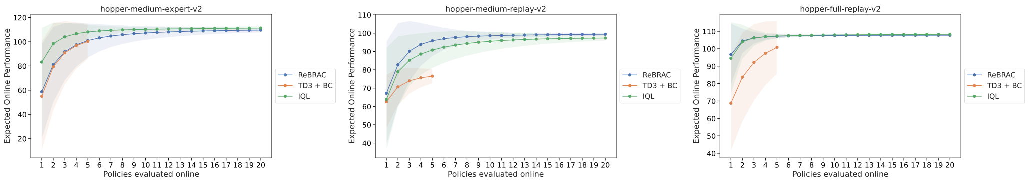
Figure 3: TD3+BC, IQL and ReBRAC visualised Expected Online Performance under uniform policy selection on HalfCheetah tasks.

Table 23: TD3+BC, IQL and ReBRAC Expected Online Performance under uniform policy selection on Hopper tasks.

	random			medium			expert			medium-expert			medium-replay			full-replay		
Policies	**TD3+BC**	**IQL**	**ReBRAC**	**TD3+BC**	**IQL**	**ReBRAC**	**TD3+BC**	**IQL**	**ReBRAC**	**TD3+BC**	**IQL**	**ReBRAC**	**TD3+BC**	**IQL**	**ReBRAC**	**TD3+BC**	**IQL**	**ReBRAC**
1	**8.3 ± 4.5**	7.5 ± 1.3	7.5 ± 0.9	39.8 ± 33.0	59.0 ± 4.9	**69.5 ± 32.6**	72.2 ± 47.1	**96.6 ± 17.3**	58.3 ± 40.5	55.0 ± 45.5	**83.3 ± 28.4**	58.7 ± 39.8	62.5 ± 14.9	63.8 ± 28.3	**67.2 ± 28.4**	68.7 ± 27.6	94.5 ± 20.8	**96.7 ± 17.8**
2	**10.8 ± 4.1**	8.1 ± 1.4	8.0 ± 0.5	57.7 ± 27.1	61.8 ± 3.8	**86.6 ± 19.7**	96.2 ± 32.3	**105.8 ± 10.5**	80.7 ± 30.8	79.3 ± 36.3	**98.5 ± 17.5**	81.1 ± 31.4	70.7 ± 10.2	78.9 ± 19.5	**82.7 ± 22.8**	83.7 ± 25.7	104.0 ± 9.9	**104.5 ± 7.2**
3	**12.2 ± 3.7**	8.4 ± 1.6	8.1 ± 0.4	66.5 ± 20.8	63.1 ± 3.2	**92.8 ± 12.8**	105.0 ± 20.6	**109.0 ± 6.4**	90.8 ± 22.1	90.9 ± 26.5	**104.1 ± 11.8**	91.8 ± 23.8	74.0 ± 7.1	85.2 ± 14.2	**90.1 ± 16.6**	92.2 ± 22.0	106.2 ± 4.7	**106.2 ± 3.4**
4	**13.1 ± 3.4**	8.7 ± 1.7	8.2 ± 0.3	71.4 ± 16.3	63.9 ± 2.9	**95.9 ± 9.3**	108.4 ± 13.0	**110.3 ± 4.1**	95.9 ± 16.0	96.8 ± 19.3	**106.7 ± 8.3**	97.5 ± 18.2	75.6 ± 5.3	88.6 ± 11.5	**93.9 ± 11.8**	97.4 ± 18.4	**107.0 ± 2.5**	106.9 ± 1.9
5	**13.7 ± 3.1**	8.9 ± 1.8	8.3 ± 0.2	74.4 ± 13.3	64.5 ± 2.6	**97.7 ± 7.3**	109.8 ± 8.2	**111.0 ± 2.7**	98.7 ± 11.8	100.2 ± 14.6	**108.1 ± 6.1**	101.0 ± 14.3	76.5 ± 4.0	90.8 ± 9.9	**95.8 ± 8.4**	100.8 ± 15.3	**107.3 ± 1.5**	107.2 ± 1.3
6	-	**9.2 ± 1.8**	8.3 ± 0.2	-	64.9 ± 2.4	**98.8 ± 5.9**	-	**111.4 ± 2.0**	100.3 ± 8.9	-	**108.9 ± 4.5**	103.2 ± 11.4	-	92.3 ± 8.7	**97.0 ± 6.1**	-	**107.5 ± 1.1**	107.4 ± 0.9
7	-	**9.4 ± 1.8**	8.3 ± 0.2	-	65.3 ± 2.1	**99.6 ± 4.9**	-	**111.6 ± 1.5**	101.4 ± 6.9	-	**109.5 ± 3.5**	104.8 ± 9.4	-	93.5 ± 7.7	**97.6 ± 4.5**	-	**107.6 ± 0.9**	107.5 ± 0.7
8	-	**9.5 ± 1.8**	8.4 ± 0.2	-	65.5 ± 2.0	**100.2 ± 4.2**	-	**111.8 ± 1.2**	102.1 ± 5.5	-	**109.8 ± 2.8**	105.8 ± 7.8	-	94.4 ± 6.8	**98.1 ± 3.4**	-	**107.8 ± 0.8**	107.5 ± 0.5
9	-	**9.7 ± 1.8**	8.4 ± 0.2	-	65.8 ± 1.8	**100.6 ± 3.6**	-	**111.9 ± 1.0**	102.7 ± 4.6	-	**110.1 ± 2.4**	106.6 ± 6.7	-	95.0 ± 6.0	**98.4 ± 2.6**	-	**107.8 ± 0.7**	107.6 ± 0.4
10	-	**9.9 ± 1.8**	8.4 ± 0.2	-	65.9 ± 1.7	**100.9 ± 3.1**	-	**112.0 ± 0.9**	103.1 ± 3.9	-	**110.3 ± 2.2**	107.3 ± 5.8	-	95.6 ± 5.3	**98.6 ± 2.1**	-	**107.9 ± 0.7**	107.6 ± 0.4
11	-	**10.0 ± 1.8**	8.4 ± 0.1	-	66.1 ± 1.6	**101.2 ± 2.7**	-	**112.0 ± 0.8**	103.4 ± 3.5	-	**110.4 ± 2.0**	107.7 ± 5.1	-	96.0 ± 4.7	**98.7 ± 1.8**	-	**108.0 ± 0.6**	107.6 ± 0.3
12	-	**10.1 ± 1.8**	8.4 ± 0.1	-	66.2 ± 1.5	**101.3 ± 2.3**	-	**112.1 ± 0.7**	103.7 ± 3.2	-	**110.6 ± 1.9**	108.1 ± 4.5	-	96.3 ± 4.1	**98.9 ± 1.5**	-	**108.0 ± 0.6**	107.7 ± 0.3
13	-	**10.2 ± 1.7**	8.4 ± 0.1	-	66.3 ± 1.4	**101.5 ± 2.0**	-	**112.2 ± 0.7**	103.9 ± 2.9	-	**110.7 ± 1.8**	108.4 ± 4.0	-	96.5 ± 3.6	**99.0 ± 1.4**	-	**108.1 ± 0.5**	107.7 ± 0.2
14	-	**10.3 ± 1.7**	8.4 ± 0.1	-	66.4 ± 1.3	**101.6 ± 1.8**	-	**112.2 ± 0.6**	104.1 ± 2.8	-	**110.8 ± 1.8**	108.7 ± 3.6	-	96.7 ± 3.2	**99.1 ± 1.2**	-	**108.1 ± 0.5**	107.7 ± 0.2
15	-	**10.4 ± 1.7**	8.4 ± 0.1	-	66.5 ± 1.3	**101.7 ± 1.6**	-	**112.2 ± 0.6**	104.3 ± 2.6	-	**110.9 ± 1.7**	108.9 ± 3.3	-	96.9 ± 2.8	**99.1 ± 1.1**	-	**108.1 ± 0.5**	107.7 ± 0.2
16	-	**10.5 ± 1.6**	8.5 ± 0.1	-	66.6 ± 1.2	**101.8 ± 1.4**	-	**112.3 ± 0.6**	104.4 ± 2.5	-	**111.0 ± 1.7**	109.1 ± 3.0	-	97.0 ± 2.5	**99.2 ± 1.1**	-	**108.2 ± 0.4**	107.7 ± 0.1
17	-	**10.6 ± 1.6**	8.5 ± 0.1	-	66.6 ± 1.2	**101.9 ± 1.3**	-	**112.3 ± 0.5**	104.6 ± 2.4	-	**111.1 ± 1.7**	109.3 ± 2.7	-	97.1 ± 2.2	**99.3 ± 1.0**	-	**108.2 ± 0.4**	107.7 ± 0.1
18	-	**10.7 ± 1.5**	8.5 ± 0.1	-	66.7 ± 1.1	**101.9 ± 1.2**	-	**112.3 ± 0.5**	104.7 ± 2.3	-	**111.2 ± 1.7**	109.4 ± 2.5	-	97.2 ± 1.9	**99.3 ± 1.0**	-	**108.2 ± 0.4**	107.7 ± 0.1
19	-	**10.8 ± 1.5**	8.5 ± 0.1	-	66.8 ± 1.1	**102.0 ± 1.1**	-	**112.4 ± 0.5**	104.8 ± 2.2	-	**111.2 ± 1.6**	109.5 ± 2.3	-	97.3 ± 1.7	**99.4 ± 0.9**	-	**108.2 ± 0.4**	107.7 ± 0.1
20	-	**10.8 ± 1.5**	8.5 ± 0.1	-	66.8 ± 1.0	**102.0 ± 1.0**	-	**112.4 ± 0.5**	104.9 ± 2.1	-	**111.3 ± 1.6**	109.6 ± 2.1	-	97.3 ± 1.5	**99.4 ± 0.9**	-	**108.2 ± 0.3**	107.7 ± 0.1



(a) hopper-random EOP. (b) hopper-medium EOP. (c) hopper-expert EOP.

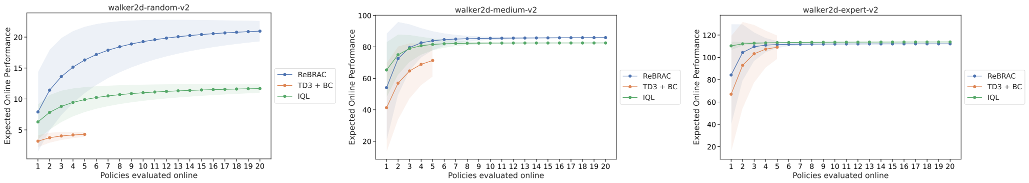


(d) hopper-medium-expert EOP. (e) hopper-medium-replay EOP. (f) hopper-full-replay EOP.

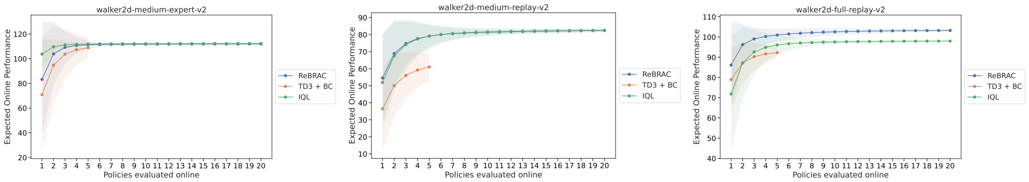
Figure 4: TD3+BC, IQL and ReBRAC visualised Expected Online Performance under uniform policy selection on Hopper tasks.

Table 24: TD3+BC, IQL and ReBRAC Expected Online Performance under uniform policy selection on Walker2d tasks.

	random			medium			expert			medium-expert			medium-replay			full-replay		
Policies	**TD3+BC**	**IQL**	**ReBRAC**	**TD3+BC**	**IQL**	**ReBRAC**	**TD3+BC**	**IQL**	**ReBRAC**	**TD3+BC**	**IQL**	**ReBRAC**	**TD3+BC**	**IQL**	**ReBRAC**	**TD3+BC**	**IQL**	**ReBRAC**
1	3.2 ± 1.0	6.3 ± 2.9	**7.9 ± 6.5**	41.3 ± 28.6	**65.3 ± 17.8**	54.1 ± 34.5	67.0 ± 52.4	**110.3 ± 4.0**	84.2 ± 45.7	70.9 ± 46.5	**103.8 ± 12.2**	83.2 ± 46.9	36.4 ± 25.2	51.9 ± 28.3	**54.5 ± 26.0**	78.9 ± 15.4	71.8 ± 28.7	**86.1 ± 21.5**
2	3.7 ± 0.8	7.8 ± 2.8	**11.4 ± 6.6**	57.0 ± 23.4	**75.0 ± 13.0**	72.5 ± 23.4	92.9 ± 39.0	**112.1 ± 1.8**	104.3 ± 25.4	94.8 ± 32.2	**109.6 ± 6.4**	103.8 ± 26.2	50.0 ± 19.0	67.5 ± 21.0	**68.8 ± 18.7**	87.1 ± 10.5	87.1 ± 17.9	**96.2 ± 10.2**
3	4.0 ± 0.7	8.8 ± 2.6	**13.6 ± 6.3**	64.7 ± 17.7	79.0 ± 8.7	**79.6 ± 14.9**	103.2 ± 26.0	**112.7 ± 1.2**	109.5 ± 13.0	103.7 ± 20.6	**111.2 ± 3.3**	109.2 ± 13.4	56.1 ± 13.8	74.2 ± 14.8	**74.8 ± 13.1**	90.3 ± 6.8	92.6 ± 11.1	**99.0 ± 5.7**
4	4.2 ± 0.6	9.4 ± 2.3	**15.1 ± 5.9**	68.9 ± 13.4	80.7 ± 5.8	**82.6 ± 9.5**	107.4 ± 16.8	**113.0 ± 1.0**	111.0 ± 6.6	107.3 ± 13.1	**111.7 ± 1.7**	110.7 ± 6.8	59.2 ± 10.5	77.4 ± 10.4	**77.6 ± 9.2**	91.6 ± 4.4	94.9 ± 7.2	**100.3 ± 3.9**
5	4.3 ± 0.4	9.9 ± 2.1	**16.3 ± 5.4**	71.3 ± 10.4	81.6 ± 3.9	**83.9 ± 6.2**	109.1 ± 10.7	**113.2 ± 0.8**	111.5 ± 3.4	108.8 ± 8.3	**111.9 ± 0.9**	111.2 ± 3.5	61.0 ± 8.5	79.1 ± 7.4	**79.1 ± 6.5**	92.3 ± 3.0	96.0 ± 4.8	**101.0 ± 3.2**
6	-	10.2 ± 1.9	**17.2 ± 5.0**	-	82.0 ± 2.7	**84.6 ± 4.1**	-	**113.3 ± 0.7**	111.7 ± 1.8	-	**112.0 ± 0.6**	111.4 ± 1.9	-	**80.1 ± 5.4**	80.0 ± 4.7	-	96.7 ± 3.4	**101.5 ± 2.8**
7	-	10.5 ± 1.7	**17.9 ± 4.6**	-	82.2 ± 1.9	**85.0 ± 2.8**	-	**113.4 ± 0.6**	111.8 ± 1.1	-	**112.1 ± 0.4**	111.5 ± 1.1	-	**80.7 ± 4.1**	80.5 ± 3.6	-	97.0 ± 2.5	**101.9 ± 2.5**
8	-	10.7 ± 1.6	**18.4 ± 4.2**	-	82.3 ± 1.3	**85.2 ± 2.0**	-	**113.5 ± 0.6**	111.9 ± 0.7	-	**112.1 ± 0.3**	111.6 ± 0.8	-	**81.1 ± 3.2**	80.9 ± 2.9	-	97.3 ± 1.9	**102.1 ± 2.3**
9	-	10.9 ± 1.4	**18.9 ± 3.8**	-	82.4 ± 0.9	**85.3 ± 1.5**	-	**113.6 ± 0.5**	111.9 ± 0.6	-	**112.2 ± 0.3**	111.7 ± 0.6	-	**81.4 ± 2.6**	81.2 ± 2.4	-	97.4 ± 1.5	**102.4 ± 2.0**
10	-	11.0 ± 1.3	**19.3 ± 3.5**	-	82.4 ± 0.7	**85.4 ± 1.2**	-	**113.6 ± 0.5**	112.0 ± 0.5	-	**112.2 ± 0.3**	111.7 ± 0.5	-	**81.6 ± 2.1**	81.4 ± 2.1	-	97.5 ± 1.2	**102.5 ± 1.8**
11	-	11.1 ± 1.3	**19.6 ± 3.2**	-	82.4 ± 0.5	**85.5 ± 1.0**	-	**113.6 ± 0.4**	112.0 ± 0.4	-	**112.2 ± 0.3**	111.7 ± 0.4	-	**81.8 ± 1.9**	81.5 ± 1.9	-	97.6 ± 1.0	**102.7 ± 1.6**
12	-	11.2 ± 1.2	**19.8 ± 3.0**	-	82.5 ± 0.3	**85.6 ± 0.8**	-	**113.7 ± 0.4**	112.0 ± 0.4	-	**112.2 ± 0.3**	111.8 ± 0.4	-	**81.9 ± 1.7**	81.7 ± 1.8	-	97.7 ± 0.9	**102.8 ± 1.5**
13	-	11.3 ± 1.1	**20.1 ± 2.8**	-	82.5 ± 0.3	**85.6 ± 0.8**	-	**113.7 ± 0.4**	112.0 ± 0.3	-	**112.3 ± 0.2**	111.8 ± 0.3	-	**82.0 ± 1.5**	81.8 ± 1.7	-	97.8 ± 0.7	**102.9 ± 1.3**
14	-	11.4 ± 1.0	**20.2 ± 2.6**	-	82.5 ± 0.2	**85.7 ± 0.7**	-	**113.7 ± 0.3**	112.1 ± 0.3	-	**112.3 ± 0.2**	111.8 ± 0.3	-	**82.1 ± 1.4**	81.9 ± 1.6	-	97.8 ± 0.7	**103.0 ± 1.2**
15	-	11.5 ± 1.0	**20.4 ± 2.4**	-	82.5 ± 0.1	**85.7 ± 0.7**	-	**113.7 ± 0.3**	112.1 ± 0.2	-	**112.3 ± 0.2**	111.8 ± 0.3	-	**82.2 ± 1.3**	82.0 ± 1.6	-	97.9 ± 0.6	**103.0 ± 1.1**
16	-	11.5 ± 0.9	**20.6 ± 2.2**	-	82.5 ± 0.1	**85.8 ± 0.6**	-	**113.8 ± 0.3**	112.1 ± 0.2	-	**112.3 ± 0.2**	111.8 ± 0.2	-	**82.3 ± 1.2**	82.1 ± 1.5	-	97.9 ± 0.5	**103.1 ± 0.9**
17	-	11.6 ± 0.9	**20.7 ± 2.1**	-	82.5 ± 0.1	**85.8 ± 0.6**	-	**113.8 ± 0.3**	112.1 ± 0.2	-	**112.3 ± 0.2**	111.9 ± 0.2	-	**82.4 ± 1.2**	82.2 ± 1.4	-	97.9 ± 0.5	**103.1 ± 0.8**
18	-	11.6 ± 0.8	**20.8 ± 1.9**	-	82.5 ± 0.1	**85.8 ± 0.6**	-	**113.8 ± 0.3**	112.1 ± 0.2	-	**112.3 ± 0.2**	111.9 ± 0.2	-	**82.4 ± 1.1**	82.3 ± 1.4	-	97.9 ± 0.5	**103.2 ± 0.8**
19	-	11.7 ± 0.8	**20.9 ± 1.8**	-	82.5 ± 0.1	**85.9 ± 0.6**	-	**113.8 ± 0.2**	112.1 ± 0.1	-	**112.3 ± 0.2**	111.9 ± 0.2	-	**82.5 ± 1.0**	82.3 ± 1.3	-	98.0 ± 0.4	**103.2 ± 0.7**
20	-	11.7 ± 0.7	**21.0 ± 1.7**	-	82.5 ± 0.1	**85.9 ± 0.5**	-	**113.8 ± 0.2**	112.1 ± 0.1	-	**112.3 ± 0.2**	111.9 ± 0.2	-	**82.5 ± 1.0**	82.4 ± 1.3	-	98.0 ± 0.4	**103.2 ± 0.6**



(a) walker2d-random EOP. (b) walker2d-medium EOP. (c) walker2d-expert EOP.

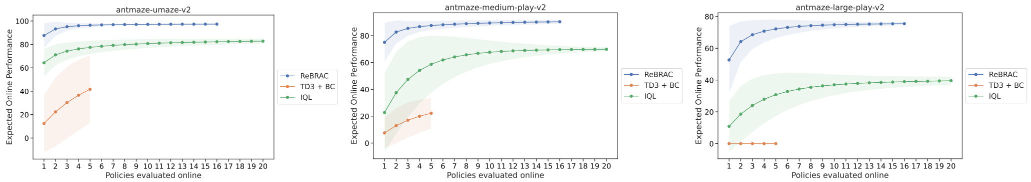


(d) walker2d-medium-expert EOP. (e) walker2d-medium-replay EOP. (f) walker2d-full-replay EOP.

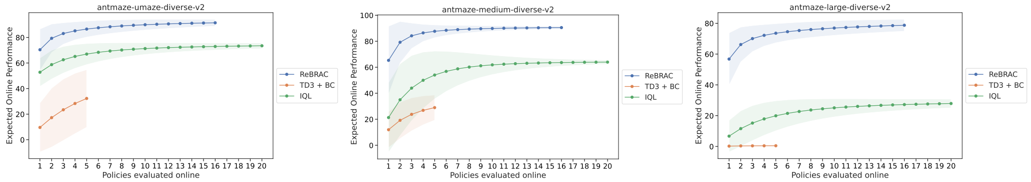
Figure 5: TD3+BC, IQL and ReBRAC visualised Expected Online Performance under uniform policy selection on Walker2d tasks.

Table 25: TD3+BC, IQL and ReBRAC Expected Online Performance under uniform policy selection on AntMaze tasks.

	umaze			medium-play			large-play			umaze-diverse			medium-diverse			large-diverse		
Policies	**TD3+BC**	**IQL**	**ReBRAC**	**TD3+BC**	**IQL**	**ReBRAC**	**TD3+BC**	**IQL**	**ReBRAC**	**TD3+BC**	**IQL**	**ReBRAC**	**TD3+BC**	**IQL**	**ReBRAC**	**TD3+BC**	**IQL**	**ReBRAC**
1	12.4 ± 24.8	64.3 ± 12.0	**87.5 ± 10.9**	7.5 ± 11.7	22.7 ± 29.5	**75.0 ± 14.8**	0.0 ± 0.0	10.9 ± 16.2	**52.7 ± 21.4**	9.6 ± 19.2	52.8 ± 11.1	**70.4 ± 16.2**	11.9 ± 13.8	21.3 ± 26.8	**65.3 ± 26.3**	0.2 ± 0.2	6.7 ± 10.3	**56.8 ± 17.0**
2	22.3 ± 29.8	71.1 ± 9.2	**93.2 ± 6.2**	13.0 ± 13.3	37.5 ± 30.9	**82.7 ± 8.9**	0.0 ± 0.0	18.6 ± 17.8	**64.1 ± 13.0**	17.3 ± 23.0	58.7 ± 11.3	**79.3 ± 11.2**	19.2 ± 14.0	35.0 ± 27.7	**79.2 ± 16.0**	0.3 ± 0.2	11.6 ± 11.6	**66.2 ± 11.3**
3	30.3 ± 31.0	74.3 ± 7.5	**95.1 ± 3.7**	17.0 ± 13.2	47.4 ± 28.5	**85.4 ± 5.7**	0.0 ± 0.0	24.0 ± 16.9	**68.5 ± 9.3**	23.4 ± 24.0	62.5 ± 10.4	**83.1 ± 8.4**	23.8 ± 12.7	44.0 ± 25.1	**84.2 ± 9.8**	0.4 ± 0.2	15.2 ± 11.4	**70.1 ± 8.5**
4	36.6 ± 30.5	76.2 ± 6.5	**96.0 ± 2.4**	20.0 ± 12.6	54.1 ± 25.0	**86.7 ± 4.2**	0.0 ± 0.0	27.9 ± 15.3	**70.8 ± 7.3**	28.3 ± 23.6	65.1 ± 9.3	**85.2 ± 7.0**	26.9 ± 11.1	50.0 ± 21.7	**86.4 ± 6.6**	0.4 ± 0.2	17.9 ± 10.6	**72.2 ± 7.0**
5	41.7 ± 29.1	77.5 ± 5.9	**96.4 ± 1.8**	22.2 ± 11.7	58.7 ± 21.4	**87.5 ± 3.5**	0.0 ± 0.0	30.8 ± 13.6	**72.2 ± 6.0**	32.3 ± 22.5	67.0 ± 8.2	**86.5 ± 6.3**	29.0 ± 9.7	54.0 ± 18.4	**87.6 ± 4.9**	0.5 ± 0.1	19.9 ± 9.7	**73.5 ± 6.2**
6	-	78.4 ± 5.4	**96.7 ± 1.4**	-	61.9 ± 18.2	**88.1 ± 3.2**	-	32.8 ± 11.8	**73.1 ± 4.9**	-	68.4 ± 7.3	**87.6 ± 5.7**	-	56.8 ± 15.4	**88.4 ± 3.9**	-	21.5 ± 8.8	**74.5 ± 5.7**
7	-	79.2 ± 4.9	**96.9 ± 1.1**	-	64.1 ± 15.3	**88.5 ± 2.9**	-	34.3 ± 10.3	**73.7 ± 4.1**	-	69.4 ± 6.4	**88.4 ± 5.3**	-	58.8 ± 12.9	**88.9 ± 3.2**	-	22.7 ± 8.0	**75.3 ± 5.4**
8	-	79.8 ± 4.6	**97.0 ± 0.9**	-	65.7 ± 12.8	**88.9 ± 2.7**	-	35.5 ± 8.9	**74.2 ± 3.5**	-	70.2 ± 5.7	**89.0 ± 4.9**	-	60.1 ± 10.7	**89.3 ± 2.8**	-	23.7 ± 7.2	**75.9 ± 5.1**
9	-	80.3 ± 4.3	**97.1 ± 0.7**	-	66.8 ± 10.7	**89.2 ± 2.5**	-	36.3 ± 7.7	**74.5 ± 2.9**	-	70.8 ± 5.1	**89.5 ± 4.5**	-	61.1 ± 9.0	**89.6 ± 2.4**	-	24.4 ± 6.5	**76.5 ± 4.9**
10	-	80.7 ± 4.0	**97.1 ± 0.6**	-	67.6 ± 9.0	**89.4 ± 2.4**	-	37.0 ± 6.7	**74.8 ± 2.5**	-	71.3 ± 4.6	**89.9 ± 4.2**	-	61.8 ± 7.5	**89.8 ± 2.1**	-	25.1 ± 5.9	**76.9 ± 4.7**
11	-	81.1 ± 3.8	**97.2 ± 0.6**	-	68.2 ± 7.5	**89.6 ± 2.2**	-	37.5 ± 5.9	**75.0 ± 2.1**	-	71.7 ± 4.2	**90.3 ± 3.9**	-	62.4 ± 6.3	**90.0 ± 1.9**	-	25.6 ± 5.4	**77.3 ± 4.6**
12	-	81.4 ± 3.5	**97.2 ± 0.5**	-	68.7 ± 6.3	**89.8 ± 2.1**	-	37.9 ± 5.2	**75.1 ± 1.8**	-	72.0 ± 3.8	**90.6 ± 3.6**	-	62.8 ± 5.3	**90.1 ± 1.6**	-	26.0 ± 4.9	**77.7 ± 4.4**
13	-	81.6 ± 3.3	**97.3 ± 0.4**	-	69.0 ± 5.2	**90.0 ± 2.0**	-	38.2 ± 4.6	**75.2 ± 1.6**	-	72.3 ± 3.5	**90.9 ± 3.3**	-	63.0 ± 4.5	**90.3 ± 1.5**	-	26.4 ± 4.5	**78.0 ± 4.3**
14	-	81.9 ± 3.1	**97.3 ± 0.4**	-	69.2 ± 4.4	**90.1 ± 1.9**	-	38.5 ± 4.2	**75.3 ± 1.4**	-	72.6 ± 3.2	**91.1 ± 3.1**	-	63.3 ± 3.8	**90.4 ± 1.3**	-	26.7 ± 4.2	**78.3 ± 4.1**
15	-	82.1 ± 3.0	**97.3 ± 0.4**	-	69.4 ± 3.7	**90.2 ± 1.8**	-	38.8 ± 3.8	**75.4 ± 1.2**	-	72.8 ± 3.0	**91.3 ± 2.9**	-	63.5 ± 3.2	**90.4 ± 1.2**	-	27.0 ± 3.9	**78.5 ± 4.0**
16	-	82.2 ± 2.8	**97.3 ± 0.3**	-	69.5 ± 3.1	**90.3 ± 1.7**	-	39.0 ± 3.5	**75.4 ± 1.1**	-	72.9 ± 2.8	**91.4 ± 2.7**	-	63.6 ± 2.8	**90.5 ± 1.1**	-	27.2 ± 3.6	**78.7 ± 3.9**
17	-	82.4 ± 2.7	-	-	69.6 ± 2.6	-	-	39.1 ± 3.3	-	-	73.1 ± 2.7	-	-	63.7 ± 2.4	-	-	27.4 ± 3.4	-
18	-	82.5 ± 2.5	-	-	69.7 ± 2.2	-	-	39.3 ± 3.1	-	-	73.2 ± 2.5	-	-	63.8 ± 2.1	-	-	27.6 ± 3.2	-
19	-	82.7 ± 2.4	-	-	69.8 ± 1.9	-	-	39.4 ± 2.9	-	-	73.4 ± 2.4	-	-	63.9 ± 1.8	-	-	27.7 ± 3.0	-
20	-	82.8 ± 2.3	-	-	69.8 ± 1.6	-	-	39.6 ± 2.8	-	-	73.5 ± 2.3	-	-	64.0 ± 1.6	-	-	27.9 ± 2.9	-



(a) antmaze-umaze EOP. (b) antmaze-medium-play EOP. (c) antmaze-large-play EOP.

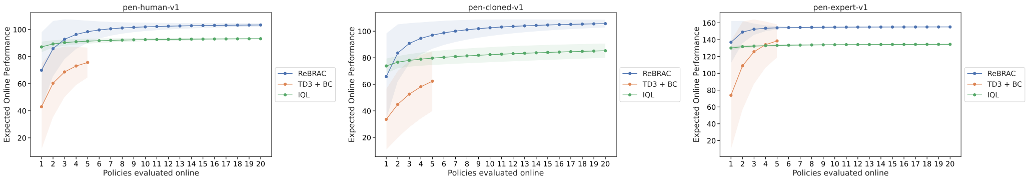


(d) antmaze-umaze-diverse EOP. (e) antmaze-medium-diverse EOP. (f) antmaze-large-diverse EOP.

Figure 6: TD3+BC, IQL and ReBRAC visualised Expected Online Performance under uniform policy selection on AntMaze tasks.

Table 26: TD3+BC, IQL and ReBRAC Expected Online Performance under uniform policy selection on Pen tasks.

	human			cloned			expert		
Policies	**TD3+BC**	**IQL**	**ReBRAC**	**TD3+BC**	**IQL**	**ReBRAC**	**TD3+BC**	**IQL**	**ReBRAC**
1	42.9 ± 32.0	**87.1 ± 4.1**	69.9 ± 28.2	33.6 ± 23.2	**73.8 ± 5.8**	65.8 ± 32.7	73.9 ± 65.2	130.1 ± 2.8	**136.9 ± 25.4**
2	60.4 ± 25.4	**89.4 ± 2.9**	85.8 ± 20.7	44.9 ± 25.4	76.7 ± 4.8	**83.6 ± 21.8**	108.9 ± 52.9	131.7 ± 2.1	**149.0 ± 13.6**
3	68.7 ± 18.9	90.3 ± 2.5	**92.8 ± 14.8**	52.6 ± 25.2	78.0 ± 4.9	**90.8 ± 15.4**	125.6 ± 38.7	132.4 ± 1.8	**152.4 ± 7.0**
4	73.0 ± 14.2	90.9 ± 2.2	**96.3 ± 10.9**	58.1 ± 24.1	79.0 ± 5.2	**94.6 ± 12.2**	134.1 ± 27.8	132.8 ± 1.6	**153.5 ± 3.8**
5	75.6 ± 11.2	91.4 ± 2.0	**98.4 ± 8.3**	62.3 ± 22.6	79.7 ± 5.4	**97.0 ± 10.4**	138.5 ± 20.0	133.1 ± 1.5	**154.1 ± 2.2**
6	-	91.7 ± 1.8	**99.7 ± 6.5**	-	80.3 ± 5.5	**98.7 ± 9.3**	-	133.4 ± 1.4	**154.3 ± 1.5**
7	-	92.0 ± 1.7	**100.5 ± 5.2**	-	80.9 ± 5.6	**100.0 ± 8.4**	-	133.6 ± 1.3	**154.5 ± 1.1**
8	-	92.2 ± 1.6	**101.1 ± 4.4**	-	81.4 ± 5.7	**101.1 ± 7.6**	-	133.7 ± 1.2	**154.7 ± 0.9**
9	-	92.3 ± 1.4	**101.6 ± 3.7**	-	81.9 ± 5.8	**101.9 ± 7.0**	-	133.9 ± 1.1	**154.8 ± 0.8**
10	-	92.5 ± 1.3	**101.9 ± 3.3**	-	82.3 ± 5.9	**102.6 ± 6.4**	-	134.0 ± 1.0	**154.8 ± 0.7**
11	-	92.6 ± 1.2	**102.2 ± 2.9**	-	82.7 ± 5.9	**103.1 ± 5.9**	-	134.1 ± 0.9	**154.9 ± 0.6**
12	-	92.7 ± 1.2	**102.4 ± 2.6**	-	83.0 ± 5.9	**103.6 ± 5.4**	-	134.1 ± 0.9	**154.9 ± 0.6**
13	-	92.8 ± 1.1	**102.6 ± 2.4**	-	83.4 ± 5.9	**104.0 ± 5.0**	-	134.2 ± 0.8	**155.0 ± 0.5**
14	-	92.9 ± 1.0	**102.8 ± 2.2**	-	83.7 ± 5.9	**104.4 ± 4.6**	-	134.2 ± 0.7	**155.0 ± 0.5**
15	-	92.9 ± 1.0	**102.9 ± 2.0**	-	84.0 ± 5.8	**104.7 ± 4.3**	-	134.3 ± 0.7	**155.0 ± 0.4**
16	-	93.0 ± 0.9	**103.0 ± 1.8**	-	84.3 ± 5.8	**104.9 ± 4.0**	-	134.3 ± 0.6	**155.1 ± 0.4**
17	-	93.0 ± 0.9	**103.1 ± 1.7**	-	84.6 ± 5.7	**105.1 ± 3.8**	-	134.4 ± 0.6	**155.1 ± 0.4**
18	-	93.1 ± 0.8	**103.2 ± 1.6**	-	84.8 ± 5.7	**105.3 ± 3.5**	-	134.4 ± 0.6	**155.1 ± 0.4**
19	-	93.1 ± 0.8	**103.3 ± 1.4**	-	85.0 ± 5.6	**105.5 ± 3.3**	-	134.4 ± 0.5	**155.1 ± 0.4**
20	-	93.2 ± 0.8	**103.3 ± 1.3**	-	85.3 ± 5.5	**105.7 ± 3.1**	-	134.5 ± 0.5	**155.1 ± 0.4**



(a) pen-human EOP.

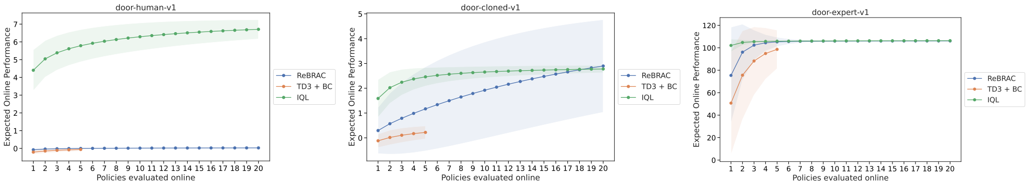
(b) pen-cloned EOP.

(c) pen-expert EOP.

Figure 7: TD3+BC, IQL and ReBRAC visualised Expected Online Performance under uniform policy selection on Pen tasks.

Table 27: TD3+BC, IQL and ReBRAC Expected Online Performance under uniform policy selection on Door tasks.

	human			cloned			expert		
Policies	**TD3+BC**	**IQL**	**ReBRAC**	**TD3+BC**	**IQL**	**ReBRAC**	**TD3+BC**	**IQL**	**ReBRAC**
1	-0.2 ± 0.1	**4.4** ± **1.2**	-0.1 ± 0.1	-0.1 ± 0.3	**1.6** ± **0.8**	0.3 ± 0.9	50.7 ± 46.3	**102.1** ± **5.7**	75.4 ± 43.0
2	-0.1 ± 0.1	**5.0** ± **1.0**	-0.0 ± 0.1	0.0 ± 0.3	**2.0** ± **0.6**	0.6 ± 1.2	75.6 ± 39.5	**104.7** ± **2.5**	96.1 ± 25.0
3	-0.1 ± 0.1	**5.4** ± **1.0**	-0.0 ± 0.0	0.1 ± 0.3	**2.2** ± **0.5**	0.8 ± 1.4	88.2 ± 30.4	**105.4** ± **1.3**	102.3 ± 13.5
4	-0.1 ± 0.1	**5.6** ± **0.9**	-0.0 ± 0.0	0.2 ± 0.3	**2.4** ± **0.4**	1.0 ± 1.6	94.9 ± 22.9	**105.7** ± **0.8**	104.4 ± 7.3
5	-0.1 ± 0.1	**5.8** ± **0.9**	-0.0 ± 0.0	0.2 ± 0.3	**2.5** ± **0.4**	1.2 ± 1.7	98.6 ± 17.2	**105.8** ± **0.6**	105.2 ± 4.1
6	-	**5.9** ± **0.9**	0.0 ± 0.0	-	**2.5** ± **0.3**	1.3 ± 1.8	-	**105.9** ± **0.5**	105.6 ± 2.4
7	-	**6.0** ± **0.9**	0.0 ± 0.0	-	**2.6** ± **0.3**	1.5 ± 1.8	-	**106.0** ± **0.5**	105.8 ± 1.5
8	-	**6.1** ± **0.8**	0.0 ± 0.0	-	**2.6** ± **0.3**	1.6 ± 1.9	-	**106.1** ± **0.4**	105.9 ± 1.0
9	-	**6.2** ± **0.8**	0.0 ± 0.0	-	**2.6** ± **0.2**	1.8 ± 1.9	-	**106.1** ± **0.4**	105.9 ± 0.7
10	-	**6.3** ± **0.8**	0.0 ± 0.0	-	**2.7** ± **0.2**	1.9 ± 1.9	-	**106.1** ± **0.4**	106.0 ± 0.5
11	-	**6.4** ± **0.8**	0.0 ± 0.0	-	**2.7** ± **0.2**	2.0 ± 2.0	-	**106.2** ± **0.4**	106.0 ± 0.4
12	-	**6.4** ± **0.7**	0.0 ± 0.0	-	**2.7** ± **0.2**	2.2 ± 2.0	-	**106.2** ± **0.3**	106.0 ± 0.3
13	-	**6.5** ± **0.7**	0.0 ± 0.0	-	**2.7** ± **0.2**	2.3 ± 2.0	-	**106.2** ± **0.3**	106.0 ± 0.2
14	-	**6.5** ± **0.7**	0.0 ± 0.0	-	**2.7** ± **0.2**	2.4 ± 2.0	-	**106.2** ± **0.3**	106.0 ± 0.2
15	-	**6.6** ± **0.7**	0.0 ± 0.0	-	**2.7** ± **0.2**	2.5 ± 2.0	-	**106.3** ± **0.3**	106.0 ± 0.2
16	-	**6.6** ± **0.6**	0.0 ± 0.0	-	**2.7** ± **0.2**	2.6 ± 1.9	-	**106.3** ± **0.3**	106.1 ± 0.2
17	-	**6.6** ± **0.6**	0.0 ± 0.0	-	**2.8** ± **0.2**	2.7 ± 1.9	-	**106.3** ± **0.3**	106.1 ± 0.2
18	-	**6.7** ± **0.6**	0.0 ± 0.0	-	**2.8** ± **0.2**	2.7 ± 1.9	-	**106.3** ± **0.3**	106.1 ± 0.1
19	-	**6.7** ± **0.6**	0.0 ± 0.0	-	2.8 ± 0.2	**2.8** ± **1.9**	-	**106.3** ± **0.3**	106.1 ± 0.1
20	-	**6.7** ± **0.5**	0.0 ± 0.0	-	2.8 ± 0.2	**2.9** ± **1.9**	-	**106.3** ± **0.2**	106.1 ± 0.1



(a) door-human EOP.

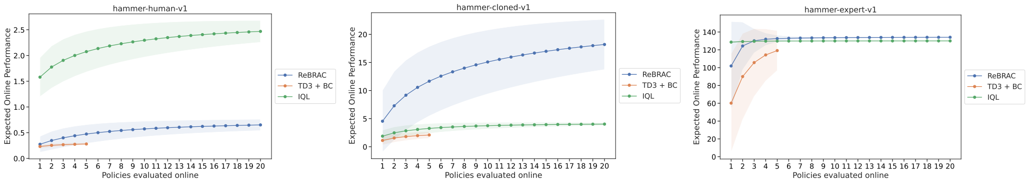
(b) door-cloned EOP.

(c) door-expert EOP.

Figure 8: TD3+BC, IQL and ReBRAC visualised Expected Online Performance under uniform policy selection on Door tasks.

Table 28: TD3+BC, IQL and ReBRAC Expected Online Performance under uniform policy selection on Hammer tasks.

	human			cloned			expert		
Policies	**TD3+BC**	**IQL**	**ReBRAC**	**TD3+BC**	**IQL**	**ReBRAC**	**TD3+BC**	**IQL**	**ReBRAC**
1	0.2 ± 0.0	**1.6 ± 0.4**	0.3 ± 0.2	1.1 ± 0.8	1.9 ± 1.1	**4.5 ± 5.5**	60.2 ± 55.4	**128.7 ± 1.1**	101.9 ± 49.9
2	0.3 ± 0.0	**1.8 ± 0.4**	0.3 ± 0.2	1.6 ± 0.8	2.5 ± 1.1	**7.3 ± 6.1**	90.0 ± 48.4	**129.3 ± 0.9**	124.3 ± 26.8
3	0.3 ± 0.0	**1.9 ± 0.4**	0.4 ± 0.2	1.8 ± 0.7	2.8 ± 1.0	**9.2 ± 6.2**	105.7 ± 38.1	129.6 ± 0.7	**130.1 ± 13.6**
4	0.3 ± 0.0	**2.0 ± 0.4**	0.4 ± 0.2	2.0 ± 0.6	3.1 ± 0.9	**10.6 ± 6.2**	114.3 ± 29.3	129.7 ± 0.5	**131.9 ± 7.0**
5	0.3 ± 0.0	**2.1 ± 0.4**	0.5 ± 0.2	2.1 ± 0.6	3.3 ± 0.9	**11.7 ± 6.2**	119.2 ± 22.4	129.8 ± 0.4	**132.6 ± 3.8**
6	-	**2.1 ± 0.4**	0.5 ± 0.2	-	3.4 ± 0.8	**12.6 ± 6.1**	-	129.9 ± 0.3	**133.0 ± 2.4**
7	-	**2.2 ± 0.4**	0.5 ± 0.2	-	3.5 ± 0.7	**13.3 ± 6.0**	-	129.9 ± 0.2	**133.2 ± 1.7**
8	-	**2.2 ± 0.4**	0.5 ± 0.2	-	3.6 ± 0.7	**14.0 ± 5.9**	-	129.9 ± 0.2	**133.4 ± 1.4**
9	-	**2.3 ± 0.3**	0.6 ± 0.2	-	3.7 ± 0.6	**14.6 ± 5.8**	-	130.0 ± 0.1	**133.5 ± 1.2**
10	-	**2.3 ± 0.3**	0.6 ± 0.2	-	3.7 ± 0.6	**15.1 ± 5.7**	-	130.0 ± 0.1	**133.6 ± 1.1**
11	-	**2.3 ± 0.3**	0.6 ± 0.1	-	3.8 ± 0.5	**15.5 ± 5.6**	-	130.0 ± 0.1	**133.7 ± 1.0**
12	-	**2.3 ± 0.3**	0.6 ± 0.1	-	3.8 ± 0.5	**16.0 ± 5.4**	-	130.0 ± 0.1	**133.8 ± 0.9**
13	-	**2.4 ± 0.3**	0.6 ± 0.1	-	3.9 ± 0.5	**16.3 ± 5.3**	-	130.0 ± 0.1	**133.9 ± 0.8**
14	-	**2.4 ± 0.3**	0.6 ± 0.1	-	3.9 ± 0.4	**16.7 ± 5.2**	-	130.0 ± 0.1	**133.9 ± 0.7**
15	-	**2.4 ± 0.3**	0.6 ± 0.1	-	3.9 ± 0.4	**17.0 ± 5.1**	-	130.0 ± 0.1	**134.0 ± 0.6**
16	-	**2.4 ± 0.3**	0.6 ± 0.1	-	4.0 ± 0.4	**17.3 ± 4.9**	-	130.0 ± 0.1	**134.0 ± 0.6**
17	-	**2.4 ± 0.2**	0.6 ± 0.1	-	4.0 ± 0.4	**17.5 ± 4.8**	-	130.0 ± 0.0	**134.0 ± 0.5**
18	-	**2.4 ± 0.2**	0.6 ± 0.1	-	4.0 ± 0.3	**17.8 ± 4.7**	-	130.0 ± 0.0	**134.0 ± 0.5**
19	-	**2.5 ± 0.2**	0.6 ± 0.1	-	4.0 ± 0.3	**18.0 ± 4.6**	-	130.0 ± 0.0	**134.1 ± 0.4**
20	-	**2.5 ± 0.2**	0.7 ± 0.1	-	4.0 ± 0.3	**18.2 ± 4.5**	-	130.0 ± 0.0	**134.1 ± 0.4**



(a) hammer-human EOP.

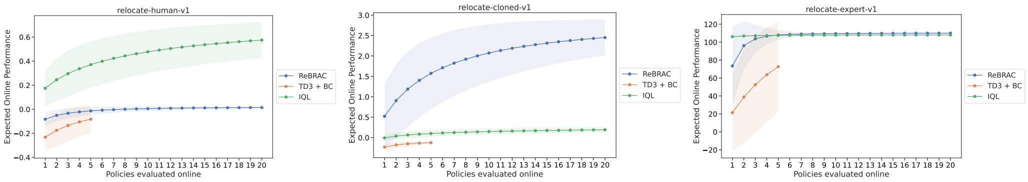
(b) hammer-cloned EOP.

(c) hammer-expert EOP.

Figure 9: TD3+BC, IQL and ReBRAC visualised Expected Online Performance under uniform policy selection on Hammer tasks.

Table 29: TD3+BC, IQL and ReBRAC Expected Online Performance under uniform policy selection on tasks.

	human			cloned			expert		
Policies	**TD3+BC**	**IQL**	**ReBRAC**	**TD3+BC**	**IQL**	**ReBRAC**	**TD3+BC**	**IQL**	**ReBRAC**
1	-0.2 ± 0.1	**0.2 ± 0.2**	-0.1 ± 0.1	-0.2 ± 0.1	-0.0 ± 0.1	**0.5 ± 0.8**	21.4 ± 43.2	**106.0 ± 1.4**	73.5 ± 44.3
2	-0.2 ± 0.1	**0.2 ± 0.2**	-0.1 ± 0.1	-0.2 ± 0.1	0.0 ± 0.1	**0.9 ± 0.9**	38.8 ± 51.9	**106.8 ± 1.0**	96.0 ± 27.4
3	-0.1 ± 0.1	**0.3 ± 0.2**	-0.0 ± 0.0	-0.2 ± 0.1	0.1 ± 0.1	**1.2 ± 0.9**	52.6 ± 54.0	**107.2 ± 0.8**	103.7 ± 15.9
4	-0.1 ± 0.1	**0.3 ± 0.2**	-0.0 ± 0.0	-0.1 ± 0.1	0.1 ± 0.1	**1.4 ± 0.9**	63.7 ± 53.1	**107.4 ± 0.7**	106.7 ± 9.4
5	-0.1 ± 0.1	**0.4 ± 0.2**	-0.0 ± 0.0	-0.1 ± 0.0	0.1 ± 0.1	**1.6 ± 0.9**	72.5 ± 50.7	107.5 ± 0.6	**107.9 ± 5.8**
6	-	**0.4 ± 0.2**	-0.0 ± 0.0	-	0.1 ± 0.1	**1.7 ± 0.8**	-	107.6 ± 0.5	**108.6 ± 3.8**
7	-	**0.4 ± 0.2**	-0.0 ± 0.0	-	0.1 ± 0.1	**1.8 ± 0.8**	-	107.7 ± 0.5	**108.9 ± 2.6**
8	-	**0.4 ± 0.2**	0.0 ± 0.0	-	0.1 ± 0.1	**1.9 ± 0.8**	-	107.7 ± 0.5	**109.2 ± 1.9**
9	-	**0.5 ± 0.2**	0.0 ± 0.0	-	0.1 ± 0.1	**2.0 ± 0.7**	-	107.8 ± 0.4	**109.3 ± 1.5**
10	-	**0.5 ± 0.2**	0.0 ± 0.0	-	0.1 ± 0.1	**2.1 ± 0.7**	-	107.8 ± 0.4	**109.4 ± 1.3**
11	-	**0.5 ± 0.2**	0.0 ± 0.0	-	0.2 ± 0.1	**2.1 ± 0.7**	-	107.8 ± 0.4	**109.5 ± 1.1**
12	-	**0.5 ± 0.2**	0.0 ± 0.0	-	0.2 ± 0.1	**2.2 ± 0.6**	-	107.9 ± 0.4	**109.6 ± 1.0**
13	-	**0.5 ± 0.2**	0.0 ± 0.0	-	0.2 ± 0.1	**2.2 ± 0.6**	-	107.9 ± 0.4	**109.7 ± 0.9**
14	-	**0.5 ± 0.2**	0.0 ± 0.0	-	0.2 ± 0.1	**2.3 ± 0.6**	-	107.9 ± 0.4	**109.8 ± 0.9**
15	-	**0.5 ± 0.2**	0.0 ± 0.0	-	0.2 ± 0.1	**2.3 ± 0.6**	-	107.9 ± 0.4	**109.8 ± 0.8**
16	-	**0.5 ± 0.2**	0.0 ± 0.0	-	0.2 ± 0.1	**2.3 ± 0.5**	-	108.0 ± 0.4	**109.9 ± 0.8**
17	-	**0.6 ± 0.2**	0.0 ± 0.0	-	0.2 ± 0.1	**2.4 ± 0.5**	-	108.0 ± 0.4	**109.9 ± 0.8**
18	-	**0.6 ± 0.2**	0.0 ± 0.0	-	0.2 ± 0.1	**2.4 ± 0.5**	-	108.0 ± 0.3	**109.9 ± 0.8**
19	-	**0.6 ± 0.2**	0.0 ± 0.0	-	0.2 ± 0.1	**2.4 ± 0.5**	-	108.0 ± 0.3	**110.0 ± 0.7**
20	-	**0.6 ± 0.1**	0.0 ± 0.0	-	0.2 ± 0.1	**2.5 ± 0.4**	-	108.0 ± 0.3	**110.0 ± 0.7**



(a) relocate-human EOP.

(b) relocate-cloned EOP.

(c) relocate-expert EOP.

Figure 10: TD3+BC, IQL and ReBRAC visualised Expected Online Performance under uniform policy selection on Relocate tasks.

F D4RL tasks ablation

Table 30: ReBRAC ablations for halfcheetah tasks. We report final normalized score averaged over 4 unseen training seeds.

Ablation	random	medium	expert	medium-expert	medium-replay	full-replay	Average
TD3+BC, paper	11.0 ± 1.1	48.3 ± 0.3	96.7 ± 1.1	90.7 ± 4.3	44.6 ± 0.5	-	-
TD3+BC, our	2.2 ± 0.0	44.6 ± 0.4	93.8 ± 0.1	91.9 ± 2.3	40.5 ± 1.6	69.3 ± 0.7	57.0
TD3+BC, tuned	30.1 ± 1.4 (+0.7%)	55.4 ± 0.9 (-15.2%)	95.5 ± 0.5 (-9.6%)	91.9 ± 2.3 (-11.4%)	45.1 ± 1.7 (-11.0%)	74.1 ± 2.9 (-9.5%)	65.3 (-10.3%)
ReBRAC w/o LN	32.0 ± 1.5 (+7.0%)	64.3 ± 4.1 (-1.5%)	61.7 ± 20.4 (-41.5%)	86.7 ± 0.9 (-16.3%)	52.8 ± 2.4 (+4.1%)	82.1 ± 2.1 (+0.2%)	63.2 (-13.1%)
ReBRAC w/o layer	27.8 ± 3.4 (-7.0%)	65.0 ± 1.6 (-0.4%)	74.4 ± 26.7 (-29.5%)	86.7 ± 8.8 (-16.3%)	50.4 ± 0.7 (-0.5%)	80.9 ± 1.1 (-1.2%)	64.1 (-11.9%)
ReBRAC w/o actor penalty	31.8 ± 4.1 (+6.3%)	64.5 ± 0.7 (-1.2%)	4.3 ± 4.3 (-95.9%)	71.6 ± 12.3 (-30.9%)	38.0 ± 27.2 (-25.0%)	59.3 ± 41.3 (-27.5%)	44.9 (-38.3%)
ReBRAC w/o critic penalty	28.1 ± 1.6 (-6.0%)	65.7 ± 1.4 (+0.6%)	104.2 ± 5.9 (-1.3%)	100.5 ± 3.1 (-3.0%)	50.7 ± 0.1 (0.0%)	81.7 ± 1.2 (-0.2%)	71.8 (-1.3%)
ReBRAC w/o large batch	21.0 ± 15.7 (-29.7%)	65.8 ± 0.7 (+0.7%)	62.6 ± 24.3 (-40.7%)	85.2 ± 7.3 (-17.8%)	50.7 ± 1.1 (0.0%)	81.9 ± 1.4 (0.0%)	61.2 (-15.9%)
ReBRAC	29.9 ± 1.2	65.3 ± 1.1	105.6 ± 1.5	103.7 ± 3.9	50.7 ± 0.6	81.9 ± 1.4	72.8

Table 31: ReBRAC ablations for hopper tasks. We report final normalized score averaged over 4 unseen training seeds.

Ablation	random	medium	expert	medium-expert	medium-replay	full-replay	Average
TD3+BC, paper	8.5 ± 0.6	59.3 ± 4.2	107.8 ± 7.0	98.0 ± 9.4	60.9 ± 18.8	-	-
TD3+BC, our	10.3 ± 1.8	53.2 ± 2.2	108.7 ± 5.3	75.8 ± 8.9	64.5 ± 24.9	49.9 ± 9.6	60.4
TD3+BC, tuned	10.3 ± 1.8 (+51.5%)	57.6 ± 6.8 (-43.2%)	110.7 ± 2.1 (+21.1%)	106.2 ± 2.5 (-3.4%)	64.5 ± 24.9 (-30.8%)	106.2 ± 2.2 (-0.6%)	75.9 (-10.6%)
ReBRAC w/o LN	12.2 ± 13.3 (+79.4%)	1.0 ± 0.6 (-99.0%)	111.1 ± 1.0 (+21.6%)	112.4 ± 0.7 (+2.3%)	57.4 ± 25.0 (-38.5%)	107.2 ± 2.0 (+0.4%)	66.8 (-21.3%)
ReBRAC w/o layer	8.8 ± 0.6 (+29.4%)	101.8 ± 0.2 (+0.4%)	103.7 ± 5.1 (+13.5%)	104.1 ± 7.7 (-5.3%)	97.5 ± 3.5 (+4.5%)	106.5 ± 0.3 (-0.3%)	87.0 (+2.5%)
ReBRAC w/o actor penalty	7.5 ± 4.6 (+10.3%)	1.7 ± 1.2 (-98.3%)	1.1 ± 0.5 (-98.8%)	1.6 ± 1.6 (-98.5%)	24.4 ± 8.7 (-73.8%)	27.7 ± 23.4 (-74.1%)	10.6 (-87.5%)
ReBRAC w/o critic penalty	7.4 ± 1.1 (+8.8%)	102.3 ± 0.5 (+0.9%)	103.4 ± 8.6 (+13.1%)	111.2 ± 0.7 (+1.2%)	83.1 ± 30.9 (-10.9%)	107.5 ± 0.1 (+0.7%)	85.8 (+1.1%)
ReBRAC w/o large batch	8.6 ± 0.5 (+26.5%)	98.9 ± 5.2 (-2.5%)	98.8 ± 13.4 (+8.1%)	107.8 ± 2.9 (-1.9%)	96.2 ± 7.6 (+3.1%)	106.6 ± 0.2 (-0.2%)	86.1 (+1.4%)
ReBRAC	6.8 ± 3.4	101.4 ± 1.5	91.4 ± 4.7	109.9 ± 3.0	93.3 ± 7.5	106.8±0.6	84.9

Table 32: ReBRAC ablations for walker2d tasks. We report final normalized score averaged over 4 unseen training seeds.

Ablation	random	medium	expert	medium-expert	medium-replay	full-replay	Average
TD3+BC, paper	1.6 ± 1.7	83.7 ± 2.1	110.2 ± 0.3	110.1 ± 0.5	81.8 ± 5.5	-	-
TD3+BC, our	4.5 ± 2.2	77.1 ± 1.9	109.1 ± 0.5	108.9 ± 0.3	50.9 ± 13.7	86.7 ± 5.1	72.8
TD3+BC, tuned	4.5 ± 2.2 (-78.8%)	77.1 ± 1.9 (-5.5%)	110.1 ± 0.1 (-2.0%)	110.2 ± 0.7 (-1.3%)	58.8 ± 28.5 (-22.2%)	89.4 ± 8.2 (-13.0%)	75.0 (-10.9%)
ReBRAC w/o LN	1.3 ± 1.5 (-93.9%)	84.3 ± 2.5 (+3.3%)	8.3 ± 3.1 (-92.6%)	52.7 ± 53.9 (-52.8%)	78.9 ± 8.4 (+4.4%)	61.1 ± 46.6 (-40.6%)	47.7 (-43.3%)
ReBRAC w/o layer	11.4 ± 11.9 (-46.5%)	86.2 ± 0.9 (+5.6%)	112.1 ± 0.1 (-0.2%)	111.9 ± 0.2 (+0.2%)	83.9 ± 5.4 (+11.0%)	101.8 ± 0.9 (-1.0%)	84.5 (+0.4%)
ReBRAC w/o actor penalty	1.1 ± 0.9 (-94.8%)	1.7 ± 2.1 (-97.9%)	0.9 ± 1.1 (-99.2%)	0.8 ± 1.3 (-99.3%)	8.9 ± 5.7 (-88.2%)	64.2 ± 29.5 (-37.5%)	12.9 (-84.7%)
ReBRAC w/o critic penalty	19.8 ± 3.6 (-7.0%)	81.6 ± 9.2 (0.0%)	112.0 ± 0.1 (-0.3%)	111.6 ± 0.3 (-0.1%)	87.0 ± 4.5 (+15.1%)	103.5 ± 1.5 (+0.7%)	85.9 (+2.0%)
ReBRAC w/o large batch	5.6 ± 0.2 (-73.7%)	84.8 ± 1.0 (+3.9%)	112.2 ± 0.2 (-0.1%)	110.9 ± 0.2 (-0.7%)	71.7 ± 20.2 (-5.2%)	97.7 ± 5.8 (-5.0%)	80.4 (-4.5%)
ReBRAC	21.3 ± 0.8	81.6 ± 3.9	112.3 ± 0.0	111.7 ± 0.3	75.6±10.3	102.8±0.9	84.2

Table 33: ReBRAC ablations for AntMaze tasks. We report final normalized score averaged over 4 unseen training seeds.

Ablation	umaze	umaze-diverse	medium-play	medium-diverse	large-play	large-diverse	Average
TD3+BC, paper	78.6	71.4	10.6	3.0	0.2	0.0	27.3
TD3+BC, our	62.0 ± 2.4	48.0 ± 11.6	0.0 ± 0.0	0.5 ± 1	0.0 ± 0.0	0.5 ± 0.5	18.5
TD3+BC, tuned	62.0 ± 2.4 (-36.8%)	48.0 ± 11.6 (-42.9%)	39.0 ± 21.7 (-54.8%)	18.5 ± 17.7 (-72.1%)	0.2 ± 0.5 (-99.6%)	0.0 ± 1.0 (-100.0%)	27.9 (-62.9%)
ReBRAC w/o γ change	0.0 ± 0.0 (-100.0%)	90.7 ± 3.2 (+7.7%)	1.0 ± 0.0 (-98.8%)	0.2 ± 0.5 (-99.7%)	19.3 ± 18.5 (-58.0%)	15.0 ± 8.0 (-79.0%)	21.0 (-72.1%)
ReBRAC w/o LN	0.0 ± 0.0 (-100%)	0.0 ± 0.0 (-100%)	0.0 ± 0.0 (-100%)	0.0 ± 0.0 (-100%)	0.0 ± 0.0 (-100%)	0.0 ± 0.0 (-100%)	0.0 (-100%)
ReBRAC w/o layer	31.0 ± 45.4 (-68.4%)	61.7 ± 25.3 (-26.7%)	0.0 ± 0.0 (-100.0%)	16.0 ± 32.0 (-75.8%)	0.0 ± 0.0 (-100.0%)	0.0 ± 0.0 (-100.0%)	18.1 (-76.0%)
ReBRAC w/o actor penalty	1.0 ± 1.1 (-99.0%)	0.0 ± 0.0 (-100%)	0.0 ± 0.0 (-100%)	0.0 ± 0.0 (-100%)	0.0 ± 0.0 (-100%)	0.0 ± 0.0 (-100%)	0.1 (-99.9%)
ReBRAC w/o critic penalty	98.2 ± 1.5 (0.0%)	78.0 ± 26.3 (-7.4%)	86.2 ± 2.6 (0.0%)	57.5 ± 24.2 (-13.1%)	56.7 ± 32.9 (+23.3%)	57.0 ± 16.4 (-20.3%)	72.2 (-4.1%)
ReBRAC w large batch	60.7 ± 31.3 (-38.2%)	68.5 ± 17.9 (-18.6%)	43.9 ± 49.9 (-49.1%)	34.0 ± 40.6 (-48.6%)	39.2 ± 45.9 (-14.8%)	0.0 ± 0.0 (-100.0%)	41.0 (-45.6%)
ReBRAC	98.2 ± 0.9	84.2 ± 18.5	86.2 ± 4.7	66.2 ± 16.3	46.0 ± 40.0	71.5 ± 12.3	75.3

Table 34: ReBRAC ablations for pen tasks. We report final normalized score averaged over 4 unseen training seeds.

Ablation	human	cloned	expert	Average
TD3+BC, paper	0.0	0.0	0.3	0.0
TD3+BC, our	65.9 ± 24.6	78.1 ± 5.7	144.9 ± 7.5	96.3
TD3+BC, tuned	77.6 ± 18.5 (-23.9%)	78.1 ± 5.7 (-8.5%)	144.9 ± 7.5 (-10.3%)	100.2 (-12.5%)
ReBRAC w/o LN	78.6 ± 14.8 (-22.9%)	21.3 ± 11.0 (-75.1%)	86.7 ± 59.8 (-44.6%)	62.1 (-45.8%)
ReBRAC w/o layer	89.1 ± 14.7 (-12.6%)	106.7 ± 13.9 (+24.9%)	147.2 ± 5.7 (-6.0%)	114.3 (-0.3%)
ReBRAC w/o actor penalty	-0.5 ± 1.3 (-100.5%)	0.6 ± 1.6 (-99.3%)	0.0 ± 3.6 (-100.0%)	0.0 (-100.0%)
ReBRAC w/o critic penalty	99.9 ± 6.1 (-2.1%)	75.0 ± 16.7 (-12.2%)	154.6 ± 1.8 (-1.3%)	109.8 (-4.2%)
ReBRAC w large batch	67.2 ± 9.0 (-34.1%)	83.2 ± 23.3 (-2.6%)	155.0 ± 6.8 (-1.0%)	101.8 (-11.2%)
ReBRAC	102.0 ± 10.8	85.4 ± 24.2	156.6 ± 1.4	114.6

Table 35: ReBRAC ablations for door tasks. We report final normalized score averaged over 4 unseen training seeds.

Ablation	human	cloned	expert	Average
TD3+BC, paper	0.0	0.0	0.0	0.0
TD3+BC, our	0.0 ± 0.1	0.4 ± 1.0	102.5 ± 2.9	34.3
TD3+BC, tuned	0.0 ± 0.1 (-)	0.4 ± 1.0 (+100.0%)	105.8 ± 0.3 (+0.8%)	35.4 (+1.1%)
ReBRAC w/o LN	-0.1 ± 0.0 (-)	-0.3 ± 0.0 (-250.0%)	106.0 ± 0.8 (+1.0%)	35.1 (+0.3%)
ReBRAC w/o layer	0.0 ± 0.0 (-)	0.1 ± 0.5 (-50.0%)	104.4 ± 2.3 (-0.5%)	34.8 (-0.6%)
ReBRAC w/o actor penalty	-0.1 ± 0.1 (-)	0.0 ± 0.0 (-100.0%)	0.0 ± 0.2 (-100.0%)	0.0 (-100.0%)
ReBRAC w/o critic penalty	0.0 ± 0.0 (-)	0.1 ± 0.0 (-50.0%)	106.1 ± 0.3 (+1.1%)	35.4 (+1.1%)
ReBRAC w large batch	-0.1 ± 0.1 (-)	0.1 ± 0.3 (-50.0%)	106.1 ± 0.1 (+1.1%)	35.3 (+0.9%)
ReBRAC	0.0 ± 0.0	0.2 ± 0.3	104.9 ± 2.2	35.0

Table 36: ReBRAC ablations for hammer tasks. We report final normalized score averaged over 4 unseen training seeds.

Ablation	human	cloned	expert	Average
TD3+BC, paper	0.0	0.0	0.0	0.0
TD3+BC, our	0.3 ± 0.4	1.1 ± 1.1	127.0 ± 0.4	42.8
TD3+BC, tuned	0.3 ± 0.4 (+50.0%)	1.1 ± 1.1 (-80.0%)	127.0 ± 0.4 (-5.3%)	42.8 (-8.1%)
ReBRAC w/o LN	0.2 ± 0.0 (0.0%)	1.0 ± 1.0 (-81.8%)	9.9 ± 19.1 (-92.6%)	3.6 (-92.3%)
ReBRAC w/o layer	0.1 ± 0.0 (-50.0%)	21.3 ± 19.7 (+287.3%)	133.1 ± 0.5 (-0.8%)	51.5 (+10.5%)
ReBRAC w/o actor penalty	0.0 ± 0.0 (-100.0%)	0.0 ± 0.1 (-100.0%)	0.0 ± 0.1 (-100.0%)	0.0 (-100.0%)
ReBRAC w/o critic penalty	0.1 ± 0.1 (-50.0%)	1.9 ± 0.7 (-65.5%)	134.1 ± 0.2 (-0.1%)	45.3 (-2.8%)
ReBRAC w large batch	0.3 ± 0.8 (+50.0%)	10.6 ± 14.0 (+92.7%)	133.4 ± 0.5 (-0.6%)	48.1 (+3.2%)
ReBRAC	0.2 ± 0.2	5.5 ± 2.5	134.2 ± 0.4	46.6

Table 37: ReBRAC ablations for relocate tasks. We report final normalized score averaged over 4 unseen training seeds.

Ablation	human	cloned	expert	Average
TD3+BC, paper	0.0	0.0	0.0	0.0
TD3+BC, our	0.0 ± 0.0	-0.1 ± 0.0	107.9 ± 0.6	35.9
TD3+BC, tuned	0.0 ± 0.0 (-)	-0.1 ± 0.0 (-105.3%)	107.9 ± 0.6 (+1.2%)	35.9 (-0.5%)
ReBRAC w/o LN	-0.2 ± 0.0 (-)	0.0 ± 0.3 (-100.0%)	-0.1 ± 0.0 (-100.1%)	-0.1 (-100.3%)
ReBRAC w/o layer	0.1 ± 0.3 (-)	1.7 ± 2.1 (-10.5%)	105.0 ± 3.1 (-1.5%)	35.6 (-1.4%)
ReBRAC w/o actor penalty	-0.1 ± 0.0 (-)	0.0 ± 0.0 (-100.0%)	-0.1 ± 0.1 (-100.1%)	0.0 (-100.0%)
ReBRAC w/o critic penalty	0.0 ± 0.1 (-)	1.9 ± 1.9 (0.0%)	109.6 ± 1.2 (+2.8%)	37.1 (+2.8%)
ReBRAC w large batch	0.0 ± 0.0 (-)	0.1 ± 0.2 (-94.7%)	109.6 ± 0.9 (+2.8%)	36.5 (+1.1%)
ReBRAC	0.0 ± 0.0	1.9 ± 2.3	106.6 ± 3.1	36.1