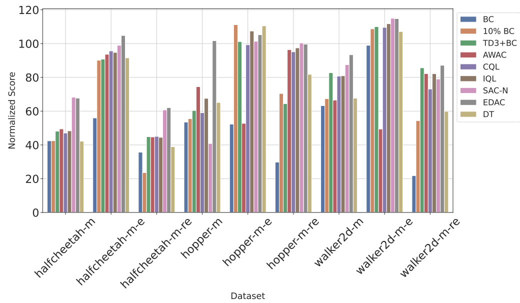
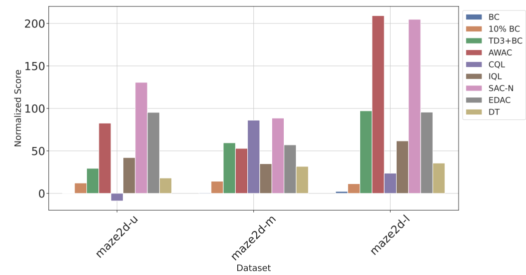
A Additional Benchmark Information

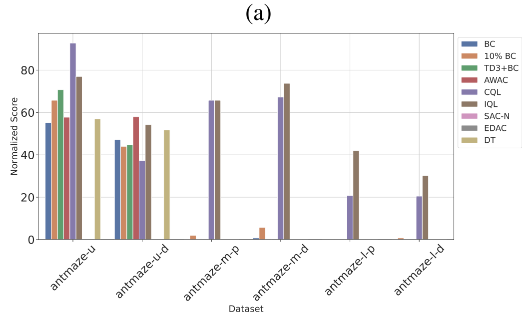
A.1 Offline



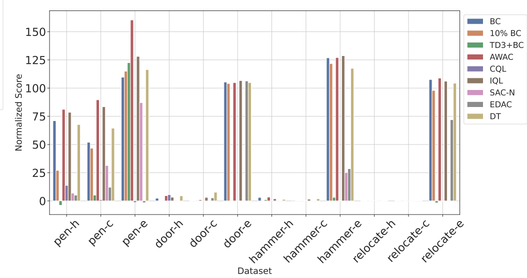
(a)



(b)

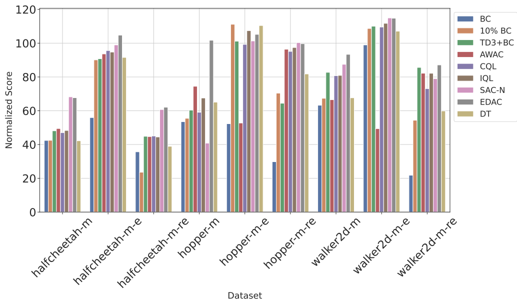


(c)

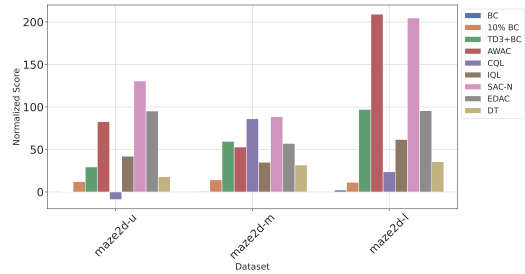


(d)

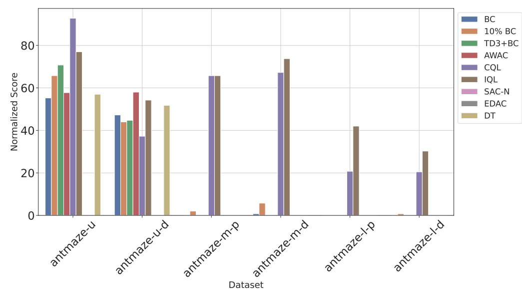
Figure 4: Graphical representation of the normalized performance of the last trained policy on D4RL averaged over 4 random seeds. (a) Gym-MuJoCo datasets. (b) Maze2d datasets (c) AntMaze datasets (d) Adroit datasets



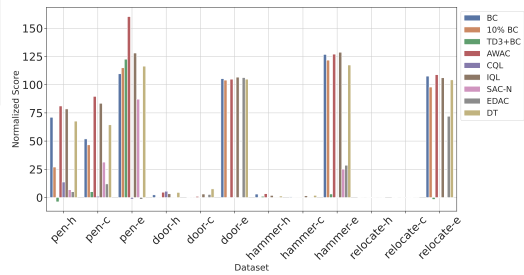
(a)



(b)

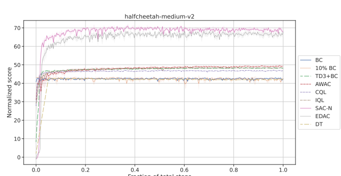


(c)

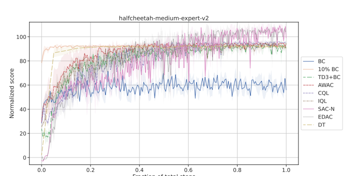


(d)

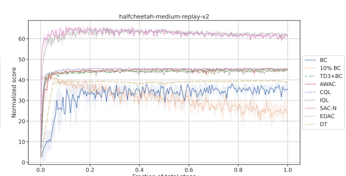
Figure 5: Graphical representation of the normalized performance of the best trained policy on D4RL averaged over 4 random seeds. (a) Gym-MuJoCo datasets. (b) Maze2d datasets (c) AntMaze datasets (d) Adroit datasets



(a)



(b)

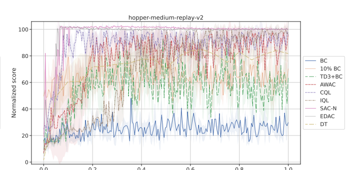


(c)

Figure 6: Training curves for HalfCheetah task.
(a) Medium dataset, (b) Medium-expert dataset, (c) Medium-replay dataset

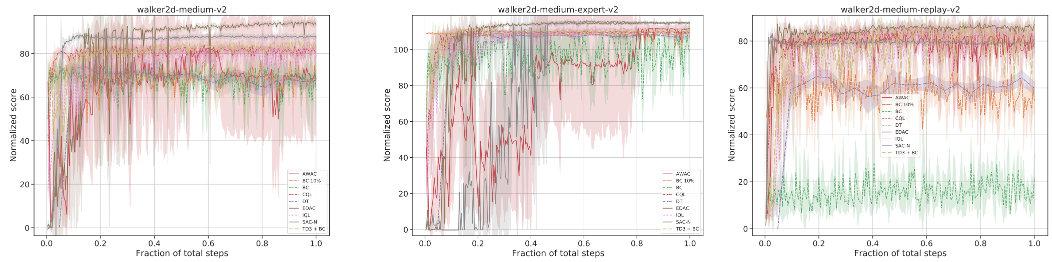
(a)

(b)



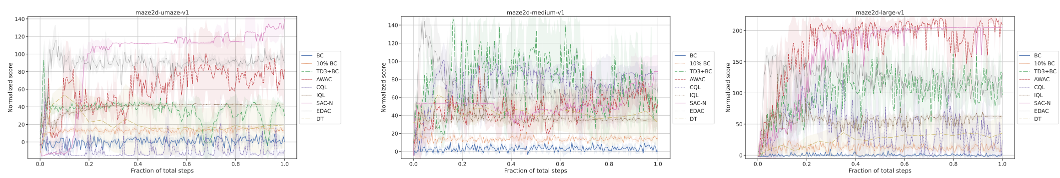
(c)

Figure 7: Training curves for Hopper task.
(a) Medium dataset, (b) Medium-expert dataset, (c) Medium-replay dataset



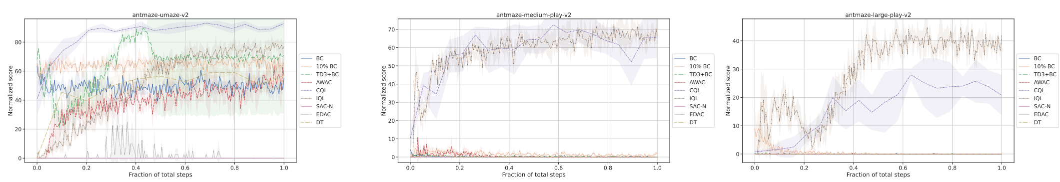
(a) (b) (c)

Figure 8: Training curves for Walker2d task.
(a) Medium dataset, (b) Medium-expert dataset, (c) Medium-replay dataset

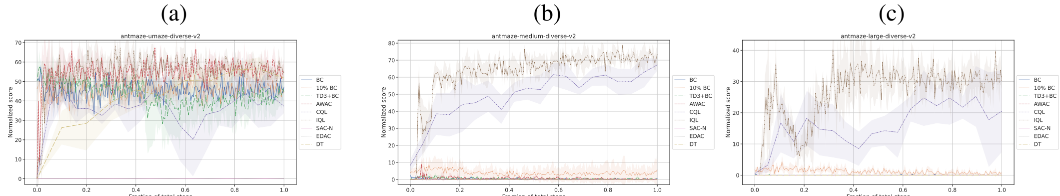


(a) (b) (c)

Figure 9: Training curves for Maze2d task.
(a) Medium dataset, (b) Medium-expert dataset, (c) Medium-replay dataset

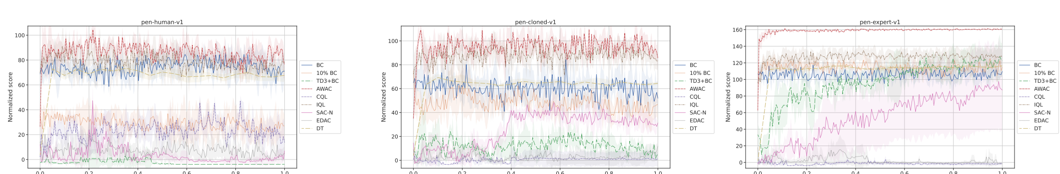


(a) (b) (c)



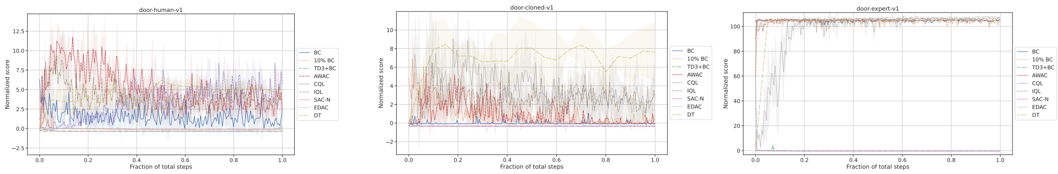
(d) (e) (f)

Figure 10: Training curves for AntMaze task.
(a) Umaze dataset, (b) Medium-play dataset, (c) Large-play dataset, (d) Umaze-diverse dataset, (e) Medium-diverse dataset, (f) Large-diverse dataset



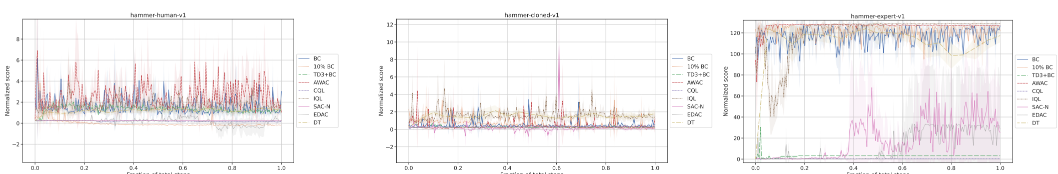
(a) (b) (c)

Figure 11: Training curves for Pen task.
(a) Human dataset, (b) Colned dataset, (c) Expert dataset



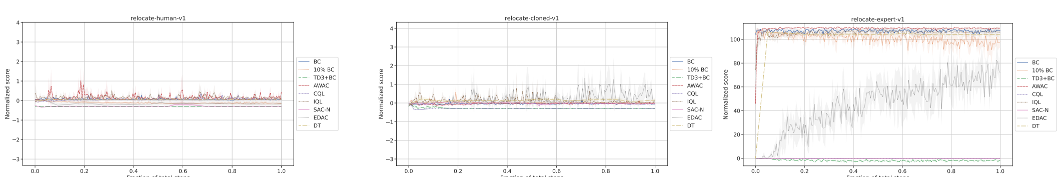
(a) (b) (c)

Figure 12: Training curves for Door task.
(a) Human dataset, (b) Colned dataset, (c) Expert dataset



(a) (b) (c)

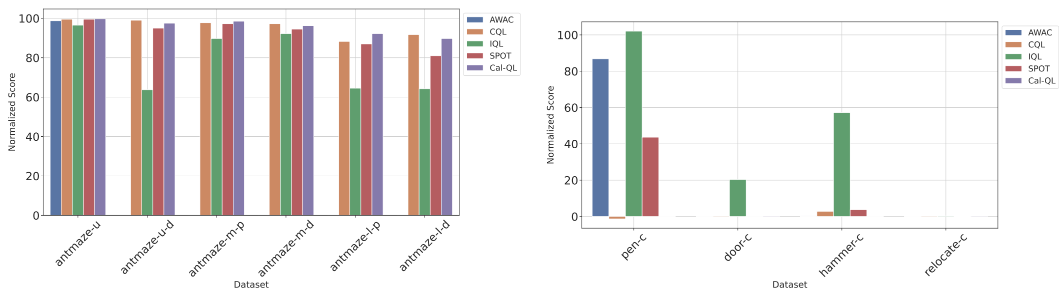
Figure 13: Training curves for Hammer task.
(a) Human dataset, (b) Colned dataset, (c) Expert dataset



(a) (b) (c)

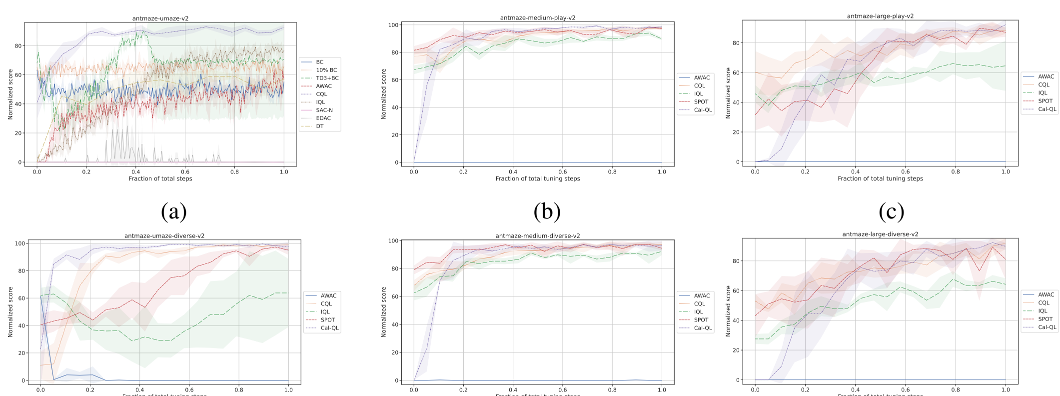
Figure 14: Training curves for Relocate task.
(a) Human dataset, (b) Colned dataset, (c) Expert dataset

A.2 Offline-to-online



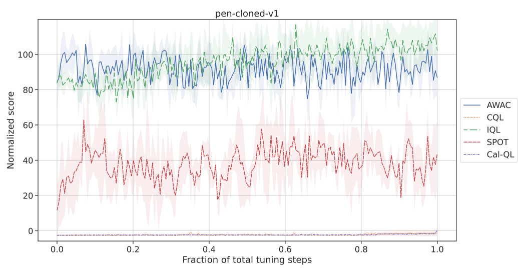
(a) (b)

Figure 15: Graphical representation of the normalized performance of the last trained policy on D4RL after online tuning averaged over 4 random seeds.
(a) AntMaze datasets (b) Adroit datasets

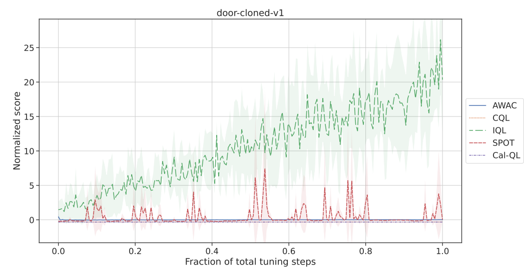


(a) (b) (c) (d) (e) (f)

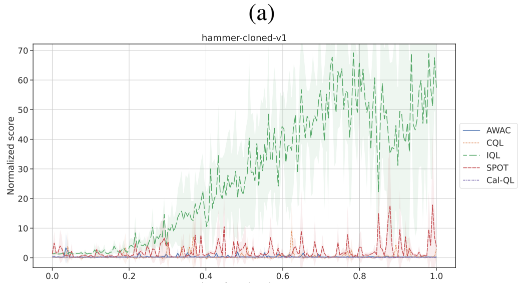
Figure 16: Training curves for AntMaze task during online tuning.
(a) Umaze dataset, (b) Medium-play dataset, (c) Large-play dataset, (d) Umaze-diverse dataset, (e) Medium-diverse dataset, (f) Large-diverse dataset



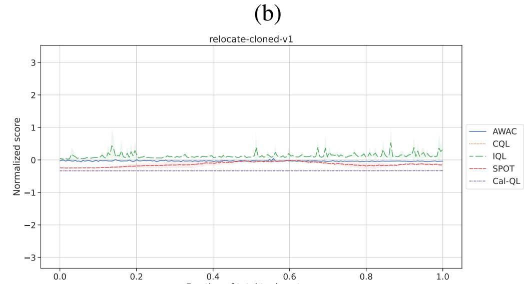
(a)



(b)



(c)



(d)

Figure 17: Training curves for Adroit Cloned task during online tuning.
(a) Pen, (b) Door, (c) Hammer, (d) Relocate

B Weights&Biases Tracking

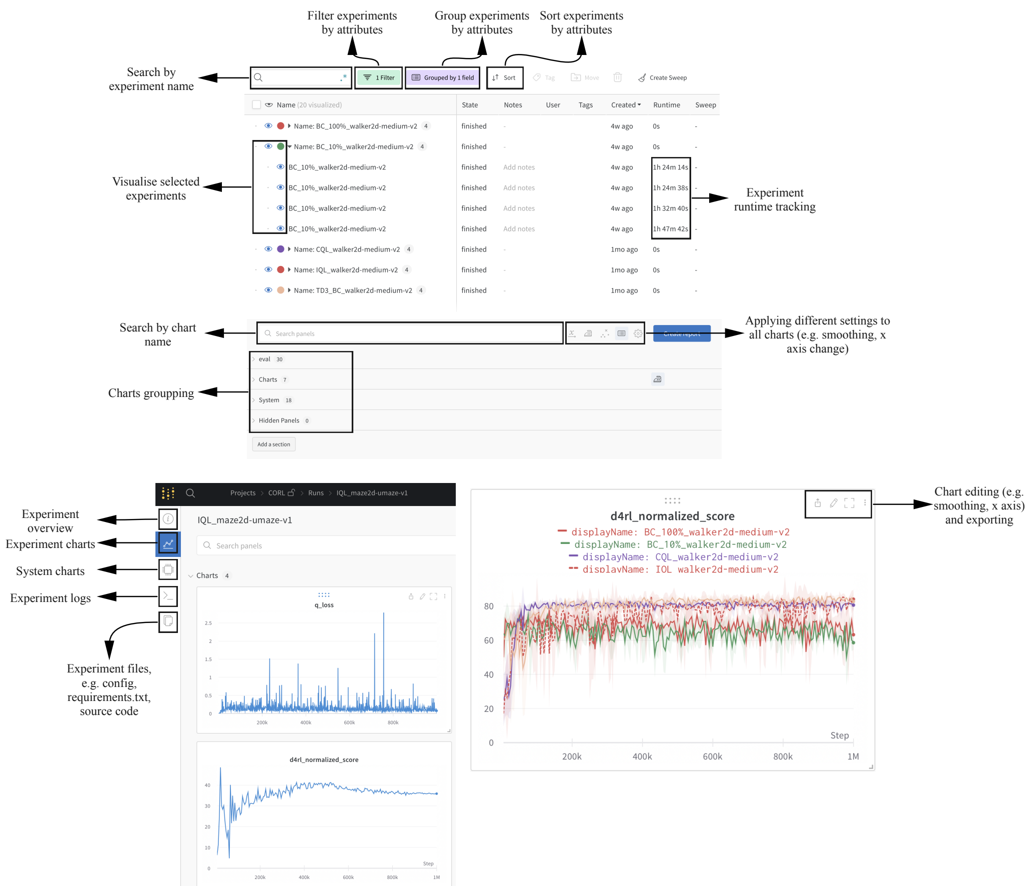


Figure 18: Screenshots of Weights&Biases experiment tracking interface.

C License

Our codebase is released under Apache License 2.0. The D4RL datasets (Fu et al., 2020) are released under Apache License 2.0.

D Experimental Details

We modify reward on AntMaze task by subtracting 1 from reward as it is done in previous works except CQL and Cal-QL, where (0, 1) are mapped into (-5, 5).

We used original implementation of TD3 + BC[14], SAC-N/EDAC[15], SPOT[16] and custom implementations of IQL[17] and CQL/Cal-QL[18] as the basis for ours.

For most of the algorithms and datasets, we use default hyperparameters if available. Configuration files for every algorithm and environment are presented in our GitHub repository. Hyperparameters are also provided in subsection D.2.

All the experiments ran using V100 and A100 GPUs, which took approximately 5000 hours of compute in total.

D.1 Number of update steps and evaluation rate

Following original work, SAC-N and EDAC are trained for 3 million steps (except AntMaze, which is trained for 1 million steps) in order to obtain state-of-the-art performance and tested every 10000 steps. Decision Transformer (DT) training is splitted into datasets pass epochs. We train DT for 50 epochs on each dataset and evaluate every 5 epochs. All other algorithms are trained for 1 million steps and evaluated every 5000 steps (50000 for AntMaze). We evaluate every policy for 10 episodes on Gym-MuJoCo and Adroit tasks and for 100 for Maze2d and AntMaze tasks.

D.2 Hyperparameters

Table 5: BC and BC-$N\%$ hyperparameters. † used for the best trajectories choice.

	Hyperparameter	Value
BC hyperparameters	Optimizer	Adam (Kingma & Ba, 2014)
	Learning Rate	3e-4
	Mini-batch size	256
Architecture	Policy hidden dim	256
	Policy hidden layers	2
	Policy activation function	ReLU
BC-$N\%$ hyperparameters	Ratio of best trajectories used	0.1
	Discount factor†	1.0
	Max trajectory length†	1000

[14] https://github.com/sfujim/TD3_BC

[15] https://github.com/snu-mllab/EDAC

[16] https://github.com/thuml/SPOT

[17] https://github.com/gwthomas/IQL-PyTorch

[18] https://github.com/young-geng/CQL

Table 6: TD3+BC hyperparameters.

	Hyperparameter	Value
TD3 hyperparameters	Optimizer	Adam (Kingma & Ba, 2014)
	Critic learning rate	3e-4
	Actor learning rate	3e-4
	Mini-batch size	256
	Discount factor	0.99
	Target update rate	5e-3
	Policy noise	0.2
	Policy noise clipping	(-0.5, 0.5)
	Policy update frequency	2
Architecture	Critic hidden dim	256
	Critic hidden layers	2
	Critic activation function	ReLU
	Actor hidden dim	256
	Actor hidden layers	2
	Actor activation function	ReLU
TD3+BC hyperparameters	α	2.5

Table 7: CQL and Cal-QL hyperparameters. Note: used hyperparameters are suboptimal on Adroit for the implementation we provide.

	Hyperparameter	Value
SAC hyperparameters	Optimizer	Adam (Kingma & Ba, 2014)
	Critic learning rate	3e-4
	Actor learning rate	1e-4
	Mini-batch size	256
	Discount factor	0.99
	Target update rate	5e-3
	Target entropy	-1 · Action Dim
	Entropy in Q target	False
Architecture	Critic hidden dim	256
	Critic hidden layers	5, AntMaze 3, otherwise
	Critic activation function	ReLU
	Actor hidden dim	256
	Actor hidden layers	3
	Actor activation function	ReLU
CQL hyperparameters	Lagrange	True, Maze2d and AntMaze False, otherwise
	Offline α	1.0, Adroit 5.0, AntMaze 10.0, otherwise
	Lagrange gap	5, Maze2d 0.8, AntMaze
	Pre-training steps	0
	Num sampled actions (during eval)	10
	Num sampled actions (logsumexp)	10
Cal-QL hyperparameters	Mixing ratio	0.5
	Online α	1.0, Adroit 5.0, AntMaze

Table 8: IQL hyperparameters.

	Hyperparameter	Value
IQL hyperparameters	Optimizer	Adam (Kingma & Ba, 2014)
	Critic learning rate	3e-4
	Actor learning rate	3e-4
	Value learning rate	3e-4
	Mini-batch size	256
	Discount factor	0.99
	Target update rate	5e-3
	Learning rate decay	Cosine
	Deterministic policy	True, Hopper Medium and Medium-replay False, otherwise
	β	6.0, Hopper Medium-expert 10.0, AntMaze 3.0, otherwise
	τ	0.9, AntMaze 0.5, Hopper Medium-expert 0.7, otherwise
Architecture	Critic hidden dim	256
	Critic hidden layers	2
	Critic activation function	ReLU
	Actor hidden dim	256
	Actor hidden layers	2
	Actor activation function	ReLU
	Value hidden dim	256
	Value hidden layers	2
	Value activation function	ReLU

Table 9: AWAC hyperparameters.

	Hyperparameter	Value
AWAC hyperparameters	Optimizer	Adam (Kingma & Ba, 2014)
	Critic learning rate	3e-4
	Actor learning rate	3e-4
	Mini-batch size	256
	Discount factor	0.99
	Target update rate	5e-3
	λ	0.1, Maze2d, AntMaze 0.3333, otherwise
Architecture	Critic hidden dim	256
	Critic hidden layers	2
	Critic activation function	ReLU
	Actor hidden dim	256
	Actor hidden layers	2
	Actor activation function	ReLU

Table 10: SAC-N and EDAC hyperparameters.

	Hyperparameter	Value
SAC hyperparameters	Optimizer	Adam (Kingma & Ba, 2014)
	Critic learning rate	3e-4
	Actor learning rate	3e-4
	α learning rate	3e-4
	Mini-batch size	256
	Discount factor	0.99
	Target update rate	5e-3
	Target entropy	-1 · Action Dim
Architecture	Critic hidden dim	256
	Critic hidden layers	3
	Critic activation function	ReLU
	Actor hidden dim	256
	Actor hidden layers	3
	Actor activation function	ReLU
SAC-N hyperparameters	Number of critics	10, HalfCheetah
		20, Walker2d
		25, AntMaze
		200, Hopper Medium-expert, Medium-replay
		500, Hopper Medium
EDAC hyperparameters	Number of critics	10, HalfCheetah
		10, Walker2d, AntMaze
		50, Hopper
	μ	5.0, HalfCheetah Medium-expert, Walker2d Medium-expert
		1.0, otherwise

Table 11: DT hyperparameters.

	Hyperparameter	Value
DT hyperparameters	Optimizer	AdamW (Loshchilov & Hutter, 2017)
	Batch size	256, AntMaze 4096, otherwise
	Return-to-go conditioning	(12000, 6000), HalfCheetah (3600, 1800), Hopper (5000, 2500), Walker2d (160, 80), Maze2d umaze (280, 140), Maze2d medium and large (1, 0.5), AntMaze (3100, 1550), Pen (2900, 1450), Door (12800, 6400), Hammer (4300, 2150), Relocate
	Reward scale	1.0, AntMaze 0.001, otherwise
	Dropout	0.1
	Learning rate	0.0008
	Adam betas	(0.9, 0.999)
	Clip grad norm	0.25
	Weight decay	0.0003
	Total gradient steps	100000
	Linear warmup steps	10000
Architecture	Number of layers	3
	Number of attention heads	1
	Embedding dimension	128
	Activation function	GELU

Table 12: SPOT hyperparameters.

	Hyperparameter	Value
VAE hyperparameters	Optimizer	Adam (Kingma & Ba, 2014)
	Learning rate	1e-3
	Mini-batch size	256
	Number of iterations	10^5
	KL term weight	0.5
VAE architecture	Encoder hidden dim	750
	Encoder layers	3
	Latent dim	2 $\times$ action dim
	Decoder hidden dim	750
	Decoder layers	3
TD3 hyperparameters	Optimizer	Adam (Kingma & Ba, 2014)
	Critic learning rate	3e-4
	Actor learning rate	1e-4
	Mini-batch size	256
	Discount factor	0.99
	Target update rate	5e-3
	Policy noise	0.2
	Policy noise clipping	(-0.5, 0.5)
	Policy update frequency	2
Architecture	Critic hidden dim	256
	Critic hidden layers	2
	Critic activation function	ReLU
	Actor hidden dim	256
	Actor hidden layers	2
	Actor activation function	ReLU
SPOT hyperparameters	λ	0.05, 0.1, 0.2, 0.5, 1.0, 2.0, AntMaze 1.0, Adroit