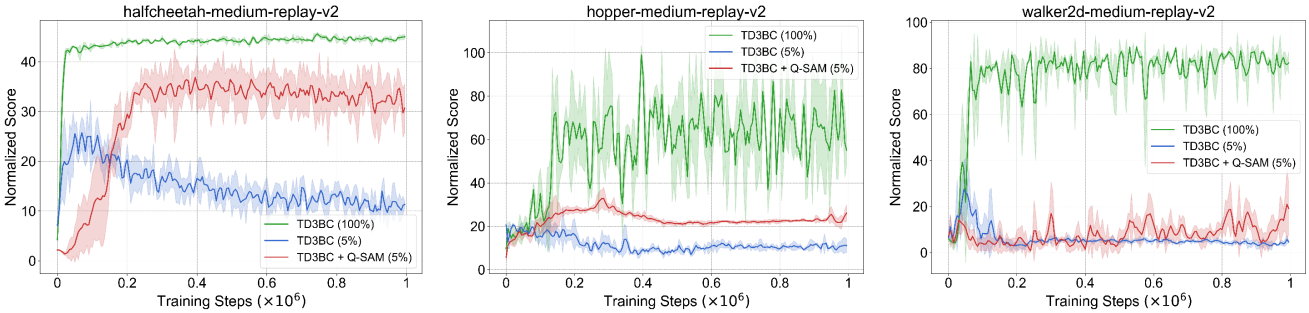

Figure 1. The small-sample performance of different algorithms. We choose TD3BC as the baseline algorithm. The ratio 100% indicates training with the full dataset, while 5% indicates training with only 1/20th of the full dataset. The shaded area comes from four different seeds {0, 10, 100, 1000}.

Table 1. Comparison of running time and performance of different algorithms. The time unit is minutes, and all experiments run on a server equipped with an Intel® Xeon® Gold 6254 CPU @ 3.10GHz and NVIDIA GeForce RTX 3090 GPU.

Algorithm	Training Steps	halfcheetah-mr		hopper-mr		walker2d-mr	
		Runtime (m)	Score	Runtime (m)	Score	Runtime (m)	Score
TD3BC	1M	142.1	44.8±0.6	137.3	64.4±21.5	135.7	85.6±4.0
	2M	290.7	45.0±0.4	271.5	65.4±21.7	261.4	68.5±11.4
TD3BC + Q-SAM	1M	276.9	**45.3±0.4**	296.1	**87.3±5.0**	298.1	**90.1±1.0**

Table 2. Performance under different parameters λ. The λ comes from WSAM in Q-SAM, and its use generally leads to performance improvements compared to using SAM alone. "Average" indicates that using CQL and IQL as a baseline can share parameters, while TD3BC uses a different set.

Algorithm	Parameter	halfcheetah			hopper			walker2d			Average	Proportion
		-m	-mr	-me	-m	-mr	-me	-m	-mr	-me		
TD3BC + Q-SAM	$\lambda = 0.1$	48.4	**45.3**	**93.9**	63.7	76.8	**105.1**	84.2	87.5	**111.5**	79.6	5/9
	$\lambda = 0.9$	48.2	45.0	91.7	**65.7**	**87.3**	104.6	**85.4**	**90.1**	111.2	**81.0**	4/9
TD3BC + SAM	–	**48.5**	44.6	90.3	60.9	69.4	101.5	82.9	86.7	109.0	77.1	–
CQL + Q-SAM	$\lambda = 0.1$	**47.3**	**45.6**	**95.6**	**70.3**	**95.5**	109.5	**84.2**	**84.4**	111.8	**82.7**	7/9
	$\lambda = 0.9$	47.1	45.5	94.9	65.5	95.3	**109.8**	82.8	80.7	**112.3**	81.5	2/9
CQL + SAM	–	47.1	45.5	94.1	59.7	94.9	108.7	81.7	74.9	109.6	79.6	–
IQL + Q-SAM	$\lambda = 0.1$	**49.2**	**45.1**	**95.9**	**75.3**	**103.0**	111.9	**87.6**	88.3	**111.7**	**85.3**	7/9
	$\lambda = 0.9$	48.8	44.6	95.5	70.9	102.1	**113.4**	87.4	**90.9**	111.6	85.0	2/9
IQL + SAM	–	48.4	44.6	95.0	68.5	100.5	109.4	86.9	87.2	111.3	83.5	–

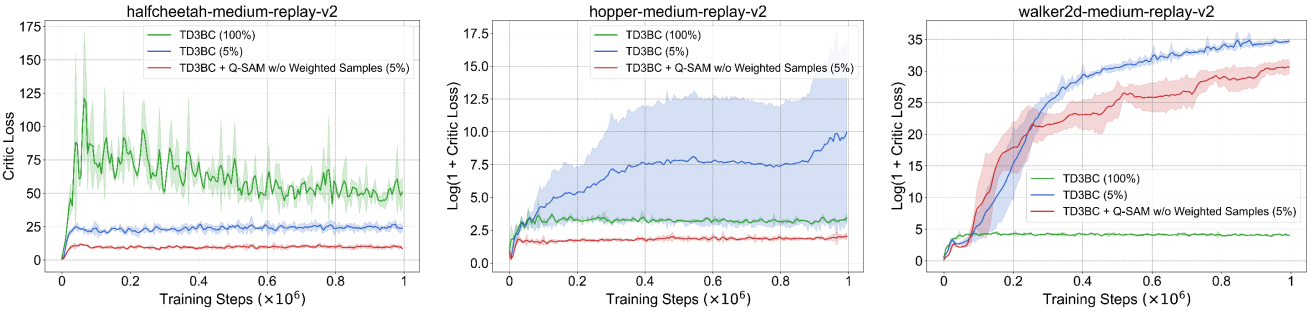

Figure 2. Critic loss values in the small-sample experiment (Figure 1 above). The critic loss value can serve as a statistical measure of the Bellman error, where Log(1 + Critic Loss) indicates taking the logarithm of the loss value to address scenarios of loss explosion. We removed the *Weighted Samples* from Q-SAM to better emphasize the impact of **Weighted Sharpness** in reducing Bellman error.