Appendix

A More toy experiments

Here we describe an additional toy experiment on a bandit task. Actions are again in a real-valued 2D space, $\boldsymbol{a} \in [-1, 1]^2$. The offline data $\mathcal{D} = \{(\boldsymbol{a}_i)\}_{i=1}^{10000}$ are collected by sampling actions equally from four Gaussian distributions with centers $\boldsymbol{\mu} \in \{(-0.8, 0.8), (0.8, 0.8), (0.8, -0.8), (-0.8, -0.8)\}$ and standard deviations $\boldsymbol{\sigma}_d = (0.05, 0.05)$, as depicted in the first panel of Figure 4. We conduct the same experiments as the ones in our main paper (see figure 1) and show the performance in Figure 4. The only difference in this experiment is that the samples are now in the corners of the ation space.

For behavior-cloning experiments, we observe that only our diffusion model could recover the original data distribution while the prior regularization methods fail in some way. For example, CVAE could only capture the two diagonal modes and place density between them, while MMD tends to align density around the boundaries because of its Tanh-Gaussian policy. For Q-learning experiments, we observe that the prior regularization methods typically push the policy to converge to sub-optimal solutions in BCQ and BEAR while preventing the policy of TD3+BC from being concentrated on the right corner. However, the policy of Diffusion-QL successfully converges to the optimal bottom corner. The ablation study experiments are consistent with our conclusion in the main paper.

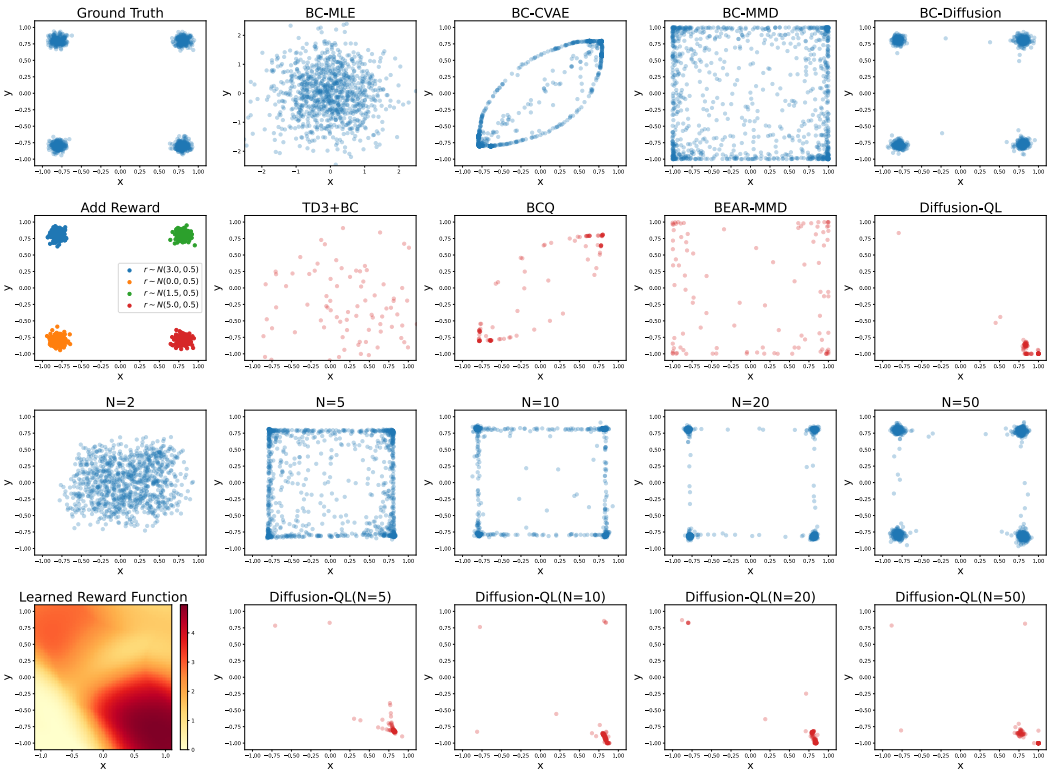

Figure 4: Bandit experiment with a strong multi-modal behavior policy. The first row shows the comparison of behavior-cloning results between our method and prior methods. The second row shows the comparison results with Q-learning involved. The third row shows the ablation study of N for BC-Diffusion. The fourth row shows the learned reward function and the ablation study of N for Diffusion-QL.

B Implementation Details

Diffusion policy. We build our policy as an MLP-based conditional diffusion model. Following the parameterization of Ho et al. (2020), the model itself is a residual model, $\boldsymbol{\epsilon}_\theta(\boldsymbol{a}^i, \boldsymbol{s}, i)$, where the i is the last timestep and $\boldsymbol{s}$ is the state condition. We model $\boldsymbol{\epsilon}_\theta$ as a 3-layer MLPs with Mish activations and we use 256 hidden units for all networks. The input of $\boldsymbol{\epsilon}_\theta$ is the concatenation of the last step action vector, the current state vector, and the sinusoidal positional embedding of timestep i. The output of $\boldsymbol{\epsilon}_\theta$ is the predicted residual at diffusion timestep i.

Q networks. We build two Q networks with the same MLP setting as our diffusion policy, which has 3-layer MLPs with Mish activations and 256 hidden units for all networks.

C Experimental Details

We train our algorithm with 2000 epochs for Gym tasks and 1000 epochs for the other tasks, where each epoch consists of 1000 gradient steps. For the Gym locomotion tasks, we average mean returns over 6 independently trained models and 10 trajectories per mode. For the other tasks, we average over 6 independently trained models and 100 evaluation trajectories. Following the convention of supervised learning, the best performed model is saved during training and used for final evaluation.

We use the original task rewards from MuJoCo Gym and Kitchen tasks. We standardize Adroit task rewards for training stability. We modify the rewards according to the suggestion of CQL (Kumar et al., 2020) for the AntMaze datasets.

D Hyperparameters

For Diffusion-QL, we consider three hyperparameters in total: learning rate, Q-learning weight η, and whether to use max Q backup from CQL (Kumar et al., 2020). For learning rate, we consider values in the grid $\{1 \times 10^{-3}, 3 \times 10^{-4}, 3 \times 10^{-5}\}$ for the policy, while we use a fixed learning rate, 3×10^{-4}, for Q-networks. For η, we consider values according to the characteristics of different domains, as we mentioned in the description of datasets that Adroit and Kitchen tasks require more policy regularization and AntMaze tasks require more Q-learning. For max Q backup, we only apply it on AntMaze tasks. Based on these considerations, we provide our hyperparameter setting in Table 3.

E Optimal Results

If a small amount of online experience is provided during the evaluation stage for model selection, we can pick the best models during training via online evaluations (similar to early stopping in supervised learning). This regime provides a further boost in the performance of Diffusion-QL as shown in Table 4.

Tasks	**learning rate**	η	**max Q backup**
halfcheetah-medium-v2	3×10^{-4}	1.0	False
hopper-medium-v2	3×10^{-4}	1.0	False
walker2d-medium-v2	3×10^{-4}	1.0	False
halfcheetah-medium-replay-v2	3×10^{-4}	1.0	False
hopper-medium-replay-v2	3×10^{-4}	1.0	False
walker2d-medium-replay-v2	3×10^{-4}	1.0	False
halfcheetah-medium-expert-v2	3×10^{-4}	1.0	False
hopper-medium-expert-v2	3×10^{-4}	1.0	False
walker2d-medium-expert-v2	3×10^{-4}	1.0	False
antmaze-umaze-v0	3×10^{-4}	0.5	False
antmaze-umaze-diverse-v0	3×10^{-4}	2.0	True
antmaze-medium-play-v0	1×10^{-3}	2.0	True
antmaze-medium-diverse-v0	3×10^{-4}	3.0	True
antmaze-large-play-v0	3×10^{-4}	4.5	True
antmaze-large-diverse-v0	3×10^{-4}	3.5	True
pen-human-v1	3×10^{-5}	0.15	False
pen-cloned-v1	3×10^{-5}	0.1	False
kitchen-complete-v0	3×10^{-4}	0.005	False
kitchen-partial-v0	3×10^{-4}	0.005	False
kitchen-mixed-v0	3×10^{-4}	0.005	False

Table 3: Hyperparameter settings of all selected tasks.

Table 4: Performance comparison with online model selection and offline model selection.

AntMaze Tasks	**Diffusion-QL (Offline)**	**Diffusion-QL (Online)**
halfcheetah-medium-v2	51.1 ± 0.5	**51.5** ± 0.3
hopper-medium-v2	90.5 ± 4.6	**96.6** ± 3.4
walker2d-medium-v2	87.0 ± 0.9	**87.3** ± 0.5
halfcheetah-medium-replay-v2	47.8 ± 0.3	**48.3** ± 0.2
hopper-medium-replay-v2	101.3 ± 0.6	**102.0** ± 0.4
walker2d-medium-replay-v2	95.5 ± 1.5	**98.0** ± 0.5
halfcheetah-medium-expert-v2	96.8 ± 0.3	**97.2** ± 0.4
hopper-medium-expert-v2	111.1 ± 1.3	**112.3** ± 0.8
walker2d-medium-expert-v2	110.1 ± 0.3	**111.2** ± 0.9
Average	88.0	**89.3**
AntMaze Tasks	**Diffusion-QL (Offline)**	**Diffusion-QL (Online)**
antmaze-umaze-v0	93.4 ± 3.4	**96.0** ± 3.3
antmaze-umaze-diverse-v0	66.2 ± 8.6	**84.0** ± 10.1
antmaze-medium-play-v0	76.6 ± 10.8	**79.8** ± 8.7
antmaze-medium-diverse-v0	78.6 ± 10.3	**82.0** ± 9.5
antmaze-large-play-v0	46.4 ± 8.3	**49.0** ± 9.4
antmaze-large-diverse-v0	56.6 ± 7.6	**61.7** ± 8.2
Average	69.6	**75.4**
Adroit Tasks	**Diffusion-QL (Offline)**	**Diffusion-QL (Online)**
pen-human-v1	72.8 ± 9.6	**75.7** ± 9.0
pen-cloned-v1	57.3 ± 11.9	**60.8** ± 11.8
Average	65.1	**68.3**
Kitchen Tasks	**Diffusion-QL (Offline)**	**Diffusion-QL (Online)**
kitchen-complete-v0	84.0 ± 7.4	**84.5** ± 6.1
kitchen-partial-v0	60.5 ± 6.9	**63.7** ± 5.2
kitchen-mixed-v0	62.6 ± 5.1	**66.6** ± 3.3
Average	69.0	**71.6**