In addition to our work, [7] have also explored the use of supplementary data in the offline setting. However, their approach (called MILO) is based on adversarial imitation learning. Specifically, MILO learns a transition model from the supplementary dataset and performs adversarial imitation learning within the learned model. In contrast, our proposed method, ISW-BC, tackles the challenge of scarce expert data by identifying and utilizing expert-style samples that are hidden within the supplementary dataset. MILO has an imitation gap bound of $\mathcal{O}(H\sqrt{\frac{|\mathcal{S}|}{N_{\mathrm{E}}}} + H^2|\mathcal{S}|\sqrt{\frac{|\mathcal{A}|}{N_{\mathrm{S}}/\mu}})$ in theory. However, MILO makes different assumptions about the data collection procedure compared with ISW-BC. Consequently, the imitation gap bounds of MILO and ISW-BC are incomparable.

The problem considered in this paper is related to IL with a single imperfect dataset [44, 6, 41, 42, 36, 25]. In particular, the supplementary dataset in our set-up can also be viewed as imperfect demonstrations. However, our problem setting differs from IL with imperfect demonstrations in two key aspects. First, in IL with imperfect demonstrations, they either pose strong assumptions [41, 36, 25] or require auxiliary information (e.g., confidence scores on imperfect trajectories) on the imperfect dataset [44, 6]. In contrast, we assume access to a small number of expert trajectories to identify in-expert-distribution data. Second, most works [44, 6, 41, 42] in IL with imperfect demonstrations require online environment interactions while we focus on the offline setting.

C Proof of Results in Section 4

Recall the objective of BC in Eq. (1):

$$\pi^{\mathrm{BC}} \in \max_{\pi} \sum_{h=1}^{H} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \widehat{d_h^{\mathrm{E}}}(s,a) \log \pi_h(a|s),$$

where $\widehat{d_h^{\mathrm{E}}}(s,a) = n_h^{\mathrm{E}}(s,a)/N_{\mathrm{tot}}$ is the empirical state-action distribution in the expert dataset, and $n_h^{\mathrm{E}}(s,a)$ is the number of expert trajectories such that their state-action pairs are equal to (s,a) in time step h. With the tabular representations, we can obtain a closed-formed solution to the above optimization problem.

$$\pi_h^{\mathrm{BC}}(a|s) = \begin{cases} \frac{n_h^{\mathrm{E}}(s,a)}{n_h^{\mathrm{E}}(s)} & \text{if } n_h^{\mathrm{E}}(s) > 0 \\ \frac{1}{|\mathcal{A}|} & \text{otherwise} \end{cases} \tag{8}$$

where $n_h^{\mathrm{E}}(s) \triangleq \sum_{a'} n_h^{\mathrm{E}}(s,a')$. Analogously, we also have a closed-form solution for NBCU in the tabular setting:

$$\pi_h^{\mathrm{NBCU}}(a|s) = \begin{cases} \frac{n_h^{\mathrm{U}}(s,a)}{n_h^{\mathrm{U}}(s)} & \text{if } n_h^{\mathrm{U}}(s) > 0 \\ \frac{1}{|\mathcal{A}|} & \text{otherwise} \end{cases} \tag{9}$$

We will discuss the generalization performance of NBCU later.

In the proof, we frequently use the notation $\lesssim$ and $\gtrsim$. In particular, $a(n) \lesssim b(n)$ means that there exist $C, n_0 > 0$ such that $a(n) \leq Cb(n)$ for all $n \geq n_0$. In our context, n usually refers to the number of trajectories. For any two distributions P and Q over a finite set $\mathcal{X}$, we define the total variation distance as

$$\mathrm{TV}(P,Q) = \frac{1}{2}\sum_{x\in\mathcal{X}} |P(x) - Q(x)| = \|P - Q\|_1 .$$

C.1 Proof of Theorem 1

When $|\mathcal{D}^{\mathrm{E}}| \geq 1$, by [32, Theorem 4.2], we have the following imitation gap bound for BC:

$$V(\pi^{\mathrm{E}}) - \mathbb{E}_{\mathcal{D}^{\mathrm{E}}}\left[V(\pi^{\mathrm{BC}})\right] \leq \frac{4|\mathcal{S}|H^2}{9|\mathcal{D}^{\mathrm{E}}|}.$$

When $|\mathcal{D}^{\mathrm{E}}| = 0$, we simply have that

$$V(\pi^{\mathrm{E}}) - \mathbb{E}_{\mathcal{D}^{\mathrm{E}}}\left[V(\pi^{\mathrm{BC}})\right] \leq H.$$

Therefore, we have the following unified bound.

$$V(\pi^{\mathrm{E}}) - \mathbb{E}_{\mathcal{D}^{\mathrm{E}}}\left[V(\pi^{\mathrm{BC}})\right] \leq \frac{|\mathcal{S}|H^2}{\max\{|\mathcal{D}^{\mathrm{E}}|, 1\}} \leq \frac{2|\mathcal{S}|H^2}{|\mathcal{D}^{\mathrm{E}}|+1}.$$

The last inequality follows that $\max\{x, 1\} \geq (x+1)/2$ for any $x \geq 0$. Finally, notice that $|\mathcal{D}^{\mathrm{E}}|$ follows a binomial distribution by Assumption 1, i.e., $|\mathcal{D}^{\mathrm{E}}| \sim \mathrm{Bin}(N_{\mathrm{tot}}, \eta)$. By Lemma 3, we have that $\mathbb{E}[1/(|\mathcal{D}|^{E}+1)] \leq N_{\mathrm{tot}}\eta$, so

$$V(\pi^{\mathrm{E}}) - \mathbb{E}\left[V(\pi^{\mathrm{BC}})\right] \leq \mathbb{E}\left[\frac{2|\mathcal{S}|H^2}{|\mathcal{D}^{\mathrm{E}}|+1}\right] \leq \frac{2|\mathcal{S}|H^2}{N_{\mathrm{tot}}\eta} = \frac{2|\mathcal{S}|H^2}{N_{\mathrm{E}}},$$

which completes the proof.

C.2 Proof of Theorem 2

For analysis, we first define the mixture state-action distribution as follows.

$$\begin{aligned} d_h^{\mathrm{mix}}(s,a) &\triangleq \eta d_h^{\pi^{\mathrm{E}}}(s,a) + (1-\eta) d_h^{\pi^{\beta}}(s,a), \\ d_h^{\mathrm{mix}}(s) &\triangleq \sum_{a \in \mathcal{A}} d_h^{\mathrm{mix}}(s,a),\ \forall (s,a) \in \mathcal{S} \times \mathcal{A},\ \forall h \in [H]. \end{aligned}$$

By Assumption 1, in the population level, the marginal state-action distribution of union dataset $\mathcal{D}^{\mathrm{U}}$ in time step h is exactly d_h^{mix}. That is, $d_h^{\mathrm{U}}(s,a) = d_h^{\mathrm{mix}}(s,a),\ \forall (s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$. Then we define the mixture policy π^{mix} induced by d^{mix} as follows.

$$\pi_h^{\mathrm{mix}}(a|s) = \begin{cases} \frac{d_h^{\mathrm{mix}}(s,a)}{d_h^{\mathrm{mix}}(s)} & \text{if } d_h^{\mathrm{mix}}(s) > 0, \\ \frac{1}{|\mathcal{A}|} & \text{otherwise.} \end{cases} \quad \forall (s,a) \in \mathcal{S} \times \mathcal{A}\ , \forall h \in [H]. \tag{10}$$

From the theory of Markov Decision Processes, we know that (see, e.g., [31])

$$\forall h \in [H], \forall (s,a) \in \mathcal{S} \times \mathcal{A}, \quad d_h^{\pi^{\mathrm{mix}}}(s,a) = d_h^{\mathrm{mix}}(s,a).$$

Therefore, we can obtain that the marginal state-action distribution of union dataset $\mathcal{D}^{\mathrm{U}}$ in time step h is exactly $d_h^{\pi^{\mathrm{mix}}}$. Then we have the following decomposition.

$$\begin{aligned} \mathbb{E}\left[V(\pi^{\mathrm{E}}) - V(\pi^{\mathrm{NBCU}})\right] &= \mathbb{E}\left[V(\pi^{\mathrm{E}}) - V(\pi^{\mathrm{mix}}) + V(\pi^{\mathrm{mix}}) - V(\pi^{\mathrm{NBCU}})\right] \\ &= \mathbb{E}\left[V(\pi^{\mathrm{E}}) - V(\pi^{\mathrm{mix}})\right] + \mathbb{E}\left[V(\pi^{\mathrm{mix}}) - V(\pi^{\mathrm{NBCU}})\right] \\ &= V(\pi^{\mathrm{E}}) - V(\pi^{\mathrm{mix}}) + \mathbb{E}\left[V(\pi^{\mathrm{mix}}) - V(\pi^{\mathrm{NBCU}})\right]. \end{aligned}$$

For $V(\pi^{\mathrm{E}}) - V(\pi^{\mathrm{mix}})$, we have that

$$\begin{aligned} V(\pi^{\mathrm{E}}) - V(\pi^{\mathrm{mix}}) &= \sum_{h=1}^{H} \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} \left(d_h^{\pi^{\mathrm{E}}}(s,a) - d_h^{\pi^{\mathrm{mix}}}(s,a) \right) r_h(s,a) \\ &= \sum_{h=1}^{H} \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} \left(d_h^{\pi^{\mathrm{E}}}(s,a) - d_h^{\mathrm{mix}}(s,a) \right) r_h(s,a) \\ &= (1-\eta) \sum_{h=1}^{H} \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} \left(d_h^{\pi^{\mathrm{E}}}(s,a) - d_h^{\pi^{\beta}}(s,a) \right) r_h(s,a) \\ &= (1-\eta) \left(V(\pi^{\mathrm{E}}) - V(\pi^{\beta}) \right). \end{aligned} \tag{11}$$

The last equation follows the dual formulation of policy value (see, e.g., [31]), i.e., $V(\pi) = \sum_{h=1}^{H} \sum_{(s,a)} d_h^{\pi}(s,a) r_h(s,a)$ for any policy π. Besides, notice that $\mathbb{E}\left[V(\pi^{\mathrm{mix}}) - V(\pi^{\mathrm{NBCU}})\right]$ is exactly the imitation gap of BC when regarding π^{mix} and $\mathcal{D}^{\mathrm{U}}$ as the expert policy and expert dataset, respectively. Note that π^{mix} may be a stochastic policy. By [32, Theorem 4.4], we have the following imtiation gap bound

$$\mathbb{E}\left[V(\pi^{\mathrm{mix}}) - V(\pi^{\mathrm{NBCU}})\right] \lesssim \frac{|\mathcal{S}|H^2 \log(N_{\mathrm{tot}})}{N_{\mathrm{tot}}}. \tag{12}$$

Combining Eq. (11) and Eq. (12) yields that

$$\mathbb{E}\left[V(\pi^{\mathrm{E}}) - V(\pi^{\mathrm{NBCU}})\right] \lesssim (1-\eta)\left(V(\pi^{\mathrm{E}}) - V(\pi^{\beta})\right) + \frac{|\mathcal{S}|H^2\log(N_{\mathrm{tot}})}{N_{\mathrm{tot}}}.$$

C.3 Proof of Proposition 1

The hard instance in Proposition 1 builds on the Standard Imitation MDP proposed in [46]; see Figure 4 for illustration. For this MDP, each state is an absorbing state, i.e., $P_h(s|s,a) = 1$ for any s and a. This property is mainly used to facilitate probability calculation and does not change the nature of our analysis. Furthermore, by only taking the action a^1 (shown in green), the agent can obtain a reward of $+1$. Otherwise, the agent obtains a reward of 0 for the other action $a \neq a^1$. The initial state distribution is a uniform distribution, i.e., $\rho(s) = 1/|\mathcal{S}|$ for any $s \in \mathcal{S}$.

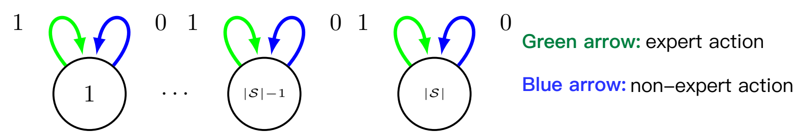

Figure 4: The Standard Imitation MDP in [46] corresponding to prove Proposition 1.

We consider that the expert policy π^{E} always takes the action a^1 (shown in green) while the behavioral policy π^{β} always takes another action a^2 (shown in blue). Formally, $\pi_h^{\mathrm{E}}(a^1|s) = 1$ and $\pi_h^{\beta}(a^2|s) = 1$ for any $s \in \mathcal{S}$ and $h \in [H]$. It is direct to calculate that $V(\pi^{\mathrm{E}}) = H$ and $V(\pi^{\beta}) = 0$. The supplementary dataset $\mathcal{D}^{\mathrm{S}}$ and the expert dataset $\mathcal{D}^{\mathrm{E}}$ are collected according to Assumption 1. The mixture state-action distribution (introduced in Appendix C.2) can be calculated as for any $s \in \mathcal{S}$ and $h \in [H]$:

$$\begin{aligned}
d_h^{\mathrm{mix}}(s,a^1) &= \eta d_h^{\pi^{\mathrm{E}}}(s,a^1) + (1-\eta) d_h^{\pi^{\beta}}(s,a^1) = \eta d_h^{\pi^{\mathrm{E}}}(s,a^1) = \eta\rho(s), \\
d_h^{\mathrm{mix}}(s,a^2) &= \eta d_h^{\pi^{\mathrm{E}}}(s,a^2) + (1-\eta) d_h^{\pi^{\beta}}(s,a^2) = (1-\eta) d_h^{\pi^{\beta}}(s,a^2) = (1-\eta)\rho(s).
\end{aligned}$$

Note that in the population level, the marginal distribution of the union dataset $\mathcal{D}^{\mathrm{U}}$ in time step h is exactly d_h^{mix}. The mixture policy induced by d^{mix} (introduced in Appendix C.2) can be formulated as

$$\pi_h^{\mathrm{mix}}(a^1|s) = \eta, \pi_h^{\mathrm{mix}}(a^2|s) = 1-\eta, \forall s \in \mathcal{S}, h \in [H].$$

Just like before, we have $d_h^{\pi^{\mathrm{mix}}}(s,a) = d_h^{\mathrm{mix}}(s,a)$. The policy value of π^{mix} can be calculated as

$$V(\pi^{\mathrm{mix}}) = \sum_{h=1}^{H} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\mathrm{mix}}(s,a) r_h(s,a) = \sum_{h=1}^{H}\sum_{s\in\mathcal{S}} d_h^{\mathrm{mix}}(s,a^1) = \eta H.$$

Recall from Eq. (9) that π^{NBCU} can be formulated as

$$\forall h \in [H], \quad \pi_h^{\mathrm{NBCU}}(a|s) = \begin{cases} \frac{n_h^{\mathrm{U}}(s,a)}{\sum_{a'} n_h^{\mathrm{U}}(s,a')} & \text{if} \sum_{a'} n_h^{\mathrm{U}}(s,a') > 0 \\ \frac{1}{|\mathcal{A}|} & \text{otherwise} \end{cases} \tag{13}$$

We can view that the BC's policy learned on the union dataset mimics the mixture policy π^{mix}. In the following part, we analyze the lower bound on the imitation gap of π^{NBCU}.

$$\begin{aligned}
\mathbb{E}\left[V(\pi^{\mathrm{E}}) - V(\pi^{\mathrm{NBCU}})\right] &= V(\pi^{\mathrm{E}}) - V(\pi^{\mathrm{mix}}) + \mathbb{E}\left[V(\pi^{\mathrm{mix}}) - V(\pi^{\mathrm{NBCU}})\right] \\
&= H - \eta H + \mathbb{E}\left[V(\pi^{\mathrm{mix}}) - V(\pi^{\mathrm{NBCU}})\right] \\
&= (1-\eta)(V(\pi^{\mathrm{E}}) - V(\pi^{\beta})) + \mathbb{E}\left[V(\pi^{\mathrm{mix}}) - V(\pi^{\mathrm{NBCU}})\right].
\end{aligned}$$

Then we consider the term $\mathbb{E}\left[V(\pi^{\mathrm{mix}}) - V(\pi^{\mathrm{NBCU}})\right]$.

$$\begin{aligned}
&V(\pi^{\mathrm{mix}}) - V(\pi^{\mathrm{NBCU}}) \\
&= \sum_{h=1}^{H} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \left(d_h^{\pi^{\mathrm{mix}}}(s,a) - d_h^{\pi^{\mathrm{NBCU}}}(s,a)\right) r_h(s,a)
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{h=1}^{H} \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} \rho(s) \left(\pi_h^{\text{mix}}(a|s) - \pi_h^{\text{NBCU}}(a|s) \right) r_h(s,a) \\
&= \sum_{h=1}^{H} \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} \rho(s) \left(\pi_h^{\text{mix}}(a|s) - \pi_h^{\text{NBCU}}(a|s) \right) r_h(s,a) \mathbb{I}\{n_h^{\text{U}}(s) > 0\} \\
&\quad + \sum_{h=1}^{H} \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} \rho(s) \left(\pi_h^{\text{mix}}(a|s) - \pi_h^{\text{NBCU}}(a|s) \right) r_h(s,a) \mathbb{I}\{n_h^{\text{U}}(s) = 0\}.
\end{aligned}
$$

We take expectation over the randomness in $\mathcal{D}^{\text{U}}$ on both sides and obtain that

$$
\mathbb{E}\left[V(\pi^{\text{mix}}) - V(\pi^{\text{NBCU}}) \right] \tag{14}
$$

$$
\begin{aligned}
&= \mathbb{E}\left[\sum_{h=1}^{H} \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} \rho(s) \left(\pi_h^{\text{mix}}(a|s) - \pi_h^{\text{NBCU}}(a|s) \right) r_h(s,a) \mathbb{I}\{n_h^{\text{U}}(s) > 0\} \right] \\
&\quad + \mathbb{E}\left[\sum_{h=1}^{H} \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} \rho(s) \left(\pi_h^{\text{mix}}(a|s) - \pi_h^{\text{NBCU}}(a|s) \right) r_h(s,a) \mathbb{I}\{n_h^{\text{U}}(s) = 0\} \right].
\end{aligned} \tag{15}
$$

For the first term in RHS, we have that

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{h=1}^{H} \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} \rho(s) \left(\pi_h^{\text{mix}}(a|s) - \pi_h^{\text{NBCU}}(a|s) \right) r_h(s,a) \mathbb{I}\{n_h^{\text{U}}(s) > 0\} \right] \\
&= \sum_{h=1}^{H} \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} \rho(s) r_h(s,a) \mathbb{E}\left[\left(\pi_h^{\text{mix}}(a|s) - \pi_h^{\text{NBCU}}(a|s) \right) \mathbb{I}\{n_h^{\text{U}}(s) > 0\} \right] \\
&= \sum_{h=1}^{H} \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} \rho(s) r_h(s,a) \mathbb{P}\left(n_h^{\text{U}}(s) > 0 \right) \mathbb{E}\left[\pi_h^{\text{mix}}(a|s) - \pi_h^{\text{NBCU}}(a|s) \mid n_h^{\text{U}}(s) > 0 \right] \\
&= 0.
\end{aligned}
$$

The last equation follows the fact that $\pi_h^{\text{NBCU}}(a|s)$ is an unbiased estimation of $\pi_h^{\text{mix}}(a|s)$, so $\mathbb{E}[\pi_h^{\text{mix}}(a|s) - \pi_h^{\text{NBCU}}(a|s) \mid n_h^{\text{U}}(s) > 0]$. For the second term in Eq. (15), we have that

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{h=1}^{H} \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} \rho(s) \left(\pi_h^{\text{mix}}(a|s) - \pi_h^{\text{NBCU}}(a|s) \right) r_h(s,a) \mathbb{I}\{n_h^{\text{U}}(s) = 0\} \right] \\
&= \sum_{h=1}^{H} \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} \rho(s) r_h(s,a) \mathbb{E}\left[\left(\pi_h^{\text{mix}}(a|s) - \pi_h^{\text{NBCU}}(a|s) \right) \mathbb{I}\{n_h^{\text{U}}(s) = 0\} \right] \\
&= \sum_{h=1}^{H} \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} \rho(s) r_h(s,a) \mathbb{P}\left(n_h^{\text{U}}(s) = 0 \right) \mathbb{E}\left[\pi_h^{\text{mix}}(a|s) - \pi_h^{\text{NBCU}}(a|s) \mid n_h^{\text{U}}(s) = 0 \right] \\
&= \sum_{h=1}^{H} \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} \rho(s) r_h(s,a) \mathbb{P}\left(n_h^{\text{U}}(s) = 0 \right) \left(\pi_h^{\text{mix}}(a|s) - \frac{1}{|\mathcal{A}|} \right) \\
&\overset{(a)}{=} \sum_{h=1}^{H} \sum_{s \in \mathcal{S}} \rho(s) \mathbb{P}\left(n_h^{\text{U}}(s) = 0 \right) \left(\eta - \frac{1}{|\mathcal{A}|} \right) \\
&\overset{(b)}{=} H \left(\eta - \frac{1}{|\mathcal{A}|} \right) \sum_{s \in \mathcal{S}} \rho(s) \mathbb{P}\left(n_1^{\text{U}}(s) = 0 \right).
\end{aligned}
$$

In the equation (a), we use the fact that $r_h(s, a^1) = 1$ but $r_h(s,a) = 0$ for any $a \neq a^1$. In the equation (b), since each state is an absorbing state, we have that $\mathbb{P}(n_h^{\text{U}}(s) = 0) = \mathbb{P}(n_1^{\text{U}}(s) = 0)$ for any $h \in [H]$. We consider two cases to address RHS of equation (b). In the first case of $\eta \geq 1/|\mathcal{A}|$, we directly have that

$$\mathbb{E}\left[\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\rho(s)\left(\pi_h^{\mathrm{mix}}(a|s)-\pi_h^{\mathrm{NBCU}}(a|s)\right)r_h(s,a)\mathbb{I}\{n_h^{\mathrm{U}}(s)=0\}\right]\geq 0.$$

By Eq. (15), we have that

$$\mathbb{E}\left[V(\pi^{\mathrm{mix}})-V(\pi^{\mathrm{NBCU}})\right]\geq 0,$$

which implies that

$$\mathbb{E}\left[V(\pi^{\mathrm{E}})-V(\pi^{\mathrm{NBCU}})\right]\geq (1-\eta)(V(\pi^{\mathrm{E}})-V(\pi^{\beta})).$$

In the second case of $\eta<1/|\mathcal{A}|$, we have that

$$\begin{aligned}
H\left(\eta-\frac{1}{|\mathcal{A}|}\right)\sum_{s\in\mathcal{S}}\rho(s)\mathbb{P}\left(n_1^{\mathrm{U}}(s)=0\right) &\overset{(a)}{\geq} -\left(\frac{1}{|\mathcal{A}|}-\eta\right)H\exp\left(-\frac{N_{\mathrm{tot}}}{|\mathcal{S}|}\right)\\
&\geq -(1-\eta)H\exp\left(-\frac{N_{\mathrm{tot}}}{|\mathcal{S}|}\right)\\
&\overset{(b)}{\geq} -\frac{(1-\eta)H}{2}.
\end{aligned}$$

In the inequality (a), we use that

$$\sum_{s\in\mathcal{S}}\rho(s)\mathbb{P}\left(n_1^{\mathrm{U}}(s)=0\right)=\sum_{s\in\mathcal{S}}\rho(s)(1-\rho(s))^{N_{\mathrm{tot}}}=\left(1-\frac{1}{|\mathcal{S}|}\right)^{N_{\mathrm{tot}}}\leq\exp\left(-\frac{N_{\mathrm{tot}}}{|\mathcal{S}|}\right).$$

The inequality (b) holds since we consider the range where $N_{\mathrm{tot}}\geq|\mathcal{S}|\log(2)$. By Eq. (15), we have that

$$\mathbb{E}\left[V(\pi^{\mathrm{mix}})-V(\pi^{\mathrm{NBCU}})\right]\geq -\frac{(1-\eta)H}{2}.$$

This implies that

$$\begin{aligned}
\mathbb{E}\left[V(\pi^{\mathrm{E}})-V(\pi^{\mathrm{NBCU}})\right] &\geq (1-\eta)(V(\pi^{\mathrm{E}})-V(\pi^{\beta}))-\frac{(1-\eta)H}{2}\\
&=\frac{(1-\eta)}{2}(V(\pi^{\mathrm{E}})-V(\pi^{\beta})).
\end{aligned}$$

In both cases, we prove that $\mathbb{E}\left[V(\pi^{\mathrm{E}})-V(\pi^{\mathrm{NBCU}})\right]\gtrsim(1-\eta)(V(\pi^{\mathrm{E}})-V(\pi^{\beta}))$ and thus complete the proof.

D Proof of Results in Section 5

D.1 Proof of Proposition 2

In the tabular case, with the first-order optimality condition, we have $c_h^{\star}(s,a)=\widehat{d_h^{\mathrm{E}}}(s,a)/(\widehat{d_h^{\mathrm{E}}}(s,a)+\widehat{d_h^{\mathrm{U}}}(s,a))$. By Eq. (5), we have

$$\widehat{d_h^{\mathrm{U}}}(s,a)w_h(s,a)=\widehat{d_h^{\mathrm{U}}}(s,a)\times\frac{\widehat{d_h^{\mathrm{E}}}(s,a)}{\widehat{d_h^{\mathrm{U}}}(s,a)}=\widehat{d_h^{\mathrm{E}}}(s,a).$$

Hence, the learning objective (3) reduces to (1).

D.2 Proof of Lemma 1

Recall that

$$\Delta_h(\theta)=\min_{(s,a)\in\mathcal{D}_h^{\mathrm{E}}\cup\mathcal{D}_h^{\mathrm{S},1}}\langle\theta,\phi_h(s,a)\rangle-\max_{(s',a')\in\mathcal{D}_h^{\mathrm{S},2}}\langle\theta,\phi_h(s',a')\rangle.$$

Then we have that

$$\Delta_h(\bar{\theta}_h)-\Delta_h(\theta)=\min_{(s,a)\in\mathcal{D}_h^{\mathrm{E}}\cup\mathcal{D}_h^{\mathrm{S},1}}\langle\bar{\theta}_h,\phi_h(s,a)\rangle-\max_{(s',a')\in\mathcal{D}_h^{\mathrm{S},2}}\langle\bar{\theta}_h,\phi_h(s',a')\rangle$$

$$
\begin{aligned}
&- \min_{(s,a)\in\mathcal{D}_h^{\mathrm{E}}\cup\mathcal{D}_h^{\mathrm{S},1}} \langle \theta, \phi_h(s,a)\rangle + \max_{(s',a')\in\mathcal{D}_h^{\mathrm{S},2}} \langle \theta, \phi_h(s',a')\rangle \\
&\overset{(a)}{\leq} \langle \bar{\theta}_h, \phi_h(s^1,a^1)\rangle - \langle \bar{\theta}_h, \phi_h(s^2,a^2)\rangle - \langle \theta, \phi_h(s^1,a^1)\rangle + \langle \theta, \phi_h(s^2,a^2)\rangle \\
&= \langle \bar{\theta}_h - \theta, \phi_h(s^1,a^1) - \phi_h(s^2,a^2)\rangle \\
&\overset{(b)}{\leq} \left\|\bar{\theta}_h - \theta\right\| \left\|\phi_h(s^1,a^1) - \phi_h(s^2,a^2)\right\|.
\end{aligned}
$$

In inequality (a), we utilize the facts that $(s^1,a^1) \in \mathrm{argmin}_{(s,a)\in\mathcal{D}_h^{\mathrm{E}}\cup\mathcal{D}_h^{\mathrm{S},1}} \langle \theta_h, \phi_h(s,a)\rangle$ and $(s^2,a^2) \in \mathrm{argmax}_{(s,a)\in\mathcal{D}_h^{\mathrm{S},2}} \langle \theta_h, \phi_h(s,a)\rangle$. Inequality (b) follows the Cauchy–Schwarz inequality. Let $L_h = \left\|\phi_h(s^1,a^1) - \phi_h(s^2,a^2)\right\|$ and we finish the proof.

D.3 Proof of Lemma 2

First, by Taylor's Theorem, there exists $\theta_h' \in \{\theta \in \mathbb{R}^d : \theta^t = \theta_h^\star + t(\bar{\theta}_h - \theta_h^\star),\ \forall t \in [0,1]\}$ such that

$$
\begin{aligned}
\mathcal{L}_h(\bar{\theta}_h) &= \mathcal{L}_h(\theta_h^\star) + \langle \nabla \mathcal{L}_h(\theta_h^\star), \bar{\theta}_h - \theta_h^\star\rangle + \frac{1}{2}\left(\bar{\theta}_h - \theta_h^\star\right)^\top \nabla^2 \mathcal{L}_h(\theta_h')\left(\bar{\theta}_h - \theta_h^\star\right) \\
&= \mathcal{L}_h(\theta_h^\star) + \frac{1}{2}\left(\bar{\theta}_h - \theta_h^\star\right)^\top \nabla^2 \mathcal{L}_h(\theta_h')\left(\bar{\theta}_h - \theta_h^\star\right). \qquad (16)
\end{aligned}
$$

The last equality follows the optimality condition that $\nabla \mathcal{L}_h(\theta_h^\star) = 0$. Then, our strategy is to prove that the smallest eigenvalue of the Hessian matrix $\nabla^2 \mathcal{L}_h(\theta_h')$ is positive, i.e., $\lambda_{\min}(\nabla^2 \mathcal{L}_h(\theta_h')) > 0$. We first calculate the Hessian matrix $\nabla^2 \mathcal{L}_h(\theta_h')$. Given $\mathcal{D}^{\mathrm{E}}$ and $\mathcal{D}^{\mathrm{U}}$, we define the function $G : \mathbb{R}^{(|\mathcal{D}^{\mathrm{E}}|+|\mathcal{D}^{\mathrm{U}}|)} \to \mathbb{R}$ as

$$
G(v) \triangleq \frac{1}{|\mathcal{D}^{\mathrm{E}}|}\sum_{i=1}^{|\mathcal{D}^{\mathrm{E}}|} g(v_i) + \frac{1}{|\mathcal{D}^{\mathrm{U}}|}\sum_{j=1}^{|\mathcal{D}^{\mathrm{U}}|} g(v_j),
$$

where v_i is the i-th element in the vector $v \in \mathbb{R}^{(|\mathcal{D}^{\mathrm{E}}|+|\mathcal{D}^{\mathrm{U}}|)}$ and $g(x) = \log(1+\exp(x))$ is a real-valued function. Besides, we use $B_h \in \mathbb{R}^{(|\mathcal{D}^{\mathrm{E}}|+|\mathcal{D}^{\mathrm{U}}|)\times d}$ to denote the matrix whose i-th row $B_{h,i} = -y_i \phi_h(s^i,a^i)^\top$, and $y_i = 1$ if $(s^i,a^i) \in \mathcal{D}_h^{\mathrm{E}}$, $y_i = -1$ if $(s^i,a^i) \notin \mathcal{D}_h^{\mathrm{E}}$. Then the objective function can be reformulated as

$$
\begin{aligned}
&\mathcal{L}_h(\theta_h) \\
&= \sum_{(s,a)} \widehat{d_h^{\mathrm{E}}}(s,a)\left[\log\left(1+\exp\left(-\langle \phi_h(s,a), \theta_h\rangle\right)\right)\right] + \sum_{(s,a)} \widehat{d_h^{\mathrm{U}}}(s,a)\left[\log\left(1+\exp\left(\langle \phi_h(s,a), \theta_h\rangle\right)\right)\right] \\
&= \frac{1}{|\mathcal{D}^{\mathrm{E}}|}\sum_{(s,a)\in\mathcal{D}^{\mathrm{E}}} \log\left(1+\exp\left(-\langle \phi_h(s,a), \theta_h\rangle\right)\right) + \frac{1}{|\mathcal{D}^{\mathrm{U}}|}\sum_{(s,a)\in\mathcal{D}^{\mathrm{U}}} \log\left(1+\exp\left(\langle \phi_h(s,a), \theta_h\rangle\right)\right) \\
&= G(B_h\theta_h).
\end{aligned}
$$

Then we have that $\nabla^2 \mathcal{L}_h(\theta_h) = B_h^\top \nabla^2 G(B_h\theta_h) B_h$, where

$$
\begin{aligned}
&\nabla^2 G(B_h\theta_h) \\
&= \mathbf{diag}\left(\frac{g''((B_h\theta_h)_1)}{|\mathcal{D}^{\mathrm{E}}|}, \ldots, \frac{g''((B_h\theta_h)_{|\mathcal{D}^{\mathrm{E}}|})}{|\mathcal{D}^{\mathrm{E}}|}, \frac{g''((B_h\theta_h)_{|\mathcal{D}^{\mathrm{E}}|+1})}{|\mathcal{D}^{\mathrm{E}}|+|\mathcal{D}^{\mathrm{U}}|}, \ldots, \frac{g''((B_h\theta_h)_{|\mathcal{D}^{\mathrm{E}}|+|\mathcal{D}^{\mathrm{U}}|})}{|\mathcal{D}^{\mathrm{E}}|+|\mathcal{D}^{\mathrm{U}}|}\right).
\end{aligned}
$$

Here $g''(x) = \sigma(x)(1-\sigma(x))$, where $\sigma(x) = 1/(1+\exp(-x))$ is the sigmoid function. The eigenvalues of $\nabla^2 G(B_h\theta_h)$ are

$$
\left\{\frac{g''((B_h\theta_h)_1)}{|\mathcal{D}^{\mathrm{E}}|}, \ldots, \frac{g''((B_h\theta_h)_{|\mathcal{D}^{\mathrm{E}}|})}{|\mathcal{D}^{\mathrm{E}}|}, \frac{g''((B_h\theta_h)_{|\mathcal{D}^{\mathrm{E}}|+1})}{|\mathcal{D}^{\mathrm{E}}|+|\mathcal{D}^{\mathrm{U}}|}, \ldots, \frac{g''((B_h\theta_h)_{|\mathcal{D}^{\mathrm{E}}|+|\mathcal{D}^{\mathrm{U}}|})}{|\mathcal{D}^{\mathrm{E}}|+|\mathcal{D}^{\mathrm{U}}|}\right\}.
$$

Notice that $\theta_h' \in \{\theta \in \mathbb{R}^d : \theta^t = \theta_h^\star + t(\bar{\theta}_h - \theta_h^\star),\ \forall t \in [0,1]\}$. For a matrix A, we use $\lambda_{\min}(A)$ to denote the minimal eigenvalue of A. Here we claim that the minimum of the minimal eigenvalues of $\nabla^2 G(B_h\theta^t)$ over $t \in [0,1]$ is achieved at $t = 0$ or $t = 1$. That is,

$$
\min\{\lambda_{\min}(\nabla^2 G(B_h\theta^t)) : \forall t \in [0,1]\} = \min\{\lambda_{\min}(\nabla^2 G(B_h\theta^0)), \lambda_{\min}(\nabla^2 G(B_h\theta^1))\}.
$$

We prove this claim as follows. For any $t \in [0,1]$, we use $\{\lambda_1(t), \ldots, \lambda_{|\mathcal{D}^{\mathrm{E}}|+|\mathcal{D}^{\mathrm{U}}|}(t)\}$ to denote the eigenvalues of $\nabla^2 G(B_h \theta^t)$. For each $i \in [|\mathcal{D}^{\mathrm{E}}| + |\mathcal{D}^{\mathrm{U}}|]$, we consider $\lambda_i(t) : [0,1] \to \mathbb{R}$ as a function of t. Specifically,

$$\lambda_i(t) = \begin{cases} \frac{g''((B_h\theta_h^\star)_i + t(B_h(\bar{\theta}_h - \theta_h^\star))_i)}{|\mathcal{D}^{\mathrm{E}}|}, & \text{if } i \in [|\mathcal{D}^{\mathrm{E}}|] \\ \frac{g''((B_h\theta_h^\star)_i + t(B_h(\bar{\theta}_h - \theta_h^\star))_i)}{|\mathcal{D}^{\mathrm{E}}|+|\mathcal{D}^{\mathrm{U}}|}, & \text{otherwise.} \end{cases}$$

We observe that $g'''(x) = \sigma(x)(1-\sigma(x))(1-2\sigma(x))$ which satisfies that $\forall x \leq 0,\ g'''(x) \geq 0$, and $\forall x \geq 0,\ g'''(x) \leq 0$. Therefore, we have that the minimum of $\lambda_i(t)$ over $t \in [0,1]$ must be achieved at $t=0$ or $t=1$. That is,

$$\min_{t \in [0,1]} \lambda_i(t) = \min\{\lambda_i(0), \lambda_i(1)\}. \tag{17}$$

For any $t \in [0,1]$, we define $i^t \in [|\mathcal{D}^{\mathrm{E}}|+|\mathcal{D}^{\mathrm{U}}|]$ as the index of the minimal eigenvalue of $\nabla^2 G(B_h\theta^t)$, i.e., $\lambda_{i^t}(t) = \lambda_{\min}(\nabla^2 G(B_h\theta^t))$. Then we have that

$$\begin{aligned} \min\{\lambda_{\min}(\nabla^2 G(B_h\theta^t)) : \forall t \in [0,1]\} &= \min\{\lambda_{i^t}(t) : \forall t \in [0,1]\} \\ &\overset{(a)}{=} \min\{\min\{\lambda_{i^t}(0), \lambda_{i^t}(1)\} : \forall t \in [0,1]\} \\ &= \min\{\lambda_{i^0}(0), \lambda_{i^1}(1)\} \\ &\overset{(b)}{=} \min\{\lambda_{\min}(\nabla^2 G(B_h\theta^0)), \lambda_{\min}(\nabla^2 G(B_h\theta^1))\} \end{aligned}$$

Equality (a) follows (17) and equality (b) follows that $\lambda_{i^0}(0)$ and $\lambda_{i^1}(1)$ are the minimal eigenvalues of $\nabla^2 G(B_h\theta^0)$ and $\nabla^2 G(B_h\theta^1)$, respectively.

In summary, we derive that

$$\min\{\lambda_{\min}(\nabla^2 G(B_h\theta^t)) : \forall t \in [0,1]\} = \min\{\lambda_{\min}(\nabla^2 G(B_h\theta^0)), \lambda_{\min}(\nabla^2 G(B_h\theta^1))\}, \tag{18}$$

which proves the previous claim.

Further, we consider $\lambda_{\min}\left(\nabla^2 \mathcal{L}_h(\theta_h)\right)$.

$$\begin{aligned} \lambda_{\min}\left(\nabla^2 \mathcal{L}_h(\theta_h)\right) &= \inf_{x \in \mathbb{R}^d : \|x\|=1} x^\top \nabla^2 \mathcal{L}_h(\theta_h) x \\ &= \inf_{x \in \mathbb{R}^d : \|x\|=1} (B_h x)^\top \nabla^2 G(B_h\theta_h)(B_h x) \\ &= \inf_{z \in \mathrm{Im}(B_h)} z^\top \nabla^2 G(B_h\theta_h) z \\ &= \left(\inf_{z \in \mathrm{Im}(B_h)} \|z\|\right)^2 \lambda_{\min}(\nabla^2 G(B_h\theta_h)) \\ &\geq \left(\inf_{z \in \mathrm{Im}(B_h)} \|z\|\right)^2 \min\{\lambda_{\min}(\nabla^2 G(B_h\theta^0)), \lambda_{\min}(\nabla^2 G(B_h\theta^1))\}. \end{aligned}$$

Here $\mathrm{Im}(B_h) = \{z \in \mathbb{R}^d : z = B_h x, \|x\| = 1\}$. The last inequality follows Eq. (18).

Recall we assume that $\mathbf{rank}(A_h) = d$, so we have that $\mathbf{rank}(B_h) = d$. Thus, $\mathrm{Im}(B_h)$ is a set of vectors with positive norms, i.e., $\inf_{z \in \mathrm{Im}(B_h)} \|z\| > 0$. Besides, since $g''(x) = \sigma(x)(1-\sigma(x)) > 0$, we also have that

$$\min\{\lambda_{\min}(\nabla^2 G(B_h\theta^0)), \lambda_{\min}(\nabla^2 G(B_h\theta^1))\} > 0.$$

In summary, we obtain that

$$\lambda_{\min}\left(\nabla^2 \mathcal{L}_h(\theta_h)\right) \geq \left(\inf_{z \in \mathrm{Im}(B_h)} \|z\|\right)^2 \min\{\lambda_{\min}(\nabla^2 G(B_h\theta^0)), \lambda_{\min}(\nabla^2 G(B_h\theta^1))\} > 0.$$

Then, with Eq. (16), there exists

$$\tau_h = \left(\inf_{z \in \mathrm{Im}(B_h)} \|z\|\right)^2 \min\{\lambda_{\min}(\nabla^2 G(B_h\theta^0)), \lambda_{\min}(\nabla^2 G(B_h\theta^1))\} > 0$$

such that

$$\mathcal{L}_h(\bar{\theta}_h) \geq \mathcal{L}_h(\theta_h^\star) + \frac{\tau_h}{2} \left\|\bar{\theta}_h - \theta_h^\star\right\|^2,$$

which completes the proof.

D.4 Proof of Theorem 3

First, invoking Lemma 1 with $\theta = \theta_h^\star$ yields that

$$\Delta_h(\theta_h^\star) \geq \Delta_h(\bar{\theta}_h) - L_h \left\| \bar{\theta}_h - \theta_h^\star \right\|.$$

Here $L_h = \|\phi_h(s,a) - \phi_h(s',a')\|$ with $(s,a) \in \operatorname{argmin}_{(s,a)\in \mathcal{D}_h^{\mathrm{E}} \cup \mathcal{D}_h^{\mathrm{S},1}} \langle \theta_h^\star, \phi_h(s,a) \rangle$ and $(s',a') \in \operatorname{argmax}_{(s,a)\in \mathcal{D}_h^{\mathrm{S},2}} \langle \theta_h^\star, \phi_h(s,a) \rangle$. Then, by Lemma 2, there exists $\tau_h > 0$ such that

$$\mathcal{L}_h(\bar{\theta}_h) \geq \mathcal{L}_h(\theta_h^\star) + \frac{\tau_h}{2} \left\| \bar{\theta}_h - \theta_h^\star \right\|^2.$$

This directly implies an upper bound of the distance between $\bar{\theta}_h$ and $\theta_h^\star$.

$$\left\| \bar{\theta}_h - \theta_h^\star \right\| \leq \sqrt{\frac{2\left(\mathcal{L}_h(\bar{\theta}_h) - \mathcal{L}_h(\theta_h^\star)\right)}{\tau_h}}.$$

If the feature is designed such that $\sqrt{\frac{2\left(\mathcal{L}_h(\bar{\theta}_h) - \mathcal{L}_h(\theta_h^\star)\right)}{\tau_h}} < \frac{\Delta_h(\bar{\theta}_h)}{L_h}$ holds, we further have that $\left\| \bar{\theta}_h - \theta_h^\star \right\| < \Delta_h(\bar{\theta}_h)/L_h$. Then we get that

$$\Delta_h(\theta_h^\star) \geq \Delta_h(\bar{\theta}_h) - L_h \left\| \bar{\theta}_h - \theta_h^\star \right\| > 0,$$

which completes the proof of the first statement.

Then we proceed to prove the imitation gap bound. We first identify the property of $\pi^{\text{ISW-BC}}$. Recall the objective of WBCU.

$$\pi^{\text{ISW-BC}} \in \operatorname*{argmax}_{\pi} \sum_{h=1}^{H} \sum_{(s,a)\in \mathcal{S}\times\mathcal{A}} \left\{ \widehat{d_h^{\mathrm{U}}}(s,a) \times \left[w_h(s,a) \log \pi_h(a|s) \right] \times \mathbb{I}\left[w_h(s,a) \geq \delta \right] \right\}.$$

For any state s with $\sum_{a\in\mathcal{A}} \widehat{d_h^{\mathrm{U}}}(s,a) w_h(s,a) \mathbb{I}\left[w_h(s,a) \geq \delta \right] > 0$, with the first-order optimality condition, we have

$$\pi_h^{\text{ISW-BC}}(a|s) = \frac{\widehat{d_h^{\mathrm{U}}}(s,a) w_h(s,a) \mathbb{I}\left[w_h(s,a) \geq \delta \right]}{\sum_{a\in\mathcal{A}} \widehat{d_h^{\mathrm{U}}}(s,a) w_h(s,a) \mathbb{I}\left[w_h(s,a) \geq \delta \right]}.$$

For an expert state s with $d_h^{\pi^{\mathrm{E}}}(s) > 0$, if $(s, \pi_h^{\mathrm{E}}(s)) \in \mathcal{D}_h^{\mathrm{E}} \cup \mathcal{D}_h^{\mathrm{S},1}$, we have that

$$\langle \theta_h^\star, \phi_h(s, \pi_h^{\mathrm{E}}(s)) \rangle > \langle \theta_h^\star, \phi_h(s,a) \rangle, \quad \forall (s,a) \in \mathcal{D}_h^{\mathrm{S},2}.$$

This is due to the first statement that $\Delta_h(\theta_h^\star) > 0$ in this theorem. Recall that

$$c_h(s,a;\theta_h^\star) = \frac{1}{1 + \exp(-\langle \phi_h(s,a), \theta_h^\star \rangle)} \quad \text{and} \quad w_h(s,a) = \frac{c_h(s,a;\theta_h^\star)}{1 - c_h(s,a;\theta_h^\star)}.$$

We can further obtain that $w_h(s, \pi_h^{\mathrm{E}}(s)) > w_h(s,a)$ for any $(s,a) \in \mathcal{D}_h^{\mathrm{S},2}$. This implies that we can find a δ such that $\mathbb{I}\left[w_h(s, \pi_h^{\mathrm{E}}(s)) \geq \delta \right] = 1$ for any $(s, \pi_h^{\mathrm{E}}(s)) \in \mathcal{D}_h^{\mathrm{E}} \cup \mathcal{D}_h^{\mathrm{S},1}$ and $\mathbb{I}\left[w_h(s,a) \geq \delta \right] = 0$ for any $(s,a) \in \mathcal{D}_h^{\mathrm{S},2}$. Based on the above analytical form of $\pi^{\text{ISW-BC}}$, we have that $\pi^{\text{ISW-BC}}(\pi_h^{\mathrm{E}}(s)|s) = 1$ for any $(s, \pi_h^{\mathrm{E}}(s)) \in \mathcal{D}_h^{\mathrm{E}} \cup \mathcal{D}_h^{\mathrm{S},1}$. In summary, for any state s with $(s, \pi_h^{\mathrm{E}}(s)) \in \mathcal{D}_h^{\mathrm{E}} \cup \mathcal{D}_h^{\mathrm{S},1}$, we have that $\pi_h^{\text{ISW-BC}}(\pi_h^{\mathrm{E}}(s)|s) = 1$.

With the above property of $\pi^{\text{ISW-BC}}$, we proceed to analyze the policy value gap. According to [32, Lemma 4.3], we have

$$V(\pi^{\mathrm{E}}) - V(\pi^{\text{ISW-BC}}) \leq H \sum_{h=1}^{H} \mathbb{E}_{s\sim d_h^{\pi^{\mathrm{E}}}(\cdot)} \left[\mathrm{TV}\left(\pi_h^{\mathrm{E}}(\cdot|s), \pi_h^{\text{ISW-BC}}(\cdot|s) \right) \right].$$

Since π^{E} is assumed to be deterministic, we have

$$V(\pi^{\mathrm{E}}) - V(\pi^{\text{ISW-BC}}) \leq H \sum_{h=1}^{H} \mathbb{E}_{s\sim d_h^{\pi^{\mathrm{E}}}(\cdot)} \left[\mathbb{E}_{a\sim \pi_h^{\text{ISW-BC}}(\cdot|s)} \left[\mathbb{I}\left\{ a \neq \pi_h^{\mathrm{E}}(s) \right\} \right] \right]$$

$$
\begin{aligned}
&\overset{(a)}{\leq} H \sum_{h=1}^{H} \mathbb{E}_{s \sim d_h^{\pi^{\mathrm{E}}}(\cdot)} \left[\mathbb{I} \left\{ (s, \pi_h^{\mathrm{E}}(s)) \notin \mathcal{D}_h^{\mathrm{E}} \cup \mathcal{D}_h^{\mathrm{S},1} \right\} \right] \\
&\overset{(b)}{=} H \sum_{h=1}^{H} \mathbb{E}_{s \sim d_h^{\pi^{\mathrm{E}}}(\cdot)} \left[\mathbb{I} \left\{ (s, \pi_h^{\mathrm{E}}(s)) \notin \mathcal{D}_h^{\mathrm{U}} \right\} \right].
\end{aligned}
$$

Inequality (a) follows the property of $\pi^{\text{ISW-BC}}$ derived above. In particular, for any state s with $(s, \pi_h^{\mathrm{E}}(s)) \in \mathcal{D}_h^{\mathrm{E}} \cup \mathcal{D}_h^{\mathrm{S},1}$, we have that $\pi_h^{\text{ISW-BC}}(\pi_h^{\mathrm{E}}(s)|s) = 1$. Equation (b) holds due to the Assumption 2. In particular, for an expert state s that $d_h^{\pi^{\mathrm{E}}}(s) > 0$, the events of $(s, \pi_h^{\mathrm{E}}(s)) \notin \mathcal{D}_h^{\mathrm{E}} \cup \mathcal{D}_h^{\mathrm{S},1}$ and $(s, \pi_h^{\mathrm{E}}(s)) \notin \mathcal{D}_h^{\mathrm{U}}$ are equivalent.

Moreover, we take the expectation over $\mathcal{D}^{\mathrm{U}}$ on both sides and obtain that

$$
\begin{aligned}
\mathbb{E}\left[V(\pi^{\mathrm{E}}) - V(\pi^{\text{ISW-BC}}) \right] &\leq H \sum_{h=1}^{H} \mathbb{E}_{s \sim d_h^{\pi^{\mathrm{E}}}(\cdot)} \left[\mathbb{P}\left((s, \pi_h^{\mathrm{E}}(s)) \notin \mathcal{D}_h^{\mathrm{U}} \right) \right] \\
&= H \sum_{h=1}^{H} \sum_{s \in \mathcal{S}} d_h^{\pi^{\mathrm{E}}}(s) \mathbb{P}\left((s, \pi_h^{\mathrm{E}}(s)) \notin \mathcal{D}_h^{\mathrm{U}} \right).
\end{aligned}
$$

According to Assumption 1, we have that

$$
\begin{aligned}
d_h^{\mathrm{U}}(s, \pi_h^{\mathrm{E}}(s)) &= \eta d_h^{\pi^{\mathrm{E}}}(s, \pi_h^{\mathrm{E}}(s)) + (1-\eta) d_h^{\pi^{\beta}}(s, \pi_h^{\mathrm{E}}(s)) \\
&\overset{(a)}{\geq} \eta d_h^{\pi^{\mathrm{E}}}(s, \pi_h^{\mathrm{E}}(s)) + \frac{(1-\eta)}{\mu} d_h^{\pi^{\mathrm{E}}}(s, \pi_h^{\mathrm{E}}(s)) \\
&= \left(\eta + \frac{(1-\eta)}{\mu} \right) d_h^{\pi^{\mathrm{E}}}(s, \pi_h^{\mathrm{E}}(s)).
\end{aligned}
$$

Inequality (a) follows the definition of μ in Theorem 3: for any $(s, h) \in \mathcal{S} \times [H]$, we have $d_h^{\pi^{\mathrm{E}}}(s, \pi_h^{\mathrm{E}}(s)) / d_h^{\pi^{\beta}}(s, \pi_h^{\mathrm{E}}(s)) \leq \mu$. Then we obtain that

$$
\begin{aligned}
\mathbb{E}\left[V(\pi^{\mathrm{E}}) - V(\pi^{\text{ISW-BC}}) \right] &\leq H \sum_{h=1}^{H} \sum_{s \in \mathcal{S}} d_h^{\pi^{\mathrm{E}}}(s) (1 - d_h^{\mathrm{U}}(s, \pi_h^{\mathrm{E}}(s)))^{N_{\mathrm{tot}}} \\
&\leq \left(\frac{1}{\eta + (1-\eta)/\mu} \right) H \sum_{h=1}^{H} \sum_{s \in \mathcal{S}} d_h^{\mathrm{U}}(s, \pi_h^{\mathrm{E}}(s)) \mathbb{P}\left((s, \pi_h^{\mathrm{E}}(s)) \notin \mathcal{D}_h^{\mathrm{U}} \right).
\end{aligned}
$$

For each $(s, h) \in \mathcal{S} \times [H]$, we observe that

$$
d_h^{\mathrm{U}}(s, \pi_h^{\mathrm{E}}(s)) \mathbb{P}\left((s, \pi_h^{\mathrm{E}}(s)) \notin \mathcal{D}_h^{\mathrm{U}} \right) = d_h^{\mathrm{U}}(s, \pi_h^{\mathrm{E}}(s)) \left(1 - d_h^{\mathrm{U}}(s, \pi_h^{\mathrm{E}}(s)) \right)^{N_{\mathrm{tot}}} \leq \frac{4}{9 N_{\mathrm{tot}}}.
$$

Here the last inequality follows Lemma 5. Consequently, we can derive that

$$
\sum_{h=1}^{H} \sum_{s \in \mathcal{S}} d_h^{\mathrm{U}}(s, \pi_h^{\mathrm{E}}(s)) \mathbb{P}\left((s, \pi_h^{\mathrm{E}}(s)) \notin \mathcal{D}_h^{\mathrm{U}} \right) \leq \frac{4H|\mathcal{S}|}{9 N_{\mathrm{tot}}},
$$

which further implies that

$$
\mathbb{E}\left[V(\pi^{\mathrm{E}}) - V(\pi^{\text{ISW-BC}}) \right] \leq \left(\frac{1}{\eta + (1-\eta)/\mu} \right) \frac{4H^2|\mathcal{S}|}{9 N_{\mathrm{tot}}} = \frac{4H^2|\mathcal{S}|}{9\left(N_{\mathrm{E}} + N_{\mathrm{S}}/\mu \right)}.
$$

We complete the proof.

D.5 An Example Corresponding to Theorem 3

In this section, we provide an example that illustrates the required feature design in Theorem 3 can hold.

Example 1. *To illustrate Theorem 3, we consider an example in the feature space $\mathbb{R}^2$. In particular, for time step $h \in [H]$, we have the expert dataset and supplementary dataset as follows.*

$$
\mathcal{D}_h^{\mathrm{E}} = \left\{ \left(s^{(1)}, a^{(1)} \right), \left(s^{(4)}, a^{(4)} \right) \right\}, \ \mathcal{D}_h^{\mathrm{S}} = \left\{ \left(s^{(2)}, a^{(2)} \right), \left(s^{(3)}, a^{(3)} \right) \right\},
$$

$$\mathcal{D}_h^{\mathrm{S},1} = \left\{ \left(s^{(2)}, a^{(2)} \right) \right\}, \; \mathcal{D}_h^{\mathrm{S},2} = \left\{ \left(s^{(3)}, a^{(3)} \right) \right\}.$$

The corresponding features are

$$\phi_h\left(s^{(1)}, a^{(1)}\right) = (0,1)^\top, \; \phi_h\left(s^{(2)}, a^{(2)}\right) = \left(-\frac{1}{2}, 0\right)^\top,$$
$$\phi_h\left(s^{(3)}, a^{(3)}\right) = \left(0, -\frac{1}{2}\right)^\top, \; \phi_h\left(s^{(4)}, a^{(4)}\right) = (-1,0)^\top.$$

Notice that the set of expert-style samples is $\mathcal{D}_h^{\mathrm{E}} \cup \mathcal{D}_h^{\mathrm{S},1} = \{(s^{(1)}, a^{(1)}), (s^{(2)}, a^{(2)}), (s^{(4)}, a^{(4)})\}$ *and the set of non-expert-style samples is* $\mathcal{D}_h^{\mathrm{S},2} = \{(s^{(3)}, a^{(3)})\}$*. It is direct to calculate that the ground-truth parameter that achieves the maximum margin among unit vectors is* $\bar{\theta}_h = (-\sqrt{2}/2, \sqrt{2}/2)^\top$ *and the maximum margin is* $\Delta_h(\bar{\theta}_h) = \sqrt{2}/2$*. According to Eq. (6), for* $\theta_h = (\theta_{h,1}, \theta_{h,2})^\top$*, the optimization objective is*

$$\begin{aligned}
&\mathcal{L}_h(\theta_h) \\
&= \sum_{(s,a)} \widehat{d_h^{\mathrm{E}}}(s,a) \left[\log\left(1+\exp\left(-\langle \phi_h(s,a), \theta_h\rangle\right)\right)\right] + \sum_{(s,a)} \widehat{d_h^{\mathrm{U}}}(s,a) \left[\log\left(1+\exp\left(\langle \phi_h(s,a), \theta_h\rangle\right)\right)\right] \\
&= \frac{1}{2}\left(\log\left(1+\exp\left(-\theta_{h,2}\right)\right) + \log\left(1+\exp\left(\theta_{h,1}\right)\right)\right) \\
&\quad + \frac{1}{4}\left(\log\left(1+\exp\left(\theta_{h,2}\right)\right) + \log\left(1+\exp\left(-\frac{1}{2}\theta_{h,1}\right)\right)\right) \\
&\quad + \frac{1}{4}\left(\log\left(1+\exp\left(-\frac{1}{2}\theta_{h,2}\right)\right) + \log\left(1+\exp\left(-\theta_{h,1}\right)\right)\right).
\end{aligned}$$

We apply CVXPY [10] to calculate the optimal solution $\theta_h^\star \approx (-0.310, 0.993)^\top$ *and the objective values* $\mathcal{L}_h(\theta_h^\star) \approx 1.287$, $\mathcal{L}_h(\bar{\theta}_h) \approx 1.309$*. Furthermore, we calculate the Lipschitz coefficient* L_h *appears in Lemma 1.*

$$(s^{(2)}, a^{(2)}) = \operatorname*{argmin}_{(s,a)\in\mathcal{D}_h^{\mathrm{E}}\cup\mathcal{D}_h^{\mathrm{S},1}} \langle \theta_h^\star, \phi_h(s,a)\rangle, \; (s^{(3)}, a^{(3)}) \in \operatorname*{argmax}_{(s,a)\in\mathcal{D}_h^{\mathrm{S},2}} \langle \theta_h^\star, \phi_h(s,a)\rangle,$$
$$L_h = \left\| \phi_h(s^{(2)}, a^{(2)}) - \phi_h(s^{(3)}, a^{(3)}) \right\| = \frac{\sqrt{2}}{2}.$$

Then we calculate the parameter of strong convexity τ_h *appears in Lemma 2. Based on the proof of Lemma 2, our strategy is to calculate the minimal eigenvalue of the Hessian matrix.*

First, for $\theta_h = (\theta_{h,1}, \theta_{h,2})^\top$*, the gradient of* $\mathcal{L}_h(\theta_h)$ *is*

$$\begin{aligned}
&\nabla \mathcal{L}_h(\theta_h) \\
&= -\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \widehat{d_h^{\mathrm{E}}}(s,a)\sigma(-\langle \phi_h(s,a), \theta_h\rangle) + \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \widehat{d_h^{\mathrm{U}}}(s,a)\sigma\left(\langle \phi_h(s,a), \theta_h\rangle\right) \\
&= \left(\frac{1}{2}\sigma(\theta_{h,1}) - \frac{1}{4}\sigma(-\theta_{h,1}) - \frac{1}{8}\sigma(-\frac{1}{2}\theta_{h,1}), \frac{1}{4}\sigma\left(\theta_{h,2}\right) - \frac{1}{2}\sigma\left(-\theta_{h,2}\right) - \frac{1}{8}\sigma(-\frac{1}{2}\theta_{h,2})\right)^\top.
\end{aligned}$$

Here $\sigma(x) = 1/(1+\exp(-x))$ *for* $x \in \mathbb{R}$ *is the sigmoid function. Then the Hessian matrix at* θ_h *is*

$$\nabla^2 \mathcal{L}_h(\theta_h) = \begin{pmatrix} \frac{3}{4}f(\theta_{h,1}) + \frac{1}{16}f\left(\frac{1}{2}\theta_{h,1}\right) & 0 \\ 0 & \frac{3}{4}f(\theta_{h,2}) + \frac{1}{16}f\left(\frac{1}{2}\theta_{h,2}\right) \end{pmatrix},$$

where $f(x) = \sigma(x)(1-\sigma(x))$ *and* $f(x) = f(-x)$*. For any* $t \in [0,1]$*, the eigenvalues of the Hessian matrix at* $\theta_h^t = \bar{\theta}_h + t(\theta_h^\star - \bar{\theta}_h)$ *are*

$$\frac{3}{4}f(\theta_{h,1}^t) + \frac{1}{16}f\left(\frac{1}{2}\theta_{h,1}^t\right), \; \frac{3}{4}f(\theta_{h,2}^t) + \frac{1}{16}f\left(\frac{1}{2}\theta_{h,2}^t\right).$$

Now, we calculate the minimal eigenvalues of $\nabla^2\mathcal{L}_h(\theta_h^t)$*. We consider the function*

$$g(x) = \frac{3}{4}f(x) + \frac{1}{16}f\left(\frac{1}{2}x\right), \; \forall x \in [a,b].$$

The gradient is

$$g'(x) = \frac{3}{4}\sigma(x)(1-\sigma(x))(1-2\sigma(x)) + \frac{1}{32}\sigma\left(\frac{1}{2}x\right)\left(1-\sigma\left(\frac{1}{2}x\right)\right)\left(1-2\sigma\left(\frac{1}{2}x\right)\right).$$

We observe that $\forall x \leq 0,\ g'(x) \geq 0$, *and* $\forall x \geq 0,\ g'(x) \leq 0$. *Thus, we have that the minimum of* $g(x)$ *must be achieved at* $x = a$ *or* $x = b$. *Besides, we have that* $g(x) = g(-x)$. *With the above arguments, we know that the minimal eigenvalue is* $g(0.993) \approx 0.163$ *and* $\tau_h \approx 0.163$. *Then we can calculate that*

$$\sqrt{\frac{2\left(\mathcal{L}_h(\bar{\theta}_h) - \mathcal{L}_h(\theta_h^\star)\right)}{\tau_h}} \approx 0.520,\ \frac{\Delta_h(\bar{\theta}_h)}{L_h} = 1.$$

The inequality in Theorem 3 holds.

E Discussion

In the main text, we focus on the tabular representations for policies. Furthermore, we consider a trajectory sampling procedure for behavior policy in collecting the supplementary dataset. We present two possible extensions in this section.

E.1 Function Approximation of Policies

Assume that the learner is access to a finite function class $\Pi = \{\pi = (\pi_1, \pi_2, \ldots, \pi_h)\}$, where $\pi_h : \mathcal{S} \to \Delta(\mathcal{A})$ could be any function (e.g., neural networks). For simplicity of analysis, we assume that Π is a finite class. Notice that the algorithms considered in this paper are BC and its variants, which all take the principle of maximum likelihood estimation (MLE). The theoretical analysis of these algorithms is based on the following inequality:

$$V(\pi^{\mathrm{E}}) - V(\pi) \leq H \sum_{h=1}^{H} \mathbb{E}_{s \sim d_h^{\pi^{\mathrm{E}}}(\cdot)}\left[\mathrm{TV}\left(\pi_h^{\mathrm{E}}(\cdot|s), \pi_h(\cdot|s)\right)\right].$$

Therefore, the key is to upper bound the TV distance. Take BC as an example (i.e., $\pi = \pi^{\mathrm{BC}}$). By using the concentration inequality in [1, Theorem 21], we obtain that for any $\delta \in (0,1)$, when $|\mathcal{D}^{\mathrm{E}}| \geq 1$, with probability at least $1-\delta$ over the randomness within $\mathcal{D}^{\mathrm{E}}$,

$$\mathbb{E}_{s \sim d_h^{\pi^{\mathrm{E}}}(\cdot)}\left[\mathrm{TV}^2\left(\pi_h^{\mathrm{E}}(\cdot|s), \pi_h^{\mathrm{BC}}(\cdot|s)\right)\right] \leq 2\frac{\log(|\Pi|/\delta)}{|\mathcal{D}^{\mathrm{E}}|}. \tag{19}$$

With additional efforts (by using union bound and Jensen's inequality), we have the following result.

Theorem 4 (BC with Function Approximation)**.** *Under Assumption 1. In the general function approximation setting, additionally assume that* $\pi^{\mathrm{E}} \in \Pi$. *If we apply BC on the expert data, we have*

$$\mathbb{E}\left[V(\pi^{\mathrm{E}}) - V(\pi^{\mathrm{BC}})\right] = \mathcal{O}\left(H^2\sqrt{\frac{\log(|\Pi| H N_{\mathrm{E}})}{N_{\mathrm{E}}}}\right),$$

where the expectation is taken over the randomness in the dataset collection.

The detailed proof is deferred to Appendix F. Compared with Theorem 1, we notice that the change in theoretical bound is that $\mathcal{O}(|\mathcal{S}|/N_{\mathrm{E}})$ is replaced by $\mathcal{O}(\sqrt{\log(|\Pi| H N_{\mathrm{E}})/N_{\mathrm{E}}})$.

NBCU can be analyzed in a similar way in the function approximation setting.

Theorem 5 (NBCU with Function Approximation)**.** *Under Assumption 1. In the general function approximation setting, additionally assume that the realizable policy class* Π *is realizable, i.e.,* $\pi^{\mathrm{mix}} \in \Pi$, *where* π^{mix} *is defined in Eq.* (10). *If we apply BC on the union dataset, we have*

$$\mathbb{E}\left[V(\pi^{\mathrm{E}}) - V(\pi^{\mathrm{NBCU}})\right] = \mathcal{O}\left((1-\eta)(V(\pi^{\mathrm{E}}) - V(\pi^{\beta})) + H^2\sqrt{\frac{\log(|\Pi| H N_{\mathrm{tot}})}{N_{\mathrm{tot}}}}\right).$$

The proof of Theorem 5 is deferred to Appendix F. We use Theorem 4 to help prove Theorem 5.

Unfornatunetly, the analysis of ISW-BC with function approximation is much more complicated since the maximum likelihood estimation is performed in a weighted manner.In the following part, we make a conjecture on the theoretical guarantee of the weighted maximum likelihood estimation. With such a conjecture, we can derive the imitation gap of ISW-BC with general function approximation. We leave the proof of the conjecture and other proof possibilities for future works.

Recall the objective of ISW-BC.

$$\pi^{\text{ISW-BC}} \in \operatorname*{argmax}_{\pi \in \Pi} \sum_{h=1}^{H} \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} \left\{ \widehat{d_h^{\text{U}}}(s,a) \times \left[w_h(s,a) \log \pi_h(a|s) \right] \times \mathbb{I}\left[w_h(s,a) \geq \delta \right] \right\},$$

Notice that the analysis of the discriminators is independent of the function approximation of policies. Therefore, we can follow the analysis of the discriminators in the proof of Theorem 3. Importantly, we can derive that there exists δ such that $\mathbb{I}\left[w_h(s, \pi_h^{\text{E}}(s)) \geq \delta \right] = 1$ for any $(s, \pi_h^{\text{E}}(s)) \in \mathcal{D}_h^{\text{E}} \cup \mathcal{D}_h^{\text{S},1}$ and $\mathbb{I}\left[w_h(s,a) \geq \delta \right] = 0$ for any $\forall (s,a) \in \mathcal{D}_h^{\text{S},2}$. Then we can obtain that

$$\begin{aligned}
&\sum_{h=1}^{H} \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} \left\{ \widehat{d_h^{\text{U}}}(s,a) \times \left[w_h(s,a) \log \pi_h(a|s) \right] \times \mathbb{I}\left[w_h(s,a) \geq \delta \right] \right\} \\
&= \sum_{h=1}^{H} \sum_{s \in \mathcal{S}_h^{\text{E}}, a = \pi_h^{\text{E}}(s)} \widehat{d_h^{\text{U}}}(s,a) \times \left[w_h(s,a) \log \pi_h(a|s) \right].
\end{aligned}$$

Here $\mathcal{S}_h^{\text{E}} = \{ s \in \mathcal{S} : d_h^{\pi^{\text{E}}}(s) > 0 \}$. Then we have that

$$\pi^{\text{ISW-BC}} \in \operatorname*{argmax}_{\pi \in \Pi} \sum_{h=1}^{H} \sum_{s \in \mathcal{S}_h^{\text{E}}, a = \pi_h^{\text{E}}(s)} \widehat{d_h^{\text{U}}}(s,a) \times \left[w_h(s,a) \log \pi_h(a|s) \right].$$

We conjecture that $\pi^{\text{ISW-BC}}$ learned by the above weighted maximum likelihood holds the following theoretical guarantee. For any $\delta \in (0,1)$, with probability at least $1-\delta$, we have that

$$\sum_{s \in \mathcal{S}_h^{\text{E}}} d_h^{\text{U}}(s, \pi_h^{\text{E}}(s)) \text{TV}^2 \left(\pi_h^{\text{E}}(\cdot|s), \pi_h^{\text{ISW-BC}}(\cdot|s) \right) = \mathcal{O}\left(\frac{\log\left(|\Pi| N_{\text{tot}}\right)}{N_{\text{tot}}} \right). \tag{20}$$

This conjecture corresponds to (19) in the unweighted maximum likelihood estimation. With this conjecture, we can derive the imitation gap of ISW-BC with function approximation.

Conjecture 1 (Imitation Gap of ISW-BC with Function Approximation)**.** *Under Assumptions 1 and 2, let* $\mu = \max_{(s,h) \in \mathcal{S} \times [H]} d_h^{\pi^{\text{E}}}(s, \pi_h^{\text{E}}(s)) / d_h^{\pi^{\beta}}(s, \pi_h^{\text{E}}(s))$*. In the general function approximation setting with the realizable policy class* Π*, i.e.,* $\pi^{\text{E}} \in \Pi$*. Furthermore, assume that the feature is designed such that* $\sqrt{\frac{2\left(\mathcal{L}_h(\bar{\theta}_h) - \mathcal{L}_h(\theta_h^{\star})\right)}{\tau_h}} < \frac{\Delta_h(\bar{\theta}_h)}{L_h}$ *holds and the conjecture in (20) holds. Then, we have the imitation gap bound*

$$\mathbb{E}[V(\pi^{\text{E}}) - V(\pi^{\text{ISW-BC}})] = \mathcal{O}\left(H^2 \sqrt{\frac{\log(|\Pi| H N_{\text{tot}})}{N_{\text{E}} + N_{\text{S}}/\mu}} \right).$$

E.2 Supplementary Data with Corruption

In the main text, we consider the trajectory sampling procedure in Assumption 1. However, in some cases, the supplementary data can be poisoned and corrupted by an adversary. For example, although the human expert demonstrates an optimal trajectory, the recorder or the recording system possibly corrupts the data by accident or on purpose. Data corruption is one of the main security threats to imitation learning methods [25]. Therefore, it is valuable to investigate the robustness of the presented algorithms in this poison setting. Supplementary data with corruption is partially investigated in our experiments under the noisy expert setting, which we argue have a large state-action distribution shift.

Assumption 3 (Poison Setting)**.** *The supplementary dataset* $\mathcal{D}^{\text{S}}$ *and expert dataset* $\mathcal{D}^{\text{E}}$ *are collected in the following way: each time, with probability* η*, we rollout the expert policy to collect a trajectory. With probability* $1-\eta$*, we still rollout the expert policy to collect a trajectory but with probability*

$1-\eta'$, *the actions along the sampled trajectory are replaced with actions uniformly sampled from the action space. Such an experiment is independent and identically conducted by* N_{tot} *times.*

Theorem 6 (NBCU in the Poison Setting)**.** *Under Assumption 3. In the tabular case, for any* $\eta \in (0, 1]$, *we have*

$$\mathbb{E}\left[V(\pi^{\mathrm{E}}) - V(\pi^{\mathrm{NBCU}})\right] = \mathcal{O}\left((1-\eta)(1-\eta')H^2\left(1-\frac{1}{|\mathcal{A}|}\right) + H^2\sqrt{\frac{|\mathcal{S}||\mathcal{A}|}{N_{\mathrm{tot}}}}\right),$$

where the expectation is taken over the randomness in the dataset collection.

Theorem 7 (ISW-BC in the Poison Setting)**.** *Under Assumptions 2 and 3, if the feature is designed such that* $\sqrt{\frac{2\left(\mathcal{L}_h(\bar{\theta}_h)-\mathcal{L}_h(\theta_h^\star)\right)}{\tau_h}} < \frac{\Delta_h(\bar{\theta}_h)}{L_h}$ *holds, we have the imitation gap bound*

$$\mathbb{E}[V(\pi^{\mathrm{E}}) - V(\pi^{\text{ISW-BC}})] = \mathcal{O}\left(\frac{H^2|\mathcal{S}|}{N_{\mathrm{E}} + N_{\mathrm{S}}\eta'}\right).$$

Proofs of Theorem 6 and Theorem 7 can be found in Appendix F. Compared with the imitation gap of NBCU, there is no non-vanishing gap due to the corrupted actions in the imitation gap of ISW-BC. This means that ISW-BC is still robust in this setting.

F Proof of Results in Section E

F.1 Proof of Theorem 4

According to [32, Lemma 4.3], we have

$$V(\pi^{\mathrm{E}}) - V(\pi^{\mathrm{BC}}) \leq H\sum_{h=1}^{H}\mathbb{E}_{s\sim d_h^{\pi^{\mathrm{E}}}(\cdot)}\left[\mathrm{TV}\left(\pi_h^{\mathrm{E}}(\cdot|s), \pi_h^{\mathrm{BC}}(\cdot|s)\right)\right].$$

With [1, Theorem 21], when $|\mathcal{D}^{\mathrm{E}}| \geq 1$, for any $\delta \in (0,1)$, with probability at least $1-\delta$ over the randomness within $\mathcal{D}^{\mathrm{E}}$, we have that

$$\mathbb{E}_{s\sim d_h^{\pi^{\mathrm{E}}}(\cdot)}\left[\mathrm{TV}^2\left(\pi_h^{\mathrm{E}}(\cdot|s), \pi_h^{\mathrm{BC}}(\cdot|s)\right)\right] \leq 2\frac{\log(|\Pi|/\delta)}{|\mathcal{D}^{\mathrm{E}}|}.$$

With union bound, with probability at least $1-\delta$, for all $h \in [H]$, it holds that

$$\mathbb{E}_{s\sim d_h^{\pi^{\mathrm{E}}}(\cdot)}\left[\mathrm{TV}^2\left(\pi_h^{\mathrm{E}}(\cdot|s), \pi_h^{\mathrm{BC}}(\cdot|s)\right)\right] \leq 2\frac{\log(|\Pi|H/\delta)}{|\mathcal{D}^{\mathrm{E}}|},$$

which implies that

$$\begin{aligned}
V(\pi^{\mathrm{E}}) - V(\pi^{\mathrm{BC}}) &\leq H\sum_{h=1}^{H}\mathbb{E}_{s\sim d_h^{\pi^{\mathrm{E}}}(\cdot)}\left[\mathrm{TV}\left(\pi_h^{\mathrm{E}}(\cdot|s), \pi_h^{\mathrm{BC}}(\cdot|s)\right)\right] \\
&\overset{(a)}{\leq} H\sum_{h=1}^{H}\sqrt{\mathbb{E}_{s\sim d_h^{\pi^{\mathrm{E}}}(\cdot)}\left[\mathrm{TV}^2\left(\pi_h^{\mathrm{E}}(\cdot|s), \pi_h^{\mathrm{BC}}(\cdot|s)\right)\right]} \\
&\leq \sqrt{2}H^2\sqrt{\frac{\log(|\Pi|H/\delta)}{|\mathcal{D}^{\mathrm{E}}|}}.
\end{aligned}$$

Inequality (a) follows Jensen's inequality. Taking expectation over the randomness within $\mathcal{D}^{\mathrm{E}}$ yields that

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}^{\mathrm{E}}}\left[V(\pi^{\mathrm{E}}) - V(\pi^{\mathrm{BC}})\right] &\leq \delta H + (1-\delta)\sqrt{2}H^2\sqrt{\frac{\log(|\Pi|H/\delta)}{|\mathcal{D}^{\mathrm{E}}|}} \\
&\overset{(a)}{=} \frac{H}{2|\mathcal{D}^{\mathrm{E}}|} + \left(1-\frac{1}{2|\mathcal{D}^{\mathrm{E}}|}\right)\sqrt{2}H^2\sqrt{\frac{\log(2|\Pi|H|\mathcal{D}^{\mathrm{E}}|)}{|\mathcal{D}^{\mathrm{E}}|}}
\end{aligned}$$

$$
\begin{aligned}
&\leq \left(\sqrt{2}+1\right) H^2 \sqrt{\frac{\log(2|\Pi|H|\mathcal{D}^{\mathrm{E}}|)}{|\mathcal{D}^{\mathrm{E}}|}} \\
&\leq 4H^2 \sqrt{\frac{\log(4|\Pi|H|\mathcal{D}^{\mathrm{E}}|)}{|\mathcal{D}^{\mathrm{E}}|}}.
\end{aligned}
$$

Equation (a) holds due to the choice that $\delta = 1/(2|\mathcal{D}^{\mathrm{E}}|)$. For $|\mathcal{D}^{\mathrm{E}}| = 0$, we directly have that

$$
\mathbb{E}_{\mathcal{D}^{\mathrm{E}}} \left[V(\pi^{\mathrm{E}}) - V(\pi^{\mathrm{BC}}) \right] \leq H.
$$

Therefore, for any $|\mathcal{D}^{\mathrm{E}}| \geq 0$, we have that

$$
\mathbb{E}_{\mathcal{D}^{\mathrm{E}}} \left[V(\pi^{\mathrm{E}}) - V(\pi^{\mathrm{BC}}) \right] \leq 4H^2 \sqrt{\frac{\log(4|\Pi|H \max\{|\mathcal{D}^{\mathrm{E}}|, 1\})}{\max\{|\mathcal{D}^{\mathrm{E}}|, 1\}}}.
$$

We consider a real-valued function $f(x) = \log(cx)/x$ for $x \geq 1$, where $c = 4|\Pi|H > 4$. Its gradient function is $f'(x) = (1 - \log(cx))/x^2 \leq 0$ when $x \geq 1$. Then we know that $f(x)$ is decreasing as x increases. Furthermore, we have that $\max\{|\mathcal{D}^{\mathrm{E}}|, 1\} \geq (|\mathcal{D}^{\mathrm{E}}| + 1)/2$ when $|\mathcal{D}^{\mathrm{E}}| \geq 0$. Then we obtain

$$
\begin{aligned}
\mathbb{E}_{\mathcal{D}^{\mathrm{E}}} \left[V(\pi^{\mathrm{E}}) - V(\pi^{\mathrm{BC}}) \right] &\leq 4H^2 \sqrt{\frac{\log(4|\Pi|H \max\{|\mathcal{D}^{\mathrm{E}}|, 1\})}{\max\{|\mathcal{D}^{\mathrm{E}}|, 1\}}} \\
&\leq 4H^2 \sqrt{\frac{2\log(4|\Pi|H(|\mathcal{D}^{\mathrm{E}}| + 1))}{|\mathcal{D}^{\mathrm{E}}| + 1}}.
\end{aligned}
$$

Taking expectation over the random variable $|\mathcal{D}^{\mathrm{E}}| \sim \mathrm{Bin}(N_{\mathrm{tot}}, \eta)$ yields that

$$
\begin{aligned}
\mathbb{E} \left[V(\pi^{\mathrm{E}}) - V(\pi^{\mathrm{BC}}) \right] &\leq 4H^2 \mathbb{E} \left[\sqrt{\frac{2\log(4|\Pi|H(|\mathcal{D}^{\mathrm{E}}| + 1))}{|\mathcal{D}^{\mathrm{E}}| + 1}} \right] \\
&\overset{(a)}{\leq} 4H^2 \sqrt{\mathbb{E} \left[\frac{2\log(4|\Pi|H(|\mathcal{D}^{\mathrm{E}}| + 1))}{|\mathcal{D}^{\mathrm{E}}| + 1} \right]}.
\end{aligned}
$$

Inequality (a) follows Jensen's inequality. We consider the function $g(x) = -x\log(x/c)$ for $x \in (0, 1]$, where $c = 4|\Pi|H$.

$$
g'(x) = -(\log(x/c) + 1) \geq 0, \; g''(x) = -\frac{1}{x} \leq 0, \quad \forall x \in (0, 1].
$$

Thus, $g(x)$ is a concave function. By Jensen's inequality, we have that $\mathbb{E}[g(x)] \leq g(\mathbb{E}[x])$. Then we can derive that

$$
\begin{aligned}
\mathbb{E} \left[V(\pi^{\mathrm{E}}) - V(\pi^{\mathrm{BC}}) \right] &\leq 4H^2 \sqrt{\mathbb{E} \left[\frac{2\log(4|\Pi|H(|\mathcal{D}^{\mathrm{E}}| + 1))}{|\mathcal{D}^{\mathrm{E}}| + 1} \right]} \\
&= 4\sqrt{2}H^2 \sqrt{\mathbb{E} \left[g\left(\frac{1}{|\mathcal{D}^{\mathrm{E}}| + 1} \right) \right]} \\
&\leq 4\sqrt{2}H^2 \sqrt{g\left(\mathbb{E}\left[\frac{1}{|\mathcal{D}^{\mathrm{E}}| + 1} \right] \right)} \\
&\overset{(a)}{\leq} 4\sqrt{2}H^2 \sqrt{g\left(\frac{1}{N_{\mathrm{E}}} \right)} \\
&\leq 4\sqrt{2}H^2 \sqrt{\frac{\log(4|\Pi|HN_{\mathrm{E}})}{N_{\mathrm{E}}}}.
\end{aligned}
$$

In inequality (a), we use the facts that $g'(x) \geq 0$ and $\mathbb{E}\left[1/(|\mathcal{D}^{\mathrm{E}}| + 1)\right] \leq 1/N_{\mathrm{E}}$ from Lemma 3. We complete the proof.

F.2 Proof of Theorem 5

Despite the function approximation scheme, we can perform the same decomposition analysis as in the proof of Theorem 2. Therefore, we can obtain that

$$\mathbb{E}\left[V(\pi^{\mathrm{E}}) - V(\pi^{\mathrm{NBCU}})\right] = (1-\eta)\left(V(\pi^{\mathrm{E}}) - V(\pi^{\beta})\right) + \mathbb{E}\left[V(\pi^{\mathrm{mix}}) - V(\pi^{\mathrm{NBCU}})\right].$$

Recall that

$$\pi^{\mathrm{NBCU}} \in \max_{\pi \in \Pi} \sum_{h=1}^{H} \sum_{(s,a)\in \mathcal{S}\times\mathcal{A}} \widehat{d_h^{\mathrm{U}}}(s,a) \log \pi_h(a|s).$$

In the proof of Theorem 2, we have shown that $d_h^{\mathrm{U}}(s,a) = d_h^{\pi^{\mathrm{mix}}}(s,a)$, meaning that the state-action distribution of the union dataset equals the state-action distribution of the policy π^{mix}. Therefore, we can regard that π^{NBCU} is obtained by performing BC on the dataset generated by π^{mix}. Consequently, we can apply Theorem 4 to obtain that[3]

$$\mathbb{E}\left[V(\pi^{\mathrm{mix}}) - V(\pi^{\mathrm{NBCU}})\right] \leq 4\sqrt{2}H^2 \sqrt{\frac{\log(4|\Pi| H N_{\mathrm{tot}})}{N_{\mathrm{tot}}}}.$$

Finally, we arrive at

$$\mathbb{E}\left[V(\pi^{\mathrm{E}}) - V(\pi^{\mathrm{NBCU}})\right] = (1-\eta)\left(V(\pi^{\mathrm{E}}) - V(\pi^{\beta})\right) + 4\sqrt{2}H^2 \sqrt{\frac{\log(4|\Pi| H N_{\mathrm{tot}})}{N_{\mathrm{tot}}}},$$

which completes the proof.

F.3 Proof of Theorem 6

We first analyze the data distribution in $\mathcal{D}^{\mathrm{U}}$. According to Assumption 3, we summarize the sampling procedure of trajectories in $\mathcal{D}^{\mathrm{U}}$ as follows. Each time, we rollout the expert policy to collect a trajectory. Furthermore, with the probability of $(1-\eta)(1-\eta')$, the actions along the sampled expert trajectory are replaced with actions uniformly sampled from the action space. Then we put this poisoned expert trajectory into $\mathcal{D}^{\mathrm{U}}$. Otherwise, with the probability of $1-(1-\eta)(1-\eta')$, we directly put the original expert trajectory into $\mathcal{D}^{\mathrm{U}}$. Therefore, we can formulate the marginal distribution of the state-action pairs in time step h in $\mathcal{D}^{\mathrm{U}}$. For each $(s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]$,

$$\begin{aligned} d_h^{\mathrm{U}}(s,a) &= \left(1-(1-\eta)(1-\eta')\right) d_h^{\pi^{\mathrm{E}}}(s,a) + (1-\eta)(1-\eta') d_h^{\pi^{\mathrm{E}}}(s)\frac{1}{|\mathcal{A}|}, \\ d_h^{\mathrm{U}}(s) &= \sum_{a\in\mathcal{A}} d_h^{\mathrm{U}}(s,a) = d_h^{\pi^{\mathrm{E}}}(s). \end{aligned}$$

Then we proceed to analyze the imitation gap. Similar to the proof of Theorem 2, according to [32, Lemma 4.3], we have

$$V(\pi^{\mathrm{E}}) - V(\pi^{\mathrm{NBCU}}) \leq H \sum_{h=1}^{H} \mathbb{E}_{s\sim d_h^{\pi^{\mathrm{E}}}(\cdot)}\left[\mathrm{TV}\left(\pi_h^{\mathrm{E}}(\cdot|s), \pi_h^{\mathrm{NBCU}}(\cdot|s)\right)\right].$$

Again, we introduce the definition of the policy π^{mix}.

$$\forall (s,a)\in\mathcal{S}\times\mathcal{A}, \forall h\in[H],\ \pi_h^{\mathrm{mix}}(a|s) = \begin{cases} \frac{d_h^{\mathrm{U}}(s,a)}{d_h^{\mathrm{U}}(s)} & \text{if } d_h^{\mathrm{U}}(s) = d_h^{\pi^{\mathrm{E}}}(s) > 0, \\ \frac{1}{|\mathcal{A}|} & \text{otherwise.} \end{cases}$$

In particular, if $d_h^{\mathrm{U}}(s) > 0$, we have that

$$\pi_h^{\mathrm{mix}}(a|s) = \frac{d_h^{\mathrm{U}}(s,a)}{d_h^{\mathrm{U}}(s)} = \left(1-(1-\eta)(1-\eta')\right)\pi_h^{\mathrm{E}}(a|s) + (1-\eta)(1-\eta')\frac{1}{|\mathcal{A}|}.$$

Then we decompose the imitation gap into two parts.

$$V(\pi^{\mathrm{E}}) - V(\pi^{\mathrm{NBCU}})$$

[3] Note that Theorem 4 holds for the case where the expert policy is stochastic.

$$
\begin{aligned}
&\leq H \sum_{h=1}^{H} \mathbb{E}_{s \sim d_h^{\pi^{\mathrm{E}}}(\cdot)} \left[\mathrm{TV}\left(\pi_h^{\mathrm{E}}(\cdot|s), \pi_h^{\mathrm{NBCU}}(\cdot|s) \right) \right] \\
&\leq H \sum_{h=1}^{H} \mathbb{E}_{s \sim d_h^{\pi^{\mathrm{E}}}(\cdot)} \left[\mathrm{TV}\left(\pi_h^{\mathrm{E}}(\cdot|s), \pi_h^{\mathrm{mix}}(\cdot|s) \right) \right] + H \sum_{h=1}^{H} \mathbb{E}_{s \sim d_h^{\pi^{\mathrm{E}}}(\cdot)} \left[\mathrm{TV}\left(\pi_h^{\mathrm{mix}}(\cdot|s), \pi_h^{\mathrm{NBCU}}(\cdot|s) \right) \right].
\end{aligned}
$$

We first analyze the first term in RHS. For certain (s,h) such $d_h^{\mathrm{U}}(s) = d_h^{\pi^{\mathrm{E}}}(s) > 0$, we have that

$$
\begin{aligned}
\mathrm{TV}\left(\pi_h^{\mathrm{E}}(\cdot|s), \pi_h^{\mathrm{mix}}(\cdot|s) \right) &= \sum_{a \neq \pi_h^{\mathrm{E}}(s)} \pi_h^{\mathrm{mix}}(a|s) \\
&= \sum_{a \neq \pi_h^{\mathrm{E}}(s)} \left(1 - (1-\eta)(1-\eta') \right) \pi_h^{\mathrm{E}}(a|s) + (1-\eta)(1-\eta') \frac{1}{|\mathcal{A}|} \\
&= (1-\eta)(1-\eta') \left(1 - \frac{1}{|\mathcal{A}|} \right).
\end{aligned}
$$

Therefore, we can derive that

$$
H \sum_{h=1}^{H} \mathbb{E}_{s \sim d_h^{\pi^{\mathrm{E}}}(\cdot)} \left[\mathrm{TV}\left(\pi_h^{\mathrm{E}}(\cdot|s), \pi_h^{\mathrm{mix}}(\cdot|s) \right) \right] \leq (1-\eta)(1-\eta') H^2 \left(1 - \frac{1}{|\mathcal{A}|} \right).
$$

Now we analyze the second term of

$$
H \sum_{h=1}^{H} \mathbb{E}_{s \sim d_h^{\pi^{\mathrm{E}}}(\cdot)} \left[\mathrm{TV}\left(\pi_h^{\mathrm{mix}}(\cdot|s), \pi_h^{\mathrm{NBCU}}(\cdot|s) \right) \right].
$$

Recall the formula of π^{NBCU}.

$$
\pi_h^{\mathrm{NBCU}}(a|s) = \begin{cases} \frac{n_h^{\mathrm{U}}(s,a)}{n_h^{\mathrm{U}}(s)} & \text{if } n_h^{\mathrm{U}}(s) > 0 \\ \frac{1}{|\mathcal{A}|} & \text{otherwise} \end{cases}
$$

Notice that π^{NBCU} is the maximum likelihood estimation of π^{mix}. According to the concentration inequality of total variation [43], for each $(s,h) \in \mathcal{S} \times [H]$, for any fixed $\delta \in (0,1)$, when $n_h^{\mathrm{U}}(s) > 0$, with probability at least $1-\delta$, we have

$$
\mathrm{TV}\left(\pi_h^{\mathrm{mix}}(\cdot|s), \pi_h^{\mathrm{NBCU}}(\cdot|s) \right) \leq \sqrt{\frac{|\mathcal{A}| \log(3/\delta)}{n_h^{\mathrm{U}}(s)}}.
$$

When $n_h^{\mathrm{U}}(s) = 0$, we have that

$$
\mathrm{TV}\left(\pi_h^{\mathrm{mix}}(\cdot|s), \pi_h^{\mathrm{NBCU}}(\cdot|s) \right) \leq 1 \leq \sqrt{|\mathcal{A}| \log(3/\delta)}.
$$

By combining the above two inequalities, for each $(s,h) \in \mathcal{S} \times [H]$, with probability at least $1-\delta$, we have

$$
\mathrm{TV}\left(\pi_h^{\mathrm{mix}}(\cdot|s), \pi_h^{\mathrm{NBCU}}(\cdot|s) \right) \leq \sqrt{\frac{|\mathcal{A}| \log(3/\delta)}{\max\{n_h^{\mathrm{U}}(s), 1\}}}.
$$

Applying union bound yields that with probability at least $1-\delta/2$, for all $(s,h) \in \mathcal{S} \times [H]$,

$$
\mathrm{TV}\left(\pi_h^{\mathrm{mix}}(\cdot|s), \pi_h^{\mathrm{NBCU}}(\cdot|s) \right) \leq \sqrt{\frac{|\mathcal{A}| \log(6|\mathcal{S}|H/\delta)}{\max\{n_h^{\mathrm{U}}(s), 1\}}}.
$$

Then we have that

$$
\begin{aligned}
&H \sum_{h=1}^{H} \mathbb{E}_{s \sim d_h^{\pi^{\mathrm{E}}}(\cdot)} \left[\mathrm{TV}\left(\pi_h^{\mathrm{mix}}(\cdot|s), \pi_h^{\mathrm{NBCU}}(\cdot|s) \right) \right] \\
&\leq H \sum_{h=1}^{H} \mathbb{E}_{s \sim d_h^{\pi^{\mathrm{E}}}(\cdot)} \left[\sqrt{\frac{|\mathcal{A}| \log(6|\mathcal{S}|H/\delta)}{\max\{n_h^{\mathrm{U}}(s), 1\}}} \right] \\
&= H \sqrt{|\mathcal{A}| \log(6|\mathcal{S}|H/\delta)} \sum_{h=1}^{H} \mathbb{E}_{s \sim d_h^{\pi^{\mathrm{E}}}(\cdot)} \left[\sqrt{\frac{1}{\max\{n_h^{\mathrm{U}}(s), 1\}}} \right]
\end{aligned}
$$

$$
\begin{aligned}
&= H\sqrt{|\mathcal{A}|\log(6|\mathcal{S}|H/\delta)}\sum_{h=1}^{H}\sum_{s\in\mathcal{S}}\sqrt{d_h^{\pi^{\mathrm{E}}}(s)}\sqrt{\frac{d_h^{\pi^{\mathrm{E}}}(s)}{\max\{n_h^{\mathrm{U}}(s),1\}}} \\
&\leq H\sqrt{|\mathcal{A}|\log(6|\mathcal{S}|H/\delta)}\sum_{h=1}^{H}\sqrt{\sum_{s\in\mathcal{S}}\frac{d_h^{\pi^{\mathrm{E}}}(s)}{\max\{n_h^{\mathrm{U}}(s),1\}}}.
\end{aligned}
$$

Here the last inequality follows Cauchy-Swartz inequality. Notice that $n_h^{\mathrm{U}}(s)$ is the number of times that the state s appears in $\mathcal{D}^{\mathrm{U}}$ in time step h and thus follows the Binomial distribution of $\mathrm{Bin}(N_{\mathrm{tot}}, d_h^{\pi^{\mathrm{E}}}(s))$. By applying Lemma 4, for each (s,h), with probability at least $1-\delta$, we have

$$
\frac{d_h^{\pi^{\mathrm{E}}}(s)}{\max\{n_h^{\mathrm{U}}(s),1\}} \leq \frac{8\log(1/\delta)}{N_{\mathrm{tot}}}.
$$

By union bound, with probability at least $1-\delta/2$, for all $(s,h)\in\mathcal{S}\times[H]$,

$$
\frac{d_h^{\pi^{\mathrm{E}}}(s)}{\max\{n_h^{\mathrm{U}}(s),1\}} \leq \frac{8\log(2|\mathcal{S}|H/\delta)}{N_{\mathrm{tot}}}.
$$

Then, with probability at least $1-\delta$, we have

$$
H\sum_{h=1}^{H}\mathbb{E}_{s\sim d_h^{\pi^{\mathrm{E}}}(\cdot)}\left[\mathrm{TV}\left(\pi_h^{\mathrm{mix}}(\cdot|s),\pi_h^{\mathrm{NBCU}}(\cdot|s)\right)\right] \leq H^2\sqrt{\frac{8|\mathcal{S}||\mathcal{A}|\log^2(6|\mathcal{S}|H/\delta)}{N_{\mathrm{tot}}}}.
$$

Finally, we upper bound the imitation gap. With probability at least $1-\delta$, we have

$$
V(\pi^{\mathrm{E}}) - V(\pi^{\mathrm{NBCU}}) \leq (1-\eta)(1-\eta')\left(1-\frac{1}{|\mathcal{A}|}\right) + H^2\sqrt{\frac{8|\mathcal{S}||\mathcal{A}|\log^2(6|\mathcal{S}|H/\delta)}{N_{\mathrm{tot}}}}.
$$

We set $\delta = H/N_{\mathrm{tot}}$ and obtain that

$$
\begin{aligned}
&\mathbb{E}\left[V(\pi^{\mathrm{E}}) - V(\pi^{\mathrm{NBCU}})\right] \\
&\leq \delta H + (1-\delta)\left((1-\eta)(1-\eta')\left(1-\frac{1}{|\mathcal{A}|}\right) + H^2\sqrt{\frac{8|\mathcal{S}||\mathcal{A}|\log^2(6|\mathcal{S}|H/\delta)}{N_{\mathrm{tot}}}}\right) \\
&\leq \frac{H^2}{N_{\mathrm{tot}}} + (1-\eta)(1-\eta')\left(1-\frac{1}{|\mathcal{A}|}\right) + H^2\sqrt{\frac{8|\mathcal{S}||\mathcal{A}|\log^2(6|\mathcal{A}|N_{\mathrm{tot}})}{N_{\mathrm{tot}}}} \\
&\leq (1-\eta)(1-\eta')\left(1-\frac{1}{|\mathcal{A}|}\right) + 4H^2\sqrt{\frac{2|\mathcal{S}||\mathcal{A}|\log^2(6|\mathcal{A}|N_{\mathrm{tot}})}{N_{\mathrm{tot}}}}.
\end{aligned}
$$

On the other hand, we directly have $\mathbb{E}[V(\pi^{\mathrm{E}}) - V(\pi^{\mathrm{NBCU}})] \leq H$. We complete the proof.

F.4 Proof of Theorem 7

In the poison setting, we can conduct the same analysis as in the proof of Theorem 3 and demonstrate that $\pi^{\text{ISW-BC}}(\pi_h^{\mathrm{E}}(s)|s) = 1,\ \forall (s,\pi_h^{\mathrm{E}}(s)) \in \mathcal{D}_h^{\mathrm{E}} \cup \mathcal{D}_h^{\mathrm{S},1}$, where $\mathcal{D}_h^{\mathrm{E}}$ is the set of state-action pairs in $\mathcal{D}^{\mathrm{E}}$ in time step h and $\mathcal{D}_h^{\mathrm{S},1} = \{(s,a)\in\mathcal{D}_h^{\mathrm{S}} : d_h^{\pi^{\mathrm{E}}}(s) > 0, a = \pi_h^{\mathrm{E}}(s)\}$. According to [32, Lemma 4.3], we have

$$
V(\pi^{\mathrm{E}}) - V(\pi^{\text{ISW-BC}}) \leq H\sum_{h=1}^{H}\mathbb{E}_{s\sim d_h^{\pi^{\mathrm{E}}}(\cdot)}\left[\mathrm{TV}\left(\pi_h^{\mathrm{E}}(\cdot|s),\pi_h^{\text{ISW-BC}}(\cdot|s)\right)\right].
$$

Since the expert policy is assumed to be deterministic, we can obtain

$$
\begin{aligned}
V(\pi^{\mathrm{E}}) - V(\pi^{\text{ISW-BC}}) &\leq H\sum_{h=1}^{H}\mathbb{E}_{s\sim d_h^{\pi^{\mathrm{E}}}(\cdot)}\left[\mathbb{E}_{a\sim\pi_h^{\text{ISW-BC}}(\cdot|s)}\left[\mathbb{I}\left\{a\neq\pi_h^{\mathrm{E}}(s)\right\}\right]\right] \\
&\leq H\sum_{h=1}^{H}\mathbb{E}_{s\sim d_h^{\pi^{\mathrm{E}}}(\cdot)}\left[\mathbb{I}\left\{(s,\pi_h^{\mathrm{E}}(s))\notin\mathcal{D}_h^{\mathrm{E}}\cup\mathcal{D}_h^{\mathrm{S},1}\right\}\right].
\end{aligned}
$$

Let $\mathcal{D}^{\mathrm{S,clean}}$ denote the non-corrupted dataset in $\mathcal{D}^{\mathrm{S}}$. Then we can obtain that

$$\begin{aligned} V(\pi^{\mathrm{E}}) - V(\pi^{\text{ISW-BC}}) &\overset{(a)}{\leq} H \sum_{h=1}^{H} \mathbb{E}_{s \sim d_h^{\pi^{\mathrm{E}}}(\cdot)} \left[\mathbb{I} \left\{ (s, \pi_h^{\mathrm{E}}(s)) \notin \mathcal{D}_h^{\mathrm{E}} \cup \mathcal{D}_h^{\mathrm{S,clean}} \right\} \right] \\ &= H \sum_{h=1}^{H} \sum_{s \in \mathcal{S}} d_h^{\pi^{\mathrm{E}}}(s) \mathbb{I} \left\{ (s, \pi_h^{\mathrm{E}}(s)) \notin \mathcal{D}_h^{\mathrm{E}} \cup \mathcal{D}_h^{\mathrm{S,clean}} \right\}, \end{aligned}$$

where $\mathcal{D}_h^{\mathrm{S,clean}}$ denotes the set of state-action pairs in $\mathcal{D}^{\mathrm{S,clean}}$ in time step h. Inequality (a) follows that $\mathcal{D}_h^{\mathrm{S,clean}} \subseteq \mathcal{D}_h^{\mathrm{S},1}$ since $\mathcal{D}^{\mathrm{S,clean}}$ is collected by the expert policy. Taking expectation over the randomness in $\mathcal{D}^{\mathrm{E}}$ and $\mathcal{D}^{\mathrm{S,clean}}$ on both sides yields that

$$\mathbb{E}_{\mathcal{D}^{\mathrm{E}}, \mathcal{D}^{\mathrm{S,clean}}} \left[V(\pi^{\mathrm{E}}) - V(\pi^{\text{ISW-BC}}) \right] \leq H \sum_{h=1}^{H} \sum_{s \in \mathcal{S}} d_h^{\pi^{\mathrm{E}}}(s) \mathbb{P} \left((s, \pi_h^{\mathrm{E}}(s)) \notin \mathcal{D}_h^{\mathrm{E}} \cup \mathcal{D}_h^{\mathrm{S,clean}} \right).$$

Notice that both $\mathcal{D}^{\mathrm{E}}$ and $\mathcal{D}^{\mathrm{S,clean}}$ are collected by the expert policy. Then if $|\mathcal{D}^{\mathrm{E}}| + |\mathcal{D}^{\mathrm{S,clean}}| \geq 1$, we can calculate that for each $(s, h) \in \mathcal{S} \times [H]$,

$$\begin{aligned} d_h^{\pi^{\mathrm{E}}}(s) \mathbb{P} \left((s, \pi_h^{\mathrm{E}}(s)) \notin \mathcal{D}_h^{\mathrm{E}} \cup \mathcal{D}_h^{\mathrm{S,clean}} \right) &= d_h^{\pi^{\mathrm{E}}}(s) \left(1 - d_h^{\pi^{\mathrm{E}}}(s) \right)^{|\mathcal{D}^{\mathrm{E}}| + |\mathcal{D}^{\mathrm{S,clean}}|} \\ &\leq \frac{4}{9(|\mathcal{D}^{\mathrm{E}}| + |\mathcal{D}^{\mathrm{S,clean}}|)}, \end{aligned}$$

where the last inequality follows Lemma 5. If $|\mathcal{D}^{\mathrm{E}}| + |\mathcal{D}^{\mathrm{S,clean}}| = 0$, we directly have that

$$d_h^{\pi^{\mathrm{E}}}(s) \mathbb{P} \left((s, \pi_h^{\mathrm{E}}(s)) \notin \mathcal{D}_h^{\mathrm{E}} \cup \mathcal{D}_h^{\mathrm{S,clean}} \right) \leq 1 = \frac{1}{\max\{|\mathcal{D}^{\mathrm{E}}| + |\mathcal{D}^{\mathrm{S,clean}}|, 1\}}.$$

We unify the above two inequalities and get that

$$d_h^{\pi^{\mathrm{E}}}(s) \mathbb{P} \left((s, \pi_h^{\mathrm{E}}(s)) \notin \mathcal{D}_h^{\mathrm{E}} \cup \mathcal{D}_h^{\mathrm{S,clean}} \right) \leq \frac{1}{\max\{|\mathcal{D}^{\mathrm{E}}| + |\mathcal{D}^{\mathrm{S,clean}}|, 1\}}.$$

Now we proceed to upper bound the imitation gap.

$$\begin{aligned} \mathbb{E}_{\mathcal{D}^{\mathrm{E}}, \mathcal{D}^{\mathrm{S,clean}}} \left[V(\pi^{\mathrm{E}}) - V(\pi^{\text{ISW-BC}}) \right] &\leq H \sum_{h=1}^{H} \sum_{s \in \mathcal{S}} d_h^{\pi^{\mathrm{E}}}(s) \mathbb{P} \left((s, \pi_h^{\mathrm{E}}(s)) \notin \mathcal{D}_h^{\mathrm{E}} \cup \mathcal{D}_h^{\mathrm{S,clean}} \right) \\ &\leq \frac{|\mathcal{S}| H^2}{\max\{|\mathcal{D}^{\mathrm{E}}| + |\mathcal{D}^{\mathrm{S,clean}}|, 1\}}. \end{aligned}$$

Note that $|\mathcal{D}^{\mathrm{E}}| + |\mathcal{D}^{\mathrm{S,clean}}| \sim \mathrm{Bin}(N_{\mathrm{tot}}, \eta + (1 - \eta)\eta')$. Taking expectation with respect to $|\mathcal{D}^{\mathrm{E}}| + |\mathcal{D}^{\mathrm{S,clean}}|$ yields that

$$\begin{aligned} \mathbb{E} \left[V(\pi^{\mathrm{E}}) - V(\pi^{\text{ISW-BC}}) \right] &\leq \mathbb{E} \left[\frac{|\mathcal{S}| H^2}{\max\{|\mathcal{D}^{\mathrm{E}}| + |\mathcal{D}^{\mathrm{S,clean}}|, 1\}} \right] \\ &\leq \mathbb{E} \left[\frac{2|\mathcal{S}| H^2}{|\mathcal{D}^{\mathrm{E}}| + |\mathcal{D}^{\mathrm{S,clean}}| + 1} \right] \\ &\overset{(a)}{\leq} \frac{2|\mathcal{S}| H^2}{N_{\mathrm{tot}}(\eta + (1 - \eta)\eta')} \\ &= \frac{2|\mathcal{S}| H^2}{N_{\mathrm{E}} + \eta' N_{\mathrm{S}}}. \end{aligned}$$

Inequality (a) follows Lemma 3. We finish the proof.

G Technical Lemmas

Lemma 3. *For any $N \in \mathbb{N}_+$ and $p \in (0, 1)$, if the random variable X follows the binomial distribution, i.e., $X \sim \mathrm{Bin}(N, p)$, then we have that*

$$\mathbb{E} \left[\frac{1}{X + 1} \right] \leq \frac{1}{Np}.$$

Proof.

$$\mathbb{E}\left[\frac{1}{X+1}\right] = \sum_{x=0}^{N}\left(\frac{1}{x+1}\right)\frac{N!}{x!(N-x)!}p^x(1-p)^{N-x}$$
$$= \frac{1}{(N+1)p}\sum_{x=1}^{N+1}\left(\frac{(N+1)!}{x!(N+1-x)!}\right)p^x(1-p)^{N+1-x}$$
$$= \frac{1}{(N+1)p}\left(1-(1-p)^{N+1}\right) \leq \frac{1}{Np}.$$

□

Lemma 4 (Binomial concentration (Lemma A.1 in [45])). *For any $N \in \mathbb{N}_+$ and $p \in (0,1)$, suppose $X \sim \mathrm{Bin}(N,p)$. Then with probability at least $1-\delta$, we have*

$$\frac{p}{\max\{X,1\}} \leq \frac{8\log(1/\delta)}{N}.$$

Lemma 5. *For any $N \in \mathbb{N}_+$ and $x \in [0,1]$, consider the function $f(x) := x(1-x)^N$, then we have*

$$\forall x \in [0,1],\ f(x) \leq \frac{4}{9N}.$$

Proof. We calculate that $f'(x) = (1-x)^{N-1}(1-(N+1)x)$. It is direct to have that $f(x)$ achieves its maximum at $x^\star = 1/(N+1)$. Furthermore, we have

$$f(\frac{1}{N+1}) = \frac{1}{N}\left(1-\frac{1}{N+1}\right)^{N+1} \overset{(a)}{\leq} \frac{1}{eN} \leq \frac{4}{9N}.$$

Inequality (a) follows that $(1+x/N)^N \leq \exp(x),\ \forall N \geq 1, |x| \leq N$. We complete the proof. □

H Experiments Details and Additional Results

H.1 Experiment Details

In this section, we present the experiment details to facilitate the replication of our results. Our codebase will be made available for public access at a later stage. The experiments are conducted on a machine comprising 48 CPU cores and 4 V100 GPU cores. We repeat each experiment 5 times using different random seeds (2021, 2022, 2023, 2024, and 2025).

H.1.1 Locomotion Control

In this study, we evaluate the performance of various imitation learning algorithms on four locomotion control tasks from the MuJoCo suite: `Ant-v2`, `HalfCheetah-v2`, `Hopper-v2`, and `Walker2d-v2`. These tasks are widely used in the literature and are considered challenging benchmarks.

To train the expert policy, we use the online Soft Actor-Critic (SAC) algorithm [16] with 1 million training steps. We implement the algorithm using the rlkit codebase, which is available at `https://github.com/rail-berkeley/rlkit`. The training curves of the online SAC agent are shown in Figure 5. We treat the resulting policy as the expert policy and use it to generate expert trajectories.

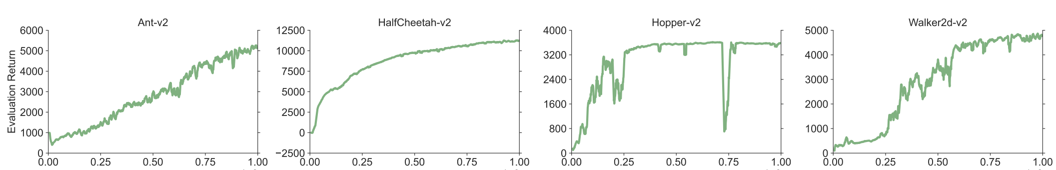

Figure 5: Training curves of online SAC on 4 locomotion control environments.

In our experimental setup, we utilize an expert dataset comprising of 1 expert trajectory collected by the trained SAC agent. Additionally, all algorithms are provided with a supplementary dataset. There are two setting of the supplementary data.

- `Full Replay`. The supplementary dataset is obtained from the replay buffer of the online SAC agent, which has over one million samples, equivalent to 1000+ trajectories. The rapid convergence of online SAC, as illustrated in Figure 5, implies that the replay buffer is enriched with a substantial number of expert-level trajectories. As a result, we expect that utilizing the supplementary data without any modification may lead to desirable results.
- `Noisy Expert`. The supplementary dataset comprises of 10 clean expert trajectories and 5 noisy expert trajectories. In this case, we replace the action labels in the noisy trajectories with random actions drawn from $[-1, 1]$. This replacement creates noisy action labels for the expert states, leading to a significant distribution shift at the state-action level, as noted in Remark 1. The high degree of distribution shift makes it challenging for using the supplementary data.

We use a 2-hidden-layer multi-layer perceptron (MLP) with hidden size 256 and ReLU activation for all algorithms, as the state information in locomotion control tasks is informative by design. The codebase of DemoDICE is based on the original authors' work, which can be accessed at `https://github.com/KAIST-AILab/imitation-dice`. For DWBC, we also use the authors' codebase, which is available at `https://github.com/ryanxhr/DWBC`. We experimented with different hyper-parameters for both algorithms but found that the default parameters provided by the authors work well. We normalize state observations in the dataset before training all algorithms, following [22]. This is crucial for achieving satisfactory performance.

In training the discriminator of ISW-BC, we use the gradient penalty (GP) regularization, as recommended by [22]. We add the following loss to the original loss (4) to enforce 1-Lipschitz continuity:

$$\min_{\theta} \sum_{(s,a) \in \mathcal{B}} \left(\|g(s, a; \theta)\| - 1 \right)^2 ,$$

where g is the gradient of the discriminator $c(s, a; \theta)$, and $\mathcal{B}$ is a mini-batch. This promotes the learning of smooth features and can improve generalization performance.

In our implementation of ISW-BC, we employ 2-hidden-layer MLPs with 256 hidden units and ReLU activation for both the discriminator and policy networks. We use a batch size of 256 and Adam optimizer with a learning rate of 0.0003 for training both networks. The training objective is to maximize the log-likelihood. We set δ to 0 and use a gradient penalty coefficient of 8 by default, unless otherwise stated. The training process is carried out for 1 million iterations. We evaluate the performance every 10k iterations with 10 episodes. The normalized score in the last column of Table 2 is computed in the following way:

$$\text{Normalized score} = \frac{\text{Expert performance} - \text{Agent performance}}{\text{Expert performance} - \text{Random policy performance}}. \tag{21}$$

H.1.2 Atari Games

We evaluate algorithms on a set of 5 Atari games from the standard benchmark: `Alien`, `MsPacman`, `Phoenix`, `Qbert`, and `SpaceInvaders`. We preprocess the game environments using a standard set of procedures, including sticky actions with a probability of 0.25, grayscaling, downsampling to an image size of [84, 84], and stacking frames of 4. These procedures follow the instructions provided by the dopamine codebase, which is available at `https://github.com/google/dopamine/blob/master/dopamine/discrete_domains/atari_lib.py`. The final image inputs are of shape (84, 84, 4).

We use the replay buffer data from an online DQN agent, which is publicly available at `https://console.cloud.google.com/storage/browser/atari-replay-datasets`, thanks to the work of [2]. The dataset consists of 200 million frames, divided into 50 indexed buckets (ranging from 0 to 49). However, using the entire dataset is computationally infeasible[4] and unnecessary for our task. Therefore, we select specific buffer buckets for imitation learning.

We choose the expert data from bucket index 49, using only the first 400K frames for training. This makes the task challenging (we find that BC performs well with 1M frames of expert data). For the full replay setting, we select supplementary data from buffer indices 45 to 48, using the first 400K frames from each bucket. This yields a supplementary dataset that is 4 times larger than the expert

[4]Loading 200M frames requires over 500GB memory.

data. In the noisy task setting, we follow the same procedure for selecting supplementary data, but replace the action labels with random labels on buffer index 45.

All agents employ the same convolutional neural network (CNN) architecture as the DQN agent, consisting of three convolutional blocks. The first block applies a filter size of 8, a stride of 4, and has a channel size of 32. The second block uses a filter size of 4, a stride of 4, and a channel size of 64, while the third block applies a filter size of 3, a stride of 4, and has a channel size of 64. All blocks use the ReLU activation function. The feature representations are flattened to a vector, on which a 1-hidden-layer MLP with a hidden size of 512 and ReLU activation function is applied. Finally, the outputs are passed through a softmax function to obtain a probability distribution.

Atari games are not considered in [22, 47] and public implementations of DemoDICE and DWBC for Atari games are not available. To use these methods in the Atari environment, we extend their original implementation by replacing the MLP used in locomotion control with the CNN described earlier. Implementing ISW-BC is a little more complicated. We use the same CNN policy network as in the other methods, but find that directly training the discriminator from scratch is less effective. This is because the discriminator tends to focus on irrelevant background information instead of the decision-centric part. To overcome this issue, we build the discriminator upon the feature extractor of the policy network, leveraging its ability to extract useful information. The discriminator is an MLP with ReLU activation and a hidden size of 1024: the image feature representation has a dimension 512 and the action feature representation also has a dimension 512 (we randomly project one-hot discrete actions to a 512-dimension space). We find that the depth of the MLP is crucial for performance, using a depth of 1 for the full replay setting and 3 for the noisy expert setting. We clip the importance sampling ratio for numerical stability, using a minimum value of 0 and a maximum value of 5 for the full replay setting, and a minimum value of 0.2 and a maximum value of 5 for the noisy expert setting. We provide ablation studies of these hyperparameters in Appendix H.2.2.

All methods were optimized using the Adam optimizer with a learning rate of 0.00025 and a batch size of 256. The training objective is to maximize the log-likelihood. The training process consisted of 200K gradient steps. Every 2K gradient steps, the algorithms were evaluated by running 10 episodes and computing the raw game scores. The normalized score in the last column of Table 3 is computed by Eq. (21).

H.1.3 Object Recognition

We utilize the publicly available DomainNet dataset [29] for our experiments, which can be accessed at `http://csr.bu.edu/ftp/visda/2019/multi-source`. This dataset comprises six sub-datasets: `clipart`, `infograph`, `painting`, `quickdraw`, `real`, and `sketch`, with 2103, 2626, 2472, 4000, 4864, and 2213 images, respectively. Our task involves recognizing objects from 10 different classes: `bird`, `feather`, `headphones`, `ice_cream`, `teapot`, `tiger`, `whale`, `windmill`, `wine_glass`, and `zebra`. We divided the images into training and test sets, with 80% for training and 20% for testing.

We employ a 2-hidden-layer neural network with a hidden size of 512 and ReLU activation as the classifier. To extract features from images, we utilize the pretrained ResNet-18 model (trained on ImageNet), which has a feature dimension of 512. The ResNet-18 model can be accessed at `https://pytorch.org/vision/main/models/generated/torchvision.models.resnet18.html`. We opted for this approach as training such a large convolutional neural network directly on the DomainNet dataset proved to be ineffective. The training objective is to minimize the cross-entropy loss. To optimize the network parameters, we use the stochastic gradient descent (SGD) optimizer with a learning rate of 0.01 and momentum of 0.9. Additionally, we apply weight decay with a coefficient of 0.0005. The models are trained for 100 epochs with a batch size of 100, following the standard practice.

The discriminators used in ISW-BC and DWBC are implemented as 2-hidden-layer neural networks with ReLU activation. It's important to note that these discriminators take both the image and label as inputs. The image input is processed by the pre-trained and fixed ResNet-18, while the label input is projected to the same dimension (512) by a random projection matrix. The hidden size for the discriminator is set to 1024 for ISW-BC and 1025 for DWBC, as the discriminator in DWBC also takes the log-likelihood as an input. For ISW-BC, the discriminator is trained independently for 100 epochs with the same optimization configuration as the classifier. Afterward, the discriminator is fixed, and its output is used to compute the importance sampling ratio, which is then used to train the classifier.

H.2 Additional Results

H.2.1 Training Curves

Training curves. The training curves on the MuJoco locomotion control tasks are displayed in Figure 6 and Figure 7. The training curves on Atari games are displayed in Figure 8 and Figure 9. The training curves on the object recognition task are displayed in Figure 10.

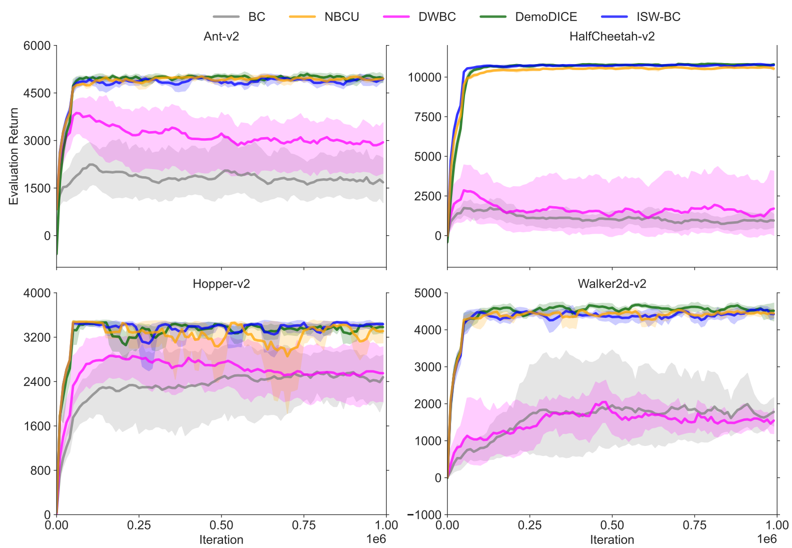

Figure 6: Training curves of algorithms on the locomotion control task in the full replay setting. Solid lines correspond to the mean performance and shaded regions correspond to the 95% confidence interval. Same as other figures.

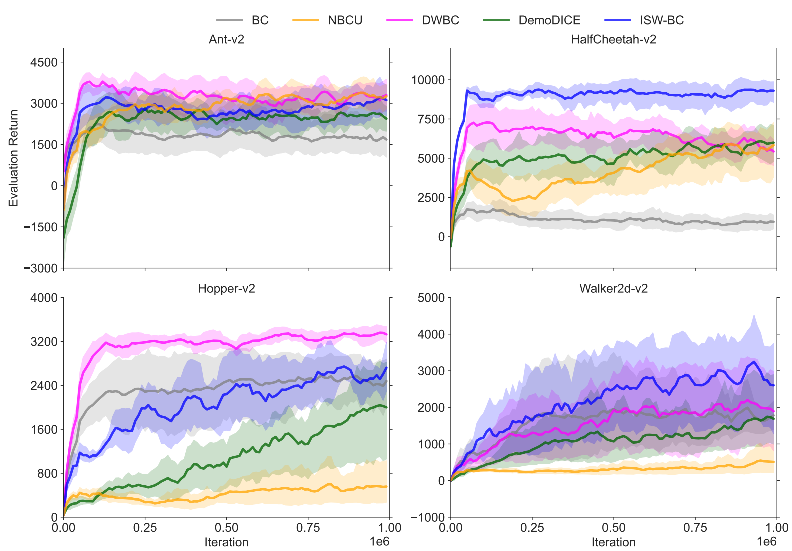

Figure 7: Training curves of algorithms on the locomotion control task in the noisy expert setting.

H.2.2 Ablation Study

In this section, we present ablation studies conducted on Atari games, aiming to provide insights into the underlying working scheme of our method. We specifically emphasize Atari games due to their high-dimensional image inputs, making these tasks particularly challenging. In contrast, the other

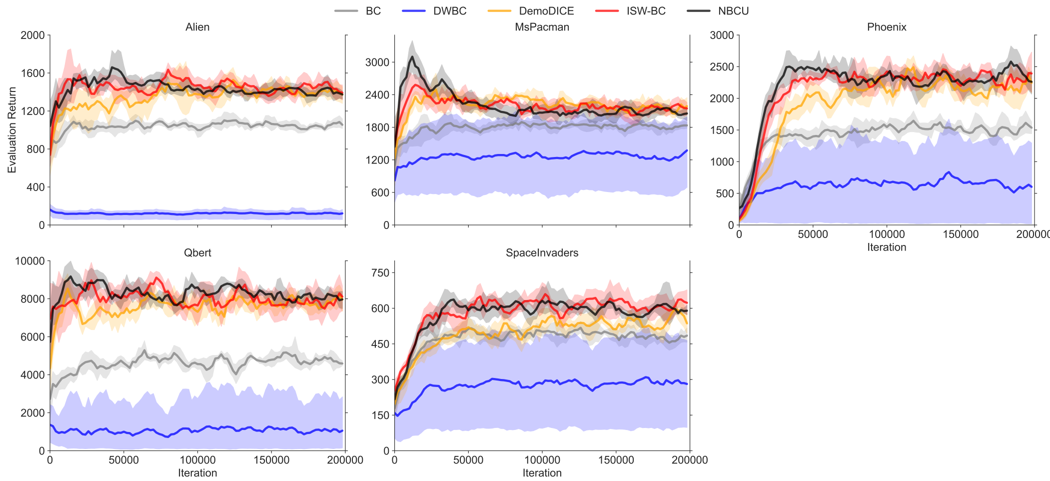

Figure 8: Training curves of algorithms on the Atari games in the full replay setting.

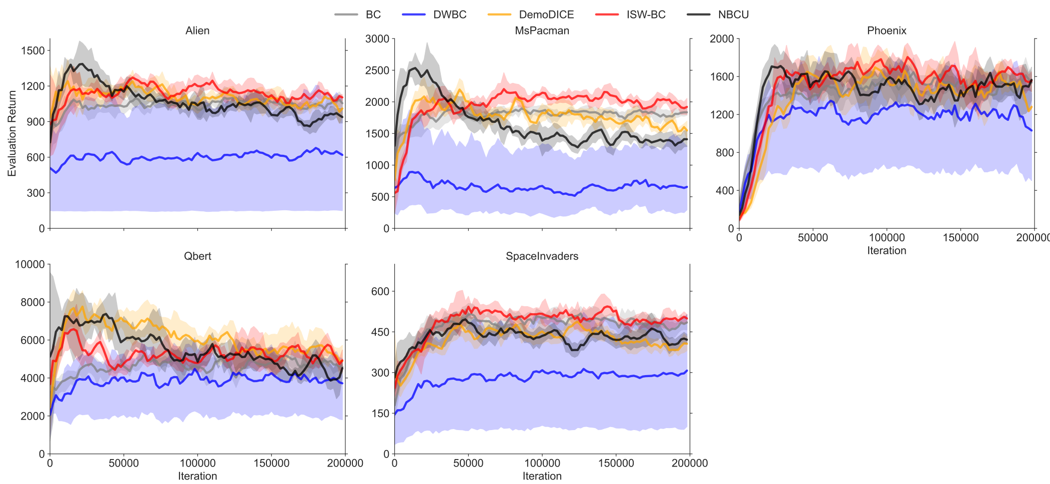

Figure 9: Training curves of algorithms on the Atari games in the noisy expert setting.

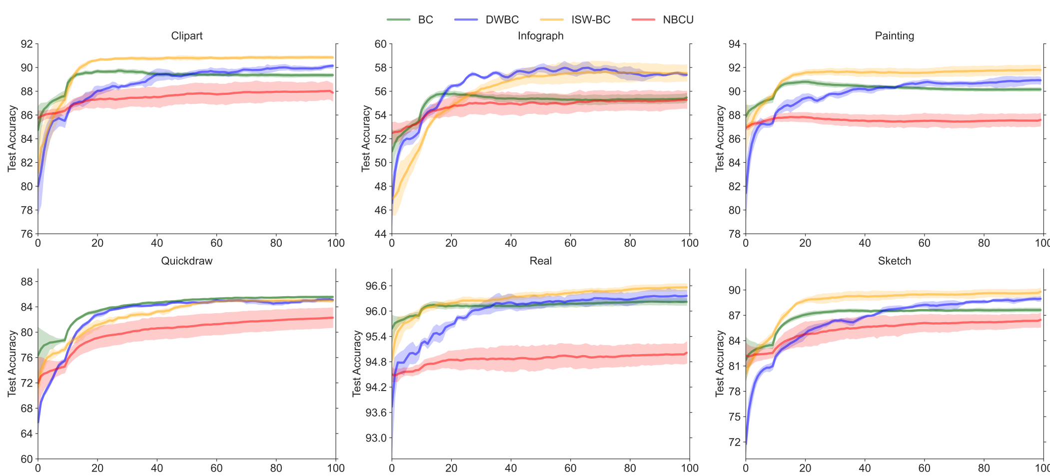

Figure 10: Training curves of algorithms on the object recognition task using the DomainNet dataset.

two tasks, locomotion control and object recognition, involve informative vector inputs, setting them apart from the unique characteristics of Atari games.

Ablation Study on Feature Representations of Discriminator Network. Our study reveals that employing a separate CNN for the discriminator yields inferior results compared to utilizing the feature extractor of the policy network. Please refer to Figure 11. Our conjecture is that training the discriminator independently may cause it to fit noise information (e.g., background). In contrast, the policy CNN network is capable of learning decision-centric information, enabling an effective approach to building the discriminator network through the feature extractor of the policy network.

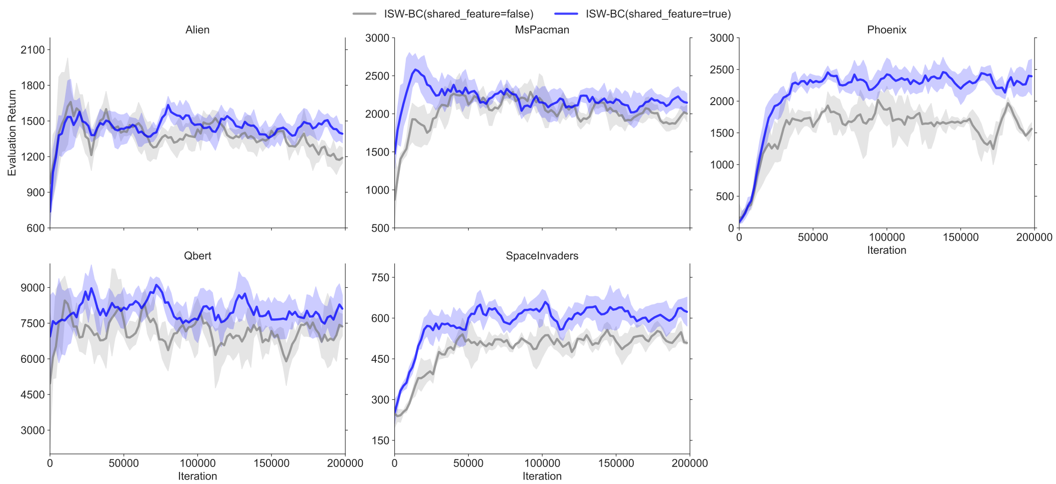

Figure 11: Training curves of ISW-BC on the Atari games in the full replay setting. We test the performance with different feature extractors of the discriminator.

Ablation Study on Depth of Discriminator Network. We have discovered that the number of discriminator layers plays a crucial role in the performance of Atari games. The training curves, depicted in both Figure 12 and Figure 13, illustrate the performance variation based on the number of layers in the discriminator network. Notably, a 1-hidden-layer neural network yields the best results for the full replay setting, while a 3-hidden-layer neural network performs optimally in the noisy expert setting. It is important to note that this phenomenon is specific to Atari games. We do not have a good explanation yet. We believe this deserves further investigation in the future work.

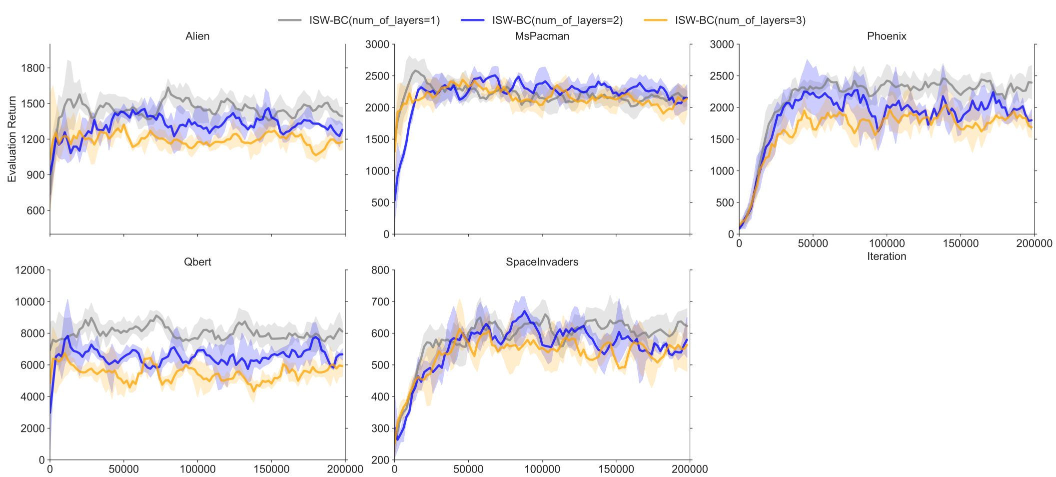

Figure 12: Training curves of ISW-BC on the Atari games in the full replay setting. We test the performance with different number of layers for the discriminator network.

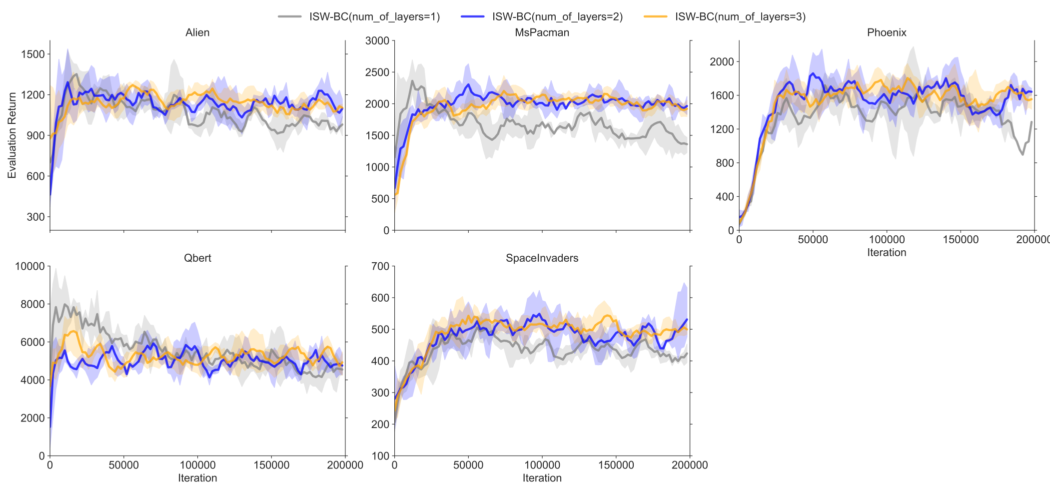

Figure 13: Training curves of ISW-BC on the Atari games with the noisy expert setting. We test the performance with different number of layers for the discriminator network.