# A Proof

**Theorem 3.1** *We denote the Q-function converged from the Q-update of EPQ using the proposed penalty $\mathcal{P}_\tau$ in (3) by $\hat{Q}^\pi$. Then, the expected value of $\hat{Q}^\pi$ underestimates the expected true policy value, i.e., $\mathbb{E}_{a \sim \pi}[\hat{Q}^\pi(s,a)] \leq \mathbb{E}_{a \sim \pi}[Q^\pi(s,a)], \forall s \in D$, with high probability $1 - \delta$ for some $\delta \in (0,1)$, if the penalizing factor $\alpha$ is sufficiently large. Furthermore, the proposed penalty reduces the average penalty for policy actions compared to the average penalty of CQL.*

## A.1 Proof of Theorem 3.1

Proof of Theorem 3.1 basically follows the proof of Theorem 3.2 in Kumar et al. [21] since $\mathcal{P}_\tau$ multiplies the penalty control factor $f_\tau^{\pi,\hat{\beta}}(s)$ to the penalty of CQL. At each $k$-th iteration, $Q$-function is updated by equation (4), then

$$Q_{k+1}(s,a) \leftarrow \hat{\mathcal{B}}^\pi Q_k(s,a) - \alpha \mathcal{P}_\tau, \ \forall s, a, \tag{A.1}$$

where $\hat{\mathcal{B}}^\pi$ is the estimation of the true Bellman operator $\mathcal{B}^\pi$ based on data samples. It is known that the error between the estimated Bellman operator $\hat{\mathcal{B}}^\pi$ and the true Bellman operator is bounded with high probability of $1 - \delta$ for some $\delta \in (0,1)$ as $|(\mathcal{B}^\pi Q)(s,a) - (\hat{\mathcal{B}}^\pi Q)(s,a)| \leq \xi^\delta(s,a), \ \forall s, a$, where $\xi^\delta$ is a positive constant related to the given dataset $D$, the discount factor $\gamma$, and the transition probability $P$ [21]. Then, with high probability $1 - \delta$,

$$Q_{k+1}(s,a) \leftarrow \mathcal{B}^\pi Q_k(s,a) - \alpha \mathcal{P}_\tau + \xi^\delta(s,a), \ \forall s, a, \tag{A.2}$$

Now, with the state value function $V(s) := \mathbb{E}_{a \sim \pi(\cdot|s)}[Q(s,a)]$

$$V_{k+1}(s) = \mathbb{E}_{a \sim \pi(\cdot|s)}[Q_k(s,a)] = \mathcal{B}^\pi V_k - \alpha \mathbb{E}_{a \sim \pi}[\mathcal{P}_\tau] + \xi^\delta(s,a)$$

$$= \mathcal{B}^\pi V_k(s) - \alpha \mathbb{E}_{a \sim \pi} \left[ f_\tau^{\pi,\hat{\beta}}(s) \cdot \left( \frac{\pi(a|s)}{\hat{\beta}(a|s)} - 1 \right) + \mathbb{E}_{a \sim \pi}[\xi^\delta(s,a)] \right]$$

$$= \mathcal{B}^\pi V_k(s) - \alpha \Delta_{EPQ}^\pi(s) + \mathbb{E}_{a \sim \pi}[\xi^\delta(s,a)] \tag{A.3}$$

Upon repeated iteration, $V_{k+1}$ converges to $V_\infty(s) = V^\pi(s) + (I - \gamma P^\pi)^{-1} \cdot \{-\alpha \Delta_{EPQ}^\pi(s) + \mathbb{E}_{a \sim \pi}[\xi^\delta(s,a)]\}$ based on the fixed point theorem, where $\Delta_{EPQ}^\pi(s) := \mathbb{E}_{a \sim \pi}[\mathcal{P}_\tau]$ is the average penalty for policy $\pi$, $I$ is the identity matrix, and $P^\pi$ is the state transition matrix where the policy $\pi$ is given. Here, we can show that the average penalty $\Delta_{EPQ}^\pi(s)$ is positive as follows:

$$\Delta_{EPQ}^\pi(s) = \mathbb{E}_{a \sim \pi} \left[ f_\tau^{\pi,\hat{\beta}}(s) \cdot \left( \frac{\pi(a|s)}{\hat{\beta}(a|s)} - 1 \right) \right]$$

$$= f_\tau^{\pi,\hat{\beta}}(s) \left[ \sum_{a \in \mathcal{A}} \pi(a|s) \left( \frac{\pi(a|s)}{\hat{\beta}(a|s)} - 1 \right) - \underbrace{\sum_{a \in \mathcal{A}} \hat{\beta}(a|s) \left( \frac{\pi(a|s)}{\hat{\beta}(a|s)} - 1 \right)}_{=0} \right]$$

$$= f_\tau^{\pi,\hat{\beta}}(s) \cdot \sum_{a \in \mathcal{A}} \frac{(\pi(a|s) - \hat{\beta}(a|s))^2}{\hat{\beta}(a|s)} \geq 0, \tag{A.4}$$

where the equality in (A.4) satisfies when $\pi = \hat{\beta}$ or $f_\tau^{\pi,\hat{\beta}} = 0$. Given that $V_{k+1}$ converges to $V_\infty = V^\pi(s) + (I - \gamma P^\pi)^{-1} \cdot \{-\alpha \Delta_{EPQ}^\pi(s) + \mathbb{E}_{a \sim \pi}[\xi^\delta(s,a)]\}$, choosing the penalizing constant $\alpha$ that satisfies $\alpha \geq \max_{s,a \in D}[\xi^\delta(s,a)] \cdot \max_{s \in D}(\Delta_{EPQ}^\pi(s))^{-1}$ will satisfy,

$$- \alpha \cdot \Delta_{EPQ}^\pi(s) + \mathbb{E}_{a \sim \pi}[\xi^\delta(s,a)]$$

$$\leq - \max_{s,a \in D}[\xi^\delta(s,a)] \cdot \underbrace{\max_{s \in D}(\Delta_{EPQ}^\pi(s))^{-1} \cdot \Delta_{EPQ}^\pi(s)}_{\geq 1} + \mathbb{E}_{a \sim \pi}[\xi^\delta(s,a)]$$

$$\leq - \max_{s,a \in D}[\xi^\delta(s,a)] + \mathbb{E}_{a \sim \pi}[\xi^\delta(s,a)] \leq 0, \quad \forall s, \tag{A.5}$$

Since $I - \gamma P^{\pi}$ is non-singular $M$-matrix and the inverse of non-singular $M$-matrix is non-negative, i.e., all elements of $(I - \gamma P^{\pi})^{-1}$ are non-negative, $V_{\infty}(s) = V^{\pi}(s) + (I - \gamma P^{\pi})^{-1} \cdot \{-\alpha \Delta_{EPQ}^{\pi}(s) + \mathbb{E}_{a \sim \pi}[\xi^{\delta}(s,a)]\} \leq V^{\pi}(s), \ \forall s$. Therefore, $V_{\infty}$ underestimates the true value function $V^{\pi}$ if the penalizing constant $\alpha$ satisfies $\alpha \geq \max_{s,a \in D}[\xi^{\delta}(s,a)] \cdot \max_{s \in D}(\Delta_{EPQ}^{\pi}(s))^{-1}$. In addition, according to [21], the average penalty of CQL for policy actions can be represented as $\Delta_{CQL}^{\pi}(s) = \mathbb{E}_{a \sim \pi}[\frac{\pi}{\hat{\beta}} - 1]$. Thus, $\Delta_{EPQ}^{\pi}(s) = f_{\tau}^{\pi,\hat{\beta}}(s)\Delta_{CQL}^{\pi}(s)$ and $f_{\tau}^{\pi,\hat{\beta}}(s) \leq 1$ from the definition in (2), so $0 \leq \Delta_{EPQ}^{\pi}(s) \leq \Delta_{CQL}^{\pi}(s)$. In addition, if $\pi = \hat{\beta}$, then $0 = \Delta_{EPQ}^{\hat{\beta}}(s) = \Delta_{CQL}^{\hat{\beta}}(s)$ from the equality condition in (A.4), which indicates that the average penalty for data actions is 0 for both EPQ and CQL. $\blacksquare$

# B  Implementation Details

In this section, we provide the implementation details of the proposed EPQ. First of all, we provide a detailed derivation of the final $Q$-loss function(4) of EPQ in Section B.1. Next, we introduce a practical implementation of EPQ to compute the loss functions for the parameterized policy and $Q$-function in Section B.2. In addition, to calculate loss functions in Section B.2, we provide the additional implementation details in Appendices B.3, B.4, and B.5. We conduct our experiments on a single server equipped with an Intel Xeon Gold 6336Y CPU and one NVIDIA RTX A5000 GPU, and we compare the running time of EPQ with other baseline algorithms in Section B.6. For additional hyperparameters in the practical implementation of EPQ, we provide detailed hyperparameter setup and additional ablation studies in Appendix C and Appendix D, respectively.

## B.1  Detailed Derivation of $Q$-Loss Function

In Section 3.3, the final $Q$-loss function with the proposed penalty $\mathcal{P}_{\tau,PD} = f_\tau^{\pi,\hat{\beta}}(\frac{\pi}{\hat{\beta}Q} - 1)$ is given by $L(Q) = \frac{1}{2}\mathbb{E}_{s,s'\sim D,a\sim\hat{\beta}Q}\left[(Q - \{\mathcal{B}^\pi Q - \alpha\mathcal{P}_{\tau,PD}\})^2\right]$. In this section, we provide a more detailed calculation of $L(Q)$ to obtain (4) as follows:

$$
\begin{aligned}
L(Q) &= \frac{1}{2}\mathbb{E}_{s,s'\sim D,a\sim\hat{\beta}Q}\left[(Q - \{\mathcal{B}^\pi Q - \alpha\mathcal{P}_{\tau,PD}\})^2\right] \\
&= \frac{1}{2}\mathbb{E}_{s,s'\sim D,a\sim\hat{\beta}Q}\left[(Q - \mathcal{B}^\pi Q)^2\right] + \alpha\mathbb{E}_{s,s'\sim D,a\sim\hat{\beta}Q}\left[\mathcal{P}_{\tau,PD}\cdot Q\right] + C \\
&= \frac{1}{2}\mathbb{E}_{s,s'\sim D,a\sim\hat{\beta}Q}\left[(Q - \mathcal{B}^\pi Q)^2\right] + \alpha\mathbb{E}_{s,s'\sim D,a\sim\hat{\beta}Q}\left[f_\tau^{\pi,\hat{\beta}}\left(\frac{\pi}{\hat{\beta}Q} - 1\right)Q\right] + C \\
&= \frac{1}{2}\mathbb{E}_{s,s'\sim D,a\sim\hat{\beta}Q}\left[(Q - \mathcal{B}^\pi Q)^2\right] + \alpha\mathbb{E}_{s,s'\sim D}\left[\int_{a\in\mathcal{A}}\hat{\beta}Q f_\tau^{\pi,\hat{\beta}}\left(\frac{\pi}{\hat{\beta}Q} - 1\right)Q da\right] + C \\
&= \frac{1}{2}\mathbb{E}_{s,s'\sim D,a\sim\hat{\beta}Q}\left[(Q - \mathcal{B}^\pi Q)^2\right] + \alpha\mathbb{E}_{s,s'\sim D}\left[\int_{a\in\mathcal{A}}f_\tau^{\pi,\hat{\beta}}\left(\pi - \hat{\beta}Q\right)Q da\right] + C \\
&= \frac{1}{2}\mathbb{E}_{s,s'\sim D,a\sim\hat{\beta}Q}\left[(Q - \mathcal{B}^\pi Q)^2\right] + \alpha\mathbb{E}_{s,s'\sim D}\left[\int_{a'\in\mathcal{A}}\pi f_\tau^{\pi,\hat{\beta}}Q da' - \int_{a\in\mathcal{A}}\hat{\beta}Q f_\tau^{\pi,\hat{\beta}}Q da\right] + C \\
&= \frac{1}{2}\mathbb{E}_{s,s'\sim D,a\sim\hat{\beta}Q}\left[(Q - \mathcal{B}^\pi Q)^2\right] + \alpha\mathbb{E}_{s,s'\sim D}\left[\mathbb{E}_{a'\sim\pi}\left[f_\tau^{\pi,\hat{\beta}}Q\right] - \mathbb{E}_{a\sim\hat{\beta}Q}\left[f_\tau^{\pi,\hat{\beta}}Q\right]\right] + C \\
&= \frac{1}{2}\mathbb{E}_{s,s'\sim D,a\sim\hat{\beta}Q}\left[(Q - \mathcal{B}^\pi Q)^2\right] + \alpha\mathbb{E}_{s,s'\sim D,a\sim\hat{\beta}Q}\left[\mathbb{E}_{a'\sim\pi}\left[f_\tau^{\pi,\hat{\beta}}Q\right] - f_\tau^{\pi,\hat{\beta}}Q\right] + C \\
&\underset{(*)}{=} \mathbb{E}_{s,s'\sim D,a\sim\hat{\beta}}\left[\frac{\hat{\beta}Q}{\hat{\beta}}\cdot\left\{\frac{1}{2}(Q - \mathcal{B}^\pi Q)^2 + \alpha f_\tau^{\pi,\hat{\beta}}\cdot(\mathbb{E}_{a'\sim\pi}[Q] - Q)\right\}\right] + C \\
&= \mathbb{E}_{s,s'\sim D,a\sim\hat{\beta},a'\sim\pi}\left[w_{s,a}^Q\cdot\left\{\frac{1}{2}(Q(s,a) - \mathcal{B}^\pi Q(s,a))^2 + \alpha f_\tau^{\pi,\hat{\beta}}(s)(Q(s,a') - Q(s,a))\right\}\right] + C,
\end{aligned}
$$

where $C$ is the remaining constant term that can be ignored for the $Q$-update since $\mathcal{B}^\pi Q$ is the fixed target value. For $(*)$, we apply the IS technique, which states that $\mathbb{E}_{x\sim p}[f(x)] = \mathbb{E}_{x\sim q}\left[\frac{p(x)}{q(x)}f(x)\right]$ for any probability distributions $p$ and $q$, and arbitrary function $f$, and $w_{s,a}^Q = \frac{\hat{\beta}Q(a|s)}{\hat{\beta}(a|s)} = \frac{\exp(Q(s,a))}{\mathbb{E}_{a'\sim\hat{\beta}(\cdot|s)}[\exp(Q(s,a'))]}$ is the importance sampling (IS) ratio between $\hat{\beta}Q$ and $\hat{\beta}$.

## B.2 Practical Implementation for EPQ

Our implementation basically follows the setup of CQL [21]. We use the Gaussian policy $\pi$ with a Tanh$(\cdot)$ layer proposed by Haarnoja et al. [4], and parameterize the policy $\pi$ and $Q$-function using neural network parameters $\phi$ and $\theta$, respectively. Then, we update the policy to maximize $Q_\theta$ with its entropy $\mathcal{H}(\pi_\phi) = \mathbb{E}_{\pi_\phi}[-\log \pi_\phi]$, following the maximum entropy principle [4] as explained in Section 3.3, to account for stochastic policies. Then, we can redefine the policy loss function $L(\pi)$ defined in (5) as the policy loss function $L_\pi(\phi)$ for policy parameter $\phi$, given by

$$L_\pi(\phi) = \mathbb{E}_{s \sim D, \, a \sim \pi_\phi}[-Q_\theta(s,a) + \log \pi_\phi(a|s)]. \tag{B.1}$$

For the $Q$-loss function in (4), we use the IS ratio $w^Q_{s,a}$ in (4) to account for prioritized sampling based on $\hat{\beta}^Q$. However, $\hat{\beta}^Q$ discards samples with low IS weights, which can reduce sample efficiency. To address this, we utilize the clipped IS weight $\max(c_{\min}, w^Q_{s,a})$, where $c_{\min} \in (0,1]$ is the IS clipping constant. This clipped IS weight is multiplied only to the term $(Q(s,a) - \mathcal{B}^\pi Q(s,a))^2$ in (4) to ensure that we can exploit all data samples for $Q$-learning while preserving the proposed penalty. The detailed analysis for $c_{\min}$ is provided in Appendix D. In addition, the optimal policy that maximizes (B.1) follows the Boltzmann distribution, proportional to $\exp(Q_\theta(s,\cdot))$. It has been proven in Kumar et al. [21] that the optimal policy satisfies $\mathbb{E}_{a \sim \pi}[Q_\theta(s,a)] + H(\pi) = \log \sum_{a \sim \mathcal{A}} \exp Q_\theta(s,a)$, so we can replace the $\mathbb{E}_{a' \sim \pi}[Q_\theta(s,a')]$ term in (4) with $\log \sum_{a' \sim \mathcal{A}} \exp Q_\theta(s,a')$, given that $H(\pi)$ does not depend on the $Q$-function. The Bellman operator $\mathcal{B}^\pi$ can be estimated by samples in the dataset as $\mathcal{B}^\pi Q_\theta \approx r(s,a) + \mathbb{E}_{a' \sim \pi} \gamma Q_{\bar{\theta}}(s',a')$, where $\bar{\theta}$ is the parameter of the target $Q$-function. The target network is updated using exponential moving average (EMA) with temperature $\eta_{\bar{\theta}} = 0.005$, as proposed in the deep Q-network (DQN) [52]. Finally, by applying IS clipping and $\log \sum_a \exp Q$ to the $Q$-loss function (4) and redefining it as the value loss function for the value parameter $\theta$, we obtain the following refined value loss function $L_Q(\theta)$ as follows:

$$L_Q(\theta) = \frac{1}{2}\mathbb{E}_{s,a,s' \sim D}\left[\max(c_{\min}, w^Q_{s,a}) \cdot (r(s,a) + \mathbb{E}_{a' \sim \pi}\gamma Q_{\bar{\theta}}(s',a') - Q_\theta(s,a))^2\right] \tag{B.2}$$
$$+ \alpha \mathbb{E}_{s,a \sim D}\left[w^Q_{s,a} f^{\pi,\hat{\beta}}_\tau(s)\left(\log \sum_{a' \in \mathcal{A}} Q_\theta(s,a') - Q_\theta(s,a)\right)\right],$$

where $\hat{\beta}$ is pre-trained by behavior cloning (BC) [18, 53] to compute $f^{\pi,\hat{\beta}}_\tau$. The parameters $\phi$ and $\theta$ are updated to minimize their loss functions $L_\pi(\phi)$ and $L_Q(\theta)$ with learning rate $\eta_\phi$ and $\eta_\theta$, respectively. Detailed implementations for estimating the behavior policy $\hat{\beta}$, the IS weight $w^Q_{s,a}$, and $\log \sum_a \exp Q$ are provided in Appendices B.3, B.4, and B.5, respectively.

## B.3 Behavior Policy Estimation Based on Variational Auto-Encoder

In Section B.2, we estimate the behavior policy $\beta$ that generates the data samples in $D$ necessary for calculating the penalty adaptation factor $f^{\pi,\hat{\beta}}_\tau$ in equation (2). To estimate the behavior policy $\hat{\beta}$, we employ the variational auto-encoder (VAE), one of the most representative variational inference methods, to approximate the underlying distribution of a large dataset based on the variational lower bound [53]. In the context of VAE, we define an encoder model $p_\psi(z|s,a)$ and a decoder model $q_\psi(a|z,s)$ parameterized by $\psi$, where $z$ is the latent variable whose prior distribution $p(z)$ follows the multivariate normal distribution, i.e., $p(z) \sim N(0,I)$. Assuming independence among all data samples, we can derive the variational lower bound for the likelihood of $\beta$ as proposed by Kingma and Welling [53]:

$$\log \beta(a|s) \geq \underbrace{\mathbb{E}_{z \sim p_\psi(\cdot|s,a)}[\log q_\psi(a|z,s)] - D_{KL}(p_\psi(z|s,a)||p(z))}_{\text{the variational lower bound}}, \; \forall s,a \in D \tag{B.3}$$

where $D_{KL}(p||q) = \mathbb{E}_p[\log p - \log q]$ is the Kullback-Leibler (KL) divergence between two distributions $p$ and $q$. In this paper, since we consider the deterministic decoder $q_\psi(z,s)$, the formal term $\mathbb{E}_{z \sim p_\psi(\cdot|s,a)}[\log q_\psi(a|z,s)]$ can be replaced with the mean square error (MSE) as $\mathbb{E}_{z \sim p_\psi(\cdot|s,a)}[\log q_\psi(a|z,s)] \approx \mathbb{E}_{z \sim p_\psi(\cdot|s,a)}[(q_\psi(z,s) - a)^2]$. At each $k$-th iteration, we update the parameter $\psi$ of VAE to maximize the lower bound in equation (B.3). The $\log \beta$ can be estimated using the variational lower-bound in (B.3) to obtain $f^{\pi,\hat{\beta}}_\tau$. The hyperparameter setup for the VAE is provided in Table 2.

Table 2: Hyperparameter setup for VAE

| VAE Hyperparameters | |
|---|---|
| $z$ dimension | 2· state dimension |
| Hidden activation function | ReLU Layer |
| Encoder network $p_\psi$ | $(512, 2 \cdot z$ dim.) <br> (512,512) <br> (state dim. + action dim., 512) |
| Decoder network $q_\psi$ | (512, action dim.) <br> (512,512) <br> ($z$ dim. + state dim., 512) |

## B.4 Implementation of IS Weight $w_{s,a}^Q$

In order to consider the prioritized data distribution $\hat\beta^Q$ proposed in Section 3.3, we use the importance sampling (IS) weight defined by

$$w_{s,a}^Q = \frac{\hat\beta^Q(a|s)}{\hat\beta(a|s)} = \frac{\exp(Q(s,a))}{\mathbb{E}_{a'\sim\hat\beta(\cdot|s)}[\exp(Q(s,a'))]}, \ \forall s,a \in D. \tag{B.4}$$

Since the computation of $\mathbb{E}_{a'\sim\hat\beta(\cdot|s)}$ makes it difficult to know the exact possible action set for state $s$, we approximately estimate the IS weight based on clustering as follows:

$$w_{s,a}^Q = \frac{\exp(Q(s,a))}{\mathbb{E}_{a'\sim\hat\beta(\cdot|s)}[\exp(Q(s,a'))]} \approx \frac{\exp(Q(s,a))}{\frac{1}{|\mathcal{C}_{s,a}|}\sum_{(s',a')\in\mathcal{C}_{s,a}}\exp(Q(s',a'))}, \ \forall s,a \in D. \tag{B.5}$$

Here, $\mathcal{C}_{s,a}$ is the cluster that contains data samples adjacent to $(s,a)$, defined by

$$\mathcal{C}_{s,a} = \{(s',a') \in D \mid ||s - s'||_2 \le \epsilon \cdot \bar{d}_{\text{closest}}\}, \tag{B.6}$$

where the cluster $\mathcal{C}_{s,a}$ can be directly obtained using the nearest neighbor (NN) algorithm [54] provided in the Python library. $\epsilon \cdot \bar{d}_{\text{closest}}$ is the radius of the cluster, and $\bar{d}_{\text{closest}}$ is the average distance between the closest states from each task. In our implementation, we control the radius parameter $\epsilon > 0$ to adjust the number of adjacent samples for the estimation of IS Weight $w_{s,a}^Q$. In addition, using the $Q$-function in the IS weight term makes the learning unstable since the $Q$-function continuously changes as the learning progresses. Thus, instead of the $Q$-function, we use the regularized empirical return $G_t/\zeta$ for each state-action pair obtained by the trajectories stored in $D$, where $\zeta > 0$ is the regularizing temperature. Upon the increase of $\zeta$, the returned difference between adjacent samples in the cluster decreases, so the effect of prioritization can be reduced. The detailed analysis for $\epsilon$ and $\zeta$ is provided in Appendix D.

## B.5  Implementation of $Q$-loss Function

In equation (B.2), the final $Q$-loss function of proposed EPQ is given by

$$L_Q(\theta) = \frac{1}{2}\mathbb{E}_{s,a,s'\sim D}\left[\max(c_{\min}, w_{s,a}^Q)\left(r(s,a) + \mathbb{E}_{a'\sim\pi}\gamma Q_{\bar{\theta}}(s',a') - Q_\theta(s,a)\right)^2\right]$$
$$+ \alpha\mathbb{E}_{s,a\sim D}\left[w_{s,a}^Q f_\tau^{\pi,\hat{\beta}}(s)\left(\log\sum_{a'\in\mathcal{A}}\exp Q_\theta(s,a') - Q_\theta(s,a)\right)\right].$$

Here, we can estimate $\log\sum_a\exp Q(s,a)$ based on the method proposed in CQL [21] as follows:

$$\log\sum_a\exp Q(s,a) = \log\left(\frac{1}{2}\sum_a\pi(a|s)\{\exp(Q(s,a) - \log\pi(a|s))\} + \frac{1}{2}\sum_a\rho_d\{\exp(Q(s,a) - \log\rho_d)\}\right)$$
$$\approx \log\left(\frac{1}{2N_a}\sum_{a_n\sim\pi}^{N_a}(\exp(Q(s,a_n) - \log\pi(a_n|s))) + \frac{1}{2N_a}\sum_{a_n\sim\mathrm{Unif}(\mathcal{A})}^{N_a}(\exp(Q(s,a_n) - \log\rho_d)))\right),$$
$$\text{(B.7)}$$

where $N_a$ is the number of action sampling, $\mathrm{Unif}(\mathcal{A})$ is a Uniform distribution on $\mathcal{A}$, and $\rho_d$ is the density of uniform distribution.

## B.6  Time comparison with other offline RL methods

In this sectrion, we compare the runtime of EPQ with other baseline algorithms: CQL, Onestep, IQL, MCQ, and MISA in Table 3 below. For a fair comparison across all algorithms, we conducted experiments on the Hopper-medium task, which is a popular dataset for comparing computational costs [48, 55], on a single server equipped with an Intel Xeon Gold 6336Y CPU and one NVIDIA RTX A5000 GPU. We measured both epoch runtime during 1,000 gradient steps and score runtime that each algorithm takes to achieve certain normalized scores.

From the epoch runtime results in Table 3, we can observe that EPQ takes approximately 2-30% more runtime per gradient step compared to the CQL baseline. Note that Onestep RL may seem to have very short execution time compared to other algorithms, but one must consider the significantly longer pretraining time required to learn the $Q$-function of behavior policy accurately. Additionally, compared to faster offline RL algorithms such as IQL and MISA, EPQ requires more runtime per step and exhibits a similar runtime to MCQ, another conservative Q-learning algorithm. However, according to the score runtime results in Table 3, we can observe that only proposed EPQ achieves a score of 100 points, while all other algorithms fail to reach this score. Particularly, compared to MCQ, which also considers CQL as a baseline, EPQ achieves the same score with significantly less runtime. Therefore, while EPQ may consume slightly more runtime per gradient step compared to other algorithms, we can conclude that proposed EPQ offers substantial advantages in terms of convergence performance over other algorithms.

Table 3: Runtime comparison: Epoch runtime and Score runtime

| epoch runtime(s) | CQL | Onestep | IQL | MCQ | MISA | EPQ |
|---|---|---|---|---|---|---|
| 1,000 gradient steps | 43.1 | 12.6 | 13.8 | 58.1 | 23.5 | 54.8 |

| score runtime(s) Normalized average return | CQL | Onestep | IQL | MCQ | MISA | EPQ |
|---|---|---|---|---|---|---|
| 60 | 3540.0 | 252.5 | 1600.2 | 31,143.4 | 4,632.7 | 3,232.2 |
| 80 | - | 568.4 | - | 49,359.7 | - | 21,920.0 |
| 100 | - | - | - | - | - | 30,633.2 |

## C   Hyperparameter Setup

The implementation of proposed EPQ basically follows the implementation of the CQL algorithm [21]. First, we provide the details of the shared algorithm hyperparameters in Table 4. In Table 4, we compare the shared algorithm hyperparameters of CQL, the revised version of CQL (revised), and proposed EPQ. CQL (revised) considers the same hyperparameter setup with our algorithm for Adroit tasks since the reproduced performance of CQL (reprod.) using the author-provided hyperparameter setup significantly underperforms compared to the result of CQL (paper) in Table 1.

For the coefficient of entropy term in the policy update (B.1), CQL automatically controls the entropy coefficient so that the entropy of $\pi$ goes to the target entropy, as proposed in Haarnoja et al. [56]. We observe that while the automatic control of policy entropy proves effective for Mujoco tasks, it adversely affects the performance in Adroit tasks since a policy with low entropy can lead to significant overestimation errors in Adroit tasks. Thus, we considered fixed entropy coefficient for Adroit tasks as in Table 4. In addition, CQL controls the penalizing constant $\alpha$ based on Lagrangian method [21] for Adroit tasks, but we also observe that the automatic control of $\alpha$ destabilizes training, leading to poor performance. Therefore, we considered fixed penalizing constant for Adroit tasks in Table 4 for stable learning.

In addition, in Table 5, we provide the details of the task hyperparameters regarding our contributions in the proposed EPQ: the penalty control threshold $\tau$ and the IS clipping factor $c_{\min}$ in the $Q$-loss implementation in (B.2), and the cluster radius $\epsilon$ and regularizing temperature $\zeta$ for the practical implementation of IS clipping factor $w^Q_{s,a}$ in Section B.4. Note that $\rho$ in Table 5 represents the log-density of uniform distribution. For the task hyperparameters, we consider various hyperparameter setups and provide the best hyperparameter setup for all considered tasks in Table 5. The results are based on the ablations studies provided in Section 4.3 and Appendix D.

Table 4: Algorithm hyperparameter setup of CQL, CQL (revised), and EPQ (ours) algorithms

| Hyperparameters | CQL | CQL (revised) (for Adroit) | EPQ |
|---|---|---|---|
| Policy learning rate $\eta_\phi$ | 1e-4 | 1e-4 | 1e-4 |
| Value function learning rate $\eta_\theta$ | 3e-4 | 3e-4 | 3e-4 |
| Soft target update coefficient $\eta_{\bar\theta}$ | 0.005 | 0.005 | 0.005 |
| Batch size | 256 | 256 | 256 |
| The number of sampling $N_a$ | 10 | 10 | 10 |
| Initial behavior cloning steps | 10000 | 10000 | 10000 |
| Gradient steps for training | 3m (0.3m for Adroit) | 0.3m | 3m (0.3m for Adroit) |
| Entropy coefficient $\eta_\theta$ | Auto | 0.5 | Auto (0.5 for Adroit) |
| Penalizing constant $\alpha$ | Auto (10 for MuJoCo) | 5 or 20 | 20 for MuJoCo<br>5 or 20 for Adroit<br>5 or Auto for AntMaze |
| Discount factor $\gamma$ | 0.99 | 0.9 or 0.95 | 0.99 (0.9 or 0.95 for Adroit) |

Table 5: Task hyperparameter setup for Mujoco tasks and Adroit tasks

| Mujoco Tasks | $\tau/\rho$ | $c_{\min}$ | $\epsilon$ | $\zeta$ |
|---|---|---|---|---|
| halfcheetah-random | 10 | 0.2 | 2 | 2 |
| hopper-random | 2 | 0.1 | 0.5 | 2 |
| walker2d-random | 1 | 0.2 | 2 | 0.5 |
| halfcheetah-medium | 10 | 0.2 | 0.5 | 2 |
| hopper-medium | 0.2 | 0.5 | 2 | 5 |
| walker2d-medium | 1 | 0.5 | 2 | 2 |
| halfcheetah-medium-expert | 1.0 | 0.2 | 0.5 | 2 |
| hopper-medium-expert | 1 | 0.2 | 0.5 | 2 |
| walker2d-medium-expert | 1.0 | 0.2 | 0.5 | 2 |
| halfcheetah-expert | 1 | 0.2 | 0.5 | 2 |
| hopper-expert | 1 | 0.2 | 0.5 | 2 |
| walker2d-expert | 0.5 | 0.2 | 2.0 | 2 |
| halfcheetah-medium-replay | 2 | 0.2 | 0.5 | 2 |
| hopper-medium-replay | 2 | 0.2 | 0.5 | 2 |
| walker2d-medium-replay | 0.2 | 0.5 | 1.0 | 2 |
| halfcheetah-full-replay | 1.5 | 0.2 | 0.5 | 2 |
| hopper-full-replay | 2.0 | 0.2 | 1.0 | 2 |
| walker2d-full-replay | 1.0 | 0.2 | 0.5 | 2 |
| **Adroit Tasks** | $\tau/\rho$ | $c_{\min}$ | $\epsilon$ | $\zeta$ |
| pen-human | 0.05 | 0.5 | 1.0 | 200 |
| door-human | 0.05 | 0.5 | 0.5 | 200 |
| hammer-human | 0.1 | 0.2 | 5 | 100 |
| relocate-human | 0.2 | 0.2 | 2 | 10 |
| pen-cloned | 0.2 | 0.2 | 5 | 50 |
| door-cloned | 0.2 | 0.5 | 1 | 10 |
| hammer-cloned | 0.2 | 0.2 | 5 | 100 |
| relocate-cloned | 0.2 | 0.2 | 5 | 10 |
| **AntMaze Tasks** | $\tau/\rho$ | $c_{\min}$ | $\epsilon$ | $\zeta$ |
| umaze | 10 | 0.2 | 2 | 2 |
| umaze-diverse | 10 | 0.2 | 2 | 2 |
| medium-play | 0.1 | 0.2 | 1 | 2 |
| medium-diverse | 0.1 | 0.2 | 1 | 2 |
| large-play | 0.1 | 0.2 | 1 | 2 |
| large-diverse | 0.1 | 0.2 | 1 | 2 |

# D  Additional Ablation Studies Related to $w_{s,a}^Q$ Estimation

In this section, we provide additional ablation studies related to IS weight $w_{s,a}^Q$ estimation in Appendix B. For analysis, Fig. 8 shows the performance plot when the IS clipping factor $c_{\min}$, the cluster radius $\epsilon$, and the temperature $\zeta$ change.

(a) IS clipping factor $c_{\min}$ &nbsp;&nbsp;&nbsp;&nbsp; (b) Cluster radius $\epsilon$ &nbsp;&nbsp;&nbsp;&nbsp; (c) Temperature $\zeta$
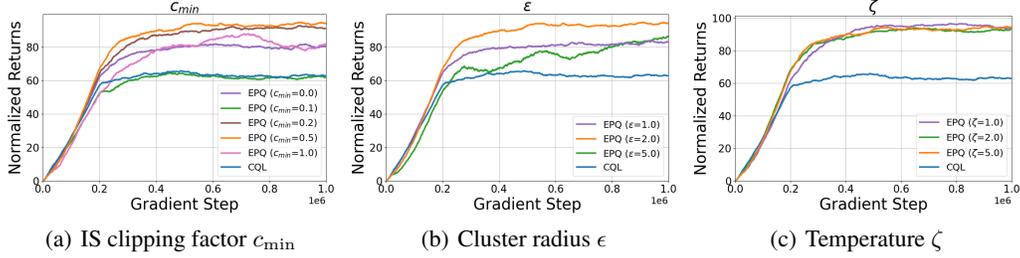
Figure 8: Additional ablation studies on Hopper-medium task

**IS Clipping Factor $c_{\min}$:** In the EPQ implementation, the IS clipping factor $c_{\min}$ is employed to clip the IS weight $w_{s,a}^Q$ to prevent the exclusion of data samples with relatively low $w_{s,a}^Q$. When $c_{\min} = 0$, low-quality samples with low $w_{s,a}^Q$ are not utilized at all based on the prioritization in Section 3.3. However, as $c_{\min}$ increases, these low-quality samples are increasingly exploited. Fig. 7(c) illustrates the performance of EPQ with varying $c_{\min}$, and EPQ achieves the best performance when $c_{\min} = 0.5$. This result suggests that it is more beneficial to use low-quality samples with proper priority rather than discarding them entirely.

**Cluster Radius $\epsilon$:** As explained in Appendix B.4, we can control the number of adjacent samples in the cluster based on the radius $\epsilon$. From the results illustrated in Fig. 8(a), we can observe that EPQ with $d = 2.0$ performs best, and a decrease or an increase in $\epsilon$ can significantly affect the performance indicating that $\epsilon$ must be chosen properly for each task to find the cluster that contains adjacent samples appropriately. If $\epsilon$ is too small, the cluster will hardly contain adjacent samples, and if $\epsilon$ is too large, samples that should be distinguished will aggregate in the same cluster, adversely affecting the performance.

**Temperature $\zeta$:** As proposed in Section 3.3, samples in the dataset are prioritized according to the definition of $w_{s,a}^Q$. Since the samples with higher $Q$ values are more likely to be selected for the update of the $Q$-function, temperature $\zeta$ controls the amount of prioritization, as explained in Appendix B.4. Increasing $\zeta$ reduces the difference in the $Q$-function between the samples, putting less emphasis on prioritization. Fig. 8(b) shows the performance change according to the change in $\zeta$, where the results state that the performance does not heavily depend on $\zeta$. From the ablation study, we can conclude that the radius $\epsilon$ has a greater influence on the performance of Hopper-medium task compared to the temperature $\zeta$.

# E   Additional Performance Comparison on Adroit Tasks

For adroit tasks, the performance of CQL (reprod.) is too low compared to CQL (paper) in Table 1, so we additionally provide the performance result of the revised version of CQL provided in Section C. We also compare the performance of EPQ with the performance of CQL (revised) on various adroit tasks, and Table 6 shows the corresponding comparison results. From the result, we can see that CQL (revised) greatly enhances the performance of CQL on adroit tasks, but EPQ still outperforms CQL (revised), which demonstrates the intact advantage of the proposed exclusive penalty and prioritized dataset well on the adroit tasks.

Table 6: Performance comparison of CQL (paper), CQL (revised), and EPQ (ours) on Adroit tasks.

| Task | CQL (paper) | CQL (revised) | EPQ |
|---|---|---|---|
| pen-human | 55.8 | 82.0±6.2 | **83.9±6.8** |
| door-human | 9.1 | 7.8±0.5 | **13.2 ± 2.4** |
| hammer-human | 2.1 | **6.4±5.4** | 3.9±5.0 |
| relocate-human | **0.4** | 0.1±0.2 | **0.3±0.2** |
| pen-cloned | 40.3 | **90.7±4.8** | **91.8±4.7** |
| door-cloned | 3.5 | 1.3±2.2 | **5.8±2.8** |
| hammer-cloned | 5.7 | 2.0±1.3 | **22.8±15.3** |
| relocate-cloned | -0.1 | 0.0±0.0 | **0.1±0.1** |
| **Adroit Tasks Total** | 116.8 | 190.3 | **221.8** |

# F   Limitations

The proposed EPQ significantly improves performance over the existing CQL baseline on various D4RL tasks, but there are many hyperparamaters that need to be optimized. We newly consider the penalty control threshold $\tau$, IS clipping factor $c_{\min}$, the cluster radius $\epsilon$, and the regularizing temperature $\zeta$. Therefore, in order for the proposed EPQ to perform well, it is necessary to find the optimal performance by considering various hyperparameter setup, which may require some interaction with the environment.

# G   Broader Impact

Nevertheless, in real-world situations, engaging with the environment can be costly. Particularly in high-risk contexts such as disaster scenarios, acquiring adequate data for learning can be quite challenging. Our research is primarily focused on offline settings and we present a novel method, EPQ, holds the potential for practical applications in real-life situations where the interaction is not available, and exhibits promise in addressing the challenges posed by offline RL algorithms. Consequently, our work carries several potential societal implications, although we believe that none require specific emphasis in this context.

# NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

**The checklist answers are an integral part of your paper submission.** They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS paper checklist",**
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims made in the abstract and introductions are well reflected in Section 3 Methodology and Section 4 Experiments.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

Justification: The limitations are addressed in the Appendix F Limitations.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: For each theoretical result, the detailed proofs and assumptions are provided in Appendix A Proof and Appendix B Implementation Details.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The specific environment descriptions and experimental setups including the hyperparameters can be found in Section 4 Experiments and Appendix C Hyperparameter Setup.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The data and code for reproducing the main experimental results are included in supplemental materials.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.

- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: The specific experimental setups including the hyperparameters can be found in Section 4 Experiments and Appendix C Hyperparameter Setup.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: The graphs included in the paper such as Figure 6 and Figure 7 in Section 4 Experiments well demonstrate the statistical significance of the experiment.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: The information on computation resources are provided in Appendix B Implementation Details.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

Answer: [Yes]

Justification: The research conducted in the paper conforms the NeurIIPS Code of Ethics

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The societal impacts of the proposed paper is included in appendix G Broader Impacts section.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The proposed paper does not pose such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The baseline code and experimental data are cited both in-text and in the References section.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: The proposed paper does not release new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The proposed paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.