A THE FULL PROCEDURE OF BATTLE

The Combined Policy. In order to address the inefficiency caused by the discrepancy between π_θ and $\pi_{\nu\circ\alpha}$ in the state distribution, we propose a strategy to construct the behavior policy π for data collection in our practical implementation. Inspired by Branched rollout (Janner et al., 2019), we combine the intention policy π_θ with the perturbed policy $\pi_{\nu\circ\alpha}$. Specifically, we define $\pi^{1:h} = \pi^{1:h}_{\nu\circ\alpha}$ and $\pi^{h+1:H} = \pi^{h+1:H}_{\theta}$, where h is sampled from a uniform distribution $U(0, H)$ and H represents the task horizon. The resulting combined policy π is responsible for data collection, which is then stored in the replay buffer during the learning process.

We present the detailed procedures of our proposed method in Algorithm 1. Our method, referred to as BATTLE, is built upon the well-established preference-based RL algorithm PEBBLE (Lee et al., 2021a).

Algorithm 1 BATTLE

Input: a fixed victim policy π_ν, frequency of human feedback K, outer loss updating frequency M, task horizon H

Initialize parameters of Q_ϕ, π_θ, $\widehat{r}_\psi$, π_α and h_ω
Initialize $\mathcal{B}$ and π_θ with unsupervised exploration
Initialize preference data set $\mathcal{D} \leftarrow \emptyset$
for each iteration **do**
  `// Construct the combined policy` π
  if episode is done **then**
    $h \sim U(0, H)$
    $\pi^{1:h} = \pi^{1:h}_{\nu\circ\alpha}$ and $\pi^{h+1:H} = \pi^{h+1:H}_{\theta}$
  end if
  Take action $a_t \sim \pi$ and collect s_{t+1}
  Store transition into dataset $\mathcal{B} \leftarrow \mathcal{B} \cup \{(s_t, a_t, \widehat{r}_\psi(s_t), s_{t+1})\}$
  `// Query preference and Reward learning`
  if iteration $\%$ $K == 0$ **then**
    for each query step **do**
      Sample pair of trajectories (σ^0, σ^1)
      Query preference y from manipulator
      Store preference data into dataset $\mathcal{D} \leftarrow \mathcal{D} \cup \{(\sigma^0, \sigma^1, y)\}$
    end for
    for each gradient step **do**
      Sample batch $\{(\sigma^0, \sigma^1, y)_i\}_{i=1}^n$ from $\mathcal{D}$
      Optimize (2) to update $\widehat{r}_\psi$
    end for
  end if
  `// Inner loss optimization`
  for each gradient step **do**
    Sample random mini-batch transitions from $\mathcal{B}$
    Optimize π_α: minimize (6) with respect to α
  end for
  `// Outer loss optimization`
  if iteration $\%$ $M == 0$ **then**
    Sample random mini-batch transitions from $\mathcal{B}$
    Optimize h_ω: minimize (7) with respect to ω
  end if
  `// Intention policy learning`
  Update Q_ϕ and π_θ according to (3) and (4), respectively.
end for

Output: adversarial policy π_α

B Derivation of the Gradient of the Outer-level Loss

In this section, we present detailed derivation of the gradient of the outer loss J_π with respect to the parameters of the weighting function ω. According to the chain rule, we can derive that

$$\begin{aligned}
&\nabla_\omega J_\pi(\hat{\alpha}(\omega))|_{\omega_t} \\
=&\frac{\partial J_\pi(\hat{\alpha}(\omega))}{\partial \hat{\alpha}(\omega)}\Big|_{\hat{\alpha}_t} \frac{\partial \hat{\alpha}_t(\omega)}{\partial \omega}\Big|_{\omega_t} \\
=&\frac{\partial J_\pi(\hat{\alpha}(\omega))}{\partial \hat{\alpha}(\omega)}\Big|_{\hat{\alpha}_t} \frac{\partial \hat{\alpha}_t(\omega)}{\partial h(\mathbf{s};\omega)}\Big|_{\omega_t} \frac{\partial h(\mathbf{s};\omega)}{\partial \omega}\Big|_{\omega_t} \\
=&-\eta_t \frac{\partial J_\pi(\hat{\alpha}(\omega))}{\partial \hat{\alpha}(\omega)}\Big|_{\hat{\alpha}_t} \sum_{\mathbf{s}\sim\mathcal{B}} \frac{\partial D_{\mathrm{KL}}\left(\pi_{\nu\circ\alpha}(\mathbf{s}) \parallel \pi_\theta(\mathbf{s})\right)}{\partial \alpha}\Big|_{\alpha_t} \frac{\partial h(\mathbf{s};\omega)}{\partial \omega}\Big|_{\omega_t} \\
=&-\eta_t \sum_{\mathbf{s}\sim\mathcal{B}} \left(\frac{\partial J_\pi(\hat{\alpha}(\omega))}{\partial \hat{\alpha}(\omega)}\Big|_{\hat{\alpha}_t}^\top \frac{\partial D_{\mathrm{KL}}\left(\pi_{\nu\circ\alpha}(\mathbf{s}) \parallel \pi_\theta(\mathbf{s})\right)}{\partial \alpha}\Big|_{\alpha_t} \right) \frac{\partial h(\mathbf{s};\omega)}{\partial \omega}\Big|_{\omega_t}.
\end{aligned} \tag{10}$$

For brevity of expression, we let:

$$f(\mathbf{s}) = \frac{\partial J_\pi(\hat{\alpha}(\omega))}{\partial \hat{\alpha}(\omega)}\Big|_{\hat{\alpha}_t}^\top \frac{\partial D_{\mathrm{KL}}\left(\pi_{\nu\circ\alpha}(\mathbf{s}) \parallel \pi_\theta(\mathbf{s})\right)}{\partial \hat{\alpha}}\Big|_{\alpha_t}. \tag{11}$$

The gradient of outer-level optimization loss with respect to parameters ω is:

$$\nabla_\omega J_\pi(\hat{\alpha}(\omega))|_{\omega_t} = -\eta_t \sum_{\mathbf{s}\sim\mathcal{B}} f(\mathbf{s}) \cdot \frac{\partial h(\mathbf{s};\omega)}{\partial \omega}\Big|_{\omega_t}. \tag{12}$$

C Connection between RSA-MDP and MDP

Lemma C.1. *Given a RSA-MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{B}, \widehat{\mathcal{R}}, \mathcal{P}, \gamma)$ and a fixed victim policy π_ν, there exists a MDP $\hat{\mathcal{M}} = (\mathcal{S}, \hat{\mathcal{A}}, \widehat{\mathcal{R}}, \widehat{\mathcal{P}}, \gamma)$ such that the optimal policy of $\hat{\mathcal{M}}$ is equivalent to the optimal adversary π_α in RSA-MDP given a fixed victim, where $\widehat{\mathcal{A}} = \mathcal{S}$ and*

$$\widehat{\mathcal{P}}(\mathbf{s}'|\mathbf{s}, \widehat{\mathbf{a}}) = \sum_{\mathbf{a}\in\mathcal{A}} \pi_\nu(\mathbf{a}|\widehat{\mathbf{a}}) \mathcal{P}(\mathbf{s}'|\mathbf{s}, \mathbf{a}) \quad \textit{for } \mathbf{s}, \mathbf{s}' \in \mathcal{S} \textit{ and } \widehat{\mathbf{a}} \in \widehat{\mathcal{A}}.$$

D Theoretical Analysis and Proofs

D.1 Theorem 1: Convergence Rate of the Outer Loss

Lemma D.1. *(Lemma 1.2.3 in Nesterov (1998)) If function $f(x)$ is Lipschitz smooth on $\mathbb{R}^n$ with constant L, then $\forall x, y \in \mathbb{R}^n$, we have*

$$\left|f(y) - f(x) - f'(x)^\top (y - x)\right| \leq \frac{L}{2} \|y - x\|^2. \tag{13}$$

Proof. $\forall x, y \in \mathbb{R}^n$, we have

$$\begin{aligned}
f(y) &= f(x) + \int_0^1 f'(x + \tau(y - x))^\top (y - x) d\tau \\
&= f(x) + f'(x)^\top (y - x) + \int_0^1 [f'(x + \tau(y - x)) - f'(x)]^\top (y - x) d\tau.
\end{aligned} \tag{14}$$

Then we can derive that

$$\begin{aligned}
\left|f(y) - f(x) - f'(x)^\top (y - x)\right| &= \left| \int_0^1 [f'(x + \tau(y - x)) - f'(x)]^\top (y - x) d\tau \right| \\
&\leq \int_0^1 \left| [f'(x + \tau(y - x)) - f'(x)]^\top (y - x) \right| d\tau \\
&\leq \int_0^1 \|f'(x + \tau(y - x)) - f'(x)\| \cdot \|y - x\| \, d\tau \\
&\leq \int_0^1 \tau L \|y - x\|^2 \, d\tau = \frac{L}{2} \|y - x\|^2,
\end{aligned} \tag{15}$$

where the first inequality holds for $\left|\int_a^b f(x)dx\right| \leq \int_a^b |f(x)|\,dx$, the second inequality holds for Cauchy-Schwarz inequality, and the last inequality holds for the definition of Lipschitz smoothness. □

Theorem D.2. *Suppose J_π is Lipschitz-smooth with constant L, the gradient of J_π and $\mathcal{L}_{att}$ is bounded by ρ. Let the training iterations be T, the inner-level optimization learning rate $\eta_t = \min\{1, \frac{c_1}{T}\}$ for some constant $c_1 > 0$ where $\frac{c_1}{T} < 1$. Let the outer-level optimization learning rate $\beta_t = \min\{\frac{1}{L}, \frac{c_2}{\sqrt{T}}\}$ for some constant $c_2 > 0$ where $c_2 \leq \frac{\sqrt{T}}{L}$, and $\sum_{t=1}^{\infty} \beta_t \leq \infty, \sum_{t=1}^{\infty} \beta_t^2 \leq \infty$. The convergence rate of J_π achieves*

$$\min_{1\leq t\leq T} \mathbb{E}\left[\left\|\nabla_\omega J_\pi(\alpha_{t+1}(\omega_t))\right\|^2\right] \leq \mathcal{O}\left(\frac{1}{\sqrt{T}}\right). \tag{16}$$

Proof. First,

$$\begin{aligned} &J_\pi(\hat{\alpha}_{t+2}(\omega_{t+1})) - J_\pi(\hat{\alpha}_{t+1}(\omega_t)) \\ =&\left\{J_\pi(\hat{\alpha}_{t+2}(\omega_{t+1})) - J_\pi(\hat{\alpha}_{t+1}(\omega_{t+1}))\right\} + \left\{J_\pi(\hat{\alpha}_{t+1}(\omega_{t+1})) - J_\pi(\hat{\alpha}_{t+1}(\omega_t))\right\}. \end{aligned} \tag{17}$$

Then we separately derive the two terms of (17). For the first term,

$$\begin{aligned} &J_\pi(\hat{\alpha}_{t+2}(\omega_{t+1})) - J_\pi(\hat{\alpha}_{t+1}(\omega_{t+1})) \\ \leq&\nabla_{\hat{\alpha}} J_\pi(\hat{\alpha}_{t+1}(\omega_{t+1}))^\top (\hat{\alpha}_{t+2}(\omega_{t+1}) - \hat{\alpha}_{t+1}(\omega_{t+1})) + \frac{L}{2}\left\|\hat{\alpha}_{t+2}(\omega_{t+1}) - \hat{\alpha}_{t+1}(\omega_{t+1})\right\|^2 \\ \leq&\left\|\nabla_{\hat{\alpha}} J_\pi(\hat{\alpha}_{t+1}(\omega_{t+1}))\right\| \cdot \left\|\hat{\alpha}_{t+2}(\omega_{t+1}) - \hat{\alpha}_{t+1}(\omega_{t+1})\right\| + \frac{L}{2}\left\|\hat{\alpha}_{t+2}(\omega_{t+1}) - \hat{\alpha}_{t+1}(\omega_{t+1})\right\|^2 \\ \leq&\rho \cdot \left\|-\eta_{t+1}\nabla_{\hat{\alpha}}\mathcal{L}_{\text{att}}(\hat{\alpha}_{t+1})\right\| + \frac{L}{2}\left\|-\eta_{t+1}\nabla_{\hat{\alpha}}\mathcal{L}_{\text{att}}(\hat{\alpha}_{t+1})\right\|^2 \\ \leq&\eta_{t+1}\rho^2 + \frac{L}{2}\eta_{t+1}^2\rho^2, \end{aligned} \tag{18}$$

where $\hat{\alpha}_{t+2}(\omega_{t+1}) - \hat{\alpha}_{t+1}(\omega_{t+1}) = -\eta_{t+1}\nabla_{\hat{\alpha}}\mathcal{L}_{\text{att}}(\hat{\alpha}_{t+1})$, the first inequality holds for Lemma D.1, the second inequality holds for Cauchy-Schwarz inequality, the third inequality holds for $\|\nabla_{\hat{\alpha}} J_\pi(\hat{\alpha}_{t+1}(\omega_{t+1}))\| \leq \rho$, and the last inequality holds for $\|\nabla_{\hat{\alpha}}\mathcal{L}_{\text{att}}(\hat{\alpha}_{t+1})\| \leq \rho$. It can be proved that the gradient of ω with respect to J_π is Lipschitz continuous and we assume the Lipschitz constant is L. Therefore, for the second term,

$$\begin{aligned} &J_\pi(\hat{\alpha}_{t+1}(\omega_{t+1})) - J_\pi(\hat{\alpha}_{t+1}(\omega_t)) \\ \leq&\nabla_\omega J_\pi(\hat{\alpha}_{t+1}(\omega_t))^\top (\omega_{t+1} - \omega_t) + \frac{L}{2}\left\|\omega_{t+1} - \omega_t\right\|^2 \\ =&-\beta_t \nabla_\omega J_\pi(\hat{\alpha}_{t+1}(\omega_t))^\top \nabla_\omega J_\pi(\hat{\alpha}_{t+1}(\omega_t)) + \frac{L\beta_t^2}{2}\left\|\nabla_\omega J_\pi(\hat{\alpha}_{t+1}(\omega_t))\right\|^2 \\ =&-(\beta_t - \frac{L\beta_t^2}{2})\left\|\nabla_\omega J_\pi(\hat{\alpha}_{t+1}(\omega_t))\right\|^2, \end{aligned} \tag{19}$$

where $\omega_{t+1} - \omega_t = -\beta_t \nabla_\omega J_\pi(\hat{\alpha}_{t+1}(\omega_t))$, and the first inequality holds for Lemma D.1. Therefore, (17) becomes

$$J_\pi(\hat{\alpha}_{t+2}(\omega_{t+1})) - J_\pi(\hat{\alpha}_{t+1}(\omega_t)) \leq \eta_{t+1}\rho^2 + \frac{L}{2}\eta_{t+1}^2\rho^2 - (\beta_t - \frac{L\beta_t^2}{2})\left\|\nabla_\omega J_\pi(\hat{\alpha}_{t+1}(\omega_t))\right\|^2. \tag{20}$$

Rearranging the terms of (20), we obtain

$$(\beta_t - \frac{L\beta_t^2}{2})\left\|\nabla_\omega J_\pi(\hat{\alpha}_{t+1}(\omega_t))\right\|^2 \leq J_\pi(\hat{\alpha}_{t+1}(\omega_t)) - J_\pi(\hat{\alpha}_{t+2}(\omega_{t+1})) + \eta_{t+1}\rho^2 + \frac{L}{2}\eta_{t+1}^2\rho^2. \tag{21}$$

Then, we sum up both sides of (21),

$$\begin{aligned}
&\sum_{t=1}^{T}(\beta_t - \frac{L\beta_t^2}{2}) \left\|\nabla_\omega J_\pi(\hat{\alpha}_{t+1}(\omega_t))\right\|^2 \\
\leq & J_\pi(\hat{\alpha}_2(\omega_1)) - J_\pi(\hat{\alpha}_{T+2}(\omega_{T+1})) + \sum_{t=1}^{T}(\eta_{t+1}\rho^2 + \frac{L}{2}\eta_{t+1}^2\rho^2) \\
\leq & J_\pi(\hat{\alpha}_2(\omega_1)) + \sum_{t=1}^{T}(\eta_{t+1}\rho^2 + \frac{L}{2}\eta_{t+1}^2\rho^2).
\end{aligned} \tag{22}$$

Therefore,

$$\begin{aligned}
&\min_{1\leq t\leq T} \mathbb{E}\left[\left\|\nabla_\omega J_\pi(\hat{\alpha}_{t+1}(\omega_t))\right\|^2\right] \\
\leq & \frac{\sum_{t=1}^{T}(\beta_t - \frac{L\beta_t^2}{2}) \left\|\nabla_\omega J_\pi(\hat{\alpha}_{t+1}(\omega_t))\right\|^2}{\sum_{t=1}^{T}(\beta_t - \frac{L\beta_t^2}{2})} \\
\leq & \frac{1}{\sum_{t=1}^{T}(2\beta_t - L\beta_t^2)}\left[2J_\pi(\hat{\alpha}_2(\omega_1)) + \sum_{t=1}^{T}(2\eta_{t+1}\rho^2 + L\eta_{t+1}^2\rho^2)\right] \\
\leq & \frac{1}{\sum_{t=1}^{T}\beta_t}\left[2J_\pi(\hat{\alpha}_2(\omega_1)) + \sum_{t=1}^{T}\eta_{t+1}\rho^2(2 + L\eta_{t+1})\right] \\
\leq & \frac{1}{T\beta_t}\left[2J_\pi(\hat{\alpha}_2(\omega_1)) + T\eta_{t+1}\rho^2(2+L)\right] \\
= & \frac{2J_\pi(\hat{\alpha}_2(\omega_1))}{T\beta_t} + \frac{\eta_{t+1}\rho^2(2+L)}{\beta_t} \\
= & \frac{2J_\pi(\hat{\alpha}_2(\omega_1))}{T}\max\{L, \frac{\sqrt{T}}{c_2}\} + \min\{1, \frac{c_1}{T}\}\max\{L, \frac{\sqrt{T}}{c_2}\}\rho^2(2+L) \\
\leq & \frac{2J_\pi(\hat{\alpha}_2(\omega_1))}{c_2\sqrt{T}} + \frac{c_1\rho^2(2+L)}{c_2\sqrt{T}} \\
= & \mathcal{O}\left(\frac{1}{\sqrt{T}}\right),
\end{aligned} \tag{23}$$

where the second inequality holds according to (22), the third inequality holds for $\sum_{t=1}^{T}\left(2\beta_t - L\beta_t^2\right) \geq \sum_{t=1}^{T}\beta_t$. □

D.2 Theorem 2: Convergence of the Inner Loss

Lemma D.3. *(Lemma A.5 in Mairal (2013)) Let $(a_n)_{n\geq 1}, (b_n)_{n\geq 1}$ be two non-negative real sequences such that the series $\sum_{n=1}^{\infty} a_n$ diverges, the series $\sum_{n=1}^{\infty} a_n b_n$ converges, and there exists $C > 0$ such that $|b_{n+1} - b_n| \leq C a_n$. Then, the sequence $(b_n)_{n\geq 1}$ converges to 0.*

Theorem D.4. *Suppose J_π is Lipschitz-smooth with constant L, the gradient of J_π and $\mathcal{L}_{att}$ is bounded by ρ. Let the training iterations be T, the inner-level optimization learning rate $\eta_t = \min\{1, \frac{c_1}{T}\}$ for some constant $c_1 > 0$ where $\frac{c_1}{T} < 1$. Let the outer-level optimization learning rate $\beta_t = \min\{\frac{1}{L}, \frac{c_2}{\sqrt{T}}\}$ for some constant $c_2 > 0$ where $c_2 \leq \frac{\sqrt{T}}{L}$, and $\sum_{t=1}^{\infty}\beta_t \leq \infty, \sum_{t=1}^{\infty}\beta_t^2 \leq \infty$. $\mathcal{L}_{att}$ achieves*

$$\lim_{t\to\infty} \mathbb{E}\left[\left\|\nabla_\alpha \mathcal{L}_{att}(\alpha_t; \omega_t)\right\|^2\right] = 0. \tag{24}$$

Proof. First,

$$\begin{aligned}
&\mathcal{L}_{\text{att}}(\alpha_{t+1}; \omega_{t+1}) - \mathcal{L}_{\text{att}}(\alpha_t; \omega_t) \\
= & \left\{\mathcal{L}_{\text{att}}(\alpha_{t+1}; \omega_{t+1}) - \mathcal{L}_{\text{att}}(\alpha_{t+1}; \omega_t)\right\} + \left\{\mathcal{L}_{\text{att}}(\alpha_{t+1}; \omega_t) - \mathcal{L}_{\text{att}}(\alpha_t; \omega_t)\right\}.
\end{aligned} \tag{25}$$

For the first term in (25),

$$\begin{aligned}
&\mathcal{L}_{\text{att}}(\alpha_{t+1};\omega_{t+1}) - \mathcal{L}_{\text{att}}(\alpha_{t+1};\omega_t) \\
\leq &\nabla_\omega \mathcal{L}_{\text{att}}(\alpha_{t+1};\omega_t)^\top(\omega_{t+1}-\omega_t) + \frac{L}{2}\left\|\omega_{t+1}-\omega_t\right\|^2 \\
= &-\beta_t \nabla_\omega \mathcal{L}_{\text{att}}(\alpha_{t+1};\omega_t)^\top \nabla_\omega J_\pi(\alpha_{t+1}(\omega_t)) + \frac{L\beta_t^2}{2}\left\|\nabla_\omega J_\pi(\alpha_{t+1}(\omega_t))\right\|^2 .
\end{aligned} \tag{26}$$

where $\omega_{t+1}-\omega_t = -\beta_t \nabla_\omega J_\pi(\alpha_{t+1}(\omega_t))$, and the first inequality holds according to Lemma D.1. For the second term in (25),

$$\begin{aligned}
&\mathcal{L}_{\text{att}}(\alpha_{t+1};\omega_t) - \mathcal{L}_{\text{att}}(\alpha_t;\omega_t) \\
\leq &\nabla_\alpha \mathcal{L}_{\text{att}}(\alpha_t;\omega_t)^\top(\alpha_{t+1}-\alpha_t) + \frac{L}{2}\left\|\alpha_{t+1}-\alpha_t\right\|^2 \\
= &-\eta_t \nabla_\alpha \mathcal{L}_{\text{att}}(\alpha_t;\omega_t)^\top \nabla_\alpha \mathcal{L}_{\text{att}}(\alpha_t;\omega_t) + \frac{L\eta_t^2}{2}\left\|\nabla_\alpha \mathcal{L}_{\text{att}}(\alpha_t;\omega_t)\right\|^2 \\
= &-(\eta_t - \frac{L\eta_t^2}{2})\left\|\nabla_\alpha \mathcal{L}_{\text{att}}(\alpha_t;\omega_t)\right\|^2 .
\end{aligned} \tag{27}$$

where $\alpha_{t+1}-\alpha_t = -\eta_t \nabla_\alpha \mathcal{L}_{\text{att}}(\alpha_t;\omega_t)$, and the first inequality holds according to Lemma (D.1). Therefore, (25) becomes

$$\begin{aligned}
&\mathcal{L}_{\text{att}}(\alpha_{t+1};\omega_{t+1}) - \mathcal{L}_{\text{att}}(\alpha_t;\omega_t) \\
\leq &-\beta_t \nabla_\omega \mathcal{L}_{\text{att}}(\alpha_{t+1};\omega_t)^\top \nabla_\omega J_\pi(\alpha_{t+1}(\omega_t)) + \frac{L\beta_t^2}{2}\left\|\nabla_\omega J_\pi(\alpha_{t+1}(\omega_t))\right\|^2 \\
&-(\eta_t - \frac{L\eta_t^2}{2})\left\|\nabla_\alpha \mathcal{L}_{\text{att}}(\alpha_t;\omega_t)\right\|^2 .
\end{aligned} \tag{28}$$

Taking expectation of both sides of (28) and rearranging the terms, we obtain

$$\begin{aligned}
&\eta_t \mathbb{E}\left[\left\|\nabla_\alpha \mathcal{L}_{\text{att}}(\alpha_t;\omega_t)\right\|^2\right] + \beta_t \mathbb{E}\left[\left\|\nabla_\omega \mathcal{L}_{\text{att}}(\alpha_{t+1};\omega_t)\right\| \cdot \left\|\nabla_\omega J_\pi(\alpha_{t+1}(\omega_t))\right\|\right] \\
\leq &\mathbb{E}\left[\mathcal{L}_{\text{att}}(\alpha_t;\omega_t)\right] - \mathbb{E}\left[\mathcal{L}_{\text{att}}(\alpha_{t+1};\omega_{t+1})\right] + \frac{L\beta_t^2}{2}\mathbb{E}\left[\left\|\nabla_\omega J_\pi(\alpha_{t+1}(\omega_t))\right\|^2\right] \\
&+ \frac{L\eta_t^2}{2}\mathbb{E}\left[\left\|\nabla_\alpha \mathcal{L}_{\text{att}}(\alpha_t;\omega_t)\right\|^2\right] .
\end{aligned} \tag{29}$$

Summing up both sides of (29) from $t=1$ to ∞,

$$\begin{aligned}
&\sum_{t=1}^{\infty} \eta_t \mathbb{E}\left[\left\|\nabla_\alpha \mathcal{L}_{\text{att}}(\alpha_t;\omega_t)\right\|^2\right] + \sum_{t=1}^{\infty} \beta_t \mathbb{E}\left[\left\|\nabla_\omega \mathcal{L}_{\text{att}}(\alpha_{t+1};\omega_t)\right\| \cdot \left\|\nabla_\omega J_\pi(\alpha_{t+1}(\omega_t))\right\|\right] \\
\leq &\mathbb{E}\left[\mathcal{L}_{\text{att}}(\alpha_1;\omega_1)\right] - \lim_{t\to\infty}\mathbb{E}\left[\mathcal{L}_{\text{att}}(\alpha_{t+1};\omega_{t+1})\right] + \sum_{t=1}^{\infty}\frac{L\beta_t^2}{2}\mathbb{E}\left[\left\|\nabla_\omega J_\pi(\alpha_{t+1}(\omega_t))\right\|^2\right] \\
&+ \sum_{t=1}^{\infty}\frac{L\eta_t^2}{2}\mathbb{E}\left[\left\|\nabla_\alpha \mathcal{L}_{\text{att}}(\alpha_t;\omega_t)\right\|^2\right] \\
\leq &\sum_{t=1}^{\infty}\frac{L(\eta_t^2+\beta_t^2)\rho^2}{2} + \mathbb{E}\left[\mathcal{L}_{\text{att}}(\alpha_1;\omega_1)\right] \leq \infty,
\end{aligned} \tag{30}$$

where the second inequality holds for $\sum_{t=1}^{\infty}\eta_t^2 \leq \infty$, $\sum_{t=1}^{\infty}\beta_t^2 \leq \infty$, $\left\|\nabla_\alpha \mathcal{L}_{\text{att}}(\alpha_t;\omega_t)\right\| \leq \rho$, $\left\|\nabla_\omega J_\pi(\alpha_{t+1}(\omega_t))\right\| \leq \rho$. Since

$$\sum_{t=1}^{\infty} \beta_t \mathbb{E}\left[\left\|\nabla_\omega \mathcal{L}_{\text{att}}(\alpha_{t+1};\omega_t)\right\| \cdot \left\|\nabla_\omega J_\pi(\alpha_{t+1}(\omega_t))\right\|\right] \leq L\rho \sum_{t=1}^{\infty}\beta_t \leq \infty. \tag{31}$$

Therefore, we have

$$\sum_{t=1}^{\infty} \eta_t \mathbb{E}\left[\left\|\nabla_\alpha \mathcal{L}_{\text{att}}(\alpha_t;\omega_t)\right\|^2\right] < \infty. \tag{32}$$

Since $|(\|a\| + \|b\|)(\|a\| - \|b\|)| \leq \|a+b\|\|a-b\|$, we can derive that

$$
\begin{aligned}
&\left|\mathbb{E}\left[\|\nabla_\alpha \mathcal{L}_{\text{att}}(\alpha_{t+1};\omega_{t+1})\|^2\right] - \mathbb{E}\left[\|\nabla_\alpha \mathcal{L}_{\text{att}}(\alpha_t;\omega_t)\|^2\right]\right| \\
=&\left|\mathbb{E}\Big[\big(\|\nabla_\alpha \mathcal{L}_{\text{att}}(\alpha_{t+1};\omega_{t+1})\| + \|\nabla_\alpha \mathcal{L}_{\text{att}}(\alpha_t;\omega_t)\|\big) + \big(\|\nabla_\alpha \mathcal{L}_{\text{att}}(\alpha_{t+1};\omega_{t+1})\| - \|\nabla_\alpha \mathcal{L}_{\text{att}}(\alpha_t;\omega_t)\|\big)\Big]\right| \\
\leq&\mathbb{E}\Big[\Big|\|\nabla_\alpha \mathcal{L}_{\text{att}}(\alpha_{t+1};\omega_{t+1})\| + \|\nabla_\alpha \mathcal{L}_{\text{att}}(\alpha_t;\omega_t)\|\Big|\Big|\|\nabla_\alpha \mathcal{L}_{\text{att}}(\alpha_{t+1};\omega_{t+1})\| - \|\nabla_\alpha \mathcal{L}_{\text{att}}(\alpha_t;\omega_t)\|\Big|\Big] \\
\leq&\mathbb{E}\Big[\|\nabla_\alpha \mathcal{L}_{\text{att}}(\alpha_{t+1};\omega_{t+1}) + \nabla_\alpha \mathcal{L}_{\text{att}}(\alpha_t;\omega_t)\| \cdot \|\nabla_\alpha \mathcal{L}_{\text{att}}(\alpha_{t+1};\omega_{t+1}) - \nabla_\alpha \mathcal{L}_{\text{att}}(\alpha_t;\omega_t)\|\Big] \\
\leq&\mathbb{E}\Big[\big(\|\nabla_\alpha \mathcal{L}_{\text{att}}(\alpha_{t+1};\omega_{t+1})\| + \|\nabla_\alpha \mathcal{L}_{\text{att}}(\alpha_t;\omega_t)\|\big)\|\nabla_\alpha \mathcal{L}_{\text{att}}(\alpha_{t+1};\omega_{t+1}) - \nabla_\alpha \mathcal{L}_{\text{att}}(\alpha_t;\omega_t)\|\Big] \\
\leq&2L\rho\mathbb{E}\Big[\|(\alpha_{t+1},\omega_{t+1}) - (\alpha_t,\omega_t)\|\Big] \\
\leq&2L\rho\eta_t\beta_t\mathbb{E}\Big[\|(\nabla_\alpha \mathcal{L}_{\text{att}}(\alpha_t;\omega_t), \nabla_\omega J_\pi(\alpha_{t+1}(\omega_t)))\|\Big] \\
\leq&2L\rho\eta_t\beta_t\sqrt{\mathbb{E}\left[\|\nabla_\alpha \mathcal{L}_{\text{att}}(\alpha_t;\omega_t)\|^2\right] + \mathbb{E}\left[\|\nabla_\omega J_\pi(\alpha_{t+1}(\omega_t))\|^2\right]} \\
\leq&2L\rho\eta_t\beta_t\sqrt{2\rho^2} \\
\leq&2\sqrt{2}L\rho^2\eta_t\beta_t.
\end{aligned}
\tag{33}
$$

Since $\sum_{t=1}^{\infty}\eta_t = \infty$, according to Lemma D.3, we have

$$
\lim_{t\to\infty}\mathbb{E}\left[\|\nabla_\alpha \mathcal{L}_{\text{att}}(\alpha_t;\omega_t)\|^2\right] = 0. \tag{34}
$$

□

E Details of PbRL

In this section, we present details of the scripted teacher and preference collection. It is a crucial part of the PbRL, and BATTLE follows these settings as Lee et al. (2021a).

Scripted Teacher. To evaluate the performance systemically, a useful way is to consider a scripted teacher that provides preferences between a pair of agent's trajectory segments according to the oracle reward function. Leveraging the preference labels from the human teacher is ideal, while it is hard to evaluate algorithms quantitatively and quickly. Specifically, the scripted teacher can immediately provide ground truth rewards based on the state $\mathbf{s}$ and action $\mathbf{a}$. It is a function designed to approximate the human's intention.

Preference Collection. During training, we need to query human preference labels at regular intervals. It samples a batch of segment pairs and calculates the cumulative reward of each segment with rewards provided by the scripted teacher. For a specific segment pair, human prefers the segment with a larger cumulative reward. The segment with a larger cumulative reward is labelled with 1, and the smaller one is labelled with 0. As for the computational cost, we assume that M preference labels are required, the segment length is N in a run, and the time complexity is $\mathcal{O}(MN)$. However, it is negligible compared with adversary training, which involves complex gradient computation.

F Experimental Details

In this section, we provide a concrete description of our experiments and detailed hyper-parameters of BATTLE. For each run of experiments, we run on a single Nvidia Tesla V100 GPUs and 16 CPU cores (Intel Xeon Gold 6230 CPU @ 2.10GHz) for training.

F.1 Tasks

In phase one of our experiments, we evaluate our method on eight robotic manipulation tasks obtained from Meta-world (Yu et al., 2020). These tasks serve as a representative set for testing the effectiveness of our approach. In phase two, we further assess our method on two locomotion tasks sourced from Mujoco (Todorov et al., 2012). By including tasks from both domains, we aim to demonstrate the versatility and generalizability of our approach across different task types. The specific tasks we utilize in our experiments are as follows:

Meta-world

- Door Lock: An agent controls a simulated Sawyer arm to lock the door.
- Door Unlock: An agent controls a simulated Sawyer arm to unlock the door.
- Drawer Open: An agent controls a simulated Sawyer arm to open the drawer to a target position.
- Drawer Close: An agent controls a simulated Sawyer arm to close the drawer to a target position.
- Faucet Open: An agent controls a simulated Sawyer arm to open the faucet to a target position.
- Faucet Close: An agent controls a simulated Sawyer arm to close the faucet to a target position.
- Window Open: An agent controls a simulated Sawyer arm to open the window to a target position.
- Window Close: An agent controls a simulated Sawyer arm to close the window to a target position.

Mujoco

- Half Cheetah: A 2-dimensional robot with nine links and eight joints aims to learn to run forward (right) as fast as possible.
- Walker: A 2-dimensional two-legged robot aims to move in the forward (right).

F.2 Hyper-parameters Setting

We adopt the PEBBLE algorithm as our baseline approach for SA-RL (Zhang et al., 2021), and we keep the same parameter settings and neural network structure as described in their work. The specific hyperparameters for SA-RL are provided in Table 4. Similarly, for PA-AD (Sun et al., 2022), we use identical hyperparameter values to those of SA-RL, ensuring a fair comparison between the two methods.

Table 3: Hyper-parameters of BATTLE for adversary training.

Hyper-parameter	**Value**	**Hyper-parameter**	**Value**
Number of layers	3	Hidden units of each layer	256
Learning rate	0.0003	Batch size	1024
Length of segment	50	Number of reward functions	3
Frequency of feedback	5000	Feedback batch size	128
Adversarial budget	0.1	(β_1, β_2)	$(0.9, 0.999)$

Table 4: Hyper-parameters of SA-RL for adversary training.

Hyper-parameter	**Value**	**Hyper-parameter**	**Value**
Number of layers	3	Hidden units of each layer	256
Learning rate	0.00005	Mini-Batch size	32
Length of segment	50	Number of reward functions	3
Frequency of feedback	5000	Feedback batch size	128
Adversarial budget	0.1	Entropy coefficient	0.0
Clipping parameter	0.2	Discount γ	0.99
GAE lambda	0.95	KL divergence target	0.01

F.3 Victim Setting

Our experiment is divided into two phases. In the first phase, we conduct experiments using a variety of simulated robotic manipulation tasks from the Meta-world environment. In the second phase, we shift our focus to two continuous control environments from the OpenAI Gym MuJoCo suite.

Meta-world. We train the victim models on the Meta-world tasks using the SAC (Soft Actor-Critic) algorithm proposed by Haarnoja et al. (2018). We employ a fully connected neural network as the

policy network for the SAC algorithm. The detailed hyperparameters used in our experiments are provided in Table 5.

Table 5: Hyper-parameters of SAC for victim training.

Hyper-parameter	**Value**	**Hyper-parameter**	**Value**
Number of layers	3	Initial temperature	0.1
Hidden units of each layer	256	Optimizer	Adam
Learning rate	0.0001	Critic target update freq	2
Discount γ	0.99	Critic EMA τ	0.005
Batch size	1024	(β_1, β_2)	$(0.9, 0.999)$
Steps of unsupervised pre-training	9000	Discount γ	0.99

Mujoco. We directly utilize the well-trained model for demonstrating the vulnerability of the Decision Transformer. Specifically, we use the Cheetah agent[4] and the Walker agent[5] with expert-level.

F.4 Scenario Designing

To validate the effectiveness of our approach, we carefully designed two experimental scenarios: the Manipulation Scenario and the Opposite Behavior Scenario. In the Manipulation Scenario, the victim policy is a well-trained policy on robotic tasks. The objective of the adversary is to manipulate the agent's behavior through targeted adversarial attacks, causing the agent to grasp objects that are far from the original target location. The successful execution of such grasping actions indicates the success of the adversarial attack. In the Opposite Behavior Scenario, the victim policy is a well-trained policy on simulated robotic manipulation tasks. The goal of the attacker is to redirect the agent's behavior towards tasks that are opposite in nature to the original objective. For instance, if the victim policy is designed to open windows, the attacker aims to modify the agent's behavior to close the windows instead.

Table 6: Success rate of different methods with varying numbers of preference labels on the Drawer Open task in the manipulation scenario and the Faucet Close task in the opposite behavior scenario. The success rate is reported as the mean and standard deviation over 30 episodes.

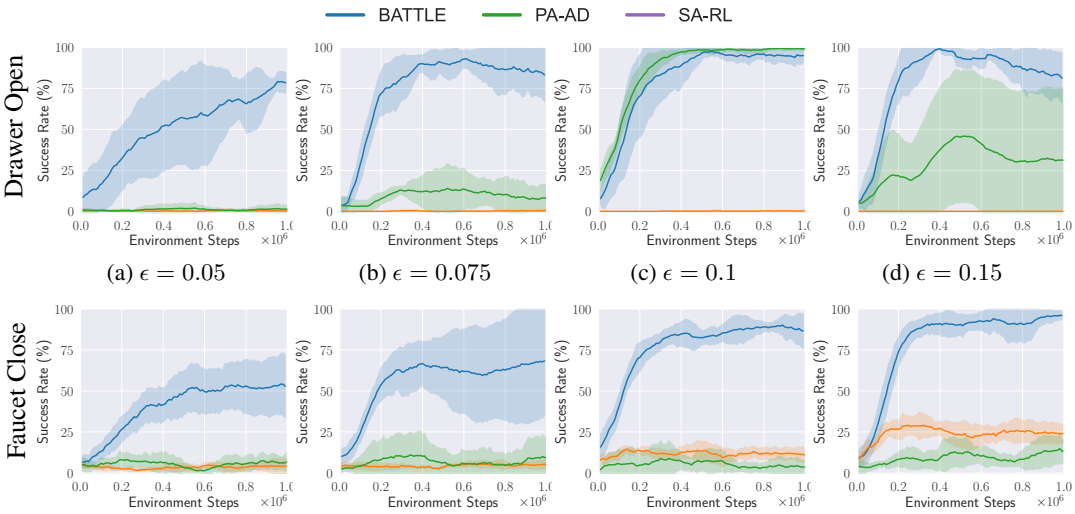
Environment	**Feedback**	**BATTLE (ours)**	**PA-AD**	**SA-RL**
Drawer Open (manipulation)	3000	$65.7\% \pm 37.1\%$	$0.0\% \pm 0.0\%$	$8.3\% \pm 13.2\%$
	5000	$86.7\% \pm 18.1\%$	$0.0\% \pm 0.0\%$	$21.3\% \pm 18.9\%$
	7000	$95.7\% \pm 13.6\%$	$0.0\% \pm 0.0\%$	$28.0\% \pm 28.1\%$
	9000	$97.0\% \pm 6.9\%$	$0.0\% \pm 0.0\%$	$13.0\% \pm 18.5\%$
Faucet Close (opposite behavior)	1000	$69.7\% \pm 35.2\%$	$16.7\% \pm 9.4\%$	$2.0\% \pm 6.0\%$
	3000	$79.0\% \pm 16.2\%$	$29.0\% \pm 14.0\%$	$6.0\% \pm 11.7\%$
	5000	$95.3\% \pm 9.2\%$	$21.3\% \pm 12.8\%$	$3.3\% \pm 12.7\%$
	7000	$95.3\% \pm 7.6\%$	$22.7\% \pm 12.4\%$	$4.0\% \pm 7.1\%$

G Extensive Experiments

Impact of Feedback Amount. We evaluate the performance of BATTLE using different numbers of preference labels. Table 6 presents the results of all methods with varying numbers of labels: $3000, 5000, 7000, 9000$ for the Drawer Open task in the manipulation scenario and $1000, 3000, 5000, 7000$ for the Faucet Close task in the opposite behavior scenario. Based on the experimental results shown in Table 6, we conclude that providing an adequate amount of human feedback improves the performance of our method, leading to a stronger adversary and a more stable attack success rate. We observe that the performance of BATTLE consistently improves as the number of preference labels increases, highlighting the crucial impact of the number of preference labels on

[4] https://huggingface.co/edbeeching/decision-transformer-gym-halfcheetah-expert
[5] https://huggingface.co/edbeeching/decision-transformer-gym-walker2d-expert

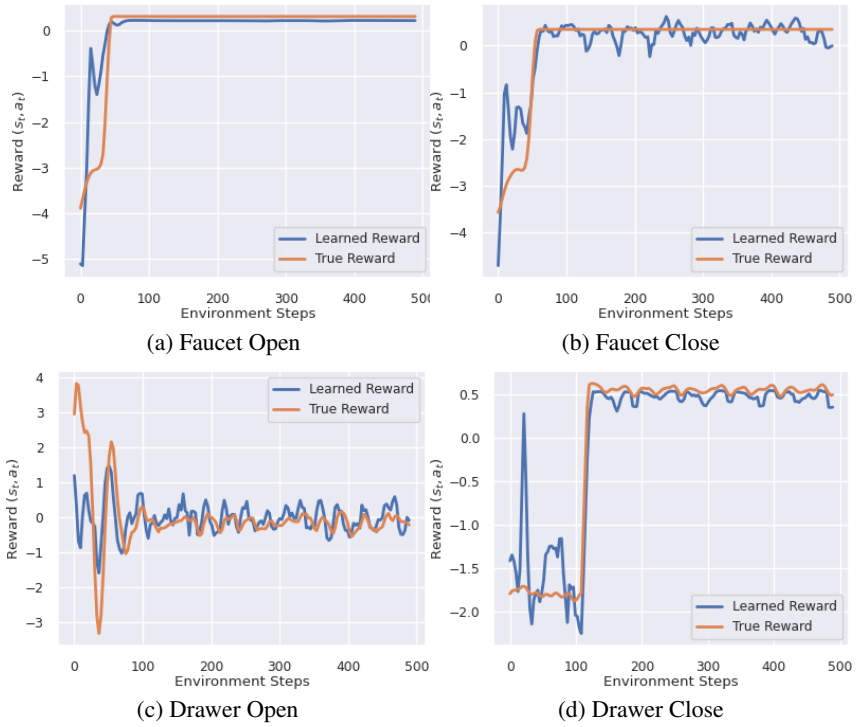
adversary learning. In contrast, SA-RL and PA-AD exhibit poor performance even with a sufficient amount of human feedback, with PA-AD failing entirely in the manipulation scenario. This can be attributed to the limited exploration space of these methods, which is constrained by the fixed victim policy. In contrast, BATTLE achieves better exploration by incorporating an intention policy, resulting in improved performance.



(a) $\epsilon = 0.05$ (b) $\epsilon = 0.075$ (c) $\epsilon = 0.1$ (d) $\epsilon = 0.15$

(e) $\epsilon = 0.02$ (f) $\epsilon = 0.05$ (g) $\epsilon = 0.075$ (h) $\epsilon = 0.1$

Figure 8: Training curves of success rate with different adversarial budgets on Drawer Open for the manipulation scenario and Faucet Close for the opposite behavior scenario. The solid line and shaded area denote the mean and the standard deviation of the success rate across five runs.



(a) Faucet Open (b) Faucet Close

(c) Drawer Open (d) Drawer Close

Figure 9: Time series of the normalized learned reward (blue) and the ground truth reward (orange). These rewards are obtained from rollouts generated by a policy optimized using BATTLE.

Impact of Different Attack Budgets. We also investigate the impact of the attack budget on the performance. To gain further insights, we conduct additional experiments with different attack budgets: $0.05, 0.075, 0.1, 0.15$ for the Drawer Open task and $0.02, 0.05, 0.075, 0.1$ for the Faucet Close task in the respective scenarios. In Figure 8, we present the performance of the baseline method and BATTLE with different attack budgets. The experimental results demonstrate that the performance of all methods improves with an increase in the attack budget.

Quality of learned reward functions. We further analyze the quality of the reward functions learned by BATTLE compared to the true reward function. In Figure 9, we present four time series plots that depict the normalized learned reward (blue) and the ground truth reward (orange). These plots represent two scenarios: opposite behaviors and manipulation tasks. The results indicate that the learned reward function aligns well with the true reward function derived from human feedback. This alignment is achieved by capturing various human intentions through the preference data.

Robust Agents Training and Evaluating. An intuitive application of BATTLE is in evaluating the robustness of a given model or enhancing the robustness of an agent through adversarial training. ATLA (Zhang et al., 2021) is a general training framework for improving robustness, which involves alternating training between an agent and an adversary. Building upon this concept, we introduce BATTLE-ATLA, which combines BATTLE with the ATLA framework by training an agent and a BATTLE attacker alternately. The robustness performance of BATTLE-ATLA for a SAC agent is presented in Table 7 and compared with state-of-the-art robust training methods. The experimental results provide two key insights: firstly, BATTLE-ATLA significantly enhances the robustness of agents, demonstrating its effectiveness in improving agent resilience to adversarial attacks. Secondly, BATTLE exhibits the capability to launch stronger attacks on robust agents, highlighting its effectiveness as an adversary in the adversarial training process.

Table 7: Average episode rewards $\pm$ standard deviation of robust agents under different attack methods, and results are averaged across 100 episodes.

Task	Model	BATTLE	PA-AD	SA-RL	Average Reward
Door Lock	BATTLE-ATLA	874±444	628±486	503±120	**668**
	PAAD-ATLA	491±133	483±15	517±129	497
	SARL-ATLA	469±11	629±455	583±173	545
Door Unlock	BATTLE-ATLA	477±203	745±75	623±60	**615**
	PAAD-ATLA	398±12	381±11	398±79	389
	SARL-ATLA	393±36	377±8	385±26	385
Faucet Open	BATTLE-ATLA	442±167	451±96	504±55	465
	PAAD-ATLA	438±53	588±222	373±32	466
	SARL-ATLA	610±293	523±137	495±305	**522**
Faucet Close	BATTLE-ATLA	1048±343	1223±348	570±453	**947**
	PAAD-ATLA	661±279	371±65	704±239	538
	SARL-ATLA	1362±149	688±196	426±120	825