
An Offline Adaptation Framework for Constrained Multi-Objective Reinforcement Learning

Qian Lin¹, Zongkai Liu¹, Danying Mo¹, Chao Yu^{1,2,3†}

¹Sun Yat-sen University, Guangzhou, China

²Pengcheng Laboratory, Shenzhen, China

³MoE Key Laboratory of Information Technology, Guangzhou, China
{linq67, liuzk, mody3}@mail2.sysu.edu.cn, yuchao3@mail.sysu.edu.cn

Abstract

In recent years, significant progress has been made in multi-objective reinforcement learning (RL) research, which aims to balance multiple objectives by incorporating preferences for each objective. In most existing studies, specific preferences must be provided during deployment to indicate the desired policies explicitly. However, designing these preferences depends heavily on human prior knowledge, which is typically obtained through extensive observation of high-performing demonstrations with expected behaviors. In this work, we propose a simple yet effective offline adaptation framework for multi-objective RL problems without assuming handcrafted target preferences, but only given several demonstrations to implicitly indicate the preferences of expected policies. Additionally, we demonstrate that our framework can naturally be extended to meet constraints on safety-critical objectives by utilizing safe demonstrations, even when the safety thresholds are unknown. Empirical results on offline multi-objective and safe tasks demonstrate the capability of our framework to infer policies that align with real preferences while meeting the constraints implied by the provided demonstrations.

1 Introduction

In the standard reinforcement learning (RL) setting, the primary goal is to obtain a policy that maximizes a cumulative scalar reward [Sutton and Barto, 2018]. However, in many real-world applications involving multiple objectives, a desired policy must not only strike a balance among potentially conflicting objectives but also consider the constraints imposed on specific safety-critical objectives. Such requirements motivate the research of Multi-objective RL (MORL) [Liu et al., 2014, Mossalam et al., 2016] and safe RL [Gu et al., 2022, Achiam et al., 2017]. In addition to reward maximization, the former aims to enable policies that cater to a target preference indicating the trade-off between competing objectives, while the latter focuses on reducing the cost measures of learned policies within a given safety threshold.

Despite the development of meticulous and effective mechanisms to achieve these goals, most existing MORL and safe RL algorithms rely on predefined target preferences or safety thresholds. These elements need to be carefully designed by human experts with prior knowledge, generalized from a large number of demonstrations through observation. Taking autonomous driving as an example, defining aggressive, stable, and conservative strategies through different preferences between minimizing energy consumption and increasing driving speed may require extensive human participation, which means that researchers need to categorize driving data into different styles based on human experience and observe the energy consumption-speed ratio to estimate appropriate preference weights and

[†]Corresponding author

safety thresholds for different strategies. Moreover, it is uncertain whether policies based on manually designed preferences will exhibit expected behaviors or whether feasible policies exist under given safety constraints. The challenge of designing appropriate preferences or safety thresholds becomes more pronounced as the number of objectives increases.

Compared to manually designing target preferences or safety thresholds based on human knowledge, it is more natural to infer expected behaviors through a few demonstrations that implicitly indicate the trade-off between multiple objectives and the constraints of safety. For instance, selecting demonstrations with conservative behaviors from the driving dataset can be easier than inferring preferences that lead to conservative policies. Therefore, in this work, we formulate a novel offline adaptation problem for constrained MORL, with the goal to leverage a few demonstrations with expected behaviors, rather than relying on handcrafted target preferences and safety thresholds, to generate a target policy that achieves desired trade-offs across various objectives and meets the constraints on safety-critical objectives under offline settings.

To achieve this, we first focus on unconstrained MORL scenarios and propose a simple yet effective offline adaptation framework **Preference Distribution Offline Adaptation (PDOA)**, which includes: 1) learning a set of policies that respond to various preferences during training, and 2) adapting a distribution of target preferences based on a few given demonstrations during deployment. Specifically, we initialize the first part with existing state-of-the-art offline MORL algorithms, and then in the second part, we propose to align the adapted policies with expected behaviors by modeling the posterior preference distribution regarding demonstrations. Moreover, we show that our framework can be extended to constrained MORL settings by converting a constrained RL problem into an unconstrained MORL counterpart, and incorporating a conservative estimate of preference weights on constrained objectives to mitigate the potential constraint violations. Lastly, we conduct several empirical experiments on classical MORL, safe RL tasks and a novel constrained MORL environment under offline settings, demonstrating the capability of our framework in generating policies that align with the real preferences and meet the constraints implied in demonstrations.

2 Related Work

Multi-objective RL Many existing MORL methods explicitly maintain a set of policies tailored to various given preferences to approximate the Pareto front of optimal policies. These works either apply a single-policy algorithm individually for each candidate preference [Rojers et al., 2014, Mossalam et al., 2016], employ evolutionary algorithms to generate a population of diverse policies [Handa, 2009, Xu et al., 2020] or simultaneously learn a set of policies represented by a single network [Abels et al., 2019, Basaklar et al., 2022]. Furthermore, due to the potential costs and risks associated with extensive online exploration, several studies [Zhu et al., 2023, Lin et al., 2024] have been proposed to leverage only offline datasets for MORL by extending offline return-conditioned methods [Chen et al., 2021, Emmons et al., 2021] or offline policy-regularized methods [Fujimoto and Gu, 2021, Wang et al., 2022] to MORL settings. Despite the ability to obtain a set of well-performing policies for various preferences, most existing MORL methods overlook the process of acquiring target preferences for identifying appropriate policies during practical deployment. In this work, we assume no online interactions and no target preferences but several demonstrations generated with expected behaviors, which are easier to access than meticulously designed target preferences.

Safe RL While maximizing the expected reward, classical safe RL methods restrict the cumulative cost to stay within a predefined safety threshold through Lagrangian primal-dual methods [Stooke et al., 2020, Chow et al., 2017] or primal approaches [Xu et al., 2021, Sootla et al., 2022]. Recently, several studies have focused on learning a set of policies that respond to various safety thresholds in both online settings [Yao et al., 2024a] and offline settings [Liu et al., 2023a, Lin et al., 2023]. Similar to MORL, these works also assume access to well-designed safety thresholds that ensure safe behaviors. Unlike prior research, our framework relies solely on safe demonstrations to indicate the implicit constraints and offers a mechanism to mitigate constraint violations through conservatism.

RL with Offline Adaptation Several studies have applied meta-learning techniques to address MORL [Chen et al., 2019] or safe RL problems [Guan et al., 2024a], but they require online interactions for task adaptation and suffer from low sample efficiency. Additionally, other research endeavors explore offline adaptation methods to mitigate environmental uncertainty by modelling a posterior distribution over all possible Markov Decision Processes (MDPs) [Ghosh et al., 2022] or

considering varying confidence levels in conservative value estimates [Hong et al., 2022]. [Mitchell et al., 2021, Xu et al., 2022a] focus on offline adaptation for multi-task problems, aiming to achieve fast adaptation to new downstream tasks after offline training on multi-task experience. Among these methods, offline meta RL [Mitchell et al., 2021] addresses a problem similar to ours, which aims to train a meta policy that can adapt to a new task with limited data. Prompt-DT [Xu et al., 2022a] achieves quick adaptation to new tasks by incorporating a few demonstrations as prompts into the decision transformer [Chen et al., 2021] framework. We present further discussions about the difference between multi-task RL and our setting in Appendix A.1.

3 Preliminaries

3.1 Constrained Multi-Objective MDP (CMO-MDP)

Both multiple-objective RL and safe RL can be discussed based on a uniform formulation: constrained multi-objective MDP (CMO-MDP) proposed by LP3 [Huang et al., 2022]. A CMO-MDP is defined as a tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathbf{r}, \mathbf{c}, \beta, \gamma)$ with state space \mathcal{S} , action space \mathcal{A} , transition distribution $\mathcal{P}(s'|s, a)$, vector reward functions $\mathbf{r} \in R^N$ for N unconstrained objectives, vector cost functions $\mathbf{c} \in R_+^K$, safety thresholds $\beta \in R_+^K$ for K constrained objectives and discount factor $\gamma \in [0, 1]$. The goal is to maximize the rewards on unconstrained objectives while ensuring the costs on constrained objectives remain within the safety threshold β . Since it is typically infeasible to maximize all task objectives simultaneously, preferences $\omega \in \Omega$ and preference functions $f_\omega(\mathbf{r})$ which map the reward \mathbf{r} to a scalar utility under a specific preference ω , are introduced to control the trade-off between unconstrained objectives. Given preferences $\omega \in \Omega$ and safety thresholds β , the goal can be formulated as follows:

$$\max_{\pi_{\omega, \beta}} \mathbb{E}_{\pi_{\omega, \beta}} [R_\omega], \text{ s.t. } \mathbf{C} \preceq \beta, \quad (1)$$

where $R_\omega = \sum_t f_\omega(\mathbf{r}_t)$ and $\mathbf{C} = \sum_t \mathbf{c}_t$ represent cumulative utility and vector cost over time t , respectively. We denote $\pi_{\omega, \beta}$ as a policy conditioned on ω, β . In this paper, we consider the linear preference setting (i.e., $f_\omega(\mathbf{r}) = \omega^\top \mathbf{r}$ where $\omega \in R^N$ and $\|\omega\|_1 = 1$), which is widely studied and applied [Mossalam et al., 2016, Abels et al., 2019] and also serves as a bridge between unconstrained and constrained MORL, as shown in Section 4.3. A CMO-MDP problem can degenerate to a standard safe RL problem when $N = 1$ and to a standard multi-objective problem when $K = 0$.

3.2 Offline MORL

Under offline MORL settings, an offline dataset $\mathcal{D} = \{(s, a, s', \mathbf{r}, \mathbf{c}, \omega)\}$ is the only data available for training, which is generated by a set of behavior policies $\pi_b(\cdot|\omega)$ with diverse behavioral preferences ω . One straightforward yet effective approach to learn a set of policies for various preferences is to adapt offline single-objective RL methods for MORL settings. An example of this approach is the **multi-objective version of Diffusion-QL (MODF)** [Lin et al., 2024], which incorporates a preference-conditioned policy (i.e., $\pi(a|s, \omega)$) and a multi-dimensional value function for N objectives (i.e., $\mathbf{Q}(s, a, \omega) = Q_1(s, a, \omega), \dots, Q_N(s, a, \omega)$) into Diffusion-QL [Wang et al., 2022]:

$$\begin{aligned} L_\pi &= -\mathbb{E}_{(s, a, \omega) \sim \mathcal{D}} [\mathbb{E}_{a' \sim \pi(\cdot|s, \omega)} [\omega^\top \mathbf{Q}(s, a', \omega)] - \\ &\quad \kappa \mathbb{E}_{i \sim \mathcal{U}, \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} [\|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_i} a + \sqrt{1 - \bar{\alpha}_i} \epsilon, s, i)\|^2]], \\ L_Q &= \mathbb{E}_{(s, a, \mathbf{r}, s', \omega) \sim \mathcal{D}} [(\mathbf{r} + \gamma \mathbb{E}_{a' \sim \pi(\cdot|s, \omega)} \mathbf{Q}(s', a', \omega) - \mathbf{Q}(s, a, \omega))^2], \end{aligned} \quad (2)$$

where i is the diffusion timestep, κ is the regularization weight, $\bar{\alpha}_i$ are pre-defined parameters of diffusion model and $\epsilon_\theta(\cdot)$ is a denoiser model. The diffusion policy generates the actions by iteratively using $\epsilon_\theta(\cdot)$ to recover actions from noise. The second term in the actor loss of Eq. (2) is a diffusion reconstruction loss, which serves as a regularization term to align the actions of diffusion policy with the behavioral actions in the dataset.

Pareto-Efficient Decision Agents (PEDA) [Zhu et al., 2023] is another MORL method based on supervised RL, which trains a policy conditioned on both target preferences and vector returns through a supervised paradigm:

$$L_\pi = -\mathbb{E}_{(\tau_{t:T}, \omega) \sim \mathcal{D}} [\log \pi(a_t | s_t, \mathbf{g}_t, \omega)], \quad (3)$$

where $\tau_{t:T} = \{(s_t, a_t, \mathbf{r}_t), \dots, (s_T, a_T, \mathbf{r}_T)\}$ is a trajectory segment, $\mathbf{g}_t = \sum_{t'=t}^T \mathbf{r}_{t'}$ is the target vector return (a.k.a., return-to-go) and ω is the behavioral preference of $\tau_{t:T}$. It is worth noting that despite the significant performance of the above two methods in offline MORL problems, both require human-provided target preferences during deployment to achieve the desired behavior, and therefore cannot be directly applied to the setting in this paper.

4 An Offline Adaptation Framework for Constrained Multi-Objective RL

4.1 Problem Formulation of Offline Adaptation for CMO-MDP

In this paper, we focus on a novel offline adaptation problem for constrained MORL, with the goal to leverage only a few demonstrations to generate the policies that exhibit expected behaviors. During training, an offline dataset $\mathcal{D} = \{(s, a, s', \mathbf{r}, \mathbf{c}, \omega)\}$ is provided for policy training. During deployment, we have access to a demonstration set corresponding to a target G , i.e.,

$$\mathcal{B}_G = \{x_i \sim \pi_G^*\}_{i=1}^M, \quad (4)$$

where x_i is defined as a tuple $(s_i, a_i, s'_i, \mathbf{r}_i, \mathbf{c}_i)$ and M is the total number of transition demonstrations. The target G can be preferences in MORL problems ($G = \omega_g$), safety thresholds in safe problems ($G = \beta_g$) or a combination of both ($G = (\omega_g, \beta_g)$), and π_G^* is the expert policy that achieves the best utility and meet the constraints under the preference ω_g and safety threshold β_g of the target G . In our setting, the real target G is inaccessible, and the goal is to obtain an adapted policy for the target G that ensures high utility $f_{\omega_g}(\mathbf{r})$ and meets the constraints with safety threshold β_g by leveraging demonstration set \mathcal{B}_G during the deployment phase.

4.2 Offline Adaptation for the Unconstrained Case

First, we set aside the constraints in CMO-MDP and focus on the unconstrained version. Our proposed framework, Preference Distribution Offline Adaptation (PDOA), solves the offline adaptation problem under unconstrained settings in two steps: **1**) learning a set of policies that respond to various preferences during training; and then **2**) adapting a distribution of target preferences based on given demonstrations during deployment.

In the first part, we directly apply existing offline MORL algorithms on the dataset \mathcal{D} to obtain a set of policies $\hat{\pi}_\omega$ that respond to varying preferences ω . In the second phase, we propose to model the distribution of target preferences and then utilize this distribution to obtain a reliable estimation of target preference. Specifically, we consider the posterior probability of the target preference ω_g with regard to demonstration set \mathcal{B}_{ω_g} , i.e., $P(\omega_g | \mathcal{B}_{\omega_g}, \mathcal{D}) = \frac{P(\mathcal{B}_{\omega_g} | \omega_g, \mathcal{D}) P(\omega_g | \mathcal{D})}{P(\mathcal{B}_{\omega_g} | \mathcal{D})}$. Here $P(\mathcal{B}_{\omega_g} | \omega_g, \mathcal{D})$ represents the probability that the optimal $\pi_{\omega_g}^*$ generates samples \mathcal{B}_{ω_g} in the real environment $\mathcal{P}(s', \mathbf{r} | s, a)$. Due to the inaccessibility of \mathcal{P} and π_ω^* under offline settings, we replace them with the empirical dynamics $\hat{\mathcal{P}}_\omega(s' | s, a)$ and its corresponding optimal policy $\hat{\pi}_\omega^*$, which can be obtained using offline data \mathcal{D} during training. Therefore, the preference posterior distribution can be approximated by

$$\begin{aligned} P(\omega | \mathcal{B}_{\omega_g}, \mathcal{D}) &= \frac{P(\mathcal{B}_{\omega_g} | \omega, \mathcal{D}) P(\omega | \mathcal{D})}{P(\mathcal{B}_{\omega_g} | \mathcal{D})} \approx \frac{P(\mathcal{B}_{\omega_g} | \hat{\mathcal{M}}_\omega, \hat{\pi}_\omega^*, \mathcal{D}) P(\omega | \mathcal{D})}{P(\mathcal{B}_{\omega_g} | \mathcal{D})} \\ &\propto P(\omega | \mathcal{D}) \prod_{i=1}^M P_{\hat{\pi}_\omega^*}(s_i) \hat{\pi}_\omega^*(a_i | s_i) \hat{\mathcal{P}}_\omega(s'_i, \mathbf{r}_i | s_i, a_i), \end{aligned} \quad (5)$$

where $(s_i, a_i, \mathbf{r}_i, s'_i) \in \mathcal{B}_{\omega_g}$. One challenge in Eq. (5) is the requirement of explicitly modeling the state distribution $P_{\hat{\pi}_\omega^*}(s_i)$ and the transition probability $\hat{\mathcal{P}}_\omega(s'_i, \mathbf{r}_i | s_i, a_i)$. Another concern is that demonstrations \mathcal{B}_{ω_g} with the target preference ω_g can be out-of-distribution samples for the estimation of $\hat{\pi}_\omega^*$ and $\hat{\mathcal{M}}_\omega$, leading to considerable discrepancy between estimated $P_{\hat{\pi}_\omega^*}(s_i)$, $\hat{\mathcal{P}}_\omega(s'_i, \mathbf{r}_i | s_i, a_i)$ and their real-world counterparts. This challenge is pronounced in multi-objective settings due to significant differences in the trajectory distributions of the optimal policy under various preferences. Therefore, following previous offline adaptation works [Ghosh et al., 2022, Hong et al.,

2022], we opt to approximate $\log P_{\hat{\pi}_\omega^*}(s_i)\hat{P}_\omega(s'_i, \mathbf{r}_i|s_i, a_i)$ with a surrogate defined by TD error of value models of $\hat{\pi}_\omega^*$ in $\hat{\mathcal{M}}_\omega$:

$$r_\omega^{\text{TD}}(s, a, \mathbf{r}, s') = -\delta\|\mathbf{Q}(s, a, \omega) - (\mathbf{r} + \gamma\mathbf{V}(s', \omega))\|_2^2, \quad (6)$$

where δ is a hyperparameter. This approximation makes sense because $r_\omega^{\text{TD}}(s, a, \mathbf{r}, s')$ not only measures how likely the sample (s, a, \mathbf{r}, s') occurred during training for preference ω but also aligns with the estimated dynamics \hat{P}_ω . In other words, $r_\omega^{\text{TD}}(s, a, \mathbf{r}, s')$ has a high value if (s, a, \mathbf{r}, s') is an in-distribution sample relative to the training datasets under preference and occurs with a high probability in $\hat{\mathcal{M}}_\omega$.

Then, we fit the posterior $P(\omega|\mathcal{B}_{\omega_g}, \mathcal{D})$ with a Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}\mathbf{I})$ with parameters $\boldsymbol{\mu} \in R^K$ and $\boldsymbol{\sigma} \in R_+^K$ by minimizing their Kullback-Leibler divergence, leading to the following adaptation loss:

$$\begin{aligned} \mathcal{L}_{\text{adpt}}(\boldsymbol{\mu}, \boldsymbol{\sigma}) &= \text{D}_{\text{KL}}(p_\theta(\omega)||P(\omega|\mathcal{B}_{\omega_g}, \mathcal{D}))/M + \eta(\|\boldsymbol{\mu}\|_1 - 1)^2 \\ &= -\mathbb{E}_{\omega \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}\mathbf{I})} \left[\frac{1}{M} \sum_{i=1}^M \left[r_\omega^{\text{TD}}(s_i, a_i, \mathbf{r}_i, s'_i) + \log \hat{\pi}_\omega^*(a_i|s_i, \omega) \right] + \right. \\ &\quad \left. \frac{\log P(\omega|\mathcal{D})}{M} \right] - \frac{H(p_\theta)}{M} + \eta(\|\boldsymbol{\mu}\|_1 - 1)^2, \end{aligned} \quad (7)$$

where the term $\eta(\|\boldsymbol{\mu}\|_1 - 1)^2$ with hyperparameter η regularizes $\|\boldsymbol{\mu}\|_1$ to be close to 1. We approximate the prior distribution $P(\omega|\mathcal{D})$ with a Gaussian distribution fit to the behavioral preferences of training data \mathcal{D} . Eq. (7) indicates that ω is likely to be the target preference ω_g if it corresponds to high $r_\omega^{\text{TD}}(s_i, a_i, \mathbf{r}_i, s'_i)$, $\log \hat{\pi}_\omega^*(a_i|s_i, \omega)$ and $P(\omega|\mathcal{D})$, which means: 1) the demonstration set \mathcal{B}_{ω_g} appears with high probability in both the training set and $\hat{\mathcal{M}}_\omega$; 2) $\hat{\pi}_\omega^*(a_i|s_i, \omega)$ is likely to generate the actions in \mathcal{B}_{ω_g} ; and 3) ω stays within the prior preference distribution with a high probability. We justify the effectiveness of this approach with the empirical results in Appendix A.9, where $r_\omega^{\text{TD}}(s_i, a_i, \mathbf{r}_i, s'_i)$ (called TD reward) and $\hat{\pi}_\omega^*(a_i|s_i, \omega)$ (called action likelihood reward) shows a strong correlation with the real target preference.

Implementation The first part of our framework PDOA is instantiated with state-of-the-art MORL algorithms: MODF and the best algorithm MORvS in PEDA mentioned in Section 3.2, to learn a set of policies that responds to various preferences. We refer to these two instances as **PDOA [MODF]** and **PDOA [MORvS]**. The well-trained policies and their corresponding value functions obtained by MORL algorithms during training are utilized to calculate the action likelihood reward $\hat{\pi}_\omega^*(\cdot)$ and TD reward $r_\omega^{\text{TD}}(\cdot)$ in Eq. (7). During deployment, the preference distribution $p_\theta(\omega)$ is updated through the adaptation loss (7) on the demonstration set \mathcal{B}_{ω_g} . Finally, we obtain an adapted target preference $\omega_a = \boldsymbol{\mu}/\|\boldsymbol{\mu}\|_1$ and policy $\hat{\pi}_\omega^*(\cdot|\omega_a)$ that aligns with the real target preference ω_g implied in demonstrations \mathcal{B}_{ω_g} . More details about implementation can be found in Appendix A.5.

4.3 Extension to Constrained Settings

Then, we consider a natural extension of our MORL adaptation framework to constrained settings. Under the linear preference setting, the constrained MORL problem in Eq. (1) can be converted to its dual form with zero duality gap [Paternain et al., 2019]:

$$\min_{\boldsymbol{\lambda} \in R_+^K} \max_{\pi} \mathbb{E}_\pi \left[\sum_t \boldsymbol{\omega}_g^\top \mathbf{r}_t - \boldsymbol{\lambda}^\top \left(\sum_t \mathbf{c}_t - \boldsymbol{\beta}_g \right) \right]. \quad (8)$$

Denoting $\boldsymbol{\lambda}^*$ as the solution of this problem, the dual problem 8 can be rewritten as:

$$\max_{\pi} \mathbb{E}_\pi \left[\sum_t [\boldsymbol{\omega}_g^\top, \boldsymbol{\lambda}^{*\top}] \cdot [\mathbf{r}_t^\top, -\mathbf{c}_t^\top]^\top \right]. \quad (9)$$

This formulation means that the constrained MORL problem under safety threshold $\boldsymbol{\beta}$ is uniquely equivalent to an unconstrained problem of finding the optimal policy under an extended preference $\tilde{\boldsymbol{\omega}}_g = [\boldsymbol{\omega}_g^\top, \boldsymbol{\lambda}^{*\top}]^\top / \|[\boldsymbol{\omega}_g^\top, \boldsymbol{\lambda}^{*\top}]^\top\|_1$ among extended $N + K$ objectives $\tilde{\mathbf{r}} = [\mathbf{r}_t^\top, -\mathbf{c}_t^\top]^\top$. However, solving for the extended preference is challenging and requires the real safety thresholds, which

are inaccessible in our setting. Nevertheless, Section 4.2 provides an approach to infer the target preferences from the demonstration set, helping us circumvent this challenge. Therefore, we can convert a constrained MORL problem to an unconstrained MORL problem, where the vector reward is defined as $\tilde{\mathbf{r}} = [\mathbf{r}_t, -\mathbf{c}_t]$ and dataset \mathcal{D} is augmented to $\tilde{\mathcal{D}}$. Here, $\tilde{\mathbf{r}}_{1:N} = \mathbf{r}$ corresponds to N unconstrained objectives, while $\tilde{\mathbf{r}}_{N+1:N+K} = -\mathbf{c}$ associated with K constrained objectives. One issue with this approach is how to set the behavior preference for augmented dataset $\tilde{\mathcal{D}}$. We present an effective scheme for approximating behavioral preferences in Appendix A.5.

However, one concern about this approach is the estimation error between the adapted preference $\tilde{\omega}_a$ obtained through Eq. (7) and the real preference $\tilde{\omega}_g$ associated with the problem 9 due to the insufficiency of demonstrations. This discrepancy can lead to constraint violations when the preference weight of the adapted preference $\tilde{\omega}_a$ on a constrained objective is less than the one of $\tilde{\omega}_g$ (i.e., $[\tilde{\omega}_a]_i < [\tilde{\omega}_g]_i$ for any $i \in \{N+1, \dots, N+K\}$). Therefore, we propose to apply a conservative estimate of preference weights on constrained objectives by neglecting the minimum value with probability $1 - \alpha$ in the adaptation distribution $\tilde{\omega} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma})$. Specifically, the preference weights on the i^{th} objective is estimated as follows:

$$\mathbf{b}_i = \text{CVaR}_\alpha([\tilde{\omega}]_i) = \frac{1}{\alpha} \int_{1-\alpha}^1 \text{VaR}_u([\tilde{\omega}]_i) du = \boldsymbol{\mu}_i + \boldsymbol{\sigma}_i \frac{\varphi(\Phi^{-1}(1-\alpha))}{\alpha}, N < i \leq N+K, \quad (10)$$

and $\mathbf{b}_i = \boldsymbol{\mu}_i$ for $1 \leq i \leq N$. Here, VaR is the value of risk and CVaR is the conditional one, and $\varphi(x) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{x^2}{2})$ is the standard normal p.d.f., and $\Phi(x)$ is the standard normal c.d.f.. The parameter α in Eq. (10) controls the conservatism of the estimation of preference weights. For all constrained objectives, $\alpha < 1$ results in a conservative estimate $\mathbf{b}_i > \boldsymbol{\mu}_i$, reducing the risk of constraint violation. Additionally, as the number of demonstrations increases, the standard deviation $\boldsymbol{\sigma}_i$ decreases, and thus \mathbf{b}_i will move towards $\boldsymbol{\mu}_i$, resulting in less conservatism. In the end, we obtain the adapted target preference $\tilde{\omega}_a = \mathbf{b} / \|\mathbf{b}\|_1$ and the adapted policy $\pi(s|a, \tilde{\omega}_a)$.

5 Experiment

In this section, we conduct several experiments on classical MORL environments in Section 5.1 and safe RL tasks in Section 5.2 to evaluate our framework in achieving preference alignment and approaching constraint satisfaction, respectively. Then, we test our method on a set of new constrained MORL (denoted as CMORL) tasks in Section 5.3. Finally, we discuss the effectiveness of the components of our method in Section 5.4.

Environments and Datasets In Section 5.1, we utilize the D4MORL datasets [Zhu et al., 2023] for MORL experiments, which involve two conflict objectives and a variety of behaviors based on preferences in multi-objective MuJoCo tasks. In Section 5.2, all algorithms are trained on the datasets from DSRL benchmark [Liu et al., 2023b], which involve a constrained objective and an unconstrained objective in BulletSafeGym tasks [Gronauer, 2022]. In Section 5.3, we develop a set of CMORL tasks by incorporating an additional velocity constraint to the original multi-objective MuJoCo environments and collect datasets from these tasks. Thus, these environments contain two unconstrained objectives and one constrained objective. More details about these environments and datasets can be found in Appendix A.2. We construct training sets and demonstration sets based on the above datasets and present the details on Appendix A.3.

Evaluation Protocols For evaluation, we define a target set $\mathcal{T} = \{G\}$, with each target G corresponding to a demonstration set \mathcal{B}_G that meets the target G . For MORL tasks in Section 5.1, G is a target preference $\boldsymbol{\omega}_g$. All target preferences are selected from the dataset’s preference support set at a fixed interval 0.01 (i.e., $\boldsymbol{\omega}_g = [0.01k, 1 - 0.01k]$, s.t. $\min \hat{\boldsymbol{\omega}} \preceq \boldsymbol{\omega}_g \preceq \max \hat{\boldsymbol{\omega}}$ where $k = 0, \dots, 100$ and $\hat{\boldsymbol{\omega}}$ is the behavioral preference of the dataset). For safe RL tasks in Section 5.2, G is a safety threshold β_g . All safety thresholds are set to 6 equidistant points within the range of possible thresholds. The target of CMORL tasks in Section 5.3 is $(\boldsymbol{\omega}_g, \beta_g)$, the Cartesian product of the target preference spaced at 0.1 interval and 6 equidistant safety thresholds.

We utilize \mathcal{B}_G to generate an adapted policy for each target G and gather 5 trajectories from environments to assess its expected vector return. For tasks without constraints, we present the average utility and Hypervolume metrics to demonstrate the overall performance across all targets and the diversity

of adapted policies. For tasks with constraints, we group the adapted policies by safety thresholds and report the maximum cost return on constrained objectives and average utility and Hypervolume on unconstrained objectives for each group. We perform 3 runs with various seeds and report average performance along with 1-sigma error bar. More details about metrics can be found in Appendix A.4.

Comparative Algorithms The algorithms for comparison are divided into two categories. The first category is preference/threshold-agnostic baselines that have no access to the real target preferences or safety thresholds implied in the demonstrations, including: 1) **BC-Finetune**, where the policy is learned on training dataset via behavior cloning and is fine-tuned on demonstration sets and 2) **Prompt-MODT**, a multi-objective version of Prompt-DT [Xu et al., 2022a] that transforms a multi-objective scenario into a multi-task problem via preference-based division and achieves offline adaptation for various tasks by taking demonstrations as prompts of the transformer. The second category is preference/threshold-informed baselines that have access to the implied preferences and thresholds and thus serve as the oracle benchmark, including: 1) original **MODF** and **MORvS** for MORL tasks, 2) **CDT** [Liu et al., 2023a] for safe RL tasks, which makes decisions based on given safety thresholds to ensure constraint satisfaction for various thresholds. For CMORL tasks, instead of adapting the preference through demonstrations, we enumerate all possible augmented preferences of MODF with a small interval and report the best performance under these preferences as the oracle. More details about comparative algorithms are placed in Appendix A.4.

5.1 Preference Alignment for MORL Tasks

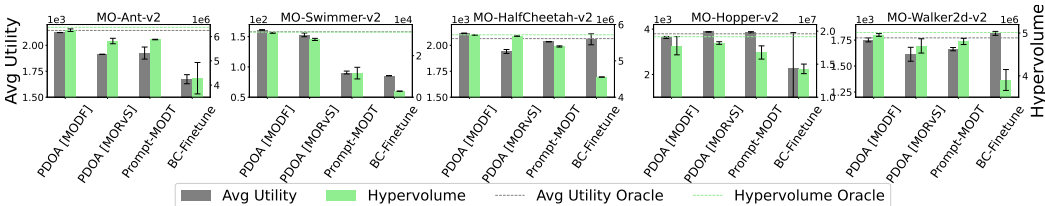


Figure 1: Results on D4MORL Amateur datasets. Higher average utility and Hypervolume are preferable. The dashed lines represent the best performance between the original MODF and MORvS.

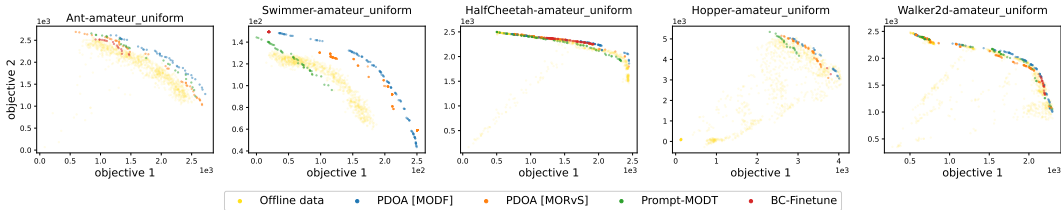


Figure 2: Pareto fronts of different algorithms on D4MORL Amateur datasets. Each point represents an adapted policy for a specific unknown target preference.

The average utility and Hypervolume of all algorithms are shown in Figure 1. Our method demonstrates superior overall performance and matches the oracle performance of preference-informed methods. Moreover, we present the Pareto fronts of different algorithms in Figure 2. Our method produces a broader and expanding Pareto front compared to BC-Finetune and Prompt-MODT, which indicates that the adapted policies of our method exhibit higher diversity and better performance. Meanwhile, we observe that the adapted policies obtained by BC-Finetune cluster around the behavior-cloned policy, and thus BC-Finetune obtains the narrow Pareto fronts and low Hypervolume, which indicates the difficulty of changing the policy’s preference through simple fine-tune with limited samples. Besides, Prompt-MODT also yields a restricted Pareto front, which can be attributed to the difficulty of partitioning tasks based on preferences. Fine-grained divisions result in data insufficiency for each task, while coarse-grained divisions lead to multiple preference behaviors being grouped into

a single task, both of which can hurt the policy performance and diversity. In contrast to BC-Finetune and Prompt-MODT, our method explicitly learns a set of policies with various preferences through MORL, thereby ensuring policy diversity. Once the target preferences are accurately identified, we can generate policies with diverse behaviors to meet various target preferences. To verify the capability of our method in preference alignment, we show the differences between the adapted preferences obtained by our method and the real target preferences in Figure 3, where we can observe a strong consistency between the adapted preferences and the real ones, especially on PDOA [MODF]. We additionally present the performance, Pareto fronts and preference comparison on the D4MORL expert datasets in Appendix A.6.1, which are consistent with the results in this section.

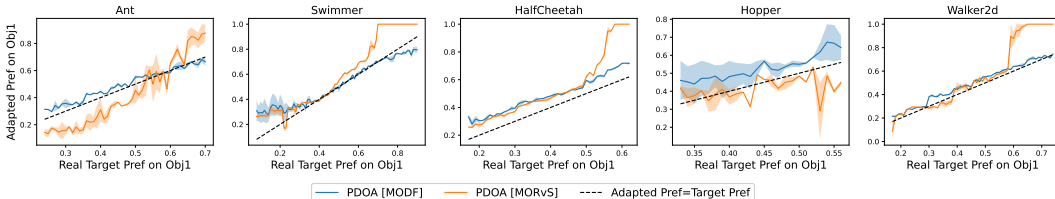


Figure 3: The comparison between the real target preferences and the adapted preferences.

5.2 Constraint Satisfaction for Safe RL Tasks

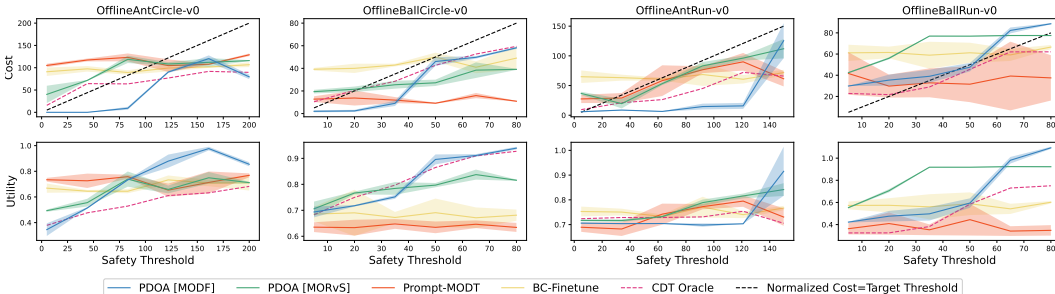


Figure 4: The adapted policies’ cost and utility of each algorithm under various safety thresholds. Here, the utility is the normalized reward, since there is only one unconstrained objective in DSRL tasks. The points above the black dashed line represent the policies that violate the constraints.

Figure 11 in Appendix A.7 presents the Pareto fronts of MODF and MORvS, i.e., the expected costs and expected rewards under various preferences, for safe RL tasks, which demonstrates that MORL algorithms can learn a variety of policies that meet various safety thresholds. Then, the reward and cost of all algorithms are shown in Figure 4. We also present the performance and Pareto fronts on other tasks in Appendix A.7. These results show that, even though the safety thresholds are inaccessible, PDOA [MODF] achieves relatively safe performance under various safety thresholds compared to other baselines, closely matching the oracle performance of the threshold-informed baseline CDT. Even with very tight safety thresholds, PDOA [MODF] can achieve constraint satisfaction or experience few constraint violations. Meanwhile, its reward performance is comparable to or even exceeds that of CDT. However, PDOA [MORvS] performs poorly on the DSRL datasets because MORvS requires accurate predictions of the target return for each preference, which is challenging due to the abundance of suboptimal trajectories in the DSRL datasets. Additionally, BC-Finetune and Prompt-MODT, which align behaviors without considering the constraints on constrained objectives, exhibit insufficient policy diversity and constraint violations in most environments.

5.3 Evaluation for Constrained MORL Tasks

The results on CMO datasets are shown in Figure 5. The CMO tasks involve more objectives than previous MORL and safe RL tasks, posing a challenge in identifying and aligning behaviors.

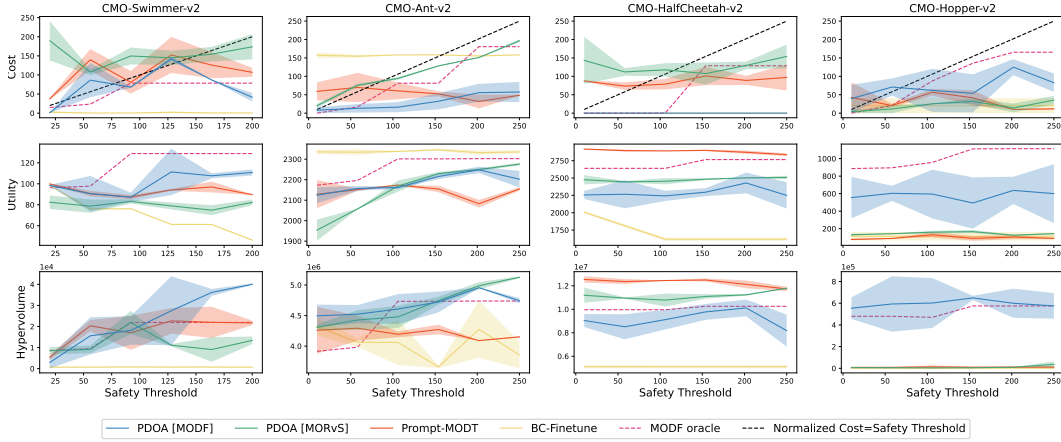


Figure 5: The maximum cost, the average utility and Hypervolume over all targets with a specific safety threshold.

This is because behaviors with various preferences increase exponentially as the number of objectives increases. We can observe that in most environments, the performance of BC-Finetune and Prompt-MODT shows low Hypervolume and a weak correlation with the safety threshold, indicating that they cannot align with the desired behaviors and exhibit high policy diversity. Nevertheless, PDOA [MODF] approaches the oracle performance with high average utility, Hypervolume, and few constraint violations, which is consistent with the results presented in Sections 5.1 and 5.2.

5.4 Ablation Study

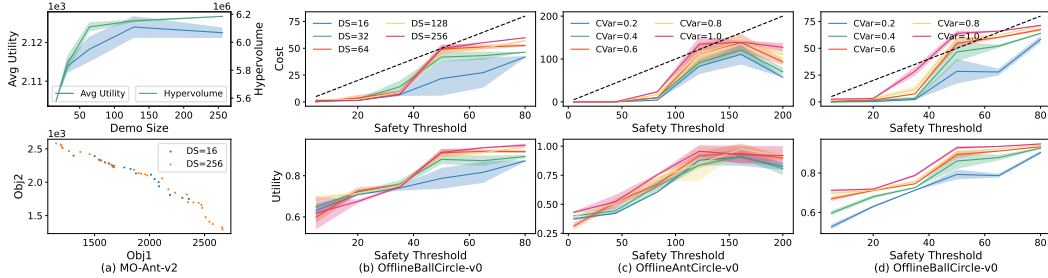


Figure 6: The performance across different demonstration sizes (abbr. DS) in Figure (a)(b) and the performance under various conservatism parameters (abbr. CVaR) in Figure (c)(d).

In this section, we aim to figure out: 1) the impact of the quantity of demonstrations on adaptation performance, and 2) the influence of the conservatism parameter α of Eq. (10) on safety performance. We conduct a set of ablation experiments for PDOA [MODF] on several MORL and safe RL tasks and present the results in Figure 6. In Figure 6 (a)(b), as the demonstration size (DS) increases, our method achieves better average utility, Hypervolume for MORL tasks and higher reward for safe RL tasks. Nevertheless, even with a very small number of demonstrations (DS = 16), the adapted policies generated by our method still exhibit sufficient diversity and safety. Figure 6 (c)(d) demonstrate that by increasing the conservatism parameters, we can further reduce constraint violations and ultimately achieve constraint satisfaction, which validates the effectiveness of our conservatism mechanism in enhancing safety.

6 Conclusion

In this paper, we present an offline adaptation framework Preference Distribution Offline Adaptation (PDOA) for constrained MORL problems where we assume no access to real target preferences or safety thresholds and only a few demonstrations with expected behaviors are available in our framework. For unconstrained MORL scenarios, we propose to 1) employ MORL methods to train a set of preference-varying policies, and then 2) align the preferences of adapted policies with expected behaviors. Furthermore, we expand our framework to accommodate constraints on specific objectives by transforming constrained problems into unconstrained counterparts. Additionally, we introduce a conservatism mechanism for preference estimation on constrained objectives to mitigate potential constraint violations. Empirical results on MORL and safe RL tasks illustrate the capability of our framework in generating diverse policies aligning with expected behaviors and approach constraint satisfaction through conservative preference estimation.

Limitation The PDOA framework involves multiple steps of preference sampling and gradient updates, which can lead to additional computational burden and latency during deployment. Besides, although the manual design of preferences and safety thresholds is not necessary, the process of constructing demonstrations may still involve human expertise.

Acknowledgements

We gratefully acknowledge the support from the National Natural Science Foundation of China (No. 62076259, 62402252), the Fundamental and Application Research Funds of Guangdong Province (No. 2023A1515012946), the Fundamental Research Funds for the Central Universities Sun Yat-sen University, and the Pengcheng Laboratory Project (PCL2023A08, PCL2024Y02).

References

- Richard S Sutton and Andrew G Barto. *Reinforcement Learning: An Introduction*. MIT press, 2018.
- Chunming Liu, Xin Xu, and Dewen Hu. Multiobjective reinforcement learning: A comprehensive overview. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 45(3):385–398, 2014.
- Hossam Mossalam, Yannis M Assael, Diederik M Roijers, and Shimon Whiteson. Multi-objective deep reinforcement learning. *arXiv preprint arXiv:1610.02707*, 2016.
- Shangding Gu, Long Yang, Yali Du, Guang Chen, Florian Walter, Jun Wang, Yaodong Yang, and Alois Knoll. A review of safe reinforcement learning: Methods, theory and applications. *arXiv preprint arXiv:2205.10330*, 2022.
- Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *International conference on machine learning*, pages 22–31. PMLR, 2017.
- Diederik M Roijers, Shimon Whiteson, and Frans A Oliehoek. Linear support for multi-objective coordination graphs. In *Proceedings of the 13rd International Conference on Autonomous Agents and Multiagent Systems*, volume 2, pages 1297–1304. IFAAMAS/ACM, 2014.
- Hisashi Handa. Solving multi-objective reinforcement learning problems by eda-rl-acquisition of various strategies. In *2009 ninth international conference on intelligent systems design and applications*, pages 426–431. IEEE, 2009.
- Jie Xu, Yunsheng Tian, Pingchuan Ma, Daniela Rus, Shinjiro Sueda, and Wojciech Matusik. Prediction-guided multi-objective reinforcement learning for continuous robot control. In *International conference on machine learning*, pages 10607–10616. PMLR, 2020.
- Axel Abels, Diederik Roijers, Tom Lenaerts, Ann Nowé, and Denis Steckelmacher. Dynamic weights in multi-objective deep reinforcement learning. In *International conference on machine learning*, pages 11–20. PMLR, 2019.
- Toygun Basaklar, Suat Gumussoy, and Umit Y Ogras. Pd-morl: Preference-driven multi-objective reinforcement learning algorithm. *arXiv preprint arXiv:2208.07914*, 2022.

- Baiting Zhu, Meihua Dang, and Aditya Grover. Scaling pareto-efficient decision making via offline multi-objective rl. *arXiv preprint arXiv:2305.00567*, 2023.
- Qian Lin, Chao Yu, Zongkai Liu, and Zifan Wu. Policy-regularized offline multi-objective reinforcement learning. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*, 2024.
- Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Misha Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. Decision transformer: Reinforcement learning via sequence modeling. *Advances in neural information processing systems*, 34:15084–15097, 2021.
- Scott Emmons, Benjamin Eysenbach, Ilya Kostrikov, and Sergey Levine. Rvs: What is essential for offline rl via supervised learning? *arXiv preprint arXiv:2112.10751*, 2021.
- Scott Fujimoto and Shixiang Shane Gu. A minimalist approach to offline reinforcement learning. *Advances in neural information processing systems*, 34:20132–20145, 2021.
- Zhendong Wang, Jonathan J Hunt, and Mingyuan Zhou. Diffusion policies as an expressive policy class for offline reinforcement learning. *arXiv preprint arXiv:2208.06193*, 2022.
- Adam Stooke, Joshua Achiam, and Pieter Abbeel. Responsive safety in reinforcement learning by pid lagrangian methods. In *International Conference on Machine Learning*, pages 9133–9143. PMLR, 2020.
- Yinlam Chow, Mohammad Ghavamzadeh, Lucas Janson, and Marco Pavone. Risk-constrained reinforcement learning with percentile risk criteria. *The Journal of Machine Learning Research*, 18(1):6070–6120, 2017.
- Tengyu Xu, Yingbin Liang, and Guanghui Lan. Crpo: A new approach for safe reinforcement learning with convergence guarantee. In *International Conference on Machine Learning*, pages 11480–11491. PMLR, 2021.
- Aivar Sootla, Alexander I Cowen-Rivers, Taher Jafferjee, Ziyang Wang, David H Mguni, Jun Wang, and Haitham Ammar. Sauté rl: Almost surely safe reinforcement learning using state augmentation. In *International Conference on Machine Learning*, pages 20423–20443. PMLR, 2022.
- Yihang Yao, Zuxin Liu, Zhepeng Cen, Jiacheng Zhu, Wenhao Yu, Tingnan Zhang, and Ding Zhao. Constraint-conditioned policy optimization for versatile safe reinforcement learning. *Advances in Neural Information Processing Systems*, 36, 2024a.
- Zuxin Liu, Zijian Guo, Yihang Yao, Zhepeng Cen, Wenhao Yu, Tingnan Zhang, and Ding Zhao. Constrained decision transformer for offline safe reinforcement learning. *arXiv preprint arXiv:2302.07351*, 2023a.
- Qian Lin, Bo Tang, Zifan Wu, Chao Yu, Shangqin Mao, Qianlong Xie, Xingxing Wang, and Dong Wang. Safe offline reinforcement learning with real-time budget constraints. In *International Conference on Machine Learning*, pages 21127–21152. PMLR, 2023.
- Xi Chen, Ali Ghadirzadeh, Mårten Björkman, and Patric Jensfelt. Meta-learning for multi-objective reinforcement learning. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 977–983. IEEE, 2019.
- Cong Guan, Ruiqi Xue, Ziqian Zhang, Lihe Li, Yi-Chen Li, Lei Yuan, and Yang Yu. Cost-aware offline safe meta reinforcement learning with robust in-distribution online task adaptation. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*, pages 743–751, 2024a.
- Dibya Ghosh, Anurag Ajay, Pulkit Agrawal, and Sergey Levine. Offline rl policies should be trained to be adaptive. In *International Conference on Machine Learning*, pages 7513–7530. PMLR, 2022.
- Joey Hong, Aviral Kumar, and Sergey Levine. Confidence-conditioned value functions for offline reinforcement learning. *arXiv preprint arXiv:2212.04607*, 2022.

- Eric Mitchell, Rafael Rafailov, Xue Bin Peng, Sergey Levine, and Chelsea Finn. Offline meta-reinforcement learning with advantage weighting. In *International Conference on Machine Learning*, pages 7780–7791. PMLR, 2021.
- Mengdi Xu, Yikang Shen, Shun Zhang, Yuchen Lu, Ding Zhao, Joshua Tenenbaum, and Chuang Gan. Prompting decision transformer for few-shot policy generalization. In *international conference on machine learning*, pages 24631–24645. PMLR, 2022a.
- Sandy Huang, Abbas Abdolmaleki, Giulia Vezzani, Philemon Brakel, Daniel J Mankowitz, Michael Neunert, Steven Bohez, Yuval Tassa, Nicolas Heess, Martin Riedmiller, et al. A constrained multi-objective reinforcement learning framework. In *Conference on Robot Learning*, pages 883–893. PMLR, 2022.
- Santiago Paternain, Luiz Chamon, Miguel Calvo-Fullana, and Alejandro Ribeiro. Constrained reinforcement learning has zero duality gap. *Advances in Neural Information Processing Systems*, 32, 2019.
- Zuxin Liu, Zijian Guo, Haohong Lin, Yihang Yao, Jiacheng Zhu, Zhepeng Cen, Hanjiang Hu, Wenhao Yu, Tingnan Zhang, Jie Tan, et al. Datasets and benchmarks for offline safe reinforcement learning. *arXiv preprint arXiv:2306.09303*, 2023b.
- Sven Gronauer. Bullet-safety-gym: A framework for constrained reinforcement learning. Technical report, mediaTUM, 2022.
- Yihang Yao, Zuxin Liu, Zhepeng Cen, Peide Huang, Tingnan Zhang, Wenhao Yu, and Ding Zhao. Gradient shaping for multi-constraint safe reinforcement learning. In *6th Annual Learning for Dynamics & Control Conference*, pages 25–39. PMLR, 2024b.
- Dohyeong Kim, Kyungjae Lee, and Songhwa Oh. Trust region-based safe distributional reinforcement learning for multiple constraints. *Advances in neural information processing systems*, 36, 2024.
- Jiayi Guan, Li Shen, Ao Zhou, Lusong Li, Han Hu, Xiaodong He, Guang Chen, and Changjun Jiang. Poce: Primal policy optimization with conservative estimation for multi-constraint offline reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26243–26253, 2024b.
- Saurabh Arora and Prashant Doshi. A survey of inverse reinforcement learning: Challenges, methods and progress. *Artificial Intelligence*, 297:103500, 2021.
- Naoya Takayama and Sachiyo Arai. Multi-objective deep inverse reinforcement learning for weight estimation of objectives. *Artificial Life and Robotics*, 27(3):594–602, 2022.
- Shehryar Malik, Usman Anwar, Alireza Aghasi, and Ali Ahmed. Inverse constrained reinforcement learning. In *International conference on machine learning*, pages 7390–7399. PMLR, 2021.
- Mohammad Ghavamzadeh, Shie Mannor, Joelle Pineau, Aviv Tamar, et al. Bayesian reinforcement learning: A survey. *Foundations and Trends® in Machine Learning*, 8(5-6):359–483, 2015.
- Haoran Xu, Xianyuan Zhan, Honglei Yin, and Huiling Qin. Discriminator-weighted offline imitation learning from suboptimal demonstrations. In *International Conference on Machine Learning*, pages 24725–24742. PMLR, 2022b.
- Geon-Hyeong Kim, Seokin Seo, Jongmin Lee, Wonseok Jeon, HyeongJoo Hwang, Hongseok Yang, and Kee-Eung Kim. Demodice: Offline imitation learning with supplementary imperfect demonstrations. In *International Conference on Learning Representations*, 2022.
- Bingyi Kang, Xiao Ma, Chao Du, Tianyu Pang, and Shuicheng Yan. Efficient diffusion policies for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 36, 2024.

A Appendix / supplemental material

A.1 Connections and Differences to Other Fields

Multi-task RL While the constrained MORL setting we focus on can align with the multi-task framework by treating policy learning under various preferences or thresholds as distinct tasks, several key differences distinguish our study from previous multi-task RL settings: (1) One primary motivation of our work is to circumvent the handcrafted design of target preferences, whereas previous studies [Mitchell et al., 2021, Xu et al., 2022a] on multi-task RL mainly focus on policy generalization to new tasks. (2) We achieve offline adaptation by updating the target preference distribution rather than the policies themselves, which distinguishes our framework from offline meta-RL methods. This adaptation pipeline not only avoids potential performance degradation due to policy parameter shifts but also allows us to incorporate conservatism to mitigate constraint violations. (3) Accurately identifying various preferences and constraints implied in demonstrations during adaptation is challenging as these demonstrations could come from the same dynamics and policies with similar preferences.

Multi-constraint RL Existing research on multi-constraint RL typically explores the correlation between multiple constraints from gradient perspectives [Yao et al., 2024b, Kim et al., 2024]. Some works [Guan et al., 2024b] attempt to meet different types of constraints simultaneously. In contrast, our approach, from the policy diversity perspective, aims to find appropriate policies for various safety requirements and thus places less focus on constraint correlation for a specific threshold. Nevertheless, previous studies on multiple constraint RL will provide important insights into managing potential correlations between multiple objectives during preference-conditioned policy learning, thereby improving overall performance.

Inverse RL Existing IRL methods [Arora and Doshi, 2021] typically assume no knowledge about the reward function and focus on learning reward functions from mere demonstrations through apprenticeship learning, maximum entropy optimization or adversarial learning. Some of these methods have been applied in MORL [Takayama and Arai, 2022] or safe RL [Malik et al., 2021]. In contrast to these studies, our work assumes knowledge of the reward information on different objectives and pays more attention to identifying the preferences of demonstrations during deployment than learning a specific decisive reward during training.

Bayesian RL Bayesian RL [Ghavamzadeh et al., 2015] is widely used for promoting online exploration and dealing with uncertainty in model-based RL or offline RL, while it serves as a cornerstone of our theoretical derivation, allowing us to incorporate prior preference information from datasets and infer the expected preferences by estimating the preference posterior based on a few demonstrations. Ghosh’s work [Ghosh et al., 2022] is the most related Bayesian RL study that inspires our approach. Nevertheless, it aims to reduce the dynamics uncertainty caused by limited offline samples, while our approach focuses on utilizing Bayesian inference to solve the uncertainty of the target preference distribution based on a few demonstrations.

Offline Imitation Learning Each adaptation can be viewed as the imitation of a expert policy with a certain preference. Therefore, offline Imitation Learning (IL) is a potential approach to solving the problem by applying an off-the-shelf offline IL algorithm [Xu et al., 2022b, Kim et al., 2022], which enables the combination of sub-optimal trajectories in the offline dataset with expert demonstrations provided by users to imitate expert behaviors. However, applying offline IL to our problem presents several challenges. We conduct comparative experiments in Appendix A.8 to further compare algorithms and illustrate these challenges.

A.2 Environments and Datasets Details

Environments and Datasets for Multi-objective RL We utilize the D4MORL dataset [Zhu et al., 2023] collected from multi-objective MuJoCo environments by a set of preference-varying policies [Xu et al., 2020]. We consider 5 MuJoCo tasks including HalfCheetah, Ant, Hopper, Walker2d, and Swimmer. Among them, the HalfCheetah, Walker2d, and Swimmer tasks involve

conflicting objectives of running speed and energy saving. The Ant task includes two speed objectives in both the x and y axes, while the Hopper considers running and jumping objectives. For each task, we test algorithms on two types of datasets: Expert and Amateur, which are collected by pure expert policies and noise-injected expert policies, respectively.

Environments and Datasets for Safe RL We utilize the datasets in DSRL benchmark [Liu et al., 2023b] that are collected by a set of behavior policies trained under various safe thresholds. We select 8 BulletSafetyGym tasks [Gronauer, 2022] for safe RL experiments, which involve 4 agents (Ball, Car, Drone, Ant) and 2 task goals (Run, Circle). In these tasks, the goal of the reward function is to navigate to target positions or complete circuits as quickly as possible, while the costs are incurred when the agent enters risky areas.

Environments and Datasets for Constrained Multi-objective RL We develop a set of CMORL tasks, namely CMO-Hopper, CMO-HalfCheetah, CMO-Ant and CMO-Swimmer. These tasks are constructed based on the multi-objective MuJoCo environments [Xu et al., 2020] and augmented by an additional constraint on robotics’ velocity, which is consistent with the constraint setting of Safety Gymnasium tasks in the DSRL benchmark. When the agent’s moving velocity exceeds a specific value, it will receive 1 cost. To gather datasets for these tasks, we create a behavioral preference set and utilize TRPO+Lagrangian [Liu et al., 2023b] to train individual policies with the goal of maximizing utility under each preference in this set while accounting for a specified velocity constraint. The safety threshold of the velocity constraint is adjusted throughout the training process to encompass a range of potential safety thresholds. The details for each CMO task are shown in Table 1. The replay buffers collected during the training of these policies constitute the CMO datasets. We present Pareto fronts of CMO datasets under different cost return intervals in Figure 7, illustrating that CMO datasets encompass a wide range of behaviors catering for various preferences and safety thresholds.

Table 1: Details of each CMO task.

Details	CMO-Ant	CMO-Hopper	CMO-HalfCheetah	CMO-Swimmer
behavioral preference Set	[0.5, 0.5], [0.6, 0.4], [0.7, 0.3], [0.8, 0.2], [0.9, 0.1], [1.0, 0.0]			
Cost function	Moving Velocity (cost=1 if the velocity exceeds a specific value)			
Threshold Range	[10, 250]			[10, 200]
Objectives	Speed on X-axis and Y-axis	Moving speed and jumping height	Moving speed and energy consumption	

A.3 Construction of Training Set and Demonstration Set

(1) For MORL experiments in Section 5.1, each target preference ω_g is associated with a demonstration set \mathcal{B}_{ω_g} , which contains M transitions² randomly sampled from K trajectories in D4MORL Expert datasets with behavioral preferences closest to the target preference ω_g . (2) For safe RL experiments in Section 5.2, each target safety threshold β_g is associated with a demonstration set \mathcal{B}_{β_g} , which contains M transitions randomly sampled from K trajectories in DSRL datasets with the highest utility among all safe trajectories (i.e., these trajectories’ cumulative cost return is less than β). (3) For CMORL experiments in Section 5.3, each combination of target preference ω_g and safety threshold β_g is associated with a demonstration set $\mathcal{B}_{\omega_g, \beta_g}$, which consists of M transitions randomly sampled from K trajectories in CMO datasets that (i) have highest utility among all safe trajectories and (ii) originate from the replay buffer of the behavior policy with preference ω_g . The unselected trajectories in datasets constitute the training set. In our experiments, $M = 128$ and $K = 2$ by default.

²For Prompt-MODT, which required trajectory segments as prompts during adaptation, we sample a trajectory segment with length= M to constitute the demonstration set.

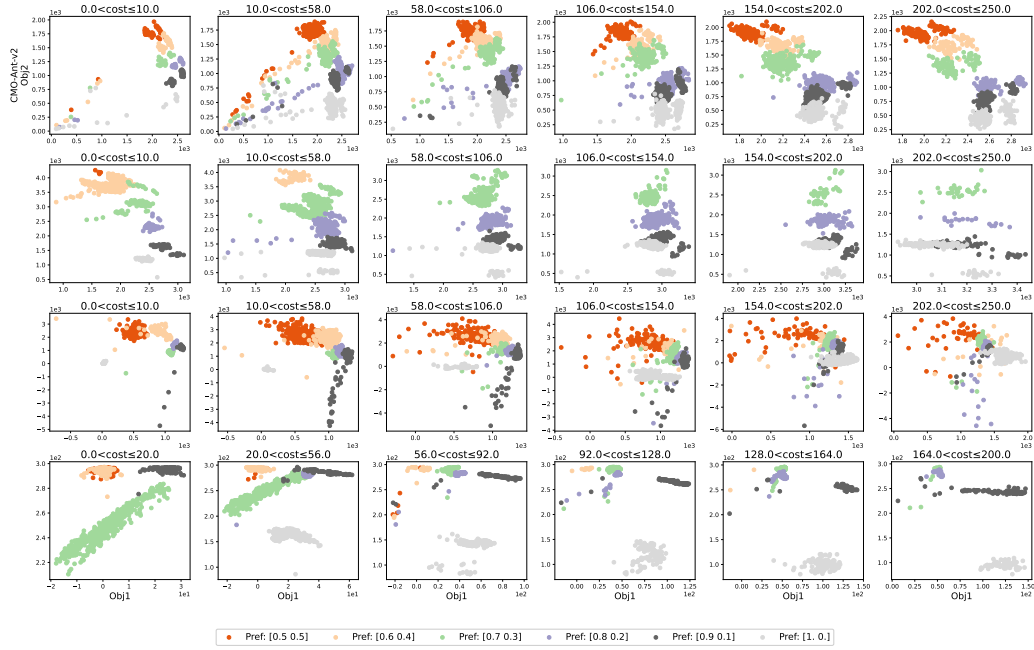


Figure 7: The reward vector distribution of CMO datasets under various safety thresholds.

A.4 Evaluation Metrics

We denote $R(\pi_G)$ as the estimated return vector of the policy π_G adapted for target G in the target set \mathcal{T} .

Average Utility For unconstrained scenarios, $G = \omega_g$ and average utility is calculated by:

$$U = \frac{1}{|\mathcal{T}|} \sum_{\omega_g \in \mathcal{T}} \omega_g^\top R(\pi_{\omega_g}). \quad (11)$$

For constrained scenarios, $G = (\omega_g, \beta_g)$. We denote $\mathcal{T}_{\beta_g} = \{\omega_g | (\omega_g, \beta_g) \in \mathcal{T}\}$ as subset of \mathcal{T} that groups all target preference ω_g by safety threshold β_g . The average utility is a function of β_g :

$$U(\beta_g) = \frac{1}{|\mathcal{T}_{\beta_g}|} \sum_{\omega_g \in \mathcal{T}_{\beta_g}} \omega_g^\top R(\pi_{\omega_g, \beta_g}). \quad (12)$$

This metric reflects the overall performance of the adapted policies across all target preferences.

Hypervolume We first consider the unconstrained settings and introduce the adapted Pareto set P that contains all adapted policies that are not dominated by any other adapted policy. Hypervolume measures the volume enclosed by the returns of policies in the adapted Pareto set P :

$$\text{HV} = \int_{R^n} \mathbb{1}_{H(P, r_0)}(z) dz, \quad (13)$$

where $H(P, r_0) = \{z \in R^n | \exists \pi \in P : r_0 \preceq z \preceq R(\pi)\}$ and $\mathbb{1}_{H(P, r_0)}$ equals 1 if $z \in H(P)$ and 0 otherwise. For constrained scenarios, we group all target preferences by safety threshold and report Hypervolume on each group. The Hypervolume reflects not only the overall performance across all target preferences but also the diversity of the policies, as a high diversity of policies corresponds to a broad Pareto front and consequently a high Hypervolume.

Sparsity, which measures the density of policies in the approximated Pareto set, is also commonly used for evaluating MORL performance. A lower sparsity indicates a denser Pareto front, which

Algorithm 1 Preference Distribution Offline Adaptation

```
1: ### Training Phase
2: Input: Offline dataset  $\mathcal{D}$ 
3: Train policy-conditioned policy  $\hat{\pi}_{\omega}^*(a|s, \omega)$  or value  $Q(s, a, \omega)$  using MODF or PEDA
4: Approximate the prior preference distribution  $P(\omega|\mathcal{D})$ 
5:
6: ### Adaptation Phase
7: Input: Demonstration set  $\mathcal{B}$ , and conservatism parameter  $\alpha$ 
8: Initialize  $\mathcal{N}(\mu, \sigma I)$  based on  $P(\omega|\mathcal{D})$ 
9: for each interaction do
10:   Sample preference  $\omega$  from  $\mathcal{N}(\mu, \sigma I)$ 
11:   Update  $\mu, \sigma$  through gradient descent (Eq. (7))
12: end for
13: if no constraint then
14:    $\tilde{\omega}_a = \mu / |\mu|$ 
15: else
16:   Compute  $\mathbf{b}$  based on Eq. (10)
17:    $\tilde{\omega}_a = \mathbf{b} / |\mathbf{b}|$ 
18: end if
19: Output: Policy  $\hat{\pi}_{\omega}^*(\cdot|\cdot, \tilde{\omega}_a)$  for decision making
```

is generally more desirable. However, the sparsity metric is not applicable in our setting. This is because, given that the number of adapted policies is fixed, an increase in utility and Hypervolume performance can result in a worse sparsity metric. Moreover, better sparsity metrics encourage a lower diversity of adapted policies, which diverges from our intended goal.

A.5 Implementation Details

The pseudocode for the training and adaptation process is provided in Algorithm 1.

Behavioral Preference Approximation Scheme in Constrained Settings In constrained settings, the vector reward is defined as $\tilde{\mathbf{r}} = [\mathbf{r}_t, -\mathbf{c}_t]$ and dataset \mathcal{D} is augmented to $\hat{\mathcal{D}}$. For behavioral preferences in $\hat{\mathcal{D}}$, we utilize $\tilde{\omega} = U(\tau) / \|U(\tau)\|_1$ as behavioral preferences for all transitions in trajectory τ . Here, $U(\tau) = [\mathbf{R}_1, \dots, \mathbf{R}_N, C_1^{\max} - C_1, \dots, C_K^{\max} - C_K]$, where \mathbf{R}_i and C_i represent the accumulations of the i^{th} reward and cost of τ , respectively, and C_i^{\max} is the maximum of C_i among all trajectories. The formulation of $\tilde{\omega}$ aims to approximate the normal vector of the Pareto front that consists of $(\mathbf{R}_1, \dots, \mathbf{R}_N, -C_1, \dots, -C_K)$.

Preference Distribution Offline Adaptation (PDOA) We utilize the policy and value function obtained by MODF and PEDA to calculate the TD reward and action likelihood reward in Eq. (7). For MODF, except the state-action value function $Q(s, a, \omega) = Q_1(s, a, \omega), \dots, Q_N(s, a, \omega)$ obtained during training, we additionally train a state value function $V(s, \omega) = \mathbb{E}_{\pi(a|s, \omega)}[Q(s, a, \omega)]$ to derive TD reward. As for the action likelihood reward, the action likelihood of the diffusion policy model is intractable but can be approximated as $\log \hat{\pi}_{\omega}^*(a|s, \omega) \approx \mathbb{E}_{i \sim \mathcal{U}, \epsilon \sim \mathcal{N}(\mathbf{0}, I)} \|a - \hat{a}^0\|_2^2$ [Kang et al., 2024], where \hat{a}^0 is reconstruction of a^i and a^i is obtained by corrupting a with noise ϵ . For PEDA, which does not involve value functions, we disregard the TD reward when calculating the adaptation loss. The prior preference distribution $P(\omega|\mathcal{D})$ in Eq. (7) is approximated with a Gaussian distribution $\mathcal{N}(\mu_{\mathcal{D}}, \sigma_{\mathcal{D}})$, where the expectation $\mu_{\mathcal{D}}$ and standard deviation $\sigma_{\mathcal{D}}$ are estimated on behavioral preferences of the training dataset \mathcal{D} . The weight η of the regularization term $(\|\mu\|_1 - 1)^2$ in Eq. (7) is set to 1.0.

During adaptation, we model the preference distribution using a Gaussian model and perform gradient updates on this distribution according to Eq. (7). For each target, the number of gradient updates is set to 1000, with 64 preferences sampled from the distribution for each gradient update. All samples in the demonstration set are used for gradient updates within one batch. We use the Adam optimizer with a learning rate of 0.05. The conservatism weight α in Eq. (10) is set to 1.0 for MORL tasks

and 0.7 for safe RL and CMORL tasks. The weight of the TD reward in Eq. (6) is set to 0.01 for PDOA [MODF].

MORL and Safe RL Algorithms For **MODF**, we use the original implementation in <https://github.com/qianlin04/PRMORL>. We remove the Regularization Weight Adaptation component in the original MODF since it requires additional online interactions during evaluation. We follow the default hyperparameters except for the regularization weight, which is set to 200 for MORL and CMORL tasks and 20 for safe RL tasks. For **PEDA**, we use the original implementation in <https://github.com/baitingzbt/PEDA>. PEDA utilizes a linear regression model to predict target vector returns given target preferences. For D4MORL datasets, it fits this model using Expert datasets. In DSRL and CMO datasets, which contain suboptimal samples, we construct a Pareto set that consists of all near-undominated trajectories with a small tolerance (5%) and use it to fit the linear regression model. For **Prompt-MODT**, we implement it based on the MODT algorithm in PEDA by incorporating prompts into the decision transformer. The target vectors predicted by the above linear regression model are used as the initialized return-to-go for Prompt-MODT. Besides, to apply Prompt-MODT, we need to transform the multi-objective problem into a multi-task problem, for which we divide trajectories with different behavioral preferences into separate tasks with a preference interval of 0.05. For example, trajectories with behavior preferences ranging from [0.00, 1.00] to [0.05, 0.95] are classified as task 1, and trajectories with behavior preferences ranging from [0.05, 0.95] to [0.10, 0.90] are classified as task 2. For **BC-Finetune**, we model the policy using a Gaussian model and employ behavioral cloning through MSE loss. During adaptation, behavior cloning is performed on the demonstration set with 0.01 learning rate and 1000 gradient update steps.

Computer Resources The training and testing were conducted on 1 NVIDIA GeForce RTX 3090 GPU. The testing utilizes 5 CPU threads to simultaneously collect test data from multiple environments. The total time for training and testing does not exceed 10 hours. The memory usage for a single run depends on the size of the dataset used, but generally does not exceed 10 GB.

A.6 Additional Results

A.6.1 Experimental Results on D4MORL Expert Datasets for MORL

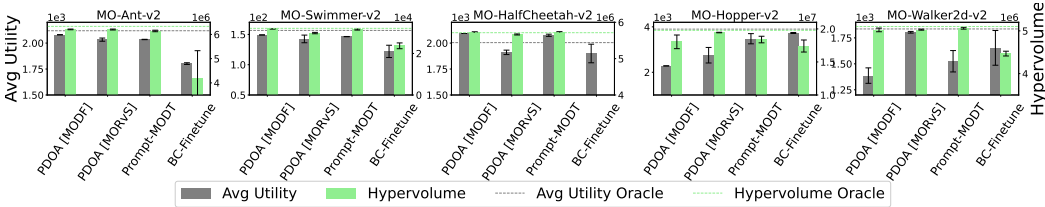


Figure 8: Average utility and Hypervolume performance of all algorithms on D4MORL Expert datasets.

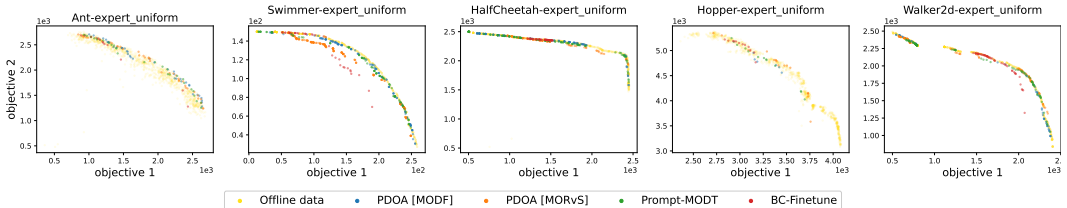


Figure 9: Pareto fronts of different algorithms on D4MORL Expert datasets. Each point represents the expected vector return of an adapted policy for a specific unknown target preference.

We present the average utility and Hypervolume performance, Pareto fronts, and the difference between real target preferences and adapted preferences on the D4MORL Expert datasets in Figure 8,

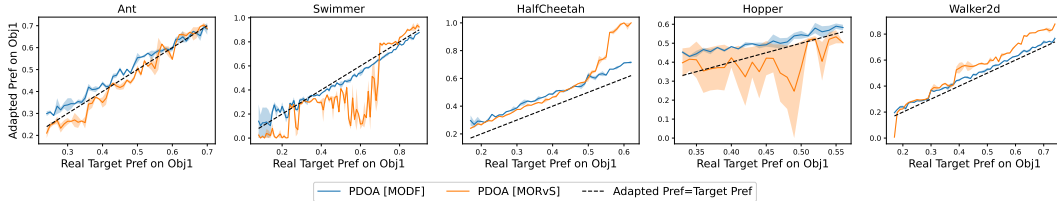


Figure 10: The comparison between real target preferences and adapted preferences on D4MORL Expert datasets.

9, and 10, respectively. In Figure 8, our method demonstrates overall competitive performance in terms of average utility and Hypervolume metrics. Figure 9 illustrates that our method achieves a broader Pareto front, indicating a high diversity of adapted policies. Figure 10 shows that our method can obtain expected behaviors due to the consistency between our adapted preferences and real target preferences. These results are consistent with those obtained on the Amateur Datasets in Section 5.1, which demonstrates that our framework achieves consistent advantages across datasets of varying quality.

A.7 Experimental Results on Additional DSRL Tasks for Safe RL

Figures 11 and 12 show the performance of MODF and MORvS under various target preferences, which demonstrate the capability of these MORL algorithms in obtaining policies that satisfy different safety thresholds. Figure 13 illustrates the normalized cost and reward for additional DSRL tasks, including CarCircle, DroneCircle, CarRun, and DroneRun. The results are consistent with those in Section 5.2, demonstrating that MORL methods can effectively learn a set of policies that meet various safety thresholds, and that PDOA [MODF] can achieve very few constraint violations.

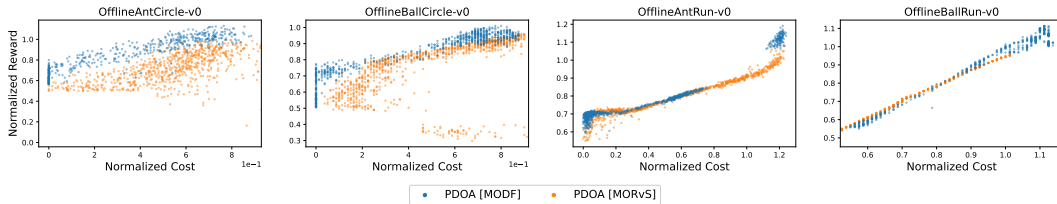


Figure 11: The normalized cost and normalized reward of policies with various preferences obtained by our framework.

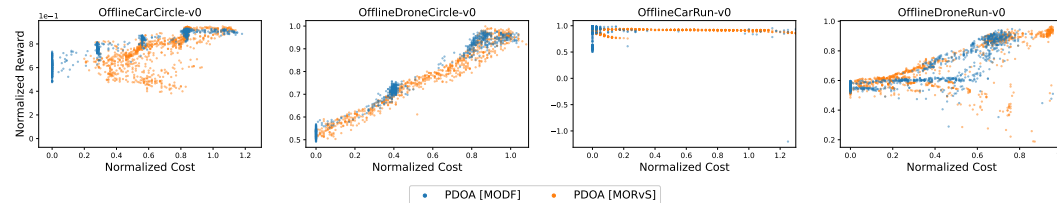


Figure 12: The normalized cost and normalized reward of policies with various preferences obtained by our framework on additional tasks.

A.8 Comparison with Offline Imitation Learning Baseline

Applying offline IL to our problem faces several challenges: 1) Offline IL requires complete policy training for each adaptation, which is resource-intensive and time-consuming. 2) Our experiments

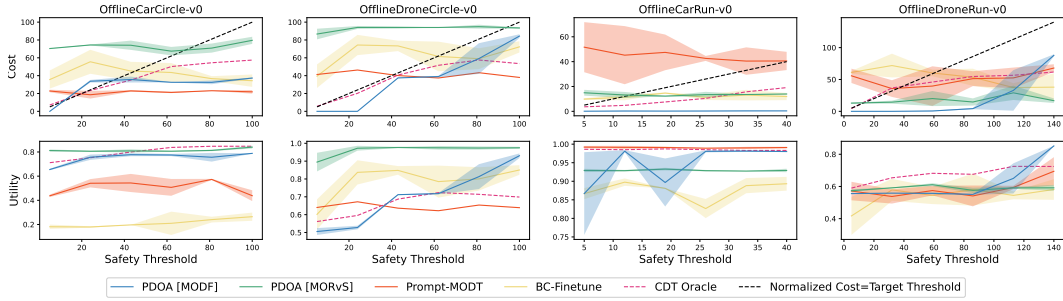


Figure 13: The adapted policies’ normalized cost and normalized reward of all algorithms under various safety thresholds on additional tasks.

involve a limited number of demonstrations (128 transitions per target preference) compared to the millions of samples in the offline dataset, leading to significant demo sufficiency and data imbalance.

We apply an OIL baseline DWBC [Xu et al., 2022b] in our MORL and safe RL experiments and present the results in Figure 14 and 15, where we limit the number of tested preferences to 10 in the D4MORL tasks because the original evaluation involves dozens of target preferences per task, and training a policy for each target preference using DWBC would be too time-consuming. In both MORL and safe RL environments, DWBC demonstrates unstable performance. It is competitive in tasks such as MO-Ant, MO-Swimmer, but performs poorly in other tasks, including MO-Walker2d. It even struggles to learn effective policies in some cases, such as MO-Hopper and AntCircle. These issues can be attributed to demo sufficiency and data imbalance. Additionally, DWBC experiences high constraint violations in most safe RL tasks, as it does not incorporate safety considerations into its policy learning.

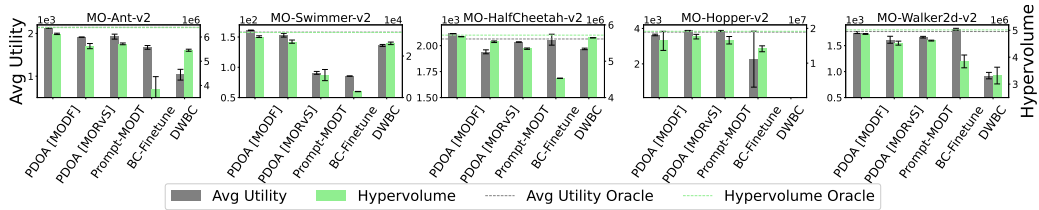


Figure 14: Comparative results with offline IL baseline on MORL tasks.

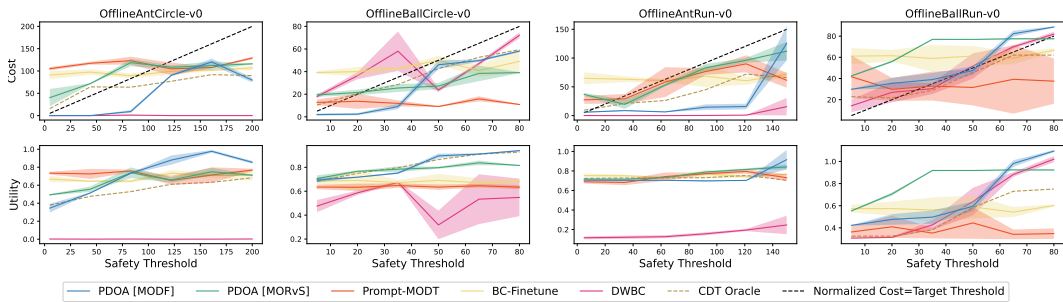


Figure 15: Comparative results with offline IL baseline on safe tasks.

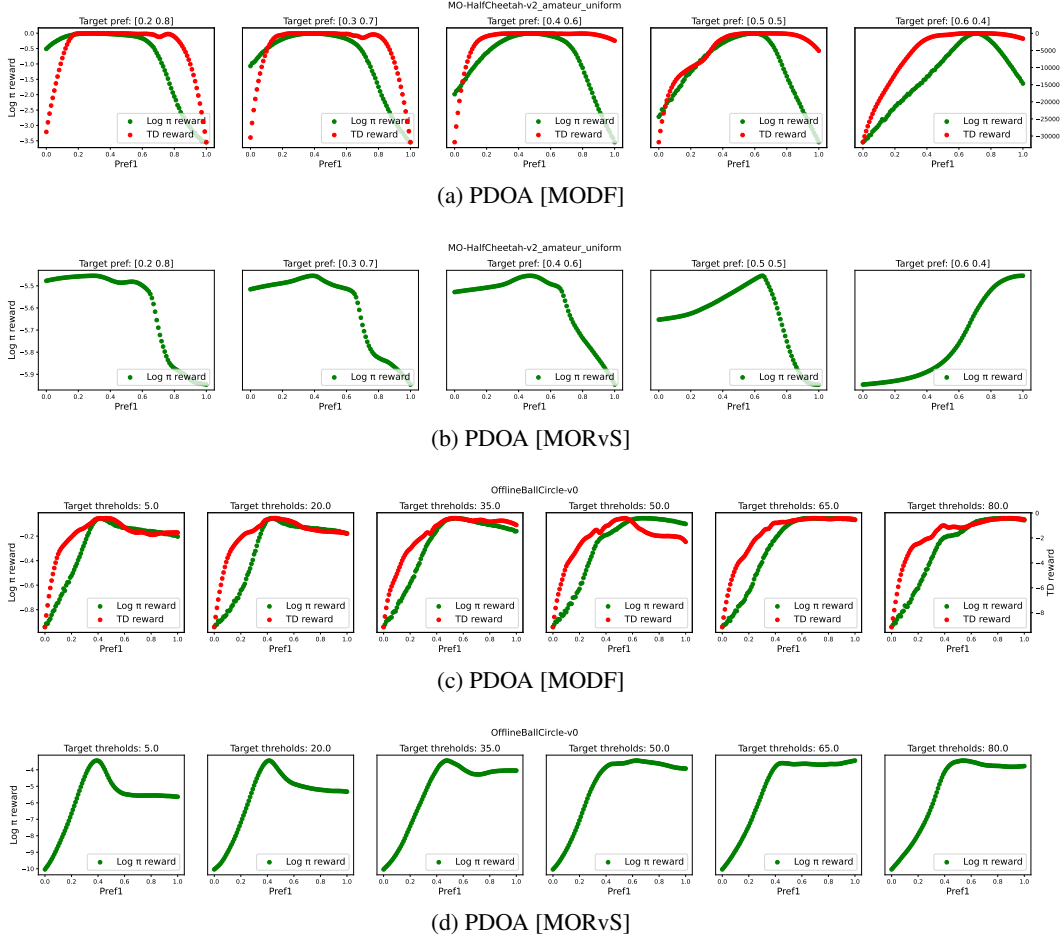


Figure 16: The relationship between TD reward, action likelihood reward, and preference. Pref1 represents the first dimension of preference.

A.9 Relationship between TD Reward, Action Likelihood Reward and Preference

For each target preference, we traverse all possible adapted preferences with a small interval and compute the corresponding TD reward and action likelihood reward. The results are presented in Figure 16, where we observe that the adapted preferences corresponding to the highest TD reward and action likelihood reward align with the target preferences. Additionally, as constraints are relaxed, the preference weights allocated to the constrained targets decrease. This demonstrates that the TD reward and action likelihood reward in our algorithm are strongly consistent with the target preferences and safety thresholds.

A.10 Licenses

- D4MORL dataset and PEDDA code [Zhu et al., 2023]: The MIT License, <https://github.com/baitingzbt/PEDA>
- Datasets and codes of DSRL, OSRL and FSRL [Liu et al., 2023b]: All datasets are licensed under the Creative Commons Attribution 4.0 License (CC BY 4.0), and code is licensed under the Apache 2.0 License. <https://github.com/liuzuxin/DSRL>, <https://github.com/liuzuxin/OSRL>, <https://github.com/liuzuxin/FSRL>
- MuJoCo: Apache 2.0 License, <https://github.com/google-deepmind/mujoco>

A.11 Broader Impacts

Our research introduces a novel framework for addressing constrained multi-objective reinforcement learning (CMORL), which is widely applicable to many real-world scenarios, such as autonomous driving, healthcare, where behavioral preferences and safety criteria are difficult to define precisely. These approaches liberate researchers from the labor required to align preferences and safety thresholds. Besides, the offline training paradigm eliminates the expenses and dangers associated with online exploration. For instance, in autonomous driving, online interactions with the environment have the risk of accidents and injuries; however, our method can mitigate this risk by learning from a pre-recorded driving dataset generated by safe and target behavior policies.

Nonetheless, our method may not be suitable for domains requiring rapid responses and frequent updates due to the additional computational load and latency caused by gradient updates during adaptation. Besides, providing personalized services with this framework necessitates a certain amount of user data as demonstrations, raising concerns about privacy breaches and data misuse. Despite the risks and challenges mentioned, we believe that CMORL holds great promise for automating and enhancing sequential decision-making in highly impactful domains. Additional work is required to make this framework robust enough for application in multi-objective and safety-critical scenarios.

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 precede 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.**

temp

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 main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope.

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 paper discusses the limitations of the work, which can be found in Section 6.

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: The derivation of the formula is clearly presented in the paper.

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 paper fully discloses all the information needed to reproduce the main experimental results of the paper. Details of our framework and implementation can be found in Section 4, 5 and Appendix A.2, A.3, A.4, A.5.

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: Codes and instructions are provided in supplemental material to generate the dataset used and reproduce the main results in the paper.

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, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: All the training and testing details necessary to understand the results can be found in Section 5 and Appendix A.2, A.3, A.4 and A.5.

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 paper reports error bars suitably.

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: We present the information of compute resources used in this paper in Appendix A.5.

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, in every respect, with the NeurIPS 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 paper discusses both potential positive societal impacts and negative societal impacts of the work performed, which can be found in Section A.11.

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 paper poses no 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: We present the license and URL of assets used in this paper in Appendix A.10.

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 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 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 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.