# AETDICE: Unified Framework and Offline Optimization for Nonlinear Multi-Objective RL

**Woosung Kim**[1*] **Youngjun Suh**[1*] **Jinho Lee**[1,4*] **Jongmin Lee**[2†] **Byung-Jun Lee**[1,3†]

[1]Korea University [2]Yonsei University [3]Gauss Labs Inc. [4]CMES Robotics Inc.
{wsk208,youngjunsuh,jinho0997,byungjunlee}@korea.ac.kr
jongminlee@yonsei.ac.kr
[*]Equal contribution [†]Corresponding authors

## Abstract

Optimizing nonlinear preferences in multi-objective reinforcement learning (MORL) captures complex trade-offs such as fairness or efficiency. However, the order of scalarization and expectation divides nonlinear MORL objectives into two distinct paradigms: Scalarized Expected Return (SER) and Expected Scalarized Return (ESR). SER requires global-level optimization and ESR requires non-Markovian policies, and the two have been addressed by separate optimization strategies. We unify them through the Aggregation–Expectation–Transformation (AET) framework. Through a tripartite decomposition of scalarization, AET recovers SER and ESR as special cases and also admits objectives that combine both levels of nonlinearity. Building on this framework, we propose AETDICE, a tractable offline RL algorithm for AET objectives. AETDICE absorbs the trajectory-level nonlinearity into a transformed per-step reward on an augmented state space and applies finite-horizon DICE-style density-ratio estimation, enabling sample-based optimization from static datasets. AETDICE addresses both requirements within a single method and captures the distinct trade-offs induced by AET objectives, including those that combine both levels of nonlinearity, which existing approaches do not address.

## 1 Introduction

Multi-objective reinforcement learning (MORL) optimizes sequential decisions under multiple, often competing objectives [1, 2, 3]. The trade-off among objectives is governed by a scalarization function that maps vector-valued returns to a scalar objective. Linear scalarization reduces to standard RL via a weighted sum, while nonlinear scalarization captures richer preferences such as diminishing returns, fairness, and efficiency [4, 5].

In nonlinear MORL, two canonical criteria arise from the ordering of scalarization and expectation: the Scalarized Expected Return (SER), which applies nonlinearity to the expected return, and the Expected Scalarized Return (ESR), which applies nonlinearity to each trajectory's return. SER and ESR represent different decision-making contexts with fundamentally different optimal policy structures, and existing methods have accordingly been developed separately for each [6, 7, 8].

We focus on the offline setting, where policies must be learned from a fixed dataset. Offline methods exist for linear scalarization [9, 10] and SER [11], but ESR remains unaddressed: trajectory-level nonlinearity makes the optimal action depend on previously accumulated rewards, and learning such reward-conditioned policies from fixed data has not been explored. Moreover, when multiple nonlinear utilities are applied to each trajectory's return vector and their expectations are aggregated linearly, lower-performing utilities can be neglected—the same imbalance that nonlinear aggregation

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

corrects in SER. Expectation-level nonlinear aggregation can counteract this imbalance, but no existing method optimizes objectives that combine trajectory-level and expectation-level nonlinearity.

We address both gaps through the Aggregation–Expectation–Transformation (AET) framework, which decomposes nonlinear scalarization into a trajectory-level transformation ($F$), expectation, and an expectation-level aggregation ($G$). AET subsumes SER and ESR as special cases while enabling a richer class of objectives that combine both levels of nonlinearity. To optimize AET objectives offline, we introduce AETDICE—offline *AET* optimization via *DI*stribution *C*orrection *E*stimation. We propose a transformed reward that absorbs nonlinear $F$ into per-step rewards on a state space augmented with the accumulated reward, which reduces offline ESR to single-objective offline RL, and develop a finite-horizon DICE formulation over this augmented space to handle the global-level optimization required by concave $G$. Empirically, AETDICE optimizes ESR, SER, and novel AET objectives within a single framework, revealing distinct policy behaviors that arise from each level of nonlinearity and their combination.

## 2 Preliminaries

### 2.1 Finite-horizon MOMDPs and MORL objectives

We consider a finite-horizon multi-objective Markov Decision Process (MOMDP) $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, H, p_0, \mathbf{r})$, where $\mathcal{S}$ and $\mathcal{A}$ denote the state and action spaces, $P(s' \mid s, a)$ is a time-homogeneous transition kernel, $H$ is the horizon, $p_0(s)$ is the initial-state distribution, and $\mathbf{r}(s, a) \in \mathbb{R}^m$ specifies an $m$-dimensional reward vector for $m$ objectives. A time-dependent policy $\pi(a|s, t)$ induces a trajectory distribution over $\tau = (s_0, a_0, \ldots, s_{H-1}, a_{H-1})$, and the cumulative multi-objective return is given by $\mathbf{R}(\tau) = \sum_{t=0}^{H-1} \mathbf{r}(s_t, a_t)$.

Multi-objective reinforcement learning (MORL) employs a scalarization function $u : \mathbb{R}^m \to \mathbb{R}$ to convert vector-valued returns into a scalar objective. The classical choice is linear scalarization, $u(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}$, which yields

$$J_{\text{lin}}(\pi; \mathbf{w}) = \mathbb{E}_{\tau\sim\pi}[\mathbf{w}^\top \mathbf{R}(\tau)], \tag{1}$$

where $\mathbf{w} \in \mathbb{R}^m$ specifies a linear trade-off between objectives. Since linear $u$ commutes with the expectation, the problem reduces to single-objective RL with scalar reward $r_{\mathbf{w}} = \mathbf{w}^\top \mathbf{r}$. To capture more general preference structures, nonlinear MORL allows $u$ to be nonlinear. In this setting, the expectation and scalarization no longer commute, giving rise to two canonical optimality criteria: the *scalarized expected return* (SER) and the *expected scalarized return* (ESR),

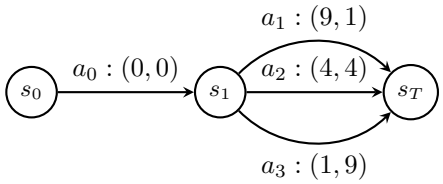
$$J_{\text{SER}}(\pi; u) = u\left(\mathbb{E}_{\tau\sim\pi}[\mathbf{R}(\tau)]\right), \qquad J_{\text{ESR}}(\pi; u) = \mathbb{E}_{\tau\sim\pi}[u(\mathbf{R}(\tau))]. \tag{2}$$

SER and ESR represent different decision-making contexts and assign different roles to the scalarization $u$. Under SER, $u_{\text{SER}}$ aggregates the expected return vector into a scalar objective and is typically concave, promoting balanced performance across objectives at the expectation level [11, 12]. Under ESR, $u_{\text{ESR}}$ encodes a trajectory-level preference and is not restricted to concave functions [2, 13]. This distinction leads to different optimal policy structures and optimization approaches, establishing the two central paradigms of nonlinear MORL.

### 2.2 Optimization challenges in nonlinear MORL

Figure 1 illustrates how optimal policies differ across MORL with linear, SER, and ESR objectives in a simple two-step MOMDP with $u_{\text{NSW}}(\mathbf{x}) = \sum_{i=1}^{m} \log x_i$ (Nash Social Welfare [14]). Nonlinear MORL produces behaviors absent in linear MORL: ESR selects an action no linear MORL policy chooses, and SER yields a stochastic optimal policy. Beyond these behavioral differences, each criterion poses a distinct optimization challenge.

**SER challenge: global-level optimization.** In standard RL, the contribution of each state-action pair can be evaluated locally through rewards or advantages, enabling Bellman-style dynamic programming. In SER, however, nonlinear scalarization is applied outside the expectation, coupling all state-action pairs globally: the marginal contribution of any action depends on the policy's overall expected return vector, not on local information alone. This precludes local improvement and necessitates optimization at the level of the policy-induced distribution. In Figure 1, action probabilities must be optimized jointly, and the SER-optimal policy is stochastic—strictly outperforming all




| Obj. | Optimal $\pi^*$ | $u_{\text{NSW}}(\mathbb{E}[\mathbf{R}])$ | $\mathbb{E}[u_{\text{NSW}}(\mathbf{R})]$ |
|---|---|---|---|
| SER | $0.5\,a_1 + 0.5\,a_3$ | $\mathbf{\log(25)}$ | $\log(9)$ |
| ESR | $a_2$ | $\log(16)$ | $\mathbf{\log(16)}$ |
| Linear | $a_1$ or $a_3$ | $\log(9)$ | $\log(9)$ |

Figure 1: Optimal policies of MORL with linear, SER, and ESR objectives in a two-step MOMDP with $u_{\text{NSW}}$. The optimal linear MORL policy selects the action corresponding to the higher-weighted objective. Full derivation in Appendix B.

deterministic policies, unlike in single-objective RL or linear MORL. Prior work [11] addresses this by reformulating the problem as convex optimization over occupancy measures, bypassing Bellman updates entirely.

**ESR challenge: non-Markovian policy.** ESR applies nonlinearity at the trajectory level, making the optimal action depend not only on the current state but also on the accumulated reward $\mathbf{R}_t^{\text{acc}} := \sum_{k=0}^{t-1} \mathbf{r}(s_k, a_k)$. For any $t \geq 1$, the ESR objective can be written as

$$J_{\text{ESR}}(\pi; u_{\text{ESR}}) = \mathbb{E}_\pi \left[ u_{\text{ESR}} \left( \mathbf{R}_t^{\text{acc}} + \sum_{k=t}^{H-1} \mathbf{r}(s_k, a_k) \right) \right].$$

Due to the nonlinearity, the future reward and $\mathbf{R}_t^{\text{acc}}$ cannot be optimized separately. The optimal action at state $s_t$ therefore depends on $\mathbf{R}_t^{\text{acc}}$: trajectories reaching the same state may require different actions depending on their reward history. ESR-optimal policies cannot, in general, be Markovian over the original state space [2, 13]. Replacing $\mathbf{r}(s_0, a_0)$ with $(2, 0)$ in Figure 1 shifts the ESR-optimal action at $s_1$ from $a_2$ to $a_3$, confirming that the optimal action depends on the reward history (derivation in Appendix B). Prior works on ESR optimization [15, 16, 17] address this by augmenting the state with $\mathbf{R}_t^{\text{acc}}$, restoring Markovianity in the augmented space.

## 3 AET: A Unified Framework for Nonlinear Scalarization in MORL

To address offline optimization of nonlinear trajectory-wise utilities and their nonlinear aggregation at the expectation level, we present AET, a unified framework that accommodates both levels of nonlinearity within a single formulation. AET decomposes nonlinear scalarization into three components: Aggregation (A), Expectation (E), and Transformation (T). The AET objective is

$$J_{\text{AET}}(\pi; F, G) = G(\mathbb{E}_{\tau\sim\pi}[F(\mathbf{R}(\tau))]), \tag{3}$$

where $F : \mathbb{R}^m \to \mathbb{R}^n$ is a trajectory-level transformation composed of $n$ trajectory-wise utilities $F(\mathbf{R}) = (f_1(\mathbf{R}), \ldots, f_n(\mathbf{R}))$, and $G : \mathbb{R}^n \to \mathbb{R}$ is an expectation-level aggregation. We assume each $f_i$ is smooth on its domain and $G$ is concave to ensure a well-behaved optimization landscape.

Each $f_i$ encodes a distinct preference over the return vector, mapping the $m$-dimensional return vector to $n$ utility dimensions whose expectations are then aggregated by $G$. For instance, $f_i(\mathbf{R}) = \log(R_i)$ captures diminishing marginal utility, amplifying gains in low-performing objectives; $f_i(\mathbf{R}) = \exp(R_i)$ captures increasing marginal utility, amplifying gains in already high-performing objectives; and non-convex Cobb-Douglas $f_i(R_1, R_2) = R_1^\rho / (R_2)^{1-\rho}$ captures the efficiency trade-off ratio between two objectives. Linear $G$ aggregates the expected utilities linearly, while concave $G$ amplifies gains in lower-performing expected utilities, encouraging balanced performance across all utility dimensions.

**Landscape of Nonlinear MORL.** The AET formulation organizes existing nonlinear MORL objectives by where nonlinearity is placed. Based on the nonlinearity of $F$ and $G$, we identify four regimes: (i) Linear MORL (linear $F$, linear $G$), recovered by $F(\mathbf{R}) = \mathbf{R}$ and $G(\mathbf{v}) = \mathbf{w}^\top \mathbf{v}$; (ii) SER-MORL (linear $F$, concave $G$), recovered by retaining the identity transformation but adopting a concave aggregation $G$; (iii) ESR-MORL (nonlinear $F$, linear $G$), recovered by setting $F = (f_1, \ldots, f_n)$ with $G(\mathbf{v}) = \mathbf{1}^\top \mathbf{v}$; and (iv) AET-MORL (nonlinear $F$, concave $G$), recovered

by aggregating $F = (f_1, \ldots, f_n)$ with concave $G$. We note that ESR-MORL can also be described as $F = f_1 + \cdots + f_n$ with $G(v) = v$, reducing to a single scalar utility; we retain the vector representation to clearly distinguish it from AET-MORL, where concave $G$ preserves the distinction among multiple utility dimensions.

This taxonomy makes explicit which regimes lack offline methods: while Linear MORL and SER-MORL are addressed by prior work [9, 10, 11], no existing offline approach handles **ESR-MORL** or **AET-MORL**. ESR-MORL has been studied in online and value-iteration settings [15, 16, 17], but (i) no offline method addresses the additional challenge of distribution shift, and (ii) current approaches cannot be extended to optimizing AET objectives with two nonlinearities, as they do not account for nonlinear aggregation $G$. We fill these gaps by addressing each source of nonlinearity in turn: Section 4 tackles nonlinear $F$, Section 5 tackles nonlinear $G$, and their combination yields AETDICE.

## 4 Augmented MOMDP for Nonlinear Transformation $F$

We present an augmented MOMDP framework for optimizing nonlinear transformation $F$. While we adopt the augmented state space from prior work [17] to restore Markovianity, we define a transformed reward that captures the incremental contribution of each action to $F$. This serves two purposes: it enables offline optimization of ESR-MORL by reducing it to standard single-objective RL in the augmented MOMDP, and it extends naturally to AET-MORL by reformulating it as SER-MORL in the augmented MOMDP.

### 4.1 Transformed Reward in the Augmented Space

We define the augmented finite-horizon MOMDP $\tilde{\mathcal{M}} = (\tilde{\mathcal{S}}, \mathcal{A}, \tilde{P}, H, \tilde{p}_0, \tilde{\mathbf{r}})$, where each augmented state is given by $\tilde{s}_t = (s_t, \mathbf{R}_t^{\text{acc}}, t)$. The initial augmented state is $\tilde{s}_0 = (s_0, \mathbf{0}, 0)$ with initial distribution $\tilde{p}_0(\tilde{s}_0) = p_0(s_0)$. The transition kernel $\tilde{P}$ is defined as

$$\tilde{P}(\tilde{s}_{t+1}|\tilde{s}_t, a_t) = P(s_{t+1}|s_t, a_t) \cdot \mathbb{I}[\mathbf{R}_{t+1}^{\text{acc}} = \mathbf{R}_t^{\text{acc}} + \mathbf{r}(s_t, a_t)], \tag{4}$$

where the timestep advances deterministically from $t$ to $t+1$. Since reward accumulation is absorbed into the transition dynamics and our objective focuses on the trajectory-wise utility $F(\mathbf{R}(\tau))$, we define a transformed reward to capture the incremental contribution of each action to $F$:

$$\tilde{\mathbf{r}}(\tilde{s}_t, a_t) = F\big(\mathbf{R}_t^{\text{acc}} + \mathbf{r}(s_t, a_t)\big) - F\big(\mathbf{R}_t^{\text{acc}}\big), \quad \text{for } t \geq 1 \tag{5}$$

with $\tilde{\mathbf{r}}(\tilde{s}_0, a_0) = F(\mathbf{r}(s_0, a_0))$. This induces an exact telescoping sum: $\sum_{t=0}^{H-1} \tilde{\mathbf{r}}(\tilde{s}_t, a_t) = F(\mathbf{R}(\tau))$, reducing the trajectory-level objective to a standard cumulative reward in the augmented MOMDP. When $F$ is linear, the transformed reward reduces to $\tilde{\mathbf{r}}(\tilde{s}_t, a_t) = F(\mathbf{r}(s_t, a_t))$, which no longer depends on $\mathbf{R}_t^{\text{acc}}$; the optimal policy is then independent of the accumulated reward, making the augmentation redundant.

Under this construction, a policy must be conditioned on the augmented state $\tilde{s}_t$ to account for the dependence on accumulated rewards. We denote such policies as $\tilde{\pi}(a|\tilde{s}_t)$, which are Markovian in $\tilde{\mathcal{M}}$ but non-Markovian in the original MOMDP $\mathcal{M}$. This reduces ESR-MORL to solving single-objective RL in $\tilde{\mathcal{M}}$, as we show below.

### 4.2 Offline ESR-MORL

As described in Section 3, when $G$ is linear, it commutes with the expectation, allowing us to collapse the AET objective into a single-utility case ($n = 1$) by defining $\tilde{F} = G \circ F \colon \mathbb{R}^m \to \mathbb{R}$. ESR-MORL then becomes standard single-objective RL in $\tilde{\mathcal{M}}$, as the trajectory-level nonlinearity is fully absorbed into the transformed reward. This enables existing offline RL algorithms, such as IQL [18], to be applied with only minor modifications to account for the augmented state $\tilde{s}_t$.

**Proposition 4.1** (Reduction to single-objective RL under linear $G$)**.** *Let $G$ be linear. Then for any policy $\tilde{\pi}$ in the augmented MOMDP $\tilde{\mathcal{M}}$,*

$$J_{\text{AET}}(\tilde{\pi}; F, G) = G(\mathbb{E}_{\tau\sim\tilde{\pi}}[F(\mathbf{R}(\tau))]) = \mathbb{E}_{\tau\sim\tilde{\pi}}\Big[\tilde{F}(\mathbf{R}(\tau))\Big] = \mathbb{E}_{\tau\sim\tilde{\pi}}\left[\sum_{t=0}^{H-1} \tilde{r}(\tilde{s}_t, a_t)\right], \tag{6}$$

*where $\tilde{r}$ is defined in* (5) *with $\tilde{F} = G \circ F$.*

**Comparison with existing ESR methods.** Prior ESR optimization methods [15, 16] modify the value-function update, rather than transforming the reward, to account for trajectory-wise nonlinearity, achieving effective online optimization. However, this approach does not align with two goals in our framework: offline optimization and AET optimization. Designing a value-function update that simultaneously handles both nonlinearity and distribution shift is nontrivial. Our transformed reward decouples these concerns: the nonlinearity of $F$ is absorbed into the reward, while the value-function update relies entirely on existing offline RL methods. Furthermore, value-function updates evaluate actions locally, which does not accommodate the global-level optimization required by concave $G$ in AET-MORL. This motivates the visitation-distribution-based formulation in the augmented MOMDP, which we develop in the next section.

In addition, prior work on ESR [17] conditions the policy on the accumulated reward vector $\mathbf{R}_t^{\text{acc}}$ and assigns the transformed return $F(\mathbf{R}(\tau))$ only as a terminal reward. In contrast, our formulation provides a denser reward signal $\tilde{\mathbf{r}}$ that captures the incremental contribution of each action to $F(\mathbf{R}(\tau))$, making it amenable to sample-based optimization.

## 5 Convex Optimization for Nonlinear Aggregation $G$

With nonlinear $F$ absorbed into the transformed reward, optimizing AET-MORL reduces to SER-MORL in $\tilde{\mathcal{M}}$: maximizing a concave $G$ over multiple expected utilities. To handle the global-level optimization required by concave $G$, we extend the Distribution Correction Estimation (DICE) framework [11, 19] to the augmented MOMDP $\tilde{\mathcal{M}}$, yielding AETDICE.

### 5.1 Visitation Distribution in the Augmented MOMDP

In the augmented MOMDP, the visitation distribution induced by a policy $\tilde{\pi}$ is defined as $d_t^{\tilde{\pi}}(\tilde{s}, a) := \Pr(\tilde{s}_t = \tilde{s}, a_t = a \mid \tilde{\pi})$, representing the probability that $\tilde{\pi}$ visits the augmented state-action pair $(\tilde{s}, a)$ at timestep $t$. While the timestep is already encoded within the augmented state $\tilde{s}$, we retain the index $t$ for notational clarity in the derivations below. Crucially, the expected utility vector becomes linear in the visitation distribution $d_t^{\tilde{\pi}}$:

$$\mathbb{E}_{\tau\sim\tilde{\pi}}[F(\mathbf{R}(\tau))] = \mathbb{E}_{\tau\sim\tilde{\pi}}\left[\sum_{t=0}^{H-1} \tilde{\mathbf{r}}(\tilde{s}_t, a_t)\right] = \sum_{t=0}^{H-1}\sum_{\tilde{s},a} d_t^{\tilde{\pi}}(\tilde{s}, a)\, \tilde{\mathbf{r}}(\tilde{s}, a). \tag{7}$$

Therefore, AET-MORL reduces to a convex optimization problem over visitation distributions in $\tilde{\mathcal{M}}$, subject to finite-horizon Bellman flow constraints that ensure the validity of $\mathbf{d} = \{d_t\}_{t=0}^{H-1}$:

$$\max_{\mathbf{d}\geq 0} \quad G\left(\sum_{t=0}^{H-1}\sum_{\tilde{s},a} d_t(\tilde{s}, a)\, \tilde{\mathbf{r}}(\tilde{s}, a)\right) \tag{8a}$$

$$\text{s.t.} \quad \sum_a d_0(\tilde{s}, a) = \tilde{p}_0(\tilde{s}), \quad \forall \tilde{s} \tag{8b}$$

$$\sum_{a'} d_{t+1}(\tilde{s}', a') = \sum_{\tilde{s},a} d_t(\tilde{s}, a)\, \tilde{P}(\tilde{s}' \mid \tilde{s}, a), \quad \forall \tilde{s}',\, t \in \{0, \ldots, H-2\}, \tag{8c}$$

where $\tilde{s}' = (s', \mathbf{R}_t^{\text{acc}} + \mathbf{r}(s, a), t+1)$ denotes the next augmented state. Since $F$ enters (8) only through $\tilde{\mathbf{r}}$, the objective is concave in $\mathbf{d}$ whenever $G$ is concave, regardless of whether $F$ is convex, concave, or neither. This formulation enables the global-level optimization required by SER-like objectives with nonlinear $G$, and the optimal policy inducing $\mathbf{d}^*$ can be recovered via $\tilde{\pi}_t^*(a|\tilde{s}) = d_t^*(\tilde{s}, a) / \sum_a d_t^*(\tilde{s}, a)$. However, solving this convex optimization directly requires access to the transition dynamics of $\tilde{\mathcal{M}}$. Moreover, since $\mathbf{R}_t^{\text{acc}}$ takes values in $\mathbb{R}^m$, the augmented state space $\tilde{\mathcal{S}}$ is in general continuous. We therefore extend the convex formulation to sample-based optimization from offline data, as developed next.

### 5.2 AETDICE: Regularized AET for sample-based optimization

To derive a sample-based objective from (8) in the offline setting, AETDICE extends the DICE framework to the augmented MOMDP with two key modifications: (i) $\phi$-divergence regularization

applied to the per-timestep data distribution $d_t^D$, correcting distribution shift at each step of the finite horizon, and (ii) a slack variable $\mathbf{k}$ that decouples the concave aggregation $G$ from the expectation over $\mathbf{d}$. The regularized primal problem is:

$$\max_{\mathbf{d}\in\mathcal{D},\mathbf{k}\in\mathbb{R}^n} G(\mathbf{k}) - \beta \sum_{t=0}^{H-1} D_\phi(d_t \| d_t^D), \quad \text{s.t.} \quad \sum_{t=0}^{H-1}\sum_{\tilde{s},a} d_t(\tilde{s},a)\tilde{\mathbf{r}}(\tilde{s},a) = \mathbf{k}, \tag{9}$$

where $\mathcal{D}$ denotes the set of non-negative visitation distributions satisfying constraints (8b)–(8c), $\beta > 0$ is a regularization coefficient, and $D_\phi(d_t \,\|\, d_t^D) := \sum_{\tilde{s},a} d_t^D(\tilde{s},a)\phi\left(d_t(\tilde{s},a)/d_t^D(\tilde{s},a)\right)$ is the $\phi$-divergence between $d_t$ and the data distribution $d_t^D$ at time step $t$. The full derivation, required assumptions, and complete algorithm are provided in Appendix C.

**Dual formulation.** We form the Lagrangian of (9) by introducing $\boldsymbol{\mu} \in \mathbb{R}^n$ for the slack constraint, and time-indexed multipliers $\nu_t(\tilde{s})$ for the Bellman flow constraints (8b)–(8c), replacing the time-independent multiplier used in standard DICE approaches [11, 19] to reflect the finite-horizon structure of the augmented MOMDP. The primal problem becomes $\max_{\mathbf{d}\geq 0,\mathbf{k}} \min_{\boldsymbol{\nu},\boldsymbol{\mu}} \mathcal{L}$, where:

$$\mathcal{L}(\mathbf{d},\mathbf{k},\boldsymbol{\nu},\boldsymbol{\mu}) := G(\mathbf{k}) - \boldsymbol{\mu}^\top\mathbf{k} + \sum_{\tilde{s}} \nu_0(\tilde{s})\tilde{p}_0(\tilde{s}) - \beta\sum_{t=0}^{H-1} D_\phi(d_t\|d_t^D) + \sum_{\tilde{s},a,t} d_t(\tilde{s},a)e_{\boldsymbol{\nu},\boldsymbol{\mu},t}(\tilde{s},a)$$

with $e_{\boldsymbol{\nu},\boldsymbol{\mu},t}(\tilde{s},a) = \boldsymbol{\mu}^\top\tilde{\mathbf{r}}(\tilde{s},a) + \mathbb{E}_{\tilde{s}'}[\nu_{t+1}(\tilde{s}')] - \nu_t(\tilde{s})$ and $\nu_H(\tilde{s}) = 0$ for consistency.

**Derivation of training objectives.** Under strong duality, we swap the optimization order to $\min_{\boldsymbol{\nu},\boldsymbol{\mu}} \max_{\mathbf{d}\geq 0,\mathbf{k}} \mathcal{L}$. The inner maximization over $\mathbf{d}$ yields a closed-form optimal visitation distribution ratio $w_t^*(\tilde{s},a) = d_t^*(\tilde{s},a)/d_t^D(\tilde{s},a) = \max\left(0, (\phi')^{-1}\left(e_{\boldsymbol{\nu},\boldsymbol{\mu},t}(\tilde{s},a)/\beta\right)\right)$. Substituting this closed form and using the Fenchel conjugates $\phi_+^*(y) := \sup_{x\geq 0}\{xy - \phi(x)\}$ and $G^*(\mathbf{y}) := \sup_{\mathbf{x}}\{\mathbf{x}^\top\mathbf{y} + G(\mathbf{x})\}$ eliminates the inner maximization analytically, leaving an unconstrained minimization over $(\boldsymbol{\nu},\boldsymbol{\mu})$ that depends only on expectations over $\tilde{p}_0$ and the offline data $d_t^D$:

$$\min_{\boldsymbol{\nu},\boldsymbol{\mu}} \mathcal{L}_{\mathrm{AET}}(\boldsymbol{\nu},\boldsymbol{\mu}) := \mathbb{E}_{\tilde{p}_0}[\nu_0(\tilde{s})] + \sum_{t=0}^{H-1} \mathbb{E}_{d_t^D}\left[\beta\phi_+^*\left(e_{\boldsymbol{\nu},\boldsymbol{\mu},t}(\tilde{s},a)/\beta\right)\right] + G^*(-\boldsymbol{\mu}). \tag{10}$$

After minimizing $\mathcal{L}_{\mathrm{AET}}$ to obtain $\boldsymbol{\nu}^*$ and $\boldsymbol{\mu}^*$, the policy that induces the optimal visitation distribution can be extracted using the optimal visitation ratio $w_t^*$ via weighted behavior cloning following [19]:

$$\max_{\tilde{\pi}} \sum_{t=0}^{H-1} \mathbb{E}_{(\tilde{s},a)\sim d_t^D}\left[w_t^*(\tilde{s},a)\log\tilde{\pi}(a|\tilde{s})\right].$$

**Comparison with prior DICE methods.** AETDICE differs from the DICE framework [11, 19] in three key aspects: (i) prior DICE approaches impose Bellman flow constraints in a discounted infinite-horizon MDP, whereas AETDICE imposes finite-horizon Bellman flow constraints over the augmented state space (8b)–(8c); (ii) the Lagrangian multipliers and policy are conditioned on the augmented state $\tilde{s}$, and offline data $d_t^D$ must be preprocessed to include the accumulated reward vector $\mathbf{R}_t^{\mathrm{acc}}$; and (iii) AETDICE extends FairDICE, which handles only SER objectives, to support a broader class of nonlinear objectives including ESR and AET, though it operates in a higher-dimensional state space. This requires the offline data to cover the accumulated reward in addition to the original state-action pairs. Since the $\phi$-divergence regularization keeps each $d_t$ within the support of $d_t^D$, limited coverage restricts the attainable objective rather than inducing extrapolation to unseen $\mathbf{R}_t^{\mathrm{acc}}$; Appendix D analyzes this requirement and its dependence on the number of objectives.

## 6 Experiments

We conduct two sets of experiments to demonstrate that AETDICE enables offline optimization of AET objectives with nonlinear $F$ and $G$. First, we focus on offline Fair MORL with concave $F$ and $G$, optimizing and comparing different fairness objectives within a single framework. Second, we go beyond concavity to optimize convex and non-convex trajectory-wise utilities, with linear or concave aggregation, on complex continuous-control domains.

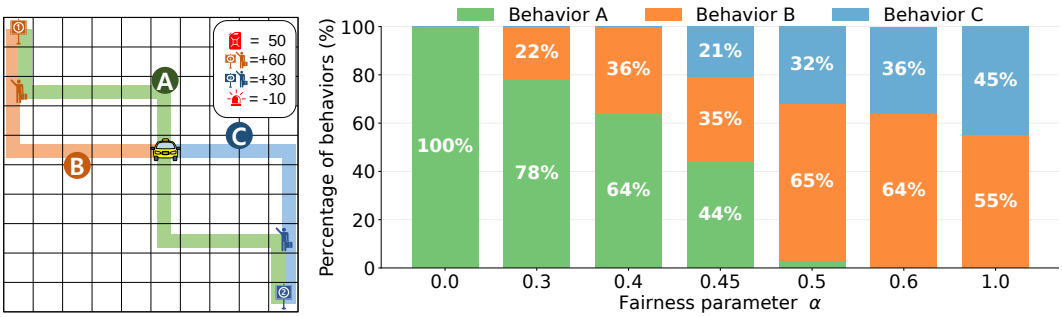



Figure 2: **(a)** Fair-Taxi: a taxi agent serving two passenger groups. Each objective corresponds to serving a passenger group. **(b)** Distribution of behavior patterns across varying nonlinear MORL objectives. Full details in Appendix F.

**Algorithms.** We evaluate offline methods corresponding to each MORL regime identified in Section 3: (1) **OptiDICE** [19], denoted OptiDICE-$\mathbf{w}$ for preference weight $\mathbf{w}$: a DICE-based offline RL method with linear $F$ and linear $G$; (2) **FairDICE** [11]: a DICE-based offline MORL method with linear $F$ and concave $G$; (3) **ESR-IQL** (ours): IQL applied to the augmented MOMDP (Section 4), for nonlinear $F$ and linear $G$; and (4) **AETDICE** (ours), denoted AET-X when trained for objective X: a DICE-based method on the augmented MOMDP for any $F$ and concave $G$.

### 6.1 Offline Fair MORL: ESR, SER, and BSR

Fair MORL [11, 20] applies concave scalarization to promote equitable performance across objectives, where improvements to lower-performing objectives yield larger gains in overall utility. As discussed in Section 3, the AET framework subsumes existing Fair MORL criteria as special cases. Using the $\alpha$-fairness utility [21], we construct **Balanced Scalarized Return (BSR)**, a novel objective family parameterized by $\alpha$ that interpolates between trajectory-level and expectation-level fairness:

$$J_{\text{BSR}}(\pi;\alpha) = G_\alpha(\mathbb{E}_{\tau\sim\pi}[F_\alpha(\mathbf{R}(\tau))])\,, \qquad F_\alpha(\mathbf{x}) = u_{1-\alpha}(\mathbf{x}), \quad G_\alpha(\mathbf{x}) = \sum_{i=1}^{n} u_\alpha(x_i), \tag{11}$$

where $u_\alpha(x) = \frac{x^{1-\alpha}-1}{1-\alpha}$ for $\alpha \neq 1$ and $u_1(x) = \log x$. By varying $\alpha \in [0, 1]$, BSR controls the relative nonlinearity applied at the trajectory level ($F_\alpha$) and the aggregation level ($G_\alpha$), recovering ESR at $\alpha = 0$ and SER at $\alpha = 1$. We analyze ESR-, BSR-, and SER-optimal policies obtained by AETDICE in the Fair-Taxi environment (Figure 2(a)), where a taxi serves two passenger groups and each objective rewards deliveries to one group. Details and full results, including MO-PointMaze-3obj, are presented in Appendices F and G.

**Behavioral differences.** Figure 2(a) illustrates three episode-level behavior patterns of optimal nonlinear MORL policies in Fair-Taxi: (A) serving both groups, (B) serving only group 1, and (C) serving only group 2. Figure 2(b) shows their distribution over 100 episodes. The ESR-optimal policy consistently produces pattern (A), balancing return across objectives within each episode. The SER-optimal policy yields a stochastic mixture of (B) and (C), where each trajectory specializes in a single objective and balance is achieved only in expectation. The BSR-optimal policy lies between these extremes: as $\alpha$ increases, specialized trajectories (B) and (C) gradually replace the balanced pattern (A).

**Non-Markovian policy.** The non-Markovian structure formalized in Section 4 appears in the ESR- and BSR-optimal policies. Since $F_\alpha$ is concave for $\alpha < 1$, the transformed reward $\tilde{\mathbf{r}}$ of an objective decreases as its accumulated reward grows, steering the policy toward the under-served objective; the optimal action at the same state therefore varies with the reward history. The SER-optimal policy is unaffected, as its transformed reward does not depend on $\mathbf{R}_t^{\text{acc}}$. Figure 3 shows a state at $t = 25$ where group 1 has already been served three times. The ESR-optimal policy switches to group 2, the SER-optimal policy continues serving group 1, and the BSR-optimal policy ($\alpha = 0.45$) chooses stochastically between the two.

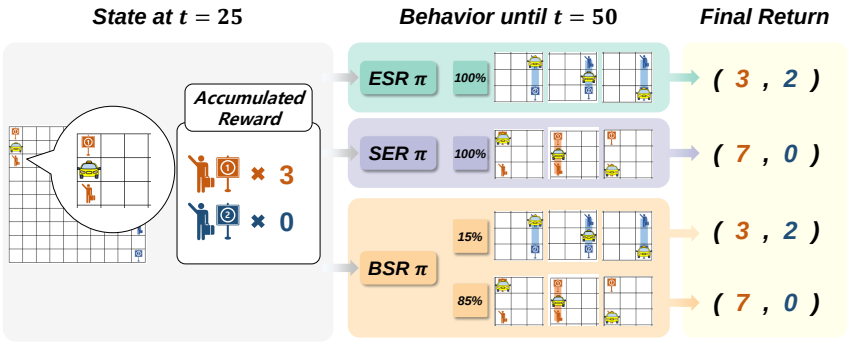



Figure 3: Actions of the ESR-, SER-, and BSR-optimal policies at the same state in the Fair-Taxi environment, after group 1 has been served three times. The ESR- and BSR-optimal policies depend on the accumulated reward; the SER-optimal policy does not.

**Quantitative evaluation.** Figure 4 shows the per-trajectory return of each objective in MO-PointMaze-3obj, a continuous navigation task in which each of three objectives rewards proximity to a distinct goal region. The SER-optimal policy attains higher expected returns with a large spread across trajectories, while the ESR-optimal policy attains lower expected returns with a much smaller spread. The BSR-optimal policy lies between the two. Prior work has noted that SER suits repeated policy execution and ESR suits single-shot deployment [4]; BSR interpolates between these two regimes through $\alpha$, retaining part of the expected return of SER while limiting the spread across trajectories.

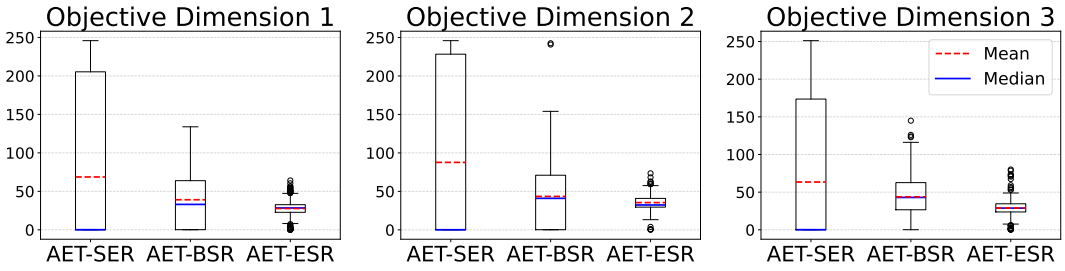
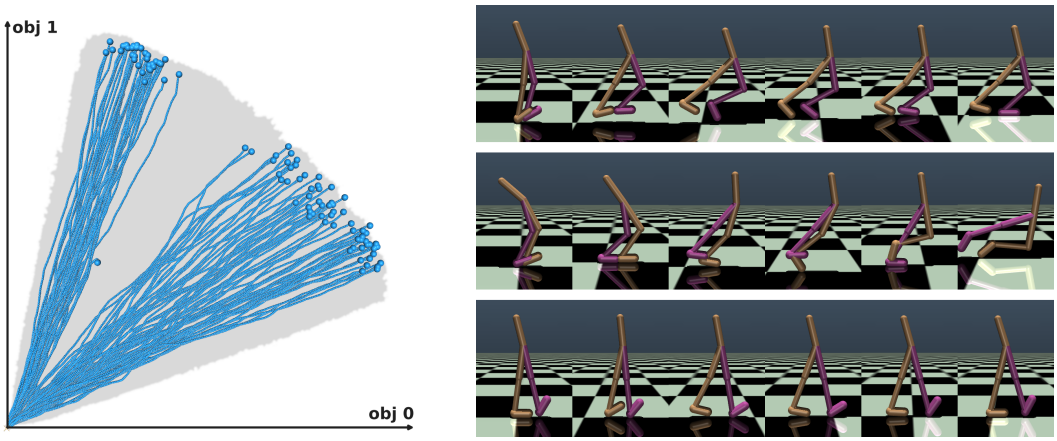



Figure 4: Per-trajectory returns of each objective in the MO-PointMaze-3obj environment under the SER-, BSR-, and ESR-optimal policies. Boxes show the interquartile range, with the mean (dashed) and median (solid).

### 6.2 Beyond Concavity: General AET Objectives

Beyond Fair MORL, the AET framework accommodates diverse trajectory-level utilities and their aggregation by varying $F$ and $G$. We demonstrate this on the D4MORL benchmark [9] of continuous-control tasks with four configurations: (i) convex $F(\mathbf{x}) = \exp(\mathbf{x}/50)$ with linear $G$, which prioritizes high-performing objectives (MO-Ant); (ii) the same convex $F$ with concave $G(\mathbf{x}) = \sum_i \log(x_i)$, which combines trajectory-level specialization with expectation-level balance (MO-Ant); (iii) the non-convex Cobb-Douglas utility $F(\mathbf{R}) = R_{\mathrm{v}}^{\rho}/R_{\mathrm{c}}^{1-\rho}$ of the velocity return $R_{\mathrm{v}}$ and the control-cost return $R_{\mathrm{c}}$, with linear $G$, which measures efficiency within each trajectory (MO-Walker2d and MO-HalfCheetah); and (iv) the Cobb-Douglas utility together with a safety utility that penalizes falls, each transformed by $\log$ within each trajectory and aggregated by linear $G$ (MO-Walker2d). Since $G$ is linear or concave in all four configurations, each remains a convex program over $\mathbf{d}$ (Section 5), regardless of the shape of $F$.

**Convex utility $F$.** We apply convex $F$ with increasing marginal utility to MO-Ant, where the two objectives correspond to movement in orthogonal directions. Convex $F$ amplifies gains from concentrating on a single objective, so under linear $G$ with equal weights (i), a policy that moves in one direction scores far higher than one that moves diagonally: in Table 1, policies that specialize in one objective, such as OptiDICE-$[0.0, 1.0]$ and ESR-IQL, score far higher under (i) than OptiDICE-$[0.5, 0.5]$. Pairing the same convex $F$ with concave $G$ (ii) restores balance at the expectation level:




(a) MO-Ant (b) MO-Walker2d ($\rho = 0.3$)

Figure 5: **(a)** MO-Ant: 100 rollout trajectories of AETDICE with convex $F$ and concave $G$ (top-down view; objectives 1 and 2 increase along the $+x$ and $+y$ directions, and the gray region is covered by the dataset). Each trajectory moves in one direction, and the policy mixes the two directions across episodes. **(b)** MO-Walker2d at $\rho = 0.3$, one episode per row. Top and middle: two episodes of the policy optimizing only the Cobb-Douglas utility, which adopts a crouched, energy-saving gait and falls in the middle episode. Bottom: the policy with the safety utility stays upright.

the concave aggregation amplifies under-served utility dimensions, producing a stochastic mixture of specialized trajectories (Figure 5(a)), where each trajectory moves in one direction while both directions are served across episodes. Achieving this with linear $G$ would require tuning the weights to offset the amplification of convex $F$. AETDICE attains the highest value under (ii) in Table 1.

**Cobb-Douglas utility.** We focus on MO-Walker2d and MO-HalfCheetah, where forward locomotion and energy consumption present a natural efficiency trade-off. We convert the energy-saving reward of D4MORL into a control cost, so that the per-step rewards are the velocity $r_{\mathrm{v}}$ and the control cost $r_{\mathrm{c}}$, with returns $R_{\mathrm{v}}$ and $R_{\mathrm{c}}$ (Appendix F.3). For $\rho \in (0, 1)$, the Cobb-Douglas utility $F(\mathbf{R}) = R_{\mathrm{v}}^{\rho} / R_{\mathrm{c}}^{1-\rho}$ is concave in $R_{\mathrm{v}}$ and convex in $R_{\mathrm{c}}$, and hence non-convex. It measures forward progress per unit of control effort within each trajectory, and $\rho$ sets the relative emphasis on the two. At low $\rho$ (e.g., $\rho = 0.3$), however, the MO-Walker2d agent falls in 21.3% of episodes, as the utility favors energy saving over stable locomotion; Figure 5(b, middle) shows such a fall.

**Cobb-Douglas utility with safety utility.** To prevent these falls, we add a third reward $r_{\mathrm{s}}$, the margin of the torso height to its termination bound, since falls in MO-Walker2d almost always cross the height bound (Appendix G.3). Its return $R_{\mathrm{s}}$, which we call the safety return, is larger for trajectories that stay upright for longer. We define the safety utility $u_{\mathrm{s}}(\mathbf{R}) = R_{\mathrm{s}} - R_{\mathrm{th}}$ with a threshold $R_{\mathrm{th}}$. Applying $\log$ to the Cobb-Douglas and safety utilities within each trajectory makes the safety term a concave penalty that grows as $R_{\mathrm{s}}$ falls short of $R_{\mathrm{th}}$, and aggregating the two with linear $G$ gives the AET objective

$$J_{\mathrm{AET}}(\tilde{\pi}) = \mathbb{E}_{\tau \sim \tilde{\pi}} \left[ \log \frac{R_{\mathrm{v}}(\tau)^{\rho}}{R_{\mathrm{c}}(\tau)^{1-\rho}} \right] + \mathbb{E}_{\tau \sim \tilde{\pi}} \left[ \log \left( R_{\mathrm{s}}(\tau) - R_{\mathrm{th}} \right) \right], \tag{12}$$

where $\log$ denotes the piecewise extension in Appendix C.5, which is defined for non-positive arguments.

Figure 5(b) compares AETDICE policies that optimize the Cobb-Douglas utility alone and with the safety utility: the safety utility keeps the agent upright and reduces the fall rate at $\rho = 0.3$ from 21.3% to 1.7% (Tables 12 and 13). These results show that the AET framework extends beyond concave transformations to practical objectives such as efficiency under a safety requirement. Full results are presented in Appendix G.

Table 1: AET objective values on MO-Ant with convex $F$, evaluated under configuration (i), linear $G$ with equal weights (left), and configuration (ii), concave $G$ (right). Each row is a policy optimized for a different objective: OptiDICE-$\mathbf{w}$ for linear scalarization with weight $\mathbf{w}$, FairDICE for SER, ESR-IQL and AET-ESR for (i), and AETDICE for (ii). Mean $\pm$ standard deviation over 5 random seeds; the best value in each column is in bold.

| METHOD | CONVEX $F$ + LINEAR $G$ | CONVEX $F$ + CONCAVE $G$ |
|---|---|---|
| OPTIDICE-$[1.0, 0.0]$ | $729.61 \pm 10.94$ | $12.91 \pm 0.03$ |
| OPTIDICE-$[0.5, 0.5]$ | $689.23 \pm 12.74$ | $13.05 \pm 0.04$ |
| OPTIDICE-$[0.0, 1.0]$ | $1221.97 \pm 26.71$ | $12.25 \pm 0.03$ |
| FAIRDICE | $682.65 \pm 20.06$ | $13.04 \pm 0.07$ |
| ESR-IQL | $\mathbf{1243.33 \pm 17.07}$ | $12.23 \pm 0.02$ |
| AET-ESR | $1067.22 \pm 6.81$ | $13.04 \pm 0.12$ |
| AETDICE | $949.73 \pm 22.02$ | $\mathbf{13.69 \pm 0.05}$ |

## 7 Conclusion

We presented AET, a unified framework that organizes nonlinear MORL objectives by the position of nonlinearity—trajectory-level transformation $F$ and expectation-level aggregation $G$—and identifies two previously unaddressed regimes: offline ESR-MORL and AET-MORL. To fill these gaps, we proposed AETDICE, which introduces a transformed reward that absorbs nonlinear $F$ into per-step rewards in an augmented MOMDP, and a finite-horizon DICE formulation that handles the global-level optimization required by concave $G$. Empirically, AETDICE optimizes ESR, SER, and novel AET objectives including BSR within a single framework, revealing distinct policy behaviors—including non-Markovian strategies and stochastic mixtures—that arise from each level of nonlinearity and their combination. Limitations and broader impact are discussed in Appendix K.

## Acknowledgments and Disclosure of Funding

This work was partly supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2022-II220311, Development of Goal-Oriented Reinforcement Learning Techniques for Contact-Rich Robotic Manipulation of Everyday Objects, No. RS-2024-00457882, AI Research Hub Project, No. RS-2019-II190079, Artificial Intelligence Graduate School Program (Korea University)), the IITP (Institute of Information & Communications Technology Planning & Evaluation)-ITRC (Information Technology Research Center) grant funded by the Korea government (Ministry of Science and ICT) (IITP-2026-RS-2024-00436857), BK21 Four project of the National Research Foundation of Korea, the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2025-00560367), the IITP under the Artificial Intelligence Star Fellowship support program to nurture the best talents (IITP-2026-RS-2025-02304828) grant funded by the Korea government (MSIT). This work was also supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant (RS-2020-II201361, Artificial Intelligence Graduate School Program (Yonsei University)), the AI Computing Infrastructure Enhancement (GPU Rental Support) User Support Program funded by the Ministry of Science and ICT (MSIT) (No. RQT-25-090109), Republic of Korea, and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2026-25505230, RS-2026-25619452). This work was also supported by the Technology Innovation Program (or Industrial Strategic Technology Development Program-Robot Industry Core Technology Development Program) (RS-2024-00507091, Development of autonomous process operation technology based on humanoid robots for autonomous manufacturing of major manufacturing industries) funded by the Ministry of Trade, Industry and Resources (MOTIR, Korea).

## References

[1] Kristof Van Moffaert and Ann Nowé. Multi-objective reinforcement learning using sets of pareto dominating policies. *The Journal of Machine Learning Research*, 15(1):3483–3512, 2014.

[2] Wang Chi Cheung. Regret minimization for reinforcement learning with vectorial feedback and complex objectives. *Advances in Neural Information Processing Systems*, 32, 2019.

[3] Runzhe Yang, Xingyuan Sun, and Karthik Narasimhan. A generalized algorithm for multi-objective reinforcement learning and policy adaptation. *Advances in neural information processing systems*, 32, 2019.

[4] Diederik M Roijers, Peter Vamplew, Shimon Whiteson, and Richard Dazeley. A survey of multi-objective sequential decision-making. *Journal of Artificial Intelligence Research*, 48:67–113, 2013.

[5] Mridul Agarwal, Vaneet Aggarwal, and Tian Lan. Multi-objective reinforcement learning with non-linear scalarization. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, pages 9–17, 2022.

[6] Diederik Roijers, Denis Steckelmacher, and Ann Nowé. Multi-objective reinforcement learning for the expected utility of the return. In *Adaptive Learning Agents Workshop 2018*, 2018.

[7] Conor F Hayes, Roxana Rădulescu, Eugenio Bargiacchi, Johan Källström, Matthew Macfarlane, Mathieu Reymond, Timothy Verstraeten, Luisa M Zintgraf, Richard Dazeley, Fredrik Heintz, et al. A practical guide to multi-objective reinforcement learning and planning. *arXiv preprint arXiv:2103.09568*, 2021.

[8] Conor F Hayes, Timothy Verstraeten, Diederik M Roijers, Enda Howley, and Patrick Mannion. Expected scalarised returns dominance: a new solution concept for multi-objective decision making. *Neural Computing and Applications*, 37(19):13079–13099, 2025.

[9] Baiting Zhu, Meihua Dang, and Aditya Grover. Scaling pareto-efficient decision making via offline multi-objective rl. *arXiv preprint arXiv:2305.00567*, 2023.

[10] 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*, pages 1201–1209, 2024.

[11] Woosung Kim, Jinho Lee, Jongmin Lee, and Byung-Jun Lee. Fairdice: Fairness-driven offline multi-objective reinforcement learning. *arXiv preprint arXiv:2506.08062*, 2025.

[12] Roxana Rădulescu, Patrick Mannion, Yijie Zhang, Diederik M Roijers, and Ann Nowé. A utility-based analysis of equilibria in multi-objective normal-form games. *The Knowledge Engineering Review*, 35:e32, 2020.

[13] Ziming Fan, Nianli Peng, Muhang Tian, and Brandon Fain. Welfare and fairness in multi-objective reinforcement learning. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, pages 1991–1999, 2023.

[14] Mamoru Kaneko and Kenjiro Nakamura. The nash social welfare function. *Econometrica: Journal of the Econometric Society*, pages 423–435, 1979.

[15] Guanbao Yu, Umer Siddique, and Paul Weng. Fair deep reinforcement learning with preferential treatment. In *ECAI*, pages 2922–2929, 2023.

[16] Umer Siddique, Peilang Li, and Yongcan Cao. Learning fair pareto-optimal policies in multi-objective reinforcement learning. In *The Seventeenth Workshop on Adaptive and Learning Agents*.

[17] Nianli Peng, Muhang Tian, and Brandon Fain. Multi-objective reinforcement learning with nonlinear preferences: Provable approximation for maximizing expected scalarized return. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems*, pages 1632–1640, 2025.

[18] Ilya Kostrikov, Ashvin Nair, and Sergey Levine. Offline reinforcement learning with implicit q-learning. *arXiv preprint arXiv:2110.06169*, 2021.

[19] Jongmin Lee, Wonseok Jeon, Byungjun Lee, Joelle Pineau, and Kee-Eung Kim. Optidice: Offline policy optimization via stationary distribution correction estimation. In *International Conference on Machine Learning*, pages 6120–6130. PMLR, 2021.

[20] Giseung Park, Woohyeon Byeon, Seongmin Kim, Elad Havakuk, Amir Leshem, and Youngchul Sung. The max-min formulation of multi-objective reinforcement learning: From theory to a model-free algorithm. In *International Conference on Machine Learning*, pages 39616–39642. PMLR, 2024.

[21] Jeonghoon Mo and Jean Walrand. Fair end-to-end window-based congestion control. *IEEE/ACM Transactions on networking*, 8(5):556–567, 2002.

[22] Kristof Van Moffaert, Madalina M Drugan, and Ann Nowé. Scalarized multi-objective reinforcement learning: Novel design techniques. In *2013 IEEE symposium on adaptive dynamic programming and reinforcement learning (ADPRL)*, pages 191–199. IEEE, 2013.

[23] Daniel J Lizotte, Michael H Bowling, and Susan A Murphy. Efficient reinforcement learning with multiple reward functions for randomized controlled trial analysis. In *ICML*, volume 10, pages 695–702, 2010.

[24] Kristof Van Moffaert, Madalina M Drugan, and Ann Nowé. Hypervolume-based multi-objective reinforcement learning. In *International Conference on Evolutionary Multi-Criterion Optimization*, pages 352–366. Springer, 2013.

[25] Joar Skalse and Alessandro Abate. On the limitations of markovian rewards to express multi-objective, risk-sensitive, and modal tasks. In *Uncertainty in Artificial Intelligence*, pages 1974–1984. PMLR, 2023.

[26] Umer Siddique, Abhinav Sinha, and Yongcan Cao. Fairness in preference-based reinforcement learning. *arXiv preprint arXiv:2306.09995*, 2023.

[27] Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. *Advances in neural information processing systems*, 33:1179–1191, 2020.

[28] 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.

[29] Runzhe Wu, Yufeng Zhang, Zhuoran Yang, and Zhaoran Wang. Offline constrained multi-objective reinforcement learning via pessimistic dual value iteration. *Advances in Neural Information Processing Systems*, 34:25439–25451, 2021.

[30] Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. D4rl: Datasets for deep data-driven reinforcement learning, 2020.

## NeurIPS Paper Checklist

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 abstract and introduction state three contributions: (i) the AET framework unifying SER and ESR (Section 3), (ii) AETDICE as a tractable offline algorithm for AET objectives (Sections 4–5), and (iii) empirical analysis of resulting policy trade-offs (Section 6). Each is supported by the corresponding section.

   Guidelines:

   - The answer [N/A] 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 [N/A] 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: Limitations are discussed in Appendix K: the larger distribution shift in offline MORL, which requires small regularization, and the coverage requirement over the accumulated reward $\mathbf{R}_t^{\text{acc}}$, analyzed in Appendix D.

   Guidelines:

   - The answer [N/A] 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: Proposition 4.1 follows directly from the linearity of $G$ and the telescoping identity below (5). The assumptions on $F$ and $G$ are stated in Section 3, and the derivation of AETDICE with its required assumptions is given in Appendix C.

   Guidelines:

   - The answer [N/A] 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: Section 5 fully describes the AETDICE algorithm including the regularized primal, dual formulation, and policy extraction. The full algorithm is given in Appendix C, hyperparameters and the ESR-IQL implementation in Appendix I, and environment and dataset details for Fair-Taxi, MO-PointMaze, and D4MORL in Appendix F.

   Guidelines:

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

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

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

   Answer: [Yes]

   Justification: The code, with instructions to reproduce the experiments, is available at `https://github.com/githubProjectAccount1/AETDICE-morl`. The D4MORL datasets are publicly available [9].

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.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://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: Network architectures, optimizer choice, regularization parameter $\beta$, $\alpha$ values for each objective, and dataset construction are reported in Appendices F and I alongside the experimental setup in Section 6.

   Guidelines:

   - The answer [N/A] 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: All quantitative results report mean and standard deviation over three or five random seeds, as stated in each table caption (standard error in Appendix D). Behavioral distribution results aggregate 100 evaluation episodes per configuration.

   Guidelines:

   - The answer [N/A] 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: Compute resources are reported in Appendix J: a single NVIDIA RTX 3090 GPU, with each AETDICE or IQL training run taking approximately 15–20 minutes.

   Guidelines:

   - The answer [N/A] 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 conforms with the NeurIPS Code of Ethics. The work is methodological, uses synthetic environments (Fair-Taxi, MO-PointMaze), and involves no human subjects, sensitive data, or high-risk dual-use elements.

   Guidelines:

   - The answer [N/A] 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: [No]

    Justification: AETDICE advances offline MORL by enabling optimization of nonlinear objectives that capture fairness and balanced performance across multiple criteria. As a methodological contribution, we do not anticipate direct negative societal impacts; further discussion is provided in Appendix K.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] 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., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release pre-trained models, generative systems, or scraped datasets that pose misuse risks.

    Guidelines:

    - The answer [N/A] 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 cite all prior work whose methods or formulations we build on, including FairDICE [11], OptiDICE [19], and IQL [18]. The Fair-Taxi environment is adapted from prior work [13]; we describe our modifications in Appendix F.

    Guidelines:

    - The answer [N/A] 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: [Yes]

    Justification: The paper introduces the AET framework, AETDICE algorithm, and BSR objective family, all of which are formally defined with full derivations in the main text and appendix. Environment details, algorithm pseudocode, and implementation specifics are provided in the appendix. The released code (`https://github.com/githubProjectAccount1/AETDICE-morl`) is documented with a README containing setup and run instructions and is released under the MIT license.

    Guidelines:

    - The answer [N/A] 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: [N/A]

    Justification: The paper does not involve crowdsourcing or human subjects research.

    Guidelines:

    - The answer [N/A] 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: [N/A]

    Justification: The paper does not involve human subjects research.

    Guidelines:

- The answer [N/A] 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.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs were not used as a component of the core method. Any use of LLMs was limited to writing assistance and does not affect methodology, scientific rigor, or originality.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.

# A Related Works

**Multi-Objective Reinforcement Learning** Multi-objective reinforcement learning (MORL) studies sequential decision-making problems with vector-valued rewards, where no single policy is universally optimal under the Pareto dominance relation [4]. As a result, MORL methods often aim to compute a set of trade-off solutions (e.g., Pareto-optimal or coverage sets) so that a decision maker can later select a policy once preferences are specified [22, 1]. A widely used instantiation is linear scalarization, which reduces the problem to a standard single-objective MDP by optimizing a weighted sum of objectives and thereby enables the use of conventional RL algorithms [23, 24]. This perspective has led to two main solution paradigms: set-based approaches, which learn a collection of Pareto-optimal policies, and preference-conditioned methods, which adapt a single policy to different trade-offs [4, 22, 1].

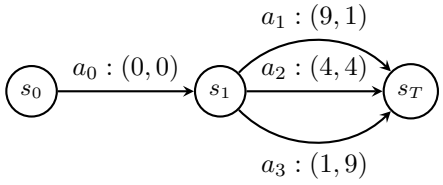
**Nonlinear Preference in MORL** However, linear scalarization cannot capture richer preference structures such as fairness, balance, or risk sensitivity [25]. To address this limitation, recent work has explored nonlinear scalarization based on concave utility functions applied to return vectors. Under appropriate smoothness and concavity conditions, such nonlinear objectives remain tractable to optimize [5, 17]. This line of work includes welfare-oriented criteria such as Nash social welfare and Gini-based measures, which encourage equitable performance across objectives [13, 26]. Related approaches also study max–min formulations that explicitly promote balanced outcomes, often combined with entropy regularization to stabilize learning in model-free online settings [20]. Together, these works reflect a broader shift from linear trade-offs toward more expressive preference modeling in MORL.

**Offline RL and Offline MORL** Offline reinforcement learning (RL) focuses on learning policies from fixed datasets without additional environment interaction, where distributional shift is a central challenge. Existing approaches address this issue through conservative value estimation [27], implicit value regularization [18], divergence-regularized optimization [19], and return-conditioned sequence modeling [28].

In the multi-objective setting, most offline approaches rely on linear scalarization and require explicit preference conditioning during training or evaluation [29, 10]. FairDICE [11] adapts stationary distribution correction to a nonlinear welfare objective, showing that certain nonlinear criteria can be handled in offline MORL. However, it is tailored to a specific welfare formulation and does not directly generalize to broader classes of nonlinear objectives with different placements of nonlinearity in the return structure.

# B Optimal Policies in the Two-Step MOMDP Example

In this section, we derive the optimal policies for linear, SER, and ESR objectives in the two-step MOMDP shown in Figure 6 (reproduced from Figure 1 for convenience). Through these derivations, we illustrate the distinct optimization challenges that arise from nonlinear MORL objectives.




| Obj. | Optimal $\pi^*$ | $u_{\text{NSW}}(\mathbb{E}[\mathbf{R}])$ | $\mathbb{E}[u_{\text{NSW}}(\mathbf{R})]$ |
|---|---|---|---|
| SER | $0.5\,a_1 + 0.5\,a_3$ | $\mathbf{\log(25)}$ | $\log(9)$ |
| ESR | $a_2$ | $\log(16)$ | $\mathbf{\log(16)}$ |
| Linear | $a_1$ or $a_3$ | $\log(9)$ | $\log(9)$ |

Figure 6: Optimal policies of MORL with linear, SER, and ESR objectives in a two-step MOMDP with $u_{\text{NSW}}(\mathbf{x}) = \sum_i \log(x_i)$.

**Linear scalarization.** For linear preference weights $[w_1, w_2] \in \mathbb{R}^2_+$, the linear MORL objective is defined as $J_{\text{lin}}(\pi; \mathbf{w}) = \mathbb{E}_\pi[w_1 R_1 + w_2 R_2]$. The optimal policy selects $a_1$ when $w_1 > w_2$ and $a_3$ when $w_1 < w_2$, with both SER $u_{\text{NSW}}(\mathbb{E}[\mathbf{R}])$ and ESR $\mathbb{E}[u_{\text{NSW}}(\mathbf{R})]$ remaining constant at $\log 9$ in either case. The exception is $w_1 = w_2$, where $a_1$ and $a_3$ are equally optimal, making any stochastic

mixture also optimal. However, Scalarized Expected Return (SER) is no longer constant in this case, as it varies with the mixing probability.

**Scalarized Expected Return (SER).** SER maximizes $J_{\text{SER}}(\pi) = u_{\text{NSW}}(\mathbb{E}_\pi[\mathbf{R}])$. Unlike the linear case, optimizing SER is not straightforward: letting $p_1$, $p_2$, $p_3$ denote the probabilities of selecting $a_1$, $a_2$, $a_3$, one must solve a convex optimization over the full policy distribution. This reflects a key challenge of SER optimization—the nonlinearity applies to the expected return, requiring global optimization over the entire policy rather than local, per-action evaluation. The convex optimization can be written as:

$$\begin{aligned} \max_{p_1,p_2,p_3} \quad & \log(9p_1 + 4p_2 + p_3) + \log(p_1 + 4p_2 + 9p_3) \\ \text{s.t.} \quad & p_1 + p_2 + p_3 = 1, \quad p_1, p_2, p_3 \geq 0. \end{aligned} \tag{13}$$

where the optimal solution is $p_1^* = 0.5$, $p_2^* = 0$, $p_3^* = 0.5$, yielding a stochastic policy that mixes $a_1$ and $a_3$ equally.

**Expected Scalarized Return (ESR).** ESR maximizes $J_{\text{ESR}}(\pi) = \mathbb{E}_\pi[u_{\text{NSW}}(\mathbf{R})]$. Since the nonlinearity is applied before the expectation, each action can be evaluated independently: $u_{\text{NSW}}(\mathbf{R}(a_1)) = \log 9$, $u_{\text{NSW}}(\mathbf{R}(a_2)) = \log 16$, $u_{\text{NSW}}(\mathbf{R}(a_3)) = \log 9$. The optimal policy is therefore deterministic: $\pi^* = a_2$.

### B.1 Demonstrating ESR non-Markovianity

By changing the initial state-action pair $(s_0, a_0)$ such that the reward at $s_0$ becomes $(2, 0)$ instead of $(0, 0)$, the previously accumulated reward at $s_1$ shifts from $(0, 0)$ to $(2, 0)$. This changes each trajectory's return vector to $\mathbf{R}(a_1) = (11, 1)$, $\mathbf{R}(a_2) = (6, 4)$, $\mathbf{R}(a_3) = (3, 9)$, yielding $u_{\text{NSW}}(\mathbf{R}(a_1)) = \log 11$, $u_{\text{NSW}}(\mathbf{R}(a_2)) = \log 24$, $u_{\text{NSW}}(\mathbf{R}(a_3)) = \log 27$. The optimal action at $s_1$ shifts from $a_2$ to $a_3$, demonstrating that the ESR-optimal policy depends on the accumulated reward and is therefore non-Markovian in the original state space.

## C Derivation of the AETDICE Loss Functions

In this section, we provide the full derivation of the AETDICE loss functions. AETDICE assumes the augmented MOMDP $\tilde{\mathcal{M}}$ defined in Section 4 with the transformed reward:

$$\tilde{\mathbf{r}}(\tilde{s}_t, a_t) = F(\mathbf{R}_t^{\text{acc}} + \mathbf{r}(s_t, a_t)) - F(\mathbf{R}_t^{\text{acc}}) \ \in \ \mathbb{R}^n. \tag{14}$$

The AET-MORL objective $\max_{\tilde{\pi}} J_{\text{AET}}(\tilde{\pi}; F, G)$ is reformulated to a convex optimization in terms of $\mathbf{d} = \{d_t(\tilde{s}, a)\}_{t=0}^{H-1}$, the visitation distributions in $\tilde{\mathcal{M}}$ as

$$\max_{\mathbf{d} \geq 0} \quad G\left( \sum_{t=0}^{H-1} \sum_{\tilde{s},a} d_t(\tilde{s}, a)\, \tilde{\mathbf{r}}(\tilde{s}, a) \right) \tag{15a}$$

$$\text{s.t.} \quad \sum_a d_0(\tilde{s}, a) = \tilde{p}_0(\tilde{s}), \quad \forall \tilde{s} \tag{15b}$$

$$\sum_{a'} d_{t+1}(\tilde{s}', a') = \sum_{\tilde{s},a} d_t(\tilde{s}, a)\, \tilde{P}(\tilde{s}' \mid \tilde{s}, a), \quad \forall \tilde{s}',\, t \in \{0, \ldots, H-2\}, \tag{15c}$$

where $\tilde{s}' = (s', \mathbf{R}_t^{\text{acc}} + \mathbf{r}(s, a), t + 1)$. No flow constraint is imposed at $t = H - 1$, as all transitions from $d_{H-1}$ lead to the terminal state.

### C.1 Regularized Primal Problem

To enable sample-based optimization of the problem in offline setting, we follow the derivation from [11] by (i) adding $\phi$-divergence regularization against per-timestep data distribution $d_t^D$ and (ii) introducing a slack variable $\mathbf{k} \in \mathbb{R}^n$ to relocate the expected utility vector outside the concave

aggregation:

$$\max_{\mathbf{d}\geq 0,\, \mathbf{k}\in\mathbb{R}^n} \quad G(\mathbf{k}) - \beta \sum_{t=0}^{H-1} D_\phi(d_t \,\|\, d_t^D) \tag{16a}$$

$$\text{s.t.} \quad \sum_{t=0}^{H-1} \sum_{\tilde{s},a} d_t(\tilde{s},a)\, \tilde{\mathbf{r}}(\tilde{s},a) = \mathbf{k}, \tag{16b}$$

$$\sum_a d_0(\tilde{s},a) = \tilde{p}_0(\tilde{s}), \qquad \forall \tilde{s} \in \tilde{\mathcal{S}}, \tag{16c}$$

$$\sum_{a'} d_{t+1}(\tilde{s}',a') = \sum_{\tilde{s},a} d_t(\tilde{s},a)\, \tilde{P}(\tilde{s}' \mid \tilde{s},a), \quad \forall \tilde{s}',\, t \in \{0,\dots,H-2\}, \tag{16d}$$

**Assumptions.** We require: (i) $G$ is concave and continuous; (ii) $\phi$ is strictly convex, with $\phi(1) = 0$; and (iii) the data distribution satisfies $d_t^D(\tilde{s},a) > 0$ for all reachable $(\tilde{s},a)$ and $t$, so that the $\phi$-divergence is well-defined and Slater's condition is satisfied. Assumption (iii) is stronger than the corresponding coverage assumption in DICE-based offline RL [19], which only requires coverage over the original state-action space. In the augmented MOMDP, coverage must additionally hold over the accumulated reward $\mathbf{R}_t^{\text{acc}}$, yet the set of $\mathbf{R}_t^{\text{acc}}$ values reachable from a given state is generally a strict subset of $\mathbb{R}^m$. We analyze this requirement, together with a relabeling scheme that enlarges coverage without new data, in Appendix D.

**Role of the slack variable $\mathbf{k}$.** The slack variable $\mathbf{k}$ is a technique introduced in [11], which decouples the expectation over $\mathbf{d}$ from the concave aggregation $G$. By enforcing (16b), the expected transformed reward is computed separately and passed to $G$ through $\mathbf{k}$. This decoupling is essential for deriving an unbiased sample-based approximation of the Lagrangian dual: without $\mathbf{k}$, estimating $G\Big(\sum_t \sum_{\tilde{s},a} d_t(\tilde{s},a)\, \tilde{\mathbf{r}}(\tilde{s},a)\Big)$ from samples would require applying $G$ to an empirical average, which is biased by Jensen's inequality.

### C.2 Lagrangian Dual

We derive the Lagrangian dual of (16) by introducing multipliers $\nu_t(\tilde{s})$ for the Bellman flow constraints (16c)–(16d), and $\boldsymbol{\mu} \in \mathbb{R}^n$ for the slack constraint (16b):

$$\begin{aligned}\max_{\mathbf{d}\geq 0,\mathbf{k}} \min_{\boldsymbol{\nu},\boldsymbol{\mu}} \mathcal{L}(\mathbf{d},\mathbf{k},\boldsymbol{\nu},\boldsymbol{\mu}) = G(\mathbf{k}) &+ \boldsymbol{\mu}^\top \left( \sum_{t,\tilde{s},a} d_t(\tilde{s},a)\, \tilde{\mathbf{r}}(\tilde{s},a) - \mathbf{k} \right) - \beta \sum_{t=0}^{H-1} D_\phi(d_t \,\|\, d_t^D) \\ &+ \sum_{\tilde{s}} \nu_0(\tilde{s}) \left( \tilde{p}_0(\tilde{s}) - \sum_a d_0(\tilde{s},a) \right) \\ &+ \sum_{t=0}^{H-2} \sum_{\tilde{s}'} \nu_{t+1}(\tilde{s}') \left( \sum_{\tilde{s},a} \tilde{P}(\tilde{s}' \mid \tilde{s},a)\, d_t(\tilde{s},a) - \sum_a d_{t+1}(\tilde{s}',a) \right).\end{aligned} \tag{17}$$

To enable sample-based evaluation of the dual function, we reorganize the flow-constraint terms of the Lagrangian into expectations with respect to the visitation distribution $\mathbf{d}$:

$$\begin{aligned}&\sum_{\tilde{s}} \nu_0(\tilde{s})\, \tilde{p}_0(\tilde{s}) \;-\; \sum_{\tilde{s},a} \nu_0(\tilde{s})\, d_0(\tilde{s},a) \\ &\quad+\; \sum_{\tilde{s},a} d_0(\tilde{s},a) \sum_{\tilde{s}'} \tilde{P}(\tilde{s}' \mid \tilde{s},a)\, \nu_1(\tilde{s}') \;-\; \sum_{\tilde{s},a} \nu_1(\tilde{s})\, d_1(\tilde{s},a) \\ &\quad+\; \sum_{\tilde{s},a} d_1(\tilde{s},a) \sum_{\tilde{s}'} \tilde{P}(\tilde{s}' \mid \tilde{s},a)\, \nu_2(\tilde{s}') \;-\; \cdots \\ &\quad-\; \sum_{\tilde{s},a} \nu_{H-1}(\tilde{s})\, d_{H-1}(\tilde{s},a) + \sum_{\tilde{s},a} d_{H-1}(\tilde{s},a) \sum_{\tilde{s}'} \tilde{P}(\tilde{s}' \mid \tilde{s},a)\, \underbrace{\nu_H(\tilde{s}')}_{:=\,0}.\end{aligned} \tag{18}$$

where we adopt $\nu_H(\tilde{s}) := 0$ for all $\tilde{s} \in \tilde{\mathcal{S}}$ for notational uniformity. This enables the derivation of the reorganized Lagrangian dual:

$$\mathcal{L}(\mathbf{d}, \mathbf{k}, \boldsymbol{\nu}, \boldsymbol{\mu}) = G(\mathbf{k}) - \boldsymbol{\mu}^\top \mathbf{k} - \beta \sum_{t=0}^{H-1} D_\phi(d_t \| d_t^D) + \sum_{\tilde{s}} \nu_0(\tilde{s}) \tilde{p}_0(\tilde{s}) + \sum_{\tilde{s},a,t} d_t(\tilde{s}, a) e_{\boldsymbol{\nu},\boldsymbol{\mu},t}(\tilde{s}, a)$$

where $e_{\boldsymbol{\nu},\boldsymbol{\mu},t}(\tilde{s}, a) = \boldsymbol{\mu}^\top \tilde{\mathbf{r}}(\tilde{s}, a) + \mathbb{E}_{\tilde{s}'}[\nu_{t+1}(\tilde{s}')] - \nu_t(\tilde{s})$.

### C.3 Switching the order of optimization via strong duality

We leverage strong duality of (16) enabled by the data coverage assumption and satisfaction of Slater's condition to swap the optimization order to $\min_{\boldsymbol{\nu},\boldsymbol{\mu}} \max_{\mathbf{d} \geq 0, \mathbf{k}} \mathcal{L}$. This yields a closed-form solution for $\mathbf{d}$ and enables simplification of the dual function via conjugate functions.

We define the per-timestep visitation ratio $w_t(\tilde{s}, a) := \frac{d_t(\tilde{s},a)}{d_t^D(\tilde{s},a)}$ prior to obtaining the closed-form solution to further simplify the Lagrangian:

$$\mathcal{L} = G(\mathbf{k}) - \boldsymbol{\mu}^\top \mathbf{k} + \mathbb{E}_{\tilde{p}_0}[\nu_0(\tilde{s})] + \sum_{t=0}^{H-1} \mathbb{E}_{(\tilde{s},a) \sim d_t^D} \Big[ w_t(\tilde{s}, a)\, e_{\boldsymbol{\nu},\boldsymbol{\mu},t}(\tilde{s}, a) - \beta\, \phi(w_t(\tilde{s}, a)) \Big]. \tag{19}$$

Given $(\boldsymbol{\nu}, \boldsymbol{\mu}, \mathbf{k})$, the inner maximization in (19) decouples pointwise across $(t, \tilde{s}, a)$:

$$\max_{w \geq 0} \Big\{ w \cdot e_{\boldsymbol{\nu},\boldsymbol{\mu},t}(\tilde{s}, a) - \beta\, \phi(w) \Big\}. \tag{20}$$

The first-order condition $e_{\boldsymbol{\nu},\boldsymbol{\mu},t}(\tilde{s}, a) - \beta\, \phi'(w) = 0$ yields the closed-form maximizer

$$w_t^*(\tilde{s}, a) = \max\left(0,\ (\phi')^{-1}\left(\frac{e_{\boldsymbol{\nu},\boldsymbol{\mu},t}(\tilde{s}, a)}{\beta}\right)\right), \tag{21}$$

where the $\max(0, \cdot)$ enforces the non-negativity constraint.

**Fenchel-conjugate substitution.** Define the non-negative-restricted Fenchel conjugate of $\phi$,

$$\phi_+^*(y) := \sup_{w \geq 0} \big\{ wy - \phi(w) \big\}. \tag{22}$$

By a direct change of variables,

$$\sup_{w \geq 0} \big\{ wz - \beta\phi(w) \big\} = \beta \sup_{w \geq 0} \Big\{ w\, \tfrac{z}{\beta} - \phi(w) \Big\} = \beta\, \phi_+^*\left(\tfrac{z}{\beta}\right). \tag{23}$$

Substituting (23) with $z = e_{\boldsymbol{\nu},\boldsymbol{\mu},t}(\tilde{s}, a)$ into (19) eliminates the inner maximization over $w_t$:

$$\max_{\mathbf{d} \geq 0} \mathcal{L} = G(\mathbf{k}) - \boldsymbol{\mu}^\top \mathbf{k} + \mathbb{E}_{\tilde{p}_0}[\nu_0(\tilde{s})] + \sum_{t=0}^{H-1} \mathbb{E}_{d_t^D}\left[ \beta\, \phi_+^*\left(\frac{e_{\boldsymbol{\nu},\boldsymbol{\mu},t}(\tilde{s}, a)}{\beta}\right)\right]. \tag{24}$$

The remaining $\mathbf{k}$-dependent part of the Lagrangian is $G(\mathbf{k}) - \boldsymbol{\mu}^\top \mathbf{k}$, maximized unconstrained over $\mathbf{k} \in \mathbb{R}^n$. Following the convention used elsewhere in the paper, define the conjugate of $G$ as $G^*(\mathbf{y}) := \sup_{\mathbf{x}} \big\{ \mathbf{x}^\top \mathbf{y} + G(\mathbf{x}) \big\}$. Then

$$\max_{\mathbf{k} \in \mathbb{R}^n} \big\{ G(\mathbf{k}) - \boldsymbol{\mu}^\top \mathbf{k} \big\} = \max_{\mathbf{k} \in \mathbb{R}^n} \big\{ \mathbf{k}^\top (-\boldsymbol{\mu}) + G(\mathbf{k}) \big\} = G^*(-\boldsymbol{\mu}). \tag{25}$$

Combining the two conjugate formulations reduces the dual problem to a sample-based loss function given by,

$$\min_{\boldsymbol{\nu},\boldsymbol{\mu}} \mathcal{L}_{\mathrm{AET}}(\boldsymbol{\nu}, \boldsymbol{\mu}) := \mathbb{E}_{\tilde{p}_0}[\nu_0(\tilde{s})] + \sum_{t=0}^{H-1} \mathbb{E}_{(\tilde{s},a) \sim d_t^D}\left[ \beta\, \phi_+^*\left(\frac{e_{\boldsymbol{\nu},\boldsymbol{\mu},t}(\tilde{s}, a)}{\beta}\right)\right] + G^*(-\boldsymbol{\mu}). \tag{26}$$

Crucially, $\mathcal{L}_{\mathrm{AET}}$ is expressed entirely as expectations over $\tilde{p}_0$ and the offline dataset $\{d_t^D\}_{t=0}^{H-1}$. It does not require knowledge of the augmented transition kernel $\tilde{P}$ (the $\tilde{P}$-expectation inside $e_{\boldsymbol{\nu},\boldsymbol{\mu},t}$ is replaced by single-sample plug-in $\nu_{t+1}(\tilde{s}_{t+1})$ at offline transitions $(\tilde{s}_t, a_t, \tilde{s}_{t+1})$), nor does it require explicit access to the primal visitation distributions $\{d_t\}$. It can therefore be optimized using standard stochastic gradient descent on neural-network parametrizations $\nu_\psi$ and $\boldsymbol{\mu}$.

### C.4 Policy Extraction

After minimizing (26) to obtain $(\boldsymbol{\nu}^*, \boldsymbol{\mu}^*)$, the optimal visitation ratio $w_t^*$ is recovered from (21) using $e_{\boldsymbol{\nu}^*,\boldsymbol{\mu}^*,t}$. The optimal augmented-state policy $\tilde{\pi}_t^*(a \mid \tilde{s})$ inducing the optimal visitation $d_t^* = w_t^*\, d_t^D$ is then extracted. In the tabular case, this reduces to direct normalization: $\tilde{\pi}_t^*(a \mid \tilde{s}) = \frac{w_t^*(\tilde{s},a)\, d_t^D(\tilde{s},a)}{\sum_{a'} w_t^*(\tilde{s},a')\, d_t^D(\tilde{s},a')}$. In the continuous case, we use weighted behavior cloning following [19]:

$$\max_{\tilde{\pi}} \sum_{t=0}^{H-1} \mathbb{E}_{(\tilde{s},a)\sim d_t^D}[w_t^*(\tilde{s},a)\, \log \tilde{\pi}(a \mid \tilde{s})]\,. \tag{27}$$

### C.5 Choice of $\phi$-Divergence and Aggregation $G$

We use the $\chi^2$-divergence for $\phi$ and a piecewise-log function for the concave aggregation $G$.

**$\chi^2$ divergence.** $\phi(x) = \frac{1}{2}(x-1)^2$ for $x \geq 0$. Its convex conjugate is

$$\phi_+^*(y) \;=\; \tfrac{1}{2}\big[\,[y+1]_+\,\big]^2 \;-\; \tfrac{1}{2}, \qquad w_t^*(\tilde{s},a) \;=\; \big[\,e_{\boldsymbol{\nu},\boldsymbol{\mu},t}(\tilde{s},a)/\beta + 1\,\big]_+, \tag{28}$$

where $[\,\cdot\,]_+ := \max(0,\cdot)$.

**Piecewise-log aggregation.** In practice, we avoid using $G(\mathbf{x}) = \sum_i \log(x_i)$ directly, as the logarithm is undefined for non-positive values that may arise during optimization. Instead, we use a piecewise extension $G(\mathbf{x}) = \sum_{i=1}^n g(x_i)$ with

$$g(x) \;=\; \begin{cases} \log(x), & x \geq 1, \\ -\frac{1}{2}(x-2)^2 + \frac{1}{2}, & x < 1. \end{cases} \tag{29}$$

which is concave, strictly increasing, and well-defined on all of $\mathbb{R}$. With $G^*(\boldsymbol{p}) := \sup_{\mathbf{x}}\{\boldsymbol{p}^\top \mathbf{x} + G(\mathbf{x})\}$, its conjugate is $G^*(-\boldsymbol{\mu}) = \sum_{i=1}^n g^*(-\mu_i)$ with

$$g^*(-\mu) \;=\; \begin{cases} -1 - \log(\mu), & 0 < \mu < 1, \\ \frac{1}{2}\mu^2 - 2\mu + \frac{1}{2}, & \mu \geq 1. \end{cases} \tag{30}$$

### C.6 Full Algorithm of AETDICE

We present the full algorithm of AETDICE in Algorithm 1. A key implementation consideration is that the Lagrangian multipliers $\nu_t(\tilde{s})$ are defined per timestep, and the neural network must be able to represent this time dependence. In standard offline RL datasets, trajectories are stored with timestep information, but each sample's contribution to the loss does not depend on when it occurs within the episode. This allows data augmentation along the time axis, which we describe in Appendix E.

## D Coverage of the Augmented State Space and Number of Objectives

AETDICE conditions the dual variables and the policy on $\tilde{s}_t = (s_t, \mathbf{R}_t^{\mathrm{acc}}, t)$, so the offline data must cover accumulated rewards in addition to state-action pairs. This appendix examines how this requirement depends on the number of objectives (Section D.1), describes a relabeling scheme that enlarges coverage without new data (Section D.2), measures the effect of coverage in a tabular environment where it can be computed exactly (Section D.3), and evaluates AETDICE with 100 objectives in a continuous domain (Section D.4).

### D.1 Objectives that enlarge the augmented state

Partition the $m$ objectives into $m_1$ objectives that enter $F$ nonlinearly and $m_2 = m - m_1$ objectives that enter $F$ only linearly. The transformed reward (5) does not depend on the accumulated reward of the latter, so only the $m_1$ objectives need to be included in $\mathbf{R}_t^{\mathrm{acc}}$. AETDICE learns $\nu_t(\tilde{s})$ over the augmented state and $\boldsymbol{\mu} \in \mathbb{R}^n$. The augmented state grows by one dimension per objective in $m_1$; neither $m_2$ nor the number of utilities $n$ affects $\nu$, and $n$ enlarges only $\boldsymbol{\mu}$, a vector of scalars rather than a function of the state. Increasing the number of objectives therefore does not enlarge the networks substantially; what grows with $m_1$ is the region of accumulated-reward space that a fixed dataset must cover.

**Algorithm 1** AETDICE: Offline AET Optimization via Distribution Correction Estimation

**Require:** Offline dataset $\mathcal{D} = \{(s_t, a_t, \mathbf{r}_t, s_{t+1}, t)\}$ from the original MOMDP $\mathcal{M}$; transformation $F : \mathbb{R}^m \to \mathbb{R}^n$; aggregation $G : \mathbb{R}^n \to \mathbb{R}$; dual network $\nu_\psi$, policy network $\tilde{\pi}_\theta$; regularization coefficient $\beta > 0$; learning rate $\eta$

**Ensure:** Augmented-state policy $\tilde{\pi}_\theta$ approximating the AET-optimal policy

**// Stage 1: Dataset preprocessing (one-time)**

Construct augmented dataset $\widetilde{\mathcal{D}}$: for each trajectory in $\mathcal{D}$, compute accumulated rewards $\mathbf{R}_t^{\mathrm{acc}} = \sum_{k=0}^{t-1} \mathbf{r}(s_k, a_k)$ and form augmented transitions $(\tilde{s}_t, a_t, \tilde{\mathbf{r}}_t, \tilde{s}_{t+1})$ with $\tilde{s}_t = (s_t, \mathbf{R}_t^{\mathrm{acc}}, t)$

Compute the transformed reward for each transition:

$$\tilde{\mathbf{r}}_t \;=\; F(\mathbf{R}_t^{\mathrm{acc}} + \mathbf{r}(s_t, a_t)) - F(\mathbf{R}_t^{\mathrm{acc}}) \;\in\; \mathbb{R}^n$$

**// Stage 2: Initialization**

Initialize $\psi$, $\theta$

**if** $G$ is nonlinear **then**

  Initialize $\boldsymbol{\mu} \in \mathbb{R}^n$

**else**

  Fix $\boldsymbol{\mu} = \mathbf{1}$ or preference weights of a choice (no slack constraint needed when $G$ is linear)

**end if**

**// Stage 3: Joint dual / policy optimization**

**while** not converged **do**

  Sample minibatch $\mathcal{B} = \{(\tilde{s}_t, a_t, \tilde{\mathbf{r}}_t, \tilde{s}_{t+1})\}$ from $\widetilde{\mathcal{D}}$ and initial states $\tilde{s}_0 \sim \tilde{p}_0$

  Compute empirical slack-augmented advantage at each transition:

$$\hat{e}_{\boldsymbol{\mu},\psi,t}(\tilde{s}_t, a_t) \;=\; \boldsymbol{\mu}^\top \tilde{\mathbf{r}}_t \;+\; \nu_\psi(\tilde{s}_{t+1}) \;-\; \nu_\psi(\tilde{s}_t)$$

  (single-sample plug-in for $\mathbb{E}_{\tilde{s}' \sim \tilde{P}}[\nu_{t+1}(\tilde{s}')]$; boundary $\nu_\psi(\tilde{s}_H) := 0$ at terminal transitions)

  Update dual parameters $\psi$ (and $\boldsymbol{\mu}$ if $G$ is nonlinear) by minimizing the empirical AET dual loss:

$$\widehat{\mathcal{L}}_{\mathrm{AET}}(\psi, \boldsymbol{\mu}) \;=\; \mathbb{E}_{\tilde{s}_0 \sim \tilde{p}_0}\big[\nu_\psi(\tilde{s}_0)\big] \;+\; \mathbb{E}_{(\tilde{s}_t, a_t, \tilde{s}_{t+1}) \sim \widetilde{\mathcal{D}}}\left[\beta\, \phi_+^*\!\left(\frac{\hat{e}_{\boldsymbol{\mu},\psi,t}(\tilde{s}_t, a_t)}{\beta}\right)\right] \;+\; G^*(-\boldsymbol{\mu})$$

  via $\psi \leftarrow \psi - \eta \nabla_\psi \widehat{\mathcal{L}}_{\mathrm{AET}}$ and $\boldsymbol{\mu} \leftarrow \boldsymbol{\mu} - \eta \nabla_{\boldsymbol{\mu}} \widehat{\mathcal{L}}_{\mathrm{AET}}$

  Compute optimal visitation ratio at sampled transitions:

$$w^*(\tilde{s}_t, a_t) \;=\; \max\!\left(0,\; (\phi')^{-1}\!\left(\frac{\hat{e}_{\boldsymbol{\mu},\psi,t}(\tilde{s}_t, a_t)}{\beta}\right)\right)$$

  Update policy parameters $\theta$ via weighted behavior cloning:

$$\mathcal{L}_\theta \;=\; -\,\mathbb{E}_{(\tilde{s}_t, a_t) \sim \widetilde{\mathcal{D}}}[w^*(\tilde{s}_t, a_t)\, \log \tilde{\pi}_\theta(a_t \mid \tilde{s}_t)]$$

  via $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}_\theta$

**end while**

**return** $\tilde{\pi}_\theta$

### D.2 Relabeling accumulated rewards

Since $\mathbf{R}_{t+1}^{\mathrm{acc}} = \mathbf{R}_t^{\mathrm{acc}} + \mathbf{r}(s_t, a_t)$ and the transition of $s_t$ does not depend on $\mathbf{R}_t^{\mathrm{acc}}$, a transition recorded at one accumulated reward remains a valid transition of $\tilde{\mathcal{M}}$ when stored at another value $\mathbf{R}'$, with next value $\mathbf{R}' + \mathbf{r}(s_t, a_t)$ and $\tilde{\mathbf{r}}$ recomputed at $\mathbf{R}'$. Each transition can thus be replicated across alternative accumulated rewards, enlarging coverage of the augmented dimensions from the same trajectories. Relabeling enlarges only these dimensions and cannot cover original states that the dataset never visits. In continuous domains, $R_i'$ can be sampled in $[0, R_i^{\max}]$ for each objective $i$, which may also produce data outside the attainable region. Relabeled values of $\mathbf{R}_t^{\mathrm{acc}}$ may or may not lie in the region traversed by the optimal visitation distribution, so the benefit of relabeling depends on the objective and the dataset. The experiments in Section 6 do not use relabeling.

### D.3 Coverage and performance in a tabular environment

**Setup.** We use a $5 \times 5$ deterministic gridworld with four actions and horizon $H = 5$, where entering a cell yields a reward vector in $\{0, 1\}^m$ fixed by the seed. The trajectory-level utility is $F(\mathbf{R}) = \sum_{i=1}^{m_1} \log(1 + R_i) + \sum_{i=m_1+1}^{m} R_i$ and $G$ is linear, so the AET objective reduces to $\mathbb{E}[F(\mathbf{R}(\tau))]$, whose exact optimum is obtained by dynamic programming over the augmented state. The dataset consists of $N \in \{30, 100\}$ trajectories collected by a uniformly random policy. Coverage is the fraction of the exactly enumerated set of attainable $(s, \mathbf{R}_t^{\mathrm{acc}}, t)$ that the dataset visits. We compare no relabeling, relabeling within a radius-1 neighborhood of the observed $\mathbf{R}_t^{\mathrm{acc}}$, and relabeling over all values attainable at that timestep (full). All runs use $\beta = 10^{-3}$ and five seeds; each seed fixes the environment and the dataset, so relabeling levels are compared within seed. Policies are evaluated exactly, and we report the normalized objective $(V - V_{\mathrm{beh}})/(V^* - V_{\mathrm{beh}})$, where $V_{\mathrm{beh}}$ and $V^*$ are the values of the behavior policy and the optimal policy.

Table 2: Normalized AET objective in the tabular environment (mean $\pm$ standard error over five seeds), with the coverage of attainable augmented states in parentheses.

| | | $N = 30$ | | | $N = 100$ | | |
|---|---|---|---|---|---|---|---|
| $m_1$ | $m_2$ | original | radius 1 | full | original | radius 1 | full |
| 3 | 0 | $0.68_{\pm 0.06}$ (32%) | $0.82_{\pm 0.08}$ (69%) | $0.80_{\pm 0.12}$ (74%) | $0.81_{\pm 0.03}$ (61%) | $0.86_{\pm 0.07}$ (93%) | $0.92_{\pm 0.03}$ (95%) |
| 3 | 97 | $0.69_{\pm 0.04}$ (32%) | $0.80_{\pm 0.07}$ (69%) | $0.83_{\pm 0.08}$ (74%) | $0.83_{\pm 0.02}$ (61%) | $0.90_{\pm 0.03}$ (93%) | $0.92_{\pm 0.03}$ (95%) |
| 4 | 0 | $0.77_{\pm 0.07}$ (26%) | $0.78_{\pm 0.07}$ (70%) | $0.82_{\pm 0.05}$ (78%) | $0.85_{\pm 0.02}$ (55%) | $0.86_{\pm 0.04}$ (93%) | $0.87_{\pm 0.03}$ (96%) |
| 5 | 0 | $0.83_{\pm 0.07}$ (24%) | $0.85_{\pm 0.06}$ (70%) | $0.86_{\pm 0.05}$ (79%) | $0.87_{\pm 0.06}$ (52%) | $0.91_{\pm 0.01}$ (93%) | $0.93_{\pm 0.02}$ (96%) |

**Effect of coverage.** Relabeling raises coverage from 24–32% to 69–79% at $N = 30$ and from 52–61% to 93–96% at $N = 100$ (Table 2). Averaged over the four configurations and paired within seed, full relabeling improves the normalized objective by $+0.086 \pm 0.025$ at $N = 30$ and $+0.070 \pm 0.021$ at $N = 100$, and radius-1 relabeling by $+0.067 \pm 0.023$ at $N = 30$. The gain is smaller at $N = 100$, where less coverage is missing initially, and does not grow with $m_1$ over the tested range; the configuration with $m = 100$ ($m_1 = 3$) behaves as the one with $m = 3$.

**Where the learned policy acts.** Table 3 reports the visitation mass of the learned policy, computed exactly over the augmented states it reaches, for $N = 30$ and $m_1 = 3$. The policy places no mass on accumulated rewards absent from the dataset, consistent with the support constraint imposed by the $\phi$-divergence. At 32% coverage, 28% of its mass falls in the 5% of covered augmented states with the lowest frequency in the dataset, which relabeling reduces to below 0.3%. Limited coverage thus restricts the attainable objective rather than leading the policy to act at unsupported accumulated rewards.

Table 3: Visitation mass of the learned policy in the tabular environment ($N = 30$, $m_1 = 3$, five seeds).

| dataset | coverage | mass at absent $\mathbf{R}_t^{\mathrm{acc}}$ | mass in lowest-density 5% |
|---|---|---|---|
| original | 31.8% | 0.00% | 28.00% |
| radius 1 | 69.2% | 0.00% | 0.28% |
| full | 74.4% | 0.00% | 0.09% |

### D.4 100 objectives on MO-Ant

On MO-Ant, with original rewards $\mathbf{r}^{(t)} = (r_1^{(t)}, r_2^{(t)})$, we construct 100 objectives by sampling weight vectors $\mathbf{w}_j = (w_{1j}, w_{2j})$ with $w_{1j} \sim \mathcal{N}(1, 0.5)$ and $w_{2j} \sim \mathcal{N}(4, 0.5)$, truncated to positive values, and defining the $j$-th reward as $r_j^{(t)} = \frac{1}{2}\big(\tanh(\mathbf{w}_j^\top \mathbf{r}^{(t)}) + 1\big) \in [0, 1]$ for $j = 1, \ldots, 100$. Each objective enters its own convex utility $f_j(\mathbf{R}) = \exp(R_j/50)$, as in Section 6, giving $n = 100$ utilities and $m_1 = 100$, so the augmented state has dimension $|s| + 100$. We aggregate the utilities with linear and concave ($\log$) $G$ and compare AETDICE with behavior cloning (BC) on the expert-uniform split of D4MORL (Appendix F.3). $\nu$ is an MLP with four hidden layers of 512 units, and $\boldsymbol{\mu} \in \mathbb{R}^{100}$.

Table 4: AET objective on MO-Ant with 100 objectives (mean $\pm$ standard error over five seeds).

| $G$ | $F$ | BC | AETDICE |
|---|---|---|---|
| linear | exp | $2378.2 \pm 71.7$ | $\mathbf{3049.4 \pm 78.7}$ |
| log | exp | $315.5 \pm 2.8$ | $\mathbf{370.2 \pm 2.3}$ |

AETDICE improves over BC under both objectives with $m_1 = 100$. The 100 objectives are derived from the two MO-Ant rewards and are therefore correlated; experiments with independently defined objectives remain for future work.

## E Time-Index Augmentation in the Augmented MOMDP

We describe the data augmentation procedure used to support learning in the augmented MOMDP. The goal is to enrich the dataset with additional time-indexed contexts while preserving the original state-action distribution.

### E.1 Dataset and Augmentation Strategy

We begin with a dataset $\mathcal{D}_0$ collected by a behavior policy $\pi_b$, consisting of tuples $(s_t, a_t, \mathbf{R}_t^{\mathrm{acc}}, t)$, where $\mathbf{R}_t^{\mathrm{acc}} = \sum_{k=0}^{t-1} \mathbf{r}(s_k, a_k)$ is recorded during data collection. No assumptions are made about the optimality of $\pi_b$. Augmentation is performed solely over the discrete time index $t \in \{0, 1, \ldots, H\}$; all other components of the augmented state—including $s$ and $\mathbf{R}^{\mathrm{acc}}$—are taken directly from $\mathcal{D}_0$ and are never resampled or synthetically modified.

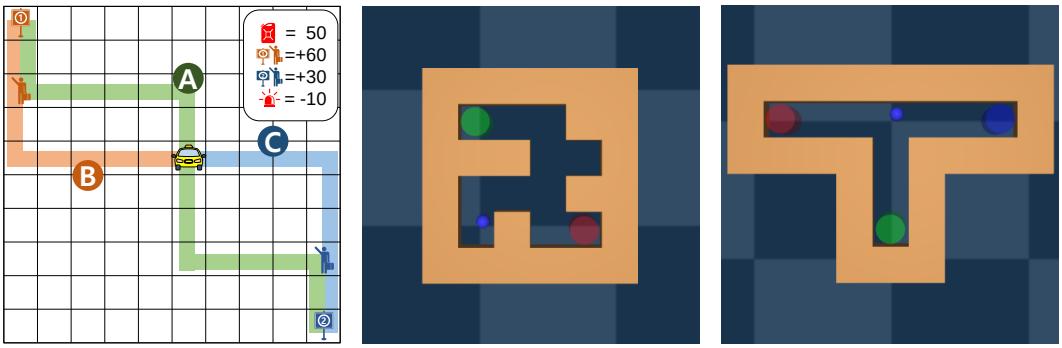
### E.2 Multi-Head Lagrangian Multiplier

Augmentation is implemented through the parameterization of the Lagrangian multiplier $\nu(s, \mathbf{R}^{\mathrm{acc}}, t)$. The neural network consists of a shared trunk processing $(s, \mathbf{R}^{\mathrm{acc}})$, followed by a multi-head output indexed by time, producing $[\nu(s, \mathbf{R}^{\mathrm{acc}}, 0), \ldots, \nu(s, \mathbf{R}^{\mathrm{acc}}, H)]$. For each data point $(s, \mathbf{R}^{\mathrm{acc}})$ from $\mathcal{D}_0$, the full set of time-indexed outputs is available during training. This exposes the dual objective to a richer collection of time-conditioned contexts without altering the underlying state-action distribution, and ensures that the multipliers are well-trained across the entire episode horizon. An alternative is to provide $t$ as a continuous input, which keeps the network size independent of $H$ but forgoes the reuse of each data point across time steps; on MO-Ant, this variant lowers the AETDICE objective from 13.67 to 13.02.

## F Environment Details

### F.1 Finite Fair-Taxi

The Finite Fair-Taxi environment is adapted from the multi-objective taxi domain of Fan et al. [13] and formulated as a finite-horizon grid-world task. A taxi operates on a square grid and serves passengers associated with different origin–destination pairs, where each passenger group defines one reward dimension. Because only one passenger can be transported at a time and episodes have a fixed horizon, the agent must allocate time across groups, introducing a fairness trade-off between objectives. All experiments use a two-objective setting.




(a) Fair-Taxi (b) MO-PointMaze-2obj (c) MO-PointMaze-3obj

Figure 7: Visualization of the multi-objective environments. Each environment balances multiple objectives over a finite horizon ($H = 50$ for Fair-Taxi and $H = 300$ for MO-PointMaze).

**State and Action Spaces.** The state is discrete and encodes the taxi location $(x, y)$, a binary indicator of whether a passenger is onboard, and the index of the passenger group currently being served (or a special value if none). The action space consists of six discrete actions: move north, south, east, west, pick up, and drop off. Movement actions are clipped at grid boundaries. Pickup is valid only when the taxi is at an origin location and not already carrying a passenger, and drop-off is valid only when the taxi is at the corresponding destination of the onboard passenger.

**Reward Function.** Rewards are vector-valued with dimension $m = 2$. At each time step, the reward vector $\mathbf{r}_t \in \mathbb{R}^2$ is defined as follows: a successful delivery for group $i$ yields $r_t^{(i)} = 30$ and $r_t^{(j)} = 0$ for $j \neq i$; invalid pickup or drop-off actions incur $r_t^{(1)} = r_t^{(2)} = -10$; all other transitions produce zero reward. To introduce asymmetry between objectives, the rewards and penalties for group 1 are scaled by a factor of two, resulting in $r_t^{(1)} \in \{60, -20, 0\}$ while group 2 retains $\{30, -10, 0\}$. Rewards are therefore sparse and event-driven.

**Episode Termination and Layout.** Episodes terminate after a fixed horizon of 50 time steps. The two origin–destination pairs are placed in distant regions of the grid, with each origin uniquely paired to a destination. Serving one group typically requires traveling in a direction that does not overlap with the other, so dedicating time to one group reduces the opportunity to serve the other. This creates a temporal trade-off between objectives within each episode.

**Offline Dataset Construction** For the Fair-Taxi domain, offline datasets are generated using a mixture of single-objective policies. Each policy is a tabular Q-function optimized for a different linear scalarization of the reward vector. During data collection, policies are alternated within each episode according to fixed time intervals, producing trajectories that contain segments optimized for different objectives. This switching scheme encourages diverse return profiles across episodes and prevents the dataset from being dominated by trajectories specialized to a single objective. Episodes start from a fixed initial taxi location to ensure comparability across policies. Each episode runs for at most 50 steps. The dataset stores state, action, next state, reward vector, timestep index, terminal flag, and accumulated return vector at each step.

### F.2 MO-PointMaze

The MO-PointMaze environment is a continuous-state multi-objective navigation task adapted from the D4RL maze domains [30]. A point-mass agent moves in a 2D maze containing multiple spatial goal regions, where each goal corresponds to one reward dimension. Because the goals are spatially separated by walls and corridors, the agent must allocate time across regions, inducing a trade-off between objectives within each episode.

**State and Action Spaces.** The observation space is inherited from the underlying D4RL PointMaze environment and includes the agent's 2D position and velocity. Experiments use a discretized action space consisting of movements along the four cardinal directions with several fixed magnitudes. This

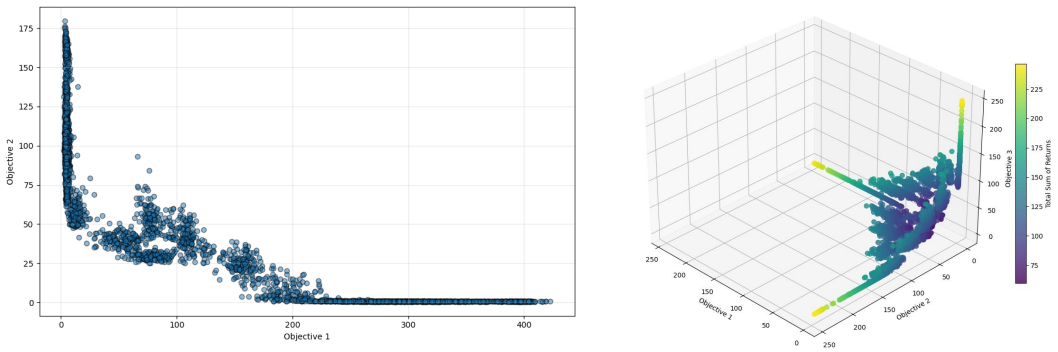



Figure 8: Visualization of the offline dataset return distributions for the MO-PointMaze environment. **(Left)** The cumulative returns in the 2-objective setting, illustrating the trade-off between the two goals. **(Right)** The spatial distribution of returns in the 3-objective setting.

discretization stabilizes training and ensures consistent coverage during offline data collection. The episode horizon is set to 300 steps.

**Maze Layouts.** Two layouts are used. In the two-objective setting (Figure 7b), the agent starts in the bottom-left region of a compact maze with two goal regions located in different corridors. In the three-objective setting (Figure 7c), the maze contains three goals positioned in distinct regions. Reaching one goal generally requires moving away from the others, making simultaneous proximity infeasible.

**Reward Function.** Rewards are vector-valued with dimension equal to the number of goals $m$. At each time step, the reward for objective $i$ depends on the Euclidean distance between the agent's position $\mathbf{s}_t$ and the $i$-th goal location $\mathbf{g}_i$:

$$r_t^{(i)} = \exp\left(-\frac{\|\mathbf{s}_t - \mathbf{g}_i\|}{d_0}\right),$$

where $d_0$ is a scaling constant proportional to the maze cell size. Rewards decay rapidly with distance, so meaningful reward in one dimension requires the agent to remain near the corresponding goal region for a sustained period.

In the two-objective setting, the first reward dimension is scaled by a factor of two to introduce asymmetry between objectives. In the three-objective setting, all reward dimensions are equally scaled.

**Multi-Objective Trade-off.** Due to the rapid spatial decay of rewards and the maze topology, the agent cannot remain close to multiple goals at the same time. Policies must therefore schedule visits to different regions within an episode, creating a temporal trade-off between objectives. This makes the environment well suited for analyzing differences between trajectory-level and expectation-level scalarization.

**Offline Dataset Collection.** Offline datasets are generated using heuristic navigation policies that move toward specific goal regions following shortest-path routes. By varying goal targets and switching patterns over time, the dataset covers diverse behaviors, including specialization on a single objective and sequential visitation of multiple goals. This diversity ensures that the dataset contains trajectories with a wide range of return trade-offs, which is essential for evaluating multi-objective offline RL methods. The resulting return distributions are visualized in Figure 8.

### F.3 MO-MuJoCo (D4MORL)

We use the multi-objective MuJoCo benchmark from D4MORL [9], which extends the standard MuJoCo locomotion suite by replacing scalar rewards with vector-valued multi-objective rewards while leaving the underlying physics, observation space, and action space unchanged. Three environments

are used: MO-Ant, MO-Walker2d, MO-HalfCheetah. For all environments we use the *expert-uniform* and *amateur-uniform* dataset splits provided by [9], in which the behavior-policy preferences are sampled uniformly from the 2-simplex during data collection; each split contains 50,000 trajectories.

**State and Action Spaces.** States and continuous actions are inherited from the underlying MuJoCo environments. State and action dimensions are Ant 27/8, HalfCheetah 17/6, and Walker2d 17/6. We follow the standard MuJoCo convention of clipping actions to $[-1, 1]$ before computing the reward.

**Reward Function.** At each time step the reward is a vector $\mathbf{r}_t \in \mathbb{R}^2$ whose two components capture distinct physical aspects of the task: forward and lateral velocity in MO-Ant, forward speed and energy saving in MO-HalfCheetah and MO-Walker2d. Both components include a shared survival bonus and a shared action-norm penalty, so the two objectives are positively oriented (higher is better) and partially correlated. We refer to [9] for exact formulae. To study the effect of asymmetry between objectives, the per-objective rewards are min–max normalized at dataset load time:

$$\bar{r}_t^{(i)} \;=\; c_i \cdot \frac{r_t^{(i)} - r_{\min}^{(i)}}{r_{\max}^{(i)} - r_{\min}^{(i)}},$$

where $r_{\min}^{(i)}, r_{\max}^{(i)}$ are computed over the offline dataset and $c_i \in \mathbb{R}_+$ is a per-objective scaling factor. We use the symmetric default $\mathbf{c} = (1, 1)$ for all environments and additionally $\mathbf{c} = (2, 1)$ in MO-Ant for an objective-1-favoring variant.

**Episode Termination and Horizon.** All environments use a fixed horizon of $H = 500$ steps. Standard MuJoCo early-termination conditions apply (height/angle bounds in MO-Walker2d, instability bounds in MO-Ant); transitions following early termination contribute zero reward and are masked out by the buffer.

**Reward Reshaping for the Cobb-Douglas Experiments.** The Cobb-Douglas utility $R_{\mathrm{v}}^{\rho}/R_{\mathrm{c}}^{1-\rho}$ measures how far the agent moves per unit of control effort, so its numerator should be a reward that the agent wants to increase and its denominator a cost that it wants to decrease. In MO-Walker2d and MO-HalfCheetah, the first D4MORL objective is already such a reward: the velocity reward $r^v = v_x + r^{\text{alive}}$. The second objective, however, is an energy-saving reward $r^e = 4 - \|\mathbf{a}\|^2 + r^{\text{alive}}$, in which the control cost $\|\mathbf{a}\|^2$ is subtracted from a constant and combined with the survival bonus $r^{\text{alive}} = 1$. Because the constant and the bonus are known, we recover the control cost exactly and keep the velocity reward unchanged:

$$r_{\mathrm{v}} \;=\; r^v, \qquad r_{\mathrm{c}} \;=\; 5 - r^e + \epsilon \;=\; \|\mathbf{a}\|^2 + \epsilon,$$

where a small constant $\epsilon > 0$ keeps the denominator positive. For the safety experiment on MO-Walker2d, we add a third reward, the height margin, i.e., the distance of the torso height $z_t$ to the nearer of its termination bounds, normalized to $[0, 1]$:

$$r_{\mathrm{s}}(s_t, a_t) \;=\; m_t^h \;=\; \mathrm{clip}\Big(\tfrac{\min(z_t - 0.8,\, 2.0 - z_t)}{0.6},\, 0,\, 1\Big),$$

which is 1 at the center of the allowed range ($z_t = 1.4$) and 0 at either bound (0.6 is half the range width). The reward vector is then $\mathbf{r} = (r_{\mathrm{v}}, r_{\mathrm{c}}, r_{\mathrm{s}})$, and its return $\mathbf{R}(\tau) = (R_{\mathrm{v}}, R_{\mathrm{c}}, R_{\mathrm{s}})$ consists of the velocity return, the control-cost return, and the safety return used in Eq. (12). Appendix G.3 explains why the safety reward uses the height rather than the angle.

# G Additional Experiment Results

## G.1 Fair-Taxi

We report quantitative performance on the Finite Fair-Taxi environment under ESR, BSR ($\alpha = 0.5$), and SER objectives. All values are log-scaled utilities (mean $\pm$ standard deviation over three seeds), so numerical differences reflect multiplicative gaps in the underlying objective values. Each policy is trained using the same AET framework with a different nonlinear objective and evaluated under all three criteria; entry $(i, j)$ in Table 5 reports the performance of a policy trained for objective $i$ when evaluated using objective $j$.

Across all objectives, policies achieve their highest performance when evaluated under the same criterion used during training, indicating that ESR, BSR, and SER correspond to distinct optimization targets in practice. Policies trained under BSR exhibit intermediate behavior: while optimized for BSR, they retain comparatively higher performance under ESR and SER evaluations than policies trained under the extreme objectives. Overall, these results demonstrate that nonlinear objectives are not interchangeable, and that objective mismatch can lead to substantial degradation even when numerical differences appear small due to log scaling.

Table 5: Performance of offline Fair MORL in Fair-Taxi environment. Rows correspond to the nonlinear objective used for training, and columns correspond to the evaluation objective. All values are log-scaled utilities (mean $\pm$ standard deviation over 3 random seeds).

| METHOD | ESR | BSR | SER |
|---|---|---|---|
| AET-ESR | $\mathbf{1.197}_{\pm 0.002}$ | $-0.797_{\pm 0.002}$ | $1.130_{\pm 0.000}$ |
| AET-BSR | $0.193_{\pm 0.011}$ | $\mathbf{-0.788}_{\pm 0.004}$ | $1.748_{\pm 0.009}$ |
| AET-SER | $0.110_{\pm 0.034}$ | $-1.045_{\pm 0.035}$ | $\mathbf{1.804}_{\pm 0.002}$ |

## G.2 MO-PointMaze

We next examine how these behavioral differences translate into return trade-offs in a continuous-state setting. Figure 9 shows the achieved return vectors. In the two-objective case, ESR produces balanced trajectories along the diagonal trade-off, whereas SER concentrates near extreme regions that prioritize a single objective per episode. BSR ($\alpha = 0.5$) spans intermediate trade-offs. The same pattern is more pronounced in the three-objective setting, where SER lies near the corners, ESR occupies the balanced interior, and BSR fills the space between. ESR-IQL aligns with ESR-style trade-offs, while AETDICE covers the full spectrum.

Table 6 reports performance under multiple evaluation criteria (ESR, BSR, SER, and LSR), reflecting different preference structures. The AETDICE variants achieve strong performance on their respective target objectives, demonstrating that the proposed framework can effectively optimize a range of nonlinear criteria within a unified approach. In the three-objective setting, ESR-IQL attains the highest ESR score, consistent with its specialization. OptiDICE-$\mathbf{w}$ achieves high LSR, as expected, but yields lower performance under nonlinear utilities, often producing policies that prioritize a single objective rather than balancing across objectives.

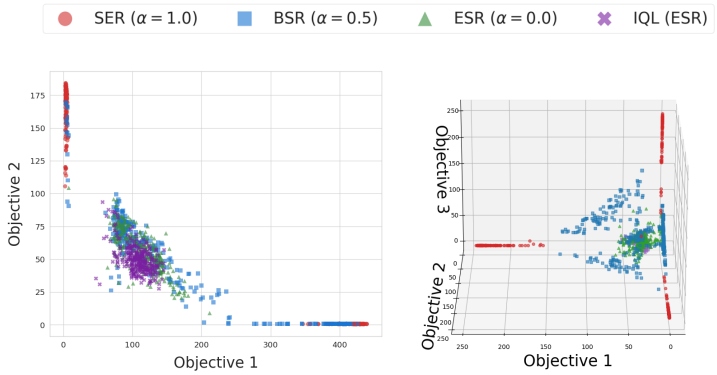



Figure 9: Performance results in the MO-PointMaze-2obj (left) and MO-PointMaze-3obj (right) environment.

## G.3 D4MORL

**Experiment setup.** The D4MORL experiments follow the four configurations of Section 6.2. Configurations (i) and (ii), convex $F$ with linear and concave $G$, are evaluated on MO-Ant and MO-Walker2d with the original D4MORL rewards. Configuration (iii), the Cobb–Douglas (CD) utility, is evaluated on MO-Walker2d and MO-HalfCheetah. Since the CD utility divides the velocity

Table 6: Comparison of ESR, BSR, SER, and LSR scores in MO-PointMaze, where LSR denotes the linear scalarized return with unit weights. Each row is a policy optimized for a different objective, evaluated under all criteria (mean ± standard deviation over 3 random seeds).

| METHOD | ESR | $\text{BSR}_{0.5}$ | SER | LSR |
|---|---|---|---|---|
| **MO-POINTMAZE-2OBJ** | | | | |
| AET-ESR | $\mathbf{6.67}_{\pm 0.08}$ | $11.90_{\pm 0.17}$ | $8.73_{\pm 0.08}$ | $164.86_{\pm 6.68}$ |
| AET-$\text{BSR}_{0.5}$ | $5.91_{\pm 0.32}$ | $\mathbf{11.93}_{\pm 0.11}$ | $9.01_{\pm 0.04}$ | $211.79_{\pm 12.29}$ |
| AET-SER | $4.06_{\pm 0.05}$ | $11.82_{\pm 0.05}$ | $\mathbf{9.74}_{\pm 0.02}$ | $284.05_{\pm 1.42}$ |
| OPTIDICE-$[0.5, 0.5]$ | $3.64_{\pm 0.01}$ | $8.77_{\pm 0.02}$ | $3.81_{\pm 0.02}$ | $\mathbf{417.01}_{\pm 3.13}$ |
| OPTIDICE-$[1.0, 0.0]$ | $3.63_{\pm 0.02}$ | $8.71_{\pm 0.04}$ | $3.82_{\pm 0.03}$ | $409.62_{\pm 5.70}$ |
| OPTIDICE-$[0.75, 0.25]$ | $3.67_{\pm 0.02}$ | $8.91_{\pm 0.19}$ | $4.53_{\pm 1.19}$ | $415.88_{\pm 6.09}$ |
| ESR-IQL | $6.56_{\pm 0.01}$ | $11.59_{\pm 0.03}$ | $8.56_{\pm 0.02}$ | $153.78_{\pm 1.49}$ |
| **MO-POINTMAZE-3OBJ** | | | | |
| AET-ESR | $6.79_{\pm 0.15}$ | $11.76_{\pm 0.10}$ | $10.15_{\pm 0.02}$ | $86.05_{\pm 0.46}$ |
| AET-$\text{BSR}_{0.5}$ | $4.72_{\pm 0.24}$ | $\mathbf{12.04}_{\pm 0.07}$ | $11.14_{\pm 0.05}$ | $120.34_{\pm 1.77}$ |
| AET-SER | $-0.19_{\pm 0.03}$ | $11.43_{\pm 0.03}$ | $\mathbf{12.81}_{\pm 0.02}$ | $\mathbf{213.78}_{\pm 1.38}$ |
| OPTIDICE-$[0.5, 0.5]$ | $-0.26_{\pm 0.07}$ | $10.61_{\pm 0.06}$ | $12.27_{\pm 0.01}$ | $176.92_{\pm 1.63}$ |
| OPTIDICE-$[1.0, 0.0]$ | $-0.34_{\pm 0.03}$ | $2.49_{\pm 0.06}$ | $2.17_{\pm 0.03}$ | $203.20_{\pm 4.01}$ |
| OPTIDICE-$[0.75, 0.25]$ | $-0.38_{\pm 0.01}$ | $2.41_{\pm 0.04}$ | $-2.20_{\pm 0.03}$ | $198.15_{\pm 3.22}$ |
| ESR-IQL | $\mathbf{7.20}_{\pm 0.03}$ | $12.03_{\pm 0.04}$ | $10.17_{\pm 0.02}$ | $87.16_{\pm 0.94}$ |

return by the control-cost return, it uses the velocity reward $r_{\text{v}}$ and the control cost $r_{\text{c}}$ in place of the original rewards from the D4MORL environment. Configuration (iv) adds the safety utility to the CD utility on MO-Walker2d. Under (iii) at low $\rho$, the agent lowers its control cost with a crouched gait and often falls (Figure 5(b)); configuration (iv) counters this with the height margin $r_{\text{s}}$ as a third reward, whose return is the safety return $R_{\text{s}}$.

**Choice of the safety reward.** An MO-Walker2d episode ends when the agent falls, i.e., when the torso height $z$ leaves $(0.8, 2.0)$ or the torso angle $\theta$ leaves $(-1, 1)$. To see which of the two causes the falls, we collect the 119 falling episodes from the evaluation rollouts of the eight trained policies that fall at least once (BC, ESR-IQL, and AET-CD at $\rho = 0.7$). In 115 of them, the torso height was closer to its bound than the angle just before termination. Table 7 also shows how early each candidate quantity warns of a fall, measured by the ROC AUC for telling apart the steps that are followed by a fall within $k$ steps from all other steps (1 is perfect, 0.5 is chance). The height margin warns best at every $k$ (0.99 at $k = 10$), while the angle-based quantities are at or below chance. We therefore use the height margin as a safety reward.

Table 7: How well each per-step quantity warns of a fall on MO-Walker2d. Each entry is the ROC AUC for telling apart the steps followed by a fall within $k$ steps from all other steps (1: perfect, 0.5: chance), averaged over the eight policies, with the minimum and maximum over the policies in brackets. A minus sign means that the quantity is negated so that a larger value means a fall is closer, e.g., a smaller height margin.

| Signal | $k = 10$ | $k = 25$ | $k = 50$ |
|---|---|---|---|
| Height margin $-m_t^h$ | **0.993** [0.97, 1.00] | **0.951** [0.88, 0.99] | **0.843** [0.70, 0.95] |
| Downward velocity $-\dot{z}_t$ | 0.953 [0.91, 1.00] | 0.888 [0.78, 0.97] | 0.715 [0.63, 0.85] |
| Boundary margin $-\min(m_t^h, m_t^\theta)$ | 0.967 [0.93, 0.99] | 0.853 [0.78, 0.92] | 0.723 [0.60, 0.82] |
| Posture score $-S_t^{\text{post}}$ | 0.840 [0.72, 0.96] | 0.703 [0.54, 0.83] | 0.588 [0.48, 0.67] |
| Torso angular speed $\lvert\dot{\theta}_t\rvert$ | 0.500 [0.37, 0.67] | 0.426 [0.30, 0.64] | 0.414 [0.32, 0.56] |
| Angle margin $-m_t^\theta$ | 0.270 [0.08, 0.63] | 0.233 [0.06, 0.51] | 0.238 [0.10, 0.46] |

**Safety utility and threshold.** The safety return $R_{\text{s}} = \sum_t m_t^h$ is the total height margin of an episode, and the safety utility is $u_{\text{s}} = R_{\text{s}} - R_{\text{th}}$. We set the threshold empirically to $R_{\text{th}} = 1000$ for both $\rho$, as higher thresholds led to fewer falls in our experiments. Since $R_{\text{s}} \leq H < R_{\text{th}}$, the argument of $\log$ in (12) is always negative; the piecewise extension of $\log$ is increasing on all of $\mathbb{R}$, so the safety utility still increases with $R_{\text{s}}$.

**Evaluation.** On MO-Walker2d, the CD utility is computed from the per-step average velocity reward and control cost, scaled by the fraction of the horizon the episode survived. The safe-CD utility evaluates Eq. (12) with $R_{\text{th}} = 279.4$, the 75th percentile of the dataset safety returns. We also report the threshold satisfaction, i.e., the fraction of episodes with $R_{\text{s}} \geq 279.4$, and the fall rate, i.e., the fraction of episodes terminated before $H$ (Table 13). All metrics use 100 episodes per seed. Baselines are trained without the safety utility.

Table 8: Results of MO-Ant Expert with concave $G$ and convex $F$. (mean ± standard deviation over 5 random seeds)

| METHOD | $w = [1, 0]$ | $w = [0.5, 0.5]$ | $w = [0, 1]$ | $G$, $F$ |
|---|---|---|---|---|
| OPTIDICE-$[1.0, 0.0]$ | $1089.56 \pm 27.60$ | $729.61 \pm 10.94$ | $369.66 \pm 11.93$ | $12.91 \pm 0.03$ |
| OPTIDICE-$[0.5, 0.5]$ | $584.69 \pm 18.29$ | $689.23 \pm 12.74$ | $793.77 \pm 23.10$ | $13.05 \pm 0.04$ |
| OPTIDICE-$[0.0, 1.0]$ | $88.83 \pm 1.53$ | $1221.97 \pm 26.71$ | $2355.10 \pm 53.65$ | $12.25 \pm 0.03$ |
| FAIRDICE | $617.17 \pm 19.68$ | $682.65 \pm 20.06$ | $748.13 \pm 30.64$ | $13.04 \pm 0.06$ |
| ESR-IQL-$[1.0, 0.0]$ | $\mathbf{1296.49 \pm 30.66}$ | $811.82 \pm 14.71$ | $327.15 \pm 6.20$ | $12.96 \pm 0.03$ |
| ESR-IQL-$[0.5, 0.5]$ | $85.54 \pm 1.61$ | $\mathbf{1243.33 \pm 17.07}$ | $\mathbf{2401.12 \pm 34.37}$ | $12.23 \pm 0.02$ |
| ESR-IQL-$[0.0, 1.0]$ | $87.32 \pm 3.82$ | $1240.07 \pm 19.36$ | $2392.82 \pm 40.01$ | $12.25 \pm 0.04$ |
| AETDICE | $816.78 \pm 61.16$ | $949.73 \pm 22.02$ | $1082.68 \pm 85.08$ | $\mathbf{13.69 \pm 0.05}$ |

Table 9: Results of MO-Ant Amateur with concave $G$ and convex $F$. (mean ± standard deviation over 5 random seeds)

| METHOD | $w = [1, 0]$ | $w = [0.5, 0.5]$ | $w = [0, 1]$ | $G$, $F$ |
|---|---|---|---|---|
| OPTIDICE-$[1.0, 0.0]$ | $869.42 \pm 27.99$ | $636.55 \pm 10.38$ | $403.67 \pm 17.70$ | $12.77 \pm 0.03$ |
| OPTIDICE-$[0.5, 0.5]$ | $540.80 \pm 29.59$ | $623.84 \pm 13.05$ | $706.89 \pm 29.54$ | $12.85 \pm 0.04$ |
| OPTIDICE-$[0.0, 1.0]$ | $138.07 \pm 6.21$ | $875.34 \pm 22.25$ | $1612.61 \pm 50.02$ | $12.31 \pm 0.02$ |
| FAIRDICE | $509.52 \pm 12.81$ | $638.80 \pm 26.01$ | $768.09 \pm 54.61$ | $12.87 \pm 0.07$ |
| ESR-IQL-$[1.0, 0.0]$ | $\mathbf{1570.58 \pm 17.03}$ | $876.82 \pm 9.37$ | $183.06 \pm 3.94$ | $12.57 \pm 0.03$ |
| ESR-IQL-$[0.5, 0.5]$ | $98.57 \pm 14.41$ | $976.31 \pm 13.21$ | $1854.05 \pm 37.46$ | $12.11 \pm 0.12$ |
| ESR-IQL-$[0.0, 1.0]$ | $88.11 \pm 2.58$ | $\mathbf{977.25 \pm 13.05}$ | $\mathbf{1866.40 \pm 27.70}$ | $12.01 \pm 0.02$ |
| AETDICE | $610.17 \pm 46.86$ | $654.87 \pm 7.66$ | $699.56 \pm 31.71$ | $\mathbf{12.96 \pm 0.03}$ |

Table 10: Results of MO-Walker2d Expert with concave $G$ and convex $F$. (mean ± standard deviation over 5 random seeds)

| METHOD | $w = [1, 0]$ | $w = [0.5, 0.5]$ | $w = [0, 1]$ | $G$, $F$ |
|---|---|---|---|---|
| OPTIDICE-$[1.0, 0.0]$ | $\mathbf{718.58 \pm 23.50}$ | $675.69 \pm 40.14$ | $632.80 \pm 93.83$ | $13.02 \pm 0.13$ |
| OPTIDICE-$[0.5, 0.5]$ | $279.96 \pm 41.59$ | $1856.50 \pm 202.64$ | $3433.03 \pm 443.57$ | $13.76 \pm 0.06$ |
| OPTIDICE-$[0.0, 1.0]$ | $10.28 \pm 1.19$ | $7292.20 \pm 402.27$ | $14574.12 \pm 805.61$ | $11.91 \pm 0.07$ |
| FAIRDICE | $307.14 \pm 15.98$ | $1733.19 \pm 97.94$ | $3159.25 \pm 207.63$ | $13.78 \pm 0.04$ |
| ESR-IQL-$[1.0, 0.0]$ | $689.95 \pm 85.09$ | $772.39 \pm 123.46$ | $854.84 \pm 328.36$ | $13.21 \pm 0.25$ |
| ESR-IQL-$[0.5, 0.5]$ | $10.53 \pm 1.45$ | $\mathbf{7366.92 \pm 326.81}$ | $\mathbf{14723.31 \pm 654.99}$ | $11.94 \pm 0.10$ |
| ESR-IQL-$[0.0, 1.0]$ | $12.35 \pm 0.83$ | $6765.08 \pm 427.43$ | $13517.81 \pm 854.75$ | $12.02 \pm 0.10$ |
| AETDICE | $246.26 \pm 70.02$ | $2566.75 \pm 850.14$ | $4887.23 \pm 1637.10$ | $\mathbf{13.87 \pm 0.78}$ |

## H Ablation Study: Impact of Regularization Strength

We study how the regularization coefficient $\beta$, which controls the strength of the $\phi$-divergence penalty in AETDICE, affects performance under different AET objectives. Figure 10 reports evaluation scores under ESR, BSR, and SER criteria as $\beta$ varies on MO-PointMaze-2obj. The effective range of $\beta$ in our experiments is considerably smaller than in single-objective DICE methods; for example, OptiDICE [19] uses $\beta$ between $10^{-4}$ and $10^{-2}$ on D4RL. This is expected: offline MORL datasets

Table 11: Results of MO-Walker2d Amateur with concave $G$ and convex $F$. (mean ± standard deviation over 5 random seeds)

| METHOD | $w = [1, 0]$ | $w = [0.5, 0.5]$ | $w = [0, 1]$ | $G, F$ |
|---|---|---|---|---|
| OPTIDICE-$[1.0, 0.0]$ | $\mathbf{540.87 \pm 11.01}$ | $491.55 \pm 24.05$ | $442.23 \pm 44.79$ | $12.38 \pm 0.11$ |
| OPTIDICE-$[0.5, 0.5]$ | $227.71 \pm 13.15$ | $2780.73 \pm 199.93$ | $5333.75 \pm 402.39$ | $14.01 \pm 0.09$ |
| OPTIDICE-$[0.0, 1.0]$ | $8.37 \pm 0.86$ | $\mathbf{8119.82 \pm 251.87}$ | $\mathbf{16231.27 \pm 504.44}$ | $11.81 \pm 0.08$ |
| FAIRDICE | $261.07 \pm 18.87$ | $1143.11 \pm 77.84$ | $2025.16 \pm 160.95$ | $13.17 \pm 0.09$ |
| ESR-IQL-$[1.0, 0.0]$ | $176.32 \pm 25.68$ | $149.61 \pm 26.89$ | $122.91 \pm 31.93$ | $9.94 \pm 0.41$ |
| ESR-IQL-$[0.5, 0.5]$ | $42.31 \pm 11.21$ | $6710.95 \pm 614.79$ | $13379.58 \pm 1237.66$ | $13.20 \pm 0.22$ |
| ESR-IQL-$[0.0, 1.0]$ | $39.94 \pm 6.93$ | $6607.13 \pm 427.73$ | $13174.33 \pm 861.51$ | $13.16 \pm 0.12$ |
| AETDICE | $277.03 \pm 42.64$ | $3376.26 \pm 532.50$ | $6475.49 \pm 1102.24$ | $\mathbf{14.37 \pm 0.08}$ |

Table 12: Results of Walker2d Cobb-Douglas objectives. CD columns report the Cobb–Douglas utility; safe-CD columns report the safe-CD utility, evaluated with the 75th-percentile safety threshold (negative values indicate a violated safety threshold). Fall rates are reported in Table 13. Baselines are re-evaluated with the safety rewards recorded (Appendix G). Mean ± standard deviation over 3 random seeds, 100 episodes each.

| METHOD | CD UTILITY | | SAFE-CD UTILITY | |
|---|---|---|---|---|
| | $\rho = 0.7$ | $\rho = 0.3$ | $\rho = 0.7$ | $\rho = 0.3$ |
| BC | $1.288 \pm 0.286$ | $0.730 \pm 0.106$ | $-5659 \pm 2841$ | $-5659 \pm 2841$ |
| ESR-IQL | $1.950 \pm 0.043$ | $0.789 \pm 0.097$ | $-2356 \pm 635$ | $-5234 \pm 3025$ |
| AET-CD | $\mathbf{2.048 \pm 0.008}$ | $0.868 \pm 0.049$ | $-1210 \pm 184$ | $-3320 \pm 1484$ |
| AET-CD(SAFE) | $1.099 \pm 0.007$ | $\mathbf{0.948 \pm 0.003}$ | $\mathbf{9.3 \pm 1.2}$ | $\mathbf{9.7 \pm 0.3}$ |

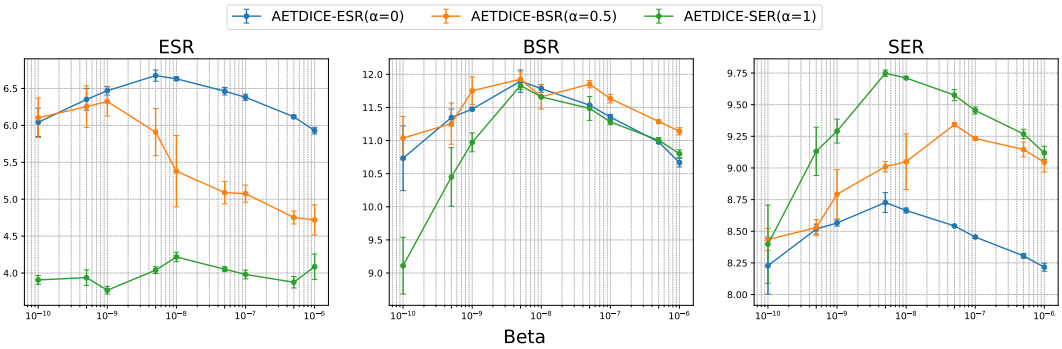



Figure 10: Performance of AETDICE with varying $\beta$ values on MO-PointMaze-2obj dataset.

are collected from policies trained under diverse preference weights, so trajectories optimal for one trade-off can be highly suboptimal for another AET objective. Smaller regularization is needed to permit the larger distribution shift required for finding the target objective's optimum within such heterogeneous data. This makes achieving strong offline performance more demanding, as the optimizer must deviate further from the data distribution. Nevertheless, with sufficiently small $\beta$, AETDICE achieves strong performance across all objectives, and performance varies smoothly as $\beta$ changes, indicating stable optimization. These results demonstrate that AETDICE enables effective optimization of AET objectives given appropriate regularization. Small $\beta$ causes no numerical difficulty with the $\chi^2$-divergence, since $w_t^* = [e_{\boldsymbol{\nu},\boldsymbol{\mu},t}/\beta + 1]_+$ is piecewise linear and involves no exponential (Appendix C.5).

# I Hyperparameter

Tables 16 and 15 summarize the hyperparameters of AETDICE on D4MORL and MO-PointMaze. The policy $\tilde{\pi}_\theta$, the dual function $\nu_\psi$, and the Lagrange multiplier $\boldsymbol{\mu}$ are optimized with Adam at a learning rate of $3 \times 10^{-4}$ for $10^6$ gradient steps. On D4MORL, $\tilde{\pi}_\theta$ and $\nu_\psi$ are four-layer MLPs with

Table 13: Safety on MO-Walker2d: fall rate (fraction of episodes terminated before $H$; lower is better) and threshold satisfaction (percentage of episodes whose safety return reaches the 75th-percentile threshold of the dataset, $R_{\text{th}} = 279.4$; higher is better). Mean $\pm$ standard deviation over 3 random seeds, 100 episodes each.

| | $\rho = 0.7$ | | $\rho = 0.3$ | |
|---|---|---|---|---|
| METHOD | FALL | SATISFACTION | FALL | SATISFACTION |
| BC | $36.0 \pm 19.8\%$ | $26.3 \pm 2.1\%$ | $36.0 \pm 19.8\%$ | $26.3 \pm 2.1\%$ |
| ESR-IQL | $19.0 \pm 2.9\%$ | $1.0 \pm 0.8\%$ | $33.3 \pm 16.0\%$ | $22.3 \pm 8.2\%$ |
| AET-CD | $5.3 \pm 0.5\%$ | $0.7 \pm 0.5\%$ | $21.3 \pm 8.1\%$ | $21.3 \pm 4.9\%$ |
| AET-CD(SAFE) | $\mathbf{1.0 \pm 1.4\%}$ | $\mathbf{99.7 \pm 0.5\%}$ | $\mathbf{1.7 \pm 1.7\%}$ | $\mathbf{99.7 \pm 0.5\%}$ |

Table 14: Results of HalfCheetah Cobb-Douglas objectives (mean $\pm$ standard deviation over 3 random seeds).

| METHOD | CD(0.7) | CD(0.3) |
|---|---|---|
| BC | $1.833 \pm 0.030$ | $1.080 \pm 0.010$ |
| ESR-IQL | $2.344 \pm 0.038$ | $\mathbf{1.150 \pm 0.003}$ |
| AET-CD | $\mathbf{2.416 \pm 0.018}$ | $1.143 \pm 0.001$ |

512 hidden units and the batch size is 512; on MO-PointMaze, they are three-layer MLPs with 768 hidden units (1024 for MO-PointMaze-3obj) and the batch size is 256.

We use the Pearson $\chi^2$ divergence for occupancy regularization. The values of $\beta$, listed in Tables 16 and 15, are smaller than those used in single-objective DICE methods (Appendix H).

### I.1 IQL Baseline Implementation

For the ESR-based IQL baseline, we adopt the same network architecture as AETDICE to ensure a fair comparison. In particular, we use identical state encoders, symlog-transformed sinusoidal embeddings for the accumulated return inputs, and MLPs of the same depth and width as AETDICE.

We tune the expectile parameter $\tau$ over $\{0.8, 0.85, 0.9\}$ and the temperature parameter $\beta_{\text{IQL}}$ over $\{10, 12, 14, 16, 18\}$. IQL models are trained for $3 \times 10^5$ gradient steps using the Adam optimizer with the same learning rate and batch size as AETDICE.

Table 15: Hyperparameters for AETDICE on MO-PointMaze.

| | |
|---|---|
| $\pi_\theta$ learning rate | $3 \times 10^{-4}$ |
| $\nu_\psi$ learning rate | $3 \times 10^{-4}$ |
| $\mu$ learning rate | $3 \times 10^{-4}$ |
| Batch size | 256 |
| Hidden dim of $\nu_\psi$ and $\pi_\theta$ | 768 (1024 for MO-PointMaze-3obj) |
| n_layer of $\nu_\psi$ and $\pi_\theta$ | 3 |
| Training steps | $10^6$ |
| Divergence $\phi$ | Pearson $\chi^2$ |
| $\beta$ | $\{10^{-6}, 5 \times 10^{-7}, 10^{-7}, 5 \times 10^{-8}, 10^{-8}, 5 \times 10^{-9}, 10^{-9}, 5 \times 10^{-10}, 10^{-10}\}$ |

## J Computational Resources

All experiments were run on a single workstation with an Intel® Xeon® Gold 6330 CPU (256 GB RAM) and an NVIDIA RTX 3090 GPU. Training one AETDICE or IQL policy per task took approximately 15-20 minutes on average, with GPU memory consumption staying below 20 GB throughout training.

Table 16: Hyperparameters for AETDICE on D4MORL [9] (MO-Ant-v2, MO-Walker2d-v2, MO-HalfCheetah-v2).

| | |
|---|---|
| $\pi_\theta$ learning rate | $3 \times 10^{-4}$ |
| $\nu_\psi$ learning rate | $3 \times 10^{-4}$ |
| $\boldsymbol{\mu}$ learning rate | $3 \times 10^{-4}$ |
| Batch size | 512 |
| Hidden dim of $\nu_\psi$ and $\pi_\theta$ | 512 |
| n_layer of $\nu_\psi$ and $\pi_\theta$ | 4 |
| Training steps | $10^6$ |
| Divergence $\phi$ | Pearson $\chi^2$ |
| $\beta$ | $\{10^{-4}, 10^{-5}, 10^{-6}\}$ |

# K Limitations and Impact Statement

**Limitations.** While AETDICE enables offline optimization of a broader class of nonlinear MORL objectives, it inherits the fundamental challenges of offline MORL. First, as discussed in the ablation study (Appendix H), offline MORL datasets are collected from policies trained under diverse preference weights, making trajectories optimal for one objective highly suboptimal for another. This requires smaller regularization to permit larger distribution shift, making offline optimization more demanding than in the single-objective setting. Second, the augmented state space adds a coverage requirement. Incorporating $\mathbf{R}_t^{\mathrm{acc}}$ increases the state dimension by the number of objectives that enter $F$ nonlinearly, and the offline data must cover the accumulated rewards of these objectives (Appendix D). This coverage can be augmented without new data by storing each transition at arbitrary values of $\mathbf{R}_t^{\mathrm{acc}}$; more effective ways of enriching coverage are left to future work.

**Impact statement.** AETDICE advances offline MORL by enabling optimization of nonlinear objectives that capture fairness and balanced performance across multiple criteria. This is relevant to real-world equitable decision-making, where policies must balance competing objectives both within and across episodes. As a methodological contribution, we do not anticipate direct negative societal impacts. However, as with any offline RL method, the quality of learned policies depends on the data, and biases present in the dataset may propagate to the optimized policy.