BECON: Belief-Conditioned Constrained Multi-Objective Reinforcement Learning under Drifting Preferences and Budgets

Bui Trong Duc* **Huynh Thi Thanh Binh**
School of Information and Communications Technology
Hanoi University of Science and Technology, Hanoi, Vietnam
{ducbt, binhht}@soict.hust.edu.vn

Abstract

Constrained preference-conditioned MORL typically assumes full task observability, treating preferences and safety budgets as stationary task signals. In real-world deployment, however, these signals are often unreliable proxies: intent may drift, safety tolerances may shift with context, and exposed descriptors may be noisy, delayed, or incomplete. We therefore formulate constrained preference-conditioned MORL as a partially observed task-specification problem, where the agent must navigate a fundamental conflict between unreliable external signals and latent interaction dynamics. We propose **BECON**, a belief-conditioned framework for constrained MORL with drifting preferences and budgets, which infers latent task context from recent histories to estimate preferences and safety budgets. These estimates are fused with the exposed task signal via uncertainty-dependent gates, enabling the policy to defer to observations under high epistemic uncertainty while correcting them when history is informative. The fused context jointly conditions preference optimization and context-adaptive constraint enforcement. Experiments on diverse multi-objective control tasks demonstrate that BECON consistently improves over state-of-the-art MORL baselines, achieving stronger robustness and constraint satisfaction under task drift.

1 Introduction

Real-world decision-making rarely reduces to optimizing a single stationary objective [1, 2]. Autonomous vehicles must balance speed, comfort, energy use, and safety; robotic assistants must trade off task efficiency, user preference, and risk; and resource allocation systems must adapt to evolving utility and feasibility requirements [3–5]. Multi-objective reinforcement learning (MORL) provides a fundamental framework for these settings by modeling returns as objective vectors and resolving trade-offs through preferences [6–8]. To avoid training a separate policy for each trade-off, preference-conditioned MORL learns a single policy that can adapt to different preferences at deployment [9, 10]. In parallel, constrained RL and budgeted Markov decision processes (MDPs) address safety and resource requirements by imposing cost constraints during policy optimization [11].

A key limitation of existing paradigms is the direct observability assumption, which regards the runtime task descriptor as a faithful proxy for the underlying objective [12, 13]. In preference-conditioned MORL, this assumption often manifests by equating the observed preference vector with the intended utility trade-off [14]. In constrained RL, the provided budget is treated as the relevant feasibility threshold, with violations controlled by a global or weakly contextualized Lagrange multiplier [15]. This assumption is fragile in deployed systems, where task descriptors are neither

*Corresponding Author

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

guaranteed to be stationary nor fully reliable. User preferences may drift within an episode as goals and priorities evolve; admissible safety tolerance may depend on recent interaction history, accumulated constraint violations, or context-specific feasibility; and exposed task contexts may be stale, noisy, delayed, or only partially specified [16].

These factors induce a signal-reality gap, in which the runtime descriptor available to the agent may no longer reflect the latent task specification that determines the desired utility trade-off and admissible risk level. An agent that conditions only on this descriptor can therefore optimize stale preferences, miss newly emerging feasibility constraints, or apply safety penalties that are inappropriate for the current context. We therefore consider constrained MORL under task drift as a partially observed task-specification problem, where the hidden context determines both the effective objective trade-off and the admissible cost level. Recent reward and cost patterns provide indirect evidence about this context, but such evidence can be ambiguous, especially over short histories. Thus, the agent must not only infer latent task context from interaction history, but also calibrate how much this inferred context should override or defer to the exposed task signal.

We introduce BECON, a belief-conditioned framework for constrained MORL under drifting task specifications. Rather than assuming that runtime preferences and budgets are directly observable task variables, BECON models them as noisy manifestations of a latent task context. A variational belief over this context is inferred from recent interactions and used to arbitrate, according to uncertainty, between exposed descriptors and history-derived evidence. This design goes beyond standard history conditioning: history is not simply added as extra input, but is used to decide whether the current task signal should be trusted or corrected. The resulting context conditions both multi-objective policy optimization and adaptive constraint enforcement, enabling safety penalties to vary with inferred feasibility rather than remain fixed across deployment conditions.

Our contributions are as follows:

- We formulate constrained MORL with drifting preferences and budgets as a partially observed task-specification problem, where exposed preference and budget signals are unreliable observations of the operative task.
- We develop a belief-conditioned constrained MORL algorithm that infers latent intent and context-dependent feasibility from recent interaction history, and fuses inferred and exposed task signals through uncertainty-dependent arbitration.
- Experiments across diverse constrained MORL benchmarks, supported by ablations and theoretical analysis, show that BECON improves adaptation, preference-aligned performance, and constraint satisfaction under task drift.

2 Related Work

MORL addresses sequential decision making with vector-valued returns, where an agent must balance competing objectives rather than optimize a single scalar reward. Linear scalarization remains a standard formulation because this approach reduces MORL to preference-indexed single-objective control while preserving access to Pareto-optimal solutions under suitable conditions [6]. However, solving a separate policy for each preference is inefficient, motivating preference-conditioned and universal policy methods that take preferences as inputs and generalize across the preference simplex [9]. Recent advances further improve Pareto-front coverage and controllability by discovering constrained Pareto sets [13], sharing latent knowledge across preferences and enabling preference-controllable optimization [17], and modeling the distribution of multi-objective returns beyond their expectations [18]. A complementary line studies continual MORL and latent preference inference, aiming to improve retention, transfer, or recovery of preference-conditioned behavior from incomplete task information [14, 16]. These works highlight the difficulty of relying solely on explicit preferences, but they do not jointly model preference and budget signals as uncertain runtime observations in constrained MORL. In parallel, constrained and safe RL formalize safety, resource, or risk requirements through budgeted MDPs, and primal-dual or Lagrangian optimization, where constraint satisfaction is typically enforced by cost thresholds or adaptive multipliers [11, 19].Recent progress further sharpens constrained RL through reward-free exploration [20], approachability-based constraint satisfaction [21], adaptive resource allocation [22], conservative off-policy learning [23], and constraint-conditioned policy selection under varying safety limits [15, 24]. When safety constraints are coupled with multiple objectives, constrained MORL offers a unified formalism for

balancing utility trade-offs against admissible risk [25]. Yet most existing methods still assume that the preference vector and safety budget are available as reliable task variables, rather than noisy observations of a latent decision context. Orthogonal progress in representation learning, non-stationary RL, and in-context RL suggests that hidden task structure can be extracted from temporal prediction [26], bisimulation metrics [27], causal shift representations [28], local regularization for changing environments [29], or recent trajectories as implicit task context [30]. Although these methods show that history can reveal hidden task structure, they do not directly address constrained preference-conditioned MORL under task specification ambiguity, where exposed preferences and budgets are imperfect observations of the latent utility trade-off and admissible safety level. A more detailed discussion is deferred to Appendix A.

3 Constrained MORL under Task Drift

We formulate a structured non-stationary variant of constrained multi-objective reinforcement learning. Consider a multi-objective Markov decision process $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, \mathbf{r}, c, \gamma)$, where $\mathcal{S}$ and $\mathcal{A}$ are the state and action spaces, $P(\mathbf{s}' \mid \mathbf{s}, \mathbf{a})$ is the transition kernel, $\mathbf{r} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^K$ is a vector reward with K objectives, $c : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is a nonnegative stage cost, i.e., $c(\mathbf{s}, \mathbf{a}) \geq 0$, and $\gamma \in [0, 1)$ is the discount factor [11, 20, 25]. The base environment is fixed, but the task specification may evolve over time and may be only partially observed. This setting isolates task drift from environment drift, so the agent need not relearn the transition dynamics but must infer the current preference and feasibility context from the exposed task signal and recent interaction history.

Let $\Omega := \{\boldsymbol{\omega} \in \mathbb{R}^K \mid \omega_i \geq 0,\ \forall i \in \{1, \ldots, K\},\ \|\boldsymbol{\omega}\|_1 = 1\}$ be the preference space. At time t, the operative task is $\tau_t = (\boldsymbol{\omega}_t, B_t)$, where $\boldsymbol{\omega}_t \in \Omega$ is the current preference vector and $B_t \in \mathcal{B}$ is the current cost budget. The preference determines the scalarized reward, while the budget defines the feasible cost level. Given $\mathbf{r}_t = \mathbf{r}(\mathbf{s}_t, \mathbf{a}_t)$ and $c_t = c(\mathbf{s}_t, \mathbf{a}_t)$, the time-varying constrained objective is

$$\max_{\pi}\ \mathbb{E}_\pi \left[\sum_{t=0}^{T-1} \gamma^t \boldsymbol{\omega}_t^\top \mathbf{r}_t \right] \qquad \text{s.t.} \qquad \mathbb{E}_\pi \left[c_t \mid \mathbf{s}_t, H_t \right] \leq B_t, \quad \forall t. \tag{1}$$

where T and H_t denote the finite decision horizon and the recent interaction history available at time t, respectively. This paper uses per-step constraints because safety tolerance may drift with the current task context, making the acceptable instantaneous cost time-varying. Additionally, non-stationarity enters through the task specification, specifically $\tau_t = (\boldsymbol{\omega}_t, B_t)$ may change over time. Therefore, task drift changes what is desirable or feasible, whereas environment drift would instead change the data-generating process, e.g., $P_t(\mathbf{s}' \mid \mathbf{s}, \mathbf{a}) \neq P_{t'}(\mathbf{s}' \mid \mathbf{s}, \mathbf{a})$, or change the reward and cost functions.

The operative task τ_t is not assumed to be fully observable. Instead, the agent receives an exposed task signal $\widehat{\tau}_t = (\widehat{\boldsymbol{\omega}}_t, \widehat{B}_t)$, where $\widehat{\boldsymbol{\omega}}_t$ and $\widehat{B}_t$ may be noisy, delayed, or incomplete measurements of the operative preference and budget. Therefore, the exposed task signal should be treated as a partial observation of the current decision context rather than as a complete task specification. Since the current state $\mathbf{s}_t$ and exposed task signal $\widehat{\tau}_t$ may not be determined fully, the policy is conditioned on a finite interaction history $H_t = (\xi_{t-L}, \ldots, \xi_{t-1})$, where L is the history length and $\xi_t = \left[\mathbf{s}_t,\ \mathbf{a}_t,\ \mathbf{r}_t,\ \widehat{\boldsymbol{\omega}}_t,\ c_t,\ \widehat{B}_t\right]$ is the transition feature stored at time t. We instantiate task drift through interpolation as follows $\boldsymbol{\omega}_{t+1} = \Pi_\Omega\left((1 - \rho_\omega)\boldsymbol{\omega}_t + \rho_\omega \bar{\boldsymbol{\omega}}_t\right)$ and $B_{t+1} = (1 - \rho_B)B_t + \rho_B \bar{B}_t$, where $\bar{\boldsymbol{\omega}}_t \sim \nu_\omega$, $\bar{B}_t \sim \nu_B$, Π_Ω projects onto the preference simplex, ν_ω and ν_B are sampling distributions for refreshed preferences and budgets, and $\rho_\omega, \rho_B \in [0, 1]$ control the drift rates. Under partial task observability, we consider history-conditioned policies of the form $\pi\left(\mathbf{a}_t \mid \mathbf{s}_t, H_t, \widehat{\boldsymbol{\omega}}_t, \widehat{B}_t\right)$.

4 BECON Framework

BECON implements the partially observed task-specification formulation through a belief-conditioned constrained actor-critic architecture. At each time step, a state encoder maps $\mathbf{s}_t$ to a latent state representation $\mathbf{e}_t$, while a history encoder maps recent transition features H_t to a variational belief $\mathbf{z}_t$ over the unobserved task context. Rather than treating the exposed descriptor $(\widehat{\boldsymbol{\omega}}_t, \widehat{B}_t)$ as a ground-truth command, BECON interprets it as a noisy observation and arbitrates between this signal and the history-derived belief according to epistemic uncertainty. The resulting fused context $(\widetilde{\boldsymbol{\omega}}_t, \widetilde{B}_t)$

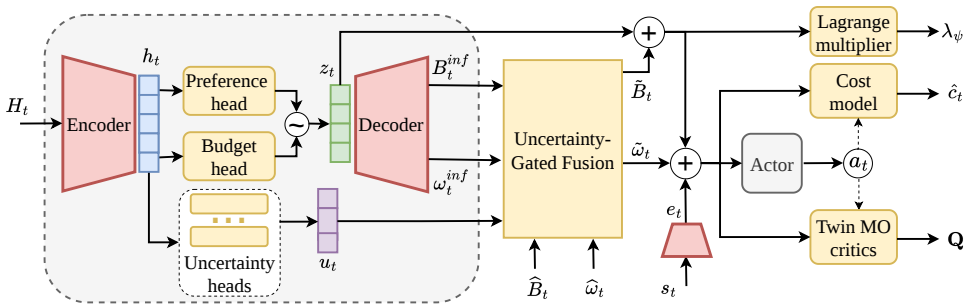

Figure 1: Overview of BECON. A history encoder maps recent transitions H_t to a variational belief $\mathbf{z}_t$ over the latent task context. Uncertainty heads gate the fusion between belief-derived estimates $(\boldsymbol{\omega}_t^{\inf}, B_t^{\inf})$ and the exposed descriptor $\widehat{\tau}_t = (\widehat{\boldsymbol{\omega}}_t, \widehat{B}_t)$. The fused context $(\widetilde{\boldsymbol{\omega}}_t, \widetilde{B}_t)$ conditions the actor, distributional critics, cost model, and context-dependent dual variable for preference-aligned control with adaptive constraint enforcement.

conditions both preference-aligned action selection and adaptive constraint enforcement. Formally, the policy is written as $\pi_\theta\left(\mathbf{a}_t \mid \mathbf{s}_t, H_t, \widehat{\boldsymbol{\omega}}_t, \widehat{B}_t\right) \equiv \pi_\theta\left(\mathbf{a}_t \mid \mathbf{e}_t, \mathbf{z}_t, \widetilde{\boldsymbol{\omega}}_t, \widetilde{B}_t\right)$. Figure 1 illustrates an overview of the proposed framework.

4.1 History-Conditioned Belief and Task Decoding

The current state $\mathbf{s}_t$ and exposed task signal $\widehat{\tau}_t = (\widehat{\boldsymbol{\omega}}_t, \widehat{B}_t)$ may not identify the operative preference and feasibility context. We therefore encode the recent history H_t into a stochastic latent belief $\mathbf{z}_t \in \mathbb{R}^{d_z}$, where d_z denotes the latent belief dimension. Specifically, a recurrent encoder produces $\mathbf{h}_t = \mathrm{GRU}_\phi(H_t)$, followed by two heads $\boldsymbol{\mu}_t = f_\phi^\mu(\mathbf{h}_t)$ and $\log \boldsymbol{\sigma}_t^2 = f_\phi^\sigma(\mathbf{h}_t)$, where $\boldsymbol{\mu}_t, \boldsymbol{\sigma}_t \in \mathbb{R}^{d_z}$. The approximate posterior is $q_\phi(\mathbf{z}_t \mid H_t) = \mathcal{N}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$ with $\boldsymbol{\Sigma}_t = \mathrm{Diag}(\boldsymbol{\sigma}_t^2)$. A belief sample is drawn by reparameterization: $\mathbf{z}_t = \boldsymbol{\mu}_t + \boldsymbol{\sigma}_t \odot \boldsymbol{\epsilon}$, where $\odot$ denotes element-wise multiplication, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{d_z})$, and $\mathbf{I}_{d_z}$ is the $d_z \times d_z$ identity matrix.

The latent belief $\mathbf{z}_t$ is intended to summarize hidden task context, including latent intent and latent feasibility, rather than merely provide an auxiliary representation. To quantify uncertainty in this inference, we use the posterior standard deviation as an aggregate uncertainty measure. Specifically, the scalar belief uncertainty is computed from the posterior standard deviation $\boldsymbol{\sigma}_t$ of $q_\phi(\mathbf{z}_t \mid H_t)$. In addition, objective-specific log-variance heads applied to the hidden belief representation produce per-objective uncertainty estimates $\mathbf{u}_t = (u_{t,1}, \ldots, u_{t,K})$. The scalar aggregate uncertainty and the per-objective uncertainty estimates control how strongly BECON trusts inferred task signals.

BECON decodes the belief $\mathbf{z}_t$ into two inferred task estimates: an inferred preference $\boldsymbol{\omega}_t^{\inf} \in \Omega$ and an inferred budget $B_t^{\inf} \in \mathcal{B}$. Let $\mathbf{q}_t^\omega = g_\omega(\mathbf{z}_t) \in \mathbb{R}^K$, and $q_t^B = g_B(\mathbf{z}_t) \in \mathbb{R}$ be the initial decoder outputs. They are mapped to valid task variables by $\omega_{t,i}^{\inf} = (\varphi(q_{t,i}^\omega) + \varepsilon) / \sum_{j=1}^K (\varphi(q_{t,j}^\omega) + \varepsilon)$ and $B_t^{\inf} = \varphi(q_t^B)$, where $\varphi(x) = \log(1 + \exp(x))$ is a positive activation and $\varepsilon > 0$ is a small constant. Thus, $\boldsymbol{\omega}_t^{\inf}$ lies on the preference simplex and $B_t^{\inf}$ is nonnegative. The decoded preference estimates the current objective trade-off, while the decoded budget estimates the current feasible cost level.

4.2 Uncertainty-Gated Task Fusion

BECON uses belief uncertainty to control how much the policy should trust inferred task signals, specifically, between the exposed task signal $\widehat{\tau}_t = (\widehat{\boldsymbol{\omega}}_t, \widehat{B}_t)$ and the belief-derived estimate $(\boldsymbol{\omega}_t^{\inf}, B_t^{\inf})$. For nonempty histories, each objective gate is defined as $\alpha_{t,i} = \alpha_{\max} \exp(-u_{t,i}), \alpha_{\max} \in [0, 1]$ and $\alpha_{t,i} = 0$ when H_t is empty. Let $\boldsymbol{\alpha}_t = (\alpha_{t,1}, \ldots, \alpha_{t,K})$. The fused preference and budget are

$$\widetilde{\boldsymbol{\omega}}_t = \Pi_\Omega\left((1 - \boldsymbol{\alpha}_t) \odot \widehat{\boldsymbol{\omega}}_t + \boldsymbol{\alpha}_t \odot \boldsymbol{\omega}_t^{\inf}\right), \quad \widetilde{B}_t = (1 - \bar{\alpha}_t)\widehat{B}_t + \bar{\alpha}_t B_t^{\inf} \tag{2}$$

where Π_Ω projects onto the preference simplex and $\bar{\alpha}_t = K^{-1}\sum_{i=1}^{K}\alpha_{t,i}$ is the average gate. Thus, uncertain beliefs defer to the exposed task signal, while confident beliefs can correct it using history-derived context.

4.3 Belief-Conditioned Constrained Actor–Critic

The current state is first mapped to a latent embedding $\mathbf{e}_t = f_{\text{enc}}(\mathbf{s}_t)$, which is shared across the actor, critics, and cost model. The actor conditions on $\mathbf{e}_t$, the belief $\mathbf{z}_t$, the fused preference $\widetilde{\boldsymbol{\omega}}_t$, and the fused budget $\widetilde{B}_t$: $\pi_\theta(\mathbf{a}_t \mid \mathbf{e}_t, \mathbf{z}_t, \widetilde{\boldsymbol{\omega}}_t, \widetilde{B}_t)$. For continuous control, it outputs Gaussian parameters $(\boldsymbol{\mu}_t^\pi, \log\boldsymbol{\sigma}_t^\pi) = f_\pi\left(\mathbf{e}_t, \mathbf{z}_t, \widetilde{\boldsymbol{\omega}}_t, \widetilde{B}_t\right)$ and samples $\tilde{\mathbf{a}}_t = \boldsymbol{\mu}_t^\pi + \boldsymbol{\sigma}_t^\pi \odot \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. The executed action is squashed as $\mathbf{a}_t = \tanh(\tilde{\mathbf{a}}_t)$.

We use twin multi-objective critics $\mathbf{Q}_j(\mathbf{e}_t, \mathbf{a}_t, \mathbf{z}_t, \widetilde{\boldsymbol{\omega}}_t, \widetilde{B}_t) \in \mathbb{R}^K$ for $j \in \{1, 2\}$. The i-th component $Q_{j,i}$ estimates the value of objective i. Under the fused preference, the scalarized critic is

$$Q_j^{\text{scal}}(\mathbf{s}_t, \mathbf{a}_t; \widetilde{\boldsymbol{\omega}}_t) = \sum_{i=1}^{K} \widetilde{\omega}_{t,i} Q_{j,i}(\mathbf{e}_t, \mathbf{a}_t, \mathbf{z}_t, \widetilde{\boldsymbol{\omega}}_t, \widetilde{B}_t). \tag{3}$$

In our implementation, each objective value is represented distributionally over atoms $\mathcal{Z} = \{\zeta_1, \ldots, \zeta_N\}$ with $\zeta_n \in [V_{\min}, V_{\max}]$. If $p_{j,i,n}$ is the predicted probability of atom ζ_n, then $Q_{j,i} = \sum_{n=1}^{N} p_{j,i,n}\zeta_n$.

The distributional cost model outputs p_ν^c, and the predicted one-step cost is its expectation $\widehat{c}_t(\mathbf{a}_t) = \mathbb{E}_{p_\nu^c}[c]$. Instead of using a single global Lagrange multiplier, which cannot adapt to time-varying feasibility requirements, BECON uses a context-dependent dual variable

$$\lambda_\psi(\widetilde{B}_t, \mathbf{z}_t) = \varphi\Big(f_\lambda(\widetilde{B}_t, \mathbf{z}_t)\Big), \tag{4}$$

where $\varphi(x) = \log(1 + \exp(x))$ ensures nonnegativity. This allows the safety penalty to vary with the inferred feasibility context. Finally, the state embedding is regularized with a latent transition model $\widehat{\mathbf{e}}_{t+1} = f_{\text{dyn}}(\mathbf{e}_t, \mathbf{a}_t)$, encouraging $\mathbf{e}_t$ to preserve predictive information about future states.

4.4 Learning Objectives

All components are trained from mini-batches sampled from the replay buffer $\mathcal{D}$. For each transition, we encode $\mathbf{e}_{t+1} = f_{\text{enc}}(\mathbf{s}_{t+1})$ and sample the next belief $\mathbf{z}_{t+1} \sim q_\phi(\mathbf{z}_{t+1} \mid H_{t+1})$. The next action is drawn under the stored fused task context $\mathbf{a}'_{t+1}$. Then, BECON selects the conservative twin-critic target, and applies the standard distributional Bellman projection onto $\mathcal{Z}$ to obtain p_i^{tar} for each objective i. The critic loss is $\mathcal{L}_{\text{critic}} = \sum_{j=1}^{2}\mathbb{E}_{\mathcal{D}}\left[\sum_{i=1}^{K}\widetilde{\omega}_{t,i}\,\text{CE}\left(p_i^{\text{tar}}, p_{j,i}\right)\right]$, where $p_{j,i}$ is the predicted categorical distribution of critic j for objective i. Let $(\widehat{\mathbf{r}}_t^b, \widehat{c}_t^b, \widehat{\boldsymbol{\omega}}_t^{\text{rec}}, \widehat{B}_t^{\text{rec}}) = d_\eta(\mathbf{e}_t, \mathbf{a}_t, \mathbf{z}_t)$ denote auxiliary prediction heads from the belief representation, where $\widehat{\mathbf{r}}_t^b$ and $\widehat{c}_t^b$ predict the observed reward vector and cost, and $(\widehat{\boldsymbol{\omega}}_t^{\text{rec}}, \widehat{B}_t^{\text{rec}})$ reconstruct the exposed task descriptor. The belief objective is $\mathcal{L}_{\text{belief}} = \lambda_{\text{pred}}\mathbb{E}_{\mathcal{D}}\left[\left\|\widehat{\mathbf{r}}_t^b - \mathbf{r}_t\right\|_2^2 + \left(\widehat{c}_t^b - c_t\right)^2\right] + \lambda_{\text{obs}}\mathbb{E}_{\mathcal{D}}\left[\left\|\widehat{\boldsymbol{\omega}}_t^{\text{rec}} - \widehat{\boldsymbol{\omega}}_t\right\|_2^2 + \left(\widehat{B}_t^{\text{rec}} - \widehat{B}_t\right)^2\right] + \beta_{\text{KL}}\mathbb{E}_{\mathcal{D}}\left[\text{KL}\left(q_\phi(\mathbf{z}_t \mid H_t) \,\|\, \mathcal{N}(\mathbf{0}, \mathbf{I}_{d_z})\right)\right]$. The dynamics and cost losses regularize the state embedding and train the distributional cost model

$$\mathcal{L}_{\text{dyn}} = \mathbb{E}_{\mathcal{D}}\left[\left\|f_{\text{dyn}}(\mathbf{e}_t, \mathbf{a}_t) - \mathbf{e}_{t+1}^-\right\|_2^2\right], \quad \mathcal{L}_{\text{cost}} = \mathbb{E}_{\mathcal{D}}\left[\text{CE}\left(p_t^{c,\text{tar}}, p_t^c\right)\right], \tag{5}$$

where $\mathbf{e}_{t+1}^-$ is a stop-gradient target and $p_t^{c,\text{tar}}$ is obtained by projecting the observed cost c_t. The actor minimizes an entropy-regularized constrained objective under the fused task context

$$\mathcal{L}_{\text{actor}} = \mathbb{E}_{\mathcal{D}}\left[\alpha_H \log\pi_\theta(\mathbf{a}_t \mid \cdot) - \min_{j\in\{1,2\}} Q_j^{\text{scal}}\left(\mathbf{s}_t, \mathbf{a}_t; \widetilde{\boldsymbol{\omega}}_t\right) + \kappa_t\lambda_\psi(\widetilde{B}_t, \mathbf{z}_t)\left[\widehat{c}_t(\mathbf{a}_t) - \widetilde{B}_t\right]_+\right], \tag{6}$$

where $\widehat{c}_t(\mathbf{a}_t) = \mathbb{E}_{p_t^c}[c]$ is the expected cost from the distributional cost model, $\kappa_t \in [0, 1]$ is the constraint warm-up coefficient. The context-dependent dual variable is updated by minimizing

$\mathcal{L}_\lambda = -\mathbb{E}_\mathcal{D}\left[\lambda_\psi(\widetilde{B}_t, \mathbf{z}_t)\left(\widehat{c}_t(\mathbf{a}_t) - \widetilde{B}_t\right)\right]$. Since λ_ψ is nonnegative, this update increases the penalty when the predicted cost exceeds the fused budget and relaxes it otherwise. The auxiliary and value-learning objectives are optimized jointly as $\mathcal{L}_{\text{main}} = \mathcal{L}_{\text{critic}} + \lambda_{\text{belief}}\mathcal{L}_{\text{belief}} + \lambda_{\text{dyn}}\mathcal{L}_{\text{dyn}} + \lambda_{\text{cost}}\mathcal{L}_{\text{cost}}$. Each iteration updates $\mathcal{L}_{\text{main}}$, $\mathcal{L}_{\text{actor}}$, and $\mathcal{L}_\lambda$, followed by soft target-network updates.

5 Theoretical Analysis

Let $x_t = (\mathbf{s}_t, H_t, \widehat{\tau}_t)$ be the observable decision context, where $\widehat{\tau}_t = (\widehat{\boldsymbol{\omega}}_t, \widehat{B}_t)$. BECON produces the fused task estimate $\widetilde{\tau}_t = (\widetilde{\boldsymbol{\omega}}_t, \widetilde{B}_t)$, where $\widetilde{B}_t$ represents the fused instantaneous admissible cost level. For a policy conditioned on this fused context, define $\bar{c}_{\pi,t} := \mathbb{E}_{a\sim\pi(\cdot|x_t,\widetilde{\tau}_t)}[c(\mathbf{s}_t, a)]$ and $\widehat{c}_{\pi,t} := \mathbb{E}_{a\sim\pi(\cdot|x_t,\widetilde{\tau}_t)}\left[\mathbb{E}_{p^c_\nu(\cdot|x_t,a)}[c]\right]$ as the true and model-predicted one-step costs, where p^c_ν is the distributional cost model parameterized by ν. The resulting perturbation terms are as follows

$$\delta^\omega_t = \left\|\widetilde{\boldsymbol{\omega}}_t - \boldsymbol{\omega}_t\right\|_2, \qquad \delta^B_t = \left|\widetilde{B}_t - B_t\right|, \qquad \epsilon^c_t(\pi) = \left|\widehat{c}_{\pi,t} - \bar{c}_{\pi,t}\right|. \tag{7}$$

Assumption 1. *For all state-action pairs $(\mathbf{s}, \mathbf{a})$, $\|\mathbf{r}(\mathbf{s}, \mathbf{a})\|_2 \le R_2$ and $0 \le c(\mathbf{s}, \mathbf{a}) \le C_{\max}$. The context-dependent dual penalty is parameterized or clipped so that it is uniformly bounded as $0 \le \lambda_\psi(\widetilde{B}_t, \mathbf{z}_t) \le \Lambda$.*

Proposition 1. *Assume that Π_Ω is the Euclidean projection onto the preference simplex and that $0 \le \alpha_{t,i} \le 1$ for all objectives. The inferred budget is parameterized as $B^{\text{inf}}_t = \varphi(g_B(\mathbf{z}_t))$ with $\varphi(x) = \log(1 + \exp(x))$, and the exposed budget $\widehat{B}_t$ is clipped to be nonnegative. Define $\alpha^-_t = \min_i\{\alpha_{t,i}\}$ and $\alpha^+_t = \max_i\{\alpha_{t,i}\}$. Let $\alpha^B_t \in [0, 1]$ denote the budget gate; in our implementation, $\alpha^B_t = \bar{\alpha}_t = K^{-1}\sum_{i=1}^K \alpha_{t,i}$. Then $\widetilde{\boldsymbol{\omega}}_t \in \Omega$ and $\widetilde{B}_t \ge 0$. Moreover,*

$$\delta^\omega_t \le (1-\alpha^-_t)\left\|\widehat{\boldsymbol{\omega}}_t - \boldsymbol{\omega}_t\right\|_2 + \alpha^+_t\left\|\boldsymbol{\omega}^{\text{inf}}_t - \boldsymbol{\omega}_t\right\|_2, \delta^B_t \le (1-\alpha^B_t)\left|\widehat{B}_t - B_t\right| + \alpha^B_t\left|B^{\text{inf}}_t - B_t\right|. \tag{8}$$

Proposition 1 captures the intended arbitration behavior of BECON: uncertain beliefs are down-weighted in favor of the exposed descriptor, while confident and accurate beliefs can override stale or noisy task signals. We next connect this local fusion property to control performance, demonstrating that both preference loss and instantaneous safety violation are controlled by three errors explicitly targeted by the method: preference-fusion error, budget-fusion error, and cost-model error. Define the true instantaneous-penalized performance and its fused model-based counterpart as follows

$$\mathcal{G}_{\tau_{0:T-1}}(\pi) := \mathbb{E}_{\pi,\widetilde{\tau}}\left[\sum_{t=0}^{T-1}\gamma^t\left\{\boldsymbol{\omega}_t^\top\mathbf{r}_t - \lambda_\psi(\widetilde{B}_t, \mathbf{z}_t)\left[\bar{c}_{\pi,t} - B_t\right]_+\right\}\right], \tag{9}$$

$$\widehat{\mathcal{G}}_{\widetilde{\tau}_{0:T-1}}(\pi) := \mathbb{E}_{\pi,\widetilde{\tau}}\left[\sum_{t=0}^{T-1}\gamma^t\left\{\widetilde{\boldsymbol{\omega}}_t^\top\mathbf{r}_t - \lambda_\psi(\widetilde{B}_t, \mathbf{z}_t)\left[\widehat{c}_{\pi,t} - \widetilde{B}_t\right]_+\right\}\right]. \tag{10}$$

Here, $[x]_+ = \max(x, 0)$ denotes the positive-part operator. Thus, $[\widehat{c}_{\pi,t} - \widetilde{B}_t]_+$ measures the predicted instantaneous constraint violation: it is zero when the predicted one-step cost is below the fused admissible cost level, and increases only when the predicted cost exceeds $\widetilde{B}_t$.

Theorem 1. *Under Assumption 1, for any history-conditioned policy π, the gap between the true instantaneous-penalized performance and the fused model-based performance is bounded by*

$$\left|\mathcal{G}_{\tau_{0:T-1}}(\pi) - \widehat{\mathcal{G}}_{\widetilde{\tau}_{0:T-1}}(\pi)\right| \le \sum_{t=0}^{T-1}\gamma^t\mathbb{E}_{\pi,\widetilde{\tau}}\left[R_2\delta^\omega_t + \Lambda\left(\epsilon^c_t(\pi) + \delta^B_t\right)\right]. \tag{11}$$

Furthermore, let Π be a comparator policy class and suppose that $\widetilde{\pi} \in \Pi$ is ϵ_{opt}-optimal for the fused model-based objective, namely $\widehat{\mathcal{G}}_{\widetilde{\tau}_{0:T-1}}(\widetilde{\pi}) \ge \sup_{\pi\in\Pi}\widehat{\mathcal{G}}_{\widetilde{\tau}_{0:T-1}}(\pi) - \epsilon_{\text{opt}}$. If the perturbation terms are uniformly bounded in expectation over Π, i.e., $\forall\,\pi \in \Pi$, $\forall t$, $\mathbb{E}_{\pi,\widetilde{\tau}}\left[\delta^\omega_t\right] \le \bar{\delta}^\omega_t, \mathbb{E}_{\pi,\widetilde{\tau}}\left[\delta^B_t\right] \le \bar{\delta}^B_t, \mathbb{E}_{\pi,\widetilde{\tau}}\left[\epsilon^c_t(\pi)\right] \le \bar{\epsilon}^c_t$, then the true suboptimality of $\widetilde{\pi}$ satisfies

$$\sup_{\pi\in\Pi}\left\{\mathcal{G}_{\tau_{0:T-1}}(\pi)\right\} - \mathcal{G}_{\tau_{0:T-1}}(\widetilde{\pi}) \le \epsilon_{\text{opt}} + 2\sum_{t=0}^{T-1}\gamma^t\left[R_2\bar{\delta}^\omega_t + \Lambda\left(\bar{\epsilon}^c_t + \bar{\delta}^B_t\right)\right]. \tag{12}$$

Theorem 1 indicates that partial task observability affects performance only through fused-context errors. Preference mismatch drives reward loss, while cost-model and budget-fusion errors govern instantaneous safety degradation. This matches BECON, which uses self-supervised belief inference, uncertainty-gated fusion, and context-dependent dual penalties to reduce the observable proxies of these errors under partial task observability.

Assumption 2. *For a fixed latent task $\tau = (\boldsymbol{\omega}, B)$, the observable history distribution induced by H_t is informative enough to identify the task context up to statistical error. Specifically, there exist belief decoders d_ω and d_B such that, when trained from observable prediction losses, their fixed-task estimation errors satisfy $\mathbb{E}\left[\|d_\omega(\mathbf{z}_t) - \boldsymbol{\omega}_t\|_2\right] \le A_\omega L^{-1/2}$ and $\mathbb{E}\left[|d_B(\mathbf{z}_t) - B_t|\right] \le A_B L^{-1/2}$.*

Theorem 2. *Assume that the drift targets satisfy $\bar{\boldsymbol{\omega}}_t \in \Omega$ and $B_t, \bar{B}_t \in [B_{\min}, B_{\max}]$, and let $D_B = B_{\max} - B_{\min}$. Then, under task drift,*

$$\mathbb{E}\left[\left\|\boldsymbol{\omega}_t^{\inf} - \boldsymbol{\omega}_t\right\|_2\right] \le A_\omega L^{-1/2} + \frac{L+1}{2}\sqrt{2}\rho_\omega, \mathbb{E}\left[\left|B_t^{\inf} - B_t\right|\right] \le A_B L^{-1/2} + \frac{L+1}{2} D_B \rho_B. \quad (13)$$

Consequently, up to constants, the preference-optimal history length satisfies $L_\omega^\star \asymp (A_\omega/\rho_\omega)^{2/3}$ and $\min_L \mathbb{E}\left[\left\|\boldsymbol{\omega}_t^{\inf} - \boldsymbol{\omega}_t\right\|_2\right] = \mathcal{O}\left(A_\omega^{2/3}\rho_\omega^{1/3}\right)$, with the analogous budget scaling $L_B^\star \asymp (A_B/(D_B\rho_B))^{2/3}$.

Theorem 2 highlights a bias-variance trade-off in history-based inference. Longer histories reduce statistical uncertainty but increase staleness bias under task drift. This motivates uncertainty-gated fusion, which downweights inferred task estimates when the history-derived belief is unreliable.

6 Experiments

6.1 Experimental Setting

Benchmarks. We evaluate BECON on a diverse set of continuous and discrete MORL benchmarks. The continuous dataset includes MuJoCo robotic-control tasks, including Ant, HalfCheetah, and Humanoid, and the Building energy-control task, which involves up to nine objectives [10, 13]. The discrete suite includes Minecart and MO-LunarLander [13]. Detailed benchmark descriptions are provided in Appendix D.

Baselines. BECON is compared against several state-of-the-art MORL baselines, including GPI-LS [31], C-MORL [13], Q-Pensieve [32], and COLA [10]. For each baseline, we use the official implementation when available, tune its hyperparameters, and report the best validated setting. All methods are trained for 1 million environment steps, with performance reported as mean $\pm$ standard deviation over 5 independent runs. It is noted that GPI-LS supports both discrete and continuous control, while the other baselines are designed mainly for continuous-control MORL. Additional implementation details are given in Appendix E.

Metrics. We evaluate all methods in both static and drift settings. In the static setting, preferences and budgets are fixed, and performance is measured by feasible-front quality. Let $\mathcal{P}$ denote the feasible Pareto front, containing only return vectors that satisfy the cost threshold.

- **Feasible Hypervolume (FHV)** [10] measures the dominated volume covered by feasible Pareto solutions with respect to a reference point $\mathbf{r}^{\text{ref}} \in \mathbb{R}^K$: $\text{FHV} = \text{HV}(\mathcal{P}; \mathbf{r}^{\text{ref}}) = \lambda_K\left(\bigcup_{\mathbf{R}\in\mathcal{P}} \prod_{k=1}^K [r_k^{\text{ref}}, R_k]\right)$, where $\lambda_K(\cdot)$ denotes the K-dimensional Lebesgue measure.
- **Feasible Utility (FUT)** [10] evaluates the best feasible scalarized return under evaluation preferences $\Omega_{\text{eval}} = \{\boldsymbol{\omega}_j\}_{j=1}^M \subset \Omega$: $\text{FUT} = \frac{1}{M}\sum_{j=1}^M \max_{\mathbf{R}\in\mathcal{P}} \boldsymbol{\omega}_j^\top \mathbf{R}$.

In the drift setting, preferences and budgets vary during evaluation, so we evaluate online performance against the current task variables. For episode $e \in \{1, \dots, E\}$ and timestep $t \in \{0, \dots, T_e - 1\}$, let $\boldsymbol{\omega}_{e,t} \in \Omega$ denote the current hidden preference, $B_{e,t}$ the current budget, $\mathbf{r}_{e,t}$ the reward vector, and $c_{e,t}$ the realized cost.

Table 1: Reference-threshold feasible-front evaluation of BECON against MORL baselines

Dataset	Metric	GPI-LS	C-MORL	Q-Pensieve	COLA	BECON (ours)
MO-Ant-2d	FHV ($\times 10^6$)	2.61 ± 0.23	0.25 ± 0.01	5.56 ± 0.34	6.37 ± 0.89	$\mathbf{7.71 \pm 0.10}$
	FUT ($\times 10^3$)	1.99 ± 0.05	0.50 ± 0.01	2.21 ± 0.07	2.35 ± 0.14	$\mathbf{2.57 \pm 0.02}$
	MC ($\times 10^{-3}$)	1.91 ± 0.10	$\mathbf{0.07 \pm 0.05}$	2.66 ± 0.17	2.60 ± 0.81	1.39 ± 0.06
MO-HalfCheetah-2d	FHV ($\times 10^6$)	3.16 ± 0.05	1.25 ± 0.00	5.61 ± 0.04	5.40 ± 0.12	$\mathbf{5.74 \pm 0.06}$
	FUT ($\times 10^3$)	2.25 ± 0.02	1.50 ± 0.00	2.28 ± 0.04	2.24 ± 0.04	$\mathbf{2.35 \pm 0.00}$
	MC ($\times 10^{-3}$)	0.59 ± 0.04	$\mathbf{0.01 \pm 0.00}$	1.49 ± 0.26	1.99 ± 0.18	0.78 ± 0.01
MO-Hopper-2d	FHV ($\times 10^7$)	0.00 ± 0.00	0.13 ± 0.11	0.24 ± 0.01	0.00 ± 0.00	$\mathbf{1.20 \pm 0.16}$
	FUT ($\times 10^3$)	0.00 ± 0.00	0.97 ± 0.74	1.48 ± 0.02	0.00 ± 0.00	$\mathbf{3.37 \pm 0.21}$
	MC ($\times 10^{-2}$)	1.85 ± 0.12	$\mathbf{0.19 \pm 0.16}$	0.46 ± 0.04	1.57 ± 0.10	0.72 ± 0.08
MO-Humanoid	FHV ($\times 10^7$)	0.00 ± 0.00	0.01 ± 0.00	1.06 ± 0.51	0.71 ± 0.27	$\mathbf{2.11 \pm 0.09}$
	FUT ($\times 10^3$)	0.00 ± 0.00	0.30 ± 0.05	3.14 ± 0.79	2.74 ± 0.48	$\mathbf{4.69 \pm 0.07}$
	MC ($\times 10^{-3}$)	1.60 ± 0.09	$\mathbf{0.04 \pm 0.00}$	0.59 ± 0.02	0.77 ± 0.04	0.07 ± 0.02
MO-Ant-3d	FHV ($\times 10^{10}$)	0.31 ± 0.03	0.00 ± 0.00	0.26 ± 0.06	0.91 ± 0.16	$\mathbf{1.01 \pm 0.03}$
	FUT ($\times 10^3$)	1.55 ± 0.01	0.69 ± 0.01	1.29 ± 0.08	2.00 ± 0.08	$\mathbf{2.05 \pm 0.01}$
	MC ($\times 10^{-3}$)	3.22 ± 0.27	$\mathbf{0.06 \pm 0.01}$	3.62 ± 0.10	3.47 ± 1.08	1.44 ± 0.08
MO-Ant-4d	FHV ($\times 10^{14}$)	3.52 ± 0.51	0.17 ± 0.00	2.35 ± 1.01	3.25 ± 0.78	$\mathbf{4.78 \pm 0.15}$
	FUT ($\times 10^3$)	4.32 ± 0.33	2.03 ± 0.01	3.79 ± 0.45	4.21 ± 0.27	$\mathbf{4.65 \pm 0.04}$
	MC ($\times 10^{-3}$)	1.56 ± 0.23	$\mathbf{0.08 \pm 0.05}$	2.73 ± 0.21	3.22 ± 1.04	1.64 ± 0.14
MO-Ant-5d	FHV ($\times 10^{16}$)	0.21 ± 0.05	0.11 ± 0.02	2.53 ± 0.54	2.08 ± 0.51	$\mathbf{3.45 \pm 0.04}$
	FUT ($\times 10^3$)	1.87 ± 0.09	1.61 ± 0.00	1.91 ± 0.09	1.86 ± 0.10	$\mathbf{2.04 \pm 0.01}$
	MC ($\times 10^{-3}$)	1.09 ± 0.13	$\mathbf{0.07 \pm 0.02}$	1.19 ± 0.14	1.11 ± 0.45	0.42 ± 0.02
MO-HalfCheetah-5d	FHV ($\times 10^{17}$)	0.22 ± 0.00	0.01 ± 0.02	0.30 ± 0.19	0.33 ± 0.05	$\mathbf{1.22 \pm 0.07}$
	FUT ($\times 10^3$)	2.29 ± 0.05	1.57 ± 0.04	2.12 ± 0.26	2.59 ± 0.08	$\mathbf{2.93 \pm 0.07}$
	MC ($\times 10^{-3}$)	2.93 ± 0.25	2.46 ± 0.68	2.96 ± 0.56	3.81 ± 0.94	$\mathbf{1.66 \pm 0.03}$
Building-9d	FHV ($\times 10^{31}$)	4.52 ± 0.11	5.02 ± 0.57	6.32 ± 0.57	5.84 ± 0.27	$\mathbf{6.98 \pm 0.15}$
	FUT ($\times 10^3$)	3.33 ± 0.02	3.36 ± 0.03	3.42 ± 0.03	3.39 ± 0.02	$\mathbf{3.45 \pm 0.01}$
	MC ($\times 10^{-3}$)	2.96 ± 0.21	$\mathbf{0.09 \pm 0.12}$	6.29 ± 0.30	5.44 ± 0.17	0.54 ± 0.28

- **Scalarized Return (SR)** [33] measures average online scalarized performance: $\mathrm{SR} = \frac{1}{E}\sum_{e=1}^{E}\frac{1}{T_e}\sum_{t=0}^{T_e-1}\boldsymbol{\omega}_{e,t}^{\top}\mathbf{r}_{e,t}$.
- **Post-Change Utility (PCU)** [34] measures utility during the adaptation window after preference or budget changes: $\mathrm{PCU} = \frac{1}{E}\sum_{e=1}^{E}\frac{1}{\max(|\mathcal{A}_e|,1)}\sum_{t\in\mathcal{A}_e}\boldsymbol{\omega}_{e,t}^{\top}\mathbf{r}_{e,t}$, where $\mathcal{A}_e$ denotes the post-change adaptation window.

Additionally, we utilize **Mean Cost (MC)** and **Violation Rate (VR)** to monitor constraint satisfaction.

6.2 Experimental Results

Table 1 demonstrates reference-threshold feasible-front performance under static preferences and budgets. It is noted that if no solution satisfies the cost threshold, the feasible front is empty and both FHV and FUT are set to zero. BECON achieves the best FHV and FUT on all benchmark tasks, indicating that the proposed belief-conditioned constraint handling improves both feasible-front coverage and preference-aligned utility. Compared with the second-best method, BECON improves FHV by 98.4% on average and FUT by 24.5% on average across all tasks. The gains are especially pronounced on challenging tasks such as Building-9d, MO-Humanoid, MO-Ant-5d, and MO-HalfCheetah-5d, where accurately identifying feasible high-return solutions becomes substantially harder. Although C-MORL often obtains the lowest mean cost, this behavior reflects conservative constraint satisfaction rather than superior feasible-front optimization. In constrained MORL, MC is used to monitor whether the learned policy respects the cost threshold, while the primary goal is to maximize feasible return quality. BECON maintains low realized costs while consistently achieving substantially higher FHV and FUT, highlighting a better balance between constraint satisfaction and multi-objective performance. Additional constraint-tightness results in Appendix F.1 illustrate that BECON preserves a nonzero FHV under stringent admissible cost levels where COLA often becomes infeasible. Table 2 further confirms this trend on discrete benchmarks.

Table 2: Feasible-front performance on discrete MORL benchmarks

Dataset	Method	FHV ↑	FUT ↑	MC
MO-LunarLander-4d	BECON	$\mathbf{3.69\times10^{10}}$	$\mathbf{71.58}$	0.00102
	COLA	3.44×10^{10}	68.98	0.00379
	Q-Pensieve	3.36×10^{10}	67.48	0.00363
	GPI-LS	1.96×10^{10}	66.72	0.00324
Minecart	BECON	$\mathbf{189.70}$	$\mathbf{-6.07}$	0.00001
	COLA	180.00	-6.67	0.00695
	Q-Pensieve	179.90	-6.70	0.00404
	GPI-LS	152.90	-8.70	0.00313

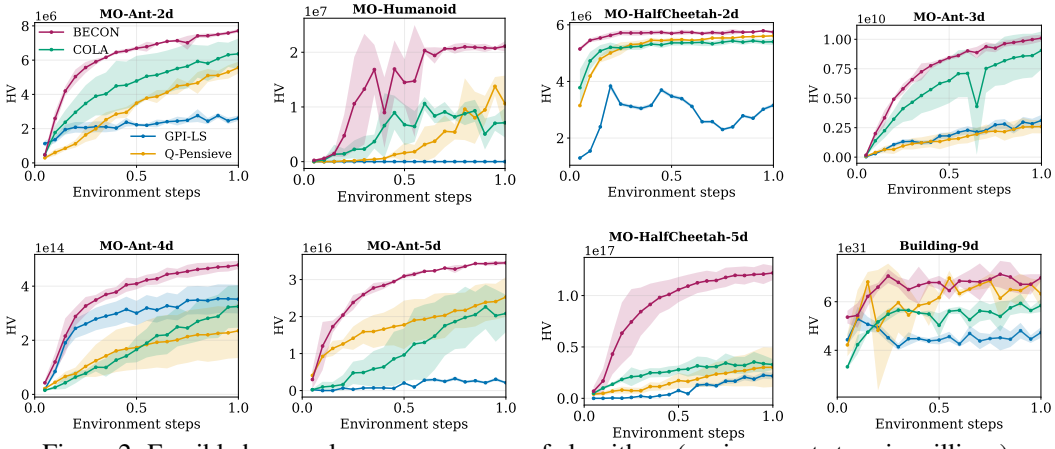

Figure 2: Feasible hypervolume convergence of algorithms (environment steps in millions).

BECON obtains the best FHV and FUT on both MO-LunarLander-4d and Minecart, while keeping MC well below the feasibility threshold. Figure 2 further shows the training dynamics in terms of feasible hypervolume. Across all environments, BECON reaches higher FHV with more stable convergence than the baselines. The smaller uncertainty bands indicate lower variance across seeds, while the consistently higher curves show better sample efficiency and more reliable discovery of feasible Pareto solutions.

Table 3 evaluates online adaptation under drifting preferences and budgets. BECON achieves the best SR and PCU on all tasks, demonstrating stronger online utility and faster post-change recovery than the baselines. BECON also achieves zero violation rate and the lowest MC across all benchmarks, suggesting that its adaptation gains are obtained under stringent constraint compliance rather than relaxed feasibility. These results demonstrate that belief-conditioned task inference and context-dependent dual control enable robust adaptation to task drift.

Additional analyses in Appendix F examine the robustness of BECON and its main design choices. In particular, Appendix F.2 validates the contribution of belief inference and feasibility-aware control, Appendix F.3 compares temporal encoders, and Appendix F.4 studies the effect of history length L. Appendix F.5 and Appendix F.6 analyze sensitivity to the fusion weight $\alpha_{\max}$ and preference drift rate ρ_ω, respectively. Finally, Appendix F.7 shows that the full context-dependent dual $\lambda_\psi(\widetilde{B}_t, \mathbf{z}_t)$ provides the most consistent balance between feasible-front quality and online adaptation.

7 Limitations

BECON focuses on stage-wise admissible cost levels rather than cumulative or remaining-budget constraints. This formulation is useful when safety tolerance is context-dependent and must be enforced locally at each decision step, such as when acceptable energy use, risk exposure, or operational cost changes with user preferences, environmental conditions, or recent interaction history. By treating B_t as an instantaneous admissible cost level, the agent can adapt constraint enforcement to the current task context instead of relying on a fixed global budget. However, this does not directly capture long-horizon resource depletion or trade-offs with future remaining

Table 3: Task-drift adaptation results

Dataset	Method	SR ↑	PCU ↑	VR ↓	MC ↓	Dataset	Method	SR ↑	PCU ↑	VR ↓	MC ↓
MO-Ant-2d	BECON	**4.68**	**4.76**	0	**0.00142**	MO-Humanoid	BECON	**4.45**	**4.49**	0	**0.00015**
	COLA	4.32	4.38	0	0.00217		COLA	2.23	2.19	0	0.00069
	Q-Pensieve	3.56	3.60	0	0.00279		Q-Pensieve	3.37	3.33	0	0.00037
	GPI-LS	3.49	3.63	0	0.00185		GPI-LS	-3.11	0.00	0	0.00160
MO-Ant-3d	BECON	**3.64**	**3.71**	0	**0.00168**	MO-HalfCheetah-2d	BECON	**4.58**	**4.58**	0	**0.00080**
	COLA	3.17	3.25	0	0.00506		COLA	4.31	4.32	0	0.00159
	Q-Pensieve	1.54	1.59	0.0008	0.00424		Q-Pensieve	4.41	4.40	0	0.00117
	GPI-LS	2.93	2.98	0	0.00346		GPI-LS	4.11	4.08	0	0.00167
MO-Ant-4d	BECON	**9.33**	**9.44**	0	**0.00151**	MO-HalfCheetah-5d	BECON	**5.63**	**5.81**	0	**0.00205**
	COLA	6.71	6.86	0	0.00368		COLA	5.03	3.51	0.0023	0.00591
	Q-Pensieve	7.31	7.51	0	0.00297		Q-Pensieve	3.28	1.38	0	0.00474
	GPI-LS	8.57	8.76	0	0.00158		GPI-LS	4.06	4.11	0.0024	0.00436
MO-Ant-5d	BECON	**3.75**	**3.73**	0	**0.00082**	Building-9d	BECON	**7.08**	**7.09**	0	**0.00017**
	COLA	2.81	2.75	0	0.00124		COLA	7.03	7.05	0	0.00566
	Q-Pensieve	3.20	3.16	0	0.00169		Q-Pensieve	6.86	6.86	0	0.00675
	GPI-LS	2.48	2.41	0	0.00194		GPI-LS	6.25	6.21	0	0.00635

budgets. Extending the framework to cumulative constraints would require augmenting the state with remaining budget dynamics and developing belief updates over both task context and residual feasibility. BECON assumes that latent preferences and budgets are sufficiently identifiable from finite interaction histories, which may not hold under uninformative or stale observations. Our experiments isolate task-specification drift under fixed environment dynamics, leaving joint task-environment drift and real-world deployment unexplored.

8 Conclusion

This paper studies constrained preference-conditioned MORL with drifting preferences and budgets as a partially observed task-specification problem. We introduced BECON, a belief-conditioned framework that infers latent task context from recent histories, uncertainty-gates inferred and exposed task signals, and adapts constraint enforcement through a context-dependent dual variable. Our analysis decomposes performance loss into preference-fusion, budget-fusion, and cost-model errors, while characterizing the history-length trade-off under drift. BECON consistently improves feasible-front quality, online adaptation, and constraint satisfaction over strong MORL baselines across various MORL benchmarks. These results suggest that preferences and budgets should be treated as uncertain task observations rather than fixed ground-truth commands.

Acknowledgments and Disclosures

Funding. This research was funded by Hanoi University of Science and Technology under project code T2025-TN-005.

Competing Interests. The authors declare no competing financial interests related to this work.

References

[1] Yibowen Zhao, Yonghui Xu, Di Wang, Yixin Zhang, Qingzhong Li, and Lizhen Cui. A multi-objective explanation framework for graph neural networks. *IEEE Trans. Knowl. Data Eng.*, 2025.

[2] Geng Sun, Jian Xiao, Jiahui Li, Jiacheng Wang, Jiawen Kang, Dusit Niyato, and Shiwen Mao. Aerial reliable collaborative communications for terrestrial mobile users via evolutionary multi-objective deep reinforcement learning. *IEEE Trans. Mobile Comput.*, 2025.

[3] Florian Felten, El-Ghazali Talbi, and Grégoire Danoy. Multi-objective reinforcement learning based on decomposition: A taxonomy and framework. *J. Artif. Intell. Res.*, 2024.

[4] Trinh Van Chien, Bui Trong Duc, Nguyen Xuan Tung, Van Duc Nguyen, Waqas Khalid, Symeon Chatzinotas, and Lajos Hanzo. Single-and multi-objective stochastic optimization for next-generation networks in the generative ai and quantum computing era. *arXiv preprint arXiv:2601.02175*, 2026.

[5] Hung Phan Duc, Duc Bui Trong, Tam Nguyen Thi, and Binh Huynh Thi Thanh. Pareto front grid guided multiobjective optimization in dynamic pickup and delivery problem considering two-sided fairness. In *Proceedings of the Genetic and Evolutionary Computation Conference*, pages 277–285, 2025.

[6] Haoye Lu, Daniel Herman, and Yaoliang Yu. Multi-objective reinforcement learning: Convexity, stationarity and pareto optimality. In *ICLR*, 2023.

[7] Ni Mu, Yao Luan, and Qing-Shan Jia. Preference-based multi-objective reinforcement learning. *IEEE Trans. Autom. Sci. Eng.*, 2025.

[8] Yifu Yuan, Zhenrui Zheng, Zibin Dong, and Jianye HAO. MODULI: Unlocking preference generalization via diffusion models for offline multi-objective reinforcement learning. In *ICML*, 2025.

[9] Toygun Basaklar, Suat Gumussoy, and Umit Ogras. PD-MORL: Preference-driven multi-objective reinforcement learning algorithm. In *ICLR*, 2023.

[10] Pengyi Li, Hongyao Tang, Yifu Yuan, Jianye HAO, Zibin Dong, and YAN ZHENG. COLA: Towards efficient multi-objective reinforcement learning with conflict objective regularization in latent space. In *NeurIPS*, 2025.

[11] Nicolas Carrara, Edouard Leurent, Romain Laroche, Tanguy Urvoy, Odalric-Ambrym Maillard, and Olivier Pietquin. Budgeted reinforcement learning in continuous state space. In *NeurIPS*, 2019.

[12] Haque Ishfaq, Thanh Nguyen-Tang, Songtao Feng, Raman Arora, Mengdi Wang, Ming Yin, and Doina Precup. Offline multitask representation learning for reinforcement learning. *NeurIPS*, 37:70557–70616, 2024.

[13] Ruohong Liu, Yuxin Pan, Linjie Xu, Lei Song, Pengcheng You, Yize Chen, and Jiang Bian. Efficient discovery of pareto front for multi-objective reinforcement learning. In *ICLR*, 2025.

[14] Lihe Li, Ruotong Chen, Ziqian Zhang, Zhichao Wu, Yi-Chen Li, Cong Guan, Yang Yu, and Lei Yuan. Continual multi-objective reinforcement learning via reward model rehearsal. In *IJCAI*, 2024.

[15] Zijian Guo, Weichao Zhou, Shengao Wang, and Wenchao Li. Constraint-conditioned actor-critic for offline safe reinforcement learning. In *ICLR*, 2025.

[16] Xianwei Cao, Dou Quan, Zhenliang Zhang, and Shuang Wang. Learning what matters now: Dynamic preference inference under contextual shifts. In *ICLR*, 2026.

[17] Yucheng Yang, Tianyi Zhou, Mykola Pechenizkiy, and Meng Fang. Preference controllable reinforcement learning with advanced multi-objective optimization. In *ICML*, 2025.

[18] Xin-Qiang Cai, Pushi Zhang, Li Zhao, Jiang Bian, Masashi Sugiyama, and Ashley Llorens. Distributional pareto-optimal multi-objective reinforcement learning. In *NeurIPS*, 2023.

[19] Zifan Wu, Bo Tang, Qian Lin, Chao Yu, Shangqin Mao, Qianlong Xie, Xingxing Wang, and Dong Wang. Off-policy primal-dual safe reinforcement learning. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun, editors, *ICLR*, 2024.

[20] Sobhan Miryoosefi and Chi Jin. A simple reward-free approach to constrained reinforcement learning. In *ICML*, 2022.

[21] Jiashuo Jiang and Yinyu Ye. Achieving $\tilde{O}(1/\epsilon)$ sample complexity for constrained markov decision process. In *NeurIPS*, 2024.

[22] Yassine Chemingui, Aryan Deshwal, Alan Fern, Thanh Nguyen-Tang, and Jana Doppa. Online optimization for offline safe reinforcement learning. In *NeurIPS*, 2025.

[23] Guopeng Li, Matthijs T. J. Spaan, and Julian F. P. Kooij. Off-policy safe reinforcement learning with cost-constrained optimistic exploration. In *ICLR*, 2026.

[24] Yassine Chemingui, Aryan Deshwal, Honghao Wei, Alan Fern, and Jana Doppa. Constraint-adaptive policy switching for offline safe reinforcement learning. In *AAAI*, 2025.

[25] Dohyeong Kim, Mineui Hong, Jeongho Park, and Songhwai Oh. Conflict-averse gradient aggregation for constrained multi-objective reinforcement learning. In *ICLR*, 2025.

[26] Max Schwarzer, Ankesh Anand, Rishab Goel, R Devon Hjelm, Aaron Courville, and Philip Bachman. Data-efficient reinforcement learning with self-predictive representations. In *ICLR*, 2021.

[27] Amy Zhang, Rowan Thomas McAllister, Roberto Calandra, Yarin Gal, and Sergey Levine. Learning invariant representations for reinforcement learning without reconstruction. In *ICLR*, 2021.

[28] Wanpeng Zhang, Yilin Li, Boyu Yang, and Zongqing Lu. Tackling non-stationarity in reinforcement learning via causal-origin representation. In *ICML*, 2024.

[29] Pouya Hamadanian, Arash Nasr-Esfahany, Malte Schwarzkopf, Siddhartha Sen, and Mohammad Alizadeh. Online reinforcement learning in non-stationary context-driven environments. In *ICLR*, 2025.

[30] Long Ma, Fangwei Zhong, and Yizhou Wang. Behavior-agnostic task inference for robust offline in-context reinforcement learning. In *ICML*, 2025.

[31] Lucas N. Alegre, Ana L. C. Bazzan, Diederik M. Roijers, Ann Nowé, and Bruno C. da Silva. Sample-efficient multi-objective learning via generalized policy improvement prioritization. In *Proc. AAMAS*, 2023.

[32] Wei Hung, Bo Kai Huang, Ping-Chun Hsieh, and Xi Liu. Q-pensieve: Boosting sample efficiency of multi-objective RL through memory sharing of q-snapshots. In *ICLR*, 2023.

[33] Nianli Peng, Muhang Tian, and Brandon Fain. Multi-objective reinforcement learning with nonlinear preferences: Provable approximation for maximizing expected scalarized return. *arXiv preprint arXiv:2311.02544*, 2023.

[34] Fan-Ming Luo, Shengyi Jiang, Yang Yu, Zongzhang Zhang, and Yi-Feng Zhang. Adapt to environment sudden changes by learning a context sensitive policy. In *AAAI*, 2022.

A Additional Related Work

MORL and preference-conditioned control. MORL studies sequential decision making with vector-valued returns, where policies must negotiate competing objectives rather than optimize a single scalar reward. Linear scalarization remains a dominant paradigm because it reduces MORL to standard RL under a preference vector, which can recover broad classes of Pareto-optimal solutions under suitable structural assumptions [6]. However, learning a separate policy for each preference is inefficient and poorly scalable. Preference-conditioned methods address this issue by training universal or meta-policies that take preferences as inputs and generalize across the preference simplex [9]. Recent work further improves Pareto-front coverage through constrained Pareto discovery, latent knowledge sharing, and preference-controllable optimization [13, 17]. Distributional MORL extends this view by optimizing over return distributions rather than only expected returns, thereby capturing richer risk-sensitive trade-offs [18]. A complementary line of work considers changing objectives, including continual MORL and latent preference inference under non-stationary priorities [14, 16]. Despite these advances, most preference-conditioned methods still assume that the operative preference is either directly observed, fixed within the decision context, or inferable without explicitly modeling feasibility.

Constrained and safe RL. Constrained RL augments reward maximization with safety, resource, or budget constraints. Budgeted MDPs formalize this trade-off through a cost signal and an adjustable threshold, enabling policies to condition their behavior on remaining risk tolerance [11]. In CMDPs, primal-dual and Lagrangian approaches are widely used because they convert constraint satisfaction into a penalty-weighted optimization problem, but their performance can be sensitive to cost-estimation bias and multiplier dynamics [19]. Recent theoretical work has refined constrained RL through reward-free exploration, approachability, and adaptive resource resolving, yielding sharper guarantees in tabular and structured settings [20, 21]. In practical safe RL, off-policy and offline methods address sample efficiency, OOD actions, and safety during deployment by using conservative cost estimation, constrained exploration, or minimax formulations [22, 23]. Adaptive-threshold methods further condition policies or policy selection on varying safety limits, but they usually treat the constraint threshold as an explicit input rather than a partially observed quantity [15, 24]. When constraints are combined with multiple objectives, recent algorithms cast objective conflicts and safety requirements as a unified constrained optimization problem [25]. Yet the safety trade-off is often governed by a global or weakly contextualized multiplier, which is restrictive when both preferences and budgets evolve along a trajectory.

Latent context inference and non-stationary RL. Latent representation learning has become a promising technique for improving generalization and sample efficiency in RL. Self-supervised temporal prediction encourages agents to learn representations that preserve control-relevant dynamics across future states [26]. Bisimulation metrics provide another principled route by learning latent spaces that retain behaviorally relevant information while discarding nuisance variation [27]. For non-stationary environments, recent methods learn representations that expose the causal origin of distributional shifts or regularize policies locally against catastrophic forgetting [28, 29]. In-context RL similarly uses recent trajectories to infer the active task at test time without gradient updates, although many offline formulations remain brittle under distribution shift in the context-collection policy [30]. These approaches show that history can encode hidden task structure, but they do not directly integrate latent context inference with constrained MORL, where both user intent and feasible safety budget may be uncertain.

B Training Procedure

Algorithm 1 summarizes the training procedure of BECON.

C Additional Theory and Proofs

This appendix provides assumptions, auxiliary results, and proofs for Section 5.

Algorithm 1 BECON Training Procedure

Require: Replay buffer $\mathcal{D}$, history length L, gate scale $\alpha_{\max}$

Initialize encoders, task decoders, actor, twin critics, cost model, dual network, and targets
for each environment step t **do**
  Encode $\mathbf{e}_t = f_{\text{enc}}(\mathbf{s}_t)$ and $\mathbf{z}_t \sim q_\phi(\mathbf{z}_t \mid H_t)$
  Decode $(\boldsymbol{\omega}_t^{\text{inf}}, B_t^{\text{inf}})$ from $\mathbf{z}_t$
  Compute uncertainty gates $\boldsymbol{\alpha}_t$ and fuse $(\widehat{\boldsymbol{\omega}}_t, \widehat{B}_t)$ with $(\boldsymbol{\omega}_t^{\text{inf}}, B_t^{\text{inf}})$ to obtain $(\widetilde{\boldsymbol{\omega}}_t, \widetilde{B}_t)$
  Act with $\pi_\theta(\cdot \mid \mathbf{e}_t, \mathbf{z}_t, \widetilde{\boldsymbol{\omega}}_t, \widetilde{B}_t)$ and store the transition in $\mathcal{D}$
  Sample mini-batches and update $\mathcal{L}_{\text{critic}}$, $\mathcal{L}_{\text{belief}}$, $\mathcal{L}_{\text{dyn}}$, $\mathcal{L}_{\text{cost}}$, $\mathcal{L}_{\text{actor}}$, and $\mathcal{L}_\lambda$
  Soft-update target networks
end for

C.1 Belief Sufficiency under Task Drift

The belief module in BECON can be viewed as an amortized, finite-dimensional approximation to belief-state filtering over the hidden task context.

Proposition 2. *Let the hidden augmented state be $m_t = (\mathbf{s}_t, \tau_t)$, where $\tau_t = (\boldsymbol{\omega}_t, B_t)$. Suppose the environment transition is fixed and Markovian, $\mathbf{s}_{t+1} \sim P(\cdot \mid \mathbf{s}_t, \mathbf{a}_t)$, the task evolves according to a Markov drift kernel $K_\tau(\tau_{t+1} \mid \tau_t)$, and the exposed descriptor is generated from an observation kernel $O(\widehat{\tau}_t \mid \tau_t)$. Then the posterior belief $b_t(m) := \mathbb{P}(m_t = m \mid o_{0:t}, \mathbf{a}_{0:t-1})$, with $o_t = (\mathbf{s}_t, \widehat{\tau}_t)$, is a sufficient statistic for optimal control. In particular, there exists an optimal policy $\pi^\star(\mathbf{a}_t \mid b_t)$.*

Proof. Given $m_t = (\mathbf{s}_t, \tau_t)$ and action $\mathbf{a}_t$, the next augmented state distribution factors as

$$\mathbb{P}(m_{t+1} \mid m_t, \mathbf{a}_t) = P(\mathbf{s}_{t+1} \mid \mathbf{s}_t, \mathbf{a}_t) K_\tau(\tau_{t+1} \mid \tau_t). \tag{14}$$

The next observation is generated from the realized environment state together with the task-observation kernel. Hence, once b_t and $\mathbf{a}_t$ are given, the full history provides no additional information about future hidden states, observations, rewards, or costs. The Bayesian update maps $(b_t, \mathbf{a}_t, o_{t+1})$ to b_{t+1}, so the belief process is Markov. Dynamic programming on this belief MDP therefore yields an optimal policy that depends on past interactions only through b_t. □

Proposition 2 justifies the use of the learned variational belief $q_\phi(\mathbf{z}_t \mid H_t)$. Since the exact belief b_t is generally intractable, BECON uses $\mathbf{z}_t$ as an amortized, finite-dimensional summary of the recent interaction history.

C.2 Proof of Proposition 1

Proof. Let $\mathbf{y}_t = (1 - \boldsymbol{\alpha}_t) \odot \widehat{\boldsymbol{\omega}}_t + \boldsymbol{\alpha}_t \odot \boldsymbol{\omega}_t^{\text{inf}}$. By construction, $\widetilde{\boldsymbol{\omega}}_t = \Pi_\Omega(\mathbf{y}_t)$, hence $\widetilde{\boldsymbol{\omega}}_t \in \Omega$. Since $\boldsymbol{\omega}_t \in \Omega$, we have $\Pi_\Omega(\boldsymbol{\omega}_t) = \boldsymbol{\omega}_t$. Euclidean projection onto a closed convex set is non-expansive, so

$$\|\widetilde{\boldsymbol{\omega}}_t - \boldsymbol{\omega}_t\|_2 = \|\Pi_\Omega(\mathbf{y}_t) - \Pi_\Omega(\boldsymbol{\omega}_t)\|_2 \le \|\mathbf{y}_t - \boldsymbol{\omega}_t\|_2 . \tag{15}$$

For each coordinate, $y_{t,i} - \omega_{t,i} = (1 - \alpha_{t,i})(\widehat{\omega}_{t,i} - \omega_{t,i}) + \alpha_{t,i}(\omega_{t,i}^{\text{inf}} - \omega_{t,i})$. Therefore,

$$\begin{aligned} \|\mathbf{y}_t - \boldsymbol{\omega}_t\|_2 &\le \|(1 - \boldsymbol{\alpha}_t) \odot (\widehat{\boldsymbol{\omega}}_t - \boldsymbol{\omega}_t)\|_2 + \left\|\boldsymbol{\alpha}_t \odot (\boldsymbol{\omega}_t^{\text{inf}} - \boldsymbol{\omega}_t)\right\|_2 \\ &\le (1 - \alpha_t^-) \|\widehat{\boldsymbol{\omega}}_t - \boldsymbol{\omega}_t\|_2 + \alpha_t^+ \left\|\boldsymbol{\omega}_t^{\text{inf}} - \boldsymbol{\omega}_t\right\|_2 . \end{aligned} \tag{16}$$

For the budget, $\widetilde{B}_t = (1 - \alpha_t^B)\widehat{B}_t + \alpha_t^B B_t^{\text{inf}}$. Thus,

$$\begin{aligned} \left|\widetilde{B}_t - B_t\right| &= \left|(1 - \alpha_t^B)(\widehat{B}_t - B_t) + \alpha_t^B (B_t^{\text{inf}} - B_t)\right| \\ &\le (1 - \alpha_t^B) \left|\widehat{B}_t - B_t\right| + \alpha_t^B \left|B_t^{\text{inf}} - B_t\right| . \end{aligned} \tag{17}$$

If $\widehat{B}_t \ge 0$ and $B_t^{\text{inf}} \ge 0$, then $\widetilde{B}_t \ge 0$ because it is a convex combination. □

C.3 Proof of Theorem 1

We use the elementary fact that the hinge function is 1-Lipschitz:

$$|[x]_+ - [y]_+| \le |x - y|, \qquad \forall x, y \in \mathbb{R}. \tag{18}$$

Proof. For any fixed policy π,

$$\begin{aligned}\left|\mathcal{G}_{\tau_{0:T-1}}(\pi) - \widehat{\mathcal{G}}_{\widetilde{\tau}_{0:T-1}}(\pi)\right| \le \mathbb{E}_{\pi,\widetilde{\tau}}\left[\sum_{t=0}^{T-1}\gamma^t\left|(\boldsymbol{\omega}_t - \widetilde{\boldsymbol{\omega}}_t)^\top \mathbf{r}_t\right|\right] \\ + \mathbb{E}_{\pi,\widetilde{\tau}}\left[\sum_{t=0}^{T-1}\gamma^t\lambda_\psi(\widetilde{B}_t, \mathbf{z}_t)\left|[\bar{c}_{\pi,t} - B_t]_+ - [\widehat{c}_{\pi,t} - \widetilde{B}_t]_+\right|\right].\end{aligned} \tag{19}$$

By Cauchy-Schwarz and bounded rewards, $\left|(\boldsymbol{\omega}_t - \widetilde{\boldsymbol{\omega}}_t)^\top \mathbf{r}_t\right| \le R_2 \left\|\boldsymbol{\omega}_t - \widetilde{\boldsymbol{\omega}}_t\right\|_2 = R_2\delta_t^\omega$. Using Eq. (18),

$$\begin{aligned}\left|[\bar{c}_{\pi,t} - B_t]_+ - [\widehat{c}_{\pi,t} - \widetilde{B}_t]_+\right| &\le \left|(\bar{c}_{\pi,t} - B_t) - (\widehat{c}_{\pi,t} - \widetilde{B}_t)\right| \\ &\le |\bar{c}_{\pi,t} - \widehat{c}_{\pi,t}| + \left|\widetilde{B}_t - B_t\right| = \epsilon_t^c(\pi) + \delta_t^B.\end{aligned} \tag{20}$$

Since $\lambda_\psi(\widetilde{B}_t, \mathbf{z}_t) \le \Lambda$, combining the two bounds gives Eq. (11).

For the near-optimality bound, fix any comparator $\pi \in \Pi$ and decompose the true performance gap through the fused model-based objective:

$$\begin{aligned}&\mathcal{G}_{\tau_{0:T-1}}(\pi) - \mathcal{G}_{\tau_{0:T-1}}(\widetilde{\pi}) \\ &= \left[\mathcal{G}_{\tau_{0:T-1}}(\pi) - \widehat{\mathcal{G}}_{\widetilde{\tau}_{0:T-1}}(\pi)\right] + \left[\widehat{\mathcal{G}}_{\widetilde{\tau}_{0:T-1}}(\pi) - \widehat{\mathcal{G}}_{\widetilde{\tau}_{0:T-1}}(\widetilde{\pi})\right] + \left[\widehat{\mathcal{G}}_{\widetilde{\tau}_{0:T-1}}(\widetilde{\pi}) - \mathcal{G}_{\tau_{0:T-1}}(\widetilde{\pi})\right].\end{aligned} \tag{21}$$

The assumed ϵ_{opt}-optimality of $\widetilde{\pi}$ bounds the middle term by ϵ_{opt}, and Eq. (11) controls the two perturbation terms. Taking the supremum over $\pi \in \Pi$ and using $\epsilon_t^c(\pi) \le \bar{\epsilon}_t^c$ yields Eq. (12). □

C.4 Proof of Theorem 2

We first quantify how task drift makes an L-step history stale relative to the current task.

Lemma 1. *Under the drift model,*

$$\frac{1}{L}\sum_{\ell=1}^{L}\left\|\boldsymbol{\omega}_t - \boldsymbol{\omega}_{t-\ell}\right\|_2 \le \frac{L+1}{2}\sqrt{2}\rho_\omega, \tag{22}$$

and

$$\frac{1}{L}\sum_{\ell=1}^{L}|B_t - B_{t-\ell}| \le \frac{L+1}{2}D_B\rho_B. \tag{23}$$

Proof. The Euclidean diameter of the probability simplex is $\sqrt{2}$. Since Euclidean projection onto a closed convex set is non-expansive,

$$\left\|\boldsymbol{\omega}_{k+1} - \boldsymbol{\omega}_k\right\|_2 = \left\|\Pi_\Omega\left((1-\rho_\omega)\boldsymbol{\omega}_k + \rho_\omega\bar{\boldsymbol{\omega}}_k\right) - \Pi_\Omega(\boldsymbol{\omega}_k)\right\|_2 \le \rho_\omega\left\|\bar{\boldsymbol{\omega}}_k - \boldsymbol{\omega}_k\right\|_2 \le \sqrt{2}\rho_\omega. \tag{24}$$

Thus $\left\|\boldsymbol{\omega}_t - \boldsymbol{\omega}_{t-\ell}\right\|_2 \le \ell\sqrt{2}\rho_\omega$. Averaging over $\ell = 1, \ldots, L$ gives Eq. (22). Similarly,

$$|B_{k+1} - B_k| = \rho_B|\bar{B}_k - B_k| \le D_B\rho_B, \tag{25}$$

so $|B_t - B_{t-\ell}| \le \ell D_B\rho_B$, and averaging gives Eq. (23). □

Proof of Theorem 2. If the task were fixed throughout the history window, the decoder errors would scale as $A_\omega L^{-1/2}$ for preferences and $A_B L^{-1/2}$ for budgets by assumption. Under task drift, however, the history window contains samples generated under past tasks $\tau_{t-L}, \ldots, \tau_{t-1}$ rather than

the current task τ_t. Thus, the decoder error decomposes into a statistical term and a staleness term. By the triangle inequality and Lemma 1,

$$\mathbb{E}\left[\left\|\boldsymbol{\omega}_t^{\inf} - \boldsymbol{\omega}_t\right\|_2\right] \le A_\omega L^{-1/2} + \frac{1}{L}\sum_{\ell=1}^{L}\left\|\boldsymbol{\omega}_{t-\ell} - \boldsymbol{\omega}_t\right\|_2 \le A_\omega L^{-1/2} + \frac{L+1}{2}\sqrt{2}\rho_\omega. \tag{26}$$

The budget bound follows identically

$$\mathbb{E}\left[\left|B_t^{\inf} - B_t\right|\right] \le A_B L^{-1/2} + \frac{L+1}{2} D_B \rho_B. \tag{27}$$

Ignoring lower-order constants, the preference error has the form

$$f(L) = A_\omega L^{-1/2} + C_\omega \rho_\omega L, \qquad C_\omega = \frac{\sqrt{2}}{2}. \tag{28}$$

Treating L as continuous,

$$f'(L) = -\frac{1}{2} A_\omega L^{-3/2} + C_\omega \rho_\omega. \tag{29}$$

Setting $f'(L) = 0$ yields

$$L_\omega^\star = \left(\frac{A_\omega}{2C_\omega \rho_\omega}\right)^{2/3} \asymp \left(\frac{A_\omega}{\rho_\omega}\right)^{2/3}. \tag{30}$$

Substituting this value into $f(L)$ gives

$$\min_L f(L) = \mathcal{O}\left(A_\omega^{2/3} \rho_\omega^{1/3}\right). \tag{31}$$

The budget result follows by replacing A_ω with A_B and $C_\omega \rho_\omega$ with $(D_B/2)\rho_B$. □

C.5 Oracle Fusion and Inverse-Uncertainty Weighting

The uncertainty gate in BECON can be interpreted as an amortized approximation to an oracle fusion rule. For a generic task variable θ_t, let $\widehat{\theta}_t$ denote the exposed estimate and $\theta_t^{\inf}$ the belief-derived estimate. We consider the gated fusion

$$\widetilde{\theta}_t(\alpha) = (1-\alpha)\widehat{\theta}_t + \alpha \theta_t^{\inf}, \qquad \alpha \in [0,1]. \tag{32}$$

Proposition 3. *Let $e^{\exp} = \widehat{\theta}_t - \theta_t$ and $e^{\inf} = \theta_t^{\inf} - \theta_t$. The mean-squared fusion risk $\mathcal{R}(\alpha) = \mathbb{E}\left[\left\|(1-\alpha)e^{\exp} + \alpha e^{\inf}\right\|_2^2\right]$ is minimized over $\alpha \in [0,1]$ by*

$$\alpha^\star = \Pi_{[0,1]}\left(\frac{\mathbb{E}\|e^{\exp}\|_2^2 - \mathbb{E}\langle e^{\exp}, e^{\inf}\rangle}{\mathbb{E}\|e^{\exp} - e^{\inf}\|_2^2}\right). \tag{33}$$

If the two errors are unbiased, independent, and have variances $\sigma_{\exp}^2$ and $\sigma_{\inf}^2$, then

$$\alpha^\star = \frac{\sigma_{\exp}^2}{\sigma_{\exp}^2 + \sigma_{\inf}^2}. \tag{34}$$

Proof. Let $d = e^{\inf} - e^{\exp}$. The fusion error can be written as $e^{\exp} + \alpha d$, hence

$$\mathcal{R}(\alpha) = \mathbb{E}\left[\|e^{\exp} + \alpha d\|_2^2\right] = \mathbb{E}\|e^{\exp}\|_2^2 + 2\alpha \mathbb{E}\langle e^{\exp}, d\rangle + \alpha^2 \mathbb{E}\|d\|_2^2. \tag{35}$$

This is a convex quadratic. The unconstrained minimizer is

$$\alpha_{\mathrm{unc}}^\star = -\frac{\mathbb{E}\langle e^{\exp}, d\rangle}{\mathbb{E}\|d\|_2^2} = \frac{\mathbb{E}\|e^{\exp}\|_2^2 - \mathbb{E}\langle e^{\exp}, e^{\inf}\rangle}{\mathbb{E}\|e^{\exp} - e^{\inf}\|_2^2}. \tag{36}$$

Projecting onto $[0,1]$ gives Eq. (33). Under unbiasedness and independence, $\mathbb{E}\langle e^{\exp}, e^{\inf}\rangle = 0$, which gives Eq. (34). □

C.6 Posterior Uncertainty and Decoded-Task Variance

Let

$$d_\omega(\mathbf{z}) = \frac{\varphi(g_\omega(\mathbf{z})) + \varepsilon \mathbf{1}_{d_\omega}}{\mathbf{1}_{d_\omega}^\top (\varphi(g_\omega(\mathbf{z})) + \varepsilon \mathbf{1}_{d_\omega})}, \qquad d_B(\mathbf{z}) = \varphi(g_B(\mathbf{z})), \tag{37}$$

where $\varphi(x) = \log(1 + \exp(x))$, $\mathbf{1}_{d_\omega} \in \mathbb{R}^{d_\omega}$ denotes the all-ones vector, d_ω is the preference dimension, and $\varepsilon > 0$ ensures strictly positive normalized preference weights.

Lemma 2. *Assume g_ω and g_B are Lipschitz with constants L_{g_ω} and L_{g_B}. Then d_B is L_{g_B}-Lipschitz. Moreover, d_ω is Lipschitz with constant*

$$L_\omega \le \frac{\sqrt{K(K+3)}}{K\varepsilon} L_{g_\omega}. \tag{38}$$

Therefore, if $\mathbf{z}_t \sim \mathcal{N}(\boldsymbol{\mu}_t, \mathrm{Diag}(\boldsymbol{\sigma}_t^2))$, then

$$\mathbb{E}\left[\|d_\omega(\mathbf{z}_t) - d_\omega(\boldsymbol{\mu}_t)\|_2^2 \mid H_t \right] \le L_\omega^2 \|\boldsymbol{\sigma}_t\|_2^2, \tag{39}$$

and

$$\mathbb{E}\left[|d_B(\mathbf{z}_t) - d_B(\boldsymbol{\mu}_t)|^2 \mid H_t \right] \le L_{g_B}^2 \|\boldsymbol{\sigma}_t\|_2^2. \tag{40}$$

This lemma shows that posterior variance controls the sensitivity of decoded task estimates to latent sampling noise, supporting its use as a practical reliability proxy.

C.7 Dual Interpretation for Instantaneous Safety

The context-dependent dual variable in BECON is

$$\lambda_\psi(\widetilde{B}_t, \mathbf{z}_t) = \Lambda \sigma(f_\lambda(\widetilde{B}_t, \mathbf{z}_t)), \tag{41}$$

where $\sigma(\cdot)$ is the sigmoid function. This parameterization ensures $0 \le \lambda_\psi(\widetilde{B}_t, \mathbf{z}_t) \le \Lambda$ and preserves monotonicity with respect to the pre-activation f_λ. For one sample, the dual loss is

$$\ell_\lambda(\psi) = -\lambda_\psi(\widetilde{B}_t, \mathbf{z}_t) \left(\widehat{c}_t(\mathbf{a}_t) - \widetilde{B}_t \right). \tag{42}$$

Its gradient is

$$\nabla_\psi \ell_\lambda(\psi) = -\Lambda \left(\widehat{c}_t(\mathbf{a}_t) - \widetilde{B}_t \right) \sigma(f_\lambda(\widetilde{B}_t, \mathbf{z}_t)) \left(1 - \sigma(f_\lambda(\widetilde{B}_t, \mathbf{z}_t)) \right) \nabla_\psi f_\lambda(\widetilde{B}_t, \mathbf{z}_t). \tag{43}$$

Thus, gradient descent on the dual loss increases the dual pre-activation when the predicted instantaneous violation $\widehat{c}_t(\mathbf{a}_t) - \widetilde{B}_t$ is positive, strengthening the actor penalty. When the predicted violation is negative, the update reduces the dual penalty and avoids excessive conservatism. The gap between predicted and true instantaneous safety is bounded by the cost-model error and the budget-fusion error, as characterized in Theorem 1.

D Environment Setup

We evaluate BECON on multi-objective MuJoCo locomotion tasks and a multi-zone building-control task. Each environment returns a vector reward $\mathbf{r}_t \in \mathbb{R}^K$, and a preference $\boldsymbol{\omega} \in \Omega$ scalarizes the return as $\boldsymbol{\omega}^\top \sum_t \mathbf{r}_t$. Feasibility is controlled by an external budget channel rather than by adding cost as an extra objective. Unless otherwise stated, the per-step cost is $c_t = 0.01 K_a^{-1} \sum_{i=1}^{K_a} a_{t,i}^2$, where K_a is the action dimension, and a trajectory is feasible when its average cost satisfies $T^{-1} \sum_{t=0}^{T-1} c_t \le B$ with $B = 0.01$. For static feasible-front evaluation, B is fixed and preferences are drawn from the evaluation grid. For task-drift evaluation, the hidden preference and budget are refreshed every 50 steps. We use $\rho_\omega = 0.25$ for preference drift, $\rho_B = 0.30$ for budget drift, $B_t \in [0.005, 0.020]$, preference observation noise 0.15, and budget observation delay 20 steps.

- **MO-HalfCheetah-2d.** The environment has $\mathcal{S} \subset \mathbb{R}^{17}$, $\mathcal{A} \subset \mathbb{R}^6$, and $K = 2$. Let v_x denote the forward velocity and $C_{\text{alive}} = 1$. The objectives are $R_1 = \min(4, v_x) + C_{\text{alive}}$ and $R_2 = 4 - \sum_i a_i^2 + C_{\text{alive}}$, corresponding to forward speed and energy-efficient control.

Table 4: Environment and drift settings

Environment	K	State dim.	Action dim.	Eval. prefs.	B	ρ_ω	ρ_B	Drift B_t range
MO-Ant-2d	2	27	8	201	0.01	0.25	0.30	$[0.005, 0.020]$
MO-HalfCheetah-2d	2	17	6	201	0.01	0.25	0.30	$[0.005, 0.020]$
MO-Hopper-2d	2	11	3	201	0.01	0.25	0.30	$[0.005, 0.020]$
MO-Humanoid	2	376	17	201	0.01	0.25	0.30	$[0.005, 0.020]$
MO-Ant-3d	3	27	8	231	0.01	0.25	0.30	$[0.005, 0.020]$
MO-Ant-4d	4	27	8	286	0.01	0.25	0.30	$[0.005, 0.020]$
MO-Ant-5d	5	27	8	126	0.01	0.25	0.30	$[0.005, 0.020]$
MO-HalfCheetah-5d	5	17	6	126	0.01	0.25	0.30	$[0.005, 0.020]$
Building-9d	9	29	23	32	0.01	0.25	0.30	$[0.005, 0.020]$

- **MO-HalfCheetah-5d.** This task uses the same state and action dimensions as MO-HalfCheetah-2d, with $K = 5$. Let h be the body height and h_0 the initial body height. The objectives are $R_1 = 0.5v_x + C_{\text{alive}}$, $R_2 = 20(h - h_0) + C_{\text{alive}}$, $R_3 = 4 - \frac{3}{2}(a_0^2 + a_3^2) + C_{\text{alive}}$, $R_4 = 4 - \frac{3}{2}(a_1^2 + a_2^2) + C_{\text{alive}}$, and $R_5 = 4 - \frac{3}{2}(a_3^2 + a_4^2) + C_{\text{alive}}$. These objectives capture forward speed, jump height, and actuator-group efficiency.
- **MO-Ant-2d.** The environment has $\mathcal{S} \subset \mathbb{R}^{27}$, $\mathcal{A} \subset \mathbb{R}^8$, and $K = 2$. Let v_x and v_y denote torso velocities along the two horizontal axes. With $R_{\text{other}} = 1 - 0.5 \sum_i a_i^2$, the objectives are $R_1 = v_x + R_{\text{other}}$ and $R_2 = v_y + R_{\text{other}}$, encouraging locomotion in two directions while penalizing large actions.
- **MO-Ant-3d.** This task uses $\mathcal{S} \subset \mathbb{R}^{27}$, $\mathcal{A} \subset \mathbb{R}^8$, and $K = 3$. The objectives are $R_1 = v_x$, $R_2 = v_y$, and $R_3 = 4 - 0.5 \sum_i a_i^2$, corresponding to movement along two horizontal axes and energy efficiency.
- **MO-Ant-4d.** This task uses the Ant state and action spaces with $K = 4$. Each objective combines forward velocity with one actuator-pair efficiency term: $R_1 = 4 - 2(a_0^2 + a_1^2) + v_x$, $R_2 = 4 - 2(a_2^2 + a_3^2) + v_x$, $R_3 = 4 - 2(a_4^2 + a_5^2) + v_x$, and $R_4 = 4 - 2(a_6^2 + a_7^2) + v_x$.
- **MO-Ant-5d.** This task separates forward speed from four actuator-group efficiency objectives: $R_1 = v_x$, $R_2 = 4 - 2(a_0^2 + a_1^2)$, $R_3 = 4 - 2(a_2^2 + a_3^2)$, $R_4 = 4 - 2(a_4^2 + a_5^2)$, and $R_5 = 4 - 2(a_6^2 + a_7^2)$. Training preferences are sampled hierarchically by first allocating mass between speed and energy and then distributing the energy mass across actuator groups.
- **MO-Hopper-2d.** The environment has $\mathcal{S} \subset \mathbb{R}^{11}$, $\mathcal{A} \subset \mathbb{R}^3$, and $K = 2$. Let v_x denote forward velocity, h the body height, and h_0 the initial height. With $R_{\text{other}} = 1 - 2 \times 10^{-4} \sum_i a_i^2$, the objectives are $R_1 = 1.5v_x + R_{\text{other}}$ and $R_2 = 12(h - h_0) + R_{\text{other}}$, rewarding forward running and jumping height. Since Hopper actions have a larger environment-scale range than $[-1, 1]$, the action-cost constraint is more demanding.
- **MO-Humanoid.** The environment has $\mathcal{S} \subset \mathbb{R}^{376}$, $\mathcal{A} \subset \mathbb{R}^{17}$, and $K = 2$. Let v_x denote mass-center velocity and $C_{\text{alive}} = 3$. The objectives are $R_1 = 1.25v_x + C_{\text{alive}}$ and $R_2 = 3 - 4 \sum_i \text{ctrl}_i^2 + C_{\text{alive}}$, capturing forward locomotion and energy-efficient control. Episodes terminate when the torso height leaves $[1.0, 2.0]$.
- **Building-9d.** The building-control task uses the OfficeLarge building under Warm Marine weather in ElPaso. The environment has $\mathcal{S} \subset \mathbb{R}^{29}$, $\mathcal{A} \subset \mathbb{R}^{23}$, and $K = 9$. The action dimensions correspond to HVAC controls for 23 zones, and observations include normalized zone temperatures and exogenous variables such as outdoor temperature, irradiance, ground temperature, occupancy power, carbon intensity, and electricity price. The zones are partitioned into bottom, middle, and top groups with sizes 9, 7, and 7. Let $e = X_{\text{new}} \odot A_{\text{C}} - 20A_{\text{C}}$ be the temperature error relative to 20°C. The comfort objectives are $R_1 = (20 \cdot 9 - \|e_{\text{bottom}}\|_1)/20$, $R_2 = (20 \cdot 7 - \|e_{\text{middle}}\|_1)/20$, and $R_3 = (20 \cdot 7 - \|e_{\text{top}}\|_1)/20$. The energy objectives are $R_4 = (20 \cdot 9 - \|a_{\text{bottom}}\|_1)/20$, $R_5 = (20 \cdot 7 - \|a_{\text{middle}}\|_1)/20$, and $R_6 = (20 \cdot 7 - \|a_{\text{top}}\|_1)/20$. The ramping objectives are $R_7 = 9 - \left| \|a_{\text{bottom}}^{t-1}\|_1 - \|a_{\text{bottom}}^{t-2}\|_1 \right|$, $R_8 = 7 - \left| \|a_{\text{middle}}^{t-1}\|_1 - \|a_{\text{middle}}^{t-2}\|_1 \right|$, and $R_9 = 7 - \left| \|a_{\text{top}}^{t-1}\|_1 - \|a_{\text{top}}^{t-2}\|_1 \right|$. For the first two transitions, the ramping rewards are set to zero.

E Implementation Details

BECON uses a shared latent state encoder, a history-conditioned belief encoder, twin distributional critics, a tanh-squashed Gaussian actor, a distributional cost model, and a context-dependent dual network. The state encoder produces a 64-dimensional latent embedding, and the GRU belief encoder infers a 64-dimensional latent task context from recent transition features $[\mathbf{s}_t, \mathbf{a}_t, \mathbf{r}_t, \widehat{\boldsymbol{\omega}}_t, c_t, \widehat{B}_t]$. All experiments are run on a workstation with an Intel Core i9-14900K CPU, 64GB RAM, and an NVIDIA GeForce RTX 5070 Ti GPU. The main implementation parameters are summarized in Table 5.

Table 5: BECON implementation details

Parameter	**Setting**
State encoder	MLP $(256, 256)$, ELU, LayerNorm output
Latent dimension	64
Belief dimension	64
Actor	Tanh-squashed Gaussian policy, MLP $(256, 256)$
Critics	Twin distributional critics with 51 atoms per objective
Value support	Initial support $[-500, 500]$, adaptive support enabled
Cost model	Distributional cost head with 51 atoms on $[0, 1]$
Dual variable	$\lambda_\psi(\widetilde{B}_t, \mathbf{z}_t)$
Optimizer	Adam
Replay buffer size	10^6 transitions
Batch size	256
Discount factor	$\gamma = 0.99$
Target update rate	0.005
Learning rates	3×10^{-4} for actor/main networks; 10^{-4} for dual
Entropy coefficient	$\alpha_H = 0.2$

F Additional Experiments

F.1 Constraint-Tightness Sensitivity

We further evaluate trained policies under stringent admissible cost levels to test whether feasible-front quality is preserved when the stage-wise constraint becomes tighter. Table 6 shows that BECON maintains low violation rates and nonzero feasible hypervolume in cases where COLA becomes partially or fully infeasible.

F.2 Component Ablation

Table 7 isolates the contribution of the main components of BECON. The full model achieves the best FHV and FUT across all tasks, indicating that history-conditioned belief inference and constraint-aware control jointly improve feasible-front quality. Using only the exposed descriptor consistently reduces FHV, especially on higher-dimensional Ant tasks, showing that recent histories provide useful evidence for recovering the latent task context. Removing the constraint penalty can reduce MC in some cases, but this does not translate into better feasible-front quality, as FHV, FUT, and $|\mathcal{P}|$ often decrease. These results suggest that BECON improves constrained MORL performance by balancing preference-aligned return with feasibility, rather than by merely minimizing cost.

F.3 History Encoder Architecture

Table 8 evaluates different temporal encoders for the belief module under task drift. GRU achieves the best overall performance on MO-HalfCheetah-2d and MO-Ant-4d, with the strongest FHV, FUT, SR, and PCU in both tasks. On MO-Ant-5d, TCN and Transformer become more competitive, suggesting that higher-dimensional preference structure can benefit from richer temporal aggregation. Overall,

Table 6: Constraint-tightness sensitivity under different admissible cost levels

Dataset	Method	B	VR ↓	FHV ↑
Building-9d	BECON	0.005	0.000	6.82×10^{31}
Building-9d	COLA	0.005	1.000	0
MO-Hopper-2d	BECON	0.010	0.005	1.29×10^{7}
MO-Hopper-2d	COLA	0.010	1.000	0
MO-HalfCheetah-5d	BECON	0.005	0.000	1.15×10^{17}
MO-HalfCheetah-5d	COLA	0.005	0.548	3.04×10^{16}

Table 7: Component ablation under feasibility constraints

Dataset	**Variant**	**FHV ↑**	**FUT ↑**	**MC** $(\times10^{-3})$ **↓**	$\lvert\mathcal{P}\rvert$ **↑**
MO-Ant-2d	Full BECON	$\mathbf{5.790\times10^{6}}$	**2224.97**	1.548	**45**
	Exposed descriptor only	5.695×10^{6}	2224.55	1.634	39
	w/o constraint penalty	5.633×10^{6}	2203.82	1.603	34
	w/o history belief	5.115×10^{6}	2100.43	**1.473**	38
MO-HalfCheetah-2d	Full BECON	$\mathbf{5.789\times10^{6}}$	**2346.69**	**0.806**	80
	Exposed descriptor only	5.675×10^{6}	2342.06	0.924	68
	w/o constraint penalty	5.737×10^{6}	2340.16	0.882	**92**
	w/o history belief	5.648×10^{6}	2341.94	0.924	82
MO-Ant-3d	Full BECON	$\mathbf{7.377\times10^{9}}$	**1823.95**	1.364	**134**
	Exposed descriptor only	6.163×10^{9}	1712.09	1.970	116
	w/o constraint penalty	7.032×10^{9}	1797.64	**1.285**	106
	w/o history belief	6.629×10^{9}	1758.62	1.888	110
MO-Ant-4d	Full BECON	$\mathbf{3.739\times10^{14}}$	**4380.32**	1.529	**17**
	Exposed descriptor only	3.561×10^{14}	4335.54	1.659	15
	w/o constraint penalty	2.942×10^{14}	4123.29	1.866	11
	w/o history belief	3.374×10^{14}	4276.89	**1.483**	4
MO-Ant-5d	Full BECON	$\mathbf{3.013\times10^{16}}$	**1991.09**	0.255	97
	Exposed descriptor only	2.594×10^{16}	1906.18	0.651	90
	w/o constraint penalty	2.563×10^{16}	1897.47	**0.239**	94
	w/o history belief	2.400×10^{16}	1889.73	0.531	**104**

Table 8: History-encoder ablation under task drift

Dataset	**Encoder**	**FHV ↑**	**FUT ↑**	**MC** $(\times10^{-3})$ **↓**	**SR ↑**	**PCU ↑**
MO-HalfCheetah-2d	GRU	$\mathbf{5.797\times10^{6}}$	**2351.60**	**0.793**	**4.437**	**4.366**
	MLP	5.673×10^{6}	2348.11	0.797	4.421	4.356
	TCN	5.778×10^{6}	2344.25	0.801	4.428	4.350
	Transformer	5.775×10^{6}	2345.66	0.820	4.413	4.360
MO-Ant-4d	GRU	$\mathbf{4.418\times10^{14}}$	**4564.29**	1.624	**9.064**	**9.198**
	MLP	4.189×10^{14}	4489.59	**1.558**	8.951	9.073
	TCN	4.103×10^{14}	4464.43	1.604	8.660	8.868
	Transformer	4.351×10^{14}	4545.02	1.647	8.999	9.059
MO-Ant-5d	GRU	3.099×10^{16}	1976.85	**0.292**	3.862	3.866
	MLP	2.804×10^{16}	1942.19	0.340	3.725	3.719
	TCN	$\mathbf{3.259\times10^{16}}$	2004.75	0.359	3.969	4.040
	Transformer	3.247×10^{16}	**2013.76**	0.426	**4.053**	**4.064**

the results indicate that BECON is not tied to a single encoder architecture, while the GRU provides a strong and stable default across tasks.

F.4 History Length Sensitivity

Table 9: History-window ablation by dataset

Dataset	L	FHV ↑	FUT ↑	MC ($\times 10^{-3}$) ↓	SR ↑	PCU ↑
MO-HalfCheetah-2d	0	5.653×10^{6}	2339.17	0.885	4.404	4.339
	2	5.725×10^{6}	2339.04	**0.845**	4.396	4.341
	4	5.676×10^{6}	**2340.38**	0.945	**4.544**	**4.563**
	8	$\mathbf{5.776 \times 10^{6}}$	2332.76	0.938	4.413	4.360
	16	5.763×10^{6}	2336.99	0.931	4.397	4.324
MO-Ant-4d	0	2.346×10^{14}	3876.40	1.816	7.334	7.515
	2	3.385×10^{14}	4270.87	**1.694**	8.340	8.492
	4	$\mathbf{3.687 \times 10^{14}}$	**4355.44**	1.696	**8.600**	**8.651**
	8	3.304×10^{14}	4248.51	1.749	7.982	8.285
	16	3.006×10^{14}	4142.39	1.754	7.943	8.234
MO-HalfCheetah-5d	0	3.325×10^{16}	2369.11	1.515	5.356	5.976
	2	$\mathbf{9.502 \times 10^{16}}$	**2682.32**	1.473	6.502	6.618
	4	7.574×10^{16}	2635.58	1.486	**6.981**	**11.477**
	8	5.762×10^{16}	2250.90	**1.149**	4.159	4.761
	16	8.435×10^{16}	2584.66	1.459	5.915	4.491

Table 9 evaluates the effect of the history window length L. The optimal value of L is task-dependent, reflecting the trade-off between useful temporal evidence and stale task information under drift. On MO-HalfCheetah-2d, $L = 8$ gives the best FHV, while $L = 4$ gives the best FUT, SR, and PCU, suggesting that short-to-moderate histories are sufficient for this lower-dimensional task. On MO-Ant-4d, $L = 4$ achieves the strongest FHV, FUT, SR, and PCU, whereas longer histories degrade performance, indicating that excessively old transitions can hurt adaptation. On MO-HalfCheetah-5d, nonzero history is essential: $L = 2$ yields the best FHV and FUT, while $L = 4$ achieves the best SR and PCU. Overall, these results support the role of history-based belief inference while demonstrating that the history window should remain short enough to avoid staleness under task drift.

F.5 Uncertainty-Gated Fusion Sensitivity

Table 10: Intent-mixing ablation by dataset

Dataset	$\alpha_{\max}$	FHV ↑	FUT ↑	MC ($\times 10^{-3}$) ↓	SR ↑	PCU ↑
MO-HalfCheetah-2d	0.00	5.655×10^{6}	2341.04	0.877	4.410	4.336
	0.25	5.703×10^{6}	2342.70	0.869	4.492	4.467
	0.50	5.694×10^{6}	**2343.06**	0.953	**4.493**	**4.459**
	0.75	$\mathbf{5.724 \times 10^{6}}$	2323.67	0.866	4.256	4.164
	1.00	5.714×10^{6}	2297.31	**0.791**	3.836	3.678
MO-Ant-4d	0.00	3.657×10^{14}	4351.41	1.719	8.457	8.520
	0.25	3.647×10^{14}	4344.34	1.863	8.465	8.639
	0.50	3.504×10^{14}	4300.74	1.734	8.447	8.701
	0.75	3.573×10^{14}	4331.54	1.717	**8.501**	**8.796**
	1.00	$\mathbf{3.684 \times 10^{14}}$	**4362.46**	**1.634**	8.417	8.582
MO-HalfCheetah-5d	0.00	3.548×10^{16}	2378.05	1.603	5.311	6.072
	0.25	7.427×10^{16}	2540.92	1.765	**6.935**	**11.461**
	0.50	$\mathbf{8.552 \times 10^{16}}$	**2555.41**	1.474	5.673	5.709
	0.75	7.622×10^{16}	2418.98	1.179	4.806	3.849
	1.00	0.000	1617.87	**0.210**	1.657	1.684

Table 10 studies the effect of the maximum fusion weight $\alpha_{\max}$, which controls the reliance on belief-derived task estimates. The best setting varies across datasets and metrics. On MO-HalfCheetah-2d, moderate fusion yields the strongest online adaptation, with $\alpha_{\max} = 0.5$ achieving the best FUT, SR, and PCU, while $\alpha_{\max} = 0.75$ gives the highest FHV and $\alpha_{\max} = 1.0$ obtains the lowest MC. On MO-Ant-4d, stronger belief reliance is beneficial for static feasible-front quality and cost control, with $\alpha_{\max} = 1.0$ achieving the best FHV, FUT, and MC, whereas $\alpha_{\max} = 0.75$ gives the best SR and PCU. On MO-HalfCheetah-5d, intermediate fusion is crucial: $\alpha_{\max} = 0.5$ achieves the best FHV and FUT, while $\alpha_{\max} = 0.25$ gives the strongest SR and PCU. Fully relying on inferred task signals can be harmful, as $\alpha_{\max} = 1.0$ collapses FHV on MO-HalfCheetah-5d despite producing the lowest MC. These results support uncertainty-gated fusion: belief-derived estimates improve adaptation and feasible-front quality, but should be balanced with the exposed descriptor to avoid overly conservative or miscalibrated task conditioning.

F.6 Preference Drift Sensitivity

Table 11: Preference-drift ablation by dataset

Dataset	ρ_ω	FHV ↑	FUT ↑	MC ($\times 10^{-3}$) ↓	SR ↑	PCU ↑
MO-HalfCheetah-2d	0.05	5.661×10^{6}	2339.41	0.899	4.219	4.146
	0.10	5.787×10^{6}	**2341.87**	0.826	4.235	4.194
	0.20	$\mathbf{5.790 \times 10^{6}}$	2339.33	**0.818**	4.405	4.349
	0.40	5.728×10^{6}	2338.07	0.875	**4.532**	**4.458**
MO-Ant-4d	0.05	3.680×10^{14}	4354.59	1.641	**8.704**	**9.070**
	0.10	3.407×10^{14}	4271.80	1.630	8.607	8.744
	0.20	$\mathbf{3.744 \times 10^{14}}$	**4369.02**	**1.572**	8.699	8.880
	0.40	3.281×10^{14}	4236.83	1.752	8.240	8.253
MO-HalfCheetah-5d	0.05	7.721×10^{16}	2521.16	1.663	6.588	5.442
	0.10	9.184×10^{16}	2662.65	1.474	**6.797**	**7.346**
	0.20	7.353×10^{16}	2405.22	1.477	4.946	4.347
	0.40	$\mathbf{9.484 \times 10^{16}}$	**2668.46**	**1.433**	5.694	6.199

Table 11 evaluates the sensitivity of BECON to the preference drift rate ρ_ω. The effect of ρ_ω is dataset- and metric-dependent, reflecting a trade-off between task diversity and history staleness. On MO-HalfCheetah-2d, moderate drift $\rho_\omega = 0.20$ gives the best FHV and lowest MC, while stronger drift $\rho_\omega = 0.40$ improves SR and PCU, suggesting that faster preference changes mainly benefit online adaptation. On MO-Ant-4d, $\rho_\omega = 0.20$ achieves the strongest FHV, FUT, and MC, whereas smaller drift $\rho_\omega = 0.05$ yields the best SR and PCU. On MO-HalfCheetah-5d, larger drift $\rho_\omega = 0.40$ gives the best feasible-front quality, while $\rho_\omega = 0.10$ provides the strongest online adaptation. Overall, these results indicate that moderate-to-large preference drift can improve the learning signal for task inference, but excessive or mismatched drift may reduce post-change adaptation quality.

F.7 Context-Dependent Dual Ablation

Table 12: Context-dependent dual ablation

Dataset	Variant	FHV ↑	FUT ↑	SR ↑	PCU ↑	VR ↓	MC ↓
MO-Ant-2d	$\lambda_\psi(\mathbf{z}_t)$	2.24×10^{6}	1.40×10^{3}	1.46	1.49	0	$\mathbf{0.59\times10^{-3}}$
	$\lambda_\psi(\widetilde{B}_t)$	5.76×10^{6}	2.22×10^{3}	3.53	3.51	0	1.21×10^{-3}
	λ	5.78×10^{6}	2.22×10^{3}	4.06	**4.19**	0	1.49×10^{-3}
	$\lambda_\psi(\widetilde{B}_t, \mathbf{z}_t)$	$\mathbf{5.88\times10^{6}}$	$\mathbf{2.25\times10^{3}}$	**4.08**	4.11	0	1.60×10^{-3}
MO-HalfCheetah-2d	$\lambda_\psi(\mathbf{z}_t)$	2.33×10^{6}	1.51×10^{3}	2.09	2.11	0	$\mathbf{0.45\times10^{-3}}$
	$\lambda_\psi(\widetilde{B}_t)$	5.61×10^{6}	2.34×10^{3}	4.48	4.45	0	0.80×10^{-3}
	λ	5.65×10^{6}	2.34×10^{3}	4.50	4.47	0	0.77×10^{-3}
	$\lambda_\psi(\widetilde{B}_t, \mathbf{z}_t)$	$\mathbf{5.66\times10^{6}}$	$\mathbf{2.35\times10^{3}}$	**4.53**	**4.52**	0	0.79×10^{-3}

Table 12 studies the role of the dual-variable parameterization. The full context-dependent dual $\lambda_\psi(\widetilde{B}_t, \mathbf{z}_t)$ achieves the strongest FHV, FUT, and SR on both tasks, and also obtains the best PCU on MO-HalfCheetah-2d. Using only the belief $\mathbf{z}_t$ produces the lowest MC, but substantially degrades FHV, FUT, SR, and PCU, indicating overly conservative constraint enforcement. The budget-only and global dual variants recover much of the performance, but remain less consistent than conditioning jointly on the fused budget and belief. These results support the use of a context-dependent dual variable that adapts the constraint penalty to both feasibility level and latent task context.

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 the main scope of the paper: constrained preference-conditioned MORL under drifting and partially observed preferences and budgets. The claims are supported by the proposed BECON framework, the theoretical error decomposition, and the empirical results in Sections 4-6.

   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: Section 7 discusses the key limitations, including stage-wise constraints, latent-task identifiability, and the scope of task drift.

   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: Section 5 states the theoretical assumptions, with complete proofs and auxiliary results provided 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 6 and Appendices D–F provide the benchmarks, training protocol, evaluation metrics, implementation details, and additional analyses needed for reproducibility.

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 is publicly available, along with instructions for reproducing the experimental results.

   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: Section 6 specifies the experimental protocol, with environment settings in Appendix D and implementation details and hyperparameters in Appendix E.

   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: The main experimental tables report mean and standard deviation over five independent runs, and Figure 2 includes uncertainty bands over seeds. These error bars capture variability due to random initialization and environment sampling.

   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: These details are provided in the appendix.

   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 study is conducted as algorithmic research and conforms to the NeurIPS Code of Ethics.

   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: [N/A]

Justification: This work is primarily a methodological contribution to constrained multi-objective reinforcement learning. Its possible societal effects depend on how future systems instantiate and deploy such algorithms, rather than on the proposed method itself. Since the paper does not introduce a specific real-world deployment, user-facing system, or high-risk data/model release, we do not identify a direct broader impact requiring separate discussion.

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 high-risk generative models, scraped datasets, or assets with substantial misuse risk.

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: The paper properly credits the creators and original owners of all existing benchmarks, baselines, and software assets used in the experiments. We respect the corresponding licenses and terms of use, with attribution and usage details provided in the paper and supplementary material.

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: [N/A]

    Justification: The paper does not introduce a new dataset, benchmark, or pretrained model as a primary asset.

    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, user studies, or experiments with human subjects.

    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, crowdsourcing, personal data collection, or user studies.

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: The proposed method does not use LLMs as an important, original, or non-standard component of the core methodology.

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.